Graphical Models for Inference and Decision Making

Instructor: Kathryn Blackmond Laskey

Room 2214 ENGR

(703) 993-1644

Office Hours: Tuesday and Thursday 4:30-5:30 PM, or by appointment

Spring 2019

Nonparametric Bayesian Models

Learning Objectives

1. Define a statistical model
2. Explain the difference between parametric and nonparametric statistical models
3. Be familiar with Dirichlet process mixtures and their use in clustering
   a) Explain how the Dirichlet process generalizes the Dirichlet distribution
   b) Express a Dirichlet process mixture model as a plate model
   c) Explain the Gibbs sampling algorithm for inference in a Dirichlet process mixture model
4. Be familiar with Gaussian processes and their use in regression
   a) Define a Gaussian process in terms of a mean and covariance function
   b) Describe how a Gaussian process regression model can be expressed as a linear regression model with a set of basis functions
   c) Describe the kernel trick and its use in machine learning
   d) Explain how inference works in Gaussian process regression

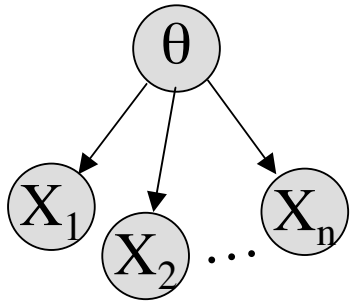
Outline

- Parametric and nonparametric statistical models
- Dirichlet process mixtures and nonparametric clustering
- Gaussian processes and nonparametric regression

Statistical Models

- A *statistical model* on a sample space $\mathbf{X}$ is a set of probability distributions on $\mathbf{X}$
  - Models are indexed by *parameter* θ, i.e., $M = \{P_\theta \mid \theta \in \mathbf{T}\}$
- The objective of inference is to use observations from P_θ to draw inferences about θ, and hence about the data-generating process P_θ
- To draw inferences about θ, a Bayesian specifies a *prior distribution* for θ and uses Bayes rule to find the *posterior distribution*
  - For example, if observations $X_1, \ldots, X_n$ are iid draws from $f(x \mid \theta)$ and prior distribution is $g(\theta)$, then the posterior distribution is

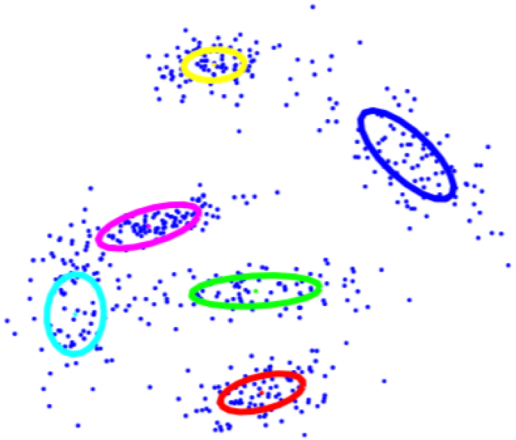
$$g(\theta \mid x_1, \ldots x_n) = \frac{f(\underline{x} \mid \theta) g(\theta)}{f(\underline{x})} = \frac{\prod_{i=1}^{n} f(x_i \mid \theta) g(\theta)}{\int \prod_{i=1}^{n} f(x_i \mid \theta) g(\theta) d\mu(\theta)}$$



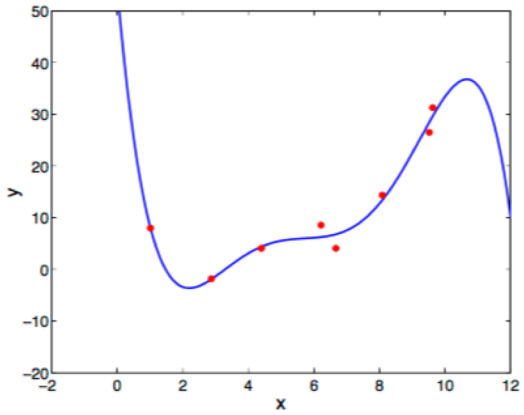
- A model is called *parametric* if the parameter θ has finite dimension, and *nonparametric* if the parameter θ has infinite dimension
  - To do Bayesian inference with nonparametric models, we need to define a prior distribution on an infinite-dimensional space and develop methods to find a posterior distribution

Why Nonparametrics?

- Statisticians often fit models of varying complexity and then choose the one that best fits the observations



e.g. selecting m, the number of Gaussians in a mixture model



e.g. selecting m the order of a polynomial in a nonlinear regression model

Figure taken from http://mlg.eng.cam.ac.uk/zoubin/talks/uai05tutorial-b.pdf

- Nonparametric models adapt their complexity according to the data
- Nonparametric models can grow in complexity as the sample size increases

Examples of Nonparametric Models

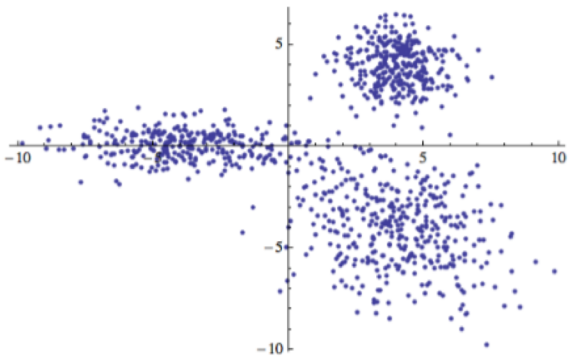
- We will consider two important classes of nonparametric Bayesian methods:
  - Dirichlet process mixtures for clustering
  - Gaussian process for nonlinear regression
- These methods are important in their own right and illustrate important concepts in Bayesian nonparametrics
  - Both methods are infinite-dimensional generalizations of commonly applied parametric models
  - Both methods exploit conjugacy for efficient inference

Outline

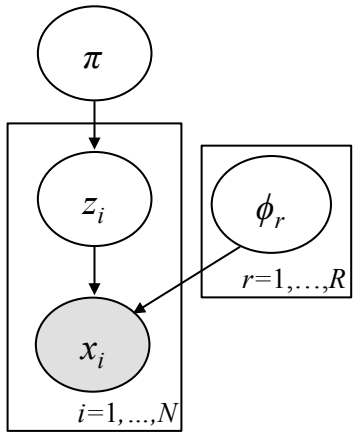
- Parametric and nonparametric statistical models
- Dirichlet process mixtures and nonparametric clustering
- Gaussian processes and nonparametric regression

Clustering

- Clustering seeks to sub-divide a set of observations into subsets called clusters
  - Each observation belongs to exactly one cluster
  - Observations in the same cluster are more similar to each other than to observations in different clusters
- Typically, we use a mixture distribution as a model for clustering
  - Latent categorical RV Z_i indicates the cluster to which observation i belongs
  - Observations in same cluster are iid
- A nonparametric model does not fix the number of clusters
  - Adapts to the observations
  - Can grow as sample size grows



Plot taken from Orbanz (2014)



$$\pi \sim Dirichlet(\alpha_1,\ldots\alpha_R)$$
$$\phi_r \sim h(\phi)$$
$$z_i \mid \pi \sim Multinomial(1,\pi)$$
$$x_i \mid z_i,\phi_{z_i} \sim f(x \mid \phi_{z_i})$$

A finite mixture model with R clusters

The Multinomial-Dirichlet Conjugate Family

- The Multinomial(n, $\underline{\pi}$) distribution is a multivariate generalization of the Binomial distribution
  - Used to model observations that fall into one of a finite number R of mutually exclusive categories
  - Parameters: total count n and probabilities $\pi_1, \ldots, \pi_R$ (where $\sum_i \pi_i = 1$)
    - » π_i is the probability that an observation falls into category i, for $i=1, \ldots, R$
  - Observation $\underline{z} = (z_1, \ldots, z_R)$ is a vector of counts of how many observations fall into each category, where $\sum_i Z_i = n$ is the total count
  - Likelihood function:

$$f(z_1,\ldots,z_R \mid \pi_1,\ldots,\pi_R) = \left(\frac{n!}{z_1!\cdots z_R!}\right)\pi_1^{z_1}\cdots\pi_R^{z_R}$$

