Module 2: Splines and Kernel Methods

Inference for Linear Smoothers

STAT/BIOSTAT 527, University of Washington

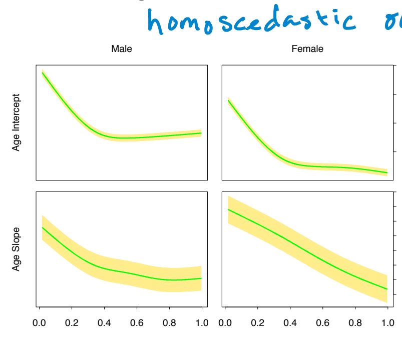
Emily Fox

April 23rd, 2013

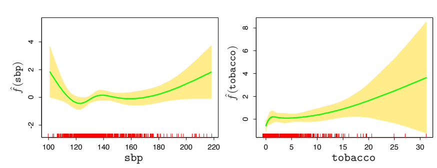
Confidence Bands

- 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***

homoscedastic $\sigma(x)=\sigma$



heterosedastic $\sigma(x)$



From Hastie, Tibshirani, Friedman book

CIs for Linear Smoothers

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

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

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

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

- 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

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

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

  - 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

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

Variance Estimation

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

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

  - 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

    $$Y - \hat{f} =$$

    and

    $$E[Y^T Q Y] = \mathrm{tr}(QV) + \mu^T Q \mu$$
  - Then,

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

$$E[\hat{\sigma}^2] =$$

  - 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:

$$y_{i+1} - y_i =$$

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

- Estimator will be inflated
- 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 $y_i = f(x_i) + \sigma(x_i)\epsilon_i$
  - Define

    $$Z_i = \log(y_i - f(x_i))^2 \quad \delta_i = \log \epsilon_i^2$$
  - Then,
- 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 $\log \sigma^2(x)$

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

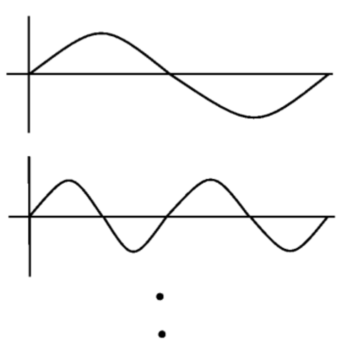
$$f(x) = E[Y \mid x]$$

- How to estimate from finite training set?

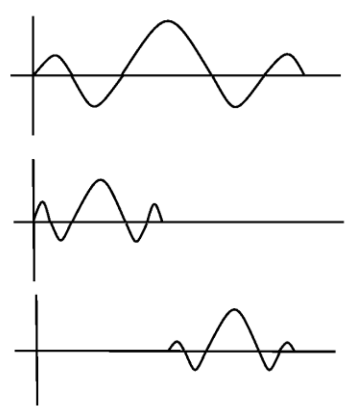
Restrict to model class

- Example = linear basis expansion
  - Standard linear
  - Polynomial
  - Splines
  - …

Other Important Basis Expansions



Fourier Basis



Wavelet Basis

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

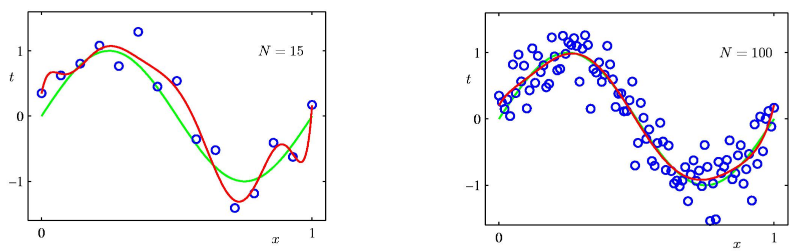
- Example = linear basis expansion

Overfitting as model complexity grows

- Penalized linear basis expansions
  - Ridge
  - Lasso
  - Smoothing splines
  - Penalized regression splines

