Module 3: Bayesian Nonparametrics

Gaussian Processes for Regression

STAT/BIOSTAT 527, University of Washington
Emily Fox

April 21st-22nd, 2014

Recap of regression so far

- Recall our regression setting

$$f(x) = E[Y \mid x]$$

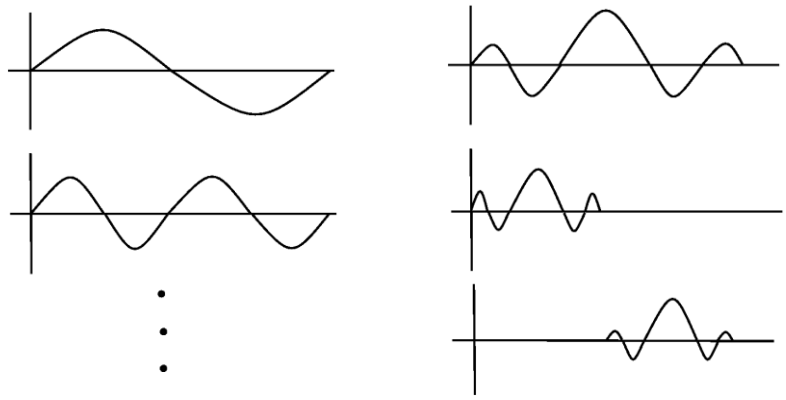
- How to estimate from finite training set?

Restrict to model class

- Example = linear basis expansion
  - Standard linear $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$
  - Polynomial $y = \sum_j \beta_j x^j + \epsilon$
  - Splines
  - …

good locally, but not globally

Other Important Basis Expansions



Fourier Basis

Wavelet Basis

not looking at these in this class

Recap of regression so far

- Recall our regression setting

$$f(x) = E[Y \mid x]$$

- How to estimate from finite training set?

Restrict to model class

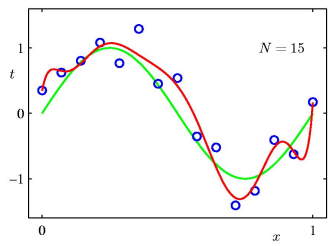
- Example = linear basis expansion

Overfitting as model complexity grows

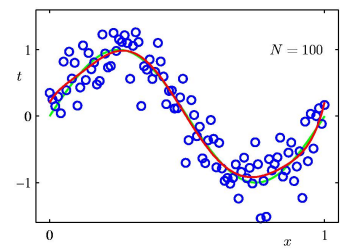
- Penalized linear basis expansions (regularized LS)
  - Ridge
  - Lasso
  - Smoothing splines ← equivalent to searching over all fcns subject to a smoothness constraint
  - Penalized regression splines

Overfitting

9th Order Polynomial



$n = 15$



$n = 100$

- model complexity is relative to sample size
- can consider more complex forms with more data

Recap of regression so far

- Recall our regression setting

$$f(x) = E[Y \mid x]$$

- How to estimate from finite training set?

Restrict to model class

Local nbhd methods

- Example = linear basis expansion

Overfitting as model complexity grows

- Penalized linear basis expansions

- Example = kernel regression

k-NN regression
local averages
Nadarya-Watson
locally weighted poly.

Again: Linear Basis Expansion

- Instead of just considering input variables *x* (potentially mult.), augment/replace with transformations = "input features"

  In this lecture, we'll focus on these forms

- ***Linear basis expansions*** maintain linear form in terms of these transformations

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$$

trans.

- What transformations should we use?
  - $h_m(x) = x_m$ → linear model
  - $h_m(x) = x_j^2, \quad h_m(x) = x_j x_k$ → polynomial reg.
  - $h_m(x) = I(L_m \leq x_k \leq U_m)$ → piecewise constant
  - …

Making Predictions

- So far, our focus has been on L_2 loss:

plus penalty

$$\min_{\beta} \ \mathrm{RSS}(\beta) + \lambda ||\beta||$$

$\sum_i (y_i - \hat{f}(x_i))^2 \quad f(x) = \beta^T h(x)$

- Here, we assumed $y = f(x) + \epsilon$ with $E[\epsilon]=0 \quad var(\epsilon)=\sigma^2$

