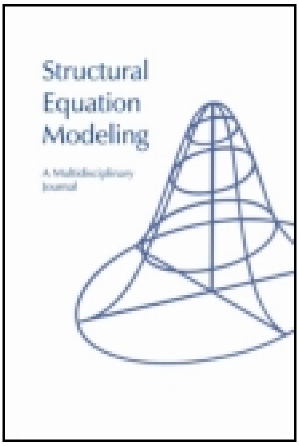

Structural Equation Modeling: A Multidisciplinary Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hsem20

Structural Equation Modeling Diagnostics Using R Package semdiag and EQS

Ke-Hai Yuan [a] & Zhiyong Zhang [a]
[a] University of Notre Dame

Version of record first published: 31 Oct 2012.

To cite this article: Ke-Hai Yuan & Zhiyong Zhang (2012): Structural Equation Modeling Diagnostics Using R Package semdiag and EQS, Structural Equation Modeling: A Multidisciplinary Journal, 19:4, 683-702

To link to this article: http://dx.doi.org/10.1080/10705511.2012.713282

Structural Equation Modeling, 19:683–702, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1070-5511 print/1532-8007 online
DOI: 10.1080/10705511.2012.713282

Psychology Press
Taylor & Francis Group

Structural Equation Modeling Diagnostics Using R Package semdiag and EQS

Ke-Hai Yuan[1] and Zhiyong Zhang[1]
[1]*University of Notre Dame*

Yuan and Hayashi (2010) introduced 2 scatter plots for model and data diagnostics in structural equation modeling (SEM). However, the generation of the plots requires in-depth understanding of their underlying technical details. This article develops and introduces an R package `semdiag` for easily drawing the 2 plots. With a model specified in EQS syntax, one only needs to supply as few as 2 parameters to generate the 2 plots using the `semdiag` package. Two examples are provided to illustrate the use of the package. Multiple figures are used to explain the elements of data and model diagnostics. Advice on selecting proper estimation methods following the diagnostics is also given.

Keywords: case-level residuals, leverage observations, outliers, robust method, `semdiag` package

Structural equation modeling (SEM) is one of the most widely used methods in social science research. Although many methods have been developed for SEM, the normal-distribution-based maximum likelihood (NML) is still routinely used regardless of the distribution of the data. When the sample is smoothly distributed without any outliers or data contamination, NML might generate reasonable but not efficient parameter estimates. When data are contaminated or contain outliers, NML estimates (NMLEs) are generally biased. Outliers can also arbitrarily change the values of standard errors and test statistics for overall model evaluation following NML (Yuan & Bentler, 1998). In practice, test statistics for evaluating the overall fit of most interesting SEM models are observed to be significant. The quality of the samples is at least partially responsible for the significant differences between data and model. A comprehensive study by Micceri (1989) indicates that no real data are normally distributed. Some of the nonnormality might be due to the underlying nature of the population. Others might be due to outliers or data contamination. In either case, it would be informative to see whether there exist a few cases that are mainly responsible for the typical lack of fit between data and model. In particular, outlying cases might contain important information on the participants and the population. Identifying those cases that significantly deviate from a theoretical model

Correspondence should be addressed to Ke-Hai Yuan, Department of Psychology, University of Notre Dame, Notre Dame, IN 46556. E-mail: kyuan@nd.edu

will facilitate treatment and intervention, correction of misconduct in data collection or coding, modification of the existing SEM model, or even reformulating the substantive theory.

A key element in data and model diagnostics is the (Mahalanobis) M-distance, which measures the distance of a random vector from its expected value and weighted by the precision or the inverse of the covariance matrix of the random vector. Using M-distances, Yuan and Hayashi (2010) proposed two plots for data and model diagnostics. The first plot contrasts the M-distances of factor scores in the horizontal direction against the M-distances of case-level residuals in the vertical direction. Cases with large residuals appear on the top of the plot and those with large factor scores are on the right. In particular, two robust methods are used when calculating these M-distances so that the results are not distorted by potentially biased NMLEs. We briefly review the two robust methods in the next section. The second plot is a quantile–quantile (QQ) plot in which the ordered M-distances of the residuals are on the vertical direction and the quantiles of a χ distribution are on the horizontal direction. Yuan and Hayashi (2010) proposed to visually identify clusters of observations on the right end of the second plot and treat each cluster of observations as a group when judging the outlier status of the observations. They also provided two SAS and two parallel R programs. One program is for confirmatory factor models with saturated factor covariances but without any correlated errors. The other is for a specific SEM model with 10 manifest and 5 latent variables. Each program generates the two M-distances for the plots. However, it is hard for applied researchers or even psychometricians to use these programs when the model or the number of variables change. The two robust methods used in Yuan and Hayashi (2010) have not been implemented in any SEM software. One of the contributions of this article is to develop an SEM diagnostics (`semdiag`) package[1] so that the two plots can be easily drawn for commonly used SEM models. In addition to the two plots, the `semdiag` package also yields the robust statistics reported in Yuan and Hayashi (2010). In particular, the package obtains robust means and covariances and feeds them into EQS[2] (Bentler, 2008) for structural parameter estimates iteratively. At the end of the iteration, parameter estimates, their standard errors, standardized solutions, and multiple fit indexes following from the robust procedures are available in the associated EQS output. Another novelty of this article is a clear step-by-step illustration of the two plots through two simulated data sets where the status of each case is known, which allows readers to fully appreciate the functionality of the two plots. We also relate the QQ plot to the scree plot in exploratory factor analysis so that the idea of examining the plot for outlier identification is easier to follow by applied researchers. Furthermore, the `semdiag` package also provides profile plots of selected cases that facilitate substantive analysis of the outliers, leverage observations, or both.

To make the article relatively self-contained, the following section includes an introduction to concepts and ideas used for model diagnostics as well as the two robust methods. A novel feature of the introduction is that we use two path diagrams to facilitate the understanding of the concepts defined in Yuan and Hayashi (2010) and Yuan and Zhong (2008). Then we illustrate the use of the `semdiag` package through an SEM model with simulated data sets. We conclude the article with some remarks on the use of the `semdiag` package in practice.

[1] The `semdiag` utilizes the `REQS` functions developed by Mair, Wu, and Bentler (2010) for running EQS (Bentler, 2008) within R. We thank Drs. Patrick Mair, Eric Wu, and Peter Bentler for allowing us to adapt the REQS functions as a part of the `semdiag` package.

[2] The `semdiag` package also works with other SEM software and more detail is given in the concluding section.

ELEMENTS FOR SEM DIAGNOSTICS AND ROBUST METHODS

Elements for SEM Diagnostics

Model and data diagnostics in regression have been well developed, where cases with large values in predictors are called leverage observations and those with large error terms are called outliers. These two concepts are best understood through Figure 1, where four regression lines are obtained when fitting different subsets of a sample with 13 observations. Point A is an outlier but not a leverage observation, point B is an outlier and also a leverage observation, and point C is a leverage observation but not an outlier. Point B is also called a bad leverage observation and point C is also called a good leverage observation. Clearly, outliers affect the intercept and slope of the regression line and consequently the distance between each observed case and the regression line. Good leverage observations have little influence on the regression line determined by the majority of the sample, but they generally lead to more accurate regression parameter estimates or estimates with smaller standard errors. Readers interested in detailed discussion of regression diagnostics are referred to Belsley, Kuh, and Welsch (1980) and Cook and Weisberg (1982).

