Gaussian Process Models

Yixuan Zou, Ye Zi

December 11, 2019

Outline

- Introduction
- Gaussian Process Regression
- Example: Birthdays and Birthdates
- Latent Gaussian Process Models
- Functional Data Analysis
- Density Estimation and Regression

Outline

Why Gaussian Process

- Splines and kernel regressions
  - It requires arbitrary set of knots. thus the choice of initial grid may need extra sensitivity analysis
  - One can prespecify a grid of many knots and then use variable selection and shrinkage to discard the useless knots
  - High dimensional grid leads to heavy computational burden, while a low-dimensional grid may not be sufficiently flexible
- **Gaussian process regression**
  - We can set up a prior distribution for the regression function using a flexible class of models for which any finite-dimensional marginal distribution is Gaussian
  - It can be viewed as a potentially infinite-dimensional generalization of Gaussian distribution
  - It has some distinct computational and theoretical advantages

Outline

Gaussian Process Regression

Realizations from a Gaussian process correspond to random functions, and hence the Gaussian process is natural as a prior distribution for an unknown regression function $\mu(x)$

Definition

Denote a Gaussian process as $\mu \sim \mathrm{GP}(m, k)$, where m is a mean function and k is a covariance function. The Gaussian process prior on μ defines it as a random function for which the values at any n prespecified points $x_1, \cdots, x_n$ are a draw from the n-dimensional normal distribution

$$\mu(x_1), \cdots, \mu(x_n) \sim \mathrm{N}((m(x_1), \cdots, m(x_n)), K(x_1, \cdots, x_n)),$$

with mean m and covariance K

Gaussian Process Regression Con't

- The Gaussian process $\mu \sim \mathrm{GP}(m, k)$ is a nonparametric model in that there are infinitely many parameters characterizing the regression function $\mu(x)$
- The mean function $m(x)$ represents an initial guess at the regression function, for example: we can use the linear model $m(x) = X\beta$ with a hyperprior for the regression coefficients β
- The covariance function k specifies the covariance between the process at any two points, the covariance function controls the smoothness of realizations from the Gaussian Process and the degree of shrinkage towards the mean

Gaussian Process Regression Con't

A common choice is the squared exponential covariance function $k(x, x') = \tau^2 \exp\left(- \frac{|x-x'|^2}{l^2} \right)$, where τ controls the magnitude and l controls the smoothness

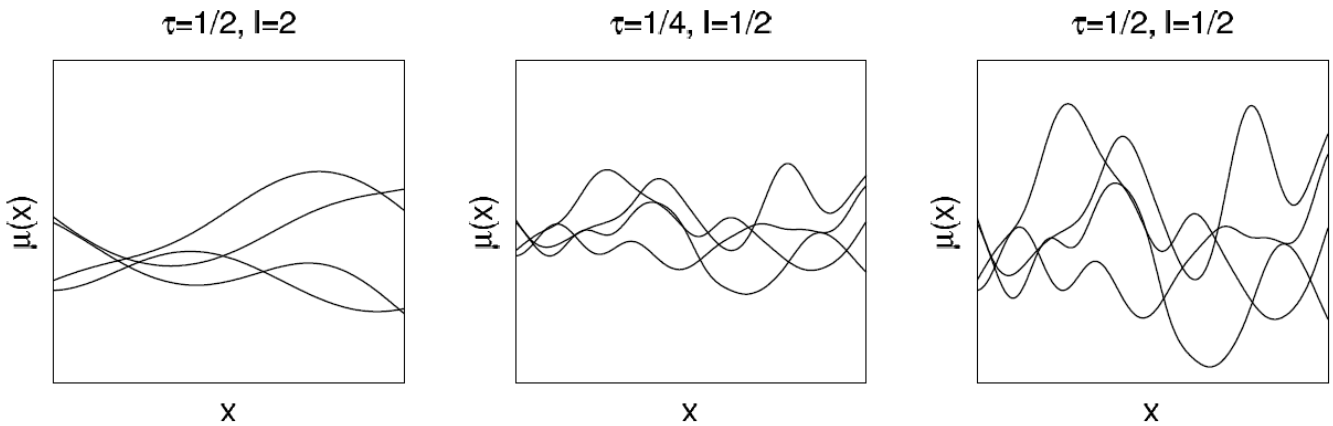

Figure 21.1 *Random draws from the Gaussian process prior with squared exponential covariance function and different values of the amplitude parameter τ and the length scale parameter l.*

