Large Sample Property of The Bayes Factor in a Spline Semiparametric Regression Model

Ameera Jaber Mohaisen and Ammar Muslim Abdulhussain
Mathematics Department College of Education for Pure Science AL-Basrah University-Iraq
E-mail: ammar.muslim1@yahoo.com

Abstract

In this paper, we consider semiparametric regression model where the mean function of this model has two part, the first is the parametric part is assumed to be linear function of p-dimensional covariates and nonparametric (second part) is assumed to be a smooth penalized spline. By using a convenient connection between penalized splines and mixed models, we can representation semiparametric regression model as mixed model. In this model, we investigate the large sample property of the Bayes factor for testing the polynomial component of spline model against the fully spline semiparametric alternative model. Under some conditions on the prior and design matrix, we identify the analytic form of the Bayes factor and show that the Bayes factor is consistent.

Keywords: Mixed Models, Semiparametric Regression Model, Penalized Spline, Bayesian Model, Marginal Distribution, Prior Distribution, Posterior Distribution, Bayes Factor, Consistent.

1. Introduction

In many applications in different fields, we need to use one of a collection of models for correlated data structures, for example, multivariate observations, clustered data, repeated measurements, longitudinal data and spatially correlated data. Often random effects are used to describe the correlation structure in clustered data, repeated measurements and longitudinal data. Models with both fixed and random effects are called mixed models. The general form of a linear mixed model for the i^{th} subject (i = 1,…, n) is given as follows (see [10,13,15]),

$$Y_i = X_i\beta + \sum_{j=1}^{r} Z_{ij}u_{ij} + \epsilon_i \quad , \quad u_{ij} \sim N(0, G_j)\,,\ \epsilon_i \sim N(0, R_i) \tag{1}$$

where the vector Y_i has length m_i, X_i and Z_{ij} are, respectively, a $m_i \times p$ design matrix and a $m_i \times q_i$ design matrix of fixed and random effects. β is a p-vector of fixed effects and u_{ij} are the q_i-vectors of random effects. The variance matrix G_j is a $q_i \times q_i$ matrix and R_i is a $m_i \times m_i$ matrix.

We assume that the random effects $\{u_{ij}\,;\ i = 1, \dots, n\,;\ j = 1, \dots, r\}$ and the set of error terms $\{\epsilon_1, \dots, \epsilon_n\}$ are independent. In matrix notation (see [13,15]),

$$Y = X\beta + Zu + \epsilon \tag{2}$$

Here $Y = (Y_1, \dots, Y_n)^T$ has length $N = \sum_{i=1}^{n} m_i$, $X = (X_1^T, \dots, X_n^T)^T$ is a $N \times p$ design matrix of fixed effects, Z is a $N \times q$ block diagonal design matrix of random effects, $q = \sum_{j=1}^{r} q_j$, $u = (u_1^T, \dots, u_r^T)^T$ is a q-vector of random effects, $R = diag(R_1, \dots, R_n)$ is a $N \times N$ matrix and $G = diag(G_1, \dots, G_r)$ is a $q \times q$ block diagonal matrix. In this paper, we consider semiparametric regression model (see [1,6,8,9,10,13,15]), for which the mean function has two part, the parametric (first part) is assumed to be linear function of p-dimensional covariates and nonparametric (second part) is assumed to be a smooth penalized spline. By using a convenient connection between penalized splines and mixed models, we can representation semiparametric regression model as mixed model. In this model we investigate large sample properties of the Bayes factor for testing the pure polynomial component of spline null model whose mean function consists of only the polynomial component against the fully spline semiparametric alternative model whose mean function comprises both the pure polynomial and the component spline basis functions. The asymptotic properties of the Bayes factor in nonparametric or semiparametric models have been studied mainly in nonparametric density estimation problems related to goodness of fit testing. These theoretical results include (see [4,7,12,14]). Compared to the previous approaches to density estimations problems for goodness of fit testing and model selection, little work has been done on nonparametric regression problems. Choi and et al 2009 studied the semiparametric additive regression models as the encompassing model with algebraic smoothing and obtained the closed form of the Bayes factor and studied the asymptotic behavior of the Bayes factor based on the closed form (see [3]).

In this paper, we obtain the closed form and studied the asymptotic behavior of the Bayes factor in spline semiparametric regression model and we proved that the Bayes factor converges to infinity under the pure polynomial model and the Bayes factor converges to zero almost surely under the spline semiparametric regression alternative and we show that the Bayes factor is consistent.

2. Semiparametric and Penalized Spline

Consider the model:

$$y_i = \sum_{j=0}^{p} \beta_j x_{ji} + m(x_{p+1,i}) + \epsilon_i \qquad , i = 1,2, \ldots, n \tag{3}$$

Where $y_1, \ldots, y_n$ response variables and the unobserved errors are $\epsilon_1, \ldots, \epsilon_n$ are known to be i.i.d. normal with mean 0 and covariance $\sigma_\epsilon^2 I$ with σ_ϵ^2 known.

The mean function of the regression model in (3) has two parts. The parametric (first part) is assumed to be linear function of p-dimensional covariates x_{ji} and nonparametric (second part) $m(x_{p+1,i})$ is function defined on some index set $T \subset R^1$.

The model (3) can be expressed as a smooth penalized spline with q degree, then it's become as:

$$y_i = \sum_{j=0}^{p} \beta_j x_{ji} + \sum_{j=1}^{q} \beta_{p+j} x_{p+1,i}^{j} + \sum_{k=1}^{K} \{u_k (x_{p+1,i} - k_k)_+^q + \epsilon_i \tag{4}$$

where $k_1, \ldots, k_K$ are inner knots $a < k_1 <,,,< k_K < b$.

By using a convenient connection between penalized splines and mixed models. Model (4) is rewritten as follows (see [13,15])

$$Y = X\beta + Zu + \epsilon \tag{5}$$

where

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \\ \beta_{p+1} \\ \vdots \\ \beta_{p+q} \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}, Z = \begin{bmatrix} (x_{p+1,1} - k_1)_+^q & \cdots & (x_{p+1,1} - k_K)_+^q \\ \vdots & \ddots & \vdots \\ (x_{p+1,n} - k_1)_+^q & \cdots & (x_{p+1,n} - k_K)_+^q \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{p1} & x_{p+1,1} & \cdots & x_{p+1,1}^q \\ 1 & x_{12} & \ldots & x_{p2} & x_{p+1,2} & \cdots & x_{p+1,2}^q \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \ldots & x_{pn} & x_{p+1,n} & \cdots & x_{p+1,n}^q \end{bmatrix}$$

We assume $n < k + p + q + 1, n + s = k + p + q + 1$, ($s$ is integer number and greater than 1), and we assume that the design matrix is orthogonal in the following way:

$$M_n^T M_n = nI_n \qquad \forall n \geq 1 \tag{6}$$

where:

$M_n = (X , Z_r)$ and Z_r be the $n - (p + q + 1) \times n - (p + q + 1)$ matrix, and let $Z = Z_r + Z_s$

Based on the above setup for regression model (5) and the assumption of the orthogonal design matrix (6), we consider a Bayesian model selection problem in a spline semiparametric regression problem. Specifically, we would like to choose between a Bayesian spline semiparametric model and its pure polynomial counterpart by the criterion of the Bayes factor for two hypotheses,

$$H_0: Y = X\beta + \epsilon \qquad \text{versus} \qquad H_1: Y = X\beta + Zu + \epsilon \tag{7}$$

