Course LNIRT: Modeling Response Accuracy and Response Times

Jean-Paul Fox

April, 2018

Introduction

When computerized tests are administered, next to response accuracy (RA) response times (RTs) can be automatically recorded. The information in the response times can help to improve routine operations in testing, such as item calibration, adaptive item selection, latent ability estimation, as well as to explore and measure factors that influence the performances on the test.

The issue of how to model response times has been approached from three different angles. One approach is to model the response times with time parameters added to a regular item response theory (IRT) model (see, e.g., Roskam, 1997; Thissen, 1983; and Verhelst *et al.*, 1997). A second approach is characterized by modeling the response times separately from the responses (see, e.g., Maris, 1993, and Scheiblechner, 1979). Van der Linden (2006) discussed a selection of these models for response times on test items. In a third approach, introduced in van der Linden (2007), the response times and responses are modeled hierarchically. At the first level, both the distributions of response accuracy and response times are assumed to follow separate models, each with a different set of person and item parameters. The person parameters represent the speed and accuracy (or ability) of the test taker on the items. A test taker's choice of speed and accuracy is generally constrained by a tradeoff. At this first level of modeling, the RTs and RA can be assumed to be conditionally independently distributed given the speed and accuracy parameters, respectively. However, at the second level, these parameters are allowed to be dependent. This leads to a hierarchical modeling framework in which the relation between speed and accuracy is defined at a higher level of modeling.

Response times have a natural lower bound at zero, the logarithm of RTs is modeled, and their distribution is assumed to be normal. The choice of a lognormal distribution is a classic one in response-time research. For response times on test items, this assumption was made earlier by, for example, Thissen (1983), Schnipke and Scrams (1997), and van der Linden *et al.* (1999). Each of these studies showed a good fit of response times to a lognormal distribution. Both the binomial distribution of response accauracy and the normal distribution of the log response times can be given a traditional item-response theory (IRT) parameterization. The binomial parameter for response accuracy has the structure of the two-parameter normal-ogive model (Lord and Novick, 1968). The distribution of the response times has a parameterization close to that of an IRT model for continuous response data (see, e.g., Samejima, 1973; Shi and Lee, 1998). RA and RTs are conditionally independently distributed. Their joint distribution is the product of a binomial and a normal distribution. This defines the level-1 model of the joint model for the analysis of RTs and RA for measuring test takers's speed and ability on test items, respectively.

In our Bayesian approach, a Gibbs sampler is used for estimating the model parameters. The approach facilitates the use of informative proper priors. The Gibbs sampler was programmed in R with an R-package of functions called *LNIRT*\footnote{LNIRT Version 0.3.0. https://cran.r-project.org/web/packages/LNIRT. This R-package enables users to model patterns of response accuracy and response times using a joint model, and to estimate and examine the model fit. A brief overview of procedures for testing the fit of the model is given. The R-package *LNIRT* is described, where the description includes a listing of the input and output variables.

The Joint Modeling Approach

A hierarchical modeling procedure is followed. At Level 1, separate measurement models are defined for the response accuracy and response times. At level 2, a distributional structure is defined for the level-1 model parameters. Subsequently, hyperprior distributions are specified for the prior parameters.

Level 1

Item responses to a set of items indexed $k = 1, \ldots, K$ are taken to be stored in an $N \times K$ data matrix $\mathbf{Y}$. The response patterns are characterized by both the test takers and the items. A two-parameter IRT model is used to define a mathematical relationship between the probabilities of the responses and the person and item parameters (see, e.g., Lord and Novick, 1968). Let θ_i denote the ability of test taker i. Then, the probability of a correct response to item k is defined as:

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k) = \Phi(a_k\theta_i - b_k), \tag{1}$$

where a_k and b_k are generally known as the discrimination parameter and difficulty parameter of item k, respectively, and Φ denotes the normal cumulative distribution function. When defining the item difficulties on the same scale as the ability scale, additional brackets need to be placed in the mean component. The probability of a correct response to item k is given by,

$$P\left(Y_{ik} = 1 \mid \theta_i, a_k, b_k\right) = \Phi\left(a_k\left(\theta_i - \tilde{b}_k\right)\right), \tag{2}$$

where θ_i and b_k are defined on the same scale. In LNIRT, both parameterizations are implemented. Note that the item difficulty parameters in Equation (1) and Equation (2) are not directly comparable, and are defined on different scales. For the three-parameter model, a guessing parameter c_k is introduced, this leads to the following success probability

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k, c_k) = c_k + (1 - c_k)\Phi(a_k\theta_i - b_k), \tag{3}$$

where c_k is the probability of guessing item k correctly.

Response-time distributions have a natural lower-bound at zero and, for that reason, are skewed to the right. A lognormal distribution is used to model the response times which are taken to be stored in an $N \times K$ matrix $\mathbf{RT}$. It is assumed that each respondent chooses to complete the items at a constant speed that can be represented by a parameter denoted as ζ_i. The time needed to complete an item also depends on item characteristic parameters. They are denoted as ϕ_k and λ_k, and can be seen as a discrimination and time-intensity parameter, respectively. The logarithm of the response times, RT_{ik}, are assumed to be normally distributed, and it follows that,

$$\begin{aligned} RT_{ik} &= \lambda_k - \phi_k\zeta_i + \epsilon_{ik} && (4)\\ \epsilon_{ik} &\sim N\left(0, \sigma^2_{\epsilon_k}\right), && (5) \end{aligned}$$

where the time intensity parameter λ_k represents the average time needed to complete the item (on a logarithmic scale), the speed parameter, ζ_i, represents the working speed of test taker i, and the time discrimination parameter, ϕ_k, the item-specific effect of working speed on the RT. Increasing the time intensity λ_k leads to a positive shift of the location of the time distribution on the item. Likewise, an increase in the speed parameter ζ_i leads to a negative shift. In the same way as for the item response model in Equation (2), the time intensities can be defined on the same scale as the speed parameter. It follows that,

$$RT_{ik} = \phi_k \left(\tilde{\lambda}_k - \zeta_i\right) + \epsilon_{ik}, \tag{6}$$

where the time discrimination parameter operates on the term $\lambda_k - \zeta_i$.

Fox *et al.* (2007); Klein Entink *et al.* (2008) introduced the time-discrimination parameter as a slope parameter for speed, which models the sensitivity of the item for different speed-levels of the test takers. This specification of the item time discrimination parameter differs from the time discrimination parameter defined by van der Linden (2007).In his approach, the reciprocal of the standard deviation of the measurement error is defined to be the time discrimination. This also allows for item-specific variances. However, the time discriminations in Equation (4) also model covariances between RTs. When considering responses to item k and l, the covariance between RTs of test taker i is given by,

$$\begin{aligned} cov\left(RT_{ik}, RT_{il}\right) &= cov\left(\lambda_k - \phi_k\zeta_i, \lambda_l - \phi_l\zeta_i\right) \\ &= cov(-\phi_k\zeta_i, \phi_l\zeta_i) \\ &= \phi_k var(\zeta_i)\phi_l \end{aligned}$$

when assuming independent errors and time intensities. So, the covariance between the two RTs is influenced by both time discriminations. Furthermore, the additional error term in Equation (4) can model variations in RTs due to stochastic behavior of the test taker. When test takers operate with different speed values, take small pauses during the test, or change their time management, the RTs might show more systematic variation than explained by the structural mean term. The item-specific error component might accommodate for these differences and avoid bias in the parameter estimates.

As already noted, modeling response times as a lognormal distribution is a classic choice. A lognormal model with a simpler decomposition of the mean parameter was proposed by Schnipke and Scrams (1997). However, their model was not used to describe the distribution of the response time for a fixed person and item but as a convenient summary of the empirical distributions of the times on the items in a bank across its history of test takers. The parameterization in (4) corresponds closely to that of the two-parameter IRT model for continuous responses developed by Samejima (1973).

An implicit assumption of the response time model is that the speed parameter remains constant during the test. This means that, whatever the conditions under which the test is taken, the test takers are assumed to settle on a level of speed at the beginning of the test and then stick to it. The joint model is unable to deal with changes in speed, for example, due to fatigue or the adoption of a new strategy during the test.

Levels 2 and 3

A bivariate normal distribution is defined for the ability and speed parameters of the test takers,

$$\left(\theta_i, \zeta_i\right) \sim \mathcal{N}_2\left(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P\right)$$

where

$$\begin{aligned} \boldsymbol{\mu}_P &= \left(\mu_\theta, \mu_\zeta\right) \\ \boldsymbol{\Sigma}_P &= \begin{pmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{pmatrix}. \end{aligned}$$

Parameter ρ denotes the covariance between the person parameters. The level-2 model for speed and ability can be considered to represent a population of persons that take the test. The distribution can then be seen

as the sampling distribution of a random test taker from the population. From a Bayesian perspective, the test takers are defined to be exchangeable, and the distribution represents the common prior for the person parameters. As a hyperprior for the covariance matrix $\boldsymbol{\Sigma}_P$, an inverse-Wishart distribution with degrees of freedom ν_P and scale parameter V_P is chosen. In the same way, a multivariate normal distribution is specified for the item parameters of the level-1 models,

$$\big(a_k, b_k, \phi_k, \lambda_k\big) \sim \mathcal{N}\big(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I\big), \tag{7}$$

where a_k and ϕ_k are restricted to be positive. This assumption allows for the fact that the item parameters within each measurement model usually correlate. The covariance matrix for the item parameters also define a correlation between the item parameters from the level-1 models. This feature models the fact that more difficult items typically require more time to complete than relatively easy items. The full covariance matrix for the item parameters is given by

$$\begin{aligned} \boldsymbol{\Sigma}_I &= \begin{pmatrix} \boldsymbol{\Sigma}_{a,b} & \boldsymbol{\Sigma}_{(a,b),(\phi,\lambda)} \\ \boldsymbol{\Sigma}_{(\phi,\lambda),(a,b)} & \boldsymbol{\Sigma}_{\phi,\lambda} \end{pmatrix} && (8) \\ &= \begin{pmatrix} \sigma_a & \sigma_{a,b} & \sigma_{a,\phi} & \sigma_{a,\lambda} \\ \sigma_{b,a} & \sigma_b & \sigma_{b,\phi} & \sigma_{b,\lambda} \\ \sigma_{\phi,a} & \sigma_{\phi,b} & \sigma_\phi & \sigma_{\phi,\lambda} \\ \sigma_{\lambda,a} & \sigma_{\lambda,b} & \sigma_{\lambda,\phi} & \sigma_\lambda \end{pmatrix}. && (9) \end{aligned}$$

As a hyperprior for $(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$, a normal-inverse-Wishart distribution is chosen. That is,

$$\begin{aligned} \boldsymbol{\Sigma}_I &\sim Inv-Wishart_{\nu_I}\big(V_I^{-1}\big) && (10) \\ \boldsymbol{\mu}_I \mid \boldsymbol{\Sigma}_I &\sim \mathcal{N}\big(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_I/\kappa\big), && (11) \end{aligned}$$

where ν_I and V_I are the degrees of freedom and scale matrix of the inverse Wishart distribution, $\boldsymbol{\mu}_0$ is the prior mean and κ the number of prior measurements.

A Beta prior is specified for the guessing parameter, c_k, where the default hyperparameter values are 2 and 6. This leads to a prior proportion of guessing of 1/7 with a variance of .02.

