Module 2: Splines and Kernel Methods

Inference for Linear Smoothers

STAT/BIOSTAT 527, University of Washington

Emily Fox

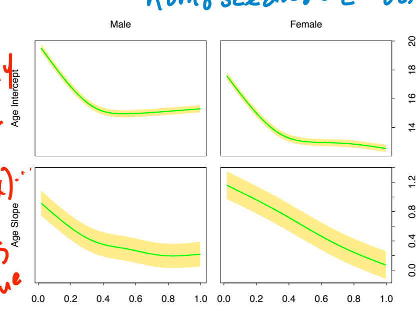
April 23rd, 2013

Confidence Bands

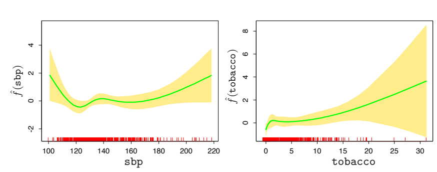
- So far we have focused on point estimation: $\hat{f}(x)$
- Often, we want to define a ***confidence interval*** for which $f(x)$ is in this interval with some pre-specified probability
- Looking over all *x*, we refer to these as ***confidence bands***

really for $\bar{f}(x)$... bias issue

homoscedastic $\sigma(x) = \sigma$



heterosedastic $\sigma(x)$



From Hastie, Tibshirani, Friedman book

CIs for Linear Smoothers

- For linear smoothers, and assuming constant variance $\sigma(x) = \sigma$

$$\hat{f}(x) = \sum_{i=1}^{n} \ell_i(x) y_i$$

$\bar{f}(x) = \sum_{i=1}^{n} \ell_i(x) f(x_i)$

$\text{var}(\hat{f}(x)) = \sigma^2 \|\ell(x)\|^2$

- Consider confidence band of the form

$CI(x) = \hat{f}(x) \pm c\,\hat{\sigma}\,\|\ell(x)\|$ $\quad a \le x \le b$

$c > 0$ $\quad$ est. of σ

- Using this, let's solve for *c*

CIs for Linear Smoothers

- Based on approach of Sun and Loader (1994)
  - Case #2: Assume σ unknown $\quad$ use est. $\hat{\sigma}$?
  - Case #3: Assume $\sigma(x)$ non-constant

$$\text{var}(\hat{f}(x)) = \sum_i \sigma^2(x_i)\, \ell_i^2(x)$$

$$\text{CI}(x) = \hat{f}(x) \pm c \sqrt{\sum_i \sigma^2(x_i)\, \ell_i^2(x)} \quad \star$$

  - If $\hat{\sigma}(x)$ varies slowly with *x*, then (Faraway and Sun 1995)

$\sigma(x_i) \approx \sigma(x)$ for those x w/ $\ell_i(x)$ large

$\Rightarrow CI(x) = \hat{f}(x) \pm c\,\hat{\sigma}(x)\,\|\ell(x)\| \quad \star$?

Variance Estimation

- In most cases σ is unknown and must be estimated
- For linear smoothers, consider the following estimator

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{f}(x_i))^2}{n - 2\nu + \tilde{\nu}}$$

$\nu = tr(L)$ $\quad \tilde{\nu} = tr(L^T L) = \sum_i \|\ell(x_i)\|^2$

  - If target function is sufficiently smooth, $\nu = o(n),\ \tilde{\nu} = o(n)$
  - Then $\hat{\sigma}^2$ is a consistent estimator of σ^2

Variance Estimation

- Proof outline:
  - Recall that

$$\underbrace{Y}_{(y_1,\dots,y_n)^T} - \hat{f} = Y - LY = (I-L)Y \triangleq \Lambda^{1/2} Y$$

  and

$$E[Y^T Q Y] = \mathrm{tr}(Q\underbrace{V}_{var(Y)}) + \underbrace{\mu^T}_{E[Y]} Q \mu$$

vec. matrix

  - Then,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{f}(x_i))^2}{n - 2\nu + \tilde{\nu}} = \frac{Y^T \Lambda Y}{tr(\Lambda)}$$

