Juho Piironen

Comparison of Bayesian predictive methods for variable selection

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo August 3, 2014

Thesis supervisor:

Prof. Jouko Lampinen

Thesis advisor:

D.Sc. (Tech.) Aki Vehtari

Aalto University
School of Science

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Juho Piironen

Työn nimi: Bayesilaisten prediktiivisten muuttujavalintamenetelmien vertailu

Päivämäärä: 3.8.2014 Kieli: Englanti Sivumäärä: 6+60

Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Professuuri: Laskennallinen tekniikka Koodi: Becs-114

Valvoja: Prof. Jouko Lampinen

Ohjaaja: TkT Aki Vehtari

Kirjallisuudessa on esitetty useita erilaisia menetelmiä bayesilaiseen mallin valintaan. Vaikka näiden menetelmien teoreettisia ominaisuuksia erityisesti mallin suorituskyvyn mittaamiseen on tutkittu runsaasti, kattavaa tutkimusta eri menetelmien eroista mallin valintaan äärelliselle aineistolle ei näytä olevan tehty. Tässä työssä käsitellään yleisimmin käytettyjä mallinvalintamenetelmiä ja vertaillaan näiden käyttäytymistä käytännön muuttujavalintaongelmissa, erityisesti tilanteissa joissa dataa on niukasti. Työn tarkoituksena on käsitellä myös valintaharhaksi kutsuttua ilmiötä ja korostaa sen merkitystä muuttujavalintaongelmissa. Vaikka työ käsittelee pääosin muuttujavalintaa, työssä esitetyt johtopäätökset ovat yleistettävissä myös muihin mallinvalintaongelmiin. Numeeriset esimerkit koostuvat simuloiduista testeistä sekä yhdestä reaalimaailman ongelmasta. Tulosten perusteella näyttää siltä, että vaikka yksittäisten mallien suorituskykyä voidaan arvioida harhattomasti, valintaharha voi vaikeuttaa mallinvalintaa huomattavasti ja johtaa ylisovittuneen mallin valintaan. Näyttää myös siltä, että referenssiprediktiiviset ja projektiomenetelmät ovat vähiten herkkiä valinnan aiheuttamalle harhalle ja kykenevät näin ollen löytämään parempia malleja kuin vaihtoehtoiset menetelmät kuten ristiinvalidointi ja informaatiokriteerit. Valintaharhasta johtuen kuitenkin myös näille menetelmille estimoitu eroavuus referenssimallin ja kandidaattimallien välillä voi antaa epäluotettavan kuvan valittujen mallien suorituskyvystä. Tästä syystä lopullinen valittujen mallien suorituskyvyn arviointi tulisi tehdä käyttäen esimerkiksi valintaprosessin ulkopuolista ristiinvalidointia.

Avainsanat: bayesilainen mallinvalinta, muuttujavalinta, valintaharha, ristiinvalidointi, informaatiokriteerit, referenssimalli, projektio

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Juho Piironen

Title: Comparison of Bayesian predictive methods for variable selection

Date: August 3, 2014 Language: English Number of pages: 6+60

Department of Biomedical Engineering and Computational Science

Professorship: Computational Engineering Code: Becs-114

Supervisor: Prof. Jouko Lampinen

Advisor: D.Sc. (Tech.) Aki Vehtari

To date, several methods for Bayesian model selection have been proposed. Although there are many studies discussing the theoretical properties of these methods for model assessment, an extensive quantitative comparison between the methods for model selection for finite data seems to be lacking. This thesis reviews the most commonly used methods in the literature and compares their performance in practical variable selection problems, especially in situations where the data is scarce. The study also discusses the selection induced bias in detail and underlines its relevance for variable selection. Although the focus of the study is on variable selection, the presented ideas are generalizable to other model selection problems as well. The numerical results consist of simulated experiments and one real world problem. The results suggest that even though there are nearly unbiased methods for assessing the performance of a given model, the high variance in the performance estimation may lead to considerable selection induced bias and selection of an overfitted model. The results also suggest that the reference predictive and projection methods are least sensitive to the selection induced bias and are therefore more robust for searching promising models than the alternative methods, such as cross validation and information criteria. However, due to the selection bias, also for these methods the estimated divergence between the reference and candidate models may be an unreliable indicator of the performance of the selected models. For this reason, the performance estimation of the found models should be done for example using cross validation outside the selection process.

Keywords: Bayesian model selection, variable selection, selection induced bias, cross validation, information criteria, reference model, projection

Preface

This work was carried out in the Bayesian Statistical Methods research group in the Department of Biomedical Engineering and Computational Science (BECS) at Aalto University. First, I would like to thank my advisor Aki Vehtari for proficient guidance and Jouko Lampinen for supervising this thesis. I would also like to thank the whole Bayes group for all the help I have had during the past two years at BECS.

Writing the thesis and completing the studies may sometimes be hard, but the right people around you can make the burden much easier to carry. I would like to express my sincere gratitude to my family and friends for all the support along my studies and during the work on this thesis. Especially I would like to thank my dear Aura who has been there for me during the stressful and even desperate times I have encountered along the way.

Otaniemi, August 3, 2014

Juho Piironen

Contents

Notation

General notation and symbols

a, b, c	Scalars (lower case)
$\mathbf{a}, \mathbf{b}, \mathbf{c}$	Column vectors (bold lower case)
$\mathbf{A}, \mathbf{B}, \mathbf{C}$	Matrices (bold capitals)
$\mathbf{A}^T$	Matrix transpose
$\mathbf{A}^{-1}$	Matrix inverse
$\mathbf{w}$	Weights of the input variables
$\mathbf{x}$	Input variables
$\tilde{\mathbf{x}}$	Input variables for future observation
$\mathbf{X}$	Matrix of input variables at training observations
$\tilde{\mathbf{X}}$	Matrix of input variables for several future observations
y	Output or target variable
$\tilde{y}$	Future output variable
$\mathbf{y}$	Output variables at training observations
$\tilde{\mathbf{y}}$	Future output variables
$\boldsymbol{\theta}$	Vector of parameters
σ^2	Noise variance
M	Model structure and model assumptions
M_*	Reference model
$\{t_s\}_{s=1}^S$	Set $\{t_1, t_2, \ldots, t_S\}$
$\mathcal{D}$	Observed dataset of n observations $\mathcal{D} = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$

Operators

$\mathrm{E}[\cdot]$	Expected value
$\mathrm{KL}(p \mid\mid q)$	Kullback-Leibler divergence between distributions p and q (from p to q)
$\mathrm{Var}[\cdot]$	Variance

Abbreviations

AIC	Akaike information criterion
BMA	Bayesian model average
CV	Cross validation
DIC	Deviance information criterion
LOO-CV	Leave-one-out cross validation
MAP	Maximum a posteriori
MCMC	Markov chain Monte Carlo
MLPD	Mean log predictive density
RJMCMC	Reversible jump Markov chain Monte Carlo
WAIC	Widely applicable information criterion

1 Introduction

Model selection is one of the fundamental problems in statistical modeling and machine learning. The generally accepted view for model selection is the famous quote by G. E. Box saying that "all models are wrong, but some are useful" (Box, 1979). Usefulness of a model is typically measured by its ability to make predictions about the future (or yet unseen) observations. This does not necessarily mean that the most important goal of modeling is to be able to predict the future as well as possible. The modeling practitioner may not only be interested in what the observations will be, but more importantly why they are like they appear. However, comparing the model predictions with the observations is still the most important tool of a statistician for figuring out whether the model makes sense or not. If the model gives poor predictions, there is not much sense in trying to interpret the model.

This thesis considers 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 estimated ability to predict unseen observations. The main focus will be on variable selection, even though most of the ideas will be generalizable to other model selection problems as well. Roughly speaking, in variable selection the goal is to select a minimal subset from a predefined set of possible input variables for model construction while maximizing the predictive ability of the model. Note, however, that there is typically a tradeoff between the number of included variables and the predictive performance of the model. Thus it depends on the situation whether the preference is the reduction of the variables or the predictive performance of the model.

Variable selection (or feature selection) is an important and widely studied problem in statistics and machine learning. The usual assumption is that the set of candidate variables contains many irrelevant variables that have no predictive power on the target variables. Some of the variables may also be highly dependent, in which case some of them can be removed without a significant loss in the predictive ability. A usual benefit from a successful variable selection is improved model interpretability. If the predictive ability of the model does not suffer or even improves when some of the variables are left out, it makes sense to believe that those variables have little to do with the underlying process that generates the observations. Some other advantages could be reduced computation time and measurement costs, if some of the variables need not be measured and subsequent model construction can be done with a smaller number of inputs.

There exists a number of proposed methods for Bayesian predictive model selection. Some of them are designed purely for variable selection, especially for linear models, and some of them are applicable also to more general model selection problems. The focus of this thesis is not to review all the proposed methods because an extensive qualitative review of the suggested methods to date is already done by Vehtari and Ojanen (2012). Instead the contribution of this thesis is to compare these methods quantitatively in practical model selection problems and discuss the differences. Our point of view will be pragmatic; the goal is to answer the question

how well the models selected by each method generalize on unseen data. This question is tackled by performing various numerical experiments. As it turns out, there are substantial differences between some of the approaches, and some of the methods are not necessarily well suited for variable selection. That is, they select variable combinations that have poor predictive ability. This is due to a phenomenon called selection bias. One of the goals in this thesis is to understand how selection bias affects model selection and why it is highly relevant to variable selection.

The thesis is structured as follows. Section 2 gives a brief introduction to Bayesian statistics. The fundamentals of Bayesian inference are discussed, such as how to update beliefs about the unknown quantities using Bayes' theorem, and how to make predictions with a Bayesian model. This section also discusses the linear regression and binary classification models that are used in the numerical experiments in section 5. Section 3 reviews many of the proposed approaches to Bayesian model selection and discusses differences and connections between them. This section also discusses the selection induced bias in detail and discusses its relevance for variable selection. Section 4 discusses special topics related to variable selection such as strategies for searching for variable combinations and sampling the model space. Section 5 is devoted to the numerical experiments. The first part of the section deals with numerical data and investigates the performance of the different model selection methods for different datasets. The second part deals with a challenging real world dataset further illustrating the key points. The discussion in section 6 concludes the thesis.

2 Bayesian modelling

This section gives a brief introduction to Bayesian modeling. The goal is to give a short recap to the basic concepts that will aid the material in the later sections. The section also briefly discusses the Bayesian linear regression and binary classification models that are used in the numerical experiments in section 5.

2.1 Basic principles

In Bayesian statistics uncertain quantities are modeled as random variables, and in essence the whole Bayesian statistics is about revising subjective beliefs about these quantities on the basis of observations (see for example Gelman et al., 2013a). This interpretation is fundamentally different from what is done in frequentist statistics, where the quantities of interest are usually considered as 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 according to the Bayes' theorem

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

The first term in the numerator $p(\mathbf{y} \mid \boldsymbol{\theta}, M)$ is the statistical observation model describing the relation between the parameters and the observations. More precisely, it defines the probability of the observations $\mathbf{y}$ given the parameters $\boldsymbol{\theta}$. Once the observations are obtained, it becomes a function of $\boldsymbol{\theta}$ and it is called 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. It is important to note the distinction between the observation model and the likelihood, even though these terms are often used interchangeably. A simple example of an observation model would be the Gaussian distribution for univariate y, where the unknown parameters are the mean and the variance $\boldsymbol{\theta} = (\mu, \sigma^2)$. The likelihood would then be a bivariate function for μ and σ^2.

The second term in the numerator is the *prior distribution* $p(\boldsymbol{\theta} \mid M)$ which describes modeler's beliefs about $\boldsymbol{\theta}$ before making the observations. The multiplication of the prior and the likelihood gives a new unnormalized distribution which combines the prior information and the information from the obtained data sample. The normalization by the constant $p(\mathbf{y} \mid M)$ gives the left hand side $p(\boldsymbol{\theta} \mid \mathbf{y}, M)$, which is called the *posterior distribution*. The posterior is the most central part of Bayesian inference because it contains all the knowledge there is about $\boldsymbol{\theta}$ after the observations, and it will be in the key role when making the predictions. The denominator $p(\mathbf{y} \mid M)$ has many names depending on the context, but is most often referred to as the *marginal likelihood*. The marginal likelihood ensures that the right hand side becomes a proper probability distribution but has no effect on its shape.

Bayesian treatment has a natural way of predicting new observations $\tilde{\mathbf{y}}$ given the old ones. The predictive distribution for future values is obtained by marginalizing

the parameters $\boldsymbol{\theta}$ out of the joint posterior $p(\tilde{\mathbf{y}}, \boldsymbol{\theta} \mid \mathbf{y}, M)$ as

$$\begin{aligned} p(\tilde{\mathbf{y}} \mid \mathbf{y}, M) &= \int p(\tilde{\mathbf{y}}, \boldsymbol{\theta} \mid \mathbf{y}, M) d\boldsymbol{\theta} \\ &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}, \mathbf{y}, M) p(\boldsymbol{\theta} \mid \mathbf{y}, M) d\boldsymbol{\theta} \\ &= \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\theta}, M) p(\boldsymbol{\theta} \mid \mathbf{y}, M) d\boldsymbol{\theta} \,. \end{aligned} \tag{2}$$

The second line follows from the definition of the conditional probability $p(A, B) = p(A \mid B)p(B)$, and the third line from the assumption that for a parametric model the future observations $\tilde{\mathbf{y}}$ are conditionally independent of the observed $\mathbf{y}$ given the parameters. The distribution $p(\tilde{\mathbf{y}} \mid \mathbf{y}, M)$ is called the *posterior predictive distribution* and according to its name, it gives the distribution for the future observations given the previous ones and the model assumptions. As is seen from (2), the predictive distribution is obtained by integrating the observation model over the posterior of the parameters. The predictive distribution takes into account the uncertainty about the model parameters as well as the stochastic randomness of the future observation. One may see the analogy between the marginal likelihood $p(\mathbf{y} \mid M)$ in (1) and posterior predictive distribution in (2); in the first case the observation model is integrated over the prior and in the latter over the posterior. For this reason, the term $p(\mathbf{y} \mid M)$ is also called the prior predictive distribution.

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

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

Here $p(M)$ and $p(M \mid \mathbf{y})$ are discrete probability distributions determining the prior and posterior probabilities, respectively, for each model M, and $p(\mathbf{y} \mid M)$ is the marginal likelihood from equation (1), also called the *model evidence*. In other words, after specifying prior probabilities for each model, $p(M \mid \mathbf{y})$ 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 one can integrate or sum over the model space to get the *Bayesian model averaging* (BMA) predictive distribution

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

BMA predictive distribution takes all the K models into account according to their relative probabilities. 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, that is, models outside this set are not possible. Usually this

assumption is not taken literally as typically none of the models can be considered to be "true". Model averaging may also lead to poor results if the set of the candidate models is poorly specified. However, BMA has been shown to have a good predictive performance both theoretically and empirically (Raftery and Zheng, 2003). From a practical point of view what matters is whether one is able to come up with a reasonably good set of candidate models that give good predictions. See review by Hoeting et al. (1999) for thorough discussion of Bayesian model averaging.

2.2 Linear regression

This section discusses Bayesian linear regression. Section 2.2.1 deals with the model with the conjugate prior for which the posterior and predictive distribution are obtained analytically. Section 2.2.2 discusses non-conjugate priors that can be used to promote sparsity in problems with possibly irrelevant variables.

2.2.1 Model with the conjugate prior

Consider a general regression problem where we want to predict a single real valued output or target variable y with p input or predictor variables $\mathbf{x} = (x_1, ..., x_p)$. Typically one fixes the first input variable to a constant $x_1 = 1$. This is a very general setting and many statistical models are of this form. In linear regression the relationship between the output variable and the predictors is modeled as linear. Given a training dataset $\mathcal{D} = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ the model is

$$y_i = \mathbf{w}^T \mathbf{x}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right) \tag{5}$$

or equivalently

$$y_i \sim \mathrm{N}\left(\mathbf{w}^T \mathbf{x}_i, \sigma^2\right) . \tag{6}$$

Here $\mathbf{w} = (w_1, \ldots, w_p)$ are the weights for each variable. The model can be written in the matrix form

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \tag{7}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} . \tag{8}$$

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

$$p(\mathbf{w}, \sigma^2 \mid \mathcal{D}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w}, \sigma^2)}{p(\mathbf{y} \mid \mathbf{X})} . \tag{9}$$

It is important to note here that the observations $\mathbf{y}$ are conditioned on the predictors $\mathbf{X}$, so these values are assumed to be known. Assuming that the observations are conditionally independent, the joint likelihood for all the observations is

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \mathbf{w}, \sigma^2) \tag{10}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{w}^T\mathbf{x}_i)^2\right). \tag{11}$$

This can be shown to be proportional to multivariate normal-inverse-gamma distribution for $\mathbf{w}$ and σ^2

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

where

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

This is the familiar least squares estimate (LS) and also the maximum likelihood solution for the parameters $\mathbf{w}$. The conjugate prior is now given by a distribution of the same form as the likelihood (12)

$$\begin{aligned} p(\mathbf{w}, \sigma^2) &= p(\mathbf{w} \mid \sigma^2)p(\sigma^2) \\ &= \mathrm{N}\left(\mathbf{w} \mid \boldsymbol{\mu}_0, \sigma^2\mathbf{A}_0^{-1}\right) \times \text{Inv-Gamma}\left(\sigma^2 \mid a_0, b_0\right). \end{aligned} \tag{14}$$

Here the covariance of the weights is formulated via the precision matrix $\mathbf{A}_0$ because this is more convenient when finding the parameters of the posterior distribution. A typical prior for the weights is $\boldsymbol{\mu}_0 = 0$ and $\mathbf{A}_0 = \tau^{-2}\mathbf{I}$, where τ^2 adjusts the prior variance for each weight determining the amount of regularization (that is, how strongly each of the weights is driven towards zero). Now one can straightforwardly multiply (12) and (14) to get the unnormalized joint posterior

$$\begin{aligned} p(\mathbf{w}, \sigma^2 \mid \mathcal{D}) &\propto p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma^2)p(\mathbf{w}, \sigma^2) \\ &\propto N(\mathbf{w} \mid \boldsymbol{\mu}_n, \sigma^2\mathbf{A}_n^{-1}) \times \text{Inv-Gamma}(\sigma^2 \mid a_n, b_n), \end{aligned} \tag{15}$$

where

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

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

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

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

Given that the unnormalized posterior is proportional to the normal-inverse-gamma distribution (15), 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 to a posterior of the same form with parameters determined by the equations (16)–(19).

After solving the posterior, one can calculate the posterior predictive distribution (2). Consider the joint predictive distribution for $\tilde{n}$ future observations $\tilde{\mathbf{y}}$, assuming that the input variables $\tilde{\mathbf{X}}$ are known. Straight from the definition (2) one obtains

$$p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathcal{D}) = \int\int p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathbf{w}, \sigma^2) p(\mathbf{w}, \sigma^2 \mid \mathcal{D}) d\sigma^2 d\mathbf{w}\,. \tag{20}$$

This integral can be calculated analytically and the result is

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

where

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

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

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

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

Distribution (21) can be shown to be a multivariate t-distribution for $\tilde{n}$ future observations $\tilde{\mathbf{y}}$, given the inputs $\tilde{\mathbf{X}}$

$$p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathcal{D}) = t_{\tilde{\nu}} \left(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}} \boldsymbol{\mu}_n, \frac{b_n}{a_n} \left(\mathbf{I} + \tilde{\mathbf{X}} \mathbf{A}_n^{-1} \tilde{\mathbf{X}}^T\right)\right), \tag{26}$$

where the degrees of freedom is $\tilde{\nu} = 2a_0 + n$. Thus for large n, the predictive distribution is nearly Gaussian. Usually one is interested in predicting one point at a time, and in this case one simply sets $\tilde{\mathbf{y}} = \tilde{y}$ and $\tilde{\mathbf{X}} = \tilde{\mathbf{x}}^T$.

For the conjugate prior, one can also calculate the marginal likelihood $p(\mathbf{y} \mid \mathbf{X})$, even though it was not needed to derive the posterior. The marginal likelihood is given by

$$p(\mathbf{y} \mid \mathbf{X}) = \int\int p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w}, \sigma^2) d\sigma^2 d\mathbf{w}\,. \tag{27}$$

Since the prior $p(\mathbf{w}, \sigma^2)$ and the posterior $p(\mathbf{w}, \sigma^2 \mid \mathcal{D})$ are of the same functional form (because of the conjugacy), this integral is identical to the one encountered 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

$$p(\mathbf{y} \mid \mathbf{X}) = (2\pi)^{-\frac{n}{2}} \frac{\Gamma(a_n)}{\Gamma(a_0)} \left(\frac{|\mathbf{A}_0|}{|\mathbf{A}_n|}\right)^{\frac{1}{2}} \frac{b_0^{a_0}}{b_n^{a_n}}\,. \tag{28}$$

The analytical result for the marginal likelihood is useful for example if one wants to handle the hyperparameter τ^2 in the prior precision $\mathbf{A}_0 = \tau^{-2}\mathbf{I}$ as an unknown. Given a prior $p(\tau^2)$, the predictions can be obtained by marginalizing over τ^2 as

$$\begin{aligned} p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathcal{D}) &= \int p(\tilde{\mathbf{y}}, \tau^2 \mid \tilde{\mathbf{X}}, \mathcal{D}) d\tilde{\mathbf{y}} \\ &= \int p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathcal{D}, \tau^2) p(\tau^2 \mid \tilde{\mathbf{X}}, \mathcal{D}) d\tilde{\mathbf{y}}, \end{aligned} \tag{29}$$

where $p(\tau^2 \mid \tilde{\mathbf{X}}, \mathcal{D})$ is the posterior of τ^2 and $p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathcal{D}, \tau^2)$ is the predictive distribution (26) which is conditioned on τ^2 (even though τ^2 was left out from the notation there). The integral (29) is not available analytically, but can be efficiently approximated for example by introducing a grid of points $\{\tau_k^2\}_{k=1}^N = \{\tau_1^2, \ldots, \tau_N^2\}$ and then writing

$$p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathcal{D}) \approx \frac{\sum_k v_k \, p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathcal{D}, \tau_k^2)}{\sum_k v_k}, \tag{30}$$

where the grid weights v_k are proportional to the posterior density of τ^2

$$v_k = p(\mathbf{y} \mid \mathbf{X}, \tau_k^2) p(\tau_k^2), \quad k = 1, \ldots, N \,. \tag{31}$$

Here $p(\mathbf{y} \mid \mathbf{X}, \tau_k^2)$ is the marginal likelihood (28) for a given τ_k^2. The approximation (30) is quick to compute and is reasonably accurate even for a relatively small number of grid points as long as the grid covers the range where $p(\tau^2 \mid \tilde{\mathbf{X}}, \mathcal{D})$ has almost all of its mass.

