Statistical Considerations in Prediction: The Role of Predicting Observables

Joe Heyse
Merck Research laboratories
UMBC – Stanford Workshop
September 24, 2016

Overview

- Estimation and testing of parameters has been the mainstay of statistical inference
- Growing interest on applications of prediction.
- Basing statistical inference on "observables" offers many advantages in these applications
  - More direct connection with the decision or objective of the analysis
  - More objective basis for model validation
  - Better capability for comparing models
- Both non-Bayesian and Bayesian methods are available and can be utilized

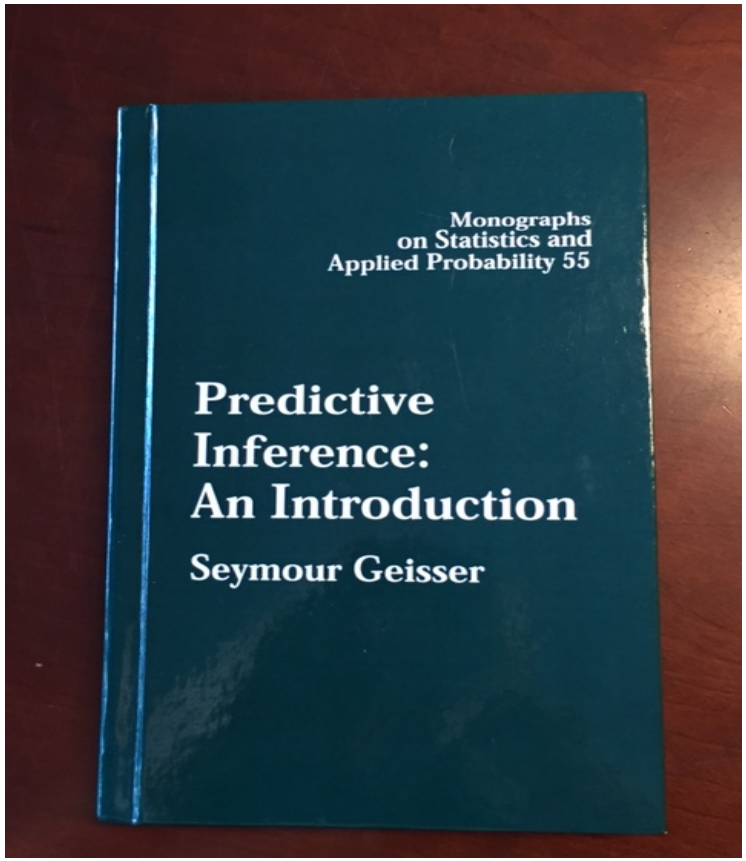
Outline

- Growing interest in prediction
- Basic elements of models for prediction
- Predicting observables as a framework
- Three illustrations
  - Analysis of Variance
  - Random Effects Meta-analysis
  - Classification
- Concluding Remarks

Growing interest in prediction

- **Predictive analytics** encompasses a variety of statistical techniques from modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events.
- **Predictive modelling** leverages statistics to predict outcomes. Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred.
- **Predictive inference**: Statistical inference in which the objective is not the estimation of parameters but the prediction of future observations from the same, or related, random system as generated the data.
- Other terms: **Predictive biomarker, predictive enrichment, predictive probability**

Seymour Geisser (October 5, 1929 – March 11, 2004) was a statistician noted for emphasizing the role of prediction in statistical inference. He held that conventional statistical inference about unobservable population parameters amounts to inference about things that do not exist. He also pioneered the theory of cross-validation.



Prediction (Geisser, 1993)

- Most statistical analysis involves inference about parameters of statistical distributions
  - Estimation
  - Testing hypotheses
- Inferences are often the basis of decisions
- A framework of predictions of future observables is often better suited for both inferences and decision making

Geisser's point of view

- Most statistical analyses use tests of significance about parameters to form conclusions
- Inference based on either point or interval estimates
- Importance is the effect on the distributions of observables

  "The real analysis starts after we have made allowance for parameters – known or unknown."

- Non-Bayesian and Bayesian approaches have the capability for implementing prediction methods
- Only Bayes is always capable of producing probability distributions for predictions

Leo Brieman's (2001) '*Two Cultures*'

- Two goals in analyzing data
  - Inference
  - Prediction
- Two cultures for modeling
  - Generative modeling: develop stochastic models which fit the data, and then make inferences about the data-generating mechanism (98%)
  - Predictive modeling: prioritizes prediction. The relatively recent discipline of Machine Learning is the "epicenter of the Predictive Modeling culture" (2%)

Validation of predictions or models

- Validation is based on $x^{(N)}$ observations divided into a model construction sample $x^{(N-n)}$ and a validation sample $x^{(n)}$
- Evaluation uses discrepancy

$$d_j = (\hat{x}_j - x_j) \text{ for } j = 1, \ldots, n$$

- Plot the d_j and summarize

Mean squared discrepancy $\frac{1}{n}\sum(d_j)^2$

Mean absolute discrepancy $\frac{1}{n}\sum|d_j|$

- Several cross-validation and systematic leave out i methods are available

Illustration 1
ANOVA model

- Have J groups or treatments and K observations per group, x_{kj}
- Usual estimate of $\theta_1, \ldots, \theta_J$ is $\bar{x}_j = \frac{1}{K}\sum_{k=1}^{K} x_{kj}$
- Stein (1962) showed that from an admissibility point of view, a shrunken estimate could yield a better set of estimates for $J \geq 3$

$$(1-\omega)\bar{x}_j + \omega\bar{x}$$

Predictive method

- Predict each of the $N = KJ$ observations x_{kj} using the remaining *N-1*

$$\bar{x}_{(kj)j} = c_{kj} = (K\bar{x}_j - x_{kj})/(K-1)$$
$$\bar{x}_{(kj)} = \bar{c}_{kj} = (N\bar{x} - x_{kj})/(N-1)$$

