*Bernoulli* **25**(2), 2019, 1326–1354
https://doi.org/10.3150/18-BEJ1022

# Strong Gaussian approximation of the mixture Rasch model

FRIEDRICH LIESE*, ALEXANDER MEISTER** and JOHANNA KAPPUS†

*Institut für Mathematik, Universität Rostock, D-18051 Rostock, Germany.*
*E-mail: *friedrich.liese@uni-rostock.de; **alexander.meister@uni-rostock.de;*
*†johanna_kappus@t-online.de*

We consider the famous Rasch model, which is applied to psychometric surveys when $n$ persons under test answer $m$ questions. The score is given by a realization of a random binary $n \times m$-matrix. Its $(j,k)$th component indicates whether or not the answer of the $j$th person to the $k$th question is correct. In the mixture, Rasch model one assumes that the persons are chosen randomly from a population. We prove that the mixture Rasch model is asymptotically equivalent to a Gaussian observation scheme in Le Cam's sense as $n$ tends to infinity and $m$ is allowed to increase slowly in $n$. For that purpose, we show a general result on strong Gaussian approximation of the sum of independent high-dimensional binary random vectors. As a first application, we construct an asymptotic confidence region for the difficulty parameters of the questions.

*Keywords:* asymptotic equivalence of statistical experiments; high-dimensional central limit theorem; item response model; Le Cam distance; psychometrics

## 1. Introduction

The Rasch model is a famous and widely used approach to analyse surveys in the field of psychometrics. It assumes that each of $n$ subjects (typically persons) are exposed to $m$ items (typically questions to be answered). For each $j = 1, \dots, n$ and $k = 1, \dots, m$ the correctness of the answer of person $j$ to the question $k$ is a binary random variable $X_{j,k}$ where the probability of a correct answer, that is, $X_{j,k} = 1$, is given by

$$P(X_{j,k} = 1) = \frac{\exp\{\beta_j - \theta_k\}}{1 + \exp\{\beta_j - \theta_k\}}, \qquad k = 1, \dots, m;\ j = 1, \dots, n.$$

The parameter $\theta_k$ characterizes the difficulty of the $k$th item and the parameter $\beta_j$ reflects the ability of the $j$th individual. The $\beta_j$ may be either considered as unknown parameters (standard Rasch model) or as realizations of i.i.d. random variables with distribution $F$. The latter case describes the situation in which the individuals are randomly selected from a large population. Then the observation vectors $Y_j = (X_{j,1}, \dots, X_{j,m})$ are i.i.d. and it holds for every binary matrix $\varepsilon = (\varepsilon_{j,k})_{j=1,\dots,n;k=1,\dots,m}$ that

$$P\big((Y_1, \dots, Y_n)^T = \varepsilon\big) = \prod_{j=1}^{n} \int \left\{ \prod_{k=1}^{m} \frac{\exp\{\varepsilon_{j,k}(\beta - \theta_k)\}}{1 + \exp\{\beta - \theta_k\}} \right\} dF(\beta), \qquad \varepsilon_{j,k} \in \{0, 1\}. \tag{1.1}$$

1350-7265 © 2019 ISI/BS

This type of psychometric model is called the mixture Rasch model which will be the central object in this paper.

For original literature, we refer to the book of Rasch [29], after whom the model has been named. Also we mention the books of Alagumalai et al. [1] and Bezruczko [5] for applications of the Rasch model. It has also confined attention in the econometric literature (Hoderlein et al. [18]). The mixture model is used in Lindsay et al. [22], Rice [31] and Strasser [36,37]. Also we refer to the books of Fischer and Molenaar [15] and von Davier and Carstensen [38].

So far most of the literature on the Rasch model has mainly focused on the estimation of the difficulty parameters, consistency and asymptotic normality (mostly for bounded $m$) where maximum likelihood (ML) or quasi-ML methods are preferred, see, for example, Haberman [17], de Leeuw and Verhelst [13] or Pfanzagl [27,28]. Lindsay et al. [22] consider semiparametric estimation in the Rasch model and related problems. Biehler et al. [6] study saddlepoint approximation of the ability parameters. Doebler et al. [14] construct confidence intervals for the ability parameters. Strasser [36,37] thoroughly investigates the covariance structure and asymptotic distribution of quasi-ML estimators in the mixture Rasch model.

In this work, we approximate the mixture Rasch model in the strong Le Cam sense by a model which contains a Gaussian observation, and – conditionally on that – another Gaussian observation whose distribution does not depend on the ability distribution $F$ (as $n \to \infty$). This investigation is motivated by the fact that, for Gaussian models, the structure of optimal estimators and tests is understood very well in both the parametric and nonparametric case.

As a first application, we will construct a uniform asymptotic confidence ellipsoid for the difficulty parameters in the asymptotically equivalent Gaussian model under potentially increasing (but restricted) dimension $m$, which, thus, also represents a uniform asymptotic confidence ellipsoid in the original mixture Rasch model. Also the asymptotic equivalence result will open a broad field of further applications as we will explain in the conclusions.

The distribution $F$ in (1.1) is not nonparametrically identified for bounded $m$, a situation that is similar to the binomial mixture models. Therefore, we allow $m = m_n$ to tend to infinity, as $n \to \infty$. Therein $m_n$ has to be of smaller order compared to $n$. This means that there are much more subjects under test compared to the total number of questions contained in the sheet, a condition that is satisfied in almost all applications and especially in the Programme for International Student Assessment (PISA), to which the Rasch model has been applied.

The model (1.1) is included in the more general framework of statistical experiments in which one observes the i.i.d. random vectors $Y_1, \dots, Y_n$ with

$$\begin{aligned} \mathcal{L}(Y_1, \dots, Y_n) &= \int \bigotimes_{j=1}^{n} \bigotimes_{i=1}^{m} P_{\beta_j - \theta_i} \, dF^{\otimes n}(\beta_1, \dots, \beta_n) \\ &= \left( \int \bigotimes_{i=1}^{m} P_{\beta - \theta_i} \, dF(\beta) \right)^{\otimes n}, \end{aligned} \tag{1.2}$$

where $\bigotimes$ is the symbol for the product measure and $P_\gamma$ is an exponential family; concretely,

$$P_\gamma(X = k) = h(k) \exp\bigl\{\gamma k - K(\gamma)\bigr\}, \qquad k \in \mathbb{N} \cup \{0\}.$$

In our setting (1.1), we have $K(\gamma) = \log(1 + \exp\{\gamma\})$ and $h = 1_{\{0,1\}}$. As an alternative example for the general setting (1.2), we provide the Poisson Rasch model, for which it holds that

$$P_\lambda(X = k) = h(k) \exp\big\{k(\beta - \theta) - \exp\{\lambda\}\big\} = \frac{(\rho\tau)^k}{k!} \exp\{-\rho\tau\},$$

where $\rho = \exp\{\beta\}$, $\tau = \exp\{-\theta\}$, $\lambda = \exp\{\beta - \theta\}$ and $h(k) = 1/k!$. Thus, $P_\lambda$ is a Poisson distribution in that model rather than a Bernoulli distribution as in our setting (1.1).

## 2. Asymptotic equivalence

In this section, we provide a brief introduction to the concept of equivalence and asymptotic equivalence of statistical experiments. Assume we have two statistical experiments $\mathcal{E}_j = (\Omega_j, \mathfrak{A}_j, (P_{\theta,j})_{\theta\in\Theta})$ with the same parameter space $\Theta$. By $\mathfrak{K}_{i,j}$ and $i, j \in \{1, 2\}$, we shall denote the set of all Markov kernels $K_{i,j} : \mathfrak{A}_j \times \Omega_i \to [0, 1]$. The application of $K_{2,1}$ on $P_{\theta,2}$, that is,

$$(K_{2,1} P_{\theta,2})(A_1) = \int K_{2,1}(A_1, \omega_2) P_{\theta,2}(d\omega_2)$$

is a probability measure on $(\Omega_1, \mathfrak{A}_1)$. The two statistical experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ are called equivalent if there are Markov kernels $K_{1,2}$ and $K_{2,1}$, both not depending on $\theta$, such that $K_{2,1} P_{\theta,2} = P_{\theta,1}$ and $K_{1,2} P_{\theta,1} = P_{\theta,2}$ for all $\theta \in \Theta$. Then the two experiments are also equivalent in the decision theoretic sense. Indeed, if $(\mathcal{D}, \mathfrak{D})$ is a decision space, $L(a, \theta)$ a loss function and $D_i(B, \omega_i)$, $B \in \mathfrak{D}$, $\omega_i \in \Omega_i$ is a (randomized) decision for the $i$th experiment then

$$D_j(B, \omega_j) := \int D_i(B, \omega_i) K_{j,i}(d\omega_i, \omega_j)$$

is a decision for the other experiment and it can be easily seen that both decisions have identical risk functions. Now suppose that $T : \Omega_1 \to \Omega_2$ is sufficient, in the sense that there exists some Markov kernel $K$ which does not depend on $\theta$ but represents a version of the conditional measure given $T$ under $P_{\theta,1}$ for all $\theta \in \Theta$; concretely $K(A_1, T) = E_{\theta,1}(1_{A_1} \mid T)$ $P_{\theta,1}$-a.s., for all $A_1 \in \mathfrak{A}_1$; and that $P_{\theta,2} = P_{\theta,1} \circ T^{-1}$ for all $\theta \in \Theta$. Let $\delta_a$ denote the Dirac measure concentrated at point $a$. Then $K_{1,2}(A_2, \omega_1) = \delta_{T(\omega_1)}(A_2)$ is a Markov kernel and it holds that $K_{1,2} P_{\theta,1} = P_{\theta,1} \circ T^{-1}$. The sufficiency of $T$ implies that there is a Markov kernel $K_{2,1}$ with $K_{2,1} P_{\theta,2} = P_{\theta,1}$. The two experiments are equivalent, therefore.

The concept of deficiency makes precise in what sense the approximate sufficiency of a statistic or, more generally, the approximate equivalence is to be understood. It is defined with the help of the total variation distance $\mathsf{TV}(P, Q) = 2 \sup_A |P(A) - Q(A)|$ between the distributions $P$ and $Q$. Put, for $i, j \in \{1, 2\}$, $i \neq j$,

$$\delta(\mathcal{E}_i, \mathcal{E}_j) = \inf_{K_{j,i} \in \mathfrak{K}_{i,j}} \sup_{\theta\in\Theta} \mathsf{TV}(K_{j,i} P_{\theta,j}, P_{\theta,i}),$$

$$\Delta(\mathcal{E}_1, \mathcal{E}_2) = \max\big(\delta(\mathcal{E}_1, \mathcal{E}_2), \delta(\mathcal{E}_2, \mathcal{E}_1)\big).$$

Therein $\delta(\mathcal{E}_i, \mathcal{E}_j)$ is called the deficiency of $\mathcal{E}_i$ and $\mathcal{E}_j$ and $\Delta(\mathcal{E}_1, \mathcal{E}_2)$ is the Le Cam distance of $\mathcal{E}_1$ and $\mathcal{E}_2$. It is a metric in the space of equivalence classes of statistical experiments with a joint parameter set. Two sequences $\mathcal{E}_{j,n} = (\Omega_{j,n}, \mathfrak{A}_{j,n}, (P_{\theta,j,n})_{\theta\in\Theta_n})$, $j = 1, 2$ of statistical experiments are called asymptotically equivalent if $\lim_{n\to\infty} \Delta(\mathcal{E}_{1,n}, \mathcal{E}_{2,n}) = 0$. By a slight abuse of language one calls the experiments $\mathcal{E}_{1,n}$ and $\mathcal{E}_{2,n}$ asymptotically equivalent while this means asymptotic equivalence of the corresponding sequences. Sometimes the sample spaces are identical, then

$$\Delta(\mathcal{E}_{1,n}, \mathcal{E}_{2,n}) \le \sup_{\theta\in\Theta_n} \mathsf{TV}(P_{\theta,1,n}, P_{\theta,2,n}).$$

Asymptotic equivalence allows to take over asymptotic properties such as convergence rates of estimators or asymptotic confidence regions from one experiment to the other.

In the local asymptotic decision theory $\Theta$ is an open subset of $\mathbb{R}^d$ and for a fixed $\theta_0 \in \Theta$ and a sequence $a_n$ tending to zero one introduce a local parameter $h \in \mathcal{H}_n = \{h : \theta_0 + a_n h \in \Theta\} \subseteq \mathbb{R}^d$. The so called LAN condition for $\mathcal{E}_{1,n}$, see Strasser [35] is equivalent to the following statement: There is a matrix $\mathcal{I}(\theta_0)$, called information matrix, such that $(\mathcal{E}_{1,n})_n$ converges weakly to the Gaussian experiment $\mathcal{E}_2 = (\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d), (\mathsf{N}(\mathcal{I}(\theta_0)h, \mathcal{I}(\theta_0))_{h\in\mathbb{R}^d})$. Weak convergence means that $\Delta(\mathcal{E}_{1,n}^{\mathcal{H}}, \mathcal{E}_2^{\mathcal{H}}) \to 0$, where the superscript $\mathcal{H}$ means that for consider the experiments only for a finite but arbitrary subset $\mathcal{H} \subseteq \mathcal{H}_n$ as parameter set. A typical situation, in which this condition holds, occurs if the family $P_\theta$ is $L_2$-differentiable and $a_n = \frac{1}{\sqrt{n}}$ and $P_{\theta,1,n} = P_{\theta,1}^{\otimes n}$, that is, if we have i.i.d. observations.