- The conjugate prior distribution for $(\pi_1, \ldots, \pi_R)$ is the Dirichlet distribution, a multivariate generalization of the Beta distribution
  - Hyperparameters $\alpha_1, \ldots, \alpha_R$ (called *virtual counts* or *pseudo-counts*)
  - Density function (restricted to $\pi_i \geq 0$ and $\sum_i \pi_i = 1$)

$$g(\pi_1,\ldots,\pi_R \mid \alpha_1,\ldots,\alpha_R) = \frac{\Gamma(\alpha_1+\ldots+\alpha_R)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_R)}\pi_1^{\alpha_1-1}\cdots\pi_R^{\alpha_R-1}$$

- The posterior distribution for $\pi_1, \ldots, \pi_R$ given $z_1, \ldots, z_R$ is Dirichlet with parameters $\alpha_1+z_1, \ldots, \alpha_R+z_R$

Some Facts About the Dirichlet Distribution

- If a k-dimensional random variable $(\pi_1, \ldots, \pi_k)$ has a Dirichlet$(\alpha_1, \ldots, \alpha_k)$ distribution then:
  - π_i has a Beta$(\alpha_i, \Sigma_{j \neq I}\, \alpha_j)$ distribution
  - $E[\pi_i] = \dfrac{\alpha_i}{\alpha_1 + \cdots + \alpha_R}$
  - $V[\pi_i] = \dfrac{\alpha_i\left(\sum_{j \neq i} \alpha_j\right)}{(\alpha_1 + \cdots + \alpha_R)^2(\alpha_1 + \cdots + \alpha_R + 1)}$ $\qquad \mathrm{Cov}[\pi_i, \pi_j] = \dfrac{-\alpha_i \alpha_j}{(\alpha_1 + \cdots + \alpha_R)^2(\alpha_1 + \cdots + \alpha_R + 1)}$
  - Dirichlet(1,…,1) is called the *uniform* distribution and puts equal probability density on all $(\pi_1, \ldots, \pi_R)$ such that $\pi_i \geq 0$ and $\Sigma_i\, \pi_i = 1$
  - We call α_i the *virtual count* or *pseudo-count* for the i^{th} category
- Marginal likelihood for vector of counts $\underline{Z}=(Z_1,\ldots,Z_R)$

$$f(z_1,\ldots,z_R \mid \alpha_1,\ldots,\alpha_R, n) = \begin{cases} \left(\dfrac{n!}{z_1!\cdots z_R!}\right)\dfrac{\Gamma(\alpha_1+\cdots+\alpha_R)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_R)}\dfrac{\Gamma(\alpha_1+z_1)\cdots\Gamma(\alpha_k+z_R)}{\Gamma(\alpha_1+\cdots+\alpha_R+n)} & \text{if } \sum_i z_i = n \\ 0 & \text{otherwise} \end{cases}$$

$$\Gamma(y) \equiv \int_0^\infty u^{y-1} e^{-u}\, du = (y-1)\Gamma(y-1)$$

Examples of Dirichlet distributions over $\mathbf{p} = (p_1, p_2, p_3)$ which can be plotted in 2D since $p_3 = 1 - p_1 - p_2$:

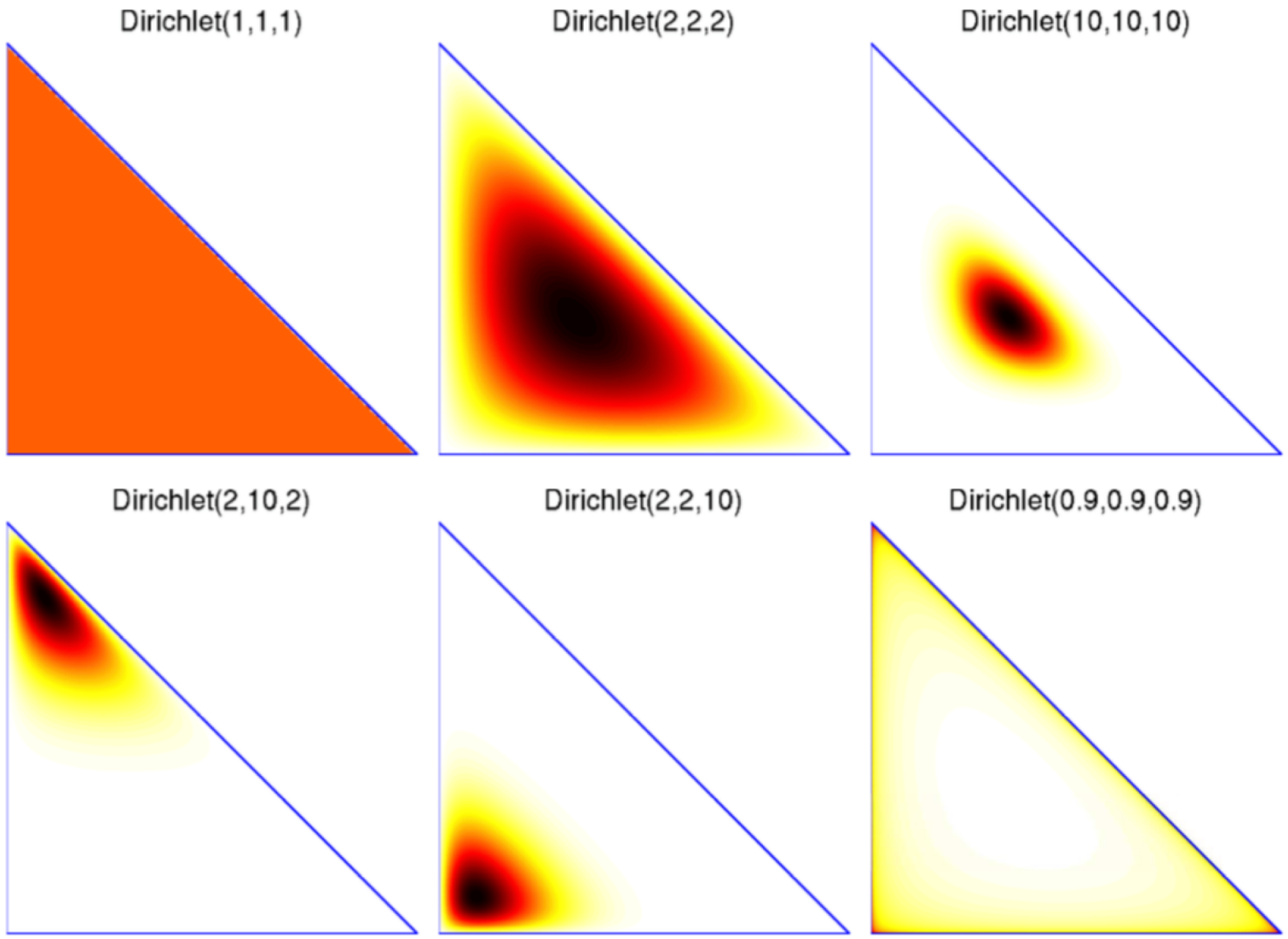
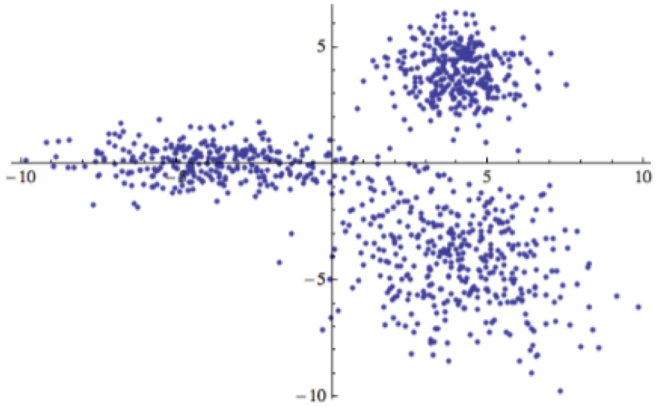


