Helsinki University of Technology, Laboratory of Computational Engineering publications. Report B. ISSN 1455-0474

Model Selection via Predictive Explanatory Power

Research report B38. ISBN 951-22-7190-7.

Aki Vehtari and Jouko Lampinen
Laboratory of Computational Engineering
Helsinki University of Technology
P.O.Box 9203, FIN-02015, HUT, Finland
{Aki.Vehtari,Jouko.Lampinen}@hut.fi

July 10, 2004

Abstract

We consider model selection as a decision problem from a predictive perspective. The optimal Bayesian way of handling model uncertainty is to integrate over model space. Model selection can then be seen as point estimation in the model space. We propose a model selection method based on Kullback-Leibler divergence from the predictive distribution of the full model to the predictive distributions of the submodels. The loss of predictive explanatory power is defined as the expectation of this predictive discrepancy. The goal is to find the simplest submodel which has a similar predictive distribution as the full model, that is, the simplest submodel whose loss of explanatory power is acceptable. To compute the expected predictive discrepancy between complex models, for which analytical solutions do not exist, we propose to use predictive distributions obtained via k-fold cross-validation. We compare the performance of the method to posterior probabilities (Bayes factors), deviance information criteria (DIC) and direct maximization of the expected utility via cross-validation.

Keywords: Bayesian model choice; covariate selection; decision theory; expected utility; cross-validation; DIC

1 Introduction

In this paper we follow the decision theoretic ideas for model selection outlined by Bernardo and Smith (1994, Ch. 6). We believe that none of our models is the "true" model, but we assume that we have been able to construct a model which is useful for describing our assumptions about the problem. For simplicity, we mainly discuss this problem from the covariate selection point of view, but the approach can also be used in non-nested model selection.

In true Bayesian spirit, we should use all observations we believe might have relevance for the predictions, that is, in classification and prediction problems we should include all observed covariates in our model and describe the related uncertainty about the relevance of the covariates with the model structure and prior distributions. Sometimes it may be difficult to come up with good model structures and priors for very high dimensional problems, and instead of thinking harder it may be easier to reduce the number of covariates beforehand. Here we assume that we have been able to construct the full model M_F, which we think gives the best predictions given the data and our prior beliefs. In the case of alternative non-nested models, we take the model uncertainty into account by integrating over the model space, that is, using Bayesian model averaging.

In order to check that the full model is adequate for practical use, the estimate of the expected utility can be compared to expert knowledge. For predictive purposes we could just use the full model, but we might have other reasons for making covariate selection. In addition to being interested in predictive performance of the model, we may be interested in the cost of observing the covariates. We might even be ready to sacrifice some of the predictive performance in order to reduce this cost. Note that computation time can be considered as a part of the cost of the covariates. In such cases, we need to use a utility that maps to same scale as the total cost of covariates. However, model construction is often made at a stage where either the future costs of covariates are yet unknown, the distribution of the future data is unknown, or the purpose of the model is to analyse some phenomenon and thus interest lies in the goodness of the predictive distribution. The goodness of the predictive distribution is best measured with logarithmic score function (Bernardo & Smith, 1994), which is incompatible with, for example, monetary costs of covariates.

Here we assume that we are not able to assign useful costs to the covariates and we do not want to sacrifice almost any predictive performance. We are interested in identifying all covariates which, based on the current data, do not contribute to the predictive distribution. These covariates are irrelevant or there is not enough information about them in the data. Identification of these covariates is useful for the application experts. Furthermore, a reduced model is easier to analyse by the application experts and thus it is easier to gain scientific insights to the phenomenon modeled.

In case of non-nested models, the full model can be constructed using Bayesian model averaging, that is, by integrating over the model space. Bayesian model averaging has also been used for nested models by averaging over all possible covariate combinations, but we consider this a single model for which there is a non-zero prior probability that the effect of a covariate can be exactly zero.

Selecting a single model corresponds to selecting a point estimate in the model space. We propose to use the Kullback-Leibler divergence from the predictive distribution of the full model $p(\cdot|D, M_F)$ to the predictive distribution of a submodel $p(\cdot|D, M_j)$. The goal is to find the simplest submodel which has a similar predictive distribution as the full model based on the expectation of this predictive discrepancy. To properly take into account the assumptions of the future data distribution in prediction problems and to compute the expected predictive discrepancy between complex models for which analytical solutions do not exist, we propose to use predictive distributions obtained via k-fold cross-validation.

The proposed method is related to the method of Dupuis and Robert (2003). They proposed using the expected Kullback-Leibler divergence from the full model $p(\cdot|\theta_F, M_F)$ to the submodel $p(\cdot|\theta_j^{\perp}, M_j)$, where the parameters $\theta_j^{\perp}$ of submodels M_j are defined as the Kullback-Leibler projections of the full model parameters and the expectation is taken over the posterior distribution $p(\theta_F|D, M_F)$. Note that the distribution of the projected parameters $\theta_j^{\perp}$ is not same as the posterior distribution $p(\theta_j|D, M_j)$. Dupuis and Robert (2003) introduced the notion of loss of explanatory power defined as the expected discrepancy, and the notion of relative explanatory power by scaling the maximal acceptable expected discrepancy in terms of the explanatory power of the covariate free model and the full model. We adopt similar notions, but since we are dealing with predictive distributions, we introduce the notions of loss of predictive explanatory power and relative predictive explanatory power.

Bayesian hypothesis testing and Bayesian Reference Criterion (BRC) by Bernardo (1999) are essentially same as the method of Dupuis and Robert (2003). Bernardo (1999) emphasized the use of reference priors, argued that the expected discrepancy is in the natural scale of the evidence in nits, and also discussed and illustrated the frequency properties of the method. Bernardo and Rueda (2002), and Bernardo and Juáres (2003) replaced the directed Kullback-Leibler divergence $k(\theta_i|\theta_F)$ with the symmetric divergence $\delta(\theta_i, \theta_F) = \min\{k(\theta_i|\theta_F), k(\theta_F|\theta_i)\}$. The relations and differences of our predictive approach to methods of Bernardo (1999), Bernardo and Rueda (2002), Bernardo and Juáres (2003), and Dupuis and Robert (2003) will be analysed in more detail elsewhere.

Lindley (1968) developed a method for Gaussian linear models, in which the quadratic distance between mean predictions of the full model and the submodels is used and thus computations can be made analytically. This approach does not take into account the width of the predictive distribution. Lindley did not use any measure of how large a distance is acceptable, and he selected the cost term *ad hoc* since there is no general way of scaling the cost to match the squared loss used for predictions. This approach has been used successfully, for example, by Brown et al. (1999) who extended this method to multivariate responses and non-conjugate priors, and by Brown et al. (2002) who extended this method to covariate selection with Bayesian model averaging.

Instead of comparing predictions of the submodels to the full model, it is possible to estimate the expected utility of the submodel directly, for example, by using partial-validation, cross-validation, or the generalized deviance information criterion (e.g. Geisser, 1975; Geisser & Eddy, 1979; Gelfand & Dey, 1994; Bernardo & Smith, 1994; Gelfand, 1996; Draper & Fouskakis, 2000; Spiegelhalter et al., 2002; Vehtari & Lampinen, 2002, 2003). Although the methods used for estimating the expected utilities of the models are usually almost unbiased, selection induced bias is a problem when comparing a large number of models. Selection induced bias is caused by selecting a model, which has the best estimated expected utility. This bias increases when the number of models increases. In section 3.1 we demonstrate how this may lead to inferior model selection. We use cross-validation and deviance information criterion as examples. The expected utility estimate of the full model is unbiased (given no pre-selection), and thus it is sensible to search for submodels which give similar predictions as the full model. The unbiased estimate of the predictive performance of the the full model works then as a safe limit for estimate of the performance of a submodel.

