Quaderni di Statistica
Vol. 11, 2009

On kernel smoothing in polytomous IRT: a new minimum distance estimator

Antonio Punzo
Dipartimento di Economia e Metodi Quantitativi, Università di Catania
E-mail: antonio.punzo@unict.it

Summary: In Item Response Theory (IRT), the Item Category Response Function (ICRF), defining the relation between ability and probability of choosing a particular option for a test item, and the Item Response Function (IRF), describing the relation between ability and probability of obtaining a particular score for an item, are both of crucial importance. In analogy with the standard statistical methodology, these functions may be estimated by using both parametric and nonparametric approaches. Here, the performance of the well-known nonparametric kernel estimator is investigated in the polytomous case giving a description of the cross-validation approach to estimate the smoothing parameter, and providing pointwise confidence intervals for IRFs. Moreover, based on the consistency of this approach, a kernel-based minimum distance estimator of parametric IRT functions is proposed and evaluated by a Monte Carlo simulation study.

Keywords: Item Response Theory, Kernel Smoothing.

1. Introduction

In psychometrics and educational testing the analysis of the relation between latent continuous variables and observed categorical variables – which can be dichotomous or (nominal/ordinal) polytomous – is known as Item Response Theory (IRT). In applications it is very common to have data that are ordinal polytomous, above all with 3 or 4 categories (*e.g.*, in aptitude testing, the response is often classified in one of the following ordinal categories: “wrong”, “partially correct”, “fully correct”). Masters (1988) and Bejar (1977) note that the purpose of using more than two categories per item is to try to obtain more information about the trait level ϑ, generically referred to as “ability”, of the people being measured. Conversely, Cohen (1983) demonstrates that reducing polytomous data to the dichotomous level leads to a systematic loss of measurement information. For these considerations, the discussion will be here concentrated on polytomous IRT models for items with *ordered categories*.

The framework that will be considered in the present paper is the following. Consider the responses of a n-dimensional set $\mathcal{S} = \{S_1, \ldots, S_v, \ldots, S_n\}$ of subjects to a k-dimensional sequence $\mathcal{I} = \{I_1, \ldots, I_i, \ldots, I_k\}$ of items. Each subject may respond to item I_i in $m + 1$ ($m \geq 1$) ordered categories, $C_0, C_1, \ldots, C_m$; the generalization of this with regard to situations in which items have different numbers $m_i + 1$, $i = 1, \ldots, k$, of categories is straightforward, but would lead to more cumbersome notation. The *score* is chosen to be h in correspondence to C_h, $h = 0, 1, \ldots, m$. The actual response of S_v to I_i can be so represented as a selection vector $\boldsymbol{y}_{vi} = (y_{vi0}, y_{vi1}, \ldots, y_{vim})'$, where $\boldsymbol{y}_{vi}$ is an observation from the random variable $\boldsymbol{Y}_{vi}$ and $y_{vih} = 1$ if the response is in category C_h, and 0 otherwise. Let $\boldsymbol{y}_{vi}$ be a single element in the $n \times k$ data matrix $\boldsymbol{y}$. From now on it will be assumed that, for each item, the subject chooses one and only one of the $m + 1$ categories; consequently, incomplete designs will be excluded from the analysis. Moreover, let

$$x_{vi} = \max_{h \in \{0,1,\ldots,m\}} h y_{vih} = \sum_{h=0}^{m} h y_{vih} \tag{1}$$

be the *score* obtained by S_v to I_i. Naturally, $x_{vi} \in \{0, 1, \ldots, m\}$. Finally, let $\boldsymbol{x} = (x_{vi})$ be the *score matrix*. It is to be noted that there is a one-to-one correspondence between the 3-dimensional data matrix $\boldsymbol{y}$ and the 2-dimensional score matrix $\boldsymbol{x}$.

Analyzing such data, with respect to dichotomous data, requires the use of a model that can adequately handle the additional information that is supplied by the greater number of response categories (for a survey of polytomous IRT models see, *e.g.*, Ostini and Nering, 2006, van der Linden and Hambleton, 1997, or Tuerlinckx and Wang, 2004). Specifically, with reference to a single item, a model is usually a mathematical function used to describe the probability of responding in a category as function of ϑ (the discussion is restricted to models for items that measure one continuous latent variable, *i.e.*, *unidimensional latent trait models*). According to Chang and Mazzeo (1994), and Weiss and Yoes (1990), this function will be referred to as Item Category Response Function (ICRF) in order to reflect its specific item category role, and it will be denoted with

$$p_{ih}(\vartheta) = \mathsf{P}\left(Y_{ih} = 1 \left| \vartheta \right.\right) = \mathsf{P}\left(X_i = h \left| \vartheta \right.\right), \tag{2}$$

$i = 1, \ldots, k$, $h = 0, 1, \ldots, m$.

In analogy with the dichotomous case, and starting from (2), in order to obtain a single function for each item in $\mathcal{I}$ it is possible to define the expected value of the score X_i, conditionally at a given value of ϑ, as follows

$$e_i(\vartheta) = \mathsf{E}\left(X_i \left| \vartheta \right.\right) = \sum_{h=0}^{m} h p_{ih}(\vartheta), \tag{3}$$

$i = 1, \ldots, k$, that takes values in $[0, mk]$. The function $e_i(\vartheta)$ is commonly known as Item Response Function (IRF) for a polytomously-scored item and it can be viewed as a regression of the item score X_i onto the ϑ scale (Lord, 1980). Naturally, for dichotomous IRT models, the IRF coincides with the ICRF referred to C_1.

Functions (2) and (3) are both of crucial importance in IRT; consequently, an adequate model specification is a preeminent problem. In such circumstances, at least two routes are possible. The first is the *parametric* one, in which a simple parametric structure is assumed so that the estimation of an ICRF is reduced to the estimation of a vector parameter $\boldsymbol{\xi}_i$, of dimension varying from model to model, for each item in $\mathcal{I}$. This vector is usually considered to be of direct interest and its estimate is often used as a summary statistic to describe items (difficulty, discrimination, and so on; Lord, 1980). In order to estimate $\boldsymbol{\xi}_i$, Marginal Maximum Likelihood (MML) procedures can be considered (see, *e.g.*, Bock and Aitkin, 1981). MML estimation assumes that persons are randomly sampled from a population in which ability is distributed according to some proper prior density function $f(\vartheta)$ – usually a $N(0,1)$ is considered – with distribution function $F(\vartheta)$. The second route is the *nonparametric* one, in which estimation is made directly on $\boldsymbol{y}$ without assuming any mathematical form for the ICRF, in order to obtain more flexible estimates. In this context, kernel smoothing is a good, and most commonly used choice because of its practical and theoretical properties. In analogy with the MML estimation procedure, and in order to make the model identifiable, a standard normal is often considered as prior ability distribution. In Section 2 this approach is retraced and pointwise confidence intervals for IRFs are defined. Cross-validation estimation of the smoothing parameter is also described. In Section 3, based on the consistency of this approach, a kernel-based minimum distance estimator of a parametric model is presented. Finally, in Section 4, a Monte Carlo simulation study is carried out in order to evaluate the performance of this estimator, with respect to both MML and kernel smoothing, in possible situations of departure from the assumption $\vartheta \sim N(0,1)$ in the population.

