Minimum ϕ-divergence estimation in constrained latent class models

A. Felipe, P. Miranda* and L. Pardo
Department of Statistics and Operations Research
Faculty of Mathematics
Complutense University of Madrid
28040 Madrid (Spain)

Abstract

The main purpose of this paper is to introduce and study the behaviour of minimum ϕ-divergence estimators as an alternative to the maximum likelihood estimator in latent class models. Minimum ϕ-divergence estimators are a natural extension of the maximum likelihood estimator. The asymptotic properties of the minimum ϕ-divergence estimator for constrained latent class models are developed. Moreover, a simulation study is carried out in order to compare the new estimators with those obtained through maximum likelihood for small and moderate sample sizes.

Keywords: Latent class models, Minimum phi-divergence estimator, Maximum likelihood estimator, Asymptotic distribution.

1. Introduction

Latent class models (LCM) were introduced in Lazarsfeld (1950) as a tool for studying categorical data analysis. Since then, many papers have been published with applications of LCM in different areas; see e.g. Hagenaars&Cutcheon (2002), Langeheine&Rost (1988), Rost&Langeheine (1997), Collins&Lanza (2010), Biemer (2011); LCM models are specially important in behavioral and social sciences (e.g. Hagenaars&Cutcheon (2002), Abar&Loken (2010), Caldwell et al. (2009), Coffman et al. (2007), Feldman et al. (2009), Gerber et al. (2009), Laska et al. (2009), Nylund et al. (2007)).

We shall formulate the problem that LCM deals with in the same way it appears in Formann (1985) and we shall use the parametrization considered in that paper. Consider a set of N people: $P_1, ..., P_N$. Each person answers to k dichotomous items $I_1, ..., I_k$; let us denote by $y_{\nu i}$ the answer of person P_ν to item I_i, i.e.

*Corresponding author: P. Miranda. pmiranda@mat.ucm.es

$$y_{\nu i} := \begin{cases} 1 & \text{if the answer of } P_\nu \text{ to } I_i \text{ is correct} \\ 0 & \text{otherwise} \end{cases}.$$

Let $y_\nu := (y_{\nu 1}, ..., y_{\nu k})$ denote a generic pattern of right and wrong answers to the k items given by P_ν. In order to explain the statistical relationship of the observed variables, a categorical latent variable (categorical unobservable variable) is postulated to exist, whose different levels partition the sample into m mutually exclusive and exhaustive latent subsamples. Let us denote these classes by $C_1, ..., C_m$ and their corresponding relative sizes by $w_1, ..., w_m$; thus, w_j denotes the probability of a randomly selected person P_ν belongs to class C_j, i.e.

$$w_j = Pr(P_\nu \in C_m),\ j = 1, ..., m.$$

We denote by p_{ji} the probability of a right answer of P_ν to the item I_i under the assumption that P_ν is in class C_j :

$$p_{ji} = Pr(y_{\nu i} = 1 / P_\nu \in C_j).$$

We shall assume that in each class the answers for the different questions are stochastically independent; therefore, we can write

$$Pr(y_\nu / P_\nu \in C_j) = \prod_{i=1}^{k} p_{ji}^{y_{\nu i}} (1 - p_{ji})^{1 - y_{\nu i}},$$

and

$$Pr(y_\nu) = \sum_{j=1}^{m} w_j \prod_{i=1}^{k} p_{ji}^{y_{\nu i}} (1 - p_{ji})^{1 - y_{\nu i}}. \tag{1}$$

There are 2^k possible answer vectors y_ν whose probability of occurrence are given by (1); they constitute the manifest probabilities for the items $I_1, ..., I_k$ in the population given by $P_1, ..., P_N$. We will denote by N_ν, $\nu = 1, ..., 2^k$, the number of times that the sequence y_ν appears in an N-sample. Let us denote by

$$\hat{\mathbf{p}} = (N_1/N, ..., N_{2^k}/N),$$

and by L the likelihood function given by

$$L(w_1, ..., w_m, p_{11}, ..., p_{mk}) = Pr(N_1 = n_1, ..., N_{2^k} = n_{2^k}) = \frac{N!}{\prod_{s=1}^{2^k} n_s!} \prod_{s=1}^{2^k} Pr(y_s)^{n_s}. \tag{2}$$

By n_s we are denoting a realization of the random variable $N_s, s = 1, ..., 2^k$. In this model the unknown parameters are $w_j, j = 1, ..., m$ and $p_{ji}, j = 1, ..., m, i = 1, ..., k$. These parameters can be estimated using the maximum likelihood estimator (e.g. McHugh (1956), Lazarsfeld&Henry (1968), Clogg (1995)). In order to avoid the problem of obtaining uninterpretable estimates for the item latent probabilities lying outside the interval $[0, 1]$, some authors (Lazarsfeld&Henry (1968), Formann (1976), Formann (1977), Formann (1978), Formann (1982), Formann (1985)) proposed a parametrization for the probabilities w_j and p_{ji} given by

$$p_{ji} = \frac{exp(\sum_{r=1}^{t} q_{jir}\lambda_r + c_{ji})}{1 + exp(\sum_{r=1}^{t} q_{jir}\lambda_r + c_{ji})},\ j = 1, ..., m,\ i = 1, ..., k,$$

and

$$w_j = \frac{exp(\sum_{r=1}^{u} v_{jr}\eta_r + d_j)}{\sum_{j=1}^{m} exp(\sum_{r=1}^{u} v_{jr}\eta_r + d_j)},\ j = 1, ..., m,$$

where

$$\mathbf{Q}_r = (q_{jir})_{\substack{j=1,...,m \\ i=1,...,k}}, r = 1, ..., t,\ \mathbf{C} = (c_{ji})_{\substack{j=1,...,m \\ i=1,...,k}},\ \mathbf{V} = (v_{jr})_{\substack{j=1,...,m \\ r=1,...,u}},\ \mathbf{d} = (d_j)_{j=1,...,m},$$

are fixed. Consequently, in this case the vector of unknown parameters $\boldsymbol{\theta}$ in the LCM is given by

$$\boldsymbol{\theta} := (\boldsymbol{\lambda}, \boldsymbol{\eta}),$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ are defined as

$$\boldsymbol{\lambda} := (\lambda_1, ..., \lambda_t),\ \boldsymbol{\eta} := (\eta_1, ..., \eta_u).$$

By $\boldsymbol{\Theta}$ we shall denote the parametric space. Thus, we have $t + u$ unknown parameters that can be estimated by maximum likelihood using (2).

