



Coherent psychometric modelling with Bayesian nonparametrics

George Karabatsos^{1*} and Stephen G. Walker²

¹College of Education, University of Illinois-Chicago, USA

²Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK

In this paper we argue that model selection, as commonly practised in psychometrics, violates certain principles of coherence. On the other hand, we show that Bayesian nonparametrics provides a coherent basis for model selection, through the use of a 'nonparametric' prior distribution that has a large support on the space of sampling distributions. We illustrate model selection under the Bayesian nonparametric approach, through the analysis of real questionnaire data. Also, we present ways to use the Bayesian nonparametric framework to define very flexible psychometric models, through the specification of a nonparametric prior distribution that supports all distribution functions for the inverse link, including the standard logistic distribution functions. The Bayesian nonparametric approach provides a coherent method for model selection that can be applied to any statistical model, including psychometric models. Moreover, under a 'non-informative' choice of nonparametric prior, the Bayesian nonparametric approach is easy to apply, and selects the model that maximizes the log likelihood. Thus, under this choice of prior, the approach can be extended to non-Bayesian settings where the parameters of the competing models are estimated by likelihood maximization, and it can be used with any psychometric software package that routinely reports the model log likelihood.

1. Introduction

Model selection plays a prominent role in psychometric practice, where often, two or more models are fitted to data, and the aim is to select the single model that best describes the data-generating process. The 'best' model would provide the most accurate summary of each examinee's performance, and the most accurate description of the items of the test (e.g. examination, questionnaire). Indeed, the psychometric

*Correspondence should be addressed to Dr George Karabatsos, University of Illinois, 1040 W. Harrison Street (MC 147), Chicago, IL 60607, USA (e-mail: georgek@uic.edu).

literature contains many articles that either propose, investigate, or advocate methods of model selection.¹

There are excellent critical reviews of psychometric modelling from the perspective of measurement theory (e.g. Luce, 2005; Michell, 1999). In this paper we argue that the practice of model selection in psychometrics is essentially incoherent, in the sense that this practice is inconsistent with the axioms of quantitative coherence. These axioms, which underpin Bayesian decision theory and rational decision-making, describe the sufficient conditions for statistical inference to be free from paradox and contradictions (see Bernardo & Smith, 1994). We also show that Bayesian nonparametric statistical inference provides a general solution to this incoherence problem in model selection. In so doing, we encourage a coherent way to think about the modelling process.

We set the notational scene for our argument:

- (1) Let X denote a random variable defined on a given finite-dimensional sample space $\mathcal{X} = \{x\} \subseteq \mathbb{R}^q$, with $\{x\}$ denoting the set of all possible realizations of that random variable.
- (2) Let $f(\cdot)$ denote the probability density function of a random variable X , with $f(x)$ the probability density of x (for all $x \in \mathcal{X}$); a density function $f(\cdot)$ corresponds² to a probability distribution function $F(\cdot)$, with $F(B) = \Pr(x : x \in B)$ denoting the probability of the event B , $0 \leq F(B) \leq 1$ for all subsets $B \subseteq \mathcal{X}$, and $F(\mathcal{X}) = 1$.
- (3) Define $\Omega_{\mathcal{X}} = \{f(\cdot)\}_{\mathcal{X}}$ as the set of all probability density functions with domain \mathcal{X} .
- (4) Let $\mathbf{x}^n = \{x_i \in \mathcal{X}\}_{i=1}^n$ denote a set of data of sample size n , which is generated from some true density f_0 from $\Omega_{\mathcal{X}}$ (with the density f_0 corresponding to some true distribution F_0).
- (5) Consider the density function f (distribution function F) as a random variable governed by some probability distribution with sample space $\Omega_{\mathcal{X}}$.

Then any model, say M_d , defines, or can be defined by, a *prior* probability distribution³ Π_d on $\Omega_{\mathcal{X}}$, given by

$$\Pi_d(A) = \Pr\{f \in A\}$$

for sets $A \subseteq \Omega_{\mathcal{X}}$. The prior distribution Π_d represents prior beliefs about the true sampling density f_0 . Given a set of data \mathbf{x}^n generated from f_0 , the prior distribution Π_d is updated⁴ to the posterior distribution Π_{nd} . On the basis of this posterior distribution, a predictive density $f_n(\cdot | M_d)$ is constructed under the assumptions of model M_d ; and this

¹ In psychometrics, model selection is used in statistical tests of item bias (i.e. differential item functioning analysis), in 'person fit analysis' where the aim is to identify examinees who provided surprising responses to items of a test (e.g. due to cheating, lucky guessing, etc.), in test equating where the objective is to match up the scores of two (or more) different tests, and in parametric point estimation where the aim is to find the optimal value of the parameter (that defines the optimal model), given the data.

² The density f and distribution function F have the relation $F(B) = \Pr(B) = \int_B f(x)dx$ (with $0 \leq f(x) < \infty$, all $x \in \mathcal{X}$) when the random variable X is continuous, and $F(B) = \Pr(B) = \sum_{x \in B} f(x)$ (with $0 \leq f(x) \leq 1$, all $x \in \mathcal{X}$) when X is discrete.

³ Our definition encompasses both the Bayesian and the frequentist approach to psychometric modelling. For example, in Section 3, we show that the maximum likelihood estimate corresponds to statistical inference under a specific choice of prior distribution.

⁴ In Section 2.1 we describe how this update is made.

predictive density provides an estimate of the true f_0 . Given a set of possible models $\{M_d\}_1^D$ that a psychometrician considers, the goal of model selection is to identify the single model M_d with predictive density $f_n(\cdot|M_d)$ that is closest (in some sense) to the unknown true sampling density f_0 .

There are many ways to perform model selection. In the Bayesian approach to statistical inference, decision theory provides a general framework for model selection (which we describe in Section 3). In this framework, models are compared on the basis of predictive density functions. Of course, when the true density f_0 is known, model selection is straightforward, since in this case the best predictive density would obviously be f_0 . Of course, in psychometric analyses of real data, the true density f_0 of the data is unknown, and as a consequence of this lack of knowledge, f_0 is estimated. Usually, the psychometrician determines this estimate through model selection, i.e. by considering a set of models $\{M_1, \dots, M_D\}$ that correspond to posterior distributions $\Pi_{n1}, \dots, \Pi_{nD}$, and then determining the model having the best estimate $f_n(\cdot|M_d)$ of the true f_0 .

However, what is the meaning of the posterior distribution, when the model is to be determined? In the attempt to answer this question, one can easily get into meaningless circular arguments. In particular, one key axiom of quantitative coherence asserts that the prior distribution must represent *actual* beliefs (e.g. of the psychometrician). Instead, in any of the existing approaches to model selection in psychometrics, a set $\{M_1, \dots, M_D\}$ of possible choices of model is considered, and these models are defined by ‘prior’ distributions $\Pi_1, \dots, \Pi_d, \dots, \Pi_D$. Since all of these models (priors) are possible at the outset, none of them can represent actual prior beliefs, and nor can any of these models be an approximation to prior beliefs. To elucidate, Π_1 cannot represent prior beliefs nor be an approximation to prior beliefs because Π_2 , and others, are being entertained. Thus, there is actually more uncertainty in the prior beliefs than the amount of uncertainty represented in the prior Π_1 alone. The same is true for the prior Π_2 , and so on for the rest of the priors (models) Π_1, \dots, Π_D .

If none of these so-called prior distributions (models) represent prior beliefs, and indeed are known not to, then any posterior distribution derived from one of them cannot represent posterior beliefs. The issue here is that the chosen prior is known not to represent actual prior beliefs, and thus the posterior distribution derived from it cannot represent actual beliefs. Therefore, such a posterior distribution is not appropriate for undertaking statistical inference. The incoherence lies in knowing the prior chosen is wrong (it does not represent actual prior beliefs), but it is used anyway.