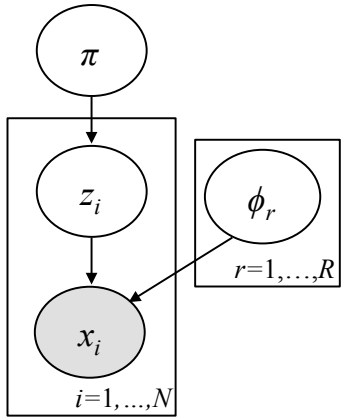
Figure taken from http://mlg.eng.cam.ac.uk/zoubin/talks/uai05tutorial-b.pdf

Dirichlet Distribution and Clustering

- Typically, we use a mixture distribution as a model for clustering
  - Latent cluster assignments z_i are iid draws from a multinomial distribution with parameter $(n, \underline{\pi})$
  - Parameter $\underline{\pi}$ is drawn from Dirichlet conjugate prior
  - Observations $i = 1, \ldots, N$ are iid draws from parametric model $f(x \mid \phi_{z_i})$ with cluster-specific parameter ϕ_{z_i}
  - Choose likelihood $f(x \mid \phi_{z_i})$ from a family with a conjugate prior
  - Cluster parameters ϕ_r, $r=1, \ldots, R$ are iid draws from conjugate prior $g(\phi_r \mid \xi)$
- Although exact inference is intractable, inference can be performed by Gibbs sampling or variational inference



Plot taken from Orbanz (2014)



$$\pi \sim Dirichlet(\alpha_1,\ldots\alpha_R)$$
$$\phi_r \sim g(\phi \mid \xi)$$
$$z_i \mid \pi \sim Multinomial(1,\pi)$$
$$x_i \mid z_i, \phi_{z_i} \sim f(x \mid \phi_{z_i})$$

A finite mixture model with R clusters

Dirichlet Process

- The Dirichlet Process (DP) is an infinite dimensional generalization of the Dirichlet distribution
- DP defines a distribution over (infinite discrete) probability distributions

$$G \sim \mathrm{DP}(\bullet \mid H, \alpha)$$

H is the *base measure* (may be continuous or discrete)

$\alpha > 0$ is the *concentration parameter*

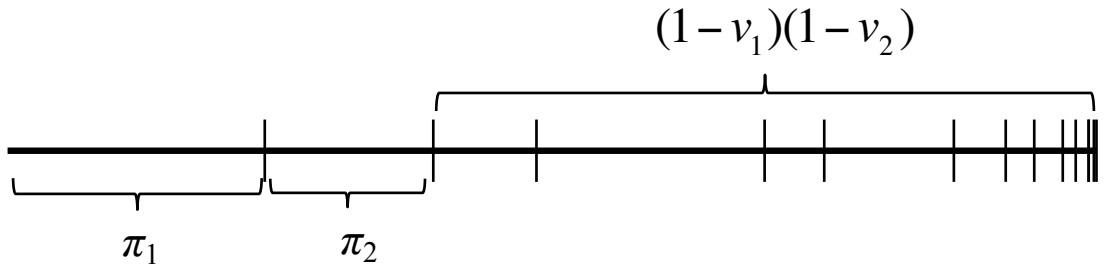
- DP is defined according to the following condition on the finite-dimensional marginals:
  - If $G \sim \mathrm{DP}(\bullet \mid H, \alpha)$ is a Dirichlet process on sample space Φ, then for any finite partition $\{B_1, \ldots, B_k\}$ of Φ (i.e., $\bigcup B_i = \Phi$ and $B_i \cap B_j = \varnothing$ for $i \neq j$), $(G(B_1), \ldots, G(B_k)) \sim \mathrm{Dirichlet}(\alpha H(B_1), \ldots, \alpha H(B_k))$
- Dirichlet process prior distribution can be used to define a natural extension of a finite mixture model to an infinite-dimensional mixture

Stick-Breaking Generative Process for DP

- The following sampling procedure draws a distribution $G \sim \text{DP}(\bullet \mid H, \alpha)$ from a Dirichlet process with base measure H and concentration α

  1. $v_1, v_2, \ldots \sim_{\text{iid}} \text{Beta}(1, \alpha)$
  2. $\pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j)$ [note that $\pi_i > 0$ and $\sum_{i=1}^{\infty} \pi_i = 1$]
  3. $\phi_1, \phi_2, \ldots \sim_{\text{iid}} H(\bullet)$
  4. G is the distribution that samples ϕ_i with probability π_i for i = 1, 2, ...

- This sampling procedure is called a *stick breaking* process



- In a clustering problem, we typically choose the base measure H to draw a parameter ϕ from family of distributions that is conjugate to the data likelihood $f(x \mid \phi)$:

$$\phi \sim g(\phi \mid \xi) \text{ where } g(\phi \mid \xi) \text{ is a conjugate family to } f(x \mid \phi)$$

Reminder: Conjugate Families of Distributions

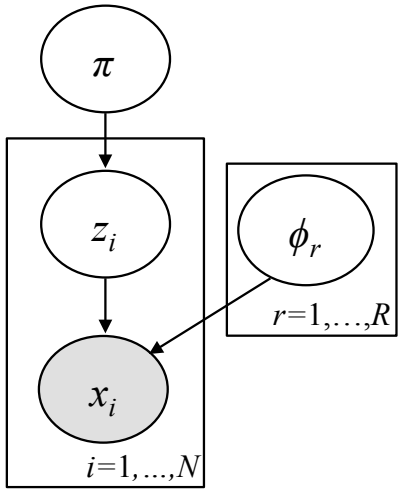
- A family $g(\theta|\alpha)$ of distributions parameterized by α is a *conjugate family* for the family of distributions $f(x|\theta)$ if it is *closed under sampling* from $f(x|\theta)$, that is:

  IF Data $X_1, \ldots, X_n$ are a random sample from $f(x|\theta)$ AND prior distribution for unknown parameter Θ is $g(\theta|\alpha)$

  THEN Posterior distribution for parameter Θ is $g(\theta|\alpha^*)$, another member of the conjugate family

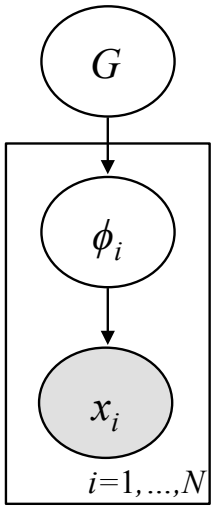
- We call α a *hyperparameter*
- Examples of conjugate families:
  - Poisson likelihood, gamma prior distribution on rate
  - Exponential likelihood, inverse gamma distribution on mean
  - Binomial likelihood, beta distribution on probability
  - Multinomial likelihood, Dirichlet distribution on probabilities
  - Normal distribution, normal-gamma distribution on precision (inverse of variance)

