Journal of Data Science 14(2016), 519-538

Fitting Exploratory Factor Analysis Models with High Dimensional Psychological Data

W. Holmes Finch, Maria E. Hernández Finch

Department of Educational Psychology, Ball State University, Muncie, Indiana, USA

Abstract: Objectives: Exploratory Factor Analysis (EFA) is a very popular statistical technique for identifying potential latent structure underlying a set of observed indicator variables. EFA is used widely in the social sciences, business and finance, machine learning, and the health sciences, among others. Research has found that standard methods of estimating EFA model parameters do not work well when the sample size is relatively small (e.g. less than 50) and/or when the number of observed variables approaches the sample size in value. The purpose of the current study was to investigate and compare some alternative approaches to fitting EFA in the case of small samples and high dimensional data. Results of both a small simulation study, and an application of the methods to an intelligence test revealed that several alternative approaches designed to reduce the dimensionality of the observed variable covariance matrix worked very well in terms of recovering population factor structure with EFA. Implications of these results for practice are discussed.

Key words: Exploratory Factor Analysis, High Dimensional Data

1. Introduction

Exploratory factor analysis (EFA) is a popular tool in many areas of research for gaining insights into the latent structure underlying observed data. Areas in which EFA is commonly used include the social sciences, computer science, business and finance, and health care, to name only some. The standard EFA model can be expressed as

$$Y = \Lambda\xi + \Psi \tag{1}$$

Where

Y =Matrix of observed indicator variables

ξ =Matrix of factor(s)

Λ =Matrix of factor loadings relating indicators to factor(s)

Ψ =Matrix of unique random errors associated with the observed indicators

Of particular interest in the context of EFA is Λ, which links the indicator variables to the latent factors. The eigenvalues associated with the model are also important for interpreting EFA results, because they can be used to ascertain the number of factors that the researcher should retain [1]. There are a number of methods available for extracting factor loadings, including maximum likelihood estimation (MLE), principal axis factoring, alpha factoring, and image factoring, to name only a few. After the initial factors are extracted, they are typically rotated in order to improve interpretability of the results through the mathematical encouragement of simple structure in Λ [1]. Simple structure simply means that each indicator is only associated with a single factor. With respect to the eigenvalues, a number of methods have been suggested for determining the optimal number of factors to retain,

including the eigenvalue greater than 1 rule, the scree plot, and parallel analysis. Of these, the latter is generally considered one of the best techniques available for determining the number of factors [2].

As with all statistical models, proper estimation of EFA parameters, such as factor loadings, requires a sufficiently large sample size (N) for the results to be accurate and efficient [3]. Despite the need for relatively large samples however, in many situations researchers are unable to obtain such samples, as for example when working with low incidence populations or specialized interventions that can only be used with a small number of individuals at a time. With such small populations, there may not be available what would generally be considered a sufficient number of individuals for inclusion in the sample to be used with EFA. The purpose of this study is to describe and compare several methods for estimating EFA models with small samples. The paper is organized as follows: First, there is a description of prior research on the performance of commonly used EFA estimation techniques with small sample sizes, then each of the alternative methods to be studied here is described, followed by a description of the study goals and hypotheses. The methods and results are presented next, after which the results discussed and suggestions for application in practice are made.

EFA and High Dimensional Data

Many recommendations exist for the minimum N necessary to accurately fit an EFA model, or for the N to number of indicators (p) ratio. These recommendations have been shown to be inconsistent, and to not account for important issues that influence results; e.g., magnitude of factor loadings; [4]. Moreover, approximately 40% of published studies had N/p ratios of less than 5 [3], and frequently involved weak factor loadings in conjunction with small samples [5]. Thus, dependable EFA methods are needed for high dimensional (small samples coupled with a relatively large number of observed indicator variables) data situations.

A number of studies have investigated the performance of standard EFA model estimation techniques with small samples, and have generally found that generally speaking accurate recovery of the latent variables requires that, at a minimum the sample size (N) be larger than the number of observed indicator variables (p), and preferably that the sample consist of minimally 50 or more individuals [6]. However, research has also shown that if the relationships between the indicators and the latent factors are large, as evidenced through the presence of large factor loadings in (1), then samples as small as 10 may accurately recover the latent structure underlying a set of observed indicators, providing that the number of indicators does not exceed the sample size [6-7]. When p exceeds N, the covariance matrix used in MLE is singular, and estimation of the parameters in (1) is not possible [8]. In addition, researchers [9] found that common estimators such as MLE have difficulty converging when N is only slightly larger than p. Thus, alternative methods for estimating the EFA model would seem to be necessary for researchers faced with high dimensional data, in which p>N, or N is barely larger than p. Following is a description of several methods that have been proposed for this purpose, followed by a comparison of their relative effectiveness at retaining the latent factor structure underlying the observed data.

Methods for high dimensional factor analysis

A number of statistical methods have been suggested for use in exploratory latent variable modeling with high dimensional datasets. These approaches have generally been applied in the areas

of bioinformatircs, financial trading, and machine learning, but not in the areas of educational and psychological research. Nor have several of these methods been compared to one another in the same study. Though very different with respect to the estimation algorithms used, these approaches all have in common the goal of identifying only those indicators that are most important with respect to defining the latent structure underlying the data, and down weighting or completely removing the others, in order to make the fitting of the factor model more tenable given a small sample size.

Graphical LASSO

One method for estimating EFA models with high dimensional data is based upon L1 regularization, or the LASSO penalty. The LASSO is widely used in the regression context for high dimensional data, when the number of independent variables approaches, or surpasses, the sample size [10], and has been widely used in a variety of contexts [11-14]. This methodology has as an overarching goal the reduction of data complexity by setting to 0 parameters that may be small, thereby creating a relatively sparse parameter space that is more amenable to the small sample size. In the context of EFA, the sparse covariance would be estimated using the LASSO penalty, after which the EFA model would be fit to the resulting sparse matrix. The method for obtaining the lasso estimator of the covariance matrix used here, known as the graphical Lasso (GLASSO) was first proposed by [15], and relies on a coordinate descent algorithm. is the data are assumed to be multivariate normal with p observed indicator variables, a sample size of N, mean vector μ, and covariance matrix Σ. S is the sample estimate of Σ, and Σ^{-1} is the inverse of the covariance matrix. The GLASSO algorithm has as its goal the maximization of the function

$$\log\det\Sigma^{-1} - \text{tr}(S\Sigma^{-1}) - \rho\|\Sigma^{-1}\|_1 \quad (2)$$

Where

$\|\Sigma^{-1}\|_1$ = Sum of the absolute values of the elements of Σ^{-1}, which is the L_1 norm.