Whether one uses direct estimation of the expected utility of the submodel, or the divergence from the full model to the submodel, there is often a problem of having very large number of models, which may lead to the search space being huge. In this paper, we do not consider different search strategies although we note that many have been proposed (e.g., Draper & Fouskakis, 2000; Fearn et al., 2002; Dupuis & Robert, 2003; Vannucci et al., 2003). Similar to the method of Dupuis and Robert (2003) we should be able to eliminate at least some of the farthest branches of the search tree by the rejection of one of their ancestors. For greatly reduced search, we constrain our search space by first ordering the covariates according to their marginal posterior probabilities. This is not optimal, but given the long computation time needed for the complex models we are usually using, this approach is a compromise between accuracy and computation time. We also compare this approach to more time consuming forward selection in a simple toy problem

Marginal likelihoods and posterior probabilities of input combinations have also been used directly for input selection (e.g., Brown et al., 1998; Ntzoufras, 1999; Han & Carlin, 2000; Sykacek, 2000; Kohn et al., 2001; Chipman et al., 2001; Barbieri & Berger, 2002). Although this kind of approach has produced useful results, it may be sensitive to prior choices (e.g., Lindley, 1957; Jeffreys, 1961; Kass & Raftery, 1995; Richardson & Green, 1997; Dellaportas & Forster, 1999; Ntzoufras, 1999; Chipman et al., 2001; Fernández et al., 2001; Kohn et al., 2001), and does not necessarily provide the model with the best expected utility as demonstrated in section 3. We also illustrate that the median probability model method of Barbieri and Berger (2002) does not work well in the case of dependent covariates.

In the next section we present the proposed method and briefly review related methods. In section 3 we illustrate the discussion with simulation results and one real world example.

2 Methods

We first present a summary of the proposed method (section 2.1). We consider the decision theoretic approach for assessing the predictive performance of a single model (section 2.2). We discuss related approximations (sections 2.3 and 2.4) and reasons why the direct utilization of these methods does not work optimally in model selection (section 2.5). Finally we describe the proposed method in detail (section 2.6).

2.1 Summary of the proposed method

1. Construct the full model $p(\cdot|\theta, M_F)$ and the prior $p(\theta|M_F)$. In the case of non-nested models, the full model is the Bayesian model average of all models.

2. Estimate the predictive performance of the full model by computing the estimate of the expected utility $\bar{u}$. Compare the result to expert knowledge to assess the practical usefulness of the model.

3. Compute the expected Kullback-Leibler divergence from the predictive distribution of the full model $p(\cdot|D, M_F)$ to the predictive distributions of the submodels $p(\cdot|D, M_j)$. Select the simplest model, whose expected predictive discrepancy in terms of relative loss of predictive explanatory power is not too large.

The proposed method corresponds to *Occam's rule* in that it selects the simplest submodel which provides predictive inference similar to the full model.

2.2 Expected utilities

We consider prediction problems with explanatory variables or covariates x and a target variable y. The posterior predictive distribution of output $y^{(n+1)}$ for the new input $x^{(n+1)}$ given the training data $D = \{(x^{(i)}, y^{(i)}); i = 1, 2, \ldots, n\}$ is obtained by

$$p(y^{(n+1)}|x^{(n+1)}, D, M) = \int p(y^{(n+1)}|x^{(n+1)}, \theta, D, M)p(\theta|x^{(n+1)}, D, M)d\theta, \quad (1)$$

where θ denotes all the model parameters and hyperparameters of the prior structures and M is all the prior knowledge in the model specification. We assume that knowing $x^{(n+1)}$ does not give more information about θ, that is, $p(\theta|x^{(n+1)}, D, M) = p(\theta|D, M)$.

It is natural to assess the predictive ability of the model by estimating the expected utilities, that is, the relative values of consequences of using the model (Bernardo & Smith, 1994). For a single model, we would like to estimate how good our model is by estimating the quality of the predictions the model makes for future observations from the same (unknown) process that generated the given set of training data D. Given a utility function u, the expected utility is obtained by taking the expectation

$$\bar{u} = \mathrm{E}_{(x^{(n+1)}, y^{(n+1)})} \left[u \left(y^{(n+1)}, x^{(n+1)}, D, M \right) \right]. \tag{2}$$

The expectation could also be replaced by some other summary quantity, such as the α-quantile. Note that considering the expected utility for the next sample is equivalent to taking the expectation over all future samples. Preferably, the utility u would be application-specific, measuring the expected benefit or cost of using the model. For example, Draper and Fouskakis (2000) discuss an example in which monetary utility is used for data collection costs and the accuracy of predicting mortality rate in health policy problem. In lack of application specific utilities, many general discrepancy and likelihood utilities can be used. Especially useful generic utility is the predictive likelihood

$$u = p(y^{(n+1)} | x^{(n+1)}, D, M), \tag{3}$$

which measures how well the model estimates the whole predictive distribution, and thus is especially useful in model comparison. It is also useful in non-prediction problems, in which the goal is to get scientific insight in modeled phenomena. For further discussion of the logarithmic scoring rule see reference (Bernardo & Smith, 1994) and for discussion of using logarithmic scoring rule in decision theoretic model selection see reference (Bernardo, 1999).

2.3 Cross-Validation Predictive Densities

Expected utilities can be estimated using cross-validation (CV) predictive densities. As the distribution of $(x^{(n+1)}, y^{(n+1)})$ in (2) is unknown, we approximate it by using the samples we already have, that is, we assume that the distribution can be reasonably well approximated using the (weighted) training data $\{(x^{(i)}, y^{(i)}); i = 1, 2, \ldots, n\}$. To simulate the fact that the future observations are not in the training data, the ith observation $(x^{(i)}, y^{(i)})$ in the training data is left out, and then the predictive distribution for $y^{(i)}$ is computed with a model that is fitted to all of the observations except $(x^{(i)}, y^{(i)})$. By repeating this for every point in the training data, we get a collection of leave-one-out cross-validation (LOO-CV) predictive densities

$$\{p(y^{(i)} | x^{(i)}, D^{(\setminus i)}, M); i = 1, 2, \ldots, n\}, \tag{4}$$

where $D^{(\setminus i)}$ denotes all the elements of D except $(x^{(i)}, y^{(i)})$. To get the expected utility estimate, these predictive densities are compared to the actual $y^{(i)}$'s using the utility u, and the expectation in (2) is approximated with

$$\bar{u}_{\mathrm{LOO}} = \frac{1}{n} \sum_{i=1}^{n} \left[u(y^{(i)}, x^{(i)}, D^{(\setminus i)}, M) \right]. \tag{5}$$

If the future distribution is expected to be different from the distribution of the training data, the observations could be weighted appropriately. By appropriate modifications of the algorithm, the cross-validation predictive densities can also be computed for data with a nested structure or other finite range dependencies. Vehtari and Lampinen (2002) discuss these issues and assumptions made on future data distributions in more detail.

For simple models, the LOO-CV-predictive densities may be computed quickly using analytical (approximative) solutions, but models that are more complex usually require a full model fitting for each of the n predictive densities. When using Monte Carlo methods, we have to sample from $p(\theta|D^{(\backslash i)}, M)$ for each i. If sampling is slow (e.g., when using MCMC methods), importance sampling LOO-CV or the k-fold-CV can be used to reduce the computational burden (Vehtari & Lampinen, 2002).

