Simultaneous confidence bands for contrasts between several nonlinear regression curves

Xiaolei Lu* and Satoshi Kuriki†

Abstract — **We propose simultaneous confidence bands of the hyperbolic-type for the contrasts between several nonlinear (curvilinear) regression curves. The critical value of a confidence band is determined from the distribution of the maximum of a chi-square random process defined on the domain of explanatory variables. We use the volume-of-tube method to derive an upper tail probability formula of the maximum of a chi-square random process, which is sufficiently accurate in commonly used tail regions. Moreover, we prove that the formula obtained is equivalent to the expectation of the Euler-Poincaré characteristic of the excursion set of the chi-square random process, and hence conservative. This result is therefore a generalization of Naiman's inequality for Gaussian random processes. As an illustrative example, growth curves of consomic mice are analyzed.**

Keywords: *Chi-square random process; expected Euler-characteristic heuristic; growth curve; Naiman's inequality; volume-of-tube method.*

1 Introduction

This paper concerns multiple comparisons of k (≥ 3) nonlinear (curvilinear) regression curves estimated from independent k groups. For each group $i = 1, \ldots, k$, let

$$\beta_i^T f(x), \quad x \in \mathcal{X}, \tag{1.1}$$

be the true regression curve, where $f(x) = (f_1(x), \ldots, f_p(x))^T$ is a known regression basis vector function, $\beta_i = (\beta_{i1}, \ldots, \beta_{ip})^T$ is an unknown parameter vector, and $\mathcal{X} \subseteq \mathbb{R}$ is the domain of explanatory variables. Suppose that an estimator $\widehat{\beta_i}$ of β_i is available whose distribution is the normal distribution $N_p(\beta_i, r_i^{-1}\Sigma)$, where r_i is a known constant representing the size of the group i. The $p \times p$ matrix Σ is supposed to be known, or at least known up to a constant $\Sigma = \sigma^2\Sigma_0$. In the case of the latter, we suppose that an independent estimator $\widehat{\sigma^2}$ of σ^2 is available.

Let C denote the set of vectors $c = (c_1, \ldots, c_k)^T$ such that $\sum_{i=1}^k c_i = 0$. The focus of this paper is the construction of $1 - \alpha$ simultaneous confidence bands for all the contrasts $\sum_{i=1}^k c_i \beta_i^T f(x)$ between the k regression curves for all $x \in \mathcal{X}$ and $c \in C$, where $\mathcal{X}$ is a given finite interval $[a, b]$. Specifically, according to the traditional form of the point estimate plus or minus a probability point times the estimated standard error, we construct a $1 - \alpha$ simultaneous confidence band of the form

$$\sum_{i=1}^k c_i \beta_i^T f(x) \in \sum_{i=1}^k c_i \widehat{\beta_i^T} f(x) \pm b_\alpha \sqrt{\left(\sum_{i=1}^k \frac{c_i^2}{r_i} \right) f(x)^T \Sigma f(x)}, \tag{1.2}$$

where $\widehat{\beta_i^T} f(x)$ is the estimator of $\beta_i^T f(x)$ in (1.1). The critical value b_α is determined such that the event in (1.2) for all $x \in \mathcal{X}$ and $c \in C$ holds with a probability of at least $1 - \alpha$.

Our problem typically arises from the following experimental design. Suppose that for each group i, and for each explanatory variable x_j, $j = 1, \ldots, n$, we have observations $y_{ij1}, \ldots, y_{ijr_i}$ as objective variables with r_i replications, which are assumed to follow the model

$$y_{ijh} = \beta_i^T f(x_j) + \varepsilon_{ijh}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n, \quad h = 1, \ldots, r_i. \tag{1.3}$$

*Department of Statistical Sciences, Graduate University for Advanced Studies, 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan

†The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan, Email: kuriki@ism.ac.jp, Tel: +81-50-5533-8423

Here, random errors ε_{ijh} are assumed to be independently distributed as the normal distribution $N(0, \sigma(x_j)^2)$. Then, the least squares estimator $\widehat{\beta_i}$ of β_i has the multivariate normal distribution $N_p(\beta_i, r_i^{-1}\Sigma)$, where

$$\Sigma = \left(\sum_{j=1}^{n} \frac{1}{\sigma(x_j)^2} f(x_j) f(x_j)^T \right)^{-1}$$

is the inverse of the $p \times p$ information matrix. Data with this structure often appear in growth curve analysis and longitudinal data analysis.