ρ =Regularization parameter, in this case the L_1.

In order to estimate the parameters that maximize the quantity in (2), [15] suggested the use of the following algorithm.

1. Start with $W = S + \rho I$ (3)

Where

I =Diagonal matrix

2. For each variable j=1, 2, …, p solve the following equation:

$$\min_\beta \left\{\frac{1}{2}\left\|W_{11}^{1/2}\beta - b\right\|^2 + \rho\|\beta\|_1\right\} \quad (4)$$

where

$b = W_{11}^{-1/2}s_{12}$

For β that minimizes equation (4), there is a value $w_{12} = W_{11}\beta$ that solves the following:

$w_{12} = \text{argmin}_y\{y^T W_{11}^{-1} y: \|y - s_{12}\| \leq \rho\}$

3. Fill in the corresponding row and column of W with $w_{12} = W_{11}\beta$

4. Continue the algorithm until convergence.

Essentially, the steps above involve regressing of each of the p variables onto the others, using the LASSO penalty, whereby each variable involved in the covariance matrix will serve as both independent and dependent variables, and the equations are linked with one another because they share a common W. The estimation of β in step 2 is based on coordinate descent using a soft threshold operator, as below.

$$\hat{\beta}_j \leftarrow S\left(s_{12} - \sum_{k \neq j} W_{kj}\,\hat{\beta}_k, \rho\right)/W_{jj} \quad (5).$$

Model terms in (5) are as defined above, with the addition of the soft threshold operator:

$$S(x, t) = \text{sign}(x)(|x| - t)_+ \quad (6).$$

Where
x=Value of predictor variable
t =Fixed threshold value of 0.001 by default, though the value can be set by the researcher.

The resulting covariance matrix W should be very sparse (contain relatively few non-0 values) in comparison with the original covariance matrix S. Thus, when it is applied to the EFA algorithm of choice, (e.g. maximum likelihood) the number of parameters to be estimated will be relatively small when compared to the model implied by the original covariance matrix. This fact in turn means that a smaller sample should be needed in order to obtain reasonable estimates for factor loadings and eigenvalues. In short, GLASSO reduces the number of covariances that are involved in estimating the factor loadings by setting those to 0 in W that are small in the original covariance matrix S. It is then W upon which EFA is conducted.

Researchers [15] conducted a small simulation study to examine the performance of GLASSO in terms of its processing speed, and found that it was faster than other similar approaches to estimating a covariance matrix, such as that proposed by [12]. In addition, they demonstrated the use of GLASSO with high dimensional cell signaling data taken from [16]. However, to our knowledge no research has examined the utility of GLASSO with high dimensional data in the context of EFA. Thus, given its clear potential advantages for such sparse data, it is included in the current study.

Principal Orthogonal Component Thresholding (POET)

A second alternative to estimating a covariance matrix in the presence of high dimensional data that is examined in the current study is Principal Orthogonal Component Threshold (POET), which was described by [8]. The GLASSO approach belongs to a set of methods that make the tacit assumption that many elements of Σ are essentially 0, although in reality this may not be the case, leading to parameter estimation bias if the covariance matrix is made too sparse [8]. One approach to dealing with this problem, POET, was proposed by [9], and was later expanded on by [8]. POET is based upon an assumption of conditional sparsity, which asserts that conditioning on a small number of common components, the observed indicator variables will have small covariances with one another [8]. Rather than setting a large portion of the covariances to 0, as is the case with GLASSO, POET

conditions the observed variable covariance matrix on a small number of components that are believed to underlie the data., through a singular value decomposition of the observed covariance matrix, keeping the covariance matrix implied by the first K principal components, and then applying a thresholding procedure to the portion of the covariance matrix that remains, yielding the POET estimator.

Given a principal components solution for a covariance matrix, the complete sample covariance matrix can be defined as

$$\Sigma = \sum_{i=1}^{K} \hat{\lambda}_i \xi_i \xi_i' + R_K \quad (7)$$

Where

$\hat{\lambda}_i$ =eigenvalue for component i (i.e. 1-*K*).

ξ_i =Eigenvector for eigenvalue *i*.

K=Number of retained components

$$R_K = \sum_{i=K+1}^{p} \hat{\lambda}_i \xi_i \xi_i'$$

Thus, R_K is the portion of the complete observed covariance matrix that is the complement of the part of the matrix that is implied by the retained components, as described by their eigenvalues and eigenvectors. In order to limit the size of the covariance matrix, [8] recommended applying a threshold to R_K, under the presumption that most, though not all, of the useful information in Σ is accounted for by the K components retained in (7). This thresholding takes the form

$$R_K^T = r_{ij}^T, r_{ij}^T = \begin{Bmatrix} r_{ii}, & i = j \\ s_{ij(r_{ij})I(|r_{ij}| \geq \tau_{ij})} & i \neq j \end{Bmatrix} \quad (8)$$

Where

r_{ij} =Element ij of R_K

s_{ij} =Generalized shrinkage function of [17]

τ_{ij} =Threshold value determining whether r_{ij} is part of the final covariance matrix, or set to 0.

The threshold τ_{ij} can be either a hard constant value that is the same for all elements of R_K, δ or an adaptive soft value that is estimated from the data. In the latter case, τ_{ij} for the covariance matrix element r_{ij} is estimated as

$$\tau_{ij} = \left(r_{ii} r_{jj}\right)^{1/2} \quad (9)$$

Where

r_{ii} =ith diagonal element of R_K

r_{jj} =jth diagonal element of R_K.

Whether τ_{ij} is a hard or soft value, it is used to identify the elements of R_K to retain, and those to constrain to 0. Thus, the POET estimator of the covariance matrix can then be defined as

$$\Sigma_{POET} = \sum_{i=1}^{K} \hat{\lambda}_i \xi_i \xi_i' + R_K^T \quad (10)$$

When $\tau_{ij} = 0$ then Σ_{POET} is simply the standard estimated sample covariance matrix. In contrast, when $\tau_{ij} = 1$ then Σ_{POET} is equivalent to the covariance matrix implied by the K retained components. The calculation of the soft adaptive value of τ_{ij} appears in (9). The optimal hard value, δ can be determined using cross-validation based on the jackknife [8]. The performance of both methods is examined in the current study.

