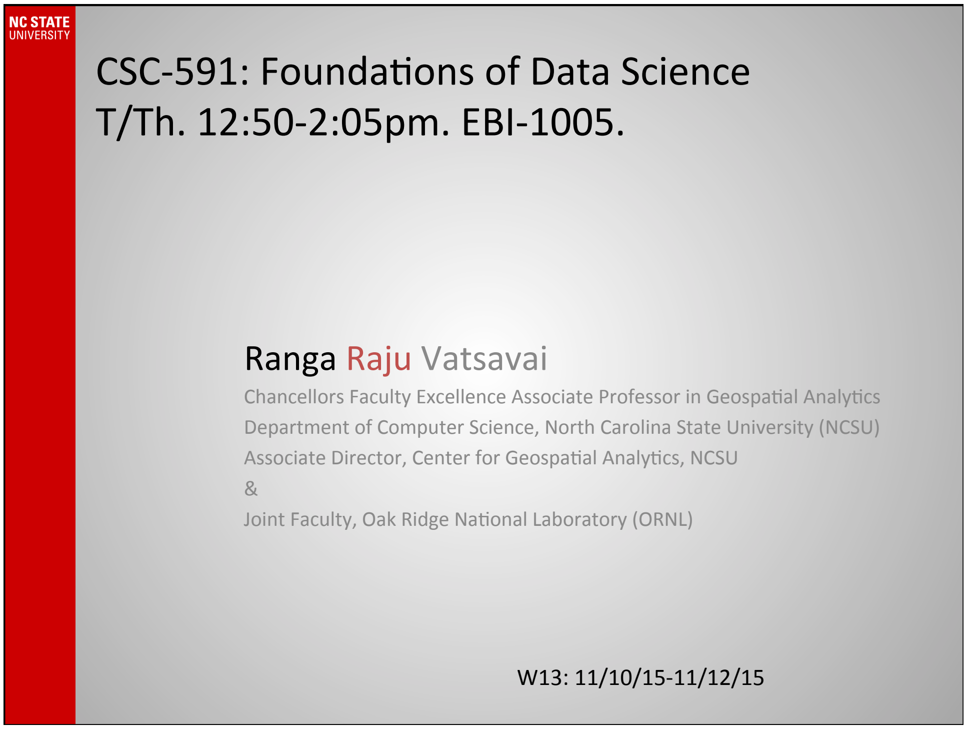
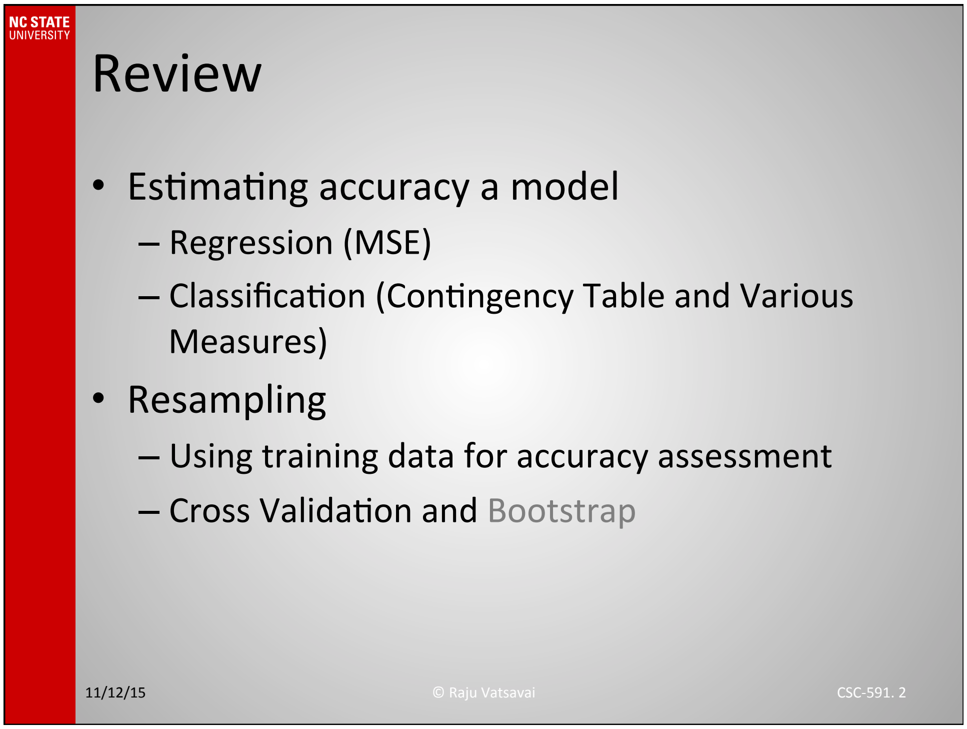


NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

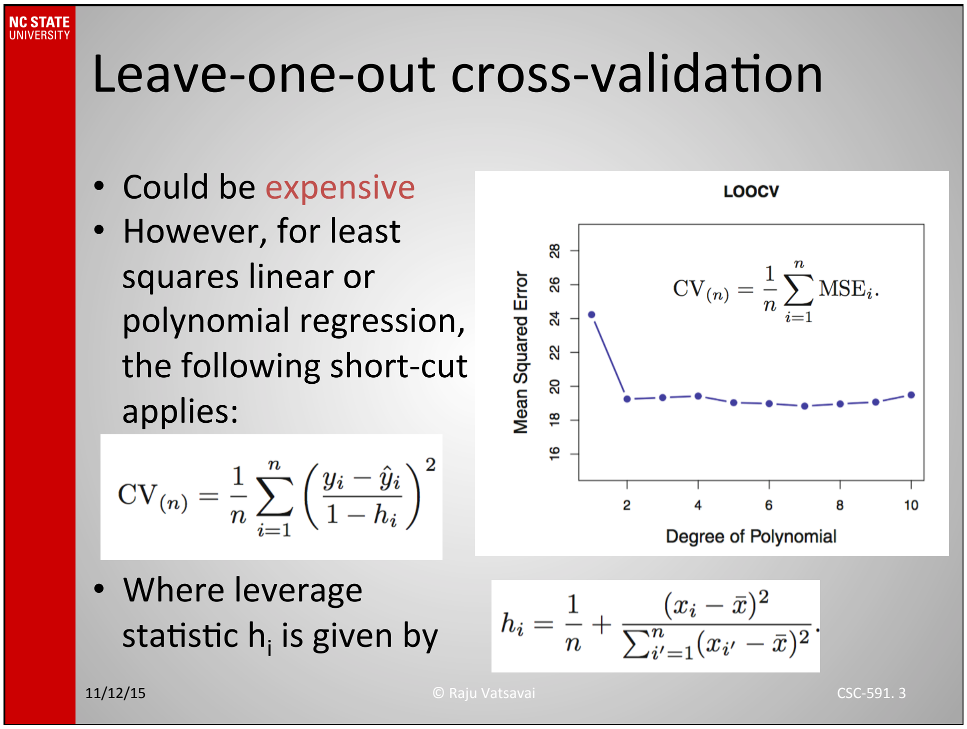
W13: 11/10/15-11/12/15



NC STATE UNIVERSITY

Review

- Estimating accuracy a model
  - Regression (MSE)
  - Classification (Contingency Table and Various Measures)
- Resampling
  - Using training data for accuracy assessment
  - Cross Validation and Bootstrap



NC STATE UNIVERSITY

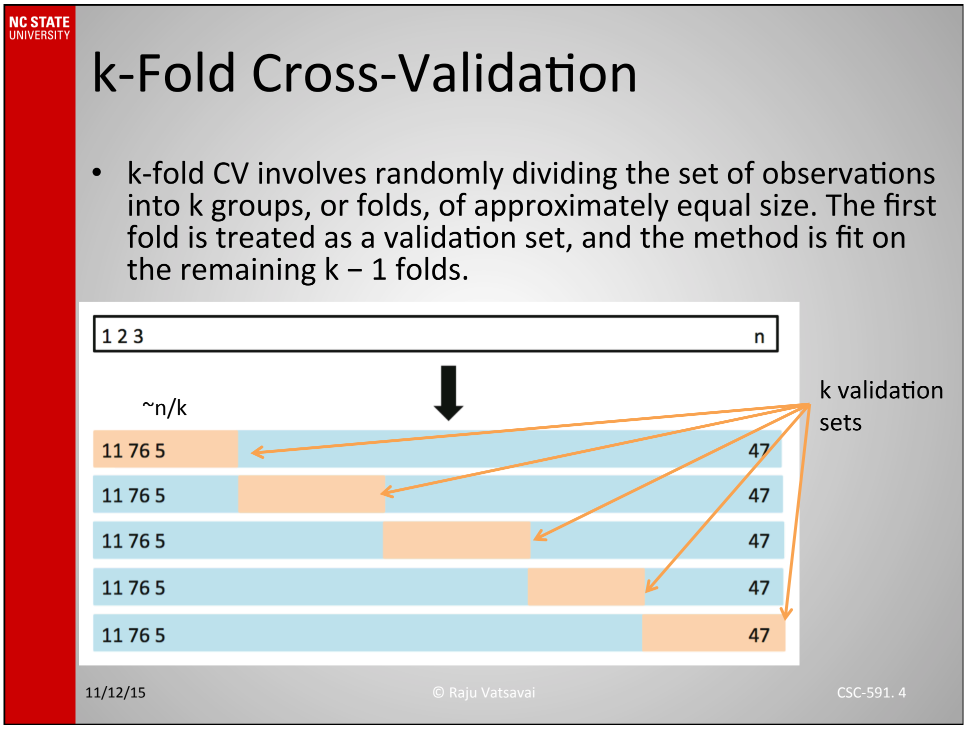
Leave-one-out cross-validation

- Could be expensive
- However, for least squares linear or polynomial regression, the following short-cut applies:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

- Where leverage statistic h_i is given by

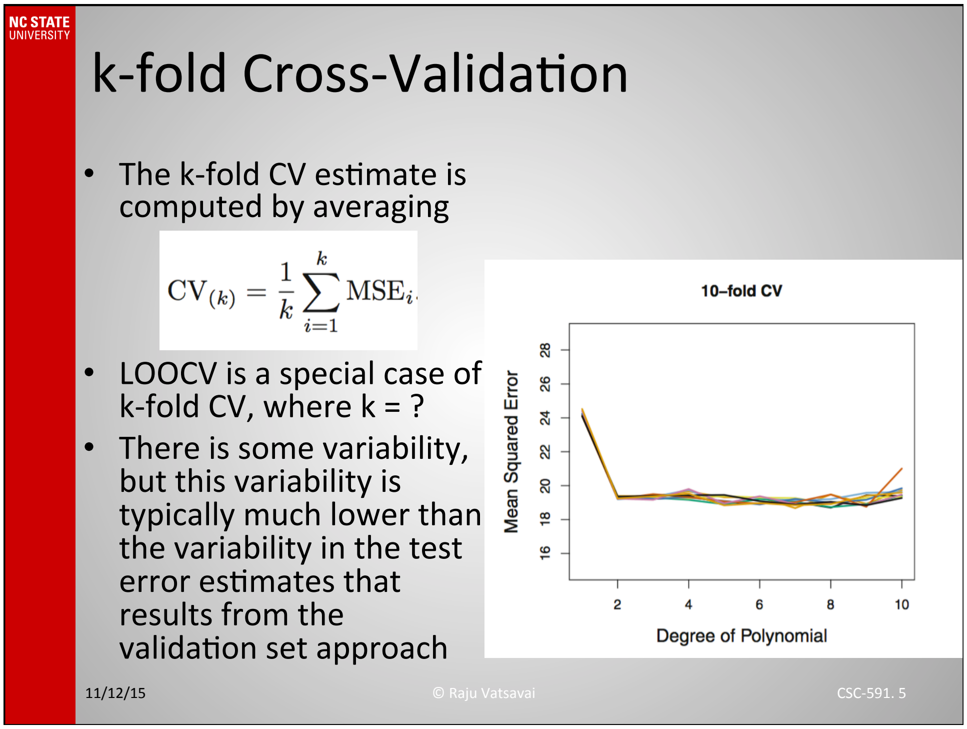
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}.$$



NC STATE UNIVERSITY

k-Fold Cross-Validation

- k-fold CV involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds.



