Bringing Rank-Minimization Back In:

An Estimator of the Number of Inputs to a Data-Generating Process

Ben Goodrich, Harvard University

September 15, 2009

Abstract

This paper derives and implements an algorithm to infer the number of 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 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 political economy literature, 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 a computer program that could answer each of 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. This algorithm is implemented in my R package (FAiR, 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 (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 algrebra and usually convex optimization theory. While none of these things are necessary 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). All the proofs are contained in the Technical Appendix, and there is a Computational Appendix that describes how the new optimization problem is solved in R.

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 one common

input, namely the true concept. In symbols, (in this paper the observations are *not* subscripted)

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

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

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

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 (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 neglible.

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. Suppose in a voter turnout experi-

ment that we have data on whether individuals voted in the previous half-dozen elections. Suppose further that voters are of essentially three types: those who almost never vote, those who almost always vote, and those who vote with probability 0.5. The latter types will tend to vote in three of the last six elections but which three is fairly random and not necessarily important. Exact matching on realized turnout is a somewhat inefficient way of matching people of the third type with each other, which is important because those people are likely the ones who are most amendable to interventions. If we think of realized turnout as a noisy measure of the true "propensity to turnout", we could estimate and then match on that.

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 Population Model

This section lays the groundwork for Thurstone's (1935) theorem but incorporates some generalizations to the setup that did not become available until after his death. Whereas Thurstone stated the theorem in terms of a factor analysis model, it is no less applicable to a more general LISREL model that includes factor analysis, linear regression, simultaneous equations and other models as special cases. In order to estimate the parameters of any of these models, a researcher has to assume a value for r. In this paper, we are trying

to estimate r as a prelude to estimating a LISREL model or any other model that conditions on r.

Although it is less common than some other parameterizations, Jöreskog and Sörbom (1996), Hayduk (1987), and others show us that we can write a general LISREL model as follows:

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

where $\boldsymbol{\eta}$ is a vector of *latent* explanatory variables with structural errors $\boldsymbol{\zeta}$, $\mathbf{y}$ is a vector of manifest variables with measurement errors $\boldsymbol{\epsilon}$, and $\boldsymbol{\Lambda}$ and $\mathbf{B}$ are coefficient matrices with the same number of columns. A latent variable is exogenous if and only if 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 if and only if it is not a function of any endogenous latent variables. All variables are expressed as deviations from their means to eliminate intercepts, but $\boldsymbol{\eta}$ has no intrinsic scale, implying that some normalizations will be necessary. Our goal is to infer the length of $\boldsymbol{\eta}$, which is the number of inputs in the data-generating process for $\mathbf{y}$.

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

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

where $\boldsymbol{\Omega}$ is a diagonal matrix such that $0 < \Omega_{ii} = \sqrt{\Sigma_{ii}} < \infty \, \forall i$, $\boldsymbol{\Omega}\boldsymbol{\Theta}\boldsymbol{\Omega} = \operatorname{cov}(\boldsymbol{\epsilon})$, and $\boldsymbol{\Psi} = \operatorname{cov}(\boldsymbol{\zeta})$. Therefore, $\operatorname{cor}(\mathbf{y}) = \boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$, where both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ are recalibrated to standardized variables and restricted so that every diagonal cell of $\boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$ is unity. Similarly, $\mathbf{B}$ and $\boldsymbol{\Psi}$ are restricted so that every diagonal cell of $\boldsymbol{\Upsilon} = \operatorname{cor}(\boldsymbol{\eta})$ is unity. However, in this paper, we are estimating $\boldsymbol{\Theta}$ given $\boldsymbol{\Sigma}$. The other parameters are not identified, so we do not concern ourselves with them.

Let $n > 3$ 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 and their eigenvalues are taken to be in non-

increasing order. If all the eigenvalues of a symmetric matrix are positive, the matrix is said to be positive definite and have full rank. If all the eigenvalues of a symmetric matrix are non-negative, the matrix is said to be positive semi-definite (PSD) and its rank is decremented by each eigenvalue that is zero.

We assume that $\boldsymbol{\Sigma}$ is not diagonal and has full rank but $\boldsymbol{\Lambda\Upsilon\Lambda}'$ does not, implying $\mathcal{R}\left(\boldsymbol{\Lambda\Upsilon\Lambda}'\right) \leq \min\left(\mathcal{R}\left(\boldsymbol{\Lambda}\right), \mathcal{R}\left(\boldsymbol{\Upsilon}\right)\right) < n$. The "reduced covariance matrix", $\boldsymbol{\Omega\Lambda\Upsilon\Lambda}'\boldsymbol{\Omega} = \boldsymbol{\Sigma} - \boldsymbol{\Omega\Theta\Omega} = \operatorname{cov}\left(\mathbf{y} - \boldsymbol{\Omega\epsilon}\right)$ has the same rank as the "reduced correlation matrix", $\boldsymbol{\Lambda\Upsilon\Lambda}'$. Both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Upsilon} = \operatorname{cor}\left(\boldsymbol{\eta}\right)$ typically have full rank, in which case $\mathcal{R}\left(\boldsymbol{\Lambda}\right) = r = \mathcal{R}\left(\boldsymbol{\Upsilon}\right)$. Thus, if r were known, it would be tempting to conclude that the length of $\boldsymbol{\eta}$ is r, which is to say that r is the number of inputs in the data-generating process for $\mathbf{y}$. There are a few exceptions where either $\boldsymbol{\Lambda}$ or $\boldsymbol{\Upsilon}$ is rank-deficient, which temper such a sweeping conclusion, but even in those cases, r is the relevant number to know for subsequent modeling. To keep things simple, we will assume that both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Upsilon}$ have full rank.

To find r we need to make some assumptions about the errors. While it is not logically necessary to assume that $\boldsymbol{\Theta}$ is diagonal, it is necessary to assume that $\boldsymbol{\Theta}$ is highly structured and not $\mathbf{0}$. Nevertheless, for expositional clarity, we will assume that $\boldsymbol{\Theta}$ is diagonal. Since both $\boldsymbol{\Theta} = \operatorname{cov}\left(\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}\right)$ and $\boldsymbol{\Sigma} - \boldsymbol{\Omega\Theta\Omega} = \operatorname{cov}\left(\mathbf{y}|\,\boldsymbol{\eta}\right)$ are covariance matrices, we require any proposal for $\boldsymbol{\Theta}$, denoted $\widetilde{\boldsymbol{\Theta}}$, to that imply that both $\widetilde{\boldsymbol{\Theta}}$ and $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ are PSD. Let $\mathcal{T}$ represent the set of admissible diagonal proposals for $\boldsymbol{\Theta}$.

Let the Ledermann (1937) bound be $L\left(n\right) = 0.5\left(2n + 1 - \sqrt{8n+1}\right)$. We can now state Thurstone's (1935) theorem:

Theorem 1. *If* $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\left(\boldsymbol{\Lambda\Upsilon\Lambda}' + \boldsymbol{\Theta}\right)\boldsymbol{\Omega}$ *such that both* $\boldsymbol{\Lambda}$ *and* $\boldsymbol{\Upsilon}$ *have full rank* r*, then* r *is the minimum rank of* $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ *when* $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$*. If, in addition,* $r < L\left(n\right)$*, then the rank-minimizing* $\widetilde{\boldsymbol{\Theta}}$ *is almost surely unique and hence equal to* $\boldsymbol{\Theta}$*.*

Proof. The proofs of these two statements are quite long and not repeated here. The first assertion is proven rigorously in Reiersøl (1950) and the second assertion is proven in Bekker and ten Berge (1997). Henceforth, proofs will be given in the Technical Appendix. □

Thus, 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, under the assumptions of the theorem, it would tell us the value of r at the optimum and, if $r < L\left(n\right)$, also the value of $\boldsymbol{\Theta}$. This problem is also known as the (or at least "a") rank-minimization

problem (RMP) in the engineering literature.

One important question raised in Guttman (1958) and Bekker and de Leeuw (1987), namely if $\boldsymbol{\Sigma}$ is any positive definite covariance matrix, how far can its rank be reduced by the appropriate choice of $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$, without regard to whether $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\left(\boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}\right)\boldsymbol{\Omega}$ holds? Guttman (1958) proves there are some cases where the minimum rank is $n-1$ and Bekker and de Leeuw (1987) states a more general theorem that gives necessary and sufficient conditions for a minimum rank of $n-1$. However, it is often possible to reduce the rank to $n-2$. Demonstrating so first requires a lemma.

Lemma 2. *If* $\mathbf{Z} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ *such that* $\mathbf{D}^{-1}$ *exists, then* $\mathcal{R}\left(\mathbf{Z}\right) = \mathcal{R}\left(\mathbf{D}\right) + \mathcal{R}\left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)$.

Theorem 3. *For any positive definite* $\boldsymbol{\Sigma}$*, the rank of* $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ *can "usually" be reduced to* $n-2$ *with an appropriate choice of* $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$.

Thus, in order to make a falsifiable claim about r, we must, at least, assume that $r < n-2$.

Whether (there is a positive probability that) the rank of $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ can be reduced below $n-2$ only if $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\left(\boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}\right)\boldsymbol{\Omega}$ is an interesting question. The closest thing to an answer is provided by Shapiro (1982), which shows that if $\widetilde{\boldsymbol{\Theta}}$ is diagonal but not required to be within $\mathcal{T}$, then the minimum-rank of $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ is almost surely greater than or equal to $L\left(n\right)$. We expect the restriction that $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$ would strengthen this result considerably but how much is unknown.

One reason why we have not been able to answer Guttman's (1958) question "To what extent can communalities reduce rank?" is because there is no feasible, general, analytical solution to the RMP, which is known to be NP-hard (see Fazel, Hindi and Boyd (2004)). Moreover, the rank function outputs an integer, so $\mathcal{R}\left(\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}\right) = n$ is locally flat in the interior of $\mathcal{T}$, implying that the RMP cannot be solved numerically. The fact that we never observe $\boldsymbol{\Sigma}$ is another problem and was the reason Thurstone himself believed his theorem to be merely theoretical.

The rest of this paper revisits the impossibility of solving the RMP and the inapplicability of the solution in a sample. It turns out that it is possible to indirectly solve the RMP by instead optimizing with respect to a more amenable function that has an equivalent optimum. Moreover, the algorithm remains useful when given an estimate of $\boldsymbol{\Sigma}$ from a sufficiently large sample.

4 Indirect Rank-Minimization Algorithm (IRMA)

The trick to solving the RMP 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 $\mathcal{R}\left(\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}\right) > r$, the computer does not know how "close" $\widetilde{\boldsymbol{\Theta}}$ is to a rank-minimizing solution or in which directions to move in $\widetilde{\boldsymbol{\Theta}}$-space. A computer needs more information to guide it toward a minimum-rank solution from any point in this n-dimensional parameter space.

The sizes of the eigenvalues implied by $\widetilde{\boldsymbol{\Theta}}$ constitute this additional information. In other words, we 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. In a sense that will be made precise in this section, we will find a minimum-rank solution if we choose $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$ to numerically maximize the "dispersion" of all the implied eigenvalues using a dispersion function that is sufficiently sensitive to all near-zero eigenvalues. In other words, we seek a particular objective function where the non-negativity constraint binds on the $r+1$th implied eigenvalue, regardless of the value of r. We call this procedure the indirect rank-minimization algorithm (IRMA).

We first develop the IRMA as it applies to the population and then consider how in fares in a sample. While the sample behavior is more relevant for applied researchers. Guttman (1977) argues that one must demonstrate the conditions under which a methodological technique is successful in the population before sampling behavior should even be considered:

> [t]he misplaced prestige of inference has been such that many researchers whose scientific problem requires only a loss function feel they must employ only the abstract theory of inference. For example, in trying to generalize Charles Spearman's scientific problem of a single-common factor, later investigators have developed something they call "maximum likelihood factor analysis". Actually, their mathematical machinery purposely fails to yield the maximum likelihood estimate of the number of common-factors — the rank number that is supposedly of basic interest to science. Maximum likelihood rank is automatically maximum rank, and those investigators do not want large rank. So, they do maximum likelihood on something other than

> rank, and use this as if it were inferential for determining rank. In effect, they debase maximum likelihood in an attempt to attain something resembling a loss function for their real concern. No reason is given for not doing direct data analysis based on a direct loss function. Nor do such investigators show that they are aware of the fact that their data analytic problem would remain even if there were no sampling error — if they had observed the population correlations at hand, and not sample estimates, so that there was no room for statistical inference in the first place. (82)

Guttman had extremely high standards for methodology, which we intend to meet in full. In contrast to the methods Guttman (1977) criticizes, the IRMA uses the dispersion as a loss (actually, gain) function, and the way in which dispersion is operationalized will be discussed below. In this section, we show that the IRMA finds a minimum-rank solution if given $\mathbf{\Sigma}$ and sampling is considered in Goodrich (2009).