In order to investigate the performance of POET, [8] conducted a small simulation study in which they examined the ability of the method to accurately recover elements of the population covariance matrix for varying sample sizes, and numbers of variables. The results showed that when the ratio of number of indicators to sample size was less than 1 (i.e. there were more variables than individuals in the sample), POET was able to yield an accurate reproduction of the population covariance matrix, if the total sample size approached 50, even when the covariances were relatively small in value. This latter point is very important, as it corresponds to the case when factor loadings would be expected to be small in value, a condition which [5] showed could lead to poor factor recovery when combined with small sample sizes. Thus, although this earlier work with POET was not directly related to its application to EFA, it does appear to hold promise for use in this context with small sample sizes, given the accurate recovery of the population covariance matrix. In the current study, POET was used to estimate a sparse covariance matrix for the observed indicators, and then EFA using MLE was applied to this sparse matrix.

Sparse Estimation via Nonconcave Penalized Likelihood in Factor Analysis Model (FANC)

Researchers [18] have proposed another alternative for estimating factor analysis models in the context of high dimensional data. Their approach involves the use of the minimax convex penalty function (MC+) by [19]. Authors [18] proposed estimating the parameters of the factor model in (1) through the maximization of the penalized log-likelihood function

$$(\Lambda, \Psi) = \text{argmax}\, l_p^{ort}(\Lambda, \Psi) \quad (11)$$

Where

$$l_p^{ort} = l^{ort}(\Lambda, \Psi, \Phi) - N \sum_{i=1}^{P} \sum_{j=1}^{m} \rho P\left(|\lambda_{ij}|\right) \quad (12)$$

$l^{ort}(\Lambda, \Psi, \Phi)$ =Standard MLE factor loading estimates, error variances, and the factor covariance matrix (Φ).

P =Penalty function

ρ =Regularization parameter

The MC+ penalty function was selected for use with FANC because it has been shown to provide somewhat sparser and more efficient estimates than either LASSO, or SCAD [18, 20-21]. The MC+ penalty function is defined as

$$MC+= \rho\left(|\theta| - \frac{\theta^2}{2\rho\gamma}\right) I(|\theta| < \rho\gamma) + \frac{\rho^2\gamma}{2} I(|\theta| \geq \rho\gamma) \quad (13)$$

Where

θ =Model parameters (e.g. factor loadings, covariances, etc.)

γ =Threshold value

An important aspect of using FANC is the selection of values for ρ and γ, which play crucial roles in yielding a sparse solution for the factor model. Based on the results of a small simulation study, [18] suggest the use of the Bayesian Information Criterion (BIC) for selecting these values. In other words, a range of γ and ρ values are assessed, and the combination that minimizes the BIC is used. This is the approach that was used in the current study.

Researchers [18] conducted a small simulation study in order to examine the performance of the FANC estimator. Factor loading estimation accuracy of the LASSO was compared with that of MC+ through the use of mean squared error (MSE) for factor loading estimates, and the proportion of cases where loadings were correctly set to be 0, or not (TPR). The researchers manipulated sample size (50, 100, 200), and underlying factor structure. Results of this simulation study demonstrated that MC+ yielded lower MSE and higher TPR than did the LASSO, leading [18] to suggest that the FANC algorithm should be examined further, and considered for use with high dimensional data.

Study goals

The primary goal of the current study was to compare the performance of GLASSO, POET, and FANC with one another in terms of factor model parameter estimation accuracy when *N* is small. In addition, the standard MLE approach was also included in order to serve as a baseline. Prior research has shown that using MLE, EFA models can be accurately fit with samples of less than 50, if the magnitude of the population factor loadings are large, the number of factors is small, and the number of indicators is large [5]. However, when these conditions are not met, the quality of the factor solution can be severely degraded. At the same time, there are several alternatives for estimating factor models when the sample is small, as described above. However, relatively little work has been done examining the performance of each of these, and no study has compared them all with one another, or with MLE. Thus, the purpose of this study is to extend the work in EFA with small samples by making these comparisons. GLASSO, POET, and FANC were selected for inclusion in this study because the research that has been done with them indicates that they have the potential to accurately recover factor models when *N* is small. However, as was mentioned previously, these earlier simulation studies tended to be small in scope, and thus may not generalize to a wide variety of situations. The current study was designed to fill this gap in the literature. Comparisons were made using both a small simulation study, and application of the methods to a high dimensional dataset. Both aspects of this work are described below.

However in many cases, claims originating in a particular year are often settled with a time delay of years or perhaps decades. Therefore, a method to estimate the expected liability is needed so that the insurer can calculate the profit of written policies, and allocate reserved assets to ensure liquidity. Since loss reserves generally represent by far the largest liability, and the greatest source of financial uncertainty in an insurance company, an appropriate valuation of insurance liabilities

including risk margin is one of the most important issues for a general insurer. Risk margin is the component of the value of claims liability that relates to the inherent uncertainty.

2. Experiment and Results

In order to compare the performance of the methods described above, a simulation study (1000 replications per combination of conditions) was conducted with all study conditions being completely crossed with one another. The observed indicator variables were generated from the N(0,1) distribution, as were the factors and the error terms. Factor loadings were set to specific values, which are described below. All data were generated, and analyses were conducted using the R software system, version 3.0.1 [22] . The factor model in (1) served as the data generating model, and a number of study conditions were manipulated, as described below. The outcome variables of interest were estimation bias, standard errors, and 95% coverage rates for factor loadings. Data were generated to represent simple structure, such that each indicator was only associated with a single factor in the multiple factor conditions.

Manipulated Simulation Conditions

Number of factors:

The number of factors underlying the observed indicators was set to be 1, 2, or 4. These values were selected so as to be reflective of common structures that are seen in practice, and are also values that have been used in prior research. For the 2 and 4 factor conditions, the interfactor correlation was 0.5.

Magnitude of primary factor loadings:

The loadings linking the observed indicators to the factors were 0.3, 0.6, or 0.9. The loadings were consistent across indicators within a given replication, so as to reflect weak (0.3), moderate (0.6), or strong factor structure (0.9).

