Bringing Rank-Minimization Back In:

Estimating the Number of Common Inputs to a Data-Generating Process

Ben Goodrich, Columbia University

September 11, 2012

Abstract

This paper derives and implements an algorithm to infer the number of common inputs to a data-generating process from the outputs. Previous working dating back to the 1930s proves that this inference can be made in theory, but the practical difficulties have been too daunting to overcome. These obstacles can now be avoided by looking at the problem from a different perspective, utilizing some insights from the study of economic inequality, and relying on modern computer technology.

Now that there is a computational algorithm that can estimate the number of variables that generated observed outcomes, the scope for applications is quite large. Examples are given showing its use for evaluating the reliability of measures of theoretical concepts, empirically testing formal models, verifying whether there is an omitted variable in a regression, checking whether proposed explanatory variables are measured without error, evaluating the completeness of multiple imputation models for missing data, and facilitating the construction of matched pairs in randomized experiments. The algorithm is used to test the main hypothesis in Esping-Andersen (1990), which has been influential in the sociology and political science literatures, namely that various welfare-state outcomes are a function of only three underlying variables.

1 Introduction

Anonymous reviewers have often been known to question whether a concept is reliable, whether the data on a key explanatory variable contain measurement error, and / or whether a model fails to control for some omitted variable that is relevant to the data-generating process in question. What if there were an algorithm that could address these charges? Conversely, editors and reviewers could insist that this tool be used by authors to bolster the plausibility of their empirical conclusions. In fact, a theorem was proven in the 1930s that provides a theoretically rigorous basis for answering all of these questions. Unfortunately, the theorem has never been of practical use to applied researchers because the theorem requires the solution to an optimization problem that in general cannot be solved directly. However, this problem can now be solved *indirectly* using the algorithm to be developed in this paper, which, for the first time in seventy five years, makes it possible to answer these critical questions under the rather general conditions presupposed by the theorem. We implemented this algorithm as part of the accompanying R package (FA*i*R, version 0.6-0).

To illustrate the immense potential of the original theorem and new algorithm, we reevaluate the evidence for the main hypothesis in Esping-Andersen (1990) that welfare-state outcomes are primarily a function of three characteristics of a nation: its liberalism, conservativism, and social-democraticness. Although Esping-Andersen (1990) has been criticized on many grounds, no one can deny that it has had a great influence on researchers in both political science and sociology that are interested in the welfare state. Recently both Hicks and Kenworthy (2003) and Scruggs and Pontusson (2008) have argued empirically that the Esping-Andersen (1990) hypothesis can be condensed into a two-dimensional theory of welfare states, although they differ on how the two dimensions should be interpreted. Conversely, Esping-Andersen (1999) acknowledges the criticism that the theory should be *expanded* to include a fourth explanatory variable — namely how the nation responds to gender issues — but concludes that this issue can be adequately addressed within the original three-dimensional theory. We illustrate how the algorithm to be developed in this paper can be used to shed light on these controversies surrounding Esping-Andersen's (1990) hypothesis.

Our goal is to describe the *data-generating process* under weak assumptions that can be moderately violated without dramatically affecting our description. The procedure is not a full model but rather a "prelude to a model" that attempts to infer the number of explanatory variables that generated the observed data and also to estimate the error variances in the data. Sometimes this information is interesting in itself, but more often it allows the researcher to confidently estimate *another* model that conditions on this information.

The original theorem was stated in terms of fairly advanced matrix algebra, even in the 1930s when only a few social scientists had the training to understand it. Moreover, all the historical progress toward solving the optimization problem it raised has utilized matrix algebra and measure theory. While little advanced training is needed to *use* the software that implements the algorithm, rigorously proving that the algorithm provides an indirect solution to this optimization problem requires the appropriate tools. However, the next section simply gives several examples in political science where the answer the algorithm produces would be useful to applied researchers.

Thereafter, there are several technical sections describing the model, restating the key theorem, deriving an indirect algorithm to realize its potential, and demonstrating that it is successful in Monte Carlo simulations. The last major section returns to the main hypothesis in Esping-Andersen (1990). After the conclusion, there is a Technical Appendix that derives some derivatives and a Computational Appendix that describes how the IRMA is executed in our R package for factor analysis and LISREL models.

2 If This Were Important, Wouldn't Someone Have Solved It Already?

Suppose that we could discover the value of r, which is the number of inputs to the data-generating process for n variables and that we can determine the residual variance for each of these n variables when predicted by the r inputs. Later sections describe how this can be accomplished. For now, we simply ask in this section how would this information be useful to applied researchers and to political scientists in particular?

This knowledge would *always* useful in some way. In any scientific investigation, the researcher can only benefit from being able to separate the noise from the signal in the data and being able to enumerate the sources underlying the signal. Even if the researcher intends to estimate a different model, the algorithm developed here can be used to build or validate a subsequent model. In this section, we present relevant examples from the political science literature, starting with simple measurement models, proceeding through more complicated parametric models, and ending with the applicability to randomized experiments.

In traditional measurement theory, n observed variables are expressed as a function of the true concept plus a random error term. In other words, the data-generating process for the n measures has $r = 1$ common input, namely the true concept. In symbols, (in this paper the observations are *not* subscripted)

$$x_j = x^* + \epsilon_j,$$

for $j = 1, 2, \ldots, n$, where x^* is the true concept, around which the n observed measures randomly deviate. The goals are to verify whether this theory holds empirically, to estimate the error variances, and ultimately to say something about the underlying concept.

If this model holds, the n measures are reliable if the error variances are small. An alternative model is

$$x_j = \sum_{k=1}^{r} f_{kj}\left(\eta_k\right) + \epsilon_j,$$

implying x_j is some function of r inputs, plus a random error term. This more general model simplifies to the simple measurement model if $r = 1$ and $f\left(\eta\right) = x^*$. Even if $r > 1$ — which is inconsistent with the simple measurement model — we want to know what r is, the error variances, and what the inputs are.

Measurement is critical to political science and other fields. For example, Ansolabehere, Rodden and Snyder (2008) argues that "averaging a large number of survey items on the same broadly defined issue area — for example, government involvement in the economy, or moral issues — eliminates a large amount of measurement error and reveals issue preferences that are well structured and stable (215)." To back this claim, Ansolabehere, Rodden and Snyder (2008) recommends subjecting n survey items to a factor analysis model with one factor to estimate the error variances and factor scores.

Our only objection is that the factor analysis model in Ansolabehere, Rodden and Snyder (2008) assumes rather than discovers that $r = 1$, which is the critical assumption in a simple measurement model. If $r > 1$, then the unidimensional factor analysis model is misspecified, the estimates of measurement error are biased, the resulting factor score predictions are not well-defined much less well-estimated, and the inferences regarding the structure and stability of issue preferences are potentially confounded. Given these stakes, it is essential to rigorously substantiate the $r = 1$ hypothesis.

Like many papers in political science, there is little (reported) evidence in Ansolabehere, Rodden and Snyder (2008) to justify its $r = 1$ assumption. In fact, only a few sentences are devoted to this critical issue: "We scaled the items using principal factors factor analysis. In all cases we find a single dominant dimension [meaning eigenvalue]...Comparing results using the first factor and the simple average of individual items reveals that the two approaches are nearly identical. This further suggests that, at least in the ANES data, preferences on each issue are mainly one-dimensional (220)."

This argument is not persuasive. The average is not defined for ordinal survey items, so treating the survey responses as integers in order to average them does nothing to suggest that issue preferences are mainly one-dimensional, regardless of their similarity to factor scores on the first factor, which are also not well-defined when the responses are ordinal but treated as integers. But even if the data were continuous, neither averaging nor a unidimensional factor model directly speaks to the alternative hypothesis that $r > 1$. Each will be within an affine transformation of the other whenever the first eigenvector of the sample correlation matrix is nearly constant, regardless of how many factors there really are.

Moreover, whether the correlation matrix has one "dominant" (however defined) eigenvalue does not imply that $r = 1$. The basis, such as it is, for the widespread practice of inferring the number of factors from the sample correlation matrix loosely stems from Guttman (1954), which showed that the number of eigenvalues of the *population* correlation matrix that are greater than or equal to unity is a *lower bound* for r. Not only is it merely a lower bound for r, Guttman (1954) proved that two lower bounds for r are tighter and sought a procedure like the one developed in this paper find r definitively. However, the appendix of Ansolabehere, Rodden and Snyder (2008) shows that during the 1990s (at least) for data on economic issues (but not moral issues), two or three eigenvalues of the (sample) correlation matrix are greater than unity.

Top psychology journals today would require much more evidence for the assumed value of r in a factor analysis. Psychologists routinely test the null hypothesis that their assumed value of r is correct in the population, or at least is "close" to correct with a (noncentral) χ^2 statistic, and these test statistics (along with other goodness-of-fit indicators) are routinely calculated by popular statistical packages or can be bootstrapped (see Mebane and Sekhon 1998). The null hypothesis that the assumed value of r is correct is usually rejected for small values of r and reasonably large values of n, and there is still some controversy over testing whether a model is "close" to correct. Nevertheless, the political science literature would be much improved if reviewers and editors insisted on the inclusion of such statistics, or better yet, insisted on the procedure developed here and in Goodrich (2009) to infer r.

The above is not intended to single out Ansolabehere, Rodden and Snyder (2008). As illustrated below, there are many papers in political science and other disciplines that follow the same route, in part because an easy-to-use, methodologically sound, analytically satisfying way to discover r has not been available. Moreover, the fundamental point of Ansolabehere, Rodden and Snyder (2008) that political scientists should use multiple survey items to get a better handle on issue preferences is unquestionably correct. However, it would be better to estimate r explicitly in order to substantiate an interesting and important claim about the stability of preferences on political issues. This paper provides a way to do so.

Another good example of a measurement model is Treier and Jackman (2008), although the working paper version first written in 2003 contains more details. Treier and Jackman (2003) emphasizes that "latent variables abound in political science" and lists "public opinion, socio-economic status, social capital, ideology, [and] democracy (1)" as a few examples. In particular, we can obtain imperfect indicators of democracy from the Polity or Freedom House scores, but we do not know how imperfect these indicators are, whether democracy is a homogeneous concept, or whether democracy is a binary or continuous variable. Unfortunately, most empirical studies of democracy have used the "overall" Polity score, which is a weighted average of $n = 5$ primitive scores and all the steps along the path to this overall score could be (and have been) debated. The key contribution of Treier and Jackman (2008) is to derive a measurement model for the primitive democracy scores that properly accounts for the fact that they are ordinal (see also Quinn 2004). However, the only justification for the assumption in Treier and Jackman (2003) that $r = 1$ is that only one eigenvalue of the sample correlation matrix is greater than unity, and the sample correlation matrix seems to have been calculated without taking the ordinal nature of the data into account.

Perhaps the best example of a (somewhat more complicated) measurement model is the process by which NOMINATE scores for US legislators are created. Poole and Rosenthal (1997) claims that — for *recent* Congresses — an $r = 1$ model is sufficient to explain the variation in voting for all members in both chambers of Congress on all roll call votes, although an $r = 2$ model is better over all Congresses. Others disagree with the low dimensionality finding, as discussed in Poole and Rosenthal (1997, chapter 3). Rather than merely relying on in-sample predictions, it would be great if we could resolve this debate simply by estimating r from the roll call data using the algorithm to be developed in this paper.

The controversy over whether $r = 1$ in the NOMINATE context illustrates the importance of the $r = 1$ assumption to much of the Empirical Implications of Theoretical Models (EITM) literature. Many formal models in political science assume that agents vary along a single-dimension in order to invoke some theorem that proves there is a Condorcet-winning outcome (see Persson and Tabellini 2002, chapter 2). As is well-known, if agents vary on multiple dimensions, then it is quite difficult to obtain a Condorcet-winner in general and issues such as agenda control and insincere voting must be considered. If preferences over $n > 3$ issues were theorized to be a function of agents' positions on the same dimension, then the $r = 1$ assumption could be empirically tested using the algorithm developed in this paper.

There are also many examples where scholars theorize that $r > 1$. In the last section, we return to Esping-Andersen's (1990) famous hypothesis that welfare-state outcomes can be explained by $r = 3$ variables, namely the population's (social) conservatism, (economic) liberalism, and social democraticness. More recently, this $r = 3$ hypothesis has been called into question by Hicks and Kenworthy (2003) and by Scruggs and Pontusson (2008), both of which believe $r = 2$ but disagree over the nature of the two dimensions. The intention of Scruggs and Pontusson (2008) is to apply the same methods as Hicks and Kenworthy (2003) to a better dataset that also allows inferences over time. However, the primary technique for inferring r in Hicks and Kenworthy (2003) is the aforementioned "eigenvalues-greater-than-one-in-the-population" procedure, which at best provides a lower bound for r. In other words, even if the population correlation matrix were available, a lower bound of two is not inconsistent with Esping-Andersen's (1990) $r = 3$ theory.

The subfield of international relations has fewer examples where someone has explicitly made a claim about r, although Treier and Jackman (2008) uses the resulting measure of democracy to reexamine the democratic peace hypothesis. However, international economic flows — such as trade, investment, aid, loans, etc. — are often modeled with some variation of the "gravity model", which predicts that the volume of the flow is increasing in the size of the two countries in question and decreasing in the distance between the two countries. The gravity model is often augmented with additional variables in a somewhat ad hoc fashion, usually in an attempt to capture other aspects of the cost of trade, so we will call it a $r \geq 2$ theory. But how do we know when this or any other model is specified completely? We have plenty of data on economic transactions; we just need an algorithm to answer this question.

In any regression model, if the explanatory variables are measured with error, the coefficient estimates are biased. In an extremely specific case — namely, a linear model with random error in exactly one explanatory variable — the estimated coefficient of that variable is biased toward zero. In any other situation, the magnitude and direction of the biases of the coefficient estimates depend on unknown population parameters. Thus, it would be useful to have a method that first estimated the measurement error in the observed variables so that we could avoid running regressions when there is (substantial) measurement error in the explanatory variables. Although the algorithm in this paper is primarily intended to estimate r, if r is sufficiently small relative to the n observed variables, it also estimates the error in the observed variables and thus can be demonstrate that the measurement error in the explanatory variables is negligible.

The error variance estimates would also be useful when using multiple imputation to fill in missing data values by drawing from their conditional distribution before estimating another model. If the estimated variances of these conditional distributions are (badly) wrong, then, at best, the standard errors of the estimates from the analysis model are suspect. However, if the variances are (badly) wrong, it suggests that there may be an important omitted variable that should be included to predict the missing values, in which case the point estimates of the analysis model are suspect as well. We should guard against these possibilities by checking whether the estimated number of inputs is less than or equal to the number of available explanatory variables and whether the variances used to draw the multiple imputations can be independently validated.

Measurement error and omitted confounders are also an issue in experiments and other studies that use matching techniques to estimate treatment effects nonparametrically. The goal is to create pairs of individuals that have the same values on all the relevant inputs to the data-generating process for the outcomes of interest. The treatment is assigned (as good as) randomly to one individual within each pair. The main obstacles to inference are the same as in parametric models. How do we know how many inputs there are to the data-generating process? Even if we have all the right inputs, what if they are measured with error? Even if they are measured without error, how do we obtain balance on all of them simultaneously?

In addition to estimating the number of inputs and perhaps the error variances, our algorithm can, under additional assumptions, produce factor scores on the r inputs, like in Ansolabehere, Rodden and Snyder (2008). This feature implies that we could match on r synthetic variables instead of n observed variables and the error in the synthetic variables (which can be estimated) may be less than in the observed variables. Thus, achieving balance on r synthetic variables could be easier than finding balance on n observed variables, a principle that is used (in a somewhat different form) in Abadie, Diamond and Hainmueller (2007).

Even if our data are measured perfectly, they can still be stochastic. In a typical voter turnout experiment, we have data on whether individuals voted in the previous few elections and perhaps a few other confounders. Elwert and Winship (2008, section 5) argues that we should *not* match on past turnout decisions because such variables are (at least partially) "colliders" with respect to the other confounders and thus conditioning on them can induce "endogenous selection bias". Of course, not matching on them can induce omitted variable bias. One way out of this dilemma is subject all pre-treatment variables to the algorithm developed here, find r, and match on the r factor scores, which are not colliders but are confounders. If the treatment and the control groups are balanced on the r factors, inference can proceed as usual.

In summary, there are many situations in which knowing how many inputs were involved in the data-generating process for n outcome variables. The aforementioned examples from political science merely illustrate general situations that arise across data-driven disciplines. We all want to know about the data-generating process for any data we observe. We all want to be able to separate signal from noise. When attempting to measure a concept such as citizen ideology, democracy, or ideal points of legislators or Justices, political scientists want to know whether $r = 1$, and the same is true when considering the empirical implications of a theoretical model where agents vary on a single-dimension. Regression models and multiple imputation models postulate that outcomes are a function of r independent variables, and we need to know r in order to plausibly claim that there is little to no omitted variable bias. The same principles hold when the explanatory variables are latent, as is the case in factor analysis and many structural equation models. When the explanatory variables are observed, we need to know that they are largely free of measurement error. Finally, even in an experiment where we are trying to estimate a treatment effect nonparametrically, it would be useful to know r in order to know how many variables must be matched on or to construct r such synthetic variables from $n > r$ observed variables. Our algorithm speaks to all of these situations.

3 General LISREL Population Model

This section lays the groundwork for Thurstone's (1935) minimum rank theorem but incorporates some subsequent generalizations to the setup. Whereas Thurstone (1935) stated the theorem in terms of a factor analysis model, with some additional assumptions, it is also applicable to a general LISREL model that includes factor analysis, linear regression, simultaneous equations, and other models as special cases.

Although it is less common than other parameterizations, Jöreskog and Sörbom (1996), Hayduk (1987), and others write a general LISREL model — that includes all other LISREL models as special cases — with just two equations, one structural equation and one measurement equation:

$$\begin{aligned} \boldsymbol{\eta} &= \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}, \\ \boldsymbol{\Omega}^{-1}\mathbf{y} &= \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \end{aligned}$$

where $\boldsymbol{\eta}$ is a column vector that includes all latent variables with structural errors $\boldsymbol{\zeta}$, $\mathbf{y}$ is a column vector of all manifest variables with measurement errors $\boldsymbol{\epsilon}$, and $\boldsymbol{\Lambda}$ and $\mathbf{B}$ are conformable coefficient matrices. A latent variable is exogenous iff it is not a function of any other latent variable, which can be ascertained by inspecting the exclusion restrictions in $\mathbf{B}$ (which necessarily has zeros along its diagonal). Consequently, a manifest variable is exogenous iff it is not a function of any endogenous latent variables.

All variables are expressed as deviations from their expectations to eliminate intercepts, but some normalizations of the variances are necessary. The latent variables have no intrinsic scale, so they are assumed to have unit variances. $\boldsymbol{\Omega}$ is a diagonal matrix whose diagonal elements contain the standard deviations of the manifest variables, which are assumed to be positive but finite. It is important to derive algorithms that do not depend on $\boldsymbol{\Omega}$, which is discussed more later.