As for the prior of $\beta, and\ \epsilon$ under H_1, we assume β and u are independent and

$$\beta \sim N(0, \textstyle\sum_0) \quad , \quad \sum_0 = \sigma_\beta^2 I_{p+q+1} \tag{8}$$

$$u \sim N(0, \textstyle\sum_1) \quad , \quad \sum_1 = \sigma_u^2 I_K \tag{9}$$

3. Bayes factor and marginal distribution

The response y_i in (4) follows the normal distribution with mean η_i and variance σ_ϵ^2 , where $\eta_i = \sum_{j=0}^{p} \beta_j x_{ji} + \sum_{j=1}^{q} \beta_j x_{p+1,i}^{j} + \sum_{k=1}^{K} \{u_k (x_{p+1,i} - K_k)_+^q$, for, $i = 1,2, \ldots, n$. Thus, given covariates and $\eta_n = (\eta_1, \ldots, \eta_n)^T$, the n-dimensional response vector $Y_n = (y_1, \ldots, y_n)^T$ follows the n-dimensional normal distribution with mean η_n and covariance matrix $\sigma_\epsilon^2 I_n$ where I_n is the $n \times n$ identity matrix. Also from the prior distributions specified in (8) and (9), we can deduce the joint distribution of η_n is the multivariate normal distribution (MVN) with mean zero and $n \times n$ covariance matrix $X \sum_0 X^T + Z \sum_1 Z^T$.

In summary, we have the following:-

$$Y_n|\eta_n \sim MVN_n(\eta_n, \sigma_\epsilon^2 I_n) , \quad \eta_n \sim MVN_n(0_n, X\Sigma_0 X^T + Z\Sigma_1 Z^T) \quad (10)$$

Result 1:

Suppose the distribution of η_n and Y_n are given by (10). Then the posterior distribution of $P(\eta_n|Y_n)$is the multivariate normal with the following mean and variance.

$E(\eta_n|Y_n, \sigma_\epsilon^2) = \Sigma_2(\Sigma_2 + \sigma_\epsilon^2 I_n)^{-1} Y_n$

$var(\eta_n|Y_n, \sigma_\epsilon^2) = (\Sigma_2^{-1} + (\sigma_\epsilon^2 I_n)^{-1})^{-1} = \Sigma_2(\Sigma_2 + \sigma_\epsilon^2 I_n)^{-1}\sigma_\epsilon^2 I_n$

Where $\Sigma_2 = X\Sigma_0 X^T + Z\Sigma_1 Z^T$. Furthermore, marginally Y_n follows

$$Y_n \sim MVN(0, \sigma_\epsilon^2 I_n + \Sigma_2) \quad (11)$$

Note the (i,j)th element of Σ_2,

$$\Sigma_2(i,j) = \sum_{k=1}^{p+q+1} \sigma_\beta^2 x_{ik} x_{kj} + \sum_{k=1}^{K} \sigma_u^2 z_{ik} z_{kj} \quad , i,j = 1, \dots, n \quad (12)$$

where $z_{ik} = (x_{p+1,i} - k_k)_+^q$

Applications of the previous result yield that, the marginal distribution of Y_n under H_0 and H_1 are respectively $N_n(0, V_0)$ and $N_n(0, V_1)$, where

$V_0 = \sigma_\epsilon^2 I_n + X\Sigma_0 X^T$

$V_1 = \sigma_\epsilon^2 I_n + X\Sigma_0 X^T + Z\Sigma_1 Z^T$

Hence, the Bayes factor for testing problem (7) is given by:

$$B_{01} = \frac{p(Y|H_0)}{p(Y|H_1)} = \frac{\frac{1}{\sqrt{2\pi \det(V_0)}} \exp\{-\frac{1}{2} Y_n^T V_0^{-1} Y_n\}}{\frac{1}{\sqrt{2\pi \det(V_1)}} \exp\{-\frac{1}{2} Y_n^T V_1^{-1} Y_n\}}$$

$$= \frac{\sqrt{\det(V_1)}}{\sqrt{\det(V_0)}} \exp\{-\frac{1}{2} Y_n^T (V_0^{-1} - V_1^{-1}) Y_n\} \quad (13)$$

4. Asymptotic behavior of Bayes factor

The Bayes factor is said to be consistent if (see[3,14])

$$lim_{n\to\infty} B_{01} = \infty \quad , \quad a.s. \quad P_0^\infty \text{ under the } H_0, \quad (14)$$

$$lim_{n\to\infty} B_{01} = 0 \quad , \quad a.s. \quad P_1^\infty \text{ under the } H_1, \quad (15)$$

We investigate the consistency of the Bayes factor, by establishing (14) and (15), under the spline semiparametric regression model described in the section 2. By truncating the regression up to the first n terms. We get $y_i = \sum_{j=0}^{p} \beta_j x_{ji} + \sum_{j=1}^{q} \beta_{p+j} x_{p+1,i}^j + \sum_{k=1}^{n-(p+q+1)} \{u_k (x_{p+1,i} - k_k)_+^q\} + \epsilon_i$

Now define $V_{1,n}$ as the covariance matrix of the marginal distribution of Y_n under H_1 as:-

$$V_{1,n} = \sigma_\epsilon^2 I_n + X\Sigma_0 X^T + Z_r \Sigma_{1,r} Z_r^T$$

$$= \sigma_\epsilon^2 I_n + M_n D_n M_n^T \quad (16)$$

Where $\Sigma_{1,r}$ is diagonal matrix with diagonal elements σ_u^2 from size $n-(p+q+1) \times n-(p+q+1)$ and

$$D_n = \begin{bmatrix} \Sigma_0 & 0_{r1} \\ 0_{1r} & \Sigma_{1,r} \end{bmatrix}$$

where 0_{r1}, 0_{1r} are the matrices for zeros element with size $(p+q+1) \times n-(p+q+1)$ and $n-(p+q+1) \times (p+q+1)$, respectively.

Then, we can rewrite (16) by using D_n as:

$$V_{1,n} = M_n \left[\frac{\sigma_\epsilon^2}{n} I_n + D_n \right] M_n^T \quad (17)$$

Let V_0 be the covariance matrix of the marginal distribution of Y_n under H_0, then V_0 can be represented as:

$$V_0 = \sigma_\epsilon^2 I_n + X\Sigma_0 X^T + Z_r 0_{rr} Z_r^T$$

$$= \sigma_\epsilon^2 I_n + M_n C_n M_n^T$$

where:

$$C_n = \begin{bmatrix} \Sigma_0 & 0_{r1} \\ 0_{1r} & 0_{rr} \end{bmatrix}$$

where 0_{rr} is the matrix for zeros element with size $n-(p+q+1) \times n-(p+q+1)$, then we can rewrite V_0 as the following:

$$V_0 = M_n \left[\frac{\sigma_\epsilon^2}{n} I_n + C_n \right] M_n^T \quad (18)$$

Then we can use the two matrices above (V_0 , $V_{1,n}$) for matrix inversions and calculating determinants as shown in following lemma.

