Economic variable selection

Koji MIYAWAKI\textsuperscript{1,2*} and Steven N. MACEACHERN\textsuperscript{3}

\textsuperscript{1}School of Economics, Kwansei Gakuin University, Nishinomiya, Japan
\textsuperscript{2}Policy Design Lab., Graduate School of Economics and Management, Tohoku University, Aoba-ku, Japan
\textsuperscript{3}Department of Statistics, The Ohio State University, Columbus, OH, U.S.A.

Key words and phrases: Decision-theoretic approach; model averaging; objective Bayes.

MSC 2020: Primary 62J05; secondary 62F15.

Abstract: Regression plays a central role in the discipline of statistics and is the primary analytic technique in many research areas. Variable selection is a classical and major problem for regression. This article emphasizes the economic aspect of variable selection. The problem is formulated in terms of the cost of predictors to be purchased for future use: only the subset of covariates used in the model will need to be purchased. This leads to a decision-theoretic formulation of the variable selection problems, which includes the cost of predictors as well as their effect. We adopt a Bayesian perspective and propose two approaches to address uncertainty about the model and model parameters. These approaches, termed the restricted and extended approaches, lead us to rethink model averaging. From an objective or robust Bayes point of view, the former is preferred. The proposed method is applied to three popular datasets for illustration. The Canadian Journal of Statistics 51: 19–37; 2023 © 2021 Statistical Society of Canada

Résumé: L’analyse de régression est un véritable pilier de la statistique et constitue, sans contexte, la principale technique analytique dans de nombreux domaines de recherche. La sélection de variables est certainement l’un de ses problèmes et défis communément rencontrés. Cette étude examine ce problème de sélection de variables à travers le prisme de l’aspect économique. Plus précisément, le problème est formulé en termes de coût des prédicteurs à acheter pour une utilisation future : seul le sous-ensemble de covariables utilisées dans le modèle devra être acheté. Ainsi, ce problème devient un problème de la théorie de la décision qui inclut le coût des prédicteurs ainsi que leurs effets. En adoptant une perspective bayésienne, les auteurs de ce travail proposent deux approches pour traiter l’incertitude au sujet du modèle et de ses paramètres. Ces approches, dites approches restreintes et étendues, incitent à revisiter les concepts de moyennage de modèles (ou intégration pondérée de modèles). D’un point de vue bayésien objectif ou robuste, la première approche est préférable à la seconde. En guise d’illustration, la méthode proposée est mise en pratique sur trois ensembles de données classiques. La revue canadienne de statistique 51: 19–37; 2023 © 2021 Société statistique du Canada

1. INTRODUCTION

Model selection with subsequent prediction is a classic and major problem in statistics. In the context of regression analysis, model selection is often equated with variable selection, to be accomplished in one of many ways, including classical hypothesis tests of full and reduced models (e.g., Vuong, 1989), use of an information criterion such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), or the deviance information criterion (DIC) (Akaike, 1998, for a reprint of the original article published in 1973 for AIC, Schwarz, 1978, for BIC, and Spiegelhalter et al., 2002, for DIC), evaluation through some form of cross-validation

* Corresponding author: miyawaki.koji@kwansei.ac.jp

© 2021 Statistical Society of Canada
(e.g., Gelfand, Dey & Chang, 1992; Gelfand & Dey, 1994), or Bayesian versions of tests based on the Bayes factor (Kass & Raftery, 1995). Subsequent prediction follows from either a separate re-fit of the data, as with model selection, followed by a least-squares fit, or is integrated into a cohesive framework involving selection and prediction as with many of the recently developed penalized likelihood methods.

Bayesian methods provide a distinct approach to model selection and prediction, as they are based on a cohesive modelling framework that allows one to simultaneously describe and work with their uncertainty across models and over parameters within a model. Its main applications in model selection are the hierarchical approach (Mitchell & Beauchamp, 1988; George & McCulloch, 1993) and the stochastic search approach (Hans, Dobra & West, 2007; Fouskakis & Draper, 2008). These methods follow the usual route from prior distribution through data to posterior distribution, with inference to follow. Model selection follows from inference designed to minimize incorrect model selection, while prediction follows inference to minimize forecasting loss. This approach separates modelling from inference, facilitating, e.g., Barbieri & Berger (2004) to distinguish model selection from variable selection. It has also led to an explosion of literature on model averaging, such as Min & Zellner (1993), Madigan et al. (1995), Raftery et al. (1997), Draper (1995), Brown et al. (2002), and Yu et al. (2011), of which the benefits are now well documented.

The split between modelling and inference has generated a novel approach to parsimony within Bayesian circles, which may be characterized as “fit in a large model space, make inference in a small model space” (Walker & Gutiérrez-Peña, 1999; MacEachern, 2001; Hahn & Carvalho, 2015). This approach moves parsimony from modelling to inference. It seeks to construct a model that reflects the full complexity of the problem and, if little benefit is shown for some variables (aspects of the model), to move to a simpler form as part of inference. In this article, we explicitly bring an economic question into the mix—namely the cost of predictors—and pursue a path suggested by the decision-theoretic formulation of the model/variable selection and prediction problem in regression. This version places our focus on two main questions:

1. Prediction, accounting for the cost of predictors. In a typical setting, predictors have costs associated with them. They cost money, take time to collect, take effort to model, or consume computational effort. These costs are real, and obtaining a slightly better prediction rule at a much higher cost or much more slowly may not be worthwhile.
2. Model uncertainty. The goal of model selection is often taken to be consistent model selection, or identification of the set of predictors with nonzero coefficients in the regression model. The economic formulation of the prediction problem suggests that a slightly inferior (in the traditional sense) model may provide a better model for practical use. This suggests a re-examination of the role of consistency in model selection. See also Clyde & George (2004) for recent approaches to model uncertainty.

Consideration of the cost of predictors and formulation of model selection as a decision problem has appeared in the literature. Lindley (1968) argues forcefully for Bayesian methods and for a full decision-theoretic formulation of the problem. Further authors have mentioned this issue (see, e.g., Brown et al., 1999; Hahn & Carvalho, 2015). Fouskakis et al. (2009a, 2009b) also take the cost into account in different directions.