Sample size:

The total sample size was set to be 10, 20, 30, 40, 50, 100, 200, 500, or 1000. These values were selected to represent extremely small samples (10, 20, 30) to very large samples (1000). In addition, prior research [5] investigated the performance of MLE with such small samples, but not with any alternatives. Thus, the current study serves as an extension of the earlier work by examining the performance of the alternatives to MLE described above, using very small samples.

Indicators per factor:

Several number of indicators were considered, including 3, 6, 9, and 12. Thus, the total number of indicators ranged between 3 (1 factor with 3 indicators) to 48 (4 factors with 12 indicators). These conditions were selected to represent a range of factor models from small to large.

Estimation Methods: MLE, POET, Soft POET, GLASSO, FANC

MLE with Promax rotation served as the baseline extraction method against which the others were compared. POET, Soft POET, and GLASSO were all first conducted in order to obtain a sparse

covariance matrix, to which MLE with Promax rotation was then applied. For POET and Soft POET, the number of principal components assumed to underlie the data equaled the number of factors that were assumed to be present. Simulations were also conducted in which only a single principal component was used, regardless of the number of factors that were believed present, and the results were virtually identical to those obtained when the number of components equaled the number of factors. Given this high degree of similarity, only results for the first case (number of components equaling number of factors) are reported below. FANC was conducted using an oblique factor rotation as well.

Simulation Study Results

In order to determine which of the manipulated factors in the simulation study influenced the estimation bias, analysis of variance (ANOVA) was used, with bias serving as the dependent variable, and the simulation study factors and their interactions as the independent terms in the model. The interaction of number of factors, sample size, number of factor indicators, and factor loading value was the highest order statistically significant term in the model ($F_{24,472} = 3.959, p < 0.001, \eta^2 = 0.168$). The factor loading estimate bias results for the different measures by the number of factors, sample size, number of observed indicator variables, and the value of the factor loadings appear in the panels of Figure 1. First, across all methods factor loading estimates were somewhat negatively biased. In addition, the results demonstrate that across all simulation conditions, MLE produced the most biased results, except when *N* was 200 or more, in which case it yielded the least biased estimates. Among the other approaches, FANC consistently had the least biased results of any of the methods studied here, again except when *N* was 200 or more. The other alternatives (POET, Soft POET, and GLASSO) also yielded less biased estimates than did MLE (except for *N* greater than or equal to 200), but had more biased estimates than did FANC, across conditions. Finally, factor loading estimation bias decreased for POET, Soft POET, and GLASSO concomitantly with increases in the population factor loading value. In other words, for these methods, the greater the population factor loading, the lower the estimation bias.

ANOVA was used to identify which of the manipulated study factors impacted the magnitude of the standard errors of the factor loadings. The interaction of sample size by number of indicators by method was the highest order effect in the model ($F_{8,472} = 17.993, p < 0.001, \eta^2 = 0.383$). All other main effects and interactions were either not statistically significant, or were subsumed in this interaction. Standard errors for all methods declined with larger sample sizes and more indicator variables. In addition, across conditions, the GLASSO estimator yielded the smallest standard error values, whereas Soft POET had the largest standard errors for 3 indicators, but comparable values for 6 or more indicators. Finally, the standard errors for the maximum likelihood approach were generally somewhat smaller than those of the POET and FANC methods, except for sample sizes of 10 or 20.

The coverage rates for the various methods were very close to the nominal 0.95 value for POET, Soft POET, FANC, and GLASSO across all study conditions. Indeed, these coverage rates ranged between 0.938 and 0.969. On the other hand, coverage rates for MLE were below the nominal level, except when *N* was 100 or greater, in which case it ranged between 0.940 and 0.958. However, for sample sizes less than 100, the coverage rate for MLE ranged between 0.712 and 0.885. Additionally, coverage for MLE was lower when more factors were present and when there were fewer indicators per factor.

EFA of Intelligence Test for Adults with Autism.

In order to demonstrate the utility of the approaches described above for estimating EFA models in the presence of small samples, they were used with a sample of 10 adults with Autism. The sample was collected from a center for Autism services at a university in the United States. The center provides employment and social/behavioral training for adults with Autism. This is a particularly difficult population for researchers to target, thereby leading to small research samples, such as the one considered here. Despite the small sample size, such research is particularly important, because young adults with Autism generally have very poor employment outcomes, with the majority never obtaining paid employment [23]. Recent research has demonstrated that cognitive functioning, as measured by intelligence tests, is crucial for such individuals to function successfully in society, though this issue requires further research in adults with Autism [24]. In order to better understand the latent structure of intelligence in such individuals, all 10 members of the sample were administered the Wechsler Adult Intelligence Scale, fourth edition (WAIS-IV) [25]. For the purposes of the current study, a total of 16 subscales were used, which appear in Table 1. The 10 subjects had a mean age of 24.3 years (standard deviation of 2.2), and had a low-average to average level of general intelligence (mean of 91.3, standard deviation of 6.6). All of the individuals in the sample were diagnosed with Autism.

The goal of this analysis was to determine whether the latent structure of the WAIS-IV for the 16 subscales that was previously confirmed empirically for the general population [26], and which is supported by theory [25], is also present for adults with Autism. This structure is represented in the first column of Table 1, in which the actual factor structure appears next to the subscale name. In order to address this issue, each of the methods for conducting EFA that were described above and compared in the simulation study were applied to the WAIS-IV subscale scores produced by the 10 adults with Autism. Factor extraction for the POET, Soft POET, and GLASSO approaches was carried out with MLE and Promax rotation on the covariance matrices produced by these methods. FANC was carried out using the original covariance matrix, with oblique factor rotation. MLE on the original covariance matrix was not possible because the number of indicators exceeded the sample size.

As an indicator to the number of factors to be retained, the eigenvalues produced by each method were examined, and found to reveal a very similar pattern. Given this similarity, only those results for POET are displayed in the scree plot in Figure 2. From these, it is clear that 4 factors capture the vast bulk of the variation in the observed indicators. Indeed, collectively the first 4 factors accounted for 98.8% of the variance in the indicator variables. This result matches with what theory and prior empirical work have shown with respect to the WAIS, namely that 4 factors should capture the latent structure of these subscales. As noted previously, the results for the other methods were very similar to those of POET.