Lemma1: Let $V_{1,n}$ and V_0 be defined as in (17) and (18) respectively, and suppose M_n satisfied (6). Then, the following hold.

i) $M_n M_n^T = M_n^T M_n = nI_n$

ii) $\det(V_0) = n^{p+q+1}(\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2)^{p+q+1}\,{\sigma_\epsilon^2}^{(n-(p+q+1))}$

iii) $\det(V_{1,n}) = n^n\,(\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2)^{p+q+1}(\frac{\sigma_\epsilon^2}{n} + \sigma_u^2)^{n-(p+q+1)}$

iv) $V_0^{-1} = \frac{1}{n^2}\,M_n\,\left[\frac{\sigma_\epsilon^2}{n} I_n + C_n\right]^{-1} M_n^T$

v) $V_{1,n}^{-1} = \frac{1}{n^2}\,M_n\,\left[\frac{\sigma_\epsilon^2}{n} I_n + D_n\right]^{-1} M_n^T$

Proof:-

i) By multiplying M_n and M_n^T to equation (6) from right and left side, we have:-

$$\begin{aligned} M_n M_n^T M_n M_n^T &= M_n n I_n M_n^T \\ &= M_n n M_n^T \\ &= n M_n M_n^T \end{aligned}$$

Multiplying $(M_n M_n^T)^{-1}$ to the above equation from left side, we get:-

$$M_n M_n^T M_n M_n^T (M_n M_n^T)^{-1} = n M_n M_n^T (M_n M_n^T)^{-1}$$

$$M_n M_n^T = nI_n$$

ii) $V_0 = M_n\left[\frac{\sigma_\epsilon^2}{n} I_n + C_n\right] M_n^T$, and by using (6), we get

$$= [X\,,\,Z_r]\begin{bmatrix} (\frac{\sigma_\epsilon^2}{n}+\sigma_\beta^2) & 0 & \cdots & 0 & 0 \\ 0 & (\frac{\sigma_\epsilon^2}{n}+\sigma_\beta^2) & \cdots & 0 & 0 \\ & & \ddots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \frac{\sigma_\epsilon^2}{n} & 0 \\ 0 & 0 & & & \\ 0 & 0 & \cdots & 0 & \frac{\sigma_\epsilon^2}{n} \end{bmatrix}[X\,,\,Z_r]^T$$

$$= \begin{bmatrix} n(\frac{\sigma_\epsilon^2}{n}+\sigma_\beta^2) & 0 & \cdots & 0 & 0 \\ 0 & n(\frac{\sigma_\epsilon^2}{n}+\sigma_\beta^2) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_\epsilon^2 & 0 \\ 0 & 0 & \cdots & 0 & \sigma_\epsilon^2 \end{bmatrix}$$

Then

$$\det(V_0) = n^{p+q+1}\,(\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2)^{p+q+1}{\sigma_\epsilon^2}^{(n-(p+q+1))}$$

iii) $V_{1,n} = M_n\left[\frac{\sigma_\epsilon^2}{n} I_n + D_n\right] M_n^T$

$$= [X\,,\,Z_r]\begin{bmatrix} (\frac{\sigma_\epsilon^2}{n}+\sigma_\beta^2) & 0 & \cdots & 0 & 0 \\ 0 & (\frac{\sigma_\epsilon^2}{n}+\sigma_\beta^2) & \cdots & 0 & 0 \\ & & \ddots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \frac{\sigma_\epsilon^2}{n}+\sigma_u^2 & 0 \\ 0 & 0 & & & \\ 0 & 0 & \cdots & 0 & \frac{\sigma_\epsilon^2}{n}+\sigma_u^2 \end{bmatrix}[X\,,\,Z_r]^T$$

$$= \begin{bmatrix} n(\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2) & 0 & \cdots & 0 & 0 \\ 0 & n(\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2) & \cdots & 0 & 0 \\ & & \ddots & \vdots & \vdots \\ \vdots & \vdots & \cdots & n(\frac{\sigma_\epsilon^2}{n} + \sigma_u^2) & 0 \\ 0 & 0 & & & \\ 0 & 0 & \cdots & 0 & n(\frac{\sigma_\epsilon^2}{n} + \sigma_u^2) \end{bmatrix}$$

Then

$$\det(V_{1,n}) = n^n\, (\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2)^{p+q+1}(\frac{\sigma_\epsilon^2}{n} + \sigma_u^2)^{n-(p+q+1)}$$

iv) $V_0^{-1} = \{M_n \left[\frac{\sigma_\epsilon^2}{n} I_n + C_n\right] M_n^T\}^{-1}$

$$= M_n^{T^{-1}} \left[\frac{\sigma_\epsilon^2}{n} I_n + C_n\right]^{-1} M_n^{-1}$$

$$= \frac{I_n^{-1}}{n} nI_n\, M_n^{T^{-1}} \left[\frac{\sigma_\epsilon^2}{n} I_n + C_n\right]^{-1} M_n^{-1} nI_n \frac{I_n^{-1}}{n}$$

$$= \frac{1}{n^2} M_n M_n^T\, M_n^{T^{-1}} \left[\frac{\sigma_\epsilon^2}{n} I_n + C_n\right]^{-1} M_n^{-1} M_n M_n^T \quad \text{by lemma 1.i}$$

$$= \frac{1}{n^2}\, M_n \left[\frac{\sigma_\epsilon^2}{n} I_n + C_n\right]^{-1} M_n^T$$

v) $V_1^{-1} = \{M_n \left[\frac{\sigma_\epsilon^2}{n} I_n + D_n\right] M_n^T\}^{-1}$

$$= M_n^{T^{-1}} \left[\frac{\sigma_\epsilon^2}{n} I_n + D_n\right]^{-1} M_n^{-1}$$

$$= \frac{I_n^{-1}}{n} nI_n\, M_n^{T^{-1}} \left[\frac{\sigma_\epsilon^2}{n} I_n + D_n\right]^{-1} M_n^{-1} nI_n \frac{I_n^{-1}}{n}$$

$$= \frac{1}{n^2} M_n M_n^T\, M_n^{T^{-1}} \left[\frac{\sigma_\epsilon^2}{n} I_n + D_n\right]^{-1} M_n^{-1} M_n M_n^T$$

$$= \frac{1}{n^2}\, M_n \left[\frac{\sigma_\epsilon^2}{n} I_n + D_n\right]^{-1} M_n^T$$

Now let

$$\tilde{B}_{01} = \frac{\sqrt{\det(V_{1,n})}}{\sqrt{\det(V_0)}} \exp\{-\frac{1}{2} Y_n^T (\,V_0^{-1} - V_{1,n}^{-1}) Y_n\} \tag{19}$$

Which an approximation of B_{01} with V_1 replaced by $V_{1,n}$ in (13)

Thus,

$$log\ \tilde{B}_{01} = \frac{1}{2}\ log \frac{\det(V_{1,n})}{\det(V_0)} - \frac{1}{2} Y_n^T (\,V_0^{-1} - V_{1,n}^{-1}) Y_n$$

$$2 \log \tilde{B}_{01} = \log \frac{\det(V_{1,n})}{\det(V_0)} - Y_n^T (\,V_0^{-1} - V_{1,n}^{-1}) Y_n$$