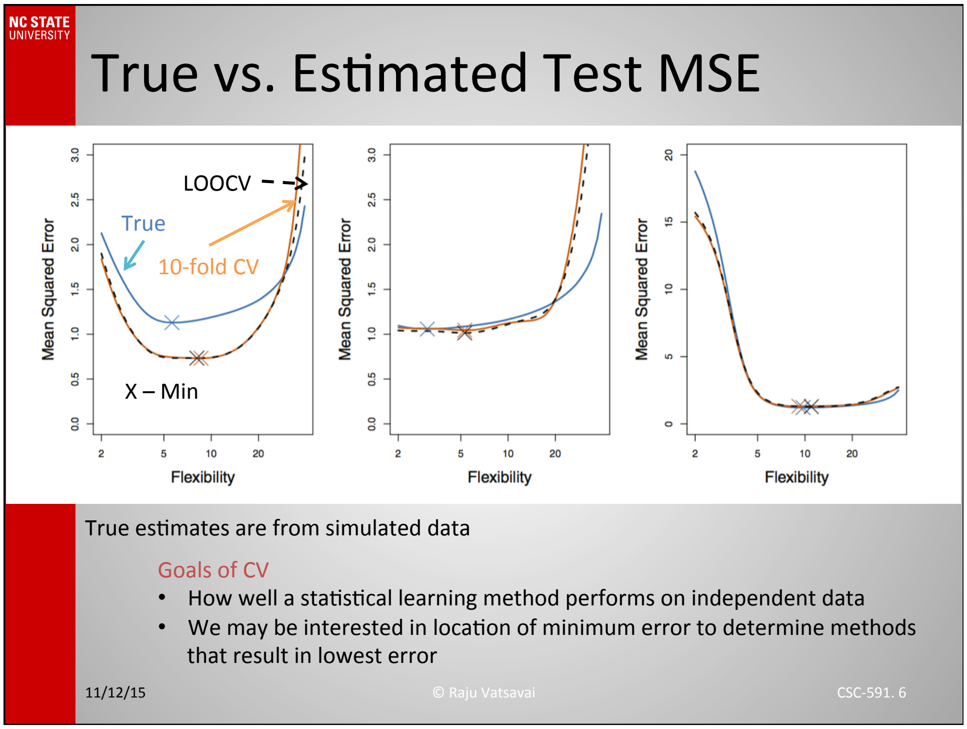
NC STATE UNIVERSITY

k-fold Cross-Validation

- The k-fold CV estimate is computed by averaging

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i.$$

- LOOCV is a special case of k-fold CV, where k = ?
- There is some variability, but this variability is typically much lower than the variability in the test error estimates that results from the validation set approach



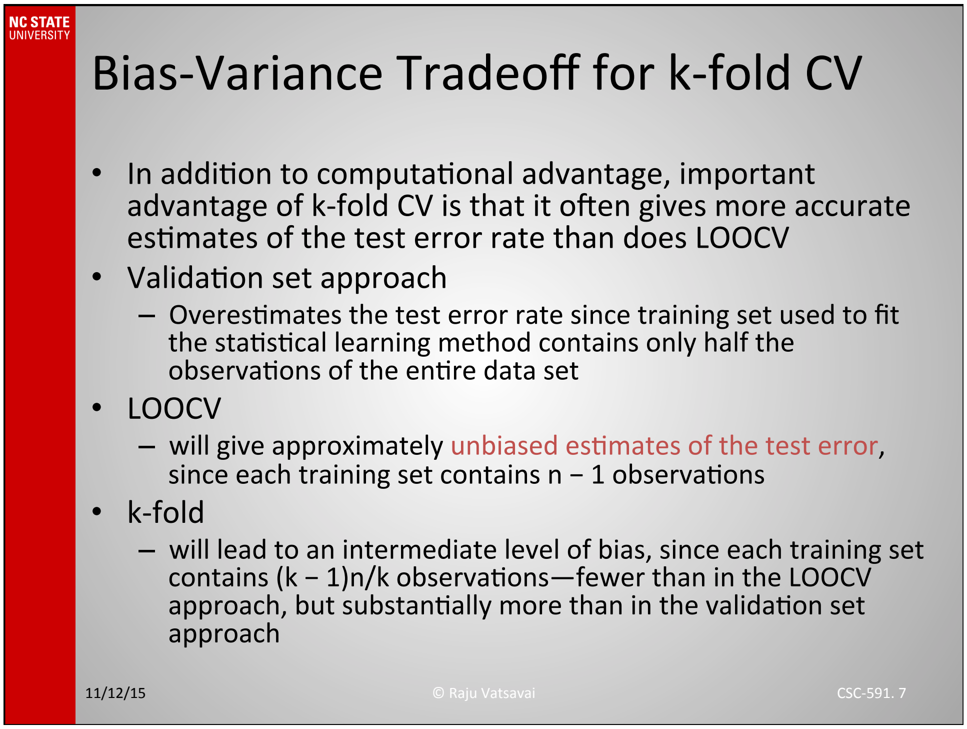
NC STATE UNIVERSITY

True vs. Estimated Test MSE

True estimates are from simulated data

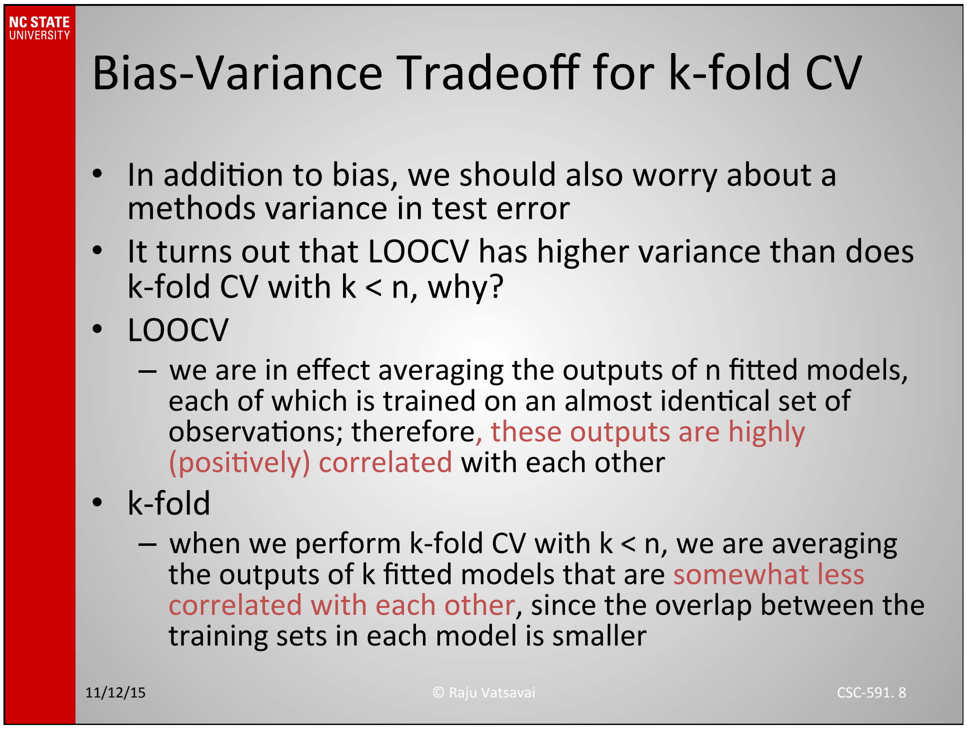
Goals of CV

- How well a statistical learning method performs on independent data
- We may be interested in location of minimum error to determine methods that result in lowest error



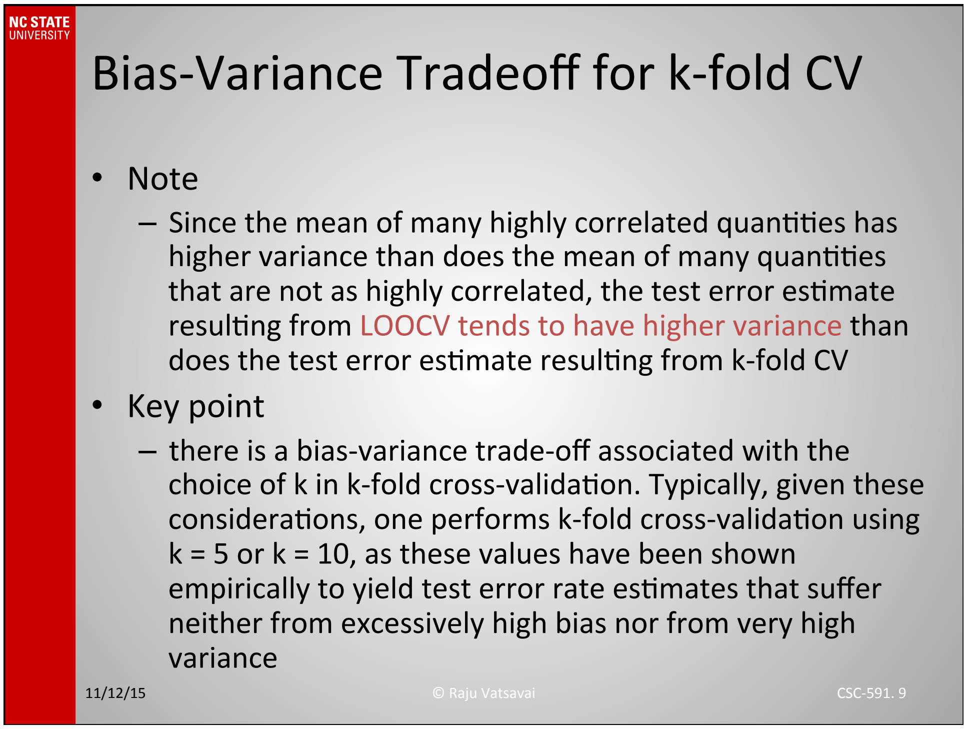
Bias-Variance Tradeoff for k-fold CV

- In addition to computational advantage, important advantage of k-fold CV is that it often gives more accurate estimates of the test error rate than does LOOCV
- Validation set approach
  - Overestimates the test error rate since training set used to fit the statistical learning method contains only half the observations of the entire data set
- LOOCV
  - will give approximately unbiased estimates of the test error, since each training set contains n − 1 observations
- k-fold
  - will lead to an intermediate level of bias, since each training set contains (k − 1)n/k observations—fewer than in the LOOCV approach, but substantially more than in the validation set approach



