Helmut Strasser

The covariance structure of conditional maximum likelihood estimates

The covariance structure of cml-estimates in the Rasch model

Helmut Strasser*

August 23, 2012

Abstract

In this paper we consider conditional maximum likelihood (cml) estimates for item parameters in the Rasch model under random subject parameters. We give a simple approximation for the asymptotic covariance matrix of the cml-estimates. The approximation is stated as a limit theorem when the number of item parameters goes to infinity. The results contain precise mathematical information on the order of approximation.

The results enable the analysis of the covariance structure of cml-estimates when the number of items is large. Let us give a rough picture. The covariance matrix has a dominating main diagonal containing the asymptotic variances of the estimators. These variances are almost equal to the efficient variances under ml-estimation when the distribution of the subject parameter is known. Apart from very small numbers n of item parameters the variances are almost not affected by the number n. The covariances are more or less negligible when the number of item parameters is large. Although this picture intuitively is not surprising it has to be established in precise mathematical terms. This has been done in the present paper.

The paper is based on previous results [5] of the author concerning conditional distributions of non-identical replications of Bernoulli trials. The mathematical background are Edgeworth expansions for the central limit theorem. These previous results are the basis of approximations for the Fisher information matrices of cml-estimates. The main results of the present paper are concerned with the approximation of the covariance matrices.

Numerical illustrations of the results and numerical experiments based on the results are presented in Strasser, [6].

*Institute for Statistics and Mathematics, WU, Vienna

1 Introduction

Let X be an item-response variable with values 0 and 1, and let θ be the logit parameter of probability of success. There are n independent items $\boldsymbol{X} = (X_1, \ldots, X_n)$ with parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ to be presented to each subject. The so-called Rasch-model (for general information see Fischer and Molenaar, [3]), is an item-response model where the logit parameter is assumed to be a sum $\boldsymbol{\theta} = \boldsymbol{\beta} + \tau$. The model parameters are identifiable if $n > 2$ and if the item parameters are restricted by a condition like $\sum_{j=1}^{n} \beta_j = 0$. Let H_n be the set of all item parameters $\boldsymbol{\beta} \in \mathbb{R}^n$ satisfying this condition.

The parameter $-\boldsymbol{\beta}$ is considered as item difficulty and τ as subject ability. By $P_{\boldsymbol{\beta},\tau}$ we denote the probability distribution of the items for fixed subject parameter τ, and by $P_{\boldsymbol{\beta},\Gamma}$ for the case of a random subject parameter with distribution Γ.

If the n items are presented to N subjects then we obtain observation vectors $\boldsymbol{X}^{(\nu)} = (X_1^{(\nu)}, \ldots, X_n^{(\nu)})$), $\nu = 1, \ldots, N$. If the distribution Γ is concentrated on a single value τ or depends smoothly on finitely many parameters then is straightforward to construct asymptotically efficient estimators of all unknown parameters. But if the distribution Γ is not known at all, then for a general item-response model estimation of the item parameter $\boldsymbol{\beta}$ is a semi-parametric problem where the construction of reasonable estimators can be difficult.

The attractive feature of the Rasch model is due to the fact that given the sum $S = \sum_{j=1}^{n} X_j$ of successes the conditional probabilities $P_{\boldsymbol{\beta},\tau}(\boldsymbol{X} = \boldsymbol{x}|S)$ do not depend on the subject parameter τ. This makes it possible to estimate item parameters by conditional maximum likelihood estimates (cml-estimates) even if Γ is unknown. The reason is that the conditional log-likelihoods

$$\ell(\boldsymbol{x}, \boldsymbol{\beta}) = \log P_{\boldsymbol{\beta},\tau}(\boldsymbol{X} = \boldsymbol{x}|S) \tag{1}$$

satisfy the Kullback-Leibler property for the item parameter $\boldsymbol{\beta}$, i.e.

$$E_{\boldsymbol{\beta},\Gamma}(\ell(.,\boldsymbol{\beta})) > E_{\boldsymbol{\beta},\Gamma}(\ell(.,\boldsymbol{b})),\ \boldsymbol{b} \in H_n,\ \boldsymbol{b} \neq \boldsymbol{\beta}.$$

Cml-estimates are thus defined by

$$\sum_{\nu=1}^{N} \ell(\boldsymbol{X}_\nu, \hat{\boldsymbol{\beta}}_N) = \max_{\boldsymbol{b} \in H_n} \sum_{\nu=1}^{N} \ell(\boldsymbol{X}_\nu, \boldsymbol{b}).$$

From Anderson, [1], it is known that under $P_{\boldsymbol{\beta},\Gamma}$ the cml-estimates $(\hat{\boldsymbol{\beta}}_N)$ for $N \to \infty$ are consistent and asymptotically normally distributed.

Although the conditional log-likelihoods do not depend on the subject parameter τ, the unconditional distribution of the cml-estimates lacks this independence. This should be clear since the unconditional distribution of the observations depends on τ resp. on Γ. The asymptotic covariance matrix $\boldsymbol{\Sigma}_{n\Gamma}$ and thus statistical precision of the cml-estimates depends on the distribution Γ of the subject parameters. Asymptotic efficiency of the cml-estimates for unknown Γ has been established by Pfanzagl, [4]. However, this kind of efficiency is rather a minimax property stating that in certain worst cases there are no better estimators than cml-estimators. Therefore it is important to know how the covariance matrix $\boldsymbol{\Sigma}_{n\Gamma}$ depends on Γ.

