Measurement and Research Department Reports 92-2

Estimation and Prediction of Individual Ability in Longitudinal Studies

F. Kamphuis

Cito

Measurement and Research Department Reports 92-2

Estimation and Prediction of Individual Ability in Longitudinal Studies

F.H. Kamphuis

Cito
Arnhem, 1992

Abstract

First, the optimal solution to the estimation or prediction problem of individual ability ('states') is discussed in the framework of the State Space Model (SSM).

Next, the consequences of replacing the linear measurement model in the SSM by a non-linear Item Response Model (IRT) are investigated.

Finally, the possibility of using estimates of latent ability in IRT as pseudo observations in the SSM are explored.

Key words: state space model, item response theory, estimates of individual ability.

The author is indebted to R.J.H. Engelen for his comments on an earlier version of this paper.

Introduction

The task of characterizing growth in level of attainment within an individual student is far from easy. One reason for this is the typical unreliability of educational measures. More structural information concerning the student, such as measurements on previous occasions of the same or other skills and background information, could be used to reduce this uncertainty.

As a general model the State Space Model (SSM) is chosen. The SSM consists of two sets of linear equations: transition (or process) equations and measurement equations. It is assumed that except for the individual abilities or 'states' all parameters in the SSM are known. The question is raised how best to estimate or predict individual ability in this setting. In the second part of the paper, the linear measurement equations in the SSM are replaced by a more proper non-linear measurement model for an educational setting, namely an Item Response Theory (IRT) model. And again the question is raised how best to estimate or predict individual ability in this context.

Estimating and Predicting Individual Ability in the SSM

Having formulated an SSM and estimated the structural parameters of this model, it may be interesting to construct scales of measurements for the latent abilities or states. For example, we might be interested in how individual ability has changed between occasions or we might be interested in the difference in ability between two individuals. The problem in fact is how to use the information present in individual's observation(s) to locate him/her on the latent dimension(s) or state(s). As Bartholomew (1987) has pointed out: " there has been a long and controversial debate about how best to do this". We totally agree with Bartholomew that this debate has often been obscured by talk of 'estimating' the latent abilities as if they were parameters in the ordinary statistical sense. Remembering that the states in the SSM are random variables, it is often forgotten that these variables are still random after the observations have been made. So, we have a prediction problem, not an estimation problem.

In the remainder of this paragraph we will elaborate the optimal solution to the prediction problem. First this will be done for the static case, and then (briefly) for the dynamic case.

For reasons of simplicity, classical test theory, as a sub-model of the SSM, is chosen as an illustration. Results, however, are easily generalized for other SSM.

The Static Case

Consider classical test theory as a sub-model of the SSM. More specific, suppose that the measurement equation is given by

$$y=\eta+\epsilon , \tag{1}$$

where y is an observable test score, η the corresponding 'true score' and the error term ϵ is distributed $N(0,\theta)$ independently of y and η. Furthermore, suppose that the transition equation is given by

$$\eta=a+\zeta , \tag{2}$$

where $\zeta \sim N(0,\Omega)$ and independent of ϵ.
Or more briefly, taking equation (1) and (2) together, y and η are related by the conditional distribution $(y|\eta) \sim N(\eta,\theta)$ and the marginal distribution (or prior distribution) of η is $N(a,\Omega)$.

Now suppose that we are unwilling to make distributional assumptions about η or in fact that we do not know a and Ω. Can we estimate η from one single observation? If we close our eyes for the fact that η is random and look at η as being fixed, the GLS estimator would be the Best Linear Unbiased Estimator (BLUE) of η. In the case of normality, it would also be the ML estimator. If we view the classical test theory model as a one-factor model, this estimator is known as Bartlett's estimator for factor scores (Lawley & Maxwell, 1971). The GLS estimator and associated error covariance in the fixed case are given by

