

# Estimation and Model Choice in Nonparametric Additive Regression

SIDDHARTHA CHIB

*Washington University in St. Louis*

IVAN JELIAZKOV

*University of California, Irvine*

(Munich, October 2005)

## 1 Setting

- Given data  $\{y_i, \mathbf{s}_i\}_{i=1}^n$ , where  $\mathbf{s}_i$  is  $p \times 1$
- Interested in analyzing nonparametric additive models (Hastie and Tibshirani 1990):

$$y_i = c + g_1(s_{i1}) + \dots + g_p(s_{ip}) + \varepsilon_i, (i = 1, \dots, n),$$

where the  $g_j(\cdot)$  are unknown functions

- This model offers a convenient way of modeling multivariate nonlinearity, at the cost of the additivity assumption
- Such additive components can be inserted in more general models, for example, in models with unobserved confounding

### The Setting (Continued)

- In general, fitting of these models is not without some difficulties
- One key issue is the modeling of the unknown  $g_j(\cdot)$  - what amounts to the prior in the Bayesian context
- Of course, the level of these functions is not identified and therefore the functions must be “anchored” in some suitable way
- Another issue is the appropriate way of computing/summarizing the resulting high-dimensional posterior distribution

### Main Goals

- A formulation and comparison of a class of (smoothness) priors under different identification schemes
- The development of efficient MCMC-based algorithms for conducting prior-posterior analysis
- Computation of marginal likelihood and Bayes factors to address model uncertainty and permit the comparison of parametric vs nonparametric and semiparametric models
- Comparison with AIC and BIC

## 2 The Prior on $g(\cdot)$

- The class of prior we focus on is the so-called Markov process smoothness prior (e.g. Fahrmeir and Lang 2001)
- Each function is assumed a priori independent
- The prior is conceptually simple
- Versions of this prior have been widely used in the literature

- Consider the  $j$ th function  $g_j(s_j)$  and let

$$v_{j1} < v_{j2} < \dots < v_{jm_j}$$

denote the  $m_j$  unique ordered values of  $s_j$  with ordinates  $g_{jt} = g_j(v_{jt})$ ,  $t = 1, \dots, m_j$

- If we define  $h_{jt} = v_{jt} - v_{j,t-1}$ , the prior is induced through a second order random walk process for the functional evaluations:

$$g_{jt} = \left(1 + \frac{h_{jt}}{h_{j,t-1}}\right) g_{j,t-1} - \frac{h_{jt}}{h_{j,t-1}} g_{j,t-2} + u_{jt},$$

and a distribution on  $(g_{j1}, g_{j2})$ , where

$$u_{jt} \sim N(0, \tau_j^2 h_{jt})$$

- $\tau_j^2$  is a smoothness parameter

- **NOTE:** In our work, we do not use an improper prior on  $(g_{j1}, g_{j2})$  (new in this context)
- We mainly work under one of two sets of assumptions:  $g_{j1} = 0$  and  $g_{j2}$  is  $N(g_{j20}, \tau_j^2)$  or

$$\begin{pmatrix} g_{j1} \\ g_{j2} \end{pmatrix} | \tau_j^2 \sim N_2 \left( \begin{pmatrix} g_{j10} \\ g_{j20} \end{pmatrix}, \tau_j^2 \mathbf{G}_{j0} \right)$$

- In each case, this leads to a convenient joint distribution on the free unknown functional evaluations (namely  $\mathbf{g}_j = (g_{j2}, \dots, g_{jm_j})$  or  $\mathbf{g}_j = (g_{j1}, \dots, g_{jm_j})$ )



- We can rewrite the Markov process as

$$\mathbf{g}_j | \tau_j^2 \sim N_{m_j} (\mathbf{g}_{j0}, \tau^2 \mathbf{K}_j^{-1})$$

where

$$\mathbf{g}_{j0} = \mathbf{\Delta}_j^{-1} \tilde{\mathbf{g}}_j,$$

$\tilde{\mathbf{g}}_j = (g_{j10}, g_{j20}, 0, \dots, 0)$  and  $\mathbf{K}_j = \mathbf{\Delta}_j' \mathbf{H}_j^{-1} \mathbf{\Delta}_j$  is a *full rank* penalty matrix

- This distribution is *proper*  $\Rightarrow$  Bayes factors are well defined
- Also easy to show that the penalty matrix  $\mathbf{K}_j$  is banded

- Other priors
  - $\tau_i^2 \sim IG(\nu_{0i}/2, \delta_{0i}/2)$
  - $\sigma^2 \sim IG(\eta_0/2, \gamma_0/2)$
  - $c \sim N(c_0, C_0)$