The subject of the present paper is the study of the asymptotic covariance matrix $\boldsymbol{\Sigma}_{n\Gamma}$ of the cml-estimates $(\hat{\boldsymbol{\beta}}_N)$. In our notation $\boldsymbol{\Sigma}_{n\Gamma}$ we suppress the dependence of the covariances on the item parameters $\boldsymbol{\beta}$, but we indicate the number n of items, this being the subject of our special attention.

As a matter of fact, the matrix $\boldsymbol{\Sigma}_{n\Gamma}$ can be evaluated numerically for moderate numbers n of items. However, such calculations turn out to be extremely computer-intensive since a large number of elementary symmetric functions have to be evaluated. The numerical complexity of such calculations increases with $n!$ and thus is not feasible for large numbers n. Moreover, any numerical evaluation does not uncover structural properties of the covariance matrix.

The main results of the present paper are approximations of the covariance matrix $\boldsymbol{\Sigma}_{n\Gamma}$ for large numbers n of items. The approximations are surprisingly simple and of high precision if $n \geq 20$. Thus, the approximation makes it possible to study the asymptotic covariance structure of cml-estimates for moderate and large numbers n of item parameters.

Let us give a brief overview of our results. As a first step we require the Fisher-information matrix

$$\boldsymbol{F}_{n\Gamma} = E_{\boldsymbol{\beta},\Gamma}(\ell'(.,\boldsymbol{\beta})\ell'(.,\boldsymbol{\beta})^t) \tag{2}$$

Let $\boldsymbol{F}_{n\tau}$ denote the Fisher-information matrix if Γ is concetrated on a single value τ. Clearly, $\boldsymbol{F}_{n\Gamma} = E_{\boldsymbol{\beta},\Gamma}(\boldsymbol{F}_{n\tau})$. Let

$$p_j(\tau) = P_{\boldsymbol{\beta},\tau}(X_j = 1) = \frac{e^{\beta_j+\tau}}{1+e^{\beta_j+\tau}} \text{ and } \sigma_j^2(\tau) = p_j(\tau)(1-p_j(\tau)).$$

From Strasser [5], 2011, it is known that the Fisher-information matrix $\boldsymbol{F}_{n\tau}$ can be approximated by a matrix $\boldsymbol{G}_{n\tau}$ with entries

$$\boldsymbol{G}_{n\tau,jk} = \sigma_j^2(\tau)\delta_{jk} - \frac{1}{n}\frac{\sigma_j^2(\tau)\sigma_k^2(\tau)}{\overline{\sigma}^2(\tau)}. \tag{3}$$

Moreover, it is proved that the maximal deviations between $\boldsymbol{F}_{n\tau}$ and $\boldsymbol{G}_{n\tau}$ are of order n^{-2}. The mathematical proof is technically complicated and is based on Edgeworth expansions for the central limit theorem in case of non-identically distributed variables (cf.

Bhattacharya and Rao [2]). As an immediate consequence we obtain that also for a random subject parameter the Fisher-information matrix $\boldsymbol{F}_{n\Gamma}$ can be approximated by the corresponding matrix

$$\boldsymbol{G}_{n\Gamma} = E_\Gamma(\boldsymbol{G}_{n\tau}), \tag{4}$$

and that the maximal deviation if of the same magnitude n^{-2} (Corollary 2.2).

It is not difficult to see (cf. Lemma 3.2), that the covariance matrix $\boldsymbol{\Sigma}_{n\Gamma}$ is identical to the Moore-Penrose pseudoinverse (MP-inverse) $\boldsymbol{F}_{n\Gamma}^{+}$ of the Fisher information $\boldsymbol{F}_{n\Gamma}$. It is therefore a natural guess that the MP-inverse $\boldsymbol{G}_{n\Gamma}^{+}$ should be a good approximation of the covariance matrix $\boldsymbol{\Sigma}_{n\Gamma}$. In Theorems 2.3 and 2.4 of the present paper we prove this suggestion and provide estimates for the magnitude of deviations.

In Theorem 2.5 we provide an explicit expression for the matrix $\boldsymbol{G}_{n\Gamma}^{+}$ which approximates the covariance $\boldsymbol{\Sigma}_{n\Gamma}$. Let us conclude with some discussion of this last result.

It is interesting to start with the case of a fixed subject parameter. In this case a very simple explicit expression for the approximate covariance matrix $\boldsymbol{G}_{n\tau}^{+}$ is available. By easy calculations it can be shown that (with simplifying notation $\sigma_j^2 := \sigma_j^2(\tau)$)

$$\boldsymbol{G}_{n\tau,jk}^{+} = \frac{1}{\sigma_j^2}\delta_{jk} - \frac{1}{n}\Big(\frac{1}{\sigma_j^2} + \frac{1}{\sigma_k^2}\Big) + \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}. \tag{5}$$

This approximation of $\boldsymbol{\Sigma}_{n\tau}$ is practically exact for $n \geq 30$.

In Theorem 2.5 we show that $\boldsymbol{G}_{n\Gamma}^{+}$ has the same structure with σ_j^2 replaced by $E_\Gamma(\sigma_j^2(\tau))$, ie.

$$\boldsymbol{G}_{n\Gamma,jk}^{+} \sim \frac{1}{E_\Gamma(\sigma_j^2(\tau))}\delta_{jk} - \frac{1}{n}\Big(\frac{1}{E_\Gamma(\sigma_j^2(\tau))} + \frac{1}{E_\Gamma(\sigma_k^2(\tau))}\Big) + \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{E_\Gamma(\sigma_i^2(\tau))}. \tag{6}$$

In fact, the exact expression is a infinite series whose leading term is (6), and the remainder terms are negligible in typical cases.

There are two important implications. The first is a practical advice and the second is a remarkable theoretical insight.