Bias-Variance Tradeoff for k-fold CV

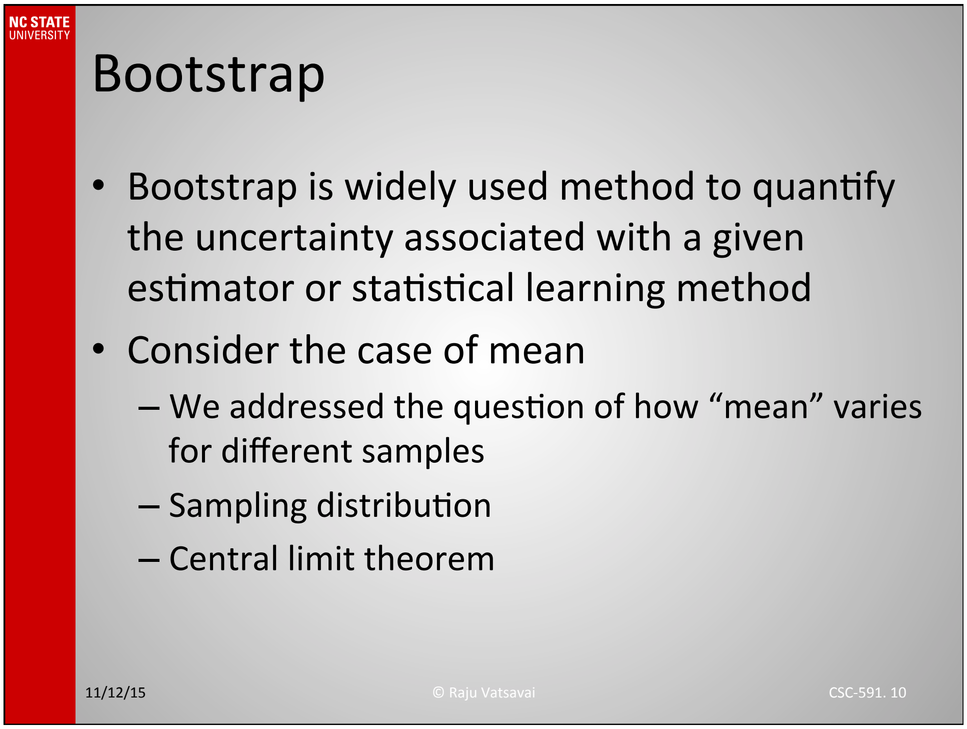
- In addition to bias, we should also worry about a methods variance in test error
- It turns out that LOOCV has higher variance than does k-fold CV with k < n, why?
- LOOCV
  - we are in effect averaging the outputs of n fitted models, each of which is trained on an almost identical set of observations; therefore, these outputs are highly (positively) correlated with each other
- k-fold
  - when we perform k-fold CV with k < n, we are averaging the outputs of k fitted models that are somewhat less correlated with each other, since the overlap between the training sets in each model is smaller



NC STATE UNIVERSITY

Bias-Variance Tradeoff for k-fold CV

- Note
  - Since the mean of many highly correlated quantities has higher variance than does the mean of many quantities that are not as highly correlated, the test error estimate resulting from LOOCV tends to have higher variance than does the test error estimate resulting from k-fold CV
- Key point
  - there is a bias-variance trade-off associated with the choice of k in k-fold cross-validation. Typically, given these considerations, one performs k-fold cross-validation using k = 5 or k = 10, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance



NC STATE UNIVERSITY

Bootstrap

- Bootstrap is widely used method to quantify the uncertainty associated with a given estimator or statistical learning method
- Consider the case of mean
  - We addressed the question of how “mean” varies for different samples
  - Sampling distribution
  - Central limit theorem

NC STATE UNIVERSITY

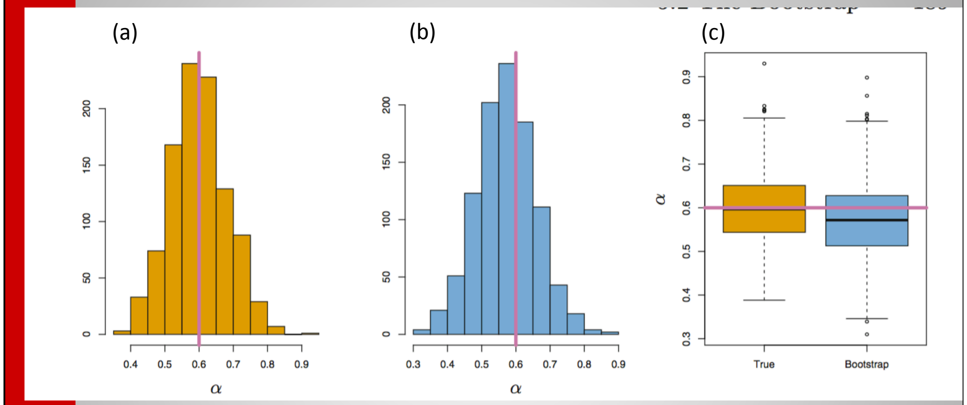
Bootstrap

- In practice it is difficult to generate new samples from the original population
- Bootstrap allows us to use computer to emulate the process of obtaining new sample sets so that we estimate the variability of a statistic without generating additional samples
- Using bootstrap, rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set.

11/12/15 © Raju Vatsavai CSC-591. 11

NC STATE UNIVERSITY

Bootstrap



(**a**) histogram of estimates of α obtained by generating 1000 simulated samples from true population

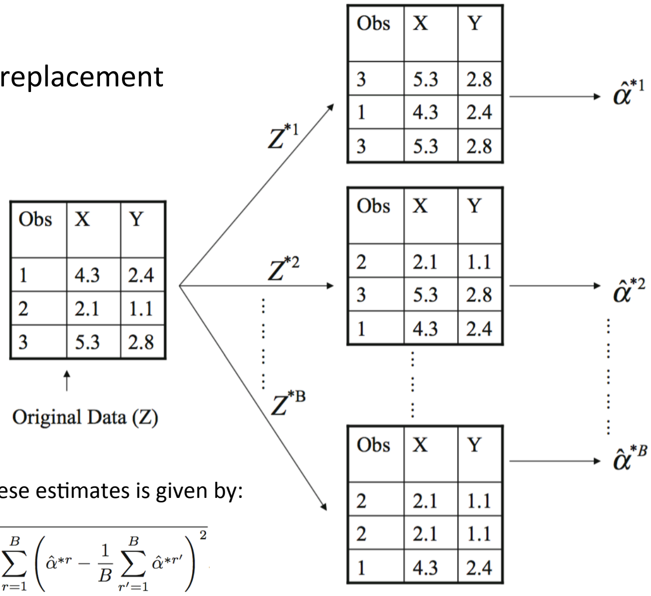
(**b**) histogram of estimates of α obtained form 1000 bootstrap samples from a single dataset

(**c**) boxplots. Pink line indicates true value of α

11/12/15 © Raju Vatsavai CSC-591. 12

Bootstrap

Sampling with replacement



Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Original Data (Z)

Z^{*1}

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

$\hat{\alpha}^{*1}$

Z^{*2}

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

$\hat{\alpha}^{*2}$

Z^{*B}

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

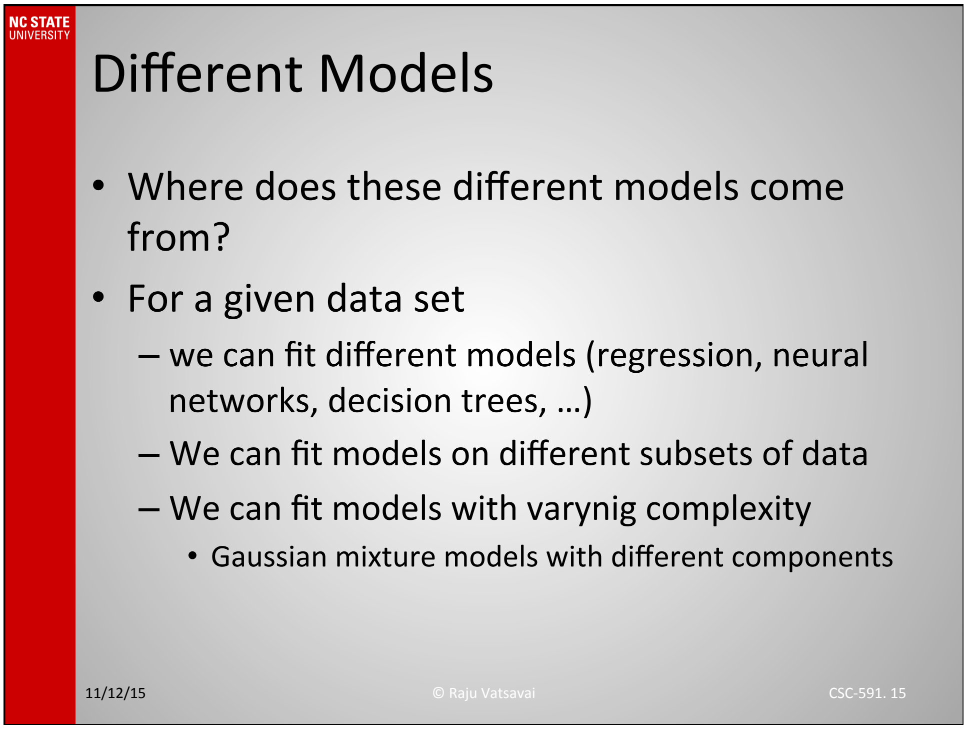
$\hat{\alpha}^{*B}$

Standard Error of these estimates is given by:

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}$$

Feature Selection Vs. Model Selection

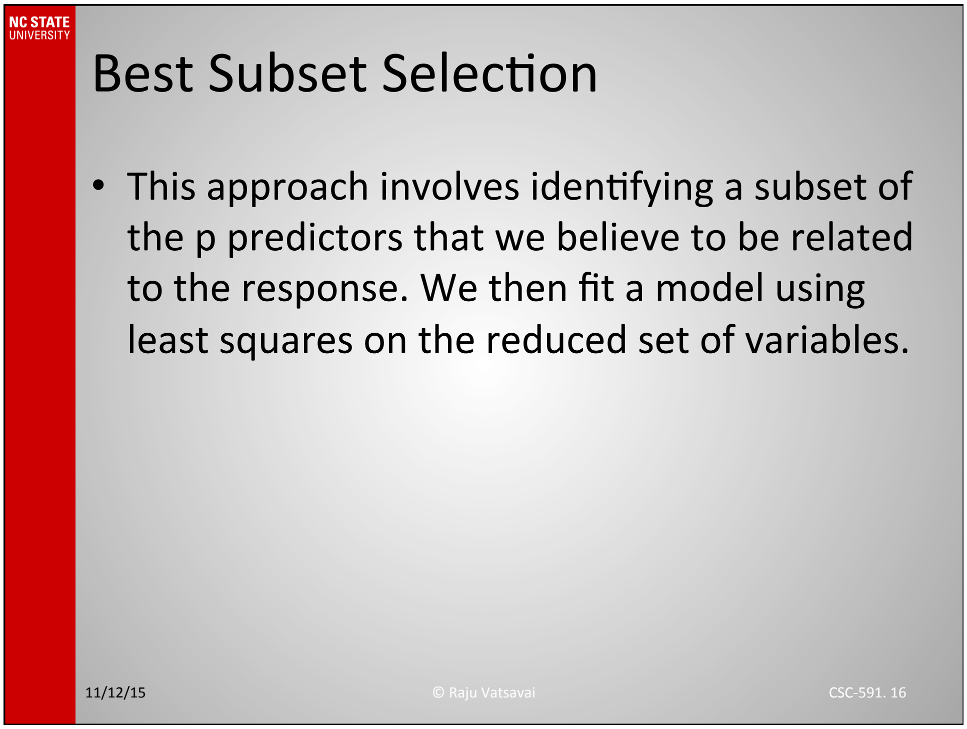
- Feature selection (variable selection or attribute selection)
  – Process of choosing a subset of relevant feature for use in the model construction
- Model selection
  – Process of choosing a sparse model that adequately explains the data



NC STATE UNIVERSITY

Different Models

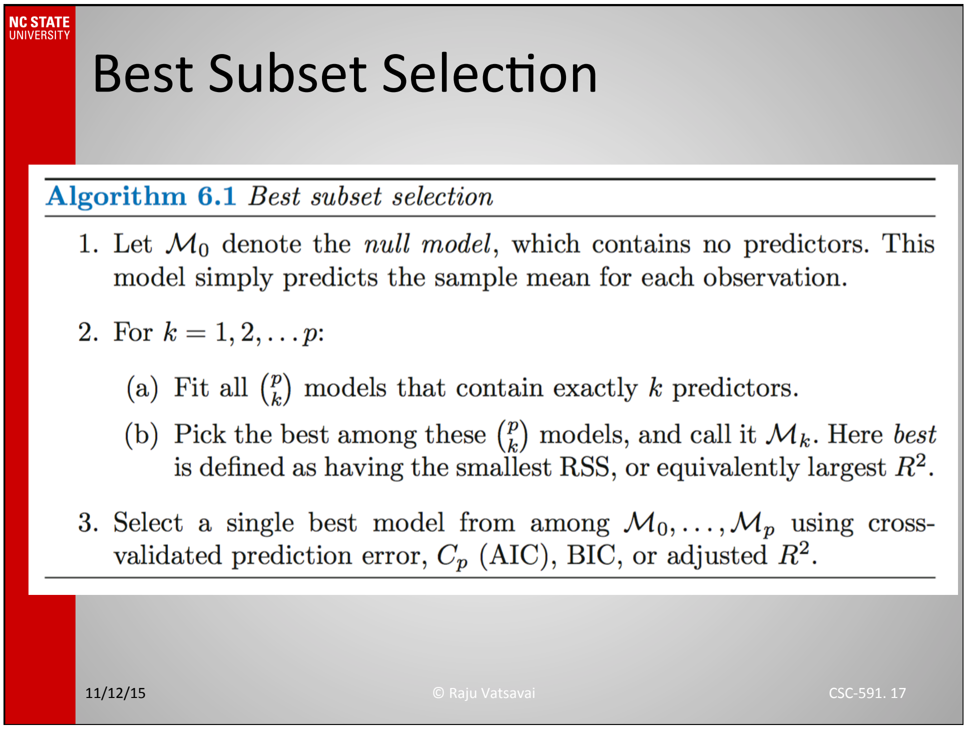
- Where does these different models come from?
- For a given data set
  - we can fit different models (regression, neural networks, decision trees, …)
  - We can fit models on different subsets of data
  - We can fit models with varynig complexity
    - Gaussian mixture models with different components



NC STATE UNIVERSITY

Best Subset Selection

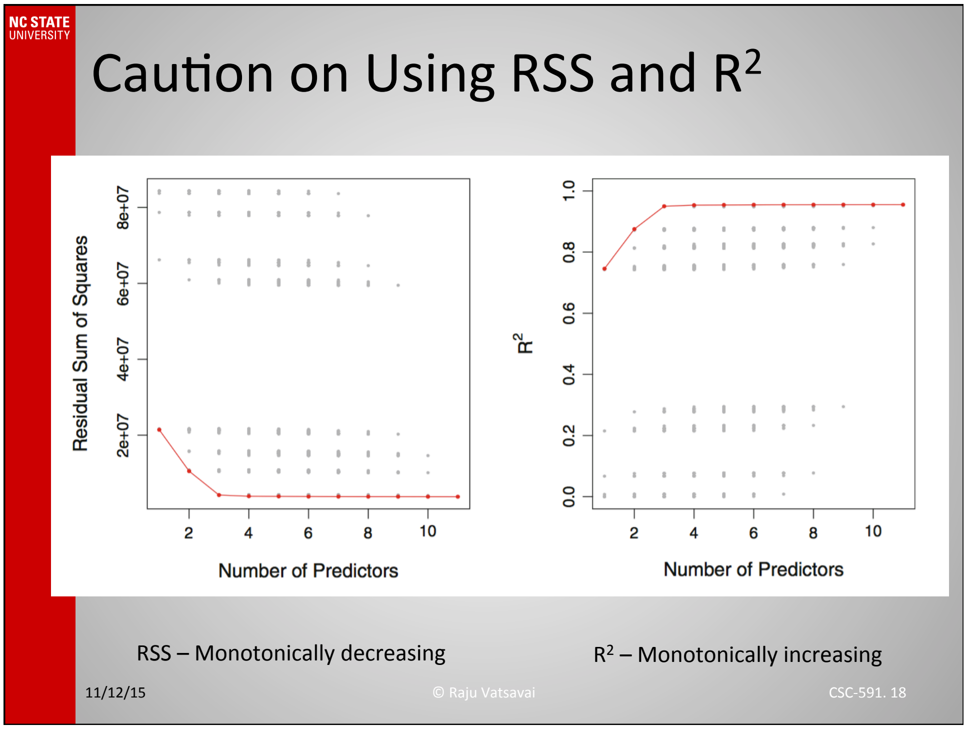
- This approach involves identifying a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.



Best Subset Selection

Algorithm 6.1 *Best subset selection*

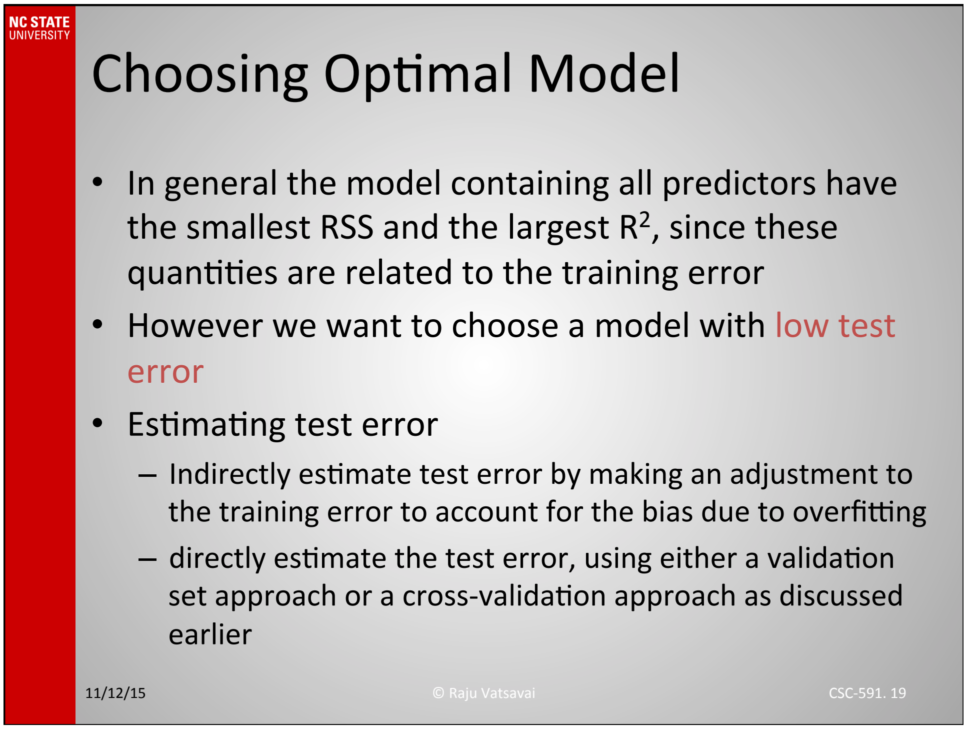
```
1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This
   model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best*
       is defined as having the smallest RSS, or equivalently largest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-
   validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.
```