For books on Le Cam theory we refer to Le Cam [19], Strasser [35], Le Cam and Yang [20], Shiryaev and Spokoiny [34] and Liese and Miescke [21]. In nonparametric literature, research mainly focuses on showing asymptotic equivalence of curve estimation problems to white noise models, in which the target curve occurs as the drift function of a Wiener process. Therein we mention for example, Nussbaum [26], Carter [10] and Brown et al. [7] for density estimation; Brown and Low [8], Rohde [32], Carter [11], Cai and Zhou [9] and Schmidt-Hieber [33] for nonparametric regression; Meister [24] for functional linear regression; Reiß [30], Genon-Catalot and Larédo [16] and Mariucci [23] for the analysis of more complex stochastic processes. The paper of Meister and Reiß [25] somehow deviates from this list as it establishes asymptotic equivalence of nonregular nonparametric regression and a specific Poisson point process. Still Gaussian limit models are most popular.

## 3. Dimension reduction by sufficiency

The sample space for the Rasch model is $\Omega = \{0, 1\}^{n\times m}$, the space of all binary $n \times m$-matrices $\omega = (\omega_{j,k})$, $1 \le j \le n$; $1 \le k \le m$. Throughout, we equip a discrete sample space $\mathcal{X}$ by the power set $\mathcal{P}(\mathcal{X})$ as $\sigma$-algebra. Let $X_{j,k}(\omega) = \omega_{j,k}$ indicate the correctness of the answer of person $j$ to question $k$. Then $Y_j = (X_{j,1}, \ldots, X_{j,m})$ is the response vector of person $j$.

We fix some $R > 0$ and a set $\mathcal{F}$ of admitted distributions $F$ in (1.1) and set

$$\Theta = \left\{\theta \in [-R, R]^m : \sum_{k=1}^m \theta_k = 0\right\}. \tag{3.1}$$

Note that the condition that the $\theta_k$ add to zero is a common calibration to ensure identifiability of the difficulty parameters. For $\theta = (\theta_1, \dots, \theta_m)$ and $F \in \mathcal{F}$, we denote by $P^A_{\theta,F}$ the joint distribution of $Y_1, \dots, Y_n$. The density $dP^A_{\theta,F}/d\kappa_{n\times m}$ of $P^A_{\theta,F}$ with respect to the counting measure $\kappa_{n\times m}$ on $\Omega$ is the probability mass function and (1.1) yields

$$\frac{dP^A_{\theta,F}}{d\kappa_{n\times m}}(\omega) = P^A_{\theta,F}\big(\{\omega\}\big) = \prod_{j=1}^n \int \left\{ \prod_{k=1}^m \frac{\exp\{\omega_{j,k}(\beta-\theta_k)\}}{1+\exp\{\beta-\theta_k\}} \right\} dF(\beta). \tag{3.2}$$

Putting together all components we arrive at the experiment, which is the mixture Rasch model,

$$\mathcal{A}_{n,m} := \big(\{0,1\}^{n\times m}, \mathcal{P}\big(\{0,1\}^{n\times m}\big), \big(P^A_{\theta,F}\big)_{(\theta,F)\in\Theta\times\mathcal{F}}\big).$$

We set

$$S_j = \sum_{k=1}^m X_{j,k}, \qquad N_k = \sum_{j=1}^n 1_{\{k\}}(S_j), \qquad T_k = \sum_{j=1}^n X_{j,k}, \tag{3.3}$$

where $1_A$ is the indicator function of the set $A$. Put

$$G(k,\theta,F) = \log\left\{ \int \left\{ \exp\{k\beta\} \prod_{l=1}^m \frac{1}{1+\exp\{\beta-\theta_l\}} \right\} dF(\beta) \right\}$$

for $k = 0, \dots, m$. The representation (3.2) yields

$$\frac{dP^A_{\theta,F}}{d\kappa_{n\times m}} = \exp\left\{ -\sum_{k=1}^m \theta_k T_k + \sum_{j=1}^n G(S_j,\theta,F) \right\} = \exp\left\{ -\sum_{k=1}^m \theta_k T_k + \sum_{k=0}^m N_k \cdot G(k,\theta,F) \right\}.$$

Then, by the Fisher–Neyman factorization criterion, we realize in a first step that the statistic $(S_1, \dots, S_n, T_1, \dots, T_m)$ which consists of the sums of the rows and of the columns is sufficient, a fact that has already been established in Andersen [2,3] or on page 41 in Fischer and Molenaar [15] for the standard Rasch model and extended to the polytomous Rasch model in Andrich [4]. But the above representation shows that one can reduce the mixture Rasch model further to the statistic $(T_1, \dots, T_m, N_0, \dots, N_m)$ in a second step. As $\sum_{k=1}^m T_k = \sum_{k=0}^m k \cdot N_k$ and $\sum_{k=0}^m N_k = n$ we may remove one of the components of $(N_0, \dots, N_m)$ and one of the components of $(T_1, \dots, T_m)$ without losing sufficiency of the statistic. Especially the statistic $(T, N) = (T_1, \dots, T_{m-1}, N_1, \dots, N_m)$ is sufficient and takes its values in $\{0, \dots, n\}^{2m-1}$. Denoting the distribution of $(T, N)$ under $P^A_{\theta,F}$ by $P^B_{\theta,F}$ we arrive at the experiment

$$\mathcal{B}_{n,m} := \big(\{0,\dots,n\}^{2m-1}, \mathcal{P}\big(\{0,\dots,n\}^{2m-1}\big), \big(P^B_{\theta,F}\big)_{(\theta,F)\in\Theta\times\mathcal{F}}\big). \tag{3.4}$$

As explained in Section 2, sufficiency implies equivalence in Le Cam's sense so that we obtain the following statement.

**Theorem 3.1.** *The experiments $\mathcal{A}_{n,m}$ in (3.2) and $\mathcal{B}_{n,m}$ in (3.4) are equivalent.*

Put $\langle\theta, b\rangle = \sum_{i=1}^m \theta_i b_i$, $b \in \{0,1\}^m$ and

$$\mathbb{S}(k,m) = \left\{ b \in \{0,1\}^m : \sum_{i=1}^m b_i = k \right\}.$$

To study the distribution of $Y_j$ on $\{0,1\}^m = \bigcup_{k=0}^m \mathbb{S}(k,m)$, we deduce from (3.2) that

$$P^A_{\theta,F}(Y_j = b) = \exp\bigl\{-\langle\theta, b\rangle + G(k,\theta,F)\bigr\}, \qquad b \in \mathbb{S}(k,m). \tag{3.5}$$

Moreover, the $Y_j$ are i.i.d. which implies that $(N_0, \dots, N_m)$ has the multinomial distribution $\mathsf{M}_{n,m,\theta,F}$ with the cell probabilities

$$\begin{aligned} q_k(\theta,F) &= \sum_{b\in\mathbb{S}(k,m)} \int \left\{ \prod_{i=1}^m \frac{\exp\{b_i(\beta-\theta_i)\}}{1+\exp\{\beta-\theta_i\}} \right\} dF(\beta) \\ &= \sum_{b\in\mathbb{S}(k,m)} \exp\bigl\{-\langle\theta,b\rangle + G(k,\theta,F)\bigr\}. \end{aligned} \tag{3.6}$$

The conditional distribution $\Gamma_{\theta,F}(\cdot|i)$ of $Y_j$ given $S_j = i$ has the probability mass function

$$\begin{aligned} \Gamma_{\theta,F}\bigl(\{b\}|i\bigr) &:= P^A_{\theta,F}(Y_j = b | S_j = i) \\ &= \frac{\exp\{-\langle\theta,b\rangle\}}{\sum_{c\in\mathbb{S}(i,m)} \exp\{-\langle\theta,c\rangle\}} \cdot 1_{\mathbb{S}(i,m)}(b) \\ &= \frac{\exp(-\sum_{k=1}^{m-1} \vartheta_k b_k)}{\sum_{c\in\mathbb{S}(i,m)} \exp(-\sum_{k=1}^{m-1} \vartheta_k c_k)} \cdot 1_{\mathbb{S}(i,m)}(b), \end{aligned}$$

where $\vartheta_k := \theta_k - \theta_m$. Writing $Y_j^{m-1} = (X_{j,1}, \dots, X_{j,m-1})$ the event $\{Y_j^{m-1} = b\}$ equals the union of $\{Y_j = (b,0)\}$ and $\{Y_j = (b,1)\}$ for any $b \in \{0,1\}^{m-1}$ so that

$$P^A_{\theta,F}\bigl(Y_j^{m-1} = b \mid S_j = i\bigr) = \frac{\exp(-\sum_{k=1}^{m-1} \vartheta_k b_k)}{\sum_{c\in\mathbb{B}(i,m)} \exp(-\sum_{k=1}^{m-1} \vartheta_k c_k)} \cdot 1_{\mathbb{B}(i,m)}(b), \tag{3.7}$$

where $\mathbb{B}(i,m) := \mathbb{S}(i-1, m-1) \cup \mathbb{S}(i, m-1)$. The conditional distribution of $Y_j^{m-1}$ given $S_j = i$ under $P^A_{\theta,F}$ is denoted by $\mathsf{U}_{\vartheta,F}(\cdot \mid i)$. As the random vectors $(Y_j^{m-1}, S_j)$, $j = 1, \dots, n$, are independent the conditional distribution of $T$ in (3.3) given $S_1, \dots, S_n$ under $P^A_{\theta,f}$ turns out to be

$$\mathcal{L}(T \mid S_1, \dots, S_n) = *_{j=1}^n \mathsf{U}_{\vartheta,F}(\cdot \mid S_j) = *_{k=1}^m \mathsf{U}_{\vartheta,F}^{*N_k}(\cdot \mid k),$$

where $*$ denotes convolution. Therein we have used that convolution is a commutative operation and that $\mathsf{U}_{\vartheta,F}(\cdot|0) = \delta_0$. Since the random measure $\mathcal{L}(T \mid S_1, \dots, S_n)$ is measurable in the $\sigma$-field

generated by $N$ we conclude that

$$\mathcal{L}(T \mid N) = *_{k=1}^{m} \mathsf{U}_{\vartheta,F}^{*N_k}(\cdot \mid k).$$

This proves the following theorem.

**Theorem 3.2.** *For the observation* $(T, N) = (T_1, \ldots, T_{m-1}, N_1, \ldots, N_m)$ *in the experiment* $\mathcal{B}_{n,m}$ *in* (3.4), *the random vector* $(N_0 := n - \sum_{i=1}^{m} N_i, N_1, \ldots, N_m)$ *has a multinomial distribution with the cell probabilities* $q_k(\theta, F)$; *and* $*_{i=1}^{m} \mathsf{U}_{\vartheta,F}^{*N_i}(\cdot \mid i)$ *is the conditional distribution of* $T$ *given* $N$.

It is remarkable that the conditional distribution of $T$ given $N$ does not depend on the ability distribution $F$ but only on the difficulty parameter $\theta$ since the distributions $\mathsf{U}_{\vartheta,F}(\cdot \mid i)$ do not depend on $F$, see (3.7) and Theorem 3.2. This fact has also been mentioned for example, in Pfanzagl [27] and Strasser [36,37].

# 4. High-dimensional Gaussian approximation

In this section, we establish a general result on the approximation of the sum of high-dimensional independent binary random vectors by Gaussian models. Later, we will apply this finding to the experiment $\mathcal{B}_{n,m}$. The results of Carter [10], which are restricted to multinomial experiments, are included in a special setting. In particular, Carter's results are not applicable to the statistic $T$ in the experiment $\mathcal{B}_{n,m}$. Moreover, we use a completely different strategy of proofs.

The starting point of this section is a triangular array of independent binary vectors $Y_{i,n} = (X_{1,i,n}, \ldots, X_{d,i,n})$ where the dimension $d = d_n$ is allowed to tend to infinity moderately with respect to $n$. That rate will be made precise later. We write $W_n := y_0 + \sum_{i=1}^{n} Y_{i,n}$ for any deterministic $y_0 \in \mathbb{Z}^d$. As $W_n$ is a discrete random vector which takes its values in $\mathbb{Z}^d$ one cannot approximate the measure $P_{W_n}$ of $W_n$ by a continuous probability measure such as a normal distribution in the total variation sense. Therefore, one has to apply a smoothing procedure to $W_n$. Concretely, a $d$-dimensional random vector $U$ is generated independently of $W_n$ and, then, $W_n$ and $U$ are added so that we consider the absolutely continuous probability measure $\mathcal{L}(W_n + U) = \mathcal{L}(W_n) * \mathcal{L}(U)$.

Now suppose that $W_n$ represents the observation in a statistical experiment. Then the Markov kernel $K(x, \cdot) := \mathcal{L}(U + x)$ transforms $\mathcal{L}(W_n)$ into $\mathcal{L}(W_n + U)$. As an attempt for the inverse transformation, one could round each component of $W_n + U$ and denote the outcome by $[W_n + U]$. Carter [10] applies this strategy where $U$ is uniformly distributed on the cube $[-1/2, 1/2]^d$. Then $[W_n + U] = W_n$ so that the original data are reconstructed by the rounding procedure. In this case the experiment in which one observes $W_n$ is equivalent to the experiment in which the observation is $W_n + U$.