The main purpose of this paper is to present a new family of estimators for the parameters $\boldsymbol{\theta}$ based on divergence measures: Minimum ϕ-divergence estimators. This family of estimators contains as a particular case the maximum likelihood estimator. Minimum ϕ-divergence estimators were introduced for the first time in Morales et al. (1995) and since then, many interesting estimation problems have been solved using minimum ϕ-divergence estimators, see Pardo (2006).

The rest of the paper is organized as follows: In Section 2 we introduce the definition of minimum ϕ-divergence estimator in the context of LCM. Its behavior for big sample sizes is presented in Section 3, as well as its asymptotic distribution. The behavior of minimum ϕ-divergence estimators for small and moderate sample sizes in LCM is carried out in Section 4 on the basis of a simulation study. Section 5 presents a case study. Last section is devoted to the conclusions. Finally, in an appendix we provide the proofs of the results developed in Section 3.

2. Minimum ϕ-divergence estimator in LCM

We are going to introduce the minimum ϕ-divergencce estimator as a natural extension of the maximum likelihood estimator. In the following, we denote $Pr(y_\nu)$ by $p(y_\nu, \boldsymbol{\lambda}, \boldsymbol{\eta})$. Based on

(2), the maximum likelihood estimator is obtained by maximizing in $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ the log-likelihood function,

$$\sum_{\nu=1}^{2^k} n_\nu \log p(y_\nu, \boldsymbol{\lambda}, \boldsymbol{\eta}). \tag{3}$$

The expression (3) can be written as

$$\begin{aligned}
\sum_{\nu=1}^{2^k} n_\nu \log p(y_\nu, \boldsymbol{\lambda}, \boldsymbol{\eta}) &= N \sum_{\nu=1}^{2^k} \frac{n_\nu}{N} \log p(y_\nu, \boldsymbol{\lambda}, \boldsymbol{\eta}) \\
&= -N \sum_{\nu=1}^{2^k} \hat{p}_\nu \log \frac{1}{p(y_\nu, \boldsymbol{\lambda}, \boldsymbol{\eta})} - N \sum_{\nu=1}^{2^k} \hat{p}_\nu \log \hat{p}_\nu + N \sum_{\nu=1}^{2^k} \hat{p}_\nu \log \hat{p}_\nu \\
&= -N \sum_{\nu=1}^{2^k} \hat{p}_\nu \log \frac{\hat{p}_\nu}{p(y_\nu, \boldsymbol{\lambda}, \boldsymbol{\eta})} + N \sum_{\nu=1}^{2^k} \hat{p}_\nu \log \hat{p}_\nu \\
&= -N D_{Kullback}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})) + constant,
\end{aligned}$$

being

$$\mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})) = (p(y_1, \boldsymbol{\lambda}, \boldsymbol{\eta}), ..., p(y_{2^k}, \boldsymbol{\lambda}, \boldsymbol{\eta})),$$

and

$$D_{Kullback}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})) = \sum_{\nu=1}^{2^k} \hat{p}_\nu \log \frac{\hat{p}_\nu}{p(y_\nu, \boldsymbol{\lambda}, \boldsymbol{\eta})}, \tag{4}$$

the Kullback-Leibler divergence measure between the probabilities $\hat{\mathbf{p}}$ and $\mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})$. Therefore, maximizing (3) in $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ is equivalent to minimizing (4) in $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$. Consequently, the value $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}})$ that minimizes $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\eta})$ in the Kullback-Leibler divergence is the maximum likelihood estimator of the parameters for the LCM or equivalently, the minimum Kullback-Leibler divergence estimator. We shall denote it by $\hat{\boldsymbol{\theta}}$ or

$$\hat{\boldsymbol{\theta}}_{Kullback} := arg \min_{(\boldsymbol{\lambda}, \boldsymbol{\eta}) \in \boldsymbol{\Theta}} D_{Kullback}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})).$$

Therefore, we can observe that the maximum likelihood estimator for the LCM consists of the values $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\eta}}$ for which the Kullback-Leibler divergence measure between the probability vectors $\hat{\mathbf{p}}$ and $\mathbf{p}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}})$ minimizes. If D is a measure of distance between $\hat{\mathbf{p}}$ and $\mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})$, we can generalize the concept of maximum likelihood estimator (or minimum Kullback-Leibler divergence estimator) by

$$\hat{\boldsymbol{\theta}}_D := arg \min_{(\boldsymbol{\lambda}, \boldsymbol{\eta}) \in \boldsymbol{\Theta}} D(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})).$$

A question now arises: which measures of distance are suitable for generalizing the Kullback-Leibler divergence? To answer this question, it is necessary to keep in mind that the Kullback-Leibler divergence measure between $\hat{\mathbf{p}}$ and $\mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})$ is a particular case of the family of divergence measures introduced in Csiszár (1967) through

$$D_{\phi}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})) = \sum_{j=1}^{2^k} p(y_j, \boldsymbol{\lambda}, \boldsymbol{\eta}) \phi\left(\frac{\hat{p}_j}{p(y_j, \boldsymbol{\lambda}, \boldsymbol{\eta})}\right). \tag{5}$$

where ϕ is a convex function for $x > 0$ satisfying $\phi(1) = 0, 0\phi(0/0) = 0$ and

$$0\phi(p/0) = p \lim_{n \to \infty} \frac{\phi(n)}{n}.$$

Let us denote the set of all functions ϕ in these conditions by Φ^*. In particular, taking $\phi_0(x) = x \log x - x + 1$, we obtain the Kullback-Leibler divergence measure, i.e.

$$D_{\phi_0}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})) = D_{Kullback}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})).$$

Let $\phi \in \Phi^*$ be differentiable at $x = 1$; then, the function given by

$$\psi(x) = \phi(x) - \phi'(1)(x - 1)$$

also belongs to Φ^* and has the additional property that $\psi'(1) = 0$. This property, together with the convexity, implies that $\psi(x) \geq 0$ for any $x \geq 0$. Moreover,

$$\begin{aligned}
D_{\psi}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})) &= \sum_{j=1}^{2^k} p(y_j, \boldsymbol{\lambda}, \boldsymbol{\eta}) \psi\left(\frac{\hat{p}_j}{p(y_j, \boldsymbol{\lambda}, \boldsymbol{\eta})}\right) \\
&= \sum_{j=1}^{2^k} p(y_j, \boldsymbol{\lambda}, \boldsymbol{\eta}) \left(\phi\left(\frac{\hat{p}_j}{p(y_j, \boldsymbol{\lambda}, \boldsymbol{\eta})}\right) - \phi'(1)\left[\frac{\hat{p}_j}{p(y_j, \boldsymbol{\lambda}, \boldsymbol{\eta})} - 1\right]\right) \\
&= \sum_{j=1}^{2^k} p(y_j, \boldsymbol{\lambda}, \boldsymbol{\eta}) \phi\left(\frac{\hat{p}_j}{p(y_j, \boldsymbol{\lambda}, \boldsymbol{\eta})}\right) \\
&= D_{\phi}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta}))
\end{aligned}$$