Comparing Finite and Infinite Mixture Plate Models



A finite mixture model with R clusters

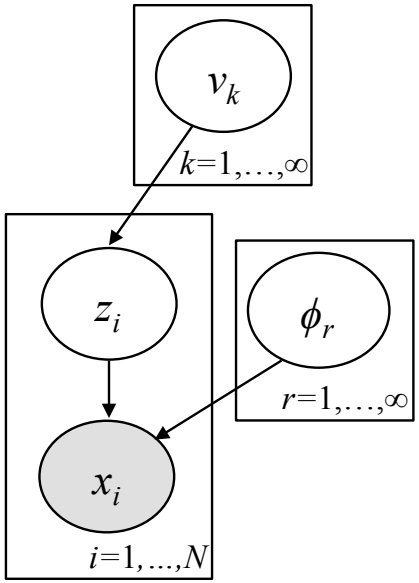
$$\begin{aligned}
\pi &\sim Dirichlet(\alpha_1,\ldots\alpha_R)\\
\phi_r &\sim g(\phi\,|\,\xi)\\
z_i\,|\,\pi &\sim Multinomial(1,\pi)\\
X_i\,|\,z_i,\phi_{z_i} &\sim f(x\,|\,\phi_{z_i})
\end{aligned}$$



A Dirichlet process mixture model

$$\begin{aligned}
G &\sim DP(H,\alpha)\\
\phi_i &\sim G\\
X_i\,|\,\phi_i &\sim f(x\,|\,\phi_{Z_i})
\end{aligned}$$

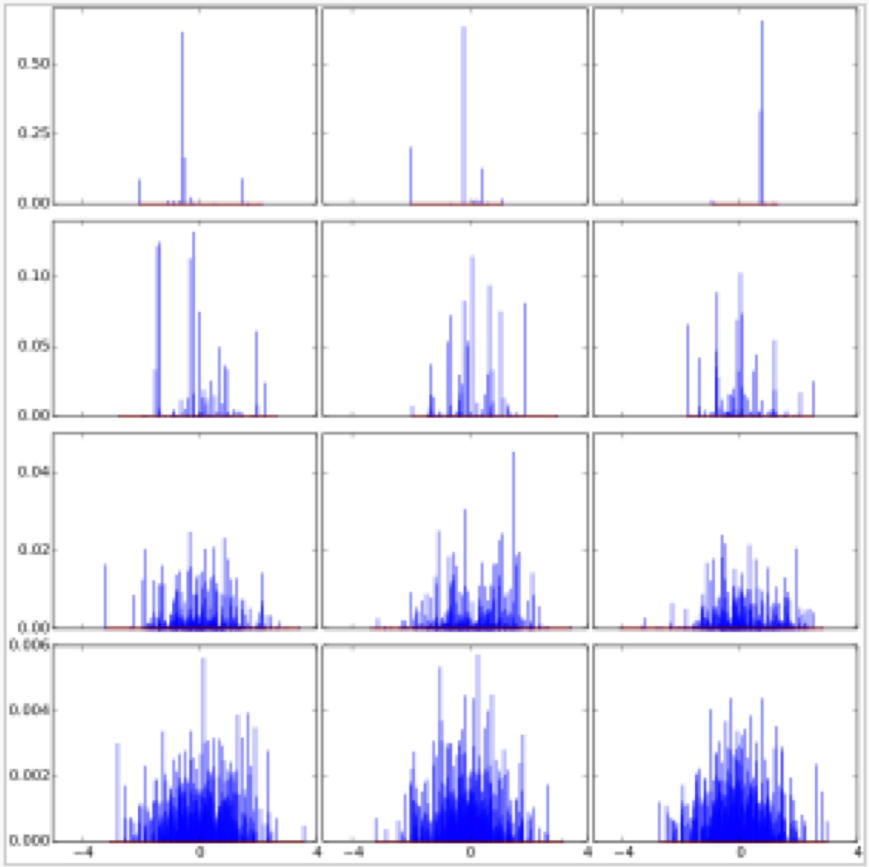
This representation leaves clusters implicit. Observations x_i and x_j belong to the same cluster if they were generated by the same parameter, i.e., $\phi_i=\phi_j$



A Dirichlet process mixture model sampled by stick-breaking

$$\begin{aligned}
v_k &\sim Beta(1,\alpha)\\
\phi_r &\sim g(\phi\,|\,\xi)\\
\pi_i &= v_i\prod_{j=1}^{i-1}(1-v_j)\\
z_i\,|\,v_{1:i} &\sim Multinomial(1,\pi)\\
x_i\,|\,z_i,\phi_{z_i} &\sim f(x\,|\,\phi_{z_i})
\end{aligned}$$

Examples of DP Draws



Draws from the Dirichlet process $\mathrm{DP}\left(N(0,1),\alpha\right)$. The four rows use different α (top to bottom: 1, 10, 100 and 1000) and each row contains 3 repetitions of the same experiment. As seen from the graphs, draws from a Dirichlet process are discrete distributions and they become less concentrated (more spread out) with increasing α. The graphs were generated using the stick-breaking process view of the Dirichlet process.

Source: https://en.wikipedia.org/wiki/Dirichlet_process

Another View of the Dirichlet Process

- If $\Phi_i \sim$ iid G and $G \sim \mathrm{DP}(\bullet \mid H, \alpha)$, then a generative model for the cluster parameters $\phi_1, \phi_2, \ldots$ (marginal over G) is:
  1. Draw ϕ_1 from H
  2. For $n > 1$,
     a. with probability $\alpha/(\alpha+n\text{-}1)$ draw ϕ_n from H
     b. with probability $(n\text{-}1)/(\alpha+n\text{-}1)$, draw ϕ_n uniformly from $\{\phi_1, \phi_2, \ldots, \phi_{n\text{-}1}\}$
- The probability of assigning a new observation to an existing component is proportional to the number of instances already assigned to it ("rich get richer")
- The probability that a new cluster is created depends on the concentration parameter and the number of observations already assigned to clusters
  - The larger the concentration parameter, the more likely a new cluster is created
  - The larger the number of observations already assigned to clusters, the less likely a new cluster will be created for the next observation
- Metaphor: Chinese Restaurant process
  - Customers arrive at a Chinese restaurant with infinitely many tables
  - Each customer sits at an occupied table with probability $(n\text{-}1)/(\alpha+n\text{-}1)$ or a vacant table with probability $\alpha/(\alpha+n\text{-}1)$
  - Popular tables attract more customers
  - Likelihood that a new customer will sit at an unoccupied table decreases as number of already-seated customers grows

Conjugate Updating with Dirichlet Process

- A Dirichlet process distribution $\mathrm{DP}(\bullet \mid H, \alpha)$ on sample space $\mathbf{\Phi}$ is a conjugate prior distribution for infinite discrete distributions on $\mathbf{\Phi}$

  IF Observations $\phi_1, \phi_2, \ldots, \phi_n$ are a random sample from the (discrete) distribution $G(\bullet)$, AND the prior distribution for G is $\mathrm{DP}(\bullet \mid H, \alpha)$

  THEN the posterior distribution for G conditional on $\Phi_1, \Phi_2, \ldots$ is $\mathrm{DP}(\bullet \mid H^*, \alpha^*)$, another member of the conjugate family

- The posterior hyperparameters are:

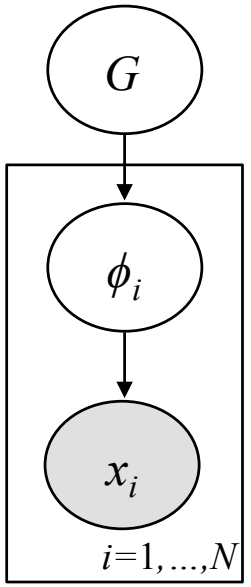
$$\alpha^* = \alpha + n$$

$$H^* = \frac{1}{\alpha + n}\left(\alpha H + \sum_{i=1}^{n} \delta_{\phi_i}\right)$$

