DE GRUYTER Int. J. Biostat. 2015; 11(1): 51–67

Daniel Commenges*, Cécile Proust-Lima, Cécilia Samieri and Benoit Liquet

A Universal Approximate Cross-Validation Criterion for Regular Risk Functions

Abstract: Selection of estimators is an essential task in modeling. A general framework is that the estimators of a distribution are obtained by minimizing a function (the estimating function) and assessed using another function (the assessment function). A classical case is that both functions estimate an information risk (specifically cross-entropy); this corresponds to using maximum likelihood estimators and assessing them by Akaike information criterion (AIC). In more general cases, the assessment risk can be estimated by leave-one-out cross-validation. Since leave-one-out cross-validation is computationally very demanding, we propose in this paper a universal approximate cross-validation criterion under regularity conditions (UACVR). This criterion can be adapted to different types of estimators, including penalized likelihood and maximum a posteriori estimators, and also to different assessment risk functions, including information risk functions and continuous rank probability score (CRPS). UACVR reduces to Takeuchi information criterion (TIC) when cross-entropy is the risk for both estimation and assessment. We provide the asymptotic distributions of UACVR and of a difference of UACVR values for two estimators. We validate UACVR using simulations and provide an illustration on real data both in the psychometric context where estimators of the distributions of ordered categorical data derived from threshold models and models based on continuous approximations are compared.

Keywords: AIC, cross-entropy, cross-validation, estimator choice, Kullback–Leibler risk, model selection, ordered categorical observations, psychometric tests

DOI 10.1515/ijb-2015-0004

1 Introduction

Selecting estimators is an essential step in modeling, and Akaike information criterion (AIC) [1] has been widely used for this purpose. AIC allows selecting maximum likelihood estimators (MLE) based on parametric models that are not too badly specified. More general criteria have been developed, in particular the Takeuchi information criterion (TIC) [2] and the general information criterion (GIC) [3]. A related criterion in the field of neural networks is the network information criterion (NIC) [4]. Two other well-known criteria are the Bayesian information criterion (BIC) and the deviance information criterion (DIC); both use Bayesian arguments and are not directly related to the present paper. A good reference book for information criteria is by Konishi and Kitagawa [5].

***Corresponding author: Daniel Commenges,** INSERM, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux F-33000, France; ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Universite de Bordeaux, Bordeaux F-33000, France, E-mail: daniel.commenges@isped.u-bordeaux2.fr
Cécile Proust-Lima, Cécilia Samieri, INSERM, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux F-33000, France; ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Universite de Bordeaux, Bordeaux F-33000, France, E-mail: cecile.proust@isped.u-bordeaux2.fr, cecilia.samieri@isped.u-bordeaux2.fr
Benoit Liquet, INSERM, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux F-33000, France; ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Universite de Bordeaux, Bordeaux F-33000, France; School of Mathematics and Physics, The University of Queensland, St Lucia, Brisbane, Queensland 4066, Australia, E-mail: benoit.liquet@isped.u-bordeaux2.fr

Likelihood cross-validation (LCV) has also been widely used for comparing parametric models. Stone [6] heuristically established that LCV was asymptotically identical to AIC. LCV, however, is more flexible in that it can be applied to other estimators than MLEs, for instance, to penalized likelihood estimators: see Golub et al. [7] and Wahba [8].

Cross-validation can also be applied to other assessment risks than Kullback–Leibler risk. The leave-one-out cross-validation is the most natural and one of the most efficient [9, 10], but it is also the most computationally demanding so that approximation formulas have been derived. Approximate cross-validation formulas have been developed for penalized splines [11, 12] or penalized likelihood [13, 14]. Commenges et al. [15] derived an approximate cross-validation criterion in the context of prognosis.

In the present paper we consider the following general framework: estimators of the true density function are defined as minimizing an estimating function; the estimating function itself can be viewed as an estimator of a risk, that we call an "estimating risk." Typically there is a model, that is a family of densities for the variable Y, $(g^\theta)_{\theta\in\Theta}$, $\Theta \subset \Re^p$, and the estimator is chosen as minimizing the estimating risk. The estimators of the true density are then assessed using an "assessment risk," which allows choosing between different available estimators. The most conventional case is when the estimating risk is $E[-\log g^\theta(Y)]$ which is estimated by the log likelihood, and the assessment risk of the obtained estimator $g^{\hat\theta}$ is $E[-\log g^{\hat\theta}(Y)]$, which can be estimated by cross-validation or in the parametric case by the normalized AIC: AIC $/2n$. These information risks are very appealing but there are cases where other risks are relevant. As an example, the MLE could be assessed by the *continuous rank probability score* (CRPS) [16]: this is detailed in Section 4.4. Another example is the estimation of the distribution of ordinal data through an approximation using models for continuous data. Models for ordinal variables that can take a large number of values are rather cumbersome; it is convenient to treat these data as continuous, using an estimating risk adapted to continuous data. However, if we wish to compare the obtained estimator to that obtained by a model for ordinal data, the assessment risk must still take into account that the data are really ordinal. Such assessment risk can be estimated by cross-validation; cross-validation has good properties but is very computationally demanding. The main aim of this paper is to find an approximation for leave-one-out cross-validation, valid whatever the estimating and assessment risks satisfying regularity conditions that will be detailed. This will be applied to the ordinal data example.

Section 2 presents the framework, the cross-validation criterion and its approximation. It is universal in the sense that it can be applied to any estimating and assessment risks satisfying regularity conditions. We denote the approximate criterion by UACVR (U for Universal, A for approximate, CV for cross-validation and R for regularity). In Section 2 the asymptotic distributions of UACVR and of a difference of two UACVR values are given. Section 4 shows how UACVR specializes to particular cases: TIC appears as a special case when cross-entropy is used for defining both estimating and assessment risks, and AIC follows if the models are close to being well specified; other important cases where estimating and assessment risks defined in a less symmetric way are given. Section 5 presents a simulation study. Section 6 presents an illustration of the use of UACVR for comparing estimators derived from threshold models and estimators obtained by continuous approximations in the case of ordered categorical data with repeated measurements; these data are psychometric scores from a large study on cognitive aging. Section 7 concludes.

2 The universal cross-validation criterion and its approximation

2.1 The estimating risk and its estimation by an estimating function

Suppose that a sample of independently identically distributed (i.i.d.) variables $\mathcal{O}_n = (Y_i, i = 1, \ldots, n)$ is available. Based on $\mathcal{O}_n$, an estimator $g^{\hat\theta}$ (where $\hat\theta$ is short for $\hat\theta_n$) of the probability density function f^* of the true distribution can be chosen in a model, that is a family of distributions $(g^\theta)_{\theta\in\Theta}$, $\Theta \subset \Re^p$. The main rules

for designing estimators of θ can be thought of as minimizing an *estimating risk*. The estimating risk $\Phi(\theta)$ is defined as the expectation under the true distribution of a loss function $\phi(\theta, Y_i)$: $\Phi(\theta) = \mathrm{E}_*\{\phi(\theta, Y_i)\}$. We would like to choose g^{θ_0} where $\theta_0 = \mathrm{argmin}_\theta \Phi(\theta)$. For making consistent estimation possible, it is natural to require that whenever the model is well specified, the risk is minimized by the true distribution. Precisely, saying that the model is well-specified amounts to say that there is a value θ_*, such that $g^{\theta_*} = f^*$. Then we require that $\theta_* = \mathrm{argmin}_\theta \Phi(\theta)$; moreover we will require that this minimum is unique. This is related to the concept of strictly proper scores [16]. In the scoring rule literature, the problem is formulated in terms of reward rather than loss; it is possible to establish a correspondence between the two theories by considering that minus a loss is a reward, and of course while one tries to minimize the expected loss, one tries to maximize the expected reward.

We cannot compute the estimating risk but a natural estimator of the estimating risk is the estimating function $\Phi_{\mathcal{O}_n}(\theta) = n^{-1}\sum_{i=1}^n \phi(\theta, Y_i)$. The estimator $\hat{\theta}$ defined as minimizing $\Phi_{\mathcal{O}_n}(\theta)$ is called an M-estimator. By the law of large numbers, $\Phi_{\mathcal{O}_n}$ converges in probability toward $\Phi(\theta) = \mathrm{E}_*\{\phi(\theta, Y_i)\}$. Under some conditions given in Van der Vaart [17] (see, e.g. Theorem 5.7), $\hat{\theta}$ converges in probability toward θ_0. A simple set of sufficient conditions is that Θ is compact, $\Phi(\theta)$ is continuous and has a unique minimizer, $\phi(\theta, y)$ is continuous for every y.