In light of these observations, we assert that Bayesian nonparametrics provides a coherent basis for Bayesian model selection. That is, the ‘correct’ Bayesian approach to model selection is based on specifying a model M_0 that is defined by a ‘large’ prior distribution Π that gives full or large support to $\Omega_{\mathcal{X}}$. Given a set of data, this prior is updated to a posterior distribution Π_n , and a predictive density f_n is obtained from the posterior Π_n .

To elaborate, a large prior distribution is a probability measure that supports a large number of density functions. For example, for density functions defined on the real line (with $\mathcal{X} = \mathbb{R}$), a prior on the normal family supports only density functions with a normal shape. A large prior, on the other hand, supports a large number of shapes of density functions, and potentially supports all density functions. Technically, the support of a prior is most conveniently defined via a distance function between

densities, say $d(f,g)$. The support of the prior Π is defined as

$$S = \{f : \Pi(N_\epsilon(f)) > 0\},$$

for all $\epsilon > 0$, where

$$N_\epsilon(f) = \{g : d(f,g) < \epsilon\}.$$

Clearly S is a subset of $\Omega_{\mathcal{X}}$ which is the set of all density functions. So a large prior means a large S and one can quantify exactly how large S is. As an illustration, consider the Pólya tree prior (Ferguson, 1974), which supports density functions, with expectation the density function f_* . Then g is in the support of the prior for every g such that the Kullback–Leibler divergence between g and f_* is finite. (The Kullback–Leibler divergence is one choice of the function $d(f,g)$, and we define this divergence measure in Section 3.)

Retaining the large prior and posterior for inference can easily be done, it is Bayesian nonparametric inference. However, while nonparametric statisticians are comfortable using only the nonparametric posterior Π_n for inference, this is not a very popular approach. We acknowledge that researchers in many disciplines, including psychometrics, prefer simple, parametric (i.e. finite-dimensional) models. Hence, in psychometric practice, there is still a strong need for a method to select among a set of parametric models, while avoiding incoherence. This is possible using the tools of decision theory.

We recommend a Bayesian nonparametric approach to model selection, based on decision theory (as described in detail in Section 3). Our approach can be briefly summarized in two steps:

- (1) Specify a large prior to define a model M_0 (with resulting posterior Π_n and predictive density f_n , given the data \mathbf{x}^n), so that there is no need to check the fit of this model to the data \mathbf{x}^n .
- (2) From a set of parametric ('reduced') models under consideration $\{M_d\}_1^D$, select the model that has a predictive density f_{nd} that is closest (in some well-defined sense) to the predictive density f_n of M_0 , with f_n providing an estimate of the true density f_0 .

These two steps provide a coherent approach to Bayesian model selection, which is clearly an important aspect of parametric inference. This approach to model selection was developed by Gutiérrez-Peña and Walker (2005) for independent and identically distributed observations, and further investigated by Karabatsos (2006). In this paper, we extend this approach to regression models that include psychometric models.

It is worth elaborating on the idea mentioned in step 1, that a large prior defines a model that does not need to be checked. A large (nonparametric) prior does not need to be checked, because it has large support, as we mentioned earlier. Essentially, for a prior with large support, nothing, or very little, has been left out, and so there is nothing to check. The 'true' density f_0 (that generated a given set of data \mathbf{x}^n) is in the support of this prior. Certainly, we agree with the perspective that model checking plays an important role in Bayesian modelling (see Gelman, Carlin, Stern, & Rubin, 2003). The idea of model checking is to replace a prior, or support, if some statistic lies in a certain critical region. However, if a prior with 'maximum' support has been used in the first

place, there would be nothing to replace it with anyway. Nonparametric priors are not checked. There is no need to check them.

While we recommend a Bayesian nonparametric approach to model selection, there exists the counter-argument that for parametric model selection there is no need to refer to a large model (large prior). There are two reasons for this counter-argument.⁵ First, each of the D parametric priors can be regarded conditional on the corresponding model structure being appropriate, when one has uncertainty about the model structure. Second, it is not necessary to believe that the prior and posterior degrees of belief on a model represent the probability of a model being ‘true’, they can simply be interpreted as subjective weights of the predictive distribution for the data, when a definite mixture model of D components is proposed. We address both parts of this popular argument.

To address the first reason for this argument, we reiterate that a Bayesian is given the task of constructing a prior distribution which genuinely reflects prior beliefs. This is the Bayesian approach and should be followed. How does a Bayesian proceed when there are a number of parametric models under consideration? The prior must reflect this level of uncertainty. None of the models themselves can adequately reflect all this uncertainty, and so the self-evident choice of prior in this case is one that covers all the models. The posterior is then the one based on this all-encompassing prior. Subsequent model selection strategies lead to a posterior which is not based on all the prior uncertainty and so is underestimating the uncertainty. We see this as a serious problem.

To address the second reason for this argument, we agree that the all-encompassing prior can be based on a mixture over all the models under consideration. But the key point is that this prior is not checked or compared, as this violates the definition of a prior distribution. So, the prior must be large enough to capture all the prior uncertainty. A Bayesian nonparametric approach could be adopted in this case, but the key is that the prior is large enough to capture all the uncertainty that exists by the very fact that a number of models are thought possible.

The following sections describe our Bayesian nonparametric framework for model selection. In Section 2, we review the Bayesian nonparametric model. In this section we also present some very flexible psychometric models that can be constructed using a large prior. In Section 3 we then present a coherent approach to model selection. Then in Section 4 we illustrate an application of our Bayesian nonparametric model selection approach on real data. We show that our approach not only has theoretical support as a coherent procedure for statistical inference, but also it can easily be implemented for the practice of model selection in psychometrics. In Section 5 we end with some conclusions, where we also discuss an extension of the model selection approach to the problem of checking the adequacy of a single parametric model.

2. Bayesian nonparametrics

2.1 Bayesian nonparametric modelling

There are many different types of nonparametric (‘large’) prior distributions that can be used for statistical inference. Of these, the most popular is Dirichlet process prior distribution (Ferguson, 1973), which has seen numerous applications. For reviews of

⁵ We thank a referee for reminding us of this argument.

Bayesian nonparametric inference, see for example Walker, Damien, Laud, and Smith (1999) and Müller and Quintana (2004).

In this section, we describe a straightforward representation of the Dirichlet process prior distribution, as this is the prior we focus on throughout this paper. Recall from Section 1 that a large prior distribution (e.g. Dirichlet process prior) is updated by a set of data to yield a posterior distribution. Then in Section 2.2, on the basis of a large, Dirichlet process prior, we suggest some possible ways to construct very flexible psychometric models.

Before we proceed to describe the Dirichlet process prior, it is first necessary to clearly establish the fact that any prior distribution generates random distribution (or density) functions. (Recall that we defined a random distribution function in Section 1.) In the following two paragraphs, we illustrate this fact in the context of a parametric model and a nonparametric model, respectively.