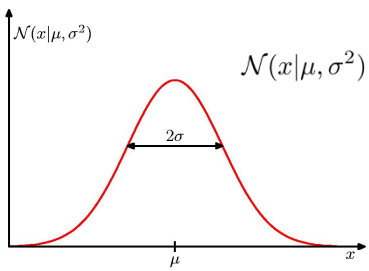
- Now, let's assume a distributional form and log-likelihood loss

$\epsilon \sim N(0,\sigma^2) \Rightarrow p(y|f(x),\sigma^2) = N(f(x),\sigma^2)$

First, recall some facts about Gaussians…

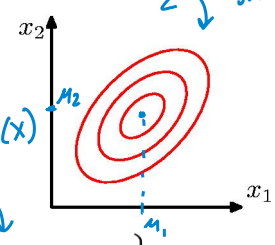
Quick Review of Gaussians

- Univariate and multivariate Gaussians



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Σ defines shape



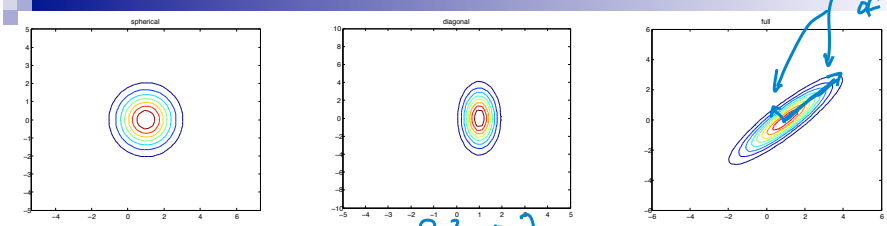
$\Sigma = \mathrm{COV}(X)$

$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

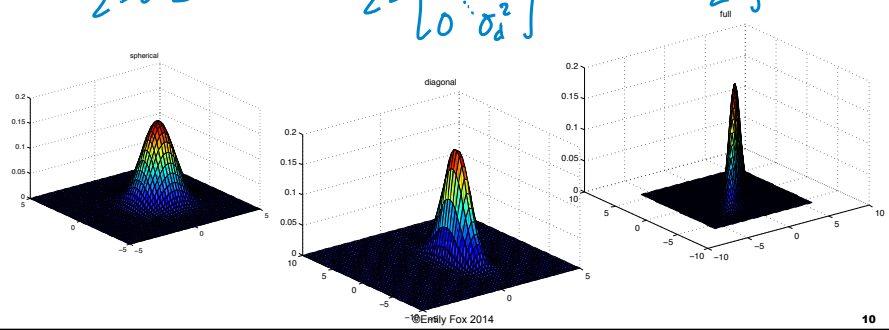
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Two-Dimensional Gaussians

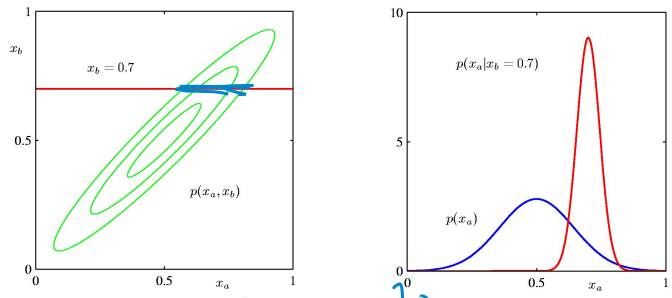
eigvecs of Σ



$\Sigma = \sigma^2 I$ $\quad$ $\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix}$ $\quad$ Σ general



Conditional & Marginal Distributions



$$\begin{pmatrix} x_a \\ x_b \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \right)$$

Marg: $x_a \sim N(\mu_a, \Sigma_{aa})$

Cond: $x_a | x_b \sim N(\mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b), \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})$

