

# Methods for detecting outliers in latent variable models

Irini Moustaki and Martin Knott  
*London School of Economics*

Dimitris Mavridis  
*University of Edinburgh*

# Outline

- Define what an outlier is in the binary case.
- Exploratory Detection Methods: Forward Search Algorithm
- A mixture model for handling outliers.
- Applications

---

## Aberrant Response Patterns

- *Continuous responses*: a response more than 3 standard deviations away from the mean is considered an unexpected response under the normal model.
- *Categorical responses*
  1. *The response style agreement*: tendency to agree/disagree with all questions regardless their content.
  2. *The extreme response style*: tendency for some individuals to use the extreme ends of a Likert type response scale
  3. *The neutral response style*: tendency to choose the middle alternative.
  4. *An unexpected response* occurs when the associated probability is low under the true model.
  5. *Guessing*

- Those response styles should either be seen as outliers and therefore controlled or removed from the estimation process or they should be treated as a manifestation of certain respondents' characteristics.
- In the literature, interest lies on how response styles differentiate among groups defined by demographic and socio-economic variables.

## Detection of outliers and extreme response patterns

- *Residuals* (Reiser 1996)
- *Person-fit statistics* (Meijer and Sijtsma 2001).
- *Backward methods*: deletion residuals and Cook's distance measure the exact effect of the deletion of a single observation on prediction and testing. Backward methods for factor analysis models (Lee and Wang 1996; Lee and Xu 2003).
- *Forward search methods* (Mavridis and Moustaki, 2008 and 2009)

## Common errors

1. *Masking* effect occurs when an outlier goes undetected because of the presence of a cluster of outliers
2. The *swamping* effect occurs when a 'good' observation is incorrectly identified as an outlier.

---

## Treating outliers

- *Robust estimation*: ML estimation assumes that the responses are exactly generated by the true model. ML was found to **break down** in the presence of outliers. Moustaki and Victoria-Feser (2006) propose a robust estimator that does not break down when outliers are present in the data set.
- *Hybrid models* (Yamamoto, 1997; Yamamoto and Everson, 1995 & 1997; Rost, 1990; Rost and von Davier, 1993 & 1995)
- *A mixture model* proposed by Knott and Moustaki: the latent variable model accommodates for outliers in the data. **Predict and account** for the proportion of individuals that guessed a response pattern.

## Forward Search Algorithm, FS

- The FS was initially developed for robust estimation of covariance matrices (Hadi 1992) and regression models (Atkinson 1994).
- The FS starts by fitting the model to a **small robustly chosen subset** of the whole data set and proceeds by adding observations until all are included.
- A book-length treatment of the FS with various applications can be found in Atkinson and Riani (2000) and Atkinson, Riani and Cerioli (2004).

---

## Steps of the FS

1. Choose an **initial subset of size  $g$**  (usually between 50-100) from the sample of size  $n$ . This is the **'basic'** set formed at the beginning of the search while the remaining  $(n - g)$  observations constitute the **'non-basic'** set. The **'basic'** and the **'non-basic'** sets are mutually exclusive throughout the search.
2. **Progress** in the FS so that eventually all observations from the **'non-basic'** set are included in the **'basic'** set.
3. **Monitor** quantities, such as parameter estimates, residuals, and goodness-of-fit tests, during the progress of the search.

## Step 1: Choosing the initial subset

Let  $\mathbf{Y}$  be the data matrix of order  $n \times p$ .

We select the initial subset,  $S_*^g$  that satisfies

$$X_{lim}^2(S_*^g, \hat{\beta}_*) = \underbrace{\min}_i \left[ X_{lim}^2(S_i^g, \hat{\beta}_i) \right], \quad (1)$$

- $X_{lim}^2(S_i^g)$  be the chi-square test statistic of the **bivariate margins** (Maydeu-Olivares and Joe, 2005) for the observations in  $S_i^g$
- $\hat{\beta}_i$  is the estimated parameter vector **using only the observations in  $S_i^g$** .
- The asymptotic distribution of the limited information criterion does not hold for subsets of the data or for small samples.