### Identification Restrictions

- **Note:** additive modelling  $\Rightarrow$   $\mathbf{g}$  should be appropriately centered for identification
- **Approach I:** to “anchor” the functions, set  $g_{j1} = 0$  in which case  $\Delta_j$  is

$$\begin{pmatrix} 1 & \dots & \dots & \dots & \dots \\ -\left(1 + \frac{h_{j3}}{h_{j2}}\right) & 1 & \dots & \dots & \dots \\ \frac{h_{j4}}{h_{j3}} & -\left(1 + \frac{h_{j4}}{h_{j3}}\right) & 1 & \dots & \dots \\ \ddots & \ddots & \ddots & \dots & \vdots \\ \dots & \dots & \dots & \frac{h_{jm_j}}{h_{j,m_j-1}} & -\left(1 + \frac{h_{jm_j}}{h_{j,m_j-1}}\right) & 1 \end{pmatrix}$$

and  $\tilde{\mathbf{g}}_j = (g_{j20}, 0, \dots, 0)$

- The model can now be written as

$$\mathbf{y} = \mathbf{i}c + \mathbf{Q}_1\mathbf{g}_1 + \mathbf{Q}_2\mathbf{g}_2 + \dots + \mathbf{Q}_p\mathbf{g}_p + \varepsilon$$

–  $\{\mathbf{Q}_j\}$  are incidence matrices

**Identification II**

- **Approach II:** “Anchor” the functions by centering  $\{\mathbf{g}_j\}$  in the likelihood and then expand by the incidence matrices  $\{\mathbf{Q}_j\}$

$$\mathbf{y} = \mathbf{ic} + \mathbf{Q}_1\mathbf{M}_1\mathbf{g}_1 + \mathbf{Q}_2\mathbf{M}_2\mathbf{g}_2 + \dots + \mathbf{Q}_p\mathbf{M}_p\mathbf{g}_p + \boldsymbol{\varepsilon}$$

- $\mathbf{M}_j$  is a mean-differencing symmetric and idempotent matrix

$$\mathbf{M}_j = \left( \mathbf{I} - \frac{\mathbf{ii}'}{m_j} \right)$$

- each vector  $\mathbf{M}_j\mathbf{g}_j$  sums to zero
- this removes any arbitrary constants from the model

### Identification II (Continued)

- Two problems: usually estimated by “centering-on-the-fly”
  - sampling proceeds *as if* the model is

$$\mathbf{y} = \mathbf{ic} + \mathbf{Q}_1\mathbf{g}_1 + \mathbf{Q}_2\mathbf{g}_2 + \dots + \mathbf{Q}_p\mathbf{g}_p + \varepsilon$$

- $\mathbf{g}_j$  is drawn and then centered as  $\mathbf{M}_j\mathbf{g}_j$  before other functions are sampled
- formally incorrect. Also sampling does not behave well
- not easily possible to fit the model directly because of non-banded precision matrices in the posterior updates

**Identification III**

- A different approach for centering  $\{\mathbf{g}_j\}$  in the likelihood: the *expanded* vectors  $\mathbf{Q}_j \mathbf{g}_j$  are centered by  $\mathbf{M} = \left(\mathbf{I} - \frac{\mathbf{ii}'}{n}\right)$ , not just the vectors  $\mathbf{g}_j$  as in Approach II, leading to

$$\mathbf{y} = \mathbf{ic} + \mathbf{MQ}_1 \mathbf{g}_1 + \mathbf{MQ}_2 \mathbf{g}_2 + \dots + \mathbf{MQ}_p \mathbf{g}_p + \boldsymbol{\varepsilon}$$

where  $\mathbf{M} = \left(\mathbf{I} - \frac{\mathbf{ii}'}{n}\right)$

- Note that
  - repeated values in  $\mathbf{Q}_j \mathbf{g}_j$  influence the centering
  - the centering constants are different when there are repeating values in  $s_i$ : otherwise the centering constants in approaches II and III are identical
  - the distinction is subtle but important
  - as we show, direct and efficient sampling of the functions from this model is possible

### MCMC Simulation

- Under identification approach I, MCMC sampling proceeds as:

1. For  $j = 1, \dots, p$ ,

$$\begin{aligned} \mathbf{g}_j | \mathbf{y}, \boldsymbol{\psi}, \{\mathbf{g}_i\}_{i \neq j}^p &\sim N_{m_j-1}(\hat{\mathbf{g}}_j, \mathbf{G}_j) \\ \hat{\mathbf{g}}_j &= \mathbf{G}_j (\tau_j^{-2} \mathbf{K}_j \mathbf{g}_{j0} + \sigma^{-2} \mathbf{Q}'_j \tilde{\mathbf{y}}_j) \\ \mathbf{G}_j &= (\tau_j^{-2} \mathbf{K}_j + \sigma^{-2} \mathbf{Q}'_j \mathbf{Q}_j)^{-1} \end{aligned}$$

where  $\tilde{\mathbf{y}}_j = \mathbf{y} - \mathbf{ic} - \sum_{k \neq j} \mathbf{Q}_k \mathbf{g}_k$  and  $\mathbf{G}_j$  is banded

2.  $\tau_j^2 | \mathbf{g}_j$
3.  $\sigma^2 | \mathbf{y}, \{\mathbf{g}_j\}, c$
4.  $c | \mathbf{y}, \{\tau_j^2\}, \sigma^2, \{\mathbf{g}_i\}$

### MCMC Simulation II

- Approach II is problematic: the resulting  $\{\mathbf{G}_j\}$  are not banded
- Sampling is not model based: centering on the fly
- Brute force sampling not possible with large data sets

### MCMC simulation III

- Important computational advantages of proposed new identification scheme
- The covariance matrix  $\mathbf{G}_j$  is not banded but is of the form

$$\begin{aligned}\mathbf{G}_j &= \left( \frac{1}{\tau_j^2} \mathbf{K}_j + \frac{1}{\sigma^2} \mathbf{Q}_j' \mathbf{M} \mathbf{Q}_j \right)^{-1} \\ &= \left( \frac{1}{\tau_j^2} \mathbf{K}_j + \frac{1}{\sigma_j^2} \mathbf{Q}_j' \mathbf{Q}_j - \frac{\mathbf{c}_j \mathbf{c}_j'}{\sigma^2 n} \right)^{-1} \\ &= (\mathbf{A}_j - \mathbf{u}_j \mathbf{u}_j')^{-1}\end{aligned}$$

where  $\mathbf{A}_j$  is banded and  $\mathbf{u}_j$  is a vector

- Useful computational shortcuts can be employed to sample in  $O(n)$  operations, rather than  $O(n^3)$  operations
  - Sherman-Morrison formula

$$\begin{aligned}\mathbf{G}_j &= (\mathbf{A}_j - \mathbf{u}_j \mathbf{u}_j')^{-1} \\ &= \mathbf{A}_j^{-1} + \frac{\mathbf{A}_j^{-1} \mathbf{u}_j \mathbf{u}_j' \mathbf{A}_j^{-1}}{1 - \lambda_j} \\ &= \mathbf{A}_j^{-1} \left( \mathbf{A}_j + \frac{\mathbf{u}_j \mathbf{u}_j'}{1 - \lambda_j} \right) \mathbf{A}_j^{-1}\end{aligned}$$

where  $\lambda_j = \mathbf{u}_j' \mathbf{A}_j^{-1} \mathbf{u}_j$

- Since  $\mathbf{A}_j$  is banded, all operations involving  $\mathbf{A}_j^{-1}$  are  $O(n)$

- Now to simulate  $\mathbf{g}_j$ , let

$$\mathbf{B}_j = \left( \mathbf{A}_j + \frac{\mathbf{u}_j \mathbf{u}_j'}{1 - \lambda_j} \right)$$

and generate  $\mathbf{x}_j \sim N_{m_j}(\mathbf{0}, \mathbf{B}_j)$  as sum of two independent random variables

- Then  $\mathbf{z}_j = \mathbf{A}_j^{-1} \mathbf{x}_j$  has a distribution  $\mathbf{z}_j \sim N(\mathbf{0}, \mathbf{G}_j)$
- Finding  $\mathbf{z}_j = \mathbf{A}_j^{-1} \mathbf{x}_j$  is done by solving  $\mathbf{A}_j \mathbf{z}_j = \mathbf{x}_j$  for  $\mathbf{z}_j$  by backsubstitution in  $O(n)$  operations because  $\mathbf{A}_j$  is banded
- A draw for  $\mathbf{g}_j$  is obtained as  $\mathbf{g}_j = \hat{\mathbf{g}}_j + \mathbf{z}_j$

### 3 Model Comparison

- Important because of model uncertainty
- Competing models  $\{\mathcal{M}_1, \dots, \mathcal{M}_L\}$ , each with model-specific parameter vector  $\boldsymbol{\theta}_l \in S_l \subseteq \mathbb{R}^{k_l}$  and sampling density  $f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l)$
- The ratio of posterior model probabilities for  $\mathcal{M}_i$  and  $\mathcal{M}_j$  is

$$\frac{\Pr(\mathcal{M}_i|\mathbf{y})}{\Pr(\mathcal{M}_j|\mathbf{y})} = \frac{\Pr(\mathcal{M}_i)}{\Pr(\mathcal{M}_j)} \times \frac{m(\mathbf{y}|\mathcal{M}_i)}{m(\mathbf{y}|\mathcal{M}_j)}$$

where

$$m(\mathbf{y}|\mathcal{M}_l) = \int f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l) \pi_l(\boldsymbol{\theta}_l|\mathcal{M}_l) d\boldsymbol{\theta}_l$$

is the marginal likelihood of  $\mathcal{M}_l$ .