By results of lemma1 we have:-

$$2 \log \tilde{B}_{01} = \log \frac{n^n\, (\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2)^{p+q+1}(\frac{\sigma_\epsilon^2}{n} + \sigma_u^2)^{n-(p+q+1)}}{n^{p+q+1}\, (\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2)^{p+q+1} \sigma_\epsilon^{2\,(n-(p+q+1))}}$$

$$- \frac{1}{n^2}\left\{Y_n^T M_n \left\{\left[\frac{\sigma_\epsilon^2}{n} I_n + C_n\right]^{-1} - \left[\frac{\sigma_\epsilon^2}{n} I_n + D_n\right]^{-1}\right\} M_n^T Y_n\right\}$$

$$= \log \frac{n^{n-(p+q+1)}\, (\frac{\sigma_\epsilon^2}{n} + \sigma_u^2)^{n-(p+q+1)}}{\sigma_\epsilon^{2\,(n-(p+q+1))}} - \frac{1}{n^2}\{Y_n^T M_n \begin{bmatrix} 0_{(p+q+1)\times(p+q+1)} & 0_{1r} \\ 0_{r1} & Q_r \end{bmatrix}\ M_n^T Y_n\}$$

$$= \log(\frac{\sigma_\epsilon^2 + n\sigma_u^2}{\sigma_\epsilon^2})^{n-(p+q+1)} - Y_n^T Q_n Y_n\}$$

where

$$Q_n = \frac{1}{n^2} M_n \begin{bmatrix} 0_{(p+q+1)\times(p+q+1)} & 0_{1r} \\ 0_{r1} & Q_r \end{bmatrix} M_n^T$$

and $Q_r = Diag\ (\frac{n^2 \sigma_u^2}{(\sigma_\epsilon^2 + n\sigma_u^2)\sigma_\epsilon^2})$

To establish the consistency of the Bayes factor we focus on $\tilde{B}_{01}$, first and the remaining terms will be considered later. Without loss of generality we assume $\sigma_\epsilon^2 = 1$ in the remainder of this paper.

Now let

$$S_{n,1} = \sum_{i=1}^{n} \frac{n\sigma_u^2}{(1+n\sigma_u^2)} = \frac{n^2\sigma_u^2}{(1+n\sigma_u^2)} \tag{20}$$

$$S_{n,2} = \sum_{i=1}^{n} \log(1+n\sigma_u^2) = n\log(1+n\sigma_u^2) \tag{21}$$

$$S_{n,3} = \sum_{i=1}^{n} (\frac{n\sigma_u^2}{1+n\sigma_u^2})^2 = n^3(\frac{\sigma_u^2}{1+n\sigma_u^2})^2 \tag{22}$$

Lemma2: let $Y_n = (y_1, \dots, y_n)^T$ where $y_i's$ are independent normal random variable with mean $\sum_{j=0}^{p} \beta_j x_{ji} + \sum_{j=1}^{q} \beta_{p+j} x_{p+1,i}^j$ and variance $\sigma_\epsilon^2 = 1$, then there exist a positive constant C_1 such that

$\frac{1}{S_{n,1}}\left[\sum_{i=1}^{n-(p+q+1)} \log(1+n\sigma_u^2) - Y_n^T Q_n Y_n\right] > C_1$, with probability tending to 1.

This implies that, under H_0.

$$\log \tilde{B}_{01} \xrightarrow{n\to\infty} \infty, \text{ in probability.}$$

Proof:

Let $X_n = (X_1, X_2, \dots, X_n)$ are independent standard normal random variables, Note that $Y_n \overset{d}{=} X_n + X\beta$ and

$$Y_n^T Q_n Y_n = (X_n + X\beta)^T Q_n (X_n + X\beta)$$
$$= X_n^T Q_n X_n + (X\beta)^T Q_n (X\beta) + X_n^T Q_n X\beta + (X\beta)^T Q_n X_n.$$

By the orthogonality of the design matrix

$$(X\beta)^T Q_n = (X\beta)^T \frac{1}{n^2} M_n \begin{bmatrix} 0_{(p+q+1)\times(p+q+1)} & 0_{1r} \\ 0_{r1} & Q_r \end{bmatrix} M_n^T = 0.$$

Then $Y_n^T Q_n Y_n = X_n^T Q_n X_n$.

Using the expectation formula of the quadratic form given in (12), we obtain

$$0 < E(X_n^T Q_n X_n) = tr(Q_n) = \sum_{i=1}^{n} \frac{n\sigma_u^2}{(1+n\sigma_u^2)} = S_{n,1}$$

Note that

$$Q_n^2 = \frac{1}{n^4} M_n \begin{bmatrix} 0_{(p+q+1)\times(p+q+1)} & 0_{1r} \\ 0_{r1} & Q_r \end{bmatrix} M_n^T M_n \begin{bmatrix} 0_{(p+q+1)\times(p+q+1)} & 0_{1r} \\ 0_{r1} & Q_r \end{bmatrix} M_n^T$$
$$= \frac{1}{n^3} M_n \begin{bmatrix} 0_{(p+q+1)\times(p+q+1)} & 0_{1r} \\ 0_{r1} & Q_r^2 \end{bmatrix} M_n^T$$

and

$$tr(Q_n^2) = \frac{1}{n^2} \sum_{i=1}^{n} (\frac{n^2\sigma_u^2}{1+n\sigma_u^2})^2 = \sum_{i=1}^{n} (\frac{n\sigma_u^2}{1+n\sigma_u^2})^2 = S_{n,3}.$$

Similarly, using the variance formula of the quadratic form of the multivariate normal variables, we have

$$var(X_n^T Q_n X_n) = 2tr(Q_n^2) = 2\sum_{i=1}^{n} (\frac{n\sigma_u^2}{1+n\sigma_u^2})^2 = 2S_{n,3}.$$

Let $c_n = \frac{S_{n,2} - S_{n,1}}{2S_{n,1}}$. Then,

$$Pr\left\{\frac{1}{S_{n,1}}\left[\sum_{i=1}^{n-(p+q+1)} \log(1+n\sigma_u^2)^2 - Y_n^T Q_n Y_n\right] \le c_n\right\}$$
$$= Pr\left[\frac{\{Y_n^T Q_n Y_n\}}{S_{n,1}} \ge \frac{S_{n,2}}{S_{n,1}} - c_n\right]$$
$$= Pr\left[\frac{\{Y_n^T Q_n Y_n\}}{S_{n,1}} - \frac{E(Y_n^T Q_n Y_n)}{S_{n,1}} \ge \frac{S_{n,2}}{S_{n,1}} - c_n - \frac{E(Y_n^T Q_n Y_n)}{S_{n,1}}\right]$$
$$= Pr\left[\frac{1}{S_{n,1}}\{(Y_n^T Q_n Y_n) - E(Y_n^T Q_n Y_n)\} \ge \frac{S_{n,2} - E(Y_n^T Q_n Y_n)}{S_{n,1}} - c_n\right]$$
$$\le \frac{1}{S_{n,1}^2} \frac{var(Y_n^T Q_n Y_n)}{(\frac{S_{n,2}-S_{n,1}}{2S_{n,1}})^2} = \frac{8S_{n,3}}{(S_{n,2}-S_{n,1})^2} \xrightarrow{n\to\infty} 0,$$

where the last statement holds from the Chebyshev inequality.

Then $S_{n,2} - S_{n,1} \ge c_1 S_{n,1}$ for some $c_1 > 0$ and let $c_n \ge c_1/2$.

Hence, we conclude that there exists a positive constant $C_1 = c_1/2$ such that

$\frac{1}{S_{n,1}}\left[\sum_{i=1}^{n-(p+q+1)}\log(1+n\sigma_u^2)^2 - Y_n^T Q_n Y_n\right] > C_1$, with probability tending to 1.