$\Lambda = (I-L)^T(I-L)$
$= I - 2L + L^T L$
$tr(\Lambda) = n - 2\nu + \tilde{\nu}$

$$E[\hat{\sigma}^2] = \frac{tr(\Lambda \sigma^2) + f^T \Lambda f}{tr(\Lambda)} = \sigma^2 + \underbrace{\frac{f^T \Lambda f}{n - 2\nu + \tilde{\nu}}}_{\text{small for large } n \text{ assuming } f \text{ smooth}}$$

  - Therefore, bias→0 for large *n* if *f* is smooth.
  - Likewise for variance.

Alternative Estimator

- Estimator:

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i-1}^{n-1} (y_{i+1} - y_i)^2$$

- Motivation:

$$y_{i+1} - y_i = \underbrace{[f(x_{i+1}) - f(x_i)]}_{\approx 0 \text{ if } f \text{ smooth}} + [\epsilon_{i+1} - \epsilon_i]$$

leads to

$$E[(y_{i+1} - y_i)^2] \approx E[\epsilon_{i+1}^2] + E[\epsilon_i^2] = 2\sigma^2$$

$$\Rightarrow E[\hat{\sigma}^2] \approx \sigma^2$$

- Estimator will be inflated ignores $f(x_{i+1}) - f(x_i)$
- Other estimators exist, too. See Wakefield or Wasserman.

Heteroscedasticity

- The point estimate $\hat{f}(x)$ is relatively insensitive to heterosced., but confidence bands need to account for non-constant variance
- Re-examine model $y_i = f(x_i) + \sigma(x_i)\epsilon_i$ ~~N(0,1)~~ $E[\epsilon]=0$ $var(1)$
  - var $\sigma^2(x_i)$
  - Define redefine obs: $Z_i = \log(y_i - f(x_i))^2 \quad \delta_i = \log \epsilon_i^2$
  - Then, $Z_i = \underbrace{\log(\sigma^2(x_i))}_{\text{est. w/ log sq. residuals}} + \delta_i$
- Algorithm:
  1. Estimate $f(x)$ using a nonparametric method w/ constant var to get $\hat{f}(x)$
  2. Define $Z_i = \log(y_i - \hat{f}(x_i))^2$
  3. Regress Z_i's on x_i's to get estimate $\hat{g}(x)$ of $\log \sigma^2(x)$ new obs.

$$\Rightarrow \hat{\sigma}^2(x) = e^{\hat{g}(x)}$$

$$Z_i = \underbrace{g(x_i)}_{\log(\sigma^2(x_i))} + \delta_i$$

Heteroscedasticity

- Drawbacks:
  - Taking log of a very small residual leads to a large outlier
  - A more statistically rigorous approach is to jointly estimate *f, g*

- Alternative = Generalized linear models

Module 3: Bayesian Nonparametrics

Gaussian Processes

STAT/BIOSTAT 527, University of Washington
Emily Fox
April 23rd, 2013

Recap of regression so far

- Recall our regression setting

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

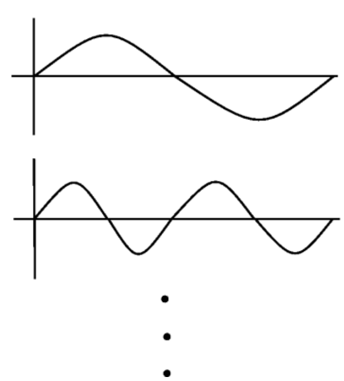
- How to estimate from finite training set?