4.1 Scale Invariance

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 $a \neq 1$, not only would Ω_{ii} change by a factor of a, the $\widetilde{\mathbf{\Theta}}$ that maximizes the eigenvalue dispersion could change in a complicated, nonlinear fashion. Optimizing with respect to a scale-dependent function is frowned upon, so we seek another matrix that is a function of $\widetilde{\mathbf{\Theta}}$ and satisfies three additional criteria: it 1) is not a function of $\mathbf{\Omega}$, 2) is PSD if and only if $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ is PSD, and 3) has the same rank as $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$. If such a matrix can be found, then we can safely maximize the dispersion of *its* eigenvalues, rather than the eigenvalues of $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$. Fortunately, two such matrices exist and are thoroughly discussed in the (somewhat dated) factor analysis literature.

The first candidate is called the "reduced correlation matrix corrected for attenuation" and is attributed to Kaiser and Caffrey (1965), which proved its scale-invariance, although it had been discussed much earlier. Here it is denoted $\mathbf{\Pi} = \operatorname{cor}(\mathbf{y}|\,\boldsymbol{\eta}) = \operatorname{diag}(\operatorname{cor}(\mathbf{y}) - \mathbf{\Theta})^{-\frac{1}{2}}(\operatorname{cor}(\mathbf{y}) - \mathbf{\Theta})\operatorname{diag}(\operatorname{cor}(\mathbf{y}) - \mathbf{\Theta})^{-\frac{1}{2}}$. We use $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$ to indicate the proposal for $\mathbf{\Pi}$ at $\widetilde{\mathbf{\Theta}}$. If $\widetilde{\Theta}_{ii} < 1\,\forall i$, then $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$ is well-defined and is just a rescaling of $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ to eliminate the dependence on $\mathbf{\Omega}$. Importantly, $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$ is PSD iff $\operatorname{cor}(\mathbf{y}) - \widetilde{\mathbf{\Theta}}$ is PSD and has the same rank, so Thurstone (1935) could have equivalently formulated the RMP in terms of $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$.

The second candidate stems from Guttman (1955) and Guttman (1956) and can be written as $\mathbf{\Delta} =$

$\text{cov}\left(\boldsymbol{\eta}|\,\mathbf{y}\right) = \mathbf{I}_r - \boldsymbol{\Phi}^{-1}$, where $\boldsymbol{\Phi}$ is a diagonal matrix containing the first r eigenvalues of $\boldsymbol{\Theta}^{-\frac{1}{2}}\text{cor}\left(\mathbf{y}\right)\boldsymbol{\Theta}^{-\frac{1}{2}}$. Although $\boldsymbol{\Delta}$ is only identified up to a rotation, its eigenvalues are unaffected by rotations, so for convenience we choose the rotation that renders $\boldsymbol{\Delta}$ diagonal. $\boldsymbol{\Delta}$ is positive-definite and is a function of $\boldsymbol{\Theta}$ but not $\boldsymbol{\Omega}$. Although $\boldsymbol{\Delta}$ is $r \times r$, $\boldsymbol{\Delta}\left(\widetilde{\boldsymbol{\Theta}}\right) = \mathbf{I}_n - \boldsymbol{\Phi}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-1}$ is $n \times n$ with minimum-rank r, where $\boldsymbol{\Phi}\left(\widetilde{\boldsymbol{\Theta}}\right)^{-1}$ is a diagonal matrix containing all n eigenvalues of $\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}\text{cor}\left(\mathbf{y}\right)\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}} = \widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}}\left(\text{cor}\left(\mathbf{y}\right) - \widetilde{\boldsymbol{\Theta}}\right)\widetilde{\boldsymbol{\Theta}}^{-\frac{1}{2}} + \mathbf{I}$. This expression implies $\boldsymbol{\Delta}\left(\widetilde{\boldsymbol{\Theta}}\right)$ is PSD iff $\text{cor}\left(\mathbf{y}\right) - \widetilde{\boldsymbol{\Theta}}$ is PSD and has the same rank, so Thurstone (1935) could have equivalently formulated the RMP in terms of $\boldsymbol{\Delta}\left(\widetilde{\boldsymbol{\Theta}}\right)$ had it been derived earlier.

To reemphasize, $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\boldsymbol{\Theta}\boldsymbol{\Omega} = \text{cov}\left(\mathbf{y}|\,\boldsymbol{\eta}\right)$ is scale-dependent, but $\boldsymbol{\Pi} = \text{cor}\left(\mathbf{y}|\,\boldsymbol{\eta}\right)$ and $\boldsymbol{\Delta} = \text{cov}\left(\boldsymbol{\eta}|\,\mathbf{y}\right)$ are scale-invariant. All three are PSD with rank r and depend on $\boldsymbol{\Theta}$, so we can conceptualize the RMP in terms of any of them. The substantive interpretations of the eigenvalues of $\boldsymbol{\Pi}$ and $\boldsymbol{\Delta}$ are interesting. Bentler (1968) interprets the former in terms of validity, while the latter sparked a highly-charged controversy over factor score indeterminacy in the 1970s (see Mulaik 2005 for a review). However, in this paper, we only care about $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ and $\boldsymbol{\Delta}\left(\widetilde{\boldsymbol{\Theta}}\right)$ insofar as they permit us to infer r by maximizing their eigenvalue dispersion.

The primary reason why we discuss both $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ and $\boldsymbol{\Delta}\left(\widetilde{\boldsymbol{\Theta}}\right)$ is to guard against various computational problems that may arise in practice due to the difficulty of the optimization problem (see the Computational Appendix). If we can find a minimum-rank solution indirectly, then both matrices will have the same rank at the optimum, namely r. Appearances to the contrary suggest that something is amiss, usually that the optimizer has gotten stuck on a local optimum rather than continuing to the global optimum. If, in addition, $r < L\left(n\right)$, we should find the same $\widehat{\boldsymbol{\Theta}}$ in both cases at the global optimum. Conversely, failing to find the same $\widehat{\boldsymbol{\Theta}}$ with a large sample gives us a hint that $r \geq L\left(n\right)$ in the population. Thus, it is always important to execute the IRMA for both $\boldsymbol{\Pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ and $\boldsymbol{\Delta}\left(\widetilde{\boldsymbol{\Theta}}\right)$.

4.2 Preliminary Theoretical Results

We use a lower-case letter subscripted by j to signify the jth *largest* eigenvalue of the corresponding matrix denoted with upper-case boldface and use tildes to distinguish proposals from population parameters. For example, $\widetilde{\pi}_j = \pi_j\left(\widetilde{\boldsymbol{\Theta}}\right)$ and $\widetilde{\delta}_j = \delta_j\left(\widetilde{\boldsymbol{\Theta}}\right)$ are admissible proposals for $\pi_j \in [0, n]$ and $\delta_j \in [0, 1)$, which are in turn the jth largest eigenvalues of $\boldsymbol{\Pi}$ and $\boldsymbol{\Delta}$. When taken as n-vectors with non-increasing elements, we will denote proposals for $\boldsymbol{\pi}$ and $\boldsymbol{\delta}$ as $\widetilde{\boldsymbol{\pi}} = \boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ and $\widetilde{\boldsymbol{\delta}} = \boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)$. When the distinctions between

$\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$ and $\mathbf{\Delta}\left(\widetilde{\mathbf{\Theta}}\right)$ are unimportant or when the distinctions between parameters and proposals thereof are unimportant, we will take x_j to be the jth largest eigenvalue of one of these matrices and similarly for $\mathbf{x}$. Sometimes we draw parallels to the literature on economic inequality where $\mathbf{x}$ is an n-vector of incomes.

When $\mathbf{x}$ is a function of $\widetilde{\mathbf{\Theta}}$, it is useful to characterize how the eigenvalues change as $\widetilde{\mathbf{\Theta}}$ changes.

Lemma 4. *If 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 $\widetilde{\mathbf{\Theta}}$. If $\widetilde{\pi}_j$ is simple, then $\frac{\partial \pi_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\left(\widetilde{\pi}_j - 1\right)\widetilde{P}_{ij}^2}{1-\widetilde{\Theta}_{ii}}$, where $\widetilde{\mathbf{P}}$ contains the orthonormal eigenvectors of $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$. Thus, $\frac{\partial \pi_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} \geq 0$ when $\widetilde{\pi}_j > 1$ and $\frac{\partial \pi_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} \leq 0$ when $\widetilde{\pi}_j < 1$. If $\widetilde{\delta}_j$ is simple, then $\frac{\partial \delta_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\left(\widetilde{\delta}_j - 1\right)\widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii}} \leq 0$, where $\widetilde{\mathbf{G}}$ contains the orthonormal eigenvectors of $\widetilde{\mathbf{\Theta}}^{-\frac{1}{2}}\mathbf{\Sigma}\widetilde{\mathbf{\Theta}}^{-\frac{1}{2}}$.*

Lemma 5. *If all eigenvalues are simple, then $\sum_{j=1}^n \frac{\partial \pi_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = 0$, and $\sum_{j=1}^n \frac{\partial \delta_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = -\left[\operatorname{cor}\left(\mathbf{y}\right)^{-1}\right]_{ii} < 0$. Thus, $\frac{\partial \overline{\pi}\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = 0$, and $\frac{\partial \overline{\delta}\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = -\frac{1}{n}\left[\operatorname{cor}\left(\mathbf{y}\right)^{-1}\right]_{ii}$.*

To summarize the implications of these lemmas (assuming all eigenvalues are simple, which generally is the case for a suboptimal proposal), as $\widetilde{\Theta}_{ii}$ increases holding the other diagonal elements of $\widetilde{\mathbf{\Theta}}$ constant, the following necessarily occurs: at least one above-average eigenvalue of $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$ *increases* while none decrease, at least one below-average eigenvalue of $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$ *decreases* while none increase, and all the changes in $\widetilde{\boldsymbol{\pi}}$ are *zero-sum* in light of the fact that $\overline{\pi} = 1$ cannot change. Thus, we might say that increasing $\widetilde{\Theta}_{ii}$ at the margin makes $\widetilde{\boldsymbol{\pi}}$ "more dispersed".

This dispersion idea can be formalized by utilizing the concept of majorization. We say that a vector $\mathbf{x}$ majorizes another vector $\mathbf{x}^*$ if and only if $\sum_{j=1}^k x_j \geq \sum_{j=1}^k x_j^* \, \forall k < n$ when both $\mathbf{x}$ and $\mathbf{x}^*$ are in non-increasing order and $\sum_{j=1}^n x_j = \sum_{j=1}^n x_j^*$. Since both $\boldsymbol{\pi}\left(\widetilde{\mathbf{\Theta}}\right)$ and $\boldsymbol{\pi}\left(\widetilde{\mathbf{\Theta}}^*\right)$ 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. Although $\boldsymbol{\delta}\left(\widetilde{\mathbf{\Theta}}\right)$ and $\boldsymbol{\delta}\left(\widetilde{\mathbf{\Theta}}^*\right)$ are ordered, they generally have different sums, so majorization is not relevant until we divide them by $\overline{\delta}\left(\widetilde{\mathbf{\Theta}}\right) = \frac{1}{n}\sum_{j=1}^n \delta_j\left(\widetilde{\mathbf{\Theta}}\right)$ and $\overline{\delta}\left(\widetilde{\mathbf{\Theta}}^*\right)$ respectively so that the sum is normalized to n in both cases. In other words, $\frac{\boldsymbol{\delta}\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{\delta}\left(\widetilde{\mathbf{\Theta}}\right)}$ may or may not majorize $\frac{\boldsymbol{\delta}\left(\widetilde{\mathbf{\Theta}}^*\right)}{\overline{\delta}\left(\widetilde{\mathbf{\Theta}}^*\right)}$, but at least they are comparable.