- Usually, we select the initial subset that gives the highest  $p$ -value of the  $X_{lim}^2(S_{\mathbf{i}}^g, \hat{\beta}_{\mathbf{i}})$  statistic among a smaller number of  $H$  subsets of size  $g$ .
- **Alternatively:** for each set of parameter estimates, the median of the absolute likelihood contributions for the whole sample is taken.

$$\text{median} [S_{*}^g, \mathbf{lc}] = \underbrace{\min}_h \left\{ \text{median} \left[ \mathbf{lc} \left( \mathbf{y}, \hat{\beta}_h \right) \right] \right\}. \quad (2)$$

---

## Step 2: Progressing in the search

- From step  $g$  to  $g + 1$  a  $q$ -factor model is fitted to the ‘basic’ set  $S_*^g$ .
- The **standard FS** sorts all  $n$  observations, from the ‘basic’ and the ‘non-basic’ set according to their closeness to the ‘basic’ set.
- **Closeness** is established via a criterion based on model estimates from  $S_*^g$ .
- The  $g + 1$  observations closest to the ‘basic’ set are selected.
- This allows observations **to enter and leave** the ‘basic’ set at each step of the FS.

## Criteria for progressing

### 1. *Likelihood contributions*

Response patterns that are observed only once are most likely to enter in the last steps of the search. To avoid that problem we weight the likelihood contributions at each step by dividing them by their sample frequency in the 'basic' set when this is different from zero.

### 2. *Residuals*

## Step 3: Monitoring the FS using forward plots

- Parameter estimates ( $\hat{\beta}$ )
- $t$ -statistics
- Goodness-of-fit tests ( $X_{lim}^2$ )
- Adjusted residuals.
- Maximum absolute residual  $\max(|e_{adj}|)$ .
- An overall measure of change for  $\hat{\beta}$  is a forward version of the Cook's distance (Atkinson and Riani 2000).

$$D_l = (\hat{\beta}_{g-1} - \hat{\beta}_g)' \{cov(\hat{\beta}_{g-1})\}^{-1} (\hat{\beta}_{g-1} - \hat{\beta}_g), \quad (3)$$

---

## Example: Data on sex role expectations, Duncan 1979

- ‘Here are some things that might be done by a boy or a girl. As I read each of these to you, I would like you to tell me if it should be done as a regular task by a boy, by a girl, or by both’
  1. Shovelling walks,
  2. Washing the car,
  3. Dusting furniture,
  4. Making beds
- Responses of ‘boy’ to items 1 and 2 and ‘girl’ to items 3 and 4 were coded as ‘0’ and are referred to as the traditional answers.
- Responses of ‘both:’, refer to as ‘egalitarian’ answers are coded as ‘1’.

- Sample size 257 mothers
- One factor model:

Statistic	<i>p</i> -value
$X^2$	0.001
$X^2_{lim}$	0.014

Label	Pattern	$e_{adj}$	Frequency
1	0 0 0 0	-1.46	86
2	1 0 0 0	0.39	20
3	0 1 0 0	0.17	12
4	1 1 0 0	2.05	8
5	0 0 1 0	1.05	7
6	1 0 1 0	-1.12	2
7	0 1 1 0	0.29	4
8	1 1 1 0	-0.96	2
9	0 0 0 1	0.46	24
10	1 0 0 1	0.74	12
11	0 1 0 1	0.14	8
12	1 1 0 1	-2.38	1
13	0 0 1 1	2.96	21
14	1 0 1 1	-3.46	7
15	0 1 1 1	-3.74	8
16	1 1 1 1	3.89	35

---

## Forward search

- **Initial subset:**  $g = 50$
- **Criterion:** lowest value of the bivariate  $X_{lim}^2$  statistic among 1000 subsets of this size.
- **Weighted likelihood contributions** (by their sample frequency in the 'basic' subset) are used for progressing in the search.

## Results

- 26 six out of the 35 response patterns (1111) enter the search in the last 28 steps.
- The remaining nine individuals with response pattern 1111 are included in the initial subset.
- The response pattern that entered in the last two iterations is 1100.
- There is a sharp increase from step 219 to 228. The response patterns that enter during that part of the search are 1110 (2 times), 0110 (1 time), 1100 (1 time), 1001 (1 time), 0011 (3 times) and again 1001 (2 times).