Since the two divergence measures coincide, we can consider the set Φ^* to be equivalent to the set

$$\Phi := \Phi^* \cap \{\phi : \phi'(1) = 0\}.$$

Based on the previous results, we can define the minimum ϕ-divergence estimator (MϕE) in LCM in the following way:

Definition 1. *Given a LCM with parameters $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_t)$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_u)$, the M$\phi$E of $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\eta})$ is any $\hat{\boldsymbol{\theta}}_{\phi}$ satisfying*

$$\hat{\boldsymbol{\theta}}_\phi = arg \min_{(\boldsymbol{\lambda},\boldsymbol{\eta})\in\boldsymbol{\Theta}} D_\phi(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta})).$$

Remark 1. *From a practical point of view, we must solve the following system of equations:*

$$\frac{\partial D_\phi(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta}))}{\partial s_j} = 0,\ j = 1, ..., u+t$$

being

$$s_j := \begin{cases} \lambda_j, & j = 1, ..., t \\ \eta_{j-t}, & j = t+1, ..., t+u \end{cases}. \tag{6}$$

It is not difficult to see that

$$\frac{\partial D_\phi(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta}))}{\partial \lambda_\alpha} = \sum_{\nu=1}^{2^k} \left\{ \frac{\partial p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})}{\partial \lambda_\alpha} \phi\left(\frac{\hat{p}_\nu}{p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})}\right) - \frac{\hat{p}_\nu}{p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})} \phi'\left(\frac{\hat{p}_\nu}{p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})}\right) \frac{\partial p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})}{\partial \lambda_\alpha} \right\}.$$

$$\frac{\partial D_\phi(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta}))}{\partial \eta_\beta} = \sum_{\nu=1}^{2^k} \left\{ \frac{\partial p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})}{\partial \eta_\beta} \phi\left(\frac{\hat{p}_\nu}{p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})}\right) - \frac{\hat{p}_\nu}{p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})} \phi'\left(\frac{\hat{p}_\nu}{p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})}\right) \frac{\partial p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})}{\partial \eta_\beta} \right\},$$

with

$$\frac{\partial p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})}{\partial \lambda_\alpha} = \sum_{j=1}^{m} w_j Pr(y_\nu / P_\nu \in C_j) \sum_{i=1}^{k} q_{ji\alpha}(y_{\nu i} - p_{ji}),\ \alpha = 1, ..., t$$

and

$$\frac{\partial p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})}{\partial \eta_\beta} = \sum_{j=1}^{m} w_j Pr(y_\nu / P_\nu \in C_j) \left[v_{j\beta} - p(y_\nu, \boldsymbol{\lambda},\boldsymbol{\eta})\right],\ \beta = 1, ..., u.$$

In next section we study the behavior of the MϕE in LCM for large sample sizes.

3. Asymptotic properties of the MϕE in LCM

Let us denote by $(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0) = (\lambda_1^0, ..., \lambda_t^0, \eta_1^0, ..., \eta_u^0)$ the true value of the parameter $(\boldsymbol{\lambda},\boldsymbol{\eta})$ and let us assume that it is an interior point of the parameter space $\boldsymbol{\Theta}$. Let us denote by $\boldsymbol{\Delta}_{2^k}$ the set

$$\boldsymbol{\Delta}_{2^k} := \left\{ \mathbf{p} = (p_1, ..., p_{2^k})^T : p_i \geq 0,\ i = 1, ..., 2^k,\ \sum_{j=1}^{2^k} p_j = 1 \right\}.$$

In this section we shall assume that Birch's conditions hold:

i) $p(y_\nu, \boldsymbol{\lambda}_0, \boldsymbol{\eta}_0) > 0$, $\nu = 1, ..., 2^k$. Thus,

$$\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0) = (p(y_\nu, \boldsymbol{\lambda}_0, \boldsymbol{\eta}_0), ..., p(y_\nu, \boldsymbol{\lambda}_0, \boldsymbol{\eta}_0))$$

is an interior point of $\boldsymbol{\Delta}_{2^k}$. In the following, and in order to avoid hard notation, we will denote $p(y_\nu, \boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)$ by $p_\nu(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)$.

ii) The mapping $\mathbf{p} : \boldsymbol{\Theta} \to \boldsymbol{\Delta}_{2^k}$ assigning to any $(\boldsymbol{\lambda}, \boldsymbol{\eta})$ the vector $\mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta})$ is continuous and totally differentiable at $(\boldsymbol{\lambda}_0, \boldsymbol{\theta}_0)$.

iii) The Jacobian matrix

$$\mathbf{J}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0) := \left(\frac{\partial p_\nu(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)}{\partial s_j} \right)_{\substack{\nu=1,...,2^k \\ j=1,...,u+t}}$$

is of rank $u + t$.

iv) The inverse mapping of $\mathbf{p}$ is continuous at $\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)$.

Now, the following can be proved:

Theorem 1. *Suppose $\phi(t)$ is twice continuously differentiable at any $t > 0$. The MϕE, $\hat{\boldsymbol{\theta}}_\phi$, for the LCM satisfies*

$$\hat{\boldsymbol{\theta}}_\phi = (\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T + (\mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0))^{-1} \mathbf{A}(\boldsymbol{\theta}_0)^T \mathbf{D}^{-\frac{1}{2}}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)} (\hat{\mathbf{p}} - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)) + o(\|\hat{\mathbf{p}} - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)\|),$$

where $\mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0) := \mathbf{D}^{-\frac{1}{2}}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)} \mathbf{J}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)$ and by $\mathbf{D}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)}$ we are denoting the diagonal matrix whose diagonal is given by $\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)$.

Proof: See Appendix.

We can observe in the previous theorem that the expansion obtained for the MϕE in LCM does not depend on the function ϕ. This fact is very important because, based on it, in next theorem we shall establish that the asymptotic distribution of $\hat{\boldsymbol{\theta}}_\phi$ does not depend on ϕ.

Theorem 2. *Under the assumptions of the previous theorem, the MϕE, $\hat{\boldsymbol{\theta}}_\phi$, for the LCM satisfies*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_\phi - (\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T) \xrightarrow{L} \mathcal{N}(\mathbf{0}, (\mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0))^{-1}).$$

Proof: See Appendix.

