The Report Committee for Mengjie Wang
certifies that this is the approved version of the following report:

Assigning g in Zellner's g Prior for Bayesian Variable Selection

APPROVED BY

SUPERVISING COMMITTEE:

Stephen Walker, Supervisor

Lizhen Lin

Assigning g in Zellner's g Prior for Bayesian Variable Selection

by

Mengjie Wang, B.S.;M.S.

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN STATISTICS

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2015

Assigning g in Zellner's g Prior for Bayesian Variable Selection

Mengjie Wang, M.S.STAT.
The University of Texas at Austin, 2015

Supervisor: Stephen Walker

There are numerous frequentist statistics variable selection methods such as Stepwise regression, AIC and BIC etc. In particular, the latter two criteria include a penalty term which discourages overfitting. In terms of the framework of Bayesian variable selection, a popular approach is using Bayes Factor (Kass & Raftery 1995), which also has a natural built-in penalty term (Berger & Pericchi 2001). Zellner's g prior (Zellner 1986) is a common prior for coefficients in the linear regression model due to its computational speed of analytic solutions for posterior. However, the choice of g is a problem which has attracted a lot of attention. (Zellner 1986) pointed out that if g is unknown, a prior can be introduced and g can be integrated out. One of the prior choices is Hyper-g Priors proposed by (Liang et al. 2008). Instead of proposing a prior for g, we will assign a fixed value for g based on controlling the Type I error for the test based on the Bayes factor. Since we are using

Bayes factor to do model selection, the test statistic is Bayes factor. Every test comes with a Type I error, so it is reasonable to restrict this error under a critical value, which we will take as benchmark values, such as 0.1 or 0.05. This approach will automatically involve setting a value of g. Based on this idea, a fixed g can be selected, hence avoiding the need to find a prior for g.

KEY WORDS: Model selection; Bayes factor; BIC; Zellner's g prior; Type I error

Table of Contents

List of Tables

List of Figures

Chapter 1

Model Selection

1.1 Linear Model

Regression analysis is arguably the most widely used dependence technique, applicable in various areas of decision making. The objective of linear regression analysis is to measure the relationship between a dependent variable and one or more independent variables (Hair 2010). For example, does protein power help people gain muscle? Or what are the features of the drilling process that affect gas production?

In linear regression, a set of weighted independent variables form the regression equation, which is a linear combination of the independent variables that best predict the dependent variable (Christensen 2011). In other words, the dependent/response variable is modeled through the linear combination of independent/explanatory variables and error.

This linear model is given by

$$Y = X\beta + \epsilon,$$

where

$$Y = (y_1, \ldots, y_n)'$$

is the response variable, n represents the number of observations. Here

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

is the design matrix with size of $n \times p$,

$$\beta = (\beta_1, \ldots, \beta_p)'$$

is the parameter vector to be estimated and

$$\epsilon = [\epsilon_1, \ldots, \epsilon_n]' \sim \mathcal{N}(0, \sigma^2 I),$$

where σ^2 is the variance of each ϵ_i. ϵ_i is the error term which adds noise to the relationship between the dependent variable and the predictors. The conditional expectation of dependent variable $\mathbb{E}[Y|X]$ is therefore equal to $X\beta$. σ is assumed to be fixed in this report.

The Ordinary Least Squares estimate of β is given by

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{1.1}$$

It is referred to as *Best Linear Unbiased Estimator* of β. We have

$$\mathbb{E}[\hat{\beta}] = \beta$$

and

$$Var[\hat{\beta}] = \hat{\sigma}^2(X'X)^{-1},$$

where

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n - 2}$$

(Christensen 2011).

In this report, for the sake of convenience and ease of exposition, I will use two predictors in the linear regression model, so for $i = 1, \ldots, n$

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \epsilon_i. \tag{1.2}$$

Without loss of generality, we can assume that

$$\sum_{i=1}^{n} x_{1i}^2 = \sum_{i=1}^{n} x_{2i}^2 = 1$$

and

$$\sum_{i=1}^{n} x_{1i} \cdot x_{2i} = \rho.$$

An important problem in linear regression is variable selection. The aim of this procedure is to reduce the whole set of predictors to a best subset. Identifying the predictors that significantly affect the response variable is crucial (Christensen 2011); it is also necessary since the redundant covariates need be removed so that the model can be precise, simple and can provide accurate predictions. To be more specific, the problem of collinearity can arise if two or more predictors are explaining the same thing.

In other words, in terms of likelihood, we would expect that the likelihood value will increase as the number of predictors goes up, *i.e.*, the complexity of the model becomes bigger. The problem is that a model with more predictor variables will always do better than the simpler model. But this leads to a problem called "overfitting", which will destroy the prediction accuracy. Hence, a good variable selection strategy is as crucial as the problem of variable selection itself.

A traditional frequentist method is Stepwise regression from (Efroymson 1960). One of the main approaches is Forward selection, which starts without any covariates in the model, importing each additional predictor into the model, testing whether it improves the model by F-test or t-test and removes any predictor which becomes insignificant as a result of introducing the new predictor, repeating this process until it reaches a equilibrium point where including a new variable will not draw a significant improvement of the model. (Kadane & Lazar 2004) mentioned that the selected model does not have to be the best, it is only the result of the algorithm applied to the particular dataset. The reason is the stepwise methods do not correspond to some specific criteria (Weisberg 1985). There are some other classical approaches or criteria such as C_p by (Mallows 1964), Akaike Information Criterion by (Akaike 1973), and Bayesian Information Criterion proposed by (Schwartz 1978). Mallows's C_p addresses the issue of overfitting while it is subject to selection bias (Mallows 1995). Recent work from (Boisbunon et al. 2014) showed that C_p and AIC are equivalent in the special case of Gaussian linear regression. Both AIC and BIC take the penalty terms into account, which penalize against the complexity of the model.