First we observe that any increase of the number n of item parameters does not change the overall covariance structure. The asymptotic covariance matrix has a dominating diagonal being independent of n. The other entries are small and vanish as $n \to \infty$. Only item parameters with extraordinary small σ^2 have to be compensated by increasing the number n. This structure answers the question whether one should estimate as many item parameters as possible together (in a single optimization procedure), or it is sufficient to estimate item parameters in parallel groups of moderate size. In the light of our results the latter approach does not deteriorate efficiency.

The second implication is related to the question of efficiency. As easy calculations show, the expressions in (5) and (6) coincide with the asymptotic covariance structures of non-conditional ml-estimators for known τ resp. known Γ. These non-conditional ml-estimators are known to be Fisher-efficient in the sense of minimizing the asymptotic variance. Thus, for the Rasch model lacking knowledge of Γ and applying cml-estimation typically is not worse a situation than knowing Γ and applying ml-estimation. This assertion is much stronger than semiparametric efficiency.

To be honest, the second implication is only true if the number of items n is large enough to neglect the residual terms of the series in Theorem 2.5 . This is the case if certain eigenvalues, which are known to be smaller than one, are noticable distant from one. At present we have numerical indications that this condition on the eigenvalues is only violated if the variance of Γ is very large. In this direction further research is required.

Let us conclude this introduction by a disclaimer. The paper does not address covariance matrices of cml-estimates for finite sample size. Only the asymptotic covariance structure of cml-estimates is studied. The problem is left open how large sample sizes must be compared with item numbers in order to get reasonable approximations to finite covariance structures.

2 The main results

Our starting point is an approximation of the Fisher-information matrix $\boldsymbol{F}_{n\tau}$ for a fixed subject parameter. From (16) and Lemma 3.1 we obtain the expression of the Fisher information matrix

$$\boldsymbol{F}_{n\tau} = E_{\boldsymbol{\beta},\tau}((\boldsymbol{X} - E_{\boldsymbol{\beta}}(\boldsymbol{X}|S))(\boldsymbol{X} - E_{\boldsymbol{\beta}}(\boldsymbol{X}|S))^T) \tag{7}$$

as covariance matrix of conditional expectations. For the approximation of such covariance matrices we apply expansions from Theorem 2.4 of Strasser [5]. Let us briefly refer this result.

Let (X_i) be a sequence of independent Bernoulli variables and let $p_i = P(X_i = 1)$. Let further $\sigma_i^2 := p_i(1-p_i)$, $\overline{\sigma}_n^2 = \frac{1}{n}\sum_{j=1}^n \sigma_j^2$ and

$$\pi_{ni} := 2\Big(p_i - \frac{\sum_{j=1}^n \sigma_j^2 p_j}{\sum_{j=1}^n \sigma_j^2}\Big). \tag{8}$$

Denote $S_n = \sum_{j=1}^n X_j$.

If $(\boldsymbol{A}_n)$ is a sequence of $n \times n$-matrices then we write

$$\boldsymbol{A}_{n,ij} = O(n^{-r}) \;\Leftrightarrow\; \max_{ij} |\boldsymbol{A}_{n,ij}| \le \frac{C}{n^r}.$$

2.1 PROPOSITION. *(Strasser [5], Theorem 2.4)*

Assume that the sequence (p_i) *is contained in a closed subinterval of* $(0,1)$*. Then*

$$E\Big((X_i - E(X_i|S_n))(X_j - E(X_j|S_n))\Big) = \sigma_i^2\delta_{ij} - \frac{1}{n}\frac{\sigma_i^2\sigma_j^2}{\overline{\sigma}_n^2} + \frac{1}{n^2}\frac{\sigma_i^2\sigma_j^2\pi_{ni}\pi_{nj}}{2\overline{\sigma}_n^4} + O(n^{-5/2}).$$

The maximal deviations between matrices with increasing row and column numbers do not give reliable information about the structural differences beween the matrices. More information is obtained by considering the matrix norm of the differences (cf. section 4, equation (17)). It is obvious how to modify Proposition 2.1 in terms of matrix norms (use inequality (18)).

Now let us turn to Rasch models. In the following we consider a sequence of Rasch models with an increasing number n of items. The global assumption of our results will be that the item parameters and the subject parameters remain bounded as $n \to \infty$. This means that the difficulties of items and the abilities of subjects do not become infinitely large or small as $n \to \infty$. To put it into mathematical terms for the subject parameter: The support of the distribution Γ is assumed to be bounded.

It is straightforward to apply Proposition 2.1 for obtaining an approximation of the Fisher-information $\boldsymbol{F}_{n\Gamma}$. The following corollary summarizes this application. Recall the definition of the matrices $\boldsymbol{G}_{n\Gamma}$ by (3) and (4). Moreover, define $\boldsymbol{H}_{n\tau}$ by

$$\boldsymbol{H}_{n\tau,jk} = \boldsymbol{G}_{n\tau,jk} + \frac{1}{n^2}\frac{\sigma_j^2(\tau)\sigma_k^2(\tau)\pi_{nj}(\tau)\pi_{nk}(\tau)}{2\overline{\sigma}_n(\tau)^4} \tag{9}$$

and

$$\boldsymbol{H}_{n\Gamma} = E_\Gamma(\boldsymbol{H}_{n\tau}) \tag{10}$$

2.2 COROLLARY. *Assume that item parameters and subject parameters are bounded. Then*

(1) $\boldsymbol{F}_{n\Gamma} = \boldsymbol{G}_{n\Gamma} + O(n^{-2})$*, and* $||\boldsymbol{F}_{n\Gamma} - \boldsymbol{G}_{n\Gamma}|| = O(n^{-1})$.

