Measurement and Research Department Reports 97-6

On the Loss of Information in Conditional Maximum Likelihood Estimation of Item Parameters

T.J.H.M. Eggen

Cito

Measurement and Research Department Reports 97-6

On the Loss of Information in Conditional Maximum Likelihood Estimation of Item Parameters

T.J.H.M. Eggen

Cito
Arnhem, 1997

Abstract

In item response models of the Rasch type (Fischer & Molenaar, 1995), item parameters are often estimated by the conditional maximum likelihood (CML) method. This paper addresses the loss of information in CML estimation by using the information concept of F-information (Liang, 1983). This concept makes it possible to specify the conditions for no loss of information and to define a quantification of information loss. For the dichotomous Rasch model, the derivations will be given in detail to show the use of the F-information concept for making efficiency comparisons for different estimation methods. It is shown that by using CML for item parameter estimation, some information is almost always lost. But compared to JML (joint maximum likelihood) as well as to MML (marginal maximum likelihood) the loss is very small. The reported efficiency of CML to JML and to MML in several comparisons is always larger than 93%, and in tests with a length of 20 items or more, larger than 99%.

Keywords: Conditional maximum likelihood, Loss of information, F-information, Rasch models, item parameter estimation

Introduction

The Rasch model for measuring a latent trait θ with dichotomously scored items, with the responses $X_{vi} = x_{vi}$ (0 or 1), persons $v = 1,\ldots,n$ and items $i = 1,\ldots,k$ is given by

$$p(x;\omega) = \prod_{v=1}^{n}\prod_{i=1}^{k} \mathrm{P}(X_{vi} = x_{vi};\beta_i,\theta_v), \tag{1}$$

$$\text{with } \mathrm{P}(X_{vi} = x_{vi};\beta_i,\theta_v) = \frac{\exp x_{vi}(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}, \tag{2}$$

and $\omega^T = (\beta^T, \theta^T)$, $\beta^T = (\beta_1,\ldots,\beta_k)$ is the vector item parameter (difficulty) and $\theta^T = (\theta_1,\ldots,\theta_n)$ the vector person parameter (ability).

In Fischer and Molenaar (1995) the foundations, the main results and an overview of recent developments and extensions of the Rasch model are given. In these type of item response theory (IRT) models conditional maximum likelihood estimation (CML) is a popular method for estimating item parameters. Although properties as consistency, asymptotic normality and sampling independence of CML estimators are well established, the justification of CML estimation with respect to the efficiency of the estimators is not clear. This topic is addressed in this paper.

Estimation of item parameters in the Rasch model

For the dichotomous Rasch model (1), three likelihood based methods for item parameter estimation are available (Molenaar, 1995).

In the first method, joint maximum likelihood (JML), the item parameters are estimated by maximizing (1) with respect to ω, given the data x. It is well known that estimating the item parameters by JML leads to inconsistent estimators (Andersen, 1973). This is caused by the fact that a limited number of interest (items) parameters (β) are to be estimated in the presence of many nuisance (ability) parameters (θ). Eliminating the nuisance parameters gives the solution for this problem (Basu, 1977). In IRT modeling this elimination is accomplished by the marginal or the conditional maximum likelihood method.

In the marginal maximum likelihood (MML) method it is assumed that the abilities θ_v in (1) constitute a random sample from an ability distribution $h(\theta;\xi)$, with ξ the

parameters of the ability distribution. In the Rasch model the joint probability of the item responses can then be written as

$$p_m(x;\beta,\xi) = \int p(x \mid \theta;\beta,\xi)\, h(\theta;\xi)\, d\theta =$$

$$\prod_{v=1}^{n} \int_{-\infty}^{\infty} \prod_{i=1}^{k} \mathrm{P}(X_{vi} = x_{vi} \mid \theta_v;\beta_i)\, h(\theta_v;\xi)\, d\theta_v = L_m(\beta,\xi;x). \qquad (3)$$

This so called marginal likelihood function is maximized with respect to β and ξ in order to get estimates of the item parameters, and, for instance, in the case of a normal ability distribution, two distribution parameters. In the MML method, the nuisance parameters are eliminated by integrating them out. In (3) $\mathrm{P}(X_{vi} = x_{vi} \mid \theta_v;\beta_i)$ is given by (2).

In Rasch type of IRT models, the CML method is an alternative solution to the inconsistency problem. If in the model there exist sufficient statistics for the nuisance parameters, it can be separated in a conditional part which is only dependent on the interest parameters and a part which models the sufficient statistics.

In the dichotomous Rasch model the sum score on the items $T_v = \sum_{i=1}^{k} X_{vi}$ is sufficient for θ_v, $v = 1,\ldots,n$, so we can rewrite (1) as

$$p(x;\omega) = \prod_{v=1}^{n} f(x_v \mid t_v;\beta) \cdot \prod_{v=1}^{n} g(t_v;\beta,\theta), \qquad (4)$$

with $X_v = (X_{v1},\ldots,X_{vk})$ the response vector of person $v = 1,\ldots,n$.
Maximizing, with respect to β, the conditional likelihood

$$L_c(\beta;x \mid t) = \prod_{v=1}^{n} f(x_v \mid t_v;\beta) \qquad (5)$$

leads to consistent estimators of the item parameters (Andersen, 1973).

Information and efficiency

In likelihood inference, the information concept plays an important role. In the case of several parameters, the Fisher information matrix is used in the evaluation of the quality of estimators. It is well known (Rao, 1973) that in the class of unbiased estimators the variance of any estimator is bounded by the inverse of the Fisher information matrix. Estimators reaching this bound, the Cramér-Rao lower bound, are called efficient. And the efficiency of any other estimator is always expressed in relation to this lower bound.

It has been shown (Andersen, 1973) that the CML estimators for the item parameters in the Rasch model are under mild conditions asymptotically efficient; the same is true for the MML estimators of the item parameters. However, these asymptotic results do not imply that there is no loss of information if CML estimation is applied. The point is that by using the conditional likelihood (5) for estimating the item parameters, the second part of the full likelihood (4), which is the marginal distribution of T, is neglected. And this distribution possibly contains some information on the item parameters. In MML estimation, where no information is discarded, on the other hand, a correct specification of the ability distribution is needed and if this distribution is not the correct one, the resulting estimates of the item parameters can be useless. Furthermore, in MML the loss in efficiency of the item parameter estimation due to the joint estimation of them with the parameters of the ability distribution is not clear. In psychometric textbooks, the efficiency problem of the item parameter estimators is mostly discarded. In some recent work, however, Engelen (1989) and Zwinderman (1991), point out that there must be some loss of information on the item parameters without giving a quantification of the loss. Assuming that the ability distribution parameters are known, Engelen (1989) has shown that the loss in CML estimation compared to MML is small.

In this paper a general treatment is given of an information concept, called F-information, which makes it possible to define a clear measure of information loss in using CML estimation. The conditions for no loss of information in separable models, like the Rasch model, will be given. The theory will be clarified with some examples. In particular, attention will be paid to the dichotomous Rasch model as specified in (1) and (2) and to the Rasch Poisson Counts model for misreadings (Rasch, 1960). For the dichotomous Rasch model the derivations will be given in detail to show the use of the general information concept for making efficiency comparisons for different estimation methods. Results will be given of the comparison of CML to JML, and of the comparison of CML to MML in case of a normal ability distribution.

Notation and Terminology

It should be noted that when the terms one- and two-parameter distribution are used in this paper, the parameters can be vectors. By a two-parameter distribution is meant that the parameters of the distribution ω can be partitioned in an interest parameter ψ and a nuisance parameter τ. In general, a scalar or vector random variable X is considered

with density (or probability in the discrete case) $p(x;\omega)$, with $\omega \in \Omega$. The parameter of interest $\psi = \psi(\omega)$ has domain Ψ and the nuisance parameter τ, the complementary part to ψ of ω, has domain T. It is assumed that $\Omega = \Psi \times \mathrm{T}$. Let $T = T(X)$ be a (vector) statistic, $f(. \mid .;.)$ a conditional density or probability of the data given a statistic and $g(.;.)$ the (marginal) density (or probability) of the statistic. Then in general the probability of the data can be factored in

$$p(x;\omega) = f(x \mid t;\omega).g(t;\omega) \tag{6}$$

In case $T(X)$ is sufficient for τ then (6) can be written as

$$p(x;\omega) = f(x \mid t;\psi).g(t;\omega) \tag{7}$$

The model is then called separable and in CML estimation the inference is based on the first part of the factorization in (7) which is only dependent on the parameter of interest ψ, while the other part is neglected.

The F-Information: Definition and Basic Properties

F-information is a generalization of the Fisher information. In an m-dimensional one-parameter distribution $p(x:\omega)$, $\omega^T = (\omega_1, \ldots, \omega_m)^T$, the *Fisher information matrix* is defined as the $m \times m$ matrix

$$\mathrm{I}_p(\omega) = \mathscr{E}\left[S_{p;\omega}.S_{p;\omega}^T\right], \tag{8}$$

in which

$$S_{p;\omega} = S_{p;\omega}(X) := \frac{\partial \ln p(X;\omega)}{\partial \omega}, \tag{9}$$

is the efficient score statistic (a $m \times 1$-vector) of a distribution p with respect to ω.

The role of this Fisher information matrix in estimation problems will be illustrated by some examples. The Cramér-Rao lower bound for the covariance matrix of any estimator in the class of unbiased estimators is given by the inverse (if it exists) of the Fisher information matrix: if V is the covariance matrix of an unbiased estimator for ω, then $V - \mathrm{I}_p^{-1}(\omega)$ is non-negative definite (nnd). Furthermore, it is known that if

there exists an efficient estimator (reaching the lower bound), the maximum likelihood estimation (MLE) procedure will produce it. MLE's need not to be unbiased, of course, but in broad classes of models and under mild regularity conditions, these estimators are asymptotically normally distributed with covariance matrix the inverse of the Fisher information matrix. In practice, the so called *observed* Fisher information matrix is used, to contrast it with the *expected* Fisher information defined in (8) and (9). The observed Fisher information matrix is obtained by disregarding the expectation in (8) and evaluating it at the MLE $\omega = \hat{\omega}$. The inverse of the observed information matrix is used as an estimator of the covariance matrix of $\hat{\omega}$. Its diagonal elements are used for obtaining standard errors of the estimates and for constructing confidence intervals.

For two-parameter distributions, the concept of F-information, originating from Efron (1977), was defined by Liang (1983) for scalar parameters and was generalized to vector parameters by Bhapkar (1989).

Definition 1.
In a two-parameter distribution $p(x:\omega)$, with $\omega^T = (\psi^T, \tau^T)$, $\psi^T = (\psi_1, \ldots, \psi_k)$ being the interest parameter and $\tau^T = (\tau_1, \ldots, \tau_n)$ the nuisance parameter, the *F-information* for ψ in p is given by

$$\mathrm{I}_p(\psi;\omega) = \underset{N}{\mathrm{Min}} \ \mathscr{E}\left[m(N).m(N)^T\right], \text{ with } m(N) = S_{p;\psi} - N^T.S_{p;\tau}. \tag{10}$$

In (10), N is a $n \times k$-matrix with constants and Min is the minimal matrix M_1 in the class $\mathbb{M}$ of non-negative definite matrices which can be written as $\mathscr{E}[m(N).m(N)^T]$ for some N, that is for any $M \in \mathbb{M}$: $M - M_1$ is nnd.