If $\boldsymbol{\epsilon}$ is uncorrelated with $\boldsymbol{\eta}$, the covariance matrices among the variables can be written as

$$\begin{aligned} \boldsymbol{\Upsilon} &= \operatorname{cor}(\boldsymbol{\eta}) = (\mathbf{I} - \mathbf{B})^{-1} \boldsymbol{\Psi} (\mathbf{I} - \mathbf{B})^{-1'}, \\ \boldsymbol{\Sigma} &= \operatorname{cov}(\mathbf{y}) = \boldsymbol{\Omega} \left(\boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \right) \boldsymbol{\Omega}, \end{aligned}$$

where $\boldsymbol{\Theta} = \operatorname{cov}(\boldsymbol{\epsilon})$ and $\boldsymbol{\Psi} = \operatorname{cov}(\boldsymbol{\zeta})$. Both $\operatorname{cor}(\mathbf{y}) = \boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$ and $\operatorname{cor}(\boldsymbol{\eta}) = \boldsymbol{\Upsilon}$ must be restricted so that all their diagonal cells are unity. $\boldsymbol{\Lambda}$ and $\boldsymbol{\Upsilon}$ are assumed to have full rank and $\boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}'$ is assumed to have the same rank. If these assumptions hold, then the second equation can be written as a reduced form factor analysis model, $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\left(\dot{\boldsymbol{\Lambda}}\dot{\boldsymbol{\Lambda}}' + \boldsymbol{\Theta}\right)\boldsymbol{\Omega}$, where $\dot{\boldsymbol{\Lambda}}$ is some matrix with same dimensions as $\boldsymbol{\Lambda}$ such that $\dot{\boldsymbol{\Lambda}}\dot{\boldsymbol{\Lambda}}' = \boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}'$. It is somewhat useful, but not strictly necessary, to eliminate the rotational indeterminacy by requiring $\dot{\boldsymbol{\Lambda}}'\boldsymbol{\Theta}^{-1}\dot{\boldsymbol{\Lambda}}$ to be diagonal, which provides a canonical basis for $\dot{\boldsymbol{\Lambda}}$. No attempt will be made to recover $\boldsymbol{\Lambda}$, $\boldsymbol{\Upsilon}$, or any of the structural parameters, so the reduced form factor analysis representation of the LISREL model is sufficient for this paper's purposes, which all hinge on $\boldsymbol{\Theta}$.

While it is not logically necessary to assume that $\boldsymbol{\Theta}$ is diagonal, this assumption is made to greatly increase the clarity of the results but should be relaxed in future work. Covariance matrices are necessarily positive semi-definite (PSD), so any proposal for $\boldsymbol{\Theta} = \operatorname{cov}(\boldsymbol{\epsilon})$, denoted $\widetilde{\boldsymbol{\Theta}}$, must imply that both $\widetilde{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega} = \operatorname{cov}(\mathbf{y}|\,\widetilde{\boldsymbol{\eta}})$ are PSD. We actually make the slightly stronger assumption that $0 < \widetilde{\Theta}_{ii} < 1 \forall i$ such that $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ is PSD, and let $\mathcal{T}$ represent the set of admissible diagonal proposals for $\boldsymbol{\Theta}$.

Let $n \geq 3$ be the order of $\boldsymbol{\Sigma}$, and let $\mathcal{R}(\cdot)$ signify the rank function, which holds a central place in Thurstone's (1935) theorem to be given below. The rank of a matrix is equal to the number of linearly independent columns it has and, in the case of a symmetric matrix, the number of non-zero eigenvalues it has. In this paper, all the relevant matrices are symmetric. If all the eigenvalues of a symmetric matrix are positive, the matrix is positive definite and has full rank. If all the eigenvalues of a symmetric matrix are non-negative, the matrix is PSD, and its rank is decremented by each eigenvalue that is zero.

When the LISREL model can be represented as a reduced form factor analysis model, results from the factor analysis case carry over. Let the Ledermann (1937) bound be $L(n) = 0.5\left(2n + 1 - (8n+1)^{\frac{1}{2}}\right)$, and let r be the minimum rank of $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ over all $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$.

Theorem 1. *(i) The minimum number of factors consistent with a factor analysis model is r.*

(ii) If $r < L(n)$, then the rank-minimizing $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$ is almost surely unique and equal to $\boldsymbol{\Theta}$.

(iii) a: If $r = L(n)$, then a rank-minimizing $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$ is almost surely locally unique. b: If $r > L(n)$, a rank-minimizing $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$ is almost surely not locally unique. c: The set of covariance matrices whose rank can be reduced below $L(n)$ by changing their diagonal cells has measure zero.

Proof. Thurstone (1935) proves (i), Bekker and ten Berge (1997) proves (ii), and Shapiro (1982) proves (iii). □

An algorithm that chooses $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$ to minimize the rank of $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ would be extremely useful because its optimum would tell us the value of r and, if $r < L(n)$, also the value of $\boldsymbol{\Theta}$. Consequently, $\min\limits_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{\mathcal{R}\left(\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}\right)\right\}$ is called a rank-minimization problem (RMP). We could equivalently conceptualize the RMP as a choice over the diagonal elements of the reduced covariance matrix, where the diagonal elements are called "communalities" in the literature.

It is also important to recognize what theorem 1 does *not* claim. It does not claim that r necessarily "is" the number of latents that generated the observed variables. There are several exceptions that prevent such a broad claim. First, recall the assumption that $\mathbf{\Lambda}$ and $\mathbf{\Upsilon}$ have full rank and $\mathbf{\Lambda\Upsilon\Lambda}'$ has the same rank. A factor analysis model entails these assumptions, but in a general LISREL model, they can fail to hold in three ways. The first is the case where $\mathcal{R}\left(\mathbf{\Upsilon}\right) < \mathcal{R}\left(\mathbf{\Lambda}\right)$, which is to say that there is at least one exact linear relationship among the latent variables. For example, in a "latent class model" $\boldsymbol{\eta}$ consists of mutually exclusive and exhaustive dummy variables, one for each of the latent classes. Only the contrasts from the reference class can be identified, which is to say that $\mathbf{\Upsilon}$ is singular. However, the model can be reparameterized by "dropping" the reference class to obtain a new model where both $\mathbf{\Lambda}$ and $\mathbf{\Upsilon}$ have full rank. The second is the case where $\mathcal{R}\left(\mathbf{\Lambda}\right) < \mathcal{R}\left(\mathbf{\Upsilon}\right)$, which is to say that there is at least one exact linear relationship among the coefficients. This situation can occur in a "second-order" factor analysis model or in "reduced-rank" regression. Although the total number of latents can be arbitrarily large in this case, the minimum rank of $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ is still relevant because it is equal to the number of first-order factors or the rank in reduced-rank regression respectively. The third is the case where $\mathcal{R}\left(\mathbf{\Lambda}\right) = \mathcal{R}\left(\mathbf{\Upsilon}\right) > \mathcal{R}\left(\mathbf{\Lambda\Upsilon\Lambda}'\right)$, which can occur when some column of $\mathbf{\Lambda}$ has only one non-zero cell. The only remedy for this case is to increase n.

Even if $\mathcal{R}\left(\mathbf{\Lambda\Upsilon\Lambda}'\right) = \mathcal{R}\left(\mathbf{\Lambda}\right) = \mathcal{R}\left(\mathbf{\Upsilon}\right)$, it is still possible that r is not the number of latents. Guttman (1958) asks "To what extent can communalities reduce rank?" and proves that if $\mathbf{\Sigma}$ is non-singular, then many $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$ can be found that reduce the rank of $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ to $n-1$. Thus, if $\mathbf{\Sigma}$ is not generated by a LISREL model or is generated by a LISREL model with at least n latents (such as a simplex model), then r is not equal to the number of latents. Bekker and de Leeuw (1987) provides the most elegant proof that a necessary and sufficient condition for $r = n-1$ is that the cells of $\mathbf{\Sigma}^{-1}$ are all positive (possibly after multiplying some manifest variables by -1). Thus, if the cells of $\mathbf{\Sigma}^{-1}$ have mixed signs, then $r \leq n-2$, regardless of whether a LISREL model holds. Shapiro's (1982) result in theorem 1 (iii) part c, which does not utilize the restriction that $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$, comes closest to answering Guttman's (1958) question by showing that communalities have zero probability of reducing the rank below $L\left(n\right)$.

The simulations toward the end of this paper cannot prove anything deductively but do suggest that the set of covariance matrices such that $\min\limits_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}}\left\{\mathcal{R}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right)\right\} \leq n-4$ has measure zero, which would be a stronger result than Shapiro's (1982) when $n \geq 10$. Also, the set of covariance matrices among

$n = 6$ variables such that $\min_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ \mathcal{R}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right)\right\} = 3$ appears to have measure zero. However, the set of covariance matrices among $n \geq 7$ variables such that $\min_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ \mathcal{R}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right)\right\} = n - 3$ seems to have positive measure that increases with n. To substantiate these three new conjectures, a general solution to the RMP must be found.

4 Existing Factor Analysis Estimators

Unfortunately, the RMP is essentially impossible to solve directly. This section discusses why and categorizes existing estimators of the factor analysis model, almost all of which make an assumption about the number of latents. Doing so sets the stage for the next section, which solves the RMP indirectly.

As has been known since the early history of multiple factor analysis, there is no general analytic solution to the RMP, although it can be solved analytically in some special cases. Intuitively, we need to choose $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$ to set $n - r$ eigenvalues of $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ equal to zero, but when $n \geq 5$ we generally cannot solve the characteristic polynomial to arrive at an analytic expression for the eigenvalues. Worse, $\min_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ \mathcal{R}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right)\right\}$ cannot be solved numerically because the objective function takes continuous parameters but outputs an integer. Essentially all optimization algorithms are intended to minimize a continuous function of continuous parameters or are intended to minimize a general function of discrete parameters whose domain is a finite set. Fazel, Hindi and Boyd 2004 notes that the RMP is NP-hard.

Consequently, some researchers have turned to *continuous* heuristics that may find something approximately equal to $\arg\min_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ \mathcal{R}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right)\right\}$. Bentler (1972), Shapiro (1982), Della Riccia and Shapiro (1982), and Shapiro and ten Berge (2000) advocate solving $\arg\min_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ \mathrm{Tr}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right)\right\}$, but the solution minimizes the sum of the eigenvalues, rather than the number of eigenvalues that are positive. Another heuristic is $\arg\min_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ \ln\left(\det\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right)\right)\right\}$, which is advocated by Fazel, Hindi and Boyd (2004). However, when $\mathcal{R}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right) < n$, this objective function is infinite. Thus, there are an infinite number of local minima that are not minimum-rank solutions but are as good as a minimum-rank solution with respect to the log-determinant heuristic. The hope is that the shortest path to a local minimum leads to an approximate minimum-rank solution. The algorithm in the next section essentially finds a happy medium between the $\arg\min_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ \mathrm{Tr}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right)\right\}$ and $\arg\min_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ \ln\left(\det\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right)\right)\right\}$ heuristics.

A much more common approach is to estimate $\mathbf{\Theta}$ conditional on k factors. These estimators can be divided into those that enforce the restriction that $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$, those that enforce the restriction that $\widetilde{\mathbf{\Theta}}$ is PSD but do not enforce the restriction that $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ is PSD, and those that enforce neither restriction. Most factor analysis estimators fall in the middle category.

Ten Berge and Kiers (1991), Shapiro and ten Berge (2002), and Sočan (2003) discuss an estimator called Minimum Rank Factor Analysis (MRFA) that enforces the restriction that $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$. MRFA is a bit of a misnomer because the rank of the reduced covariance matrix is not the function being minimized by the algorithm. Rather, MRFA minimizes what Shapiro and ten Berge (2002) calls the Unexplained Common Variance (UCV), which is the sum of the trailing $n - k$ eigenvalues of the reduced covariance or reduced correlation matrix for some value of k that is specified in advance. The primary advantage of estimators that enforce the restriction that $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$ is that the communalities are "proper", which permits rigorous statements like "a factor analysis model with k factors would leave only five percent of the manifest variance unexplained". Such statements would not be coherent if any of the eigenvalues are negative, which is the case for most factor analysis estimators.

In summary, in the absence of an algorithm to solve the RMP, researchers have turned to other ways to estimate factor analysis models, most of which estimate $\mathbf{\Theta}$ conditional on k factors. Of course, it becomes imperative to set k equal to r, but history has shown that doing so can be difficult despite dozens of proposed ways to do so. It is important to distinguish between estimators that enforce the restriction that $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$ and those that do not. The algorithm to be developed in the next section does so but does not condition on a prespecified number of factors.

5 Indirect Rank-Minimization Algorithm (IRMA)

The trick to minimizing the rank of a matrix numerically is to approach it indirectly. A direct approach entails the fundamental difficulty that the rank function is knife-edged: If an eigenvalue of the reduced covariance matrix is "barely" greater than zero, it increments the rank just as much as an eigenvalue that is "substantially" greater than zero. In other words, if $\mathcal{R}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right) > r$, the computer does not know how "close" $\widetilde{\mathbf{\Theta}}$ is to a rank-minimizing solution or in which directions to move in $\widetilde{\mathbf{\Theta}}$-space. A numerical optimization algorithm needs more information to guide it toward a minimum-rank solution.

The sizes of the eigenvalues implied by $\widetilde{\mathbf{\Theta}}$ constitute this additional information. In other words, the objective function should take into account not merely how many implied eigenvalues are exactly zero but how far they are from zero. Most factor analysis estimators utilize this idea but depend only on the trailing $n - k$ eigenvalues for a given value of k. This section shows that a minimum-rank solution will be found indirectly if $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$ is chosen to numerically maximize the "dispersion" of *all* the implied eigenvalues using a dispersion function that is sufficiently sensitive to all near-zero eigenvalues; hence the acronym IRMA.

Guttman (1977) persuasively argues that one must demonstrate the conditions under which a methodological technique is successful *in the population* before sampling behavior should even be considered. In contrast to the methods Guttman (1977) criticizes, the IRMA finds a minimum-rank solution if given $\mathbf{\Sigma}$ without conditioning on r. Sampling is considered only briefly here and further in Goodrich (2009).

5.1 Four Scale Invariant Matrices

The non-zero eigenvalues of $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ depend on $\mathbf{\Omega}$ as well as $\widetilde{\mathbf{\Theta}}$. Thus, if a researcher were to multiply the ith manifest variable by $b \neq 1$, not only would Ω_{ii} change by a factor of b, the $\widetilde{\mathbf{\Theta}}$ that maximizes the eigenvalue dispersion does, in some cases, change in a complicated, nonlinear fashion. Optimizing with respect to a scale-dependent function is frowned upon, necessitating another matrix that is a function of $\widetilde{\mathbf{\Theta}}$ and satisfies three additional criteria: it 1) is PSD iff $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ is PSD, 2) has the same rank as $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$, and 3) entails cancellation of each $\mathbf{\Omega}$ with an $\mathbf{\Omega}^{-1}$. If such a matrix were available, then the dispersion of *its* eigenvalues can be safely maximized, rather than the eigenvalues of $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$. Fortunately, four such matrices exist.[1]

The first candidate is called the "reduced correlation matrix corrected for attenuation". Recalling that $\mathbf{\Omega}$ is diagonal and that multiplication of diagonal matrices is commutative, the reduced correlation matrix

[1] There are actually two more scale-invariant matrices that have the same theoretical properties as the four discussed here. Whether the additional two have any benefits over the four discussed here is currently under evaluation.

corrected for attenuation is

$$\begin{aligned}
\boldsymbol{\Pi} &= \mathrm{diag}\left(\boldsymbol{\Sigma} - \boldsymbol{\Omega\Theta\Omega}\right)^{-\frac{1}{2}} \left(\boldsymbol{\Sigma} - \boldsymbol{\Omega\Theta\Omega}\right) \mathrm{diag}\left(\boldsymbol{\Sigma} - \boldsymbol{\Omega\Theta\Omega}\right)^{-\frac{1}{2}}, \\
&= \mathrm{diag}\left(\boldsymbol{\Omega\Lambda\Upsilon\Lambda}'\boldsymbol{\Omega}\right)^{-\frac{1}{2}} \left(\boldsymbol{\Omega\Lambda\Upsilon\Lambda}'\boldsymbol{\Omega}\right) \mathrm{diag}\left(\boldsymbol{\Omega\Lambda\Upsilon\Lambda}'\boldsymbol{\Omega}\right)^{-\frac{1}{2}}, \\
&= \boldsymbol{\Omega}^{-\frac{1}{2}}\mathrm{diag}\left(\boldsymbol{\Lambda\Upsilon\Lambda}'\right)^{-\frac{1}{2}} \boldsymbol{\Omega}^{-\frac{1}{2}} \left(\boldsymbol{\Omega\Lambda\Upsilon\Lambda}'\boldsymbol{\Omega}\right) \boldsymbol{\Omega}^{-\frac{1}{2}}\mathrm{diag}\left(\boldsymbol{\Lambda\Upsilon\Lambda}'\right)^{-\frac{1}{2}} \boldsymbol{\Omega}^{-\frac{1}{2}}, \\
&= \mathrm{diag}\left(\boldsymbol{\Lambda\Upsilon\Lambda}'\right)^{-\frac{1}{2}} \boldsymbol{\Omega}^{-1} \left(\boldsymbol{\Omega\Lambda\Upsilon\Lambda}'\boldsymbol{\Omega}\right) \boldsymbol{\Omega}^{-1}\mathrm{diag}\left(\boldsymbol{\Lambda\Upsilon\Lambda}'\right)^{-\frac{1}{2}}, \\
&= \mathrm{diag}\left(\boldsymbol{\Lambda\Upsilon\Lambda}'\right)^{-\frac{1}{2}} \left(\boldsymbol{\Lambda\Upsilon\Lambda}'\right) \mathrm{diag}\left(\boldsymbol{\Lambda\Upsilon\Lambda}'\right)^{-\frac{1}{2}}, \\
&= \mathrm{diag}\left(\mathrm{cor}\left(\mathbf{y}\right) - \boldsymbol{\Theta}\right)^{-\frac{1}{2}} \left(\mathrm{cor}\left(\mathbf{y}\right) - \boldsymbol{\Theta}\right) \mathrm{diag}\left(\mathrm{cor}\left(\mathbf{y}\right) - \boldsymbol{\Theta}\right)^{-\frac{1}{2}},
\end{aligned}$$

which demonstrates the irrelevance of $\boldsymbol{\Omega}$. $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ indicates the *proposal* for $\boldsymbol{\Pi} = \mathrm{cor}\left(\mathbf{y}|\,\boldsymbol{\eta}\right)$ implied by $\widetilde{\boldsymbol{\Theta}}$ such that $0 < \widetilde{\Theta}_{ii} < 1\,\forall i$. $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ is PSD iff $\mathrm{cor}\left(\mathbf{y}\right) - \widetilde{\boldsymbol{\Theta}}$ is PSD and has the same rank.

Let π_j be the jth largest eigenvalue of $\boldsymbol{\Pi}$. Kaiser and Caffrey (1965) notes that $\alpha_j = \frac{n}{n-1}\left(1 - \frac{1}{\pi_j}\right) \in (-\infty, 1]$ is (a lower bound to) the "generalizeability" of the jth factor to the universe of variables that the manifest variables are taken from. Unfortunately, $\pi_j < 1 \iff \alpha_j < 0$, which is a logical problem because a generalizeability cannot be negative and is a computational problem because the IRMA requires non-negative quantities. To circumvent this issue, we define a non-negative quantity that is "like" α_j — particularly when π_j is large — namely, $\alpha_j^* = \frac{n+1}{n}\left(1 - \frac{1}{\pi_j+1}\right) \in [0, 1]$, where α_j^* is the jth largest eigenvalue of $\mathbf{A}^* = \frac{n+1}{n}\left(\mathbf{I} - \left(\boldsymbol{\Pi} + \mathbf{I}\right)^{-1}\right)$. $\mathbf{A}^*\left(\widetilde{\boldsymbol{\Theta}}\right)$ inherits the three critical properties from $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$.

