2012

Resampling-Based Information Criteria for Best-Subset Regression

Philip T. Reiss, *New York University*
Lei Huang, *Columbia University*
Joseph E. Cavanaugh, *University of Iowa*
Amy Krain Roy, *Fordham University*

Annals of the Institute of Statistical Mathematics manuscript No.
(will be inserted by the editor)

Resampling-Based Information Criteria for Best-Subset Regression

Philip T. Reiss · Lei Huang · Joseph E. Cavanaugh · Amy Krain Roy

Received: date / Revised: date

Abstract When a linear model is chosen by searching for the best subset among a set of candidate predictors, a fixed penalty such as that imposed by the Akaike information criterion may penalize model complexity inadequately, leading to biased model selection. We study resampling-based information criteria that aim to overcome this problem through improved estimation of the effective model dimension. The first proposed approach builds upon previous work on bootstrap-based model selection. We then propose a more novel approach based on cross-validation. Simulations and analyses of a functional neuroimaging data set illustrate the strong performance of our resampling-based methods, which are implemented in a new R package.

Keywords Adaptive model selection · Covariance inflation criterion · Cross-validation · Extended information criterion · Functional connectivity · Overoptimism

1 Introduction

A popular strategy for model selection is to choose the candidate model minimizing an estimate of the expected value of some loss function for a hypo-

P. T. Reiss
Department of Child and Adolescent Psychiatry, New York University, New York, NY 10016, USA, and Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY 10962, USA
E-mail: phil.reiss@nyumc.org

L. Huang
Department of Child and Adolescent Psychiatry, New York University, New York, NY 10016, USA

J. E. Cavanaugh
Department of Biostatistics, University of Iowa College of Public Health, Iowa City, IA 52242, USA

A. K. Roy
Department of Psychology, Fordham University, Bronx, NY 10458, USA

thetical future set of outcomes. That estimate may take the form of the value of the loss function for the given data plus a penalty or correction term. The latter term compensates for the fact that the loss function will tend to have a lower value for the data from which the model fit was obtained than for a different data set to which the same fitted model is applied. In other words, the correction represents the "overoptimism" (Efron, 1983; Pan and Le, 2001) inherent in using the observed loss as an estimate of the expected loss in a future data set. This general template can lead to ostensibly very different model selection criteria, such as the Akaike (1973, 1974) information criterion (AIC), Mallows' (1973) C_p statistic, and more recent "covariance penalty" methods (Efron, 2004).

In this paper we are concerned with the problem of subset selection for linear models (Miller, 2002), which can be stated as follows. Suppose we are given an n-dimensional outcome vector $\boldsymbol{y}$, and an $n \times P$ matrix $\boldsymbol{X}_{full}$, where the first column of $\boldsymbol{X}_{full}$ consists of 1s, and the remaining columns correspond to $P-1 \geq 1$ candidate predictors. Let $\mathcal{A} = \left\{ \{1\} \cup \tilde{A} : \tilde{A} \subset \{2, \ldots, P\} \right\}$. Any $A \in \mathcal{A}$ defines a model matrix $\boldsymbol{X} = \boldsymbol{X}_A$ of those columns $\boldsymbol{X}_{full}$ indexed by the elements of A, and thereby defines a model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon}$ is an n-dimensional vector of mean-zero errors. (The definition of $\mathcal{A}$ implies that we consider only models that include an intercept.) Our goal is to choose $A \in \mathcal{A}$ for which the estimated expected loss for a future data set is lowest. Following Akaike, we use -2 times the log likelihood as the loss function, which favors a model whose predictive distribution is closest to the true distribution in the sense of Kullback-Leibler information (Konishi and Kitagawa, 2008). The Akaike paradigm is built upon a solid foundation, thanks to its connection with likelihood theory and information theory. In part for this reason, AIC remains the best-known and most widely used criterion for model selection.

Our contribution addresses a problem that is often overlooked in practice: the fact that, when selecting one among many possible models, the overoptimism is inflated. For example, AIC adds the penalty $2p$ to -2 times the log likelihood, where $p = |A|$. Selecting the AIC-minimizing model is best suited to settings in which there is only one candidate model of each size p. But when searching among *all* size-p subsets for each p, the overoptimism associated with selecting the best size-p subset is greater than $2p$. "Adaptive" model selection procedures, in the sense of Tibshirani and Knight (1999), increase the overoptimism penalty to account for searching among a number of possible models of each size (cf. Shen and Ye, 2002, who define adaptive model selection somewhat differently).

We develop an approach to adaptive model selection that is particularly relevant to applications with a moderate number of candidate predictors, i.e., P/n less than 1 but not necessarily near zero. Whereas AIC is based on an asymptotic approximation that assumes $p \ll n$, the corrected AIC (AIC_c)

(Sugiura, 1978; Hurvich and Tsai, 1989) applies an exact penalty term that is not proportional to p. This penalty term (see (7) below) indicates that as p grows while n remains fixed, the rate of growth in overoptimism increases; in other words, complexity is more costly for small-to-moderate than for large samples, so that complexity penalization should depend on both p and n.

This important case of moderate predictor dimension has not been fully addressed in previous work on adaptive linear model selection. Since in this case it is problematic to view the overoptimism as proportional to p, adaptive model selection methods that seek to replace the AIC penalty $2p$ with λp for some $\lambda > 2$ (e.g., Foster and George, 1994; Ye, 1998; Shen and Ye, 2002) may be less than ideal. Our proposed methods more closely resemble the covariance inflation criterion of Tibshirani and Knight (1999), which uses permuted versions of the data to estimate the overoptimism. While this estimate works well when the true model is null, these authors acknowledge (p. 543) that it is biased when the true model is non-null. Instead of data permutation, the adaptive methods that we describe rely on two alternative resampling approaches, bootstrapping and cross-validation, to produce overoptimism estimates that are appropriate whether or not the true model is the null model.

This work was motivated by research relating psychological outcomes to functional connectivity (FC), the temporal correlation of activity levels in different brain regions of interest. In a study of eight left-hemisphere regions, Stein *et al.* (2007) identified 10 between-region connections (i.e., pairs of regions; see Figure 1) whose FC, assessed using functional magnetic resonance imaging (fMRI), may be related to psychological measures. We were interested in exploring FC for these 10 connections as predictors of two such measures: the Rosenberg (1965) self-esteem score, and the General Distress: Depressive Symptoms (GDD) subscale of the Mood and Anxiety Symptom Questionnaire (MASQ) (Watson *et al.*, 1995). Given the paucity of scientific theory to guide model building, it is natural to turn to automatic model selection criteria to find the best subset of the 10 candidate predictors.

Section 2 introduces information criteria for linear model selection, and explains in more detail the need for adaptive approaches. Section 3 describes the bootstrap-based "extended information criterion" of Ishiguro *et al.* (1997), and its extension to adaptive linear model selection. Section 4 proposes an alternative adaptive criterion, based on cross-validatory estimation of the overoptimism, which may overcome some of the limitations of the bootstrap method. Simulations in Section 5 demonstrate that our adaptive methods perform well compared with previous approaches. Section 6 presents analyses of our functional connectivity data set, and Section 7 offers concluding remarks.

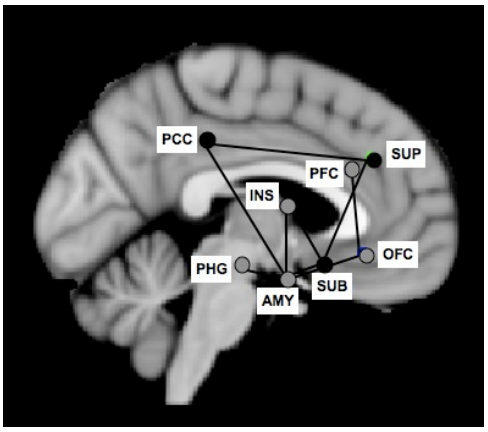

Fig. 1 The eight brain regions studied by Stein *et al.* (2007), who identified the 10 indicated connections (pairs of regions joined by edges) as potential predictors of psychological outcomes. Black dots indicate regions that lie on the mid-sagittal plane shown, while grey dots indicate left-hemisphere regions that have been projected onto this plane for illustration. Abbreviations: AMY, amygdala; INS, insula; OFC, orbitofrontal cortex; PCC, posterior cingulate cortex; PFC, prefrontal cortex; PHG, parahippocampal gyrus; SUB, subgenual cingulate; SUP, supragenual cingulate.

2 Information criteria for linear models

2.1 AIC and corrected AIC

In what follows, we shall refer to model (1) with $\boldsymbol{X} = \boldsymbol{X}_A$ as model A. The notation $\boldsymbol{X}$ for the design matrix, and $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ for the parameter estimates, will refer to an *a priori* model A with $|A| = p$, fitted to the n observations. We shall add subscripts to $\boldsymbol{X}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ (i) to denote a model selected as the best model of a particular dimension, as opposed to *a priori*, and/or (ii) to refer to resampled data sets and the associated model fits.

Imagine a future realization of the data with the same matrix of predictors $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ as in (1), but a new outcome vector $\boldsymbol{y}^+$ that is independent of $\boldsymbol{y}$, conditionally on $\boldsymbol{X}$. A good model will give rise to maximum likelihood estimates (MLEs) $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ such that the expected -2 times log likelihood for the fitted model at the future data $(\boldsymbol{y}^+, \boldsymbol{X})$, i.e.,

$$E\left[-2l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 | \boldsymbol{y}^+, \boldsymbol{X})\right], \tag{2}$$

will be as small as possible; here the expectation is with respect to the joint likelihood of $(\boldsymbol{y}, \boldsymbol{y}^+)$ conditional on $\boldsymbol{X}$. Information criteria estimate the expected loss (2) by

$$-2l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 | \boldsymbol{y}, \boldsymbol{X}) + \hat{C}, \tag{3}$$

where $\hat{C}$ is an estimate of the overoptimism

$$C \equiv E\left[-2l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 | \boldsymbol{y}^+, \boldsymbol{X})\right] - E\left[-2l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 | \boldsymbol{y}, \boldsymbol{X})\right]. \tag{4}$$

The idea of (3) is that $-2l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2|\boldsymbol{y}, \boldsymbol{X})$ has a downward bias as an estimate of (2), since $(\boldsymbol{y}, \boldsymbol{X})$ is the original data set for which the likelihood was maximized to obtain $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$. C is this bias, and the penalty term $\hat{C}$ in (3) is an estimate of it.