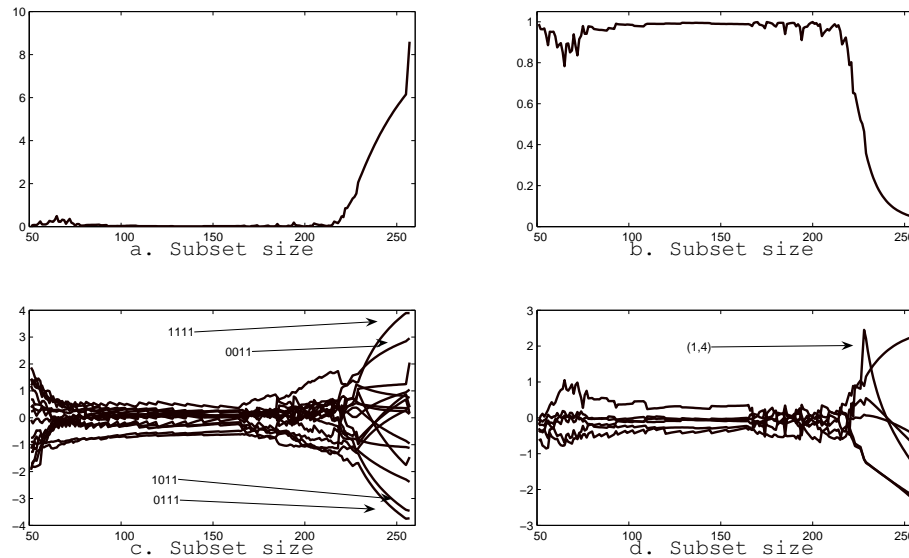


Figure 1: a) limited information goodness-of-fit statistic,  $X_{lim}^2$  b) asymptotic  $p$ -value of  $X_{lim}^2$ , c) adjusted residuals for the 16 distinct patterns, d) limited information adjusted residuals referring to positive responses to pair of items.

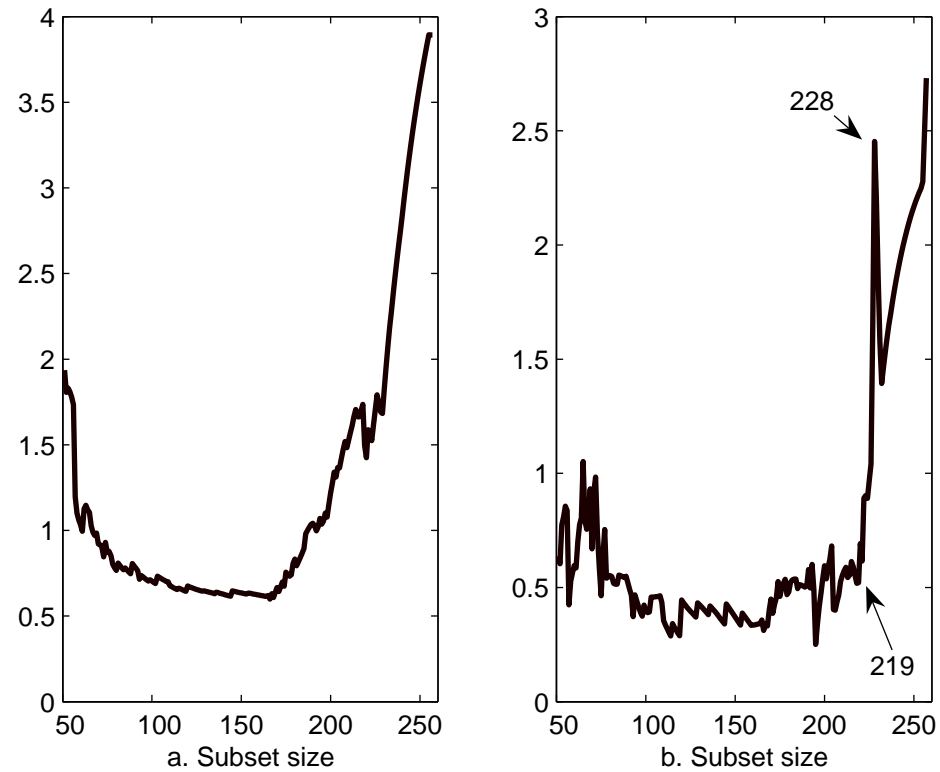


Figure 2: a) maximum absolute overall residual,  $\max(|e_{adj}|)$ , b) limited information residuals,  $\max(|e_{adj}^l|)$  where  $l = 2$ .