In this article, we seek to reconcile economic considerations with current practice in Bayesian model selection and model averaging. We find that this perspective provides strong commentary on current practice, we present several reasons to believe that current practice is generally reasonable, and we identify settings where improvements can be made. Overall, we find that Bayesian model averaging (BMA) is a valuable technique but also that care should be taken in its implementation.
2023 ECONOMIC VARIABLE SELECTION

The rest of the article is organized as follows. The next section lays out our methodology, including two approaches (Sections 2.4 and 2.5). Three datasets are used to illustrate our method in Section 3. Section 4 points to future directions and concludes the article.

2. ECONOMIC VARIABLE SELECTION

2.1. Normal Linear Model with \( g \) Prior

Suppose we have the response \( Y_i \) and \( p \) potential predictors \( x_i \) for each \( i = 1, \ldots, n \) observation (\( p < n - 1 \)). All predictors are standardized. A subset of \( p \) predictors is indexed by \( \gamma \) and is denoted by \( x_{i,\gamma} \), where \( \gamma = (\gamma_1, \ldots, \gamma_p) \) is a vector of 1’s and 0’s to indicate which predictors are in the model. The \( j \)th predictor is in the model \( \gamma \), the \( j \)th element \( \gamma_j \) is set to 1, and it is zero otherwise \( (j = 1, \ldots, p) \). Let \( p_\gamma \) be the dimension of \( x_{i,\gamma} \) and \( \Gamma \) be the set of all possible \( \gamma \)s.

This article focuses on the model, details of which are given below. For \( \gamma \in \Gamma \) and \( i = 1, \ldots, n \),

\[
Y_i = \beta_0 + x_i' \beta_{1,\gamma} + \epsilon_i,
\]

where the error term \( \epsilon_i \) independently and identically follows the normal distribution with mean 0 and variance \( \sigma^2 \), i.e., \( \epsilon_i \sim N(0, \sigma^2) \). Because the model is indexed by \( \gamma \) and \( \gamma \) is associated with \( x_i \), every model includes the intercept. The design matrix \( X_\gamma = (x_{1,\gamma}, \ldots, x_{n,\gamma})' \) is assumed to be of full column rank, which is satisfied in all examples in Section 3.

The model parameters are \( \phi_\gamma = (\beta_0, \beta_{1,\gamma}, \sigma^2) \). The subscript for \( \beta_0 \) and \( \sigma^2 \) is suppressed because they are commonly used in all models. In particular, because the predictors are standardized, \( \beta_0 \) is interpreted as the mean of the response variable. Thus, it is reasonable to assume its prior knowledge to be the same and non-informative across all models. While the error variance does not have such an interpretation, we make a similar assumption because it is a nuisance parameter. For the other parameters, proper prior distributions are assumed. There are \( 2^p \) possible models (\( |\Gamma| = 2^p \)). The prior distribution on this model space is assumed as non-informative because we have little information on which models are better in general. In summary, the following prior distributions are assumed:

\[
\pi(\beta_0) \propto 1, \quad \beta_{1,\gamma} \sim N_{p_\gamma}(\mathbf{0}, g\sigma^2 \left( \sum_{i=1}^{n} x_{i,\gamma} x_{i,\gamma}' \right)^{-1}),
\]

\[
\pi(\sigma^2) \propto \sigma^{-2}, \quad \pi(\gamma) = \prod_{j=1}^{k} s^j (1 - s)^{1-j},
\]

where \( g \) and \( s \) are known constants, and \( N_k(\mu, \Sigma) \) is the \( k \)-dimensional multivariate normal distribution with mean vector \( \mu \) and variance covariance matrix \( \Sigma \).

The prior for slope coefficients \( (\beta_{1,\gamma}) \) is called the \( g \) prior (see Zellner & Siow, 1980; Zellner, 1986). The constant \( g \) is set equal to the number of observations \( n \), which is recommended by Fernandez et al. (2001) when the number of observations is greater than the squared number of predictors. This prior specification is often called a class of benchmark priors. We choose it for operational simplicity, but other choices are possible.

When \( g = n \), the resulting \( g \)-prior has the unit information, where information in the prior is equal to information from one observation with respect to the Fisher information matrix. Such a prior was proposed by Kass & Wasserman (1995) to show the relationship between the Bayes factor and BIC. When \( g \) is equal to the squared number of predictors, the resulting \( g \)-prior satisfies the risk-inflated criterion (RIC), proposed by Foster & George (1994) in relation to the minimaxity.
It is possible to assume a hyperprior on $g$. Popular choices are the hyper-$g$ and hyper-$g/n$ priors proposed by Liang et al. (2008) and the robust prior proposed by Bayarri et al. (2012). In Section 2.7, we examine the prior sensitivity under the restricted approach (see Section 2.4) and empirically show it is robust to the prior choice between the above specification and other choices mentioned above.

The benchmark Beta prior by Ley & Steel (2012) and the block hyper-$g$ prior by Som et al. (2015) would be other choices. See also Ley & Steel (2012) for other hyperpriors including Maruyama & George (2011) as well as their performances on the numerical and empirical dataset. Priors from the objective perspective are extensively reviewed in Consonni et al. (2018).

Owing to the lack of knowledge about models, we set $s = 0.5$, leading to the uniform prior over models. Other specifications of the prior model probability are found in, e.g., Steel (2019).

Under the above specifications, the (marginal) posterior of $(\beta_0, \beta_1, \gamma)$ is a generalized multivariate $t$ distribution and the posterior model probability is proportional to the marginal likelihood, $m \left( y \mid X, \gamma \right)$, details of which are given in the Appendix.

