PSYCHOMETRIKA—VOL. 76, NO. 4, 564–583
OCTOBER 2011
DOI: 10.1007/S11336-011-9230-8

ON THE RELATION BETWEEN THE LINEAR FACTOR MODEL AND THE LATENT PROFILE MODEL

PETER F. HALPIN, CONOR V. DOLAN, RAOUL P.P.P. GRASMAN, AND PAUL DE BOECK

UNIVERSITY OF AMSTERDAM

The relationship between linear factor models and latent profile models is addressed within the context of maximum likelihood estimation based on the joint distribution of the manifest variables. Although the two models are well known to imply equivalent covariance decompositions, in general they do not yield equivalent estimates of the unconditional covariances. In particular, a 2-class latent profile model with Gaussian components underestimates the observed covariances but not the variances, when the data are consistent with a unidimensional Gaussian factor model. In explanation of this phenomenon we provide some results relating the unconditional covariances to the goodness of fit of the latent profile model, and to its excess multivariate kurtosis. The analysis also leads to some useful parameter restrictions related to symmetry.

Key words: linear factor model, latent profile model, maximum likelihood, Kullback–Leibler divergence, multivariate kurtosis.

1. Introduction

A well known result in psychometric modeling is the relationship between the second-order moments of the J-dimensional linear factor model (LF-J) and those of the K-class latent profile model (LP-K) (see, e.g., Bartholomew & Knott, 1999; McDonald, 1967; Molenaar & von Eye, 1997). The models imply equivalent decompositions of the unconditional covariance matrix when $K = J + 1$. This has direct consequences for model misspecification in factor analysis, and in particular it means that LF-J can be mistakenly fitted to covariances generated by LP-$(J + 1)$. This misspecification problem has led to some discussion about how to best distinguish the two models (e.g., Halpin & Maraun, 2010; Lubke & Neale, 2006; Steinley & McDonald, 2007). We consider this topic further here.

Although their unconditional second-order moments are equivalent, a Gaussian factor model cannot reproduce many of the other properties associated with mixture models (e.g., skewness, multimodality). It is therefore reasonable to expect that maximum likelihood estimation based on unconditional densities (*raw data ML*) will often serve to address the model misspecification problem described above. Here the simulation studies of Lubke and Neale (2006) have shown that, when LP-$(J + 1)$ is used to generate data, the models are quite reliably distinguished by fit statistics based on raw data ML (e.g., AIC, BIC). As one would expect, latent profile models with larger class separations are more readily distinguished from Gaussian factor models.

There remains the question of whether and under what conditions LP-$(J + 1)$ can reproduce the unconditional density of LF-J. Here we are primarily concerned with mistakenly fitting LP-$(J + 1)$ when the data are generated by LF-J, which is the reverse of the model misspecification problem considered above. Lubke and Neale (2006) again provided insight into the problem, finding that information criteria also reliably distinguish the two models when data are generated under LF-J. However, unlike the case where a Gaussian model is fitted to a well separated

Requests for reprints should be sent to Peter F. Halpin, Psychological Methods, University of Amsterdam, Roetersstraat 15, 5th floor, 1018 WB Amsterdam, The Netherlands. E-mail: p.f.halpin@uva.nl

© 2011 The Psychometric Society

mixture, it is not clear precisely why such a result should occur. This paper adds to the findings of Lubke and Neale by showing that when LP-$(J+1)$ is fitted to data generated by LF-J, larger observed covariances lead to worse fit.

In addition to overall goodness of fit, Loken and Molenaar (2007) considered recovery of observed moments. They described a situation in which covariances that are known to fit a factor model are poorly reproduced by the latent profile model. This result is surprising, since it appears to contradict the equivalence between the second-order moments of the models. Here we also consider recovery of observed moments, showing that larger covariances imply more excess negative kurtosis in the latent profile model. This explains the apparent contradiction and provides an intuitive explanation of our findings concerning goodness of fit.

In general, the relation between factor analysis and latent profile analysis is not straightforward when considered from a raw data ML perspective, rather than from the perspective of covariance-based estimation. Therefore we provide some empirical and analytic results to describe further this relationship. The analysis requires that we restrict consideration to the case of $K = J + 1 = 2$, but the main results also apply to more general covariance structures.

The first section of this paper describes an empirical example from aptitude testing. The purpose of the example is to address a case in which LF-1 fits the data and consider the performance of LP-2. The following three observations are made: (a) LP-2 underestimates the observed covariances but not the variances, and the degree of underestimation is increasing in the observed covariances; (b) the excess kurtosis of the LP-2 bivariate densities is decreasing (i.e., more negative) in the observed covariances; (c) the LP-2 bivariate distributions are symmetrical about their unconditional means. The first two findings describe how LP-2 fails to fit data that are compatible with LF-1, while the observed symmetry implies that both models agree at all odd central moments.

The second section provides some analytic results by which to interpret and generalize these three empirical observations. We begin by discussing parameter constraints on LP-2 that are consistent with the observed symmetry. It is shown that the minimum Kullback–Leibler (KL) divergence of LP-2 from LF-1 is obtained when the unconditional density of LP-2 is symmetrical. These constraints are of interest in their own right, and they also facilitate subsequent analysis by allowing for a simplified parameterization of LP-2. Using this parameterization we derive an upper bound on the KL divergence of the two models as well as an expression for the excess multivariate kurtosis of LP-2. The former is seen to be increasing in the factor loadings of LF-1, while the latter is decreasing (i.e., becomes more negative). The mathematical details are collected in the Appendix.

The analytic results are all asymptotic in the sense that they concern population quantities and do not address finite samples. Therefore we also include a small simulation that illustrates the same conclusions. Here it is found that the models are well distinguished by their likelihoods at reasonable sample sizes ($N \geq 200$) so long as the observed correlations are not too small (e.g., $\rho \geq 0.3$). Stable estimation of the multivariate kurtosis requires larger samples.

Before proceeding the models of interest are specified and some remarks are made about their unconditional second-order moments.

1.1. Model Specification

We assume that the continuous latent distributions of both models are Gaussian and write $f_X(x; \mu_X, \Sigma_X)$ to denote the density of a Gaussian random variable $X \in \mathbb{R}^P$ with mean vector μ_X having elements $\mu_{X_i} = \mu_i$ and covariance matrix Σ_X having elements $\sigma_{X_i X_j} = \sigma_{ij}$, for $i, j \in \{1, \ldots, P\}$.

Using this notation the unidimensional factor model is given as

$$f_X(x; \gamma, C) = \int_{-\infty}^{\infty} f_{X|Y}(x; \Lambda y + \gamma, \Psi) f_Y(y; 0, 1)\, dy. \tag{1}$$

$Y \in \mathbb{R}$ is the unidimensional factor and its scale is fixed by setting $\mu_Y = 0$ and $\sigma_Y^2 = 1$. The vector Λ contains the factor loadings and the error covariance matrix Ψ is assumed to be positive definite. We also assume that Ψ is diagonal, although this serves mainly to simplify the discussion of excess kurtosis; more general assumptions are considered in the appendix. The vector of intercepts γ trivially reproduces the unconditional means. The present parameterization implies that $C = \Lambda\Lambda' + \Psi$, which is a non-trivial assertion about unconditional covariances. All other properties of f_X are implicit in the requirement that latent distributions, and hence f_X, are Gaussian. The model has a total of $3P$ free parameters. Bartholomew and Knott (1999) provide further discussion of Gaussian factor models.

The two-class latent profile model can be defined as a convex combination of Gaussian distributions,

$$g_X^{(2)}(x) = \eta_1 g_1(x; \mu_1, \Theta_1) + \eta_2 g_2(x; \mu_2, \Theta_2) \tag{2}$$

with $\eta_k \in [0, 1]$ and $\eta_k = 1 - \eta_l$, for $k, l \in \{1, 2\}$. The model is formally identified by requiring that $\eta_k \notin \{0, 1\}$ and that $\mu_1 \neq \mu_2$ (see, e.g., Frühwirth-Schnatter, 2006; McLachlan & Peel, 2000). For convenience the kth component of the mixture is denoted as g_k instead of invoking a discrete latent variable Z and treating the component distributions as conditioned on $Z = z_k$. The notation $g_X^{(2)}$ serves to emphasize that the unconditional density is not Gaussian. In parallel with LF-1 it is assumed that the Θ_k are positive definite and diagonal, although the latter is not always essential. The model has a total of $4P + 1$ free parameters.

We now note the equivalence between the unconditional second-order moments of LF-1 and LP-2 and show the underidentification of LP-2 at the level of observed covariances. We assume throughout that P is fixed. These results are contrasted with those to be presented in the remainder of this article.