Majorization is important in this paper because the relevant vector of population eigenvalues has $n - r$ trailing zeros (as do all other minimum-rank solutions when $r \geq L\left(n\right)$) 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 $f(\mathbf{x}) > f(\mathbf{x}^*)$ when $\mathbf{x}$ majorizes $\mathbf{x}^*$. Also, $f(\mathbf{x})$ is said to be symmetric if it is invariant to a permutation of the order of the n elements of $\mathbf{x}$. 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 among $\frac{\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}$. Thus, we can use known results from the economics literature to help find a suitable function of eigenvalues to optimize. The potential of this approach is immediately suggested by the following

Theorem 6. *If $\widehat{\boldsymbol{\Theta}} \in \mathcal{T}$ and $\widehat{\boldsymbol{\pi}} = \boldsymbol{\pi}\left(\widehat{\boldsymbol{\Theta}}\right)$ majorizes $\widetilde{\boldsymbol{\pi}} = \boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ $\forall \widetilde{\boldsymbol{\Theta}} \in \mathcal{T} \neq \widehat{\boldsymbol{\Theta}}$, then $\widehat{\boldsymbol{\Theta}}$ is a minimum-rank solution that can be found by maximizing a Schur-convex, symmetric function of $\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$. A similar result holds for $\frac{\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}$.*

Corollary 7. *If $r = 1$, then $\boldsymbol{\pi}$ majorizes $\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ $\forall \widetilde{\boldsymbol{\Theta}} \in \mathcal{T} \neq \boldsymbol{\Theta}$, and $\frac{\boldsymbol{\delta}}{\bar{\delta}}$ majorizes $\frac{\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}$ $\forall \widetilde{\boldsymbol{\Theta}} \in \mathcal{T} \neq \boldsymbol{\Theta}$. Hence, $\boldsymbol{\Theta}$ maximizes a Schur-convex, symmetric function of either.*

Thus, if $\boldsymbol{\Sigma}$ were available and $r = 1$, then we could find the minimum-rank solution by maximizing *any* Schur-convex, symmetric function of $\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ or $\frac{\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}$. Of course, if $r = 1$, $\boldsymbol{\Sigma}$ were available, and all its elements were positive, we could find $\boldsymbol{\Theta}$ 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 we will now attempt to do with some help from the literature on economic inequality.

4.3 The Generalized Entropy Dispersion Function

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

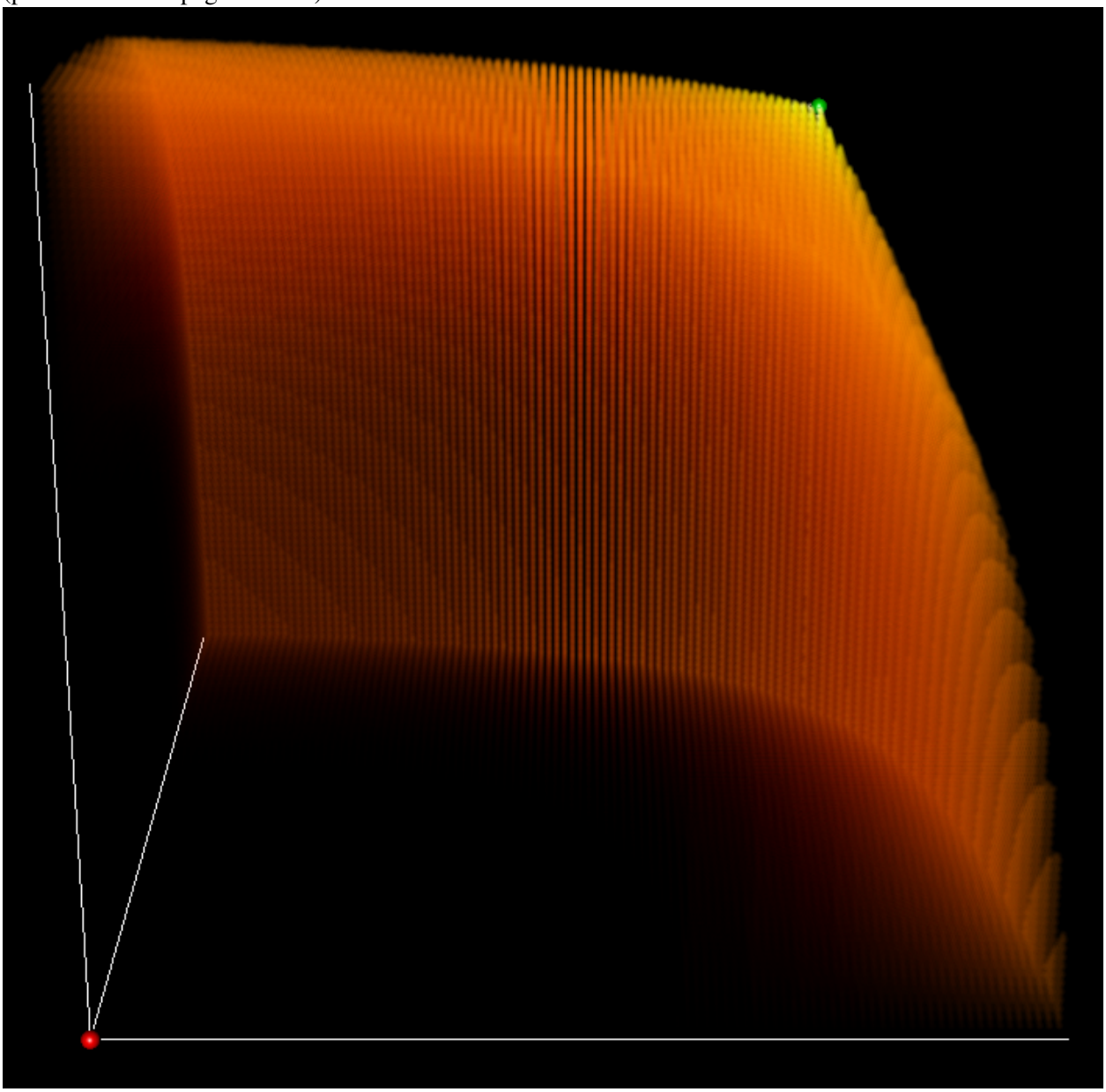
$$D_c(\mathbf{x}) = \begin{cases} \frac{1}{n}\sum_{j=1}^{n}\frac{x_j}{\overline{x}}\ln\left(\frac{x_j}{\overline{x}}\right) & \text{as } c \to 1 \text{ (Theil Index)} \\ \frac{1}{c(c-1)n}\sum_{j=1}^{n}\left[\left(\frac{x_j}{\overline{x}}\right)^c - 1\right] & \text{if } c \in (0,1) \\ \frac{1}{n}\sum_{j=1}^{n}\ln\left(\frac{\overline{x}}{x_j}\right)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

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

When $c = 1$, $D_1(\mathbf{x})$ is the Theil (1967) index, which is well-known in the income inequality literature, and when $c = 0$, $D_0(\mathbf{x})$ 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 $D_c(\mathbf{x})$ is the only continuous, additively decomposable, "relative" function of $\mathbf{x}$ that outputs zero if and only if $x_j = x\,\forall j$. "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 $D_c(\mathbf{x})$. Several properties of $D_c(\mathbf{x})$ are summarized in Cowell (2008, Appendix A). For example, if $c > 0$, then the upper bound of $D_c(\mathbf{x})$ is $\frac{n^{c-1}-1}{c(c-1)}$ — which approaches $\ln(n)$ from the right as $c \to 1$ — and this upper bound is reached if and only if $x_j = 0\,\forall j > 1$, i.e. the $r = 1$ case, which constitutes a simple proof of corollary 7.

Figure 1 shows $D_1\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)$ when $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$ for the special case where $n = 3$ and $r = 1$, which has a unique minimum-rank solution represented by the green dot in the upper right of the figure. The colors in figure 1 correspond to the value of $D_1\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)$ when $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ is PSD but rank deficient. Lighter colors correspond to higher values of $D_1\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)$. We know that the minimum-rank solution occurs at a point on

Figure 1: Visualization of $\widetilde{\boldsymbol{\Theta}}$-space Illuminated by $D_1\left(\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)$ when $n = 3$ and $r = 1$
(please view this page in color)



The three axes represent $\widetilde{\Theta}_{11}$, $\widetilde{\Theta}_{22}$, and $\widetilde{\Theta}_{33}$, so the red dot in the lower left is the origin. This cone represents the region where $\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Theta}}$ is PSD, and in its interior$\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Theta}}$ is positive definite. Outside the frontier of the cone, $\widetilde{\boldsymbol{\Theta}}$ is not in $\mathcal{T}$ because $\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Theta}}$ is not PSD. On the frontier, $\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Theta}}$ is PSD with at least one null eigenvalue. The green dot at the top right represents the point where $\widetilde{\boldsymbol{\Theta}} = \boldsymbol{\Theta}$, which implies $\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Theta}}$ has two null eigenvalues and one positive eigenvalue. The color of the frontier represents $D_1\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)$ with lighter colors indicating higher values. The lightest point corresponds to $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{D_1\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\right\} = \arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{D_1\left(\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\right\} = \boldsymbol{\Theta}$, which is the unique minimum-rank solution.

the frontier of $\mathcal{T}$ where two eigenvalues of $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ are null. Since the values of $D_1\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)$ are higher in the neighborhood of the minimum-rank solution and lower farther away from minimum-rank solution, figure 1 suggests we could find the minimum-rank solution by choosing $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$ to maximize $D_1\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)$. This intuition generalizes and can be made more precise by the following

Theorem 8. *For some sufficiently small* $c \in (0, 1]$, $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{D_c\left(\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\right\}$ *is a minimum-rank solution, and similarly for* $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{D_c\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\right\}$ *with perhaps a different critical value of* c.

Although this theorem is reassuring, for any particular value of $c > 0$, $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$ may not be a minimum-rank solution. In general, the value of c that renders $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$ a minimum-rank solution depends on n and r, among other things. However, we can make some further progress analytically.

Lemma 9. *If there are* k *positive elements in* $\mathbf{x}$, $D_1(\mathbf{x}) = D_1(x_1 \ldots x_k) + \ln\left(\frac{n}{k}\right)$.

This lemma implies that if $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_1(\mathbf{x})\}$ is a minimum-rank solution, then it is the minimum-rank solution among minimum-rank solutions with the maximum $D_1(x_1 \ldots x_r)$. We can state a condition that is sufficient but not necessary for $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_1(\mathbf{x})\}$ to be a minimum-rank solution, namely

Theorem 10. *As* $\frac{n}{r} \to \infty$, $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_1(\mathbf{x})\} = \boldsymbol{\Theta}$ *is the minimum-rank solution.*

However, this theorem 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 7 implies $c = 1$ is small enough, but in general this question must be answered with simulations.

4.4 Sampling Behavior

If c is sufficiently small, then $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$ would be a minimum-rank solution if given $\boldsymbol{\Sigma}$. Instead, we observe $\mathbf{S}$, 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 $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\boldsymbol{\Theta}\boldsymbol{\Omega}$. In other words, $\mathcal{R}(\mathbf{S} - \boldsymbol{\Omega}\boldsymbol{\Theta}\boldsymbol{\Omega}) > r$. Moreover, Shapiro (1982) shows that it is generally impossible to choose $\widetilde{\boldsymbol{\Theta}}$ to reduce the rank of $\mathbf{S} - \widetilde{\boldsymbol{\Theta}}$ to r; it is only possible to do so when given $\boldsymbol{\Sigma}$. However,

Theorem 11. *If* $c \in (0, 1]$ *is sufficiently small and* $\mathbf{S} \overset{\text{plim}}{\to} \boldsymbol{\Sigma}$, *then* $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\} \overset{\text{plim}}{\to} \widehat{\boldsymbol{\Theta}}$, *such that* $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widehat{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ *has minimum-rank. If, in addition,* $r < L(n)$, *then* $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\} \overset{\text{plim}}{\to} \boldsymbol{\Theta}$.

