Computational Statistics (2020) 35:1133–1152
https://doi.org/10.1007/s00180-020-00960-w

ORIGINAL PAPER

Discrete factor analysis using a dependent Poisson model

Rolf Larsson [1]

Received: 16 April 2019 / Accepted: 21 January 2020 / Published online: 31 January 2020

Abstract
In this paper, we present a method for factor analysis of discrete data. This is accomplished by fitting a dependent Poisson model with a factor structure. To be able to analyze ordinal data, we also consider a truncated Poisson distribution. We try to find the model with the lowest AIC by employing a forward selection procedure. The probability to find the correct model is investigated in a simulation study. Moreover, we heuristically derive the corresponding asymptotic probabilities. An empirical study is also included.

Keywords AIC · Model selection · Ordinal data

1 Introduction

The main idea of classical factor analysis (see e.g. Jöreskog et al. 2016) is to describe a random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ as a linear combination of unknown factors $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)'$ plus some independent random error $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$ say, where $k < n$. Introducing the $n \times k$ matrix of factor loadings $\boldsymbol{\Lambda}$, we then have the equation

$$\mathbf{Y} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}. \tag{1}$$

Restricting the ξ_j to be uncorrelated with zero mean and unit variance, and uncorrelated with $\boldsymbol{\delta}$, the covariance matrix of $\mathbf{Y}$ is

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \tag{2}$$

where $\boldsymbol{\Psi}$ is the covariance matrix of $\boldsymbol{\delta}$, usually assumed to be diagonal.

✉ Rolf Larsson
rolf.larsson@math.uu.se

[1] Department of Mathematics, Uppsala University, P.O. Box 480, 751 06 Uppsala, Sweden

The main focus of (explanatory) factor analysis is to find out about the structure of the loading matrix $\mathbf{\Lambda}$. A common way to deal with this is to let $\mathbf{Y}$ be multivariate normal with zero mean and covariance matrix $\mathbf{\Sigma}$ as in (2), and estimate the parameters by maximum likelihood, see Jöreskog (1967). Observe that the $\mathbf{\Lambda}$ matrix is unique only up to rotation, i.e. $\mathbf{\Lambda}\mathbf{\Lambda}' = \mathbf{\Lambda}^*\mathbf{\Lambda}^{*\prime}$ for any $\mathbf{\Lambda}^* = \mathbf{\Lambda}\mathbf{T}$ where $\mathbf{T}$ is some orthogonal $k \times k$ matrix.

If the assumption of uncorrelated (and unit variance) ξ_j is relaxed, so that the covariance matrix of $\boldsymbol{\xi}$, $\mathbf{\Phi}$ say, is not necessarily diagonal, then the covariance matrix of $\mathbf{Y}$ is $\mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Psi}$. In this case, the rotation matrix $\mathbf{T}$ does not have to be orthogonal, but just invertible. To preserve the model structure, we add the factor transformation $\boldsymbol{\xi}^* = \mathbf{T}^{-1}\boldsymbol{\xi}$. This is so, because then,

$$\mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' = \mathbf{\Lambda}\mathbf{T}\mathbf{T}^{-1}\mathbf{\Phi}\mathbf{T}^{-1'}\mathbf{T}'\mathbf{\Lambda}' = \mathbf{\Lambda}^*\mathbf{\Phi}^*\mathbf{\Lambda}^{*\prime}$$

where $\mathbf{\Phi}^* = \mathbf{T}^{-1}\mathbf{\Phi}\mathbf{T}^{-1'}$ is the covariance matrix of $\boldsymbol{\xi}^*$. In the factor anaysis literature, this is called oblique rotation.

Despite for the many appealing features of maximum likelihood, searching for the 'best' factor analysis model given data involves some more or less 'arbitrary' steps such as choosing the number of factors k and a suitable rotation matrix $\mathbf{T}$.

In applications, it is common that the data $\mathbf{x}$ are observed on an ordinal scale. The continuous variable factor analysis model in (1) can still be applied to this situation, see e.g. Jöreskog and Moustaki (2001). To this end, the observed data are considered as outcomes from an underlying continuous variable (preferrably normal) that may be described by the factor model (1). Here, a certain (integer) value of the data corresponds to an interval on the continuous scale, defined by threshold parameters. As with all the other parameters, the thresholds may be estimated by maximum likelihood. In terms of numerics, this is a quite formidable task. Hence, alternative procedures have been proposed, for example using polychoric correlations, see Olsson (1979), or likelihood approximations, see Katsikatsou et al. (2012).

Factor analysis with discrete data is performed by Zhou et al. (2012) and Wedel et al. (2003). The former proposes a fully Bayesian method where the parameter vector of the observed discrete variates is modelled with a factor structure similar to the classical Jöreskog model. The latter approach uses a generalized linear regression model, with a link function that is a function of covariates in a factor form. Like the classical method for continuous data, these two approaches are rather involved numerically and contain issues about factor rotation and determination of the number of factors.

In the present paper, we propose a completely different approach to discrete and ordinal data factor analysis. The basis of our approach is the dependent Poisson model, described in e.g. Karlis (2003). In particular, let U, X_1 and X_2 be independent Poisson variates. Then, the variates $Y_1 = U + X_1$ and $Y_2 = U + X_2$ are also Poisson, but they are now linked through the common factor U. Of course, this idea may be extended to arbitrary dimensions, and we could also think of a vector of variables $(Y_1, \ldots, Y_N)$ which may be split up into a number of independent sub systems of the type described.

This is then a way to construct a discrete factor model. To deal with ordinal data, we consider truncated distributions. To relax the requirement of independent sub systems, we propose a mixed model approach.

In fact, this factor model idea extends to any (combinations of) discrete distributions, but as a start, we only pursue Poisson in the present paper. This is because Poisson seems to be the simplest discrete distribution with only one parameter that may be useful for our purposes. Observe that there are many other ways to construct dependent systems of discrete random variables, e.g. via copulas, mixing (compound Poisson) and graphical models. However, none of these seems to produce a system with a factor structure. See further Inouye et al. (2017) for a recent overview.

As will be seen in the sequel, the construction of factor models in the way outlined here, as well as maximum likelihood estimation of them, is fairly straightforward. The issue that may be problematic and time consuming is how to choose the 'best' possible model among the very many possible suggestions for a given dimension k. In this paper, by the 'best' model we mean the one with the lowest value of the Akaike information criterion (AIC), see Akaike (1974). We propose to resolve this by employing a forward search algorithm. We will study the probability to find the 'correct' model (if there is one) by simulations in dimensions five (where we compare to selection among all possible models) and seven, and we also heuristically calculate the corresponding asymptotic probabilities.

The rest of the paper is as follows. In Sect. 2 we lay out the model and its estimation via maximum likelihood. The selection algorithm is presented and discussed in Sect. 3. Section 4 contains a simulation study. In Sect. 5, we give an empirical example with ordinal data that previously has been analysed by Jöreskog et al. (2016). Section 6 concludes.