Let $\mathcal{C}$ denote the set of all LF-1 covariance matrices and $\mathcal{D}$ denote the set of LP-2 covariance matrices. We show that $\mathcal{C} = \mathcal{D}$ by showing $\mathcal{C} \subset \mathcal{D}$ and $\mathcal{D} \subset \mathcal{C}$.

It has already been seen that an arbitrary member of $\mathcal{C}$ can be written as $C = \Lambda\Lambda' + \Psi$. For all $D \in \mathcal{D}$ we have

$$D = \sum_k \eta_k D_k = \sum_k \eta_k (\delta_k \delta_k' + \Theta_k) \tag{3}$$

with $\delta_k = \mu_k - \sum_k \eta_k \mu_k$ for $k \in \{1, 2\}$. Setting $\delta_1 = -\delta_2 = \Lambda$ and $\Theta_1 = \Theta_2 = \Psi$ in Equation (3) yields $D = \Lambda\Lambda' + \Psi$ and hence $\mathcal{C} \subset \mathcal{D}$.

The inclusion of $\mathcal{D}$ in $\mathcal{C}$ is established by observing that

$$\eta_1 \delta_1 \delta_1' + \eta_2 \delta_2 \delta_2' = \eta_1 \eta_2 (\delta_1 - \delta_2)(\delta_1 - \delta_2)' \tag{4}$$

and that $\Theta^* = \eta_1 \Theta_1 + \eta_2 \Theta_2$ is positive definite and diagonal. Then writing $\Lambda = (\eta_1 \eta_2)^{1/2} \times (\delta_1 - \delta_2)$ and $\Psi = \Theta^*$ for an arbitrary matrix D implies that $\mathcal{D} \subset \mathcal{C}$.

The foregoing argument shows that $\mathcal{C} = \mathcal{D}$. Because the factor model is identified at the level of second-order moments (up to the sign of the factor loadings; see, e.g., Bollen, 1989, Chap. 7), we can conclude that there are just as many realizations of LF-1 as are there members of $\mathcal{C}$. However, for each realization of LF-1 there are many realizations of LP-2 that produce numerically identical covariances. This can be seen by letting $\Lambda\Lambda' + \Psi$ be any fixed matrix in $\mathcal{C}$ and choosing an arbitrary value of $\eta_1 \in (0, 1)$. Then it is always possible to find values of δ_k and Θ_k such that $\Lambda\Lambda' + \Psi = \sum_k \eta_k (\delta_k \delta_k' + \Theta_k)$. This is verified by setting

$$\delta_1 = \frac{1}{\sqrt{\eta_1 \eta_2}} \Lambda + \delta_2, \qquad \theta_{1i}^2 = \frac{1}{\eta_1} \left(\psi_i^2 - \eta_2 \theta_{2i}^2 \right) \quad \text{and} \quad \theta_{2i}^2 > \frac{1}{\eta_2} \psi_i^2 \tag{5}$$

for $i \in \{1, \ldots, P\}$. Note that in Equation (5) and throughout we use the notation $\Psi = \text{diag}(\psi_1^2, \ldots, \psi_P^2)$ and $\Theta_k = \text{diag}(\theta_{k1}^2, \ldots, \theta_{kP}^2)$.

These results on moments are standard fare; LF-1 and LP-2 yield numerically identical covariance matrices, but LP-2 is not identified at this level. We reiterate them here because quite different conclusions are reached through a raw data ML perspective. In particular, the following section provides a empirical example in which LF-1 reproduces the observed covariances much better than LP-2. Moreover, a highly constrained parameterization of LP-2 occurs when it is fitted to the data. Like the factor model, this parameterization has $3P$ free parameters and is identified at the level of observed covariances. Thus a raw data ML perspective leads to the opposite conclusions as those outlined above: LP-2 is identified at the level of second-order moments but it does not, in general, yield numerically identical covariances as LF-1. In the second half of the paper we explain these findings analytically.

2. Empirical Comparison of LF-1 and LP-2

This section addresses an example from aptitude testing. The purpose of the example is to consider a case in which LF-1 fits the data quite well but LP-2 does not, and thereby to illustrate some relations between the models. The data were obtained from the 1997 Armed Services Vocational Aptitude Battery (ASVAB; see Segall, 2004). The battery was administered to a total of $n + 1 = N = 7127$ youths between ages 12 and 18 as part of the National Longitudinal Survey of Youth 1997. We made use of $P = 4$ subtests, namely Arithmetic Reasoning (AR), Mathematics Knowledge (MK), Paragraph Comprehension (PC), and Assembling Objects (AO). The data were standardized for ease of interpretation. Visual inspection of the univariate distributions did not reveal incompatibility with the assumptions of normality, and Table 1 shows that the univariate skewness and excess kurtosis deviated only slightly from zero. Nonetheless, normal scores were computed using Tukey's method to increase agreement with the distributional assumptions of LF-1. All analyses were carried out on the standardized normal scores. Table 2 reports the correlations of the transformed scales.

Both models were estimated by optimizing the raw data log-likelihood, and we denote the optimized value by l. The latent profile model was estimated using 200 sets of random starting

TABLE 1.
Univariate skew and kurtosis of ASVAB subscales.

	AR	MK	PC	AO
Skewness (SE = 0.03)	−0.373	−0.067	−0.119	0.089
Kurtosis (SE = 0.06)	−0.065	−0.511	−0.799	−0.855

Note: ASVAB = Armed Services Vocational Aptitude Battery; AR = Arithmetic Reasoning; MK = Mathematics Knowledge; PC = Paragraph Comprehension; AO = Assembling Objects; SE = standard error.

TABLE 2.
Correlations between normalized ASVAB subscales.

	AR	MK	PC
MK	0.809		
AC	0.745	0.735	
AO	0.638	0.631	0.609

Note: ASVAB = Armed Services Vocational Aptitude Battery; AR = Arithmetic Reasoning; MK = Mathematics Knowledge; PC = Paragraph Comprehension; AO = Assembling Objects.

values, and after 10 expectation-maximization (EM) iterations the 20 best values of l were selected and iterated until convergence. The convergence criterion was set to a difference of less than 1e–6 in successive values of l on the M-step. The 20 starting values that ran to completion all converged within 500 iterations to the same maximum.

The relative fit of the two models favored LF-1 ($-2l_{\text{LF-1}} = 62248.03$, 12 free parameters) over LP-2 ($-2l_{\text{LP-2}} = 69371.74$, 17 free parameters) by any of the usual likelihood-based criteria for non-nested models. That the models are well distinguished by their likelihoods is consistent with the results reported by Lubke and Neale (2006).

It is also insightful to compare the goodness of fit of the second-order moments. As discussed above, LP-2 can exactly reproduce any population covariance matrix implied by LF-1, and it would therefore be natural to expect that both models fit the second-order moments equally well. We computed the ML fit function (FML) from structural equation modeling (SEM; see Bollen, 1989) using the raw data ML estimates of both models. For LF-1, the fit was acceptable (nFML $= 24.02$; $df = 2$; RMSEA $= 0.039$; 90% CI $= [0.026, 0.054]$). However, the fit of LP-2 to the second-order moments was remarkably worse (nFML $= 3475.80$). Note that $4P + 1 > P(P + 1)/2$ if $P < 8$, so the degrees of freedom for the (unidentified) LP-2 model are negative and only its raw fit index is available. Nonetheless, it may be concluded that the observed covariances were well reproduced by LF-1, but not by LP-2. A similar finding was reported by Loken and Molenaar (2007).

We now consider the three points mentioned in the introduction. The notation $S = \{s_{ij}\}$ denotes the sample covariance matrix and $S_{(M)} = \{s_{ij(M)}\}$ denotes the unconditional covariance matrix computed from the parameter estimates of the model M.

2.1. *Recovery of Observed Covariances*

Figure 1 shows that there was a systematic pattern in the misfit of LP-2 to the observed covariances. The figure plots the sample covariances against those computed from the fitted LF-1 and LP-2 models. The solid diagonal line represents perfect agreement between the observed and model-implied covariances. The residuals of LP-2, $r_{ij} = s_{ij} - s_{ij\,(\text{LP-2})}$, are also shown. The dotted line shows the simple linear regression of the r_{ij} on the s_{ij} ($R^2 = 0.99, b = 0.65$). The linear trend makes clear that the degree to which LP-2 underestimated the observed covariances increased with the observed covariances. In terms of the parameter estimates of the two models this implies $|\hat{\lambda}_i| - \sum_k \hat{\eta}_k |\hat{\delta}_{ki}| > 0$ for all i.