In k-fold-CV, we sample only from k (e.g., $k = 10$) k-fold-CV distributions $p(\theta|D^{(\backslash s(i))}, M)$ and get a collection of k-fold-CV predictive densities

$$\{p(y^{(i)}|x^{(i)}, D^{(\backslash s(i))}, M); i = 1, 2, \ldots, n\}, \tag{6}$$

where $s(i)$ is a set of data points as follows: the data are divided into k groups so that their sizes are as nearly equal as possible and $s(i)$ is the set of data points in the group where the ith data point belongs. In the case of data with nested structure, the grouping needs to respect the hierarchical nature of the data, and in the case of non-structured finite range dependency, the group size should be selected according to the range of the dependency.

Since the k-fold-CV predictive densities are based on smaller training data sets $D^{(\backslash s(i))}$ with n_c observations instead of n, the expected utility estimate is slightly biased. The amount of this bias depends on the shape of the learning curve, that is, how much predictions improve given n observations instead of n_c observations. This bias can be corrected using a first order bias correction (Burman, 1989)

$$\begin{aligned}
\bar{u}_{\mathrm{tr}} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}[u(y^{(i)}, x^{(i)}, D, M)] && (7)\\
\bar{u}_{\mathrm{cvtr}} &= \tfrac{1}{k}\textstyle\sum_{j=1}^{k}\left[\tfrac{1}{n}\sum_{i=1}^{n}[u(y^{(i)}, x^{(i)}, D^{(\backslash s_j)}, M)]\right] \quad ; \quad j = 1, \ldots, k && (8)\\
\bar{u}_{\mathrm{CCV}} &= \bar{u}_{\mathrm{CV}} + \bar{u}_{\mathrm{tr}} - \bar{u}_{\mathrm{cvtr}}, && (9)
\end{aligned}$$

where $\bar{u}_{\mathrm{tr}}$ is the expected utility evaluated with the training data given the training data, that is, the training error or the expected utility computed with the marginal posterior predictive densities, and $\bar{u}_{\mathrm{cvtr}}$ is the average of the expected utilities evaluated with the training data given the k-fold-CV training sets.

Instead of just making a point estimate, it is important to obtain the distribution of the expected utility estimate in order to describe the associated uncertainty. We use a quick and generic approach based on the Bayesian bootstrap for obtaining samples from the distributions of the expected utility estimates (Rubin, 1981; Vehtari & Lampinen, 2002). Bayesian bootstrap makes a non-parametric Dirichlet distribution approximation of the distribution of a random variable. This approach can handle arbitrary summary quantities and gives a useful approximation also in non-Gaussian cases, unless extreme tail-areas are considered.

2.4 Other approaches

2.4.1 Partial Predictive Densities

Sometimes (4) is replaced with collection of partial predictive densities

$$\{p(y^{(i)}|x^{(i)}, D^{(1,\ldots,n_p)}, M); i = n_p + 1, \ldots, n\}, \tag{10}$$

where usually $n_p \geq n/2$. This approach is very conservative unless the learning curve of the model is essentially flat between n_p and n, that is the posterior distribution given $D^{(1,\ldots,n_p)}$ is similar to the posterior distribution given D. In the case of complex models with small or moderate size data set this is not true, and this approach would favor simpler models than necessary. Additionally, in this approach the distribution of the $(x^{(n+1)}, y^{(n+1)})$ is approximated using only $n - n_p$ data points, which reduces its accuracy. However, due to being conservative, this approach is much safer than marginal posterior predictive densities, for example.

2.4.2 Marginal Posterior Predictive Densities

Sometimes (4) is replaced with a collection of marginal posterior predictive densities

$$\{p(y^{(i)}|x^{(i)}, D_*, M); i = n_p + 1, \dots, n\}, \tag{11}$$

where D_* is an exact replicate of D, that is, the predictive density is conditioned twice on $x^{(i)}$. It is well known that this approach underestimates the generalization error of flexible models as it does not correctly measure the out-of-sample performance and thus favors too complex overfitted models. Only if the effective number of parameters is relatively small (i.e., if $p_{\text{eff}} \ll n$), the marginal posterior predictive densities may be useful approximations to cross-validation predictive densities, and may be used to save computational resources. In the case of flexible non-linear models p_{eff} is usually relatively large.

2.4.3 Deviance information criterion

The deviance information criterion (DIC) (Spiegelhalter et al., 2002) can also be used to estimate expected utility (2). As the name indicates, DIC was originally defined using deviance as utility. The approximation used in DIC can be generalized to arbitrary utilities (Vehtari, 2002; Vehtari & Lampinen, 2003). Given a utility function u, it is possible to use Monte Carlo samples to estimate $\mathrm{E}_\theta[\bar{u}(\theta)]$ and $\bar{u}(\mathrm{E}_\theta[\theta])$, and then compute an expected utility estimate as

$$\bar{u}_{\text{DIC}} = \bar{u}(\mathrm{E}_\theta[\theta]) + 2\left(\mathrm{E}_\theta[\bar{u}(\theta)] - \bar{u}(\mathrm{E}_\theta[\theta])\right). \tag{12}$$

Vehtari (2002), and Vehtari and Lampinen (2003) argue that cross-validation has several advantages over DIC. However, both DIC and cross-validation suffer from selection induced bias if used directly for model selection as discussed in section 2.5 and illustrated in section 3.1.

2.5 Model selection via expected utility estimates

For model selection, it has been shown that cross-validation is inconsistent in the sense that the probability of selecting the model with best predictive ability does not converge to one as the total number of observations $n \to \infty$, unless the division of the data satisfies $n_v/n \to 1$ and $n_c \to \infty$ as $n \to \infty$, where n_v is the number of samples left out for each fold and $n_c = n - n_v$ (Shao, 1993). For a single finite case, however, this does not have any relevance since any division fulfills these conditions.

In the finite case, there are uncertainties in expected utility estimation and comparison and thus variance in model selection. Part of the variance can be reduced by increasing the number of the different data divisions in k-fold-CV or by using Monte Carlo CV with a large number of repeats, but the improvement is usually small since the uncertainty due to not knowing the future data distribution dominates this variance (Vehtari & Lampinen, 2002).

There is a common belief that cross-validation tends to select an unnecessarily large model (see, e.g., Shao, 1993). However, since cross-validation is unbiased, this is not caused by the inherent properties of the cross-validation. The tendency to select an unnecessarily large model depends on the differences between the true expected utilities of the models and the variance in the estimates. Consider covariate selection with the possible models divided to three categories:

1. At least one of the relevant covariates is not included.

2. All and only relevant covariates are included (optimal choice).

3. All relevant and some irrelevant covariates are included.

The selection of an unnecessarily large model is more probable if predictive performance declines more when one of the relevant covariates is not included than when one of the irrelevant covariates is included. With sensible priors it is often possible to have many irrelevant covariates included in the model without great loss of predictive performance and thus, due to the variance in the expected utility estimates, it is probable that an unnecessarily large model is selected. Selection of a smaller than optimal model may happen, for example, if the relevant covariates are highly dependent, and thus removing one of the relevant covariates does not significantly reduce the predictive performance (see example in section 3.1).

