

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/33204> holds various files of this Leiden University dissertation

**Author:** Ommen, Thijs van

**Title:** Better predictions when models are wrong or underspecified

**Issue Date:** 2015-06-10

## Chapter 2

# Extra-Sample Prediction in Linear Models

Methods for combining predictions from a number of models in a supervised learning setting must somehow estimate/predict the quality of a model's predictions at unknown future inputs. Many of these methods (often implicitly) make the assumption that the test inputs are identical to the training inputs, which is seldom reasonable. By failing to take into account that prediction will generally be harder for test inputs that did not occur in the training set, this can sometimes lead to the selection of too complex models.

Based on a novel, unbiased expression for KL divergence, we propose XAIC and its special case FAIC as versions of AIC intended for prediction that use different degrees of knowledge of the test inputs. Both methods substantially differ from and may outperform all the known versions of AIC *even when the training and test inputs are i.i.d.*, and are especially useful for deterministic inputs and under covariate shift. Our experiments on linear models suggest that if the test and training inputs differ substantially, then XAIC and FAIC predictively outperform AIC, BIC and several other methods including Bayesian model averaging.

*Terminology* In this chapter we freely use machine learning terminology; let us quickly compare the main notions to their counterpart in statistics. By *supervised* learning problems, we mean problems such as regression and classification in which we observe a *training set* (i.e. a sample) of  $(X, Y)$  pairs, where  $X$  are *inputs* (covariates) and  $Y$  is the *output variable*, and we are interested in learning, based on the sample, a (set of) probability distribution(s) or a (set of) decision rule(s) that can help us to make predictions of  $Y$  given *test input*  $X$ .

### 2.1 Introduction

In the statistical problem of model selection, we are given a set of models  $\{\mathcal{M}_i \mid i \in \mathcal{I}\}$ , each of the form  $\mathcal{M}_i = \{g_i(\cdot \mid \theta) \mid \theta \in \Theta_i\}$ , where the  $g_i(\cdot \mid \theta)$  are

density functions on (sequences of) data. We wish to use one of these models to explain our data and/or to make predictions of future data, but do not know which model explains the data best. It is well known that simply selecting the model containing the maximum likelihood distribution from among all the models leads to overfitting, so any expression of the quality of a model must somehow avoid this problem. One way to do this is by estimating each model's ability to predict *unseen* data (this will be made precise below). This approach is used by many methods for model selection, including cross-validation, AIC (Akaike, 1973) and its many variants, Gelfand and Ghosh's  $D_k$  (1998), BPIC (Ando, 2007), and WAIC (Watanabe, 2010). However, none of these methods takes into account that for supervised learning problems, the generalization error being estimated will vary with the test input variables. Instead, they implicitly assume that the test inputs will be *identical* to the training inputs.

In this chapter, we derive an estimate of the generalization error that does take the input data into account, and use this to define a new model selection criterion XAIC, its special case FAIC, and the variants XAIC<sub>C</sub> and FAIC<sub>C</sub> (small sample corrections). We use similar assumptions as AIC, and thus our methods can be seen as relatives of AIC that are adapted to supervised learning when the training and test inputs differ. Our experiments show that our methods have excellent predictive performance, better even than Bayesian model averaging in some cases. Also, we show theoretically that AIC's unawareness of input variables leads to a bias in the selected model order, even in the seemingly safe case where the test inputs are drawn from the same distribution as the training inputs. No existing model selection method seems to address this issue adequately, making XAIC and FAIC more than "yet another version of AIC".

It is in fact quite surprising that, more than forty years after its original invention, all the forms of AIC currently in use are biased in the above sense, and in theoretical analyses, conditional model selection methods are often even compared on a new point  $x$  constrained to be one of the  $x$  values in the training data (see e.g. Yang, 2005), even though in most practical problems, a new point  $x$  will *not* be drawn from this empirical training data distribution, but rather should be regarded as falling in one of the three cases considered in this chapter: (a) it is drawn from the same distribution as the training data (but not necessarily equal to one of the training inputs); (b) it is drawn from a different distribution (covariate shift); (c) it is set to a fixed, observable value, usually not in the training set, but the process that gave rise to this value may not be known.

### 2.1.1 Goals of model selection

When choosing among or combining predictions from different models, one can have different goals in mind. Whereas BIC and BMS (Bayesian model selection) focus on finding the most probable model, methods like AIC, cross-validation and SRM (structural risk minimization; Vapnik, 1998) aim to find the model that leads to the best *predictions* of future data. While AIC and cross-

validation typically lead to predictions that converge faster to optimal in the sense of KL divergence than those of BIC and BMS, it is also well-known that, unlike BIC and BMS, such methods are not statistically consistent (i.e. they do not find the smallest submodel containing the truth with probability 1 as  $n \rightarrow \infty$ ); there is an inherent conflict between these two goals, see for example Yang (2007a); Van Erven et al. (2007, 2012). Like AIC, the XAIC and FAIC methods developed here aim for predictive optimality rather than consistency, thus, if consistency is the main concern, they should not be used. We also stress at the outset that, unlike most other model selection criteria, the model selected by FAIC may *depend* on the new  $x$  whose corresponding  $y$  value is to be predicted; for different  $x$ , a different model may be selected based on the same training data. Since — as in many other model selection criteria — our goal is predictive accuracy rather than ‘finding the true model’, and since the dependence on the test  $x$  helps us to get substantially better predictions, we are not worried by this dependency.

FAIC thus cannot be said to select a ‘single’ model for a given training set — it merely outputs a *function* from  $x$  values to models. As such, it is more comparable with BMA (Bayesian model *averaging*) rather than BMS (*selection*). BMA is of course a highly popular method for data prediction; like FAIC, it adapts its predictions to the test input  $x$  (as we will see, FAIC tends to select a simpler model if there are not many training points near  $x$ ; BMA predicts with a larger variance if there are not many training points near  $x$ ). BMA leads to the optimal predictions in the idealized setting where one takes expectation under the prior (i.e., in frequentist terms, we imagine nature to draw a model, and then a distribution within the chosen model, both from the prior used in BMA, and then data from the drawn distribution), and usually performs very well in practice as well. It is of considerable interest then that our XAIC and FAIC outperform Bayes by a fair margin in some of our experiments in Section 2.5.

### 2.1.2 In-sample and extra-sample error

Many methods for model selection work by computing some estimate of how well each model will do at predicting unseen data. This generalization error may be defined in various ways, and methods can further vary in the assumptions used to find an estimate. AIC (Akaike, 1973) is based on the expression for the generalization error

$$-2 \mathbf{E}_{\mathbf{U}} \mathbf{E}_{\mathbf{V}} \log g_i(\mathbf{V} \mid \hat{\theta}_i(\mathbf{U})), \quad (2.1)$$

for model  $\mathcal{M}_i = \{g_i(\cdot \mid \theta) \mid \theta \in \Theta_i\}$ , where  $\hat{\theta}_i(\mathbf{U})$  denotes the element of  $\Theta_i$  which maximizes the likelihood of data  $\mathbf{U}$ , and where both random variables are independent samples of  $n$  data points each, both following the true distribution of the data. (In this chapter, we use capitals to denote sequences of data points, and boldface for random variables. Throughout this chapter,  $\log$  denotes the natural logarithm.) Up to an additive term which is the same for all models, the inner expectation is the KL divergence from the true distribution to  $g_i(\cdot \mid \hat{\theta}_i(\mathbf{U}))$ . An interpretation of (2.1) is that we first estimate

the model's parameters using a random sample  $\mathbf{U}$ , then judge the quality of this estimate by looking at its performance on an independent, identically distributed sample  $\mathbf{V}$ . AIC then works by estimating (2.1) for each model by the asymptotically unbiased estimator

$$-2 \log g_i(\mathbf{U} \mid \hat{\theta}(\mathbf{U})) + 2k \quad (2.2)$$

where  $k$  is the number of parameters in the model, and selecting the model minimizing this estimate. Thus AIC selects the model whose maximum likelihood estimate is expected to be closest to the truth in terms of KL divergence. In the sequel, we will consider only one model at a time, and therefore omit the model index.