- Chib (1995) suggests an alternative to direct marginalization using

$$m(\mathbf{y}|\mathcal{M}_l) = \frac{f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l)\pi(\boldsymbol{\theta}_l|\mathcal{M}_l)}{\pi(\boldsymbol{\theta}_l|\mathbf{y}, \mathcal{M}_l)}$$

- $m(\mathbf{y}|\mathcal{M}_l)$  is obtained by finding an estimate of the posterior ordinate  $\pi(\boldsymbol{\theta}_l^*|\mathbf{y}, \mathcal{M}_l)$
- Suppose that  $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_B^*)$
- Then, by the law of total probability we have

$$\pi(\boldsymbol{\theta}^*|\mathbf{y}) = \pi(\boldsymbol{\theta}_1^*|\mathbf{y})\pi(\boldsymbol{\theta}_2^*|\mathbf{y}, \boldsymbol{\theta}_1^*) \cdots \pi(\boldsymbol{\theta}_B^*|\mathbf{y}, \boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_{B-1}^*)$$

- Above ordinates estimated in reduced runs

- Because  $\{\mathbf{g}_j\}$  are high-dimensional, care is needed in estimating the marginal likelihood
- Two ways of dealing with this issue
  - Direct analytical marginalization over  $\{\mathbf{g}_j\}$
  - Reduced run method: carefully place the  $\{\mathbf{g}_j\}$  in the posterior decomposition used for the marginal likelihood computation

- **Direct marginalization method** relies on

$$m(\mathbf{y}) = \frac{f(\mathbf{y} | \{\tau_i^{2*}\}, \sigma^{2*}) \pi(\tau_i^{2*}, \sigma^{2*})}{\pi(\{\tau_i^{2*}\}, \sigma^{2*} | \mathbf{y})}$$

- All densities are marginalized over  $\{\mathbf{g}_j\}, c$
- $f(\mathbf{y} | \{\tau_i^{2*}, \sigma^{2*}\})$  is analytically available in  $O(n^3)$  operations for Gaussian problems
  - involves an  $n \times n$  non-banded matrix
  - saves further reduced runs
  - useful when  $n$  is small

- **Reduced run method** relies on

$$m(\mathbf{y}) = \frac{f(\mathbf{y} | \{\tau_i^{2*}\}, \sigma^{2*}, \{\mathbf{g}_j^*\}, c^*) \pi(\tau_i^{2*}, \sigma^{2*}, \{\mathbf{g}_j^*\}, c^*)}{\pi(\{\tau_i^{2*}\}, \sigma^{2*}, \{\mathbf{g}_j^*\}, c^* | \mathbf{y})},$$

- requires  $p - 1$  reduced runs
- $\{\mathbf{g}_j\}$  are placed last in the posterior decomposition to improve statistical efficiency.
- applicable and efficient even for large dimensional  $\{\mathbf{g}_j\}$  (e.g.  $m_j$  in the thousands)
- Little computational cost because sampling of  $\{\mathbf{g}_j\}$  is  $O(n)$ 
  - useful for large data sets

