Ann Inst Stat Math (2017) 69:571–597
DOI 10.1007/s10463-016-0552-2

Improving the convergence rate and speed of Fisher-scoring algorithm: ridge and anti-ridge methods in structural equation modeling

Ke-Hai Yuan[1] · Peter M. Bentler[2]

Received: 12 May 2015 / Revised: 2 November 2015 / Published online: 23 February 2016

Abstract In structural equation modeling (SEM), parameter estimates are typically computed by the Fisher-scoring algorithm, which often has difficulty in obtaining converged solutions. Even for simulated data with a correctly specified model, non-converged replications have been repeatedly reported in the literature. In particular, in Monte Carlo studies it has been found that larger factor loadings or smaller error variances in a confirmatory factor model correspond to a higher rate of convergence. However, studies of a ridge method in SEM indicate that adding a diagonal matrix to the sample covariance matrix also increases the rate of convergence for the Fisher-scoring algorithm. This article addresses these two seemingly contradictory phenomena. Using statistical and numerical analyses, the article clarifies why both approaches increase the rate of convergence in SEM. Monte Carlo results confirm the analytical results. Recommendations are provided on how to increase both the speed and rate of convergence in parameter estimation.

Keywords Fisher-scoring algorithm · Coefficient of variation · Condition number · Speed of convergence

The research is supported by the National Science Foundation under Grant No. SES-1461355.

✉ Ke-Hai Yuan
kyuan@nd.edu

[1] Department of Psychology, University of Notre Dame, Notre Dame, IN 46556, USA

[2] Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095, USA

1 Introduction

Obtaining a set of parameter estimates for a theoretically-plausible model is the first step in any statistical analysis. For structural equation modeling (SEM), however, parameter estimates have to be computed iteratively, and there is a good chance that a researcher is unable to obtain a set of converged solutions in practice, especially when the sample size is not large enough. Various factors can contribute to non-convergence of an iterative algorithm, including bad model or bad data. But only after a set of solutions is obtained can we possibly distinguish different causes. In particular, non-convergences not just occur to poorly formulated models, as they have been repeatedly reported for correctly specified models with simulated data in Monte Carlo studies (Bentler and Yuan 1999; Hu et al. 1992; Jackson 2001) and in bootstrap replications (Ichikawa and Konishi 1995; Yuan and Hayashi 2003). Obtaining converged solutions is equally important to Monte Carlo studies although replications are essentially cost free. This is because non-converged replications cannot be regarded as identically distributed as the converged ones (e.g., Yuan and Hayashi 2003). Similar to missing data analysis that ignores missing not at random mechanism, when the percentage of non-converged replications is substantial, results obtained based on just the converged replications may not correctly reflect the properties of the methodology being studied. The main purpose of this article is to examine factors affecting the convergence properties of the Fisher-scoring (FS) algorithm, which is used in most SEM packages. Strategies for achieving higher convergence rates in both simulation and real data analysis are explored as well.

In conducting Monte Carlo studies with confirmatory factor models using LISREL (Jöreskog and Sörbom 1981), Anderson and Gerbing (1984) observed that non-converged replications occurred more often with small factor loadings or indicators with low reliability. Boomsma (1985) also observed that in simulation studies the convergence rate of LISREL increased with greater measurement reliabilities. Jackson (2001) noted a similar phenomenon in simulation studies using SAS Calis (SAS Institute 1996). However, recent studies of a ridge method in SEM indicate that adding a diagonal matrix to the sample covariance matrix $\mathbf{S}$ increases the rate of convergence (Yuan and Chan 2008). Because manipulation for larger factor loadings is in the opposite direction of the ridge method, we will call it an anti-ridge method. Thus, the findings in the literature are seemingly in conflict. By examining factors affecting the convergence properties of the FS algorithm, we will clarify why these two seemingly contradictory methods both lead to higher convergence rate. Our analysis of the ridge and anti-ridge methods also applies to other algorithms for minimizing the normal-distribution-based maximum likelihood (NML) discrepancy function and similarly is relevant to other discrepancy functions.

The convergence properties of the FS algorithm are affected by many factors. One of them is multicolinearity among the observed variables. If the sample or model implied covariance matrix is close to being singular, then the FS algorithm may have difficulty to reach a set of converged solutions. The ridge method developed in Yuan and Chan (2008) aims to address the problem of near singular covariance matrices. Instead of fitting $\mathbf{S}$ by a structural model $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, the ridge method fits $\mathbf{S}_a = \mathbf{S} + a\mathbf{I}$ by $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ through minimizing the NML-based discrepancy function, where $a > 0$ is a constant that may depend on the sample size N and the number of variables p but not the observed data. Let $\hat{\boldsymbol{\theta}}_a$ be the resulting estimates from fitting $\mathbf{S}_a$. Then the final parameter estimates are obtained by deducting a from the elements of $\hat{\boldsymbol{\theta}}_a$ that correspond to the variances of errors.

In the literature of numerical analysis, the ratio of the largest over the smallest eigenvalues of a matrix is called the condition number of the matrix (Golub and Van Loan 1983). A large condition number not only causes computations involving the inverse of the matrix to be less accurate, it may also cause the algorithm to fail to converge if the computation is iterative. The condition number of $\mathbf{S}_a = \mathbf{S} + a\mathbf{I}$ is always smaller than that of $\mathbf{S}$. Yuan and Chan (2008) briefly discussed how the ridge method improves the convergence properties of the FS algorithm. In this article, we further study the changes in condition numbers as well as the convergence rate and speed between modeling $\mathbf{S}$ and $\mathbf{S}_a$ using numerical examples and Monte Carlo simulation.

The population or sample covariance matrix corresponding to a SEM model might be close to singular if certain error variances are tiny. Thus, we would expect the convergence rate of the FS algorithm to decrease with the increase of factor loadings, opposite to what Anderson and Gerbing (1984) and Boomsma (1985) have found. However, as we shall show, with the increase of factor loadings, the relative sampling errors in the sample covariances become smaller. Our analysis and results further show that smaller sampling errors in $\mathbf{S}$ will positively affect the convergence rate as well as the speed of convergence of the FS algorithm. Smaller sampling errors also tend to improve the condition of the sample covariance matrix $\mathbf{S}$. Because the condition number is a rather complicated function of the elements of $\mathbf{S}$, we will use numerical examples and Monte Carlo simulation to evaluate the change of condition numbers following the anti-ridge method.

In addition, we will discuss how to use the findings in practice when FS fails to obtain a set of converged solutions. As we shall see, the ridge method can be applied to all the models where error variances are subject to estimation. In contrast, the anti-ridge method is mostly usable a priori in Monte Carlo studies where the factor loadings are subject to manipulation, and can also be applied to special models in post-hoc analysis when the space of the common factors are known or when alternative items are available.

We will not study the properties of parameter estimates or test statistics for overall model evaluation with the ridge method. These have been studied in Yuan and Chan (2008). In particular, Kamada (2011) and Kamada and Kano (2012) found that the ridge method can yield parameter estimates that are substantially more accurate than MLEs at smaller N, even when the population is normally distributed. Since the applicability of the anti-ridge method is limited, we will not study properties of parameter estimates following the anti-ridge method.

In Sect. 2 of the article, we review the formulation of the FS algorithm and examine the factors that affect its convergence properties. In Sect. 3, we obtain formulas that show how the relative errors in $\mathbf{S}$ are affected by population factor loadings and error/unique variances. In Sect. 4, using examples, we numerically illustrate how converged solutions are obtained with ridge and/or anti-ridge methods. Monte Carlo results on the effectiveness of ridge and anti-ridge methods are presented in Sect. 5. Recommendation and discussions regarding the applications of ridge and anti-ridge methods are given in the concluding section.

2 Fisher-scoring algorithm

In this section, we will first present a formulation of the FS algorithm in SEM. Factors that affect the speed of convergence of FS as well as whether there exists a vector of parameters that satisfies a given convergence criterion are then examined.

Let $\mathbf{S} = (s_{jk})$ be a sample covariance matrix of size N from a p-variate population. We are interested in modeling the covariance matrix $\mathbf{\Sigma} = E(\mathbf{S})$ by a structural model $\mathbf{\Sigma}(\boldsymbol{\theta})$ using NML, which defines parameter estimate $\hat{\boldsymbol{\theta}}$ by minimizing

$$F_{ML}(\mathbf{S}, \mathbf{\Sigma}(\boldsymbol{\theta})) = \mathrm{tr}[\mathbf{S}\mathbf{\Sigma}^{-1}(\boldsymbol{\theta})] - \log|\mathbf{S}\mathbf{\Sigma}^{-1}(\boldsymbol{\theta})| - p. \tag{1}$$

Let vec$(\mathbf{S})$ be the vector of stacking the columns of $\mathbf{S}$, and $\mathbf{s} = \mathrm{vech}(\mathbf{S})$ be the vector of stacking the lower-triangular part of $\mathbf{S}$. Then, with $p^* = p(p+1)/2$, there exists a $p^2 \times p^*$ matrix $\mathbf{D}_p$ such that $\mathbf{D}_p\mathrm{vech}(\mathbf{S}) = \mathrm{vec}(\mathbf{S})$, and $\mathbf{D}_p$ is called a duplication matrix (see e.g., Schott 2005, p. 313). Further let $\boldsymbol{\sigma}(\boldsymbol{\theta}) = \mathrm{vech}[\mathbf{\Sigma}(\boldsymbol{\theta})]$,

$$\dot{\boldsymbol{\sigma}}(\boldsymbol{\theta}) = \frac{\partial\boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}, \quad \text{and} \quad \mathbf{W}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{D}_p'\left[\mathbf{\Sigma}^{-1}(\boldsymbol{\theta}) \otimes \mathbf{\Sigma}^{-1}(\boldsymbol{\theta})\right]\mathbf{D}_p.$$

With initial value $\boldsymbol{\theta}^{(0)}$, the FS algorithm for computing $\hat{\boldsymbol{\theta}}$ is given by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \Delta\boldsymbol{\theta}^{(t)}, \tag{2}$$

where

$$\Delta\boldsymbol{\theta}^{(t)} = [\mathbf{H}(\boldsymbol{\theta}^{(t)})]^{-1}\dot{\boldsymbol{\sigma}}'(\boldsymbol{\theta}^{(t)})\mathbf{W}(\boldsymbol{\theta}^{(t)})[\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}^{(t)})] \tag{3}$$

with $\mathbf{H}(\boldsymbol{\theta}) = \dot{\boldsymbol{\sigma}}'(\boldsymbol{\theta})\mathbf{W}(\boldsymbol{\theta})\dot{\boldsymbol{\sigma}}(\boldsymbol{\theta})$ being the information matrix. The NML estimate $\hat{\boldsymbol{\theta}}$ is obtained when the algorithm converges, which is typically defined as the absolute values of all the elements of $\Delta\boldsymbol{\theta}^{(t)}$ being smaller than a given number. A variant of (2) is

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \alpha\Delta\boldsymbol{\theta}^{(t)}, \tag{4}$$

where the scalar α is to control the size of the step so that

$$F_{ML}(\mathbf{S}, \mathbf{\Sigma}(\boldsymbol{\theta}^{(t+1)})) < F_{ML}(\mathbf{S}, \mathbf{\Sigma}(\boldsymbol{\theta}^{(t)})).$$

The value of α can be chosen using step halving (.50, .25, ...) or a line search method (e.g., chapter 2 of Everitt 1987; chapter 3 of Nocedal and Wright 1999). We will call (4) the FS algorithm with step-size adjustment.