The Dirac measure δ_ϕ assigns probability 1 to the value ϕ

Inference with DP Mixtures: Gibbs Sampling

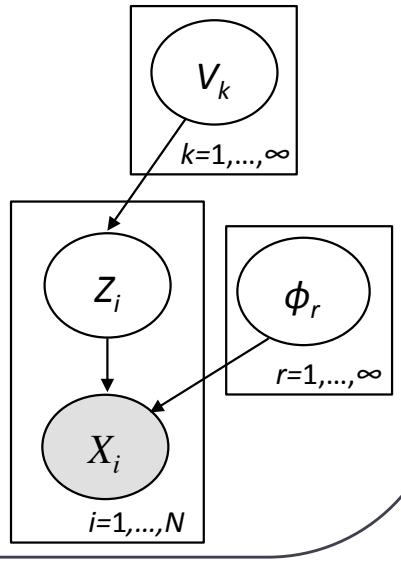
1. Initialize number of clusters $R < n$ and cluster parameters $\phi_1, \phi_2, \ldots, \phi_R$
2. Initialize cluster memberships $1 \le z_1, z_2, \ldots, z_n \le R$
3. At each iteration k do the following:
   - Let K be the current number of clusters that have observations assigned, and define a 0^{th} cluster that has no current observations
   - For $i = 1, ..., n$:
     a. Find probabilty that the i^{th} observation is in each cluster $0, 1, ..., r$, conditional on parameters $\phi_1, \phi_2, \ldots, \phi_R$ and the cluster assignments of observations other than i
     b. Sample new value for i^{th} cluster membership $0 \le z_i \le R$ using these probabilities
   - For $r = 0, \ldots R$:
     a. For each occupied cluster r, find the posterior distribution of ϕ_r conditional on the observations assigned to cluster r
     b. Sample a new value of ϕ_r from this posterior distribution



If the atoms of G are distributions $g(\phi \mid \xi)$ from a family conjugate to $f(x \mid \phi)$ then the conditional distributions for sampling are easily obtained

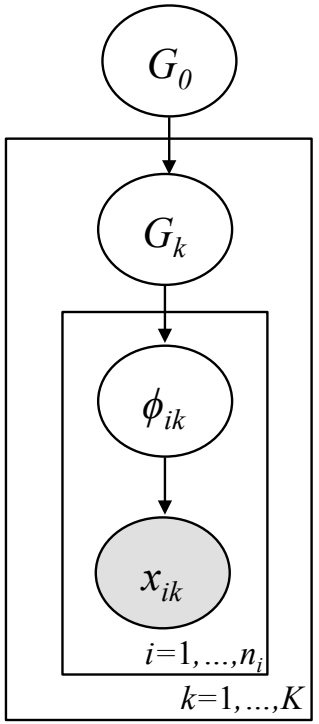
Variational Inference for DP Mixtures

- In Gibbs sampling we repeatedly sample each latent variable from its distribution given the observations, the hyperparameters, and the cluster assignments of the other variables
- Variational inference replaces sampling with an optimization problem:
  - Choose a family of distributions $q(z,v,\phi|\xi)$ on the latent cluster assignments z, stick-breaking variables v, and cluster parameters ϕ
  - A specific distribution $q(z,v,\phi|\xi^*)$ is chosen from this family to optimize a lower bound on the log-likelihood of the observations
  - Variational family $q(z,v,\phi|\xi)$ is chosen as a class of distributions for which the optimization problem is tractable
- A common approach is to choose a fully factored distribution in which all the variables are independent
- Variational EM algorithm repeatedly cycles through the latent variables, optimizing the parameters for each distribution given data and current estimates for other parameters



Hierarchical Dirichlet Processes

- Data is divided into J groups
  - Each group consists of clusters
  - Clusters are shared between groups (data points in same group can belong to same cluster



$$G_0 \mid H, \gamma \sim \mathrm{DP}(\bullet \mid H, \gamma)$$

$$G_k \mid G_0, \alpha \sim \mathrm{DP}(\bullet \mid G_0, \alpha), k=1, \ldots, K$$

- Common DP H forms base measure for drawing G_0
- G_0 is base measure for DP draw for each group

Source: Gharamani, 2005

Outline

- Parametric and nonparametric statistical models
- Dirichlet process mixtures and nonparametric clustering
- Gaussian processes and nonparametric regression

Regression

- Regression seeks to model a *dependent variable* as a function of one or more *independent variables* plus random noise

  $y_i = f(x_i) + \varepsilon_i$ where ε_i are iid RVs with $\mathrm{E}[\varepsilon_i] = 0$

- The most common regression model is normal linear regression
- e.g., Simple linear regression (one predictor):

  $f(x) = \beta_0 + \beta_1 x$ is a linear function of the predictor variable

  ε has a zero-mean normal distribution

- If the dependent variable has a nonlinear relationship to the independent variable(s), we can include polynomial terms in the regression, e.g.:

  $$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon$$

- Bayesian regression places a prior distribution on the regression coefficients $(\beta_0, \beta_1, \ldots, \beta_n)$
- We can view this as placing a prior distribution on polynomial functions of degree n
- We would like to allow a more flexible class of functions than polynomials of a fixed degree

Bayesian Linear Regression

Bayesian simple linear regression places a prior distribution on functions of the form $f(x) = w_1 + w_2 x$ and uses observations to infer a posterior distribution on this class of functions