Thus, if $r < L(n)$, the IRMA consistently estimates $\boldsymbol{\Theta}$ if $c \in (0, 1]$ is sufficiently small. If $r \geq L(n)$, then $\boldsymbol{\Theta}$ is not identified and so the IRMA does not "estimate" $\boldsymbol{\Theta}$. However, if c is sufficiently small, then the difference between $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$ and *some* minimum-rank solution in the population becomes arbitrarily small as $N \to \infty$, allowing us to infer r at the optimum.

In a sense, this result justifies Guttman's (1977) approach to methodology. "All" that is needed is a continuous objective function that will yield the right answer in the population, and it will yield an increasingly close approximation to the right answer in a sample as the sample size increases. Moreover, there is nothing special about the distortions caused by sampling variation. If $\boldsymbol{\Sigma} \approx \boldsymbol{\Omega}(\boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}' + \boldsymbol{\Theta})\boldsymbol{\Omega}$ due to some minor violation of the LISREL model's assumption, then the IRMA will find a solution where the rank of $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widehat{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ is "approximately" r in the sense that more than r eigenvalues will be positive but only r of them will be "very positive". Since the eigenstructure of a matrix is reasonably robust to small changes to the matrix, we expect that the IRMA will be reasonably robust to sampling and to minor violations of the strict assumptions needed to justify it theoretically.

The dilemma revealed in the simulations to follow is that the smaller is c, the larger is the variance of finite-sample estimates of $\boldsymbol{\Theta}$ when $r < L(n)$. Thus, while one might want to use a very small value of c if $\boldsymbol{\Sigma}$ were available, doing so entails an efficiency cost when only $\mathbf{S}$ is available. However, this increase in sampling variance of the estimates is fairly small.

5 Monte Carlo Simulations

The previous section has shown analytically that there are circumstances in which $\widehat{\boldsymbol{\Theta}} = \arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$ is a minimum-rank solution when given $\boldsymbol{\Sigma}$ and asymptotic conditions under which $\widehat{\boldsymbol{\Theta}}$ finds a minimum-rank solution when given $\mathbf{S}$. 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.

5.1 Simulations with Random Populations

The first task is to assess how well $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\pi}} = \arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_c \left(\boldsymbol{\pi} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \right\}$ and $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}} = \arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_c \left(\boldsymbol{\delta} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \right\}$ correspond to a minimum-rank solution when given $\boldsymbol{\Sigma}$. Thus, we apply the IRMA to a representative selection of population matrices where all the assumptions of the two-equation LISREL model are satisfied.

Davies and Higham (2000) presents an algorithm to randomly draw a correlation matrix with specified eigenvalues, which, although not specifically mentioned in the article, has profound methodological implications for covariance structure analysis. In short, Davies and Higham (2000) makes it easy to draw a random $\boldsymbol{\Pi}$ once $\boldsymbol{\pi}$ is specified, where $\boldsymbol{\pi}$ has r positive eigenvalues that sum to n and $n - r$ null eigenvalues. Once a random $\boldsymbol{\Pi}$ is in hand, we draw each Θ_{ii} uniformally and independently to create a random $\boldsymbol{\Theta}$. Given a random $\boldsymbol{\Pi}$ and a random $\boldsymbol{\Theta}$, we can solve for the implied $\operatorname{cor}(\mathbf{y}) = (\mathbf{I} - \boldsymbol{\Theta})^{\frac{1}{2}} \boldsymbol{\Pi} (\mathbf{I} - \boldsymbol{\Theta})^{\frac{1}{2}} + \boldsymbol{\Theta}$ and feed it to any scale-invariant optimization procedure, such as the IRMA. Thus, we can experimentally manipulate n, r, $\boldsymbol{\pi}$, $\boldsymbol{\Theta}$ and c to evaluate the performance of the IRMA for both $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_c \left(\boldsymbol{\pi} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \right\}$ and $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_c \left(\boldsymbol{\delta} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \right\}$. Note that $\boldsymbol{\Pi}$ and $\boldsymbol{\Theta}$ randomly change for each simulation while n, r, and c are fixed over a set of simulations.

In this paper, we draw each Θ_{ii} from the $[0.05, 0.95]$ interval in most of the simulations. Some results where each Θ_{ii} is drawn from the $[0.05, 0.50]$ interval are reported to show that the IRMA performs even better when the measurement errors are smaller. The critical remaining question is what $\boldsymbol{\pi}$ to use to generate a random $\boldsymbol{\Pi}$. To avoid prejudicing the results, we construct the first r eigenvalues of $\boldsymbol{\Pi}$ by drawing uniformally from the unit simplex and multiplying the resulting r-vector by n. A random draw from a Dirichlet distribution with all its r parameters equal to unity is a draw from the appropriate unit simplex. This procedure is fairly conservative in the sense that researchers choose their variables non-randomly in an attempt to make π_r reasonably large and to make the first r eigenvalues of $\boldsymbol{\Pi}$ fairly equal. Thus, these random simulations may not be perfectly representative of the variables used in the literature, but the simulated performance of the IRMA should be no worse than its performance in the wild.

We first conduct simulations for 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 in the population, so it is easy to assess whether $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}$. If $r < L(n)$ and the eigenvalue dispersion at $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{ D_c(\mathbf{x}) \}$ is greater than the dispersion of the population eigenvalues, then c is too large, which we classify as a "logical error" regardless of how inconsequential it

is. In other words, a logical error occurs whenever $D_c(\widehat{\mathbf{x}}) > D_c(\mathbf{x})$ where $\mathbf{x}$ temporarily signifies the population eigenvalues and $\widehat{\mathbf{x}}$ signifies the implied eigenvalues at the optimum. An "optimization error" occurs when the dispersion of the population eigenvalues is greater than the eigenvalue dispersion at the optimum, which is to say that $D_c(\widehat{\mathbf{x}}) < D_c(\mathbf{x})$.

Strictly speaking, success only occurs when $D_c(\widehat{\mathbf{x}}) = D_c(\mathbf{x})$, which is impossible with floating point numbers. We are not worried about optimization errors that are only evident past the first few decimal places. Sometimes there is an egregious optimization error in the simulations, but egregious optimization errors can usually be avoided in practice by responsible researchers. First, researchers should always find both $\arg\max_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_c\left(\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\right\}$ and $\arg\max_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\left\{D_c\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\right\}$ and verify that they suggest the same r. It is unlikely that both would be compromised by an egregious optimization error. Second, simply running the IRMA a few times with different pseudo-random number generator seeds is usually sufficient to root out optimization difficulties. All that said, it is not feasible to be as meticulous in simulations as we would be in actual research situations. For this reason, we focus on the medians over the simulations to evaluate the IRMA, which presumably are better measures of central tendency than are the means because the mean is distorted by egregious optimization errors that would otherwise be preventable.

Since there are multiple minimum-rank solutions when $r \geq L(n)$, we need a metric to quantify how close $\operatorname{cor}(\mathbf{y}) - \widehat{\boldsymbol{\Theta}}$ is to *some* minimum-rank solution without regard to whether that minimum-rank solution is $\boldsymbol{\Theta}$. To do so, we take the root mean squared difference (RMSD) between the cells of $\operatorname{cor}(\mathbf{y})$ and $\widehat{\boldsymbol{\Lambda}\boldsymbol{\Upsilon}\boldsymbol{\Lambda}'} + \widehat{\boldsymbol{\Theta}}$ where the first term is calculated using the only first r eigenvalues of $\operatorname{cor}(\mathbf{y}) - \widehat{\boldsymbol{\Theta}}$. At any minimum-rank solution, this RMSD, denoted $m\left(\widehat{\boldsymbol{\Theta}}\right)$, is zero. When $m\left(\widehat{\boldsymbol{\Theta}}\right) > 0$, it could be due to logical or optimization errors, but we cannot be confident as to which error occurred unless $r < L(n)$.

Table 1 indicates the percentage of logical errors when $c = 1$ and $r < L(n)$. There main point is that logical failures are rare: about 0.50% for $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\pi}}$ and about 0.75% for $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}$ over the whole table. Moreover, $\arg\max_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\{D_1(\mathbf{x})\}$ is often quite close to $\boldsymbol{\Theta}$ even when $c = 1$ is too large. The median $m\left(\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\pi}}\right)$ among logical failures is about 0.0005, and the median $m\left(\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}\right)$ among logical failures is about 0.0007, both of which are tiny relative to the other sources of error that would be expected in any real research situation. When logical failures do occur, it is almost always the case that r is barely below $L(n)$, and the rth eigenvalue is very small. For example, in the worst case — namely, $n = 9$ and $r = 5$ — the median π_r among

Table 1: Percentage of Simulations Where $c = 1$ Is Too Large when $r < L(n)$

	$\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_1 \left(\boldsymbol{\pi} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \right\}$									$\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_1 \left(\boldsymbol{\delta} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \right\}$							
$r \rightarrow$ $n \downarrow$	2	3	4	5	6	7	8	9		2	3	4	5	6	7	8	9
5	5.5									7.0							
6	0									1.5							
7	0	0.5								0	5.5						
8	0	0.5	6.0							0	0	10.5					
9	0	0	1.0	11.0						0	0	3.0	20.0				
10	0	0	1.5	1.5						0	0	1.0	0.5				
11	0	0	0	0	2.5					0	0	0	0	2.0			
12	0	0	0	0	0	5.5				0	0	0	0	0.5	7.0		
13	0	0	0	0	0	0	5.5			0	0	0	0	0	0.5	7.5	
14	0	0	0	0	0	0	1.0	5.5		0	0	0	0	0	0	1.0	6.0
15	0	0	0	0	0	0	0	0.5		0	0	0	0	0	0	0	2.0

There were no logical failures for $16 \leq n \leq 19$ when $r \leq 9$. Blank cells indicate $r \geq L(n) = 0.5\left(2n + 1 - \sqrt{8n+1}\right)$, implying there is no unique minimum-rank solution.

logical failures was about 0.20 and the median δ_r among logical failures was about 0.11. In contrast, the rth eigenvalue among logical successes when $n = 9$ and $r = 5$ tended to be twice as big.

The logical failure rate drops to 0.13% for $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}$ when Θ_{ii} is restricted to the $[0.05, 0.50]$ interval rather than the $[0.05, 0.95]$ interval, although the results for $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\pi}}$ become slightly worse with a logical failure rate of 0.79%. Also, when $\Theta_{ii} \in [0.05, 0.95]\, \forall i$ the logical failure rate drops sharply for both $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\pi}}$ and $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}$ when c is smaller: As c decreases to 0.5, to 0.25, and to 0.1, the logical failure rate falls to about 0.20%, 0.10%, and 0.02% respectively for both $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\pi}}$ and $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}$.

While the logical failure rate is essentially zero for $c = 0.1$, the degree of optimization error is worse. When $c = 0.1$, the median $m\left(\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\pi}}\right)$ is about 0.0006, as compared to 0.000007 when $c = 1$. The precision for $\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}$ is essentially the same but slightly worse. The median $m\left(\widehat{\boldsymbol{\Theta}}\right)$ increases as n and r increase when $r < L(n)$ but at worst is zero through two decimal places. Thus, we can usually find the minimum-rank solution to reasonable precision when $r < L(n)$, and the precision is highest when $c = 1$.