Since all the elements of the $\Delta\boldsymbol{\theta}^{(t)}$ in (3) must be small enough for FS to converge, we further examine its two major components: the inverse of the information matrix $\mathbf{H}^{(t)} = \mathbf{H}(\boldsymbol{\theta}^{(t)})$ and the score vector $\boldsymbol{\nu}^{(t)} = \dot{\boldsymbol{\sigma}}'(\boldsymbol{\theta}^{(t)})\mathbf{W}(\boldsymbol{\theta}^{(t)})[\mathbf{s} - \boldsymbol{\sigma}(\boldsymbol{\theta}^{(t)})]$. Denote $\boldsymbol{\Sigma}^{(t)} = \boldsymbol{\Sigma}(\boldsymbol{\theta}^{(t)})$ and $\dot{\boldsymbol{\Sigma}}_j^{(t)} = \partial\boldsymbol{\Sigma}(\boldsymbol{\theta}^{(t)})/\partial\theta_j^{(t)}$, then the jth element of $\boldsymbol{\nu}^{(t)}$ can be further written as

$$\nu_j^{(t)} = \mathrm{tr}\left[(\boldsymbol{\Sigma}^{(t)})^{-1}\dot{\boldsymbol{\Sigma}}_j^{(t)}(\boldsymbol{\Sigma}^{(t)})^{-1}(\mathbf{S} - \boldsymbol{\Sigma}^{(t)})\right].$$

Clearly, causes for FS to fail to converge must be through $\boldsymbol{\nu}^{(t)}$ and $\mathbf{H}^{(t)}$, and they might be classified into four categories: (C1) The first cause is when $\boldsymbol{\Sigma}^{(t)}$ is near singular. Then $\boldsymbol{\Sigma}^{(t)}$ may not be invertible, which will cause problems to the computation of both $\mathbf{H}^{(t)}$ and $\boldsymbol{\nu}^{(t)}$. (C2) The second is when $\dot{\boldsymbol{\sigma}}(\boldsymbol{\theta}^{(t)})$ is close to rank deficient. Rank deficient $\dot{\boldsymbol{\sigma}}(\boldsymbol{\theta}^{(t)})$ does not cause operational problems to the calculation of $\nu_j^{(t)}$ because matrix multiplication is very robust to the values of the elements of the matrices, but it may cause $\mathbf{H}^{(t)}$ to be close to singular. Then the calculation of $\Delta\boldsymbol{\theta}^{(t)}$ cannot proceed. (C3) The third is when there exist large sampling and/or systematic differences between $\mathbf{S}$ and $\boldsymbol{\Sigma}^{(t)}$ relative to the size of the elements of $\boldsymbol{\Sigma}^{(t)}$. Notice that the matrix $(\boldsymbol{\Sigma}^{(t)})^{-1}(\mathbf{S} - \boldsymbol{\Sigma}^{(t)})$ in the expression of $\nu_j^{(t)}$ essentially represents the relative errors in $\mathbf{S}$. When the relative errors are large enough, certain elements of $\Delta\boldsymbol{\theta}^{(t)}$ may never satisfy a given convergence criterion. (C4) The fourth is the effect of interactions between the relative errors in $\mathbf{S}$ and the conditions of $\boldsymbol{\Sigma}^{(t)}$ and/or $\mathbf{H}^{(t)}$. A matrix $\mathbf{A}$ is said to be ill-conditioned if its condition number $\kappa(\mathbf{A})$ is huge. An ill-conditioned matrix may not be near singular if its smallest eigenvalue is not close to zero. However, according to Golub and Van Loan (1983, Sect. 2.5), the relative errors in $\mathbf{x}$ resulting from $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ can be $\kappa(\mathbf{A})$ times those in $\mathbf{A}$ and $\mathbf{b}$. Moderate errors in $\mathbf{S}$ together with large condition numbers of $\boldsymbol{\Sigma}^{(t)}$ and/or $\mathbf{H}^{(t)}$ can result in substantial fluctuations in $\Delta\boldsymbol{\theta}^{(t)}$ from iteration to iteration, which will not satisfy a given convergence criterion.

In addition to the four noted causes, the convergence properties of the FS algorithm are also affected by the initial value $\boldsymbol{\theta}^{(0)}$. In the following section, we will examine how ridge and anti-ridge methods change the formulation of $\Delta\boldsymbol{\theta}^{(t)}$ so that the convergence properties of FS improve. We will not discuss initial values because they are not unique to either the ridge or the anti-ridge method.

3 Relative errors in sample covariances, ridge and anti-ridge methods

In this section, we will first quantify the relative errors in $\mathbf{S}$ using the coefficient of variation (CV). Then we examine how the relative errors change in the ridge and anti-ridge methods. Condition numbers of covariance and information matrices following ridge and anti-ridge methods will also be discussed. Since our interest is in the effect of the size of factor loadings versus that of the size of error variances, we will mainly consider factor models. Another reason for us to consider factor models is that an SEM model can be equivalently expressed as a factor model with structured factor variances-covariances. Notice that, in SEM or factor analysis, the size of factor loadings and that of factor variances-covariances cannot be distinguished before fixing the scales of latent variables. Unless stated otherwise, we assume that the variance of each factor is fixed at 1.0 from now on.

3.1 Relative errors in s_{jk}

Let $\mathbf{y} = (y_1, y_2, \ldots, y_p)'$ represent a population with p random variables. Suppose $\mathbf{y}$ follows a confirmatory factor model

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{\mu} = E(\mathbf{y})$; $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings; $\boldsymbol{\xi}$ is a vector of q latent factors with $E(\boldsymbol{\xi}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\xi}) = \boldsymbol{\Phi} = (\phi_{lm})$ being a correlation matrix; and $\boldsymbol{\varepsilon}$ is a vector of p errors or uniquenesses with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi} = \mathrm{diag}(\psi_{11}, \psi_{22}, \ldots, \psi_{pp})$. When $\boldsymbol{\xi}$ and $\boldsymbol{\varepsilon}$ are uncorrelated, the covariance matrix of $\mathbf{y}$ is given by

$$\boldsymbol{\Sigma} = (\sigma_{jk}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{6}$$

In this section, we will further assume that $\boldsymbol{\xi}$ and $\boldsymbol{\varepsilon}$ are independent to avoid overly complicated analytical results. In Monte Carlo studies in Sect. 5, we will further evaluate relative errors in $\mathbf{S}$ when $\boldsymbol{\xi}$ and $\boldsymbol{\varepsilon}$ are uncorrelated but dependent.

Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})'$, $i = 1, 2, \ldots, N$, be a random sample of the $\mathbf{y}$ in (5), then the sample covariance matrix is given by $\mathbf{S} = (s_{jk})$ with

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{N}(y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k),$$

where $n = N - 1$. Notice that each s_{jk} is a 2nd-order sample moment. Standard large sample theory shows that the asymptotic variance of $\sqrt{n}s_{jk}$ is given by $\gamma_{jk} = \mathrm{Var}(y_{j0}y_{k0})$ (e.g., Ferguson 1996), where $y_{j0} = y_j - \mu_j$ and $y_{k0} = y_k - \mu_k$. Because the exact expression for $\mathrm{Var}(\sqrt{n}s_{jk})$ is rather complicated and its difference from γ_{jk} is in the order of $1/N$, we treat γ_{jk} as the variance of $\sqrt{n}s_{jk}$ for simplicity. Let CV_{jk} denote the coefficient of variation of s_{jk}. Then $\mathrm{CV}_{jk} = \gamma_{jk}^{1/2}/(\sqrt{n}\sigma_{jk})$. We next quantify CV_{jk} with respect to the population values of the parameters in (6). For the obtained formulas to have relatively simple forms, we only consider unidimensional measurement where each variable loads on only one factor.

With q factors, suppose $y_j = \lambda_j\xi_{j^*} + \varepsilon_j$ and $y_k = \lambda_k\xi_{k^*} + \varepsilon_k$, where $1 \leq j^* \leq k^* \leq q$. It follows from the results in Sect. 7 (Appendix) that, when $j^* \neq k^*$ $(\sigma_{jk} = \lambda_j\lambda_k\phi_{j^*k^*})$,

$$n\mathrm{CV}_{jk}^2 = \frac{1}{\phi_{j^*k^*}^2}\left\{\left[E\left(\xi_{j^*}^2\xi_{k^*}^2\right) - \phi_{j^*k^*}^2\right] + \frac{\psi_{kk}}{\lambda_k^2} + \frac{\psi_{jj}}{\lambda_j^2} + \frac{\psi_{jj}}{\lambda_j^2}\frac{\psi_{kk}}{\lambda_k^2}\right\}; \tag{7}$$

when $j^* = k^*$ but $j \neq k$ $(\sigma_{jk} = \lambda_j\lambda_k)$,

$$n\mathrm{CV}_{jk}^2 = \left[E\left(\xi_{j^*}^4\right) - 1\right] + \frac{\psi_{kk}}{\lambda_k^2} + \frac{\psi_{jj}}{\lambda_j^2} + \frac{\psi_{jj}}{\lambda_j^2}\frac{\psi_{kk}}{\lambda_k^2}; \tag{8}$$

and when $j = k$ ($\sigma_{jj} = \lambda_j^2 + \psi_{jj}$),

$$n\mathrm{CV}_{jj}^2 = \frac{\left[E\left(\xi_{j*}^4\right) - 1\right] + 4\psi_{jj}/\lambda_j^2 + \left[E\left(\varepsilon_j^4\right) - \psi_{jj}^2\right]/\lambda_j^4}{\left(1 + \psi_{jj}/\lambda_j^2\right)^2}. \tag{9}$$

Let $\varepsilon_j = \psi_j \varepsilon_{j0}$ with $\psi_j = \psi_{jj}^{1/2}$, then we can further write (9) as

$$n\mathrm{CV}_{jj}^2 = \frac{\left[E\left(\xi_{j*}^4\right) - 3\right]}{\left(1 + \psi_{jj}/\lambda_j^2\right)^2} + \frac{\left[E\left(\varepsilon_{j0}^4\right) - 3\right]\left(\psi_{jj}/\lambda_j^2\right)^2}{\left(1 + \psi_{jj}/\lambda_j^2\right)^2} + 2. \tag{10}$$

It is clear from (7) and (8) that, when $j \neq k$, the relative error in s_{jk} is an increasing function of ψ_{jj} and ψ_{kk}, and a decreasing function of $|\lambda_j|$ and $|\lambda_k|$. The relationship of CV_{jj}^2 with λ_j and ψ_{jj} in (9) or (10) depends on the kurtoses of ξ_{j^*} and ε_j. When both ξ_{j^*} and ε_j are normally distributed, $E(\xi_{j*}^4) = E(\varepsilon_{j0}^4) = 3$, then $n\mathrm{CV}_{jj}^2$ is unrelated to λ_j or ψ_{jj}. Otherwise, $n\mathrm{CV}_{jj}^2$ will depend on the values of factor loadings and error variances. Suppose $E(\varepsilon_{j0}^4) > 3$ and $E(\xi_{j*}^4) > 3$, then the first term on the ride side of (10) increases with $|\lambda_j|$ and decreases with ψ_{jj}, whereas the second term on the right side of (10) changes in the opposite direction.

For normally distributed populations considered in Anderson and Gerbing (1984) and Boomsma (1985), relative errors in s_{jj} are not affected by the values of factor loadings or error variances, but relative errors in s_{jk} become smaller with larger factor loadings. Since each element of $\Delta\boldsymbol{\theta}^{(t)}$ is proportional to relative errors in $\mathbf{S}$, results in (7) and (8) explain why larger factor loadings lead to smaller elements of $\Delta\boldsymbol{\theta}^{(t)}$ and consequently more convergent replications, as reported in the literature.

3.2 Smaller relative errors via the ridge and anti-ridge methods

The results in the previous subsection characterize the relationship of relative errors in s_{jk} with the population factor loadings and error variances that generated the data. The ridge method of modeling $\mathbf{S}_a = \mathbf{S} + a\mathbf{I} = (s_{jka})$ is a post-hoc technique after $\mathbf{S} = (s_{jk})$ is obtained. Since a is a constant, the variance of s_{jka} is the same as that of s_{jk}. However, $E(s_{jja}) = E(s_{jj}) + a$. Consequently, the relative error in s_{jja} monotonically decreases with a. Thus, the $\Delta\boldsymbol{\theta}^{(t)}$ in (2) following the ridge method becomes smaller, which increases both the speed and rate of convergence of the FS algorithm.

