AALTO UNIVERSITY
SCHOOL OF SCIENCE

Juho Piironen

Bayesian predictive methods for model selection

Independent research project in applied mathematics (Mat-2.4108)

April 3, 2013

Supervisor:

Prof. Ahti Salo

Instructor:

M. Sc. Janne Ojanen

Contents

1 Introduction **2**

2 Bayesian inference **3**
2.1 Bayes' theorem and basic principles 3
2.2 Example: linear regression 4

3 Predictive model selection **8**
3.1 Predictive ability as an expected utility 8
3.2 Estimation of the expected utility 9
3.2.1 Utility estimation when a reference model is available 11
3.2.2 Utility estimation when no reference model is available 13
3.2.3 Other methods 15

4 Example problem: covariate selection **16**
4.1 Data and model 16
4.2 Methods 17
4.3 Results 18

5 Conclusions **21**

Appendix A Linear regression model with conjugate prior **24**

1 Introduction

Mathematical modelling plays an important role in the natural sciences and engineering, but nowadays it is used also in many other fields such as economics, sociology and psychology. The applicability of any model is usually measured by its predictive performance; if the model's predictions about some phenomenon do not agree with our observations, there is not much sense in trying to interpret the model. On the other hand, if the model is able to make reasonable predictions, it may be useful even though no model can be concluded to be "correct" (Box, 1979). Some other features of a good model are generally considered to be simplicity, transparency and interpretability.

In this study we consider Bayesian statistical models and predictive model selection. Predictive model selection refers to a problem where one is choosing a model from a set of candidate models based on their ability to predict unseen observations. In practice, the methods for assessing the predictive performance of a given model may vary depending on the modelling task and available data. We shall describe several methods which estimate the predictive ability of a model via expected utilities and illustrate their use in a covariate selection problem. The methods we consider are cross validation and two different information criteria, and also Bayes and Gibbs variants for reference, training and test utilities.

We begin with a brief introduction to Bayesian statistics in section 2. We introduce the concept of belief updating with Bayes' theorem and the prediction with a Bayesian model. After the general discussion, we derive the Bayesian linear regression model as an example. Section 3 deals with model selection and explains how the predictive performance of a model is defined and how it can be estimated. In section 4 we illustrate the use of these methods by applying them to a covariate selection problem, in which we use the derivation of the regression model from section 2.2. Through the example, we discuss some key differences and similarities between the methods, such as biasness or unbiasness, as well as other relevant concepts related to model selection.

2 Bayesian inference

2.1 Bayes' theorem and basic principles

In Bayesian statistics uncertain quantities are treated as random variables, and in essence the whole Bayesian statistics is about revising subjective beliefs about these quantities on the basis of observations. This interpretation is fundamentally different from what is done in frequentist statistics, where the quantities of interest are usually considered to be fixed constants even though they are unknown. For a parametric model, given all the model assumptions M the belief updating about parameter or parameters $\boldsymbol{\theta}$ on the basis of the observations $\mathbf{y}$ is done by the Bayes' theorem

$$p(\boldsymbol{\theta}|\mathbf{y}, M) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)}{p(\mathbf{y}|M)} = \frac{p(\mathbf{y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)}{\int p(\mathbf{y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}}. \quad (1)$$

Prior probability distribution $p(\boldsymbol{\theta}|M)$ describes our beliefs about $\boldsymbol{\theta}$ before making the observations. $p(\mathbf{y}|\boldsymbol{\theta}, M)$ is the actual statistical model describing the relation between the parameters and the observations. Once the observations are obtained, it becomes a function of $\boldsymbol{\theta}$ and we call it the *likelihood*. The likelihood describes which of the parameter values are more likely based on the observations, but it is not a genuine probability distribution since in general it does not integrate into 1. The denominator $p(\mathbf{y}|M)$, often referred to as the *marginal likelihood*, is a normalizing constant that makes the right hand side a proper probability distribution. The result $p(\boldsymbol{\theta}|\mathbf{y}, M)$ is called the *posterior distribution* and it combines the prior information and the observations giving us the updated belief about $\boldsymbol{\theta}$. All the inference is done using the posterior because it contains all the knowledge there is about $\boldsymbol{\theta}$ after the observations.

Bayesian treatment has a natural way of predicting new observations $\tilde{\mathbf{y}}$ given the old ones. This is done by using the statistical model $p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, M)$ together with the posterior beliefs about the parameters $\boldsymbol{\theta}$. Because we do not know the exact value of $\boldsymbol{\theta}$, we use all the possible values with respect to their probabilities, which leads us to take expectation of the probability model over the posterior $p(\boldsymbol{\theta}|\mathbf{y}, M)$:

$$p(\tilde{\mathbf{y}}|\mathbf{y}, M) = \int p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{y}, M)d\boldsymbol{\theta}. \quad (2)$$

$p(\tilde{\mathbf{y}}|\mathbf{y}, M)$ is called the *posterior predictive distribution*, and it takes into account the uncertainty about $\boldsymbol{\theta}$ as well as the stochastic randomness about the future observation. One may see the analogy between the formulas

of $p(\mathbf{y}|M)$ and $p(\tilde{\mathbf{y}}|\mathbf{y}, M)$; $p(\mathbf{y}|M)$ predicts the data $\mathbf{y}$ on the basis of the prior beliefs $p(\boldsymbol{\theta}|M)$, whereas $p(\tilde{\mathbf{y}}|\mathbf{y}, M)$ predicts the future observations $\tilde{\mathbf{y}}$ on the basis of the prior beliefs *and* the previous observations $\mathbf{y}$. For this reason, $p(\mathbf{y}|M)$ is also called as the *prior predictive distribution* before the observations are made.

So far we have conditioned all the terms on model assumptions M. However, we can treat M exactly in the same way as we treat the other parameters $\boldsymbol{\theta}$. This is done by specifying a model space, i.e. a set of candidate models $\{M_k\}_{k=1}^K$, and writing using the Bayes' theorem

$$p(M|\mathbf{y}) = \frac{p(\mathbf{y}|M)p(M)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|M)p(M)}{\sum_{k=1}^K p(\mathbf{y}|M_k)p(M_k)}. \tag{3}$$

Here $p(M)$ and $p(M|\mathbf{y})$ are discrete probability distributions determining the prior and posterior probabilities, respectively, for each model M_k, and $p(\mathbf{y}|M)$ is the marginal likelihood from equation (1) also called the *model evidence*. In other words, after specifying prior probabilities for each model, $p(M|\mathbf{y})$ indicates which of the models are more likely and which of them are unlikely based on the prior beliefs and the data. After calculating posterior probabilities and posterior predictive distributions we can integrate or sum over the model space to get the *Bayesian model averaging* (BMA) predictive distribution (Hoeting et al., 1999)

$$p(\tilde{\mathbf{y}}|\mathbf{y})_{\text{BMA}} = \sum_{k=1}^{K} p(\tilde{\mathbf{y}}|\mathbf{y}, M_k)p(M_k|\mathbf{y}). \tag{4}$$

BMA predictive distribution takes all the K models into account according to their probabilities and is therefore richer than any of the candidate models alone. BMA has been shown to have a good predictive performance (Raftery and Zheng, 2003). However, it is important to note that setting prior probabilities for a set $\{M_k\}_{k=1}^K$ means stating a belief that the true data producing model belongs to this group, i.e. models outside this set are not possible. Thus, a poorly specified set of candidate models is also likely to lead to poor results. We shall use BMA as the reference model in our numerical example in section 4. The concept of the reference model is discussed in section 3.2.

2.2 Example: linear regression

We now consider the linear regression as an example of the Bayesian inference. Here we derive the model and in section 4 we give a model selection

example by fitting a set of linear models to a simulated data. We shall not show all the steps here, but present only the main results. See appendix A for more detailed treatment.

