Chapter 20: Basis function models

Kai Tan

Department of Statistics

University of Kentucky

December 2, 2019

Outline

Outline

Model Setting

- $Y = \mu(x) + \epsilon$,
- $\mu(x) = \sum_{h=1}^{H} \beta_h b_h(x)$
- Prespecified basis functions: $b_1(x), b_2(x), \ldots, b_H(x)$.
- Goal: to estimate the basis coefficients: $\beta_1, \beta_2, \ldots, \beta_H$.

Common basis functions I

i) Gaussian radial basis:

$$b_h(x) = \exp\left(-\frac{|x - x_h|^2}{l^2}\right),$$

x_h are centers of the basis functions, l is a common width parameter.

ii) B-Spline: a piecewise continuous function that is defined conditional on some set of *knots*. For example, cubic B-Spline

- is continuous at merging points.
- has continuous first and second derivatives at the merging points.
- The 2nd derivative at end points are 0.

Cubic B-spline

- Assume $x_{h+k} = x_h + \delta k$, the cubic B-spline basis functions is

$$b_h(x) = \begin{cases} \frac{1}{6}u^3 & \text{for } x \in (x_h, x_{h+1}),\ u = \frac{(x-x_h)}{\delta} \\ \frac{1}{6}(1 + 3u + 3u^2 - 3u^3) & \text{for } x \in (x_{h+1}, x_{h+2}), u = \frac{(x-x_{h+1})}{\delta} \\ \frac{1}{6}(4 - 6u^2 + 3u^3) & \text{for } x \in (x_{h+2}, x_{h+3}), u = \frac{(x-x_{h+2})}{\delta} \\ \frac{1}{6}(1 - 3u + 3u^2 - u^3) & \text{for } x \in (x_{h+3}, x_{h+4}), u = \frac{(x-x_{h+3})}{\delta} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

- The width of the basis function is determined by distance δ between knots.
- B-splines are more complicated than Gaussian radial basis function, but each B-spline basis function has compact support, so the design matrix of the linear model is sparse which can be exploited in computation.

Gaussian v.s. B-spline

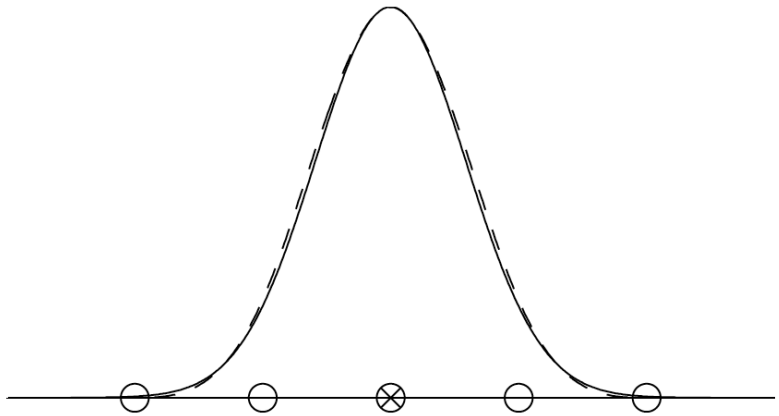

Figure 20.1 *Single Gaussian (solid line) and cubic B-spline (dashed line) basis functions scaled to have the same width. The X marks the center of the Gaussian basis function, and the circles mark the location of knots for the cubic B-spline.*

- Gaussian radial basis is smoother, as they are infinitely differentiable.
- The cubic B-spline is only three times differentiable.