---

## A generated data set with 8 items

- Generated a sample of 300 obs. and 8 items using a one-factor model.
- $\beta_0 = (0.29, -1.34, 0.71, 1.62, -0.69, 0.86, 1.25, -1.59)'$
- $\beta_1 = (0.49, 1.33, 0.82, 1.41, 1.50, 1.43, 1.91, 1.40)'$
- The sample proportions of positive responses: 0.55, 0.24, 0.61, 0.73, 0.39, 0.62, 0.64 and 0.23 respectively.
- We artificially create an outlier by substituting a random subject with the response pattern  $[01001001] = '147'$ .

- The initial subset of size 150 was selected among 75 randomly constructed sub-samples according to the  $p$ -value of the bivariate  $X_{lim}^2$  test statistic.
- Likelihood contributions are used for adding observations in the 'basic' set.

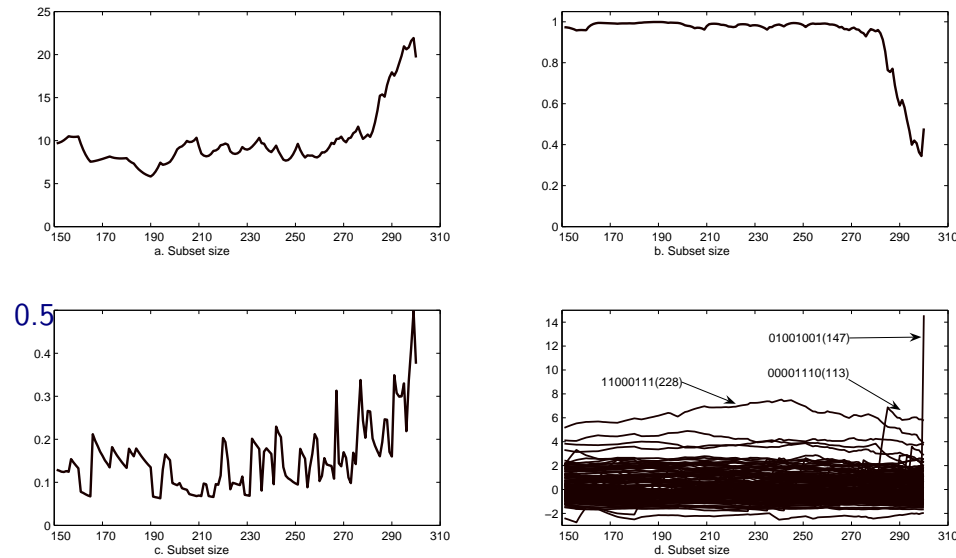


Figure 3: a) limited information goodness-of-fit statistic,  $X_{lim}^2$  b) asymptotic  $p$ -value of  $X_{lim}^2$ , c) Cook's distance, d) adjusted residuals.

## Simulation envelopes

- The distribution of test statistics is unknown for subsets of the data.
- Simulation envelopes are constructed via parametric bootstrapping.
- Points found outside the simulation envelopes need to be further examined.

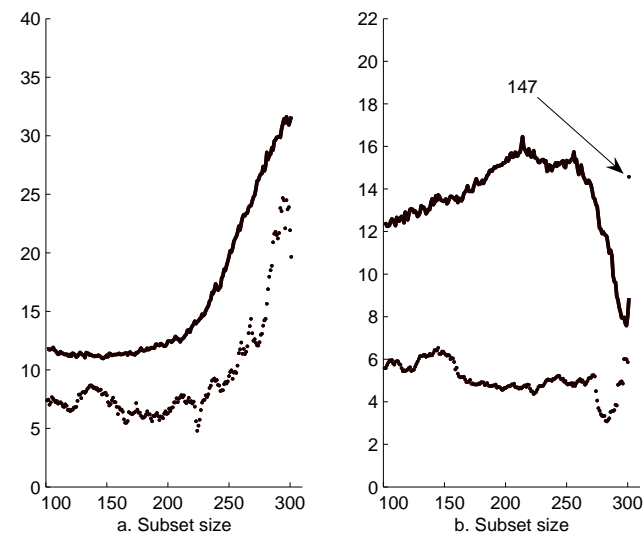


Figure 4: a) limited information goodness-of-fit statistic,  $X^2_{lim}$ , b) maximum absolute adjusted residual,  $\max(|e_{adj}|)$  accompanied by 95% simulation envelopes (solid lines).