2.2. Predictive Loss

Predictive regression modelling is often formulated as a decision problem, and it can be argued that this formulation underlies BMA. The traditional formulation of the problem is driven by a predictive loss of the form $L \left( y, \hat{y}(x) \right) = \left( y - \hat{y}(x) \right)^2$, where $y$ is a response to be predicted and $\hat{y}(x)$ is the predicted value associated with the predictors $x$. Using standard models such as the one given in the previous subsection and integrating over the conditional distribution of the future $y$, this loss becomes a loss taking parameter and action as arguments, namely

$$L \left( E[Y \mid x], \hat{y}(x) \right) = \left( y(x) - E[Y \mid x] \right)^2 = E \left[ L \left( Y, \hat{y}(x) \right) \mid x \right] - V[Y \mid x].$$  \hspace{1cm} (1)

The variance term in Equation (1) does not depend on the decision rule and hence can be ignored in the determination of the optimal rule. Under the normal linear regression model described in the previous subsection, this predictive loss is estimated by Equation (A3) in the Appendix.

Our focus is on Bayesian procedures, and the Bayes rule is the Bayesian’s optimal decision rule. It is typically constructed from the Bayesian posterior conditional viewpoint (Berger, 1985), by moving from prior distribution to posterior distribution and then choosing the action to minimize posterior expected (against the posterior distribution) loss.

BMA focuses on the setting where the prior distribution cuts across slices of the parameter space that are naturally described as models. In the case of linear regression with a set of $p$ potential predictors, the prior distribution is across the entirety of $R^p$ for the predictors’ regression coefficients. A model is defined by the set of nonzero regression coefficients, and the set of $2^p$ potential models partition $R^p$. The prior distribution on these coefficients is of mixed form. It typically assigns positive probability to each element of the partition—i.e., to each subset of $R^p$ that corresponds to a model. The support of the prior is the entirety of each element, leading to an overall support of all of $R^p$. The prior distributions that underlie BMA are thus seen to be of slightly non-standard form, but they are prior distributions.

From this perspective, BMA follows directly as a standard Bayesian procedure: pass from prior distribution to posterior distribution via the Bayes theorem. Once arriving at the posterior distribution, the optimal (posterior) action is found. In this case, the action happens to be expressed as a summary of the model-averaged posterior distribution, or, for squared-error loss, as model-averaged posterior predictive means. BMA is nothing more (nor less) than sound application of the Bayes theorem and choice of an appropriate action. As such, it inherits all of the optimality properties of Bayesian inference.
2.3. Decision with Cost

In the variable selection problem, the basic decision involves two sets of possibly overlapping and possibly null sets of predictors. The analyst must decide which to purchase, knowing the state of nature. For this decision, it is important to consider their costs as well as their predictive adequacies.

To this end, following Lindley (1968), we modify the original predictive loss to include the cost of data acquisition, modelling, and processing, including the cost to purchase information (as in credit history for a customer), cost of time (as in the delay in obtaining results from a medical lab test), cost in processing time (as in variables that are computationally expensive in conjunction with their use in a model), or other.

More specifically, suppose we have \((Y, x)\), a single future case. Let \(c(\gamma)\) be the non-negative cost function for the model \(\gamma\) that uses the single future case. The cost depends on the set of predictors, but it does not depend on the values of those predictors. Without knowledge of predictors, a typical choice is a function of the number of predictors. Section 3 provides specific forms of the cost function.

The total cost, or the negative utility, from purchasing predictors \(x\) is expressed as the sum of \(E \left[ L \left( Y, \hat{y}(x) \right) \mid x \right] \) and \(c(\gamma)\). It is better to purchase \(x_1\) than \(x_2\) if

\[
E \left[ L \left( Y, \hat{y}(x_1) \right) \mid x_1 \right] + c(\gamma_1) < E \left[ L \left( Y, \hat{y}(x_2) \right) \mid x_2 \right] + c(\gamma_2),
\]

where \(\gamma_1\) and \(\gamma_2\) are models associated with \(x_1\) and \(x_2\), respectively. In general, the best predictor to purchase is chosen by solving the following minimization problem:

\[
\min_{\gamma \in \Gamma} E \left[ L \left( Y, \hat{y}(x_\gamma) \right) \mid x_\gamma \right] + c(\gamma).
\]

If all predictors are free of charge, it is clear that the best combination of predictors is the one that minimizes the loss.

The idea of model selection (or variable selection in the normal linear regression model) as above goes back to Lindley (1968). A recent study is by Gelfand & Ghosh (1998). They propose a model selection criterion by using the weighted sum of losses based on the future and current data and discuss its properties and generalizations. The cost function in our case can be interpreted as a specific form of the loss based on the current data. More general discussion on this utility-based approach is found in Bernardo & Smith (2000), for example.

2.4. Two Approaches

In reality, the state of nature is unknown. Its uncertainty is specified as the form of distribution about model parameters and about models. Let \(x\) be the \(k\) purchased predictors, and let \(w\) denote the \(p - k\) unpurchased predictors. The predictors may or may not be relevant to predict the response, and we expect future data of the form \((Y, x)\) to reveal the relationship between the response and the purchased predictors. There are two main approaches to provide forecasts for future \(Y\) as a function of the future covariate \(x\).

**The restricted approach.** The restricted approach confines us to the small world of predictors \(x\) and response \(Y\). BMA applied to this world results in model averaging across \(2^k\) potential models, with individual predictors in \(x\) either active or not.

**The extended approach.** The extended approach considers the large world of models determined by predictors \(x\) and \(w\) for the response \(Y\). BMA applied to this world results in model averaging across \(2^p\) potential models, with individual predictors in \(x\) and \(w\) either active or not.
The first approach makes use of information only on purchased predictors. The second approach makes use of information on both purchased and unpurchased predictors. Information on the unpurchased predictors is available through the conditional distribution of the unpurchased predictors given the purchased (and maybe less expensive) predictors. It can be considered as an extreme of imputation in the missing value problem, where all cases are missing for some predictors (see also Boone, Ye & Smith, 2011). The measurement error model also has a similar structure, in that the true value is unobserved (see also Doppelhofer, Hansen & Weeks, 2016; Zhang, Ma & Carroll, 2019).