The hierarchical structure induces a shrinkage estimation method for all measurement-model parameters. In fact, the two covariance structures introduce a relationship between the RA and RT data. The simultaneous estimation procedure includes the available collateral information about each of the parameters: The RTs serve as collateral information that is used to estimate the parameters of the response model. Conversely, RA are used as collateral information when estimating the parameters of the response-time model (van der Linden *et al.*, 2010).

Explanatory Variables

The multivariate models for persons and item parameters can be extended to include explanatoray variables. Let $\mathbf{X}_\theta$ denote the predictors for the ability parameter and $\mathbf{X}_\zeta$ for the speed parameter. The mean component for the person parameters can be expressed as

$$\begin{aligned} \mu_\theta &= \mathbf{X}_\theta \boldsymbol{\beta}_\theta \\ \mu_\zeta &= \mathbf{X}_\zeta \boldsymbol{\beta}_\zeta. \end{aligned}$$

For the mean component of the item parameters a similar extension can be defined,

$$
\begin{aligned}
\mu_a &= \mathbf{X}_a \boldsymbol{\beta}_a \\
\mu_b &= \mathbf{X}_b \boldsymbol{\beta}_b. \\
\mu_\phi &= \mathbf{X}_\phi \boldsymbol{\beta}_\phi. \\
\mu_\lambda &= \mathbf{X}_\lambda \boldsymbol{\beta}_\lambda.
\end{aligned}
$$

Explanatory information can be included to explain differences between persons and item characteristics. Noninformative normal priors are defined for the regression parameters with a mean of zero and a large variance. In the *LNIRT* software, the predictors for discrimination and time-discrimination are not implemented.

Estimation and Identification

Two-parameter IRT models are usually identified by fixing the mean and variance of the latent scale to zero and one, respectively. Typically, this can be done directly by setting the prior mean, μ_θ and variance σ^2_θ equal to a fixed value, or by putting restrictions on the item parameters.

The joint model can be identified in the same way; the restrictions are now imposed on the mean vector and a covariance matrix. For example, it is sufficient to set $\boldsymbol{\mu}_P = 0$ and $\sigma^2_\theta = 1$, and $\prod_k \phi_k = 1$. The first restriction sets the mean of the speed and ability parameters equal to zero, which implies that the mean of the time-intensity parameters of the items is equated to the mean log response times and the mean ability is absorbed in the mean of the item difficulties, respectively. The prior for the covariance matrix is modified, when elements of the covariance matrix are restricted. The inverse-Wishart distribution does not apply to a restricted covariance matrix.

The proposed procedure is Bayesian estimation of all parameters through Gibbs sampling of their joint posterior distribution. The procedure involves the division of all unknown parameters into blocks, with iterative sampling of the conditional posterior distributions of the parameters in each block given the preceding draws for the parameters in all other blocks (Fox, 2010).

Examples

A simulate data function (simLNIRT) can be used to generate data for the joint model. A few examples can be given to illustrate the options. First, data are simulated for $N = 500$ persons and $K = 10$ items, where the correlation between abiltiy and speed is assumed to be .7. The remaining input are given their default values, which means that the time discrimination is restricted to be 1, and no explanatory variables are simulated. The variance of the ability and speed parameter is set to one.

Example: Log-Normal Response Times

```
library(LNIRT)
N <- 500
K <- 10
rho <- 0.7
data <- simLNIRT(N=500,K=10,rho=0.7)
```

The object data contains an object $data\$Y$, which is simulated according to the parameterisation in Equation (1) and $data\$Y1$, according to Equation (2). Furthermore, objects $data\$Yg$ and $data\$1g$ are created, which are the corresponding simulated data objects with 10% guessing. The object data also contains simulated resonse times on a logarithmic scale, $data\$RT$, according to Equation (4), and $data\$RT1$ according to Equation (6).

The object *data*$*theta* contains the simulated persons parameters ability and speed, and the object *data*$*ab* contains the simulated discrimination, difficulty, time discrimination, and time intensity parameters (column-wise also in this order), respectively. The simulated measurement error variances are stored in the object *data*$*sigma*2. The product of discriminations and time discriminations is set to one, and the mean of the difficulty and time intensity parameters is set to zero.

The parameters of the response time model can be estimated using the LNRT function, which only considers the log of response times as data input. The default states that time discriminations are estimated as well as item-specific measurement error variances. When including the simulated data object, the summary report will also report the true simulated parameter values.

```
XG <- 5000 #number of MCMC iterations
out <- LNRT(RT=data$RT,data=data,XG=5000)
```

```
#report output
summary(out)
```

```
##
##  Log-Normal RT Modeling, 2013, J.P. Fox
##  Summary of results
##
##   Time Discrimination          Time Intensity           Measurement Error Variance
##   item    EAP     SD      Sim     item    EAP     SD      Sim     EAP     SD      Sim
##
##   1   1.171   0.049   1.187   1   0.768   0.048   0.788   1.140   0.080   1.120
##   2   1.292   0.054   1.147   2  -0.012   0.053  -0.013   1.430   0.100   1.375
##   3   0.969   0.053   0.936   3   0.841   0.052   0.762   1.370   0.092   1.280
##   4   0.680   0.059   0.707   4  -0.426   0.063  -0.368   1.974   0.127   1.787
##   5   1.329   0.045   1.317   5  -1.111   0.040  -1.168   0.791   0.061   0.778
##   6   0.687   0.039   0.756   6  -0.475   0.040  -0.467   0.795   0.053   0.828
##   7   0.811   0.043   0.808   7  -0.058   0.045  -0.051   1.001   0.068   0.959
##   8   1.172   0.040   1.197   8   1.330   0.036   1.407   0.650   0.051   0.755
##   9   0.784   0.041   0.760   9   0.284   0.041   0.314   0.835   0.057   0.822
##   10      1.492   0.045   1.515   10     -1.197   0.040  -1.205   0.802   0.064   0.749
##
##   Mean and Covariance matrix Items (phi,lambda)
##
##   --- Population Mean Item ---
##   mu_phi   SD     mu_lam   SD
##   1.033    0.349 -0.009    0.624
##
##   --- Covariance matrix Items ---
##   phi     SD      Cov     SD      lambda   SD
##   0.781   0.542  -0.045   0.786   4.417   2.593
##
##   Mean and Covariance matrix Persons
##
##   --- Population Mean Person ---
##   mu_P     SD
##    0.00    0.04
##
##   --- Covariance matrix Person ---
##   Sigma_P SD
##   1.009   0.072
```

The summary report provides for each item the time discrimination, time intensity, and measurement eror

variance estimates with the estimated posterior standard deviations. Subsequently, in Figure 1, the estimated values are plotted against the simulated values. The object $out\$MAB$ contains the simulated parameter values for the time discriminations and time intensities. The default burn-in period for the summary report is set to 10% of the total number of MCMC iterations.

```
par(mfrow=c(2,1))
plot(apply(out$MAB[500:XG,,1],2,mean),data$ab[,3],xlab="estimated",
     ylab="simulated",main="Time Discrimination",bty="l",
     pch=15,cex=.75,cex.main=.8,cex.axis=.7,cex.lab=.8,
    ylim=c(min(data$ab[,3])-.5,max(data$ab[,3])+.5),
    xlim=c(min(data$ab[,3])-.5,max(data$ab[,3])+.5))
abline(0,1)
plot(apply(out$MAB[500:XG,,2],2,mean),data$ab[,4],xlab="estimated",
     ylab="simulated",main="Time Intensity",bty="l",
     pch=15,cex=.75,cex.main=.8,cex.axis=.7,cex.lab=.8,
    ylim=c(min(data$ab[,4])-.5,max(data$ab[,4])+.5),
    xlim=c(min(data$ab[,4])-.5,max(data$ab[,4])+.5))
abline(0,1)
```

The same simulation study can be done using the parameterization with the reciprocal of the standard deviation of the measurement error as the time discrimination. Therefore, set $WL = TRUE$ as the indicator for this parameterization. When including the simulated data object, it will use the variable RT in that data object.

```
set.seed(1234)
data <- simLNIRT(N=500,K=10,rho=0.7,WL=TRUE)
out <- LNRT(RT,data=data,XG=5000,WL=TRUE)
summary(out)
```

In Figure 2, the results are plotted in the same way, where the reciprocal of the squared estimated time discriminations are plotted against the simulated measurement error variance.

```
par(mfrow=c(2,1))
plot(1/(apply(out$Msigma2[500:XG,],2,mean))**2,data$sigma2,xlab="estimated",
     ylab="simulated",main="Measurement Error Variance",bty="l",
     pch=15,cex=.75,cex.main=.8,cex.axis=.7,cex.lab=.8,
    ylim=c(min(data$sigma2)-.5,max(data$sigma2)+.5),
    xlim=c(min(data$sigma2)-.5,max(data$sigma2)+.5))
abline(0,1)
plot(apply(out$MAB[500:XG,,2],2,mean),data$ab[,4],xlab="estimated",
     ylab="simulated",main="Time Intensity",bty="l",
     pch=15,cex=.75,cex.main=.8,cex.axis=.7,cex.lab=.8,
    ylim=c(min(data$ab[,4])-.5,max(data$ab[,4])+.5),
    xlim=c(min(data$ab[,4])-.5,max(data$ab[,4])+.5))
abline(0,1)
```

A residual analysis can be performed by including the argument $residual = TRUE$. At least 1000 MCMC iterations are needed, since the residual analysis will only use the sampled values after 1000 iterations.

```
set.seed(1234)
data <- simLNIRT(N=500,K=10,rho=0.8)
out <- LNRT(RT=data$RT,data=data,XG=5000,residual=TRUE)
```

```
summary(out)
```

```
##
##  Log-Normal RT Modeling, 2013, J.P. Fox
```

