UvA ECONOMETRICS

Discussion Paper: 2007/05

Prediction of latent variables in a mixture of structural equation models, with an application to the discrepancy between survey and register data

Erik Meijer, Susann Rohwedder, and Tom Wansbeek

www.feb.uva.nl/ke/UvA-Econometrics

Amsterdam School of Economics
Department of Quantitative Economics
Roetersstraat 11
1018 WB AMSTERDAM
The Netherlands

UvA UNIVERSITEIT VAN AMSTERDAM

Prediction of latent variables in a mixture of structural equation models, with an application to the discrepancy between survey and register data

Erik Meijer, Susann Rohwedder, and Tom Wansbeek[*]

October 31, 2007

Abstract

We study the prediction of latent variables in a finite mixture of linear structural equation models. The latent variables can be viewed as well-defined variables measured with error or as theoretical constructs that cannot be measured objectively, but for which proxies are observed. The finite mixture component may serve different purposes: it can denote an unobserved segmentation in subpopulations such as market segments, or it can be used as a nonparametric way to estimate an unknown distribution. In the first interpretation, it forms an additional discrete latent variable in an otherwise continuous latent variable model. Different criteria can be employed to derive "optimal" predictors of the latent variables, leading to a taxonomy of possible predictors. We derive the theoretical properties of these predictors. Special attention is given to a mixture that includes components with degenerate distributions. We then apply the theory to the optimal estimation of individual earnings when two independent observations are available: one from survey data and one from register data. The discrete components of the model represent observations with or without measurement error, and with either a correct match or a mismatch between the two data sources.

1 Introduction

Since the earliest days of factor analysis (FA), researchers have used the estimation results to assign values to the factor scores for each individual observation. Intelligence measurement is the classical context for this, see, e.g., Carroll (1993). A recent example from economics is De Haan,

[*] Meijer: RAND Corporation and University of Groningen; Rohwedder: RAND Corporation; Wansbeek: University of Amsterdam and University of Groningen, t.j.wansbeek@uva.nl. We would like to thank Jelmer Ypma and Arie Kapteyn for sharing their code with us and for discussions about their model, and Jos ten Berge and conference participants at Aston Business School, Birmingham, in particular Battista Severgnini, for stimulating discussions and comments on an earlier version of this paper. The National Institute on Aging supported the collection of the survey data and subsequent match to administrative records (grant R03 AG21780).

Leertouwer, Meijer and Wansbeek (2003), who rated the degree of independence of central banks. Since the values are assigned to variables rather than to parameters, factor score *prediction* is a better expression than factor score *estimation*, although the latter is often used, as is factor score computing.

The use of factor score prediction is not restricted to cases like the above two examples, where the scores are interesting per se. They may also be used in further analysis. So, it may be desired to reduce a set of dependent variables in an experiment to a smaller set of underlying factors. These factors can then be used in subsequent analysis, taking their reliability into account (see, e.g., Kapteyn and Wansbeek, 1984). This paper is concerned with factor score prediction in a model that extends the FA model in two ways, one of which is essential.

First, we extend the usual FA model to encompass mixtures. This is an increasingly widespread way to allow for heterogeneity, with applications most notably in psychology (e.g., Arminger, Stein and Wittenberg, 1999) and marketing (e.g., Wedel and Kamakura, 2000).

The heterogeneity may stem from two sources. One is that the data may come from different classes (subpopulations or samples), in which different models are assumed to hold. Each observation comes from one particular class, but its origin is typically unknown to the researcher and has to be inferred from the data. We notice that handling heterogeneity by a discrete representation through mixing a finite number of classes is not the only approach feasible, and continuous representations may offer an attractive alternative, see Wedel et al. (1999).

Alternatively, heterogeneity is basically a different word for nonnormality. One way to model nonnormality is to flexibly approximate distributions by mixtures of normals. This kind of heterogeneity is of lesser importance since in many cases meaningful inference is possible with robust methods.

The other extension is that we consider structural equation models (SEMs) in general. The FA model is a member of the entire family of SEMs. However, an SEM can always be rewritten as an FA model, cf. section 3 below. Hence, prediction of latent variables in an SEM comes down to prediction of factor scores in an FA model. Of course, estimation of elaborate and highly structured SEMs is much more complicated than the estimation of a simple exploratory FA model; so in that respect the two are different. But in this paper we abstract from estimation issues and take all parameters to be known. It follows that our results—derived in an FA setting—apply to the entire class of SEMs.

Mixtures of SEMs have recently spawned a rather elaborate literature. A formulation of the SEM mixture model plus an algorithm for estimation of the parameters was given by Jedidi et al. (1997a), with an application in the field of marketing in Jedidi et al. (1997b). The model they consider is the classical LISREL model adapted to a multiclass context. The two sets of measurement equations, one for the indicators of the endogenous variables and one for those of the exogenous variables, have parameters that are allowed to differ across classes. The corresponding factors are class dependent and related through a system of simultaneous equations with parameters that may vary across classes. Estimation is through maximum likelihood, assuming normality within a class, and the likelihood is maximized by applying the EM algorithm.

To that end, unobservable variables z_{nj} are introduced, with $z_{nj} = 1$ if observation n belongs to class j and $z_{nj} = 0$ otherwise. In the E-step, the z_{nj} are estimated by the posterior probability of belonging to class j given the current parameter estimates. These estimates are substituted in the likelihood which is then maximized. This constitutes the M-step. The EM algorithm, plus a simplified scoring algorithm, is also described by Yung (1997), for the case of the mixture FA model. He considers a more general sampling scheme, allowing for the inclusion of observations with known class membership. Arminger et al. (1999) discuss estimation where normality is assumed conditional on the regressors only. A recent overview of mixture FA models was given by Lubke and Muthén (2005). Lee and Song (2003) discuss the handling of missing data, and Song and Lee (2004) present a sensitivity analysis. Estimation can be performed using the programs MECOSA (Arminger, Wittenberg and Schepers, 1996) and M*plus* (Muthén, 2004, appendix 8). Both programs handle conditional models, that is, they assume normality conditional on the regressors.

The current paper is concerned with the prediction of latent variables in a mixture of SEMs. As far as we know, this topic has not yet been described explicitly in the literature. The only paper we are aware of to present such predictors is Zhu and Lee (2001). They offer a Bayesian approach to the estimation of a mixture of SEMs, using the Gibbs sampler, which produces inter alia predictors for the latent variables. In contrast to this, we offer predictors that do not depend on an assumed prior distribution for the parameters.

The model we consider is described in section 2, and we derive the first- and second-order implications of the model. Section 3 presents a brief summary of the methods to predict factor scores for the simplest case where the model is normal. In section 4 we describe an array of predictors for the mixture case, as follows: section 4.1 deals with the predictors per class. These are not very useful per se, but they can be weighted over the classes to obtain potentially sensible predictors, see section 4.2. Predictors obtained by considering the full model are elaborated in section 4.3.

Up to this point in the paper it is assumed that no extraneous information is available about class membership and that all covariance matrices are nonsingular. In section 5 we adapt the analysis to cover the cases where these assumptions do not hold.

The remainder of the paper is dedicated to the study of a case that motivated the current research. Self-reported earnings information obtained in a survey of Swedish individuals was matched with administrative data. As a result, for about 400 individuals there are two independent observations of earnings. Individual-level comparison of the values from the two data sources reveals sizeable differences for the vast majority of cases. Only 59 cases show no or very small differences of up to \$100. Kapteyn & Ypma (2007) proposed a mixture model for these data, which can be cast in the form of a mixture SEM. For the empirical researcher who is interested in measuring the true value of individuals' earnings, the question arises how best to use the available information. This is what the method presented in this paper is designed to do: based on this mixture model with a latent variable, true earnings, combine the information from both data sources to predict the latent variable in some optimal way.

We present the application in section 6. The model for this case is discussed in section 6.1 and formulated as a special case of the general model considered in the current paper in section 6.2; theoretical predictors are given in section 6.3 and the empirical predictors in section 6.4. Section 7 concludes.

2 The model

In each class the standard FA model is assumed to hold, under normality. Classes are denoted by j. Let π_j be the (prior) probability that an observation is from class j. It is assumed that an observation from class j is generated from the model

$$y = \tau_j + B_j\xi + \varepsilon, \tag{1}$$

with $\xi \sim \mathcal{N}(\kappa_j, \Phi_j)$ and $\varepsilon \sim \mathcal{N}(0, \Omega_j)$. Moreover, ξ and ε are assumed independent of each other. Notice that in this formulation all parameters are allowed to differ across classes. In line with the factor score prediction literature, we assume the parameters to be known and we abstract from estimation issues, and hence from issues of identification.

A convenient side effect of this is that parameters are allowed to depend on covariates, e.g., $\tau_j = A_jx$, where x may vary across individuals. Another convenient side effect is that the parameters π_j, τ_j, B_j, κ_j, Φ_j, and Ω_j may be functions of a deeper parameter vector θ_j. This is especially relevant since any linear SEM with covariates x can be written as a submodel of the specification consisting of the measurement equation

$$y = Ax + B\xi + \varepsilon,$$

with $\mathrm{Cov}(\varepsilon) = \Omega$, and the structural equation

$$\xi = \Gamma\xi + \Delta x + \zeta,$$

with $\mathrm{Cov}(\zeta) = \Psi$. Solving the latter for ξ yields

$$\xi = (I - \Gamma)^{-1}\Delta x + (I - \Gamma)^{-1}\zeta,$$

which we can substitute back into the measurement equation. Omitting the class subscripts j, this gives our general model with structure on the parameters,

$$\begin{aligned}\tau &= Ax\\ \kappa &= (I - \Gamma)^{-1}\Delta x\\ \Phi &= (I - \Gamma)^{-1}\Psi(I - \Gamma')^{-1}.\end{aligned}$$

Combining both side effects, we conclude that our setup allows for an arbitrary linear structural equation model specification within each class.

Throughout the paper all expressions pertain to a generic observation. Hence we do not use subscripts to indicate individual observations. If the parameters are known (or once they have been consistently estimated), prediction of ξ for a given observation does not depend on the other observations since we assume independence of observations.