In supervised learning problems such as regression and classification, the data points consist of two parts  $u_i = (x_i, y_i)$ , and the models are sets of distributions on the *output variable*  $\mathbf{y}$  conditional on the *input variable*  $x$  (which may or may not be random). We call these *conditional* models. The conditionality expresses that we are not interested in explaining the behaviour of  $x$ , only that of  $\mathbf{y}$  given  $x$ . Then (2.1) can be adapted in two ways: in the terminology of Hastie et al. (2001), as the *extra-sample error*

$$-2 \mathbf{E}_{\mathbf{Y}|\mathbf{X}} \mathbf{E}_{\mathbf{Y}'|\mathbf{X}'} \log g(\mathbf{Y}' \mid \mathbf{X}', \hat{\theta}(\mathbf{X}, \mathbf{Y})), \quad (2.3)$$

and, replacing both  $\mathbf{X}$  and  $\mathbf{X}'$  by a single variable  $\mathbf{X}$ , as the *in-sample error*

$$-2 \mathbf{E}_{\mathbf{Y}|\mathbf{X}} \mathbf{E}_{\mathbf{Y}'|\mathbf{X}} \log g(\mathbf{Y}' \mid \mathbf{X}, \hat{\theta}(\mathbf{X}, \mathbf{Y})), \quad (2.4)$$

where capital letters again denote sequences of data points. Contrary to (2.1), these quantities capture that the expected quality of a prediction regarding  $\mathbf{y}$  may vary with  $x$ .

An example of a supervised learning setting is given by *linear models*. In a linear model, an input variable  $x$  is represented by a *design vector* and a sequence of  $n$  inputs by an  $n \times p$  *design matrix*; with slight abuse of notation, we use  $x$  and  $\mathbf{X}$  to represent these. Then the densities  $g(\mathbf{Y} \mid \mathbf{X}, \beta)$  in the model are Gaussian with mean  $\mathbf{X}\beta$  and covariance matrix  $\sigma^2 I_n$  for some fixed  $\sigma^2$ . Because  $g$  is of the form  $e^{-\text{squared error}}$ , taking the negative logarithm as in (2.1) produces an expression whose main component is a sum of squared errors; the residual sum of squared errors  $\text{RSS}(\mathbf{Y})$  is the minimum for given data, which is attained by the maximum likelihood estimator. Alternatively,  $\sigma^2$  may be another parameter in addition to  $\beta$  if the true variance is unknown.

It is standard to apply ordinary AIC to supervised learning problems. For example, for linear models with fixed variance, (2.2) takes the well-known form

$$\frac{1}{\sigma^2} \text{RSS}(\mathbf{Y}) + 2k. \quad (2.5)$$

But because the standard expression behind AIC (2.1) makes no mention of  $\mathbf{X}$  or  $\mathbf{X}'$ , this corresponds to the tacit assumption that  $\mathbf{X} = \mathbf{X}'$ , so that the in-sample error is being estimated.

However, the extra-sample error is more appropriate as a measure of the expected performance on new data. AIC was intended to correct the bias that results from evaluating an estimator on the data from which it was derived, but because it uses the in-sample error, AIC evaluates estimators on new output data, but old input data. So we see that in supervised problems, a bias similar to the one it was intended to correct is still present in AIC.

### 2.1.3 Contents

The remainder of this chapter is structured as follows. In Section 2.2, we develop our main results about the extra-sample error and propose a new model selection criterion based on this. It involves  $\kappa_{Y'}$ , a term which can be calculated explicitly for linear models; we concentrate on these models in the remainder of the chapter. Special cases of our criterion, including a focused variant, are presented in Section 2.3. In Section 2.4 we discuss the behaviour of our estimate of the extra-sample error, and find that without our modification, AIC's selected model orders are biased. Several experiments on simulated data are described in Section 2.5. Section 2.6 contains some further theoretical discussion regarding Bayesian prediction and covariate shift. Finally, Section 2.7 concludes. Technical regularity conditions and proofs are in the appendix.

## 2.2 Estimating the extra-sample error

In this section, we will derive an estimate for the extra-sample error. Our assumptions will be similar to those used in AIC to estimate the in-sample error; therefore, we start with some preliminaries about the setting of AIC.

### 2.2.1 Preliminaries

In the setting of AIC, the data points are independent but not necessarily identically distributed. The number of data points in  $\mathbf{Y}$  and  $\mathbf{Y}'$  is  $n$ . We define the Fisher information matrix  $I(\theta)$  as  $-\mathbf{E}_{\mathbf{Y}'} \frac{\partial^2}{\partial \theta^2} \log g(\mathbf{Y}' | \theta)$ , and define the conditional Fisher information matrix  $I(\theta | X')$  analogously. We write  $\text{Cov}(\hat{\theta}(X, \mathbf{Y}) | X)$  for the conditional covariance matrix  $\mathbf{E}_{\mathbf{Y}|X}[\hat{\theta}(X, \mathbf{Y}) - \mathbf{E}_{\mathbf{Y}|X} \hat{\theta}(X, \mathbf{Y})][\hat{\theta}(X, \mathbf{Y}) - \mathbf{E}_{\mathbf{Y}|X} \hat{\theta}(X, \mathbf{Y})]^\top$ .

Under standard regularity assumptions, there exists a unique parameter value  $\theta_0$  that minimizes the KL divergence from the true distribution, and this is what  $\hat{\theta}(\mathbf{Y})$  converges to. Under this and other (not very restrictive) regularity assumptions (Shibata, 1989), it can be shown that (Burnham and Anderson, 2002)

$$-2 \log g(\mathbf{Y} | \hat{\theta}(\mathbf{Y})) + 2 \widehat{\text{tr}} \{ I(\theta_0) \text{Cov}(\hat{\theta}(\mathbf{Y})) \} \quad (2.6)$$

(where  $\widehat{\text{tr}}$  represents an appropriate estimator of that trace) is an asymptotically unbiased estimator of (2.1). The model selection criterion TIC (Takeuchi's information criterion) selects the model which minimizes (2.6).

The estimator of the trace term that TIC requires has a large variance, making it somewhat unreliable in practice. AIC uses the very simple estimate  $2k$  for TIC's trace term. This estimate is generally biased except when the true data-generating distribution is in the model, but obviously has 0 variance. Also, if some models are more misspecified than others, those models will have a worse log-likelihood. This term in AIC grows linearly in the sample size, so that asymptotically, those models will be disqualified by AIC. Thus AIC selects good models even when its penalty term is biased due to misspecification of the models.

This approach corresponds to making the following assumption in the derivation leading to AIC's penalty term:

**Assumption 2.1.** *The model contains the true data-generating distribution.*

It follows that  $\theta_0$  specifies this distribution. We emphasize that this assumption is only required for AIC's derivation and does not mean that AIC necessarily works badly if applied to misspecified models. Under this assumption, the two matrices in (2.6) cancel, so the objective function becomes (2.2), the standard formula for AIC (Burnham and Anderson, 2002).

We now move to supervised learning problems, where the true distribution of the data and the distributions  $g$  in the models are conditional distributions of output values given input values. In this setting, the data are essentially i.i.d. in the sense that  $g(\mathbf{Y} \mid X, \theta) = \prod_{i=1}^n g(\mathbf{y}_i \mid x_i, \theta)$ . That is, the outputs are independent given the inputs, and if two input variables are equal, the corresponding output variables are identically distributed. Also, the definition of  $\theta_0$  would need to be modified to depend on the training inputs, but since Assumption 2.1 now implies that  $g(\mathbf{y} \mid x, \theta_0)$  defines the true distribution of  $\mathbf{y}$  given  $x$  for all  $x$ , we can take this as the definition of  $\theta_0$  for supervised learning when Assumption 2.1 holds.

For supervised learning problems, AIC and TIC silently assume that  $X'$  either equals  $X$  or will be drawn from its empirical distribution. We want to remove this assumption.

## 2.2.2 Main results

We will need another assumption:

**Assumption 2.2.** *For training data  $(X, \mathbf{Y})$  and (unobserved) test data  $(X', \mathbf{Y}')$ ,*

$$-\frac{1}{n} \mathbf{E}_{\mathbf{Y} \mid X} \log g(\mathbf{Y} \mid X, \theta_0) = -\frac{1}{n'} \mathbf{E}_{\mathbf{Y}' \mid X'} \log g(\mathbf{Y}' \mid X', \theta_0),$$

where  $n$  and  $n'$  denote the number of data points in  $X$  and  $X'$ , respectively.