It is more difficult to make strong conclusions when $r \geq L(n)$ because there are multiple minimum-rank solutions that the IRMA can gravitate toward. The few cases where $r = L(n)$ are interesting because the minimum-rank solutions are locally unique and because we can compare the results to "neighboring" simulations where there is a unique minimum-rank solution, either due to r being smaller by one or due to

Table 2: Simulated Performance when $r = L(n)$ as Compared to Similar Simulations where $r < L(n)$

n	r	$L(n)$	$D_1(\widehat{\boldsymbol{\pi}}) > D_1(\boldsymbol{\pi})$	Median $m\left(\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\pi}}\right)$	$D_1\left(\widehat{\boldsymbol{\delta}}\right) > D_1(\boldsymbol{\delta})$	Median $m\left(\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}\right)$
6	2	3.0	0	3×10^{-7}	1.5	4×10^{-8}
6	3	3.0	22.5	2×10^{-4}	31.5	2×10^{-4}
7	3	3.7	0.5	6×10^{-6}	5.5	2×10^{-5}
10	5	6.0	1.5	3×10^{-4}	0.5	6×10^{-4}
10	6	6.0	21.5	8×10^{-4}	27.5	1×10^{-3}
11	6	6.8	2.5	7×10^{-4}	2.0	1×10^{-3}
15	9	10.0	0.5	9×10^{-4}	2.0	1×10^{-3}
15	10	10.0	7.0	1×10^{-3}	10.0	2×10^{-3}
16	10	10.8	0.5	1×10^{-3}	2.0	2×10^{-3}
21	14	15.0	0	1×10^{-3}	0	2×10^{-3}
21	15	15.0	1.0	1×10^{-3}	2.5	1×10^{-3}
22	15	15.8	0	1×10^{-3}	0	2×10^{-3}

$D_1(\widehat{\mathbf{x}}) > D_1(\mathbf{x})$ signifies the percentage of simulations where the eigenvalue dispersion at the optimum is greater than the eigenvalue dispersion in the population. If $r = L(n)$, then this percentage is merely an upper bound on the logical errors because there are multiple minimum-rank solutions.

n being larger by one. These results are shown in table 2. For example, when $n = 6$, $r = 3 = L(n)$ and $c = 1$, the median $m\left(\widehat{\boldsymbol{\Theta}}\right)$ is about 0.0002, which is quite close to a minimum-rank solution but not nearly as close as when $n = 6$ and $r = 2$ or when $n = 7$ and $r = 3$. For larger values of n and r such that $r = L(n)$, the performance with respect to $m\left(\widehat{\boldsymbol{\Theta}}\right)$ is similar to "neighboring" simulations where $r < L(n)$.

There are many instances where the eigenvalue dispersion at the optimum is greater than the eigenvalue dispersion in the population, but some of these are situations where a "false" minimum-rank solution has more eigenvalue dispersion than the "true" minimum-rank solution implied by $\boldsymbol{\Theta}$. Thus, when $r = L(n)$, these numbers in the fourth and sixth columns of table 2 are only an upper bound for the percentage of logical errors when $c = 1$. This upper bound shrinks as the difference between n and r grows to the point where the upper bound is almost zero in the case where $n = 21$ and $r = 15 = L(n)$. A decent *lower* bound for the percentage of logical errors when $r = L(n)$ is the percentage of logical errors when n is one larger (holding r constant), which unfortunately leaves a very wide interval for all cases except the last.

Thus, there are an unknown number of situations where $r = L(n)$ and $c = 1$ is too large to find a

minimum-rank solution. However, even in those cases, $\arg\max_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\{D_1(\mathbf{x})\}$ may be extremely close to a minimum-rank solution. When $r = L(n)$ the median $m\left(\widehat{\boldsymbol{\Theta}}\right)$ in simulations where $D_1(\hat{\mathbf{x}}) > D_1(\mathbf{x})$ is *smaller* than the median $m\left(\widehat{\boldsymbol{\Theta}}\right)$ in simulations where $D_1(\hat{\mathbf{x}}) < D_1(\mathbf{x})$, which suggests that optimization errors are a greater concern in practice than are logical errors.

Next, we conduct simulations for all values of $2 \leq r \leq 9$ and $4 \leq n \leq 13$ such that $n - 1 > r > L(n)$ to judge how well the IRMA finds *some* minimum-rank solution when $c = 1$ but $\boldsymbol{\Theta}$ is not even locally identified. Table 3 reports the median over the simulations for $m\left(\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\pi}}\right)$ and $m\left(\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}\right)$, and a striking pattern is immediately apparent: As r becomes farther away from $L(n)$, $\arg\max_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\{D_1(\mathbf{x})\}$ gets closer to a minimum-rank solution. However, even in the worst case scenarios where $r \gtrapprox L(n)$, $\arg\max_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\{D_1(\mathbf{x})\}$ is usually a minimum-rank solution to at least three decimal places, as measured by $m\left(\widehat{\boldsymbol{\Theta}}\right)$.

To summarize, the IRMA with $c = 1$ works well in the population. When $r \approx L(n)$, $c = 1$ can work somewhat less well, in the sense that the maximum of $D_1(\mathbf{x})$ may not exactly coincide with a minimum-rank solution or the numerical precision may be slightly lessened. If $\boldsymbol{\Sigma}$ were available, a smaller value of c would make us extra confident that $\arg\max_{\widetilde{\boldsymbol{\Theta}}\in\mathcal{T}}\{D_c(\mathbf{x})\}$ is a minimum-rank solution. However, as shown in the next subsection, the lower is c, the larger is the variance of the sampling distribution of the parameters.

5.2 Simulations with Random Samples

Given that the IRMA with $c = 1$ works reasonably well in the population, the next question is how well it works when only $\mathbf{S}$ is available, rather than $\boldsymbol{\Sigma}$. In theory, the performance of the estimator in a sample is closely related to how quickly the sample covariance estimator converges in probability to the population covariance matrix as the sample size increases, so simulations are conducted for $N \in \{100, 300, 600, 1000\}$.

The procedure for generating $\mathbf{S}$ is as follows. First, in every simulation a random $\mathrm{cor}(\mathbf{y})$ is generated as before but only for combinations of n and r such that $r < L(n)$. Then, $\mathbf{S}$ is drawn from a Wishart distribution with $N - 1$ degrees of freedom and expectation $\mathrm{cor}(\mathbf{y})$. 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 simulational convenience. It also facilitates comparisons between the IRMA and the maximum likelihood exploratory factor analysis (ML) estimator, which assumes the data are multivariate normal.

Table 3: Root Mean-Squared Differences from a Minimum-Rank Solution when $n-1>r>L\left(n\right)$

	Median over the Simulations for $m\left(\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\pi}}\right)$							
$r\rightarrow$ $n\downarrow$	2	3	4	5	6	7	8	9
4	2×10^{-10}							
5		2×10^{-10}						
6			2×10^{-10}					
7			6×10^{-5}	2×10^{-10}				
8				9×10^{-5}	2×10^{-10}			
9					2×10^{-5}	2×10^{-10}		
10						4×10^{-5}	2×10^{-10}	
11						7×10^{-4}	4×10^{-5}	2×10^{-10}
12							6×10^{-4}	3×10^{-5}
13								6×10^{-4}
	Median over the Simulations for $m\left(\widehat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}\right)$							
$r\rightarrow$ $n\downarrow$	2	3	4	5	6	7	8	9
4	7×10^{-11}							
5		1×10^{-10}						
6			1×10^{-10}					
7			1×10^{-4}	1×10^{-10}				
8				4×10^{-5}	1×10^{-10}			
9					5×10^{-5}	2×10^{-10}		
10						2×10^{-5}	1×10^{-10}	
11						1×10^{-3}	2×10^{-5}	2×10^{-10}
12							9×10^{-4}	2×10^{-5}
13								1×10^{-3}

Blank cells in the lower triangles imply $r<L\left(n\right)$, while blank cells in the upper triangles imply $r\geq n-1$.

Table 4: Root Mean-Squared Error Comparison between IRMA and ML

$N \rightarrow$ $r \downarrow$	100	300	600	1000
2	6 0.151	6 0.096	5 0.098	5 0.074
3	7 0.192	8 0.112	8 0.075	8 0.069
4	9 0.201	10 0.126	9 0.101	10 0.068
5	11 0.208	11 0.141	12 0.094	12 0.074
6	13 0.231	14 0.137	14 0.104	14 0.073
7	14 0.238	16 0.149	16 0.105	16 0.084
8	16 0.247	17 0.160	18 0.108	18 0.087
9	18 0.258	> 19 0.158	> 19 0.115	> 19 0.090

The RMSE pertains to the diagonal of $\mathbf{\Lambda \Upsilon \Lambda}'$ when it is estimated with r latents at $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ D_1 \left(\boldsymbol{\pi} \left(\widetilde{\mathbf{\Theta}} \right) \right) \right\}$. The integers in the table body indicate the smallest value of n such that the mean RMSE of the IRMA is *smaller* than the mean RMSE of the ML estimator for a value of r and N, and the non-integers indicate the mean RMSE of the IRMA. When $r = 9$ and $N \geq 300$, the IRMA is preferable to the ML estimator for at least $n = 19$ and values of n greater than 19 were not tried. The number below the > 19 symbols is the mean RMSE for the IRMA when $n = 19$. When $r = 1$, the ML estimator is always preferable to the IRMA.

The root mean-squared error (RMSE) in estimating $\mathbf{\Theta}$ could be used to compare the IRMA and the ML estimator. However, this comparison is not that instructive because the ML estimator conditions on r while the IRMA is primarily intended to infer r, and its estimates of the diagonal elements of $\mathbf{\Theta}$ are biased downward in finite samples . We can obtain a somewhat more apples-to-apples comparison by calculating not the RMSE with respect to $\mathbf{\Theta}$, but the RMSE with respect to the diagonal elements of $\mathbf{\Lambda \Upsilon \Lambda}$, which are called the communalities in the factor analysis literature. In other words, we assume that the researcher automatically infers the correct value of r at the IRMA optimum, and we compare it to the MLE estimator that conditions on the correct value of r. This comparison is still not perfect because the ML estimator is given this valuable piece of information to use *during* the optimization, which is not the case for the IRMA.

The main result in table 4 is that the ML estimator outperforms the IRMA on this criterion when r is much less than $L(n)$, and the IRMA outperforms the ML estimator when r is barely less than $L(n)$. The top halves of the cells in table 4 indicate the smallest value of n such that the IRMA is better than the ML estimator for given values of r and N. The bottom halves report the mean RMSE values for that value of r, N, and n. When $N = 100$, the ML estimator is preferable to the IRMA unless $r \approx \frac{n}{2}$, and the RMSE is poor in absolute terms. For larger values of N, the critical value of n tends to rise, and the RMSE of the IRMA improves in absolute terms.

Although there is no clear winner with respect to RMSE, 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).

Next, we must compare the RMSE for the IRMA for different values of c. Recall that $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{D_1(\mathbf{x})\}$ has some difficulty with logical errors when $r \lessapprox L(n)$, which can be eliminated by making c smaller. So, what is the cost to using a small c? For $N = 100$ and all values of r and n such that $r < L(n)$, the RMSE of $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ D_1 \left(\boldsymbol{\pi} \left(\widetilde{\mathbf{\Theta}} \right) \right) \right\}$ is about the same between $c = 1$ and $c = 0.5$. Otherwise, for larger values of N, the RMSE is slightly lower when $c = 1$. This pattern continues for lower values of c. Table 5 investigates the situations where $r \lessapprox L(n)$ more closely, where there is no clear advantage in RMSE for $c = 1$ or $c = 0.5$. Thus, it does not seem to matter that much what value of c is used. Purists can use a lower c to be more confident that $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$ would be a minimum-rank solution if $\mathbf{\Sigma}$ were available. Pragmatists can use $c = 1$ to save a few tenths in RMSE. Anyone can try different values of c to see if the optimum is qualitatively affected, but that is unlikely. No one needs to worry about whether $c = 1$ is too large unless $L(n) - 1 \leq r \leq L(n) + 1$.