- Predictor for x_{kj}

$$\hat{x}_{kj} = (1-\omega)c_{kj} + \omega\bar{c}_{kj}$$

- Stein estimate chooses ω that minimizes the mean squared discrepancy

$$\frac{1}{N}\sum_j\sum_k \left(\hat{x}_{kj} - x_{kj}\right)^2$$

Example (Geisser, page 39)

- Samples of size K= 10 were drawn from J= 4 normal populations with means $\theta_1 = 0.35, \theta_2 = 0.1, \theta_3 = -0.1, \theta_4 = -0.35$ and variance $\sigma^2 = 1$
- Considered shrinkage estimates of the form

$$(1 - \omega)\bar{x}_j + \omega\bar{x} \text{ for } 0 \leq \omega \leq 1$$

- ω was chosen to minimize the mean squared discrepancy computed by predicting each of the 40 observations using the remaining 39
- Two models were evaluated:
  - One way ANOVA
  - Mixed model randomly sampling the row vectors across the 4 groups

Mean squared discrepancy over a range of weights ω

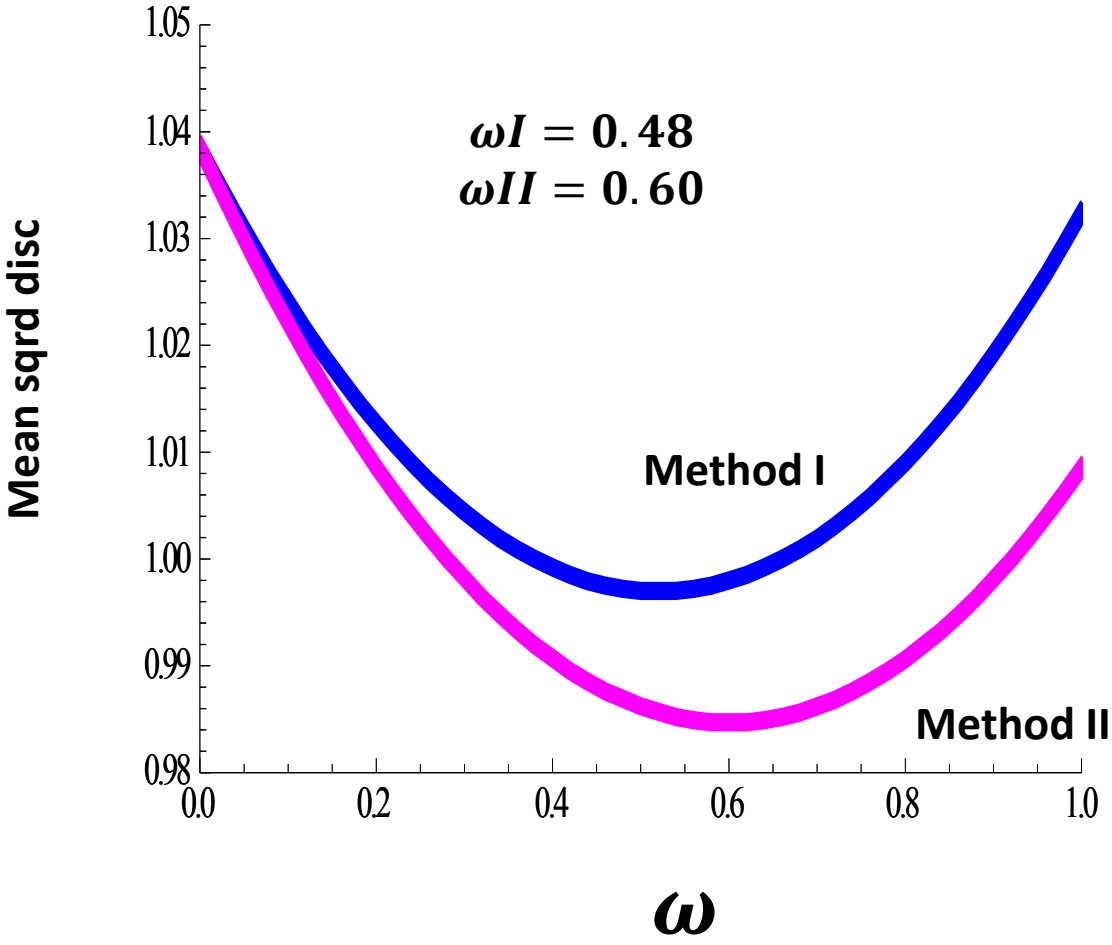


Illustration 2
Random Effects Meta-analysis

- ATLAS: The Assessment of Treatment with Lisinopril and Survival
  - N = 1,545 patients in K = 17 countries
  - Analysis of incremental costs (drug plus CV events)
  - Overall $\hat{\mu}$= -154.9 and $\hat{\tau}$ = 751.6

Economic Evaluation of Mean Costs in Multinational Clinical Trial (REMA, Random Effects Meta-Analysis)

- Heterogeneity among individual country estimates (τ^2)
- Overall mean net cost measure ($\hat{\mu}$)
- Country specific mean net costs ($\tilde{\theta}_i$)
- Net cost predictions for countries that did not participate in the trial ($\tilde{\theta}_{new}$)

