Profiling Time Course Expression of a Single Virus Gene

I-Shou Chang[1,2], Li-Chu Chien[2], Pramod K. Gupta[2], Chi-Chung Wen[3], Yuh-Jenn Wu[4], and Chao A. Hsiung[2*]

[1]*Institute of Cancer Research,* [2]*Division of Biostatistics and Bioinformatics, National Health Research Institutes, 35 Keyan Road, Zhunan Town, Miaoli County 350, Taiwan*
[3]*Department of Mathematics, Tamkang University, 151 Ying-chuan Road, Tamsui Town, Taipei County 251, Taiwan*
[4]*Department of Applied Mathematics, Chung Yuan Christian University, 200 Chung Pei Road, Chung Li City 320, Taiwan*

Abstract

We illustrate a Bayesian shape restricted regression method in making inference on the time course expression profile of a virus gene, using data from microarray experiments. The prior is introduced through Bernstein polynomials so as to take into consideration the geometry of the regression functions, which are assumed to be zero initially, increasing after a while and staying positive later on. A reversible jump Metropolis-Hastings algorithm is used to generate the posterior distribution. We evaluate the performance of this method in a simulation study and illustrate its use by analyzing the microarray data of a virus gene. One advantage of this method is that it offers an assessment of the strength of the evidence provided by the data in favor of hypothesis on the shape of the regression function; for example, the hypothesis that it is unimodal. Another advantage of this approach is that estimates of many salient features of the profile like onset time, inflection point, maximum value, time to maximum value, etc. can be obtained immediately.

Key words: Bayesian shape restricted regression model; Bernstein polynomials; microarray data; reversible jump Metropolis-Hastings algorithm.

*Corresponding author. Email address: hsiung@nhri.org.tw. Phone: 886-37-246-166, ext. 36100. Fax: 886-37-586-467.

November 19, 2008

1. Introduction

It is generally believed that genes of a virus have their time course expression level being zero initially, then increasing after a while and finally decreasing; because viruses don't have cells, their genes start to express only after getting into cells, and cells may eventually malfunction when infected. Based on microarray expression data of virus genes taken at several time points, it is now possible and of interest to study the time course expression profile of a virus gene; for example, to assess whether it is unimodal and to estimate salient features of the profile like onset time, inflection point, maximum value, time to maximum value, etc. Using expression data described in Jiang et al. (2006), this paper addresses these problems in the framework of Bayesian shape restricted regression; we will introduce priors on a space of continuous functions satisfying certain shape restrictions, use Markov chain Monte Carlo methods for the inference, and then analyze the expression data so as to profile the time course expression of a virus gene. In fact, it seems to us that the profile of a virus gene provides an excellent example for the illustration of the strength of Bayesian approach to shape restricted inference.

The data are obtained from single color cDNA microarray experiments with external controls and are normalized using these external controls; mRNA samples of baculovirus genes were taken at 16 different time points during the 72 hours following infection; the sample for each time point is hybridized to a single chip that has exactly four spots for each of the 156 genes of baculovirus. Details of the experiments are in Jiang et al. (2006). Preliminary examination of the data suggests that 2 of these genes seem to have their expression levels being zero finally as well as initially and the rest 154 genes being zero only initially, probably because no data were taken at time points beyond 72 hours and the life cycle of baculovirus is longer than 72 hours, according to Friesen and Miller (2001). Since the main purpose of this paper is to illustrate a Bayesian method under shape restriction, we only study the profile of the gene *lef2*, whose expression at the 72 hour time point is positive, based only on the data for this gene. Other genes, including those two whose expression levels being zero before the 72 hour time point, can be studied similarly or with some modification and

refinement of the method.

In order to facilitate the discussion, we introduce some notations. Let $\mathcal{A}$ denote the set of all continuously differentiable functions on [0,1] that are zero initially, start to increase after a while, and stay positive onward except possibly at the point 1. We assume that given F in $\mathcal{A}$,

$$Y_{jk} = F(X_k) + \epsilon_{jk}, \tag{1.1}$$

where $\{X_k \,|\, k = 0, \cdots, K\}$ are constant design points in [0,1], $\{Y_{jk} \,|\, j = 1, \cdots, m_k, k = 0, \cdots, K\}$ are response variables, and $\{\epsilon_{jk} \,|\, j = 1, \cdots, m_k, k = 0, \cdots, K\}$ are independent errors with ϵ_{jk} being normal with mean μ and variance

$$\sigma_k^2 = \sigma^2(F(X_k) + \mu)(1 + X_k(1 - X_k)^2), \tag{1.2}$$

for every $j = 1, \cdots, m_k$.

In the microarray experiments in this paper, X_k represents a time point at which the mRNA sample is taken; for $j = 1, \cdots, m_k$, Y_{jk} is the expression level, in terms of fluorescent intensity, obtained at the jth spot of the gene for the sample taken at time point X_k. More specifically, let [0,1] denote the time period of 72 hours, then $K = 15$, $m_k = 4$, $(X_0, X_1, \cdots, X_{15})$=(0, 1/216, 1/108, 1/72, 1/36, 1/24, 1/12, 1/8, 1/6, 5/24, 1/4, 1/3, 5/12, 2/3, 5/6, 1).