Finally, let us present a result in relation to the estimated manifest probabilities, $\mathbf{p}(\hat{\boldsymbol{\theta}}_\phi)$.

Theorem 3. *Under the assumptions of the previous theorems, the estimated manifest probabilities satisfy*

$$\sqrt{n}(\mathbf{p}(\hat{\boldsymbol{\theta}}_\phi) - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \mathbf{J}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T \left(\mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^t \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0) \right)^{-1} \mathbf{J}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)).$$

Proof: See Appendix.

4. **Simulation study**

In this section we will carry out a simulation study to search some alternative minimum ϕ-divergence estimators to the maximum likelihood estimator. As explained before, the maximum likelihood estimator corresponds to the value $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ satisfying

$$\hat{\boldsymbol{\theta}} = arg \min_{(\boldsymbol{\lambda},\boldsymbol{\eta})\in\boldsymbol{\Theta}} D_{Kullback}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta})),$$

being $D_{Kullback}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta}))$ the Kullback-Leibler divergence between the probability vectors $\hat{\mathbf{p}}$ and $\mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta}))$. We shall consider in our simulation study the family of divergences introduced in Cressie and Read (1984). This family of divergence measures, called the *power-divergence family*, is obtained from (5) with

$$\phi(x) \equiv \phi_a(x) = \begin{cases} \frac{1}{a(a+1)}(x^a - x - a(x-1)) & a \neq 0, a \neq -1 \\ x\log x - x + 1 & a = 0 \\ -\log x + x - 1 & a = -1 \end{cases} \tag{7}$$

Based on (7), we get the minimum power-divergence estimator by

$$\hat{\boldsymbol{\theta}}_a = arg \min_{(\boldsymbol{\lambda},\boldsymbol{\eta})\in\boldsymbol{\Theta}} D_a(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta})),$$

where by $D_a(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta}))$ we denote $D_{\phi_a}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta}))$, whose expression is

$$D_a(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta})) = \begin{cases} \frac{1}{a(a+1)}\sum_{j=1}^{2^k}\left(\frac{\hat{p}_j^{a+1}}{p_j(\boldsymbol{\lambda},\boldsymbol{\eta})^a} - 1\right) & a \neq 0, a \neq -1 \\ D_{Kullback}(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta})) & a = 0 \\ \sum_{j=1}^{2^k} p_j(\boldsymbol{\lambda},\boldsymbol{\eta}))\log\frac{p_j(\boldsymbol{\lambda},\boldsymbol{\eta})}{\hat{p}_j} & a = -1 \end{cases}$$

Observe that for $a = 0$ we recover the maximum likelihood estimator. In order to do our simulation study we have considered a theoretical model with 5 dichotomous questions and 10 latent classes; next, 7 parameters λ_j and 6 parameters η_k are considered; the corresponding matrices of the model are

$$Q_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, Q_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}, Q_3 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}, Q_4 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$Q_5 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \; Q_6 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \; Q_7 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Matrix $\mathbf{C}$ is the null matrix. Matrix V is given by

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

while $\mathbf{d} = \mathbf{0}$. The theoretical values for vector $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ are

$$\boldsymbol{\lambda}_0 = (\lambda_1^0, ..., \lambda_7^0) = (-3, -2, -1, 0, 1, 2, 3), \; \boldsymbol{\eta}_0 = (\eta_1^0, ..., \eta_6^0) = (0.5, 1, 1.5, 2, 2.5, 3).$$

We consider three different values of the sample size N ($N = 50, 100, 200$) and several values of a ($a = -1, -\frac{1}{2}, 0, \frac{2}{3}, 1, 1.5, 2, 2.5, 3$). For each combination (N, a) we have conducted $n = 1000$ simulations.

The method used to obtain the estimators was a multistart minimization algorithm. For each sample size N we generate 1000 random samples (simulations). These 1000 samples are the same for each parameter. We apply the following algorithm to estimate the parameters:

Step 1. Initialization. Let

$$R = \{(\boldsymbol{\lambda}, \boldsymbol{\eta}) : \lambda_{i,lo} \leq \lambda_i \leq \lambda_i^{up}, \; i = 1, ..., t, \; \eta_{j,lo} \leq \eta_j \leq \eta_j^{up}, \; j = 1, ..., u\}$$

be the region considered to seek the estimators, where $\lambda_{i,lo}, \lambda_i^{up}, \eta_{j,lo}, \eta_j^{up}, \; \forall i, j$ are fixed values. Let us denote by N_{in} the number of points to be randomly generated in R. Initialize $D_\phi^{min} = +\infty, D_\phi^{in} = +\infty, i = 1$.

Step 2. Rough improvement. Generate $(\boldsymbol{\lambda}, \boldsymbol{\eta})_i$ and perform a full iteration of a variant[1] of the Hooke and Jeeves algorithm (Hooke&Jeeves (1961)) with an additional improvement step. Let $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\eta}})_i$ be the point obtained through this procedure.

[1]The used variant consists in permuting randomly the direction of the parameters. The additional improvement

If $D_\phi(\hat{\mathbf{p}}, (\boldsymbol{\lambda}, \boldsymbol{\eta})_i) < D_\phi^{in}$, set $D_\phi^{in} = D_\phi(\hat{\mathbf{p}}, (\boldsymbol{\lambda}, \boldsymbol{\eta})_i)$ and go to step 3; otherwise, go to step 4.

Step 3. Fine improvement. From $(\boldsymbol{\lambda}, \boldsymbol{\eta})_i$ as initial point, perform the limited memory quasi-Newton conjugate gradient algorithm (Gill and Murray (1979)) and let us denote by $(\boldsymbol{\lambda}^0, \boldsymbol{\eta}^0)_i$ the point obtained through this procedure.

From $(\boldsymbol{\lambda}^0, \boldsymbol{\eta}^0)_i$ as initial point, solve the system $\nabla D_\phi(\hat{\mathbf{p}}, (\boldsymbol{\lambda}, \boldsymbol{\eta})) = 0$ using the hybrid algorithm of Powell (Powell (1970)) and let us denote by $(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}})_i$ the solution obtained through this procedure. If $D_\phi(\hat{\mathbf{p}}, (\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}})_i) < D_\phi^{min}$, set $D_\phi^{min} = D_\phi(\hat{\mathbf{p}}, (\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}})_i)$, $(\boldsymbol{\lambda}, \boldsymbol{\eta})^{min} = (\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}})_i$.

Step 4. Stop. If $i = N_{in}$, stop; otherwise let $i = i + 1$ and go to step 2.