Maximum Likelihood Estimation

- Model:

$$y = f(x) + \epsilon \quad \text{where} \quad \epsilon \sim N(0, \sigma^2)$$

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$$

- Equivalently,

$$p(y \mid x, \beta, \sigma^2) = N(y \mid f(x), \sigma^2)$$

- For our training data (independent obs) $(x_1, y_1), \ldots, (x_n, y_n)$

$$p(y \mid X, \beta, \sigma^2) = \prod_{i=1}^{n} N(y_i \mid f(x_i), \sigma^2)$$

Maximum Likelihood Estimation

$$p(y \mid X, \beta, \sigma^2) = \prod_i N(y_i \mid \beta^T h(x_i), \sigma^2)$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Taking the log

$$\log p(y \mid X, \beta, \sigma^2) = \sum_i -\frac{1}{2}(y_i - \beta^T h(x_i))^2 \underbrace{- \frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2}_{\text{const. wrt } \beta}$$

- Equivalent objective to RSS *(Gaussian log-like loss = L_2 loss)*

- Taking the gradient and setting to zero, we have already shown

$$\hat{\beta}^{ML} = (H^T H)^{-1} H^T y$$

$$H = \begin{pmatrix} h_1(x_1) \dots h_M(x_1) \\ \vdots \\ h_1(x_n) \dots h_M(x_n) \end{pmatrix}$$

Look at penalized regression

A Bayesian Formulation

- Consider a model with likelihood

$$y_i \mid \beta \sim N(\beta_0 + x_i^T \beta, \sigma^2)$$

if $\varepsilon \sim N(0, \sigma^2)$

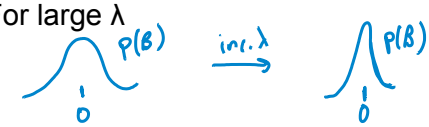
and prior

$$\beta \sim N\left(0, \frac{\sigma^2}{\lambda} I_p\right)$$

$\beta_j \sim N(0, \frac{\sigma^2}{\lambda})$

prior places penalty

- For large λ



prior peaked around $\beta = 0$ ⇒ penalizing β far from 0

- The posterior is

$$\beta \mid y \sim N\left(\hat{\beta}^{ridge}, \sigma^2(X^T X + \lambda I)^{-1} X^T X \sigma^2 (X^T X + \lambda I)^{-1}\right)$$

$\boxed{\hat{\beta}^{MAP} = \hat{\beta}^{ridge}}$

works against overfitting of MLE

easy to show $var(\hat{\beta}^{ridge})$

Bayesian Linear Regression

- More generally, consider a conjugate prior on the basis expansion coefficients:

$$p(\beta) = N(\beta \mid \mu_0, \Sigma_0)$$

prior mean + cov.

- Combining this with the Gaussian likelihood function, and using standard Gaussian identities, gives posterior

$$p(\beta \mid y) = N(\beta \mid \mu_n, \Sigma_n)$$

posterior (updated) mean + cov. after obs. y

where

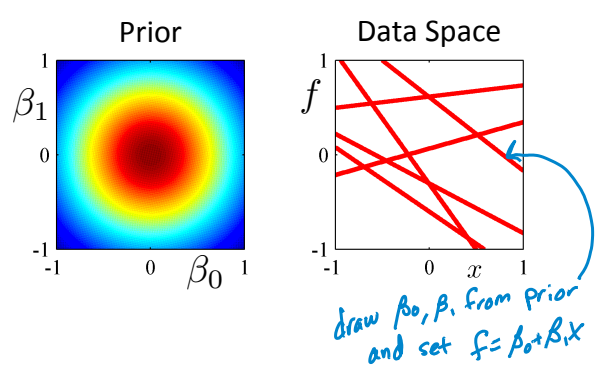
$$\mu_n = \Sigma_n (\Sigma_0^{-1}\mu_0 + \sigma^{-2} H^T y)$$