The manipulations on the size of factor loadings and error variances in Anderson and Gerbing (1984) and Boomsma (1985) are a priori. We may consider applying the anti-ridge method in a post-hoc manner after $\mathbf{S}$ is observed. Suppose we know the variance-covariance matrix of the common scores $\boldsymbol{\Lambda\xi}$. Then we may consider fitting $\mathbf{S}_c = (s_{jkc}) = \mathbf{S} + (c\boldsymbol{\Lambda})\boldsymbol{\Phi}(c\boldsymbol{\Lambda})' = \mathbf{S} + c^2\boldsymbol{\Lambda\Phi\Lambda}'$ by the structural model $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Since $\boldsymbol{\Sigma}_c = E(\mathbf{S}_c) = (1 + c^2)\boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi}$ and $\mathrm{Var}(s_{jkc}) = \mathrm{Var}(s_{jk})$, relative errors in all the elements of $\mathbf{S}_c$ are smaller than those of $\mathbf{S}$. Thus, we expect that the post-hoc use of the anti-ridge method is more effective than the ridge method in improving the convergence properties of FS. However, such a technique can only be used in certain applications where the factor loadings are not subject to estimation.

Example 1 Consider the linear latent growth curve model (Preacher et al. 2008) $y_j = \xi_1 + (j-1)\xi_2 + \varepsilon_j$, $j = 1, 2, \ldots, p$, where ξ_1 is the latent intercept, ξ_2 is the latent slope, with $E(\xi_1) = \tau_1$, $E(\xi_2) = \tau_2$, $\mathrm{Var}(\xi_1) = \phi_{11}$, $\mathrm{Var}(\xi_2) = \phi_{22}$, and $\mathrm{Cov}(\xi_1, \xi_2) = \phi_{12}$, resulting in a covariance structure as in (6), where all the elements of the first column of $\mathbf{\Lambda}$ are 1.0, and those of the second column of $\mathbf{\Lambda}$ are $0, 1, 2, \ldots, p-1$ in sequence; $\mathbf{\Phi}$ is a free matrix subject to estimation; and $\mathbf{\Psi}$ is a diagonal matrix. For growth curve modeling, there is also a mean structure $\mu_j = \tau_1 + (j-1)\tau_2$. Then, with the same rationale as for just covariance structure analysis, keeping the sample means the same and treating $\mathbf{S}_c = \mathbf{S} + c^2\mathbf{\Lambda}\mathbf{\Lambda}'$ as the new sample covariance matrix will increase the likelihood for the Fisher-scoring algorithm to converge. Except for the estimate of $\mathbf{\Phi}$, all other estimates obtained from fitting $(\bar{\mathbf{y}}, \mathbf{S}_c)$ by the mean and covariance structure model are consistent, and one can get a consistent estimate of $\mathbf{\Phi}$ by $\hat{\mathbf{\Phi}} = \hat{\mathbf{\Phi}}_c - c^2\mathbf{I}$, where $\hat{\mathbf{\Phi}}_c$ is the estimate of $\mathbf{\Phi}$ under modeling $\mathbf{S}_c$.

In summary, both the ridge and the anti-ridge methods alleviate the non-convergence problems caused by (C3), as discussed in Sect. 2. If the problem of a nearly rank deficient $\dot{\boldsymbol{\sigma}}(\boldsymbol{\theta}^{(t)})$ is due to extreme or improper values of $\boldsymbol{\theta}^{(t)}$, caused by large sampling errors, then the two methods also alleviate the non-convergence problems caused by (C2). When non-convergence is due to the fluctuations of certain elements of $\Delta\boldsymbol{\theta}^{(t)}$ from iteration to iteration caused by the interaction of sizeable errors in $\mathbf{S}$ and the conditions of $\mathbf{\Sigma}^{(t)}$ and/or $\mathbf{H}^{(t)}$, then the two methods also address the problems caused by (C4).

3.3 Condition numbers following the ridge and anti-ridge methods

In addition to affecting relative errors in FS, the ridge and anti-ridge methods also affect condition numbers of the model covariance and information matrices. In the following discussion, $\mathbf{\Sigma}^{(t)}$ and $\mathbf{H}^{(t)}$ are used to denote the model covariance and information matrices corresponding to modeling $\mathbf{S}$; $\mathbf{\Sigma}_a^{(t)}$ and $\mathbf{H}_a^{(t)}$, and $\mathbf{\Sigma}_c^{(t)}$ and $\mathbf{H}_c^{(t)}$ are used to denote those corresponding to modeling $\mathbf{S}_a$ and $\mathbf{S}_c$, respectively.

When $\mathbf{S}$ is close to being singular or ill-conditioned, because the algorithm is to approximate $\mathbf{S}$ by $\mathbf{\Sigma}^{(t)}$ as close as possible, $\mathbf{\Sigma}^{(t)} \approx \mathbf{S}$ will be very likely close to singular or not invertible from iteration to iteration, and so will be $\mathbf{W}(\boldsymbol{\theta}^{(t)})$ and the corresponding $\mathbf{H}^{(t)}$. Since $\mathbf{S}_a = \mathbf{S} + a\mathbf{I}$, there exists $\mathbf{\Sigma}_a^{(t)} \approx \mathbf{\Sigma}^{(t)} + a\mathbf{I}$. Thus, with an appropriate a, $\mathbf{\Sigma}_a^{(t)}$ in the ridge method is always well-conditioned. But this is not always true for $\mathbf{H}_a^{(t)}$. When an ill-conditioned $\mathbf{H}^{(t)}$ is due to an ill-conditioned $\mathbf{\Sigma}^{(t)}$, not a rank deficient $\dot{\boldsymbol{\sigma}}(\boldsymbol{\theta}^{(t)})$, then $\mathbf{H}_a^{(t)}$ will be well-conditioned. However, if an ill-conditioned $\mathbf{H}^{(t)}$ is caused by a rank deficient $\dot{\boldsymbol{\sigma}}(\boldsymbol{\theta}^{(t)})$, then the condition of $\mathbf{H}_a^{(t)}$ may not improve. This is because the ridge constant a mainly affects only the variances of errors, and their values do not have any effect on the Jacobian matrix $\dot{\boldsymbol{\sigma}}(\boldsymbol{\theta}^{(t)})$.

Another scenario in which the ridge method improves the condition numbers of $\mathbf{\Sigma}^{(t)}$ and $\mathbf{H}^{(t)}$ is when $\mathbf{S}$ is well-conditioned but its elements contain substantial sampling or systematic errors, which cause some $\psi_{jj}^{(t)}$ to be close to zero or negative during the iterative process. Then $\mathbf{\Sigma}(\boldsymbol{\theta}^{(t)})$ might be close to singular and so might be $\mathbf{H}^{(t)}$. Again, because $\mathbf{\Sigma}_a^{(t)} \approx \mathbf{\Sigma}^{(t)} + a\mathbf{I}$, both $\mathbf{\Sigma}_a^{(t)}$ and $\mathbf{H}_a^{(t)}$ will be well-conditioned.

A third scenario where the ridge method improves the condition numbers of $\mathbf{\Sigma}^{(t)}$ and $\mathbf{H}^{(t)}$ is through reducing relative errors in $\mathbf{S}_a$. Certain elements of $\boldsymbol{\theta}^{(t)}$ can become extreme when a model is not a good representation of the data, due to sampling or systematic errors. Extreme elements other than variances of errors in $\boldsymbol{\theta}^{(t)}$ can also make $\mathbf{\Sigma}^{(t)}$ close to singular or $\dot{\boldsymbol{\sigma}}(\boldsymbol{\theta}^{(t)})$ close to rank deficient. Due to smaller relative errors in $\mathbf{S}_a$, $\boldsymbol{\theta}_a^{(t)}$ becomes less extreme and results in well-conditioned $\mathbf{\Sigma}_a^{(t)}$ and $\mathbf{H}_a^{(t)}$.

The changes in condition numbers of the model covariance and information matrices following the anti-ridge method are much more complicated than those following the ridge method. This is because there is no simple relationship between the eigenvalues of $\mathbf{S}$ and those of $\mathbf{S}_c = \mathbf{S} + c^2\mathbf{\Lambda\Phi\Lambda}'$. Although $\kappa(\mathbf{S}_c)$ is not necessarily always greater than $\kappa(\mathbf{S})$, we would expect $\kappa(\mathbf{S}_c)$ to be greater than $\kappa(\mathbf{S})$ most of the times in practice. However, $\kappa(\mathbf{H}_c^{(t)})$ is not necessarily greater than $\kappa(\mathbf{H}^{(t)})$ even if $\kappa(\mathbf{S}_c) > \kappa(\mathbf{S})$. A scenario where $\mathbf{\Sigma}_c^{(t)}$ and $\mathbf{H}_c^{(t)}$ are better conditioned than $\mathbf{\Sigma}^{(t)}$ and $\mathbf{H}^{(t)}$ is when elements of $\boldsymbol{\theta}^{(t)}$ are extreme due to substantial relative errors in $\mathbf{S}$. Similar to the third scenario with the ridge method, with smaller relative errors in $\mathbf{S}_c$, $\boldsymbol{\theta}_c^{(t)}$ becomes less extreme and results in well-conditioned $\mathbf{\Sigma}_c^{(t)}$ and $\mathbf{H}_c^{(t)}$. We will show the changes in condition numbers due to anti-ridge manipulations using examples and Monte Carlo simulations in the next two sections.

Another interesting fact is that the condition number $\kappa(\mathbf{H}^{(t)})$ is not invariant with respect to rescaling of $\mathbf{\Sigma}$ by a constant. That is, $\kappa(\mathbf{H}^{(t)})$ may change when all the elements in $\mathbf{S}$ or $\mathbf{\Sigma}(\boldsymbol{\theta})$ change proportionally. We will illustrate such a property of condition numbers in Sect. 5.

Having discussed how non-convergence problems are affected by condition numbers of $\mathbf{\Sigma}^{(t)}$ and/or $\mathbf{H}^{(t)}$ caused through (C1) and (C4), we would like to note that, when the model $\mathbf{\Sigma}(\boldsymbol{\theta})$ is in a neighborhood of $\mathbf{S}$, a large but not extreme $\kappa(\mathbf{S})$ alone may not cause a non-convergence problem although it affects the accuracy of parameter estimates. The value of $\kappa(\mathbf{S})$ affects but does not determine the value of $\kappa(\mathbf{\Sigma}^{(t)})$ or $\kappa(\mathbf{H}^{(t)})$. The effect of $\kappa(\mathbf{S})$ on convergence is mostly through its interactions with the model and/or the relative errors in $\mathbf{S}$.

4 Numerical examples

In this section, we consider two examples. Fisher-scoring algorithm repeatedly has non-convergence problems in simulation studies, especially at smaller N. The data (sample covariance matrices) for the examples are just two samples (replications) from our simulation studies. The first example involves a one-factor model with $p = 4$ variables, and the second example involves a structural equation model with $p = 6$ variables and two latent factors. When estimating each of the models, the FS algorithm implemented in SAS IML[1] cannot reach convergence. In addition to the FS algorithm, we will also use the commercial programs EQS (Bentler 2008) and SAS Calis (SAS Institute Inc 2011) to estimate the models in the two examples.

Example 2 The sample covariance matrix

$$\mathbf{S} = \begin{pmatrix} 1.436 & .176 & .506 & .120 \\ .176 & 1.681 & 1.153 & .616 \\ .506 & 1.153 & 1.278 & .243 \\ .120 & .616 & .243 & 1.946 \end{pmatrix} \tag{11}$$

is obtained from a normally distributed population with sample size $N = 30$. The population covariance matrix satisfies a one-factor model with all the factor loadings, the factor variance, and all the error variances being at 1.0. In estimating the model, we fix the factor variance at 1.0 for model identification. Thus, the model has 8 parameters with $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \psi_{11}, \psi_{22}, \psi_{33}, \psi_{44})'$. The population counterpart of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_0 = (1, 1, 1, 1, 1, 1, 1, 1)'$, which is used as the initial value of the FS and other algorithms described below, even when $\mathbf{S}$ is replaced by $\mathbf{S}_a = \mathbf{S}+a\mathbf{I}$ or $\mathbf{S}_c = \mathbf{S}+c^2\mathbf{1}\mathbf{1}'$, where $\mathbf{1} = (1, 1, 1, 1)'$ with $\mathbf{1}\mathbf{1}'$ being the population covariance matrix of the common scores. The criterion for convergence of the FS algorithm is defined as

$$\max |\Delta\boldsymbol{\theta}^{(t)}| < .0001 \text{ within 300 iterations,} \tag{12}$$

where $\max |\cdot|$ is the maximum absolute value on all the elements of $\Delta\boldsymbol{\theta}^{(t)}$. EQS and SAS Calis have their own convergence criteria that are different from (12).