Example 1: If we take as loss function $\phi(\theta, Y_i) = [Y_i - E_{g^\theta}(Y_i)]^2$, the estimating risk is $\Phi(\theta) = E_*[Y_i - E_{g^\theta}(Y_i)]^2$; the estimating function is $\Phi_{\mathcal{O}_n}(\theta) = n^{-1}\sum_{i=1}^n [Y_i - E_{g^\theta}(Y_i)]^2$ and $\hat{\theta}$ is the least-square estimator.

Example 2: If we take as loss function $\phi(\theta, Y_i) = -\log g^\theta(Y_i)$, the estimating risk is $\Phi(\theta) = E_*[-\log g^\theta(Y_i)]$ which is the cross-entropy of g^θ with respect to f^*; the estimating function is $\Phi_{\mathcal{O}_n}(\theta) = -n^{-1}\sum_{i=1}^n \log g^\theta(Y_i)$ and $\hat{\theta}$ is the MLE.

2.2 The assessment risk and its estimation by cross-validation

When several estimators are available, we wish to assess their performance by estimating an *assessment risk*. Estimators with small assessment risks will be preferred. For constructing the risk of an estimator $g^{\hat{\theta}}$ we may use a loss function $\psi(g^{\hat{\theta}}, Y)$. The assessment risk is the expectation under f^* of $\psi(g^{\hat{\theta}}, Y)$, where both Y and $g^{\hat{\theta}}$ are random:

$$\Psi(g^{\hat{\theta}}) = E_*\{\psi(g^{\hat{\theta}}, Y)\}. \tag{1}$$

The problem is to estimate the assessment risk (without knowing the true density f^*). A natural, albeit naive, estimator is

$$\Psi_{\mathcal{O}_n}(g^{\hat{\theta}}) = n^{-1}\sum_{i=1}^n \psi(g^{\hat{\theta}}, Y_i). \tag{2}$$

However $\Psi_{\mathcal{O}_n}(g^{\hat{\theta}})$ is not completely satisfying because it does not take into account that $g^{\hat{\theta}}$ depends on the observations; as a result $\Psi_{\mathcal{O}_n}(g^{\hat{\theta}})$ underestimates $\Psi(g^{\hat{\theta}})$ (the well-known overoptimism bias).

If another sample $\mathcal{O}'_n = (Y'_i, i = 1, \ldots, n)$ i.i.d. with respect to $\mathcal{O}_n$ were available, a natural estimator of the assessment risk would be $\Psi_{\mathcal{O}'_n}(g^{\hat{\theta}}) = n^{-1}\sum_{i=1}^n \psi(g^{\hat{\theta}}, Y'_i)$. We call $\Psi_{\mathcal{O}'_n}(g^{\hat{\theta}})$ the "oracle estimator." This is an unbiased estimator of the assessment risk but cannot be computed based on $\mathcal{O}_n$. Its variance is $\mathrm{var}_*\Psi_{\mathcal{O}'_n}(g^{\hat{\theta}}) = n^{-1}\mathrm{var}_*\{\psi(g^{\hat{\theta}}, Y'_i)|\hat{\theta}\}$, which tends toward $n^{-1}\kappa_*^2$, where κ_*^2 is the variance of $\psi(g^{\theta_0}, Y'_i)$.

A pseudo-oracle estimator of the assessment risk is often used by practitioners who split their original sample in a training and a validation sample. However, this practice leads to a loss of efficiency since only half of the data is used for computing the estimator $g^{\hat{\theta}}$ and half of the data also for estimating its assessment risk. Cross-validation estimators of the assessment risk make a more efficient use of the information. In particular the leave-one-out cross-validation criterion is

$$\mathrm{CV}(g^{\hat{\theta}}) = n^{-1}\sum_{i=1}^{n}\psi(g^{\hat{\theta}_{-i}}, Y_i),$$

where $\hat{\theta}_{-i} = \mathrm{argmin}\ \Phi_{\mathcal{O}_{n|i}}$ and $\Phi_{\mathcal{O}_{n|i}} = \frac{1}{n-1}\sum_{j\neq i}^{n}\phi(\theta, Y_j)$. $\mathrm{CV}(g^{\hat{\theta}})$ does nearly as well as if another sample $\mathcal{O}'_n$ were available, in terms of both bias and variance. Indeed it can immediately be seen that $E\{\mathrm{CV}(g^{\hat{\theta}})\} = \Psi(g^{\hat{\theta}_{n-1}})$. We shall see in Section 3 that the asymptotic variance of the approximate cross-validation criterion $\mathrm{UACVR}(g^{\hat{\theta}})$ is precisely $n^{-1}\kappa_*^2$, the same as that of the oracle estimator.

For comparing two estimators, the difference of assessment risks is relevant. This can be estimated by the difference of cross-validation estimates of the assessment risks.

2.3 The universal approximate cross-validation criterion

The leave-one-out cross-validation criterion may be computationally demanding since it is necessary to run the maximization algorithm n times for finding the $\hat{\theta}_{-i}, i = 1, \ldots, n$. For this reason an approximate formula is very useful. In this section we propose a universal approximation of the cross-validation (UACVR) criterion for regular loss functions ϕ and ψ.

Definition 1 (Universal approximation of the cross-validation)

$$\mathrm{UACVR}(g^{\hat{\theta}}) = \Psi_{\mathcal{O}_n}(g^{\hat{\theta}}) + \mathrm{Trace}(H^{-1}_{\Phi_{\mathcal{O}_n}}K), \tag{3}$$

where $H_{\Phi_{\mathcal{O}_n}} = \frac{\partial^2\Phi_{\mathcal{O}_n}}{\partial\theta^2}|_{\hat{\theta}}$ and $K = n^{-1}\sum_{i=1}^{n}\hat{v}_i\hat{d}_i^T$, with

$$\hat{v}_i = \frac{\partial\psi(g^{\theta}, Y_i)}{\partial\theta}|_{\hat{\theta}}$$

and

$$\hat{d}_i = \frac{1}{n-1}\frac{\partial\phi(\theta, Y_i)}{\partial\theta}|_{\hat{\theta}}.$$

The leading term in eq. (3) is the naive estimator of $\Psi(g^{\hat{\theta}})$ defined in eq. (2) while the second term is a correction accounting for parameter estimation. This correction term involves $H_{\Phi_{\mathcal{O}_n}}$, the Hessian of the estimating function, and $\hat{v}_i$ and $\hat{d}_i$ which are the gradients of the assessment and estimating functions (up to the multiplicative constant $1/(n-1)$ for the latter).

Under regularity assumptions on $\phi(.,.)$ and $\psi(.,.)$, we have that the leave-one-out cross-validation criterion differs from UACVR by an asymptotically negligible term in $o_p(n^{-1})$, which makes UACVR a good approximation for n relatively large, when leave-one-out cross-validation becomes computationally too demanding. The regularity conditions are detailed in the Appendix and are essentially: A1: $\Phi(\theta)$ has a unique maximizer; A2: thrice differentiability of $\phi(\theta, y)$; A3: twice differentiability of $\psi(\theta, y)$.

Theorem 1 *Under assumptions A1, A2, A3, we have*

$$\mathrm{CV}(g^{\hat{\theta}}) = \Psi_{\mathcal{O}_n}(g^{\hat{\theta}}) + \mathrm{Trace}(H^{-1}_{\Phi_{\mathcal{O}_n}}K) + o_p(n^{-1}), \tag{4}$$

UACVR applies only to regular parametric problems. Thus it does not apply to non- or semi-parametric estimators and more generally to singular problems as treated by Watanabe [18]. Also, some assessment functions do not satisfy the regularity assumptions: for instance, a non-parametric estimator of the area under the ROC curve can be used for assessing the discriminating ability of an estimator, and this is not continuous in the parameter θ. Nevertheless, UACVR may be useful in various important contexts as detailed in Section 4, including penalized likelihood estimators approximated on a spline basis, which is a way to avoid strong parametric assumptions.

