Multivariate Behavioral Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

Weighted Structural Regression: A Broad Class Of Adaptive Methods For Improving Linear Prediction

Robert M. Pruzek & Greg M. Lepak

Available online: 10 Jun 2010

To cite this article: Robert M. Pruzek & Greg M. Lepak (1992): Weighted Structural Regression: A Broad Class Of Adaptive Methods For Improving Linear Prediction, Multivariate Behavioral Research, 27:1, 95-129

To link to this article: http://dx.doi.org/10.1207/s15327906mbr2701_7

Multivariate Behavioral Research, 27 (1), 95-129

Weighted Structural Regression: A Broad Class Of Adaptive Methods For Improving Linear Prediction

Robert M. Pruzek
Department of Educational Psychology and Statistics
State University of New York at Albany

Greg M. Lepak
Department of Business Administration
Le Moyne College

Given a criterion variable and two or more predictors, applied linear prediction usually entails some form of OLS regression. But when there are several predictors, and especially when these are subject to non-ignorable errors of measurement, applications of OLS methods are often fraught with problems. Weighted structural regression (WSR) methods can mitigate many difficulties through the incorporation of prior structural models into analyses. WSR methods are sufficiently general to include OLS, ridge, reduced rank regression, as well as most covariance structural regression models, as special cases; many other regression methods, heretofore not available, are also included. In this article adaptive forms of WSR are developed and discussed. According to our bootstrapping studies the new methods have potential to recover known population regression weights and predict criterion score values routinely better than OLS with which they are compared. These new methods are scale free as well as simple to compute; they seem well suited to many prediction applications in behavioral research.

Although ordinary least squares (OLS) regression methods are widely used in prediction, a number of problems are known to restrict their applicability, especially in circumstances commonly encountered in behavioral research. Unlike prediction in the physical or natural sciences, three major problems typically characterize regression applications in the social and behavioral sciences: (a) there are usually numerous social and behavioral variables from which one must select or compose prospective predictors for analyses; (b) sample size n is often limited, frequently so much that the ratio of p, the number of variables, to n may approach – and can even exceed – unity; and (c) there are generally non-ignorable measurement errors associated with individual observations, for both predictors and criterion variables, errors that tend to weaken or confuse predictive relationships.

These features of social and behavioral data systems have a major role in complicating researchers' abilities to find stable or interpretable predictive

models. However, as we shall show, regression methods can be devised to account for measurement errors, and that do not necessarily deteriorate as p grows larger for a given n.

In this article we focus on regression methods intended for general exploratory applications where there may be several correlated predictors, where sample sizes may be limited, and where one's prior knowledge of relationships among variables is likely to be relatively vague or diffuse. We are not concerned with fixed predictors, as typically employed in experimental contexts, but rather with random or stochastic variables, which properly characterize most predictors used in observational studies.

The principal purpose of this article is to describe and discuss a class of stochastic variable linear prediction methods that provides ways for analysts to incorporate prior or collateral information into analyses. It will be assumed that the prior information to be used takes the form of a structural model, one that generally assumes the existence of measurement errors in all observed variables. Unlike recently developed structural analytic approaches to regression which are designed to estimate parameters within the context of specific, prechosen structural models, the methods to be examined here incorporate structural information in an adaptive fashion, capitalizing on it only to the extent that extant data are consistent with that information. Thus, in relation to conventional structural methods the new ones generally offer more flexible ways for analysts to incorporate prior information into analyses. As we shall see below, the role of prior information tends to be of distinctive value in situations where there are numerous predictors, especially when sample sizes are limited. (See Laughlin, 1986, for an excellent review of the role of prior information in selected regression methods.)

Our general approach uses both psychometric and statistical principles to generate a class of methods that can be supported with either frequentist or Bayesian arguments. However, in relation to most previously developed statistical methods for incorporating prior information the present system based on using prior structural information will often provide substantial advantages to the analyst even if the prior structural model is relatively vague or diffuse. Although it is possible to use the new methods with virtually any type of covariance structure model, ranging from exploratory to confirmatory, in this article we have chosen to emphasize the exploratory mode.

The class of methods under consideration is described as weighted structural regression (WSR). Although WSR methods include OLS methods within their broad framework, most WSR estimators have properties that distinguish them from OLS estimators. All non-OLS forms of WSR coefficient estimators are generally biased, at least when there are no observation errors

in the variables. Yet our simulation results suggest that reductions in sampling variability will often more than compensate for the bias, and can lead to results that are more readily interpretable than their OLS counterparts. In particular, our simulations, which are based on a real data prediction problem, suggest that the new adaptive regression estimators can typically recover population weights (βs) with lower mean squared errors than their least squares stochastic variable counterparts; and predictor score values (*y*s) can generally be estimated more accurately in cross-validation samples when using WSR methods. Furthermore, the algorithms for these methods need not break down even when the original covariance or correlation matrix is singular.

Chen (1979), in an article that strongly influenced our thinking, developed a theoretically sophisticated Bayesian approach to regression using structural priors in combination with empirical data to generate posterior distributions from which prediction equations can be derived. But despite the theoretical appeal and formal rigor of his methods, we have been unable to find a single prediction application based on Chen's methods in the decade following his publication. Although our general approach parallels Chen's, we do not use maximum likelihood estimation techniques for structural estimation since their use with small samples often tends to be problematic. Instead, we concentrate on minimizing a risk function. Coefficient estimators thus derived will be called MinRisk (MR) estimators. Although we do not use Chen's estimation methods, by connecting our work with Chen's theoretical framework, it is possible to provide a basis for estimating standard errors for MR coefficient estimators, thus providing a general operational system for inference that in many respects parallels conventional OLS methodology.

The central principle of our approach to WSR estimation is that when the joint predictor-criterion covariance or correlation matrix can be even roughly approximated using a relatively parsimonious structural model, for example, a common factor model of low dimensionality, one can incorporate information from that model to stabilize, and also possibly enhance the interpretability of, derived regression equations.

The first step in producing WSR estimators is to generate a convex sum estimator of the *joint* predictor-criterion covariance matrix. This estimator has the form

$$\hat{\Sigma}_{cs} = w\hat{\Sigma}_{MF} + (1 - w)(\hat{\Sigma}_{model\text{-}based}), \tag{1}$$

where $\hat{\Sigma}_{MF}$ represents the conventional model-free covariance estimator, and $\hat{\Sigma}_{model\text{-}based}$ refers to a model-derived covariance estimator based on the same data used to generate $\hat{\Sigma}_{MF}$. The scalar $w(0 \leq w \leq 1)$ will generally be computed on

the basis of lack of fit of the structural model to observed data. WSR regression coefficients are generated from $\hat{\Sigma}_{cs}$ in a manner that directly parallels OLS procedures for random variable prediction.

The remainder of this article is divided into six sections. In the first we provide a brief introduction to weighted structural regression, and review sampling assumptions used in the following technical sections. Next, we present two distinctive forms of covariance estimators based on minimizing risk; this entails optimizing weights based on loss functions, in one case for populations with no covariance structure and in another, for populations that are taken to have approximate common factor structure. The third section applies the basic results of risk theory to regression estimation. In the fourth section we describe a scaling system and propose a corresponding common factor algorithm, aspects that bear directly on how the new methods can be used effectively in practice. Next, we present a modest bootstrapping study, using real data to compare several forms of the new methods with their OLS counterparts. In the sixth section we briefly discuss the rationale of the new prediction methods, compare them with conventional methods, and make some suggestions concerning applications.

Some Technical Background for Weighted Structural Regression

For any set of k predictor variables and a criterion, OLS regression equations can be simply derived from the joint covariance matrix of all observed variables. In particular, if y designates the criterion variable and x the predictors, then the $(1 \mid k)$ symmetrically partitioned sample covariance matrix

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{yy} & \hat{\sigma}_{yx} \\ \hat{\sigma}_{xy} & \hat{\Sigma}_{xx} \end{bmatrix}, \tag{2}$$

can provide the basis for computing $\hat{\beta}$, the $k \times 1$ vector of ordinary least squares regression coefficients,

$$\hat{\beta}_{ols} = \hat{\Sigma}_{xx}^{-1}\hat{\sigma}_{xy}, \tag{3}$$

where $\hat{\Sigma}_{xx}$ represents the $k \times k$ covariance matrix for the independent variables, and $\hat{\sigma}_{xy}$ is the vector of k predictor-criterion covariances. (We assume without loss of generality that all variables been converted to deviation score form; thus the intercept term is necessarily zero. Moreover, any variable in a system can be moved to the first position so this also is no restriction.)

WSR estimators are of the same form as Equation 3, except the covariance matrix they derive from is a convex sum estimator of the form of expression in Equation 1. In particular, for present purposes we shall suppose that the model-based covariance estimator in Equation 1 takes the form $\hat{\Sigma}_{cfa} = \hat{\mathbf{F}}\hat{\mathbf{F}}' + \hat{\mathbf{U}}^2$, for $\hat{\mathbf{F}}$ of order $p \times m$ with m orthogonal factors. If the matrices $\hat{\mathbf{F}}$ and $\hat{\mathbf{U}}^2$ are partitioned to correspond to the criterion and predictor variables where,

$$(4) \qquad \hat{\mathbf{F}} = \begin{bmatrix} \hat{f}'_y \\ \hat{\mathbf{F}}_x \end{bmatrix} \text{ and } \hat{\mathbf{U}}^2 = \begin{bmatrix} \hat{u}_{yy}^2 & 0 \\ 0 & \hat{\mathbf{U}}_{xx}^2 \end{bmatrix},$$

and the partitioned form of the joint predictor covariance matrix in Equations 2 and 3 is replaced with the convex sum estimator in Equation 1, then with some algebraic manipulation, it is seen that the WSR coefficient vector takes the form

$$(5) \qquad \hat{\beta}_{WSR\text{-}cfa} = \{w\hat{\Sigma}_{xx} + (1 - w)(\hat{\mathbf{F}}_x\hat{\mathbf{F}}'_x + \hat{\mathbf{U}}_{xx}^2)\}^{-1}\{w\hat{\sigma}_{xy} + (1 - w)\hat{\mathbf{F}}_x\hat{f}_y\}.$$

Note that if the scalar $w = 1$, Equation 5 is equivalent to Equation 3; if $w = 0$, then Equation 5 generally yields a reduced rank representation for the vector of regression coefficients (Lawley & Maxwell, 1973; Pruzek & Frederick, 1978). If $m = 0$ (so that each term with $\hat{\mathbf{F}}_x$ drops out), with an intermediate value of w, then Equation 5 depicts a ridge representation for the vector of regression weights. Generally, however, for any given number of common factors, m, the scalar w can be preset according to how much weight the analyst chooses to give to the respective covariance estimators in the convex sum; alternatively, w can be estimated from the data adaptively, thus forming a convex sum based on how well the data support the given common factor model. In this article we focus on the latter concept.