In summary, it appears that, if anything, the IRMA works better in practice than one might expect from theory. The primary concern from the previous section is that all the interesting results come with "for

Table 5: Root Mean-Squared Error Comparison between $c = 1$ and $c = 0.5$ when $r \lessapprox L(n)$

n	r	$N = 100$	$N = 300$	$N = 600$	$N = 1000$
5	2	0.167 0.171	0.111 0.112	0.098 0.086	0.074 0.080
7	3	0.192 0.179	0.120 0.132	0.099 0.111	0.081 0.083
8	4	0.230 0.203	0.150 0.151	0.119 0.127	0.104 0.115
9	5	0.229 0.225	0.167 0.166	0.127 0.127	0.112 0.119
11	6	0.238 0.243	0.159 0.169	0.130 0.127	0.111 0.113
12	7	0.246 0.239	0.165 0.170	0.129 0.147	0.112 0.126
13	8	0.254 0.252	0.167 0.180	0.137 0.143	0.116 0.128
14	9	0.255 0.260	0.180 0.186	0.145 0.153	0.124 0.130

The RMSE pertains to the diagonal of $\mathbf{\Lambda \Upsilon \Lambda}'$ when it is estimated with r latents at $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \left\{ D_c \left(\boldsymbol{\pi} \left(\widetilde{\mathbf{\Theta}} \right) \right) \right\}$.
The top number in each cell is for $c = 1$, while the bottom number is for $c = 0.5$.

sufficiently small c" caveat, which turns out to be very minor concern in these simulations. Even in the rare cases where $c = 1$ is too large, the optimum is usually so close to a minimum-rank solution in the population that sampling variation, rounding errors in the data, etc. swamp the potential impact of logical errors. Some may be surprised that the IRMA is competitive with the ML estimator in finite samples, especially when $r \lessapprox n$, despite the ML estimator conditioning on r and exploiting the assumed multivariate normality of the data. Of course, a large sample is always helpful, but the IRMA is quite viable regardless.

6 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 $\mathcal{R}\left(\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}\right) > 1 \, \forall \widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$. 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 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 about $r = 3$, then $\boldsymbol{\Theta}$ is identified.

Table 6 shows the main result, which is that $r = 4$ because the fourth eigenvalue of $\mathbf{S} - \arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_1 \left(\boldsymbol{\pi} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \right\}$ is 0.43 rather than (near) zero. While it is certainly possible that Esping-Andersen's (1990) three dimensions plus one minor dimension comprise the $r = 4$ inputs, it is difficult to make any conclusion about the nature of the r inputs until $\boldsymbol{\Theta}$ is identified. Although not shown, the results for $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_1 \left(\boldsymbol{\delta} \left(\widetilde{\boldsymbol{\Theta}} \right) \right) \right\}$ also suggest that $r = 4$.

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

Table 6: IRMA Results for Esping-Andersen's (1990) $n = 7$ Variables

Eigenvalues at optimum:	1.90	1.05	0.76	0.43	0.01	0.00	0.00

Results for $\arg\max_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}}\left\{D_1\left(\boldsymbol{\delta}\left(\widetilde{\mathbf{\Theta}}\right)\right)\right\}$ are qualitatively similar but quite different numerically, presumably due to the lack of identification of $\mathbf{\Theta}$, because the minimum rank appears to be four and $4 > L(7)$. $\mathbf{S}$ is obtained using a shrinkage covariance estimator.

Table 7: IRMA Results for $n = 13$ Variables

Eigenvalues of	2.77	1.90	1.25	0.71	0.52	0.26	0.16
$\mathbf{S} - \arg\max_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}}\left\{D_1\left(\boldsymbol{\pi}\left(\widetilde{\mathbf{\Theta}}\right)\right)\right\}$	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
$\arg\max_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}}\left\{D_1\left(\boldsymbol{\pi}\left(\widetilde{\mathbf{\Theta}}\right)\right)\right\}$	0.40	0.32	0.38	0.37	0.42	0.54	

Results for $\arg\max_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}}\left\{D_1\left(\boldsymbol{\delta}\left(\widetilde{\mathbf{\Theta}}\right)\right)\right\}$ are extremely similar because $\mathbf{\Theta}$ now appears to be identified. $\mathbf{S}$ is estimated using a shrinkage covariance estimator.

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 $L(13) \approx 8.4$, $\mathbf{\Theta}$ is identified as long as $r \leq 8$.

Table 7 shows the results, the most important of which is that $r \leq 8$. The ninth eigenvalue of $\mathbf{S} - \arg\max_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}}\left\{D_1\left(\boldsymbol{\pi}\left(\widetilde{\mathbf{\Theta}}\right)\right)\right\}$ is about 0.02, which is entirely consistent with it being zero in the population, given that the sample size is only 18. Also, the results for $\arg\max_{\widetilde{\mathbf{\Theta}}\in\mathcal{T}}\left\{D_1\left(\boldsymbol{\delta}\left(\widetilde{\mathbf{\Theta}}\right)\right)\right\}$ are extremely similar, strongly suggesting that $r < L(n)$. 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 $r > 3$ 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 easily 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

easily 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* 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 $\boldsymbol{\Theta}$ is diagonal implausible. However, it would be simple to utilize an autoregressive specification where $\boldsymbol{\Theta}$ 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 $\boldsymbol{\Theta}$ is not diagonal, but the fundamental principle of (indirect) rank-minimization remains appropriate.

7 Conclusion

Although this paper is quite long, we have made only one accomplishment, namely figuring out how to solve Thurstone's (1935) optimization problem. Instead of choosing $\widetilde{\boldsymbol{\Theta}}$ to minimize the rank of $\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Theta}}$ directly — which is essentially impossible — we choose $\widetilde{\boldsymbol{\Theta}}$ 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. Although this paper makes only one contribution to the literature, it is fundamental.

The recognition that the rank of $\boldsymbol{\Sigma} - \boldsymbol{\Theta}$ was equal to the number of factors was the pillar on which Thurstone (1935) built factor analysis with $r \geq 1$ factors as a generalization of Spearman's (1904) single-factor model. It is important to keep in mind that Thurstone's (1935) primary motivation was to challenge the single most influential theory in psychology for the previous three decades. 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 such as $\mathbf{\Theta} = \theta\mathbf{I}$ or as $n \to \infty$. Guttman (1954, 1956, 1958) did more to advance our understanding of the rank-minimization problem than 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 $\mathbf{\Delta} \neq \mathbf{I}_r$). 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 we could find a minimum-rank solution if we were 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 fine unless r is near the Ledermann (1937) bound. Also, it is probably possible to make further improvements to the optimization algorithm in order to reduce the already small number of cases where it fails to converge to the global optimum. The IRMA can now be used to investigate the open question of how far the rank of $\mathbf{\Sigma}$ can be reduced by $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$ when $\mathbf{\Sigma}$ is *not* generated according to a LISREL model with r latents.

These remaining questions should not stop anyone from using the IRMA today, even in a sample. Of course, the IRMA can be used to choose the number of factors in a factor analysis or more generally (with some full rank assumptions) to choose the number of latents in a LISREL model. In measurement and EITM models where the main hypothesis is that $r = 1$, the IRMA will likely be very fruitful because $c = 1$ is proven to be small enough if this hypothesis is true. Simply execute the IRMA and show that $r = 1$ is a reasonable conclusion at the optimum. The IRMA can be used as a precursor to a regression model to check whether r is less than or equal to the number of covariates in the model and — if $\mathbf{\Theta}$ is identified — whether the covariates are largely free of measurement error. Similarly, if $\mathbf{\Theta}$ is identified, the IRMA can be used to validate multiple imputation models. The key question is whether $\left[\mathbf{\Sigma}^{-1}\right]_{ii}^{-1} \approx \Theta_{ii}\, \forall i$, where the left-hand side is the error variance when the ith manifest variable is predicted by the other $n - 1$ manifest variables.

If this variance is (approximately) equal to the true error variance, then the missing values are being drawn from a conditional distribution that has the (approximately) correct variance. If not, then researchers should try to add variables to their multiple imputation models.

Finally, the IRMA can be used to aid experiments, particularly if multiple outcome variables are of interest and measured at baseline. Simply apply the IRMA to the union of the baseline outcomes and any available covariates and hope that $r < L(n)$. If so, generate scores on r factors, create matched pairs to achieve multivariate balance on the r factors, and randomly assign the treatment to one individual in each pair. If there are many background covariates, but they are measured with error and / or largely proxies for fewer concepts that are thought to affect outcomes, then it will presumably be easier to find balance on r factor scores than on the n observed variables. This same recipe can of course also be followed in imitations of experiments.

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* .

Abadir, K.M. and J.R. Magnus. 2005. *Matrix algebra*. Cambridge University Press.

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. 1968. "Alpha-maximized factor analysis (alphamax): its relation to alpha and canonical factor analysis." *Psychometrika* 33(3):335–345.

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`.

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.

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.

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

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.

Reiersøl, O. 1950. "On the identifiability of parameters in Thurstone's multiple factor analysis." *Psychometrika* 15(2):121–149.

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.

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

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.

Technical Appendix

Lemma from Abadir and Magnus (2005, exercise 5.48) on page 11

Lemma 1. *If* $\mathbf{Z} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ *such that* $\mathbf{D}^{-1}$ *exists, then* $\mathcal{R}(\mathbf{Z}) = \mathcal{R}(\mathbf{D}) + \mathcal{R}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})$.

Proof. Since $\mathcal{R}(\mathbf{Z}) = \mathcal{R}(\mathbf{EZF})$, where both $\mathbf{E} = \begin{bmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$ and $\mathbf{F} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}$ are unit tringular and thus have full rank, and $\mathbf{EZF} = \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}$, $\mathcal{R}(\mathbf{Z}) = \mathcal{R}(\mathbf{D}) + \mathcal{R}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})$. □

Falsifiability theorem on page 3

Theorem 2. *The rank of* $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ *can "usually" be reduced to* $n - 2$ *with the appropriate choice of* $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$.

Proof. Let $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega} = \mathbf{Z} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{bmatrix}$ where $\mathbf{D}$ is of order and rank $n - 3$. According to the previous lemma, $\mathcal{R}(\mathbf{Z}) = n - 3 + \mathcal{R}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}')$, where $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}'$ is of order 3. The diagonal of $\mathbf{A}$ is "mostly free" in the sense that it is only subject to the inequality restrictions implied by $\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}$. Hence given any admissible choice for the diagonal of $\mathbf{D}$, the diagonal of $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}'$ is also mostly free. As is well-known, the diagonal elements of a 3×3 symmetric matrix can "usually" be chosen to reduce its rank to one. Thus, the rank $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ can "usually" be reduced to $n - 2$. □

Eigenvalue derivatives on page on page 15

Lemma 3. *If 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 $\widetilde{\mathbf{\Theta}}$. If $\widetilde{\pi}_j$ is simple, then $\frac{\partial \pi_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\left(\widetilde{\pi}_j - 1\right)\widetilde{P}_{ij}^2}{1-\widetilde{\Theta}_{ii}}$, where $\widetilde{\mathbf{P}}$ contains the orthonormal eigenvectors of $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$. Thus, $\frac{\partial \pi_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} \geq 0$ when $\widetilde{\pi}_j > 1$ and $\frac{\partial \pi_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} \leq 0$ when $\widetilde{\pi}_j < 1$. If $\widetilde{\delta}_j$ is simple, then $\frac{\partial \delta_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\left(\widetilde{\delta}_j - 1\right)\widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii}} \leq 0$, where $\widetilde{\mathbf{G}}$ contains the orthonormal eigenvectors of $\widetilde{\mathbf{\Theta}}^{-\frac{1}{2}}\mathbf{\Sigma}\widetilde{\mathbf{\Theta}}^{-\frac{1}{2}}$.*

Proof. First, we seek the derivative of the jth eigenvalue of $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) = \mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)\left(\text{cor}\left(\mathbf{y}\right) - \widetilde{\mathbf{\Theta}}\right)\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)$ with respect to $\widetilde{\Theta}_{ii}$, where $\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right) = \text{diag}\left(\text{cor}\left(\mathbf{y}\right) - \widetilde{\mathbf{\Theta}}\right)^{-\frac{1}{2}}$. 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 $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) - \mathbf{I} = \mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)\left(\text{cor}\left(\mathbf{y}\right) - \mathbf{I}\right)\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)$.

de Leeuw (2007) summarizes several known results on the derivatives of parameterized eigenvalues, the most important of which in this case is that it $\widetilde{\pi}_j - 1$ is distinct, then $\frac{\partial \widetilde{\pi}_j - 1}{\partial \widetilde{\Theta}_{ii}} = \widetilde{\mathbf{p}}_j' \frac{\partial \mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) - \mathbf{I}}{\partial \widetilde{\Theta}_{ii}} \widetilde{\mathbf{p}}_j$, where $\widetilde{\mathbf{p}}_j'$ is the jth normalized eigenvector of $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) - \mathbf{I}$. 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 $\mathbf{D} = \mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)$ and $\mathbf{C} = \mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) - \mathbf{I}$, we can write $\frac{\partial \mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) - \mathbf{I}}{\partial \widetilde{\Theta}_{ii}} = \frac{\partial \mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)\left(\text{cor}(\mathbf{y}) - \mathbf{I}\right)\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)}{\partial\left(1-\widetilde{\Theta}_{ii}\right)^{-\frac{1}{2}}} \times \frac{\partial\left(1-\widetilde{\Theta}_{ii}\right)^{-\frac{1}{2}}}{\partial \widetilde{\Theta}_{ii}} = \frac{2\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)\left(\text{cor}(\mathbf{y}) - \mathbf{I}\right)\mathbf{J}^{ii}}{2\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{3}{2}}}$. Now, premultiply $\mathbf{J}^{ii}$ by $\mathbf{I}$ in the form of $\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)^{-1}$ so that $\frac{\partial \mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) - \mathbf{I}}{\partial \widetilde{\Theta}_{ii}} = \frac{\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)\left(\text{cor}(\mathbf{y}) - \mathbf{I}\right)\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)^{-1}\mathbf{J}^{ii}}{\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{3}{2}}} = \frac{\left(\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) - \mathbf{I}\right)\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)^{-1}\mathbf{J}^{ii}}{\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{3}{2}}}$.