(2) $\boldsymbol{F}_{n\Gamma} = \boldsymbol{H}_{n\Gamma} + O(n^{-5/2})$*, and* $||\boldsymbol{F}_{n\Gamma} - \boldsymbol{H}_{n\Gamma}|| = O(n^{-3/2})$.

Proof: These assertions are obvious for fixed subject parameter (Γ replaced by τ). The general assertions follow by taking expectations w.r.t. to Γ and applying inequality (20). □

The remaining mathematical problem is how to use these approximations in order to obtain corresponding approximations for the covariance matrices of cml-estimates. This is the content of the following two main results of the present paper.

Let $\boldsymbol{e} := (1, \ldots, 1)^t \in \mathbb{R}^n$ and recall that $H_n \subseteq \mathbb{R}^n$ is the orthogonal complement of $\boldsymbol{e}$. The letter $\boldsymbol{E}_n$ denotes the identity matrix of order n.

2.3 THEOREM. *Assume that item parameters and subject parameters are bounded. Then*

$$||\boldsymbol{\Sigma}_{n\Gamma} - \boldsymbol{G}_{n\Gamma}^{+}|| = O(n^{-1}).$$

Proof: It is known from 2.2 that $||\boldsymbol{F}_{n\Gamma} - \boldsymbol{G}_{n\Gamma}|| = O(n^{-1})$. In view of 4.1 it remains to show that the matrix norms $||\boldsymbol{\Sigma}_{n\Gamma}|| = ||\boldsymbol{F}_{n\Gamma}^{+}||$ und $||\boldsymbol{G}_{n\Gamma}^{+}||$ are bounded.

Let us start with $||\boldsymbol{G}_{n\Gamma}^{+}||$. The positive eigenvalues of $\boldsymbol{G}_{n\Gamma}^{+}$ are the reciprocals of the positive eigenvalues of $\boldsymbol{G}_{n\Gamma}$. Therefore we need a positive lower bound of the non-vanishing eigenvalues of $\boldsymbol{G}_{n\Gamma}$. By Lemma 4.2 it is sufficient to show

$$\inf_{||\boldsymbol{x}||=1,\, \boldsymbol{x}\in H_n} \boldsymbol{x}^t \boldsymbol{G}_{n\tau} \boldsymbol{x} \geq \inf_{i,\tau} \sigma_i^2(\tau). \tag{11}$$

since this lower bound is positive by assumption.

In order to prove (11) let $||\boldsymbol{x}|| = 1$ and $\boldsymbol{x} \in H_n$, i.e. $\boldsymbol{x}^t\boldsymbol{e} = 0$. Then

$$(\boldsymbol{G}_{n\tau}\boldsymbol{x})_i = \sigma_i^2(\tau)(x_i - \lambda(\tau)), \text{ where } \lambda(\tau) = \frac{\sum_j \sigma_j^2(\tau) x_j}{\sum_j \sigma_j^2(\tau)}.$$

It is easy to see that $\sum_j \sigma_j^2(\tau)(x_j - \lambda(\tau)) = 0$. Hence

$$\boldsymbol{x}^t \boldsymbol{G}_{n\tau} \boldsymbol{x} = \sum_{i=1}^{n} \sigma_i^2(\tau) x_i (x_i - \lambda(\tau)) = \sum_{i=1}^{n} \sigma_i^2(\tau)(x_i - \lambda(\tau))^2.$$

By $\boldsymbol{x}^t\boldsymbol{e} = 0$ we have

$$\sum_{i=1}^{n} (x_i - \lambda(\tau))^2 \geq \sum_{i=1}^{n} x_i^2 = ||\boldsymbol{x}||^2 = 1.$$

This implies

$$\boldsymbol{x}^t \boldsymbol{G}_{n\tau} \boldsymbol{x} \geq \inf_{i,\tau} \sigma_i^2(\tau) \sum_{i=1}^{n} (x_i - \lambda(\tau))^2 \geq \inf_{i,\tau} \sigma_i^2(\tau).$$

Next we consider $||\boldsymbol{F}_{n\Gamma}^{+}||$. By $||\boldsymbol{F}_{n\Gamma}^{+}|| \leq 1/||\boldsymbol{F}_{n\Gamma}||$ we need a positive lower bound of $||\boldsymbol{F}_{n\Gamma}||$. This is immediate from

$$||\boldsymbol{F}_{n\Gamma}|| \geq ||\boldsymbol{G}_{n\Gamma}|| - ||\boldsymbol{F}_{n\Gamma} - \boldsymbol{G}_{n\Gamma}||$$

since $||\boldsymbol{F}_{n\Gamma} - \boldsymbol{G}_{n\Gamma}|| = O(n^{-1})$. $\square$

2.4 THEOREM. *Assume that item parameters and subject parameters are bounded. Then*

$$||\boldsymbol{\Sigma}_{n\Gamma} - \boldsymbol{H}^{+}_{n\Gamma}|| = O(n^{-3/2}).$$

Proof: We know from 2.2 that $||\boldsymbol{F}_{n\Gamma} - \boldsymbol{H}_{n\Gamma}|| = O(n^{-3/2})$. In view of Lemma 4.1 it is sufficient to show that the matrix norms $||\boldsymbol{\Sigma}_{n\Gamma}|| = ||\boldsymbol{F}^{+}_{n\Gamma}||$ and $||\boldsymbol{H}^{+}_{n\Gamma}||$ are bounded.

Boundedness of $||\boldsymbol{\Sigma}_{n\Gamma}|| = ||\boldsymbol{F}^{+}_{n\Gamma}||$ and $||\boldsymbol{G}^{+}_{n\Gamma}||$ have been shown in the proof of Theorem 2.3 .