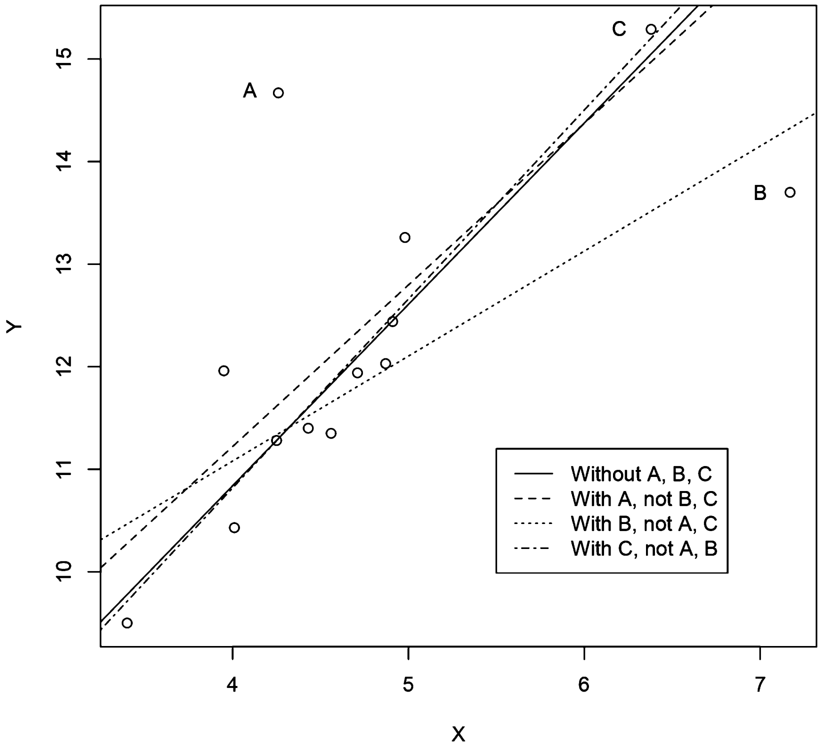

FIGURE 1 Simple regression with outliers and leverage observations.

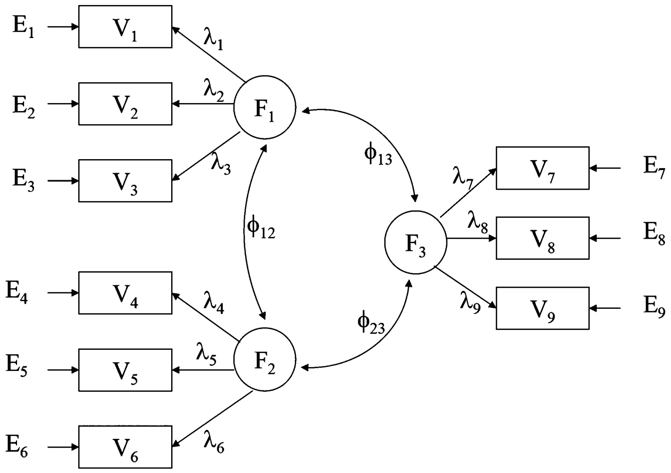

FIGURE 2 Confirmatory factor model: Cases with extreme Fs are leverage observations and those with extreme Es are outliers.

Because the factor analysis model can be regarded as a multivariate regression model with latent predictors, Yuan and Zhong (2008) defined leverage observations in factor analysis as cases with large values in latent factors. Outliers are observations with large measurement errors.[3] Figure 2 contains the path diagram for a confirmatory factor model with nine manifest and three latent variables, and we could imagine that there is such a path diagram for each case in a sample. If any of the Es is large, then the corresponding case is an outlier. If any of the Fs is large, then the corresponding case is a leverage observation. Good leverage observations are those with large Fs but ordinary Es, and bad leverage observations have large values in both Fs and Es. Like in regression illustrated in Figure 1, both leverage observations and outliers in factor analysis are influential in the sense that their exclusion or inclusion greatly affects either the parameter estimates or the assessment of the overall model. In particular, outliers lead to biased parameter estimates and an inflated likelihood ratio (LR) statistic[4] following NML, and leverage observations lead to greater values of variances and covariances of the Fs in Figure 2 and they do not affect the estimates of the other parameters nor the LR statistic (see Yuan & Zhong, 2008; Zhong & Yuan, 2011).

These concepts can also be generalized to SEM models, as in Figure 3 where the Es are measurement errors and the Ds are prediction errors. Although there are five variables describing the relationships among the latent variables[5] in Figure 3, only F_1, D_2, and D_3

[3] More formal definitions of outliers using population distributions with saturated and structured models are given in Yuan and Bentler (2001) and Yuan and Hayashi (2010).

[4] The likelihood ratio statistic is commonly referred to as the chi-square statistic although its true distribution is seldom chi-square in practice.

[5] Parallel to the regression literature, we do not call the Es latent variables in this article although they are not observable in practice.

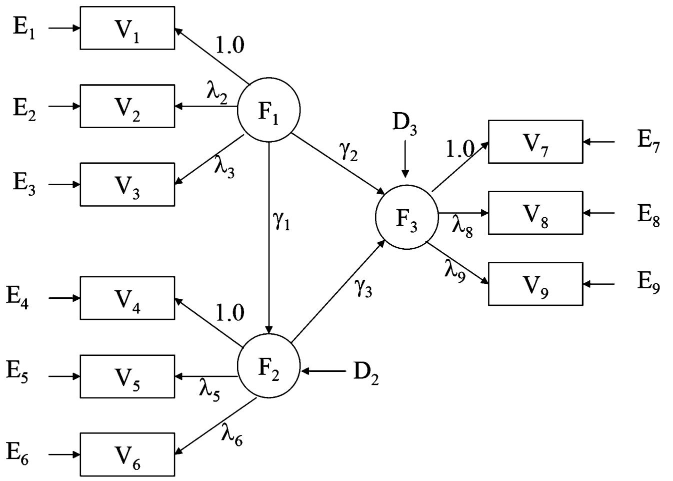

FIGURE 3 Structural equation model: Cases with extreme F_1s are leverage observations and those with extreme D_2s and D_3s can be either leverage observations or outliers; cases with extreme Es are outliers.

independently vary; F_2 and F_3 are dependent variables with

$$F_2 = \gamma_1 F_1 + D_2, \quad \text{and} \quad F_3 = \gamma_2 F_1 + \gamma_3 F_2 + D_3 = (\gamma_2 + \gamma_3\gamma_1)F_1 + \gamma_3 D_2 + D_3. \tag{1}$$

Clearly, D_2 and D_3 in Figure 3 jointly play the function of F_2 and F_3 in Figure 2. Actually, we can rewrite the structural equation model in Figure 3 as a factor model in Figure 2, where the factor variances–covariances are determined by the two equations in Equation 1. Cases with large values of D_2 and D_3 can be regarded as leverage observations. Alternatively, we can also regard cases with large values of D_2 or D_3 as outliers. Then, we need to combine the terms containing D_2 and D_3 in Equation 1 with the Es in Figure 3 so that only cases with large F_1s are leverage observations. The covariance matrix of the combined errors (Ds and Es) is not diagonal anymore even if the Es are not correlated in Figure 3.