2.2.2 Sparsity promoting priors

The conjugate normal-inverse-gamma prior has the advantage of being computationally appealing because all the calculations are obtained analytically (except the integration over the hyperparameter τ^2). However, for problems where it is known a priori that some of the variables are irrelevant and have almost a zero weight, one would like to incorporate this information to the prior structure to improve the model fitting. The conjugate prior for a single weight

$$p(w_j \mid \sigma^2) = \mathrm{N}\left(w_j \mid 0, \tau^2\sigma^2\right) \tag{32}$$

has the drawback that it does not promote sparsity because all the variables have a zero probability of being exactly zero. An alternative is to use the *spike-and-slab* prior (Mitchell and Beauchamp, 1988)

$$p(w_j) = \pi_j \mathrm{N}\left(w_j \mid 0, \tau^2\right) + (1 - \pi_j)\delta_0(w_j) \tag{33}$$

where δ_0 is Dirac's delta function having mass only at zero. In other words, the variable x_j is included with probability π_j and its weight has a noninformative

Gaussian prior ("slab"), and with probability $1 - \pi_j$ it is excluded by setting its weight to zero ("spike"). Using the conjugate Gaussian in the slab-part

$$p(w_j \mid \sigma^2) = \pi_j \mathrm{N}\left(w_j \mid 0, \tau^2\sigma^2\right) + (1 - \pi_j)\delta_0(w_j) \tag{34}$$

is equivalent to forming the Bayesian model average from the individual models with the conjugate priors (see section 4.2). The spike-and-slab is a widely used prior and often considered as the "golden standard" approach for sparse estimation (see e.g. George and McCulloch, 1993, 1997; Raftery et al., 1997; Hoeting et al., 1999; Carvalho et al., 2009; Titsias and Lázaro-Gredilla, 2011). This comes with the cost that the predictions are no more analytically available, and one must use approximative solutions instead. Typically this is done by drawing a posterior sample $\{(\mathbf{w}_s, \sigma_s^2)\}_{s=1}^S$ with a suitable Markov chain Monte Carlo -method (MCMC) and then approximating the predictive distribution (20) as

$$p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S} p(\tilde{\mathbf{y}} \mid \tilde{\mathbf{X}}, \mathbf{w}_s, \sigma_s^2)\,. \tag{35}$$

The sampling of the model parameters under the spike-and-slab prior is discussed in section 4.2. Although the sampling scheme is used in this thesis, it is noted that also analytic approximations based on the variational Bayes (VB) and expectation propagation (EP) have been proposed by Titsias and Lázaro-Gredilla (2011) and Hernández-Lobato et al. (2013), respectively.

The discrete mixing nature of different variable combinations makes the spike-and-slab prior computationally rather challenging. To ease the computations, many continuous so called *shrinkage priors* for weights w_j have been proposed. These include for example Student-t (Tipping, 2001), Laplace (Park and Casella, 2008) and Horseshoe (Carvalho et al., 2009, 2010) distributions. Laplace prior is the Bayesian version of the frequentist L_1-regularizer, the Lasso (Tibshirani, 1996). All these different priors are connected by the fact that they can be expressed as scale mixtures of the Gaussian density

$$\begin{gathered} p(w_j \mid \lambda_j^2) = \mathrm{N}\left(w_j \mid 0, \lambda_j^2\tau^2\right) \\ \lambda_j^2 \sim p(\lambda_j^2) \end{gathered}$$

with mixing distributions

$\lambda_j^2 = 1$	for Gaussian
$\lambda_j^2 \sim$ Bernoulli	for spike-and-slab
$\lambda_j^2 \sim$ Inverse-Gamma	for Student-t
$\lambda_j^2 \sim$ Exponential	for Laplace
$\lambda_j \sim$ Half-Cauchy	for Horseshoe.

Note that for the Horseshoe, the mixing distribution is for for λ_j, not for λ_j^2. Ideally, the mixing distribution would have mass both for small and large values of λ_j^2

allowing $w_j = 0$ for some weights and leaving the others unshrinked. Although none of the continuous mixing distributions give a nonzero probability for a weight being exactly zero, the empirical results suggest that the Horseshoe is able to produce behaviour close to the spike-and-slab outperforming the other shrinkage priors (Carvalho et al., 2010). For a detailed comparison and nice illustration of the differences, see the papers by Carvalho et al. (2009, 2010).

2.3 Linear binary classification

Section 2.2 discussed the linear regression model where the target variable y is real valued. Other important problems are the classification problems, where y is one of the C different class labels. In other words, one is interested in predicting the outcome probabilities of each class given the associated input variables $\mathbf{x}$. For the relevance concerning this thesis, only the simplest case, the binary classification $y \in \{0, 1\}$ is considered here. The generalization of the following inference to the multiclass problems is straightforward and is discussed for example by Bishop (2006).

For binary classification problems, the standard approach is to use a real valued auxiliary latent variable $\eta = \eta(\mathbf{x})$ and then feed this to a continuous response or activation function h that maps it between 0 and 1. The probability of class $y = 1$ is then given by

$$p(y = 1 \mid \mathbf{x}) = h(\eta(\mathbf{x}))\,. \tag{36}$$

Note that many authors prefer to formulate the model using the link function g which is the inverse of the response function $g = h^{-1}$. A common choice for the response function is the logistic sigmoid

$$s(\eta) = \frac{1}{1 + \exp(-\eta)}\,. \tag{37}$$

An alternative choice is the cumulative function of normal distribution

$$\Phi(\eta) = \int_{-\infty}^{\eta} \mathrm{N}\,(t \mid 0, 1)\,dt, \tag{38}$$

which is called the probit function. The two response functions are plotted in figure 1. The logistic function has slightly longer tails but apart from that, the two curves are not markedly different.

If the response function is chosen to be the logistic sigmoid (37) and the latent variable η is modeled as a linear function of the input variables $\eta = \mathbf{w}^T\mathbf{x}$, the model is called the linear logistic regression (or just logistic regression)

$$p(y = 1 \mid \mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}\,. \tag{39}$$

Given a set of independent and identically distributed (i.i.d.) observations $\mathcal{D} = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and a prior $p(\mathbf{w})$ for the weights, the unnormalized posterior

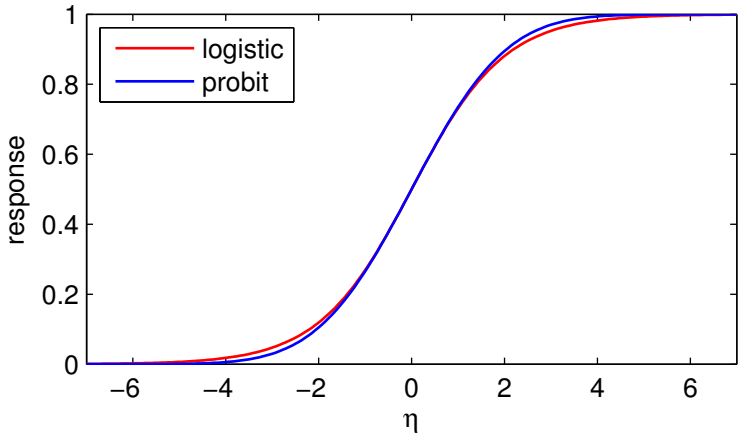


Figure 1: Two response functions: logistic $s(\eta)$ (equation (37)) and probit $\Phi(a\eta)$ (equation (38)) where $a = \sqrt{\pi/8}$. The scaling factor a is chosen so that the derivatives of the two curves coincide at the origin. As can be seen, the logistic function has slightly longer tails.

is then given by

$$\begin{aligned} p(\mathbf{w} \mid \mathcal{D}) &\propto p(\mathbf{w})p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) \\ &= p(\mathbf{w}) \prod_{i=1}^{n} p(y_i = 1 \mid \mathbf{x}_i, \mathbf{w})^{y_i} p(y_i = 0 \mid \mathbf{x}_i, \mathbf{w})^{1-y_i} \\ &= p(\mathbf{w}) \prod_{i=1}^{n} \left(\frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x}_i)} \right)^{y_i} \left(\frac{\exp(-\mathbf{w}^T \mathbf{x}_i)}{1 + \exp(-\mathbf{w}^T \mathbf{x}_i)} \right)^{1-y_i} . \end{aligned} \tag{40}$$

The simplest choice for the prior is the Gaussian $p(\mathbf{w}) = \mathrm{N}(\mathbf{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, typically of form $\boldsymbol{\mu}_0 = 0$, $\mathbf{A}_0 = \tau^{-2}\mathbf{I}$. Also priors discussed in section 2.2.2 can be used to promote sparsity. For the Gaussian and also to other priors, the posterior will be non-Gaussian and the computations are not as straightforward as with the regression model in section 2.2. The posterior must be approximated somehow, that is, by drawing samples using some MCMC method or by making an analytic approximation. There are a great number of different MCMC methods for drawing posterior samples and they are not covered in this thesis. The implementation and use of these techniques can be found in textbooks, such as the ones by Bishop (2006) and Gelman et al. (2013a).

An alternative to sampling is to make an analytic approximation to the posterior distribution, typically a Gaussian one. The simplest Gaussian approximation is obtained using the Laplace method, which is discussed next briefly. It is noted that also approximations based on variational Bayes and expectation propagation are available (e.g. Bishop, 2006; Gelman et al., 2013a) but they are not discussed here. Consider the Taylor series of the log posterior $f(\mathbf{w}) = \log p(\mathbf{w} \mid \mathcal{D})$ around $\mathbf{w} = \mathbf{w}_0$

$$f(\mathbf{w}) = f(\mathbf{w}_0) + (\mathbf{w} - \mathbf{w}_0)^T \mathbf{g}(\mathbf{w}_0) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T \mathbf{H}(\mathbf{w}_0)(\mathbf{w} - \mathbf{w}_0) + \ldots \tag{41}$$

where $\mathbf{g}(\mathbf{w}_0)$ and $\mathbf{H}(\mathbf{w}_0)$ are the gradient and Hessian of $f(\mathbf{w})$ at $\mathbf{w} = \mathbf{w}_0$

$$\mathbf{g}(\mathbf{w}_0) = \left.\frac{\partial f(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}_0} \qquad \mathbf{H}(\mathbf{w}_0) = \left.\frac{\partial^2 f(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^T}\right|_{\mathbf{w}=\mathbf{w}_0} . \tag{42}$$

If the expansion is made at the posterior mode $\mathbf{w} = \hat{\mathbf{w}}$ leaving out the third and higher order terms, the approximation becomes

$$f(\mathbf{w}) \approx f(\hat{\mathbf{w}}) + \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^T \mathbf{H}(\hat{\mathbf{w}})(\mathbf{w} - \hat{\mathbf{w}}), \tag{43}$$

because the gradient will be zero at the mode (note that the mode of the log posterior coincides with the true posterior mode). This second order approximation for the log posterior leads to a Gaussian approximation for the posterior

$$p(\mathbf{w} \mid \mathcal{D}) \propto \exp\left(\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^T \mathbf{H}(\hat{\mathbf{w}})(\mathbf{w} - \hat{\mathbf{w}})\right) \propto \mathrm{N}\left(\mathbf{w} \mid \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right), \tag{44}$$

where

$$\boldsymbol{\mu}_n = \hat{\mathbf{w}} \tag{45}$$

$$\boldsymbol{\Sigma}_n = -\mathbf{H}(\hat{\mathbf{w}})^{-1} . \tag{46}$$

In other words, the Laplace approximation replaces the true posterior with a Gaussian whose mean is set to the posterior mode and precision to the negative Hessian of the log posterior at the mode (the covariance is the inverse of the precision matrix). The theoretical justification for the Laplace approximation comes from the fact that under some regularity conditions the posterior converges to a Gaussian distribution as the number of observations $n \to \infty$. This is due to the fact that the posterior becomes increasingly peaked at the mode and the higher order terms can be ignored in the expansion (41) (e.g. Gelman et al., 2013a). Therefore Laplace approximation improves when the size of the dataset gets larger. However, for small sample sizes and skewed distributions the approximation might be poor, and the approximation should not be used recklessly.

Laplace approximation is appealing because the implementation requires only the first two derivatives of the log posterior which are straightforwardly calculated from (40). Note that one can work with the unnormalized posterior because any normalization constant would cancel out in the differentiation of the log posterior. The mode can be found, for example, by Newton's iteration (e.g. Boyd and Vandenberghe, 2004)

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \mathbf{H}(\mathbf{w}^{(k)})^{-1} \mathbf{g}(\mathbf{w}^{(k)}) \tag{47}$$

with a suitable starting point $\mathbf{w} = \mathbf{w}^{(0)}$. Note also that when the prior $p(\mathbf{w})$ is Gaussian (or any other log-concave function), the log of the unnormalized posterior (40) can be shown to be a strictly concave function of $\mathbf{w}$. Thus for a moderate dimensionality of $\mathbf{w}$, the iteration (47) converges very fast to a unique maximum, which is the main advantage of the analytic approximation over the MCMC solution.

For priors that are not log-concave such as the spike-and-slab (33), the posterior is not guaranteed to be unimodal and may indeed be multimodal. For such priors, either sampling methods or more sophisticated analytical approximations should be employed. The sampling under the spike-and-slab prior is discussed in section 4.2. For probit model, the inference with spike-and-slab prior can be carried out efficiently using expectation propagation which has been suggested to give comparable results to the MCMC solution but with significantly less computational burden (Hernández-Lobato et al., 2010).

For the Gaussian posterior approximation (44), the predictions are obtained from

$$p(\tilde{y} = 1 \mid \tilde{\mathbf{x}}, \mathcal{D}) = \int s(\mathbf{w}^T \tilde{\mathbf{x}}) p(\mathbf{w} \mid \mathcal{D}) d\mathbf{w} = \int s(\eta) p(\eta) d\eta, \tag{48}$$

where $\eta = \mathbf{w}^T \tilde{\mathbf{x}}$ and $p(\eta)$ is its distribution. Now, if the posterior $p(\mathbf{w} \mid \mathcal{D})$ is (approximated by) Gaussian $\mathrm{N}(\mathbf{w} \mid \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, then also $\eta = \mathbf{w}^T \tilde{\mathbf{x}}$ has a Gaussian distribution because it is a linear combination of jointly Gaussian variables. The mean and variance of $p(\eta)$ can be shown to be (Bishop, 2006)

$$\mu_\eta = \boldsymbol{\mu}_n^T \tilde{\mathbf{x}} \tag{49}$$

$$\sigma_\eta^2 = \tilde{\mathbf{x}}^T \boldsymbol{\Sigma}_n \tilde{\mathbf{x}} \,. \tag{50}$$

Thus the prediction has reduced to evaluating the integral

$$p(\tilde{y} = 1 \mid \tilde{\mathbf{x}}, \mathcal{D}) = \int s(\eta) \mathrm{N}\left(\eta \mid \mu_\eta, \sigma_\eta^2\right) d\eta \,. \tag{51}$$

This is not analytically available, but as a one dimensional integral it poses no problem as it can be easily calculated using standard numerical integration techniques.

3 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 some criterion or criteria. As already discussed in the introduction, in the context of statistical modeling, the best available model is usually the one giving the most precise predictions about the future observations. Model selection that is based on assessing the predictive performance of the candidate models is called *predictive model selection* (Vehtari and Ojanen, 2012), and this thesis considers predictive model selection only. How the predictive performance of a model is defined is discussed in section 3.1.

This section somewhat follows the ideas by Bernardo and Smith (1994) and Vehtari and Ojanen (2012) in how to categorize the model selection methods proposed in the literature. Bernardo and Smith (1994) presented the idea of $\mathcal{M}$-closed, $\mathcal{M}$-completed and $\mathcal{M}$-open views. $\mathcal{M}$-closed means assuming that one of the candidate models is the true data generating model without explicit knowledge which one it is. Under this assumption, one can set prior probabilities for each model and form the Bayesian model average (4). In practical modeling situations adopting the $\mathcal{M}$-closed perspective is questionable in the literal sense. Other than in controlled situations (such as in computer simulations) it is hard to justify the belief that one of the models is the true model. A relaxed version is the $\mathcal{M}$-completed view where one abandons the idea of the true model, but still forms a *reference model* which is believed to be the best available description of the future observations. In $\mathcal{M}$-open view one does not assume one of the models being true and also rejects the idea of constructing the reference model.

From a practical point of view, the ideas of $\mathcal{M}$-closed, $\mathcal{M}$-completed and $\mathcal{M}$-open views should not be taken too strictly. In practice, the distinction between $\mathcal{M}$-closed and $\mathcal{M}$-completed is somewhat hazy. In some cases, the construction of the Bayesian model average may be reasonable even though it is theoretically unjustified because it explicitly assumes that one of the models is the true model. One such a situation is the sparse variable selection problem, where one has a set of predefined variables with prior knowledge that many of the variables are irrelevant, but it is just not known which. As discussed in section 2.2.2, in such a situation the Bayesian model averaging over the variable combinations is often considered the "golden standard" solution.

Throughout section 3, the notation assumes a model that predicts a single output variable y given the input variables $\mathbf{x}$. Both of these can be either continuous or discrete. However, most of the time y is treated as continuous and the expectations are written using integrals. If not explicitly stated, the treatment for discrete y is analogous and readily obtained by replacing the integrals with summations. To make the notation simpler, the training data is denoted by $\mathcal{D} = (\mathbf{X}, \mathbf{y})$.

The section is organized as follows. Section 3.1 discusses how the predictive ability of a model is defined in terms of an expected utility. Sections 3.2–3.4 review methods for Bayesian model selection proposed in the literature. The methods are categorized according to their type of approach and this division is sketched in table

Section 3.2	Utility estimation with data reuse ($\mathcal{M}$-open view)	Cross validation Information criteria Other predictive criteria
Section 3.3	Reference methods ($\mathcal{M}$-completed view)	Reference predictive approach Projection
Section 3.4	Estimation of model probabilities	Maximum a posteriori model Median model

Table 1: Categorization of the different model selection methods.

1. Section 3.5 discusses the concept of selection induced bias and motivates why it is relevant in the variable selection context.

3.1 Predictive ability as an expected utility

The predictive performance of a model is typically defined in terms of a utility function that describes the quality of the predictions. This means, that one defines a utility function u that maps each prediction $a_k \in \mathcal{A}_k$ of the candidate model M_k to a scalar utility 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 $u : \mathcal{A}_k \times \tilde{\mathcal{Y}} \mapsto \mathbb{R}$. Note, that many authors prefer to use loss functions instead of utilities. However, the two approaches are equivalent, because any utility function can be made a loss function by only changing its sign. Thus, it is merely a matter of taste or convenience which one is preferred.

The prediction $a_k = a(M_k)$ may be either a scalar or a distribution depending on the situation. Sometimes one might be interested in obtaining the best possible point prediction for future observation $\tilde{y}$, and in such a situation a common loss function is the squared error

$$e(M_k, \tilde{y}) = (a_k - \tilde{y})^2, \tag{52}$$

which clearly obtains its minimum at $a_k = \tilde{y}$. The squared error is an often used loss function especially in frequentist literature where the predictions are commonly described via point estimates. Often the motivation for using the squared error loss function is its mathematical convenience, because in addition to its simplicity, it is also everywhere differentiable and strictly convex with respect to a_k. However, as discussed in section 2, in Bayesian modelling the predictions are described in terms of the posterior predictive distributions $a(M_k) = p(\tilde{y} \mid \mathbf{y}, M_k)$ instead of point predictions. Here and in the subsequent notation the input variables $\mathbf{x}$ are left out for simplicity. An often used utility function measuring the quality of a predictive distribution is the logarithmic score (Good, 1952)

$$u(M_k, \tilde{y}) = \log p(\tilde{y} \mid \mathbf{y}, M_k)\,. \tag{53}$$

The logarithmic score will be used throughout this thesis and its connection to the information theory will be discussed shortly. However, at this point it is stressed that in principle any other utility functions could be considered as well, and the choice of a suitable utility function might also be application specific.

