Lecture 10, 12.S990

Regression Machines

$$\text{DATA } \{\underset{\substack{\downarrow\\ \text{Predictor}}}{\underline{u_i}}\,,\ \underset{\substack{\downarrow\\ \text{Predictands}}}{\underline{y_i}}\} \xrightarrow{Model} y \approx f(x)$$

Rich Literature

Examples:

Recall:
$$\underset{\substack{\uparrow\\ \text{Output}}}{\underline{Y}} = \underset{\substack{\uparrow\\ \text{Det.}}}{\underline{\underline{U}}}\ \underset{\substack{\uparrow\\ \text{Stochas.}}}{\underline{X}} \longleftarrow \begin{array}{l} ROM \\ POD \longrightarrow \text{Eigen basis} \\ \left.\begin{array}{l} gPC \\ RSM \end{array}\right\} \rightarrow \text{Polynomial Chaos} \end{array}$$

EnKF:
$$\underline{\underline{A}}^a = \underline{\underline{A}}^f\ \underline{\underline{\mathscr{X}}} \longrightarrow \text{Update Equation}$$

AR-model etc.

You know the basic idea:

$$y^{(j)} \underset{\substack{\text{We are}\\ \text{modeling}}}{\overset{=}{\downarrow}} \sum_i x_i^{(j)}\beta_i + \underline{\eta}^{(j)}$$

$$\underbrace{y^{(j)}}_{\substack{\downarrow\\ \text{R.V.}}} \approx \sum_i x_i^{(j)}\beta_i + \underbrace{\underline{\eta}^{(j)}}_{\substack{\downarrow\\ \text{Homosedastic}\\ \underline{\eta}^{(j)} = \underline{\eta}}}$$

where $x_i^{(j)}$ can be R.V. too, but generally not considered as such.

GLM

$$g\Big(E[y^{(j)}]\Big) \approx \sum_i x_i^{(j)}\beta_i$$

$$h^{(j)} \doteq \sum_i x_i^{(j)}\beta_i$$

$$g(\mu^{(j)}) = h^{(j)}$$

$g(\mu^{(j)})$ is a link function

$\mu^{(j)}$ is mean

h is a canonical variable

$$\begin{aligned} Y &\longrightarrow \text{Exponential family} \\ g &\longrightarrow \text{Somewhat arbitrary but a few hints} \end{aligned}$$

1. $\mu = g^{-1}(h)$, so a nicely invertible g.
2. g^{-1} maps $\underline{\underline{x}}^T \; \underline{\beta}$ into admissible ranges for μ.

Why bother?

$$Y^{(j)} = \sum_{i=1}^{n} x_i^{(j)} \beta_i + \eta$$

Ordinary linear regression

$$E\left[Y^{(j)} | x_1^{(j)} \ldots x_n^{(j)}\right]$$

If/for:

$$Y \sim N(...)$$

$$\bar{Y} + \tilde{Y} = \sum_{i=1}^{n} x_i \beta_i + \sum_{i=1}^{n} \delta x_i \beta_i$$

Essentially vary around a central value. GLM extends it to a range of distributions - the Exponential Family.

Link Functions.

Y	**g**
Poisson	$\ln(\mu) = \underline{x}^T \underline{\beta};$
	$\underset{\underset{\Re}{\uparrow}}{\mu} = \exp\left(\ \underset{\underset{\{0,1,..\}}{\uparrow}}{\underline{x}^T}\ \ \underline{\beta}\right)$
Categorical	Logit
$x \in \begin{cases} 1 - P_1 \\ \vdots \\ k - P_k \end{cases}$	$\ln\left(\frac{\mu}{1-\mu}\right) = \underline{x}^T \underline{\beta}$
	$\mu = \frac{1}{1+e^{-\underline{x}^T \underline{\beta}}}$
Many others	

How to solve?