Notice that the relationship among the three Fs in Figure 3 is saturated. The estimated covariance matrix of the three Fs can always recover any change in the sample covariance matrix of the observed variables caused by the values of extreme Ds. Thus, cases with extreme Ds only affect parameter estimates describing the relationship of the Fs and standard errors of all model parameter estimates, and they do not interfere with evaluation of the overall model fit using NML (see Yuan & Zhong, 2008). When the relationship among latent variables is not saturated, outliers in Ds will lead to biased NMLEs for all the model parameters as well as an inflated LR statistic.

Our semdiag package contains the option for users to choose between setting the Ds as factors or errors. In practice, one might need to consult the literature regarding the theory around the substantive model. If the substantive model is strongly supported by theory, then any deviance from the structural model can be regarded as an error. In such a case, observations with large prediction errors should be regarded as outliers with Ds being treated equivalent to Es. If the theory around the model is not well-established or the study is still ongoing, then large prediction errors are very likely caused by weak or improper predictors, indicators of important predictors being excluded from the model, or a misspecified model structure. In such a case, we can regard the Ds as equivalent to Fs. Because the two classifications of the Ds will very likely lead to different results in practice, we recommend letting the program run for both options. If the approach of classifying the Ds as errors yields better overall fit between the data and model, then we will have the additional information that the lack of fit in the other approach is due to cases with large prediction errors. We can also distinguish cases with large prediction errors from those with large measurement errors by visually examining the four plots following from running the two options. This is illustrated using examples in the next section when introducing the `semdiag` package.

Two Robust Methods and the M-Distances d_f and d_r

Both outliers and leverage observations introduced in the previous section are model based. What we have in practice is a sample of the observable Vs, not Fs or Es. For a given model, because the number of Fs and Es is typically greater than the number of Vs, we still do not know the values of Fs or Es even when the values of the model parameters are known. Of course, the values of model parameters are never known in practice. What we can do is to estimate the model parameters and use the estimated values to further evaluate leverage observations and outliers. The same practice is used in regression as illustrated in Figure 1, where slopes and intercepts have to be estimated to draw the regression lines. Clearly, the values of slopes and intercepts have direct consequences on the evaluation of residuals. For example, point C is much further away from the dotted line than point B and is more likely to be judged as an outlier if the evaluation is based on the dotted line. Similarly, evaluation of outliers and leverage observations in factor analysis and SEM also needs reliable parameter estimates. NML-based analysis that treats all the observations equally results in biased parameter estimates when outliers exist (Yuan & Bentler, 1998, 2001). Robust methods are needed so that the resulting parameter estimates are determined by the majority of the observations rather than a few influential cases.

The two-stage robust method. Because an extreme F, D, or E in either Figure 2 or 3 will be reflected in the manifest Vs, the effect of leverage observations and outliers can be minimized by a method in which cases with extreme Vs are assigned a tiny weight. Such a method was developed in Yuan and Bentler (1998), where the saturated means and covariance matrix of the manifest variables are robustly estimated first and then fitted by the structural model, called a two-stage robust (TSR) method. Let $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_n$ be a p-dimensional sample with $E(\mathbf{v}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{v}_i) = \boldsymbol{\Sigma}$, $i = 1, 2, \ldots, n$. At the first stage one starts by selecting a set of initial values for $\boldsymbol{\mu}^{(j)}$ and $\boldsymbol{\Sigma}^{(j)}$ with $j = 0$, and calculates an M-distance of each case by

$$d_i^{(j)} = [(\mathbf{v}_i - \boldsymbol{\mu}^{(j)})'(\boldsymbol{\Sigma}^{(j)})^{-1}(\mathbf{v}_i - \boldsymbol{\mu}^{(j)})]^{1/2}. \tag{2}$$

Then new $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are obtained using

$$\boldsymbol{\mu}^{(j+1)} = \frac{1}{\sum_{i=1}^{n} w_1(d_i^{(j)})} \sum_{i=1}^{n} w_1(d_i^{(j)})\mathbf{v}_i, \quad \boldsymbol{\Sigma}^{(j+1)} = \frac{1}{n} \sum_{i=1}^{n} w_2(d_i^{(j)})(\mathbf{v}_i - \boldsymbol{\mu}^{(j)})(\mathbf{v}_i - \boldsymbol{\mu}^{(j)})', \quad (3)$$

where $w_1(d_i)$ and $w_2(d_i)$ are chosen as the popular Huber-type weights (see Yuan & Hayashi, 2010). Clearly, the two formulas in Equation 3 are weighted averages. The weights $w_1(d_i) = 1$ and $w_2(d_i)$ are a constant when d_i is smaller than a certain threshold that is determined by a tuning parameter $0 < \varphi < 1$; and they start to decrease with d_i when d_i is greater than the threshold. Thus, cases with d_i greater than the threshold get smaller weights in Equation 3 and are called downweighted. The tuning parameter roughly controls the percentage of cases being downweighted in Equation 3. A greater φ implies more cases being downweighted. Results and analyses in Yuan and Hayashi (2003) and Yuan and Zhong (2008) indicate that a φ between 0.05 and 0.25 can effectively control outlying cases in practice. The default value of φ in our package is set at 10%, and it can be changed by users as described in Appendix C. When evaluating Equation 3 with $j = 0, 1, 2, \ldots$, the values on the left sides of the equations are the needed input values on the right sides. Continuously updating the $\boldsymbol{\mu}^{(j)}$ and $\boldsymbol{\Sigma}^{(j)}$ until convergence leads to a set of robust estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. Let $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ represent the structural model with $\boldsymbol{\theta}$ being the vector of unknown parameters. At the second stage, the structural model is estimated by minimizing the normal-distribution-based discrepancy function between $\hat{\boldsymbol{\Sigma}}$ and the structural model $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, resulting in a vector of robust estimates $\hat{\boldsymbol{\theta}}$. Because $w_1(d_i)$ and $w_2(d_i)$ corresponding to cases with large d_i are small, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\theta}}$ from the TSR method are barely affected by either outliers or leverage observations.

The M-distance d_f. Let $\hat{\mathbf{f}}_i$ be the Bartlett factor score estimator for the ith case and $\boldsymbol{\Sigma}_f$ be its covariance matrix. With $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\theta}}$, we can evaluate $\boldsymbol{\Sigma}_f$ and $\hat{\mathbf{f}}_i$ for each case. The M-distances for measuring leverage observations are obtained when replacing the $(\mathbf{v}_i - \boldsymbol{\mu})$ by $\hat{\mathbf{f}}_i$ and $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}}_f$ in Equation 2. Let the resulting distance be denoted as d_{fi}. When F_1, the Ds and Es in Figure 3 are normally distributed and the population values of the parameters are known, d_{fi} follows the chi distribution χ_q, where q is the number of variables in the factor score estimator. In Figure 2, $q = 3$ and in Figure 3 $q = 3$ or 1 depending on whether we treat D_2 and D_3 as factors or errors. Thus, we can refer d_{fi} to χ_q when judging whether an observation has extreme Fs or Ds.

The direct robust method and the M-distance d_r. The M-distances for measuring outliers are obtained following a direct robust (DR) method. In this method, a set of initial values of $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ is chosen first. Next, a vector of case-level residuals when $\mathbf{v}_i$ is predicted by the Bartlett factor scores $\hat{\mathbf{f}}_i$ is obtained, and so is the M-distance d_{ri} corresponding to the residuals. Then a new set of $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ is obtained using weighted averages where cases with larger d_{ri} are assigned a smaller weight according to the Huber-type weight functions, parallel to Equation 3. The obtained values of $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ are further used to obtain the next set of $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$, and continuing the iteration until convergence leads to the direct robust estimates $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\theta}}$

as well as d_{ri} for the ith case. When data are normally distributed and the population values of the parameters are known, $d_r \sim \chi_{p-q}$, where p is the number of manifest variables. Thus, we can compare d_r to χ_{p-q} to facilitate the determination of outliers. Readers interested in the detail of the DR method are referred to Yuan, Fung, and Reise (2004) and Yuan and Zhong (2008).