We also note that both models reproduced the univariate second-order moments reasonably well, with $\max\{|s_i^2 - s_{i(\text{LP-2})}^2|\} = 5.5\text{e-}4$ and $\max\{|s_i^2 - s_{i(\text{LF-1})}^2|\} = 5.2\text{e-}4$. In combination with the underestimated covariances, this implies that $\hat{\psi}_i^2 - \sum_k \hat{\eta}_k \hat{\theta}_{ki}^2 < 0$. Thus the two models recover the observed variances equally well, but they partition it differently: LP-2 attributes more of the variance to error (i.e., within-class sources).

In summary, LP-2 underestimated the observed second-order moments at the bivariate margins but not the univariate margins. Moreover, the degree of underestimation was increasing in the observed covariances. In the second half of this paper this finding is considered analytically. In particular we show that, when LF-1 is the data-generating model, increasing the overall magnitude of the observed covariances leads to worse fit of LP-2 in terms of raw data ML.

In order to explain the misfit of the bivariate margins we sought more information about the moment structure of LP-2. Figure 2 shows a contour plot of the distribution of AR and MK implied under the estimated LP-2 model. The dashed lines in the plots are the principal components of the implied covariance matrix and the heavy dots represent the conditional means. Parameter estimates and contour values are given in the caption. The following two subsections describe two properties of Figure 2 that were shared by all six of the estimated LP-2 bivariate distributions, namely bimodality and symmetry. We find that (a) excess kurtosis of the LP-2

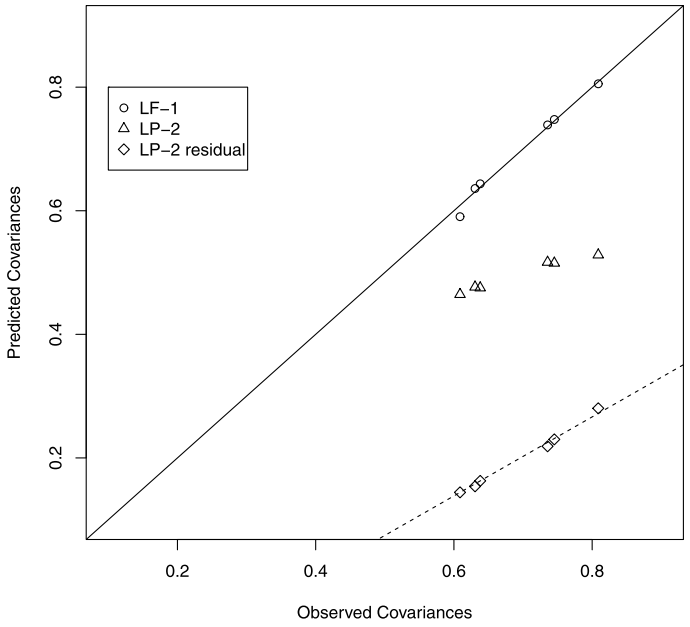


FIGURE 1.
Predicted versus observed covariances for the Armed Services Vocational Aptitude Battery example. LF-1 = factor model; LP-2 = latent profile model. The *dotted line* is the linear regression of the residuals of the LP-2 covariances on the observed covariances.

distributions became more negative as the observed covariances increased, and (b) an LP-2 model that is constrained to be symmetrical provides a more parsimonious explanation of the ASVAB data. We relate these finding to the observed misfit of the second-order moments.

2.2. *Excess Kurtosis*

One striking feature of Figure 2 is that the distribution is clearly bimodal. This graphical observation was confirmed analytically using the "pi-function" of Ray and Lindsay (2005). There is an intuitive explanation of this observed bimodality in terms of the model-implied covariances. Equation (4) implies that $\sigma_{ij} = \eta_1\eta_2(\delta_{1i} - \delta_{2i})(\delta_{1j} - \delta_{2j})$. Thus, for fixed mixing proportions, larger model-implied covariances imply larger class separations. When the class separation is sufficiently large relative to the within-class variances, one would anticipate bimodality such as that seen in Figure 2.

Although modality is an intuitive descriptor, it is not useful for distinguishing among the estimated LP-2 bivariate distributions since, as noted, these were all bimodal. It would therefore be convenient to adopt some other measure of this observed misfit between LP-2 and LF-1. The location of the modes of LP-2 is an obvious candidate, although we are not aware of any method for analytically describing these locations in terms of the model parameters. Similarly, there are various measures of class separation available (see Frühwirth-Schnatter, 2006, Chap. 1), although these have been related to observed modality only under less general models than LP-2 (e.g., univariate and homoscedastic mixtures). We therefore turn to multivariate kurtosis. Again there are many choices, and in particular we note that Mardia's (1970) statistic does not lend itself to defining a model-implied quantity under LP-2. However, as shown in the appendix, the excess kurtosis tensor of LP-2 in comparison to LF-1 is readily derived. We employed the average over the elements of this tensor, K, as a summary measure of discrepancy between the kurtosis of LF-1 and LP-2. In the following section we show that all of the terms in K are non-positive,

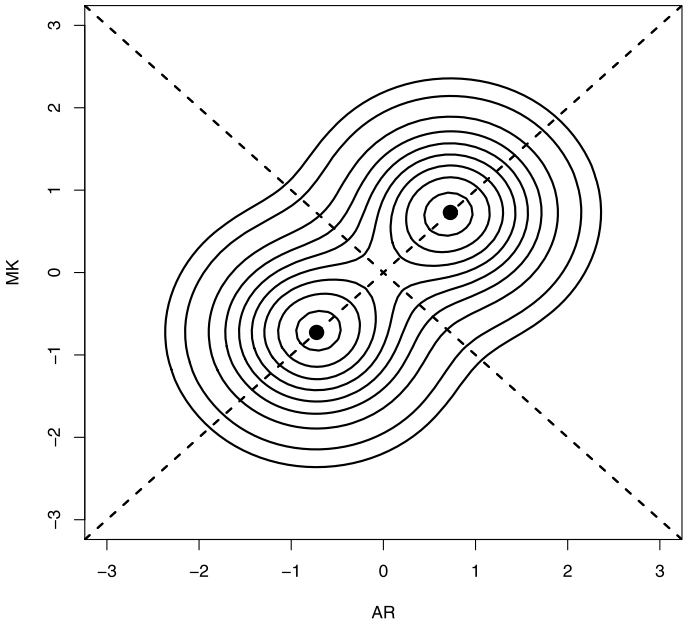


FIGURE 2.
Bivariate density contour plots under LP-2. AR = Arithmetic Reasoning; MK = Mathematics Knowledge. Right/upper component: centroid = (0.727, 0.729), variances = (0.470, 0.467). Left/lower component: centroid = (−0.725, −0.727), variances = (0.476, 0.473). Contours from outer to inner: 0.01, then 0.02 to 0.16 in increments of 0.02.

and so it is numerically equivalent to more usual measures of discrepancy such as the mean square error or absolute deviation, while also having the advantage of showing the direction of the excess (i.e., negative).

Figure 3 shows that the excess kurtosis of the estimated LP-2 bivariate distributions became more negative as the observed covariances increased. The dotted line shows the linear regression of $\hat{K}$ on the s_{ij} ($R^2 = 0.91, b = 0.63$). Thus, as with the covariances of LP-2, its kurtosis was also underestimated more severely when the observed covariances were larger. This leads us to consider the following generalization in the second half of this paper: when LF-1 is the data-generating model, increasing the overall magnitude of the observed covariances leads to more excess negative kurtosis in LP-2. It was also the case that $\hat{K}$ decreased with $s_{ij\,(\text{LP-2})}$. This is again intuitive from Equation (4), and we demonstrate this intuition analytically in Equation (11) below.

2.3. Symmetry

It was noted above that all of the estimated LP-2 bivariate distributions were markedly symmetrical about both the first and the second principal components. The parameter constraints suggested by the observed symmetry are homoscedasticity across the components and equality of the mixing proportions; the latter is equivalent to $\delta_1 = -\delta_2$. These parameter restrictions imply that LP-2 is symmetric about its unconditional mean (i.e., $g_X^{(2)}(x-\mu) = g_X^{(2)}(-x+\mu)$). We refer to an LP-2 model with these constraints as a symmetric latent profile model (SLP).