This assumption ensures that the log-likelihood on the test data can be estimated from the training data. If  $X$  and  $X'$  are random and mutually i.i.d., this is automatically satisfied when the expectations are taken over these inputs as well. While this assumption of randomness is standard in machine

learning, there are other situations where  $X$  and  $X'$  are not random and Assumption 2.2 holds nevertheless. For instance, this is the case if  $g(\mathbf{y} \mid x, \theta)$  is such that  $\mathbf{y}_i = f_\theta(x_i) + \mathbf{z}_i$ , where the noise terms  $\mathbf{z}_i$  are zero-mean and i.i.d. (their distribution may depend on  $\theta$ ). This additive noise assumption is common in regression-like settings. Then Assumption 2.1 implies that Assumption 2.2 holds for all  $X, X'$ .

To get an estimator of the extra-sample error (2.3), we do not make any assumptions about the process generating  $X$  and  $X'$  but leave the variables free. We allow  $n \neq n'$ .

**Theorem 2.1.** *Under Assumptions 2.1 and 2.2 and some standard regularity conditions (detailed in Assumption 2.3 in the appendix), and for  $n'$  either constant or growing with  $n$ ,*

$$\begin{aligned} -2 \frac{n}{n'} \mathbf{E}_{\mathbf{Y}|X} \mathbf{E}_{\mathbf{Y}'|X'} \log g(\mathbf{Y}' \mid X', \hat{\theta}(X, \mathbf{Y})) \\ = -2 \mathbf{E}_{\mathbf{Y}|X} \log g(\mathbf{Y} \mid X, \hat{\theta}(X, \mathbf{Y})) + k + \kappa_{X'} + o(1), \end{aligned} \quad (2.7)$$

where  $\kappa_{X'} = \frac{n}{n'} \text{tr} \{ I(\theta_0 \mid X') \text{Cov}(\hat{\theta}(X, \mathbf{Y}) \mid X) \}$ .

Moreover, if the true conditional distribution of  $\mathbf{Y}$  given  $X$  is Gaussian with fixed variance and the conditional distributions in the models are also Gaussian with that same variance (as is the case in linear models with known variance), then the above approximation becomes exact.

We wish to use (2.7) as a basis for model selection. To do this, first note that (2.7) can be estimated from our training data using

$$-2 \log g(\mathbf{Y} \mid X, \hat{\theta}(X, \mathbf{Y})) + k + \kappa_{X'}. \quad (2.8)$$

Theorem 2.1 expresses that this is an asymptotically unbiased estimator of the extra-sample error. We see that the difference with standard AIC (2.2) is that the penalty  $2k$  has been replaced by  $k + \kappa_{X'}$ . We propose to use (2.8) as the basis for a new model selection criterion *extra-sample AIC* (XAIC), which chooses the model that minimizes an estimator of (2.8). What remains for this is to evaluate  $\kappa_{X'}$ , which may depend on the unknown true distribution, and on the test set through  $X'$ . This is done below for the case of linear models; further intuition about  $\kappa_{X'}$  for linear models and a single test input  $x'$  is given in Section 2.4 (where it is related to a norm of  $x'$ ) and Section 2.6.1 (where it is given a Bayesian interpretation).

### 2.2.3 The $\kappa_{X'}$ and $o(1)$ terms for linear models

If the densities  $g$  are Gaussian, then  $\kappa_{X'}$  does not depend on the unknown  $\theta_0$  because the Fisher information is constant, so no additional estimation is necessary to evaluate it. Thus for a linear model with fixed variance,  $\kappa_{X'}$  becomes

$$\kappa_{X'} = \frac{n}{n'} \text{tr} \left\{ \left[ \frac{1}{\sigma^2} X'^\top X' \right] \left[ \sigma^2 (X^\top X)^{-1} \right] \right\} = \frac{n}{n'} \text{tr} \left[ X'^\top X' (X^\top X)^{-1} \right].$$

If the variance is also to be estimated, it can be easily seen that  $\kappa_{X'}$  will become this value plus one. In that case, the approximation in Theorem 2.1 is not exact (as it is in the known variance case), but the  $o(1)$  term can be evaluated explicitly:

**Theorem 2.2.** *For a linear model with unknown variance,*

$$\begin{aligned} & -2 \frac{n}{n'} \mathbf{E}_{\mathbf{Y}|\mathbf{X}} \mathbf{E}_{\mathbf{Y}'|\mathbf{X}'} \log g(\mathbf{Y}' | \mathbf{X}', \hat{\theta}(\mathbf{X}, \mathbf{Y})) \\ & = -2 \mathbf{E}_{\mathbf{Y}|\mathbf{X}} \log g(\mathbf{Y} | \mathbf{X}, \hat{\theta}(\mathbf{X}, \mathbf{Y})) + k + \kappa_{X'} + \frac{(k + \kappa_{X'})(k + 1)}{n - k - 1}, \end{aligned}$$

where  $\kappa_{X'}$  can be computed from the data and equals  $(n/n') \text{tr}(\mathbf{X}'^\top \mathbf{X}' (\mathbf{X}^\top \mathbf{X})^{-1}) + 1$ , and  $k$  is the number of parameters including  $\sigma^2$ .

Theorem 2.2 presents an extra-sample analogue of the well-known small sample correction  $\text{AIC}_C$  (Hurvich and Tsai, 1989), which is derived similarly and uses a penalty of  $2k + 2k(k + 1)/(n - k - 1)$ . We define  $\text{XAIC}_C$  accordingly. Though the theorem holds exactly only in the specific case described, we believe that the extra penalty term will lead to better results in much more general settings in practice, as is the case with  $\text{AIC}_C$  (Burnham and Anderson, 2002).

## 2.3 Model selection for extra-sample prediction

In this section, we discuss several concrete model selection methods, all based on the  $\text{XAIC}$  formula (2.8) and thus correcting  $\text{AIC}$ 's bias.

### 2.3.1 Nonfocused versions of $\text{XAIC}$

Except in trivial cases, the extra-sample error (2.3) and its estimate (2.8) depend on the test inputs  $\mathbf{X}'$ , so some knowledge of  $\mathbf{X}'$  is required when choosing a model appropriate for extra-sample prediction. In a semi-supervised learning setting where  $\mathbf{X}'$  itself is known at the time of model selection, we could evaluate (2.8) directly for each model. However,  $\mathbf{X}'$  might not yet be known when choosing a model.

If  $\mathbf{X}'$  is not known but its distribution is, we can replace  $\kappa_{X'}$  by its expectation; for i.i.d. inputs, computing this reduces to computing  $\mathbf{E}_{\mathbf{X}'} I(\theta_o | \mathbf{x}')$ .

If the distribution of  $\mathbf{X}'$  is also unknown, we need to estimate it somehow. If it is believed that  $\mathbf{X}$  and  $\mathbf{X}'$  follow the same distribution, the empirical distribution of  $\mathbf{X}$  could be used as an estimate of the distribution of  $\mathbf{X}'$ . Then  $\text{AIC}$  is retrieved as a special case. Section 2.4 will show that this is a bad choice even if  $\mathbf{X}$  and  $\mathbf{X}'$  follow the same distribution, so a smoothed estimate is recommended instead.

Of course, we are not restricted to the case where  $\mathbf{X}$  and  $\mathbf{X}'$  follow similar distributions. In the setting of covariate shift (Sugiyama and Kawanabe, 2012), the distributions are different but known (or can be estimated). This variant of

XAIC is directly applicable to that setting, yielding an unbiased analogue of AIC.

### 2.3.2 Focused model selection

It turns out there is a way to apply (2.8) even when nothing is known about the processes generating  $X$  and  $X'$ . If our goal is prediction, we can set  $X'$  to the single point  $x'$  for which we need to predict the corresponding  $y'$ . Contrary to standard model selection approaches, we thus use  $x'$  already at the stage of model selection, rather than only inside the models. We define the model selection criterion *Focused AIC* (FAIC) as this special case of XAIC, and FAIC<sub>C</sub> as its small sample correction.