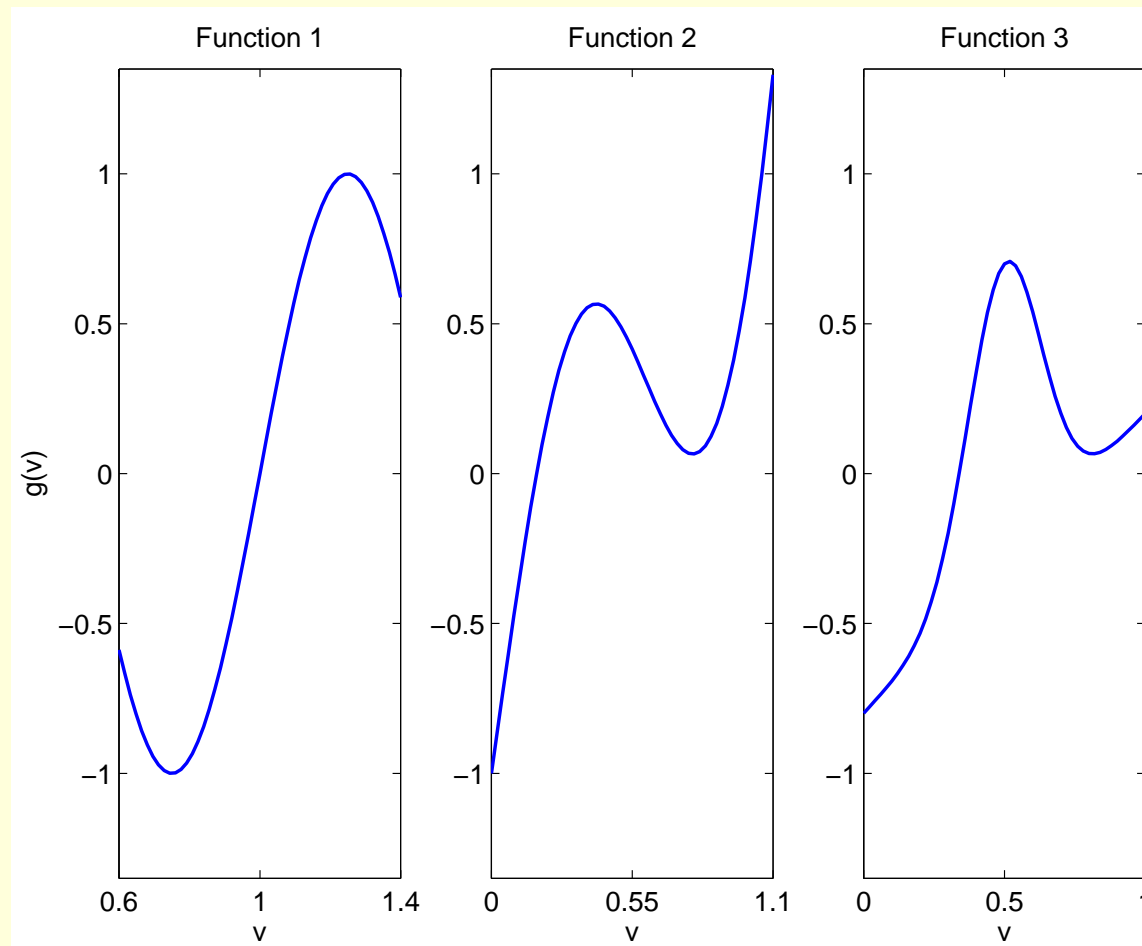
- The identification scheme is also useful in marginal likelihood computation
- Calculating the determinants of the full conditional precision matrices for  $\{\mathbf{g}_j\}$  takes  $O(n)$ , rather than the usual  $O(n^3)$ , operations
- Determinant computed as follows

$$\begin{aligned}\det(\mathbf{A}_j - \mathbf{u}_j \mathbf{u}_j') &= \det\{\mathbf{A}_j (\mathbf{I} - \mathbf{A}_j^{-1} \mathbf{u}_j \mathbf{u}_j')\} \\ &= \det(\mathbf{A}_j) (1 - \mathbf{u}_j' \mathbf{A}_j^{-1} \mathbf{u}_j)\end{aligned}$$

## 4 Simulation Study

- An artificial data study suggests that the method performs well
- The estimation method recovers the parameters and functions used to generate the data
- The model selection method selects the correct model
- Estimation is very efficient (computationally and statistically)

## True Functions in the Study



- data are generated from an additive model with  $\sigma^2 = 0.25^2 = 0.0625$
- some resulting descriptive signal-to-noise statistics for the functions in the simulation study:

Generated Functions			
	$g_1$	$g_2$	$g_3$
$SD(g_i)/\sigma$	3.10	1.81	1.85
$Range(g_i)/\sigma$	7.99	9.32	6.03

### Priors Used in Simulation Study

- $\sigma^2 \sim IG(3, 0.125) \Rightarrow E(\sigma^2) = SD(\sigma^2) = 0.0625$
- $\tau_j^2 \sim IG(3, 0.05) \Rightarrow E(\tau_j^2) = SD(\tau_j^2) = 0.025$ , for  $j = 1, 2, 3$ .
- The initial conditions were given by

$$\begin{pmatrix} g_{j1} \\ g_{j2} \end{pmatrix} | \tau_j^2 \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tau_j^2 \mathbf{G}_{j0} \right),$$