To summarize, if we select a model based on the estimates, we may select model which has slightly worse true predictive performance as optimal model, but just by chance it is estimated to be better due to the variance. When the number of considered models increases, the probability of selecting a worse model increases. Due to the selection process, the estimate for the selected model is not unbiased anymore; it is now too optimistic and the selected model is overfitted to the data. This is especially a problem when selecting a model from a large number of models having similar predictive performances.

Note that the Bayes factor suffers from this same problem in the finite data case. The evidence term divided by n can be written as $\frac{1}{n} \log p(D|M) = \frac{1}{n} \sum_i \log p(y^{(i)}|y^{(s_i)}, M)$, where $y^{(s_i)}$ is a set of data points so that $y^{(s_1)} = \emptyset$, $y^{(s_2)} = y^{(1)}$, and $y^{(s_i)} = y^{(1,\ldots,i-1)}$; $i = 3, \ldots, n$. In an expected utility sense this is an average of predictive log-likelihoods with the number of data points used for fitting ranging from 0 to $n-1$. As there are terms which are conditioned on none or very few data points, the Bayes factor is sensitive to prior changes. With more vague priors and more flexible models the few first terms dominate the expectation unless n is very large.

An almost unbiased estimate of the expected utility of the selected model could be obtained via two level cross-validation, in which the additional cross-validation level estimates the effect of the model selection. This would not affect which models are finally selected, but would require k^2 more time than sampling from the full posterior distribution when using k-fold-CV. Furthermore, this would not prevent suboptimal model selection.

For a few models the selection induced bias is smaller, and thus, the distributions of the expected utility estimates can be used to compare models, for example, by computing the probability of one model having a better expected utility than some other model (Vehtari & Lampinen, 2002).

2.6 Model selection via predictive explanatory power

Selecting a submodel can be seen as point estimation in the model space. If the full model describes our knowledge of the problem and related uncertainties in the best possible way, the predictive distribution using the full model describes our beliefs on the future observations in the best possible way. Thus, it is natural to select a submodel M_j for which the predictive distribution $p(y^{(n+1)}|x^{(n+1)}, D, M_j)$ is the most similar to the predictive distribution of the full model $p(y^{(n+1)}|x^{(n+1)}, D, M_F)$.

A natural information and decision theoretic choice for measuring the discrepancy between distributions is Kullback-Leibler divergence (e.g. Bernardo, 1979; Bernardo & Smith, 1994; Robert, 1996; Bernardo, 1999)

$$k(g|f) = \int f(z) \log \left(\frac{f(z)}{g(z)} \right) dz. \tag{13}$$

Since we are comparing predictive distributions, we define the *predictive discrepancy* from M_F to M_j as

$$\delta(M_j|M_F) = \int p(y|x^{(n+1)}, D, M_F) \log \left(\frac{p(y|x^{(n+1)}, D, M_F)}{p(y|x^{(n+1)}, D, M_j)} \right) dy. \tag{14}$$

The *expectation of the predictive discrepancy* or the *loss of predictive explanatory power* is defined as

$$d(M_j|M_F) = \mathrm{E}_{x^{(n+1)}}\,\delta(M_j|M_F), \tag{15}$$

which we approximate by using cross-validation (compare to (6))

$$d(M_j|M_F) \approx \frac{1}{n}\sum_{i=1}^{n}\left[\int p(y|x^{(i)}, D^{(\backslash s(i))}, M_F)\log\left(\frac{p(y|x^{(i)}, D^{(\backslash s(i))}, M_F)}{p(y|x^{(i)}, D^{(\backslash s(i))}, M_j)}\right)dy\right]. \tag{16}$$

In the case of complex models, it is usually not possible to compute the integral in (14) analytically, and we approximate further

$$d(M_j|M_F) \approx \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m}\sum_{l=1}^{L}\log\left(\frac{p(\dot{y}^{(l)}|x^{(i)}, D^{(\backslash s(i))}, M_F)}{p(\dot{y}^{(l)}|x^{(i)}, D^{(\backslash s(i))}, M_j)}\right)\right], \tag{17}$$

where $\dot{y}^{(l)}$ are samples from the predictive distribution $p(y|x^{(i)}, D^{(\backslash s(i))}, M_F)$, obtained using Monte Carlo methods. This later approximation induces a small constant to the discrepancy estimate. Due to Monte Carlo uncertainty, even if we compare an exact replicate of the full model M_{F_*} to the full model M_F, the expected discrepancy estimate in (17) is not exactly zero. The estimate of $d(M_{F_*}|M_F)$ can be used to estimate the smallest observable discrepancy due to the Monte Carlo uncertainty.

Although predictive densities are not compared to the actually observed $y^{(i)}$, we still need to use cross-validation to simulate the fact that $x^{(n+1)}$ can have values different from the training data. Similar to the estimation of expected utilities, in the case of simple models, we could replace the cross-validation predictive densities with the marginal predictive densities (see 2.4.2). In theory, we could also make a density estimate (e.g. kernel density estimate) of $p(x^{(n+1)})$ using $p(x^i)$, and then use samples from that estimate to estimate the expectation in (15). In practice, however, it is very difficult to make a good density estimate if x has a large number of dimensions, and thus we prefer the non-parametric cross-validation approach.

The expected predictive discrepancy corresponds to the amount of information lost in the description of predictive densities when replacing the full model M_F with submodel M_j, and thus it is natural to use notion of loss of predictive explanatory power. Bernardo (1999), and Bernardo and Rueda (2002) argued that the expected discrepancy from $p(\cdot|\theta_F, M_F)$ to the submodel $p(\cdot|\theta_i^{\perp}, M_j)$ corresponds to the expected log-likelihood ratio against the sub-model, and thus conventionally used values of about 2.5 or 5.0 should respectively be regarded as mild and strong evidence against the hypothesis $\theta = \theta_0$. In hypothesis testing, the goal is to reject an alternative only if there is strong evidence against it. In model selection for predictive purposes, the goal is to select a model which is very similar to the full model, but values of about 2.5 or 5.0 would lead to the selection of models quite different to the full model. Thus we need an alternative calibration method for deciding how small a difference is acceptable.

Dupuis and Robert (2003) proposed to scale the expected discrepancy from $p(\cdot|\theta_F, M_F)$ to submodel $p(\cdot|\theta_i^{\perp}, M_j)$ by scaling the discrepancies with the expected discrepancy to the covariate free model M_0. They denote the expected discrepancy from the model M_0 the *explanatory power* and the scaled discrepancy from the model M_0 the *relative explanatory power*. We use the same idea in problems in which M_0 is sensible, for example, in covariate selection, by defining the *relative loss of predictive explanatory power*

$$d_r(M_j|M_F) = d(M_j|M_F)/d(M_0|M_F), \tag{18}$$

and the *relative predictive explanatory power*

$$d_r(M_j|M_F) = 1 - d(M_j|M_F)/d(M_0|M_F). \tag{19}$$

Naturally, we still have the problem of deciding how much predictive explanatory power we are ready to sacrifice. In a toy example we used a cut-off at 1% loss of predictive explanatory power, and in real world problem, we decided the cut-off based on an elbow in plotted d_r values (see section 3).

To summarize, we propose to select the simplest submodel for which the relative predictive explanatory power is not less than, for example, 99%, or in other words, the loss of relative predictive explanatory power is less than 1%. Since the selection is based on finding a model with a similar predictive distribution to the predictive distribution of the full model, there is no mechanism in the selection process which would cause overfitting to the data.