In summary, because both d_f and d_r involve unknown model parameters, we choose to use two robust methods in our `semdiag` package so that both the effects of outliers and leverage observations are minimized when evaluating the two M-distances. The choice is closely related to the formulations of d_f and d_r. All model parameters are involved in formulating d_f, and TSR minimizes the effect of both leverage observations and outliers. The formulation of d_r does not involve the variances–covariances of the latent factors (see Yuan & Zhong, 2008). Good leverage observations mainly affect the values of the variances–covariances of the latent factors and are not downweighted by the DR method to achieve the highest efficiency of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\theta}}$. One might wonder whether substituting the latent variables by their Bartlett estimates leads to biased analysis. Yuan and Hayashi (2010) showed that, among estimators of factor scores, the Bartlett estimator is the unique formulation such that the vector of resulting residuals is orthogonal to the space of the true scores, which implies that substituting Bartlett estimates for the latent predictors does not lead to biased analysis. Even in the Bartlett formulation of factor score, parameters still need to be estimated, and thus, model diagnostics are subject to sampling errors, which are much smaller when compared to the effect of outliers or leverage observations (see Yuan & Zhong, 2008).

In obtaining the two M-distances for the plots, we also obtain parameter estimates $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$, which are typically less biased and more efficient than the NMLEs for practical data (Zhong & Yuan, 2011). Similarly, test statistics for overall model evaluation following TSR and DR are also more reliable (Yuan & Zhong, 2008). The output of running our package also contains NMLEs, $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$, and the associated test statistics for overall model evaluation.

R PACKAGE `semdiag` FOR SEM DIAGNOSTICS

In this section, we illustrate the use of the `semdiag` package using the SEM model in Figure 3 through two sets of simulated data. Let all the λs and the variances of the Es be 1.0; $\gamma_1 = .70$, $\gamma_2 = .50$, and $\gamma_3 = .4$; the variances of D_2 and D_3 are .51 and .31, respectively, so that the variances of F_1, F_2, and F_3 are all at 1.0; and all the means of the nine Vs are at 1.0. A sample of size 100 from the normally distributed population with the preceding parameters is generated and can be downloaded at http://rpackages.psychstat.org/data/semdiag/N100.dat.[6] The code for estimating this model using EQS is given in Appendix A, where each $*$ represents a free parameter. The EQS file is named semplot.eqs. The line numbers in the right margin of the appendix are for the convenience of explaining the codes, not part of EQS syntax. The line numbers should be removed when running the program. An electronic version of Appendix A without the line numbers can be obtained at http://rpackages.psychstat.org/examples/semdiag/

[6] The data set is also built into the R packages. To access the data, use `data(N100)` within R after loading the package.

semplot.eqs. Because EQS does not have the capability of talking with R before version 6.1, one needs to have EQS 6.1 for Windows (build 97) and onward for properly applying the package. To use EQS within R, EQS has to save certain output in files as described in Lines 33 to 38 of Appendix A. Note that the file name semplot.ETS should be the same as the EQS script file except for the extension name. Readers are referred to Bentler (2008) for detailed instruction on specifying different models within EQS.

To use the `semdiag` package for the first time, install it by issuing the following command:

```
install.packages("semdiag", repos="http://rpackages.psychstat.org")
```

In the command `repos="http://rpackages.psychstat.org"` is the server address where the package will be downloaded. To select a different server through a graphical interface, one can use the command `install.packages("semdiag")`.

The R codes on Lines 1 to 14 in Appendix B illustrate a typical routine for using our R package for SEM diagnostics. Again, the line numbers in the right margin are not part of the R codes, and an electronic version of Appendix B without the line numbers is available at http://rpackages.psychstat.org/examples/semdiag/Examples.R. The # sign indicates comments in R. Specifically, `library(semdiag)` loads our R package. The code `setwd("C:/research/SEMDiagnostics")` sets the working directory to the folder where the data file and the EQS model file reside. Line 4 uses the R function `read.table` to read raw data in the file `N100.dat` into R. Line 8 uses the function `semdiag` from the package `semdiag` to calculate the M-distances d_r and d_f as well as parameter estimates and statistics for measuring model fit. The first argument `N100` specifies the name of the data. The second argument `"semplot.eqs"` is the name of the EQS input file.[7] The third argument tells that the Ds in Figure 3 are treated the same as F_1. If it is reasonable to treat Ds as error Es, one can remove `D="F"` as on Line 17 or change the option to `D="E"`.

Running the codes on Line 11 generates the two plots in Figure 4. The scatter plot in Figure 4a is divided into four areas by a dashed vertical line and a dashed horizontal line. The vertical line is drawn at x =.01-level critical value of χ_q, and the horizontal line is drawn at y =.01-level critical value of χ_{p-q}. Data points on the right side of the vertical line are leverage observations, those above the horizontal line are potential outliers. As expected, no observations belong to these two categories in Figure 4a because the data are simulated from a normal population. The two cases with most outstanding d_fs are 63 and 9. A few observations have d_rs close to the horizontal line with case 51 on top.

Figure 4b contrasts the ordered d_rs in the vertical axis against the quantiles of χ_{p-q} in the horizontal axis. Because d_r approximately follows χ_{p-q}, without data contamination we would expect that the plot can be roughly described by a line with slope 1.0 (the $x = y$ line). A sudden upperward increase on the right tail indicates at least heavy-tail-distributed errors. The corresponding cases are suspects of outliers. This is very much like the scree plot in exploratory factor analysis, where points beyond a linear trend represent a systematic change in the eigenvalues and the eigenvalues in the linear part are mostly due to unique variances

[7] One can use any name for the EQS input file. However, the file name on Line 35 of Appendix A should be the same as the input file name except for different extensions.

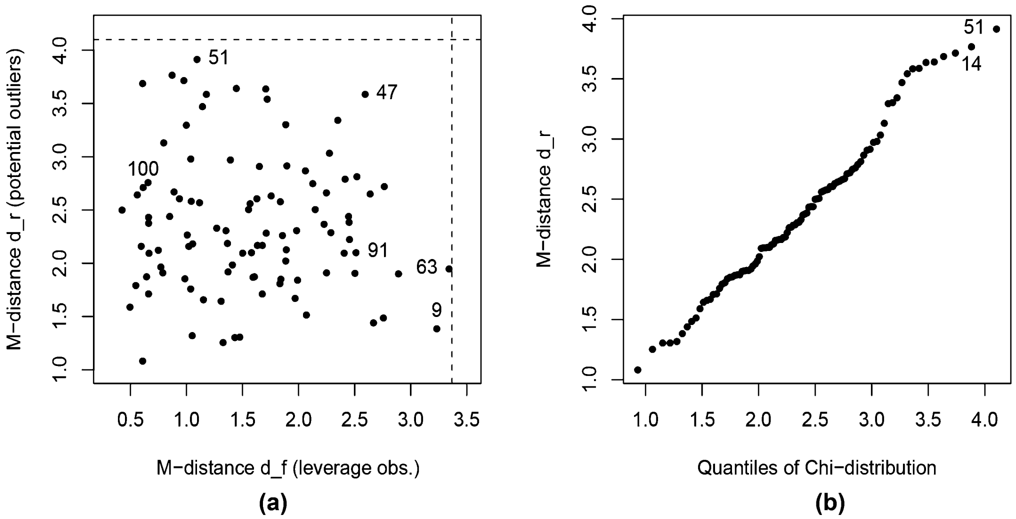

FIGURE 4 Structural equation model with normally distributed data; cases with extreme D_2 and D_3 are treated as leverage observations.

with sampling errors. Actually, the rationale for outliers to cause a significant departure from a linear trend on the right tail of a QQ plot is well-established (Gnanadesikan, 1997), whereas for the scree plot it is mostly empirical (Gorsuch, 1983). Because the errors are ordered, it is impossible to see a down-turn on the right tail in the plot of d_r against the quantile of χ_{p-q}. However, it is possible to have a relative flat tail on the right, which indicates that the errors follow a distribution with lighter tail than that of a normal distribution.