Overfitting

9th Order Polynomial



$n = 15$ $\qquad$ $n = 100$

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

Again: Linear Basis Expansion

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

- ***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

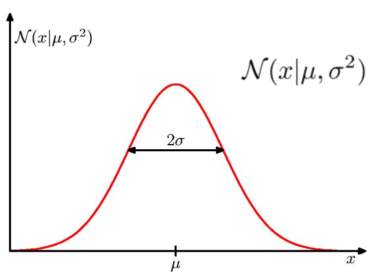
- So far, our focus has been on L_2 loss:

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

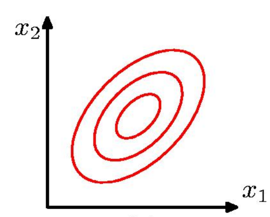
- Here, we assumed $y = f(x) + \epsilon$ with
- Now, let's assume a distributional form and log-likelihood loss

Quick Review of Gaussians

- Univariate and multivariate Gaussians

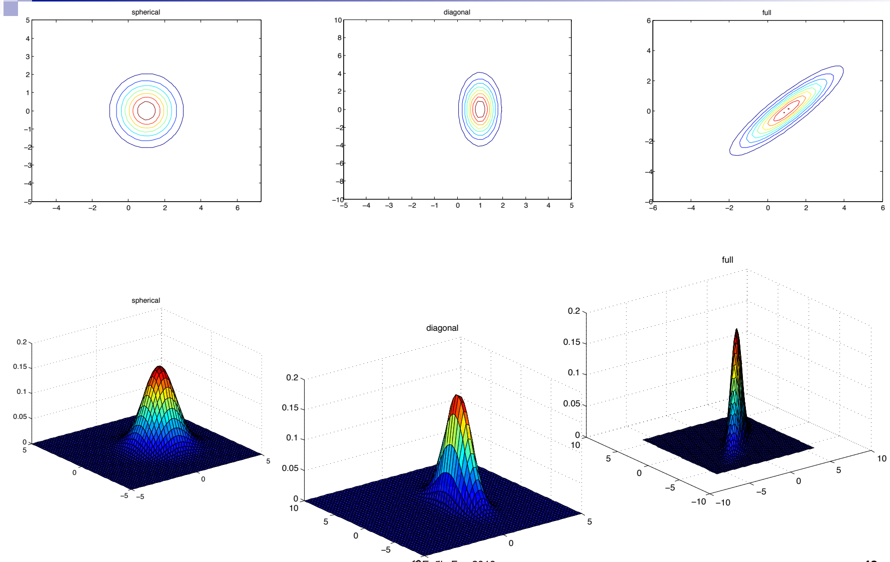


$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

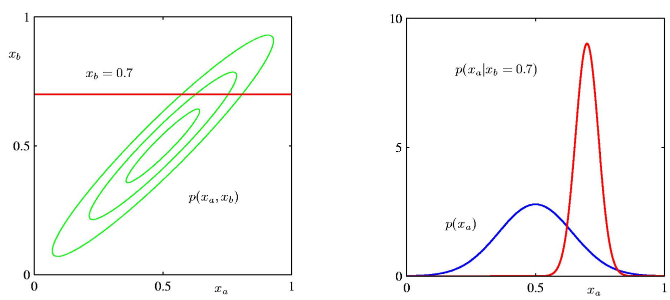


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Two-Dimensional Gaussians



Conditional & Marginal Distributions



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)

$$p(y \mid X, \beta, \sigma^2) =$$

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

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

- Equivalent objective to RSS *(Gaussian log-like loss = L_2 loss)*
- Taking the gradient and setting to zero, we have already shown

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

A Bayesian Formulation of Ridge

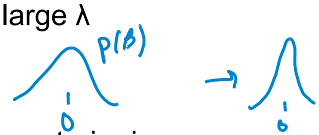
- Consider a model with likelihood　　If $\epsilon \sim N(0,\sigma^2)$

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

and prior

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

- For large λ



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

- The posterior is

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

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

Bayesian Linear Regression

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

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

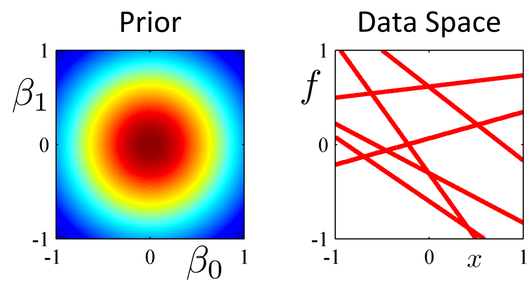
- Combining this with the Gaussian likelihood function, and using standard Gaussian identities, gives posterior

