Bayesian Statistics

ddebarr@uw.edu

2016-05-05

Happy Cinco de Mayo! Celebrating victory at the Battle of Puebla and Mexican Heritage!

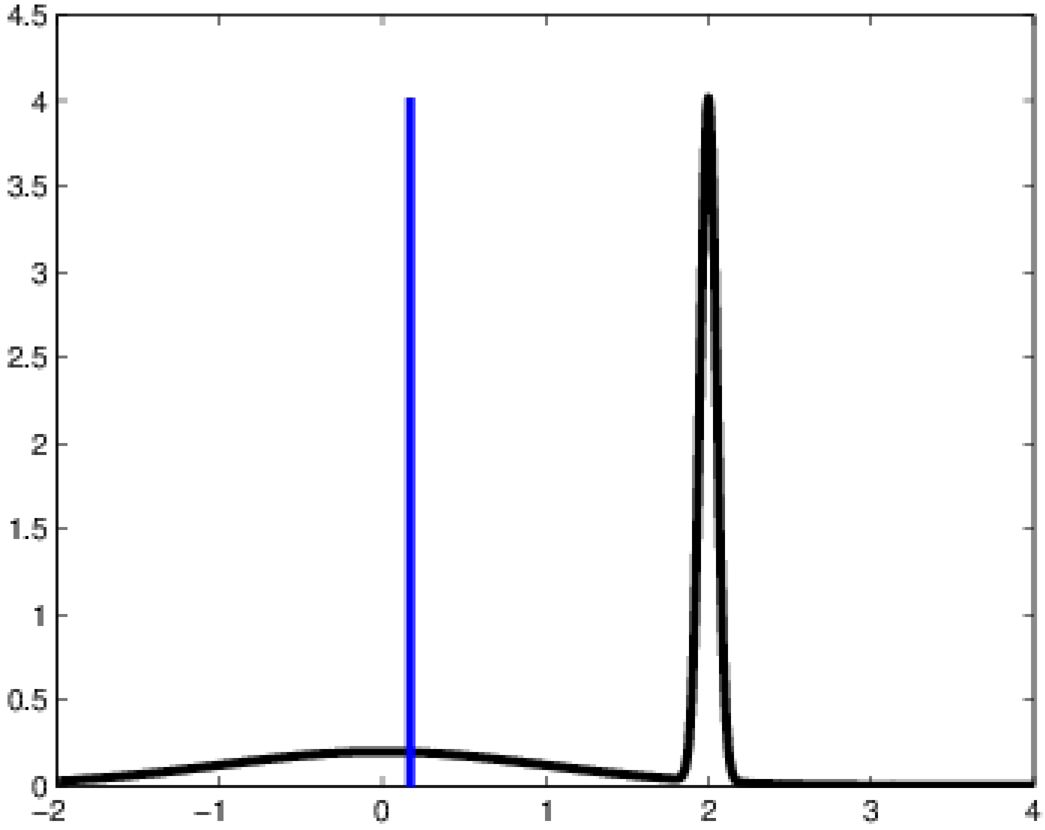
Agenda

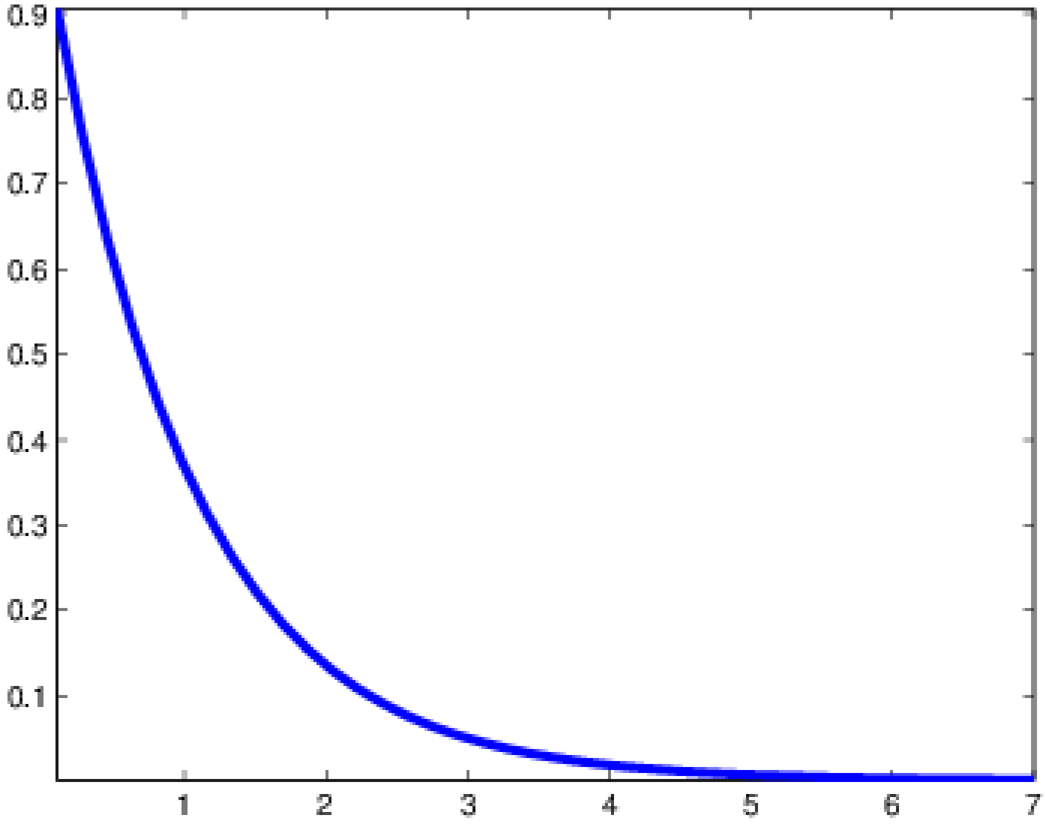
- Summarizing Posterior Distributions
- Bayesian Model Selection
- Priors
- Hierarchical Bayes
- Empirical Bayes
- Bayesian Decision Theory