3 Asymptotic distribution and tracking interval

3.1 Asymptotic distribution of UACVR

Commenges et al. [19] using results of Vuong [20] studied the asymptotic distribution of a difference of normalized AIC's as an estimator of a difference of Kullback–Leibler risks: the normalized AIC is defined as $\frac{1}{2n}$AIC. Here similar arguments are applied to study the asymptotic distribution of UACVR and a difference of two UACVR values. By the continuous mapping theorem, the asymptotic distribution of $\mathrm{UACVR}(g^{\hat{\theta}})$ is the same as that of $\Psi(g^{\theta_0})$. Since the latter quantity is a mean, it immediately follows by the central limit theorem that

$$n^{1/2}\{\mathrm{UACVR}(g^{\hat{\theta}}) - \Psi(g^{\theta_0})\} \xrightarrow{D} N(0, \kappa_*^2), \tag{5}$$

where $\kappa_*^2 = \mathrm{var}_*\psi(g^{\theta_0}, Y)$ and var_* stands for the variance under the true distribution. We can also write:

$$n^{1/2}\{\mathrm{UACVR}(g^{\hat{\theta}}) - \Psi(g^{\hat{\theta}})\} \xrightarrow{D} N(0, \kappa_*^2), \tag{6}$$

and κ_*^2 can be estimated by the empirical variance of $\psi(g^{\hat{\theta}}, Y_i)$, $i = 1, \ldots, n$.

3.2 Asymptotic distribution of a difference between UACVR values of two estimators

If two estimators $g^{\hat{\theta}}$ and $h^{\hat{\gamma}}$ are available, we would like to know which is the best according to the chosen assessment risk. Thus, we have to estimate the difference of their assessment risks: $\Delta^{\psi}(g^{\hat{\theta}}, h^{\hat{\gamma}}) = \Psi(g^{\hat{\theta}}) - \Psi(h^{\hat{\gamma}})$. The obvious estimator is: $D_{\mathrm{UACVR}}(g^{\hat{\theta}}, h^{\hat{\gamma}}) = \mathrm{UACVR}(g^{\hat{\theta}}) - \mathrm{UACVR}(h^{\hat{\gamma}})$. We focus on the case where $g^{\theta_0} \neq h^{\gamma_0}$. We obtain in that case using the same arguments as above:

$$n^{1/2}\{D_{\mathrm{UACVR}}(g^{\hat{\theta}_n}, h^{\hat{\gamma}_n}) - \Delta(g^{\hat{\theta}_n}, h^{\hat{\gamma}_n})\} \xrightarrow{D} N(0, \omega_*^2), \tag{7}$$

where $\omega_*^2 = \mathrm{var}_*\{\psi(g^{\theta_0}, Y) - \psi(h^{\gamma_0}, Y)\}$, and this can be estimated by the empirical variance of $\left\{\psi(g^{\hat{\theta}}, Y_i) - \psi(h^{\hat{\gamma}}, Y_i)\right\}$.

Based on the same type of results, Commenges et al. [19] proposed to construct a "tracking interval" for a difference of normalized AIC values. The tracking interval is a kind of confidence interval for the difference of risks. Because the variability of estimators of difference of risks is rather large in general, it is useful to have an interval estimate rather than just a point estimate. However, in the conventional theory of point and interval estimation, the target parameter is fixed; here, it changes with n. Thus, we have a moving target: hence the name of tracking interval. Some simulations in Commenges et al. [19] showed that the variance of the difference of AIC was correctly estimated and the corresponding tracking interval had good coverage properties. The same idea can be applied in the more general case treated here. The tracking interval is given by (A_n, B_n), where $A_n = D_{\mathrm{UACVR}}(g^{\hat{\theta}_n}, h^{\hat{\gamma}_n}) - z_{\alpha/2}n^{-1/2}\hat{\omega}_n$ and $B_n = D_{\mathrm{UACVR}}(g^{\hat{\theta}_n}, h^{\hat{\gamma}_n}) + z_{\alpha/2}n^{-1/2}\hat{\omega}_n$, where z_u is the u^{th} quantile of the standard normal variable.

Note that ω_* is in general much lower than κ_*. This has been shown by Commenges et al. [13] for the expected cross-entropy assessment risk and comes from the fact that $\psi(g^{\hat{\theta}}, Y_i)$ and $\psi(h^{\hat{\gamma}}, Y_i)$ are often positively correlated.

4 Particular cases of UACVR

In this section we give seven frameworks in which UACVR applies (a non-exhaustive list).

4.1 MLEs and information assessment risk: TIC and AIC

Suppose we take: $\phi(\theta, Y_i) = \psi(g^\theta, Y_i) = -\log g^\theta(Y_i)$. Then, the estimating function is minus the log-likelihood. It estimates the estimating risk, here the cross-entropy [21] of g^θ with respect to the true density f^*: $\mathrm{E}_*\{-\log g^\theta(Y)\} = H(f^*) + \mathrm{KL}(g^\theta; f^*)$, where $H(f^*) = -E_*\{\log f^*(Y)\}$ is the entropy of f^* and $\mathrm{KL}(g^\theta; f^*) = E_*\left\{\log \frac{f^*(Y)}{g^\theta(Y)}\right\}$ the Kullback–Leibler divergence of g^θ relative to f^*. The assessment risk is here the expected cross-entropy:

$$\mathrm{ECE}(g^{\hat{\theta}}) = E_*[E_*\{-\log g^{\hat{\theta}}(Y)|\mathcal{O}_n\}] = H(f^*) + \mathrm{EKL}(g^{\hat{\theta}}; f^*), \tag{8}$$

where $\mathrm{EKL}(g^{\hat{\theta}}; f^*) = E_*\left\{\log \frac{f^*(Y)}{g^{\hat{\theta}}(Y)}\right\}$ is the expected Kullback–Leibler risk. It differs from the conventional Kullback–Leibler risk defined for a fixed density because it is applied here to an estimator: it was mentioned by Hall [22] under the name of "expected Kullback–Leibler loss." So, although the loss functions for estimating and assessment are the same, there is a dissymmetry in that the estimating risk is a cross-entropy while, because $g^{\hat{\theta}}$ is random, the assessment risk is an *expected* cross-entropy.

In that case the leading term of eq. (3) is minus the maximized (normalized) log-likelihood. We have also that $\hat{v}_i$ is the individual score and $\hat{d}_i = \frac{1}{n-1}\hat{v}_i$ so that UACVR is identical to a normalized version of TIC [5]. If the model is well specified K tends in probability toward $I(\theta_0)$. The Hessian $H_{\Phi_{\mathcal{O}_n}}$ also tends toward $I(\theta_0)$ so that the correction term tends toward p, the number of parameters. Thus, if the model is not too badly specified, TIC is approximately equal to AIC. We have $\mathrm{UACVR} = \frac{1}{2n}\mathrm{TIC} \approx \frac{1}{2n}\mathrm{AIC}$, and this estimates the *expected* cross-entropy of the estimator, $\mathrm{ECE}(g^{\hat{\theta}})$. In practice, Burnham and Anderson [23] do not recommend the use of TIC if n is small because of the variability of the correction term. On the other hand, Konishi and Kitagawa [5] show (see their Table 3.3) that the correction terms can be rather different when the models are misspecified.

4.2 M-estimators and information assessment risk: GIC

Konishi and Kitagawa [3] have generalized TIC and AIC to cases where $g^{\hat{\theta}}$ was an M-estimator. The criterion they proposed, obtained by correcting the bias of the log-likelihood, is the GIC. GIC is also a special case of UACVR, obtained when the assessment risk is the expected cross-entropy. They apply GIC in particular to penalized likelihood estimators. Thus UACVR, as GIC, can be applied to maximum a posteriori, maximum penalized likelihood and hierarchical likelihood estimators.

4.3 Restricted AIC

Liquet and Commenges [24] have proposed a modification of AIC and LCV when estimators are based on the full information while they are assessed on a smaller (more targeted) information. More specifically, the estimator is based on the sample $\mathcal{O}_n = (Y_i, i = 1, \ldots, n)$ but the assessment risk is based on a random variable Z which is a coarsened version of Y. For instance Z is a dichotomization of Y: $Z = 1_{Y>l}$. For this case, the restricted AIC (RAIC) was derived by both direct approximation of the risk and by approximation of the LCV. RAIC is a particular case of UACVR for the case: $\phi(\theta, Y_i) = -\log g^\theta(Y_i)$ and $\psi(g^\theta(Y_i)) = -\log g^\theta(Z_i)$.

4.4 Estimators assessment by CRPS

Gneiting and Raftery [16] studied scoring rules and particularly the CRPS. Its inverse that can be used as a loss function is defined as

$$\mathrm{CRPS}^*(G(.,\theta),Y)=\int_{-\infty}^{+\infty}\{G(u,\theta)-1_{u\geq Y}\}^2du,$$

where $G(.,\theta)$ is the cumulative distribution function (c.d.f.) of a distribution in the model. The risk is a Cramer–von Mises-type distance: $d(G,G^*)=\int\{G(u)-G^*(u)\}^2du$. In some cases, it may be interesting to assess MLE's using this assessment risk rather than the logarithmic loss which may be too sensitive to low values of the density. UACVR can be used for estimating this risk. In that case, the leading term of UACVR is $n^{-1}\sum_{i=1}^n\mathrm{CRPS}^*(G(.,\hat\theta),Y_i)$; for the correcting term, $H_{\Phi_{\mathcal{O}_n}}$ is the Hessian of the log-likelihood (since $\hat\theta$ is the MLE) and K must be computed with $\hat v_i=\frac{\partial\psi}{\partial\theta}|_{\hat\theta}=2\int_{-\infty}^{+\infty}\{G(u,\hat\theta)-1_{u\geq Y_i}\}\frac{\partial G(u,\theta)}{\partial\theta}|_{\hat\theta}du$; $\hat d_i$ is the individual score (gradient of the individual log-likelihood) divided by $n-1$. Thus the computation of $\hat v_i$, for each i, involves the computation of p simple integrals, which can be done numerically.