When fitting the $\mathbf{S}$ in (11) by the one-factor model, at the 60th iteration ($t = 60$), FS as implemented in SAS IML declares that the information matrix $\mathbf{H}^{(t)}$ cannot be inverted, with $\kappa(\boldsymbol{\Sigma}^{(t)}) \approx 3.74 \times 10^8$ and $\kappa(\mathbf{H}^{(t)}) \approx 1.56 \times 10^{17}$. Initial values other than $\boldsymbol{\theta}_0$ are also used and FS runs into the same problem. The problem of singular information matrix encountered by FS in this example may also occur to other iterative methods. We may want to address such a problem by adjusting the step size as in (4) with a proper value of α, which is the default implementation in EQS. With the default convergence criterion of EQS (conv = .001), the program converges in 261 iterations and yields

$$\hat{\boldsymbol{\theta}} = (.049, .142, 8.267, -.027, 1.434, 1.661, -67.031, 1.945)'. \tag{13}$$

The output of EQS indicates that during the iteration process step size is adjusted many times with the α in (4) ranging from 1.0 to .001. However, using the $\hat{\boldsymbol{\theta}}$ in (13) as the initial value $\boldsymbol{\theta}^{(0)}$, FS declares that $\mathbf{H}^{(t)}$ is singular at $t = 16$, with $\kappa(\mathbf{H}^{(t)}) = -3.1 \times 10^{16}$. To better understand the problem, we reset the convergence criterion in EQS to conv = .000001. Then EQS cannot reach convergence in 1000 iterations.

[1] The implementation of the FS algorithm as described in Eqs. (2) and (3) is straightforward. Readers are welcome to contact the authors to obtain an electronic copy of the SAS IML code.

The program SAS Calis uses the so-called Levenberg-Marquardt optimization method (Nocedal and Wright 1999, pp. 262–266) with step size adjustment. With the default convergence criterion of Calis and setting the maximum number of iterations at 1000, SAS Calis gives an error message "LEVMAR Optimization cannot be completed". at the end of the 1000 iterations.

For the $\mathbf{S}$ in (11), fitting the $\mathbf{S}_a$ at $a = .5$ by the one-factor model, FS converges in 82 iterations and yields

$$\hat{\boldsymbol{\theta}} = (.388, .933, 1.236, .243, 1.285, .811, -.250, 1.887)', \tag{14}$$

where the estimates of error variances are obtained by subtracting a from each of their values directly from the FS algorithm for the purpose of consistency. Notice that the solution in (14) contains a negative estimate of error variance (called Heywood case in factor analysis). Fitting this $\mathbf{S}_a$ by the one-factor model with conv $= .000001$, after 126 iterations EQS obtains estimates with the first 3 decimals identical to those in (14).

Fitting the $\mathbf{S}_a$ at $a = 1$ by the one-factor model, FS converges in 29 iterations and yields

$$\hat{\boldsymbol{\theta}} = (.335, 1.089, 1.059, .380, 1.324, .495, .157, 1.801)'. \tag{15}$$

Fitting this $\mathbf{S}_a$ by the one-factor model with conv $= .000001$, EQS converges in 46 iterations and yields estimates with the first 3 decimals identical to those in (15).

We next apply the anti-ridge method by fitting the one-factor model to $\mathbf{S}_c = \mathbf{S} + c^2\mathbf{1}\mathbf{1}'$ at $c^2 = .5$ and 1.0. These values of c^2 are chosen so that the variances-covariances of the common scores are increased by respectively 50 and 100 %, corresponding to parallel increases of error variances at $a = .5$ and 1.0. Let $\hat{\boldsymbol{\lambda}}_c$ be the vector of estimates of the factor loadings corresponding to the solution of fitting $\mathbf{S}_c$. The anti-ridge estimates of factor loadings reported below are obtained by $\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}_c - [(1 + c^2)^{1/2} - 1]\mathbf{1}$, while the estimates of error variances are not changed. Thus, all the reported estimates following the anti-ridge method are consistent.

Fitting the one-factor model to $\mathbf{S}_c$ at $c^2 = .5$, FS converges in 37 iterations and yields

$$\hat{\boldsymbol{\theta}} = (.529, .992, 1.134, .285, 1.368, .700, -.069, 2.186)', \tag{16}$$

which also contains a Heywood case. For this $\mathbf{S}_c$ and the one-factor model, with conv $= .000001$, EQS converges in 63 iterations and yields the same estimates as in (16) for the first 3 decimal places. Fitting the one-factor model to the $\mathbf{S}_c$ at $c^2 = 1.0$, FS converges in 20 iterations and yields

$$\hat{\boldsymbol{\theta}} = (.583, 1.025, 1.081, .438, 1.441, .609, .042, 2.220)'. \tag{17}$$

Fitting this $\mathbf{S}_c$ by the one-factor model with conv $= .000001$, EQS converges in 34 iterations and yields estimates with the first 3 decimals identical to those in (17).

Table 1 contains the condition numbers of $\mathbf{S}$, $\mathbf{S}_a$ ($a = .5$, 1.0) and $\mathbf{S}_c$ ($c^2 = .5$, 1.0), called input covariance (ICov) matrix in the table. Condition numbers of the estimated covariance (ECov) matrix $\hat{\boldsymbol{\Sigma}}$ or $\boldsymbol{\Sigma}^{(60)}$ as well as the corresponding estimated information (EInf) matrix are also reported in the table. With $\kappa(\mathbf{S}) = 14.645$, the sample

Table 1 Condition numbers of the input covariance (ICov) matrix, the estimated covariance (ECov) matrix, and the estimated information (EInf) matrix corresponding to $\mathbf{S}$, $\mathbf{S}_a$ and $\mathbf{S}_c$ in Example 3

	$\mathbf{S}$	$\mathbf{S}_{a=.5}$	$\mathbf{S}_{a=1}$	$\mathbf{S}_{c^2=.5}$	$\mathbf{S}_{c^2=1}$
ICov	14.645	5.097	3.410	23.571	32.754
ECov		4.474	3.158	16.950	23.231
EInf		95.454	47.880	58.328	48.462
$\mathrm{ECov}_{(60)}$	3.74×10^8				
$\mathrm{EInf}_{(60)}$	1.56×10^{17}				

covariance matrix is not ill conditioned. However, the sampling errors in $\mathbf{S}$ cause $\mathbf{\Sigma}^{(t)}$ and $\mathbf{H}^{(t)}$ to be close to singular, which further causes FS and other algorithms to fail to converge. The results in Table 1 indicate that $\kappa(\hat{\mathbf{\Sigma}}_a)$ decreases as a increases and $\kappa(\hat{\mathbf{\Sigma}}_c)$ increases as c^2 increases. However, the condition number of the information matrix corresponding to $\mathbf{S}_c$ at $c^2 = 1.0$ is smaller than that at $c^2 = .5$. The condition number of the information matrix corresponding to $\mathbf{S}_a$ at $a = .5$ ($c^2 = 0$) is also much larger than that corresponding to $\mathbf{S}_c$ at either $c^2 = .5$ or 1.0.

In Example 2, neither FS nor the default algorithm in SAS Calis is able to reach convergence when fitting $\mathbf{S}$ by the 1-factor model. The seeming convergence of EQS at conv $= .001$ is just a coincidence, with the $\hat{\boldsymbol{\theta}}$ in (13) being not a stationary point. By using either the ridge or anti-ridge method, FS easily yields converged solutions. However, the estimates by different methods are quite different. Although both the ridge and the anti-ridge methods yield estimates that are consistent with the model and population, the $\mathbf{S}$ in (11) contains substantial sampling errors, which cause the differences among the estimates in (14) to (17). The example shows that, in addition to yielding converged solutions, ridge and anti-ridge methods can be also effective in removing Heywood cases.

Notice that the convergence criterion in EQS is defined differently from that in (12). The reason for us to choose conv $= .000001$ when using EQS to fit $\mathbf{S}_a$ and $\mathbf{S}_c$ is because the program was unable to reach convergence when working with $\mathbf{S}$ under the same value of conv. If setting conv at a larger number, it will take fewer iterations for EQS to reach convergence.

The previous example is on the convergence issue of the FS algorithm with a confirmatory factor model. The FS algorithm has similar problems when fitting a structural equation model, as illustrated by the following example.

Example 3 Consider six variables with y_1, y_2 and y_3 being indicators for the first factor ξ_1; y_4, y_5 and y_6 being indicators for the second factor ξ_2; and ξ_2 is predicted by ξ_1 according to

$$\xi_2 = \gamma_{21}\xi_1 + \zeta_2,$$

where ξ_1 and ζ_2 are independent with $\phi_{11} = \mathrm{Var}(\xi_1)$ and $\varphi_{22} = \mathrm{Var}(\zeta_2)$. Letting $\boldsymbol{\xi} = (\xi_1, \xi_2)'$, then the covariance structure of $\mathbf{y} = (y_1, y_2, \ldots, y_6)'$ is given by Eq. (6) with

$$\boldsymbol{\Lambda}' = \begin{pmatrix} 1.0 & \lambda_2 & \lambda_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.0 & \lambda_5 & \lambda_6 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Phi} = \begin{pmatrix} \phi_{11} & \gamma_{21}\phi_{11} \\ \gamma_{21}\phi_{11} & \gamma_{21}^2\phi_{11} + \varphi_{22} \end{pmatrix}. \tag{18}$$

Note that, for the above structural equation model, we cannot fix the variance of ξ_2 at 1.0 becuase its value is subject to prediction. So we put $\lambda_1 = \lambda_4 = 1.0$ for the purpose of model identification. In this model, there are 13 free parameters with

$$\boldsymbol{\theta} = (\lambda_2, \lambda_3, \lambda_5, \lambda_6, \phi_{11}, \gamma_{21}, \varphi_{22}, \psi_{11}, \psi_{22}, \psi_{33}, \psi_{44}, \psi_{55}, \psi_{66})'.$$

Like for Example 2, the sample covariance matrix

$$\mathbf{S} = \begin{pmatrix} 3.498 & 1.686 & .679 & 1.033 & 1.426 & .286 \\ 1.686 & 1.964 & 1.062 & .718 & .654 & .642 \\ .679 & 1.062 & 1.754 & .863 & 1.036 & .720 \\ 1.033 & .718 & .863 & 2.061 & 1.402 & .779 \\ 1.426 & .654 & 1.036 & 1.402 & 2.284 & 1.380 \\ .286 & .642 & .720 & .779 & 1.380 & 2.646 \end{pmatrix} \tag{19}$$

is obtained from a normally distributed population with $N = 30$. Except $\gamma_{21} = .5$ and $\varphi_{22} = .75$, all the other elements of $\boldsymbol{\theta}$ in the population are 1.0; and we denote the vector of these values as $\boldsymbol{\theta}_0$. The initial value of the FS and other algorithms described below are set at $\boldsymbol{\theta}_0$ regardless of whether the ridge or anti-ridge method is used when estimating the model in (18). The criterion for convergence of the FS algorithm is the same as defined in (12).

When fitting the model in (18) to the sample covariance matrix in (19), the FS algorithm does not converge. Starting at 66th iteration ($t = 66$), the $\boldsymbol{\theta}^{(t)}$ in Eq. (2) oscillates between

$$\boldsymbol{\theta}^{(t)} = (.668, .633, 1.346, .972, 2.054, .366, .690, 1.444, .832, \\ .926, 1.040, .384, 1.688)' \tag{20}$$

and

$$\boldsymbol{\theta}^{(t+1)} = (.983, .640, 1.535, .948, 1.379, .509, .514, 2.119, .483, 1.184, \\ 1.148, .139, 1.828)'. \tag{21}$$

Other initial values are also used but FS eventually runs into the same problem.