Under general regularity conditions, which entail the justification of interchanging the order of differentiation and integration (taking expectations), the F-information has the following properties:

1. F-information is a generalization of the Fisher information.
   Partition the parameter vector in itself and an empty part $\omega^T = (\omega^T, \varnothing)$ and observe that

$$\mathrm{I}_p(\omega;\omega) = \underset{N}{\mathrm{Min}} \ \mathscr{E}\left[\left(S_{p;\omega} - N^T S_{p;\varnothing}\right).\left(S_{p;\omega} - N^T S_{p;\varnothing}\right)^T\right] = \mathscr{E}\left[S_{p;\omega}.S_{p;\omega}^T\right] = \mathrm{I}_p(\omega)$$

2. F-information can be defined in terms of Fisher information.
It can be shown (Bhapkar, 1989) that if the Fisher information matrix (8) is positive definite and is rewritten as a partitioned matrix:

$$I_p(\omega) = \mathcal{E}\left[S_{p;\omega}.S_{p;\omega}^T\right] = \mathcal{E}\begin{bmatrix} S_{p;\psi}S_{p;\psi}^T & S_{p;\psi}S_{p;\tau}^T \\ S_{p;\tau}S_{p;\psi}^T & S_{p;\tau}S_{p;\tau}^T \end{bmatrix} =: \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \tag{11}$$

and its inverse as

$$I_p^{-1}(\omega) = \begin{bmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{bmatrix},$$

then the F-information matrix for ψ in p is equal to

$$I_p(\psi;\omega) = I_{11} - I_{12}I_{22}^{-1}I_{21} = (I^{11})^{-1}, \tag{12}$$

which is the inverse of the upper left submatrix of the inverse of the Fisher information matrix.
Suppose both ψ and τ are scalar, then $N = N^T$ and all efficient scores are scalar. The F-information is found by minimizing $\mathcal{E}(S_{p,\psi}^2 - 2NS_{p,\psi}S_{p,\tau} + N^2S_{p,\tau}^2)$ with respect to N. After differentiation, the minimum is easily seen to be reached at $N = \mathcal{E}(S_{p,\psi}\,S_{p,\tau}).(\mathcal{E}(S_{p,\tau})^2)^{-1}$, and the F-information is given by (12). Generalizing to the multidimensional case, the same result follows for: $N^T = \mathcal{E}(S_{p,\psi}S_{p,\tau}^T).\,\mathcal{E}(S_{p,\tau}S_{p,\tau}^T)^{-1}$.

3. F-information has a clear geometrical interpretation.
To enable a geometrical interpretation of the F-information observe: $I_p(\psi;\omega)$ is the minimal matrix in the class $\mathbb{M}$ of non-negative definite $k \times k$ matrices: $\forall K \in \mathbb{M} : K - I_p(\psi;\omega) \in \mathbb{M}$. Consider a real valued function U from $\mathbb{M} \to \mathbb{R}$, so from the class of nnd matrices to the real line. Suppose this function satisfies:

a. $U(K) = 0$ if and only if $K = \mathbf{0}$ and (13a)

b. for every $K_1, K_2 \in \mathbb{M}$ with $K_1 - K_2 \in \mathbb{M}$ implies $U(K_1) \geq U(K_2)$. (13b)

Then it follows from (13b) that for every $K \in \mathrm{M}$, by definition $\mathrm{I}_p(\psi;\omega) \in \mathrm{M}$ and $K - \mathrm{I}_p(\psi;\omega) \in \mathrm{M}$, which implies $U(K) \geq U(\mathrm{I}_p(\psi;\omega))$.
So it can be concluded that finding the F-information is equivalent to finding the minimum of a function U. A matrix function which satisfies (13a) and (13b) is tr(.), the trace function. So minimizing $\mathscr{E}[S_{p;\psi} - N^T S_{p;\tau}).(S_{p;\psi} - N^T S_{p;\tau})^T]$, see (10), is equivalent to minimizing the trace of this matrix and then from standard matrix algebra it follows

$$\min_N \mathrm{tr}\left\{\mathscr{E}\left[\left(S_{p;\psi} - N^T S_{p;\tau}\right).\left(S_{p;\psi} - N^T S_{p;\tau}\right)^T\right]\right\} =$$

$$\min_N \mathrm{tr}\left\{\mathscr{E}\left[\left(S_{p;\psi} - N^T S_{p;\tau}\right)^T.\left(S_{p;\psi} - N^T S_{p;\tau}\right)\right]\right\} = \min_N \mathscr{E}\left[\left(S_{p;\psi} - N^T S_{p;\tau}\right)^T.\left(S_{p;\psi} - N^T S_{p;\tau}\right)\right]$$

The last expression makes it possible to give the F-information the same geometric interpretation as a least squares solution in a regression problem. As can be seen in Figure 1, the F-information follows from projecting the efficient score of the interest parameter onto the space of the efficient score of the nuisance parameter.

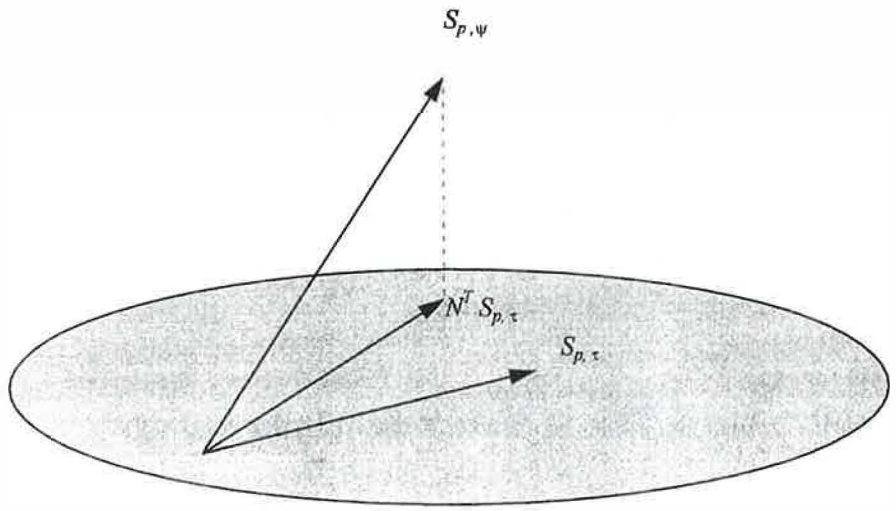

Figure 1
Geometrical interpretation F-information

4. F-information is the Cramér-Rao bound.

   For a finite number of parameters, $I_p(\psi;\omega)$ is the Cramér-Rao lower bound in the class of unbiased estimators of ψ, that is, for every covariance matrix V of unbiased estimators of ψ $V - (I_p(\psi;\omega))^{-1}$ is non-negative definite (Rao, 1973, p. 326). So when an estimator of ψ in a two-parameter distribution $p(x;(\psi,\tau))$ is evaluated with respect to reaching this lower bound, it is seen that all the relevant information on ψ in this distribution is included in the F-information.

5. A scalar measure of information.

   The F-information for a parameter ψ is in general a $k \times k$-matrix. And comparing F-information matrices is not straightforward. However, following Bhapkar (1989, p. 147), a scalar measure of information can be defined by a function on the F-information matrix satisfying (13a) and (13b). Possible choices could be the largest eigenvalue of the F-information, the trace of the F-information, and the determinant of the F-information in case the matrix is non-singular. In this paper only the trace of the F-information matrix, $\mathrm{tr}(I_p(\psi;\omega))$, will be used as a scalar measure of information on ψ in p.

6. F-information in case Fisher information is not positive definite.

   Bhapkar (1991) remarks that even in case the Fisher information matrix is not positive definite the F-information exists, although (12) cannot be used directly. In general, it yields that for any $n \times k$ matrix B which is a solution of $I_{22}\, B = I_{21}$ the F-information is given by $I_p(\psi;\omega) = I_{11} - B^T I_{22} B$, which is the same for any B.

   Note that in the special case that the submatrix I_{22} of $I_p(\omega)$ is positive definite, a solution is given by $B = I_{22}^{-1}\, I_{21}$, and it follows that

$$I_p(\psi;\omega) = I_{11} - I_{21}^T \left(I_{22}^{-1}\right)^T I_{22}\, I_{22}^{-1}\, I_{21} = I_{11} - I_{12}\, I_{22}^{-1}\, I_{21}, \qquad (14)$$

   which gives the same expression as (12).

F-Information in Separable Models

The F-information can be used to quantify the loss of information in estimating the interest parameters ψ in case the original two-parameter model can be factored as in (7), $p(x;\omega) = f(x \mid t;\psi).g(t;\omega)$, by just considering the conditional distribution $f(x \mid t;\psi)$. Moreover, conditions can be specified for no loss of information using this method.

In these separable models, some properties of the F-information will be shown to hold. Throughout it is assumed, that regularity conditions are met which guarantee the existence of the Fisher information and allow interchanging the order of differentiation and integration of the logarithm of the model.

Properties of the efficient score statistics

Three basic properties of the efficient score statistics, involved in the definition of F-information, are given first.

Property 1

$$\mathcal{E}\, S_{p;\psi} = \mathcal{E}\, S_{p;\tau} = 0.$$

Property 2

a. $\mathrm{COV}\left(S_{p;\psi}, S_{p;\psi}^T\right) = \mathcal{E}\left(S_{p;\psi} S_{p;\psi}^T\right) = -\mathcal{E}\, S_{p;\psi,\psi^T} := -\mathcal{E}\left[\dfrac{\partial^2 \log p}{\partial\psi\, \partial\psi^T}\right]$; and

b. $\mathrm{COV}\left(S_{p;\tau}, S_{p;\tau}^T\right) = \mathcal{E}\left(S_{p;\tau} S_{p;\tau}^T\right) = -\mathcal{E}\, S_{p;\tau,\tau^T}$.

Property 3

a. $\mathrm{COV}\left(S_{p;\psi}, S_{p;\tau}^T\right) = \mathcal{E}\left(S_{p;\psi} S_{p;\tau}^T\right) = -\mathcal{E}\, S_{p;\psi,\tau^T} := -\mathcal{E}\left[\dfrac{\partial^2 \log p}{\partial\psi\, \partial\tau^T}\right]$; and

b. $\mathrm{COV}\left(S_{p;\tau}, S_{p;\psi}^T\right) = \mathcal{E}\left(S_{p;\tau} S_{p;\psi}^T\right) = -\mathcal{E}\, S_{p;\tau,\psi^T}$.

These three properties are valid in each two-parameter distribution. So when a decomposition as in (6) or (7) is considered, the properties also apply to the efficient score statistics of $f(x \mid t;\omega)$ and $g(t;\omega)$. Then the following properties, relating the efficient score statistics of p, f and g, are easily deduced.

Property 4

a. $\mathscr{E}\left(S_{p;\psi} \mid T = t\right) = S_{g;\psi}$; and

b. $\mathscr{E}\left(S_{p;\tau} \mid T = t\right) = S_{g;\tau}$.