The fit indices presented in Table 3 show that SLP-2 was favored over LP-2 for ASVAB data. Here it should be noted that traditional tests of the parameter constraints are not available in the present context. As described earlier, the ASVAB data were fitted much better by LF-1 than LP-2, and therefore LP-2 is non-trivially misspecified. This implies that the regularity conditions

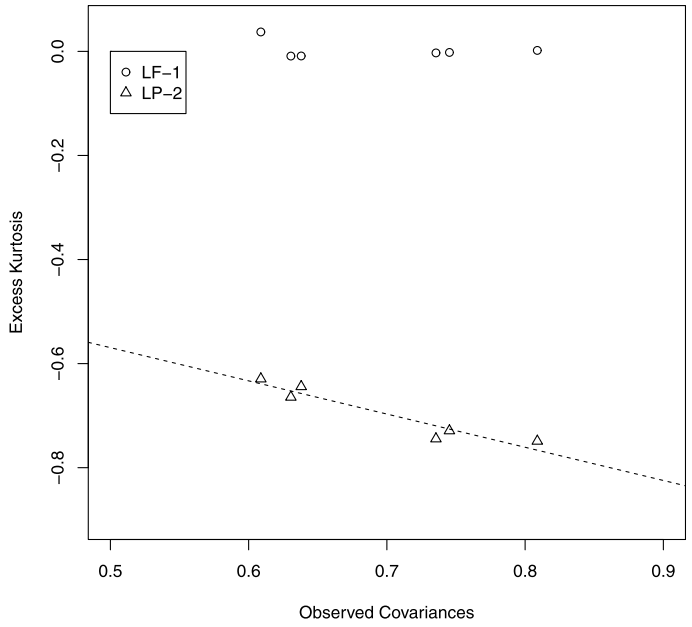


FIGURE 3.
Averaged, model-implied kurtosis versus observed covariances for the bivariate distributions of the Armed Services Vocational Aptitude Battery example. LF-1 = factor model; LP-2 = latent profile model. The *dotted line* is the linear regression of the LP-2 average kurtoses on the observed covariances.

TABLE 3.
Comparison of LP-2 and SLP-2 for the ASVAB data.

Model	AIC	BIC	nBIC
LP-2	69405	69522	69468
SLP-2	69392	69447	69422

Note: LP-2 = 2 class latent profile model; SLP-2 = symmetry constrained LP-2; ASVAB = Armed Services Vocational Aptitude Battery; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; nBIC = Sample size adjusted BIC.

for the usual asymptotic tests of LP-2 against SLP-2 do not hold. We discuss this point further in the data simulation and conclusion.

There are two main implications of accepting the SLP-2 constraints for the ASVAB data. Firstly, the constraints imply that unconditional distributions of the fitted LP-2 and LF-1 models share all their odd moments, and therefore any discrepancies in their goodness of fit cannot be described in terms of these moments. Secondly, it suggests that the SLP-2 constraints are a valid parameterization of LP-2 when the data-generating distribution is LF-1. We begin the following section by proving this claim.

2.4. Summary

The present section has described the misfit of LP-2 to data that are compatible with LF-1. At the level of second-order moments, LP-2 underestimated the observed covariances, although the variances were recovered well. This finding apparently contradicts previous results relating the covariances of the two models. However, examination of the model-implied bivariate distributions of LP-2 showed that larger covariances also corresponded to more excess negative kurtosis.

This indicates that a consideration of second-order moments alone can provide an incomplete and potentially misleading picture of the relation between the two models.

On the basis of the ASVAB example we proposed the following two generalizations about the relation between LF-1 and LP-2. When the data-generating model is LF-1, increasing the overall magnitude of the observed covariances will lead to (a) worse overall goodness of fit of LP-2 and (b) more excess negative kurtosis of LP-2. These claims are the focus of the following section.

3. Some Results on the Misfit Between LF-1 and LP-2

This section relates the goodness of fit of LP-2 and its excess negative kurtosis to the observed covariances when LF-1 is the data-generating model. We derive an upper bound on the KL divergence of LP-2 from LF-1 and show that this is increasing in the factor loadings, given mild constraints on the error variances of both models. Next we derive the average excess multivariate kurtosis of LP-2 compared to LF-1 and show that this is decreasing in the factor loadings. This section ends with a small simulation study that illustrates these same conclusions.

Before proceeding, we consider the SLP-2 parameter constraints. Theorem 1 shows that the constraints are valid when the data-generating model is LF-1. We assume the regularity conditions for the existence and consistency of the maximum likelihood estimator under model misspecification (see, e.g., White, 1994, Chaps. 2 & 3).

Theorem 1. *Let ξ contain the parameters of LP-2. The Kullback–Leibler divergence of LP-2 from LF-1,*

$$KL\big(f_X \,\|\, g_X^{(2)}\big) = E_{f_X}\left(\ln\left(\frac{f_X(X)}{g_X^{(2)}(X)}\right)\right) \tag{6}$$

obtains its global minimum in ξ when the following three conditions on LP-2 are satisfied:

$$\eta_1 = \eta_2 = 1/2, \qquad \delta_1 = -\delta_2 = \delta, \quad \textit{and} \quad \Theta_1 = \Theta_2 = \Theta. \tag{7}$$

The proof of Theorem 1 is given in the appendix. On the basis of the theorem we use the parameter constraints (7) to derive the results that follow. That is, we employ SLP-2 as a valid parameterization of LP-2 under LF-1 data.

The SLP-2 constraints are also of interest in their own right. SLP-2 has a total of $3P$ free parameters, and referring to Equation (3) we have $D = D_1 = D_2$. Thus SLP-2 is identified at the level of second-order moments by the argument given under Equation (3). This means that fitting SLP-2 using the methods of factor analysis would yield a model that numerically agrees with LF-1 not only at the level of second-order moments, but also at all odd central moments. These similarities with LF-1 are striking and extend previous considerations about the moments of the models. However, they also provide a basis to establish that the models are not equivalent from the perspective of raw data ML or at the level of fourth-order moments.

3.1. Goodness of Fit

This section addresses the overall goodness of fit of SLP-2 to LF-1 data. We derive an upper bound on the KL divergence of SLP-2 from LF-1 and show that it is increasing in Λ, given certain constraints on the error variances. Thus we conclude that, when the data-generating model is LF-1, increasing the overall magnitude of the observed covariances leads to worse fit of LP-2.

In the Appendix, Equations (7) are used to derive the following upper bound on Equation (6):

$$KL\big(f_X \,\|\, g_X^{(2)}\big) \leq \frac{1}{2}\left(\ln\left(\frac{|\Theta|}{|\Psi|}\right) + \mathrm{Tr}(\Theta^{-1}\Psi - I) - \ln(\Lambda'\Psi^{-1}\Lambda + 1) + \Lambda'\Theta^{-1}\Lambda + \delta'\Theta^{-1}\delta\right). \tag{8}$$

We denote the right side of Equation (8) as KL^* and consider its behavior in Λ. First note that if $\Lambda = \mathbf{0}$, then setting $\delta = \mathbf{0}$ and $\Psi = \Theta$ implies that KL^* (and KL) are equal to zero. In this case both models are trivial, reducing to a Gaussian distribution with diagonal covariance matrix, and LP-2 is not formally identified. However, this provides an intuitive starting point from which to consider the behavior of KL^*. Inspection of Equation (8) also shows that KL^* is increasing linearly in $\Lambda\Theta^{-1}\Lambda$ but decreasing logarithmically in $\Lambda\Psi^{-1}\Lambda$. The former term will dominate unless the error variances of LF-1 are very much smaller than those of LP-2. This condition is made explicit in the following.

Because the sign of the factor loadings is arbitrary, let $\Lambda \in \mathbb{R}_+^P$. Define the direction of increase in Λ with respect to $\Lambda + \epsilon$, where ϵ is any P-vector whose elements are all non-negative. Then a sufficient condition for KL^* to be strictly increasing in Λ is that each component of the gradient vector

$$\frac{\partial}{\partial \Lambda} KL^* = \Lambda'\left(\Theta^{-1} - \frac{1}{\Lambda'\Psi^{-1}\Lambda + 1}\Psi^{-1}\right) = \Lambda' Q \tag{9}$$

is strictly positive. The latter condition is satisfied if and only if the matrix Q is positive definite. Manipulation of Q shows that this is equivalent to requiring

$$\frac{\psi_i^2}{\theta_i^2} > \frac{1}{\Lambda'\Psi^{-1}\Lambda + 1} \tag{10}$$

for $i \in \{1, \ldots, P\}$.

Inequality (10) implies that KL^* is increasing in Λ. The right side of the inequality is familiar as the coefficient of alienation of the factor model (i.e., $1 - R^2$). It has a range of (0, 1), and so the condition is satisfied whenever $\theta_i^2 < \psi_i^2$. When $\theta_i^2 > \psi_i^2$, the requirement essentially comes down to the magnitude of the factor loadings, since for any fixed values of Ψ and Θ, it is possible to select a value of Λ such that inequality (10) holds for all $\Lambda + \epsilon$.