2 Model and estimation

2.1 General

At first, let us repeat the Karlis bivariate model,

$$\begin{cases} Y_1 = U + X_1, \\ Y_2 = U + X_2, \end{cases} \tag{3}$$

where U, X_1 and X_2 are independent random variables that may attain non negative integer values. (At this stage, we do not impose the Poisson assumption.) We say that U is the "common factor" that "loads" on the variables Y_1 and Y_2.

It is easy to imagine a setup of a number of possibly dependent variables $Y_1, \ldots, Y_N$ which may be "linked" by a set of common factors $U_1, \ldots, U_k$ where $k < N$. If these factors are only allowed to load on one variable each, this gives the general model

$$\begin{cases} Y_1 = X_1, \\ Y_2 = X_2, \\ \vdots \\ Y_{n_0} = X_{n_0}, \\ Y_{n_0+1} = U_1 + X_{n_0+1}, \\ \vdots \\ Y_{n_0+n_1} = U_1 + X_{n_0+n_1}, \\ Y_{n_0+n_1+1} = U_2 + X_{n_0+n_1+1}, \\ \vdots \\ Y_{n_0+n_1+n_2} = U_2 + X_{n_0+n_1+n_2}, \\ \vdots \\ Y_{n_0+\cdots+n_k+1} = U_k + X_{n_0+\cdots+n_{k-1}+1}, \\ \vdots \\ Y_{n_0+\cdots+n_k} = U_k + X_{n_0+\cdots+n_k}, \end{cases} \tag{4}$$

where $N = n_0 + \cdots + n_k$ and $U_1, \ldots, U_k, X_1, \ldots, X_N$ are all assumed to be independent non negative integer valued random variables. Here, n_0 is the number of variables that are not linked to any others, n_1 is the number of variables linked to the first common factor (U_1), and so on. Moreover, we assume that $n_i \geq 2$ for $i = 1, 2, \ldots, k$. Observe that this setup also allows for a set of independent components $Y_1, \ldots, Y_{n_0}$. In the following, we will refer to this as a model of type $(n_1, n_2, \ldots, n_k, 1, \ldots, 1)$, where there are n_0 ones at the end. The variables may be shuffled around so that $n_1 \geq n_2 \geq \ldots \geq n_k$. For example, the model of type (3, 2, 1, 1) is given by

$$\begin{cases} Y_1 = X_1, \\ Y_2 = X_2, \\ Y_3 = U_1 + X_3, \\ Y_4 = U_1 + X_4, \\ Y_5 = U_1 + X_5, \\ Y_6 = U_2 + X_6, \\ Y_7 = U_2 + X_7. \end{cases} \tag{5}$$

We want to estimate the parameters of (4) by maximum likelihood. This is very feasible, since (4) consists of $n_0 + k$ simultaneously independent systems. Hence, the likelihood is the product of the likelihoods of all these systems, and the maximum likelihood is the product of the corresponding maximum likelihoods, which all may be evaluated separately. For example, in (5), the likelihood is a product of likelihoods of one three-dimensional system with one common factor, one two-dimensional system with a common factor and two separate one-dimensional variates.

We need to add distributional assumptions on the U_js and X_js. For example (cf Karlis 2003), we could assume that the U_j are Poisson with parameters λ_j and that the X_j are Poisson with parameters μ_j. Then, by the additivity property of the Poisson distribution, the Y_j are also Poisson, but dependent. The degree of dependence, mea-

sured by e.g. the correlation, is a function of the parameters. In the simplest example, (3) with $U \sim \text{Po}(\lambda)$ and $X_j \sim \text{Po}(\mu_j)$ for $j = 1, 2$, the correlation between Y_1 and Y_2 is given by

$$\text{corr}(Y_1, Y_2) = \frac{\lambda}{\sqrt{(\lambda + \mu_1)(\lambda + \mu_2)}}.$$

Observe that in this way, only positive correlations are allowed for.

In (3), introduce $f(u; \lambda)$ and $g(x; \mu_j)$ as the probability mass functions of U and the X_j respectively. We have a set of observation pairs $(y_{11}, y_{12}), \ldots, (y_{n1}, y_{n2})$. Since Y_1 and Y_2 are conditionally independent given U, the likelihood is

$$L(\lambda, \mu_1, \mu_2) = \prod_{i=1}^{n} \sum_{u=0}^{\min(y_{i1}, y_{i2})} f(u; \lambda) g(y_{i1} - u; \mu_1) g(y_{i2} - u; \mu_2). \tag{6}$$

Imposing the Poisson assumption, this becomes

$$\begin{aligned} L(\lambda, \mu_1, \mu_2) &= \prod_{i=1}^{n} \sum_{u=0}^{\min(y_{i1}, y_{i2})} \frac{\lambda^u e^{-\lambda}}{u!} \frac{\mu_1^{y_{i1}-u} e^{-\mu_1}}{(y_{i1} - u)!} \frac{\mu_2^{y_{i2}-u} e^{-\mu_2}}{(y_{i2} - u)!} \\ &= e^{-n(\lambda+\mu_1+\mu_2)} \prod_{i=1}^{n} \sum_{u=0}^{\min(y_{i1}, y_{i2})} \frac{\lambda^u}{u!} \frac{\mu_1^{y_{i1}-u}}{(y_{i1} - u)!} \frac{\mu_2^{y_{i2}-u}}{(y_{i2} - u)!}. \end{aligned} \tag{7}$$

The right-hand side of (7) (of rather the log of it) is readily maximized over the parameters with standard numerical iteration methods. In fact, because of Proposition 1 below, we only need to maximize over λ since it turns out that $\hat{\lambda} + \hat{\mu}_k = \bar{y}_k$ for $k = 1, 2$ where $\hat{\lambda}$ and $\hat{\mu}_k$ are the MLEs of λ and the μ_k, respectively.

For any numerical maximization in this paper, we use the Matlab routine `fmincon`.

Next, consider a model with one common factor and an arbitrary number of variables, m say, i.e.

$$\begin{cases} Y_1 = U + X_1, \\ Y_2 = U + X_2, \\ \quad \vdots \\ Y_m = U + X_m. \end{cases} \tag{8}$$

Let $f(u; \lambda)$ and $g(x; \mu_j)$ be the probability mass functions of U and the X_j respectively, $j = 1, 2, \ldots, m$. Then, with m-dimensional observations $(y_{i1}, \ldots, y_{in})$ for $i = 1, 2, \ldots, n$, we get the likelihood

$$L(\lambda, \mu_1, \ldots, \mu_m) = \prod_{i=1}^{n} \sum_{u=0}^{\min(y_{i1}, \ldots, y_{im})} f(u; \lambda) g(y_{i1} - u; \mu_1) \cdots g(y_{im} - u; \mu_m), \tag{9}$$

and, imposing the Poisson assumption,

$$L(\lambda, \mu_1, \ldots, \mu_m) = e^{-n(\lambda+\mu_1+\cdots+\mu_m)} \prod_{i=1}^{n} \sum_{u=0}^{\min(y_{i1},\ldots,y_{im})} \frac{\lambda^u}{u!} \frac{\mu_1^{y_{i1}-u}}{(y_{i1}-u)!} \cdots \frac{\mu_m^{y_{im}-u}}{(y_{im}-u)!}. \tag{10}$$

Again, to perform numerical maximization of (10) over the parameters, we only need to maximize w.r.t. λ. This is a simple consequence of the following proposition. [This fact was also pointed out by Karlis (2003).]