2. The kernel smoothing approach

Nonparametric estimation of ICRFs have been popularized by proposing nonparametric regression methods, based on kernel smoothing approaches, which are implemented in `TestGraf` program (Ramsay, 1991, 1997, 2000). The basic idea of kernel smoothing is to obtain a nonparametric estimate of the ICRF by taking a (local) weighted average:

$$\widehat{p}_{ih}^{\text{ker}}(\vartheta) = \sum_{v=1}^{n} w_v(\vartheta)\, Y_{vih}, \tag{4}$$

at each evaluation point, where the weights $w_v(\vartheta)$ are defined so as to be maximal when $\vartheta = \vartheta_v$ and to be smoothly non-increasing as $|\vartheta - \vartheta_v|$ increases (Altman, 1992; Eubank, 1988; Härdle, 1990, 1991; Simonoff, 1996). The need to keep $\widehat{p}_{ih}^{\text{ker}}(\vartheta) \in [0,1]$, for each $\vartheta \in \mathrm{I\!R}$, argues for the additional constraints $w_v(\vartheta) \geq 0$ and $\sum_{v=1}^{n} w_v(\vartheta) = 1$, and as a consequence, it is preferable to use Nadaraya-Watson weights (Nadaraya, 1964; Watson, 1964):

$$w_v(\vartheta) = \frac{K\left(\dfrac{\vartheta - \vartheta_v}{\lambda_i}\right)}{\sum\limits_{v=1}^{n} K\left(\dfrac{\vartheta - \vartheta_v}{\lambda_i}\right)}, \tag{5}$$

where:

- $\lambda_i > 0$ is the so-called *smoothing parameter* controlling the amount of smoothness. It is chosen to obtain a desirable trade-off between the bias and the variance of estimation (see, *e.g.*, Härdle, 1990). As λ_i decreases, the bias will decrease and the variance of the estimated function at each evaluation point will increase. If λ_i increases, the reverse is true. Naturally, λ_i can vary from item to item, and this is underlined by the subscript "i". Ordinarily this trade-off is stated in terms of Mean Squared Error (MSE) of the estimator

  $$\begin{aligned} \mathsf{MSE}\left[\widehat{p}_{ih}^{\,\text{ker}}(\vartheta)\right] &= \mathsf{E}\left\{\left[\widehat{p}_{ih}^{\,\text{ker}}(\vartheta) - p_{ih}(\vartheta)\right]^2\right\} \\ &= \left\{\mathsf{Bias}\left[\widehat{p}_{ih}^{\,\text{ker}}(\vartheta)\right]\right\}^2 + \mathsf{Var}\left[\widehat{p}_{ih}^{\,\text{ker}}(\vartheta)\right], \end{aligned}$$

  where $p_{ih}(\vartheta)$ is the real but unknown ICRF, and λ_i is chosen minimizing this quantity. As suggested by Ramsay (2000), it turns out that MSE is minimized in a wide range of situations by letting λ_i be proportional to $n^{-1/5}$. Nevertheless, in Subsection 2.1, a simple "objective" risk-based method to select the best value for λ_i will be described.

- K is the so-called *kernel function*, a nonnegative, continuous ($\widehat{p}_{ih}^{\,\text{ker}}$ inherits the continuity from K) and usually symmetric function that is non-increasing as its argument moves further from zero. Since the performance of (5) largely depends on the choice of λ_i, rather than on the kernel function, a simple Gaussian kernel $K(u) = \exp(-u^2/2)$ is often preferred.

Consequently, the ICRF smoothing estimate becomes

$$\widehat{p}_{ih}^{\text{ker}}\left(\vartheta\right)=\frac{\sum\limits_{v=1}^{n}K\left(\dfrac{\vartheta-\vartheta_v}{\lambda_i}\right)Y_{vih}}{\sum\limits_{v=1}^{n}K\left(\dfrac{\vartheta-\vartheta_v}{\lambda_i}\right)}. \tag{6}$$

In (6), unlike the standard kernel regression estimators, the dependent variable is a binary variable Y_{ih} and the independent one is the latent ability variable ϑ_v. Unfortunately, ϑ_v cannot be directly observed. Kernel smoothing can still be used, but each ϑ_v in (6) must be replaced with a reasonable estimate $\widehat{\vartheta}_v$ (Ramsay, 1991), resulting in an estimate of the form:

$$\widehat{p}_{ih}^{\text{ker}}\left(\vartheta\right)=\sum_{v=1}^{n}\widehat{w}_v\left(\vartheta\right)Y_{vih}, \tag{7}$$

where

$$\widehat{w}_v\left(\vartheta\right)=\frac{K\left(\dfrac{\vartheta-\widehat{\vartheta}_v}{\lambda_i}\right)}{\sum\limits_{v=1}^{n}K\left(\dfrac{\vartheta-\widehat{\vartheta}_v}{\lambda_i}\right)}.$$

However, it is critical to deal with ability estimation when nonparametric assumptions are made concerning the form of the ICRFs because, in this situation, a test cannot yield anything more than rank order information about examinees. To see this, consider any strictly monotonic transformation $\tau=g\left(\vartheta\right)$ of the ability continuum. Then

$$p_{ih}\left(\vartheta\right)=p_{ih}\left\{g^{-1}\left[g\left(\vartheta\right)\right]\right\}=p_{ih}\left[g^{-1}\left(\tau\right)\right]=p_{ih}^{*}\left(\tau\right), \tag{8}$$

where the function $p_{ih}^{*}=p_{ih}\circ g^{-1}$ is the equivalent ICRF relative to the new ability continuum τ; thus, the choice of scale becomes perfectly arbitrary. This *lack of identifiability*, implicitly recognized in the MML estimation procedure, implies that what is being estimated are the values of the functions $p_{ih}(\vartheta)$, which are invariant with respect to monotone transformations of their domain. Consequently, as far as the n ability value estimates are concerned, only rank order considerations make sense. In particular, as suggested by Ramsay (1991, 2000, p. 102), to determine the estimates $\widehat{\vartheta}_v$ one could:

1. estimate the rank r_v of the v-th examinee by ranking the values T_v of some statistic T, $v = 1, \dots, n$. The total score of an examinee is the most obvious and used statistic for ranking examinees;

2. sort examinee response patterns according to the estimated ability ranking;

3. replace the rank r_v by the "r_v-th quantile" $\widehat{\vartheta}_v$ of some distribution function F that is seen to be appropriate to the contemplated application. The r_v-th quantile is such that $F\left(\widehat{\vartheta}_v\right) = r_v/(n+1)$, where the denominator $n+1$ is chosen in order to avoid an infinity value for the biggest $\widehat{\vartheta}_v$. Thus, the estimated ability value for S_v becomes $\widehat{\vartheta}_v = F^{-1}\left[r_v/(n+1)\right]$.

   The choice of F is equivalent to the choice of the ϑ metric. Historically, the standard normal distribution $F = \Phi$ has been heavily used even because it is one of the most used in applications of the parametric models to which the kernel model is often compared. Logically, other distributions are not excluded. For example, users who think of ability as percentages may prefer a distribution on $[0, 1]$ such as the Beta; a Beta$(2.5, 2.5)$ looks very much like a standard normal.