More on B-spline

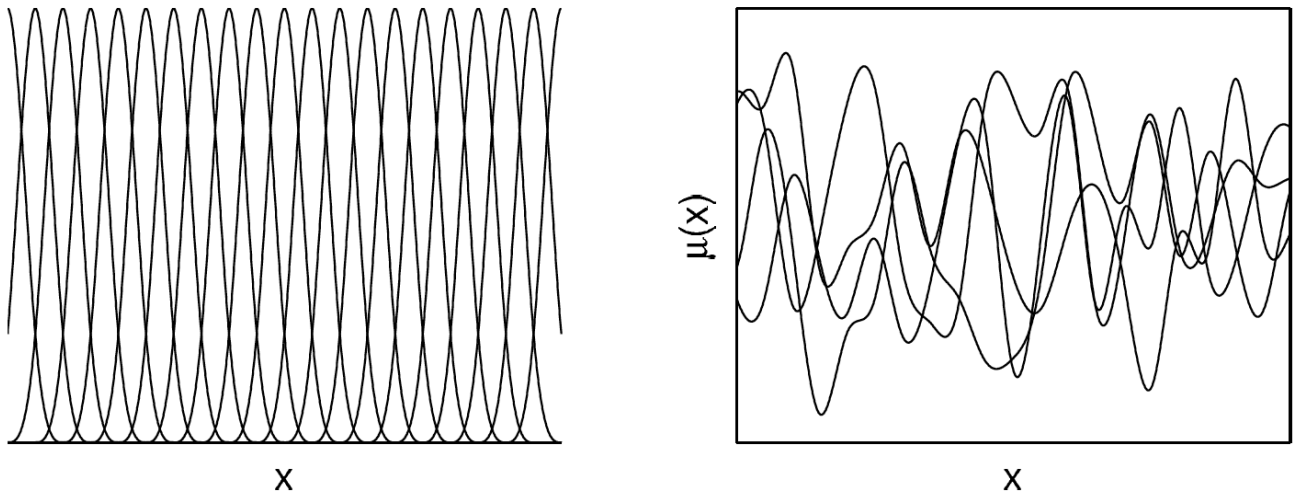

Figure 20.2 *(a) A set of cubic B-splines with equally spaced knots. (b) A set of random draws from the B-spline prior for $\mu(x)$ based on the basis functions in the left graph, assuming independent standard normal priors for the basis coefficients.*

- The number of splines H impacts the flexibility of the resulting model for $\mu(x)$.
- As one cannot characterize finer scale features in $\mu(x)$ than the splines chosen.

Connection with linear model

$$Y = \mu(x) + \epsilon = \sum_{h=1}^{H} \beta_h b_h(x) + \epsilon$$

- Equivalent to linear model when basis functions $b_h(x)$ were selected:

$$y_i = \mu(x_i) + \epsilon_i = w_i\beta + \epsilon_i, \quad \epsilon_i \ iid, \sim N(0, \sigma^2)$$

  where $w_i = (b_1(x_i), b_2(x_i), \ldots, b_H(x_i))$.
- The likelihood $y|x, \beta, \sigma^2 \sim N(w\beta, \sigma^2)$
- Assume multivariate normal-inverse-χ^2 prior for β, σ^2.
- Then the posterior $\beta, \sigma^2|(x, y)$ will also be multivariate normal-inverse-χ^2.
- It is often useful to center the basis function model to linear model $\mu(x) = \beta_1 + \beta_2 x + \sum_{h=3}^{H} \beta_h b_h(x)$.

Example: Chloride concentration I

- A small dataset from a biology experiment containing 54 measurements of the concentration of chloride taken over a short time interval.
- Choose $H = 21$.
- There are 21 coefficients to be estimated $(\beta_1, ..., \beta_{21})$
- But only 54 data points so it becomes problematic to estimate all of the basis coefficients without incorporating prior information.