Since the future observations $\tilde{y}$ are unknown, the utility function $u(M_k, \tilde{y})$ can not be evaluated beforehand. In other words, one does not know how well the model will predict yet unseen observations. Because of this, one usually works with expected utilities instead

$$\bar{u}_t(M_k) = \mathrm{E}\left[u(M_k, \tilde{y})\right] = \int p_t(\tilde{y}) u(M_k, \tilde{y}) d\tilde{y}, \tag{54}$$

where $p_t(\tilde{y})$ is the true distribution of the future observation $\tilde{y}$. This expression will be referred to as the *generalization utility* or more loosely as the generalization performance of model M_k. With the logarithmic score, the generalization utility becomes

$$\bar{u}_t(M_k) = \int p_t(\tilde{y}) \log p(\tilde{y} \mid \mathbf{y}, M_k) d\tilde{y}\,. \tag{55}$$

Also the true data generating distribution $p_t(\tilde{y})$ is in practice unknown and exact evaluation of (55) is impossible. This formula is still of central importance, because it can be estimated in various ways, and some of the proposed methods for Bayesian model selection are based on maximizing an estimate of it. The definition of the generalization utility with the logarithmic score is related to the information theory via the definition of Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951)

$$\mathrm{KL}\left(p \,||\, q\right) = \int p(\tilde{y}) \log\left(\frac{p(\tilde{y})}{q(\tilde{y})}\right) d\tilde{y}\,. \tag{56}$$

KL-divergence measures the similarity of distributions p and q and can be considered as the information loss of using distribution q when the true distribution is p. In this context the information means the Shannon information (Shannon, 1948). KL-divergence satisfies two important properties, that is, for all distributions p and q $\mathrm{KL}\left(p \,||\, q\right) \geq 0$ and $\mathrm{KL}\left(p \,||\, q\right) = 0$ only if $p = q$. However, KL-divergence is not a valid distance metric because it is directed, meaning that usually $\mathrm{KL}\left(p \,||\, q\right) \neq \mathrm{KL}\left(q \,||\, p\right)$. Applying the definition (56) one can decompose the generalization utility (55) as

$$\begin{aligned}
\bar{u}_t(M_k) &= \int p_t(\tilde{y}) \log p(\tilde{y} \mid \mathbf{y}, M_k) d\tilde{y} \\
&= \int p_t(\tilde{y}) \left(\log p_t(\tilde{y}) + \log\left(\frac{p(\tilde{y} \mid \mathbf{y}, M_k)}{p_t(\tilde{y})}\right)\right) d\tilde{y} \\
&= \int p_t(\tilde{y}) \log p_t(\tilde{y}) d\tilde{y} - \int p_t(\tilde{y}) \log\left(\frac{p_t(\tilde{y})}{p(\tilde{y} \mid \mathbf{y}, M_k)}\right) d\tilde{y} \\
&= \bar{u}_t(M_t) - \mathrm{KL}\left(p_t(\tilde{y}) \,||\, p(\tilde{y} \mid \mathbf{y}, M_k)\right),
\end{aligned} \tag{57}$$

where the constant $\bar{u}_t(M_t)$ is the expected utility of the true data generating model M_t. Note that since the KL-divergence is always nonnegative, the maximization of the generalization utility (55) is equivalent to minimizing the divergence between the predictive distributions of the true data generating model M_t and the candidate model M_k.

3.2 Utility estimation with sample reuse

This section reviews approaches for estimating the generalization utility (55) that are based on sample reuse. These methods can be considered to reflect the $\mathcal{M}$-open view in the sense that none of the models is believed to be the true model but one is interested in finding the best model available. The section begins with a short discussion of estimating the model fit at training observations, and then proceeds to methods that are designed to correct the bias in this approach, namely cross validation (section 3.2.2) and information criteria (section 3.2.3). Also a few other related predictive approaches are considered (section 3.2.4).

3.2.1 Posterior and hold-out predictive approach

A natural idea for estimating the generalization performance (55) would be to use the already obtained sample $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ and approximate the integral with the Monte Carlo method as

$$\bar{u}_{\text{tr}} = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i, \mathcal{D}, M_k), \tag{58}$$

which is called the training utility of model M_k. This, however, overestimates the generalization utility, because the model is tested on the same data that was used for training it. The model may fit to the training observations very well, but has no predictive ability outside the training data. A better idea is to split the training data into two $\mathcal{D} = \{\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{ts}}\}$, and use the first part for training and the second part for testing. This gives the hold-out utility

$$\bar{u}_{\text{hold-out}} = \frac{1}{n_{\text{ts}}} \sum_{i \in I_{\text{ts}}} \log p(y_i \mid \mathbf{x}_i, \mathcal{D}_{\text{tr}}, M_k)\,. \tag{59}$$

where I_{ts} denotes the indices of points in $\mathcal{D}$ that belong to the test set $\mathcal{D}_{\text{ts}}$, and n_{ts} is the size of $\mathcal{D}_{\text{ts}}$. This approach is safer in the sense that it does not overestimate the generalization performance but instead it typically underestimates it. This is because the model is trained with only part of the data, and the predictions would be better if all the data was used for training. Of course one could do the division so that almost all of the points are used for training, but the small test set size leads to inaccuracy and high variance in the utility estimate. Let us next discuss the idea of cross validation which tackles this very issue.

3.2.2 Cross validation

Cross validation (CV) is a widely used method for frequentist and Bayesian model selection and assessment (from frequentist perspective see Stone (1974) and from Bayesian perspective see Geisser and Eddy (1979)). It extends the idea of hold-out (section 3.2.1) by dividing the original dataset of n points into k subsets $I_1, \ldots, I_k$ each of which is then used in turn for validation while the points in the remaining $k-1$ sets form the training set. This way the same part of the data is never used simultaneously for model training and validation, but still a utility or an error estimate is obtained for the whole dataset. This is referred to as k-fold cross validation

$$k\text{-fold-CV} = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i, \mathcal{D}_{\setminus I(i)}, M_k), \tag{60}$$

where $I(i)$ denotes the validation set that contains the ith point and $D_{\setminus I(i)}$ the training data from which this subset has been removed. A natural idea would be to set $k = n$, that is, the model is trained each time using all the points except the one which is used for validation. This is referred to as leave-one-out cross validation (LOO-CV). For LOO-CV the utility estimate is given by

$$\text{LOO-CV} = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i, \mathcal{D}_{\setminus i}, M_k), \tag{61}$$

where $\mathcal{D}_{\setminus i}$ denotes the original training data excluding ith point. Watanabe (2010) showed that LOO-CV with logarithmic utility function is asymptotically equal to the expected true utility (55) and the error is $o(1/n)$. Thus when n is not very small, LOO-CV is a nearly unbiased estimate of the true generalization utility. The obvious drawback in LOO-CV is that it requires training the model n times which is computationally infeasible in many practical modeling situations. In some cases, however, there might be computational shortcuts, which allow the computation of LOO-CV with only small additional computations to a single model training. Examples of such special cases are the linear and Gaussian process regression with fixed hyperparameters (e.g. Orr, 1996; Sundararajan and Keerthi, 2001).

Gelfand et al. (1992) proposed a computationally convenient way of estimating the LOO-CV, called importance sampling (IS) LOO-CV. IS-LOO-CV is based on the general importance sampling formula, where the expectation of an arbitrary function $h(\boldsymbol{\theta})$ over distribution $f(\boldsymbol{\theta})$ is approximated as

$$\int h(\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int \frac{h(\boldsymbol{\theta}) f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} g(\boldsymbol{\theta}) d\boldsymbol{\theta} \approx \frac{\sum_s w_s h(\boldsymbol{\theta}_s)}{\sum_s w_s}, \tag{62}$$

where $\{\boldsymbol{\theta}_s\}_{s=1}^{S}$ is a sample from distribution $g(\boldsymbol{\theta})$ and

$$w_s = \frac{f(\boldsymbol{\theta}_s)}{g(\boldsymbol{\theta}_s)}\,.$$

Conditions under which the approximation (62) converges to the true value are discussed by Geweke (1989). The leave-one-out predictions are approximated by a straightforward application of the above formula by using the full model posterior $p(\boldsymbol{\theta} \mid \mathcal{D}, M_k)$ as the importance sampling distribution $g(\boldsymbol{\theta})$. One gets

$$\begin{aligned} p(y_i \mid \mathbf{x}_i, \mathcal{D}_{\setminus i}, M_k) &= \int p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}, M_k) p(\boldsymbol{\theta} \mid \mathcal{D}_{\setminus i}, M_k) d\boldsymbol{\theta} \\ &\approx \frac{\sum_s w_s p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_s, M_k)}{\sum_s w_s}, \end{aligned} \tag{63}$$

where $\{\boldsymbol{\theta}_s\}_{s=1}^S$ is a sample from $p(\boldsymbol{\theta} \mid \mathcal{D}, M_k)$ and

$$w_s = \frac{p(\boldsymbol{\theta}_s \mid \mathcal{D}_{\setminus i}, M_k)}{p(\boldsymbol{\theta}_s \mid \mathcal{D}, M_k)} \propto \frac{1}{p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_s, M_k)} \,. \tag{64}$$

Plugging this into (63) gives the IS-LOO-CV formula

$$\text{IS-LOO-CV} = \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{1}{\frac{1}{S} \sum_s \frac{1}{p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_s, M_k)}} \right) . \tag{65}$$

The reliability of IS-LOO-CV estimate can be studied by the variability in the weights w_s, for example by plotting the cumulative normalized weights. If for some point i most of the mass of $\sum_s w_s$ is concentrated on a few of the weights only, this may indicate that the approximation is unreliable. The reliability estimation is discussed in more detail by Vehtari and Lampinen (2002).

Another way of overcoming the computational cost of LOO-CV is simply setting $k \ll n$, that is, using fewer validation sets. The computational ease comes with the disadvantage that conditioning the predictions for less than n observations introduces an extra bias in the utility estimate. However, this bias can be estimated (Burman, 1989). A conventional choice is to set $k = 10$, that is, each time 90% of the data is used for training and 10% for validation.

Usually LOO-CV is considered as a robust method for model assessment, but in model selection it suffers from the selection induced bias (section 3.5) and this is demonstrated in section 5.

3.2.3 Information criteria

Information criteria (IC) are a computationally convenient alternative for estimating the model fit. Several methods falling into this category have been proposed and all of them are not covered here. Instead this section discusses the general idea and goes through a few of them which are best known.

AIC Information criteria are designed to estimate the fit of the candidate model, and can often be written in the form

$$\text{IC} = \text{training utility} + \text{bias correction} \,.$$

The very first information criterion by Akaike (1974) (AIC) estimates the expected logarithmic score of the model at the training observations with maximum likelihood estimate $\boldsymbol{\theta}_{\mathrm{ML}}$ for the parameters

$$\mathrm{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\int p_t(\tilde{y} \mid \mathbf{x}_i)\log p(\tilde{y} \mid \mathbf{x}_i, \boldsymbol{\theta}_{\mathrm{ML}}, M_k)d\tilde{y}\right]. \tag{66}$$

The expectation is taken over all the possible training sets $\mathcal{D}$, and $p_t(\tilde{y} \mid \mathbf{x}_i)$ is the true data generating distribution conditional on the input variables. In the derivation of Akaike's criterion it is assumed that $p_t(\tilde{y} \mid \tilde{\mathbf{x}})$ is well approximated by a parametric model $p(\tilde{y} \mid \tilde{\mathbf{x}}, \boldsymbol{\theta}_t, M_k)$, and after using the Taylor expansion for $\boldsymbol{\theta}_{\mathrm{ML}}$ one gets the estimate

$$\mathrm{AIC} = \frac{1}{n}\sum_{i=1}^{n}\log p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}_{\mathrm{ML}}, M_k) - \frac{p}{n}, \tag{67}$$

where p is the number of parameters in the model. The first part is the logarithmic training utility of the model with the maximum likelihood estimate for the parameters, and the second part is the bias correction due to estimating the fit at the training observations. AIC does not fit very well to the Bayesian paradigm since it is based on estimating the goodness of the maximum likelihood parameters. In Bayesian setting, one is generally interested in the goodness of the predictions obtained by integrating over the parameters.

DIC Another well-known information criterion is the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002). DIC estimates the generalization performance of the model at the training inputs with parameters fixed to the posterior mean $\bar{\boldsymbol{\theta}} = \mathrm{E}\left[\boldsymbol{\theta} \mid \mathcal{D}\right]$. DIC can be written as

$$\mathrm{DIC} = \frac{1}{n}\sum_{i=1}^{n}\log p(y_i \mid \mathbf{x}_i, \bar{\boldsymbol{\theta}}, M_k) - \frac{p_{\mathrm{eff}}}{n}, \tag{68}$$

where p_{eff} is the effective number of parameters in the model. p_{eff} can be estimated in two different ways the first one being

$$p_{\mathrm{eff}} = 2\sum_{i=1}^{n}\left(\log p(y_i \mid \mathbf{x}_i, \bar{\boldsymbol{\theta}}, M_k) - \mathrm{E}\left[\log p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}, M_k)\right]\right), \tag{69}$$

where the expectation in the latter term is taken over the posterior $p(\boldsymbol{\theta} \mid \mathcal{D}, M_k)$. The alternative formula is

$$p_{\mathrm{eff}} = 2\mathrm{Var}\left[\log p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}, M_k)\right], \tag{70}$$

where the variance is also calculated over the posterior $p(\boldsymbol{\theta} \mid \mathcal{D}, M_k)$. The first one (69) is argued to be numerically more stable (Gelman et al., 2013b), but due to the use of a point estimate for $\boldsymbol{\theta}$ the estimation becomes in general variant to the

parametrization. The latter formula (70) on the other hand avoids the use of the point estimate being therefore invariant to the parametrization.

Investigating the formulas (67) and (68) one observes that DIC bears a resemblance to AIC in the sense that it replaces the maximum likelihood estimate $\boldsymbol{\theta}_{\mathrm{ML}}$ with the posterior mean $\bar{\boldsymbol{\theta}}$ and the number of parameters p by a data based bias correction. Based on this connection Gelman et al. (2013b) stated that "DIC is a somewhat Bayesian version of AIC". Even though the ML estimate of AIC is replaced by posterior mean in DIC, this does not make DIC a fully Bayesian approach, because it is still based on a point estimate.

WAIC In contrast to AIC and DIC, a fully Bayesian criterion is the widely applicable information criterion (WAIC) by Watanabe (2009, 2010). WAIC can be calculated as

$$\mathrm{WAIC} = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid \mathbf{x}_i, \mathcal{D}, M_k) - \frac{V}{n}, \tag{71}$$

where the first term is the training utility (58) and V is the functional variance given by

$$V = \sum_{i=1}^{n} \left\{ \mathrm{E}\left[(\log p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}, M_k))^2 \right] - \mathrm{E}\left[\log p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}, M_k) \right]^2 \right\} \tag{72}$$

Here both of the expectations are taken over the posterior $p(\boldsymbol{\theta} \mid \mathcal{D}, M_k)$. Watanabe (2010) proved that WAIC is asymptotically equal to the leave-one-out cross validation (61) and to the generalization utility (55) and the error is $o(1/n)$. In other words, both LOO-CV and WAIC converge to the generalization utility as $n \to \infty$, and the error is bounded by $o(1/n)$. Watanabe's proof gives WAIC a solid theoretical justification in the sense that it measures the predictive ability of the model including the uncertainty in the parameters, not just goodness of a point estimate as DIC does. The use of a point estimate is questionable not only because of Bayesian principles, but also from a practical point of view especially when the model is singular. For singular models the set of the "true parameters" consists of more than one point, and therefore the point estimate does not represent the posterior completely even when $n \to \infty$. WAIC can also be used for singular models which is the reason for the name "widely applicable".

3.2.4 Other posterior and CV-predictive approaches

***L*-criterion** Laud and Ibrahim (1995) proposed a model selection criterion which is derived by considering replicated measurements $\tilde{\mathbf{y}}$ at training inputs $\mathbf{X}$. The criterion measures the expected squared error between the new observations and the old ones $\mathbf{y}$ over the posterior predictive distribution of the candidate model M_k

$$L^2(M_k) = \mathrm{E}\left[(\mathbf{y} - \tilde{\mathbf{y}})^T (\mathbf{y} - \tilde{\mathbf{y}}) \mid \mathbf{X}, \mathcal{D}, M_k \right] . \tag{73}$$

The error can be decomposed as

$$L^2(M_k) = \sum_{i=1}^{n}(y_i - \mathrm{E}\left[\tilde{y} \mid \mathbf{x}_i, \mathcal{D}, M_k\right])^2 + \sum_{i=1}^{n}\mathrm{Var}\left[\tilde{y} \mid \mathbf{x}_i, \mathcal{D}, M_k\right], \tag{74}$$

that is, sum of the squared errors for mean predictions plus sum of the predictive variances. For convenience, the authors defined the final criterion as the square root of L^2, so that the criterion has the same units as the target variable. This yields the L-criterion

$$L(M_k) = \sqrt{L^2(M_k)}. \tag{75}$$

The preferred model is then the one which minimizes (75). As is seen from (74), L-criterion (or L^2) is not the same as the squared error at the training inputs. It also measures the predictive variance, and therefore favors models having a narrow posterior predictive distribution and a good mean predictive fit to training data. As complex models tend to have larger predictive variance, L-criterion penalizes complex models more than just the squared training error. However, L-criterion is still somewhat problematic because the same data is used both for the model training and testing.

L_q-criterion Marriott et al. (2001) proposed a criterion which is a cross validated version of the L^2-criterion (74). The authors named their measure as L_q-criterion

$$L_q(M_k) = \sum_{i=1}^{n}(y_i - \mathrm{E}\left[\tilde{y} \mid \mathbf{x}_i, \mathcal{D}_{\setminus i}, M_k\right])^2 + \sum_{i=1}^{n}\mathrm{Var}\left[\tilde{y} \mid \mathbf{x}_i, \mathcal{D}_{\setminus i}, M_k\right]. \tag{76}$$

L_q-criterion sounds intuitively better than L-criterion, because it does not use the same data for training and testing. However, as is demonstrated in section 5, also L_q-criterion may perfom badly in variable selection due to the selection bias.

L_{GG}-criterion Gelfand and Ghosh (1998) presented an approach that is based on defining an optimal point prediction for a replicated measurement $\tilde{\mathbf{y}}$. The optimal point prediction is defined to be close both to the observed $\mathbf{y}$ and to the future $\tilde{\mathbf{y}}$. Given a loss function $l(a, \tilde{y})$, the optimal point prediction at inputs $\mathbf{x}_i$ is defined as

$$a_i^{\mathrm{opt}} = \arg\min \mathrm{E}\left[l(a_i, \tilde{y}_i) + kl(a_i, y_i) \mid \mathbf{x}_i, \mathcal{D}, M_k\right], \tag{77}$$

where the expectation is taken over the posterior predictive distribution of model M_k, and k is a free parameter adjusting the relative importance of the future and observed data. The criterion is then defined as

$$L_{GG}(M_k) = \sum_{i=1}^{n}\mathrm{E}\left[l(a_i^{\mathrm{opt}}, \tilde{y}_i) + kl(a_i^{\mathrm{opt}}, y_i) \mid \mathbf{x}_i, \mathcal{D}, M_k\right]. \tag{78}$$

The framework is quite general allowing in principle any loss function $l(a, \tilde{y})$ to be used. However, for arbitrary loss functions the optimal point prediction (77) may

not be easily calculated. Gelfand and Ghosh (1998) discuss the squared error loss $l = (a-\tilde{y})^2$ for which a_i^{opt} can be derived analytically. For squared error, the optimal point prediction becomes

$$a_i^{\text{opt}} = \frac{\mathrm{E}\left[\tilde{y} \mid \mathbf{x}_i, \mathcal{D}, M_k\right] + k y_i}{k+1}. \tag{79}$$

This result is obtained simply by plugging the loss function into (77), differentiating with respect to a_i and setting the derivative to zero. The criterion then becomes

$$L_{GG}(M_k) = \frac{k}{k+1} \sum_{i=1}^{n} (y_i - \mathrm{E}\left[\tilde{y} \mid \mathbf{x}_i, \mathcal{D}, M_k\right])^2 + \sum_{i=1}^{n} \mathrm{Var}\left[\tilde{y} \mid \mathbf{x}_i, \mathcal{D}, M_k\right] . \tag{80}$$

When $k \to \infty$, L_{GG} is the same as L^2-criterion (74). On the other hand, when $k = 0$ the criterion reduces to the sum of the predictive variances, and the model with the narrowest predictive distribution is chosen. In their experiment, the authors reported that the results were not very sensitive to the choice of k.

3.3 Utility estimation with the reference model

Section 3.2 reviewed methods for utility estimation that are based on sample reuse without any assumptions on the true model ($\mathcal{M}$-open view). An alternative way is to construct a reference model, which is believed to best describe our knowledge about the future observations, and perform the utility estimation as if it was the true data generating distribution ($\mathcal{M}$-closed/$\mathcal{M}$-completed view). In this thesis, all the methods based on evaluating the predictive performance of the submodels with respect to the reference model are referred to as reference methods. There are basically two somewhat different but related approaches that fit into this category, namely the reference predictive approach and the projection, which will be discussed separately.