Proof 4a

$$\mathscr{E}\left(S_{p;\psi} \mid T=t\right) = \int_{\{x:t(x)=t\}} \frac{\partial \log p(x;\omega)}{\partial \psi} . f(x \mid t;\omega)\, dx = \int_{\{x:t(x)=t\}} \frac{\partial p(x;\omega)/\partial \psi}{p(x;\omega)} \cdot \frac{p(x;\omega)}{g(t;\omega)}\, dx =$$

$$\int_{\{x:t(x)=t\}} \frac{\partial p(x;\omega)/\partial \psi}{g(t;\omega)}\, dx = \frac{\partial g(t;\omega)/\partial \psi}{g(t;\omega)} = \frac{\partial \log g(t;\omega)}{\partial \psi} = S_{g;\psi} .$$

The fourth equality yields because $g(t;\omega) = \int_{\{x:t(x)=t\}} p(x;\omega)\, dx$. □

Property 5

a. $\mathscr{E}\left(S_{p;\psi}\, S_{g;\psi}^{T}\right) = \mathscr{E}\left(S_{g;\psi}\, S_{g;\psi}^{T}\right)$; and

b. $\mathscr{E}\left(S_{p;\tau}\, S_{g;\tau}^{T}\right) = \mathscr{E}\left(S_{g;\tau}\, S_{g;\tau}^{T}\right)$.

Proof 5a

Following a general property of conditional expectations it yields

$$\mathscr{E}\left(S_{p;\psi}\, S_{g;\psi}^{T}\right) = \mathscr{E}_{t}\Big(\mathscr{E}_{x \mid t}\left(S_{p;\psi}\, S_{g;\psi}^{T} \mid T = t\right)\Big),$$

but in the conditional distribution $S_{g;\psi}$ is a constant. So

$$\mathscr{E}_{x \mid t}\left(S_{p;\psi}\, S_{g;\psi}^{T} \mid T = t\right) = \mathscr{E}_{x \mid t}\left(S_{p;\psi} \mid T = t\right) S_{g;\psi}^{T}.$$

Then using property 4a, 5a follows. □

In the special case of a separable two-parameter distribution that can be decomposed as in (7), the following properties can be added.

Property 6

a. $S_{p;\psi} = S_{f;\psi} + S_{g;\psi}$, by definition.

The same yields for score statistics with respect to τ. But in this case:

b. $S_{f,\tau} = 0$ and $S_{p;\tau} = S_{g;\tau}$.

Property 7

a. $\mathscr{E}\left(S_{f;\psi} S_{g;\tau}^T\right) = 0$; and

b. $\mathscr{E}\left(S_{f;\tau} S_{g;\psi}^T\right) = 0$.

Proof 7a

Because of property 3 we have:

$$\mathscr{E}\left(S_{p;\psi} S_{p;\tau}^T + S_{p;\psi,\tau^T}\right) = \mathscr{E}\left[S_{p;\psi} S_{p;\tau}^T + \frac{\partial}{\partial \psi} S_{p;\tau^T}\right] = 0.$$

Applying property 6 this can be rewritten as:

$$\mathscr{E}\left[\left(S_{f;\psi} + S_{g;\psi}\right) S_{g;\tau}^T + \frac{\partial}{\partial \psi} S_{p;\tau^T}\right] = 0,$$

$$\mathscr{E}\left[S_{f;\psi} S_{g;\tau}^T + S_{g;\psi} S_{g;\tau}^T + \frac{\partial}{\partial \psi} S_{g;\tau^T}\right] = 0,$$

$$\mathscr{E}\left(S_{f;\psi} S_{g;\tau}^T\right) + \mathscr{E}\left(S_{g;\psi} S_{g;\tau}^T\right) + \mathscr{E}\, S_{g;\psi,\tau^T} = 0.$$

And the last two terms in the left hand side of the last equation sum to 0, because of property 3a. □

Property 8

a. $\mathscr{E}\left(S_{f;\psi} S_{g;\psi}^T\right) = 0$, and

b. $\mathscr{E}\left(S_{f;\tau} S_{g;\tau}^T\right) = 0$.

Proof 8a

Because of property 5a

$$\mathscr{E}\left(S_{p;\psi}\, S_{g;\psi}^{T} - S_{g;\psi}\, S_{g;\psi}^{T}\right) = 0,$$

$$\mathscr{E}\left[\left(S_{p;\psi} - S_{g;\psi}\right) S_{g;\psi}^{T}\right] = 0.$$

Using property 6 this becomes

$$\mathscr{E}\left[\left(S_{f;\psi} + S_{g;\psi} - S_{g;\psi}\right) S_{g;\psi}^{T}\right] = \mathscr{E}\left(S_{f;\psi}\, S_{g;\psi}^{T}\right) = 0.\square$$

Theorems on the F-information in separable models

Theorem 1

For a two-parameter distribution which can be factored in $p(x;\omega) = f(x \mid t;\psi).g(t;\omega)$

a. the F-information for ψ in p is the sum of the F-information in f and the F-information in g:

$$\mathrm{I}_{p}(\psi;\omega) = \mathrm{I}_{f}(\psi;\omega) + \mathrm{I}_{g}(\psi;\omega). \tag{15a}$$

b. Moreover, the F-information in the conditional distribution is the same as the expectation of the Fisher information in the conditional distribution, with the expectation taken with respect to the distribution of the sufficient statistic T:

$$\mathrm{I}_{f}(\psi;\omega) = \mathscr{E}\,\mathrm{I}_{f}(\psi \mid T).\ \text{So,}\ \mathrm{I}_{p}(\psi;\omega) = \mathscr{E}\,\mathrm{I}_{f}(\psi \mid T) + \mathrm{I}_{g}(\psi;\omega). \tag{15b}$$

Proof Theorem 1:

a. Using (11) and (12) we write

$$\mathrm{I}_{p}(\psi;\omega) = \mathscr{E}\left[S_{p;\psi}.S_{p;\psi}^{T}\right] - \mathscr{E}\left[S_{p;\psi}\, S_{p;\tau}^{T}\right] \left(\mathscr{E}\left[S_{p;\tau}\, S_{p;\tau}^{T}\right]\right)^{-1} \mathscr{E}\left[S_{p;\tau}\, S_{p;\psi}^{T}\right]. \tag{16}$$

Replacing in (16) the efficient score statistics in p by those in f and g, using property 6 above, gives:

$$I_p(\psi;\omega) = \mathcal{E}\left(S_{f;\psi} S_{f;\psi}^T + S_{f;\psi} S_{g;\psi}^T + S_{g;\psi} S_{f;\psi}^T + S_{g;\psi} S_{g;\psi}^T\right) -$$

$$\mathcal{E}\left(S_{f;\psi} S_{g;\tau}^T + S_{g;\psi} S_{g;\tau}^T\right) \left(\mathcal{E}\left[S_{g;\tau} S_{g;\tau}^T\right]\right)^{-1} \mathcal{E}\left(S_{g;\tau} S_{f;\psi}^T + S_{g;\tau} S_{g;\psi}^T\right)$$

In this expression, due to properties 7 and 8,

$\mathcal{E}\left(S_{f;\psi} S_{g;\psi}^T\right) = \mathcal{E}\left(S_{g;\psi} S_{f;\psi}^T\right) = \mathcal{E}\left(S_{f;\psi} S_{g;\tau}^T\right) = \mathcal{E}\left(S_{g;\tau} S_{f;\psi}^T\right) = 0$. So,

$$I_p(\psi;\omega) = \mathcal{E}\left(S_{f;\psi} S_{f;\psi}^T\right) + \mathcal{E}\left(S_{g;\psi} S_{g;\psi}^T\right) - \mathcal{E}\left(S_{g;\psi} S_{g;\tau}^T\right) \left(\mathcal{E}\left[S_{g;\tau} S_{g;\tau}^T\right]\right)^{-1} \mathcal{E}\left(S_{g;\tau} S_{g;\psi}^T\right).$$

In the right hand side of this expression, according to (12), the last two terms give the F-information in g, and $\mathcal{E}(S_{f;\psi} S_{f;\psi}^T) = I_f(\psi;\omega)$, the F-information in f, because $S_{f,\tau} = 0$.
This results in:

$I_p(\psi;\omega) = I_f(\psi;\omega) + I_g(\psi;\omega)$.□

b. From the definitions and a property of conditional expectations follows:

$I_f(\psi;\omega) = \mathcal{E}_x\left[S_{f;\psi} . S_{f;\psi}^T\right] = \mathcal{E}_t\left[\mathcal{E}_{x|t}(S_{f;\psi} . S_{f;\psi} \mid T)\right] = \mathcal{E} I_f(\psi|T)$, which gives:

$I_p(\psi;\omega) = \mathcal{E} I_f(\psi \mid T) + I_g(\psi;\omega)$.□

In separable two-parameter distributions therefor, the F-information with respect to the interest parameter ψ can be written as the sum of the F-information in the conditional distribution and the F-information in the marginal distribution of the statistic $T(X)$ This additivity of the components of the F-information is a very attractive. It will be clear that the loss of information by using the conditional model instead of the full two-parameter model in an inference on ψ is given by the F-information in T: $I_g(\psi;\omega)$. The efficiency of using the conditional instead of the full model can be expressed as $I_f(\psi;\omega)/I_p(\psi;\omega)$. However, it should be remembered, that in general, having a k-vector interest parameter ψ, the F-informations are $k \times k$ matrices. For comparing matrices and computing efficiencies, the trace function of the matrices will be used. The efficiency of two models with respect to the information on ψ will be computed as:

$$\mathrm{EFF}(\psi;f:p) = \frac{\mathrm{tr}(\mathrm{I}_f(\psi;\omega))}{\mathrm{tr}(\mathrm{I}_p(\psi;\omega))} \tag{17}$$

Example

An example of the computation of the F-information is given for a problem in which both the interest and the nuisance parameter are scalar. Consider a random sample from a normal distribution $X = (X_1, \ldots, X_n)$, $X_i \sim N(\mu, \sigma^2)$, with both the mean μ and the variance σ^2 unknown. The interest is in estimating σ^2, while μ is considered as a nuisance parameter. In this problem, the sample mean $T = \Sigma_i X_i / n$ is sufficient for μ and the full two-parameter model can be decomposed as in (7):

$$p(x;\mu,\sigma^2) = f(x \mid t;\sigma^2)\, g(t;\mu,\sigma^2).$$

For estimating σ^2, the full model p or the conditional model f can be used.

The Fisher information matrices for the full model p and the distribution g of T, $(T \sim N(\mu, \sigma^2/n))$, are

$$\mathrm{I}_p(\mu, \sigma^2) = \begin{bmatrix} n\sigma^{-2} & 0 \\ 0 & \dfrac{n\sigma^{-4}}{2} \end{bmatrix}, \text{ and } \mathrm{I}_g(\mu, \sigma^2) = \begin{bmatrix} n\sigma^{-2} & 0 \\ 0 & \dfrac{\sigma^{-4}}{2} \end{bmatrix}.$$

The F-information with respect to σ^2 in p and g follow from (12):
$\mathrm{I}_p(\sigma^2;(\mu, \sigma^2)) = n/2\sigma^4$, and $\mathrm{I}_g(\sigma^2;(\mu, \sigma^2)) = 1/2\sigma^4$. From Theorem 1 (15a), the F-information in the conditional model equals: $\mathrm{I}_f(\sigma^2;(\mu, \sigma^2)) = (n-1)/2\sigma^4$. So in using CML instead of ML for estimating σ^2, there is a small loss of information which amounts to $1/2\sigma^4$. The efficiency of CML versus ML is given by: $\mathrm{I}_f(\sigma^2;(\mu, \sigma^2)) / \mathrm{I}_p(\sigma^2;(\mu, \sigma^2)) = (n-1)/n$. In this example it can be remarked that, although there is a little loss of information in using CML, the resulting CML estimator is unbiased, whereas the ML estimator is not: $\hat{\sigma}^2_{cml} = \Sigma_i (x_i - t)^2/(n-1)$ and $\hat{\sigma}^2_{ml} = \Sigma_i (x_i - t)^2/n$. □

The conditions for no loss of information

Next, two definitions are given which make it possible to specify the conditions under which there is no loss of information when the conditional model is used instead of the full model.