Example: Chloride concentration II

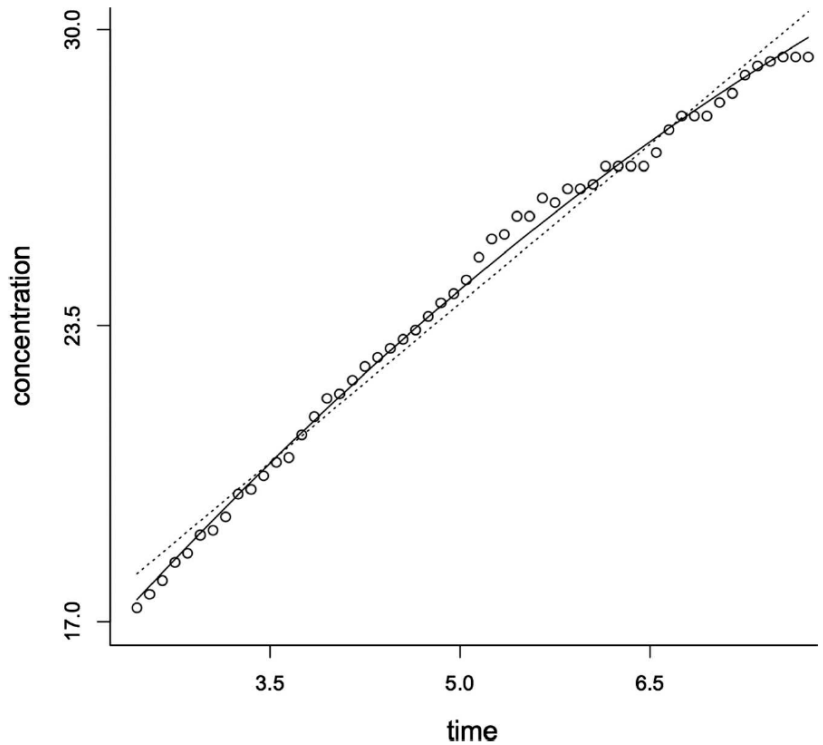

Figure 20.3 *A small dataset of concentration of chloride over time in a biology experiment. Data points are circles, the linear regression estimate is shown with a dotted line, and the posterior mean curve using B-splines is the curved solid line.*

Example: Chloride concentration III

Fig 20.3 shows raw data, a fitted straight line regression, and the posterior mean of the regression function (that is, $E(\mu(x)|y)$) as a function of x, averaging over the posterior distribution of the parameters β) from a fitted B-spline model.

Apply Bayesian model to this data

- $y \sim N(\beta' W, \sigma^2)$.
- Prior $\beta|\sigma \sim N(\beta_0, \sigma^2 \lambda^{-1} I_H)$ and $\sigma^2 \sim \text{Inv-gam}(a_0, b_0)$
- Recall that $\mu(x) = \sum_{h=1}^{H} \beta_h b_h(x)$, thus the prior mean is

$$\mu_0(x) = E(\mu(x)) = \sum_{h=1}^{H} \beta_{0h} b_h(x).$$

- Assume $\mu_0(x) = \alpha + \psi x$, thus it's linear.
- Use LSE to estimate the β_0, such that the resulting $\mu_0(x)$ is as close as possible to $\alpha + \psi x$.
- Plug in the least squares estimates for α and ψ to obtain $\hat{\mu}_0(x)$.
- $\hat{\mu}(x) = (W'W + \lambda I_H)^{-1}(W'y + \lambda \hat{\mu}_0(x))$.

Implementation of Spline methods

- Need to specify the number of knots and their locations.
- Several different Bayesian approaches are available for accommodating uncertainty in basis function specification.
  - Free knot approach
  - Relax the priors from set the $\beta_h = 0$ to shrink $\beta_h \approx 0$.

Outline

Variable selection mixture prior I

Consider the nonparametric regression model

$$y_i \sim N(w_i\beta, \sigma^2),$$

where $w_i = (b_1(x_i), b_2(x_i), \ldots, b_H(x_i))$

- Model index $\gamma = (\gamma_1, \ldots, \gamma_H) \in \Gamma$, with $\gamma_h = 1$ denotes $b_h(x)$ should be included and $\gamma_h = 0$ otherwise.
- Γ is a model space with size 2^H.
- Prior: $\beta_h \sim \pi_h\delta_0 + (1-\pi)N(0, \kappa_h^{-1}\sigma^2)$, $\sigma^2 \sim \text{Inv-gam}(a, b)$. δ_0 is a degenerate distribution with all mass at 0.
- That is, $P(\beta_h = 0) = \pi_h$, and $\beta_h \sim N(0, \kappa_h^{-1}\sigma^2)$ otherwise.
- $P(\gamma_h \neq 0) = 1 - \pi_h$, thus $\gamma_h \sim B(1-\pi_h)$.
  $\beta_\gamma \sim N_{p_\gamma}(0, V_\gamma\sigma^2)$, $p_\gamma = \sum_h \gamma_h$, $V_\gamma = \text{diag}(\kappa_h : \gamma_h = 1)$.
- When there are no prior information on β_h, let $\pi_h = \pi$ and $\pi \sim Beta(a_\pi, b_\pi)$.