Gaussian Process Regression Con't

- Gaussian process priors are appealing in being able to fit a wide range of smooth surfaces while being computationally tractable even for moderate to large numbers of predictors
- We can also introduce Gaussian process prior that depends on the hyperparameters and basis functions. To demonstrate this, let

$$\mu(x) = \sum_{h=1}^{H} \beta_h b_h(x), \quad \beta = (\beta_1, \cdots, \beta_n) \sim \mathrm{N}(\beta_0, \Sigma_\beta),$$

which is a basis function model with a multivariate normal prior on the coefficients. Then,

$$(\mu(x_1), \cdots, \mu(x_n)) \sim \mathrm{N}_n((m(x_1), \cdots, m(x_n)), K(x_1, \cdots, x_n)),$$

with mean and covariance function

$$m(x) = b(x)\beta_0, \quad k(x, x') = b(x)^T \Sigma_\beta b(x')$$

and $b(x) = (b_1(x), \cdots, b_H(x))$

Covariance Functions

- Different covariance functions can be used to add structural prior assumptions like smoothness, nonstationarity, periodicity, and multiscale or hierarchical structures
- When there is more than one predictor and the focus is on a multivariate regression, it is typically not ideal to use a single parameter l to control the smoothness of μ in all directions, the covariance function can be modified as

$$k(x, x') = \text{cov}(\mu(x), \mu(x')) = \tau^2 \text{exp}\left(-\Sigma_{j=1}^{p} \frac{(x_j - x_j')^2}{l_j^2}\right)$$

- We can do nonparametric variable selection by choosing hyperpriors for these l_j's so that predictors that are not needed drop out with large l_j

Inference

- Given a Gaussian observation model, $y_i \sim \mathrm{N}(\mu_i, \sigma^2), i = 1, \cdots, n$, Gaussian process priors are appealing in being conditionally conjugate given I, σ, τ, so that the conditional posterior for μ given $(x_i, y_i)_{i=1}^n$ is again a Gaussian process with updated mean and variance
- In practice, we cannot estimate μ at infinitely many locations and hence the focus is on the realizations at the data points and any additional locations $\tilde{x}$
- Given Gaussian process prior $\mathrm{GP}(0, k)$, the joint density of y and $\tilde{\mu}$ is

$$\mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}K(x,x)+\sigma^2 I & K(\tilde{x},x)\\ K(x,\tilde{x}) & K(\tilde{x},\tilde{x})\end{pmatrix}\right)$$

- The posterior for $\tilde{\mu}$ at a new value $\tilde{x}$ not in the original dataset x is

$$\tilde{\mu}|x, y, \tau, I, \sigma \sim \mathrm{N}(\mathrm{E}(\tilde{\mu}), \mathrm{cov}(\tilde{\mu}))$$
$$\mathrm{E}(\tilde{\mu}) = K(\tilde{x}, x)(K(x,x) + \sigma^2 I)^{-1} y$$
$$\mathrm{cov}(\tilde{\mu})) = K(\tilde{x}, \tilde{x}) - K(\tilde{x}, x)(K(x,x) + \sigma^2 I)^{-1} K(x, \tilde{x})$$

Two Main Hurdles for Posterior Computation

1. Computation of the mean and covariance in the n-variate normal conditional posterior distribution for $\tilde{\mu}$ involves matrix inversion that requires $\mathcal{O}(n^3)$ computation
2. Computation needs to be repeated, for example, at each MCMC step with changing hyperparameters, and hence the computation expense increases so rapidly with n that it becomes challenging to fit Gaussian process regression models

Covariance Function Approximations by Reducing the Matrix Inversion Burden

- Some Gaussian process can be represented as Markov random fields. When there are three or fewer predictors, Markov random fields can be computed efficiently by exploiting conditional independence to produce a sparse precision matrix
- In low-dimensional cases it is also possible to approximate Gaussian processes with basis function approximations where the number of basis functions m is much smaller than n
- The above-mentioned approximations can be used when the number of predictors is large, if the latent function is modeled as additive
- If there are rapid changes in the function, the length scale of the dependency is relatively short. Then sparse covariance matrices can be obtained by using compact support covariance functions
- If the function is smooth, the length scale of the dependency is relatively long. Then reduced rank approximations of the covariance matrices can be obtained in many different ways, reducing the time needed for the inversion to $\mathcal{O}(mn^2)$, where $m \ll n$