It turns out that, in the experiment $\mathcal{B}_{n,m}$, the approach which involves uniformly distributed $U$ would require $d_n$ to increase only at a logarithmic rate in $n$ in order to obtain asymptotic equivalence to a Gaussian model. It is really surprising that polynomial rates can be achieved by a simple idea. We replace the uniform distribution by a centered normal distribution where the

scaling factor of the covariance matrix tends to infinity in a suitable manner. We consider $\mathcal{L}(U) = \mathsf{N}(0, b_n I)$ where $I$ denotes the $d \times d$-identity matrix and the sequence $(b_n)_n$ is allowed to tend to infinity. Now the random vector $W_n$ cannot be identified from $W_n + U$ but we will show that the total variation distance between $\mathcal{L}(W_n)$ and $\mathcal{L}([W_n + U])$ still tends to zero (uniformly with respect to the parameter) under some constraints so that the experiment in which one observes $W_n$ is asymptotically equivalent to the experiment which describes the observation of $W_n + U$.

We introduce the notation

$$Y_{i,n}^{-j} := (X_{1,i,n}, \dots, X_{j-1,i,n}, X_{j+1,i,n}, \dots, X_{d,i,n}),$$

and

$$p_{j,i} := E\big(X_{j,i,n} \mid Y_{i,n}^{-j}\big), \qquad \mu_j := \sum_{i=1}^{n} p_{j,i}, \qquad \sigma_j^2 := \sum_{i=1}^{n} p_{j,i}(1 - p_{j,i}).$$

Moreover, we define

$$\kappa := \inf_{\theta' \in \Theta'} \min_{i=1,\dots,n} \min_{j=1,\dots,d} E p_{j,i}(1 - p_{j,i}), \tag{4.1}$$

when the distributions of the $Y_{i,n}$ and $y_0 \in \mathbb{Z}^d$ are indexed by a parameter $\theta'$, which lies in a set $\Theta'$. Small values of $\kappa$ indicate that there are strong dependencies between some components. Moreover, we will give lower bounds on the smallest eigenvalue of the covariance matrix of $\sum_{i=1}^{n} Y_{i,n}$ in Lemma 4.3. We derive the following central theorem (CLT) which allows to approximate statistical experiments, in which one observes a sum of independent binary random vectors, by Gaussian experiments. Then the experiment $\mathcal{X}_n$ describes the observation of the random vector $W_n$. Furthermore, we define the Gaussian experiment $\mathcal{Z}_n$ by

$$\mathcal{Z}_n := \big(\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d), \{\mathsf{N}(\overline{\mu}_{\theta'}, \overline{\Lambda}_{\theta'})\}_{\theta' \in \Theta'}\big).$$

The above consideration leads to the following theorem, which is one of our main results.

**Theorem 4.1.** *Suppose that $\kappa = \kappa_n > n^{-1/2+\delta}$ for some fixed $\delta > 0$ and $n$ sufficiently large; and that $\inf(b_n)_n > 0$. Then the Le Cam distance between the experiments $\mathcal{X}_n$ and $\mathcal{Z}_n$ satisfies*

$$\Delta(\mathcal{X}_n, \mathcal{Z}_n) \le \text{const.} \cdot \big(\{\log(n\kappa_n)\}^{1/2} n^{-1/2} \kappa_n^{-1/2} d_n b_n^{1/2} + b_n^{-1/2} d_n^4/\kappa_n + b_n n^{-1} \kappa_n^{-1} d_n^{3/2}\big),$$

*for some universal constant.*

**Remark 4.1.** The Markov kernel which transforms $\mathcal{X}_n$ into $\mathcal{Z}_n$ in Theorem 4.1 equals $x \mapsto \mathsf{N}(x, b_n I)$, $x \in \mathbb{Z}^d$; and the inverse transformation is carried out by rounding each component of the observation from $\mathcal{Z}_n$.

Pointing out the dominating terms, the upper bound on the Le Cam distance which is provided in Theorem 4.1 converges to zero as $n \to \infty$ whenever

$$\lim_{n\to\infty} \{\log(n\kappa_n)\} n^{-1} \kappa_n^{-1} d_n^2 b_n + b_n^{-1} d_n^8 \kappa_n^{-2} = 0. \tag{4.2}$$

Note that, in order to satisfy (4.2), the scaling factor $(b_n)_n$ must be chosen such that $(b_n)_n$ tends to infinity (as $n \to \infty$), but not too fast with respect to the sequences $(\kappa_n)_n$ and $(d_n)_n$.

Theorem 4.1 is a consequence of a series of lemmata. First, let us show asymptotic proximity between $\mathcal{L}(W_n)$ and its shifted versions.

**Lemma 4.1.** *Fix any $\delta > 0$ such that $\kappa > n^{-1/2+\delta}$ and $n\kappa > 2$. Then the total variation distance between $\mathcal{L}(W_n)$ and $\mathcal{L}(W_n + l)$, for some deterministic $l \in \mathbb{Z}^d$, obeys the following upper bound*

$$\mathsf{TV}\big(\mathcal{L}(W_n), \mathcal{L}(W_n + l)\big) \le A\big\{\log(n\kappa)\big\}^{1/2} n^{-1/2}\kappa^{-1/2} \sum_{j=1}^{d} |l_j|,$$

*for a universal constant $A \in (0, \infty)$.*

Lemma 4.1 represents a robustness property of $\mathcal{L}(W_n)$ with respect to shifting the measure on the $\mathbb{Z}^d$-grid. That provides the major tool to establish the upper bound on the total variation distance between $\mathcal{L}(W_n)$ and $\mathcal{L}([W_n + U])$ as provided in the next lemma.

**Lemma 4.2.** *Under the conditions of Lemma 4.1 we have, for $\mathcal{L}(U) = \mathsf{N}(0, b_n I)$, $b_n > 0$, that*

$$\mathsf{TV}\big(\mathcal{L}(W_n), \mathcal{L}\big([W_n + U]\big)\big) \le A\big\{\log(n\kappa)\big\}^{1/2} n^{-1/2}\kappa^{-1/2} d_n\big(1/2 + b_n^{1/2}\big)$$

*with $A$ as in Lemma 4.1.*

Clearly, $b_n$ must not be chosen too large as it represents the noise level of the smoothing normal random vector. If the regime of blurring the observation becomes too intensive, then the total variation distance between the laws of $W_n$ and its contaminated and, afterwards, componentwise rounded version will not tend to zero. That is made precise by the statement of the upper bound in the lemma. Thus, if this right hand side of the inequality in Lemma 4.2 tends to zero (uniformly with respect to a family of admitted measures of the $Y_{i,n}$, $i = 1, \dots, n$), the observation of $W_n$, on the one hand, and of $W_n + U$, on the other hand, represent asymptotically equivalent experiments.

In the next step, we will approximate the smoothed distribution $\mathcal{L}(W_n + U)$ by the normal distribution whose expectation vector and covariance matrix coincide with those of $W_n + U$. We establish a CLT for independent binary random vectors with increasing dimension in the total variation sense. We write $\overline{\mu}_i$ and $\overline{\Lambda}_i$ for the expectation vector and the covariance matrix of $Y_{i,n}$, respectively. Accordingly, $\overline{\mu} = y_0 + \sum_{i=1}^n \overline{\mu}_i$ and $\overline{\Lambda} = \sum_{i=1}^n \overline{\Lambda}_i$ are the corresponding quantities of $W_n$. Preparatory to this CLT, we provide a positive lower bound on the eigenvalues of partial sums of the matrices $\overline{\Lambda}_i$.

**Lemma 4.3.** *For any $\mathcal{N} \subseteq \{1, \dots, n\}$, all eigenvalues of the matrix $\sum_{i \in \mathcal{N}} \overline{\Lambda}_i$, are bounded from below by $(\#\mathcal{N})\kappa/d$.*

Besides Lemma 4.3 also yields invertibility of the matrix $\overline{\Lambda}$ whenever $\kappa > 0$. Another important result which will be used to derive the CLT is the asymptotic proximity of the smoothed version of each $\mathcal{L}(Y_{i,n})$ (i.e. $\mathcal{L}(Y_{i,n})$ convolved with some normal distribution $\mathsf{N}(0, \tilde{\Lambda})$) and the

normal distribution with the same expectation vector and covariance matrix as $\mathcal{L}(Y_{i,n}) * \mathsf{N}(0, \tilde{\Lambda})$. We provide

**Lemma 4.4.** *Let $\tilde{\Lambda}$ be some positive definite $d \times d$-matrix. Then,*

$$\mathsf{TV}\big(\mathcal{L}(Y_{i,n}) * \mathsf{N}(0, \tilde{\Lambda}), \mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i + \tilde{\Lambda})\big) \leq B\lambda^{-3/2} d_n^3,$$

*for a universal constant $B \in (0, \infty)$ where $\lambda$ denotes the smallest eigenvalue of the matrix $\tilde{\Lambda}$.*

Lemma 4.4 forms the main tool for the proof of the following lemma. We are now ready to prove a strong CLT for sums of independent binary random vectors.

**Lemma 4.5.** *If $(b_n)_n$ is bounded away from zero, then*

$$\mathsf{TV}\big(\mathcal{L}(W_n) * \mathsf{N}(0, b_n I), \mathsf{N}(\overline{\mu}, \overline{\Lambda} + b_n I)\big) \leq C b_n^{-1/2} d_n^4 / \kappa_n,$$

*for $\kappa = \kappa_n$ with a universal constant $C$.*

By Lemma 4.5, the blurred version of $W_n$ (normal mixture) may be replaced by a fully Gaussian random variable with the law $\mathsf{N}(\overline{\mu}, \overline{\Lambda} + b_n I)$ where $\overline{\mu}$ and $\overline{\Lambda}$ represent the expectation vector and the covariance matrix of the original random vector $W_n$. For that purpose, the noise level $b_n$ shall be selected as large as possible to make the upper bound on the total variation as small as possible; while respecting the constraints on $b_n$ which come from Lemma 4.2.

The term $b_n I$ remains to be removed in the covariance matrix of the new random vector. By a famous formula which governs the Hellinger distance between normal distributions we deduce

**Lemma 4.6.** *We have that*

$$\mathsf{TV}\big(\mathsf{N}(\overline{\mu}, \overline{\Lambda} + b_n I), \mathsf{N}(\overline{\mu}, \overline{\Lambda})\big) \leq 2\sqrt{2} b_n n^{-1} \kappa_n^{-1} d_n^{3/2}.$$

This upper bound resembles that of Lemma 4.2. It represents an analogue of Lemma 4.2 in the Gaussian world. Again the noise level $b_n$ must be chosen as small as possible so that the two normal distributions become close to each other (in the total variation sense), which is consistent with the conditions on $b_n$ from that lemma.

Piecing together the Lemmata 4.2, 4.5 and 4.6 yields Theorem 4.1.

# 5. Gaussian approximation of the mixture Rasch model

In this section, we apply the general Gaussianization scheme provided in Section 4 and, in particular, in Theorem 4.1 to the experiment $\mathcal{B}_{n,m}$ in (3.4). Therein we distinguish between the

statistics $T$ and $N$. Obviously $d_n$ from Theorem 4.1 equals $m-1$ and $m$ for the statistic $T$ and $N$, respectively, while the quantity $\kappa_n$ has to be studied in details in both settings.

## 5.1. Gaussian model for the difficulty parameters

The new statistical experiment, which is denoted by $\mathcal{C}_{n,m}$, describes the observation of $(T^*, N)$ where $N$ is as in the experiment $\mathcal{B}_{n,m}$ and $T^*$ is an $(m-1)$-dimensional random vector whose conditional distribution given $N$ is $\mathsf{N}(E^B_{\theta,F}(T|N), \operatorname{cov}^B_{\theta,F}(T|N))$. We define the experiments

$$\mathcal{B}^{n'}_{n,m} := \big(\{0,\dots,n\}^{m-1}, \mathcal{P}\big(\{0,\dots,n\}^{m-1}\big), \big(\mathcal{L}^B_{\theta,F}\big(T \mid N = n'\big)\big)_{\theta,F}\big),$$
$$\mathcal{C}^{n'}_{n,m} := \big(\mathbb{R}^{m-1}, \mathfrak{B}\big(\mathbb{R}^{m-1}\big), \big(\mathcal{L}^B_{\theta,F}\big(T^* \mid N = n'\big)\big)_{\theta,F}\big).$$

Now we consider sequences of experiments indexed by the random vector $N$. Note that

$$\Delta(\mathcal{B}_{n,m}, \mathcal{C}_{n,m}) \le \sup_{\theta,F} E^B_{\theta,F} \Delta\big(\mathcal{B}^N_{n,m}, \mathcal{C}^N_{n,m}\big). \tag{5.1}$$

By Theorem 3.2 the observation in the experiment $\mathcal{B}^{n'}_{n,m}$ can be written as the sum of $n'_1 + \cdots + n'_{m-1}$ independent binary random vectors so that it has the structure of the random vector $W_n$ from Section 4 when putting $y_0 = n'_m \cdot (1,\dots,1)$. The following lemma gives us a lower bound on $\kappa$ in (4.1).