Recall (EM for Exponential Family):

$\underline{IWLS}$

$$
\begin{aligned}
h_i^{(t)} &= \underline{x}_i^T \beta^{(t-1)} \\
\mu_i^{(t)} &= g^{-1}(h_i^{(t)}) \\
Z_i^{(t)} &= h_i^{(t)} + (y_i - \mu_i^{(t)}) \left[\frac{\partial g}{\partial \mu_i}\right]^{(t)} \\
w_i^{(t)} &= \left[\left(\frac{\partial g}{\partial \mu_i}\right)^2 \mathcal{O}(\mu_i)\right]^{-1}
\end{aligned}
$$

$V(\mu_i) \equiv$ Variance function

Binomial	$\mu(1-\mu)$
Gamma	μ^2
Normal	1
Poisson	μ
etc	

$w_i^{(t)}\left[Z_i^{(t)} - x_i^{(t)}\beta^{(t)}\right]^2 \longleftarrow$ Weighted Least Squares

Repeat.

Generalized Additive Model (GAM)

$$\text{GLM:} \qquad g\big(E[y^{(i)}]\big) = \underline{x}^T \underline{\beta}$$

$$\text{GAM:} \qquad g\big(E[y^{(i)}]\big) = f_0 + \sum_{i=1}^{P} \underbrace{f_i(x_i^{(j)})}_{\text{Covariates}}$$

$$\text{and} \qquad E\Big[f_i(x_i)\Big] = 0 \;\forall\, i$$

Example:

$$g(\mu) = a_0 + \underbrace{xa_1 + a_2x^2 + a_3x^3}_{f(x)}$$

Polynomials

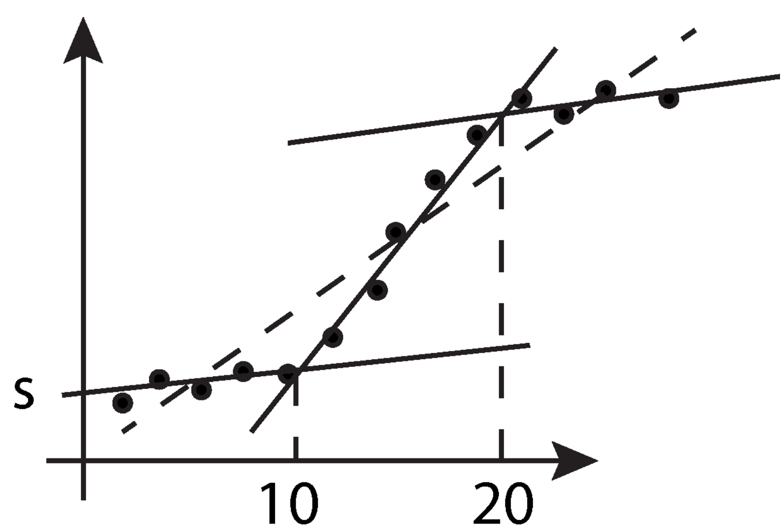
Splines...

Spline Reconstruction:

$$\sum_j \left[y_i^{(j)} - \sum_i f_i(x_i^{(j)}) - f_0 \right]^2 + \sum_i \lambda_i \underbrace{\int f_i''(x_i)^2 dx_i}_{\text{Require Smoothness}}$$

Also → Look up MARS

MARS



$$y = S + a_1 \max(0, x - 10) + a_2 \max(0, x - 20)$$

$$f(x) = \sum_{i=1}^{K} \underset{\substack{\uparrow \\ \text{weights}}}{c_i} \; \underset{\substack{\uparrow \\ \text{Basis}}}{B_i(x)}$$

$\mathbf{B_i}$:

1. Constant function
2. Hinge function at knots
3. Product of hinge functions

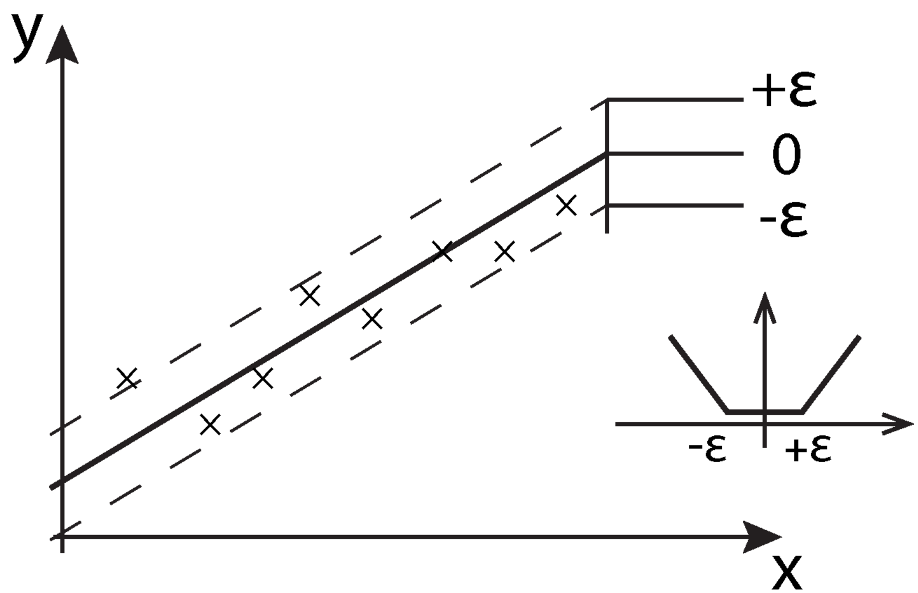
How it works:

- Start with intercept (mean of y_i):
- Then: (Forward Pass):
  - A pair of basis function that gives maximum reduction in fit $\max(0, x - c)$; $\max(0, c - x)$.
  - New basis function:
    - Has all "Parent" (previous) basis
    - Requires additional search through variables and values

Backward Pass:

- Forward pass $\rightarrow$ overfits
- Backward pass $\rightarrow$ (pruning): Eliminate terms "one by one" and pick best model from pruning.

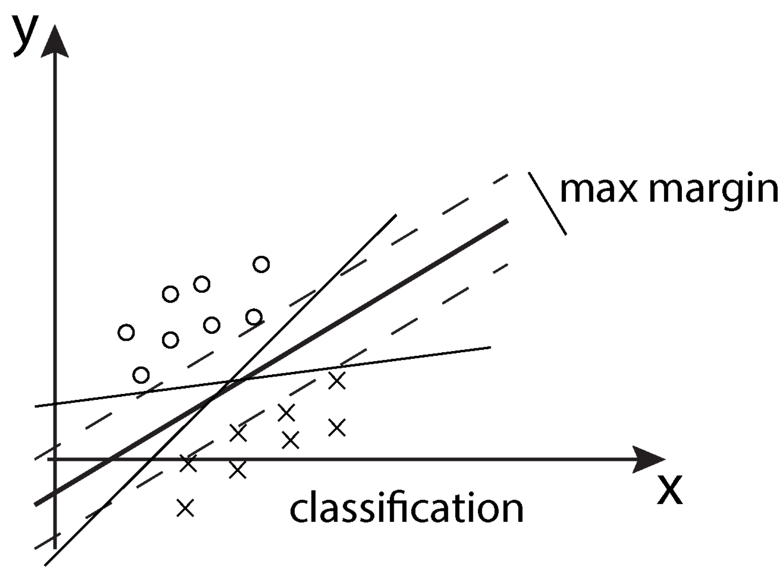
Regression by support vectors:



$$y = \omega x + b$$

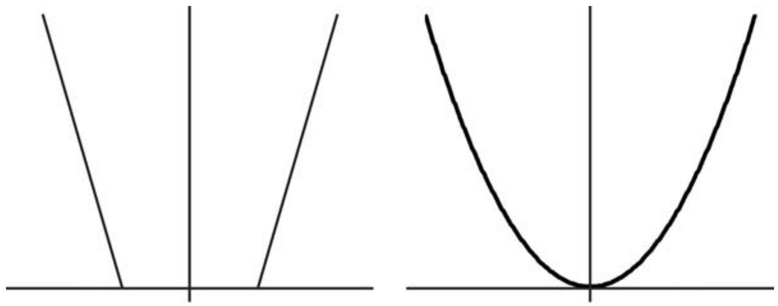
$$\alpha \equiv \begin{cases} 0, & |y_i - f(x_i)| \leq \epsilon \\ |y_i - f(x_i)| - \epsilon & \text{otherwise} \end{cases}$$

Some "intuition:"