4.5 Estimators assessment by Brier score

Brier score [25] can be used to assess estimators of the distribution of categorical variables, say Y, taking values $1,\ldots,m$. Consider a model for this distribution: we write $g^\theta(j)=P(Y=j)$. Brier score is defined as $\sum_{j=1}^m(\delta_{Y,j}-g^\theta(j))^2$, where δ is the Kronecker symbol ($\delta_{Y,j}=1$ if $Y=j$, zero otherwise). Assume that we estimate θ by maximum likelihood and use the Brier score for assessment. In this case, the leading term of UACVR is $n^{-1}\sum_{i=1}^n\sum_{j=1}^m(\delta_{Y_i,j}-g^\theta(j))^2$; for the correcting term, $H_{\Phi_{\mathcal{O}_n}}$ is the Hessian of the log-likelihood (since $\hat\theta$ is the MLE) and K must be computed with $\hat v_i=\frac{\partial\psi}{\partial\theta}|_{\hat\theta}=-2\frac{\partial g^\theta}{\partial\theta}|_{\hat\theta}(Y_i)+2\sum_{j=1}^m g^{\hat\theta}(j)\frac{\partial g^\theta}{\partial\theta}|_{\hat\theta}(j)$; $\hat d_i$ is the individual score (gradient of the individual log-likelihood) divided by $n-1$.

4.6 Conditional AIC

A referee suggested that UACVR might be useful for selecting random effect models based on conditional assessment functions, that is when the target is the density conditional on random effects. Conditional Akaike criterion was proposed by Vaida and Blanchard [26]; Greven and Kneib [27] proposed a correction taking into account uncertainty on the covariance matrix of the random effects; Braun et al. [28] proposed a predictive cross-validation criterion. UACVR could directly apply to this case by considering that the assessment loss is $-\log g^\theta(Y|\hat b)$, where b is the random effect and $\hat b$ its estimator. Since $\hat b$ is a function of θ and Y, the assessment loss can indeed be written $\psi(\theta,y)$. For computing UACVR, the main task would be here to compute the gradient $\frac{\partial\psi(\theta,Y_i)}{\partial\theta}$, not forgetting the dependence of $\hat b$ on θ. This could be easily done by numerical differentiation.

4.7 Estimators based on continuous approximation of categorical data

Assume Y is an ordered categorical variable taking values $l=0,1,\ldots,L$. Here for simplicity we consider that Y is univariate. Several models are available for this type of variables. Cumulative probit models, further called "threshold link models," assume that $Y_i=l$ if a latent variable Λ_i takes values in the interval (c_l,c_{l+1}) for $l=0,\ldots,L$, with $c_0=-\infty$ and $c_{L+1}=+\infty$:

$$Y_i=\sum_{l=0}^L 1_{\{\Lambda_i\in(c_l,c_{l+1})\}}l. \tag{9}$$

Λ_i itself can be modeled as a noisy linear form of explanatory variables $\Lambda_i=\beta x_i+\varepsilon_i$, where ε_i has a normal distribution of mean zero and variance σ^2, and where x_i are explanatory variables. The parameters are

$\theta = (c_1, \ldots, c_L, \beta, \sigma)$. For identifiability one must add some constraints, for instance $\sigma = 1$ and null intercept in the linear model for Λ_i. An estimator of the distribution can be obtained by maximum likelihood leading to define $g^{\hat{\theta}}$. The assessment risk can be ECE($g^{\hat{\theta}}$). Note that since Y is discrete, the densities are defined with respect to a counting measure that is, $g^{\hat{\theta}}(l)$ defines the probability that $Y = l$.

One may also make a continuous approximation which leads to simpler computations and may be more parsimonious, especially if Y is multivariate as in the illustration of Section 6. For example we can consider the model $Y_i = \beta x_i + \varepsilon_i$. Maximizing the likelihood of this model for observations of Y_i leads to a probability measure specified by the density $h_c^{\hat{\gamma}}$. This is however a density relative to Lebesgue measure. This probability measure gives zero probabilities to $\{Y_i = l\}$ for all l, and this yields infinite value for ECE (meaning strong rejection of this estimator). However from h_c a natural estimator of f^* can be constructed by gathering at l the mass around l: $h^{\hat{\gamma}}(l) = \int_{l-1/2}^{l+1/2} h_c^{\hat{\gamma}}(u)du$, for $l = 1, \ldots, L-1$, and $h^{\hat{\gamma}}(0) = \int_{-\infty}^{1/2} h_c^{\hat{\gamma}}(u)du$, $h^{\hat{\gamma}}(L) = \int_{L-1/2}^{+\infty} h_c^{\hat{\gamma}}(u)du$. UACVR can be computed for this estimator for estimating its ECE. The leading term of UACVR($h^{\hat{\gamma}}$) can be interpreted as the log-likelihood obtained by this estimator with respect to the counting measure. For the correcting term we need the Hessian of the log-likelihood of $h_c^{\hat{\gamma}}$ and we have to compute $\hat{v}_i = \frac{\partial \psi(h^\gamma, Y_i)}{\partial \gamma}|_{\hat{\theta}}$. For instance if $Y_i = l$ for $l = 1 \ldots, L-1$ we have

$$\hat{v}_i = -\frac{\int_{l-1/2}^{l+1/2} \frac{\partial h_c^{\hat{\gamma}}}{\partial \gamma}(u)du}{\int_{l-1/2}^{l+1/2} h_c^{\hat{\gamma}}(u)du}.$$

Since the denominator is the probability under $h_c^{\hat{\gamma}}$ that $Y \in (l-1/2, l+1/2)$, $\hat{v}_i$ can be interpreted as the conditional expectation (under $h_c^{\hat{\gamma}}$) of the individual score. Thus if $h_c^{\hat{\gamma}}$ does not vary much on $(l-1/2, l+1/2)$, $\hat{v}_i$ is close to $-(n-1)\hat{d}_i$. Using the same arguments as in Section 4.1 we obtain that UACVR is close to correcting by the number of parameters as in AIC; such a criterion that we call AIC$_d$ was proposed by Proust-Lima et al. [29], and this is likely to be a good approximation if the number of modalities of Y is large.

5 Simulation: choice of estimators for ordered categorical data

5.1 Design

We conducted a simulation study to illustrate the use of UACVR for comparing estimators derived from threshold link models and estimators obtained by a linear continuous approximation in the case of ordered categorical data (see Section 4.7). The aim was to assess the performance of UACVR as an estimator of ECE defined in eq. (8), and to compare it to the normalized naive AIC criterion (noted AIC) and the normalized AIC criterion computed on the counting measure (noted AIC$_d$). Performances of these criteria were studied in the case where the number of modalities ($L+1$) of the response variable Y is small (Section 5.2.1) and when it is large (Section 5.2.2).

5.1.1 True distributions

For all the simulations, the data came from a cumulative probit model where the relationship between Y_i and Λ_i is as in eq. (9) and the linear form of Λ_i is specified by

$$\Lambda_i = \beta_1 X_i^1 + \beta_2 X_i^2 + \varepsilon_i; \quad i = 1, \ldots, n, \tag{10}$$

where ε_i and the two explanatory variables X_i^1 and X_i^2 were generated from independent standard normal distributions. In order not to disadvantage the linear continuous approximation compared to the threshold link model, the parameters $c_1, \ldots, c_L$ were chosen as the solution of the following equations:

$$\begin{cases} P(\Lambda_i < c_1) = P(\Lambda_i > c_L) \\ P(\Lambda_i < c_1) = P(c_1 < \Lambda_i < c_2), \\ c_{i+1} = c_i + m \ \text{ with } \ m = (c_L - c_1)/(L-1) \end{cases}$$

5.1.2 The different models

For each generated sample, we fitted the cumulative probit model as previously defined, and a linear model assuming a linear continuous approximation of the response variable Y, $Y_i = \gamma_0 + \gamma_1 X_i^1 + \gamma_2 X_i^2 + \varepsilon_i$, with ε_i being independent zero mean normal variables with variance τ^2. Both models were fitted by maximum likelihood using a Fortran program which was checked to be correct by comparing the results with those obtained by the R package lcmm [30].