The variance structure in (1.2) is a simple way to take into consideration the observation that for single color cDNA microarray experiments, larger intensities often incur larger variances, and the observation that, according to the data, given the same expression level, larger $X_k(1 - X_k)^2$ seems to correspond to larger variance of intensities at X_k. The reason for not assuming ϵ_{jk} having zero mean is that there are always background intensities due to non-specific hybridization and hence $E(Y_{jk})$ may not be zero even when the expression level $F(X_k)$ is 0. Another possible way to take care of background noise is to assume $Y_{jk} = \mu + F(X_k) + \epsilon_{jk}$, with μ the intercept and ϵ_{jk} has mean 0. These are further elaborated in the Remarks in Section 4.

Formulating as an inference problem for Bayesian shape restricted regression model, we introduce priors by Bernstein polynomials. For integers $0 \le i \le n$, let $\varphi_{i,n}(t) = C_i^n t^i (1-t)^{n-i}$,

where $C_i^n = n!/(i!(n-i)!)$. The set $\{\varphi_{i,n} \,|\, i = 0, \cdots, n\}$ is called the Bernstein basis for polynomials of order up to n. Let $\mathcal{B} = [0,1] \times \bigcup_{n=1}^{\infty}(\{n\} \times \mathbb{R}^{n+1})$. Define $\mathbf{F} : \mathcal{B} \times [0,1] \longrightarrow \mathbb{R}^1$ by

$$\mathbf{F}(c, n, b_{0,n}, \cdots, b_{n,n}; t) = \sum_{i=0}^{n} b_{i,n}\varphi_{i,n}(\frac{t-c}{1-c})I_{(c,1]}(t), \tag{1.3}$$

where $(c, n, b_{0,n}, \cdots, b_{n,n}) \in \mathcal{B}$ and $t \in [0,1]$. We also denote (1.3) by $F_{c,b_n}(t)$ if $b_n = (b_{0,n}, \cdots, b_{n,n})$. We will see in Section 2 that $F_{c,b_n}(\cdot)$ is a member of $\mathcal{A}$ if $0 = b_{0,n} = b_{1,n} \leq \min_{l=2,\cdots,n} b_{l,n} < \max_{l=2,\cdots,n} b_{l,n}$, and every member of $\mathcal{A}$ can be approximated by $F_{c,b_n}(\cdot)$ satisfying these restrictions on b_n. These suggest that by means of (1.3), Bernstein polynomials can be used to introduce priors on $\mathcal{A}$ with large enough support.

We note that Bayesian shape restricted inference with priors introduced by Bernstein polynomials was studied by Chang et al. (2005), which provides a smooth estimate of an increasing failure rate based on right censored data, and by Chang et al. (2006), which compares the Bernstein polynomial method with the density-regression method (Dette et al. 2006) in estimating an isotonic regression function and a convex regression function. It was shown there that these priors easily take into consideration geometric information, select only smooth functions, can have large support, and can be easily specified.

The present paper indicates that the expression profile of a virus gene can also be studied by random Bernstein polynomials. In particular, we will test the hypothesis on the shape of the time course expression profile; for example, we will examine whether it is unimodal on the region $[0,\tau]$ for some $\tau < 1$. In fact, by calculating both the posterior probability and the prior probability that it is unimodal on $[0,\tau]$, we offer an assessment of the strength of the evidence in favor of the hypothesis. We note that this direct approach to hypothesis testing is markedly different from the frequentist p-value approach, as discussed in Kass and Raftery (1995) and Lavine and Schervish (1999), for example. We will also estimate salient features of the profile like onset time, inflection point, maximum value, time to maximum value, etc., utilizing the fact that derivative of a polynomial has a closed form. We note that these properties and features are of particular interest to biologists; for example, based on onset time and time to maximum, Jiang et al. (2006) clustered the 156 genes of baculovirus; it

would be desirable to make use of all these properties and features in the study of baculovirus.

There is a large literature on shape restricted inference since Hildreth (1954) and Brunk (1955). Most of them treat isotonic and concave regressions from the frequentist viewpoint. Readers are referred to Gijbels (2003) for an excellent review and to Dette et al. (2006) for some of the more recent developments. For Bayesian approach, there are the works of Lavine and Mockus (1995), Dunson (2005) and Chang et al. (2006), among others. This paper illustrates the use of Bernstein polynomial in investigating the strength of the evidence provided by the data in favor of hypothesis on the shape of the regression function and in estimating its salient features.

This paper is organized as follows. Section 2 presents the Bernstein polynomial geometry and the regression model. Algorithms for Bayesian inference are given in the Appendix. Section 3 presents a simulation study to demonstrate the numerical performance. Section 4 illustrates the method by analyzing the data for the gene *lef2*. Section 5 provides a brief discussion on future investigations.