1.2 Bayes Factor

A common approach in Bayesian statistics is using Bayes Factor (Kass & Raftery 1995). Denote the model,

$$\mathcal{M} = \{f(y|\theta), \pi(\theta)\},$$

where $f(y|\theta)$ is the probability density function and $\pi(\theta)$ is the prior for the parameters. Suppose two models are of interest to us, $\mathcal{M}_0$ and $\mathcal{M}_1$. Typically, we usually assume

$$P(\mathcal{M}_1) = P(\mathcal{M}_2) = \frac{1}{2}.$$

$P(Y|\mathcal{M})$ is denoted as the marginal likelihood. So by Bayes Theorem,

$$P(\mathcal{M}_1|Y) = \frac{P(Y|\mathcal{M}_1)P(\mathcal{M}_1)}{P(Y|\mathcal{M}_1)P(\mathcal{M}_1) + P(Y|\mathcal{M}_0)P(\mathcal{M}_0)}$$

and so

$$\frac{P(\mathcal{M}_0|Y)}{P(\mathcal{M}_1|Y)} = \frac{P(Y|\mathcal{M}_0)}{P(Y|\mathcal{M}_1)} \times \frac{P(\mathcal{M}_0)}{P(\mathcal{M}_1)}.$$

So in words,

$$the\, posterior\, odds = \frac{P(Y|\mathcal{M}_0)}{P(Y|\mathcal{M}_1)} \times prior\, odds$$

and

$$\mathcal{B} = \frac{P(Y|\mathcal{M}_0)}{P(Y|\mathcal{M}_1)}$$

is known as the Bayes factor. Notice that the key to update of the odds is the Bayes factor.

For example, if we want to test two candidate models, $\mathcal{M}_0$ (Log-normal) against $\mathcal{M}_1$ (Weibull). Θ_0 is denoted as the parameter space for $\mathcal{M}_0$ and Θ_1 is denoted as the parameter space for $\mathcal{M}_1$. So this is testing

$$H_0 : \text{data comes from}\, \mathcal{M}_0\ against\ H_1 : \text{data comes from}\, \mathcal{M}_1.$$

The Bayes factor is calculated by

$$\mathcal{B}_{10}(Y) = \frac{\int_{\Theta_1} \pi(\theta_1|\mathcal{M}_1) P(Y|\theta_1, \mathcal{M})\, d\theta_1}{\int_{\Theta_0} \pi(\theta_0|\mathcal{M}_0) P(Y|\theta_0, \mathcal{M})\, d\theta_0},$$

where $\pi(\theta|\mathcal{M})$ is the conditional prior for the parameter and $P(Y|\theta, \mathcal{M})$ is the likelihood function. Notice that larger $\mathcal{B}_{10}$ supports the model on the numerator.

Chapter 2

Zellner's g prior

2.1 Zellner's g prior

In normal linear multiple regression model, (Zellner 1986) in his paper mentioned that assessing the informative prior distribution for the coefficient parameters is important. He proposed a reference informative prior called "g prior" which is easy to evaluate the prior covariance for the elements of β. As Zellner mentioned, this g prior is relatively simple to use.

The g prior for β is given by

$$\beta|\sigma^2 \sim \mathcal{N}\left(\beta_0, \sigma^2 g(X'X)^{-1}\right).$$

Notice that typically, $\beta_0 = 0$. This conjugate prior yields the Gaussian posterior for β, which is given by

$$\beta|\sigma^2, X, Y \sim \mathcal{N}(\Sigma^{-1}a, \Sigma^{-1}), \tag{2.1}$$

where

$$\Sigma = \frac{1}{g}\left(\sigma^2(X'X)^{-1}\right)^{-1} + \frac{1}{\sigma^2}X'X,$$

and

$$a = \frac{1}{g}\left(\sigma^2(X'X)^{-1}\right)^{-1}\beta_0 + \frac{1}{g}X'Y.$$

Recall that

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Hence, the posterior is

$$\beta|Y, \sigma^2, X \sim \mathcal{N}\left(\frac{1}{g+1}(\beta_0 + g\hat{\beta}), \frac{g\sigma^2}{g+1}(X'X)^{-1}\right).$$

So the posterior mean

$$\begin{aligned}\mathbb{E}[\beta|Y, \sigma^2, X] =& \Sigma^{-1}a \\ =& \frac{1}{1+g}\beta_0 + \frac{g}{1+g}(X'X)^{-1}X'Y \\ =& \frac{1}{1+g}\beta_0 + \frac{g}{1+g}\hat{\beta}.\end{aligned}$$

Zellner's informative g prior intuitively determines how much the prior distribution of β contributes to the posterior. For instance, if $g = 0$, the posterior mean fully shrinkages to the prior mean; if $g = 1$, the posterior mean shrinkages 50% to the prior mean; if g goes to ∞, the prior is a diffuse prior (Geinitz 2009).

Zellner's g prior is popular in variable selection. It provides a closed form for the marginal likelihood and an explicit expression for Bayes factor, which ensures a fast computation. However, the choice of g is problematic. (Zellner 1986) himself mentioned that g can depend on the sample size n, *e.g.*, $g \propto \frac{1}{n}$ or put a prior on g, and g can be integrated out. g could be chosen by finding the maximum posterior probability and corresponding to some popular selection criteria mentioned above, such as BIC and AIC. (George & Foster

2000) proposed the empirical Bayes selection criteria which have dimensionality penalties depending on the data. (Liang et al. 2008) proposed Hyper-g priors, which provides robustness to misspecification of g while maintaining the computational efficiency.

2.2 Choice for g

In terms of this two variables setting, I will implement the calibration idea that equalizing BIC and posterior probability to find the value for g.

2.2.1 Bayesian Information Criterion

A frequentist statistics strategy for model selection is called BIC (Bayesian Information Criterion) (Schwarz 1978). The model with the lowest BIC is preferred.

The BIC is formally defined as

$$BIC = -2 \cdot \log(\hat{\boldsymbol{L}}) + k \cdot \log(n),$$

where n is the number of observations, k is the number of parameters and $\hat{\boldsymbol{L}}$ is the maximized value of the likelihood function of the model with respect to the parameter θ.