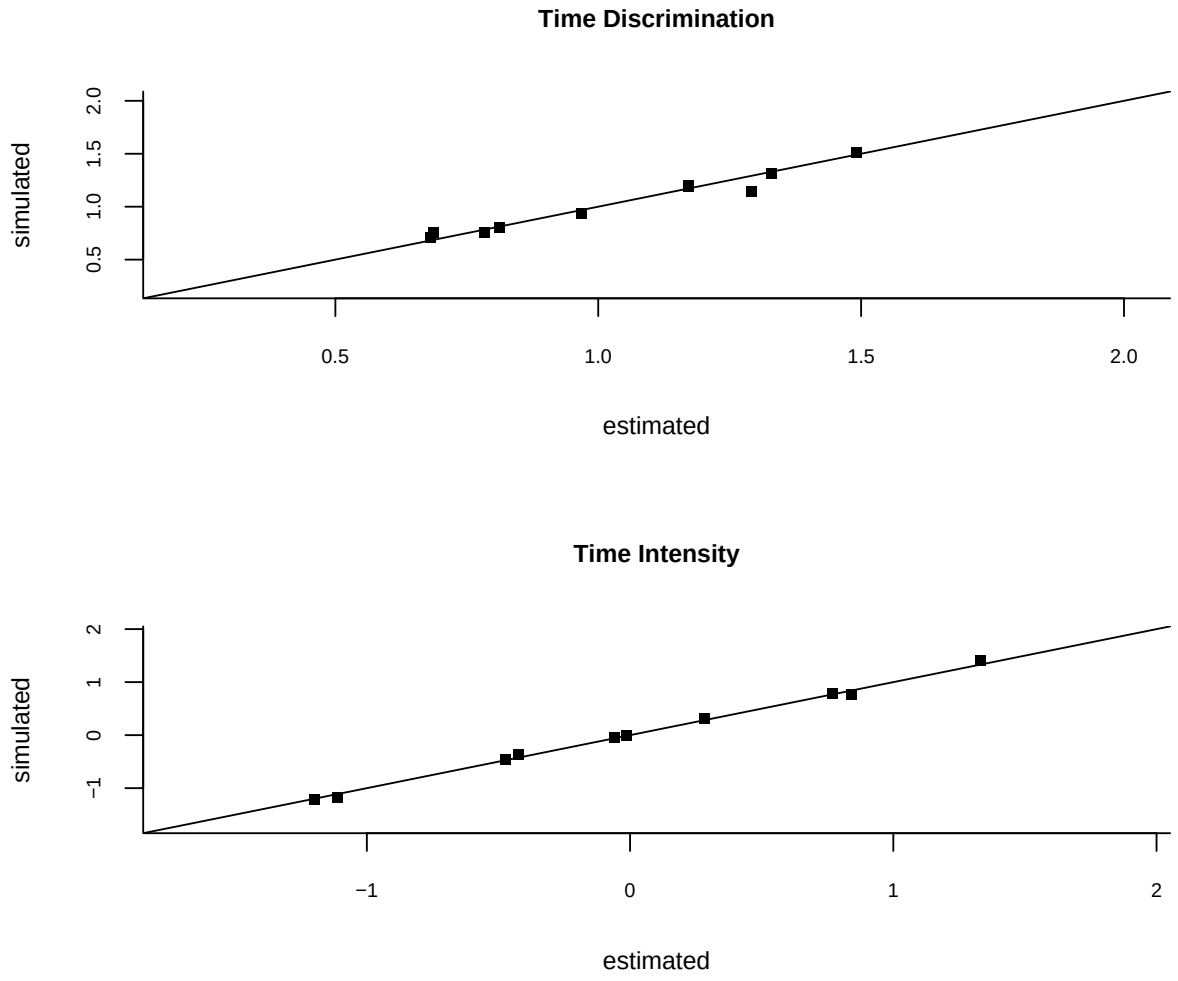

Figure 1: Estimated time discriminations and time intensities against the simulated values.

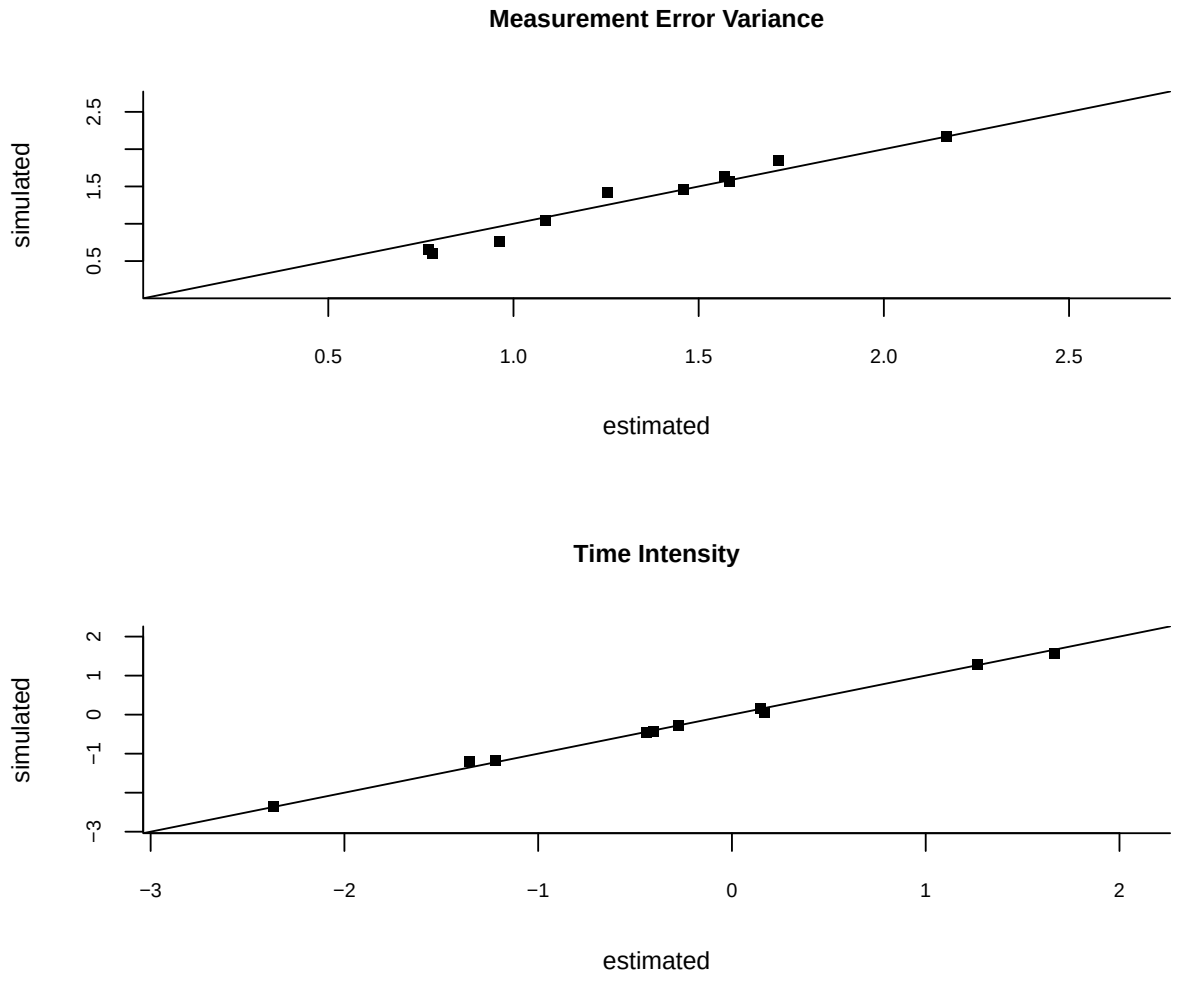

Figure 2: Estimated measurement error variances and time intensities against the simulated values

```
##  Summary of results
##
##   Time Discrimination          Time Intensity           Measurement Error Variance
##   item    EAP     SD      Sim     item    EAP     SD      Sim     EAP     SD      Sim
##
##   1   1.384   0.042   1.357   1   1.550   0.036   1.563   0.655   0.054   0.697
##   2   0.675   0.045   0.747   2   1.917   0.046   1.848   1.065   0.071   1.095
##   3   0.811   0.035   0.809   3  -1.006   0.034  -0.918   0.590   0.041   0.630
##   4   0.953   0.048   0.909   4  -0.124   0.050  -0.155   1.232   0.083   1.210
##   5   0.866   0.047   0.861   5   0.432   0.049   0.429   1.210   0.082   1.235
##   6   0.885   0.035   0.878   6   0.410   0.034   0.349   0.592   0.042   0.565
##   7   1.365   0.053   1.416   7  -0.154   0.052  -0.162   1.356   0.097   1.325
##   8   1.081   0.045   1.074   8  -2.084   0.045  -2.059   0.991   0.069   0.935
##   9   0.957   0.039   0.924   9   0.008   0.038   0.004   0.746   0.053   0.817
##   10      1.291   0.051   1.265   10     -0.946   0.051  -0.899   1.292   0.090   1.143
##
##   Mean and Covariance matrix Items (phi,lambda)
##
##   --- Population Mean Item ---
##   mu_phi   SD     mu_lam   SD
##   1.026    0.381  0.008    0.667
##
##   --- Covariance matrix Items ---
##   phi     SD      Cov     SD      lambda   SD
##   0.795   0.668  -0.057   0.912   5.241   2.845
##
##   Mean and Covariance matrix Persons
##
##   --- Population Mean Person ---
##   mu_P     SD
##    0.00    0.04
##
##   --- Covariance matrix Person ---
##   Sigma_P SD
##   1.000   0.072
##
##
##   *** Person Fit Analysis ***
##
##   Percentage Outliers Persons (5% level)
##
##   lZ
##   2.6 %
##   95% Posterior Probability:  2 %
##
##
##   *** Item Fit Analysis ***
##
##   Misfitting Items (5% level)
##   No Misfitting Items
##
##   *** Residual Analysis ***
##   No Extreme Residuals
##
```

```
##   Kolmogorov Smirnov Test (5% level)
##   0 %
```

Marianti *et al.* (2014) and Fox and Marianti (2017) developed a person-fit statistic to identify extreme RT patterns. The test is referred to as the lZ, which is known to be chi-squared distributed given the model's parameter values. The object $out\$lZP$ contains the estimated posterior probability to observe a more extreme lZ value then the estimated value. This is a posterior predictive test. In the summary report, the percentage of RT patterns with a posterior probability of significance of less than 5% is printed under the label lZ. The object $out\$EAPCP$ represents for each pattern the posterior probability that the pattern is flagged to be extreme given significance level of 5%. In the summary report, the percentage of patterns flagged with a posterior probability of more than 95% to be extreme is reported. It can be seen that around 2.6% of the patterns have an estimated significance probability of less than 5%, and around 2% of the patterns were flagged as extreme.

A significance posterior probability is also computed for each vector of RT observations, and the proportion of items with a significance probability of less than 5% is reported. In the summary report, no items were reported as misfitting. A standardized residual is computed and the proportion of residuals with a significance probability of less than 5% is computed and reported. Finally, the Kolmogorov-Smirnov test is used to compare the empirical distribution of the residuals to the normal distribution. The percentage of items showing significant violations (i.e., significance probability is less than 5%) of normality is reported. In Figure 3, a plot of the estimated person-fit statistics with respect to the RT patterns against the corresponding posterior significance probability can be given, where the extreme RT patterns are marked in red.

```
plot(out$lZPT,out$lZP,xlab=expression(paste(l[z])),
     cex=.75,cex.main=.8,cex.axis=.7,cex.lab=.8,
     ylab="Bayesian significance level",pch=1,bty="l")
set <- which(out$lZP < .05)
points(out$lZPT[set],out$lZP[set],col="red",pch=16,cex=.75)
```

Finally, predictor variables can be simulated for the time intensities and the speed parameter. In the summary report, the posterior mean regression effects and the posterior standard deviations are reported. The following R-code shows the simulation of RT data with a predictor for time intensity and speed.

```
data <- simLNIRT(N=500,K=10,rho=0.7,kpt=1,kit=1)

out <- LNRT(RT=data$RT,data=data,XG=5000,residual=FALSE,XPT=data$XPT,XIT=data$XIT)
```

Example: Response Accuracy and Response Times

The simulated data object from simLNIRT also contains the accuracy data. The joint model can be fitted by making a call to LNIRT. The default considers the two-parameter IRT model in Equation (1), with the log-normal response time model represented in Equation (4), and the multivariate priors for the person and item parameters.

```
set.seed(1234)
data <- simLNIRT(N=500,K=10,rho=0.8)
out1 <- LNIRT(RT,Y,data=data,XG=5000)
```

```
summary(out1)
```

```
##
##  Log-Normal RT-IRT Modeling, 2013, J.-P. Fox
##  Summary of results
##
##   Item Discrimination parameter   Item Difficulty parameter
##   item    EAP     SD      Sim     item    EAP     SD      Sim
```

