Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Applications of discrete factor analysis

Rolf Larsson & Jesper Rydén

To cite this article: Rolf Larsson & Jesper Rydén (2023) Applications of discrete factor analysis, Communications in Statistics - Simulation and Computation, 52:10, 4592-4602, DOI: 10.1080/03610918.2021.1964528

To link to this article: https://doi.org/10.1080/03610918.2021.1964528

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
2023, VOL. 52, NO. 10, 4592–4602
https://doi.org/10.1080/03610918.2021.1964528

Applications of discrete factor analysis

Rolf Larsson[a] and Jesper Rydén[b]

[a]Department of Mathematics, Uppsala University, Uppsala, Sweden; [b]Department of Energy and Technology, Swedish University of Agricultural Sciences, Uppsala, Sweden

ABSTRACT

A recently proposed method for factor analysis of discrete data is extended to better handle overdispersion. Three empirical examples from veterinary sciences, musicology and agriculture are investigated, involving true count data as well as ordinal data. Comparisons are made with results from related statistical techniques, e.g., principal component analysis.

ARTICLE HISTORY
Received 11 May 2021
Accepted 31 July 2021

KEYWORDS
Poisson distribution; Negative binomial distribution; Model selection

1. Introduction

Factor analysis is a statistical technique that can be employed in several fields of application. The variability among observed correlated variables is described in terms of a lower number of unobserved variables called factors. Recently, a method for factor analysis of discrete data was presented (Larsson 2020), along with an empirical example: a seven-dimensional data set of ordinal data (survey study). In this paper, the main purpose is to illustrate the usefulness of that method by studying three different empirical examples. In the sequel of this introductory section, we first review the method from Larsson (2020) and also present an extension for analysis of data with overdispersion, introducing the negative binomial distribution.

The method is implemented in Matlab, and codes are available upon request.

1.1. Dependent Poisson models

In Larsson (2020), a method of performing factor analysis for discrete data using a dependent Poisson distribution is presented. A short description of the method is as follows. In the literature, there are many versions of dependent Poisson models, see Larsson (2020) and the references therein. The model studied by Larsson (2020) has been proposed by e.g., Karlis (2003). In is simplest form, it is given as a bivariate model

$$\begin{cases} Y_1 &= U + X_1, \\ Y_2 &= U + X_2, \end{cases} \tag{1}$$

where U, X_1 and X_2 are independent Poisson variables, with intensities λ, μ_1 and μ_2 (say), respectively. Then Y_1 and Y_2 are also Poisson, but with intensities $\lambda + \mu_1$ and $\lambda + \mu_2$, respectively. The variables U, X_1 and X_2 are considered latent, only Y_1 and Y_2 are observed.

CONTACT Rolf Larsson rolf.larsson@math.uu.se Department of Mathematics, Uppsala University, Uppsala, Sweden.

It is easily seen that

$$\operatorname{corr}(Y_1, Y_2) = \frac{\lambda}{\sqrt{(\lambda+\mu_1)(\lambda+\mu_2)}}. \tag{2}$$

In particular we see that (as is natural), the correlation is bigger when λ is large relative to μ_1, μ_2 than otherwise. Another important observation is that the model only allows for a positive correlation.

Suppose we have observation pairs $(y_{11}, y_{12}), \ldots, (y_{n1}, y_{n2})$. In Larsson (2020), it is laid out how the parameters may be estimated numerically by maximum likelihood. Indeed, for the model in (1), the likelihood takes the form

$$L(\lambda, \mu_1, \mu_2) = \prod_{i=1}^{n} \sum_{u=0}^{\min(y_{i1}, y_{i2})} f(u; \lambda) g(y_{i1} - u; \mu_1) g(y_{i2} - u; \mu_2), \tag{3}$$

where f and g are Poisson probability functions, e.g.,

$$f(u; \lambda) = \frac{\lambda^u}{u!} e^{-\lambda} \tag{4}$$

and we used that Y_1 and Y_2 are conditionally independent, given $U = u$.