The predictive loss for both approaches is

$$E \left[ (Y - h(x, w))^2 \mid x \right],$$

where $h(\cdot)$ is the action as a function of potential predictors. When the normal linear regression model is used, this loss corresponds to a Bayesian version of the Mallows $C_p$ (see Mallows, 1973).

The restricted approach removes $w$ from the problem, restricting $h$ to be a function of $x$ alone, and it averages over a reduced set of models. This approach leads to the following expression of the loss:

$$E \left[ \left\{ Y - \sum_{\gamma \in \Gamma} h(x_\gamma) \pi(\gamma) \right\}^2 \mid x \right],$$

where $x_\gamma$ is a subset of purchased predictors.

The extended approach marginalizes the loss over $w$ by its conditional distribution $g(w \mid x)$, leading to

$$E \left[ \left\{ Y - \sum_{\gamma \in \Gamma, \lambda \in \Lambda} \left( h(x_\gamma, w_\lambda) \int g(w_\lambda \mid x_\gamma) \pi(\gamma, \lambda) \right)^2 \mid x \right\} \right],$$

where $w_\lambda$ is a subset of unpurchased predictors indexed by $\lambda$, which is defined in a manner similar to $\gamma$, and $\Lambda$ is a set of all possible $\lambda$s. In either approach, the optimal action minimizes the sum of predictive loss and cost of predictors.

Both these approaches can be implemented with standard computational methods. The restricted approach is standard BMA based on the purchased predictors. The extended approach is easiest to follow if we assume that we know the joint distribution of potential predictors. In this case, the unpurchased predictors are merely missing data to be imputed (distributionally) as we fit our model. When BMA is accomplished by means of Markov chain Monte Carlo (MCMC), standard methods allow us to draw the missing values in each iterate of the algorithm. If the conditional distribution of $w \mid x$ is not fully determined, and it follows a probability model governed by hyperparameters, it is merely part of the larger Bayesian model and MCMC or other techniques can be used to perform model averaging over the full set of $2^p$ models.

2.5. Choice of Approach

One central question is whether the restricted approach or the extended approach is to be preferred. Our first take on this question is motivated by the subjective Bayesian viewpoint expressed, for example, in Savage (1972). He constructs Bayesian methods from the principles of rational behaviour. This leads to the notion of personal probability, and along with it the ability to specify a prior distribution on unknown parameters (tied to $Y \mid \phi$). The same argument
allows one to specify a prior on models and a prior on the distribution of \( w \mid x \). This provides a complete description of uncertainty over models, parameters within a model, and missing predictors. Coupling this with standard results from decision theory—which state that the Bayes risk is the minimum possible risk when the parameter follows a given distribution and that the Bayes rule achieves the Bayes risk (Result 1, p. 159 of Berger, 1985)—we arrive at the usual Bayesian destination. In other words, the extended approach is preferred over the subjective view.

The implications of this choice run contrary to mainstream Bayesian practice. Consider a standard BMA problem where one has a set of \( k \) predictors, say \( x \), and a response \( Y \). The usual practice is to apply BMA to the set of all \( 2^k \) models. While this may appear to agree with the preceding paragraph, we can certainly envision further unobserved predictors \( w \) that may well be connected to the response at a low cost. The extended approach averages over these predictors as well, with the analyst’s prior beliefs governing the relationship between \( x \) and \( w \) and the extended set of model probabilities.

This leads us to ask why BMA is practiced in its current form. Objective Bayesian methods provide a counterpoint to the subjective Bayesian perspective. The typical BMA implementation is far from subjective. Rather than using elicitation procedures to carefully specify a prior distribution across models, and for each model a prior distribution over the parameters within the model, one resorts to a rule to determine the prior distribution. The rule may assign a set probability to each model of a given size, and it may routinely specify the distribution on the parameters given the model. Popular rules include the conjugate priors on model parameters along with the uniform prior model probability (Raftery, Madigan & Hoeting, 1997), the benchmark prior (Fernández, Ley & Steel, 2001), and the mixture of \( g \) priors (Ley & Steel, 2012). See Steel (2019) for other choices.

Many of these prior distributions are improper, negating the subjective Bayesian argument. These prior distributions are not constructed in the careful fashion appropriate for smaller scale problems, and they are not accompanied by the claim that all can be modelled, including unseen \( w \). A typical attempt at such a specification of the distribution of \( w \mid x \) would lead to an improper distribution for \( w \). To see this, replace \( w \) with \( Y \) and note that the marginal distribution on \( Y \) is improper for many objective specifications: in particular, for those in which a regression makes use of a uniform improper prior distribution on the intercept or an improper prior distribution on the error variance. For unseen \( w \), we may be left without a distribution, and this precludes the use of the extended approach.

In addition to the question of whether the extended approach can be applied under a chosen version of the Bayesian paradigm, there is a question of whether it should be used. The major concerns surround our inability to check aspects of the model for future data—our inability to check the form of \( Y \mid (x, w) \) when \( w \) is unavailable and our inability to check the form of \( w \mid x \)—and our inability to consistently estimate the distribution of \( w \mid x \) as future data accrue. This last implies that, even as the future dataset size tends to \( \infty \), there will always be some uncertainty about the value of observing \( w \).

The robustness to priors is also an issue when comparing approaches. The conditional distribution \( w \mid x \) is an additional (subjective) prior. This specification may or may not be correct, bringing additional sensitivity to the analysis. If it is based on scientific theory, it should help us to obtain accurate prediction at a lower cost. However, especially in the area of social sciences, it is unstable due to, for example, the advance of technology or the change of legislation. Such a misspecification may contaminate the inference, as shown by examples provided in Section 2.1 of Liu et al. (2009) and Hahn (2019). They compare two models with and without biased samples, and discuss how the benefit from the former is less than its cost, which in turn suggests preferring the restricted approach over the extended approach from an objective perspective.