Consequently, it follows that

$2\log \tilde{B}_{01} = \sum_{i=1}^{n-(p+q+1)}\log(1+n\sigma_u^2)^2 - Y_n^T Q_n Y_n \xrightarrow{n\to\infty} \infty$ in probability under H_0

$\log \tilde{B}_{01} \xrightarrow{n\to\infty} \infty$

Lemma3

Suppose we have the penalized spline smoothing, then

$$\frac{1}{S_{n,1}} tr\left(V_{1,n}^{-1} - V_1^{-1}\right) \xrightarrow{n\to\infty} 0$$

Proof:

Since $V_1 - V_{1,n}$ is the covariance matrix of $\eta^*(x) = \sum_{k=n-(p+q)+1}^{n+s} z_k\, u_k = Z_s u$ it is a positive definite matrix. Thus $V_{1,n}^{-1} - V_1^{-1}$ is also a positive definite matrix. Thus, $tr\left(V_{1,n}^{-1}\right) \geq tr(V_1^{-1})$, also from the matrix analogue of the Cauchy-Schwarz inequality, we have

$$tr\left(Z_s^T V_{1,n}^{-1}\right) \leq \sqrt{tr(Z_s^T Z_s) tr\left(V_{1,n}^{-2}\right)} = \|Z_s\| \left\|V_{1,n}^{-1}\right\|$$

where

$$\left\|V_{1,n}^{-1}\right\| = \{ tr\left(V_{1,n}^{-2}\right) \}^{\frac{1}{2}} = [\frac{1}{n^2}\{ \sum_{i=1}^{p+q+1}\left(\frac{\sigma_\epsilon^2}{n}+\sigma_\beta^2\right)^{-2} + \sum_{i=1}^{n-(p+q+1)}\left(\frac{\sigma_\epsilon^2}{n}+\sigma_u^2\right)^{-2} \}]^{\frac{1}{2}}$$

Note that

$$\sum_{i=1}^{p+q+1}\left(\frac{\sigma_\epsilon^2}{n}+\sigma_\beta^2\right)^{-2} = \frac{p+q+1}{\left(\frac{\sigma_\epsilon^2}{n}+\sigma_\beta^2\right)^2} = \frac{p+q+1}{\left(\frac{\sigma_\epsilon^2+n\,\sigma_\beta^2}{n}\right)^2} = \frac{n^2(p+q+1)}{(\sigma_\epsilon^2+n\,\sigma_\beta^2)^2} \leq \frac{p+q+1}{\sigma_\beta^4} = O(1)$$

And

$$\sum_{i=1}^{n-(p+q+1)}\left(\frac{\sigma_\epsilon^2}{n}+\sigma_u^2\right)^{-2} = \frac{n-(p+q+1)}{\left(\frac{\sigma_\epsilon^2}{n}+\sigma_u^2\right)^2} \leq \frac{n}{\left(\frac{\sigma_\epsilon^2}{n}+\sigma_u^2\right)^2} = \frac{n^3}{(\sigma_\epsilon^2+n\sigma_u^2)^2} \leq \frac{n^3}{(n\sigma_u^2)^2} = O(n)$$

Thus, it follows that $\left\|V_{1,n}^{-1}\right\| = \left\{O\left(\frac{1}{n}\right)\right\}^{1/2} = O(n)^{-1/2}$

Since $\sup_{x\in[0,1]}\|Z_s\| < 1$ from the distances between truncated power functions, it follows that $\|Z_s\|^2 = \sum_{i=1}^{n} Z_{sk}^2(x_i) \leq n$, where $Z_{sk}^2(x_i) = (x_{p+1,i} - k_k)_+^q$ is element in Z_s

Thus, $\frac{1}{S_{n,1}} tr\left(V_{1,n}^{-1} - V_1^{-1}\right) = \frac{1}{S_{n,1}} tr\left(V_{1,n}^{-1}(V_1 - V_{1,n})\, V_1^{-1}\right)$

$$= \frac{1}{S_{n,1}}\sum_{k=n-(p+q+1)}^{n+s}\sigma_u^2 (Z_k^T\ V_{1,n}^{-1}(Z_k^T\, V_1^{-1})^T)$$

$$\leq \frac{\sigma_u^2}{S_{n,1}}\left\|V_{1,n}^{-1}\right\|^2 \sum_{k=n-(p+q+1)}^{n+s}\|Z_k\|^2$$

$$\leq \frac{n\sigma_u^2}{n\,S_{n,1}}\sum_{k=n-(p+q+1)}^{n+s} 1 \leq \frac{K\,\sigma_u^2}{S_{n,1}} = \frac{K\sigma_u^2}{\frac{n^2\sigma_u^2}{(1+n\sigma_u^2)}} = \frac{K(1+n\sigma_u^2)}{n^2}$$

Thus

$$\frac{1}{S_{n,1}} tr\left(V_{1,n}^{-1} - V_1^{-1}\right) \leq \frac{K(1+n\sigma_u^2)}{n^2} \xrightarrow{n\to\infty} 0$$

Theorem(1):

Consider the testing problem in (7), suppose that the true model is H_0 (pure polynomial model) and let P_0^n denote the true distribution of the whole data, with p.d.f $P_0(y|\beta_0) = \emptyset\{(y - X^T\beta_0)/\sigma_\epsilon\}$, where $\emptyset(\cdot)$ is the standard normal density, then, the Bayes factor is consistent under the null hypothesis H_0:

$\lim_{n\to\infty} B_{01} = \infty$ in P_0^n probability.

Proof:

$$\log B_{01} = \frac{1}{2}\log\frac{\det(V_1)}{\det(V_0)} - \frac{1}{2}Y_n^T(V_0^{-1} - V_1^{-1})Y_n$$

$$= \tfrac{1}{2}\log\frac{\det(V_1)}{\det(V_{1,n})} + \tfrac{1}{2}\log\frac{\det(V_{1,n})}{\det(V_0)} - \tfrac{1}{2}Y_n^T(V_0^{-1}-V_{1,n}^{-1})Y_n \; - \tfrac{1}{2}Y_n^T(V_{1,n}^{-1}-V_1^{-1})Y_n$$

$$= \log\tilde{B}_{01} + \tfrac{1}{2}\log\frac{\det(V_1)}{\det(V_{1,n})} \; - \tfrac{1}{2}Y_n^T(V_{1,n}^{-1}-V_1^{-1})Y_n$$

$$= \log\tilde{B}_{01} + A_1 + A_2$$

where

$$A_1 = \tfrac{1}{2}\log\frac{\det(V_1)}{\det(V_{1,n})} \quad , \quad A_2 = -\tfrac{1}{2}Y_n^T(V_{1,n}^{-1}-V_1^{-1})Y_n$$

Since $V_1 - V_{1,n}$ is a positive definite matrix, $det\,(V_1) \;\geq\; det(V_{1,n})$, thus $A_1 \geq 0$

$E(\,Y_n^T\big(V_{1,n}^{-1}-V_1^{-1}\big)Y_n\,) = \sigma_\epsilon^2\, tr\big(\,V_{1,n}^{-1} - V_1^{-1}\big) + (X\beta)^T\big(\,V_{1,n}^{-1} - V_1^{-1}\big)\, X\beta,$

since $V_{1,n}^{-1} - V_1^{-1}$, $V_{1,n}^{-1}$ and V_1^{-1} are all nonnegative definite matrices, it follows that

$0 \leq\; (X\beta)^T(V_{1,n}^{-1} - V_1^{-1})\, X\beta \;\leq\; (X\beta)^T\, V_{1,n}^{-1}\, X\beta.$

By orthogonality of the design matrix (6), we have $X^T M_n = \; M_n^T X = n\,\big(\,I_{p+q+1}\,, 0_{n-(p+q+1)}\big)$.