Definition 2.

In a two-parameter distribution, satisfying (7), the statistic $T = T(X)$ is *strongly ancillary* for ψ, if $g(t;\omega)$ only depends on τ: $p(x;\omega) = f(x \mid t;\psi).g(t;\tau)$.

It is easily understood that in case the statistic T is strongly ancillary for ψ, there will be no loss of information if in an inference on ψ the conditional distribution $f(x \mid t;\psi)$ is used instead of $p(x;\omega)$. In this case the part which is neglected does not depend on the parameter ψ and $I_g(\psi;\omega) = 0$.

Bhapkar (1989), generalizing an earlier result of Liang (1983) for scalar parameters to vector parameters, has proven that there is also no loss of information if weaker conditions on the distribution of T are fulfilled. Consider the following definition:

Definition 3.

In a two-parameter distribution, satisfying (7), the statistic $T = T(X)$ is *weakly ancillary* for ψ, if there exists a one-to-one reparametrization between (ψ,τ) and (ψ,δ) such that $g(t;\omega) = g(t;\delta)$, that is, only depends on δ.

Example

X_1 and X_2 are independent Poisson distributed variables: $X = (X_1, X_2), \mathcal{E}(X_1) = \lambda$ and $\mathcal{E}(X_2) = \lambda\rho$. The interest and nuisance parameter are respectively $\ln\rho$ and $\ln\lambda$. $T(X) = X_1 + X_2$ is also Poisson distributed with $\mathcal{E}(T) = \lambda + \lambda\rho$. Factorization (7) yields:

$$p(x;\lambda,\rho) = \frac{\exp\left[(x_1 + x_2)\ln\lambda + x_2\ln\rho - (\lambda + \lambda\rho)\right]}{x_1!\,x_2!} =$$

$$\frac{t!\exp\left[x_2\ln\rho\right]}{x_1!\,x_2!\exp\left[t\ln(1+\rho)\right]} \cdot \frac{\exp\left[t\ln\lambda + t\ln(1+\rho) - (\lambda + \lambda\rho)\right]}{t!} = f(x \mid t;\rho)\ .\ g(t;\lambda,\rho)\,.$$

Apply the one-one reparametrization $\ln\rho = \ln\rho$ and $\ln\lambda = \ln\delta - \ln(1+\rho)$. The distribution of T is then given by: $g(t;\rho,\delta) = (\exp[t\ln\delta - \delta])/t!$, which only depends on δ.□

Next, Bhapkar's theorem (1989) is given.

Theorem 2

A two-parameter distribution which can be factored in $p(x;\omega) = f(x \mid t;\psi) \cdot g(t;\omega)$ and in which the statistic $T = T(X)$ is weakly ancillary for ψ, there is no loss of information in using the conditional distribution $f(x \mid t;\psi)$ instead of $p(x;\omega)$ for inference on ψ: $I_g(\psi;\omega) = 0$.

The condition of weak ancillarity, although slightly differently defined, was also the key condition under which Andersen (1970, 1973) established the asymptotic efficiency of conditional maximum likelihood estimators. According to Andersen (1973, p. 99), weak ancillarity of the statistic T means that no inference about the interest parameter ψ can be drawn from the distribution of T, $g(t;\omega)$, that is not completely dependent on the specification of the nuisance parameter τ. In other words, nothing can be learned from the data about the interest parameter from the sole observation of the statistic. Although intuitively weak ancillarity may be an appealing concept, it is not easy to show in general that a distribution has this property. However, for models belonging to the exponential family, necessary and sufficient conditions for weak ancillarity, which are easily checked, are given in the next theorem (Andersen, 1973; Liang, 1983; Bhapkar 1989).

Theorem 3

If X has a distribution belonging to the exponential family with natural parameters $\omega = (\psi, \tau)$:

$$p(x;\omega) = c(\psi, \tau) \cdot \exp\left\{u^T(x) \cdot \psi + t^T(x) \cdot \tau + b(x)\right\} \tag{18}$$

and the distribution of $T = T(X)$ is given by

$$g(t;\omega) = c(\psi, \tau) \cdot \gamma(t, \psi) \cdot \exp(t^T\tau), \tag{19}$$

then T is weakly ancillary for ψ <u>if and only if</u> there exist functions of ψ only and independent of the data: $w_i(\psi)$, $i = 1,\ldots,k$, and $v(\psi)$, such that:

$$\frac{\partial \ln \gamma(t,\psi)}{\partial \psi_i} = w_i(\psi)\, t + v(\psi), \text{ for } i = 1,\ldots,k \tag{20}$$

for all (ae) t.

Checking the Conditions for no Loss of Information using CML in Two Rasch Models

The Rasch Poisson Counts model

In this model, proposed by Rasch (1960), the number of failures X_{vi} of person $v = 1,\ldots,n$ on test $i = 1,\ldots,k$, which each consist of a number of items with low error probabilities, is considered. A well known application of the model is the number of misreadings in texts. The model assumes that X_{vi} is Poisson distributed with parameter $\lambda_{vi} = \beta_i\theta_v$, β_i being the difficulty parameter of text i and θ_v the ability parameter of person v. Writing the model in the exponential family form (18) gives:

$$P(X_{vi} = x_{vi};\beta_i\theta_v) = \frac{\exp(-\beta_i\theta_v)\,.\,\exp\{x_{vi}\ln\beta_i + x_{vi}\ln\theta_v\}}{x_{vi}!},\quad x_{vi} = 0,1,2,\ldots.$$

With the assumption of independence over texts and persons we get:

$$p(x;\beta,\theta) = \prod_i \prod_v P(X_{vi} = xvi;\beta_i,\theta_v) =$$

$$\frac{\exp\left\{\sum_i \sum_v -\beta_i\theta_v\right\}\,.\,\exp\left\{\sum_i\left(\sum_v x_{vi}\right)\ln\beta_i + \sum_v\left(\sum_i x_{vi}\right)\ln\theta_v\right\}}{\prod_i \prod_v x_{vi}!}.$$

It is easily seen that the model is a member of the exponential family (18), in which the statistic $T_v = \sum_i X_{vi}$, the number of failures of person v is sufficient for $\ln\theta_v$, for $v = 1,\ldots,n$.

So, the model is separable:

$$p(x;\beta,\theta) = \prod_{v=1}^{n} f(x_v \mid t_v;\beta)\,.\,\prod_{v=1}^{n} g(t_v;\beta,\theta_v),$$

and for estimating the text parameters β CML estimation can be considered to use. Instead of the full model $p(x;\beta,\theta)$, only the conditional model, the first part of the factorization, is used. Whether neglecting the distribution of T causes loss of information is now easily checked.

Observe that T_v is also Poisson distributed, with parameter $\theta_v\Sigma_i\beta_i$:

$$g(t;\beta,\theta) = \prod_{v=1}^{n} g(t_v;\beta,\theta_v) =$$

$$\frac{\exp\left\{\sum_i \sum_v - \beta_i \theta_v\right\} \cdot \exp\left\{\left(\sum_v t_v\right) \ln \sum_i \beta_i + \sum_v t_v \ln \theta_v\right\}}{\prod_v t_v!}.$$

So the function $\gamma(t,\psi)$ in (19) is given by

$$\gamma(t;\beta) = \prod_{v=1}^{n} \left[\exp\left\{t_v \ln \sum_i \beta_i\right\}\right] / t_v!,$$

and because $\ln \gamma(t;\beta) = \sum_v \left\{t_v \ln \sum_i \beta_i - \ln(t_v!)\right\}$

$$\frac{\partial \ln \gamma(t;\beta)}{\partial \beta_j} = \sum_v t_v \frac{1}{\sum_i \beta_i}, \text{ for } j = 1,\ldots,k.$$

The condition, (20), for T being weakly ancillary for β is seen to be fulfilled: $\partial \ln \gamma(t;\beta) / \partial \beta_j$ is a linear function of t, and $w_j(\psi) = 1 / \Sigma_i \beta_i$, $j = 1,\ldots,k$ only depends on the text parameter β. So, there will be no loss of information if the text parameters are estimated with CML. In terms of F-information, with $\omega^T = (\beta^T, \theta^T)$: $I_p(\beta;\omega) = I_f(\beta;\omega)$ or $I_g(\beta;\omega) = 0$.

The Rasch Model for Dichotomously Scored Items

The same conditions will be checked for the model which was presented in the introduction. Writing this model, (1) and (2), as

$$p(x;\beta,\theta) = \frac{\exp\left\{-\sum_i \left(\sum_v x_{vi}\right)\beta_i + \sum_v \left(\sum_i x_{vi}\right)\theta_v\right\}}{\prod_i \prod_v \left\{1 + \exp(\theta_v - \beta_i)\right\}}, \tag{21}$$

it can be seen to belong to the exponential family. $T_v = \sum_i X_{vi}$ is sufficient for θ_v, for $v = 1,\ldots,n$. The distribution of T is checked for weak ancillarity. This distribution is given by

$$g(t;\beta,\theta) = \prod_{v=1}^{n} g(t_v;\beta,\theta_v) = \prod_{v=1}^{n} \frac{\exp(\theta_v t_v) \cdot \gamma_{t_v}(\beta)}{\prod_{i=1}^{k} \left\{1 + \exp(\theta_v - \beta_i)\right\}}, \tag{22}$$

with $\gamma_{t_v}(\beta) = \sum_{\sum x_{vi}=t_v} \exp\left[-\sum_{i=1}^{k} \beta_i x_{vi}\right]$.

For, $j = 1,\ldots,k$, this gives

$$\frac{\partial \ln \prod_{v=1}^{n} \gamma_{t_v}(\beta)}{\partial \beta_j} = \sum_{v=1}^{n} \frac{\partial \ln \gamma_{t_v}(\beta)}{\partial \beta_j} = \sum_{v=1}^{n} \frac{e^{-\beta_j} \cdot \gamma_{t_v-1}^{(j)}}{\gamma_{t_v}(\beta)} \cdot \frac{t_v}{t_v} = \sum_{v=1}^{n} \frac{e^{-\beta_j} \cdot \gamma_{t_v-1}^{(j)}}{\sum_{i=1}^{k} e^{-\beta_i} \cdot \gamma_{t_v-1}^{(i)}} \cdot t_v, \qquad (23)$$

in which

$\gamma_{t_v-1}^{(j)} := \partial \gamma_{t_v}(\beta) / \partial e^{-\beta_j}$, for $j = 1,\ldots,k$.

Note that the last equality in (23) uses an expression given by Fischer (1974, p. 242):

$$\gamma_{t_v}(\beta) \,.t_v = \sum_{i=1}^{k} e^{-\beta_i} \gamma_{t_v-1}^{i}.$$

It is seen that condition (20) of Theorem 3 is not fulfilled, since the function $w_j(\psi)$ is not only dependent on the item parameters β, but also on the statistic t_v. T is therefore not weakly ancillary for β, which means that using CML estimation is possibly accompanied with loss of information compared to the situation in which the full model is used. In terms of F-information: $I_p(\beta;\omega) \geq I_f(\beta;\omega)$ or $I_g(\beta;\omega) \geq 0$.