The expected factor structure, based upon theory and prior empirical evidence, and the actual factor structures estimated by the various methods, when they were set to retrieve 4 factors, appear in Table 1. From these results, it is clear that each of the methods generally provided accurate recoveries of the expected factor structure. POET and Soft POET correctly grouped all but 2 of the 15 variables together, whereas GLASSO grouped all but 3 of the variables together in the expected fashion. The best performer was FANC, which accurately grouped all but one of the variables together in the appropriate theoretical construct. It is interesting to note that all of the methods incorrectly grouped the Matrix Reasoning subscale together with Digit Span, Arithmetic, and Letter-Number Sequencing. This confluence of results could indicate that this variable behaves differently for individuals with Autism than it does for the general population.

Given that it yielded the most accurate recovery of the latent structure of the WAIS-IV, the factor loadings estimated by FANC are examined, and appear in Table 2. From these results, it is clear that the loadings for the primary variables were typically more than twice as large as the cut-value used to determine to which factor each indicator belonged (0.3). In addition, the loadings for the non-primary

factors were all well below this cut-value, demonstrating the clear factor separation in the results. In short, with one exception FANC grouped the WAIS subscales in a manner consistent with theory, and did so in a clear fashion, with no cross loadings, or indicators that did not load on any factors. Taken together with the eigenvalue results presented above, it can be concluded that FANC, and to a slightly lesser degree POET, Soft POET, and GLASSO have accurately recovered the 4 factor structure underlying the 16 WAIS subscales examined here, using a sample consisting of only 10 adults with Autism.

3. Discussion

The primary goal of this study was to investigate the performance of several methods for estimating covariance matrices and use them to perform EFA in the presence of sparse data. Sparse data is a potential problem faced by researchers in a variety of disciplines, such as with small and difficult to sample populations, as was the case in this study. The results of the simulation study demonstrated that when *N* was less than 200, the greatest bias in factor loading estimates was exhibited by MLE, while FANC had the lowest estimation bias of the methods studied here. In addition, POET, Soft POET, and GLASSO all yielded estimates that had only slightly more bias than did FANC. In addition, neither the number of factors, nor the number of indicators per factor were found to be related to factor loading estimation bias of any of the alternative methods, though they were related to bias in MLE. Specifically, the more factors that were present the greater the bias in MLE estimates, whereas the more indicators per factor, the less such bias was present. Sample size was also related to the level of bias in MLE factor loading estimates, such that when *N* was 200 or greater, the level of bias was 0.005 or less, and was lower than that of the alternatives. In contrast, sample size had a very modest impact on the loading bias of the alternatives. Finally, the magnitude of the population factor loading value was not related to the level of bias of either FANC or MLE, but was associated with estimation bias for POET, Soft POET, and GLASSO, whereby larger values were associated with less bias with these methods. This latter result was likely associated with the fact that the regularization methods are designed to identify and constrain to 0 relatively low covariances among the variables. In turn, low covariances are associated with smaller factor loadings. Therefore, when the factor loadings are relatively small, this is indicative of small covariances among the indicators, which in turn appears to make it somewhat more difficult for the regularization procedures to correctly identify which covariances to constrain to be 0. However, even with this difficulty, POET, Soft POET, and GLASSO all produced less biased estimates than did MLE.

In addition to the simulation study, results of the EFA on the WAIS further demonstrated the superiority of the alternative methods studied here when compared to the MLE, when the sample size is small. Indeed, although MLE could not estimate the EFA model for the 10 adults with Autism using the 16 WAIS subscales, FANC, POET, Soft POET, and GLASSO all yielded estimated models that were very close in latent structure to what was anticipated, based on prior research and theory. Taken together, the results of this study have demonstrated that it is possible to obtain accurate EFA results even for very small sample *N* and high dimensional data situations. The accuracy of FANC in particular would appear to hold even in situations where the population factor loadings are fairly weak, and the data are quite high dimensional. Thus, we would conclude by recommending that researchers faced with a high dimensional data problem consider using FANC as their method for estimating EFA models. It appears to yield low parameter estimation bias, and to accurately recover the latent variable structure across a wide variety of high dimensional data conditions. One final issue to address is that, where possible, researchers using regularization methods should carefully employ several criteria in the selection of the regularization parameter values. Each of the methods discussed above require the

setting of a tuning parameter in order to establish the penalty that will be used. Typically these approaches utilize information indices such as the BIC, or resampling methods such as the jackknife. Ideally, researchers making use of these regularization methods should refer to multiple of these criteria when possible, including BIC, AIC, and CAIC, as well as mean square error of prediction, for example.

Conflict of Interest:

None declared.

References

[1] Thomposon, B. (2004). Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications. Washington, D.C.: American Psychological Association.

[2] Green, S.B., Redell, N., Thompson, M.S., & Levy, R. (in press). Accuracy of Revised and Traditional Parallel Analyses for Assessing Dimensionality with Binary Data. Educational and Psychological Measurement.

[3] Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best Practices in Exploratory Factor Analysis. In J. W. Osborne (Ed.), Best Practices in Quantitative Methods (pp. 205-213). Thousand Oaks, CA: Sage Publishing.

[4] MacCallum, R.C., Widaman, K.F., Preacher, K.J., & Hong, S. (2001). Sample Size in Factor Analysis: The Role of Model Error. Multivariate Behavioral Resarch, 36(4), 611-637.

[5] de Winter, J.C.F., Dodou, D., & Wieringa, P.A. (2009). Exploratory Factor Analysis with Small Sample Sizes. Multivariate Behavioral Research, 44, 147-181.

[6] Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research & Evaluation, 10, 1–9.

[7] Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. Behavior Genetics, 32, 153–161.

[8] Fan, J., Liao, Y., Mincheva, M. (2013). Large Covariance Estimation by Thresholding Principal Components. Journal of the Royal Statistical Society, Series B, 75(4), 603-680.

[9] Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. Journal of Econometrics, 147, 186–197.

[10] Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society, Ser. B, 58, 267–288.

[11] Yuan, M. & Lin, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model. Biometrika, 94, 19-35.

[12] Banerjee, O., Ghaoui, L.E., & d'Aspremont, A. (2008). Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. Journal of Machine Learning Research, 9 485-516.