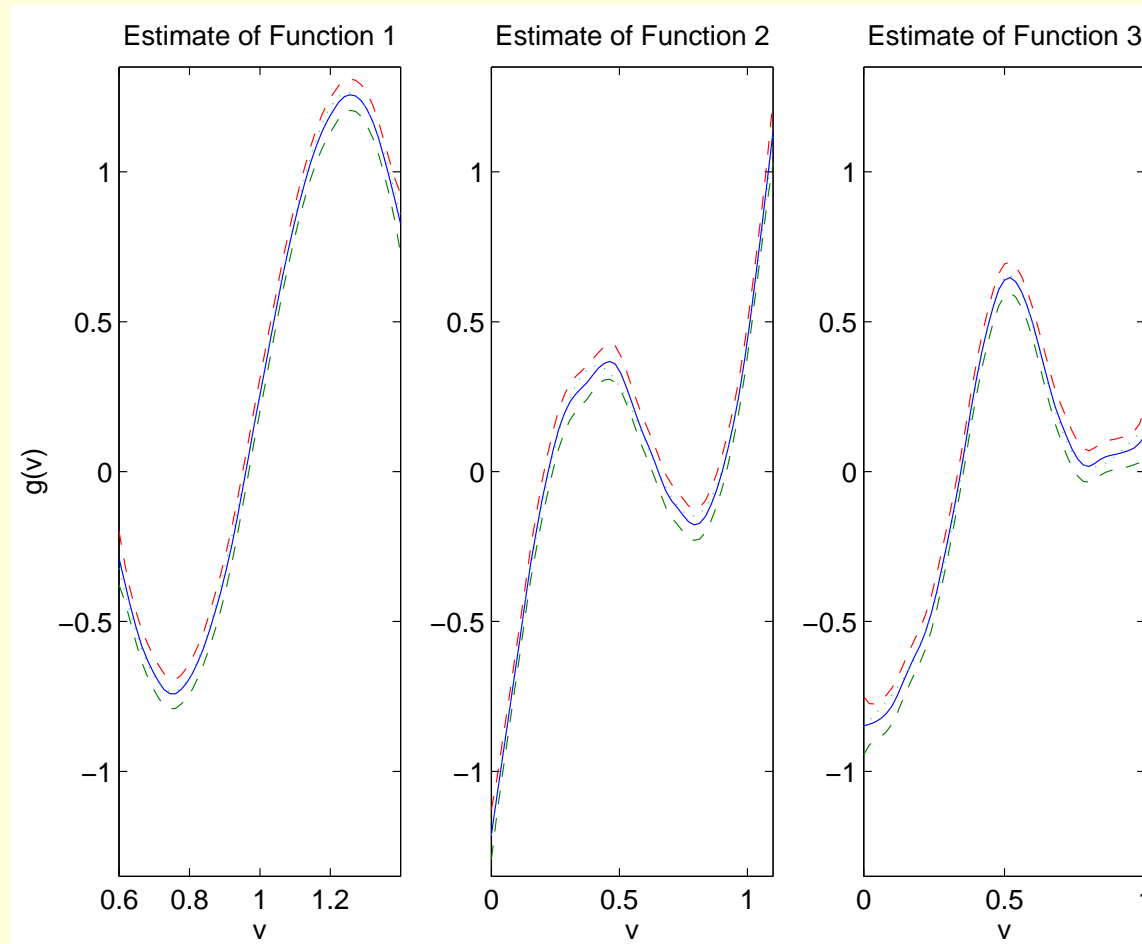
where  $\mathbf{G}_{j0} = \frac{10 * \mathbf{I}_2}{E(\tau_j^2)}$ .

### Function Estimates

- Average mean squared errors based on 15 samples, with estimated standard errors in parentheses

Average Mean Squared Errors			
Observations	$g_1$	$g_2$	$g_3$
$n = 250$	0.00260 (0.00157)	0.00514 (0.00437)	0.00837 (0.00848)
$n = 500$	0.00088 (0.00052)	0.00292 (0.00486)	0.00302 (0.00268)
$n = 1000$	0.00055 (0.00043)	0.00235 (0.00187)	0.00245 (0.00301)