$\boldsymbol{\Pi}$ and $\mathbf{A}^*$ scale $\mathrm{cor}\left(\mathbf{y}\right) - \boldsymbol{\Theta}$ from both sides by root-communalities. The alternative is to scale by root-uniquenesses. Let $\boldsymbol{\Phi}$ be a diagonal matrix that contains the eigenvalues of $\left(\boldsymbol{\Omega\Theta\Omega}\right)^{-\frac{1}{2}} \boldsymbol{\Sigma} \left(\boldsymbol{\Omega\Theta\Omega}\right)^{-\frac{1}{2}} = \boldsymbol{\Theta}^{-\frac{1}{2}}\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega}\left(\boldsymbol{\Lambda\Upsilon\Lambda}' + \boldsymbol{\Theta}\right)\boldsymbol{\Omega}\right)\boldsymbol{\Omega}^{-1}\boldsymbol{\Theta}^{-\frac{1}{2}} = \boldsymbol{\Theta}^{-\frac{1}{2}}\left(\boldsymbol{\Lambda\Upsilon\Lambda}' + \boldsymbol{\Theta}\right)\boldsymbol{\Theta}^{-\frac{1}{2}} = \boldsymbol{\Theta}^{-\frac{1}{2}}\left(\mathrm{cor}\left(\mathbf{y}\right) - \boldsymbol{\Theta}\right)\boldsymbol{\Theta}^{-\frac{1}{2}} + \mathbf{I}$. Although $\boldsymbol{\Phi} - \mathbf{I}$ is not readily interpretable, it does not depend on $\boldsymbol{\Omega}$. Assuming $\widetilde{\Theta}_{ii} > 0\,\forall i$, $\boldsymbol{\Phi}\left(\widetilde{\boldsymbol{\Theta}}\right) - \mathbf{I}$ is PSD iff $\mathrm{cor}\left(\mathbf{y}\right) - \widetilde{\boldsymbol{\Theta}}$ is PSD and has the same rank, so Thurstone (1935) could have equivalently formulated the RMP in terms of $\boldsymbol{\Phi}\left(\widetilde{\boldsymbol{\Theta}}\right) - \mathbf{I}$.

The final candidate is derived in Guttman (1955, 1956) and can be written as $\boldsymbol{\Delta}\left(\widetilde{\boldsymbol{\Theta}}\right) = \mathrm{cov}\left(\widetilde{\boldsymbol{\eta}}|\,\mathbf{y}\right) = \mathbf{I} - \boldsymbol{\Phi}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-1}$ (iff the loadings are canonical). If $\widetilde{\boldsymbol{\Theta}} = \boldsymbol{\Theta}$, then $\boldsymbol{\Delta}\left(\widetilde{\boldsymbol{\Theta}}\right) = \begin{bmatrix} \boldsymbol{\Delta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, where the jth largest eigenvalue of $\boldsymbol{\Delta} = \dot{\boldsymbol{\Lambda}}'\mathrm{cor}\left(\mathbf{y}\right)^{-1}\dot{\boldsymbol{\Lambda}}$ can be interpreted as the squared correlation between the jth factor and the predicted scores on the jth factor in the population. $\mathbf{\Delta}$ (re)ignited the highly-charged controversy over factor score indeterminacy in the 1970s (see Mulaik 2005 for a review). $\mathbf{\Delta}\left(\widetilde{\mathbf{\Theta}}\right)$ inherits scale independence from $\mathbf{\Phi}\left(\widetilde{\mathbf{\Theta}}\right)$, is PSD iff $\mathbf{\Phi}\left(\widetilde{\mathbf{\Theta}}\right)-\mathbf{I}$ is PSD, and has the same rank, so it too is IRMA-eligible.

To reemphasize, $\mathbf{\Sigma}-\mathbf{\Omega\Theta\Omega}=\text{cov}\left(\mathbf{y}|\,\boldsymbol{\eta}\right)$ is scale-dependent, but $\mathbf{\Pi}=\text{cor}\left(\mathbf{y}|\,\boldsymbol{\eta}\right)$, $\mathbf{A}^{*}$, $\mathbf{\Phi}-\mathbf{I}$, and $\mathbf{\Delta}=\text{cov}\left(\boldsymbol{\eta}|\,\mathbf{y}\right)$ are scale-invariant. All five matrices are PSD with rank r and can be written as functions of $\mathbf{\Theta}$, so the RMP could be conceptualized in terms of any of them. Everyone prefers scale-invariant estimators, but which of the four scale-invariant matrices is the best candidate for indirect rank-minimization? In the population, it makes no difference, at least in theory, although in practice the performance with $\mathbf{\Phi}\left(\widetilde{\mathbf{\Theta}}\right)-\mathbf{I}$ is so horrific that we removed this option from FA*i*R. In a sample, one of the remaining three scale-invariant matrices could yield better performance in particular situations, which is assessed by simulations in section 6, and the differences are small. Interpretability is not a valid consideration because once the optimum with respect to any of these matrices is found, *any* (identified) quantity can be interpreted.

5.2 Preliminary Theoretical Results on the Road to the IRMA

A lower-case Greek letter subscripted by j signifies the jth largest eigenvalue of the corresponding matrix denoted with an upper-case, boldface Greek letter. Tildes are used to distinguish proposals from population parameters, and hats are used to distinguish optimal proposals from generic proposals. Hence, $\widetilde{\pi}_j=\pi_j\left(\widetilde{\mathbf{\Theta}}\right)$, $\widetilde{\alpha}_j^{*}=\alpha_j^{*}\left(\widetilde{\mathbf{\Theta}}\right)$, $\widetilde{\phi}_j-1=\phi_j\left(\widetilde{\mathbf{\Theta}}\right)-1$, and $\widetilde{\delta}_j=\delta_j\left(\widetilde{\mathbf{\Theta}}\right)$ are admissible proposals for $\pi_j\in[0,n]$, $\alpha_j^{*}\in[0,1]$, $\phi_j-1\in[0,\infty)$, and $\delta_j\in[0,1)$, which are in turn are the jth largest eigenvalues of $\mathbf{\Pi}$, $\mathbf{A}^{*}$, $\mathbf{\Phi}-\mathbf{I}$, and $\mathbf{\Delta}$. When taken as n-vectors with non-increasing elements, proposals for $\boldsymbol{\pi}$, $\boldsymbol{\alpha}^{*}$, $\boldsymbol{\phi}-\mathbf{1}$, and $\boldsymbol{\delta}$ are denoted by $\widetilde{\boldsymbol{\pi}}=\boldsymbol{\pi}\left(\widetilde{\mathbf{\Theta}}\right)$, $\widetilde{\boldsymbol{\alpha}}^{*}=\boldsymbol{\alpha}^{*}\left(\widetilde{\mathbf{\Theta}}\right)$, $\widetilde{\boldsymbol{\phi}}-\mathbf{1}=\boldsymbol{\phi}\left(\widetilde{\mathbf{\Theta}}\right)-\mathbf{1}$, and $\widetilde{\boldsymbol{\delta}}=\boldsymbol{\delta}\left(\widetilde{\mathbf{\Theta}}\right)$ respectively.

As a shorthand for when the distinctions among the four scale-invariant matrices are unimportant, let x_j to be the jth largest eigenvalue of any of these population matrices with $\widetilde{x}_j=x_j\left(\widetilde{\mathbf{\Theta}}\right)$ being the corresponding proposal implied by $\widetilde{\mathbf{\Theta}}$. Let $\mathbf{x}$ be a vector of population eigenvalues and let $\widetilde{\mathbf{x}}=\mathbf{x}\left(\widetilde{\mathbf{\Theta}}\right)$ be a proposal for it. Let $\overline{x}$ be the mean eigenvalue in the population and $\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)$ be the mean of the proposed eigenvalues. Finally, let $\widehat{x}_j=x_j\left(\widehat{\mathbf{\Theta}}\right)$ and $\widehat{\mathbf{x}}=\mathbf{x}\left(\widehat{\mathbf{\Theta}}\right)$ be optimal proposals for their population counterparts. We also draw occasional parallels to economic inequality literature where $\mathbf{x}$ is instead a vector of incomes.[2]

[2] If $\mathbf{X}\left(\widetilde{\mathbf{\Theta}}\right)$ is positive semi-definite, then it can be factored as $\mathbf{X}\left(\widetilde{\mathbf{\Theta}}\right)=\mathbf{L}\left(\widetilde{\mathbf{\Theta}}\right)\mathbf{D}\left(\widetilde{\mathbf{\Theta}}\right)\mathbf{L}\left(\widetilde{\mathbf{\Theta}}\right)'$, in which case it would be

Appendix A characterizes how $x_j\left(\widetilde{\boldsymbol{\Theta}}\right)$ changes as $\widetilde{\boldsymbol{\Theta}}$ changes, resulting in

Lemma 2. *If $\widetilde{x}_j$ is not simple, which is to say that it is equal to another eigenvalue, its derivative with respect to $\widetilde{\Theta}_{ii}$ does not exist at $\widetilde{\boldsymbol{\Theta}}$. Otherwise,*

(i) $\frac{\partial \pi_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\left(\widetilde{\pi}_j - 1\right)\widetilde{P}_{ij}^2}{1 - \widetilde{\Theta}_{ii}}$*, where $\widetilde{\mathbf{P}}$ contains the orthonormal eigenvectors of $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$. Thus,* $\frac{\partial \pi_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} \geq 0$ *when $\widetilde{\pi}_j > 1$ and* $\frac{\partial \pi_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} \leq 0$ *when $\widetilde{\pi}_j < 1$.*

(ii) $\frac{\partial \alpha_j^*\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\left(\widetilde{\pi}_j - 1\right)\widetilde{P}_{ij}^2 (n+1)}{\left(\widetilde{\pi}_j + 1\right)^2 \left(1 - \widetilde{\Theta}_{ii}\right) n}$*. Thus,* $\frac{\partial \alpha_j^*\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} \geq 0$ *when $\widetilde{\pi}_j > 1$ and* $\frac{\partial \alpha_j^*\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} \leq 0$ *when $\widetilde{\pi}_j < 1$.*

(iii) $\frac{\partial \phi_j\left(\widetilde{\boldsymbol{\Theta}}\right) - 1}{\partial \widetilde{\Theta}_{ii}} = \frac{\widetilde{\phi}_j \widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii}} \geq 0$*, where $\widetilde{\mathbf{G}}$ contains the orthonormal eigenvectors of $\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}} \operatorname{cor}\left(\mathbf{y}\right) \widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}$.*

(iv) $\frac{\partial \delta_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\left(\widetilde{\delta}_j - 1\right)\widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii}} \leq 0.$

The eigenvalues of $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ are the most straightforward to discuss. Assuming they are all simple (which generally is the case for a suboptimal proposal), as $\widetilde{\Theta}_{ii}$ increases holding the other diagonal elements of $\widetilde{\boldsymbol{\Theta}}$ constant, the following necessarily occurs: at least one above-average eigenvalue of $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ *increases* while none decrease, at least one below-average eigenvalue of $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ *decreases* while none increase, and all the changes in $\widetilde{\boldsymbol{\pi}}$ are *zero-sum* in light of the fact that $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ is a correlation matrix for any admissible $\widetilde{\boldsymbol{\Theta}}$. Thus, increasing $\widetilde{\Theta}_{ii}$ at the margin makes $\widetilde{\boldsymbol{\pi}}$ "more dispersed".

This dispersion idea can be formalized with the concept of majorization. A vector $\mathbf{z}$ majorizes another vector $\mathbf{z}^* \neq \mathbf{z}$ iff $\sum_{j=1}^{k} z_j \geq \sum_{j=1}^{k} z_j^* \, \forall k < n$ when both $\mathbf{z}$ and $\mathbf{z}^*$ are in non-increasing order and $\sum_{j=1}^{n} z_j = \sum_{j=1}^{n} z_j^*$. Since both $\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ and $\boldsymbol{\pi}\left(\boldsymbol{\Theta}^*\right)$ are ordered appropriately and necessarily sum to n, one may majorize the other or perhaps neither does. However, a marginal increase in $\widetilde{\Theta}_{ii}$ would imply majorization of $\widetilde{\boldsymbol{\pi}}$ in light of the derivatives above. The eigenvalues of $\mathbf{A}^*\left(\widetilde{\boldsymbol{\Theta}}\right)$, $\boldsymbol{\Phi}\left(\widetilde{\boldsymbol{\Theta}}\right) - \mathbf{I}$, and $\boldsymbol{\Delta}\left(\widetilde{\boldsymbol{\Theta}}\right)$ are ordered but do not have constant sums. Thus, majorization is not a relevant concept until we divide their eigenvalues by their respective mean eigenvalues so that the cumulative sum is normalized to n in all four cases. Hence, the eigenvalues of $\frac{1}{\overline{\alpha}^*\left(\widetilde{\boldsymbol{\Theta}}\right)}\mathbf{A}^*\left(\widetilde{\boldsymbol{\Theta}}\right)$, $\frac{1}{\overline{\phi}\left(\widetilde{\boldsymbol{\Theta}}\right) - 1}\left(\boldsymbol{\Phi}\left(\widetilde{\boldsymbol{\Theta}}\right) - \mathbf{I}\right)$, and $\frac{1}{\overline{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}\boldsymbol{\Delta}\left(\widetilde{\boldsymbol{\Theta}}\right)$ all sum to n and can be compared on the majorization metric to the eigenvalues implied by $\boldsymbol{\Theta}^*$, as can the eigenvalues of $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$.

possible to define $\widetilde{\mathbf{x}} = \mathbf{x}\left(\widetilde{\boldsymbol{\Theta}}\right)$ as the diagonal of $\mathbf{D}\left(\widetilde{\boldsymbol{\Theta}}\right)$ instead of as the eigenvalues of $\mathbf{X}\left(\widetilde{\boldsymbol{\Theta}}\right)$. The theoretical properties of the IRMA would be the same either way, so the question is whether the defining $\widetilde{\mathbf{x}} = \mathbf{x}\left(\widetilde{\boldsymbol{\Theta}}\right)$ as the diagonal of $\mathbf{D}\left(\widetilde{\boldsymbol{\Theta}}\right)$ has any practical advantages. This question is currently under evaluation, and the preliminary evidence suggests that it does.

Majorization is important in this paper because $\mathbf{x}$ has $n - r$ trailing zeros and should majorize *most* admissible proposals for it, which have fewer trailing zeros but the same sum. In other words, an optimization algorithm could perhaps move in the general direction of a minimum-rank solution by maximizing a Schur-convex function, which is defined by the property that $f(\mathbf{z}) > f(\mathbf{z}^*)$ when $\mathbf{z}$ majorizes $\mathbf{z}^*$. Also, $f(\mathbf{z})$ is said to be a symmetric function if it is invariant to a permutation of the order of the n elements of $\mathbf{z}$. Symmetry is an essential property for a function of eigenvalues, whose order is substantively arbitrary.

In the literature on economic inequality, a function that is both Schur-convex and symmetric is said to satisfy the "transfer axiom" (see, for example, Foster and Shneyerov 1999), which requires that economic inequality must strictly increase under zero-sum transfers of money from poorer people to richer people. These "regressive" transfers are precisely what transpires among elements of $\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ as $\widetilde{\Theta}_{ii}$ increases holding the other diagonal elements of $\widetilde{\boldsymbol{\Theta}}$ constant, and the same is often true for the other matrices. Thus, long-standing results from the economics literature can point toward a suitable function of eigenvalues to optimize. The potential of this approach is immediately suggested by the following

Theorem 3. *If $\widehat{\boldsymbol{\Theta}} \in \mathcal{T}$ and $\frac{\mathbf{x}\left(\widehat{\boldsymbol{\Theta}}\right)}{\bar{x}\left(\widehat{\boldsymbol{\Theta}}\right)}$ majorizes $\frac{\mathbf{x}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{x}\left(\widetilde{\boldsymbol{\Theta}}\right)} \forall \widetilde{\boldsymbol{\Theta}} \in \mathcal{T} \neq \widehat{\boldsymbol{\Theta}}$, then $\widehat{\boldsymbol{\Theta}}$ is a minimum-rank solution that can be found by maximizing any Schur-convex, symmetric function of $\frac{\mathbf{x}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{x}\left(\widetilde{\boldsymbol{\Theta}}\right)}$.*

Proof. Assume $\widehat{\boldsymbol{\Theta}} \in \mathcal{T}$ is not a minimum-rank solution. If so, then $\sum_{j=1}^{r} \frac{x_j\left(\widehat{\boldsymbol{\Theta}}\right)}{\bar{x}\left(\widehat{\boldsymbol{\Theta}}\right)} < n$, in which case $\frac{\mathbf{x}\left(\widehat{\boldsymbol{\Theta}}\right)}{\bar{x}\left(\widehat{\boldsymbol{\Theta}}\right)}$ would not majorize $\frac{\mathbf{x}}{\bar{x}}$, which contradicts the premise that $\frac{\mathbf{x}\left(\widehat{\boldsymbol{\Theta}}\right)}{\bar{x}\left(\widehat{\boldsymbol{\Theta}}\right)}$ majorizes $\frac{\mathbf{x}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{x}\left(\widetilde{\boldsymbol{\Theta}}\right)} \forall \widetilde{\boldsymbol{\Theta}} \in \mathcal{T} \neq \widehat{\boldsymbol{\Theta}}$. By definition, $\frac{\mathbf{x}\left(\widehat{\boldsymbol{\Theta}}\right)}{\bar{x}\left(\widehat{\boldsymbol{\Theta}}\right)}$ is superior to all other admissible proposals with respect to a Schur-convex function. □

Corollary 4. *If $r = 1$, then $\frac{\mathbf{x}}{\bar{x}}$ majorizes all other admissible proposals for it.*

Proof. If $r = 1$, then $\frac{x_1}{\bar{x}} = n \iff x_j = 0 \,\forall j > 1$, so $\frac{\mathbf{x}}{\bar{x}}$ majorizes all other admissible proposals for it. □

Hence, if $\boldsymbol{\Sigma}$ were available and $r = 1$, then a minimum-rank solution could be found by maximizing *any* Schur-convex, symmetric function of the eigenvalues of any of the four scale-invariant matrices. Of course, if $r = 1$, $\boldsymbol{\Sigma}$ were available, and all its elements were positive, $\boldsymbol{\Theta}$ could be found using the tetrad method (see Bekker and de Leeuw (1987) for a historical review and its theorem 1). Thus, the potential of these results depends on picking a "good" Schur-convex, symmetric function to maximize that will yield a minimum-rank solution under more general conditions, which the next subsection attempts to do with some help from the literature on economic inequality.