Variable selection mixture prior II

- Full conditional posterior distribution for π

$$\pi|- \sim \mathit{Beta}\left(a_\pi + \sum_h (1-\gamma_h), b_\pi + \sum_h (\gamma_h)\right)$$

- Include a heavy-tailed Cauchy prior for nonzero β_h with $\kappa_h \sim_{iid} \mathit{Gamma}(0.5, 0.5)$, $h = 1, \ldots, H$.
- Assuming fixed π and $\kappa_h = \kappa$ for simplicity, the full joint posterior distribution is conjugate with

$$\Pr(\gamma|y,X) = \frac{\pi^{k-p_\gamma}(1-\pi)^{p_\gamma} p(y|X,\gamma)}{\sum_{\gamma^* \in \Gamma} \pi^{k-p_{\gamma^*}}(1-\pi)^{p_{\gamma^*}} p\left(y|X,\gamma^*\right)}, \qquad \text{for all } \gamma \in \Gamma$$

where $p(y|X,\gamma)$ is the marginal likelihood of the data under model γ.

Variable selection mixture prior III

- The marginal likelihood of the data under model γ

$$p(y|X,\gamma)$$
$$= \int \prod_{i=1}^{n} \mathrm{N}\left(y_i|w_{i,\gamma}\beta_\gamma, \sigma^2\right) \mathrm{N}\left(\beta_\gamma|0, V_\gamma\sigma^2\right) \mathrm{Inv}\, G\left(\sigma^2|a,b\right) d\beta_\gamma d\sigma^2$$

  with $w_{i,\gamma} = (w_{ih} : \gamma_h = 1)$

- It's normal linear regression model under a jointly conjugate multivariate normal-gamma prior.
- The posterior distribution of $\beta_\gamma, \sigma^2|\gamma$ is multivariate normal-inverse-gamma.

Curse of dimensionality

- When H is large, it's hard to calculate the denominator.
- E.g., when $H = 50$, there are $2^{50} = 1.1 \times 10^{15}$ possible models.

Two possible solutions via approximation:

- MCMC-based stochastic search algorithm (George and McCulloch, 1993, 1997) to identify high posterior probability models in Γ, and model-average across these models.
- Apply Gibbs sampling to update γ_h from its Bernoulli full conditional posterior distribution

$$\Pr\left(\gamma_h = 1 | \gamma_{(-h)}, \pi\right) = \left(1 + \frac{p\left(y|X, \gamma_h = 0, \gamma_{(-h)}\right)}{p\left(y|X, \gamma_h = 1, \gamma_{(-h)}\right)}\right)^{-1}.$$

One cycle of the Gibbs sampler would update γ_h given $\gamma_{-h}$for $h = 1, \ldots, H$.

Example: Chloride concentration (continued)

- If all 21 basis are includes, and with a prior $N(0, I)$ or $N(0, 2^2 I)$ for the basis coefficients β_h, lead to an extremely poor fit.
- Apply Bayesian variable selection to account for uncertainty in the B-spline basis functions that are needed to characterize the curve (Smith and Kohn (1996).
- If assigned each basis function a prior inclusion probability of 0.5, the coefficients for the basis functions that are included were given independent $N(0, 2^2)$ priors, and σ^2 with a Inv-gam$(1, 1)$ prior.
  - The posterior mean for the number of included basis functions is 12.0 with a 95% posterior interval of [8.0, 16.0].
  - The posterior mean of the residual standard deviation is $\hat{\sigma} = 0.27$ with 95% interval [0.23, 0.33], suggesting that the measurement error variance is small.

Potential drawback for Bayesian variable selection

- There may be some sensitivity to the initial choice of basis.
- For example, using $H = 21$ prespecified cubic B-splines conveys some implicit prior information that the curve is quite smooth, and there are not sharp changes and spikes;
- In many applications, this is well justified but when spike functions are expected a priori one may want to use wavelets or another choice of basis.
- Include multiple types of basis functions in the initial collection of potential basis functions, with Bayesian variable selection used to select the subset of basis functions doing the best job at parsimoniously characterizing the curve.
- However, whenever possible prior information should strongly inform basis choice as well as the choice of prior on the coefficients.