In fact the likelihood only needs to be numerically maximized over λ, since it can be proved that $\hat{\lambda} + \hat{\mu}_k = \bar{y}_k$ for $k = 1, 2$, where $\hat{\lambda}$ and $\hat{\mu}_k$ are the MLEs and $\bar{y}_k = n^{-1} \sum_i y_{ik}$, cf. Larsson (2020).

We may view (1) as a simple two-variable factor model, with a common factor U. The idea of Larsson (2020) is to extend this model to a system with many variables, with different groups of variables. For each group, the variables are connected to each other through a specific common factor. For example, a model with five variables and two factors may be constructed as

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

where $U_1, U_2, X_1, X_2, X_3, X_4, X_5$ are independent.

Here, the variables Y_1, Y_2, Y_3 are connected (correlated) through U_1 and Y_4, Y_5 are connected through U_2. It is clear that any of Y_1, Y_2, Y_3 is uncorrelated to (independent of) any of Y_4 and Y_5. The parameters of the model (5) may be estimated by maximum likelihood in much the same way as for the parameters of (1). In fact, this estimation is quite simple because of the independence of the two sub systems (Y_1, Y_2, Y_3) and (Y_4, Y_5).

Now, it is quite obvious how to go on to form many other systems in the style of (5), see further Larsson (2020) for a general formulation of the model.

The next question is how to find the dependent Poisson model that best fits the observed data. Larsson (2020) suggests that the model giving the smallest value on the Akaike information criteria (AIC) should be selected. An obstacle is that, in systems with many variables, there are a lot of possible models to go through. To alleviate this, a kind of forward search algorithm is proposed, starting with the simple independence model and then successively moving to models of higher complexity. This is also the approach that we will adapt in the present paper.

Larsson (2020) also extends the model to deal with truncated Poisson distributions (which is suitable in presence of ordinal data). Another extension is the so-called mixed model, where it is allowed that the same factor can 'load' on more than one group of variables. This gives the same kind of flexibility as in traditional factor analysis models. However, in order to be able to

compare to cluster analysis ('kmeans') and principal component analysis, we will not consider mixed discrete factor models in this paper.

In Larsson (2020), empirical data from a questionnaire is analyzed. This data is ordinal, and it is modeled by dependent truncated Poisson distributions. The results from this exercise are shown to very much agree with the outcomes of the traditional factor analysis performed by Jöreskog, Olsson, and Wallentin (2016).

1.2. Dependent negative binomial models

Many empirical count data sets are subject to overdispersion, i.e., the variances of the variables are much larger than their means, making Poisson modeling insufficient. To alleviate this, we suggest in the present paper to extend the models in Larsson (2020), simply by replacing the Poisson distribution with the negative binomial. For this, we replace the probability functions in e.g., (4) by

$$f(u;r,p)=\binom{r+u-1}{u}p^r(1-p)^u. \tag{6}$$

There are many different parametrisations, see e.g., Hilbe (2011), but we choose this one for convenience, since this is the one implemented in Matlab. This parametrisation has the property that the expectation is $r(1-p)/p$ and the variance is $r(1-p)/p^2$. Moreover, the correlations are given by the formula (cf. (2))

$$\operatorname{corr}(Y_i,Y_j)=\frac{v_0}{\sqrt{(v_0+v_i)(v_0+v_j)}}, \tag{7}$$

where v_0 is the variance of the factor (U) and v_i is the variance of X_i.

For r large and p close to one, the distribution is close to a Poisson with parameter $r(1-p)$, while for small p, the Poisson approximation breaks down and there is a considerable overdispersion.

Observe that, when going from Poisson to negative binomial, the number of parameters to estimate in our models such as in (1) and (5) is doubled. This is expected to cause numerical problems for systems of large dimensions. However, so far it seems from our calculations that, analogous to the Poisson case, the number of parameters that need to be estimated is reduced by the fact that the empirical means equal the estimated expectations under the model. To our best knowledge, this is a fact that remains to be proved.

In the remainder of the article, three examples of data analysis follow in Secs. 2–4. In Sec. 5, a concluding discussion is provided.