If model (1) holds with $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $-2l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2|\boldsymbol{y}, \boldsymbol{X}) = n\log\hat{\sigma}^2 + n$, so that

$$\begin{aligned} C &= E\left[(n\log\hat{\sigma}^2 + \|\boldsymbol{y}^+ - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/\hat{\sigma}^2) - (n\log\hat{\sigma}^2 + n)\right] \\ &= E(\|\boldsymbol{y}^+ - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/\hat{\sigma}^2) - n, \end{aligned} \tag{5}$$

and the generic information criterion (3) reduces to

$$n\log\hat{\sigma}^2 + n + \hat{C}. \tag{6}$$

Sugiura (1978) and Hurvich and Tsai (1989) derive the exact value

$$C = C_{AIC_c}(p) \equiv \frac{n(2p+2)}{n-p-2} \tag{7}$$

(see also the succinct treatment of Davison, 2003, pp. 402–403). Substituting this value for $\hat{C}$ in (6) gives the AIC_c

$$n\log\hat{\sigma}^2 + \frac{n(n+p)}{n-p-2}. \tag{8}$$

If $p \ll n$, (7) is approximately $2(p+1)$ or equivalently $2p$, the ordinary AIC penalty.

The fact that formula (8) assumes that (1) is a correct model, i.e. either the true data-generating model or a larger model, is sometimes seen as a limitation of choosing among candidate models by minimizing AIC_c. In practice, though, model selection by minimal AIC_c has proved effective, since it weeds out both overfitted models (which are heavily penalized) and underfitted models (which have low likelihood). However, this procedure is less than ideal when choosing among all possible subsets, as we explain next.

2.2 Selection bias in best-subset regression

For $p \in \{1, \ldots, P\}$, let $M(p)$ denote the set in $\mathcal{A}$ of size p such that model $M(p)$ attains the highest maximized likelihood of any $A \in \mathcal{A}$ with $|A| = p$. A naïve approach to subset selection would choose the AIC_c-minimizing subset $M(p_{AIC_c})$ where p_{AIC_c} minimizes $n\log\hat{\sigma}^2_{M(p)} + \frac{n(n+p)}{n-p-2}$ over $p \in \{1, \ldots, P\}$. Note, however, that this criterion equals $\min_{|A|=p}(n\log\hat{\sigma}^2_A) + \frac{n(n+p)}{n-p-2}$, which is smaller on average than (8) for a fixed *a priori* size-p model, and hence is a downward-biased estimate of (2). Equivalently, whereas the overoptimism (5) is equal to $C_{AIC_c}(p) = \frac{n(2p+2)}{n-p-2}$ for a fixed size-p model, it is larger when considering the best size-p model. We shall denote this larger value by $C_{ad}(p)$.

The difference $C_{AIC_c}(p) - C_{ad}(p)$ can be thought of as the "selection bias" associated with using (7) to estimate the overoptimism of the selected model of size p.

Of course, this selection bias would have no impact on model selection if it did not depend on p. But it does: in particular, it equals zero for $p \in \{1, P\}$, since there are only one size-1 (null) model and one size-P (full) model, but it is negative for $1 < p < P$, so that AIC_c minimization will be tilted against the null model. This situation is somewhat akin to multiple hypothesis testing (cf. George and Foster, 2000): with many candidate predictors, even if the null model is true, there will be an unacceptably high probability of choosing a non-null model, unless some sort of correction is applied. An appropriate correction is to use an information criterion (6) in which, instead of $C_{AIC_c}(p)$, we take $\hat{C}$ to be an estimate of $C_{ad}(p)$. This is the strategy pursued by the resampling-based information criteria described in the next two sections.

3 The extended (bootstrap) information criterion

3.1 The fixed-model case

Ishiguro *et al.* (1997) propose a nonparametric bootstrap approach to estimating C in a much more general setting than model (1). (See also Konishi and Kitagawa (1996), and the bootstrap model selection criterion of Shao (1996), which is based on prediction error loss rather than likelihood.) Suppose we sample n pairs $(y_i, \boldsymbol{x}_i)$ from the data, with replacement, B times, and denote the bth bootstrap data set thus generated by $(\boldsymbol{y}_b^*, \boldsymbol{X}_b^*)$ and the associated MLEs by $\hat{\boldsymbol{\beta}}_b^*$, $\hat{\sigma}_b^{*2}$. Ishiguro *et al.*'s extended information criterion (EIC) uses

$$\hat{C}_{boot} = \frac{1}{B}\sum_{b=1}^{B}\left[-2l(\hat{\boldsymbol{\beta}}_b^*, \hat{\sigma}_b^{*2}|\boldsymbol{y}, \boldsymbol{X}) + 2l(\hat{\boldsymbol{\beta}}_b^*, \hat{\sigma}_b^{*2}|\boldsymbol{y}_b^*, \boldsymbol{X}_b^*)\right] \tag{9}$$

as an estimate of the overoptimism (4). Intuitively, bootstrapping creates simulated replicates of the data-generating process in which the empirical distribution of $(\boldsymbol{y}_b^*, \boldsymbol{X}_b^*)$ replaces that of $(\boldsymbol{y}, \boldsymbol{X})$, whereas the empirical distribution of $(\boldsymbol{y}, \boldsymbol{X})$ replaces the population distribution; these correspondences motivate using (9) to estimate (4). In the linear model setting this reduces to estimating $E(\|\boldsymbol{y}^+ - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/\hat{\sigma}^2)$, the first term of (5), by $\frac{1}{B}\sum_{b=1}^{B}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_b^*\|^2/\hat{\sigma}_b^{*2}$, so that (9) and (6) yield

$$\hat{C}_{boot} = \frac{1}{B}\sum_{b=1}^{B}\frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_b^*\|^2}{\hat{\sigma}_b^{*2}} - n, \tag{10}$$

$$EIC = n\log\hat{\sigma}^2 + \frac{1}{B}\sum_{b=1}^{B}\frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_b^*\|^2}{\hat{\sigma}_b^{*2}}. \tag{11}$$

3.2 The best-subset case

To correct for the "selection bias" problem described in Section 2.2, one can modify (11) to define an information criterion associated with selection of the best subset of size p. Whereas the bootstrap overoptimism estimate (10) is appropriate for a fixed model, a natural extension to the best-subset case is to estimate $C_{ad}(p)$ by

$$\hat{C}_{ad,boot}(p) = \frac{1}{B}\sum_{b=1}^{B} \frac{\|\boldsymbol{y} - \boldsymbol{X}_{M_b^*(p)}\hat{\boldsymbol{\beta}}^*_{b;M_b^*(p)}\|^2}{\hat{\sigma}^{*2}_{b;M_b^*(p)}} - n, \tag{12}$$

where $M_b^*(p)$ is the best subset of size p for the bth resampled data set, and $\hat{\boldsymbol{\beta}}_{b;M_b^*(p)}$ and $\hat{\sigma}^{*2}_{b;M_b^*(p)}$ are the MLEs from the corresponding model for that data set. Substituting into the generic criterion (6) gives an adaptive EIC

$$EIC_{ad}(p) = n\log\hat{\sigma}^2_{M(p)} + \frac{1}{B}\sum_{b=1}^{B} \frac{\|\boldsymbol{y} - \boldsymbol{X}_{M_b^*(p)}\hat{\boldsymbol{\beta}}^*_{b;M_b^*(p)}\|^2}{\hat{\sigma}^{*2}_{b;M_b^*(p)}}, \tag{13}$$

which is to be minimized with respect to p.

To understand the motivation for (12), (13), recall that in bootstrap estimation of the overoptimism, each bootstrap data set takes the place of the original data, while the original data stands in for the new data set. Estimate (12) substitutes the bootstrap data for the original data both for selection of the best size-p model $M_b^*(p)$ and for the resulting parameter estimates $\hat{\boldsymbol{\beta}}^*_{b;M_b^*(p)}, \hat{\sigma}^{*2}_{b;M_b^*(p)}$, in order to capture, and compensate for, the overoptimism arising from the entire procedure of choosing the best size-p subset. See Appendix A.1 for an alternative definition of best-subset EIC.

Criterion (13) is defined only for the best size-p model for each p. To be able to assign a score to model A for any set $A \in \mathcal{A}$, we extend the definition by applying penalty (12) not only to the best p-term model but to every such model, and thus obtain

$$EIC_{ad}(A) = n\log\hat{\sigma}^2_{A} + \frac{1}{B}\sum_{b=1}^{B} \frac{\|\boldsymbol{y} - \boldsymbol{X}_{M_b^*(|A|)}\hat{\boldsymbol{\beta}}^*_{b;M_b^*(|A|)}\|^2}{\hat{\sigma}^{*2}_{b;M_b^*(|A|)}}. \tag{14}$$

The best size-p model $M_b^*(p)$ for each bootstrap data set can be computed efficiently using the branch-and-bound algorithm implemented in the package `leaps` (Lumley, 2009) for R (R Development Core Team, 2010). This algorithm would be difficult to integrate with parametric bootstrap samples; hence our preference for nonparametric bootstrapping.

3.3 Small-sample performance of EIC

Returning to the fixed-model setting of Section 3.1, we undertook to study the performance of the bootstrap overoptimism estimate $\hat{C}_{boot}$ (10) when the true

Table 1 Median (and range from 5th to 95th percentile) of $E(\hat{C}_{boot}) - C$ in 300 simulations.

	p				
n	5	10	15	20	25
50	0.8 (-4.2-6.8)	5.1 (-6.1-23)	29 (-3-88.5)	110.6 (17.4-411)	609.8 (123.3-5801.7)
100	0.3 (-3-3.9)	1 (-4.6-11.1)	5.4 (-5.8-19)	12.8 (-3.2-35.8)	23.4 (-0.4-65.8)
150	-0.1 (-2.3-2.5)	0.8 (-3.9-6.2)	2.2 (-5.8-11)	5.4 (-5.3-17.7)	10.6 (-4.6-28.7)

expected loss (2) is known: namely, when model (1) is correct and the errors are independent and identically distributed (IID) normal. As noted above, we then have $C = \frac{n(2p+2)}{n-p-2}$. We were thus able to compare the penalty $\hat{C}_{boot}$ to this "gold standard" in 300 Monte Carlo simulations for each of the values of n and p displayed in Table 1 (see Appendix Appendix B: for details of the simulation procedure). For fixed n, $E(\hat{C}_{boot}) - C$ is seen to increase with p, so that EIC will be biased toward smaller models.