Thus similarly to the proof of lemma (3), we get:

$$(X\beta)^T\, V_{1,n}^{-1}\, X\beta = \frac{1}{n^2}\,(X\beta)^T M_n\,\left[\frac{\sigma_\epsilon^2}{n}I_n + D_n\right]^{-1} M_n^T\, X\beta$$

$$= \frac{1}{n^2}\,\beta^T n\,\big(\,I_{p+q+1}\,, 0_{n-(p+q+1)}\big)\,\left[\frac{\sigma_\epsilon^2}{n}I_n + D_n\right]^{-1} n\,\big(\,I_{p+q+1}\,, 0_{n-(p+q+1)}\big)\,\beta$$

$$= \beta^T\,\big(\,I_{p+q+1}\,, 0_{n-(p+q+1)}\big)\,\left[\frac{\sigma_\epsilon^2}{n}I_n + D_n\right]^{-1}\big(\,I_{p+q+1}\,, 0_{n-(p+q+1)}\big)\,\beta$$

$$= \sum_{i=1}^{p+q+1}\beta_i^2\,\left[\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2\right]^{-1} = O(1).$$

Let $\varepsilon\; > 0$, by the Markov inequality when r = 1 and lemma (3), we have:

$$pr\left\{\frac{1}{S_{n,1}}Y_n^T\big(V_{1,n}^{-1}-V_1^{-1}\big)Y_n \;>\; \varepsilon\right\} \;\leq\; \frac{1}{S_{n,1}}\left\{\frac{E(\,Y_n^T\big(V_{1,n}^{-1}-V_1^{-1}\big)Y_n\,)}{\varepsilon}\right\}$$

$$\leq \frac{1}{S_{n,1}}\,\frac{tr\big(V_{1,n}^{-1}-V_1^{-1}\big)}{\varepsilon} + \frac{1}{S_{n,1}}\,\frac{(X\beta)^T\, V_{1,n}^{-1}\, X\beta}{\varepsilon} \xrightarrow{n\to\infty} 0.$$

Note that $V_{1,n}^{-1} - V_1^{-1}$ is nonnegative definite and thus, $Y_n^T(V_{1,n}^{-1}-V_1^{-1})Y_n$ is nonnegative random variable. Therefore, $A_2 = \; o\big(S_{n,1}\big)$.

The above results with lemma (2) imply there exists a constant C_1 such that

$$\frac{1}{S_{n,1}}\,\log B_{01} \geq \frac{1}{S_{n,1}}\,\log \tilde{B}_{01} + \,o(1)$$

$$> \,C_1 + \;o(1).$$

That is,

$B_{01} \geq \,\exp(\,S_{n,1}\{\,C_1 + \,o(1)\})\; \xrightarrow{n\to\infty} \infty$ in P_0^n probability.

This completes the proof.

Now we consider asymptotic behavior of the Bayes factor under H_1 and assuming that the true model for Y_n is in H_1.

Lemma(4):

Let $Y_n\; = (y_1\,, \dots, y_n)$ where $y_i{}'s$ are independent normal random variables with mean $\sum_{j=0}^{p}\beta_j x_{ji} + \sum_{j=1}^{q}\beta_{p+j}x_{p+1,i}^{j} + \sum_{k=1}^{n-(p+q+1)} z_{ik}\, u_k$ and variance $\sigma_\epsilon^2 = 1$ and let $w_n = \; E(\,Y_n\,)^T\, Q_n\, E(\,Y_n\,)$.

Then, $\lim_{n\to\infty}\frac{1}{w_n}\, Y_n^T Q_n Y_n = 1$, in P_1^∞ probability.

Proof:

From the moment formula of the quadratic form of normal random variables, the expectation and variance of quadratic form $Y_n^T Q_n Y_n$ are given by:

$$E(Y_n^T Q_n Y_n\,) = tr\,(\,Q_n\,) + E(\,Y_n\,)^T\, Q_n\, E(\,Y_n\,)$$

$$= \sum_{i=1}^{n-(p+q+1)}\frac{n\sigma_u^2}{(1+n\sigma_u^2)} + E(\,Y_n\,)^T\, Q_n\, E(\,Y_n\,)$$

$$= S_{n,1} + w_n \;\text{, and}$$

$$var(Y_n^T Q_n Y_n\,) = 2tr\,\big(\,{Q_n}^2\,\big) + 4E(\,Y_n\,)^T\, {Q_n}^2\, E(\,Y_n\,).$$

Note

$$w_n = E(Y_n)^T Q_n E(Y_n)$$
$$= (\beta^T, u^T_{n-(p+q+1)}) M_n^T \frac{1}{n^2} M_n \begin{bmatrix} 0_{(p+q+1)\times(p+q+1)} & 0_{1r} \\ 0_{r1} & Q_r \end{bmatrix} M_n^T M_n (\beta^T, u^T_{n-(p+q+1)})^T$$
$$= (\beta^T, u^T_{n-(p+q+1)})\, n.\frac{1}{n^2} \begin{bmatrix} 0_{(p+q+1)\times(p+q+1)} & 0_{1r} \\ 0_{r1} & Q_r \end{bmatrix} n\, (\beta^T, u^T_{n-(p+q+1)})^T$$
$$= (\beta^T, u^T_{n-(p+q+1)}) \begin{bmatrix} 0_{(p+q+1)\times(p+q+1)} & 0_{1r} \\ 0_{r1} & Q_r \end{bmatrix} (\beta^T, u^T_{n-(p+q+1)})^T$$
$$= u^T_{k-(p+q+1)} Q_r\, u_{n-(p+q+1)} = \sum_{i=1}^{n-(p+q+1)} \frac{n^2\sigma_u^2 u_i^2}{1+n\sigma_u^2},$$

since $sup_i\, u_i < \infty$, we have

$$w_n \le sup_i\, u_i^2\, n \sum_{i=1}^{n} \frac{n\sigma_u^2}{1+n\sigma_u^2} = O(n\, S_{n,1}).$$

On the other hand, consider a fixed positive integer $N \ge 1$ such that $u_N^2 > 0$ and a sufficiently large n with $\frac{n\sigma_u^2}{1+n\sigma_u^2} \ge \frac{1}{2}$. For such n and N.

$$w_n = \sum_{i=1}^{n-(p+q+1)} \frac{n^2\sigma_u^2 u_i^2}{1+n\sigma_u^2} \ge \frac{n^2\sigma_u^2}{1+n\sigma_u^2}\, u_N^2$$
$$= n\, \frac{n\sigma_u^2}{1+n\sigma_u^2}\, u_N^2 \ \ge n\frac{u_N^2}{2}.$$