2.2.2 Model Selection Using Posterior Probabilities

Assume the prior for each model is the same, we are interested in the posterior probability of each model $\mathcal{M}$ given the data, so we have

$$P(\mathcal{M}|Y) = \frac{P(Y|\mathcal{M}) \cdot P(\mathcal{M})}{\sum_{i=1}^{4} P(\mathcal{M}_i) \cdot P(Y|\mathcal{M}_i)},$$

where the denominator is the normalizing constant that is same for all of the four models.

In terms of the numerator, we can write $p(Y|\mathcal{M})$ as $\int p(Y|\mathcal{M}, \beta)\pi(\beta)\, d\beta$. Assume that $P(\mathcal{M}_1) = p(\mathcal{M}_2) = p(\mathcal{M}_3) = P(\mathcal{M}_4)$. We are left with figuring out the integral.

For the basic model,

$$p(Y|\mathcal{M}_1) = \int p(Y|\mathcal{M}_1, \beta)\pi(\beta)\, d\beta = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{n} y_i^2}{\sigma^2}\right]$$

and

$$p(\mathcal{M}_1|Y) \propto p(\mathcal{M}_1) \cdot p(y|\mathcal{M}_1) \propto p(\mathcal{M}_1) \cdot \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{n} y_i^2}{\sigma^2}\right]$$

For model with X_1,

$$p(\mathcal{M}_2|Y) \propto p(\mathcal{M}_2) \cdot \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sqrt{1+g}} \cdot \exp\left[\frac{1}{2\sigma^2}\frac{a^2}{1+\frac{1}{g}}\right] \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right] \tag{2.2}$$

Similarly, for model with X_2,

$$p(\mathcal{M}_3|Y) \propto p(\mathcal{M}_3) \cdot \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sqrt{1+g}} \cdot \exp\left[\frac{1}{2\sigma^2}\frac{b^2}{1+\frac{1}{g}}\right] \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right]$$

For the full model,*i.e.*, model with both variables,

$$p(\mathcal{M}_4|Y) \propto p(\mathcal{M}_4) \cdot \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \frac{1}{1+g} \cdot \exp\left[\frac{1}{2\sigma^2} \cdot \frac{a^2+b^2-2ab\rho}{(1-\rho^2)(1+\frac{1}{g})}\right] \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right] \tag{2.3}$$

2.2.3 Model Selection Using `BIC`

The basic model is the one without any independent variables. That is to say, $y_i = \epsilon_i$, for $i = 1, \ldots, n$.

Assume $\sum_{i=1}^n x_{1i}y_i = a, \sum_{i=1}^n x_{2i}y_i = b$. Since we have two predictors, we will have four different candidate models which are the basic model, *i.e.*, the one without any predictors, the model with only one predictor and the full model, *i.e.*, the one with both predictors.

Without any predictors,

$$BIC_1 = n\log(2\pi) + 2n\log(\sigma) + \frac{\sum_{i=1}^n y_i^2}{\sigma^2} \tag{2.4}$$

With only X_1 in the model,

$$BIC_2 = n\log(2\pi) + 2n\log(\sigma) + \frac{\sum_{i=1}^n y_i^2}{\sigma^2} - \frac{a^2}{\sigma^2} + \log(n) \tag{2.5}$$

With only X_2 in the model,

$$BIC_3 = n\log(2\pi) + 2n\log(\sigma) + \frac{\sum_{i=1}^n y_i^2}{\sigma^2} - \frac{b^2}{\sigma^2} + \log(n) \tag{2.6}$$

With both X_1 and X_2 in the model,

$$BIC_4 = n\log(2\pi) + 2n\log(\sigma) + \frac{\sum_{i=1}^n y_i^2}{\sigma^2} - \frac{a^2+b^2}{(1-\rho^2)\sigma^2} + \frac{2ab\rho}{(1-\rho^2)\sigma^2} + 2\log(n) \tag{2.7}$$

2.2.4 Choose g to Make Both Criteria Equivalent

In this section, four models are compared to see what condition needs be met to select the most appropriate one. And how g needs to be chosen to make both criteria equivalent to each other.

1. If the null model $\mathcal{M}_1$ is most preferred, *i.e.*, BIC_1 is the lowest,

$$\begin{cases} \frac{a^2}{\sigma^2} - \log(n) < 0 \\ \frac{b^2}{\sigma^2} - \log(n) < 0 \\ \frac{a^2-2ab\rho+b^2}{(1-\rho^2)\sigma^2} - 2\log(n) < 0 \end{cases} \Longrightarrow \begin{cases} a^2 < \sigma^2 \log(n) \\ b^2 < \sigma^2 \log(n) \\ \frac{a^2+b^2-2ab\rho}{1-\rho^2} < 2\sigma^2 \log(n) \end{cases}$$

   Equivalently, from Bayesian perspective, $P(\mathcal{M}_1|Y)$ must be the biggest to make $\mathcal{M}_1$ the preferred model. So,

$$\begin{cases} \exp[\frac{1}{2\sigma^2} \cdot \frac{a^2}{1+1/g}] \cdot \frac{1}{\sqrt{1+g}} < 1 \\ \exp[\frac{1}{2\sigma^2} \cdot \frac{b^2}{1+1/g}] \cdot \frac{1}{\sqrt{1+g}} < 1 \\ \exp[\frac{1}{2\sigma^2} \cdot \frac{a^2+b^2-2ab\rho}{(1-\rho^2)(1+\frac{1}{g})}] \cdot \frac{1}{1+g} < 1 \end{cases} \Longrightarrow \begin{cases} a^2 < \sigma^2(1+\frac{1}{g})\log(1+g) \\ b^2 < \sigma^2(1+\frac{1}{g})\log(1+g) \\ \frac{a^2+b^2-2ab\rho}{1-\rho^2} < 2\sigma^2(1+\frac{1}{g})\log(1+g) \end{cases}$$

2. If the model with X_1 is most preferred, *i.e.*, BIC_2 is the lowest,

$$\begin{cases} a^2 > \sigma \log(n) \\ a^2 > b^2 \\ \frac{(\rho a-b)^2}{1-\rho^2} < \sigma^2 \log(n) \end{cases}$$