Restrict to model class

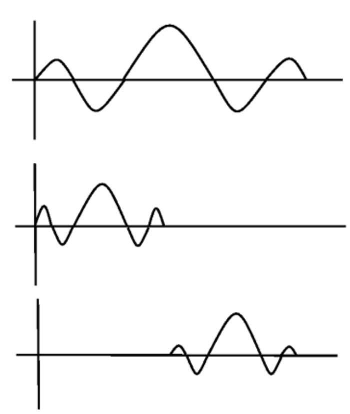
- Example = linear basis expansion
  - Standard linear $y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$
  - Polynomial $y = \sum \beta_j x^j + \epsilon$
  - Splines
  - …

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

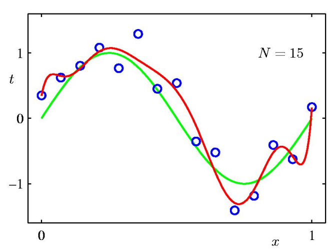
- Example = linear basis expansion

Overfitting as model complexity grows

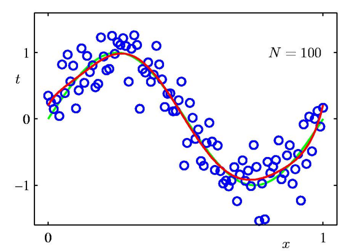
- Penalized linear basis expansions (regularized LS)
  - Ridge
  - Lasso
  - Smoothing splines ← equivalent to searching over all fcns s.t. smoothness constraints
  - Penalized regression splines

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
Nadaraya-Watson
local 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

some penalty

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

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

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

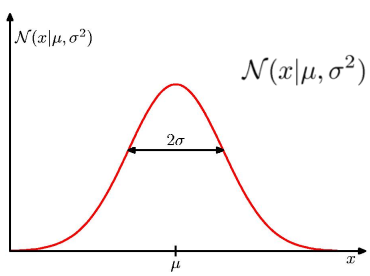
- Now, let's further assume a distributional form and log-likelihood loss

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

First, recall some facts about Gaussians...

