Bayesian Nonparametric Modelling with the Dirichlet Process Regression Smoother

J. E. Griffin and M. F. J. Steel

University of Warwick

Introduction

- Nonparametric regression offers flexibility that many real applications require

Introduction

- Nonparametric regression offers flexibility that many real applications require
- Nonlinear relationships with minimal assumptions
- Interest could be in various aspects (cond. location, cond. spread, . . .)

Introduction

- Nonparametric regression offers flexibility that many real applications require
- Nonlinear relationships with minimal assumptions
- Interest could be in various aspects (cond. location, cond. spread, . . .)
- Existing Bayesian approaches: flexible location modelling (Gaussian Processes, splines, wavelets) and local modelling (partial exchangeability, BPM)

Introduction

- Nonparametric regression offers flexibility that many real applications require
- Nonlinear relationships with minimal assumptions
- Interest could be in various aspects (cond. location, cond. spread, . . .)
- Existing Bayesian approaches: flexible location modelling (Gaussian Processes, splines, wavelets) and local modelling (partial exchangeability, BPM)
- Here we attempt to combine Bayesian NP function estimation and density estimation
- We also want to allow for centring over parametric models

Bayesian Nonparametric Modelling

Usual hierarchical Bayesian model for $y_1, \ldots, y_n$:

$$y_i \sim k(\psi_i), \quad \psi_i \sim F, \quad F \sim \Pi,$$

where $k(\cdot)$ is pdf and Π is flexible class

Bayesian Nonparametric Modelling

Usual hierarchical Bayesian model for $y_1, \ldots, y_n$:

$$y_i \sim k(\psi_i), \quad \psi_i \sim F, \quad F \sim \Pi,$$

where $k(\cdot)$ is pdf and Π is flexible class

Here concentrate on stick-breaking class

$$F \stackrel{d}{=} \sum_{i=1}^{\infty} p_i \delta_{\theta_i},$$

δ_θ Dirac measure at θ and $p_i = V_i \prod_{j<i}(1 - V_j)$
$V_1, V_2, V_3, \ldots$ independent with $V_i \sim \text{Be}(a_i, b_i)$
$\theta_1, \theta_2, \theta_3, \ldots$ i.i.d. from centring distribution H

Bayesian Nonparametric Modelling

Usual hierarchical Bayesian model for $y_1, \ldots, y_n$:

$$y_i \sim k(\psi_i), \quad \psi_i \sim F, \quad F \sim \Pi,$$

where $k(\cdot)$ is pdf and Π is flexible class

Here concentrate on stick-breaking class

$$F \stackrel{d}{=} \sum_{i=1}^{\infty} p_i \delta_{\theta_i},$$

δ_θ Dirac measure at θ and $p_i = V_i \prod_{j<i}(1 - V_j)$
$V_1, V_2, V_3, \ldots$ independent with $V_i \sim \text{Be}(a_i, b_i)$
$\theta_1, \theta_2, \theta_3, \ldots$ i.i.d. from centring distribution H

Dirichlet process when $a_i = 1$ and $b_i = M$ for all i

Bayesian Nonparametric Regression

For pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ a natural extension is

$$y_i \sim k(\psi_i), \quad \psi_i \sim F_x, \quad F_x \stackrel{d}{=} \sum_{j=1}^{\infty} p_j(x)\delta_{\theta_j(x)}.$$

Covers existing processes:

Bayesian Nonparametric Regression

For pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ a natural extension is

$$y_i \sim k(\psi_i), \quad \psi_i \sim F_x, \quad F_x \stackrel{d}{=} \sum_{j=1}^{\infty} p_j(x)\delta_{\theta_j(x)}.$$

Covers existing processes:

- $p_i(x) = p_i$: single p DDP (MacEachern, 2001; De Iorio *et al.*, 2004; Gelfand *et al.*, 2005)
- $\theta_i(x) = \theta_i$: *e.g.* πDDP (Griffin and Steel, 2006)