We would hope that the problem of oscillation between two points encountered by FS in this example would be solved by adjusting the step size via the value of α in (4). With the default convergence criterion of EQS (conv = .001), the program converges in 12 iterations and yields

$$\hat{\boldsymbol{\theta}} = (.887, .622, 1.445, .963, 1.799, .415, .662, 1.699, .549, 1.058, \\ 1.089, .256, 1.745)'. \tag{22}$$

The output of EQS indicates that, at the 4th and 10th iterations, step halving is used. However, with the $\hat{\boldsymbol{\theta}}$ in (22) as the initial values, the FS algorithm in (2) returns to oscillating between the two sets of values in (20) and (21), starting at $t = 134$. To better understand the problem, we reset the convergence criterion in EQS to conv = .00001. Then EQS cannot reach convergence in 1000 iterations.

With the default convergence criterion, SAS Calis yields a vector of converged values essentially the same as that of EQS. However, using the converged values of SAS Calis as the initial values for FS and let the algorithm continue to run, the iteration returns to oscillating between the two sets of values in (20) and (21).

Since the relative errors in s_{jk} do not depend on how the structure model is identified, the results and properties regarding ridge and anti-ridge methods obtained in the previous sections still hold, as illustrated below.

For the $\mathbf{S}$ in (19), fitting the $\mathbf{S}_a = \mathbf{S} + a\mathbf{I}$ at $a = .5$ by the model in (18), the FS algorithm converges in 50 iterations and yields

$$\hat{\boldsymbol{\theta}} = (.877, .633, 1.438, .928, 1.758, .462, .609, 1.740, .613, 1.050,\\ 1.077, .248, 1.798)'. \tag{23}$$

Fitting the $\mathbf{S}_a$ with $a = .5$ and conv = .00001 by the model in (18), EQS converges in 59 iterations and yields estimates with the first 3 decimals identical to those in (23).

Fitting the $\mathbf{S}_a$ at $a = 1$ by the SEM model in (18), the FS algorithm converges in 28 iterations and yields

$$\hat{\boldsymbol{\theta}} = (.875, .644, 1.430, .904, 1.725, .484, .593, 1.773, .643, 1.039,\\ 1.063, .244, 1.830)'. \tag{24}$$

Fitting the $\mathbf{S}_a$ at $a = 1$ by the SEM model and set conv = .00001, EQS converges in 33 iterations and yields estimates with the first 3 decimals identical to those in (24).

Let $\mathbf{S}_c = \mathbf{S} + c^2\boldsymbol{\Lambda\Phi\Lambda}'$, where $\boldsymbol{\Lambda\Phi\Lambda}'$ is the population covariance matrix of the common scores. Because λ_1 and λ_4 are fixed at 1.0 in the formulation of the SEM model in (18), the population counterpart of $\boldsymbol{\Phi}$ corresponding to $\mathbf{S}_c$ becomes $(1+c^2)\boldsymbol{\Phi}$, and those of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ remain the same. Thus, consistent estimates of ϕ_{11} and φ_{22} are obtained by $\hat{\phi}_{11} = \hat{\phi}_{c11} - c^2\phi_{110}$ and $\hat{\varphi}_{c22} = \hat{\varphi}_{22} - c^2\varphi_{220}$, where $\hat{\phi}_{c11}$ and $\hat{\varphi}_{c22}$ are the estimates of ϕ_{11} and φ_{22} under modeling $\mathbf{S}_c$; and ϕ_{110} and φ_{220} are the population values of ϕ_{11} and φ_{22}, respectively.

Fitting the SEM model to the $\mathbf{S}_c$ at $c^2 = .5$, FS converges in 135 iterations and yields

$$\hat{\boldsymbol{\theta}} = (.973, .715, 1.318, .977, 1.690, .417, .687, 1.808, .389, 1.135,\\ 1.119, .278, 1.770)'. \tag{25}$$

With conv = .00001, fitting the $\mathbf{S}_c$ by the SEM model in (18) using EQS takes 160 iterations and yields estimates with the first 3 decimals identical to those in (25).

Fitting the SEM model to the $\mathbf{S}_c$ at $c^2 = 1.0$, FS converges in 44 iterations and yields

Table 2 Condition numbers of the input covariance (ICov) matrix, the estimated covariance (ECov) matrix, and the estimated information (EInf) matrix corresponding to $\mathbf{S}$ ($t \geq 22$), $\mathbf{S}_a$ and $\mathbf{S}_c$ in Example 2

	$\mathbf{S}$	$\mathbf{S}_{a=.5}$	$\mathbf{S}_{a=1}$	$\mathbf{S}_{c^2=.5}$	$\mathbf{S}_{c^2=1}$
ICov	31.381	10.662	6.744	38.504	46.123
ECov		6.652	4.999	13.496	16.722
EInf		76.999	85.052	99.001	150.535
$\text{ECov}_{(t)}$	9.462				
$\text{ECov}_{(t+1)}$	10.915				
$\text{EInf}_{(t)}$	71.713				
$\text{EInf}_{(t+1)}$	78.224				

$$\hat{\boldsymbol{\theta}} = (1.006, .769, 1.249, .984, 1.633, .421, .705, 1.865, .299, 1.196, 1.139.287, 1.786)'. \tag{26}$$

With conv = .00001, fitting the SEM model to the $\mathbf{S}_c$ at $c^2 = 1$ by EQS takes 53 iterations and yields estimates with the first 3 decimals identical to those in (26).

Parallel to Table 1, the condition numbers of $\mathbf{S}$, $\mathbf{S}_a$ ($a = .5$, 1.0) and $\mathbf{S}_c$ ($c^2 = .5$, 1.0) as well as those of the corresponding $\hat{\boldsymbol{\Sigma}}$ or $\boldsymbol{\Sigma}^{(t)}$ and the associated information matrices for this example are reported in Table 2. The sample covariance matrix is not ill conditioned. But $\kappa(\mathbf{S}) = 31.381$ is several times of that in Table 1. The results in Table 2 indicate that $\kappa(\hat{\boldsymbol{\Sigma}}_a)$ decreases as a increases and $\kappa(\hat{\boldsymbol{\Sigma}}_c)$ increases as c^2 increases, which are similarly observed in Table 1. However, in Table 2 the condition number of the information matrix in fitting $\mathbf{S}_c$ increases with c^2, and is the largest at $c^2 = 1.0$. This may explain why anti-ridge method in this example is not as effective as in the previous example, and it took 135 iterations for the FS algorithm to converge when fitting $\mathbf{S}_c$ at $c^2 = .5$, compared to 50 iterations when fitting $\mathbf{S}_a$ at $a = .5$.

In this example, FS algorithm is unable to reach convergence when $\mathbf{S}$ is fitted by the SEM model. Step size adjustment does not solve the problem, as shown by running EQS with conv = .00001. With or without step size adjustment, FS has no problem in reaching a converged solution with either the ridge or the anti-ridge method. Although the $\mathbf{S}$ in (19) contains substantial sampling errors, the 4 sets of estimates in (23) to (26) are comparable.

Notice that the convergence criterion of EQS with fitting $\mathbf{S}_a$ and $\mathbf{S}_c$ in this example is set at conv = .00001 while in the previous example it was set at conv = .000001. This is because, when working with $\mathbf{S}$, EQS could not reach convergence at these specified values.

5 Monte Carlo results

In this section, we empirically compare the convergence rate and speed of the FS algorithm in fitting $\mathbf{S}$, $\mathbf{S}_a$ and $\mathbf{S}_c$. In Sect. 3.1, our characterization of the relative errors in $\mathbf{S}$ is based on the assumption that factors and errors in the factor model are independent, and the results are derived using asymptotics. We will empirically evaluate the size of relative errors in $\mathbf{S}$ when factors and errors are dependent but uncorrelated. Because the convergence properties of FS are related to the condition numbers of $\mathbf{S}$ and/or the information matrix in (3), we will also evaluate how these condition numbers and sampling errors jointly affect the convergence rate and speed of the FS algorithm.

5.1 Conditions

The population distributions are specified through a confirmatory factor model with $p = 15$ observed variables

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}(r\boldsymbol{\xi}) + r\boldsymbol{\varepsilon}, \tag{27}$$

where $\boldsymbol{\mu}$ is a 15×1 vector of means;

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\lambda} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\lambda} \end{pmatrix}$$

with $\boldsymbol{\lambda}$ being a 5×1 vector of factor loadings; $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{15})'$ are independent with $E(\boldsymbol{\xi}) = \mathbf{0}$,

$$\boldsymbol{\Phi} = \mathrm{Var}(\boldsymbol{\xi}) = (\phi_{jk}) = \begin{pmatrix} 1.0 & .3 & .4 \\ .3 & 1.0 & .5 \\ .4 & .5 & 1.0 \end{pmatrix},$$

$E(\boldsymbol{\varepsilon}) = \mathbf{0}$, and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi} = \mathrm{diag}(\psi_{11}, \psi_{22}, \ldots, \psi_{pp})$. The multiplier r in (27), to be further specified, is to make the factors $(r\boldsymbol{\xi})$ and errors $(r\boldsymbol{\varepsilon})$ dependent but uncorrelated. Such a condition was used in Hu et al. (1992) to invalidate the so-called asymptotic robustness conditions in studying the likelihood ratio statistic. Three sets of population parameters are used: (P1) $\boldsymbol{\lambda} = (1, 1, 1, 1, 1)'$ or $\lambda_j = 1$ and $\psi_{jj} = 1$, $j = 1$ to 15; (P2) $\boldsymbol{\lambda} = (2, 2, 2, 2, 2)'$ or $\lambda_j = 2$ and $\psi_{jj} = 1$, $j = 1$ to 15; (P3) $\boldsymbol{\lambda} = (1, 1, 1, 1, 1)'$ or $\lambda_j = 1$ and $\psi_{jj} = 2$, $j = 1$ to 15. Four population distribution conditions of $\mathbf{y}$, as described in Table 3a, are used. Each distribution of $\mathbf{y}$ is defined through $\boldsymbol{\xi} = \boldsymbol{\Phi}^{1/2}\mathbf{z}_\xi$ and $\boldsymbol{\varepsilon} = \boldsymbol{\Psi}^{1/2}\mathbf{z}_\varepsilon$, where the elements in $\mathbf{z}_\xi = (z_{\xi 1}, z_{\xi 2}, z_{\xi 3})'$ and $\mathbf{z}_\varepsilon = (z_{\varepsilon 1}, z_{\varepsilon 2}, \ldots, z_{\varepsilon 15})'$ are independent and each follows a standardized distribution described in the table. In condition D1, $r = 1$, $\boldsymbol{\xi} \sim N(\mathbf{0}, \boldsymbol{\Phi})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi})$, and thus $\mathbf{y}$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix as given in (6). In conditions D2, D3 and D4, $r \sim (3/\chi_5^2)^{1/2}$. Since $E(r^2) = 1$, each $\boldsymbol{\Sigma}$ in D2, D3 or D4 is also given by (6). In D2, $\mathbf{y}$ follows an elliptical distribution. In D3, $\mathbf{y}$ follows a skewed distribution due to a skewed $\boldsymbol{\xi}$. In D4, $\mathbf{y}$ follows a skewed distribution due to a skewed $\boldsymbol{\varepsilon}$.

Since non-convergence is typically associated with smaller sample sizes, we choose $N = 30, 50, 100$ and 200. The number of replications is $N_r = 500$.