A focused model selection method implements the intuition that those test points whose input is farther away from the training inputs should be predicted with more caution; that is, with less complex models. As discussed in Section 2.1.1, methods that optimize predictive performance often are not consistent; this hurts in particular for test inputs far away from the training inputs. We expect that extra-sample adaptations of such methods (like XAIC) are also inconsistent, but that using the focused special case helps to guard against this small chance of large loss.

Choosing a model specifically for the task at hand potentially lets us end up with a model that performs this task much better than a model found by a non-focused model selection method. However, there are situations in which focus is not a desirable property: the mapping from input values to predictions given by a focused model selection method will be harder to interpret than that of a non-focused method, as it is a combination of the models under consideration rather than a single one of them. Thus, if the experimenter's goal is interpretation/transparency, a focused model selection method is not recommended; these methods are best applied when the goal is prediction.

Evaluating the  $x'$ -dependent model selection criterion separately for each  $x'$  leads to a regression curve which in general will not be from any one of the candidate models, but only piecewise so. It will usually have discontinuities where it switches between models. If the models contain only continuous functions and such discontinuities are undesirable, Akaike weights (Akaike, 1979; Burnham and Anderson, 2002) may be used to get a continuous analogue of the FAIC regression curve.

## 2.4 AIC vs. XAIC ( $k$ vs. $\kappa_x$ ) in linear models

Intuitively, the quantity  $\kappa_x$  that appears as a penalty term in the XAIC formula (2.8) expresses a measure of dissimilarity between the test input  $x$  and the training inputs  $X$ . This measure is determined fully by the models and does not have to be chosen some other way. However, its properties are not readily apparent. Because  $\kappa_x$  can be computed exactly for linear models, we investigate some of its properties in that case.

One useful characterization of  $\kappa_x$  is the following: if we express the design vector  $x$  of the test point in a basis that is orthonormal to the empirical measure of the training set  $X$ , then  $\kappa_x = \|x\|^2$ .

For given  $X$ ,  $x$  may exist such that  $\kappa_x$  is either greater or smaller than the number of parameters  $k$ . An example of  $\kappa_x < k$  occurs for the linear model consisting of all linear functions with known variance (so  $k = 2$ ). Then  $\kappa_x$  will be minimized when  $x$  lies at the mean of the input values in the training set, where  $\kappa_x = 1$ .

We will now consider the case where  $X$  and  $x$  are random and i.i.d. We showed that the XAIC expression (2.8) is an unbiased estimator of the extra-sample error. AIC uses  $k$  in place of  $\kappa_x$ , and the above suggests the possibility that maybe the instances where  $\kappa_x > k$  and those where  $\kappa_x < k$  cancel each other out, so that AIC would also be approximately unbiased as an estimate of the extra-sample error. However, the following proposition shows that, except in a trivial case,  $\kappa_x$  is on average greater than  $k$ . This means that in those settings, AIC underestimates the model's extra-sample error.

(We should mention here that if  $X$  and  $x$  are random and mutually i.i.d., then as  $n \rightarrow \infty$ , AIC's bias goes to 0. The bias we show below concerns all finite  $n$ ; additionally, without focus, an extreme  $x$  could result in a very biased AIC value even for large  $n$ .)

**Proposition 2.3.** *Consider a linear model  $\mathcal{M}$  with training inputs  $X$  and test input  $x$  i.i.d. such that  $X^\top X$  is almost surely invertible. Let  $\mathcal{M}'$  be the submodel obtained by removing the final entry from every design vector. Then these models are related by  $E \kappa_x \geq E \kappa_{x'} + 1$ , with strict inequality if  $x$  has at least two entries.*

It follows by induction on  $k$  that for random input data, AIC is biased as an estimate of the extra-sample error except in a special case with  $k = 1$ . Also, the bias becomes worse for larger models. This last fact is distressing, as it shows that when AIC assesses a sequence of nested models, the amount by which it overestimates their generalization ability grows with the model order. Thus the biases in the AIC scores lead to a bias in the selected model order, which was not evident from earlier work.

The XAIC formula (2.8) contains two terms that depend on the data: minus two times the log-likelihood, and the penalty term  $\kappa_{x'}$ . The log-likelihood measures distances between output values and is independent of  $X'$ , while  $\kappa_{x'}$  expresses a property of input values and is largely unaffected by output values; in fact, in linear models its computation does not involve any (estimates based on) output values. Hence the variance of XAIC is no greater than that of AIC when comparing the two on fixed  $X, X'$ , so that XAIC's reduction in bias does not come at the price of an increase in variance. However, focused model selection demands that  $X'$  is *not* held fixed, so that FAIC may have a larger variance than AIC. Similarly, if the distribution of  $X'$  is being estimated as part of applying XAIC, the used estimator's quality will affect the accuracy of the estimated generalization error.

## 2.5 Experiments

We will now experimentally compare XAIC and FAIC (or more precisely, their small-sample corrected versions  $\text{XAIC}_C$  and  $\text{FAIC}_C$ ) to several other model selection methods, in univariate and multivariate problems.

### 2.5.1 Description of experiments

In the univariate experiments, linear models  $\mathcal{M}_1, \dots, \mathcal{M}_7$  with unknown variance were considered. Model  $\mathcal{M}_i$  contained polynomials of degree  $i - 1$  (and so had  $i + 1$  parameters). The input values  $x$  of the training data were drawn from a Gaussian distribution with mean 0 and variance 1, while the output values were generated as  $\mathbf{y}_i = f(x_i) + \mathbf{z}$  with  $\mathbf{z}_i$  i.i.d. Gaussians with mean 0 and variance 0.1, and  $f$  some unknown true function. Given 100 training data points, each of the eight model selection methods under consideration had to select a model. The squared risk  $(\hat{y} - f(x))^2$  of the chosen model's prediction  $\hat{y}$  was computed for each of a range of values of the test point's  $x$ , averaged over 100 draws of the training data. This experiment was performed for two different true functions:  $f_1(x) = x + 2$  and  $f_2(x) = |x|$ .

In the multivariate experiments, each input variable was a vector  $(u_1, \dots, u_6)$ , and the models corresponded to all possible subsets of these 6 variables. Each model also included an intercept and a variance parameter. The true function was given by  $f(u) = 2 + u_1 + 0.1u_2 + 0.03u_3 + 0.001u_4 + 0.003u_5$ , and the additive noise was again Gaussian with variance 0.1. A set of  $n' = 400$  test inputs was drawn from a standard Gaussian distribution, but the training inputs were generated differently in each experiment: from the same Gaussian distribution as the test inputs; from a uniform distribution on  $[-\sqrt{3}, \sqrt{3}]^6$ ; or from a uniform 'spike-and-slab' mixture of two Gaussians with covariance matrices  $(1/5)I_6$  and  $(9/5)I_6$ . Note that all three distributions have the same mean and covariance as the test input distribution, making these mild cases of covariate shift. For the Gaussian training case, we report the results for  $n = 60$  and, after extending the same training set, for  $n = 100$ . Squared risks were averaged over the test set and further over 50 repeats of these experiments.

The experiments used the version of XAIC that is given a distribution of the test inputs, but not the test inputs themselves. In the multivariate experiments, XAIC used the actual (Gaussian) distribution of the test inputs. In the univariate case, two instances of XAIC were evaluated: one for test inputs drawn from the same distribution as the training inputs (standard Gaussian), and another (labelled  $\text{XAIC}_C2$ ) for a Gaussian test input distribution with mean 0 and variance 4.

Bayesian model averaging (BMA) differs from the other methods in that it does not select a single model, but formulates its prediction as a weighted average over them; in our case, its prediction corresponds to the posterior mean over all models. Weighted versions exist of other model selection methods as well, such as Akaike weights (Akaike, 1979; Burnham and Anderson, 2002) for AIC and variants. In our experiments we saw that these usually perform sim-

ilar to but somewhat better than their originals. In our univariate experiments, we decided against reporting these, as they are less standard. However, in the multivariate experiments, the weighted versions were all better than their selection counterparts, so both are reported separately to allow fair comparisons.

In our experiments, BMA used a uniform prior over the models. Within the models, Jeffreys' noninformative prior (for which the selected  $\beta$  would correspond to the maximum likelihood  $\hat{\beta}$  used by other methods) was used for the variable selection experiments; for the polynomial case, it proved too numerically unstable for the larger models, so there BMA uses a weakly informative Gaussian prior (variance  $10^2$  on  $\beta_2, \dots, \beta_7$  with respect to the Hermite polynomial basis, and Jeffreys' prior on  $\sigma^2$ ).