Bayesian Nonparametric Regression

Here focus on models with $\theta_i(x) = \theta_i$

Bayesian Nonparametric Regression

Here focus on models with $\theta_i(x) = \theta_i$

Often undersmooth posterior mean (piecewise constant)

Bayesian Nonparametric Regression

Here focus on models with $\theta_i(x) = \theta_i$

Often undersmooth posterior mean (piecewise constant)

So consider:

$$y_i - g(x_i) - m(x_i) \sim k(\psi_i), \quad \psi \sim F_x, \quad F_x \stackrel{d}{=} \sum_{i=1}^{\infty} p_i(x)\delta_{\theta_i},$$

conditional regression function: parametric part $g(x)$ and a nonparametric part $m(x)$

For $m(x)$: Gaussian process prior with mean 0 and covariance $\sigma_0^2\rho(x_i, x_j)$ where $\rho(x_i, x_j)$ is a Matèrn correlation function

Bayesian Density Smoother

Stick-breaking prior: consider the atoms and their ordering at each $x \in \mathbb{R}^p$

Bayesian Density Smoother

Stick-breaking prior: consider the atoms and their ordering at each $x \in \mathbb{R}^p$

Define closed, convex sets in $\mathbb{R}^p$, say $I_1, I_2, \ldots$ and construct $F(x)$ by only considering $\{(V_j, \theta_j) | x \in I_j\}$

Bayesian Density Smoother

Stick-breaking prior: consider the atoms and their ordering at each $x \in \mathbb{R}^p$

Define closed, convex sets in $\mathbb{R}^p$, say $I_1, I_2, \ldots$ and construct $F(x)$ by only considering $\{(V_j, \theta_j) | x \in I_j\}$

Ordering determined by associated $t_j > 0$ (smallest first)

So prior is defined by $(V_1, \theta_1, I_1, t_1), (V_2, \theta_2, I_2, t_2), \ldots$

Bayesian Density Smoother

Stick-breaking prior: consider the atoms and their ordering at each $x \in \mathbb{R}^p$

Define closed, convex sets in $\mathbb{R}^p$, say $I_1, I_2, \ldots$ and construct $F(x)$ by only considering $\{(V_j, \theta_j) | x \in I_j\}$

Ordering determined by associated $t_j > 0$ (smallest first)

So prior is defined by $(V_1, \theta_1, I_1, t_1), (V_2, \theta_2, I_2, t_2), \ldots$

If $p_{s,w} = P(s, w \in I_k | s \in I_k \text{ or } w \in I_k)$ is given, then

$$\text{Corr}(F_s, F_w) = \frac{2(M+1)p_{s,w}}{2 + M(1 + p_{s,w})}$$

Bayesian Density Smoother

For I_k choose a ball of radius r_k around C_k

Bayesian Density Smoother

For I_k choose a ball of radius r_k around C_k

$(C_1, r_1, t_1), (C_2, r_1, t_1), \ldots$: Poisson process on $\mathbb{R}^p \times \mathbb{R}_+^2$ with intensity $p(r)$ (pdf on $\mathbb{R}_+$)

Bayesian Density Smoother

For I_k choose a ball of radius r_k around C_k

$(C_1, r_1, t_1), (C_2, r_1, t_1), \ldots$: Poisson process on $\mathbb{R}^p \times \mathbb{R}_+^2$ with intensity $p(r)$ (pdf on $\mathbb{R}_+$)

Some results for case where $x \in \mathbb{R}$:

$$p_{s,s+u} = \frac{2\mu_2 - uI}{4\mu - 2\mu_2 + uI}$$

where $\mu = \mathsf{E}[r]$, $I = \int_{u/2}^{\infty} p(r)\, dr$ and $\mu_2 = \int_{u/2}^{\infty} rp(r)\, dr$

If $r \sim \mathsf{Ga}(\alpha, \cdot)$, F_x is mean square differentiable of order $q = 1, 2, \ldots$ if and only if $\alpha \geq 2q - 1$.