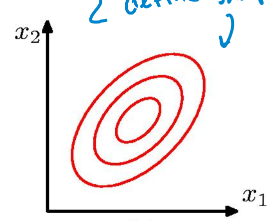
Quick Review of Gaussians

- 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 define shape

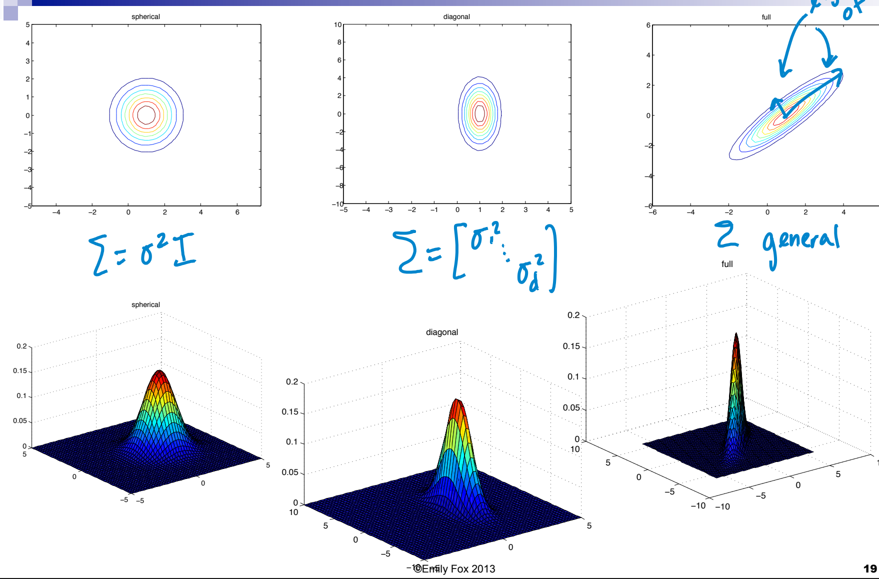


$\Sigma = cov(x)$

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

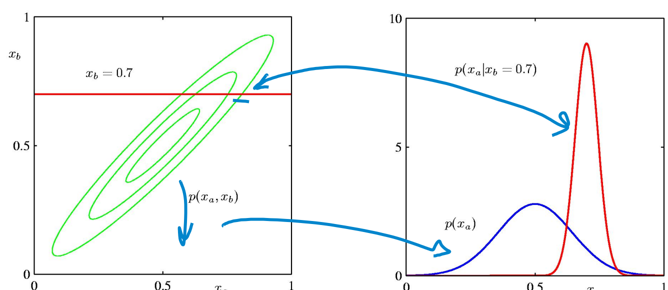
Two-Dimensional Gaussians

eigvecs of Σ



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

Conditional & Marginal Distributions



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

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

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

- Taking the log

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

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

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

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

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

$$H = \begin{pmatrix} h_1(x_1) & \cdots & h_m(x_1) \\ \vdots & \ddots & \vdots \\ h_1(x_n) & \cdots & h_m(x_n) \end{pmatrix}$$

Look at penalized regression

A Bayesian Formulation of Ridge

- Consider a model with likelihood

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

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

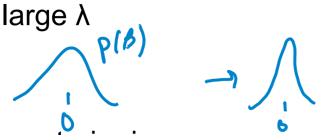
and prior

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

prior places penalty

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

- For large λ



prior is peaked around $\beta = 0$

$\Leftrightarrow$ penalizing β far from 0

- The posterior is

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

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:

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

- 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 $\propto$ likelihood $\times$ prior

where

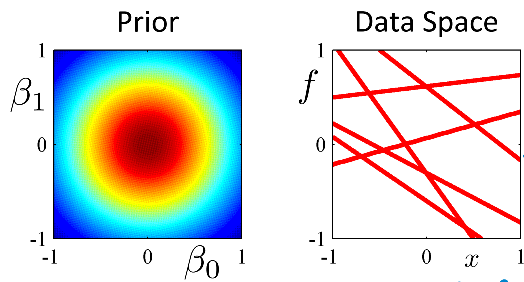
$$\mu_n = \Sigma_n (\Sigma_0^{-1} \mu_0 + \sigma^{-2} H^T y)$$