Original:	$y = \omega x + b$
Mini:	$\frac{1}{2}\|\|\omega\|\|^2,$ such that
	$\|y_i - \omega x_i - b\| \leq \epsilon$
With soft margin:	$\frac{1}{2}\|\|\omega\|\|^2,$ such that
	$\begin{cases} y_i - \omega x_i - b \leq \epsilon + \xi_i \\ \omega x_i + b - y_i \leq \epsilon + \xi_i^* \end{cases}$

Soft Margins are a way to relax constraints:



Other: Huber etc.

$$
\begin{aligned}
L = \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \\
- \sum_{i=1}^{n}\alpha_i\left[\epsilon + \xi_i - y_i + \omega x_i + b\right] \\
- \sum_{i=1}^{n}\alpha_i^*\left[\epsilon + \xi_i^* + y_i - \omega x_i - b\right] \\
- \sum_{i=1}^{n}\left(\eta_i\xi_i + \eta_i^*\xi_i^*\right)
\end{aligned}
$$

$$\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$$

$$\begin{aligned}
\frac{\partial L}{\partial b} &= \quad \sum_{i=1}^{n} \left(\alpha_i^* - \alpha_i\right) = 0 \\
\frac{\partial L}{\partial \omega} &= \quad \omega - \sum_{i=1}^{n} \left(\alpha_i^* - \alpha_i\right) x_i = 0 \\
&\text{So:} \\
y_i &= \quad \underbrace{\Big(\sum_j \left(\alpha_j^* - \alpha_j\right)\Big) x_i}_{\text{Support vector Expansion}} + b
\end{aligned}$$

To Calculate b:

Karush-Kuhn-Tucker(KKT) Conditions:

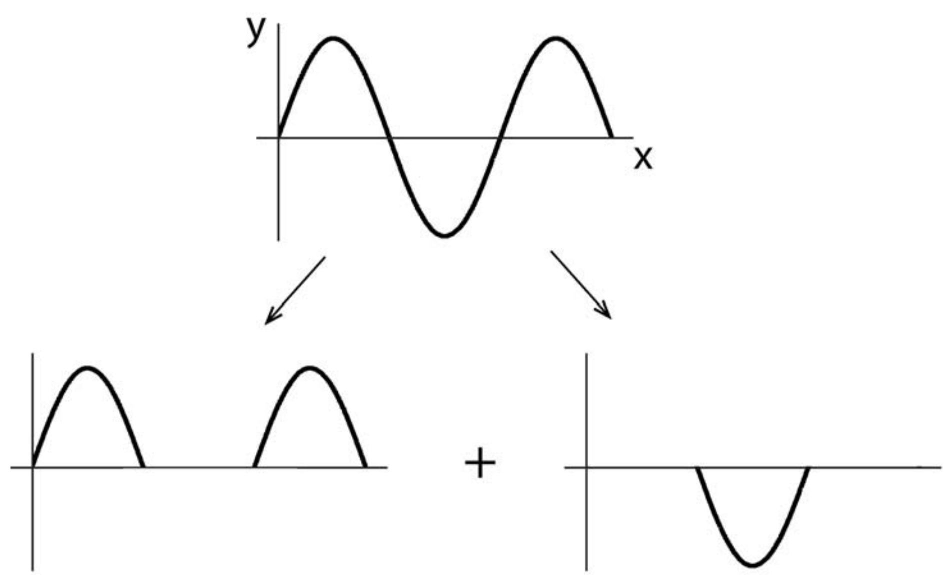
$$\begin{gathered}
(C - \alpha_i)\xi_i = 0 \\
(C - \alpha_i^*)\xi_i^* = 0 \\
\alpha_i(\epsilon + \xi_i - y_i + \omega x + b) = 0 \\
\alpha_i^*(\epsilon + \xi_i^* + y_i - \omega x - b) = 0
\end{gathered}$$