In linear regression with one dependent variable y and p predictor variables $\mathbf{x} = (x_1, ..., x_p)$ the idea is to fit to the given training dataset $\boldsymbol{\mathcal{X}} = (\mathbf{X}, \mathbf{y}) = \{(y^{(i)}, x_1^{(i)}, ..., x_p^{(i)})\}_{i=1}^n$ a model

$$y^{(i)} = \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + ... + \beta_p x_p^{(i)} + \varepsilon^{(i)}, \quad \varepsilon^{(i)} \sim N(0, \sigma^2). \tag{5}$$

This is in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{6}$$

where

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \vdots \\ \mathbf{x}^{(n)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{bmatrix}, \tag{7}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon^{(1)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix}. \tag{8}$$

In this case the unknown parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma^2 \in \mathbb{R}_+$. The joint posterior is

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{\mathcal{X}}) = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2)}{p(\mathbf{y}|\mathbf{X})}. \tag{9}$$

It is important to note that here the observations $\mathbf{y}$ are conditioned on the covariates $\mathbf{X}$, so the covariate values are assumed to be known. First, we write down the likelihood for a single observation which is now the normal distribution

$$p(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{x}^{(i)}\boldsymbol{\beta})^2\right). \tag{10}$$

Assuming that the observations are conditionally independent, the joint likelihood for all the observations is

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{x}^{(i)}\boldsymbol{\beta})^2\right), \tag{11}$$

which can be shown to be proportional to multivariate normal-inverse-gamma distribution for $\boldsymbol{\beta}$ and σ^2:

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) &= \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right) \\ &\quad \times \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}})\right) \\ &\propto N(\boldsymbol{\beta};\, \cdot, \cdot) \times \text{Inv-Gamma}(\sigma^2;\, \cdot, \cdot), \end{aligned} \tag{12}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{13}$$

Next we define a prior for the parameters. For convenience, we choose a *conjugate prior*, which leads to a posterior of the same functional form as the prior. The advantage is that we do not need to calculate the normalizing constant (marginal likelihood) since we know the form of the posterior. In this case also the posterior predictive distribution (2) can be solved analytically. The conjugate prior is given by a joint distribution which is of the same form as the likelihood (12):

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2) &= p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \\ &= N(\boldsymbol{\beta};\, \boldsymbol{\mu}_0,\, \sigma^{-2}\mathbf{A}_0) \times \text{Inv-Gamma}(\sigma^2;\, a_0,\, b_0). \end{aligned} \tag{14}$$

Note that here $\sigma^{-2}\mathbf{A}_0$ denotes the precision matrix, i.e the inverse of the covariance matrix of the coefficients $\boldsymbol{\beta}$. Now we straightforwardly multiply (12) and (14) to get for the joint posterior

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\mathcal{X}}) &\propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2) \\ &\propto N(\boldsymbol{\beta};\, \boldsymbol{\mu}_n, \sigma^{-2}\mathbf{A}_n) \times \text{Inv-Gamma}(\sigma^2;\, a_n, b_n) \end{aligned} \tag{15}$$

where

$$\mathbf{A}_n = \mathbf{A}_0 + \mathbf{X}^{\mathrm{T}}\mathbf{X} \tag{16}$$

$$\boldsymbol{\mu}_n = \mathbf{A}_n^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{A}_0\boldsymbol{\mu}_0) \tag{17}$$

$$a_n = a_0 + \frac{n}{2} \tag{18}$$

$$b_n = b_0 + \frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\mu}_0^{\mathrm{T}}\mathbf{A}_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^{\mathrm{T}}\mathbf{A}_n\boldsymbol{\mu}_n). \tag{19}$$

Given that the unnormalized posterior is proportional to $N(\boldsymbol{\beta};\, \boldsymbol{\mu}_n, \sigma^{-2}\mathbf{A}_n) \times \text{Inv-Gamma}(\sigma^2;\, a_n, b_n)$, the posterior must be exactly this distribution. So the prior definition (14) with parameters $\boldsymbol{\mu}_0, \mathbf{A}_0, a_0, b_0$ indeed leads in a posterior of the same form with parameters determined by the equations

(16)–(19).

After solving the posterior we can calculate the posterior predictive distribution (2). We shall calculate the joint predictive distribution for $\tilde{n}$ future observations $\tilde{\mathbf{y}}$, assuming that the covariates $\tilde{\mathbf{X}}$ in those points are known. Straight from the definition (2) we obtain

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \boldsymbol{\mathcal{X}}) = \int\int p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\mathcal{X}}) d\sigma^2 d\boldsymbol{\beta}. \tag{20}$$

This integral can be calculated analytically and the result is

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \boldsymbol{\mathcal{X}}) = (2\pi)^{-\tilde{n}/2} \frac{\Gamma(\tilde{a})}{\Gamma(a_n)} \left(\frac{|\mathbf{A}_n|}{|\tilde{\mathbf{A}}|}\right)^{\frac{1}{2}} \frac{b_n^{a_n}}{\tilde{b}^{\tilde{a}}}, \tag{21}$$

where

$$\tilde{\mathbf{A}} = \mathbf{A}_n + \tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}} \tag{22}$$

$$\tilde{\boldsymbol{\mu}}_n = \tilde{\mathbf{A}}^{-1}(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{y}} + \mathbf{A}_n\boldsymbol{\mu}_n) \tag{23}$$

$$\tilde{a} = a_n + \frac{\tilde{n}}{2} \tag{24}$$

$$\tilde{b} = b_n + \frac{1}{2}(\tilde{\mathbf{y}}^{\mathrm{T}}\tilde{\mathbf{y}} + \boldsymbol{\mu}_n^{\mathrm{T}}\mathbf{A}_n\boldsymbol{\mu}_n - \tilde{\boldsymbol{\mu}}^{\mathrm{T}}\tilde{\mathbf{A}}\tilde{\boldsymbol{\mu}}). \tag{25}$$

Distribution (21) is a multivariate t-distribution for $\tilde{n}$ future observations $\tilde{\mathbf{y}}$, given the covariates $\tilde{\mathbf{X}}$. Later in the model selection example in section 4 we are interested in only one prediction at a time, and in this case we simply set $\tilde{n} = 1$ and $\tilde{\mathbf{y}} = \tilde{y}$.

Finally, we derive the model evidence $p(\mathbf{y}|\mathbf{X})$. This is needed for computing the posterior probabilities (3) for a set of different models to obtain the BMA predictive distribution (4). The marginal likelihood is given by

$$p(\mathbf{y}|\mathbf{X}) = \int\int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) d\sigma^2 d\boldsymbol{\beta}. \tag{26}$$

Since the prior $p(\boldsymbol{\beta}, \sigma^2)$ and posterior $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\mathcal{X}})$ are of the same functional form (because of the conjugacy), the integral is exactly the same as when calculating the posterior predictive distribution (20), only the parameters are changed. It is therefore easy to verify that the result for the marginal likelihood is given by

$$p(\mathbf{y}|\mathbf{X}) = (2\pi)^{-n/2} \frac{\Gamma(a_n)}{\Gamma(a_0)} \left(\frac{|\mathbf{A}_0|}{|\mathbf{A}_n|}\right)^{\frac{1}{2}} \frac{b_0^{a_0}}{b_n^{a_n}}. \tag{27}$$

3 Predictive model selection

The term *model selection* in general can be considered as a decision problem where one needs to select a model from a set of candidate models $\{M_k\}_{k=1}^K$ on the basis of certain criteria. As already discussed in the introduction, when we are talking about a scientific theory or a statistical model, the best available model is usually the one giving the most precise predictions about the future observations, and model selection that is based on assessing the predictive performance of the candidate models is called *predictive model selection* (Vehtari and Ojanen, 2013). The predictive performance of a model can be defined via expected utilities which we shall discuss next. Throughout section 3 we shall deal with models that predict a single output variable y given a covariate vector $\mathbf{x}$ as an input. The linear regression model from section 2.2 is an example of this type of model. Note, however, that the following model selection methods can be used basically for any Bayesian models, but we fix the model structure because of the notation. Moreover, we shall again denote the training data as $\boldsymbol{\mathcal{X}} = (\mathbf{X}, \mathbf{y})$.

3.1 Predictive ability as an expected utility

A natural way of assessing the predictive performance of a model M_k is to use a utility function to describe the quality of the predictions. This means that we define a suitable utility function u that maps each prediction $a_k \in \mathcal{A}_k$ of the model to a utility value for each possible future observation $\tilde{y} \in \tilde{\mathcal{Y}}$ so that the utility is higher for predictions that are closer to the possibly later observed state of the world. In mathematical terms the utility function is $u : \mathcal{A}_k \times \tilde{\mathcal{Y}} \mapsto \mathbb{R}$.