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

Example: Standard Linear Basis

0 data points observed

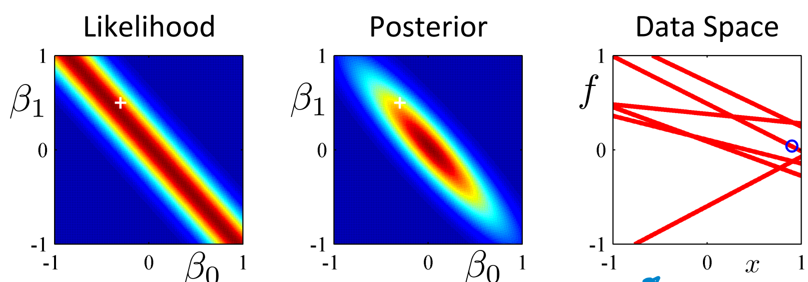
$y = \beta_0 + \beta_1 x + \epsilon$



draw β_0, β_1 from prior and set $f = \beta_0 + \beta_1 x$

Example: Standard Linear Basis

1 data point observed

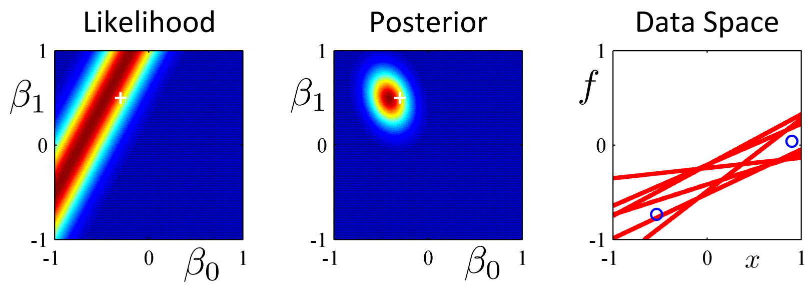


post. $\propto$ like. $\times$ prior

draws of $\underline{\beta}$ from posterior

Example: Standard Linear Basis

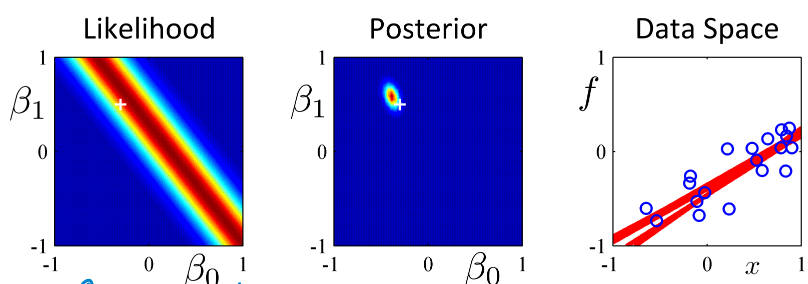
2 data points observed



posterior
concentrates

Example: Standard Linear Basis

20 data points observed



$p(y|\beta, x, \sigma^2)$
$= N(y|\beta^T x, \sigma^2)$

Predictive Distribution

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

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

x^*, x

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

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

$\beta^T h(x)$

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

$\beta \overset{|y}{\sim} N(\mu_n, \Sigma_n)$

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

fcn of obs. locations x

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

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

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

var of our params β

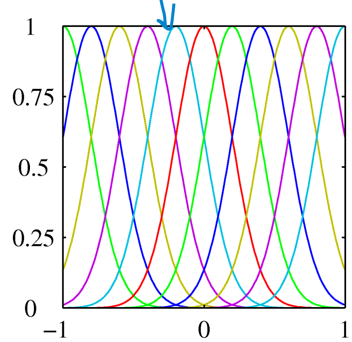
var of obs

Example: Gaussian Basis Expansion

- Gaussian basis functions:

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

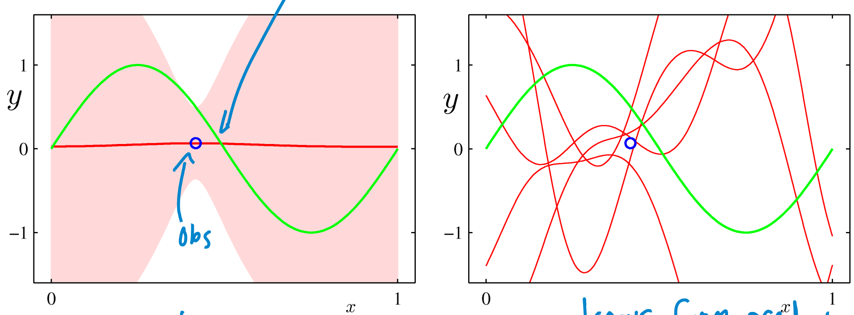
- These are local;
  a small change in x
  only affects nearby
  basis functions.
  Parameters control
  location and scale (width)



Example: Gaussian Basis Expansion

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

true fcn

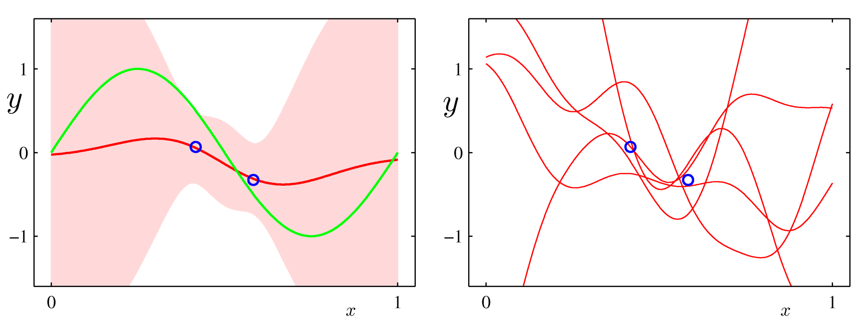


pred. dist. shows pointwise uncertainty

draws from pred. dist.