Case 51 in Figure 4b has the largest residual, but it is relatively close to Case 14 in the vertical direction. Actually, the d_{ri} for these observations is slightly smaller than the corresponding values of χ_{p-q}, implying that no observations are outliers.

Running the R codes on Lines 17 and 20 in Appendix B yields the two plots in Figure 5, where the Ds in Figure 3 are treated the same as the Es. Figure 5a contains different information from that of Figure 4a. For example, Case 91 has the largest d_f in Figure 5a while case 63 is on the right in Figure 4a. Case 47 has the largest d_r in Figure 5a and it also has a relatively large d_f in Figure 4a, indicating that it might have large values in D_2 or D_3. When D_2 and D_3 are treated as errors, Case 47 beats Case 51 in the value of d_r. Figure 5a suggests that the value of F_1 for Case 47 is not outstanding at all. Figure 5b parallels Figure 4b when the Ds in Figure 3 are treated the same as the Es. Although Case 47 is relatively outstanding in d_r, it is expected because the corresponding quantile of χ_{p-q} is also proportionally outstanding, and the point is well predicted by the $x = y$ line.

After seeing these plots with normally distributed data, we continue to illustrate their shapes when data are contaminated. A data set, called N85.dat (http://rpackages.psychstat.org/data/semdiag/N85.dat), is created for such a purpose. The first 85 cases in this data set are the same as in N100.dat. Cases 86 to 100 are created by adding a number 2_s to the Fs, Ds, or Es as described later, where $2_s = 2$ if the corresponding F, D, or E is positive and $2_s = -2$

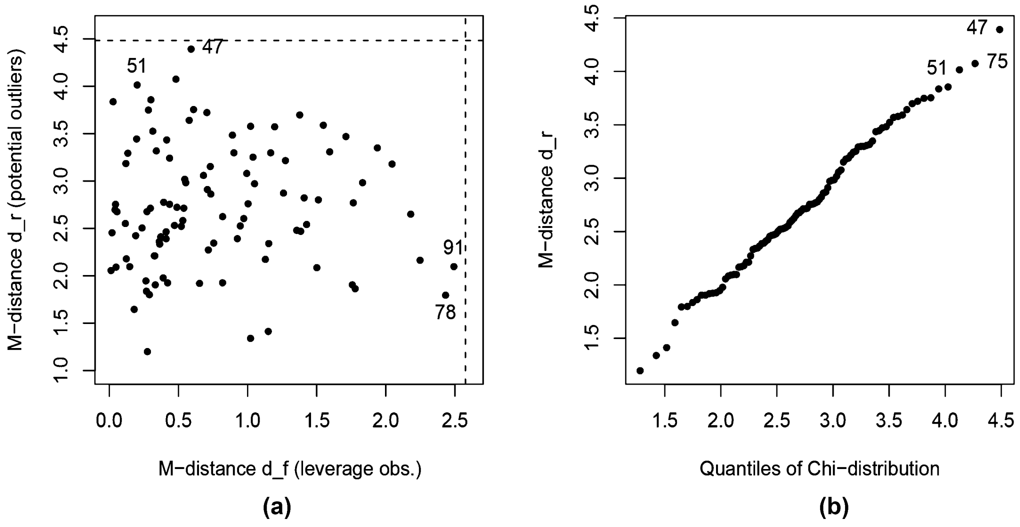

FIGURE 5 Structural equation model with normally distributed data; cases with extreme D_2 and D_3 are treated as outliers.

otherwise. Cases 86 to 90 in N85.dat are created by adding a number 2_s to each of F_1, D_2, and D_3, whereas the Es are the same as those in N100.dat; Cases 91 to 95 in N85.dat are created by adding the number 2_s to each of E_1, E_2, ..., E_9, whereas F_1, D_2, and D_3 are the same as those in N100.dat; Cases 96 to 100 in N85.dat are created by adding the number 2_s to each of F_1, D_2, D_3, E_1, E_2, ..., E_9. According to the definitions in the previous section, Cases 86 to 90 are good leverage observations when D_2 and D_3 are treated the same as F_1, Cases 91 to 95 are outliers, and Cases 96 to 100 are bad leverage observations and also outliers.

Treating D_2 and D_3 the same as F_1 in Figure 3 and applying the semdiag package to N85.dat (Lines 30 and 33 in Appendix B) yields the two plots in Figure 6. All five good leverage observations are correctly identified in Figure 6a, Cases 92 to 95 are correctly identified as outliers, and Cases 96 to 99 are also correctly identified as leverage observations and outliers. But Case 91 is also identified as having a large d_f, whereas Case 100 is identified as having a smaller d_f. This is because Case 91 has a relatively large d_f in the sample N100.dat (see Figure 4a), and the added number 2_s in each of the nine Es also makes the nine manifest variables greater proportionally to the factor loadings. Any change in the manifest variables proportional to the loadings will change the values of the estimated F_1, D_2, and D_3. Similarly, Case 100 has a small d_f in N100.dat as reflected in Figure 4a. Adding the number 2_s to each of F_1, D_2, and D_3 makes the corresponding d_f greater than the majority of the cases, but not reaching the .01-level critical value of χ_3. The right tail of Figure 6b definitely cannot be predicted by a linear trend anymore. In particular, the five leverage observations are not on the right tail. In this example, we know Cases 91 to 100 are outliers. Otherwise, we would classify the 10 cases into four clusters: {96}, {92, 94, 100, 97, 93}, {95, 91, 98}, {99}. Yuan and Hayashi (2010) proposed to remove each cluster of observations and study the change in the resulting LR