2 Main result

Recall that our aim is to obtain the threshold b_α such that the event in (1.2) for all $x \in \mathcal{X}$ and $c \in C$ holds with a probability of at least $1 - \alpha$. For this purpose, we adopt the volume-of-tube method (Takemura and Kuriki, 2002). This is an integral geometric approach originating from Naiman (1986), and is equivalent to the expected Euler-characteristic heuristic (Adler and Taylor, 2007) in our setting.

In the unit sphere $\mathbb{S}^{p-1}$ of the p-dimensional Euclidean space, define a trajectory Γ of a normalized basis vector function by

$$\Gamma = \overline{\{\psi(x) \mid x \in \mathcal{X}\}} \subset \mathbb{S}^{p-1} \quad \text{with} \quad \psi(x) = \frac{\Sigma^{\frac{1}{2}} f(x)}{\|\Sigma^{\frac{1}{2}} f(x)\|}.$$

Then, b_α can be obtained from geometric quantities of Γ.

Theorem (Lu and Kuriki, 2016). *Let*

$$\widehat{P}(b) = \frac{\Gamma(\frac{k}{2})}{\sqrt{\pi}\Gamma(\frac{k-1}{2})} |\Gamma| \left\{ P\left(\chi_k^2 > b^2\right) - P\left(\chi_{k-2}^2 > b^2\right) \right\} + P\left(\chi_{k-1}^2 > b^2\right),$$

where $|\Gamma|$ is the length of the curve Γ. Let $b = b_\alpha$ be the solution of $\widehat{P}(b) = \alpha$. Then, b_α is a conservative and asymptotically exact (when $\alpha \to \infty$) threshold for the $1 - \alpha$ simultaneous confidence bands (1.2).

Note that when the map $\psi : \mathcal{X} \to \mathbb{S}^{p-1}$ is one-to-one,

$$|\Gamma| = \int_{\mathcal{X}} \left\| \frac{d}{dx} \psi(x) \right\| dx. \tag{2.1}$$

3 Growth curve analysis

Mice are one of the most popular model organisms, and are often used in genomic research. Figure 3.1 depicts the average body weights of male mice from four different strains measured from 2 to 20 weeks after birth. The four strains are C57BL/6 (referred to as B6), MSM/Ms (MSM), B6-Chr17$^{\mathrm{MSM}}$(B6-17), and B6-ChrXT$^{\mathrm{MSM}}$(B6-XT). Among these, B6 is the most common laboratory strain and serves as the standard. MSM is a wild-derived strain that has contrasting properties to B6 such as non-black color, small size, and aggressive behavior. B6-17 and B6-XT are artificial strains known as consomic mice made from B6 and MSM. B6-17 has all the chromosomes from B6, but only chromosome 17 from MSM; B6-XT has all the chromosomes from B6, but only half of the X chromosome from MSM. By comparing the consomic strains with B6, we expect to reveal the role of each chromosome.

The dataset we used is publicly available as Supplemental Table S1 of Takada, et al. (2008). In their experiments, the weight (unit: gram) y_{ijh} of the hth individual from strain i was measured at time point x_j. The measurement time points were $\{x_1, \ldots, x_{10}\} = \{2, 4, \ldots, 20\}$ ($n = 10$). This dataset includes the average body weight y_{ij} of strain i at time x_j, and its standard error

$$y_{ij} = \frac{1}{r_i} \sum_{h=1}^{r_i} y_{ijh}, \quad \widehat{\text{s.e.}}(y_{ij}) = \sqrt{\frac{1}{r_i^2} \sum_{h=1}^{r_i} (y_{ijh} - y_{ij})^2},$$