Since latent ability estimates are based on ranked total scores, they are usually referred to as *ordinal ability estimates*.

A further remark should be noted. The denominator of equation (7) is in effect (proportional to) a Rosenblatt-Parzen kernel estimator (see, *e.g.*, Silverman, 1986) of the ability density function $f(\vartheta)$. Although this density is already known, in the sense of being determined by the choice of the quantile distribution F, and consequently could be replaced by the actual density, this substitution is not recommended because it might result in occasional values of $\widehat{p}_{ih}^{\text{ker}}$ slightly outside of the natural interval $[0,1]$.

Finally note that, starting from (3), it is straightforward to define the kernel IRF estimator as follows

$$\widehat{e}_{i}^{\text{ker}}(\vartheta)=\sum_{h=0}^{m} h\widehat{p}_{ih}^{\text{ker}}(\vartheta)=\sum_{h=0}^{m} h\sum_{v=1}^{n}\widehat{w}_{v}(\vartheta)\,Y_{vih}=\sum_{v=1}^{n}\widehat{w}_{v}(\vartheta)\sum_{h=0}^{m} hY_{vih}. \tag{9}$$

2.1. Choosing the smoothing parameter

The choice of the smoothing parameter λ_i is important. Although it is informative to choose the smoothing parameter by trial and error, and although, as previously said, Ramsay (1991, 2000) suggests using a value proportional to $n^{-1/5}$ (in `TestGraf` the value $1.1n^{-1/5}$ is used by default), it is also convenient to have an objective, risk-based method for selecting the best value for λ_i. The literature on data-driven methods for selecting the optimal value for the smoothing parameter is vast. Cross-validation (Stone, 1974) is without doubt commonly used and simple to understand. Here, a description of cross-validation in the context of the kernel smoothing approach in IRT will be presented.

Before going on, let $\boldsymbol{y}_i=(\boldsymbol{y}_{1i},\dots,\boldsymbol{y}_{vi},\dots,\boldsymbol{y}_{ni})$ be the $(m+1)\times n$ selection matrix, referred to I_i, in which the v-th column contains the selection vector $\boldsymbol{y}_{vi}$. Moreover, let

$$\widehat{\boldsymbol{p}}_{i}^{\text{ker}}(\vartheta)=\left(\widehat{p}_{i0}^{\text{ker}}(\vartheta),\widehat{p}_{i1}^{\text{ker}}(\vartheta),\dots,\widehat{p}_{im}^{\text{ker}}(\vartheta)\right)'$$

be the $(m+1)$-dimensional vector of kernel-estimated probabilities, for item I_i, at the evaluation point ϑ. The probability kernel estimator evaluated in ϑ, for I_i, can thus be rewritten in the following form

$$\widehat{\boldsymbol{p}}_{i}^{\text{ker}}(\vartheta)=\sum_{v=1}^{n}\widehat{w}_{v}(\vartheta)\,\boldsymbol{y}_{vi}=\boldsymbol{y}_{i}\widehat{\boldsymbol{w}}(\vartheta),$$

where $\widehat{\boldsymbol{w}}(\vartheta)=\left(\widehat{w}_{1}(\vartheta),\dots,\widehat{w}_{v}(\vartheta),\dots,\widehat{w}_{n}(\vartheta)\right)'$.

In detail, cross-validation simultaneously fits and smooths the data contained in $\boldsymbol{y}_i$ by removing one “data point” $\boldsymbol{y}_{vi}$ at a time, estimating the value of $\boldsymbol{p}_i$ at the correspondent ordinal ability estimate $\widehat{\vartheta}_v$, and then comparing the estimate to the omitted, observed value. So the cross-validation statistic or score, $CV(\lambda_i)$, is

$$CV(\lambda_i) = \frac{1}{n}\sum_{v=1}^{n}\left(\boldsymbol{y}_{vi} - \widehat{\boldsymbol{p}}_i^{\text{ker},(-v)}\left(\widehat{\vartheta}_v\right)\right)'\left(\boldsymbol{y}_{vi} - \widehat{\boldsymbol{p}}_i^{\text{ker},(-v)}\left(\widehat{\vartheta}_v\right)\right), \quad (10)$$

where

$$\widehat{\boldsymbol{p}}_i^{\text{ker},(-v)}\left(\widehat{\vartheta}_v\right) = \frac{\sum\limits_{\substack{u=1\\u\neq v}}^{n} K\left(\dfrac{\widehat{\vartheta}_v - \widehat{\vartheta}_u}{\lambda_i}\right)\boldsymbol{y}_{ui}}{\sum\limits_{\substack{u=1\\u\neq v}}^{n} K\left(\dfrac{\widehat{\vartheta}_v - \widehat{\vartheta}_u}{\lambda_i}\right)}$$

is the estimated vector of probabilities at $\widehat{\vartheta}_v$ computed by removing the observed selection vector $\boldsymbol{y}_{vi}$. The value of λ_i that minimizes $CV(\lambda_i)$ is referred to as the cross-validation smoothing parameter, $\widehat{\lambda}_i^{CV}$, and it is possible to find it by systematically searching across a suitable smoothing parameter region.

2.2. Pointwise confidence intervals

In visual inspection, and graphical interpretation, of the estimated kernel curves, pointwise confidence intervals at the evaluation points $\vartheta \in \mathbb{R}$ provide relevant information because they indicate the extent to which the kernel ICRFs and IRFs are well defined across the range of ϑ considered. Moreover, they are useful when nonparametric models are compared with respect to parametric models. Here, the confidence limits provided by Ramsay (1991) for the ICRFs will be extended also to the IRFs.

Pointwise confidence intervals for ICRFs

Since $\widehat{p}_{ih}^{\,\mathsf{ker}}\left(\vartheta\right)$ is a linear function of the data, as can be easily seen from (7), and being $Y_{vih} \sim \text{Ber}\left[p_{ih}\left(\widehat{\vartheta}_v\right)\right]$,

$$\begin{aligned}\mathsf{Var}\left[\widehat{p}_{ih}^{\,\mathsf{ker}}\left(\vartheta\right)\right] &= \sum_{v=1}^{n}\left[\widehat{w}_v\left(\vartheta\right)\right]^2\mathsf{Var}\left(Y_{vih}\right)\\ &= \sum_{v=1}^{n}\left[\widehat{w}_v\left(\vartheta\right)\right]^2 p_{ih}\left(\widehat{\vartheta}_v\right)\left[1-p_{ih}\left(\widehat{\vartheta}_v\right)\right],\end{aligned}$$

holding if independence of the Y_{vih}s is assumed and possible error variation in the arguments, $\widehat{\vartheta}_v$, are ignored. Substituting p_{ih} with $\widehat{p}_{ih}^{\,\mathsf{ker}}$ yields the $(1-\alpha)\cdot 100\%$ pointwise confidence intervals

$$\widehat{p}_{ih}^{\,\mathsf{ker}}\left(\vartheta\right) \mp z_{1-\frac{\alpha}{2}}\sqrt{\sum_{v=1}^{n}\left[\widehat{w}_v\left(\vartheta\right)\right]^2 \widehat{p}_{ih}^{\,\mathsf{ker}}\left(\widehat{\vartheta}_v\right)\left[1-\widehat{p}_{ih}^{\,\mathsf{ker}}\left(\widehat{\vartheta}_v\right)\right]}, \tag{11}$$

where $z_{1-\frac{\alpha}{2}}$ is such that $\Phi\left[z_{1-\frac{\alpha}{2}}\right] = 1-\frac{\alpha}{2}$.