## Examples of Estimated Functions

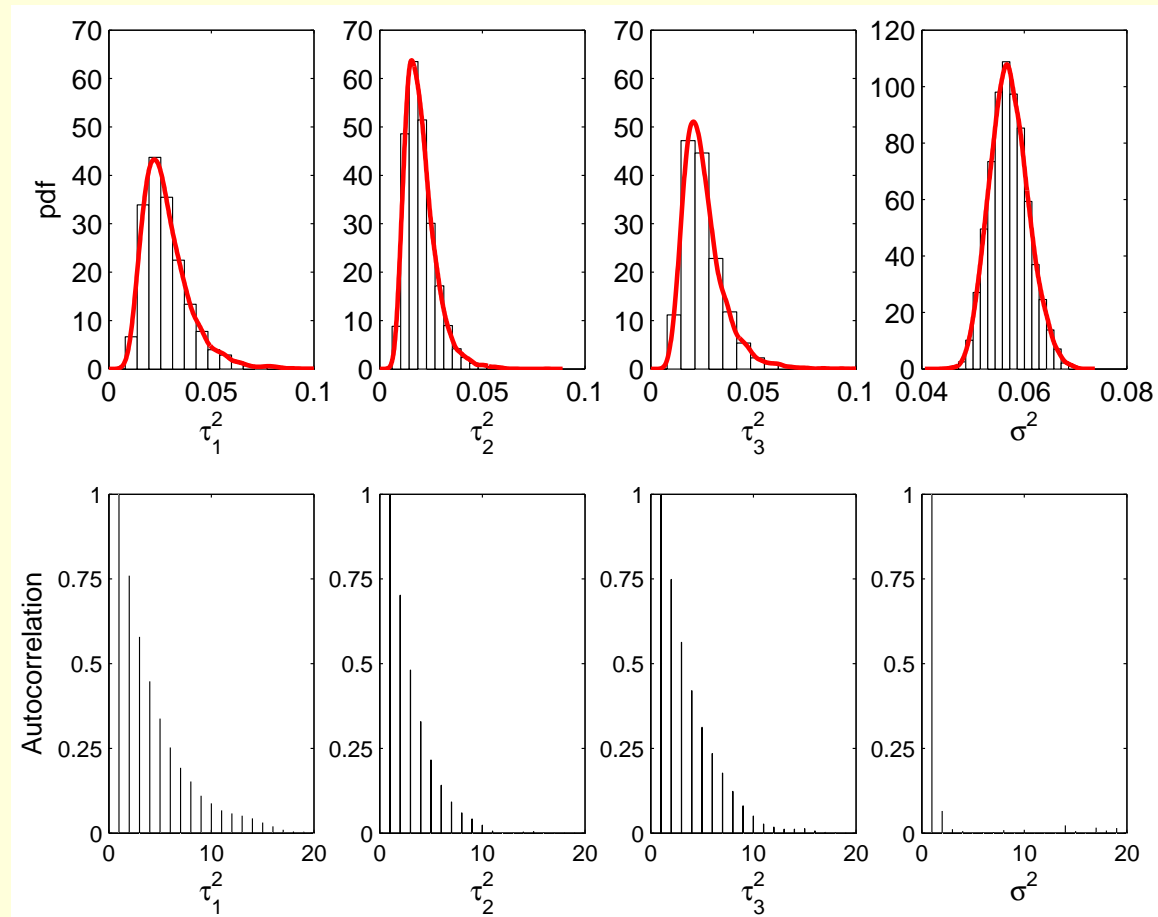


### Parameter Estimates

- Examples of estimated inefficiency factors (autocorrelation times) for the parameters of the additive model for various sample sizes

	Inefficiency Factors			
	$\tau_1^2$	$\tau_2^2$	$\tau_3^2$	$\sigma^2$
$n = 250$	8.885	6.261	8.286	1.220
$n = 500$	6.684	4.872	5.488	1.068
$n = 1000$	5.758	5.689	4.966	1.000

## Posterior Distributions of Parameters



### Performance of the Model Selection Technique

- Compared three nonparametric models
- Realizations generated from an additive model using the first two functions above
- Three fitted models
  - first model has only one covariate
  - second model used both relevant covariates
  - third model includes an additional irrelevant covariate
- Correct model selected every time

### Model Selection (Continued)

- **Note:** Placing  $\{\mathbf{g}_j\}$  last in the posterior decomposition reduces the variability of the marginal likelihood estimates:
  - reduced run method almost as efficient as direct marginalization
  - alternative decomposition, where  $\{\mathbf{g}_j\}$  are placed first in the posterior decomposition is not as precise
  - direct marginalization takes more time than reduced run method when the sample size  $n > 2000$  (approximately).