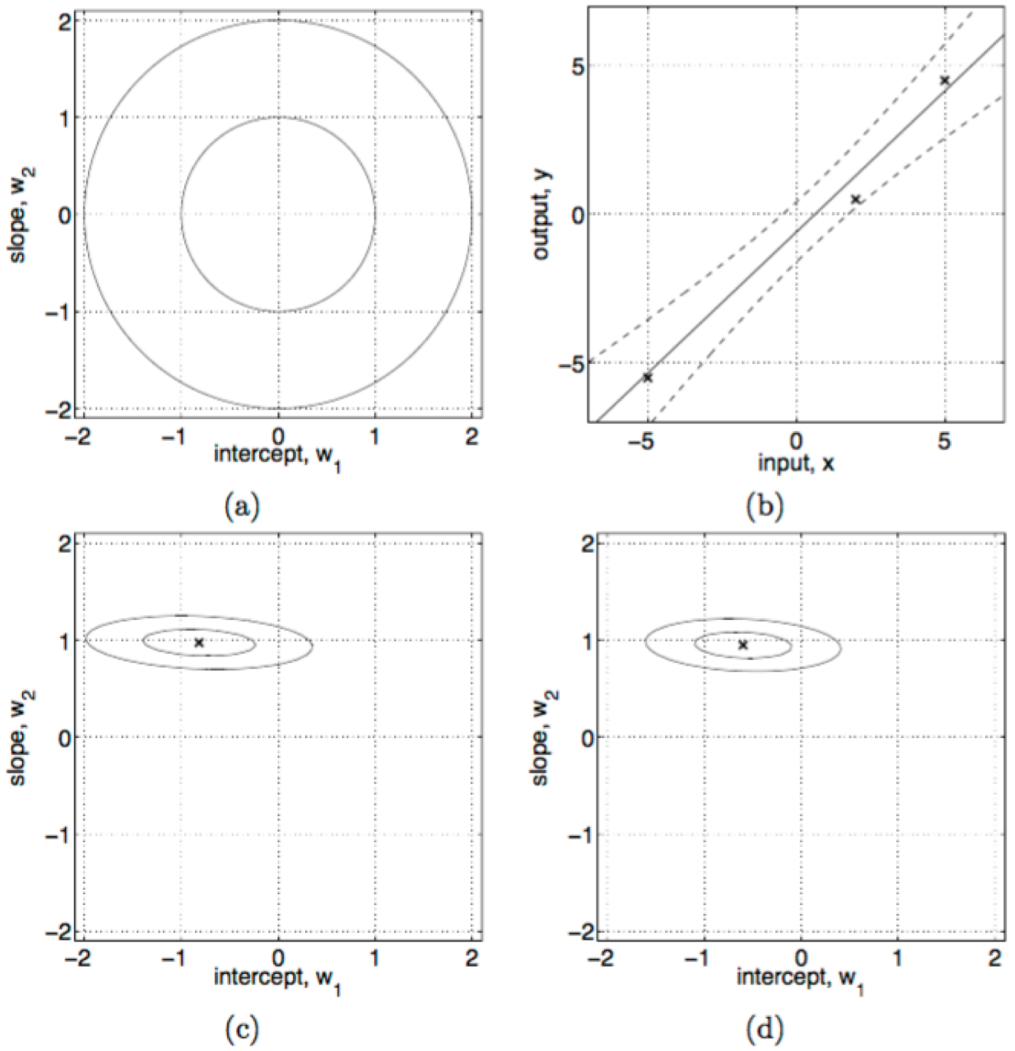


Figure 2.1: Example of Bayesian linear model $f(x) = w_1 + w_2 x$ with intercept w_1 and slope parameter w_2. Panel (a) shows the contours of the prior distribution $p(\mathbf{w}) \sim \mathcal{N}(\mathbf{0}, I)$, eq. (2.4). Panel (b) shows three training points marked by crosses. Panel (c) shows contours of the likelihood $p(\mathbf{y}|X, \mathbf{w})$ eq. (2.3), assuming a noise level of $\sigma_n = 1$; note that the slope is much more "well determined" than the intercept. Panel (d) shows the posterior, $p(\mathbf{w}|X, \mathbf{y})$ eq. (2.7); comparing the maximum of the posterior to the likelihood, we see that the intercept has been shrunk towards zero whereas the more 'well determined' slope is almost unchanged. All contour plots give the 1 and 2 standard deviation equi-probability contours. Superimposed on the data in panel (b) are the predictive mean plus/minus two standard deviations of the (noise-free) predictive distribution $p(f_*|\mathbf{x}_*, X, \mathbf{y})$, eq. (2.9).

Source: Rasmussen and Williams (2006)

Gaussian Process

- A linear or polynomial regression places a prior distribution on a restricted class of functions:

  $f(x) = \beta_0 + \beta_1 x$ for linear regression

  $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k$ for polynomial regression

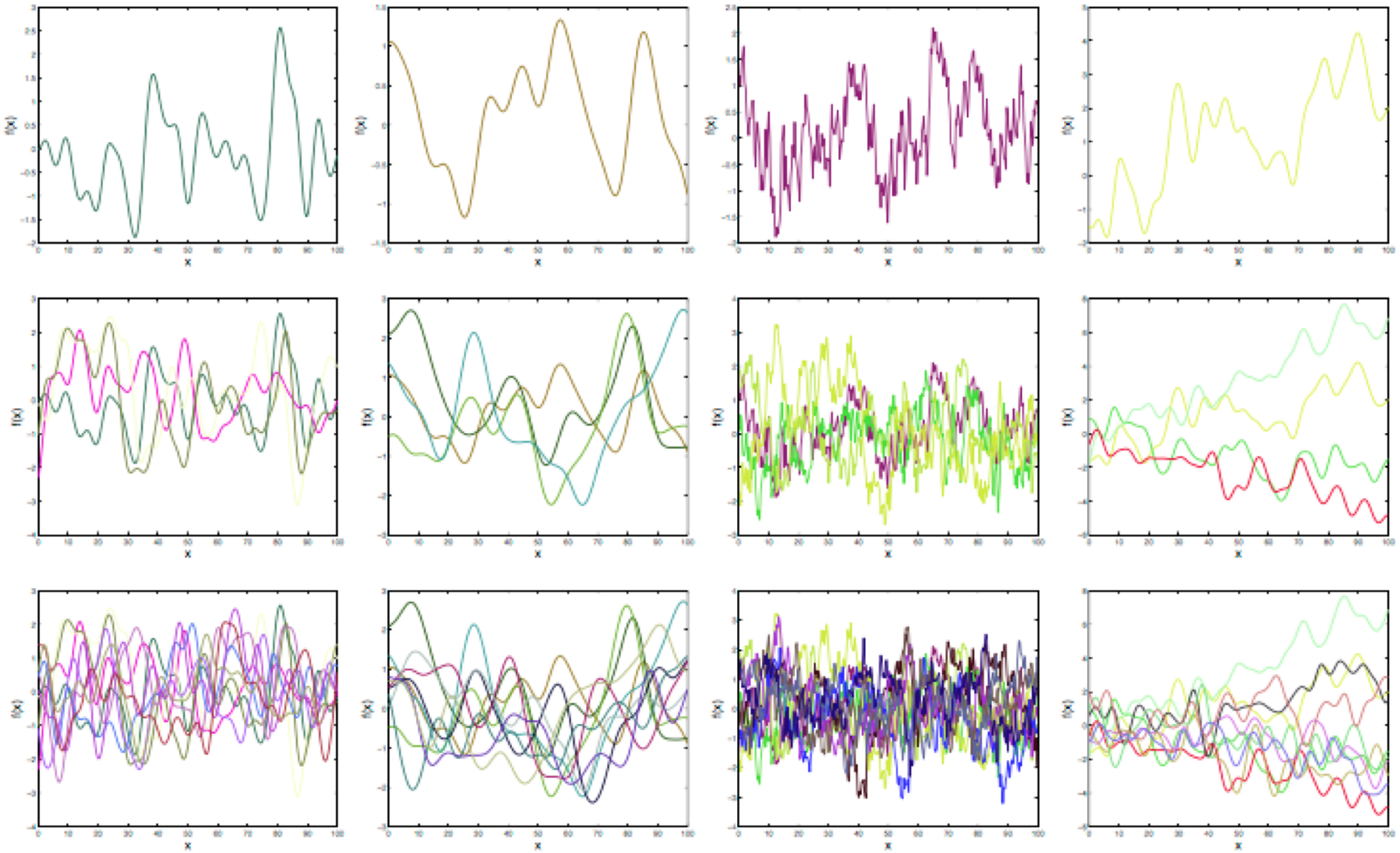
- A Gaussian process can place a distribution on a richer class of functions

Definition: A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution

  - Must satisfy *consistency property*: if $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$ then $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$, where μ_1 and Σ_{11} are the relevant subvector and submatrix of μ and Σ

- A Gaussian process is completely specified by its mean function $m(x)$ and covariance function $k(x,x')$ (aka *kernel*)
  - $m(x) = E[f(x)]$
  - $k(x,x') = E((f(x) - m(x))\,(f(x') - m(x'))]$
- A GP with mean function $m(x)$ and covariance function $k(x,x')$ satisfies the consistency property

Samples from Gaussian Processes with Different k(x,x')



Source: Gharamani, 2005

Features and Basis Functions

- The mean function for Bayesian linear regression is a Gaussian process:
  - $m(x) = \mathrm{E}[f(x)] = \mathrm{E}[\beta_0] + x\,\mathrm{E}[\beta_1]$
  - $k(x,x') = \mathrm{E}((\beta_0 + \beta_1 x - m(x))\,(\beta_0 + \beta_1 x' - m(x'))] = (1, x)\Sigma(1, x')^{\mathsf{T}}$, where $\Sigma = \mathrm{Cov}(\beta)$ is the covariance matrix of the regression coefficients
  - For any n, the function values $f(x_1), f(x_2), \ldots, f(x_n)$ are jointly Gaussian, although the covariance matrix is singular for $n > 2$
- The mean function for polynomial regression is also a Gaussian process
  - $m(x) = \mathrm{E}[f(x)] = \mathrm{E}[\beta_0] + x\,\mathrm{E}[\beta_1] + \ldots + x^k\,\mathrm{E}[\beta_1]$
  - $k(x,x') = (1, x, \ldots, x^k)\Sigma(1, x', \ldots, x'^k)^{\mathsf{T}}$, where $\Sigma = \mathrm{Cov}(\beta)$ is the covariance matrix of the polynomial regression coefficients
  - For any n, the function values $f(x_1), f(x_2), \ldots, f(x_n)$ are jointly Gaussian, although the covariance matrix is singular for $n > k+1$
- We can generalize this idea, mapping the dependent variable(s) into a high-dimensional space using a set of *basis functions* $(\varphi_1(x), \ldots, \varphi_k(x))$
  - $f(x) = \beta_1\varphi_1(x) + \beta_2\varphi_2(x) + \ldots + \beta_k\varphi_k(x)$
  - $m(x) = \varphi_1(x)\mathrm{E}[\beta_1] + \varphi_2(x)\mathrm{E}[\beta_2] + \ldots + \varphi_k(x)\mathrm{E}[\beta_k]$
  - $k(x,x') = \varphi(x)\Sigma\varphi(x')^{\mathsf{T}}$, where $\Sigma = \mathrm{Cov}(\beta)$ is the covariance matrix of the regression coefficients for the basis functions