   Equivalently, $P(\mathcal{M}_2|Y)$ is believed to be the biggest to make $\mathcal{M}_2$ the selected one, that is,

$$\begin{cases} a^2 > \sigma^2(1+g)\log(1+g) \\ a^2 > b^2 \\ \frac{(\rho a-b)^2}{1-\rho^2} < \sigma^2(1+\frac{1}{g})\log(1+g) \end{cases}$$

3. Similarly, in terms of model $\mathcal{M}_3$, From classic point of view,

$$\begin{cases} b^2 > \sigma^2 \log(n) \\ b^2 > a^2 \\ \frac{(\rho b - a)^2}{(1-\rho^2)} < \sigma^2 \log(n) \end{cases}$$

From Bayesian point of view,

$$\begin{cases} b^2 > \sigma^2 (1 + \frac{1}{g}) \log(1+g) \\ b^2 > a^2 \\ \frac{(\rho b - a)^2}{1-\rho^2} < \sigma^2 (1 + \frac{1}{g}) \log(1+g) \end{cases}$$

4. If the full model $\mathcal{M}_4$ is most preferred, we have

$$\begin{cases} \frac{a^2 - 2ab\rho + b^2}{1-\rho^2} > 2\sigma^2 \log(n) \\ \frac{(\rho a - b)^2}{1-\rho^2} > \sigma^2 \log(n) \\ \frac{(\rho b - a)^2}{1-\rho^2} > \sigma^2 \log(n) \end{cases}$$

and

$$\begin{cases} \frac{a^2 - 2ab\rho + b^2}{1-\rho^2} < 2\sigma^2 (1 + \frac{1}{g}) \log(1+g) \\ \frac{(\rho b - a)^2}{1-\rho^2} < \sigma^2 (1 + \frac{1}{g}) \log(1+g) \\ \frac{(\rho a - b)^2}{1-\rho^2} < \sigma^2 (1 + \frac{1}{g}) \log(1+g) \end{cases}$$

It is obvious to see that if $(1+\frac{1}{g})\log(1+g)$ is set to $\log(n)$, posterior probability selection criterion will get the same results as BIC.

Chapter 3

Hypothesis Testing

3.1 Statistical Hypothesis Testing

The statistical hypothesis is a testable assumption about the unknown parameters in the model. A statistical hypothesis testing is a procedure that is used to decide whether rejecting or not rejecting the hypothesis.

There are two hypotheses in statistical hypothesis testing, null hypothesis and alternative hypothesis. The null hypothesis is that there is no difference or relationship between the two measured quantities. The alternative hypothesis is the rival opponent or opposite counterpart against the null hypothesis.

3.2 Type I Error

The type I error happens when you falsely reject the null while it is true. The probability of making the type I error is usually denoted by α. For example, testing $\beta = 0$ against $\beta \neq 0$, the type I error in this case is rejecting $\beta = 0$ while the truth is $\beta = 0$.

In this report, σ is set to 1, a constant. The test statistics for model selection is Bayes factor, here the type I error is incorrectly rejecting the basic

model when the basic model is the true model. Assume

$$(1+\frac{1}{g})\log(1+g)=\lambda(g)$$

and

$$A=\{(a,b)|a^2<\lambda(g),b^2<\lambda(g),\frac{a^2+b^2-2ab\rho}{(1-\rho^2)}<2\lambda(g)\}.$$

So,

$$P_{\mathcal{M}_1}\left[P(\mathcal{M}_1|Y)>P(\mathcal{M}_2|Y),P(\mathcal{M}_1|Y)>P(\mathcal{M}_3|Y),P(\mathcal{M}_1|Y)>P(\mathcal{M}_4|Y)\right] \tag{3.1}$$

$$=P_{\mathcal{M}_1}\left[a^2<\lambda(g),b^2<\lambda(g),\frac{a^2+b^2-2ab\rho}{(1-\rho^2)}<2\lambda(g)\right]$$

$$=1-\alpha.$$

3.3 Finding A

The idea is to find the value of g to get the probability given by (3.1). Hence, all the three inequalities must be met simultaneously. Notice that (3.1) can be also written as

$$\int\int_A f(a,b)\,dadb,$$

where A is the constraints area in which a and b satisfy all the inequalities. $f(x,y)$ is the probability density function of the bivariate normal distribution, namely,

$$f(a,b)=\frac{1}{\sqrt{1-\rho^2}}\cdot\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\times\frac{1}{1-\rho^2}[a^2+b^2-2ab\rho]\right].$$

There are two ways to get the probability, one approach is to express the probability explicitly by calculating out the integral, and the other approach is by simulation. In this report, the second approach is applied to solve for g.

3.3.1 Visualization of Constraints

The boundary of the first two inequalities in (3.1) is a square, we have

$$(1 + \frac{1}{g}) \log(1 + g) = \lambda(g).$$

Here I take $\lambda(g)$ as 5 for illustration. Since σ is fixed, the constraints change only as ρ changes. See Figure (3.1), (3.2) and (3.3).

3.3.2 Simulation

The other approach to get the value of g with controlling type I error under the predetermined level is through simulation. We know that (a, b) is bivariate normally distributed. Next, one thousand *i.i.d.* samples of (a_i, b_i) are generated and a vector from 0.1 to 300 of g is set, so

$$\mathbb{P}(A) \approx \sum_{i=1}^{1000} \frac{1}{1000} \times \mathbb{1}\left[(a_i, b_i) \in A\right].$$

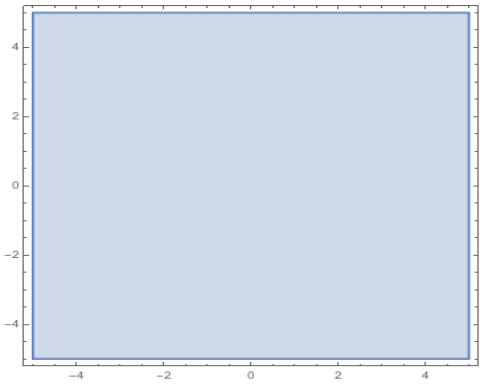
That is calculating the ratio of the points that meet all inequalities and the total points to estimate the true probability.