2. Example: veterinary science, rearing of broilers

2.1. Background

We study data from organic rearing of broilers in Sweden, more precisely data concerning the welfare of chickens. Data from eight farms in Sweden were investigated. From visual inspections, various signs of injury to the chickens had been registred, with the value 0 meaning no injury, and with numbers on an ordinal scale to indicate the degree of worsened injury. The scope of the initial study was to gather new information regarding health and other welfare aspects, housing and management routines in order to describe the present situation on organic broiler farms in Sweden (Göransson, Yngvesson, and Gunnarsson 2020).

After initial cleaning of the data set, disregarding some missing values, measurements of 300 individuals remained. For the analysis below, we chose to study four variables. In the table below, these variables are stated along with short descriptions and the values taken by each variable (ordinal scales):

x_1	Feathers (feather injury)	0, 1, 2
x_2	Hock burns (injury marks)	0, 1
x_3	FPD (footpad dermatitis)	0, 1, 2
x_4	Plumage cleanliness	0, 1, 2

Recall that 0 means no injury at all and observed increased injury on a bird means higher levels on the ordinal scale. Actually, the values taken by variables were $x_1 : (0-2)$, $x_2 : (0-4)$, $x_3 : (0-4)$, $x_4 : (0-3)$, but in the table above are listed the levels where counts actually occurred in the data set.

2.2. *Statistical* analysis

To check the assumption of Poisson distribution, we simply present empirical means and variances for all variables. In Table 1, we see that the Poisson assumption is not unreasonable. Note a slight underdispersion for the fourth variable.

We also give empirical correlations, see Table 2. The method of Larsson (2020) only works when all variables have non-negative correlation. We here find one negative, but its magnitude is quite small.

2.3. Discrete factor analysis

After using the forward search method, discrete factor analysis proposes as the best model to link the variables x_2 and x_4 in one group with the same factor, and then to have the variables x_1 and x_3 each in separate groups. The parameter estimates are given in Table 3.

Note that, as we should have, the observed means are obtained as $\hat{\mu}_1$ and $\hat{\mu}_3$ for the variables x_1 and x_3, while for the variable x_2, $\hat{\lambda} + \hat{\mu}_2 \approx 0.12$, which is the observed mean for x_2, and for x_4 we similarly have $\hat{\lambda} + \hat{\mu}_4 \approx 0.54$.

As for estimated correlations from the factor model, we get them as in Table 4. Here, we have used (2) to calculate the estimated correlation between x_2 and x_4.

We find that the estimated correlation between x_2 and x_4 agrees well with the corresponding sample correlation. Naturally, all the other estimated correlations from the model are zero since these are between variables that are considered independent. In particular, this means that when comparing to sample values, we miss out on the correlation between x_3 and x_4 by some margin.

2.4. Discussion. Other techniques

With cluster analysis on these data, via the procedure kmeans, it was the variables x_1 and x_2 that fell into the same cluster when specifying the number of clusters to three. Given the low sample correlation between x_1, and x_2, we find this result to be a little surprising. Supposedly, one reason for this result could be the relative similarity between x_1 and x_2 in terms of their sample means and sample variances.

Finally, we compare the result with discrete factor analysis to a principal-component analysis (PCA). Denote by z_1, z_2, z_3 the first three principal components. Usually, cumulative percentage of the total variance is reported. Here, z_1 accounts for 36%, z_1 and z_2 for 62% and the three first principal components account for 83%.

In Table 5, we note that the first principal component loads high on variables x_2, x_3 and x_4, not exactly the result from the discrete factor analysis but somehow pointing in that direction.

We conclude with an interpretation from an etological point of view of the finding that variables 2 and 4 form a factor. In earlier analyses, researchers have found relationships between hock burns and plumage cleanliness. The reason could be that birds with hock burns possibly are to a large extent seated in the litter material, and hence get dirtier (if the litter quality is bad).

Table 1. Veterinary data, means and variances.