5.3 The Generalized Entropy Dispersion Function

The generalized entropy dispersion function is Schur-convex, symmetric, non-negative, and defined as

$$D_c\left(\widetilde{\mathbf{x}}\right) = \begin{cases} \frac{1}{n}\sum_{j=1}^{n}\frac{x_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\overline{x}\left(\widetilde{\boldsymbol{\Theta}}\right)}\ln\left(\frac{x_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\overline{x}\left(\widetilde{\boldsymbol{\Theta}}\right)}\right) & \text{as } c\rightarrow 1 \text{ (Theil Index)} \\ \frac{1}{c(c-1)n}\sum_{j=1}^{n}\left[\left(\frac{x_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\overline{x}\left(\widetilde{\boldsymbol{\Theta}}\right)}\right)^{c}-1\right] & \text{if } c\in(0,1) \\ \frac{1}{n}\sum_{j=1}^{n}\ln\left(\frac{\overline{x}\left(\widetilde{\boldsymbol{\Theta}}\right)}{x_j\left(\widetilde{\boldsymbol{\Theta}}\right)}\right) & \text{as } c\rightarrow 0 \text{ (second Theil Index),} \end{cases}$$

where c is a tuning constant set *in advance* by the researcher that controls the relative sensitivity of $D_c\left(\widetilde{\mathbf{x}}\right)$ to different sized eigenvalues. The choice of c must be justified as much as possible, but reasonable values of c tend to yield very similar results. We reject $c \leq 0$ because $D_c\left(\widetilde{\mathbf{x}}\right)$ would be infinite whenever $x_n = 0$, which is inappropriate when seeking to maximize the number of null eigenvalues. It may appear as if $D_c\left(\widetilde{\mathbf{x}}\right)$ would also be (negatively) infinite whenever $\widetilde{x}_j = 0$ if $c = 1$, but $\frac{\widetilde{x}_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\overline{x}\left(\widetilde{\boldsymbol{\Theta}}\right)}\ln\left(\frac{\widetilde{x}_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\overline{x}\left(\widetilde{\boldsymbol{\Theta}}\right)}\right) \rightarrow 0^-$ as $\widetilde{x}_j \rightarrow 0^+$, so $0\ln(0)$ is always defined to be zero. The economics literature usually rejects $c > 1$ because it would paradoxically make $D_c\left(\widetilde{\mathbf{x}}\right)$ more sensitive to inequality among those with above-average incomes.

When $c = 1$, $D_1\left(\widetilde{\mathbf{x}}\right)$ is the Theil (1967) index, which is well-known in the income inequality literature, and when $c = 0$, $D_0\left(\widetilde{\mathbf{x}}\right)$ is the "second" Theil (1967) index. There are several axiomatic approaches to defining an income dispersion measure in the economics literature that eliminate all but the generalized entropy function. In particular, Shorrocks (1984) shows that $D_c\left(\widetilde{\mathbf{x}}\right)$ is the only continuous, additively decomposable, "relative" function of $\mathbf{x}$ that outputs zero iff all the elements of $\widetilde{\mathbf{x}}$ are equal. "Relative" has a specific, technical meaning that goes slightly beyond Schur-convexity and symmetry but one that accords with intuition, given the ratios involved in $D_c\left(\widetilde{\mathbf{x}}\right)$. Several properties of $D_c\left(\widetilde{\mathbf{x}}\right)$ are summarized in Cowell (2008, Appendix A). In particular, if $c > 0$, then the upper bound of $D_c\left(\widetilde{\mathbf{x}}\right)$ is $\frac{n^{c-1}-1}{c(c-1)}$ — which approaches $\ln(n)$ from the right as $c \rightarrow 1$ — and this upper bound is reached iff $\widetilde{x}_j = 0\,\forall j > 1$, i.e. the $r = 1$ case, as in corollary 4. Importantly, this indirect rank-minimization result can be extended to the $r \geq 1$ case.

Theorem 5. *If $c \in (0, 1]$ is small enough,* $\arg\max_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_c\left(\widetilde{\mathbf{x}}\right)\middle|\,\boldsymbol{\Sigma}\right\} = \arg\min_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{\mathcal{R}\left(\boldsymbol{\Sigma}-\boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}\right)\right\}.$

Proof. Let $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$ imply $\widetilde{\mathbf{x}}$ has $k > r$ positive eigenvalues, while $\widehat{\mathbf{\Theta}} \in \mathcal{T}$ implies $\widehat{\mathbf{x}}$ has maximum eigenvalue dispersion among minimum-rank solutions. The theorem will be proven if $\exists c \in (0,1] : D_c(\widehat{\mathbf{x}}) > D_c(\widetilde{\mathbf{x}})$ for every admissible $\widetilde{\mathbf{\Theta}}$ that is not a minimum-rank solution. By straightforward manipulation,

$$\begin{aligned}
D_c(\widehat{\mathbf{x}}) = \frac{1}{c(c-1)n} \sum_{j=1}^{n} \left[\left(\frac{x_j\left(\widehat{\mathbf{\Theta}}\right)}{\overline{x}\left(\widehat{\mathbf{\Theta}}\right)} \right)^c - 1 \right] &\gtreqless \frac{1}{c(c-1)n} \sum_{j=1}^{n} \left[\left(\frac{x_j\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \right)^c - 1 \right] = D_c(\widetilde{\mathbf{x}}), \\
\frac{1}{c(c-1)n} \sum_{j=1}^{r} \left(\frac{x_j\left(\widehat{\mathbf{\Theta}}\right)}{\overline{x}\left(\widehat{\mathbf{\Theta}}\right)} \right)^c - \frac{n}{c(c-1)n} &\gtreqless \frac{1}{c(c-1)n} \sum_{j=1}^{k} \left(\frac{x_j\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \right)^c - \frac{n}{c(c-1)n}, \\
\frac{1}{c(c-1)n} \left(\sum_{j=1}^{r} \left(\frac{x_j\left(\widehat{\mathbf{\Theta}}\right)}{\overline{x}\left(\widehat{\mathbf{\Theta}}\right)} \right)^c - \sum_{j=1}^{k} \left(\frac{x_j\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \right)^c \right) &\gtreqless 0, \\
\frac{1}{c(1-c)n} \left(\sum_{j=1}^{k} \left(\frac{x_j\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \right)^c - \sum_{j=1}^{r} \left(\frac{x_j\left(\widehat{\mathbf{\Theta}}\right)}{\overline{x}\left(\widehat{\mathbf{\Theta}}\right)} \right)^c \right) &\gtreqless 0.
\end{aligned}$$

As $c \to 0^+$, the left-hand side approaches $\frac{k-r}{0} = \infty$, so $\arg\max\limits_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{ D_c(\widetilde{\mathbf{x}}) | \mathbf{\Sigma} \}$ is a minimum-rank solution as $c \to 0^+$. Thus, a sufficiently small value of $c \in (0,1]$ always exists. □

As $c \to 0^+$, $\arg\max\limits_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{ D_c(\widetilde{\mathbf{x}}) | \mathbf{\Sigma} \}$ and $\arg\min\limits_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ \mathcal{R}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega} \right) \right\}$ are essentially the same problem, but $c > 0$ implies that $D_c(\widetilde{\mathbf{x}})$ is smooth and thereby somewhat amenable to numeric optimization. However, any particular value of c may be too large to imply that $\arg\max\limits_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{ D_c(\widetilde{\mathbf{x}}) | \mathbf{\Sigma} \}$ is a minimum-rank solution. The critical value of c depends on $\mathbf{\Sigma}$, but some further progress can be made analytically.

Lemma 6. *If there are k positive elements in $\widetilde{\mathbf{x}}$, $D_1(\widetilde{\mathbf{x}}) = D_1(\widetilde{x}_1 \ldots \widetilde{x}_k) + \ln\left(\frac{n}{k}\right)$.*

Proof. Recall that $0 \ln 0 = 0$. Thus, when calculating $D_1(\widetilde{\mathbf{x}})$, it is only necessary to sum over the k positive eigenvalues, which is to say that $D_1(\widetilde{\mathbf{x}}) = \frac{1}{n} \sum_{j=1}^{n} \frac{\widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \ln\left(\frac{\widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \right) = \frac{1}{n} \sum_{j=1}^{k} \frac{\widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \ln\left(\frac{\widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \right)$. $D_c(\widetilde{\mathbf{x}})$ is clearly invariant to a rescaling of all the eigenvalues by a constant, b. If $b = \frac{n}{k}$, then $D_1(\widetilde{\mathbf{x}}) = \frac{1}{n} \sum_{j=1}^{k} \frac{\frac{n}{k}\widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\frac{n}{k}\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \ln\left(\frac{\frac{n}{k}\widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\frac{n}{k}\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \right) = \frac{1}{k} \sum_{j=1}^{k} \frac{\widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\frac{n}{k}\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \left(\ln\left(\frac{\widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\frac{n}{k}\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} \right) + \ln\left(\frac{n}{k}\right) \right)$. Next, note that $\frac{n}{k}\overline{x}\left(\widetilde{\mathbf{\Theta}}\right) = \frac{1}{k} \sum_{j=1}^{k} \widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)$, which is the mean over the k positive eigenvalues. Thus, $D_1(\widetilde{\mathbf{x}}) = D_1(\widetilde{x}_1 \ldots \widetilde{x}_k) + \ln\left(\frac{n}{k}\right) \frac{1}{k} \sum_{j=1}^{k} \frac{\widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\frac{n}{k}\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} = D_1(\widetilde{x}_1 \ldots \widetilde{x}_k) + \ln\left(\frac{n}{k}\right) \frac{1}{n} \sum_{j=1}^{k} \frac{\widetilde{x}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{x}\left(\widetilde{\mathbf{\Theta}}\right)} = D_1(\widetilde{x}_1 \ldots \widetilde{x}_k) + \ln\left(\frac{n}{k}\right)$. □

This lemma implies that if $\arg\max\limits_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{ D_1(\widetilde{\mathbf{x}}) | \mathbf{\Sigma} \}$ is a minimum-rank solution, then it is the minimum-rank solution with maximum $D_1(\widetilde{x}_1 \ldots \widetilde{x}_r)$, which is somewhat appealing. A sufficient condition for

$\arg\max\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_1(\widetilde{\mathbf{x}})\middle|\,\boldsymbol{\Sigma}\right\}$ to be a minimum-rank solution is given by the following

Theorem 7. *Let $\{\boldsymbol{\Sigma}_n\}$ be a sequence of covariance matrices indexed by increasing n. As $\frac{n}{r}\to\infty$, the sequence of maxima, $\left\{\arg\max\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_1(\widetilde{\mathbf{x}})\middle|\,\boldsymbol{\Sigma}_n\right\}\right\}$, converges to* $\operatorname{diag}\left(\operatorname{cor}(\mathbf{y})_n^{-1}\right)^{-1}$.

Proof. According to lemma 6, in the population, $D_1(\mathbf{x}) = D_1(x_1\ldots x_r) + \ln\left(\frac{n}{r}\right)$, implying $\ln\left(\frac{n}{r}\right) \le D_1(\mathbf{x}) \le \ln(n)$ because $\ln(n)$ is the upper bound for the Theil index. As $\frac{n}{r}\to\infty$, the middle expression in $\ln\left(\frac{n}{r}\right) \le D_1(x_1\ldots x_r) + \ln\left(\frac{n}{r}\right) \le \ln(n)$ gets squeezed, implying $\lim\limits_{\frac{n}{r}\to\infty} D_1(x_1\ldots x_r) = 0$. Since $\{D_1(\mathbf{x})\}$ converges to its upper bound as $\frac{n}{r}\to\infty$, the probability that any proposal which is not a minimum-rank solution is superior with respect to $D_c(\widetilde{\mathbf{x}})$ vanishes. Thus, $\left\{\arg\max\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_1(\widetilde{\mathbf{x}})\middle|\,\boldsymbol{\Sigma}_n\right\}\right\}$ converges to a minimum-rank solution, and this minimum-rank solution is proven to be $\operatorname{diag}\left(\operatorname{cor}(\mathbf{y})_n^{-1}\right)^{-1}$ by Guttman (1956) and Krijnen (2006). □

However, theorem 7 merely pushes back the question of "what value of c is sufficiently small" to "what value of $\frac{n}{r}$ is sufficiently large to render $c = 1$ sufficiently small?" If $r = 1$, then theorem 4 implies $c = 1$ is small enough, but in general this question must be answered with simulations.

5.4 Sampling Behavior of the IRMA

Theorem 5 shows that if c is sufficiently small, then $\arg\max\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_c(\widetilde{\mathbf{x}})\middle|\,\boldsymbol{\Sigma}\right\}$ is a minimum-rank solution. Instead, $\mathbf{S}$ is observed, which is an estimate of $\boldsymbol{\Sigma}$ based on a sample of size N, and thus every cell of $\mathbf{S}$ is afflicted with random sampling variation, which destroys the exact linear relationships that exist among the columns of $\boldsymbol{\Sigma} - \boldsymbol{\Omega\Theta\Omega}$. When referring to a sample, we reinterpret $\mathcal{T}$ to be the set of diagonal proposals for $\boldsymbol{\Theta}$ such that $\mathbf{S} - \widehat{\boldsymbol{\Omega}}\widetilde{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Omega}}$ is PSD and $0 < \widetilde{\Theta}_{ii} < 1\,\forall i$. The continuity of $D_c(\widetilde{\mathbf{x}})$ implies the following

Theorem 8. *If $c\in(0,1]$ is sufficiently small and* $\operatorname{plim}\mathbf{S} = \boldsymbol{\Sigma}$, *then* $\operatorname{plim}\,\arg\max\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_c(\widetilde{\mathbf{x}})\middle|\,\mathbf{S}\right\} = \widehat{\boldsymbol{\Theta}}$*, such that $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widehat{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ has minimum-rank. If, in addition, $r < L(n)$, then* $\operatorname{plim}\,\arg\max\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_c(\widetilde{\mathbf{x}})\middle|\,\mathbf{S}\right\} = \boldsymbol{\Theta}$*, which is to say that $\boldsymbol{\Theta}$ is estimated consistently.*

Proof. The result follows from the Slutsky theorem, which says that if $\mathbf{V}_N$ is a random variable and $g(\mathbf{V}_N)$ is a continuous function that does not depend on N, then $\operatorname{plim} g(\mathbf{V}_N) = g(\operatorname{plim}\mathbf{V}_N)$. Let $\mathbf{V}_N = \mathbf{S}$ and let $g()$ be $\arg\max\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_c(\mathbf{S})\middle|\,\boldsymbol{\Sigma}\right\}$. If $\operatorname{plim}\mathbf{S} = \boldsymbol{\Sigma}$, then $\operatorname{plim} g(\mathbf{S}) = \arg\max\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_c(\widetilde{\mathbf{x}})\middle|\,\boldsymbol{\Sigma}\right\}$ which theorem

5 shows is a minimum-rank solution if c is sufficiently small and theorem 1 shows is almost surely unique if $r < L(n)$. □

If $r \geq L(n)$, then $\mathbf{\Theta}$ is not identified, implying the IRMA cannot consistently estimate $\mathbf{\Theta}$. However, if c is small enough, the difference between $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{ D_c(\widetilde{\mathbf{x}}) | \mathbf{S}\}$ and a minimum-rank solution in the population vanishes as $N \to \infty$, permitting inferences about r despite the nonidentification of $\mathbf{\Theta}$.

The asymptotic sampling distribution of $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{ D_c(\widetilde{\mathbf{x}}) | \mathbf{S}\}$ could be derived using the results in Shapiro and ten Berge (2002). In short, it is asymptotically normally distributed and its sampling variance is $\mathcal{O}(N^{-1})$ but depends on the fourth-order moment matrix of the data. However, some closed-form results that Shapiro and ten Berge (2002) obtains for MRFA have not yet been established for the IRMA, largely due to the fact that $D_c(\widetilde{\mathbf{x}})$ depends both on c and nonlinear transformations to induce scale-invariance. If the researcher is interested in the sampling distribution of $\widehat{\mathbf{\Theta}}$, it would probably be best to use nonparametric bootstrapping, which is supported in FA*i*R.

6 Monte Carlo Simulations

The previous section shows that there are circumstances in which $\widehat{\mathbf{\Theta}} = \arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{ D_c(\widetilde{\mathbf{x}}) | \mathbf{\Sigma}\}$ is a minimum-rank solution and asymptotic conditions under which this result extends to samples. But aside from the $r = 1$ case and the case where $\frac{n}{r} \to \infty$, we do not know analytically when c is sufficiently small to obtain a minimum-rank solution in the population. Nor do we know what value of N is sufficiently large to render sample approximations viable. To answer these questions, we have to conduct Monte Carlo simulations.

As described in Appendix B, the IRMA as implemented in FA*i*R can be executed with a "fast" engine (a hacked simplex algorithm) or a "slow" engine (a genetic algorithm). In any particular situation, one may produce a higher value of $D_c(\widehat{\mathbf{x}})$ than the other, so both should be executed. In the all simulations in this paper, we use both the fast and slow engines and retain the result that yields more eigenvalue dispersion. Although the fast engine would probably be adequate for most simulations or bootstrapping, the slow engine produces more eigenvalue dispersion more often than not.

6.1 Simulations with Random but Structured Populations

The first task is to assess how well $\widehat{\boldsymbol{\Theta}} = \arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_c\left(\widetilde{\mathbf{x}}\right) \middle| \boldsymbol{\Sigma} \right\}$ corresponds to a minimum-rank solution. Davies and Higham (2000) presents an algorithm that turns out to have profound implications for simulations of LISREL models. Simply set n and r to desired values and repeat the following steps:

1. Set or randomly draw $\boldsymbol{\pi}$, such that it has r positive values that sum to n and $n - r$ null values.

2. Use Davies and Higham's (2000) algorithm to randomly draw $\boldsymbol{\Pi}$ from the set of correlation matrices whose eigenvalues are *exactly* $\boldsymbol{\pi}$ (hence a LISREL model with r latents fits perfectly).

3. Set or randomly draw $\Theta_{ii} \in (0, 1)\ \forall i$.

4. Find $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_c \left(\mathrm{cor}\left(\mathbf{y}\right) = \left(\mathbf{I} - \boldsymbol{\Theta}\right)^{\frac{1}{2}} \boldsymbol{\Pi} \left(\mathbf{I} - \boldsymbol{\Theta}\right)^{\frac{1}{2}} + \boldsymbol{\Theta} \right) \middle| \boldsymbol{\Sigma} \right\}$

In other words, n, r, $\boldsymbol{\pi}$, $\boldsymbol{\Theta}$ and c can be experimentally manipulated to evaluate the IRMA (or any other scale-invariant algorithm) *in the population* for each of the four scale-invariant ways to operationalize $\widetilde{\mathbf{x}}$. This subsection conducts 500 simulations for each combination of n and r.

The critical question is how to execute step 1. To avoid the appearance of tainted results, we chose to randomly draw the positive eigenvalues uniformally from the r-dimensional unit simplex (multiplied by n) in *each* simulation.[3] This procedure is conservative in the sense that actual researchers select their manifest variables non-randomly in an attempt to make π_r reasonably large. In these simulations, π_r can be arbitrarily close to zero, which is a very daunting challenge for the IRMA. Thus, on average, the IRMA in actual research situations where $\pi_r \gg 0$ should be better than in these simulations.