The computations are performed by our own Fortran programs with the NAG Fortran Library. For simulation l we get the values

$$\hat{\lambda}_{a,l}^j,\ j = 1, ..., t, \quad \hat{\eta}_{a,l}^k,\ k = 1, ..., u,$$

i.e. we obtain two vectors $\hat{\boldsymbol{\lambda}}_a^{(l)} = (\hat{\lambda}_{a,l}^1, ..., \hat{\lambda}_{a,l}^t)$ and $\hat{\boldsymbol{\eta}}_a^{(l)} = (\hat{\eta}_{a,l}^1, ..., \hat{\eta}_{a,l}^u)$ being $\hat{\boldsymbol{\theta}}_a^{(l)} = (\hat{\boldsymbol{\lambda}}_a^{(l)}, \hat{\boldsymbol{\eta}}_a^{(l)})$ the minimum power-divergence estimator obtained for the l-th simulation using $D_a(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta}))$. Next, $\hat{\boldsymbol{\theta}}_a = (\hat{\boldsymbol{\lambda}}_a, \hat{\boldsymbol{\eta}}_a)$ is defined as

$$\hat{\lambda}_a^j = \frac{1}{n}\sum_{l=1}^n \hat{\lambda}_{a,l}^j, \quad \hat{\eta}_a^k = \frac{1}{n}\sum_{l=1}^n \hat{\eta}_{a,l}^k.$$

We compute

$$mse(\lambda_j) = \frac{1}{n}\sum_{l=1}^n (\hat{\lambda}_a^j - \lambda_j^0)^2, \quad mse(\eta_k) = \frac{1}{n}\sum_{l=1}^n (\hat{\eta}_a^k - \eta_k^0)^2$$

and at the same time

$$mse_{\boldsymbol{\lambda}} = \frac{1}{t}\sum_{j=1}^t mse(\lambda_j), \quad mse_{\boldsymbol{\eta}} = \frac{1}{u}\sum_{k=1}^u mse(\eta_k).$$

In Table 1 we present the values of $mse_{\boldsymbol{\lambda}}, mse_{\boldsymbol{\eta}}$ and

$$mse = \frac{1}{t+u}(t\ mse(\boldsymbol{\lambda}) + u\ mse(\boldsymbol{\eta}))$$

for each combination (N, a).

From this table, we can observe that there are some power divergence estimators with better behavior than the MLE for any sample size N; for example, this is the case for $a = \frac{3}{2}$.

5. A numerical example

In order to study the extension proposed in this paper we have considered the interview data colected by Coleman and analized in Goodman (1974); this model is explained in Formann

consists in exploring the direction from the initial point to the final point in legs with double, exterior half or interior half spacing between both points. At most we need $2(t+u)+4$ evaluations of D_ϕ. This criterion is used in order to discard non promising initial points from a finer and most costly improvement.

N	a	$mse_{\boldsymbol{\lambda}}$	$mse_{\boldsymbol{\eta}}$	mse
50	-1	14.53738	12.85120	13.75914
	$-\frac{1}{2}$	9.39724	8.57172	9.01623
	0	8.82961	8.47212	8.66462
	$\frac{2}{3}$	8.47918	8.58022	8.52582
	1	8.50968	8.55418	8.53022
	1.5	8.28130	8.47194	**8.36929**
	2	8.22638	8.59321	8.39569
	2.5	8.43137	8.67074	8.54185
	3	8.23336	8.59487	8.40021
100	-1	14.05111	12.62465	13.39274
	$-\frac{1}{2}$	8.81442	8.32502	8.58854
	0	8.19294	8.63299	8.39604
	$\frac{2}{3}$	8.01635	8.61440	8.29237
	1	7.90969	8.54711	8.20389
	1.5	7.80300	8.53064	**8.13883**
	2	7.83062	8.58989	8.18105
	2.5	7.85925	8.51367	8.16129
	3	7.77588	8.56704	8.14103
200	-1	12.12742	11.05020	11.63024
	$-\frac{1}{2}$	8.10155	8.52116	8.29522
	0	7.72814	8.33506	8.00826
	$\frac{2}{3}$	7.67556	8.47668	8.04531
	1	7.61732	8.65570	8.09657
	1.5	7.51919	8.53508	7.98806
	2	7.50521	8.69846	8.05594
	2.5	7.52651	8.50773	**7.97938**
	3	7.45609	8.62044	7.99348

Table 1: Results of the simulation study

(1982); the experiment consists in evaluating the answers of 6658 schoolboys to two questions about their membership in the "leading crowd" on two occasions. Thus, in this model we have 4 dichotomous questions; next, 4 latent classes are considered; then, there are 16 probability values p_{ji} to be estimated. A model with 8 parameters λ_i is considered; the corresponding matrices Q_i are

$$Q_1 := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \; Q_2 := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$Q_3 := \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \; Q_4 := \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Moreover, $c_{ij} = 0$, $\forall i, j$. Next, 4 parameters η_j are considered, taking matrix $V = Id$ and $d_j = 0$, $\forall j$.

If we look for the solution when $a = 1.5$, we obtain the following results:

Parameter	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
Value	-3.51036	0.45943	3.63001	-4.04896	-2.95335	-0.68536	-4.28890	0.31269

Parameter	η_1	η_2	η_3	η_4
Value	0.74496	-0.65399	-3.08647	-4.32671

Parameter	$p_{1,1}$	$p_{1,2}$	$p_{1,3}$	$p_{1,4}$	$p_{2,1}$	$p_{2,2}$	$p_{2,3}$	$p_{2,4}$
Value	0.02902	0.97417	0.04958	0.01353	0.02902	0.01714	0.04958	0.57754

Parameter	$p_{3,1}$	$p_{3,2}$	$p_{3,3}$	$p_{3,4}$	$p_{4,1}$	$p_{4,2}$	$p_{4,3}$	$p_{4,4}$
Value	0.61288	0.97417	0.33507	0.01353	0.61288	0.01714	0.33507	0.57754

Parameter	w_1	w_2	w_3	w_4
Value	0.78443	0.19364	0.01701	0.00492

6. **Conclusions**

From a classical point of view the unknown parameters in LCM have been estimated using Maximum Likelihood Estimators. In this paper, using the parametrization for LCM proposed in Formann (1985), we introduce and study the family of minimum power divergence estimators. This family can be considered as an extension of the MLE in the sense that the MLE is an estimator included in this family. The assymptotic distribution of all these estimators is the same, but their behavior may vary for small or moderate sample sizes. This fact is pointed out on the basis of a simulation study. From this study, we think that some estimators in this family could exhibit a better behavior than the MLE in LCM.