Covariance Functions and Basis Functions

- A commonly used covariance function is the squared exponential*

$$k(x,x') = v\exp\left\{-\tfrac{1}{2}\left(\frac{(x-x')^2}{\lambda}\right)\right\}$$

  - » v controls the overall variance of the process
  - » λ controls the length scale
  - » GP with squared exponential covariance function is infinitely mean-square differentiable
- Squared exponential covariance function corresponds to a Bayesian linear regression with infinitely many basis functions
- For any positive definite covariance function, there is an expansion in terms of (possibly infinitely many) basis functions for a Bayesian linear regression
- Other examples of basis functions and corresponding covariance functions can be found at

  http://www.gaussianprocess.org/gpml/chapters/RW4.pdf

*Also called radial basis function (RBF) or Gaussian

Bayesian Inference with Noise-Free Observations

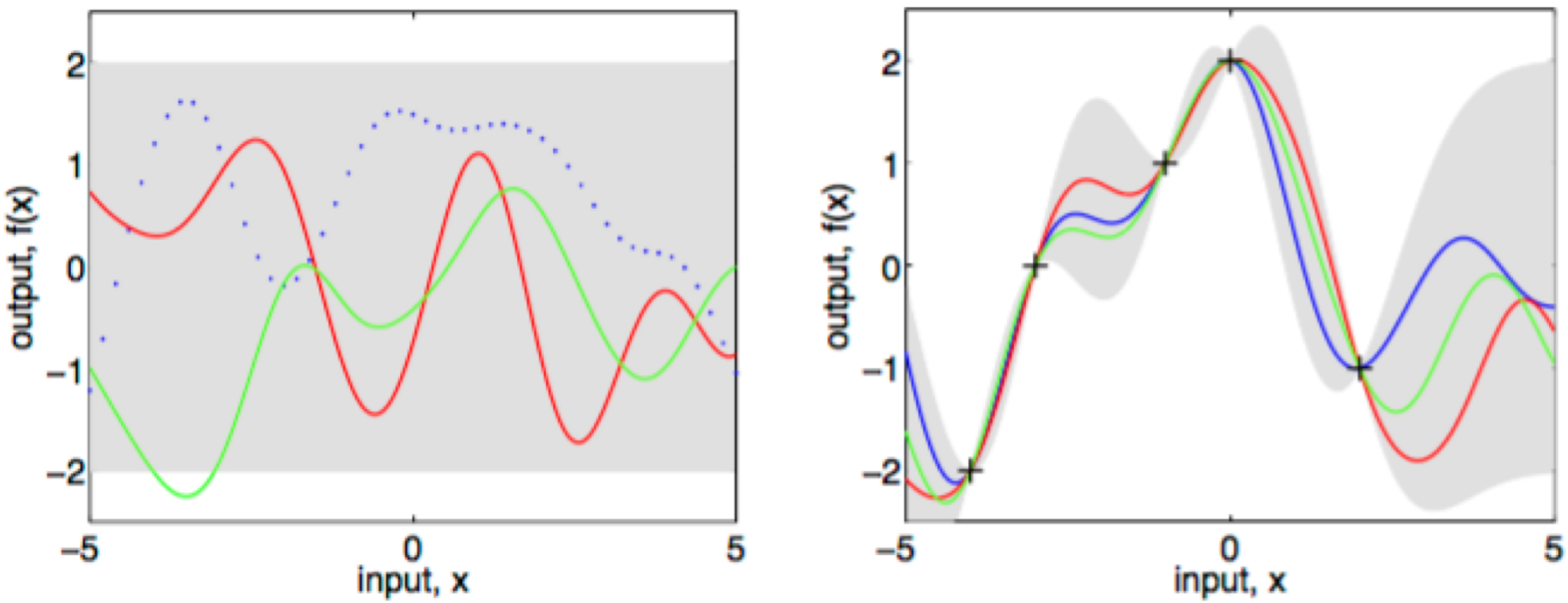
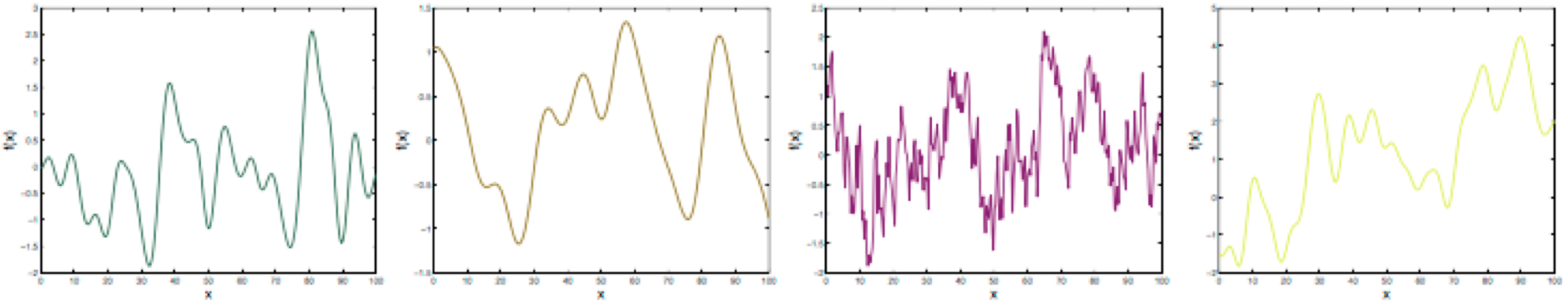


Figure 2.2: Panel (a) shows three functions drawn at random from a GP prior; the dots indicate values of y actually generated; the two other functions have (less correctly) been drawn as lines by joining a large number of evaluated points. Panel (b) shows three random functions drawn from the posterior, i.e. the prior conditioned on the five noise free observations indicated. In both plots the shaded area represents the pointwise mean plus and minus two times the standard deviation for each input value (corresponding to the 95% confidence region), for the prior and posterior respectively.