Since u depends on $\tilde{y}$ which is typically not known before making the prediction, the utilities for the predictions cannot be evaluated beforehand. For this reason, instead of the actual utilities we use expected utilities, which are defined simply as the expected value of the utility function u over the true future observation distribution $p_t(\tilde{y}|\tilde{\mathbf{x}})$. We call this the *Bayes generalization utility*

$$u_{\text{gen}}^{\text{B}}(M_k) = \text{E}_{\tilde{y}|\tilde{\mathbf{x}}}[u(a_k, \tilde{y})] = \int u(a_k, \tilde{y}) p_t(\tilde{y}|\tilde{\mathbf{x}}) d\tilde{y}. \tag{28}$$

According to its name, the generalization utility measures how well the model generalizes on unseen data. The definition (28) holds when the model predictions a_k are not conditioned on the values of the model parameters $\boldsymbol{\theta}_k$. However, one can condition the predictions on the parameters

$a_k = a_k(\boldsymbol{\theta}_k)$ and then take the expectation of these parametric predictions over the model posterior as $\mathrm{E}_{\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}}}[u(a_k(\boldsymbol{\theta}_k), \tilde{y})]$. The expectation of this over the future observation distribution is called the *Gibbs generalization utility* (following Watanabe (2009)):

$$\begin{aligned} u^{\mathrm{G}}_{\mathrm{gen}}(M_k) &= \mathrm{E}_{\tilde{y}|\tilde{\mathbf{x}}}\left[\mathrm{E}_{\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}}}[u(a_k(\boldsymbol{\theta}_k), \tilde{y})]\right] \\ &= \int \left[\int u(a_k(\boldsymbol{\theta}_k), \tilde{y}) p(\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}}, M_k) d\boldsymbol{\theta}_k\right] p_t(\tilde{y}|\tilde{\mathbf{x}}) d\tilde{y}. \end{aligned} \tag{29}$$

Even though both (28) and (29) measure the predictive ability of the model, they differ in a fundamental respect. The Gibbs utility measures the average predictive performance of the parametric probability densities of the model M_k and does not tell how to actually predict the future observations. On the other hand, the Bayes utility is a measure of the goodness of the actual prediction made by the model. Because of this difference the formulation of the Bayes utility appears more intuitive and natural.

3.2 Estimation of the expected utility

In practice the true data generating distribution $p_t(\cdot)$ is also unknown and the direct evaluation of (28) and (29) is not possible. However, there are still several methods that can be employed to estimate the expected utility of a model, and in sections 3.2.1–3.2.3 we review some of these methods in detail. In some cases it is possible to construct a good model, which is believed to describe well the knowledge about the modelling task and no significant flaws in the model are detected. In this case we call such a model the *reference model* or the *actual belief model* (Bernardo and Smith, 1994) and denote it by M_*. If the reference model is available, we may believe that it describes well the distribution of the future observation, and we can calculate the expected utilities over this distribution and end up with the *reference utility*. This is illustrated in Figure 1.

The reference model approach, however, leads to a natural question: if we already have a model whose predictions we believe are the best available, why is there need for any model selection? There may be several different reasons. For example, the reference model may be very complex, it may not be expressed in a closed form and the integrals over it may be difficult and computationally heavy. Such a model may also be difficult to interpret if we are interested in the model parameters. In practice it would be more satisfying to obtain a model which gives almost as good predictions as the

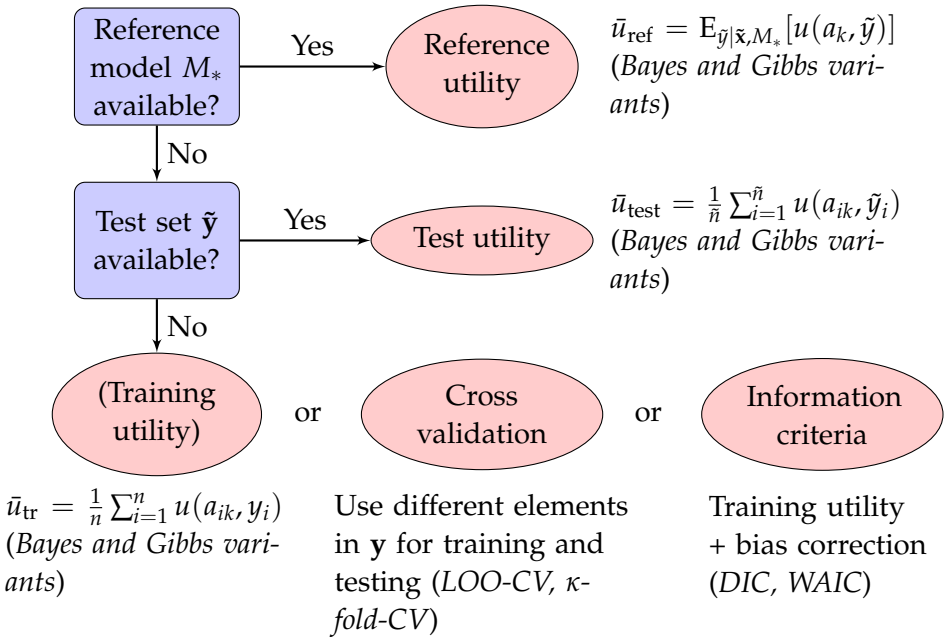

Figure 1: Different ways of estimating the expected utility of model M_k depending on the situation. The methods that we consider in this study are classified according their type of approach and are given in parenthesis in italic. The training utility is set in parenthesis due to its biasness and incompetence in estimating the generalization utility.

reference model but is as simple as possible and also easy to interpret and use in the calculations. Calculating the expected utilities over the reference predictive distribution reveals if there exists such a model in the set of the candidate models. The utility estimation with the reference model is discussed in section 3.2.1.

A completely different situation occurs when no reference model is available. This is typically the more general case when facing a new modelling task. In this case the performance of the candidate models must be assessed using the available data. The naive approach is to use the same data for model training and testing which is referred to as *training utility*, but this leads in biased, overoptimistic utility estimates. A better approach is to separate a set of points from the data and use them for testing. This leads to a *test utility* (figure 1). However, many times the amount of the available data is relatively small and one cannot afford the luxury of splitting the data into two, as this would reduce the number of points for model training too much. In these cases the utility estimation can be done for ex-

ample using *cross validation* or *information criteria* (figure 1). The methods based on the reuse of the training data are discussed in section 3.2.2.

3.2.1 Utility estimation when a reference model is available

In theory, the utility estimation when the reference model M_* is available is rather straightforward. As already mentioned in section 3.2, the idea is to replace the true distribution of the future observation $p_t(\cdot)$ with the posterior predictive distribution of the reference model M_* and evaluate the integrals in equations (28) and (29) directly.

Bayes reference utility When estimating the Bayes generalization utility (28), the expected utility for any prediction $a = a(\tilde{y})$ about the future observation $\tilde{y}$ can be expressed as

$$\bar{u}(a) = \int u(a, \tilde{y}) p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\mathcal{X}}, M_*) d\tilde{y}, \tag{30}$$

where $p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\mathcal{X}}, M_*)$ is the posterior predictive distribution of the reference model. As a utility function, we shall use the logarithmic score proposed by Good (1952):

$$u(a, \tilde{y}) = \log a(\tilde{y}). \tag{31}$$

The logarithmic score is a widely used utility function, which has also information theoretical grounds (Kullback and Leibler, 1951). In this case the expected utility of prediction a becomes

$$\bar{u}(a) = \int \log a(\tilde{y}) p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\mathcal{X}}, M_*) d\tilde{y}, \tag{32}$$

which is maximized by $\hat{a} = \hat{a}(\tilde{y}) = p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\mathcal{X}}, M_*)$. In other words, the optimal prediction under the reference model M_* is the prediction made by the reference model itself. This is evident, given that we believe $p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\mathcal{X}}, M_*)$ describes best the future observations. With similar logic, the optimal prediction under model specification M_k is the posterior predictive distribution $\hat{a}_k = \hat{a}_k(\tilde{y}) = p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\mathcal{X}}, M_k)$. Hence, the expected utility for the optimal prediction of candidate model M_k is given by the *Bayes reference utility*

$$u^{\mathrm{B}}_{\mathrm{ref}}(M_k) = \int \log p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\mathcal{X}}, M_k) p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\mathcal{X}}, M_*) d\tilde{y}. \tag{33}$$

This utility is greatest for the models whose predictions are closest to the predictions of M_*, so the candidate models can be ranked according to their utilities. Naturally it follows that the quality of the model selected

this way depends on the quality of the reference model. For example if the reference model is overfitted to the training data, the model selection favors also overfitted models.