To study the convergence properties of FS with ridge and anti-ridge methods, we fit $\mathbf{S}, \mathbf{S}_a = \mathbf{S} + a\mathbf{I}$ and $\mathbf{S}_c = \mathbf{S} + c^2\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$, where $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$ is set at the population values of the variances-covariances of the common scores; $a = 1$ and $c^2 = 1$ for parameterizations P1 and P2; and $a = 2$ and $c^2 = 1$ for parameterization P3. Thus, the choice of a makes the error variances corresponding to $\mathbf{S}_a$ doubled those corresponding to $\mathbf{S}$ in P1, P2 and P3; whereas the choice of c makes the common-score variances-covariances

Table 3 (a) Population distributions of $\mathbf{y}$ for the Monte Carlo study, $\boldsymbol{\xi} = \boldsymbol{\Phi}^{1/2}\mathbf{z}_\xi$ with $\mathbf{z}_\xi = (z_{\xi 1}, z_{\xi 2}, z_{\xi 3})'$, and the $z_{\xi j}$ are independent; $\boldsymbol{\varepsilon} = \boldsymbol{\Psi}^{1/2}\mathbf{z}_\varepsilon$ with $\mathbf{z}_\varepsilon = (z_{\varepsilon 1}, z_{\varepsilon 2}, \ldots, z_{\varepsilon 15})'$, and the $z_{\varepsilon j}$ are independent. (b) Values of parameters in the population, and condition numbers of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_a = E(\mathbf{S}_a)$ and $\boldsymbol{\Sigma}_c = E(\mathbf{S}_c)$ as well as those of the corresponding information matrix $\mathbf{H} = \dot{\boldsymbol{\sigma}}'\mathbf{W}\dot{\boldsymbol{\sigma}}$

(a)

Condition	$z_{\xi j}$	$z_{\varepsilon j}$	r	$\mathbf{y}$
D1	$N(0, 1)$	$N(0, 1)$	1.0	Normal
D2	$N(0, 1)$	$N(0, 1)$	$(3/\chi_5^2)^{1/2}$	Elliptical
D3	Standardized χ_1^2	$N(0, 1)$	$(3/\chi_5^2)^{1/2}$	Skewed $\boldsymbol{\xi}$
D4	$N(0, 1)$	Standardized χ_1^2	$(3/\chi_5^2)^{1/2}$	Skewed $\boldsymbol{\varepsilon}$

(b)

Parameter	$\kappa(\boldsymbol{\Sigma})$	$\kappa(\boldsymbol{\Sigma}_a)$	$\kappa(\boldsymbol{\Sigma}_c)$	$\kappa(\mathbf{H})$	$\kappa(\mathbf{H}_a)$	$\kappa(\mathbf{H}_c)$
P1 ($\lambda_j = 1, \psi_{jj} = 1$)						
$a = 1, c^2 = 1$	10.028	5.514	19.056	7.038	12.387	15.800
P2 ($\lambda_j = 2, \psi_{jj} = 1$)						
$a = 1, c^2 = 1$	37.112	19.056	73.223	33.983	31.029	70.796
P3 ($\lambda_j = 1, \psi_{jj} = 2$)						
$a = 2, c^2 = 1$	5.514	3.257	10.028	12.387	24.456	19.616

corresponding to $\mathbf{S}_c$ doubled those corresponding to $\mathbf{S}$ in the three conditions. Clearly, the conditions contain both post-hoc and a priori implementations of the ridge and anti-ridge methods. Condition numbers for the population covariance matrix and the information matrix corresponding to $\mathbf{S}$, $\mathbf{S}_a$ and $\mathbf{S}_c$ are listed in Table 3b. Notice that the $\kappa(\boldsymbol{\Sigma})$ under P1 equals the $\kappa(\boldsymbol{\Sigma}_c)$ under P3 because $\boldsymbol{\Sigma}_c = 2\boldsymbol{\Sigma}$, but the condition numbers of their corresponding information matrices are not equal.

For each condition in Table 3, the model is the same. That is, a confirmatory 3-factor model as in Eq. (6), each factor has 5 unidimensional indicators and the factors are freely correlated, and the errors are uncorrelated. In the estimation, each factor variance is fixed at 1.0. Thus, there are 15 factor loadings, 3 factor correlations, and 15 error variances.

For a given condition, let s_{ijk} be the sample covariance between the jth and kth variables in the ith replication. Since $E(s_{ijk}) = \sigma_{jk}$ is positive in all the conditions, we use the average

$$\mathrm{RE}_{od} = \frac{1}{N_r}\sum_{i=1}^{N_r}\left[\sum_{j=1}^{p-1}\sum_{k=j+1}^{p}\frac{|s_{ijk} - \sigma_{jk}|}{\sigma_{jk}}\right]/[p(p-1)/2]$$

to measure the relative errors in the off-diagonal elements of the sample covariance matrix $\mathbf{S}$. Similarly, we use

$$\mathrm{RE}_{d} = \frac{1}{N_r}\sum_{i=1}^{N_r}\left[\sum_{j=1}^{p}\frac{|s_{ijj} - \sigma_{jj}|}{\sigma_{jj}}\right]/p$$

to measure the relative errors in the diagonal elements of $\mathbf{S}$.

When fitting $\mathbf{S}$, $\mathbf{S}_a$ or $\mathbf{S}_c$, the population values of $\boldsymbol{\theta}$ corresponding to $\boldsymbol{\Sigma} = E(\mathbf{S})$, $\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma} + a\mathbf{I}$ and $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma} + c^2\boldsymbol{\Lambda\Phi\Lambda}'$ are used as initial values respectively. For some replications, the FS algorithm cannot reach the convergence criterion defined in (12). These non-converged replications can be classified into two types. One is that (12) is still not satisfied after completing 300 iterations, call it type A; and the other is that, for $t < 300$, the ratio of the largest absolute eigenvalue over the smallest absolute eigenvalue of $\mathbf{H}^{(t)}$ is so large that SAS IML declares $\mathbf{H}^{(t)}$ as singular, call it type B. The numbers of replications for each type are recorded to measure the rate of non-convergence/convergence. For each condition, the average number of iterations across all the converged replications (out of 500) is also recorded as an indicator of the speed of convergence of FS.

Average condition numbers of $\mathbf{S}$, $\mathbf{S}_a$ and $\mathbf{S}_c$ for each condition across the 500 replications are recorded in order to examine their relationship to the convergence properties of FS. Similarly, the average condition numbers of the information matrices corresponding to $\mathbf{S}$, $\mathbf{S}_a$ and $\mathbf{S}_c$ across the converged replications are also recorded for each condition.

5.2 Results

Table 4 contains the averages of relative errors in sample variances (RE$_d$) and covariances (RE$_{od}$) across 500 replications of $\mathbf{S}$ for each condition. It is clear that, regardless of whether the population distribution is symmetric or skewed, or whether the errors and factors are independent or just uncorrelated, the RE$_{od}$s in condition P2 ($\lambda_j = 2$,

Table 4 Relative errors in the sample variances (RE$_d$) and covariances (RE$_{od}$) as factor loadings and/or unique variances vary: (D1) normally distributed population; (D2) elliptically distributed population; (D3) distribution with skewed factors and symmetrically distributed errors; (D4) distribution with skewed errors and symmetrically distributed factors

	RE$_d$				RE$_{\text{od}}$			
N	30	50	100	200	30	50	100	200
P1 ($\lambda_j = 1,\ \psi_{jj} = 1$)								
D1	.205	.163	.113	.080	.656	.505	.359	.250
D2	.314	.254	.188	.136	.887	.723	.537	.383
D3	.365	.306	.228	.170	.904	.762	.558	.409
D4	.392	.328	.246	.180	.829	.687	.505	.371
P2 ($\lambda_j = 2,\ \psi_{jj} = 1$)								
D1	.205	.162	.113	.080	.433	.329	.236	.163
D2	.315	.252	.188	.137	.594	.469	.359	.255
D3	.440	.373	.283	.214	.625	.531	.392	.291
D4	.320	.261	.192	.140	.570	.466	.339	.252
P3 ($\lambda_j = 1,\ \psi_{jj} = 2$)								
D1	.206	.163	.113	.080	.963	.745	.528	.369
D2	.313	.255	.188	.136	1.292	1.051	.783	.558
D3	.335	.279	.207	.152	1.301	1.090	.796	.576
D4	.447	.377	.286	.209	1.176	.980	.724	.530

Table 5 The number of replications that the Fisher-scoring algorithm cannot reach convergence within 300 iterations (in front of /) or the information matrices are singular during iterations (after /); (D1) normally distributed population; (D2) elliptically distributed population; (D3) distribution with skewed factors and symmetrically distributed errors; (D4) distribution with skewed errors and symmetrically distributed factors

		P1 ($\lambda_j = 1, \psi_{jj} = 1$)			P2 ($\lambda_j = 2, \psi_{jj} = 1$)			P3 ($\lambda_j = 1, \psi_{jj} = 2$)		
	N	$\mathbf{S}$	$\mathbf{S}_{a=1}$	$\mathbf{S}_{c^2=1}$	$\mathbf{S}$	$\mathbf{S}_{a=1}$	$\mathbf{S}_{c^2=1}$	$\mathbf{S}$	$\mathbf{S}_{a=2}$	$\mathbf{S}_{c^2=1}$
D1	30	5/1						49/24	3/12	
	50							5/1		
D2	30	15/4	1/0					74/38	10/15	4/0
	50	1/0						25/8	4/1	2/0
	100							1/0	1/0	
D3	30	31/23	4/6	2/0				100/62	6/26	8/2
	50	5/6	0/3	1/0				18/25	6/9	1/1
	100							7/1	1/3	
D4	30	17/5	2/2	5/0	2/0		2/0	56/28	8/11	7/2
	50	3/0		1/0				20/5	3/1	3/0
	100			1/0				2/0		
Total		77/39	7/11	10/0	2/0		2/0	357/192	42/78	25/5

$\psi_{jj} = 1$) are smallest and those in condition P3 ($\lambda_j = 1$, $\psi_{jj} = 2$) are largest, consistent with the analytical results obtained in Sect. 3.1. The relative errors in sample variances (RE$_d$) do not follow the same pattern as that for the sample covariances. In particular, under normally or elliptically distributed population, RE$_d$s barely change from P1 to P3. With skewed factors, RE$_d$s are greatest in P2 and smallest in P3. However, with skewed errors/uniquenesses, RE$_d$s are greatest in P3 and smallest in P2. These results are also consistent with our analysis in Sect. 3.1.

Table 5 contains the numbers of non-converged replications of type A (in front of /) and type B (after /), where empty cells correspond to conditions in which all 500 replications converged. We did not include the results for $N = 200$ because all 500 replications converged. It is clear from Table 5 that the number of non-converged replications in fitting $\mathbf{S}$ is closely related to the size of RE$_{od}$s reported in Table 4. In particular, least numbers of non-converged replications occurred under P2 and largest numbers occurred under P3. Within P3, the three conditions with largest RE$_{od}$ in Table 4 (D3, D2, and D4 under $N = 30$) correspond to most non-converged replications in Table 5 (100/62, 74/38, 56/28). The three largest entries of RE$_{od}$ under $N = 50$ in Table 4 (D3, D2, D4 following P3) also correspond to most non-converged conditions following $N = 50$ in Table 5 (18/25, 25/8, 20/5).

Results in Table 5 also show that both post-hoc ridge and anti-ridge methods are effective in addressing the problem of non-convergence with fitting $\mathbf{S}$. Under P1, the anti-ridge method is slightly more effective than the ridge method for conditions D2 and D3 but not for D4. Under P2, only two replications in D4 could not converge with fitting $\mathbf{S}$, and the ridge method solves the problem whereas the anti-ridge method does not. This is because the variances-covariances of the common-scores ($\mathbf{\Lambda\Phi\Lambda}'$)

Table 6 Average number of iterations across converged replications; (D1) normally distributed population; (D2) elliptically distributed population; (D3) distribution with skewed factors and symmetrically distributed errors; (D4) distribution with skewed errors and symmetrically distributed factors

		P1 ($\lambda_j = 1, \psi_{jj} = 1$)			P2 ($\lambda_j = 2, \psi_{jj} = 1$)			P3 ($\lambda_j = 1, \psi_{jj} = 2$)		
	N	$\mathbf{S}$	$\mathbf{S}_{a=1}$	$\mathbf{S}_{c^2=1}$	$\mathbf{S}$	$\mathbf{S}_{a=1}$	$\mathbf{S}_{c^2=1}$	$\mathbf{S}$	$\mathbf{S}_{a=2}$	$\mathbf{S}_{c^2=1}$
D1	30	15.858	10.666	10.264	9.912	7.110	9.004	32.679	19.144	14.146
	50	10.074	8.150	7.944	7.700	6.114	7.266	17.271	12.876	9.856
	100	7.142	6.260	6.226	6.110	5.122	5.862	9.698	8.530	7.306
	200	5.712	5.174	5.150	5.090	4.464	4.936	7.126	6.550	5.926
D2	30	22.154	15.076	12.978	12.432	8.678	10.974	42.655	27.480	19.192
	50	15.389	11.378	10.430	10.010	7.428	9.042	28.084	19.927	13.233
	100	9.988	8.542	7.998	7.678	6.226	7.222	16.355	12.892	9.916
	200	7.492	6.616	6.542	6.268	5.334	5.998	10.198	9.230	7.674
D3	30	26.011	19.351	14.090	14.582	9.726	11.438	41.985	30.376	20.469
	50	15.851	13.372	10.509	10.244	7.930	8.836	31.823	22.390	14.010
	100	10.336	9.030	7.898	7.738	6.640	7.162	16.909	14.621	9.944
	200	7.542	6.904	6.308	6.260	5.496	5.862	10.486	9.482	7.500
D4	30	22.469	14.974	15.123	14.468	8.616	13.309	37.579	23.391	21.149
	50	14.557	11.700	11.220	10.394	7.450	9.592	24.251	18.028	15.183
	100	9.908	8.246	8.333	7.950	6.216	7.574	14.414	12.324	10.206
	200	7.606	6.636	6.708	6.466	5.406	6.234	10.186	9.218	8.194
Ave		13.006	10.130	9.233	8.956	6.747	8.144	21.981	16.029	12.119

corresponding to fitting $\mathbf{S}$ are already rather large in P2, and further enlarging their values does not make much difference. In other words, non-converged replications in fitting $\mathbf{S}$ under P2 are not due to large relative errors in $\mathbf{S}$ but something related to condition numbers of $\mathbf{\Sigma}^{(t)}$ and/or $\mathbf{H}^{(t)}$, and the ridge method directly addresses the problem. In contrast, under P3, because the relative errors (RE_{od}) are rather large (see Table 4) and $\kappa(\mathbf{S})$ (to be discussed) is already quite small, reducing the relative errors by modeling $\mathbf{S}_c$ is more effective than further improving the condition number of $\mathbf{S}$.