3 Illustrative examples

3.1 A toy example

To illustrate the frequency properties of the proposed method and reference methods in the finite data case, we consider a simple linear Gaussian problem with high dependencies in covariates and some irrelevant covariates. The data is distributed as follows

$$\begin{aligned}
z_1, z_2, z_3, z_4 &\sim \mathrm{U}(-1.73, 1.73)\\
x_{1,2,3,4} &\sim \mathrm{N}(z_1, .045^2)\\
x_{5,6,7,8} &\sim \mathrm{N}(z_2, .05^2)\\
x_{9,10,11,12} &\sim \mathrm{N}(z_3, .055^2)\\
x_{13,14,15,16} &\sim \mathrm{N}(z_4, .06^2)\\
y &= z_1 + .5z_2 + .25z_3 + \epsilon\\
\epsilon &\sim \mathrm{N}(0, 0.5^2),
\end{aligned}$$

that is, x's are noisy observations of z so that there are four groups of highly correlated covariates and four of the covariates have no effect on y. One thousand replications of the training data sets with $n = 20$ were simulated and the performance was evaluated with independent test data sets with $n_t = 400$. The basic model used was

$$\begin{aligned}
\tilde{y} &= \sum_{j=1}^{16} \alpha_j x_j + e\\
\alpha_j &\sim \mathrm{N}(0, \sigma_\alpha^2)\\
\sigma_\alpha &\sim \text{Inv-}\chi^2(0.5, 0.5^2)\\
e &\sim \mathrm{N}(0, \sigma_e^2)\\
\sigma_e &\sim \text{Inv-}\chi^2(0.5, 0.5^2).
\end{aligned}$$

Furthermore, the full model was constructed as a Bayesian model average of all possible submodels, that is, there was a non-zero prior probability that the effect of a covariate can be exactly zero. Bayesian model averaging was not used for submodels, since the goal was to select a submodel which would have only relevant covariates with non-zero weights. Sampling of parameters was made using Gibbs sampling and Bayesian model averaging was made using reversible jump MCMC (RJMCMC).

In this toy example, we used the root mean square error (RMSE) as the cost function to relate the performance of the selected models to the amount of noise in the data. The proposed method was compared to the following reference methods: expected cost minimization via CV, expected cost minimization via

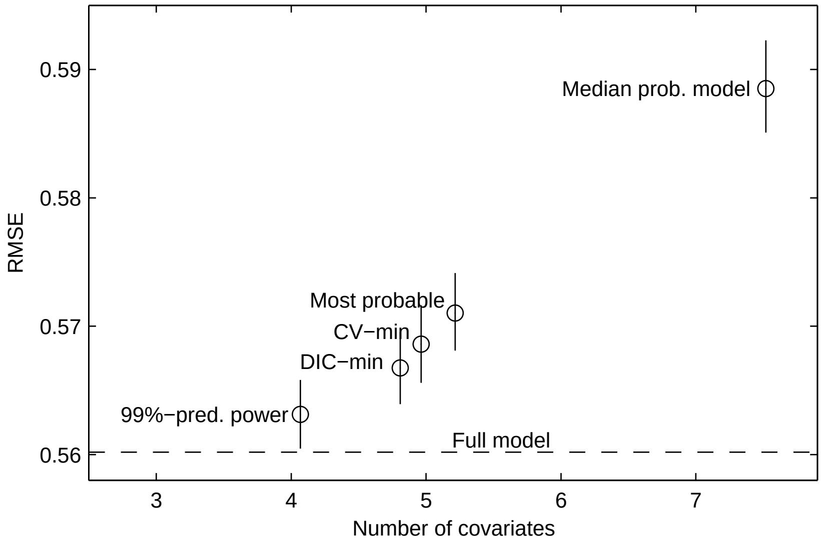

Figure 1: Toy example: The mean and the 90% credible interval of root mean square error (RMSE) for the test data and the mean number of selected covariates for models selected using the proposed and reference methods. The dashed horizontal line shows the performance of the full model.

DIC, selection of the most probable model, and the median probability model method. For the proposed method, CV-minimization, and DIC-minimization, the search was made using basic forward search. The most probable model corresponds to model selection with Bayes factors. The median probability model is defined as the model consisting of those covariates whose overall posterior probability of being in a model is greater than or equal to 1/2 (Barbieri & Berger, 2002). The most probable model and the median model were estimated using samples from RJMCMC.

Figure 1 shows the mean and the 90% credible interval of the root mean square error for independent test data for the full model and submodels selected using the proposed method and the reference methods. Figure 1 also shows the mean number of selected covariates for each selection method. The models selected with the proposed method give on average better predictions than models selected with the reference methods. The data was generated using 12 covariates, but all methods selected models with fewer covariates. This is due to the strong correlations between covariates and thus, using more covariates than what were selected could improve performance only slightly or negligibly.

Figure 2 illustrates the selection induced bias by showing the mean of the root mean square error for the test data and the mean of the CV-estimated expected root mean square error for each method. The difference in these values is the selection induced bias which also results in worse performance in model selection, as illustrated by the results of the reference methods. Even in this simple small problem, the effect is clear and it grows larger with more flexible models (e.g., non-parametric models) and with increasing number of covariates. The CV estimate of the expected cost of models selected by the proposed method is almost unbiased, because the CV estimate of the expected cost for the full model is almost unbiased, and the proposed method selects the model having a similar predictive distribution as the full model. Thus, an additional benefit of using the proposed method is that it is possible to get an almost unbiased estimate of the effect of the reduced predictive explanatory power in terms of application specific utility, if such is available. Although it is not shown in the figures, replacing the cross-validation predictive densities with marginal posterior predictive densities also induces selection bias. Even in this simple problem this bias was visible, although not as large as for the other reference methods.

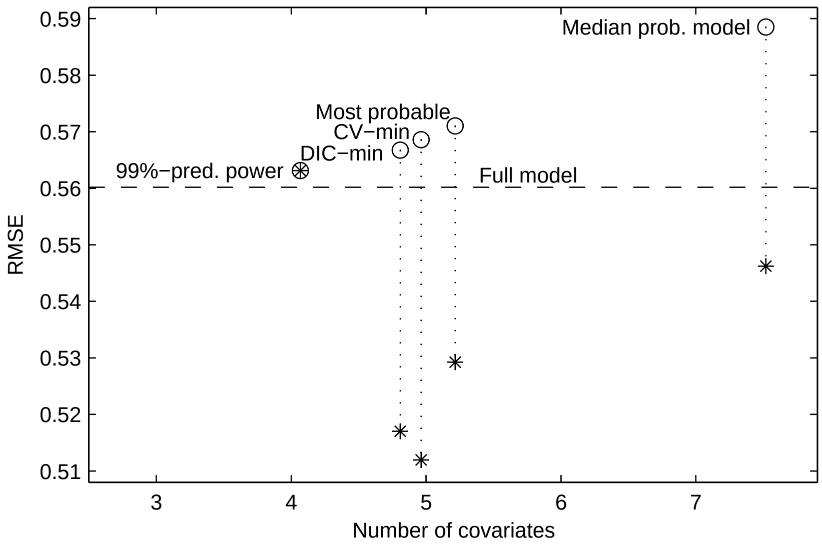

Figure 2: Toy example: The mean of the root mean square error for the test data (circle) and the mean of the CV-estimated expected root mean square error (star) for submodels selected using the proposed and reference methods. The dashed horizontal line shows the performance of the full model.

In this simple toy example, it was feasible to make a forward search through the model space. In the case of more complex models, for which Monte Carlo sampling from the posterior distribution may take much longer time, it may be necessary to use a greatly reduced search. We tested an approach in which the covariates are ordered based on the marginal posterior probabilities and only the models having the k most probable covariates are compared. Marginal posterior probabilities do not help to reduce highly similar covariates, but it can help to reduce completely irrelevant covariates. This approach could be combined with a more complete search after the most probably irrelevant covariates have been removed.