Samples of $300,\ 500,\ 3,000$ subjects were generated. For all simulations, $N = 10{,}000$ samples were generated. The true assessment risk, ECE, which is available only in a simulation study, was computed by a Monte Carlo approach: for each sample $\mathcal{O}_n^j$ we computed $g^{\hat{\theta}(j)}$; we generated a large number $M = 100{,}000$ observations Y_k independent of $\mathcal{O}_n^j, j = 1, \ldots, N$; we estimated ECE by the global mean $\frac{1}{NM}\sum_{j=1}^{N}\sum_{k=1}^{M} -\log g^{\hat{\theta}(j)}(Y_k)$.

5.2 Results of the simulation

5.2.1 Small number of modalities

We consider here the case where the number of modalities of Y is relatively small ($L+1=5$). In this simulation, we fixed $\beta_1 = -1.05$, $\beta_2 = -1.85$. In Table 1, we present, for different sample sizes n, the results for the different empirical criteria AIC, AIC_d and UACVR which can be compared with ECE. For any sample size, the cumulative probit model provided a better ECE than the linear model (positive difference). It

Table 1: Performance of the criteria for a small number of modalities ($L + 1 = 5$) and different sample sizes.

	ECE	UACV	AIC_d	AIC	Bias UACV	Bias AIC_d	Bias AIC
$n = 300$							
Linear	1.266	1.266	1.270	1.658	0.0003	0.0044	0.3919
Threshold	0.986	0.983	0.984	0.984	−0.0029	−0.0028	−0.0028
Difference	0.279	0.283	0.287	0.674	0.0032	0.0072	0.3947
Agreement ECE		100%	100%	100%			
$n = 500$							
Linear	1.264	1.263	1.266	1.652	−0.0007	0.0017	0.3876
Threshold	0.981	0.981	0.981	0.981	−0.0002	−0.0002	−0.0002
Difference	0.283	0.282	0.285	0.671	−0.0006	0.0019	0.3878
Agreement ECE		100%	100%	100%			
$n = 1{,}000$							
Linear	1.262	1.262	1.262	1.649	−0.0001	0.0003	0.3875
Threshold	0.975	0.975	0.975	0.9751	0.0000	0.0000	0.0000
Difference	0.287	0.287	0.287	0.674	−0.0001	0.0003	0.3874
Agreement ECE		100%	100%	100%			

Note: Mean over 1,000 replications of the criteria UACVR, AIC_d, AIC. ECE is the true risk; the biases of the criteria as estimator of ECE are given, as well as the percentage of agreement with ECE for model choice.

appeared that UACVR had a very small bias for all the sample sizes (of order 10^{-3}). The two other criteria AIC and AIC_d were also in favor of a threshold model. However, as expected, the naive normalized AIC did not correctly estimate ECE due to the wrong probability measure (Lebesgue measure instead of a counting measure). We note that the criterion AIC_d estimated ECE relatively well, with a small bias around 10^{-2} and 10^{-3}. All the criteria were in agreement with ECE for the choice of the model.

5.2.2 Large number of modalities

We consider here the case where the number of modalities of Y is relatively large ($L+1=20$). In this simulation, we fixed $\beta_1 = -0.15$, $\beta_2 = -0.85$. The results of this simulation are presented in Table 2. For any sample size, the linear model provided a better ECE than the threshold model (negative difference). It appeared that UACVR had a small bias for all the sample sizes (of order 10^{-3} and 10^{-4}). The AIC_d criterion gave similar results as the UACVR criterion while the AIC criterion failed to find the best estimator (positive difference).

Table 2: Performance of the criteria for a large number of modalities ($L + 1 = 20$) and different sample sizes.

	ECE	UACV	AIC_d	AIC	BIAS.UACV	BIAS.AIC_d	BIAS.AIC
$n = 300$							
Linear	2.678	2.678	2.678	2.737	−0.0003	0.0001	0.0595
Threshold	2.709	2.705	2.705	2.705	−0.0036	−0.0037	−0.0037
Difference	−0.031	−0.027	−0.027	0.0325	0.0033	0.0038	0.0632
Agreement ECE		99.0%	99.0%	3.7%			
$n = 500$							
Linear	2.6752	2.6752	2.6754	2.7347	−0.0001	0.0002	0.0595
Threshold	2.6922	2.6911	2.6910	2.6910	−0.0012	−0.0012	−0.0012
Difference	−0.0170	−0.0159	−0.0156	0.0437	0.0011	0.0014	0.0607
Agreement ECE		99.2%	99.1%	0%			
$n = 1{,}000$							
Linear	2.672	2.672	2.672	2.731	−0.0000	−0.0000	0.0596
Threshold	2.673	2.673	2.673	2.673	−0.0001	−0.0001	−0.0001
Difference	−0.0009	−0.0009	−0.0009	0.0588	0.0000	0.0000	0.0597
Agreement ECE		81.2%	80.6%	0%			

Note: Mean over 1,000 replications of the criteria UACVR, AIC_d, AIC. ECE is the true risk; the biases of the criteria as estimator of ECE are given, as well as the percentage of agreement with ECE for model choice.

5.2.3 Coverage of tracking intervals

Finally we looked at the coverage of the tracking intervals and the percentage of cases where 0 was inside of the tracking interval. The results are given in Table 3. The coverage rates appear to be too large. We checked that the distributions of UACVR were approximately normal. We found however that the estimated standard deviations were too large by a factor varying from 1.2 to 1.8 for small and large number of modalities respectively, but we were unable to find the reason of this overestimation. Nevertheless, the estimate gives the order of magnitude of the variability of UACVR.

For small number of modalities, 0 was always outside of the tracking interval, leading to an unequivocal choice. For large number of modalities, the percentage increased with n. This may seem paradoxical but illustrates well the difference between a tracking interval and a confidence interval. What happens is that the misspecification risk of the linear model is rather large for small number of modalities and is very

Table 3: Performance of the 95% tracking interval in both situations ($L+1=5$ and $L+1=20$) and for the different sample sizes ($n=300,\ 500$ and 3,000).

	$L+1=5$		$L+1=20$	
	ECE ∈ TI	0∈ TI	ECE ∈ TI	0∈ TI
$n=300$	98.3%	0%	99.6%	51.1%
$n=500$	98.0%	0%	99.7%	55.7%
$n=3{,}000$	98.2%	0%	100%	99.4%

Note: Number of times that the tracking interval (TI) includes the true value ECE over 1,000 replications. Percentages of tracking intervals (TI) including the value 0.

small for large number of modalities. Thus the global risk is driven by the statistical risk. The latter decreases with n, so that the difference of risks, which is the target, decreases with n, becoming very small for $n=3,000$; in this case the two models are nearly equivalent and there is no point to choose one rather than the other according to the chosen risk.

6 Illustration on the choice of estimators for psychometric tests

In epidemiological studies, cognition is measured by psychometric tests which usually consist in the sum of items measuring one or several cognitive domains. A common example is the Mini-Mental State Examination (MMSE) score [31], computed as the sum of 30 binary items evaluating memory, calculation, orientation in space and time, language, and word recognition; for this reason it is called a "sumscore" and ranges from 0 to 30. Although in essence psychometric tests are ordered categorical data, they are most often analyzed as continuous data. Indeed, they usually have a large number of different levels and, especially in longitudinal studies, models for categorical data are numerically complex. Recently, Proust-Lima et al. [29] defined a latent process mixed model to analyze repeated measures of discrete outcomes involving either a threshold link model or an approximation of it using continuous parameterized increasing functions. Comparison of models assuming either categorical data (using the threshold model) or continuous data (using continuous functions) was done with an AIC_d, computed with respect to the counting measure. In this illustration, we use UACVR to compare such latent process mixed models assuming either continuous or ordered categorical data when applied on the repeated measures of the MMSE and its calculation subscore in a large sample from a French prospective cohort study.

6.1 Latent process mixed models

In brief, the latent process mixed model assumes that a latent process $(\Lambda_i^*(t))_{t\geq 0}$ underlies the repeated measures of the observed variable Y_{ij} for subject i ($i=1,\ldots,n$) and occasion j ($j=1,\ldots,n_i$). The latent process $\Lambda_i^*(t)$ is defined as a standard linear mixed model: $\Lambda_i^*(t)=X_i(t)^T\beta+Z_i(t)^Tb_i$ for $t\geq 0$ where $X_i(t)$ and $Z_i(t)$ are distinct vectors of time-dependent covariates associated, respectively, with the vector of fixed effects β and the vector of random effects b_i ($b_i\sim\mathcal{N}(\mu,D)$). We further assume that b_{i0}, the first component of b_i that usually represents the random intercept, is $\mathcal{N}(0,1)$ for identifiability; except for the variance of b_{i0}, D is an unstructured variance matrix.