Caution on Using RSS and R^2

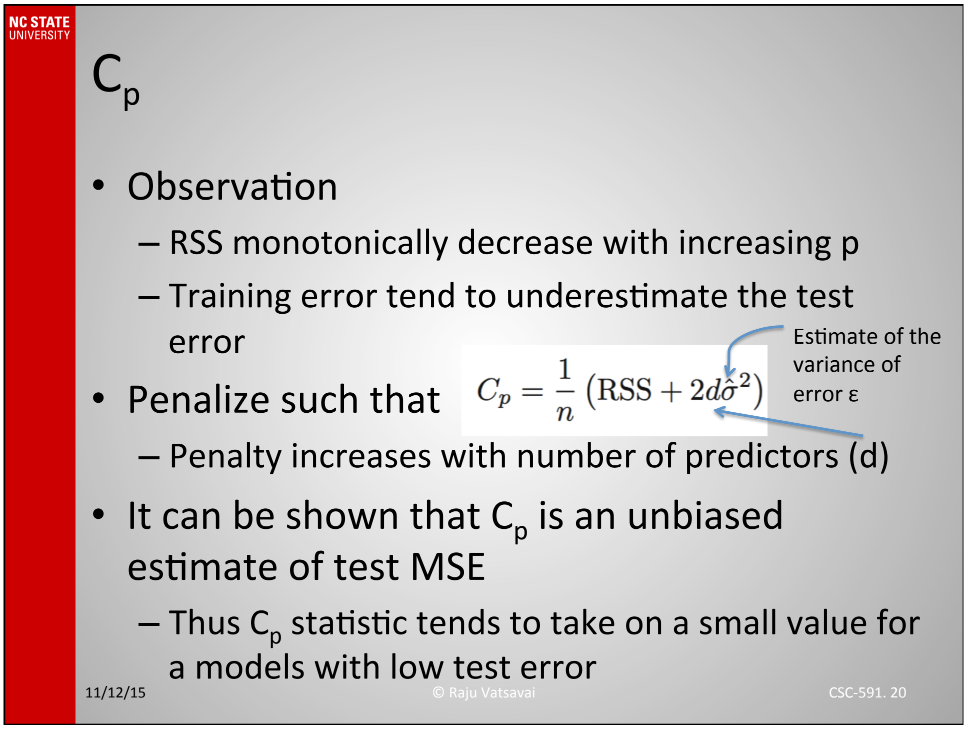
RSS – Monotonically decreasing

R^2 – Monotonically increasing



Choosing Optimal Model

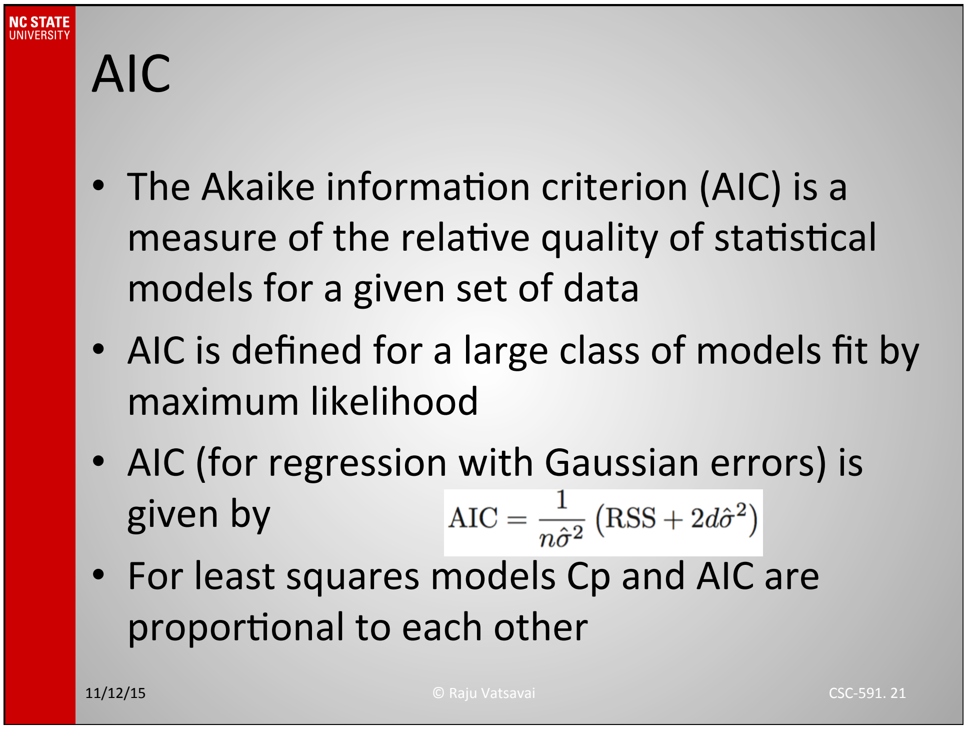
- In general the model containing all predictors have the smallest RSS and the largest R^2, since these quantities are related to the training error
- However we want to choose a model with low test error
- Estimating test error
  - Indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting
  - directly estimate the test error, using either a validation set approach or a cross-validation approach as discussed earlier



C_p

- Observation
  - RSS monotonically decrease with increasing p
  - Training error tend to underestimate the test error
- Penalize such that $C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$

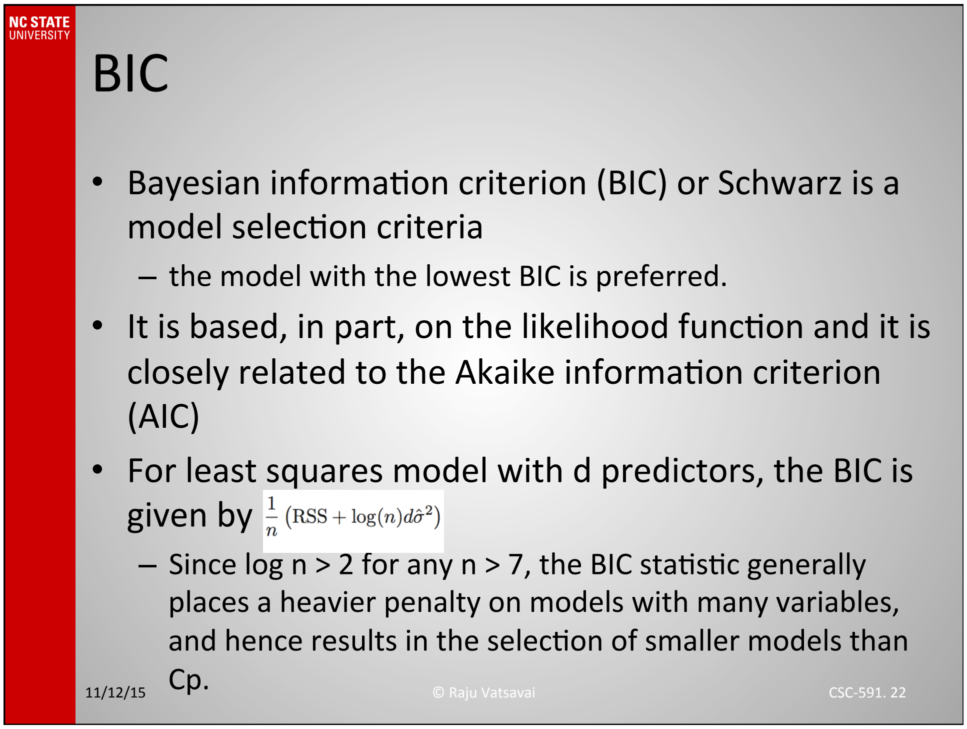
  Estimate of the variance of error ε
  - Penalty increases with number of predictors (d)
- It can be shown that C_p is an unbiased estimate of test MSE
  - Thus C_p statistic tends to take on a small value for a models with low test error



NC STATE UNIVERSITY

AIC

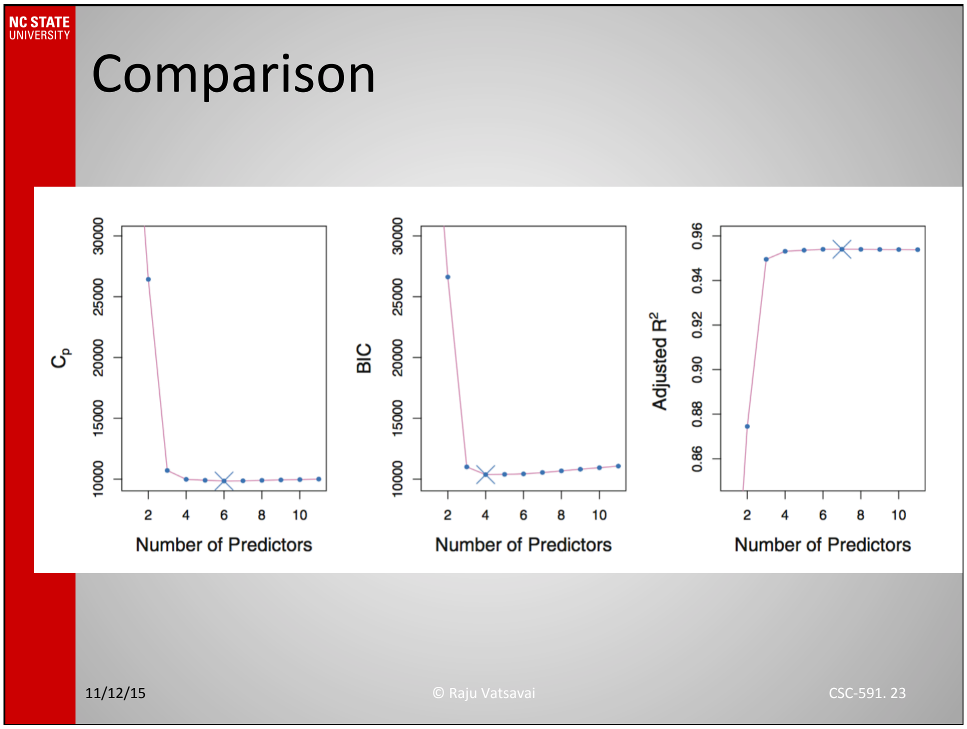
- The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data
- AIC is defined for a large class of models fit by maximum likelihood
- AIC (for regression with Gaussian errors) is given by $\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$
- For least squares models Cp and AIC are proportional to each other



NC STATE UNIVERSITY

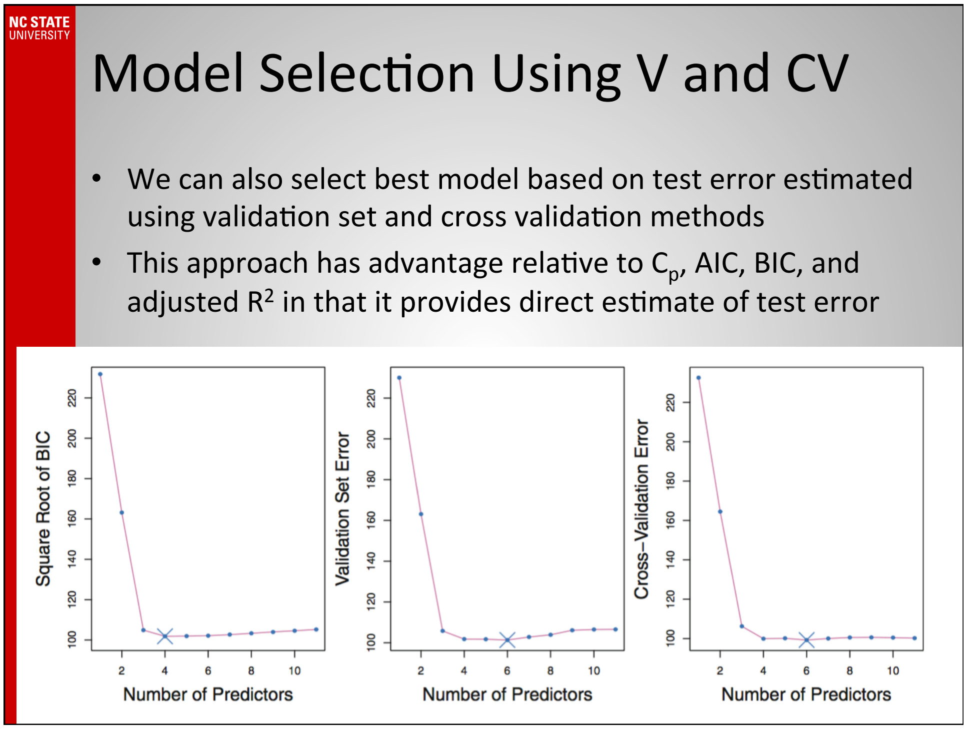
BIC

- Bayesian information criterion (BIC) or Schwarz is a model selection criteria
  - the model with the lowest BIC is preferred.
- It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC)
- For least squares model with d predictors, the BIC is given by $\frac{1}{n}\left(\mathrm{RSS} + \log(n)d\hat{\sigma}^2\right)$
  - Since log n > 2 for any n > 7, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than Cp.