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

where

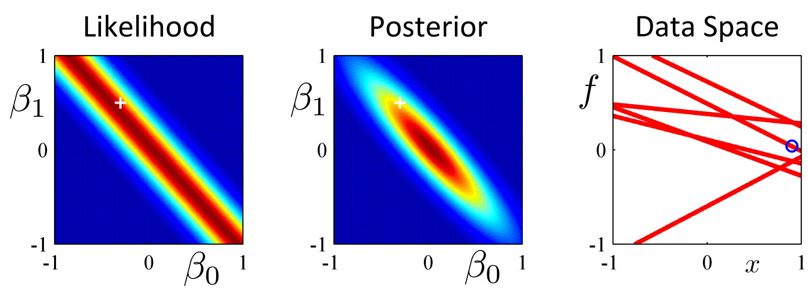
Example: Standard Linear Basis

0 data points observed



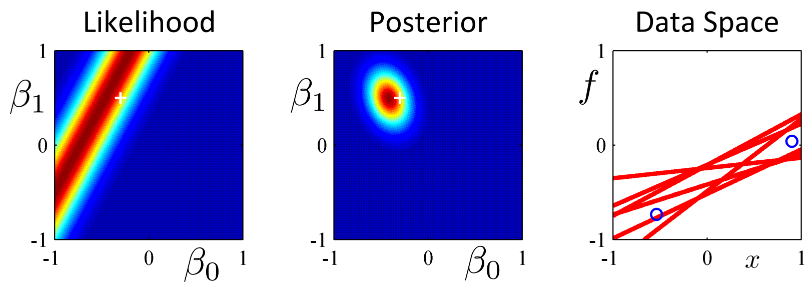
Example: Standard Linear Basis

1 data point observed



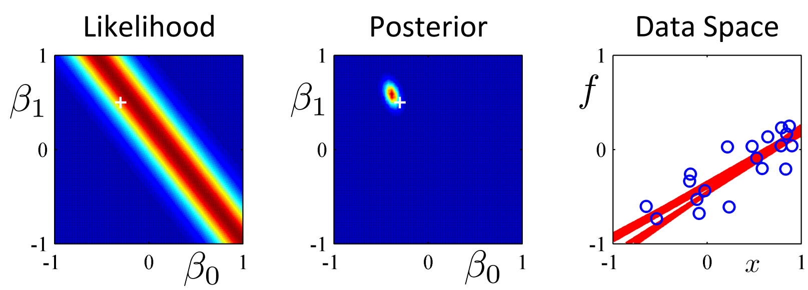
Example: Standard Linear Basis

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

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

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

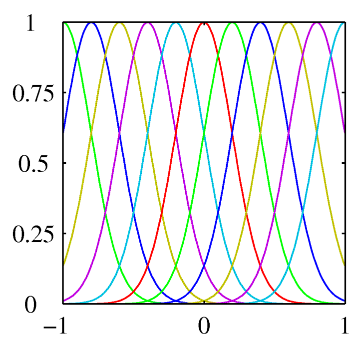
$$p(y \mid x, \beta, \sigma^2) = N(y \mid f(x), \sigma^2)$$

Example: Gaussian Basis Expansion

- Gaussian basis functions:

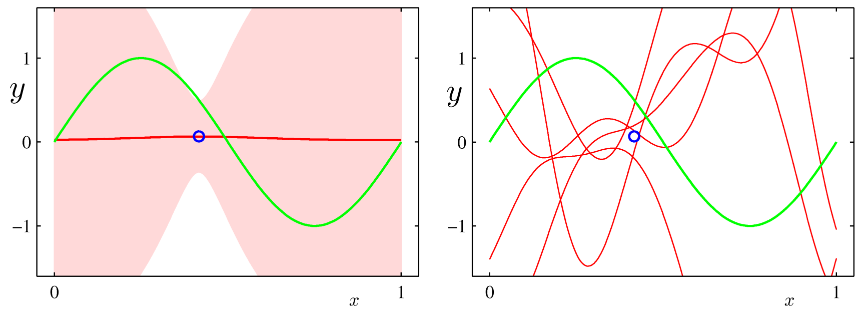
$$h_j(x) = \exp\left\{ -\frac{(x-\mu_j)^2}{2s^2} \right\}$$