Unlike the standard FA case where observations are routinely centered and hence intercepts play no role, we explicitly incorporate intercepts since they will generally vary across classes and constitute an essential and complicating ingredient of the predictors.

A direct implication of (1) is

$$\begin{pmatrix} y \\ \xi \end{pmatrix} \Big| \, j \sim \mathcal{N}\left[\begin{pmatrix} \mu_j \\ \kappa_j \end{pmatrix}, \begin{pmatrix} \Sigma_j & B_j\Phi_j \\ \Phi_j B_j' & \Phi_j \end{pmatrix}\right], \tag{2}$$

with

$$\mu_j \equiv \tau_j + B_j\kappa_j \tag{3a}$$

$$\Sigma_j \equiv B_j\Phi_j B_j' + \Omega_j. \tag{3b}$$

The first-order and second-order moments of y and ξ follow directly:

$$\mu_y \equiv \mathrm{E}(y) = \sum_j \pi_j\mu_j = \sum_j \pi_j\left(\tau_j + B_j\kappa_j\right) \tag{4a}$$

$$\mu_\xi \equiv \mathrm{E}(\xi) = \sum_j \pi_j\kappa_j \tag{4b}$$

$$\begin{aligned}\Sigma_y &\equiv \mathrm{Var}(y) = \mathrm{E}(yy') - \mathrm{E}(y)\mathrm{E}(y)' \\ &= \sum_j \pi_j(\Sigma_j + \mu_j\mu_j') - \mu_y\mu_y'\end{aligned} \tag{4c}$$

$$\begin{aligned}\Sigma_\xi &\equiv \mathrm{Var}(\xi) = \mathrm{E}(\xi\xi') - \mathrm{E}(\xi)\mathrm{E}(\xi)' \\ &= \sum_j \pi_j(\Phi_j + \kappa_j\kappa_j') - \mu_\xi\mu_\xi'\end{aligned} \tag{4d}$$

$$\begin{aligned}\Sigma_{\xi y} &\equiv \mathrm{Cov}(\xi, y) = \mathrm{E}(\xi y') - \mathrm{E}(\xi)\mathrm{E}(y)' \\ &= \sum_j \pi_j(\Phi_j B_j' + \kappa_j\mu_j') - \mu_\xi\mu_y'.\end{aligned} \tag{4e}$$

From these formulas it becomes clear that the consideration of more than one class leads to rather unelegant expressions.

This very general formulation nests many special cases that have received attention in the literature. We mention a few examples. Schneeweiss and Cheng (2006) use mixtures to approximate the density of regressors in a (possibly nonlinear) structural measurement error model. For the special case of a linear relationship, this fits into our framework with y being a scalar, and only the κ_j and Φ_j varying across classes. Phillips (2003) uses a mixture of normals to approximate the distribution of the disturbances in an error-components model for panel data. For a generic observation, the model is

$$y = X\beta + \iota_T\mu + v,$$

where T is the time dimension of the panel, ι_T a vector of T ones, X is a matrix of regressors varying across individuals, μ is a random individual effect and v is a random error term, both mixture normal with mean zero and variance $\sigma^2_{\mu j}$ and $\sigma^2_{vj}I_T$, respectively. We can translate this into our framework by taking ξ equal to μ, the τ_j's equal to $X\beta$, the B_j's equal to ι_T, the κ_j's equal to zero, all constant across classes, and for all j separately Φ_j equal to $\sigma^2_{\mu j}$ and Ω_j equal to $\sigma^2_{vj}I_T$.

Notice that this translation into our framework is not unique. There are various ways to express this model in terms of our general model, depending on the choice for the factors ξ. We took it equal to μ since, in case there is interest in factor score prediction, it is likely to concern the individual effects. However, we could also have taken ξ equal to $(\mu, v')'$ by some appropriate adaptations in the translation scheme, most notably by setting all Ω_j equal to zero. Alternatively, we could interchange the role of μ as the latent variable and v as the disturbance term.

3 The basics of factor score prediction

For reference later on, we recapitulate in this section the basics of factor score prediction when there is just one class. Hence, given our earlier assumptions, all variables are normally distributed, although a large part of the theory of factor score prediction still holds under nonnormality, see, e.g. Wansbeek and Meijer (2000, section 7.3). We use the notation as above, but without the class identifier j.

Before we discuss the various factor score predictors we observe that we can present the expressions in two forms. One form is based on the covariance matrix Σ of the marginal distribution of y, and the other form is based on the covariance matrix Ω of the conditional distribution of y given ξ. Hence we dub the first form the "marginal" one, and the other one the "conditional" one. The link between the two forms is provided by the following. We assume that Φ and Ω are nonsingular and let

$$\Lambda \equiv \Phi^{-1} + B'\Omega^{-1}B. \tag{5}$$

Then $\Sigma\Omega^{-1}B = B + B\Phi B'\Omega^{-1}B = B\Phi\Lambda$, from which

$$\Sigma^{-1}B\Phi = \Omega^{-1}B\Lambda^{-1} \tag{6}$$

follows, or $\Omega\Sigma^{-1}B = B\Lambda^{-1}\Phi^{-1}$. This fact allows for shifting between expressions in the marginal and the conditional form.

Now consider the conditional distribution of ξ given y. Let

$$\hat{\xi} \equiv \kappa + \Phi B'\Sigma^{-1}(y - \mu) \tag{7}$$

$$F \equiv \Phi - \Phi B'\Sigma^{-1}B\Phi. \tag{8}$$

Using the expression for the conditional normal distribution we obtain

$$\xi \mid y \sim \mathcal{N}(\hat{\xi}, F). \tag{9}$$

We consider the minimization of the mean-squared error of a function of y as a predictor of ξ. Using a standard result in statistics, the MSE is minimized by the conditional expectation. So $\hat{\xi}$ is the minimum MSE predictor of ξ.

Alternative expressions in the conditional rather than the marginal form are

$$\begin{aligned}\hat{\xi} &= \kappa + \Lambda^{-1}B'\Omega^{-1}(y-\mu)\\ F &= \Phi(\Lambda - B'\Omega^{-1}B)\Lambda^{-1} = \Lambda^{-1},\end{aligned}$$

which follows directly by applying (5) and (6).

From (7) it is clear that this predictor is linear in y. Had we imposed the restriction that our minimum MSE predictor be linear, we would have obtained the same result, $\hat{\xi}_{\mathrm{L}} = \hat{\xi}$, in self-evident notation.

The predictor $\hat{\xi}$ is biased in the sense that

$$\mathrm{E}(\hat{\xi} - \xi \mid \xi) = -(I - \Phi B'\Sigma^{-1}B)(\xi - \kappa) \neq 0.$$

This may be considered a drawback of $\hat{\xi}$. An unbiased predictor can be obtained as follows. The model, rewritten as $y - \tau = B\xi + \varepsilon$, can be interpreted as a regression model with ξ as the vector of regression coefficients, B the matrix of regressors, and $y - \tau$ as the dependent variable. In this model the GLS estimator is the best linear unbiased estimator (BLUE) of ξ, and under normality it is also BUE as it attains the Cramér-Rao bound (e.g., Amemiya, 1985, section 1.3.3). So, extending our self-evident notation,

$$\begin{aligned}\hat{\xi}_{\mathrm{LU}} &= \hat{\xi}_{\mathrm{U}}\\ &= (B'\Omega^{-1}B)^{-1}B'\Omega^{-1}(y-\tau)\\ &= \kappa + (B'\Omega^{-1}B)^{-1}B'\Omega^{-1}(y-\mu)\end{aligned} \tag{10}$$

minimizes the conditional MSE, $\mathrm{E}[(\hat{\xi} - \xi)^2 \mid \xi]$, for every possible value of ξ, conditional on unbiasedness. This implies that it must also minimize the unconditional MSE, $\mathrm{E}[(\hat{\xi} - \xi)^2] = \mathrm{E}\{\mathrm{E}[(\hat{\xi} - \xi)^2 \mid \xi]\}$, among unbiased predictors and thus that it is the best unbiased predictor of ξ. Using (6) we obtain

$$\hat{\xi}_{\mathrm{LU}} = \kappa + (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}(y-\mu)$$

as a reformulation of the conditional into the marginal form. This predictor is called the Bartlett predictor, after Bartlett (1937). Although this predictor is based on the interpretation of the FA model as a regression model, $\hat{\xi}$ rather than $\hat{\xi}_{\mathrm{LU}}$ is (somewhat confusingly) called the regression predictor as it is based on the notion of regression as a conditional expectation.

When computing factor score predictors it is sometimes deemed desirable to impose a priori the restrictions that the predictors and the factors in the model have the same covariance, i.e.,

$$\mathrm{Cov}(\hat{\xi}) = \Phi;$$

see e.g. Ten Berge et al. (1999). This is usually called correlation preservation although it involves the covariance rather than the correlation. In the present context, where means play a more explicit role than is usually the case, it is natural to impose mean preservation as well, i.e.,

$$\mathrm{E}(\hat{\xi}) = \kappa,$$

and consider predictors that are mean and covariance preserving (MCP). After some algebra one can show that the trace of the MSE matrix is minimized by

$$\hat{\xi}_{\mathrm{MCP}} = \kappa + \Phi^{1/2}(\Phi^{1/2}B'\Sigma^{-1}B\Phi^{1/2})^{-1/2}\Phi^{1/2}B'\Sigma^{-1}(y-\mu),$$

as expressed in the marginal form. So we now have three predictors, which can be parameterized as follows:

$$\hat{\xi}(a) = \kappa + \Phi^{1/2}(\Phi^{1/2}B'\Sigma^{-1}B\Phi^{1/2})^{-1/a}\Phi^{1/2}B'\Sigma^{-1}(y-\mu),$$

with $\hat{\xi}(1) = \hat{\xi}_{\mathrm{LU}}, \hat{\xi}(2) = \hat{\xi}_{\mathrm{MCP}}$, and $\hat{\xi}(\infty) = \hat{\xi}$. In the context of a mixture model considering an MCP predictor does not seem to be very meaningful unless the means and variances are the same over the classes.