By definition of $\boldsymbol{H}_{n\Gamma}$ in (9) und (10) we have $\boldsymbol{H}_{n\Gamma} = \boldsymbol{G}_{n\Gamma} + \boldsymbol{J}_{n\Gamma}$, where

$$\boldsymbol{J}_{n\Gamma} = \frac{1}{n^2} E\Big(\frac{1}{2\overline{\sigma}^4_n(\tau)} \boldsymbol{a}_n(\tau)\boldsymbol{a}^t_n(\tau)\Big), \text{ where } \boldsymbol{a}_n(\tau)_j := \sigma^2_j(\tau)\pi_{nj}(\tau),$$

(see (8)). From $||\boldsymbol{a}_n(\tau)\boldsymbol{a}^t_n(\tau)|| = ||\boldsymbol{a}_n(\tau)||^2 \leq C\, n$ it follows that $||\boldsymbol{J}_{n\Gamma}|| = O(n^{-1})$. This proves the assertion. $\square$

Our last result clarifies the structure of the approximating covariance matrix $\boldsymbol{G}^{+}_{n\Gamma}$. Let $\boldsymbol{P}_n = \boldsymbol{E}_n - \boldsymbol{e}\boldsymbol{e}^t/n$ the orthogonal projection to H_n and let $\boldsymbol{G}_{n\Gamma} = \boldsymbol{D}_{n\Gamma} - \boldsymbol{B}_{n\Gamma}$, thus defining $\boldsymbol{D}_{n\Gamma}$ as the leading diagonal matrix in (3) and $\boldsymbol{B}_{n\Gamma}$ the residual matrix in (3).

In the following theorem the matrix

$$\overline{\boldsymbol{B}}_{n\Gamma} := \boldsymbol{D}^{-1/2}_{n\Gamma} \boldsymbol{B}_{n\Gamma} \boldsymbol{D}^{-1/2}_{n\Gamma}$$

plays a crucial role.

2.5 THEOREM. *The approximating covariance matrix $\boldsymbol{G}^{+}_{n\Gamma}$ is given by*

$$\boldsymbol{G}^{+}_{n\Gamma} = (\boldsymbol{P}_n \boldsymbol{D}^{-1/2}_{n\Gamma})\Big(\boldsymbol{E}_n + \sum_{k=1}^{\infty} \overline{\boldsymbol{B}}^k_{n\Gamma}\Big)(\boldsymbol{D}^{-1/2}_{n\Gamma}\boldsymbol{P}_n)$$

Before we turn to the proof of Theorem 2.5 let us make some remarks on the assertion.

First, it is easy to see that in the case of a fixed subject parameter τ we have $\boldsymbol{P}_n \boldsymbol{D}^{-1}_{n\tau} \boldsymbol{B}_{n\tau} = \boldsymbol{O}$. Hence, in this case the infinite series vanishes and we have $\boldsymbol{G}^{+}_{n\tau} = \boldsymbol{P}_n \boldsymbol{D}^{-1}_{n\tau} \boldsymbol{P}_n$ which gives formula (5).

In general, the series does not vanish. We have to show that the series is convergent on that subspace of $\mathbb{R}^n$ which is the image of $\boldsymbol{D}^{-1/2}_{n\Gamma}\boldsymbol{P}_n$, i.e. on the subspace $V_n := \boldsymbol{D}^{-1/2}_{n\Gamma}(H_n)$. For this we will show that norm $||\overline{\boldsymbol{B}}_{n\Gamma}||_{V_n} < 1$. In typical cases the norm is actually rather small. Therefore, $\boldsymbol{G}_{n\Gamma}^{+} \sim \boldsymbol{P}_n \boldsymbol{D}_{n\Gamma}^{-1} \boldsymbol{P}_n$ provides a reasonable approximation in most cases (see formula (6)). If we additionally use the first term of the series then the approximation becomes satisfactory.

Proof: (of Theorem 2.5) For notational convenience let us omit the indices n and Γ. Note that for any matrix $\boldsymbol{A}$ with vanishing row- and column sums we have $\boldsymbol{PA} = \boldsymbol{AP} = \boldsymbol{A}$.

By definition of the MP-inverse we have $(\boldsymbol{D} - \boldsymbol{B})\boldsymbol{G}^{+} = \boldsymbol{P}$. This can be written as

$$\boldsymbol{D}^{1/2}(\boldsymbol{E} - \overline{\boldsymbol{B}})\boldsymbol{D}^{1/2}\boldsymbol{G}^{+} = \boldsymbol{P}$$

If we consider the matrices as linear operators on H then everything is invertible and we have

$$\boldsymbol{G}^{+} = \boldsymbol{D}^{-1/2}(\boldsymbol{E} - \overline{\boldsymbol{B}})^{-1}\boldsymbol{D}^{-1/2}$$

We want to show that on $V := \boldsymbol{D}^{-1/2}(H)$ we have

$$(\boldsymbol{E} - \overline{\boldsymbol{B}})^{-1} = \boldsymbol{E} + \sum_{k=1}^{\infty} \overline{\boldsymbol{B}}^{k}. \tag{12}$$

This implies the assertion since the extension of the operators from H to $\mathbb{R}^n$ is done by multiplication with the orthogonal projection $\boldsymbol{P}$.

For proving (12) we have to show that $||\overline{\boldsymbol{B}}||_V < 1$.