DOI: 10.1002/cjs.11675
At the end of this section, we compare these two approaches with empirical datasets. Figure 1 shows the empirical difference of these two approaches by using the ozone dataset (see Section 3.1). We use the normal linear model with $g$-prior described in Section 2.1 and draw a pair of losses estimated by taking two approaches (see the next subsection for the loss estimation). The prior distribution of $w | x$ for the extended approach is constructed by using the normal approximation: use the entire dataset to construct the multivariate normal distribution for $(x, w)$ and derive conditional distributions of unpurchased predictors given purchased ones. This figure shows that these two approaches result in similar losses because they are gathering around the 45 degree line. Thus, in addition to the points listed above, as well as the computational aspect, we recommend the restrictive approach, and the article focuses on it hereafter.

### 2.6. Cross-Validated Loss

When the data are observed, we are able to estimate the loss. Let $(y_i, x_i)$ be the response to be predicted and the purchased predictors for case $i (i = 1, \ldots, n)$, respectively. Let $D_\gamma = \{y_i, x_{i,\gamma}\}_{i=1}^n$ denote the data for each model $\gamma$.

The uncertainties about models and parameters are estimated by their posterior distributions. Then, the loss is estimated as

$$
\left\{ \bar{y} - \sum_{\gamma \in \Gamma} h(\bar{x}_\gamma) \pi(\gamma | D_\gamma) \right\}^2,
$$

where $(\bar{y}, \bar{x}_\gamma)$ is the new response and predictors for the subset $\gamma$, $\pi(\gamma | D_\gamma)$ is the posterior model probability, and $h(\cdot)$ is an action to be chosen. Under the squared-error loss, the best action is the posterior conditional expectation $E(\bar{Y} | \bar{x}_\gamma, D_\gamma)$.

When the new data are not available, the cross-validated loss is an alternative. The data are split into two parts: training data and validation data. Then, the conditional expectation and the
posterior distributions are estimated based on the training data, and by using the validation data in place of the new data, we have the estimated loss. When the data are divided into several groups, this process is repeated by treating one of them as the validation and the remaining ones as the training. The cross-validated loss is the average of these losses. The rest of this article applies 10-fold cross-validation to estimate the predictive loss. Equation (A3) in the Appendix provides the analytical form of loss under the normal linear model with $g$-prior.

2.7. Prior Sensitivity

The next section will illustrate our methodology by using real datasets. The results may depend on the prior specification we choose. This subsection examines the sensitivity from this possibility by using the ozone data to be used in Section 3.1.

Taking the restrictive approach, normal linear models with four prior specifications are considered: (i) $g = k^2$, (ii) the hyper-$g$ prior

$$\pi(g) = \frac{1}{2} (1 + g)^{-3/2}, \quad (g > 0),$$

(iii) the hyper-$g/n$ prior

$$\pi(g) = \frac{1}{2n} \left(1 + \frac{g}{n}\right)^{-3/2}, \quad (g > 0),$$

and (iv) the robust prior

$$\pi(g) = \frac{1}{2} \sqrt{\frac{1 + n}{1 + k}} \left(1 + \frac{g}{n}\right)^{-3/2}, \quad (g > (1 + k)^{-1}(1 + n) - 1),$$

in addition to the $g$ prior for the last three specifications.

Figure 2 shows a pair of losses based on different prior specifications, along with the 45 degree line. For losses under the hyper-$g/n$ prior, the standard Laplace approximation is used as suggested by Liang et al. (2008). All panels show that these five specifications do not have any substantial difference in terms of the squared predictive loss. Thus, we use the $g$ prior with $g = \max\{n, k^2\}$ as recommended by Fernandez et al. (2001) for operational simplicity in the following illustrative examples.

3. ILLUSTRATIVE EXAMPLES

This section illustrates the economic variable selection with three real datasets. All predictors are standardized, and the loss is estimated by using 10-fold cross-validation.

3.1. Ozone Dataset

The first dataset was originally analyzed by Breiman & Friedman (1985) to develop a model between the daily ozone concentration level and meteorological variables in Los Angeles. We use the data provided by the R package “bfp” (Sabanes Bove & Held, 2011). The number of observations is 330.

The response is the log daily ozone concentration level in 1976 measured at Upland, California. There are 10 possible predictors: (i) 500-millibar-pressure height; (ii) wind speed; (iii) relative humidity; (iv) temperature at Sandberg; (v) inversion base height; (vi) a binary variable (which is set to 1 if the inversion base height is 5000); (vii) pressure gradient from Los Angeles International Airport to Daggett; (viii) inversion base temperature; (ix) square root of visibility; and (x) day of year.
Figure 2: Loss plots of four different prior specifications. (a) $g = p^2$, (b) hyper-$g$ prior, (c) hyper-$g/n$ prior, and (d) robust prior.

Figure 3 summarizes the results. Both panels consist of two parts: the upper part is the (estimated) squared predictive loss plot in ascending order, and each column of the lower part represents a corresponding combination of predictors purchased. When a cell of a column in the lower part is filled by black, the predictor labelled on the $y$-axis is purchased. When, on the other hand, it is white, the corresponding predictor is not purchased, which is denoted by NP in the legend provided under the panel. The marginal posterior probability that the coefficient is nonzero is discretized by the four intervals: $[0, 0.25)$, $(0.25, 0.5]$, $(0.5, 0.75]$, $(0.75, 1]$, and is expressed by the brightness of the cell as shown in the legend. See Clyde (2003) for the marginal posterior nonzero probability.

The least loss combination is $(x_3, x_4, x_5, x_6, x_9, x_{10})$. Among them, $x_6$ and $x_9$ are less relevant in terms of their marginal posterior nonzero probability (less than 0.6). Thus, cells corresponding
to these predictors are coloured light grey. Compared to the selection by Breiman & Friedman (1985), we choose $x_3$ (and $x_6$) instead of $x_7$.

This difference is partly due to the transformation of variables. The response is logged in our example, while it is not in Breiman & Friedman (1985) (see their Figure 5(a), p. 588). Among the predictors, $x_7$ is transformed by a highly nonlinear function in Breiman & Friedman (1985) (see their Figure 5(d), p. 588), while we do not transform it. These transformations would be a source of such a difference between their result and ours.