## Model the outliers using a LVM

LVM examine associations among a set of  $p$  observed variables  $(y_1, y_2, \dots, y_p)$  using  $q$  latent variables  $(z_1, z_2, \dots, z_q)$  where  $q$  is much less than  $p$

## Theoretical Framework

Bartholomew and Knott (1999)

As only  $\mathbf{y}$  can be observed any inference must be based on the joint distribution of  $\mathbf{y}$ :

$$f(\mathbf{y}) = \int_{R_{\mathbf{z}}} g(\mathbf{y} | \mathbf{z}) \phi(\mathbf{z}) d\mathbf{z}$$

$\phi(\mathbf{z})$ : prior distribution of  $\mathbf{z}$

$g(\mathbf{y} | \mathbf{z})$ : conditional distribution of  $\mathbf{y}$  given  $\mathbf{z}$ .

Note that  $\phi(\mathbf{z})$  and  $g(\mathbf{y} | \mathbf{z})$  are not uniquely determined.

What we want to know:  $\phi(\mathbf{z} | \mathbf{y})$

If correlations among the  $y$ 's can be explained by a set of latent variables then when all  $z$ 's are accounted for the  $y$ 's will be independent (local independence).

$q$  must be chosen so that:

$$g(\mathbf{y} \mid \mathbf{z}) = \prod_{i=1}^p g(y_i \mid \mathbf{z})$$

The question is whether  $f(\mathbf{y})$  admit the presentation:

$$f(\mathbf{y}) = \int_{R_{\mathbf{z}}} \prod_{i=1}^p g(y_i \mid \mathbf{z}) h(\mathbf{z}) d\mathbf{z}$$

for some small value of  $q$ .

---

## Notation

Response patterns are divided into:

‘Extreme response pattern’ denoted by  $\mathbf{y}^e$ . (e.g. pattern 11111)

Non-extreme response pattern denoted by  $\mathbf{y}^{\bar{e}}$ .

A pseudo item is used to indicate whether an extreme response pattern,  $\mathbf{y}^e$ , is guessed or not.

The pseudo item is denoted with  $u$  and it takes the value 1 when a response pattern is guessed and 0 otherwise.

Note that the pseudo item is not observed in the data since we do not know in advance the number of extreme response patterns that have been guessed.

Table 1: Response mechanism

	Guessing ( $u = 1$ )	No Guessing ( $u = 0$ )	Total
Extreme Response ( $\mathbf{y}^e$ )	$n_{e,g}$	$n_{e,\bar{g}}$	$n_e$
Non-Extreme Response ( $\mathbf{y}^{\bar{e}}$ )	0	$n_{\bar{e},\bar{g}}$	$n_{\bar{e}}$
Total	$n_g$	$n_{\bar{g}}$	$n$

- $P(u = 1 | \mathbf{y}^{\bar{e}}) = 0$  and  $P(u = 0 | \mathbf{y}^{\bar{e}}) = 1$ .
- $P(\mathbf{y}^e | u = 1) = 1$ .

## Modelling the guessing response mechanism

We define the distributions of the responses to the items  $(y_1, \dots, y_p)$  and the unobserved guessing item  $u$  conditional on a single latent variable  $z$ .

Under the assumption of conditional independence

$$f(\mathbf{y}^e, u = 0 \mid z) = \left[ \prod_{i=1}^p f_{y_i}(y_i^e \mid z) \right] f_u(u = 0 \mid z) \quad (4)$$

$$f(\mathbf{y}^e, u = 1 \mid z) = f_u(u = 1 \mid z) \quad (5)$$

$$f(\mathbf{y}^{\bar{e}}, u = 0 \mid z) = \left[ \prod_{i=1}^p f_{y_i}(y_i^{\bar{e}} \mid z) \right] f_u(u = 0 \mid z), \quad (6)$$

$$f(\mathbf{y}^{\bar{e}}, u = 1 \mid z) = 0 \quad (7)$$

The density  $f_{y_i}(y_i | z)$  of each binary item taken to be the Bernoulli:

$$f_{y_i}(y_i | z) = [\pi_i(z)]^{y_i} [1 - \pi_i(z)]^{1-y_i}, \quad i = 1, \dots, p \quad (8)$$

where  $\pi_i(z) = P(y_i = 1 | z)$ .