On the one hand, a parametric model generates a random (finite-dimensional) parameter which then fits into a family of distributions. To give a very simple example of this idea, consider the case where the sample space is defined by the real line, with $\mathcal{X} = \mathbb{R}$, and that the family of distributions is defined by the set Ψ of all normal distributions, given by

$$\Psi = \{\text{Normal}(\cdot|\mu, \sigma^2) : (\mu, \sigma^2) \in \mathbb{R} \times [0, \infty)\}.$$

Of course, for all $x \in \mathbb{R}$ and all $(\mu, \sigma^2) \in \mathbb{R} \times [0, \infty)$, the normal distribution function is defined by

$$\text{Normal}(x|\mu, \sigma^2) = \Pr(X \leq x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt,$$

and this normal distribution function corresponds to the density function defined by

$$\text{normal}(x|\mu, \sigma^2) = f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Clearly, Ψ is a strict subset of $\Omega_{\mathbb{R}}$, because the set $\Omega_{\mathbb{R}}$ also contains asymmetric distributions, multimodal distributions, logistic distributions, Cauchy distributions, and so forth. Then, it is said that the normal distribution is a parametric model that generates a random finite-dimensional parameter $\theta = (\mu, \sigma^2)$ that is induced by a prior distribution $P(\mu, \sigma^2)$ on the finite-dimensional parameter space $\Theta = \mathbb{R} \times [0, \infty)$. In turn, this prior distribution $P(\mu, \sigma^2)$ induces a random distribution F governed by a prior distribution Π on $\Omega_{\mathbb{R}}$, with this prior distribution satisfying $\Pi(\Psi) = 1$ and $(\Pi(\Omega_{\mathbb{R}} - \Psi) = 0)$. Of course, while we exemplify with the normal model, it is possible to characterize any parametric model as a random distribution F (random density f) governed by a prior distribution Π on $\Omega_{\mathcal{X}}$, to any particular corresponding to of sample space \mathcal{X} , and any particular choice of finite-dimensional parameter θ .

On the other hand, a nonparametric ('large') prior distribution Π on $\Omega_{\mathcal{X}}$, such as the Dirichlet process prior, can be specified to satisfy $\Pi(\Omega_{\mathcal{X}}) = 1$, i.e. can be specified to support (assign positive probability to) the set of all distribution functions.

Now we proceed to describe the Dirichlet process prior distribution. There are many ways to describe the Dirichlet process. In describing this nonparametric ('large') prior distribution, it is helpful to view it as a probability distribution governing a stochastic

process whose sample paths behave almost surely as a distribution function. The most convenient way to describe the Dirichlet process is through a representation based on a countably infinite sampling strategy. So let θ_j , $j = 1, 2, \dots$, denote values of a random variable that is independent and identically distributed from a fixed distribution function G . Also, let v_j , $j = 1, 2, \dots$, be a random variable that is independent and identically distributed from the beta distribution with parameters $(1, c)$, for some positive constant $c > 0$. It is known that this beta distribution has a density defined by

$$f(v) = c(1 - v)^{c-1},$$

for all $v \in (0, 1)$. Now denote $\mathbf{1}(x_i \leq x)$ as the indicator function that equals 1 whenever $x_i \leq x$ and equals 0 otherwise. Then a random distribution function F chosen from a Dirichlet process prior with parameters (c, G) can be constructed via

$$F(x) = \sum_{j=1}^{\infty} w_j \mathbf{1}(\theta_j \leq x),$$

where $w_1 = v_1$ and, for $j > 1$,

$$w_j = v_j \prod_{l < j} (1 - v_l).$$

We denote such a model (prior) by $\Pi(c, G)$. Under the Dirichlet process prior distribution, the (prior) mean of the random distribution F is defined by $E\{F(B)\} = G(B)$ for any measurable set B , and the (prior) variance of F is defined by

$$\text{Var}\{F(B)\} = \frac{G(B)[1 - G(B)]}{c + 1}.$$

Hence, the parameter c in the Dirichlet process prior reflects the amount of uncertainty in the random distribution F such that the variance of F increases as c becomes small.

When the prior distribution Π is specifically a Dirichlet process, the posterior distribution of F is also a Dirichlet process, with updated parameters given by $c \rightarrow c + n$, and

$$G \rightarrow \frac{cG + nG_n}{c + n} = F_n = E(F|\mathbf{x}^n),$$

where G_n is the empirical distribution function of the observed data $\mathbf{x}^n = \{x_i\}_{i=1}^n$. (The empirical distribution function is defined by the distribution that assigns probability $1/n$ to every observation x_i from a set of data \mathbf{x}^n .) Hence, under the Dirichlet process prior, the posterior average of F is a mixture of the data, via the empirical distribution function G_n , and the prior mean G .

The Bayesian bootstrap (Rubin, 1981) is defined by the posterior distribution of F based on setting $c = 0$. Of course, there is no prior distribution which allows this. However, one can always put c as close to 0 as one likes, in which case the Bayesian bootstrap can be seen as providing an approximate posterior distribution of F which only utilizes the data (via the empirical distribution G_n). A random distribution function from this 'Bayesian bootstrap' posterior distribution can be obtained via

$$F(x) = \sum_{i=1}^n w_i \mathbf{1}(x_i \leq x),$$

where $\{x_1, \dots, x_n\}$ are the data, and where the vector of weights $\mathbf{w} = (w_1, \dots, w_n)$ (satisfying $\sum_{i=1}^n w_i = 1$) are from a Dirichlet distribution with parameters $(\alpha_1, \dots, \alpha_n)$ all set to 1. (This particular Dirichlet distribution has a density defined by $f(\mathbf{w}) = \Gamma(n)$, for all values of \mathbf{w} .) Under this ‘Bayesian bootstrap’ posterior distribution, the posterior average ($F_n = E(F|\mathbf{x}^n)$) of the random distribution function F is equal to the empirical distribution function G_n . Hence, the Bayesian bootstrap procedure can be viewed as corresponding to inference under a Dirichlet process prior distribution that is ‘non-informative’, in the sense that this prior assigns equal prior probabilities to all members of the set of distribution functions $\Omega_{\mathcal{X}}$.

It is clear from the construction of a Dirichlet process prior that this prior supports only discrete, random distribution functions. Continuous random distribution functions are often preferred, and a way to achieve this using the Dirichlet process is to make use of a mixture model. In this so-called Dirichlet process mixture model, the (continuous) random density function f is constructed via

$$f_P(x) = \int K(x; \boldsymbol{\theta}) P(d\boldsymbol{\theta}).$$

Here, P is a random distribution governed by a Dirichlet process prior distribution, and $K = (x; \boldsymbol{\theta})$ is a kernel density function. An example of a kernel density function $K = (x; \boldsymbol{\theta})$ is provided by the density of the normal distribution, normal $(x|\mu, \sigma^2)$, in which case $\boldsymbol{\theta} = (\mu, \sigma^2)$.

2.2 Flexible psychometric modelling

As this paper is primarily concerned with regression (psychometric) models, it is incumbent on us to specify Bayesian nonparametric regression models. We will describe some flexible models for item response data, which can be constructed using a large, Dirichlet process prior.

Consider the case when the item-response data x_i ($i = 1, \dots, n$) are binary outcomes, with $x_i \in \{0, 1\}$ for all i . For such data, a popular model is based on a monotonicity assumption, and in this case we can model

$$\Pr(x_i = 1 | \mathbf{z}_i, \boldsymbol{\lambda}) = G(\boldsymbol{\lambda}^T \mathbf{z}_i),$$

where $\boldsymbol{\lambda}$ denotes a column vector of parameters (with the transpose $\boldsymbol{\lambda}^T$ being a row vector), \mathbf{z}_i is a vector of covariates that is associated the i th outcome x_i , and G is a cumulative distribution that defines the inverse-link function. In a parametric model, G often is specified as the logistic distribution function (with location parameter 0 and scale parameter 1), and this distribution function is defined by

$$G(\eta) = \Pr(H \leq \eta) = \frac{\exp \eta}{1 + \exp \eta} = \text{Logistic}(\eta|0, 1),$$

for $-\infty < \eta < +\infty$. In the above formulation, G is specified to have a specific parametric (logistic) form. Instead, we can define a more flexible psychometric model by allowing G to be random and to take on any form. This can be achieved by modelling G nonparametrically as a Dirichlet process. See Newton, Czado, and Chappell (1996) for further details.