Pointwise confidence intervals for IRFs

Consider the IRF function defined in (9). In analogy with the previous case, the $(1-\alpha)\cdot 100\%$ pointwise confidence interval for the IRF $\widehat{e}_i^{\,\mathsf{ker}}$ is given by

$$\widehat{e}_{ih}^{\,\mathsf{ker}}\left(\vartheta\right) \mp z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathsf{Var}\left[\widehat{e}_i^{\,\mathsf{ker}}\left(\vartheta\right)\right]}}, \tag{12}$$

where, since $Y_{ih}Y_{it} \equiv 0$ for $h \neq t$, one has

$$\mathsf{Var}\left[\widehat{e}_i^{\,\mathsf{ker}}\left(\vartheta\right)\right] = \sum_{v=1}^{n}\left[\widehat{w}_v\left(\vartheta\right)\right]^2\mathsf{Var}\left(\sum_{h=0}^{m} hY_{vih}\right)$$

$$
\begin{aligned}
&= \sum_{v=1}^{n} \left[\widehat{w}_v(\vartheta)\right]^2 \left[\sum_{h=0}^{m} h^2 \mathsf{Var}\left(Y_{vih}\right) + \sum_{h=0}^{m} \sum_{t \neq h} ht \mathsf{Cov}\left(Y_{vih}, Y_{vit}\right)\right] \\
&= \sum_{v=1}^{n} \left[\widehat{w}_v(\vartheta)\right]^2 \left[\sum_{h=0}^{m} h^2 \mathsf{Var}\left(Y_{vih}\right) - \sum_{h=0}^{m} \sum_{t \neq h} ht \mathsf{E}\left(Y_{vih}\right) \mathsf{E}\left(Y_{vit}\right)\right] \\
&= \sum_{v=1}^{n} \left[\widehat{w}_v(\vartheta)\right]^2 \left\{\sum_{h=0}^{m} h^2 p_{ih}\left(\widehat{\vartheta}_v\right)\left[1 - p_{ih}\left(\widehat{\vartheta}_v\right)\right]\right. \\
&\quad \left. - \sum_{h=0}^{m} \sum_{t \neq h} ht p_{ih}\left(\widehat{\vartheta}_v\right) p_{it}\left(\widehat{\vartheta}_v\right)\right\}.
\end{aligned}
$$

Substituting p_{ih} with $\widehat{p}_{ih}^{\text{ker}}$ in $\mathsf{Var}\left[\widehat{e}_i^{\text{ker}}(\vartheta)\right]$, one obtains $\widehat{\mathsf{Var}\left[\widehat{e}_i^{\text{ker}}(\vartheta)\right]}$, quantity that has to be inserted in (12).

Really, intervals in (11) and (12) are, respectively, strictly speaking intervals for $\mathsf{E}\left[\widehat{p}_{ih}^{\text{ker}}(\vartheta)\right]$ and $\mathsf{E}\left[\widehat{e}_i^{\text{ker}}(\vartheta)\right]$, rather than for $p_{ih}(\vartheta)$ and $e_{ih}(\vartheta)$. Because of this, they share the bias present in $\widehat{p}_{ih}^{\text{ker}}$ and $\widehat{e}_i^{\text{ker}}$, respectively (for the ICRF case, cfr. Ramsay, 1991).

3. A kernel-based minimum distance estimator

Douglas (1997), in the dichotomous case, shows that although any $\widehat{p}_{i1}^{\text{ker}}(\vartheta)$ is an empirical regression estimate of Y_{i1} on a total score transformation, it can consistently estimate the true $p_{i1}(\vartheta)$. The author argues that this asymptotic result can easily be extended to the polytomous case. Moreover, Douglas (2001) proves that, for long tests, there is only one correct IRT model for a given choice of F, and nonparametric methods (including the kernel estimation approach) can consistently estimate it. Thus, following the idea of Douglas and Cohen (2001), if nonparametric estimated curves are meaningfully different from parametric ones, this parametric model – defined on the particular scale determined by F – is an uncorrected model for the data. In order to make valid this comparison, it is fundamental that the same F be used for both nonparametric and parametric curves. For example, if MML (that typically assumes a normal distribution for ϑ) is selected to fit a parametric model, kernel estimates represented on this same distribution can be compared to it.

Here the above mentioned idea is adopted in most general polytomous frame and in slightly different way: $\widehat{p}_{ih}^{\text{ker}}(\vartheta)$, $h=0,1,\ldots,m$, are estimated and parametric ICRFs are computed by finding the nearest approximation to the kernel ICRFs within the parametric family that is seen to be appropriate to the application. In the choice of the parametric family, visual inspections of the estimated kernel curves can be useful.

In detail, let $\widehat{\boldsymbol{\xi}}_i^{\text{ker}}$ be the parameter vector that minimizes an opportune "global" distance measure between kernel-estimated ICRFs, $\widehat{p}_{ih}^{\text{ker}}(\vartheta)$, and theoretical parametric ICRFs, $p_{ih}(\vartheta;\boldsymbol{\xi}_i)$, $h=1,\ldots,m$, over all values of $\boldsymbol{\xi}_i$ in the parameter space Ξ. Let $p_{ih}\left(\vartheta;\widehat{\boldsymbol{\xi}}_i^{\text{ker}}\right)$ be the resulting parametric approximation to $\widehat{p}_{ih}^{\text{ker}}(\vartheta)$. Among the possible choices for this "global" measure of distance, the Mean Root Integrated Squared Error (MRISE) is used here. For a generic item $I_i \in \mathcal{I}$, MRISE is defined as follows:

$$\mathsf{MRISE}_i = \frac{1}{m}\sum_{h=1}^{m}\sqrt{\int_{\mathrm{IR}}\left[\widehat{p}_{ih}^{\text{ker}}(\vartheta) - p_{ih}(\vartheta;\boldsymbol{\xi}_i)\right]^2 f(\vartheta)\,d\vartheta}. \qquad (13)$$

The weighting of squared error by $f(\vartheta)$ permits the measure to be most sensitive to departures for values of ϑ that are most commonly observed. MRISE is simple to compute and can be considered as a natural polytomous generalization of the RISE definable in the dichotomous case (cfr. Douglas and Cohen, 2001). Because of this natural generalization, MRISE values can be compared with RISE ones. The vector $\widehat{\boldsymbol{\xi}}_i^{\text{ker}}$ can be obtained by minimizing the MRISE_i, that is:

$$\widehat{\boldsymbol{\xi}}_i^{\text{ker}} = \underset{\boldsymbol{\xi}_i\in\Xi}{\arg\min}\,\mathsf{MRISE}_i. \qquad (14)$$

For computational convenience, MRISE_i can be minimized by discretizing the integral in (13) over some finite grid $\vartheta_1^*,\ldots,\vartheta_q^*,\ldots,\vartheta_s^*$ of ϑ values, considering:

$$\mathsf{MRISE}_i = \frac{1}{m}\sum_{h=1}^{m}\sqrt{\frac{\sum_{q=1}^{s}\left[\widehat{p}_{ih}^{\text{ker}}\left(\vartheta_q^*\right) - p_{ih}\left(\vartheta_q^*;\boldsymbol{\xi}_i\right)\right]^2 f\left(\vartheta_q^*\right)}{\sum_{q=1}^{s} f\left(\vartheta_q^*\right)}}. \qquad (15)$$

The Newton-Raphson algorithm, implemented in many environments such as, for example, `Mathematica`, can be applied to minimize (15) over parameters.