$$\tilde{\eta}=y \quad \text{and} \tag{3}$$

$$P=\theta . \tag{4}$$

The advantage often claimed for this estimator is that it is conditionally unbiased, $E(\tilde{\eta}\,|\,\eta)=\eta$. In other words, if we average $\tilde{\eta}$ over all individuals whose true scores are given by η, this average equals η. To this kind of reasoning one could object that this estimate is not relevant for an individual without replications. Or to put it in other words, with a small sample (n=1) the estimator $\tilde{\eta}$ will always be biased (small sample bias). In addition we neglected the fact that η is a random variable. If we take η is random, $\tilde{\eta}$ as defined above again has some optimal properties, which now must be expressed in terms of the estimation error, $\tilde{\eta}-\eta$. This estimation error has expectation zero and variance θ, equivalent to the properties of the GLS estimator in the fixed case. It can be shown that $\tilde{\eta}$ is best within the class of linear and <u>un</u>conditionally unbiased estimators (extended Gauss-Markov theorem). So there is some evidence to use the GLS estimator in the case that we do not know the mean and the variance of η.

How to proceed if we know the mean and the variance of η? Now the conditional distribution $\eta\,|y$ (or posterior) becomes of interest. Given the prior distribution of η and the conditional distribution $y\,|\eta$, use of standard results from multivariate analysis or application of Bayes' theorem will give us the posterior distribution of η given y. As a prediction of an individual's η we could take the posterior mean, and the estimated true score, $\hat{\eta}$, would then be

$$E(\eta\,|y)=a+\Omega(\Omega+\theta)^{-1}(y-a) \quad . \tag{5}$$

A measure of variation of the η's can be obtained from the posterior variance,

$$P=\Omega-\Omega(\Omega-\theta)^{-1}\Omega \quad . \tag{6}$$

Remarks:

1 The estimator $\hat{\eta}$ as a result of the prediction of the random variable η from y is unconditionally unbiased, i.e., we can see predictors as estimators of future values of random variables.

2 Referring to the posterior variance, it is apparent that we can calculate this variance before any data is collected! More specific, the posterior variance is a statement about the properties of the estimator $\hat{\eta}$ based on the random vector y. If we make many estimates

of η, using measurements y, the sample variance $\eta - \hat{\eta}$ would be approximately equal to the posterior variance (presuming η is known).

3 In classical test theory $\hat{\eta}$ is called Kelley's estimator of the true score and the ratio $\Omega(\Omega + \theta)^{-1}$ is known as the test reliability (Lord & Novick, 1968).

Now lets return to the Minimum Mean Square Estimation (MMSE) or Minimum Variance (MV) estimation. Dropping the assumption of normality, the question is now: Can we incorporate prior information about η in the GLS estimator? To do this, we have to express the prior information in terms of a distribution for $a - \eta$ with mean zero and variance Ω (see Harvey, 1981). The next step is to construct an augmented model

$$\begin{pmatrix} a \\ y \end{pmatrix} = I\eta + \begin{pmatrix} a - \eta \\ \epsilon \end{pmatrix} . \tag{7}$$

Updating the GLS estimator gives

$$\eta^* = P(\Omega^{-1}a + \theta^{-1}y) \quad \text{and} \tag{8a}$$

$$P = (\Omega^{-1} + \theta^{-1})^{-1} . \tag{8b}$$

It is easy to show that the estimators $\hat{\eta}$ and η^* and their associated variances are equal (use the matrix inversion lemma). So, tackling the problem from different directions we end up with the same results. In factor analysis, the estimates η^* are known as 'regression' factor scores (Lawley & Maxwell, 1971).

Suppose now that we have two or more observations. In the prediction setup, the posterior becomes the new prior and in the MMSE approach the GLS estimator is updated to incorporate the new information. For n observations this boils down to

$$\eta^*_n = P_n(\Omega^{-1}a + \theta^{-1}\sum_{i=1}^{n} y_i) \quad \text{and} \tag{9a}$$

$$P_n=(\Omega^{-1}+n\theta^{-1})^{-1} \quad . \tag{9b}$$

This is in fact a recursive linear estimation scheme (Bayesian estimation). If we take η random, in the sense that it is randomly drawn from a prior distribution and keep it fixed before we repeatedly generate the observations in y, the influence of the prior information will die out in the long run when n becomes large. In the limit, assuming inverses exists, equation (9a) will converge to the mean of the observations. In the case η is also redrawn every time an observation y is generated, in the log run, we end up, with the mean of the prior distribution, namely a.