As we can see from both the table and the graph, the type I error decreases as g gets larger, which makes sense because as g goes up, the area of the constraints gets larger, there are more points tend to land in it. Notice

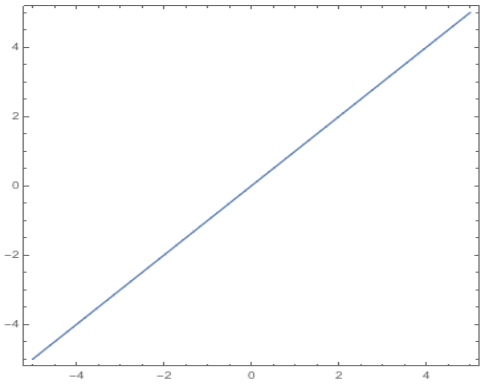
that the increasing speed of $1-\alpha$ gets slower as g increases, this simply results from the probability cannot be bigger than 1. In the graph, the horizontal coordinate of the point where the red line crosses the blue line is the value of g that satisfies our condition. See Table (3.1), Figure (3.4) and (3.5).

Table 3.1: Choice of g under different type I error and ρ

ρ	Type I Error	Value of g
0.20	0.20	11.5
	0.15	18.5
	0.10	41.2
	0.05	178.7
0.50	0.20	11.0
	0.15	20.1
	0.10	38.9
	0.05	232.6
-0.50	0.20	8.0
	0.15	14.8
	0.10	30.8
	0.05	150.4
-0.75	0.20	6.9
	0.15	12.5
	0.10	32.2
	0.05	148.4
0.00	0.20	10.0
	0.15	18.3
	0.10	42.9
	0.05	148.1