2. Bayesian inference

2.1. Bernstein polynomial geometry

Let $F_{c,a}(t) = \sum_{i=0}^{n} a_i \varphi_{i,n}(\frac{t-c}{1-c}) I_{(c,1]}(t)$, where $a = (a_0, \cdots, a_n)$. Proposition 1 provides a sufficient condition on a under which $F_{c,a}$ is a smooth function that is zero initially, increases after a while and stays positive later on. Proposition 2 complements Proposition 1 and provides Bernstein-Weierstrass type approximations for functions of the desired shape restriction. In this paper, derivatives at 0 and 1 are meant to be one-sided. All the proofs of the propositions in this paper are omitted, because they are similar to those in Chang et al. (2005) and Chang et al. (2006).

Proposition 1. *Let $n \geq 3$ and $c \in [0, 1)$. If $0 = a_0 = a_1 \leq min_{l=2,\cdots,n} a_l < max_{l=2,\cdots,n} a_l$, then $F_{c,a}$ is continuously differentiable, constantly 0 on $[0, c]$, and larger than 0 on $(c, 1)$.*

Let $I_n = \{F_{c,a} \mid c \in [0, 1), a = (a_0, \cdots, a_n)$ satisfying $0 = a_0 = a_1 \leq \min_{l=2,\cdots,n} a_l < \max_{l=2,\cdots,n} a_l\}$. Then we have

Proposition 2. *Let $\mathcal{D} = \bigcup_{n=3}^{\infty} I_n$. Let $\mathcal{A}$ denote the set of all continuously differentiable real-valued functions F defined on $[0,1]$ satisfying the property that there exists $c \in [0,1)$ such that $F = 0$ on $[0,c]$ and $F(t) > 0$ for $c < t < 1$. For two continuously differentiable functions f and g, define $d(f,g) = \|f-g\|_\infty + \|f'-g'\|_\infty$, where $\|\cdot\|_\infty$ is the sup-norm for functions on $[0,1]$. Then $\mathcal{D}$ is dense in $\mathcal{A}$, under d.*

2.2. Bayesian regression

We now introduce probability distributions on $\mathcal{A}$ by the mapping $\mathbf{F}$ defined in (1.3) and probabilities on $\mathcal{B}$. The following represents a convenient and flexible way to introduce priors. Let π_1 be a probability density function on [0,1], π_2 be a probability mass function on the set of positive integers $\{3,4,\cdots\}$, $\pi_3(\cdot|\{n\}\times\mathbb{R}^{n+1})$ be a probability density function on $\mathbb{R}^{n+1}$. The probability density/mass functions π_1, π_2 and π_3 jointly define a probability on $\mathcal{B}$ by the product $\pi_1(c)\times\pi_2(n)\times\pi_3(b_n|\{n\}\times\mathbb{R}^{n+1})$. This in turn defines a probability measure $\tilde{\pi}$ on $\mathcal{A}$ by (1.3). To complete the prior specification, we need also a probability density π_4 on $\mathbb{R}^1$ for μ, the mean of ϵ_{jk}. Then $\pi = \tilde{\pi}\times\pi_4$ is the prior we use for Bayesian inference. In accordance with (1.1), given $\mathbf{B} = (c,n,b_n,\mu) \in \mathcal{B}\times\mathbb{R}^1$, the likelihood for the data $\{(X_k, Y_{jk}) \,|\, j = 1,\cdots,m_k, k = 0,\cdots,K\}$ is

$$\prod_{k=0}^{K}\prod_{j=1}^{m_k} g_k(Y_{jk} - F_{c,b_n}(X_k)),$$

where g_k is the normal density of ϵ_{jk} specified in (1.2).

Thus the posterior density ν of the parameter (c,n,b_n,μ) given the data is proportional to

$$\prod_{k=0}^{K}\prod_{j=1}^{m_k} g_k(Y_{jk} - F_{c,b_n}(X_k))\pi(c,n,b_n,\mu),$$

where $(c,n,b_n) \in \mathcal{B}$ and $\mu \in \mathbb{R}^1$.

Because the parameter space consists of subspaces of different dimension, we propose a reversible jump Metropolis-Hastings (RJMH) algorithm, as discussed in Green (1995), to generate posterior distributions for inference. Details of the algorithm are in the Appendix.

The following proposition shows that the support of the Bernstein prior can be quite large.

Proposition 3. *Assume* π_1 *has support* $[0,1]$, $\pi_2(n) > 0$ *for every* $n = 3, 4, \cdots$, *and* $\pi_3(b_n|\{n\} \times \mathbb{R}^{n+1})$ *has support* B_n *on an infinite subsequence of* n. *Here* $B_n = \{b_n \in \mathbb{R}^{n+1} : F_{c,b_n} \in I_n$ *for some* $c \in [0,1)\}$. *Let* F *be a continuously differentiable real-valued function defined on* $[0,1]$ *satisfying the property that there exists* $\tilde{c} \in [0,1)$ *such that* $F = 0$ *on* $[0,\tilde{c}]$ *and* $F(t) > 0$ *for* $\tilde{c} < t < 1$. *Then* $\tilde{\pi}\{(c,n,b_n) \in [0,1) \times \bigcup_{n=3}^{\infty}(\{n\} \times B_n) : \|F_{c,b_n} - F\|_\infty < \epsilon\} > 0$ *for every* $\epsilon > 0$.