4 A taxonomy of predictors

After this brief summary of prediction in the single-class model under normality we are now in a position to discuss a whole array of predictors for the mixture model. This array can be grouped into a taxonomy, whose first node is whether the predictors are designed from below or above, so to speak. That is, we can start with predictors per class, as discussed in the previous section, and then combine them with some kind of weighting, or we can derive system-wide predictors right away by applying the minimum MSE criterion with or without the restrictions of linearity or unbiasedness.

The kind of weighting of predictors per class defines another node. The weighting can be done by employing the prior class probabilities π_j or it can be based on posterior weights. More specifically, we employ the posterior probabilities, conditional on y, of being in a particular class. Using Bayes' rule, these probabilities are

$$p_j(y) \equiv \Pr(\text{class} = j \mid y) = \frac{\pi_j f_j(y)}{\sum_k \pi_k f_k(y)}, \tag{11}$$

where $f_j(y)$ is the normal density, evaluated in y, with mean μ_j and variance Σ_j.

The predictors are grouped in table 1. The columns of the table reflect two obvious restrictions that one could impose: linearity and unbiasedness. (Meijer and Wansbeek, 1999, consider quadratic rather than linear predictors, but this generalization is not pursued here.)

4.1 Predictors per class

The first row of the table presents the predictors from the classes. Within a class the distribution is normal, so the imposition of linearity is redundant, as argued above, and just two instead of four predictors are given. We extend the notation of section 3 by adding the class subscript j.

It should be noted that the first predictor, $\hat{\xi}_{\mathrm{U},j}$ is placed under the heading 'unbiased' due to the reasoning along the lines of a single class; if class j were the only class present, this predictor would be unbiased, but this quality, of course, is lost in a multiclass context. It is of interest to elaborate this. From (10)

$$\hat{\xi}_{\mathrm{U},j} = (B_j'\Omega_j^{-1}B_j)^{-1}B_j'\Omega_j^{-1}(y - \tau_j) \equiv B_j^{-}(y - \tau_j),$$

where the notation B_j^{-} is chosen since it is a generalized inverse of B_j, we have

$$\mathrm{E}(\hat{\xi}_{\mathrm{U},j} \mid \xi) = B_j^{-}(\mathrm{E}(y \mid \xi) - \tau_j).$$

Let

$$q_j(\xi) \equiv \Pr(\text{class} = j \mid \xi) = \frac{\pi_j g_j(\xi)}{\sum_k \pi_k g_k(\xi)}, \tag{12}$$

where $g_j(\xi)$ is the normal density, evaluated in ξ, with mean κ_j and variance Φ_j. Then

$$\mathrm{E}(y \mid \xi) = \sum_j q_j(\xi)\,\mathrm{E}(y \mid \xi, j) = \sum_j q_j(\xi)\left(\tau_j + B_j\xi\right) = \bar{\tau} + \bar{B}\xi, \tag{13}$$

with

$$\begin{aligned}\bar{\tau} &\equiv \sum_j q_j(\xi)\,\tau_j \\ \bar{B} &\equiv \sum_j q_j(\xi)\,B_j.\end{aligned}$$

and

$$\mathrm{E}(\hat{\xi}_{\mathrm{U},j} \mid \xi) = B_j^{-}(\bar{\tau} - \tau_j + \bar{B}\xi).$$

For unbiasedness, this should be equal to ξ for all ξ. (Notice that ξ also enters into $\bar{\tau}$ and $\bar{B}$.) This seems possible only in the special case where the τ_j and B_j do not vary across classes. The heterogeneity is then restricted to the case where the factor loadings are homogeneous, and only the distributions of ξ and ε have a mixture character.

4.2 Weighted predictors

The next row of the table contains the weighted averages of the predictors per class,

$$\hat{\xi}_{\mathrm{U}} = \sum_j \pi_j \hat{\xi}_{\mathrm{U},j} \quad \text{and} \quad \hat{\xi} = \sum_j \pi_j \hat{\xi}_j.$$

Table 1: A taxonomy of predictors.

<table>
<tr><th></th><th colspan="4">Restrictions</th></tr>
<tr><th></th><th>Linear
unbiased</th><th>Unbiased</th><th>Linear</th><th>None</th></tr>
<tr><td>Within class (j)</td><td colspan="2">$\hat{\xi}_{\mathrm{U},j}$</td><td colspan="2">$\hat{\xi}_j$</td></tr>
<tr><td>Weighted</td><td colspan="2">$\hat{\xi}_{\mathrm{U}}$</td><td colspan="2">$\hat{\xi}$</td></tr>
<tr><td>Posterior-weighted</td><td colspan="2">$\hat{\xi}^*_{\mathrm{U}}$</td><td colspan="2">$\hat{\xi}^*$</td></tr>
<tr><td>System-wide</td><td>$\left[\tilde{\xi}_{\mathrm{LU}}\right]$</td><td>$\left[\tilde{\xi}_{\mathrm{U}}\right]$</td><td>$\tilde{\xi}_{\mathrm{L}}$</td><td>$\tilde{\xi}\ (=\hat{\xi}^*)$</td></tr>
</table>

The same remark as to unbiasedness applies at the aggregate level as well.

The results for posterior weighting are displayed in the next row of the table. We use an asterisk to distinguish these predictors from the previous ones. Then

$$\hat{\xi}^*_{\mathrm{U}} = \sum_j p_j(y)\,\hat{\xi}_{\mathrm{U},j} \quad \text{and} \quad \hat{\xi}^* = \sum_j p_j(y)\,\hat{\xi}_j. \tag{14}$$

Notice that the weights involve y and hence the linearity property is lost when weighting.

4.3 System-wide predictors

The bottom row of the table presents the system-wide predictors that are obtained by minimizing the MSE with or without the restrictions. Since in the mixture model normality does not hold any longer, linearity is not obtained as a byproduct, and the full range of four different predictors has to be considered. We use a tilde to describe system-wide predictors, in otherwise evident notation.

Prediction under linearity and unbiasedness The starting point for finding an unbiased predictor that is linear in y is to find a matrix L and a vector b such that

$$\tilde{\xi}_{\mathrm{LU}} = L'y + b$$

satisfies $\mathrm{E}(\tilde{\xi}_{\mathrm{LU}} \mid \xi) = \xi$ for all ξ. Using (13) we obtain

$$\mathrm{E}(\tilde{\xi}_{\mathrm{LU}} \mid \xi) = L'\mathrm{E}(y \mid \xi) + b = L'(\bar{\tau} + \bar{B}\xi) + b.$$

This should be equal to ξ for any ξ, so also for $\xi = 0$. So $b = -L'\bar{\tau}_0$, with $\bar{\tau}_0 \equiv \sum_j q_j(0)\tau_j$. Substituting this back leads to the requirement

$$L'(\bar{\tau} - \bar{\tau}_0 + \bar{B}\xi) = \xi$$

for all ξ. Because of the (nonlinear) dependence of $\bar{\tau}$ and $\bar{B}$ on ξ, we conjecture that there does not exist a matrix L with this property, and hence there does not exist a linear unbiased predictor.

Prediction under unbiasedness Imposing linearity may preclude a solution. We relax this requirement, but maintain unbiasedness. We then consider a predictor $\tilde{\xi}_{\mathrm{U}} = h(y)$, where $h(y)$ is a function of y such that $\mathrm{E}(h(y) \mid \xi) = \xi$ for all ξ. Using (12), this requirement for $h(y)$ can be rewritten as

$$\sum_j \pi_j g_j(\xi) \left\{ \mathrm{E}[h(y) \mid \xi, j] - \xi \right\} = 0$$

for all ξ. This does not seem to lead to anything tractable.

Prediction under linearity This case is conceptually the simplest one, and is based on the idea of linear projection. According to a well-known result, the MSE is minimized over all linear functions by

$$\tilde{\xi}_{\mathrm{L}} = \mu_\xi + \Sigma_{\xi y} \Sigma_y^{-1} (y - \mu_y). \tag{15}$$

The expressions for the means μ_ξ and μ_y and the covariance matrices $\Sigma_{\xi y}$ and Σ_y were given in section 2. Their *Gestalt* precludes further elaboration.

Prediction without restrictions The approach here rests on the fact that the MSE is minimized by the mean of the conditional distribution. Adapting results from section 3 we have

$$\xi \mid y, j \sim \mathcal{N}\left(\hat{\xi}_j, F_j\right),$$

with

$$\begin{aligned} \hat{\xi}_j &\equiv \kappa_j + \Phi_j B_j' \Sigma_j^{-1} (y - \mu_j) \\ F_j &\equiv \Phi_j - \Phi_j B_j' \Sigma_j^{-1} B_j \Phi_j. \end{aligned}$$

So $\xi \mid y$ is mixture normal and the system-wide predictor without restrictions is given by

$$\tilde{\xi} = \mathrm{E}(\xi \mid y) = \sum_j p_j(y) \hat{\xi}_j,$$

with $p_j(y)$ given by (11). Notice that $\tilde{\xi}$ equals the predictor obtained by posterior weighting of the unrestricted predictors per class, $\hat{\xi}_j$. As to the variance of this predictor, we notice

$$\mathrm{Var}(\xi \mid y) = \sum_j p_j(y) F_j + \sum_j p_j(y) \hat{\xi}_j \hat{\xi}_j' - \tilde{\xi}\tilde{\xi}'.$$

The first term on the right-hand side is the weighted average of the variances within the classes, and the other terms represent the between-class variance.

4.4 Summary

The above discussion leaves us with five predictors. These are the prior weighted predictors $\hat{\xi}_{\mathrm{U}}$ and $\hat{\xi}$, their posterior-weighted counterparts $\hat{\xi}^*_{\mathrm{U}}$ and $\hat{\xi}^*$, the latter being also the system-wide conditional expectation, and the linear projection-based predictor $\tilde{\xi}_{\mathrm{L}}$.

Although some of these predictors were built on predictors that would be unbiased in a single-class setting under normality, none of these predictors is unbiased themselves, and attempts to construct unbiased estimators seem futile.