In practice, the reference utility (33) can be evaluated by using the covariates in the training set and calculating the average utility as

$$u_{\text{ref}}^{\text{B}}(M_k) = \frac{1}{n}\sum_{i=1}^{n}\int \log p(y|\mathbf{x}^{(i)}, \boldsymbol{\mathcal{X}}, M_k)\, p(y|\mathbf{x}^{(i)}, \boldsymbol{\mathcal{X}}, M_*)\, dy. \tag{34}$$

See section 3.2.3 for estimating the out-of-sample performance of the models with a cross validation reference utility (cross validation is discussed in section 3.2.2).

Gibbs reference utility In the second approach, one does not form the explicit predictive distribution a, i.e. $\boldsymbol{\theta}$ is not integrated out to obtain the posterior predictive distribution. Instead the logarithm score for model M_k is calculated straight from the likelihood for the future observation

$$u(M_k, \boldsymbol{\theta}_k, \tilde{y}) = \log p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\theta}_k, M_k), \tag{35}$$

and then the expectation of this is taken over the model posterior $p(\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}}, M_k)$

$$u(M_k, \boldsymbol{\theta}_k) = \int \log p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\theta}_k, M_k)\, p(\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}}, M_k)\, d\boldsymbol{\theta}_k. \tag{36}$$

This is a function of $\tilde{y}$ which gives higher values to those $\tilde{y}$ that are more probable according to the model. The average predictive performance of model M_k is thereby this function integrated over the posterior predictive distribution of the reference model, which gives us the *Gibbs reference utility*:

$$u_{\text{ref}}^{\text{G}}(M_k) = \int \left[\int \log p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\theta}_k, M_k)\, p(\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}}, M_k)\, d\boldsymbol{\theta}_k\right] p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\mathcal{X}}, M_*)\, d\tilde{y}. \tag{37}$$

In mathematical terms, the difference between (37) and (33) is the order of the logarithm and the inner integration. By Jensen's inequality the Gibbs utility is a lower bound for the Bayes utility. The Gibbs utility can be evaluated the same way as the Bayes utility, i.e. by averaging the expression (37) in the training points

$$u_{\text{ref}}^{\text{G}}(M_k) = \frac{1}{n}\sum_{i=1}^{n}\int \left[\int \log p(y|\mathbf{x}^{(i)}, \boldsymbol{\theta}_k, M_k)\, p(\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}}, M_k)\, d\boldsymbol{\theta}_k\right] p(y|\mathbf{x}^{(i)}, \boldsymbol{\mathcal{X}}, M_*)\, dy. \tag{38}$$

3.2.2 Utility estimation when no reference model is available

Training and test utilities When the reference model M_* is not available, the training data $\boldsymbol{\mathcal{X}} = (\mathbf{X}, \mathbf{y})$ can be reused as a proxy for the true distribution of the future observation. The simplest approach is to use the same data for model training and testing. This is done by evaluating the mean value of the utility function in the training points, which is called the *training utility*. With the logarithmic utility function the Bayes training utility is

$$u^{\mathrm{B}}_{\mathrm{train}}(M_k) = \frac{1}{n}\sum_{i=1}^{n} \log p(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\mathcal{X}}, M_k). \tag{39}$$

This estimate, however, is biased, since the testing points are not independent of the training points (we shall see this in section 4.3). A better approach is to use a separate test set where the data points are independent of the points used for training. If there is enough data, such a set can be formed simply by dividing the original training data into a new training and test sets. If such an independent test set $\tilde{\boldsymbol{\mathcal{X}}} = (\tilde{\mathbf{X}}, \tilde{\mathbf{y}})$ can be formed in one way or another, one can obtain the *test utility*

$$u^{\mathrm{B}}_{\mathrm{test}}(M_k) = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \log p(\tilde{y}^{(i)}|\tilde{\mathbf{x}}^{(i)}, \boldsymbol{\mathcal{X}}, M_k). \tag{40}$$

Test utility approaches the true generalization utility (28) as the number of testing points $\tilde{n}$ increases. For both of these, training and test utilities we can define the Gibbs utilities the same way as we did with the reference utility. The analogous Gibbs utilities are defined as

$$u^{\mathrm{G}}_{\mathrm{train}}(M_k) = \frac{1}{n}\sum_{i=1}^{n} \int \log p(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta}_k, M_k)\, p(\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}}, M_k) d\boldsymbol{\theta}_k \tag{41}$$

$$u^{\mathrm{G}}_{\mathrm{test}}(M_k) = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \int \log p(\tilde{y}^{(i)}|\tilde{\mathbf{x}}^{(i)}, \boldsymbol{\theta}_k, M_k)\, p(\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}}, M_k) d\boldsymbol{\theta}_k. \tag{42}$$

Cross validation Even if no separate test set $\tilde{\boldsymbol{\mathcal{X}}}$ is available, a better utility estimate than the training utility can be obtained with the cross validation. The idea in cross validation is to use each data point in the original training set for testing, but the same points are never used simultaneously for training and testing to achieve the independence between the training and testing points. Probably the most natural choice is to pick up one point at a time for testing, and train the model with the rest of the

data. This is referred to as *leave-one-out cross validation (LOO-CV)* and its utility estimate with the logarithmic score is given by

$$u_{\text{LOO}}(M_k) = \frac{1}{n}\sum_{i=1}^{n} \log p(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\mathcal{X}}^{(\setminus i)}, M_k), \tag{43}$$

where $\boldsymbol{\mathcal{X}}^{(\setminus i)}$ denotes the original training data set excluding ith point. Watanabe (2010) showed that LOO-CV with logarithmic utility function is asymptotically equal to the expected true utility and the error is $o(1/n)$. The drawback in LOO-CV is that in large modelling tasks the computation of (43) may be time consuming, since it requires solving the posterior predictive distribution n times. There are some ways to reduce the computational effort, such as importance sampling LOO-CV (IS-LOO-CV) but they are not considered in this study. Another cross validation method requiring less computing than LOO-CV is the *κ-fold-CV*, where the original training data is divided into κ subsets. Each of these subsets is used at a time for testing whereas the rest of the points are used for the training. The utility for κ-fold-CV is given by

$$u_{\kappa\text{-fold}}(M_k) = \frac{1}{n}\sum_{i=1}^{n} \log p(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\mathcal{X}}^{(\setminus I(i))}, M_k), \tag{44}$$

where $I(i)$ denotes the subset where ith point belongs to and hence $\boldsymbol{\mathcal{X}}^{(\setminus I(i))}$ refers to points in $\boldsymbol{\mathcal{X}}$ except those which belong to $I(i)$. In κ-fold-CV the computational burden is smaller than in LOO-CV since the predictive distribution needs to be solved only $\kappa < n$ times, but the flip side is that the smaller the κ, the smaller the number of points for training and the poorer the results. Note also that there is no unique way of dividing the data into κ sets, and thus also the results for κ-fold-CV are not unique.

Information criteria Another way of improving the utility estimate from the training utility (39) without a separate test set $\tilde{\boldsymbol{\mathcal{X}}}$ is to use *information criteria (IC)* methods, which can be usually defined in the form IC = training utility + bias correction. One often used method in Bayesian literature is the *deviance information criterion (DIC)* (Spiegelhalter et al., 2002) which uses the deviance function $\delta(\boldsymbol{\theta}_k) = -2\log p(y|\mathbf{x}, \boldsymbol{\theta}_k, M_k)$ to estimate the utility. DIC utility can be calculated as

$$\begin{aligned} u_{\text{DIC}}(M_k) &= -\frac{1}{2n}\left[2\text{E}_{\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}},M_k}[\delta(\boldsymbol{\theta}_k)] - \delta(\text{E}_{\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}},M_k}[\boldsymbol{\theta}_k])\right] \\ &= -\frac{1}{2n}\left[2\widehat{\delta(\boldsymbol{\theta}_k)} - \delta(\widehat{\boldsymbol{\theta}}_k)\right], \end{aligned} \tag{45}$$

where $\mathrm{E}_{\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}},M_k}[\cdot]$ denotes expectation over the model posterior $p(\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}}, M_k)$. One advantage in DIC is that it is easy to calculate, only samples from the posterior are needed in order to evaluate the expectations. The downside is, however, that it is dependent on a point estimate which means rejecting the uncertainty of the parameters. Due to the use of the plug-in estimate, the DIC value is also variant to the parametrization.