F-Information in the Dichotomous Rasch Model: Comparing JML and CML

In order to determine the loss of information in the Rasch model, with both the item- and ability parameters considered fixed, the expressions for the F-information in the full model p, the distribution g of the sufficient statistic and conditional distribution f, are given first. The Fisher information matrix of the model $p(x;\beta,\theta)$, given in (21), with respect to both parameters β and θ, is written in the partitioned form as in (11):

$$\mathrm{I}_p(\omega) = \mathscr{E}\begin{bmatrix} S_{p;\beta}S_{p;\beta}^T & S_{p;\beta}S_{p;\theta}^T \\ S_{p;\theta}S_{p;\beta}^T & S_{p;\theta}S_{p;\theta}^T \end{bmatrix} =: \begin{bmatrix} \mathrm{I}_p^{\beta\beta^T} & \mathrm{I}_p^{\beta\theta^T} \\ \mathrm{I}_p^{\theta\beta^T} & \mathrm{I}_p^{\theta\theta^T} \end{bmatrix} \tag{24}$$

Using the properties 1, 2 and 3 of the efficient score statistics, the submatrices, $\mathrm{I}_p^{\beta\beta^T}$ $(k \times k)$, $\mathrm{I}_p^{\beta\theta^T} = [\mathrm{I}_p^{\theta\beta^T}]^T$ $(k \times n)$ and $\mathrm{I}_p^{\theta\theta^T}$ $(n \times n)$, which determine the F-information (see (12)), are easily seen to be the negative expectation of the second derivatives of $\ln p$ with respect to the parameters. The elements of the submatrices in (24) are then, writing

$$p_{vi} := \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}, \tag{25}$$

given by:

$$\left(\mathrm{I}_p^{\beta\beta^T}\right)_{ii} = \sum_{v=1}^{n} p_{vi}(1 - p_{vi}) \text{ for } i = 1,\ldots,k; \tag{26a}$$

$$\left(\mathrm{I}_p^{\beta\beta^T}\right)_{ij} = 0 \text{ for } i \neq j = 1,\ldots,k; \tag{26b}$$

$$\left(\mathrm{I}_p^{\beta\theta^T}\right)_{iv} = -p_{vi}(1 - p_{vi}) \text{ for } i = 1,\ldots,k\,;\, v = 1,\ldots,n; \tag{26c}$$

$$\left(\mathrm{I}_p^{\theta\theta^T}\right)_{vv} = \sum_{i=1}^{k} p_{vi}(1 - p_{vi}) \text{ for } v = 1,\ldots,n; \tag{26d}$$

$$\left(\mathrm{I}_p^{\theta\theta^T}\right)_{vw} = 0 \text{ for } v \neq w = 1,\ldots,n. \tag{26e}$$

The Fisher information matrix is clearly singular. For every row and column

$$\sum_{i=1}^{k+n}\left(\mathrm{I}_p(\omega)\right)_{ij} = \sum_{j=1}^{k+n}\left(\mathrm{I}_p(\omega)\right)_{ij} = 0 \text{ for } i,j = 1,\ldots,k+n.$$

But because the diagonal submatrix $\mathrm{I}_p^{\theta\theta^T}$ is positive definite, and the elements of the inverse are given by the reciprocals of (26d), using (14), the F-information in the full Rasch model $\mathrm{I}_p(\beta;\omega)$, is given by

$$\left(\mathrm{I}_p(\beta;\omega)\right)_{ii} = \sum_{\nu=1}^{n} p_{\nu i}(1 - p_{\nu i}) - \sum_{\nu=1}^{n} \frac{\left[p_{\nu i}(1 - p_{\nu i})\right]^2}{\sum_{\ell=1}^{k} p_{\nu\ell}(1 - p_{\nu\ell})} \quad \text{for } j = 1,\ldots,k, \tag{27a}$$

$$\left(\mathrm{I}_p(\beta;\omega)\right)_{ij} = -\sum_{\nu=1}^{n} \frac{p_{\nu i}(1 - p_{\nu i}) \cdot p_{\nu j}(1 - p_{\nu j})}{\sum_{i=1}^{k} p_{\nu\ell}(1 - p_{\nu\ell})} \quad \text{for } i \neq j = 1,\ldots,k. \tag{27b}$$

Next, the F-information in the distribution $g(t;\beta,\theta)$ of the sufficient statistic (see (22)) is determined. The Fisher information matrix, analogous to (24), is given by

$$\mathrm{I}_g(\omega) = \mathscr{E}\begin{bmatrix} S_{g;\beta} S_{g;\beta}^T & S_{g;\beta} S_{g;\theta}^T \\ S_{g;\theta} S_{g;\beta}^T & S_{g;\theta} S_{g;\theta}^T \end{bmatrix} =: \begin{bmatrix} \mathrm{I}_g^{\beta\beta^T} & \mathrm{I}_g^{\beta\theta^T} \\ \mathrm{I}_g^{\theta\beta^T} & \mathrm{I}_g^{\theta\theta^T} \end{bmatrix}. \tag{28}$$

Because of the properties 3 and 6 ($S_{g,\theta} = S_{p,\theta}$) of the efficient score statistics, $\mathrm{I}_g^{\beta\theta^T} = \mathrm{I}_p^{\beta\theta^T}$ and $\mathrm{I}_g^{\theta\theta^T} = \mathrm{I}_p^{\theta\theta^T}$, which are given by (26c), (26d) and (26e). This is, of course, a consequence of the fact that T_ν is sufficient for θ_ν in $p(x;\beta,\theta)$. The second derivatives of $\ln g$ with respect to β are given by:

$$\partial^2 \ln g / \partial \beta_i^2 = \sum_{\nu=1}^{n} \left\{ \frac{e^{-\beta_i} \cdot \gamma_{t_\nu - 1}^{(i)}}{\gamma_{t_\nu}} - \left[\frac{e^{-\beta_i} \cdot \gamma_{t_\nu - 1}^{(i)}}{\gamma_{t_\nu}}\right]^2 - p_{\nu i}(1 - p_{\nu i}) \right\} \quad \text{for } i = 1,\ldots,k,$$

and

$$\frac{\partial^2 \ln g}{\partial \beta_j \partial \beta_i} = \sum_{\nu=1}^{n} \left\{ \frac{e^{-\beta_i - \beta_j} \cdot \gamma_{t_\nu - 2}^{(i,j)}}{\gamma_{t_\nu}} - \frac{e^{-\beta_i - \beta_j} \cdot \gamma_{t_\nu - 1}^{(i)} \gamma_{t_\nu - 1}^{(j)}}{\gamma_{t_\nu}^2} \right\} \quad \text{for } i \neq j = 1,\ldots,k,$$

with

$$\gamma_{t_\nu - 2}^{(i,j)} := \frac{\partial^2 \gamma_{t_\nu}(\beta)}{\partial e^{-\beta_i} \, \partial e^{-\beta_j}}.$$

Observe that in the dichotomous Rasch model

$$\mathrm{P}(X_{\nu i} = 1 \mid T_\nu = t_\nu;\beta) = \frac{e^{-\beta_i} \cdot \gamma_{t_\nu - 1}^{(i)}}{\gamma_{t_\nu}} =: p_{\nu i \mid t_\nu} \quad \text{for } i = 1,\ldots,k$$

$$P\left(X_{vi}=1, X_{vj}=1 \mid T_v=t_v;\beta\right) = \frac{e^{-\beta_i-\beta_j}\cdot \gamma_{t_v-2}^{(ij)}}{\gamma_{t_v}} =: p_{vi,vj\mid t_v} \text{ for } i \neq j = 1,\ldots,k\,.$$

So, the elements of the submatrix $I_g^{\beta\beta^T}$ of (28) are given by:

$$\left(I_g^{\beta\beta^T}\right)_{ii} = \mathscr{E}\sum_{v=1}^{n}\left\{-p_{vi\mid t_v}\left(1-p_{vi\mid t_v}\right) + p_{vi}\left(1-p_{vi}\right)\right\},\ i=1,\ldots,k \text{ and} \tag{29a}$$

$$\left(I_g^{\beta\beta^T}\right)_{ij} = \mathscr{E}\sum_{v=1}^{n}\left\{-p_{vi,vj\mid t_v} + p_{vi\mid t_v}\cdot p_{vj\mid t_v}\right\},\ i\neq j=1,\ldots,k. \tag{29b}$$

The Fisher information matrix of g with respect to β and θ, as specified in (29ab) and (26cde), is not positive definite, but $I_g^{\theta\theta^T}$ is positive definite, and (14) can be used again to find the F-information in g with respect to β. This matrix is for $i=1,\ldots,k$ given by:

$$\left(I_g(\beta;\omega)\right)_{ii} = \mathscr{E}\sum_{v=1}^{n}\left\{-p_{vi\mid t_v}\left(1-p_{vi\mid t_v}\right) + p_{vi}\left(1-p_{vi}\right)\right\} - \sum_{v=1}^{n}\frac{\left[p_{vi}\left(1-p_{vi}\right)\right]^2}{\sum_{\ell=1}^{k} p_{v\ell}\left(1-p_{v\ell}\right)}, \tag{30a}$$

and for $i\neq j=1,\ldots,k$

$$\left(I_g(\beta;\omega)\right)_{ij} = \mathscr{E}\sum_{v=1}^{n}\left\{-p_{vi,vj\mid t_v} + p_{vi\mid t_v}\cdot p_{vj\mid t_v}\right\} - \sum_{v=1}^{n}\frac{p_{vi}\left(1-p_{vi}\right)\cdot p_{vj}\left(1-p_{vj}\right)}{\sum_{\ell=1}^{k} p_{v\ell}\left(1-p_{v\ell}\right)}. \tag{30b}$$

Using (15ab) on the additivity of the F-information, the F-information in the conditional distribution $f(x \mid t;\beta)$ simply is found by subtracting the expressions in (30ab) from (27ab) and this gives:

$$\left(\mathscr{E}\, I_f(\beta \mid T)\right)_{ii} = \mathscr{E}\sum_{v=1}^{n} p_{vi\mid t_v}\left(1-p_{vi\mid t_v}\right) \text{ for } i=1,\ldots,k \text{ and} \tag{31a}$$

$$\left(\mathscr{E}\, I_f(\beta \mid T)\right)_{ij} = \mathscr{E}\sum_{v=1}^{n}\left(p_{vi,vj\mid t_v} - p_{vi\mid t_v}\cdot p_{vj\mid t_v}\right) \text{ for } i\neq j=1,\ldots,k. \tag{31b}$$

All three relevant F-information matrices in the Rasch model are now specified in (27ab), (30ab) and (31ab). Once the expectations in (30ab) and (31ab) over the distribution of T_v (22) are determined, they are easily computed for a given set of item and ability parameters. The diagonal of the F-information in the conditional distribution (31a), for instance, is

$$\left(\mathcal{E}\, \mathrm{I}_f(\beta \mid T)\right)_{ii} = \sum_{v=1}^{n} \sum_{t_v=0}^{k} p_{vi \mid t_v}\left(1 - p_{vi \mid t_v}\right) \exp(\theta_v t_v)\, \gamma_{t_v} \prod_{i=1}^{k} \left(1 - p_{vi}\right) \text{ for } i = 1, \ldots, k.$$

The expressions (30ab) and (31b) change similarly.
Note that the expressions for the F-information matrices are all sums of independent contributions of the persons $v = 1, \ldots, n$.