Proposition 1 *The parameters that maximize* (10)*,* $\hat{\lambda}, \hat{\mu}_1, \ldots, \hat{\mu}_m$*, satisfy the equalitites*

$$\bar{y}_k = \hat{\mu}_k + \hat{\lambda}, \quad k = 1, 2, \ldots, m, \tag{11}$$

where $\bar{y}_k = n^{-1} \sum_{i=1}^{n} y_{ik}$ *for all* k*.*

Proof See the "Appendix". □

2.2 Truncated distributions

Considering the situation with ordinal data, the Poisson assumption does not seem to fit perfectly well because of the finite number of classes. However, it can still be considered to provide an approximation. Alternatively, the truncated Poisson distribution could be tried. This means that we condition the Poisson variable to at most attain a maximum value, A say. The probability mass function of a $\text{Po}(\lambda)$ variable truncated in such a way is

$$f(y; \lambda) = \frac{\lambda^y / y!}{\sum_{j=0}^{A} \lambda^j / j!}.$$

The formulae in the previous section may be readily adjusted to cover this case. However, there does not seem to be any counterpart to Proposition 1. Thus, numerical maximization of the likelihood must be performed simultaneously over all parameters, not only over λ.

2.3 A mixed model

Comparing our setup to traditional factor analysis models, a potential obstacle is the restriction that more than one factor cannot load on the same Y variable. In the literature, an ANOVA like extension to the outlined model here that permits this is proposed, see e.g. Karlis (2003) and Loukas and Kemp (1983).

For the purposes of the present paper, the ANOVA like model seems to be quite complicated. We suggest another type of model, that extends the model of the previous sections in an easy way and leads to a relatively simple likelihood function. For example, consider the (3, 2) model:

$$\begin{cases} Y_1 = U_1 + X_1, \\ Y_2 = U_1 + X_2, \\ Y_3 = U_1 + X_3, \\ Y_4 = U_2 + X_4, \\ Y_5 = U_2 + X_5. \end{cases} \tag{12}$$

We can think about this as two groups, the first group (Y_1, Y_2, Y_3) sharing the common factor U_1 and the second group (Y_4, Y_5) sharing U_2. But maybe Y_1 should rather belong to the second group? This would give us the alternative model

$$\begin{cases} Y_1 = U_2 + X_1, \\ Y_2 = U_1 + X_2, \\ Y_3 = U_1 + X_3, \\ Y_4 = U_2 + X_4, \\ Y_5 = U_2 + X_5. \end{cases} \tag{13}$$

Now, a mixed model that allows for both of these possibilities is a model that is described by (12) with probability π and by (13) with probability $1-\pi$. Such a model may be interpreted as having both factors U_1 and U_2 loading on Y_1. Here, in a sense, π describes the extent to which the first factor, U_1, is relatively more important than U_2 as a loading on Y_1. Of course, $\pi = 1$ gives us the model (12) as special case, and $\pi = 0$ gives us (13).

As all the other parameters, π may be estimated by maximum likelihood. With notation as above, the likelihood for the mixed model described here is

$$L(\pi, \lambda_1, \lambda_2, \mu_1, \ldots, \mu_5) = \prod_{i=1}^{n} \{\pi s_{i1} s_{i2} + (1-\pi) s_{i3} s_{i4}\}, \tag{14}$$

where

$$\begin{aligned} s_{i1} &= \sum_{u_1=0}^{\min(y_{i1}, y_{i2}, y_{i3})} f(u_1; \lambda_1) g(y_{i1} - u_1; \mu_1) g(y_{i2} - u_1; \mu_2) g(y_{i3} - u_1; \mu_3), \\ s_{i2} &= \sum_{u_2=0}^{\min(y_{i4}, y_{i5})} f(u_2; \lambda_2) g(y_{i4} - u_2; \mu_4) g(y_{i5} - u_2; \mu_5), \\ s_{i3} &= \sum_{u_1=0}^{\min(y_{i2}, y_{i3})} f(u_1; \lambda_1) g(y_{i2} - u_1; \mu_2) g(y_{i3} - u_1; \mu_3), \\ s_{i4} &= \sum_{u_2=0}^{\min(y_{i1}, y_{i4}, y_{i5})} f(u_2; \lambda_2) g(y_{i1} - u_2; \mu_1) g(y_{i4} - u_2; \mu_4) g(y_{i5} - u_2; \mu_5). \end{aligned}$$

Observe that this was just one example of a mixed model. More complicated structures are allowed for by mixing more than two models, or by mixing several mixed models.

2.4 Pros and cons

Comparing to the traditional factor analysis setup with an underlying multivariate normal distribution, there are several immediate advantages with our approach. First, our model is more explicit and does not take the route over some underlying continuous distribution. This is beneficial, since to go from a continuous distribution to a discrete one, there is a need to estimate threshold parameters. With our approach, such computer intensive tasks are avoided.

Second, we do not run into factor rotation issues. Once we have found the most suitable model (more on that in Sect. 3), it may be readily interpreted by looking at its structure and parameter estimates.

It might seem that our model implies that the variables linked to the same factor, U_1 say, have the same loadings. But this is not so, since the corresponding independent X variables are allowed to have different parameters. For example, say that $Y_j = U_1 + X_j$ for $j = 1, 2, 3$, where U_1, X_1, X_2, X_3 are independent Poisson with parameters $\lambda_1, \xi_1, \xi_2, \xi_3$. This means that the covariances between any pair of Y_js are all equal to λ_1. However, the correlation between e.g. Y_1 and Y_2 is $\lambda_1/\sqrt{(\lambda_1+\mu_1)(\lambda_1+\mu_2)}$, and this is distinct from the correlation between e.g. Y_1 and Y_3 as long as $\mu_2 \neq \mu_3$.

One drawback with our approach is that negative correlations are not allowed for. To some extent this can be alleviated by reordering the categories within one or more of the variables. If it is the case that one variable has a negative correlation with all the other variables, these correlations can be turned to positive by reversing the order of categories for the variable. That is, if the variable in question, Y_j say, may take the values $0, 1, \ldots, q$, then replace it by $q - Y_j$. This method also works for a group of variables that correlate negatively to all the others, as long as all correlations within this group of variables are all positive.