Since $\mathbf{H}\left(\widetilde{\mathbf{\Theta}}\right)^{-1}\mathbf{J}^{ii} = \left(1 - \widetilde{\Theta}_{ii}\right)^{\frac{1}{2}}\mathbf{J}^{ii}$, $\frac{\partial \widetilde{\pi}_j}{\partial \widetilde{\Theta}_{ii}} = \widetilde{\mathbf{p}}_j' \frac{\left(\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) - \mathbf{I}\right)\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{1}{2}}\mathbf{J}^{ii}}{\left(1-\widetilde{\Theta}_{ii}\right)^{\frac{3}{2}}}\widetilde{\mathbf{p}}_j = \frac{\left(\widetilde{\mathbf{p}}_j'\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) - \widetilde{\mathbf{p}}_j'\right)\mathbf{J}^{ii}\widetilde{\mathbf{p}}_j}{1-\widetilde{\Theta}_{ii}}$. Recall from the definition of eigenvalues that $\widetilde{\mathbf{p}}_j'\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right) = \widetilde{\mathbf{p}}_j'\widetilde{\pi}_j$, which can be substituted into the previous expression to yield $\frac{\partial \widetilde{\pi}_j}{\partial \widetilde{\Theta}_{ii}} = \frac{\left(\widetilde{\pi}_j - 1\right)\widetilde{\mathbf{p}}_j'\mathbf{J}^{ii}\widetilde{\mathbf{p}}_j}{1-\widetilde{\Theta}_{ii}} = \frac{\left(\widetilde{\pi}_j - 1\right)\widetilde{P}_{ij}^2}{1-\widetilde{\Theta}_{ii}}$. Thus, $\frac{\partial \widetilde{\pi}_j}{\partial \widetilde{\Theta}_{ii}} \geq 0$ when $\widetilde{\pi}_j > 1$ and $\frac{\partial \widetilde{\pi}_j}{\partial \widetilde{\Theta}_{ii}} \leq 0$ when $\widetilde{\pi}_j < 1$.

To derive $\frac{\partial \delta_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}}$, we note that $\delta_j\left(\widetilde{\mathbf{\Theta}}\right) = 1 - \frac{1}{\phi_j\left(\widetilde{\mathbf{\Theta}}\right)}$ where $\phi_j\left(\widetilde{\mathbf{\Theta}}\right)$ is the jth largest eigenvalue of $\mathbf{\Phi}\left(\widetilde{\mathbf{\Theta}}\right) = \widetilde{\mathbf{\Theta}}^{-\frac{1}{2}}\text{cor}\left(\mathbf{y}\right)\widetilde{\mathbf{\Theta}}^{-\frac{1}{2}}$. The chain rule implies $\frac{\partial \delta_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\partial \widetilde{\phi}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} \times \frac{1}{\widetilde{\phi}_j\left(\widetilde{\mathbf{\Theta}}\right)^2}$, and we just need to derive $\frac{\partial \widetilde{\phi}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}}$. Proceeding as before, but now with $\mathbf{D} = \widetilde{\mathbf{\Theta}}^{-\frac{1}{2}}$ and $\mathbf{C} = \text{cor}\left(\mathbf{y}\right)$, we can write $\frac{\partial \widetilde{\phi}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \widetilde{\mathbf{g}}_j'\left(\frac{\partial \mathbf{\Phi}\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}^{-\frac{1}{2}}} \times \frac{\partial \widetilde{\Theta}_{ii}^{-\frac{1}{2}}}{\partial \widetilde{\Theta}_{ii}}\right)\widetilde{\mathbf{g}}_j = -\widetilde{\mathbf{g}}_j'\left(\frac{2\widetilde{\mathbf{\Theta}}^{-\frac{1}{2}}\text{cor}(\mathbf{y})\mathbf{J}^{ii}}{2\widetilde{\Theta}_{ii}^{\frac{3}{2}}}\right)\widetilde{\mathbf{g}}_j$ where $\widetilde{\mathbf{g}}_j$ is the jth normalized eigenvector of $\mathbf{\Phi}\left(\widetilde{\mathbf{\Theta}}\right)$. Again premultiplying $\mathbf{J}^{ii}$ by $\mathbf{I}$ in the form of $\widetilde{\mathbf{\Theta}}^{-\frac{1}{2}}\widetilde{\mathbf{\Theta}}^{\frac{1}{2}}$ and noting that $\widetilde{\mathbf{\Theta}}^{\frac{1}{2}}\mathbf{J}^{ii} = \widetilde{\Theta}_{ii}^{\frac{1}{2}}\mathbf{J}^{ii}$,

we obtain $\frac{\partial \widetilde{\phi}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\widetilde{\mathbf{g}}_j' \mathbf{\Phi}\left(\widetilde{\mathbf{\Theta}}\right) \mathbf{J}^{ii} \widetilde{\mathbf{g}}_j}{\widetilde{\Theta}_{ii}}$. Since $\mathbf{\Phi}\left(\widetilde{\mathbf{\Theta}}\right) = \widetilde{\mathbf{G}} \begin{bmatrix} \phi_1\left(\widetilde{\mathbf{\Theta}}\right) & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \phi_n\left(\widetilde{\mathbf{\Theta}}\right) \end{bmatrix} \widetilde{\mathbf{G}}'$, $\widetilde{\mathbf{g}}_j' \widetilde{\mathbf{G}}$ is a row-vector with zero everywhere except for unity in the jth cell, and $\widetilde{\mathbf{G}}' \mathbf{J}^{ii} \widetilde{\mathbf{g}}_j$ is a column vector with $\widetilde{G}_{ij}^2$ in its jth cell, $\frac{\partial \widetilde{\phi}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\widetilde{\phi}_j\left(\widetilde{\mathbf{\Theta}}\right) \widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii}}$. Thus, $\frac{\partial \delta_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = \frac{\partial \widetilde{\phi}_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} \times \frac{1}{\widetilde{\phi}_j\left(\widetilde{\mathbf{\Theta}}\right)^2} = \frac{\widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii} \widetilde{\phi}_j\left(\widetilde{\mathbf{\Theta}}\right)} = \frac{\left(\widetilde{\delta}_j\left(\widetilde{\mathbf{\Theta}}\right) - 1\right) \widetilde{G}_{ij}^2}{\widetilde{\Theta}_{ii}}$. Since $\widetilde{\delta}_j\left(\widetilde{\mathbf{\Theta}}\right) \in [0, 1)$, this derivative is non-positive. □

Sum of eigenvalue derivatives on page on page 15

Lemma 4. *If all eigenvalues are simple, then* $\sum_{j=1}^n \frac{\partial \pi_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = 0$, *and* $\sum_{j=1}^n \frac{\partial \delta_j\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = -\left[\mathrm{cor}\left(\mathbf{y}\right)^{-1}\right]_{ii} < 0$. *Thus,* $\frac{\partial \overline{\pi}\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = 0$, *and* $\frac{\partial \overline{\delta}\left(\widetilde{\mathbf{\Theta}}\right)}{\partial \widetilde{\Theta}_{ii}} = -\frac{1}{n}\left[\mathrm{cor}\left(\mathbf{y}\right)^{-1}\right]_{ii}$.

Proof. The trace of a matrix is also equal to the sum of its eigenvalues, and thus the derivative of the trace with respect to $\widetilde{\Theta}_{ii}$ is equal to the sum of the n eigenvalue derivatives with respect to $\widetilde{\Theta}_{ii}$. The first result follows from the fact that $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$ is a correlation matrix with ones on its diagonal regardless of the value of $\widetilde{\Theta}_{ii}$. Since the trace of $\mathbf{\Pi}\left(\widetilde{\mathbf{\Theta}}\right)$ is a constant, namely n, the sum of the eigenvalue derivatives with respect to $\widetilde{\Theta}_{ii}$ must be zero. The second result follows from Guttman (1956), which shows that the trace of $\mathbf{\Delta}\left(\widetilde{\mathbf{\Theta}}\right)$ is $n - \sum_{i=1}^n \widetilde{\Theta}_{ii}\left[\mathrm{cor}\left(\mathbf{y}\right)^{-1}\right]_{ii}$. Thus, the derivative of its trace with respect to $\widetilde{\Theta}_{ii}$ and the sum of the eigenvalue derivatives is $-\left[\mathrm{cor}\left(\mathbf{y}\right)^{-1}\right]_{ii} < 0$. Since the average eigenvalue of a matrix is equal to the ratio of its trace to n, the derivatives of the average eigenvalues follow from the derivatives of the traces. □

Majorization on page on page 16

Theorem 5. *If* $\widehat{\mathbf{\Theta}} \in \mathcal{T}$ *and* $\widehat{\boldsymbol{\pi}} = \boldsymbol{\pi}\left(\widehat{\mathbf{\Theta}}\right)$ *majorizes* $\widetilde{\boldsymbol{\pi}} = \boldsymbol{\pi}\left(\widetilde{\mathbf{\Theta}}\right)$ $\forall \widetilde{\mathbf{\Theta}} \in \mathcal{T} \neq \widehat{\mathbf{\Theta}}$, *then* $\widehat{\mathbf{\Theta}}$ *is a minimum-rank solution that can be found by maximizing a Schur-convex, symmetric function of* $\widetilde{\mathbf{\Theta}}$. *A similar result holds for* $\frac{\boldsymbol{\delta}\left(\widetilde{\mathbf{\Theta}}\right)}{\overline{\delta}\left(\widetilde{\mathbf{\Theta}}\right)}$.