3. A simulation study

We now explore the numerical performance of the Bayesian method in a simulation study, whose model and design points are motivated by the virus gene expression data. We consider the Bayesian regression model (1.1) with the true regression function

$$F(t) = 200(t - 0.02)^{5/2}(1 - 0.7t)^5 I_{(0.02,1]}(t).$$

As in Section 1, $(X_0, X_1, \cdots, X_{15})$=(0, 1/216, 1/108, 1/72, 1/36, 1/24, 1/12, 1/8, 1/6, 5/24, 1/4, 1/3, 5/12, 2/3, 5/6, 1), $m_k = 4$, for every $k = 0, 1, \cdots, 15$; for $j = 1, 2, 3, 4$ and $k = 0, 1, \cdots, 15$, we assume that ϵ_{jk} are independently normally distributed with mean $E(\epsilon_{jk}) = 0.0039$ and variance given by

$$\sigma_k^2 = 0.0366(F(X_k) + 0.0039)(1 + X_k(1 - X_k)^2).$$

Using the notations in Section 2, we specify the following priors for inference. Let $\pi_2(3) = \sum_{n=0}^{3} C_n^{n_0} p^n (1-p)^{n_0-n}$ and $\pi_2(n) = C_n^{n_0} p^n (1-p)^{n_0-n}$, for $n = 4, 5, \cdots, n_0$. Let q be *Uniform*(q_1, q_2) whose support contains the true regression function F and is non-negative; let $a_0 = a_1 = 0$ and let $a_2, a_3, \cdots, a_n$ be a random sample of size $n-1$ from q; the conditional distribution of $\pi_3(\cdot|\{n\} \times \mathbb{R}^{n+1})$ is defined to be that of $(a_0, a_1, \cdots, a_n)$. Let π_1 be *Uniform*(π_{11}, π_{12}) and π_4 be *Uniform*(π_{41}, π_{42}). Here q_1, q_2, π_{11}, π_{12}, π_{41} and

π_{42} are defined as follows. Let $\overline{Y}_{(0)} \leq \overline{Y}_{(1)} \leq \cdots \leq \overline{Y}_{(15)}$ be the order statistics for $\{\overline{Y}_0, \overline{Y}_1, \cdots, \overline{Y}_{15}\}$, where $\overline{Y}_k = \sum_{j=1}^4 Y_{jk}/4$. Let $Y_{j[k']} = Y_{jk}$ and $X_{[k']} = X_k$, if $\overline{Y}_{(k')} = \overline{Y}_k$. Denote by $Y_{(1[k])} \leq Y_{(2[k])} \leq Y_{(3[k])} \leq Y_{(4[k])}$ the order statistics of $\{Y_{1[k]}, Y_{2[k]}, Y_{3[k]}, Y_{4[k]}\}$. Then $q_1 = \pi_{11} = \pi_{41} = 0$, $q_2 = Y_{(4[15])}$, $\pi_{12} = X_{[15]}$ and $\pi_{42} = 2\overline{Y}_0$. We choose $p = 0.5$ and $n_0 = 25$, 45, 60 in this simulation study.

We use the algorithm (RJMH), in the Appendix, with $\gamma = 0.35$, $M = q_2$, and estimated σ_k^2's to generate the posterior distribution. We note that this choice of γ allows relatively large probabilities of changing the order of the polynomial and, for the sampling of $x^{(t+1)}$, σ_k^2 is defined to be

$$\hat{\sigma}_k^2 = \hat{\sigma}^2(\hat{F}(X_k) + \mu)(1 + X_k(1 - X_k)^2),$$

where $\hat{\sigma}^2 = \sum_{k=0}^{15}(\tilde{\sigma}_k^2/\tilde{\mu}_k)/16$, $\hat{F}(X_k) = F_{x^{(t)}}(X_k)$ with $x^{(t)}$ denoting the current state and μ is the background noise in the current state $x^{(t)}$. Here $\tilde{\sigma}_k^2 = \sum_{j=1}^4 (Y_{jk} - \overline{Y}_k)^2/3$ and $\tilde{\mu}_k = \overline{Y}_k(1 + X_k(1 - X_k)^2)$.

We run 5 RJMH chains with initial values chosen randomly from the prior and monitor convergence by the Gelman-Rubin statistic $\hat{R}$, following the suggestion in Gelman and Rubin (1992) and Gelman et al. (2004), pp. 294–297. The Gelman-Rubin statistics $\hat{R}$ is calculated for six estimands of interest, which are onset time (Ton), time to maximum (Tmax), maximum (Max), time at which the slope is highest (Tslope), the highest slope (Slope), and the area under curve (L_1-norm). Each of the five chains was run with 200,000 updates; all the $\hat{R}$ based on the second halves of these 5 sequences are less that 1.1; the 500,000 updates from the second halves of these 5 sequences are considered sample from the posterior distribution, which form the basis for inference. Table 1a presents the initial values of the onset time, the order of the polynomial and the average of the background noise μ in the 5 chains; Table 1b gives the $\hat{R}$.