The matrix $\overline{\boldsymbol{B}}$ has entries

$$\frac{1}{\sqrt{E(\sigma_i^2)}\sqrt{E(\sigma_j^2)}} E\Big(\frac{\sigma_i^2\sigma_j^2}{\sum_k \sigma_k^2}\Big)$$

For any $\boldsymbol{x} \in \mathbb{R}^n$ we get

$$\boldsymbol{x}^t\overline{\boldsymbol{B}}\boldsymbol{x} = \sum_{i,j} \frac{1}{\sqrt{E(\sigma_i^2)}\sqrt{E(\sigma_j^2)}} E\Big(\frac{\sigma_i^2 x_i \sigma_j^2 x_j}{\sum_k \sigma_k^2}\Big) = E\Big(\frac{[\sum_i \sigma_i^2 x_i/\sqrt{E(\sigma_i^2)}]^2}{\sum_k \sigma_k^2}\Big)$$

The Cauchy-Schwarz inequality implies that

$$\Big(\sum_i \sigma_i^2 x_i/\sqrt{E(\sigma_i^2)}\Big)^2 \leq \Big(\sum_i \sigma_i^2\Big)\Big(\sum_i \sigma_i^2 x_i^2/E(\sigma_i^2)\Big)$$

where $<$ is true whenever the vector (σ_i) is not proportional to the vector $(\sigma_i x_i/\sqrt{E(\sigma_i^2)})$. This is the case iff $\boldsymbol{x}$ is not proportional to the vector $(\sqrt{E(\sigma_i^2)})$. Thus, for all vectors $\boldsymbol{x}$ which are not proportional to $(\sqrt{E(\sigma_i^2)})$ we have

$$\boldsymbol{x}^t\overline{\boldsymbol{B}}\boldsymbol{x} < E\Big(\sum_i \sigma_i^2 x_i^2/E(\sigma_i^2)\Big) = ||\boldsymbol{x}||^2 \tag{13}$$

Note, that $(\sqrt{E(\sigma_i^2)})$ is an eigenvector of $\overline{\boldsymbol{B}}$ for the eigenvalue 1, and that V is the orthogonal complement of this eigenvector. It follows, that for all eigenvectors of $\overline{\boldsymbol{B}}$ in V the inequality (13) is true.

Recall that

$$||\overline{\boldsymbol{B}}||_V = \sup_{\boldsymbol{x}\in V, ||x||=1} \boldsymbol{x}^t \overline{\boldsymbol{B}} \boldsymbol{x}.$$

The set $\{\boldsymbol{x} \in V, ||\boldsymbol{x}|| = 1\}$, where the supremum is taken, is compact. Hence, there exists $\boldsymbol{x}_0 \in V$ where the supremum is attained. From the preceding we get $\boldsymbol{x}_0^t \overline{\boldsymbol{B}} \boldsymbol{x}_0 < 1$. This proves the assertion $||\overline{\boldsymbol{B}}||_V < 1$. $\square$

3 Review on cml-estimation for the Rasch model

The following is well-known in the field of item-response analysis. For the reader's convenience we present some details where references are difficult to provide.

Let X be an item-response variable with values 0 and 1. If θ is the logit parameter then $P(X = x) = e^{\theta x}/(1 + e^{\theta})$. It we present n independent items $\boldsymbol{X} = (X_1, \ldots, X_n)$ to a subject then

$$P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{\sum_{j=1}^n \theta_j x_j}}{\prod_{j=1}^n (1 + e^{\theta_j})}.$$

By

$$\gamma(\boldsymbol{\theta}, s) := \sum \left\{ e^{\sum_{j=1}^n \theta_j y_j} : \boldsymbol{y} \in \{0,1\}^n, \sum_{j=1}^n y_j = s \right\}$$

we denote the elementary symmetric functions of e^{θ_i}, $i = 1, \ldots, n$. Then the conditional probabilities given the sum $S = \sum_{j=1}^n X_j$ are

$$P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} | S = s) = \frac{e^{\sum_{j=1}^n \theta_j x_j}}{\gamma(\boldsymbol{\theta}, s)}, \; s = \sum_{j=1}^n x_i.$$

The numerical complexity of exact calculations mentioned before is due to the evaluation of the functions $\gamma(\boldsymbol{\theta}, s)$.

We assume that the Rasch model holds, i.e. $\boldsymbol{\theta} = \boldsymbol{\beta} + \tau$. Let H_n be the set of all item parameters $\boldsymbol{\beta} \in \mathbb{R}^n$ satisfying $\sum_{j=1}^n \beta_j = 0$.

The conditional probabilities given the sum

$$P_{\boldsymbol{\beta},\tau}(\boldsymbol{X} = \boldsymbol{x} | S = s) = \frac{e^{\sum_{j=1}^n \beta_j x_j}}{\gamma(\boldsymbol{\beta}, s)} =: q(\boldsymbol{x}, \boldsymbol{\beta})$$

do not depend on the subject parameter. Cml-estimates are based on the conditional loglikelihood

$$\ell(\boldsymbol{X},\boldsymbol{\beta}) = \log q(\boldsymbol{X},\boldsymbol{\beta}) = \sum_{j=1}^{n} \beta_j X_j - \log \gamma(\boldsymbol{\beta}, S). \tag{14}$$

This conditional loglikelihood shares with any loglikelihood the Kullback-Leibler property

$$E_{\boldsymbol{\beta}}(\ell(.,\boldsymbol{\beta})|S) > E_{\boldsymbol{\beta}}(\ell(.,\boldsymbol{b})|S),\ \boldsymbol{b} \in H_n,\ \boldsymbol{b} \neq \boldsymbol{\beta}. \tag{15}$$

The conditional loglikelihood is therefore a contrast function for the estimation of the item parameter $\boldsymbol{\beta}$, i.e.