This nonparametric idea can easily be extended to the analysis of the item response data that take values in two or more categories, say, $K + 1$ of them, with $k = 0, \dots, K$.

In particular, we can adopt the idea that

$$\Pr(x_i = k | \mathbf{z}_i, \boldsymbol{\lambda}) = b(G(\boldsymbol{\lambda}^T \mathbf{z}_i), k, K)$$

for some chosen function b , for example

$$b(G, k, K) = \binom{K}{k} G^k (1 - G)^{K-k},$$

thus ensuring that

$$\sum_{k=0}^K b(G, k, K) = 1.$$

As in the binary case, a nonparametric model would have G as a Dirichlet process.

It is known that any psychometric model can be classified as either a continuation-ratio model, an adjacent-category model, or a cumulative probability model (Van der Ark, Hemker, & Sijtsma, 2002). As before, let G be a cumulative distribution function defined by $G(\eta) = \Pr(H \leq \eta)$, and also let $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_0, \dots, \boldsymbol{\lambda}_K)$ be a parameter vector and \mathbf{z} be a vector of covariates. The continuation-ratio model is defined by

$$\Pr(x_i = k | \mathbf{z}_i, \boldsymbol{\lambda}) = \left[\frac{1 - G(\boldsymbol{\lambda}_k^T \mathbf{z}_i)}{1 - G(\boldsymbol{\lambda}_{k-1}^T \mathbf{z}_i)} \right] \prod_{c=1}^k \left\{ \frac{1 - G(\boldsymbol{\lambda}_{c-1}^T \mathbf{z}_i)}{1 - G(\boldsymbol{\lambda}_{c-2}^T \mathbf{z}_i)} \right\}, \quad (1)$$

for ordinal categories $k = 0, \dots, K$, with $G(\boldsymbol{\lambda}_{-1}^T \mathbf{z}_i) \equiv 0$. The adjacent-category model is defined by:

$$\Pr(x_i = k | \mathbf{z}_i, \boldsymbol{\lambda}) = \frac{\prod_{j=0}^k G(\boldsymbol{\lambda}_j^T \mathbf{z}_i) \prod_{l=k+1}^K [1 - G(\boldsymbol{\lambda}_l^T \mathbf{z}_i)]}{\sum_{y=0}^K \left\{ \prod_{j=0}^y G(\boldsymbol{\lambda}_j^T \mathbf{z}_i) \prod_{l=y+1}^K [1 - G(\boldsymbol{\lambda}_l^T \mathbf{z}_i)] \right\}}, \quad (2)$$

for ordinal categories $k = 0, \dots, K$, where $G(\boldsymbol{\lambda}_0^T \mathbf{z}_i) \equiv 1$ and $G(\boldsymbol{\lambda}_{K+1}^T \mathbf{z}_i) \equiv 0$. Finally, the cumulative probability model is given by

$$\Pr(x_i = k | \boldsymbol{\lambda}, \mathbf{z}_i) = G(\boldsymbol{\lambda}_k^T \mathbf{z}_i) - G(\boldsymbol{\lambda}_{k-1}^T \mathbf{z}_i), \quad (3)$$

for $k = 1, \dots, K - 1$, with $\Pr(x_i = 0 | \mathbf{z}_i, \boldsymbol{\lambda}) = G(\boldsymbol{\lambda}_0^T \mathbf{z}_i)$ and $\Pr(x_i = K | \mathbf{z}_i, \boldsymbol{\lambda}) = 1 - G(\boldsymbol{\lambda}_{K-1}^T \mathbf{z}_i)$. It is known that with either equation (1), (2), or (3), parametric psychometric models are defined on the basis of G being a parametric distribution function, such as the Logistic(0,1) distribution or the Normal(0,1) distribution. To give just some examples of such parametric psychometric models, the family of Rasch models (Fischer & Molenaar, 1995) and the generalized partial credit model (Muraki, 1992) are adjacent-category models that fit the form (2), the graded response model (Samejima, 1969) is a cumulative probability model having the form (3), while the sequential model (Tutz, 1990) is a continuation-ratio model that fits the form (1).

Instead, using either the form (1), (2), or (3), we may specify G as a Dirichlet process in order to define a more flexible, nonparametric psychometric model. In principle, any nonparametric prior can be used as a model for G . Recently, an alternative mixture modelling approach was presented in a technical report by San Martín, Jara, Rolin, and Mouchart (2007), where the subvector of examinee parameters in $\boldsymbol{\lambda}$ is modelled by a nonparametric prior, with G a parametric function (e.g. Logistic(0,1)).

3. Coherent model selection

Bayesian statistical inference is a mathematical consequence of the axioms of quantitative coherence, in which the aim is to find the optimal decision that maximizes the posterior expected utility (e.g. Bernardo & Smith, 1994). Let

- $\mathcal{D} = \{d = 1, \dots, D\}$ denote the set of possible decisions in a decision problem,
- $\Omega_{\mathcal{X}}$, the set of all distribution functions on \mathcal{X} , define the ‘possible states of the world’,
- $\mathcal{D} \times \Omega_{\mathcal{X}} = \{(d, F) : d \in \mathcal{D}, F \in \Omega_{\mathcal{X}}\}$ denote the set of all possible pairings of a decision (d) with a possible states of the world (F).

In the Bayesian nonparametric approach to decision theory, the decision maker (e.g. statistician) specifies a utility function $u(M, F)$, over $\mathcal{D} \times \Omega_{\mathcal{X}}$, that measures his/her desirability of decision (i.e. model) $M \in \mathcal{D}$ given a possible ‘state of the world’ $F \in \Omega_{\mathcal{X}}$. Also, the posterior probability $\prod_n(dF)$ represents his/her degree of belief that F is the true state of the world (i.e. is the true distribution F_0), given the available evidence \mathbf{x}^n (a set of data).

We will consider the log utility function. Other utility functions are possible but we will use the log utility for its connection to information theory and the Kullback–Leibler divergence, which will become clear later. For models without covariates, the utility of selecting model M , with ‘best’ density $f(x|M)$, when the correct distribution is $F(dx)$, is given by

$$u(M, F) = \int \log f(x|M) F(dx). \quad (4)$$

We will explain what we mean by the ‘best’ density $f(x|M)$ for a family of densities shortly. This utility function provides an attractive measure of information that has desirable mathematical properties (Bernardo, 1979). In particular, if \hat{M} maximizes $u(M, F)$, then \hat{M} minimizes the Kullback–Leibler divergence between $F(\cdot)$ and $F(\cdot|M)$. Recall that $F(\cdot)$ is assigned the nonparametric prior and represents the Bayesian model, and $f(\cdot|M)$ is a parametric model for which the utility of selecting this parametric model when the correct model is F is to be calculated.

Extending the logarithmic utility to include covariate information is straightforward; we have

$$u(M, F) = \iint_{\mathcal{Z}\mathcal{X}} \log f(x|\mathbf{z}, M) F(dx|\mathbf{z}) w(d\mathbf{z}), \quad (5)$$

where, $w(d\mathbf{z})$ is a weight function that measures the relative importance of a covariate value \mathbf{z} . We can consider consider three cases of this logarithmic utility:

- (1) If the \mathbf{z} are randomly generated from some density $\pi(d\mathbf{z})$, then we would take $w \equiv \pi$.
- (2) When the data $\mathbf{x}^n = \{(x_i, \mathbf{z}_i)\}_{i=1}^n$ are composed of n covariate vectors \mathbf{z}_i , then we can take

$$w(d\mathbf{z}) = n^{-1} \sum_{i=1}^n \delta_{\mathbf{z}_i}(d\mathbf{z}),$$

where δ_{z_i} is the degenerate distribution that assigns probability 1 to the covariate value $\mathbf{z} = \mathbf{z}_i$. This weight function is appropriate when the covariates are part of a fixed design (i.e. when the covariates are fixed and known).

