Title

Beyond Standard Assumptions - Semiparametric Models, A Dyadic Item Response Theory Model, and Cluster-Endogenous Random Intercept Models

Author
Sim, Nicholas

Publication Date
2019

Peer reviewed|Thesis/dissertation

Beyond Standard Assumptions - Semiparametric Models, A Dyadic Item Response Theory Model, and Cluster-Endogenous Random Intercept Models

by

Nicholas Sim

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sophia Rabe-Hesketh, Chair
Professor Bruce Fuller
Associate Professor Maya Petersen
Assistant Professor Avi Feller

Fall 2019

Beyond Standard Assumptions - Semiparametric Models, A Dyadic Item Response Theory Model, and Cluster-Endogenous Random Intercept Models

Abstract

Beyond Standard Assumptions - Semiparametric Models, A Dyadic Item Response Theory Model, and Cluster-Endogenous Random Intercept Models

by

Nicholas Sim

Doctor of Philosophy in Education

University of California, Berkeley

Professor Sophia Rabe-Hesketh, Chair

In most statistical analyses, quantitative education researchers often make simplifying assumptions regarding the manner in which their data was generated in order to answer some of these questions. These assumptions can help to reduce the complexity of the problem, and allow the researcher to describe their data using a simpler, and often times more interpretable, statistical model. However, making some of these assumptions when they are not true can lead to biased estimates and misleading answers. While the standard sets of assumptions associated with commonly-used statistical models are usually sufficient in a wide range of contexts, it will always be beneficial for education researchers to understand what they are, when they are reasonable, and how to modify them if necessary.

This dissertation focuses on three of the most common models used in quantitative education research (viz. parametric models like Linear Models (LMs), Item Response Theory (IRT) models, and Random-Intercept Models (RIMs)), discusses the standard sets of assumptions that accompany these models, and then describes related models with less stringent sets of assumptions. In each of the following three chapters, we either explicitly unpack existing models that are useful but are currently still uncommon in the field of education research, or propose novel models and/or estimation strategies for these models.

We begin in Chapter 1 with a common parametric model known as the Gaussian LM, and use it as a scaffold to better understand semiparametric models and their estimation. We begin by reviewing how the coefficients of the Gaussian LM are usually estimated using Maximum Likelihood (ML) or Least-Squares (LS). We then introduce the notion of an m-estimator as well as that of a Regular Asymptotically Linear estimator, and show how they relate to the ML estimator. In particular, we introduce the notion of influence functions/curves and discuss their geometry together with concepts such as Hilbert spaces and tangent spaces. We then demonstrate, concretely, how to derive the so-called efficient influence function under the Gaussian LM, and show that it is precisely the influence function of the ML and

(Ordinary) LS estimators. This shows that the ML estimator (at least under the Gaussian LM) is efficient. Using the foundation built, we move on from the Gaussian LM by relaxing both the assumption that the residuals are normally distributed, as well as the assumption that they have a constant variance, and define this as the Heteroskedastic Linear Model. Unlike the Gaussian LM, this is a semiparametric model. Where possible, we make use of intuition and analogous results from the parametric setting to help describe the workflow for obtaining an efficient estimator for the coefficients of the Heteroskedastic Linear Model. In particular, we derive the nuisance tangent space for this semiparametric model, and use it to obtain the efficient influence function for our model. We then show how to use the efficient influence function to obtain an efficient estimator (which happens to be the Weighted LS estimator) from the (Ordinary) LS estimator via a one-step approach as well as an estimating equations approach. We then conclude by directing readers to more advanced material, including references on more modern approaches to estimating more general semiparametric models such as Targeted Maximum Likelihood Estimation.

In Chapter 2, we focus on a class of measurement models known as Item Response Theory models which are useful for measuring latent traits of a subject based on the subject's response to items. We relax the condition that the responses are only a result of the individual's latent trait (and possibly an external rater), and propose a dyadic Item Response Theory (dIRT) model for measuring interactions of pairs of individuals when the responses to items represent the actions (or behaviors, perceptions, etc.) of each individual (actor) made within the context of a dyad formed with another individual (partner). Examples of its use in education include the assessment of collaborative problem solving among students, or the evaluation of intra-departmental dynamics among teachers. The dIRT model generalizes both Item Response Theory models for measurement and the Social Relations Model for dyadic data. Here, the responses of an actor when paired with a partner are modeled as a function of not only the actor's inclination to act and the partner's tendency to elicit that action, but also the unique relationship of the pair, represented by two directional, possibly correlated, interaction latent variables. We discuss generalizations such as accommodating triads or larger groups, but focus on demonstrating the key idea in the dyadic case. We show that estimation may be performed using Markov-chain Monte Carlo implemented in `Stan`, making it straightforward to extend the dIRT model in various ways. Specifically, we show how the basic dIRT model can be extended to accommodate latent regressions, random effects, distal outcomes. We perform a simulation study that demonstrates that our estimation approach performs well. In the absence of educational data of this form, we demonstrate the usefulness of our proposed approach using speed-dating data instead, and find new evidence of pairwise interactions between participants, describing a mutual attraction that is inadequately characterized by individual properties alone.

Finally, in Chapter 3, we consider the often implicit assumption made when estimating the coefficients of structural Random Intercept Models (RIMs) that covariates at all levels do not co-vary with the random intercepts. A violation of this assumption (called cluster-level endogeneity) leads to inconsistent estimates when using standard estimation procedures. For two-level RIMs with such endogeneity, Hausman and Taylor (HT) devised a consistent multi-step instrumental variable estimator using only internal instruments. We, instead, approach this problem by explicitly modeling the endogeneity using a Structural Equation Model (SEM). In this chapter, we compare, through simulation, the HT and SEM estimators, and evaluate their asymptotic and finite sample properties. We show that the SEM approach is also flexible enough to deal with different exchangeability assumptions for the covariates (e.g., whether the correlations between pairs of all units in a cluster are the same) and investigate how these exchangeability assumptions affect finite sample properties of the HT estimator. For the simulations, we propose a new procedure for generating cluster- and unit-level covariates and random intercepts with a fully flexible covariance structure. We also compare our approach to another common approach known as Multilevel Matching using data from the High School and Beyond survey.

Contents

Acknowledgments

The past four and a half years have been an incredible learning experience for me, and I would like to especially thank my advisor, Sophia Rabe-Hesketh, for her patience, her kindness, and her tutelage during this time. I would also like to thank Anders Skrondal, Bruce Fuller, Maya Petersen, and Avi Feller in particular for their constant encouragement, advice, and belief in me. I am also deeply grateful for the support from family and friends both here in Berkeley as well as back in Singapore.

Lastly, I am thankful for the funding received from the Ministry of Education, Singapore via its Postgraduate Scholarship.

Chapter 1

Using the Linear Model to Understand Semiparametric Efficient Estimation

1.1 Introduction

For many decades, the Linear Model (LM) has served as the workhorse of quantitative research in Education and other Social Science disciplines. The popularity of the LM may, in part, be due to the fact that the model parameters have a simple and clear interpretation, are computationally easy to estimate, and have estimators whose finite-sample and asymptotic behaviors are well-understood. Under some additional assumptions, the LM also may be used in conjunction with the Potential Outcomes (PO) or the Structural Equation Modeling (SEM) frameworks to imbue parameter estimates with a causal interpretation. This has, in particular, been immensely useful to researchers interested in evaluating educational programs, policies and initiatives. However, as with all statistical models, utilizing an LM to analyze data presupposes a set of assumptions on the distribution of the data. The validity and interpretability of the parameter estimates of an LM thus depend very much on how reasonable these assumptions are.

Much like the LM, the underlying theory behind semiparametric models is also well-developed. These models serve as a middle ground between parametric models like the LM, which make assumptions that in certain situations the researcher may not be comfortable with making, and nonparametric models, which make no assumptions on the distribution of the data even when there could be some knowledge available. For example, in the causal inference setting where all confounders are measured, the average causal effect of a treatment on an outcome is identified using the G-computation formula (Robins, 1986) which only uses the conditional distribution of the outcome given the treatment and the covariates, as well as the marginal distribution of the covariates, and ignores the conditional distribution of the treatment given the covariates (otherwise known as the treatment mechanism). Since no assumptions are made regarding these distributions under the nonparametric model, complex density estimators as well as large amounts of data are sometimes required to estimate the causal effect of interest. If there is some knowledge on the treatment mechanism, say for example if the experiment is a Randomized Controlled Trial (RCT), then this knowledge could be incorporated into the model, resulting in a semiparametric model, to improve statistical efficiency.

We posit that the main reason why semiparametric models are not part of a typical education researcher's arsenal lies in their perceived complexity. In this chapter, we aim to utilize the reader's existing knowledge of the LM, as well as Maximum Likelihood (ML) and Least-Squares (LS) theories to try to elucidate some of concepts behind the estimation of parameters in semiparametric models. To this end, we follow the general theory and notation in Tsiatis (2007) closely, but also draw inspiration from Bickel, Klaassen, Ritov, and Wellner (1993) and Newey (1988). Since our main target audience is quantitative education researchers and methodologists, we focus less on mathematical precision and regularity conditions, and more on developing the intuition for these concepts. We refer advanced readers interested in these details to the running example on the restricted moment model in Tsiatis (2007) instead.

We contribute to the vast literature of efficient estimation in semiparametric models not with novel theory, but in directly drawing links of existing theory with the LM as well as ML and LS theories. We also work through proofs for the specific concepts for the case of the LM, which to our knowledge has not been presented concretely, and provide our own analogous derivation of the so-called nuisance tangent space of the Heteroskedastic Linear Model.

The rest of this chapter proceeds as follows. In section 2, we begin by defining the parametric Gaussian Linear Model, and demonstrate how parameter estimates are typically obtained via ML estimation. In section 3, we introduce m-estimators and discuss how these estimators are a class of Regular Asymptotically Linear (RAL) estimators. In section 4, we then introduce the notion of influence functions/curves, Hilbert spaces, tangent spaces, and show how to obtain the efficient influence function/curve using the geometry of influence functions/curves. In section 5, we generalize concepts introduced in section 3 and 4 to semiparametric models, and in particular, show how to derive the nuisance tangent space, and how to obtain efficient estimators of the regression coefficients via both the estimating equation approach as well as the one-step estimator. As some of the results in the semiparametric case are analogous to the parametric case, we will use them directly rather than focus on proving them. Lastly, in section 6, we conclude by directing readers to what they can read next, including modern approaches to estimating more general semiparametric models such as Targeted Maximum Likelihood Estimation.

Throughout this chapter, we will assume that regularity conditions hold as and when we need them to and refer the reader specifically to Newey (1988) for more details. These assumptions include, but are not limited to, the existence of inverses, the existence of moments, smoothness of functions, those required by to interchange derivatives and integrals, etc. We will also ignore any measure-theoretic niceties in our explanation.

For mathematical consistency, we will define all vectors as column vectors. We also define the derivative of a scalar u with respect to a q-dimensional vector $\beta = (\beta_1, \beta_2, \dots, \beta_q)^\intercal$ to be the q-dimensional vector given by

$$\frac{\partial}{\partial \beta} u := \begin{pmatrix} \frac{\partial}{\partial \beta_1} u \\ \frac{\partial}{\partial \beta_2} u \\ \vdots \\ \frac{\partial}{\partial \beta_q} u \end{pmatrix},$$

and the derivative of a q-dimensional vector $\beta = (\beta_1, \beta_2, \dots, \beta_q)^\intercal$ with respect to a scalar u to be the q-dimensional vector given by

$$\frac{\partial}{\partial u} \beta := \begin{pmatrix} \frac{\partial}{\partial u} \beta_1 \\ \frac{\partial}{\partial u} \beta_2 \\ \vdots \\ \frac{\partial}{\partial u} \beta_q \end{pmatrix}.$$

Note that in the cases where we take the derivative of a scalar with respect to the transpose of a vector, and the case where we take the derivative of the transpose of a vector with respect to a scalar, we would obtain results similar to above except that they would be transposed into row vectors instead.

Compatible to the definitions above, we will likewise define the derivative of a q-dimensional vector $\beta = (\beta_1, \beta_2, \dots, \beta_q)^\intercal$ with respect to the transpose of an r-dimensional vector $\eta = (\eta_1, \eta_2, \dots, \eta_r)^\intercal$ to be given by the $q \times r$ matrix

$$\frac{\partial}{\partial \eta^\intercal} \beta := \begin{pmatrix} — \frac{\partial}{\partial \eta^\intercal} \beta_1 — \\ — \frac{\partial}{\partial \eta^\intercal} \beta_2 — \\ \vdots \\ — \frac{\partial}{\partial \eta^\intercal} \beta_q — \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \frac{\partial}{\partial \eta_1} \beta & \frac{\partial}{\partial \eta_2} \beta & \cdots & \frac{\partial}{\partial \eta_r} \beta \\ | & | & & | \end{pmatrix}.$$

1.2 The Gaussian Linear Model and the Maximum Likelihood and Ordinary Least Squares Estimators

For the remainder of this chapter, we consider the scenario where we are interested in a random variable, $Y \in \mathbb{R}$, representing a continuous outcome, and a q-dimensional random vector, $X \in \mathbb{R}^q$, comprising $q-1$ covariates of interest and a constant. We also suppose that we have access to n independent observations $Z_1, Z_2, \ldots, Z_n$ of this $k+1$-tuple $Z = (Y, X^\intercal)^\intercal \in \mathbb{R}^{q+1}$.

1.2.1 The Gaussian Linear Model

For this section, let the Gaussian LM be defined as the set of conditional distributions of an outcome Y given the covariates X with the following assumptions. Given the random vector X, these conditional distributions, written as $p_{Y|X;\beta,\sigma^2}(y \mid x)$, (i) take the form of a normal/Gaussian distribution (commonly known as the "Normality" assumption), (ii) have mean equal to $X^\intercal\beta$ where $\beta \in \mathbb{R}^q$ (commonly called the "Linearity" assumption), and (iii) have variance given by a constant $\sigma^2 \in \mathbb{R}_{\geq 0}$ (commonly called the "Homoscedasticity" assumption). Since all normal distributions are fully characterized by their mean and variance, every distribution in the Gaussian LM is fully characterized by the vector of parameters $\theta = (\beta^\intercal, \sigma^2)^\intercal$. When all distributions in a model can be characterized by a finite set of parameters, we say that the model is parametric. The Gaussian LM is therefore a parametric model that is indexed by $\theta = (\beta^\intercal, \sigma^2)^\intercal$.

We can represent the Gaussian LM as

$$\mathcal{M}_{\text{Gauss}} := \left\{ p_{Y|X;\beta,\sigma^2}(y \mid x) : Y = X^\intercal\beta + \epsilon, \epsilon \mid X \sim \mathcal{N}(0, \sigma^2) \right\}.$$

Implicitly, by assuming the Gaussian LM, we assume that the probability distribution $p_{Y|X;\beta_0,\sigma_0^2}(y \mid x)$ that gave rise to the data (sometimes called the data-generating distribution) for some fixed values of β_0 and σ_0^2, is indeed normal with mean $X^\intercal\beta_0$ and variance σ_0^2. In most situations, we are interested in β_0, and hence, the goal is to now find estimators $\widehat{\beta}$ that have 'good' (asymptotic) properties. The preferred estimator in this setting is the Maximum Likelihood (ML) estimator which is identical to the Ordinary Least-Squares (LS) estimator implemented in most statistical software.

Remark 1 (Fixed/Random Designs). *The typical approach when looking at Gaussian LMs is to consider a fixed design. That is, we treat X as fixed for each observation, and so the randomness in our system only comes from Y. We choose, instead, to consider a random design with X drawn from some distribution to facilitate the discussion of asymptotics.*

Remark 2 (Joint/Conditional Distributions)**.** *In the random design setting, the Gaussian LM is typically defined to be the set of all joint distributions of Y and X (e.g., in Tsiatis (2007)), rather than the set of conditional distributions. In such a situation, the Gaussian LM is actually a Semiparametric Model unless further assumptions are made on the marginal distribution of X. We chose to focus on conditional distributions of Y given X in this section for pedagogical purposes as we in fact do not lose anything by doing so.*

Remark 3 (Expectations)**.** *Throughout the rest of this section, expectations involving X are taken over that true unspecified distribution of X unless stated otherwise. When we write $\mathbb{E}(\cdot)$, we implicitly mean that the expectation is taken over the joint distribution of all random variables within the parentheses, whereas when we write $\mathbb{E}_\theta(\cdot)$ we mean that the expectation is taken only over the conditional distribution of Y given X (i.e., treating X as fixed). In general, we will use the former when evaluating properties of an estimator, and the latter when deriving an estimator.*

1.2.2 Maximum Likelihood Estimation

Suppose we believe that each observation of the data we have at hand is generated first by a draw of X from its distribution, and then by a draw of Y from its conditional distribution given X in $\mathcal{M}_{\text{Gauss}}$ with parameter values $\theta_0 = (\beta_0^\intercal, \sigma_0^2)^\intercal$. Then, we can write

$$Y = X^\intercal \beta_0 + \epsilon$$

where $\epsilon := Y - X^\intercal \beta_0$. Since $\mathbb{E}(\epsilon X) = 0$ by construction, we have, under regularity conditions, that

$$\beta_0 = (\mathbb{E}[XX^\intercal])^{-1}\,\mathbb{E}[XY]. \tag{1.1}$$

Suppose, that β_0 is the parameter that answers our scientific query. Then, σ_0^2 is relevant insofar as it is needed to fully characterize the data-generating distribution, but is of little interest to us. As such, in the Gaussian LM, the parameter σ^2 is called a nuisance parameter. We label nuisance parameters as $\eta = \sigma^2$ following the notation in Tsiatis (2007).

Within the context of a Gaussian LM, one popular method to estimate β_0 is via ML. That is, we can first define the likelihood function

$$\mathcal{L}(\beta, \eta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\eta}} \exp\left(-\frac{(Y_i - X_i^\intercal \beta)^2}{2\eta}\right),$$

and then define the ML estimator $\widehat{\beta}$ as the β that maximizes this function.

To do so, one typically tries to find the β that maximizes the log-likelihood function given below instead:

$$\ell(\theta) = \log \mathcal{L}(\beta, \eta) = \sum_{i=1}^{n} \left[-\log \sqrt{2\pi\eta} - \frac{(Y_i - X_i^\intercal \beta)^2}{2\eta}\right].$$

This is equivalent, in almost all cases, to taking the first derivative of the log-likelihood function, and finding the β such that the first derivative evaluates to zero. That is, we want to find β such that

$$\frac{\partial \ell(\theta)}{\partial \beta} = \frac{1}{\eta} \sum_{i=1}^{n} X_i \left[Y_i - X_i^{\intercal} \beta \right] = \frac{1}{\eta} \sum_{i=1}^{n} X_i \epsilon_i = 0 \in \mathbb{R}^q. \tag{1.2}$$

With a little arithmetic, and some regularity assumptions, the ML estimator for β_0 which solves (1.2) is given by

$$\widehat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i^{\intercal} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right). \tag{1.3}$$

It is worthwhile to note from (1.2) that the ML estimator for β_0 does not depend on knowing or estimating η. As such, the ML estimator is actually identical to the Ordinary Least-Squares estimator which we will elaborate on in subsequent sections.

ML estimators are favored because they have 'good' asymptotic properties. They are consistent, asymptotically normal, and asymptotically efficient. It is instructive at this point to look at the proof for why the ML estimator for β_0 is consistent and asymptotically normal. We will discuss its asymptotic efficiency in a subsequent section.

1.2.2.1 Consistency

To establish consistency of the ML estimator, we first appeal to the Law of Large Numbers to obtain

$$\left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i^{\intercal} \right) \xrightarrow{p} \mathbb{E}[XX^{\intercal}], \text{ and} \tag{1.4}$$

$$\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \xrightarrow{p} \mathbb{E}[XY]. \tag{1.5}$$

Then, under regularity conditions, $\widehat{\beta}$ is consistent by applying the Continuous Mapping Theorem together with (1.4) and (1.5) to (1.1).

1.2.2.2 Asymptotic Normality

To determine the asymptotic distribution of $\widehat{\beta}$, we first consider that

$$\begin{aligned}\widehat{\beta} - \beta_0 &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\intercal\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) - \beta_0 \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\intercal\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i \left(X_i^\intercal \beta_0 + \epsilon_i\right)\right) - \beta_0 \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\intercal\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i \epsilon_i\right).\end{aligned}$$

Looking at the second term on the right-hand side, we then appeal to the Central Limit Theorem to obtain

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i - \mathbb{E}(X\epsilon)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathbb{E}(\epsilon^2 X X^\intercal)\right)$$

since $\mathbb{E}(X\epsilon) = 0$ by construction. Hence, together with (1.4), we see that

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta} - \beta_0\right) &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\intercal\right)^{-1} \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i\right) \\ &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \left(\mathbb{E}[XX^\intercal]\right)^{-1} \mathbb{E}(\epsilon^2 XX^\intercal) \left(\mathbb{E}[XX^\intercal]\right)^{-1}\right) \qquad (1.6)\end{aligned}$$

by Slutsky's Theorem under regularity conditions. We thus also have an expression for the asymptotic variance of the ML estimator. It is useful to point out here that the plug-in estimator for the expression of the asymptotic variance in (1.6) which is given by

$$\left(\frac{1}{n}\sum_{i=1}^{n}[X_i X_i^\intercal]\right)^{-1} \frac{1}{n}\sum_{i=1}^{n}(\widehat{\epsilon}_i^2 X_i X_i^\intercal) \left(\frac{1}{n}\sum_{i=1}^{n}[X_i X_i^\intercal]\right)^{-1}$$

is in fact the Huber-White Sandwich Estimator for the variance of the sampling distribution of $\widehat{\beta}$.

In the case of the Gaussian LM where $\mathbb{E}\left(\epsilon^2 \mid X\right) = \sigma^2$, we see that

$$\mathbb{E}(\epsilon^2 XX^\intercal) = \mathbb{E}\,\mathbb{E}(\epsilon^2 XX^\intercal \mid X) = \sigma^2\,\mathbb{E}(XX^\intercal),$$

which yields a variance formula of $\sigma^2\left(\mathbb{E}[X_i X_i^\intercal]\right)^{-1}$ and a corresponding estimator

$$\widehat{\sigma}^2\left(\frac{1}{n}\sum_{i=1}^{n}[X_i X_i^\intercal]\right)^{-1}$$

which is the familiar "model-based" estimator for the variance of the sampling distribution of $\widehat{\beta}$.

1.3 m-estimators and Regular Asymptotically Linear Estimators

In this section, we introduce the concept of an m-estimator and intuit some of its asymptotic properties. In particular, we demonstrate that it is an asymptotically linear estimator and define the notion of an influence function/curve in that context.

1.3.1 m-estimators

The so-called m-estimators can be thought of as generalizations of ML estimators. In fact, the "m" stands for "Maximum-Likelihood-type". In our setting, suppose we have a p-dimensional function $m(Y, X, \theta)$ of data and parameters such that

$$\mathbb{E}_\theta \, m(Y, X, \theta) = 0 \in \mathbb{R}^p.$$

Then, under some regularity conditions, the solution $\widehat{\theta}$ to the equation

$$\sum_{i=1}^{n} m(Y_i, X_i, \widehat{\theta}) = 0 \in \mathbb{R}^p$$

is defined to be an m-estimator for the true parameter θ_0.

As an example for what the function $m(\cdot)$ may look like, consider the p-dimensional score function defined to be

$$\begin{aligned} S_\theta(Y, X, \theta) :&= \frac{\partial}{\partial \theta} \ell(\theta) \\ &= \frac{\partial}{\partial \theta} \log p_{Y,X;\theta}(Y, X) \\ &= \frac{\frac{\partial}{\partial \theta} p_{Y,X;\theta}(Y, X)}{p_{Y,X;\theta}(Y, X)}. \end{aligned}$$

Then since it can be easily verified that $\mathbb{E}_\theta(S_\theta(Y, X, \theta)) = 0$ under regularity conditions, the solution $\widehat{\theta}$ to the equation

$$\sum_{i=1}^{n} S_\theta(Y_i, X_i, \widehat{\theta}) = 0 \in \mathbb{R}^p \tag{1.7}$$

is an m-estimator. However, by simple arithmetic, we see that any solution that maximizes the log-likelihood function must be a solution to (1.7). Hence, all ML estimators are in fact m-estimators for the special case where $m(Y, X, \theta) \equiv S_\theta(Y, X, \theta)$.

To illustrate this point in our running example concretely, let us consider the score functions for the Gaussian LM with respect to β and η separately as follows. We call $S_\beta(Y_i, X_i, \theta)$ the score with respect to the target parameter, and $S_\eta(Y_i, X_i, \theta)$ the score with respect to the nuisance parameter.

$$\begin{aligned} S_\beta(Y_i, X_i, \theta) &:= \frac{\partial}{\partial \beta} \log p_{Y|X;\beta,\eta}(Y_i, X_i) = \frac{1}{\eta} X_i \left[Y_i - X_i^\intercal \beta\right], \text{ and} \\ S_\eta(Y_i, X_i, \theta) &:= \frac{\partial}{\partial \eta} \log p_{Y|X;\beta,\eta}(Y_i, X_i) = -\frac{1}{2\eta} + \frac{\left(Y_i - X_i^\intercal \beta\right)^2}{2\eta^2}. \end{aligned}$$

Then, the overall score function is given by

$$S_\theta(Y_i, X_i, \theta) = \left(S_\beta(Y_i, X_i, \theta)^\intercal, S_\eta(Y_i, X_i, \theta)^\intercal\right)^\intercal .$$

It is thus clear that the solution to (1.2) corresponds to an m-estimator where $m(\cdot) = S_\theta$ which also solves (1.7). Hence, our ML estimator for β_0 is also an m-estimator.

Remark 4 (Subscripts/Arguments for Score Functions)**.** *In the notation for the score function $S_\theta(Y_i, X_i, \theta)$, we distinguish using θ as a subscript to indicate which parameters the score function is for and using θ as an argument to indicate at what parameter value θ the score function is evaluated.*

1.3.2 Consistency, Asymptotic Linearity and Asymptotic Normality

In much the same way as for ML Estimators, m-estimators can be shown to be consistent. As such, we will omit the proof here.

Additionally, m-estimators, and as a consequence ML Estimators, are also asymptotically linear. That is, there exists a random function $\varphi(Y, X) = \varphi_{\widehat{\beta}}(Y, X, \theta_0) \in \mathbb{R}^q$ known as the influence function/curve of the estimator $\widehat{\beta}$ such that $\mathbb{E}[\varphi(Y, X)] = 0 \in \mathbb{R}^q$ and

$$\left(\widehat{\beta} - \beta_0\right) = \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i, X_i) + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{1.8}$$

That is, as n gets large, we can approximate the difference between the estimator and the parameter of interested as an empirical mean of its influence function evaluated at all data points. The remainder term $o_p\left(\frac{1}{\sqrt{n}}\right)$ diminishes to 0 very quickly as n increase. This is sometimes referred to as the $\sqrt{n}$-rate.

Remark 5 (Subscripts/Arguments for Influence Functions)**.** *When we suppress the subscript $\widehat{\beta}$ and the argument θ_0, we mean that in influence function pertains to our estimator for β evaluated at the true vector of parameters θ_0.*

From here, it is clear from the Central Limit Theorem that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi(Y_i,X_i)=\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\varphi(Y_i,X_i)-\mathbb{E}[\varphi(Y,X)]\right]\xrightarrow{\mathcal{D}}\mathcal{N}\left(0,\mathbb{E}[\varphi(Y,X)\varphi^{\intercal}(Y,X)]\right)$$

and as such, from (1.8) and by Slutsky's Theorem, we have

$$\sqrt{n}\left(\widehat{\beta}-\beta_0\right)\xrightarrow{\mathcal{D}}\mathcal{N}\left(0,\mathbb{E}[\varphi(Y,X)\varphi^{\intercal}(Y,X)]\right).$$

That is, we have established that m-estimators are asymptotically normal with asymptotic variance given by $\frac{1}{n}\,\mathbb{E}[\varphi(Y,X)\varphi^{\intercal}(Y,X)]$.

All that is left is to show is that m-estimators are in fact asymptotically linear, which we will do for the Gaussian LM case. To do so, and to find the influence function of our estimator, we start by computing the derivative $S_{\theta\theta}(Y,X,\theta_0)$ of the score function component-wise. To make notations compact, we write

$$\begin{aligned}
S_\beta\equiv S_\beta(Y,X,\theta_0)&=\frac{1}{\eta_0}X\left[Y-X^{\intercal}\beta_0\right]=\frac{1}{\eta_0}X\epsilon\\
S_\eta\equiv S_\eta(Y,X,\theta_0)&=-\frac{1}{2\eta_0}+\frac{\left(Y-X^{\intercal}\beta_0\right)^2}{2\eta_0^2}=-\frac{1}{2\eta_0}+\frac{\epsilon^2}{2\eta_0^2}\\
S_\theta\equiv S_\theta(Y,X,\theta_0)&=\left(S_\beta^{\intercal},S_\eta\right)^{\intercal}
\end{aligned}$$

Remark 6 (Further Notation for Score Functions)**.** *When we suppress the arguments of the score functions, we implicitly mean that the score function for the random variable Y and the random vector X is evaluated at the vector of true parameters θ_0.*

By straightforward computations, we have

$$\begin{aligned}
S_{\beta\beta}\equiv\frac{\partial}{\partial\beta^{\intercal}}S_\beta&=-\frac{1}{\eta_0}XX^{\intercal},\\
S_{\beta\eta}\equiv\frac{\partial}{\partial\eta}S_\beta&=-\frac{1}{\eta_0^2}X\left[Y-X^{\intercal}\beta_0\right]=-\frac{1}{\eta_0^2}X\epsilon,\\
S_{\eta\beta}\equiv\frac{\partial}{\partial\beta^{\intercal}}S_\eta&=-\frac{1}{\eta_0^2}X^{\intercal}\left[Y-X^{\intercal}\beta_0\right]=-\frac{1}{\eta_0^2}X^{\intercal}\epsilon,\\
S_{\eta\eta}\equiv\frac{\partial}{\partial\eta}S_\eta&=\frac{1}{2\eta_0^2}-\frac{\left(Y-X^{\intercal}\beta_0\right)^2}{\eta_0^3}=\frac{1}{2\eta_0^2}-\frac{\epsilon^2}{\eta_0^3}.
\end{aligned}$$

And so, we have

$$S_{\theta\theta}\equiv S_{\theta\theta}(Y,X,\theta_0)=\frac{\partial}{\partial\theta^{\intercal}}S_\theta=\begin{pmatrix}S_{\beta\beta}&S_{\eta\beta}\\S_{\beta\eta}&S_{\eta\eta}\end{pmatrix}.$$

Since the m-estimator $\widehat{\theta}$ (and equivalently, the ML estimator) solves the score equation (1.7), we can consider the Mean Value expansion of the sum of the scores to achieve

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left\{ S_\theta(Y_i, X_i, \widehat{\theta}) \right\} \\ &= \sum_{i=1}^{n} \left\{ S_\theta(Y_i, X_i, \theta_0) + S_{\theta\theta}(Y_i, X_i, \theta^*) \left[\widehat{\theta} - \theta_0 \right] \right\} \end{aligned}$$

for some θ^* that lies between θ and $\widehat{\theta}$ component-wise.