$$\Sigma_n^{-1} = \Sigma_0^{-1} + \sigma^{-2} H^T H$$

Example: Standard Linear Basis

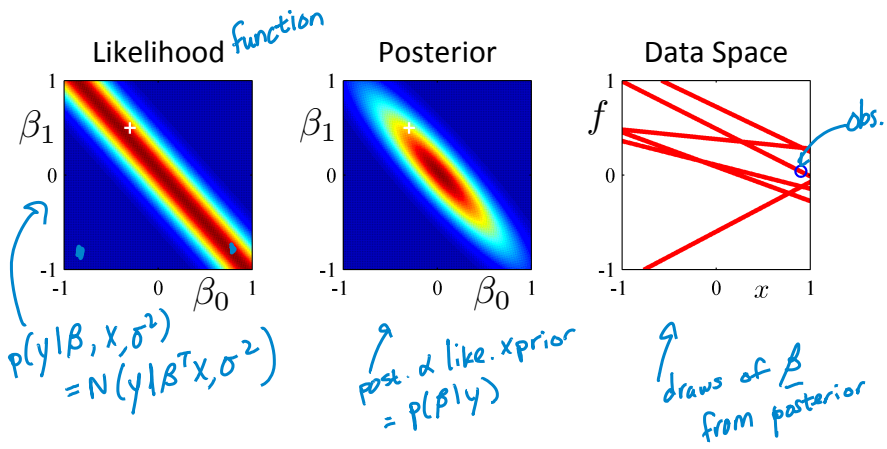
0 data points observed

$y = \beta_0 + \beta_1 X + \epsilon$



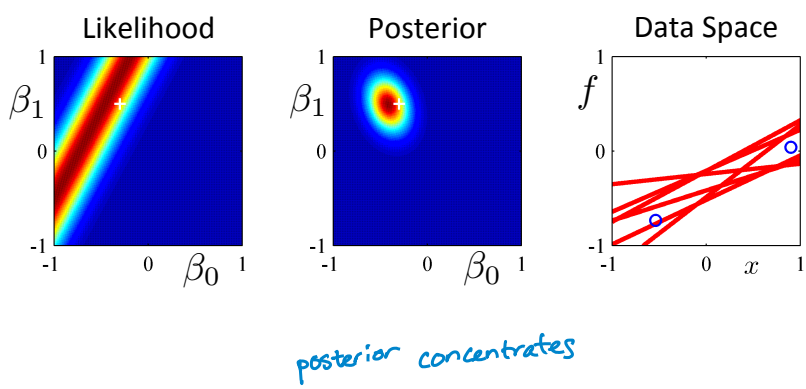
Example: Standard Linear Basis

1 data point observed



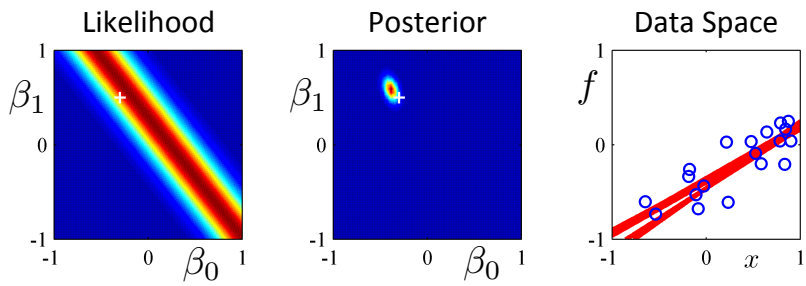
Example: Standard Linear Basis

2 data points observed



Example: Standard Linear Basis

20 data points observed



Predictive Distribution

- Predict y^* at new locations x^* by integrating over parameters β

$y_1, \dots, y_n$

$$p(y^* \mid y) = \int p(y^* \mid \beta) p(\beta \mid y) d\beta$$

x^*, x

likelihood

posterior:

$$p(\beta \mid y) = N(\beta \mid \mu_n, \Sigma_n)$$

$$p(y \mid x, \beta, \sigma^2) = N(y \mid f(x), \sigma^2)$$

$\beta^T h(x)$

$y^* = h(x^*)^T \beta + \epsilon$

$\beta \mid y \sim N(\mu_n, \Sigma_n)$

$\epsilon \sim N(0, \sigma^2)$

Σ_n fcn of our obs. locations $x_1, \dots, x_n$

$$= N(\mu_n^*(x^*), \Sigma_n^*(x^*))$$

$$\mu_n^*(x^*) = E[y^* \mid y] = \mu_n^T h(x^*)$$

$$\Sigma_n^*(x^*) = \mathrm{cov}(y^* \mid y) = h(x^*)^T \mathrm{cov}(\beta \mid y) h(x^*) + \sigma^2$$