Marginal Likelihood and Posterior

- If the data model is Gaussian, the log marginal likelihood is:

$$\log p(y|\tau, I, \sigma^2) = -\frac{n}{2} log(2\pi) - \frac{1}{2}\log|K(x,x) + \sigma^2 I| - \frac{1}{2} y^T (K(x,x) + \sigma^2 I)^{-1} y$$

- The marginal likelihood is combined with the prior to get the unnormalized marginal posterior, and inference can proceed with methods described in Chapters 10–13

Outline

Introduction

Gaussian processes can be directly fit to data, but more generally they can be used as components in a larger model. We illustrate with an analysis of patterns in birthday frequencies in a dataset containing records of all births in the United States on each day during the years 1969–1988

- We originally read about these data being used to uncover a pattern of fewer births on Halloween and excess births on Valentine's Day (due, presumably, to choices involved in scheduled deliveries, along with decisions of whether to induce a birth for health reasons)
- We thought it would be instructive to fit a model to look not just at special days but also at day-of-week effects, patterns during the year, and longer-term trends

Decomposing the Time Series

Based on the structural knowledge of the calendar and, we started with an additive model,

$$y_t(t) = \Sigma_{i=1}^{n} f_i(t) + \epsilon_t,$$

where t is the time in days, and $f_i(t)$ represent variations with different scales and periodicity:

- Long-term trends:

$$f_1(t) \sim \mathrm{GP}(0, k_1), k_1(t, t') = \sigma_1^2 \exp\left(- \frac{|t - t'|^2}{l_1^2} \right)$$

- Short-term trends:

$$f_2(t) \sim \mathrm{GP}(0, k_2), k_2(t, t') = \sigma_2^2 \exp\left(- \frac{|t - t'|^2}{l_2^2} \right)$$

Decomposing the Time Series

- Weekly quasi-periodic pattern:

$$f_3(t) \sim \mathrm{GP}(0, k_3), k_3(t, t') = \sigma_3^2 \exp\Big(- \frac{2\sin^2\big(\frac{\pi(t-t')}{7}\big)}{l_{3,1}^2}\Big)\exp\Big(- \frac{|t-t'|^2}{l_{3,2}^2}\Big)$$

- Yearly smooth seasonal pattern:

$$f_4(t) \sim \mathrm{GP}(0, k_4), k_4(t, t') = \sigma_4^2 \exp\Big(- \frac{2\sin^2\big(\frac{\pi(s-s')}{365.25}\big)}{l_{4,1}^2}\Big)\exp\Big(- \frac{|s-s'|^2}{l_{4,2}^2}\Big)$$

- Special days including an interaction term with weekend:

$$f_5(t) = I_{special\ \ day}(t)\beta_a + I_{weekend}(t) I_{special\ \ day}(t)\beta_b$$

- $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$ represents the unstructured residuals

Some Priors and Data Pre-process

- We set weakly informative log-t priors for the time-scale parameters l (to improve identifiability of the model) and log-uniform priors for all the other hyperparameters
- We normalized the number of daily births y to have mean 0 and standard deviation 1

Results

- We analytically determined the marginal likelihood and its gradients for hyperparameters
- We used the marginal posterior mode for the hyperparameters
- As n was relatively high, this posterior mode was fine in practice