By extracting only the first q components which correspond to the parameter β, we have

$$0 = \sum_{i=1}^{n} \left\{ S_\beta(Y_i, X_i, \theta_0) + S_{\beta\beta}(Y_i, X_i, \theta^*) \left[\widehat{\beta} - \beta_0 \right] \right\}$$

Assuming regularity, rearranging the equation above yields

$$\begin{aligned} \widehat{\beta} - \beta_0 &= \left\{ -\frac{1}{n} \sum_{i=1}^{n} S_{\beta\beta}(Y_i, X_i, \theta^*) \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} S_\beta(Y_i, X_i, \theta_0) \right\} \\ &= \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\eta^*} X_i X_i^\intercal \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\eta_0} X_i \left[Y_i - X_i^\intercal \beta_0 \right] \right\} \end{aligned}$$

Under regularity conditions (where there is uniform convergence of all θ to θ_0 within a neighborhood of the true value containing θ^*), we have

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\eta^*} X_i X_i^\intercal \right\}^{-1} \xrightarrow{p} \left\{ \mathbb{E}\left[\frac{1}{\eta_0} X X^\intercal \right] \right\}^{-1}.$$

And so,

$$\sqrt{n}\left(\widehat{\beta} - \beta_0 \right) = \left\{ \mathbb{E}\left[\frac{1}{\eta_0} X X^\intercal \right] \right\}^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{\eta_0} X_i \left[Y_i - X_i^\intercal \beta_0 \right] \right\} + o_p(1).$$

Hence, we have shown that $\widehat{\beta}$ is an asymptotically linear estimator with influence function equal to

$$\begin{aligned} \varphi(Y_i, X_i) &= \left\{ \mathbb{E}\left[\frac{1}{\eta_0} X X^\intercal \right] \right\}^{-1} \frac{1}{\eta_0} X_i \left[Y_i - X_i^\intercal \beta_0 \right] \\ &= \left\{ \mathbb{E}\left[X X^\intercal \right] \right\}^{-1} X_i \left[Y_i - X_i^\intercal \beta_0 \right]. \end{aligned} \tag{1.9}$$

Remark 7 (DFBETA)**.** *It turns out that the estimated influence function evaluated at a point (Y_i, X_i) is in fact similar to a quantity commonly used to measure the influence of the data point (Y_i, X_i) on the estimator $\widehat{\beta}$ - the DFBETA (unscaled version). The DFBETA and the estimated influence functions for observation i can be expressed as*

$$\begin{aligned} DFBETA_i = \widehat{\beta} - \widehat{\beta}_{[-i]} &= \left(\sum_{i=1}^{n} X_i X_i^{\intercal}\right)^{-1} X_i \frac{\widehat{\epsilon}_i}{1-h_{ii}} = \left(\sum_{i=1}^{n} X_i X_i^{\intercal}\right)^{-1} X_i \widehat{\epsilon}_{[-i]} \\ \widehat{\varphi}(Y_i, X_i) &= \left(\sum_{i=1}^{n} X_i X_i^{\intercal}\right)^{-1} X_i \frac{\widehat{\epsilon}_i}{n}, \end{aligned}$$

where $\widehat{\beta}_{[-i]} = \left(\frac{1}{n}\sum_{j=1,j\neq i}^{n} X_j X_j^{\intercal}\right)^{-1} \left(\frac{1}{n}\sum_{j=1,j\neq i}^{n} X_j Y_j\right)$, $\widehat{\epsilon}_{[-i]} = Y_i - X_i^{\intercal}\widehat{\beta}_{[-i]}$ and $\widehat{\epsilon}_i = Y_i - X_i^{\intercal}\widehat{\beta}$. Other than the additional scaling by the size of the sample, the main difference between these two quantities is the choice of residuals. Where the estimated influence function is evaluated using the residual with $\widehat{\beta}$ as the estimator for β, the DFBETA is evalued using $\frac{\widehat{\epsilon}_i}{1-h_{ii}} = \widehat{\epsilon}_{[-i]}$ which is the residual with $\widehat{\beta}_{[-i]}$ as the estimator for β.

Correspondingly, the asymptotic variance of $\widehat{\beta}$ is given by

$$\begin{aligned} &\left\{\mathbb{E}\left[XX^{\intercal}\right]\right\}^{-1} \mathbb{E}\left(X\left[Y - X^{\intercal}\beta_0\right]^2 X^{\intercal}\right) \left\{\mathbb{E}\left[XX^{\intercal}\right]\right\}^{-1} \\ =&\left\{\mathbb{E}\left[XX^{\intercal}\right]\right\}^{-1} \mathbb{E}\left[\epsilon^2 XX^{\intercal}\right] \left\{\mathbb{E}\left[XX^{\intercal}\right]\right\}^{-1}, \end{aligned}$$

which is exactly the same as the asymptotic variance computed in (1.6). We can also derive exactly the same influence functions and asymptotic variance using the general formula given in (1.11) and (1.12) below.

Some General Results. In general, we state here, without proof, that for independent and identically distributed points $Z_1, Z_2, \ldots, Z_n$, any m-estimator with respect to the function $m(Z, \theta_0)$ has an influence function evaluated at each point given by

$$\varphi(Z_i) = -\left(\mathbb{E}\left[\frac{\partial m(Z, \theta_0)}{\partial \theta^{\intercal}}\right]\right)^{-1} m(Z_i, \theta_0), \tag{1.10}$$

and hence, have asymptotic variance given by

$$\left(\mathbb{E}\left[\frac{\partial m(Z, \theta_0)}{\partial \theta^{\intercal}}\right]\right)^{-1} \mathbb{E}\left[m(Z_i, \theta_0) m^{\intercal}(Z_i, \theta_0)\right] \left(\mathbb{E}\left[\frac{\partial m(Z, \theta_0)}{\partial \theta^{\intercal}}\right]\right)^{-1\intercal}$$

It is also useful to point out here that if $m(Z, \theta) \equiv S_\theta(Z, \theta)$, then the influence function of the point Z_i is given by the expected Fisher Information Matrix multiplied by the score evaluated at the point Z_i. That is,

$$\varphi(Z_i) = -\left(\mathbb{E}\left[S_{\theta\theta}(Z, \theta_0)\right]\right)^{-1} S_\theta(Z_i, \theta_0), \tag{1.11}$$

where $S_{\theta\theta}(Z, \theta_0) = \frac{\partial}{\partial \theta^{\intercal}} S_\theta(Z, \theta_0) = \frac{\partial^2}{\partial \theta \partial \theta^{\intercal}} \log p_{Z;\theta}(Z)$, yielding an asymptotic variance of

$$\begin{aligned}&\left(\mathbb{E}\left[S_{\theta\theta}(Z, \theta_0)\right]\right)^{-1} \mathbb{E}\left[S_\theta(Z, \theta_0) S_\theta^{\intercal}(Z, \theta_0)\right] \left(\mathbb{E}\left[S_{\theta\theta}(Z, \theta_0)\right]\right)^{-1} \\ &= -\left(\mathbb{E}\left[S_{\theta\theta}(Z, \theta_0)\right]\right)^{-1}, \end{aligned} \tag{1.12}$$

where the equality is due to a commonly know fact under regularity conditions that

$$-\mathbb{E}\left[S_{\theta\theta}(Z, \theta_0)\right] = \mathbb{E}\left[S_\theta(Z, \theta_0) S_\theta^{\intercal}(Z, \theta_0)\right]. \tag{1.13}$$

Here, readers familiar with ML theory will recognize that (1.12) is in fact the Cramer-Rao Lower Bound, indicating that the ML estimator is in fact asymptotically efficient, a result which we will verify by finding the so-called efficient influence function later.

We consider m-estimators because they belong to the class of Regular Asymptotically Linear (RAL) estimators (where regularity is a technical condition that excludes super-efficient estimators). RAL estimators also come about by applying the functional delta-method, under some regularity conditions, to functionals of estimators whose influence functions are known. One example of this is the so-called substitution or plug-in estimator where for a parameter $f(P)$ that is a functional $f(\cdot)$ of the data generating distribution P, the plug-in estimator is simply the functional $f(\cdot)$ applied to the empirical distribution $\widehat{P}$ of the data. The estimator in (1.3) is precisely such an estimator.

1.4 Finding the Efficient Influence Function

In this section, with the goal of finding the 'best' influence function within a certain class of influence function, we introduce, within the context of a parametric Gaussian LM, concepts such as tangent spaces that are relevant to semiparametric efficient estimation. We aim to build intuition for these concepts in this section as the results largely hold true for semiparametric models in the next section even though the proofs thereafter require a little more technical finesse which we will avoid presenting.

1.4.1 Influence Functions

It is worthwhile to point out here, without proof, that every asymptotically linear estimator has a unique influence function (almost surely). This means that once we are in possession of an asymptotically linear estimator, we can assess its asymptotic properties by looking at its influence function. Conversely, once we know the influence function of an estimator, we can attempt to recover what that estimator actually is.

One important problem in statistics is to identify an estimator that is 'best' in some predefined sense among a specific class of estimators. If we only consider RAL estimators, we can then define the 'best' estimator among this class of estimators to be the one that has the smallest asymptotic variance[1]. We call such an estimator the asymptotically efficient RAL estimator. Since the asymptotic variance of an asymptotically linear estimator is precisely the variance of its influence function, our problem reduces to finding the influence function with the smallest variance which we will call the Efficient influence function. Recovering an estimator that has the efficient influence function as its influence function may be done by the One-Step Estimator or Estimating Equations approaches (Bickel et al., 1993; Liang & Zeger, 1986; Tsiatis, 2007). One of the more modern approaches for recovering an efficient estimator using the Efficient Influence Function is via the Targeted Maximum Likelihood Estimation approaches (van der Laan & Rose, 2011) which we encourage the reader to explore, but leave out of the chapter for brevity.

In order to find the efficient influence function, we first introduce some brief concepts about the relationship between influence functions and score functions, as well as their geometry.

[1] We say that a q-dimensional estimator $\widehat{\beta}$ has smallest asymptotic variance avar $\left(\widehat{\beta}\right)$ if for any other estimator $\widetilde{\beta}$, we have avar $\left(\widehat{\beta}\right)$ − avar $\left(\widetilde{\beta}\right) \preceq 0$. That is, the difference between their Variance-Covariance matrices is negative semi-definite.

1.4.2 The Geometry of Influence Functions and Scores

Some General Results. In a model indexed by $\theta = (\beta^\intercal, \eta)^\intercal \in \mathbb{R}^{q+r}$, $\varphi(Y, X)$ is the influence function of an RAL estimator $\widehat{\beta}$ for the parameter β if and only if it satisfies the following identities:

$$\mathbb{E}\left[\varphi(Y, X) S_\beta^\intercal(Y, X, \theta_0)\right] = I \in \mathbb{R}^{q \times q} \tag{1.14}$$

$$\mathbb{E}\left[\varphi(Y, X) S_\eta^\intercal(Y, X, \theta_0)\right] = 0 \in \mathbb{R}^{q \times r} \tag{1.15}$$

Instead of a general proof, we show why (1.14) and (1.15) hold in our example. First, we note that the score functions of our model evaluated at the true parameter vector θ_0 satisfy (1.13). Rearranging, we have

$$-\left[\mathbb{E}\left[S_{\theta\theta}(Y, X, \theta_0)\right]\right]^{-1} \mathbb{E}\left[S_\theta(Y, X, \theta_0) S_\theta^\intercal(Y, X, \theta_0)\right] = I \in \mathbb{R}^{(q+r)\times(q+r)}. \tag{1.16}$$

Separately, we also know that the influence function of the m-estimator $\widehat{\theta}$ for the parameter θ_0 is given by (1.11). That is,

$$\varphi_{\widehat{\theta}}(Y, X) = -\left(\mathbb{E}\left[S_{\theta\theta}(Y, X, \theta_0)\right]\right)^{-1} S_\theta(Y, X, \theta_0)$$

Post-multiplying this by $S_\theta^\intercal(Z_i, \theta_0)$, and then taking expectations, we have

$$\mathbb{E}\left[\varphi_{\widehat{\theta}}(Y, X) S_\theta^\intercal(Y, X, \theta_0)\right] = -\left(\mathbb{E}\left[S_{\theta\theta}(Y, X, \theta_0)\right]\right)^{-1} \mathbb{E}\left[S_\theta(Y, X, \theta_0) S_\theta^\intercal(Y, X, \theta_0)\right]. \tag{1.17}$$

Notice that the left-hand side of (1.16) is identical to the right-hand side of (1.17). Equating the two equations yields

$$\mathbb{E}\left[\varphi_{\widehat{\theta}}(Y, X) S_\theta^\intercal(Y, X, \theta_0)\right] = I \in \mathbb{R}^{(q+r)\times(q+r)}.$$

By expressing the equation above as block matrices,

$$\begin{pmatrix} \mathbb{E}\left[\varphi_{\widehat{\beta}}(Y, X) S_\beta^\intercal(Y, X, \theta_0)\right] & \mathbb{E}\left[\varphi_{\widehat{\beta}}(Y, X) S_\eta^\intercal(Y, X, \theta_0)\right] \\ \mathbb{E}\left[\varphi_{\widehat{\eta}}(Y, X) S_\beta^\intercal(Y, X, \theta_0)\right] & \mathbb{E}\left[\varphi_{\widehat{\eta}}(Y, X) S_\eta^\intercal(Y, X, \theta_0)\right] \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

we see that the conditions in (1.14) and (1.15) must hold for any RAL estimator by looking at the upper-left and lower-left blocks.

With necessary and sufficient conditions in hand for a random function with zero mean and finite variance to be an influence functions, we can try to better understand their properties and how they relate to each other by studying their geometry.

1.4.3 The Hilbert Space of Random Functions

We begin by considering the space $\mathcal{H}$ of all q-dimensional measurable random functions of (Y, X) with zero mean and finite variance. It is useful to think of this space in the same way one would think of a three-dimensional Euclidean space where each element of the space is a point defined by a 3-dimensional vector from the origin to the point. In $\mathcal{H}$, each element just represents a random function.

In Euclidean space $\mathbb{R}^n$, there is also the concept of an inner product, $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$, of two points u and v in that space which is defined to be the sum of the coordinate-wise products of the two points. While this inner product does satisfy some additional criteria, the most important property is that it also induces the concept of orthogonality and of distance. In Euclidean space, the inner product of two orthogonal vectors is zero, and the inner product of a vector with itself gives the squared length of that vector or, in other words, the squared distance of that point from the origin. That is, $\langle u, v \rangle_{\mathbb{R}^n} = 0$ if $u \perp v$, and $\langle u, u \rangle_{\mathbb{R}^n} = \|u\|_2^2$ for all $u \in \mathbb{R}^n$.

It turns out that in $\mathcal{H}$, for any two random functions of Z, we can define the inner product to be the expectation of the sum of the component-wise product of the two functions. That is, for $f, g \in \mathcal{H}$, we have $\langle f, g \rangle_{\mathcal{H}} = \mathbb{E}\left(f^\intercal(Y, X) g(Y, X)\right)$. Since these are all random functions with zero mean, the inner product thus corresponds to the sum of the component-wise covariances between the random functions f and g. Like in the case of Euclidean space, two functions in $\mathcal{H}$ are said to be orthogonal if their inner product is zero. Also interestingly, the inner product $\langle f, f \rangle_{\mathcal{H}} = \mathbb{E}\left(f^\intercal(Y, X) f(Y, X)\right)$ of a random function with itself yields the sum of the component-wise variances of the random function. This is equivalent to the trace of the variance-covariance matrix of the random function given by the outer product $\mathbb{E}\left(f(Y, X) f^\intercal(Y, X)\right)$. Finally, the inner product also allows us to define the notion of a unique orthogonal projection of the random function $f \in \mathcal{H}$ onto a subspace $S \subset \mathcal{H}$ spanned by the r-dimensional function g. We denote this projection as $\Pi(f \mid S)$ and it takes the form

$$\Pi(f \mid S) = \left(\mathbb{E}\, f g^\intercal\right) \left(\mathbb{E}\, g g^\intercal\right)^{-1} g. \tag{1.18}$$

This is similar in spirit to projections in Euclidean space where we want to project a vector u onto the line spanned by the vector v. Since the resulting projection $\Pi(u \mid Bv)$ must be a scalar multiple of v for a unique scalar B_0, the projection can be represented by $B_0 v$. Furthermore, the residual from the projecion, given by $u - B_0 v$ must then be orthogonal to the all vectors on the line. Hence, we have that

$$\langle u - B_0 v, Bv \rangle_{\mathbb{R}^n} = 0$$

for all $B \in \mathbb{R}$. In order to find a formula for the projection, we need only solve for B_0 to get

$$\Pi(u \mid Bv) = B_0 v = \langle u, v \rangle_{\mathbb{R}^n} \langle v, v \rangle_{\mathbb{R}^n}^{-1} v.$$

The observant reader will notice that the projection formula here consists of inner products, whereas what is provided in (1.18) involves outer products. They may also notice that in solving for B_0, the choice of B does not play any role.

We now contrast this with the case in our Hilbert space $\mathcal{H}$ where we want to project a function f onto a space S spanned by g. Like the Euclidean case, since the projection of f onto S must lie on S, we can represent the projection as $\Pi(f \mid S) = B_0 g$ for some scalar $q \times r$ matrix B_0. The projection residual, given by $f - B_0 g$ must also be orthogonal to the space S and hence,

$$\langle f - B_0 g, Bg \rangle_{\mathcal{H}} = 0$$

for all $B \in \mathbb{R}^{q \times r}$. Unlike the Euclidean case, B now plays an important role since it cannot easily be eliminated from the equation, and it turns out that the statement above being true for all B is equivalent to (1.18) being true. We refer the reader to Tsiatis (2007) for more details on the derivation, but will state that this discrepancy also arises because of how we define the term "span" in these two spaces. In $\mathbb{R}^n$, we usually consider spaces spanned by vectors in $\mathbb{R}^n$ (i.e., our generator $v \in \mathbb{R}^n$). However, in our Hilbert space $\mathcal{H}$, notice that S is spanned by g which is an r-dimensional function rather than a q-dimensional function, and hence does not have to be in $\mathcal{H}$. All g provides, is a set of r unidimensional functions of which each of the q components of an element in S is a linear combination of. We describe more about the spaces spanned by functions in the next subsection.

With these tools, and knowing that the necessary and sufficient conditions (1.14) and (1.15) for a random function in $\mathcal{H}$ to be the influence function of an RAL estimator depend on its relationship with the score functions, we can try to identify the class of these influence functions by exploring the geometry of the spaces spanned by these score functions.

1.4.4 Tangent Spaces

For the parameter $\theta = (\beta^\intercal, \eta^\intercal)^\intercal \in \mathbb{R}^{q+r}$, we define the subspace $\mathcal{T}$ of $\mathcal{H}$ spanned by the score function S_θ to be

$$\mathcal{T} := \left\{ B_\theta S_\theta \mid B_\theta \in \mathbb{R}^{q \times (q+r)} \right\}$$

We can interpret this space to contain all random q-dimensional functions with each component being linear combination of the $(q+r)$ components of the score function S_θ. We call this subspace the tangent space. The space of functions that are orthogonal to all functions in $\mathcal{T}$ is called the orthogonal compliment of $\mathcal{T}$ which we represent by $\mathcal{T}^\perp$.

Since $S_\theta = \left(S_\beta^\intercal, S_\eta\right)^\intercal$, we can likewise define two other subspaces similarly as follows:

$$\begin{aligned} \mathcal{T}_\beta &:= \left\{ B_\beta S_\beta \mid B_\beta \in \mathbb{R}^{q \times q} \right\} \\ \Lambda &:= \left\{ B_\eta S_\eta \mid B_\eta \in \mathbb{R}^{q \times r} \right\} \end{aligned}$$

In particular, we call $\mathcal{T}_\beta$ and Λ the tangent space of the target parameter and the nuisance tangent space respectively, because they are spanned by the score function of the target parameter and the nuisance parameter respectively.

In these tangent spaces, each of the q components of any random function can be thought of as being themselves a linear combination of the relevant score function. As such, the tangent spaces are therefore called q-replicating linear spaces. It turns out that this is important because a well-known fact that the variance of the sum of two univariate functions f and g that are orthogonal to each other is equal to the sum of their individual variances, can be extended for q-dimensional functions if they lie in a q-replicating linear space (see Tsiatis (2007) for more details). In short, if $f \in \mathcal{T}$ and $g \in \mathcal{T}^\perp$, then

$$\mathrm{Var}(f+g) = \mathrm{Var}(f) + \mathrm{Var}(g) \tag{1.19}$$

where we note that here, $\mathrm{Var}(\cdot)$ represents the Variance-Covariance matrix of the random functions. This will become important when we try to find the influence function of an RAL estimator (and by extension, the estimator itself) that has the lowest variance among all other influence functions of RAL estimators.

In our running example, the tangent spaces take the forms

$$\begin{aligned}
\mathcal{T}_\beta &= \left\{ B_\beta \frac{1}{\eta_0} X\epsilon \,\middle|\, B_\beta \in \mathbb{R}^{q\times q} \right\} \\
\Lambda &= \left\{ B_\eta \left(-\frac{1}{2\eta_0} + \frac{\epsilon^2}{2\eta_0^2} \right) \middle|\, B_\eta \in \mathbb{R}^{q\times 1} \right\} \\
\mathcal{T} &= \left\{ B_\beta \frac{1}{\eta_0} X\epsilon + B_\eta \left(-\frac{1}{2\eta_0} + \frac{\epsilon^2}{2\eta_0^2} \right) \middle|\, B_\beta \in \mathbb{R}^{q\times q}, B_\eta \in \mathbb{R}^{q\times 1} \right\}
\end{aligned}$$

1.4.5 The Linear Variety of Influence Functions

Now suppose we are given an initial RAL estimator $\widehat{\beta}^*$ with influence function $\varphi^*(Y,X)$. Then, it is easy to check that for any random function $h(Y,X) \in \mathcal{T}^\perp$, the random function given by $\varphi^*(Y,X)+h(Y,X)$ also satisfies conditions (1.14) and (1.15), i.e., $\varphi^*(Y,X)+h(Y,X)$ is also an influence function. Conversely, it is also easy to check that given any two RAL estimators with influence functions $\varphi^{(1)}(Y,X)$ and $\varphi^{(2)}(Y,X)$, we have that $\varphi^{(1)}(Y,X) - \varphi^{(2)}(Y,X) \in \mathcal{T}^\perp$. This implies that if one has an initial RAL estimator with influence function $\varphi^*(Y,X)$, the set containing all possible influence functions of RAL estimators is given by

$$\varphi^*(Y,X) + \mathcal{T}^\perp = \left\{ \varphi^*(Y,X) + h(Y,X) \mid h(Y,X) \in \mathcal{T}^\perp \right\}.$$

$\varphi^*(Y,X) + \mathcal{T}^\perp$ while not a subspace, since it does not contain zero, is called a linear variety.

1.4.6 Finding the Efficient Influence Function

Suppose we had an RAL estimator and its associated influence function $\varphi^*(Y, X)$. Then, since the projection $\Pi(\varphi^*(Y, X) \mid \mathcal{T})$ can be rewritten as $\varphi^*(Y, X) - \Pi(\varphi^*(Y, X) \mid \mathcal{T}^\perp)$, and $-\Pi(\varphi^*(Y, X) \mid \mathcal{T}^\perp) \in \mathcal{T}^\perp$, we see that it is a valid influence function since it belongs to the linear variety of influence functions.

Let $\varphi_{\text{eff}}(Y, X) := \Pi(\varphi^*(Y, X) \mid \mathcal{T})$. Then, any influence function $\varphi(Y, X)$ can be written as $\varphi(Y, X) = \varphi_{\text{eff}}(Y, X) + h(Y, X)$ where $h(Y, X) \in \mathcal{T}^\perp$. However, since $\varphi_{\text{eff}}(Y, X) \in \mathcal{T}$ which is a q-replicating space, and $\varphi_{\text{eff}}(Y, X) \perp h(Y, X)$, we know from (1.19) that

$$\operatorname{Var}(\varphi(Y, X)) = \operatorname{Var}(\varphi_{\text{eff}}(Y, X)) + \operatorname{Var}(h(Y, X)).$$

And so,

$$\operatorname{Var}(\varphi_{\text{eff}}(Y, X)) - \operatorname{Var}(\varphi(Y, X)) = -\operatorname{Var}(h(Y, X)) \preceq 0,$$

which shows that $\varphi_{\text{eff}}(Y, X)$ has the smallest variance out of all influence functions of RAL estimators. We call $\varphi_{\text{eff}}(Y, X)$ the efficient influence function. The argument above critically works only for $\varphi_{\text{eff}}(Y, X)$ because it lies in $\mathcal{T}$.

It is useful to point out here that since the starting choice of $\varphi^*(Y, X)$ is arbitrary, the projection of any influence function onto $\mathcal{T}$ is unique. This means that the efficient influence function is the only influence function that lies on $\mathcal{T}$.

In our running example, the efficient influence function can be computed using the projection formula in (1.18) as follows. First we note that

$$\varphi_{\text{eff}}(Y, X)) = \Pi(\varphi^*(Y, X) \mid \mathcal{T}) = (\mathbb{E}\, \varphi^* S_\theta^\intercal)\, (\mathbb{E}\, S_\theta S_\theta^\intercal)^{-1}\, S_\theta,$$

where we write $\varphi^* \equiv \varphi^*(Y, X)$.

Looking at the parts separately, we see that

$$\mathbb{E}\, \varphi^* S_\theta^\intercal = \mathbb{E}\, \varphi^* \begin{bmatrix} S_\beta^\intercal & S_\eta \end{bmatrix} = \begin{bmatrix} \mathbb{E}\, \varphi^* S_\beta^\intercal & \mathbb{E}\, \varphi^* S_\eta \end{bmatrix} = \begin{bmatrix} I^{q\times q} & 0^{q\times 1} \end{bmatrix} \in \mathbb{R}^{q\times(q+1)}$$

from conditions (1.14) and (1.15) of influence functions.

Next, we see that

$$\begin{aligned}
\left(\mathbb{E}\, S_\theta S_\theta^\intercal\right)^{-1} &= -\left(\mathbb{E}\, S_{\theta\theta}\right)^{-1} \\
&= -\left(\mathbb{E}\begin{bmatrix} S_{\beta\beta} & S_{\eta\beta} \\ S_{\beta\beta} & S_{\beta\eta} \end{bmatrix}\right)^{-1} \\
&= \begin{bmatrix} \mathbb{E}\left[\frac{1}{\eta_0} X X^\intercal\right] & \mathbb{E}\left[\frac{1}{\eta_0^2} X \epsilon\right] \\ \mathbb{E}\left[\frac{1}{\eta_0^2} X^\intercal \epsilon\right] & \mathbb{E}\left[-\frac{1}{2\eta_0^2} + \frac{\epsilon^2}{\eta_0^3}\right] \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \left(\mathbb{E}\left[\frac{1}{\eta_0} X X^\intercal\right]\right)^{-1} & 0 \\ 0 & \left(\mathbb{E}\left[\frac{1}{2\eta_0^2}\right]\right)^{-1} \end{bmatrix}
\end{aligned}$$

Finally, putting the pieces together, we have

$$\begin{aligned}
\varphi_{\text{eff}}(Y, X) &= \begin{bmatrix} I^{q\times q} & 0^{q\times 1} \end{bmatrix} \begin{bmatrix} \left(\mathbb{E}\left[\frac{1}{\eta_0} X X^\intercal\right]\right)^{-1} & 0 \\ 0 & \left(\mathbb{E}\left[\frac{1}{2\eta_0^2}\right]\right)^{-1} \end{bmatrix} \begin{bmatrix} \frac{1}{\eta_0} X \left[Y - X^\intercal \beta_0\right] \\ -\frac{1}{2\eta_0} + \frac{(Y - X^\intercal\beta_0)^2}{2\eta_0^2} \end{bmatrix} \\
&= \left(\mathbb{E}\left[\frac{1}{\eta_0} X X^\intercal\right]\right)^{-1} \left(\frac{1}{\eta_0} X \left[Y - X^\intercal \beta_0\right]\right) \\
&= \left(\mathbb{E}\left[X X^\intercal\right]\right)^{-1} \left(X \left[Y - X^\intercal \beta_0\right]\right), \qquad (1.20)
\end{aligned}$$

which is identical to what we derived in (1.9).

1.4.7 Finding the Efficient Influence Function via the Efficient Score

Another way of deriving the Efficient influence function, and the manner which we will adopt in the case of a semiparametric model in the next section, is via the Efficient Score which we will define now. Recall that since any influence function must satisfy (1.15), it must be orthogonal to every element in the nuisance tangent space Λ. That is, the influence function of any RAL estimator must lie in $\Lambda^\perp$.

We then note that the projection of the score function of the target parameter, S_β, onto the orthogonal complement of the Nuisance Tangent Space, $\Lambda^\perp$, represented by $\Pi(S_\beta \mid \Lambda^\perp)$, also satisfies (1.15) in that $\mathbb{E}\left[\Pi(S_\beta \mid \Lambda^\perp) S_\eta^\intercal\right] = 0$. We call this projection the Efficient Score. That is

$$S_{\text{eff}} = \Pi(S_\beta \mid \Lambda^\perp) = S_\beta - \Pi(S_\beta \mid \Lambda). \qquad (1.21)$$

We note that S_{eff} may not be the efficient influence function since it may not satisfy (1.14). Therefore, our goal now is to find $B \in \mathbb{R}^{q\times q}$ such that $\mathbb{E}\left(B S_{\text{eff}} S_\beta^\intercal\right) = I$, and in so doing, BS_{eff} will be an influence function. Since S_β and $\Pi(S_\beta \mid \Lambda)$ both lie in $\mathcal{T}$, we also know that $BS_{\text{eff}} \in \mathcal{T}$, and will hence be the efficient influence function.

As seen from above, it is clear that

$$\begin{aligned} B &= \left(\mathbb{E}\left(S_{\text{eff}} S_\beta^\intercal\right)\right)^{-1} \\ &= \left(\mathbb{E}\left(S_{\text{eff}}\left(S_{\text{eff}}^\intercal - \Pi(S_\beta \mid \Lambda)^\intercal\right)\right)\right)^{-1} \\ &= \left(\mathbb{E}\left(S_{\text{eff}} S_{\text{eff}}^\intercal\right)\right)^{-1}, \end{aligned}$$

since $S_{\text{eff}} \perp \Pi(S_\beta \mid \Lambda)$.