In the following sections the model-based covariance estimator, $\hat{\Sigma}_{model\text{-}based} = \hat{\Sigma}_{cfa}$, is computed using specialized common factor procedures that tend to ensure scale-freeness of derived factors. Development of MinRisk methods will entail generating a specific scalar weight, w_m, to provide a particular convex sum using an m-factor model. Given this convex sum covariance matrix of the form of $\hat{\Sigma}_{cs}$ in Equation 1, one can construct the vector of corresponding regression weights, as well as $\hat{y}$ regression estimates. Although variations on the initial MinRisk procedure are considered, both in principle and in simulation studies, the main emphasis in what follows derives from minimizing risk for the joint predictor-criterion covariance matrix.

In the foregoing paragraphs the sampling model is only implicit. To make it explicit, we shall assume that a system of n observation vectors (each

composed of one criterion score, and $k = p - 1$ predictor values), $z_1, z_2, \ldots, z_n$, for $n > p$, represents a random sample of n p-dimensional scores from a multivariate normal distribution with mean zero and positive definite covariance matrix Σ; that is, $z_i \sim \text{MN}(0, \Sigma)$. (To keep the notation relatively simple we continue to assume that the first observation z_{i1} in each vector z_i represents criterion measurements so that the remaining observations $z_{i2}, \ldots, z_{ip}$ correspond to scores on stochastic predictor variables. Sample vectors z_i are also assumed to have been *centered* with respect to the vector of sample means.)

From the foregoing sampling assumption it follows that sample covariance estimators of the form $\hat{\Sigma}_{\text{MF}} = (1/n)\sum_{i=1}^{n} z_i z_i'$ follow a Wishart form, namely, $n\hat{\Sigma}_{\text{MF}} \sim W_p(\Sigma, n\text{-}1)$.

In deriving a general class of covariance estimators, we shall find it useful to consider an eigenanalysis, or singular value decomposition, of an appropriately scaled covariance estimator. The question of scaling becomes crucial in analysis since the eigen decomposition of a particular covariance matrix is not invariant to multiplicative changes in the scales of the different variables. In order to provide a means to associate the eigenstructure of an appropriately scaled covariance matrix to the structure of $\hat{\Sigma}$, the observed (raw metric) covariance matrix, we shall assume — at the outset — that the original $\hat{\Sigma}$ has been rescaled in the metric of a known diagonal scaling matrix, $\mathbf{D}^2$. Thus, $\mathbf{D}^{-1}\hat{\Sigma}\mathbf{D}^{-1} = \hat{\Sigma}^*$ will represent the scaled covariance matrix for the rescaled variables. (The asterisk denotes rescaled form; for simplicity, the symbol $\hat{\Sigma}$ will be used interchangeably with $\hat{\Sigma}_{\text{MF}}$, as used above.)

As the following developments make clear the proposed forms of covariance/regression estimators will generally depend upon the choice of $\mathbf{D}^2$, and one's prior structural model will include specification for $\mathbf{D}^2$. In the next section we shall concentrate first (Case I) on estimating population covariance matrices in the situation where no covariance model has been advanced; for Case II we shall assume an approximate factor analytic structure for the population covariance matrix.

Covariance Structure and Covariance Risk Estimation

CASE I: (No structural model restrictions on Σ)

Many of the concepts in this initial discussion are introduced primarily to provide a basis for our Case II development. The first case of interest entails adding a constant to each diagonal entry in the usual model-free maximum likelihood estimator $\hat{\Sigma}^*$ of the rescaled population covariance matrix, using the fixed diagonal $\mathbf{D}^2$ for rescaling. In applications which are discussed later, $\mathbf{D}^2$

will be estimated from the sample covariance matrix; we shall have more to say about this later.

Our first procedure for generating a covariance estimator rests on previous unpublished work of others, notably that of Press (1975) and especially Chen (1976). Press introduced the basic loss function on which we focus, namely

$$L(\tilde{\Sigma}^* - \Sigma^*) = tr(\tilde{\Sigma}^* - \Sigma^*)^2 = \sum(\tilde{\sigma}^*_{jk} - \sigma^*_{jk})^2, \tag{6}$$

where the latter summation is understood to include all rows and columns of Σ^*. The quantities $\tilde{\sigma}^*_{jk}$ and σ^*_{jk} represent entries in the matrices $\tilde{\Sigma}^*$ and Σ^*, respectively. The associated risk is defined as $R(\tilde{\Sigma}^*) = Etr(\tilde{\Sigma}^* - \Sigma^*)^2$, where $E(\cdot)$ denotes expected value operator.

Chen (1976) discussed a class of estimators of a covariance matrix in the form of a convex sum as

$$\hat{\Sigma}^*_{cs\text{-}w_0} = w_0\hat{\Sigma}^* + (1 - w_0)(g\mathbf{I}), \tag{7}$$

where $w_0 = n/(n + \gamma_0)$, $\gamma_0 > 0$, g is a specified scalar, and $\mathbf{I}$ is the identity matrix. The risk of $\hat{\Sigma}^*_{cs\text{-}w_0}$ as an estimator for the population Σ^* using the loss function in Equation 6 is minimized for this particular choice of w_0. To elaborate, define $g = \hat{g}_0 = tr(\hat{\Sigma}^*)/p = \sum\hat{\lambda}_j/p$, with $\hat{\mathbf{G}} = \hat{g}_0\mathbf{I}$, where $\hat{\lambda}_j$ denotes the eigenvalues of $\hat{\Sigma}^*$. From Equation 7, we have

$$\hat{\Sigma}^*_{cs\text{-}w_0} = \hat{\mathbf{Q}}\{w_0\hat{\Lambda} + (1 - w_0)\hat{\mathbf{G}}\}\hat{\mathbf{Q}}', \tag{8}$$

where the eigenvalues $\hat{\Lambda}$ and eigenvectors $\hat{\mathbf{Q}}$ of $\hat{\Sigma}^*$ result from $\mathbf{D}^{-1}\hat{\Sigma}\mathbf{D}^{-1} = \hat{\mathbf{Q}}\hat{\Lambda}\hat{\mathbf{Q}}'$. The risks for various covariance estimators are given in Appendix A. In particular, Equation 8 in Appendix A shows an expression for $\{R(\hat{\Sigma}^*) - R(\hat{\Sigma}^*_{cs\text{-}w_0})\}$. Denoting this difference by $\Delta R(\gamma_0)$, for each given $\gamma_0 > 0$, the optimal choice for γ_0, setting the $\partial R(\gamma_0)/\partial\gamma_0 = 0$, results in

$$\gamma_0 = \{p(1 + r_0) - 2\}/(p - r_0), \tag{9}$$

where $r_0 = (tr\Sigma^*)^2/tr(\Sigma^{*2})$ and $1 < r_0 \leq p$. For present purposes, since Σ^* is not observable, we use $\hat{r}_0 = (tr\hat{\Sigma}^*)^2/tr(\hat{\Sigma}^{*2})$ as an estimator for r_0. It can be shown that $\hat{r}_0$ tends toward p, from below, as the eigenvalues $\hat{\lambda}_j$ approach their average value, $\hat{g}_0$. Consequently, $\hat{r}_0$ provides an index on a scale from 1 to p of whether the sample $\hat{\Sigma}^*_{cs\text{-}w_0}$ lies nearer some unknown non-diagonal matrix, say $\Sigma^*_{unknown}$ (index = 1), versus $g_0\mathbf{I}$ (index = p). Given $\hat{\gamma}_0 = \{p(1 + \hat{r}_0) - 2\}/(p - \hat{r}_0)$ and $w_0 = n/(n + \hat{\gamma}_0)$, the index w_0 can be described as a badness of fit index on a scale

from zero to unity, representing how poorly $\hat{\Sigma}^*_{cs\text{-}w_0}$ in Equation 8 fits the scalar diagonal $g_0\mathbf{I}$. The complement 1 - w_0 is thus a goodness of fit index.

Chen showed that for any given positive definite covariance matrix, with $p \geq 2$, $\Delta R(\gamma_0) > 0$ for all n if and only if

$$0 < \gamma_0 < 2\{p(1 + r_0) - 2\}/(p - r_0). \tag{10}$$

This expression and Equation 9 will both be used to define regression estimators later.

The reader can see that estimators of the form of Equation 8 generally shrink all eigenvalues of the covariance estimator *toward the middle*. Consequently, use of an estimator such as Equation 8 is consistent with the fact that the largest eigenvalues of a sample covariance matrix are generally positively biased for their population analogs, and the smallest eigenvalues are generally negatively biased estimators of their population counterparts (Anderson, 1984).

It is useful for Case I to think of Σ as being approximated by a factor structure with zero common factors, so that provisionally, $\Sigma = g_0\mathbf{D}^2$, a diagonal matrix. Thus, the rescaled form $\Sigma^* = \mathbf{D}^{-1}\Sigma\mathbf{D}^{-1} = g_0\mathbf{I}$, which shows that covariance estimators of the form Equation 8 result in all sample eigenvalues being shrunk symmetrically toward $\hat{g}_0$.

The indices r_0 and γ_0 have previously been used as measures of sphericity of distribution (cf. Dempster, 1969). The covariance estimator $\hat{\Sigma}^*_{cs\text{-}w_0}$ is used in the derivation of a generalized ridge version of our WSR procedures (compare Equation 5 with Equation 22 below). We next examine the case where the population covariance matrix is assumed to have some *non-null* off-diagonal structure.

CASE II: (Using a Common Factor Model for Σ*)*

For this Case, we introduce a new assumption, namely that the population covariance matrix Σ can be *approximated* using a common factor structure (in practice, one with relatively low rank). Through use of this assumption, a new covariance estimator shall be devised that is in the spirit of $\hat{\Sigma}^*_{cs\text{-}w_0}$ in Equation 8, but where the number, m, of common factors is prespecified to exceed zero.

Lawley (cf. Lawley & Maxwell, 1971) and Rao (1955) have demonstrated that it is a useful statistical principle in common factor analyses to rescale covariance matrices in the metric of the unique-factor variances. Following their lead, and noting Harris' (1962) development of connections between statistical and psychometric approaches to factor analysis, we use an eigenvalue/

eigenvector decomposition of the rescaled covariance estimator of Σ with $\mathbf{D}^2 = \mathbf{U}^2$. In this case

$$\hat{\Sigma}^* = \mathbf{U}^{-1}\hat{\Sigma}\mathbf{U}^{-1} = \hat{\mathbf{Q}}\hat{\Lambda}\hat{\mathbf{Q}}', \tag{11}$$

where $\mathbf{U}^2$ represents a *fixed* diagonal matrix of uniqueness variances, $\hat{\mathbf{Q}}$ is a $p \times p$ matrix of column unit-length eigenvectors, and $\hat{\Lambda} = \text{diag}(\hat{\lambda}_j$ depicts eigenvalues of $\mathbf{U}^{-1}\hat{\Sigma}\mathbf{U}^{-1}$.