Another IC method is the *widely applicable information criterion (WAIC)* for which the utility estimate can be written as

$$u_{\mathrm{WAIC}}(M_k) = u^{\mathrm{B}}_{\mathrm{train}}(M_k) - \frac{V}{n}, \tag{46}$$

where V is the functional variance

$$V = \sum_{i=1}^{n} \left\{ \mathrm{E}_{\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}},M_k} \left[(\log p(y^{(i)}|\boldsymbol{\theta}_k, M_k))^2 \right] - \left(\mathrm{E}_{\boldsymbol{\theta}_k|\boldsymbol{\mathcal{X}},M_k} \left[\log p(y^{(i)}|\boldsymbol{\theta}_k, M_k) \right] \right)^2 \right\}. \tag{47}$$

Watanabe (2010) showed that WAIC is different from DIC but instead asymptotically equal to the LOO-CV and to the expected true utility, and therefore asymptotically unbiased. Note that WAIC is not dependent on a point estimate in contrast to DIC.

3.2.3 Other methods

Cross validation reference utility Even if the reference model is available, one may still be interested in estimating the out-of-sample performance of the candidate models, i.e. the performance outside the training set. In this case one can combine the idea of cross validation to the reference utility calculation and obtain a *cross validation reference utility*

$$u^{\mathrm{CV}}_{\mathrm{ref}}(M_k) = \frac{1}{n} \sum_{i=1}^{n} \int \log p(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\mathcal{X}}^{(\setminus I(i))}, M_k) p(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\mathcal{X}}^{(\setminus I(i))}, M_*) dy, \tag{48}$$

where $I(i)$ denotes the subset where ith point belongs to, and hence $\boldsymbol{\mathcal{X}}^{(\setminus I(i))}$ refers to points in $\boldsymbol{\mathcal{X}}$ except those which belong to $I(i)$. Note that (48) is a mixing of Bayes reference utility (34) and κ-fold cross validation (44). The major drawback in the reference cross validation is that it requires solving both the candidate models and the reference model κ times, which may lead to a huge computational effort. The advantage is, however, the use of cross validation should avoid the selection of an overfitted model.

4 Example problem: covariate selection

In this section we consider a covariate selection problem for a linear regression model derived in section 2.2. The used data and the model are explained first and we then consider how the various methods are calculated for a regression model (section 4.2) and then discuss about the results (section 4.3).

4.1 Data and model

We use the following simulation data (Vehtari and Lampinen, 2004):

$$\begin{aligned} z_1, ..., z_4 &\sim U(-1.73, 1.73) \\ x_1, ..., x_4 &\sim N(z_1, 0.045^2) \\ x_5, ..., x_8 &\sim N(z_2, 0.05^2) \\ x_9, ..., x_{12} &\sim N(z_3, 0.055^2) \\ x_{13}, ..., x_{16} &\sim N(z_4, 0.06^2) \\ y &= z_1 + 0.5z_2 + 0.25z_3 + \varepsilon \\ \varepsilon &\sim N(0, 0.5^2). \end{aligned}$$

In other words, in the data, we have one dependent variable y and 16 possible covariates $x_1, ..., x_{16}$. The x-variables do not determine y but instead are noisy indicators of the underlying true parameters z that are not observed. The covariates are divided into four groups and in each group the covariates are highly correlated with each other. We have a hundred replications of training and testing data $D_1, ..., D_{100}$. In each replication, the training set $\boldsymbol{\mathcal{X}}$ consists of $n = 20$ points and the test set $\tilde{\boldsymbol{\mathcal{X}}}$ of $\tilde{n} = 400$ points of data. Note that the number of training points in each replication is relatively small and close to the maximum number of covariates.

Given that the number of predictors is 16, the number of different subsets is 2^{16} so we have now $K = 2^{16}$ linear candidate models for the data. The task is to try to find one which would have a good predictive ability but would be as simple as possible, i.e. would contain only relevant covariates. Since the number of candidate models is still relatively small, we can apply the 'brute force'-approach and calculate all the utilities for all the models. However, in real world problems it may be impossible to go through all the candidate models since in difficult modelling tasks the number of possible models K may be far too great. For these purposes more advanced searching methods have been developed but they are not

considered in this study.

The basic model is given by the equation (6), where the dimensions of $\mathbf{X}$ and $\boldsymbol{\beta}$ vary according to the number of covariates in each model. For each model we use the conjugate prior (14), where the parameters are defined as:

$$\begin{aligned}\boldsymbol{\mu}_0 &= \begin{bmatrix}0 & \cdots & 0\end{bmatrix}^{\mathrm{T}} \\ \mathbf{A}_0 &= \tau_0 \mathbf{I}, \quad \tau_0 = 0.3 \\ a_0 &= 1 \\ b_0 &= 1.\end{aligned}$$

That is, all the regression coefficients β_i are defined to have zero prior mean and precision of 0.3 (variance 0.3^{-1}), and no covariances between the predictors are set a priori. The prior parameters $a_0 = 1$ and $b_0 = 1$ in the inverse gamma distribution lead in a relatively flat prior and an expectation of about 1.2^2 for the noise parameter σ^2. The sizes of vector $\boldsymbol{\mu}_0$ and matrix $\mathbf{A}_0$ naturally vary according to the number of covariates in each model, but other than that the prior parameters are similar for all the candidate models. With the prior specified, the posterior parameters (16)–(19) and the posterior predictive parameters (22)–(25) can be calculated.

4.2 Methods

For each of the models, we calculate the utilities using the methods described in sections 3.2.1–3.2.3. As a reference model M_* we use the BMA predictive distribution (4), i.e. $p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \boldsymbol{\mathcal{X}}, M_*) = p_{\mathrm{BMA}}(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \boldsymbol{\mathcal{X}})$ which can be justified if there is strong belief that these 16 predictors contain all the necessary information to describe y and in addition that the relationships between x_i and y are linear. In order to calculate $p_{\mathrm{BMA}}(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \boldsymbol{\mathcal{X}})$ we need the posterior predictive distributions (21) and model evidences (marginal likelihoods) (27) for each model M_k. In addition, we need to specify prior probabilities $p(M_k)$ for all the models. For simplicity, we set equal prior probabilities for each model, $p(M_k) = 1/K$. In this case we get for posterior probabilities (3)

$$\begin{aligned}p(M_k|\boldsymbol{\mathcal{X}}) &= \frac{p(\mathbf{y}|\mathbf{X}, M_k)p(M_k)}{\sum_{k=1}^{K} p(\mathbf{y}|\mathbf{X}, M_k)p(M_k)} = \frac{\frac{1}{K}p(\mathbf{y}|\mathbf{X}, M_k)}{\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{y}|\mathbf{X}, M_k)} \\ &= \frac{p(\mathbf{y}|\mathbf{X}, M_k)}{\sum_{k=1}^{K} p(\mathbf{y}|\mathbf{X}, M_k)}.\end{aligned}$$

Thus the posterior probabilities are proportional to the model evidences $p(\mathbf{y}|\mathbf{X}, M_k)$ (equation (27)). After these, the BMA is straightforward to calculate.

After the BMA and the predictive distributions for each of the models are solved, the Bayes and cross validation utilities are obtained in a straightforward fashion by plugging in the predictive densities and evaluating the necessary integrals numerically. The expressions for Gibbs utilities and information criterion can be expanded as some of the expectations can be calculated analytically. The derivation is, however, long and detailed and is not presented here. In addition to these, we shall evaluate the true Bayes and Gibbs generalization utilities (28) and (29) to illustrate the differences between the utility estimates of the different methods and the true utilities. This can be done since the data is determined by the simulations and we know the true data generating process is a normal distribution $p_t(y|\mathbf{z}) = N(y;\, m_t, s_t^2)$, where $m_t = z_1 + 0.5z_2 + 0.25z_3$ and $s_t^2 = 0.5^2$.