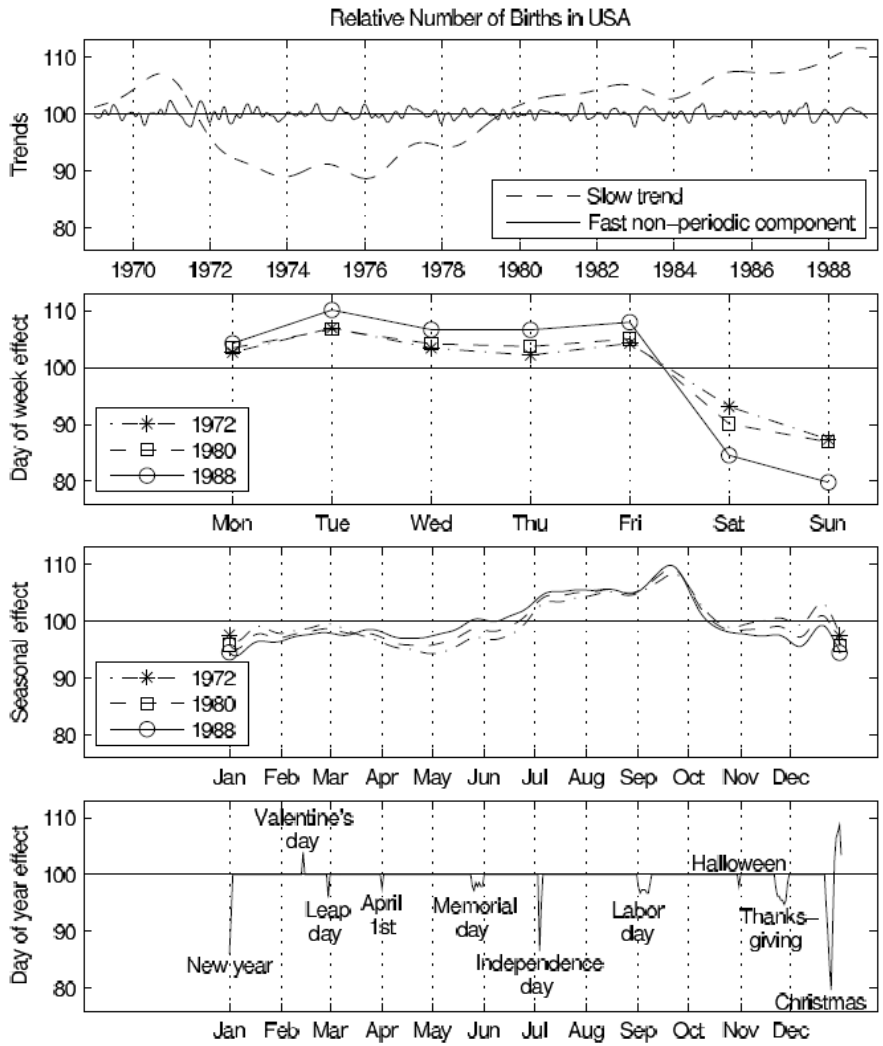

Figure 21.4 *Relative number of births in the United States based on exact data from each day from 1969 through 1988, divided into different components, each with an additive Gaussian process model. The estimates from an improved model are shown in Figure 21.5.*

Outline

Latent Gaussian Process

- **Latent Gaussian Process** is used when likelihood are not Gaussian, the Gaussian process prior is set to a latent function f which through a link function determines the likelihood $p(y|f,\phi)$
- The conditional posterior density of the latent f id $p(f|x,y,\theta,\phi) \propto p(y|f,\phi)p(f|x,\theta)$
- As the prior distribution for latent values is multivariate Gaussian, the posterior disitribution of the latent values is also often close to Gaussian, thus we can use

$$p(f|x,y,\theta,\phi) \approx \mathrm{N}(f|\hat{f},\Sigma),$$

where $\hat{f}$ is the posterior mode and

$$\Sigma^{-1} = K(x,x) + W$$

where $K(x,x)$ is the prior covariance matrix and W is a diagonal matrix with $W_{ii} = \frac{d^2}{df^2}\log p(y|f_i,\phi)|_{f_i=\hat{f}_i}$

Latent Gaussian Process Con't

- The approximate predictive density

$$p(\tilde{y}_i|\tilde{x}_i, x, y, \theta, \phi) \approx \int p(\tilde{y}_i|\tilde{f}_i, \phi)\mathrm{N}(\tilde{f}_i|\tilde{x}_i, x, y, \theta, \phi)d\tilde{f}_i$$

  can be evaluated with quadrature integration
- Log marginal likelihood can be approximated by integrating over f using Laplace's method

$$\log p(y|x, \theta, \phi) \approx \log g(y|x, \theta, \phi) \propto \log p(y|\hat{f}, \phi) - \frac{1}{2}\hat{f}^T K(x, x)\hat{f} - \frac{1}{2}\log|B|,$$

  where $|B| = |I + W^{1/2}K(x, x)W^{1/2}|$
- If the likelihood contribution is heavily skewed, as can be the case with the logistic model, expectation propagation (Section 13.8) can be used instead