We assume — at the outset — that $\mathbf{U}^2$ is a known diagonal matrix consisting of uniqueness variances, and g_m is a scaling constant. In particular, we shall reexpress Equation 11 as

$$\hat{\Sigma}^* = (w_m)\hat{\mathbf{Q}}\,\hat{\Lambda}\,\hat{\mathbf{Q}}' + (1 - w_m)\hat{\mathbf{Q}}\,(\hat{\Lambda} - g_m\mathbf{I}_p + g_m\mathbf{I}_p)\hat{\mathbf{Q}}', \tag{12}$$

this being just an identity for $\hat{\Sigma}^*$. Now the last term in Equation 12 is $\hat{\mathbf{Q}}(\hat{\Lambda} - g_m\mathbf{I}_p + g_m\mathbf{I}_p)\hat{\mathbf{Q}}' = \hat{\mathbf{Q}}(\hat{\Lambda} - g_m\mathbf{I}_p)\hat{\mathbf{Q}}' + g_m\mathbf{I}_p$ whenever $\hat{\mathbf{Q}}$ is square, that is, when all eigenvectors are used in $\hat{\mathbf{Q}}$. (If there are multiple eigenvalues the vectors corresponding to them can be taken as mutually orthogonal.) In turn, we can write

$$\hat{\mathbf{Q}}(\hat{\Lambda} - g_m\mathbf{I}_p)\hat{\mathbf{Q}}' = \hat{\mathbf{Q}}_m(\hat{\Lambda}_m - g_m\mathbf{I}_m)\hat{\mathbf{Q}}'_m + \hat{\mathbf{Q}}_{p-m}(\hat{\Lambda}_{p-m} - g_m\mathbf{I}_{p-m})\hat{\mathbf{Q}}'_{p-m} \tag{13}$$

where

$$\hat{\mathbf{Q}}' = \begin{bmatrix} \hat{\mathbf{Q}}'_m \\ \hat{\mathbf{Q}}'_{p-m} \end{bmatrix} \text{ and } \hat{\Lambda} = \begin{bmatrix} \hat{\Lambda}_m & 0 \\ 0 & \hat{\Lambda}_{p-m} \end{bmatrix}$$

represent a partitioning of $\hat{\mathbf{Q}}$ and $\hat{\Lambda}$ into sets of m and $p - m$ eigenvectors and values, respectively.

The final expression in Equation 13 now involves the so-called rejected eigenvalues; the m factors that remain correspond to the largest eigenvalues of $\hat{\Sigma}^*$, and thus the accepted common factors. If we define $g_m = \hat{g}_m = \{1/(p - m)\}\Sigma_{j=m+1}^{p}\hat{\lambda}_j$, the average of the rejected eigenvalues, the second term in Equation 13 tends to vanish as $n \to \infty$ if the m-factor model holds in the population and $\mathbf{U}^2$ is known (cf. Lawley & Maxwell, 1971). In this case the last term in Equation 13 reflects mere *noise* in the data in the sense that the smallest eigenvalues and corresponding vectors of the rescaled sample matrix will tend to represent only stochastic variation in the sample relative to the $p - m$ smallest population eigenvalues, all of which are equal. If Equation 13

is rewritten to exclude the *noise* term then the covariance estimator in Equation 12 reduces to

$$(14) \qquad \hat{\Sigma}^{*}_{cs\text{-}cfa\text{-}m} = (w_m)\hat{\mathbf{Q}}\,\hat{\Lambda}\,\hat{\mathbf{Q}}' + (1 - w_m)\{\hat{\mathbf{Q}}_m(\hat{\Lambda}_m - \hat{g}_m\mathbf{I}_m)\hat{\mathbf{Q}}'_m + \hat{g}_m\mathbf{I}_p\}$$

which for $0 \leq w_m \leq 1$ generally yields a convex sum of two covariance estimators of Σ^{*}. When the common factor model does not hold exactly then it would be desirable for w_m to exceed zero *to the extent that the m-factor model is inappropriate.*

At this point it would be convenient to be able to demonstrate an analytical solution for w_m in Equation 14, the analog of w_0 in Equation 8. However, this problem has proven difficult. Reasoning by analogy with the Case I solution, however, the application of minimum risk principles to Case II yields a procedure that is theoretically attractive and which has worked well in numerous trials. We present this mechanism next.

By modifying the previous terms r_0 and γ_0, one can define the scalar w_m so as to achieve the desired type of convex sum weighting in Equation 14. The basic argument to be presented ensues from the discussion following Equation 9.

The major new requirement for the Case II solution is that instead of assuming that all eigenvalues of $\hat{\Sigma}^{*}$ approach the same limit as $n \rightarrow \infty$ when the statement $\Sigma^{*} = g_0\mathbf{I}$ is true, we instead assume that only the $p - m$ smallest eigenvalues of the properly scaled covariance matrix approach the same value $\hat{g}_m$ as $n \rightarrow \infty$ when the more liberal statement $\Sigma^{*} = \mathbf{F}^{*}\mathbf{F}^{*\prime} + g_m\mathbf{I}$ is true. (The notation $\mathbf{F}^{*}$ is used to be consistent with the * used to symbolize the rescaled covariance matrix.) When the population *cfa-m* model is true (and the uniqueness diagonal known) then the smallest $p - m$ eigenvalues of interest will equal one another (Lawley & Maxwell, 1971). Consequently, by setting $\hat{r}_m = (\Sigma^{p}_{j=m+1}\hat{\lambda}_j)^2/\Sigma^{p}_{j=m+1}\hat{\lambda}_j^2$, and letting $q = p - m$, we see that $\hat{r}_m \rightarrow q$ from below as the *rejected* eigenvalues approach their common value $\hat{g}_m$. Given that m is chosen to be *relatively small*, for what may be termed a parsimonious structural model, the index $\hat{r}_m$ varies between the limits 1 and $q = p - m$ and indicates whether the sample matrix $\hat{\Sigma}^{*}_{cs\text{-}cfa\text{-}m}$ lies nearer some $\Sigma^{*}_{unknown}$ (index = 1) versus the relatively parsimonious form $\Sigma^{*} = \mathbf{F}_m^{*}\mathbf{F}_m^{*\prime} + g_m\mathbf{I}$ (index = q). Using the representations $\hat{\gamma}_m = \{p(1 + \hat{r}_m - 2\}/(p - \hat{r}_m)$, and $w_m = n/(n + \hat{\gamma}_m)$, the index w_m can be described as a badness of fit index on a scale from zero to unity, representing how poorly $\hat{\Sigma}^{*}_{cs\text{-}cfa\text{-}m}$ in Equation 14 fits the common factor form $\mathbf{F}_m^{*}\mathbf{F}_m^{*\prime} + g_m\mathbf{I}$. The complementary term $1 - w_m$ represents a goodness of fit index.

Although these specifications for γ_m and w_m are consistent with those for Case I, it is clear that the above arguments do not constitute an analytical proof.

We invite readers to provide analytic solutions for γ_m and w_m using this or other appropriate rationale.

Given either of the covariance estimators, Equations 8 or 14, one can construct *unscaled* covariance representations of the population Σ in any arbitrary metric. Thus, from Equation 8, we see that for the Case I solution

$$\hat{\Sigma}_{cs\text{-}w_0} = \mathbf{D}\hat{\mathbf{Q}}\{w_0\hat{\Lambda} + (1 - w_0)\hat{g}_0\mathbf{I}_p\}\hat{\mathbf{Q}}'\mathbf{D} \tag{15}$$

or

$$\hat{\Sigma}_{cs\text{-}w_0} = w_0\hat{\Sigma} + (1 - w_0)\hat{g}_0\mathbf{D}^2 \tag{16}$$

where $\hat{\Sigma} = \mathbf{D}\hat{\mathbf{Q}}\hat{\Lambda}\hat{\mathbf{Q}}'\mathbf{D}$.

From Equation 14 we have the parallel forms for Case II,

$$\hat{\Sigma}_{cs\text{-}cfa\text{-}m} = w_m\mathbf{D}\hat{\mathbf{Q}}\hat{\Lambda}\hat{\mathbf{Q}}'\mathbf{D} + (1 - w_m)\mathbf{D}\{\hat{\mathbf{Q}}_m(\hat{\Lambda}_m - \hat{g}_m\mathbf{I}_m)\hat{\mathbf{Q}}'_m + \hat{g}_m\mathbf{I}_p\}\mathbf{D}, \tag{17}$$

or its equivalent,

$$\hat{\Sigma}_{cs\text{-}cfa\text{-}m} = w_m\hat{\Sigma} + (1 - w_m)(\hat{\mathbf{F}}_m\hat{\mathbf{F}}'_m + \hat{g}_m\mathbf{D}^2), \tag{18}$$

where $\hat{\mathbf{F}}_m = \mathbf{D}\hat{\mathbf{Q}}_m(\hat{\Lambda}_m - \hat{g}_m\mathbf{I}_m)^{1/2}$, a standard form for the matrix of common factor coefficients. Another way to represent the latter form is to write

$$\hat{\Sigma}_{cs\text{-}cfa\text{-}m} = (\hat{\mathbf{F}}_m\hat{\mathbf{F}}'_m + \hat{g}_m\mathbf{D}^2) + w_m\{\hat{\Sigma} - (\hat{\mathbf{F}}_m\hat{\mathbf{F}}'_m + \hat{g}_m\mathbf{D}^2)\} \tag{19}$$

The latter expression shows explicitly that $\hat{\Sigma}_{cs\text{-}cfa\text{-}m}$ can be thought of as the usual reduced rank covariance estimator to which part of the residual covariance matrix is added, according to the weight w_m.

Note that if $m = 0$ the representation in Equation 18 includes the expression of Equation 16 as a special case. Consequently, we shall ignore the less general Equations 15 and 16 in what follows. It should be clear, however, that we will do well to continue to make conceptual distinctions between the cases $m = 0$ and $m > 0$. Equations 18 and 19 define a covariance estimator of exactly the same form as the structural Bayesian estimator given by Chen (1979, p. 236). However, our methods for estimating the convex sum weighting coefficient differ in that our system is non-iterative since we use available eigenvalues directly, given $\hat{\Sigma}^*$, whereas Chen's solution requires maximum likelihood estimation and an iterative process based on the EM algorithm. Of course we have required — so far — that $\mathbf{D}^2$ is known apriori, an unrealistic assumption in practice. However, this assumption will be relaxed later. Regardless of the

value of m, Equation 18 combined with the general Case II method for estimating w_m, provides a closed-form solution for an estimator of the population covariance matrix, an estimator that will serve as a basis for regression estimation.

Covariance Estimation as a Basis for Weighted Structural Regression

Given that the basic procedure for generating covariance estimators of the general form in Equation 18 has been developed, $\hat{\Sigma}_{cs\text{-}cfa\text{-}m}$ can be used as a basis for computing a vector of structural regression coefficients. In particular, if the covariance estimator $\hat{\Sigma}_{cs\text{-}cfa\text{-}m}$ is partitioned as in Equation 2, then

$$\hat{\Sigma}_{cs\text{-}cfa\text{-}m} = \begin{bmatrix} \hat{\sigma}_{yy\text{-}cfa\text{-}m} & \hat{\sigma}_{yx\text{-}cfa\text{-}m} \\ \hat{\sigma}_{xy\text{-}cfa\text{-}m} & \hat{\Sigma}_{xx\text{-}cfa\text{-}m} \end{bmatrix}, \tag{20}$$

from which it is straightforward to obtain structural regression estimators of the form of Equation 3. (Again, to simplify notation, but without loss of generality, the first variable is designated as the criterion, to be predicted from the remaining variables.)

