Multivariate Behavioral Research, 25 (2), 173-180

Structural Model Evaluation and Modification: An Interval Estimation Approach

James H. Steiger
University of British Columbia

Procedures for evaluation and sequential modification of structural models have attracted much interest in the recent psychometric literature. Kaplan (1990) proposes to extend a procedure (post hoc model modification, or PMM) popularized by Jöreskog and Sörbom (1984). PMM is designed for cases where one's best attempt at an a priori theoretical model has been found to have poor or marginal fit to the sample data. The researcher, in desperation, may wonder if there is any model which fits the data. The PMM approach uses *modification indices* to predict which path, if added to a structural diagram, would decrease the chi-square fit statistic the most. One frees the parameter associated with that path to obtain an improved model.

Kaplan (1990) recognizes some technical problems with the procedure. In particular, a parameter when freed may not change much from zero, even though the chi-square changes a lot. This can occur for at least two reasons. First, sample size may be huge, and so the big change in chi-square was really associated with a trivial imperfection in the model. (This is a tip-off that one has *too much power.*) Second, the model may have been badly misspecified. In that case, adding a parameter alters the estimated variance-covariance structure of the parameter estimates significantly, thus altering the weighting of residuals in computing the loss function. In either case, freeing the parameter is a bad idea.

By using the EPC statistic, Kaplan (1990) hopes to diagnose such occurrences in advance.

The procedure described by Kaplan is not radical, and in the context of mind-withering technical articles on structural modeling, surprisingly simple. But does it make good statistical sense?

I don't think so. The problem is not so much with Kaplan (1990) as with the tradition he is trying to improve. The PMM procedure violates some of the most basic statistical principles:

1. It performs post hoc selection without post hoc protection;

2. It implicitly assumes that a model remains unspecified in a particular way. If this assumption is wrong, the procedure may function in a bizarre fashion.

These points are easy to grasp in context of the following case study.

> You have sampled a 6 × 6 covariance matrix based on 100 observations from a multivariate normal distribution. You do not know it, but in the population your variables have a correlation matrix which is an equicorrelation matrix (i.e., all correlations are equal to each other). In this case the single value all correlations are equal to is, in the population, .20. You do not know this either.

Suppose you choose to examine the data as a correlation structure. You begin by testing the hypothesis that all correlations are zero, as a *null model*, using LISREL. (My own program, EzPATH, would allow you, starting from scratch, to set up *and* test such a model in less than 5 minutes.) Suppose (as would usually happen in the case study) you encounter a relatively high chi-square statistic with low probability. You are then ready to indulge in the kind of statistical exercise described by Kaplan (1990).

How will Kaplan's (1990) procedure behave in this situation? Very badly. The procedure will pick the largest non-zero element in the correlation matrix, and set it equal to a free parameter. It will keep doing this until the chi-square statistic becomes *insignificant.*

The *correct* model, of course, requires freeing only one parameter, and setting all the correlations equal to it. Kaplan's (1990) procedure, (and the *automatic modification* system of LISREL) generates, for any given sample, a few free parameters corresponding to the largest correlations, and keeps the others (erroneously) constrained to zero.

Just to see *how* badly this procedure would behave, I generated two simulated samples in the situation I described, using a normal random number generator. In one, LISREL freed the correlation between variables 6 and 4. This correlation was, in the Monte Carlo sample, .336. In the second example LISREL freed *four* different correlations before stopping, and the largest was .366.

Thus, in a situation where all correlations are .20 in the population, the PMM procedure ended up deciding in my first sample that an acceptable model is that all correlations but one are actually zero, and the other is .336.

This example dramatizes the dangers inherent in performing post hoc analysis without appropriate statistical protection. This problem generalizes to most situations where (a) a model has been misspecified, and (b) many of the zero parameters have moderate, non-zero, relatively equal values. In such cases, unless sample size is very large, sampling error generates an appreciable spread of estimated values for these parameters in the sample. If you select data post hoc, you can get a very misleading value of your *largest* parameter if you do not take this into account. Tukey recognized this more than 30 years ago in the context of the analysis of variance, and we have taught his insight, and his test, to a generation of undergraduates. Unfortunately, our statistical powers of generalization seem to have been quite limited.