The response probability  $\pi_i(z)$  for the  $p$  observed items is modelled with the two-parameter logistic model.

$$\text{logit}\pi_i(z) = \alpha_{0i} + \alpha_{1i}z, \quad i = 1, \dots, p, \quad (9)$$

where the parameters  $\alpha_{0i}$  and  $\alpha_{1i}$  are the intercepts and factor loadings respectively.

The model for the **pseudo item**  $u$  becomes:

$$\text{logit}\pi_u(z, \mathbf{x}) = \alpha_{0,u} + \alpha_{1,u}z + \sum_{j=1}^s \beta_{j,u}x_s, \quad (10)$$

where  $\beta_{j,u}$  are regression coefficients.

## Estimation

The complete likelihood for a random sample of size  $n$  is:

$$l = \prod_{m=1}^n f(\mathbf{y}_m, u_m, z_m) \quad (11)$$

with  $f(\mathbf{y}_m, u_m, z_m)$ .

The only observed part is  $\mathbf{y}'_m = (y_{1m}, \dots, y_{pm})$ , where both the pseudo item  $u_m$  and the latent variable  $z_m$  are not observed.

For the maximization of the log-likelihood the E-M algorithm is used.

## Model interpretation

What do we expect when we adjust for guessed extreme patterns?

- Improve the fit (study goodness-of-fit measures)
- Interpret the factor loadings taking into account the existence of outliers
- Relate guessing mechanism to covariates and identify demographic groups that are more inclined to guess than other groups.

## A simulation study: conditions

1. We generated 1000 observations for four correlated attitudinal items and one pseudo item.
2. 1500 simulations were conducted.
3. Based on the generated values of the simulated pseudo item, a proportion of each pattern is taken to be guessed.
4. Each response pattern in turn was considered to be the 'extreme pattern'. All the guessed response patterns according to the generated values of the pseudo item were replaced with the 'extreme pattern' at the time.
5. At each simulation run the model with the guessing mechanism was fitted 16 times ( $2^4$ ).