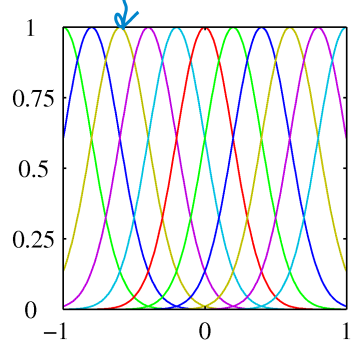
$$= \underbrace{h(x^*)^T \Sigma_n h(x^*)}_{\text{var of } \beta} + \underbrace{\sigma^2}_{\text{var of obs.}}$$

Example: Gaussian Basis Expansion

- Gaussian basis functions:

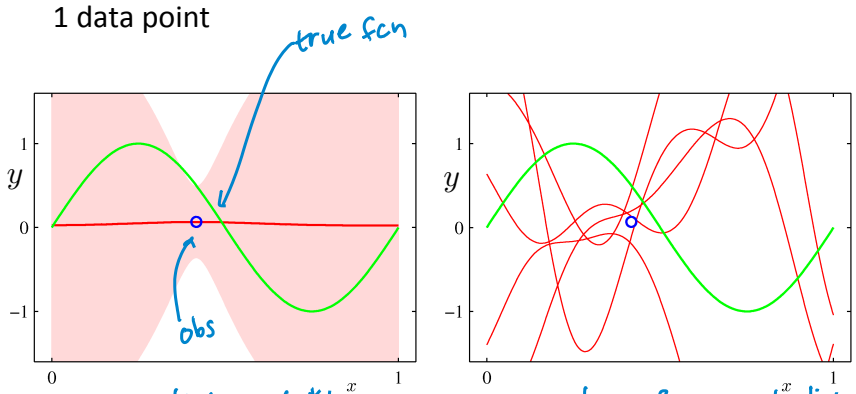
$$h_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

- These are local; a small change in *x* only affects nearby basis functions. Parameters control location and scale (width)



Example: Gaussian Basis Expansion

- Example: Sinusoidal data, 9 Gaussian basis functions, 1 data point

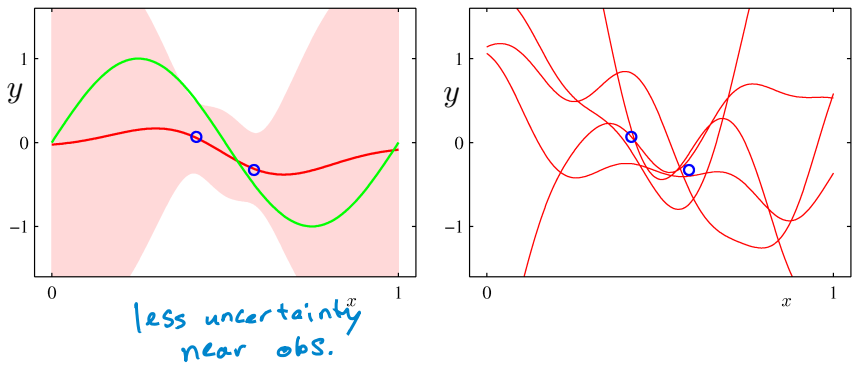


pred. dist. $p(y^*|y)$ shows pointwise uncertainty

draw from pred. dist.

Example: Gaussian Basis Expansion

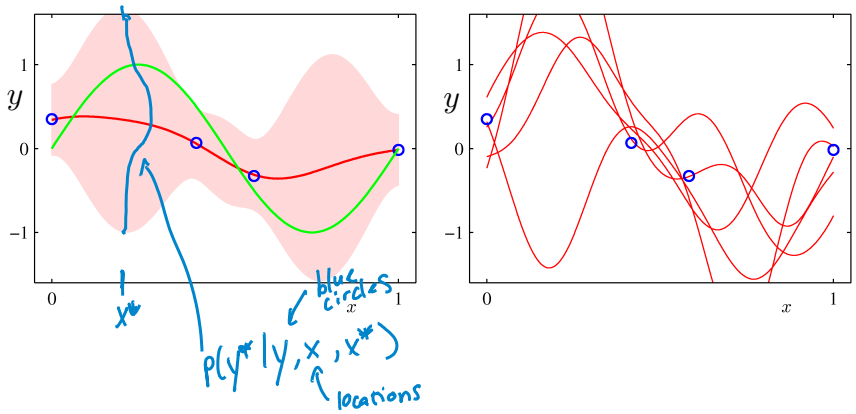
- Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points



less uncertainty near obs.