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



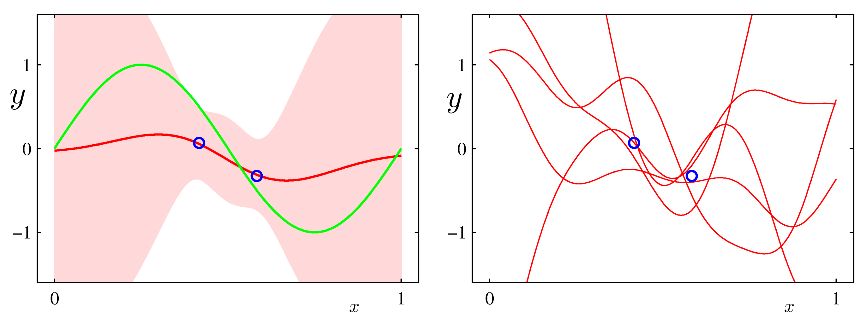
Example: Gaussian Basis Expansion

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



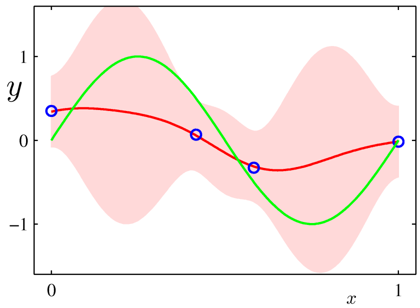
Example: Gaussian Basis Expansion

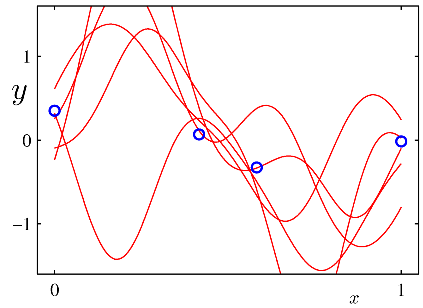
- Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points



Example: Gaussian Basis Expansion

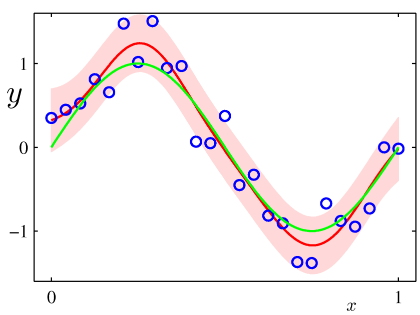
- Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points

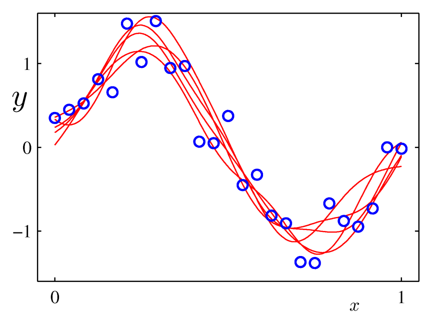




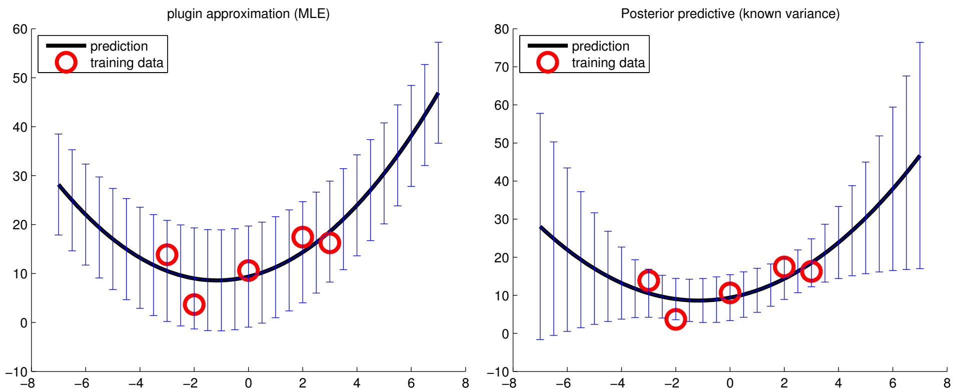
Example: Gaussian Basis Expansion

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





Estimation vs. Predictive Distributions



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

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

- Compare models:

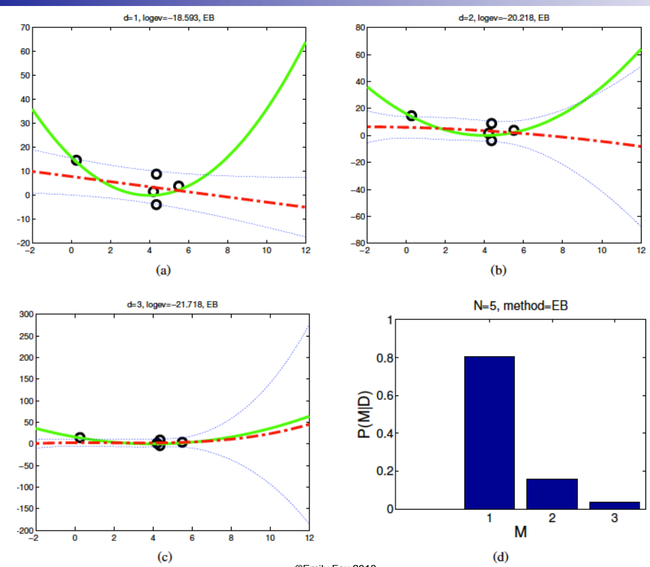
post. odds

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

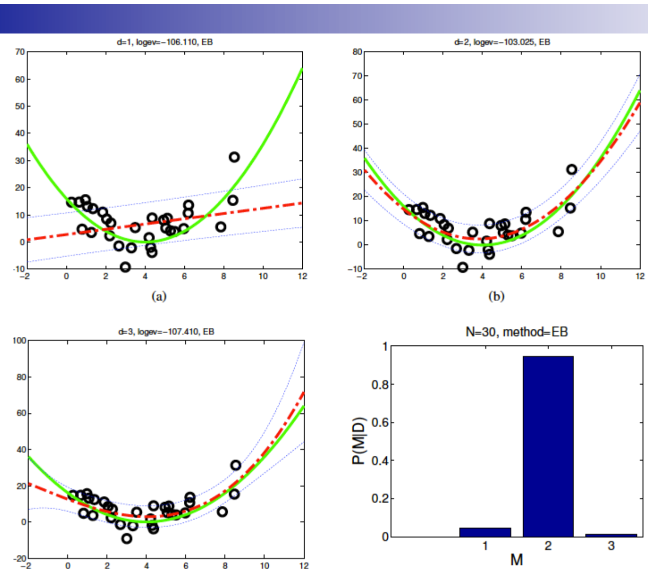
often, uniform prior

Bayes factor

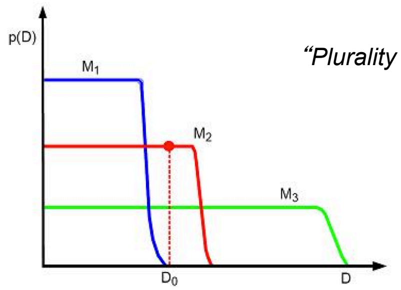
BMS Example (n=5)



BMS Example (n=30)



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

Going Infinite…

Change of notation:

$$h(x) \to \phi(x)$$

- Nonparametric Gaussian regression:
  Would like to let the number of "features" $M \rightarrow \infty$
- *Prior:* $p(\beta \mid 0, \alpha^{-1} I_M)$
- *Predictions:* $f = \Phi\beta$
- Gaussian process models replace explicit basis function representation with a direct specification in terms of a ***positive definite kernel function***

Mercer Kernel Functions

- Predictions are of the form

$$p(f) = N(f \mid 0, \alpha^{-1}\Phi\Phi^T)$$

  where the ***Gram matrix*** *K* is defined as

$$K_{ij} =$$

- *K* is a ***Mercer kernel*** if the Gram matrix is positive definite for any *n* and any $x_1, \ldots, x_n$

Mercer's Theorem

- If *K* is positive definite, we can compute the eigendecomp:
- Then $K_{ij} =$
- Define $\phi(x) = \Lambda^{\frac{1}{2}} U_{\cdot i}$ so that

$$K_{ij} =$$

- If a kernel is Mercer, there exists a function $\phi : \mathcal{X} \to \mathbb{R}^d$ s.t.