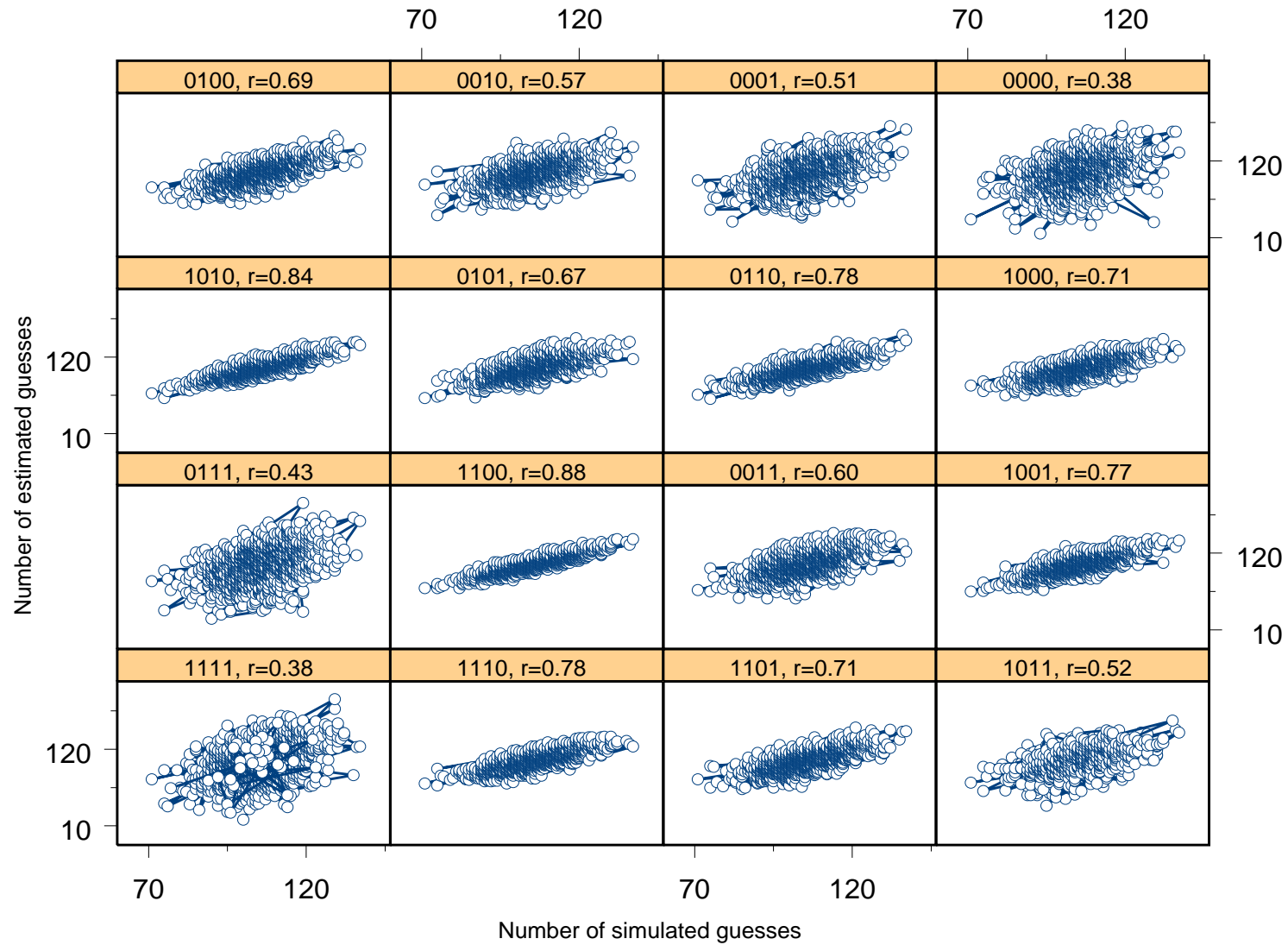


Figure 5: Scatter plots for the estimated and simulated number of guesses

Response pattern	Pattern frequency (mean)	Simulated mean number of guesses	Estimated mean number of guesses	Simulated % of guesses (mean)	Estimated % of guesses (mean)
1 1 1 1	207.65	105.73 (9.46)	104.25 (24.52)	50.92	50.00
1 1 1 0	129.96	105.73 (9.46)	104.94 (11.93)	81.36	80.61
1 1 0 1	144.87	105.73 (9.46)	104.96 (13.75)	72.99	72.26
1 0 1 1	173.89	105.73 (9.46)	105.30 (16.88)	60.80	60.36
0 1 1 1	205.06	105.73 (9.46)	104.78 (23.80)	51.57	50.87
1 1 0 0	124.36	105.73 (9.46)	105.24 (10.74)	85.02	84.51
0 0 1 1	194.03	105.73 (9.46)	105.32 (15.78)	54.49	54.06
1 0 0 1	147.63	105.73 (9.46)	105.26 (12.40)	71.62	71.11
1 0 1 0	134.00	105.73 (9.46)	105.22 (11.60)	78.91	78.36
0 1 0 1	160.31	105.73 (9.46)	105.52 (13.04)	65.96	65.63
0 1 1 0	141.75	105.73 (9.46)	105.32 (11.94)	74.59	74.13
1 0 0 0	139.61	105.73 (9.46)	104.91 (13.05)	75.74	74.99
0 1 0 0	144.26	105.73 (9.46)	104.53 (14.76)	73.30	72.24
0 0 1 0	160.62	105.73 (9.46)	104.37 (16.51)	65.83	64.78
0 0 0 1	181.85	105.73 (9.46)	103.57 (20.52)	58.15	56.69
0 0 0 0	196.04	105.73 (9.46)	104.33 (22.51)	53.94	52.99