Figure 3 shows the mean of the root mean square error for the test data and the mean number of selected covariates for models selected using the proposed method with forward search and the reduced search. The average performance of the models having the k most probable covariates is also shown. Forward search is able to find models with much less covariates than the reduced search based on ordering the covariates according to the marginal probabilities. The reduced search is able to remove some of the irrelevant covariates, but since the search is restricted, it is not able to reduce the number of highly correlating covariates. Note that if we knew how many relevant covariates there were, and we selected a model having 9–12 covariates, we would get better performance than with the full model. Since in real problems we do not know the number of relevant covariates, on the average we are not able to select better submodels than the full model.

Figure 4 illustrates that using the reduced search does not avoid selection induced bias, although this bias is reduced compared to CV minimization. The bias in the reduced search is caused by the pre-ordering of the covariates based on the marginal probabilities. Since the model selection is based on comparison to the full model, this bias does not have such an effect on performance of the selected model as in the direct minimization of the expected cost estimation.

3.2 Real world problem I: Concrete quality estimation

The goal of the project was to develop a model for predicting the quality properties of concrete, as a part of a large quality control program of the industrial partner of the project (Järvenpää, 2001). The

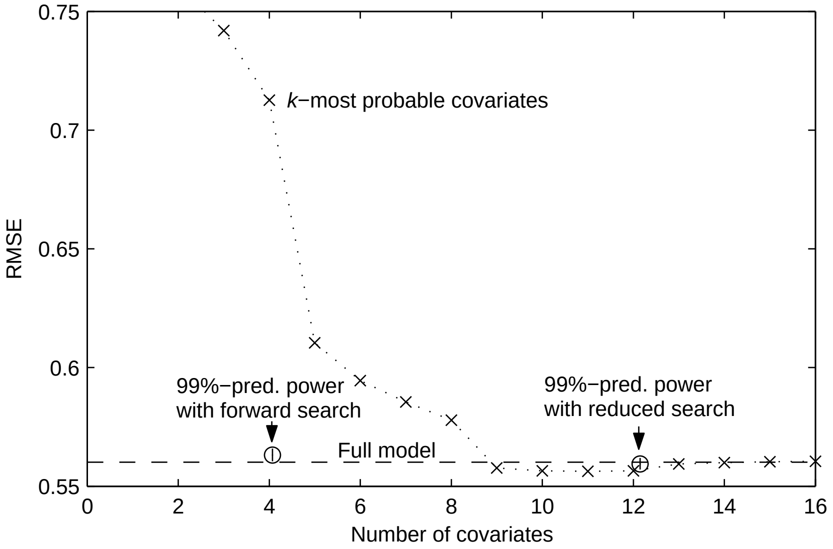

Figure 3: Toy example: The mean of the root mean square error (RMSE) for the test data and the mean number of selected covariates for models selected using the proposed method with forward search and the reduced search. The dashed horizontal line shows the performance of the full model. The dotted line with x's shows the average performance of the models having the *k* most probable covariates.

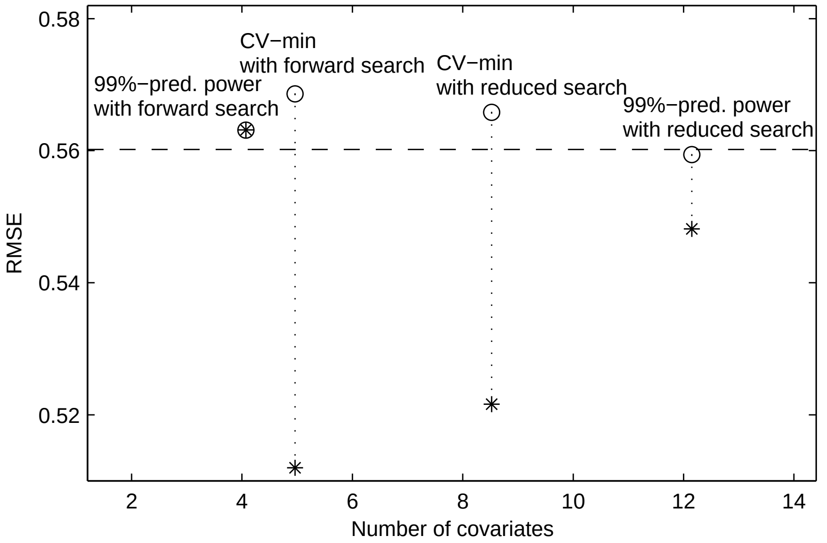

Figure 4: Toy example: The mean of the root mean square error for the test data (circle) and the mean of the CV-estimated expected root mean square error (star) for submodels selected using proposed method with forward search or the reduced search, and CV minimization with forward search or the reduced search. The dashed horizontal line shows the performance of the full model.

quality variables included, for example, compressive strengths and densities for 1, 28 and 91 days after casting, and bleeding (water extraction), flow value, slump, and air-%, that measure the properties of fresh concrete. These quality measurements depend on the properties of the stone material (natural or crushed, size and shape distributions of the grains, mineralogical composition), additives, and the amount of cement and water. In the study we had 27 explanatory variables and 215 samples designed to cover the practical range of the variables, collected by the concrete manufacturing company.

The aim of the study was to identify which properties of the stone material are important, and additionally, examine the effects that properties of the stone material have on concrete. It was desirable to select the minimal set of covariates required to get a model with similar predictive capability as the full model. A smaller model is easier to analyse and there is no need to make possibly costly or toxic measurements in the future of properties having negligible effect.

The prediction models made in the project are already in use by concrete industry. The reduction of the covariates helped the concrete expert to successfully use graphical tools to visualize the effects and interactions of the inputs (Järvenpää, 2001). By using the models and conclusions based on them it was possible to reduce the proportion of the natural gravel in concrete from 50% to 5%-20% and achieve 5-15% savings in concrete manufacturing.

The covariate selection reported in reference (Järvenpää, 2001) was made using the DIC and heuristic backward selection, and the covariate selection for the current models in use was made using CV-based expected utility maximization (Vehtari & Lampinen, 2001a). The method presented in this paper was used to check the effect of the selection induced bias in CV-based expected utility maximization. Here we report results for the volume percentage of air in the concrete.

Based on expert knowledge and preliminary analysis, it was known that there were nonlinear effects, strong interactions and dependencies in the covariates. An useful model in such situations is a non-linear nonparametric Gaussian process model, which can handle interactions implicitly (Neal, 1999). This alleviates the covariate selection problem, since we only need to select whether a covariate should be included in the model or not, and if it should, the model automatically handles the possible interactions. We used a Gaussian process with a quadratic covariance function producing smooth functions. Each variable had its own length scale parameter describing the characteristic length of the function for a given direction. This prior effectively controls the relevance of the covariates and thus allows use of a large number of potentially useful covariates.

The residual model used was an input dependent Student's t_ν with an unknown number of degrees of freedom ν. As the size of the residual variance seemed to vary depending on three inputs, which were zero/one covariates indicating the use of additives, the parameters of the Student's t_ν were made dependent on these three inputs with a common hyperprior.

Posterior and predictive distributions were computed with MCMC methods. Details of the models and the computation are described in references (Vehtari & Lampinen, 2001b, Appendix; Vehtari, 2001). The MCMC sampling was done with Matlab-code partly derived from the FBM software[1] and the Netlab toolbox[2]. The sampling for one model took about eight hours with a fast work station, and thus there was a need to perform a reduced search in model space.