Dirichlet Process Regression Smoother

Definition: Let (t_i, C_i, r_i) be a Poisson process with intensity $\frac{\beta^\alpha}{\Gamma(\alpha)} r_i^{\alpha-1} \exp\{-\beta r_i\}$ on $\mathbb{R}_+ \times \mathbb{R}^p \times \mathbb{R}_+$ with associated marks (V_i, θ_i) which are i.i.d. from Be$(1, M)$ and H. If

$$F_x = \sum_{\{i|x \in B_{r_i}(C_i)\}} V_i \prod_{\{j|x \in B_{r_j}(C_j), t_j < t_i\}} (1 - V_j)\delta_{\theta_i}$$

then $\{F_x | x \in \mathbb{R}^p\}$ follows a *Dirichlet Process Regression Smoother (DPRS)*, represented as *DPRS*(M, H, α, β)

Centring over Models

Centre nonparametric model over nontrivial parametric model:

Centring over Models

Centre nonparametric model over nontrivial parametric model:

- nonparametric model can indicate flaws in common parametric models
- can aid interpretation and prior elicitation

Centring over Models

Centre nonparametric model over nontrivial parametric model:

- nonparametric model can indicate flaws in common parametric models
- can aid interpretation and prior elicitation

Regression errors: $\epsilon_i = y_i - g(x_i)$

All models centred over $\epsilon_i \sim \mathsf{N}(0, \sigma^2)$

Centring over Models

Model 1(a):

$$\epsilon_i \sim \mathsf{N}(\mu_i, a\sigma^2), \quad \mu_i \sim F_{x_i},$$

$$F_x \sim \mathsf{DPRS}(M, H, \alpha, \beta), \quad H \sim \mathsf{N}(0, (1-a)\sigma^2), \quad 0 < a < 1$$

a close to zero: nonparametric modelling crucial

Centring over Models

Model 1(a):

$$\epsilon_i \sim \mathsf{N}(\mu_i, a\sigma^2), \quad \mu_i \sim F_{x_i},$$

$$F_x \sim \mathsf{DPRS}(M, H, \alpha, \beta), \quad H \sim \mathsf{N}(0, (1-a)\sigma^2), \quad 0 < a < 1$$

a close to zero: nonparametric modelling crucial

Model 1(b):

$$\epsilon_i - m(x_i) \sim \mathsf{N}(\mu_i, a{\sigma_\star}^2), \quad \mu_i \sim F_{x_i},$$

$$F_x \sim \mathsf{DPRS}(M, H, \alpha, \beta), \quad H \sim \mathsf{N}(0, (1-a){\sigma_\star}^2),$$

with $\sigma^2 = {\sigma_\star}^2 + \sigma_0^2$ and $b = \sigma_0^2/\sigma^2$

b indicates relative importance $m(x)$ (GP)

Centring over Models

Large a, small b: nonparametric modelling less critical, and $g(x)$ is a good parametric model

Interpretation of $g(x)$ nonstandard (given F_x), so consider fixing median

Centring over Models

Large a, small b: nonparametric modelling less critical, and $g(x)$ is a good parametric model

Interpretation of $g(x)$ nonstandard (given F_x), so consider fixing median

Model 2:

$$\epsilon_i - m(x_i) \sim \mathsf{U}\left(-\sigma_\star\sqrt{u_i}, \sigma_\star\sqrt{u_i}\right), \quad u_i \sim F_{x_i},$$

$$F_x \sim \mathsf{DPRS}(M, H, \alpha, \beta),$$

which leads to symmetric error distributions

Choose $H \sim \mathsf{Ga}(3/2, 1/2)$

Computational Issues

- DPRS allows for much simpler MCMC sampling scheme than in Griffin and Steel (2006)
- Adapt Retrospective sampling methods from Papaspiliopoulos and Roberts (2004) (no truncation)