4. Simulation study

A Monte Carlo simulation study, under various conditions on the number of categories ($m = 2, 3$), on the number of items ($k = 15$ if $m = 2$, and $k = 12$ if $m = 3$), and on the ability distribution F in the population (normal distributions $N(0, \sigma)$ with different values of σ, and uniform distributions $U(a, b)$ defined on different intervals $[a, b]$), was performed. The aim was twofold: firstly, to investigate the performance of the kernel approach in estimating ICRFs and IRFs; secondly, to compare this approach with respect to both the kernel-based minimum distance estimator and the standard MML one. The simulation factors m and k were selected to reflect practical testing conditions, while the different distributions F were chosen to emulate situations in which the assumption $\vartheta \sim N(0, 1)$ may be untenable.

In detail, as data generator, a flexible parametric Generalized Partial Credit Model (GPCM; Muraki, 1992):

$$p_{ih}(\vartheta; \boldsymbol{\xi}_i) = \frac{\exp\left[\sum_{l=0}^{h} (\alpha_i \vartheta - \delta_{il})\right]}{\sum_{t=0}^{m} \exp\left[\sum_{l=0}^{t} (\alpha_i \vartheta - \delta_{il})\right]}, \tag{16}$$

was considered, where α_i and δ_{ih} are respectively referred to slope and item-category parameters for item $I_i \in \mathcal{I}$; for notational convenience, $\delta_{i0} = 0$. Thus, the parameter vector becomes $\boldsymbol{\xi}_i = (\alpha_i, \delta_{i1}, \ldots, \delta_{im})'$. Although the range of the slope parameters α_i is in principle the real line $\mathbb{R}$, they are usually positive and the values seen in practice are typically less than 2.5; for this reason α_i, $i = 1, \ldots, k$, were randomly drawn from a $U(0.5, 2.5)$. An $N(0, 1)$ was instead used to generate item-category parameters δ_{ih}, $i = 1, \ldots, k$, $h = 1, \ldots, m$.

Several distributions – three normal distributions $N(0,1)$, $N(0,0.5)$ and $N(0,2)$, and two uniform distributions $U(-1,1)$ and $U(-2,2)$ – were considered to randomly generate the true ability values ϑ_v, referred to $n = 800$ subjects. A (selection) response pattern $\boldsymbol{y}_{vi}$ was thus generated by sampling from a multinomial distribution with parameters:

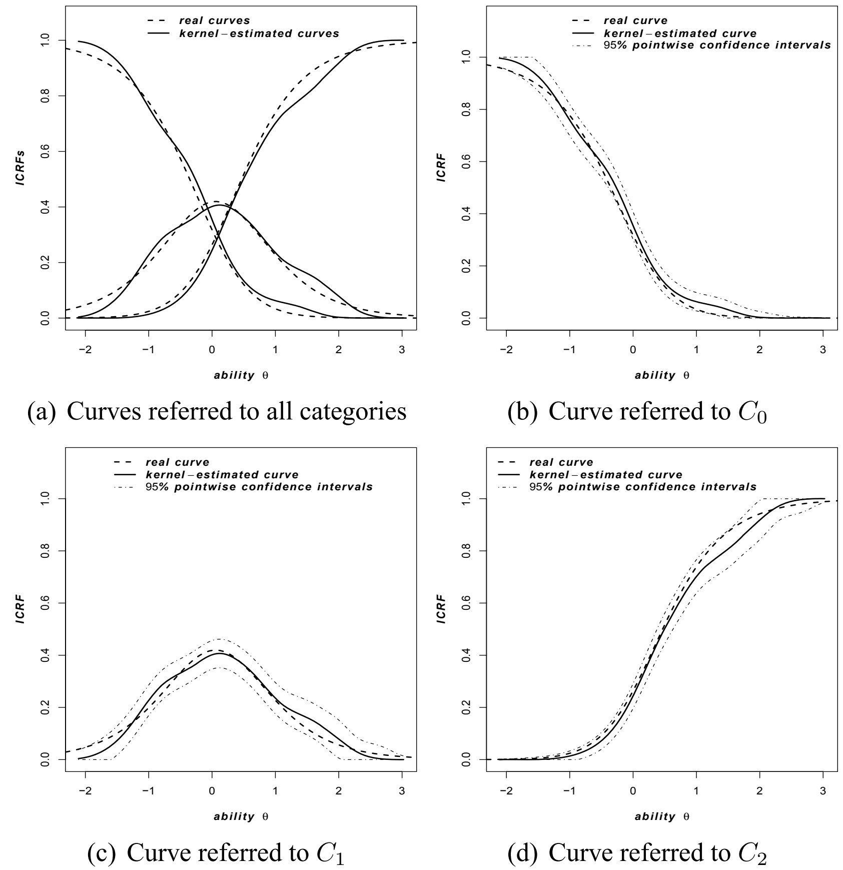
$$p_{i0}\left(\vartheta_v;\boldsymbol{\xi}_i\right), p_{i1}\left(\vartheta_v;\boldsymbol{\xi}_i\right), \dots, p_{im}\left(\vartheta_v;\boldsymbol{\xi}_i\right).$$

Starting from $\boldsymbol{y}$, and based on (1), the score matrix $\boldsymbol{x}$ was easily obtained. This process was replicated $M = 1000$ times in order to obtain a sufficient number of Monte Carlo samples.

These simulated data were firstly analyzed as indicated in Section 2: the 800 total scores were calculated, and the rows of the $800 \times k$ score matrix $\boldsymbol{x}$ were then sorted by these total scores. Thus, the total scores were replaced by the 800 quantiles of the standard normal distribution Φ, and for each item, kernel ICRF estimates, with a Gaussian kernel, were fitted using the cross-validation approach described in Subsection 2.1. The R program necessary to implement these estimates is available from the author upon request. The performance evaluation was at first made by a graphical comparison between underlying true curves in (16) and kernel smoothing estimated curves. In order to make "valid" such a comparison, 95% confidence intervals, defined in (11) and (12), were computed at the evaluation points and superimposed on the kernel-estimated curves. In all the displays and computations these evaluation points were $s = 101$ values equally-spaced between the ordinal ability estimates $\widehat{\vartheta}_1$ and $\widehat{\vartheta}_n$ inclusive.