NC STATE UNIVERSITY

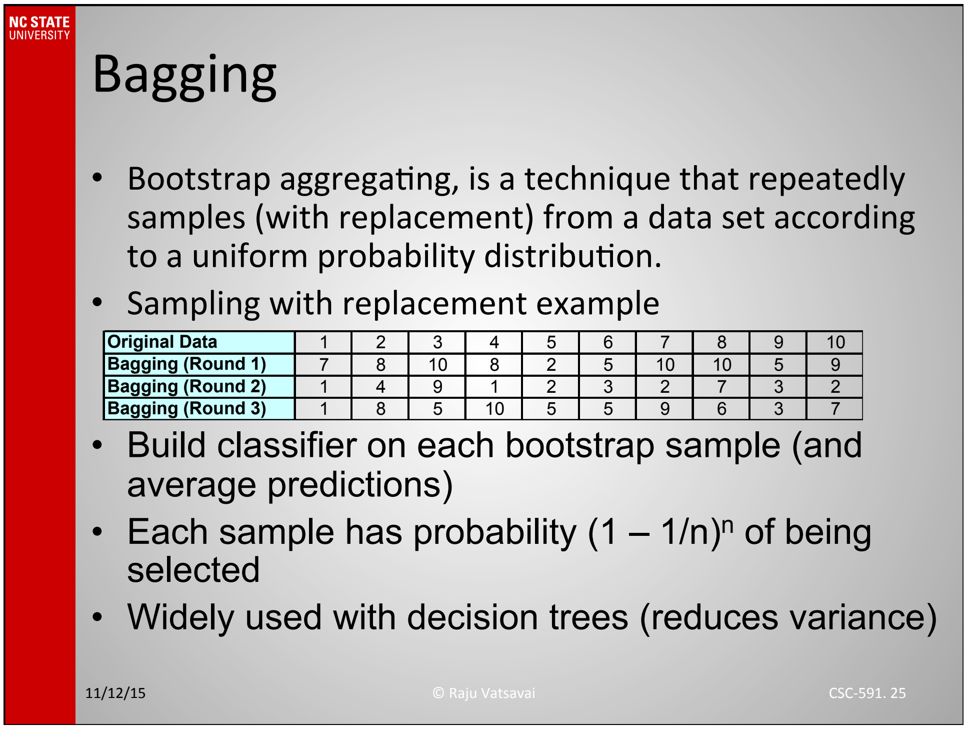
Comparison



NC STATE UNIVERSITY

Model Selection Using V and CV

- We can also select best model based on test error estimated using validation set and cross validation methods
- This approach has advantage relative to C_p, AIC, BIC, and adjusted R^2 in that it provides direct estimate of test error

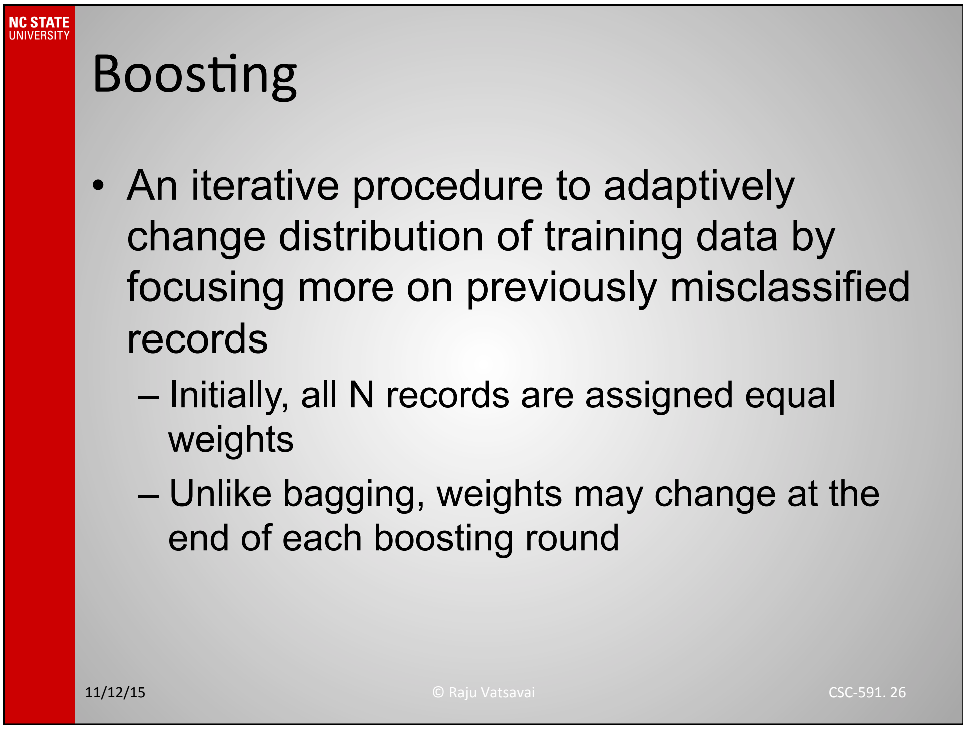


Bagging

- Bootstrap aggregating, is a technique that repeatedly samples (with replacement) from a data set according to a uniform probability distribution.
- Sampling with replacement example

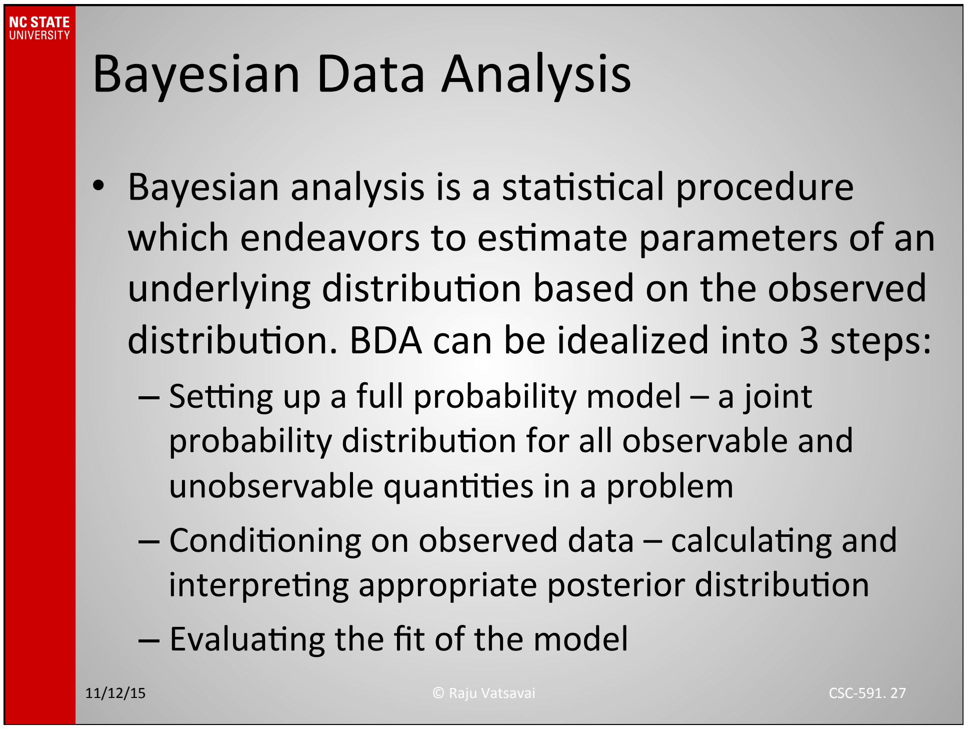
Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- Build classifier on each bootstrap sample (and average predictions)
- Each sample has probability $(1 – 1/n)^n$ of being selected
- Widely used with decision trees (reduces variance)



Boosting

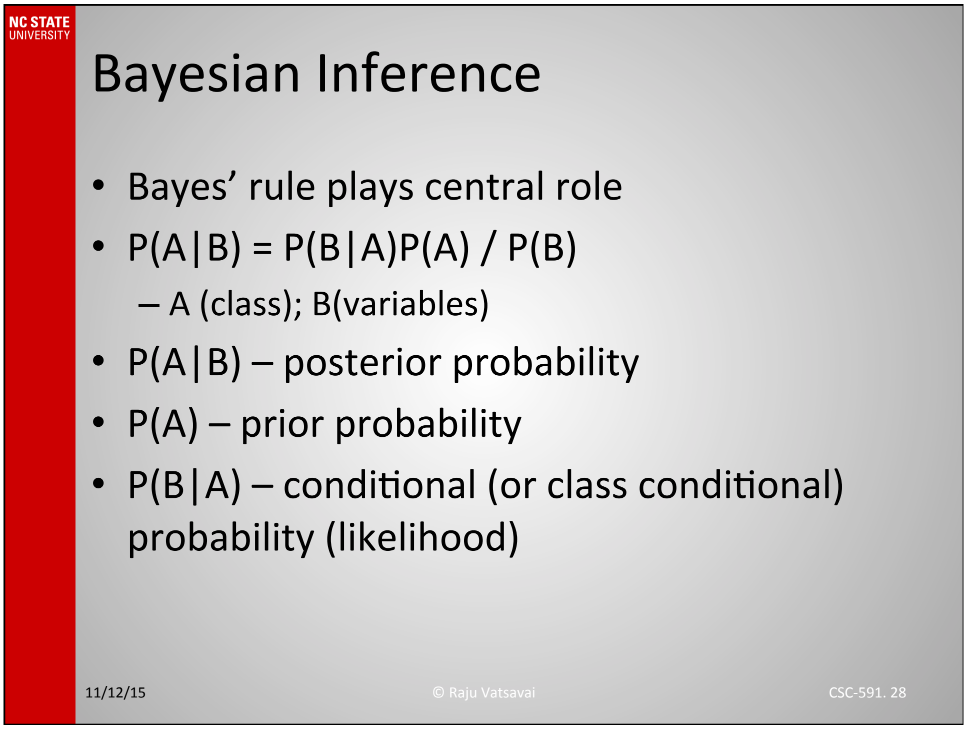
- An iterative procedure to adaptively change distribution of training data by focusing more on previously misclassified records
  - Initially, all N records are assigned equal weights
  - Unlike bagging, weights may change at the end of each boosting round