Summarizing Posterior Distributions

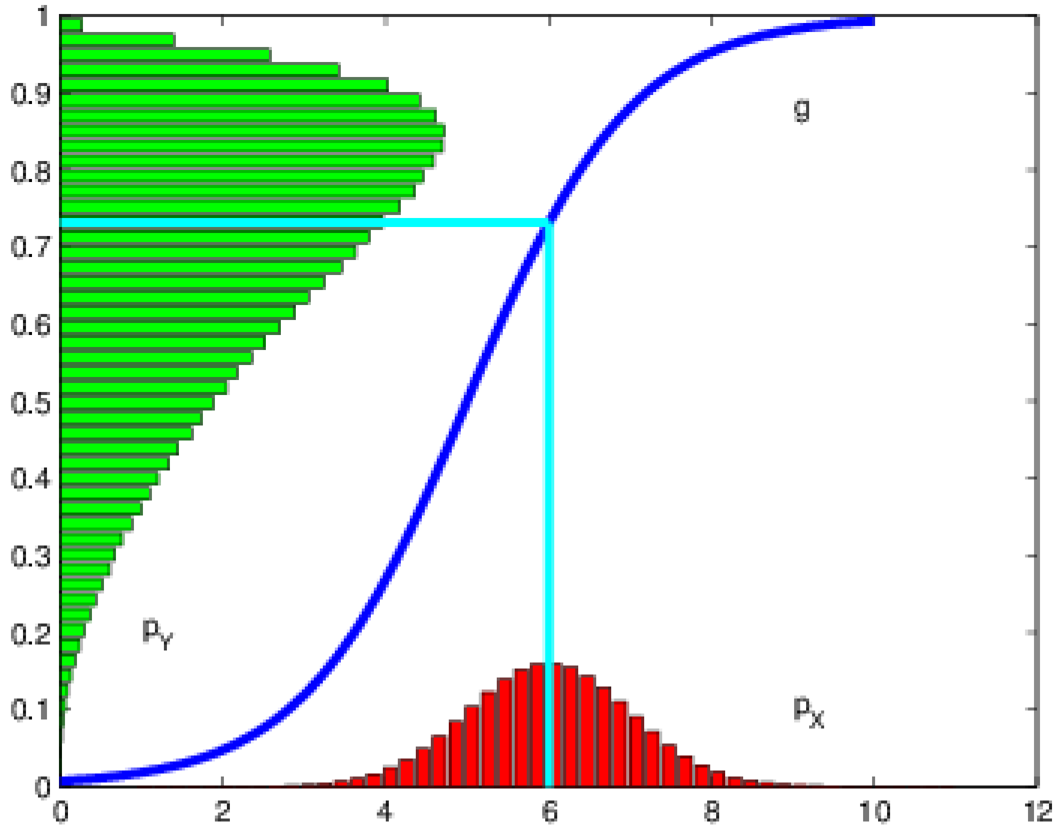
- Point Estimates
  - Mode
  - Mean
  - Median
- Interval Estimates
  - Central Interval
  - Highest Posterior Density Region