- Marginal likelihood NSEs for three  $n$  (holding  $m = 51$ )

NSE of the Marginal Likelihood Estimate			
	Reduced Run	Direct Marginalization	Alt. Decomposition
$n = 250$	0.013	0.013	0.150
$n = 500$	0.009	0.009	0.116
$n = 1000$	0.008	0.008	0.083

- Marginal likelihood NSEs for three  $m$  (holding  $n = 2500$ )

NSE of the Marginal Likelihood Estimate		
	Reduced Run	Direct Marginalization
$m = 501$	0.049	0.049
$m = 1001$	0.061	0.059
$m = 1501$	0.086	0.065

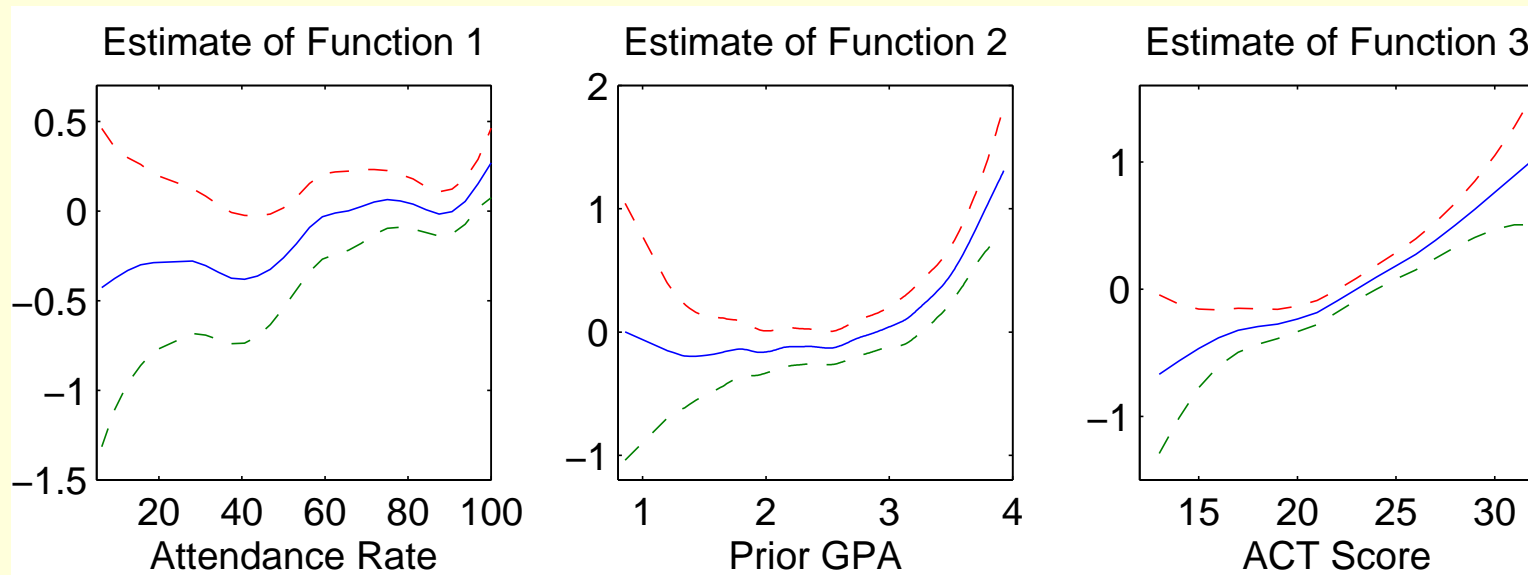
## 5 Marginal Likelihoods vs. AIC and BIC

- Estimating the marginal likelihood is less computationally intensive – compare with Shively et al. (1999), Wood et al. (2002), DiMatteo et al. (2001), Hansen and Kooperberg (2002), who do not calculate Bayes factors.
  - Why? Estimation of the marginal likelihood does not require maximization.
- AIC and BIC estimation not always feasible
  - maximization is demanding or infeasible in many settings
- Bayes factors provide finite sample model probabilities

## 6 Application to Exam Scores

- 680 college students in Introduction to Microeconomics
  - data from Wooldridge (2002)
- Predict *relative performance*,  $y$ , as a function of
  - attendance ( $s_1$ )
  - cumulative GPA ( $s_2$ )
  - ACT scores ( $s_3$ )
- Model is

$$y_i = g_1(s_1) + g_2(s_2) + g_3(s_3) + \varepsilon_i$$

**Estimated Functions**

### Model Comparison

- Alternative specifications also considered
  - hypothesis: the effect of class attendance might be different for “good” vs “bad” students – estimated additive models including  $g_4(s_1s_2)$  or  $g_4(s_1s_3)$
  - these models did not perform competitively based on Bayes factors
  - additive separability appears to be a reasonable restriction in this context

## 7 Concluding Remarks

- Examined the specification, estimation, and comparison of nonparametric additive models based on proper smoothness priors
- New identification scheme allows for efficient MCMC estimation
- Two methods for calculating the marginal likelihood
  - allow for computationally and statistically efficient model selection
- Techniques can be applied in many other settings, including binary and censored responses, longitudinal outcomes, and problems with unobserved confounders