Table 2: Estimated and true parameters,  $\hat{\alpha}_{i0}$ , 24000 simulations

Item	True	Mean	Median	Q1	Q3	M-estimator
1	-0.50	-0.52	-0.51	-0.57	-0.45	-0.51
2	-0.25	-0.26	-0.26	-0.32	-0.20	-0.26
3	0.20	0.19	0.19	0.12	0.25	0.19
4	0.60	0.60	0.59	0.52	0.67	0.59
5	-2.50	-3.48	-2.51	-2.71	-2.37	-2.52

Table 3: Estimated and true parameters,  $\hat{\alpha}_{i1}$ , 24000 simulations

Item	True	Mean	Median	Q1	Q3	M-estimator
1	0.60	0.62	0.61	0.51	0.73	0.61
2	0.80	0.83	0.81	0.69	0.95	0.82
3	1.00	1.05	1.02	0.89	1.20	1.02
4	1.20	1.27	1.22	1.04	1.44	1.23
5	-1.00	-1.86	-0.99	-1.06	-0.93	-1.00

Table 4: Estimated asymptotic and simulated **standard errors**, 24000 simulations

Item	$\hat{\alpha}_{i0}$ Asymptotic	$\hat{\alpha}_{i0}$ Simulated	$\hat{\alpha}_{i1}$ Asymptotic	$\hat{\alpha}_{i1}$ Simulated
1	0.11	0.09	0.16	0.18
2	0.13	0.09	0.19	0.20
3	0.15	0.01	0.26	0.27
4	0.19	0.12	0.35	0.38
5	1.00	2.64	1.52	2.35

## Conclusions

- The model itself adjusts for guessed extreme response patterns.
- Extensions to other types of extreme responses.
- Extensions to ordinal and nominal responses is trivial.
- Use covariates for strengthening the power of the model. e.g. guessing can be a function of demographic characteristics.
- Link and compare this method with other methods available such as robust estimation and subset regression methods as well as outlier detection methods (FS).

## References

1. Mavridis and Moustaki (2008) Detecting outliers in factor analysis using the forward search algorithm. *Multivariate Behavioral Research*.
2. Mavridis and Moustaki (2009) The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of Computational and Graphical Statistics*.
3. Knott and Moustaki (submitted) Latent variable models that account for atypical responses.
4. Moustaki, I., and Victoria-Feser, M. P. (2006) Bounded-influence robust estimation in Generalized Linear Latent Variable Models. *Journal of the American Statistical Association*.

## **Limitations in instrumental activities of daily living, 2004, Switzerland**

- Data Source: Survey of Health, Ageing and Retirement in Europe.
- It is a cross-national panel database of more than 30,000 individuals aged 50 or over.

The seven binary items analyzed comprise the scale on limitations in instrumental activities of daily living (IADL) and are:

Respondents are asked to report whether they have any difficulties with the activities below:

1. difficulties using a map in a strange place
2. difficulties preparing a hot meal
3. difficulties shopping for groceries
4. difficulties making telephone calls
5. difficulties taking medications
6. difficulties doing work around the house or garden
7. difficulties managing money

Responses are coded as '1' if the respondent selected the item and '0' otherwise.

All eleven countries show low levels on the IADL scale.

The sample size for Switzerland is 952.

Table 5: Observed and expected frequencies for the one-factor model, Switzerland

Observed	Expected	Pattern
873	871.98	0 0 0 0 0 0 0
4	7.97	0 0 1 0 0 0 0
18	21.07	0 0 0 0 0 1 0
18	17.31	1 0 0 0 0 0 0
<b>11</b>	<b>2.49</b>	<b>0 0 1 0 0 1 0</b>

We carried on the analysis by fitting the model with the guessing mechanism and gender (1=male, 0=female) as an explanatory variable for the guessing pseudo item.

Table 6: AIC and BIC

	AIC	BIC
One-factor model	1043.56	1111.59
One-factor with guessing and covariate	1019.85	1097.59

---

Item	$\hat{\alpha}_{i0}$	$\hat{\alpha}_{i1}$	$\hat{\beta}_i$
1	-7.53	3.89	
2	-12.41	4.85	
3	-8.62	3.65	
4	-10.12	3.38	
5	-11.84	4.27	
6	-4.69	1.73	
7	-8.10	3.28	
8	-7.39	-2.89	-1.36

The higher the difficulty level the less likely is a respondent to guess.

Furthermore, men are less likely to guess than women.

The model estimated 10.24 respondents to have guessed out of the 11.