Example: Gaussian Basis Expansion

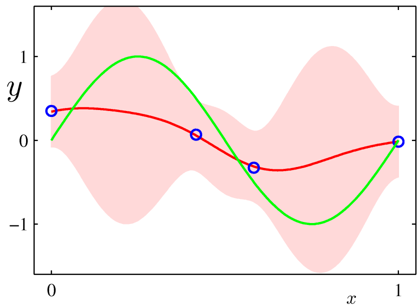
- Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points

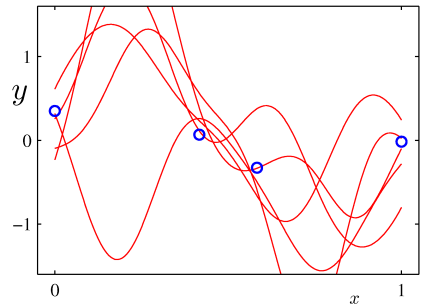


less uncertainty near obs.

Example: Gaussian Basis Expansion

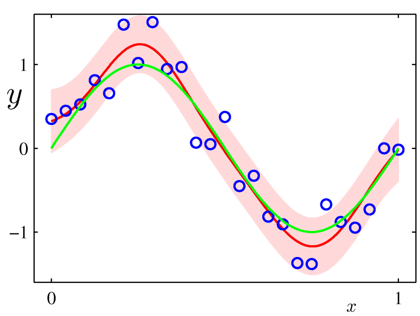
- Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points

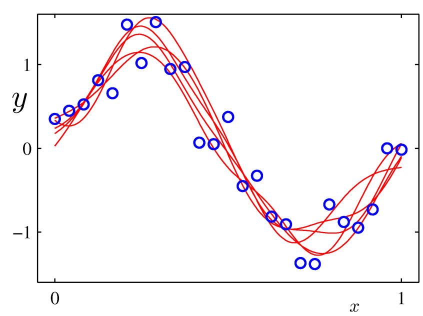




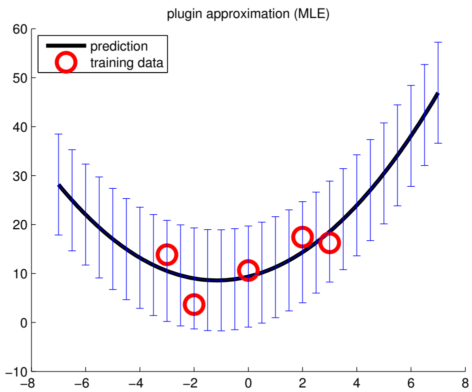
Example: Gaussian Basis Expansion

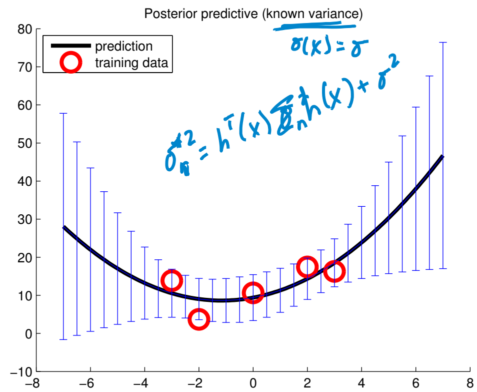
- Example: Sinusoidal data, 9 Gaussian basis functions, 25 data points





Estimation vs. Predictive Distributions





$p(y^* \mid x^*, X, y, \sigma^2)$

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

$= N(y \mid \hat{\beta}^T h(x), \sigma^2)$ ← constant error bars

What # of basis fcns should we use?