Hence, the efficient influence function may also be written as a function of the Efficient Score as follows:

$$\varphi_{\text{eff}}(Y,X) = BS_{\text{eff}}(Y,X) = \left(\mathbb{E}\left(S_{\text{eff}}(Y,X) S_{\text{eff}}^\intercal(Y,X)\right)\right)^{-1} S_{\text{eff}}(Y,X) \tag{1.22}$$

In our running example, the efficient score is given by

$$S_{\text{eff}} = S_\beta - \Pi(S_\beta \mid \Lambda) = S_\beta - \mathbb{E}(S_\beta S_\eta)\left(\mathbb{E}\left(S_\eta^2\right)\right)^{-1} S_\eta.$$

Furthemore, since $\epsilon \mid X \sim \mathcal{N}(0,\eta)$, we know that $\mathbb{E}(\epsilon \mid X) = \mathbb{E}(\epsilon^3 \mid X) = 0$. Hence,

$$\begin{aligned} \mathbb{E}(S_\beta S_\eta) &= \mathbb{E}\left(\frac{1}{\eta_0} X\epsilon\left(-\frac{1}{2\eta_0} + \frac{\epsilon^2}{2\eta_0^2}\right)\right) \\ &= -\frac{1}{2\eta_0^2}\,\mathbb{E}(X\epsilon) + \frac{1}{2\eta_0^3}\,\mathbb{E}(X\epsilon^3) \\ &= -\frac{1}{2\eta_0^2}\,\mathbb{E}(X\,\mathbb{E}(\epsilon \mid X)) + \frac{1}{2\eta_0^3}\,\mathbb{E}(X\,\mathbb{E}(\epsilon^3 \mid X)) \\ &= 0. \end{aligned}$$

As such,

$$S_{\text{eff}} = S_\beta.$$

And the efficient influence function is given by

$$\begin{aligned} \varphi_{\text{eff}}(Y,X) &= \left(\mathbb{E}\left(S_{\text{eff}}(Y,X) S_{\text{eff}}^\intercal(Y,X)\right)\right)^{-1} S_{\text{eff}}(Y,X) \\ &= \left(\mathbb{E}\left(S_\beta S_\beta^\intercal\right)\right)^{-1} S_\beta \\ &= \left(-\,\mathbb{E}\left(S_{\beta\beta}\right)\right)^{-1} S_\beta \\ &= \left(\mathbb{E}\left(XX^\intercal\right)\right)^{-1}\left(X\left[Y - X^\intercal\beta_0\right]\right), \end{aligned}$$

which is the same as in (1.9) and (1.20).

Since the OLS estimator has influence function which is equal to the efficient influence function, we can conclude that it is the efficient estimator in our parametric model. We will show in Section 1.5.6 how to derive an efficient estimator using the efficient influence function if the original estimator is not efficient to begin with.

1.5 The Heteroskedastic Linear Model and Semiparametric Efficient Estimation

In this section, even though we will be working within a semiparametric model that cannot be indexed by a finite set of parameters, we will treat the results obtained in the previous section as true and point out distinctions in the proof where it will aid in understanding. In particular, the workflow for finding the efficient estimator in our model will remain the same. That is, we need to work out what the score of the target parameter is, determine the nuisance tangent space, project the score onto the nuisance tangent space to obtain the efficient score, and re-scale the efficient score to obtain the efficient influence curve. We will then show how to obtain the efficient estimator once we have the efficient influence function via a one-step estimator approach and the estimating equation approach.

1.5.1 The Heteroskedastic Linear Model

We consider a statistical model of the joint distribution of Y and X, and also relax the assumptions that (i) the conditional distribution of Y given X is Gaussian, and (ii) that the conditional variance of the residuals ϵ is constant across X. We call this the Heteroskedastic Linear Model and represent it as

$$\mathcal{M}_{\text{Het}} := \left\{p_{Y,X}(y,x) : Y = X^{\intercal}\beta + \epsilon, \mathbb{E}(\epsilon \mid X) = 0\right\}.$$

We point out that since we do not make any assumptions on the shape of each distribution in $\mathcal{M}_{\text{Het}}$, each of these distributions cannot be indexed by a finite set of parameters. However, since our interest is still in the parameter β, we can instead index each distribution by β and an infinite nuisance parameter, η. To some extent, though it can be helpful to think of η as an infinite 'vector', strictly speaking, it is actually more mathematically precise to think of it as an infinite dimensional vector function $\eta(\cdot)$ instead.

To get a better handle on what $\eta(\cdot)$ is, we consider the following. For each distribution, since there is a one-to-one relationship between Y and ϵ given X, for any density $p_{Y,X}(y,x)$ in $\mathcal{M}_{\text{Het}}$ we can write

$$\begin{aligned} p_{Y,X}(y,x) &= p_{\epsilon,X}(\epsilon,x) \\ &= p_{\epsilon|X}(\epsilon \mid x)p_X(x). \end{aligned}$$

Now let the functions $\eta^*(\cdot)$, $\eta^{**}(\cdot)$ and $\eta(\cdot)$ be defined in the following manner: $\eta^*(\epsilon, x) = p_{\epsilon|X}(\epsilon \mid x)$, $\eta^{**}(x) = p_X(x)$, and $\eta(\epsilon, x) = \eta^*(\epsilon, x)\eta^{**}(x)$. Then treating $\eta^*(\cdot)$ as a function (rather than a density), we see that for all x,

$$\int \eta^*(\epsilon, x)d\epsilon = 1 \text{ and} \int \epsilon\eta^*(\epsilon, x)d\epsilon = 0.$$

Similarly, $\eta^{**}(\cdot)$ as a function must satisfy

$$\int \eta^{**}(x)dx = 0.$$

We can thus think of the joint distributions of Y and X in the model $\mathcal{M}_{\text{Het}}$ as being indexed by the parameter β as well as the functions $\eta^*(\cdot)$, and $\eta^{**}(\cdot)$. That is, for each $p_{Y,X}(y, x) \in \mathcal{M}_{\text{Het}}$, we can write

$$p_{Y,X}(y, x) \equiv p_{Y,X;\beta,\eta^*,\eta^{**}}(y, x),$$

and in particular, the true data generating $p_0(y, x)$ distribution is indexed by β_0, $\eta_0^*(\cdot)$, and $\eta_0^{**}(\cdot)$ as follows:

$$p_0(y, x) = p_{Y,X;\beta_0,\eta_0^*,\eta_0^{**}}(y, x).$$

1.5.2 Parametric Submodels

At this point, we should note that it is incredibly difficult to work with infinite dimensional parameters. Furthermore, the concepts and tools we developed in the previous section were applicable only to models that could be indexed by a finite number of parameters. As such, in order to make progress, we first consider working with submodels within $\mathcal{M}_{\text{Het}}$ that are indexed only by a finite number of parameters and then consider what happens in the limit when we take the union over all these submodels.

To that end, we define a *parametric submodel* $\mathcal{M}_{\beta,\gamma^*,\gamma^{**}}$ as a collection of joint distributions of Y and X in $\mathcal{M}_{\text{Het}}$ that are indexed by a finite vector of parameters $(\beta^\intercal, \gamma^{*\intercal}, \gamma^{**\intercal})^\intercal \in \mathbb{R}^{q+r_1+r_2}$, and **that contains the true data-generating distribution**. Note that since the definition of a parametric submodel above necessitates that it contains the true data generating distribution, we are, in practice, usually not able to write down an explicit form of the distributions contained in any parametric submodel. We will only be using parametric submodels purely as a conceptual tool to get a better understanding of the larger semiparametric model.

Since every parametric submodel contains the true data generating distribution by definition, we can now index that distribution by a finite vector of true parameters $(\beta_0^\intercal, \gamma_0^{*^\intercal}, \gamma_0^{**^\intercal})^\intercal$, and write

$$p_0(y, x) \equiv p_{Y,X;\beta_0,\gamma_0^*,\gamma_0^{**}}(y, x).$$

For each parametric submodel $\mathcal{M}_{\beta,\gamma^*,\gamma^{**}}$, we can then define an associated nuisance tangent space $\Lambda_{\gamma^*,\gamma^{**}}$ as the space of q-dimensional mean-zero random functions that are spanned by the score function of the nuisance parameters. That is,

$$\Lambda_{\gamma^*,\gamma^{**}} = \left\{B^* S_{\gamma^*} + B^{**} S_{\gamma^{**}} \mid B^* \in \mathbb{R}^{q\times r_1}, B^{**} \in \mathbb{R}^{q\times r_2}\right\}.$$

In any parametric submodel, we can define the score function of the nuisance parameters like we did before as the first derivative of the log-likelihood of the density. In particular, when these score functions are evaluated at the truth, we have

$$\begin{aligned}
S_{\gamma^*} &\equiv S_{\gamma^*}(Y, X, \beta_0, \gamma_0^*, \gamma_0^{**}) \\
&= \frac{\partial}{\partial\gamma^*} \log p_{Y,X;\beta_0,\gamma_0^*,\gamma_0^{**}}(y, x) \\
&= \frac{\partial}{\partial\gamma^*} \left(\log p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x) + \log p_{X;\gamma_0^{**}}(x)\right) \\
&= \frac{\partial}{\partial\gamma^*} \log p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x), \text{ and} \qquad (1.23) \\
S_{\gamma^{**}} &\equiv S_{\gamma^{**}}(Y, X, \beta_0, \gamma_0^*, \gamma_0^{**}) \\
&= \frac{\partial}{\partial\gamma^{**}} \log p_{Y,X;\beta_0,\gamma_0^*,\gamma_0^{**}}(y, x) \\
&= \frac{\partial}{\partial\gamma^{**}} \left(\log p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x) + \log p_{X;\gamma_0^{**}}(x)\right) \\
&= \frac{\partial}{\partial\gamma^*} \log p_{X;\gamma_0^{**}}(x).
\end{aligned}$$

We can then define the nuisance tangent space Λ for $\mathcal{M}_{\text{Het}}$ to be the *mean-square closure* of the union of the nuisance tangent spaces $\Lambda_{\gamma^*,\gamma^{**}}$ associated with all parametric submodels. The notion of the mean-square closure is a technical detail which the reader can read more about in Tsiatis (2007). In essence, by 'closure' we mean that Λ is not only the union of the nuisance tangent spaces of *all* parametric submodels, but also contains the limits of sequences of elements from the union, and by 'mean-square' we mean that the limit is defined by the mean of the squared-distance induced by the inner product.

This, together with the regularity assumption that the resultant space Λ is still a linear space, is important as it guarantees the existence and uniqueness of the projection of a random function onto Λ, and in particular, the existence and uniqueness of the projection $\Pi(S_\beta \mid \Lambda)$ of the score function of the target parameter onto the nuisance tangent space Λ.

1.5.3 The Nuisance Tangent Space Λ

One way of deriving the nuisance tangent space Λ is to first hypothesize a set of functions the space should encompass by looking at properties of the scores that span them, and then show that the set is precisely the mean-square closure of the nuisance tangent spaces of all parametric submodels. While this is the approach taken in Tsiatis (2007), he does so in a way that also explores the geometric relationship between the different conditions implied by conditional distribution of ϵ given X and the marginal distribution of X. Since our goal here is simply to didactically derive an expression of Λ, we present a more compact and direct derivation of our own.

1.5.3.1 Our Conjecture

In $\mathcal{M}_{\text{Het}}$, we know that $\mathbb{E}[\epsilon \mid X] = 0$ for all X. This implies that $\mathbb{E}[\epsilon \mid X] = 0$ also holds in any parametric submodel $\mathcal{M}_{\gamma^*,\gamma^{**}}$. As such, we have

$$\begin{aligned}
& \int \epsilon p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x) d\epsilon = 0 \\
\overset{\text{(i)}}{\Longrightarrow} \quad & \frac{\partial}{\partial \gamma^*} \int \epsilon p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x) d\epsilon = 0 \\
\overset{\text{(ii)}}{\Longrightarrow} \quad & \int \epsilon \frac{\partial}{\partial \gamma^*} p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x) d\epsilon = 0 \\
\overset{\text{(iii)}}{\Longrightarrow} \quad & \int \epsilon \frac{\frac{\partial}{\partial \gamma^*} p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x)}{p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x)} p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x) d\epsilon = 0 \\
\overset{\text{(iv)}}{\Longrightarrow} \quad & \int \epsilon \frac{\partial}{\partial \gamma^*} \left(\log p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x)\right) p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x) d\epsilon = 0 \\
\overset{\text{(v)}}{\Longrightarrow} \quad & \int \epsilon S_{\gamma^*} p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x) d\epsilon = 0 \\
\overset{\text{(vi)}}{\Longrightarrow} \quad & \mathbb{E}\left[\epsilon S_{\gamma^*} \mid X\right] = 0,
\end{aligned}$$

where in (i), we take partial derivatives with respect to γ^* on both sides, in (ii), we assume sufficient regularity to swap the integral and derivative, in (iii), we multiply and divide the integrand by $p_{\epsilon|X;\beta_0,\gamma_0^*}(\epsilon \mid x)$, in (iv) we use the identity that $\frac{\partial}{\partial \gamma} \log p_{Z,\gamma}(z) = \frac{\frac{\partial}{\partial \gamma} p_{Z,\gamma}(z)}{p_{Z,\gamma}(z)}$, in (v) we use (1.23), and finally, in (vi), we use the definition of the conditional expectation of a function.

Next, we also notice that $S_{\gamma^{**}}$ is a function only of X. Hence,

$$\mathbb{E}[\epsilon S_{\gamma^{**}} \mid X] = S_{\gamma^{**}}\, \mathbb{E}[\epsilon \mid X] = 0,$$

since $\mathbb{E}[\epsilon \mid X] = 0$.

And so, for any element $l(\epsilon, X)$ in the nuisance tangent space $\Lambda_{\gamma^*,\gamma^{**}}$ of an arbitrary parametric submodel $\mathcal{M}_{\beta,\gamma^*,\gamma^{**}}$, since

$$l(\epsilon, X) = B^* S_{\gamma^*} + B^{**} S_{\gamma^{**}}$$

for some $B^* \in \mathbb{R}^{q\times r_1}$ and $B^{**} \in \mathbb{R}^{q\times r_2}$, we know that

$$\mathbb{E}\left[\epsilon l(\epsilon, X) \mid X\right] = B^*\,\mathbb{E}\left[\epsilon S_{\gamma^*} \mid X\right] + B^{**}\,\mathbb{E}\left[\epsilon S_{\gamma^{**}} \mid X\right] = 0. \tag{1.24}$$

Since Λ is the mean-square closure of the nuisance tangent spaces $\Lambda_{\beta,\gamma^*,\gamma^{**}}$ of all parametric submodels, it is a reasonable conjecture that Λ should contain all mean-zero random functions whose conditional expectation given X when multiplied by ϵ is zero. That is, we hypothesize that $\Lambda^{\text{conj}} = \Lambda$ where

$$\Lambda^{\text{conj}} = \left\{l(\epsilon, X) \in \mathcal{H} : \mathbb{E}\left[\epsilon l(\epsilon, X) \mid X\right] = 0\right\}.$$

1.5.3.2 Proving Our Conjecture

To show that $\Lambda^{\text{conj}} = \Lambda$, we need to show that Λ^{conj} is indeed the mean-square closure of the nuisance tangent spaces of all parametric submodels. We do this via the usual approach of showing that two sets are equal by showing that they are subsets of each other. In this case, we want to show that the mean-square closure is a subset of Λ^{conj} and vice-versa. This can be done by showing that

1. any element in the nuisance tangent space of any parametric submodel is an element in Λ^{conj}, and
2. any element in Λ^{conj} is an element in the nuisance tangent space of at least one parametric submodel or is the limit of elements of the nuisance tangent spaces of parametric submodels.

The first statement is an immediate consequence, by construction, of (1.24) since the choice of parametric submodel there was arbitrary.

To show the second statement, we will first choose an arbitrary bounded function in Λ^{conj} and then try to construct a parametric submodel that has a nuisance tangent space that contains it. In the following steps, we make the distribution explicit that expectations are taken over by using subscripts.

We first let $l(\epsilon, X) \in \Lambda^{\text{conj}}$ be a bounded, (jointly) mean-zero random function such that

$$\mathbb{E}_{\epsilon|X}\left[\epsilon l(\epsilon, X) \mid X\right] = 0.$$

In order to define a parametric submodel that contains the true joint distributions of ϵ and X more easily, we consider perturbations of the conditional distribution of ϵ given X and the marginal distribution of X about the true distributions separately.

Decomposing the function $l(\epsilon, X)$. We begin by defining

$$l^{**}(X) := \mathbb{E}_{\epsilon|X}\left[l(\epsilon, X) \mid X\right].$$

It is easy to see that $l^{**}(X)$ satisfies

$$\mathbb{E}_X\left[l^{**}(X)\right] = \mathbb{E}_X\left[\mathbb{E}_{\epsilon|X}\left[l(\epsilon, X) \mid X\right]\right] = 0, \tag{1.25}$$

and

$$\mathbb{E}_{\epsilon|X}\left[\epsilon l^{**}(X) \mid X\right] = l^{**}(X)\,\mathbb{E}\left[\epsilon \mid X\right] = 0,$$

since $\mathbb{E}\left[\epsilon \mid X\right] = 0$ in $\mathcal{M}_{\text{Het}}$.

Now, we define

$$l^{*}(\epsilon, X) := l(\epsilon, X) - l^{**}(X),$$

and similarly see that

$$\mathbb{E}_{\epsilon|X}\left[l^{*}(\epsilon, X) \mid X\right] = \mathbb{E}_{\epsilon|X}\left[l(\epsilon, X) \mid X\right] - \mathbb{E}_{\epsilon|X}\left[l^{**}(X) \mid X\right] = l^{**}(X) - l^{**}(X) = 0, \tag{1.26}$$

and

$$\mathbb{E}_{\epsilon|X}\left[\epsilon l^{*}(\epsilon, X) \mid X\right] = \mathbb{E}_{\epsilon|X}\left[\epsilon l(\epsilon, X) \mid X\right] - \mathbb{E}_{\epsilon|X}\left[\epsilon l^{**}(X) \mid X\right] = 0. \tag{1.27}$$

Defining a set of functions $\mathcal{M}_{\gamma^*,\gamma^{**}}$. With $l^*(\epsilon, X)$ and $l^{**}(X)$, we can now use two finite-dimensional parameters γ^* and γ^{**} to define a set $\mathcal{M}_{\gamma^*,\gamma^{**}}$ that contains functions of the form

$$\underbrace{p_{\epsilon|X;\eta_0^*}(\epsilon \mid X)\left(1 + \gamma^{*\intercal} l^*(\epsilon, X)\right)}_{p_{\epsilon|X;\gamma^*}(\epsilon|X)}\underbrace{p_{X;\eta_0^{**}}(X)\left(1 + \gamma^{**\intercal} l^{**}(X)\right)}_{p_{X;\gamma^{**}}(X)}. \tag{1.28}$$

Verifying that $\mathcal{M}_{\gamma^*,\gamma^{**}}$ is a parametric submodel. In the expression above, we treat the true densities $p_{\epsilon|X;\eta_0^*}(\epsilon \mid X)$ and $p_{X;\eta_0^{**}}(X)$ as known functions. As such, we see that $\mathcal{M}_{\gamma^*,\gamma^{**}}$ contains the density of the true data-generating distribution given by $p_{\epsilon|X;\eta_0^*}(\epsilon \mid X)p_{X;\eta_0^{**}}(X)$ for parameter values $\gamma^* = \gamma^{**} = 0$. This is an important first criteria to ensure that $\mathcal{M}_{\gamma^*,\gamma^{**}}$ is indeed a valid parametric submodel.

Next, we have to check that the functions in $\mathcal{M}_{\gamma^*,\gamma^{**}}$ are indeed densities. That is, we need to check that are non-negative, and integrate to 1.

To guarantee the functions in (1.28) are non-negative, we simply restrict the range of γ^* and γ^{**} defining $\mathcal{M}_{\gamma^*,\gamma^{**}}$ to be small enough values to ensure that $(1 + \theta^{*\intercal} l^*(\epsilon, X)) \geq 0$ and $(1 + \theta^{**\intercal} l^{**}(X)) \geq 0$ for all ϵ and X.

We can then check that the functions integrate to 1 directly by first noting that

$$\int\int p_{\epsilon|X;\gamma^*}(\epsilon \mid X)p_{X;\gamma^{**}}(X)d\epsilon dX = \int p_{X;\gamma^{**}}(X)\left[\int p_{\epsilon|X;\gamma^*}(\epsilon \mid X)d\epsilon\right]dX.$$

Looking at the inner integral, we have that

$$\begin{aligned}\int p_{\epsilon|X;\gamma^*}(\epsilon \mid X)d\epsilon &= \int p_{\epsilon|X;\eta_0^*}(\epsilon \mid X)\left(1+\gamma^{*\intercal} l^*(\epsilon,X)\right)d\epsilon \\ &= \int p_{\epsilon|X;\eta_0^*}(\epsilon \mid X)d\epsilon + \gamma^{*\intercal}\int l^*(\epsilon,X)p_{\epsilon|X;\eta_0^*}(\epsilon \mid X)d\epsilon \\ &\overset{\text{(i)}}{=} 1+\gamma^{*\intercal}\,\mathbb{E}_{\epsilon|X}\left[l^*(\epsilon,X) \mid X\right] \\ &\overset{\text{(ii)}}{=} 1,\end{aligned}$$

where in (i) we make use of the fact that $p_{\epsilon|X;\eta_0^*}(\epsilon, X)$ is a valid density, and in (ii) we make use of (1.26).

Looking at the outer integral, we have that

$$\begin{aligned}\int p_{X;\gamma^{**}}(X)dX &= \int p_{X;\eta_0^{**}}(X)\left(1+\gamma^{**\intercal} l^{**}(X)\right)dX \\ &= \int p_{X;\eta_0^{**}}(X)dX + \gamma^{**\intercal}\int l^{**}(X)p_{X;\eta_0^{**}}(X)dX \\ &\overset{\text{(i)}}{=} 1+\gamma^{**\intercal}\,\mathbb{E}_X\left[l^{**}(X)\right] \\ &\overset{\text{(ii)}}{=} 1,\end{aligned}$$

where in (i) we similarly make use of the fact that $p_{X;\eta_0^{**}}(X)$ is a valid density, and in (ii) we make use of (1.25).

We have thus verified that under the restricted set of values for γ^* and γ^{**}, the functions of the form in (1.28) are in fact densities.

The last thing we have to check in order to be sure that $\mathcal{M}_{\gamma^*,\gamma^{**}}$ is in fact a parametric submodel is whether $\mathcal{M}_{\gamma^*,\gamma^{**}} \subset \mathcal{M}_{\text{Het}}$. That is, we need to verify that all densities $p_{\epsilon|X}(\epsilon \mid X)p_X(X) \in \mathcal{M}_{\gamma^*,\gamma^{**}}$ satisfy the condition that

$$\mathbb{E}_{\epsilon|X}[\epsilon \mid X] = \int \epsilon p_{\epsilon|X}(\epsilon \mid X)d\epsilon = 0.$$

Evaluating the integral above directly, we have

$$\begin{aligned}\int \epsilon p_{\epsilon|X;\gamma^*}(\epsilon \mid X)d\epsilon &= \int \epsilon p_{\epsilon|X;\eta_0^*}(\epsilon \mid X)\left(1+\gamma^{*\intercal} l^*(\epsilon,X)\right)d\epsilon \\ &= \int \epsilon p_{\epsilon|X;\eta_0^*}(\epsilon \mid X)d\epsilon + \gamma^{*\intercal}\int \epsilon l^*(\epsilon,X)p_{\epsilon|X;\eta_0^*}(\epsilon \mid X)d\epsilon \\ &\overset{\text{(i)}}{=} 0+\gamma^{*\intercal}\,\mathbb{E}_{\epsilon|X}\left[\epsilon l^*(\epsilon,X) \mid X\right] \\ &\overset{\text{(ii)}}{=} 0,\end{aligned}$$

where in (i) we make use of the fact that $p_{\epsilon|X;\eta_0^*}(\epsilon \mid X) \in \mathcal{M}_{\text{Het}}$, and in (ii) we make use of (1.27).

Showing that the nuisance tangent space of our parametric submodel contains the function $\boldsymbol{l(\epsilon, X)}$. Now, all that is left is to show that $l(\epsilon, X)$ lies in the nuisance tangent space of this parametric submodel. To do so, we first need to evaluate the scores of the nuisance parameters γ^* and γ^{**} as follows:

$$\begin{aligned}
S_{\gamma^*} &= \frac{\partial}{\partial \gamma^*} \log \left\{ p_{Y|X;\beta_0,\eta_0^*}(\epsilon, X) \left(1 + \gamma^{*\intercal} l^*(\epsilon, X)\right) p_{X;\eta_0^{**}}(X) \left(1 + \gamma^{**\intercal} l^{**}(X)\right) \right\} \Big|_{\gamma^*=0} \\
&= \frac{\partial}{\partial \gamma^*} \log \left(1 + \gamma^{*\intercal} l^*(\epsilon, X)\right) \Big|_{\gamma^*=0} \\
&= l^*(\epsilon, X) \\
S_{\gamma^{**}} &= \frac{\partial}{\partial \gamma^{**}} \log \left\{ p_{Y|X;\beta_0,\eta_0^*}(\epsilon, X) \left(1 + \gamma^{*\intercal} l^*(\epsilon, X)\right) p_{X;\eta_0^{**}}(X) \left(1 + \gamma^{**\intercal} l^{**}(X)\right) \right\} \Big|_{\gamma^{**}=0} \\
&= \frac{\partial}{\partial \gamma^{**}} \log \left(1 + \gamma^{**\intercal} l^{**}(X)\right) \Big|_{\gamma^{**}=0} \\
&= l^{**}(X).
\end{aligned}$$

Since the nuisance tangent space of $\mathcal{M}_{\gamma^*,\gamma^{**}}$ is spanned by the scores of the nuisance parameters, it must contain $l^*(\epsilon, X) + l^{**}(X) = l(\epsilon, X)$. Since our choice of $l(\epsilon, X)$ was arbitrary, we can choose them to also be the limits of a sequence of bounded mean-zero random variables and arrive at the same result.

Hence, we have shown that

$$\Lambda = \Lambda^{\text{conj}} = \{l(\epsilon, X) \in \mathcal{H} : \mathbb{E}\left[\epsilon l(\epsilon, X) \mid X\right] = 0\}$$

1.5.4 The Efficient Score

Recall that we define in (1.21) that the efficient score S_{eff} is the projection of the score of the target parameter S_β onto the orthogonal complement of the nuisance tangent space $\Lambda^\perp$ which we can think of as the collection of all the residuals $h(\epsilon, X) - \Pi(h \mid \Lambda)$ for every $h \in \mathcal{H}$.

At this point, we do not know what the space $\Lambda^\perp$ comprises, but we can hypothesize that it contains all functions of the form $A(X)\epsilon$ since we observe that

$$\mathbb{E}\left[(A(X)\epsilon)^\intercal l(\epsilon, X)\right] = \mathbb{E}\left[(A^\intercal(X)\, \mathbb{E}\left[\epsilon l(\epsilon, X) \mid X\right]\right] = 0$$

for all $l(\epsilon, X) \in \Lambda$.

To show that $\Lambda^{\perp}$ does indeed comprise elements of the form $A(x)\epsilon$, we need to find a unqiue random function $A(X) \in \mathcal{H}$ such that for any element $h(\epsilon, X) \in \mathcal{H}$,

$$\Pi(h \mid \Lambda^{\perp}) = A(X)\epsilon. \tag{1.29}$$

Since we also know that the residual $h(\epsilon, X) - \Pi(h \mid \Lambda^{\perp})$ of the projection of $h(\epsilon, X)$ onto the orthogonal complement $\Lambda^{\perp}$ is by definition an element of Λ, we have that

$$\mathbb{E}\left[\epsilon\left(h(\epsilon, X) - \Pi(h \mid \Lambda^{\perp})\right) \mid X\right] = 0.$$

By rearranging the equation above, and using (1.29) we thus have

$$\begin{aligned}
\mathbb{E}\left[\epsilon h(\epsilon, X) \mid X\right] &= \mathbb{E}\left[\epsilon \Pi(h \mid \Lambda^{\perp})) \mid X\right] \\
&= \mathbb{E}\left[\epsilon(A(X)\epsilon) \mid X\right] \\
&= A(X)\,\mathbb{E}\left[\epsilon^2 \mid X\right] \\
&= A(X)\sigma^2(X),
\end{aligned}$$

where we use $\sigma^2(X)$ to represent $\mathbb{E}\left[\epsilon^2 \mid X\right]$ to simplify notation. Note that the notation is intentional because $\mathbb{E}\left[\epsilon^2 \mid X\right]$ is really the conditional variance of the residual within strata of X.

Solving for $A(X)$, and assuming $\sigma^2(X) \neq 0$, we thus see that we have a unique solution for $A(X)$ is given by

$$A(X) = \frac{1}{\sigma^2(X)}\,\mathbb{E}\left[\epsilon h(\epsilon, X) \mid X\right].$$

We have thus shown that the orthogonal complement $\Lambda^{\perp}$ consists of elements of the form $A(X)\epsilon$. In particular, if we replace $h(\epsilon, X)$ in the derivation above with S_β, we see that the projection of S_β onto $\Lambda^{\perp}$, which is the efficient score S_{eff}, is given by

$$S_{\text{eff}} = \Pi(S_\beta \mid \Lambda^{\perp}) = \frac{1}{\sigma^2(X)}\,\mathbb{E}\left[\epsilon S_\beta \mid X\right]\epsilon.$$

All that remains to determine the exact form of the efficient score is to evaluate $\mathbb{E}\left[\epsilon S_\beta \mid X\right]$. To do so, we first make use of the fact that S_β is the derivative of the log-likelihood with respect the β evaluated at β_0 and $\eta_0(\cdot)$. That is,

$$\begin{aligned} S_\beta &= \frac{\partial}{\partial\beta}\log p_{Y,X;\beta,\eta_0(\cdot)}(y,x)\Big|_{\beta=\beta_0} \\ &= \frac{\partial}{\partial\beta}\left(\log\eta_0^*(y-x^\intercal\beta,x)+\log\eta_0^{**}(x)\right)\Big|_{\beta=\beta_0} \\ &= \frac{\partial}{\partial\beta}\log\eta_0^*(\epsilon,x)\Big|_{\beta=\beta_0} \\ &= \frac{\frac{\partial}{\partial\beta}\eta_0^*(\epsilon,x)\Big|_{\beta=\beta_0}}{\eta_0^*(\epsilon,x)}. \end{aligned}$$

And so, we have that

$$S_\beta\eta_0^*(\epsilon,x) = \frac{\partial}{\partial\beta}\eta_0^*(\epsilon,x)\Big|_{\beta=\beta_0}. \tag{1.30}$$

Secondly, we know that $\int\epsilon\eta^*(\epsilon,x)d\epsilon = 0$ since $\mathbb{E}\left[\epsilon \mid X\right]=0$. Taking the derivative with respect to β on both sides, we thus have

$$\begin{aligned} & \frac{\partial}{\partial\beta}\int\epsilon\eta^*(\epsilon,x)d\epsilon\Big|_{\beta=\beta_0} = 0 \\ \overset{\text{(i)}}{\Longrightarrow} & \int\frac{\partial}{\partial\beta}\epsilon\eta^*(\epsilon,x)\Big|_{\beta=\beta_0}d\epsilon = 0 \\ \overset{\text{(ii)}}{\Longrightarrow} & \int\frac{\partial}{\partial\beta}\epsilon\Big|_{\beta=\beta_0}\eta^*(\epsilon,x)d\epsilon + \int\epsilon\frac{\partial}{\partial\beta}\eta^*(\epsilon,x)\Big|_{\beta=\beta_0}d\epsilon = 0 \\ \overset{\text{(iii)}}{\Longrightarrow} & \int(-X)\eta^*(\epsilon,x)d\epsilon + \int\epsilon S_\beta\eta_0^*(\epsilon,x)d\epsilon = 0 \\ \overset{\text{(iv)}}{\Longrightarrow} & -X+\mathbb{E}\left[\epsilon S_\beta \mid X\right] = 0, \end{aligned}$$

where in (i) we swap the order of the integral and the derivative, in (ii), we apply the product rule, in (iii), we note that $\epsilon = Y-X^\intercal\beta$, find its derivative with respect to β and then evaluate the result at β_0 for the term on the left, and make use of (1.30) for the term on the right, and finally in (iv), we use the fact that $\eta^*(\epsilon,x)$ is a (conditional) density which integrates to 1.