For the case $m = 2$ and $k = 15$, with reference to a generic item, some exemplary ICRF kernel estimates, along with the true underlying parametric ICRFs, are displayed in Figure 1(a); 95% pointwise confidence intervals, computed using (11), are also shown in Figure 1(b)-1(d) for each single ICRF. In Figure 2(a), starting from data used for graphics in Figure 1, the real and the kernel-estimated IRFs are together plotted along with 95% pointwise confidence intervals computed on the basis of equation (12). In Figure 2(b) each kernel-estimated triple of probabilities:

$$\widehat{\boldsymbol{p}}_i^{\text{ker}}\left(\widehat{\vartheta}_v\right) = \left(\widehat{p}_{i0}^{\text{ker}}\left(\widehat{\vartheta}_v\right), \widehat{p}_{i1}^{\text{ker}}\left(\widehat{\vartheta}_v\right), \widehat{p}_{i2}^{\text{ker}}\left(\widehat{\vartheta}_v\right)\right)',$$



(a) Curves referred to all categories (b) Curve referred to C_0

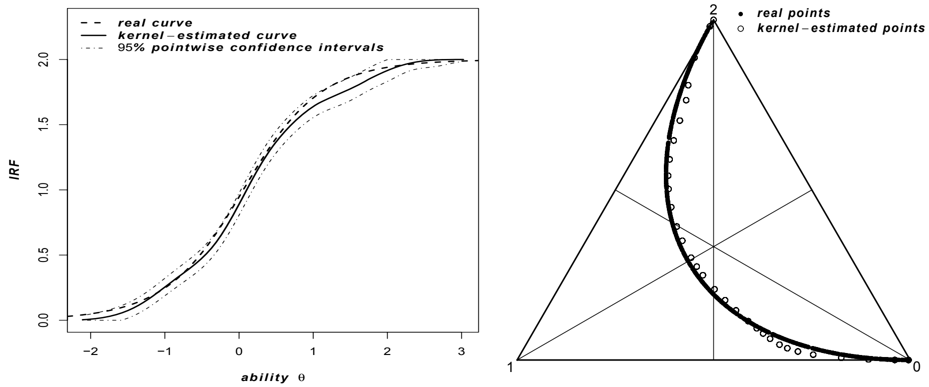
(c) Curve referred to C_1 (d) Curve referred to C_2

Figure 1. Kernel ICRF estimates (bold line), true underlying parametric ICRFs (dotted line), and 95% *pointwise confidence intervals (dashed thin lines)*

at varying of $v = 1, \ldots, 800$, is plotted as a point in the probability simplex along with the 800 true underlying triples of probabilities:

$$\left(p_{i0}\left(\vartheta_v; \boldsymbol{\xi}_i\right), p_{i1}\left(\vartheta_v; \boldsymbol{\xi}_i\right), p_{i2}\left(\vartheta_v; \boldsymbol{\xi}_i\right)\right)',$$

for more details on this kind of representation, see Aitchison (2003, pp. 5–9). It is to be noted that the "apparent" low number of kernel-estimated points, in the probability simplex, reflects the great number of ties between total scores (in this experiment the total score can only assume



(a) IRF (b) Probability simplex

Figure 2. On the left, kernel IRF estimate (bold line), true underlying parametric IRF (dotted line), and 95% *pointwise confidence intervals (dashed thin line), are plotted. On the right, the solid circles represent the true location in the probability simplex (equilateral triangle with unitary height) whereas the open circles indicate the kernel-estimated locations of* 800 *examinees*

integer values ranging from 0 to 30) and, consequently, between ordinal ability estimates referred to different subjects.

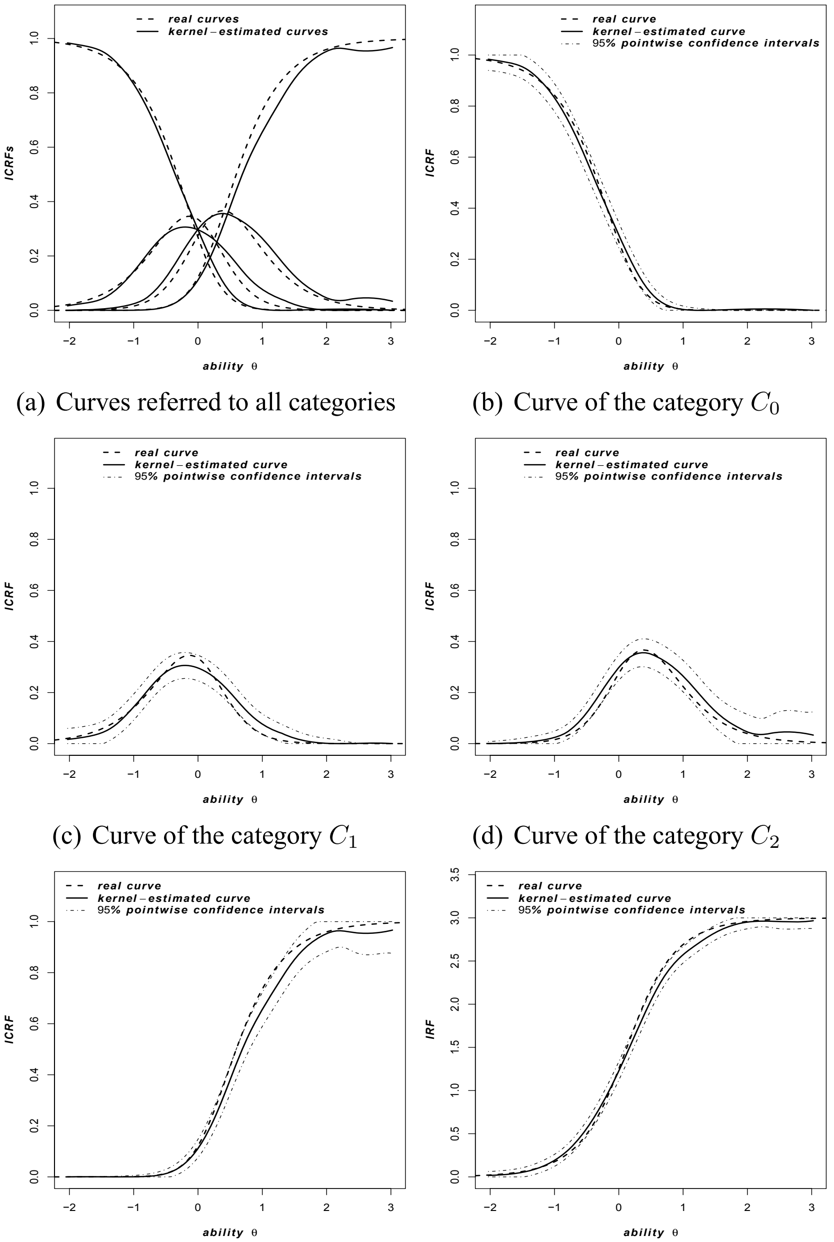
For the case $m = 3$ and $k = 12$, always with reference to a generic item, some exemplary ICRF and IRF kernel estimates, compared with the true underlying parametric ICRFs and IRF, respectively, are presented in Figure 3 along with 95% pointwise confidence intervals. In analogy with the previous case, in Figure 4 each 4-dimensional kernel-estimated probability vector:

$$\widehat{\boldsymbol{p}}_i^{\text{ker}}\left(\widehat{\vartheta}_v\right) = \left(\widehat{p}_{i0}^{\text{ker}}\left(\widehat{\vartheta}_v\right), \widehat{p}_{i1}^{\text{ker}}\left(\widehat{\vartheta}_v\right), \widehat{p}_{i2}^{\text{ker}}\left(\widehat{\vartheta}_v\right), \widehat{p}_{i3}^{\text{ker}}\left(\widehat{\vartheta}_v\right)\right)',$$

along with each true underlying probability vector:

$$\left(p_{i0}\left(\vartheta_v; \boldsymbol{\xi}_i\right), p_{i1}\left(\vartheta_v; \boldsymbol{\xi}_i\right), p_{i2}\left(\vartheta_v; \boldsymbol{\xi}_i\right), p_{i3}\left(\vartheta_v; \boldsymbol{\xi}_i\right)\right)',$$

$v = 1, \ldots, 800$, is represented as a point in the probability simplex using the 3-dimensional representation suggested by Aitchison (2003, p. 9).