$$E_{\boldsymbol{\beta},\Gamma}(\ell(.,\boldsymbol{\beta})) > E_{\boldsymbol{\beta},\Gamma}(\ell(.,\boldsymbol{b})),\ \boldsymbol{b} \in H_n,\ \boldsymbol{b} \neq \boldsymbol{\beta}.$$

3.1 LEMMA. *The gradient of the conditional loglikelihood satisfies*

$$\ell'_j(\boldsymbol{X},\boldsymbol{\beta}) := \frac{\partial}{\partial \beta_j} \ell(\boldsymbol{X},\boldsymbol{\beta}) = X_j - E_{\boldsymbol{\beta}}(X_j|S).$$

Proof: Note that (15) is an extremal property which is satisfied by every $\boldsymbol{\beta} \in H_n$. The inequality (15) is even valid for all $\boldsymbol{b} \in \mathbb{R}^n$, since the conditional loglikelihood is invariant w.r.t. shifts of the subject parameter. Therefore the extremal property is global and it follows that $E_{\boldsymbol{\beta}}(\ell'_j(\boldsymbol{X},\boldsymbol{\beta})|S) = 0$ for every $\boldsymbol{\beta} \in H_n$. From (14) we obtain that $\ell'_j(\boldsymbol{X},\boldsymbol{\beta}) = X_j - f_j(S)$. Hence $f_j(S) = E_{\boldsymbol{\beta}}(X_j|S)$. □

The item parameter $\boldsymbol{\beta}$ can be estimated from independent observations $\boldsymbol{X}_\nu$, $\nu = 1, \ldots, N$ by maximization of

$$\sum_{\nu=1}^{N} \ell(\boldsymbol{X}_\nu, \hat{\boldsymbol{\beta}}_N) = \max_{\boldsymbol{b} \in H_n} \sum_{\nu=1}^{N} \ell(\boldsymbol{X}_\nu, \boldsymbol{b}).$$

This is cml-estimation. From Anderson, [1], it is known that under $P_{\boldsymbol{\beta},\Gamma}$ the cml-estimates $(\hat{\boldsymbol{\beta}}_N)$ for $N \to \infty$ are consistent and asymptotically normally distributed. By $\sum_{i=1}^{n} \beta_i = 0$ the asymptotic covariance matrix $\boldsymbol{\Sigma}_{n\Gamma}$ is of rank $n-1$ since all row and column sums vanish.

It is not difficult to write $\boldsymbol{\Sigma}_{n\Gamma}$ as a mathematical expression. We have to start with the Fisher-information matrix

$$\boldsymbol{F}_{n\Gamma} = E_{\boldsymbol{\beta},\Gamma}(\ell'(.,\boldsymbol{\beta})\ell'(.,\boldsymbol{\beta})^t) \tag{16}$$

of the cml-estimates. The Fisher-information matrix is of rank $n-1$, too, since all row and column sums vanish (by the invariance property of the conditional probabilities).

The following assertion is then a basic fact for the present paper.

3.2 LEMMA. *The covariance matrix $\Sigma_{n\Gamma}$ is identical to the Moore-Penrose inverse $\boldsymbol{F}_{n\Gamma}^{+}$ of $\boldsymbol{F}_{n\Gamma}$.*

Proof: From Anderson, [1], we obtain consistency of the sequence of cml-estimates. By familiar arguments this leads to the expansion

$$E_{\boldsymbol{\beta},\Gamma}(\ell''(\boldsymbol{X},\boldsymbol{\beta}))\cdot\sqrt{N}(\hat{\boldsymbol{\beta}}_N-\boldsymbol{\beta}) = -\frac{1}{\sqrt{N}}\sum_{\nu=1}^{N}\ell'(\boldsymbol{X}_\nu,\boldsymbol{\beta})+\boldsymbol{R}_N$$

where $\boldsymbol{R}_N\to\boldsymbol{o}$ in $P_{\boldsymbol{\beta},\Gamma}$-probability. Differentiating the extremal property of the loglikelihoods twice gives

$$E_{\boldsymbol{\beta},\Gamma}(\ell''(\boldsymbol{X},\boldsymbol{\beta})) = -E_{\boldsymbol{\beta},\Gamma}(\ell'(.,\boldsymbol{\beta})\ell'(.,\boldsymbol{\beta})^t) = -\boldsymbol{F}_{n\Gamma}.$$

Thus we obtain

$$\boldsymbol{F}_{n\Gamma}\cdot\sqrt{N}(\hat{\boldsymbol{\beta}}_N-\boldsymbol{\beta}) = \frac{1}{\sqrt{N}}\sum_{\nu=1}^{N}\ell'(\boldsymbol{X}_\nu,\boldsymbol{\beta})+\boldsymbol{R}_N.$$

Since both $\hat{\boldsymbol{\beta}}_N-\boldsymbol{\beta}$ and $\ell'(\boldsymbol{X}_\nu,\boldsymbol{\beta})$ are contained in H_n multiplication by $\boldsymbol{F}_{n\Gamma}^{+}$ yields

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_N-\boldsymbol{\beta}) = \boldsymbol{F}_{n\Gamma}^{+}\cdot\frac{1}{\sqrt{N}}\sum_{\nu=1}^{N}\ell'(\boldsymbol{X}_\nu,\boldsymbol{\beta})+\boldsymbol{F}_{n\Gamma}^{+}\cdot\boldsymbol{R}_N$$

Since $\boldsymbol{F}_{n\Gamma}$ is the covariance matrix of $\ell'(\boldsymbol{X}_\nu,\boldsymbol{\beta})$ the assertion follows. □

4 Review on matrices

In this section we collect some well-known facts on matrix norm and the Moore-Penrose pseudo-inverse (MP-inverse).