Of the model selection methods included in our experiments, AIC was extensively discussed in Section 2.2.1; as with XAIC and FAIC, we use here the small sample correction AIC<sub>C</sub> (see Section 2.2.3). BIC (Schwarz, 1978) and BMS were mentioned in Section 2.1.1 as methods that attempt to find the most probable model given the data rather than aiming to optimize predictive performance; both are based on BMA, which computes the Bayesian posterior probability of each model. Three other methods were evaluated in our experiments; these are discussed below.

Like AIC, the much more recent focused information criterion (FIC; Claeskens and Hjort, 2003) is designed to make good predictions. Unlike other methods, these predictions are for a *focus parameter* which may be any function of the model's estimate, not just its prediction at some input value (though we only used the latter in our experiments). Unlike FAIC, it uses this focus not just for estimating a model's variance, but also its bias; FAIC on the other hand uses a global estimate of a model's bias based on Assumption 2.2. A model's bias for the focus parameter is evaluated by comparing its estimate to that of the most complex model available.

Another more recent method for model selection is the subspace information criterion (SIC; Sugiyama and Ogawa, 2001), which is applicable to supervised learning problems when our models are subspaces of some Hilbert space of functions, and our objective is to minimize the squared norm. Like FIC, SIC estimates the models' biases by comparing their estimates to that of a larger model, but it includes a term to correct for this large model's variance. In our experiments, we used the corrected SIC (cSIC) which truncates the bias estimate at 0.

Generalized cross-validation (GCV; Golub et al., 1979) can be seen as a computationally efficient approximation of leave-one-out cross-validation for linear models. We included it in our experiments because Leeb (2008) shows that it performs better than other model selection methods when the test input variables are newly sampled.

## 2.5.2 Results

Results from the two univariate experiments are shown in Figures 2.1 and 2.2 (squared risks) and in Table 2.1 (selected models). Squared risk results for the

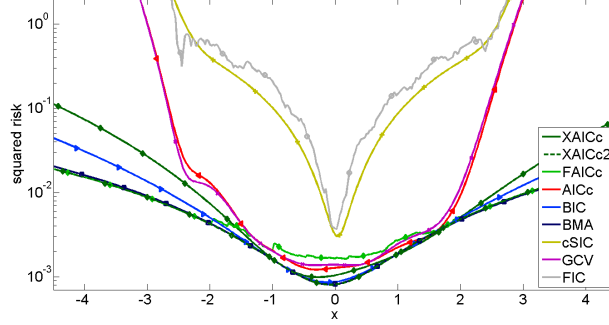


Figure 2.1: Squared risk of different model selection methods as a function of  $x$  when the true function is  $f_1(x) = x + 2$

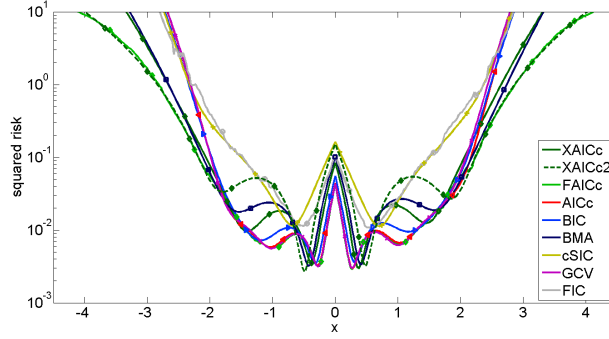


Figure 2.2: Squared risk of different model selection methods as a function of  $x$  when the true function is  $f_2(x) = |x|$

multivariate experiments are given in Table 2.2 for the model selection methods, and in Table 2.3 for the model weighting/averaging variants.

**XAIC and FAIC** The characteristic behaviours of our methods are clearly visible in the univariate experiments. Both instances of XAIC perform well overall in both experiment. Of the two, XAIC<sub>C2</sub> was set up to expect test inputs farther away from the centre. As a result, it selects models more conservatively, and obtains smaller risk at such off-centre test inputs. Its selections were very stable: in both experiments, XAIC<sub>C2</sub> selected the same model in each of the 100 runs.

We see that in the centre of Figure 2.2, the simple model chosen by XAIC<sub>C2</sub> was outperformed by more complex models. FAIC exploits this by choosing a model adaptively for each test input. This resulted in good risk performance at all test inputs.

In the multivariate experiments, FAIC was the best method for the spike-

Table 2.1: Average selected model index per method for  $f_1$  and  $f_2$ , at test inputs  $x' = 0$  and 4 (if different)

| $x'$               |   | $f_1$ | $f_2$ |
|--------------------|---|-------|-------|
| XAIC <sub>C</sub>  |   | 2.10  | 4.57  |
| XAIC <sub>C2</sub> |   | 2.00  | 3.00  |
| FAIC <sub>C</sub>  | 0 | 2.94  | 6.56  |
|                    | 4 | 2.00  | 1.54  |
| AIC <sub>C</sub>   |   | 2.33  | 6.38  |
| BIC                |   | 2.02  | 5.70  |
| BMS                |   | 2.00  | 4.05  |
| cSIC               |   | 2.94  | 4.70  |
| GCV                |   | 2.38  | 6.49  |
| FIC                | 0 | 2.66  | 5.29  |
|                    | 4 | 3.12  | 5.35  |

and-slab training data, where there are pronounced differences in training point density surrounding different test points, so that selecting a different model for each pays off. The performance of XAIC was more reliable overall, comparing very favourably to each of its competitors.

**AIC** Our methods XAIC and FAIC were derived as adaptations of AIC, and share its tendency to go for complex models as soon as there is some indication that their predictions might be worthwhile. This leads to good predictions on average, but also to inconsistency: when a simpler model contains the true distribution, AIC will continue to select more complex models with positive probability, no matter how large  $n$  grows. This may sometimes hurt predictive performance, because the accuracy of the estimated parameter will be smaller for more complex models; for details, we refer to Yang (2007a); Van Erven et al. (2007, 2012). XAIC makes a better assessment of the generalization error, even when the training and test inputs follow the same distribution, so that it overfits less than AIC and may achieve much better risks. FAIC differs from AIC in another way: its tendency to choose more complex models is strengthened in areas where many data points are available (so that the potential damage of picking an overly complex model is smaller), while it is suppressed when few data points are available (and the potential damage is much greater).

This tendency is also apparent in Table 2.1. In the first experiment, where a small model contains the true distribution, it causes FAIC to perform worse than AIC near  $x = 0$ . However, note that the vertical axis is logarithmic, so the difference appears larger than it is: when we average over the training input distribution, we find that FAIC performs better by a factor 20 in terms of squared risk.

In the multivariate experiments, XAIC again performs better than AIC, though the difference eventually disappears as  $n$  grows. With the notable ex-

Table 2.2: Multivariate: squared risk for different training sets; model selection

|                   | Gaussian<br>( $n = 60$ ) | uniform<br>( $n = 60$ ) | spike-<br>and-slab<br>( $n = 60$ ) | Gaussian<br>( $n = 100$ ) |
|-------------------|--------------------------|-------------------------|------------------------------------|---------------------------|
| XAIC <sub>C</sub> | 0.0119                   | 0.0123                  | 0.0144                             | 0.0070                    |
| FAIC <sub>C</sub> | 0.0123                   | 0.0127                  | 0.0133                             | 0.0077                    |
| AIC <sub>C</sub>  | 0.0125                   | 0.0126                  | 0.0156                             | 0.0070                    |
| BIC               | 0.0113                   | 0.0128                  | 0.0140                             | 0.0073                    |
| BMS               | 0.0120                   | 0.0126                  | 0.0138                             | 0.0075                    |
| cSIC              | 0.0119                   | 0.0134                  | 0.0138                             | 0.0074                    |
| GCV               | 0.0129                   | 0.0131                  | 0.0153                             | 0.0072                    |
| FIC               | 0.0196                   | 0.0189                  | 0.0241                             | 0.0111                    |

Table 2.3: Multivariate: squared risk for different training sets; model weighting/averaging

|                    | Gaussian<br>( $n = 60$ ) | uniform<br>( $n = 60$ ) | spike-<br>and-slab<br>( $n = 60$ ) | Gaussian<br>( $n = 100$ ) |
|--------------------|--------------------------|-------------------------|------------------------------------|---------------------------|
| XAIC <sub>CW</sub> | 0.0099                   | 0.0108                  | 0.0114                             | 0.0063                    |
| FAIC <sub>CW</sub> | 0.0100                   | 0.0110                  | 0.0110                             | 0.0066                    |
| AIC <sub>CW</sub>  | 0.0101                   | 0.0108                  | 0.0119                             | 0.0063                    |
| BIC <sub>w</sub>   | 0.0096                   | 0.0106                  | 0.0111                             | 0.0062                    |
| BMA                | 0.0100                   | 0.0107                  | 0.0113                             | 0.0061                    |

ception of the spike-and-slab experiment, FAIC does not perform well here: in two of the experiments, it does worse than AIC. Part of the reason must be our observation at the end of Section 2.4: FAIC's estimate of the generalization error, while unbiased, may potentially have a larger variance than (X)AIC's estimate, and this is not always a good trade-off.