This being said, there are of course many situations where we can not get rid of all negative correlations by order reversion, see for example Table 6. The hope is then that the negative correlations are so close to zero that they do not really matter in practice.

Another drawback is that we will have to search for the best model within a very large set of possible models. This issue will be discussed at some length in Sect. 3.

3 Model selection

3.1 A proposed method

When choosing between different models, one may for example use information criteria such as AIC or BIC, see e.g. Akaike (1974) and Schwarz (1978), respectively. When possible, sequential likelihood ratio tests may be employed as well.

In the following, we have chosen to stick to AIC. In presence of data sets of moderately large sizes, this seems to be the most common choice for model selection in the literature. We will use the definition

$$\text{AIC} = -2 \log L_{max} + 2p, \tag{15}$$

where L_{max} is the maximum likelihood value and p is the number of parameters. The selected model is the one with lowest AIC.

The main obstacle with our method is that there are so many potential models (combinations of factors). For large dimensions (numbers of Y variables) N, it is completely unrealistic to try them all, even for the fastest computer.

In the rest of this subsection, we will only discuss how to find models where only one factor loads to each variable. How to relax this constraint by allowing for mixed models will be the topic of the next subsection.

Below, we will consider dimensions 5 and 7. In dimension 5, there are 52 possible models: the (1, 1, 1, 1, 1) model, $\binom{5}{2} = 10$ models of type (2, 1, 1, 1), $\binom{5}{3} = 10$ models of type (3, 1, 1), $\binom{5}{4} = 5$ models of type (4, 1), $5 * \binom{4}{2}/2 = 15$ models of type (2, 2, 1), $\binom{5}{3} = 10$ models of type (3,2), and the model of type (5), where the same factor loads on all the five variables.

In dimension 7, it can be shown that the number of possible models is 877. In fact, the number of possible models in dimension N is described by the Bell number, cf Flajolet and Sedgewick (2009), pp. 560–562. The Bell number gives the number of partitions of the set of integers from 1 to N. Calling this number B_N, it holds that $\log B_N$ behaves like $N \log N$ as N tends to infinity. Hence, B_N increases with more than an exponential rate with N. Thus, for large dimensions, it is not practically feasible to consider all possible models.

The way out of this dilemma is to try some sort of model selection algorithm. In this paper, we suggest to start with the independence model $(1, 1, \ldots, 1)$, and compare it with all possible $(2, 1, \ldots, 1)$ models. (A total of $\binom{N}{2}$ models.) If the independence model is the best (has the lowest AIC), the algorithm stops. If not, we go on by estimating all $(3, 1, \ldots, 1)$ models where the pair of variables that had the same factor in the first step is joined by one of the other variables ($N - 2$ models) as well as all $(2, 2, 1, \ldots, 1)$ models where we add a new pair of variables that consists of any two that were not in the first pair ($\binom{N-2}{2}$ models). If none of the $(3, 1, \ldots, 1)$ of $(2, 2, 1, \ldots, 1)$ models tried is better than the previously chosen $(2, 1, \ldots, 1)$ model, we stop and choose the previous model. If not, we go on to test new models, and so it goes on.

The principle in all steps is to take the favorite model of the previous step and then merge any two groups (considering the ones to be groups of their own). For example, if the previously selected model was of type $(2, 2, 1, \ldots, 1)$, the new models tried are of types $(3, 2, 1, \ldots, 1)$, $(2, 2, 2, 1, \ldots, 1)$ and $(4, 1, \ldots, 1)$.

For dimension $N = 5$, the algorithm is illustrated in Fig. 1. Note that in this figure, we have simplified the last step of the algorithm (if it reaches that far) to test model (5) together with (3, 2) and (4, 1).

Note that for our algorithm, in dimension five the maximum number of model estimations is 21, out of the 52 possible models. This may not be considered to be a really substantial reduction. However, in dimension seven, the maximum number of model estimations turns out to be 57 out of 877 possible models.

For an arbitrary dimension N, the number of steps in the selection algorithm is of the order N^3, see Proposition 2. This is in contrast to the exponential rate of increase of the number of possible models as N increases.

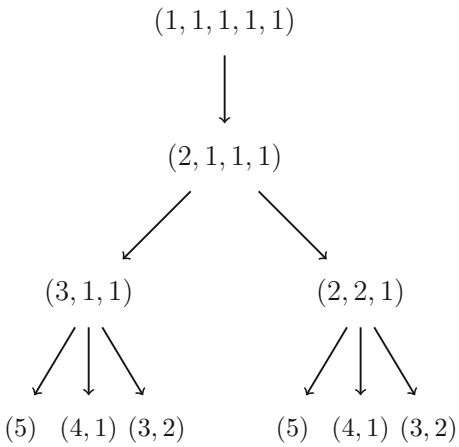

Fig. 1 Model selection algorithm, dimension 5

Proposition 2 *In the forward selection algorithm, for dimension N the maximum number of tested models is*

$$1+\sum_{k=2}^{N}\binom{k}{2}=1+\binom{N+1}{3}=\frac{1}{6}(N+2)(N^2-2N+3). \tag{16}$$

Proof At first, we estimate the model $(1, 1, \ldots, 1)$. The second step is to estimate all possible $(2, 1, \ldots, 1)$ models, the number of which is $\binom{N}{2}$. If one of them is the best so far, we go on estimating all models of the forms $(3, 1, \ldots, 1)$ or $(2, 2, 1, \ldots, 1)$ that may be constructed by merging any two of the $N-1$ subsets in the $(2, 1, \ldots, 1)$ model. This number of subsets equals $\binom{N-1}{2}$. If each step in the algorithm results in a better model than previously, the procedure goes on until a model with 3 subsets is tested against all of its submodels, the number of which is $4=3+1=\binom{3}{2}+\binom{2}{2}$.

This shows that the maximum number of estimated models is as in the left hand side of (16). The equalities of (16) follow from simple algebra. □

3.2 Finding a mixed model

Clearly, the amount of possible mixed models increases enormously with the dimension. Thus, it does not seem feasible to involve these in an automatic search procedure. Instead, the proposal is to at first select the best model where only one factor loads on each variable. Then, the mixed models that are most "nearby" to this model can be tried. The one with smallest AIC (if any one has smaller AIC than the non mixed one) is selected. See further the empirical example in Sect. 5.

3.3 Asymptotic properties

In this section, we heuristically derive the asymptotic probabilities to select the correct model for the outlined selection algorithm. This discussion will not take mixed models into account.

Take dimension 5 as an example. At first, consider testing the model (1, 1, 1, 1, 1) versus a specific (2, 1, 1, 1) alternative. Here, the null model has five parameters, while the alternative model has six, the "extra" parameter, λ say, being that of the common factor. Seeking to minimize AIC [cf (15)], we reject the null model and choose the alternative one if the difference of their -2 log likelihood values is more than 2.