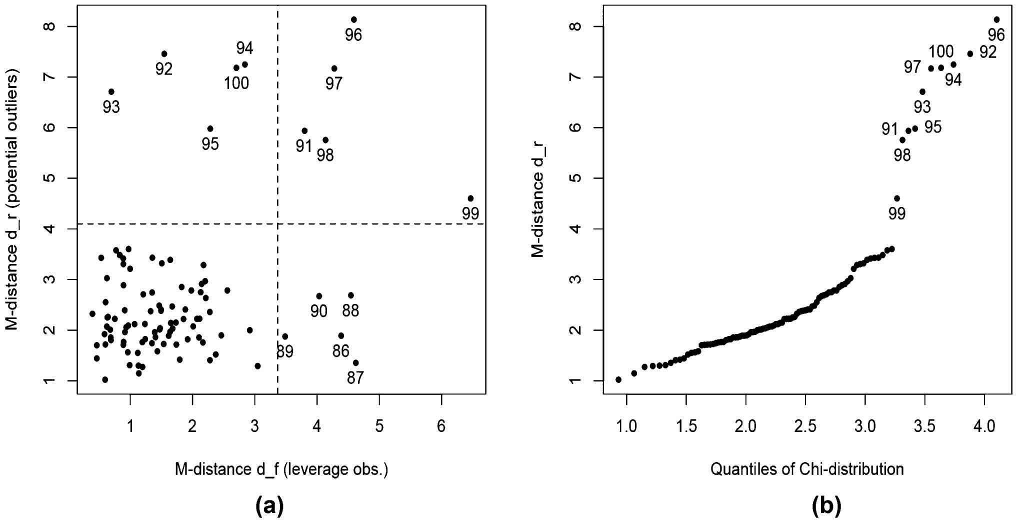

FIGURE 6 Structural equation model with contaminated data; cases with extreme D_2 and D_3 are treated as leverage observations.

statistic in sequence.[8] Clusters corresponding to a substantial drop of the LR statistic should be treated as outliers, and observations within a cluster should not be discriminated when determining their outlier status even if part of them are greater than a critical value. Removing cases in the R package is performed by including an argument `delete=c(n1,n2,...,nm)` in the `semdiag` function, where `n1,n2,...,nm` are case numbers as sequentially appeared in the data matrix, not participants' ID numbers when collecting data. For example, Cases 99 and 100 are not included in the analysis when running Line 51 of Appendix B.

Figure 7 contains the two plots for the analysis of N85.dat and treating D_2 and D_3 the same as the Es in Figure 3. The 15 artificially created cases are all outstanding in Figure 7a. However, their positions are quite different from those in Figure 6a. This is because cases with positively large Ds can have negatively large Es or vice versa. Their combination does not have the same character as their individual scores. Among the 15, the least outstanding case in Figure 7a is Case 89, whose F_1 was added a number 2_s. Comparing Figures 7a and 6a, we will notice that the d_r of Case 89 changes from smaller than those of most cases in Figure 6a to greater than those of most cases in Figure 7a. The 10 artificially created outliers all appear on the right tail of Figure 7b, which cannot be explained by a linear trend. In particular, Cases 88, 86, 89, and 90 also appear on the right tail below the 10 outliers in Figure 7b. In contrast, these cases are not outstanding at all in Figure 6b.

[8] The idea proposed in Yuan and Hayashi (2010) might seem like the idea of jackknifing. However, there exist basic differences between the two approaches. Jackknifing is to estimate the bias and standard error of a statistic by systematically recomputing the statistic estimate leaving out one or more observations at a time from the sample set. The proposal in Yuan and Hayashi is to study the change on the LR statistic on a few selected clusters of observations, not to calculate its bias or standard error. Also, the number of observations in each cluster is determined by data rather than predetermined as in jackknifing.

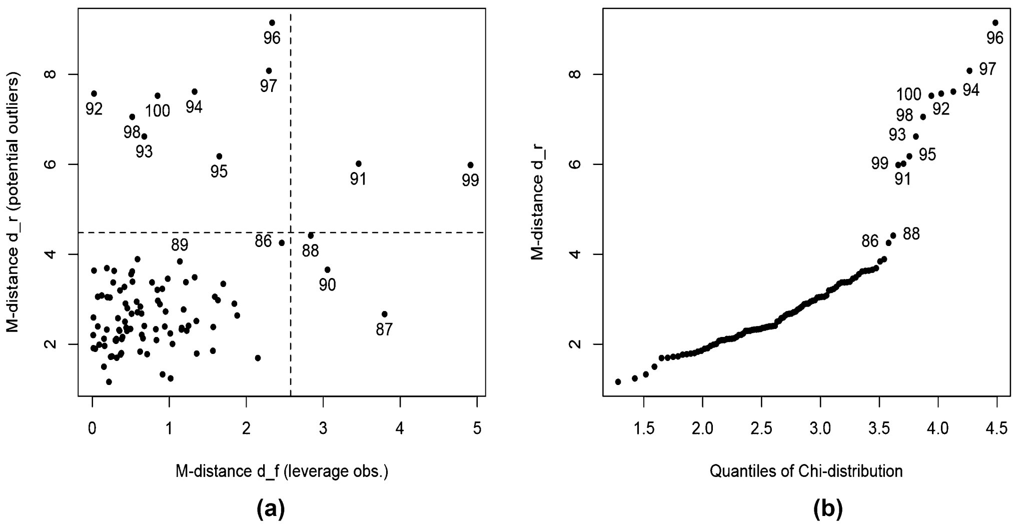

FIGURE 7 Structural equation model with contaminated data; cases with extreme D_2 and D_3 are treated as outliers.

Additional information on model fit statistics, parameter estimates, and their standard errors following NML, TSR and DR can be obtained using the function `semdiag.summary`. An example output is given in Appendix B from Line 55 to Line 91. The working directory also contains three standard EQS output files following running the `semdiag` package. The first one (`nml.out`) contains the results of NML-based analysis; the second one (`tsr.out`) contains the results of TSR when both outliers and leverage observations are downweighted; and the third one (`dr.out`) contains the results of DR when only outliers are downweighted. The R functions `semdiag`, `semdiag.plot`, and `semdiag.summary` can be tuned for customized analysis. A detailed account of the three functions is given in Appendix C.

Figures 4 to 7 have illustrated the functions of the R package for data and model diagnostics. When outliers are confirmed after examining the plots and studying their effect on the NML-based LR statistic, we recommend further examination of the profile of each identified case. For researchers who are interested in intervention or prevention, each identified case needs to be substantively analyzed to develop proper treatment procedures. For researchers who are only interested in fitting the model to data, examining the profiles of the outliers also leads to a better understanding of the model and the population. To facilitate profile analysis, a case-profile-plot (CPP) function `semdiag.cpp` in our R package plots the values of the variables in the vertical direction against the order of the variables in the horizontal direction for selected cases, where each variable is centered at its estimated mean following the two-stage robust method. For example, executing Line 48 of Appendix B yields the CPPs in Figure 8 for Cases 86, 90, 98, 99, and 100 of data set N85. Because a crowded figure does not facilitate visual examination, we recommend not including more than five cases in running the CPP function.

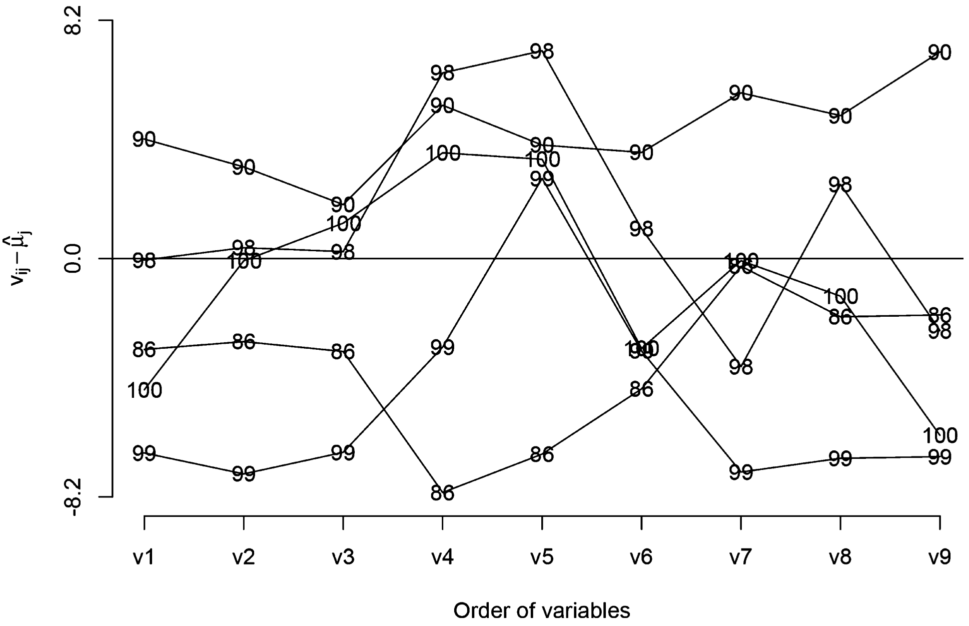

FIGURE 8 Profile plots for Cases 86, 90, 98, 99, and 100 of data set N85.