NC STATE UNIVERSITY

Bayesian Data Analysis

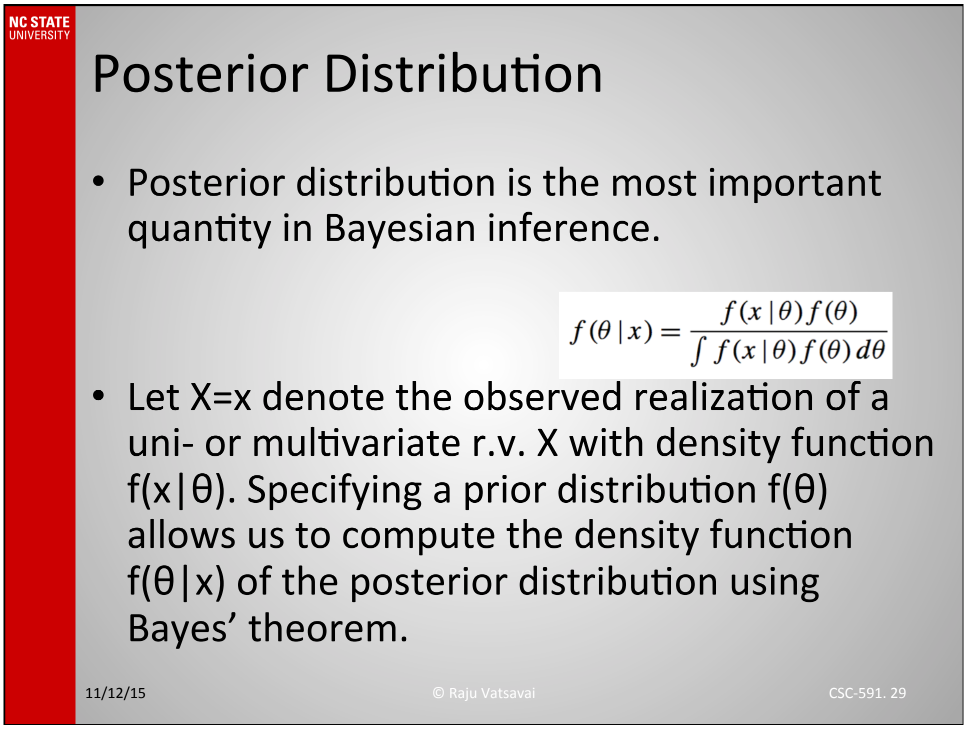
- Bayesian analysis is a statistical procedure which endeavors to estimate parameters of an underlying distribution based on the observed distribution. BDA can be idealized into 3 steps:
  - Setting up a full probability model – a joint probability distribution for all observable and unobservable quantities in a problem
  - Conditioning on observed data – calculating and interpreting appropriate posterior distribution
  - Evaluating the fit of the model



NC STATE UNIVERSITY

Bayesian Inference

- Bayes' rule plays central role
- P(A|B) = P(B|A)P(A) / P(B)
  - A (class); B(variables)
- P(A|B) – posterior probability
- P(A) – prior probability
- P(B|A) – conditional (or class conditional) probability (likelihood)



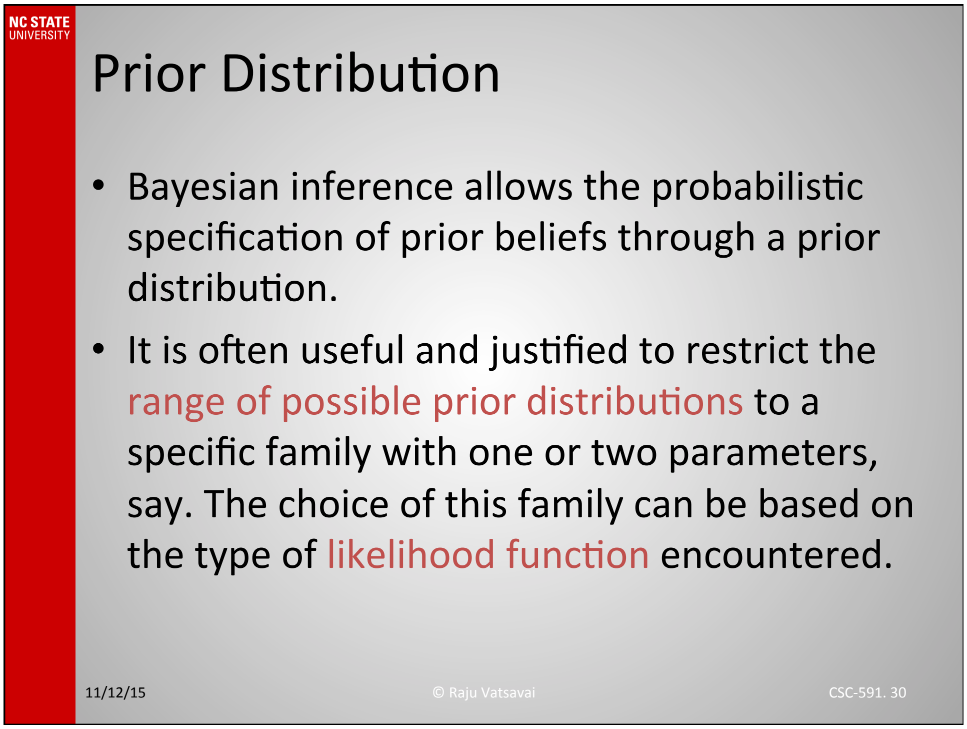
NC STATE UNIVERSITY

Posterior Distribution

- Posterior distribution is the most important quantity in Bayesian inference.

$$f(\theta \mid x) = \frac{f(x \mid \theta)\, f(\theta)}{\int f(x \mid \theta)\, f(\theta)\, d\theta}$$

- Let X=x denote the observed realization of a uni- or multivariate r.v. X with density function f(x|θ). Specifying a prior distribution f(θ) allows us to compute the density function f(θ|x) of the posterior distribution using Bayes' theorem.



NC STATE UNIVERSITY

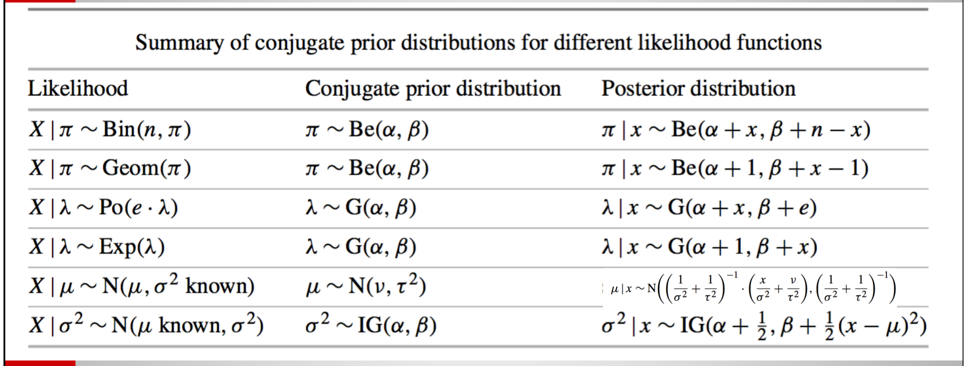
Prior Distribution

- Bayesian inference allows the probabilistic specification of prior beliefs through a prior distribution.
- It is often useful and justified to restrict the range of possible prior distributions to a specific family with one or two parameters, say. The choice of this family can be based on the type of likelihood function encountered.

Conjugate Prior Distributions

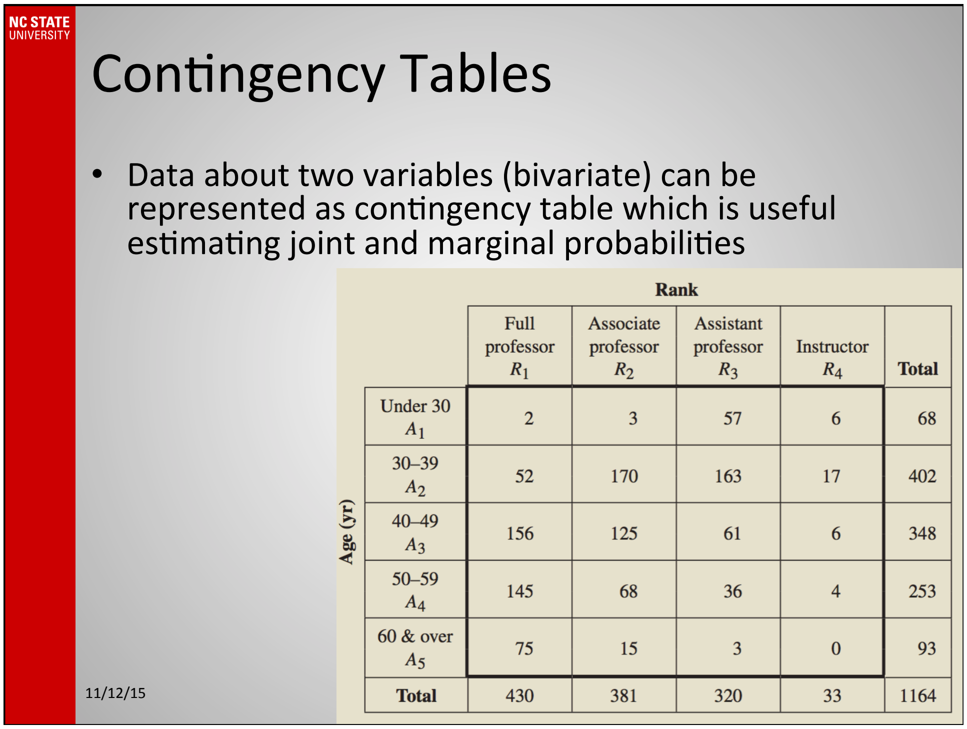
- A pragmatic approach to choosing a prior distribution is to select a member of a specific family of distributions such that the posterior distribution belongs to the same family. This is called a *conjugate prior distribution*.
- Let L(θ) = f (x | θ) denote a likelihood function based on the observation X = x. A class ζ of distributions is called *conjugate with respect to* L(θ) if the posterior distribution f (θ | x) is in ζ for all x whenever the prior distribution f (θ) is in ζ .