To calculate the asymptotic probability (*asp* in the following) for this to happen, we may employ classical results on the maximum likelihood ratio (MLR) test. Here, observe that we are testing H_0: $\lambda = 0$ versus H_1: $\lambda > 0$, so we are testing the null that the parameter lies on the boundary of the parameter space. Under H_0, the asymptotic distribution of the MLR test is given by e.g. Self and Liang (1987) as $V = Z^2 I(Z > 0)$ where Z is standard normal and I is the indicator function. In other words, asymptotically and under H_0, in our case the *asp not to reject* is given by

$$\gamma = P(V \leq 2) = P(Z < \sqrt{2}) \approx 0.92135. \tag{17}$$

To get further, we need to employ the following unproved postulates (cf Voung 1989, for a theory of this type for the standard case when the null value of the parameter is not at the boundary of the parameter space):

1. Tests performed at the same step in the algorithm are asymptotically independent.
2. Tests of a null model with fewer parameters than the alternative model have asymptotic power 1.
3. Tests of a null model with as many as or fewer parameters than the alternative model have asymptotic probability 1 not to reject.

Now, consider testing the (1, 1, 1, 1, 1) versus *any* (2, 1, 1, 1) model. There are $\binom{5}{2} = 10$ alternative models. By postulate 1, we get that the *asp* not to reject is $\gamma^{10} \approx 0.44$. Hence, we have heuristically derived the *asp* to correctly find the (1, 1, 1, 1, 1) model to be approximately 0.44.

Next, consider the case when the (2, 1, 1, 1) model is true. Asymptotically, by postulate 2 the probability to get from the (1, 1, 1, 1, 1) model to (2, 1, 1, 1) in the first step tends to one. Moreover, this must be the true one among the 10 possible similar models, because from postulate 3, the *asp* to accept the true (2, 1, 1, 1) model over a false (2, 1, 1, 1) model is one.

Coming this far, we will in fact select the true (2, 1, 1, 1) model if we do not reject it when testing versus the three (2, 2, 1) models that are accomplished by merging the single items as well as testing versus the three (3, 1, 1) models that we get by putting any single item together with the pair. Testing (2, 1, 1, 1) versus (2, 2, 1) gives one extra factor parameter, so by postulate 1, the *asp* not to reject in any of these three cases is $\gamma^3 \approx 0.78$. Testing versus a (3, 1, 1) model, however, gives no extra parameter, and so, by postulate 3, the *asp* to keep the (2, 1, 1, 1) model is one in this case. To sum up, the *asp* to correctly select a (2, 1, 1, 1) model is approximately 0.78.

We may go on in the same fashion to calculate the *asp* of correctly selecting any possible model. In particular, one may note that the *asp* of correctly selecting any model containing at most one single item is one.

All of this was also done for dimension 7. The *asp* values of correct selection are given together with the corresponding finite sample simulated probabilities in the tables of the next section (the $n = \infty$ columns).

Table 1 Estimated probability to find the correct model, dimension 5, parameters 0.5 for the factors and 1 for the observed variables, 5000 replicates

Model	$n = 25$		$n = 50$		$n = 100$		$n = \infty$
	Test all	Selection	Test all	Selection	Test all	Selection	
(5)	0.97	0.97	1.00	1.00	1.00	1.00	1.00
(4, 1)	0.93	0.91	0.99	0.99	1.00	1.00	1.00
(3, 2)	0.79	0.78	0.98	0.98	1.00	1.00	1.00
(3, 1, 1)	0.76	0.76	0.90	0.90	0.92	0.92	0.92
(2, 2, 1)	0.62	0.62	0.90	0.90	0.99	0.99	1.00
(2, 1, 1, 1)	0.49	0.51	0.66	0.68	0.74	0.76	0.78
(1, 1, 1, 1, 1)	0.30	0.39	0.32	0.42	0.33	0.42	0.44

4 Simulations

The main question to be asked in this section is: What is the probability that the selection algorithm finds the correct model? We check this with simulations. As a start, we will consider dimension 5, where it is feasible to compare the algorithm to the method of estimating all possible models (there are "only" 52 of them here). We then go on to dimension 7, which is also the dimension of the empirical example. In this dimension, we only study the selection algorithm. For this dimension, we also check what happens when the distribution is truncated.

All simulations are performed in Matlab2016a. We maximize the likelihood by minimizing the minus log likelihood using the function `fmincon`. As starting values for the function, we take 0 for the parameters of the factors and the means of the corresponding Y_i observations for the X_i.

Inspired by the empirical example in the next section, we take the common factors U_i to be Po(0.5). The X_i that sum with a common factor to give the observed Y_i are also Po(0.5). For the X_i that are not (so $X_i = Y_i$ in these cases), we take Po(1). This means that all Y_i are Po(1).

We simulate models of all possible types and then check the proportion of times that AIC is smallest for the model simulated. We also check the proportion of times that the selection algorithm finds the correct model. This always means not only that it is of the correct type, but also that it places the variables correctly into the different groups that have the same common factor.

The results are given in Tables 1 and 2. Comparing to testing all models, it is seen that the selection method works remarkably well. As expected, we also find that the selection probabilities increase with n, and that they approach the *asp* derived in the previous section. Moreover, as is also natural, for models with relatively many factors they are smaller when the parameter is relatively smaller for the factors compared to the independent components. Cf Table 2.

For dimension 7, we only consider the selection algorithm and one parameter combination, see Table 3. The conclusions are similar to dimension 5.

In Tables 4 and 5, we consider truncated distributions. The truncation at 3 of Table 4 is the same as in the empirical example, whereas the truncation at 2 in Table 5 illustrates

Table 2 Estimated probability to find the correct model, dimension 5, parameters 0.5 for the factors and 2 for the observed variables, 5000 replicates

Model	$n = 25$		$n = 50$		$n = 100$		$n = \infty$
	Test all	Selection	Test all	Selection	Test all	Selection	
(5)	0.44	0.45	0.77	0.77	0.96	0.96	1.00
(4, 1)	0.39	0.36	0.69	0.67	0.92	0.91	1.00
(3, 2)	0.19	0.18	0.42	0.40	0.76	0.75	1.00
(3, 1, 1)	0.31	0.28	0.54	0.52	0.76	0.76	0.92
(2, 2, 1)	0.12	0.12	0.27	0.27	0.56	0.56	1.00
(2, 1, 1, 1)	0.19	0.21	0.31	0.33	0.48	0.50	0.78
(1, 1, 1, 1, 1)	0.31	0.38	0.33	0.41	0.34	0.42	0.44

Table 3 Estimated probability to find the correct model, the non truncated case, dimension 7, parameters 0.5 for the factors and 1 for the observed variables, 5000 replicates