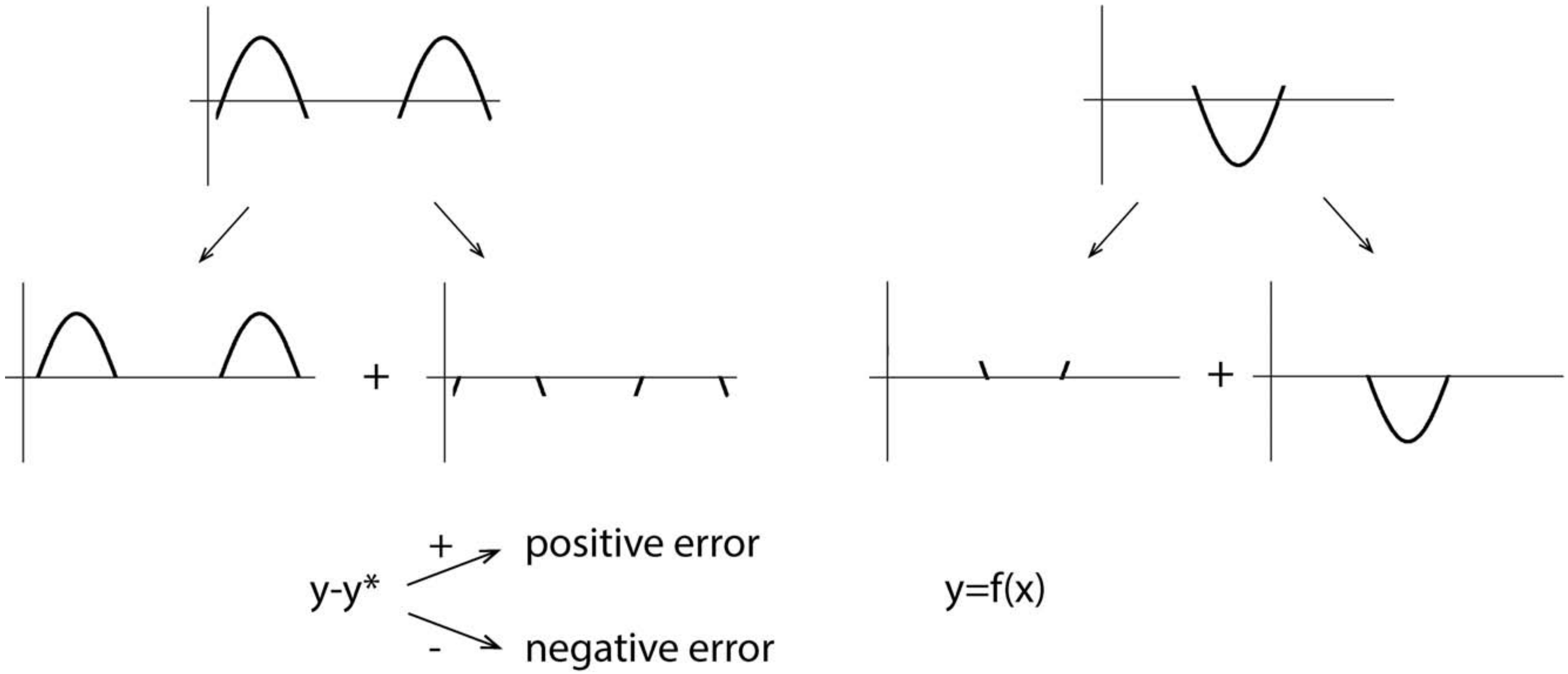
Note:
Only samples (x_i, y_i) with $\alpha_i^* = C$ lie outside $\epsilon-$ insensitive region.

$$\alpha_i \alpha_i^* = 0 \quad \rightarrow \text{Satisfy KKT} \rightarrow \text{Support vectors.}$$

This implies that vectors that satisfy KKT condition are the support vectors.