**Lemma 5.1.** *Assuming that* $\mathcal{L}(W_n - y_0 \mid N) = *_{k=1}^{m-1} \mathsf{U}^{*N_k}_{\vartheta,F}(\cdot \mid k)$ *in the notation of Section* 4; *and that* $m \ge 3$, *the quantity* $\kappa$ *in* (4.1) *satisfies*

$$\kappa \ge \frac{\exp(-6R)}{(m-1)(1+\exp(2R))}.$$

Note that the number of $Y_{i,n}$, which is denoted by $n$ in Section 4, equals $n'_1 + \cdots + n'_{m-1}$ in the experiment $\mathcal{B}^{n'}_{n,m}$. To show that it has at least the order $n$ from the mixture Rasch model, the following assumption and lemma are required. We impose that every distribution $F$ in $\mathcal{F}$ has a Lebesgue density $f$; and that there exists an enveloping function $\overline{f}$ with $\int \overline{f}(x)\,dx < \infty$ such that

$$f \le \overline{f} \qquad \text{a.e.}, \forall F \in \mathcal{F}. \tag{5.2}$$

Condition (5.2) represents a tightness property of $\mathcal{F}$. Then is the following lemma.

**Lemma 5.2.** *Under the conditions* (5.2), $m \ge 3$ *and* $\rho \in (0,1)$ *sufficiently close to* 1, *we have that*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta, F\in\mathcal{F}} P^B_{\theta,F}(N_0 + N_m > \rho n) = 0.$$

By (5.1), we deduce for some $\rho \in (0,1)$ from Lemma 5.2 that

$$\Delta(\mathcal{B}_{n,m}, \mathcal{C}_{n,m}) \le 2 \sup_{\theta,F} P^B_{\theta,F}(N_0 + N_m > \rho n) + \sup_{\theta,F} E^B_{\theta,F} 1_{[(1-\rho)n,\infty)} \left( \sum_{j=1}^{m-1} N_j \right) \cdot \Delta\big(\mathcal{B}^N_{n,m}, \mathcal{C}^N_{n,m}\big),$$

where the latter term tends to zero as $n \to \infty$ whenever

$$\lim_{n\to\infty} n^{-1} m^3 b_n + b_n^{-1} m^{10} = 0,$$

thanks to Lemma 5.1, Theorem 4.1 and equation (4.2). The convergence of the first term is guaranteed by Lemma 5.2. We establish asymptotic equivalence between the experiments $\mathcal{B}_{n,m}$ and $\mathcal{C}_{n,m}$ under some constraints.

**Theorem 5.1.** *Assume* (5.2); $m = m_n \ge 3$; *that there is some* $\beta > 13$ *such that* $\sup_n m_n^\beta / n < \infty$. *Then the selection* $b_n \asymp n^{(7+\beta)/(2\beta)}$ *yields asymptotic equivalence of the experiments* $\mathcal{B}_{n,m}$ *and* $\mathcal{C}_{n,m}$; *moreover*

$$\Delta(\mathcal{B}_{n,m}, \mathcal{C}_{n,m}) = \mathcal{O}\big(n^{(13-\beta)/(4\beta)} (\log n)^{1/2}\big).$$

Let us consider the conditional Gaussian distribution of the statistic $T^*$ given $N$ in the experiment $\mathcal{C}_{n,m}$. Since $\mathcal{L}(T \mid N) = *_{k=0}^m \mathsf{U}_{\vartheta,F}^{*N_k}(\cdot \mid N)$ with $(T, N)$ as in the experiment $\mathcal{B}_{n,m}$ we have that $E^B_{\theta,F}(T \mid N) = -\nabla \Psi_N(\vartheta)$ and $\operatorname{cov}^B_{\theta,F}(T \mid N) = \Delta \Psi_N(\vartheta)$ where

$$\Psi_{n'}(\vartheta) := \sum_{k=0}^m n'_k \cdot \log\left( \sum_{b \in \mathbb{S}(k,m)} \exp\left\{ -\sum_{l=1}^{m-1} \vartheta_l b_l \right\} \right), \qquad \vartheta = (\vartheta_1, \dots, \vartheta_{m-1}),$$

and $\nabla$ and $\Delta$ denote the gradient and the Hessian matrix, respectively.

We introduce the experiment $\mathcal{D}_{n,m}$ by

$$\mathcal{D}_{n,m} := \big(\mathbb{R}^{2m-1}, \mathfrak{B}\big(\mathbb{R}^{2m-1}\big), \big(\mathcal{L}_{\theta,F}\big(N, T^{**}\big)\big)_{\theta\in\Theta, F\in\mathcal{F}}\big),$$

where $N$ is as in the experiment $\mathcal{C}_{n,m}$ and the conditional distribution of $T^{**}$ given $N$ equals $\mathsf{N}(\vartheta, \{\Delta\Psi_N(\vartheta)\}^{-1})$ if $N_0 + N_m < n$; otherwise put $T^{**} = 0$. By the Lemmata 4.3 and 5.1, the matrix $\Delta\Psi_N(x)$ is invertible for all $x \in \mathbb{R}^{m-1}$ on the event $\{N_0 + N_m < n\}$. Therein note that, for any $x \in \mathbb{R}^{m-1}$, there exist some $R > 0$ and $\theta \in \Theta$ such that $\vartheta = x$. That also implies injectivity of the mapping $x \mapsto \nabla\Psi_N(x)$ on the domain $\mathbb{R}^{m-1}$ in the case of $N_0 + N_m < n$. Now define the function $\Phi$ by

$$\Phi\big(x, n'\big) := \begin{cases} \big(\nabla\Psi_{n'}(x), n'\big), & \text{if } n'_1 + \dots + n'_{m-1} \ne 0, \\ \big(0, n'\big), & \text{otherwise.} \end{cases}$$

By $\mathcal{D}'_{n,m}$ we define the experiment in which one observes $\Phi(T^{**}, N)$ with $(T^{**}, N)$ as in $\mathcal{D}_{n,m}$. Clearly $N$ is uniquely reconstructible from $\Phi(T^{**}, N)$. If $N_1 + \cdots + N_{m-1} = 0$ then $T^{**} = 0$; otherwise the injectivity of $x \mapsto \nabla\Psi_N(x)$ enables us to identify $T^{**}$. Therefore, the experiments $\mathcal{D}_{n,m}$ and $\mathcal{D}'_{n,m}$ are equivalent in Le Cam's sense. Then it suffices to establish that

$$\lim_{n\to\infty} \sup_{\theta\in\Theta, F\in\mathcal{F}} \mathsf{TV}\big(\mathcal{L}^D_{\theta,F}(\Phi), P^C_{\theta,F}\big) = 0, \tag{5.3}$$

in order to show the following theorem.

**Theorem 5.2.** *Under the conditions of Theorem 5.1 the experiments $\mathcal{C}_{n,m}$ and $\mathcal{D}_{n,m}$ are asymptotically equivalent as $n \to \infty$.*

The experiment $\mathcal{D}_{n,m}$ has the advantage compared to $\mathcal{C}_{n,m}$ that the directly observed statistic $T^{**}$ represents an asymptotically unbiased estimator of $\vartheta$. This will be exploited in Section 6.

## 5.2. Gaussian model for the ability distribution

We focus on the statistic $N$ in the experiment $\mathcal{D}_{n,m}$. Note that $(N_0, N)$ is a multinomial random vector by Theorem 3.2. If we can show that the sub-experiment in which only $N$ is observed is asymptotically equivalent to the experiment which describes the observation of $N^*$ with

$$\mathcal{L}\big(N^*\big) = \mathsf{N}\big(E^D_{\theta,F} N, \operatorname{cov}^D_{\theta,F}(N)\big),$$

then we have asymptotic equivalence of $\mathcal{D}_{n,m}$ and the experiment $\mathcal{E}_{n,m}$ which is defined by

$$\mathcal{E}_{n,m} := \big(\mathbb{R}^{2m-1}, \mathfrak{B}\big(\mathbb{R}^{2m-1}\big), \big(\mathcal{L}_{\theta,F}\big(T^{**}, N^*\big)\big)_{\theta\in\Theta, F\in\mathcal{F}}\big),$$

such that $\mathcal{L}^E_{\theta,F}(T^{**} \mid N^*) = \mathsf{N}(\vartheta, \{\Delta\Psi_{[N^*]_+}(\vartheta)\}^{-1})$ if $[N^*_1]_+ + \cdots + [N^*_{m-1}]_+ > 0$ (put $T^{**} := 0$ otherwise) where $[x]_+$ denotes $(\max\{[x_j], 0\})_{j=1,\ldots,m}$ for any $x \in \mathbb{R}^m$. Note that, for all $\theta \in \Theta$ and $F \in \mathcal{F}$, we have that

$$\mathcal{L}^E_{\theta,F}\big(T^{**} \mid N^* = N\big) = \mathcal{L}^D_{\theta,F}\big(T^{**} \mid N\big), \qquad \text{a.s.},$$

for $N$ as in the experiment $\mathcal{D}_{n,m}$. Moreover, by the multinomial distribution of $N$, we immediately derive that

$$\begin{aligned} E^D_{\theta,F} N &= n \cdot \tilde{q}(\theta, F) := n \cdot \big(q_1(\theta, F), \ldots, q_m(\theta, F)\big)^T, \\ \operatorname{cov}^D_{\theta,F}(N) &= n \cdot \big(\tilde{Q}(\theta, F) - \tilde{q}(\theta, F)\tilde{q}(\theta, F)^T\big), \end{aligned}$$

where $q_k(\theta, F)$ is as in (3.6) and $\tilde{Q}(\theta, F)$ denotes the $(m-1) \times (m-1)$-diagonal matrix whose $(k, k)$th entry equals $q_k(\theta, F)$. The asymptotic equivalence of $\mathcal{D}_{n,m}$ and $\mathcal{E}_{n,m}$ is shown by a direct application of Theorem 4.1 where the quantity $\kappa$ in (4.1) has to be bounded from below again.

Therefore, a constraint on the tail behaviour of the Lebesgue density $f$ of the ability distribution $F$ is required; concretely we assume that

$$f(x) \geq D_0 \exp\{-D_1|x|\}, \qquad \forall x \in \mathbb{R}, F \in \mathcal{F}, \tag{5.4}$$

for some universal positive constants $D_0$ and $D_1$. As an alternative for condition (5.4), we may consider $m = m_n$ as bounded with respect to $n$. Then Gaussian models for $F$ are still included. In the notation of Section 4, one may conclude that

$$\kappa \geq \inf_{\theta, F} \inf_{k=1,\dots,m} q_0(\theta, F) q_k(\theta, F) / \big(q_0(\theta, F) + q_k(\theta, F)\big).$$

Thus, a lower bound on the $q_k(\theta, F)$ is needed.

**Lemma 5.3.** *Under condition* (5.4), *we obtain that*

$$\inf_{\theta, F} \inf_{k=0,\dots,m} q_k(\theta, F) \geq \text{const.} \cdot m^{-3/2-D_1},$$

*for a universal constant factor.*

Hence, $\kappa_n \geq \text{const.} \cdot m^{-3/2-D_1}$ so that, by (4.2), the following statement follows.

**Theorem 5.3.** *Assume the constraints of Theorem* 5.1; *condition* (5.4); *and the existence of some* $\beta > 3D_1 + 29/2$ *such that* $\sup_n m_n^\beta / n < \infty$. *Then the selection* $b_n \asymp n^{(2\beta+2D_1+15)/(4\beta)}$ *yields that*

$$\Delta(\mathcal{D}_{n,m}, \mathcal{E}_{n,m}) = \mathcal{O}\big(n^{(29+6D_1-2\beta)/(8\beta)} (\log n)^{1/2}\big),$$

*so that the experiments* $\mathcal{D}_{n,m}$ *and* $\mathcal{E}_{n,m}$ *are asymptotically equivalent as* $n \to \infty$.

Instead of condition (5.4) one can assume that $m = m_n$ is bounded in $n$ and the claim of Theorem 5.3 remains valid.

Thanks to the marginal multinomial distribution of the statistic $N$ in the experiment $\mathcal{B}_{n,m}$ a transformation of the experiment $\mathcal{E}_{n,m}$ (in particular, of the statistic $N^*$) is possible in order to obtain independent components. Similar arguments have been used in Carter [10]. We introduce the $(m+1)$-dimensional random vector $N^{**}$ with $\mathcal{L}(N^{**}) = \mathsf{N}(nq(\theta, F), nQ(\theta, F))$ where

$$q(\theta, F) := \big(q_0(\theta, F), \dots, q_m(\theta, F)\big)^T,$$
$$Q(\theta, F) := \big\{1_{\{j\}}(k) q_k(\theta, F)\big\}_{j,k=0,\dots,m}.$$

Then the conditional distribution of $T^{**}$ given $N^{**}$ equals $\mathsf{N}(\vartheta, \{\Delta\Psi_{[\tau(N^{**})]_+}(\vartheta)\}^{-1})$ on the event $\{[\tau(N^{**})]_{+,1} + \dots + [\tau(N^{**})]_{+,m+1} > 0\}$ (again $T^{**} := 0$ otherwise), where the function $\tau$ from $\mathbb{R}^{m+1}$ to $\mathbb{R}^m$ is defined by

$$\tau : x = (x_0, \dots, x_m) \mapsto (x_1, \dots, x_m) \cdot n \Big/ \max\left\{\zeta, \sum_{j=0}^m x_j\right\},$$

for some deterministic $\zeta > 0$ still to be chosen. We consider the experiment

$$\mathcal{F}_{n,m} := \big(\mathbb{R}^{2m}, \mathfrak{B}\big(\mathbb{R}^{2m}\big), \big(\mathcal{L}_{\theta,F}\big(T^{**}, N^{**}\big)\big)_{\theta,F}\big).$$

In order to show asymptotic equivalence of $\mathcal{E}_{n,m}$ and $\mathcal{F}_{n,m,}$ we consider the statistic $N^{**}$ from the experiment $\mathcal{F}_{n,m}$ and the sum of its components, which we call $V$. As $N^{**}$ can be uniquely reconstructed from $(\tau(N^{**}), V)$ we derive equivalence of $\mathcal{F}_{n,m}$ and the experiment $\mathcal{F}'_{n,m}$ in which $(T^{**}, \tau(N^{**}), V)$ is observed. It can be shown that

$$\begin{aligned}
&\mathcal{L}_{\theta,F}\big(N^{**} \mid V\big) = \mathsf{N}\big(Vq(\theta,F), nQ(\theta,F) - nq(\theta,F)q(\theta,F)^T\big),\\
&\mathcal{L}_{\theta,F}\big(\tau\big(N^{**}\big) \mid V\big)\\
&\quad = \mathsf{N}\big(n\tilde{q}(\theta,F)V/\max\{T,\zeta\}, n^3\big(\tilde{Q}(\theta,F) - \tilde{q}(\theta,F)\tilde{q}(\theta,F)^T\big)/\big(\max\{V,\zeta\}\big)^2\big).
\end{aligned}$$

The following asymptotic approximation is required.