To get an empirical sense for inequality (10), let us reconsider the ASVAB data. For LF-1, $1 - \hat{R}^2 = 0.08$, so that the inequality reduces to $\hat{\theta}_i^2 < 12.5\hat{\psi}_i^2$ (i.e., the LP-2 error variances can be no greater than 12.5 times the LF-1 error variances). In fact, $\max\{\hat{\theta}_i^2\} = 2.55\hat{\psi}_i^2$, so that the requirement was satisfied for the ASVAB data. In the simulation study that follows, inequality (10) was always satisfied for correlations between 0.1 and 0.9. Thus we interpret inequality (10) as a mild restriction on the discrepancy between the error variances of LF-1 and LP-2. We leave as an open question the conditions under which inequality (10) is necessarily true.

3.2. *Excess Kurtosis*

A measure of excess multivariate kurtosis, K, was introduced above. The measure is derived in the appendix, where it is also shown that it reduces to the usual definitions in the univariate case and in the case of a single Gaussian distribution. It is notable that in the multivariate case the kurtosis of the Gaussian distribution is not constant, but varies with the second-order moments. Therefore the definition of excess kurtosis requires specification of the particular Gaussian distribution from which the departure is measured. The analytic expression we present here is for

the excess kurtosis of SLP-2 from an LF-1 model with an identical unconditional covariance matrix. We thereby quantify the excess kurtosis necessary for SLP-2 to exactly reproduce the unconditional covariance matrix of LF-1. This is given by

$$K = \frac{-2}{P^2} \sum_{i,j} \left(\frac{1}{\frac{\psi_i^2}{\lambda_i^2} + 1} \right) \left(\frac{1}{\frac{\psi_j^2}{\lambda_j^2} + 1} \right). \tag{11}$$

This formulation of K assumes that $\lambda_i > 0$, $i \in \{1, \ldots, P\}$. From inspection of Equation (11), it is clear that $K \in (0, -2)$ and that it is decreasing in each component of Λ. Thus we have an explanation of the misfit of the observed covariances found in ASVAB data. Underestimation of the observed covariances by SLP-2 lessens its excess negative kurtosis

Parameterizing Equation (11) in terms of LP-2 shows that there is a functional dependency between the model-implied covariances and kurtosis of LP-2. This dependency makes explicit why consideration of only the second-order moments of LP-2 provides an incomplete picture of the relation between the two models.

3.3. Simulation Study

The foregoing has provided some quantities by which to interpret the results of ASVAB example data. These pertain only to population-level relations between the models. Moreover, in the case of KL divergence, the analysis has only partially described its numerical behavior (i.e., in terms of an upper bound). For these reasons we now present a brief simulation study that illustrates the relations described above. The simulations show that, when the population correlations are reasonably large ($\rho \geq 0.3$) the two models can be effectively distinguished at realistic samples size (i.e., $N \geq 200$) using raw data ML. On the other hand, estimation of K can only be recommended for larger sample sizes.

The simulation addresses two factors, namely sample size, $N \in \{100, 200, 300\}$, and the magnitude of the data-generating correlations, $\rho \in \{0.1, 0.2, \ldots, 0.9\}$. The data were generated from LF-1 and fitted to LF-1, LP-2 and SLP-2 using the Fortran program from Dolan and van der Maas (1998). For each combination of N and ρ, $NR = 500$ simulations were conducted, yielding a total of 13,500 simulated data sets. The value of the correlations was uniform over the data-generating correlation matrix, and $P = 5$ variables were used in each data set. For each replication and each model M, we recorded the maximized log-likelihoods, $l_{(M)}$, the observed excess kurtosis, $\hat{K}_{(M)}$, and the root mean square error of the model-implied correlations, $RMSE_{(M)}$.

For both very small correlations ($\rho < 0.2$) and very large ($\rho \geq 0.8$), LP-2 did not always converge; and, when it did converge, it did not always arrive at the global maximum. These data sets were omitted from the following analysis. In the worst case 53 (10.3%) of the samples were omitted for $\rho = 0.9$ and $N = 100$. In total, 164 (1.2%) of the samples were dropped, leaving 13,336 usable data sets. Convergence problems were less severe for the larger sample sizes and non-existent at other values of ρ.

Consider first the estimated minimum of the KL divergence:

$$\widehat{KL}_{\min}\left(f_X \,\|\, g_X^{(2)}\right) = \frac{1}{N}\left(l_{(\text{LF-1})} - l_{(\text{LP-1})}\right). \tag{12}$$

Figure 4 plots the mean of $\widehat{KL}_{\min}$ against each value of ρ and for $N = 200$. The error bars were computed as $\pm 2 \times SD$. The distributions of $\widehat{KL}_{\min}$ at each value of ρ were unimodal and symmetrical, so that the displayed statistics provide a reasonable summary of the observed distributions. We note that the depicted values can be rescaled by affine transformation to differences in information criteria between LF-1 and LP-2.

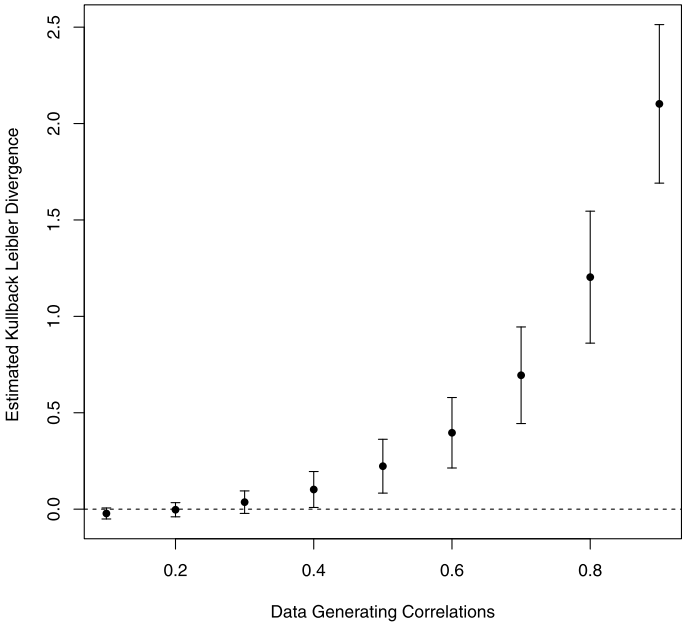


FIGURE 4.
Estimated minimum Kullback–Leibler divergence versus data-generating correlations. Plot depicts averages and probable errors. Minimum Kullback–Leibler divergence was estimated between a correctly specified linear factor model and a misspecified latent profile model using a sample size of 200.

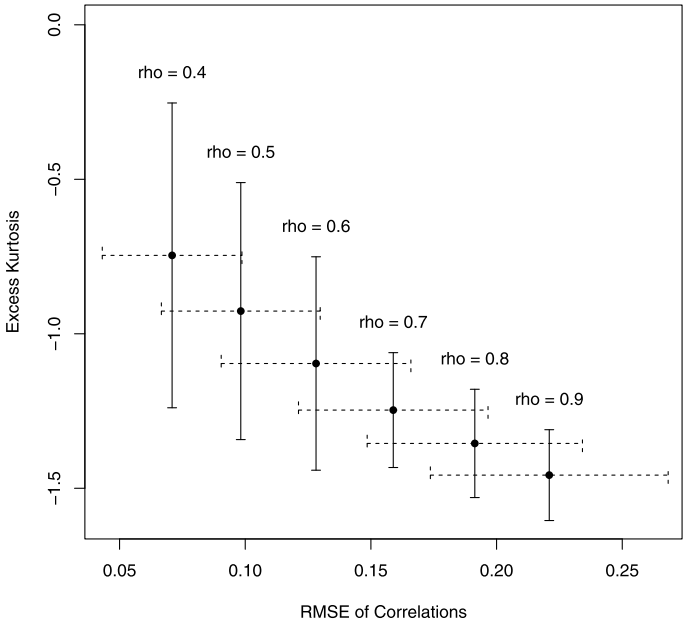


FIGURE 5.
Model-implied excess multivariate kurtosis versus root means square error (RMSE) of correlations for latent profile model. Plot depicts averages and probable errors at various values of data-generating correlations (rho). Kurtosis and RSME of correlations were computed using sample sizes of 300.