The posterior probability and the prior probability that the parameter represents a unimodal curve on the interval [0,τ] for $\tau = 0.6667, 0.8333, 1.0000$ are reported in Table 1c; the last two rows give respectively the ratio of the posterior probability to the prior probability and the Bayes factor. Table 1c presents a strong evidence, provided by the data, in

favor of the unimodality of the regression function. The posterior probability and the prior probability that the parameter represents a curve that is increasing before reaching to its global maximum are reported in Table 1d; the last two rows give respectively the ratio of the posterior probability to the prior probability and the Bayes factor.

Table 1e reports the Ton, Tmax, Max, Tslope, Slope, L_1-norm and Tend of the mode of the posterior distribution; Table 1e also reports the mean, standard deviation and support of Ton, Tmax, Max, Tslope, Slope, L_1-norm and Tend on the sample respectively from the posterior and prior distributions. Here Tend is the largest time point τ so that the profile is unimodal on $[0,\tau]$. The true values of these quantities are also included in Table 1e, which shows that these estimates are quite accurate.

We also carried out the analysis with the order $n_0 = 45, 60$ and all the other specifications in the prior unchanged. The results are quite close to each other and thus omitted.

4. Profile for the gene *lef2*

We now analyze the microarray data described in the introduction; the data is presented in Table 2. This is the normalized data and it was preprocessed as follows. The average of the fluorescence signal intensities of the external RNA spiked-in controls with predetermined quantities of input mRNA was used to carry out a global normalization; namely, the value in each entry of Table 2 is equal to the fluorescence intensity at the spot divided by this average. We note this is different from the normalization procedure described in Jiang et al. (2006).

The prior distribution and the parameters in the algorithm (RJMH) in this analysis are exactly the same as those in Section 3. We also run 5 chains with randomly chosen initial values; each chain has length 200,000 and the 500,000 updates from the second halves of these 5 sequences are considered sample from the posterior distribution, which form the basis for inference. The results are contained in Table 3a–e; each entry in these tables bears exactly the same meaning as that in the corresponding entry in Table 1a–e; the polynomial order n_0 for Table 3a–e is 25. The results for polynomial order $n_0 = 45$ and 60 are very close to those in Table 3 and hence omitted.

Since biological knowledge together with properties of Bernstein polynomial given in Section 2 indicates that the prior is reasonable for the study of time course expression profile of a virus gene, we think it is appropriate to study the hypothesis of the unimodality of the expression profile on $[0,\tau]$ based on the posterior probability and the prior probability of its being unimodal on $[0,\tau]$; in particular, it seems that both its ratio and the Bayes factor can be used to represent the strength of the evidence provided by the data in favor of the hypothesis.

According to Table 3c, the Bayes factor and the ratio of the posterior probability to the prior probability are quite large for the hypothesis that the expression profile is unimodal on [0, 0.6667], or before the 48 hour time point; according to Table 3e, the standard deviation and support of the posterior distribution of each estimand are much smaller than those of the corresponding prior distribution. The former indicates that the data strongly suggest the expression profile is unimodal before the 48 hour time point, and the latter indicates that the data help capture the features on the expression profile.

Knowing that there are relatively less design points near 1 and the infected cells, which are synchronized at the beginning of the experiment, become less synchronized at later stage, it seems reasonable to see in Table 3c that the posterior probability decreases as τ approaches 1. It would be interesting to incorporate the information regarding synchronization into the model to improve this approach.

Remarks on the model.

i) A more sophisticated model that takes into consideration both the technological and biological variations might replace (1.1) by

$$Y_{jk} = F(X_k) + \mu_j + \epsilon_{jk}, \tag{4.1}$$

with ϵ_{jk} being independent and normal with mean 0 and variance

$$\sigma^2_{jk} = \sigma^2(F(X_k) + \mu_j)(1 + X_k(1 - X_k)^2).$$

The reason that we use (1.1), instead of (4.1), are two fold. First, data in Table 2 do

not indicate obvious spot bias. Second, we wish to use a simpler model so as to keep the main ideas of the paper in focus.

ii) The variance structure (1.2) and the details of the prior specification in the above data analysis are motivated by the crude estimates obtained from the data points given in Table 2. In fact, we considered $\tilde{\sigma}_k^2/\tilde{\mu}_k^\xi$ for $\xi = 0, 0.5, 1$, and 2 for $\tilde{\mu}_k$ larger than the background noise value, and found, when $\xi = 1$, $\tilde{\sigma}_k^2/\tilde{\mu}_k^\xi$ as a function in k varies the least. We also note that we analyzed the data with $X_k(1-X_k)^2$ replaced by $X_k(1-X_k)$ and obtained similar results.

iii) If there are biological replicates available, we may extend (1.1) as follows. Let Y_{jkh} denote the expression level for the hth biological replicate at the jth spot for the sample taken at time point X_k. An appropriate model might be

$$Y_{jkh} = F(X_k) + \mu_h + \epsilon_{jkh}.$$

Here $\{\epsilon_{jkh} \,|\, j = 1, \cdots, m_k, k = 0, \cdots, K, h = 1, \cdots, H\}$ are independent normal random variables with mean 0 and variance

$$\sigma_{kh}^2 = \sigma_h^2(F(X_k) + \mu_h)(1 + X_k(1 - X_k)^2).$$

iv) Readers are referred to Altman (2005), Lee (2004) and Parmigiani et al. (2003) and references therein for statistical approaches dealing with replication in microarray experiments.