Model	$n = 25$	$n = 50$	$n = 100$	$n = \infty$
(7)	0.97	1.00	1.00	1.00
(6, 1)	0.94	1.00	1.00	1.00
(5, 2)	0.81	0.97	0.99	1.00
(5, 1, 1)	0.84	0.92	0.91	0.92
(4, 3)	0.82	0.97	0.99	1.00
(4, 2, 1)	0.70	0.93	0.99	1.00
(4, 1, 1, 1)	0.65	0.75	0.77	0.78
(3, 3, 1)	0.73	0.95	0.99	1.00
(3, 2, 2)	0.64	0.96	1.00	1.00
(3, 2, 1, 1)	0.54	0.82	0.91	0.92
(3, 1, 1, 1, 1)	0.43	0.56	0.59	0.61
(2, 2, 2, 1)	0.45	0.83	0.98	1.00
(2, 2, 1, 1, 1)	0.33	0.60	0.74	0.78
(2, 1, 1, 1, 1, 1)	0.22	0.35	0.41	0.44
(1, 1, 1, 1, 1, 1, 1)	0.13	0.16	0.16	0.18

what happens when the truncation probability is relatively high. To get the same expected value of the Y variables as in the untruncated case, we have chosen parameter values 1.08 and 1.414, respectively, instead of 1 (and half of these values instead of 0.5). As before, the probabilities to find the correct model increase with n. Comparing to the case without truncation, we see that the probabilities are smaller, and even more so in case of the more severe truncation in Table 5.

5 Empirical example

In this section, we analyze a seven-dimensional data set taken from Jöreskog et al. (2016). The data come from the Eurobarometer Survey of 1992, where citizens

Table 4 Estimated probability to find the correct model, dimension 7, parameters 0.54 for the factors and 1.08 for the observed variables, truncated at 3, 1000 replicates

Model	$n = 25$	$n = 50$	$n = 100$	$n = \infty$
(7)	0.93	1.00	1.00	1.00
(6, 1)	0.87	0.99	1.00	1.00
(5, 2)	0.69	0.94	0.99	1.00
(5, 1, 1)	0.74	0.90	0.92	0.92
(4, 3)	0.69	0.93	0.99	1.00
(4, 2, 1)	0.55	0.88	0.97	1.00
(4, 1, 1, 1)	0.55	0.73	0.76	0.78
(3, 3, 1)	0.60	0.91	0.98	1.00
(3, 2, 2)	0.48	0.89	1.00	1.00
(3, 2, 1, 1)	0.41	0.75	0.90	0.92
(3, 1, 1, 1, 1)	0.34	0.54	0.56	0.61
(2, 2, 2, 1)	0.31	0.73	0.96	1.00
(2, 2, 1, 1, 1)	0.24	0.51	0.70	0.78
(2, 1, 1, 1, 1, 1)	0.18	0.32	0.37	0.44
(1, 1, 1, 1, 1, 1, 1)	0.14	0.14	0.16	0.18

Table 5 Estimated probability to find the correct model, dimension 7, parameters 0.707 for the factors and 1.414 for the observed variables, truncated at 2, 1000 replicates

Model	$n = 25$	$n = 50$	$n = 100$	$n = \infty$
(7)	0.50	0.87	0.98	1.00
(6, 1)	0.37	0.79	0.97	1.00
(5, 2)	0.27	0.66	0.94	1.00
(5, 1, 1)	0.35	0.67	0.88	0.92
(4, 3)	0.29	0.66	0.94	1.00
(4, 2, 1)	0.20	0.54	0.86	1.00
(4, 1, 1, 1)	0.23	0.53	0.72	0.78
(3, 3, 1)	0.23	0.57	0.90	1.00
(3, 2, 2)	0.13	0.47	0.91	1.00
(3, 2, 1, 1)	0.17	0.44	0.74	0.92
(3, 1, 1, 1, 1)	0.18	0.37	0.50	0.61
(2, 2, 2, 1)	0.09	0.33	0.75	1.00
(2, 2, 1, 1, 1)	0.09	0.27	0.52	0.78
(2, 1, 1, 1, 1, 1)	0.11	0.20	0.31	0.44
(1, 1, 1, 1, 1, 1, 1)	0.11	0.16	0.17	0.18

of Great Britain were asked about their attitudes towards Science and Technology. The answers are collected on an ordinal scale with values $1, 2, 3, 4$. The sample size is $n = 392$. The variables are called `Comfort`, `Environment`, `Work`, `Future`, `Technology`, `Industry` and `Benefit`, but in the following, we will just refer to them as $\tilde{y}_1, \ldots, \tilde{y}_7$. Because the means of all variables are closer to 4 than to 1, we have chosen to transform them according to $y_j = 4 - \tilde{y}_j$, to get a better fit to a truncated Poisson distribution. The truncation point is then at 3.

Table 6 Descriptive statistics for the empirical data set (four minus the original data)

	Mean	Variance	Correlations						
			y_1	y_2	y_3	y_4	y_5	y_6	y_7
y_1	0.88	0.35	1.00						
y_2	1.05	0.85	0.08	1.00					
y_3	1.28	0.65	0.15	− 0.07	1.00				
y_4	1.01	0.57	0.28	− 0.03	0.40	1.00			
y_5	1.00	0.74	0.07	0.39	− 0.09	− 0.03	1.00		
y_6	0.76	0.58	0.13	0.33	− 0.02	0.06	0.35	1.00	
y_7	1.16	0.64	0.33	− 0.03	0.17	0.31	− 0.01	0.09	1.00

In Table 6, we give descriptive statistics: mean, variance and the correlation matrix. Observe that the means are larger than the variances. This is in accord with the truncated Poisson distribution. For example, a Poisson variable with parameter 1.08 has expectation 1.00 and variance 0.84 and a parameter value of 0.836 corresponds to expectation 0.80 and variance 0.56. In view of this, we find that most of the variables have a little smaller variances than expected from the truncated Poisson, but not much smaller.

Looking at correlations, it can be seen that some are negative. This is impossible under the dependent Poisson model. However, all negative correlations are small in absolute value. Hence, in a factor analysis context they should be relatively unimportant anyway.

Next, we try our model selection method, applied for Poisson variables truncated at 3, to the data $(y_1, \ldots, y_7)$. The model found has the same factor structure as the one given in the explanatory analysis chapter of Jöreskog et al. (2016) when estimated with maximum likelihood. It is a (4, 3) model where the variables are grouped as (y_1, y_3, y_4, y_7) and (y_2, y_5, y_6). The estimates are found in the first column of Table 7. (The standard errors are obtained from the empirical Fisher information, which in turn is calculated as numerical second derivatives of the observed minus log likelihood w.r.t. the parameters. The standard errors are the Fisher informations to the power of $-1/2$.)

We find that the estimates reflect the means of Table 6 fairly well. (Recall that the expected value of a truncated Poisson is greater than the parameter.) Moreover, note that all the negative correlations of Table 6 correspond to correlations between variables that belong to different groups (hence are independent) in the selected model. Hence, under the model, these correlations are zero.