If minimum risk principles are used to generate $w_m = \{n/(n + \hat{\gamma}_m)\}$, then, as in Equation 3, regression weights can be obtained using

$$\hat{\beta}_{WSR\text{-}cfa\text{-}m} = \hat{\Sigma}^{-1}_{xx\text{-}cfa\text{-}m}\hat{\sigma}_{xy\text{-}cfa\text{-}m} \tag{21}$$

where the subscript *WSR-cfa-m* refers to weighted structural regression for *cfa* models with m factors. Other principles could of course be used to estimate w in the context of further varieties of weighted structural regression.

Consistent with Equation 5 above, some algebraic manipulation based on a $(1 \mid k)$ row partitioning of the matrix $\hat{\mathbf{F}}_m$, as in Equation 18 above, reveals a representation equivalent to Equation 21 as

$$\hat{\beta}_{WSR\text{-}cfa\text{-}m} = \{n\hat{\Sigma}_{xx\text{-}cfa\text{-}m} + \hat{\gamma}_m(\hat{\mathbf{F}}_x\hat{\mathbf{F}}'_x + \hat{\mathbf{U}}_{xx}{}^2)\}^{-1}(n\,\hat{\sigma}_{xy\text{-}cfa\text{-}m} + \hat{\gamma}_m\hat{\mathbf{F}}_x\hat{\mathbf{f}}_y). \tag{22}$$

This makes explicit the ridge form, since when $m = 0$, the terms $\hat{\mathbf{F}}_x\hat{\mathbf{F}}'_x$ and $\hat{\mathbf{F}}_x\hat{\mathbf{f}}_y$ drop out, and $\hat{\gamma}_m$ becomes a scalar multiplier for the diagonal matrix $\hat{\mathbf{U}}_{xx}{}^2$.

The essential assumption for Case II is that the population Σ can be approximated using a (low rank) common factor structure, an assumption that must be distinguished from one that says the population Σ has exact common factor form and the sample is intended to approximate this structure. Equivalently, since for a sufficient number of factors any Σ can be taken to have

common factor form (although generally, for a non-identifiable *cfa* model), one could say that the use of a *low rank cfa model* for Case II involves specifying *too few* factors, or equivalently, *too few model parameters*; but as we shall see in the forthcoming numerical study, *too few* parameters for one purpose may be a reasonable number for another purpose.

Examination of Equation 18 shows that as $\hat{\gamma}_m$ gets larger and larger in relation to n, then w_m approaches zero, resulting ultimately in a covariance estimator $\hat{\Sigma}_{cs\text{-}cfa\text{-}m}$ that derives wholly from the m-factor *cfa* representation. That is, for a specified value of $m < p$, say $\hat{m}$, $\hat{\Sigma}_{cs\text{-}cfa\text{-}m}$ in Equation 18 alters $\hat{\Sigma}$ by shrinking the smallest $p - \hat{m}$ eigenvalues of the rescaled covariance matrix toward their average value, $\hat{g}_m$, leaving the largest eigenvalues unchanged. The eigenvalues of interest can be seen to have the general form

$$\hat{\lambda}_j = \left\{ \begin{array}{ll} \hat{\lambda}_j & \text{for } j = 1, \ldots, m \\ \dfrac{n\hat{\lambda}_j + \hat{\gamma}_m \hat{g}_m}{n + \hat{\gamma}_m} & \text{for } j = m+1, \ldots, p \end{array} \right\} .$$

If Equations 18 and 21 are used for regression estimation, but the m-factor model is poorly supported by extant data, then one sees that $\hat{\Sigma}_{cs\text{-}cfa\text{-}m}$ converges toward the standard model-free estimator, $\hat{\Sigma}$, as $w_m \rightarrow 1.0$. Also, noting that the *cfa* model in this context necessarily fits any Σ perfectly for $m = p - 1$, we see that as $m \rightarrow p - 1$, the term $\hat{\mathbf{F}}_m \hat{\mathbf{F}}'_m + \hat{g}_m \mathbf{D}^2$ necessarily approaches $\hat{\Sigma}$, so that the convex sum in Equation 18 necessarily approaches $\hat{\Sigma}$. In either case, $w_m \rightarrow 1.0$, or $m \rightarrow p - 1$, it is thus clear that *WSR* estimators will tend toward standard OLS estimators (assuming the appropriate regular inverses exist).

Theoretical work suggests that *WSR* results based on minimizing risk will typically, though not necessarily, resemble those of OLS when sample sizes are *sufficiently large*. This is because the procedures for estimating γ_m from data tend to ensure that for fixed p, as n increases $\hat{\gamma}_m$ will not increase systematically and so $w_m = n/(n + \hat{\gamma}_m) \rightarrow 1.0$ as $n \rightarrow \infty$. Yet, as seems desirable, the new methods generally give an advantage to the investigator whose prior knowledge is greatest in situations where sample sizes are limited, or when p/n is relatively large.

For any covariance structure model that might be used with real data, not just common factor forms, the weighted structural regression problem can be described as that of choosing a parsimonious structural model that is likely to be supported by data. When the model is supported by extant data, in the sense that $w \rightarrow 0$, then covariance estimators based on the model will tend to derive from parameter estimates associated with the model. In our final section we discuss such matters further.

Provided that the estimator of w, the weighting coefficient, is appropriately generated, and conceiving of the convex sum estimator in Equation 18 as a counterpart of the mode of Chen's (1979, p. 236) derived posterior distribution for $\hat{\Sigma}$, we can write a general expression for the sampling distribution of sample *WSR* coefficients. Although his theory was more general, and was derived explicitly from Bayesian arguments, Chen's theorem 5.1 (p. 241) and his subsequent analysis can be used directly to secure the case for saying $\hat{\beta}_{WSR\text{-}cfa\text{-}m}$ has a multivariate t-distribution. Using Chen's analysis, the covariance matrix of the vector of regression estimators can be estimated as

$$\widehat{Cov}(\hat{\beta}_{WSR\text{-}cfa\text{-}m}) = \{\hat{h}^{o}/(n + \hat{\gamma}_m)\}\hat{\Sigma}^{-1}_{xx\text{-}cfa\text{-}m}, \tag{23}$$

where $\hat{h}^{o}$ denotes the reciprocal of the first diagonal term in the inverse of the joint covariance estimator, $\hat{\Sigma}_{cs\text{-}cfa\text{-}m}$. Thus for any element in $\hat{\beta}_{WSR\text{-}cfa\text{-}m}$, say $\hat{b}_{WSR\text{-}cfa\text{-}j}$, the estimated standard error can be computed as

$$\widehat{s.e.}(\hat{b}_{WSR\text{-}cfa\text{-}j}) = diag[\{\hat{h}^{o}/(n + \hat{\gamma}_m)\}\hat{\Sigma}^{-1}_{xx\text{-}cfa\text{-}m}]^{1/2}_{jj}, \tag{24}$$

the square root of the j^{th} diagonal element of the matrix in braces.

The sampling theory underlying these expressions is based on the same multivariate normal assumptions as given previously. The theory associated with the construction of $\hat{\Sigma}_{cs\text{-}cfa\text{-}m}$ should be regarded as approximate of course, particularly since in the general case, where $m > 0$, the analytical solution for w_m is not available. Nevertheless, several bootstrap trials with a number of data sets lend support to the validity of these equations. We shall present some results below that are suggestive of the usefulness of Equation 24.

The choice of m and the relative sizes of n and γ are clearly pivotal; for different values of m and γ or w, *WSR* can be seen to include OLS, a general form of ridge regression, as well as reduced rank regression as special cases. In some instances of course we are suggesting that the scalar w is prespecified, as say zero or unity, but we have concentrated on using WSR principles for estimating w where the structural model used has common factor form. Recent literature of covariance structure estimation implies that other choices for w may be reasonable to consider. Indeed, we use and discuss one other form in our subsequent numerical studies. In the next section we present a method for scaling the covariance estimator, with a focus on small sample applications.

Scaling the Joint Predictor-Criterion Covariance Matrix and Choosing a Common Factor Method

An important role of covariance matrix rescaling in this context is that when based on appropriate uniqueness variance estimators, covariance matrices

of the form of Equation 18 can be used to make the structural analysis *scale-free*. This means that *WSR* estimators that use uniqueness rescaling can, like their common factor counterparts, be described as invariant with respect to arbitrary linear rescalings of the original covariance matrix. It is interesting to note that nearly all previously developed non-OLS forms of multiple regression are generally not invariant to changes in predictor rescalings (cf. Smith & Campbell, 1980, as well as the published commentaries).

Given the foregoing correspondences between common factor analysis and regression it is natural to consider linkages between these methods and Guttman's (1953) image analysis. Further, noting the role of rescaling in this framework, Harris' (1962) developments of Rao-Guttman relationships are also of special interest. Given these articles, however, the key to making further progress seemed to lie in finding an effective way to rescale the joint predictor-criterion covariance matrix when sample size is relatively small in relation to p.

Muirhead (1985), starting from the same multinormality assumption used earlier, uses risk minimization principles to produce specialized procedures for improving sample estimates of the ratio $M = R^2/(1 - R^2)$, the ratio of the population squared multiple correlation coefficient to its complement. This ratio is essentially a correlation-based signal-to-noise ratio that can be generated for each variable separately in the context of a system of stochastic or random variables. Specifically, Muirhead (1985, p. 924, Equation 8) presented an estimator of this ratio of the form

$$M^* = a_1\hat{M} - a_2, \tag{25}$$

for $\hat{M} = smc/(1 - smc)$, $a_1 = (n - p - 4)/(n + 1)$ and $a_2 = (p - 1)(n - p - 4)/(n + 1)(n - p - 2)$, where *smc* is the sample estimate of R^2. He showed that M^* is optimal in the sense that no other linear estimator of M dominates this one in terms of squared error loss. Since M^* could be negative for some values of a_1 and a_2, even though population values of M cannot be negative, Muirhead pointed out that negative estimates of M^* should be reset to zero.

Because Muirhead's (1985) linear estimator can be used to *correct* estimates of various different functions of *smc*, consistent with minimizing a tractable form of squared quadratic loss, this estimator provides an appealing basis in the present context for constructing an estimate of the diagonal rescaling matrix $\mathbf{D}^2$. There is a well-established basis in the literature of factor analysis (cf. Guttman, 1956; Jöreskog, 1969) for using complements of sample *smc*s as a basis for estimating uniqueness variances. However, when samples are relatively small, or the ratio p/n is relatively large, then the well-known bias of sample *smc*s appears to require correction.