Dynamic Case

The above updating scheme was static in the sense that we were updating the same random vector all the time. Now we consider the case where η changes between measurements according to a specified statistical dynamic, i.e., a discrete random process or time series model. The transition equation is given by

$$\eta_{t+1}=a_t+\phi_t\eta_t+\zeta_t \qquad t=0,1,2,\ldots \quad . \tag{10}$$

The estimation procedure for η follows the same steps as compared to the static case. The only thing we have to do is update the posterior before it becomes the new prior according to the time series model. The equations for these predictions and associated estimation error variances are given by

$$\eta^*_{t+1|t}=a_t+\phi_t\eta^*_{t|t} \quad \text{and} \tag{11a}$$

$$P_{t+1|t}=\phi_t P_{t|t}\phi_t'+\Omega_t \quad . \tag{11b}$$

If we have m measurements, $z_m \triangleq [y_0, y_1, y_2, \ldots, y_m]$, the following estimators for ability can be considered:

$$\eta^*_{s|s}=E(\eta \mid z_s;\omega) \quad \text{and} \tag{12a}$$

$$\eta^*_{s|m}=E(\eta\,|z_m;\omega) \quad \text{for} \quad s<m \;. \tag{12b}$$

Here ω stands for the known parameters in the model. The estimates $\eta^*_{s|s}$ and $\eta^*_{s|m}$ are known as Kalman filtered estimates and Kalman smoothed estimates respectively. For a thorough derivation and discussion of the Kalman filter model and equations for the estimates see Kalman (1960), Harvey (1981), Jazwinsky (1970) or Catlin (1989).

Comment

The estimation or prediction procedure described above typically has a two-level character. In the present case, there are two levels of sampling involved: occasions within individuals, and individuals within the population. Especially it is worthwhile noting that the filtering procedure optimally combines the information in both levels of sampling. Dependent on the magnitude of the information present in the levels, weights will be accordingly attached to the levels for extracting information. We end up with estimates which are optimally in the sense that they minimize the MSE in the population.

As an estimate for an individual's ability, all the estimates discussed above come into consideration. These estimates differ in the amount of information they incorporate. So if we want an estimate for an individual, choose the one which incorporates all the available information. So, a smoothed estimate will be better than a filtered estimate.

IRT as a Measurement Model in the SSM

The linear measurement model in the SSM is now replaced by a more proper non-linear measurement model for an educational setting, namely an IRT model. First, an example of an IRT model will be given, the One Parameter Logistic Model (OPLM), and some of its benefits will be discussed. Then, estimation of an individual's ability in the context of the SSM, with the OPLM as the measurement part, will be discussed. Finally, we discuss estimates of latent ability in IRT models which could be utilized as 'pseudo' observations in the SSM.

Item Response Theory

An IRT model supposes that the ability or latent trait, commonly denoted by θ, is not directly observable. An IRT model specifies a probabilistic relationship between this latent trait and the observable responses on test items of individuals from a well defined population of individuals. Consider, as is appropriate in this example, dichotomously scored items, and we define the random variables X_{ij}, the response of student i (i=1,...,n) on item j (j=1,...,k), with possible values 0 (item is wrong) and 1 (item is correct). The basic equation of the IRT model we use is

$$P(X_{ij}=x_{ij} \mid \theta_i,a_j,\beta_j) = \frac{\exp[a_j(\theta_i - \beta_j)x_{ij}]}{1+\exp[a_j(\theta_i - \beta_j)]} \quad . \tag{13}$$

In this equation β_j is the difficulty parameter and a_j the discrimination index of item j. The model further assumes unidimensionality of the latent trait, which implies local independence (given θ) of the item responses:

$$P(X_{ij}=x_{ij},X_{ik}=x_{ik} \mid \theta_i,a_j,\beta_j)=$$

$$P(X_{ij}=x_{ij} \mid \theta_i,a_j,\beta_j) . P(X_{ik}=x_{ik} \mid \theta_i,a_j,\beta_j) \quad j\neq k \tag{14}$$

and independence of the responses between students:

$$P(X_{i_1j}=x_{i_1j},X_{i_2j}=x_{i_2j}) = P(X_{i_1j}=x_{i_1j}) \, . \, P(X_{i_2j}=x_{i_2j}) \quad i_1\neq i_2 \, . \tag{15}$$