3.3.1 Reference predictive approach

In the reference predictive approach, one forms the reference model M_* and uses it almost as if it was the true data generating model M_t. Thus, the utilities of the candidate models M_k can be estimated by replacing the true distribution $p_t(\tilde{y})$ in (55) by the reference distribution $p(\tilde{y} \mid \mathcal{D}, M_*)$

$$\int p(\tilde{y} \mid \mathcal{D}, M_*) \log p(\tilde{y} \mid \mathcal{D}, M_k) d\tilde{y}. \tag{81}$$

By averaging this over the training inputs $\{\mathbf{x}_i\}_{i=1}^n$ one gets the reference utility

$$\bar{u}_{\text{ref}}(M_k) = \frac{1}{n} \sum_{i=1}^{n} \int p(\tilde{y} \mid \mathbf{x}_i, \mathcal{D}, M_*) \log p(\tilde{y} \mid \mathbf{x}_i, \mathcal{D}, M_k) d\tilde{y} \,. \tag{82}$$

As the reference model is in practice different from the true data generating model, the reference utility is a biased estimate of the true performance of the candidate

model. By equation (57), the maximization of the reference utility is equivalent to minimizing the predictive KL-divergence between the reference model M_* and the candicate model M_k at the training inputs

$$\delta(M_* \mid\mid M_k) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}\left(p(\tilde{y} \mid \mathbf{x}_i, \mathcal{D}, M_*) \mid\mid p(\tilde{y} \mid \mathbf{x}_i, \mathcal{D}, M_k)\right). \tag{83}$$

In other words, in this approach one seeks a candidate model whose predictive distribution is as close as possible to the predictive distribution of the reference model. The model choice can then be based on the strict minimization of the discrepancy measure (83), or choosing the simplest model that has an acceptable discrepancy. What is meant by "acceptable" may be somewhat arbitrary and context dependent (see the concept of relative explanatory power in the next section, eq. (88)). In section 5 it is also demonstrated that due to the selection induced bias the estimated discrepancy may be a poor estimate of the difference in the predictive performance between the reference model and the selected submodels.

Formula (83) measures the predictive discrepancy at the training inputs. If the future predictions will be conditioned on different inputs, one would be interested in measuring the out-of-sample discrepancy. In this case the discrepancy from the reference model could be calculated using cross validation densities as

$$\delta(M_* \mid\mid M_k) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}\left(p(\tilde{y} \mid \mathbf{x}_i, \mathcal{D}_{\setminus i}, M_*) \mid\mid p(\tilde{y} \mid \mathbf{x}_i, \mathcal{D}_{\setminus i}, M_k)\right). \tag{84}$$

The above notation is for leave-one-out densities, but one could also use for example 10-fold cross validation to reduce the computations.

The reference predictive approach is inherently a less straightforward approach to model selection than the methods presented in section 3.2 because it requires the construction of the reference model. The quality of the model selected by minimizing the reference discrepancy (either (83) or (84)) is naturally dependent on the quality of the reference model M_*. This emphasizes the importance of the model criticism when constructing the reference model. There is no automated way of coming up with a good reference model (otherwise all the modelling problems would be solved). San Martini and Spezzaferri (1984) proposed using the Bayesian model average (4) as the reference model. This has been criticized because, as discussed in the introduction of section 3, this means assuming that one of the candidate models is the true data generating model, which is debatable. In principle, however, averaging over the discrete model space (i.e. forming the model average) does not differ in any sense from integrating over the continuous parameters which is the standard procedure in Bayesian modeling. Moreover, especially in variable selection problems the integration over the different variable combinations, that is, the use of the spike-and-slab prior (section 2.2.2) is widely accepted. Thus, even if one did not believe in one of the candidate models being true, forming the model average might still be reasonable.

It is worth mentioning, however, that even though BMA is a natural choice for the reference model, one can in principle let M_* to be anything as long it is believed that M_* is the model presenting one's beliefs the best possible way.

3.3.2 Projection approach

A related method to the reference predictive approach (section 3.3.1) is the parametric projection proposed by Goutis and Robert (1998) and further discussed by Dupuis and Robert (2003). The idea is to project the parameters of the reference model to the parameter space of a candidate model such that the associated parametric KL-divergence is minimized. More specifically, given the parameter of the reference model $\boldsymbol{\theta}^*$, the projected parameter $\boldsymbol{\theta}^\perp$ in the parameter space of model M_k is defined via

$$\boldsymbol{\theta}^\perp = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}\left(p(\tilde{y} \mid \mathbf{x}_i, \boldsymbol{\theta}^*, M_*) \mid\mid p(\tilde{y} \mid \mathbf{x}_i, \boldsymbol{\theta}, M_k)\right) . \tag{85}$$

The discrepancy between the reference model M_* and the candidate model M_k is then defined to be the expectation of the divergence over the posterior of the reference model

$$\delta(M_* \mid\mid M_k) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{\theta}^* \mid \mathcal{D}, M_*} \left[\mathrm{KL}\left(p(\tilde{y} \mid \mathbf{x}_i, \boldsymbol{\theta}^*, M_*) \mid\mid p(\tilde{y} \mid \mathbf{x}_i, \boldsymbol{\theta}^\perp, M_k)\right)\right] . \tag{86}$$

The posterior expectation in (86) is in general not available analytically. Dupuis and Robert (2003) proposed calculating the discrepancy by drawing samples $\{\boldsymbol{\theta}_s^*\}_{s=1}^S$ from the posterior of the reference model, calculating the projected parameters $\{\boldsymbol{\theta}_s^\perp\}_{s=1}^S$ individually according to (85), and then approximating (86) as

$$\delta(M_* \mid\mid M_k) \approx \frac{1}{nS} \sum_{i=1}^{n} \sum_{s=1}^{S} \mathrm{KL}\left(p(\tilde{y} \mid \mathbf{x}_i, \boldsymbol{\theta}_s^*, M_*) \mid\mid p(\tilde{y} \mid \mathbf{x}_i, \boldsymbol{\theta}_s^\perp, M_k)\right) . \tag{87}$$

Moreover, Dupuis and Robert (2003) introduced a measure called relative explanatory power

$$\phi(M_k) = 1 - \frac{\delta(M_* \mid\mid M_k)}{\delta(M_* \mid\mid M_0)} , \tag{88}$$

where M_0 denotes the empty model, that is, the model that has the largest discrepancy to the reference model. In terms of variable selection, M_0 is the variable free model. By definition, the relative explanatory power obtains values between 0 and 1, and Dupuis and Robert (2003) proposed choosing the simplest model with enough explanatory power, for example 90%. The effect of this threshold is discussed in section 5.

Even though the idea of the parametric projection is quite general, the method is designed for variable selection for generalized linear models. In general, the projected parameters (85) are not easily calculated (if at all). However, when the observation model is in the exponential family

$$p(y \mid \mathbf{x}, \boldsymbol{\theta}^*, M) = h(y) \exp\left(\boldsymbol{\eta}(\mathbf{x}, \boldsymbol{\theta}^*)^T \mathbf{u}(y) - a(\mathbf{x}, \boldsymbol{\theta}^*)\right) , \tag{89}$$

the KL-divergence is conveniently minimized. Using the shorthand notation $\boldsymbol{\eta}_i^* = \boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\theta}^*)$, $\boldsymbol{\eta}_i^\perp = \boldsymbol{\eta}(\mathbf{x}_i, \boldsymbol{\theta}^\perp)$, $a_i^* = a(\mathbf{x}_i, \boldsymbol{\theta}^*)$, $a_i^\perp = a(\mathbf{x}_i, \boldsymbol{\theta}^\perp)$ one can write the KL-divergence in (85) as

$$\mathrm{KL}\left(p(\tilde{y} \mid \mathbf{x}_i, \boldsymbol{\theta}^*, M_*) \mid\mid p(\tilde{y} \mid \mathbf{x}_i, \boldsymbol{\theta}^\perp, M_k)\right) = (\boldsymbol{\eta}_i^* - \boldsymbol{\eta}_i^\perp)^T \nabla_{\boldsymbol{\theta}^*} a_i^* + a_i^\perp - a_i^* \,. \tag{90}$$

This follows from the fact that the expectation of the sufficient statistics $\mathbf{u}(y)$ over the distribution (89) is $\nabla_{\boldsymbol{\theta}^*} a^*$. By plugging (90) into (85) and setting the gradient with respect to $\boldsymbol{\theta}^\perp$ to zero, one obtains a system of equations

$$\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}^\perp} \left((\boldsymbol{\eta}_i^\perp)^T \nabla_{\boldsymbol{\theta}^*} a_i^*\right) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}^\perp} a_i^\perp \,. \tag{91}$$

These are essentially the same equations as obtained when finding the maximum likelihood solution for the parameters. This can be seen by considering the gradient of the log-likelihood for the candidate model

$$\begin{aligned} \nabla_{\boldsymbol{\theta}^\perp} \log \mathcal{L}(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}^\perp} \left(\sum_{i=1}^{n} \left(\log h(y_i) + (\boldsymbol{\eta}_i^\perp)^T \mathbf{u}(y_i) - a_i^\perp \right) \right) \\ &= \sum_{i=1}^{n} \left(\nabla_{\boldsymbol{\theta}^\perp} \left((\boldsymbol{\eta}_i^\perp)^T \mathbf{u}(y_i) \right) - \nabla_{\boldsymbol{\theta}^\perp} a_i^\perp \right) \,. \end{aligned} \tag{92}$$

By setting this to zero one obtains the same equations as (91) but $\mathbf{u}(y_i)$ in place of $\nabla_{\boldsymbol{\theta}^*} a_i^*$. Thus projecting the parameters of the reference model to the parameter space of the candidate model can be seen as "fitting to fit". If both $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^\perp$ belong to the same subspace, then naturally the solution to the projection equations is $\boldsymbol{\theta}^\perp = \boldsymbol{\theta}^*$. The idea is typically to restrict the projection space in some way. Dupuis and Robert (2003) proposed to project the weights $\mathbf{w}$ of the full generalized linear model by restricting $w_k = 0$ for some components k. If some of the variables can be set to have a zero weight without a significant increase in the discrepancy (86), then these variables can be left out of the model.

For the normal linear model (section 2.2) the system (91) is linear and has an analytical solution. For other models the equations are nonlinear and the solution is not analytically available, but the equations can be efficiently solved using standard numerical techniques such as Newton's method. The projection has the advantage that it allows the computation of the discrepancy between the reference model and the submodels without having to place priors for each model separately; the prior information of the reference model is contained in the projected samples. The projection is also typically much faster to calculate than having to sample the parameters of each model separately. However, it is noted that the projected parameters are determined only when the maximum likelihood equations have a unique solution. For instance, for the normal linear model the projection is undetermined when there are more variables in the candidate model than observations. Moreover, it could be argued that the projection approach does not correspond to "actual" model selection because strictly speaking it does not compare the candidate models trained by the data; it compares the candidate models trained by the fit of the reference model.

3.4 Estimation of model probabilities

MAP / Marginal likelihood As discussed in section 2.1, Bayesian formalism has a natural way of describing the uncertainty with respect to the used model specification. Omitting the input variables in the notation, the distribution over the model space is given by

$$p(M \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid M)p(M)}{p(\mathbf{y})} \propto p(\mathbf{y} \mid M)p(M)\,. \tag{93}$$

One may then choose the model with the highest posterior probability which gives the maximum a posteriori (MAP) model

$$M_{\text{MAP}} = \arg\max_M p(M \mid \mathbf{y})\,. \tag{94}$$

This involves setting prior probabilities for each model $p(M)$. If the models are given equal prior probabilities $p(M) \propto 1$, the posterior probabilities become proportional to the marginal likelihood $p(M \mid \mathbf{y}) \propto p(\mathbf{y} \mid M)$. The maximization of the marginal likelihood is a fairly popular model selection method in Bayesian modelling and is sometimes also referred to as the type II maximum likelihood.

Figure 2 illustrates the marginal likelihood based model selection. The horizontal axis describes all the possible measurements and the curves show the marginal likelihoods (i.e. prior predictive distributions) of three different models. Under the simplest model, only a small range of different datasets are possible, whereas the most complex model allows a much wider range of measurements. However, because the marginal likelihood is a normalized density function of $\mathbf{y}$, the density value will be higher for the simpler model for datasets $\mathbf{y}$ it can explain. If one obtained data marked by the vertical dashed line, then the model with intermediate complexity would be chosen, because it is the simplest model that can account for the observed data. Thus model selection based on maximum marginal likelihood corresponds to *Occam's razor*: out of equally good explanations, choose the simplest.

While this sounds intuitively nice, marginal likelihood is in general sensitive to prior assumptions. To see this, consider the factorized form

$$\begin{aligned} p(\mathbf{y} \mid M) &= p(y_1, \ldots, y_n \mid M) \\ &= p(y_1 \mid M)p(y_2 \mid y_1, M) \ldots p(y_n \mid y_1, \ldots, y_{n-1}, M)\,. \end{aligned} \tag{95}$$

The first term is the predictive distribution for the first observation and depends on the prior only $p(y_1 \mid M) = \int p(y_1 \mid \boldsymbol{\theta}, M)p(\boldsymbol{\theta} \mid M)d\boldsymbol{\theta}$. If the prior is very uninformative, the predictive distribution for the first observation will be very flat and therefore $p(y_1 \mid M) \approx 0$ and also $p(\mathbf{y} \mid M) \approx 0$. The prior sensitivity of marginal likelihood is sometimes referred to as Lindley's paradox (Shafer, 1982; Kass and Raftery, 1995).

Another difficulty is that the computation of the marginal likelihood is in general nontrivial. The integral

$$p(\mathbf{y} \mid M) = \int p(\mathbf{y} \mid \boldsymbol{\theta}, M)p(\boldsymbol{\theta} \mid M)d\boldsymbol{\theta} \tag{96}$$

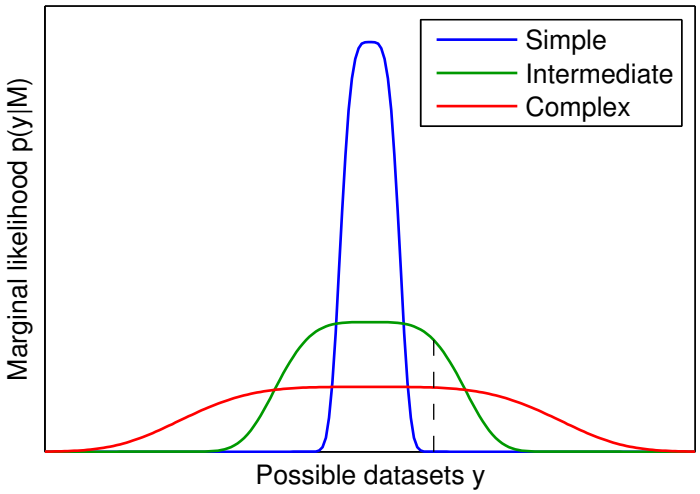


Figure 2: Schematic illustration of model selection based on maximum marginal likelihood. The horizontal axis describes all the possible datasets and the curves depict the marginal likelihoods of three different models. The maximum marginal likelihood principle favors the simplest model that is is able to explain the data. For the measurements marked by the dashed line the chosen model would be the model with intermediate complexity. The figure is reproduced from Rasmussen and Williams (2006).

is analytically available only for the simplest models and must usually be approximated somehow. Some of the proposed methods for this purpose are Laplace approximation, Schwarz criterion (also known as Bayes information criterion, BIC) and different MCMC methods. The first two are reviewed by Kass and Raftery (1995) and the MCMC methods by Han and Carlin (2001). There are also general methods for posterior approximation such as variational Bayes and expectation propagation that give an estimate of the marginal likelihood as a "by-product" (see for example Bishop, 2006). Section 4.2 discusses a method called reversible jump MCMC (RJMCMC) which can be used to sample parameters from different models according to their posterior probabilities. The relative posterior odds

$$\frac{p(M_i \mid \mathbf{y})}{p(M_j \mid \mathbf{y})} = \frac{p(\mathbf{y} \mid M_i)p(M_i)}{p(\mathbf{y} \mid M_j)p(M_j)} \tag{97}$$

can then be calculated according to the number of visits in each model. When the uniform prior $p(M) \propto 1$ is used, the relative posterior odds reduce to the ratio of the marginal likelihoods $p(\mathbf{y} \mid M_i)/p(\mathbf{y} \mid M_j)$ which is referred to as the Bayes factor. With the uniform prior, the choice of the most frequently occurring model is equivalent to maximizing marginal likelihood, even though the marginal likelihoods are not calculated explicitly.

Median model Barbieri and Berger (2004) proposed a variable selection method for normal linear model, and they called it the Median model. Median model is

defined as the model containing all the variables with overall posterior probability greater than $\frac{1}{2}$. Let binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ denote which of the variables are included in the model ($\gamma_j = 1$ meaning that variable j is included). The posterior inclusion probability of variable j is then

$$p_j = \sum_{M_k | \gamma_j = 1} p(M_k \mid \mathbf{y}) \tag{98}$$

that is, the sum of the posterior probabilities of the models which include variable j. Median model $\boldsymbol{\gamma}^{\text{med}}$ is then defined componentwise as

$$\gamma_j^{\text{med}} = \begin{cases} 1, & \text{if } p_j \geq \frac{1}{2} \\ 0, & \text{otherwise}\,. \end{cases} \tag{99}$$

The authors showed that under some specific assumptions the median model is optimal. By optimal the authors mean the model whose mean prediction for future $\tilde{y}$ is closest to the mean of the Bayesian model averaging prediction in the squared error sense. The key assumption is that the predictors are mutually uncorrelated, that is $\mathbf{Q} = \mathrm{E}\left[\mathbf{x}\mathbf{x}^T\right]$ is diagonal, where the expectation is taken over the true data generating distribution. This is a strong assumption, and the authors themselves admit that it does not often apply. Median model is also built on the assumption that the optimality is defined in terms of the mean predictions. More precisely, the theory does take into account the uncertainty in the predictive distributions. It might be that the median model has, for instance, much narrower predictive distribution than BMA underestimating therefore the uncertainty in the future observation, which might be undesirable.

3.5 Selection bias

As discussed in section 3.1, the performance of a model is usually defined in terms of an expected utility (54), and often with the logarithmic score function (55). Many of the proposed selection criteria rewieved in sections 3.2 and 3.3 are based on estimating this quantity in one way or another, even though some of the methods use different utility function than the logarithmic score. As discussed in section 3.4, also the selection of maximum a posteriori model measures the predictive ability of the candidate models, even though it can not be considered as a direct estimation of (55).

Consider a hypothetical utility estimation method. For a fixed training dataset $\mathcal{D}$, its utility estimate $g_k = g(M_k, \mathcal{D})$ for model M_k can be decomposed as

$$g_k = u_k + \varepsilon_k, \tag{100}$$

where $u_k = u(M_k, \mathcal{D})$ represents the true generalization utility of the model, and $\varepsilon_k = \varepsilon(M_k, \mathcal{D})$ is the error in the utility estimate. Note that also u_k depends on the observed dataset $\mathcal{D}$; favourable datasets lead to better generalization performance. If the utility estimate is correct on average over the different datasets $\mathrm{E}_{\mathcal{D}}[g_k] = \mathrm{E}_{\mathcal{D}}[u_k]$

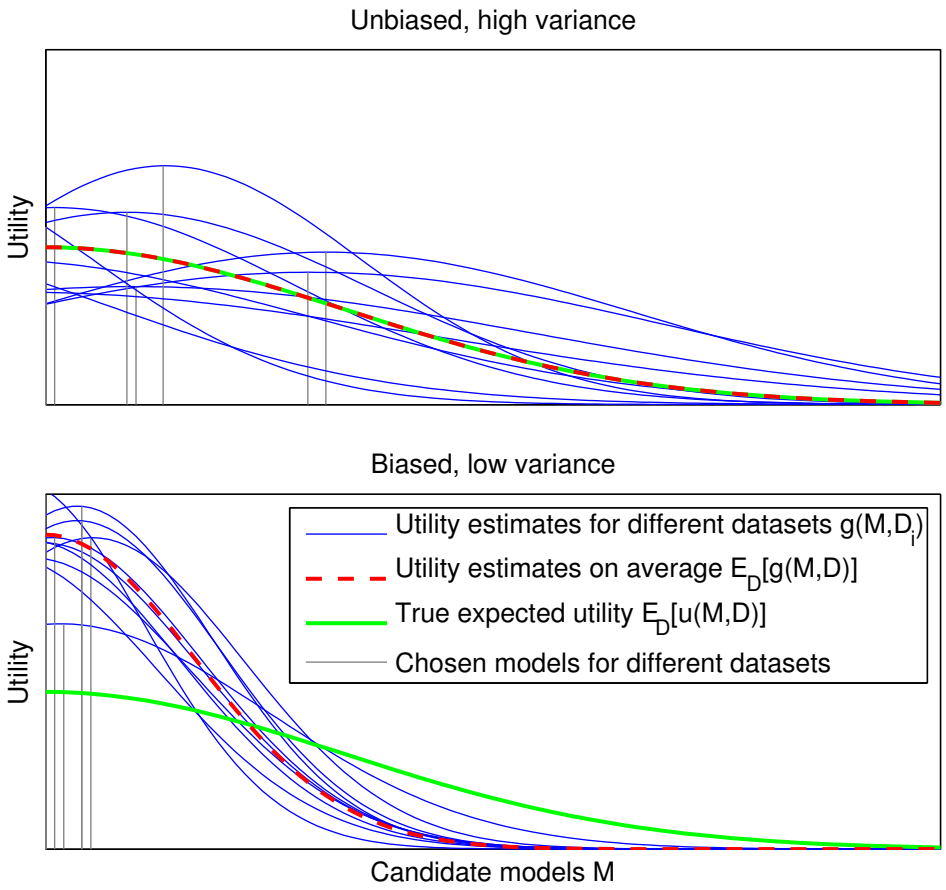


Figure 3: Schematic illustration of unbiased (top) and biased (bottom) utility estimation methods. In this case, the biased method is likely to choose better models due to smaller variance in the utility estimates. The figure mimics the ideas presented by Cawley and Talbot (2010).

or equivalently $\mathrm{E}_{\mathcal{D}}[\varepsilon_k] = 0$, the estimate g is said to be unbiased. If $\mathrm{E}_{\mathcal{D}}[\varepsilon_k] \neq 0$, the estimate is biased. The unbiasedness of a utility estimate is typically considered as beneficial for model selection. While this may sound intuitively sensible, it may not be the case in reality. The unbiasedness of a utility estimate is intrinsically unimportant for model selection, and a successful model selection does not necessarily require unbiased utility estimates (the unbiasedness can actually be disadvantageous as we shall see in a moment). To see this, note that the only requirement for perfect model selection criterion is that $g_k > g_l$ implies $u_k > u_l$ for all models M_k and M_l. In other words, the higher utility estimate implies higher generalization performance. This condition can be satisfied even if $\mathrm{E}_{\mathcal{D}}[g_k] \neq \mathrm{E}_{\mathcal{D}}[u_k]$.