And so, our expression for the efficient score reduces to

$$S_{\text{eff}} = \frac{1}{\sigma^2(X)}X\epsilon.$$

1.5.5 The Efficient Influence Function

Following the result from (1.22), we know that the efficient influence function can be obtained by scaling the efficient score such that it lies on the tangent space $\mathcal{T}$. Hence,

$$
\begin{aligned}
\varphi_{\text{eff}}(Y,X) &= \left\{\mathbb{E}\left[S_{\text{eff}}(Y,X)S_{\text{eff}}^{\intercal}(Y,X)\right]\right\}^{-1} S_{\text{eff}}(Y,X) \\
&= \left\{\mathbb{E}\left[\left(\frac{1}{\sigma^2(X)}X\epsilon\right)\left(\frac{1}{\sigma^2(X)}X\epsilon\right)^{\intercal}\right]\right\}^{-1}\left(\frac{1}{\sigma^2(X)}X\epsilon\right) \\
&= \left\{\mathbb{E}\left[\left(\frac{1}{\sigma^2(X)}\right)^2 XX^{\intercal}\epsilon^2\right]\right\}^{-1}\frac{1}{\sigma^2(X)}X\epsilon \\
&= \left\{\mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{\sigma^2(X)}\right)^2 XX^{\intercal}\epsilon^2 \mid X\right]\right]\right\}^{-1}\frac{1}{\sigma^2(X)}X\epsilon \\
&= \left\{\mathbb{E}\left[\left(\frac{1}{\sigma^2(X)}\right)^2 XX^{\intercal}\,\mathbb{E}\left[\epsilon^2 \mid X\right]\right]\right\}^{-1}\frac{1}{\sigma^2(X)}X\epsilon \\
&= \left\{\mathbb{E}\left[\left(\frac{1}{\sigma^2(X)}\right)^2 XX^{\intercal}\sigma^2(X)\right]\right\}^{-1}\frac{1}{\sigma^2(X)}X\epsilon \\
&= \left\{\mathbb{E}\left[\frac{1}{\sigma^2(X)}XX^{\intercal}\right]\right\}^{-1}\frac{1}{\sigma^2(X)}X\epsilon \qquad (1.31)
\end{aligned}
$$

1.5.6 The Efficient Estimator

Unlike the parametric case, the OLS estimator given by (1.3) is not the efficient estimator since its influence function is not given by (1.31). In fact, in our semiparametric model, it is commonly known that the efficient estimator is in fact the Weighted Least-Squares (WLS) estimator given by

$$
\widehat{\beta}_{\text{WLS}} = \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i X_i^{\intercal}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i Y_i\right)
$$

where $W_i = \frac{1}{\sigma^2(X_i)}$. We now show how to derive the WLS estimator from the efficient influence function via two approaches: the one-step estimator approach and the estimating equation approach.

Remark 8 (Estimating the Weights)**.** *As is commonly known, an important step in evaluating $\widehat{\beta}_{WLS}$ is figuring out what the weights W_i are. This can be done via parametric or semi-/non-parametric modeling, but not to stray from the intent of this section, we will treat the function $\sigma^2(X)$ as known or simply assume that it can be consistently estimated. It should be noted, however, that the WLS being the efficient estimator hinges on $\sigma^2(X)$ being known.*

1.5.6.1 The One-Step Estimator

In the one-step approach, we start with a plug-in estimator of the parameter β. Then, it is known that we can derive the efficient estimator by simply adding the empirical mean of the estimated efficient influence function evaluated at each data points to this plug-in estimator. In our case, the OLS estimator is a plug-in estimator, and we will show that

$$\widehat{\beta}_{\text{WLS}} = \widehat{\beta}_{\text{OLS}} + \frac{1}{n}\sum_{i=1}^{n} \widehat{\varphi}_{\text{eff}}(Y_i, X_i).$$

Firstly, since we note that we can estimate the efficient influence function by its plug-in analogue

$$\begin{aligned} \widehat{\varphi}_{\text{eff}}(Y_i, X_i) &= \left\{ \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\sigma^2(X_j)} X_j X_j^{\intercal} \right\}^{-1} \frac{1}{\sigma^2(X_i)} X_i \widehat{\epsilon}_i \\ &= \left\{ \frac{1}{n}\sum_{j=1}^{n} W_j X_j X_j^{\intercal} \right\}^{-1} W_i X_i \widehat{\epsilon}_i \end{aligned}$$

where $\widehat{\epsilon}_i = Y_i - X_i^{\intercal}\widehat{\beta}_{\text{OLS}}$.

Then, we see that

$$\begin{aligned} &\widehat{\beta}_{\text{WLS}} - \frac{1}{n}\sum_{i=1}^{n} \widehat{\varphi}_{\text{eff}}(Y_i, X_i) \\ =& \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i X_i^{\intercal} \right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i Y_i \right) - \frac{1}{n}\sum_{i=1}^{n} \left[\left(\frac{1}{n}\sum_{j=1}^{n} W_j X_j X_j^{\intercal} \right)^{-1} W_i X_i \widehat{\epsilon}_i \right] \\ =& \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i X_i^{\intercal} \right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i Y_i \right) - \left(\frac{1}{n}\sum_{j=1}^{n} W_j X_j X_j^{\intercal} \right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i \widehat{\epsilon}_i \right) \\ =& \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i X_i^{\intercal} \right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i \left(Y_i - \widehat{\epsilon}_i \right) \right) \\ =& \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i X_i^{\intercal} \right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i X_i^{\intercal} \widehat{\beta}_{\text{OLS}} \right) \\ =& \widehat{\beta}_{\text{OLS}}. \end{aligned}$$

Hence, we have shown that the one-step estimator approach applied to the OLS estimator yields the efficient WLS estimator.

1.5.6.2 The Estimating Equation Approach

Another approach for deriving the efficient estimator is to identify the estimating equation that produces an estimator with the efficient influence function as its influence function. Recalling from (1.10) that the influence curve of an m-estimator for β with the function $m(\cdot)$ has the form

$$\varphi(Y_i, X_i) = -\left(\mathbb{E}\left[\frac{\partial m(Y, X, \beta_0)}{\partial \beta^\intercal}\right]\right)^{-1} m(Y_i, X_i, \beta_0).$$

A good guess for the from for $m(Y_i, X_i, \beta_0)$ would be taking $m(Y_i, X_i, \beta_0) = \frac{1}{\sigma^2(X)} X\epsilon = \frac{1}{\sigma^2(X)} X\left(Y - X^\intercal\beta_0\right)$. We can then check that its influence function $\varphi_m(Y_i, X_i)$ associated with the m-estimator with function $m(\cdot)$ is precisely the efficient influence function $\varphi_{\text{eff}}(Y_i, X_i)$ as follows.

$$\begin{aligned}
\varphi_m(Y_i, X_i) &= -\left(\mathbb{E}\left[\frac{\partial m(Y, X, \beta_0)}{\partial \beta^\intercal}\right]\right)^{-1} m(Y_i, X_i, \beta_0) \\
&= -\left(\mathbb{E}\left[\frac{\partial}{\partial \beta^\intercal}\left(\frac{1}{\sigma^2(X)} X\left(Y - X^\intercal\beta\right)\right)\Bigg|_{\beta=\beta_0}\right]\right)^{-1}\left(\frac{1}{\sigma^2(X_i)} X_i\left(Y_i - X_i^\intercal\beta_0\right)\right) \\
&= -\left(\mathbb{E}\left[-\frac{1}{\sigma^2(X)} X X^\intercal\right]\right)^{-1}\left(\frac{1}{\sigma^2(X_i)} X_i\left(Y_i - X_i^\intercal\beta_0\right)\right) \\
&= \left(\mathbb{E}\left[\frac{1}{\sigma^2(X)} X X^\intercal\right]\right)^{-1}\left(\frac{1}{\sigma^2(X_i)} X_i\epsilon_i\right) \\
&= \varphi_{\text{eff}}(Y_i, X_i).
\end{aligned}$$

Furthermore, $m(Y_i, X_i, \theta_0)$ is a valid function for an m-estimator since

$$\begin{aligned}
\mathbb{E}\left[m(Y_i, X_i, \beta_0)\right] &= \mathbb{E}\left[\frac{1}{\sigma^2(X)} X\epsilon\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{\sigma^2(X)} X\epsilon \mid X\right]\right] \\
&= \mathbb{E}\left[\frac{1}{\sigma^2(X)} X\, \mathbb{E}\left[\epsilon \mid X\right]\right] \\
&= 0
\end{aligned}$$

since $\mathbb{E}\left[\epsilon \mid X\right]$ by our model assumption.

Hence, the efficient estimator can also be expressed at the solution, $\widehat{\beta}$ to the estimating equation

$$\sum_{i=1}^{n} \frac{1}{\sigma^2(X_i)} X_i\left(Y - X^\intercal\widehat{\beta}\right) = 0,$$

which is clearly the WLS estimator after rearranging the equations.

Remark 9. *We note that replacing* $\frac{1}{\sigma^2(X_i)}$ *with a function* $f(X_i)$ *yields an estimating equation with a solution that is still a consistent estimator for the parameter* β. *In fact, letting* $f(X_i) = 1$ *yields the estimating equation with the OLS estimator as the solution. The function* f *is sometimes known as a projection function (for example in the Generalized Estimating Equation literature) and affects only the efficiency of the estimator and not its consistency so long as the modeling assumptions (in this case that the mean structure is linear and that* $\mathbb{E}(\epsilon \mid X) = 0$*) hold.*

1.6 Concluding Remarks

The purpose of this chapter was to shed some light on the mechanics behind semiparametric efficient estimation using the familiar Linear Model as a scaffold. We have done so at the expense of not showcasing its benefits concretely beyond that of the Linear Model. Therefore to conclude, we briefly mention three important reasons for adopting a semiparametric approach in modern statistical analysis.

Firstly, adopting a semiparametric approach frees researchers from the constraints of thinking about parameters merely as coefficients of a regression, but more generally as functionals of a data-generating distribution. This gives researchers access to a much larger class of parameters that may provide more targeted and direct answers to their scientific questions of interest. Secondly, semiparametric methods are becoming increasingly important in high-dimensional data analysis. They allow the researcher to partition a problem into a low-dimensional parameter of interest that can be easily understood and interpreted, and a high dimensional nuisance parameter that provides greater flexibility in model fitting and usually does not require many assumptions on the data-generating distribution. Lastly, many semiparametric estimators make use of modern machine learning methods while still providing valid inference. The class of Targeted Maximum Likelihood Estimators for example has been increasingly popular in the estimation of causal effects.

Readers interested to learn more may hopefully find the transition to reading Tsiatis (2007) smoother. We direct readers also to an excellent pedagogical piece by Fisher and Kennedy (2018) for more insight and intuition on influence function based estimation. For even more advanced readers, Bickel et al. (1993) and Newey (1988) are classical but challenging references. And finally, for a modern approach to semiparametric estimation, we direct the reader to van der Laan and Rose (2011) and van der Laan and Rose (2018) which cover the Targeted Maximum Likelihood Estimation framework.

Chapter 2

A Dyadic Item Response Theory Model

2.1 Introduction

The study of how individuals interact within a group has been and continues to be of interest to researchers in the behavioral sciences. Even in this setting, the majority of statistical models focus primarily on how each individual behaves isolated from the influences of other group members. However, one model developed to handle the simplest case of two individuals interacting in a dyad is the Social Relations Model (SRM) (e.g., Kenny, Kashy, & Cook, 2006; Kenny & La Voie, 1984; Warner, Kenny, & Stoto, 1979). Here, the ways one individual (often called an actor or perceiver) of a dyad behaves when paired with the other (often called the partner or target) and vice-versa are analyzed to infer individual-level and dyad-level effects. The behavior of the actors can be directed towards the partner (e.g., an individual's perception of the partner's attractiveness) or undirected (e.g., the number of times an individual takes the lead in a collaborative problem solving task), and can be measured during or after socially interacting with the partner. Compared with traditional "isolated" models, the innovative SRM considers both members of the dyad as contributors to the eventual observed behavior. The SRM model has been most often used in social psychology (e.g., Kenny & Kashy, 1994), but is increasingly being used in other fields. A diverse set of examples include relationships in pharmacy and therapeutics hospital-committee decision-making (Bagozzi & Ascione, 2005), social media ties among basketball teammates (Koster & Brandy, 2018), and militarized interstate disputes (Dorff & Ward, 2013).

In the original formulation of the SRM, the specific behavior of an actor when paired with a partner depends on a composite dyad-level latent trait that can be decomposed into three parts: (i) an individual-level latent trait reflecting a general inclination of the actor to behave in a certain way when paired with a partner, (ii) an individual-level latent trait reflecting the general tendency of the partner to elicit such a behavior, and (iii) a dyad-level latent trait that characterizes the effect of the unique (directed) relationship between both parties on the behavior of the actor that is independent of the two individual-level latent traits (Back & Kenny, 2010). More concretely, if one is interested in the level of physical attraction of an actor towards a partner, then the three components reflect (i) how, on average, an actor tends to find others attractive, (ii) how, on average, the partner tends to be found attractive, and how (iii) the actor uniquely finds the partner attractive. As a result of this formulation, the SRM is identifiable only if individuals belong to multiple pairs.

While the SRM is a useful tool in the analysis of dyads, it has not yet been extended for the case where a set of behaviors or responses of an actor can be viewed as measuring a latent variable, such as the actor's perception of or disposition towards a partner. Multivariate SRM (e.g., Card, Little, & Selig, 2008; Kenny, 1994; Nestler, 2018) accommodates multiple measures, but it effectively corresponds to a set of univariate SRMs with additional correlations of individual-level and dyad-level latent traits across measures. When there are more than two or three measures, the multivariate SRM has an abundance of cross-variable correlations that are not easy to interpret. More importantly, multivariate SRMs do not provide a means for predicting or scoring actor, partner and dyad effects on an underlying latent trait. With existing methodology, a better alternative would be to specify a univariate SRM for some summary of the measures, such as a sum-score or mean. However, this could result in a loss of information analogous to educational testing where the scores on different items of the test are sometimes summed up, and only the sum score is used. Our proposed dyadic Item Response Theory (dIRT) model therefore incorporates an Item Response Theory (IRT) model. Advantages include having the ability to account for differences in item difficulty, allowing for missing responses in subsets of items (under the Missing-at-Random assumption), and having individualized standard errors of the latent trait scores (e.g., Embretson & Reise, 2000).

IRT is the standard approach for modeling the relationship between the latent traits of individuals and their responses to a set of items in educational testing. There are a variety of IRT models that may differ, among other things, in terms of the numbers of parameters in the model, the type of link function used, or the approach taken (e.g., confirmatory or exploratory) (e.g., van der Linden, 2016). However, existing models treat the latent trait as a property of the individuals who responded to the items, and perhaps an external party like a rater, but do not include a unique interaction between individuals in a dyad. That is, traditional IRT can be used to model the behavior of an actor when paired with a partner as a function of the items/stimuli, the actor's tendency to behave in a certain way and perhaps the partner's tendency to elicit the behavior, but does not accommodate the unique dyadic effect due to both individuals interacting in a social setting. Thus, if individuals interact with one another, and the manner and effect of this interaction is of interest, then existing IRT models are not useful.

Although SRM and IRT models each have limitations that could be overcome by the other, there is, to our knowledge, no prior work on integrating the models. Only two related cases appear to exists: Alexandrowicz (2015) extended the Actor-Partner Interdependence Model (APIM) and Common Fate Model (CFM) that we describe in Section 2.2.4 to work within an IRT framework. While these models relax the condition that only an individual's latent ability affects the individual's response to an item, neither of them models the dyadic interaction as a latent trait of the dyad. Furthermore, the APIM and CFM are limited to a dyadic design where each individual is paired with only one other partner whereas the SRM handles the case where individuals belong to multiple pairs (Kenny et al., 2006).

Our contributions include the following. First, we describe our proposed dIRT model that incorporates the key features of both the SRM and IRT. The model includes individual and dyad-level latent traits and corresponding variance and covariance parameters afforded by the SRM, while retaining all the important measurement properties of IRT. We also indicate how the model can be extended to larger groupings than dyads, such as triads. Second, we provide a literature review of related classes of models and discuss data designs and conditions for identifiability. Importantly, unlike the SRM, the dIRT model is identified for cross-sectional data. Third, we extend the basic dIRT model to let the latent traits affect a distal outcome and depend on observed covariates and cluster-level random effects. Finally, we demonstrate the practical utility of the model by applying it to a speed dating dataset and making `Stan` code available, together with a case-study explaining the code. While univariate SRMs for one Likert scale item at a time, treated as continuous, have been applied to speed-dating data (e.g., Ackerman, Kashy, & Corretti, 2015), our multivariate model accommodates the ordinal nature of the responses and allows estimation of the unique interaction variance separate from the error variance. We hope that our contributions will inspire researchers to collect and analyze dyadic data in new settings.

The structure of the chapter is as follows. In Section 2.2, we introduce the basic dIRT model, discuss data design and identification, propose various extensions of the basic model, and provide a review of related models. We present a Markov-chain Monte Carlo approach to estimating the model in Section 2.3, using `Stan` for estimation. In Section 2.4, we apply our model and estimation method to a publicly available speed-dating dataset. In Section 2.5 we conduct a simulation study to evaluate the performance of our estimator under a variety of conditions. Finally, we make some concluding remarks in Section 2.6.

2.2 Dyadic Item Response Theory (dIRT)

2.2.1 Basic dIRT Model

In a social setting where groups of individuals interact, it is likely that the behavior of individual $a \in \{1, 2, \ldots, n\}$ (called the actor) in group g is affected not only by his/her own latent traits, but also those of the individuals he/she interacts with. Additionally, there could also be a "unique" component attributable to the specific composition of the group that could affect the actor's behavior above and beyond the effects at the individual level. We can extend any IRT model to deal with such a setting by replacing the latent trait θ_a of individual a, with a composite latent trait $\theta_{a,g}$ of individual a in the context of group g of size n:

$$\theta_{a,g} \;\equiv\; \alpha_a + \sum_{\substack{j=1 \\ j\neq a}}^{n} \beta_j + \sum_{k\in K} \gamma_{a,g(k)}. \tag{2.1}$$

Here, α_a represents the inclination of the actor to behave in a certain way, β_j represents the tendency of another member j of the group to elicit the behavior, and $\gamma_{a,g(k)}$ represents the unique way members of subgroup $g(k)$ interacted to elicit the behavior from actor a. The last sum above is over all possible subgroups $g(k)$ of sizes 1 to $n-1$ excluding the actor. (The index set is defined as $K := \{A \subseteq \{1, 2, \ldots, n\} \setminus \{a\} \mid |A| \geq 1\}$, i.e., the set of all subsets of $\{1, 2, \ldots, n\} \setminus \{a\}$ except the empty set). Note that $\gamma_{a,g(k)}$ includes not only physical interactions between actor a and the other members of the group, but also how the behavior of actor a is altered by the mere presence of the rest of the group. For example, in a collaborative problem solving task, α_a could represent the inclination of actor a to be vocal, β_j how much partner j tends to elicit opinions from actors, and $\gamma_{a,g(k)}$ how vocal the actor is due to the composition of the group $g(k)$. In practice, it may not be necessary to include anything more than pairwise and possibly three-way interactions.

To simplify notation, in the rest of the chapter, we focus on the case when $n = 2$ as it is clear how the model can be extended when working with larger group sizes. In this dyadic setting, for actor a and partner p, the composite latent trait is modeled as

$$\theta_{a,p} \;\equiv\; \alpha_a + \beta_p + \gamma_{a,p}.$$

Unlike (2.1) where the composite latent variable θ, and in particular the dyad-level latent trait γ, are indexed by the actor and the group, we can instead index θ and γ by both individuals a and p since the index set K reduces to the singleton set $\{\{p\}\}$. Here, α_a is the actor latent trait (sometimes called actor effect), β_p the partner latent trait (sometimes called partner effect), and $\gamma_{a,p}$ the dyadic latent trait (sometimes called interaction or relationship effect) which represents the unique contribution of pairing actor a with partner p to the behavior of the actor. Note that $\gamma_{a,p}$ is not assumed to be identical to $\gamma_{p,a}$ when the roles of actor and partner are reversed.

We could consider any traditional IRT model for measuring $\theta_{a,p}$. The model for response $y_{a,p,i}$ to item i by actor a, when paired with the partner p, is of the form

$$g(\mathbb{P}(y_{a,p,i} = j \mid \theta_{a,p}, \boldsymbol{\xi}_{i,j})) \;=\; f(\theta_{a,p}, \boldsymbol{\xi}_{i,j})$$

for some link function $g(\cdot)$, item parameters $\boldsymbol{\xi}_{i,j}$, and functional form $f(\cdot)$. For instance, for ordinal responses we can obtain the standard partial credit model (Masters, 1982) by using the adjacent-category logit link, letting $\boldsymbol{\xi}_{i,j}$ represent (unidimensional) step difficulty parameters, and taking $f(\cdot)$ to be the identity function. If item i has m_i categories (from 0 to $m_i - 1$), the model becomes

$$\log\left(\frac{\mathbb{P}_{\text{PCM}}(y_{a,p,i} = j \mid \theta_{a,p}, \delta_{i,j})}{\mathbb{P}_{\text{PCM}}(y_{a,p,i} = j-1 \mid \theta_{a,p}, \delta_{i,j})}\right) \;=\; \theta_{a,p} - \delta_{i,j} \;\equiv\; (\alpha_a + \beta_p + \gamma_{a,p}) - \delta_{i,j}, \tag{2.2}$$

subject to the constraint that $\sum_{j=0}^{m_i-1} \mathbb{P}_{\text{PCM}}(y_{a,p,i} = j \mid \theta_{a,p}, \delta_{i,j}) = 1$, where $j \in \{1, 2, \ldots, m_i - 1\}$, and $\delta_{i,j}$ are item step difficulties. Note that we condition on $\delta_{i,j}$ because we will adopt a (pragmatic) Bayesian perspective (see Section 3).

In the dIRT model, we assume that the latent traits (or random effects) have bivariate normal distributions:

$$\begin{aligned} \begin{bmatrix} \alpha_a \\ \beta_a \end{bmatrix} &\sim \text{N}\left(\begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta \\ \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{bmatrix}\right), \\ \begin{bmatrix} \gamma_{a,p} \\ \gamma_{p,a} \end{bmatrix} &\sim \text{N}\left(\begin{bmatrix} \mu_\gamma \\ \mu_\gamma \end{bmatrix}, \begin{bmatrix} \sigma_\gamma^2 & \rho_\gamma\sigma_\gamma^2 \\ \rho_\gamma\sigma_\gamma^2 & \sigma_\gamma^2 \end{bmatrix}\right). \end{aligned} \tag{2.3}$$

The parameters are (i) the variances σ_α^2, σ_β^2, and σ_γ^2 of the individual and dyad latent traits, (ii), the expectations μ_α, μ_β, and μ_γ of each of the individual and dyadic latent traits, and (iii) the correlations $\rho_{\alpha\beta}$ and ρ_γ.

The individual-level correlation $\rho_{\alpha\beta}$ (sometimes called the general or individual reciprocity) relates the tendency of an individual to behave in a certain way (i.e., α_a or α_p) to that same individual's tendency to elicit the behavior from his/her partner (i.e., β_a or β_p). The dyad-level correlation ρ_γ (sometimes called dyadic reciprocity) relates the two (directed) latent traits of each dyad (i.e., $\gamma_{a,p}$ and $\gamma_{p,a}$) to each other.

We will extend the dIRT model in Section 2.2.3 after discussing data design and identification issues that will motivate and justify some of the extensions.

2.2.2 Data Design and Identification

The dIRT model has five variance-covariance parameters for the individual and dyadic latent traits that imply five "reduced-form parameters" for the composite latent trait: one constant variance, $\text{Var}(\theta_{a,p}) = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2$, and four distinct non-zero covariances, $\text{cov}(\theta_{a,p}, \theta_{p,a}) = 2\rho_{\alpha\beta}\sigma_\alpha\sigma_\beta + \rho_\gamma\sigma_\gamma^2)$, $\text{cov}(\theta_{a,p}, \theta_{a,q}) = \sigma_\alpha^2$, $\text{cov}(\theta_{a,p}, \theta_{b,p}) = \sigma_\beta^2$, and $\text{cov}(\theta_{a,p}, \theta_{b,a}) = \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta$ (where a, p, b, q are all different individuals). It is straightforward to find unique solutions for the five variance-covariance parameters from the five equations above, showing that the they are identified if the reduced-form parameters (variance and covariances) are identified.

The reduced-form parameters are identified if all the pairs of dyads involved in the covariances exist, i.e., actor/partner role reversal (sometimes referred to as "reciprocals") must occur to identify $\text{cov}(\theta_{a,p}, \theta_{p,a})$ and it must be possible to belong to more than one dyad. Specifically, it must be possible for actors to be paired with several partners to identify $\text{cov}(\theta_{a,p}, \theta_{a,q})$, for partners to be paired with several actors to identify $\text{cov}(\theta_{a,p}, \theta_{b,p})$, and for an actor paired with a partner p to also occur in a dyad as a partner of an actor $b \neq p$ to identify $\text{cov}(\theta_{a,p}, \theta_{b,a})$. It is necessary to set some mean parameters and/or step difficulty parameters to constants for identification. Here, we set the expectations of the latent traits to zero ($\mu_\alpha = \mu_\beta = \mu_\gamma = 0$), and allow the item step difficulties $\delta_{i,j}$ to be unconstrained (anchoring on latent trait scores instead of item difficulties), except that $\delta_{i,0} = 0$.

We have implicitly assumed that dyads and individuals within dyads are exchangeable by restricting the mean of $\theta_{a,p}$ to be constant (set to 0 to identify the step-difficulties) and allowing for only five distinct second-order reduced form parameters (one variance and four covariances), i.e., by assuming that the variance is constant and that covariances between dyadic composite latent variables depend only on the actor/partner roles of the individuals that are present in both dyads. The corresponding five parameters σ^2_α, σ^2_β, σ^2_γ, $\rho_{\alpha\beta}$, and ρ_γ enforce no other constraints besides exchangeability and positive semi-definiteness. Li and Loken (2002) make the point, for a regular SRM, that the model is in that sense justified by exchangeability.

When dyads occur naturally, such as in families or work settings, and where different individuals have different roles (e.g., father and daughter) or when interest centers on asymmetric relationships (e.g., supervisor and trainee), the exchangeability restrictions enforced by the model are no longer justified and we discuss how to relax them in Section 2.2.3.1. A special case of non-exchangeability is where each dyad is composed of individuals from two different groups, such as husbands and wives, and these groups are the same across dyads, so that there cannot, for example, be husband and wife dyads as well as father and daughter dyads. Kenny et al. (2006) refer to this design as distinguishable dyads.

We now explore several dyadic designs for which the SRM is identified, following Kenny and La Voie (1984) and Malloy and Kenny (1986). The simplest and most common design is the round robin design. In this design, each individual belongs to a dyad with every other member of the study, and there are a total of $\frac{n(n-1)}{2}$ dyads and $n(n-1)$ directed dyads. In graph theoretic language where we view each individual as a node, the round-robin design is represented by a complete graph in the undirected case (see upper-left panel of Figure 2.1), and a complete directed graph (digraph) in the directed case.

One immediate extension of the round-robin design is the block design where the n individuals are split into two blocks of sizes p and q respectively, and $p + q = n$. Then, each individual from every block forms a dyad with every individual from the other block, but not with individuals in his/her block. That is, there are a total of pq undirected dyads and $2pq$ directed dyads. In graph theoretic terms, such a design can be represented by a complete bipartite graph (see upper-right panel of Figure 2.1). This occurs most naturally for distinguishable dyads, for example when interactions only occur between members of the opposite gender. In this case, the n individuals are split into two blocks by their gender. Kenny et al. (2006) refer to such a design as an asymmetric block design.

When individuals are nested in groups, such as families, work groups, or social networks, where each individual from the group forms a dyad with each other individual of the group, we have a "k-group round-robin design" (see lower-left panel of Figure 2.1). In addition to such naturally occurring groups, the groups can also be created by the researcher to reduce response burden and costs by reducing the number of partners per actor and the number of dyads, respectively. Another reason for creating groups artificially is to allow individuals to interact within a group as a way to create the context for the dyadic responses. For example, Christensen and Kashy (1998) created an initial social situation for groups of four lonely individuals that involved problem-solving tasks and subsequently collected dyadic ratings on personal characteristics. There can also be a block design within each group, resulting in the "k-group block design" (see lower-right panel of Figure 2.1). This is the data design for the speed-dating application in Section 2.4.

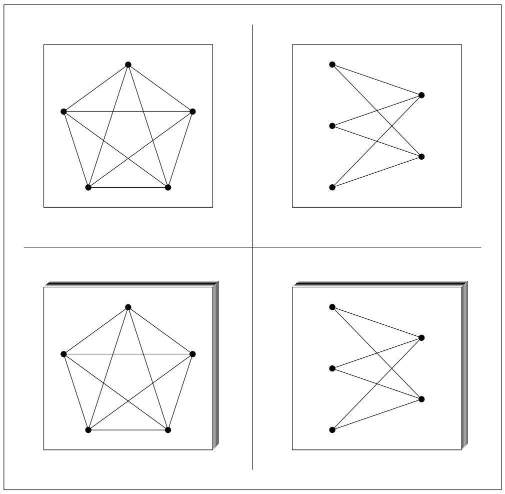


Figure 2.1: Graphs representing round-robin design (upper-left panel), block design (upper-right panel), k-group round-robin design (lower-left panel), and k-group block design (lower-right panel). For the k-group designs each group is represented by a layer.