**BIC and BMS/BMA** BIC and BMS do not try to identify the model that will give the best predictions now, but instead attempt to find the most probable model given the data, which usually amounts to the simplest model containing the true distribution. This leads them to be conservative about selecting complex models. For similar reasons, Bayesian model averaging (BMA) puts only small weight on complex models. We see this in Figure 2.1, where BIC and BMA have good performance because they most often select the optimal second model (or in the case of BMA, give it the largest weight). However, for  $f_2$  in Figure 2.2, it may be outperformed by FAIC or XAIC for test inputs away from the centre. In the multivariate experiments, XAIC often performs better than BMS/BMA, and rarely much worse; the only instance of the latter is for the spike-and-slab data, where FAIC outperforms both. (See Section 2.6.1 for further discussion of BMA.)

**FIC** In all our experiments, FIC obtained large squared risks, and we see in Table 2.1 that its selection behaviour was the opposite of FAIC: for extreme  $x$ , FIC often selects a more complex model than near  $x = 0$ . This seems to happen because FIC uses the most complex model's prediction at a given  $x$  to estimate each other model's bias. Because the most complex model will usually have a significant variance, this resulted in FIC being misled in many of the experiments we examined. In particular, in areas with few training inputs, FIC apparently usually believes the simpler models will perform badly because it attributes to them a large bias, so that the same model as elsewhere (or even a more complex one) is selected. Conversely, FIC was often observed to switch to an overly simple model near some input value where this model's estimate happened to coincide with that of the most complex model.

**SIC** SIC obtained large risks in the univariate experiments due to underfitting. Its results in three of the four multivariate experiments were competitive, however.

**GCV** Based on Leeb (2008), we expected GCV might be one of the strongest competitors to XAIC. This was not clearly reflected in our experiments, where its performance was very similar to that of AIC.

## 2.6 Discussion

### 2.6.1 Relation to the Bayesian predictive distribution

The quantity  $\kappa_{x'}$  that occurs in FAIC has an interpretation in the Bayesian framework. If we do linear regression with known variance and a noninformative prior on  $\beta$ , then after observing  $X$ ,  $\mathbf{Y}$  and  $x'$ , the predictive distribution of  $\mathbf{y}'$  is  $\mathbf{y}' \mid \mathbf{Y}, X, x' \sim \mathcal{N}(x'^\top \hat{\beta}, \sigma^2(1 + x'^\top (X^\top X)^{-1} x'))$ . We see that  $\kappa_{x'}$  and the variance of this predictive distribution obey a linear relation. Thus if BMA is allowed to give a distribution over output values as its prediction, then this distribution (a mixture of Gaussians with different variances) will reflect that some models' predictions are more reliable than others. However, if the predictive distribution must be summarized by a point prediction, then such information is likely to be lost. For instance, if the point prediction  $\hat{y}'$  is to be evaluated with squared loss and  $\hat{y}'$  is chosen to minimize the expected loss under the predictive distribution (as in our experiments in Section 2.5), then  $\hat{y}'$  is a weighted average of posterior means for  $\mathbf{y}'$  given  $x'$  (one mean for each model, weighted by its posterior probability). The predictive variances are not factored into  $\hat{y}'$ , so that in this scenario, BMA does not use the information captured by  $\kappa_{x'}$ , so that XAIC and FAIC rely on.

This is not to say that BMA *should* use this information: the consideration of finding the most probable model (BMS, BIC) or the full distribution over models (BMA) is not affected by the purpose for which the model will be used, so it should not depend on the input values in the test data through  $\kappa_{x'}$ . This suggests that there is no XBIC analogue to XAIC. For Bayesian methods that

aim for good predictions, such as DIC (Spiegelhalter et al., 2002), BPIC (Ando, 2007) and WAIC (Watanabe, 2010), on the other hand, extra-sample and focused equivalents may exist. Gelman et al. (2014) give a theoretical and experimental comparison between AIC, DIC and WAIC.

### 2.6.2 Relation to covariate shift methods

We observed at the end of Section 2.4 that of the two data-dependent terms in XAIC, the log-likelihood is independent of  $X'$ , while  $\kappa_{X'}$  is (largely) unaffected by output values. An important practical consequence of this split between input and output values is that XAIC and FAIC look for models that give a good overall fit, not just a good fit at the test inputs.  $X'$  is then used to determine how well we can expect these models to generalize to the test set. So if we have two models and believe each to be able to give a good fit in a different region of the input space, then FAIC is not the proper tool for the task of finding these regions: FAIC considers global fit rather than local fit when evaluating a model, and within the model selects the maximum likelihood estimator, not an estimator specifically chosen for a local fit at input point  $x$ .

In this respect, our methods differ from those commonly used in the covariate shift literature (see Sugiyama and Kawanabe (2012); Pan and Yang (2010); some negative results are in Ben-David et al. (2010)), where typically a model (and an estimator within that model) is sought that will perform well on the test set only, using for example importance weighting. This is appropriate if we believe that no available model can give satisfactory results on both training and test inputs simultaneously. In situations where such models are believed to exist, our methods try to find them using all information in the training set.

## 2.7 Conclusions and future work

We have shown a bias in AIC when it is applied to supervised learning problems, and proposed XAIC and FAIC as versions of AIC which correct this bias. We have experimentally shown that these methods give better predictive performance than other methods in many situations.

We see several directions for future work. First, the practical usefulness of our methods needs to be confirmed by further experiments. Other future work includes considering other model selection methods: determining whether they are affected by the same bias that we found for AIC, whether such a bias can be removed (possibly leading to extra-sample and focused versions of those methods), and how these methods perform in simulation experiments and on real data. In particular, BPIC (Ando, 2007) is a promising candidate, as its derivation starts with a Bayesian equivalent of (2.1). (The same may be true for WAIC (Watanabe, 2010).) An XBPIC method would also be better able to deal with more complex models that a variant of AIC would have difficulty with, such as hierarchical Bayesian models, greatly increasing its practical applicability.