5 Labeled classes and degeneracy

The derivations in the previous sections assumed that all covariance matrices are nonsingular and that no extraneous information is available about class membership. However, in some cases, including the empirical case to be discussed further on, these assumptions are not met. We then have to adapt the analysis somewhat. If we know from which class an observation is drawn we have the situation of *labeled classes*. *Degeneracy* denotes singular covariance matrices.

Labeled classes In the case of a labeled class, it is known to which class an observation belongs. For example, we may have different factor analysis models for males and females for the same dependent variables, and we have observed whether the individual is male or female. Analogous to conditioning on any possible exogenous variables x, it appears most fruitful to proceed by conditioning on the given class. But then the model for a given observation is not a mixture anymore, but an ordinary factor analysis model, and the basic factor score predictors from section 3 can be used straightforwardly.

A generalization of this is when it is known that the observation belongs to a certain subset of the classes, but it is not known to which particular class within the subset. For example, we might have a model that postulates three market segments for women and two for men, so that there are five classes. For each observation, the gender of the individual is known, but not the precise market segment. Then we can restrict attention to the given subset of classes. After renormalizing the class probabilities so that they sum to 1 for the given subset, and disregarding classes that are not in the given subset, this situation is the same as the general situation.

The fully general extension of this is when the information of class membership takes the form of individual-specific prior probabilities $\Pr(\text{class} = j) = P_j$. Note that we still omit the index denoting the individual, because for prediction we only consider one individual at a time. It is now immediately clear that, for prediction purposes, this situation is simply the general model as presented earlier, with $\pi_j = P_j$. So the situation of labeled classes is subsumed in the general model.

Degeneracies Several types of degeneracies are conceivable in a single-class factor analysis model. For substantive reasons, the errors ε of some of the indicators may be assumed to be identically zero, leading to a singular covariance matrix Ω. This may also be the result of the estimation procedure without prior restriction, i.e., a *Heywood case*. Similarly, the covariance matrix Φ of the factors may be singular. Finally, the factor loadings matrix B may not be of full column rank. The latter does not necessarily lead to a degenerate distribution, but it does imply that some of the formulas presented earlier cannot be applied. In a single-class factor analysis setting, singularity of Φ and/or deficient column rank of B are not particularly interesting or relevant, because the model is equivalent to a model with fewer factors and thus can be reduced to a more parsimonious one. However, in a mixture setting, such forms of degeneracy may occur quite naturally.

If Ω is nonsingular, then evidently the covariance matrix Σ of the observed indicators is also nonsingular. Thus, a necessary, but by no means sufficient, condition for singularity of Σ is singularity of Ω.

The formulas for the single-class predictors cannot be directly applied in the case of degeneracies, because the required inverses do not exist. In appendix A, expressions are given for the unbiased and unrestricted predictors that can be used in this case. These expressions can be used for the within-class predictors in a mixture model with degeneracies.

In a mixture factor analysis model, the degeneracies mentioned here may be limited to only a subset of the classes, or all classes may have degenerate distributions. In the former case, the marginal distribution of the indicators is nondegenerate. In the latter case, it may be either nondegenerate or degenerate, depending on the types of degeneracies and the parameter values. When there are classes with degenerate distributions of the indicators y, the formula (11) is invalid, because y does not have a proper density in such classes. Appendix B gives the appropriate expressions for the posterior probabilities in this case.

With the adapted expressions for the within-class predictors and posterior probabilities, prediction of the factor scores in mixture models with degenerate distributions is now relatively straightforward. For the non-system-wide predictors, we can apply the standard expressions, using the adapted formulas for the within-class predictors and the posterior probabilities. For the system-wide linear predictor, expressions (2)–(4) are still valid, and thus the predictor $\tilde{\xi}_{\mathrm{L}}$ is still given by (15), with Σ_y^{-1} replaced by Σ_y^{+} if it is singular. The system-wide unrestricted predictor $\tilde{\xi}$ is again equivalent to the posterior-weighted unrestricted predictor $\hat{\xi}^{*}$.

6 Application to Swedish earnings data

To illustrate our proposed methods we apply the above to an empirical example taken from Kapteyn and Ypma (2007). They study the relationship between observations of individuals' earnings obtained from two different sources: self-reports collected in survey interviews which were subsequently matched to administrative records, presumably pertaining to the same individuals. We first present the model in Kapteyn and Ypma's notation and then show how it is a special case of a mixture SEM. We next present the various predictors for this special case.

6.1 Combining register and survey data

The data used in Kapteyn and Ypma (2007) come from Sweden. Since we take the parameter estimates presented in their work as given we only provide a short overview. The data were collected as part of a validation study to inform subsequent data gathering efforts in the Survey of Health, Ageing and Retirement in Europe (SHARE). The objective was to compare self-reported information obtained in a survey with information obtained from administrative data sources (registers) to learn about measurement error, in particular in respondents' self-reports. The survey was fielded in the spring of 2003, shortly after tax preparation season, when people are most likely to have relevant financial information fresh in their minds. The sample was designed to cover individuals age 50 and older and their spouses; it was drawn from an existing data base of administrative records (LINDA) which contains a five percent random subsample of the entire Swedish population. Out of 1431 sampled individuals 881 responded. Kapteyn and Ypma investigate the relation between survey data and register data for three different variables. We use their model on individual earnings without covariates for our illustrative example, which is based on 400 observations with positive values recorded both in the register and in the survey data.

In the Kapteyn-Ypma model, there is a single latent variable ξ, representing the true value of the logarithm of earnings of an individual. (As before, we omit the index indicating the individual.) This true value is not measured directly. However, there are two indicators available, one coming from administrative data and the other from individuals' self-reports collected in a survey. Figure 1 depicts the observations: the logs of the survey earnings are plotted against the logs of the register earnings. In this figure, we see that most points are close to the $r = s$ line that is also plotted. Thus, for most respondents, earnings as reported in the survey are close to the earnings observed in the administrative records. But a noticeable number of points is not located close to the $r = s$ line. Unlike most of the literature in this area, the Kapteyn-Ypma model does not assume that this is necessarily due to measurement error in the survey earnings. Rather, it is assumed that both indicators may contain error; the administrative data are not a priori assumed to be error-free. However, the structure of the error differs between the indicators.

Let ξ and three other random variables, ζ, η, and ω, be all independent of each other and normally distributed with means and variances (μ_ξ, σ_ξ^2), $(\mu_\zeta, \sigma_\zeta^2)$, $(\mu_\eta, \sigma_\eta^2)$, and $(\mu_\omega, \sigma_\omega^2)$, respectively. Let r denote the value of log earnings from the administrative data and let s be the corresponding value from the survey.

It is assumed that, whenever there are errors in the administrative data, they are only due to mismatching. Let π_r be the probability by which the observed administrative value, r, equals the true value of individual i's earnings, ξ. $1 - \pi_r$ is the probability of a mismatch. In case of a mismatch the administrative value r corresponds to the true value of someone else. This mismatched value will be denoted by ζ. According to the assumptions made above, ζ is not correlated with ξ. According to the same assumptions, the means and variances of ζ and ξ are allowed to differ. This reflects the background of the data collection. The survey data only cover a subset of the population restricted to individuals age 50 and older and their spouses, but a mismatch

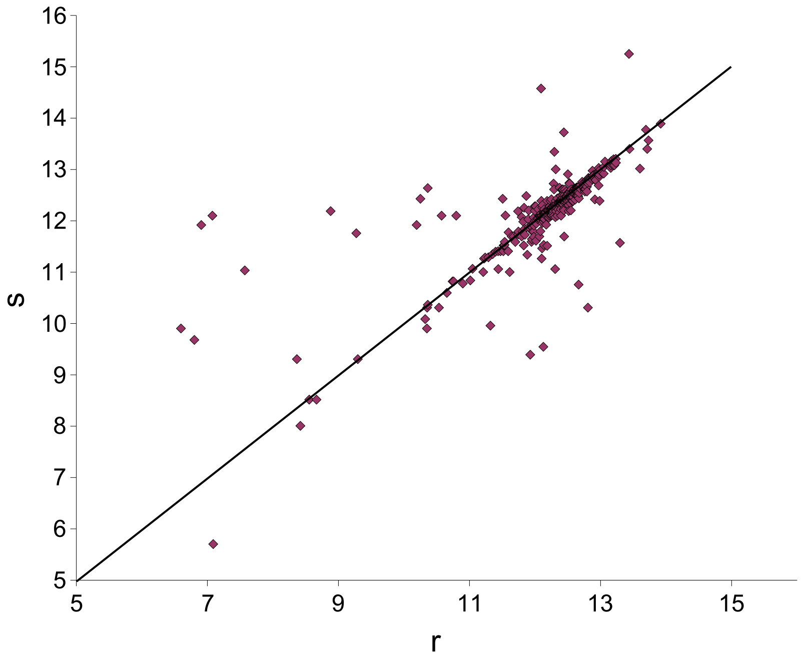

Figure 1: Survey data versus register data.

may involve an individual from the entire administrative database which is a random sample of the entire Swedish population. To formalize this part of the model, for the register data, the observed values r are mixture normal with

$$r = \begin{cases} \xi & \text{with probability } \pi_r \qquad \text{(R1)} \\ \zeta & \text{with probability } 1 - \pi_r, \qquad \text{(R2)} \end{cases}$$

the first case reflecting a match and the second case reflecting a mismatch.

The second part of the model covers the survey data. Three cases are distinguished. We let π_s denote the probability that the observed survey value is correct. With probability $1 - \pi_s$ the survey data contain response error, part of which is mean-reverting, as expressed by the term $\rho(\xi - \mu_\xi)$, where $\rho < 0$ implies a mean-reverting response error in the sense of Bound and Krueger (1991). When the data contain response error, there is a π_ω probability that the data are contaminated, modeled by adding an extra error-term, ω. Contamination can, e.g., result from erroneously reporting monthly earnings as annual or vice versa. So, collecting the elements of this second part of the model, for the survey data, the observed values s are mixture normal with

$$s = \begin{cases} \xi & \text{with probability } \pi_s \qquad \text{(S1)} \\ \xi + \rho(\xi - \mu_\xi) + \eta & \text{with probability } (1 - \pi_s)(1 - \pi_\omega) \qquad \text{(S2)} \\ \xi + \rho(\xi - \mu_\xi) + \eta + \omega & \text{with probability } (1 - \pi_s)\pi_\omega. \qquad \text{(S3)} \end{cases}$$

Table 2: Parameterization of the Swedish earnings data model in terms of a mixture SEM.