2.2.3 Extended dIRT Model

2.2.3.1 Including Covariates for the Latent Traits

The dIRT model can be extended to take into account individual and dyadic covariates that may affect the latent traits at both the individual and dyadic levels by generalizing the idea of explanatory item response models (e.g., De Boeck & Wilson, 2004). One way that this can be accomplished is by specifying how the means μ_α, μ_β, and μ_γ depend on covariates, such as

$$\mu_{\alpha,a} \; = \; \boldsymbol{x}'_{\alpha,a}\boldsymbol{c}_\alpha, \quad \mu_{\beta,p} \; = \; \boldsymbol{x}'_{\beta,p}\boldsymbol{c}_\beta, \quad \mu_{\gamma,a,p} \; = \; \boldsymbol{x}'_{\gamma,a,p}\boldsymbol{c}_\gamma,$$

where $\boldsymbol{x}_{\alpha,a}$ are the covariates for α_a, $\boldsymbol{x}_{\beta,p}$ are the covariates for β_p, $\boldsymbol{x}_{\gamma,a,p}$ are the covariates for $\gamma_{a,p}$, and $\boldsymbol{c}_\alpha$, $\boldsymbol{c}_\beta$ and $\boldsymbol{c}_\gamma$ are the corresponding regression coefficients. If the dyads are all pairs of individuals within a family (k-group round-robin design), the covariates can include dummy variables for the roles, e.g., for the actor being a father (Snijders & Kenny, 1999).

Keeping in mind that the response probability for actor a when combined with partner p is a function of the composite latent variable $\theta_{a,p}$, whose mean is $\mu_{\alpha,a} + \mu_{\beta,p} + \mu_{\gamma,a,p}$, care must be taken to ensure that the regression coefficients are identified. For instance, if one of the covariates for $\mu_{\gamma,a,p}$ is the difference in some attribute, $z_a - z_p$, between the actor and partner, it is not possible to also include both the attribute for the actor, z_a, in the model for $\mu_{\alpha,a}$ and the attribute for the partner, z_p, in the model for $\mu_{\beta,p}$. Another example where identification is impossible is where dyads are males paired only with females (i.e., if the actor is a male, then the partner must be a female and vice versa) and gender is included as a covariate in the models for both $\mu_{\alpha,a}$ and $\mu_{\beta,p}$. Such an example is described in greater detail in Section 2.4.

It is also possible to allow the variances of the latent traits to depend on covariates, for instance to have different variances for different roles within families (Snijders & Kenny, 1999). Such an approach allows modeling non-exchangeable dyads in general.

2.2.3.2 Including Random Effects for the Latent Traits

If the individuals are clustered in different ways, e.g., in schools and/or neighborhoods, it may make sense to include cluster-level random effects into the models for α_a and β_p, to allow the actor and partner effects to be higher, on average, in some clusters than others, or, in other words, to have intraclass correlations. (For simplicity, we do not consider random effects in the model for $\gamma_{a,p}$.)

An obvious specification would be to introduce corresponding cluster-level actor and partner effects, A_j and B_j, respectively, for cluster j. The corresponding expression for $\theta_{a,p}$ then becomes

$$\theta_{a,p} \; \equiv \; \alpha_a + \beta_p + \gamma_{a,p} + A_{j[a]} + B_{j[p]},$$

where $j[a]$ is the cluster that individual a belong to. We could specify a bivariate normal distribution for A_j, B_j with variances σ_A^2, σ_B^2 and correlation $\rho_{A,B}$.

These three additional parameters are identified if individuals in the same dyad can belong to different clusters. In this case, $\text{cov}(\theta_{a,p}, \theta_{b,q}) = 0$ if the four different individuals a, p, b, q come from four different clusters. Otherwise, we add σ_A^2 to the covariance if and only if (iff) $j[a] = j[b]$, σ_B^2 iff $j[p] = j[q]$, $\rho_{A,B}\sigma_A\sigma_B$ iff either $j[a] = j[q]$ or $j[p] = j[b]$ but not both, and $2\rho_{A,B}\sigma_A\sigma_B$ iff $j[a] = j[q]$ and $j[p] = j[b]$. Depending on the cluster memberships of these four individuals, each of these terms can be added in isolation or in combination, producing eight distinct covariances. The parameters σ_A^2, σ_B^2 and $\rho_{A,B}$ are identified from these reduced form parameters alone. Further distinct covariances arise if, for instance, the actor is the same individual in both dyads. In this case, $\text{cov}(\theta_{a,p}, \theta_{a,q}) = \sigma_a^2 + \sigma_A^2$ if the different individuals, a, p and q, all belong to different clusters and we follow the same rules as above for adding the other terms besides σ_A^2.

However, if dyads are formed only among individuals within the same cluster, e.g., students are paired only with other students from the same school, then the term $\sigma_A^2 + \sigma_B^2 + 2\rho_{A,B}\sigma_A\sigma_B$ appears in all variances and covariances unless the two dyads belong to different clusters. This can occur only if the two dyads do not share any individuals in common, in which case we obtain $\text{cov}(\theta_{a,p}, \theta_{b,q}) = \sigma_A^2 + \sigma_B^2 + 2\rho_{A,B}\sigma_A\sigma_B$ if dyad (a, p) belongs to the same cluster as dyad (b, q) and $\text{cov}(\theta_{a,p}, \theta_{b,q}) = 0$, otherwise. It follows that only the sum $\sigma_A^2 + \sigma_B^2 + 2\rho_{A,B}\sigma_A\sigma_B$ is identified and therefore it makes sense to define $u_j \equiv A_j + B_j$, with one variance parameter σ_u^2, and to include u_j directly in the model for $\theta_{a,p}$.

It is of course possible to handle multiple nested or non-nested classifications by adding the corresponding random intercepts u if dyads are formed within a classification and A and B if dyads are formed across classifications (e.g., neighborhood when dyads are formed within schools or firms). Non-exchangeability can be handled by specifying different (co)variances for u or for A and B for different groups of individuals.

2.2.3.3 Distal Outcomes

The dIRT model can be extended by using, for instance, Generalized Linear Models to model one or more distal outcomes, where α_a, β_p, and $\gamma_{a,p}$ are latent covariates.

For example, we can consider a binary distal outcome $d_{a,p}$ of a dyad (a, p) taking the value of 1 with the conditional probability $\pi_{a,p}$ given the latent traits, and 0 otherwise. For the speed-dating application considered in Section 4, the distal outcome is whether each actor in a dyad elected to see the partner again. Here, $\pi_{a,p}$ can be modeled using the logistic regression

$$\log\left(\frac{\pi_{a,p}}{1-\pi_{a,p}}\right) = b_0 + b_1\alpha_a + b_2\alpha_p + b_3\beta_a + b_4\beta_p + b_5\gamma_{a,p} + b_6\gamma_{p,a} + b_7\alpha_a\alpha_p + b_8\beta_a\beta_p + b_9\gamma_{a,p}\gamma_{p,a}. \tag{2.4}$$

Distal outcome regressions can also include covariates of both the individual and the dyad if necessary.

Notice that in the above example, the distal outcome is directed in the sense that it depends on which individual in the dyad plays the role of the actor, and the individuals are therefore not exchangeable in the sense that the effect of α_a on the distal outcome for actor a is not necessarily the same as the effect of α_p. If the distal outcome is undirected, however, and individuals within dyads are exchangeable (e.g., in the case of pairs of individuals participating in a collaborative problem solving task where the outcome of interest is how well the task was completed per pair), then, (2.4) should be constrained to have $b_1 = b_2$, $b_3 = b_4$, and $b_5 = b_6$. If there is one undirected outcome per dyad and the individuals in the dyad are non-exchangeable (e.g., males paired with females), such a constraint is not needed if, for instance, a represents the male and p the female in the dyad.

2.2.4 Relationship with Other Models

We first review models for dyadic designs for which the dIRT and SRM are not identified, either because individuals can belong only to one dyad or because actor/partner role reversal is not possible.

Starting with the situation where individuals belong to only one grouping (dyad or larger group), the dIRT model reduces to a multilevel IRT where $\theta_{a,g} = \zeta_g + \zeta_{a,g}$, sometimes called a variance components factor/IRT model (e.g., Rabe-Hesketh, Skrondal, & Pickles, 2004). Here ζ_g is a group-level random intercept and $\zeta_{a,g}$ an individual-level random intercept.

In the dyadic data literature, the most popular model for this case is the Actor-Partner-Interdependence Model (APIM) proposed by Kenny (1996). The APIM for distinguishable dyads is basically a bivariate regression model where the actor's and partner's continuous responses y_a and y_p are both regressed on the covariates x_a and x_p of both the actor and the partner:

$$y_a = b_1x_a + c_1x_p + \zeta_a, \quad y_p = c_2x_a + b_2x_p + \zeta_p,$$

where the disturbances ζ_a and ζ_p are correlated. Here, b_1 and b_2 are interpreted as actor effects and c_1 and c_2 as partner effects. In the exchangeable APIM the actor effects are constrained to be equal, $b_1 = b_2$, as are the partner effects, $c_1 = c_2$, and the variances, $\text{Var}(\zeta_a) = \text{Var}(\zeta_p)$. Generalizations of the classical APIM have also been proposed. For example, Loeys and Molenberghs (2013) used generalized linear mixed models for categorical

y_a and y_p. Alexandrowicz (2015) replaced the observed variables, x_a, x_p and y_a, y_p, in the APIM by latent variables measured by multiple items via IRT models.

The mutual-influence model (Kenny, 1996) has no partner covariate effects which allows a reciprocal or mutual relationship between the responses of the actor and partner:

$$y_a \;=\; d_1 y_p + b_1 x_a + \zeta_a, \quad y_p \;=\; d_2 y_a + b_2 x_p + \zeta_p,$$

where ζ_a and ζ_p are correlated. This is a simultaneous equation model where d_1 and d_2 represent the mutual influence between the responses in a pair and x_a and x_p serve as instrumental variables for the endogenous explanatory variables y_a and y_p, respectively. In the exchangeable version (Duncan, Haller, & Portes, 1968), the actor effects are constrained to be equal, $b_1 = b_2$, as are the mutual effects, $d_1 = d_2$, and the variances, $\text{Var}(\zeta_a) = \text{Var}(\zeta_p)$.

In the Common-Fate Model (CFM) of Kenny and La Voie (1984) a dyad-level latent variable η_g for dyad g, measured by the continuous responses $y_{a,g}$ and $y_{p,g}$, is regressed on a dyad-level latent variable ξ_g, measured by the continuous covariates $x_{a,g}$ and $x_{p,g}$:

$$x_{a,g} \;=\; \xi_g + \delta_{a,g}, \quad x_{p,g} \;=\; \xi_g + \delta_{p,g}, \quad y_{a,g} \;=\; \eta_g + \epsilon_{a,g}, \quad y_{p,g} \;=\; \eta_g + \epsilon_{p,g},$$

$$\eta_g \;=\; \gamma \xi_g + \zeta_g.$$

The unique factors $\delta_{a,g}$ and $\epsilon_{a,g}$ for the actor-variables are correlated as are $\delta_{p,g}$ and $\epsilon_{p,g}$ for the partner-variables. Hence, the relationships between the variables is decomposed into a dyad-level relation (represented by γ) and two individual-level relations (represented by the error covariances, $\text{cov}(\delta_{a,g}, \epsilon_{a,g})$ and $\text{cov}(\delta_{p,g}, \epsilon_{p,g})$). In the exchangeable case, the following constraints are necessary: $\text{Var}(\delta_{a,g}) = \text{Var}(\delta_{p,g})$, $\text{Var}(\epsilon_{a,g}) = \text{Var}(\epsilon_{p,g})$, $\text{cov}(\delta_{a,g}, \epsilon_{a,g}) = \text{cov}(\delta_{p,g}, \epsilon_{p,g})$. To use the CFM in an IRT framework, Alexandrowicz, 2015 simply allowed all items measuring the latent versions of $x_{a,g}$ and $x_{p,g}$ to load on ξ_g and all items measuring the latent versions of $y_{a,g}$ and $y_{p,g}$ to load on η_g. We believe that a more appropriate approach would have been to replace each of $x_{a,g}$, $x_{p,g}$, $y_{a,g}$ and $y_{p,g}$ by a separate (first-order) latent variable, so that ξ_g and η_g become second-order latent variables and the error covariances of the CFM can be directly accommodated as covariances among the disturbances of the first-order latent variables.

We now discuss the situation where individuals appear in multiple dyads but actor/partner role reversals (or reciprocals) do not occur. For example, if the dyads are raters and examinees (with each examinee rated by several raters and each rater rating several examinees) only the raters provide responses so that the raters are always the actors and the examinees are always the partners. Then α_a is the rater leniency, β_p the examinee ability and $\gamma_{a,p}$, interpretable as person-specific rater leniency, can be included only if raters assesses several items by the same examinee (see, e.g., Shin, Rabe-Hesketh, and Wilson, 2019). In such a design, $\rho_{\alpha\beta}$ and ρ_γ are not defined because examinees and raters never switch roles.

We now turn to designs of the kind discussed in Section 2.2.2, where the SRM or dIRT are identified. Kenny and La Voie (1984) defined the original SRM for a continuous observed outcome y of actor a in the presence of partner p measured over multiple time points t as

$$y_{a,p,t} \;=\; \alpha_a + \beta_p + \gamma_{a,p} + \epsilon_{a,p,t}.$$

Here, $\epsilon_{a,p,t}$ can be viewed as test-retest measurement error. Since this term reduces to $\epsilon_{a,p}$ if there is only one time point, the identifiability of the above model, and in particular the variance of $\gamma_{a,p}$ separately from that of $\epsilon_{a,p,i}$, hinges crucially on measurements of the same dyad across multiple time points. In the dIRT model, multiple items essentially play the role of multiple time points, allowing for identification of the variance of $\gamma_{a,p}$. If one does not have multiple items or time-points, the model may still be identifiable if we assume that for two individuals a and p, $\gamma_{a,p} = \gamma_{p,a}$. That is, we assume that the dyadic effect of the pair is symmetric across the role that the individuals play and in that sense is no longer directional. In this case, $\gamma_{a,p}$ simply induces additional dependence between the responses for a given dyad and we can alternatively replace $\gamma_{a,p} + \epsilon_{a,p}$ by a single error term, typically denoted $\gamma_{a,p}$, that is correlated across members of the same dyad. In a k-group round-robin design, it may also make sense to include a group-level random intercept, for instance, when the groups are families, with each pair of family members forming a dyad (Loncke et al., 2018; Snijders & Kenny, 1999). Such a model is described at the end of Section 2.2.3.2.

In genetic experiments, a diallel cross is the set of all possible matings between several genotypes. The genotypes may be defined as individuals, clones, homozygous lines, etc. (Hayman, 1954). Some quantitative trait is measured for offspring from father a and mother p, and there are reciprocal crosses, with the role of mother and father reversed. Li and Loken (2002) show the correspondence between the SRM and a diallel model used in genetics (e.g., Cockerham & Weir, 1977):

$$y_{a,p} \;=\; \mu + g_a + g_p + s_{a,p} + d_a - d_p + r_{a,p},$$

where all terms are uncorrelated, except for g_a and d_a, and where $s_{a,p} = s_{p,a}$ and $r_{a,p} = -r_{p,a}$. The correspondence with the SRM is that $\alpha_a = g_a + d_a$, $\beta_p = g_p - d_p$, and $\gamma_{a,p} = s_{a,p} + r_{a,p}$.

Multivariate extensions of the SRM have been proposed for the situation where actors provide ratings on several continuous variables (Lüdtke, Robitzsch, & Trautwein, 2018; Nestler, 2018). For the case with a single time-point, the model can be written as

$$y_{a,p,i} \;=\; \alpha_{a,i} + \beta_{p,i} + \gamma_{a,p,i}$$

for variable i, where $\epsilon_{a,p,i}$ has been removed because only one error term (correlated across members of the same dyad) can be included. Unstructured covariance matrices are specified for each of these terms across variables and, in addition to the same-variable covariances between $\alpha_{a,i}$ and $\beta_{a,i}$ and between $\gamma_{a,p,i}$ and $\gamma_{p,a,i}$ that are part of a univariate SRM, the model allows for all corresponding cross-variable covariances as well. As far as we know, the common factor analogue to our dIRT model (where the measurement model for $\theta_{a,p}$ is a univariate factor model) has not been discussed in the literature.

We are aware of only very few papers that extend the classic SRM model to handle non-continuous responses, such as Koster and Leckie (2014) who used bivariate Poisson models for counts and Koster and Brandy (2018) who used bivariate probit models for binary responses.

2.3 Estimation

The dIRT model includes crossed random effects so that the marginal likelihood involves high-dimensional integrals. For example, in a k-group block design, the dimensionality of integration for the likelihood contribution of a group is $p + 1$ or $q + 1$, whichever is smaller (Goldstein, 1987). Numerical integration or Monte Carlo integration quickly becomes prohibitive and approximate methods are often not satisfactory (see, e.g., Jeon, Rijmen, and Rabe-Hesketh, 2017 and references therein). Fortunately, Bayesian estimation via Markov-chain Monte Carlo (MCMC) is feasible, and we adopt this approach here. Specifically, we use the the "No-U-Turn" sampler (Hoffman & Gelman, 2014) implemented in `Stan` (Stan Development Team, 2018). The `Stan` language affords us great flexibility in extending the basic dIRT model. We also verified all results using `Matlab` (version r2016b) via custom-written code based on the Metropolis-Hastings algorithm (Metropolis & Ulam, 1949).

To use MCMC, we define prior distributions for the parameters in (2.3) as well as the item parameters in (2.2) (and potentially the coefficients of the distal outcome regression in (2.4)). In our approach, we take the distributions of all hyperparameters $\sigma^2_\alpha, \sigma^2_\beta, \sigma^2_\gamma, \rho_{\alpha\beta}$ and ρ_γ to be noninformative by assuming uniform distributions for the variances $[0, +\infty)$ and correlations $[-1, 1]$. For step difficulties $\delta_{i,j}$ and regression coefficients $b_0, b_1, \ldots, b_9$ in the distal outcome model (2.4), we specify noninformative uniform priors $(-\infty, +\infty)$.

All parameter estimates were obtained using MCMC simulations of 4 chains with $2,000$ iterations, with a burn-in period of $1,000$ iterations. The parameter and hyper-parameter estimates are expected a posteriori (EAP) values obtained as means over the converged (post burn-in) MCMC draws for the four chains, i.e., they are based on an MCMC sample size of 4,000. Convergence was assessed by monitoring the $\hat{R}$ statistic (Gelman & Rubin, 1992).

The distal outcome model in (2.4) can be estimated jointly with the dIRT model by combining the log-likelihood contributions from the dIRT (ℓ_{dIRT}) and distal outcome (ℓ_{distal}) models in forming the joint log posterior of all parameters, given the dIRT item responses and distal outcome.

Joint estimation of the dIRT and distal outcome models is consistent and asymptoticaly efficient if both models are correctly specified. However, to protect against misspecification of the distal outcome (or "structural") model, a sequential approach could be used where the parameters of the dIRT ("measurement") model are estimated in step 1 and subsequent steps are used to obtain estimates of the structural (distal outcome) model parameters. If the measurement model is correctly specified, the estimates from step 1 are consistent even if the structural model is misspecified. However, if the structural model is correctly specified, joint estimation is more efficient than sequential approaches. From a conceptual point of view, it has been argued in the structural equation modeling and latent class literature, that altering the structural model by, for instance, adding or removing distal outcomes, affects the interpretation of the measurement model because these distal outcomes play a similar role to the items or indicators that define the latent traits. Sequential modeling can protect against such "interpretational confounding" (Burt, 1976) where the meaning of a construct is different from the meaning intended by the researcher (see Bakk and Kuha (2018) for further discussion).

The most obvious sequential approach is to use factor score regression (Skrondal & Laake, 2001) where one estimates the measurement model (step 1), obtains judiciously chosen scores for the latent traits from the measurement model (step 2), and substitutes these scores for the latent traits to estimate the structural model as if the latent traits were observed (step 3). This approach was adopted by Loncke et al. (2018) for SRMs. However, factor score regression is only consistent for link functions that are rarely of relevance in IRT (such as the identity) and naive standard errors from this approach are moreover underestimated. To address these limitations, a multiple imputation approach can be used, where multiple draws of the latent traits are obtained from their posterior distribution and the estimates for the structural model are combined using Rubin's formula (Rubin, 1987). Lüdtke et al. (2018) use such an approach in an SRM to estimate covariate effects on individual-level latent traits (i.e., as discussed in Section 2.2.3.1). Multiple imputation is natural in a Bayesian setting where full posteriors of the latent traits are available. A more straightforward pseudo-likelihood estimator, in the sense of Gong and Samaniego (1981), was proposed by Skrondal and Kuha (2012) (see also Bakk and Kuha (2018)). In this case the measurement model is first estimated, followed by joint estimation of the measurement and structural models under the constraint that the parameters of the measurement model are set equal to the estimates from the first stage.

We present the results of the joint approach in this chapter and include results for the sequential approach with multiple imputation in Appendix B.

2.4 Speed-Dating Application

We use a speed-dating dataset (Fisman, Iyengar, Kamenica, & Simonson, 2006) to examine the mutual attractiveness ratings of both individuals in a dyad to look for evidence of interactions that cannot be explained solely by the individuals' attractiveness or rating preferences. We also considered whether males and females differ in how they perceive their interactions. Additionally, by treating the final dating decision of whether the actor wants to see the partner again as a distal outcome, we investigate to what extent it relates to the dyadic latent trait.

The data was collected at 21 separate researcher-organized speed-dating sessions, over a period of 2 years, with 10-44 students from graduate and professional schools at Columbia University in each session. During these sessions, attended by nearly an equal number of male and female participants, all members of one gender would meet and interact with every member of the opposite gender for 5 minutes each. At the end of the 5 minute session, participants would rate their partner based on five attractiveness factors on a form attached to a clipboard that they were provided with. They also indicated whether or not they would like to see the person again.

After data cleaning, we had a total of 551 individuals, interacting in 4,184 distinct pairs, leading to 8,368 surveys completed (twice the number of pairs, given that both members of a pair rated each other). This corresponds to the "k-group block-dyadic" design described in Section 2.2.2. An illustrative example of data collected for one item in a balanced group of 10 individuals is shown in Figure 2.2.

		partner, p									
		F_1	F_2	F_3	F_4	F_5	M_6	M_7	M_8	M_9	M_{10}
actor, a	F_1						$y_{1,6}$	$y_{1,7}$	$y_{1,8}$	$y_{1,9}$	$y_{1,10}$
	F_2						$y_{2,6}$	$y_{2,7}$	$y_{2,8}$	$y_{2,9}$	$y_{2,10}$
	F_3						$y_{3,6}$	$y_{3,7}$	$y_{3,8}$	$y_{3,9}$	$y_{3,10}$
	F_4						$y_{4,6}$	$y_{4,7}$	$y_{4,8}$	$y_{4,9}$	$y_{4,10}$
	F_5						$y_{5,6}$	$y_{5,7}$	$y_{5,8}$	$y_{5,9}$	$y_{5,10}$
	M_6	$y_{6,1}$	$y_{6,2}$	$y_{6,3}$	$y_{6,4}$	$y_{6,5}$					
	M_7	$y_{7,1}$	$y_{7,2}$	$y_{7,3}$	$y_{7,4}$	$y_{7,5}$					
	M_8	$y_{8,1}$	$y_{8,2}$	$y_{8,3}$	$y_{8,4}$	$y_{8,5}$					
	M_9	$y_{9,1}$	$y_{9,2}$	$y_{9,3}$	$y_{9,4}$	$y_{9,5}$					
	M_{10}	$y_{10,1}$	$y_{10,2}$	$y_{10,3}$	$y_{10,4}$	$y_{10,5}$					

Figure 2.2: Example of responses $y_{a,p}$ of actor a rating partner p in a single-group block-dyadic structure consisting of 5 females $F_1, \ldots, F_5$ and 5 males $M_6, \ldots, M_{10}$.

In the data, the rating by actor a of partner p on item i is given by $y_{a,p,i}$. Each item was rated on a 10-point Likert-scale, which we collapsed to a 5-point scale by combining pairs of adjacent response categories to mitigate sparseness. Participants rated each other on 5 different items, all related to the overall attractiveness of the partner (viz. physical attractiveness, ambition, how fun they were, intelligence, and sincerity). We dropped all invalid ratings from an actor of a partner and the corresponding ratings from the partner of the actor even if the latter was valid. This amounted to a loss of less than 5% of the data.

In addition to each individual's rating of his/her partner, we also had access to an indicator $d_{a,p}$ for whether actor a elected to see partner p again. Note that this indicator is directional and $d_{a,p}$ may therefore differ from $d_{p,a}$. However, embedding the dIRT model within a distal outcome regression where the distal outcome is non-directional is also possible. For example, if we knew whether the dyad did in fact go on a date, this outcome would be unique to the dyad.

Using the joint MCMC estimation approach described in Section 2.3, the results of estimating the basic dyadic partial credit model (2.2) and the model also including a distal regression (2.4) are presented in Tables 2.1 and 2.2 under the heading "without gender". The code used to obtain the subsequent results is provided in Appendix A and explained in a `Stan` case-study (Sim, Gin, Skrondal, & Rabe-Hesketh, 2019). We estimate two versions of the distal regression, one with all 10 parameters $b_0, \ldots, b_9$ (labeled "with interactions"), and another model with $b_7 = b_8 = b_9 = 0$ (labeled "without interactions"). The estimates presented are the posterior means of the MCMC draws, and the values in parentheses represent the 2.5th and the 97.5th quantiles of the posterior distribution of the MCMC draws.

Table 2.1: Estimates of Standard Deviations and Correlations of Individual and Dyadic Latent Traits (Joint Approach)

	without gender				with gender	
	with interactions		without interactions		without interactions	
μ_{male}					0.08	(-0.10,0.24)
σ_α	1.03	(0.96,1.10)	1.03	(0.96,1.10)	1.03	(0.96,1.10)
σ_β	0.63	(0.58,0.68)	0.63	(0.58,0.68)	0.63	(0.58,0.69)
σ_γ	0.98	(0.95,1.02)	0.98	(0.95,1.01)	0.98	(0.95,1.02)
$\rho_{\alpha\beta}$	-0.06	(-0.17,0.04)	-0.06	(-0.16,0.04)	-0.07	(-0.17,0.03)
ρ_γ	0.46	(0.42,0.51)	0.46	(0.41,0.51)	0.46	(0.42,0.51)

2.4.1 Partitioning of Variance between Individual and Dyadic Latent Traits

Standard deviation and correlation estimates are reported in Table 2.1. In the dIRT, the variance of the composite latent variable $\theta_{a,p}$ is the sum of the variances of the individual and dyad-level latent traits, α_a, β_p and $\gamma_{a,p}$. It is instructive to examine the relative contributions of these latent traits to the composite. The percentage of the variance of $\theta_{a,p}$ that is due to α_a, β_p and $\gamma_{a,p}$ is estimated as 44%, 16% and 40%, respectively.

Table 2.2: Estimates for Distal Outcome Regression (Joint Approach)

	without gender				with gender	
	with interactions		without interactions		without interactions	
b_0	-0.87	(-1.03,-0.71)	-0.88	(-1.04,-0.73)	-0.88	(-1.04,-0.73)
b_1	0.15	(-0.04, 0.33)	0.14	(-0.05, 0.32)	0.14	(-0.04, 0.32)
b_2	-0.02	(-0.13, 0.09)	-0.02	(-0.13, 0.09)	-0.03	(-0.13, 0.08)
b_3	-3.03	(-3.62,-2.58)	-2.92	(-3.46,-2.49)	-2.91	(-3.44,-2.45)
b_4	3.56	(3.17, 4.01)	3.48	(3.12, 3.93)	3.48	(3.11, 3.91)
b_5	3.50	(3.06, 4.06)	3.42	(3.00, 3.95)	3.42	(2.99, 3.94)
b_6	0.17	(0.00, 0.35)	0.13	(-0.04, 0.29)	0.13	(-0.04, 0.29)
b_7	-0.01	(-0.13, 0.09)				
b_8	0.45	(0.06, 0.87)				
b_9	-0.28	(-0.53,-0.02)				

Interestingly, the variance of α_a is larger than that of β_p, implying that the actor's perception of the partner is more influenced by the actor's average tendency to rate others as attractive, which we could call actor leniency, than by the partner's average tendency to be rated as attractive, which we could think of as the partner's "universal" attractiveness. While the majority (60%) of the variance is accounted for by the individual effects (α_a and β_p), the dyadic effect ($\gamma_{a,p}$) accounts for a substantial proportion of the total variance, at 40%. A traditional IRT model, measuring individual latent traits only, would ignore this contribution, which can be thought of as the "eye-of-the-beholder" effect. In particular, this dyadic component would not be identifiable for standard IRT data where the individual only belongs to a single dyad.

2.4.2 Correlations

The within-person correlation $\rho_{\alpha\beta}$ of α_a and β_a reflects the relationship between how willing an individual was to rate someone else as attractive ("leniency"), and his/her own attractiveness. If this correlation is positive, it indicates that the more attractive an individual is, the more lenient he/she is in his/her ratings. If negative, it indicates that more attractive an individual is, the harsher he/she tends to be in rating his/her partners' attractiveness.

The between-person correlation ρ_γ of a dyad reflects the extent to which the (directed) dyadic trait is correlated between members of a given pair. If positive, it indicates that when an individual is affected by a social interaction with his/her partner, the partner will be more likely to also be affected in a similar manner. If negative, it suggests that members of a pair perceive their interaction in opposing ways.

Table 2.1 shows that the estimate of the correlation ρ_γ is positive with a 95% credible interval that does not contain zero. In contrast, the estimate of the correlation $\rho_{\alpha\beta}$ is negative with a 95% credible interval containing zero. The relatively larger estimated between-individual correlation indicates that members of each pair were likely to perceive their interaction similarly.

2.4.3 Distal Outcome Regression

We estimate the distal outcome regression for each individual's dating decision in (2.4) using the joint approach described in Section 2.3 and compare the regression estimates in Table 2.2 for the full model (under "with interactions") and a reduced model without interaction terms (under "without interactions").

We see that the estimated distal outcome regression coefficients are largest, in absolute value, for: a) the individual attractiveness α of both the actor ($\hat{b}_3$) and the partner ($\hat{b}_4$), and b) the unique relationship of the dyad γ from the actor's perspective ($\hat{b}_5$) but *not* for that from the partner's perspective ($\hat{b}_6$). Finding b) is consistent with our expectations given that the distal outcome reflects the viewpoint of the actor, rather than that of the partner. However, a less obvious finding is a) because the rater's own attractiveness, $\hat{b}_3$, *negatively* influences their dating decision. This suggests that the more attractive a rater was, the less likely they were to want to see the partner again. The estimated coefficients are tiny for the leniency of both the actor and the partner, as well as for the unique relationship of the dyad from the perspective of the partner, and have 95% credible intervals either containing zero or having one limit close to zero.

2.4.4 Gender Differences

Both the basic dIRT model and the model with a distal outcome can be extended to account for differences in the way females and males perceived their social interactions. In (2.3), we assumed that male and female participants shared the same expected leniency μ_α and attractiveness μ_β, by setting both of these expectations to zero. We can relax this by allowing the genders to have a different expectation for one of these parameters whilst setting the other to zero. The distribution for μ_α and μ_β becomes:

$$\begin{bmatrix} \alpha_a \\ \beta_a \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} m_a \mu_{\text{male}} \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta \\ \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{bmatrix} \right).$$

Here, μ_{male} is the difference between the expected attractiveness of males and females respectively, and m_a is an indicator for whether individual a is male. This gender parameter can also be interpreted as the difference between the expected leniency of females and males. Hence, a positive μ_{male} would suggest that males were on average more attractive than females, and/or females were more lenient in their ratings of males. We note that these effects could be disentangled if males rated other males and females rated other females. However, because we do not have such data, μ_{male} can only be interpreted as a linear combination (with unknown constants) of the average additional male attractiveness and average additional female rater leniency.