7. **Acknowledgements**

This work was partially supported by Grant MTM2012-33740.

Appendix

Proof of Theorem 1.

Let l^{2^k} be the interior of the 2^k-dimensional unit cube; then, the interior of Δ_{2^k} is contained in l^{2^k}. Let V be a neighborhood of $(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)$, the true value of the unknown parameter $(\boldsymbol{\lambda}, \boldsymbol{\eta})$, on which

$$\begin{array}{rccl} \mathbf{p}: & \Theta & \to & \Delta_{2^k} \\ & (\boldsymbol{\lambda}, \boldsymbol{\eta}) & \mapsto & \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta}) := (p_1(\boldsymbol{\lambda}, \boldsymbol{\eta}), ..., p_{2^k}(\boldsymbol{\lambda}, \boldsymbol{\eta})) \end{array}$$

has continuous second partial derivatives. Let

$$\mathbf{F} := (F_1, ..., F_{t+u}) : l^{2^k} \times V \to \mathbb{R}^{t+u}$$

whose components F_j, $j = 1, ..., t+u$ are defined by

$$F_j(\tilde{p}_1, ..., \tilde{p}_{2^k}; \lambda_1, ..., \lambda_t; \eta_1, ..., \eta_u) := \frac{\partial D_\phi(\tilde{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta}))}{\partial s_j},\ j = 1, ..., t+u,$$

where s_j is defined in (6).

It holds

$$F_j(p_1(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0), ..., p_{2^k}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0); \lambda_1^0, ..., \lambda_t^0; \eta_1^0, ..., \eta_u^0) = 0,\ \forall j = 1, ..., t+u$$

due to

$$\frac{\partial D_\phi(\tilde{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta}))}{\partial \lambda_s} = \sum_{l=1}^{2^k} \left\{ \phi\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})}\right) - \frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})} \phi'\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})}\right) \right\} \frac{\partial p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})}{\partial \lambda_s},\ s = 1, ..., t.$$

$$\frac{\partial D_\phi(\hat{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta}))}{\partial \eta_a} = \sum_{l=1}^{2^k} \left\{ \phi\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})}\right) - \frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})} \phi'\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})}\right) \right\} \frac{\partial p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})}{\partial \eta_a},\ a = 1, ..., u.$$

In the following we shall include the two previous expressions by

$$\frac{\partial D_\phi(\tilde{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta}))}{\partial s_j},\ j = 1, ..., t+u.$$

Since

$$\begin{aligned}\frac{\partial}{\partial s_r}\left(\frac{\partial D_\phi(\tilde{\mathbf{p}},\mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta}))}{\partial s_j}\right) &= -\sum_{l=1}^{2^k}\phi'\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}\right)\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda},\boldsymbol{\eta})^2}\frac{\partial p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}{\partial s_r}\frac{\partial p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}{\partial s_j}\\ &+\sum_{l=1}^{2^k}\phi''\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}\right)\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda},\boldsymbol{\eta})^2}\frac{\partial p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}{\partial s_r}\frac{\partial p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}{\partial s_j}\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}\\ &+\sum_{l=1}^{2^k}\phi'\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}\right)\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda},\boldsymbol{\eta})^2}\frac{\partial p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}{\partial s_r}\frac{\partial p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}{\partial s_j}\\ &+\sum_{l=1}^{2^k}\frac{\partial^2 p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}{\partial s_r s_j}\left\{\phi\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}\right)-\phi'\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}\right)\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda},\boldsymbol{\eta})}\right\},\end{aligned}$$

and denoting $\pi_i = p_i(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0), i = 1,...,2^k$, the $(t+u)\times(t+u)$ matrix $\mathbf{J}_\mathbf{F}$ associated with function $\mathbf{F}$ at point $(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0),(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))$ is given by

$$\begin{aligned}\frac{\partial \mathbf{F}}{\partial(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)} &= \left(\frac{\partial \mathbf{F}}{\partial(\boldsymbol{\lambda},\boldsymbol{\eta})}\right)_{(\tilde{\mathbf{p}},(\boldsymbol{\lambda},\boldsymbol{\eta}))=(\pi_1,...,\pi_{2^k};\lambda_1^0,...,\lambda_t^0;\eta_1^0,...,\eta_u^0)}\\ &= \left(\left(\frac{\partial}{s_r}\left(\frac{\partial D_\phi(\tilde{\mathbf{p}},\mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta}))}{\partial s_j}\right)\right)_{\substack{j=1,...,t+u\\ r=1,...,t+u}}\right)_{(\tilde{\mathbf{p}},(\boldsymbol{\lambda},\boldsymbol{\eta}))=(\pi_1,...,\pi_{2^k};\lambda_1^0,...,\lambda_t^0;\eta_1^0,...,\eta_u^0)}\\ &= \phi''(1)\left(\sum_{l=1}^{2^k}\frac{1}{p_l(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}\frac{\partial p_l(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}{\partial s_r}\frac{\partial p_l(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}{\partial s_j}\right)_{\substack{j=1,...,t+u\\ r=1,...,t+u}}\end{aligned}$$

To get the last expression we are using that $\phi(1) = \phi'(1) = 0$. Recall that if $\mathbf{B}$ is a $p\times q$ matrix with $rank(\mathbf{B}) = p$ and $\mathbf{C}$ is a $q\times s$ matrix with $rank(\mathbf{C}) = q$, then $rank(\mathbf{BC}) = p$. Taking

$$\mathbf{B} = \mathbf{J}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T,\ \mathbf{C} = \mathbf{D}^{-\frac{1}{2}}_{\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)},$$

it follows that $\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T = \mathbf{BC}$ has rank $u+t$ applying the fourth condition of Birch. Also,

$$rank(\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)) = rank(\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T) = rank(\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)) = u+t.$$

Therefore, the $(u+t)\times(u+t)$ matrix $\frac{\partial \mathbf{F}}{\partial(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}$ is nonsingular at $(\pi_1,...,\pi_{2^k};\lambda_1^0,...,\lambda_t^0;\eta_1^0,...,\eta_u^0)$.