Conjugate Prior Distributions



Summary of conjugate prior distributions for different likelihood functions

Likelihood	Conjugate prior distribution	Posterior distribution
$X \mid \pi \sim \mathrm{Bin}(n, \pi)$	$\pi \sim \mathrm{Be}(\alpha, \beta)$	$\pi \mid x \sim \mathrm{Be}(\alpha + x, \beta + n - x)$
$X \mid \pi \sim \mathrm{Geom}(\pi)$	$\pi \sim \mathrm{Be}(\alpha, \beta)$	$\pi \mid x \sim \mathrm{Be}(\alpha + 1, \beta + x - 1)$
$X \mid \lambda \sim \mathrm{Po}(e \cdot \lambda)$	$\lambda \sim \mathrm{G}(\alpha, \beta)$	$\lambda \mid x \sim \mathrm{G}(\alpha + x, \beta + e)$
$X \mid \lambda \sim \mathrm{Exp}(\lambda)$	$\lambda \sim \mathrm{G}(\alpha, \beta)$	$\lambda \mid x \sim \mathrm{G}(\alpha + 1, \beta + x)$
$X \mid \mu \sim \mathrm{N}(\mu, \sigma^2 \text{ known})$	$\mu \sim \mathrm{N}(\nu, \tau^2)$	$\mu \mid x \sim \mathrm{N}\left(\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \cdot \left(\frac{x}{\sigma^2} + \frac{\nu}{\tau^2}\right), \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right)$
$X \mid \sigma^2 \sim \mathrm{N}(\mu \text{ known}, \sigma^2)$	$\sigma^2 \sim \mathrm{IG}(\alpha, \beta)$	$\sigma^2 \mid x \sim \mathrm{IG}(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(x - \mu)^2)$



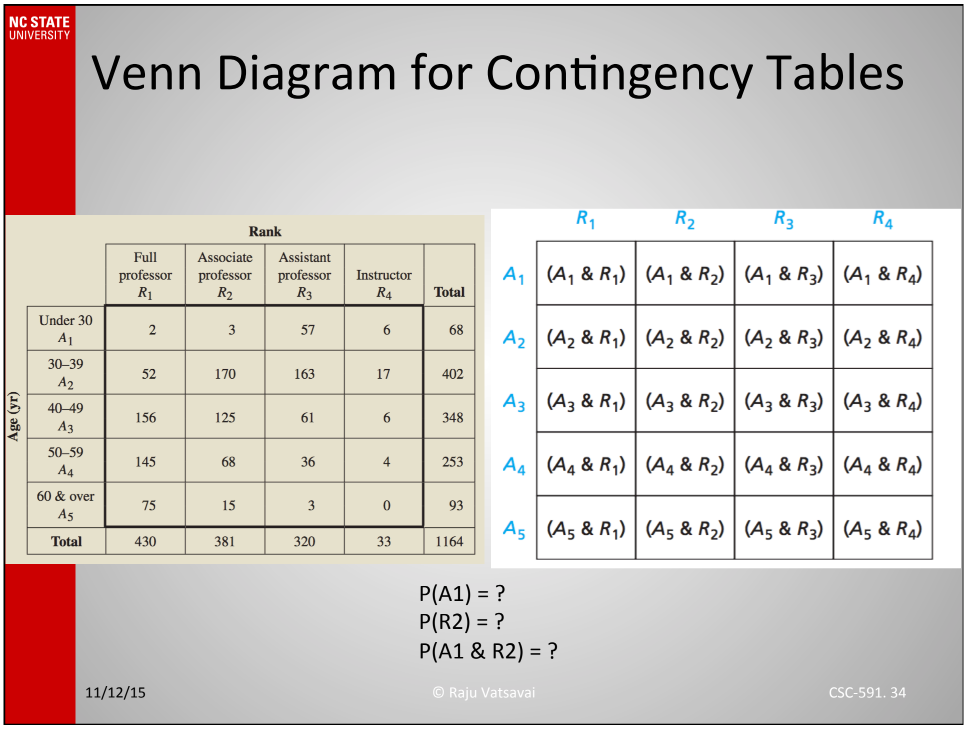
NC STATE UNIVERSITY

Contingency Tables

- Data about two variables (bivariate) can be represented as contingency table which is useful estimating joint and marginal probabilities

		Rank				
		Full professor R_1	Associate professor R_2	Assistant professor R_3	Instructor R_4	Total
Age (yr)	Under 30 A_1	2	3	57	6	68
	30–39 A_2	52	170	163	17	402
	40–49 A_3	156	125	61	6	348
	50–59 A_4	145	68	36	4	253
	60 & over A_5	75	15	3	0	93
	Total	430	381	320	33	1164

11/12/15



NC STATE UNIVERSITY

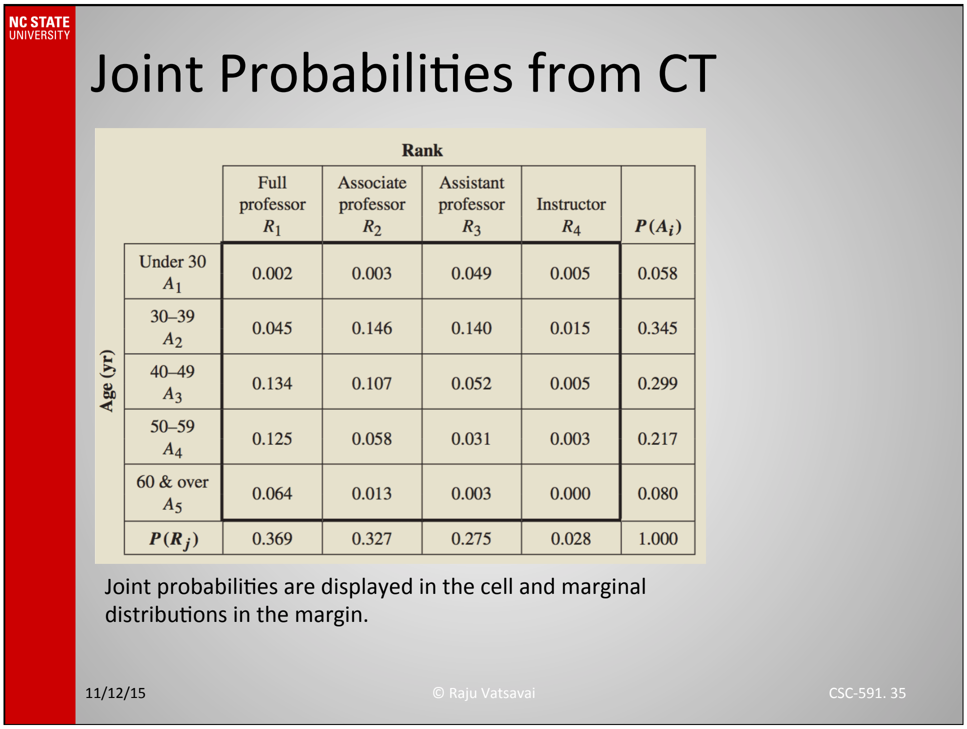
Venn Diagram for Contingency Tables

		Rank				
		Full professor R_1	Associate professor R_2	Assistant professor R_3	Instructor R_4	Total
Age (yr)	Under 30 A_1	2	3	57	6	68
	30–39 A_2	52	170	163	17	402
	40–49 A_3	156	125	61	6	348
	50–59 A_4	145	68	36	4	253
	60 & over A_5	75	15	3	0	93
	Total	430	381	320	33	1164

	R_1	R_2	R_3	R_4
A_1	$(A_1$ & $R_1)$	$(A_1$ & $R_2)$	$(A_1$ & $R_3)$	$(A_1$ & $R_4)$
A_2	$(A_2$ & $R_1)$	$(A_2$ & $R_2)$	$(A_2$ & $R_3)$	$(A_2$ & $R_4)$
A_3	$(A_3$ & $R_1)$	$(A_3$ & $R_2)$	$(A_3$ & $R_3)$	$(A_3$ & $R_4)$
A_4	$(A_4$ & $R_1)$	$(A_4$ & $R_2)$	$(A_4$ & $R_3)$	$(A_4$ & $R_4)$
A_5	$(A_5$ & $R_1)$	$(A_5$ & $R_2)$	$(A_5$ & $R_3)$	$(A_5$ & $R_4)$

P(A1) = ?
P(R2) = ?
P(A1 & R2) = ?

11/12/15 © Raju Vatsavai CSC-591. 34

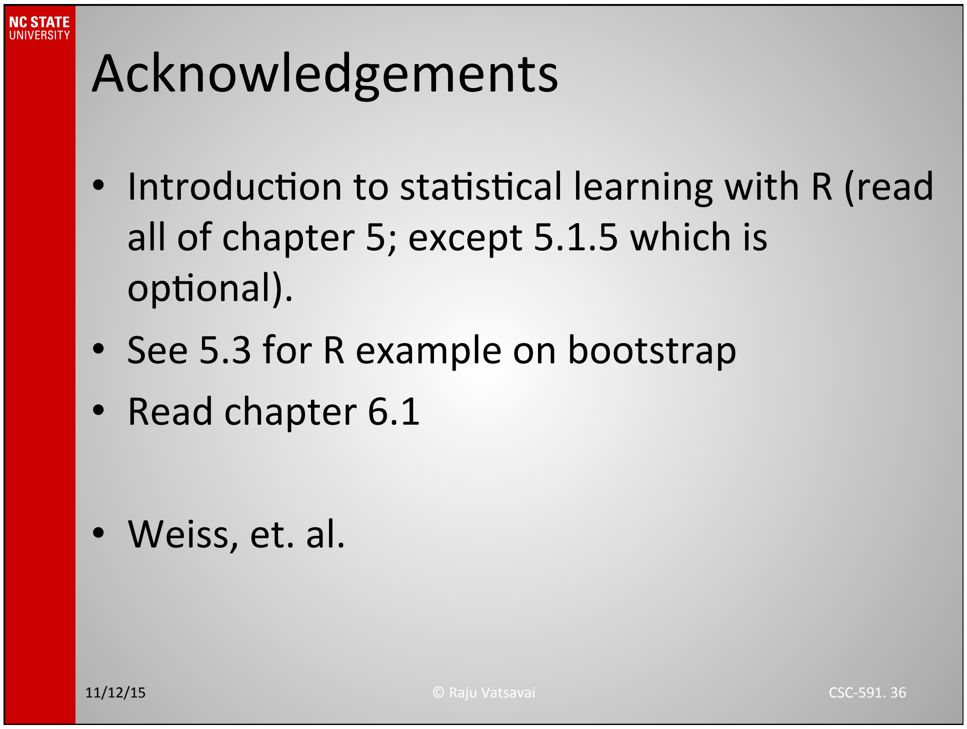


NC STATE UNIVERSITY

Joint Probabilities from CT

		Rank				
		Full professor R_1	Associate professor R_2	Assistant professor R_3	Instructor R_4	$P(A_i)$
Age (yr)	Under 30 A_1	0.002	0.003	0.049	0.005	0.058
	30–39 A_2	0.045	0.146	0.140	0.015	0.345
	40–49 A_3	0.134	0.107	0.052	0.005	0.299
	50–59 A_4	0.125	0.058	0.031	0.003	0.217
	60 & over A_5	0.064	0.013	0.003	0.000	0.080
	$P(R_j)$	0.369	0.327	0.275	0.028	1.000

Joint probabilities are displayed in the cell and marginal distributions in the margin.



NC STATE UNIVERSITY

Acknowledgements

- Introduction to statistical learning with R (read all of chapter 5; except 5.1.5 which is optional).
- See 5.3 for R example on bootstrap
- Read chapter 6.1

- Weiss, et. al.