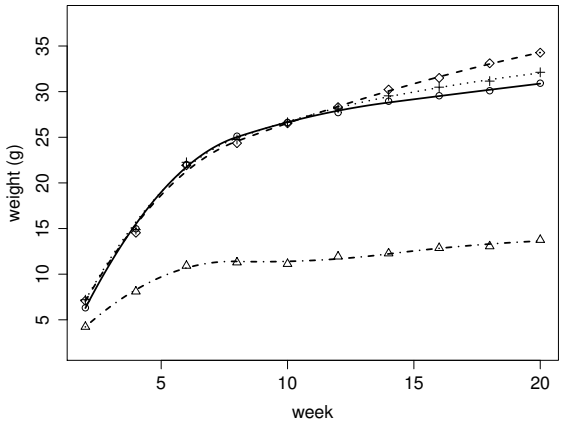

Figure 3.1: Average body weights of mice from four strains.

(sample mean: ∘ (B6), + (B6-17), ⋄ (B6-XT), ▵ (MSM);
fitted curve: — (B6), ⋯ (B6-17), – – (B6-XT), – · – (MSM))

as well as the number r_i of individuals of strain i.

In the following analysis, we use $k = 3$ groups (strains) B6 ($i = 1$), B6-17 ($i = 2$), and B6-XT ($i = 3$). The number of individuals are $r_1 = 12$, $r_2 = 24$, and $r_3 = 12$.

We fit the model (1.3) to these data. We estimate the variance as

$$\widehat{\sigma}(x_j)^2 = \frac{1}{\sum_{i=1}^k (r_i - 1)} \sum_{i=1}^k \sum_{h=1}^{r_i} (y_{ijh} - y_{ij})^2 = \frac{1}{\sum_{i=1}^k (r_i - 1)} \sum_{i=1}^k r_i^2 \widehat{\mathrm{s.e.}}(y_{ij})^2,$$

which is used as the true value $\sigma(x_j)^2$ hereafter. One particular feature of this dataset is that the experiment is well controlled and the measurement errors are quite small.

As the basis function $f(x)$, we consider a family of basis functions

$$f(x) = f_{d,m}(x) = \left(B_d \left(\frac{x-2}{20-2}(m-d) - (i-d-1) \right) \right)_{1 \le i \le m},$$

where

$$B_d(x) = \sum_{r=0}^{d+1} (-1)^{d+1-r} \binom{d+1}{r} \frac{(r-x)_+^d}{d!}$$

(de Boor, 1978, page 89). $f_{d,m}(x)$ consists of m B-spline bases with equally-spaced knots at intervals of $(20-2)/(m-d)$. Note that $f_{d,m}(x)$ is piecewise of class C^d.

In the range $d = 2, 3, 4$ and $m = d+1, d+2, \ldots, n\,(= 10)$, we searched for the best model that minimizes AIC and BIC. In both criteria, the minimizer was $(d, m) = (2, 5)$, which we use as the true value hereafter.

Suppose that we are interested in the period $\mathcal{X} = [a, b] = [2, 20]$. An approximate value of the length of Γ in (2.1) is given by

$$|\Gamma| \approx \sum_{t=1}^N \left\| \psi(x_t) - \psi(x_{t-1}) \right\|,$$

where $x_t = a + t(b-a)/N$, $t = 0, 1, \ldots, N$. When $N = 10,000$, the approximate value of $|\Gamma|$ is $6.989 = 2.225\pi$. Using this value, the critical value is $b_\alpha = 3.258$ ($\alpha = 0.05$).