5. Discussion

We illustrated a Bayesian shape restricted regression method for the inference on the time course expression profile of a virus gene. The simulation study and the analysis of a real dataset seem to suggest that this method is useful in the study of virus gene expression, which in turn can be used to study the mechanism of the regulation of virus gene transcription.

As a method for shape restricted regression model, it features an assessment of the strength of the evidence provided by the data in favor of a hypothesis on its shape and convenient estimates of its salient features. We hope these features would make Bayesian approach to other shape restricted inference problems, including isotonic regression and concave regression, even more appealing.

Our Bayesian method is suitable for the inference on the time expression profile of a single virus gene. To extend our method for the simultaneous analysis of many genes deserves our attention and is under investigation in the framework of hierarchical Bayesian models. Simultaneous analysis of all 156 genes would allow the use of shrinkage-type approaches that pool information across genes, and thus lead to more efficient estimators. This, in turn, would enhance our understanding of the regulation of virus gene transcription. The results of these studies will be reported elsewhere.

Acknowledgements

We are grateful to Prof. Xiao-Li Meng for his comments on an earlier version of this paper, which leads to improvements of the paper in several ways. Part of this work is supported by NSC grant NSC 94-3112-B-400-002-Y.

Appendix: algorithm (RJMH for the posterior)

Let $B_{(n)} = \{(c, n, a_0, \cdots, a_n, \mu) \,|\, c \in [0,1), (a_0, \cdots, a_n) \in B_n, \mu \in \mathbb{R}^1\}$ and the current state $x^{(t)} = (c, n, a_0, \cdots, a_n, \mu) \in B_{(n)}$. We describe the transition from $x^{(t)} \in B_{(n)}$ to a new point $x^{(t+1)}$ as follows.

Randomly select one of three types of moves, say H, H^+, or H^-. Here H is a transition of element in $B_{(n)}$, H^+ a transition of element from $B_{(n)}$ to $B_{(n+1)}$, and H^- a transition of element from $B_{(n)}$ to $B_{(n-1)}$, respectively. The probabilities of selecting the three different types of moves H, H^+, and H^-, when the current state of the Markov chain is in $B_{(n)}$, are respectively denoted by P^n_H, $P^n_{H^+}$, and $P^n_{H^-}$. We set $P^3_{H^-} = P^{n_0}_{H^+} = 0$, $P^n_H = 1 - P^n_{H^+} - P^n_{H^-}$, $P^n_{H^+} = \gamma \min\{1, \frac{\pi_2(n+1)}{\pi_2(n)}\}$, and $P^n_{H^-} = \gamma \min\{1, \frac{\pi_2(n-1)}{\pi_2(n)}\}$, where γ is a sample parameter.

Suppose that $0 \leq F < M$.

If the move of type H is selected, then

1. select k randomly from $\{0, 1, \cdots, n\}$ so that there is $1/3$ probability of choosing 0 or 1; there is $1/3(n-1)$ probability of choosing any one of $\{2, 3, \cdots, n\}$;

2. if $k = 0$, then generate $W \sim \pi_4$ and let $y^{(t)}$ be the vector $x^{(t)}$ with μ replaced by W; if $k = 1$, then generate $W \sim \pi_1$ and let $y^{(t)}$ be the vector $x^{(t)}$ with c replaced by W; if $2 \leq k \leq n$, then generate $W \sim$ *Uniform*(0,M) and let $y^{(t)}$ be the vector $x^{(t)}$ with a_k replaced by W;

3. set the next state

$$x^{(t+1)} = \begin{cases} y^{(t)} & \text{, with prob. } \rho = \min\{1, \dfrac{\nu(y^{(t)})}{\nu(x^{(t)})}\}, \\ x^{(t)} & \text{, o.w.} \end{cases}$$

If the move of type H^+ is selected, then

1. select k randomly from $\{1, 2, \cdots, n\}$ and generate $W \sim$ *Uniform*(0,M);

2. let

$$y^{(t)} = \begin{cases} (c, n+1, a_0, a_1, \cdots, a_k, W, a_{k+1}, \cdots, a_n, \mu), & \text{if } k = \{1, 2, \cdots, n-1\}, \\ (c, n+1, a_0, a_1, \cdots, a_k, a_{k+1}, \cdots, a_n, W, \mu), & \text{if } k = n; \end{cases}$$

3. set the next state

$$x^{(t+1)} = \begin{cases} y^{(t)} & \text{, with prob. } \rho = \min\{1, \dfrac{\nu(y^{(t)}) \times \pi_2(n) \times M}{\nu(x^{(t)}) \times \pi_2(n+1)}\}, \\ x^{(t)} & \text{, o.w.} \end{cases}$$