A measurement model links the latent process with the observed repeated measures through intermediary variables which are noisy versions of the latent process at time t_{ij}: $\Lambda_{ij}=\Lambda_i^*(t_{ij})+\varepsilon_{ij}$, where the ε_{ij}'s are i.i.d. normal variables with zero expectation. For ordered categorical data, a standard threshold link model as defined in eq. (9) (Section 4.7) for the univariate case is well adapted, leading to a cumulative probit mixed model. For continuous data, the link has been modeled as $H(Y_{ij};\eta)=\Lambda_{ij}$ where $H(.;\eta)$ is a monotonic

increasing transformation. Three families of such transformations are considered: (i) $H(y;\eta) = \frac{h(y;\eta_1,\eta_2)-\eta_3}{\eta_4}$ where $h(.;\eta_1,\eta_2)$ is the beta c.d.f. with parameters (η_1,η_2); (ii) $H(y;\eta) = \eta_1 + \sum_{l=2}^{m+2} \eta_l B_l^I(y)$ where $(B_l^I)_{l=2,m+2}$ is a basis of quadratic I-splines with m nodes; (iii) $H(y;\eta) = \frac{y-\eta_1}{\eta_2}$ which gives the standard linear mixed model.

Latent process mixed models are estimated within the maximum likelihood framework using the lcmm function of lcmm R package [30]. When assuming continuous data, the likelihood can be computed analytically using the Jacobian of H [32]. In contrast, when assuming ordered categorical data, an integration over the random effects has to be done numerically [29].

UACVR is computed from the log-likelihood $\Psi_{\mathcal{O}_n}$ obtained for the MLEs $\hat{\theta}$ with respect to the counting measure:

$$\begin{aligned}\Psi_{\mathcal{O}_n}(\hat{\theta}) &= -n^{-1}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\prod_{j=1}^{n_i} P(Y_{ij}|b_i)f_b(b_i)db_i \\ &= -n^{-1}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\prod_{j=1}^{n_i}\prod_{l=0}^{L}\left(P(Y_{ij}=l|b_i)\right)^{1_{\{Y_{ij}=l\}}} f_b(b_i)db_i \\ &= -n^{-1}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\prod_{j=1}^{n_i}\prod_{l=0}^{L}\left(P(c_l \le \Lambda_{ij} < c_{l+1}|b_i)\right)^{1_{\{Y_{ij}=l\}}} f_b(b_i)db_i,\end{aligned} \tag{11}$$

where $c_0 = -\infty$, $c_{L+1} = +\infty$, and either c_l $(l = 1,\ldots,L)$ are the estimated thresholds when a threshold model is considered, or $c_l = H(l-\frac{1}{2},\hat{\eta})$ $(l = 1,\ldots,L)$ when monotonic increasing families of transformations are used. We also need to compute $\hat{v}_i$ similarly as in Section 4.7. The integral is approximated by Gaussian quadrature.

6.2 Application: categorical psychometric tests

Data come from the French prospective cohort study PAQUID initiated in 1988 to study normal and pathological aging [33]. Subjects included in the cohort were 65 and older at initial visit and were followed up to 10 times with a visit at 1, 3, 5, 8, 10, 13, 15, 17 and 20 years after the initial visit. At each visit, a battery of psychometric tests including the MMSE was completed. In the present analysis, all the subjects free of dementia at the 1-year visit and who had at least one MMSE measure during the whole follow-up were included: this resulted in a sample size of 2,914 subjects. Data from baseline were removed to avoid modeling the first-passing effect. The observed distributions of the MMSE sumscore and of its calculation subscore are displayed in Figure 1.

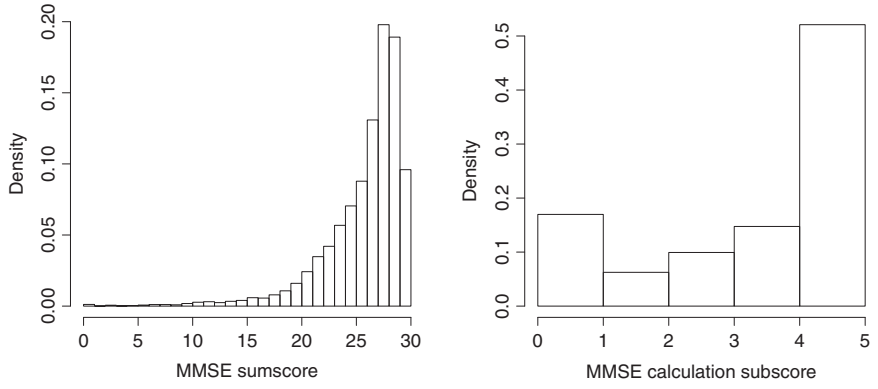

Figure 1: Distributions of MMSE sumscore and MMSE calculation subscore in the PAQUID sample ($n = 2{,}914$). Data were pooled from all available visits for a total of 10,846 observations.

The trajectory of the latent process was modeled as an individual quadratic function of age with correlated random effects for intercept, slope and quadratic slope ($Z_i(t)^{\mathrm{T}} = (1, \mathrm{age}_i(t), \mathrm{age}_i^2(t))$), and an adjustment for binary covariates educational level (EL $=$ 1 if the subject graduated from primary school) and gender (SEX $=$ 1 if the subject is a man) plus their interactions with age and quadratic age (so that $X_i(t)^{\mathrm{T}} = Z_i(t)^{\mathrm{T}} \otimes (1, \mathrm{EL}_i, \mathrm{SEX}_i)$). For MMSE sumscore, in addition to the threshold link, the linear, beta c.d.f. and I-splines (with five equidistant nodes) continuous link functions were considered. For calculation subscore, in addition to the threshold link, only the linear link was considered.

6.3 Results

Table 4 gives the assessment criteria for estimators based on the different models, and Table 5 provides the differences in UACVR or AIC_d and their 95% tracking interval. For the MMSE sumscore, the mixed model assuming the standard linear transformation yielded a clearly worse UACVR than other models accounting for nonlinear relationships with the underlying latent process. The model involving a beta c.d.f. gave a similar risk as the one involving the less parsimonious I-splines transformation ($D_{\mathrm{UACVR}} = -0.0070$ and 0 in the 95% tracking interval). Finally, the mixed model considering a threshold link model, which is numerically demanding (because of a three-dimensional integral in the likelihood), gave the best assessment risk but remained relatively close to the simpler ones assuming a beta c.d.f. ($D_{\mathrm{UACVR}} = 0.0200$) or a I-splines

Table 4: Number of parameters (p), naive normalized AIC (AIC), AIC_d, and UACVR for latent process mixed models involving different transformations H and applied on either the MMSE sumscore or its calculation subscore.

Transformation H	p	AIC	AIC_d	UACVR
MMSE				
Linear	16	8.752	8.529	8.536
Beta c.d.f.†	18	7.758	7.786	7.786
I-splines ‡	21	7.832	7.793	7.793
Thresholds	44	7.762	7.762	7.766
Calculation				
Linear	16	6.011	4.821	4.820
Thresholds	19	4.368	4.368	4.369

Note: † c.d.f. for cumulative distribution function.
‡ Quadratic I-splines with five equidistant nodes located at 0, 7.5, 15, 22.5 and 30.

Table 5: Difference of AIC_d (D_{AIC_d}), difference of two UACVR values(D_{UACVR}) and its 95% tracking interval between latent process mixed models involving different transformations H_1 and H_2, and applied on either the MMSE sumscore or its calculation subscore.