- (3) We can also consider the situation where we put the prior on densities/distributions on the joint space (x, \mathbf{z}) , so we have $F(dx|\mathbf{z})w(d\mathbf{z}) = F(dx, d\mathbf{z})$.

In any of these cases, the optimal decision $\hat{M} \in \mathcal{D}$ is defined as the action that maximizes the posterior expected utility:⁶

$$\hat{M} = \arg \max_{M \in \mathcal{D}} \bar{u}(M|\mathbf{x}^n),$$

where

$$\begin{aligned} \bar{u}(M|\mathbf{x}^n) &= \int_{\Omega_x} u(M, F) \Pi_n(dF) = \int_{\Omega_x} \left[\iint_{\underline{\mathcal{Z}}\mathcal{X}} \log f(x|\mathbf{z}, M) F(dx|\mathbf{z}) w(d\mathbf{z}) \right] \pi_n(dF) \\ &= \iint_{\underline{\mathcal{Z}}\mathcal{X}} \log f(x|\mathbf{z}, M) F_n(dx|\mathbf{z}) w(d\mathbf{z}). \end{aligned}$$

The optimal decision $\hat{M} \in \mathcal{D}$ also minimizes⁷ the Kullback–Leibler divergence (when it exists):

$$\hat{M} = \arg \min_{M \in \mathcal{D}} \int \log \left\{ \frac{f_n(x|\mathbf{z})}{f(x|\mathbf{z}, M)} \right\} F_n(dx|\mathbf{z}) w(d\mathbf{z})$$

which can be written as

$$\hat{M} = \arg \min_{M \in \mathcal{D}} \int D(F_n(\cdot|\mathbf{z}), F(\cdot|\mathbf{z}, M)) w(d\mathbf{z}).$$

Thus the optimal decision \hat{M} is associated with the family of sampling densities $f(\cdot|\mathbf{z}, \hat{M})$ which explains the most information about the (nonparametric) predictive density

$$f_n(\cdot|\mathbf{z}) = \int f(\cdot|\mathbf{z}) \Pi_n(d\mathbf{f}),$$

which is the Bayes estimate of the (unknown) true sampling density $f_0(\cdot|\mathbf{z})$.

We just need to know exactly what $f(\cdot|\mathbf{z}, M)$ is. Suppose that model M is parameterized by λ_M . For each M we find the parameter λ_M , the point estimate, which maximizes

$$\int \int \log f(x|\mathbf{z}, \lambda_M, M) F_n(dx|\mathbf{z}) w(d\mathbf{z})$$

⁶ The term $\arg \max_{M \in \mathcal{D}} \bar{u}(M|\mathbf{x}^n)$ means: ‘the value of M (from the set of possible values of M , denoted by \mathcal{D}) that has the highest value of the posterior expected utility $\bar{u}(M|\mathbf{x}^n)$ ’.

⁷ The following equation means: ‘the value of M (from the set of possible values of M , denoted by \mathcal{D}) that has the lowest value of the Kullback–Leibler divergence’.

or, equivalently, minimizes

$$\int D(F_n(\cdot|\mathbf{z}), F(\cdot|\mathbf{z}, \boldsymbol{\lambda}_M, M)) w(d\mathbf{z}).$$

Then $f(\cdot|\mathbf{z}, M) = f(\cdot|\mathbf{z}, \boldsymbol{\lambda}_M, M)$. In essence, we pick the model which possesses the distribution within it which is closest, in the Kullback–Leibler sense, to F_n .

For example, suppose we consider a prior on densities on the joint space (x, \mathbf{z}) (this is case 3 of the possibilities discussed above), and take the prior to be the non-informative Dirichlet process. Then the posterior for the distribution on (x, \mathbf{z}) is the Bayesian bootstrap. This gives

$$F_n(dx, d\mathbf{z}) = n^{-1} \sum_{i=1}^n \delta_{x_i, \mathbf{z}_i}(dx, d\mathbf{z}).$$

Consequently, we have

$$\bar{u}(M|\mathbf{x}^n) = n^{-1} \sum_{i=1}^n \log f(x_i|\mathbf{z}_i, M)$$

and so it is easy to see that, under the non-informative Dirichlet process prior, the point estimate $\boldsymbol{\lambda}_M$ is the maximum likelihood estimator. Thus, the optimal model M is the one which maximizes

$$n^{-1} \sum_{i=1}^n \log f(x_i|\mathbf{z}_i, \boldsymbol{\lambda}_M, M),$$

equal to n^{-1} times the log likelihood.

For example, let us reconsider the binary response model discussed in Section 2. Also, let N denote the number of examinees and J be the number of items in a test (e.g. examination, questionnaire), so that the total number of observations, n , is defined by $n = NJ$. Then, using the logistic model as the parametric model,

$$f(x_i|\mathbf{z}_i, \boldsymbol{\lambda}_M) = \frac{\exp\{\mathbf{1}(x_i = 1)(\boldsymbol{\lambda}_M^T \mathbf{z}_i)\}}{1 + \exp(\boldsymbol{\lambda}_M^T \mathbf{z}_i)},$$

with $\mathbf{1}(x_i = 1)$ the indicator function that equals 1 whenever $x_i = 1$, and equals 0 otherwise. Hence,

$$\bar{u}(M|\mathbf{x}^n) = n^{-1} \sum_{i=1}^n \{\mathbf{1}(x_i = 1)\boldsymbol{\lambda}_M^T \mathbf{z}_i + \log(1 + \exp(\boldsymbol{\lambda}_M^T \mathbf{z}_i))\}.$$

For the continuation-ratio model, with G defined by a parametric distribution, we have

$$\begin{aligned} f(x_i = k|\boldsymbol{\lambda}_M, \mathbf{z}_i) &= \Pr(x_i = k|\boldsymbol{\lambda}_M, \mathbf{z}_i) \\ &= \left[\frac{1 - G(\boldsymbol{\lambda}_{kM}^T \mathbf{z}_i)}{1 - G(\boldsymbol{\lambda}_{k-1, M}^T \mathbf{z}_i)} \right] \prod_{c=1}^k \left\{ \frac{1 - G(\boldsymbol{\lambda}_{c-1, M}^T \mathbf{z}_i)}{1 - G(\boldsymbol{\lambda}_{c-2, M}^T \mathbf{z}_i)} \right\}, \end{aligned}$$

for ordinal categories $k = 0, \dots, K$, with $G(\boldsymbol{\lambda}_{-1}^T \mathbf{z}_i) \equiv 0$ (thus, there is no λ_{-1} in the model). Hence, we define the posterior expected utility by

$$\bar{u}(M|\mathbf{x}^n) = n^{-1} \sum_{i=1}^n \sum_{k=0}^K \mathbf{1}(x_i = k) \log f(k|\boldsymbol{\lambda}_M, \mathbf{z}_i),$$

where $\mathbf{1}(x_i = k)$ is the indicator function that equals 1 when $x_i = k$, and equals 0 otherwise.

In general, we need to construct a nonparametric regression model, and evaluate the posterior expected utility $\bar{u}(M|\mathbf{x}^n)$. This should be feasible using Markov chain Monte Carlo methods which involve sampling from the predictive $F_n(dx|\mathbf{z})$; see Robert and Casella (2005). This is non-trivial, particularly when combined with the need to maximize over the parameter space. This is why we advocate the procedure of relying on the Bayesian bootstrap, which provides an easy-to-implement approach to model selection that requires no more than evaluating maximum likelihood estimators. Indeed, in this section, we have already shown that model selection under the ‘non-informative’ Dirichlet process prior leads to selecting the model with the highest log likelihood, and that this approach is coherent Bayesian practice. However, it is worth mentioning that in our model selection framework we can still consider more general priors than the non-informative Dirichlet process prior, and also we can generalize a logarithmic utility function such as (4) or (5) to include terms that penalize the dimension of the model (see Karabatsos & Walker, 2006).