The positive bias of $\hat{C}_{boot}$ can be better understood by rewriting $\hat{C}_{boot}$ as

$$\frac{1}{B}\sum_{b=1}^{B}\frac{\sum_{i=1}^{n}(1-r_i^b)(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_b^*)^2}{\hat{\sigma}_b^{*2}}$$

$$= \frac{1}{B}\sum_{b=1}^{B}\left[\sum_{i:r_i^b=0}\frac{(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_b^*)^2}{\hat{\sigma}_b^{*2}} - \sum_{i:r_i^b>1}\frac{(r_i^b-1)(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_b^*)^2}{\hat{\sigma}_b^{*2}}\right], \qquad (15)$$

where r_i^b is the number of occurrences of the ith observation $(\boldsymbol{x}_i, y_i)$ in the bth bootstrap sample. The first term within the brackets is the sum of scaled squared errors for those observations not included in the bth bootstrap sample; the second is a weighted sum of scaled squared errors for observations which are included multiple times, and which therefore have high influence on the estimate $\hat{\boldsymbol{\beta}}_b^*$. Consequently, the summands in this second term will tend to be atypically small, making (15) a positively biased estimate of the overoptimism. The key point for our purposes is that this bias increases with p, especially when n is small (in agreement with the "EIC_1" results of Konishi and Kitagawa (2008), p. 209), so that EIC will tend to underfit in the fixed-model case. The simulation results in Section 5 suggest that the same is true of EIC_{ad} in the best-subset case.

4 A cross-validation information criterion

4.1 Motivation and initial definition

We have seen that EIC seeks to gauge the extent to which the original-data likelihood overestimates the expected likelihood for an independent data set, using bootstrap samples as surrogates for the original data, and the full data as a surrogate for an independent data set. But clearly the full-data outcomes are not independent of the bootstrap-sample outcomes, and the argument at

the end of the previous section suggests that this lack of independence is what makes EIC biased in small samples. This led us to consider an alternative: using cross-validation (CV) to estimate the overoptimism. We remark that the bootstrap .632 and .632+ estimators of prediction error (Efron, 1983; Efron and Tibshirani, 1997; cf. Efron, 2004), which seek to improve upon CV, use bootstrap samples in a similar spirit to CV—essentially, the data points excluded from bootstrap samples are used to correct for overoptimism. However, this work is concerned with a different class of loss functions and with prediction error for a fixed model, and it is not clear how to adapt the .632 and .632+ estimators to our context of selecting among all possible subsets.

Let $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_K \end{pmatrix}$ and $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_K \end{pmatrix}$ where, for $k = 1, \ldots, K$, $\boldsymbol{y}_k \in \mathcal{R}^{n_v}$ and $\boldsymbol{X}_k$ has n_v rows, with $n_v = n/K$. Let $(\boldsymbol{y}_{-k}, \boldsymbol{X}_{-k})$ denote the kth "training set", i.e. the $n_t = n - n_v$ observations not included in $(\boldsymbol{y}_k, \boldsymbol{X}_k)$ (the kth "validation set"), and let $\hat{\boldsymbol{\beta}}_{-k}, \hat{\sigma}^2_{-k}$ be the MLEs obtained by fitting model (1) to $(\boldsymbol{y}_{-k}, \boldsymbol{X}_{-k})$. We can then define a CV-based analogue of the overoptimism—

$$C^* = E\left(\sum_{k=1}^{K} \frac{\|\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{-k}\|^2}{\hat{\sigma}^2_{-k}}\right) - n, \tag{16}$$

where the expectation is with respect to the joint distribution of $(\boldsymbol{y}, \boldsymbol{X})$—along with its natural unbiased estimate $\hat{C}^*_{CV} = \sum_{k=1}^{K} \|\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{-k}\|^2/\hat{\sigma}^2_{-k} - n$, a CV-based analogue of the EIC overoptimism estimate (10). Two remarks are in order:

1. C^* is approximately equal to

$$E(\|\boldsymbol{y}^+ - \boldsymbol{X}^+\hat{\boldsymbol{\beta}}\|^2/\hat{\sigma}^2) - n, \tag{17}$$

   where $(\boldsymbol{y}^+, \boldsymbol{X}^+)$ is an independent set of n observations drawn from the same joint distribution, and the expectation is with respect to the distribution of $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{y}^+, \boldsymbol{X}^+)$. But (16) is slightly larger than (17) because the estimates $\hat{\boldsymbol{\beta}}_{-k}, \hat{\sigma}^2_{-k}$ are based on $n_t < n$ observations, inflating the mean prediction error.
2. (17) is similar to the overoptimism C (5), but the definition of C assumes the original predictor matrix $\boldsymbol{X}$ is fixed and we draw an independent set of outcomes $\boldsymbol{y}^+$ from the same *conditional* (on $\boldsymbol{X}$) distribution as $\boldsymbol{y}$.

In order to apply C^* to selecting the model dimension, we will need to express it as $C^*(p)$, an explicit function of p, analogous to the overoptimism formula $C_{AIC_c}(p)$ of (7). Since C^* depends on the distribution of $\boldsymbol{X}$ as well as that of $\boldsymbol{y}$, our derivation of $C^*(p)$ (see below, Section 4.2) will rely on assumptions regarding the predictor distribution.

Similarly to (12), we can define adaptive extensions of $\hat{C}^*_{CV}$ and C^* for $M(p)$, the best model of dimension p:

$$\hat{C}^*_{ad,CV}(p) = \sum_{k=1}^{K} \frac{\|\boldsymbol{y}_k - \boldsymbol{X}_{k;M_{-k}(p)}\hat{\boldsymbol{\beta}}_{-k;M_{-k}(p)}\|^2}{\hat{\sigma}^2_{-k;M_{-k}(p)}} - n, \qquad C^*_{ad}(p) = E[\hat{C}^*_{ad,CV}(p)],$$

where $M_{-k}(p)$ is the best size-p model for the kth training set; $\hat{\boldsymbol{\beta}}_{-k;M_{-k}(p)}$, $\hat{\sigma}^2_{-k;M_{-k}(p)}$ are the associated MLEs for this training set; and $\boldsymbol{X}_{k;M_{-k}(p)}$ comprises the corresponding columns of $\boldsymbol{X}_k$.

The function $C^*(p)$, given explicitly in (23), is strictly increasing on its domain and hence invertible, so we have $C_{AIC_c}(p) = C_{AIC_c} \circ C^{*-1} \circ C^*(p)$. This suggests the following plug-in estimator of the overoptimism of $M(p)$:

$$\begin{aligned} \hat{C}_{ad,CV}(p) &= C_{AIC_c} \circ C^{*-1} \circ \hat{C}^*_{ad,CV}(p) \\ &= C_{AIC_c}(df_p) = \frac{n(2df_p + 2)}{n - df_p - 2} \end{aligned} \tag{18}$$

where

$$df_p = C^{*-1}[\hat{C}^*_{ad,CV}(p)]. \tag{19}$$

One can think of df_p as the effective degrees of freedom, or effective model dimension, of $M(p)$. As we show in Appendix Appendix C:, generally speaking $df_p \geq p$.

The overoptimism estimate (18) leads to

$$CVIC(p) = n \log \hat{\sigma}^2_{M(p)} + n + \hat{C}_{ad,CV}(p) = n \log \hat{\sigma}^2_{M(p)} + \frac{n(n + df_p)}{n - df_p - 2} \tag{20}$$

—i.e., AIC_c for $M(p)$, but with p replaced by df_p in the penalty—as our model selection criterion. Like (13), this criterion is defined only for the best model of each size; but it can be extended to all candidate models, analogously to (14).

4.2 Derivation of $C^*(p)$

To evaluate (18) and hence (20), we must derive the aforementioned function $C^*(p)$. We begin by writing

$$\boldsymbol{H} = (h_{ij})_{1 \leq i,j \leq n} \equiv \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \begin{pmatrix} \boldsymbol{H}_{11} & \dots & \boldsymbol{H}_{1K} \\ \vdots & \ddots & \vdots \\ \boldsymbol{H}_{K1} & \dots & \boldsymbol{H}_{KK} \end{pmatrix},$$

where each of the above blocks is $n_v \times n_v$. We can then state the following result, which gives the expectation of $\hat{C}^*_{CV}$, conditional on the K "folds" that make up $\boldsymbol{X}$. Theorems 1 and 2 are proved in Appendix Appendix D:.

Theorem 1 *Suppose model (1) holds where* $\boldsymbol{X}$ *is an* $n \times p$ *matrix and* $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$*. Assume that* $n_t > p+2$ *and that* $\boldsymbol{I}_{n_v} - \boldsymbol{H}_{kk}$ *is invertible for each* k*. Then*

$$E\left(\sum_{k=1}^{K} \frac{\|\boldsymbol{y}_k - \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_{-k}\|^2}{\hat{\sigma}^2_{-k}} \middle| \boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\right) = \frac{n_t}{n_t - p - 2} \sum_{k=1}^{K} tr[(\boldsymbol{I}_{n_v} - \boldsymbol{H}_{kk})^{-1}]. \tag{21}$$

To derive $C^*(p)$, i.e. an expression for $E(\hat{C}^*_{CV})$ that depends only on the model dimension p but not on $\boldsymbol{X}$, we require distributional assumptions on the rows of $\boldsymbol{X}$.

Theorem 2 *Let* $\boldsymbol{X} = (\boldsymbol{1} \mid \tilde{\boldsymbol{X}})$ *where* $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n$*, the rows of* $\tilde{\boldsymbol{X}}$*, are IID* $(p-1)$*-variate normal. Under the assumptions of Theorem 1, if* $\boldsymbol{X}_{-k}$ *has rank* p *for each* k*, then*

$$E\left(\sum_{k=1}^{K} \frac{\|\boldsymbol{y}_k - \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_{-k}\|^2}{\hat{\sigma}^2_{-k}}\right) = \frac{n(n_t+1)(n_t-2)}{(n_t-p-2)^2}. \tag{22}$$

By (16) and (22), we may take

$$C^*(p) = n\left[\frac{(n_t+1)(n_t-2)}{(n_t-p-2)^2} - 1\right] \tag{23}$$

—provided that the predictor vectors are IID multivariate normal. In practice this condition may not hold; but Appendix Appendix E: offers evidence that (23) is a reasonable approximation in most realistic settings. We therefore adopt (23) as our formula for $C^*(p)$ in what follows.