Example
Consider the Rasch model with 3 items, $\beta_1 = -.5$, $\beta_2 = 0.0$ and $\beta_3 = 1.75$, and 4 persons with θ_v respectively -1.0, 0.0, 1.0, and 2.0.
The Fisher information matrix ($k + n \times k + n$) is given by:

$$\mathrm{I}_p(\omega) = \begin{bmatrix} .689 & .000 & .000 & -.235 & -.235 & -.149 & -.070 \\ .000 & .748 & .000 & -.197 & -.250 & -.197 & -.105 \\ .000 & .000 & .647 & -.056 & -.126 & -.218 & -.246 \\ -.235 & -.197 & -.056 & .488 & .000 & .000 & .000 \\ -.235 & -.250 & -.126 & .000 & .611 & .000 & .000 \\ -.149 & -.197 & -.218 & .000 & .000 & .564 & .000 \\ -.070 & -.105 & -.246 & .000 & .000 & .000 & .421 \end{bmatrix}$$

The structure of the matrix is clear: The item parameter submatrix as well as the ability parameter submatrix is diagonal and all coformations between items and abilities are negative.
The symmetric F-Information matrix in p ($k \times k$), using (27ab), is given by

$$\mathrm{I}_p(\beta; \omega) = \begin{bmatrix} .435 & -.260 & -.174 \\ -.260 & .472 & -.212 \\ -.174 & -.212 & .386 \end{bmatrix}.$$

Note that in JML estimation in the Rasch model the diagonal of the item parameter submatrix of the Fisher information matrix is sometimes used for estimating the standard errors of the item parameters (Wright, 1977). (In the computations the matrix is evaluated at the estimates). This practice leads to an underestimate of the standard error, because a better estimate, taking account of the negative coformation between the items and abilities, would use the diagonal of the F-information matrix.
The F-information matrices in the conditional distribution f and the marginal distribution g are given by

$$I_f(\beta;\omega) = \begin{bmatrix} .409 & -.272 & -.137 \\ -.272 & .451 & -.180 \\ -.137 & -.180 & .316 \end{bmatrix} \text{ and } I_g(\beta;\omega) = \begin{bmatrix} .026 & .012 & -.037 \\ .012 & .021 & -.032 \\ -.037 & -.032 & .070 \end{bmatrix}.$$

The loss of information in using the conditional distribution $f(x \mid t;\beta)$ instead of the full model $p(x;\beta,\theta)$, that is, in using CML instead of JML estimation for the item parameters is given by $I_g(\beta;\omega)$, the F-information matrix in g. As a scalar measure of this loss, the trace of this matrix is used. Note that in our application it is easily checked that the determinant function is useless for this, because all F-information matrices are singular:

$$\sum_i I_p(\beta;\omega) = \sum_j I_p(\beta;\omega) = \sum_i I_f(\beta;\omega) = \sum_j I_f(\beta;\omega) = \sum_i I_g(\beta;\omega) = \sum_j I_g(\beta;\omega) = 0.$$

The loss of information in the example appears to be very small. In the example, the efficiency measure, defined in (17), $\mathrm{EFF}(\beta;f:p) = \mathrm{tr}(I_f(\beta;\omega)) / \mathrm{tr}(I_p(\beta;\omega))$, expressing the loss in using the conditional instead of the full model is $\mathrm{EFF}(\beta;f:p) = 1.176 / 1.293 = .910$.□

Efficiency comparison of JML and CML estimation

The efficiency of CML versus JML will be reported for some typical item parameter and ability parameters sets. For these sets, 100 ability parameters were drawn from a normal ability distribution $N(\mu,\sigma^2)$. Table 1 gives the efficiencies for 10 items with β =(-3, -2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2, 3) and 100 abilities drawn from the normal distribution with varying μ and σ^2.

Table 1

EFF(β;f:p) for β = (-3,-2,-1.5,-1,-0.5,0.5,1,1.5,2,3) and 100 abilities from $N(\mu,\sigma^2)$

	σ^2			
μ	.25	1	2.25	4
-2	.984	.985	.985	.986
-1	.988	.987	.987	.986
0	.988	.988	.987	.986
1	.988	.988	.986	.986
2	.985	.985	.986	.985

It should be noted that the traces of $I_f(\beta;\omega)$ as well as of $I_p(\beta;\omega)$ decrease, when σ^2 is increased, and also when the distance between the mean of the ability- and the mean of the item parameters, $|\mu - \frac{1}{k}\Sigma_k\beta_i|$, is increased. However, the efficiency of CML versus JML, reported in Table 1, hardly shows any variation and amounts to at least 98.4%.

Table 2 shows the relation between the efficiency and the spread in the item parameters. For 100 abilities drawn from N(0,1), the efficiency for estimating 10 equidistant item parameters with $\sum_i\beta_i = 0$ and varying stepsize, $\beta_{i+1} - \beta_i, i = 1,\ldots,9$ is given.

Table 2

EFF(β;f:p) for 10 equidistant items with $\sum_i\beta_i = 0$ and varying stepsize and 100 abilities from $N(0,1)$

step	EFF(β;f:p)
1	.9451
.5	.9891
.25	.9968
.125	.9992
.0625	.9998
0	1.000

The efficiency of CML versus JML estimation increases with decreasing spread of the item parameters. In case all item parameters are equal there is no loss of information. The theoretical explanation for this is that the weak ancillarity condition in Theorem 3

(23) is fulfilled in this special case: $\partial \ln \prod_v \gamma_{t_v}(\beta)/\partial\beta_j = \sum_v t_v/k$ for $j = 1,\ldots,k$ and $I_g(\beta;\omega) = 0$.

In Table 3 the relation between the efficiency and the test length is illustrated. For 100 abilities drawn from N(0.5,1), the efficiency is reported for estimating the item parameters $\beta = (0,1,2)$, which are n times in a test. The test length is equal to $3n$.

Table 3

EFF$(\beta;f:p)$ for $\beta = (0,1,2)$ with varying test length and 100 abilities from $N(0.5,1)$

length	EFF$(\beta;f:p)$
3	.9360
6	.9909
9	.9965
12	.9982
15	.9989
18	.9992
21	.9995

An increasing efficiency is observed when the test length is increased. It is clear that already at a relative short test length of 21 items, CML estimation is almost as efficient as JML estimation of the item parameters.

F-Information in the Dichotomous Rasch Model: Comparing MML and CML

In order to compare CML with MML estimation of the item parameters in the Rasch model, the marginal likelihood function (3) in which the θ_v's are not fixed ability parameters but random draws from the ability distribution $h(\theta;\xi)$, is considered. With again $T_v = \Sigma_i X_{vi}$, this can be rewritten as:

$$p_m(x;\beta,\xi) = \prod_v p_m(x_v;\beta,\xi) = \prod_v \int_{-\infty}^{\infty} \prod_i \mathrm{P}(X_{vi} = x_{vi} \mid \theta_v;\beta_i)\, h(\theta_v;\xi)\, d\theta_v =$$

$$\prod_v f(x_v \mid t_v;\beta) \prod_v \int_{-\infty}^{\infty} g(t_v \mid \theta_v;\beta) h(\theta_v;\xi)\, d\theta_v. \qquad (32)$$

In (32) the parameters are the item parameter β and the parameter of the ability distribution ξ. This two-parameter distribution is factored in a part which is only dependent on the interest parameter β, and a part which is the distribution of T, which depends on both parameters β and ξ. For estimating the item parameters with CML, as seen in (5), only the first part of the factorization is used. In MML estimation the full model (32) is used. And for comparing CML estimation with MML estimation, the F-information in f, $\mathcal{E}\, \mathrm{I}_f(\beta \mid T)$, given in (31ab), can be compared with the F-information in p_m with respect to β.

Note that because in the model (32) every person is a random draw from an ability distribution, it suffices to derive the F-information in p_m for only one person, indexed with v. The same is true for the F-information in f (see (31ab)).

In order to determine these F-information matrices completely, an ability distribution $h(\theta_v, \xi)$ has to be specified. Here it is assumed that the ability distribution is normal with parameter $\xi = (\mu, \sigma^2)$, and the density is given by

$$h(\theta_v; \xi) = \sigma^{-1} \phi\left((\theta_v - \mu)/\sigma\right), \tag{33}$$

with $\phi(y)$ being the standard normal density:

$$\phi(y) = (2\pi)^{-\frac{1}{2}} \exp(-y^2/2). \tag{34}$$

F-information in p_m

As before, this matrix can be determined once the Fisher information matrix has been obtained. If the length of the vector parameter of the ability distribution is ℓ, this matrix can be written in the following partitioned form:

$$\mathrm{I}_{p_m}(\omega) = \mathcal{E} \begin{bmatrix} S_{p;\beta} S_{p;\beta}^T & S_{p;\beta} S_{p;\xi}^T \\ S_{p;\xi} S_{p;\beta}^T & S_{p;\xi} S_{p;\xi}^T \end{bmatrix} =: \begin{bmatrix} \mathrm{I}_{p_m}^{\beta\beta^T} & \mathrm{I}_{p_m}^{\beta\xi^T} \\ \mathrm{I}_{p_m}^{\xi\beta^T} & \mathrm{I}_{p_m}^{\xi\xi^T} \end{bmatrix} \tag{35}$$

The submatrices, $\mathrm{I}_{p_m}^{\beta\beta^T}$ $(k \times k)$, $\mathrm{I}_{p_m}^{\beta\xi^T} = [\mathrm{I}_{p_m}^{\xi\beta^T}]^T$ $(k \times \ell)$ and $\mathrm{I}_{p_m}^{\xi\xi^T}$ $(\ell \times \ell)$ are again the negative expectation of the second derivatives of $\ln p_m$ with respect to the parameters. For notational convenience define:

$$Q_v := Q_v\left(\theta_v, x_{vi}, t_v, \beta, \mu, \sigma^2\right) = \exp(\theta_v t_v) \prod_{i=1}^{k} (1 - p_{vi})\, h(\theta_v; \xi), \tag{36}$$

in the normal density ((33) with (34)), can be substituted for $h(\theta_v; \xi)$.
So, considered is the model:

$$p_m(x_v; \beta, \xi) = f(x_v \mid t_v; \beta) \cdot \int_{-\infty}^{\infty} g(t_v \mid \theta_v; \beta) h(\theta_v; \xi) d\theta_v, \tag{37}$$

in which (see (22), with (25) and (36)),

$$g(t_v \mid \theta_v; \beta) = \exp(\theta_v t_v) \cdot \gamma_{t_v}(\beta) \cdot \prod_{i=1}^{k} (1 - p_{vi}) = Q_v \gamma_{t_v} / h(\theta_v; \xi), \tag{38}$$

and (divide (21) by (22) for one person)

$$f(x_v \mid t_v; \beta) = \frac{\exp\left(-\sum_i \beta_i x_{vi}\right)}{\gamma_{t_v}(\beta)}. \tag{39}$$

The marginal distribution of T_v is given by

$$\mathrm{P}\left(T_{v=t_v}; \beta, \xi\right) = \int_{-\infty}^{\infty} g(t_v \mid \theta_v; \beta) h(\theta_v; \xi) d\theta_v = \gamma_{t_v} \int Q_v d\theta_v. \tag{40}$$

Inserting (38) and (39) in (37) and taking the logarithm, gives

$$\ln p_m(x_v; \beta, \xi) = \sum_i \beta_i x_{vi} + \ln \int_{-\infty}^{\infty} \exp(\theta_v t_v) h(\theta_v; \xi) \prod_{i=1}^{k} (1 - p_{vi})\, d\theta_v = \sum_i \beta_i x_{vi} + \ln \int Q_v d\theta_v.$$