Individual Country Estimates

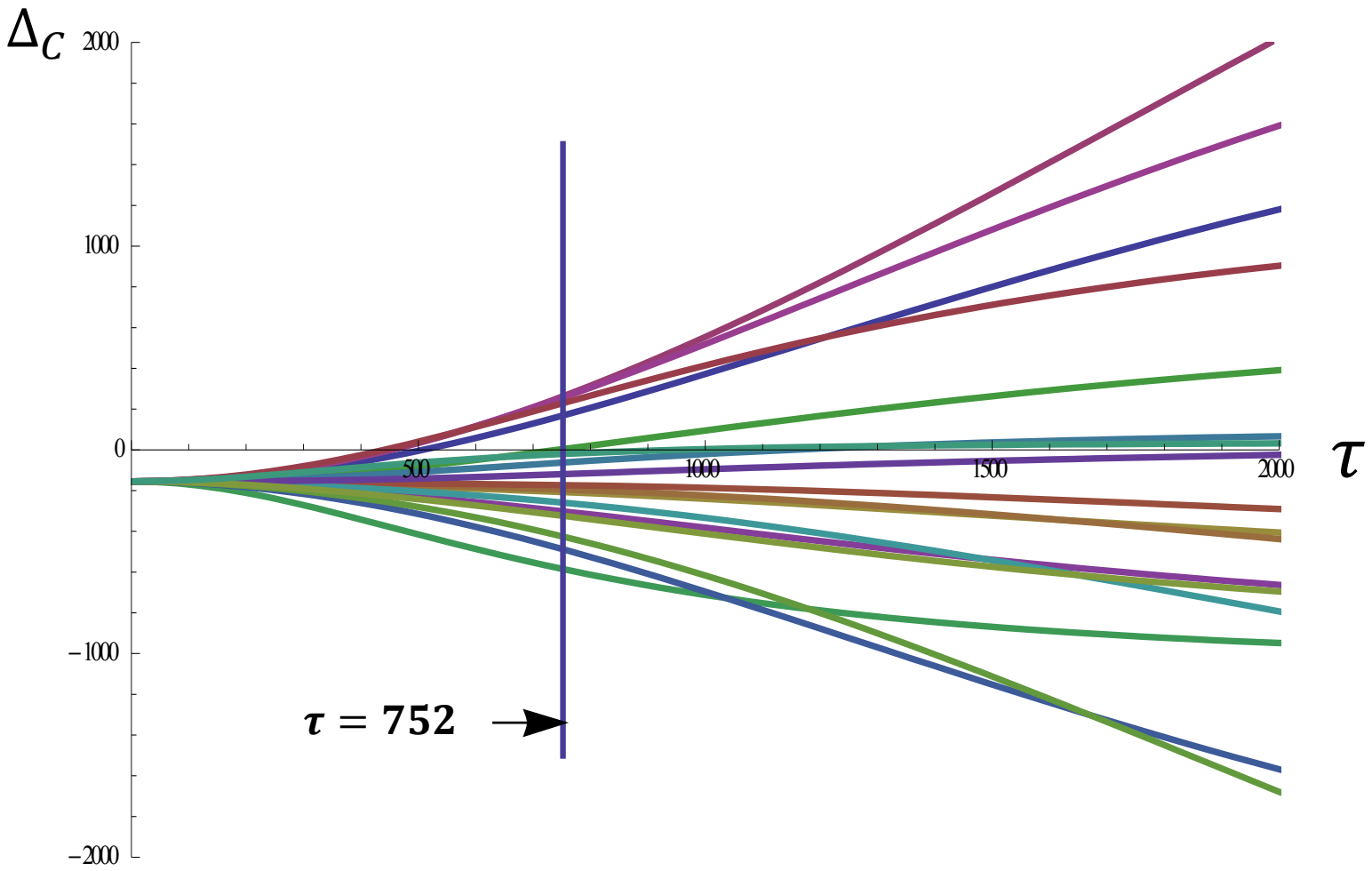
- Heterogeneity factor (τ^2) estimated using REMA
- "Shrinkage" factors based on variance components

$$\hat{\omega}_i = \frac{\hat{\sigma}_i^2}{(\hat{\sigma}_i^2 + \hat{\tau}^2)}$$

$$\tilde{\theta}_i = (1 - \hat{\omega}_i)\hat{\theta}_i + \hat{\omega}_i\hat{\mu}$$

❖Address uncertainty in τ^2 using Bayesian analysis and simulation

ATLAS Study: Estimated Mean Cost for each of 17 countries for increasing values of τ



Prediction

- Predictive ability is an important feature of economic models
- What exactly is being predicted?
  - Overall mean of a distribution?
  - What would country C have been had they been in the trial?
- There is an extra (within country) component of variance for each country that needs to be taken into account
- Using ATLAS the predictive ability of each model can be assessed by systematically predicting each country based on the data from the remaining countries

Validation/Model Checking for ATLAS 17 Countries

- The NREM, jackknife (JK), and Bayes models were checked by predicting each country individually based on the remaining 16
- 95% C.I. for NREM and JK covered 16 of 17 country averages
- 95% credible interval for Bayes prediction covered 16 of 17 country averages
- 50% internals for all methods covered 9 of 17 country averages
- 8 (9) of 17 observed averages were greater (less) than predictions

Predicting ATLAS Individual Country Mean Costs (Leave One Out)

Method	Standard Prediction Error	Predicted Pr(Average Cost < 0)
NREM	2888.2	0.56
Jackknife	2857.5	0.56
Bayes	2904.8	0.49

Actual # < 0 was 9/17 = 0.53

Prediction Error = $\sqrt{(\sum(\theta_i - \hat{\theta}_i)^2)/N}$

Classification and Prediction

- Classification and prediction are fundamental components of medical practice
  - Diagnosing the presence of disease
  - Patient disease subgroups
  - Predicting prognosis
  - Patient-specific treatments
  - Identifying at-risk subsets