4.3 Refining the criterion by constrained monotone smoothing

Three obvious limitations of CVIC (20) are:

1. It is based on an overoptimism estimate $\hat{C}_{ad,CV}(p)$ that is stochastic, unlike the fixed penalty $C_{AIC_c}(p)$.
2. For $p = 2, \ldots, P-1$, we may be willing to accept the stochastic estimate $\hat{C}_{ad,CV}(p)$, since no closed-form expression for $C_{ad}(p)$ exists. But as noted in Section 2.2, for $p = 1, P$, there is only one candidate model of size p, so $C_{ad}(p)$ reduces to $C_{AIC_c}(p) = \frac{n(2p+2)}{n-p-2}$. In effect, for $p = 1, P$, df_p is just a noisy version of p, and CVIC is just a noisy version of AIC_c—and hence inferior to simply using AIC_c.
3. Whereas C_{ad} is an increasing function of p, $\hat{C}_{ad,CV}$ need not be. This may cause CVIC to favor overfitting in some cases.

We can mitigate the first of these problems, and eliminate the other two, by means of the constrained penalized spline algorithm implemented in the R package `mgcv` (Wood, 2006). This algorithm allows us to compute a smooth function $\tilde{df}(p)$, approximating df_p at $p = 1, \ldots, P$, such that (i) $\tilde{df}(1) = 1$ and

$\tilde{df}(P) = P$, and (ii) $\tilde{df}$ is constrained to be monotonically increasing, by the method of Wood (1994). (Figure 8 displays the raw df_p plotted against the smoothed $\tilde{df}(p)$ for a real data example.) We can then replace the raw CVIC (20) with the monotonic variant

$$CVIC_{mon}(p) = n \log \hat{\sigma}^2_{M(p)} + \frac{n[n + \tilde{df}(p)]}{n - \tilde{df}(p) - 2}.$$

4.4 Summary of the proposed adaptive methods

We provide here a brief summary of the resampling-based information criteria considered above. Section 3 introduced Ishiguro *et al.*'s (1997) EIC, based on using bootstrap samples to estimate the overoptimism (5), and proposed the adaptive extension EIC_{ad}. In Section 3.3 we showed that the bootstrap tends to overpenalize larger models in small samples, and this motivated an alternative, cross-validatory approach to adaptive linear model selection. Section 4.1 defined the CVIC in terms of a quantity $C^*(p)$ which we derived in Section 4.2. In Section 4.3 we sought to overcome some of CVIC's limitations by applying constrained monotone smoothing to the effective degrees of freedom, resulting in the new criterion CVIC_{mon}.

A somewhat unappealing feature of the CVIC overoptimism estimate (18) is that it estimates $C_{ad}(p)$ indirectly, by applying $C_{AIC_c} \circ C^{*-1}$ to $\hat{C}^*_{ad,CV}(p)$, in contrast to the direct bootstrap estimate $\hat{C}_{ad,boot}(p)$ (12). On the other hand, whereas $\hat{C}_{ad,boot}(p)$ is an adaptive extension of $\hat{C}_{boot}(p)$, which we have shown to be a biased overoptimism estimator, $\hat{C}^*_{ad,CV}(p)$ is an adaptive extension of the *unbiased* estimator $\hat{C}^*_{CV} = \hat{C}^*_{CV}(p)$ of $C^*(p)$—offering some hope that the resulting model selection criterion CVIC will outperform the bootstrap criterion EIC_{ad}. Appendix A.2 discusses two alternative CV-based approaches to subset selection.

5 Simulation study

5.1 Setup

We conducted a simulation study to compare the performance of minimization of (1) AIC; (2) AIC_c; (3) BIC; (4) CIC; (5) EIC_{ad}; (6) CVIC; and (7) CVIC_{mon}. Four sets of 300 simulations were performed. Each set began by choosing a 50×20 predictor matrix $\boldsymbol{X} = (\boldsymbol{x}_1 \ldots \boldsymbol{x}_{50})^T$ whose rows were independently generated from a multivariate normal distribution with mean zero and 20×20 covariance matrix having (j, k) entry $0.7^{|j-k|}$. (Note that here, as in Tibshirani and Knight's (1999) simulation study, $\boldsymbol{X}$ does not include a column of 1s, since the true intercept is 0.) Then, for each individual simulation, outcomes $y_1, \ldots, y_{50}$ were generated from the model $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i$, where $\varepsilon_1, \ldots, \varepsilon_{50}$ are independent normal variates with mean 0 and variance

$\sigma^2 = 1$. In the first set of simulations the true coefficient vector $\boldsymbol{\beta} \in \mathcal{R}^{20}$ was set to zero. For the remaining simulations, there were two sets of nonzero coefficients centered around the 5th and 15th predictors. These were set initially to $\beta_{5+j} = \beta_{15+j} = \sqrt{h - |j|}$ for $|j| < h$, where h was 1, 2, 3 in the second, third, and fourth sets of simulations, resulting in 2, 6, and 10 nonzero coefficients. We then multiplied $\boldsymbol{\beta}$ by a constant chosen so that the final $\boldsymbol{\beta}$ would satisfy $R^2 \equiv \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} / (n\sigma^2 + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}) = 0.75$. This quantity is the "theoretical R^2" used by Tibshirani and Knight (1999), i.e., the regression sum of squares divided by the expected total sum of squares for the true model; note that the more general expression given by Luo et al. (2006) would be required if the true model had a nonzero intercept. For CIC and EIC_{ad}, resampling was performed 40 times per simulation; for the two CVIC variants, leave-one-out CV was used.

5.2 Comparative performance

Figures 2–5 display the true coefficients for each set of simulations, along with the proportion of simulations in which each coefficient was included in the model—i.e., the "detection rate" for truly nonzero coefficients, and the "false alarm rate" for zero coefficients—for each method. Figure 6 shows boxplots of the model error $\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ (i.e., the mean prediction error minus σ^2) for each method, in each set of simulations.

For the true models with 0 or 2 nonzero coefficients, the proposed methods EIC_{ad}, CVIC and CVIC_{mon} are the best performers, achieving near-perfect variable selection. The constrained smoothing makes CVIC_{mon} somewhat superior to CVIC, but EIC_{ad} attains the lowest model error of all the methods. For the models with 6 or 10 nonzero coefficients, the three proposed methods are again notably resistant to false alarms, but have difficulty detecting the truly nonzero coefficients; hence these methods' model error is no better than that of the other methods, and indeed EIC_{ad}, which tends to overpenalize large models (see Section 3.3), has the highest model error when the true model has 10 nonzero coefficients. Since the selection bias discussed in Section 2.2 tends to favor non-null over null models, it is unsurprising that our adaptive methods provide the greatest benefit when all or most of the true coefficients are zero.

CIC, like our proposed methods, aims to take into account the process of searching among numerous candidate models, but the simulation results suggest that its ability to protect against spurious predictors diminishes as the true model grows: CIC's false alarm rates are lower than AIC_c's when the true model is null, but higher for the larger models.

5.3 Variability

A key criticism of resampling-based overoptimism estimators is that their inherent variability leads to unstable model selection. To investigate the vari-

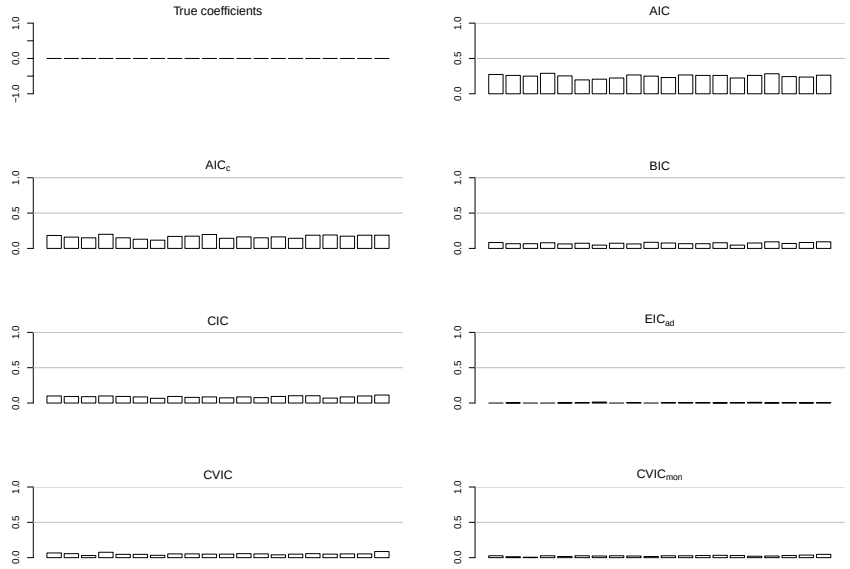

Fig. 2 This figure and the three figures that follow show detection rates and false alarm rates for four true models with 20 candidate predictors. The top left subfigure shows the true coefficients, which are all zero in this case, while the other subfigures display the relative frequency of inclusion of each predictor, in 300 simulations, based on the seven methods listed in the text.

Table 2 Mean, over 100 simulated data sets, of the mean and standard deviation of the overoptimism estimates from 30 replicates of each of the proposed adaptive procedures. The last row shows C_{AIC_c} (7) for comparison. Note that the value 4.26 in the first column is the true overoptimism, and is recovered exactly by CVIC$_{mon}$. In the remaining columns, the true overoptimism C_{ad} is greater than C_{AIC_c} due to the effect of model selection.

	Number of predictors with nonzero coefficients			
	0	2	6	10
EIC$_{ad}$	4.18 (1.13)	10.76 (1.5)	72.16 (5.05)	147.38 (12.17)
CVIC	3.17 (1.24)	7.86 (2.09)	43.59 (7.06)	62.98 (8.20)
CVIC$_{mon}$	4.26 (0)	12.18 (2.19)	42.88 (5.91)	62.58 (7.23)
AIC$_c$	4.26	8.89	19.51	32.43

ability of our estimators of C_{ad}, we performed another four sets of 100 simulations, using the exact same specifications for the sample size, predictors and outcomes as above. For each simulated data set, we computed 30 replicates of the EIC$_{ad}$ overoptimism estimate for the true model size using 100 bootstrap samples, and 30 replicates of the CVIC and CVIC$_{mon}$ overoptimism estimates using 10-fold CV; we then obtained the mean and standard deviation (SD) of the 30 estimates by each method. The means of these values over the 100 data sets are given in Table 2.