Variable	1	2	3	4
Mean	0.23	0.12	0.26	0.54
Variance	0.18	0.10	0.24	0.36

Table 2. Veterinary data, empirical correlations.

Variable number	1	2	3	4
1	1	0.047	−0.059	0.064
2	.	1	0.107	0.297
3	.	.	1	0.224
4	.	.	.	1

Table 3. Veterinary data, estimated parameters.

$\hat{\lambda}$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$
0.079	0.230	0.038	0.257	0.458

Table 4. Veterinary data, estimated correlations from the factor model.

Variable number	1	2	3	4
variable 1	1	0	0	0
variable 2	·	1	0	0.31
variable 3	·	·	1	0
variable 4	·	·	·	1

Table 5. Veterinary data, estimated principal components.

	x_1	x_2	x_3	x_4
z_1	0.097	0.580	0.471	0.658
z_2	−0.868	−0.170	0.464	−0.055
z_3	0.470	−0.586	0.659	−0.025

3. Example: musicology, features of fugue subjects

3.1. Background

Rydén (2020) has considered a problem arising in musicology. In the musical art form called fugue, a so-called subject is first presented. By quantifying a fugue subject, comparisons can be made on a statistical basis between J.S. Bach and composers from later epochs, a priori dividing works into three categories depending on the background of the composition in music history.

A subject could be seen as a melody, which needs to be quantified by some numerical measures. Rydén (2020) chose the following integer-valued variables:

x_1	Length, expressed in number of notes written
x_2	Range (in semitones)
x_3	Number of unique pitch classes
x_4	Initial interval (in semitones)
x_5	Number of unique intervals between successive notes
x_6	Maximum interval between successive notes (in semitones)

Consider for instance, in Figure 1, the following subject by Bach, from Fantasia and Fugue in C minor (BWV 537):

In this example, we find the following observed values of the variables:

x_1	x_2	x_3	x_4	x_5	x_6
17	9	7	7	7	9

Note that the tied note in bar 2 implies one single count of the note (A flat).

For details on the data collection, see Rydén (2020). In all, 238 fugue subjects were collected and features compiled.

3.2. Statistical analysis: prerequisites

In our analysis, it feels natural to subtract the respective minima (over individuals i) and consider the so transformed data $y_j = x_j - \min x_{ij}$.

We start by performing some descriptive analysis of the data. The means and variances are given in Table 6.

We find signs of overdispersion for y_1 in particular, and also for y_2 and y_4. The large variance in y_1 is due to some outliers.

Next, we give the empirical Pearson correlations in Table 7. We find one negative correlation (between y_3 and y_4), but this one is so close to zero that we consider it to be a minor problem. This might well be a zero correlation that has been estimated negative by pure chance.

3.3. Statistical analysis: discrete factor analysis

Next, we go on to search for the 'best' discrete factor analysis model, by seeking the one with the smallest possible AIC using the forward search method of Larsson (2020).

We find a model where variables 1,2,5,6 have a common factor, while variables 3 and 4 turn out as independent of the others, as well as of each other. This is also in accord with Table 7, where we see that the estimated correlation between y_3 and y_4 is very close to zero (as already mentioned), and also that y_3 and y_4 are the variables that correlate the least with all the other variables.

The parameter estimates for the model are given in Table 8. In particular, note that for variables y_j with a common factor, $\hat{\lambda} + \hat{\mu}_j$ equals the observed mean of y_j, and that for unrelated variables (like y_4), $\hat{\mu}_j$ equals the observed mean.

In order to validate the model, we also give the estimated correlations from the models in Table 9, cf Eq. (2). Comparing to Table 7, we see that in almost all cases, the model estimates of correlations are lower than the corresponding empirical ones.

To account for the overdispersion that is present for some of the variables, we also fitted a negative binomial factor model to the group (y_1, y_2, y_5, y_6) as well as to y_4. (The variable y_3 did not show signs of over dispersion, so it was left as it is.) The corresponding parameter estimates are given in Table 10.