But why might the unbiasedness be a problem and biasedness a benefit? Take a look at figure 3. The top plot shows an imaginary prototype of unbiased but high variance utility estimates. The blue curves represent the estimated utilities for each model M with different data realizations. On average (red dashed) these curves coincide with the true expected utility over all datasets (green). However, due to

the high variance, the maximization of the utility estimate may lead to choosing a model with nonoptimal expected true utility (the maxima become more scattered to the right). This phenomenon of choosing a nonoptimal model due to the variance in the utility estimates is called the *selection induced bias* or just *selection bias*. In other words, the selection procedure fits to the noise in the utility estimates, and therefore it is expected that the chosen model has a nonoptimal true utility. The lower plot then shows a biased utility estimation method that overestimates the ability of models on the left and underestimates the ability of the models on the right. However, due to smaller variance, the probability of choosing a model with better true performance is increased (the maxima focus more to the left). This example demonstrates that even though the unbiasedness is beneficial for the performance evaluation of a *particular* model, it is not necessarily important for model *selection*. For the selection, it is more important to have a right combination of bias and variance.

The top plot in figure 3 also demonstrates that, even though the utility estimates are unbiased for each model before the selection, the utility estimate for the *selected* model is no longer unbiased and is typically optimistic (the maxima of each individual curve tend to lie over the average curve). To get an intuitive idea how this phenomenon arises, consider the following simple example. Ask a thousand people to throw dice ten times and then choose the person who got the best total score. Then ask him or her to throw again and record the score. It is likely that the result for the second round is less than in the first round, because the result in the first round was probably higher that it could be expected on average. The phenomenon seems obvious when presented this way but may be easy to forget when put into another context.

The selection bias is an important concept that has received relatively little attention compared to the vast literature on the model selection in general. However, the topic has been discussed for example by Reunanen (2003), Varma and Simon (2006), and Cawley and Talbot (2010). These authors discuss mainly the model selection using cross validation, but the ideas apply also to other utility estimation methods. As discussed in section 3.2.2, cross validation gives a nearly unbiased estimate of the generalization performance of a single model, but the selection based on cross validation may become biased when the variance in the utility estimates is high (as depicted in the top plot of figure 3). This will be demonstrated empirically in section 5. The experiments will also demonstrate that the reference methods (section 3.3) seem to be less vulnerable to selection induced bias, which is possibly explained by smaller variance in the utility estimates. For this reason the reference model based model selection may yield better results even though the reference utilities are not unbiased estimates of the actual generalization performance (being therefore closer to the case in the bottom plot of figure 3). The variance in the utility estimate is different for different estimation methods, but generally the noise is large when the data is scarce. The probability of fitting to noise increases also with a large number of models being compared, which is the case for example in variable selection.

4 Variable selection

Variable selection is an important special case of the model selection problem discussed in section 3. As pointed out in the introduction, loosely speaking, the goal in variable selection is to select a minimal subset from a set of possible input variables $\mathbf{x} = (x_1, \ldots, x_p)$ while maximizing the predictive ability of the model with respect to the target variable y. It is assumed here that other aspects of the model than the included variables are fixed. In other words, the model structure (for example linear or Gaussian process model) is fixed and one is comparing only different variable combinations within this model. Typically there is a tradeoff between the number of included variables and the predictive performance of the model. Thus it depends on the situation whether the preference is the reduction of the variables or the predictive performance of the model. If one is simply interested in maximizing the predictive ability of the model, the formally correct Bayesian approach is to include all the variables in the model, place a prior that best reflects the beliefs of the modelling task, and then integrate over the posterior uncertainty. However, often the goal is to be able to learn something about the underlying process that generates the data. In such a situation the identification of the relevant variables is of central importance, because this improves the interpretability of the model. An example would be the survival studies where it would be of interest to identify the factors that increase the probability of one getting the disease. In some applications the variable selection might also lead to reduced measurement costs if some of the variables can be left unobserved in the subsequent measurements. This might be the case for example in industrial or medical research. An additional benefit of variable selection could also be the reduced computational costs in the subsequent model fitting.

In principle the variable selection could be carried out in the same fashion as any other model selection problem, that is, by choosing the model that has the highest estimated expected utility (according to some criterion discussed in section 3). However, variable selection poses a few extra challenges. First, the number of different variable combinations $K = 2^p$ grows exponentially with the number of variables p. For instance, with $p = 30$ the number of different models is $K \approx 10^9$. Thus the exhaustive search through the model space becomes infeasible already for a relatively small number of variables and heuristic search strategies must be employed. A few of these approaches are discussed in section 4.1. Another difficulty is the overfitting of the selection process due to the selection bias. As discussed in section 3.5, the risk of overfitting due to the selection bias increases when the number of models being compared is large. The selection process is vulnerable to overfitting especially when the data is scarce, because the variance in the utility estimates becomes then high. This is demonstrated in section 5.

4.1 Search strategies

Variable selection (or more generally the subset selection problem) can be stated formally as follows. Given a set of variable indices $S = \{1, \ldots, p\}$ and a utility func-

tion f that maps each subset $S_k \subset S$ to a scalar value $f : S_k \mapsto \mathbb{R}$, find the subset S_* that maximizes f, that is $f(S_*) \geq f(S_k)$ for all $S_k \subset S$. As discussed already, the exact solution would require going through all the possible subsets of S and becomes therefore quickly infeasible as p gets larger. This section presents heuristic search strategies that can be used for searching candidate models and finding an approximate solution to the maximization problem.

Forward selection One simple but useful search strategy is the forward selection (also known as forward search, greedy search and greedy algorithm). In forward search one starts from the empty set (no variables) and adds variables one at a time such that the objective function is maximized at each step. The forward search can be stated as follows:

1. Initialize the sets of added and not added variable sets as $S_{\text{in}} = \emptyset$, $S_{\text{out}} = S$.
2. Repeat until all the variables have been added $S_{\text{in}} = S$, $S_{\text{out}} = \emptyset$:
   (a) Choose variable $j \in S_{\text{out}}$ so that $f(S_{\text{in}} \cup j)$ is maximized. Remove j from S_{out} and add it to S_{in}.

The algorithm requires p objective function evaluations at the first round, $p - 1$ at the second round, then $p - 2$ and so on. Thus the algorithm goes through $1 + 2 + \cdots + p = \frac{p(p+1)}{2} = O(p^2)$ objective function evaluations. In variable selection terms, this is the number of compared models. Depending on how long one function evaluation (or model training) takes, the full forward search might be feasible up to a few hundred or thousand variables. As a benchmark, if one objective function evaluation takes one second, then going through a full forward search for $p = 1000$ variables takes about 6 days. As the objective function evaluations are not typically smaller than this by several orders of magnitude, one can conclude that this type of searching is applicable to small and medium sized problems, but infeasible for very large problems (tens of thousands of variables or more). However, one can always stop the iteration when a predefined number of variables or a suitable stopping criterion is satisfied, which may reduce the computational time considerably.

Backward selection A closely related search strategy to the forward search is the backward selection, which does the search in the opposite direction. In backward selection one starts from the full set of variables and reduces the variables one at a time by optimizing the utility function. The backward selection proceeds as follows:

1. Initialize the sets of added and not added variables as $S_{\text{in}} = S$, $S_{\text{out}} = \emptyset$.
2. Repeat until all the variables have been removed $S_{\text{in}} = \emptyset$, $S_{\text{out}} = S$:
   (a) Choose index $j \in S_{\text{in}}$ so that $f(S_{\text{in}} \setminus j)$ is maximized. Remove j from S_{in} and add it to S_{out}.

The computational complexity in backward selection is equivalent to the forward selection, but the search paths are in general different. In section 5.1 the experimental difference between the forward and backward search is briefly discussed.

Probability based search An alternative to the forward and backward searches is to order the variables according to their probability, and then choose some suitable number of variables. Recall that the Median model presented in section 3.4 is defined to include all the variables that have marginal probability of 0.5 or more. However, one can use marginal probabilities only for sorting the variables, and then decide the number of variables by maximizing the utility estimate. One could expect that the selection induced bias would be smaller with this type of approach compared to the forward and backward searches, because the number of models under comparison is smaller (only p models are compared). However, ordering the variables accoring to their probabilities introduces selection bias because the order is already fitted to the data. The experimental results for probability based sorting in relation to forward and backward searches are briefly considered in section 5.

Typically the marginal probabilities of the different variables are estimated by sampling from the model space and estimating the relative probabilities of the different models (that is, the different variable combinations). After this it is straightforward to estimate the marginal probability of a particular variable by summing up the probabilities of the models that include this variable. Section 4.2 discusses how the probabilities of the different models are estimated via sampling based methods.

4.2 Sampling the model space

Several methods for sampling between the different variable combinations have been proposed, mainly for the linear and generalized linear model (George and McCulloch, 1993, 1997; Kuo and Mallick, 1998; Dellaportas et al., 2000). These methods sample the different variable combinations by setting a spike-and-slab prior (section 2.2.2) for each input weight. The spike-and-slab assumption is equivalent to formulating the posterior using the different variable combinations

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \sum_{\boldsymbol{\gamma}} p(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma} \mid \mathcal{D}) = \sum_{\boldsymbol{\gamma}} p(\boldsymbol{\theta}_{\boldsymbol{\gamma}} \mid \mathcal{D}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma} \mid \mathcal{D}), \tag{101}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ denotes the variable combination so that $\gamma_i = 1$ if variable i is in the model and $\gamma_i = 0$ otherwise. In the above formula the summation runs over all the different variable combinations $\boldsymbol{\gamma}$, and the notation $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ emphasizes that the parameter values depend on the variable combination. From the latter form of (101) one can see that the full posterior can be written as a mixture of the individual model posteriors each weighted by the corresponding posterior probability $p(\boldsymbol{\gamma} \mid \mathcal{D})$. The different proposed methods are designed to sample the distribution (101) in various ways. All the different approaches are not discussed here, see instead the study by O'Hara and Sillanpää (2009) for a review.

Let us, however, discuss a general method called reversible jump Markov Chain Monte Carlo (RJMCMC) (Green, 1995) which can be used to sample from and

arbitrary mixture of individual model posteriors

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \sum_{M_k} p(\boldsymbol{\theta}_k \mid \mathcal{D}, M_k) p(M_k \mid \mathcal{D}) \,. \tag{102}$$

RJMCMC samples the parameters of the different models such that each model is sampled according to its probability. In other words, the chain travels randomly from model to another such that the more probable models are visited more often and therefore also most of the samples are from these models. In practice, if there are a huge number of different models, then only models with considerable posterior probability are sampled and models with very small probability may not be visited at all.

Since the different models have in general different number of parameters, one must introduce an auxiliary variable $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)$ that does the dimension matching between any two models. The algorithm can be summarized as follows:

1. Let the current state be $(M_1, \boldsymbol{\theta}_1)$, that is, model M_1 with parameters $\boldsymbol{\theta}_1$. A jump to another model M_2 is proposed with probability J_{12} and with probability $1 - J_{12}$ an MCMC jump within the current model is taken.

2. If a jump to model M_2 is proposed, generate the associated auxiliary variable from proposal density as $\mathbf{u}_1 \sim q(\mathbf{u}_1 \mid \boldsymbol{\theta}_1, M_1, M_2)$ and calculate the parameters of the proposed model as $(\boldsymbol{\theta}_2, \mathbf{u}_2) = h_{12}(\boldsymbol{\theta}_1, \mathbf{u}_1)$. Here h_{12} is an invertible vector valued function that defines the transformation between the different parameter spaces.

3. Define the ratio

$$r = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}_2, M_2)\, p(\boldsymbol{\theta}_2 \mid M_2)\, p(M_2)\, J_{12}\, q(\mathbf{u}_2 \mid \boldsymbol{\theta}_2, M_2, M_1)}{p(\mathcal{D} \mid \boldsymbol{\theta}_1, M_1)\, p(\boldsymbol{\theta}_1 \mid M_1)\, p(M_1)\, J_{21}\, q(\mathbf{u}_1 \mid \boldsymbol{\theta}_1, M_1, M_2)} \left| \frac{\partial h_{12}(\boldsymbol{\theta}_1, \mathbf{u}_1)}{\partial(\boldsymbol{\theta}_1, \mathbf{u}_1)} \right|$$

   and accept the new sample with probability $\alpha = \min(1, r)$. The last factor in r is the determinant of the Jacobian matrix of h_{12}.

The above presents the algorithm in its general form. For sampling between the different input variable combinations, one can simplify the above expression for the acceptance probability. Let the current state be $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$. Propose to change the state of a randomly selected bit γ_k with some probability, for example 0.5. If a new input k is being added, one can set the new parameters as $\boldsymbol{\theta}_2 = (\boldsymbol{\theta}_1, \mathbf{u}_1)$, where $\mathbf{u}_1$ is the parameter corresponding to the new input, for example the associated weight $\mathbf{u}_1 = w_k$ in case of the linear model. In this case there is no variable $\mathbf{u}_2$, the determinant of the Jacobian will be 1 (because the transformation is identity) and $J_{12} = J_{21}$. Moreover, using the conditional prior as the proposal distribution $q(\mathbf{u}_1 \mid \boldsymbol{\theta}_1, M_1, M_2) = p(\mathbf{u}_1 \mid \boldsymbol{\theta}_1, M_2)$, the prior terms cancel out, and the acceptance probability simplifies to

$$\alpha = \left(1, \frac{p(\mathcal{D} \mid \boldsymbol{\theta}_2, M_2)\, p(M_2)}{p(\mathcal{D} \mid \boldsymbol{\theta}_1, M_1)\, p(M_1)}\right) .$$

On the hand, if an input is being removed, one sets $(\boldsymbol{\theta}_2, \mathbf{u}_2) = \boldsymbol{\theta}_1$ where $\mathbf{u}_2$ is the parameter corresponding to the removed variable. In this case one does not have $\mathbf{u}_1$ and hence there is no need to generate any sample. All the other things stay the same, and the acceptance probability will be given by the above formula.

As presented above, the calculation of the jumping probability between the different variable combinations involves only the evaluation of the likelihood function and the prior probabilities of the models. The sampling from the conditional priors is also straightforward if the priors are of simple form such as Gaussian. The advantages of RJMCMC are its applicability to different types of models and a relatively easy implementation. It is also typically quite fast even for many variables. However, for a large number of input variables, one must typically run very long chains for a good mixing between different combinations. This is true especially if only a single input variable is being set on or off at a time. Discussion about the properties of RJMCMC in relation to other sampling methods for linear models can be found from the review by O'Hara and Sillanpää (2009).

4.3 Priors on the model space

The sampling of different variable combinations involves setting prior probabilities for each of the combinations. The easiest approach would be to use the uninformative prior $p(\boldsymbol{\gamma}) = 1/2^p \propto 1$ giving equal weight for all the models. This approach has the problem that, even though being uninformative with respect to the variable combination, it is quite informative about the number of variables included in the model. This can be seen by writing the prior in the form

$$p(\boldsymbol{\gamma} \mid \pi) = \pi^k (1-\pi)^{p-k}, \tag{103}$$

where k is the number of included variables (number of nonzeros in $\boldsymbol{\gamma}$) and $\pi = 0.5$ denotes the probability of including a single variable. The total probability of having k variables in the model is then the sum of all the models of size k

$$p(k \mid \pi) = \binom{p}{k} \pi^k (1-\pi)^{p-k}, \tag{104}$$

which is a binomial distribution. Because the binomial distribution has most of its mass around $k = \pi p = 0.5p$, this prior favors models having about half of the variables. In other words, although all the models are equally probable, it is likely that a randomly chosen model would have about $0.5p$ variables simply because there are many more such combinations than any other. This prior is illustrated in figure 4 (black dots) when the number of variables is $p = 100$.

Typically in variable selection problems one would like to favor smaller models to incorporate the prior knowledge of possibly irrelevant variables. This can be done by adjusting π and fixing it to a smaller value. For example $\pi = 0.2$ would then favor models having about 20% of the variables (figure 4, blue dots). However, this requires a careful estimation of π because the prior is still quite informative about the model size. An alternative and more flexible approach is to set a hyperprior for

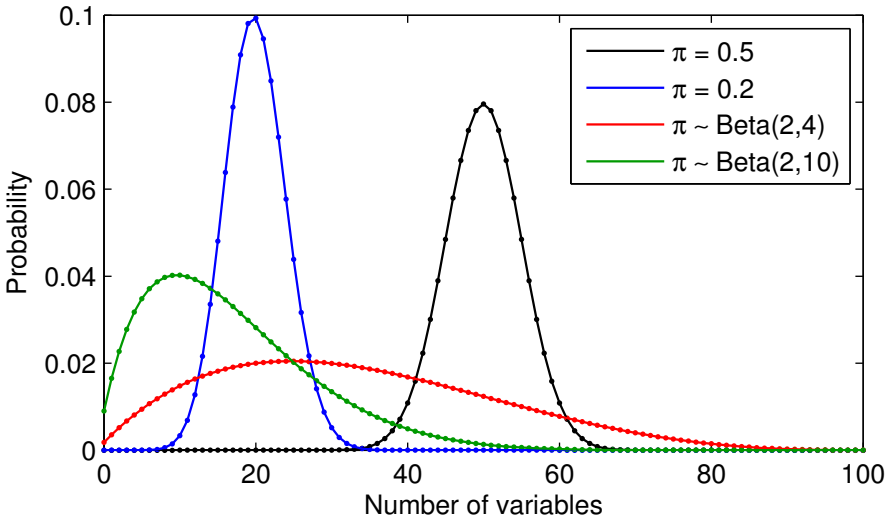


Figure 4: Different priors for the number of variables k when the prior for variable combination $\boldsymbol{\gamma}$ is $p(\boldsymbol{\gamma}) = \pi^k(1-\pi)^{p-k}$ and the total number of variables is $p = 100$. The "uninformative" prior $\pi = 0.5$ is actually quite informative with respect to the model size.

π and integrate over it. A convenient choice is the Beta distribution

$$p(\pi) = \text{Beta}\,(\alpha, \beta) = \frac{1}{\text{B}(\alpha, \beta)}\pi^{\alpha-1}(1-\pi)\pi^{\beta-1}, \tag{105}$$

where $\text{B}(\alpha, \beta)$ is the normalization constant, called the Beta-function

$$\text{B}(\alpha, \beta) = \int_0^1 \pi^{\alpha-1}(1-\pi)\pi^{\beta-1}d\pi\,. \tag{106}$$

This yields

$$p(\boldsymbol{\gamma}) = \int_0^1 p(\boldsymbol{\gamma}, \pi)d\pi = \int_0^1 p(\boldsymbol{\gamma} \mid \pi)p(\pi)d\pi = \frac{\text{B}(\alpha + k, \beta + p - k)}{\text{B}(\alpha, \beta)}\,.$$

The prior for model size becomes then

$$p(k) = \binom{p}{k}\frac{\text{B}(\alpha + k, \beta + p - k)}{\text{B}(\alpha, \beta)}, \tag{107}$$

which is called the Beta-binomial distribution. One can choose the hyperparameters α and β by plotting (107) for several values of α and β and then choosing the combination that best reflects the prior beliefs of the model size. Figure 4 shows this prior when $(\alpha, \beta) = (2, 4)$ and $(\alpha, \beta) = (2, 10)$. As can be seen, the Beta-binomial allows much more flexible distribution for model size and this approach can be useful when the information about the "correct" model size is vague.

5 Numerical experiments

This section experiments with the model selection methods presented in section 3. The first example in section 5.1 consist of simulated data for linear regression. Simulated data is convenient as it allows an arbitrary adjustment of the problem size and especially the ratio between the number of variables and the training set size. These experiments will illustrate the main points of this thesis. The other example in section 5.2 considers a binary classification problem with a real world dataset. This will serve as a realistic example of a challenging variable selection problem in practice and will further confirm the conclusions from section 5.1.