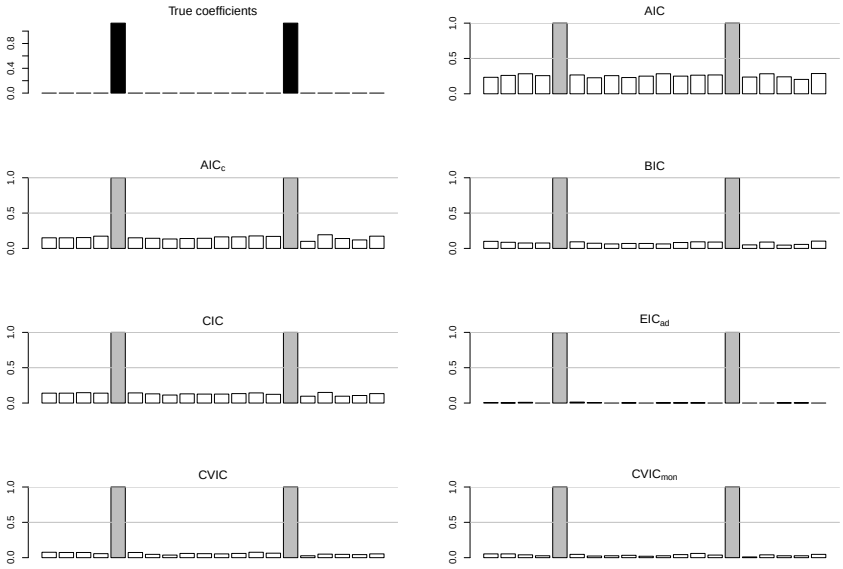

Fig. 3 Detection and false alarm rates for the true coefficient vector shown at top left, i.e., two equal positive coefficients and all other coefficients equal to zero. Grey bars indicate detection rates for truly nonzero coefficients, and white bars give false alarm rates for zero coefficients.

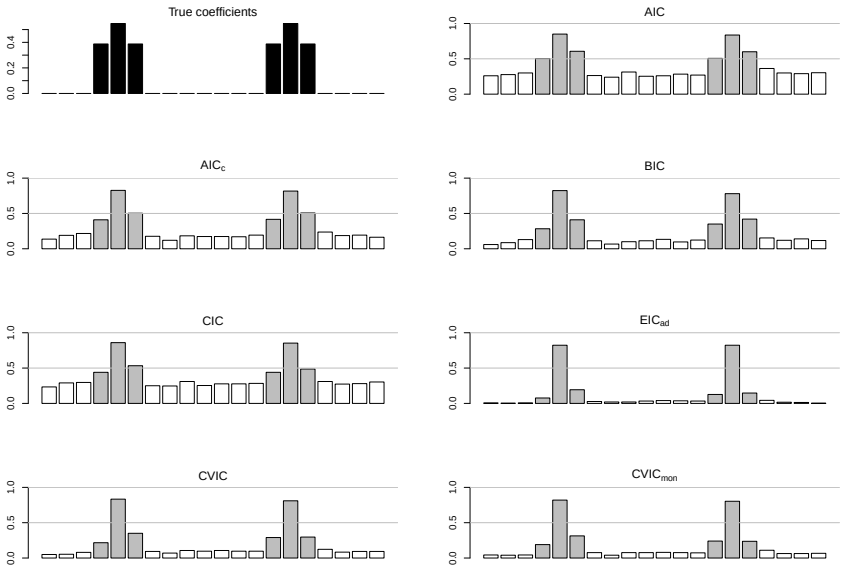

Fig. 4 Detection and false alarm rates for the true coefficient vector shown at top left (six nonzero coefficients).

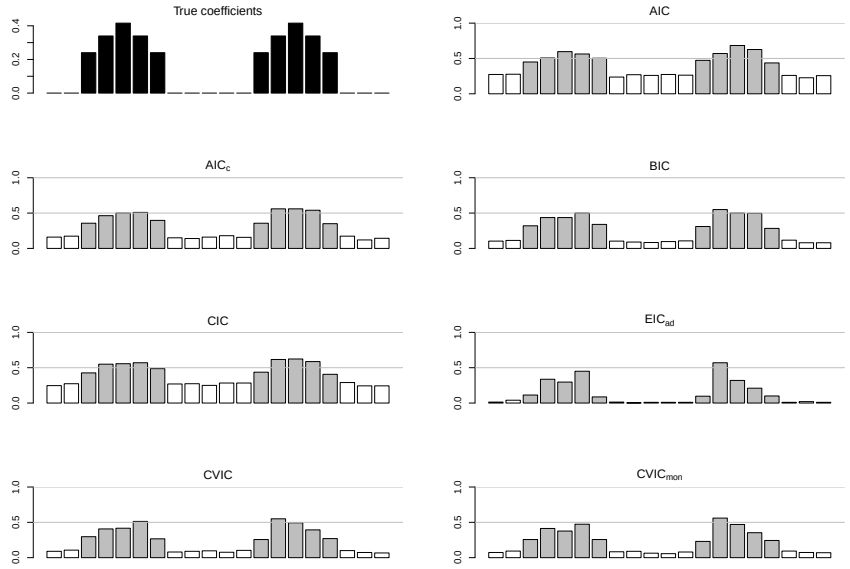

Fig. 5 Detection and false alarm rates for the true coefficient vector shown at top left (ten nonzero coefficients).

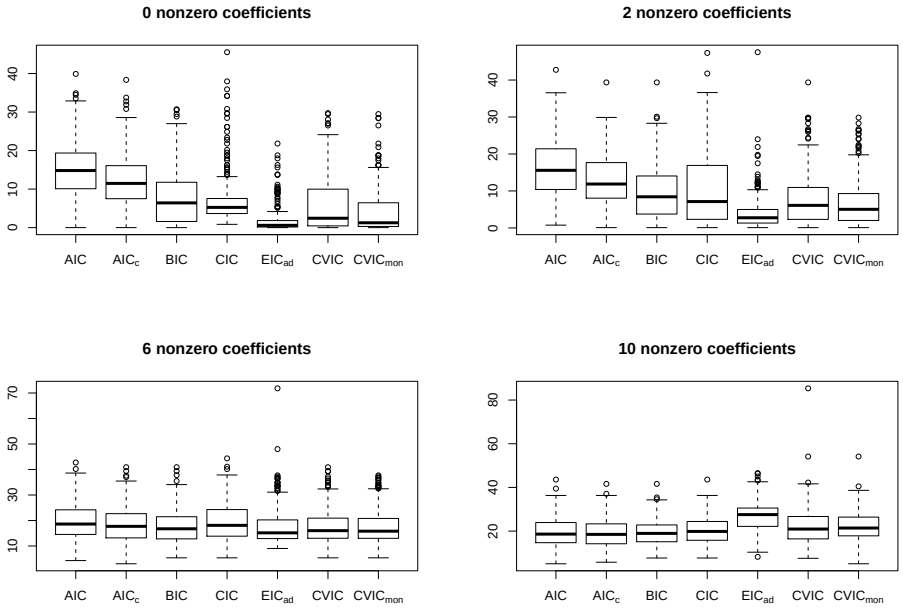

Fig. 6 Model error $\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ under the four simulation settings.

The raw CVIC penalty is seen to have somewhat higher SD than that of EIC_{ad}, except in the column for 10 true predictors, where EIC_{ad} has much higher SD. Note that in that column the mean is much higher for EIC_{ad} than for CVIC; although the true value C_{ad} is unknown, the EIC_{ad} penalty seems to be positively biased here, in line with the fixed-model results in Table 1. The SD values for CVIC_{mon} suggest that the constrained monotone smoothing succeeds in reducing the penalty variability; the only exception is that with 2 true predictors, the SD is slightly higher for CVIC_{mon} than for raw CVIC, but here the mean is also notably higher.

6 Application: functional connectivity in the human brain

We now turn to the application outlined in the Introduction. Self-esteem and MASQ-GDD (depression) scores, and FC values for the 10 connections described above, were acquired in a sample of 43 participants scanned with resting-state fMRI (Biswal *et al.*, 1995) at New York University. The regions of interest were defined, and FC was computed, as in Stark *et al.* (2008). Approximately optimal Box-Cox transformations were applied to the two psychological outcomes (the third power for self-esteem, and the logarithm for MASQ-GDD), which were then regressed on subsets of the 10 FC predictors. We compare the subset selection results for AIC, AIC_c, EIC_{ad}, and CVIC_{mon}.

For regression of self-esteem score on the 10 connections, there are 13 models, with 1–5 predictors, having somewhat lower AIC than the null model. The best model (AIC value 771.2, versus 771.7 for the second-best model and 772.5 for the null model) includes the AMY-SUB, OFC-AMY, SUB-INS, and SUP-PCC connections. But since a number of models have AIC near the minimum, it seems reasonable to adopt a "pluralistic" approach that considers all of the near-optimal models and asks which predictors appear most often in them. Of the 13 models that outperform the null model, two of the above four connections stand out in terms of inclusion frequency: SUB-INS appears in all 13 models, while OFC-AMY occurs in all but two—suggesting that these two connections may be most strongly associated with self-esteem. However, AIC_c, EIC_{ad} and CVIC_{mon} choose the null model as the best model, suggesting that chance variation accounts for the AIC-based findings.

For the MASQ-GDD score, on the other hand, the adaptive criteria yield rather different results than either AIC or AIC_c. The number of models outscoring the null model is 78 for AIC and 56 for AIC_c, but only 3 for CVIC_{mon} (the models including PCC-AMY only, SUB-INS only, or both), and 1 for EIC_{ad} (the model including both PCC-AMY and SUB-INS). Figure 7 displays the 10 best models according to AIC_c, and the 3 models with lower CVIC_{mon} values than the null model. All four criteria choose the model with the PCC-AMY and SUB-INS connections as the best. Moreover, these two connections occur in all 10 of the lowest-AIC_c models, whereas no other connection occurs in more than 3 of these models. But whereas the AIC_c results indicate that one or two additional predictors might improve the model, the EIC_{ad} and CVIC_{mon}

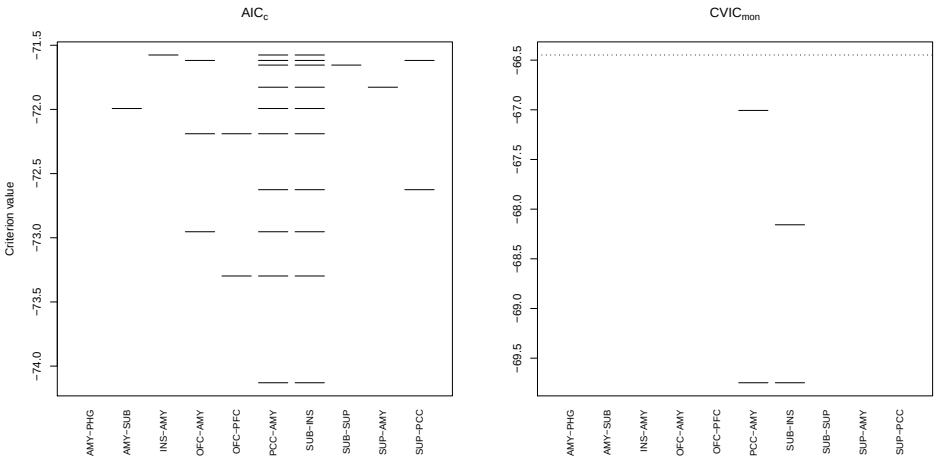

Fig. 7 Information criterion values for the best models for MASQ-GDD score, based on AIC_c and on CVIC_{mon}. The set of line segments at a given y-coordinate indicates the connections included in the model attaining that value of the information criterion, and the dotted line in the right plot indicates CVIC_{mon} for the null model.

results imply that such predictors would be spurious. The 10 lowest-AIC models (not shown in Figure 7), like the lowest-AIC_c models, consistently include PCC-AMY and SUB-INS, but the former models are generally even larger, including as many as 6 of the connections. Figure 8 shows the effective degrees of freedom obtained by the CVIC_{mon} method.