In using an IRT model as the measurement model, two major problems can be distinguished: the first problem is the scaling (or calibration) of the total set of items which measure the ability, whereas the second problem is the measurement of the ability of (groups of) individuals on the established scale. In the calibration phase, the item parameters are estimated and the validity of the model is tested on the basis of responses of students to the items. It is decided in this phase whether there are items which do not follow the model and must be rejected from the item pool and whether the assumptions of the model yield. The procedure we prefer for estimating the item parameters is conditional maximum likelihood estimation (CML). The CML procedure has several attractive features:

a item parameters can be estimated without making any assumptions about the distribution of the latent ability;

b no need for a random sample from a population, i.e., the consistency of the item parameter estimates is not influenced by the sampling method;

c item calibration in incomplete longitudinal designs is possible without complications;

d test statistics at the item level are available which have well known good statistical properties.

With CML, the measurement problem, i.e., estimating item parameters and assessing model fit, is separated from the structural problem, i.e., estimating latent abilities.

The application of CML, however, is restricted to the case in which there exists a sufficient statistic $S_i=S(X_{i1},...,X_{ik})$ for the individual parameter θ_i. The model we use is the One Parameter Logistic Model, which model was introduced and first applied on a large scale in the Dutch National Assessment Program (Verhelst & Eggen, 1989). The model is a special case of the Birnbaum model (Lord & Novick, 1968), from which it differs that in the model the discriminating power of an item is not modelled as a parameter which is to be estimated but as a hypothesized known integer constant. The model belongs to the exponential family, with the weighted sum score $S = \sum_{j=1}^{k} a_j X_j$ as the sufficient statistic for θ_i, which makes it possible to develop sound statistical procedures for estimating and testing the model. In the remainder of the paper we consider the item parameters known.

Individual Ability

Suppose we have a large item bank from which we have constructed a suitable test for each occasion t $(t=1,2...T)$. Furthermore we have registered, for each occasion, the (weighted) test scores, s_t, for an individual drawn form some population with corresponding latent abilities θ_t. How to estimate θ_t?

First, we deal with the cross-sectional case. As before, we have two sources of information. One source of information comes from the population level, θ_t is drawn from some population with known ability distribution $G_\lambda(\theta_t)$ (prior), whereas the other source is from the individual level, the observed test score s_t. As an estimator of individual ability, θ_t^*, we once again take the mean of the posterior

$$E(\theta_t|s_t,\omega,\lambda)=\frac{\int\theta_t\ L(\theta_t|s_t,\omega)G_\lambda(\theta_t)d\theta_t}{\int L(\theta_t|s_t,\omega)G_\lambda(\theta_t)d\theta_t}\ . \tag{16}$$

Here, λ and ω are known parameters, and $\mathbf{L}$ stands for the likelihood of the IRT model. The conditional mean, θ_t^*, called Expected A Posteriori (EAP) estimate of ability (Bock & Mislevy, 1982), in IRT is comparable to the estimate considered in the random case in the linear model as discussed before.

And also for the dynamic case we use the obtained posterior to make an prediction according to the time series model which in turn becomes the new prior to obtain filtered estimates. And finally, as before, smoothed estimates can be obtained using a set of backward recursions with these filtered estimates. For reasons of duplication and space we will not give the expressions for these estimators and associated error covariances.
In summary, replacing the measurement model in the SSM by a non-linear IRT model will not change the estimation procedures for the latent abilities much.

Pseudo Observations

EAP estimates (filtered or smoothed) for individual ability seem the right thing. However, the use of EAP estimates gives some practical disadvantages, especially in combination with time series models. Suppose for instance that we want to use several time series models for an individual, or, in the cross-sectional case, suppose that one wants to choose several priors. The evaluation of the integrals, however, involved in calculating EAP estimates becomes rather expensive. Also, we have to do lot of bookkeeping, since we have to record which items an individual has been presented, during recalculations. Therefore, it would be nice to have some baseline measure which utilizes no prior (population) information and could serve as a pseudo observation in the SSM. Such measure could then be combined with population level information in the way discussed before. There are two candidates in IRT models. The first candidate is the Weighted Maximum Likelihood (WML; Warm, 1989) estimate of ability and the second candidate is the Weighted Expected A Posteriori (WEAP;