Using Muirhead's (1985) estimator Equation 25 in the context of image factor analysis, the uniqueness diagonal $\mathbf{U}^2$ can be estimated as a function of the sample *smc*s. Specifically, algebraic manipulation of M^* yields an alternative estimator of $\mathbf{U}^2$ consistent with Muirhead's Equation 25; it is

$$\hat{\mathbf{S}}^{*2} = [a^1\mathbf{D}\{smc/(1 - smc)\} + (1 - a_2)\mathbf{I}]^{-1}. \tag{26}$$

where $\mathbf{D}\{smc/(1 - smc)\}$ is a diagonal matrix whose non-zero entries consist of each variate's sample squared multiple correlation with all other variates in the set, divided by its complement. The a_1 and a_2 are the same as given following Equation 25. The rescaling matrix for $\mathbf{R}_S$, the sample correlation matrix, is thus $\hat{\mathbf{S}}^{*-1} = [a_1\mathbf{D}\{smc/(1 - smc)\} + \{1 - a_2\}\mathbf{I}]^{1/2}$. Following Harris (1962), as well as Jöreskog (1969), $\hat{\mathbf{S}}^{*-1}\mathbf{R}_S\hat{\mathbf{S}}^{*-1}$ can be used for eigenanalysis. Since for fixed p, $a_1 \rightarrow 1$ and $a_2 \rightarrow 0$ as $n \rightarrow \infty$, this expression converges to that of Harris' $\hat{\mathbf{S}}^{-1}\mathbf{R}_S\hat{\mathbf{S}}^{-1}$ as the ratio $n/p \rightarrow \infty$. Consistent with Muirhead's recommendations concerning negative values of the estimate M^*, when estimates in Equation 26 are smaller than unity, they should be set to unity.

Given the foregoing, it appears that an effective common factor method for use in many regression applications, and perhaps for other small sample applications as well, is to start from the sample correlation matrix $\mathbf{R}_S$ and compute $\hat{\mathbf{S}}^{*-1}\mathbf{R}_S\hat{\mathbf{S}}^{*-1}$ from Equation 26, then generate the m largest eigenvalues $\hat{\Lambda}_m$ and corresponding eigenvectors $\hat{\mathbf{Q}}_m$ of this matrix. Thus, the (untransformed) matrix of common factor coefficients can be written as

$$\hat{\mathbf{F}}_{cfa\text{-}m} = \hat{\mathbf{S}}^{*}\hat{\mathbf{Q}}_m(\hat{\Lambda}_m - \hat{g}_m\mathbf{I}_m)^{1/2}, \tag{27}$$

a form that can be used in Equation 18 to provide a basis for *WSR* estimation. As $n/p \rightarrow \infty$ the expression Equation 27 converges to the factor coefficients matrix given by Jöreskog (1969) in his version of image factor analysis. This is the form of cfa that is employed numerically with four different regression methods in the next section.

Since Jöreskog (1969) has shown image factor analysis to be scale free, this feature is retained for Equation 27 as long as all estimates in Equation 26 exceed unity. If the latter condition prevails, the row-rescaled form $\hat{\mathbf{F}}_{\hat{\Sigma}\text{-}cfa\text{-}m} = \mathbf{D}_s\hat{\mathbf{F}}_{cfa\text{-}m}$, for $\mathbf{D}_s$ the diagonal matrix of standard deviations of the original variables, represents in obvious notation the common factor coefficients matrix that would be obtained if the analysis had begun from the sample covariance matrix $\hat{\Sigma}$ instead of being generated from the matrix $\mathbf{R}_S$ of sample correlations. The basic arguments are given in Jöreskog (1969).

The value of scale freeness or scale separability is distinctive. It means that either correlation or covariance metric can be used for *WSR* estimation

whenever scale free methods are used for estimating covariance structures; conversion between standardized and raw score weights thus takes the same form for *WSR* as for OLS weights. Standardized weights are usually found to be more easily interpretable, whereas raw score weights are more conducive to comparison across independent samples with the same variables. Neither of these virtues should be ignored.

In the next section we study the merits and demerits of the various alternative regression estimators with real data.

A Numerical Study of Alternative WSR Estimators

In this section we present and discuss selected results of a bootstrapping study that provides numerical comparisons among six different forms of *WSR* estimation for a selected set of data. The forms of regression analysis used were: OLS, three versions of *WSR* based on minimizing risk, an alternative convex sum procedure, labeled GFI, and reduced rank regression. The GFI procedure was prompted by consideration of one of the goodnees of fit indices recommended by Jöreskog and Sörbom (1986) in the general context of covariance structure analysis. These methods are described below.

Data and Methods

The data set used for bootstrapping and simulation is taken from Gunst and Mason (1980, p. 367) and represents body measurements of $n = 33$ female applicants for positions as police trainees. The criterion measure was defined as applicants' heights, and the predictors were seven body measures such as arm, foot and leg length.

Appendix B contains the product-moment correlation matrix for this set of data, as well as its eigenvalues, and the squared multiple correlation coefficients, for each variable with all seven others.

Since bootstrap procedures (tacitly) take the initial data system as defining a fixed population, its metric can be chosen arbitrarily. In this study all variables were scaled initially to z-score, that is, correlation metric, to make the various bootstrap weight coefficients and standard errors generally comparable with one another. All sample regression coefficients were then computed in so-called *raw* metric to make them comparable across samples.

Two different methods were used for bootstrapping: The first used a conventional bootstrap procedure, starting from the available raw data matrix, as given in Gunst and Mason (1980, p. 363); the second was based on the *normal bootstrap*, where data vectors were simulated to be stochastically

consistent with the Gunst-Mason covariance (i.e., correlation) matrix (cf. Efron & Tibshirani, 1986).

Conventional Bootstrap

In the case of the conventional bootstrap method each bootstrap sample, of size $n = 33$, was generated by sampling randomly with replacement from the original matrix of observation vectors. One hundred such bootstrap samples were obtained, and each of these samples was then used to generate regression coefficients using six different methods, as described below.

Recognizing that most empirical prediction work uses somewhat larger samples, it was decided to conduct a second *bootstrapping* study to examine how the same regression methods would compare using a larger sample size. For this second set of analyses, the same Gunst-Mason prediction problem was utilized on the basis of what Efron and Tibshirani (1986) call the *normal bootstrap.*

Normal Bootstrap

For this procedure, the eight variable Gunst-Mason population covariance (here, correlation) matrix $\mathbf{R}_{pop}$ was used as a starting point to simulate observation vectors using the following procedure: Given $\mathbf{R}_{pop}$, a common factor analysis was employed (using the same image-factor methods as described above) with $m = p - 1$ common factors. This ensured that the derived matrix of factor coefficients $\mathbf{F}_p$, of order $p \times (p - 1)$, would exactly reproduce all off-diagonals of $\mathbf{R}_{pop}$. (Lack of uniqueness of this $\mathbf{F}_p$ is of no consequence here.) From $\mathbf{F}_p$, a diagonal matrix was constructed as $\mathbf{D}_u^2 = \mathbf{R}_{pop} - \mathbf{F}_p\mathbf{F}'_p$. Given these population-based matrices, 100 normal bootstrap samples were created, each of size $n = 100$, using the construction $\mathbf{X}_s = \mathbf{X}_c\mathbf{F}'_p + \mathbf{X}_u\mathbf{D}_u$. Entries in the matrices $\mathbf{X}_c$ and $\mathbf{X}_u$, of order $n \times m$ and $n \times p$ respectively, were computer-generated using a pseudo random normal generator, that is, $x_{ij} \sim iidN(0,1)$. Use of this process ensured that simulated *observation* vectors associated with each sample data matrix $\mathbf{X}_s$ would follow a multivariate normal distribution, that is, $\text{Row}(\mathbf{X}_s)_i \sim MN(0,\mathbf{R}_{pop})$.

For each combination of an estimation method and a bootstrap sample regression weights were compared with their full sample OLS counterparts as described in the first footnote in Table 1. Three evaluative criteria were used to compare regression methods. The first two were: average mean squared errors of sample versus population weights, computed as $\text{MSE} = \text{Avg}(\hat{\beta}_j - \beta_{ols})^2$; and averages of estimable parts of predictive mean squared errors, computed as $\text{EPMSE} = \text{Avg}(\hat{\beta}_j - \beta_{ols})'\mathbf{R}_{pop}(\hat{\beta}_j - \beta_{ols})$. Here $\hat{\beta}_j$ represents the weights for each

combination of bootstrap sample and method, β_{ols} denotes the full sample (bootstrap population) regression weights, and $\mathbf{R}_{pop}$ references the population covariance (here, correlation) matrix. MSE is a conventional measure of how well sample weights estimate their population counterparts. The logic of the PMSE is that repeated use of $\hat{y}$s from samples in place of $\hat{y}$ from the population to predict y will result in a mean squared prediction error, per prediction, of σ^2 + E(EPMSE) where the latter term is the expected value of EPMSE and σ^2 represents the residual population variance (Browne, 1975).

A third criterion was computed to assess the quality of the criterion score estimators associated with each regression method. For this criterion, each set of original bootstrap sample weights were used for *bootstrap cross-validation*. For each sample, initial bootstrap weights were applied to the predictor variates in the corresponding bootstrap *holdout* sample to compute predicted y scores, that is, $\hat{y}$s. Each holdout sample consisted of the rows not selected with replacement for the corresponding bootstrap sample; on average approximately 63% of the rows of the bootstrap population are used in the bootstrap sample leaving 37% in the holdout sample (cf. Efron, 1983). Thus, average squared differences of the form Avg$(y - \hat{y})^2$ were computed for each method for every sample and then averaged across bootstrap samples. This method apparently has not often been used in the past.

For the normal bootstrap study, the same evaluative criteria and methods were employed as described above. However, to obtain the cross-validation evaluative indices Avg$(y - \hat{y})^2$ for the normal bootstrap, an additional independent random sample (of size $2 \times p = 16$) was constructed to parallel each normal bootstrap sample; these were used for cross-validation purposes. Thus, in the case of the second bootstrapping experiment, Avg$(y - \hat{y})^2$ was computed across 100 cross-validations based on these independent samples.

The methods used to compute weights for each bootstrap sample were the following:

Method 1: OLS

Ordinary Least Squares regression, where the (raw score) weights were computed using an expression of the form of Equation 3 for each bootstrap sample;

Method 2: WSR-RDG

The WSR version of ridge regression, obtained using expression Equation 16, with m set at zero, and Equation 21 used to generate weights. Use of this method corresponds to choosing a minimum risk estimate of the population

covariance matrix and using this covariance matrix to provide a closed-form solution for the corresponding regression weights in the context of *inverse-based* rescaling; thus, like all methods used here, WSR-RDG is properly described as scale free;

Method 3: WSR-MR

The third method may be declared to be the *conventional* cfa version of WSR based on minimizing risk. This method uses expression Equation 18 with $w_m = n/(n + \hat{\gamma}_m)$ to generate the covariance estimator with the common factor method based on Equation 27, for $m = 1, 2, 3$ and 4. Equation 21 was used in each case to generate regression weights; this, and the two methods below can be described as convex sum methods since for them w is generally *intermediate* between zero and unity unlike OLS and RedRk;

Method 4: WSR-MR2

This method used a modified WSR-MR method for which $w_m = n/(n + 2\hat{\gamma}_m)$ in the convex sum Equation 18, but is otherwise equivalent to Method 3. This choice was motivated by examination of the upper limit of equation 10, the use of which yields a sample covariance estimator that should generally be no worse than the conventional mle estimator, and generally better to the extent to which the m-factor model is *reasonable* for the population;

Method 5: GFI

This method was based on the choice $1 - w_{js} = gfi = 1 - (d_1/d_2)\{\mathrm{tr}(\hat{\Sigma}^{-1}_{cs\text{-}cfa\text{-}m}\hat{\Sigma} - \mathbf{I})^2/\mathrm{tr}(\hat{\Sigma}^{-1}_{cs\text{-}cfa\text{-}m}\hat{\Sigma})^2\}$ where (d_1/d_2), a correction term based on degrees of freedom, is $p(p + 1)/\{(p - m)^2 + p - m + 2\}$ in this context (cf. Jöreskog & Sörbom, 1986, p. I.40, exp. AGFI); $\hat{\Sigma}_{cs\text{-}cfa\text{-}m}$ represents the model-based cfa estimate derived from Equations 18 and 27, and $\hat{\Sigma}$ is the model-free covariance estimator. Thus $w_{js} = (d_1/d_2)\{\mathrm{tr}(\hat{\Sigma}^{-1}_{cs\text{-}cfa\text{-}m}\hat{\Sigma} - \mathbf{I})^2/\mathrm{tr}(\hat{\Sigma}^{-1}_{cs\text{-}cfa\text{-}m}\hat{\Sigma})^2\}$. Although w_{js} has a general upper bound of unity, theoretically it can attain negative values for some covariance structure models — although this eventuality will virtually never occur with *sufficiently small* values of the ratio m/p;

Method 6: RedRk

This is the regression method resulting from setting the coefficient $w = 0$ in Equation 18 and using $m = 1, 2, 3$ or 4, to generate the population covariance estimator which was then used to compute weights as in Equation 21.