Example Mercer Kernels

- Example #1: (non-stationary) ***polynomial kernel***

$$\kappa(x,x') = (\gamma x^T x' + r)^M$$

- For M=2, $\gamma = r = 1$,

$$(1 + x^T x')^2 = (1 + x_1 x_1' + x_2 x_2')^2$$

- This can be written as $\phi(x)^T \phi(x')$, with

$$\phi(x) =$$

  - Equivalent to working in a 6-dimensional feature space
  - For general *M,* basis contains all terms up to degree *M*
- Example #2: ***Gaussian kernel***

$$\kappa(x,x') = \exp\left(-\frac{1}{2}(x - x')^T \Sigma^{-1} (x - x')\right)$$

  - Feature map lives in an infinite-dimensional space

Gaussian Processes

- Dispense of parametric view (prior on β) and consider prior on functions themselves (prior on *f)*
- Seems hard, but we have shown that it is feasible when we look at a finite set of values $x_1, \ldots, x_n$

$$p(f) = N(f \mid 0, K)$$

- Defined by a *Mercer kernel*
- More generally, a ***Gaussian process*** provides a distribution over functions

Gaussian Processes

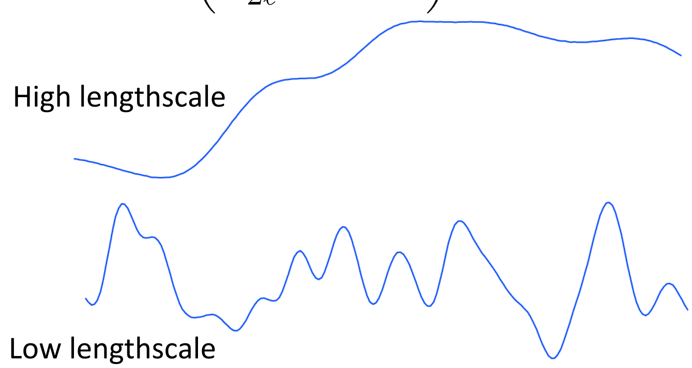
- Distribution on functions
  - $f \sim GP(m, \kappa)$
    - m: mean function
    - κ: covariance function

$$\Updownarrow$$

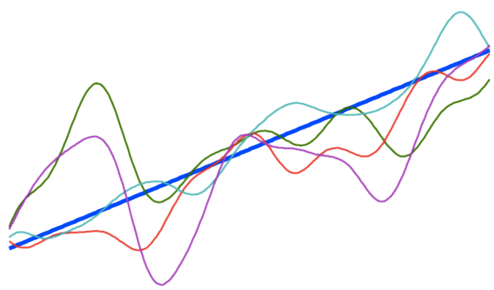
  - $p(f(x_1), \ldots, f(x_n)) \sim N_n(\mu, K)$
    - $\mu = [m(x_1), \ldots, m(x_n)]$
    - $K_{ij} = \kappa(x_i, x_j)$
- Idea: If x_i, x_j are similar according to the kernel, then $f(x_i)$ is similar to $f(x_j)$