Examples

Example 1: Sine wave

100 observations from $y_i = \sin(2\pi x_i) + \epsilon_i$ with $x_i \sim \mathsf{U}(0, 1)$ and

Examples

Example 1: Sine wave

100 observations from $y_i = \sin(2\pi x_i) + \epsilon_i$ with $x_i \sim \mathsf{U}(0, 1)$ and

- *Error 1*: ϵ_i is t with 2.5 d.f. and conditional variance $\sigma^2(x) = (x - \frac{1}{2})^2$
- *Error 2*:
  $p(\epsilon_i|x_i) = 0.3\mathsf{N}(0.3, 0.01) + 0.7\mathsf{N}(-0.3 + 0.6x_i, 0.01)$

  Bimodal at $x_i = 0$ and unimodal (and normal) at $x_i = 1$

Examples

Example 1: Sine wave

100 observations from $y_i = \sin(2\pi x_i) + \epsilon_i$ with $x_i \sim \mathsf{U}(0, 1)$ and

- *Error 1*: ϵ_i is t with 2.5 d.f. and conditional variance $\sigma^2(x) = (x - \frac{1}{2})^2$
- *Error 2*:
  $p(\epsilon_i|x_i) = 0.3\mathsf{N}(0.3, 0.01) + 0.7\mathsf{N}(-0.3 + 0.6x_i, 0.01)$