[13] Dahl, J., Roychowdhury, V. & Vandenberghe, L.. (2008). Covariance selection for non-chordal graphs via chordal embedding. Optimization Methods and Software, 23(4), 501-520.

[14] Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the lasso. Annals of statistics, 34:1436–1462.

[15] Friedman, J., Hastie, T., Tibshirani, R. (2007). Sparse Inverse Covariance Estimation with the Lasso. Biostatistics, http://www-stat.stanford.edu/~tibs/ftp/graph.pdf.

[16] Sachs, K., Perez, O., Pe'er, D., Lauffenbuerger, D.A., & Nolan, G.P. (2005). Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science, 308, 523-529.

[17] Antoniadis, A. & Fan, J. (2001). Regularization of Wavelet Approximations. Journal of the American Statistical Association, 96, 939-967.

[18] Hirose, K. & Yamamoto, M. (2015). Sparse Estimation via Nonconcave Penalized Likelihood in Factor Analysis Model. Statistical Computing, 25, 863-875.

[19] Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics, 38, 894–942.

[20] Zhao, P. & Yu, B. (2007). On model selection consistency of lasso. Journal of Machine Learning Research, 7(2), 2541.

[21] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101, 1418–1429.

[22] R Core Development Team. (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[23] Roux, A.M., Shattuck, P.T., & Cooper, B.P., (2013) Postsecondary employment experiences among young adults with an autism spectrum disorder. Journal of the American Academy of Child and Adolescent Psychiatry, 52(9), 931–939.

[24] Corbett, B.A., Schupp, C.W., Levine, S., & Mendoza, S..(2009). Comparing cortisol, stress and sensory sensitivity in children with autism. Autism Research. 2:32–39.

[25] Wechsler D. (2009). Wechsler Memory Scale–Fourth Edition. Pearson; San Antonio, TX.

[26] Holdnack, J.A., Zhou, X., Larrabee, G.J., Millis, S.R., & Salthouse, T.A. (2011). Confirmatory Factor Analysis of the WAIS-IV/WMS-IV. Assessment, 18(2), 178-91.

Received March 15, 2013; accepted November 10, 2013.

Table 1: Factor Structure for the WAIS in the Population, and as Recovered by EFA Models

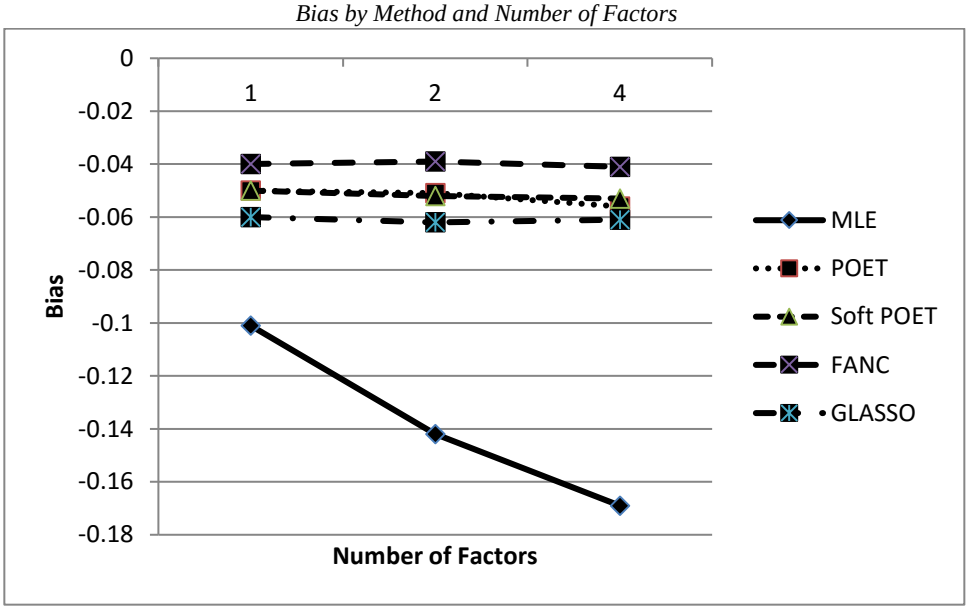
WAIS subscale	Actual	POET	SoftPOET	GLASS0	FANC
Similarities	1	3	3	1	4
Vocabulary	1	3	3	1	4
Information	1	3	3	1	4
Comprehension	1	3	3	4	4
Block Design	2	2	2	3	2
Matrix Reasoning	2	1	1	2	1
Visual Puzzles	2	2	2	3	2
Figure Weights	2	2	2	3	2
Picture Completion	2	2	2	3	2
Digit Span	3	1	1	2	1
Arithmetic	3	1	1	2	1
Letter-Number Sequence	3	1	1	2	1
Symbol Search	4	4	4	4	2
Coding	4	4	4	4	2
Cancellation	4	2	2	3	2