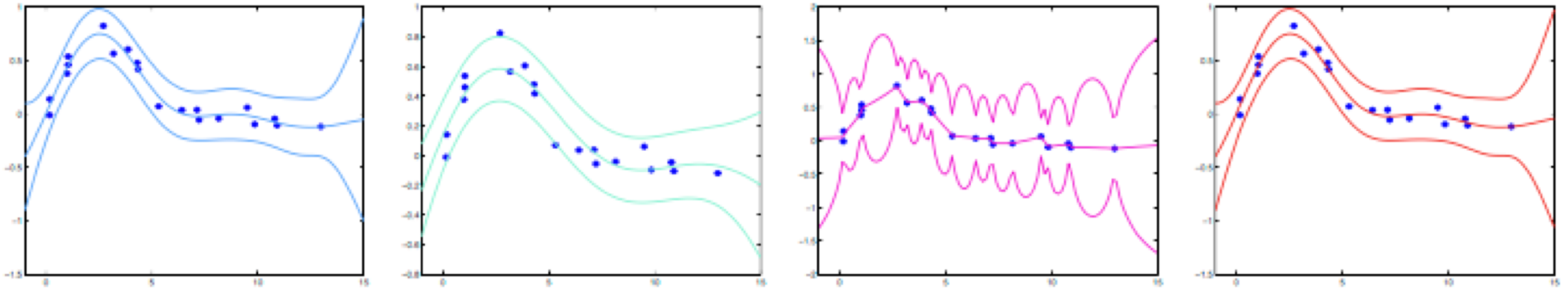
Source: Rasmussen and Williams (2006)

Bayesian Inference with Noisy Observations

A sample from the prior for each covariance function:



Corresponding predictions, mean with two standard deviations:



Source: Gharamani, 2005

Graphical Model Representation of GP Regression

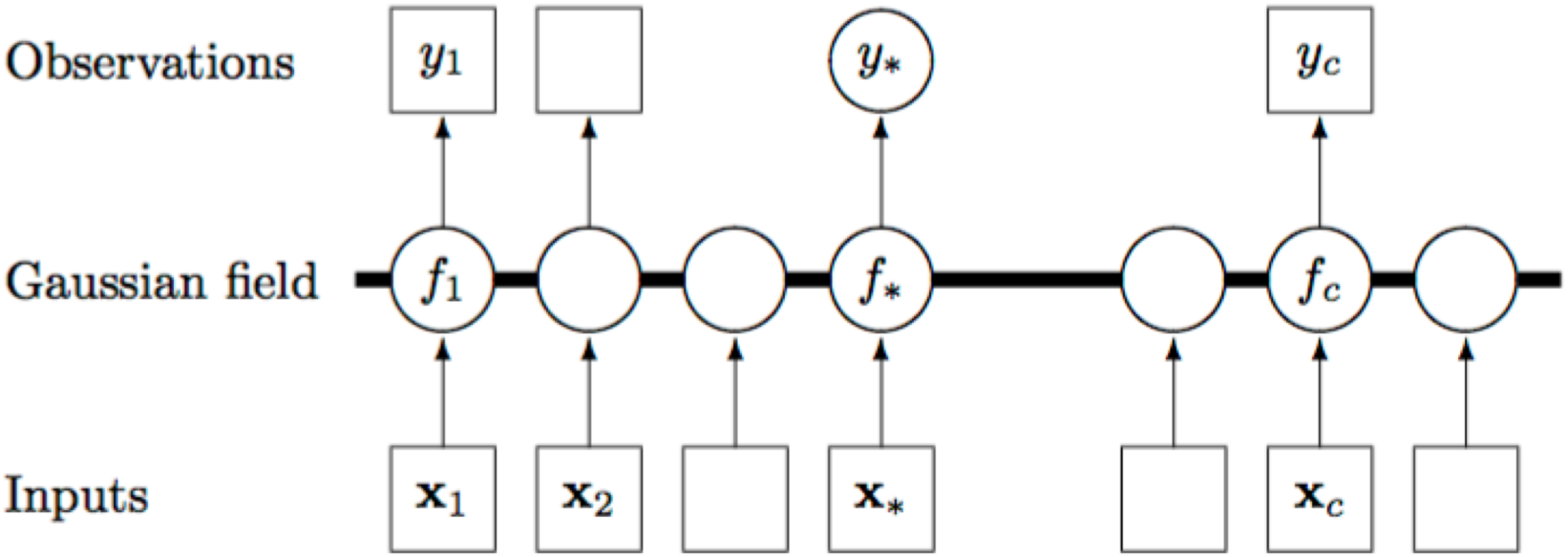


Figure 2.3: Graphical model (chain graph) for a GP for regression. Squares represent observed variables and circles represent unknowns. The thick horizontal bar represents a set of fully connected nodes. Note that an observation y_i is conditionally independent of all other nodes given the corresponding latent variable, f_i. Because of the marginalization property of GPs addition of further inputs, $\mathbf{x}$, latent variables, f, and *unobserved* targets, y_*, does not change the distribution of any other variables.

Source: Rasmussen and Williams (2006)

Inference with Gaussian Process

- Gaussian process regression models dependent variable as a Gaussian process plus iid normal zero-mean noise

$$y_i = f(x_i) + \varepsilon_i$$
$$\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0,\sigma^2)$$

- Prior distribution for f is a Gaussian process

$$f \sim \mathcal{GP}(0, k) \qquad \text{(for simplicity assume zero-mean prior)}$$

- Note that the prior distribution for y is also a Gaussian process with mean function m and covariance function k_y, where

$$k_y(x,x) = k(x,x) + \sigma_2 \text{ and } k_y(x,x') = k(x,x') \text{ for } x \neq x'$$

- Observe data $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$
- Posterior predictive distribution for $y(x)$ is Gaussian process with

$$m^*(x) = (k(x,x_1), \ldots, k(x,x_n))\,(K(X,X) + \sigma^2 \mathrm{I})^{-1}\,(y_1, \ldots, y_n)^{\mathsf{T}}$$
$$k_y^*(y,y') = K(Y,Y) - K(Y,X)\,(K(X,X) + \sigma^2 \mathrm{I})^{-1}\,K(X, (y,y'))$$

$K(X,X)$ is the matrix with entries $k(x_i,x_j)$ for $i,j=1,\ldots,n$
$K(Y,X)$ is the matrix with entries $k(y,x_j)$ and $k(y',x_j)$ for $j=1,\ldots,n$
$K(X,Y) = K(Y,X)^{\mathsf{T}}$
$K(Y,Y)$ is the matrix with diagonal elements $k(y,y)$ and $k(y',y')$ and off-diagonal elements $k(y,y')$

Working With High-Dimensional Feature Spaces: the Kernel Trick

- Gaussian process regression can be viewed as mapping the original dependent variable(s) into an implicit infinite-dimensional feature space (the basis functions)
  - The learning algorithm operates with the *kernel function* $k(x,x')$
  - $k(x,x')$ is the inner product of projections of pairs of data points into the infinite-dimensional feature space:

    $k(x,x') = \varphi(x)\Sigma\varphi(x')^T$, where Σ is the covariance matrix of the regression coefficients for the basis functions
  - We never actually compute the features (basis functions) in this infinite-dimensional space
  - Working with the kernel is computationally cheaper than working directly with the basis functions
- The approach of working with kernels rather than explicitly with features in a high (or infinite) dimensional space is known as the *kernel trick*
- There are many machine learning approaches that make use of the kernel trick

Summary and Synthesis

- Nonparametric methods allow the dimension of the model to adapt to the data and for dimensionality to grow as the number of observations increases
- Distributions used for nonparametric models are often generalizations of commonly applied finite-dimensional models
  - Conjugacy is inherited from finite-dimensional case
  - Inference methods are similar to finite-dimensional case
- We studied two of the most commonly applied nonparametric models
  - Dirichlet process mixture models for clustering
  - Gaussian process for Bayesian nonlinear regression
- Bayesian nonparametrics is an active research topic and many new methods are being developed

References

- David Blei and Michael Jordan. Variational Analysis for Dirichlet Process Mixtures. Bayesian Analysis 1(1), pp. 121-144, 2006 http://www.cs.columbia.edu/~blei/papers/BleiJordan2004.pdf
- Zoubin Gharamani, Non-parametric Bayesian Methods, Tutorial at Conference on Uncertainty in Artificial Intelligence, 2005. http://mlg.eng.cam.ac.uk/zoubin/talks/uai05tutorial-b.pdf
- Peter Orbanz, Lecture Notes on Bayesian Nonparametrics, May 16, 2014, http://stat.columbia.edu/~porbanz/papers/porbanz_BNP_draft.pdf
- C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X http://www.GaussianProcess.org/gpml