Eggen, Engelen & Verhelst, 1992) estimate. Both estimates utilize the weighted likelihood (by the square root of the information) in IRT models, with the WML estimate being the mode and the WEAP estimate being the mean of this function. In both estimation procedures, individual ability is considered as a fixed (unknown) constant (compare section 2).
The measurement model with pseudo observations takes a simple linear form

$$\tilde{\theta}=\theta+\epsilon \qquad \epsilon \sim N(0,VAR(\tilde{\theta})) \ , \tag{17}$$

where $\tilde{\theta}$ is the WML estimate or the WEAP estimate. The transition equations are as before. In order to evaluate the WML and WEAP estimates, a simulation study was carried out. A data set comprising 5 occasions (**T1**,...**T5**) was simulated for 10000 individuals according to a latent first order autoregressive process. Given these abilities, responses were simulated according to the Rasch model (40 items at each occasion). The autoregressive coefficient was chosen equal to .8 throughout the time series. The variances of the latent variables were chosen constant (var=1), whereas the means where -2, -1, 0 , 1 and 2 respectively. The item parameters in the Rasch model for each occasion where drawn uniform on the interval of two standard deviations from the mean ability in the population.

For the cross-sectional, the filtered and the smoothed case the MSE, the bias and the mean standard errors of estimation of true and estimated ability are presented in Table 1,2 and 3. For comparison, the results for the theoretical preferred EAP estimates are also presented.

TABLE 1

Comparison of estimated abilities using EAP, WML and WEAP in the cross-sectional and longitudinal (filtered and smoothed) case with the Mean Squared Errors (MSE) as criterium

CROSS-SECTIONAL	T1	T2	T3	T4	T5
EAP	.12407	.12393	.13570	.12778	.13826
WARM	.14426	.14719	.16047	.14818	.16176
WEAP	.15663	.16248	.17330	.15903	.17040
FILTERED	**T1**	**T2**	**T3**	**T4**	**T5**
EAP	.12407	.10720	.11505	.10938	.11663
WARM	.12832	.11121	.11833	.11322	.11899
WEAP	.12657	.10955	.11694	.11172	.11791
SMOOTHED	**T1**	**T2**	**T3**	**T4**	**T5**
EAP	.10563	.09391	.10051	.09541	.11663
WARM	.10894	.09687	.10357	.09884	.11899
WEAP	.10968	.09741	.10398	.09884	.11791

TABLE 2

Comparison of estimated abilities using EAP, WML and WEAP in the cross-sectional and longitudinal (filtered and smoothed) case with the bias as criterium

CROSS-SECTIONAL	T1	T2	T3	T4	T5
EAP	.00788	.00687	-.00340	.00334	.00005
WARM	.00849	.00757	-.00417	.00320	-.00112
WEAP	.00448	.00151	.00509	.00627	-.00075
FILTERED	**T1**	**T2**	**T3**	**T4**	**T5**
EAP	.00788	.00688	-.00323	.00368	.00042
WARM	.01471	.01636	-.01387	-.00283	-.00153
WEAP	.01194	.01240	-.00892	.00008	-.00053
SMOOTHED	**T1**	**T2**	**T3**	**T4**	**T5**
EAP	.00845	.00671	-.00365	.00290	.00042
WARM	.01521	.01234	-.01360	-.00281	-.00153
WEAP	.01598	.01291	-.01466	-.00310	-.00053

TABLE 3

Comparison of estimated abilities using EAP, WML and WEAP in the cross-sectional and longitudinal (filtered and smoothed) case with the mean standard errors (s.e.) as criterium

CROSS-SECTIONAL					
	T1	T2	T3	T4	T5
EAP	.35080	.35227	.36399	.35840	.36601
WARM	.38206	.38476	.39742	.39028	.39826
WEAP	.38322	.38611	.39842	.39117	.39880
FILTERED					
	T1	T2	T3	T4	T5
EAP	.35080	.32766	.33551	.33157	.33732
WARM	.35553	.33103	.33831	.33439	.33974
WEAP	.35623	.33158	.33870	.33477	.33999
SMOOTHED					
	T1	T2	T3	T4	T5
EAP	.32473	.30657	.31295	.31091	.33732
WARM	.32845	.30921	.31518	.31323	.33974
WEAP	.32897	.30964	.31549	.31354	.33999

The results show that the WEAP estimate and the WML estimate could serve as pseudo observations. The loss in MSE for the WEAP and WML, in comparison with the preferred EAP, is minor. The WEAP outperforms the WML slightly. The reason we prefer the WML estimate as a base line measurement instead of the WEAP estimator, is because of the additional feature of conditional unbiasedness (in the fixed case), which is convenient in calculating groups means.