Class (j)	r	s	π_j	τ_j	B_j	Ω_j
1	R1	S1	$\pi_r \pi_s$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$
2	R1	S2	$\pi_r(1-\pi_s)(1-\pi_\omega)$	$\begin{pmatrix} 0 \\ \mu_\eta - \rho\mu_\xi \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1+\rho \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix}$
3	R1	S3	$\pi_r(1-\pi_s)\pi_\omega$	$\begin{pmatrix} 0 \\ \mu_\eta + \mu_\omega - \rho\mu_\xi \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1+\rho \end{pmatrix}$	$\begin{pmatrix} 0 & 0 \\ 0 & \sigma_\eta^2 + \sigma_\omega^2 \end{pmatrix}$
4	R2	S1	$(1-\pi_r)\pi_s$	$\begin{pmatrix} \mu_\zeta \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & 0 \end{pmatrix}$
5	R2	S2	$(1-\pi_r)(1-\pi_s)(1-\pi_\omega)$	$\begin{pmatrix} \mu_\zeta \\ \mu_\eta - \rho\mu_\xi \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1+\rho \end{pmatrix}$	$\begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix}$
6	R2	S3	$(1-\pi_r)(1-\pi_s)\pi_\omega$	$\begin{pmatrix} \mu_\zeta \\ \mu_\eta + \mu_\omega - \rho\mu_\xi \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1+\rho \end{pmatrix}$	$\begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & \sigma_\eta^2 + \sigma_\omega^2 \end{pmatrix}$

Having established the distributions of r and s separately, we notice that the distribution of (r, s) is a mixture of $2 \times 3 = 6$ classes.

6.2 Formulation as a special case of the general model

We can now express the parameters of the general model in this paper in terms of the parameters of the Swedish earnings data model as given above. We have

$$y = \begin{pmatrix} r \\ s \end{pmatrix}$$

and it is easy to see that $\kappa_j = \mu_\xi$ and $\Phi_j = \sigma_\xi^2$ for all classes and that for the rest the correspondence between the parameters is shown in table 2. A striking feature is that Ω_j is singular for classes $1, \ldots, 4$. So there is a degree of degeneracy in the model, cf. the discussion in section 5.

Given these expressions, it is straightforward to combine them to obtain the mean and covariance matrix of the observations y per class according to (3a) and (3b). The result is presented in table 3. It is interesting to see that modeling a few simple, reasonable points of departure implies a rather formidable looking structure. As indicated above, Σ_1 is singular, because y takes on values in the subspace $y_1 = y_2$, i.e., $r = s$, whereas Σ_j is nonsingular for $j = 2, \ldots, 6$.

6.3 Predictors

Given this structure, we now turn to the issue of prediction. We discuss a number of predictors, following the taxonomy presented in table 1.

Table 3: Within-class means and covariance matrices.

Class (j)	μ_j	Σ_j
1	$\begin{pmatrix} \mu_\xi \\ \mu_\xi \end{pmatrix}$	$\begin{pmatrix} \sigma_\xi^2 & \sigma_\xi^2 \\ \sigma_\xi^2 & \sigma_\xi^2 \end{pmatrix}$
2	$\begin{pmatrix} \mu_\xi \\ \mu_\xi + \mu_\eta \end{pmatrix}$	$\begin{pmatrix} \sigma_\xi^2 & (1+\rho)\sigma_\xi^2 \\ (1+\rho)\sigma_\xi^2 & (1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 \end{pmatrix}$
3	$\begin{pmatrix} \mu_\xi \\ \mu_\xi + \mu_\eta + \mu_\omega \end{pmatrix}$	$\begin{pmatrix} \sigma_\xi^2 & (1+\rho)\sigma_\xi^2 \\ (1+\rho)\sigma_\xi^2 & (1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2 \end{pmatrix}$
4	$\begin{pmatrix} \mu_\zeta \\ \mu_\xi \end{pmatrix}$	$\begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & \sigma_\xi^2 \end{pmatrix}$
5	$\begin{pmatrix} \mu_\zeta \\ \mu_\xi + \mu_\eta \end{pmatrix}$	$\begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & (1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 \end{pmatrix}$
6	$\begin{pmatrix} \mu_\zeta \\ \mu_\xi + \mu_\eta + \mu_\omega \end{pmatrix}$	$\begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & (1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2 \end{pmatrix}$

Table 4: Within-class predictors.

Class (j)	$\hat{\xi}_{\mathrm{U},j}$	$\hat{\xi}_j$
1	$\frac{1}{2}(r+s)$	$\frac{1}{2}(r+s)$
2	r	r
3	r	r
4	s	s
5	$\mu_\xi + \dfrac{1}{1+\rho}(s - \mu_\xi - \mu_\eta)$	$\mu_\xi + \dfrac{(1+\rho)\sigma_\xi^2}{(1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2}(s - \mu_\xi - \mu_\eta)$
6	$\mu_\xi + \dfrac{1}{1+\rho}(s - \mu_\xi - \mu_\eta - \mu_\omega)$	$\mu_\xi + \dfrac{(1+\rho)\sigma_\xi^2}{(1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2}(s - \mu_\xi - \mu_\eta - \mu_\omega)$

Within-class predictors We consider the predictors per class first. They are given in table 4, and are obtained after some straightforward derivations involving the inversion of some 2×2 matrices. Notice that in the cases involving a singular Ω the expressions for the predictors in the marginal rather than the conditional form have to be used, cf. the discussion in section 3. Notice also that for classes 5 and 6, the expressions for $\hat{\xi}_{\mathrm{U},j}$ can be obtained from the corresponding expressions for $\hat{\xi}_j$ by putting $\sigma_\eta^2 = \sigma_\omega^2 = 0$, i.e., by ignoring the measurement error variances in the computation.

Weighted predictors Moving down the rows of table 1, we obtain overall predictors by weighting the $\hat{\xi}_{\mathrm{U},j}$ and $\hat{\xi}_j$ with the probabilities π_j to obtain

$$\hat{\xi}_{\mathrm{U}} = ar + bs + c$$

with

$$
\begin{aligned}
a &= \pi_r(1 - \tfrac{1}{2}\pi_s) \\
b &= \tfrac{1}{2}\pi_r\pi_s + (1 - \pi_r)\left\{\pi_s + (1 - \pi_s)\frac{1}{1+\rho}\right\} \\
c &= (1 - \pi_r)(1 - \pi_s)\left\{\mu_\xi - \frac{1}{1+\rho}\left[\mu_\xi + (1 - \pi_s)(\mu_\eta + \pi_\omega\mu_\omega)\right]\right\}
\end{aligned}
$$

and

$$\hat{\xi} = ar + bs + c$$

with

$$
\begin{aligned}
a &= \pi_r(1 - \tfrac{1}{2}\pi_s) \\
b &= \tfrac{1}{2}\pi_r\pi_s + (1 - \pi_r)\Bigg\{\pi_s + (1 - \pi_s)\Bigg[(1 - \pi_\omega)\frac{(1+\rho)\sigma_\xi^2}{(1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2} \\
&\qquad + \pi_\omega\frac{(1+\rho)\sigma_\xi^2}{(1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2}\Bigg]\Bigg\} \\
c &= (1 - \pi_r)(1 - \pi_s)\Bigg\{\mu_\xi - (1 - \pi_\omega)\frac{(1+\rho)\sigma_\xi^2}{(1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2}(\mu_\xi + \mu_\eta) \\
&\qquad - \pi_\omega\frac{(1+\rho)\sigma_\xi^2}{(1+\rho)^2\sigma_\xi^2 + \sigma_\eta^2 + \sigma_\omega^2}(\mu_\xi + \mu_\eta + \mu_\omega)\Bigg\},
\end{aligned}
$$

after again some straightforward algebraic manipulations.

Posterior-weighted predictors We continue the discussion in section 5 and elaborate it for the present case. Consider the event $r = s$. If the observation is drawn from class 1, the conditional probability of observing this event is 1, whereas this conditional probability is 0 if the observation is drawn from any of the other five classes. Hence, the conditional probability that $j = 1$ given this event is 1: $p_1((r, r)') = 1$ for all r and $p_j((r, r)') = 0$ for $j = 2, \ldots, 6$. Conversely, using the same logic, $p_1((r, s)') = 0$ for all $s \neq r$. In the latter case,

$$p_j(y) = \frac{\pi_j f_j(y)}{\sum_{k=2}^{6} \pi_k f_k(y)}, \qquad j = 2, \ldots, 6.$$

It follows that

$$\begin{aligned}
\hat{\xi}^*_{\mathrm{U}} &= \begin{cases} \frac{1}{2}(r+s) = r = s & \text{if } r = s \\ \sum_{j=2}^{6} p_j(y)\,\hat{\xi}_{\mathrm{U},j} & \text{if } r \neq s \end{cases} \\
\hat{\xi}^* &= \begin{cases} \frac{1}{2}(r+s) = r = s & \text{if } r = s \\ \sum_{j=2}^{6} p_j(y)\,\hat{\xi}_j & \text{if } r \neq s. \end{cases}
\end{aligned} \tag{16}$$

Evidently, these expressions cannot be further elaborated.

System-wide predictors The last row of table 1 presents four predictors. As argued in section 4.3, expressions for the two unbiased predictors $\tilde{\xi}_{\mathrm{LU}}$ and $\tilde{\xi}_{\mathrm{U}}$ could not be found, so here we only elaborate the projection-based predictor $\tilde{\xi}_{\mathrm{L}}$ and the conditional expectation-based predictor $\tilde{\xi}$.