**Lemma 5.4.** *Assume the conditions of Theorem 5.3 and select $\zeta = n/2$. Then,*

$$\lim_{n\to\infty} \sup_{\theta,F} E_{\theta,F}\mathsf{TV}\big(\mathcal{L}_{\theta,F}\big(\tau\big(N^{**}\big) \mid V\big), \mathsf{N}\big(n\tilde{q}(\theta,F), n\tilde{Q}(\theta,F) - n\tilde{q}(\theta,F)\tilde{q}(\theta,F)^T\big)\big) = 0.$$

As the conditional distribution of $T^{**}$ given $\tau(N^{**})$ and $V$ equals that given $\tau(N^{**})$, Lemma 5.4 provides asymptotic equivalence of $\mathcal{F}'_{n,m}$ and $\mathcal{F}''_{n,m}$ where the latter experiment describes the observation of $(T^{**}, N^*, V^*)$ where $V^*$ and $V$ are identically distributed but $V^*$ and $(T^{**}, N^*)$ are independent. As, in addition, $\mathcal{L}_{\theta,F}(V^*) = \mathsf{N}(n,n)$, the distribution of $V^*$ does not depend on $\theta$ or $F$ and, thus, $V^*$ can be omitted without losing any information on $(\theta, F)$. Therefore, $\mathcal{F}''_{n,m}$ and $\mathcal{E}_{n,m}$ are equivalent so that the following result has been established.

**Theorem 5.4.** *Under the conditions of Theorem 5.3, the experiments $\mathcal{E}_{n,m}$ and $\mathcal{F}_{n,m}$ are asymptotically equivalent as $n \to \infty$.*

# 6. Applications

In this section, we apply the Gaussian models of Section 5, which have now been proved to be asymptotically equivalent to the mixture Rasch model $\mathcal{A}_{n,m}$, to develop asymptotic inference. In particular we will construct an asymptotic confidence ellipsoid for the difficulty parameters. Thus, the results carry over to the original mixture Rasch model.

Let $T^{**}$ be the part of the observation from the experiment $\mathcal{D}_{n,m}$ where $\mathcal{L}(T^{**} \mid N) = \mathsf{N}(\vartheta, \{\Delta\Psi_N(\vartheta)\}^{-1})$ if $N_1 + \cdots + N_{m-1} > 0$. We define the random ellipsoid

$$\hat{E} := \left\{x \in \mathbb{R}^m : \sum_{j=1}^m x_j = 0, \big(x - ZT^{**}\big)^T Z\big(Z^T Z\big)^{-1}\Delta\Psi_N\big(T^{**}\big)\big(Z^\dagger Z\big)^{-1} Z^\dagger\big(x - ZT^{**}\big) \le \iota\right\},$$

for some threshold $\iota > 0$ to be determined and the $m \times (m-1)$-matrix

$$Z := \begin{pmatrix} 1-1/m & -1/m & \cdots & -1/m \\ -1/m & 1-1/m & \cdots & -1/m \\ \vdots & \vdots & \vdots & \vdots \\ -1/m & -1/m & \cdots & 1-1/m \\ -1/m & -1/m & \cdots & -1/m \end{pmatrix}.$$

Note that $\theta = Z\vartheta$ for all $\theta \in \Theta$ thanks to the definition (3.1). In order to motivate the selection of $\iota$, we give an oracle version of $\hat{E}$ by

$$\tilde{E} := \left\{ x \in \mathbb{R}^m : \sum_{j=1}^m x_j = 0, \left(x - ZT^{**}\right)^\dagger Z\left(Z^\dagger Z\right)^{-1} \Delta\Psi_N(\vartheta)\left(Z^\dagger Z\right)^{-1} Z^\dagger\left(x - ZT^{**}\right) \le \iota \right\}.$$

Concretely, we have replaced $T^{**}$ by its conditional expectation given $N$ in the argument of $\Delta\Psi_N$. Conditionally on $N$ under $\{N_1 + \cdots + N_{m-1} > 0\}$, we may represent $T^{**}$ by $T^{**} = \vartheta + \{\Delta\Psi_N(\vartheta)\}^{-1/2}\varepsilon$ where here $\varepsilon$ denotes an $(m-1)$-dimensional random vector with independent $\mathsf{N}(0,1)$-distributed components. On the event $\{N_1 + \cdots + N_{m-1} > 0\}$, it follows that

$$P^D_{\theta,F}(\theta \in \tilde{E} \mid N) = P\left(|\varepsilon|^2 \le \iota\right).$$

That inspires us to choose $\iota$ as the $(1-\alpha)$-quantile of the $\chi^2(m-1)$-distribution, i.e. $\iota = F^{-1}_{m-1}(1-\alpha)$ where $F_{m-1}$ denotes the $\chi^2(m-1)$-distribution function, for some given $\alpha \in (0,1)$. Then,

$$\liminf_{n\to\infty} \inf_{\theta,F} P^D_{\theta,F}(\theta \in \tilde{E}) \ge 1 - \alpha,$$

as $P^D_{\theta,F}(N_0 + N_m < n)$ tends to zero uniformly in $\theta$ and $F$. Focusing on the ellipsoid $\hat{E}$, we provide the following result.

**Theorem 6.1.** *In the experiment $\mathcal{D}_{n,m}$ we have*

$$\limsup_{n\to\infty} \sup_{\theta,F} P^D_{\theta,F}(\theta \notin \hat{E}) \le \alpha,$$

*under the assumptions of Theorem 5.1 for any fixed $\alpha \in (0,1)$. The maximal length of an axis $\hat{e}$ of $\hat{E}$ obeys the following asymptotic upper bound*

$$\lim_{c\to\infty} \limsup_{n\to\infty} \sup_{\theta,F} P^D_{\theta,F}(\hat{e} > c \cdot m_n/\sqrt{n}) = 0.$$

**Remark 6.1.** Theorem 6.1 shows that $\hat{E}$ is an asymptotic $(1-\alpha)$-confidence ellipsoid for $\theta$ in the experiment $\mathcal{D}_{n,m}$. The maximal axis of this ellipsoid shrinks to zero at the rate $\mathcal{O}_P(m_n/\sqrt{n})$ as $n \to \infty$. By the Theorems 3.1, 3.2, 5.1 and 5.2 the properties established in Theorem 6.1 extend to the original mixture Rasch model (experiment $\mathcal{A}_{n,m}$) after applying the appropriate

Markov kernel which transforms experiment $\mathcal{D}_{n,m}$ to $\mathcal{A}_{n,m}$. Note that the asymptotic confidence region is uniform with respect to the parameter $\theta \in \Theta$ and $F \in \mathcal{F}$. Thus, we have developed a stronger version of asymptotic confidence regions than in the usual setting where $\theta$ and $F$ are viewed as fixed, that is, $\theta$ and $F$ must not change in $n$. This is thanks to the fact that we have used asymptotic approximation with respect to the Le Cam distance rather than central limit laws for specific estimators in terms of weak convergence, where the latter results are commonly used to construct asymptotic confidence regions.

# 7. Conclusions and outlook

In this paper, we derive asymptotically equivalent Gaussian experiments for the mixture Rasch model. In Section 6, asymptotic statistical inference on the difficulty parameters is provided based on these Gaussian experiments. But the asymptotic equivalence of the experiment $\mathcal{F}_{n,m}$ and the original mixture Rasch model $\mathcal{A}_{n,m}$ also opens the perspective for nonparametric inference on the ability distribution. While this goal exceeds the framework of the current paper the authors are working on this issue and intend to present their future results in a separate paper.

# 8. Proofs

**Proof of Lemma 4.1.** Thanks to the shift-invariance of the total variation distance we may put $y_0 = 0$ without any loss of generality. Note that we may write

$$\mathcal{L}(W_n + l) = \mathcal{L}(W_n) * \big(*_{j=1}^{d} \delta_{l_j e_j}\big),$$

where $e_j$ denotes the vector with its $j$th component equal to 1 while all other components vanish. By a telescoping sum, we deduce that

$$\mathsf{TV}\big(\mathcal{L}(W_n + l), \mathcal{L}(W_n)\big) \le \sum_{j=1}^{d} \mathsf{TV}\big(\mathcal{L}(W_n), \mathcal{L}(W_n) * \delta_{l_j e_j}\big) \le \sum_{j=1}^{d} |l_j| \mathsf{TV}\big(\mathcal{L}(W_n), \mathcal{L}(W_n) * \delta_{e_j}\big).$$

We have that

$$\begin{aligned} \mathsf{TV}\big(\mathcal{L}(W_n), \mathcal{L}(W_n) * \delta_{e_j}\big) &= \sum_{w \in \mathbb{Z}^d} |P(W_n = w) - P(W_n = w - \delta_{e_j})| \\ &\le E \sum_{u \in \mathbb{Z}} |P(W_{n,j} = u \mid \mathcal{Y}_{n,j}) - P(W_{n,j} = u - 1 \mid \mathcal{Y}_{n,j})|, \end{aligned} \tag{8.1}$$

where $\mathcal{Y}_{n,j}$ denotes the $\sigma$-field generated by $Y_{1,n}^{-j}, \dots, Y_{n,n}^{-j}$. By Fourier inversion we obtain that

$$P(W_{n,j} = u \mid \mathcal{Y}_{n,j}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp\{-iux\} \psi_{W_{n,j} \mid \mathcal{Y}_{n,j}}(x)\, dx,$$

for all $u \in \mathbb{Z}$ where $\psi_Z$ denotes the characteristic function of a random variable $Z$. Since

$$\begin{aligned}\left|\psi_{W_{n,j}|\mathcal{Y}_{n,j}}(x)\right| &= \prod_{i=1}^{n}\left|\psi_{X_{j,i,n}|Y_{i,n}^{-j}}(x)\right| = \prod_{i=1}^{n}\left|\exp\{ix\}p_{j,i} + 1 - p_{j,i}\right| \\ &\le \prod_{i=1}^{n}\exp\left\{-2p_{j,i}(1-p_{j,i})x^2/\pi^2\right\} = \exp\left\{-2\sigma_j^2 x^2/\pi^2\right\},\end{aligned}$$

for all $x \in [-\pi, \pi]$ it follows that

$$\begin{aligned}&\left|P(W_{n,j} = u \mid \mathcal{Y}_{n,j}) - P(W_{n,j} = u-1 \mid \mathcal{Y}_{n,j})\right| \\ &\quad = \frac{1}{2\pi}\left|\int_{-\pi}^{\pi}\exp\{-iux\}\big(1-\exp\{ix\}\big)\psi_{W_{n,j}|\mathcal{Y}_{n,j}}(x)\,dx\right| \\ &\quad \le \frac{1}{\pi}\int_0^{\pi} x\exp\left\{-2\sigma_j^2 x^2/\pi^2\right\}dx \le \pi/\big(4\sigma_j^2\big).\end{aligned}$$