(a) Curves referred to all categories

(b) Curve of the category C_0

(c) Curve of the category C_1

(d) Curve of the category C_2

(e) Curve of the category C_3

(f) IRF

Figure 3. Kernel-estimated curves (bold line), true underlying parametric curves (dotted line), and 95% *pointwise confidence intervals (dashed thin lines)*

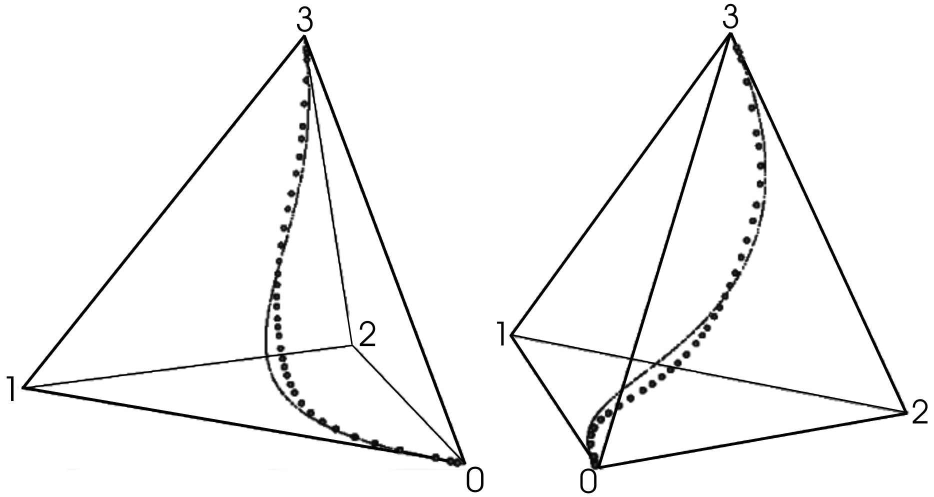


Figure 4. Two different angles of the 3*-dimensional probability simplex (tetrahedron with unitary height). The true (small points) and the kernel-estimated (big points) location of* 800 *examinees are represented*

Summarizing, in all simulations carried out, even if the entire set of graphical representations is obviously omitted, the confidence intervals and in general the agreement between real and estimated functions, suggests that the estimated curves were reasonably precise, above all when the ability values were drawn from an $N(0,1)$. However, although the phenomenon is not noticed in the displayed plots, confidence intervals could be somewhat wider at the ends of the curves, near to $\widehat{\vartheta}_1$ and $\widehat{\vartheta}_n$, when F arises from a unimodal density with a high variance and the probability is still substantially different from 0 and 1. This is due to the metric induced by this kind of distribution function that generates sparse ordinal ability estimates in the tails.

The same simulated data were also analyzed by two other methods; indeed, considering the GPCM in (16), both MML and kernel-based minimum distance estimation approaches were applied to estimate $\boldsymbol{\xi}_i$. MML parameter estimates $\widehat{\boldsymbol{\xi}}_i^{\text{MML}}$ were obtained by the `R` package *gpcm* (Johnson, 2007). The kernel-based minimum distance estimates $\widehat{\boldsymbol{\xi}}_i^{\text{ker}}$ were instead obtained in `Mathematica` minimizing, through an opportune implemented function, the discretized MRISE_i in (15). MRISE_i was also used as numerical index of estimation precision between the underlying true ICRFs and three estimated ICRFs, $p_{ih}\left(\vartheta; \widehat{\boldsymbol{\xi}}_i^{\text{MML}}\right)$, $\widehat{p}_{ih}^{\text{ker}}\left(\vartheta\right)$, and $p_{ih}\left(\vartheta; \widehat{\boldsymbol{\xi}}_i^{\text{ker}}\right)$, $h = 1, \dots, m$, referred to I_i. To allow these comparisons, and in order to conform with the usual applications, a prior standard normal ability distribution Φ was also considered in the MML estimation procedure. The global results of this study are shown in Table 1. Rough comparisons of performance between adopted estimation techniques were assessed by the following quantity

$$\overline{\mathsf{MRISE}} = \sum_{i=1}^{k} \mathsf{MRISE}_i.$$

The averages of these values, across the $M = 1000$ Monte Carlo replications, are reported in the fifth column of Table 1. The values shown in roman bold underline the best results for each combination of simulation factors. In the sixth column, standard deviations of these values, always with respect to the $M = 1000$ Monte Carlo replications, are also displayed. Other global information is reported in the last three columns of Table 1.

The results in Table 1 can thus be summarized. When abilities were randomly generated from an $N\left(0, 1\right)$, MML behaved slightly better than the other two considered estimation procedures in terms of the average of the $M = 1000$ $\overline{\mathsf{MRISE}}$-values; conversely, kernel estimation was the worst. The latter result is not surprising: unlike MML and kernel-based minimum distance estimation methods, the kernel approach did not use the additional information about the true underlying model in (16). On the other hand, when abilities were not randomly generated from an $N\left(0, 1\right)$, the considerations change. First of all, not surprisingly, the average values of $\overline{\mathsf{MRISE}}$ for the three considered estimation methods were worse with respect to the previous ones. Moreover, the performance of the kernel-based minimum distance estimator, in terms of $\overline{\mathsf{MRISE}}$, was always better than the two other estimators. This improvement in efficiency is substantial in the case $m = 2$ with true ability distribution $N\left(0, 2\right)$.

Table 1. Global comparison, based on $\overline{\mathsf{MRISE}}$, between MML ICRFs, kernel ICRFs, and kernel-based minimum distance ICRFs. Various simulation factors are considered.