We consider $\tilde{\xi}_{\mathrm{L}}$ first. It requires the mean and covariance matrices of y and ξ. We weight the rows of table 2 with the π's to obtain

$$\begin{aligned}
\mu_y &= \begin{pmatrix} \pi_r\mu_\xi + (1-\pi_r)\mu_\zeta \\ \mu_\xi + (1-\pi_s)(\mu_\eta + \pi_\omega\mu_\omega) \end{pmatrix} \\
\Sigma_y &= \begin{pmatrix} \sigma_{rr} & \sigma_{rs} \\ \sigma_{rs} & \sigma_{ss} \end{pmatrix} \\
\Sigma_{\xi y} &= \sigma^2_\xi(\pi_r, 1 + (1-\pi_s)\rho)
\end{aligned}$$

where

$$\begin{aligned}
\sigma_{rr} &= \pi_r\sigma^2_\xi + (1-\pi_r)\sigma^2_\zeta + \pi_r(1-\pi_r)(\mu_\xi - \mu_\zeta)^2 \\
\sigma_{rs} &= \pi_r\left[\pi_s + (1-\pi_s)(1+\rho)\right]\sigma^2_\xi \\
\sigma_{ss} &= \left[\pi_s + (1-\pi_s)(1+\rho)^2\right]\sigma^2_\xi + (1-\pi_s)\sigma^2_\eta + (1-\pi_s)\pi_\omega\sigma^2_\omega \\
&\quad + (1-\pi_s)\left[\pi_s(\mu_\eta + \pi_\omega\mu_\omega)^2 + \pi_\omega(1-\pi_\omega)\mu^2_\omega\right].
\end{aligned}$$

Substitution in (15) yields, after some straightforward manipulations,

$$\tilde{\xi}_{\mathrm{L}} = ar + bs + c$$

with

$$\begin{aligned}
(a,b) &= \sigma^2_\xi(\pi_r, 1 + (1-\pi_s)\rho)\Sigma_y^{-1} \\
c &= \mu_\xi - a[\pi_r\mu_\xi + (1-\pi_r)\mu_\zeta] - b[\mu_\xi + (1-\pi_s)(\mu_\eta + \pi_\omega\mu_\omega)],
\end{aligned}$$

where we have not given an explicit expression for Σ_y^{-1}. We finally turn to $\tilde{\xi}$. As was already stated in section 4.3, $\tilde{\xi} = \hat{\xi}^*$. The formula for the latter has been given by (16).

Table 5: Parameter estimates.

	ξ	ζ	ω	η
μ	12.28	9.20	−0.31	−0.05
σ^2	0.51	3.27	1.52	0.01

Table 6: Predictions, with weights, $\hat{\xi}_{\mathrm{U},j}$ (or $\hat{\xi}_j$) $= ar + bs + c$.

			$\hat{\xi}_{\mathrm{U},j}$				$\hat{\xi}_j$		
Class (j)	π_j		a	b	c		a	b	c
1	.15		0.5	0.5	0		0.5	0.5	0
2	.69		1	0	0		1	0	0
3	.13		1	0	0		1	0	0
4	.01		0	1	0		0	1	0
5	.03		0	1.01	−0.11		0	0.99	0.12
6	.01		0	1.01	0.19		0	0.25	9.32
		$\hat{\xi}_{\mathrm{U}}$:	0.89	0.11	−0.00	$\hat{\xi}$:	0.89	0.11	0.05

6.4 Prediction

We now turn to the empirical outcomes of the prediction process. In essence we are after something in fact quite simple. We have two possibly conflicting figures on the same variable, and we want to combine the information from the two sources in an optimal way. We do so on the basis of a simple model, and of the estimation results from that model. These are from Kapteyn and Ypma (2007), where also an extensive discussion of these results can be found. Means and variances of the normal distributions are given in table 5. Further, $\pi_r = .96$, $\pi_s = .15$, $\pi_\omega = .16$, and $\rho = -.01$.

Table 6 contains a first series of prediction results. The top panel of the table presents results per class. The second column gives the class probabilities. The largest class is the second one, representing the case of measurement error, but no mismatch. The remaining columns present the numerical results corresponding with table 4. The predictors in that table have been rewritten as a linear function $ar + bs + c$ of r and s, the three weights a, b and c being given for $\hat{\xi}_{\mathrm{U},j}$ first and next for $\hat{\xi}_j$.

As was already apparent from table 4, the last two rows of the top panel are the most interesting ones. These represent the cases of a mismatch, hence the value from the register is discarded and only the value from the survey is used; and the latter suffers from measurement error. If there is no (further) contamination, the unbiased predictor is obtained from the survey figure by slightly overweighting it, and the unrestricted predictor is obtained by slightly downweighting it. If there is contamination, the results diverge strongly. The unrestricted predictor is predominantly a constant, plus some effect from the survey, whereas the unbiased predictor is very similar to the predictor obtained when there is no contamination.

The bottom panel of the table gives the results for the predictors $\hat{\xi}_{\mathrm{U},j}$ and $\hat{\xi}_j$ that are obtained by weighting the preceding rows by the prior probabilities given in the second column. Apart from the constant c, the results from the two predictors only differ from the third digit onwards, and imply a 89% weight to the register data and a 11% weight to the survey data.

The linear projection-based predictor $\hat{\xi}_{\mathrm{L}}$ can also be expressed as a linear combination of r and s:

$$\tilde{\xi}_{\mathrm{L}} = .22r + .55s + 2.85.$$

This is a striking result, with only very low weight given to the register data, vastly unlike the case with $\hat{\xi}_{\mathrm{U}}$ and $\hat{\xi}$.

These predictors are based on uniform weights across the (r, s) space. The predictors $\hat{\xi}^*_{\mathrm{U}}$ and $\tilde{\xi}$ are based on posterior weighting and cannot be expressed as a simple function of r and s since the weights now depend on $y = (r, s)'$ through the densities $p_j(y)$, cf. (14). In order to get a feeling for these predictors, we have computed the relative weight given to r (as compared to s) over the (r, s) space. The result for $\hat{\xi}^*_{\mathrm{U}}$ is given in figure 2 and the result for $\tilde{\xi}$ is given in figure 3. We see that both figures are quite similar. In the middle of each figure the predictor is (almost) fully dominated by r, but further from the mean of ξ ($\mu_\xi = 12.28$), more weight is attached to s. Remarkably, the relative weight seems to be rather insensitive to the value of s and largely depends on the value of r only.

In order to obtain an overall comparison of the performance of the five predictors we have computed their reliabilities. The reliability of a proxy is a measure for the precision with which it measures the latent variable of interest. When ξ is one-dimensional, the reliability of a proxy x is the squared correlation between ξ and x. Thus, the reliability of r is $\pi_r^2\sigma_\xi^2/\sigma_{rr}$ and the reliability of s is $[1 + (1 - \pi_s)\rho]^2\sigma_\xi^2/\sigma_{ss}$. The reliability φ^2 of a linear combination $x = ar + bs + c$ is

$$\varphi^2 = \frac{\left\{a\pi_r + b[1 + (1 - \pi_s)\rho]\right\}^2 \sigma_\xi^2}{a^2\sigma_{rr} + b^2\sigma_{ss} + 2ab\sigma_{rs}},$$

which, of course, does not depend on c. For posterior-weighted predictors, reliabilities cannot be computed analytically since they are nonlinear combinations of r and s. We have computed the reliabilities of those predictors by simply drawing a large sample (100,000 observations) from the model and computing the squared sample correlation between factor and predictor.

The results are given in table 7, for the five predictors considered throughout, plus r and s as a reference. We also report the mean squared errors of the predictors, MSE = E(predictor $-\ \xi)^2$, which are, of course, strongly negatively related to the reliabilities, although they are not perfect substitutes. Note, e.g., that MSE does depend on c. In addition, we present the bias E(predictor$-\xi$) and variance Var(predictor $-\ \xi$).

The table shows some striking results. The register data r, which in a sense could be interpreted as representing the true data, have a (squared) correlation with the true data as defined implicitly by the model of less than a half. Informally stated, r is a clear loser, which is very surprising.

Table 7: Precision of the predictors

predictor	reliability	MSE	bias	variance
r	.47	.54	−.13	.52
s	.69	.23	−.08	.22
$\hat{\xi}_{\mathrm{U}}$	.53	.43	−.12	.41
$\hat{\xi}$	.53	.43	−.12	.41
$\hat{\xi}^*_{\mathrm{U}}$	.97	.01	.00	.01
$\tilde{\xi}_{\mathrm{L}}$	.76	.12	.00	.12
$\tilde{\xi}$	.98	.01	.00	.01

Apparently a high probability of being exactly correct is not sufficient and the small probability of being an independent arbitrary drawing from a different distribution has a dramatic consequence for the quality of the indicator. The survey data s perform considerably better than r. The predictors obtained by prior-weighting per class perform poorly, but the predictor based on linear projection performs much better. However, both predictors that employ posterior weighting are nearly perfect, again, of course, against the background of the model postulated. We stress that these empirical findings are only meant to illustrate latent variable prediction in mixture models. We have taken the Kapteyn-Ypma model as given.

In the last two columns of table 7, we see that the biases of the first four predictors are sizeable, but the squared biases are negligible compared to the variances, so that the variances dominate the MSEs. It is also noteworthy that both r and s have negative bias. The bias of r is due to mismatch, whereas the bias of s is due to the negative means of the measurement error η and contamination ω. The estimated negative biases of r and s are identified on the basis of the $r = s$ cases, which are higher up on the earnings distribution (on average) than the other cases.

7 Discussion

In a sense, the paper has dealt with a very simple question. We are given two numerical values for the same quantity. What is the best guess as to the truth? As always, getting to a sensible answer involves some model building, necessarily based on some assumptions. The model in case was of the mixture factor analysis type, and we were hence drawn into the field of factor score prediction in mixture SEMs. It appeared that, to the best of our knowledge, this field had not yet been systematically explored, leading to the current paper.

Taking the standard factor score prediction literature as our point of departure, we systematically explored the options for constructing predictors for the mixture case. This produced five predictors for further consideration. Two of these are based on weighting the predictors per class by prior weights, and two more are based on posterior weights. One of these appears to be the predictor that is obtained as the conditional expectation of the latent variables given the indicators. A fifth predictor is simply the linear projection of the latent variables on the indicators.