To focus on combinations that yield smaller predictive losses, the left panel of Figure 3 is magnified to the right by picking up the top 128 combinations of the left. In this right panel, $(x_4, x_{10})$ are always included in terms of the nonzero probability. Among others, $x_3$ is in combination with smaller predictive losses.

Two special combinations are considered: the intercept-only combination and the combination that purchases all predictors. Their respective position is denoted by the vertical solid lines labelled by $N$ and $F$ in Figure 3. The former yields high predictive loss, although it is not the worst (the fourth from the worst). On the other hand, the latter performs much better. This loss is achieved when we use the usual BMA. Its predictive performance is closer to the best (see the predictive loss plot of the left panel). However, there are combinations that yield low predictive losses and purchase fewer predictors.

Next, two cost structures are considered. The first one is the uniform cost structure, where all predictors are set at the same price. That is, when $k$ predictors are purchased, the total cost is $c \cdot k$, where $c$ is the price. This structure is used when a decision maker has no information about the cost of predictors.

When the uniform cost structure is applied, the predictors purchased are shown by the left panel of Figure 4. Each column is the least-loss combination for a fixed $c$, which is indicated by the $x$-axis. Similarly to the previous plots, the brightness represents the discretized marginal posterior probability that the corresponding coefficient is nonzero. As $c$ increases, fewer predictors are purchased. When $c$ is sufficiently high, the optimal purchase is the one with no predictors.
It is reasonable to consider that a decision is made with some knowledge about the cost of predictors. A possible decision maker for this dataset is a researcher who is interested in global warming. As a part of his or her interest, the researcher would like to predict the ozone level. He or she probably knows the cost of predictors. One reasonable cost structure for the researcher is cost per predictor. Because \((x_2, x_3, x_4, x_9, x_{10})\) are often reported on regular weather news, it is natural to assume that their costs are zero, while the remaining predictors require positive prices. To simplify the structure, we assume each of them requires the constant price \(c\). The results are shown in the right panel of Figure 4. As \(c\) increases, the optimal purchase is the one without \(x_5\) and \(x_6\) because of their higher cost.

### 3.2. Diabetes Dataset

The next dataset is the diabetes data, used in Efron et al. (2004) and provided through Professor Trevor Hastie’s webpage (https://web.stanford.edu/~hastie/Papers/LARS/diabetes.data. Accessed 4 November 2021). The data are used to predict the progression of the disease 1 year ahead of the baseline when predictors related to patients are collected. In this dataset, 442 observations are included.

The response is the logarithm of the diabetes progression measure. Ten possible predictors are included: age, sex, body mass index, blood pressure, and six blood serum measures.

The results without cost are summarized in Figure 5. The least-loss combination is \((x_2, x_3, x_4, x_5, x_6, x_9)\). From the top 128 combinations of predictors purchased, \((x_3, x_9)\) is useful to predict the disease progression because it is always included in the combinations. In addition, \(x_2\) and \(x_4\) perform well because they are in low-loss combinations. Among the six blood serum measures, \(x_7\) is useful as well because it is almost always included in the combinations. Efron et al. (2004) applied the least angle regression and found that variables enter into the active set in the order of \(x_3, x_9, x_4, x_7\), where they are selected in combination with higher predictive losses in our results.
The performance of two special combinations is examined. The intercept-only combination is the second worst. The combination that purchases all predictors does not perform well in this dataset, suggesting purchases that include fewer predictors would be better to predict the disease progression.

Two specific cost structures are examined. The first one is the uniform cost structure, and its results are on the left panel of Figure 6. As $c$ (the uniform price) increases, the number of predictors in the optimal purchase decreases. The optimal purchase with sufficiently high price is the one only with the intercept.
A possible decision maker for this dataset is a person who is at the risk of diabetes. If he or she considers it low, the cost of the blood test is high. On the other hand, if he or she considers it high, it becomes cheap. To simplify this decision problem, the constant cost $c$ is introduced if any of the six blood serum measures is included in the combination. Other predictors are assumed to be free of charge.

The results are shown on the right panel of Figure 6. As the price for the blood test increases, $(x_5, x_6, x_9)$ are excluded in the optimal purchase because they become more expensive. When it is sufficiently high, $(x_3, x_4)$ are selected to predict the progression of diabetes. For a person who is at low risk of diabetes, these predictors are enough for the purpose.

### 3.3. Wage Dataset

The last dataset focuses on how wage is determined by attributes of workers, such as education level and ability.

The dataset to be used in the analysis is the one taken from the National Longitudinal Survey of Youth and is the panel data from 1979 to 1993. This was analyzed by Koop & Tobias (2004), and is provided from *The Journal of Applied Econometrics* data archive. The response is the log of hourly wage for white males. Koop & Tobias (2004) excluded individuals who were aged less than 16 years or who reported small wages, short working hours, or inappropriate education years. There are seven possible predictors: (i) education in years; (ii) potential experience in years (age − years of education − 5); (iii) the ability measure ranging from about −4 to about 2, constructed on 10 component tests of the Armed Services Vocational Aptitude Battery; (iv) mother’s education in years; (v) father’s education in years; (vi) binary variable for broken home until the age of 14; and (vii) number of siblings. The response and the first two variables are time-varying, while the remaining five are time-invariant. More details of this dataset are given in Section 4 of Koop & Tobias (2004).

The least-loss combination of predictors purchased for each wave is aligned in Figure 7. The ability measure $(x_3)$ comes into the set of predictors purchased after the fifth wave in terms of

---

**Figure 7:** Wage data: least-loss combinations by waves.
FIGURE 8: Wage data: when to purchase. NP denotes the combination that purchases no predictor for all waves. (a) No cost and (b) cost introduced.

its marginal posterior nonzero probability more than 0.75. A possible reason is as follows: For the first 4 years, companies mainly set wages by the education level \((x_1)\) and the experience \((x_2)\) because the ability is unobservable at this moment. Ability will become clear over time. After about 4 years, companies start to use their information to set wages more accurately.