Figure 5 shows the CV-estimated expected predictive likelihoods for models with the k most probable covariates. Note that the estimates for the submodels are biased and for most submodels, the expected performance is estimated to be better than the performance of the full model.

Figure 6 shows the relative loss in predictive explanatory power for models with the k most probable covariates. Based on this plot, we selected the model with nine most probable covariates. Similar

[1] http://www.cs.toronto.edu/~radford/fbm.software.html

[2] http://www.ncrg.aston.ac.uk/netlab/

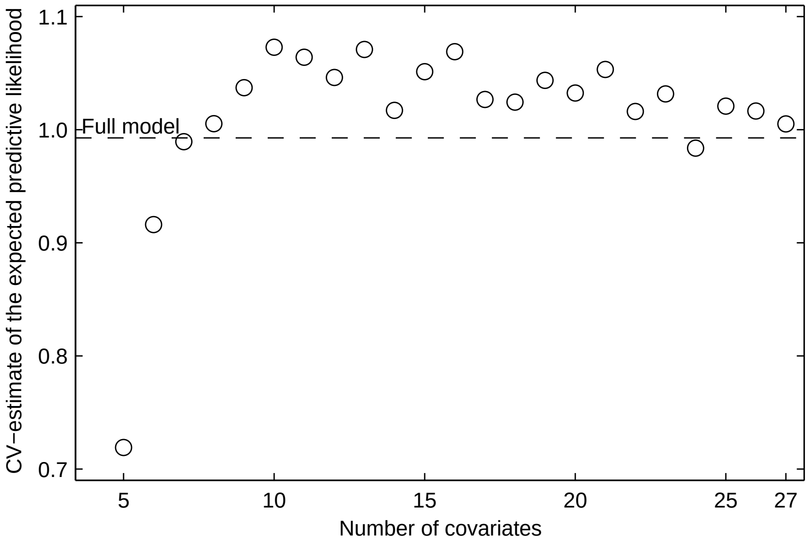

Figure 5: Concrete example: CV-estimated expected predictive likelihoods for models with the k most probable covariates. The dashed horizontal line shows the performance of the full model.

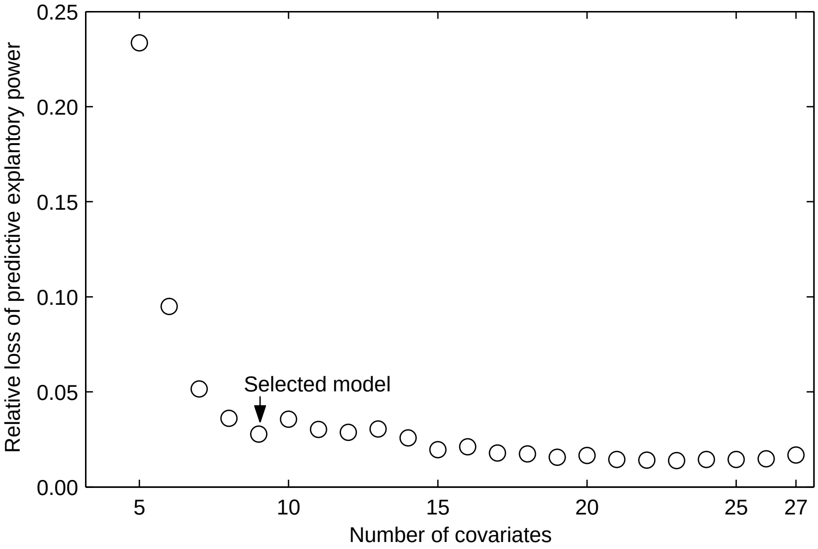

Figure 6: Concrete example: The relative loss of predictive explanatory power estimated with the proposed method for models having the k most probable covariates.

plots could be generally used to find out an elbow in predictive explanatory power instead of using a predefined limit for acceptable loss. Note that in Figure 5 the expected utility of the model with the seven most probable covariates was estimated to be almost same as the full model, but the relative loss of predictive explanatory power for that model is 5%, almost twice as much as the relative loss of predictive explanatory power for the model with the nine most probable covariates.

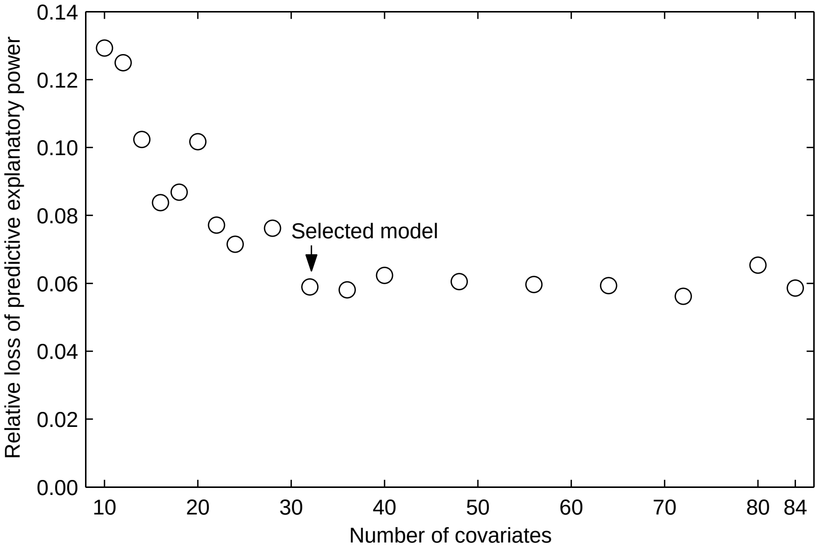

Figure 7: Forest example: The relative loss of predictive explanatory power estimated with the proposed method for models having the *k* most probable covariates.

3.3 Real world problem II: Forest scene classification

The case study here is the classification of forest scenes with MLP networks (Vehtari, Heikkonen, Lampinen, & Juujärvi, 1998). The final objective of the project was to assess the accuracy of estimating the volumes of growing trees from digital images. To locate the tree trunks and to initialize the fitting of the trunk contour model, a classification of the image pixels to tree and non-tree classes was necessary. Training data was 4800 samples from 48 images (100 pixels from each image) with 84 different Gabor and statistical features as covariates. The goal of the covariate selection was to reduce the computational burden, but without explicitly defined cost for the covariates.

We used a non-linear nonparametric one hidden layer MLPs with 20 tanh hidden units and a logistic likelihood model. We used Gaussian priors on weights and a hierarchical Gaussian prior on input weights. This hierarchical prior allows less relevant inputs to have a smaller effect in the model (Neal, 1996; Lampinen & Vehtari, 2001). Details of the models and the priors are described in the references (Vehtari & Lampinen, 2001b, Appendix; Vehtari, 2001). The sampling for one model took more than 12 hours with a fast workstation.

Figure 6 shows the relative loss in predictive explanatory power for a part of the models with the *k* most probable covariates. Because of the heavy computational burden, the discrepancy to the full model was not computed for every value of *k*. Selecting the model with the 32 most probable covariates would reduce the required computational time for predictions with more than half without measurable loss of predictive explanatory power.

4 Discussion and Conclusion

We have proposed a decision theoretic predictive model selection approach, based on the expected Kullback-Leibler divergence from the full model to a submodel. The goal is to find the simplest submodel which has a similar predictive distribution as the full model, which we believe describes our knowledge of the phenomenon in the best possible way.