CONCLUSION AND RECOMMENDATION

In practice, one has a sample and tries to fit the sample by a substantively interesting model. As pointed out in the introduction, most substantive models in the SEM literature have statistically significant differences from their samples. If all the observations in a sample are well collected and correctly coded, one might need to modify the model to minimize its difference from the sample. However, if certain observations are contaminated or the sample contains outliers, just modifying the model might not be fruitful. Our `semdiag` package for data and model diagnostics provides case-wise information on the lack of fit between data and model. In particular, the graphic output of the package facilitates visual examination of outliers and leverage observations. If the outlying cases are not totally unreasonable after the examination, they might be due to heavy tails of the underlying population. Then parameter estimates following from the robust methods are preferred. If the profiles of the outlying cases are totally not expected from the population, then we should delete them and report the results from the subsequent NML or either of the robust analyses. Without outliers, all three methods yield reliable results while those following a robust procedure are more efficient for typical practical data (Zhong & Yuan, 2011). Without leverage observations, the two robust methods should lead to comparable results. Otherwise, the DR method will lead to more efficient parameter estimates because only cases that do not fit the model are downweighted.

We would like to note that, with typical unknown population distributions, the statistics following from any of the three estimation methods, as appear on Line 61 of Appendix B, do not exactly follow a chi-square distribution. However, unless data are normally distributed, parameter estimates, their standard errors, statistics, and fit indexes following from the robust methods are more reliable than those following from NML (Yuan & Zhong, 2008; Zhong & Yuan, 2011). For those who are interested in rigorous model inference, we recommend using the bootstrap to study the performance of each of the statistics. With bootstrapping, there is no need to assume a chi-square distribution for each of the statistics. For TSR and DR methods, one can also study the empirical distributions of the statistics by changing the tuning parameter φ (see Yuan & Hayashi, 2003).

We want to emphasize the specification of models using EQS syntax so that `semdiag` can correctly parse the EQS input file. EQS allows specification of multiple equations separated by semicolons on the same line; and it also allows specification of multiple factor or error variances–covariances using one equation. To ensure the correct communication between `semdiag` and EQS during the iterations for robust estimation, it is required that each line of EQS syntax contain one equation in the `/EQUATIONS` section, and each free variance or covariance be specified by a separate equation in the `/VARIANCES` section. For example, although `D2-D3=*;` is allowed by EQS to specify free variances for D_2 and D_3, they are better specified as `D2=*;` and `D3=*;` on two separate lines.

In addition to EQS, the `semdiag` package also works with the R `sem` package for model and data diagnostics. Interested readers are welcome to modify our R source codes to utilize other SEM software. By default, EQS is used as specified by `software="EQS"`. To use the R `sem` package, one needs to change this argument to `software="sem"` and supply an RAM path object using the argument `ram.path`. Sample R codes based on the `sem` package for the analysis used in this article are provided in the help document of `semdiag`. Information on how to use the `sem` package to specify an SEM model can be found in Fox (2006).

We also would like to note that the model diagnostics tools as well as the `semdiag` package introduced in this article aim for situations when one substantive population distribution is behind the sample. The sample might be contaminated or contain outliers, but there is no foreseeable interesting population for the outliers. When multiple samples come from multiple populations and the membership of each observation is known, we can apply the developed technique to each of the samples before conducting a standard multiple group analysis. When a mixture population is behind the sample and the membership of each observation is unknown, one could use the technique of finite mixture modeling for data analysis (McLachlan & Peel, 2000). If data contamination or outliers exist, they also create problems for model evaluation and lead to biased estimates with finite normal mixture models. Currently, model diagnostics parallel to the development in this article do not exist for finite mixture SEM models. Once developed, they will provide better tools for statistical modeling with a heterogeneous population.

ACKNOWLEDGMENT

This research was supported by Grants DA00017 and DA01070 from the National Institute on Drug Abuse. We would like to thank Xin Tong and Rebecca Cheung for testing the package. We also would like to thank Dr. Deborah Bandalos and three reviewers for constructive comments on an earlier version of the article.

REFERENCES

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York, NY: Wiley.

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

Cook, D. R., & Weisberg, S. (1982). *Residuals and influence in regression*. New York, NY: Chapman & Hall.

Fox, J. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling, 13,* 465–486.

Gnanadesikan, R. (1997). *Methods for statistical data analysis of multivariate observations* (2nd ed.). New York, NY: Wiley.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Mair, P., Wu, E., & Bentler, P. M. (2010). EQS goes R: Simulations for SEM using the package REQS. *Structural Equation Modeling, 17,* 333–349.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Yuan, K.-H., & Bentler, P. M. (1998). Structural equation modeling with robust covariances. *Sociological Methodology, 28,* 363–396.

Yuan, K.-H., & Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 54,* 161–175.

Yuan, K.-H., Fung, W. K., & Reise, S. (2004). Three Mahalanobis-distances and their role in assessing unidimensionality. *British Journal of Mathematical and Statistical Psychology, 57,* 151–165.

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.

Yuan, K.-H., & Hayashi, K. (2010). Fitting data to model: Structural equation modeling diagnosis using two scatter plots. *Psychological Methods, 15,* 335–351.

Yuan, K.-H., & Zhong, X. (2008). Outliers, leverage observations and influential cases in factor analysis: Minimizing their effect using robust procedures. *Sociological Methodology, 38,* 329–368.

Zhong, X., & Yuan, K.-H. (2011). Bias and efficiency in structural equation modeling: Maximum likelihood versus robust methods. *Multivariate Behavioral Research, 46,* 229–265.