Transformations H_1/H_2	D_{AIC_d}	D_{UACVR}	95% tracking interval
MMSE			
Linear/Beta c.d.f.†	0.7421	0.7495	[0.6619; 0.8372]
Linear/I-splines ‡	0.7357	0.7425	[0.6526; 0.8325]
Beta c.d.f.†/I-splines ‡	−0.0064	−0.0070	[−0.0152; 0.0012]
I-splines ‡/thresholds	0.0306	0.0270	[0.0166; 0.0374]
Beta c.d.f.†/thresholds	0.0241	0.0200	[0.0097; 0.0303]
Linear/thresholds	0.7662	0.7696	[0.6784; 0.8607]
Calculation			
Linear/thresholds	0.4515	0.4523	[0.4127; 0.4919]

Note: † c.d.f. for cumulative distribution function.
‡ Quadratic I-splines with five equidistant nodes located at 0, 7.5, 15, 22.5 and 30.

transformation ($D_{\mathrm{UACVR}} = 0.0270$). For the interpretation of these values Commenges et al. [19] suggested to qualify values of order 10^{-1}, 10^{-2} and 10^{-3} as "large," "moderate" and "small," respectively; moreover for multivariate observations, it was suggested to divide by the total number of observations rather by the number of independent observations. With this correction (which amounts to divide the current values by a factor of $3.7 = 10,846/2,914$) the differences between the linear model and the other models can be qualified as "large," and the differences between the threshold model and both beta c.d.f. and I-splines are between "moderate" and "small." Of course, this gives only an idea of the difference of risks between estimators; a more intuitive and reliable interpretation scale is still to be found. Figure 2 displays the estimated link functions in (A) and the predicted mean trajectories of the latent process according to educational level in (B) from the models involving either a linear, a beta c.d.f., I-splines or a threshold link function. The estimated link functions as well as the predicted trajectories of the latent process are very close when assuming either beta c.d.f., I-splines or a threshold link function but they greatly differ when assuming a linear link.

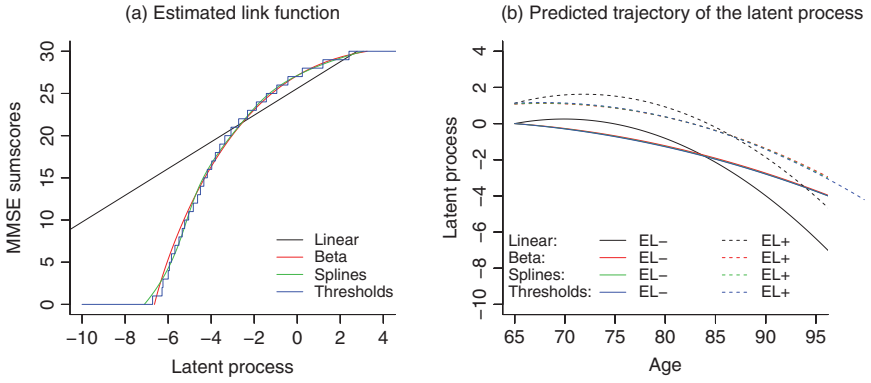

Figure 2: (A) Estimated inverse link functions between MMSE sumscores and the underlying latent process and (B) predicted trajectories of the latent process of a woman according to educational level (with EL+ and EL− for, respectively, validated or non-validated primary school diploma) in latent process mixed models assuming either linear, beta c.d.f., I-splines or threshold link functions (PAQUID sample, $n = 2{,}914$); the trajectories for the latter three transformations are indistinguishable.

For the calculation subscore also, the standard linear mixed model again gave a clearly higher risk than the mixed model assuming a threshold link model ($D_{\mathrm{UACVR}}(\mathrm{linear}, \mathrm{thresholds}) = 0.452$, 95% tracking interval: $[0.413, 0.492]$).

7 Conclusion

We have proposed a universal approximate formula for leave-one-out cross-validation under regularity conditions: it is universal in the sense that it applies to any couple of estimating and assessment risks which can be correctly estimated from the observations. UACVR is often a very good approximation of leave-one-out cross-validation which itself does nearly as well as an "oracle estimator" of the assessment risk which would be computable if we assessed the estimator on an independent replica of the sample. Another asset is that UACVR does not need the assumption that the models are well specified, and non-nested models can be compared. The result is in principle restricted to parametric models but extends to smooth semi- or non-parametric ones through spline representation of penalized likelihood estimators. The approximate formula not only allows fast computation, because the model is fitted only once, but also allows deriving the asymptotic distribution.

Estimating this distribution is important since the variability of UACVR, as that of any criterion used for estimator choice, may be large. Hopefully, as noted in Section 3, the variability of a difference of UACVR values between two estimators is smaller, but still remains non-negligible. A simple formula allows to estimate these variances and to construct so-called tracking intervals; our simulation study however shows that the coverage of these tracking intervals is too large, due to an overestimation of the variances. This is an open question to find why this happened here while in other contexts [15, 19] the coverage rates were correct, and possibly to find a correction to this overestimation; nevertheless, the estimates get the correct order of magnitude and the tracking intervals may be useful.

In this paper, UACVR has been applied to the issue of choice between estimators of the distribution of longitudinal categorical data based on cumulative probit mixed models or on mixed models based on a continuous approximation. It has been shown that the naive AIC can be misleading while a procedure called AIC_d (which had not been validated yet) yields results very close to UACVR, even if the latter is slightly better. Both quantities can be computed in the lcmm R package.

Appendix: Proof of Theorem 1

Under Assumptions A1–A3 below, we have formula (4).

In the proof, we apply the O_p concept to vectors and matrices. Saying that a matrix H is $O_p(1)$ means that all its elements are $O_p(1)$. The proof is partly heuristic in that we need at the end an assumption for obtaining that a mean of n $O_p(n^{-2})$ remainder terms is itself an $O_p(n^{-2})$ or at least an $o_p(n^{-1})$ term.

We assume:

A1 θ_0 is the unique minimizer of $\Phi(\theta)$ and the M-estimator $\hat{\theta}$ is consistent for θ_0.

A2 $\phi(\theta, y)$ is thrice differentiable for every y and the third derivative is dominated by a fixed function in a neighborhood of θ_0.

A3 $\psi(\theta, y)$ is twice differentiable for every y and the second derivative is dominated by a fixed function in a neighborhood of θ_0.

The proof is as follows. Assumption A2 is the essential assumption in the so-called classical conditions [17] for obtaining that $\sqrt{n}(\hat{\theta} - \theta_0)$ has an asymptotic normal distribution. It implies that $\hat{\theta}_{-i} - \hat{\theta} = O_p(n^{-1/2})$. A Taylor expansion of $\frac{\partial \Phi_{\mathcal{O}_{n|i}}}{\partial \theta}|_{\hat{\theta}_{-i}}$ around $\hat{\theta}$ yields

$$0 = \frac{\partial \Phi_{\mathcal{O}_{n|i}}}{\partial \theta}|_{\hat{\theta}} + H_{\Phi_{\mathcal{O}_{n|i}}}(\hat{\theta}_{-i} - \hat{\theta}) + R_n^1,$$

where $H_{\Phi_{\mathcal{O}_{n|i}}} = \frac{\partial^2 \Phi_{\mathcal{O}_{n|i}}}{\partial \theta^2}|_{\hat{\theta}}$ and R_n^1 is a quadratic form of $\hat{\theta}_{-i} - \hat{\theta}$ involving third derivatives of $\Phi_{\mathcal{O}_{n|i}}$ taken in $\tilde{\theta}$ so that $||\tilde{\theta}_n - \hat{\theta}|| \leq ||\hat{\theta}_{-i} - \hat{\theta}||$. Thus $||\tilde{\theta}_n - \hat{\theta}||$ is also an $O_p(n^{-1/2})$. Under Assumption A2 and using Lemma 2.12 of Van der Vaart [17], R_n^1 is an $O_p(n^{-1})$. Assumptions A1 and A2 imply that $I(\theta) = \frac{\partial^2 \Phi}{\partial \theta^2}|_{\theta}$ exists and is invertible in a neighborhood of θ_0. By the strong law of large numbers, $H_{\Phi_{\mathcal{O}_n}} = \frac{\partial^2 \Phi_{\mathcal{O}_n}}{\partial \theta^2}|_{\hat{\theta}}$ and $H_{\Phi_{\mathcal{O}_{n|i}}} = \frac{\partial^2 \Phi_{\mathcal{O}_{n|i}}}{\partial \theta^2}|_{\hat{\theta}}$ converge toward $I(\theta_0)$ and thus are invertible for sufficiently large n. It also follows that both these matrices and their inverses are $O_p(1)$. Thus, from the above development we obtain

$$\hat{\theta}_{-i} - \hat{\theta} = -H^{-1}_{\Phi_{\mathcal{O}_{n|i}}} \frac{\partial \Phi_{\mathcal{O}_{n|i}}}{\partial \theta}|_{\hat{\theta}} + R_n,$$

where $R_n = -H^{-1}_{\Phi_{\mathcal{O}_{n|i}}} R_n^1$ is an $O_p(n^{-1})$.