Although this method is *new* in the sense that it was based on the common factor coefficients defined in Equation 27, it is of the same form as that given in Lawley and Maxwell (1973).

Bootstrap Results

Results for the data set are summarized in Tables 1 through 4, where attention is focused on prediction of the first variate from all others. We recall however that these methods are all symmetric in that the same methods could be (and were in fact) used to summarize regression results for each variate, as if it were the criterion variable with all others as predictors. Subject-matter interpretations are ignored here although the reader is encouraged to examine the original source for these details.

Table 1 (next page) presents selected sets of average values and standard deviations of the weights for each set of bootstrap samples. The table also provides in the last four rows the full sample counterparts of the bootstrap values for two of the methods, OLS and WSR-MR2-2 (WSR-MR2 with m = 2), the *standard* method and the *generally best* cfa-based WSR method; equation 24 was used to estimate standard errors for the weighted structural regression estimators. Table 2 presents the MSE, PMSE and cross-validation badness of fit indices for the data set.

From Table 1, in which bootstrap results are presented only for the case m = 2 for the cfa-based methods, it can be seen that weights obtained from the alternative regression methods are similar to one another, and that the standard errors tended to be largest for OLS and smallest for the reduced rank method. It is also notable that the two vectors of weights computed from the full sample are generally comparable with their bootstrap counterparts above in rows one and four. Comparing bootstrap results with those computed by formula for the bootstrap population, standard errors for both methods are somewhat less similar than are means, but this is to be expected for *secondary* as distinct from *primary* statistics. Non-OLS weights appear considerably less erratic across bootstrap samples than those of the OLS method. Also, the evidence suggests that bootstrap standard errors are slightly larger than those computed by formula for OLS from full samples.

It is interesting to note that when using OLS methods, only variables 2 and 7 appear to be the established *significant* predictor variables in the sense that their average bootstrap weights divided by their estimated standard errors exceeded 2.5 (with reference to conventional t-statistic standards). However, using, say the MinRisk Ridge method, or the MR2-2 method, two additional variables move into the picture in terms of nominally significant t's.

Table 1
Bootstrap Averages[a] and Standard Errors of Regression Weights Results for Gunst-Mason Data, for m = 2, CFA-based Method Body Measurement Variables, n = 33, p = 8

	Averages for 100 Samples Independent Variable Number (Criterion is Variable 1)						
Method	2	3	4	5	6	7	8
OLS	.431	-.091	.066	.148	.126	.484	.138
WSR-RDG	.376	-.027	.142	.115	.163	.342	.149
WSR-MR	.385	-.038	.121	.109	.143	.411	.144
WSR-MR2	.362	-.013	.139	.098	.140	.387	.144
GFI	.349	.001	.148	.093	.136	.376	.144
RedRk	.268	.068	.192	.087	.097	.333	.151

	Standard Errors for 100 Samples Independent Variable Number (Criterion is Variable 1)							
Method	2	3	4	5	6	7	8	row avg
OLS	.071	.125	.118	.128	.163	.140	.106	.122
WSR-RDG	.048	.078	.062	.072	.090	.060	.065	.068
WSR-MR	.064	.078	.070	.081	.111	.086	.077	.081
WSR-MR2	.066	.062	.059	.068	.096	.076	.070	.071
GFI	.069	.055	.057	.063	.090	.074	.068	.068
RedRk	.100	.037	.061	.054	.079	.093	.067	.070

	Weights and Standard Errors for Full Sample[b] Vector of Multiple Regression Weights							
Method	2	3	4	5	6	7	8	smc's[c]
OLS	.439	-.084	.103	.157	.114	.464	.119	.892
WSR-MR2-2	.368	.004	.152	.101	.120	.384	.138	.886
	Corresponding Multiple Regression Standard Errors							
OLS	.074	.105	.110	.101	.131	.141	.087	
WSR-MR2-2	.058	.078	.074	.070	.085	.095	.061	

Note. Legend for Methods -- OLS is Ord. Least Squares, WSR-RDG is Min. Risk Ridge, WSR-MR is Min. Risk Regression, WSR-MR2 is Mod. Min. Risk, GFI is Weighted Str. Regression for GFI from Jöreskog & Sörbom, 1986, RedRk is Red. Rank Reg.

[a] Each average based on 100 bootstrap samples drawn with replacement from full sample; original data (Gunst & Mason, 1980, p. 367) were transformed to z-score metric to facilitate comparisons. [b] Last four rows present formula-based values of the vector of βs and corresponding standard errors computed from full sample using two Methods: OLS and WSR-MR2 with m = 2. [c] The first smc is the uncorrected squared multiple correlation associated with the OLS vector to its left; the second smc is the analog of the OLS smc, computed using the vector to its left.

Table 2 (next page) presents results for four different values of m in the case of the cfa-based WSR methods and shows that for these data the *best* result for each m is in most cases a convex sum method, regardless of which badness of fit index is used for judgment. The best method in recovering full sample, that is, *population* (OLS) weights, tended to be WSR-MR2. Study of the MSE's in Table 2 suggests that despite the *bias* of the non-OLS WSR weights in Table 1 the reductions in sampling variability more than compensated for the apparent bias. The ridge method, WSR-RDG, worked well in terms of prediction mean squared error yielding fit statistics that were generally comparable to the best cfa-based convex sum method. All the non-OLS methods produced relatively low cross-validation squared errors. The RedRk method is so poor as to be wholly unsatisfactory for $m = 1$, but for every fit criterion it achieves an advantage over OLS for $m = 3$ and 4.

Although not presented here, standard deviations of the MSE's, EPMSE's, and Cross-Validation Mean Squared Errors were, like the means, smaller for non-OLS regression methods, generally 30 to 70% smaller in the case of the conventional bootstrapping exercise, and 10 to 30% smaller in the case of the larger n for the normal bootstrap studies. Also not presented, but clear from the summary statistics, was the finding that RedRk regression weights were on average distinctly different from the OLS bootstrap population weights, particularly for small m. This of course suggests a risk of possibly substantial bias for RedRk weights.

Table 3 is of the same form as Table 2 except the entries in Table 3 are averages based on 100 normal bootstrap samples, each of size $n = 100$. Again, four different choices were used for m, the number of common factors in the model-based covariance estimator in Equation 18.

It can be seen from Table 3 that regression results based on the ridge method and the convex sum covariance estimators, MR2 especially, are again generally better than those based on OLS regression. However, with the larger sample size of $n = 100$ sample data is routinely able to support the estimation of more common factors with the consequence that the best results tend to be associated with the largest values of m, specifically, $m = 3$ or 4. The smallest mean EPMSE's amd MSE's are notably smaller than their OLS counterparts, as much as 40% smaller for MSE's, 25% smaller for EPMSE's. The indices of criterion score fit, Cross-Validation Mean Squared Error, are less sensitive indicators of differences among the regression methods, although these too favor the same methods that yield the smallest EPMSE's, especially MR2. The GFI method appears to work somewhat better for $n = 100$ than for $n = 33$, but only for the larger values of m; even the MR2 method does poorly with $m = 1$ and 2, depending on the evaluative criterion examined. The RedRk method worked very poorly for $m = 1$ and 2, but improved to the point of being almost

Table 2
Bootstrap Results for Gunst-Mason (1980) Data, Body Measurement Variables, $n = 33$, $p = 8$.

Summary Statistics for Six Regression Coefficient Estimators[a]								
Method[c]	Mean Squared Errors βs				Pred. Mean Squared Errors [b]			
OLS	.107	.110	.111	.088	.034	.034	.033	.030
WSR-RDG	.063	.061	.061	.051	.023	.023	.023	.019
	m=1	*m*=2	*m*=3	*m*=4	*m*=1	*m*=2	*m*=3	*m*=4
WSR-MR	.057	.059	.060	.050	.026	.023	.022	.020
WSR-MR2	.066	.059	.055	.045	.036	.025	.022	.019
GFI	.083	.063	.057	.051	.051	.028	.023	.021
RedRk	.266	.120	.084	.056	.193	.063	.034	.022
Method	Cross-Valid. Mean Squared Errors[d]							
OLS	.019	.019	.029	.018				
WSR-RDG	.015	.015	.024	.014				
	m=1	*m*=2	*m*=3	*m*=4				
WSR-MR	.016	.016	.025	.016				
WSR-MR2	.016	.016	.024	.015				
GFI	.017	.016	.024	.015				
RedRk	.029	.018	.024	.014				

[a] Each MSE and corresponding table entry is based on 100 conventional bootstrap samples drawn with replacement as described in the test; MSE represents the average of squared differences between sample bootstrap weights and the full sample OLS weights. [b] Predictive mean squared errors are actually estimable parts of predictive mean squared errors $(y - \hat{y})^2$ (all would be incremented by a constant to equal actual PMSEs, cf. Browne, 1975). [c] Because each set of bootstrap runs was completed for a specified value of *m*, OLS and Ridge statistics are duplicated for each *m*; *m* is not relevant to these two statistics, but the variation in averages indiates the degree of variability in bootstrap means. [d] All cross validation averages are based on bootstrap samples versus their holdout counterparts as described in the text.