Estimates are reported under the heading "with gender" in the tables. The gender difference μ_{male} is estimated to be 0.08 with a 95% credible interval containing zero. There is therefore insufficient evidence to suggest a gender difference. The variance and correlation estimates are virtually the same for the models with and without μ_{male}.

2.5 Simulations

We first present the results of a simulation study exploring Bayesian properties of the MCMC estimator for an extended dIRT model that includes a distal outcome. We generated data for the same data design, size and parameter estimates as in the previous section. Starting with the estimated values of the variance and correlation hyperparameters, we generated 551 pairs of individual latent traits (α_a, β_a), and 8,126 directed dyadic latent traits $\gamma_{a,p}$. Using the estimated item step-difficulties from Section 2.4, we then generated responses from the dIRT model (2.2). Using the estimated regression coefficients, we finally generated the distal outcomes according to model (2.4). We summarize our findings regarding parameter recovery in the figures below.

Figure 2.3 depicts the difference between the estimated hyperparameters and the actual parameters across all 4,000 draws after convergence. The square represents the posterior mean of these estimates while the whiskers represent the bounds for the 95% credible intervals based on the 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ quantiles. Similarly, Figures 2.4 and 2.5 provide the analogous comparison for estimates of the item step parameters and the distal outcome regression parameters, respectively. We see that all credible intervals contains the true value and our procedure hence has good Bayesian performance.

In order to evaluate frequentist properties such as the bias of point estimates and the validity of model-based standard errors, we generated 50 datasets based on the same procedure as above, and estimated the same model for each dataset. Based on these 50 replications, we then estimated (i) the absolute bias of parameter estimates using the difference between the mean (over replications) of the estimated parameters and the true values, and (ii) the relative bias of standard error estimates using the mean (over replications) of the estimated standard errors divided by the empirical standard deviation (over replications) of the point estimates minus 1. Monte Carlo errors for these quantities were estimated using the formulae in White (2010).

In Figure 2.6 we show the estimated absolute bias of the parameter estimates (top) and relative error of the standard error estimates (bottom), together with error bars of ±1.96 times their Monte Carlo error estimates, representing approximate 95% confidence intervals

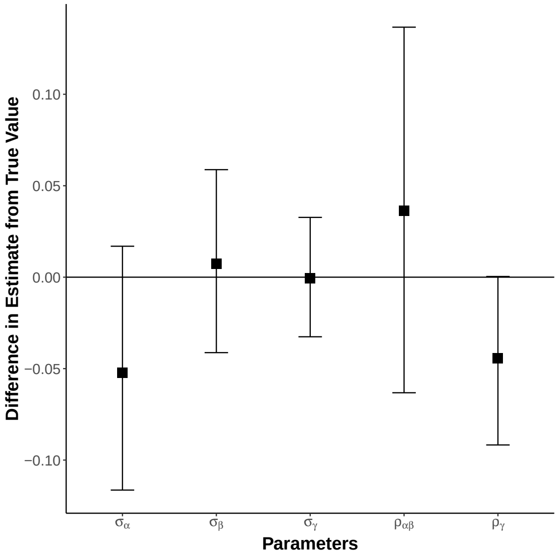


Figure 2.3: Difference between hyperparameter estimates and true parameter values.

if the sampling distributions are approximately normal. We see that there is small absolute bias in our point estimates across parameters, most of which can be attributed to Monte Carlo error with the exception of b_2, b_4 and b_5. There is also small relative bias for the standard error estimates, most of which can be attributed to Monte Carlo error with the exception of $\delta_{2,2}$ and $\delta_{3,4}$. In summary, our procedure has good frequentist properties.

2.6 Concluding Remarks

We have proposed a dyadic Item Response Theory (dIRT) model that integrates Item Response Theory (IRT) models for measurement and the Social Relations Model (SRM) for dyadic data by modeling the responses of an actor as a function of the actor's inclination to act and the partner's tendency to elicit that action as well as the unique relationship of the pair. We described how the model can be extended to larger group settings, include covariates for the individual and dyad, include cluster-level random effects, and accommodate distal outcomes. We also discussed data designs for which the dIRT model is identified, emphasizing that longitudinal data is not required, and described how the model can be estimated using standard software for Bayesian inference. The proposed estimation approach was shown to have good performance in simulation studies.

The practical utility of the dIRT model was demonstrated by applying it to speed dating data with ordinal items. The estimated variance of the actor effect suggests that there was some variation in the way different individuals rated the same sets of partners, or

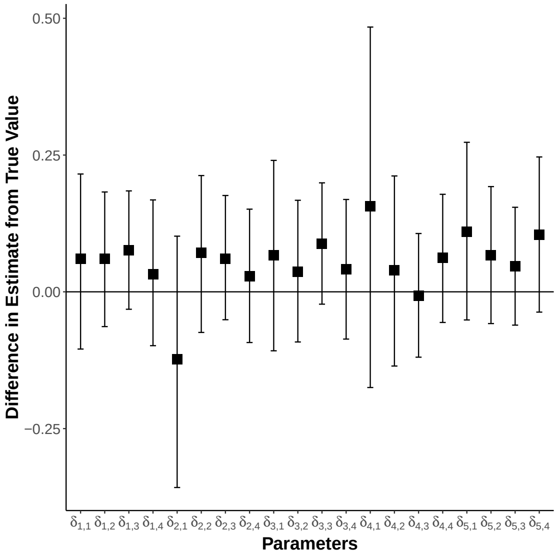


Figure 2.4: Difference between item step difficulty estimates and true parameter values.

in other words that there was a large variation in how lenient individuals were in rating their partners. The estimated variance of the partner effect can be thought of as reflecting how attractive the partner is, on average, to all other individuals, and indicates that there is some degree of universal attractiveness. We found that there is evidence for a unique interaction effect (dyadic latent trait) and that the magnitude of this effect helps predict whether the individuals want to see each other again. This finding suggests that the dyadic latent trait has predictive validity, a conclusion that can perhaps be more easily justified when a sequential estimation approach is used. A traditional IRT model, measuring individual latent traits only, would ignore this dyadic latent trait, which can be thought of as the "eye-of-the-beholder" effect. The dyadic latent traits were positively correlated within dyads, suggesting that both members of a dyad tended to perceive their interaction similarly.

In the speed-dating application, the dyadic latent trait was of particular interest from the point-of-view of matchmaking. In other applications where the actors can be viewed as the raters, "perceivers" or informants used to make inferences regarding the partners, the partner latent trait is of greatest interest. In this case, the advantage of the dIRT is that it purges the measurement of the partner latent trait from both the global rater bias α and the target-specific rater bias γ. In a collaborative problem-solving task, both the actor and partner latent traits may be of interest, in which case it becomes important to accommodate the dyadic latent trait in the model to prevent it from contaminating the individual latent traits of interest. The dyadic latent trait could in this case be viewed as a nuisance reflecting a fortunate or unfortunate choice of collaborator. For all these types of applications, dyadic designs that permit estimation of the dIRT are essential.

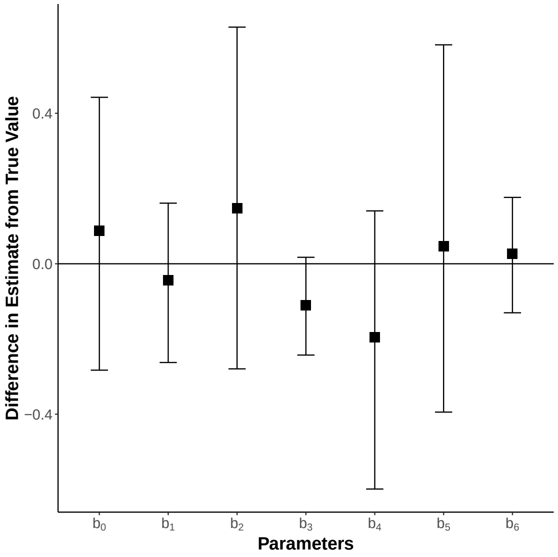


Figure 2.5: Difference between distal outcome regression estimates and true parameter values.

The formulation of the dIRT model, and providing a viable estimation approach for it, provides researchers with the impetus to collect appropriate data for investigating dyadic interactions or individual latent traits, free from such interaction effects, in a measurement context.

Acknowledgements

This chapter comprises work done by Brian Gin, the author of this dissertation, Anders Skrondal, and the chair of this dissertation committee. Brian Gin and the author of this dissertation are joint first authors of this chapter. Permission was sought from all authors as well as from the Berkeley Graduate Division for the work to be included as a chapter in this dissertation.

This work was partly funded by the Research Council of Norway through its Centres of Excellence funding scheme, project number 262700, and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D140059.

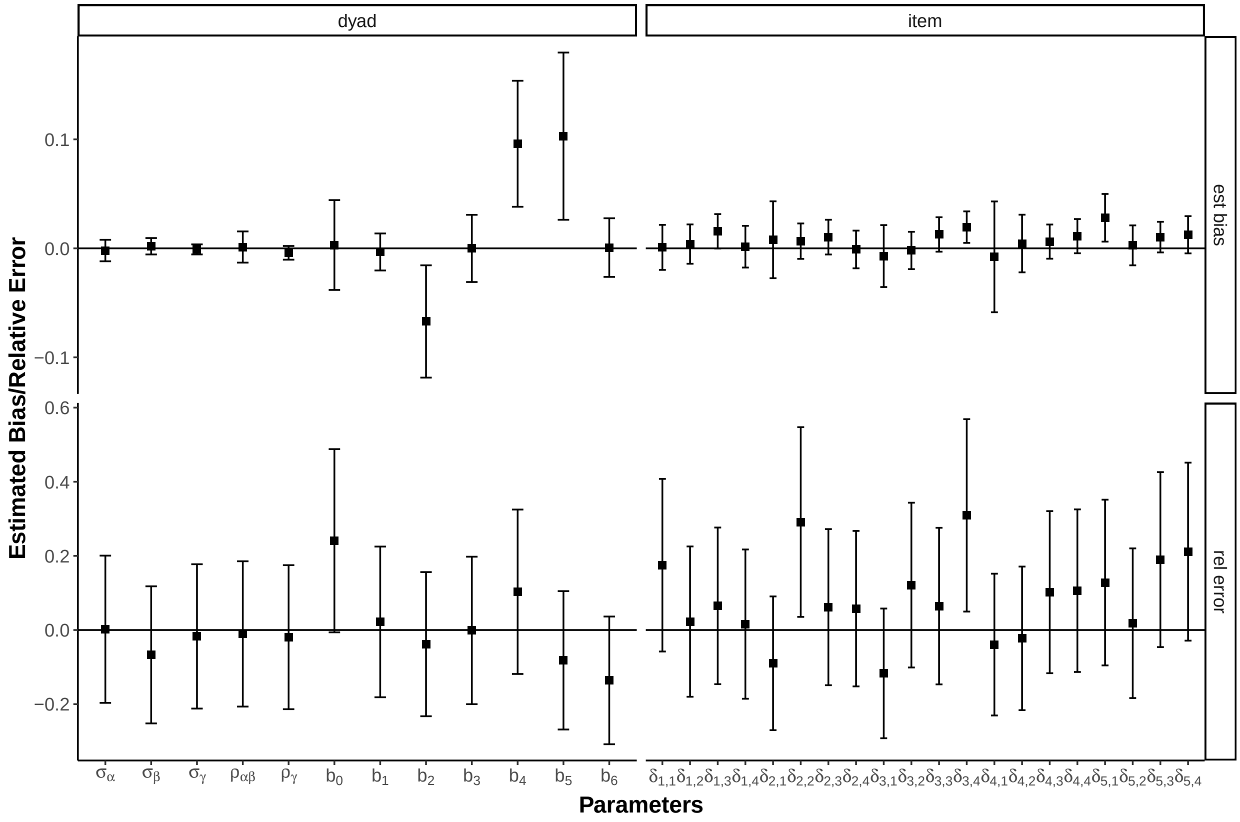


Figure 2.6: Performance of point estimates and standard errors across 50 replications.

Appendix A: Stan Code

```
# clears workspace:
rm(list = ls())

library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = 8)

library(tidyverse)

# load dataset:
load(file = "df.complete.Rdata")
load(file = "dpair.specific.Rdata")

# no gen with int stan model
modelngwi <- "
functions {
    real pcminteract(int x, real alpha, real beta, real gamma, vector delta) {
      vector[rows(delta) + 1] unsummed;
      vector[rows(delta) + 1] probs;
      unsummed = append_row(rep_vector(0.0, 1), alpha + beta + gamma - delta);
      probs = softmax(cumulative_sum(unsummed));
      return categorical_lpmf(x+1 | probs);
    }
  }
data {
  int<lower = 1> I;                  // # items
  int<lower = 1> A;                  // # actors (or partners)
  int<lower = 1> U;                  // # undirected pairs
  int<lower = 1> N;                  // # responses
  int<lower = 1> D;                  // # decisions
  int<lower = 1> B;                  // integer value for # distal regression parameters
  int<lower = 1, upper = A> aa[N];   // size N array to index actors for each response
  int<lower = 1, upper = A> pp[N];   // size N array to index partners for each response
  int<lower = 1, upper = I> ii[N];   // size N array to index items for each response
  int<lower = 0> x[N];               // size N array for responses; x = 0, 1 ... m_i
  int<lower = 1, upper = U> dd[N];   // size N array to index undirected pairs for each response
  int<lower = 1, upper = 2> mm[N];   // size N array to index match for each response
  int<lower = 1, upper = A> aaa[D];  // size D array to index actors for each decision
  int<lower = 1, upper = A> ppp[D];  // size D array to index partners for each decision
  int<lower = 1, upper = U> ddd[D];  // size D array to index undirected pairs for each decision
  int<lower = 1, upper = 2> mmm[D];  // size D array to index match for each decision
  int<lower = 0, upper = 1> zzz[D];  // size D array for decisions
}
transformed data {
  int M;                             // # parameters per item (same for all items)
    M = max(x);
}
parameters {
  vector[M] delta[I];                // length m vector for each item i
  vector[2] AB[A];                   // size 2 vector of alpha and beta for each person;
  vector[2] GG[U];                   // size 2 vector of gammas for each undirected pair;
  real<lower = 0> sigmaA;            // real sd of alpha
  real<lower = 0> sigmaB;            // real sd of beta
  real<lower = 0> sigmaG;            // real sd of gamma
  real<lower = -1, upper = 1> rhoAB; // real cor between alpha and beta (within person)
  real<lower = -1, upper = 1> rhoG;  // real cor between gammas (within pair)
  real beta[B];                      // B-dimensional array of real valued of beta
                                     // (distal regression parameters)
}
transformed parameters {
  cov_matrix[2] SigmaAB;             // 2x2 covariance matrix of alpha and beta
  cov_matrix[2] SigmaG;              // 2x2 covariance matrix of gammas
  SigmaAB[1, 1] = sigmaA^2;
  SigmaAB[2, 2] = sigmaB^2;
  SigmaAB[1, 2] = rhoAB * sigmaA * sigmaB;
  SigmaAB[2, 1] = rhoAB * sigmaA * sigmaB;
  SigmaG[1, 1] = sigmaG^2;
  SigmaG[2, 2] = sigmaG^2;
  SigmaG[1, 2] = rhoG * sigmaG^2;
  SigmaG[2, 1] = rhoG * sigmaG^2;
}
model {
  AB ~ multi_normal(rep_vector(0.0, 2), SigmaAB);
  GG ~ multi_normal(rep_vector(0.0, 2), SigmaG);
  for (n in 1:N){
    target += pcminteract(x[n], AB[aa[n],1], AB[pp[n],2], GG[dd[n], mm[n]], delta[ii[n]]);
  }
  for (d in 1:D){
    //distal logistic regression
    target += bernoulli_logit_lpmf(zzz[d] | (beta[1]
    + beta[2]*AB[aaa[d],1]
    + beta[3]*AB[ppp[d],1]
    + beta[4]*AB[aaa[d],2]
    + beta[5]*AB[ppp[d],2]
    + beta[6]*GG[ddd[d], mmm[d]]
    + beta[7]*GG[ddd[d], (3-mmm[d])]
    + beta[8]*AB[aaa[d],1]*AB[ppp[d],1]
    + beta[9]*AB[aaa[d],2]*AB[ppp[d],2]
    + beta[10]*GG[ddd[d], mmm[d]]*GG[ddd[d], (3-mmm[d])]));
  }
}
"

# no gen with int model
I <- max(df.complete$item)
A <- max(df.complete$actor)
U <- max(df.complete$unique.pair)
N <- nrow(df.complete)
D <- nrow(dpair.specific)
B <- 10

data <- list(I = I,
             A = A,
             U = U,
             N = N,
             D = D,
             B = B,
             aa = as.numeric(df.complete$actor),
             pp = as.numeric(df.complete$partner),
             ii = as.numeric(df.complete$item),
             x = as.numeric(df.complete$x),
             dd = as.numeric(df.complete$unique.pair),
             mm = as.numeric(df.complete$selector),
             aaa = as.numeric(dpair.specific$actor),
             ppp = as.numeric(dpair.specific$partner),
             ddd = as.numeric(dpair.specific$unique.pair),
             mmm = as.numeric(dpair.specific$selector),
             zzz = as.numeric(dpair.specific$decision))

set.seed(349)
samples <- stan(model_code=modelngwi,
                 data=data,
                 iter=2000,
                 chains=4,
                 seed = 349)

pcm_estimated_values <- summary(samples,
                                pars = c("sigmaA",
                                         "sigmaB",
                                         "sigmaG",
                                         "rhoAB",
                                         "rhoG",
                                         "beta"),
                                probs = c(.025, .975))
View(pcm_estimated_values$summary)

# no gen no int stan model
modelngni <- "
functions {
    real pcminteract(int x, real alpha, real beta, real gamma, vector delta) {
      vector[rows(delta) + 1] unsummed;
      vector[rows(delta) + 1] probs;
      unsummed = append_row(rep_vector(0.0, 1), alpha + beta + gamma - delta);
      probs = softmax(cumulative_sum(unsummed));
      return categorical_lpmf(x+1 | probs);
    }
  }
data {
  int<lower = 1> I;                  // # items
  int<lower = 1> A;                  // # actors (or partners)
  int<lower = 1> U;                  // # undirected pairs
  int<lower = 1> N;                  // # responses
  int<lower = 1> D;                  // # decisions
  int<lower = 1> B;                  // integer value for # distal regression parameters
  int<lower = 1, upper = A> aa[N];   // size N array to index actors for each response
  int<lower = 1, upper = A> pp[N];   // size N array to index partners for each response
  int<lower = 1, upper = I> ii[N];   // size N array to index items for each response
  int<lower = 0> x[N];               // size N array for responses; x = 0, 1 ... m_i
  int<lower = 1, upper = U> dd[N];   // size N array to index undirected pairs for each response
  int<lower = 1, upper = 2> mm[N];   // size N array to index match for each response
  int<lower = 1, upper = A> aaa[D];  // size D array to index actors for each decision
  int<lower = 1, upper = A> ppp[D];  // size D array to index partners for each decision
  int<lower = 1, upper = U> ddd[D];  // size D array to index undirected pairs for each decision
  int<lower = 1, upper = 2> mmm[D];  // size D array to index match for each decision
  int<lower = 0, upper = 1> zzz[D];  // size D array for decisions
}
transformed data {
  int M;                             // # parameters per item (same for all items)
    M = max(x);
}
parameters {
  vector[M] delta[I];                // length m vector for each item i
  vector[2] AB[A];                   // size 2 vector of alpha and beta for each person;
  vector[2] GG[U];                   // size 2 vector of gammas for each undirected pair;
  real<lower = 0> sigmaA;            // real sd of alpha
  real<lower = 0> sigmaB;            // real sd of beta
  real<lower = 0> sigmaG;            // real sd of gamma
  real<lower = -1, upper = 1> rhoAB; // real cor between alpha and beta (within person)
  real<lower = -1, upper = 1> rhoG;  // real cor between gammas (within pair)
  real beta[B];                      // B-dimensional array of real valued of beta
                                     // (distal regression parameters)
}
transformed parameters {
  cov_matrix[2] SigmaAB;             // 2x2 covariance matrix of alpha and beta
  cov_matrix[2] SigmaG;              // 2x2 covariance matrix of gammas
  SigmaAB[1, 1] = sigmaA^2;
  SigmaAB[2, 2] = sigmaB^2;
  SigmaAB[1, 2] = rhoAB * sigmaA * sigmaB;
  SigmaAB[2, 1] = rhoAB * sigmaA * sigmaB;
  SigmaG[1, 1] = sigmaG^2;
  SigmaG[2, 2] = sigmaG^2;
  SigmaG[1, 2] = rhoG * sigmaG^2;
  SigmaG[2, 1] = rhoG * sigmaG^2;
}
model {
  AB ~ multi_normal(rep_vector(0.0, 2), SigmaAB);
  GG ~ multi_normal(rep_vector(0.0, 2), SigmaG);
  for (n in 1:N){
    target += pcminteract(x[n], AB[aa[n],1], AB[pp[n],2], GG[dd[n], mm[n]], delta[ii[n]]);
  }
  for (d in 1:D){
    //distal logistic regression
    target += bernoulli_logit_lpmf(zzz[d] | (beta[1]
    + beta[2]*AB[aaa[d],1]
    + beta[3]*AB[ppp[d],1]
    + beta[4]*AB[aaa[d],2]
    + beta[5]*AB[ppp[d],2]
    + beta[6]*GG[ddd[d], mmm[d]]
    + beta[7]*GG[ddd[d], (3-mmm[d])]));
  }
}
"

# no gen with int model
I <- max(df.complete$item)
A <- max(df.complete$actor)
U <- max(df.complete$unique.pair)
N <- nrow(df.complete)
D <- nrow(dpair.specific)
B <- 7

data <- list(I = I,
             A = A,
             U = U,
             N = N,
             D = D,
             B = B,
             aa = as.numeric(df.complete$actor),
             pp = as.numeric(df.complete$partner),
             ii = as.numeric(df.complete$item),
             x = as.numeric(df.complete$x),
             dd = as.numeric(df.complete$unique.pair),
             mm = as.numeric(df.complete$selector),
             aaa = as.numeric(dpair.specific$actor),
             ppp = as.numeric(dpair.specific$partner),
             ddd = as.numeric(dpair.specific$unique.pair),
             mmm = as.numeric(dpair.specific$selector),
             zzz = as.numeric(dpair.specific$decision))

set.seed(349)
samples <- stan(model_code=modelngni,
                data=data,
                iter=2000,
                chains=4,
                seed = 349)

pcm_estimated_values <- summary(samples,
                                pars = c("sigmaA",
                                         "sigmaB",
                                         "sigmaG",
                                         "rhoAB",
                                         "rhoG",
                                         "beta",
                                         "delta"),
                                probs = c(.025, .975))
View(pcm_estimated_values$summary)

# with gen no int stan model
modelwgni <- "
functions {
    real pcminteract(int x, real alpha, real beta, real gamma, vector delta) {
      vector[rows(delta) + 1] unsummed;
      vector[rows(delta) + 1] probs;
      unsummed = append_row(rep_vector(0.0, 1), alpha + beta + gamma - delta);
      probs = softmax(cumulative_sum(unsummed));
      return categorical_lpmf(x+1 | probs);
    }
  }
data {
  int<lower = 1> I;                  // # items
  int<lower = 1> A;                  // # actors (or partners)
  int<lower = 1> U;                  // # undirected pairs
  int<lower = 1> N;                  // # responses
  int<lower = 1> D;                  // # decisions
  int<lower = 1> B;                  // integer value for # distal regression parameters
  int<lower = 1, upper = A> aa[N];   // size N array to index actors for each response
  int<lower = 1, upper = A> pp[N];   // size N array to index partners for each response
  int<lower = 1, upper = I> ii[N];   // size N array to index items for each response
  int<lower = 0> x[N];               // size N array for responses; x = 0, 1 ... m_i
  int<lower = 1, upper = U> dd[N];   // size N array to index undirected pairs for each response
  int<lower = 1, upper = 2> mm[N];   // size N array to index match for each response
  int<lower = 0, upper = 1> gg[N];   // size N array to index gender for each response
  int<lower = 1, upper = A> aaa[D];  // size D array to index actors for each decision
  int<lower = 1, upper = A> ppp[D];  // size D array to index partners for each decision
  int<lower = 1, upper = U> ddd[D];  // size D array to index undirected pairs for each decision
  int<lower = 1, upper = 2> mmm[D];  // size D array to index match for each decision
  int<lower = 0, upper = 1> zzz[D];  // size D array for decisions
}
transformed data {
  int M;                             // # parameters per item (same for all items)
```

```
    M = max(x);
}
parameters {
  vector[M] delta[I];                // length m vector for each item i
  vector[2] AB[A];                   // size 2 vector of alpha and beta for each person;
  vector[2] GG[U];                   // size 2 vector of gammas for each undirected pair;
  real<lower = 0> sigmaA;            // real sd of alpha
  real<lower = 0> sigmaB;            // real sd of beta
  real<lower = 0> sigmaG;            // real sd of gamma
  real<lower = -1, upper = 1> rhoAB; // real cor between alpha and beta (within person)
  real<lower = -1, upper = 1> rhoG;  // real cor between gammas (within pair)
  real mu;                           // real value of mean of theta for males
  real beta[B];                      // B-dimensional array of real valued of beta
                                     // (distal regression parameters)
}
transformed parameters {
  cov_matrix[2] SigmaAB;             // 2x2 covariance matrix of alpha and beta
  cov_matrix[2] SigmaG;              // 2x2 covariance matrix of gammas
  SigmaAB[1, 1] = sigmaA^2;
  SigmaAB[2, 2] = sigmaB^2;
  SigmaAB[1, 2] = rhoAB * sigmaA * sigmaB;
  SigmaAB[2, 1] = rhoAB * sigmaA * sigmaB;
  SigmaG[1, 1] = sigmaG^2;
  SigmaG[2, 2] = sigmaG^2;
  SigmaG[1, 2] = rhoG * sigmaG^2;
  SigmaG[2, 1] = rhoG * sigmaG^2;
}
model {
  AB ~ multi_normal(rep_vector(0.0, 2), SigmaAB);
  GG ~ multi_normal(rep_vector(0.0, 2), SigmaG);
  for (n in 1:N){
    target += pcminteract(x[n], AB[aa[n],1] - mu*gg[n], AB[pp[n],2], GG[dd[n], mm[n]], delta[ii[n]]);
  }
  for (d in 1:D){
    //distal logistic regression
    target += bernoulli_logit_lpmf(zzz[d] | (beta[1]
    + beta[2]*AB[aaa[d],1]
    + beta[3]*AB[ppp[d],1]
    + beta[4]*AB[aaa[d],2]
    + beta[5]*AB[ppp[d],2]
    + beta[6]*GG[ddd[d], mmm[d]]
    + beta[7]*GG[ddd[d], (3-mmm[d])]));
  }
}
"

# no gen with int model
I <- max(df.complete$item)
A <- max(df.complete$actor)
U <- max(df.complete$unique.pair)
N <- nrow(df.complete)
D <- nrow(dpair.specific)
B <- 7

data <- list(I = I,
             A = A,
             U = U,
             N = N,
             D = D,
             B = B,
             aa = as.numeric(df.complete$actor),
             pp = as.numeric(df.complete$partner),
             ii = as.numeric(df.complete$item),
             x = as.numeric(df.complete$x),
             dd = as.numeric(df.complete$unique.pair),
             mm = as.numeric(df.complete$selector),
             gg = as.numeric(df.complete$male),
             aaa = as.numeric(dpair.specific$actor),
             ppp = as.numeric(dpair.specific$partner),
             ddd = as.numeric(dpair.specific$unique.pair),
             mmm = as.numeric(dpair.specific$selector),
             zzz = as.numeric(dpair.specific$decision))

set.seed(349)
samples <- stan(model_code=modelwgni,
                data=data,
                iter=2000,
                chains=4,
                seed = 349)

pcm_estimated_values <- summary(samples,
                                pars = c("sigmaA",
                                         "sigmaB",
                                         "sigmaG",
                                         "rhoAB",
                                         "rhoG",
                                         "mu",
                                         "beta"),
                                probs = c(.025, .975))
View(pcm_estimated_values$summary)
```

Appendix B: Sequential Estimation

Using the sequential estimation approach with multiple imputation described in Section 2.3, the results of first estimating the dyadic partial credit model (ignoring the distal outcome), and subsequently estimating the distal regression are presented in Tables 2.3 and 2.4. We report estimates as means of draws, and the values in parentheses represent the 2.5$^{\text{th}}$ and the 97.5$^{\text{th}}$ quantiles of the parameter estimates.

MCMC estimates for the standard deviations and correlations of the individual and dyadic latent traits are shown in Table 2.3. The estimates are qualitatively similar to the estimates from the joint approach reported in Table 2.1.

Estimates for the distal regression based on multiple draws of the latent traits from their posterior distribution are shown in Table 2.4. While the sign of the coefficient estimates are the same as for the joint approach in Table 2.2, their magnitudes differ substantially. Overall, the estimates using the sequential approach are smaller in absolute value, particularly for b_3, b_4 and b_5. This may be because the joint approach effectively treats the distal outcome as an item in the measurement model, and therefore makes the latent traits highly predictive of the distal outcome.

Table 2.3: Sequential Estimation Approach: Estimates of Standard Deviations and Correlations of Individual and Dyadic Latent Traits

	without gender				with gender	
	with interactions		without interactions		without interactions	
μ_{male}					-0.15	(-0.36,0.06)
σ_α	1.05	(0.98,1.13)	1.05	(0.98,1.13)	1.05	(0.98,1.13)
σ_β	0.71	(0.66,0.76)	0.71	(0.66,0.76)	0.71	(0.66,0.76)
σ_γ	0.89	(0.86,0.92)	0.89	(0.86,0.92)	0.89	(0.86,0.91)
$\rho_{\alpha\beta}$	0.03	(-0.06,0.13)	0.03	(-0.06,0.13)	0.04	(-0.06,0.13)
ρ_γ	0.35	(0.30,0.40)	0.35	(0.30,0.40)	0.35	(0.30,0.40)

Table 2.4: Sequential Estimation Approach: Estimates for Distal Outcome Regression

	without gender				with gender	
	with interactions		without interactions		without interactions	
b_0	-0.36	(-0.43,-0.29)	-0.36	(-0.43,-0.29)	-0.36	(-0.43,-0.28)
b_1	0.40	(0.36, 0.44)	0.40	(0.36, 0.44)	0.38	(0.34, 0.43)
b_2	-0.03	(-0.07, 0.00)	-0.03	(-0.07, 0.00)	-0.02	(-0.06, 0.02)
b_3	-0.35	(-0.44,-0.26)	-0.34	(-0.42,-0.25)	-0.32	(-0.42,-0.23)
b_4	1.29	(1.19, 1.38)	1.28	(1.19, 1.37)	1.26	(1.16, 1.36)
b_5	0.82	(0.76, 0.89)	0.82	(0.77, 0.89)	0.82	(0.76, 0.88)
b_6	0.06	(0.01, 0.11)	0.06	(0.01, 0.11)	0.06	(-0.01, 0.11)
b_7	-0.01	(-0.04, 0.01)				
b_8	0.18	(0.09, 0.28)				
b_9	-0.02	(-0.07, 0.04)				

Chapter 3

Handling Endogeneity with Structural Equation Modeling

3.1 Introduction

Hierarchical Linear Modeling is a common modeling approach for clustered data. In particular, the Random Intercept Model (RIM) is a straightforward model for cross-sectional data with a nested structure (say students within schools, or patients within hospitals), or longitudinal data (where repeated observations of the same individual are nested within that individual). Like other linear models, there are occasions when an analyst would like to treat the parameter estimates of a RIM causally. In this setting, we call the model and its parameters "structural", and the error terms now represent the effects of omitted covariates (Castellano, Rabe-Hesketh, & Skrondal, 2014). Covariates in such a structural RIM are called cluster-endogenous if they are correlated with the random intercept, and their presence could lead to biased estimation of their structural coefficients in the RIM.