4. A psychometric application of model selection

Here we provide a practical illustration for the Bayesian nonparametric approach to model selection, in the comparison of three well-known psychometric models for ordered item response data, namely, the Rasch rating scale model (Andrich, 1978), the Rasch partial credit model (Masters, 1982), and the graded response model (Samejima, 1969). Recall from Section 2.2 that these two Rasch models are adjacent-category models having the form defined in equation (2), and that the graded response model is a cumulative probability model having the form (3), and where in any of these three models, G is assumed to be a parametric cumulative distribution function. In typical psychometric applications of these models, G is assumed to be the Logistic(0,1) parametric distribution. Also, while these models do not have a special status within the Bayesian nonparametric approach to model selection, the same approach can be used to perform model selection on any number of the available psychometric models.

We proceed to describe these three models in more detail. Let N denote the number of examinees and J be the number of items in a test (e.g. examination, questionnaire), with the items scored in ordinal categories $k = 0, \dots, K$. Label the Rasch rating scale model as M_1 and the Rasch partial credit model as M_2 . These models are most clearly presented using Andersen’s (1995) representation. The Rasch rating scale model is defined by

$$f(x_{sj} = k|\boldsymbol{\lambda}_1, \mathbf{z}_{sj}, M_1) = \Pr(x_{sj} = k|\boldsymbol{\lambda}_1, \mathbf{z}_{sj}, M_1) = \frac{\exp(k(\theta_s + \beta_j) + \gamma_k)}{\sum_{l=0}^K \exp(l(\theta_s + \beta_j) + \gamma_l)},$$

for ordinal categories $k = 0, \dots, K$, with parameter vector

$$\boldsymbol{\lambda}_1 = (\theta_1, \dots, \theta_N, \beta_1, \dots, \beta_J, \gamma_0, \dots, \gamma_K),$$

where θ_s is the ability of subject s ($s = 1, \dots, N$), β_j is the parameter of item j ($j = 1, \dots, J$), and the γ_k ($k = 0, \dots, K$) are category parameters. The Rasch partial credit model is defined by

$$f(x_{sj} = k | \boldsymbol{\lambda}_2, \mathbf{z}_{sj}, M_2) = \Pr(x_{sj} = k | \boldsymbol{\lambda}_2, \mathbf{z}_{sj}, M_2) = \frac{\exp(k(\theta_s + \beta_j) + \gamma_k)}{\sum_{l=0}^K \exp(k(\theta_s + \beta_j) + \gamma_l)},$$

for $k = 0, \dots, K$, with parameter vector

$$\boldsymbol{\lambda}_2 = (\theta_1, \dots, \theta_N, \beta_{10}, \dots, \beta_{J0}, \dots, \beta_{1K}, \dots, \beta_{JK}),$$

where β_{jk} is the parameter of category k in item j (for all $j = 1, \dots, J$ and all $k = 1, \dots, K$). The key difference between these two Rasch models is that in the Rasch rating scale model the parametric spacing between ordinal categories is assumed to be the same for all J test items, whereas in the Rasch partial credit model the spacing is allowed to vary over the different test items. Finally, the graded response model, which we label as model M_3 , is defined by

$$\begin{aligned} f(x_{sj} = k | \boldsymbol{\lambda}_3, \mathbf{z}_{sj}, M_3) &= \Pr(x_{sj} = k | \boldsymbol{\lambda}_3, \mathbf{z}_{sj}, M_3) \\ &= \frac{\exp[\alpha_j(\theta_s - \lambda_{jk})]}{1 + \exp[\alpha_j(\theta_s - \lambda_{jk})]} - \frac{\exp[\alpha_j(\theta_s - \lambda_{j,k+1})]}{1 + \exp[\alpha_j(\theta_s - \lambda_{j,k+1})]}, \end{aligned}$$

for $k = 0, \dots, K$, and the parameter vector is defined by

$$\boldsymbol{\lambda}_3 = (\theta_1, \dots, \theta_N, \alpha_1, \dots, \alpha_J, \lambda_{10}, \dots, \lambda_{J0}, \dots, \lambda_{1K}, \dots, \lambda_{JK}),$$

with item slope parameters α_j , $j = 1, \dots, J$, and category location parameters satisfying

$$-\infty \equiv \lambda_{j0} < \lambda_{j1} \leq \dots \leq \lambda_{jK} < \lambda_{j,K+1} \equiv \infty, \quad j = 1, \dots, J.$$

Thus, in the graded response model, the category spacing and the slopes are allowed to vary over the different test items.

We applied each of the three models in the analysis of a data set collected in 2001 from students attending a Louisiana medical university. Before attending a course that teaches communication with terminally ill patients, 60 second-year medical students were asked to answer 11 questionnaire items concerning their level of comfort with communicating with terminally ill patients about their health (see Taylor, Hammond, DiCarlo, Karabatsos, & Deblieux, 2003). For each of these items, each student gave an ordinal response, and the possible responses for each questionnaire item were 0 = 'Not at all comfortable', 1 = 'A little comfortable', 2 = 'Somewhat comfortable', 3 = 'Quite comfortable', 4 = 'Very comfortable' and thus the five possible ordinal responses are indexed by $k = 0, \dots, 4$, with $K = 4$. The questionnaire items are presented in Table 1, along with the frequencies of the five ordinal categories in the data. This data set, denoted by

$$\mathbf{x}^n = \{(x_{sj}, \mathbf{z}_{sj}) : s = 1, \dots, N = 60, j = 1, \dots, J = 11\},$$

Table 1. The questionnaire items, along with the frequency of responses in the data

	Response count				
	0	1	2	3	4
How comfortable are you discussing the . . .					
1. Patient's hope for survival?	7	17	17	11	8
2. Patient's likely progression of illness before death?	9	11	27	10	3
3. Patient's chances for survival?	12	13	24	10	1
4. Patient's concerns about what the future holds?	5	14	23	9	9
5. Patient's experiences with the terminal condition?	5	6	21	20	8
6. Patient's pain issues?	2	3	20	20	15
7. Patient's worries about dying?	3	12	20	18	7
8. Patient's loss of functional independence?	1	16	16	19	8
9. Patient's fears about spreading their disease to other people?	1	7	15	22	15
10. Patient's concern about what the future holds?	2	9	24	16	9
11. Uncertainty of death with the patient?	3	14	22	13	8

with $x_{sj} \in \{0, \dots, 4\}$ for all students s and all items j , consists of a total of $n = NJ = 60 \cdot 11 = 660$ observations, from the response of $N = 60$ students on the $J = 11$ questionnaire items.

For each of the two Rasch models, maximum likelihood estimates of the item parameters were obtained by conditioning on the sufficient statistics $\sum_{j=1}^J x_{sj}$ ($s = 1, \dots, N$) of the models, and thus, the ability parameters $\theta_1, \dots, \theta_N$ are eliminated as 'nuisance' parameters (Andersen, 1973). This conditioning maintains the asymptotic consistency of the item parameter estimates. To identify the Rasch rating scale model, the constraint $\beta_1 = \tau_0 = \tau_1 \equiv 0$ is imposed, and this leaves the model with a total of $13 = J + K - 2$ parameters. To identify the Rasch partial credit model, the constraint $\beta_{10} = \dots = \beta_{j0} = \beta_{11} \equiv 0$ is imposed, leaving the model with a total of $43 = JK - 1$ parameters. For each of these two models, the conditional maximum likelihood estimates of the parameters were obtained through the eRm package of the R statistical software (Mair & Hatzinger, 2007). For the graded response model, estimates of the item parameters were obtained using the marginal maximum likelihood method (Bock & Aitkin, 1981), where the ability parameters $(\theta_1, \dots, \theta_N)$ are integrated out, with the assumption that they are samples from a Normal(0,1) population distribution of examinees. This approach to eliminating the examinee parameters also maintains the asymptotic consistency of the item parameter estimates. In this data set, with the ability parameters integrated out, the graded response model has a total of $55 = JK + 1$ parameters. The marginal maximum likelihood estimates of the graded response model were obtained through the ltm package of R (Rizopoulos, 2006).