Therefore the total variation distance between $\mathcal{L}(W_{n,j} \mid \mathcal{Y}_{n,j})$ and $\mathcal{L}(W_{n,j}+1 \mid \mathcal{Y}_{n,j})$ is bounded from above by

$$\begin{aligned}&\sum_{u\in\mathbb{Z}}\left|P(W_{n,j} = u \mid \mathcal{Y}_{n,j}) - P(W_{n,j} = u-1 \mid \mathcal{Y}_{n,j})\right| \\ &\quad \le \sum_{|u-\mu_j|\le\tau\sigma_j+1}\left|P(W_{n,j} = u \mid \mathcal{Y}_{n,j}) - P(W_{n,j} = u-1 \mid \mathcal{Y}_{n,j})\right| \\ &\qquad + 2P\big(|W_{n,j}-\mu_j| > \tau\sigma_j \mid \mathcal{Y}_{n,j}\big) \\ &\quad \le (2\tau\sigma_j+3)\pi^2/\big(2\sigma_j^2\big) + 4\exp\left\{-\tau^2/4\right\} + 4\exp\{-3\sigma_j\tau/4\},\end{aligned}$$

for any $\tau > 0$ where Bernstein's inequality has been used in the last step. We introduce the event $A_j := \{\sigma_j^2 > E\sigma_j^2/2\}$ where $E\sigma_j^2 \ge n\kappa$ and we put $\tau := \sqrt{c\cdot\log(n\kappa)}$ with a constant $c > 0$ sufficiently large so that

$$\begin{aligned}&E\sum_{u\in\mathbb{Z}}\left|P(W_{n,j} = u \mid \mathcal{Y}_{n,j}) - P(W_{n,j} = u-1 \mid \mathcal{Y}_{n,j})\right| \\ &\quad \le 2\big(1-P(A_j)\big) + A^*\left\{\log(n\kappa)\right\}^{1/2}n^{-1/2}\kappa^{-1/2},\end{aligned}$$

for a universal constant $A^* \in (0, \infty)$. Finally, Hoeffding's inequality yields that

$$1 - P(A_j) \le 2\exp\left\{-\big(E\sigma_j^2\big)^2/(2n)\right\} \le 2\exp\left\{-n\kappa^2/2\right\},$$

which completes the proof of the lemma. □

**Proof of Lemma 4.2.** As $W_n$ is $\mathbb{Z}^d$-valued it holds that $[W_n + U] = W_n + [U]$ so that

$$\begin{aligned}\mathsf{TV}\big(\mathcal{L}([W_n+U]), \mathcal{L}(W_n)\big) &\le \sum_{l\in\mathbb{Z}^d} \mathsf{TV}\big(\mathcal{L}(W_n+l), \mathcal{L}(W_n)\big)\cdot P\big([U]=l\big)\\ &\le A\big\{\log(n\kappa)\big\}^{1/2} n^{-1/2}\kappa^{-1/2}\sum_{j=1}^{d} E\big|[U_j]\big|,\end{aligned}$$

where $E|[U_j]| \le 1/2 + E|U_j| \le 1/2 + b_n^{1/2}$. □

**Proof of Lemma 4.3.** Let $\lambda$ be an arbitrary eigenvalue of the matrix $\sum_{i\in\mathcal{N}} \overline{\Lambda}_i$ with the corresponding unit eigenvector $v$. As $\overline{\Lambda}_i$ is the covariance matrix of $Y_{i,n}$ we deduce that

$$\begin{aligned}\lambda &= v^T \sum_{i\in\mathcal{N}} \overline{\Lambda}_i v = \sum_{i\in\mathcal{N}} v^T \overline{\Lambda}_i v = \sum_{i\in\mathcal{N}} \operatorname{var}\left(\sum_{k=1}^{d} v_k X_{k,i,n}\right)\\ &= \sum_{i\in\mathcal{N}} EE\left\{\left(\sum_{k=1}^{d} v_k (X_{k,i,n} - EX_{k,i,n})\right)^2 \mid Y_{i,n}^{-l}\right\}\\ &\ge \sum_{i\in\mathcal{N}} E\operatorname{var}\big(v_l X_{k,l,n} \mid Y_{i,n}^{-l}\big) \ge v_l^2 \cdot (\#\mathcal{N})\cdot\kappa,\end{aligned}$$

for all $l = 1, \dots, d$. Summing up both sides of the above inequality over $l = 1, \dots, d$ we obtain that

$$d\lambda \ge (\#\mathcal{N})\cdot\kappa,$$

which completes the proof. □

**Proof of Lemma 4.4.** Note that $\tilde{\Lambda} - \lambda I$ is a positive semi-definite matrix so that

$$\mathsf{N}(0, \tilde{\Lambda}) = \mathsf{N}(0, \lambda I) * \mathsf{N}(0, \tilde{\Lambda} - \lambda I),$$

from what follows that

$$\mathsf{TV}\big(\mathcal{L}(Y_{i,n}) * \mathsf{N}(0, \tilde{\Lambda}), \mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i + \tilde{\Lambda})\big) \le \mathsf{TV}\big(\mathcal{L}(Y_{i,n}) * \mathsf{N}(0, \lambda I), \mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i + \lambda I)\big).$$

The distribution $\mathcal{L}(Y_{i,n}) * \mathsf{N}(0, \lambda I)$ has the $d$-dimensional Lebesgue density

$$g_0(x) = E(2\pi\lambda)^{-d/2}\exp\big\{-|x - Y_{i,n}|^2/(2\lambda)\big\}.$$

Since $\mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i + \lambda I) = \mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i) * \mathsf{N}(0, \lambda I)$ the Lebesgue density $g_1$ of the distribution $\mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i + \lambda I)$ may be written as

$$g_1(x) = E(2\pi\lambda)^{-d/2}\exp\big\{-|x - Z_{i,n}|^2/(2\lambda)\big\},$$

where $\mathcal{L}(Z_{i,n}) = \mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i)$. The total variation distance between $\mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i + \lambda I)$ and $\mathcal{L}(Y_{i,n}) * \mathsf{N}(0, \lambda I)$ equals the $L_1(\mathbb{R}^d)$-distance between the densities $g_0$ and $g_1$. Thus,

$$\begin{aligned}
&\mathsf{TV}\big(\mathcal{L}(Y_{i,n}) * \mathsf{N}(0, \tilde{\Lambda}), \mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i + \lambda I)\big) \\
&\quad = (2\pi)^{-d/2} \int \big| E \exp\big\{-\big|x - \lambda^{-1/2} Y_{i,n}\big|^2/2\big\} - E \exp\big\{-\big|x - \lambda^{-1/2} Z_{i,n}\big|^2/2\big\}\big| \, dx.
\end{aligned}$$

Taylor expansion around $x$ yields that

$$\exp\big(-|x - \Delta|^2/2\big) = P_{2,x}(\Delta) + R_{2,x}(\Delta),$$

for all $\Delta \in \mathbb{R}^d$ and any fixed $x \in \mathbb{R}^d$ where $P_{2,x}$ is a $d$-variate quadratic polynomial and $R_{2,x}$ is the corresponding remainder term. As the expectation vectors and the covariance matrices of $Y_{i,n}$ and $Z_{i,n}$ coincide we deduce that

$$E P_{2,x}\big(\lambda^{-1/2} Y_{i,n}\big) = E P_{2,x}\big(\lambda^{-1/2} Z_{i,n}\big).$$

Therefore,

$$\begin{aligned}
&\mathsf{TV}\big(\mathcal{L}(Y_{i,n}) * \mathsf{N}(0, \lambda I), \mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i + \lambda I)\big) \\
&\quad \le (2\pi)^{-d/2} \bigg( \int E\big|R_{2,x}\big(\lambda^{-1/2} Y_{i,n}\big)\big| \, dx + \int E\big|R_{2,x}\big(\lambda^{-1/2} Z_{i,n}\big)\big| \, dx \bigg).
\end{aligned}$$

Calculating the third order partial derivatives of $x \mapsto \exp(-|x|^2/2)$ we deduce that

$$\begin{aligned}
(2\pi)^{-d/2} E \int |R_{2,x}\big(\lambda^{-1/2} Y_{i,n}\big)| \, dx &\le \lambda^{-3/2} E \bigg( \sum_{k=1}^{d} |X_{k,i,n}| \bigg)^3 \\
&\le B^* \cdot \lambda^{-3/2} \cdot d^3,
\end{aligned} \tag{8.2}$$

for some universal constant $B^* \in (0, \infty)$ where $\mathcal{L}(Z) = \mathsf{N}(0, 1)$. When replacing $X_{k,i,n}$ by the $k$th component of $Z_{i,n}$ the identical upper bound applies (with a different constant $B^*$). Note that this component is $\mathsf{N}(E X_{k,i,n}, \operatorname{var} X_{k,i,n})$-distributed. □

**Proof of Lemma 4.5.** Again the shift-invariance of the total variation distance allows us to restrict to the case of $y_0 = 0$. As a telescoping sum, we consider

$$\begin{aligned}
&\mathsf{TV}\big(\mathcal{L}(W_n) * \mathsf{N}(0, b_n I), \mathsf{N}(\overline{\mu}, \overline{\Lambda} + b_n I)\big) \\
&\quad = \mathsf{TV}\big(\mathsf{N}(0, b_n I) * \big\{ *_{i=1}^n \mathcal{L}(Y_{i,n}) \big\}, \mathsf{N}(0, b_n I) * \big\{ *_{i=1}^n \mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i) \big\}\big) \\
&\quad \le \sum_{k=1}^{n} \mathsf{TV}\big(\big\{ *_{i=k+1}^n \mathcal{L}(Y_{i,n}) \big\} * \mathsf{N}(0, b_n I) * \big\{ *_{i=1}^k \mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i) \big\}, \\
&\qquad \big\{ *_{i=k}^n \mathcal{L}(Y_{i,n}) \big\} * \mathsf{N}(0, b_n I) * \big\{ *_{i=1}^{k-1} \mathsf{N}(\overline{\mu}_i, \overline{\Lambda}_i) \big\}\big)
\end{aligned}$$

$$\leq \sum_{k=1}^{n} \mathsf{TV}\big(\mathsf{N}(\overline{\mu}_k, \overline{\Lambda}_k + \tilde{\Lambda}_{k-1}), \mathcal{L}(Y_{k,n}) * \mathsf{N}(0, \tilde{\Lambda}_{k-1})\big),$$

where $\tilde{\Lambda}_l := b_n I + \sum_{i=1}^{l} \overline{\Lambda}_i$. By Lemma 4.3 the smallest eigenvalue of $\tilde{\Lambda}_l$ is bounded from below by $b_n + l\kappa/d_n$. Then Lemma 4.4 provides that

$$\begin{aligned}
&\mathsf{TV}\big(\mathcal{L}(W_n) * \mathsf{N}(0, b_n I), \mathsf{N}(\overline{\mu}, \overline{\Lambda} + b_n I)\big) \\
&\quad \leq B d_n^3 \sum_{k=1}^{n} \big(b_n + (k-1)\kappa/d\big)^{-3/2} \\
&\quad \leq B d_n^3 \left( b_n^{-3/2} + \int_0^\infty (b_n + x\kappa/d_n)^{-3/2}\, dx \right) = B d_n^3 b_n^{-3/2} + 2 B d_n^4 b_n^{-1/2}/\kappa.
\end{aligned}$$

The lemma has been shown. □

**Proof of Lemma 4.6.** The total variation distance between two distributions is bounded from above by twice their Hellinger distance. It follows from for example, eq. (A.4) in Reiß [30] that

$$\mathsf{TV}^2\big(\mathsf{N}(\overline{\mu}, \overline{\Lambda} + b_n I), \mathsf{N}(\overline{\mu}, \overline{\Lambda})\big) \leq 8 b_n^2 \big\| \overline{\Lambda}^{-1} \big\|_F^2 = 8 b_n^2 \sum_{j=1}^{d_n} \lambda_j^{-2},$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\lambda_j$, $j = 1, \ldots, d_n$, are the eigenvalues of the matrix $\overline{\Lambda}$. Applying the lower bound on the eigenvalues provided in Lemma 4.3 completes the proof of this lemma. □

**Proof of Lemma 5.1.** In the notation of Section 4, we assume some random vector $Y_{i,n} = (X_{1,i,n}, \ldots, X_{d,i,n})$ with $d = m - 1$ and $\mathcal{L}(Y_{i,n}) = \mathsf{U}_{\vartheta,F}(\cdot \mid k)$. Then

$$p_{l,i} = E\big(X_{l,i,n} \mid Y_{i,n}^{-l}\big) = \frac{\exp(-\vartheta_l)}{1 + \exp(-\vartheta_l)} \cdot 1_{\{k-1\}} \left( \sum_{q \neq l}^{m-1} X_{q,i,n} \right) + 1_{\{k-2\}} \left( \sum_{q \neq l}^{m-1} X_{q,i,n} \right),$$

so that

$$\begin{aligned}
E p_{l,i}(1 - p_{l,i}) &= \frac{\exp(-\vartheta_l)}{(1 + \exp(-\vartheta_l))^2} \cdot \mathsf{U}_{\vartheta,F} \left( \sum_{q \neq l}^{m-1} X_{q,i,n} = k - 1 \mid k \right) \\
&= \frac{\exp(-\vartheta_l)}{1 + \exp(-\vartheta_l)} \cdot \sum_{b \in \mathbb{B}'(l,k,m)} \exp\left( - \sum_{i \neq l}^{m-1} \vartheta_i b_i \right) \Big/ \sum_{c \in \mathbb{B}(k,m)} \exp\left( - \sum_{i=1}^{m-1} \vartheta_i c_i \right),
\end{aligned}$$

where $\mathbb{B}'(l,k,m)$ collects all $b \in \mathbb{B}(k,m)$ such that $\sum_{i\neq l}^{m-1} b_i = k-1$. Note that

$$\mathbb{B}(k,m) = \bigcup_{l=1}^{m-1} \mathbb{B}'(l,k,m),$$

as $k \in \{1,\dots,m-1\}$. Also we have

$$\sum_{b\in\mathbb{B}'(l,k,m)} \exp\left(-\sum_{i\neq l}^{m-1} \vartheta_i b_i\right) \le \exp(4R) \sum_{b\in\mathbb{B}'(l',k,m)} \exp\left(-\sum_{i\neq l'}^{m-1} \vartheta_i b_i\right),$$

for all $l, l' \in \{1,\dots,m-1\}$. It follows that

$$Ep_{l,i}(1-p_{l,i}) \ge \frac{\exp(-6R)}{(m-1)(1+\exp(2R))},$$

which completes the proof. □

**Proof of Lemma 5.2.** Note that $N_0 + N_m$ has a binomial distribution with the parameters $n$ and $q_0(\theta,F) + q_m(\theta,F)$ where $q_k(\theta,F)$ is defined in (3.6). For any $s > 0$, we have

$$\begin{aligned} q_0(\theta,F) &\le \int_{x<-s} \overline{f}(x)\,dx + \big(1+\exp(-s-R)\big)^{-m}, \\ q_1(\theta,F) &\le \int_{x>s} \overline{f}(x)\,dx + \big(1+\exp(-s-R)\big)^{-m}, \end{aligned} \tag{8.3}$$

for all $\theta \in \Theta$. For any fixed $\varepsilon > 0$, we choose $s$ sufficiently large such that the first summands in both lines of (8.3) are bounded from above by $\varepsilon/2$ and, then, $M$ sufficiently large such that for all $m > M$ the second summands in (8.3) are smaller than $\varepsilon/2$. Thus, for all $m > M$, we obtain that $q_0(\theta,F) + q_m(\theta,F) < \varepsilon$. On the other hand, if $m \le M$, we fix $s' > 0$ sufficiently large such that

$$\int_{|x|\ge s'} \overline{f}(x)\,dx < 1/2,$$

so that $\int_{|x|\le s'} f(x)\,dx \ge 1/2$ holds true for all $F \in \mathcal{F}$. Then we consider the continuous positive mapping $T_m$, $m = 2,\dots,M$, with

$$T_m(x,\theta) := \sum_{b\in\mathbb{S}(1,m)} \prod_{l=1}^{m} \frac{\exp(b_l[x-\theta_l])}{1+\exp(x-\theta_l)},$$

which take its positive minimum on the compact domain $[-s',s'] \times [-R,R]^m$. Hence,

$$\inf_{\theta\in\Theta, F\in\mathcal{F}} q_1(\theta,F) > 0, \qquad \forall m = 2,\dots,M,$$

so that $\sup_{\theta\in\Theta, F\in\mathcal{F}} q_0(\theta, F) + q_m(\theta, F) < 1$. Thus, we have shown that

$$q := \sup_{\theta\in\Theta, F\in\mathcal{F}} \sup_{m\geq 3} q_0(\theta, F) + q_m(\theta, F) < 1.$$