Note that the *free one parameter and assign it to one coefficient* approach of PMM rules out a whole class of hypotheses, that is, those allowing assignment of the same free parameter to more than one coefficient. Kaplan (1990) might argue that LISREL ignores this possibility because such a model specifying parameters are precisely equal is not realistic. But is a model where many factor loadings are (precisely) zero any more realistic?

Kaplan (1990) might also respond that his procedure is not intended to be used unless *all important internal errors have been removed.* This raises the interesting question of how, when one obtains a *good* fit (one would obtain a reasonably *good* fit with the model LISREL gives for Case 1 data), one is supposed to know that serious internal errors have *not* been removed.

In the hope of provoking discussion and controversy, I will pose the following three unsolved riddles of structural modeling.

1. Why do *LISRELITES* insist on performing an inappropriate post hoc analysis without post hoc protective techniques?

2. Why do the modification procedures favored by Kaplan (1990) and others rule out hypotheses which assign one free parameter to several coefficients?

3. What is the best *protected* post hoc procedure (analogous to the Scheffe test in analysis of variance) for structural modeling?

Kaplan (1990), perhaps sensitive to riddle 1, states that one should only free a parameter when "there is sufficient theoretical justification for doing so." This raises a fourth riddle.

4. What percentage of researchers would ever find themselves unable to think up a *theoretical justification* for freeing a parameter?

In the absence of empirical information to the contrary, I assume that the answer to riddle 4 is “near zero.” Call me a cynic if you like, but before judging me too harshly, read the rich (and somewhat contradictory) psychological literature on smoking and personality. You will see many interesting mini-theories developed to explain correlations selected post hoc from huge matrices, and declared *significant* without appropriate protective techniques.

I conclude that Kaplan’s (1990) procedure, attempting to improve on a technique which is statistically bankrupt, is rather like trying to get a 20 year old car to run like a new one by pumping more air in the tires. Probably the procedure will work in the highly restricted class of situations where several paths in a path diagram are zero, only one path (or perhaps two or three) needs to be added to produce the *true* model, and the others are really zero paths. (Someone will probably do a Monte Carlo study to prove this.) Does anyone out there think this is a prototypic situation?

Is there some way to employ PMM, while avoiding publication of incorrect models? Possibly. The vast majority of incorrect models generated by unprotected post hoc model modification would fail to fit a cross-validation sample. Until we get a strong answer to riddle 3, it might be well to remember the following adage: *An ounce of replication is worth a ton of inferential statistics.*

Perhaps a moratorium should be declared on publication of causal modeling articles using any PMM procedure similar to Kaplan's (1990), unless such articles provide evidence of cross-validation.

The literature of causal modeling already provides massive evidence that practitioners will imitate what software manual writers do, rather than what they preach. Mild cautions (e.g., Sörbom, 1989) about the post hoc fallacy, rendered almost as an afterthought along with suggestions to cross-validate, will have no effect unless violators are attacked in print. Strong action is necessary to clean up the PMM mess. I have additional thoughts on this, but space here is too limited to register them effectively.

My second major objection to Kaplan's (1990) article is that it accepts, rather uncritically, post hoc power analysis as a useful tool in covariance structure modeling. Such analysis is not necessary to protect against cases where a significant chi-square statistic results from *too much power.*

I have argued recently (Steiger, 1989) and not-so-recently (Steiger & Lind, 1980) that the appropriate question for statistical analysis in covariance structure modeling is not whether fit is perfect. Rather, we should ask three questions. They are:

1. How well does my model fit my statistical *population*?
2. How *precisely* have we determined population fit from our sample data?
3. Does the fit still appear good when we take into account the complexity of the model and its number of free parameters?