Table 3
Bootstrap Results for Funst-Mason (1980) Data, Body Measurement Variables, $n = 100, p = 8$

Summary Statistics for Six Regression Coefficient Estimators[a]

Method	Mean Squared Errors βs				Pred. Mean Squared Errors			
OLS	.025	.027	.026	.027	.010	.012	.011	.012
WSR-RDG	.020	.021	.020	.022	.009	.010	.010	.011
	m=1	m=2	m=3	m=4	m=1	m=2	m=3	m=4
WSR-MR	.020	.020	.017	.019	.010	.010	.009	.010
WSR-MR2	.020	.018	.016	.017	.012	.010	.008	.009
GFI	.047	.027	.018	.016	.041	.018	.009	.009
RedRk	.154	.052	.021	.017	.160	.037	.011	.009

Method	Cross-Valid. Mean Squared Errors			
OLS	.015	.015	.014	.015
WSR-RDG	.015	.015	.014	.015
	m=1	m=2	m=3	m=4
WSR-MR	.015	.015	.014	.015
WSR-MR2	.016	.015	.014	.014
GFI	.018	.016	.014	.014
RedRk	.029	.018	.014	.014

[a] Each MSE and corresponding entry in this table is based on 100 normal bootstrap samples as described in the text.

as good as the best convex sum methods for $m = 4$. The ridge method continued to be superior to OLS, but did not fare as well as the better convex sum methods for the larger value of n.

Comparing the two sets of bootstrap results, Table 2 versus Table 3, some systematic differences can be found. In the case of results for the conventional

bootstrap, where $n = 33$, the relative advantage in terms of both sampling variability and cross-validation accuracy of MR and MR2 methods over OLS regression seems essentially not to depend on the choice of m, the number of common factors, for the cfa method used. The same can be said of these methods for the normal bootstrap, for $n = 100$, only if the cases $m = 1$ or 2 are excluded. Of course a major difference between the two bootstrapping methods is that the first uses the empirical distribution of the given observed data as the starting point; the second imposes the concept of multivariate normal sampling. It is possible that some of the differences observed between Tables 2 and 3 have more to do with this difference than with the differences in n's (although the observed Gunst-Mason data appeared generally *well-behaved*; marginals were generally symmetric, and bivariate plots exhibited no obvious curvilinearity).

Relative invariance to the choice of m seems important since such a finding is suggestive of model robustness. It is most desirable that adaptive procedures generally take advantage of the information in the structural model to the extent to which it is present, to downweight the structural (cfa) model if it doesn't fit well, and weight it more heavily if the data support the structural model. To the extent that this feature characterizes adaptive WSR methods it bodes well in those common situations where the investigator has only vague or imperfect knowledge of what structural model to employ, where the use of regression is largely exploratory in form.

It has been useful to learn that relatively simple structural model-based methods can work effectively to make linear predictions with real data. Compared with mainstream OLS methods, the new regression methods yielded systematically more stable weights, and better independent sample criterion score estimation, for both bootstrapping studies. Despite the fact that target weights for each bootstrap sample were the full sample (bootstrap population) OLS weights for the MSE and PMSE criteria, it is interesting that non-OLS procedures have worked best, often substantially so, in recovering these weights for each problem. Also, although more studies are needed to assess the appropriateness of Equation 24 for estimating weight standard errors for convex sum methods, this expression worked essentially as well for MR, and MR2 as did the theoretically-based expression for OLS weight standard errors.

Table 4 displays averages and standard deviations of the goodness of fit criteria, each of the form $1 - w$, for the methods that were summarized in Tables 2 and 3. As can be seen from Table 4, weights given to the structural model covariance estimator tend to be substantially larger for the GFI method than for the WSR-MR and WSR-MR2 convex sum methods. However, it is evident that a better overall fit, in terms of the goodness of fit index which controls the

Table 4
Goodness of Fit Summary for Bootstrapping Study[a], Gunst-Mason (1980) Data

	n = 33 (conventional bootstrap)					
	Method WSR-MR		Method WSR-MR2		Method GFI	
m	Avg-gfi	(sd)	Avg-gfi	(sd)	Avg-gfi	(sd)
1	.234	(.047)	.379	(.054)	.497	(.078)
2	.283	(.049)	.441	(.054)	.532	(.088)
3	.302	(.084)	.464	(.091)	.557	(.133)
4	.393	(.103)	.564	(.105)	.670	(.163)
	n = 100 (normal bootstrap)					
	Method WSR-MR		Method WSR-MR2		Method GFI	
m	Avg-gfi	(sd)	Avg-gfi	(sd)	Avg-gfi	(sd)
1	.108	(.020)	.196	(.026)	.555	(.056)
2	.177	(.039)	.301	(.047)	.686	(.066)
3	.295	(.066)	.455	(.072)	.848	(.048)
4	.388	(.076)	.559	(.077)	.901	(.037)

[a] Each of the Avg-gfi indices is based on an average across 100 bootstrap samples of a goodness of fit term of the form 1 - w for a particular convex sum form of WSR; the value in parenthesis is the corresponding standard deviation. Each combination of Method and m is associated with a set of statistics in Tables 2 and 3.

weighting in the convex sum for Equation 18, does not necessarily imply an advantage in terms of other fit criteria such as MSE or PMSE. Of course, the WSR-MR2 method yields a goodness of fit index which is always larger, on average, than the WSR-MR method.

In many cfa-based WSR regression applications it would seem reasonable to recommend a *scree plot* of the eigenvalues of the matrix $\hat{\mathbf{S}}^{*-1}\mathbf{R}_S\hat{\mathbf{S}}^{*-1}$, to choose the number of factors for the cfa model, principally on the basis of where the first major break occurs, reading left-to-right in the plot. One might also pay special attention to the residual covariances, or more likely correlations,

between the criterion of interest and the predictors, for each of several values of m. A plot of the sums of squares of such residual correlations against m could also provide evidence of the appropriate number of factors.

Although the procedure was too computationally intensive to be used for all bootstrap samples, Bayesian structural regression (BSR) weights (Chen, 1979) were generated for each of our initial samples and some of the findings can be reported. The main point was that the value of the goodness of model fit, $(1 - w)$ in current notation, exceeded that of even the GFI procedure, which as noted in Table 4, was in every case studied larger than that of either the WSR-MR or modified WSR-MR method. Such findings suggest that the BSR method will generally yield MSE's, PMSE's and Cross-Validation fit indices somewhere between those of the GFI and the reduced rank method. Further evidence on this matter would be useful.

Rabinowitz (1990) recently completed a study that demonstrates similar findings for four other population systems. He employed the same regression methods as were used here. Rabinowitz used normal bootstrap procedures, systematically varying m, number of factors, and n, sample size. He found systematic and distinctive advantages for the adaptive convex sum regression methods, results that essentially paralleled those reported above. For small sample sizes especially, results strongly favored adaptive WSR methods; but even for samples as large as $n = 140$ with 7 to 11 predictors, the OLS results were always systematically worse than the best regression methods based on convex sum procedures.

For those who are interested, we have worked out other examples comparable to the applications above, and also have available the principal results of Rabinowitz (1990). These results and the software to implement adaptive WSR (and bootstrapping) procedures are available from the first author.

Discussion and Conclusions

Despite the rich variety of methods available to support applied linear prediction, it has been a major premise of this work that there are distinctive new ways in which one might approach prediction analysis, some of which have the potential to improve, perhaps substantially, the generalizability of the predictive equations we derive from our data, as well as the interpretability of conclusions.

The principal aim of this article has been to develop a new approach to regression that accounts for variable unreliability and permits an analyst to incorporate psychometric knowledge or information into analysis. As a general class WSR methods have been shown to include OLS, ridge and reduced rank methods as special cases within its framework.

Compared to other forms of regression methodology, adaptive forms of weighted structural regression differ in a fundamental way: they provide a flexible mechanism with which the analyst can advance a prior structural model and incorporate it into analysis; however, for any model selected, these methods use the structural information only to the extent that the prior model is supported by extant data. If data are not consistent with the model the analysis automatically tends to discount it, moving toward the model-free situation, namely, OLS regression.

The specific procedure discussed above for adaptive weighting in the context of Equation 18 assumes a multivariate normal sampling process, however, preliminary sensitivity analyses suggest that adaptive WSR methods are reasonably robust to at least some violations of this assumption. Chen also suggests that his closely related methods may be robust to violations of distributional assumptions and more resistant than OLS methods to outliers. Given the reciprocity relationships inherent in systems that derive from simultaneous prediction of each variable from all others, there also seems to be special potential in this methodology for estimating missing data.

Although we have focused on WSR methods using the assumption that predictors are random or stochastic, many applications of regression include both fixed and random predictors. In those applications where some predictor variables are regarded as fixed and others random, various approaches to *partialing* out the fixed or design variates will generally be warranted when applying adaptive WSR methods.

Many articles and monographs have been written to expound on Bayesian regression estimation, but most (cf. Laughlin, 1979, and Vinod, 1982) have aimed directly at incorporating priors on betas, that is, the weights themselves, rather than on predictor-criterion covariances or correlations. Adaptive WSR methods seem more natural and easier to use since no burden is placed on the investigator to provide prior information about complex *multiple-partial* parameters, the betas. Rather, one need only provide information about predictor-criterion covariances, information that may be either generic or specific.

If an investigator is in a position to advance a parsimonious structural model which in turn is strongly supported by data, then the results can be interpreted in the context of the relatively small number of related parameter estimates that are associated with the structural model. Moreover, if the prior model has been based on a *meaningful theoretical structure*, regression estimates derived to depend on this structure should be meaningful to the investigator. In principle, the new methods thus reinforce the use of relevant prior information or knowledge by enhancing the interpretability and generalizability of results.

Just as importantly, if one's prior knowledge is in fact vague or diffuse, one can use adaptive WSR methods that account for this vagueness. Exploratory common factor models, the focus of this article, serve just this purpose since their use generally entails the assumption of vague prior knowledge. In relation to recently developed structural equation systems, those based on highly general modeling software, adaptive WSR methods seem less likely to force imputation of knowledge where it does not exist, or to induce model fabrication in situations where the investigation is genuinely exploratory.

Although nearly every applied scientist has been sensitized to statistical inferential thinking, most pay less heed to principles of domain sampling or psychometric inference. It seems, however, that neither form of inference can be ignored with impunity. Because some arbitrariness seems inevitable whenever regressors are selected from a larger domain of possible variables, it seems important to make this process explicit. Through WSR methodology it seems possible that both inferential concepts can be accommodated in a single, unified approach to regression. As shown above, adaptive WSR methods can account for measurement errors, they can be made to be scale free, and their algorithms generally do not break down even when there is an abundance of predictors.

Many authors have argued that one can expect routinely to loose little predictive accuracy if one just weights the — assumedly unreliable — predictors equally, and forgets about the complexities. It is easy to demonstrate that this can be poor advice, however, depending on the nature and extent of interdependencies among the various pairs of predictors and between them and the criterion variable. Such a cavalier attitude seems difficult to justify once convenient, easy-to-use software is available to facilitate adaptive forms of structural regression. Once measurement error has been taken into account, the details concerning structure of joint predictor-criterion relations may be of value not only for developing future hypotheses about how best to describe the variables, but also for estimating the upper limits to predictability of various criteria.

Equal or simple weighting has seemed like a pleasing idea to many students of regression methodology, and is compelling when it can be made to work (cf. Green, 1977). However, the question is: *Given a large number of predictors, each of whose predictive usefulness is in doubt, how does one decide which ones to weight say, one versus zero, or differentially?* Pruzek and Frederick (1978) showed that the choice of $m = 1$, with a reduced rank method, is roughly consistent with the use of equal weights, or perhaps simple zero-one weights. Yet, as shown above, this choice is demonstrably inferior to several other possibilities for some data systems. It is always partly an empirical question as to whether simple or more complicated weighting systems are

desirable. Although several varieties of OLS regression, including several forms of stepwise and all subset methodology have been developed, none of these methods account in any explicit way for measurement errors or for prior structural information.