- Interest in classification and prediction has increased recently in biopharmaceutical applications

- Two step process

  - **Discrimination** uses a learning data set of labeled observations to construct a classifier

  - **Classification** uses the measurements on a new unlabeled observation to predict the class

Summary of Existing Methods

- Pepe (2003) is almost entirely dedicated to ROC analysis for the single variable and 2 group case.
  - The single variable may be a function of several predictors e.g., a logistic risk score.
  - The application of 2 tests run in parallel or sequentially is discussed.
- Zhou (2002) considers the Sensitivity and Specificity of multiple tests.
- Krzanowski and Hand (2009) consider K classes and recommend either using K ROC curves for each class contrasted with all others, or using all K(K-1) pairwise ROC curves.
- The problem seems to be that ROC methodology does not lend itself easily to these expanded situations.
- Statistical learning approaches have addressed the problem of classification using parametric, nonparametric, and Bayesian methods.

Focus for Today

- General case for $K \geq 2$ class predictions based on $M \geq 1$ biomarkers
- Measures of distribution separation for multivariate densities
- Interpreting posterior probabilities of classification based on the Bayes classifier
- Strength of evidence method over M candidate biomarkers

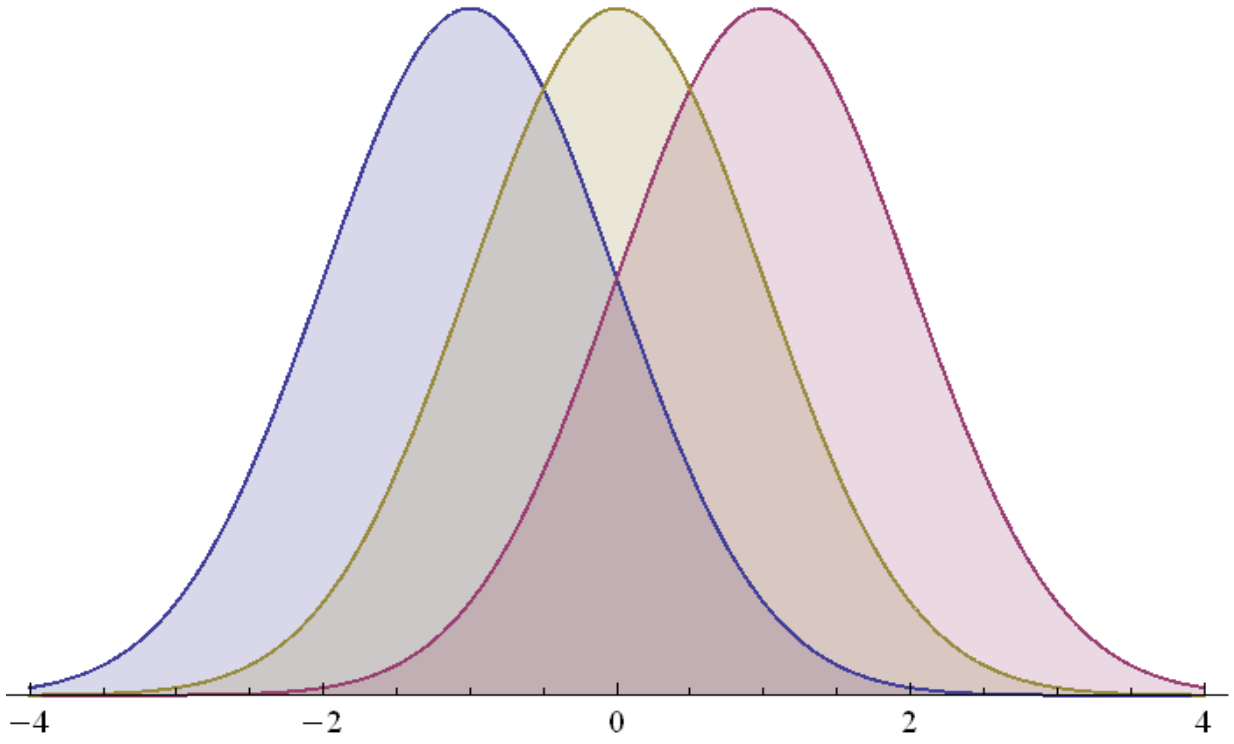
Measuring Multiple Distribution Similarity

- Lachenbruch et al (2004) proposed a measure of similarity between distributions for evaluating vaccine lot consistency
- For K groups

$$\gamma = \frac{1}{K}\int Max\left[f_j(x)\right]dx \text{ over } j = 1, \ldots, K$$

- γ is the probability that a random observation is assigned to the correct group
- γ ranges from $^1/_K$ when all distributions are identical to 1 if they are distinct
- **γ is not the AUROC when K= 2!**
- Estimation can assume normality or use NPKD (estimates can also be derived using the empirical distribution function)

Three Normal Distributions
($\mu_1 = -1, \mu_2 = 0, \mu_3 = 1, \sigma^2 = 1$)
$\Upsilon = 0.59$



Illustrative Values of ϒ when K= 3 Variance = 1

μ_1, μ_2, μ_3	ϒ	OVL
$\mu_1 = 0, \mu_2 = 0, \mu_3 = 0$	0.33	1.00
$\mu_1 = -0.5, \mu_2 = 0, \mu_3 = 0.5$	0.46	0.86
$\mu_1 = -1, \mu_2 = 0, \mu_3 = 1$	0.59	0.72
$\mu_1 = -2, \mu_2 = 0, \mu_3 = 2$	0.74	0.41
$\mu_1 = -3, \mu_2 = 0, \mu_3 = 3$	0.91	0.18
$\mu_1 = -4, \mu_2 = 0, \mu_3 = 4$	0.97	0.06