The matrix norm of a positive semidefinite symmetric matrix is defined by

$$||\boldsymbol{A}|| = \sup\{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} : ||x|| = 1\}. \tag{17}$$

For this matrix norm we have

$$\max_{ij}|\boldsymbol{A}_{ij}| \le ||\boldsymbol{A}|| \le n\max_{ij}|\boldsymbol{A}_{ij}|. \tag{18}$$

Apart from usual norm properties the matrix norm satisfies

$$||\boldsymbol{AB}|| \leq ||\boldsymbol{A}||\, ||\boldsymbol{B}||. \tag{19}$$

and by convexity of the norm

$$||E(\boldsymbol{A})|| \leq E(||\boldsymbol{A}||). \tag{20}$$

The matrix norm $||\boldsymbol{A}||$ is equal to the maximal eigenvalue of $\boldsymbol{A}$. Let $\boldsymbol{E}$ be the identiy matrix and $\boldsymbol{a}$ any vector. Then $\boldsymbol{Q} = \boldsymbol{E} - \boldsymbol{aa}^t/||\boldsymbol{a}||^2$ is the orthogonal projection on the orthoronal complement of $\boldsymbol{a}$ and $||\boldsymbol{Q}|| = 1$.

Let $\boldsymbol{e} = (1, 1, \ldots, 1)^t \in \mathbb{R}^n$ and let $\boldsymbol{A}$ be a positive semidefinite symmetric $n \times n$-Matrix such that $\boldsymbol{Ae} = \boldsymbol{o}$. This means that $\boldsymbol{A}$ has vanishing row and column sums. Assume that $\boldsymbol{A}$ is of rank $n-1$. Let H_n be the orthogonal complement of $\boldsymbol{e}$, i.e. the set of all vectors with mean value zero. Denote by $\boldsymbol{P}$ the orthogonal projection onto H_n, i.e. $\boldsymbol{P} = \boldsymbol{E} - \frac{1}{n}\boldsymbol{ee}^t$. This projection $\boldsymbol{P}$ satisfies $\boldsymbol{Pe} = \boldsymbol{o}$, and $\boldsymbol{Px} = \boldsymbol{x}$ whenever $\boldsymbol{x} \in H_n$.

The Moore-Penrose inverse $\boldsymbol{A}^+$ is defined by $\boldsymbol{AA}^+ = \boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{P}$. It satisfies $||\boldsymbol{A}^+|| = 1/||\boldsymbol{A}||$.

4.1 LEMMA. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be positive semidefinite symmetric matrices. Then*

$$||\boldsymbol{A}^+ - \boldsymbol{B}^+|| \leq ||\boldsymbol{A}^+||\, ||\boldsymbol{B} - \boldsymbol{A}||\, ||\boldsymbol{B}^+||$$

Proof: Both for $\boldsymbol{x} \in H_n$ and for $\boldsymbol{x} = \boldsymbol{e}$ we have

$$\boldsymbol{A}^+\boldsymbol{x} - \boldsymbol{B}^+\boldsymbol{x} = \boldsymbol{A}^+(\boldsymbol{BB}^+)\boldsymbol{x} - (\boldsymbol{A}^+\boldsymbol{A})\boldsymbol{B}^+\boldsymbol{x}.$$

This implies

$$\boldsymbol{A}^+ - \boldsymbol{B}^+ = \boldsymbol{A}^+\boldsymbol{BB}^+ - \boldsymbol{A}^+\boldsymbol{AB}^+ = \boldsymbol{A}^+(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{B}^+.$$

The assertion follows from (19). □

4.2 LEMMA. *Let $\boldsymbol{A}$ be a random matrix with positive semidefinite symmetric values and such that $\boldsymbol{Ae} = \boldsymbol{o}$. Then $E(\inf_{||x||=1, \boldsymbol{x} \in H_n} \boldsymbol{x}^t\boldsymbol{Ax})$ is a lower bound for the positive eigenvalues of $E_\Gamma(\boldsymbol{A})$.*

Proof: Since $\boldsymbol{Ae} = \boldsymbol{o}$ wehave

$$||E(\boldsymbol{A})|| = \sup_{||x||=1, \boldsymbol{x} \in H_n} \boldsymbol{x}^t E(\boldsymbol{A})\boldsymbol{x} = \sup_{||x||=1, \boldsymbol{x} \in H_n} E(\boldsymbol{x}^t\boldsymbol{Ax}) \geq E(\inf_{||x||=1, \boldsymbol{x} \in H_n} \boldsymbol{x}^t\boldsymbol{Ax}).$$

□

References

[1] E. Andersen. Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 32:283–301, 1970.

[2] R. N. Bhattacharya and R. Ranga Rao. *Normal approximation and asymptotic expansions.* Wiley Series in Probability and Mathematical Statistics. New York etc.: John Wiley&Sons, Inc. XIV, 274 p., 1976.

[3] G. H. (ed.) Fischer and I. W. (ed.) Molenaar. *Rasch models. Foundations, recent developments, and applications*. New York, Springer-Verlag, 1995.

[4] J. Pfanzagl. *Estimation in semiparametric models, Some recent developments*, volume 63 of *Lecture Notes in Statistics*. Springer Verlag, New York, 1990.

[5] H. Strasser. Asymptotic expansions for conditional moments of Bernoulli trials. Technical report, Institute of Statistics and Mathematics, WU, 2011.

[6] H. Strasser. Numerical studies for the rasch model with many item parameters. Technical report, https://www.dropbox.com/s/x7iecn6dxwp88sp/download001.pdf, 2011.