In Jöreskog et al. (2016), a second model is fitted (using polychoric correlations and weighted least squares). In this model, y_1 is allowed to belong to both variable groups. To see if we can obtain something similar, we fit a mixed model to the data, where y_1 belongs to the first group with probability π. We give the corresponding estimates in the second column of Table 7.

We find that for the mixed model, the log likelihood is more than 2 units higher than the log likelihood for the (4, 3) model. Hence, AIC is lower for the mixed model. The interpretation of $\hat{\pi} = 0.74$ is that y_1 is more strongly connected to the (y_3, y_4, y_7)

Table 7 Estimated parameters and log likelihood, empirical data set (standard errors in parenthesis)

	Model (4, 3)	Mixed model
$\hat{\pi}$	–	0.74 (0.06)
$\hat{\lambda}_1$	0.67 (0.06)	0.75 (0.06)
$\hat{\lambda}_2$	0.48 (0.04)	0.48 (0.04)
$\hat{\mu}_1$	0.40 (0.04)	0.37 (0.04)
$\hat{\mu}_3$	0.89 (0.06)	0.81 (0.06)
$\hat{\mu}_4$	0.55 (0.04)	0.48 (0.04)
$\hat{\mu}_7$	0.74 (0.05)	0.66 (0.05)
$\hat{\mu}_2$	0.70 (0.05)	0.70 (0.05)
$\hat{\mu}_5$	0.64 (0.05)	0.64 (0.05)
$\hat{\mu}_6$	0.36 (0.03)	0.36 (0.03)
log L	− 3064.6	− 3052.3

Table 8 Estimated log likelihood for different mixed models

Mixing variable	Group 1	Group 2	log L
1	3 4 7	2 5 6	− 3052.3
3	1 4 7	2 5 6	− 3060.9
4	1 3 7	2 5 6	− 3061.4
7	1 3 4	2 5 6	− 3053.9
2	1 3 4 7	5 6	− 3058.9
5	1 3 4 7	2 6	− 3064.1
6	1 3 4 7	2 5	− 3042.8

group than to (y_2, y_5, y_6). The latter finding is in accord with the estimates of Jöreskog et al. (2016), where y_1 loads two to three times stronger on the first group than on the second.

Moreover, observe that $\hat{\lambda}_1 + \hat{\mu}_j$ for $j = 3, 4, 7$ is about the same for both models and, in fact, they are equal up to two decimal points for $\hat{\lambda}_2 + \hat{\mu}_j$ for $j = 2, 5, 6$.

We have also tried the other most "nearby" mixed models, to see if any one of them fits better, see Table 8. (In this table, "group 1" refers to the variables that have the first factor in common, and correspondingly for "group 2".) Here, we can in fact observe a better fit (with lower log likelihood, hence also lower AIC because of equally many estimated parameters) for the model where variable 6 loads on both groups, with variables 1, 3, 4 and 7 in the first group. The parameter estimates for this model are given in Table 9.

Finally, to enable a further comparison with the Jöreskog model, we give estimated factor loadings for the latter in Table 10. (The estimated Φ matrix is close to the identity matrix in both cases, not shown here.) In this table, model 1 corresponds to the one obtained by maximum likelihood, with separate factors for the two groups of variables. The second one is a model where y_1 loads on both groups. This model is estimated by a least squares method using polycoric correlations. The estimates for the two models are not directly comparable to each other, since the variables are

Table 9 Estimated parameters and log likelihood for the best mixing model, empirical data set (standard errors in parenthesis)

$\hat{\pi}$	0.33 (0.05)
$\hat{\lambda}_1$	0.68 (0.06)
$\hat{\lambda}_2$	0.58 (0.05)
$\hat{\mu}_6$	0.28 (0.03)
$\hat{\mu}_1$	0.40 (0.04)
$\hat{\mu}_3$	0.89 (0.06)
$\hat{\mu}_4$	0.54 (0.04)
$\hat{\mu}_7$	0.73 (0.05)
$\hat{\mu}_2$	0.60 (0.05)
$\hat{\mu}_5$	0.54 (0.04)
$\log L$	-3042.8

Table 10 Estimated factor loadings, Jöreskog method (standard errors in parenthesis)

	Model 1	Model 2
$\hat{\lambda}_1$	1.046 (0.190)	0.528 (0.074)
$\hat{\lambda}_3$	1.221 (0.182)	0.536 (0.064)
$\hat{\lambda}_4$	2.289 (0.484)	0.768 (0.061)
$\hat{\lambda}_7$	1.095 (0.183)	0.535 (0.064)
$\hat{\lambda}_1$	–	0.197 (0.076)
$\hat{\lambda}_2$	1.622 (0.250)	0.661 (0.069)
$\hat{\lambda}_5$	1.744 (0.279)	0.692 (0.061)
$\hat{\lambda}_6$	1.530 (0.242)	0.631 (0.067)

Loadings on the first/second factor are above/below the line

standardized in different ways. Also, a direct comparison to our estimates in Tables 7 and 9 is not feasible, since the Poisson parameters reflect expectations and variances at the same time.

What could be compared is the relation between parameter estimates within a certain model versus the corresponding relations within other models. Here, we can see that the large factor loading for variable 4 stands out in model 1 of Jöreskog. Comparing to the empirical variances of Table 6, this seems a bit peculiar. But apart from this, no dramatic differences between the models are seen.

6 Concluding remarks

In this paper, we have proposed a method for performing factor analysis on discrete data. In principle, the method should work for any choice of discrete distribution. As a first try, we have chosen the Poisson distribution. Among the very many candidate models, we look for the one with smallest AIC in a forward search algorithm. We have found, both by heuristic calculations and simulations, that this method works well in the sense that it has a high probability to find the correct model (if there is one) for moderately large to large sample sizes.

Since most real life examples of discrete data factor analysis concern ordinal data, we modify our method to deal with this by looking at truncated discrete distributions. So far, ordinal data factor analysis has been performed as in Jöreskog et al. (2016), who assume an underlying normal distribution. Numerically, the Jöreskog methodology can be rather complicated, at least when employing maximum likelihood, because in addition to the parameters of main interest, the threshold parameters need to be estimated. Also, as is always the case with traditional factor analysis, factor rotations of more or less arbitrary nature are imposed.

The method proposed in the present paper is more straightforward. The model is fully specified once the factor structure has been found. The difficulty lies in finding this structure among the very many possible ones. To this end, we have outlined a forward selection method which seems to work well for small and moderately large dimensions. However, model selection for very large dimensions seems to be a challenge that calls for further development of the selection method. This issue is left for future research.

Another aspect that needs further investigation is the choice of distribution. One could of course replace the Poisson by something else like the geometric, the binomial or the negative binomial distribution. Different mixture distributions (different distributions on factors and independent components) are also possible, for example the mixture of the Binomial and Poisson distributions as discussed in Karlis (2003) among others.