Example: Gaussian Basis Expansion

- Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points



x^*

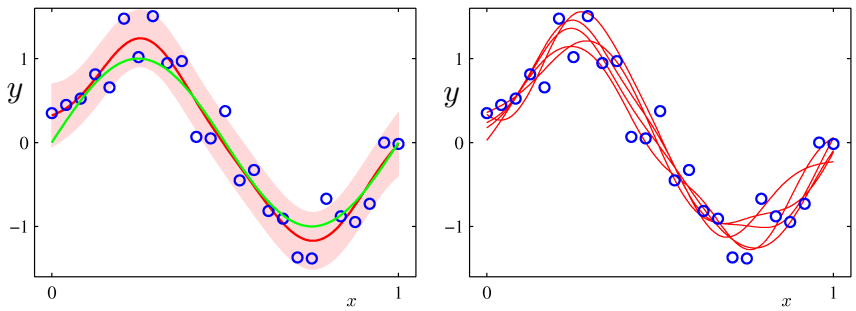
blue circles

$p(y^* | y, x, x^*)$

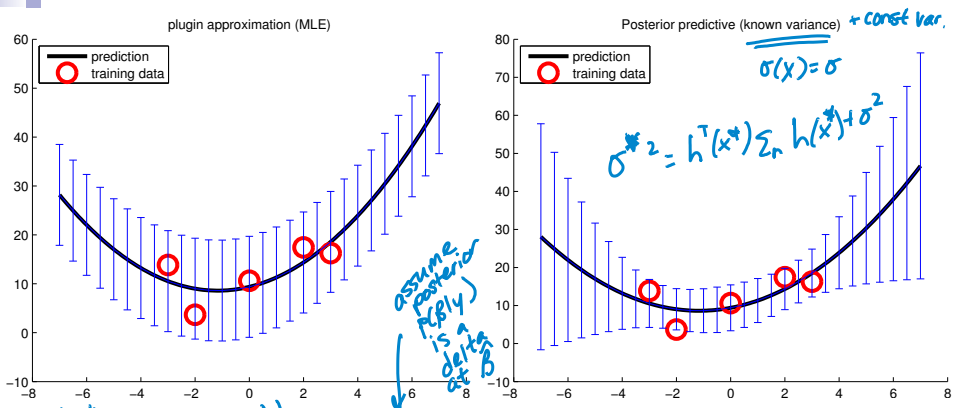
locations

Example: Gaussian Basis Expansion

- Example: Sinusoidal data, 9 Gaussian basis functions, 25 data points



Estimation vs. Predictive Distributions



$$p(y^* | x^*, x, y, \sigma^2) \approx \int N(y^* | \beta^T h(x^*), \sigma^2)\, \delta_{\hat{\beta}}(\beta)\, d\beta = N(y^* | \hat{\beta}^T h(x^*), \sigma^2)$$

constant error bars

What # of basis fcns should we use?

Bayesian Model Selection

- Assume some *M* possible models
  - Model M_m *m=1,…,M* has parameters θ_m and prior $p(\theta_m \mid M_m)$
  - Prior over models $p(M_m)$
- Model posterior — training data $Z=(Y,X)$

$$p(M_m \mid Z) \propto p(M_m)p(Z \mid M_m)$$

$$\propto p(M_m) \int p(Z \mid \theta_m, M_m)p(\theta_m \mid M_m)d\theta_m$$

eg mean + cov of Gauss or $\{\beta_j\}$

- Compare models:

posterior odds

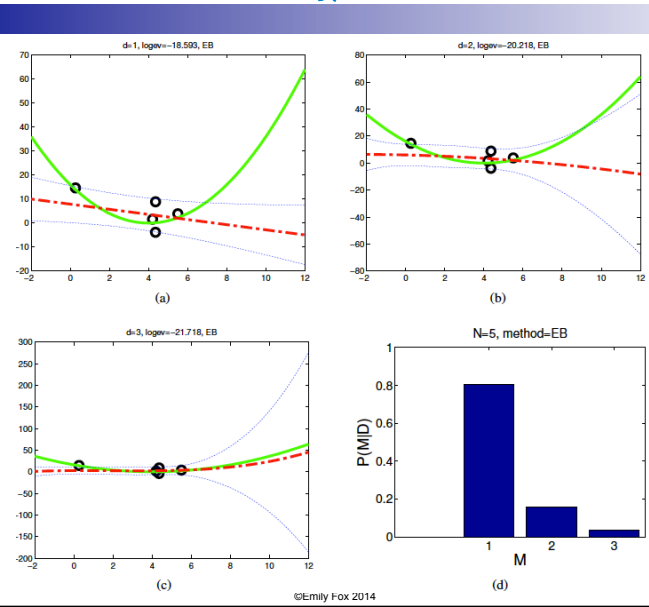
$$\frac{p(M_m \mid Z)}{p(M_\ell \mid Z)} = \frac{p(M_m)\boxed{p(Z \mid M_m)}}{p(M_\ell)\boxed{p(Z \mid M_\ell)}} \gtrless 1$$

"marginal likelihood"

often, uniform prior

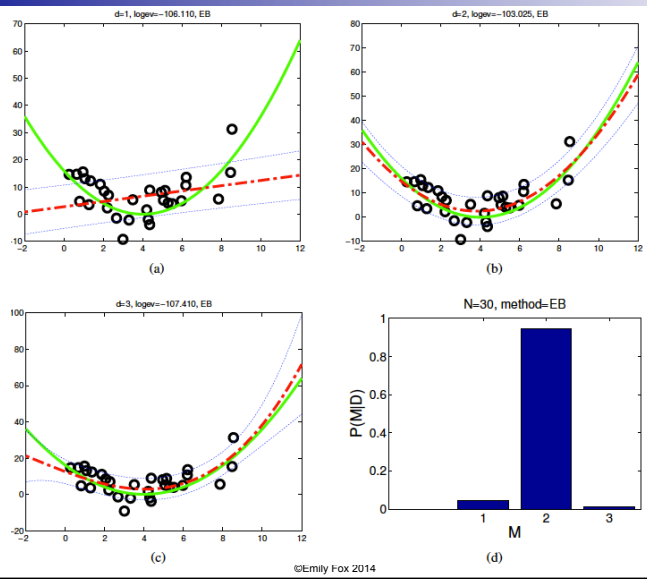
Bayes factor

BMS Example (n=5)



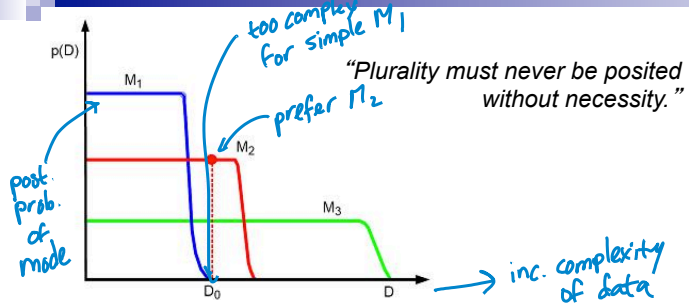
high post. prob. on simple models

BMS Example (n=30)



favors slightly more complex models

Bayesian Ockham's Razor



"Plurality must never be posited without necessity."

William of Ockham

- **Parametric Bayes:** Consider a finite list of possible models, average according to posterior probability (or in practice, just select the most probable)
- **Nonparametric Bayes:** Consider a single infinite model, integrate over parameters when making predictions or infer which finite subset is exhibited in your dataset

@9:00 am

$M \to \infty$

Acknowledgements

Many figures courtesy Kevin Murphy's textbook
Machine Learning: A Probabilistic Perspective,
and Chris Bishop's textbook
Pattern Recognition and Machine Learning

Slides based on parts of the lecture notes of Erik Sudderth for "Applied Bayesian Nonparametrics" at Brown University