Observe that, for the underdispersed variable y_5, the estimated r is very large and the estimated p is extremely close to one. This is because for large r and p close to one, the negative binomial distribution is close to Poisson. Also, for y_6, which is close to being underdispersed, we have a large estimated r and an estimated p fairly close to one.

A model check of negative binomial may be done both for variances and correlations. (As for Poisson, the expectations are estimated without errors.) Estimated variances are given in Table 11,

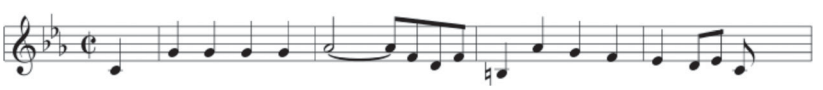

Figure 1. J.S. Bach: subject from Fugue in C minor (BWV 537).

Table 6. Music data, means and variances.

Variable number	1	2	3	4	5	6
Mean	15.9	6.9	3.6	2.9	5.0	7.3
Variance	143.5	10.3	2.6	6.4	4.2	7.7

Table 7. Music data, empirical correlations.

Variable number	1	2	3	4	5	6
variable 1	1	0.55	0.35	0.12	0.67	0.40
variable 2	·	1	0.41	0.21	0.58	0.59
variable 3	·	·	1	−0.004	0.34	0.28
variable 4	·	·	·	1	0.20	0.31
variable 5	·	·	·	·	1	0.68
variable 6	·	·	·	·	·	1

Table 8. Music data, estimated parameters, Poisson model.

$\hat{\lambda}$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$	$\hat{\mu}_6$
2.7	13.2	4.2	3.6	2.9	2.3	4.6

Table 9. Music data, estimated correlations from the Poisson factor model.

Variable number	1	2	3	4	5	6
variable 1	1	0.26	0	0	0.30	0.25
variable 2	·	1	0	0	0.46	0.38
variable 3	·	·	1	0	0	0
variable 4	·	·	·	1	0	0
variable 5	·	·	·	·	1	0.44
variable 6	·	·	·	·	·	1

Table 10. Music data, estimated parameters, negative binomial model.

$\hat{r}_0$	2.5	$\hat{p}_0$	0.44
$\hat{r}_1$	1.5	$\hat{p}_1$	0.11
$\hat{r}_2$	9.2	$\hat{p}_2$	0.71
$\hat{r}_4$	4.7	$\hat{p}_4$	0.57
$\hat{r}_5$	$1.4 \cdot 10^7$	$\hat{p}_5$	0.99999986
$\hat{r}_6$	273	$\hat{p}_6$	0.985

Table 11. Music data, estimated variances from the negative binomial model.

Variable number	1	2	3	4	5	6
Variance	127.4	12.3	3.6	5.9	9.0	11.3

Table 12. Music data, estimated correlations from the negative binomial model.

Variable number	1	2	3	4	5	6
variable 1	1	0.18	0	0	0.21	0.19
variable 2	·	1	0	0	0.68	0.60
variable 3	·	·	1	0	0	0
variable 4	·	·	·	1	0	0
variable 5	·	·	·	·	1	0.71
variable 6	·	·	·	·	·	1

and estimated correlations are in Table 12. Comparing to Table 6, the agreement of variances is not too bad, with exceptions for variables 5 and 6, where the model over-estimates by some margin. As for correlations, they fit better to the empirical ones than for Poisson as long as variable 1 is not involved. The correlations between variable 1 and other variables are more under-estimated here than in the Poisson case.

Table 13. Music data, estimated principal components.

	x_1	x_2	x_3	x_4	x_5	x_6
z_1	0.44	0.47	0.31	0.19	0.49	0.46
z_2	−0.19	−0.040	−0.49	0.82	−0.0067	0.22
z_3	0.33	−0.055	−0.78	−0.41	0.32	0.08

3.4. Discussion. Other techniques

It might be interesting to compare with the result given by PCA. Concerning the cumulative percentage of the total variance, for this data set, the first principal component accounts for 51%, the first two principal components 68% and the first three 80%. An interpretation of the coefficients could be made (see Table 13); the first is essentially a weighted linear combination of the variables, with positive weights. Less weight is put on x_4, initial interval in semitones. Turning to the second principal component, we find a contrast between variables x_1 and x_3 (length and number of pitch classes, in a sense overall measures of the subject) against x_4 and x_6 (interval features, inner construction of subject).