Bayesian Model Selection

- Assume some *M* possible models
  - Model M_m $m=1,\dots,M$ has parameters θ_m and prior $p(\theta_m \mid M_m)$
  - Prior over models $p(M_m)$
- Model posterior

training data

"marginal likelihood"

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

- Compare models:

post. odds

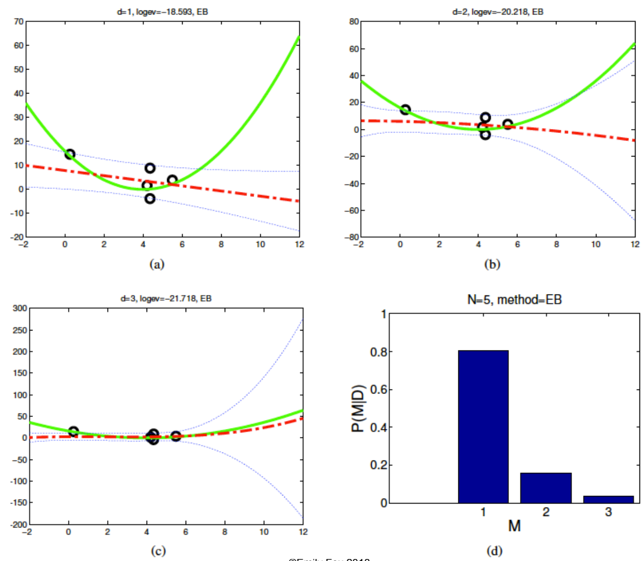
$$\frac{p(M_m \mid Z)}{p(M_\ell \mid Z)} = \frac{p(M_m)}{p(M_\ell)} \boxed{\frac{p(Z \mid M_m)}{p(Z \mid M_\ell)}} \overset{m}{\underset{\ell}{\gtrless}} 1$$

often, uniform prior

Bayes factor

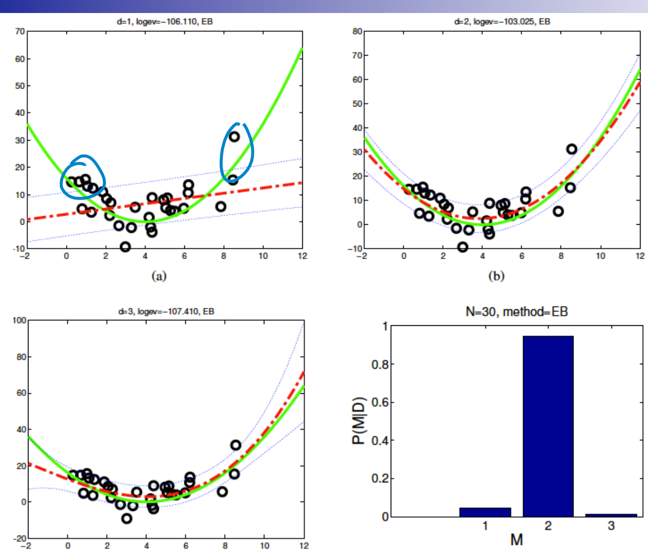
BMS Example (n=5)

uniform prior on models



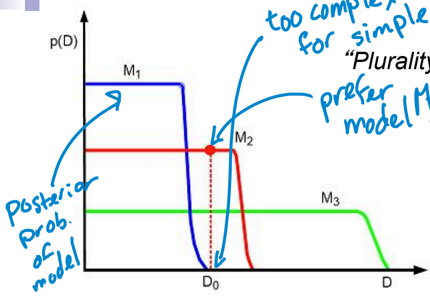
high posterior prob. on simple models

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

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

Announcements

- Upcoming changes…
- Lectures:
  - Instead of lecture next Tuesday, Shirley will provide an examples section
  - Instead of recitation on Tuesday May 9, I will do a lecture on nonparametrics for generalized linear models (GLM)
- Homeworks:
  - Starting this Thursday, homeworks will be 2 weeks long
  - Provides extra flexibility on timing to accommodate project
  - Each homework (HW4 and HW5) will count the same as two 1-wk assignments
  - Should be slightly shorter than two 1-wk assignments