By definition of $\Phi_{\mathcal{O}_n}(\theta)$ we have the relation

$$n\Phi_{\mathcal{O}_n}(\theta) = (n-1)\Phi_{\mathcal{O}_{n|i}}(\theta) + \phi(\theta, Y_i). \tag{12}$$

Taking derivatives of the terms of this equation and taking the values at $\hat{\theta}$ we find $0 = (n-1)\frac{\partial \Phi_{\mathcal{O}_{n|i}}}{\partial \theta}|_{\hat{\theta}} + \frac{\partial \phi(\theta, Y_i)}{\partial \theta}|_{\hat{\theta}}$ and we obtain that $\frac{\partial \Phi_{\mathcal{O}_{n|i}}}{\partial \theta}|_{\hat{\theta}} = -\hat{d}_i$

Hence we have

$$\hat{\theta}_{-i} - \hat{\theta} = H^{-1}_{\Phi_{\mathcal{O}_{n|i}}} \hat{d}_i + R_n, \tag{13}$$

Note that this implies that $\hat{\theta}_{-i} - \hat{\theta} = O_p(n^{-1})$ because $H_{\Phi_{\mathcal{O}_{n|i}}} = O_p(1)$ and both $\hat{d}_i$ and R_n are $O_p(n^{-1})$. But this in turn implies that R_n is in fact an $O_p(n^{-2})$ (as a quadratic form of $O_p(n^{-1})$ terms). Now we show that $H_{\Phi_{\mathcal{O}_{n|i}}}$ can be replaced by $H_{\Phi_{\mathcal{O}_n}} = \frac{\partial^2 \Phi_{\mathcal{O}_n}}{\partial \theta^2}|_{\hat{\theta}}$ in eq. (13). By twice derivating eq. (12) we obtain $H_{\Phi_{\mathcal{O}_n}} = \frac{n-1}{n} H_{\Phi_{\mathcal{O}_{n|i}}} + \frac{1}{n} H_{\phi_i}$ where $H_{\phi_i} = \frac{\partial^2 \phi(\theta, Y_i)}{\partial \theta^2}|_{\hat{\theta}}$; since the last term is an $O_p(n^{-1})$, we can write $H_{\Phi_{\mathcal{O}_{n|i}}} = H_{\Phi_{\mathcal{O}_n}} + O_p(n^{-1})$. Equation (13) can be written $H_{\Phi_{\mathcal{O}_{n|i}}}(\hat{\theta}_{-i} - \hat{\theta}) = \hat{d}_i + O_p(n^{-2})$ or replacing $H_{\Phi_{\mathcal{O}_{n|i}}}$ by $H_{\Phi_{\mathcal{O}_n}} + O_p(n^{-1})$, $H_{\Phi_{\mathcal{O}_n}}(\hat{\theta}_{-i} - \hat{\theta}) = \hat{d}_i + O_p(n^{-1})(\hat{\theta}_{-i} - \hat{\theta}) + O_p(n^{-2})$. Using the fact that $\hat{\theta}_{-i} - \hat{\theta} = O_p(n^{-1})$ we obtain

$$\hat{\theta}_{-i} - \hat{\theta} = H^{-1}_{\Phi_{\mathcal{O}_n}} \hat{d}_i + O_p(n^{-2}). \tag{14}$$

Developing now the assessment loss function for $\hat{\theta}_{-i}$ around $\hat{\theta}$ yields (using Assumption A3):

$$\psi(g^{\hat{\theta}_{-i}}, Y_i) = \psi(g^{\hat{\theta}}, Y_i) + (\hat{\theta}_{-i} - \hat{\theta})^T \hat{v}_i + O_p(n^{-2}).$$

Replacing in this equation $\hat{\theta}_{-i} - \hat{\theta}$ by its approximation in eq. (14) we obtain $\psi(g^{\hat{\theta}_{-i}}, Y_i) = \psi(g^{\hat{\theta}}, Y_i) + \hat{d}_i^T H^{-1}_{\Phi_{\mathcal{O}_n}} \hat{v}_i + O_p(n^{-2})$. Taking the mean of the left-hand terms of these equations yields $\mathrm{CV}(g^{\hat{\theta}})$. Taking the mean of the terms on the right-hand side gives us a development with an error term which is the mean of n error terms in $O_p(n^{-2})$. Because the number of error terms to consider increases with n, it is not true in general that such a mean preserves the order of the error terms. This is true assuming some boundedness conditions of the expectations of these terms. At this stage the proof is heuristic: we assume conditions such that the mean of these $O_p(n^{-2})$ terms is also an $O_p(n^{-2})$, or at least $o_p(n^{-1})$. When this holds, we obtain the announced result given in formula (4).

References

1. Akaike H. Information theory and an extension of the maximum likelihood principle. In BN Petrov and F Csâki, editors. Proc. of the 2nd Int. symp. on information theory, Budapest: Akademiai Kiâdo, 1973: 267–81.
2. Takeuchi K. Distributions of information statistics and criteria for adequacy of models. Math Sci 1976;153:12–18.
3. Konishi S, Kitagawa G. Generalised information criteria in model selection. Biometrika 1996;83:875–90.
4. Murata N, Yoshizawa S, Amari S-I. Network information criterion-determining the number of hidden units for an artificial neural network model. Neural Networks IEEE Trans 1994;5:865–72.
5. Konishi S, Kitagawa G. Information criteria and statistical modeling. New York: Springer Series in Statistics, 2008.
6. Stone M. Cross-validatory choice and assessment of statistical predictions (with discussion). J R Stat Soc B 1974;39:111–47.
7. Golub G, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 1979;21:215–23.
8. Wahba G. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Ann Stat 1985;13:1378–402.
9. Van Der Laan M, Dudoit S, Keles S. Asymptotic optimality of likelihood-based cross-validation. Stat Appl Genet Mol Biol 2004;3:1036.
10. Xu G, Huang JZ. Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. Ann Stat 2012;40:3003–30.
11. Gu C, Xiang D. Cross-validating non-Gaussian data. J Comput Graphical Stat 2001;10:581–91.
12. Xiang D, Wahba G. A generalized approximate cross validation for smoothing splines with non-Gaussian data. Stat Sin 1996;6:675–92.
13. Commenges D, Joly P, Gegout-Petit A, Liquet B. Choice between semi-parametric estimators of Markov and non-Markov multi-state models from generally coarsened observations. Scand J Stat 2007;34:33–52.
14. O'Sullivan F. A statistical perspective on ill-posed inverse problems. Stat Sci 1986;1:502–18.

15. Commenges D, Liquet B, Proust-Lima C. Choice of prognostic estimators in joint models by estimating differences of expected conditional Kullback-Leibler risks. Biometrics 2012;68:380–7.
16. Gneiting T, Raftery A. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 2007;102:359–78.
17. Van der Vaart A. Asymptotic statistics. Cambridge: Cambridge University Press, 2000.
18. Watanabe S. Algebraic geometry and statistical learning theory. Vol. 25. Cambridge: Cambridge University Press, 2009.
19. Commenges D, Sayyareh A, Letenneur L, Guedj J, Bar-Hen A. Estimating a difference of Kullback-Leibler risks using a normalized difference of AIC. Ann Appl Stat 2008;2:1123–42.
20. Vuong Q. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 1989;57:307–33.
21. Cover T, Thomas J. Elements of information theory. New York: John Wiley and Sons, 1991:542.
22. Hall P. On Kullback-Leibler loss and density estimation. Ann Stat 1987;15:1491–519.
23. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. 2nd ed. New York: Springer-Verlag, 2002.
24. Liquet B, Commenges D. Choice of estimators based on different observations: modified AIC and LCV criteria. Scand J Stat 2011;38:268–87.
25. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev 1950;78:1–3.
26. Vaida F, Blanchard S. Conditional Akaike information for mixed-effects models. Biometrika 2005;92:351–70.
27. Greven S, Kneib T. On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika 2010;97:773–89.
28. Braun J, Held L, Ledergerber B. Predictive cross-validation for the choice of linear mixed-effects models with application to data from the Swiss HIV cohort study. Biometrics 2012;68:53–61.
29. Proust-Lima C, Amieva H, Jacqmin-Gadda H. Analysis of multivariate mixed longitudinal data: a flexible latent process approach Br J Math Stat Psychol 2012;66:470–87.
30. Proust-Lima C, Philipps V, Diakite A, Liquet B. LCMM: Estimation of extended mixed models using latent classes and latent processes. R package version 1.6.6, 2014.
31. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–98.
32. Proust C, Jacqmin-Gadda H, Taylor JM, Ganiayre J, Commenges D. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. Biometrics 2006;62:1014–24.
33. Letenneur L, Commenges D, Dartigues JF, Barberger-Gateau P. Incidence of dementia and Alzheimer's disease in elderly community residents of South-Western France. Int J Epidemiol 1994;23:1256–61.