Shrinkage Prior

- Allowing basis functions to drop out of the model adaptively by allowing their coefficients to be zero with positive probability is conceptually appealing, but comes with a computational price.
- When the number of models 2^H in Γ is enormous, (1) MCMC algorithms cannot converge in that only a small percentage of the models will be visited even in several hundreds of thousands of iterations. (2) there can be slow mixing due to the one at a time updating of the elements of Y.
- One possible solution (philosophical appeal): to avoid $\beta_h = 0$, but instead use a regularization or shrinkage prior.
- An appropriate prior would have **high density at zero**, corresponding to basis functions that can be effectively excluded as their coefficients are close to zero.

- Most useful shrinkage priors can be expressed as scale mixtures of Gaussians as follows:

$$\beta_h \sim N(0, \sigma_h^2), \quad \sigma_h^2 \sim G,$$

  with G corresponding to a mixture distribution for the variances.
- E.g., if $G = \text{Inv-gamma}(\nu/2, \nu/2)$, then $\beta_h|\nu \sim t_\nu$. When $\nu \to 0$ leads to a common prior for shrinkage of basis coefficient.
- To obtain a proper posterior and accommodate uncertainty, a common approach is to instead choose equal to a small nonzero value, such as $\nu = 10^{-6}$.
- For $\nu > 0$, the posterior mode will not be exactly zero but the posterior for β_h can still be concentrated at zero for unnecessary basis functions as long as the number of degrees of freedom is sufficiently small.

Generalized double Pareto prior I

$$\text{gdP}(\beta|\xi,\alpha) = \frac{1}{2\xi}\left(1+\frac{|\beta|}{\alpha\xi}\right)^{-(\alpha+1)},$$

where $\xi > 0$ is a scale parameter and $\alpha > 0$ is a shape parameter.

- We can sample from the generalized double Pareto by instead drawing $\beta \sim N(0,\sigma^2), \sigma \sim \text{Exp}(\lambda^2/2)$, and $\lambda \sim \textit{Gamma}(\alpha,\eta)$, where $\xi \sim \eta/\alpha$.
- Let $\alpha = \eta = 1$, which leads to Cauchy-like tails.

$$p(\beta|\sigma) = \prod_{h=1}^{H} \frac{\alpha}{2\sigma\eta}\left(1+\frac{|\beta_h|}{\sigma\eta}\right)^{-(\alpha+1)},$$

which is equivalent to $\beta_h \sim N(0,\sigma^2\tau_h)$, with $\tau_h \sim \text{Exp}(\lambda_h^2/2)$, and $\lambda_h \sim \text{Gamma}(\alpha,\eta)$.

Generalized double Pareto prior II

- Placing the prior $p(\sigma) \propto 1/\sigma$ on the error variance, we then obtain a simple block Gibbs sampler having the following conditional posterior distributions:

$$\beta|- \sim \mathrm{N}\left(\left(W^T W + T^{-1}\right)^{-1} W^T y, \sigma^2 \left(W^T W + T^{-1}\right)^{-1}\right)$$

$$\sigma^2|- \sim \mathrm{Inv-gamma}\left((n+k)/2, (y - W\beta)^T (y - X\beta)/2 + \beta^T T^{-1} \beta/2\right)$$

$$\lambda_h|- \sim \mathrm{Gamma}\left(\alpha + 1, |\beta_h|/\sigma + \eta\right)$$

$$\tau_h^{-1}|- \sim \mathrm{Inv\text{-}Gaussian}\left(\mu = \left(\lambda_h \sigma/\beta_h, \rho = \lambda_h^2\right)\right.$$

where $W = (w_1, \ldots, w_n)$ and $T = \mathrm{Diag}(\tau_1, \ldots, \tau_H)$.

- After convergence, one can obtain draws from the posterior distribution for the nonparametric regression curve $\mu(x)$.

Outline

Heteroscedastic model

- To accommodate heavier-tailed residual densities that allow outliers by instead using a scale mixture of normals.
- $$y_i \sim N(\mu(x_i), \psi_i \sigma^2), \psi_i \sim \text{Inv-gam}(\nu/2, \nu/2),$$

  which induces a t_ν distribution for the residual density.
- For low ν, the t density is substantially heavier-tailed than the normal density, thus downweighting the influence of outliers on the posterior distribution of $\mu(x)$ without needing to discard outlying points.