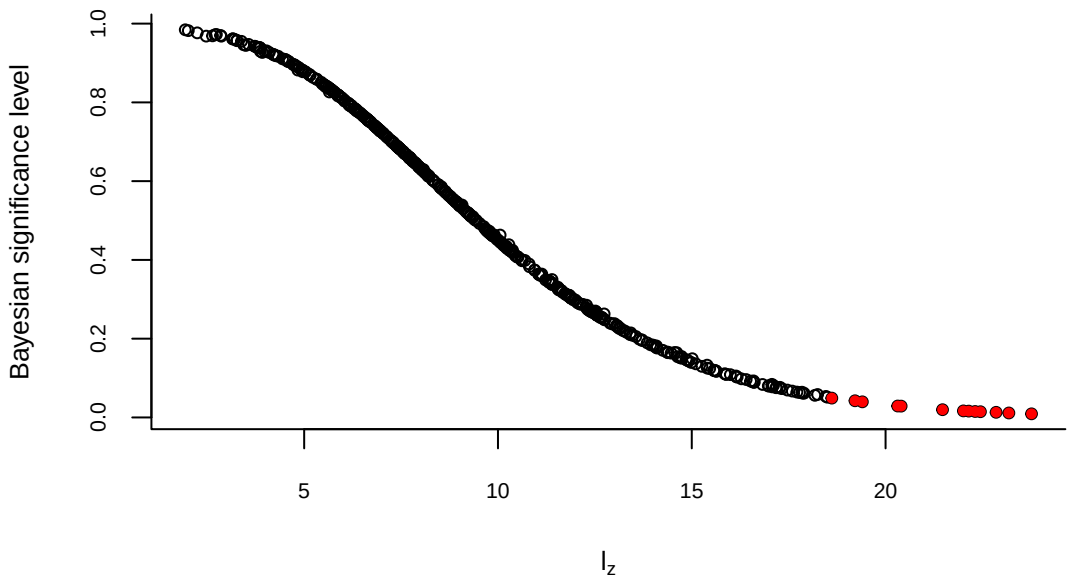

Figure 3: Person-fit statistic for RT patterns against the posterior significance probability.

```
##
##    1   1.139   0.113   1.058   1   0.369   0.071   0.219
##    2   0.762   0.085   0.737   2  -0.588   0.066  -0.651
##    3   0.948   0.093   0.886   3  -0.017   0.064   0.069
##    4   1.160   0.110   1.134   4  -0.481   0.075  -0.555
##    5   1.536   0.163   1.455   5  -0.085   0.074  -0.222
##    6   1.268   0.118   1.371   6  -0.212   0.070  -0.191
##    7   1.257   0.144   1.203   7   1.533   0.135   1.515
##    8   1.077   0.106   1.010   8   0.951   0.087   0.903
##    9   0.523   0.070   0.622   9   0.002   0.058  -0.046
##    10      0.807   0.097   0.847   10     -0.988   0.084  -1.040
##
##    Time Discrimination         Time Intensity          Measurement Error Variance
##    item    EAP     SD      Sim     item    EAP     SD      Sim     EAP     SD      Sim
##
##    1   1.369   0.040   1.357   1   1.549   0.036   1.563   0.660   0.053   0.697
##    2   0.692   0.044   0.747   2   1.914   0.047   1.848   1.055   0.070   1.095
##    3   0.814   0.033   0.809   3  -1.006   0.034  -0.918   0.593   0.041   0.630
##    4   0.960   0.046   0.909   4  -0.123   0.050  -0.155   1.214   0.083   1.210
##    5   0.873   0.047   0.861   5   0.432   0.049   0.429   1.206   0.083   1.235
##    6   0.888   0.034   0.878   6   0.411   0.034   0.349   0.588   0.041   0.565
##    7   1.350   0.052   1.416   7  -0.156   0.052  -0.162   1.364   0.096   1.325
##    8   1.076   0.045   1.074   8  -2.079   0.045  -2.059   1.009   0.071   0.935
##    9   0.957   0.038   0.924   9   0.007   0.038   0.004   0.747   0.052   0.817
##    10      1.266   0.052   1.265   10     -0.943   0.050  -0.899   1.284   0.091   1.143
##
```

```
##   Mean and Covariance matrix Items (mu_a,mu_b,mu_phi,mu_lambda)
##
##   --- Population Mean Item ---
##   mu_a    SD      mu_b    SD      mu_phi  SD      mu_lam  SD
##   1.037   0.142   0.029   0.217   1.020   0.094  -0.002   0.267
##
##   --- Covariance matrix Items (a,b,phi,lambda)---
##       SigmaI          SD SigmaI          SigmaI (Correlation)
##   0.129  0.080  0.009  0.014  0.087  0.123  0.043  0.192  1.000  0.280  0.093  0.032
##   0.080  0.634  0.080 -0.227  0.123  0.348  0.087  0.371  0.280  1.000  0.372 -0.231
##   0.009  0.080  0.073 -0.046  0.043  0.087  0.050  0.142  0.093  0.372  1.000 -0.138
##   0.014 -0.227 -0.046  1.523  0.192  0.371  0.142  0.889  0.032 -0.231 -0.138  1.000
##
##
##   Mean and Covariance matrix Persons (ability,speed)
##
##   --- Population Mean Person (Ability - Speed)---
##           muP     SD
##   Ability     0.000   0.057
##   Speed       0.000   0.029
##
##   SigmaP      SD SigmaP   SigmaP (Correlation)
##   1.018  0.783    0.093  0.065    1.000  0.771
##   0.783  1.012    0.065  0.069    0.771  1.000
```

The summary report provides the estimated and simulated values, since the simulated data object is also given in the call. In the first block, the item parameter estimates of the IRT model are given, and in the second block those of the RT model. Subsequently, the multivariate prior estimates are given.

The joint model in the parameterization of Equation (2) and (6) can be estimated by providing the argument that $par1 = TRUE$, and using the data objects $Y1$ and $RT1$;

```
out1 <- LNIRT(RT1,Y1,data=data,XG=5000,par1=TRUE)
```

When the log-normal model needs to be parameterized with the time discrimination equal to the reciprocal of the standard deviation of the measurement error, the argument $WL = TRUE$ should be provided:

```
data <- simLNIRT(N=500,K=10,rho=0.8,WL=TRUE)

out1 <- LNIRT(RT1,Y1,data=data,XG=2000,par1=TRUE,WL=TRUE)
```

Joint Model Evaluation Tools

When giving the argument residual=TRUE, a residual analysis is computed for the joint model, which includes person- and item-fit tests for both patterns, residual estimates and posterior probability estimates of extremeness of residuals, and the evaluation of distributional assumptions (Fox and Marianti, 2017).

```
set.seed(1234)
data <- simLNIRT(N=500,K=10,rho=0.7)
out <- LNIRT(RT,Y,data=data,XG=5000,residual=TRUE)
```

```
summary(out)
```

```
##
##  Log-Normal RT-IRT Modeling, 2013, J.-P. Fox
##  Summary of results
##
```

```
##   Item Discrimination parameter   Item Difficulty parameter
##   item    EAP     SD      Sim     item    EAP     SD      Sim
##
##   1   1.131   0.116   1.058   1   0.377   0.070   0.219
##   2   0.800   0.085   0.737   2  -0.570   0.067  -0.651
##   3   0.971   0.094   0.886   3  -0.032   0.065   0.069
##   4   1.153   0.108   1.134   4  -0.483   0.074  -0.555
##   5   1.425   0.152   1.455   5  -0.037   0.073  -0.222
##   6   1.310   0.124   1.371   6  -0.220   0.071  -0.191
##   7   1.265   0.136   1.203   7   1.497   0.124   1.515
##   8   1.072   0.112   1.010   8   0.937   0.083   0.903
##   9   0.516   0.070   0.622   9   0.003   0.059  -0.046
##   10      0.800   0.096   0.847   10     -1.001   0.082  -1.040
##
##   Time Discrimination         Time Intensity          Measurement Error Variance
##   item    EAP     SD      Sim     item    EAP     SD      Sim     EAP     SD      Sim
##
##   1   1.369   0.040   1.357   1   1.547   0.036   1.563   0.659   0.054   0.697
##   2   0.684   0.044   0.747   2   1.916   0.046   1.848   1.051   0.071   1.095
##   3   0.818   0.034   0.809   3  -1.006   0.034  -0.918   0.594   0.041   0.630
##   4   0.953   0.047   0.909   4  -0.123   0.049  -0.155   1.219   0.083   1.210
##   5   0.874   0.047   0.861   5   0.431   0.049   0.429   1.207   0.081   1.235
##   6   0.889   0.034   0.878   6   0.411   0.034   0.349   0.590   0.041   0.565
##   7   1.354   0.053   1.416   7  -0.155   0.052  -0.162   1.369   0.098   1.325
##   8   1.084   0.044   1.074   8  -2.079   0.045  -2.059   1.006   0.071   0.935
##   9   0.959   0.038   0.924   9   0.006   0.038   0.004   0.747   0.052   0.817
##   10      1.265   0.053   1.265   10     -0.943   0.051  -0.899   1.289   0.092   1.143
##
##   Mean and Covariance matrix Items (mu_a,mu_b,mu_phi,mu_lambda)
##
##   --- Population Mean Item ---
##   mu_a    SD      mu_b    SD      mu_phi  SD      mu_lam  SD
##   1.036   0.143   0.035   0.213   1.021   0.100  -0.002   0.276
##
##   --- Covariance matrix Items (a,b,phi,lambda)---
##       SigmaI          SD SigmaI           SigmaI (Correlation)
##   0.118  0.082  0.009  0.014  0.080  0.118  0.042  0.176  1.000  0.301  0.096  0.033
##   0.082  0.627  0.082 -0.211  0.118  0.350  0.089  0.367  0.301  1.000  0.381 -0.215
##   0.009  0.082  0.074 -0.055  0.042  0.089  0.052  0.137  0.096  0.381  1.000 -0.164
##   0.014 -0.211 -0.055  1.529  0.176  0.367  0.137  0.841  0.033 -0.215 -0.164  1.000
##
##
##   Mean and Covariance matrix Persons (ability,speed)
##
##   --- Population Mean Person (Ability - Speed)---
##           muP     SD
##   Ability     0.000   0.055
##   Speed       0.000   0.033
##
##   SigmaP      SD SigmaP   SigmaP (Correlation)
##   1.009  0.679    0.095  0.062    1.000  0.672
##   0.679  1.011    0.062  0.070    0.672  1.000
##
##
```

```
##
##   ***                          ***
##   *** Person Fit Analysis (Log-Normal Speed) ***
##   ***                          ***
##
##   Percentage Outliers Persons (5% level)
##
##   lZ
##   2.4 %
##   95% Posterior Probability:  2 %
##
##
##   *** Item Fit Analysis ***
##
##   Misfitting Items (5% level)
##   No Misfitting Items
##
##   *** Residual Analysis ***
##   No Extreme Residuals
##
##   Kolmogorov Smirnov Test (5% level)
##   0 %
##
##
##
##   ***                          ***
##   *** Person Fit Analysis (IRT Model For Ability) ***
##   ***                          ***
##
##   Percentage Outliers Persons (5% level)
##
##   Log-likelihood Statistic
##   2.6 %
##   95% Posterior Probability:  2.6 %
##   95% Posterior Probability (Ability and Speed):  0 %
##
##
##   *** Item Fit Analysis ***
##
##   Misfitting Items (5% level)
##   No Misfitting Items
##
##   *** Residual Analysis ***
##
##   Percentage Extreme Residuals (.95 Posterior Probability)
##   0.02 % (general average across persons and items)
##
##   Extreme Residuals
##   Person  Item     Response   EAP Theta
##     247      10       0    1.3879
##
##   Kolmogorov Smirnov Test (5% level)
##   70 % of items has non-normally distributed latent residuals
##   Item    P-value
```