For the model selection analysis, we assume the non-informative Dirichlet process prior that we discussed in Sections 2 and 3, and recall that this prior corresponds to the Bayesian bootstrap posterior distribution. Also, we specify that prior on distributions on the joint space (x, \mathbf{z}) (Case 3 of Section 3), so we have $F(dx|\mathbf{z})w(d\mathbf{z}) = F(dx, d\mathbf{z})$. Under these assumptions, and for the data set, the posterior expected utility for the Rasch rating scale model (denoted M_1), the Rasch partial credit model (denoted M_2), or the graded response model (denoted M_3) is defined by

$$\begin{aligned}\bar{u}(M_d|\mathbf{x}^n) &= \frac{1}{NJ} \sum_{s=1}^{N=60J=11} \sum_{j=1}^{K=4} \sum_{k=0} \mathbf{1}(x_{sj} = k) \log f(k|\boldsymbol{\lambda}_d, \mathbf{z}_{sj}) \\ &= \frac{1}{NJ} \text{LogLik}(\boldsymbol{\lambda}_d|\mathbf{x}^n)\end{aligned}\quad (6)$$

where $\boldsymbol{\lambda}_d$ denotes the parameter point estimate, for models $d = 1, 2, 3$, and recall that $\text{LogLik}(\boldsymbol{\lambda}_d|\mathbf{x}^n)$ denotes the log likelihood. This estimate may be a maximum likelihood estimate or, more generally, any type of point estimate. Recall that the total number of observations is $n = NJ = 60 \cdot 11 = 660$.

Interestingly, in cases where the non-informative Dirichlet process prior is assumed, the posterior expected utility $\bar{u}(M_d|\mathbf{x}^n)$ can easily be computed using any software for parametric IRT models that reports either the log likelihood statistic, $\text{LogLik}(\boldsymbol{\lambda}_d|\mathbf{x}^n)$, or the deviance statistic, $\text{Dev}(\boldsymbol{\lambda}_d|\mathbf{x}^n)$ which equals -2 times the log likelihood. In particular, $\bar{u}(M_d|\mathbf{x}^n) = \text{LogLik}(\boldsymbol{\lambda}_d|\mathbf{x}^n)/n$ and $\bar{u}(M_d|\mathbf{x}^n) = -\text{Dev}(\boldsymbol{\lambda}_d|\mathbf{x}^n)/2n$. Moreover, two popular criteria for model selection include the Akaike information criterion (AIC; Akaike, 1973), which is defined by

$$\text{AIC}_d = \text{Dev}(\boldsymbol{\lambda}_d|\mathbf{x}^n) + 2\text{dim}(\Theta_d) = -2 \cdot \text{LogLik}(\boldsymbol{\lambda}_d|\mathbf{x}^n) + 2\text{dim}(\Theta_d)$$

and the Bayesian information criterion (BIC; Schwarz, 1978), defined by

$$\text{BIC}_d = \text{Dev}(\boldsymbol{\lambda}_d|\mathbf{x}^n) + \log(N)\text{dim}(\Theta_d) = -2 \cdot \text{LogLik}(\boldsymbol{\lambda}_d|\mathbf{x}^n) + \log(N)\text{dim}(\Theta_d),$$

for a set of models M_d , $d = 1, \dots, D$, where $\text{dim}(\Theta_d)$ denotes the dimensionality of the parameter space Θ_d of model M_d . In using the AIC criterion, the aim is to select the model with the lowest AIC, and likewise for BIC. Notice that both AIC and BIC penalize according to model dimensionality. In fact, it has been shown (Karabatsos & Walker, 2006) that model selection under AIC (BIC) corresponds to model selection with the posterior expected utility under a non-informative Dirichlet process prior, such that the logarithmic utility is subtracted by a penalty $2\text{dim}(\Theta_d) (\log(N) \text{dim}(\Theta_d))$ that reflects the decision maker's preference for 'simpler', lower-dimensional models. For example, assuming the AIC penalty in the logarithmic utility function, and assuming a non-informative Dirichlet process prior, the posterior expected utility is defined by $\bar{u}(M_d|\mathbf{x}^n) = (1/n)\text{LogLik}(\boldsymbol{\lambda}_d|\mathbf{x}^n) - 2\text{dim}(\Theta_d)$, and this quantity is proportional to $-\text{AIC}_d$. The deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van der Linde, 2002) provides yet another popular criterion for model selection, and is a Bayesian version of the AIC. For a given model M_d , the DIC is defined by:

$$\text{DIC}_d = \text{Dev}(\bar{\boldsymbol{\lambda}}_d) + 2\{\overline{\text{Dev}}_d - \text{Dev}(\bar{\boldsymbol{\lambda}}_d|\mathbf{x}^n)\},$$

where

$$\overline{\text{Dev}}_d = \int \text{Dev}(\boldsymbol{\lambda}_d) p(\boldsymbol{\lambda}_d|\mathbf{x}^n) d\boldsymbol{\lambda}_d$$

is the posterior mean of the deviance, $\text{Dev}(\bar{\boldsymbol{\lambda}}_d|\mathbf{x}^n)$ denotes the deviance observed at the posterior mean $\bar{\boldsymbol{\lambda}}_d$ of the parameter, and the second term in the above equation is a penalty for model dimensionality. Interestingly, it can be easily shown that, under a non-informative Dirichlet process prior, the posterior expected utility $\bar{u}(M_d|\mathbf{x}^n)$ of our Bayesian nonparametric approach to model selection is obtained

by taking $\bar{u}(M_d|\mathbf{x}^n) = -\overline{\text{Dev}}_d/2n$. Thus, this posterior expected utility can be calculated using any software for Bayesian analysis that reports the $\overline{\text{Dev}}_d$ statistic of the DIC.

In Table 2, we compare three psychometric models, in terms of the posterior expected utility ($\bar{u}(M_d|\mathbf{x}^n)$) under the non-informative Dirichlet process prior (calculated from equation (6)), and in terms of the log likelihood, the deviance, AIC, and BIC. From the results of the posterior expected utility, we conclude that of the three models, the Rasch partial credit model explains the most information in the (unknown) true sampling distribution that generated the questionnaire data \mathbf{x}^n . Table 3 presents the maximum likelihood estimates of the item parameters in the partial credit model. Given the result of this model comparison in Table 2, we infer that there is evidence that the category spacing is different over the 11 questionnaire items. Also, while the graded response model is more flexible than the Rasch partial credit model in terms of its functional form (Van der Ark *et al.*, 2002), it seems that the assumption of a Normal(0,1) ability distribution, used in marginal maximum likelihood estimation, negatively impacts the fit of the model.