A decision problem in this dataset is when to purchase the ability measure as a manager of a company. When it is free of charge, purchasing at the beginning of the \(t\)th wave yields the prediction loss as \(\sum_{s=1}^{t-1} l_s + \sum_{s=t}^{15} l_s\), where \(l_s\) and \(l_s^*\) are the least losses with and without purchasing the ability measure, respectively.

The left panel of Figure 8 plots this loss changing when to purchase. The minimum loss is reached at the beginning of the first wave (represented by the black dot), although losses at the beginning of the second through fifth waves are comparable.

The decision changes when cost is introduced. There are two kinds of cost in this problem: the discount factor \((\delta)\) and the price of the ability measure \((c)\). Because the present and future utilities/costs are not equivalent, the discount factor is introduced to evaluate the future value in terms of the present value. If the ability measure was purchased at the beginning of the \(t\)th wave, the prediction loss with cost adjusted by the discount factor would be

\[
\sum_{s=1}^{t-1} \frac{l_s}{(1 + \delta)^{t-1}} + \sum_{s=t}^{15} \frac{l_s^*}{(1 + \delta)^{t-1}} + \frac{c}{(1 + \delta)^{t-1}}.
\]

The loss is discounted because it is interpreted as the utility in the statistical decision problem or because it is measured by dollars so that it is additive to the cost.

The decision with cost is shown in the right panel of Figure 8. Each panel is for a fixed discount factor and 10 different prices of the ability measure, ranging from 0.01 to 0.2. A grey line is the plot of the loss described above for a fixed price and discount factor, and a black dot indicates the minimum.

As the price gets higher, the loss increases. When we see the top left panel as an example, the grey line moves up along the \(y\)-axis as the price increases. Thus, in this panel, the optimal
decision stays at the beginning of the fifth wave. With the positive discount factor, it moves to a later wave or no purchase as the price increases. See the top right panel, for example.

As the discount factor becomes larger, the loss plot shows a downward trend because the decision maker values the present more than the future. Then the optimal decision tends to be to purchase the measure at a later wave even if it decreases the loss. When the discount factor and price are sufficiently large, the optimal decision is not to purchase. It is reasonable that the manager of a company purchases the ability measure later or decides not to purchase when it is expensive and/or the discount factor is large.

4. DISCUSSION

The variable selection problem depends on the person who chooses predictors. This aspect is formulated as a decision problem, and the optimal decision is the BMA with purchased predictors. Following our terminology, it is the restricted approach. The extended approach is another methodology to select predictors when the subjective prior information about the distribution of unpurchased predictors, conditional on purchased ones, is available. As discussed in Section 2.5, the restricted approach is our recommendation.

Empirical results that employ the restricted approach show that the predictive loss is improved with a subset of predictors, compared with the one with all predictors. Cost structures specific to the dataset are also considered. We find that the optimal decision (the optimal set of predictors to be purchased in this case) changes according to the structure or the level of the price for predictors.

Finally, a computational issue is noted. The method is computationally feasible when the number of predictors is moderate. However, for example, the economic growth model usually includes more than 50 predictors (67 in Sala-i-Martin, Doppelhofer & Miller, 2004). In this case, a simple method such as stepwise selection would be useful to remove less relevant predictors before proceeding to apply our method (see, e.g., James et al., 2013, for stepwise selection). However, a decision-theoretic variable selection in high dimensions is an interesting research question and will be left to future work.

ACKNOWLEDGEMENTS

We are grateful to the editor, the review editor, and three anonymous referees for giving us invaluable comments to improve this article. We thank Professor Carlos Carvalho (University of Texas) for his constructive discussion at O-Bayes17 and Professor Yasuhiro Omori (University of Tokyo) for his helpful comments at EAC-ISBA 2019. We also appreciate comments from Professor Mark Steel (University of Warwick), which significantly improved this article and a comment from Professor Heleno Bolfarine (University of Sao Paulo) which brought to our notice the necessary reference. One of the authors acknowledges support through a scholarship from Kwansei Gakuin University and a grant from the Japan Society for the Promotion of Science (JSPS) (no. 25245035). This study was conducted as part of a project implemented at the Policy Design Laboratory, Tohoku University.

REFERENCES

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. Selected Papers of Hirotugu Akaike. Springer Series in Statistics, Springer, New York.

Arellano-Valle, R. B. & Bolfarine, H. (1995). On some characterizations of the t-distribution. Statistics & Probability Letters, 25, 79–85.

Barbieri, M. M. & Berger, J. O. (2004). Optimal predictive model selection. The Annals of Statistics, 32, 870–897.

Bayarri, M. J., Berger, J. O., Forte, A., & Gracia-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. The Annals of Statistics, 40, 1550–1577.
Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics, 2nd ed., Springer-Verlag, New York.

Bernardo, J. & Smith, A. F. M. (2000). *Bayesian Theory*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Boone, E. L., Ye, K., & Smith, E. P. (2011). Assessing environmental stressors via Bayesian model averaging in the presence of missing data. *Environmetrics*, 22, 13–22.

Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580–598.

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

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, 519–536.

Clyde, M. (2003). Model averaging. *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. Wiley Series in Probability and Mathematical Statistics, Chapter 13, 2nd ed., Wiley, Hoboken, NJ, 320–335.

Clyde, M. & George, E. I. (2004). Model uncertainty. *Statistical Science*, 19, 81–94.

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13, 627–679.

Doppelhofer, G., Hansen, O.-P. M., & Weeks, M. (2016). *Determinants of long-term economic growth redux: A measurement error model averaging (MEMA) approach*. NHH Department of Economics Discussion Paper No. 19.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 57, 45–97.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32, 407–499.

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

Foster, D. P. & George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22, 1947–1975.

Fouskakis, D. & Draper, D. (2008). Comparing stochastic optimization methods for variable selection in binary outcome prediction, with application to health policy. *Journal of the American Statistical Association*, 103, 1367–1381.

Fouskakis, D., Ntzoufras, I., & Draper, D. (2009a). Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *The Annals of Applied Statistics*, 3, 663–690.