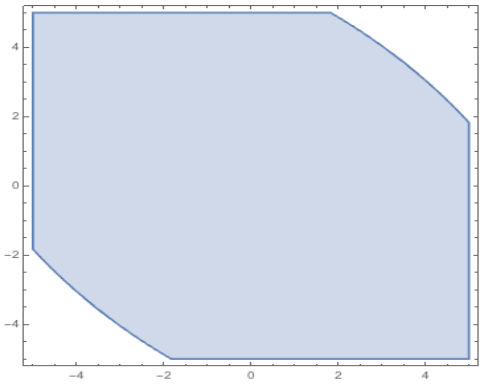


(a) Plot of constraints when $\rho = 0$

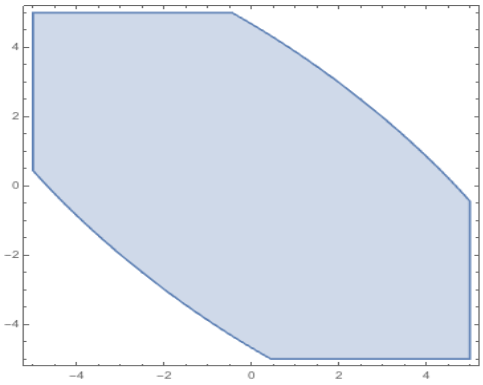


(b) Plot of constraints when $\rho = 1$

Figure 3.1: Plots of constraints when ρ is 0 or 1

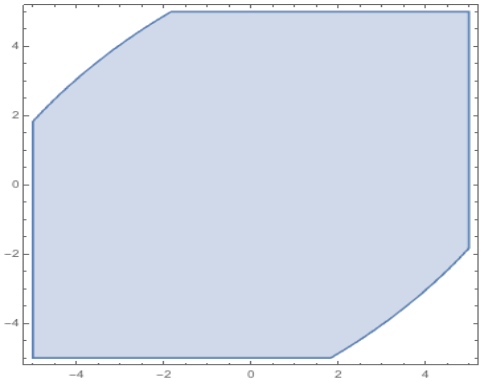


(a) Plot of constraints when $\rho = -0.5$

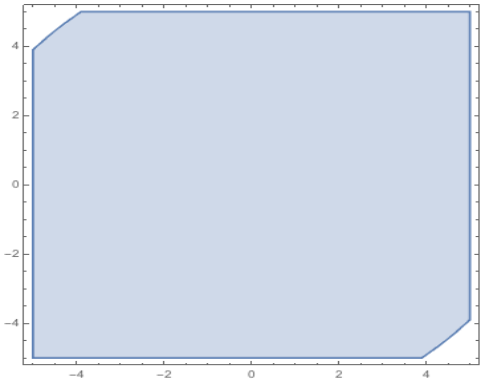


(b) Plot of constraints when $\rho = -0.75$

Figure 3.2: Plots of constraints when ρ is negative

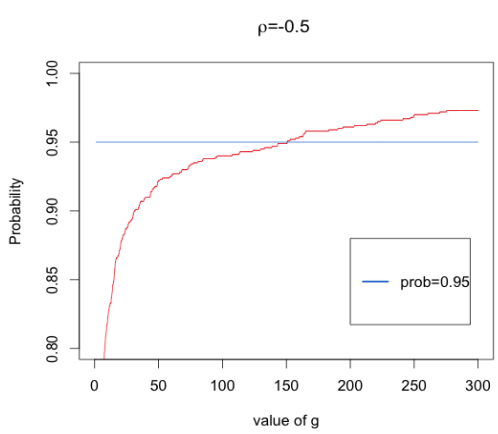


(a) Plot of constraints when $\rho = 0.5$

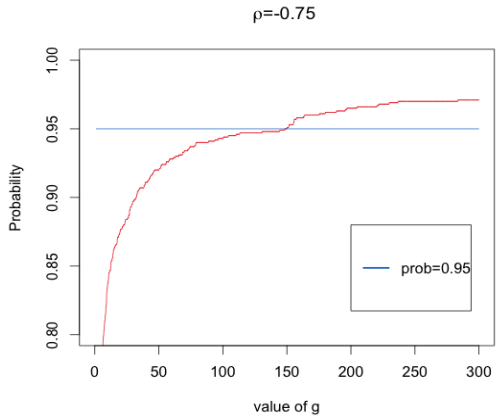


(b) Plot of constraints when $\rho = 0.2$

Figure 3.3: Plots of constraints when ρ is positive

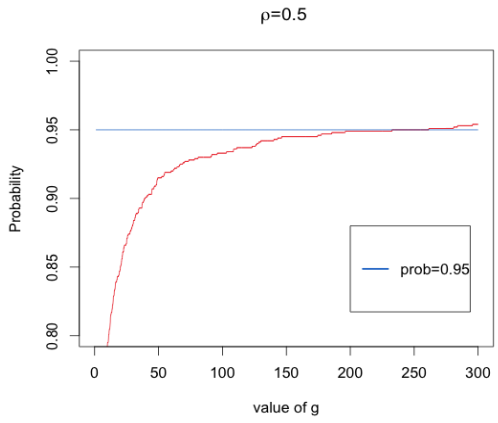


(a) Plot of g when $\rho = -0.5$

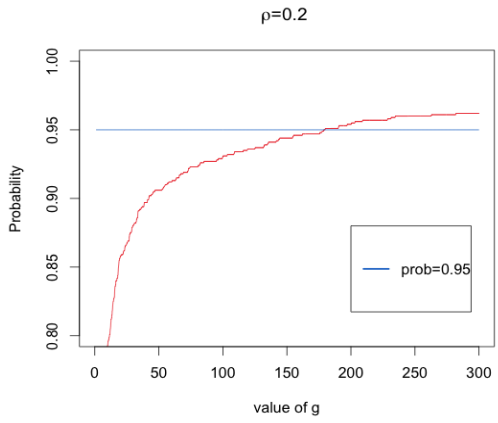


(b) Plot of g when $\rho = -0.75$

Figure 3.4: Plots of g when ρ is negative



(a) Plot of g when $\rho = 0.5$



(b) Plot of g when $\rho = 0.2$

Figure 3.5: Plots of g when ρ is positive

Chapter 4

Data Analysis and Conclusion

4.1 Data Analysis

In this part, the technique of selecting g proposed in the previous chapter will be used accordingly to do data analysis. I use the Boston housing dataset which is from UC Irvine Machine Learning Repository. The response variable is Median value of owner-occupied homes in $\$1000's$. Two predictors are weighted distances to five Boston employment centers and crime rate by town respectively. Equal prior probability is assigned to each model. After normalizing the predictors, ρ is found to be 0.169 and σ^2 is estimated to be 155.46 by OLS estimation. Since g depends on ρ and σ, we get $g = 177.1$ through simulation by making the α to be 0.05. We can see from the following Figure (4.1).

The Bayes factor is the same as posterior odds. The log ratio of it is calculated by plugging in 177.1 for g. $\mathcal{M}_1$ is the basic model without any predictors. $\mathcal{M}_2$ is the model with predictor "crime ratio", while $\mathcal{M}_3$ is the model with predictor "weighted distance" and $\mathcal{M}_4$ is the full model. The estimate of β under Zellner's g prior is

$$\beta = (44.205, 458.239)'$$

while under OLS estimation,

$$\beta = (44.454, 460.826)'.$$

Recall that the model with the biggest posterior probability is preferred and the model with the lowest BIC is preferred. So, as we can see from the following Table (4.1), the full model is preferred in each model selection criterion. However, the comparison result between $\mathcal{M}_3$ and $\mathcal{M}_1$ has stronger evidence under the posterior probability criterion than under BIC. So does the comparison between $\mathcal{M}_2$ and $\mathcal{M}_3$. We can also see that there is not much difference when deciding to choose an appropriate model between $\mathcal{M}_4$ and $\mathcal{M}_2$.

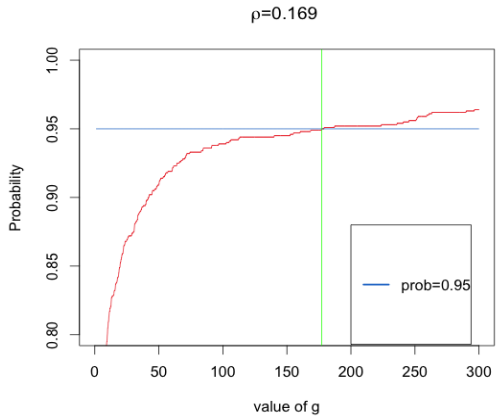


Figure 4.1: Plots of g when $\rho = 0.169, \sigma^2 = 155.46$

Table 4.1: Model comparison under different criterion

Candidate models	Log ratio of posterior probability	Ratio of BIC
$\mathcal{M}_3$&$\mathcal{M}_1$	45.40259	0.9957683
$\mathcal{M}_2$&$\mathcal{M}_3$	15.39491	0.8623403
$\mathcal{M}_4$&$\mathcal{M}_2$	1.005076	0.9984994

4.2 Conclusion

In this report, a new way of determining g is proposed. Instead of trying to assign a prior on g and maintain the good properties at the same time, g can be easily found by controlling the type I error of the test. A test is a test, it always has a type I error. It is intuitive to come up with a value for g through the type I error. We usually do not consider type I error in Bayesian hypothesis test, people just calculate the Bayes factor and make conclusions based on that. Here I incorporate the type I error and calculate the Bayes factor. It is a mixture of Bayesian and frequentist methods. For convenience, only two predictors are allowed in the linear regression equation. However, simulation speed may go down when dealing with more and more predictors; when σ^2 is unknown, a prior on it should be given.