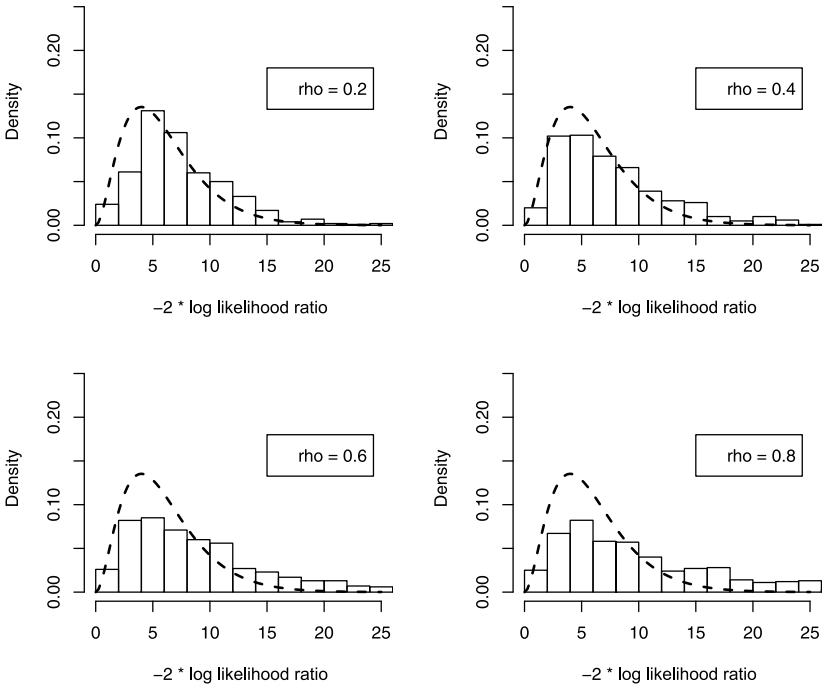


FIGURE 6.
Sampling distributions of the likelihood ratio test of a symmetry constrained latent profile against unconstrained latent profile model. Panels show distributions for various values of data-generating correlations (rho). Samples were generated under linear factor model with $N = 300$. *Dotted curve* represents chi-square distribution with six degrees of freedom.

The trend in ρ is apparent. For $\rho < 0.3$ it was not uncommon that $\widehat{KL}_{\min} < 0$, which can be attributed to LP-2 overfitting random sampling fluctuations. However, with moderate to large correlations, the error bars do not include zero. The results for $N = 300$ led to identical conclusions. For $N = 100$, the probable error included zero for $\rho \leq 0.5$. We conclude that goodness of fit based on raw data ML will distinguish the models when moderate sample sizes are used and the true correlations are not too small.

Figure 5 shows that larger observed correlations also lead to worse estimation of the second- and fourth-order moments under LP-2. The figure depicts the means of the excess kurtosis against those of the *RMSE* of the correlations for LP-2. The error bars for $\hat{K}$ and for *RMSE* are also shown ($\pm 2 \times SD$); the dispersion within levels of ρ was approximately elliptical in all cases. The plot shows values for $N = 300$ and $\rho \geq 0.4$. Smaller correlations and sample sizes were not considered since both of these conditions led to large instability in the estimates of K. Figure 5 required omission of 39 (7.8%) cases for $\rho = 0.4$ and 10 (2.0%) cases for $\rho = 0.5$; these cases were outside of the theoretical range of K. Despite these limitations, Figure 5 depicts the theoretical relationship described in Equation (11): Larger observed correlations led to more excess negative kurtosis, which in turn was also associated with worse fit at the level of second-order moments. Visual inspection of the second-order moments revealed that the misfit was uniformly due to underestimation.

As a final point, we consider the validity of SLP-2 constraints in finite samples. Figure 6 shows some histograms of the log-likelihood ratio statistic for SLP-2 against LP-2. The distributions are for $N = 300$ since this best reflects the asymptotic properties of the distribution, and values of ρ are indicated in the panels. The dotted curve represents the chi-square distribution on

6 degrees of freedom, which would be the reference distribution under conventional likelihood theory.

It can be seen that the sampling distributions in Figure 6 are affected by the magnitude of the observed correlations. In particular, the observed type I error rates at the 5% level ranged from 11.0% ($\rho = 0.2$) to 32.6% ($\rho = 0.8$). Corrections based on Satorra and Bentler's (2001) theory did not do much to remedy this situation. The distributions proposed by Vuong (1989) also did not describe the observed probabilities; these almost never led to a rejection of the null hypothesis even at very large sample sizes ($N \geq 2000$). Information criteria consistently favoured SLP-2. For instance, in the 13,336 replications in which the unconstrained model converged, SLP-2 was favoured 73.9% of the time by AIC and 99% of the time by BIC. Thus, while we may conclude that SLP-2 is the preferable model, the asymptotic distribution of the likelihood ratio statistic in this situation remains an interesting question.

4. Conclusions

This paper has revisited the relationship between linear factor analysis and latent profile analysis from the perspective of maximum likelihood estimation based on their unconditional densities. The prevailing wisdom about the relation between the two models, which is based on consideration of only their unconditional second-order moments, was found to be lacking in the present context. An upper bound on the KL divergence of the latent profile model from the factor model was shown to be increasing in the observed covariances. We also discussed the higher-order moments of the latent profile model, showing that its multivariate kurtosis decreases (becomes more negative) with the observed covariances. The latter result provides a reasonably intuitive explanation of the non-equivalence between the two models: The latent profile model cannot simultaneously reproduce the covariances and kurtosis of a non-trivial Gaussian factor model. In terms of practical consequences, the models were well distinguished on the basis of their likelihoods provided that the observed correlations and sample size were moderate ($\rho \geq 0.3$; $N \geq 200$). Stable estimation of the excess kurtosis of LP-2 required larger sample sizes.

This paper also demonstrated the validity of a constrained latent profile model under linear factor data. The constraints imply an unconditional density that is symmetrical about its mean and has the same number of free parameters as the factor model. As described in the Appendix, the result applies to any case where one mistakenly fits a two-component Gaussian mixture to a homogeneous Gaussian distribution. As such it applies directly to models with more general covariance structures (e.g., factor mixture models). While tests based on the asymptotic distribution of the likelihood ratio of these constraints are not currently available, in the present case information criteria consistently favored the constrained model.

It is also interesting to consider the validity of the symmetry constraints in more general cases, for example where $J \geq 1$, $K > 2$, and without requiring that $K = J + 1$. In these cases we have found that observed distributions of fitted LP-K models are symmetrical and that increasing the value of K for fixed J leads to better reproduction of the observed second- and fourth-order moments. Thus, extensions of the present research seem to suggest that SLP-K can be treated as a discrete approximation to LF-J for any value of $K \geq J + 1$. Analytic results have so far proven intractable for larger numbers of classes.

Further research directions include addressing the implications of the present results for the original model selection problem described at the outset of this paper. For instance, given that factor models and latent profile models differ with regard to their model-implied multivariate kurtosis, it seems that goodness of fit statistics that take into account fourth-order moments as well as second-order moments (e.g., weighted least squares and distribution free tests) may also successfully distinguish the two models. There is also the interesting possibility of extending these results to the case of discrete manifest variables.

Acknowledgements

This research was funded in part by a postdoctoral grant from the National Science and Engineering Research Council of Canada.

Appendix: Derivation of Excess Multivariate Kurtosis of SLP-2 from LP-2

We obtain Equation (11) through the central moments of a P-dimensional, K-component Gaussian mixture. These are derived in a straightforward manner from the component moment generating functions,

$$M_k(t) = \exp\{0.5t'\Theta_k t + t'\delta_k\},$$

with $t \in \mathbb{R}^P$ and $k \in \{1, \ldots, K\}$. As in Equation (3), the δ_k are deviations of the component means from the unconditional mean. For the moment we place no restrictions on the component covariance matrices, Θ_k.

We introduce the following notation. Let

$$M_k^{ab\cdots n}(t) = \frac{\partial}{\partial t_a}\left(\frac{\partial}{\partial t_b}\cdots\left(\frac{\partial}{\partial t_n}\big(M_k(t)\big)\right)\right) \tag{13}$$

denote the derivatives of $M_k(t)$ with respect to the elements $t_a, t_b, \ldots, t_n$ of t. Note that the number of superscripts in Equation (13) denotes the order of the derivative, whereas the indices themselves denote elements of t. We use partition notation to represent summation over all possible permutations of indices (see, e.g., McCullagh, 1987). In this notation, the summation is implicit and the number of permutations summed over is given in square brackets. We also require that the indices for the components of the mixture are distinguished from those for the elements of a tensor. For this purpose we place the former in parentheses and require that permutation notation does not apply to indices that appear in parentheses. For example,

$$x_{(k)ab}y_{(k)c}[3] = x_{(k)ab}y_{(k)c} + x_{(k)ac}y_{(k)b} + x_{(k)bc}y_{(k)a}.$$

Using these conventions the central fourth-order moments of a Gaussian mixture, G, can be written:

$$\begin{aligned}\phi_G^{abcd} &= \sum_{k=1}^{K}\eta_k\big(M_k^{abcd}(0)\big)\\ &= \sum_{k=1}^{K}\eta_k\big(\theta_{(k)ab}\theta_{(k)cd}[3] + \theta_{(k)ab}\delta_{(k)c}\delta_{(k)d}[6] + \delta_{(k)a}\delta_{(k)b}\delta_{(k)c}\delta_{(k)d}\big).\end{aligned} \tag{14}$$

Equation (14) follows immediately from computation of the necessary derivatives. We note the following special cases of Equation (14). Setting $P = 1$, it is equivalent to the expression found in Frühwirth–Schnatter (2006, p. 11). Setting $K = 1$ yields the central fourth-order moment of a homogeneous Gaussian model: $\phi_G^{abcd} = \theta_{ab}\theta_{cd}[3]$. Setting $P = 1$ and $K = 1$ gives $\phi_G^{abcd} = 3\theta^4$.