Example. Leukemia Survival Times

- To illustrate a Gaussian process model with non-Gaussian data, we analyze survival in acute myeloid leukemia (AML) in adults
- As data we have survival times t and censoring indicator z (0 for observed and 1 for censored) for 1043 cases recorded between 1982 and 1998 in the North West Leukemia Register in the United Kingdom. Some 16% of cases were censored
- Predictors are age, sex, white blood cell count (WBC) at diagnosis with 1 unit $= 50 \times 10^9$/L, and the Townsend score which is a measure of deprivation for district of residence

Model

- We found log-logistic model fit the data the most, the corresponding likelihood is

$$p(y|-) = \prod_{i=1}^{n} \left(\frac{r y^{r-1}}{\exp(f(X_i))}\right)^{1-z_i} \left(1 + \left(\frac{y}{\exp(f(X_i))}\right)^{r}\right)^{z_i - 2},$$

  where r is the shape parameter and z_i is the censoring indicators
- We center the Gaussian process on a linear model to get a latent model

$$f_i(X_i) = \alpha + X_i\beta + \mu(X_i),$$

  where $\mu \sim \mathrm{GP}(0, k)$ with squared exponential covariance function

$$k(x, x') = \sigma_g^2 \exp\left(-\Sigma_{j=1}^{p} \frac{|x_j - x_j'|^2}{l_j}\right)$$