Further, similarly to the previous calculation, we have

$$E(Y_n)^T Q_n^2 E(Y_n) = (\beta^T, u^T_{n-(p+q+1)}) M_n^T \frac{1}{n^3} M_n \begin{bmatrix} 0_{(p+q+1)\times(p+q+1)} & 0_{1r} \\ 0_{r1} & Q_r^2 \end{bmatrix} M_n^T M_n (\beta^T, u^T_{n-(p+q+1)})^T$$
$$= \frac{1}{n} u^T_{n-(p+q+1)} Q^2_{1,n-(p+q+1)}\, u_{n-(p+q+1)}$$
$$= \frac{1}{n} \sum_{i=1}^{n-(p+q+1)} \frac{n^4\sigma_u^4 u_i^2}{(1+n\sigma_u^2)^2}.$$

As in case of $E(Y_n)^T Q_n E(Y_n)$

$$E(Y_n)^T Q_n^2 E(Y_n) \le \sum_{i=1}^{n} \frac{n^3\sigma_u^4 u_i^2}{(1+n\sigma_u^2)^2} \le n\ sup_i\, u_i^2 \sum_{i=1}^{n} \frac{n^2\sigma_u^4}{(1+n\sigma_u^2)^2} = O(nS_{n,3}).$$

Consider a fixed positive integer $N \ge 1$ such that $u_N^2 > 0$ and a sufficiently large n with $\frac{n\sigma_u^2}{1+n\sigma_u^2} \ge \frac{1}{2}$. Again, for such n and N,

$$E(Y_n)^T Q_n^2 E(Y_n) = \sum_{i=1}^{n-(p+q+1)} \frac{n^3\sigma_u^4 u_i^2}{(1+n\sigma_u^2)^2} \ge n\, u_N^2 \left(\frac{n\sigma_u^2}{1+n\sigma_u^2}\right)^2 \ge \frac{u_N^2}{4}\, n.$$

By combining value of $S_{n,1}$ and the previous result, we have

$$var(Y_n^T Q_n Y_n) \le O(n\, S_{n,1}).$$

Furthermore, note that $E(Y_n)^T Q_n E(Y_n) \ge E(Y_n)^T {Q_n}^2 E(Y_n)$.

Let $\varepsilon > 0$, by the Chebshev inequality.

$$pr\left\{ \left| \frac{1}{w_n} Y_n^T Q_n Y_n - \frac{E(Y_n^T Q_n Y_n)}{w_n} \right| > \varepsilon \right\} \le \frac{var(Y_n^T Q_n Y_n)}{w_n^2 \varepsilon^2} = O(n^{-2}) \xrightarrow{n\to\infty} 0.$$

Since $\lim_{n\to\infty} \frac{E(Y_n^T Q_n Y_n)}{w_n} = \lim_{n\to\infty} \left(\frac{S_{n,1}}{w_n} + \frac{w_n}{w_n}\right) = 1$

$\lim_{n\to\infty} \frac{1}{w_n} Y_n^T Q_n Y_n = 1$, in probability

This completes the proof.

Lemma(5):

Suppose we have the penalized spline smoothing, then

$$\lim_{n\to\infty} \frac{1}{w_n} \log \frac{det(V_1)}{det(V_{1,n})} = 0$$

Proof:

From a property of the determinant of the positive definite matrix, we have

$$det(V_1) \le \prod_{i=1}^{n}\left(\sigma_\epsilon^2 + \sum_{k=1}^{p+q+1}\sigma_\beta^2 x_{ik}^2 + \sum_{k=1}^{n+s}\sigma_u^2 z_{ik}^2\right)$$
$$= n^n \prod_{i=1}^{n}\left(\frac{\sigma_\epsilon^2}{n} + \sum_{k=1}^{p+q+1}\frac{\sigma_\beta^2 x_{ik}^2}{n} + \sum_{k=1}^{n+s}\frac{\sigma_u^2 z_{ik}^2}{n}\right),$$

where x_{ik}, z_{ik} is the k th element of x_i, z_i respectively.

From lemma (1), we have:

$$\det(V_{1,n}) = n^n \left(\frac{\sigma_\epsilon^2}{n} + \sigma_u^2\right)^{n-(p+q+1)} \left(\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2\right)^{p+q+1}.$$

Thus

$$\log\frac{det(V_1)}{det(V_{1,n})} = \log(det(V_1)) - \log(det(V_{1,n}))$$
$$\le n\log n + \sum_{i=1}^{n}\log\left(\frac{\sigma_\epsilon^2}{n} + \sum_{k=1}^{p+q+1}\frac{\sigma_\beta^2 x_{ik}^2}{n} + \sum_{k=1}^{n}\frac{\sigma_u^2 z_{ik}^2}{n}\right) - \left[n\log n + (p+q+1)\log\left(\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2\right) + (n-(p+q+1))\left(\frac{\sigma_\epsilon^2}{n} + \sigma_u^2\right)\right] = O(1) - O(1) = O(1).$$

Therefore, it follows that

$$\frac{1}{w_n}\log\frac{det(V_1)}{det(V_{1,n})} \xrightarrow{n\to\infty} 0$$

Theorem(2):

Consider the testing problem in (7), suppose that the true model is in H_1 (the penalized spline semiparametric model) and let P_1^n denote the true distribution of the whole data, with p.d.f. $p_1(y|\beta_1, u_1) = \emptyset\{(y - (X^T\beta_1 + Zu_1))/\sigma_\epsilon\}$, where $\emptyset(\cdot)$ is the standard normal density. Then the Bayes factor is consistent under the alternative hypothesis H_1:

$\lim_{n\to\infty} B_{01} = 0$ in P_1^n probability.

Proof:

$$\log B_{01} = \frac{1}{2}\log\frac{\det(V_1)}{\det(V_0)} - \frac{1}{2}Y_n^T(V_0^{-1} - V_1^{-1})Y_n$$
$$= \frac{1}{2}\log\frac{\det(V_1)}{\det(V_{1,n})} + \frac{1}{2}\log\frac{\det(V_{1,n})}{\det(V_0)} - \frac{1}{2}Y_n^T(V_0^{-1} - V_{1,n}^{-1})Y_n - \frac{1}{2}Y_n^T(V_{1,n}^{-1} - V_1^{-1})Y_n$$
$$= \log\tilde{B}_{01} - \frac{1}{2}Y_n^T(V_0^{-1} - V_{1,n}^{-1})Y_n - \frac{1}{2}Y_n^T(V_{1,n}^{-1} - V_1^{-1})Y_n$$
$$= \log\tilde{B}_{01} + A_1 + A_2,$$

where

$$\log\tilde{B}_{01} = \frac{1}{2}\log\frac{\det(V_{1,n})}{\det(V_0)} - \frac{1}{2}Y_n^T(V_0^{-1} - V_{1,n}^{-1})Y_n$$

$$A_1 = \frac{1}{2}\log\frac{\det(V_1)}{\det(V_{1,n})}, \quad A_2 = -\frac{1}{2}Y_n^T(V_{1,n}^{-1} - V_1^{-1})Y_n.$$