Notice that, under conditions P1 and P3 in Table 5, there are more non-converged replications of type A than type B when fitting $\mathbf{S}$, whereas it is the other way around when fitting $\mathbf{S}_a$. This suggests that the ridge method is less effective in dealing with type B non-converged replications. This is because, as reported in Table 3b, the condition numbers $\kappa(\mathbf{\Sigma})$ for the two conditions are already rather small, and $\kappa(\mathbf{H}_a)$ is even greater than $\kappa(\mathbf{H})$. Although the condition number $\kappa(\mathbf{H}_c)$ under P1 or P3 is also greater than the corresponding $\kappa(\mathbf{H})$, the relative errors in $\mathbf{S}$ are effectively reduced by the anti-ridge method, and thus, the number of non-converged replications due to singular $\mathbf{H}^{(t)}$ caused by the size of relative errors in $\mathbf{S}$ becomes smaller.

For each condition, the average number of iterations across the converged replications is reported in Table 6; and a further average across the 4 sample sizes and 4 distribution conditions is reported in the last line of the table. It is clear that both the ridge and anti-ridge methods accelerate the speed of convergence of FS. Under

Table 7 Average condition number of $\mathbf{S}$, $\mathbf{S}_a$ and $\mathbf{S}_c$ across 500 replications; (D1) normally distributed population; (D2) elliptically distributed population; (D3) distribution with skewed factors and symmetrically distributed errors; (D4) distribution with skewed errors and symmetrically distributed factors

		P1 ($\lambda_j = 1$, $\psi_{jj} = 1$)			P2 ($\lambda_j = 2$, $\psi_{jj} = 1$)			P3 ($\lambda_j = 1$, $\psi_{jj} = 2$)		
	N	$\mathbf{S}$	$\mathbf{S}_{a=1}$	$\mathbf{S}_{c^2=1}$	$\mathbf{S}$	$\mathbf{S}_{a=1}$	$\mathbf{S}_{c^2=1}$	$\mathbf{S}$	$\mathbf{S}_{a=2}$	$\mathbf{S}_{c^2=1}$
D1	30	89.655	10.595	128.235	312.983	35.314	481.561	52.861	6.525	69.582
	50	40.088	9.120	66.968	143.114	30.807	253.812	23.034	5.532	35.890
	100	22.913	7.803	41.337	83.316	26.658	157.821	12.883	4.673	21.942
	200	16.905	6.975	31.370	62.009	23.963	120.156	9.402	4.149	16.579
D2	30	140.995	11.799	184.214	470.879	37.857	683.019	88.350	7.630	102.182
	50	57.645	10.278	90.743	199.283	33.473	340.226	34.805	6.437	49.474
	100	30.065	8.619	52.336	106.914	28.898	198.543	17.481	5.298	28.078
	200	20.591	7.563	37.496	74.685	25.747	143.161	11.643	4.555	19.914
D3	30	146.509	12.072	183.669	490.817	39.061	683.367	91.851	7.762	101.878
	50	58.959	10.469	90.425	205.039	34.609	340.444	35.477	6.570	49.236
	100	30.164	8.609	52.037	107.976	29.079	197.952	17.470	5.271	27.869
	200	20.671	7.580	37.313	75.185	25.900	142.590	11.669	4.556	19.805
D4	30	210.594	12.452	275.199	692.170	39.132	1014.366	136.183	8.277	155.461
	50	78.375	10.907	121.109	263.658	35.026	450.915	49.113	7.083	67.183
	100	37.351	9.092	64.639	131.849	30.311	244.648	22.125	5.670	34.896
	200	23.737	7.859	43.330	85.766	26.670	165.269	13.553	4.771	23.070
Ave		64.076	9.487	93.776	219.103	31.407	351.116	39.244	5.922	51.440

P1, the anti-ridge method is slightly faster than the ridge method on average. Under P2, the ridge method is uniformly faster than the anti-ridge method; and, under P3, the anti-ridge method is uniformly faster than the ridge method. Comparing Tables 6 and 5, we may notice that conditions under which FS converges faster also tend to have smaller numbers of non-converged replications. This implies that both the speed and rate of convergence of FS are strongly affected by the relative errors in sample covariances.

The average condition numbers of $\mathbf{S}$, $\mathbf{S}_a$ and $\mathbf{S}_c$ across the 500 replications for each condition are reported in Table 7. Due to sampling errors, each of the averages is much greater than the corresponding population condition number reported in Table 3b, especially in condition D4 when errors follow a skewed distribution. The average condition numbers monotonically decrease as N increases, but they are still substantially above the population values even at $N = 200$. Further averages across the 4 sample sizes and 4 distribution conditions are reported in the last row of Table 7, and those under the ridge method are less than two times of the population condition number, while those under the anti-ridge method are about 5 times of the corresponding population value.

Corresponding to $\mathbf{S}$, $\mathbf{S}_a$ and $\mathbf{S}_c$, the average condition numbers of the information matrices across the converged replications for each condition are reported in Table 8.

Table 8 Average condition number of the information matrix across the converged replications; (D1) normally distributed population; (D2) elliptically distributed population; (D3) distribution with skewed factors and symmetrically distributed errors; (D4) distribution with skewed errors and symmetrically distributed factors

		P1 ($\lambda_j = 1, \psi_{jj} = 1$)			P2 ($\lambda_j = 2, \psi_{jj} = 1$)			P3 ($\lambda_j = 1, \psi_{jj} = 2$)		
	N	$\mathbf{S}$	$\mathbf{S}_{a=1}$	$\mathbf{S}_{c^2=1}$	$\mathbf{S}$	$\mathbf{S}_{a=1}$	$\mathbf{S}_{c^2=1}$	$\mathbf{S}$	$\mathbf{S}_{a=2}$	$\mathbf{S}_{c^2=1}$
D1	30	22.655	26.834	21.730	66.606	60.874	84.535	18858.867	492.230	48.087
	50	13.262	20.926	18.201	49.963	47.206	77.991	34.873	52.460	37.347
	100	9.391	16.981	16.854	41.156	38.305	74.434	21.952	36.778	29.779
	200	7.931	15.049	16.299	37.244	34.168	72.589	17.595	31.479	25.921
D2	30	298.909	1132.487	47.973	208.928	178.923	126.882	23236427.300	210643.772	130.480
	50	34.959	34.622	28.133	85.404	81.493	93.303	2748.604	986.102	63.177
	100	17.047	23.368	20.513	57.299	54.736	80.882	59.743	60.032	44.091
	200	10.571	18.049	17.399	43.386	41.055	75.363	25.114	41.453	32.691
D3	30	240945.570	104.712	37.036	173.907	123.476	108.591	8007681.450	29129034.400	90.231
	50	84.616	34.759	27.214	100.254	87.207	93.766	8741361.590	81911.339	111.607
	100	16.859	21.811	20.749	57.909	54.704	81.330	1513.458	85.893	44.446
	200	10.618	17.106	17.554	44.339	41.876	75.926	23.989	38.338	32.104
D4	30	285497.050	110.341	228.583	310.685	185.647	319.755	828416.539	2205436.440	390.626
	50	126.168	69.397	108.763	148.250	127.856	162.301	103156.938	77129.931	247.058
	100	42.786	36.965	42.627	74.393	79.098	95.740	507.287	336.608	111.481
	200	19.687	24.784	24.154	45.815	52.775	76.558	47.097	56.650	63.011
Ave		32947.380	106.762	43.361	96.596	80.587	106.247	2558806.400	1981648.400	93.884

Many numbers in the table are huge, and far above their population values reported in Table 3b. Since condition numbers corresponding to type B non-converged replications are so large that SAS cannot properly store them, including one of them in the calculation of the average can make the result larger than any of the numbers reported in Table 8. Comparing Tables 8 and 5, we may notice that most larger numbers in Table 8 correspond to conditions with many non-converged replications in Table 5, although the numbers in Table 8 are the averages of only the converged replications. Comparing Table 8 with Tables 3b and 7 suggests that condition numbers of the information matrices are affected much more by the relative errors in the sample covariances than by the population condition numbers or those of the sample covariance matrices. The rapid decline of condition number with increasing sample size in Table 8 is also due to smaller relative errors in $\mathbf{S}$.

In summary, the results in this section suggest that the size of sampling errors in $\mathbf{S}$ affects the convergence properties of FS most. The sampling errors strongly affect the condition numbers of $\mathbf{S}$, $\mathbf{S}_a$ and $\mathbf{S}_c$ as well as those of the corresponding information matrices, which are key components of the FS algorithm. For majority of the replications with $\mathbf{H}^{(t)}$ being singular, the ridge method solves the problem by fitting $\mathbf{S}_a$, which improves the condition number of $\mathbf{S}$. However, the ridge method is more effective when the non-convergence is caused by fluctuations of $\boldsymbol{\theta}^{(t)}$ from iteration to iteration rather than by a singular $\mathbf{H}^{(t)}$. By directly reducing the size of the relative errors in $\mathbf{S}$, the anti-ridge method is more effective in improving the convergence properties of the FS algorithm. However, if the non-convergence is not due to the size of relative errors in $\mathbf{S}$, the ridge method will outperform the anti-ridge method.

6 Discussion and conclusion

In the context of SEM, the FS or other algorithms are not always able to reach a set of converged solutions. Methods to improve the convergence properties of FS or its variants have been explored empirically, and one of the findings is to use more reliable indicators. However, such a finding is in the opposite direction of the ridge method, which improves the convergence properties of the FS algorithm by working with $\mathbf{S}_a = \mathbf{S}+a\mathbf{I}$. In this article we clarified the mechanisms behind the convergence properties of the two seemingly contradicting methods. Our analytical results indicate that, when the population follows a SEM or a confirmatory factor model, the size of relative errors in sample covariances increases with error variances, and decreases with the size of factor loadings or common-score variances. The improved convergence properties of FS or other algorithms following the anti-ridge method are due to smaller relative errors in $\mathbf{S}$. On the other hand, the ridge method is a post-hoc method, and convergence properties of FS improve because both relative errors in $\mathbf{S}_a$ and its condition number become smaller. For majority of the cases where the information matrices $\mathbf{H}^{(t)}$ corresponding to $\mathbf{S}$ are singular, the ridge method is able to solve the problem.