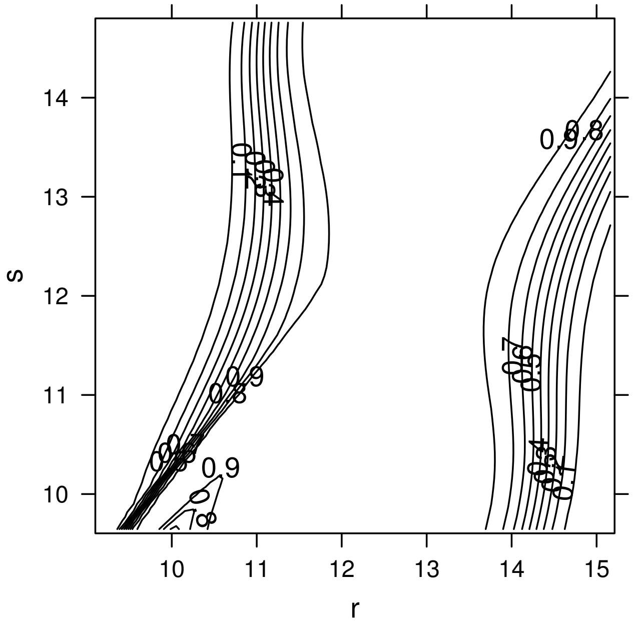

Figure 2: Relative weight of r in $\hat{\xi}^*_{\mathrm{U}}$.

We applied the thus filled toolkit to the case of Swedish earnings data, where two measurements were available per individual. For the cases where the two measurements coincided, we took the corresponding value as the true value, which hence was assumed known with certainty. To incorporate this element of certainty required an extension of the model formulation. This led to the consideration of labeled classes and degenerate distributions.

In the case study of the Swedish earnings data, we needed a yardstick for comparison of performance. For this, we took the reliability of the predictors. The major conclusion for this case is that posterior-weighted predictors perform extremely well, much better than the indicators themselves or the linear-projection based estimator. A topic for further research is to check the generality of this finding.

Appendix

A Single-class prediction with degenerate distributions

When degeneracies are present in the model, (9) is still correct, provided that in (7) and (8) Σ^{-1} is replaced by Σ^{+}, the Moore-Penrose generalized inverse of Σ. This encompasses the nondegenerate case, because $\Sigma^{+} = \Sigma^{-1}$ if Σ is nonsingular. The unrestricted minimum MSE predictor is $\hat{\xi}$ with

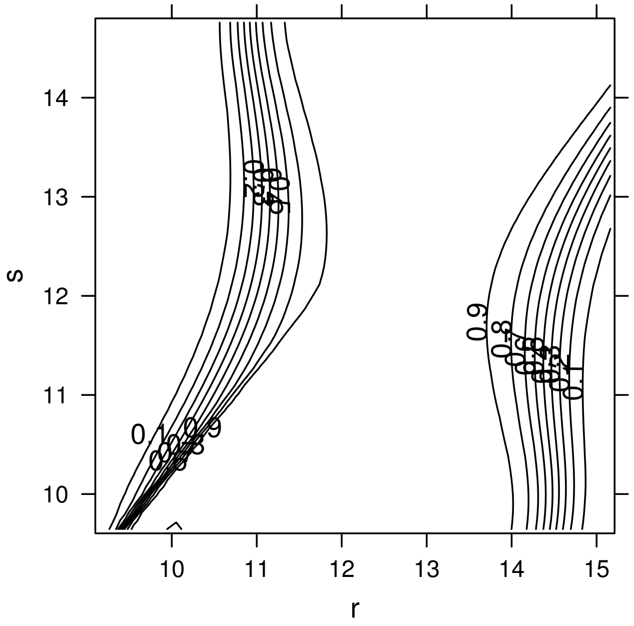

Figure 3: Relative weight of r in $\tilde{\xi}$.

this minor change in its definition.

An unbiased predictor may not exist when degeneracies are present. A trivial example is when $B = 0$ and Φ is nonzero. Then y contains no information about ξ and $\mathrm{E}(h(y) - \xi \mid \xi) = \mathrm{E}(h(y)) - \xi$, where $\mathrm{E}(h(y))$ does not depend on ξ. Hence, $\mathrm{E}(h(y) - \xi \mid \xi)$ does depend on ξ and, in particular, it cannot be zero irrespective of the true value of ξ.

The other extreme is obtained with $\Omega = 0$ and $B = I$, so that $y = \xi$, and thus y is an unbiased predictor of ξ with zero MSE. A necessary and sufficient condition for the existence of an unbiased predictor is that the matrix BQ has full column rank, where Q is the matrix whose columns are the eigenvectors corresponding to the nonzero eigenvalues of Φ. If $\Phi = 0$, an unbiased predictor $\hat{\xi}_{\mathrm{U}} = \kappa$ also trivially exists. In the following we will assume that $\Phi \neq 0$.

Using this condition and the condition that BQ has full column rank, we can adapt the analyses of Amemiya (1985, sections 1.4 and 6.1.5) to our notation and, if Φ is singular, extend them with the restriction $\xi = \kappa + Q\xi_1$ for some ξ_1, or, equivalently, $Q'_{\perp}\xi = Q'_{\perp}\kappa$. Here and in the following, if A is a $p \times q$ matrix of rank $r \leq q$, $A_{\perp}$ denotes a $p \times (p - r)$ matrix such that $A'_{\perp}A = 0$ and $A'_{\perp}A_{\perp} = I_{p-r}$, provided $r < p$. We can now distinguish the following cases:

1. Ω is nonsingular. Then $\hat{\xi}_{\mathrm{U}} = \kappa + Q(Q'B'\Omega^{-1}BQ)^{-1}Q'B'\Omega^{-1}(y - \mu)$. The MSE is nonzero in this case.

2. $\Omega = 0$. Then $\hat{\xi}_{\mathrm{U}} = \kappa + Q(Q'B'BQ)^{-1}Q'B'(y - \mu)$. This is equal to ξ with probability 1.

3. Ω is nonzero, but singular and $H'_\perp BQ$ has full column rank, where H is the matrix whose columns are the eigenvectors corresponding to the nonzero eigenvalues of Ω. Then $\hat{\xi}_{\mathrm{U}} = \kappa + Q(Q'B'H_\perp H'_\perp BQ)^{-1}Q'B'H_\perp H'_\perp(y-\mu)$. Again, this is equal to ξ with probability 1.

4. Ω is singular, $H'_\perp BQ$ does not have full column rank, and Q is square and nonsingular. Let $R \equiv B'H_\perp$. Then $R_\perp$ exists and is well-defined. Let $\kappa^* \equiv (R^+)'H'_\perp(y-\tau) = (H'_\perp B)^+ H'_\perp(y-\tau)$. Then
$$\hat{\xi}_{\mathrm{U}} = \kappa^* + R_\perp(R'_\perp B'\Omega^+ BR_\perp)^{-1}R'_\perp B'\Omega^+(y-\tau-B\kappa^*).$$
Now the MSE is nonzero.

5. Ω is singular, $H'_\perp BQ$ does not have full column rank, and Q is not square. Then $R \equiv (B'H_\perp, Q_\perp)$ and $R_\perp$ both exist and are well-defined. Let
$$\kappa^* \equiv (R^+)'\begin{pmatrix} H'_\perp(y-\tau) \\ Q'_\perp\kappa \end{pmatrix}.$$
Then $\hat{\xi}_{\mathrm{U}} = \kappa^* + R_\perp(R'_\perp B'\Omega^+ BR_\perp)^{-1}R'_\perp B'\Omega^+(y-\tau-B\kappa^*)$. The MSE is nonzero in this case as well.

These expressions for $\hat{\xi}_{\mathrm{U}}$ all give a best linear unbiased predictor (BLUP) of ξ, which is best unbiased (BUP) as well given the normality assumption for ε. However, other BLUPs may exist. Consider, for example, $\tau = 0$, $B = (1,1)'$, and $\Omega = 0$. Then the relevant expression above gives $\hat{\xi}_{\mathrm{U}} = \frac{1}{2}(y_1 + y_2)$, but $\xi^* \equiv ay_1 + (1-a)y_2$ is BLUP for every a. As long as the data are consistent with the model, such other BLUPs are necessarily equal to the expressions given here (with probability 1). However, in a mixture model, the weighted predictors using prior probabilities as weights may be different for different "BLUPs". We will not study this in more detail and only use the expressions given here.

B Posterior probabilities in mixture models with degenerate distributions

In a mixture of normals, if all the class-specific covariance matrices Σ_j are nonsingular, the posterior probabilities of the classes are given by (11). Note that this does not require Ω_j or Φ_j to be nonsingular or B_j to have full column rank. On the other hand, singularity of Σ_j necessarily implies singularity of Ω_j; so there is a relationship.

If Σ_j is singular, $(y \mid \text{class} = j)$ does not have a proper density and thus (11) is not well-defined. Hence, if this is the case for some or all of the classes in the model, the computation of the posterior probabilities has to be adapted. The basic principles are easiest to understand by looking at some simple examples.

1. Let there be two classes, $\mu_1 = 0$, $\Sigma_1 = 0$, $\mu_2 = 1$, $\Sigma_2 = 0$, and $0 < \pi_1 = 1 - \pi_2 < 1$. We have a mixture of two degenerate normal distributions. If an observation is drawn from class 1, it is equal to 0 with probability 1, whereas if an observation is drawn from class 2, it is equal to 1 with probability 1. Hence, if $y = 0$, it follows immediately (or from Bayes' rule for probabilities) that $p_1(y) = 1$ and $p_2(y) = 0$. Conversely, if $y = 1$, then $p_1(y) = 0$ and $p_2(y) = 1$. Other events have (prior) probability zero of occurring and can therefore be ignored.