A more mundane question is how best to execute step 3. We chose to randomly draw each Θ_{ii} uniformally from the $[0.1, 0.9]$ interval in *each* simulation. Thus, the average communality in these simulations is 0.5. An alternative not pursued here would be to draw each Θ_{ii} from a generalized Beta distribution with its parameters set to auspicious values to further analyze how the IRMA is affected by $\boldsymbol{\Theta}$.[4] A preliminary analysis suggests that optimization with respect to $\widetilde{\boldsymbol{\delta}}$ is more sensitive to $\boldsymbol{\Theta}$ than are the other choices. Dividing by near-zero quantities is hazardous, so if $\Theta_{ii} \approx 0$, optimization with respect to $\widetilde{\boldsymbol{\phi}} - \mathbf{1}$ or $\widetilde{\boldsymbol{\delta}}$ becomes more

[3] We have discovered a better way to draw a random, but positive semi-definite $\boldsymbol{\Pi}$ with rank r, which is only a slight variation on the algorithm in Lewandowski, Kurowicka and Joe (2009). The algorithm is in FAiR 0.6-0 already and a paper is in progress.

[4] Another approach that should have been pursued here would be to take each Θ_{ii} from a sequence of quasi-random numbers

difficult but not fatal. Conversely, if $\Theta_{ii} \approx 1$, optimization with respect to $\widetilde{\boldsymbol{\pi}}$ or $\widetilde{\boldsymbol{\alpha}}^*$ is more difficult but not fatal.

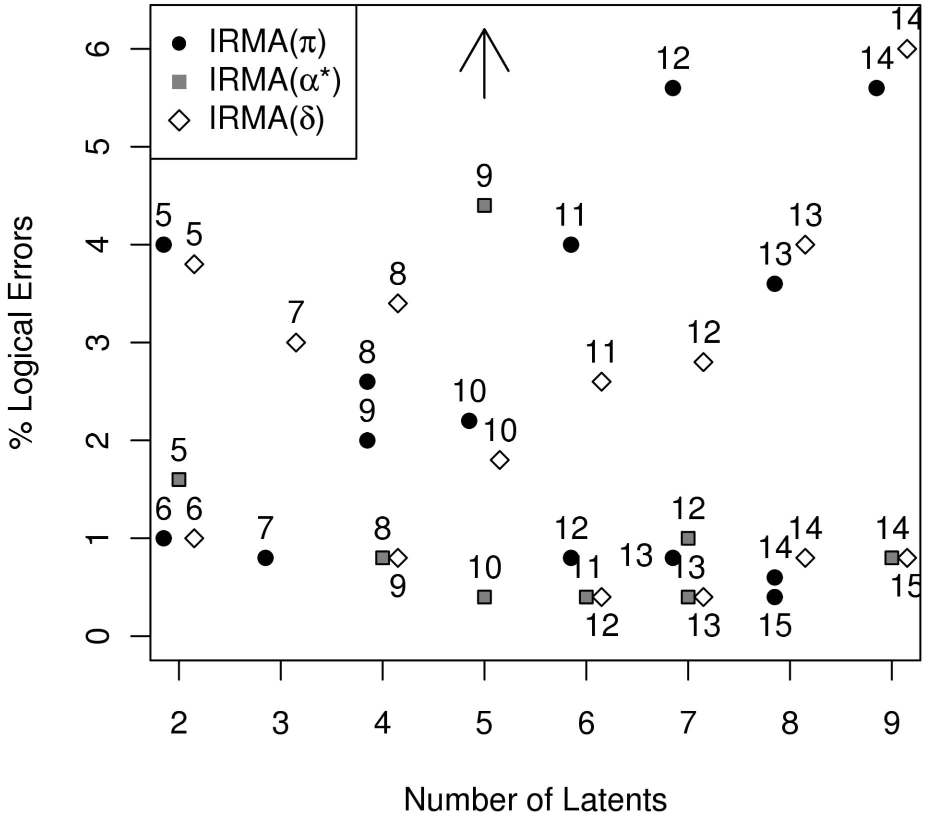
Since there are multiple minimum-rank solutions when $r \geq L(n)$, a metric is needed to quantify how close $\operatorname{cor}(\mathbf{y}) - \widehat{\boldsymbol{\Theta}}$ is to *some* minimum-rank solution without regard to whether $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}$. A good candidate is the ratio of the UCV to the total manifest variance, which is defined as $u\left(\widehat{\boldsymbol{\Theta}}\right) = \frac{1}{n}\sum_{j=r+1}^{n} \widehat{\lambda}_j \geq 0$, where $\widehat{\lambda}_j \geq 0$ is the jth largest eigenvalue of $\operatorname{cor}(\mathbf{y}) - \widehat{\boldsymbol{\Theta}}$. Iff $\widehat{\boldsymbol{\Theta}}$ is a minimum-rank solution, then $u\left(\widehat{\boldsymbol{\Theta}}\right) = 0$.

The first batch of simulations covers all values of $1 \leq r \leq 9$ and $4 \leq n \leq 19$ such that $r < L(n)$, implying there is a unique minimum-rank solution. If $r < L(n)$ and $D_c(\widehat{\mathbf{x}}) > D_c(\mathbf{x})$, then c is too large for theorem 5 to hold, which is classified as a "logical error" regardless of how inconsequential it is. Conversely, an "optimization error" occurs when $D_c(\widehat{\mathbf{x}}) < D_c(\mathbf{x})$, which is to say that c is small enough for theorem 5 to hold but neither engine got all the way to the maximum of $D_c(\widetilde{\mathbf{x}})$. Strictly speaking, success only occurs when $D_c(\widehat{\mathbf{x}}) = D_c(\mathbf{x})$, which is impossible with floating point numbers, but logical or optimization errors that are only evident past the first few decimal places are of little concern.

Corollary 4 implies that there can be no logical errors when $r = 1$, so $u\left(\widehat{\boldsymbol{\Theta}}\right)$ reflects optimization error only. The average $u\left(\widehat{\boldsymbol{\Theta}}\right)$ over the simulations where $r = 1 = c$ and $n \leq 19$ is on the order of 10^{-7} or 10^{-8}, depending on which scale-invariant matrix is utilized. Optimization error increases slightly as n increases. In the worst case, $u\left(\widehat{\boldsymbol{\Theta}}\right)$ is on the order of 10^{-5} or 10^{-6}, which is still effectively zero. Thus, the IRMA is capable of finding $\arg\max\limits_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_1(\widetilde{\mathbf{x}}) \middle| \boldsymbol{\Sigma} \right\}$ to a high degree of precision — at least when $r = 1$ — despite the difficulty of this optimization problem. Things are much more complicated when $r > 1$.

When $r > 1$, it is possible that $c = 1$ is too large for theorem 5 to hold. In fact, $c = 1$ is almost always too large when maximizing the dispersion of $\boldsymbol{\phi}\left(\widetilde{\boldsymbol{\Theta}}\right) - \mathbf{1}$. This result is presumably due to the lack of an upper bound on $\phi_1\left(\widetilde{\boldsymbol{\Theta}}\right)$. Henceforth, we focus on the other three ways to execute the IRMA, which will be denoted IRMA($\boldsymbol{\pi}$), IRMA($\boldsymbol{\alpha}^*$), and IRMA($\boldsymbol{\delta}$) for convenience with $c = 1$ unless explicitly stated otherwise. As shown in figure 1, logical errors typically occur in less than five percent of the simulations when $2 \leq r \leq 9$ and $5 \leq n \leq 19$ such that $r < L(n)$ and essentially only occur when $r \lessapprox L(n)$ (where $\lessapprox$ means "barely less than"). The logical failure rates for the IRMA($\boldsymbol{\pi}$) and IRMA($\boldsymbol{\delta}$) are considerably larger than the logical failure rate for IRMA($\boldsymbol{\alpha}^*$) in relative terms but all are small in absolute terms, except when $n = 9$ and $r = 5$, when the logical failure rate for the IRMA($\boldsymbol{\alpha}^*$) is 4.4% and in excess of 20% for the other

Figure 1: Logical Errors when $c = 1$ and $r \lessapprox L(n)$



Integers indicate the value of n (number of manifest variables) that the point corresponds to. A logical error occurs whenever the optimal value of the objective function exceeds the value of the objective function at the minimum-rank solution. Other combinations of $r \leq 9$ (number of latents) and $n \leq 19$ such that $r < L(n) = 0.5\left(2n + 1 - (8n+1)^{\frac{1}{2}}\right)$ yielded logical errors in at most 0.2% of the simulations and are not plotted. The other exception, which is indicated by the big arrow, is when $r = 5$ and $n = 9$, where the IRMA($\boldsymbol{\pi}$) and IRMA($\boldsymbol{\delta}$) respectively produced logical errors in 20.4% and 25.4% of simulations.

two.

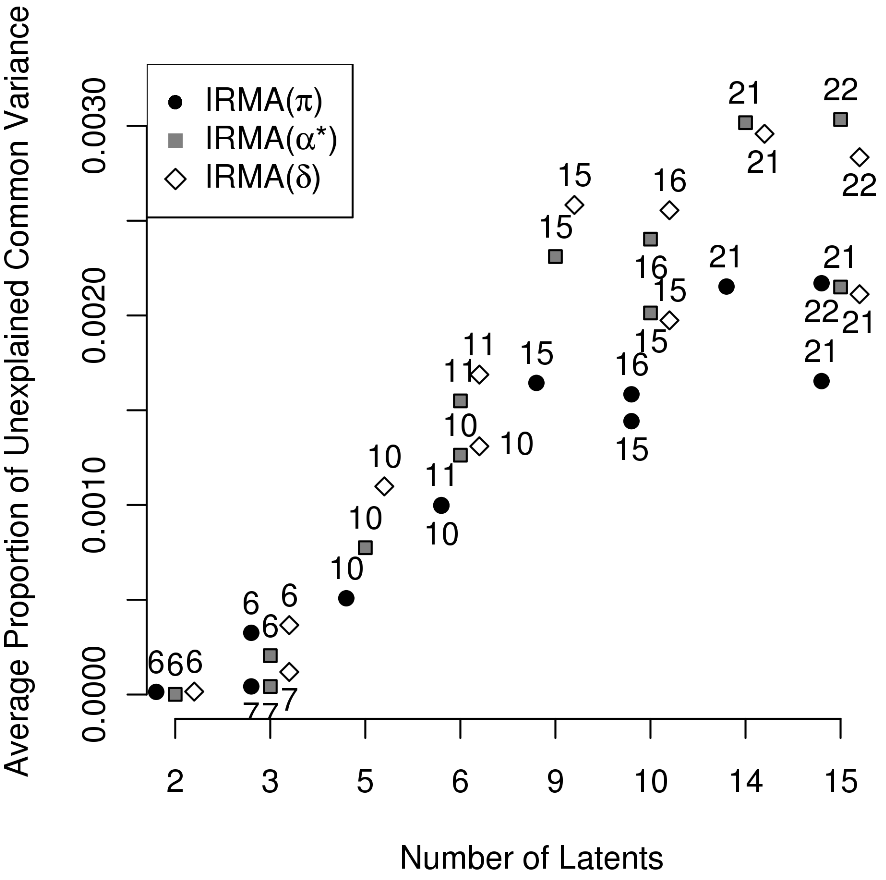
As n increases holding r constant, the probability of a logical error falls because the positive eigenvalues tend to be farther from zero. This trend can be quantified with a logistic regression (not shown) of the probability of a logical error as a function of x_r. Since logical errors are rare, the "rare event" bias in the estimated intercept is corrected using King and Zeng (2001), implying that $\widehat{\Pr}\left(D_c\left(\widehat{\mathbf{x}}\right) > D_c\left(\mathbf{x}\right)\right) = \left(1 + \exp\left(-\widehat{\beta_0} - x_r\widehat{\beta_1}\right)\right)^{-1}$ can be estimated for any value of the corresponding x_r. For example, if $n = 8$, $r = 4$, and $\delta_r = 0.25$, the estimated probability that $D_c\left(\widehat{\boldsymbol{\delta}}\right) > D_c\left(\boldsymbol{\delta}\right)$ is about 0.14, whereas the probability falls to about 0.02 when $\delta_r = 0.5$.

When a logical error occurs, it tends to be the case that x_r is small. Thus, it is not surprising that logical errors do not necessarily affect any relevant inference. When a logical error does occur, the average value of $u\left(\widehat{\boldsymbol{\Theta}}\right)$ tends to be about 0.001. In fact, optimization error is just as concerning as logical error. To assess "normal" optimization error, we calculate the median $u\left(\widehat{\boldsymbol{\Theta}}\right)$ over the simulations for each combination of n and r such that $1 < r < L\left(n\right)$ when $c = 1$. These medians are little influenced by logical errors or egregious optimization errors. When $r \lessapprox L\left(n\right)$, the median $u\left(\widehat{\boldsymbol{\Theta}}\right)$ is about 0.0005. In other words, when a logical error does occur, the $u\left(\widehat{\boldsymbol{\Theta}}\right)$ is roughly the same (in absolute terms) as the $u\left(\widehat{\boldsymbol{\Theta}}\right)$ when there is no logical error but normal optimization error. In general, optimization error is increasing in both r and n.

The few cases where $r = L\left(n\right)$ are interesting because the multiple minimum-rank solutions are locally unique. Logical and optimization errors cannot be distinguished because there may be *some* minimum-rank solution that implies more eigenvalue dispersion than does $\boldsymbol{\Theta}$, which is a problem due to lack of identification rather than a logical error. Nevertheless, logical errors are possible, albeit undetectable, because the IRMA could bypass all minimum-rank solutions in favor of a solution with more eigenvalue dispersion. In figure 2, it is clear that when $r = L\left(n\right)$, the IRMA gets about as close to a minimum-rank solution as when $r \lessapprox L\left(n\right)$. In short, whether $u\left(\widehat{\boldsymbol{\Theta}}\right)$ reflects logical or optimization error, it tends to be negligible.

Finally, simulations are conducted for all values of $2 \leq r \leq 9$ and $4 \leq n \leq 19$ such that $n - 1 > r > L\left(n\right)$ to judge how well the IRMA finds *some* minimum-rank solution when $c = 1$ but $\boldsymbol{\Theta}$ is not even locally identified. Again, logical and optimization errors cannot be distinguished, but $u\left(\widehat{\boldsymbol{\Theta}}\right)$ can be calculated. These simulations are not shown in a figure because the $u\left(\widehat{\boldsymbol{\Theta}}\right)$ values are simply tiny and become smaller as r becomes *farther* away from $L\left(n\right)$. In the worst case scenarios where $r \gtrapprox L\left(n\right)$, $\arg\max\limits_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_1\left(\widetilde{\mathbf{x}}\right) \middle| \boldsymbol{\Sigma}\right\}$

Figure 2: Average Proportion of Unexplained Common Variance when $r = L\left(n\right)$ as Compared to $r \lessapprox L\left(n\right)$



Integers indicate the value of n (number of manifest variables) that the point corresponds to. A minimum-rank solution implies the proportion of Unexplained Common Variance $\left(u\left(\widehat{\mathbf{\Theta}}\right)\right)$ is zero. The combinations of r (number of latents) and n such that $r = L\left(n\right) = 0.5\left(2n + 1 - \left(8n + 1\right)^{\frac{1}{2}}\right)$ are $(3, 6)$, $(6, 10)$, $(10, 15)$, and $(15, 21)$. Other plotted combinations of r and n imply $r \lessapprox L\left(n\right)$. All simulations use $c = 1$.

is usually a minimum-rank solution to at least two decimal places, as measured by $u\left(\widehat{\mathbf{\Theta}}\right)$.

6.2 Simulations with Random Samples from Random but Structured Populations

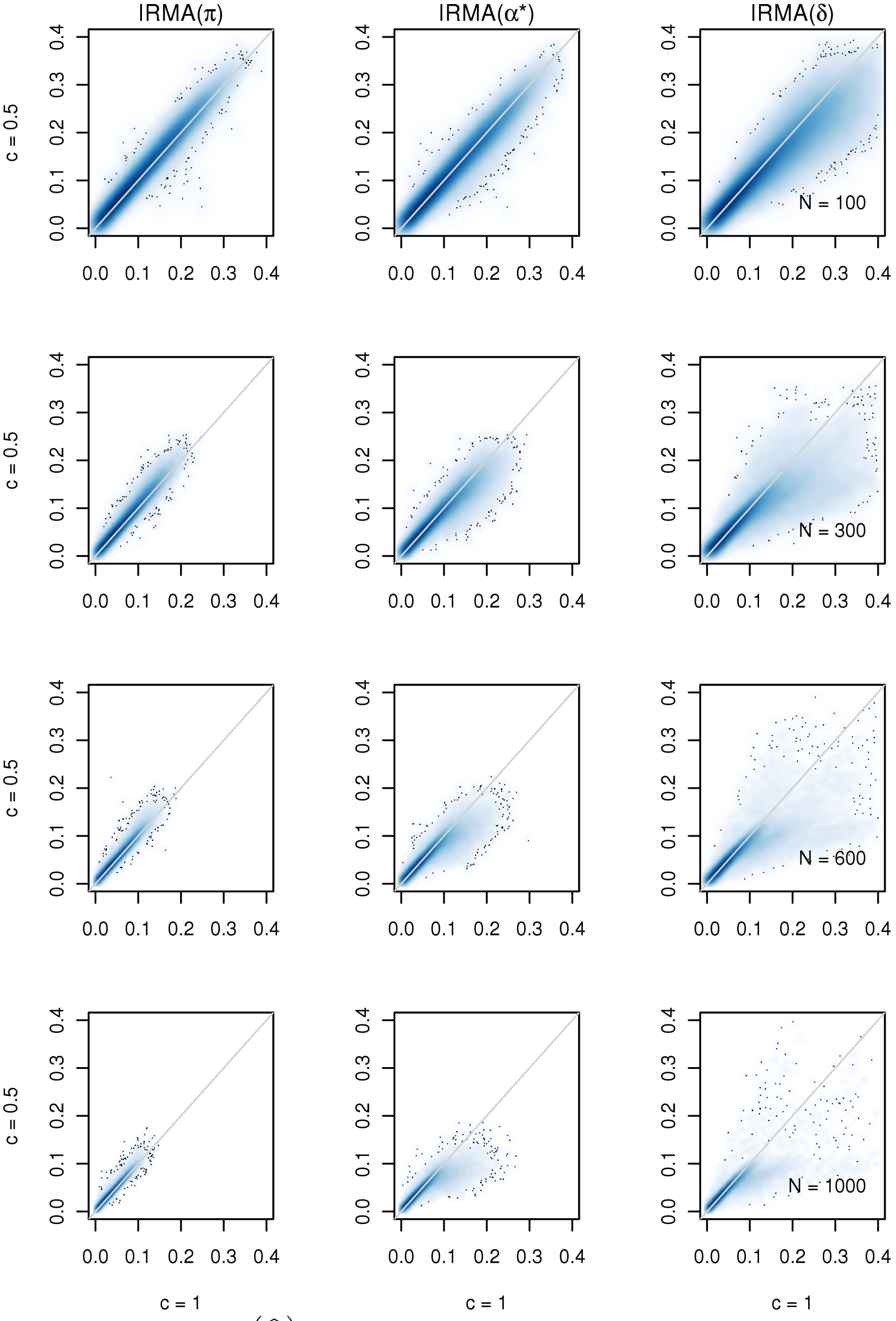
Given that the IRMA with $c = 1$ works well in the population, the next question is how well it works when only $\mathbf{S}$ is available. In theory, the finite-sample performance of the IRMA depends on N, so 500 simulations are conducted for each combination of $n \leq 19$, $r \leq 9$, and $N \in \{100, 300, 600, 1000\}$ such that $r < L(n)$.