First, by lemma (1),

$$\frac{\det(V_{1,n})}{\det(V_0)} = \frac{n^n\left(\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2\right)^{p+q+1}\left(\frac{\sigma_\epsilon^2}{n} + \sigma_u^2\right)^{n-(p+q+1)}}{n^{p+q+1}\left(\frac{\sigma_\epsilon^2}{n} + \sigma_\beta^2\right)^{p+q+1}\sigma_\epsilon^{2(n-(p+q+1))}} = \frac{n^{n-(p+q+1)}\left(\frac{\sigma_\epsilon^2}{n} + \sigma_u^2\right)^{n-(p+q+1)}}{\sigma_\epsilon^{2(n-(p+q+1))}} = \left(\frac{\sigma_\epsilon^2 + n\sigma_u^2}{\sigma_\epsilon^2}\right)^{n-(p+q+1)}$$
$$= \left(1 + \frac{n\sigma_u^2}{\sigma_\epsilon^2}\right)^{n-(p+q+1)} = O(n) \Rightarrow \log\left(1 + \frac{n\sigma_u^2}{\sigma_\epsilon^2}\right)^{n-(p+q+1)} = O(n).$$

And from lemma (4) $w_n \le O(n^2)$. Then

$$\frac{1}{w_n}\log\frac{\det(V_{1,n})}{\det(V_0)} \le O(n^{-1}) \xrightarrow{n\to\infty} 0.$$

Therefore,

$$\frac{1}{w_n}\log\tilde{B}_{01} = \frac{1}{2}\left(\frac{1}{w}\log\frac{\det(V_{1,n})}{\det(V_0)} - \frac{1}{w}Y_n^T(V_0^{-1} - V_{1,n}^{-1})Y_n\right)$$

Since $V_0^{-1} - V_{1,n}^{-1} = Q_n$, then

$\frac{1}{w_n}Y_n^T(V_0^{-1} - V_{1,n}^{-1})Y_n = \frac{1}{w_n}Y_n^T Q_n Y_n$, from lemma (4),

$\lim_{n\to\infty}\frac{1}{w_n}Y_n^T Q_n Y_n = 1$. Then

$$\frac{1}{w_n}\log\tilde{B}_{01} \xrightarrow{n\to\infty} = \frac{1}{2}\left(\frac{1}{w}\log\frac{\det(V_{1,n})}{\det(V_0)} - \frac{1}{w}Y_n^T(V_0^{-1} - V_{1,n}^{-1})Y_n\right) \xrightarrow{n\to\infty} = \frac{1}{2}(0-1) = \frac{-1}{2}$$

$\therefore \frac{1}{w_n}\log\tilde{B}_{01} \xrightarrow{n\to\infty} \frac{-1}{2}$, in the P_1^n probability.

And from lemma (5),

$\frac{1}{w_n}A_1 \xrightarrow{n\to\infty} 0$, finally, note that A_2 is nonpositive random variable since $V_{1,n}^{-1} - V_1^{-1}$ is a nonnegative definite matrix.

Therefore, by combining the above results, there exist $C_2 > 0$ such that

$\lim_{n\to\infty} \sup \frac{1}{w_n} \log B_{01} \leq \frac{-C_2}{2}$, with probability tending to 1.

Hence, $B_{01} \xrightarrow{n\to\infty} 0$, in P_1^n probability.

This completes the proof.

5.Conclusion

1- The Bayes factor for testing problem $H_0: Y = X\beta + \epsilon$ versus $H_1: Y = X\beta + Zu + \epsilon$ is given by: $B_{01} = \frac{\sqrt{\det(V_1)}}{\sqrt{\det(V_0)}} \exp\{-\frac{1}{2} Y_n^T (V_0^{-1} - V_1^{-1}) Y_n\}$.

2- If $Y_n = (y_1, \dots, y_n)^T$ where $y_i's$ are independent normal random variable with mean $\sum_{j=0}^{p} \beta_j x_{ji} + \sum_{j=1}^{q} \beta_{p+j} x_{p+1,i}^{j}$ and variance $\sigma_\epsilon^2 = 1$, then there exist a positive constant C_1 such that $\frac{1}{S_{n,1}} \left[\sum_{i=1}^{n-(p+q+1)} \log(1 + n\sigma_u^2) - Y_n^T Q_n Y_n \right] > C_1$, with probability tending to 1.

This implies that, under H_0. $\log \tilde{B}_{01} \xrightarrow{n\to\infty} \infty$, in probability.

3- If the true model is H_0 (the pure polynomial model), then, the Bayes factor is consistent under the null hypothesis H_0. This implies that, $\lim_{n\to\infty} B_{01} = \infty$ in P_0^n probability.

4- If $Y_n = (y_1, \dots, y_n)$ where $y_i's$ are independent normal random variables with mean $\sum_{j=0}^{p} \beta_j x_{ji} + \sum_{j=1}^{q} \beta_{p+j} x_{p+1,i}^{j} + \sum_{k=1}^{n-(p+q+1)} z_{ik} u_k$ and variance $\sigma_\epsilon^2 = 1$ and if $w_n = E(Y_n)^T Q_n E(Y_n)$. Then, $\lim_{n\to\infty} \frac{1}{w_n} Y_n^T Q_n Y_n = 1$, in P_1^∞ probability.

5- If the true model is in H_1 (the penalized spline semiparametric model) ,then the Bayes factor is consistent under the alternative hypothesis H_1. This implies that, $\lim_{n\to\infty} B_{01} = 0$ in P_1^n probability.

6.References

[1] Angers, J.-F. and Delampady, M. (2001). Bayesian nonparametric regression using wavelets. Sankhyã Ser. B 63, 287–308.

[2] Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis. Springer, New York.

[3] Choi, T., Lee, J. and Roy, A. (2008). A note on the Bayes factor in a semiparametric regression model, Journal of Multivariate Analysis. 1316-1327.

[4] G.Gosal, J. Lember, A. Van der Vaart, Nonparametric Bayesian model selection and averaging, Electron.J. Stat. 2 (2008) 63-89.

[5] Ghosh, J. K., Delampady, M. and Samanta, T. (2006). Introduction to Bayesian Analysis: Theory and Methods. Springer, New York.

[6] Ghosh, J. K. and Ramamoorthi, R. V. (2003). Bayesian Nonparametric. Springer, New York.

[7] I. Verdinelli, L. Wasserman, Bayesian goodness –of-fit testing using infinite- dimentional exponential families, Ann. Stat.26 (4) (1998)1215-1241.

[8] L. H. Zhao, Bayesian aspects of some nonparametric problems., Ann.Statist. 28 (2000) 532-552.

[9] P. J. Lenk, Bayesian inference for semiparametric regression using a Fourier representation, J. R, Stat. Soc. Ser. B Stat. Methodol. 61 (4) 1999) 863-879.

[10] Natio, k. (2002) " Semiparametric regression with multiplicative adjustment " Communications in statistics, Theory and methods 31, 2289-2309.

[11] Pelenis, J (2012) Bayesian Semiparametric Regression. Institute for Advanced Studies, Vienna 1-39.

[12] R.McVinish, J. Rousseau,K. Mengerson, Bayesian goodness –of-fit testing with mixtures of triangular distributions, Scand. J, statist. (2008), in press (doi:10.1111/j.1467-9469.2008.00620.x).

[13] Ruppert, D, Wand, M. P. and Carroll, R. J.(2003) Semiparametric Regression. Cambridge University Press.

[14] S.C. Dass, lee, A note on the consistency of Bayes factors for testing point null versus non-parametric alternative, J.Statist. 26 (4) (1998) 1215-1241.

[15] Tarmaratram, K. (2011) Robust Estimation and Model Selection in Semiparametric Regression Models, proefschrift voorgedragen tot het behalen van de grad van Doctor.