5.1 Simulated data

To get an idea of how the different selection methods perform for different data configurations, simulated datasets were generated for different training set sizes n, total number of input variables p and level of correlations ρ between the variables. The data is distributed as follows

$$\begin{aligned} \mathbf{x} &\sim \mathrm{N}(0, \mathbf{R}), \quad \mathbf{R} \in \mathbb{R}^{p \times p} \\ y \mid \mathbf{x} &\sim \mathrm{N}\left(\mathbf{w}^T \mathbf{x}, \sigma^2\right), \quad \sigma^2 = 1 \end{aligned}$$

where the covariance matrix $\mathbf{R}$ is block diagonal consisting of blocks $\mathbf{R}_1, \ldots, \mathbf{R}_q \in \mathbb{R}^{5 \times 5}$. The elements in the blocks are defined as

$$[\mathbf{R}_k]_{ij} = \begin{cases} 1, & i = j \\ \rho, & i \neq j \,. \end{cases}$$

In other words, the input variables are divided into q groups of 5 variables. Each variable has a unit variance and is correlated with other variables in the same group with coefficient ρ but uncorrelated with variables in the other groups. The number of groups q depends on the total amount of variables p which is varied. The true weights $\mathbf{w} = (w_1, \ldots, w_p)$ for the inputs are defined as

$$w_i = \begin{cases} w_0, & i = 1, \ldots, 15 \\ 0, & \text{otherwise} \,. \end{cases}$$

In other words, the variables in the first three groups have equal nonzero weight w_0, and the rest of the variables are irrelevant. Now the output variable has variance $\mathrm{Var}[y] = \mathbf{w}^T \mathbf{R} \mathbf{w} + \sigma^2$ which depends on $\mathbf{R}$ and therefore on ρ. To get comparable results with different values of ρ, the amount of variance in y explained by the predictors is fixed as $\mathbf{w}^T \mathbf{R} \mathbf{w} / \mathrm{Var}[y] = 0.8$. For $\rho = 0, 0.5, 0.9$ this is satisfied by setting approximately $w_0 = 0.52, 0.30, 0.24$ respectively.

For a fixed variable combination, the model specification is

$$\begin{aligned} y \mid \mathbf{x}, \mathbf{w}, \sigma^2 &\sim \mathrm{N}\left(\mathbf{w}^T \mathbf{x}, \sigma^2\right) \\ \mathbf{w} \mid \sigma^2, \tau^2 &\sim \mathrm{N}\left(0, \tau^2 \sigma^2 \mathbf{I}\right) \\ \sigma^2 &\sim \text{Inv-Gamma}\left(0.5, 0.5^2\right) \\ \tau^2 &\sim \text{Gamma}\left(0.25, 4^2\right) \,. \end{aligned}$$

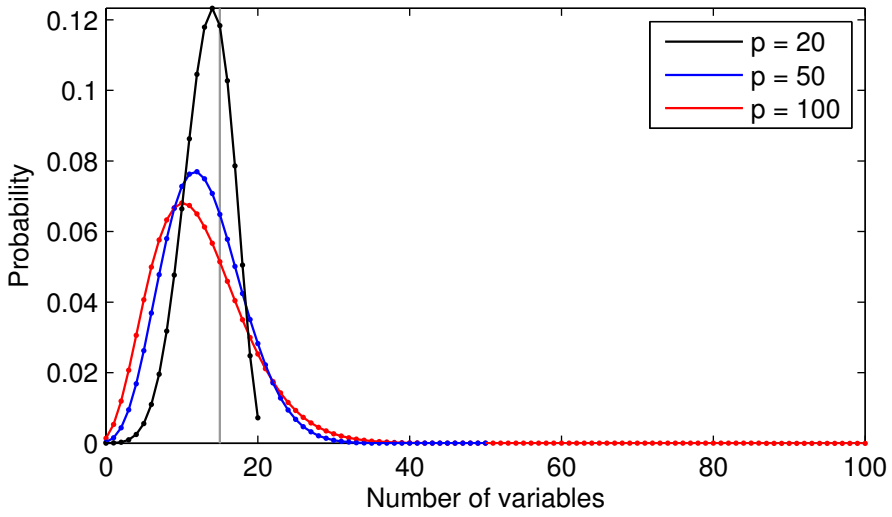


Figure 5: Simulated data: Priors on the number of input variables when the total number of variables varies $p = 20, 50, 100$. The true number of variables (15) in the data generating distribution is marked by the grey line.

For a fixed prior variance for the weights τ^2 this is the normal linear model with conjugate prior discussed in section 2.2.1. Now τ^2 is handled as an unknown and it is given a relatively uninformative prior. For submodels the integration over τ^2 is carried out using the grid approximation as explained in section 2.2.1. The intercept term is handled by adding a vector of ones to the design matrix $\mathbf{X}$ and it is given the same prior as the other weights. For notational convenience, the binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ denoting which of the variables are included in the model is omitted in the above formulas. The prior is the same for each submodel $\boldsymbol{\gamma}$. The reference model is constructed as the Bayesian model average from the submodels using the reversible jump MCMC as discussed in section 4.2, which corresponds to setting the spike-and-slab prior for the full model. In the RJMCMC algorithm, the samples from the posteriors of the submodels (15) are easily drawn consecutively by first sampling τ^2, then σ^2 given τ^2, and finally $\mathbf{w}$ given σ^2. The prior for the number of variables is set to be a Beta-binomial distribution (section 4.3) with parameters α and β varying according to the total number of variables. The exact values of α and β will be discussed shortly.

The experiments were repeated varying the training set size $n = 20, 50, 100$, the total number of variables $p = 20, 50, 100$ and the correlation coefficient $\rho = 0, 0.5, 0.9$. Given that the true number of relevant variables is 15, the cases $p = 20$ and $p = 100$ correspond to relatively dense and sparse problems (ratio of completely irrelevant variables is 25% and 85%, respectively). Some of the settings also have less training points than variables corresponding to "large p, small n" -problems. For different problem sizes $p = 20, 50, 100$, the hyperparameters adjusting the prior on the number of variables were set to $(\alpha, \beta) = (10, 5), (7, 20), (5, 35)$ respectively. These priors are plotted in figure 5. For each combination of n, p and ρ, the variable selection was performed 100 times for different data realizations with each method listed in table

Abbreviation	Method
LOO-CV	Leave-one-out cross validation maximization (eq. (61))
WAIC	WAIC maximization (eq. (71))
DIC	DIC maximization (eq. (68))
L-CRIT	L-criterion minimization (eq. (75))
LQ	L_q-criterion minimization (eq. (76))
LGG	L_{GG}-criterion minimization with $k = 1$ (eq. (78))
MAP	Maximum a posteriori model (estimated from RJMCMC) (eq. (94))
Median	Median model (estimated from RJMCMC) (eq. (99))
Projection	Projection of BMA to submodels, smallest model having 95% explanatory power (eq. (87) and (88))
Ref-KL	Posterior predictive discrepancy minimization from BMA (eq. (83))
Ref-KL-CV	CV-predictive discrepancy minimization from BMA, 10-fold-CV (eq. (84))

Table 2: Compared variable selection methods for simulated experiments. The methods are discussed in section 3.

2, and then the performance of the selected models were tested on independent test set of size $n_t = 1000$. As a proxy of the generalization performance, the mean log predictive density (MLPD) was used.

Figure 6 shows the average number of selected variables (6a) and the generalization performance of the selected models in each data setting in comparison to the BMA (6b). In 6b zero indicates the same predictive ability as the BMA and negative values worse. A striking observation is that when $n \leq p$ many of the methods perform poorly and choose overfitted models with bad predictive performance, in some settings clearly worse than the empty model (the colored dotted lines). This holds especially for LOO-CV, WAIC, DIC, L-CRIT, LQ and LGG, and the conclusion stays unchanged for all the levels of correlation between the variables (blue, red and green circles), albeit the high dependency between the variables seems to reduce the effect. These methods choose models with predictive performance close to BMA only when $p < n$. The observed behaviour is caused by the selection induced bias, which – due to scarce data and therefore high variance in the utility estimates – leads to selecting overfitted models as discussed in section 3.5. MAP and Median models perform better, but also these choose models with suboptimal predictive performance in some of the settings, for instance when $p = 50$ and $n = 50$. Overall, the reference methods Ref-KL, Ref-KL-CV and the Projection seem to perform best often choosing models with predictive performance close to BMA, but in some cases they are also choosing more variables than the other methods. The Projection seems the most effective in reducing the number of variables but still maintaining the predictive ability. However, for all the reference methods (Ref-KL, Ref-KL-CV, Projection) the number of chosen variables and performance of the chosen submodels are highly sensitive to the applied selection rule as will be discussed shortly (the selection rules used in these experiments are listed in table 2).

To get more insight to the problem let us now study the setting $p = 100$, $\rho = 0.5$ more closely. This should roughly correspond to a problem expected in reality; a moderate number of predictors out of which many are irrelevant, and some of the variables are moderately correlated with each other. Figure 7 shows the averaged forward search paths for LOO-CV, Ref-KL, Ref-KL-CV and Projection for training set sizes $n = 20, 50, 100, 200$. The solid black line shows the test utilities, red dashed the CV-utilities within the data used for selection, and the horizontal dashed line the test utility of the BMA. The search paths for LOO-CV (top row) demonstrate the overfitting due to the selection induced bias; starting from the empty model and adding variables one at a time one finds models that have high CV-utility but much worse test utility. In other words, the performance of the models at the search path is dramatically overestimated. Note, however, that for the empty model and model with all the variables the CV-utility and test utility are the same. If one is choosing the model based on the CV-maximum, this leads to choosing a model with a bad predictive performance if the data is scarce. The selection induced bias decreases when the size of the training set grows, but the effect is still visible for $n = 200$. It is noted that the behaviour is very similar also for the other methods based on sample reuse (WAIC, DIC, L-CRIT, LQ, LGG), but for convenience only results for LOO-CV are shown here for demonstration.

The reference methods show better performance as can be seen from the more desirable test utilities along the search paths. For the reference methods, the test utility curves also converge faster to a sharp elbow at the true number of variables (15). The projection works very well when $n = 100$ or $n = 200$, quite well for $n = 50$, but breaks down when $n = 20$. However, when $n = 20$ none of the methods is able to remove variables without significantly compromising the predictive ability. Note that due to the technical issues discussed in section 3.3.2 the projection is defined only when the number of parameters in the model is less than the number of datapoints. Therefore the utility curves can be calculated only up to n variables. The reason why both the CV-utility and the test utility decrease for Projection when the number of variables $k \approx n$ is that the associated KL-divergence shrunks then to (almost) zero regardless of the variable combination, and there is no guarantee of any generalization of the projected samples. In other words, given enough variables, one can project to completely irrelevant variables with no predictive ability but the associated KL-divergence is equal to zero. Yet another important observation from figure 7 is that the CV-utility is a biased estimate of the true predictive ability for the chosen models also when the search is performed using the reference methods or projection. Thus, this example demonstrates that even if one uses the reference predictive approach or projection for searching promising candidate models, the cross validation utility is in general an unreliable indicator about a suitable stopping point at the search path.

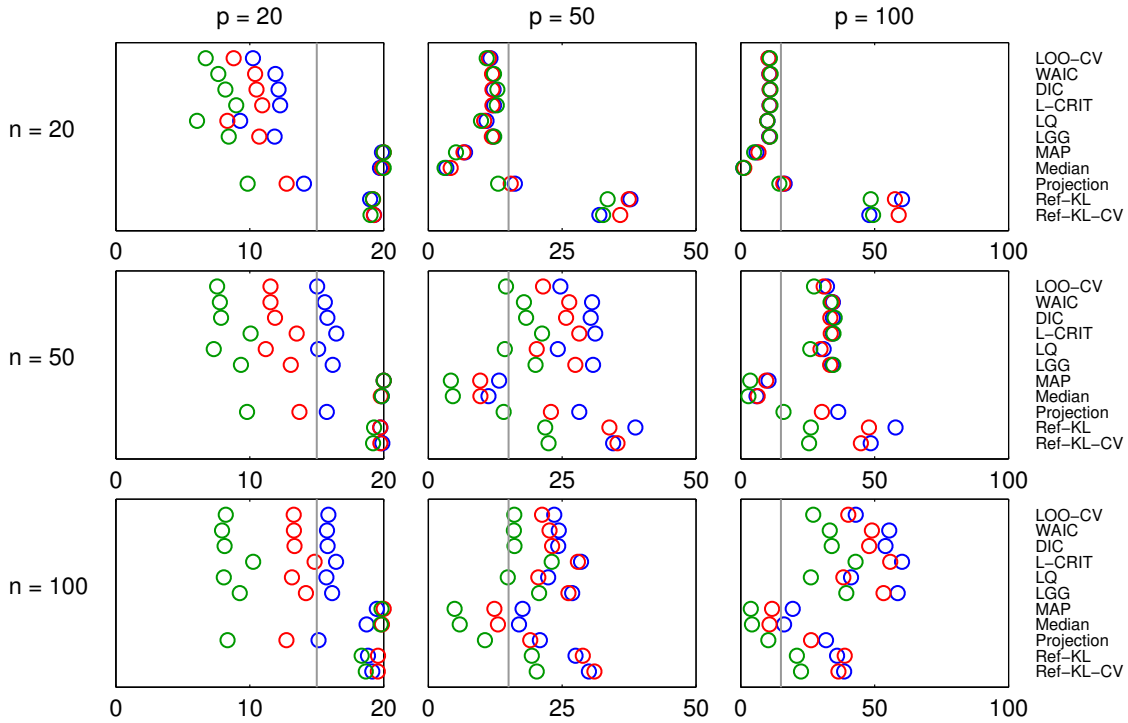
Figure 8 shows the estimated relative explanatory powers (eq. (88)) and the test utilities for the reference methods for the same problem. The test utilities are shown with respect to the BMA because the BMA obtains different test utility when the training set size is varied. The figure demonstrates that it may be very difficult to decide the number of chosen variables based on the estimated explanatory

power. This is because the correspondence between the explanatory power and the predictive ability may be very different for different sizes of the dataset. For large enough datasets ($n = 100, 200$) the correspondence may be good enough, and the number of selected variables could be chosen based on, for example, finding the elbow of the explanatory power curve. However, for smaller datasets this may work poorly. For instance, for Ref-KL the explanatory power curves seem quite similar for all the different training set sizes but the test utilities are markedly different. Especially for $n = 20$ the estimated explanatory power is a bad indicator of the predictive performance, and this holds also for Ref-KL-CV and Projection. Thus in practice there is no guarantee for the submodel having a good predictive performance even if the estimated explanatory power is close to one. The reason for this behaviour is again due to the selection induced bias. When the models are searched by minimizing the discrepancy-estimate, the obtained estimates are no longer unbiased for the models on the search path, and therefore they may be bad indicators of the performance of the submodels. Another difficulty with the explanatory power for Ref-KL and Ref-KL-CV is that none of the submodels has explanatory power equal to 100%, because for all the submodels the discrepancy from the BMA is nonzero. It seems that especially for Ref-KL-CV the maximum explanatory power might be as low as 80%, so it is problematic to decide whether this corresponds to acceptable discrepancy from the reference model or not (and if not, what should one then do?). Projection does not have this problem because the zero discrepancy is always obtained by allowing enough variables, but as is seen, the smaller discrepancy does not necessarily mean better predictions.

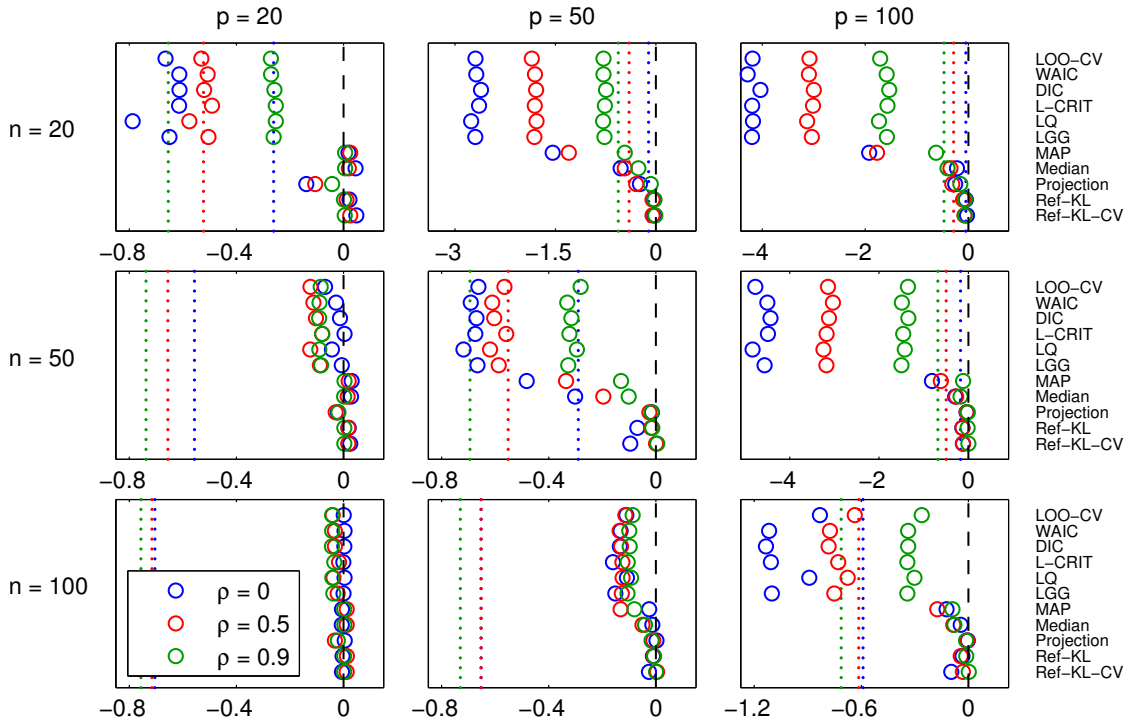
For completeness, also the effect of the used search strategy and the selection of MAP or Median model was considered for this same problem. The forward and backward searches were compared along with sorting the variables according to their estimated marginal probabilities. Figure 9 shows the results for LOO-CV and Ref-KL-CV. One can see that ordering the variables according to their probabilities reduces the overfitting effect with LOO-CV markedly. This makes sense; when the variables are sorted based on their probabilities one is comparing only p models, whereas in forward and backward searches the number of compared models is $\frac{p(p+1)}{2}$. The fewer model comparisons give smaller chance for the selection procedure to overfit to the noise in the utility estimates. However, sorting the variables according to their probabilities introduces also selection bias which can be seen from the drop in the test utility curve after a certain amount of variables have been added. Minimizing the KL-divergence from the BMA gives even better search paths when $n = 50, 100, 200$, although the performance is slightly inferior in the case $n = 20$. The results also suggest that the backward search works somewhat better in this example compared to the forward search. This seemed to be the case also for other combinations of p and n even though the results are not shown. Figure 9 also suggests that selecting the MAP or Median model may lead to inferior selection. For instance, the Median model seems to choose too small models with suboptimal predictive ability especially when $n = 20$ or $n = 50$. On the other hand allowing the inclusion of variables with smaller probability would not necessarily improve the selection because the predictive ability may even decrease when more

variables are added (note the shape of the green curve). The MAP model seems to have even worse predictive performance. It is also noted that for a moderate number of variables the estimation of the MAP model from the RJMCMC samples is problematic, because the number of visits in even the most probable models in the MCMC chains is typically low, which may lead to highly varying and unreliable results.

To shortly summarize these experiments, one can conclude that the "blind" maximization of a utility estimate such as LOO-CV or WAIC may lead to bad results due to the selection induced bias if the data is scarce and the number of models being compared is large. The selection bias reduces when the size of the dataset becomes larger and many of the methods are likely to choose models with good predictive performance (figure 6). The reference predictive methods and the projection seem to be less sensitive to the selection bias being therefore able to find better models. Especially the projection approach shows desirable behaviour even for quite small datasets (figure 7). However, also for the reference methods the estimated discrepancy from the reference model may be an unreliable indicator of the submodel's predictive performance. Thus the results may be highly sensitive to the actual selection rule (e.g. 95% explanatory power or minimum discrepancy). In fact, based on the figure 8, the obtained results may be more dependent on the selection rule than to which of these methods (Ref-KL, Ref-KL-CV, Projection) is used for searching the models.



(a) Average number of selected variables. The true number of variables (15) is marked by the grey line.



(b) Mean log predictive densities (MLPD) of the selected models in comparison to the BMA; zero means the same performance as BMA, negative values worse than BMA. The dotted lines denote the performance of the empty model (intercept term only) in comparison to BMA for different levels of correlation (note that the high correlation between the variables improves the fitting of the BMA and therefore the difference to the empty model increases).

Figure 6: Simulated data: Number of chosen variables (6a) and predictive performance of the chosen models (6b) on average. The results are divided into subfigures according to the total number of variables $p = 20, 50, 100$ and training set size $n = 20, 50, 100$. Colours mark the correlations between the variables in the same group $\rho = 0$ (blue), $\rho = 0.5$ (red), $\rho = 0.9$ (green). The search through model space was done using forward selection and the results are averaged over 100 data realizations. See the text for more precise description of the data.