The procedure for generating $\mathbf{S}$ is as follows. First, in each of the 500 simulations a random $\operatorname{cor}(\mathbf{y})$ is generated using the same algorithm as before. Then, $\mathbf{S}$ is drawn from a Wishart distribution with $N - 1$ degrees of freedom and expectation $\operatorname{cor}(\mathbf{y})$. This design implies that the manifest variables are multivariate normal, which is the usual assumption in the literature. However, no distributional assumptions were made in deriving the IRMA, so the multivariate normal assumption is used primarily for convenience. It also facilitates comparisons between the IRMA and the maximum likelihood (ML) exploratory factor analysis estimator, which assumes the data are multivariate normal and conditions on k factors.

Figure 3 plots values of $u\left(\widehat{\mathbf{\Theta}}\right)$ for the IRMA when $c = 1$ on the x-axis and values of $u\left(\widehat{\mathbf{\Theta}}\right)$ when $c = 0.5$ on the y-axis, pooling over all combinations of n and r such that $r < L(n)$. The three columns correspond to the three flavors of IRMA and the four rows correspond to the four sample sizes. Darkness indicates greater density and visible points on the edge of the cloud are "outliers" in the sense that they are far from the mode but more likely reflect egregious optimization error. As would be expected, as N increases, $u\left(\widehat{\mathbf{\Theta}}\right)$ decreases, as indicated by the southwest shifts in the point clouds. An unexpected result is that the $u\left(\widehat{\mathbf{\Theta}}\right)$ values for the IRMA($\boldsymbol{\pi}$) tend to be smaller than for the IRMA($\boldsymbol{\alpha}^*$) and IRMA($\boldsymbol{\delta}$), and the IRMA($\boldsymbol{\delta}$) is especially prone to very large values of $u\left(\widehat{\mathbf{\Theta}}\right)$. Also somewhat surprising is that whether $c = 1$ or $c = 0.5$ does not have a major effect on $u\left(\widehat{\mathbf{\Theta}}\right)$. For the IRMA($\boldsymbol{\pi}$), the results for $c = 1$ are slightly but consistently better, for the IRMA($\boldsymbol{\delta}$), the results for $c = 0.5$ are better in the right tail, and for the IRMA($\boldsymbol{\alpha}^*$) the results are virtually the same.

Figure 3 does not break down the results by n and r. Nor does it provide a comparison with the ML estimator because the ML estimator does not enforce the restriction that $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$, which is necessary for meaningful UCV values. The root mean-squared error (RMSE) in estimating $\mathbf{\Theta}$ could be used to compare the IRMA and the ML estimator, but such a comparison is not that instructive because the ML estimator

Figure 3: Comparison of Proportions of Unexplained Common Variance



Each (x, y) is a combination of $u\left(\widehat{\mathbf{\Theta}}\right)$, the proportion of Unexplained Common Variance, when $c = 1$ (x-axis) and $c = 0.5$ (y-axis) for the same $\mathbf{S}$. The grey lines correspond to $y = x$. Columns indicate the relevant version of the IRMA. Rows indicate different sample sizes. Each plot represents $48,500$ total simulations from random samples where $1 \leq r \leq 9$ and $n \leq 4 \leq 19$ such that $r < L\left(n\right)$.

conditions on $k = r$ while the IRMA is primarily intended to infer r without conditioning on it. A somewhat better (but still imperfect) comparison involves the RMSE with respect to the diagonal elements of $\mathbf{\Lambda \Upsilon \Lambda}$. In other words, we assume that the researcher automatically infers the correct value of r at the IRMA optimum and compare it to the communality-RMSE of the ML estimator.

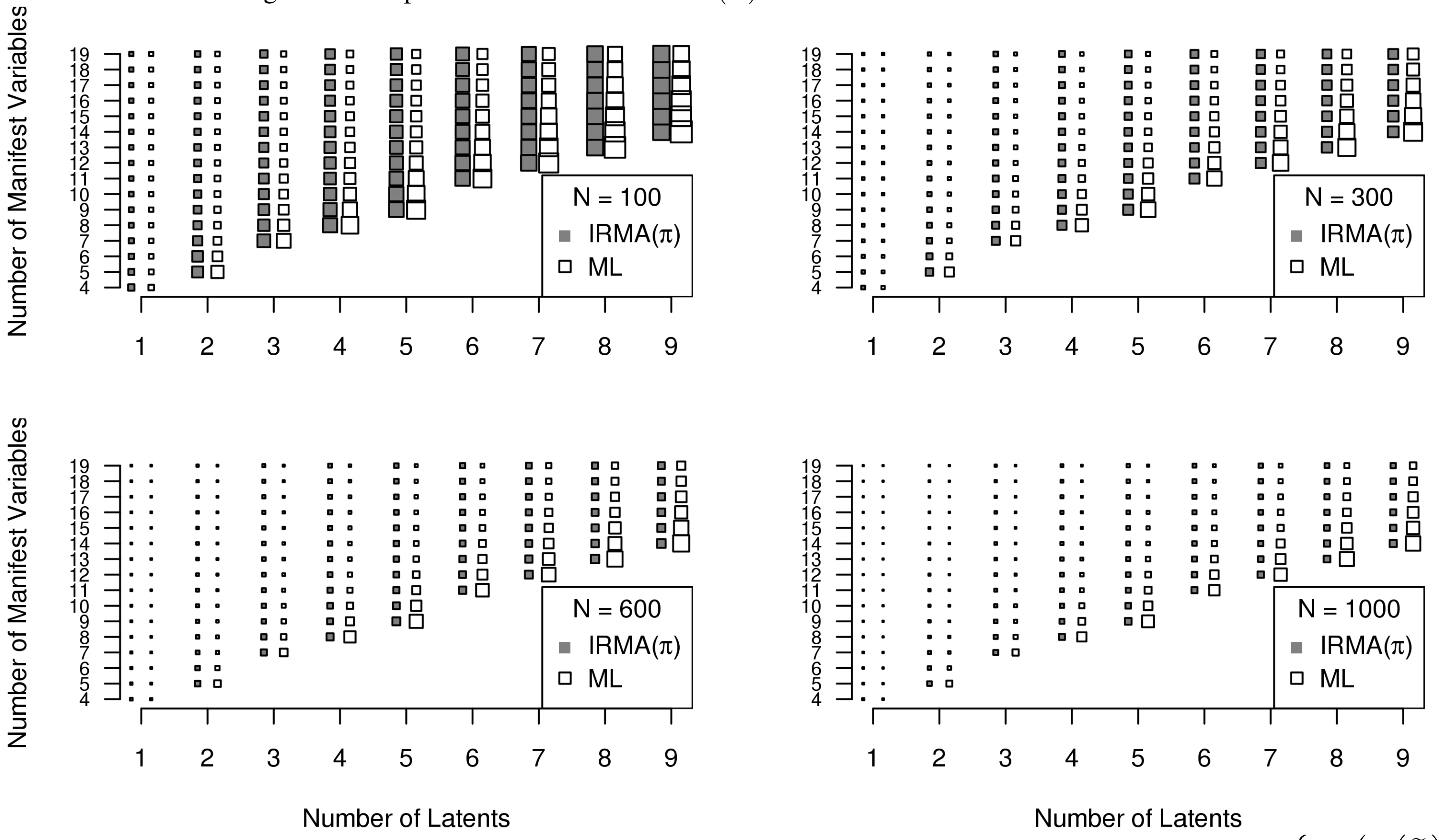
The main result in figure 4 is that the ML estimator outperforms IRMA($\boldsymbol{\pi}$) on this communality-RMSE criterion when r is much less than $L(n)$, and the IRMA($\boldsymbol{\pi}$) outperforms the ML estimator when r is barely less than $L(n)$. Of course, for both estimators, the RMSE becomes smaller as N increases, and the RMSE becomes smaller as r becomes smaller relative to n. The finite-sample performances of the IRMA($\boldsymbol{\alpha}^*$) and IRMA($\boldsymbol{\delta}$) (not shown) are somewhat worse than for the IRMA($\boldsymbol{\pi}$) on the RMSE criterion but are still sometimes better than the ML estimator if $r \lessapprox L(n)$.

Although there is no clear winner with respect to RMSE, the IRMA is more than competitive. Moreover, the RMSE is not the only relevant consideration for an estimator. The virtues of ML are well-known. The IRMA has the advantage that its estimates are admissible in the sense that both $\widehat{\mathbf{\Theta}}$ and $\mathbf{S} - \widehat{\mathbf{\Omega}}\widehat{\mathbf{\Theta}}\widehat{\mathbf{\Omega}}$ are PSD, while the latter matrix is indefinite for the ML estimator. There is also strong reason to believe the performance of the IRMA is more robust because it does not make distributional assumptions (although the ML estimator remains consistent when the multivariate normal assumption fails to hold). Thus, while one can make a case for either, the decisive factor should be whether the IRMA puts the researcher in a better position to infer r in finite samples, which is its primary purpose and is the sole topic of Goodrich (2009).

6.3 Simulations with Arbitrary Correlation Matrices and Three New Conjectures

The IRMA performs decently in a sample when the LISREL model holds in the population. The next question is how does the IRMA fare when the LISREL model is inappropriate. To answer this question, the IRMA is subjected to 3000 *uniform* draws from the set of correlation matrices of size n using the algorithm in Lewandowski, Kurowicka and Joe (2009). Since each correlation matrix of order n is as likely as any other, such a draw is referred to as an “arbitrary correlation matrix”. Unlike the algorithm based on the insight of Davies and Higham (2000), the minimum rank of arbitrary correlation matrices cannot be controlled, implying the LISREL model cannot hold for an arbitrary correlation matrix, although it could approximately hold to greater or lesser degrees by chance. To judge “approximate”, the UCV proportion is

Figure 4: Comparison of RMSE for IRMA($\boldsymbol{\pi}$) and Maximum Likelihood with r Latents



The root mean-squared error (RMSE) pertains to the diagonal of $\boldsymbol{\Lambda\Upsilon\Lambda}'$ when it is estimated with r latents at $\arg\max_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_1\left(\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\middle|\,\mathbf{S}\right\}$, or in the case of the maximum likelihood estimator, when conditioning on r latents. The edges of the squares are proportional to the RMSE. Thus, if $r \ll L\left(n\right) = 0.5\left(2n+1-\left(8n+1\right)^{\frac{1}{2}}\right)$, the ML estimator yields lower RMSE but if $r \lessapprox L\left(n\right)$, the IRMA($\boldsymbol{\pi}$) yields lower RMSE.

redefined as $u\left(\widehat{\boldsymbol{\Theta}}\right)=\frac{1}{n}\sum_{j=k+1}^{n}\widehat{\lambda}_j$ for a specified value of k retained factors.

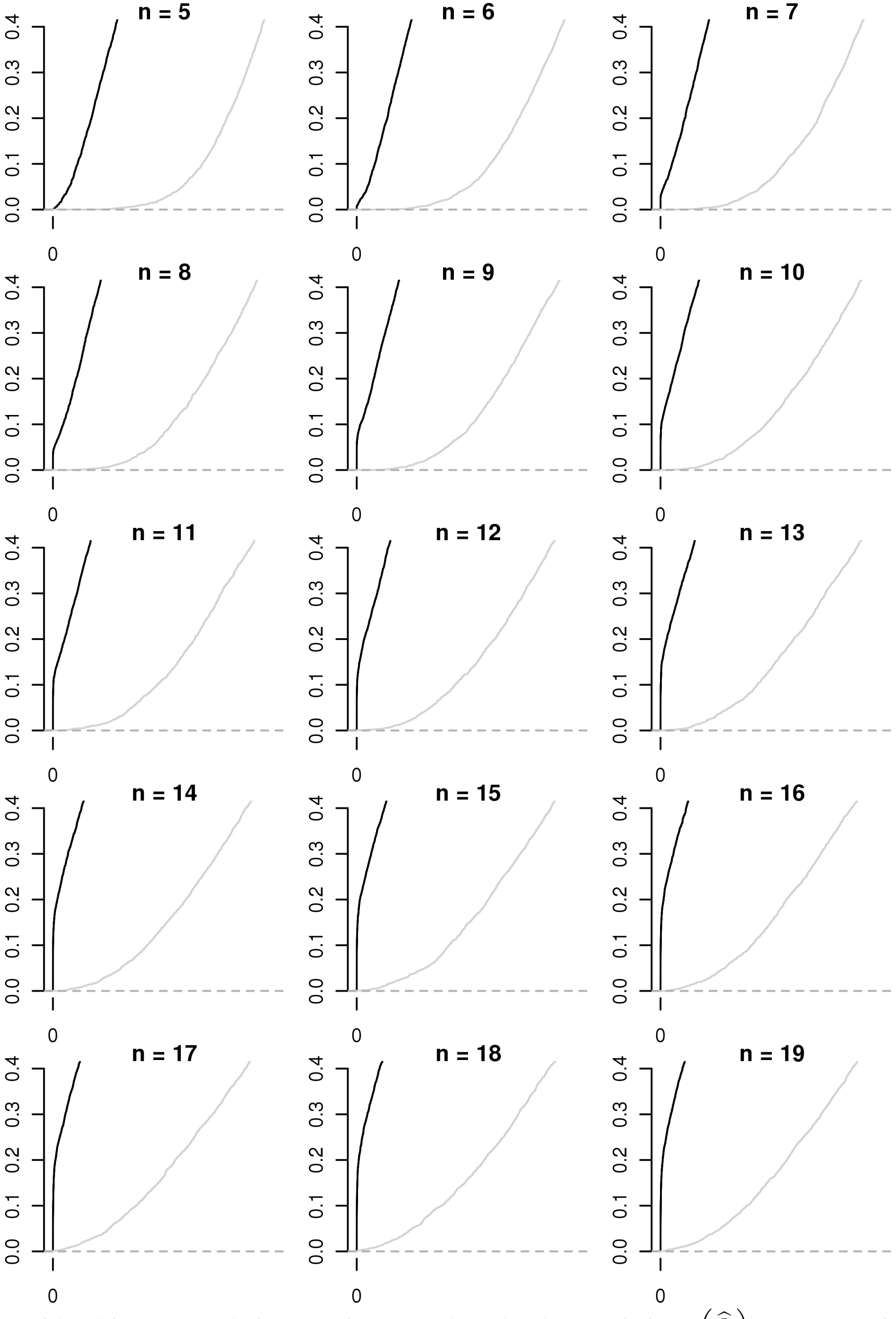
Guttman (1958) essentially asks: How far can the rank of an arbitrary correlation matrix be reduced when $\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}$?. Shapiro (1982) proves that the answer is "at best to $L(n)$" but does not utilize the constraint that $\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}$, so we might expect that the minimum-rank is larger when this constraint is imposed. Figure 5 takes up this question by plotting $u\left(\widehat{\boldsymbol{\Theta}}\right)$ values for the IRMA($\boldsymbol{\pi}$) in simulations with arbitrary correlation matrices of various sizes. The black lines trace the empirical CDF of $u\left(\widehat{\boldsymbol{\Theta}}\right)$ when $k=n-3$ factors are retained. The top left plot where $n=5$ is conceptually important. Shapiro's (1982) result implies that when $n=5$ the black line cannot literally touch zero on the x-axis, but we see that it can occasionally be quite close to zero. For example, the smallest value of $u\left(\widehat{\boldsymbol{\Theta}}\right)$ is 0.004 and five percent are less than 0.0291 (a result that is used in Goodrich (2009) as a critical value to reject the null hypothesis that $r\geq L(5)$). But the key point is that the top left plot provides a visual representation of Shapiro's (1982) main result: the proportion of covariances matrices among $n=5$ variables such that $\min\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{\mathcal{R}\left(\boldsymbol{\Sigma}-\boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}\right)\right\}=2$ is zero, which is indicated by the lack of mass at zero.

When $n=6$, the plot appears essentially the same, suggesting it is impossible to reduce the rank of an arbitrary correlation matrix among $n=6$ variables to three when $\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}$ (although it is generally possible without this restriction). Of course, plots of simulations do not constitute rigorous measure theory proofs, but they can be suggestive or provide counter-examples. We conjecture that the set of covariance matrices among $n=6$ variables such that $\min\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{\mathcal{R}\left(\boldsymbol{\Sigma}-\boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}\right)\right\}=3=L(6)$ has measure zero.

Conversely, when $n\geq 7$, it appears that the rank of *some* arbitrary correlation matrices can be reduced to $n-3$; their proportion is indicated by the height of the point mass at zero. Hence, we conjecture that the set of covariance matrices among $n\geq 7$ variables such that $\min\limits_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{\mathcal{R}\left(\boldsymbol{\Sigma}-\boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}\right)\right\}=n-3>L(n)$ has positive measure. Such matrices constitute a relatively small set, particularly when n is small, but are about a fifth of the arbitrary correlation matrices among $n=19$ variables. However, it is also true that many arbitrary correlation matrices can have their rank effectively but not literally reduced to $n-3$.

The grey lines in figure 5 trace the empirical CDF of $u\left(\widehat{\boldsymbol{\Theta}}\right)$ when $k=n-4$ factors are retained. Shapiro's (1982) result implies that if $n\leq 9$, the gray lines cannot literally touch zero on the x-axis. The plots for $n\geq 10$ are both novel and interesting in the sense that they look qualitatively similar to the relevant lines when Shapiro's (1982) result is applicable but look qualitatively different from the black lines when

Figure 5: Empirical CDFs of $u\left(\widehat{\mathbf{\Theta}}\right)$ Retaining $n-3$ (black) and $n-4$ (grey) Factors



Simulations with arbitrary correlation matrices are plotted. The x-axis is, $u\left(\widehat{\mathbf{\Theta}}\right)$, the proportion of Unexplained Common Variance of the IRMA$(\boldsymbol{\pi})$ when retaining $k = n-3$ (black line) or $k = n-4$ (grey line) factors. The y-axis is the proportion of simulations less than or equal to x. The point mass at zero is the estimated probability that the rank can be reduced to $n-3$.

$n \geq 7$. It appears as if no arbitrary correlation matrix can have its rank reduced to $n-4$ by any $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$, in contrast to the black lines where there is a distinct point mass at zero when $n \geq 7$. Thus, we conjecture that the set of covariance matrices such that $\min\limits_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ \mathcal{R}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right)\right\} \leq n-4$ has measure zero, while again emphasizing that many can have their rank effectively reduced to $n-4$.

If the last conjecture were true, the null hypothesis that $r \geq n-3$ becomes testable, which would satisfy Guttman's (1958) plea to avoid making the null hypothesis that r is small.[5] If the null hypothesis that $r \geq n-3$ were rejected in a particular case, then the researcher can be reasonably confident that some LISREL model holds in the population. Conversely, if this null hypothesis could not be rejected, it would alert the researcher to a problem with insisting on a LISREL model with a small number of latents.

7 Empirical Example

For any substantive theory where k explanatory variables are said to explain n outcomes, the IRMA can be used to assess a necessary condition for that substantive theory to be true, namely whether $k = r$ or at least whether k explanatory variables explain the vast majority of the common variation in the manifest variables. In fact, the motivation for Thurstone's (1935) theorem was to challenge Spearman's (1904) $r = 1$ theory of intelligence by showing that $\mathcal{R}\left(\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}\right) > 1 \, \forall \widetilde{\mathbf{\Theta}} \in \mathcal{T}$. Similarly, Esping-Andersen (1990) famously argues that $k = 3$ explanatory variables explain a variety of welfare-state outcomes. This section illustrates how this claim can be evaluated with the IRMA.