Lachenbruch's γ for Multivariate Densities

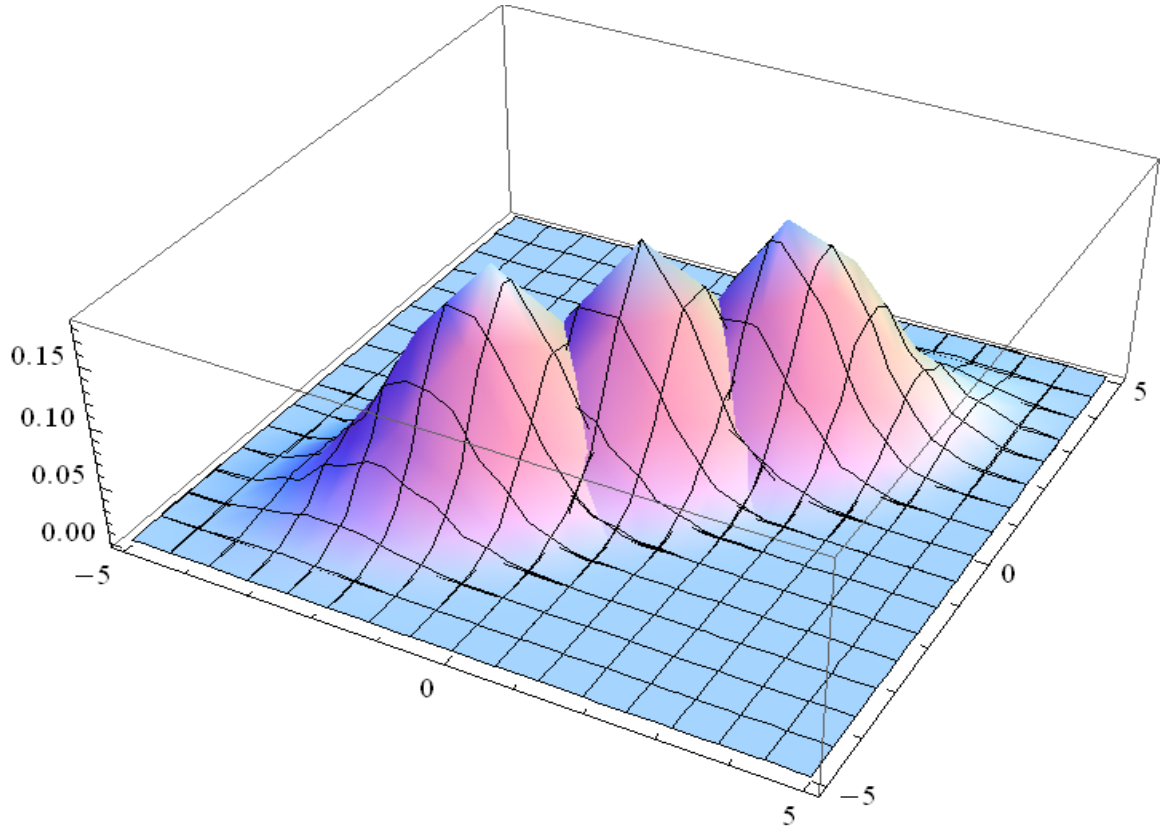
- For K multivariate densities of dimension M

$$\gamma = \frac{1}{K}\int \ldots \int Max\{f_k(x_1, \ldots, x_M)\}dx_1 \ldots dx_M$$

over $j = 1, \ldots, K$

- γ is the probability that a random M-dimensional observation is assigned to the correct group
- γ ranges from $1/K$ when all distributions are identical to 1 if they are distinct
- Estimation can use normality, MV NPKD, or empirical distribution functions
- **γ is not the AUROC when K= 2 and M= 1! γ is potentially useful as an AUROC type measure.**

Three Bivariate Normal Distributions
Means {(-1.5, -1.5), (0, 0), (1.5, 1.5)}
Variances 1 and Correlation 0.5
γ = 0.74



Illustrations of γ for 3 densities and varying means (variance = 1 and correlation = 0.5)

Mean vectors $\{(-\mu,-\mu),(0,0),(\mu,\mu)\}$	γ
$\mu = 0$	0.33
$\mu = 0.25$	.41
$\mu = 0.5$	.48
μ =1.0	.62
μ =1.5	.74
μ =2.0	.83

For $M \geq 1$ Biomarkers and $K \geq 2$ Classes

- Most methods construct a single summary using a linear combination over the $M > 1$ variables
  - Linear/Quadratic discriminant analysis
  - Multiple regression
  - Principle components analysis
  - Logistic regression
- For $K > 2$ classes analysis uses sets of pairwise classifications
  - Each class compared to all others
  - All pairwise classifications
- Bayes classifier can handle $K \geq 2$ and $M \geq 1$

Bayesian Framework

- For a given value of x define

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_j \pi_j f_j(x)} \text{ for } k = 1, \ldots, K$$

  where $f_k(x)$ is the value of the density function for group k and may be multivariate

- The densities can be estimated using a Gaussian assumption, multivariate nonparametric kernel density estimation, or K-nearest neighbor
- Recognize that the $p_k(x)$ are posterior probabilities