κ: covariance function

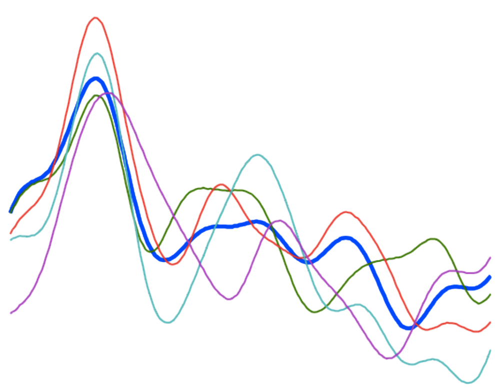
$$\kappa(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}(x - x')^2\right)$$



m: mean function

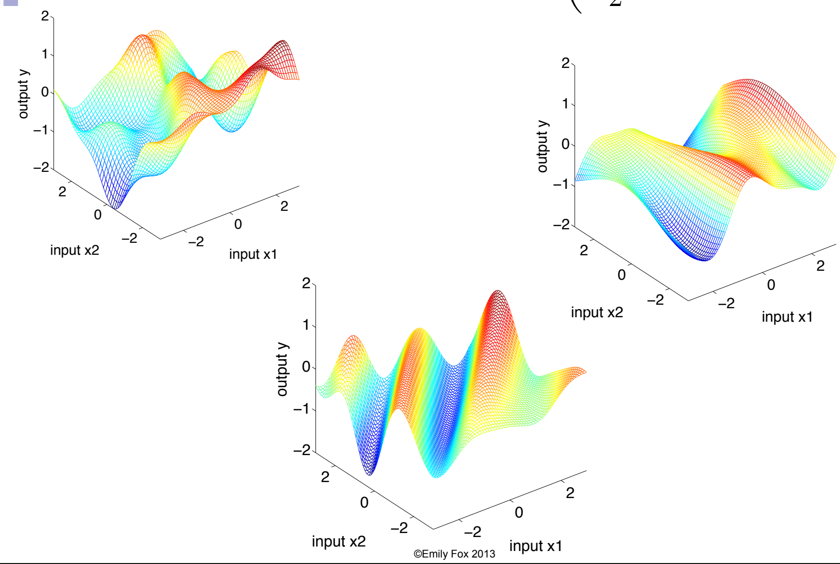


m: mean function



2D Gaussian Processes

$$\kappa(x_p, x_q') = \sigma_f^2 \exp\left(-\frac{1}{2}(x_p - x_q')^T M (x_p - x_q')\right)$$



GPs for Regression

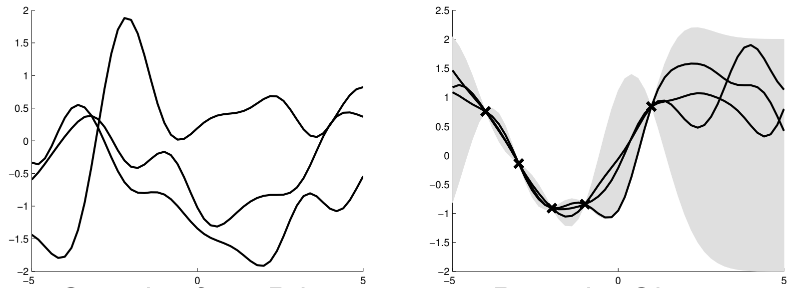
- Start with noise-free scenario: directly observe the function
- Training data $\mathcal{D} = \{(x_i, f_i), i = 1, \ldots, n\}$
- Test data locations X^* → predict *f**
- Jointly, we have

$$\begin{pmatrix} f \\ f^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu_* \end{pmatrix}, \begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix}\right)$$

- Therefore,

$$p(f^* \mid X^*, X, f) =$$

1D Noise-Free Example



Samples from Prior $\kappa(x, x') = \sigma_f^2 \exp(-\frac{1}{2\ell^2}(x - x')^2)$

Posterior Given 5 Noise-Free Observations

- Interpolator, where uncertainty increases with distance
- Useful as a computationally cheap proxy for a complex simulator
  - Examine effect of simulator params on GP predictions instead of doing expensive runs of the simulator

GPs for Regression

- Noisy scenario: observe a noisy version of underlying function

$$y = f(x) + \epsilon \quad \epsilon \sim N(0, \sigma_y^2)$$

  - Not required to interpolate, just come “close” to observed data

$$\mathrm{cov}(y|X) =$$

- Training data $\mathcal{D} = \{(x_i, y_i), i = 1, \dots, n\}$
- Test data locations X^* → predict *f**
- Jointly, we have $\begin{pmatrix} y \\ f^* \end{pmatrix} \sim N\left(0, \begin{pmatrix} K_y & K_* \\ K_*^T & K_{**} \end{pmatrix}\right)$
- Therefore, $p(f^* \mid X^*, X, y) =$