Pursuing these questions leads to the conclusion that fit coefficients should be (a) based on a population rationale, rather than heuristic arguments or a sample rationale, (b) relatively unbiased, (c) relatively uninfluenced by sample size, (d) reported with a confidence interval, and (e) adjusted for model complexity.

An interval estimate reduces the chance of (a) rejecting a model which fits very well (but not perfectly) because sample size is *too large*, and (b) accepting a fit as good because the sample goodness of fit coefficient is high, when in fact precision of estimation of the fit is too low to warrant such confidence.

My computer program EzPATH (Steiger, 1989) gives indices of fit which are relatively uninfluenced by sample size. *It also calculates and reports confidence intervals for these coefficients.* Situations (a) and (b) in the preceding paragraph are detected in examples from the published literature. For example, Jöreskog (1978) tested a circumplex hypothesis on a 6 × 6 correlation matrix with n=710. Power was extremely high in this situation. Although these data were often considered to fit a circumplex well, Jöreskog concluded they did not, because the chi-square statistic reached a probability level of .0076. This is a classic case of too much power causing the chi-square test to be too sensitive to minor departures from perfect fit. One could do an elaborate power analysis to demonstrate this, but it is not necessary. The confidence interval on a population equivalent of the Goodness-of-Fit Index (GFI) reported by LISREL is computed by EzPATH. The 90% confidence interval ranges from .984 to .998! This demonstrates, simply and eloquently, that goodness of fit, though not perfect, has been determined with high precision to be outstanding. The significant chi-square can be ignored.

My interest in the statistical estimation of fit indices dates back more than a decade. In 1980 (Steiger & Lind, 1980), I presented an original notion — that the *population noncentrality index* Φ (the value of the maximum likelihood or generalized least squares discrepancy function obtained when a model is fit to the *population covariance matrix* Σ) could be used as a natural measure of badness of fit of a covariance structure model. I pointed out that this notion can be improved on, however, because such a measure did not take into account the inevitable effect of model complexity on population fit. Following a logic similar to that of James, Mulaik, and Brett (1982), I suggested correcting this index for model complexity by dividing by degrees of freedom, v. Taking the square root of the resulting ratio gives an index which has much in common with a root mean square standardized effect measure which can be calculated as a summary measure of effect size in the analysis of variance (Steiger, 1990). The resulting RMS index is, in the population

$$\text{RMS} = (\Phi/v)^{1/2}$$

I am pleased that a number of colleagues and friends who attended the Steiger-Lind paper presentation in 1980 and did not find noncentrality estimation a particularly compelling notion at the time have now embraced the idea and/or suggested variants of it (See, e.g., Bentler, 1989, 1990; Browne & Du Toit, 1989; McDonald, 1989; McDonald & Marsh, 1990). It is doubly gratifying that some of those authors (Browne & Du Toit, 1989; McDonald & Marsh, 1990), as well as others (Mels, 1989) have cited the Steiger-Lind (1980) paper and recognized its priority.

There are many possible indices of fit for structural models. Indeed, there are infinitely many ways to transform the RMS index (or, indeed, Φ itself) onto the interval from 0 to 1. If a particular fit index can be expressed as a function of a single noncentrality index, and that function satisfies an obvious monotonicity condition, it is easy to obtain a confidence interval for the index (see, e.g., Cox and Hinckley, 1974, p. 212). In the user's guide for my computer program EzPATH (Steiger, 1989), I proved a surprising result which makes good use of this fact. If a structural model is ICSF (invariant under a constant scaling factor) in the sense of Browne (1982), then the *population equivalents* of the widely used Jöreskog-Sörbom (1984) GFI and AGFI fit indices can be shown to be a simple function of the population noncentrality index. Specifically, let Γ_1 be the population equivalent of the GFI and Γ_2 the population equivalent of the AGFI (i.e., the values which would be obtained if these indices were computed on the population covariance matrix Σ). Steiger (1989), relying heavily on results from Browne (1974) and Tanaka and Huba (in press), proved that for a $p \times p$ covariance matrix Σ, with $p^* = p(p + 1)/2$,

$$\Gamma_1 = p / [2\Phi + p]$$

and

$$\Gamma_2 = 1 - (p^*/v)(1 - \Gamma_1).$$

Because both coefficients are simple functions of a single noncentrality index, we can obtain consistent estimates, *and confidence intervals*, for them. It turns out (Steiger, 1989) that at small to moderate sample sizes the sample GFI and AGFI computed by LISREL can be seriously biased, and underestimate the corresponding population quantity substantially.