Esping-Andersen (1990) argues that particular variables best represent the dimensions that welfare-state outcomes are a function of. The "conservative" dimension is defined by two variables: the number of public pension schemes intended for different occupational groups and how much a country spends on pensions for public employees as a percentage of gross domestic product (GDP). The "liberal" dimension is defined by three variables: the percentage of private health spending in total health spending, the percentage of private pensions in total pensions, and the percentage of means-tested poor relief in total social spending. The "social democracy" dimension is defined by two variables: the (average) percentage of the labor force covered under unemployment insurance, sickness insurance, and public pensions and the (average) ratio of the normal to maximum replacement rate for unemployment insurance, sickness insurance, and public

[5] Work in progress is attempting to prove or disprove these three conjectures using techniques from real algebraic geometry.

pensions. Esping-Andersen (1990) believes that each of these seven variables is only a function of the one dimensions that it is associated with.

As noted in Scruggs and Allan (2008) and Scruggs and Pontusson (2008), there are significant questions about the observed data on these variables used in Esping-Andersen (1990). Scruggs and Allan (2008) attempts to replicate Esping-Andersen's (1990) data collection on these seven variables (excluding percentage of private pensions) and finds several important differences. Thus, we use the "circa 1980" data from Scruggs and Allan (2008) on the variables proposed by Esping-Andersen (1990) to test its main hypothesis.

Esping-Andersen (1990) constructs three indices, one for each of the dimensions, using the aforementioned seven variables and weights that are at best subjective and at worst potentially arbitrary. Scruggs and Pontusson (2008) use these three indices, along with some other variables, in their analysis, as do Hicks and Kenworthy (2003) using Esping-Andersen's (1990) data in part. Both Hicks and Kenworthy (2003) and Scruggs and Pontusson (2008) conclude that $r = 2$ but differ in how they consolidate Esping-Andersen's (1990) three dimensions into two. However, the two papers are methodologically similar, and in particular arrive at the conclusion that the number of dimensions is two by noting that two eigenvalues of the sample correlation matrix are greater than unity. Even if this were true for the population correlation matrix, it only provides a lower bound for r and thus is not inconsistent with Esping-Andersen's (1990) theory.

This decision to use Esping-Andersen's (1990) three indices rather than the seven variables that comprise them is perhaps too deferential to Esping-Andersen (1990) since it has not been empirically established that these indices are reliable, particularly using the higher quality data from Scruggs and Allan (2008). There are at least two ways in which these indices could be questioned. First, the constituent variables may not measure the concept the index is intended to represent and second, the weights Esping-Andersen (1990) uses could be misspecified. To investigate these possibilities, we subject the covariance matrix among the seven variables to the IRMA to see if $r = 3$, which is a necessary but not sufficient condition for Esping-Andersen's (1990) indices to be valid. Since $n = 7$ and $L(n) \approx 3.7$, if Esping-Andersen (1990) is correct that $r = 3$, then $\boldsymbol{\Theta}$ is identified, and according to simulations in Goodrich (2009), we reject the null hypothesis that $r \geq 4$ if $u\left(\widehat{\boldsymbol{\Theta}}\right) < 0.0562$ when three factors are retained from the IRMA($\boldsymbol{\pi}$).

Table 1 shows the main result, which is that we fail to reject the null hypothesis that $r \geq 4$, and the same conclusion is reached (but not shown) for the IRMA($\boldsymbol{\alpha}^*$) and IRMA($\boldsymbol{\delta}$). Intuitively, the fourth eigenvalue

Table 1: IRMA($\boldsymbol{\pi}$) Results for Esping-Andersen's (1990) $n = 7$ Variables

Implied Eigenvalues:	1.90	1.05	0.76	0.43	0.01	0.00	0.00

Since the test statistic is about $\frac{0.44}{7} \approx 0.063$, we fail to reject the null hypothesis that $r \geq 4$.

Table 2: IRMA($\boldsymbol{\pi}$) Results for $n = 13$ Variables

Eigenvalues of	2.77	1.90	1.25	0.71	0.52	0.26	0.16
$\mathbf{S} - \arg\max\limits_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_1 \left(\boldsymbol{\pi} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \middle| \mathbf{S} \right\}$	0.09	0.03	0.01	0.00	0.00	0.00	
Diagonal of	0.42	0.43	0.47	0.45	0.43	0.43	0.25
$\arg\max\limits_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_1 \left(\boldsymbol{\pi} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \middle| \mathbf{S} \right\}$	0.40	0.32	0.38	0.37	0.42	0.54	

Since the test statistic is about $\frac{0.04}{13} \approx 0.003$, we reject the null hypothesis that $r \geq 9$.

of $\mathbf{S} - \arg\max\limits_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_1 \left(\boldsymbol{\pi} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \middle| \boldsymbol{\Sigma} \right\}$ is 0.43 rather than (near) zero, so it seems that four dimensions are needed. While it is certainly possible that Esping-Andersen's (1990) three dimensions plus one minor dimension comprise the four dimensions, it is difficult to make any substantive conclusion about the nature of the r inputs until $\boldsymbol{\Theta}$ is identified.

The best thing to do, if possible, is to increase n until $r < L(n)$. Fortunately, it is easy to do so in this case. Using unemployment insurance, sickness insurance, and pensions data from Scruggs (2004), we add three measures of "decommodification" (as defined in Esping-Andersen 1990) and three measures of benefit generosity. According to Esping-Andersen's (1990) theory, decommodification and other variables like benefit generosity are a function of where a nation sits with respect to conservatism, liberalism, and social democracy. Thus, the minimum rank should still be three (or four) but since $L(13) \approx 8.4$, $\boldsymbol{\Theta}$ is identified as long as $r \leq 8$. According to simulations in Goodrich (2009), we reject the null hypothesis that $r \geq 9$ if $u\left(\widehat{\boldsymbol{\Theta}}\right) < 0.0188$ when eight factors are retained from the IRMA($\boldsymbol{\pi}$).

Table 2 shows the results, the most important of which is that we decisively reject the null hypothesis that $r \geq 9$, and the same conclusion is reached (but not shown) for the IRMA($\boldsymbol{\alpha}^*$) and IRMA($\boldsymbol{\delta}$). Intuitively, the ninth eigenvalue of $\mathbf{S} - \arg\max\limits_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_1 \left(\boldsymbol{\pi} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \middle| \boldsymbol{\Sigma} \right\}$ is about 0.02, which is entirely consistent with it being zero in the population, given that the sample size is only 18. So what is the value of r? To answer this question more precisely, we would need to use some of the finite-sample techniques discussed in Goodrich (2009) for inferring r at the optimum. However, it is reasonably clear that r is at least four or five because the fourth eigenvalue is about 0.71 and the fifth is about 0.52, both of which are rather far from zero.

It is possible that Esping-Andersen's (1990) $r = 3$ theory is more-or-less appropriate for these $n = 13$ variables, but leaves a little common variation to be explained by a fourth or fifth input to the data-generating process. At this point, we could follow the route of Hicks and Kenworthy (2003) and Scruggs and Pontusson (2008), which rotate the factors in an attempt to interpret the major dimensions. Or we could estimate a confirmatory or semi-exploratory (see Goodrich 2008) factor model that conditions on r but does not need rotation. A semi-exploratory analysis suggests that Esping-Andersen's (1990) three dimensions are *not* all among the inputs to the data-generating process for these variables, but such an inference — while substantively interesting — is not the focus of this paper. The main point of this paper is that the IRMA puts us into a position to infer r, which (with some help from the methods in Goodrich 2009) appears to be four or five in this case. To analyze a given number of inputs, another model is usually necessary.

It would also be easy to add a few more observed variables, as is done in Hicks and Kenworthy (2003) and Scruggs and Pontusson (2008), or to utilize the variation over time in the data, as is done in Scruggs and Pontusson (2008). Of course, it is implausible to argue that the observations for the same country are conditionally independent from year-to-year, making the assumption that $\mathbf{\Theta}$ is diagonal implausible. However, it would be simple to utilize an autoregressive specification where $\mathbf{\Theta}$ is not diagonal but only requires the estimation of the additional autoregressive parameter, ρ. Some of the theorems in this paper would need to be modified slightly for the case where $\mathbf{\Theta}$ is not diagonal, but the fundamental principle of (indirect) rank-minimization remains appropriate.

8 Conclusion

Although this paper is quite long, it has made only one major accomplishment, namely figuring out how to solve Thurstone's (1935) rank-minimization problem. Instead of choosing $\widetilde{\mathbf{\Theta}}$ to minimize the rank of $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ directly — which is essentially impossible — $\widetilde{\mathbf{\Theta}}$ is chosen to maximize the eigenvalue dispersion of PSD matrices with the same rank to obtain a boundary solution where as many eigenvalues as possible pile up at the boundary of zero. This singular contribution to the literature is fundamental.

The recognition that the minimum rank of $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ was the smallest number of factors consistent with a factor analysis model was the pillar on which Thurstone (1935) founded multiple factor analysis as a generalization of Spearman's (1904) single-factor model. Since then, factor analysis has been used more than a million times across dozens of disciplines, been generalized into LISREL modeling, and continues to thrive despite persistent criticism that the choice of the number of factors is fairly arbitrary.

The critics have a point: Thurstone did not think his theorem for discovering r was useful in practice because $\mathbf{\Sigma}$ was never observed. Guttman was among the critics, although his primary criticism was that r could not be discovered even if $\mathbf{\Sigma}$ were observed, except in special cases. Guttman (1954, 1956, 1958) did more to advance the understanding of the rank-minimization problem than did anyone else, but eventually, Guttman essentially abandoned common factor analysis, despite his respect for Spearman and Thurstone, in part because r was undiscoverable (and in part because $\mathbf{\Delta} \neq \mathbf{I}$). Thus, it is not hyperbole to say that three of the foremost social scientists of all time — Spearman, Thurstone, and Guttman — each spent a considerable portion of their professional careers trying to find r.

Decades later, this fundamental problem is now solved, at least in the population and as the sample size tends toward infinity. If c is sufficiently small (and $c = 1$ typically is so), then the IRMA would find a minimum-rank solution if given $\mathbf{\Sigma}$. The primary remaining question is how to infer r from a finite-sample solution that is only approximately a minimum-rank solution in the population. This question is taken up in Goodrich (2009), and several ideas, some from the historical literature and some new ones, are promising.

Some secondary aspects of the hypothetical case where $\mathbf{\Sigma}$ is observed merit future research. One partially unresolved question is what value of c is best to use, although it seems that $c = 1$ is adequate, especially if r is far from the Ledermann (1937) bound in either direction. Also, it is probably possible to make further improvements to the code that executes the IRMA in order to reduce the already small number of cases where it fails to converge to the global optimum, particularly for the fast engine. Future work should investigate the IRMA when $\mathbf{\Sigma} \approx \mathbf{\Omega} \left(\mathbf{\Lambda \Upsilon \Lambda}' + \mathbf{\Theta}\right) \mathbf{\Omega}$ and / or when $\widetilde{\mathbf{\Theta}}$ is not diagonal.

These remaining questions should not stop anyone from using the IRMA today, even in a sample. The IRMA can be used for many purposes. In measurement and EITM models where the main hypothesis is that $r = 1$, the IRMA can be used to test this hypothesis. Simply obtain the covariance matrix among $n > 3$ variables, execute the IRMA both ways, and observe whether the results suggest that $r = 1$ in the population. If so, then the analysis can proceed. If not, it may be possible to find different variables or some subset of the variables where the $r = 1$ hypothesis is reasonable. The same recipe can be followed when the hypothesis is that r is some number greater than one, as in the Esping-Andersen (1990) case.

The IRMA can also be used as a precursor to a regression. Obtain the covariance matrix among all outcome variables and proposed explanatory variables, execute the IRMA both ways, and check whether r is less than or equal to the proposed number of explanatory variables. If $r < L\left(n\right)$, it is also necessary to check that the diagonal elements of $\widehat{\mathbf{\Theta}}$ that correspond to explanatory variables are nearly zero, which is to say that they are largely free of measurement error. If so, continue with the regression. If r is larger than the proposed number of explanatory variables, omitted variable bias will likely be a problem. If the explanatory variables are measured with error, their coefficients will be biased. In either case, it would be better to find an instrument for the key explanatory variable or to estimate a LISREL model where the explanatory variables are considered latent.

Of course, the IRMA can be used to choose the number of factors in a factor analysis or, more generally, to choose the number of latents in a LISREL model (with some full-rank assumptions). This was the original impetus for finding a rank-minimization algorithm. LISREL models have never been popular in political science, except for the special case of linear regressions, which are ubiquitous. Perhaps one reason for the unpopularity of LISREL models with latent variables is that there has not been a good way to choose r until now.

The IRMA can be used to validate multiple imputation models. The multiple imputation algorithm will typically produce an estimate of $\mathbf{\Sigma}$, which is the covariance matrix among all observed variables and can be fed to the IRMA. The key question is whether $\left[\mathbf{\Sigma}^{-1}\right]_{ii}^{-1} = \Theta_{ii} \, \forall i$ in the population, where the left-hand side is the error variance when the ith manifest variable is predicted by the other $n-1$ manifest variables and the right-hand side is the true error variance. If this equation holds (approximately), then the missing values are being drawn from a conditional distribution that (approximately) has the correct variance. If not, then researchers should try to add variables to their multiple imputation models.

Finally, the IRMA can be used to in conjunction with experiments or other situations where matching is used. Simply execute the IRMA both ways to the covariance matrix among baseline outcomes and any available covariates. If $r < L\left(n\right)$, generate scores on r factors, create matched pairs to achieve multivariate balance on the r factors, and — if the treatment has not already been assigned — randomly assign the treatment to one individual in each pair. If there are many background covariates, but they are measured with error and / or redundant, then it will presumably be easier to achieve balance on r factor scores than on the n observed variables. If there are baseline outcomes, this procedure can also be used to avoid the "endogenous selection bias" discussed in Elwert and Winship (2008).

References

Abadie, A., A. Diamond and J. Hainmueller. 2007. "Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco Control Program." *NBER Working Paper* .

Ansolabehere, S., J. Rodden and J.M. Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102(02):215–232.

Bekker, P.A. and J. de Leeuw. 1987. "The rank of reduced dispersion matrices." *Psychometrika* 52(1):125–135.

Bekker, P.A. and J.M.F. ten Berge. 1997. "Generic global indentification in factor analysis." *Linear Algebra and its Applications* 264:255–263.

Bentler, P.M. 1972. "A lower-bound method for the dimension-free measurement of internal consistency." *Social Science Research* 1(3):343–357.

Cowell, F.A. 2008. *Measuring inequality*. Prentice Hall.

Davies, P.I. and N.J. Higham. 2000. "Numerically stable generation of correlation matrices and their factors." *BIT Numerical Mathematics* 40(4):640–651.

de Leeuw, J. 2007. "Derivatives of Generalized Eigen Systems with Applications." Unpublished paper available from http://preprints.stat.ucla.edu/download.php?paper=528.

Della Riccia, G. and A. Shapiro. 1982. "Minimum rank and minimum trace of covariance matrices." *Psychometrika* 47(4):443–448.

Elwert, F. and C. Winship. 2008. "Endogenous Selection Bias." Unpublished paper, Harvard University.

Esping-Andersen, G. 1990. *The three worlds of welfare capitalism*. Polity Press Cambridge.

Fazel, M., H. Hindi and S. Boyd. 2004. Rank minimization and applications in system theory. In *American Control Conference, 2004. Proceedings of the 2004*. Vol. 4.

Foster, J.E. and A.A. Shneyerov. 1999. "A general class of additively decomposable inequality measures." *Economic Theory* 14(1):89–111.

Goodrich, B. 2008. "Semi-Exploratory Factor Analysis and Software to Estimate It." Paper presented at the annual meeting of the APSA 2008 Annual Meeting, Boston, MA.

Goodrich, B. 2009. "Choosing the Number of Latents with the Indirect Rank-Minimization Algorithm." Unpublished paper, Harvard University.

Guttman, L. 1954. "Some necessary conditions for common-factor analysis." *Psychometrika* 19(2):149–161.

Guttman, L. 1955. "The determinacy of factor score matrices with implications for five other basic problems of common-factor theory." *British Journal of Statistical Psychology. Vol* 8:65–81.

Guttman, L. 1956. ""Best possible" systematic estimates of communalities." *Psychometrika* 21(3):273–285.

Guttman, L. 1958. "To what extent can communalities reduce rank?" *Psychometrika* 23(4):297–308.

Guttman, L. 1977. "What is not what in statistics." *The Statistician* pp. 81–107.

Hayduk, L.A. 1987. *Structural equation modeling with LISREL: essentials and advances*. Johns Hopkins Univ Pr.

Hicks, A. and L. Kenworthy. 2003. "Varieties of welfare capitalism." *Socio-Economic Review* 1(1):27–61.

Irwin, L. 1966. "A method for clustering eigenvalues." *Psychometrika* 31(1):11–16.

Jöreskog, KG and D. Sörbom. 1996. *LISREL 8: User's reference guide*. Scientific Software.

Kaiser, H.F. and J. Caffrey. 1965. "Alpha factor analysis." *Psychometrika* 30(1):1–14.

King, G. and L. Zeng. 2001. "Logistic regression in rare events data." *Political Analysis* 9(2):137.

Krijnen, W.P. 2006. "Convergence of Estimates of Unique Variances in Factor Analysis, Based on the Inverse Sample Covariance Matrix." *Psychometrika* 71(1):193–199.

Ledermann, W. 1937. "On the rank of the reduced correlational matrix in multiple-factor analysis." *Psychometrika* 2(2):85–93.

Lewandowski, D., D. Kurowicka and H. Joe. 2009. "Generating random correlation matrices based on vines and extended onion method." *Journal of Multivariate Analysis* 100(9):1989–2001.

Mebane, W.R. and J.S. Sekhon. 1998. "GENBLIS: GENetic optimization and Bootstrapping of LInear Structures." computer program.
URL: *http://sekhon.berkeley.edu/genblis/*

Mebane, W.R. and J.S. Sekhon. 2009. "Genetic Optimization Using Derivatives: The rgenoud package for R." *Journal of Statistical Software* 13(9). forthcoming , available from http://sekhon.berkeley.edu/papers/rgenoudJSS.pdf.

Mulaik, S.A. 2005. Looking back on the indeterminacy controversies in factor analysis. In *Contemporary psychometrics: A festschrift for Roderick P. McDonald*, ed. J.J. McArdle and A. Maydeu-Olivares. Lawrence Erlbaum Associates pp. 173–206.

Persson, T. and G.E. Tabellini. 2002. *Political economics: explaining economic policy*. The MIT press.

Poole, K.T. and H. Rosenthal. 1997. *Congress: A political-economic history of roll call voting*. Oxford University Press, USA.

Quinn, K.M. 2004. "Bayesian factor analysis for mixed ordinal and continuous responses." *Political Analysis* 12(4):338–353.