GPs for Regression

$$p(f^* \mid X^*, X, y) = N(K_*^T K_y^{-1} y, K_{**} - K_*^T K_y^{-1} K_*)$$

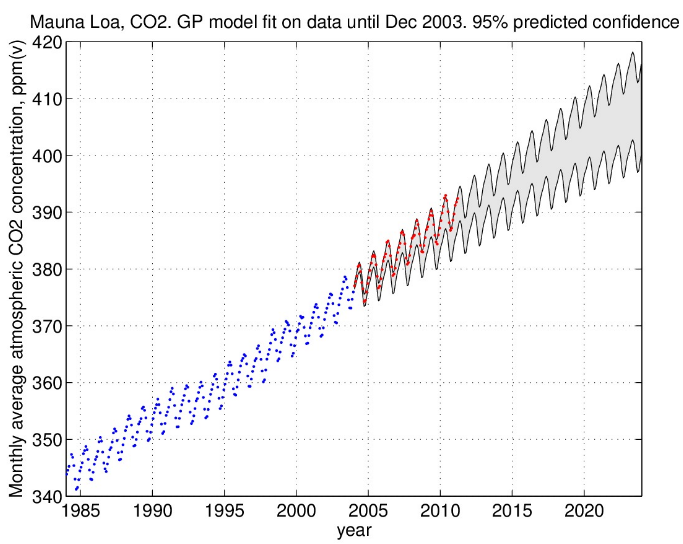
- For a single point *x**

$$p(f^* \mid X^*, X, y) = N(k_*^T K_y^{-1} y, k_{**} - k_*^T K_y^{-1} k_*)$$

so

$$\bar{f}^* = k_*^T K_y^{-1} y =$$

CO2 Concentration Over Time



Mauna Loa Observatory in Hawaii, analyzed by Rasmussen & Williams 2006

Mixing Kernels for CO2 GP Analysis

Smooth global trend

$$\kappa_1(x, x') = \theta_1^2 \exp\left(-\frac{(x - x')^2}{2\theta_2^2}\right)$$

Seasonal periodicity

$$\kappa_2(x, x') = \theta_3^2 \exp\left(-\frac{(x - x')^2}{2\theta_4^2} - \frac{2\sin^2(\pi(x - x'))}{\theta_5^2}\right)$$

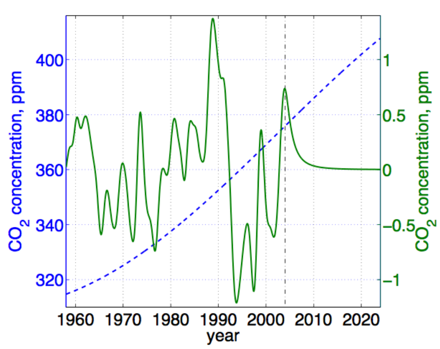
Medium term irregularities

$$\kappa_3(x, x') = \theta_6^2 \left(1 + \frac{(x - x')^2}{2\theta_8\theta_7^2}\right)^{-\theta_8}$$

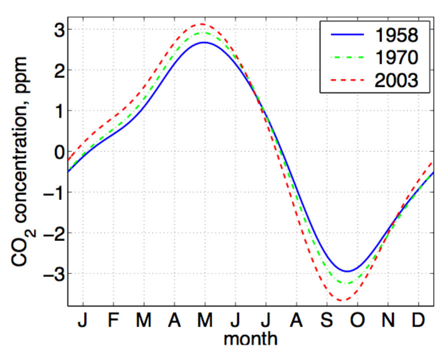
Correlated Observation Noise

$$\kappa_4(x_p, x_q) = \theta_9^2 \exp\left(-\frac{(x_p - x_q)^2}{2\theta_{10}^2}\right) + \theta_{11}^2\delta_{pq}$$

CO2 Concentration Over Time



(a)



(b)

Mauna Loa Observatory in Hawaii, analyzed by Rasmussen & Williams 2006

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