Results

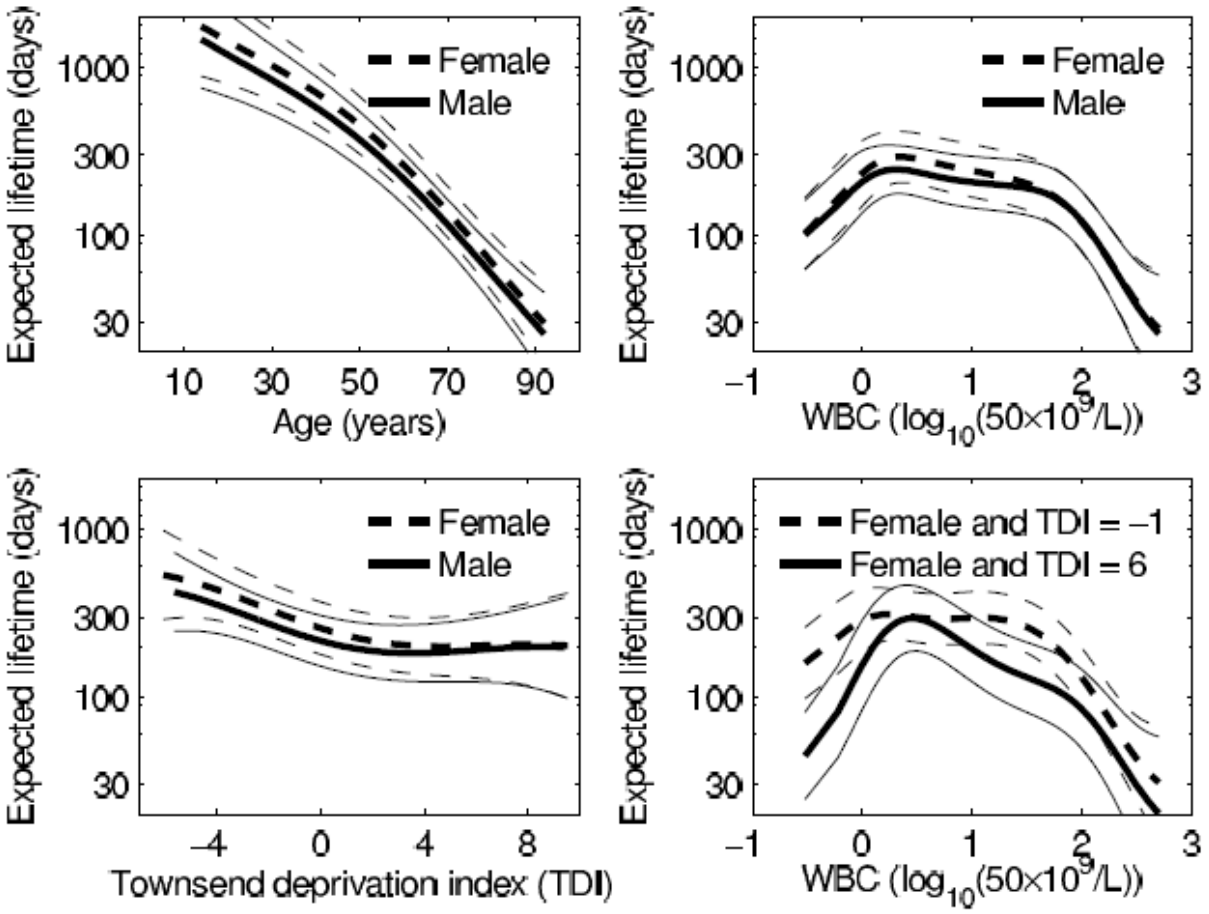

Figure 21.6 *For the leukemia example, estimated conditional comparison for each predictor with other predictors fixed to their mean values or defined values. The thick line in each graph is the posterior median estimated using a Gaussian process model, and the thin lines represent pointwise 90% intervals.*

Outline

Functional Data Analysis

- **Functional data analysis** considers responses and predictors for a subject as random functions defined at infinitely-many points. Let $y_i = (y_{i1}, \ldots, y_{in_i})$ denote the observations on function f_i for subject i, where y_{ij} is an observation at point t_{ij}, with $t_{ij} \in \mathcal{T}$.

$$y_{ij} \sim N(f_i(t_{ij}), \sigma^2).$$

- Gaussian processes can be easily used for functional data analysis. For example, in normal regression we have

$$y_{ij} \sim N(f(x_i, t_{ij}), \sigma^2),$$

where x_i are subject specific predictors. We set $f \sim GP(m, k)$, with squared exponential covariance as

$$\tau^2 \exp\left(-\left[\sum_{j=1}^{p} \frac{(x_j - x_j')^2}{l_j^2} + \frac{(t - t')^2}{l_{p+1}^2}\right]\right).$$

Outline

Density Estimation

- **Logistic Gaussian process** (LGP) generates a random surface from a Gaussian process and then transforming the surface to the space of probability densities.
- Assume y_i iid, and $y_1 \sim p$. Aim: estimate p.
- We can use the continuous logistic transformation, and build the model as

$$p(y|f) = \frac{e^{f(y)}}{\int e^{f(y')}dy'},$$

where $f \sim GP(m, k)$ is a generalization from a continuous Gaussian process.
- choose m to be a log density of elicited parametric distribution
- choose k to be squared exponential:

$$k(y, y') = \tau^2 \exp\left(-\frac{|y - y'|^2}{l^2}\right).$$

Density Estimation

- An alternative specification uses a zero-mean Gaussian process $W(t)$ on $[0, 1]$ an defines

$$p(y) = g_0(y) \frac{e^{W(G_0(y))}}{\int e^{W(\nu)} d\nu}$$

  where g_0 is some elicited parametric distribution with cumulative distribution function G_0.
- Integral in the denominator: computed using a finite basis function representation or a discretization of a chosen finite region.
- Inference for f and parameters of m and k: various Markov chain simulation methods, or using a combination of Laplace's method for the latent values f and quadrature integration for parameters.