Applying the Implicit Function Theorem, there exists a neighborhood U of $(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0),(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))$ such that the matrix $\mathbf{J}_\mathbf{F}$ is nonsingular (in our case $\mathbf{J}_\mathbf{F}$ at $(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0),(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))$ is positive definite and then it is continuously differentiable). Also, there exists a continuously differentiable function

$$\tilde{\boldsymbol{\theta}} : A\subset l^{2^k}\to\mathbb{R}^{u+t}$$

such that $\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)\in A$ and

$$\{(\tilde{\mathbf{p}},(\boldsymbol{\lambda},\boldsymbol{\eta}))\in U:\mathbf{F}(\tilde{\mathbf{p}},(\boldsymbol{\lambda},\boldsymbol{\eta}))=0\}=\left\{(\tilde{\mathbf{p}},\tilde{\boldsymbol{\theta}}(\tilde{\mathbf{p}})):\tilde{\mathbf{p}}\in A\right\}. \tag{8}$$

We can observe that $\tilde{\boldsymbol{\theta}}(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))$ is an argmin of

$$\psi(\boldsymbol{\lambda},\boldsymbol{\eta}):=D_\phi(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0),\mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta}))$$

because $\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)\in A$ and then

$$\mathbf{F}(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0),\tilde{\boldsymbol{\theta}}(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)))=\frac{\partial D_\phi(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0),\mathbf{p}(\tilde{\boldsymbol{\theta}}(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))))}{\partial(\boldsymbol{\lambda},\boldsymbol{\eta})}=\mathbf{0}.$$

On the other hand, applying (8),

$$(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0),\tilde{\boldsymbol{\theta}}(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)))\in U,$$

and then $\mathbf{J_F}$ is positive definite at $(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0),\tilde{\boldsymbol{\theta}}(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)))$. Therefore,

$$D_\phi(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0),\mathbf{p}(\tilde{\boldsymbol{\theta}}(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))))=\inf_{(\boldsymbol{\lambda},\boldsymbol{\eta})\in\Theta}D_\phi(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0),\mathbf{p}(\boldsymbol{\lambda},\boldsymbol{\eta})),$$

and by the ϕ-divergence properties $\tilde{\boldsymbol{\theta}}(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))=(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T$, and

$$\frac{\partial\mathbf{F}}{\partial\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}+\frac{\partial\mathbf{F}}{\partial(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}\frac{\partial(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}{\partial\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}=\mathbf{0}.$$

Further, we know that

$$\frac{\partial\mathbf{F}}{\partial(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}=\phi''(1)\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)$$

and we shall establish later that the $(u+t)\times 2^k$ matrix $\frac{\partial\mathbf{F}}{\partial\boldsymbol{\pi}}$ is

$$\frac{\partial\mathbf{F}}{\partial\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}=-\phi''(1)\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T\mathbf{D}_{\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}^{-\frac{1}{2}}. \tag{9}$$

Therefore, the $(u+t)\times 2^k$ matrix $\frac{\partial(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}{\partial\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}$ is

$$\frac{\partial(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}{\partial\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}=(\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))^{-1}\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T\mathbf{D}_{\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)}^{-\frac{1}{2}}.$$

The Taylor expansion of the function $\tilde{\boldsymbol{\theta}}$ around $\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)$ yields

$$\tilde{\boldsymbol{\theta}}(\tilde{\mathbf{p}})=\tilde{\boldsymbol{\theta}}(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))+\left(\frac{\partial\tilde{\boldsymbol{\theta}}(\tilde{\mathbf{p}})}{\tilde{\mathbf{p}}}\right)_{\tilde{\mathbf{p}}=\boldsymbol{\pi}}(\tilde{\mathbf{p}}-\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))+o(\|\tilde{\mathbf{p}}-\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)\|).$$

As $\tilde{\boldsymbol{\theta}}(\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))=(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T$, we obtain from here

$$\tilde{\boldsymbol{\theta}}(\tilde{\mathbf{p}})=(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T+(\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))^{-1}\mathbf{A}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)^T\mathbf{D}_{\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0}^{-\frac{1}{2}}(\tilde{\mathbf{p}}-\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0))+o(\|\tilde{\mathbf{p}}-\mathbf{p}(\boldsymbol{\lambda}_0,\boldsymbol{\eta}_0)\|).$$

We know that $\hat{\mathbf{p}} \xrightarrow{a.s.} \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)$, so that $\hat{\mathbf{p}} \in A$ and, consequently, $\tilde{\boldsymbol{\theta}}(\hat{\mathbf{p}})$ is the unique solution of the system of equations

$$\frac{\partial D_\phi(\hat{\mathbf{p}}, \mathbf{p}(\tilde{\boldsymbol{\theta}}(\hat{\mathbf{p}})))}{s_j} = 0,\ j = 1, ..., u+t,$$

and also $(\hat{\mathbf{p}}, \tilde{\boldsymbol{\theta}}(\hat{\mathbf{p}})) \in U$. Therefore, $\tilde{\boldsymbol{\theta}}(\hat{\mathbf{p}})$ is the minimum ϕ-divergence estimator, $\hat{\boldsymbol{\theta}}_\phi$, satisfying the relation

$$\hat{\boldsymbol{\theta}}_\phi = (\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T + (\mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0))^{-1} \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T \mathbf{D}^{-\frac{1}{2}}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)} (\hat{\mathbf{p}} - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)) + o(\|\hat{\mathbf{p}} - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)\|).$$

Finally, we are going to establish (9). We compute the (i,j)-th element of the $(u+t) \times 2^k$ matrix $\frac{\partial \mathbf{F}}{\partial \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)}$.

$$\begin{aligned}
\frac{\partial}{\partial p_i}\left(\frac{\partial D_\phi(\tilde{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta}))}{\partial s_j}\right) &= \frac{\partial}{\partial p_i}\left(\sum_{l=1}^{2^k}\left\{\phi\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})}\right) - \phi'\left(\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})}\right)\frac{\tilde{p}_l}{p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})}\right\}\frac{\partial p_l(\boldsymbol{\lambda}, \boldsymbol{\eta})}{\partial s_j}\right) \\
&= \frac{1}{p_i(\boldsymbol{\lambda}, \boldsymbol{\eta})}\left(-\frac{p_i}{p_i(\boldsymbol{\lambda}, \boldsymbol{\eta})}\phi''\left(\frac{p_i}{p_i(\boldsymbol{\lambda}, \boldsymbol{\eta})}\right)\right)\frac{\partial p_i(\boldsymbol{\lambda}, \boldsymbol{\eta})}{\partial s_j}
\end{aligned}$$

and for $(\pi_1, ..., \pi_{2^k}; \lambda_1^0, ..., \lambda_t^0; \eta_1^0, ..., \eta_u^0)$ we have

$$\frac{\partial}{\partial p_i}\left(\frac{\partial D_\phi(\tilde{\mathbf{p}}, \mathbf{p}(\boldsymbol{\lambda}, \boldsymbol{\eta}))}{\partial s_j}\right) = \frac{1}{p_i(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)}\phi''(1)\frac{\partial p_i(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)}{\partial s_j}.$$