A summary measure of the excess kurtosis of a Gaussian mixture G_1 relative to any other Gaussian mixture G_2 can be obtained by standardizing Equation (14) relative to the corresponding elements of the unconditional covariance matrix and averaging over their difference:

$$K(G_1, G_2) = \frac{1}{P^4}\sum_{a,b,c,d,}\left(\frac{\phi_{G_1}^{abcd}}{\phi_{G_1}^{ab}\phi_{G_1}^{cd}} - \frac{\phi_{G_2}^{abcd}}{\phi_{G_2}^{ab}\phi_{G_2}^{cd}}\right). \tag{15}$$

Here ϕ_G^{ab} denotes the (a, b)-th element of the unconditional covariance matrix of G.

Equation (15) can be used for computation. In particular, it was used with the ASVAB data by setting G_2 to a homogeneous Gaussian model with parameters estimated directly from the data. G_1 was then substituted for the estimated LF-1 and LP-2 models. Note that Equation (15) can be easily converted to a mean square error or a similar measure, and this would be preferable when the sign of the differences is not known or not of interest.

The expression for K given in Equation (11) is obtained by substituting into Equation (15) using SLP-2 for G_1 and LF-1 for G_2 and setting the parameters of SLP-2 to equal those of LF-1. Here we note that

$$\phi_{\text{LF-1}}^{abcd} = (\psi_{ab} + \lambda_a\lambda_b)(\psi_{cd} + \lambda_c\lambda_d)[3] = \psi_{ab}\psi_{cd}[3] + \psi_{ab}\lambda_c\lambda_d[6] + 3\lambda_a\lambda_b\lambda_c\lambda_d, \tag{16}$$

so that $\phi_{\text{LP-2}}^{abcd} - \phi_{\text{LF-1}}^{abcd} = -2\lambda_a\lambda_b\lambda_c\lambda_d$ and

$$K(\text{LP-2, LF-1}) = \frac{-2}{P^4} \sum_{a,b,c,d,} \frac{\lambda_a\lambda_b\lambda_c\lambda_d}{(\lambda_a\lambda_b + \psi_{ab})(\lambda_c\lambda_d + \psi_{cd})}. \tag{17}$$

Assuming the factor loadings are non-zero, we can divide through by $\lambda_a\lambda_b\lambda_c\lambda_d$ and then Equation (11) follows by simplifying.

A.1. Proof of Theorem 1

This section provides several results that together imply Theorem 1. Throughout $f(x) = f_X(x; \mathbf{0}, \Sigma_X)$ denotes an arbitrary Gaussian density centered at the origin, and $g(x) = g(x; \xi) = g_X^{(2)}(x; \xi)$ denotes the LP-2 model given in Equation (2) with parameter vector ξ. Because Lemma 3 (see below) shows that the unconditional mean of $g(x)$ is equal to that of $f(x)$ when $KL(\xi) = KL(f \| g)$ is minimized in ξ, we further simplify notation by writing $\mu = \delta$ for the constraints in Equation (7). Other than positive definiteness, no explicit restrictions on the within-class covariance matrices of g are required. We do, however, assume the existence of the maximum likelihood estimator $\hat{\xi}$ of ξ under $f(x)$ and that $\hat{\xi}$ converges in probability to the unique argument ξ^* that maximizes $L(\xi) = E_f(\ln(g(X; \xi)))$ (see White, 1994, Chaps. 3 & 4). Then ξ^* is also the argument that minimizes $KL(\xi)$. Therefore we prove Theorem 1 by showing that ξ^* satisfies the constraints given in Equations (7). The argument centers on the asymptotic log-likelihood equations

$$\frac{\partial L}{\partial \xi} = \int_{\mathbb{R}^P} f(x) \frac{\partial \ln(g(x; \xi))}{\partial \xi}\, dx = \mathbf{0}, \tag{18}$$

and in particular on

$$\frac{\partial L}{\partial \eta_1} = \int \frac{f(x)}{g(x)} \big(g_1(x; \mu_1, \Theta_1) - g_2(x; \mu_2, \Theta_2)\big)\, dx = 0; \tag{19}$$

$$\frac{\partial L}{\partial \mu_k} = \int \frac{f(x)}{g(x)} \eta_k g_k(x; \mu_k, \Theta_k)(x - \mu_k)' \Theta_k^{-1}\, dx = \mathbf{0}. \tag{20}$$

Definition 1 and Proposition 1 remind the reader of some basic results employed in the following.

Definition 1 (Odd and Even Functions). Let $h : \mathbb{R}^m \to \mathbb{R}$. We call h *even* if for all $x \in \mathbb{R}^m$, $h(x) = h(-x)$, and we call h *odd* if for all $x \in \mathbb{R}^m$, $h(-x) = -h(x)$.

Proposition 1. *Let* $h, h_1, h_2 : \mathbb{R}^m \to \mathbb{R}$. *Then*

1.1 $h(x) + h(-x)$ *is even for all* h.
1.2 $h(x) - h(-x)$ *is odd for all* h.
1.3 *If* $h_1(x)$ *is odd and* $h_2(x)$ *is even,* $h_1(x)h_2(x)$ *is odd.*
1.4 *If* $h(x)$ *is odd,* $\int_{\mathbb{R}^m} h(x)\,dx = 0$.

Note that $f(x)$ is even in the present context. The following lemma concerns Equation (19).

Lemma 1. *Equations* (7) *imply* $\frac{\partial L}{\partial \eta_1} = 0$.

Proof: Substituting Equations (7) into Equation (19) yields

$$2\int f(x)\frac{g_1(x;\mu,\Theta) - g_2(x;-\mu,\Theta)}{g_1(x;\mu,\Theta) + g_2(x;-\mu,\Theta)}\,dx = 2\int f(x)\frac{g_1(x;\mu,\Theta) - g_2(-x;\mu,\Theta)}{g_1(x;\mu,\Theta) + g_2(-x;\mu,\Theta)}\,dx. \quad (21)$$

The denominator in Equation (21) is an even function (Proposition 1.1) and the numerator is an odd function (Proposition 1.2), so the fraction is an odd function (Proposition 1.3). Because $f(x)$ is even, the integrand is odd (Proposition 1.3), and so the integral evaluates to zero (Proposition 1.4). □

Lemma 1 does not imply that Equations (7) correspond to a solution of Equation (18). It simply shows that $\eta_1 = 1/2$ is one solution of Equation (19). However,

$$\frac{\partial^2 L}{\partial \eta_1^2} = -\int \frac{f(x)}{g(x)^2}\big(g_1(x;\mu_1,\Theta_1) - g_2(x;\mu_2,\Theta_2)\big)^2\,dx, \quad (22)$$

is negative for all $\eta_1 \in [0, 1]$ and so L is convex in η_1. Therefore there is only one solution to Equation (19), and this is a global maximum. Lemma 1 and Equation (22) then imply

Corollary 1. $\eta_1^* = 1/2$ *and this is the unique solution of Equation* (19).

The following two lemmas show that Corollary 1 also yields $\mu_1^* = -\mu_2^*$.