## Appendix 2.A Regularity conditions and proofs

**Assumption 2.3** (Regularity conditions). *Items 1–4 correspond to the regularity assumptions behind AIC given by Shibata (1989), but rewritten to take the input variables into account. Item 5 is the assumption of asymptotic normality of the maximum likelihood estimator, which is also standard.*

1.  $\Theta \subseteq \mathbf{R}^k$  is open, and for sufficiently large  $n$  the gradient and Hessian of the log-likelihood function  $\ell(\theta) = \log g(\mathbf{Y} \mid X, \theta)$  are well-defined for all  $\theta \in \Theta$  with probability 1, and both are continuous;
2. For sufficiently large  $n$ ,  $\mathbf{E}_{\mathbf{Y} \mid X} \left| \frac{\partial}{\partial \theta} \ell(\theta) \right| < \infty$  and  $\mathbf{E}_{\mathbf{Y} \mid X} \left| \frac{\partial^2}{\partial \theta^2} \ell(\theta) \right| < \infty$ ;
3. For sufficiently large  $n$ , there exists a unique  $\theta_o \in \Theta$  such that  $\mathbf{E}_{\mathbf{Y} \mid X} \frac{\partial}{\partial \theta} \ell(\theta_o) = 0$ . For all  $\epsilon > 0$ , it satisfies

$$\inf_{\theta: \|\theta - \theta_o\| > \epsilon} \ell(\theta_o) - \ell(\theta) \rightarrow \infty \quad \text{almost surely}$$

as  $n \rightarrow \infty$ ;

4. For all  $\epsilon > 0$ , there is a  $\delta > 0$  such that for sufficiently large  $n$ ,

$$\sup_{\|\theta - \theta_o\| < \delta} \left| \mathbf{E}_{\mathbf{Y} \mid X} [\hat{\theta}(\mathbf{Y} \mid X) - \theta_o]^\top I(\theta \mid X) [\hat{\theta}(\mathbf{Y} \mid X) - \theta_o] - \text{tr} \left[ J(\theta_o \mid X) I(\theta_o \mid X)^{-1} \right] \right| < \epsilon,$$

where  $I(\theta \mid X) = -\mathbf{E}_{\mathbf{Y} \mid X} \frac{\partial^2}{\partial \theta^2} \ell(\theta)$  and  $J(\theta_o \mid X) = \mathbf{E}_{\mathbf{Y} \mid X} \left[ \frac{\partial}{\partial \theta} \ell(\theta_o) \right] \left[ \frac{\partial}{\partial \theta} \ell(\theta_o) \right]^\top$  are continuous and positive definite.

5.  $\sqrt{n}(\hat{\theta}(\mathbf{Y} \mid X) - \theta_o) \xrightarrow{D} \mathcal{N}(0, \Sigma)$  for some  $\Sigma$ .

**Proof of Theorem 2.1.** This proof is adapted from the one in Burnham and Anderson (2002), with modifications to take  $X$  and  $X'$  into account. Derivation of an estimator for (2.3) starts with a Taylor expansion:

$$\begin{aligned} & -2 \log g(\mathbf{Y}' \mid X', \hat{\theta}(X, \mathbf{Y})) \\ &= -2 \log g(\mathbf{Y}' \mid X', \theta_o) - 2 \left[ \frac{\partial}{\partial \theta} \log g(\mathbf{Y}' \mid X', \theta_o) \right]^\top [\hat{\theta}(X, \mathbf{Y}) - \theta_o] \\ & \quad - [\hat{\theta}(X, \mathbf{Y}) - \theta_o]^\top \left[ \frac{\partial^2}{\partial \theta^2} \log g(\mathbf{Y}' \mid X', \theta_o) \right] [\hat{\theta}(X, \mathbf{Y}) - \theta_o] + r(\hat{\theta}), \end{aligned}$$

where  $r(\hat{\theta}) / \|\hat{\theta}(X, \mathbf{Y}) - \theta_o\|^2 \rightarrow 0$  as  $\hat{\theta}(X, \mathbf{Y}) \rightarrow \theta_o$ .

Next, we take the expectation  $\mathbf{E}_{\mathbf{Y}' \mid X'}$ . Given the regularity conditions on the model,  $\theta_o$  minimizes  $\mathbf{E}_{\mathbf{Y}' \mid X'} \log g(\mathbf{Y}' \mid X', \theta)$ , so the linear term vanishes. (Note that we need this vanishing to hold for any  $X'$  (or equivalently, for any

single point  $x$ ); this follows from the assumption that  $\theta_o$  represents the true conditional data-generating distribution.) The coefficient of the quadratic term now becomes the conditional Fisher information at  $\theta_o$  given  $X'$ , so we have

$$\begin{aligned} -2 \mathbf{E}_{\mathbf{Y}'|X'} \log g(\mathbf{Y}' | X', \hat{\theta}(X, \mathbf{Y})) &= -2 \mathbf{E}_{\mathbf{Y}'|X'} \log g(\mathbf{Y}' | X', \theta_o) \\ &\quad + [\hat{\theta}(X, \mathbf{Y}) - \theta_o]^\top I(\theta_o | X') [\hat{\theta}(X, \mathbf{Y}) - \theta_o] + r(\hat{\theta}). \end{aligned}$$

Rearranging the quadratic term and taking the other expectation, we obtain

$$\begin{aligned} -2 \mathbf{E}_{\mathbf{Y}|X} \mathbf{E}_{\mathbf{Y}'|X'} \log g(\mathbf{Y}' | X', \hat{\theta}(X, \mathbf{Y})) &= -2 \mathbf{E}_{\mathbf{Y}'|X'} \log g(\mathbf{Y}' | X', \theta_o) \\ &\quad + \text{tr} \left\{ I(\theta_o | X') \left[ \mathbf{E}_{\mathbf{Y}|X} [\hat{\theta}(X, \mathbf{Y}) - \theta_o] [\hat{\theta}(X, \mathbf{Y}) - \theta_o]^\top \right] \right\} + \mathbf{E}_{\mathbf{Y}|X} r(\hat{\theta}). \end{aligned} \quad (2.9)$$

The other matrix in the trace is the conditional covariance matrix of  $\hat{\theta}(X, \mathbf{Y})$ .

To proceed with the first term on the right hand side, we use Assumption 2.2. Then we have

$$-2 \frac{n}{n'} \mathbf{E}_{\mathbf{Y}'|X'} \log g(\mathbf{Y}' | X', \theta_o) = -2 \mathbf{E}_{\mathbf{Y}|X} \log g(\mathbf{Y} | X, \theta_o)$$

for a sample  $(X, \mathbf{Y})$  of size  $n$ . (Here  $X$  still represents the values of the input variable in the training set, but  $\mathbf{Y}$  conceptually represents a new sample.) Now only one  $X$  remains, so the rest of the derivation corresponds to that of standard AIC, which gives us

$$-2 \mathbf{E}_{\mathbf{Y}|X} \log g(\mathbf{Y} | X, \theta_o) = -2 \mathbf{E}_{\mathbf{Y}|X} \log g(\mathbf{Y} | X, \hat{\theta}(X, \mathbf{Y})) + k + o(1). \quad (2.10)$$

Multiplying (2.9) by  $n/n'$  and plugging in the above, we get

$$\begin{aligned} -2 \frac{n}{n'} \mathbf{E}_{\mathbf{Y}|X} \mathbf{E}_{\mathbf{Y}'|X'} \log g(\mathbf{Y}' | X', \hat{\theta}(X, \mathbf{Y})) &= -2 \mathbf{E}_{\mathbf{Y}|X} \log g(\mathbf{Y} | X, \hat{\theta}(X, \mathbf{Y})) \\ &\quad + k + \frac{n}{n'} \text{tr} \left\{ I(\theta_o | X') \text{Cov}(\hat{\theta}(X, \mathbf{Y}) | X) \right\} + \mathbf{E}_{\mathbf{Y}|X} \frac{n}{n'} r(\hat{\theta}) + o(1). \end{aligned}$$

The term with the trace is what we called  $\kappa_{X'}$ .

By the assumed asymptotic normality of the maximum likelihood estimator,  $\mathbf{E}_{\mathbf{Y}|X} n \|\hat{\theta}(X, \mathbf{Y}) - \theta_o\|^2$  converges to a constant, so that the first remainder term  $\mathbf{E}_{\mathbf{Y}|X} (n/n') r(\hat{\theta}) = (1/n') o(1)$ ; because we additionally assumed the test set is either fixed or grows with the training set, this is again  $o(1)$ . This proves (2.7).