Bernardo and Smith (1994, Ch. 6) discuss how cross-validation approximates the formal Bayes procedure of computing the expected utilities. Decision theory states that the optimal action is the one which maximizes the expected utility. The problem is that in model selection we have only an estimate of the expected utility and as discussed and illustrated in sections 2.5 and 3.1, the variance in this estimate makes the model selection suboptimal. For covariate selection Bernardo and Smith (1994, Ch. 6) propose identifying optimal submodels $M_{(k)}$ for each number k of covariates included. Observing that

$$\bar{u}(M_{(1)}) \leq \bar{u}(M_{(2)}) \leq \cdots \leq \bar{u}(M_{(k)}) \leq \cdots , \tag{20}$$

is typically concave, reflecting the marginal expected utility for the incorporation of further covariates, they propose to select $M_{(k)}$ for which $\bar{u}(M_{(k+1)}) - \bar{u}(M_{(k)})$ is less than some appropriately predefined small constant. Again, the problem is that we have only an estimate of the expected utility, and thus selection of $M_{(k)}$ for each k is suboptimal and the estimates $\hat{u}(M_{(k)})$ are biased.

The proposed method overcomes these problems by replacing the direct estimation of the expected utility of a submodel with the estimation of the expected predictive discrepancy from the full model to a submodel. Although, there is variance also in the estimate of the expected discrepancy, this does not have the same effect in the selection of the best submodels $M_{(k)}$ with k covariates and with the proposed method, the estimates $\hat{u}(M_{(k)})$ are almost unbiased.

In covariate selection, ordering of the models is straightforward. In non-nested cases it is not always clear, what is a simpler model. For example, using coordinate transforms it may be possible to change a non-linear problem to a linear problem which may be considered simpler. Also, it is not easy to compare how simple different non-nested hierarchical Bayesian models are. For example, many random effect models are easy to describe, although they contain large numbers of parameters and also the effective number of parameters may be very different from the available number of parameters. If the primary goal is to obtain good predictions and the selection is based on the discrepancy to the full model, we may choose any of the submodels having similar predictions as the full model, and thus we can select the model which we or the application expert feel is simpler.

Acknowledgements

The authors would like to thank Toni Tamminen for helpful comments on the manuscript.

References

Barbieri, M. M., & Berger, J. O. (2002). Optimal predictive model selection. ISDS Discussion Paper 02-02, Duke University, Institute of Statistics and Decision Sciences.

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics*, 7(3), 686–690.

Bernardo, J. M. (1999). Nested hypothesis testing: The Bayesian reference criterion. In J. M. Bernardo, J. O. Berger, & A. P. Dawid (eds.), *Bayesian Statistics 6*, (pp. 101–130). Oxford University Press.

Bernardo, J. M., & Juáres, M. A. (2003). Intrinsic estimation. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (eds.), *Bayesian Statistics 7*, (pp. 456–476). Oxford University Press.

Bernardo, J. M., & Rueda, R. (2002). Bayesian hypothesis testing: a reference approach. *International Statistical Review*, 70(3), 351–372.

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

Brown, P. J., Fearn, T., & Vannucci, M. (1999). The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika*, 86(3), 635–648.

Brown, P. J., Vannucci, M., & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(3), 627–641.

Brown, P. J., Vannucci, M., & Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3), 519–536.

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.

Chipman, H., George, E. I., & McCulloch, R. E. (2001). Practical implementation of Bayesian model selection (with discussion). In P. Lahiri (ed.), *Model Selection*, vol. 38 of *IMS Lecture Notes – Monograph Series*, (pp. 65–134). Institute of Mathematical Statistics.

Dellaportas, P., & Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86(3), 615–633.

Draper, D., & Fouskakis, D. (2000). A case study of stochastic optimization in health policy: Problem formulation and preliminary results. *Journal of Global Optimization*, 18, 399–416.

Dupuis, J. A., & Robert, C. P. (2003). Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference*, 111, 77–94.

Fearn, T., Brown, P. J., & Besbeas, P. (2002). A Bayesian decision theory approach to variable selection for discrimination. *Statistics and Computing*, 12(3), 253–260.

Fernández, C., Ley, E., & Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100(2), 381–427.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350), 320–328.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365), 153–160.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (eds.), *Markov Chain Monte Carlo in Practice*, (pp. 145–162). Chapman & Hall.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3), 501–514.

Han, C., & Carlin, B. P. (2000). MCMC methods for computing Bayes factors: A comparative review. Research Report 2000-001, Division of Biostatistics, University of Minnesota.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, 3rd ed.

Järvenpää, H. (2001). *Quality characteristics of fine aggregates and controlling their effects on concrete*. Acta Polytechnica Scandinavica, Civil Engineering and Building Construction Series No. 122. The Finnish Academy of Technology.

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

Kohn, R., Smith, M., & Chan, D. (2001). Nonparametric regression using linear combination of basis functions. *Statistics and Computing*, 11(4), 313–322.

Lampinen, J., & Vehtari, A. (2001). Bayesian approach for neural networks – review and case studies. *Neural Networks*, 14(3), 7–24.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2), 187–192.

Lindley, D. V. (1968). The choice of variables in multiple regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(1), 31–66.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag.

Neal, R. M. (1999). Regression and classification using Gaussian process priors (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (eds.), *Bayesian Statistics 6*, (pp. 475–501). Oxford University Press.

Ntzoufras, I. (1999). *Aspects of Bayesian model and variable selection using MCMC*. Ph.D. thesis, Department of Statistics, Athens University of Economics and Business.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(4), 731–792.

Robert, C. P. (1996). Intrinsic lossess. *Theory and decision*, 40(2), 191–214.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9(1), 130–134.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422), 486–494.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3), 583–639.

Sykacek, P. (2000). On input selection with reversible jump Markov chain Monte Carlo sampling. In S. A. Solla, T. K. Leen, & K.-R. Müller (eds.), *Advances in Neural Information Processing Systems 12*, (pp. 638–644). MIT Press.

Vannucci, M., Brown, P. J., & Fearn, T. (2003). A decision theoretical approach to wavelet regression on curves with a high number of regressors. *Journal of Statistical Planning and Inference*, 112(1–2), 195–212.

Vehtari, A. (2001). *Bayesian Model Assessment and Selection Using Expected Utilities*. Dissertation for the degree of Doctor of Science in Technology, Helsinki University of Technology.

Vehtari, A. (2002). Discussion of “Bayesian measures of model complexity and fit” by Spiegelhalter et al. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3), 620.

Vehtari, A., Heikkonen, J., Lampinen, J., & Juujärvi, J. (1998). Using Bayesian neural networks to classify forest scenes. In D. P. Casasent (ed.), *Intelligent Robots and Computer Vision XVII: Algorithms, Techniques, and Active Vision*, (pp. 66–73). SPIE.

Vehtari, A., & Lampinen, J. (2001a). Bayesian input variable selection using posterior probabilities and expected utilities. Tech. Rep. B31, Helsinki University of Technology, Laboratory of Computational Engineering.

Vehtari, A., & Lampinen, J. (2001b). On Bayesian model assessment and choice using cross-validation predictive densities. Tech. Rep. B23, Helsinki University of Technology, Laboratory of Computational Engineering.

Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10), 2439–2468.

Vehtari, A., & Lampinen, J. (2003). Expected utility estimation via cross-validation. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (eds.), *Bayesian Statistics 7*, (pp. 701–710). Oxford University Press.