Scruggs, L. 2004. “Welfare State Entitlements Data Set: A Comparative Institutional Analysis of Eighteen Welfare States, version 1.2.”.
URL: *http://sp.uconn.edu/ scruggs/wp.htm*

Scruggs, L. and J. Pontusson. 2008. “New Dimensions of Welfare State Regimes in Advanced Democracies.” Paper presented at the American Political Science Association conference.
URL: *http://www.princeton.edu/ jpontuss/ScruggsPontussonAPSA08.pdf*

Scruggs, L.A. and J.P. Allan. 2008. “Social Stratification and Welfare Regimes for the Twenty-first Century: Revisiting The Three Worlds of Welfare Capitalism.” *World Politics* 60(4):642–664.

Sekhon, J.S. and W.R. Mebane. 1998. “Genetic optimization using derivatives.” *Political Analysis* 7(1):187–210.

Shapiro, A. 1982. “Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis.” *Psychometrika* 47(2):187–199.

Shapiro, A. and J.M.F. ten Berge. 2000. “The asymptotic bias of minimum trace factor analysis, with applications to the greatest lower bound to reliability.” *Psychometrika* 65(3):413–425.

Shapiro, A. and J.M.F. ten Berge. 2002. “Statistical inference of minimum rank factor analysis.” *Psychometrika* 67(1):79–94.

Shorrocks, A.F. 1984. “Inequality decomposition by population subgroups.” *Econometrica: Journal of the Econometric Society* pp. 1369–1385.

Sočan, Gregor. 2003. The incremental value of minimum rank factor analysis. PhD thesis University of Groningen. http://irs.ub.rug.nl/ppn/254817777.

Ten Berge, J.M.F. and H.A.L. Kiers. 1991. “A numerical approach to the approximate and the exact minimum rank of a covariance matrix.” *Psychometrika* 56(2):309–315.

Theil, H. 1967. *Economics and Information Theory*. Rand McNally.

Thurstone, L.L. 1935. *The vectors of mind: multiple-factor analysis for the isolation of primary traits*. The University of Chicago Press.

Treier, S. and S. Jackman. 2003. “Democracy as a latent variable.” Available from http://jackman.stanford.edu/papers/master.pdf.

Treier, S. and S. Jackman. 2008. “Democracy as a latent variable.” *American Journal of Political Science* 52(1):201–217.

Appendix A: Eigenvalue derivatives

First, we seek the derivative of the jth eigenvalue of $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$ with respect to $\widetilde{\Theta}_{ii}$, assuming it is simple. Since the derivative of $\widetilde{\pi}_j - 1$ with respect to $\widetilde{\Theta}_{ii}$ is the same as $\frac{\partial \widetilde{\pi}_j}{\partial \widetilde{\Theta}_{ii}}$, we can instead find the derivative of

$\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)-\mathbf{I}=\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-\frac{1}{2}}\left(\mathrm{cor}\left(\mathbf{y}\right)-\mathbf{I}\right)\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-\frac{1}{2}}$, where $\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)=\mathbf{I}-\widetilde{\boldsymbol{\Theta}}$ is a diagonal matrix of proposed communalities. Several known results on the derivatives of parameterized eigenvalues are summarized in de Leeuw (2007), the most important of which in this case is that if $\widetilde{\pi}_j-1$ is simple, then $\frac{\partial\widetilde{\pi}_j-1}{\partial\widetilde{\Theta}_{ii}}=\widetilde{\mathbf{p}}_j'\frac{\partial\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)-\mathbf{I}}{\partial\widetilde{\Theta}_{ii}}\widetilde{\mathbf{p}}_j$, where $\widetilde{\mathbf{p}}_j'$ is the jth normalized eigenvector of $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)-\mathbf{I}$.

If $\mathbf{D}$ is diagonal and $\mathbf{C}$ is symmetric, then $\frac{\partial\mathbf{DCD}}{\partial D_{ii}}=2\mathbf{DCJ}^{ii}$, where $\mathbf{J}^{ii}$ is a square matrix with zero everywhere except 1.0 in the ith diagonal cell. Letting $\mathbf{D}=\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-\frac{1}{2}}$ and $\mathbf{C}=\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)-\mathbf{I}$, we can write $\frac{\partial\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)-\mathbf{I}}{\partial\widetilde{\Theta}_{ii}}=\frac{\partial\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-\frac{1}{2}}(\mathrm{cor}(\mathbf{y})-\mathbf{I})\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-\frac{1}{2}}}{\partial\left(1-\widetilde{\Theta}_{ii}\right)^{-\frac{1}{2}}}\times\frac{\partial\left(1-\widetilde{\Theta}_{ii}\right)^{-\frac{1}{2}}}{\partial\widetilde{\Theta}_{ii}}=\frac{2\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-\frac{1}{2}}(\mathrm{cor}(\mathbf{y})-\mathbf{I})\mathbf{J}^{ii}}{2\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{3}{2}}}$. Now, premultiply $\mathbf{J}^{ii}$ by $\mathbf{I}$ in the form of $\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-\frac{1}{2}}\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{\frac{1}{2}}$ so that $\frac{\partial\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)-\mathbf{I}}{\partial\widetilde{\Theta}_{ii}}=\frac{\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-\frac{1}{2}}(\mathrm{cor}(\mathbf{y})-\mathbf{I})\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-\frac{1}{2}}\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{\frac{1}{2}}\mathbf{J}^{ii}}{\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{3}{2}}}=\frac{\left(\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)-\mathbf{I}\right)\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{\frac{1}{2}}\mathbf{J}^{ii}}{\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{3}{2}}}$. Since $\mathbf{H}\left(\widetilde{\boldsymbol{\Theta}}\right)^{\frac{1}{2}}\mathbf{J}^{ii}=\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{1}{2}}\mathbf{J}^{ii}$, we get $\frac{\partial\widetilde{\pi}_j}{\partial\widetilde{\Theta}_{ii}}=\widetilde{\mathbf{p}}_j'\frac{\left(\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)-\mathbf{I}\right)\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{1}{2}}\mathbf{J}^{ii}}{\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{3}{2}}}\widetilde{\mathbf{p}}_j=\frac{\left(\widetilde{\mathbf{p}}_j'\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)-\widetilde{\mathbf{p}}_j'\right)\mathbf{J}^{ii}\widetilde{\mathbf{p}}_j}{1-\widetilde{\Theta}_{ii}}$. Recall from the definition of eigenvalues that $\widetilde{\mathbf{p}}_j'\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)=\widetilde{\mathbf{p}}_j'\widetilde{\pi}_j$, which can be substituted into the previous expression to yield $\frac{\partial\widetilde{\pi}_j}{\partial\widetilde{\Theta}_{ii}}=\frac{\left(\widetilde{\pi}_j-1\right)\widetilde{\mathbf{p}}_j'\mathbf{J}^{ii}\widetilde{\mathbf{p}}_j}{1-\widetilde{\Theta}_{ii}}=\frac{\left(\widetilde{\pi}_j-1\right)\widetilde{P}_{ij}^2}{1-\widetilde{\Theta}_{ii}}$. Recall that $\widetilde{\alpha}_j^*=\frac{n+1}{n}\left(1-\frac{1}{\widetilde{\pi}_j+1}\right)$, so again assuming that $\widetilde{\pi}_j$ is distinct, the chain rule implies $\frac{\partial\widetilde{\alpha}_j^*}{\partial\widetilde{\Theta}_{ii}}=\frac{\partial\widetilde{\pi}_j}{\partial\widetilde{\Theta}_{ii}}\times\frac{\partial\widetilde{\alpha}_j}{\partial\widetilde{\pi}_j}=\frac{\left(\widetilde{\pi}_j-1\right)\widetilde{P}_{ij}^2(n+1)}{\left(\widetilde{\pi}_j+1\right)^2\left(1-\widetilde{\Theta}_{ii}\right)n}$.

Turning now to $\frac{\partial\phi_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial\widetilde{\Theta}_{ii}}$, let $\mathbf{D}=\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}$ and $\mathbf{C}=\mathrm{cor}\left(\mathbf{y}\right)$, so $\frac{\partial\widetilde{\phi}_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial\widetilde{\Theta}_{ii}}=\widetilde{\mathbf{g}}_j'\left(\frac{\partial\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}\mathrm{cor}(\mathbf{y})\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}}{\partial\widetilde{\Theta}_{ii}^{-\frac{1}{2}}}\times\frac{\partial\widetilde{\Theta}_{ii}^{-\frac{1}{2}}}{\partial\widetilde{\Theta}_{ii}}\right)\widetilde{\mathbf{g}}_j$ $=-\widetilde{\mathbf{g}}_j'\left(\frac{2\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}\mathrm{cor}(\mathbf{y})\mathbf{J}^{ii}}{2\widetilde{\Theta}_{ii}^{\frac{3}{2}}}\right)\widetilde{\mathbf{g}}_j$ where $\widetilde{\mathbf{g}}_j$ is the jth normalized eigenvector of $\boldsymbol{\Phi}\left(\widetilde{\boldsymbol{\Theta}}\right)$. Again premultiplying $\mathbf{J}^{ii}$ by $\mathbf{I}$ in the form of $\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}\widetilde{\boldsymbol{\Theta}}^{\frac{1}{2}}$ and noting that $\widetilde{\boldsymbol{\Theta}}^{\frac{1}{2}}\mathbf{J}^{ii}=\widetilde{\Theta}_{ii}^{\frac{1}{2}}\mathbf{J}^{ii}$, we obtain $\frac{\partial\widetilde{\phi}_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial\widetilde{\Theta}_{ii}}=\frac{\widetilde{\mathbf{g}}_j'\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}\mathrm{cor}(\mathbf{y})\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}\mathbf{J}^{ii}\widetilde{\mathbf{g}}_j}{\widetilde{\Theta}_{ii}}$.

Since $\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}\mathrm{cor}\left(\mathbf{y}\right)\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}=\widetilde{\mathbf{G}}\begin{bmatrix}\phi_1\left(\widetilde{\boldsymbol{\Theta}}\right) & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots\\ \mathbf{0} & \cdots & \phi_n\left(\widetilde{\boldsymbol{\Theta}}\right)\end{bmatrix}\widetilde{\mathbf{G}}'$, $\widetilde{\mathbf{g}}_j'\widetilde{\mathbf{G}}$ is a vector with zero everywhere except for unity in the jth cell, and $\widetilde{\mathbf{G}}'\mathbf{J}^{ii}\widetilde{\mathbf{g}}_j$ is a vector with $\widetilde{G}_{ij}^2$ in its jth cell, $\frac{\partial\widetilde{\phi}_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial\widetilde{\Theta}_{ii}}=\frac{\widetilde{\phi}_j\left(\widetilde{\boldsymbol{\Theta}}\right)\widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii}}$. Recalling that $\widetilde{\delta}_j=1-\frac{1}{\widetilde{\phi}_j}$, we obtain $\frac{\partial\delta_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial\widetilde{\Theta}_{ii}}=\frac{\partial\widetilde{\phi}_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial\widetilde{\Theta}_{ii}}\times\frac{\partial\delta_j\left(\widetilde{\boldsymbol{\Theta}}\right)}{\partial\widetilde{\phi}_j\left(\widetilde{\boldsymbol{\Theta}}\right)}=\frac{\widetilde{\phi}_j\left(\widetilde{\boldsymbol{\Theta}}\right)\widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii}}\times\frac{1}{\widetilde{\phi}_j\left(\widetilde{\boldsymbol{\Theta}}\right)^2}=\frac{\widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii}\widetilde{\phi}_j\left(\widetilde{\boldsymbol{\Theta}}\right)}=\frac{\left(\widetilde{\delta}_j\left(\widetilde{\boldsymbol{\Theta}}\right)-1\right)\widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii}}$.

Appendix B: Computational Details

Finding $\arg\max_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}}\left\{ D_c\left(\widetilde{\mathbf{x}}\right)\middle| \mathbf{\Sigma}\right\}$ is a non-trivial problem, regardless of whether if c is small enough to render it a minimum-rank solution. Although $D_c\left(\widetilde{\mathbf{x}}\right)$ is a continuous function of $\widetilde{\mathbf{\Theta}}$, it is not maximized at a mode in the interior of $\mathcal{T}$ but rather on the the frontier of $\mathcal{T}$ where $\mathbf{\Sigma}-\mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ is PSD but singular. Hill-climbing algorithms exploit the fact that the gradient of the objective function is a zero vector at an interior mode and thus are not useful for finding $\arg\max_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}}\left\{ D_c\left(\widetilde{\mathbf{x}}\right)\middle| \mathbf{\Sigma}\right\}$. The IRMA requires "gradient-free" optimization algorithms that only depend on the values of the objective function.

In doing so, the constraint that $\widetilde{\mathbf{\Theta}}\in\mathcal{T}$ must be enforced. One unsatisfactory approach is to "define" $D_c\left(\widetilde{\mathbf{x}}\right)$ as $-\infty$ if $\widetilde{\mathbf{\Theta}}\notin\mathcal{T}$. However, this approach creates a "cliff" in the parameter space where proposals that barely violate the PSD constraint are inferior to all proposals that satisfy the PSD constraint no matter how bad an admissible proposal is. Such cliffs make it difficult to find a minimum-rank solution, which is necessarily on the very edge of the cliff.

A better approach is to use an eigenvalue shifting strategy such that $\widetilde{\mathbf{\Theta}}=\dot{\mathbf{\Theta}}+\dot{\lambda}_n\mathbf{I}$, where $\dot{\mathbf{\Theta}}$ is a diagonal proposal for $\mathbf{\Theta}$ that is not required to be within $\mathcal{T}$ and $\dot{\lambda}_n$ is the smallest (and possibly negative) eigenvalue of $\operatorname{cor}\left(\mathbf{y}\right)-\dot{\mathbf{\Theta}}$. Irwin (1966) proves that if $\widetilde{\mathbf{\Theta}}$ is defined in this way, then $\mathbf{\Sigma}-\mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ is PSD and singular. While there is still a possibility that $\widetilde{\Theta}_{ii}\notin(0,1)$, "defining" $D_c\left(\widetilde{\mathbf{x}}\right)$ as $-\infty$ in that case does not place the optimum on the edge of a cliff, unless $\Theta_{ii}=0$ or $\Theta_{ii}=1$. This parameterization of $\widetilde{\mathbf{\Theta}}$ entails a performance penalty because the objective function must calculate eigenvalues twice, the first time to calculate $\dot{\lambda}_n$ at $\dot{\mathbf{\Theta}}$ and the second time to calculate $\mathbf{x}$ at $\widetilde{\mathbf{\Theta}}=\dot{\mathbf{\Theta}}+\dot{\lambda}_n\mathbf{I}$. However, doing so tends to find $\arg\max_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}}\left\{ D_c\left(\widetilde{\mathbf{x}}\right)\middle| \mathbf{\Sigma}\right\}$ more reliably and with fewer total calls to the objective function.

Most optimization algorithms are designed to handle only monotone transformations of the parameters, which is decidedly not the case with $\widetilde{\mathbf{\Theta}}=\dot{\mathbf{\Theta}}+\dot{\lambda}_n\mathbf{I}$. Maximizing a function in $\dot{\mathbf{\Theta}}$-space is daunting because the transformation is not unique — both $\dot{\mathbf{\Theta}}$ and $\dot{\mathbf{\Theta}}+a\mathbf{I}$ transform to the same $\widetilde{\mathbf{\Theta}}$ — which implies plateaus in $\dot{\mathbf{\Theta}}$-space. We overcame this problem by contributing a transformation option to Mebane and Sekhon's (2009) RGENOUD optimization algorithm that was accepted into the source code after that article went to press. RGENOUD is a genetic algorithm that has been in development for about fifteen years (see Sekhon and Mebane 1998), is capable of solving many difficult optimization problems, and has already been fruitfully used to estimate LISREL models (see Mebane and Sekhon 1998). As explained in Mebane and

Sekhon (2009), RGENOUD can find the global optimum of an objective function for essentially the same reasons that an appropriate Markov chain converges to its stationary distribution. Our key modification is to breed the $g + 1$th generation from the gth generation *after* it has been transformed to $\widetilde{\mathbf{\Theta}}$-space. This modification facilitates the convergence of the population to $\arg\max\limits_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}} \{ D_c(\widetilde{\mathbf{x}}) | \mathbf{\Sigma}\}$ because the population has greater fitness in $\widetilde{\mathbf{\Theta}}$-space than in $\dot{\mathbf{\Theta}}$-space. Put in different terms, with this modification, RGENOUD behaves as if the parameter space consists only of those proposals such that the reduced covariance matrix is PSD but *singular*, which is a $n - 1$ dimensional subspace of $\mathcal{T}$.

RGENOUD can take a relatively long time to find $\arg\max\limits_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}} \{ D_c(\widetilde{\mathbf{x}}) | \mathbf{\Sigma}\}$. Even when n is small, it can easily take 30 seconds to find $\arg\max\limits_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}} \{ D_c(\widetilde{\mathbf{x}}) | \mathbf{\Sigma}\}$, and the expected duration increases rapidly as n increases. In actual research situations, computation times of a few minutes are negligible, but for Monte Carlo simulations, bootstrapping, and demonstrations, it is useful to have a faster algorithm. A simplex optimization algorithm is somewhat similar to a genetic algorithm, but with a population of $n + 1$ and much simpler, deterministic reproductive rules. Thus, it can converge much faster (generally within a few seconds) than RGENOUD, even with the $\widetilde{\mathbf{\Theta}} = \dot{\mathbf{\Theta}} + \dot{\lambda}_n \mathbf{I}$ transformations.

The simplex algorithm produced more eigenvalue dispersion than RGENOUD in less than fifteen percent of the population simulations when $1 < r < L(n)$. However, when $r \geq L(n)$, the simplex algorithm was better in almost forty percent of the population simulations. In samples with $N = 1000$, the simplex algorithm was better in about forty percent of the simulations for the IRMA($\boldsymbol{\pi}$) and IRMA($\boldsymbol{\alpha}^*$) but better in about sixty five percent of the simulations for the IRMA($\boldsymbol{\delta}$). With smaller sample sizes, the balance between the simplex algorithm and RGENOUD tilted toward the latter. For simulations with arbitrary correlation matrices, the simplex algorithm was rarely preferable to RGENOUD for the IRMA($\boldsymbol{\pi}$) and IRMA($\boldsymbol{\alpha}^*$). However, the simplex algorithm was preferable for the IRMA($\boldsymbol{\delta}$) in more than fifty percent of the simulations with arbitrary correlation matrices when $n = 5$, falling gradually to about forty percent when $n = 18$. In short, in a real research situation, it would be very difficult to guess whether the simplex algorithm or RGENOUD will produce more eigenvalue dispersion, so both should be executed, perhaps multiple times each with different starting values and random number seeds.

As can be seen from the generally small values of $u\left(\widehat{\mathbf{\Theta}}\right)$ reported in the body of this paper, FAiR is generally successful at getting extremely close to $\arg\max\limits_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}} \{ D_c(\widetilde{\mathbf{x}}) | \mathbf{\Sigma}\}$. However, it is still possible to experience an optimization error, so researchers should use proper precautions whenever the results are important. In particular, use population sizes of at least 1000 with strict convergence thresholds, verify that the IRMA($\boldsymbol{\pi}$), IRMA($\boldsymbol{\alpha}^*$), and IRMA($\boldsymbol{\delta}$) imply the same value of r, do so a few times with different pseudo-random number seeds, check the results with the simplex engine, etc.