In addition, a conventional factor analysis with two factors was carried out. The first factor put less weight on variable x_4 (cf. the discrete factor analysis).

Finally in this section, we want to mention that we have also tried traditional cluster analysis on these data, using the procedure kmeans in Matlab. As an initial normalization (after possible transformation), we divided the variables by their respective means. Then, instructing kmeans to form two clusters on the original data, we got one cluster with the variables 1,2,3,5,6 and one with variable 4 only. For transformed data and with three clusters, we got 1,2,5,6, then 3 alone and 4 alone. The latter result exactly agrees with what we got from our discrete factor analysis procedure.

4. Example: Agriculture, damage to potato tubers

4.1. Background

When harvesting, potatoes can be damaged by the lifter device. In experiments performed at Wageningen, the Netherlands, eight types of lifting rods were compared (Keen and Engel 1997). Two energy levels, six genotypes/varieties and three weight classes were used. Most combinations of treatments involved about 20 potato tubers. Tubers were rated as undamaged to severely damaged. Data are found in the R package agridat (Wright 2018). In the following, we will only consider the four variables in the dataset that are either ordinal or counts.

In all, we consider a controlled experiment and face a data frame with 1152 observations on 4 variables:

x_1	Energy factor (1, 2)
x_2	Weight class (1–3)
x_3	Damage category (1–4)
x_4	Count of tubers in each combination of categories (integer)

Here, variables x_1–x_3 are ordinal and we present them simply as integers.

4.2. Statistical analysis

In order to adapt to the Poisson distribution (or negative binomial), and to get positive correlations, we transformed the original variables, $x_1, \ldots, x_4$ according to $y_1 = x_1 - \min x_{i1}$ and $y_j = \max x_{ij} - x_j$ for $j = 2, 3$, while $y_4 = x_4$ was left untransformed.

Table 14. Potato data, means and variances.

Variable number	1	2	3	4
Mean	0.50	1.00	1.50	4.68
Variance	0.25	0.67	1.25	31.71

Table 15. Potato data, empirical correlations.

Variable number	1	2	3	4
variable 1	1	0	0	0.03
variable 2	·	1	0	0.04
variable 3	·	·	1	0.58
variable 4	·	·	·	1

A preliminary check of means and variances, see Table 14, reveals that variable 4 is overdispersed. The other variables are underdispersed.

As for empirical correlations, these are given in Table 15. Because the data stem from a controlled experiment, the empirical correlations between y_1, y_2, y_3 are all zero. The interest is the empirical correlations of y_4 with the other variables. We find that, as accomplished by the initial transformation, all these are positive. It is also clear that the only correlation that might be of interest is the one between y_4 and y_3.

4.3. Discrete factor analysis

We went on to search for the best Poisson factor model. As expected, this model identified a common factor for variables 3 and 4, while all other variables turned out as independent. For these variables, as usual the parameter estimates equal their corresponding means. See Table 16.

Moving on to handle the overdispersion, we estimated negative binomial models for the pair (y_3, y_4). The results of these estimations are given in Table 17.

Like the music example in Sec. 3, for the underdispersed variable y_3, the large estimated r and the estimated p extremely close to one are striking.

Moreover, just like for the Poisson model, the estimated expectations of the negative binomial model equal the empirical means. However, the estimated variances are not equal to their sample counterparts. These are given in Table 18. As expected, the model overestimates the variance for the underdispersed variable 3. For variable 4, it over estimates, but the estimate may still be considered better than the corresponding Poisson model variance, which equals the mean and is much too low.

The estimated correlations for the Poisson model are zero except for the correlation between y_3 and y_4, which is 0.24. This is an underestimation of the empirical correlation from Table 15: 0.58. The corresponding number for the negative binomial model is 0.06, an even more severe underestimation.