With some algebra the following expressions are easily derived:

$$\frac{\partial Q_v}{\partial \beta_j} = Q_v p_{vj}, \text{ for } j = 1, \ldots, k,$$

$$\frac{\partial Q_v}{\partial \mu} = Q_v(\theta_v - \mu)\sigma^{-2},$$

$$\frac{\partial Q_v}{\partial \sigma^2} = Q_v\left[-\frac{1}{2}\sigma^{-2} + \frac{1}{2}(\theta_v - \mu)^2\sigma^{-4}\right],$$

$$\frac{\partial^2 Q_v}{\partial \beta_j^2} = Q_v(2p_{vj} - 1)p_{vj}, \text{ for } j = 1,\ldots,k, \tag{41a}$$

$$\frac{\partial^2 Q_v}{\partial \beta_j \partial \beta_m} = Q_v p_{vj} p_{vm}, \text{ for } m \neq j = 1,\ldots,k, \tag{41b}$$

$$\frac{\partial^2 Q_v}{\partial \mu^2} = Q_v\left[(\theta_v - \mu)^2\sigma^{-4} - \sigma^{-2}\right], \tag{42a}$$

$$\frac{\partial^2 Q_v}{\partial (\sigma^2)^2} = Q_v\left[\frac{3}{4}\sigma^{-4} - \frac{3}{2}(\theta_v - \mu)^2\sigma^{-6} + \frac{1}{4}(\theta_v - \mu)^4\sigma^{-8}\right], \tag{42b}$$

$$\frac{\partial^2 Q_v}{\partial \sigma^2 \partial \mu} = Q_v\left[-\frac{3}{2}(\theta_v - \mu)\sigma^{-4} + \frac{1}{2}(\theta_v - \mu)^3\sigma^{-6}\right], \tag{42c}$$

$$\frac{\partial^2 Q_v}{\partial \mu \partial \beta_j} = Q_v p_{vj}(\theta_v - \mu)\sigma^{-2}, \text{ for } j = 1,\ldots,k, \tag{43a}$$

$$\frac{\partial^2 Q_v}{\partial \sigma^2 \partial \beta_j} = Q_v p_{vj}\left[-\frac{1}{2}\sigma^{-2} + \frac{1}{2}(\theta_v - \mu)^2\sigma^{-4}\right] \text{ for } j = 1,\ldots,k. \tag{43b}$$

The second derivatives of $\ln p_m(x_v; \beta, \xi)$ are generally written as (see Appendix):

$$\frac{\partial^2 \ln p_m(x_v;\beta,\xi)}{\partial a\,\partial b} = \frac{\int \frac{\partial^2 Q_v}{\partial a\,\partial b}\,d\theta_v}{\int Q_v d\theta_v} - \frac{\int \frac{\partial Q_v}{\partial a}d\theta_v \int \frac{\partial Q_v}{\partial b}d\theta_v}{\left(\int Q_v d\theta_v\right)^2}, \tag{44}$$

with a and b being any pair from the parameters μ, σ^2, and $\beta_j, j = 1,\ldots,k$. Taking minus the expectation of (44) over the distribution of T_v (40) gives:

$$-\mathscr{E}\frac{\partial^2 \ln p_m(x_v;\beta,\xi)}{\partial a\,\partial b} = -\sum_{t_v=0}^{k}\gamma_{t_v}\left[\int \frac{\partial^2 Q_v}{\partial a\,\partial b}\,d\theta_v - \frac{\int \frac{\partial Q_v}{\partial a}d\theta_v \int \frac{\partial Q_v}{\partial b}d\theta_v}{\int Q_v d\theta_v}\right], \tag{45}$$

which is the general expression for all parts of the Fisher information matrix (35). Using (41ab) in (45) gives the item parameter part:

$$\left(I_{p_m}^{\beta\beta^T}\right)_{jj} = -\sum_{t_v=0}^{k}\gamma_{t_v}\left[\int (2p_{vj}-1)p_{vj}Q_v d\theta_v - \frac{\left(\int p_{vj}Q_v d\theta_v\right)^2}{\int Q_v d\theta_v}\right] \tag{46a}$$

for $j = 1,\ldots,k$, and

$$\left(I_{p_m}^{\beta\beta^T}\right)_{mj} = -\sum_{t_v=0}^{k}\gamma_{t_v}\left[\int p_{vm}p_{vj}Q_v d\theta_v - \frac{\int p_{vm}Q_v d\theta_v \int p_{vj}Q_v d\theta_v}{\int Q_v d\theta_v}\right] \tag{46b}$$

for $m \neq j = 1,\ldots,k$.

Using (42abc) in (45) gives the ability distribution parameter part:

$$\left(I_{p_m}^{\xi\xi^T}\right)_{11} = -\sum_{t_v=0}^{k}\gamma_{t_v}\left[\int Q_v\left[(\theta_v-\mu)^2\sigma^{-4} - \sigma^{-2}\right]d\theta_v - \frac{\left(\int Q_v(\theta_v-\mu)\sigma^{-2}\,d\theta_v\right)^2}{\int Q_v d\theta_v}\right] \tag{47a}$$

$$\left(I_{p_m}^{\xi\xi^T}\right)_{22} = -\sum_{t_\nu=0}^{k}\gamma_{t_\nu}\left\{\int Q_\nu\left[\frac{3}{4}\sigma^{-4}-\frac{3}{2}(\theta_\nu-\mu)^2\sigma^{-6}+\frac{1}{4}(\theta_\nu-\mu)^4\sigma^{-8}\right]d\theta_\nu - \frac{\left[\int Q_\nu\left[-\frac{1}{2}\sigma^{-2}+\frac{1}{2}(\theta_\nu-\mu)^2\sigma^{-4}\right]d\theta_\nu\right]^2}{\int Q_\nu d\theta_\nu}\right\} \tag{47b}$$

$$\left(I_{p_m}^{\xi\xi^T}\right)_{12} = -\sum_{t_\nu=0}^{k}\gamma_{t_\nu}\left\{\int Q_\nu\left[-\frac{3}{2}(\theta_\nu-\mu)\sigma^{-4}+\frac{1}{2}(\theta_\nu-\mu)^3\sigma^{-6}\right]d\theta_\nu - \frac{\int Q_\nu(\theta_\nu-\mu)\sigma^{-2}d\theta_\nu \int Q_\nu\left[-\frac{1}{2}\sigma^{-2}+\frac{1}{2}(\theta_\nu-\mu)^2\sigma^{-4}\right]d\theta_\nu}{\int Q_\nu d\theta_\nu}\right\}. \tag{47c}$$

Finally, using (43ab) in (45) gives the elements of $I_{p_m}^{\beta\xi^T}$. For $j=1,\ldots,k$

$$\left(I_{p_m}^{\beta\xi^T}\right)_{1j} = -\sum_{t_\nu=0}^{k}\gamma_{t_\nu}\left\{\int Q_\nu p_{\nu j}(\theta_\nu-\mu)\sigma^{-2}d\theta_\nu - \frac{\int Q_\nu p_{\nu j}d\theta_\nu \int Q_\nu(\theta_\nu-\mu)\sigma^{-2}d\theta_\nu}{\int Q_\nu d\theta_\nu}\right\}, \tag{48a}$$

and

$$\left(I_{p_m}^{\beta\xi^T}\right)_{2j} = -\sum_{t_\nu=0}^{k}\gamma_{t_\nu}\left\{\int Q_\nu p_{\nu j}\left[-\frac{1}{2}\sigma^{-2}+\frac{1}{2}(\theta_\nu-\mu)^2\sigma^{-4}\right]d\theta_\nu - \frac{\int Q_\nu p_{\nu j}d\theta_\nu \int Q_\nu\left[-\frac{1}{2}\sigma^{-2}+\frac{1}{2}(\theta_\nu-\mu)^2\sigma^{-4}\right]d\theta_\nu}{\int Q_\nu d\theta_\nu}\right\}. \tag{48b}$$

With respectively (46ab), (47abc), and (48ab), all expressions for the submatrices of the Fisher information matrix (35) are obtained. From these expressions the F-information matrix $I_{p_m}(\beta;\omega)$ can be computed. Because $I_{p_m}^{\xi\xi^T}$ is positive definite, the F-information matrix with respect to β is given by (14), that is,

$$I_{p_m}(\beta;\omega) = I_{p_m}^{\beta\beta^T} - I_{p_m}^{\beta\xi^T}[I_{p_m}^{\xi\xi^T}]^{-1}I_{p_m}^{\xi\beta^T}.$$

F-information in $f(x|t)$

In order to make the relevant comparisons, the expressions for the F-information in the conditional distribution are required. Taking expectations over the distribution of T_v (40) in (31ab) gives:

$$\left(\mathrm{I}_f(\beta;\omega)\right)_{jj} = \left(\mathscr{E}\,\mathrm{I}_f(\beta \mid T)\right)_{jj} = \sum_{t_v=0}^{k} \left\{ e^{-\beta_j}.\gamma_{t_v-1}^{(j)} - \frac{\left[e^{-\beta_j}.\gamma_{t_v-1}^{(j)}\right]^2}{\gamma_{t_v}} \right\} \int Q_v d\theta_v$$

for $j = 1,\ldots,k$, and

$$\left(\mathrm{I}_f(\beta;\omega)\right)_{mj} = \left(\mathscr{E}\,\mathrm{I}_f(\beta \mid T)\right)_{mj} = \sum_{t_v=0}^{k} \left\{ -\frac{e^{-\beta_m-\beta_j}.\gamma_{t_v-1}^{(m)}\gamma_{t_v-1}^{(j)}}{\gamma_{t_v}} + e^{-\beta_m-\beta_j}.\gamma_{t_v-2}^{(mj)} \right\} \int Q_v d\theta_v$$

for $m \neq j = 1,..,k$.

Example

Consider the example similar to the one used in the computations of the information matrices in the Rasch model with fixed ability parameters. For the Rasch model with 3 items, $\beta_1 = -.5$, $\beta_2 = 0.0$ and $\beta_3 = 1.75$, and assuming that θ is distributed as $N(\mu,\sigma^2)$, with $\mu=0$ and $\sigma^2=1$, the expected Fisher information matrix, $(k + \ell) \times (k + \ell)$ (35), for one random draw of the ability distribution is given by:

$$\mathrm{I}_{p_m}(\omega) = \begin{pmatrix} \mathrm{I}_{p_m}^{\beta\beta^T} & \mathrm{I}_{p_m}^{\beta\xi^T} \\ \mathrm{I}_{p_m}^{\xi\beta^T} & \mathrm{I}_{p_m}^{\xi\xi^T} \end{pmatrix} = \begin{pmatrix} .172 & -.027 & -.016 & -.129 & .016 \\ -.027 & .178 & -.018 & -.134 & .002 \\ -.016 & -.018 & .119 & -.085 & -.029 \\ -.129 & -.134 & -.085 & .348 & .011 \\ .016 & .002 & -.029 & .011 & .038 \end{pmatrix}$$

As expected, the upper left 3×3 submatrix $\mathrm{I}_{p_m}^{\beta\beta^T}$ is symmetric, and the coformation elements between the items are negative. This matrix is the information matrix with respect to β, when μ and σ^2 are considered known. In the submatrix $\mathrm{I}_{p_m}^{\beta\xi^T}$, negative coformations between the item parameters and the mean of the ability distribution are obtained, while the coformation between items and the variance of the ability

distribution has no constant sign. In the part of the parameters of the ability distribution, a non-zero coformation between μ and σ^2 can be observed. Note that only in case the mean of the item parameters equals the mean of the ability distribution, the coformation between μ and σ^2 is zero.