If the move of type H^- is selected, then

1. select k uniformly from $\{2, 3, \cdots, n\}$;

2. let $y^{(t)} = (c, n-1, a_0, a_1, \cdots, a_{k-1}, a_{k+1}, \cdots, a_n, \mu)$;

3. set the next state

$$x^{(t+1)} = \begin{cases} y^{(t)} & \text{, with prob. } \rho = \min\{1, \dfrac{\nu(y^{(t)}) \times \pi_2(n)}{\nu(x^{(t)}) \times \pi_2(n-1) \times M}\}, \\ x^{(t)} & \text{, o.w.} \end{cases}$$

References

ALTMAN, N. (2005). Replication, variation and normalisation in microarray experiments. *Applied Bioinformatics* **4**, 33-44.

BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics* **26**, 607-616.

CHANG, I. S., CHIEN, L. C., HSIUNG, C. A., WEN, C. C. AND WU, Y. J. (2006). Shape restricted regression with random Bernstein polynomials. Accepted for the Vardi Volume, *IMS Lecture Notes — Monograph Series*.

CHANG, I. S., HSIUNG, C. A., WU, Y. J. AND YANG, C. C. (2005). Bayesian survival analysis using Bernstein polynomials. *Scandinavian Journal of Statistics* **32**, 447-466.

DETTE, H., NEUMEYER, N. AND PILZ, K. F. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli* **12**, 469-490.

DUNSON, D. B. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association* **100**, 618-627.

FRIESEN, P. D. AND MILLER, L. K. (2001). Insect viruses. In Knipe, D. M., Howley, P. M., Griffin, D. E., Martin, M. A., Lamb, R. A., Roizman, B. and Straus, S. E. (eds), *Fields' Virology*, 4nd ed. Philadelphia: Lippincott Williams & Wilkins, pp. 608-609.

GELMAN, A., CARLIN, J. B., STERN, H. S. AND RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Boca Raton: Chapman & Hall/CRC.

GELMAN, A. AND RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457-511.

GIJBELS, I. (2003). Monotone regression. Discussion paper 0334, Institute de Statistique, Université Catholique de Louvain. http://www.stat.ucl.ac.be.

GREEN, P. G. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

HILDRETH, C. (1954). Point estimate of ordinates of concave functions. *Journal of the American Statistical Association* **49**, 598-619.

JIANG, S. S., CHANG, I. S., HUANG, L. W., CHEN, P. C., WEN, C. C., LIU, S. C., CHIEN, L. C., LIN, C. Y., HSIUNG, C. A. AND JUANG, J. L. (2006). Temporal transcription program of recombinant *Autographa californica* multiple nucleopolyhedrosis virus. *Journal of Virology* **80**, 8989-8999.

KASS, R. E. AND RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

LAVINE, M. AND MOCKUS, A. (1995). A nonparametric Bayes method for isotonic regression. *Journal of Statistical Planning and Inference* **46**, 235-248.

LAVINE, M. AND SCHERVISH, M. J. (1999). Bayes factors: what they are and what they are not. *The American Statistician* **53**, 119-122.

LEE, M.-L.T. (2004). *Analysis of Microarray Gene Expression Data.* Boston: Kluwer Academic Publishers.

PARMIGIANI, G., GARRETT, E. S., IRIZARRY, R. A. AND ZEGER, S. L. (2003). *The Analysis of Gene Expression Data: Methods and Software.* New York: Springer-Verlag.

Table 1. Simulation study for $F(t) = 200(t - 0.02)^{5/2}(1 - 0.7t)^5 I_{(0.02,1]}(t)$ with $n_0 = 25$.

Chain	c	n	μ
1	0.20123	5	0.00003
2	0.04928	7	0.00025
3	0.28551	10	0.00038
4	0.39481	14	0.00028
5	0.16573	18	0.00005

Table 1a. Initial values of the five chains.

Estimand	$\hat{R}$
Ton	1.0232
Tmax	1.0035
Max	1.0016
Tslope	1.0051
Slope	1.0106
L_1-norm	1.0004

Table 1b. Gelman-Rubin statistics of the six estimands of interest.

$[0, \tau]$	[0, 0.6667]	[0, 0.8333]	[0, 1.0000]
Po	1.0000	0.9961	0.7641
Pr	0.4567	0.3555	0.1738
Po/Pr	2.1896	2.8020	4.3964
Bf	∞	463.0354	15.3950

Table 1c. Posterior probability (Po), prior probability (Pr), the ratio of Po to Pr, and the Bayes factor (Bf) of being unimodal on $[0,\tau]$.

Po	1.0000
Pr	0.4299
Po/Pr	2.3261
Bf	∞

Table 1d. Posterior probability (Po), prior probability (Pr), the ratio of Po to Pr, and the Bayes factor (Bf) that it is increasing before reaching its global maximum.