To compare k groups, various types of contrasts are used. For a pairwise comparison between group i and group j, we choose $c = (\ldots, 0, \underset{i\text{th}}{1}, 0, \ldots, 0, \underset{j\text{th}}{-1}, 0, \ldots)$. For the comparison of groups $\{i, j\}$ and group k, we use

$$c = \Big(\ldots, 0, \underset{i\text{th}}{\frac{r_i}{r_i + r_j}}, 0, \ldots, 0, \underset{j\text{th}}{\frac{r_j}{r_i + r_j}}, 0, \ldots, 0, \underset{k\text{th}}{-1}, 0, \ldots\Big).$$

Figure 3.2 depicts the difference curves of strains B6-17 vs. B6 (left) and B6-XT vs. B6 (right), and their 95% simultaneous confidence bands. In the left panel, the horizontal line representing zero difference is almost between the confidence bands. This means that there is no significant difference between B6-17 and B6. In contrast, in the right panel, after about week 14, the horizontal line is outside the confidence bands, thereby indicating that B6-XT and B6 are different during this period.

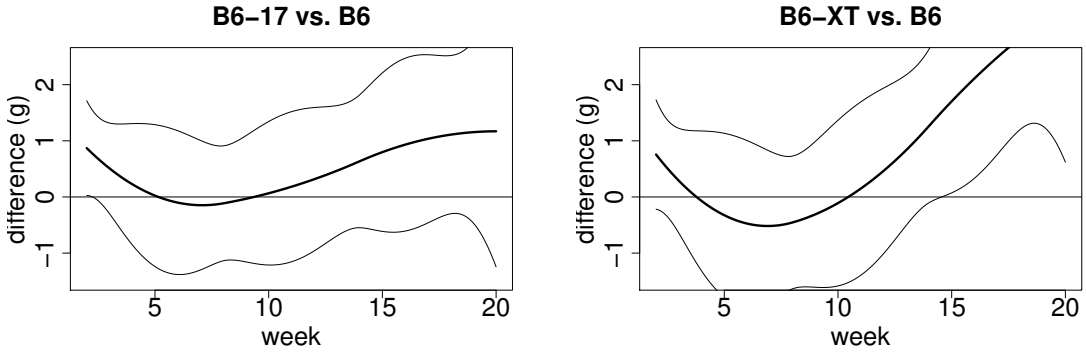

Figure 3.2: Differences of body weights and 95% confidence bands.

For a fixed x, the test statistic for the null hypothesis $H_{0,x} : \beta_1^T f(x) = \cdots = \beta_k^T f(x)$ is

$$\chi^2(x) = \frac{1}{f(x)^T \Sigma f(x)} \sum_{i=1}^{k} r_i \left(\widehat{\beta}_i^T f(x) - \frac{\sum_{i=1}^{k} r_i \widehat{\beta}_i^T f(x)}{\sum_{i=1}^{k} r_i} \right)^2 .$$

For a fixed x, the null distribution is the chi-square distribution with $k-1$ degrees of freedom. However, for the overall null hypothesis $H_0 : \beta_1^T f(x) = \cdots = \beta_k^T f(x)$ for all $x \in \mathcal{X}$, the distribution of the maximum of the chi-square random process should be used. The critical value b_α^2 can be used for this purpose. In this dataset, as already shown in Figure 3.2, after about week 14, the hypothesis of equality is rejected.

References

[1] Adler, R. J. and Taylor, J. E. (2007). *Random Fields and Geometry*, Springer.

[2] de Boor, C. (1978). *A Practical Guide to Splines*, Springer.

[3] Lu, X. and Kuriki, S. (2016). Simultaneous confidence bands for contrasts between several nonlinear regression curves, arXiv:1510.05077.

[4] Naiman, D. Q. (1986). Conservative confidence bands in curvilinear regression. *The Annals of Statistics*, **14** (3), 896–906.

[5] Takada, T., Mita, A., Maeno, A., Sakai, T., Shitara, H., Kikkawa, Y., Moriwaki, K, Yonekawa, H., and Shiroishi, T. (2008). Mouse inter-subspecific consomic strains for genetic dissection of quantitative complex traits, *Genome Research*, **18** (3), 500–508. `http://genome.cshlp.org/content/18/3/500`

[6] Takemura, A. and Kuriki, S. (2002). On the equivalence of the tube and Euler characteristic methods for the distribution of the maximum of Gaussian fields over piecewise smooth domains. *The Annals of Applied Probability*, **12** (2), 768–796.