Appendices

Appendix A

Derivation of Posterior Probability for β

For (2.1)

$$
\begin{aligned}
&P(\beta|\sigma^2, X, Y) \propto P(Y|X, \beta, \sigma^2) \cdot P(\beta|\sigma^2, X) \\
&\propto \exp\left[-\frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^2}\right] \cdot \exp\left[-\frac{(\beta - \beta_0)'(X'X)(\beta - \beta_0)}{2\sigma^2 g}\right] \\
&\propto \exp\left[-\frac{\beta' X'X\beta - 2\beta' X'Y + \frac{1}{g}\beta' X'X\beta - \frac{2}{g}\beta_0 X'X\beta}{2\sigma^2}\right] \\
&\propto \exp\left[-\frac{1}{2}\left[\beta - \frac{(\beta_0 + g(X'X)^{-1}X'Y)}{1+g}\right]' \frac{X'X}{\frac{\sigma^2 g}{1+g}}\left[\beta - \frac{(\beta_0 + g(X'X)^{-1}X'Y)}{1+g}\right]\right].
\end{aligned}
$$

So

$$
\beta|\sigma^2, X, Y \sim \mathcal{N}\left(\frac{\beta_0 + g(X'X)^{-1}X'Y}{1+g}, \frac{X'X}{\frac{\sigma^2 g}{1+g}}\right).
$$

Appendix B

Derivation of BIC

For (2.4)

$$\begin{aligned} BIC_1 &= -2 \cdot \log(\hat{\boldsymbol{L}}) + 0 \cdot \log(n) \\ &= -2 \cdot \log\left[\left(\frac{1}{\sqrt{2\pi}\cdot\sigma}\right)^n \cdot \exp\left(-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right)\right] \\ &= -2 \cdot \left[-\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{\sum i = 1^n y_i^2}{2\sigma^2}\right] \\ &= n\log(2\pi) + 2n\log(\sigma) + \frac{\sum_{i=1}^n y_i^2}{\sigma^2}. \end{aligned}$$

For (2.5)

$$\begin{aligned} BIC_2 &= -2 \cdot \log(\hat{\boldsymbol{L}}) + 1 \cdot \log(n) \\ &= -2 \cdot \log\left[\left(\frac{1}{\sqrt{(2\pi)}\cdot\sigma}\right)^n \cdot \exp\left(-\frac{\sum_{i=1}^n (y_i - \hat{\beta}_1 x_{2i})^2}{2\sigma^2}\right)\right] + \log(n). \end{aligned} \quad \text{(B.1)}$$

The oridnary least square estimate from (1.1) for β_1 is

$$\hat{\beta}_1 = (X_1' X_1)^{-1} \cdot X_1' \cdot Y = \frac{\sum_{i=1}^n x_{1i} y_i}{\sum_{i=1}^n x_{1i}^2},$$

which is equivalent to the maximized likelihood estimate for β_1, see (Christensen 2011). Plug it into (B.1), we have

$$BIC_2 = n\log(2\pi) + 2n\log(\sigma) + \log(n) + \frac{\sum_{i=1}^{n}\left(y_i - \frac{\sum_{i=1}^{n} x_{1i}y_i}{\sum_{i=1}^{n} x_{1i}^2}\cdot x_{1i}\right)^2}{\sigma^2}$$
$$= n\log(2\pi) + 2n\log(\sigma) + \log(n) + \frac{\sum_{i=1}^{n} y_i^2 - \frac{2\cdot a^2}{\sum_{i=1}^{n} x_{1i}^2} + \sum_{i=1}^{n} x_{1i}^2 \cdot \frac{a^2}{(\sum_{i=1}^{n} x_{1i}^2)^2}}{\sigma^2}$$
$$= n\log(2\pi) + 2n\log(\sigma) + \log(n) + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} - \frac{a^2}{\sum_{i=1}^{n} x_{1i}^2\cdot\sigma^2}$$
$$= n\log(2\pi) + 2n\log(\sigma) + \log(n) + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} - \frac{a^2}{\sigma^2}.$$

Similarly, for (2.6),

$$BIC_3 = -2\cdot\log\left[\left(\frac{1}{\sqrt{(2\pi)}\cdot\sigma}\right)^n \cdot \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_2 x_{2i})^2}{2\sigma^2}\right)\right] + \log(n)$$
$$= n\log(2\pi) + 2n\log(\sigma) + \log(n) + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} - \frac{b^2}{\sigma^2}.$$

For (2.7),

$$BIC_4 = -2\cdot\log(\hat{\boldsymbol{L}}) + 2\cdot\log(n) \tag{B.2}$$
$$= n\log(2\pi) + 2n\log(n) + \frac{\sum_{i=1}^{n}(y_i - x_{1i}\cdot\hat{\beta}_1 - x_{2i}\cdot\hat{\beta}_2)^2}{\sigma^2} + 2\log(n).$$

Notice that

$$X'X = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$
$$(X'X)^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$
$$|(X'X)^{-1}| = \frac{1}{|X'X|} = \frac{1}{1-\rho^2}.$$

So

$$\hat{\beta}_1 = \sum_{i=1}^{n} (\frac{1}{1-\rho^2} x_{i1} y_i - \frac{\rho}{1-\rho^2} x_{i2} y_i),$$

and

$$\hat{\beta}_2 = \sum_{i=1}^{n} (\frac{1}{1-\rho^2} x_{i2} y_i - \frac{\rho}{1-\rho^2} x_{i1} y_i).$$

(B.2) can be rewritten as

$$\begin{aligned} BIC_4 &= n\log(2\pi) + 2n\log(n) + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} + \frac{\left[\sum_{i=1}^{n} (\frac{1}{1-\rho^2} x_{i1} y_i - \frac{\rho}{1-\rho^2} x_{i2} y_i)\right]^2}{\sigma^2} + \\ &\frac{\left[\sum_{i=1}^{n} (\frac{1}{1-\rho^2} x_{i2} y_i - \frac{\rho}{1-\rho^2} x_{i1} y_i)\right]^2}{\sigma^2} - \frac{2\sum_{i=1}^{n} x_{i1} y_i \left[\sum_{i=1}^{n} (\frac{1}{1-\rho^2} x_{i1} y_i - \frac{\rho}{1-\rho^2} x_{i2} y_i)\right]}{\sigma^2} \\ &- \frac{2\sum_{i=1}^{n} x_{2i} y_i \left[\sum_{i=1}^{n} (\frac{1}{1-\rho^2} x_{i2} y_i - \frac{\rho}{1-\rho^2} x_{i1} y_i)\right]}{\sigma^2} + 2\log(n) \\ &= n\log(2\pi) + 2n\log(n) + 2\log(n) + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} - \frac{a^2 - 2ab\rho + b^2}{(1-\rho^2)\sigma^2}. \end{aligned}$$