The fitted models regressing log-transformed GDD on PCC-AMY and/or SUB-INS suggest that both are positively related with depression: a one-standard-deviation increase in either connectivity score is associated with approximately a 10% increase in the GDD subscale. The coefficients of determination are roughly 11% for PCC-AMY alone, 13% for SUB-INS alone, and 25% for the model containing both.

7 Discussion

In the Introduction, we argued that our resampling approach to model selection should be particularly relevant when the predictor dimension p is not negligible compared with the sample size n. On the other hand, when p becomes too large to compute a score such as AIC for all subsets, it is more common to turn to other methods such as partial least squares (Helland, 1988), ridge regression (Hoerl and Kennard, 1970) or the lasso (Tibshirani, 1996). Indeed, such methods have been applied successfully in moderate-p applications. Nevertheless, when it is feasible to evaluate a score such as CVIC_{mon} for all subsets, there are advantages to doing so—including the fact that this approach provides a natural framework for obtaining not just a single model

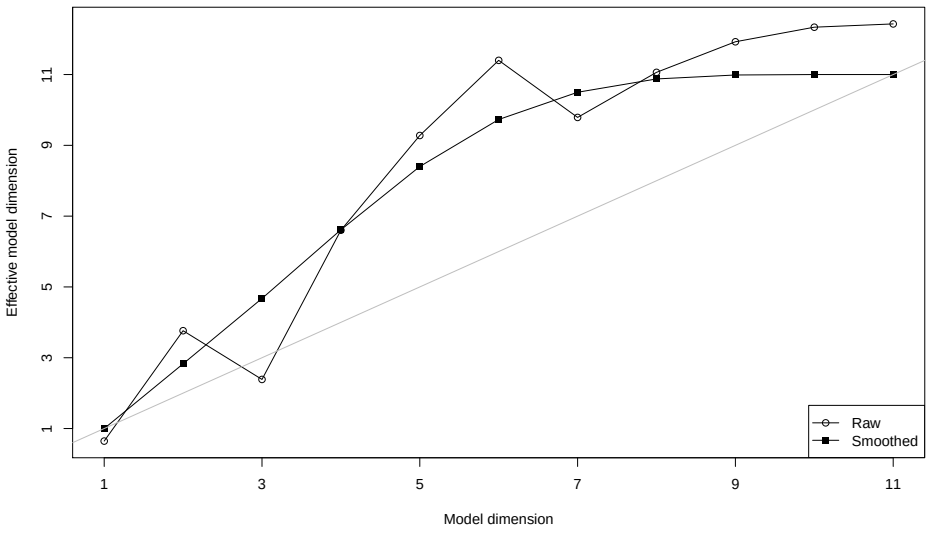

Fig. 8 CVIC_{mon} for the GDD model: effective model dimension, raw (df_p) and smoothed ($\hat{df}(p)$), for the selected model of each size. The line of identity is shown in grey.

estimate but a collection of leading candidate models, as illustrated in our functional connectivity example (Section 6 above).

The above data analysis also demonstrates that the relatively low false-alarm rate of resampling-based information criteria makes them particularly useful for highly exploratory "discovery science" studies in which preventing spurious findings is of central importance. In other settings, however, these criteria's tendency to err toward non-detection of true predictors may represent a serious limitation. Another concern is the variability of our methods' overoptimism estimates, which cannot be reduced by the technique of Konishi and Kitagawa (1996) since that device is applicable only in the fixed-model case. In ongoing research, we are seeking ways to surmount these limitations.

To extend our resampling-based information criteria from linear to generalized linear models (GLMs), two difficulties must be surmounted. First, finding the best model for each resampled data set requires efficient subset selection algorithms, which are readily available for linear models (e.g., the above-cited "leaps" algorithm) but much less developed for GLMs. Second, the closed-form expressions for expected overoptimism used in the CVIC criterion are valid only for the linear case. The methods of Lawless and Singhal (1978) and Cerdeira et al. (2009) have enabled us to overcome the first difficulty and implement a logistic regression version of EIC_{ad}, but this criterion appears to suffer from a marked tendency to underfit. This preliminary finding adds to our motivation to overcome the second difficulty and extend CVIC to the generalized linear case.

The methods of this paper have been implemented in an R package called `reams` (*re*sampling-based *a*daptive *m*odel *s*election), available at `http://cran.r-project.org/web/packages/reams` .

Appendix A: Some alternatives to our adaptive criteria

A.1 An alternative form of best-subset EIC

An alternative to (13) would be

$$n \log \hat{\sigma}^2_{M(p)} + \frac{1}{B} \sum_{b=1}^{B} \frac{\|\boldsymbol{y} - \boldsymbol{X}_{M(p)} \hat{\boldsymbol{\beta}}^*_{b;M(p)}\|^2}{\hat{\sigma}^{*2}_{b;M(p)}}. \tag{24}$$

This criterion differs from (13) in that, although new parameter estimates $\hat{\boldsymbol{\beta}}^*_{b;M(p)}, \hat{\sigma}^{*2}_{b;M(p)}$ are obtained for each bootstrap sample, we reuse the original-data selected subset $M(p)$, rather than selecting a new subset $M^*_b(p)$ for each bootstrap sample as in (13). Criterion (24) is more computationally efficient than (13), and may be equivalent to the "EIC_2" criterion of Konishi and Kitagawa (2008, pp. 208-209). However, because this criterion uses the bootstrap data as a surrogate for the real data only for parameter estimation, but not for model selection, it may not fully capture the overoptimism associated with parameter estimates from a selected model. Indeed, in simulations similar to those reported in Section 5, we found criterion (24) to be susceptible, like nonadaptive criteria, to a high rate of false detections.

A.2 Two alternative CV methods

An alternative to the CVIC (20), which at first glance may seem much simpler, is to compute the usual K-fold CV estimate of prediction error $\sum_{k=1}^{K} \|\boldsymbol{y}_k - \boldsymbol{X}_{-k;A} \boldsymbol{\beta}_{-k;A}\|^2$ for every possible subset A, and choose A for which this quantity is minimized. However, to the best of our knowledge, it is nontrivial to extend efficient all-subsets regression algorithms to perform all-subsets CV. Moreover, this approach would be biased in favor of model dimensions p for which the number of subsets of size p is highest.

Another alternative to CVIC is to minimize

$$n \log \hat{\sigma}^2_{M(p)} + n + \hat{C}^*_{ad,CV}(p), \tag{25}$$

i.e., to define our criterion directly in terms of $\hat{C}^*_{ad,CV}(p)$ instead of the overoptimism estimate $\hat{C}_{ad,CV}(p) = C_{AIC_c} \circ C^{*-1} \circ \hat{C}^*_{ad,CV}(p)$ appearing in (20). However, we chose the latter quantity for consistency with the standard definitions (4), (5) of the overoptimism C. Had we opted for (25), it would not have been clear whether our criterion performed differently from AIC_c due to its adaptive nature, or due to substituting the estimand C^* for the traditional C.

Appendix B: The simulation procedure of Section 3.3

Suppose the bth bootstrap sample consists of cases $i^b_1, i^b_2, \ldots, i^b_n$. The resampled data set can be written as $\boldsymbol{y}^*_b = \boldsymbol{S}_b \boldsymbol{y}$ and $\boldsymbol{X}^*_b = \boldsymbol{S}_b \boldsymbol{X}$ where $\boldsymbol{S}_b$ is the sampling-with-replacement matrix $(\boldsymbol{e}_{i^b_1}\ \boldsymbol{e}_{i^b_2}\ \ldots\ \boldsymbol{e}_{i^b_n})^T$; here $\boldsymbol{e}_k$ is the n-dimensional vector with 1 in the kth position and 0 elsewhere. At first glance, Monte Carlo estimation of $E(\hat{C}_{boot})$ for a given design matrix $\boldsymbol{X}$

entails randomly generating (i) the bootstrap sampling matrix $\boldsymbol{S}_b$ and (ii) the error vector $\boldsymbol{\varepsilon}$. However, since

$$E(\hat{C}_{boot}) = E\left(\frac{1}{B}\sum_{b=1}^{B}\frac{\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_b^*\|^2}{\hat{\sigma}_b^{*2}}\right) - n$$
$$= E\left[\frac{1}{B}\sum_{b=1}^{B}E\left(\frac{\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_b^*\|^2}{\hat{\sigma}_b^{*2}}\,\middle|\,\boldsymbol{S}_b\right)\right] - n, \tag{26}$$

we can eliminate the second source of simulation error by means of an analytic expression for the inner expectation in (26). We do this in two steps: expressing the fractional expression in (26) as a quotient of quadratic forms, and exploiting a formula for the expectation of such a quotient.