Now we choose $\rho := (1+q)/2 \in (0,1)$ so that simple application of Chebyshev's inequality completes the proof. □

**Proof of Theorem 5.2.** Fix some $\rho \in (0,1)$ from Lemma 5.2. Thus, the probability of $\{N_0 + N_m > \rho n\}$ converges to zero uniformly with respect to $\theta$ and $F$. Therefore, it suffices to show that the mean total variation distance between $\mathcal{L}^D_{\theta,F}(\Phi \mid N)$ and $P^C_{\theta,F}(\cdot \mid N)$, restricted to the event $\mathcal{N} := \{N_0 + N_m \leq \rho n\}$, tends to zero uniformly in $\theta$ and $F$ as well. The first (conditional) probability measure has the Lebesgue density

$$\begin{aligned} h_\theta(x \mid N) &= (2\pi)^{1/2-1/m}\big(\det \Delta\Psi_N(\vartheta)\big)^{1/2}\big(\det \Delta\Psi_N(x)\big)^{-1} \\ &\quad \cdot \exp\big\{-\big((\nabla\Psi_N)^{-1}(-x) - \vartheta\big)^T \Delta\Psi_N(\vartheta)\big((\nabla\Psi_N)^{-1}(-x) - \vartheta\big)/2\big\}, \end{aligned}$$

on the range $\mathcal{R}$ of $-\nabla\Psi_N$, on which $h_\theta(\cdot \mid N)$ is supported and on which the function $\nabla\Psi_N$ has an inverse mapping. We write $g_\theta(\cdot \mid N)$ for the density of $P^C_{\theta,F}(\cdot \mid N) = \mathsf{N}(-\nabla\Psi_N(\vartheta), \Delta\Psi_N(\vartheta))$. Moreover, note that

$$E1_{\mathcal{N}} \int \big|h_\theta(x \mid N) - g_\theta(x \mid N)\big|\, dx \leq 2E1_{\mathcal{N}} \int_{\mathcal{R}} \big|h_\theta(x \mid N) - g_\theta(x \mid N)\big|\, dx.$$

Applying the integral substitution via $-\nabla\Psi_N$ the right-hand side of the above inequality equals $2E1_{\mathcal{N}}|Y-1|$ where

$$\begin{aligned} Y := \frac{\det \Delta\Psi_N(X)}{\det \Delta\Psi_N(\vartheta)} \exp\bigg\{ -\frac{1}{2}\big(\big(\nabla\Psi_N(X) - \nabla\Psi_N(\vartheta)\big)^T \big\{\Delta\Psi_N(\vartheta)\big\}^{-1} \\ \times \big(\nabla\Psi_N(X) - \nabla\Psi_N(\vartheta)\big) - (X-\vartheta)^T \Delta\Psi_N(\vartheta)(X-\vartheta)\big)\bigg\}, \end{aligned}$$

where $\mathcal{L}(X \mid N) = \mathsf{N}(\vartheta, \{\Delta\Psi_N(\vartheta)\}^{-1})$. All third-order partial derivatives of $\Psi_N$ are bounded by $6n$ so that

$$\nabla\Psi_N(X) - \nabla\Psi_N(\vartheta) = \Delta\Psi_N(\vartheta)(X-\vartheta) + R_1,$$

where the remainder term $R_1$ satisfies $|R_1| \leq 6nm^{3/2}|X-\vartheta|^2$. The matrix-valued function $\Delta\Psi_N$ has the following Lipschitz property (with respect to the Frobenius norm),

$$\big\|\Delta\Psi_N(X) - \Delta\Psi_N(\vartheta)\big\|_F \leq 6nm^{3/2}|X-\vartheta|.$$

The Theorem of Courant–Fischer yields that

$$\sup_{j=1,\dots,m-1} \big|\lambda_j(X) - \lambda_j(\vartheta)\big| \leq 6nm^{3/2}|X-\vartheta|,$$

where $\lambda_j(X)$ and $\lambda_j(\vartheta)$ denote the eigenvalues of the matrices $\Delta\Psi_N(X)$ and $\Delta\Psi_N(\vartheta)$, respectively, in decreasing order. We learn from the Lemmata 4.3 and 5.1 that

$$\inf_n \big(m_n^2/n\big) \cdot \inf_\theta \inf_j \lambda_j(\vartheta) > 0,$$

for $m = m_n$. Thus, on the event $\mathcal{C} := \mathcal{N} \cap \{|X - \vartheta| \le \alpha_n m_n^{-9/2}\}$, for any sequence $(\alpha_n) \downarrow 0$, we deduce that

$$\left|1 - \frac{\det \Delta\Psi_N(X)}{\det \Delta\Psi_N(\vartheta)}\right| = \left|1 - \prod_{j=1}^{m_n-1}\left(1 + \frac{\lambda_j(X) - \lambda_j(\vartheta)}{\lambda_j(\vartheta)}\right)\right| \le \text{const.} \cdot \alpha_n,$$

where, in the sequel, const. stands for a constant only depending on $\rho$ and $R$. Furthermore,

$$\begin{aligned}&\big(\nabla\Psi_N(X) - \nabla\Psi_N(\vartheta)\big)^T \big\{\Delta\Psi_N(\vartheta)\big\}^{-1}\big(\nabla\Psi_N(X) - \nabla\Psi_N(\vartheta)\big) - (X-\vartheta)^T \Delta\Psi_N(\vartheta)(X-\vartheta)\\ &\quad = 2R_1^T(X-\vartheta) + R_1^T\big\{\Delta\Psi_N(\vartheta)\big\}^{-1} R_1 \le \text{const.} \cdot \big(n\alpha_n^3 m_n^{-12} + n\alpha_n^4 m_n^{-13}\big),\end{aligned}$$

holds true on the event $\mathcal{C}$. Any selection of $(\alpha_n)_n$ such that

$$\lim_{n\to\infty} \alpha_n m_n^{-4} n^{1/3} = 0,$$

guarantees uniform convergence of $E1_{\mathcal{C}}|Y - 1|$ to zero. On the other hand the probability of $\mathcal{N}\backslash\{|X - \vartheta| \le \alpha_n m_n^{-9/2}\}$ also tends to zero uniformly with respect to $\theta$ and $F$ if

$$\alpha_n n^{1/2} m_n^6 \to \infty,$$

as $n \to \infty$ since $\mathcal{L}(X - \vartheta) = \mathsf{N}(0, \{\Delta\Psi_N(\vartheta)\}^{-1})$. As $\sup_n m_n^\beta n < \infty$ for some $\beta > 13$ such a choice of $(\alpha_n)_n$ exists. Then,

$$\lim_{n\to\infty} \sup_{\theta,F} E1_{\mathcal{N}}|Y \cdot 1_{\mathcal{C}} - 1| = 0.$$

As $Y$ is non-negative, $E1_{\mathcal{N}}Y \le 1$ and $\lim_{n\to\infty}\sup_{\theta,F}(1 - P_{\theta,F}^{\mathcal{C}}(\mathcal{N})) = 0$ we arrive at

$$\lim_{n\to\infty} \sup_{\theta,F} E1_{\mathcal{N}}|Y - 1| = 0,$$

which completes the proof. □

**Proof of Lemma 5.3.** Setting

$$\eta_k(\beta) := \exp\{\beta - \theta_k\}/\big(1 + \exp\{\beta - \theta_k\}\big), \qquad k = 1, \dots, m,$$

we may write

$$q_k(\theta, F) = \int \big\{*_{k=1}^m \mathsf{B}\big(1, \eta_k(\beta)\big)\big\}\, dF(\beta).$$

As $\theta \in [-R, R]^m$ we have that

$$\eta_k(\beta) \le \exp(2R) \cdot \eta_1(\beta),$$
$$1 - \eta_k(\beta) \le \exp(2R) \cdot \big(1 - \eta_1(\beta)\big),$$

for all $\beta \in \mathbb{R}$ and $k = 2, \dots, m$.

On the sets $A_1 := \{\beta : \eta_1(\beta) \le cm^{-3/4}\}$ and $A_2 := \{\beta : 1 - \eta_1(\beta) \le cm^{-3/4}\}$ for some constant $c > 0$, we apply Poisson approximation of binomial distributions. Precisely, an inequality of Le Cam (see p. 657 in DasGupta [12], for instance) yields that

$$\mathsf{TV}\big(\mathsf{P}\big(Q(\beta)\big), *_{k=1}^m \mathsf{B}\big(1, \eta_k(\beta)\big)\big) \le 2\exp(4R)c^2 m^{-1/2}, \qquad \forall \beta \in A_1,$$

where $Q(\beta) := \sum_{k=1}^m \eta_k(\beta)$. Put $I_k(\delta) := Q^{-1}([k - \delta, k + \delta])$ for $k = 1, \dots, m - 1$; $I_0(\delta) := Q^{-1}([0, \delta])$ and $I_m(\delta) := Q^{-1}([m - \delta, m - \delta/2])$ for some fixed $\delta \in (0, 1)$. By Stirling's approximation,

$$\begin{aligned} b(m, k, \beta) &:= \big\{*_{l=1}^m \mathsf{B}\big(1, \eta_l(\beta)\big)\big\}(k) \ge \exp\big\{-Q(\beta)\big\} Q^k(\beta)/k! - 2\exp(4R) \cdot c^2 m^{-1/2} \\ &\ge \exp\big\{-1/(12k) - \delta\big\}(1 - \delta/k)^k/\sqrt{2\pi k} - 2\exp(4R) \cdot c^2 m^{-1/2} \\ &\ge \text{const.} \cdot m^{-1/2}, \end{aligned}$$

for all $\beta \in A_1 \cap I_k(\delta)$, $k \ge 1$, and a constant factor only depending on $R$, when choosing the constant $c > 0$ sufficiently small. For $k = 0$, this bound applies as well.

For $\beta \in A_2 \cap I_k(\delta)$ the identical lower bound applies since $q_k(\theta, F)$ is viewed as the density of $*_{l=1}^m \mathsf{B}(1, 1 - \eta_l(\beta))$ at $m - k$.

Finally, we consider the complement $A_3 := \mathbb{R} \backslash (A_1 \cup A_2)$. Clearly $\eta_1(\beta) \in (cm^{-3/4}, 1 - cm^{-3/4})$ holds for all $\beta \in A_3$. By Fourier inversion,

$$b(m, k, \beta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp\big\{-it\big(k - Q(x)\big)\big\} f_m(t, \beta)\, dt,$$

with

$$f_m(t, \beta) := \prod_{l=1}^m \exp\big\{it\big(1 - \eta_l(\beta)\big)\big\} \cdot \eta_l(\beta) + \exp\big\{-it\eta_l(\beta))\big\} \cdot \big(1 - \eta_l(\beta)\big),$$
$$\eta(\beta) := \sum_{l=1}^m \eta_l(\beta)\big(1 - \eta_l(\beta)\big).$$

As in the proof of Lemma 4.1, we derive that

$$\big|f_m(t, \beta)\big| \le \exp\left\{-\frac{2}{\pi^2}\eta(\beta)\right\},$$

for all $t \in [-\pi, \pi]$. Moreover, for all $\beta \in A_3$, we have

$$\eta(\beta) \geq c \exp\{-2R\} m^{1/4} \big(1 - c \exp\{-2R\} m^{-3/4}\big).$$

Put $\nu := \{D(\log \eta(\beta))/\eta(\beta)\}^{1/2}$ for some constant $D > 0$ sufficiently large. Then $|f_m(t, \beta)| \leq m^{-D/(2\pi^2)}$ if $|t| \in (\nu, \pi]$. Otherwise, for $t \in [-\nu, \nu]$, Taylor approximation yields that

$$f_m(t, \beta) = \exp\big\{-t^2 \eta(\beta)/2\big\} \cdot \big(1 + \Delta_m(t, \beta)\big),$$

with the remainder $|\Delta_m(t, \beta)| \leq \text{const.} \cdot \nu$ for some universal constant factor. Combining these facts with

$$\big|1 - \exp\big\{-it\big(k - Q(\beta)\big)\big\}\big| \leq |t|\delta,$$

for all $\beta \in I_k(\delta)$, we deduce that

$$b(m, k, \beta) \geq \text{const.} \cdot m^{-1/2}, \tag{8.4}$$

for all $\beta \in I_k(\delta) \cap A_3$ and some universal positive constant. Summarising, the inequality (8.4) has been verified for all $\beta \in I_k(\delta)$ where the constant factor only depends on $R$ and $\delta$.