Lemma 2. *For* $k \in \{1, 2\}$, *Equation* (19) *implies*

$$\int \frac{f(x)}{g(x)} g_k(x;\mu_k,\Theta_k)\,dx = 1. \quad (23)$$

Proof: Rearranging Equation (19) we have

$$\int \frac{f(x)}{g(x)} g_1(x;\mu_1,\Theta_1)\,dx = \int \frac{f(x)}{g(x)} g_2(x;\mu_2,\Theta_2)\,dx \quad (24)$$

and consequently

$$\begin{aligned}\int \frac{f(x)}{g(x)} g_k(x;\mu_k,\Theta_k)\,dx &= (\eta_1+\eta_2)\int \frac{f(x)}{g(x)} g_k(x;\mu_k,\Theta_k)\,dx \\ &= \eta_1 \int \frac{f(x)}{g(x)} g_1(x;\mu_1,\Theta_1)\,dx + \eta_2 \int \frac{f(x)}{g(x)} g_2(x;\mu_2,\Theta_2)\,dx \\ &= \int \frac{f(x)}{g(x)} \big(\eta_1 g_1(x;\mu_1,\Theta_1) + \eta_2 g_2(x;\mu_2,\Theta_2)\big)\,dx \\ &= \int \frac{f(x)}{g(x)} g(x)\,dx = \int f(x)\,dx = 1. \end{aligned} \tag{25}$$

□

Lemma 3. *Equations* (19) *and* (20) *imply* $E_f(X) = E_g(X)$.

Proof: Manipulation of Equation (20) gives

$$\eta_k \int x \frac{f(x)}{g(x)} g_k(x;\mu_k,\Theta_k)\,dx = \eta_k \mu_k \int \frac{f(x)}{g(x)} g_k(x;\mu_k,\Theta_k)\,dx. \tag{26}$$

By Lemma 2 the integral on the right hand side of Equation (26) evaluates to 1, and thus the integral on the left hand side is equal to μ_k. Then from adding together Equation (26) for $k = 1, 2$ it follows that

$$\int x f(x)\,dx = \sum_{k=1}^{2} \eta_k \mu_k = \mathbf{0}. \tag{27}$$

□

Lemma 3 shows that the asymptotic maximum likelihood Equations (19) and (20) imply the first-order moment equation $E_f(X) = E_g(X)$. The latter is equivalent to $\eta_1\mu_1 = -\eta_2\mu_2$, and in combination with Corollary 1 this leads to:

Corollary 2. *Any solution to Equations* (19) *and* (20) *satisfies* $\mu_1 = -\mu_2$.

Of course, this holds also for ξ^*.

It remains to show that $\Theta_1^* = \Theta_2^*$. For reference we note that the univariate second-order moment equations can be derived in a manner similar to Lemma 3, but this is not the most convenient way to obtain the desired result. Instead we consider Lemma 4. We make use of the following notation. For any fixed value of ξ, the vector ξ_+ is obtained from ξ by switching Θ_1 and Θ_2. For explicitness, if $\xi = (\eta_1, \mu_1, \mu_2, \Theta_1, \Theta_2)$ then $\xi_+ = (\eta_1, \mu_1, \mu_2, \Theta_2, \Theta_1)$.

Lemma 4. *For all* ξ *that satisfy Equations* (19) *and* (20), $L(\xi) = L(\xi_+)$.

Proof: First note that for $y = -x$

$$L(\xi) = \int f(x) \ln\big(g(x;\xi)\big)\,dx = \int f(y) \ln\big(g(y;\xi)\big)\,dy, \tag{28}$$

But under Corollaries 1 and 2,

$$\begin{aligned} 2g(y;\xi) &= g_1(-x,\mu,\Theta_1) + g_2(-x,-\mu,\Theta_2) \\ &= g_1(x,-\mu,\Theta_1) + g_2(x,\mu,\Theta_2) = 2g(x;\xi_+). \end{aligned} \tag{29}$$

Substituting $g(x;\xi_+)$ into Equation (28) and using the evenness of f give the result. □

Roughly speaking, Lemma 4 states that it makes no difference whether one evaluates L for ξ or ξ_+, since $g(x;\xi)$ is just the reflection of $g(x,\xi_+)$ about the origin (given Corollaries 1 and 2) and $f(x)$ is symmetrical about the origin. The lemma holds in particular for ξ^*. Thus we have the following corollary:

Corollary 3. *If ξ^* is unique, then $\Theta_1^* = \Theta_2^*$.*

If the uniqueness of the asymptotic ML solution is assumed, as we have done here, Corollaries 1 thru 3 directly imply Theorem 1 as a special case. Theorem 1 is a special case because the results presented in this section require no structural restrictions on the covariance matrix of $f(x)$ or on the component covariance matrices of $g(x)$.

*A.2. Derivation of KL**

In this section we derive Equation (8). For simplicity of notation we assume that f_X and $g_X^{(2)}$ are centered at the origin and write μ in place of δ for the restrictions in Equations (7). Using Equation (7), $g_X^{(2)}$ can be factored as follows:

$$g_X^{(2)}(x) = \left(2^{(p+2)}\pi^p|\Theta|\right)^{-1/2} \exp\left\{-\frac{1}{2}\left(x'\Theta^{-1}x + \mu'\Theta^{-1}\mu\right)\right\} \times \exp\{A(x)\} \tag{30}$$

where $\exp\{A(x)\} = B(x) + B(x)^{-1}$ and

$$B(x) = \exp\left\{x'\Theta^{-1}\mu\right\}. \tag{31}$$

The KL divergence of g_X^2 from f_X is gotten in a manner similar to the usual method for two Gaussian distributions, with a remainder term $R = \ln(1/2) + E_{f_X}(A(X))$:

$$\begin{aligned} KL\left(f_X \| g_X^{(2)}\right) &= E_{f_X}\left(\ln\left(f_X(X)\right) - \ln\left(g_X^{(2)}(X)\right)\right) \\ &= \frac{1}{2}\left(\ln\left(\frac{|\Theta|}{|\Lambda\Lambda' + \Psi|}\right) + \mathrm{Tr}\left(\Theta^{-1}(\Lambda\Lambda' + \Psi) - I\right) + \mu'\Theta^{-1}\mu\right) - R. \end{aligned} \tag{32}$$

The upper bound is obtained by using the fact that $A(x)$ is concave in x. This is readily verified by setting $x = u + yt$ for $u, t \in \mathbb{R}^P$ and $y \in \mathbb{R}$ and noting that $\frac{\partial^2}{\partial y^2} A(u + yt) \leq 0$ (see Boyd & Vandenberghe, 2004). Then application of Jensen's inequality leads to the following lower bound on R:

$$R \geq \ln(1/2) + \ln\left(A\left(E_{f_X}(X)\right)\right) = \ln(1/2) + \ln\left(\exp\{0\} + \exp\{0\}\right) = 0. \tag{33}$$

Therefore, omitting the term R from Equation (32) yields an upper bound on KL. Equation (8) follows by rearranging the remaining terms. Note that equality holds in (33) when $\mu = \mathbf{0}$ (i.e., when the components are not separated).

References

Bartholomew, D.J., & Knott, M. (1999). *Latent variable models and factor analysis*. London: Arnold.
Bollen, K.A. (1989). *Structural equations with latent variables*. New York: John Wiley and Sons.
Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. New York: Cambridge University Press.
Dolan, C.V., & van der Maas, H. (1998). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika*, *63*, 227–253.
Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.
Halpin, P.F., & Maraun, M.D. (2010). Selection between linear factor models and latent profile models using conditional covariances. *Multivariate Behavioral Research*, *46*, 1–25.
Loken, E., & Molenaar, P. (2007). On the arbitrary nature of latent variables. In G.R. Hancock, & K.M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 277–298). Charlotte: Information Age Publishing.
Lubke, G., & Neale, M. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood. *Multivariate Behavioral Research*, *41*, 499–532.
Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*, 519–530.
McCullagh, P. (1987). *Tensor methods in statistics*. New York: Chapman Hall.
McDonald, R.P. (1967). *Psychometric Monographs: Vol. 15. Nonlinear factor analysis*.
McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley and Sons.
Molenaar, P.C.M., & von Eye, A. (1997). On the arbitrary nature of latent variables. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Latent variable analysis* (pp. 226–242). New York: Springer.
Ray, S., & Lindsay, B.G. (2005). The topography of multivariate normal mixtures. *Annals of Statistics*, *33*, 2042–2065.
Satorra, A., & Bentler, P.M. (2001). A scaled difference chi-square test for moment structure analysis. *Psychometrika*, *66*, 507–514.
Segall, D.O. (2004). Development and evaluation of the 1997 ASVAB score scale (Technical report no. 2004-002). Seaside, CA: Defense Manpower Data Center.
Steinley, D., & McDonald, R.P. (2007). Examining factor score distributions to determine the nature of latent spaces. *Multivariate Behavioral Research*, *42*, 133–156.
Vuong, Q.A. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrika*, *57*, 307–333.
White, H. (1994). *Estimation, inference, and specification analysis*. New York: Cambridge University Press.

Manuscript Received: 26 FEB 2011
Final Version Received: 26 FEB 2011
Published Online Date: 1 OCT 2011