Bayes Classification Methods

- Method 1: Compute single multivariate density estimate $\tilde{P}_k(x)$
- Method 2:
  - Compute $\hat{P}_k(x_m)$ for each of the $k = 1, \ldots, K$ classes and each of the $m = 1, \ldots, M$ variables in $x = (x_1, \ldots, x_M)$
  - Synthesize the $\hat{P}_k(x_m)$ using Fisher's or Stouffer's method into a single $\tilde{P}_k(x)$
- Classify observation x to group J

$$J = Max\{\tilde{P}_k(x)\}, k = 1, \ldots, K$$

Synthesis of Evidence: Method II

1. Compute $\tilde{p}_k(x_m) = \frac{\pi_k f_k(x_m)}{\sum_j \pi_j f_j(x_m)}$ for each of the $k = 1, \ldots, K$ classes and each of the $m = 1, \ldots, M$ variables in $x = (x_1, \ldots, x_M)$
2. Compute $\tilde{Z}_j = \frac{1}{\sqrt{M}} \sum_m \Phi^{-1}(\tilde{p}_k(x_m))$ for each of the classes. $\Phi^{-1}(\,)$ is the inverse normal function.
3. Compute $\tilde{P}_k = \Phi(\tilde{Z}_k)$
4. Method II: Classify observation x to the group J where $J = Max\{\tilde{P}_k\}$ over $k = 1, \ldots, K$

Example: Wisconsin Diagnostic Breast Cancer (Woberg et al., 1994)

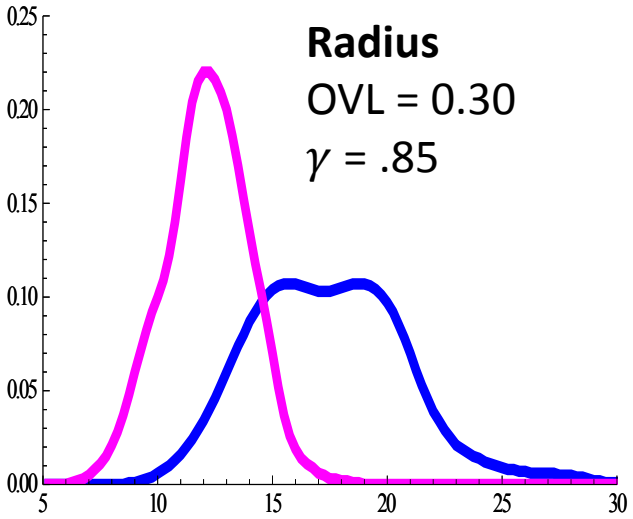
- Breast cancer features for 569 women with benign (N = 357) or malignant (N=212) diagnosis
- Ten real-valued features are computed for each cell nucleus:
  1. radius (mean of distances from center to points on the perimeter)
  2. texture (standard deviation of gray-scale values)
  3. perimeter
  4. area
  5. smoothness (local variation in radius lengths)
  6. compactness (perimeter^2 / area - 1.0)
  7. concavity (severity of concave portions of the contour)
  8. concave points (number of concave portions of the contour)
  9. symmetry
  10. fractal dimension ("coastline approximation" - 1)

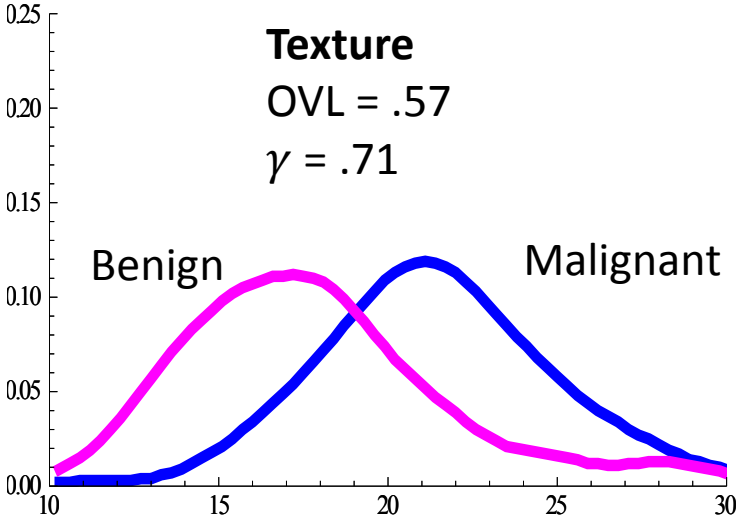
Methods

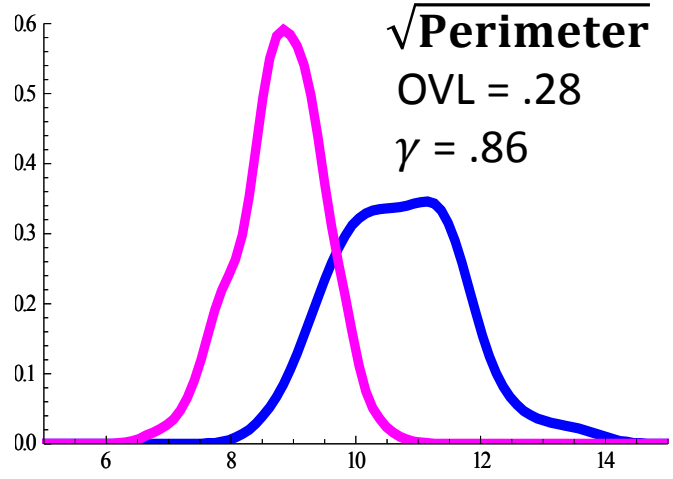
- Individual classification of 569 subjects to Π_1= Benign or Π_2 = Malignant based on the remaining 568 subjects
- Four variables
  - Radius
  - Texture
  - SQRT(Perimeter)
  - CBRT(Area)
- Prevalence estimates π_1 = 0.627 and π_2 = 0.373
- Bayes posterior probability classification (Method I)
  - Assuming Gaussian sampling
  - Multivariate NPKD
- Individual variable classification (Method II)
  - Assuming Gaussian sampling
  - Multivariate NPKD