4.3 Results

Figure 2 shows the maximum mean utilities in all the datasets for different number of covariates (red dots). In other words, we calculated the utilities in all the datasets for each model, and figure 2 shows the mean utilities of the best models of each size. The genaralization utilities (calculated exactly the same way) are plotted for comparison (black dots). The differences between the red and black dots describe the asymptotic biasness of the utility estimates, and we see that both of the training utilities are significantly biased. The training utility increases with the model complexity which demonstrates the overfitting effect: the model fits better to the training data, but the generalization utility does not increase after 3 covariates, and even reduces after 12 covariates. On the other hand one can see the effect of using independent test data; the test utility gives almost exactly the same estimates as the true generalization utility. The reason that the reference utilities deviate from the generalization utilities is that the reference model is not good enough. In closer inspection, the BMA predictive distribution seemed to be somewhat flatter than the true data generating distribution (not shown) which at least partly explains the difference in the results. Nevertheless, an important point is that in contrast to the training utility, the reference utility does not increase monotonically with the number of covariates, therefore protecting from choosing an overfitted and unnecessarily complex model.

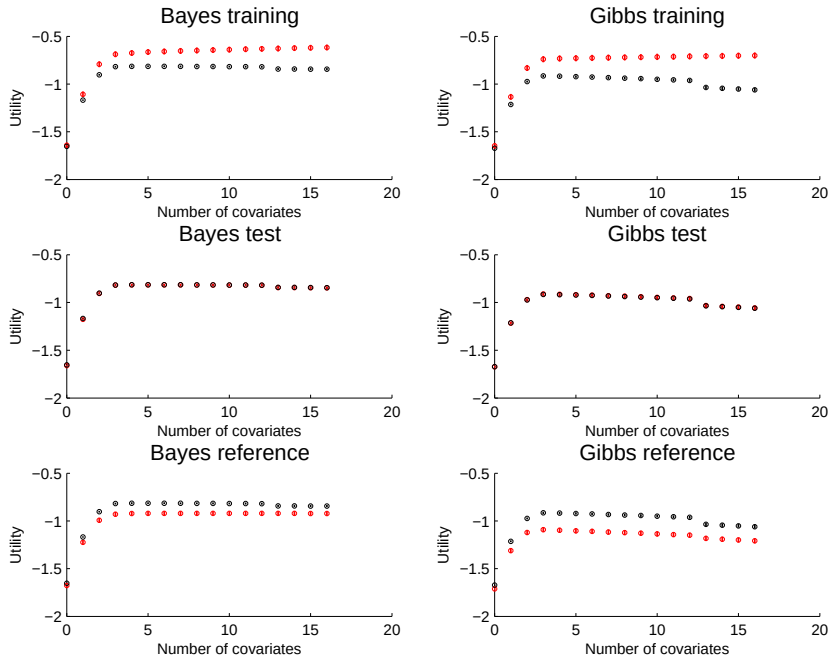

Figure 2: Maximum mean utility over all the datasets $D_1, ..., D_{100}$ for different number of covariates (red). In other words, with each number of covariates, the utility value is given by the best model of that size. Black dots show the corresponding Bayes or Gibbs true mean generalization utilities. 95%-confidence intervals are smaller than the dots.

Figure 3 shows the analogous results for the cross validations and information criteria. As can be seen, the cross validations and information criteria produce practically unbiased utility estimates as was discussed in section 3.2.2, and no significant differences can be seen. Again, the utility estimates of the reference cross validation method differ from the generalization utility for the same reason as the reference utility – the reference model is imperfect.

Given the asymptotical properties of the methods, we then consider what happens in the model selection. Figure 4 shows the maximum cross validation and information criteria utilities with different number of covariates for a single dataset D_1 (red dots). Red crosses mark the chosen models, i.e. the models with highest utility in the dataset D_1. One can see that unnecessarily large models may be selected when the selection is based on the utility maximization. For instance the 5-fold-CV chooses a model with

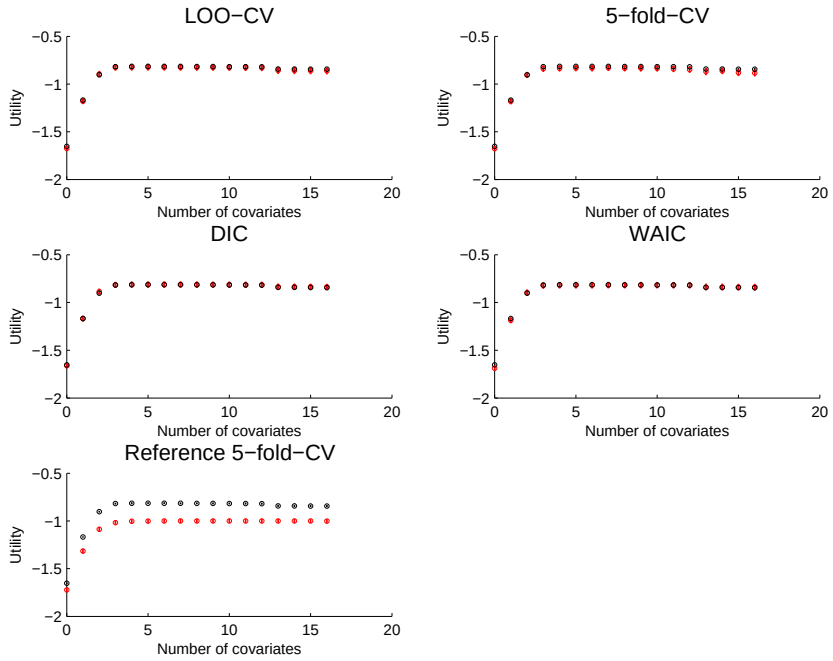

Figure 3: The same as in figure 2 but for cross validations and information criteria.

9 covariates even though the utility does not increase significantly after 3 covariates. This is an example of *the selection bias*. In this case the selection bias means that even if the real predictive performance does not increase after a certain amount covariates (3 in this case), some model containing for example 9 covariates may give the highest utility just by chance, since there are so many candidate models containing 9 predictors. The selection induced bias becomes even clearer when we consider how the best models with different number of covariates in D_1 perform on average in all the datasets $D_1, ..., D_{100}$; green dots show the mean over all the datasets and black lines denote the interval containing 95% of the values. This demonstrates how the models performing well in D_1 give on average significantly lower utilities when we change the dataset. The phenomenon seems to be clearest for the cross validations and information criteria due to their high variance in the utility estimates. The use of the reference model or independent test data reduces the variance, and in these cases also the selection bias is smaller (reference-CV in figure 4, and reference and test utilities in figure 5).

The selection bias is an important phenomenon in model selection (see

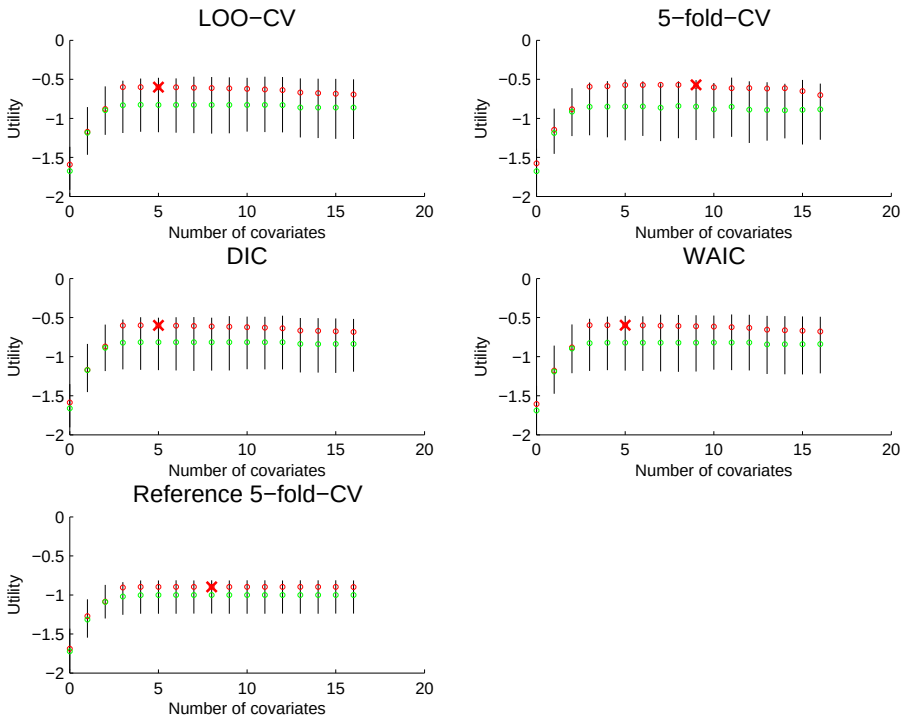

Figure 4: The maximum utilities for different number of covariates in dataset D_1 (red dots) and the chosen model (red cross). Green dots show the performance of those models on average in all the datasets $D_1, ..., D_{100}$ and black lines show the 95%-interval

e.g. Vehtari and Lampinen, 2004; Cawley and Talbot, 2010) which cannot be removed. There are, however, methods of estimating its magnitude but we do not consider them here. In short, the conclusion from this example is that there are asymptotically unbiased methods for estimating the true utilities of the given models, but after selecting a model by maximizing the utility estimate, the expected performance of the chosen model on unseen data is weaker than the utility estimate suggests.