Appendix C

Derivation of Posterior Probability for Different Models

For (2.2)

$$
\begin{aligned}
& p(\mathcal{M}_2|Y) \propto p(\mathcal{M}_2) \cdot p(Y|\mathcal{M}_2) \\
\propto & P(\mathcal{M}_2) \cdot \int \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n}(y_i - x_{1i}\beta_1)^2\right] \cdot \\
& \exp\left[-\frac{1}{2\sigma^2}\beta_1^2 \cdot \frac{1}{g}\right] \cdot \frac{1}{\sqrt{2\pi}\sigma\sqrt{g}}\, d\beta \\
\propto & p(\mathcal{M}_2) \cdot \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sqrt{2\pi}\sigma\sqrt{g}} \cdot \int \exp\left[\frac{1}{2\sigma^2} \cdot (1+\frac{1}{g}) \cdot (\beta_1 - \frac{a}{1+\frac{1}{g}})^2\right] d\beta \cdot \\
& \exp\left[\frac{1}{2\sigma^2} \cdot \frac{a^2}{1+\frac{1}{g}}\right] \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right] \\
= & p(\mathcal{M}_2) \cdot \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sqrt{2\pi}\sigma\sqrt{g}} \cdot \frac{\sqrt{2\pi}\sigma}{\sqrt{1+\frac{1}{g}}} \cdot \exp\left[\frac{1}{2\sigma^2} \cdot \frac{a^2}{1+\frac{1}{g}}\right] \\
& \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right] \\
= & p(\mathcal{M}_2) \cdot \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sqrt{1+g}} \cdot \exp\left[\frac{1}{2\sigma^2}\frac{a^2}{1+\frac{1}{g}}\right] \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right].
\end{aligned}
$$

For (2.3)

$$
\begin{aligned}
&p(\mathcal{M}_4|Y) \propto p(\mathcal{M}_4)\cdot p(Y|\mathcal{M}_4)\\
=&p(\mathcal{M}_4)\cdot\int\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\cdot\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-x_{1i}\beta_1-x_{2i}\beta_2)^2\right]\\
&\cdot\exp\left[-\frac{1}{2\sigma^2}\frac{\beta'(X'X)\beta}{g}\right]\cdot\frac{1}{2\pi\sqrt{(\sigma^2g)^2|(X'X)^{-1}|}}\,d\beta\\
=&P(\mathcal{M}_4)\cdot\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\cdot\int\exp\left[-\frac{[(Y-X\beta)'(Y-X\beta)+\frac{\beta'(X'X)\beta}{g}]}{2\sigma^2}\right]d\beta\\
&\cdot\frac{1}{2\pi(\sigma^2g)\sqrt{|(X'X)^{-1}|}}\\
=&P(\mathcal{M}_4)\cdot\int\exp\left[-\frac{[Y'Y-Y'X\beta-\beta'X'Y+\beta'X'X\beta+\frac{1}{g}\beta'X'X\beta]}{2\sigma^2}\right]d\beta\\
&\cdot\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\cdot\frac{1}{2\pi\sqrt{(\sigma^2g)^2|(X'X)^{-1}|}}\\
=&P(\mathcal{M}_4)\cdot\int\exp\left[-\frac{(1+\frac{1}{g})(\beta-\frac{(X'X)^{-1}X'Y}{1+\frac{1}{g}})'(X'X)(\beta-\frac{(X'X)^{-1}X'Y}{1+\frac{1}{g}})}{2\sigma^2}\right]d\beta\\
&\cdot\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\cdot\exp\left[\frac{1}{2\sigma^2}\frac{Y'X(X'X)^{-1}X'Y}{1+\frac{1}{g}}\right]\cdot\frac{1}{2\pi\sigma^2g\sqrt{|(X'X)^{-1}|}}\\
&\cdot\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}y_i^2\right]\\
=&P(\mathcal{M}_4)\cdot\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\cdot\frac{1}{1+g}\cdot\exp\left[\frac{1}{2\sigma^2}\frac{Y'X(X'X)^{-1}X'Y}{1+\frac{1}{g}}\right]\\
&\cdot\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}y_i^2\right]\\
\propto&p(\mathcal{M}_4)\cdot\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\cdot\frac{1}{1+g}\cdot\exp\left[\frac{1}{2\sigma^2}\cdot\frac{a^2+b^2-2ab\rho}{(1-\rho^2)(1+\frac{1}{g})}\right]\\
&\cdot\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}y_i^2\right].
\end{aligned}
$$

Bibliography

[1] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[2] Aurélie Boisbunon, Stephane Canu, Dominique Fourdrinier, William Strawderman, and Martin T Wells. Aic, cp and estimators of loss for elliptically symmetric distributions. *arXiv preprint arXiv:1308.2766*, 2013.

[3] Ronald Christensen. *Plane answers to complex questions: the theory of linear models*. Springer Science & Business Media, 2011.

[4] MA Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, 1:191–203, 1960.

[5] Steve Geinitz. Prior covariance choices and the g prior, 2009.

[6] EdwardI George and Dean P Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

[7] Joseph F Hair. Multivariate data analysis. 2010.

[8] Joseph B Kadane and Nicole A Lazar. Methods and criteria for model selection. *Journal of the American Statistical Association*, 99(465):279–290, 2004.

[9] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[10] Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481), 2008.

[11] M. Lichman. UCI machine learning repository, 2013.

[12] C. L. Mallows. Choosing variables in a linear regression: a graphical aid. 1964.

[13] Cohn L Mallows. More comments on cp. *Technometrics*, 37(4):362–372, 1995.

[14] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[15] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[16] Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243, 1986.