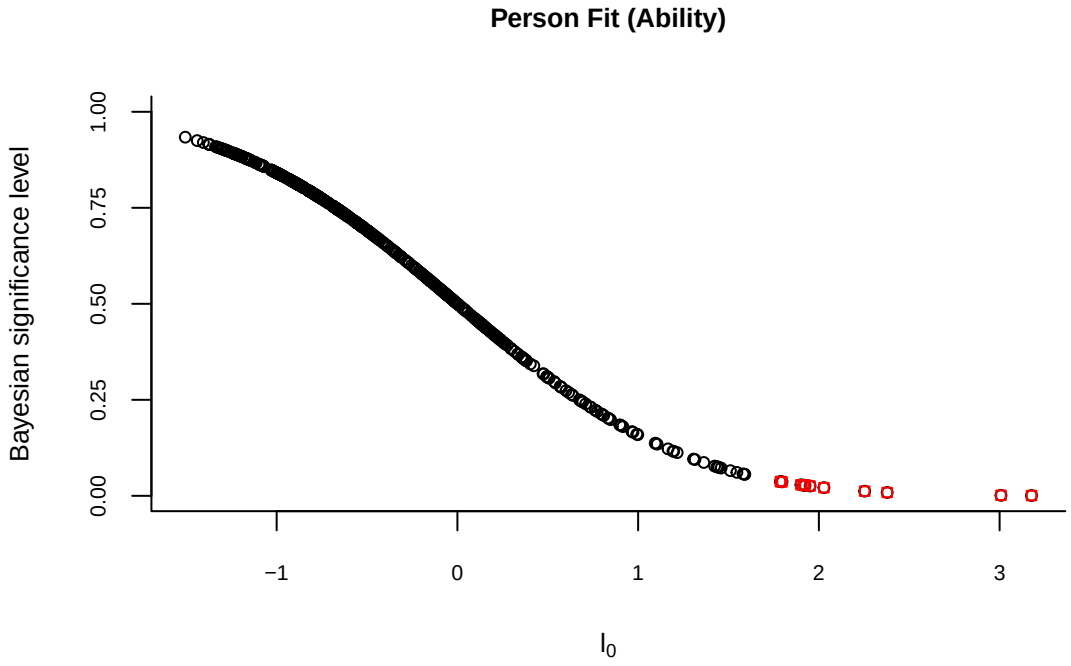

Figure 4: Person-fit statistic for RA patterns plotted against the posterior significance probability.

```
##        1   0.011
##        2   0.001
##        4   0.002
##        6   0.040
##        7   0.000
##        8   0.000
##       10   0.000
```

The same diagnostic information is obtained about the response time model, since it is a component of the joint model. In Figure 4, the estimated person fit statistic values for response accuracy patterns is plotted against the posterior significance probability. The critical area is above the statistic value of 1.645, when considering a significance level of .05. Response accuracy patterns with a statistic value higher than 1.645, are located in the critical region.

```
plot(out$PFl,out$PFlp,xlab=expression(paste(l[0])),ylab="Bayesian significance level",
     pch=1,cex=.75,bty="l",ylim=c(0,1),yaxp=c(0,1,4),cex.axis=.7,cex.lab=.8,
     cex.main=.8,main="Person Fit (Ability)")
set1 <- which(out$PFlp < .05)
set2 <- which(out$EAPCP3 > .95)
points(out$PFl[set1],out$PFlp[set1],col="red",pch=22,cex=.75)
points(out$PFl[set2],out$PFlp[set2],col="blue",pch=17,cex=.95)
```

The speed-accuracy trade-off in the population can be investigated by plotting the estimated abilities against speed for each set of simulated patterns of response time and accuracy, see Figure 5. The RT patterns marked as aberrant are marked in the plot with a triangle. It can be seen that the relationship between speed and ability in the population are not influenced by the aberrant observations.

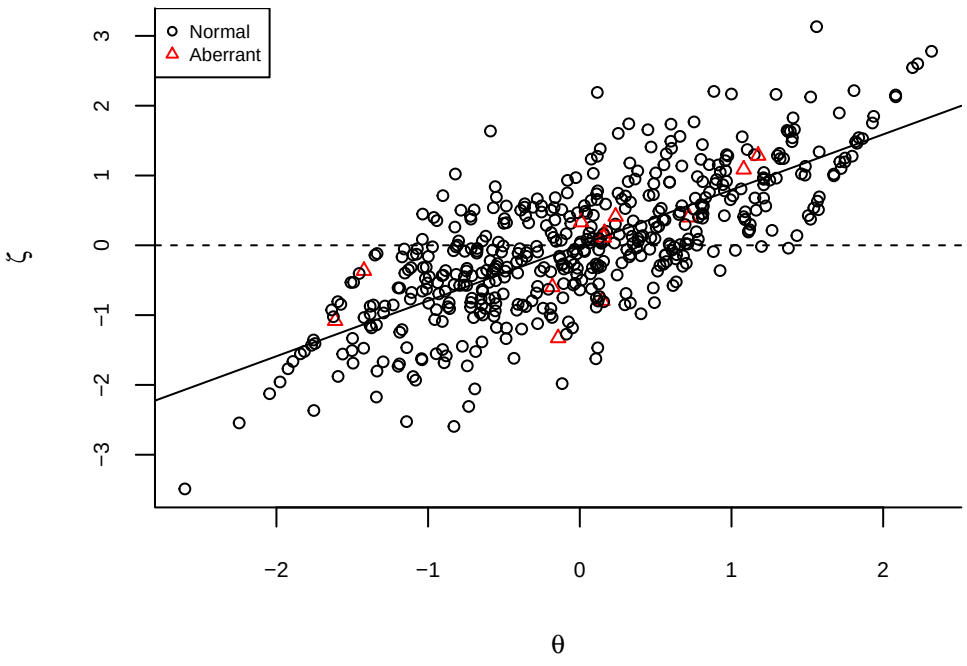

Figure 5: Ability plotted against speed for the non-aberrant and aberrant test takers given the person-fit statistic for RT patterns

```
aberrant <-out$lZP
aberrant[which(aberrant<.05)]<-2 ## group indicator aberrant
aberrant[which(aberrant != 2)]<-1 ## group indicator not aberrant
pch.list <- as.numeric(aberrant)

par(mar=c(4,5,3,4),xpd=F)
plot(out$Mtheta[,1],out$Mtheta[,2],xlab=expression(paste(theta)),
     ylab=expression(paste(zeta)),xlim=c(min(out$Mtheta[,1]),max(out$Mtheta[,1])),
     ylim=c(min(out$Mtheta[,2]),max(out$Mtheta[,2])),bty="l",cex.axis=.7,cex.lab=.8,
     cex.main=.8,pch=pch.list,col=pch.list,cex=.75)
abline(lm(out$Mtheta[,2]~out$Mtheta[,1]))
abline(h = mean(out$Mtheta[,2]),lty = 2)
legend("topleft",c("Normal","Aberrant"),
       horiz=FALSE,pch=c(1,2),col=c(1,"red"),cex=.6)
```

In Figure 6, the person-fit statistic for response patterns is plotted against the person-fit statistic for RT patterns. For both statistics the threshold value of the significant area is marked with a dotted line. The extremeness of each reponse accuracy and response time pattern can be quantified by computing how likely it is that the pattern is flagged under the log-normal RT model and under the IRT model, respectively. The patterns that have a posterior probability of .95 or higher of being extreme are flagged and they are plotted with filled points. Notet that the probability of making a Type-I error is reduced, since the posterior probability quantifies the extremeness of each RT pattern, instead of classifying the pattern based on a chosen significance level (Fox and Marianti, 2017).

It can be seen that the aberrant patterns have a statistic value greater than the critical value. However, a

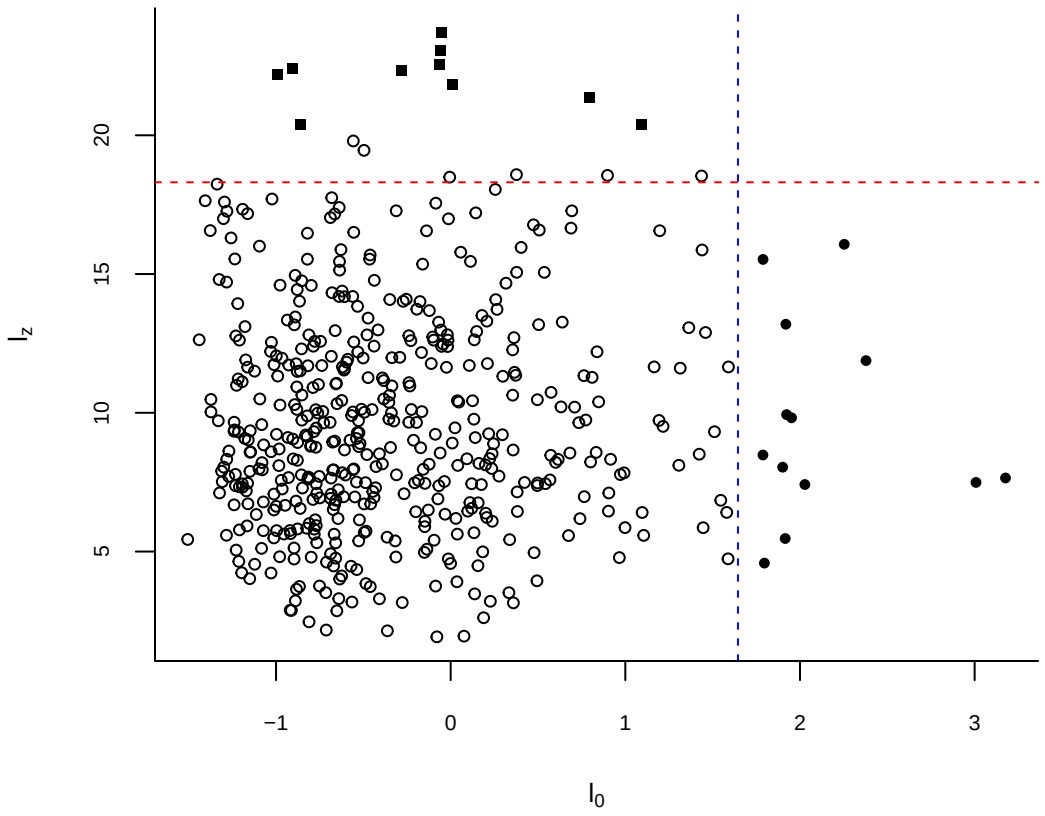

Figure 6: Person-fit statistic for RA patterns plotted against the statistic for RT patterns.

few patterns have an estimated statistic value above the critical value but are not flagged as aberrant, since the posterior probability that the pattern is extreme is less than 95%. More MCMC iterations can increase the accuracy of the estimated significance and classification probabilities.

```
pch.list <- ifelse(out$EAPCP1 > .95,15,1) #flagged aberrant response accuracy
pch.list[which(out$EAPCP2 > .95)] <- 16 #flagged aberrant response times
pch.list[which(out$EAPCP3 > .95)] <- 17 #flagged for both patterns
par(mar=c(5,5,3,2), xpd=F)
plot(out$PFl,out$lZPT,xlab=expression(paste(l[0])),ylab=expression(paste(l[z])),
     xlim=c(min(out$PFl),max(out$PFl)),ylim=c(min(out$lZPT),max(out$lZPT)),bty="l",
     pch=pch.list,cex=.75,cex.main=.8,cex.axis=.7,cex.lab=.8)
abline(h = qchisq(.95, df= nrow(out$data$ab)),lty = 2,col="red")
abline(v = qnorm(.95),lty = 2,col="blue")
```

The object *out*$*EAPCP*1 represents the posterior probability of an extreme RT pattern, *out*$*EAPCP*2 represents the posterior probability of an extreme accuracy pattern, and *out*$*EAPCP*3 represents the posterior probability that the response time and accuracy pattern are extreme.