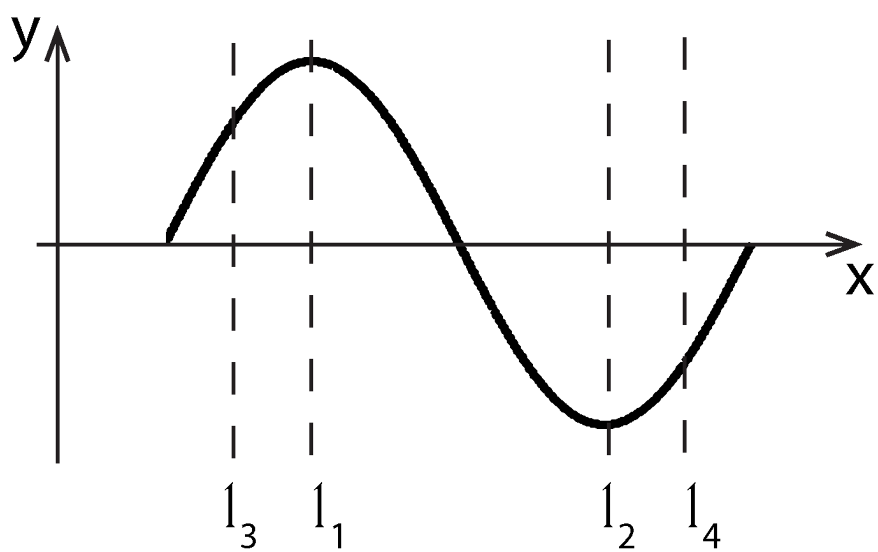
Regression Traces





Why does this not work?

The splits must come from the feature:



- $x < \mathrm{l}_1$
  - $x < \mathrm{l}_3$
    - $\mathrm{m}_1,\mathrm{b}_1$
  - $x > \mathrm{l}_2$
    - $\mathrm{m}_2,\mathrm{b}_2$
    - $x > \mathrm{l}_4$
      - $\mathrm{m}_3,\mathrm{l}_3$
      - $\mathrm{m}_4,\mathrm{b}_4$

Mechanism/Method

a. Consider all $\underset{\text{Response}}{\underset{\downarrow}{\textit{values}}}$ of all $\underset{\text{Covariate}}{\underset{\downarrow}{\textit{features}}}$ in Set.

b. Pick a $\underline{\text{feature}}$ and $\underline{\text{value}}$ that splits data into two, such that the total variance of Splits is reduced the most.

c. Continue till some termination criterion.

Mechanisms for Regression

Machine Learning		
	Divide and Con.:	Regression Trees
	Margin Maximization:	Support Vectors
$\Downarrow$	Smoothness:	Spline model, MARS
	Randomness:	GLM and GAM
	Nonlinearity:	Kernel Machines
Statistics		

SLR $\in$ MLR $\in$ GLM $\in$ GAM.

Some Limitations

1. Overfitting (We saw this is density estimation).
2. Greedy algorithms - local convergence.

How to fix?

Bagging and Boosting

$\Downarrow$

Randomization + Aggregation

Bagging $\Rightarrow$ Bootstrap Aggregating

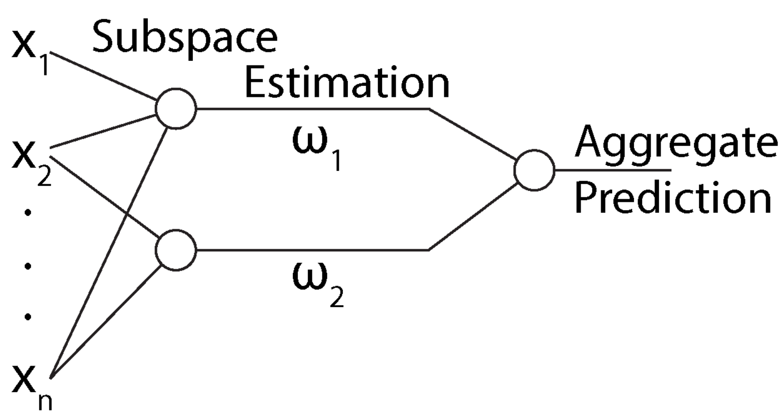
Bagging (Breiman)

a. Generate Bootstrap Samples (Sampling with replacement)

b. "Train" Regression Machine on Each

c. Average the predictors, quantify variability

Boosting



For classification (response is $\{0, 1\}$), AdaBoost (and variants).

Gradient boosting (Regression)

.

- Train a "tree" $\{f_1, ...f_M\}$
- Compute residuals for each $m = 1...M$
- $r_i = y_i - f_{M-1}(x_i)$ - sum all the way to $M - 1$.
- Fit a "tree" to $r_i : f_M$
- Add f_M

12.S990 Quantifying Uncertainty
Fall 2012