In this chapter, we explore some common approaches to achieve consistent estimation of the structural model parameters. In particular, we highlight the most common approach in Econometrics, the Hausman-Taylor estimator (Hausman & Taylor, 1981), that allows for consistent parameter estimation in a RIM even in the face of cluster-level endogeneity.

We then propose setting a RIM up as a SEM (Muthén, 1994). This allows us to treat covariates in the model as random rather than fixed, and in so doing, enables us to extend the model naturally to allow endogenous covariates to co-vary with the random intercept. Additionally, previous approaches (Allison, 2005, 2009; Allison, Williams, & Moral-Benito, 2017; Bollen & Brand, 2010; Teachman, Duncan, Yeung, & Levy, 2001) that use SEMs to model RIMs have implicitly treated the units as non-exchangeable. We explore modeling RIMs using SEMs that enforce exchangeability among units which are more suitable for most cross-sectional, clustered data the units within each cluster are exchangeable.

As this approach has never been compared to the Hausman-Taylor estimator, we simulate data to evaluate both the asymptotic as well as finite sample properties of both estimators. In particular, we show that while both estimators are consistent, they may perform differently in finite samples depending on (i) whether the data are balanced, and (ii) whether the data are assumed to be exchangeable. We also posit that with data that is assumed to be non-exchangeable, the SEM-based estimator performs better.

We highlight here that most simulations for clustered data are often overly simplistic and are not adequately explicit about assumptions regarding the covariance structure of the data. In this study, we propose a sequential simulation scheme that can be used to generate two-level data for any covariance structure. This is particularly important for us because unlike most of the previous work in this area which focused on panel data, we are interested in hierarchical data which necessitates additional exchangeability assumptions.

Additionally, we point out that in order to assess the asymptotic properties of the SEM estimator, we adopt an approach highlighted in Muthén, Kaplan, and Hollis (1987), which treats the model-implied population covariance matrix as data to determine how the estimator will perform as the number of clusters tends to infinity.

The structure of this chapter is as follows. We begin in Section 2 by reviewing concepts of endogeneity in Linear Models as well as in RIMs and the approaches one can take to estimate the structural parameters consistently. We then provide a detailed description of the Hausman-Taylor estimator and describe how one can also model the cluster-level endogeneity using SEM. In Section 3, we set up a simulation study accounting for the exchangeability assumptions in simulating the data. To generate the simulated data, we first carefully specify the population covariance matrix for all observed variables making sure to account for all variance-covariance parameters and extra exchangeability constraints. We then present findings regarding the asymptotic as well as finite-sample properties of both the Hausman-Taylor and Maximum-Likelihood SEM estimators. Finally in Section 4, we conclude with the pros and cons of using our proposed SEM approach, as well as suggest related areas where more work needs to be done.

3.2 Endogeneity

3.2.1 Endogeneity in a Linear Regression Model

When considering the linear regression model, $y_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \epsilon_i$, as a structural model, the Ordinary Least Squares (OLS) and Maximum Likelihood (ML) estimators for $\boldsymbol{\beta}$ are consistent under the assumption that $\boldsymbol{x}_i$ is exogenous - that is, $\mathbb{E}(\boldsymbol{x}_i\epsilon_i) = \mathbf{0}$, and hence, $\boldsymbol{x}_i$ is uncorrelated with ϵ_i. In some instances, however, such an assumption is not valid. For example, when (i) covariates are measured with error; (ii) the outcome and covariates are jointly related through simultaneous equations; or (iii) variables correlated with the outcome and the covariates are omitted from the model, it can be shown that $\mathbb{E}(\boldsymbol{x}_i \epsilon_i) \neq \mathbf{0}$. In these cases, we call $\boldsymbol{x}_i$ endogenous, and both the OLS and ML estimators for $\boldsymbol{\beta}$ are inconsistent.

In econometrics, the most common way of handling endogeneity (the existence of endogenous covariates in a model) is through the use of Instrumental Variables (IV) where the instruments $\boldsymbol{z}_i$ are correlated with $\boldsymbol{x}_i$ and not correlated with ϵ_i. It is worthwhile to note that the instrumental variables $\boldsymbol{z}_i$ can be existing variables in the assumed model. In this case, we call these variables *internal* instrumental variables.

3.2.2 Endogeneity in a Random Intercept Model

In the case where the structural model is hierarchical in nature (consisting of units nested within clusters), it is sometimes useful to consider a structural Random Intercept Model (RIM) among other competing models. However, just like the regular Linear Regression Model, any RIM that includes endogenous covariates will likewise suffer from having inconsistent OLS and ML estimators of its parameters. Dealing with endogeneity in the setting of a RIM, however, is more complex.

For illustration, we consider a two-level RIM given by

$$y_{ij} = \boldsymbol{x}_{ij}^{\intercal} \boldsymbol{\beta} + \boldsymbol{z}_j^{\intercal} \boldsymbol{\gamma} + \zeta_j + \epsilon_{ij}, \text{ for } i = 1, 2, \ldots, n_j, \text{ and } j = 1, 2, \ldots, J$$

where the outcome y_{ij} of individual i in cluster j depends on both unit-level covariates $\boldsymbol{x}_{ij}$ as well as cluster-level covariates $\boldsymbol{z}_j$ with coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively. In this model, the error term has been decomposed into two parts to account for variation at the unit-level (ϵ_{ij}) and at the cluster-level (ζ_j) separately. We typically assume that ϵ_{ij} is normally distributed with mean 0 and variance θ, and ζ_j is normally distributed also with mean zero and variance ψ. In the context of longitudinal or panel data, the cluster is often a person, whereas the unit is a measurement taken of that person at a given time. In both cross-sectional or longitudinal data, there are two ways to define endogeneity: when covariates at either the unit or cluster level are correlated with ϵ_{ij}, we call this "unit-level endogeneity"; and when they are correlated with ζ_j, we call this "cluster-level endogeneity". In this chapter, we assume that there is no "unit-level endogeneity", but only "cluster-level endogeneity".

Much like the single-level case, it can be shown that the OLS and ML estimators for both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are inconsistent when there is cluster-level endogeneity.

One traditional approach to obtain consistent estimators for $\boldsymbol{\beta}$ is by "de-meaning" each variable in the model. That is for each cluster j, we subtract from each variable its cluster specific mean as shown below.

$$\begin{aligned} y_{ij} - \overline{y}_{\cdot j} &= (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{\cdot j})^{\intercal} \boldsymbol{\beta} + (\boldsymbol{z}_j - \boldsymbol{z}_j)^{\intercal} \boldsymbol{\gamma} + (\zeta_j - \zeta_j) + (\epsilon_{ij} - \overline{\epsilon}_{\cdot j}) \\ \Longrightarrow y_{ij} - \overline{y}_{\cdot j} &= (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{\cdot j})^{\intercal} \boldsymbol{\beta} + (\epsilon_{ij} - \overline{\epsilon}_{\cdot j}). \end{aligned}$$

The estimator of the model above is known as the within-effects estimator. Notice that in this model, all cluster-level covariates $\boldsymbol{z}_j$ and the random intercept ζ_j are eliminated from the model. Since none of the covariates in this model are endogenous, the OLS and ML estimators for $\boldsymbol{\beta}$ are consistent. In fact, it can be shown that the within-effects estimator for $\boldsymbol{\beta}$ is identical to the so-called "fixed-effects" estimator where we include a dummy variable for each cluster instead of including a random intercept. Mundlak (1978) and Chamberlain (1982) also came up with alternative, consistent estimators for $\boldsymbol{\beta}$ that make use of an "auxiliary model" where the random intercept is projected onto either the cluster-mean or a full set of leads and lags (in the longitudinal setting) of each of the unit-level covariates. The downside to all these estimators is that they do not allow for the consistent estimation of the coefficients $\boldsymbol{\gamma}$ of cluster-level covariates z_j which may be of interest, even if z_j is exogenous. In this chapter, we focus specifically on the Hausman-Taylor estimator (Hausman & Taylor, 1981) that allows for the consistent estimation of both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. One nice feature of this estimator is that it is an IV estimator that makes use of IVs that are only internal to the model.

3.2.3 Hausman-Taylor Estimator

Hausman and Taylor (1981) described an estimation method to (i) make use of unit-level covariates to obtain consistent estimates for their own coefficients, and (ii) make use of exogenous covariates at both levels to serve as internal instruments for consistent estimation of the coefficients of endogenous cluster-level covariates. Their method assumes the existence of at least the same number of exogenous unit-level covariates (to be used as instruments) as endogenous cluster-level covariates, and although not explicitly stated in their paper, that there is sufficient correlation between these instruments and the endogenous cluster-level covariates. They showed, among other things, that this estimator, called the Hausman-Taylor (HT) estimator, is both consistent and asymptotically efficient.

At its core, the HT estimator makes use of the so-called Generalized Least Squares (GLS) or Fuller-Battese Transformation, to transform the model from

$$y_{ij} = \boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\beta} + \boldsymbol{z}_j^{\intercal}\boldsymbol{\gamma} + \zeta_j + \epsilon_{ij} = \boldsymbol{x}_{ij}^{\intercal}\boldsymbol{\beta} + \boldsymbol{z}_j^{\intercal}\boldsymbol{\gamma} + \xi_{ij} \tag{3.1}$$

to

$$\underbrace{y_{ij} - \kappa_j \overline{y}_{\cdot j}}_{\tilde{y}_{ij}} = \underbrace{(\boldsymbol{x}_{ij} - \kappa_j \overline{\boldsymbol{x}}_{\cdot j})^{\intercal}}_{\tilde{\boldsymbol{x}}_{ij}^{\intercal}}\boldsymbol{\beta} + \underbrace{(1-\kappa_j)\boldsymbol{z}_j^{\intercal}}_{\tilde{\boldsymbol{z}}_j^{\intercal}}\boldsymbol{\gamma} + \underbrace{(\xi_{ij} - \kappa_j \overline{\xi}_{\cdot j})}_{\tilde{\xi}_{ij}} \tag{3.2}$$

so that with an appropriate choice of κ_j, the transformed error terms, $\tilde{\xi}_{ij}$, can be shown to be uncorrelated. With uncorrelated errors, the model can now be estimated consistently using an IV approach (e.g., via two-stage least squares) with mean-centered level-1 covariates, the means of level-1 exogenous covariates, and the exogenous level-2 covariates as instruments.

Fuller and Battese (1973) showed that the choice of κ_j for this transformation depends on the variances of ζ_j and ϵ_{ij} and is given by

$$\kappa_j = 1 - \left(\frac{\theta}{\theta + n_j\psi}\right)^{\frac{1}{2}}. \tag{3.3}$$

Hence, the HT estimator can be obtained by finding consistent estimators for θ and ψ, which in turn depend on finding consistent (though not necessarily efficient) estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Hausman and Taylor thus suggested the following procedure:

Step 1: Obtain consistent estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\theta}$.
In order to obtain consistent estimators for $\boldsymbol{\beta}$ and θ, both of which at the unit-level, Hausman and Taylor suggested estimating the de-meanded model

$$y_{ij} - \overline{y}_{\cdot j} = (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{ij})^\intercal \boldsymbol{\beta} + (\epsilon_{ij} - \overline{\epsilon}_{\cdot j}).$$

In this transformed model, the cluster-level covariates as well as the random intercept disappears, and we are left with a model where the covariates are all uncorrelated with the error term by the unit-level exogeneity assumption. The OLS estimator for $\boldsymbol{\beta}$ as well as $\mathbb{V}(\epsilon_{ij} - \overline{\epsilon}_{\cdot j})$ (and by extension θ) in this model are thus consistent.

Step 2: Obtain consistent estimator $\widehat{\boldsymbol{\gamma}}$.
Next, a consistent estimator for $\boldsymbol{\gamma}$ can be obtained by considering the between-cluster model, obtained by taking the cluster means of the original model:

$$\overline{y}_{\cdot j} = \overline{\boldsymbol{x}}_{\cdot j}^\intercal \boldsymbol{\beta} + \boldsymbol{z}_j^\intercal \boldsymbol{\gamma} + \zeta_j + \overline{\epsilon}_{\cdot j} \tag{3.4}$$

Since we have a consistent estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, with some rearranging of (3.4), we can use the resultant estimates of $\overline{\boldsymbol{x}}_{\cdot j}^\intercal \widehat{\boldsymbol{\beta}}$ to transform $\overline{y}_{\cdot j}$ as follows:

$$\overline{y}_{\cdot j} - \overline{\boldsymbol{x}}_{\cdot j}^\intercal \widehat{\boldsymbol{\beta}} \approx \boldsymbol{z}_j^\intercal \boldsymbol{\gamma} + \zeta_j + \overline{\epsilon}_{\cdot j}.$$

However, since there may still be endogenous cluster-level covariates, instead of performing OLS or ML estimation, we make use of an IV approach (via two-stage least-squares) using the cluster means of the unit-level exogenous covariates as well as the cluster-level exogenous covariates as instruments. This results in a consistent estimator for $\boldsymbol{\gamma}$.

Step 3: Obtain consistent estimator $\widehat{\psi}$.
With consistent (but not necessarily efficient) estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we now return to (3.1), and compute the residuals $\widehat{\xi}_{ij} = y_{ij} - \boldsymbol{x}_{ij}^\intercal \widehat{\boldsymbol{\beta}} + \boldsymbol{z}_j^\intercal \widehat{\boldsymbol{\gamma}}$. Using these residuals, Hausman and Taylor, 1981 showed that as $J \to \infty$,

$$\frac{1}{J}\sum_{j=1}^{J} \widehat{\xi}_{ij}^2 \xrightarrow{p} \frac{J}{\sum_{j=1}^{J}\frac{1}{n_j}}\psi + \theta.$$

Hence, a consistent estimator for ψ is given by

$$\frac{\sum_{j=1}^{J} \frac{1}{n_j}}{J} \left[\frac{1}{J} \sum_{j=1}^{J} \widehat{\xi}_{ij}^{2} - \theta \right]. \tag{3.5}$$

While (3.5) is a common consistent estimator for ψ implemented in existing software packages, Castellano et al. (2014) also showed that another consistent estimator can be achieved by replacing the reciprocal of the harmonic mean of cluster sizes, $\frac{\sum_{j=1}^{J} \frac{1}{n_j}}{J}$, in (3.5) with the reciprocal of the arithmetic mean, $\frac{J}{\sum_{j=1}^{J} n_j}$. In our own implementation of the Hausman-Taylor estimator, we consider the MLE of ψ.

Step 4: Obtain efficient estimators.
Finally, with consistent estimates for ψ and θ, we can now obtain a consistent estimate of κ_j in (3.3). We then perform the Fuller-Battese Transformation in (3.2) and estimate all parameters consistently using two-stage least-squares as described above. This yields consistent **and efficient** estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

3.2.4 Modeling endogeneity using SEM

In this chapter, we propose estimating a RIM with cluster-level endogeneity using Structural Equation Modeling (SEM) which enables us to treat all variables in the models as responses, and in so doing, provides us with the flexibility to model the endogeneity naturally. For example, within a SEM framework, a standard two-level RIM with 20 units per cluster, two unit-level covariates x_{ij}^a and x_{ij}^b, as well as two cluster-level covariates z_j^a, z_j^b can be represented by the graph in Figure 3.1 as per Teachman et al. (2001). Using the reduced form in a typical RIM setting, this graph represents the model

$$y_{ij} = \alpha + \beta_1 x_{ij}^a + \beta_2 x_{ij}^b + \gamma_1 z_j^a + \gamma_2 z_j^b + \zeta_j + \epsilon_{ij}. \tag{3.6}$$

In this path diagram, the loadings from η_j to $y_{1j}, y_{2j}, \ldots, y_{20j}$ are set to one because ζ_j is a random intercept. The regression coefficients for y_{ij} on x_{ij}^a and x_{ij}^b for $i = 1, 2, \ldots 20$ are denoted β_1 and β_2 respectively. η_j is also regressed on z_j^a and z_j^b with coefficients denoted by γ_1 and γ_2, and a disturbance term ζ_j.

As such, the measurement and structural components of the SEM depicted by Figure 3.1 are given by

$$\begin{aligned} y_{ij} &= \alpha + \beta_1 x_{ij}^a + \beta_2 x_{ij}^b + \eta_j + \epsilon_{ij} \text{ for all } i = 1, 2, \ldots, 20 \\ \eta_j &= \gamma_1 z_j^a + \gamma_2 z_j^b + \zeta_j. \end{aligned}$$

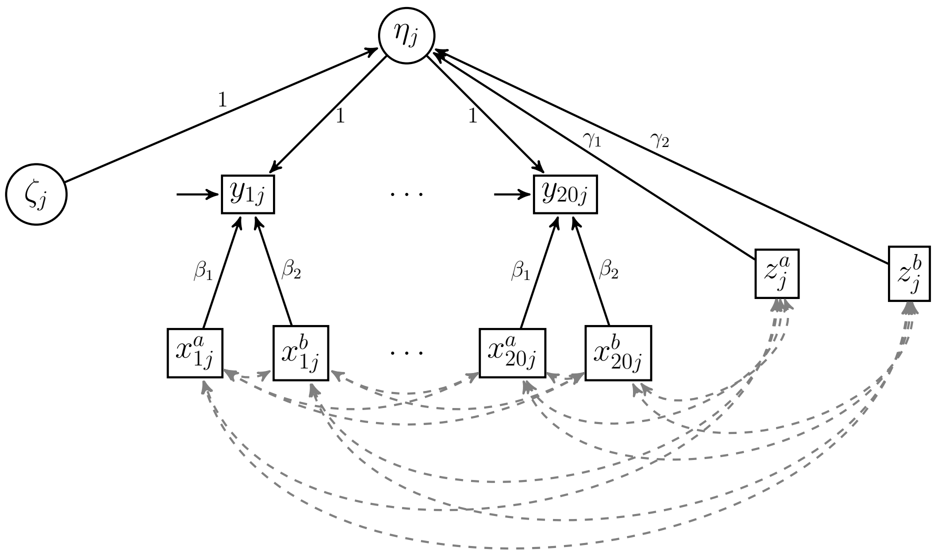


Figure 3.1: Structural Model for Two-Level RIM with no Endogeneity

These regression coefficients coincide precisely with the parameters β_1, β_2, γ_1, and γ_2 in Equation 3.6. In fact, by substituting the equation for η_j into the equations for y_{ij}, we recover Equation 3.6 exactly.

Note that in Figure 3.1, the covariates x_{ij}^a, x_{ij}^b, z_j^a, and z_j^b are allowed to covary with each other, but not with ζ_j as per the cluster-level exogeneity assumption.

At this junction, it is important to highlight a subtlety in interpreting a RIM expressed as a SEM. In a traditional RIM with students clustered in schools for example, the unit of analysis is the student. That is, for a school with n_j students, each set of random variables y_{ij}, x_{ij}^a, x_{ij}^b, z_j^a, z_j^b, and ζ_j represents characteristics of student i in school j. Another way to think of this is that data for a traditional RIM are usually represented in 'long' form, where each row represents characteristics of a student in a school (in this case, there are n_j rows of data for this school). However, when viewing the RIM as a SEM, the unit of analysis is now the school rather than the student. For example, in a school with n_j students, each k-tuple, consisting of the outcome of interest and the unit-level covariates (in this case a triple $(y_{ij}, x_{ij}^a, x_{ij}^b)$) is now treated as a set of variables for the school together with the cluster-level covariates and the random intercept. Another way to think of this is that the data for RIM expressed as a SEM is usually in 'wide' form where each row represents a school. Mehta and Neale (2005) highlight this distinction best in their paper with a tongue-in-cheek title "People are variables too".

Since panel data may not be exchangeable, previous approaches to model RIMs using SEM implicitly treat units as non-exchangeable (Allison, 2005, 2009; Allison et al., 2017; Bollen & Brand, 2010). This assumption, however, does not make sense for cross-sectional data, and this fact has not been pointed out or incorporated in previous estimators using SEM. We set up our simulation and estimator specifically for this case. In the later sections, we compare the results for different estimators on data with an exchangeable and an unstructured covariance structure. In the meantime, we thus impose the following constraints on the covariances: for all $i = 1, 2, \ldots, 20$, $i' = 1, 2, \ldots, 20$ but $i \neq i'$

- among different unit-level covariates, within the same unit (within the same cluster): $\text{cov}(x_{ij}^a, x_{ij}^b) = c_{\text{within}}$;
- among different unit-level covariates, across different units (within the same cluster: $\text{cov}(x_{ij}^a, x_{i'j}^b) = c_{x_a x_b}$;
- among the same unit-level covariates, across different units (within the same cluster:
  - $\text{cov}(x_{ij}^a, x_{i'j}^a) = c_{x_a}$;
  - $\text{cov}(x_{ij}^b, x_{i'j}^b) = c_{x_b}$;
- among cluster-level covariates and unit-level-covariates:
  - $\text{cov}(x_{ij}^a, z_j^a) = c_{x_a z_a}$;
  - $\text{cov}(x_{ij}^a, z_j^b) = c_{x_a z_b}$;
  - $\text{cov}(x_{ij}^b, z_j^a) = c_{x_b z_a}$;
  - $\text{cov}(x_{ij}^b, z_j^b) = c_{x_b z_b}$; and
- among cluster-level covariates: $\text{cov}(z_j^a, z_j^b) = c_{z_a z_b}$.

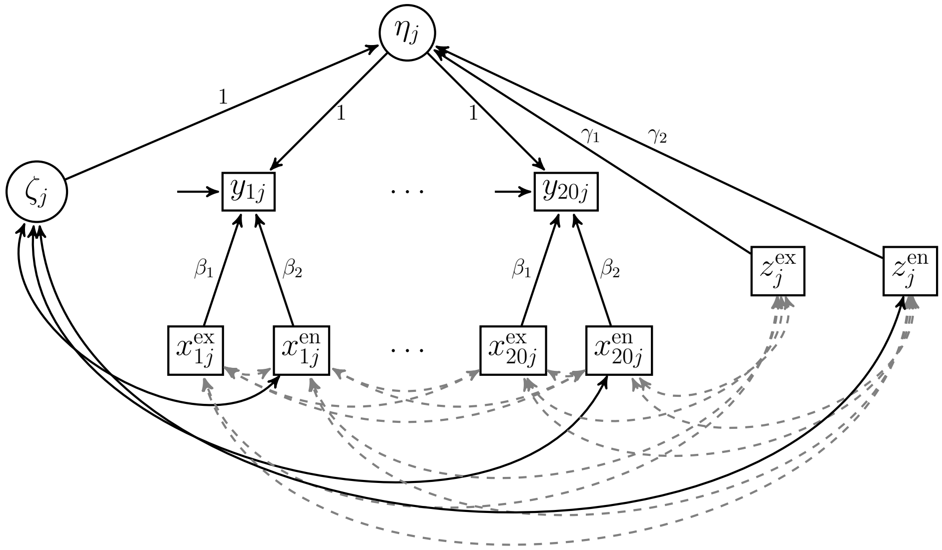
We make analogous constraints on the variances $\mathbb{V}(x_{ij}^a) = \sigma_a^2$ and $\mathbb{V}(x_{ij}^b) = \sigma_b^2$ for all $i = 1, 2, \ldots 20$. The remaining variance-covariance parameters are the variances of the covariates at the cluster level (viz. $\mathbb{V}(z_j^a)$ and $\mathbb{V}(z_j^b)$). As such, the SEM model we have defined has 13 parameters from the variance-covariance structure for the explanatory variables on top of the 5 parameters from the mean-structure, as well as 2 coming from ψ and θ for a total of 20 parameters to estimate.

Since the model we have defined in Figure 3.1 is equivalent to the RIM in Equation 3.6, we can extend it to deal with cluster-level endogenous covariates by simply allowing the endogenous covariates to covary with ζ_j in our model. For instance, if we consider x_{ij}^b and z_j^b to be endogenous, then the pictorial representation of the modified SEM will be given by Figure 3.2. For ease of interpretation, we now label the exogenous covariates $x_{ij}^a \equiv x_{ij}^{\text{ex}}$ and $z_j^a \equiv z_j^{\text{ex}}$, and the endogenous covariates $x_{ij}^b \equiv x_{ij}^{\text{en}}$ and $z_j^b \equiv z_j^{\text{en}}$.

We continue to impose the exchangeability constraints described previously for this new SEM, but by allowing x_{ij}^{en} and z_j^{en} to covary with ζ_j, we now have to estimate the covariances between these variables. Analogous to the previous model, we also impose exchangeability constraints that for all $i = 1, 2, \ldots 20$, $\text{cov}(x_{ij}^{\text{en}}, \zeta_j) = c_{x_{ij}^{\text{en}}\zeta}$. Using a similar notation, we denote $\text{cov}(z_j^{\text{en}}, \zeta_j) = c_{z_j^{\text{en}}\zeta}$. This adds an additional 2 parameters of the variance-covariance structure to be estimated for a total of 22 parameters.

It should be pointed out here that we adopt the ML SEM estimator throughout the rest of this chapter. When the data is unbalanced, the SEM described above is defined based on the largest cluster size, and as a result will contain clusters with missing k-tuples for some units (where $k - 1$ is the number of unit-level covariates). In such cases, we make use of the Full-Information Maximum Likelihood SEM estimator (Muthén et al., 1987) to deal with the missingness. Additionally, when $\max_j n_j > J$, the implied covariance matrix has dimensions greater than the total number of clusters J but is identifiable because of the exchangeability constraints we placed on the covariances. However, many existing SEM packages evaluate the identifiability of the SEM without accounting for these constraints, which may make the model practically inestimable.

Figure 3.2: Structural Model for Two-Level RIM with Endogeneity



3.3 Simulation

In this simulation, we are interested in validating the consistency of both the SEM and HT estimators. Additionally, we also consider the finite sample performance of both these estimators in terms of their bias and variability. Throughout this simulation, we implement the HT estimator using (i) the `pht` function (henceforth called the HT1 estimator) in the `plm` package (Croissant & Millo, 2008), as well as (ii) a direct implementation using the `ivreg` function (henceforth called the HT2 estimator) in the `AER` package (Kleiber & Zeileis, 2008) in `R`. We point out here that the estimate for ψ in the `plm` package was obtained using the method described earlier using the harmonic mean as in Castellano et al. (2014), whereas when we implement the HT estimator directly using the `AER` package, we make use of the ML estimate for ψ. We also implement the ML SEM estimator in `R` using the `lavaan` package (Rosseel, 2012).

We assess properties of the SEM estimator versus the HT estimator via simulation by considering a two-level RIM as in Subsection 3.2.4 and Figure 3.2. That is, our data-generating process is given by

$$y_{ij} = \alpha + \beta_1 x_{ij}^{\text{ex}} + \beta_2 x_{ij}^{\text{en}} + \gamma_1 z_j^{\text{ex}} + \gamma_2 z_j^{\text{en}} + \zeta_j + \epsilon_{ij}. \tag{3.7}$$

We generate data for our simulation based on the model given in (3.7) with the exchangeability constraints for the variance-covariance matrix discussed in Section 3.2.4. In generating the data, we considered a range of values for J, as well as both the balanced case with an equal number of units per cluster (where $n_j = \frac{N}{J}$), and the unbalanced case where the size of cluster j was drawn from a scaled-uniformed distribution.

As we proceed, we also count the number of mean-structure and variance-covariance structure parameters we fixed to ensure that we did not inadvertently constrain the data with fewer than the 22 parameters described in Subsection 3.2.4. For this chapter, we selected true values $\alpha = 1.4$, $\beta_1 = 2.3$, $\beta_2 = -1.1$, $\gamma_1 = -0.7$, and $\gamma_2 = 1.8$ for the main parameters of the model (for a total of 5 specified mean-structure parameters), as well as $\psi = 1.2$ and $\theta = 0.8$ for the variance components.

3.3.1 Data-Generating Process

We begin this section with a brief overview of the approach we will take to simulate the data. Note that in Steps 1 to 3, we create the variables ζ_j, z_j^{ex}, z_j^{en}, x_{ij}^{ex}, and x_{ij}^{en}.

As a brief overview, in Step 1, we generate independent values of ζ_j, two cluster-level variables, and two unit-level variables, and collect them in the $N \times 5$ matrix X. At this point, ζ_j and each covariate is independent from each other. In Step 2, we impose a covariance structure onto X by post-multiplying X with the Cholesky decomposition of a correlation matrix with the specified structure. In Step 3, we induce additional intraclass correlations (ICCs) for unit level covariates by adding cluster-level variables generated from a Multivariate Normal distribution. Finally, in Step 4, we generate the observed outcome values Y_{ij}.

3.3.1.1 Step 1

For J clusters, we generated two independent variables with J values, $s_j^{(1)}$ and $s_j^{(2)}$ for $j = 1, \ldots, J$, each drawn from independent standard normal distributions, as well as another variable with J values, ζ_j drawn from an independent normal distribution with mean 0 and variance $\psi = 1.2$ (hence, fixing 3 variance-covariance parameters). For $j = 1, \ldots, J$, we then expanded $s_j^{(1)}$, $s_j^{(2)}$ and ζ_j by replicating each variable depending on the cluster size n_j for a total of $N = \sum_{j=1}^{J} n_j$ observations.

We then generated two variables with N values, $t_{ij}^{(1)}$ and $t_{ij}^{(2)}$ for $j = 1, \ldots, J$, $i = 1, \ldots, n_j$, each drawn from independent standard normal distributions, as well as another variable with N values, ϵ_{ij} for $j = 1, \ldots, J$, $i = 1, \ldots, n_j$, drawn from a normal distribution with mean 0 and variance θ (for an additional 3 variance-covariance parameters).

We collected ζ_j, $s_j^{(1)}$, $s_j^{(2)}$, $t_{ij}^{(1)}$ and $t_{ij}^{(2)}$ in an $N \times 5$ matrix X. The columns of X are thus independent of each other in the population.