We conclude that

$$q_k(\theta, F) \geq \text{const.} \cdot m^{-1/2} \cdot \int_{I_k(\delta)} f(x)\, dx,$$

for all $k = 0, \ldots, m$ where the constant does not depend on $k$. As the derivative of $Q$ is bounded by $m$ the length of the interval $I_k(\delta)$ has the lower bound $\delta/(2m)$. Moreover,

$$\sup\big\{|x| : x \in I_k(\delta)\big\} \leq R + \big|\log(\delta/2)\big| + \log m,$$

for any $\delta \in (0, 1/2)$ and all $k = 0, \ldots, m$. Finally, by the tail condition (5.4) on $f$, the proof is completed. □

**Proof of Lemma 5.4.** The total variation distance between $\mathsf{N}(n\tilde{q}(\theta, F), n\tilde{Q}(\theta, F) - n\tilde{q}(\theta, F) \times \tilde{q}(\theta, F)^T)$ and $\mathcal{L}(\tau(N^{**}) \mid V)$ is bounded from above by

$$\begin{aligned}
&1_{(-\infty,\zeta)}(V) + 1_{[\zeta,\infty)}(V) \\
&\quad \cdot \mathsf{TV}\big(\mathsf{N}\big(n\tilde{q}(\theta, F), n^3\big(\tilde{Q}(\theta, F) - \tilde{q}(\theta, F)\tilde{q}(\theta, F)^T\big)/V^2\big), \\
&\quad \mathsf{N}\big(n\tilde{q}(\theta, F), n\big(\tilde{Q}(\theta, F) - \tilde{q}(\theta, F)\tilde{q}(\theta, F)^T\big)\big)\big) \\
&\leq 1_{(-\infty,\zeta)}(V) + 1_{[\zeta,\infty)}(V) \cdot \mathsf{H}\big(\mathsf{N}\big(0, \big(n^2/V^2\big)I\big), N(0, I)\big) \\
&\leq 1_{(-\infty,\zeta)}(V) + \sqrt{2(m+1)} \cdot 1_{[\zeta,\infty)}(V) \cdot \big|n^2/V^2 - 1\big|,
\end{aligned}$$

where $\mathsf{H}$ denotes the Hellinger distance and $I$ the $(m+1) \times (m+1)$-identity matrix. Therein equation (A.4) in Reiß [30] has been used to bound the Hellinger distance between normal distributions. Applying the expectation to the above term, we obtain

$$P^F_{\theta,F}(V < \zeta) + \sqrt{2(m+1)}\zeta^{-2} E\big|n^2 - V^2\big| \leq 4/n + 4\sqrt{2(m+1)}\big(2n^{3/2} + n\big)/n^2,$$

as $\mathcal{L}^F_{\theta,F}(V) \sim N(n,n)$ and $\zeta = n/2$. Thanks to the conditions on $m_n$ in Theorem 5.3, the above expression tends to zero uniformly in $\theta \in \Theta$ and $F \in \mathcal{F}$. □

**Proof of Theorem 6.1.** We consider that

$$\begin{aligned} P^D_{\theta,F}(\theta \notin \hat{E} \mid N) &\le 1_{[0,n)}(N_0+N_m) P^D_{\theta,F}\big(\varepsilon^T \Delta\Psi_N(\vartheta)^{-1/2} \Delta\Psi_N\big(T^{**}\big)\Delta\Psi_N(\vartheta)^{-1/2}\varepsilon > \iota \mid N\big) \\ &\quad + P^D_{\theta,F}(N_0+N_m = n) \\ &\le 1_{[0,n)}(N_0+N_m) P^D_{\theta,F}\big(|\varepsilon|^2 \cdot \big\{1 + 6nm^{3/2}\big\|\big\{\Delta\Psi_N(\vartheta)\big\}^{-1/2}\big\|^3 |\varepsilon|\big\} > \iota \mid N\big) \\ &\quad + P^D_{\theta,F}(N_0+N_m = n), \end{aligned}$$

where $\|\cdot\|$ denotes the usual matrix norm which is induced by the Euclidean metric; we have used that

$$\sup_{x\in\mathbb{R}^{m-1}} \sup_{l,j,j'} \left| \frac{\partial^3}{\partial x_l \partial x_j \partial x_{j'}} \Psi_N(x) \right| \le 6n.$$

By the Lemmata 4.3 and 5.1, we deduce that

$$\begin{aligned} P^D_{\theta,F}(\theta \notin \hat{E}) &\le 2P^D_{\theta,F}(N_0+N_m > \rho n) + P\big(|\varepsilon| > h_n n^{1/2} m^{-9/2}\big) \\ &\quad + P\big(|\varepsilon|^2 \cdot \big(1 + 6c(1-\rho)^{-3/2} h_n\big) > \iota\big), \end{aligned}$$

for some sequence $(h_n)_n$ such that $(h_n m_n)_n \downarrow 0$ (with $m = m_n$) but $(h_n^2 n m_n^{-10})_n \uparrow \infty$, some constant $c > 0$ only depending on $R$ and some fixed $\rho \in (0,1)$ from Lemma 5.2. Taking the supremum over $\theta \in \Theta$ and $F \in \mathcal{F}$ and, then, the limit superior, as $n \to \infty$, on both sides of the above inequality we conclude that

$$\limsup_{n\to\infty} \sup_{\theta,F} P^D_{\theta,F}(\theta \notin \hat{E}) \le \limsup_{n\to\infty} P\big(|\varepsilon|^2 > \iota\big) = \alpha,$$

where we have used Lemma 5.2 and the fact that the $\chi^2(m-1)$-density, as a consecutive sequence of convolutions, is bounded uniformly with respect to $m$ where $\iota \asymp m$ (since $\alpha \in (0,1)$ is fixed). Moreover the choice of $\iota$ is crucial in the last step.

The maximal length of an axis $\hat{e}$ of $\hat{E}$ turns out to be $2/\sqrt{\lambda_{\min}}$ where $\lambda_{\min}$ is the smallest positive eigenvalue of the matrix $Z(Z^\dagger Z)^{-1}\Delta\Psi_N(T^{**})(Z^\dagger Z)^{-1}Z^\dagger$. For all $x \in \mathbb{R}^m$ with $\sum_{j=1}^m x_j = 0$, we have $|(Z^\dagger Z)^{-1} Z^\dagger x| \ge |x|$. Therefore, $\lambda_{\min}$ is bounded from below by the smallest eigenvalue $\lambda'_{\min}$ of the matrix $\Delta\Psi_N(T^{**})$. Then it follows from the Lemmata 4.3 and 5.1 that

$$\lambda'_{\min} \ge \text{const.} \cdot \big\{(1-\rho)n/m_n^2 - (1-\rho)^{-1/2}|\varepsilon| \cdot m_n/n^{1/2}\big\},$$

holds on the event $\{N_0 + N_m \le \rho n\}$. Using Lemma 5.2, we establish that

$$1/\lambda_{\min} = \mathcal{O}_P\big(m_n^2/n\big),$$

uniformly with respect to $\theta$ and $F$. □

## Acknowledgments

The authors are grateful to the editors and a referee for their inspiring and helpful comments.

## References

[1] Alagumalai, S., Curtis, D.D. and Hungi, N. (2005). *Applied Rasch Measurement*: *A Book of Exemplars*. Berlin: Springer.
[2] Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika* **42** 69–81. MR0483255
[3] Andersen, E.B. (1980). Comparing latent distributions. *Psychometrika* **45** 121–134. MR0570773
[4] Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika* **75** 292–308. MR2719929
[5] Bezruczko, N. (2005). *Rasch Measurement in Health Sciences*. Maple Grove, MN: JAM Press.
[6] Biehler, M., Holling, H. and Doebler, P. (2015). Saddlepoint approximations of the distribution of the person parameter in the two parameter logistic model. *Psychometrika* **80** 665–688. MR3392024
[7] Brown, L.D., Carter, A.V., Low, M.G. and Zhang, C.-H. (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.* **32** 2074–2097. MR2102503
[8] Brown, L.D. and Low, M.G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958
[9] Cai, T.T. and Zhou, H.H. (2009). Asymptotic equivalence and adaptive estimation for robust nonparametric regression. *Ann. Statist.* **37** 3204–3235. MR2549558
[10] Carter, A.V. (2002). Deficiency distance between multinomial and multivariate normal experiments. *Ann. Statist.* **30** 708–730. MR1922539
[11] Carter, A.V. (2006). A continuous Gaussian approximation to a nonparametric regression in two dimensions. *Bernoulli* **12** 143–156. MR2202326
[12] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. *Springer Texts in Statistics*. New York: Springer. MR2664452
[13] de Leeuw, J. and Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *J. Educ. Behav. Stat.* **11** 183–196.
[14] Doebler, A., Doebler, P. and Holling, H. (2012). Optimal and most exact confidence intervals for person parameters in item response theory models. *Psychometrika* **77** 98–115. MR3042821
[15] Fischer, G.H. and Molenaar, I.W., eds. (1995). *Rasch Models*. New York: Springer. MR1367341
[16] Genon-Catalot, V. and Larédo, C. (2014). Asymptotic equivalence of nonparametric diffusion and Euler scheme experiments. *Ann. Statist.* **42** 1145–1165. MR3224284
[17] Haberman, S.J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5** 815–841. MR0501540
[18] Hoderlein, S., Mammen, E. and Yu, K. (2011). Non-parametric models in binary choice fixed effects panel data. *Econom. J.* **14** 351–367. MR2853231
[19] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. *Springer Series in Statistics*. New York: Springer. MR0856411
[20] Le Cam, L. and Yang, G.L. (2000). *Asymptotics in Statistics*: *Some Basic Concepts*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR1784901
[21] Liese, F. and Miescke, K.-J. (2008). *Statistical Decision Theory*. *Springer Series in Statistics*. New York: Springer. MR2421720
[22] Lindsay, B., Clogg, C.C. and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Amer. Statist. Assoc.* **86** 96–107. MR1137102

[23] Mariucci, E. (2016). Asymptotic equivalence for pure jump Lévy processes with unknown Lévy density and Gaussian white noise. *Stochastic Process. Appl.* **126** 503–541. MR3434992

[24] Meister, A. (2011). Asymptotic equivalence of functional linear regression and a white noise inverse problem. *Ann. Statist.* **39** 1471–1495. MR2850209

[25] Meister, A. and Reiß, M. (2013). Asymptotic equivalence for nonparametric regression with nonregular errors. *Probab. Theory Related Fields* **155** 201–229. MR3010397

[26] Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430. MR1425959

[27] Pfanzagl, J. (1993). On the consistency of conditional maximum likelihood estimators. *Ann. Inst. Statist. Math.* **45** 703–719. MR1252949

[28] Pfanzagl, J. (1994). On the identifiability of structural parameters in mixtures: Applications to psychological tests. *J. Statist. Plann. Inference* **38** 309–326. MR1261806

[29] Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests* (*Expanded Edition*). Chicago: Univ. Chicago Press.

[30] Reiß, M. (2011). Asymptotic equivalence for inference on the volatility from noisy observations. *Ann. Statist.* **39** 772–802. MR2816338

[31] Rice, K.M. (2004). Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *J. Amer. Statist. Assoc.* **99** 510–522. MR2062836

[32] Rohde, A. (2004). On the asymptotic equivalence and rate of convergence of nonparametric regression and Gaussian white noise. *Statist. Decisions* **22** 235–243. MR2125610

[33] Schmidt-Hieber, J. (2014). Asymptotic equivalence for regression under fractional noise. *Ann. Statist.* **42** 2557–2585. MR3277671

[34] Shiryaev, A.N. and Spokoiny, V.G. (2000). *Statistical Experiments and Decisions*: *Asymptotic theory*. *Advanced Series on Statistical Science & Applied Probability* **8**. River Edge, NJ: World Scientific. MR1791434

[35] Strasser, H. (1985). *Mathematical Theory of Statistics*: *Statistical Experiments and Asymptotic Decision Theory*. *De Gruyter Studies in Mathematics* **7**. Berlin: de Gruyter. MR0812467

[36] Strasser, H. (2012a). The covariance structure of cml-estimates in the Rasch model. *Stat. Risk Model.* **29** 315–326. MR2997038

[37] Strasser, H. (2012b). Asymptotic expansions for conditional moments of Bernoulli trials. *Stat. Risk Model.* **29** 327–343. MR2997039

[38] von Davier, M. and Carstensen, C.H. (2007). *Multivariate and Mixture Distribution Rasch Models – Extensions and Applications*. *Statistics for Social and Behavioral Sciences*. New York: Springer.

*Received February 2017 and revised August 2017*