Explanatory Variables for the Joint Models

Predictor variables can be included for the item (difficulty and time intensity) and person parameters (ability and speed). Data can be simulated using simLNIRT function, where arguments *kia* and *kit* represent the number of predictors for the difficulty and time intensity parameter, respectively. The arguments *kpa* and *kpt* represent the number of predictors for the ability and speed parameter, respectively. In the simulated object, objects XIA and XIT contain the simulated predictor variables for the difficulty and the time intensity parameter, respectively. The objects XPA and XPT the predictor variables for speed and accuracy. As an example, data are simulated with a predictor for the difficulty parameter, two for the time-intensity parameters, one for the ability parameter, and one for the speed parameter. The R-code for this example is given by;

```
set.seed(2224)
data <- simLNIRT(N=1000,K=50,rho=0.7,kpa=1,kpt=1,kia=1,kit=2)
outp <- LNIRT(RT,Y,data=data,XG=2000,XPA=data$XPA,XPT=data$XPT,XIA=data$XIA,XIT=data$XIT)
```

```
summary(outp)
```

```
##
##  Log-Normal RT-IRT Modeling, 2013, J.-P. Fox
##  Summary of results
##
##   Item Discrimination parameter   Item Difficulty parameter
##   item    EAP     SD      Sim     item    EAP     SD      Sim
##
##   1   1.126   0.065   1.133   1   0.397   0.050   0.455
##   2   0.872   0.058   0.817   2   1.071   0.059   1.124
##   3   0.746   0.047   0.866   3  -0.511   0.047  -0.516
##   4   0.992   0.057   0.921   4  -0.492   0.049  -0.406
##   5   1.201   0.068   1.250   5   0.057   0.050   0.014
##   6   0.769   0.050   0.773   6   0.653   0.050   0.643
##   7   0.943   0.060   0.988   7  -1.123   0.063  -1.124
##   8   0.848   0.053   0.788   8  -0.315   0.047  -0.302
##   9   0.841   0.053   0.844   9   0.415   0.047   0.478
##   10      0.691   0.047   0.730   10      0.508   0.045   0.536
##   11      1.173   0.070   1.064   11     -0.372   0.052  -0.304
##   12      0.798   0.051   0.837   12     -0.379   0.049  -0.416
##   13      0.993   0.055   0.988   13     -0.113   0.047  -0.001
##   14      1.468   0.080   1.622   14      0.363   0.057   0.415
##   15      1.127   0.059   1.054   15     -0.580   0.052  -0.550
##   16      0.936   0.059   0.958   16      0.617   0.051   0.623
##   17      1.017   0.060   0.966   17      0.188   0.048   0.257
##   18      1.094   0.068   1.136   18      0.491   0.051   0.498
##   19      1.326   0.074   1.239   19     -0.383   0.053  -0.308
##   20      1.024   0.067   0.995   20     -0.699   0.056  -0.634
##   21      1.100   0.062   1.140   21      0.167   0.048   0.165
##   22      1.032   0.059   0.986   22     -0.145   0.047  -0.099
##   23      1.214   0.072   1.099   23     -0.697   0.059  -0.726
##   24      0.957   0.058   0.918   24     -0.480   0.049  -0.403
##   25      0.894   0.055   0.920   25      0.654   0.052   0.733
##   26      1.187   0.070   1.215   26     -0.497   0.054  -0.486
##   27      1.200   0.067   1.227   27     -0.318   0.049  -0.356
##   28      0.912   0.054   0.933   28      0.180   0.047   0.205
##   29      1.084   0.063   0.913   29     -0.736   0.058  -0.634
##   30      1.333   0.079   1.348   30      0.436   0.056   0.486
```

```
##   31      1.222   0.071   1.317   31     -0.366   0.053  -0.374
##   32      1.082   0.060   1.074   32      0.016   0.050  -0.058
##   33      1.235   0.064   1.176   33      0.053   0.050   0.110
##   34      1.098   0.061   1.094   34      0.059   0.048   0.080
##   35      0.796   0.051   0.841   35     -0.031   0.043  -0.017
##   36      0.926   0.057   0.787   36     -0.362   0.048  -0.354
##   37      0.912   0.057   0.843   37      0.459   0.049   0.480
##   38      1.095   0.068   1.104   38     -1.050   0.064  -0.974
##   39      0.815   0.052   0.805   39      0.440   0.048   0.463
##   40      1.161   0.070   1.120   40      0.209   0.050   0.261
##   41      1.464   0.082   1.405   41     -0.904   0.063  -0.784
##   42      0.803   0.050   0.870   42     -0.285   0.047  -0.305
##   43      0.754   0.049   0.758   43      0.158   0.044   0.114
##   44      0.996   0.063   1.038   44      0.289   0.048   0.303
##   45      1.079   0.069   1.073   45      1.202   0.071   1.151
##   46      1.217   0.070   1.264   46     -0.439   0.052  -0.383
##   47      1.126   0.069   1.194   47      0.483   0.053   0.545
##   48      0.830   0.052   0.914   48      0.504   0.047   0.553
##   49      0.788   0.053   0.835   49      0.601   0.051   0.612
##   50      0.684   0.051   0.700   50     -0.804   0.050  -0.787
##
##   Time Discrimination         Time Intensity          Measurement Error Variance
##   item    EAP     SD      Sim     item    EAP     SD      Sim     EAP     SD      Sim
##
##   1   0.647   0.016   0.673   1  -0.283   0.026  -0.263   0.666   0.030   0.643
##   2   0.874   0.020   0.852   2   0.831   0.032   0.831   1.005   0.045   0.970
##   3   0.887   0.017   0.910   3   0.047   0.026   0.039   0.696   0.032   0.681
##   4   1.117   0.021   1.148   4  -0.238   0.032  -0.233   1.063   0.050   1.107
##   5   1.084   0.021   1.088   5  -0.106   0.033  -0.098   1.092   0.050   1.045
##   6   1.079   0.023   1.088   6  -0.787   0.035  -0.773   1.180   0.055   1.087
##   7   0.843   0.020   0.847   7  -0.049   0.031  -0.033   0.999   0.045   1.054
##   8   0.765   0.028   0.782   8  -0.104   0.045  -0.085   1.922   0.087   1.937
##   9   1.026   0.018   1.041   9  -0.148   0.029  -0.109   0.822   0.039   0.745
##   10      1.128   0.019   1.101   10      0.487   0.028   0.561   0.830   0.037   0.839
##   11      0.807   0.025   0.830   11      0.411   0.039   0.452   1.487   0.067   1.429
##   12      0.942   0.020   0.918   12     -0.052   0.032  -0.081   1.072   0.048   1.048
##   13      1.078   0.024   1.087   13      0.152   0.037   0.161   1.398   0.065   1.272
##   14      0.927   0.018   0.930   14     -0.836   0.029  -0.842   0.843   0.040   0.842
##   15      1.190   0.024   1.155   15      0.047   0.039   0.056   1.492   0.066   1.445
##   16      1.054   0.020   1.050   16      0.330   0.030   0.326   0.910   0.041   0.885
##   17      1.049   0.024   1.015   17      0.356   0.038   0.377   1.502   0.069   1.566
##   18      1.390   0.022   1.381   18      0.130   0.034   0.176   1.150   0.051   1.227
##   19      0.919   0.017   0.921   19      0.216   0.029   0.229   0.784   0.034   0.773
##   20      1.245   0.019   1.248   20     -1.070   0.029  -1.002   0.860   0.039   0.915
##   21      1.215   0.022   1.213   21     -1.133   0.033  -1.160   1.163   0.052   1.240
##   22      1.292   0.016   1.310   22      1.230   0.026   1.231   0.654   0.030   0.669
##   23      0.836   0.023   0.823   23     -1.071   0.036  -1.104   1.338   0.061   1.306
##   24      1.179   0.021   1.172   24      0.594   0.031   0.557   1.026   0.046   1.023
##   25      1.065   0.024   1.068   25      0.801   0.037   0.740   1.436   0.065   1.434
##   26      1.056   0.021   1.032   26     -0.339   0.032  -0.364   1.083   0.049   0.100
##   27      0.711   0.026   0.739   27      0.390   0.044   0.384   1.765   0.079   1.723
##   28      0.666   0.019   0.656   28      0.636   0.031   0.674   0.962   0.043   0.968
##   29      0.821   0.015   0.820   29     -1.420   0.023  -1.427   0.522   0.025   0.497
##   30      0.798   0.019   0.805   30     -0.581   0.030  -0.558   0.925   0.042   0.844
```

```
##   31      1.122   0.024   1.124   31      0.102   0.039   0.118   1.462   0.066   1.484
##   32      1.312   0.020   1.344   32      0.339   0.030   0.313   1.001   0.047   1.022
##   33      0.997   0.019   0.970   33      0.154   0.029   0.152   0.840   0.039   0.832
##   34      1.112   0.023   1.086   34      0.865   0.037   0.905   1.262   0.059   1.225
##   35      1.003   0.018   1.019   35     -0.722   0.028  -0.679   0.799   0.036   0.763
##   36      1.517   0.017   1.515   36     -0.613   0.025  -0.607   0.641   0.030   0.681
##   37      1.047   0.021   1.063   37     -0.004   0.035   0.018   1.143   0.052   1.161
##   38      0.857   0.025   0.860   38     -0.774   0.039  -0.801   1.478   0.066   1.490
##   39      0.790   0.016   0.804   39     -1.240   0.026  -1.265   0.688   0.031   0.655
##   40      1.313   0.019   1.309   40      1.119   0.028   1.086   0.775   0.036   0.770
##   41      0.938   0.024   0.924   41     -0.705   0.038  -0.734   1.364   0.062   1.334
##   42      1.108   0.020   1.085   42      1.134   0.031   1.170   0.930   0.043   0.970
##   43      0.937   0.020   0.930   43     -0.170   0.030  -0.145   0.968   0.045   0.100
##   44      1.044   0.022   1.049   44     -0.653   0.034  -0.639   1.171   0.053   1.057
##   45      1.420   0.025   1.408   45      0.050   0.039   0.021   1.498   0.068   1.516
##   46      0.965   0.023   0.969   46      1.358   0.034   1.362   1.271   0.056   1.261
##   47      1.270   0.019   1.239   47      0.762   0.031   0.722   0.935   0.044   0.994
##   48      0.887   0.023   0.893   48     -0.688   0.037  -0.653   1.331   0.061   1.383
##   49      0.848   0.025   0.842   49      0.565   0.041   0.578   1.672   0.076   1.557
##   50      0.783   0.018   0.778   50      0.416   0.029   0.416   0.887   0.040   0.860
##
##   Item Effects and Covariance matrix Items
##
##   --- Population Mean Item ---
##           EAP     SD
##   mu_a      1.019   0.029
##   mu_phi    1.019   0.029
##       Item Difficulty Predictor Effects
##           EAP     SD      Sim
##   Intercept  -0.024   0.140
##   X 1        -0.256   0.474  -0.555
##       Time Intensity Predictor Effects
##           EAP     SD      Sim
##   Intercept  -0.008   0.139
##   X 2        -1.331   0.541  -1.455
##   X 3         0.110   0.457   0.021
##
##
##   --- Covariance matrix Items (a,b,phi,lambda)---
##   SigmaI              SD SigmaI           SigmaI (Correlation)
##   0.043 -0.018  0.002 -0.009  0.010  0.019  0.007  0.022  1.000 -0.152  0.047 -0.072
##  -0.018  0.328  0.020  0.053  0.019  0.076  0.018  0.054 -0.152  1.000  0.170  0.154
##   0.002  0.020  0.042  0.034  0.007  0.018  0.009  0.021  0.047  0.170  1.000  0.277
##  -0.009  0.053  0.034  0.359  0.022  0.054  0.021  0.089 -0.072  0.154  0.277  1.000
##
##
##   Person Effects and Covariance matrix Persons (ability,speed)
##
##   --- Person Effects (Ability - Speed)---
##       Ability Predictor Effects
##       EAP     SD      Sim
##   X 1    -1.032   0.068  -0.965
##       Speed Predictor Effects
##       EAP     SD      Sim
```

```
##   X 2     2.411   0.066   2.541
##
##   SigmaP      SD SigmaP   SigmaP (Correlation)
##   1.018  0.698    0.052  0.040    1.000  0.691
##   0.698  1.001    0.040  0.045    0.691  1.000
```