The F-information matrices in p_m and in the conditional distribution f are given by

$$\mathrm{I}_{p_m}(\beta;\omega) = \begin{bmatrix} .114 & -.080 & -.034 \\ -.080 & .126 & -.046 \\ -.034 & -.046 & .080 \end{bmatrix} \quad \mathrm{I}_f(\beta;\omega) = \mathcal{E}I_f(\beta \mid T) = \begin{bmatrix} .114 & -.080 & -.034 \\ -.080 & .122 & -.042 \\ -.034 & -.042 & .076 \end{bmatrix}.$$

The efficiency for estimating the item parameters β using the conditional model (CML) instead of the full model (MML) can be computed by dividing the traces of the F-information matrices (17). In this example $\mathrm{EFF}(\beta;f:p_m) = .3115/.3197 = .9744$.

Efficiency comparison CML and MML estimation

The efficiency of CML versus MML will be reported for some typical item parameter sets. In Table 4, the traces of the F-information matrices and the efficiency for 10 items with β =(-3,-2,-1.5,-1,-0.5,0.5,1,1.5,2,3) with varying μ and σ^2 are presented.

It can be seen that the traces of the F-information matrices, and, of course, also the efficiencies, are symmetric in μ. Furthermore it is seen, that, as expected, the traces of $\mathrm{I}_f(\beta;\omega)$ as well as $\mathrm{I}_{p_m}(\beta;\omega)$ decrease, σ^2 is increased, also when the distance between the mean of the ability- and the mean of the item parameters, $|\mu-\Sigma_k\beta_i/k|$, is increased.

Table 4

$\mathrm{tr}(\mathrm{I}_f(\beta;\omega))$, $\mathrm{tr}(\mathrm{I}_{p_m}(\beta;\omega))$ and $\mathrm{EFF}(\beta;f{:}p_m)$ for β=(-3,-2,-1.5,-1,-0.5,0.5,1,1.5,2,3) and ability from $N(\mu,\sigma^2)$

	σ^2			
μ	.25	1	2.25	4
		$\mathrm{tr}(\mathrm{I}_f(\beta;\omega)$		
-2	.976	.939	.894	.843
-1	1.188	1.136	1.061	.976
0	1.255	1.204	1.122	1.024
1	1.888	1.136	1.061	.976
2	.976	.939	.894	.843
		$\mathrm{tr}(\mathrm{I}_{p_m}(\beta;\omega))$		
-2	.980	.947	.903	.854
-1	1.193	1.144	1.073	.989
0	1.258	1.212	1.134	1.039
1	1.193	1.144	1.073	.989
2	.980	.947	.903	.854
		$\mathrm{EFF}(\beta;f{:}p_m)$		
-2	.995	.992	.989	.987
-1	.997	.993	.989	.987
0	.997	.993	.989	.989
1	.997	.993	.989	.987
2	.995	.992	.989	.987

The efficiency of CML versus MML decreases somewhat with increasing σ^2, and decreasing $|\mu - \Sigma_k \beta_i / k|$. However, the efficiency of CML versus MML amounts to at least 98.7%.

In Table 5, the relation between the efficiency and the spread in the item parameters is given. For 1 ability drawn from N(0,1), the efficiency for estimating 10 equidistant item parameters with $\Sigma_i \beta_i = 0$ and varying stepsize, $\beta_{i+1} - \beta_i$, $i = 1, \ldots 9$ is given.

Table 5

$\mathrm{EFF}(\beta;f{:}p_m)$ for $\Sigma_i \beta_i = 0$ with varying stepsize and 1 ability from $N(0,1)$

step	$\mathrm{EFF}(\beta;f{:}p_m)$
1	.9585
.5	.9945
.25	.9994
.125	.9999
.0625	1.0000

Although there is hardly any loss in information, the efficiency of CML versus MML estimation increases with a decrease in the spread between the item parameters.

In Table 6, the relation between the efficiency and the test length is illustrated. For 1 ability drawn from $N(0.5,1)$, the efficiency is reported for estimating the item parameters $\beta=(0,1,2)$, which are n times in a test. The test length is $3n$.

Table 6

EFF$(\beta;f:p_m)$ for $\beta_i=0,1,2$ with varying test length and 1 ability from $N(0.5,1)$

length	EFF$(\beta;f:p_m)$
3	.9822
6	.9983
9	.9993
12	.9996
15	.9998
18	1.0000

An increasing efficiency is observed when the test length is increased. Already with a relative short test of 18 items, CML estimation is as efficient as MML estimation of the item parameters.

Conclusion

In this paper it is shown that the concept of F-information, a generalization of Fisher information, is a useful tool for evaluating the loss of information in conditional maximum likelihood estimation. In separable two-parameter models with a sufficient statistic for the nuisance parameter, the conditional model which only depends on the interest parameter is often used instead of the full two-parameter model for estimating the interest parameter. With the F-information concept it is possible to investigate the conditions under which there is no loss of information in CML estimation, and furthermore, if there is a loss, to quantify this.

In this paper the main properties of F-information are presented and the conditions for no loss of information are specified. Especially in the case of exponential family models, it is shown that these conditions can be easily checked. It is shown that in the Poisson Counts Model these conditions are met. For the Rasch model for dichotomously

scored items these conditions are not fulfilled, which means that loss of information in CML estimation of the item parameters in this model is to be expected.

For the dichotomous Rasch model the expressions needed to be able to investigate the loss of information using CML estimation of the item parameters are derived in detail. For the Rasch model with fixed ability parameters, a comparison was made between JML and CML estimation, and, under the assumption of a normal ability distribution, a comparison of MML and CML estimation. The comparisons were made in several conditions. Varied were: the spread of the item difficulty parameters, the mean and the variance of the ability distribution and the test length. In almost all comparisons, some loss of information in using CML appeared. However, in all the comparisons of CML to JML as well as to MML the loss was very small. The reported efficiencies are always larger than 93%, and for tests with 20 or more items larger than 99%.

On basis of these results it can be concluded that CML item parameter estimation in the Rasch model, which has some other known practical and theoretical attractive properties, is also from an information point of view a sound practice. Hardly any loss of information is to be expected compared to alternative estimation methods.

The method of efficiency comparison described in the paper, applies to any separable two-parameter model. For popular extensions of the dichotomous Rasch model (Fischer & Molenaar, 1995), for example the partial credit model and the one-parameter logistic model, comparable work is in progress.

References

Andersen, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B, 32*, 283-301

Andersen, E.B. (1973). *Conditional inference and models for measuring*. Copenhagen: Mentalhygiejnisk Forlag.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Society, 72*, 355-366.

Bhapkar, V.P. (1989). Conditioning on ancillary statistics and loss of information in the presence of nuisance parameters. *Journal of Statistical Planning and Inference, 21*, 139-160.

Bhapkar, V.P. (1991). Loss of information in the presence of nuisance parameters and partial sufficiency. *Journal of Statistical Planning and Inference, 28*, 185-203.

Efron, B. (1977). On the efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Society, 72*, 557-565.

Engelen, R.J.H. (1989). *Parameter estimation in the logistic item response model.* (Doctoral Thesis.) Enschede: University of Twente.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* [Introduction to mental test theory.] Bern: Huber.

Fischer, G.H., & Molenaar, I.W. (Eds.) (1995). *Rasch Models*. New York: Springer.

Glas, C.A.W. (1989). Contributions to estimating and testing Rasch models. (Doctoral Thesis.) Enschede: University of Twente.

Liang, K. (1983). On information and ancillary in the presence of a nuisance parameter. *Biometrika, 70*, 607-612.

Louis, T.A. (1982). Finding the observed information matrix with the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44*, 226-233.

Molenaar, I.W. (1995). Estimation of item parameters. In: Fischer, G.H., & Molenaar, I.W. (Eds.). *Rasch Models* (pp.39-51). New York: Springer.

Rao, C.R. (1973). *Linear statistical inference and its applications*. New York: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

Wright, B.J. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.

Zwinderman, A.H. (1991). *Studies of estimating and testing Rasch models.* (NICI Technical Report 91-01.) (Doctoral Thesis.) Nijmegen: University of Nijmegen.

Appendix

If the ability distribution $h(\theta_v;\xi)$ is considered as a prior distribution, then the a posteriori distribution of the ability of person v, θ_v, given the score t_v, is by Bayes rule equal to:

$$k(\theta_v \mid t_v;\beta,\xi) =: \frac{g(t_v \mid \theta_v;\beta)\,h(\theta_v;\xi)}{\int_{-\infty}^{\infty} g(t_v \mid \theta_v;\beta)\,h(\theta_v,\xi)d\theta_v} = \frac{Q_v}{\int Q_v\,d\theta_v}.$$

The second derivatives with respect to the item parameters can be rewritten as:

$$\frac{\partial^2 \ln p_m(x_v;\beta,\xi)}{\partial \beta_j^2} = \int_{-\infty}^{\infty} (2p_{vj}-1)\,p_{vj}\,k(\theta_v \mid t_v;\beta,\xi)\,d\theta_v - \left[\int_{-\infty}^{\infty} p_{vj}\,k(\theta_v \mid t_v;\beta,\xi)\,d\theta_v\right]^2 \quad \text{(i)}$$

for $j = 1,\ldots,k$.

and

$$\frac{\partial^2 \ln p_m(x_v;\beta,\xi)}{\partial \beta_m\,\partial \beta_j} = \int_{-\infty}^{\infty} p_{vm}p_{vj}\,k(\theta_v \mid t_v;\beta,\xi)\,d\theta_v -$$

$$\left[\int_{-\infty}^{\infty} p_{vm}\,k(\theta_v \mid t_v;\beta,\xi)\,d\theta_v\right]\left[\int_{-\infty}^{\infty} p_{vj}\,k(\theta_v \mid t_v;\beta,\xi)\,d\theta_v\right] \quad \text{(ii)}$$

for $m \neq j = 1,\ldots,k$.

(i) and (ii) can be written as moments in the a posteriori distribution of θ_v given t_v.

Because $(2p_{vj}-1)p_{vj} = -(1-p_{vj})p_{vj} + p_{vj}^2$, it yields that:

$$\frac{\partial^2 \ln p_m(x_v;\beta,\xi)}{\partial \beta_j^2} = -\mathscr{E}\left(p_{vj}(1-p_{vj}) \mid t_v\right) + \mathrm{Var}(p_{vj} \mid t_v)$$

$$\frac{\partial^2 \ln p_m(x_v;\beta,\xi)}{\partial\beta_m\,\partial\beta_j} = \mathrm{Cov}(p_{vm}, p_{vj} \mid t_v).$$

Evaluating the negatives of these expressions in the estimates of the parameters, the item parameter part of the observed information matrix of the Rasch model using MML estimation is obtained. The same expressions were deduced by Glas (1989, p. 50-55, Appendix D), using the theory of Louis (1982) on the relation between the observed and expected information matrix

Recent Measurement and Research Department Reports:

97-1 H.H.F.M. Verstralen. A Logistic Latent Class Model for Multiple Choice Items.

97-2 N.D. Verhelst. A New Heuristic for Estimating the Reliability of a Test.

97-3 N.D. Verhelst and R.J.H. Engelen. Estimating Latent Abilities from Raw Test Scores.

97-4 G.J.J.M. Straetmans and T.J.H.M. Eggen. Comparison of Test Administration Procedures for Placement Decisions in a Mathematics Course.

97-5 H.H.F.M. Verstralen. A Latent IRT Model for Options of Multiple Choice Items.