Estimand (True)		Mode	Mean	Standard deviation	Support
Ton	Posterior	0.0004	0.0170	0.0079	(0.0000, 0.0358)
(0.0200)	Prior		0.2084	0.1202	(0.0000, 0.4166)
Tmax	Posterior	0.5000	0.4897	0.0222	(0.3750, 0.6157)
(0.4895)	Prior		0.8156	0.2008	(0.1759, 1.0000)
Max	Posterior	3.5274	3.3187	0.1430	(2.6498, 3.8282)
(3.7077)	Prior		2.9024	0.6259	(0.8620, 4.2663)
Tslope	Posterior	0.2176	0.2057	0.0142	(0.0972, 0.3194)
(0.2291)	Prior		0.4949	0.3264	(0.0463, 1.0000)
Slope	Posterior	12.1214	11.9005	0.7933	(8.5898, 15.7563)
(12.9943)	Prior		19.5399	12.7743	(2.7513, 116.0165)
L_1-norm	Posterior	2.0014	1.9265	0.0723	(0.9078, 2.2280)
(2.0281)	Prior		1.3447	0.3653	(0.2556, 2.5731)
Tend	Posterior	1.0000	0.9941	0.0210	(0.5231, 1.0000)
(1.0000)	Prior		0.9548	0.0970	(0.3333, 1.0000)

Table 1e. The mode, mean, standard deviation and support of the posterior probability distribution and the prior probability distribution of Ton, Tmax, Max, Tslope, Slope, L_1-norm and Tend.

Table 2. The data of the gene *lef2*.

X_k	y_{1k}	y_{2k}	y_{3k}	y_{4k}
0	0.0024	0.0057	0.0022	0.0044
1/216	0.0011	0.0004	0.0011	0.0005
1/108	0.0015	0.0010	0.0018	0.0013
1/72	-0.0004	0.0000	0.0004	0.0008
1/36	0.0012	0.0012	0.0017	0.0010
1/24	0.0017	0.0021	0.0020	0.0020
1/12	0.0123	0.0136	0.0107	0.0107
1/8	0.1576	0.1573	0.1291	0.1449
1/6	0.2715	0.2246	0.2644	0.2584
5/24	0.1237	0.1251	0.0850	0.1186
1/4	1.8829	1.5837	1.1372	1.2720
1/3	2.7286	2.2877	3.0697	2.7389
5/12	4.2684	2.7731	3.1413	3.0366
2/3	2.1904	2.1640	2.4436	2.4903
5/6	1.1919	1.1907	1.2578	1.2287
1	1.0117	0.9564	0.9913	0.8631

Table 3. Data analysis for the gene *lef2* with $n_0 = 25$.

Chain	c	n	μ
1	0.2012	5	0.0005
2	0.0493	7	0.0044
3	0.2855	10	0.0067
4	0.3948	14	0.0050
5	0.1657	18	0.0009

Table 3a. Initial values of the five chains.

Estimand	$\hat{R}$
Ton	1.0217
Tmax	1.0036
Max	1.0002
Tslope	1.0033
Slope	1.0135
L_1-norm	1.0028

Table 3b. Gelman-Rubin statistics of the six estimands of interest.

$[0, \tau]$	[0, 0.6667]	[0, 0.8333]	[0, 1.0000]
Po	0.9999	0.8618	0.0346
Pr	0.4567	0.3555	0.1738
Po/Pr	2.1894	2.4242	0.1991
Bf	11895.0700	11.3051	0.1702

Table 3c. Posterior probability (Po), prior probability (Pr), the ratio of Po to Pr, and the Bayes factor (Bf) of being unimodal on [0,τ].

Po	1.0000
Pr	0.4299
Po/Pr	2.3261
Bf	∞

Table 3d. Posterior probability (Po), prior probability (Pr), the ratio of Po to Pr, and the Bayes factor (Bf) that it is increasing before reaching its global maximum.

Estimand		Mode	Mean	Standard deviation	Support
Ton	Posterior	0.0739	0.0734	0.0037	(0.0497, 0.0804)
	Prior		0.2084	0.1202	(0.0000, 0.4166)
Tmax	Posterior	0.4444	0.4570	0.0140	(0.4028, 0.5093)
	Prior		0.8156	0.2008	(0.1759, 1.0000)
Max	Posterior	3.4330	3.3976	0.1208	(2.9256, 3.8310)
	Prior		2.9035	0.6261	(0.8623, 4.2679)
Tslope	Posterior	0.2778	0.2790	0.0054	(0.2593, 0.3009)
	Prior		0.4949	0.3264	(0.0463, 1.0000)
Slope	Posterior	17.4965	16.7663	0.6313	(14.7583, 20.0061)
	Prior		19.5474	12.7792	(2.7524, 116.0608)
L_1-norm	Posterior	1.5539	1.5832	0.0699	(0.9294, 1.8167)
	Prior		1.3452	0.3655	(0.2557, 2.5741)
Tend	Posterior	0.8519	0.8956	0.0566	(0.6019, 1.0000)
	Prior		0.9548	0.0970	(0.3333, 1.0000)

Table 3e. The mode, mean, standard deviation and support of the posterior probability distribution and the prior probability distribution of Ton, Tmax, Max, Tslope, Slope, L_1-norm and Tend.