It follows that the estimated predictor effects are given in the output together with the standard deviations. The true simulated values are also given. The estimates of the predictor effects for the person parameter are quite accurate with a posterior standard deviation of .07, since 1000 response patterns were simulated. The estimated predictor effects for the item parameters show larger standard deviations (around .50), since only 50 items were considered. The standard regression plots can be made by considering the estimated item and person parameters as outcomes of the regression.

When dealing with categorical predictors, dummy coded variables (also known as indicator or design variables which take on values of zero or one) of the categorical (qualitative) predictors are needed to account for the discrete nature of the observed predictor values. For predictors with more than two levels it is better to use effect coding, where the dummy variable takes on values of one, zero or minus one. With effect coding, the constant or baseline is equal to the grand mean of all of the observations. This will make sure that the scale of the dependent variable is not affected by the scale of the predictor variable. As an example, consider a predictor X for ability, which has a qualitative scale of three levels. Then, two dummy variables X_1 and X_2 are needed, and regurlarly they would be coded as $X_1 = 1$ if $X = 1$, and zero otherwise, and $X_2 = 1$ if $X = 2$, and zero otherwise. Then, the mean of the latent ability scale is equal to the mean ability of the baseline group, which are represented by those with $X = 3$. For reasons of intepretation or numerical reasons, this might not be desirable. To restrict the population mean of the ability scale to zero, effect coding can be used. In that case, $X_1 = 1$ if $X = 1$, $X_1 = -1$ if $X = 3$, and zero otherwise. In the same way, $X_2 = 1$ if $X = 2$, $X_2 = -1$ if $X = 3$, and zero otherwise. The effect of X_1 and X_2 are interpreted as the group effects relative to the general mean.

The following code can be used to generate a qualitative predictor with three levels, which is represented by two dummy variables using effect coding. Subsequently, the dummy coded variables are used as predictors to simulate ability parameters. The LNIRT model is fitted to the simulated data.

```
#simulate categorical predictor for ability
kpa <- 1
XPA <- matrix(factor(sample(1:3,N,replace=TRUE)),ncol=kpa,nrow=N)
dummy1<- dummy2 <- rep(0,N)
dummy1[XPA==1]<-  1
dummy1[XPA==3]<- -1
dummy2[XPA==2]<-  1
dummy2[XPA==3]<- -1
XPA <- cbind(dummy1,dummy2)
Ba <- matrix(rnorm(2),ncol=1,nrow=2)
## manipulated simLNIRT code to generate dummy predictors ##

set.seed(1234)
data <- simLNIRT(N=1000,K=10,rho=0.7,kpa=1)
out1 <- LNIRT(RT,Y,data=data,XG=2000,XPA=data$XPA)
summary(out1)
```

The output concerning the estimated effects of the dummy coded predictors for ability are given in the figure below. The estimated values correspond to the simulated values stored in object $data\$Ba$. The effect of the dummy coded variables are reported, but the effect of the group coded as minus one on both variables can also be computed, including the posterior standard deviation:

```
mean(-out1$MmuP[500:XG,1]-out1$MmuP[500:XG,2])

sqrt(var(-out1$MmuP[500:XG,1]-out1$MmuP[500:XG,2]))
```

When including predictor variables for ability, then the intercept for speed is also reported as a predictor effect. In that case, the LNIRT program generates estimates for effects of the predictors for the multivariate outcomes ability and speed. In this multivariate regression all variables (including the intercept) are reported as predictor variables.

```
Person Effects and Covariance matrix Persons (ability,speed)

--- Person Effects (Ability - Speed)---
        Ability Predictor Effects
        EAP     SD      Sim
X 1    -0.628   0.054  -0.625
X 2    -0.573   0.053  -0.552
        Speed Predictor Effects
        EAP     SD      Sim
X 3     0.000   0.032

SigmaP          SD SigmaP       SigmaP (Correlation)
1.039  0.730    0.081  0.048    1.000  0.713
0.730  1.010    0.048  0.051    0.713  1.000
```

Real Data Study

In Fox and Marianti (2017), a real data set (referred to as the credentialing data) concerning 1,636 test takers who applied for licensure were analysed (Cizek and Wollack, 2016). The candidates made Form 1 of the test, which consisted of 170 items, and their RA and RT data were stored. The collected data followed from a year of testing using a computer-based program that tests continuously. Besides the RA information, background information of each candidate was available, for instance, the country where the candidate received his/her educational training, the state in which the test taker applied for licensure, and the center where the candidate took the exam. The test takers were pretested using three different item sets. The average scores varied significantly across the differently pretested groups.

In this study, RT and RA patterns of 1636 test takers were analyzed using the joint model. Fox and Marianti (2017) only considered responses of those who were pretested with the same item set. The person-fit tests were used to detect aberrant response behavior, without using any background information. The LNIRT program was used to estimate all model parameters and to compute the person-fit statistics.

The joint model was identified by restricting the population means of ability and speed to zero and by restricting the product of time discriminations and discriminations to one. The MCMC convergence diagnostics were used to evaluate the convergence of the chains. According to the diagnostics, a burn-in period of 1,000

iterations and a total of 5,000 MCMC iterations were made to estimate the model parameters. The object out from a workspace contains the simulated parameter values.

The following call was made,

out <- LNIRT(RT=RT1,Y=Yf1,XG=10000,residual=TRUE)

summary(out)

and object out was stored in a workspace credential.Rdata

The estimated mean of the item difficulties is $-.70$ and the item difficulties vary with a variance of .27. The estimated mean of the time intensities is around 4.00 and the time intensities vary with a variance of .11. It can be seen that the range in item difficulties is relatively large, which gives support to accurate estimation of test-takers' ability. Item 159 and item 169discriminate poorly with a value of .15). There is a negative correlation between the item discrimination and item difficulty. The variety in time discriminations is not very high with a variance ofaround .05. The average population level of ability and speed was fixed to zero to identify the scale.

The covariance and correlation estimates are given of the population parameters of the joint model. For all test takers, it can be seen that the estimated correlation between ability and speed, when speed is constant, is around .395. The positive correlation indicates that the high-ability test takers worked faster than the low-ability test takers. The variation in speed values across test takers is around .027, which is rather small, since the variation in time intensities is .11 and almost 5 times larger. Most of the variation between RTs is explained by the differences in time intensities.

There exists a high correlation between item discrimination and time discrimination, and item difficulty and time intensity, around .485 and .464, respectively. This means that the discriminating items with respect to ability also discriminate well with respect to speed. The positive relation between the item difficulty and time intensity means that the time-intensive items are the more difficult items.

The person-fit statistics to detect aberrant response behavior given response accuracy and responser times, respectively, were computed. In Figure 7, the estimated person-fit statistic values are plotted against the posterior probability of significance. The statistic values are chi-square distributed with 170 degrees of freedom, under the joint model. Subsequently, the critical statistic value is 201.4, when the level of significance equals .05. Estimated statistic values higher than 201.4 are located in the critical region. Given this significance level, a total of 19.5% of the response time patterns are identified as aberrant. For the response accuracy patterns, the critical area is above the statistic value of 1.645, when considering a significance level of .05. Test takers with a statistic value higher than 1.645, are located in the critical region. In this study, around 1.4% are identified in the critical region and hence, are detected as persons with aberrant RA patterns.

The speed-accuracy trade-off in the population was investigated by plotting the estimated ability against the speed values. In Figure 8, the relationship between speed and ability for the identified non-aberrant and aberrant test takers is plotted. An aberrant group of test takers was identified according to the person-fit test using response times (significance level of .05). It can be seen that both groups show a comparable positive correlation between speed and ability. The aberrant RT patterns do not strongly influence the estimated relationship between speed and ability.

In Figure 8, the person-fit statistic for response accuracy (x-axis) is plotted against the person-fit statistic for response times (y-axis). For both statistics, the threshold value of the significant area is marked with a dotted line. It can be seen that with respect to aberrant RT patterns, a serious number of test takers are marked as aberrant, since their value is above the threshold of 201.4. A few test takers are marked as aberrant with respect to their RA pattern, since their statistic value is above 1.645. Those marked as aberrant with respect to their RA and RT pattern are represented by a triangle. Only 5 test takers are marked as aberrant for both patterns. The plotted statistic scores concerning RT and RA patterns do not seem to be related. In theory, this relationship is possible, since in the computation of the person-fit statistics structural relationships between parameters are taken into account. Therefore, it would be possible that differences between aberrant and non-aberrant patterns are explained by a relationship between speed and ability or by a relationship between item characteristics.