Simulation factors				$\overline{\mathsf{MRISE}}$				
Number of categories	Number of items	True underlying ability distribution	ICRF Estimation methods	Average	Stand. Dev.	Min	Max	Range
3	**15**	$N(0,1)$	MML	**0.017238**	0.007102	0.002281	0.037456	0.035175
			Kernel	0.018192	0.007843	0.004805	0.039889	0.035084
			Minimum distance	0.017889	0.007114	0.003625	0.038506	0.034881
		$N(0,0.5)$	MML	0.097101	0.026981	0.078772	0.122054	0.043282
			Kernel	0.098858	0.029780	0.078122	0.122244	0.044122
			Minimum distance	**0.093443**	0.027881	0.074955	0.120088	0.045125
		$N(0,2)$	MML	0.111212	0.040011	0.090789	0.134933	0.044144
			Kernel	0.098853	0.032913	0.079981	0.122507	0.042526
			Minimum distance	**0.094078**	0.031944	0.072344	0.114522	0.042178
		$U(-1,1)$	MML	0.065868	0.019735	0.051009	0.082395	0.031386
			Kernel	0.070779	0.020411	0.054982	0.085858	0.030876
			Minimum distance	**0.063010**	0.018681	0.044241	0.084686	0.040445
		$U(-2,2)$	MML	0.040555	0.015826	0.026942	0.058062	0.031120
			Kernel	0.043017	0.014119	0.028004	0.060704	0.032700
			Minimum distance	**0.038188**	0.013117	0.019288	0.052189	0.032901
4	**12**	$N(0,1)$	MML	**0.014115**	0.009991	0.003804	0.029862	0.026058
			Kernel	0.016121	0.010408	0.012148	0.044000	0.031852
			Minimum distance	0.014880	0.010169	0.008724	0.041188	0.032464
		$N(0,0.5)$	MML	0.082654	0.029939	0.058411	0.104408	0.045997
			Kernel	0.084002	0.031000	0.063571	0.105811	0.042240
			Minimum distance	**0.079812**	0.030799	0.060917	0.099600	0.038683
		$N(0,2)$	MML	0.070199	0.020999	0.054799	0.088972	0.034173
			Kernel	0.065998	0.021488	0.051144	0.081811	0.030667
			Minimum distance	**0.062455**	0.021326	0.050811	0.078327	0.027516
		$U(-1,1)$	MML	0.064588	0.027616	0.042277	0.073521	0.031244
			Kernel	0.066881	0.028009	0.053744	0.080901	0.027157
			Minimum distance	**0.062308**	0.028444	0.050608	0.075277	0.024669
		$U(-2,2)$	MML	0.052334	0.014008	0.037081	0.070988	0.033907
			Kernel	0.046808	0.012854	0.028881	0.060914	0.032033
			Minimum distance	**0.042588**	0.013111	0.028211	0.058123	0.029912

5. Concluding remarks

In the literature a lot of works deal with the kernel smoothing approach in IRT; attention is often given to dichotomously-scored items even if, from a practical point of view, it is very common to have data that are polytomous, above all with 3 or 4 categories. Thus, in this paper, the kernel approach is discussed for polytomously-scored items describing a cross-validation approach to estimate the smoothing parameter. Moreover, based on Ramsay (1991), pointwise confidence intervals for IRFs are provided.

Motivated by the consideration that, if no *a priori* information is available, but a parametric model is however preferred, a preliminary nonparametric analysis could give valuable indication of features useful either in suggesting simple parametric formulations or, vice versa, in rejecting any "rigid" parametric specification, a kernel-based minimum distance estimator of parametric IRT models was also proposed. This proposal was justified by the consistency of the kernel method (Douglas, 1997).Through a Monte Carlo simulation study, a comparison with the standard MML approach was also accomplished. In order to make this valid, a standard normal distribution – as usual – was assumed in both methods as a prior for the ability distribution in the population. From this study it appears that the proposed kernel-based minimum distance estimator behaves better in situations of departure from the assumption of standard normality, while MML behaves slightly better otherwise.

Acknowledgements: The author would like to thank the anonymous referee for his/her useful suggestions.

References

Aitchison, J. (2003), *The statistical analysis of compositional data*, Chapman & Hall, London.

Altman, N. S. (1992), An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician*, 46, 175–185.

Bejar, I. I. (1977), An application of the continuous response level model to personality measurement, *Applied Psychological Measurement*, 1, 509–521.

Bock, R. D., Aitkin, M. (1981), Marginal maximum likelihood estimation of item parameters: application of an EM algorithm, *Psychometrika*, 46, 443–459.

Chang, H-H., Mazzeo, J. (1994), The unique correspondence of the item response function and item category response functions in polytomously scored item response models, *Psychometrika*, 59, 391–404.

Cohen, J. (1983), The cost of dichotomization, *Applied Psychological Measurement*, 7, 249–253.

Douglas, J. A. (1997), Joint consistency of nonparametric item characteristic curve and ability estimation, *Psychometrika*, 62, 7–28.

Douglas, J. A. (2001), Asymptotic identifiability of nonparametric item response models, *Psychometrika*, 66, 531–540.

Douglas, J. A., Cohen, A. (2001), Nonparametric item response function estimation for assessing parametric model fit, *Applied Psychological Measurement*, 25, 234–243.

Eubank, R. L. (1988), *Spline smoothing and nonparametric regression*, Marcel Dekker, New York.

Härdle, W. (1990), *Smoothing techniques with implementation in S*, Springer-Verlag, New York.

Härdle, W. (1991), *Applied nonparametric regression*, Cambridge University Press, Cambridge.

Johnson, M. S. (2007), Marginal maximum likelihood estimation of item response models in R, *Journal of Statistical Software*, 20, 1–24.

Lord, F. M. (1980), *Application of item response theory to practical testing problems*, Erlbaum, Hillsdale.

Masters, G. N. (1988), Measurement models for ordered response categories, in R. Langeheine and J. Rost (eds.), *Latent trait and latent class models*, Plenum Publ. Corp., New York, 11–29.

Muraki, E. (1992,. A generalized partial credit model: application of an EM algorithm, *Applied Psychological Measurement*, 16, 159–176.

Nadaraya, E. A. (1964), On estimating regression, *Theory of Probability and Its Applications*, 9, 141–142.

Ostini, R., Nering, M. L. (2006), *Polytomous item response theory models*, Sage Publications, London.

Ramsay, J. O. (1991), Kernel smoothing approaches to nonparametric item characteristic curve estimation, *Psychometrika*, 56, 611–630.

Ramsay, J. O. (1997), A functional approach to modeling test data, in W. J. van der Linden and R. K. Hambleton (eds.), *Handbook of modern item response theory*, Springer-Verlag, New York, 381–394.

Ramsay, J. O. (2000), TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data, Available from http://www.psych.-mcgill.ca/faculty/ramsay/ramsay.html.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis*, Chapman & Hall, London.

Simonoff, J. S. (1996), *Smoothing methods in statistics*, Springer-Verlag, New York.

Stone, M. (1974), Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society: Series B*, 36, 111–147.

Tuerlinckx, F., Wang, W.-C. (2004), Models for polytomous data. in M. Wilson and P. De Boeck (eds.), *Explanatory item response models – a generalized linear and nonlinear approach*, Springer-Verlag, New York, 75–109.

van der Linden, W. J., Hambleton, R. K. (1997), *Handbook of modern item response theory*, Springer-Verlag, New York.

Watson, G. S. (1964), Smooth regression analysis, *Sankhy aitclose, Series A*, 26, 359–372.

Weiss, D. J., Yoes, M. E. (1990), Item response theory, in R. K. Hambleton and J. N. Zaal (eds.), *Advances in educational and psychological testing: theory and applications*, Kluwer Academic Publishers, Norwell, 69–95.