5 Conclusions

In this study we discussed the theory and methods of predictive Bayesian model selection using expected utilities. We presented also a numerical example about applying these methods to a covariate selection problem. As we saw, the estimation of the predicitive ability and thereby the goodness of the model, can be done in several ways. How this is done depends on

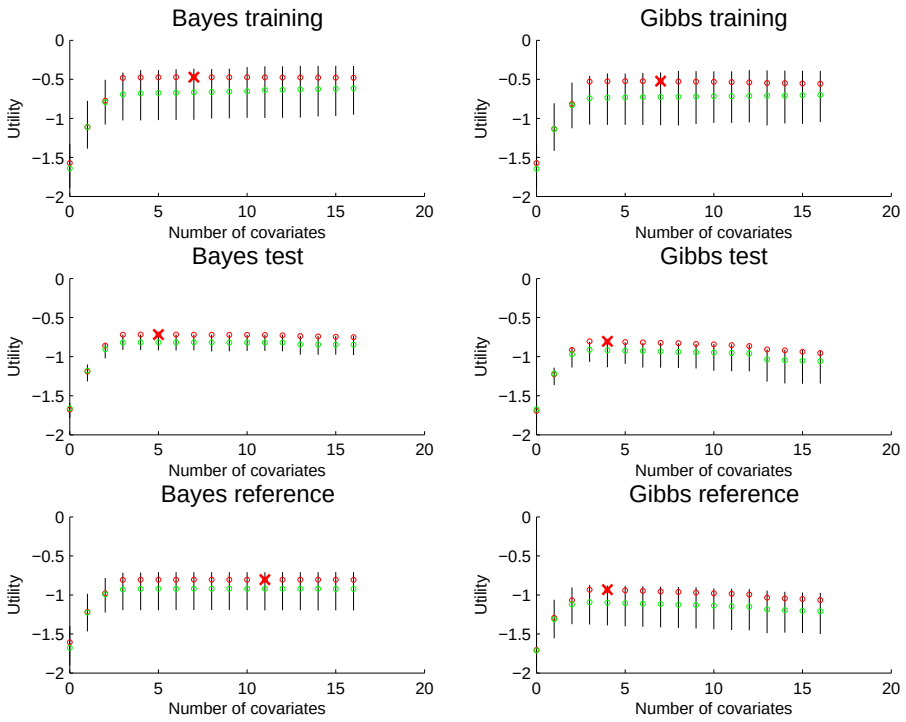

Figure 5: Same as in figure 4 but for Bayes and Gibbs utilities.

the available data and the modelling task (section 3).

In section 4 we demonstrated that if the predictive performance of a model is estimated by using the same data points that are used for model training the utility estimate becomes biased. A better approach is to avoid this by dividing the original data into the training and validation sets by using cross validation techniques, or to use a completely separate test set if possible. If there is not enough test data the biasness of the training utility can be corrected also by using the information criteria approaches. However, we demonstrated that even when using (almost) unbiased utility estimates, the model selection itself induces bias when the selection is based on maximizing the expected utility. This means that an unnecessarily large or complex model may be selected just by change, because it happens to fit well to the used dataset, but on average it is expected give lower utilities. The selection bias comes from overfitting to data and the probability of selecting a nonoptimal model increases with the number of considered models (Vehtari and Lampinen, 2004). The magnitude of selection bias can be reduced by choosing a method whose utility estimates have small variance (Cawley and Talbot, 2010).

References

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons Ltd., Chichester.

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics*, pages 201–236, New York. Academic Press.

Cawley, G. C. and Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, 11:2079–2107.

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B. Methodological*, 14:107–114.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: a tutorial. *Statistical Science*, 14(4):382–417.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Raftery, A. E. and Zheng, Y. (2003). Discussion: Performance of Bayesian Model Averaging. *Journal of the American Statistical Association*, 98(464):931–938.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 64(4):583–639.

Vehtari, A. and Lampinen, J. (2004). Model Selection via Predictive Explanatory Power. Technical report, Helsinki University of Technology, Laboratory of Computational Engineering.

Vehtari, A. and Ojanen, J. (2013). A Survey of Bayesian Predictive Methods for Model Assessment, Selection and Comparison. *Statistics Surveys*. Accepted for publication.

Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, Cambridge.

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11:3571–3594.

A Linear regression model with conjugate prior

Let the training data $\mathcal{X} = (\mathbf{X}, \mathbf{y})$ and the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ be defined as in section 2.2. The likelihood for a single observation is then the normal distribution

$$p(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{x}^{(i)}\boldsymbol{\beta})^2\right).$$

Assuming that the observations are conditionallly independent, the joint likelihood for all the observations is

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{x}^{(i)}\boldsymbol{\beta})^2\right) \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^{(i)} - \mathbf{x}^{(i)}\boldsymbol{\beta})^2\right) \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right). \end{aligned}$$

Now we need to modify the likelihood to get a distribution for $\boldsymbol{\beta}$ and σ^2. The inner product in the exponent can be written as

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y} \\ &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{M}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{M}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{M}\hat{\boldsymbol{\beta}} + C \\ &= (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{M}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + C, \end{aligned}$$

where

$$\begin{aligned} \mathbf{M} &= \mathbf{X}^{\mathrm{T}}\mathbf{X} \\ \mathbf{M}\hat{\boldsymbol{\beta}} &= \mathbf{X}^{\mathrm{T}}\mathbf{y} \quad \Rightarrow \hat{\boldsymbol{\beta}} = \mathbf{M}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ C &= \mathbf{y}^{\mathrm{T}}\mathbf{y} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{M}\hat{\boldsymbol{\beta}} \end{aligned}$$

assuming that $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$ exists. Hence the likelihood becomes

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} &\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}})\right) \\ &\times \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right), \end{aligned}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

Now this is a multivariate normal inverse gamma distribution for $\boldsymbol{\beta}$ and σ^2.

The conjugate prior is now of the same functional form as the likelihood, that is

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2) &= p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \\ &= N(\boldsymbol{\beta}; \boldsymbol{\mu}_0, \sigma^{-2}\mathbf{A}_0) \times \text{Inv-Gamma}(\sigma^2; a_0, b_0) \\ &= (2\pi)^{-\frac{p}{2}} \left|\sigma^{-2}\mathbf{A}_0\right|^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^{\mathrm{T}}\mathbf{A}_0(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\right) \\ &\quad \times \frac{b_0^{a_0}}{\Gamma(a_0)} \left(\sigma^2\right)^{-a_0-1} \exp\left(-\frac{b_0}{\sigma^2}\right). \end{aligned}$$

Here $\sigma^{-2}\mathbf{A}_0$ is the precision matrix, i.e the inverse of the covariance matrix. Now, dropping out the constants and completing the square for $\boldsymbol{\beta}$, the joint posterior becomes

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\mathcal{X}}) &\propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2) \\ &\propto \left(\sigma^2\right)^{-\frac{p}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_n)^{\mathrm{T}}\mathbf{A}_n(\boldsymbol{\beta}-\boldsymbol{\mu}_n)\right) \\ &\quad \times \left(\sigma^2\right)^{-a_n-1} \exp\left(-\frac{b_n}{\sigma^2}\right) \\ &\propto N(\boldsymbol{\beta}; \boldsymbol{\mu}_n, \sigma^{-2}\mathbf{A}_n) \times \text{Inv-Gamma}(\sigma^2; a_n, b_n), \end{aligned}$$

where

$$\begin{aligned} \mathbf{A}_n &= \mathbf{A}_0 + \mathbf{X}^{\mathrm{T}}\mathbf{X} \\ \boldsymbol{\mu}_n &= \mathbf{A}_n^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{A}_0\boldsymbol{\mu}_0) \\ a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\mu}_0^{\mathrm{T}}\mathbf{A}_0\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^{\mathrm{T}}\mathbf{A}_n\boldsymbol{\mu}_n). \end{aligned}$$

Thus the prior definition with parameters $\boldsymbol{\mu}_0, \mathbf{A}_0, a_0, b_0$ leads to a posterior of the same functional form with parameters defined above.