Step 1. It is easily seen that $\boldsymbol{S}_b^T\boldsymbol{S}_b = \boldsymbol{D}_b \equiv \mathrm{diag}(r_1^b, r_2^b, \ldots, r_n^b)$, where r_k^b is the number of occurrences of k in the bth bootstrap sample. If fewer than p of $r_1^b, r_2^b, \ldots, r_n^b$ are positive, i.e., if the bootstrap sample contains fewer than p distinct cases, then $\hat{\sigma}_b^{*2} = 0$ and EIC is undefined. On the other hand, if the bootstrap sample contains at least p distinct cases, then $\boldsymbol{X}^T\boldsymbol{D}_b\boldsymbol{X}$ would ordinarily be nonsingular. Assuming this to be the case, one can show that $\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_b^*\|^2/\hat{\sigma}_b^{*2} = (n\boldsymbol{y}^T\boldsymbol{Q}_b^T\boldsymbol{Q}_b\boldsymbol{y})/(\boldsymbol{y}^T\boldsymbol{Q}_b^T\boldsymbol{D}_b\boldsymbol{Q}_b\boldsymbol{y})$, where $\boldsymbol{Q}_b = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{D}_b\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{D}_b$. If $\boldsymbol{X}$ represents a correct model, i.e. (1) holds, then since $\boldsymbol{Q}_b\boldsymbol{X} = \boldsymbol{0}$, we obtain

$$\|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_b^*\|^2/\hat{\sigma}_b^{*2} = \frac{n\boldsymbol{\varepsilon}^T\boldsymbol{Q}_b^T\boldsymbol{Q}_b\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T\boldsymbol{Q}_b^T\boldsymbol{D}_b\boldsymbol{Q}_b\boldsymbol{\varepsilon}}. \tag{27}$$

Step 2. If $\boldsymbol{\varepsilon}$ is a vector of n IID normal variates with mean zero, $\boldsymbol{A}$ is a symmetric $n\times n$ matrix, and $\boldsymbol{B}$ is a positive semidefinite $n\times n$ matrix with singular value decomposition $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$, then by Theorem 6 of Magnus (1986),

$$E\left(\frac{\boldsymbol{\varepsilon}^T\boldsymbol{A}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T\boldsymbol{B}\boldsymbol{\varepsilon}}\right) = \int_0^\infty \left|(\boldsymbol{I}+2t\boldsymbol{\Lambda})^{-1/2}\right| \mathrm{tr}\left[(\boldsymbol{I}+2t\boldsymbol{\Lambda})^{-1}\boldsymbol{P}^T\boldsymbol{A}\boldsymbol{P}\right] dt, \tag{28}$$

provided the expectation exists. By (27), the inner expectation in (26) is given by this integral with $\boldsymbol{A} = n\boldsymbol{Q}_b^T\boldsymbol{Q}_b, \boldsymbol{B} = \boldsymbol{Q}_b^T\boldsymbol{D}_b\boldsymbol{Q}_b$.

One can thus estimate $E(\hat{C}_{boot})$ for a given $\boldsymbol{X}$ by numerically computing the integral (28) for each of a large set of bootstrap sampling matrices $\boldsymbol{S}_b$. In our simulation, for each (n,p), we generated 300 design matrices $\boldsymbol{X}$ by sampling each element independently from the standard normal distribution, and drew a single bootstrap sample for each $\boldsymbol{X}$.

Appendix C: An explicit expression for df_p

We attempt here to provide some intuition regarding df_p. By (19), $C^*(df_p) = \hat{C}^*_{ad,CV}(p)$; (23) then leads to

$$df_p = n_t - \sqrt{\frac{n(n_t+1)(n_t-2)}{n+\hat{C}^*_{ad,CV}(p)}} - 2,$$

or alternatively

$$df_p = \frac{p}{\sqrt{\kappa}} + \left(1-\frac{1}{\sqrt{\kappa}}\right)(n_t-2), \tag{29}$$

where $\kappa = \frac{n+\hat{C}^*_{ad,CV}(p)}{n+C^*(p)}$, i.e., the ratio of $\sum_{k=1}^K \|\boldsymbol{y}_k - \boldsymbol{X}_{k;M_{-k}(p)}\hat{\boldsymbol{\beta}}_{-k;M_{-k}(p)}\|^2/\hat{\sigma}^2_{-k;M_{-k}(p)}$ to $E(\sum_{k=1}^K \|\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{-k}\|^2/\hat{\sigma}^2_{-k})$ as given by (22). The convex combination formula (29) makes it clear that $df_p \geq p$ if and only if $\hat{C}^*_{ad,CV}(p) \geq C^*(p)$, as would ordinarily be case.

Appendix D: Proofs of Theorems 1 and 2

D.1 Theorem 1

We suppress the conditioning on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ in what follows. Since $\hat{\boldsymbol{\beta}}_{-k}$ and $\hat{\sigma}^2_{-k}$ are independent, so are the numerator and denominator on the left side of (21). Therefore, since $\hat{\sigma}^2_{-k} \sim \sigma^2 \chi^2_{n_t-p}/n_t$,

$$E\left(\sum_{k=1}^K \frac{\|\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{-k}\|^2}{\hat{\sigma}^2_{-k}}\right) = \sum_{k=1}^K E\left(\frac{1}{\hat{\sigma}^2_{-k}}\right) E(\|\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{-k}\|^2) = \frac{n_t \sum_{k=1}^K E(\|\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{-k}\|^2)}{\sigma^2(n_t - p - 2)}.$$

It therefore suffices to show that

$$\sum_{k=1}^K E(\|\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{-k}\|^2) = \sigma^2 \sum_{k=1}^K \mathrm{tr}[(\boldsymbol{I}_{n_v} - \boldsymbol{H}_{kk})^{-1}]. \tag{30}$$

To prove (30) we use the identity

$$\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{-k} = (\boldsymbol{I}_{n_v} - \boldsymbol{H}_{kk})^{-1}(\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}),$$

the generalization to K-fold CV of a well-known result for leave-one-out CV (e.g., Reiss et al., 2010). Combining this with $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\varepsilon}$ implies that the left side of (30) equals $E[\boldsymbol{\varepsilon}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{B}(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\varepsilon}] = \sigma^2 \mathrm{tr}[\boldsymbol{B}(\boldsymbol{I}_n - \boldsymbol{H})]$ where

$$\boldsymbol{B} = \begin{pmatrix} (\boldsymbol{I}_{n_v} - \boldsymbol{H}_{11})^{-2} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{I}_{n_v} - \boldsymbol{H}_{22})^{-2} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \ldots & \ldots & (\boldsymbol{I}_{n_v} - \boldsymbol{H}_{KK})^{-2} \end{pmatrix}.$$

Clearly $\mathrm{tr}[\boldsymbol{B}(\boldsymbol{I}_n - \boldsymbol{H})] = \sum_{k=1}^K \mathrm{tr}[(\boldsymbol{I}_{n_v} - \boldsymbol{H}_{kk})^{-1}]$, so (30) follows.

D.2 Theorem 2

Since $E\left(\sum_{k=1}^K \|\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{-k}\|^2/\hat{\sigma}^2_{-k}\right) = E\left[E\left(\sum_{k=1}^K \|\boldsymbol{y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_{-k}\|^2/\hat{\sigma}^2_{-k}\,\middle|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\right)\right]$, (21) implies that it suffices to prove

$$\sum_{k=1}^K E\left[\mathrm{tr}\{(\boldsymbol{I}_{n_v} - \boldsymbol{H}_{kk})^{-1}\}\right] = \frac{n(n_t+1)(n_t-2)}{n_t(n_t-p-2)}. \tag{31}$$

By the Sherman-Morrison-Woodbury lemma (e.g., Harville, 2008) and the fact that $\boldsymbol{H}_{kk} = \boldsymbol{X}_k(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_k$,

$$\begin{aligned} \mathrm{tr}[(\boldsymbol{I}_{n_v} - \boldsymbol{H}_{kk})^{-1}] &= \mathrm{tr}\left[\boldsymbol{I}_{n_v} + \boldsymbol{X}_k(\boldsymbol{X}_{-k}^T\boldsymbol{X}_{-k})^{-1}\boldsymbol{X}_k^T\right] \\ &= n_v + \sum_{i \in V_k} \boldsymbol{x}_i^T(\boldsymbol{X}_{-k}^T\boldsymbol{X}_{-k})^{-1}\boldsymbol{x}_i \\ &= n_v + \sum_{i \in V_k} (1 \quad \tilde{\boldsymbol{x}}_i^T) \begin{pmatrix} n_t & n_t\bar{\tilde{\boldsymbol{x}}}_{-k}^T \\ n_t\bar{\tilde{\boldsymbol{x}}}_{-k} & \tilde{\boldsymbol{X}}_{-k}^T\tilde{\boldsymbol{X}}_{-k} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ \tilde{\boldsymbol{x}}_i \end{pmatrix}, \end{aligned}$$

where $V_k = \{i : \boldsymbol{x}_i$ belongs to the kth validation set$\}$, $\tilde{\boldsymbol{X}}_{-k}$ comprises the last $p-1$ columns of $\boldsymbol{X}_{-k}$, and $\bar{\tilde{\boldsymbol{x}}}_{-k}$ is the vector of column means of $\tilde{\boldsymbol{X}}_{-k}$. By a standard formula for the inverse of a partitioned matrix, the above yields

$$\begin{aligned}\mathrm{tr}[(\boldsymbol{I}_{n_v} - \boldsymbol{H}_{kk})^{-1}] = n_v + \sum_{i \in V_k} \Big[& 1/n_t + \bar{\tilde{\boldsymbol{x}}}_{-k}^T (\tilde{\boldsymbol{X}}_{-k}^{cT} \tilde{\boldsymbol{X}}_{-k}^{c})^{-1} \bar{\tilde{\boldsymbol{x}}}_{-k} \\ & - 2\tilde{\boldsymbol{x}}_i^T (\tilde{\boldsymbol{X}}_{-k}^{cT} \tilde{\boldsymbol{X}}_{-k}^{c})^{-1} \bar{\tilde{\boldsymbol{x}}}_{-k} + \tilde{\boldsymbol{x}}_i^T (\tilde{\boldsymbol{X}}_{-k}^{cT} \tilde{\boldsymbol{X}}_{-k}^{c})^{-1} \tilde{\boldsymbol{x}}_i \Big],\end{aligned}$$

where $\tilde{\boldsymbol{X}}_{-k}^{c}$ is the column-centered version of $\tilde{\boldsymbol{X}}_{-k}$. If $\bar{\tilde{\boldsymbol{x}}}_{-k} = \boldsymbol{\mu} + \bar{\boldsymbol{\eta}}_{-k}$ and $\tilde{\boldsymbol{x}}_i = \boldsymbol{\mu} + \boldsymbol{\eta}_i$ where $\boldsymbol{\mu} \in \mathcal{R}^{p-1}$ is the mean of the predictor distribution, then some algebra leads to

$$\begin{aligned}E\left[\mathrm{tr}\{(\boldsymbol{I}_{n_v} - \boldsymbol{H}_{kk})^{-1}\}\right] = n_v + \sum_{i \in V_k} \Big[& 1/n_t + E\{\bar{\boldsymbol{\eta}}_{-k}^T (\tilde{\boldsymbol{X}}_{-k}^{cT} \tilde{\boldsymbol{X}}_{-k}^{c})^{-1} \bar{\boldsymbol{\eta}}_{-k}\} \\ & + E\{\boldsymbol{\eta}_i^T (\tilde{\boldsymbol{X}}_{-k}^{cT} \tilde{\boldsymbol{X}}_{-k}^{c})^{-1} \boldsymbol{\eta}_i\} \Big].\end{aligned}$$

The two quadratic forms above are distributed as $\frac{1}{n_t(n_t-1)}$ times the Hotelling $T^2(p, n_t-1)$ distribution and as $\frac{1}{n_t-1}$ times the $T^2(p, n_t-1)$ distribution, respectively. Hence

$$E\left[\mathrm{tr}\{(\boldsymbol{I}_{n_v} - \boldsymbol{H}_{kk})^{-1}\}\right] = n_v + \sum_{i \in V_k} \left[\frac{1}{n_t} + \left\{ \frac{1}{n_t(n_t-1)} + \frac{1}{n_t - 1} \right\} \frac{(n_t-1)p}{n_t - p - 2} \right].$$

Summing over k leads to (31).