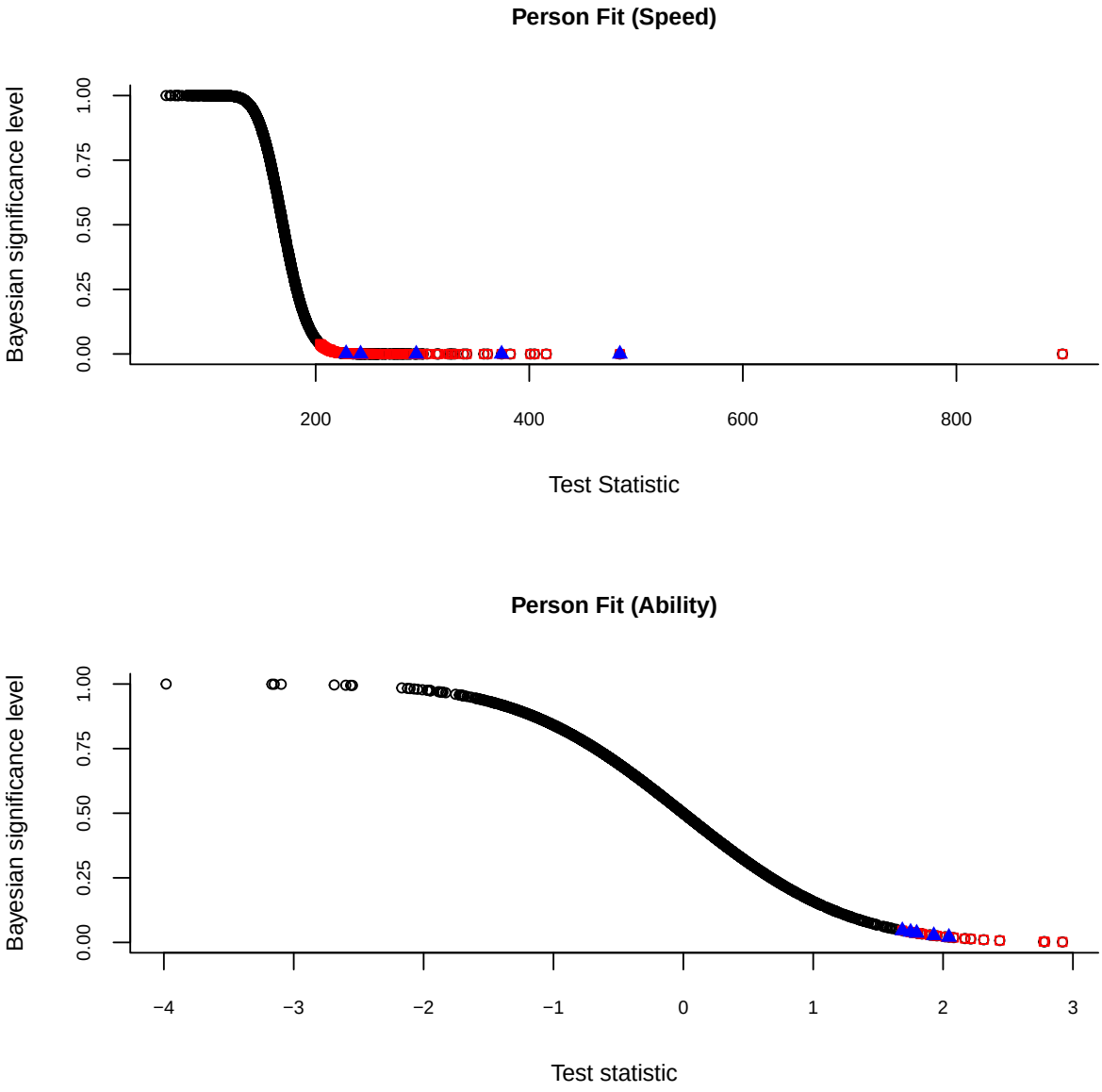

Figure 7: Person-fit statistic plotted against the corresponding posterior significance probability

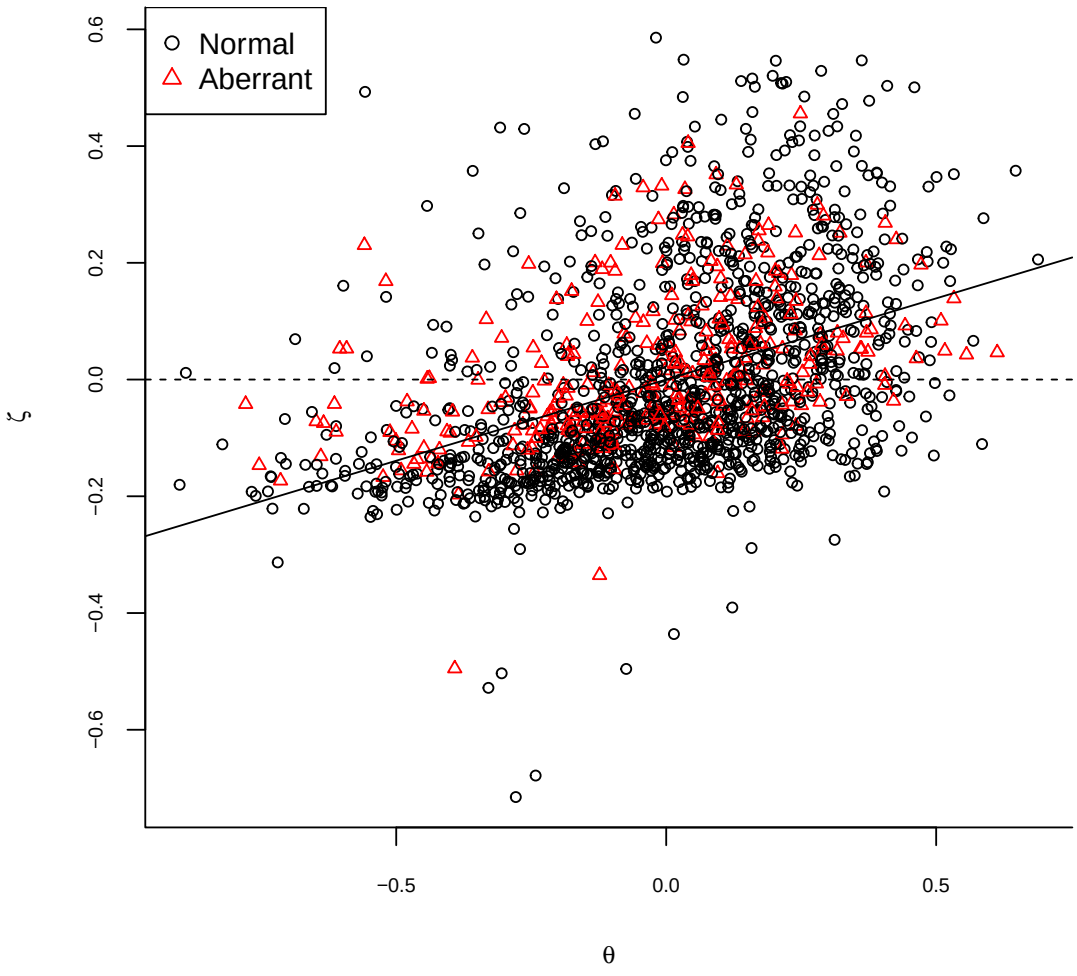

Figure 8: Ability plotted against speed for the identified non-aberrant and aberrant test takers.

Discussion

The joint modeling of responses and RTs can be used to make inferences about ability and speed given educational test data. The relationship between speed and ability provides information about the test taker and items. It increasingly receives attention due to the increase in computer-based testing. To make correct inferences from the joint model, statistical tests have been developed to evaluate the fit. The R-package *LNIRT* can be used to fit the joint model for RTs and RA.

Bayesian significance testing is used to evaluate person fit for observed RT and RA patterns. The developed person-fit statistics can be used to identify aberrant test takers with respect to their RT pattern or their response pattern or both patterns.

In practice, aberrant response behavior can seriously diminish the validity of the test results and affect test results of other test takers. Test companies and programs require advanced technology such as video surveillance, but also seating charts and follow-up interviews, to prevent and detect inappropriate behavior of test takers. They should support test integrity and actively prevent and detect fraudulent or deceptive response behavior, since the test results can have important consequences for test takers. The tools for the joint model can be used by test companies to analyze their test data and to identify statistical irregularities. The person-fit tests can be used to detect inappropriate behavior by identifying patterns which show irregularities and/or extreme responses and/or RTs.

References

Cizek GJ, Wollack JA (2016). *Handbook of quantitative methods for detecting cheating on tests*. Taylor & Francis.

Fox JP (2010). *Bayesian Item Response Modeling: Theory and Applications*. Springer Science & Business Media. ISBN 978-1-4419-0742-4.

Fox JP, Klein Entink R, van der Linden WJ (2007). "Modeling of Responses and Response Times with the Package CIRT." *Journal of Statistical Software*, **20**(7), 1–14. ISSN 1548-7660.

Fox JP, Marianti S (2017). "Person-Fit Statistics for Joint Models for Accuracy and Speed." *Journal of Educational Measurement*, **54**(2), 243–262. ISSN 1745-3984. doi:10.1111/jedm.12143. URL http://dx.doi.org/10.1111/jedm.12143.

Klein Entink RH, Fox JP, van der Linden WJ (2008). "A Multivariate Multilevel Approach to the Modeling of Accuracy and Speed of Test Takers." *Psychometrika*, **74**(1), 21. ISSN 0033-3123, 1860-0980.

Lord FM, Novick MR (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

Marianti S, Fox JP, Avetisyan M, Veldkamp BP, Tijmstra J (2014). "Testing for Aberrant Behavior in Response Time Modeling." *Journal of Educational and Behavioral Statistics*, **39**(6), 426–451. doi:10.3102/1076998614559412. https://doi.org/10.3102/1076998614559412, URL https://doi.org/10.3102/1076998614559412.

Maris E (1993). "Adaptive and Multiplicative Models for Gamma Distributed Variables, and Their Application as Psychometric Models for Response Times." *Psychometrika*, **58**, 445–469.

Roskam EE (1997). "Models for Speed and Time-Limit Tests." In WJ van der Linden, RK Hambleton (eds.), "Handbook of Modern Item Response Theory," pp. 187–208. Springer, New York.

Samejima F (1973). "Homogeneous Case of the Continuous Response Level." *Psychometrika*, **38**, 203–219.

Scheiblechner H (1979). "Specific Objective Stochastic Latency Mechanisms." *Journal of Mathematical Psychology*, **19**, 18–38.

Schnipke DL, Scrams DJ (1997). "Representing Response Time Information in Item Banks." *LSAC Computarized Testing Report 97-09*, Law School Admission Council, Newton, PA.

Shi JQ, Lee SY (1998). "Bayesian Sampling-based Approach for Factor Analysis Models with Continuous and Polytomous Data." *British Journal of Mathematical and Statistical Psychology*, **51**, 233–252.

Thissen D (1983). "Timed Testing: An Approach Using Item Response Theory." In DJ Weiss (ed.), "Latent Trait Test Theory and Computarized Adaptive Testing," pp. 179–203. Academic Press, New York.

van der Linden WJ (2006). "A Lognormal Model for Response Times on Test Items." *Journal of Educational and Behavioural Statistics*, **31**, 181–204.

van der Linden WJ (2007). "A Hierarchical Framework for Modeling Speed and Accuracy on Test Items." *Psychometrika*, **72**(3), 287. ISSN 0033-3123, 1860-0980.

van der Linden WJ, Entink RHK, Fox JP (2010). "IRT Parameter Estimation With Response Times as Collateral Information." *Applied Psychological Measurement*, **34**(5), 327–347. doi:10.1177/0146621609349800. https://doi.org/10.1177/0146621609349800, URL https://doi.org/10.1177/0146621609349800.

van der Linden WJ, Scrams DJ, Schnipke DL (1999). "Using Response-time Constraints to Control for Speededness in Computarized Adaptive Testing." *Applied Psychological Measurement*, **23**, 195–210.

Verhelst ND, Verstraalen HHFM, Jansen MG (1997). "A Logistic Model for Time Limit Tests." In WJ van der Linden, RK Hambleton (eds.), "Handbook of Modern Item Response Theory," pp. 169–185. Springer, New York.