In the case of a linear model with fixed variance  $\sigma^2$ , the second-order Taylor approximation and the approximation in (2.10) are actually exact.  $\square$

**Proof of Theorem 2.2.** This proof follows a different path from the one above. It is adapted from the derivation of AIC<sub>C</sub> in Burnham and Anderson (2002, Section 7.4.1). We first consider the case where the training set size  $n' = 1$ . Then  $X'$  becomes a vector (we choose to make it a column vector) and  $\mathbf{Y}'$  a

scalar; we write  $x$  and  $\mathbf{y}$  for these. Hats denote maximum likelihood estimates. For Gaussian densities, we get

$$\begin{aligned} T &= -2 \mathbf{E}_{\mathbf{Y}|X} \mathbf{E}_{\mathbf{y}|x} \log g(\mathbf{y} | x, \hat{\theta}(X, \mathbf{Y})) \\ &= \mathbf{E}_{\mathbf{Y}|X} \mathbf{E}_{\mathbf{y}|x} \left[ \log 2\pi \hat{\sigma}^2(X, \mathbf{Y}) + \frac{1}{\hat{\sigma}^2(X, \mathbf{Y})} (\mathbf{y} - x^\top \hat{\beta}(X, \mathbf{Y}))^2 \right] \\ &= \mathbf{E}_{\mathbf{Y}|X} \log 2\pi \hat{\sigma}^2(X, \mathbf{Y}) + \mathbf{E}_{\mathbf{Y}|X} \frac{1}{\hat{\sigma}^2(X, \mathbf{Y})} \mathbf{E}_{\mathbf{y}|x} (\mathbf{y} - x^\top \hat{\beta}(X, \mathbf{Y}))^2. \end{aligned}$$

We will call the final term  $T'$ . Writing  $y_o$  for  $\mathbf{E}_{\mathbf{y}|x} y$  and  $\sigma_o^2$  for  $\mathbf{y}$ 's unknown variance, the inner expectation becomes

$$\begin{aligned} &\mathbf{E}_{\mathbf{y}|x} (\mathbf{y} - x^\top \hat{\beta}(X, \mathbf{Y}))^2 \\ &= \mathbf{E}_{\mathbf{y}|x} (\mathbf{y} - y_o)^2 + 2 (\mathbf{y}_o - x^\top \hat{\beta}(X, \mathbf{Y})) \mathbf{E}_{\mathbf{y}|x} (\mathbf{y} - y_o) + (\mathbf{y}_o - x^\top \hat{\beta}(X, \mathbf{Y}))^2 \\ &= \sigma_o^2 + x^\top (\beta_o - \hat{\beta}(X, \mathbf{Y})) (\beta_o - \hat{\beta}(X, \mathbf{Y}))^\top x. \end{aligned}$$

Using the fact that  $\hat{\beta}(X, \mathbf{Y})$  and  $\hat{\sigma}^2(X, \mathbf{Y})$  are independent in this setting,

$$T' = \left[ \mathbf{E}_{\mathbf{Y}|X} \frac{1}{\hat{\sigma}^2(X, \mathbf{Y})} \right] \cdot \left[ \sigma_o^2 + x^\top \text{Cov}(\hat{\beta}(X, \mathbf{Y}) | X) x \right].$$

The covariance matrix equals  $\sigma_o^2 (X^\top X)^{-1}$ . Then we use that  $n\hat{\sigma}^2/\sigma_o^2$  follows a chi-squared distribution with  $n - k + 1$  degrees of freedom ( $k$  is the number of free parameters in the model, which includes  $\sigma^2$ ), and that  $\mathbf{E} 1/\chi_{n-k}^2 = 1/(n - k - 1)$ :

$$\begin{aligned} T' &= \left[ \mathbf{E}_{\mathbf{Y}|X} \frac{1}{\hat{\sigma}^2(X, \mathbf{Y})} \right] \left[ \sigma_o^2 + \sigma_o^2 x^\top (X^\top X)^{-1} x \right] \\ &= \left[ \mathbf{E}_{\mathbf{Y}|X} \frac{\sigma_o^2}{n\hat{\sigma}^2(X, \mathbf{Y})} \right] \left[ n + nx^\top (X^\top X)^{-1} x \right] \\ &= \frac{n + nx^\top (X^\top X)^{-1} x}{n - k - 1} \\ &= 1 + \frac{n + nx^\top (X^\top X)^{-1} x - (n - k - 1)}{n - k - 1} = 1 + \frac{k + \kappa_x}{n - k - 1}, \end{aligned}$$

where  $\kappa_x = nx^\top (X^\top X)^{-1} x + 1$ . The reason for splitting off the 1 from the fraction is that  $n(\mathbf{E}_{\mathbf{Y}|X} \log 2\pi \hat{\sigma}^2(X, \mathbf{Y}) + 1)$  is  $-2$  times the maximized log-likelihood. Then we multiply by  $n$  and get the result in the stated form:

$$\begin{aligned} nT &= -2 \mathbf{E}_{\mathbf{Y}|X} \log g(\mathbf{Y} | X, \hat{\theta}(X, \mathbf{Y})) + \frac{n(k + \kappa_x)}{n - k - 1} \\ &= -2 \mathbf{E}_{\mathbf{Y}|X} \log g(\mathbf{Y} | X, \hat{\theta}(X, \mathbf{Y})) + k + \kappa_x + \frac{(k + 1)(k + \kappa_x)}{n - k - 1}. \end{aligned}$$

The result for  $n' > 1$  now follows by taking the average over all sample points in the test set on both sides.  $\square$

**Proof of Proposition 2.3.** Assume without loss of generality that the variance is known (as its inclusion does not affect the statement of the theorem) and that the basis is orthonormal with respect to the measure underlying  $\mathbf{x}$  (that is, that  $\mathbf{E}_\mathbf{x} \mathbf{x} \mathbf{x}^\top = I_k$ ). Then

$$\begin{aligned} \mathbf{E}_\mathbf{x} \kappa_\mathbf{x} &= n \mathbf{E}_\mathbf{x} \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x} \\ &= n \operatorname{tr}(\mathbf{X}^\top \mathbf{X})^{-1} = \operatorname{tr}\left(\frac{1}{n} \mathbf{X}^\top \mathbf{X}\right)^{-1}, \end{aligned}$$

where orthonormality was used in the second equality. To compare the  $\kappa_\mathbf{x}$  for this model with that of a submodel with one fewer entry in its design vectors, write

$$\frac{1}{n} \mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{v} \\ \mathbf{v}^\top & \mathbf{d} \end{bmatrix}.$$

Note that by orthonormality, the expected value of this matrix is the identity matrix. We require that its inverse exists. Then for  $\mathbf{d}' = (\mathbf{d} - \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{v})^{-1}$ ,

$$\begin{aligned} \mathbf{E} \kappa_\mathbf{x} &= \mathbf{E} \operatorname{tr} \begin{bmatrix} \mathbf{A} & \mathbf{v} \\ \mathbf{v}^\top & \mathbf{d} \end{bmatrix}^{-1} \\ &= \mathbf{E} \operatorname{tr} \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1} \mathbf{v} \mathbf{d}' \mathbf{v}^\top \mathbf{A}^{-1} & -\mathbf{A}^{-1} \mathbf{v} \mathbf{d}' \\ -\mathbf{d}' \mathbf{v}^\top \mathbf{A}^{-1} & \mathbf{d}' \end{bmatrix} \\ &= \mathbf{E} \operatorname{tr} \mathbf{A}^{-1} + \mathbf{E} \frac{\mathbf{v}^\top \mathbf{A}^{-2} \mathbf{v} + 1}{\mathbf{d} - \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{v}} \\ &\geq \mathbf{E} \operatorname{tr} \mathbf{A}^{-1} + \mathbf{E} \frac{1}{\mathbf{d} - \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{v}} \\ &\geq \mathbf{E} \operatorname{tr} \mathbf{A}^{-1} + \frac{1}{1 - \mathbf{E} \mathbf{v}^\top \mathbf{A}^{-1} \mathbf{v}} \\ &\geq \mathbf{E} \operatorname{tr} \mathbf{A}^{-1} + 1. \end{aligned}$$

This shows that adding an element to the design vector increases  $\mathbf{E} \kappa_\mathbf{x}$  by at least one. For  $k = 1$  (so that  $\mathbf{A}$  is a  $0 \times 0$  matrix), we have equality if and only if  $\mathbf{d} = 1$  almost surely, which means that for  $\mathbf{x}_1$  (the first and only entry of design vector  $\mathbf{x}$ ), we must have  $\mathbf{x}_1 = \pm 1$  almost surely. For  $k \geq 2$ , because  $\mathbf{A}^{-1}$  is positive definite, equality requires that  $\mathbf{v}$  is the zero vector almost surely (in addition to the same requirement as above on all  $\mathbf{x}_i$ ). But this can only be satisfied if  $\mathbf{x}_i \mathbf{x}_k = 0$  almost surely for all  $i < k$ , which is incompatible with the conditions on  $\mathbf{x}_1$  and  $\mathbf{x}_k$ .  $\square$