Appendix E: Robustness of (22)

E.1 Expectation of $n + \hat{C}^*_{CV}$ in general

Our formula for $C^*(p)$ is based on (22), which gives the expectation of $n + \hat{C}^*_{CV} = \sum_{k=1}^{K} \|\boldsymbol{y}_k - \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_{-k}\|^2 / \hat{\sigma}^2_{-k}$ assuming IID multivariate normal predictor vectors. Here we consider the robustness of (22) to departures from this distribution, for the most popular form of CV: leave-one-out CV, i.e., $K = n$. In this case the right-hand expressions in (21) and (22) reduce to $\frac{n-1}{n-p-3} \sum_{i=1}^{n} \frac{1}{1-h_{ii}}$ and $\frac{n^2(n-3)}{(n-p-3)^2}$, respectively. Let $r_{n,p}(\boldsymbol{X})$ be the ratio of these two values, i.e.,

$$r_{n,p}(\boldsymbol{X}) = \frac{(n-1)(n-p-3)}{n^2(n-3)} \sum_{i=1}^{n} \frac{1}{1-h_{ii}}. \tag{32}$$

The ratio of the actual expectation of $n + \hat{C}^*_{CV}$ to that given by Theorem 2 is then $\int r_{n,p}(\boldsymbol{X}) d\boldsymbol{X}$, where the integration is with respect to the distribution of the random matrix $\boldsymbol{X}$ (if one wishes to view $\boldsymbol{X}$ as fixed, this distribution becomes a point mass). We then have the following result.

Proposition 1 *Let $h_{max} = h_{max}(\boldsymbol{X})$ be the largest diagonal element of $\boldsymbol{H}$. If the assumptions of Theorem 2 hold and $h_{max} < 1$, then*

$$\frac{(n-1)(n-p-3)}{(n-3)(n-p)} \leq r_{n,p}(\boldsymbol{X}) \leq \frac{c_X (n-1)(n-p-3)}{n(n-3)},$$

where $c_X = \min\left\{ \frac{1}{1-h_{max}},\ 1 + \frac{p}{n(1-h_{max})^2},\ 1 + \frac{p}{n} + \frac{p h_{max}}{n(1-h_{max})^3} \right\}$.

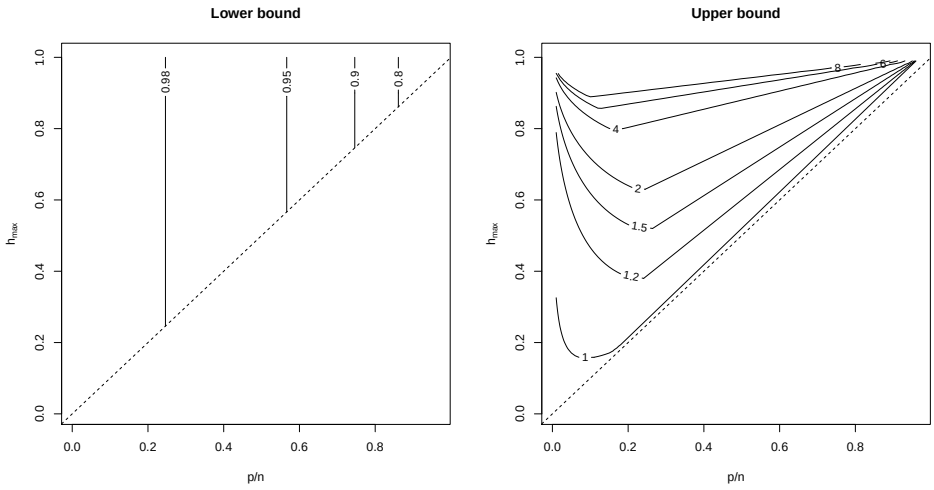

Fig. 9 Contour plots, for $n = 100$, of the bounds imposed by Proposition 1 on the ratio $r_{n,p}(\boldsymbol{X})$. Note that points below the line of identity are excluded, since h_{max} must be at least p/n.

We can gain insight into the practical implications of Proposition 1 by plotting the above bounds on $r_{n,p}(\boldsymbol{X})$ with respect to p/n and h_{max}. The bounds depend only weakly on n; Figure 9 displays them for $n = 100$. For p not too large, the lower bound is just slightly below 1, and the random quantity h_{max} will ordinarily have most of its mass in the region where the upper bound is also near 1; hence $\int r_{n,p}(\boldsymbol{X})d\boldsymbol{X} \approx 1$, i.e., (22) should provide a good approximation. The bounds deviate markedly from 1 only when either p/n or h_{max} is very high, i.e., either the model dimension is very large or there are inordinately high-leverage observations—two scenarios in which the entire enterprise of linear modeling is generally unreliable.

E.2 Proof of Proposition 1

By (32), it suffices to show that

$$\frac{1}{1-p/n} \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1-h_{ii}} \leq c_X .$$

The first inequality holds since the first expression is the harmonic mean, whereas the second expression is the arithmetic mean, of $\frac{1}{1-h_{11}}, \ldots, \frac{1}{1-h_{nn}}$. The second inequality is derived from the following three inequalities:

$$\sum_{i=1}^{n}\frac{1}{1-h_{ii}} \leq \frac{n}{1-h_{max}};$$

$$\begin{aligned}\sum_{i=1}^{n}\frac{1}{1-h_{ii}} &= \sum_{i=1}^{n}\left[1+\frac{h_{ii}}{(1-h_{ii}^{*})^2}\right] \text{ where } 0 < h_{ii}^{*} < h_{ii} \text{ for each } i \\ &< n + \frac{p}{(1-h_{max})^2};\end{aligned}$$

$$\sum_{i=1}^{n} \frac{1}{1-h_{ii}} = \sum_{i=1}^{n} \left[1 + h_{ii} + \frac{h_{ii}^2}{(1-h_{ii}^{**})^3} \right] \text{ where } 0 < h_{ii}^{**} < h_{ii} \text{ for each } i$$
$$< n + p + \frac{p h_{max}}{(1-h_{max})^3}.$$

Acknowledgements The first author's research is supported in part by National Science Foundation grant DMS-0907017. The authors thank Mike Milham, Eva Petkova, Thad Tarpey, Lee Dicker and Tao Zhang, for illuminating discussions; Zarrar Shehzad, for assistance with the functional connectivity data; and the Associate Editor and referee, whose incisive comments led to major improvements in the paper.

References

1. Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (eds B. N. Petrov and F. Csàki), pp. 267–281. Budapest: Akademiai Kiàdo.
2. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
3. Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34, 537–541.
4. Cerdeira, J. O., Duarte Silva, P., Cadima, J., and Minhoto, M. (2009). subselect: Selecting variable subsets. R package version 0.10-1. http://CRAN.R-project.org/package=subselect
5. Davison, A. C. (2003). *Statistical Models*. Cambridge: Cambridge University Press.
6. Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78, 316–331.
7. Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation (with discussion). *Journal of the American Statistical Association*, 99, 619–642.
8. Efron, B., and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92, 548–560.
9. Foster, D. P., and George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, 22, 1947–1975.
10. George, E. I., and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87, 731–747.
11. Harville, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.
12. Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in Statistics: Theory and Methods*, 17, 588–607.
13. Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12, 69–82.
14. Hurvich, C. M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297–307.
15. Ishiguro, M., Sakamoto, Y., and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics*, 49, 411–434.
16. Konishi, S., and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83, 875–890.
17. Konishi, S., and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York: Springer.
18. Lawless, J. F., and Singhal, K. (1978). Efficient screening of nonnormal regression models. *Biometrics*, 34, 318–327.
19. Lumley, T., using Fortran code by A. Miller (2009). `leaps`: regression subset selection. R package version 2.9, http://CRAN.R-project.org/package=leaps

20. Luo, X., Stefanski, L. A., and Boos, D. D. (2006). Tuning variable selection procedures by adding noise. *Technometrics*, 48, 165–175.
21. Magnus, J. R. (1986). The exact moments of a ratio of quadratic forms in normal variables. *Annales d'Économie et de Statistique*, 4, 95–109.
22. Miller, A. (2002). *Subset Selection in Regression*, 2nd ed. Boca Raton: Chapman & Hall/CRC.
23. Pan, W., and Le, C. T. (2001). Bootstrap model selection in generalized linear models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6, 49–61.
24. R Development Core Team (2010). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.
25. Reiss, P. T., Huang, L., and Mennes, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *International Journal of Biostatistics*, 6, article 28.
26. Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press.
27. Shao, J. (1996). Bootstrap model selection. *Journal of the American Statistical Association*, 91, 655–665.
28. Shen, X., and Ye, J. (2002). Adaptive model selection. *Journal of the American Statistical Association*, 97, 210–221.
29. Stark, D. E., Margulies, D. S., Shehzad, Z., Reiss, P. T., Kelly, A. M. C., Uddin, L. Q., Gee, D., Roy, A. K., Banich, M. T., Castellanos, F. X., and Milham, M. P. (2008). Regional variation in interhemispheric coordination of intrinsic hemodynamic fluctuations. *Journal of Neuroscience*, 28, 13754–13764.
30. Stein, J. L., Wiedholz, L. M., Bassett, D. S., Weinberger, D. R., Zink, C. F., Mattay, V. S., and Meyer-Lindenberg, A. (2007). A validated network of effective amygdala connectivity. *NeuroImage*, 36, 736–745.
31. Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communicatons in Statistics: Theory and Methods*, 7, 13–26.
32. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.
33. Tibshirani, R., and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, Series B*, 61, 529–546.
34. Watson, D., Weber, K., Assenheimer, J. S., Clark, L. A., Strauss, M. E., and McCormick, R. A. (1995). Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. *Journal of Abnormal Psychology*, 104, 3–14.
35. Wood, S. N. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, 15, 1126–1133.
36. Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman and Hall/CRC.
37. Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93, 120–131.