2. Let there be two classes, $\mu_1 = 0$, $\Sigma_1 = 0$, $\mu_2 = 0$, $\Sigma_2 = 1$, and $0 < \pi_1 = 1 - \pi_2 < 1$. We have a mixture of a degenerate normal distribution and a nondegenerate one. Again, if an observation is drawn from class 1, it is equal to 0 with probability 1, whereas if an observation is drawn from class 2, it is unequal to 0 with probability 1. Hence, if $y = 0$, $p_1(y) = 1$ and $p_2(y) = 0$ and if $y \neq 0$, $p_1(y) = 0$ and $p_2(y) = 1$.

3. Let there be three classes, $\mu_1 = 0$, $\Sigma_1 = 0$, $\mu_2 = -1$, $\Sigma_2 = 1$, $\mu_3 = 1$, $\Sigma_3 = 1$, and π_1, π_2, and π_3 are all nonzero. We have a mixture of one degenerate normal distribution and two nondegenerate ones. Again, if an observation is drawn from class 1, it is equal to 0 with probability 1, whereas if an observation is drawn from classes 2 or 3, it is unequal to 0 with probability 1. Hence, if $y = 0$, we find that $p_1(y) = 1$ and $p_2(y) = p_3(y) = 0$. Conversely, if $y \neq 0$, $p_1(y) = 0$, but the observation could be from either class 2 or class 3. The posterior probability for class 2 can be written as

$$p_2(y) = \Pr(\text{class} = 2 \mid y, \text{class} \in \{2,3\}) \Pr(\text{class} \in \{2,3\} \mid y).$$

   If $y \neq 0$, the second factor is equal to 1. The first factor is simply (11) restricted to classes 2 and 3, with the "prior" probabilities π_j replaced by $\pi_{j|\{2,3\}} = \pi_j/(\pi_2 + \pi_3)$. But the denominator of this drops out of the resulting expression, so we can simply use the π_j themselves as well.

4. Let there be two classes, $\mu_1 = (-1, 0)'$, $\mu_2 = (1, 0)'$,

$$\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

   and $0 < \pi_1 = 1 - \pi_2 < 1$. We have a mixture of two degenerate normal distributions in two dimensions, so the density of y does not exist and we cannot use (11) directly. However, it is clear that the second dimension does not give any information about class membership. The marginal distribution of the first element of y, y_1, is a mixture in one dimension of two nondegenerate normal distributions with means -1 and 1, respectively, and both having variance 1. The prior probabilities of these are π_1 and π_2. Thus, the posterior probabilities are the posterior probabilities associated solely with the marginal distribution (and outcome) of y_1. These posterior probabilities are obtained by applying (11) to this one-dimensional problem.

To analyze the general case, let K_j be the matrix whose columns are the (orthonormal) eigenvectors of Σ_j corresponding to its nonzero eigenvalues. Then, with probability 1, if y is drawn from class j, it can be written as $y = \mu_j + K_j x_j$ for some random vector x_j, which has lower dimension than y if Σ_j is singular. In fact, x_j can be defined as $x_j = K_j'(y - \mu_j)$. Conditional on being drawn from class j, x_j is normally distributed with a positive definite covariance matrix. Now define the hyperplane $\mathcal{P}_j$ that consists of all points of the form $\mu_j + K_j x$, i.e.,

$$\mathcal{P}_j \equiv \{z \mid \exists x \text{ such that } z = \mu_j + K_j x\} = \{z \mid (I - K_j K_j')(z - \mu_j) = 0\}.$$

The dimension d_j of $\mathcal{P}_j$ is the number of columns of K_j, or, equivalently, the number of elements of x. The hyperplane $\mathcal{P}_j$ and the outcome y are called *inconsistent* (with each other) if $y \notin \mathcal{P}_j$.

The posterior probabilities $p_j(y)$ can now be computed as follows:

1. Given the outcome y, $p_j(y) = 0$ if $\mathcal{P}_j$ and y are inconsistent with each other, as in Examples 1–3 above. Eliminate these classes from further analysis.

2. Let $d_{\min}$ be the smallest dimension of the hyperplanes that are consistent with y, i.e., it is the minimum of the d_j for the classes that have not been eliminated in the previous step.

3. For all remaining classes, if $d_j > d_{\min}$, then $p_j(y) = 0$, as in Examples 2 and 3 above. Eliminate these classes from further analysis.

4. We now have a set of classes whose hyperplanes $\mathcal{P}_j$ are the same. (Mathematically, it is possible that there are intersecting hyperplanes and y lies on the intersection, so that the remaining $\mathcal{P}_j$ are not the same, but such an event has probability zero of occurring and can thus be safely ignored.) In general, their μ_j, Σ_j, and K_j will be different, but they are related through

$$y = \mu_j + K_j x = \mu_i + K_i w$$

for some x and w, from which we derive

$$w = K_i'(\mu_j - \mu_i + K_j x)$$

and $K_i' K_j$ must be nonsingular, so that there exists a nonsingular matrix Γ_{ij} such that $K_j = K_i \Gamma_{ij}$. Choose a point μ^* that lies on the hyperplane and a set of basis vectors of the hyperplane, collected in the matrix K^*. Obvious candidates are one of the μ_j's in the current set and the corresponding K_j. The posterior probabilities do not depend on this choice. In Example 4 above, we could choose $\mu^* = (-1, 0)'$ and $K^* = (1, 0)'$.

5. Compute $y^* \equiv K^{*\prime}(y - \mu^*)$ This is distributed as a mixture of normal distributions with probabilities proportional to π_j for the classes still under consideration and corresponding means and covariance matrices equal to $\mu_j^* = K^{*\prime}(\mu_j - \mu^*)$ and $\Sigma_j^* = K^{*\prime}\Sigma_j K^*$, respectively. In Example 4 above, with $\mu^* = (-1, 0)'$ and $K^* = (1, 0)'$, we have $y^* = y_1 + 1$, $\mu_1^* = 0$, $\mu_2^* = 2$, and $\Sigma_1^* = \Sigma_2^* = 1$.

6. By construction, Σ_j^* is nonsingular for the classes under consideration, and thus the posterior probabilities for the classes still under consideration are found by applying (11) to the reduced problem, i.e., where $f_j(y)$ is replaced by the density function of the $\mathcal{N}_{d_{\min}}(\mu_j^*, \Sigma_j^*)$ distribution evaluated in y^*.

References

Amemiya, T. (1985), *Advanced econometrics*, Harvard University Press, Cambridge, MA.

Arminger, G., P. Stein, and J. Wittenberg (1999), “Mixtures of conditional mean- and covariance-structure models”, *Psychometrika*, **64**, 475–494.

Arminger, G., J. Wittenberg, and A. Schepers (1996), *MECOSA 3: Mean and covariance structure analysis*, Additive, Friedrichsdorf, Germany.

Bartlett, M. S. (1937), “The statistical conception of mental factors”, *British Journal of Psychology*, **28**, 97–104.

Bound, J. and A. B. Krueger (1991), “The extent of measurement error in longitudinal earnings data: do two wrongs make a right?”, *Journal of Labor Economics*, **9**, 1–24.

Carroll, J. B. (1993), *Human cognitive abilities*, Cambridge University Press, Cambridge.

De Haan, J., E. Leertouwer, E. Meijer, and T. J. Wansbeek (2003), “Measuring central bank independence: a latent variables approach”, *Scottish Journal of Political Economy*, **50**, 326–340.

Jedidi, K., H. S. Jagpal, and W. S. DeSarbo (1997a), “STEMM: a general finite mixture structural equation model”, *Journal of Classification*, **14**, 23–50.

Jedidi, K., H. S. Jagpal, and W. S. DeSarbo (1997b), “Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity”, *Marketing Science*, **16**, 39–59.

Kapteyn, A. and T. J. Wansbeek (1984), “Errors in variables: Consistent Adjusted Least Squares (CALS) estimation”, *Communications in Statistics - Theory and Methods*, **13**, 1811–1837.

Kapteyn, A. and J. Y. Ypma (2007), “Measurement error and misclassification: A comparison of survey and administrative data”, *Journal of Labor Economics*, **25**, 513–551.

Lee, S.-Y. and X.-Y. Song (2003), “Maximum likelihood estimation and model comparison for mixtures of structural equation models with ignorable missing data”, *Journal of Classification*, **20**, 221–255.

Lubke, G. H. and B. O. Muthén (2005), “Investigating population heterogeneity with factor mixture models”, *Psychological Methods*, **10**, 21–39.

Meijer, E. and T. J. Wansbeek (1999), “Quadratic factor score prediction”, *Psychometrika*, **64**, 495–508.

Muthén, B. O. (2004), *Mplus technical appendices*, Muthén & Muthén, Los Angeles.

Phillips, R. F. (2003), “Estimation of a stratified error-components model”, *International Economic Review*, **44**, 501–521.

Schneeweiss, H. and C.-L. Cheng (2006), “Bias of the structural quasi-score estimator of a measurement error model under misspecification of the regressor distribution”, *Journal of Multivariate Analysis*, **97**, 455–473.

Song, X.-Y. and S.-Y. Lee (2004), “Local influence analysis for mixture of structural equation models”, *Journal of Classification*, **21**, 111–137.

Ten Berge, J. M. F., W. P. Krijnen, T. J. Wansbeek, and A. Shapiro (1999), “Some new results on correlation preserving factor scores prediction methods”, *Linear Algebra and its Applications*, **289**, 311–318.

Wansbeek, T. J. and E. Meijer (2000), *Measurement error and latent variables in econometrics*, North-Holland, Amsterdam.

Wedel, M. and W. A. Kamakura (2000), *Market segmentation: conceptual and methodological foundations*, second edition, Kluwer, Boston.

Wedel, M., W. A. Kamakura, N. Arora, A. C. Bemmaor, J. Chiang, T. Elrod, R. Johnson, P. Lenk, S. Neslin, and C. S. Poulsen (1999), “Discrete and continuous representation of heterogeneity”, *Marketing Letters*, **10**, 219–232.

Yung, Y.-F. (1997), “Finite mixtures in confirmatory factor-analysis models”, *Psychometrika*, **62**, 297–330.

Zhu, H.-T. and S.-Y. Lee (2001), “A Bayesian analysis of finite mixtures in the LISREL model”, *Psychometrika*, **66**, 133–152.