Also, other information criteria than AIC could be used, e.g. BIC. Moreover, in many applications it would be helpful to avoid the requirement that correlations can not be negative, see e.g. Famoye (2015) and Berkhout and Plug (2004). On the theoretical side, a full proof that the heuristic calculations on asymptotic probabilities to find the correct model are valid is called for.

Acknowledgements Open access funding provided by Uppsala University. I would like to thank the referees, the associate editor and the editor for their careful reading and valuable suggestions that greatly helped to improve the paper.

Appendix: Proof of Proposition 1

Suppose that it is not the case that all $y_{i1} = \min(y_{i1}, \ldots, y_{im})$. Rewrite (10) as

$$L(\lambda, \mu_1, \ldots, \mu_m) = e^{-n(\lambda+\mu_1+\cdots+\mu_m)} \sum_{u_1=0}^{z_1} \cdots \sum_{u_n=0}^{z_n} \frac{\lambda^{\sum_{j=1}^n u_j}}{\prod_{i=1}^n u_j!} g_1(\mu_1) \cdots g_m(\mu_m), \tag{18}$$

where $z_i = \min(y_{i1}, \ldots, y_{im})$ for all i and

$$g_k(\mu_k) = \frac{\mu_k^{n\bar{y}_k - \sum_{i=1}^n u_i}}{\prod_{i=1}^n (y_{ki} - u_i)!}, \quad k = 1, 2, \ldots, m.$$

Without loss of generality, pick $k = 1$. Now, suppressing the arguments of L, differentiation w.r.t μ_1 in (18) yields

$$\begin{aligned}\frac{\partial L}{\partial \mu_1} &= -nL \\ &\quad + e^{-n(\lambda+\mu_1+\cdots+\mu_m)} \sum_{u_1=0}^{z_1} \cdots \sum_{u_n=0}^{z_n} \frac{\lambda^{\sum_{j=1}^n u_j}}{\prod_{i=1}^n u_j!} \left\{ \frac{\partial}{\partial \mu_1} g_1(\mu_1) \right\} \\ &\quad \cdot g_2(\mu_2) \cdots g_m(\mu_m), \end{aligned} \tag{19}$$

where

$$\frac{\partial}{\partial \mu_1} g_1(\mu_1) = \frac{n\bar{y}_1 - \sum_{i=1}^n u_i}{\mu_1} g_1(\mu_1).$$

Hence, inserting into (19) and in view of (18),

$$\frac{\partial L}{\partial \mu_1} = -nL + \frac{n\bar{y}_1}{\mu_1} L - \frac{1}{\mu_1} e^{-n(\lambda+\mu_1+\cdots+\mu_m)} \sum_{i=1}^n A_i, \tag{20}$$

where e.g.

$$A_n = \sum_{u_1=0}^{z_1} \cdots \sum_{u_{n-1}=0}^{z_{n-1}} \frac{\lambda^{\sum_{j=1}^{n-1} u_j}}{\prod_{i=1}^{n-1} u_j!} \sum_{u_n=0}^{z_n} u_n \frac{\lambda^{u_n}}{u_n!} g_1(\mu_1) \cdots g_m(\mu_m). \tag{21}$$

But since

$$u_n \lambda^{u_n} = \lambda \frac{d}{d\lambda} \lambda^{u_n},$$

it follows from (21) and analogous equations for the other A_i that

$$\sum_{i=1}^n A_i = \lambda \frac{d}{d\lambda} \sum_{u_1=0}^{z_1} \cdots \sum_{u_n=0}^{z_n} \frac{\lambda^{\sum_{j=1}^n u_j}}{\prod_{i=1}^n u_j!} g_1(\mu_1) \cdots g_m(\mu_m).$$

In view of (18), this is

$$\sum_{i=1}^n A_i = \lambda \frac{d}{d\lambda} \left\{ e^{n(\lambda+\mu_1+\cdots+\mu_m)} L \right\} = \lambda \left\{ n e^{n(\lambda+\mu_1+\cdots+\mu_m)} L + e^{n(\lambda+\mu_1+\cdots+\mu_m)} \frac{dL}{d\lambda} \right\},$$

and (20) yields

$$\frac{\partial L}{\partial \mu_1} = -nL + \frac{n\bar{y}_1}{\mu_1} L - \frac{\lambda}{\mu_1}\left(nL + \frac{dL}{d\lambda}\right).$$

But since $dL/d\lambda = 0$ at $\lambda = \hat{\lambda}$, we get for this λ that

$$\frac{\partial L}{\partial \mu_1} = nL\left(-1 + \frac{\bar{y}_1}{\mu_1} - \frac{\hat{\lambda}}{\mu_1}\right) = \frac{nL}{\mu_1}(-\mu_1 + \bar{y}_1 - \hat{\lambda}),$$

which is zero for $\mu_1 = \hat{\mu}_1 = \bar{y}_1 - \hat{\lambda}$, as was to be shown.

The proof for the case that all $y_{i1} = \min(y_{i1}, \ldots, y_{im})$ follows similarly.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723

Berkhout B, Plug E (2004) A bivariate Poisson count data model using conditional probabilities. Stat Neerl 3:349–364

Famoye F (2015) A multivariate generalized Poisson regression model. Commun Stat Theory Methods 44:497–511

Flajolet P, Sedgewick R (2009) Analytic combinatorics. Cambridge University Press, Cambridge

Inouye DI, Yang E, Allen GI, Ravikumar P (2017) A review of multivariate distributions for count data derived from the Poisson distribution. WIREs Comput Stat 9:e1398. https://doi.org/10.1002/wics.1398

Jöreskog KG (1967) Some contributions to maximum likelihood factor analysis. Psychometrica 32:443–482

Jöreskog KG, Moustaki I (2001) Factor analysis of ordinal variables: a comparison of three approaches. Multivar Behav Res 36:347–387

Jöreskog KG, Olsson UH, Wallentin FY (2016) Multivariate analysis with LISREL. Springer, Berlin

Karlis D (2003) An EM algorithm for multivariate Poisson distribution and related models. J Appl Stat 30:63–77

Katsikatsou M, Moustaki I, Yang-Wallentin F, Jöreskog KG (2012) Pairwise likelihood estimation for factor analysis models with ordinal data. Comput Stat Data Anal 56:4243–4258

Loukas S, Kemp CD (1983) On computer sampling from trivariate and multivariate discrete distributions. J Stat Comput Simul 17:113–123

Olsson U (1979) Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrica 44:443–460

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610

Voung QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57:307–333

Wedel M, Böckenholt U, Kamakura WA (2003) Factor models for multivariate count data. J Multivar Anal 87:356–369

Zhou M, Hannah LA, Dunson DB, Carin L (2012) Beta-negative binomial process and poisson factor analysis. In: Proceedings of the 15th international conference on artificial intelligence and statistics