4.4. Discussion. Other techniques

As for other methods, kmeans (with two clusters) grouped variables 1-3 together and variable 4 separately. This might be because of the large sample variance of variable 4.

In PCA, the first two principal components are given in Table 19. We find that the first principal component mainly loads on variable 4, but it also agrees with our analysis, in the sense that it also loads a little bit on variable 3. The other components, of which the second one is shown here, loads heavily on one of the variables each. Given the design of the experiment, this should be considered natural. Finally, let us remark that The first principal component accounts for 40% of the variance, while the two first stand for 65%.

Table 16. Potato data, estimated parameters, Poisson model.

$\hat{\lambda}$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$
0.63	0.50	1.00	0.87	4.04

Table 17. Potato data, estimated parameters, negative binomial model.

$\hat{r}_0$	0.51	$\hat{p}_0$	0.61
$\hat{r}_3$	$7.2 \cdot 10^6$	$\hat{p}_3$	0.99999983
$\hat{r}_4$	0.42	$\hat{p}_4$	0.09

Table 18. Potato data, estimated variances from the negative binomial model.

Variable number	1	2	3	4
Variance	0.25	0.67	1.75	49.97

Table 19. Potato data, estimated principal components.

	y_1	y_2	y_3	y_4
z_1	0.00	0.01	0.12	0.99
z_2	−0.02	−0.14	0.98	−0.12

Finally, let us mention that for this data set a conventional factor analysis with one factor gives a factor that loads heavily on y_3 and y_4 but basically not on any other variable. This is quite in accord with our results.

5. Discussion

Investigating (possible) relationships between various variables is common in many domains of science. Factor analysis is then one of the classical tools from multivariate statistical analysis, here presented in terms of the methodology given by Larsson (2020) suitable for discrete data.

The three data sets considered in this article face different situations from the point of view of data type. In Sec. 3, all four variables were on the ordinal scale, while the data set in Sec. 4 included exclusively "true" count data. In Sec. 5, the situation with three ordinal variables and one count variable was investigated (a controlled experiment). Interpreting the results of a factor analysis from the applied user's point of view is often not straight-forward, but comparing with e.g., PCA the results and practical conclusions are valid for the new methodology presented.

In this paper, in order to handle overdispersed data, we have generalized the Poisson assumption of Larsson (2020) to negative binomial. However, the search algorithm builds on Poisson models, so bringing in negative binomial here is a generalization that is left to make. Also, one could think of other models, for example different combinations of the Poisson with other distributions.

Acknowledgements

We would like to thank the editor and the anonymous referee for helpful suggestions to improve the manuscript.

References

Göransson, L., J. Yngvesson, and S. Gunnarsson. 2020. Bird health, housing and management routines on Swedish organic broiler chicken farms. *Animals* 10 (11):2098. doi:10.3390/ani10112098.

Hilbe, J. M. 2011. *Negative binomial regression*. 2nd ed. New York: Cambridge University Press.

Jöreskog, K. G., U. H. Olsson, and F. Y. Wallentin. 2016. *Multivariate analysis with LISREL*. Switzerland: Springer International Publishing.

Karlis, D. 2003. An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics* 30 (1):63–77. doi:10.1080/0266476022000018510.

Keen, A., and B. Engel. 1997. Analysis of a mixed model for ordinal data by iterative re-weighted REML. *Statistica Neerlandica* 51 (2):129–44. doi:10.1111/1467-9574.00044.

Larsson, R. 2020. Discrete factor analysis using a dependent Poisson model. *Computational Statistics* 35 (3): 1133–52. doi:10.1007/s00180-020-00960-w.

Rydén, J. 2020. On features of fugue subjects. A comparison of J.S. Bach and later composers. *Journal of Mathematics and Music* 14 (1):1–20. doi:10.1080/17459737.2019.1610193.

Wright, K. 2018. Agridat: Agricultural Datasets. R package version 1.16. https://CRAN.R-project.org/package=agridat