Multivariate regression surfaces

- To accommodate multiple predictors, one must keep in mind the curse of dimensionality.
  - computational methods may not scale well as predictors are added.
  - require to have enormous amounts of data to reliably estimate a multivariate regression surface without parametric assumptions, substantial prior information or some restrictions.
- As p increases for a given sample size n, observations become much more sparsely distributed across the domain of the predictors $\mathcal{X} \in \mathcal{R}^p$ and hence there are typically subregions of $\mathcal{X}$ having few observations.
- The choice of prior for μ is crucial in developing Bayesian approaches for producing accurate interpolations across sparse data regions.

Additivity model

The regression surface $\mu(x)$ is is characterized as a sum of univariate regression

$$\mu(x) = \mu_0 + \sum_{j=1}^{p} \beta_j(x_j), \quad \beta_j(x_j) = \sum_{h=1}^{H_j} \theta_{jh} b_{jh}(x_j)$$

- E.g., b_j my corresponds to B-splines.
- Bayesian variable selection or shrinkage priors can be applied exactly as described above in the $p = 1$ case without complications.
- Additive models can often reduce the curse of dimensionality, and efficiency can be further improved by including prior information (e.g., shape constraints).

Shape constraints

- For example, it may be reasonable to assume a prior that the mean of the response variable $\mu(x)$ is nondecreasing in one or more of the predictors leading to a nondecreasing constraint on certain $\beta_j(x_j)$
- Easy to incorporate within a Bayesian approach by using piecewise linear or monotone splines b_j and then constraining the regression coefficients θ_j to be nonnegative.
- If a sparse model is preferred, impose a prior distribution that is a mixture of a point mass at zero and a truncated normal distribution on the θ_{jh}'s, leading to nondecreasing β_j functions that can be flat across regions of the predictor space.
- More involved shape restrictions, such as unimodality and convexity, can also be incorporated through an appropriate prior.

Example: nondecreasing nonparametric regression function

- Study the impact of DDE (a persistent metabolite of the pesticide DDT) on the risk of premature delivery.
- Data from pregnant women in the U.S. Collaborative Perinatal Project. Out of 2380 pregnancies in the dataset, there were 361 preterm births.
- Response: Probability of preterm birth.
- Predictors: Serum DDE concentration (x), and potentially confounding maternal characteristics (z) including cholesterol and triglyceride levels, age, BMI and smoking status (yes or no).
- Incorporate a **nondecreasing** constraint on the regression function relating level of DDE to the probability of preterm birth in order to improve efficiency in assessing the dose response trend.

semiparametric probit additive model

$$\Pr\left(y_i = 1 | \theta, x_i, z_i\right) = \Phi\left(\alpha_0 + \sum_{l=1}^{5} z_{il}\alpha_l + f\left(x_i\right)\right) = \Phi\left(z_i\alpha + f\left(x_i\right)\right)$$

- y_i is an indicator of preterm birth, x_i is DDE level,
- $z_i = (1, z_{i1}, z_{i2}, \ldots, z_{i5})$ is a vector of the five predictor in the order listed above.
- Φ is the CDF of standard normal distribution.
- The covariate adjustment is parametric.
- $f(x)$ is characterized nonparametrically as a **nondecreasing** but potentially flat curve using splines with a carefully structured prior on the basis coefficients.

$$\Pr\left(y_i = 1|\theta, x_i, z_i\right) = \Phi\left(\alpha_0 + \sum_{l=1}^{5} z_{il}\alpha_l + f\left(x_i\right)\right) = \Phi\left(z_i\alpha + f\left(x_i\right)\right)$$

- Choose $N(0, 10^2)$ priors independently for the α values.
- For $f(x)$, we simply used a piecewise linear function with a dense set of knots and β_j representing the slope within the j-th interval.
- By choosing a prior for the β_j's that does not allow negative values, we enforce the nondecreasing constraint.