Because certain WSR methods can provide a means to capitalize on psychometric redundancy, and can be tied to the concept of sampling variables from content (sub)domains, the use of these methods may carry implications for the design of prediction batteries. The key design principle would seem to be: observe as many regressor variables as content considerations or prior knowledge suggests are necessary to cover the reliable criterion variance, and follow with structural models for estimation that are tailored to use of the information collected.

The foregoing principle is clearly quite different from the standard operative principle in OLS regression applications: namely, use as few predictor variables as possible, attempting simultaneously to ensure that each will be individually valid for the criterion, as well as relatively uncorrelated with other predictors. OLS regression methods tend by their nature to discourage use of numerous predictors, except as summarized by a few preselected *composites*.

Although we have used language that seems to require a strictly Bayesian approach to analysis, we note that de Finetti himself, perhaps the greatest of all twentieth century Bayesians, has argued that as techniques, Bayesian methods are no more trustworthy for applications than other statistical techniques since all techniques can easily be misused or abused (cf. de Finetti, 1974). Use of adaptive WSR procedures can be made to be generally consistent with the Bayesian outlook, a la de Finetti, even though the methods as such are founded on classical statistical arguments. Some may prefer to think of adaptive WSR methods as having an empirical Bayes form. In any event, the strong connections between the foregoing theory and the results of Chen (1979) suggests a strong bridge between the two classes of methods.

Finally, it should be recognized that although MinRisk and some alternative WSR methods have been developed here using common factor models to convey a priori structures, a wide variety of structural models remain unexamined as a basis for estimators of the form of Equation 21. The use of common factor models seems generally consistent with exploratory applications where one's prior knowledge is diffuse or vague, a characterization that seems appropriate in exploratory contexts where multiple regression methods are so often employed. After all, if one *had firm knowledge* that a particular a priori model was appropriate for a certain predictor/criterion system, that knowledge could be used to construct a composite predictor, and perhaps dispense with multiple regression completely. Nevertheless, we encourage studies of all structural

forms that seem likely to support sound applications of prediction methodology, and recognize that the foregoing developments provide only an introduction to the possibilities.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis, 2nd edition.* New York: Wiley.

Browne, M. W. (1975). A comparison of single sample and cross-validation methods for estimating the mean squared error of prediction in multiple linear regression. *British Journal of Mathematical and Statistical Psychology, 28,* 112-120.

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

Chen, C. (1976). *Estimation of covariance matrices under a quadratic loss function* (Research Report S-46). Albany, NY: Department of Mathematics, State University of New York at Albany.

Chen, C. (1979). Bayesian inference for a normal dispersion matrix and its applications to stochastic multiple regression analysis. *Journal of the Royal Statistical Society, Series B, 41,* 235-248.

de Finetti, B. (1974). Bayesianism: It's unifying role for both the foundations and applications of statistics. *International Statistical Review, 42,* 117-130.

Dempster, A. P. (1969). *Elements of continuous multivariate analysis.* Boston: Addison Wesley.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement in cross-validation. *Journal of the American Statistical Association, 78,* 316-331.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1,* 54-77.

Green, B. F. (1977). Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research, 12,* 263-287.

Gunst, R. F., & Mason, R. L. (1980). *Regression analysis and its application: A data-oriented approach.* New York: Marcel Dekker.

Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika, 18,* 277-296.

Guttman, L. (1956). 'Best possible' systematic estimates of communalities. *Psychometrika, 21,* 273-285.

Harris, C. W. (1962). Some Rao-Guttman relationships. *Psychometrika, 27,* 247-263.

Jöreskog, K. G. (1969). Efficient estimation in image factor analysis. *Psychometrika, 34,* 51-75.

Jöreskog, K. G., & Sörbom, D. (1986). *Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares, 4th edition.* Mooresville, IN: Scientific Software.

Laughlin, J. E. (1979). A Bayesian alternative to least squares and equal weighting coefficients in regression. *Psychometrika, 44,* 271-288.

Laughlin, J. E. (1986). Contrasting alternatives to least squares in regression using diagnostics for identifying influential data. *Multivariate Behavioral Research, 21,* 103-118.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method.* London: Butterworths.

Lawley, D. N., & Maxwell, A. E. (1973). Regression and factor analysis. *Biometrika, 60*, 331-338.

Muirhead, R. J. (1985). Estimating a particular function of the multiple correlation coefficient. *Journal of the American Statistical Association, 80*, 923-925.

Press, S. J. (1975). Estimation of a normal covariance matrix. Unpublished manuscript.

Pruzek, R. M., & Frederick, B. C. (1978). Weighting predictors in linear models: Alternatives to least squares and limitations of equal weights. *Psychological Bulletin, 85*, 254-266.

Rabinowitz, S. N. (1990). *A simulation study of a class of random variable linear regression methods*. Unpublished doctoral dissertation, State University of New York at Albany, Albany, New York.

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika, 20*, 93-111.

Smith, G., & Campbell, F. (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association, 75*, 74-80.

Vinod, H. D. (1982). *Recent advances in regression methods*. New York: Marcel Dekker.

Appendix A

Proposition

Define $\hat{\Sigma}_G^* = w_0\hat{\Sigma}^* + (1 - w_0)\mathbf{G}$, where $\mathbf{G}$ is a fixed positive definite diagonal matrix. The difference between the risk for $\hat{\Sigma}^*$, with typical element $\hat{\sigma}_{jk}^*$, and the risk for $\hat{\Sigma}_G^*$ is given by

$$(1) \qquad R(\hat{\Sigma}^*) - R(\hat{\Sigma}_G^*) = \{(1 - w_0^2)/n\}\{\mathrm{tr}\Sigma^{*2} + (\mathrm{tr}\Sigma^*)^2\} - (1 - w_0)^2\mathrm{tr}(\Sigma^* - \mathbf{G})^2.$$

Proof

Write

$$(2) \qquad R(\hat{\Sigma}^*) = \sum \mathrm{Var}(\hat{\sigma}_{jk}^*) = (1/n)\{\mathrm{tr}\Sigma^{*2} + (\mathrm{tr}\Sigma^*)^2\},$$

which is a consequence of the fact that

$$(3) \qquad \mathrm{Var}(\hat{\sigma}_{jk}^*) = (1/n)(\sigma_{jk}^{*2} + \sigma_{jj}^*\sigma_{kk}^*),$$

(cf. Dempster, 1969), which holds for any data distribution where all fourth order cumulants are zero (such as the multivariate normal distribution) (cf. Browne, 1982), and

$$(4) \qquad R(\hat{\Sigma}_G^*) = w_0^2 E\{\mathrm{tr}(\hat{\Sigma}^* - \Sigma^*)^2\} + (1 - w_0)^2\mathrm{tr}(\Sigma^* - \mathbf{G})^2.$$

Proposition 1 follows easily using Equations 2 and 3, above.

Proposition 2

For a given Σ^*, with typical element σ_{jk}^*, let the $p \times p$ matrix $\mathbf{G} = \ddot{g}_0\mathbf{I}$, where $\mathbf{I}$ is the identity matrix and $\ddot{g}_0 = (1/p)\sum\hat{\sigma}_{jj}^*$. Define $\hat{\Sigma}^*_{cs\text{-}w_0} = \hat{\mathbf{Q}}\{w_0\hat{\Lambda} + (1 - w_0)\hat{\mathbf{G}}\}\hat{\mathbf{Q}}'$ (as in Equation 8) where $\hat{\mathbf{G}} = \hat{g}_0\mathbf{I}$, $g_0 = \sum\hat{\lambda}_j/p$. The difference between the risk for $\hat{\Sigma}^*_{cs\text{-}w_0}$ and the risk for $\hat{\Sigma}_G^*$ defined in Proposition 1 is

$$(5) \qquad R(\hat{\Sigma}^*_{cs\text{-}w_0}) - R(\hat{\Sigma}_G^*) = (1 - w_0^2)E\{\mathrm{tr}(\hat{\mathbf{G}} - \mathbf{G})^2\}.$$

Proof

$$(6) \qquad R(\hat{\Sigma}^*_{cs\text{-}w_0}) = E\{\mathrm{tr}(\hat{\Sigma}_G^* - \Sigma^*)^2\} + (1 - w_0)^2E\{\mathrm{tr}(\hat{\mathbf{G}} - \mathbf{G})^2\} + 2(1 - w_0)E\{\mathrm{tr}\,(\hat{\mathbf{G}} - \mathbf{G})(\hat{\Sigma}_G^* - \Sigma^*)\},$$

and since we can write

$$(7) \qquad \begin{aligned} E\{\mathrm{tr}(\hat{\mathbf{G}} - \mathbf{G})(\hat{\Sigma}_G^* - \Sigma^*)\} &= w_0E\{\mathrm{tr}\,(\hat{\mathbf{G}} - \mathbf{G})\hat{\Sigma}^*\} \\ &= w_0p\mathrm{Var}\,(\hat{g}_0) \\ &= w_0E\{\mathrm{tr}\,(\mathbf{G} - \mathbf{G})^2\}, \end{aligned}$$

we have the result.

Proposition 3

The difference between the risk for $\hat{\Sigma}^*$ and the risk for $\hat{\Sigma}^*_{cs\text{-}w_0}$ is given by

$$(8) \quad R(\hat{\Sigma}^*) - R(\hat{\Sigma}^*_{cs\text{-}w_0}) = \{(1 - w_0^2)/n\}[\{(p - 2)/p\}\mathrm{tr}\Sigma^{*2} + \{\mathrm{tr}\Sigma^*\}^2] - (1 - w_0)^2\{\mathrm{tr}\Sigma^{*2} - (1/p)(\mathrm{tr}\Sigma^*)^2\}.$$

Proof

Write

$$(9) \qquad R(\hat{\Sigma}^*) - R(\hat{\Sigma}^*_{cs\text{-}w_0}) = R(\hat{\Sigma}^*) - R(\hat{\Sigma}_G^*) - \{R(\hat{\Sigma}^*_{cs\text{-}w_0}) - R(\hat{\Sigma}_G^*)\}$$

and use Propositions 1 and 2.

Appendix B

Population System — Gunst-Mason Data

Matrix of Product-Moment Correlations

	1	2	3	4	5	6	7	8
1	1.000							
2	.648	1.000						
3	.535	.144	1.000					
4	.695	.279	.471	1.000				
5	.579	.144	.642	.502	1.000			
6	.595	.186	.716	.366	.592	1.000		
7	.783	.226	.662	.728	.542	.715	1.000	
8	.495	.368	.147	.428	.349	-.030	.282	1.000

Eigenvalues

1	2	3	4	5	6	7	8
4.366	1.409	.813	.594	.326	.299	.125	.067

Squared multiple correlations

1	2	3	4	5	6	7	8
.892	.674	.623	.657	.617	.758	.849	.471

Accepted December, 1990.