APPENDIX A

```
/TITLE
Model built by EQS 6.1 for Windows (build 97) and onward
/SPECIFICATIONS
DATA=data.txt; VARIABLES=9; Cases=100;
METHOD=ML; ANALYSIS=COVARIANCE; MATRIX=COVARIANCE;
/EQUATIONS
 V1 =  1F1 + E1;
 V2 =  1*F1 + E2;
 V3 =  1*F1 + E3;
 V4 =  1F2 + E4;
 V5 =  1*F2 + E5;
 V6 =  1*F2 + E6;
 V7 =  1F3 + E7;
 V8 =  1*F3 + E8;
 V9 =  1*F3 + E9;
 F2=*F1+D2;
 F3=*F1+*F2+D3;
/VARIANCES
 F1 = *;
 D2 = *;
 D3 = *;
 E1 = *;
 E2 = *;
 E3 = *;
 E4 = *;
 E5 = *;
 E6 = *;
 E7 = *;
 E8 = *;
 E9 = *;
/PRINT
FIT=ALL;
/OUTPUT
CODEBOOK;
DATA=semplot.ETS;
PARAMETER ESTIMATES;
STANDARD ERRORS;
LISTING;
/END
```

APPENDIX B

```
library(semdiag)
setwd("C:/research/SEMDiagnostics")
## Example 1. Normally distributed data
N100<-read.table("N100.dat")

## The EQS input file is semplot.eqs
## Model 1: treating prediction errors as factors
N100out.1<-semdiag(N100, "semplot.eqs", D="F")

## Diagnostics plot
semdiag.plot(N100out.1)

## Summary output
semdiag.summary(N100out.1)

## Model 2: treating prediction errors the same as measurement errors
N100out.0<-semdiag(N100, "semplot.eqs")

## Diagnostics plot
semdiag.plot(N100out.0)

## Summary output
semdiag.summary(N100out.0)

## Example 2. Contaminated data
N85<-read.table("N85.dat")

```

```
## The EQS input file is semplot.eqs
## Model 1: treating prediction errors as factors
N85out.1<-semdiag(N85, "semplot.eqs", D="F")

## Diagnostics plot
semdiag.plot(N85out.1)

## Summary output
semdiag.summary(N85out.1)

## Model 2: treating prediction errors the same as measurement errors
N85out.0<-semdiag(N85, "semplot.eqs", D="E")

## Diagnostics plot
semdiag.plot(N85out.0)

## Summary output
semdiag.summary(N85out.0)

## Case profile plot
semdiag.cpp(N85out.0, cases=c(86, 90, 98:100))

## Delete the 99th and 100th observations
N85out.1.del<-semdiag(N85, "semplot.eqs", D="F", delete=c(99,100))

## Sample output from semdiag.summary function
> semdiag.summary(N85out.1)
Leverage observations and outliers = 91 96 97 98 99
Leverage observations not outliers = 86 87 88 89 90
Outliers not leverage observations = 92 93 94 95 100

Model fit comparison
              NML      TSR       DR
Statistics 91.681 45.63700 36.81900
df         24.000 24.00000 24.00000
p-value     0.000  0.00489  0.04561

Parameter estimates
                NML              TSR               DR
Label    Est.  S.E.     z  Est.  S.E.     z  Est.  S.E.     z
(F1,F1)  2.34  0.54  4.32  1.39  0.35  3.95  1.81  0.41  4.46
(E1,E1)  1.62  0.29  5.60  1.18  0.22  5.46  1.09  0.21  5.26
(E2,E2)  1.41  0.27  5.18  1.21  0.22  5.49  1.19  0.21  5.57
(E3,E3)  1.81  0.31  5.89  1.22  0.23  5.35  1.19  0.23  5.27
(E4,E4)  1.98  0.47  4.25  1.51  0.30  5.08  1.43  0.30  4.81
(E5,E5)  1.86  0.42  4.41  1.13  0.27  4.16  1.20  0.27  4.42
(E6,E6)  2.36  0.40  5.86  1.39  0.26  5.44  1.29  0.24  5.27
(E7,E7)  2.80  0.42  6.62  1.50  0.25  5.95  1.39  0.24  5.74
(E8,E8)  0.92  0.34  2.75  1.04  0.25  4.20  1.01  0.24  4.31
(E9,E9)  2.76  0.43  6.43  1.61  0.28  5.84  1.53  0.27  5.70
(D2,D2)  1.90  0.54  3.54  1.09  0.32  3.39  1.42  0.37  3.82
(D3,D3)  0.17  0.12  1.38  0.30  0.14  2.11  0.40  0.16  2.48
(V2,F1)  1.05  0.13  8.06  1.01  0.15  6.63  0.95  0.12  7.71
```

```
(V3,F1)  0.95  0.13  7.30  1.05  0.16  6.74  1.04  0.13  8.01
(F2,F1)  0.83  0.16  5.29  0.87  0.17  5.07  0.81  0.15  5.50
(F3,F1)  0.66  0.16  4.23  0.73  0.19  3.92  0.79  0.16  4.96
(V5,F2)  0.94  0.13  7.08  1.05  0.14  7.46  1.00  0.12  8.23
(V6,F2)  0.72  0.12  6.16  0.87  0.13  6.87  0.85  0.11  7.76
(V8,F3)  1.75  0.32  5.41  1.32  0.20  6.49  1.23  0.15  8.13
(V9,F3)  1.20  0.26  4.55  1.08  0.19  5.75  1.06  0.15  7.26
(F3,F2) -0.01  0.08 -0.19  0.09  0.12  0.79  0.10  0.11  0.90

Note: NML=Normal ML, TSR=Two-stage robust method, DR=Direct robust
method, Est.=Parameter estimates,  S.E.=Standard error, z=Z-score.
```

APPENDIX C

1. semdiag Function

The function semdiag calculates the two M-distances d_f and d_r, and obtains parameter estimates for the model. The full specification of the semdiag function is

```
semdiag(x, EQSmodel, varphi = 0.1, EQSdata = "data.txt", model = "E",
    delete = integer(0), max_it = 1000, EQSprog ="C:/Progra~1/EQS61/WINEQS",
    serial = "111111 222222 333333", ram.path, software = "EQS")
```

The first argument x specifies the data to be used. The second argument EQSmodel provides the name of the EQS input file. These two arguments are required. The third augment varphi=0.1 specifies the Huber-type weight function that gives the approximate proportion of cases to be down-weighted. The default value is 10%. The fourth argument EQSdata specifies the file name to save the estimated covariance matrix for use in EQS and should be the same as the file name for the argument data in the EQS input file (e.g., Line 4 in Appendix A). The fifth argument model tells how to treat Ds in the analysis (D="E" or "F"). The next argument delete supplies the case or row numbers as appeared in the data matrix for the observations to be deleted, which can be a single number such as delete=99 or a vector such as delete=c(95, 97, 100). If not provided, no data will be deleted. The argument max_it defines the maximum number of iterations for the robust algorithms. The default is 1,000. The argument EQSprog is the string containing the default installation path of EQS software and the argument serial provides the serial number for the EQS software, which is typically three sets of numbers separated by space. The argument ram.path represents an SEM model specified in R using the sem package. The last argument software specifies which software is used. By default, EQS is used and changing it to "sem" will utilize the sem package for the analysis.

2. semdiag.plot Function

The function semdiag.plot generates the two plots. Its full specification is

```
semdiag.plot(d, alpha=.01, label=0, cex=1)
```

The first argument d is the output of the function semdiag. The second argument alpha is the α level for drawing the two dashed lines in the scatter plot of Figure 4a to 7a. The default value is 0.01. The argument label allows the labeling of the observations in the two plots by their case numbers. By default, label=0 labels cases that are outstanding according to the two dashed lines. By setting label=1, one can label the observations manually by clicking on the points in the plots. If label=2, in addition to automatically labeling the outstanding cases, one can also label extra observations manually. The last argument cex=1 controls the size of the labels in the plots.

3. semdiag.summary Function

The function semdiag.summary prints the case numbers of the leverage observations and potential outliers on the screen. It also outputs the parameter estimates and their standard errors following from the NML, TSR, and DR methods. The full specification of semdiag.summary is

```
semdiag.summary(d, alpha=.01, digits=2)
```

The first argument d is the output of the function semdiag. The second argument alpha is the same as that in semdiag.plot. The default value is 0.01. The third argument digits controls the number of decimals for parameter estimates and their standard errors being printed.