Fouskakis, D., Ntzoufras, I., & Draper, D. (2009b). Population-based reversible jump Markov chain Monte Carlo methods for Bayesian variable selection and evaluation under cost limit restrictions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 58, 383–403.

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

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling based methods. In Bernardo, J. M., Berger, J. O., Dawid, A. P., & Smith, A. F. M. (Eds.) *Bayesian Statistics 4*, Oxford University Press, New York, 147–167.

Hans, C., Dobra, A., & West, M. (2007). Shotgun stochastic search for “large p” regression. *Journal of the American Statistical Association*, 102, 507–516.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer, New York.
Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Kass, R. E. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90, 928–934.

Koop, G. & Tobias, J. L. (2004). Learning about heterogeneity in returns to schooling. *Journal of Applied Econometrics*, 19, 827–849.

Ley, E. & Steel, M. F. (2012). Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics*, 171, 251–266.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103, 410–423.

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

Liu, F., Bayarri, M. J., & Berger, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4, 119–150.

MacEachern, S. N. (2001). Decision theoretic aspects of dependent nonparametric processes. *Bayesian Methods: With Applications to Science, Policy and Official Statistics*, Eurostat, 551–560.

Madigan, D., York, J., & Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63, 215–232.

Mallows, C. L. (1973). Some comments on c_p, *Technometrics*, 15, 661–675.

Maruyama, Y. & George, E. I. (2011). Fully Bayes factors with a generalized g-prior. *The Annals of Statistics*, 39, 2740–2765.

Min, C.-K. & Zellner, A. (1993). Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics*, 56, 89–118.

Mitchell, T. J. & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83, 1023–1032.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92, 179–191.

Sabañes Bove, D. & Held, L. (2011). Bayesian fractional polynomials. *Statistics and Computing*, 21, 309–324.

Sala-i-Martin, X., Doppelhofer, G., & Miller, R. I. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (Bace) approach. *The American Economic Review*, 94, 813–835.

Savage, L. J. (1972). *The Foundations of Statistics*. Dover Publications, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Som, A., Hans, C. M., & MacEachren, S. N. (2015). Block hyper-g priors in Bayesian regression. https://arxiv.org/abs/1406.6419.

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

Steel, M. F. (2019). Model averaging and its use in economics. *Journal of Economic Literature*, 53, 644–719.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57, 307–333.

Walker, S. G. & Gutiérrez-Peña, E. (1999). Robustifying Bayesian procedures (with discussion). In Bernardo, J. M., Berger, J. O., Dawid, A. P., & Smith, A. F. M. (Eds.) *Bayesian Statistics 6*, Oxford University Press, New York, 685–710.

Yu, Q., MacEachern, S. N., & Peruggia, M. (2011). Bayesian synthesis: Combining subjective analyses, with an application to ozone data. *The Annals of Applied Statistics*, 5, 1678–1698.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics and Statistics, Chapter 15*, Vol. 6, North-Holland/Elsevier, Amsterdam, 233–243.

Zellner, A. & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., & Smith, A. F. M. (Eds.) *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia (Spain)*, University Press, Valencia, 585–603.
This section derives the posterior, the marginal likelihood, and the loss under the normal linear regression model specified in Section 2.1. The subscript \( \gamma \) is suppressed in this section to simplify notation, except for the number of predictors. We use \( k \) instead of \( p_\gamma \).

Suppose we have the training data \( D = \{ (y_i, x_i) \}_{i=1}^n \). The matrix representation gives \( y = (y_1, \ldots, y_n)' \) and \( X = (x_1, \ldots, x_n)' \). Consider the following normal linear regression and prior distribution:

\[
Y_i = \beta_0 + x_i' \beta_1 + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n, \\
\pi (\beta_0) \propto 1, \quad \beta_1 \sim N_k \left\{ 0, g\sigma^2 (X'X)^{-1} \right\}, \quad \pi (\sigma^2) \propto \sigma^{-2}.
\]

Then we have an analytical form for the (marginal) posterior distribution of the regression coefficients, \( \beta = (\beta_0, \beta_1)' \). The posterior is the Arellano-Valle and Bolfarine’s generalized \( t \) distribution, which is given by

\[
\beta | D \sim t (b, B; S, n - 1),
\]

where \( \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad d_y^2 = y'y - ny^2, \)

\[
b = \left( \frac{\bar{y}}{1 + g} (X'X)^{-1} X'y \right), \quad B = \left( \frac{1}{n} \quad 0' \quad g \left( \frac{1}{1 + g} \left( X'X \right)^{-1} \right) \right), \\
R^2 = \frac{y'X (X'X)^{-1} X'y}{d_y^2}, \quad S = \frac{d_y^2}{1 + g} \left\{ 1 + g \left( 1 - R^2 \right) \right\}.
\]

The probability density function is given in Arellano-Valle & Bolfarine (1995). The posterior expectation and variance matrix of \( \beta \) are \( b \) and \( \frac{S}{n-3} B \), respectively.

The marginal likelihood is derived as

\[
m (y | X) = \frac{\Gamma \left( \frac{(n-1)/2}{2} \right)}{\sqrt{\pi^{n-1} / n}} \left( 1 + g \right)^{(n-k-1)/2} d_y^{n-(n-1)} \left\{ 1 + g \left( 1 - R^2 \right) \right\}^{-n/2},
\]

where \( \Gamma(x) \) is the gamma function (see also Steel (2019) for this expression).

Finally, the squared predictive loss given the model is estimated by

\[
\frac{1}{m} \sum_{i=1}^m \left\{ \bar{y}_i - \bar{y} - \frac{g}{1 + g} \bar{y}_i (X'X)^{-1} X'y \right\}^2,
\]

where \( \bar{y}_1, \ldots, \bar{y}_m)' \) and \( \bar{x}_1, \ldots, \bar{x}_m)' \) are the response and predictors in the validation set with \( m \) observations.

---

Zhang, X., Ma, Y., & Carroll, R. J. (2019). Malmem: Model averaging in linear measurement error models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 81, 763–779.