3.3.1.2 Step 2

Next, we needed to generate an initial population covariance matrix $\Sigma^{(0)}$ for ζ_j, z_j^{ex}, z_j^{en}, x_{ij}^{ex}, and x_{ij}^{en}. Since it is non-trivial to directly generate a valid correlation/covariance matrix that is positive semi-definite, we instead approached this by specifying a lower triangular matrix L (whose elements we generated randomly), and considered the covariance matrix associated with L given by $\Sigma^{(0)} = LL^T$. We then obtained the correlation matrix R by standardizing $\Sigma^{(0)}$ with its diagonal elements. That is, if $D = \sqrt{\text{diag}(\Sigma^{(0)})}$ where $\sqrt{\text{diag}(\Sigma^{(0)})}$ is a 5×5 diagonal matrix with diagonal entries equal to the square-root of the diagonal entries of Σ, then $\rho = D^{-1} \Sigma D^{-1}$.

For the purposes of our simulation, we only considered matrices $\Sigma^{(0)}$ that were strictly positive-definite so that in the population, ζ_j and the other covariates were not perfectly collinear. We checked this in our simulation by using the `matrixcalc` package (Novomestky, 2012) in `R`. This meant that the matrix L was precisely the unique lower triangular matrix associated with the Cholesky Decomposition of Σ. This one-one correspondence between L and $\Sigma^{(0)}$ ensures that our approach to generating Σ covers the space of all possible positive-definite covariance matrices that satisfies the exogeneity constraints below.

Using this approach, each element in L was drawn from a normal distribution with mean 0 and standard deviation 0.2. In L, we then set the two elements corresponding to the relationship between ζ_j and the exogenous covariates x_{ij}^{ex} and z_j^{ex} to zero and generated R as described above. The resulting population correlation matrix, R and the associated lower triangular matrix $L_R = D^{-1}L$ are presented below.

$$R = \begin{pmatrix} 1.000 & & & & \\ 0.000 & 1.000 & & & \\ -0.570 & -0.456 & 1.000 & & \\ 0.000 & -0.208 & 0.380 & 1.000 & \\ 0.714 & -0.185 & -0.232 & 0.562 & 1.000 \end{pmatrix}$$

$$L_R = \begin{pmatrix} 1.000 & & & & \\ 0.000 & 1.000 & & & \\ -0.570 & -0.456 & 0.683 & & \\ 0.000 & -0.208 & 0.417 & 0.885 & \\ 0.714 & -0.185 & -0.133 & 0.529 & 0.398 \end{pmatrix}$$

In order to impose the correlations summarized by R on X, we considered the Cholesky decomposition of $R = L_R L_R^T$, and post-multiplied X with L_R^T. That is, we considered the new dataset $X^* = [X_1^*, X_2^*, X_3^*, X_4^*, X_5^*] = X L_R^T$ which has a sample correlation matrix as if it came from a population with a population correlation matrix equal to R. We note that the variances of the elements of X^* are the same as the variances of X and equal to $(\psi, 1, 1, 1, 1)$ since L_R is the Cholesky deomposition of a correlation matrix R. Further, through this transformation, X_1^*, X_2^*, and X_3^* continue to vary only at the cluster-level, while X_4^* and X_5^* vary at the unit-level, but with ICCs induced by the transformation. Note that the order of the columns of X are important. As a result of this transformation, we have fixed another 8 parameters for the non-zero correlations induced in the data by L_R for a total of 14 variance-covariance structure parameters and 5 mean structure parameters, and are hence short of 3 parameters.

3.3.1.3 Step 3

At this junction, we note that while X^* has population correlations equal to R, the ICCs for x_{ij}^{ex} and x_{ij}^{en} are induced through the contributions from ζ_j, $s_j^{(1)}$, and $s_j^{(2)}$ via corresponding non-zero weights in L_R. As such, to ensure that a portion of the ICCs of x_{ij}^{ex} and x_{ij}^{en} are induced independently of the other variables, we modify X^* in the following way.

We first generate another two variables with J values, $r_j^{(1)}$ and $r_j^{(2)}$ for $j = 1, \ldots, J$, drawn from a bivariate normal distribution with mean 0, variances 1 and covariance 0.2, and then replicate each variable depending on the cluster sizes n_j for a total of $N = \sum_{j=1}^{J} n_j$ observations. We then add $r_j^{(1)}$ and $r_j^{(2)}$ to X_4^* and X_5^* respectively to induce an additional intraclass correlation that does not arise due only to ζ_j, $s_j^{(1)}$, and $s_j^{(2)}$. This fixes an additional

3 parameters for a total of 22 parameters as required.

$$\begin{aligned}\zeta_j &= X_1^*;\\ z_j^{\text{ex}} &= X_2^*;\\ z_j^{\text{en}} &= X_3^*;\\ x_{ij}^{\text{ex}} &= X_4^* + r_j^{(1)}; \text{ and}\\ x_{ij}^{\text{en}} &= X_5^* + r_j^{(2)}.\end{aligned}$$

3.3.1.4 Step 4

Finally, using the parameters generated earlier and Equation 3.7, we generated values for the observed outcome y_{ij}.

3.3.2 The Population Covariance Matrix for Observed Variables

As described in our data-generating process, we obtained X^* by post multiplying X with L_ρ^T. This means that the variables in X^* are linear combinations of the variables in X with coefficients given by L_ρ^T. As such, the covariance matrix of X^*, Σ_{X^*} can be obtained from the covariance matrix of X, $\Sigma_X = \text{diag}(\psi, 1, 1, 1, 1)$ by the quadratic form $\Sigma_{X^*} = L_\rho \Sigma_X L_\rho^T$. Since the addition of $r_j^{(1)}$, and $r_j^{(2)}$ only changes the variances and covariances of the affected variables, we can obtain the population covariance matrix Σ_P in terms of ζ_j, z_j^{ex}, z_j^{en}, x_{ij}^{ex}, and x_{ij}^{en} by adding to Σ_{X^*}, a matrix 5×5 matrix C whose elements are all zero except for

$$C[4,5;4,5] = \text{cov}(r_j^{(1)}, r_j^{(2)}) = \begin{pmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{pmatrix}$$

In our study, we thus have

$$\begin{aligned}\Sigma_P &= L_\rho \text{diag}(\psi, 1, 1, 1, 1) L_\rho^T + C \\ &= \begin{pmatrix} 1.200 & & & & \\ 0.000 & 1.000 & & & \\ -0.684 & -0.456 & 1.065 & & \\ 0.000 & -0.208 & 0.380 & 2.000 & \\ 0.857 & -0.185 & -0.313 & 0.762 & 2.102 \end{pmatrix}.\end{aligned}$$

In this matrix, we make note particularly of the fact that $\mathbb{V}(x_{ij}^a) = 2.000$, $\mathbb{V}(x_{ij}^b) = 2.102$ and $c_{\text{within}} = 0.762$.

In order to compute the population covariances between unit-level covariates within the same cluster but between *different units*, say $\text{cov}(x_{ij}^a, x_{i'j}^a) = c_{x_a}$, we first realize that the contribution to the covariances of the unit-level factors $t_{ij}^{(1)}$ and $t_{ij}^{(2)}$ is 0.2. As such, we consider the sub-matrix Σ_B obtained by taking the 4th and 5th rows and columns from the analogue of the matrix Σ_P above by using diag$(\psi, 1, 1, 0, 0)$ instead of Σ_X. We thus obtain

$$\begin{aligned}\Sigma_B &= (L_\rho \mathrm{diag}(\psi, 1, 1, 0, 0) L_\rho^T + C)[4, 5; 4, 5] \\ &= \begin{pmatrix} 1.217 & \\ 0.294 & 1.664 \end{pmatrix}.\end{aligned}$$

In this matrix, we make note particularly of the fact that $c_{x_a} = 1.217$, $c_{x_b} = 1.664$ and $c_{x_a x_b} = 0.294$. The ICCs for x_a and x_b are $\frac{1.217}{2} = 0.609$ and $1.664/2.102 = 0.792$, respectively.

Since ϵ_{ij} is uncorrelated with all other variables, we can augment Σ_P and Σ_B (which we denote as Σ_P^* and Σ_B^* respectively) to include ϵ_{ij} by including an additional row and column consisting of all zeros for the covariance terms, and θ for the variance term.

Now, all that is left to obtain a covariance matrix of only the observed variables y_{ij}, z_j^{ex}, z_j^{en}, x_{ij}^{ex}, and x_{ij}^{en}. Since y_{ij} is a linear combination of the variables whose covariance structure is captured by Σ_P^* and Σ_B^*, we can consider $\Sigma_P^{**} = A\Sigma_P^* A^T$, and $\Sigma_B^{**} = A\Sigma_B^* A^T$ where

$$A = \begin{pmatrix} 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 1.0 & \gamma_1 & \gamma_2 & \beta_1 & \beta_2 & 1.0 \end{pmatrix}.$$

Hence, the population covariance matrix in terms of observed variables is given by

$$\Sigma_P^{**} = \begin{pmatrix} 1.000 & & & & \\ -0.456 & 1.065 & & & \\ -0.208 & 0.380 & 2.000 & & \\ -0.185 & -0.313 & 0.762 & 2.102 & \\ -1.796 & 2.771 & 4.591 & -0.137 & 16.782 \end{pmatrix}$$

and

$$\Sigma_B^{**} = \Sigma_B = \begin{pmatrix} 1.217 & \\ 0.294 & 1.664 \end{pmatrix}.$$

3.3.3 Results

3.3.3.1 Evaluating Consistency

An estimator of a parameter is said to be (weakly) consistent if it converges in probability to the true value of the parameter as the sample size (which, in this case, is the number of clusters) approaches infinity. In order to assess the consistency of both the HT and the ML SEM estimators, we considered a single unbalanced dataset generated by the method described in the previous section for each condition of 100, 1000, 10,000, 100,000 and 1,000,000 clusters with 5, 10 or 15 units per cluster, each occurring with equal probability. We summarize the results in Table 3.1 and Fig 3.3 below.

Table 3.1: Summary of Parameter Estimates using Different Estimators and Packages for Unbalanced Data Assuming Exchangeability

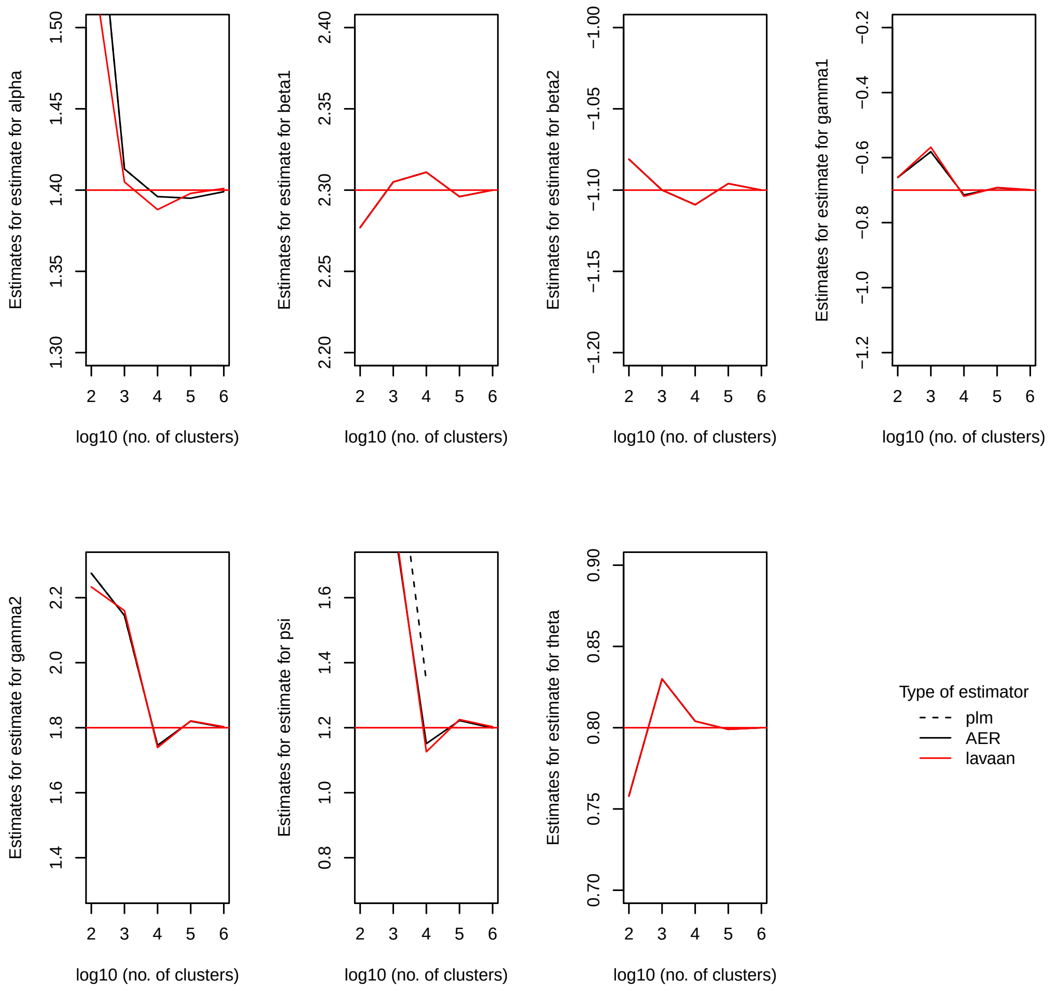
parameter	**no. of**	**HT1**	**HT2**	**ML SEM**
(true value)	**clusters**	`plm`	`AER`	`lavaan`
α (1.4)	100	1.626	1.626	1.544
	$1,000,000$	—	1.399	1.401
β_1 (2.3)	100	2.277	2.277	2.277
	$1,000,000$	—	2.300	2.300
β_2 (-1.1)	100	-1.081	-1.081	-1.081
	$1,000,000$	—	-1.100	-1.100
γ_1 (-0.7)	100	-0.661	-0.661	-0.661
	$1,000,000$	—	-0.700	-0.699
γ_2 (1.8)	100	2.275	2.275	2.233
	$1,000,000$	—	1.801	1.803
ψ (1.2)	100	2.537	2.199	1.853
	$1,000,000$	—	1.199	1.203
θ (0.8)	100	0.758	0.758	0.758
	$1,000,000$	—	0.800	0.800

Based on Table 3.1 and Figure 3.3, we note that as the number of clusters increased, both the SEM and the HT estimates for all the parameters approached the true value. This provides us with some evidence that, like the HT estimator, the SEM estimator is also consistent. We note that the `plm` package in `R` did not converge when the number of clusters exceeded 10,000. We also note here that the parameter estimates are not identical in finite samples.

3.3.3.2 Population Study for Consistency

We also assessed consistency of the SEM estimator (with and without the exchangeability constraints) by making use of the population level covariance structures as per Muthén et al. (1987). To do so, we transform the true population covariance matrix Σ_P^{**} from long from (where each row represents a unit-cluster combination) to wide form (where each row represents a cluster), making sure to use the covariances from Σ_B^{**} for covariances between unit-level covariates from different clusters. Since the largest cluster in our data had 15 units, we then transformed Σ_P^{**} and Σ_B^{**} to account for 15 repetitions of y_{ij}, x_{ij}^{ex}, and x_{ij}^{en}.

Figure 3.3: HT and SEM Estimates from Table 1 for Model Parameters



This block-matrix in wide-form, Σ_L, is then used as input to `lavaan` together with a vector of means consisting of all zeros except the last 15 entries being $\alpha = 1.4$. In estimating a SEM using only a covariance matrix and a vector of means, `lavaan` also requires a specification of the number of observations in the dataset. Since we input a population covariance matrix and a population vector of means, instead of the sample analogues, we arbitrarily declare our sample size to be 1000. This choice did not affect our point estimates.

The results of our SEM estimators are presented in Table 3.2 below. Both the models with and without the exchangeability constraints produced the same parameter estimates. The SEs for the parameters estimates are also the same for both models for $N = 1000$. However, the SEs are greater for the variance-covariance parameters that were unrestricted in the non-exchangeable SEM. The SEs from the exchangeable SEM estimator versus the unrestricted SEM estimator are as follows. For $c_{x_a z_a}$ it is 0.036 vs 0.045, for $c_{x_b z_a}$ it is 0.042 vs 0.046, for $c_{x_a z_b}$ it is 0.039 vs 0.044, for $c_{x_b z_b}$ it is 0.039 vs 0.047, for c_{x_a} it is 0.057 vs 0.065, for c_{x_b} it is 0.076 vs 0.078, $c_{x_a x_b}$ it is 0.047 vs 0.056.

Table 3.2: Summary of SEM Estimates using the Population Covariance Matrix

parameter	**exchangeable SEM**	**unrestricted SEM**
α (1.4)	1.400 (0.022)	1.400 (0.022)
β_1 (2.3)	2.300 (0.014)	2.300 (0.014)
β_2 (-1.1)	-1.100 (0.019)	-1.100 (0.019)
γ_1 (-0.7)	-0.700 (0.075)	-0.700 (0.075)
γ_2 (1.8)	1.800 (0.145)	1.800 (0.145)
ψ (1.2)	1.199 (0.203)	1.199 (0.203)
θ (0.8)	0.799 (0.010)	0.799 (0.010)

3.3.3.3 Finite Sample Properties

While both the Hausman-Taylor Estimator and the SEM estimators are consistent, in this chapter, we also explore their finite sample properties. In a simulation using the same data-generating mechanism described previously, we simulated 100 sets of 100 clusters with 5, 10, 15 and 20 units. The estimated bias and empirical standard errors along with their respective Monte Carlo errors are presented in Table 3.3 below. We see that the ML SEM estimator performs no worse compared to the standard Hausman-Taylor estimators. In fact, due to a small number of clusters, we notice that the estimated bias for ψ is higher when using the `plm` function compared to the other two approaches. This could due to the fact that it uses the harmonic mean of the residuals in Step 3 of the approach by Hausman and Taylor.

3.4 Application to Estimating Causal Effect of Catholic Schools

One of the benefits of the Hausman-Taylor approach, and by extension, our SEM approach, is that it allows the causal effect of a cluster-level treatment that is cluster-endogenous to be estimated consistently provided we have at least one unit-level exogenous covariate. That is, we can estimate the causal effect even if we have have not measured and accounted for all the necessary confounders in our model. This is in contrast to another approach called “multilevel matching”. In this approach, we need to assume that all the necessary confounders at both levels are measured, and then try to balance these measured covariates by matching. A consistent estimate of the causal effect can then be obtained by estimating

Table 3.3: Evaluating bias and variance of different estimators (100 replications of 100 clusters of size 5, 10, 15 or 20) when the data satisfy exchangeability

parameter		**HT1**	**HT2**	**ML exch SEM**
(true value)		`plm`	`AER`	`lavaan`
α (1.4)	Estimated Bias	0.008	0.008	0.003
	MCE Bias	0.013	0.013	0.013
	Empirical SD	0.131	0.131	0.128
	MCE SD	0.009	0.009	0.009
β_1 (2.3)	Estimated Bias	-0.008	-0.008	-0.008
	MCE Bias	0.005	0.005	0.005
	Empirical SD	0.052	0.052	0.052
	MCE SD	0.004	0.004	0.004
β_2 (-1.1)	Estimated Bias	0.003	0.003	0.003
	MCE Bias	0.007	0.007	0.007
	Empirical SD	0.074	0.074	0.074
	MCE SD	0.005	0.005	0.005
γ_1 (-0.7)	Estimated Bias	-0.020	-0.020	-0.014
	MCE Bias	0.032	0.032	0.032
	Empirical SD	0.314	0.314	0.318
	MCE SD	0.022	0.022	0.023
γ_2 (1.8)	Estimated Bias	-0.017	-0.018	-0.008
	MCE Bias	0.057	0.057	0.057
	Empirical SD	0.569	0.569	0.569
	MCE SD	0.041	0.041	0.041
ψ (1.2)	Estimated Bias	0.420	0.132	0.133
	MCE Bias	0.051	0.042	0.045
	Empirical SD	0.504	0.418	0.449
	MCE SD	0.036	0.030	0.032
θ (0.8)	Estimated Bias	-0.002	-0.002	-0.002
	MCE Bias	0.004	0.004	0.004
	Empirical SD	0.036	0.036	0.036
	MCE SD	0.003	0.003	0.003

a three-level RIM of the outcome on the treatment with a random intercept for clusters and another for matched pairs.

We emphasize here, however, that there is no free lunch in the sense that our approach requires that the effect of the treatment and the covariates on the outcome is indeed linear, and that the variables we declare to be exogenous are precisely so. Since these two assumptions as well as the no unmeasured confounding assumption are fundamentally untestable, the choice of which method to utilize will depend on contextual knowledge.

As a proof of concept, we make use of the publicly-available "High School and Beyond" dataset that can be found in the `multiMatch` package (Pimentel, Page, & Keele, 2016). Following Pimentel et al. (2016), we are interested in determining the effect of a student's enrolment in a Catholic school versus a non-Catholic school on his/her math achievement. In our demonstration, we use the following covariates: (i) an indicator of each student's gender (which we assume is exogenous), (ii) a measure of each student's Socio-Economic Status (SES) (which we allow to be endogenous), (iii) an indicator of whether a student's school is Catholic (which we allow to be endogenous), and (iv) a measure of the disciplinary culture in the school (which we assume to be exogenous). We argue that there are important school-level covariates that we have omitted such as the size of the school - which may affect a student's math achievement through various mechanisms such as the distribution of resources to students or peer-effects. The size of the school may also be correlated with (i) the SES of a student since higher SES students may opt to enroll in smaller schools where teachers can provide more dedicated attention to their students, and (ii) whether the school is Catholic since Catholic schools tend to be smaller. Additionally, instead of using the full dataset, we randomly sampled only 14 students for each cluster. This was done to avoid technical difficulties with the current implementation of `lavaan` which does not take parameter constraints into consideration when evaluating the identifiability of the model. We point out again that this is purely a software issue which can be overcome.

We present the estimates of the "Catholic School Effect" using the different approaches with the 95% confidence intervals appended in parentheses next to the point estimates. As a baseline, we regressed the outcome on the treatment and the three other covariates and included a random intercept for the school. In this model, the estimated mean math achievement score for students in Catholic schools was 0.39 $(-0.38, 1.15)$ points higher than those from non-Catholic schools, after adjusting for the other covariates. The multilevel matching approach yielded an estimated difference in means of about 1.21 $(0.05, 2.38)$. The built-in HT estimator and our SEM estimator produced similar estimates of about 9.38 $(-2.48, 21.20)$ and 9.38 $(-2.47, 21.24)$ respectively.

While none of the estimates were significant, the point estimates themselves as well as their precision heavily depended on the method of choice and by extension the assumptions made. It is unsurprising that the estimates were so different, and clearly illustrates the point that in empirical studies like this, contextual knowledge must be used to guide the researcher in choosing his/her estimation strategy.

3.5 Concluding Remarks

Based on our simulation study with exchangeable data, we see that the exchangeable-SEM estimator produces similar estimates to the Hausman-Taylor estimator with balanced data. With unbalanced data, it performs no worse than the Hausman-Taylor in terms of consistency, and finite sample properties. This is important because under the SEM framework, estimation of structural RIMs with endogeneity can be extended to situations that are more naturally handled with SEM. For example, our approach can (i) incorporate measurement error in the covariates, (ii) be extended to deal with endogeneity involving random slopes, and (iii) be extended to deal with binary outcomes (although consistency of the resultant estimator is not guaranteed, and more work has to be done to study the performance of this estimator).

There are, however, two main drawbacks to the exchangeable-SEM estimator. Firstly, the exchangeable-SEM estimator is computationally more expensive compared to existing packages that implement the Hausman-Taylor estimator. This is especially so in the unbalanced case where we make use of Full Information Maximum Likelihood to fit the SEM model. Secondly, even though the number of parameters to be estimated in the SEM model is reduced by constraining some parameters to be the same, current software packages do not take this into account when determining the identifiability of the model. As such, it is substantially easier to fit models with smaller cluster sizes. However, these technical problems can be overcome by software that is optimized to estimate such models.

Despite the drawbacks, we believe that, on balance, there is still value in the exchangeable-SEM estimator. Chief among the next steps would be to explore the finite sample performance of the exchangeable-SEM estimator compared to the Hausman-Taylor estimator on panel data where the true data-generating covariance structure is not exchangeable.

Bibliography

Ackerman, R. A., Kashy, D. A., & Corretti, C. A. (2015). A tutorial on analyzing data from speed-dating studies with heterosexual dyads. *Personal Relationships*, *22*, 92–110.

Alexandrowicz, R. W. (2015). Analyzing dyadic data with IRT models. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.), *Dependent data in social sciences research* (pp. 173–202). New York: Springer.

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using sas.* SAS Institute.

Allison, P. D. (2009). *Fixed effects regression models.* Sage.

Allison, P. D., Williams, R., & Moral-Benito, E. (2017). Maximum likelihood for cross-lagged panel models with fixed effects. *Socius*, *3*, 1–17.

Back, M. D., & Kenny, D. A. (2010). The social relations model: How to understand dyadic processes. *Social and Personality Psychology Compass*, *4*, 855–870.

Bagozzi, R. P., & Ascione, F. J. (2005). Inter-role relationshis in hospital-based pharmacy and therapeutics committee decision making. *Journal of Health Psychology*, *10*, 45–64.

Bakk, Z., & Kuha, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika*, *83*, 871–892.

Bickel, P. J., Klaassen, C. A., Ritov, Y., & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins University Press Baltimore.

Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces*, *89*, 1–34.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods & Research*, *5*, 3–52.

Card, N. A., Little, T. D., & Selig, J. P. (2008). Using the bivariate social relations model to study dyadic relationships: Early adolescents' perceptions of friends' aggression and prosocial behavior. In N. A. Card, T. D. Little, & J. P. Selig (Eds.), *Modeling dyadic and interdependent data in the developmental and behavioral sciences* (pp. 245–276). New York: Routledge.

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics*, *39*, 333–367.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, *18*, 5–46.

Christensen, P. N., & Kashy, D. A. (1998). Perceptions of and by lonely people in initial social interactions. *Personality and Social Psychology Bulletin*, *24*, 322–329.

Cockerham, C. C., & Weir, B. S. (1977). Quadratic analysis of reciprocal crosses. *Biometrics*, *33*, 187–203.

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, *27*, 1–43.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Dorff, C., & Ward, M. D. (2013). Networks, dyads, and the social relations model. *Political Science Research and Methods*, *1*, 159–178.

Duncan, O. D., Haller, O. A., & Portes, A. (1968). Peer influences on aspirations: A reinterpretation. *American Journal of Sociology*, *74*, 119–137.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New York, NY: Psychological Press.

Fisher, A., & Kennedy, E. H. (2018). Visually communicating and teaching intuition for influence functions. *arXiv preprint arXiv:1810.03260*.

Fisman, R., Iyengar, S. S., Kamenica, E., & Simonson, I. (2006). Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics*, *121*, 673–697.

Fuller, W. A., & Battese, G. E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, *68*, 626–632.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

Goldstein, H. (1987). Multilevel variance components models. *Biometrika*, *74*, 430–431.

Gong, G., & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, *74*, 861–869.

Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica*, *49*, 1377–1398.

Hayman, B. I. (1954). The theory and analysis of diallel crosses. *Genetics*, *39*, 789–809.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2017). A variational maximization-maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika*, *82*, 693–716.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York: Guilford.

Kenny, D. A. (1996). Models of nonindependence in dyadic research. *Journal of Social and Personal Relationships*, *13*, 279–294.

Kenny, D. A., & Kashy, D. A. (1994). Enhanced co-orientation in the perception of friends: A social relations analysis. *Journal of Personality and Social Psychology*, *67*, 1024–1033.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis.* New York: Guilford.

Kenny, D. A., & La Voie, L. (1984). The social relations model. *Advances in Experimental Social Psychology*, *18*, 141–182.

Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R.* Springer.

Koster, J. M., & Brandy, A. (2018). The effects of individual status and group performance on network ties among teammates in the National Basketball Association. *PLoS ONE*, *13*(e0196013).

Koster, J. M., & Leckie, G. (2014). Food sharing networks in lowland Nicaragua: An application of the social relations model to count data. *Social Networks*, *38*, 100–110.

Li, H., & Loken, E. (2002). A unified theory of statistical analysis and inference for variance components models for dyadic data. *Statistica Sinica*, *12*, 519–535.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Loeys, T., & Molenberghs, G. (2013). Modeling actor and partner effects in dyadic data when outcomes are categorical. *Psychological Methods*, *18*, 220–236.

Loncke, J., Eichelsheim, V. I., Branje, S. J. T., Buysse, A., Meeus, W. H. J., & Loeys, T. (2018). Factor score regression with social relations model components: A case study exploring antecedents and consequences of perceived support in families. *Frontiers in Psychology*, *9*(1699).

Lüdtke, O., Robitzsch, A., & Trautwein, U. (2018). Integrating covariates into social relations models: A plausible value approach for handling measurement error in perceiver and target effects. *Multivariate Behavioral Research*, *53*, 102–124.

Malloy, T. E., & Kenny, D. A. (1986). The social relations model: An integrative method for personality research. *Journal of Personality*, *54*, 199–225.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*, 335–341.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*, 69–85.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431–462.

Nestler, S. (2018). Likelihood estimation of the multivariate social relations model. *Journal of Educational and Behavioral Statistics*, *43*, 387–406.

Newey, W. K. (1988). Adaptive estimation of regression models via moment restrictions. *Journal of Econometrics*, *38*(3), 301–339.

Novomestky, F. (2012). *Matrixcalc: Collection of functions for matrix calculations.* R package version 1.0-3. Retrieved from https://CRAN.R-project.org/package=matrixcalc

Pimentel, S. D., Page, L. C., & Keele, L. (2016). An overview of optimal multilevel matching using network flows with the matchmulti package in R.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical modelling*, *7*(9-12), 1393–1512.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Shin, H. J., Rabe-Hesketh, S., & Wilson, M. (2019). Trifactor models for multiple-ratings data. *Multivariate Behavioral Research*, *54*, 1–22.

Sim, N., Gin, B., Skrondal, A., & Rabe-Hesketh, S. (2019). A dyadic item response theory model: Stan case study. Retrieved from https://github.com/education-stan/example-models/tree/master/education/dyadic_irt_model.

Skrondal, A., & Kuha, J. (2012). Improved regression calibration. *Psychometrika*, *77*, 649–669.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*, 563–575.

Snijders, T. A. B., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships*, *6*, 471–486.

Stan Development Team. (2018). *Rstan: The r interface to stan.* R package version 2.17.3. Retrieved from http://mc-stan.org

Teachman, J., Duncan, G. J., Yeung, W. J., & Levy, D. (2001). Covariance structure models for fixed and random effects. *Sociological Methods & Research*, *30*, 271–288.

Tsiatis, A. (2007). *Semiparametric theory and missing data.* Springer.

van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data.* Springer Science & Business Media.

van der Laan, M. J., & Rose, S. (2018). *Targeted learning in data science.* Springer.

van der Linden, W. J. (2016). *Handbook of item response theory, volume one: Models.* Boca Raton: CRC Press.

Warner, R., Kenny, D. A., & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, *37*, 1742–1757.

White, I. R. (2010). Simsum: Analysis of simulation studies including Monte Carlo error. *The Stata Journal*, *10*, 369–385.