Summary

In the first part of this paper, the estimation or prediction of individual ability in the SSM was discussed. The main point stated there was that the states in the SSM are random variables. Therefore, we must realize that always two (or more) levels of sampling are involved: occasions within individuals and individuals within the population. So, in our opinion it is necessary, in assigning scores to an individual, whether we call this a prediction or an estimation problem, to make use of information concerning both levels of sampling.

In the second part of the paper, the linear measurement model of the SSM was replaced by a non-linear IRT model. Following the setup developed in the first part of the paper, estimators for the latent abilities could be derived in the same fashion.

In addition we discussed estimates of latent ability in IRT models which could be utilized as 'pseudo' observations in the SSM. It turned out that both candidates, WML and WEAP estimates, could serve, with minor loss of information, as such pseudo observations. Throughout the paper it was assumed that all the parameters in the SSM were known, except of course for the latent abilities. How to estimate these structural parameters will be discussed elsewhere.

References

Bartholomew, D.J.(1987). *Latent variable models and factor analysis*. London: Charles Griffin.

Bock, R.D. and Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Catlin, D.E.(1989). *Estimation control, and the discrete Kalman filter*. New York: Springer Verlag.

Eggen, T.J.H.M., Engelen, R.J.H. and Verhelst, N.D. (1992). *Weighted expected a posteriori estimation of ability in the Rasch model*. Arnhem: Cito (unpublished paper).

Harvey, A.C.(1981). *Time series models*. Oxford: Phillip Allan.

Jazwinsky, A.H.(1970). *Stochastic processes and filtering theory*. New York: Academic Press.

Kalman, R.E.(1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, Transactions ASME. Series D*, *82*, 35-45.

Lawley, D.N. and Maxwell, A.E.(1971). *Factor analysis as a statistical method*. London: Butterworths.

Lord, F.M. and Novick, M.R.(1968). *Statistical theories of mental test scores*. Reading (Mass.): Addison-Wesley.

Verhelst, N.D. and Eggen, T.J.H.M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek*. Arnhem: Cito.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, *54*, 427-450.

Recent Measurement and Research Department Reports:

91-1 N.D. Verhelst & N.H. Veldhuijzen. A New Algorithm for Computing Elementary Symmetric Functions and Their First and Second Derivatives.

91-2 C.A.W. Glas. Testing Rasch Models for Polytomous Items: With an Example Concerning Detection of Item Bias.

91-3 C.A.W. Glas & N.D. Verhelst. Using the Rasch Model for Dichotomous Data for Analyzing Polytomous Responses.

91-4 N.D. Verhelst & C.A.W. Glas. A Dynamic Generalization of The Rasch Model.

91-5 N.D. Verhelst & H.H.F.M. Verstralen. The Partial Credit Model with Non-Sequential Solution Strategies.

91-6 H.H.F.M. Verstralen & N.D. Verhelst. The Sample Strategy of a Test Information Function in Computerized Test Design.

91-7 H.H.F.M. Verstralen & N.D. Verhelst. Decision Accuracy in IRT Models.

91-8 P.F. Sanders & T.J.H.M. Eggen. The Optimum Allocation of Measurements in Balanced Random Effects Models.

91-9 P.F. Sanders. Alternative Solutions for Optimization Problems in Generalizability Theory.

91-10 N.D. Verhelst, H.H.F.M. Verstralen & T.J.H.M. Eggen. Finding Starting Values for the Item Parameters and Suitable Discrimination Indices in the One-Parameter Logistic Model.

92-1 N.D. Verhelst, H.H.F.M. Verstralen & M.G.H. Jansen. A Logistic Model for Time Limit Tests.