$$f(x) = \sum_{j=1}^{H} \beta_j b_j(x)$$

- In order to borrow information across the adjacent intervals, we use a latent threshold prior.
- Defined a first-order normal random walk autoregressive prior for latent slope parameters, $\beta_j^* \sim N(\beta_{j-1}^*, \sigma_\beta^2)$, with σ_β^2 assigned an inverse-gamma hyperprior to allow the data to inform about the level of smoothness.
- Let $\beta_j = 1_{(\beta_j \geq \delta)} \beta_j^*$, with a small positive threshold parameter, which is assigned a gamma hyperprior.
- As δ increases, it becomes more likely to sample $\beta_j = 0$ and the resulting function has more flat regions.

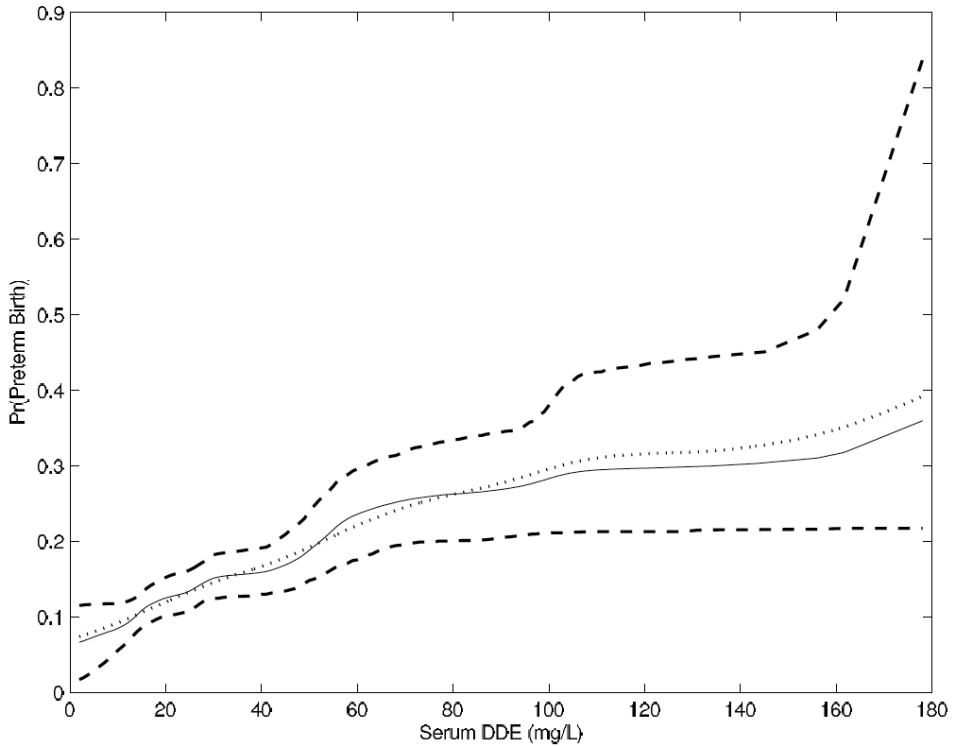

Figure 20.4 *Estimated probability of preterm birth as a function of DDE dose. The solid line is the posterior mean based on a Bayesian nonparametric regression constrained to be nondecreasing, and the dashed lines are 95% posterior intervals for the probability at each point. The dotted line is the maximum likelihood estimate for the unconstrained generalized additive model.*

Tensor product model (Pati and Dunson (2011))

$$\mu(x) = \sum_{h_1=1}^{H} \cdots \sum_{h_p=1}^{H} \left[\theta_{h_1 \cdots h_p} \prod_{j=1}^{p} b_{jh_j}(x_j) \right]$$

- Assume $H_j = H$ for simplicity, and $b_j = \{b_{jh}\}$ is a prespecified set of basis functions for the j-th predictor.
- $\theta = (\theta_{h_1,\ldots,h_p})$ is a p-way array (tensor) containing unknown coefficients.
- The number of coefficients in the tensor θ (H^p) can be large, particularly as p grows.
- Bayesian variable selection or shrinkage priors to favor many elements of θ close to zero.
- Conditionally on the basis functions, it's actually linear function, so that efficient computation is possible using Gibbs sampling.

Thank You!