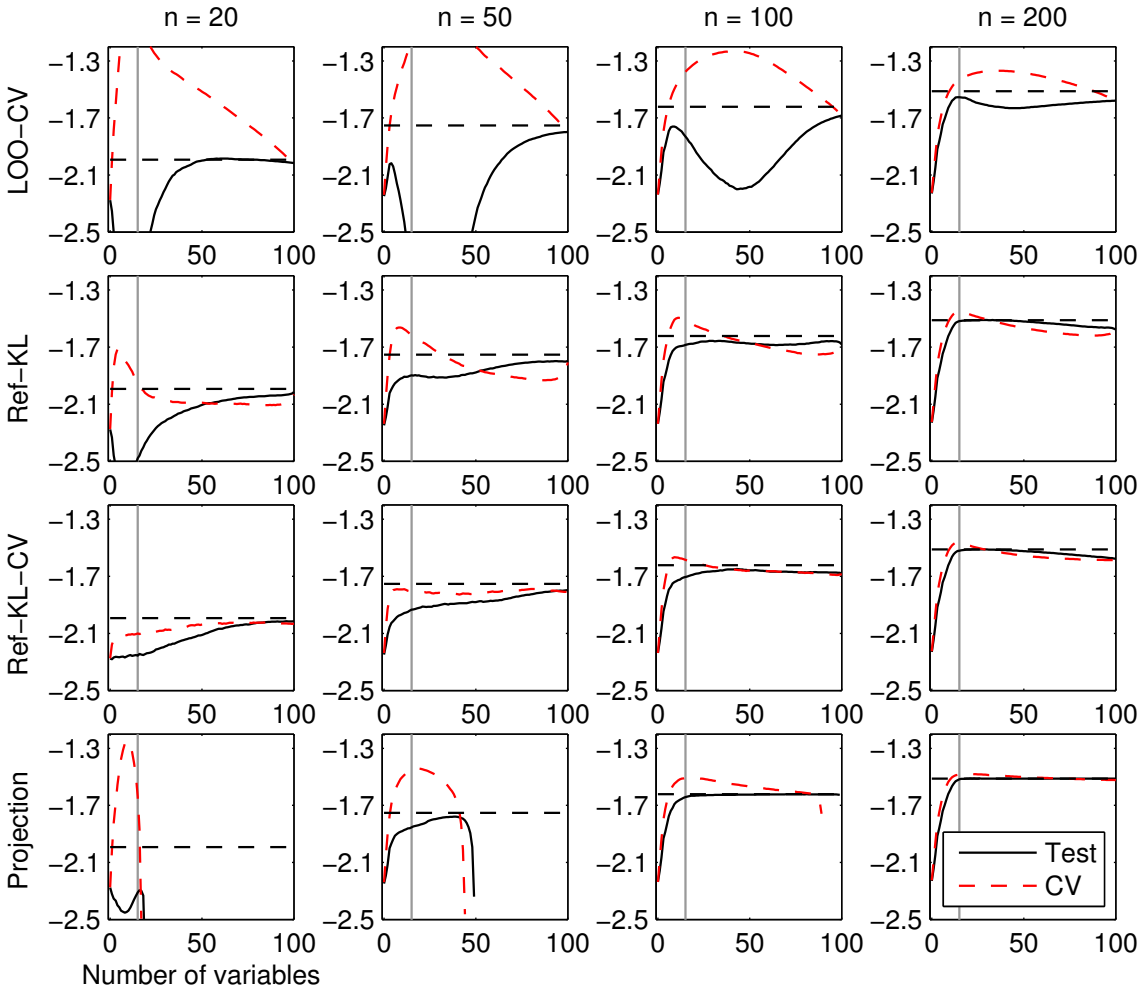


Figure 7: Simulated data: Average search paths for LOO-CV, Ref-KL, Ref-KL-CV and Projection when the total number of variables is $p = 100$, within group correlation $\rho = 0.5$ and size of the training dataset varies. Dashed red shows the cross validation MLPD and solid black the test MLPD for the submodels along the forward search path. Dashed black denotes the test MLPD of the BMA. Note that the most extreme values are limited outside the figures for convenience (for instance when $n = 20$). The true number of variables (15) is denoted by the grey vertical line. The results are averaged over 100 data realizations.

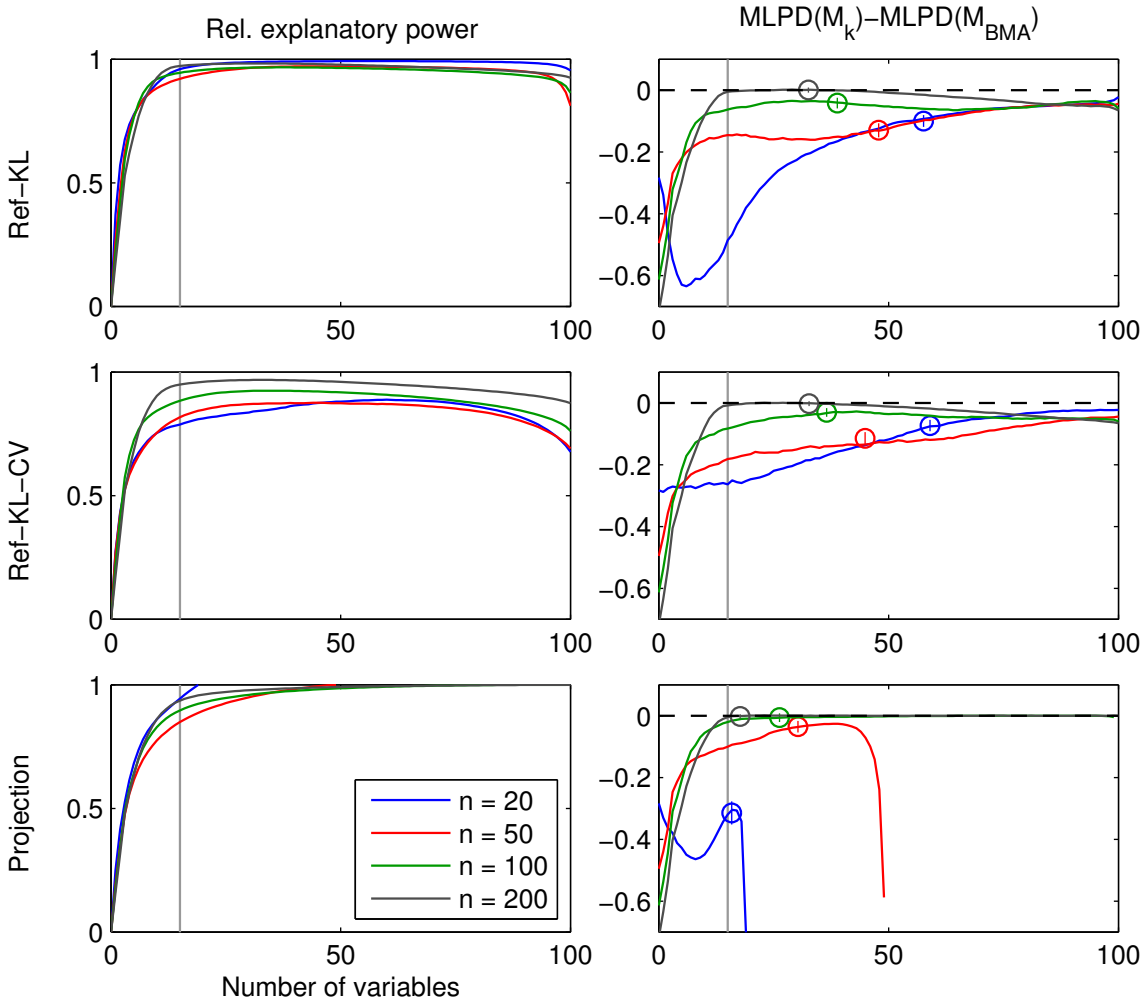


Figure 8: Simulated data: Estimated relative explanatory power (eq. (88)) of the submodels with respect to the BMA (left column), and corresponding predictive performances compared to the BMA (right column) along the forward search paths with different training set sizes $n = 20, 50, 100, 200$. The circles denote the average number of selected variables and obtained test utility with 95% credible interval for the actually selected submodels. For Projection, the selection rule was to choose the smallest model with 95% explanatory power, and model with maximum explanatory power for Ref-KL and Ref-KL-CV (the maximum explanatory power means the same as the minimum discrepancy). Total number of variables is $p = 100$, within group correlation $\rho = 0.5$ and the true number of variables is 15 (the grey line). The results are averaged over 100 data realizations.

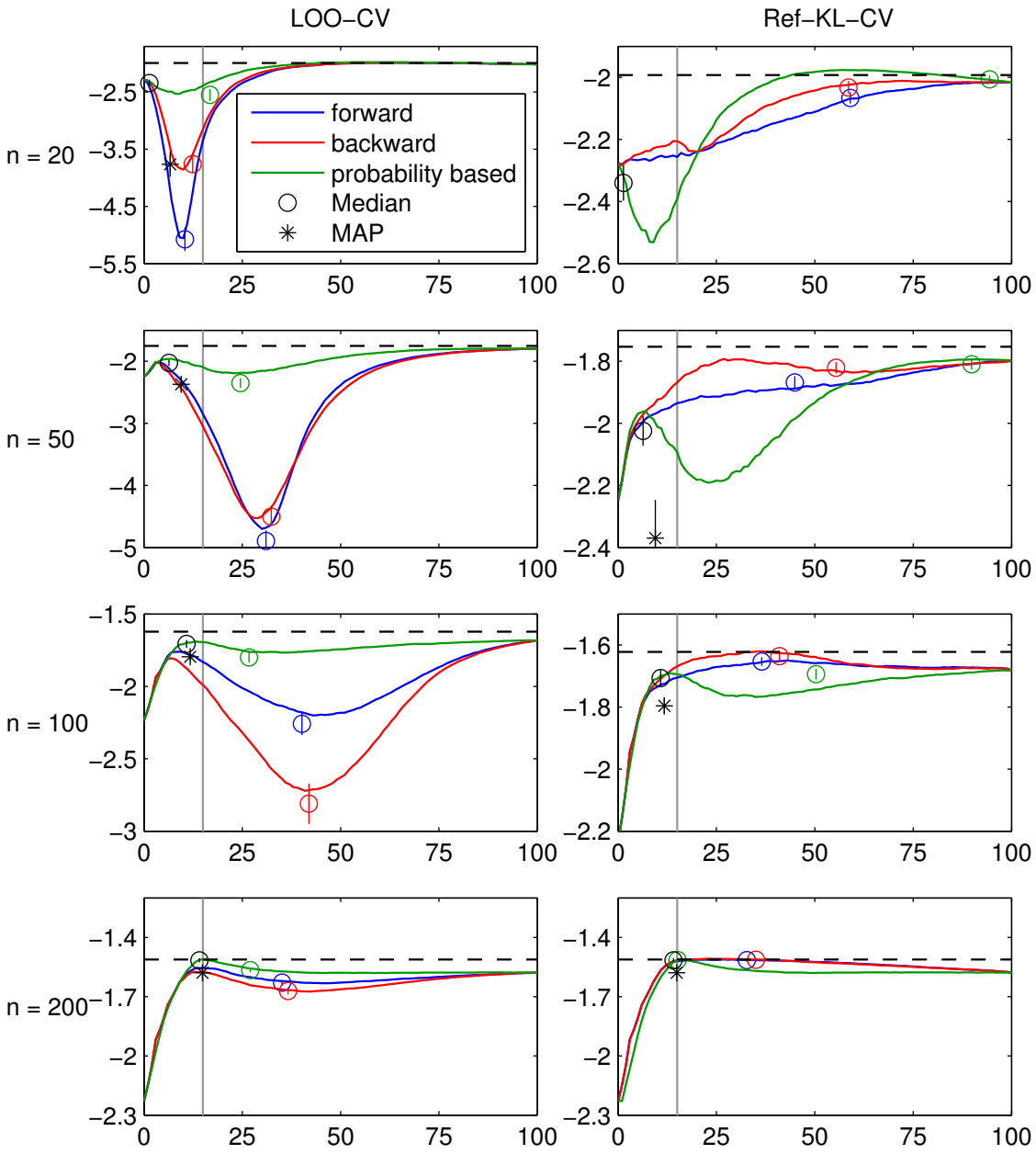


Figure 9: Simulated data: Comparison of different search strategies for LOO-CV (left column) and Ref-KL-CV (right column); the curves show the test MLPD for forward (blue) and backward searches (red), and for sorting the variables according to their probabilities (green). Note the different scales on the vertical axes, and that for each row the green curves are the same in both columns. The circles denote the average number of selected variables and obtained test utility with 95% credible interval for the actually selected models (LOO-CV-max for left and maximum explanatory power with respect to the BMA for right column). The average size and performance of the MAP and Median models are also shown for comparison. Total number of variables is $p = 100$, within group correlation $\rho = 0.5$ and true number of variables is 15 (grey line). The results are averaged over 100 data realizations.

5.2 Ovarian cancer data

The second example deals with binary classification. The dataset is the ovarian cancer microarray data by Schummer et al. (1999). The data covers measurements of 54 patients out of whom 30 had an ovarian tumour and 24 were healthy. Each patient is associated by an input vector of 1536 features derived from the patient's cDNA microarray. The dataset has been considered before at least by Li et al. (2002) and Hernández-Lobato et al. (2010) and can be found at http://www.dcs.gla.ac.uk/~srogers/lpd/lpd.html. The dataset did not require any further preprocessing (Li et al., 2002).

The applied model was the logistic regression model (39) with a relatively vague prior for the weights

$$p(\mathbf{w}) = \mathrm{N}\left(\mathbf{w} \mid 0, \tau^2\mathbf{I}\right)$$
$$\tau^2 = 10\,.$$

The intercept term was handled by adding a vector of ones to the original data matrix $\mathbf{X}$, and it was given the same prior as all the other weights. The reference model was again constructed as a Bayesian model average using RJMCMC. Given that the number of variables is much larger than the number of training instances, an informative prior was placed for the number of variables. The used prior was the Beta-binomial with parameters $(\alpha, \beta) = (1, 1200)$ which restricts the number of variables to less or equal to 10 with probability 99.8%. The prior and posterior probabilities for the number of variables are shown in figure 10. Also the marginal posterior probabilities for each variable are shown. One can see that under this prior, most likely the number of variables is around 1–4, but at least 6 of the variables have nonnegligible inclusion probability.

The variable selection was then carried out by performing forward search up to 15 variables using the same methods as for the simulated experiments in section 5.1 (listed in table 2), except that the methods based on the squared error (L-, L_q and L_{GG}-criteria) were left out as unsuitable for the classification model. To estimate the generalization performance of the selected models and the reference model, leave-one-out cross validation was used. In other words, the selection and reference model construction was repeated $n = 54$ times each time using 53 training observations and then testing the performance of the models with the point outside the data used for selection. The results obtained this way will be referred to as the out-of-sample estimates. To reduce the computational burden, Laplace approximation was used for the submodels (except for the projected submodels, which use the projected posterior samples of the BMA model). For the reference methods (Ref-KL, Ref-KL-CV, Projection), the selection rule was chosen to be the smallest model having 95% explanatory power. The sensitivity of the results with respect to this selection rule will be discussed later on. MAP and Median models were estimated from the RJMCMC-samples, and for the other methods (LOO-CV, WAIC, DIC) the model with the highest utility estimate was chosen.

The performance of the BMA, and the chosen submodels are listed in table 3. BMA obtains an estimated classification accuracy of about 94% which is quite sim-

ilar to the MCMC result by Hernández-Lobato et al. (2010), although they used the probit model and a slightly different prior. Roughly speaking, one can conclude that Ref-KL and Projection work well choosing models with predictive performance comparable to BMA. Ref-KL-CV and MAP perform the second best, and the other methods work somewhat poorer. Median model breaks down in this example, because in all except one datasplit none of the variables had marginal probability of 0.5 or more, and thus Median model is practically the same as the empty model (model having only the intercept term). According to these results it seems that Ref-KL-CV performs poorer than Ref-KL and Projection, but the results are highly dependent on the applied selection rule as will be discussed shortly.

To get more insight to the selection process, let us consider the forward search paths for each of the methods. Figures 11 and 12 show the results for MLPD and classification accuracies, respectively. The figures again demonstrate the effect of the selection induced bias. LOO-CV, WAIC and DIC identify variable combinations that fit very well to the training data in cross validation sense (dashed line) but have suboptimal predictive performance on independent data (dots). More specifically, these methods identify variable combinations that have cross validation classification accuracy close or equal to 100% within the data used for selection, but out-of-sample accuracy of about 80–90%. Note also the high variance in the estimated out-of-sample performance of the selected models, which is caused by high variability in the quality of the selected models in the different datasplits. Ref-KL, Ref-KL-CV and Projection seem to perform somewhat better especially in terms of the estimated out-of-sample MLPD. All these methods are able to find models of only 2 variables that have similar out-of-sample MLPD than the BMA. Note also that these methods choose models that have quite similar within and out-of-sample fit (the dashed line coincides with the dots) indicating that the chosen models are not overfitted. Ref-KL seems to find even a bit better models than the BMA in terms of the out-of-sample MLPD but also for these models the out-of-sample accuracies are slightly worse or equal to the BMA. It must be noted, however, that the point estimates for the classification accuracies may be somewhat misleading, because the dataset consists of only 54 points and thus even a single misclassification reduces the accuracy estimate by almost 2 percentage points. Furthermore, some of the points correctly classified by the BMA and misclassified by the submodels had class probabilities close to 50% describing the uncertainty in the classification. Thus in terms of the classification accuracy the differences between the BMA and the submodels may appear larger than they actually are.

Also this example demonstrates that the performance of the model actually selected might be sensitive to the selection rule. Consider the relative explanatory power curves for the reference methods plotted in figure 11. For instance, when searching variables using Ref-KL the relative explanatory power increases monotonically and reaches value close to one with 15 variables. However, the maximum out-of-sample MLPD is obtained around 3 variables after which the performance starts to decrease slowly. This indicates that more explanatory power (with respect to the reference model) does not necessarily imply better predictions. Also for Ref-KL-CV the submodels reach out-of-sample MLPD indistinguishable from the BMA

with only 2 variables, but for this model size the estimated explanatory power is clearly less than 1. These ideas are illustrated even better in figure 13. This figure shows that for Ref-KL and Projection, a good stopping rule for this data would be about 95% relative explanatory power with respect to the reference model, but for Ref-KL-CV the optimal selection rule would be somewhere between 80–85%. For Ref-KL-CV after this is reached, the predictive ability actually starts to decrease when more variables are added. For this reason, the results in table 3 are misleading when they might suggest that Ref-KL-CV chooses larger and worse models than Ref-KL and Projection; the optimal selection rules for the different reference methods are just different (at least for this dataset). Nevertheless, it is worth mentioning that any of the threshold values over 80% for any of the reference methods would lead to selecting a better model than the maximization of LOO-CV, WAIC or DIC.

To summarize, the maximization of the utility estimate based on LOO-CV, WAIC or DIC leads to suboptimal results as could be expected given the results in section 5.1. MAP model works better, but Median model fails. Even better results than MAP model are obtained with the reference methods, although also these are somewhat sensitive to the selection rule. In practice the suggested approach would be to perform the cross validation outside the selection process (as was done here), and decide the selection rule based on those results (like figures 11–13). It is relevant to point out that even though the actual selection decision would be based on the outer cross validation, the results here suggest that it is advisable to do the model searching using one of the reference methods (rather than LOO-CV, WAIC or DIC), because these seem to find better models with less variability in the predictive performance (figures 11–12). However, the results are not quite conclusive for ranking between the different reference methods, and the results here call for more research on this topic.

Model	MLPD	Classification %	# of variables on avg.
BMA	−0.245 (−0.337 ... −0.167)	94.4 (87.6 ... 98.8)	–
Ref-KL	−0.206 (−0.332 ... −0.120)	92.6 (84.9 ... 98.1)	3.15
Ref-KL-CV	−0.274 (−0.381 ... −0.181)	87.0 (77.9 ... 94.7)	6.70
Projection	−0.240 (−0.354 ... −0.158)	90.7 (82.4 ... 97.0)	2.18
MAP	−0.283 (−0.422 ... −0.168)	88.9 (79.9 ... 95.9)	1.00
Median	−0.696 (−0.725 ... −0.662)	57.4 (45.8 ... 71.4)	0.019
LOO-CV	−0.341 (−0.535 ... −0.186)	83.3 (73.5 ... 92.3)	4.13
WAIC	−0.420 (−0.655 ... −0.227)	81.5 (71.1 ... 91.0)	3.37
DIC	−0.377 (−0.614 ... −0.195)	83.3 (73.0 ... 92.2)	3.19

Table 3: Ovarian data: Estimated out-of-sample mean log predictive densities (MLPD) and classification accuracies for the BMA and selected submodels. For the selection methods also the average number of selected variables are shown. The values in the parenthesis give the 95% credible intervals estimated using the Bayesian bootstrap. The out-of-sample performances were estimated using leave-one-out cross validation (the selection and BMA construction was repeated $n = 54$ times and the performance was evaluated with the single point not in the training set).

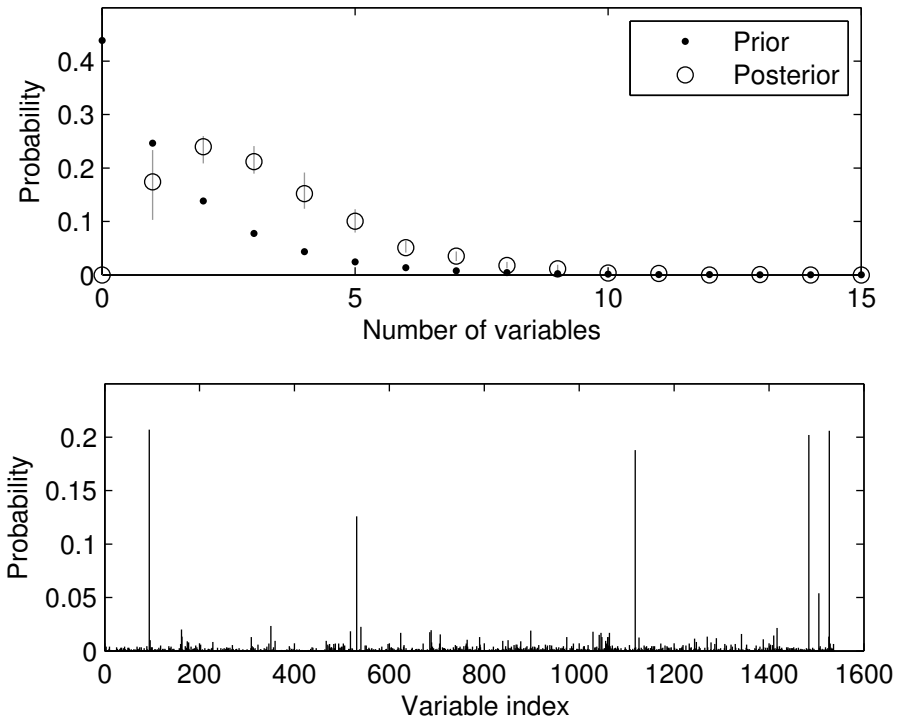


Figure 10: Ovarian data: Prior and posterior probabilities for different number of variables (top) and marginal posterior probabilities for each variable (bottom). Posterior probabilities for different number of variables are shown with 95%-credible intervals, which are estimated from variability between different RJMCMC-chains using Bayesian bootstrap. The results are calculated using the full dataset.