Proof. Assume $\widehat{\mathbf{\Theta}} \in \mathcal{T}$ is not a minimum-rank solution. If so, then $\sum_{j=1}^r \pi_j\left(\widehat{\mathbf{\Theta}}\right) < n$, in which case $\widehat{\boldsymbol{\pi}}$ would not majorize $\boldsymbol{\pi}$, which contradicts the premise that $\widehat{\boldsymbol{\pi}}$ majorizes all admissible $\widetilde{\boldsymbol{\pi}} \neq \widehat{\boldsymbol{\pi}}$. Then, if $\widehat{\boldsymbol{\pi}}$ were inferior to any $\widetilde{\boldsymbol{\pi}} \neq \widehat{\boldsymbol{\pi}}$ with respect to a Schur-convex function, it would contradict the definition of a Schur-convexity, which proves the result. However, if there are multiple minimum-rank solutions, it is

possible that $\widehat{\boldsymbol{\pi}} \neq \boldsymbol{\pi}$. The same line of argument holds for $\frac{\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}$. □

$r = 1$ case on page on page 16

Corollary 6. *If $r = 1$, then $\boldsymbol{\pi}$ majorizes $\widetilde{\boldsymbol{\pi}} = \boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right) \forall \widetilde{\boldsymbol{\Theta}} \in \mathcal{T} \neq \boldsymbol{\Theta}$, and $\frac{\boldsymbol{\delta}}{\bar{\delta}}$ majorizes $\frac{\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)} \forall \widetilde{\boldsymbol{\Theta}} \in \mathcal{T} \neq \boldsymbol{\Theta}$. Thus, $\boldsymbol{\Theta}$ maximizes a Schur-convex, symmetric function of either.*

Proof. If $r = 1$ and $n \geq 3$, then there is almost surely a unique rank-minimizing solution. Furthermore, $\pi_1 = n \iff \pi_j = 0 \, \forall j > 1$, and $\frac{\delta_1}{\bar{\delta}} = n \iff \delta_j = 0 \, \forall j > 1$. Hence, $\boldsymbol{\pi}$ majorizes all admissible $\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right) \neq \boldsymbol{\pi}$, and $\frac{\boldsymbol{\delta}}{\bar{\delta}}$ majorizes all admissible $\frac{\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)} \neq \frac{\boldsymbol{\delta}}{\bar{\delta}}$. In this case, the maximum of any Schur-convex, symmetric function of $\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)$ or $\frac{\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}{\bar{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)}$ occurs uniquely at $\widetilde{\boldsymbol{\Theta}} = \boldsymbol{\Theta}$. □

Indirect rank-minimization on page on page 19

Theorem 7. *For some sufficiently small $c \in (0, 1]$, $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_c\left(\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\right\}$ is a minimum-rank solution, and similarly for $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{ D_c\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\right\}$ with perhaps a different critical value of c.*

Proof. Let $\mathbf{x}$ correspond to a minimum-rank solution and let $\mathbf{x}^*$ correspond to any admissible proposal with $\infty > k > r$ null eigenvalues. Contrarily assume that as $c \to 0^+$, $D_c(\mathbf{x}) \leq D_c(\mathbf{x}^*)$, implying

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} \ln\left(\frac{\overline{x}}{x_j}\right) &\leq \frac{1}{n}\sum_{j=1}^{n} \ln\left(\frac{\overline{x}^*}{x_j^*}\right), \\
\frac{1}{n}\sum_{j=1}^{r} \ln\left(\frac{\overline{x}}{x_j}\right) + \frac{n-r}{n}\ln\left(\frac{\overline{x}}{0}\right) &\leq \frac{1}{n}\sum_{j=1}^{k} \ln\left(\frac{\overline{x}^*}{x_j^*}\right) + \frac{n-k}{n}\ln\left(\frac{\overline{x}^*}{0}\right), \\
\ln(\infty)\left(\frac{n-r}{n} - \frac{n-k}{n}\right) &\leq \frac{1}{n}\sum_{j=1}^{k} \ln\left(\frac{\overline{x}^*}{x_j^*}\right) - \frac{1}{n}\sum_{j=1}^{r} \ln\left(\frac{\overline{x}}{x_j}\right) < \ln(k), \\
\ln(\infty)\left(\frac{k-r}{n}\right) &< \ln(k),
\end{aligned}$$

which is a contradiction that implies $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_0(\mathbf{x})\}$ is a minimum-rank solution. Since $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$ is continuous in c, it remains a minimum-rank solution in some neighborhood to the right of $c = 0$. □

Decomposition of Theil index on page on page 19

Lemma 8. *If there are k positive eigenvalues, $D_1(\mathbf{x}) = D_1(x_1 \ldots x_k) + \ln\left(\frac{n}{k}\right)$.*

Proof. First, recall that $0 \ln 0 = 0$. Thus, when calculating $D_1(\mathbf{x})$, it is only necessary to sum over the k positive eigenvalues, which is to say that $D_1(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} \frac{x_j}{\overline{x}} \ln\left(\frac{x_j}{\overline{x}}\right) = \frac{1}{n}\sum_{j=1}^{k} \frac{x_j}{\overline{x}} \ln\left(\frac{x_j}{\overline{x}}\right)$. Second, it is well-known that $D_c(\mathbf{x})$ is invariant to a rescaling of all the eigenvalues by a constant, a. To see this, simply note that $D_c(\mathbf{x})$ depends on the ratio of x_j to $\overline{x}$ and $\frac{x_j}{\overline{x}} = \frac{ax_j}{a\overline{x}} \forall j$. If $a = \frac{n}{k}$, then $D_1(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{k} \frac{\frac{n}{k}x_j}{\frac{n}{k}\overline{x}} \ln\left(\frac{\frac{n}{k}x_j}{\frac{n}{k}\overline{x}}\right) = \frac{1}{k}\sum_{j=1}^{k} \frac{x_j}{\frac{n}{k}\overline{x}} \left(\ln\left(\frac{x_j}{\frac{n}{k}\overline{x}}\right) + \ln\left(\frac{n}{k}\right)\right)$. Next, note that $\frac{n}{k}\overline{x} = \frac{1}{k}\sum_{i=1}^{k} x_i$, which is the mean over the k positive eigenvalues. Therefore, $D_1(\mathbf{x}) = D_1(x_1 \dots x_k) + \ln\left(\frac{n}{k}\right) \frac{1}{k}\sum_{j=1}^{k} \frac{x_j}{\frac{n}{k}\overline{x}} = D_1(x_1 \dots x_k) + \ln\left(\frac{n}{k}\right)$. □

Justification for $c = 1$ on page on page 19

Theorem 9. *As* $\frac{n}{r} \to \infty$, $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{D_1(\mathbf{x})\} = \mathbf{\Theta}$ *is the minimum-rank solution.*

Proof. Let $\mathbf{x}$ correspond to a minimum-rank solution and let $\mathbf{x}^*$ correspond to any admissible proposal with $k > r$ null eigenvalues. Assume that $D_1(\mathbf{x}) \leq D_1(\mathbf{x}^*)$, so the previous lemma implies $D_1(x_1 \dots x_r) + \ln\left(\frac{n}{r}\right) \leq D_1(x_1^* \dots x_k^*) + \ln\left(\frac{n}{k}\right)$. As $\frac{n}{r} \to \infty$, this inequality is false. There are four cases to consider:

Case 1. r and k fixed — Since $D_1(x_1 \dots x_r) < \ln(r)$ and $D_1(x_1^* \dots x_k^*) < \ln(k)$, they are irrelevant in the limit, leaving $\ln\left(\frac{n}{r}\right) > \ln\left(\frac{n}{k}\right)$.

Case 2. r and k increase at the same rate but at a slower rate than n — In that case, $D_1(x_1 \dots x_r)$ and $D_1(x_1^* \dots x_k^*)$ increase at the same rate and cancel, leaving $\ln\left(\frac{n}{r}\right) > \ln\left(\frac{n}{k}\right)$.

Case 3. r increases at a faster rate than k but at a slower rate than n — In that case, the left-hand side of the purported inequality increases at a faster rate than the right-hand side, so it is again false.

Case 4. r increases at a slower rate than k and both increase at a slower rate than n — In the limit, this case contradicts the premise that $k > r$.

These contradictions prove $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{D_1(\mathbf{x})\}$ is a minimum-rank solution in the limit as $\frac{n}{r} \to \infty$, and $\frac{n}{r} \to \infty \Longrightarrow r < L(n)$, implying the minimum-rank solution is unique. □

Consistency on page on page 19

Theorem 10. *If* $c \in (0,1]$ *is sufficiently small and* $\mathbf{S} \overset{\text{plim}}{\rightarrow} \mathbf{\Sigma}$*, then* $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\} \overset{\text{plim}}{\rightarrow} \widehat{\mathbf{\Theta}}$*, such that* $\mathbf{\Sigma} - \mathbf{\Omega}\widehat{\mathbf{\Theta}}\mathbf{\Omega}$ *has minimum-rank. If, in addition,* $r < L(n)$*, then* $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\} \overset{\text{plim}}{\rightarrow} \mathbf{\Theta}$.

Proof. The result follows from the Slutsky theorem, which says that if $\mathbf{A}$ is a random variable and $g(\mathbf{A})$ is a continuous function that does not depend on the sample size, then $\text{plim}\, g(\mathbf{A}) = g(\text{plim}\mathbf{A})$. Let $\mathbf{A} = \mathbf{S}$ and let $g()$ be $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$. Under the premise that $\mathbf{S} \overset{\text{plim}}{\rightarrow} \mathbf{\Sigma}$, $\text{plim}\, g(\mathbf{A})$ is whatever it would be if $\mathbf{\Sigma}$ were available, namely a minimum-rank solution in the population when c is sufficiently small, which is unique if $r < L(n)$. □

Computational Appendix

Finding $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$ is a non-trivial problem, even if c is small enough to make it a minimum-rank solution. Although $D_c(\mathbf{x})$ is a continuous (and, for what it is worth, differentiable) 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, as illustrated in figure 1. 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 $\arg\max_{\widetilde{\mathbf{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$. We must use a "gradient-free" optimization algorithm that only depends on the values of the objective function, and only one algorithm, which has been tweaked for this purpose, has been found to yield acceptable performance.

In doing so, we must respect the admissibility constraints on $\widetilde{\mathbf{\Theta}}$. One approach to enforcing the constraint that $\widetilde{\mathbf{\Theta}} \in \mathcal{T}$ is to return some negative number if $\widetilde{\mathbf{\Theta}} \notin \mathcal{T}$ since $D_c(\mathbf{x})$ is non-negative but only defined when the admissibility constraints are satisfied. 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 formed by the n-dimensional cone of matrices where $\mathbf{\Sigma} - \mathbf{\Omega}\widetilde{\mathbf{\Theta}}\mathbf{\Omega}$ is PSD, as seen in figure 1.

Instead, we utilize 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

$\text{cor}\,(\mathbf{y}) - \dot{\boldsymbol{\Theta}}$. Irwin (1966) proves that if $\widetilde{\boldsymbol{\Theta}}$ is defined in this way, then $\boldsymbol{\Sigma} - \boldsymbol{\Omega}\widetilde{\boldsymbol{\Theta}}\boldsymbol{\Omega}$ is PSD and singular. While there is still a possibility that $\widetilde{\boldsymbol{\Theta}}$ is not PSD, returning some negative number in that case does not cause any of the aforementioned difficulties. This parameterization of $\widetilde{\boldsymbol{\Theta}}$ entails a severe performance penalty because the objective function must calculate eigenvalues twice, the first time to calculate $\dot{\lambda}_n$ at $\dot{\boldsymbol{\Theta}}$ and the second time to calculate $\mathbf{x}$ at $\widetilde{\boldsymbol{\Theta}} = \dot{\boldsymbol{\Theta}} + \dot{\lambda}_n \mathbf{I}$. However, doing so tends to find $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$ 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 here. Maximizing a function in $\dot{\boldsymbol{\Theta}}$-space is daunting because the transformation is not unique — both $\dot{\boldsymbol{\Theta}}$ and $\dot{\boldsymbol{\Theta}} + a\mathbf{I}$ transform to the same $\widetilde{\boldsymbol{\Theta}}$ — which implies plateaus in $\dot{\boldsymbol{\Theta}}$-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). A genetic algorithm works by creating many (1000 by default) proposed vectors of free parameters and using reproductive rules that mimic the process of evolution to eventually find the parameters that are most "fit" with respect to the objective function. 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 $\widetilde{\boldsymbol{\Theta}}$-space. This modification facilitates the convergence of the population to $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$ because the population has greater fitness in $\widetilde{\boldsymbol{\Theta}}$-space than in $\dot{\boldsymbol{\Theta}}$-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 $\mathcal{T}$.

As can be seen from the generally small values of $m\left(\widehat{\boldsymbol{\Theta}}\right)$ reported in the body of this paper, RGENOUD is generally successful at getting extremely close to $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \{D_c(\mathbf{x})\}$. However, it is still possible to experience an optimization error, so researchers should use proper precautions. In particular, use population sizes of at least 1000, find both $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{D_c\left(\boldsymbol{\pi}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\right\}$ and $\arg\max_{\widetilde{\boldsymbol{\Theta}} \in \mathcal{T}} \left\{D_c\left(\boldsymbol{\delta}\left(\widetilde{\boldsymbol{\Theta}}\right)\right)\right\}$, perhaps a few times with different pseudo-random number seeds.