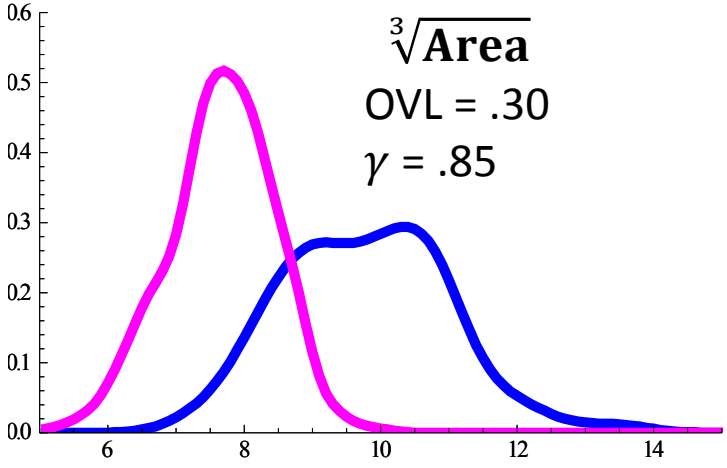
WDBC Study Data

Benign (N = 357) and Malignant (N=212)









Performance of Classifier

- The classification is summarized in the form of an M x M confusion matrix

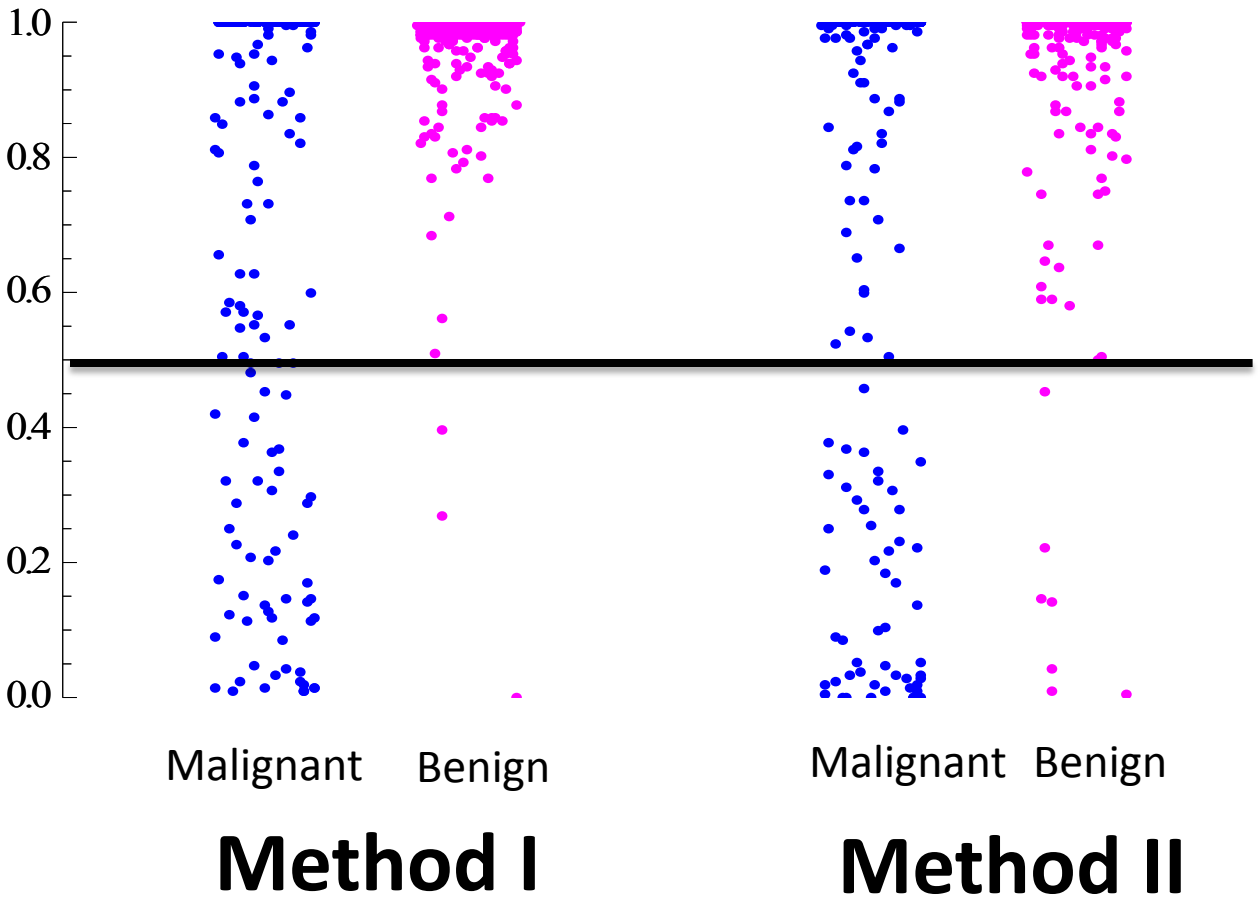
$$C = \{c_{ij}\}$$

where c_{ij} is the number of validation samples assigned to class i which were actually class j