From Table 2, we also notice that selecting the model with the highest posterior expected utility (here, the Rasch partial credit model) corresponds to selecting the model that maximizes the log likelihood, and thus corresponds to selecting the model that minimizes the deviance. This is not surprising because, as we showed earlier, the posterior expected utility (under the non-informative Dirichlet process prior) is

Table 2. Comparing the psychometric models by posterior expected utility, log likelihood, deviance, AIC, and BIC

Model	$\bar{u}(M_d \mathbf{x}^n)$	Log likelihood	Deviance	$\dim(\Theta_d)$	AIC	BIC
Rasch partial credit	-0.72	-474.84	949.63	43	1,035.63	1,125.68
Rasch rating scale	-0.75	-494.93	989.86	13	1,015.86	1,043.08
Graded response	-1.07	-703.60	1,407.20	55	1,517.20	1,632.38

Table 3. Item parameter estimates of the Rasch partial credit model

Item _(j)	Item parameter estimate			
	$\hat{\beta}_{j1}$	$\hat{\beta}_{j2}$	$\hat{\beta}_{j3}$	$\hat{\beta}_{j4}$
1	0	-2.04	-5.74	-11.20
2	-0.72	-1.93	-6.68	-14.49
3	-1.08	-2.83	-7.88	-33.83
4	0.37	-1.01	-5.11	-10.12
5	-0.23	-0.43	-3.15	-8.86
6	0.66	1.60	-0.55	-4.74
7	1.02	-0.04	-3.02	-9.01
8	2.54	1.01	-1.61	-7.32
9	2.15	1.92	0.16	-4.09
10	1.38	0.96	-2.18	-7.51
11	1.10	-0.15	-3.71	-9.24

obtained by a simple transformation of the log likelihood, and it is obtained by a simple transformation of the deviance. Also, notice that according to either the AIC or the BIC, the Rasch rating scale model is preferred to the Rasch partial credit model. As mentioned, both the AIC and the BIC imply the use of a logarithmic utility that reflects the decision maker's preference for simpler models, via a term that penalizes the dimension of the model.

5. Conclusions

We have presented a coherent basis for parametric model selection via the use of nonparametric priors, which act as the Bayesian prior. The parametric models can then be compared using Bayesian decision theory. Also, in Section 2.2, we have presented a general way to define a very flexible psychometric model through the use of a nonparametric prior. In the task of model selection, we have demonstrated that under the non-informative Dirichlet process prior on the space of sampling densities, the posterior expected utility of any psychometric model is easily calculated, and can even be directly obtained from output from standard software that routinely reports the likelihood (or deviance) for each psychometric model. Also, while we have restricted our attention to psychometric models in this paper, our Bayesian nonparametric approach to model selection can be applied to any statistical model that defines a likelihood function. This includes both the class of generalized linear models and the class of multi-level regression models (e.g. Raudenbush & Bryk, 2003), which have many applications in the social sciences. We should point out that the psychometric models applied in Section 4 are multi-level ordinal regression models with random intercepts, such that item responses are nested within examinees, a random intercept represents a parameter of examinee ability, and the test items are fixed effects that correspond to dummy-coded covariates (Raudenbush & Bryk, 2003, p. 365).

Finally, in some situations involving psychometric analysis, only a single parametric psychometric model (M_1) is considered, and it is of interest to determine whether the model provides an adequate description of the (unknown) true sampling density that has generated a given set of data \mathbf{x}^n . Then, the decision problem is whether or not to reject the parametric model. One useful extension of our Bayesian nonparametric approach, for the evaluation of model adequacy, is to include the true model in the decision space, the model denoted by M_0 and represented by the chosen nonparametric prior. While the non-informative Dirichlet process provides one possible choice of prior for the problem of testing model adequacy, it may not be the best possible choice. It is an improper prior which gives full support to the observed data points, and in effect it does not support distributions defined over the entire sample space. Thus, for testing model adequacy, it seems advisable to select a proper prior that supports all densities defined on the entire sample space (see Walker, 2004). On the basis of this prior, it is possible in principle to use Markov chain Monte Carlo techniques to estimate the posterior expected utilities $\bar{u}(M_0|\mathbf{x}^n)$ and $\bar{u}(M_1|\mathbf{x}^n)$. Then one may proceed to reject the model whenever the positive difference $\bar{u}(M_0|\mathbf{x}^n) - \bar{u}(M_1|\mathbf{x}^n) \geq 0$ is large enough, and in fact this difference is equivalent to the Kullback-Leibler divergence $\int D(F_n(\cdot|\mathbf{z}, M_0), F(\cdot|\mathbf{z}, M_1)) w(d\mathbf{z})$ which measures the amount of information that the null model M_1 does not explain in the true model M_0 . One can then use any of the available approaches to calibrate the Kullback-Leibler divergence (e.g. McCulloch, 1989) to define a critical value for making the decision as to whether or not to reject the parametric model (M_1).

Acknowledgements

The authors thank the Editor and three anonymous referees for valuable suggestions which improved the paper. An earlier version was presented in July 2006 at the Bayesian Nonparametric Workshop in Jeju Island, South Korea.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov & F. Csáki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akadémiai Kiadó.
- Andersen, E. (1973). *Conditional inference and models for measuring* (pp. 271–292). Copenhagen: Mentalhygiejnisk Forlag.
- Andersen, E. (1995). Polytomous Rasch models and their estimation. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 357–374.
- Bernardo, J. (1979). Expected information as expected utility. *Annals of Statistics*, *7*, 686–690.
- Bernardo, J., & Smith, A. (1994). *Bayesian theory*. Chichester: Wiley.
- Bock, R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of items parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.
- Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*, 209–230.
- Ferguson, T. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, *2*, 615–629.
- Fischer, G., & Molenaar, I. (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer.
- Gelman, A., Carlin, A., Stern, H., & Rubin, D. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.
- Gutiérrez-Peña, E., & Walker, S. (2005). Statistical decision problems and Bayesian nonparametric methods. *International Statistical Review*, *73*, 309–330.
- Karabatsos, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology*, *50*, 123–148.
- Karabatsos, G., & Walker, S. (2006). On the normalized maximum likelihood and Bayesian decision theory. *Journal of Mathematical Psychology*, *50*, 517–520.
- Luce, R. (2005). Measurement analogies: Comparisons of behavioral and physical measures. *Psychometrika*, *70*, 227–251.
- Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for an application of IRT models in R. *Journal of Statistical Software*, *20*, 1–20.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–173.
- McCulloch, R. (1989). Local model influence. *Journal of the American Statistical Association*, *84*, 473–478.
- Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. New York: Cambridge University Press.
- Müller, P., & Quintana, F. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, *19*, 95–110.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–177.
- Newton, M., Czado, C., & Chappell, R. (1996). Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association*, *91*, 142–153.
- Raudenbush, S., & Bryk, A. (2003). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

- Rizopoulos, D. (2006). Ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17, 1-25.
- Robert, C., & Casella, G. (2005). *Monte Carlo Statistical Methods* (2nd ed.). New York: Springer.
- Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34 (Monograph 17), 100-114.
- San Martín, E., Jara, A., Rolin, J.-M., & Mouchart, M. (2007). On the analysis of Bayesian semiparametric IRT-type models. *Interuniversity Attraction Pole Technical Report 08029*, Institut de Statistique, Université Catholique de Louvain, Belgium.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Spiegelhalter, D., Best, N., Carlin, B., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64, 1-34.
- Taylor, L., Hammond, J., DiCarlo, R., Karabatsos, G., & Deblieux, P. (2003). A student-initiated elective on end-of-life care: A unique perspective. *Journal of Palliative Medicine*, 6, 86-92.
- Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43, 39-55.
- Van der Ark, L., Hemker, B., & Sijtsma, K. (2002). Hierarchically related nonparametric IRT models, and practical data analysis methods. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 41-62). Mahwah, NJ: Erlbaum.
- Walker, S. (2004). New approaches to Bayesian consistency. *Annals of Statistics*, 32, 2028-2043.
- Walker, S., Damien, P., Laud, P., & Smith, A. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society, Series B*, 61, 485-527.

Received 1 February 2007; revised version received 25 July 2007