  Bimodal at $x_i = 0$ and unimodal (and normal) at $x_i = 1$

Take $g(x) = 0$ throughout

Example: Sine wave, error 1

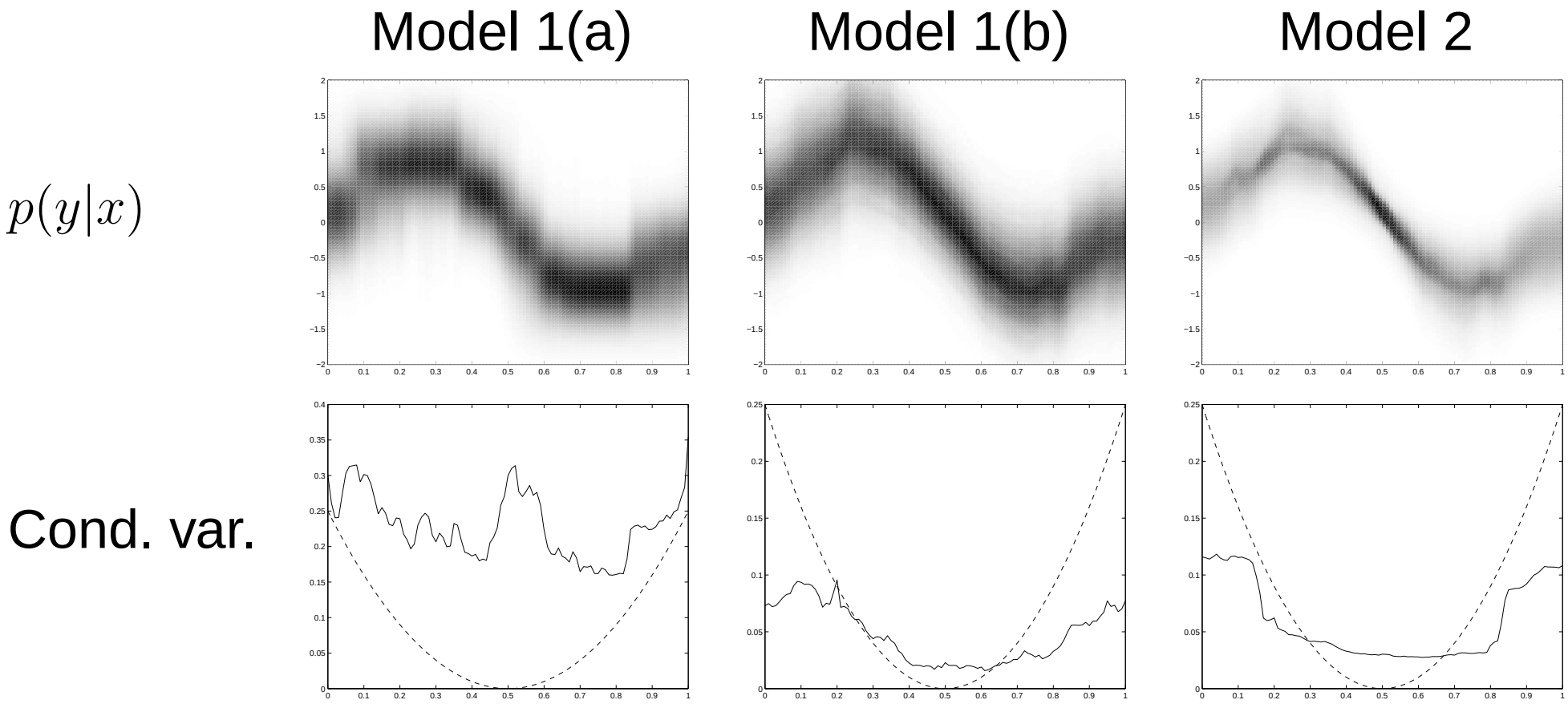


Figure 1: posterior predictive and $\sigma^2(x)$ (true value dashed)

Example: Sine wave

Results with Error 1:

- Small values of a indicate lack of normality
- Model 1(a): “blocky” as expected
- Models with GP regression function do better
- Cond. variance reasonably captured by latter models

Example: Sine wave, error 2

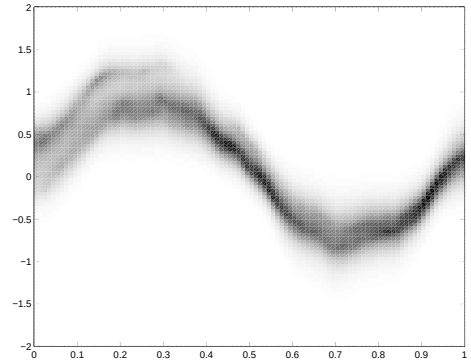
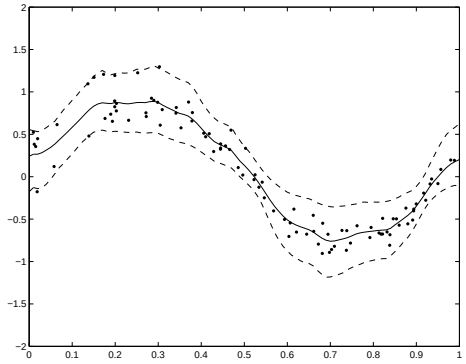
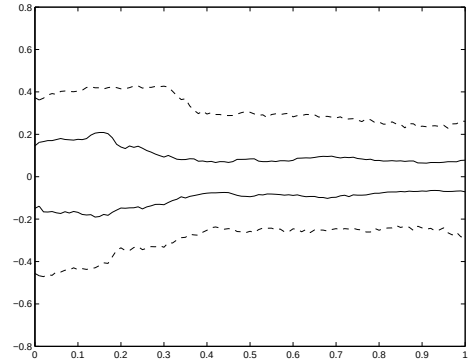
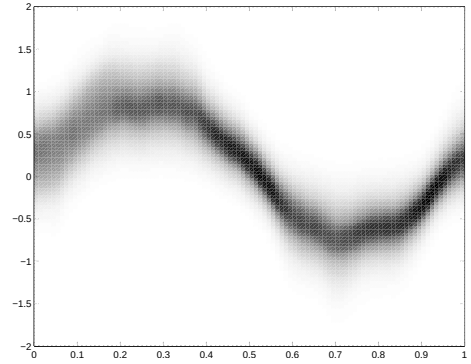
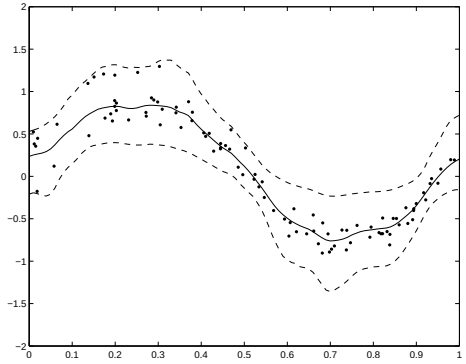
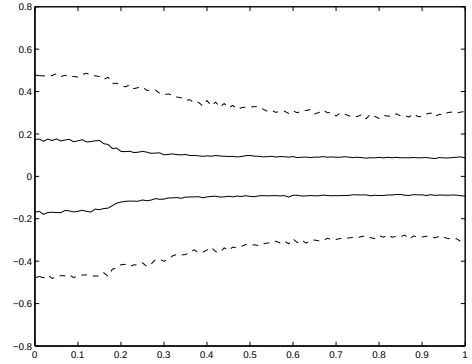
	Predictive $p(y\|x)$	Posterior of $m(x)$	Predictive error
Model 1(b)			
Model 2			