Table 2: FANC Factor Loading Estimates for the 4 Factor Solution

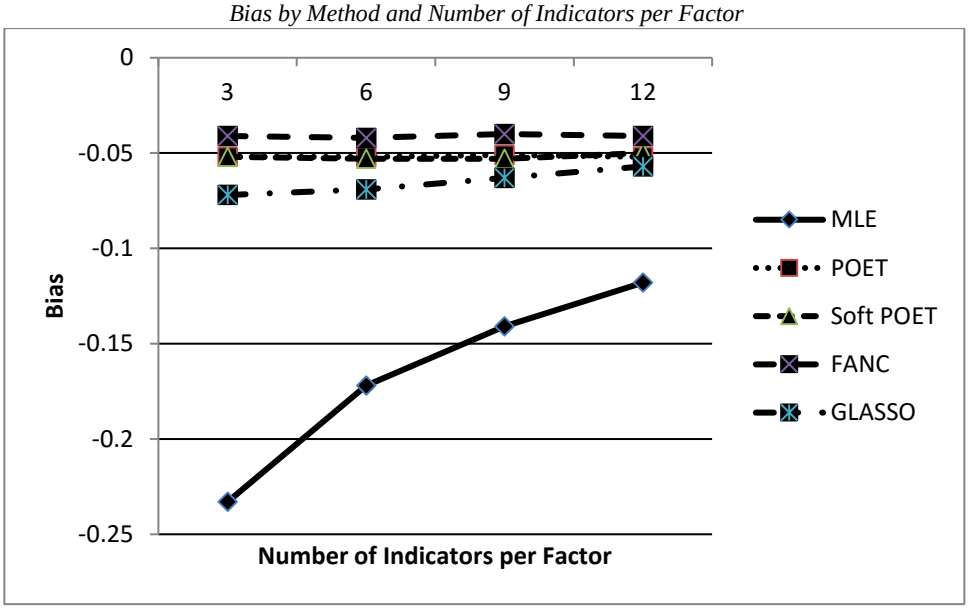
WAIS subscale	Factor 1	Factor 2	Factor 3	Factor 4
Similarities	.16	.02	.21	.80
Vocabulary	.20	-.22	-.02	.57
Information	-.06	-.03	-.20	.84
Comprehension	.03	-.01	.11	.76
Block Design	.02	.65	-.10	-.20
Matrix Reasoning	.91	.01	.19	.01
Visual Puzzles	0	.88	.05	0
Figure Weights	.04	.77	-.05	0
Picture Completion	.23	.81	-.06	.03
Digit Span	.75	-.22	-.19	0
Arithmetic	.72	0	-.20	0
Letter-Number Sequence	.66	.10	-.14	.09
Symbol Search	.14	-.84	.26	-.28
Coding	-.09	-.86	.06	0
Cancellation	.04	-.81	.08	-.21

Figure 1: Factor Loading Bias by Number of Factors, Number of Indicators per Factor, Sample Size, and Population Factor Loading Value

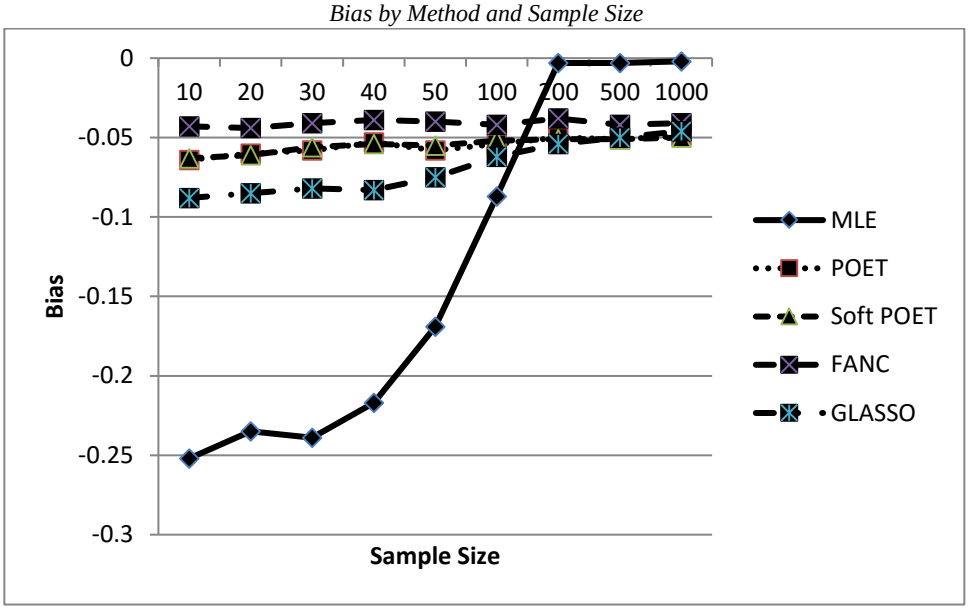
Bias by Method and Number of Factors



Bias by Method and Number of Indicators per Factor



Bias by Method and Sample Size



Bias by Method and Population Factor Loading Value

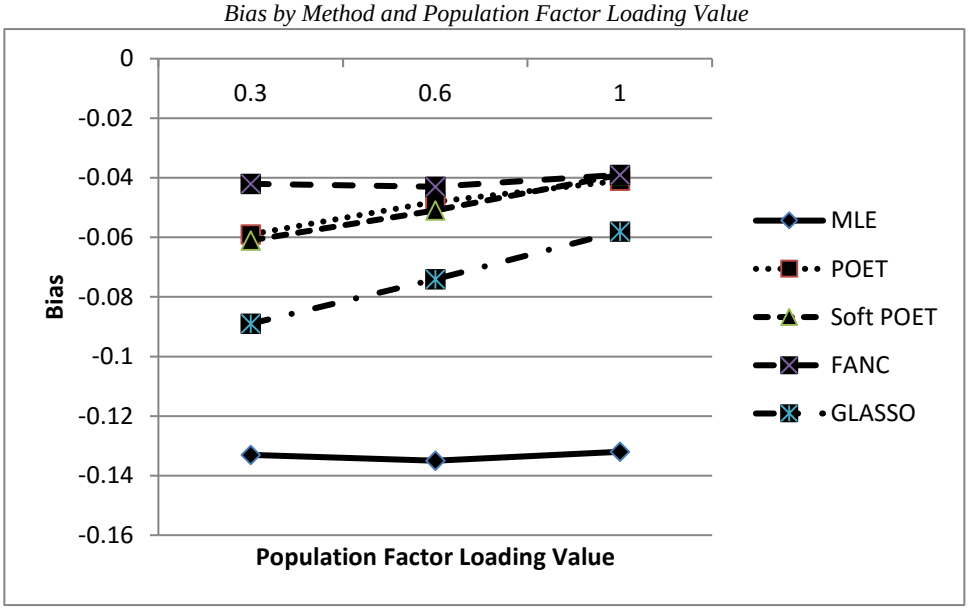


Figure 2: Standard Error Estimates by Method of Estimation, Number of Indicators, and Sample Size