Comparing the ridge and anti-ridge methods, the latter is more effective in improving the convergence properties of FS or other algorithms. However, the scope of the applicability of the anti-ridge method is limited. In addition to models with fixed factor loadings as in Example 1, anti-ridge strategy can also be used when multiple indicators for each construct are available and we have the freedom to choose a subset of them. Then more reliable indicators will correspond to a higher likelihood of obtaining a set of converged solutions. However, if one has only a limited number of indicators for each construct or the indicators are not exchangeable, then the ridge method can be used. Even if all the indicators have fine reliabilities, the ridge method can still be used. In particular, as Kamada (2011) and Kamada and Kano (2012) showed, the ridge method can substantially improve the accuracy of parameter estimates when sample size is small even when data are normally distributed.

When the model is correctly specified or the difference between data and model is due to sampling error, the ridge method or a priori use of the anti-ridge method still yields consistent parameter estimates. When a model is incorrectly misspecified, then parameter estimates corresponding to the ridge or a priori use of the anti-ridge method may be systematically different from those obtained by modeling $\mathbf{S}$, and the differences also will be related to the values of a or c. But it is not clear which estimate will be more biased. We briefly explored the effect of a and c numerically in Sect. 4. Yuan and Chan (2008) used $a = p/N$ in their empirical study of mean square errors of ridge estimates. Kamada (2011) and Kamada and Kano (2012) obtained more refined formula of a that depends on data. However, these results are for correctly specified models and normally distributed data. More studies for the effect of a on the convergence properties of FS and other algorithms as well as on the properties of parameter estimates are needed.

The current research addressed sources of non-convergence problems in the FS algorithm for minimizing the NML-based discrepancy function (1). A modification to the weight matrix $\mathbf{W}(\boldsymbol{\theta})$ in (3) yields an iterative algorithm for minimizing the normal-distribution-based generalized least squares (GLS) discrepancy function (Browne 1974). Thus, most of our analyses and discussions also apply to computing the GLS estimator. There also exists a GLS function that does not depend on normal-distribution assumption, called asymptotically distribution free (ADF) function (Browne 1984). Huang and Bentler (2015) recently showed that the condition numbers of the ADF weight matrices in both covariance and correlation structures are strongly affected by sample size, and are closely related to the performance of test statistics in the ADF method. We expect that a ridge method applying to the weight matrix will improve the convergence properties of the corresponding FS algorithm for minimizing the ADF function. Further studies in this direction are needed.

Our study of relative errors in sample covariances might be extended to relative errors in sample correlations, which are commonly used in exploratory factor analysis. Since it has been reported that the size of factor loadings is closely related to factor pattern recovery (e.g., Velicer and Fava 1998), we suspect that the positive effect of larger loadings on factor pattern recovery occurs because the relative errors in sample correlations become smaller. Further study is needed in this direction.

The development of this article is around Fisher scoring algorithm for SEM with complete data. With incomplete data, two methods for SEM were found to perform well in practice (Savalei and Falk 2014). One is a two-stage procedure (Yuan and Bentler 2000) in which saturated means and covariance matrix are obtained via the EM-algorithm (Dempster et al. 1977) in the first stage. In the second stage, the $\mathbf{S}$ in Eq. (1) is replaced by the estimated covariance matrix, and then estimate of the structural parameter $\boldsymbol{\theta}$ is obtained by minimizing the resulting function F_{ML}. It is clear that the ridge and anti-ridge methods equally apply to the second stage of this two-stage procedure. However, they may not be applicable to the first stage when estimating the saturated means and covariance matrix by the EM-algorithm. This is because the E-step involves conditional expectation of the missing variables given the observed values, and the formulation of the conditional expectation depends on the current values of the covariance matrix. When the covariance matrix is changed as in the ridge or anti-ridge method, the conditional expectation will also be different. Then the resulting algorithm may no longer possess the properties as described in Wu (1983). Another method for SEM with incomplete data is via direct maximum likelihood, and Jamshidian and Bentler (1999) developed an EM algorithm for this approach. The E-step is performed based on the structured means and covariances, and the M-step is performed by maximizing a counterpart of Eq. (1) that also includes the mean structure. The convergence properties of this EM-algorithm might be improved by apply the ridge or anti-ridge method at the M-step. For example, one may change the $\mathbf{S}^*$ in Eq. (4) of Jamshidian and Bentler (1999) by $\mathbf{S}_a^* = \mathbf{S}^* + a\mathbf{I}$. However, for a similar reason with estimating the saturated means and covariance matrix, the ridge or anti-ridge method may not be applicable at the E-step of the EM algorithm. More studies for applying the ridge or anti-ridge idea to improve the convergence of EM algorithm for SEM with missing data is worth further studying.

The focus of the article is ridge and anti-ridge techniques in improving the convergence properties of the Fisher-scoring algorithm in SEM. However, both Fisher-scoring and EM can apply to many other models beyond SEM. In particular, the EM-algorithm is not sensitive to starting values, and the sequence of the iterated values of EM always converges to a stationary point of the likelihood function (see Wu 1983), whereas FS does not possess such properties. But in most cases, when convergence is not an issue, the speed of FS is much faster than EM (Bentler and Tanaka 1983).

7 Appendix

This appendix provides the details leading to the coefficient of variations (CV) of the sample covariances as given in (7), (8) and (9). With $y_{j0} = y_j - \mu_j$ and $y_{k0} = y_k - \mu_k$, the main work is to obtain $\gamma_{jk} = \text{Var}(y_{j0} y_{k0})$.

It follows from (5) that

$$y_{j0} y_{k0} = \lambda_j \lambda_k \xi_{j^*} \xi_{k^*} + \lambda_j \xi_{j^*} \varepsilon_k + \lambda_k \xi_{k^*} \varepsilon_j + \varepsilon_j \varepsilon_k$$

and

$$\begin{aligned} y_{j0}^2 y_{k0}^2 = {} & \lambda_j^2 \lambda_k^2 \xi_{j^*}^2 \xi_{k^*}^2 + \lambda_j^2 \xi_{j^*}^2 \varepsilon_k^2 + \lambda_k^2 \xi_{k^*}^2 \varepsilon_j^2 + \varepsilon_j^2 \varepsilon_k^2 + 2 \Big(\lambda_j^2 \lambda_k \xi_{j^*}^2 \xi_{k^*} \varepsilon_k \\ & + \lambda_j \lambda_k^2 \xi_{j^*} \xi_{k^*}^2 \varepsilon_j + 2 \lambda_j \lambda_k \xi_{j^*} \xi_{k^*} \varepsilon_j \varepsilon_k + \lambda_j \xi_{j^*} \varepsilon_j \varepsilon_k^2 + \lambda_k \xi_{k^*} \varepsilon_j^2 \varepsilon_k \Big). \quad (28) \end{aligned}$$

Taking the expected value of (28) term by term, we have:

when $j^* \neq k^*$,

$$E\left(y_{j0}^2 y_{k0}^2\right) = \lambda_j^2 \lambda_k^2 E\left(\xi_{j*}^2 \xi_{k^*}^2\right) + \lambda_j^2 \psi_{kk} + \lambda_k^2 \psi_{jj} + \psi_{jj} \psi_{kk}; \quad (29)$$

when $j^* = k^*$ but $j \neq k$,

$$E\left(y_{j0}^2 y_{k0}^2\right) = \lambda_j^2 \lambda_k^2 E\left(\xi_{j^*}^4\right) + \lambda_j^2 \psi_{kk} + \lambda_k^2 \psi_{jj} + \psi_{jj} \psi_{kk}; \quad (30)$$

and when $j = k$,

$$E\left(y_{j0}^4\right) = \lambda_j^4 E\left(\xi_{j^*}^4\right) + 6 \lambda_j^2 \psi_{jj} + E\left(\varepsilon_j^4\right). \quad (31)$$

Notice that $\sigma_{jk} = \lambda_j \lambda_k \phi_{j^* k^*}$ when $j^* \neq k^*$, it follows from (29) that

$$\gamma_{jk} = \lambda_j^2 \lambda_k^2 \left[E\left(\xi_{j*}^2 \xi_{k^*}^2\right) - \phi_{j^* k^*}^2 \right] + \lambda_j^2 \psi_{kk} + \lambda_k^2 \psi_{jj} + \psi_{jj} \psi_{kk}. \quad (32)$$

The CV in (7) is obtained using (32) and $\text{CV}_{jk} = \gamma_{jk}^{1/2} / (\sqrt{n} \sigma_{jk})$.

When $j^* = k^*$ but $j \neq k$, $\sigma_{jk} = \lambda_j \lambda_k$, it follows from (30) that

$$\gamma_{jk} = \lambda_j^2 \lambda_k^2 \left[E\left(\xi_{j^*}^4\right) - 1 \right] + \lambda_j^2 \psi_{kk} + \lambda_k^2 \psi_{jj} + \psi_{jj} \psi_{kk}. \tag{33}$$

The CV in (8) directly follows from (33).

When $j = k$, $\sigma_{jj} = \lambda_j^2 + \psi_{jj}$, it follows from (31) that

$$\gamma_{jj} = \lambda_j^4 \left[E\left(\xi_{j^*}^4\right) - 1 \right] + 4\lambda_j^2 \psi_{jj} + \left[E\left(\varepsilon_j^4\right) - \psi_{jj}^2 \right]. \tag{34}$$

With (34), the result in (9) is obtained from $\mathrm{CV}_{jj} = \gamma_{jj}^{1/2} / (\sqrt{n} \sigma_{jj})$.

References

Anderson, J. C., Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*, 155–173.

Bentler, P. M. (2008). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., Tanaka, J. S. (1983). Problems with EM algorithms for ML factor analysis. *Psychometrika*, *48*, 247–251.

Bentler, P. M., Yuan, K.-H. (1999). Structural equation modeling with small samples: Test statistics. Multivariate Behavioral Research, 34, 181–197.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *50*, 229–242.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, *8*, 1–24.

Browne, M. W. (1984). Asymptotic distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–38.

Everitt, B. S. (1987). *Introduction to optimization methods and their application in statistics*. London: Chapman & Hall.

Ferguson, T. (1996). *A course in large sample theory*. London: Chapman & Hall.

Golub, G. H., Van Loan, C. F. (1983). *Matrix computations*. Baltimore, MD: Johns Hopkins University Press.

Hu, L. T., Bentler, P. M., Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351–362.

Huang, Y., Bentler, P. M. (2015). Behavior of asymptotically distribution free test statistics in covariance versus correlation structure analysis. *Structural Equation Modeling*, *22*, 489–503.

Ichikawa, M., Konishi, S. (1995). Application of the bootstrap methods in factor analysis. *Psychometrika*, *60*, 77–93.

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling*, *8*, 205–223.

Jamshidian, M., Bentler, P. M. (1999). Using complete data routines for ML estimation of mean and covariance structures with missing data. *Journal Educational and Behavioral Statistics*, *23*, 21–41.

Jöreskog, K. G., Sörbom, D. (1981). *LISREL: Analysis of linear structural relationships by the method of maximum likelihood (version V)*. Chicago, IL: National Educational Resources Inc.

Kamada, A. (2011). *Selecting a shrinkage parameter in structural equation modeling with a near singular covariance matrix by the GIC minimization method*. Technical report 11-04, Statistical Research Group, Hiroshima University, Hiroshima. (http://www.math.sci.hiroshima-u.ac.jp/~stat/TR/TR11/TR11-04).

Kamada, A., Kano, Y. (2012). Statistical inference in structural equation modeling with a near singular covariance matrix. In *Paper presented at The 2nd institute of mathematical statistics Asia Pacific Rim meeting*. Tsukuba, Japan.

Nocedal, J., Wright, S. J. (1999). *Numerical optimization* (2nd ed.). New York: Springer.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage.

SAS Institute (1996). *SAS/STAT Software: Changes and enhancements through release 6.11*. Cary, NC: Author.

SAS Institute Inc. (2011). *SAS/STAT 9.3 user's guide*. Cary, NC: Author.

Savalei, V., Falk, C. F. (2014). Robust two-stage approach outperforms robust full information maximum likelihood with incomplete nonnormal data. *Structural Equation Modeling*, *21*, 280–302.

Schott, J. (2005). *Matrix analysis for statistics* (2nd ed.). New York: Wiley.

Velicer, W. F., Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231–251.

Wu, C. F. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, *11*, 95–103.

Yuan, K.-H., Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 167–202.

Yuan, K.-H., Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics and Data Analysis*, *52*, 4842–4858.

Yuan, K.-H., Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology*, *56*, 93–110.