Figure 2: posterior predictive density, $m(x)$ with data (dots), and posterior predictive error distribution

Example: Sine wave

Results with Error 2:

- Model 1(b) can deal with bimodality
- Model 2 can not, by construction
- Small a: nonnormality
- Large b: constant centring model is poor

Examples

Example 2: Scale economies

Cost function for electricity distribution

$$\text{tc} = f(\text{cust}) + \beta_1 \text{wage} + \beta_2 \text{pcap} + \beta_3 \text{PUC} + \beta_4 \text{kwh} + \beta_5 \text{life} + \beta_6 \text{lf} + \beta_7 \text{kmwire} + \epsilon,$$

tc: log total cost per customer
cust: log number of customers

Data: 81 municipal distributors in Ontario, Canada during 1993

Interest: effect of cust and other regressors on tc

Example: Scale economies

DPRS model with cust as the covariate for ϵ and the GP

Centre the model over two parametric regression models by choosing $f(\text{cust})$ to be:

- Parametric 1: $\gamma_1 + \gamma_2 \text{cust}$
- Parametric 2: $\gamma_1 + \gamma_2 \text{cust} + \gamma_3 \text{cust}^2$

Example: Scale economies

Results with Parametric 1:

- Inference on β, γ quite different for parametric and nonparametric models
- Small a for Model 1(a), much larger for Model 1(b)

Example: Scale economies

Results with Parametric 1:

- Inference on β, γ quite different for parametric and nonparametric models
- Small a for Model 1(a), much larger for Model 1(b)

Results with Parametric 2:

- Inference on β, γ similar for parametric and nonparametric models
- Small a for Model 1(a), much larger for Model 1(b)
- Now b smaller than with Parametric 1