Density Estimation Con't

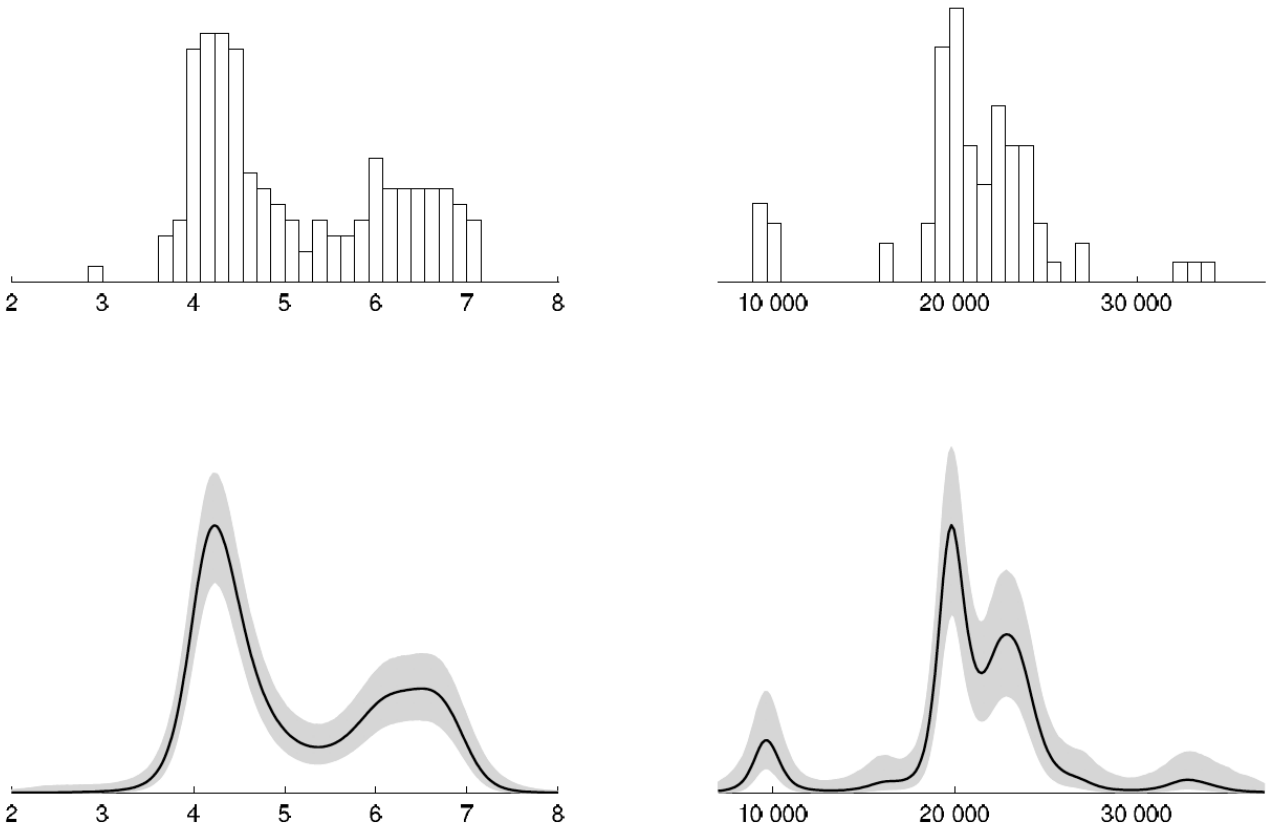

Figure 21.7 *Two simple examples of density estimation using Gaussian processes. Left column shows acidity data and right column shows galaxy data. Top row shows histograms and bottom row shows logistic Gaussian process density estimate means and 90% pointwise posterior intervals.*

Density Regression

- LGP prior can be easily generalized to density regression by setting $p_{\mathcal{X}} = \{p(y|x), x \in \mathcal{R}, y \in \mathcal{R}\}$ as

$$p(y|x) = \frac{e^{f(x,y)}}{\int e^{f(x,y')dy'}},$$

where f is drawn from a Gaussian process with

$$k((x,y),(x',y')) = \tau^2 \exp\left(-\left[\sum_{j=1}^{p} \frac{(x_j - x_j')^2}{l_j} + \frac{(y-y')^2}{l_{p+1}}\right]\right)$$

- Letting $s = (s_1, \ldots, s_p) \in [-1,1]^p$ and $t \in [0,1]$ and prespecifying monotone continuous functions $F_j : \mathcal{R} \to [-1,1]$, for $j = 1, \ldots, p$,

$$p(y|x) = g_0(y) \frac{e^{W(F(x),G_0(y))}}{\int e^{W(F(x),\nu)} d\nu},$$

where $F(x) = (F_1(x_1), \ldots, F_p(x_p))$ and W is drawn from a Gaussian process.

Latent Variable Regression

- Latent-variable regression model

$$y_i \sim N(\mu(u_i), \sigma^2), \ u_i \sim U(0,1),$$

where u_i is a uniform latent variable and $\mu : [0,1] \to [\mathcal{R}]$ is an unknown regression function.
- drawing μ from a Gaussian process centered on μ_0 with a squared exponential covariance kernel.

- Generalize for the density regression problem by

$$y_i \sim N(\mu(u_i, x_i), \sigma^2), \ u_i \sim U(0,1),$$

where $x_i = (x_{i1}, \ldots, x_{ip})$ is the vector of observed predictors.
- μ is a $(p+1)$-dimensional surface drawn from a Gaussian process and the covariance function can be chosen to be squared exponential with a different spatial-range parameter for each dimension.