Since $\mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0) = \mathbf{D}^{-\frac{1}{2}}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)} \mathbf{J}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)$, then (9) holds. ■

Proof of Theorem 2.
Applying the previous theorem it holds

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_\phi - (\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T) = \left(\mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)\right)^{-1} \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0) \mathbf{D}^{-\frac{1}{2}}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)} \sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)) + \sqrt{n}\ o(\|\hat{\mathbf{p}} - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)\|).$$

Note that

$$\sqrt{n}\ o(\|\hat{\mathbf{p}} - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)\|) = o_p(1).$$

On the other hand, as $\hat{\mathbf{p}}$ is the sample proportion, we can apply the Central Limit Theorem to conclude

$$\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)}),$$

where $\boldsymbol{\Sigma}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)}$ is given by

$$\boldsymbol{\Sigma}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)} = \mathbf{D}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)} - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T.$$

Therefore, it follows

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_\phi - (\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^*),$$

where $\boldsymbol{\Sigma}^*$ is given by

$$\boldsymbol{\Sigma}^* = \left(\mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)\right)^{-1} - \mathbf{B}\mathbf{B}^T$$

with $\mathbf{B} := \left(\mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)\right)^{-1} \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T \mathbf{D}^{\frac{1}{2}}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)}$.

It is not difficult to see that

$$\mathbf{D}^{\frac{1}{2}}_{\mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)} \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0) = 0,$$

whence $\mathbf{B} = \mathbf{0}$ and the result holds. ■

Proof of Theorem 3.

Using Theorem 2, it suffices to apply the delta method. Then, we can conclude that

$$\sqrt{n}(\mathbf{p}(\hat{\boldsymbol{\theta}}_\phi) - \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \nabla \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^T \left(\mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)^t \mathbf{A}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)\right)^{-1} \nabla \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)).$$

Now, as $\nabla \mathbf{p}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0) = \mathbf{J}(\boldsymbol{\lambda}_0, \boldsymbol{\eta}_0)$, we have the result. ■

References

- Abar, B. and Loken, E. (2010). Self-regulated learning and self-directed study in a pre-college sample. *Learning and Individual Differences*, 20:25–29.

- Biemer, P. (2011). *Latent Class Analysis and Survey Error*. John Wiley and Sons.

- Caldwell, L., Bradley, S., and Coffman, D. (2009). A person-centered approach to individualizing a scool-based universal preventive intervention. *American Journal of Drug and Alcohol Abuse*, 35(4):214–219.

- Clogg, C. (1995). Latent class models: Recent developments and prospects for the future. In Arminger, C. G. and Sobol, M., editors, *Handbook of statistical modeling for the social and behavioral sciences*, pages 311–352. Plenum, New York (USA).

- Coffman, D., Patrick, M., Polen, L., Rhoades, B., and Ventura, A. (2007). Why do high school seniors drink? Implication for a targeted approach to intervention. *Prevention Science*, 8:1–8.

- Collins, L. and Lanza, S. (2010). *Latent class and latent transition analysis for the social, behavioral, and health sciences*. Wiley, New York (USA).

- Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

- Feldman, B., Masyn, K., and Conger, R. (2009). New approaches to studying behaviors: A comparison of methods for modelling longitudinal, categorical and adolescent drinking data. *Development Psycology*, 45(3):652–676.

- Formann, A. (1976). Schätzung der Parameter in Lazarsfeld Latent-Class Analysis. In *Res. Bull.*, number 18. Institut für Psycologie der Universität Wien. In German.

- Formann, A. (1977). Log-linear Latent Class Analyse. In *Res. Bull.*, number 20. Institut für Psycologie der Universität Wien. In German.

- Formann, A. (1978). A note on parametric estimation for Lazarsfeld's latent class analysis. *Psychometrika*, 48:123–126.

- Formann, A. (1982). Linear logistic latent class analysis. *Biometrical Journal*, 24:171–190.

- Formann, A. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematics and Statistical Psicology*, 38:87–111.

- Gerber, M., Witterkind, A., Grote, G., and Staffelbach, B. (2009). Exploring types of career orientation: a latent class analysis approach. *Journal of Vocational Behavior*, 75:303–318.

- Gill, P. E. and Murray, W. (1979). Conjugate-gradient methods for large-scale nonlinear optimization. *Technical Report SOL 79-15*. Department of Operations Research, Stanford University.

- Goodman, L. A. (1974). Exploratory latent structure analysis using Goth identifiable and unidentifiable models. *Biometrika*, 61:215–231.

- Hagenaars, J. A. and Cutcheon, A. L. M. (2002). *Applied Latent Class Analysis*. Cambridge University Press, Cambridge (UK).

- Hooke, R. and Jeeves, T. A. (1961). Direct Search Solution of Numerical and statistical Problems. *Journal of the Association for Computing Machinery*, 8:212–229.

- Langeheine, R. and Rost, J. (1988). *Latent Trait and Latent Class Models*. Plenum Press, New York (USA).

- Laska, M., Pash, K., Lust, K., Story, M., and Ehlinger, E. (2009). Latent class analysis of lifestyle characteristics and health risk behaviors among college youth. *Prevention Sciences*, 10:376–386.

- Lazarsfeld, P. and Henry, N. (1968). *Latent structure analysis*. Houghton-Mifflin, Boston (USA).

- Lazarsfeld, P. (1950). The logical and mathematical foundation of latent structure analysis. In *Studies in Social Psycology in World War II, vol. IV: Measurement and prediction*, pages 362–412. Princeton University Press.

- McHugh, R. (1956). Efficient estimation and local identification in Latent Class Analysis. *Psychometrika*, 21:331–347.

- Morales, D., Pardo, L., and Vajda, I. (1995). Asymptotic divergence of estimators of discrete distributions. *Jounal of Statistical Planning and Inference*, 48:347–369.

- Nylund, K., Bellmore, A., Nishina, A., and Grahan, S. (2007). Subtypes, severity and structural stability of peer victimization: What does latent class analysis say? *Child Prevention*, 78:1706–1722.

- Pardo, L. (2006). *Statistical Inference based on Divergence Measures*. Chapman & Hall CRC.

- Powell M.(1970). A hybrid method for nonlinear algebraic equations. In Rabinowitz, P. editor. *Numerical Methods for Nonlinear Algebraic Equations*. Gordon and Breach.

- Rost, J. and Langeheine, R. (1997). *Applications of Latent trait and Latent Class Models in the Social Sciences*. Waxmann, Münster (Germany).