Example: Scale economies

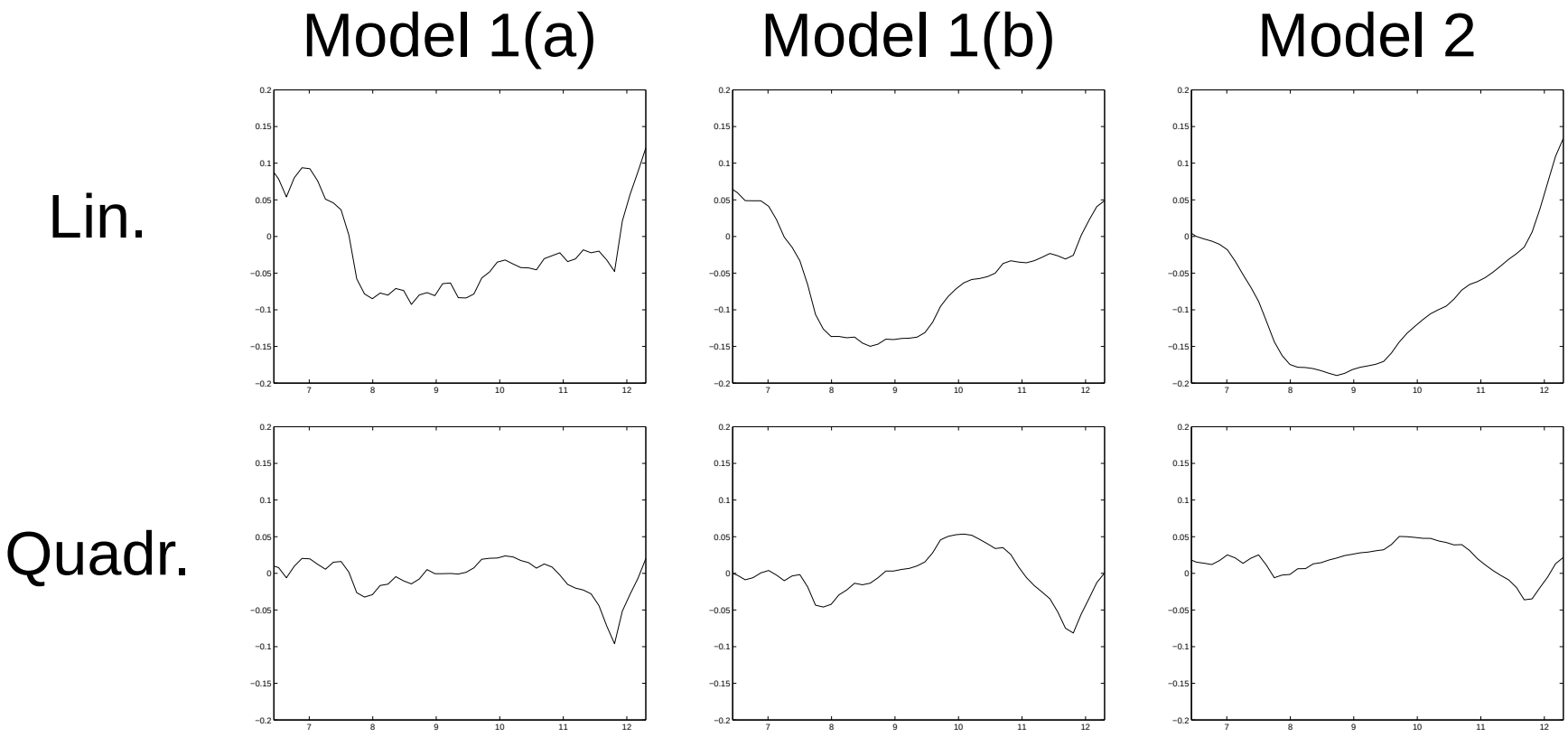


Figure 3: Posterior mean of the nonparametric component(s) of the model

Example: Scale economies

Nonparametric correction to parametric fit: linear model suggests problems; quadratic is better

Conclusion: Quadratic parametric model is not bad; linear is inappropriate

Discussion

- Combine Bayesian NP density estimation and regression modelling

Discussion

- Combine Bayesian NP density estimation and regression modelling
- Separate modelling of components: NP smoother needs to do less

Discussion

- Combine Bayesian NP density estimation and regression modelling
- Separate modelling of components: NP smoother needs to do less
- Centring over parametric models:
  - More structured approach
  - Can identify specific problems of parametric models

Discussion

- Combine Bayesian NP density estimation and regression modelling
- Separate modelling of components: NP smoother needs to do less
- Centring over parametric models:
  - More structured approach
  - Can identify specific problems of parametric models
- Modelling ideas can be used in combination with any NP prior that allows for dependence on covariates