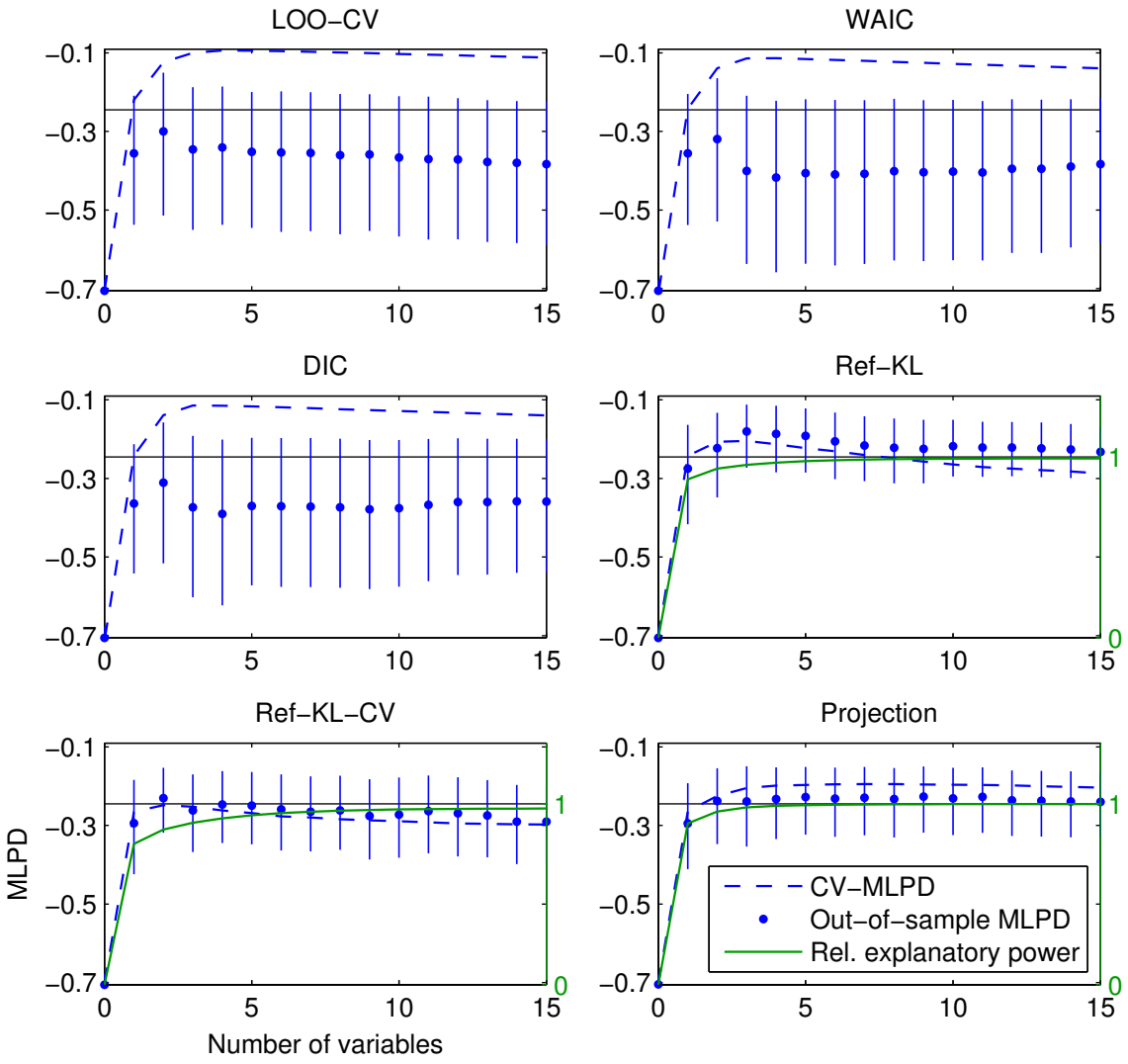


Figure 11: Ovarian data: Estimated mean log predictive densities (MLPD) for different number of variables on average along the forward search paths. The out-of-sample MLPDs were estimated using leave-one-out cross validation so that the search was repeated $n = 54$ times each time using 53 points and then measuring the predictive ability of the models with the point that was not used for selection. The results are given with 95% credible intervals calculated using Bayesian bootstrap. CV-MLPD denotes the average cross validation MLPD within those 53 points that were used for selection. The solid black denotes the out-of-sample performance of the BMA. For reference methods and projection also the average relative explanatory power (eq. (88)) for different number of variables over the $n = 54$ selections is shown on the right y-axis.

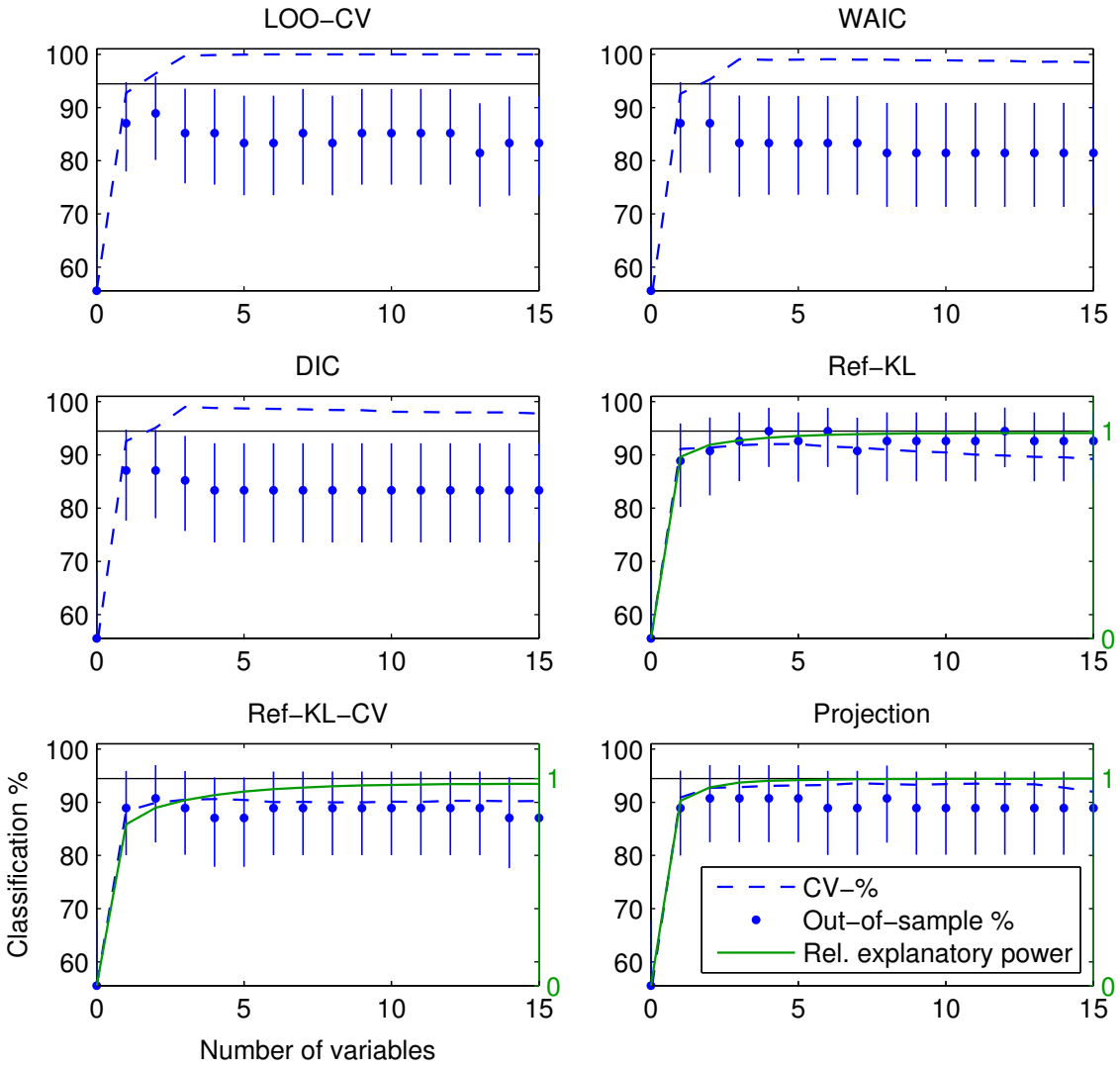


Figure 12: Ovarian data: The estimated CV and out-of-sample classification accuracies on average along the forward search paths calculated similarly as in figure 11. The solid black shows the estimated out-of-sample classification accuracy of the BMA. Green lines show the relative explanatory power for the reference methods.

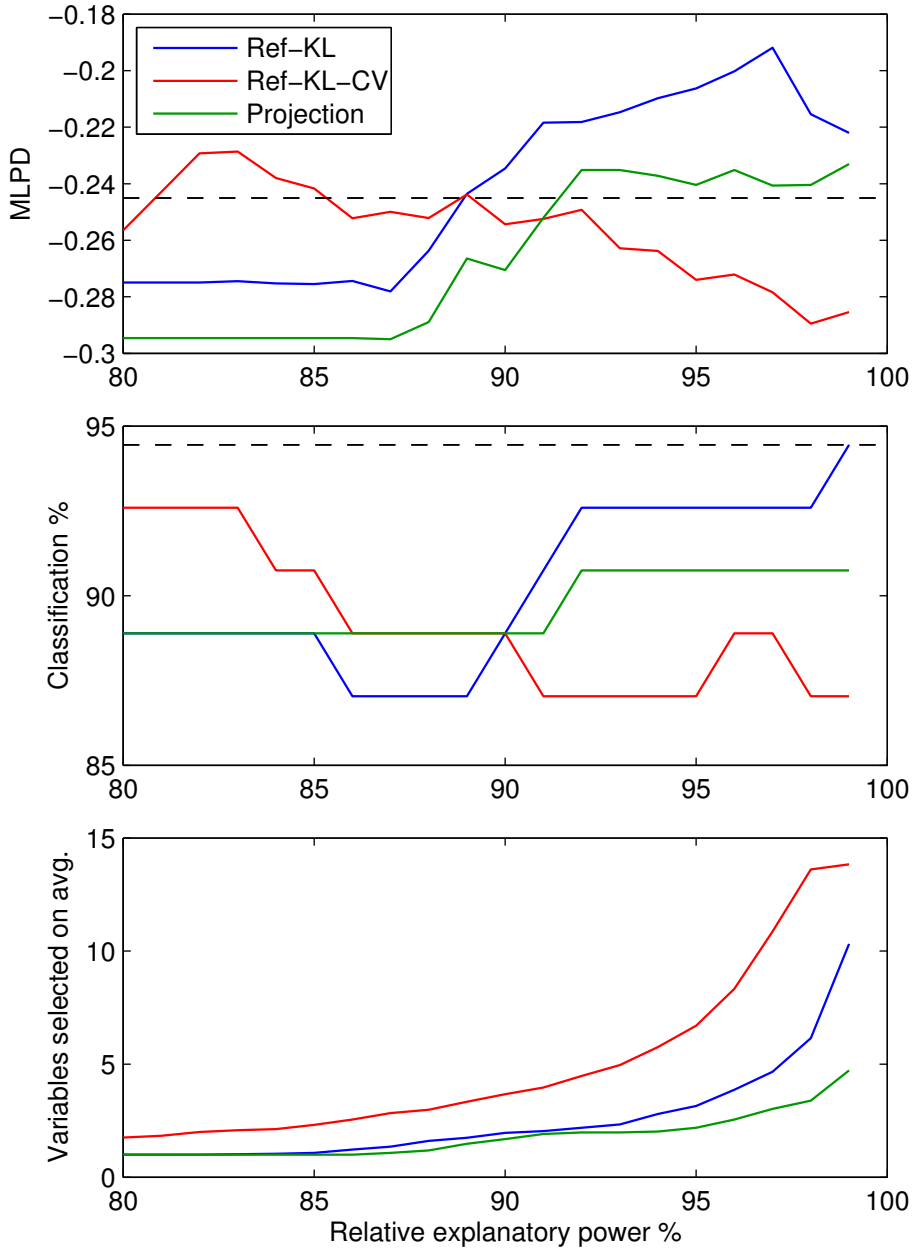


Figure 13: Ovarian data: Out-of-sample MLPD, classification accuracy and average number of selected variables for the reference methods as a function of the estimated explanatory power. In other words, if the selection rule was to choose the smallest model with the given explanatory power, the curves show what would be the corresponding average number of selected variables and estimates for the out-of-sample MLPD and classification accuracy. In the top and middle plots the dashed black line denotes the out-of-sample performance of the BMA.

6 Conclusions and discussion

This thesis has discussed Bayesian variable selection and model selection in general. The thesis has reviewed many of the proposed model selection methods in the wide literature and demonstrated their use in practical variable selection problems for linear regression and binary classification. The thesis has also discussed and illustrated in detail the selection induced bias, which is an important concept concerning the variable selection.

As discussed in section 3, there are methods that give an almost unbiased estimate of the generalization performance for a given model, but due to the selection induced bias, these methods may not necessarily work well for model selection. The selection induced bias means choosing a model with a nonoptimal generalization performance because its ability is overestimated just by chance. It is important to note that this may happen even if the utility estimate for a single model is completely unbiased. The probability of selecting a nonoptimal model increases when the number of considered models is large and the variance in the utility estimates is high. The variance in the utility estimates is different for different methods, but generally increases when the number of training observations becomes small.

The numerical experiments suggest that for small datasets the selection bias may hinder the variable selection considerably. This happens because the variance in the utility estimates is then high and the number of models under comparison becomes quickly very large even for a relatively small number of variables. At worst, it may happen that one is not only overestimating the ability of the chosen model but also finding a variable combination with no predictive ability at all. According to the experiments, it seems that the reference model based variable selection (reference predictive approach and projection) is less sensitive to the selection bias than the methods based on sample reuse, such as cross validation and information criteria. This is most likely due to smaller variance in the utility estimates. The experiments suggest also that assessing the marginal probabilities of the different variables and variable combinations (MAP and Median models) is unreliable in general although it may yield good results in some cases.

Although the reference predictive and projection approaches seem the most robust way of searching for good variable combinations, the estimated discrepancy between the reference model and a submodel is in general an unreliable indicator of the predictive performance of the submodel. The reason for this is again due to the selection bias; when the models are searched based on their estimated discrepancy from the reference model, this estimate is no longer an unbiased estimate of the true discrepancy for the models on the searchpath. Especially for small datasets, the estimate may underestimate the true discrepancy considerably. It is important to note that also the cross validation utility within the data used for selection seems to give a biased estimate of the true generalization performance even if the searching is done using the reference methods. For this reason, a more reliable method for assessing the performance of the models on the search path would be to use cross validation outside the selection process. This gives a nearly unbiased estimate of the true performance for the different number of selected variables, and the number

of included variables could be done using this estimate, although for the ultimately selected model the estimate would no longer be unbiased.

Finally, although the reference model approach appears the most promising approach to difficult variable selection problems, the results in this study are insufficient for making a clear distinction between the different reference approaches as these methods seemed to behave differently with respect to the applied selection rule. Therefore the results presented in this thesis call for more understanding about these methods and their behaviour, preferably also for nonlinear models such as Gaussian processes.

References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions Automatic Control*, AC-19:716–723. System identification and time-series analysis.

Barbieri, M. M. and Berger, J. O. (2004). Optimal Predictive Model Selection. *The Annals of Statistics*, 32(3):870–897.

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

Bishop, C. M. (2006). *Pattern recognition and Machine Learning*. Springer.

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

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Burman, P. (1989). A Comparative Study of Ordinary Cross-validation, v-fold Cross-validation and the Repeated Learning-Testing Methods. *Biometrika*, 76(3):503–514.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling Sparsity via the Horseshoe. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The Horseshoe Estimator for Sparse Signals. *Biometrika*, 97(2):73–80.

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.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2000). Bayesian Variable Selection Using the Gibbs Sampler. In *Generalized linear models: a Bayesian perspective*, pages 273–286.

Dupuis, J. A. and Robert, C. P. (2003). Variable Selection in Qualitative Models via an Entropic Explanatory Power. *Journal of Statistical Planning and Inference*, 111(1-2):77–94.

Geisser, S. and Eddy, W. F. (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model Determination Using Predictive Distributions with Implementation via Sampling-Based Methods. In *Bayesian Statistics 4*, pages 147–167.

Gelfand, A. E. and Ghosh, S. K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika*, 85(1):1–11.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013a). *Bayesian Data Analysis*. Chapman & Hall, Third edition.

Gelman, A., Hwang, J., and Vehtari, A. (2013b). Understanding Predictive Information Criteria for Bayesian Models. *Statistics and Computing*, pages 1–20.

George, E. I. and McCulloch, R. E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, 7:339–374.

Geweke, J. (1989). Bayesian Inference in Econometric Models using Monte Carlo Integration. *Econometrica*, 57(6):1317–1339.

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

Goutis, C. and Robert, C. P. (1998). Model Choice in Generalised Linear Models: a Bayesian Approach via Kullback-Leibler Projections. *Biometrika*, 85(1):29–37.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711–732.

Han, C. and Carlin, B. P. (2001). Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review. *Journal of the American Statistical Association*, 96(455):1122–1132.

Hernández-Lobato, D., Hernández-Lobato, J. M., and Dupont, P. (2013). Generalized Spike-and-Slab Priors for Bayesian Group Feature Selection Using Expectation Propagation. *Journal of Machine Learning Research*, 14:1891–1945.

Hernández-Lobato, D., Hernández-Lobato, J. M., and Suárez, A. (2010). Expectation Propagation for Microarray Data Classification. *Pattern Recognition Letters*, 31(12):1618–1626.

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

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

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

Kuo, L. and Mallick, B. (1998). Variable Selection for Regression Models. *Sankhya: The Indian Journal of Statistics*, 60:65–81.

Laud, P. W. and Ibrahim, J. G. (1995). Predictive Model Selection. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):247–262.

Li, Y., Campbell, C., and Tipping, M. (2002). Bayesian Automatic Relevance Determination Algorithms for Classifying Gene Expression Data. *Bioinformatics*, 18(10):1332–1339.

Marriott, J. M., Spencer, N. M., and Pettitt, A. N. (2001). A Bayesian Approach to Selecting Covariates for Prediction. *Scandinavian Journal of Statistics*, 28:87–97.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1036.

O'Hara, R. B. and Sillanpää, M., J. (2009). A Review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Analysis*, 4(1):85–118.

Orr, M. J. L. (1996). Introduction to Radial Basis Function Networks. Technical report, Centre for Cognitive Science, University of Edinburgh. Available at http://www.anc.ed.ac.uk/~mjo/papers/intro.ps.gz.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92(437):179–191.

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

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Reunanen, J. (2003). Overfitting in Making Comparisons Between Variable Selection Methods. *Journal of Machine Learning Research*, 3:1371–1382.

San Martini, A. and Spezzaferri, F. (1984). A Predictive Model Selection Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):296–303.

Schummer, M., Ng, W. V., Bumgarner, R. E., Nelson, P. S., Schummer, B., Bednarski, D. W., Hassell, L., Baldwin, R. L., Karlan, B. Y., and Hood, L. (1999). Comparative Hybridization of an Array of 21,500 Ovarian cDNAs for the Discovery of Genes Overexpressed in Ovarian Carcinomas. *Gene*, 238(2):375–385.

Shafer, G. (1982). Lindley's Paradox. *Journal of the American Statistical Association*, 77(378):325–334.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 623–656.

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 (Methodological)*, 64(4):583–639.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.

Sundararajan, S. and Keerthi, S. S. (2001). Predictive Approaches for Choosing Hyperparameters in Gaussian Processes. *Neural Computation*, 13(5):1103–1118.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244.

Titsias, M. K. and Lázaro-Gredilla, M. (2011). Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning. In *Advances in Neural Information Processing Systems 24*, pages 2339–2347.

Varma, S. and Simon, R. (2006). Bias in Error Estimation when using Cross-Validation for Model Selection. *BMC Bioinformatics*, 7(91). Available at http://www.ncbi.nlm.nih.gov/pubmed/16504092.

Vehtari, A. and Lampinen, J. (2002). Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities. *Neural Computation*, 14(10):2439–2468.

Vehtari, A. and Ojanen, J. (2012). A Survey of Bayesian Predictive Methods for Model Assessment, Selection and Comparison. *Statistics Surveys*, 6:142–228.

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

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.