Let us calculate the joint posterior predictive distribution for $\tilde{n}$ future observations $\tilde{\mathbf{y}}$, assuming that the predictor values $\tilde{\mathbf{X}}$ in those points are

known. We obtain

$$\begin{aligned} p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \boldsymbol{\mathcal{X}}) &= \int\int p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\mathcal{X}}) d\sigma^2 d\boldsymbol{\beta} \\ &= \int\int (2\pi)^{-\frac{1}{2}(\tilde{n}+p)} |\mathbf{A}_n|^{\frac{1}{2}} \frac{b_n^{a_n}}{\Gamma(a_n)} \left(\sigma^2\right)^{-(\tilde{a}+\frac{p}{2})-1} \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\mu}})^{\mathrm{T}}\tilde{\mathbf{A}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\mu}})\right) \exp\left(-\frac{\tilde{b}}{\sigma^2}\right) d\sigma^2 d\boldsymbol{\beta}, \end{aligned}$$

where

$$\begin{aligned} \tilde{\mathbf{A}} &= \mathbf{A}_n + \tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}} \\ \tilde{\boldsymbol{\mu}}_n &= \tilde{\mathbf{A}}^{-1}(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{y}} + \mathbf{A}_n\boldsymbol{\mu}_n) \\ \tilde{a} &= a_n + \frac{\tilde{n}}{2} \\ \tilde{b} &= b_n + \frac{1}{2}(\tilde{\mathbf{y}}^{\mathrm{T}}\tilde{\mathbf{y}} + \boldsymbol{\mu}_n^{\mathrm{T}}\mathbf{A}_n\boldsymbol{\mu}_n - \tilde{\boldsymbol{\mu}}^{\mathrm{T}}\tilde{\mathbf{A}}\tilde{\boldsymbol{\mu}}). \end{aligned}$$

The inner integral can be calculated by taking the constants out and writing the rest in a form of a normalized inverse Gamma distribution for σ^2:

$$\begin{aligned} p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \boldsymbol{\mathcal{X}}) &= \int (2\pi)^{-\frac{1}{2}(\tilde{n}+p)} |\mathbf{A}_n|^{\frac{1}{2}} \frac{b_n^{a_n}}{\Gamma(a_n)} \frac{\Gamma(A)}{B^A} \\ &\qquad \underbrace{\int \frac{B^A}{\Gamma(A)} \left(\sigma^2\right)^{-A-1} \exp\left(-\frac{B}{\sigma^2}\right) d\sigma^2}_{=1} d\boldsymbol{\beta} \\ &= \int (2\pi)^{-\frac{1}{2}(\tilde{n}+p)} |\mathbf{A}_n|^{\frac{1}{2}} \frac{b_n^{a_n}}{\Gamma(a_n)} \frac{\Gamma(A)}{B^A} d\boldsymbol{\beta}, \end{aligned}$$

where

$$\begin{aligned} A &= \tilde{a} + \frac{p}{2} \\ B &= \tilde{b} + \frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\mu}})^{\mathrm{T}}\tilde{\mathbf{A}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\mu}}). \end{aligned}$$

After rewriting, the resulting part becomes a multiplication of a constant and a multivariate t–distribution for $\boldsymbol{\beta}$ with location $\tilde{\boldsymbol{\mu}}$, precision matrix $\frac{\nu\tilde{\mathbf{A}}}{2\tilde{b}}$ and ν degrees of freedom:

$$\begin{aligned} A = \tilde{a} + \frac{p}{2} &:= \frac{\nu + p}{2} \\ \Rightarrow \nu &= 2\tilde{a} \end{aligned}$$

and

$$\begin{aligned} B &= \tilde{b} + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}})^{\mathrm{T}} \tilde{\mathbf{A}} (\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}}) \\ &= \tilde{b}\left(1 + \frac{1}{\nu}(\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}})^{\mathrm{T}} \frac{\nu \tilde{\mathbf{A}}}{2\tilde{b}} (\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}})\right) \\ \Rightarrow B^A &= \tilde{b}^A \left(1 + \frac{1}{\nu}(\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}})^{\mathrm{T}} \frac{\nu \tilde{\mathbf{A}}}{2\tilde{b}} (\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}})\right)^A \\ &= \tilde{b}^{(\nu+p)/2} \left(1 + \frac{1}{\nu}(\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}})^{\mathrm{T}} \frac{\nu \tilde{\mathbf{A}}}{2\tilde{b}} (\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}})\right)^{(\nu+p)/2}. \end{aligned}$$

Hence

$$\begin{aligned} p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}, \boldsymbol{\mathcal{X}}) &= (2\pi)^{-(\tilde{n}+p)/2} |\mathbf{A}_n|^{\frac{1}{2}} \frac{b_n^{a_n}}{\Gamma(a_n)} \tilde{b}^{-(\nu+p)/2} \Gamma(\frac{\nu}{2}) (\pi\nu)^{p/2} \left|\frac{\nu \tilde{\mathbf{A}}}{2\tilde{b}}\right|^{-\frac{1}{2}} \\ &\quad \times \underbrace{\int \frac{\Gamma(\frac{\nu+p}{2}) \left|\frac{\nu \tilde{\mathbf{A}}}{2\tilde{b}}\right|^{\frac{1}{2}}}{\Gamma(\frac{\nu}{2})(\pi\nu)^{p/2} \left(1 + \frac{1}{\nu}(\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}})^{\mathrm{T}} \frac{\nu \tilde{\mathbf{A}}}{2\tilde{b}} (\boldsymbol{\beta} - \tilde{\boldsymbol{\mu}})\right)^{(\nu+p)/2}} d\boldsymbol{\beta}}_{=1} \\ &= (2\pi)^{-(\tilde{n}+p)/2} |\mathbf{A}_n|^{\frac{1}{2}} \frac{b_n^{a_n}}{\Gamma(a_n)} \tilde{b}^{-(\nu+p)/2} \Gamma(\frac{\nu}{2}) (\pi\nu)^{p/2} \left(\frac{\nu}{2\tilde{b}}\right)^{-p/2} |\tilde{\mathbf{A}}|^{-\frac{1}{2}} \\ &= (2\pi)^{-\tilde{n}/2} \frac{\Gamma(\frac{\nu}{2})}{\Gamma(a_n)} \left(\frac{|\mathbf{A}_n|}{|\tilde{\mathbf{A}}|}\right)^{\frac{1}{2}} \frac{b_n^{a_n}}{\tilde{b}^{\nu/2}} \\ &= (2\pi)^{-\tilde{n}/2} \frac{\Gamma(\tilde{a})}{\Gamma(a_n)} \left(\frac{|\mathbf{A}_n|}{|\tilde{\mathbf{A}}|}\right)^{\frac{1}{2}} \frac{b_n^{a_n}}{\tilde{b}^{\tilde{a}}} \end{aligned}$$

This is the joint posterior predictive distribution for $\tilde{n}$ future observations, given the predictors $\tilde{\mathbf{X}}$. The marginal likelihood $p(\mathbf{y}|\mathbf{X})$ is now given by

$$p(\mathbf{y}|\mathbf{X}) = \int\int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) d\sigma^2 d\boldsymbol{\beta}.$$

Because the prior and posterior are of the same form, this integral is exactly the same as with the predictive distribution. Thus simply replacing the posterior parameters with the prior parameters and $\mathbf{X} \to \tilde{\mathbf{X}}$, $\mathbf{y} \to \tilde{\mathbf{y}}$ and $\tilde{n} \to n$ in the likelihood we obtain

$$p(\mathbf{y}|\mathbf{X}) = (2\pi)^{-n/2} \frac{\Gamma(a_n)}{\Gamma(a_0)} \left(\frac{|\mathbf{A}_0|}{|\mathbf{A}_n|}\right)^{\frac{1}{2}} \frac{b_0^{a_0}}{b_n^{a_n}}.$$