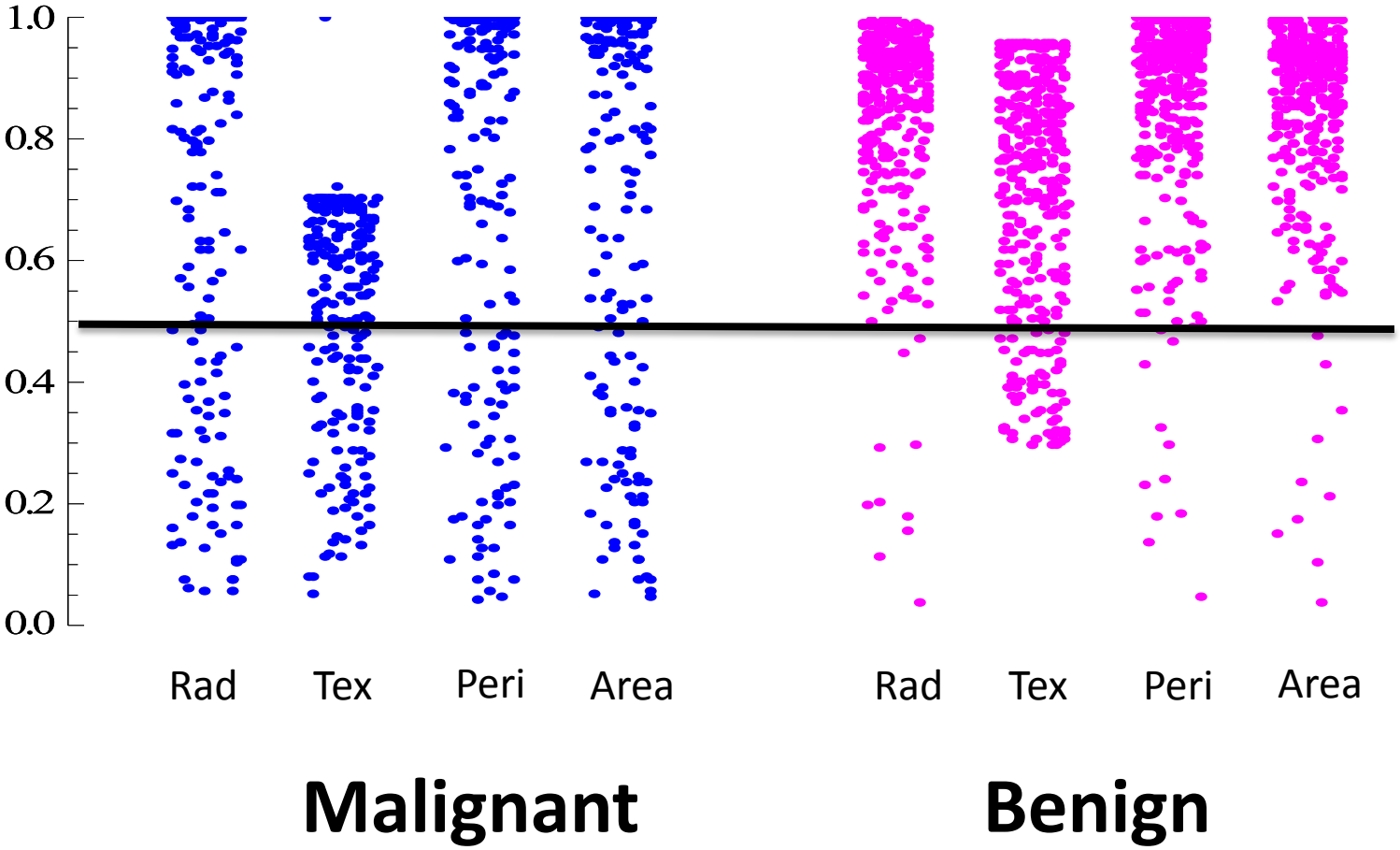
- Proportion of Correct Classification

$$\frac{1}{N}\sum_{m=1}^{M} c_{mm}$$

Probability of Correct Classification WDBC Data MV NPKD



Probability of Correct Classification WDBC Data Using 4 Variables MV NPKD



Concluding Remarks

- Prediction methods are quickly becoming the mainstay of precision medicine applications
- Basing statistical inference on "observables" offers many advantages in these applications
  - More direct connection with the decision or objective of the analysis
  - More objective basis for model validation
  - Better capability for comparing models
- Methods and software exist to address problems of prediction, and to use prediction based estimation for inference
- While the gold standard methods involve using separate training and test/validation data sets, cross-validation and leave out k methods can be applied

References

- Brieman L: Statistical modeling: The two cultures. *Statistical Science* 2001, Vol. 16, 199-231.
- Geisser S. *Predictive Inference: An Introduction*. New York: CRC Press, 1993.
- Giacoletti KED and Heyse JF: Using proportion of similar response to evaluate correlates of protection for vaccine efficacy. Statistical Methods in Medical Research 2011.
- Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York: Springer, 2013.
- Heyse J, Chen J, Cook J: Use of multinational randomized clinical trials in economic evaluations of health care. *Multiregional Clinical Trials for Simultaneous Global New Drug Devalopment* (Chapman & Hall/CRC Biostatistics Series, Chapter 24, 2016.
- Lachenbruch PA, Rida W and Kou J: Lot consistency as an equivalence problem. *Journal of Biopharmaceutical Statistics* 2004:14(2):275-290.
- Pepe M: The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press 2004.
- Pepe MS: Evaluating technologies for classification and prediction in medicine. Statistics in Medicine 2005; 24:3687-3696.
- Wolberg WH, Street WN , and Mangasarian OL. Machine learning techniques to diagnose breast cancer from fine-needle aspirates. Cancer Letters 77 (1994) 163-171.
- Zhou X-H, Obuchowski NA and McClish DK: *Statistical Methods in Diagnostic Medicine*. Wiley-Interscience, New York 2002.
- Zou KH, Hall WJ and Shapiro DE: Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* 1997; 16:2143-2156.