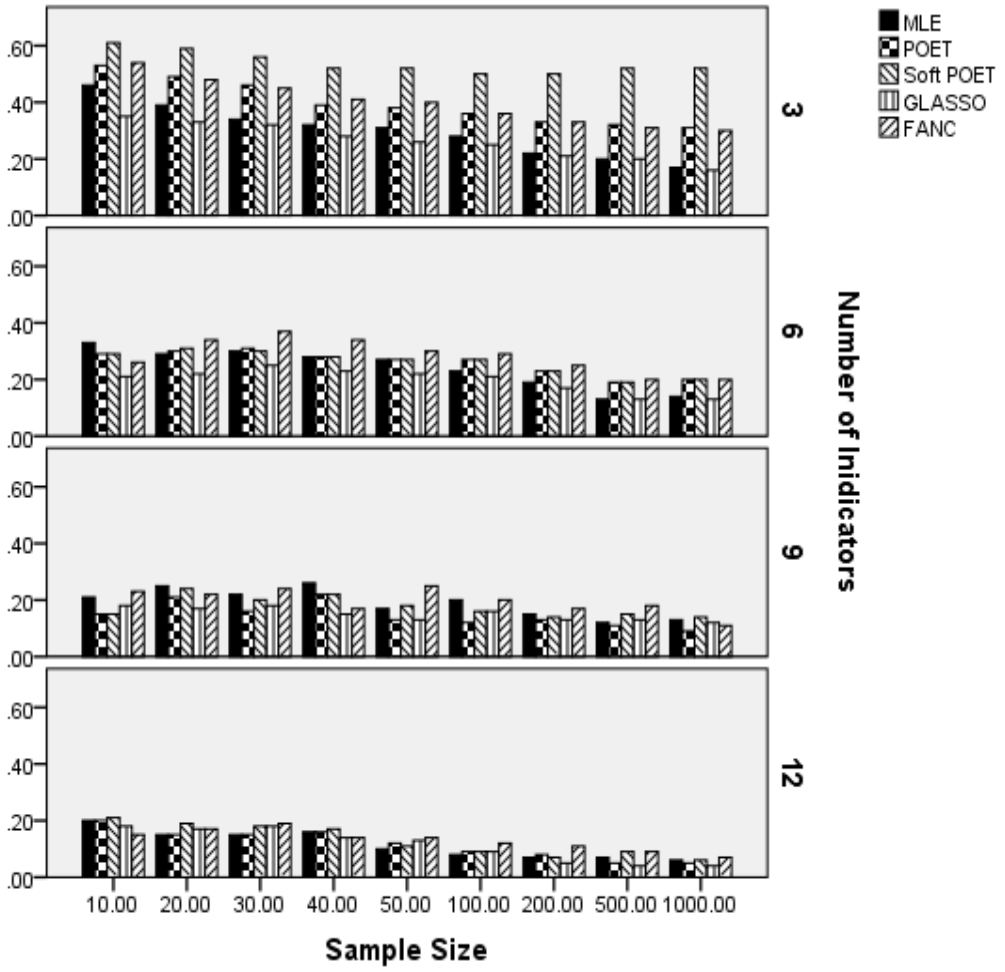
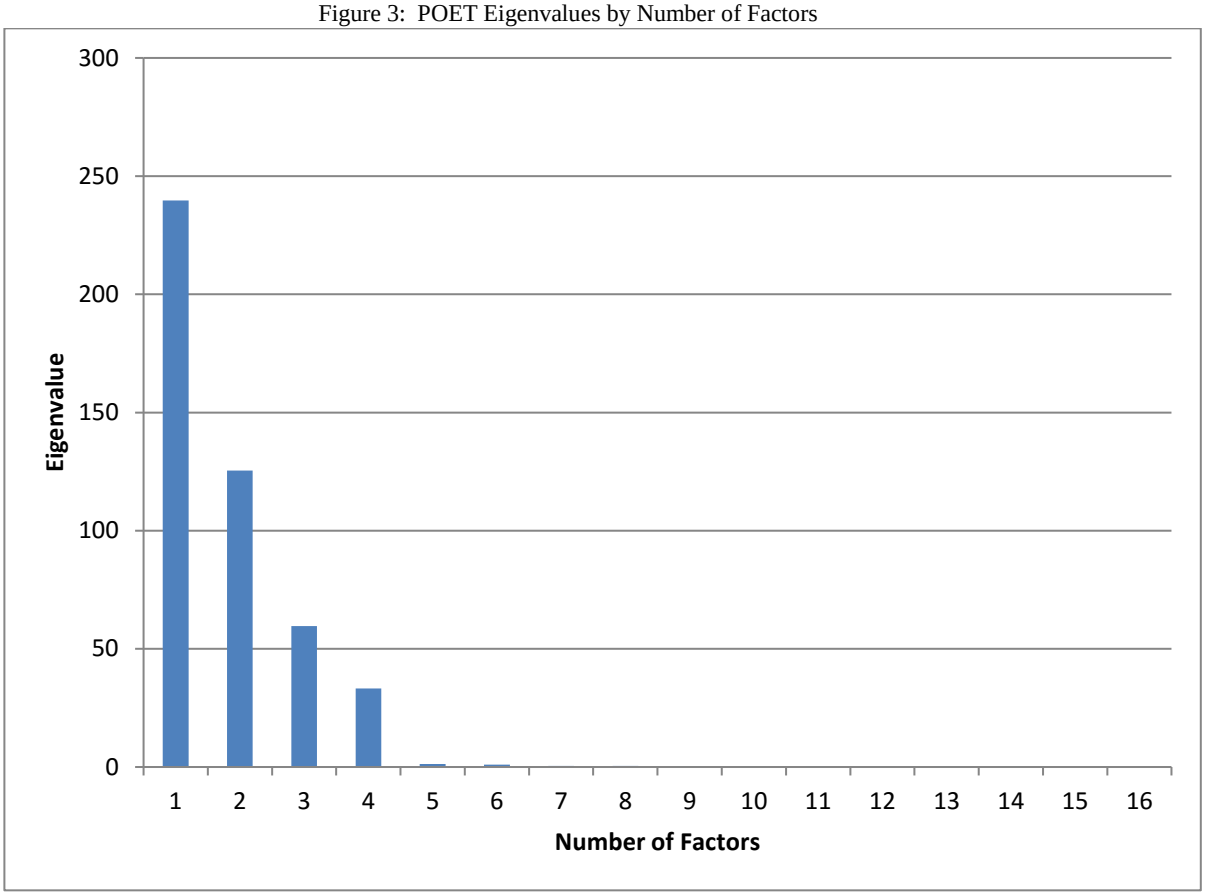


Figure 3: POET Eigenvalues by Number of Factors



W. Holmes Finch
Department of Educational Psychology
Ball State University
Muncie, IN, USA

Maria Herandez E. Finch
Department of Educational Psychology
Ball State University
Muncie, IN, USA