The interval estimation approach will help avoid the fallacies of testing with too much *or* too little power. It will not compensate for inappropriate post hoc testing.

Even if we avoid PMM, some major problems remain. Suppose we are testing a nested sequence of models. Adding parameters to a model will generally improve fit. One view of the model fitting process is that the model and data are like pieces of a jigsaw puzzle. We want *goodness of fit* which is a result of a special match of model to data, rather than an inevitable consequence of model complexity. If, in general, one model tends to fit data better than another model, we should be prepared to compensate. But...how? This calls to mind one of the great unsolved riddles of structural modeling.

5. How do we properly compensate for the fact that increasing complexity of a model almost inevitably results in improvement in fit?

Providing a coherent rationale for solving riddle 5 may require a broader perspective than that offered by the data at hand. One approach might be to address the problem: In general, how well does this model fit covariance matrices of this order? Examining this notion, Botha, Shapiro, and Steiger

(1988) fit common factor models to randomly generated *population* matrices. Simpler approaches (Steiger & Lind, 1980; James et al., 1982; Steiger, 1989) assume that tautological improvement in fit is a simple function of degrees of freedom lost (i.e., of free parameters gained). This assumption is possibly too simplistic.

A closing thought: In the final analysis, it may be, in a sense, impossible to define one *best* way to combine measures of complexity and measures of badness-of-fit in a single numerical index, because the precise nature of the *best* numerical tradeoff between complexity and fit is, to some extent, a matter of personal taste. The choice of a model is a classic problem in the two-dimensional analysis of preference. From that point of view, there can never be one *best* coefficient for assessing fit (or a model which is indisputably *best*) any more than there is a single *best* automobile. On the other hand, I agree completely with the sentiments expressed in James et al. (1982) – model parsimony should be addressed in any reasonable model selection procedure.

References

Bentler, P. M. (1989). *EQS structural equations program manual. Los Angeles*: BMDP Software.

Bentler, P. M. (1990). Fit indices, LaGrange multipliers, constraint changes, and incomplete data in structural models. *Multivariate Behavioral Research, 25*, 163-172.

Botha, J. D., Shapiro, A., & Steiger, J. H. (1988). Uniform indices of fit for factor analysis models. *Multivariate Behavioral Research, 23*, 443-450.

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

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72-141). London: Cambridge University Press.

Browne, M. W., & Du Toit, S. H. C. (1989, October). *Models for learning data.* Paper presented at the Conference on Best Methods for the Analysis of Change, University of Southern California, Los Angeles.

Cox, D. R., & Hinckley, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data.* Beverly Hills, CA: Sage Publications.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43*, 443-477.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods* [Computer program]. Mooresville, IN: Scientific Software.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research, 25*, 137-155.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6*, 97-103.

McDonald, R. P., & Marsh, H. W. (in press). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin.*

Mels, G. (1989). *A general system for path analysis with latent variables.* Unpublished masters thesis. University of South Africa, Department of Statistics, Pretoria, South Africa.

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371-384.

Steiger, J. H. (1990). Noncentrality interval estimation and the evaluation of statistical models. Manuscript in preparation.

Steiger, J. H. (1989). *EZPATH: A supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors.* Paper presented at the annual Spring meeting of the Psychometric Society, Iowa City, IA.

Tanaka, J. S., & Huba, G. J. (in press). A general coefficient of determination structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology.*