Mean Versus Mode



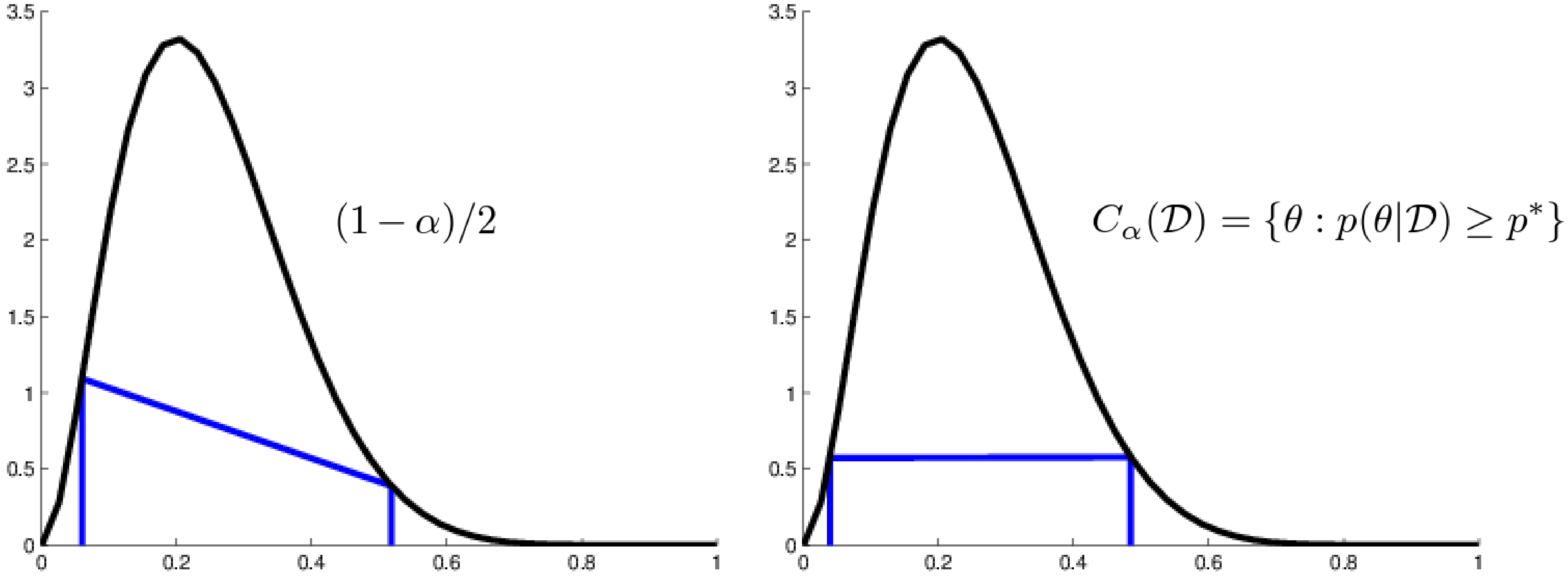


Transformation of a Mode



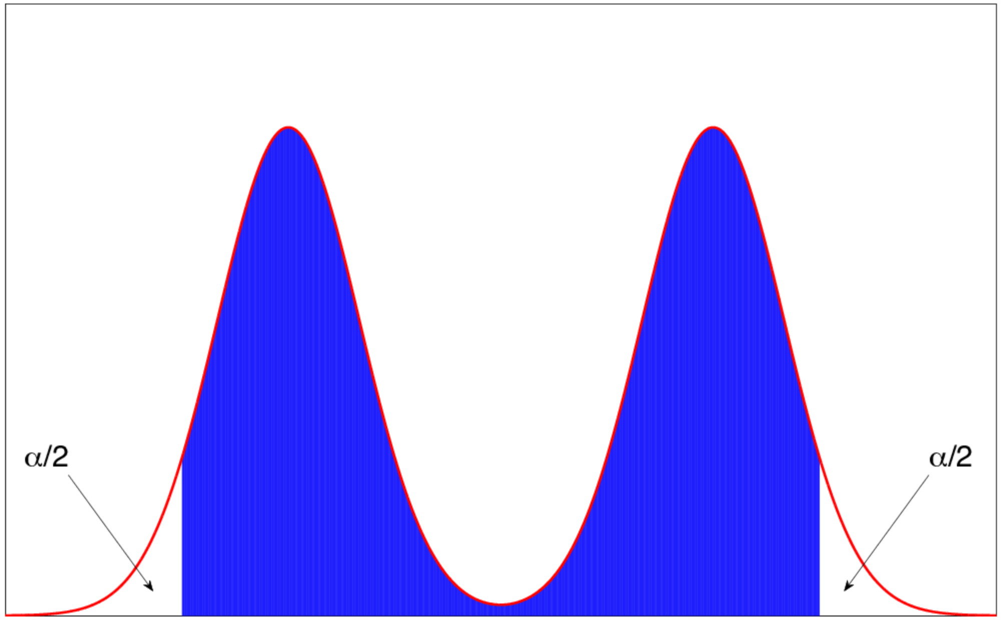
MAP estimation is not invariant to reparameterization; e.g. y = 1 / (1 + exp(-x + 5))

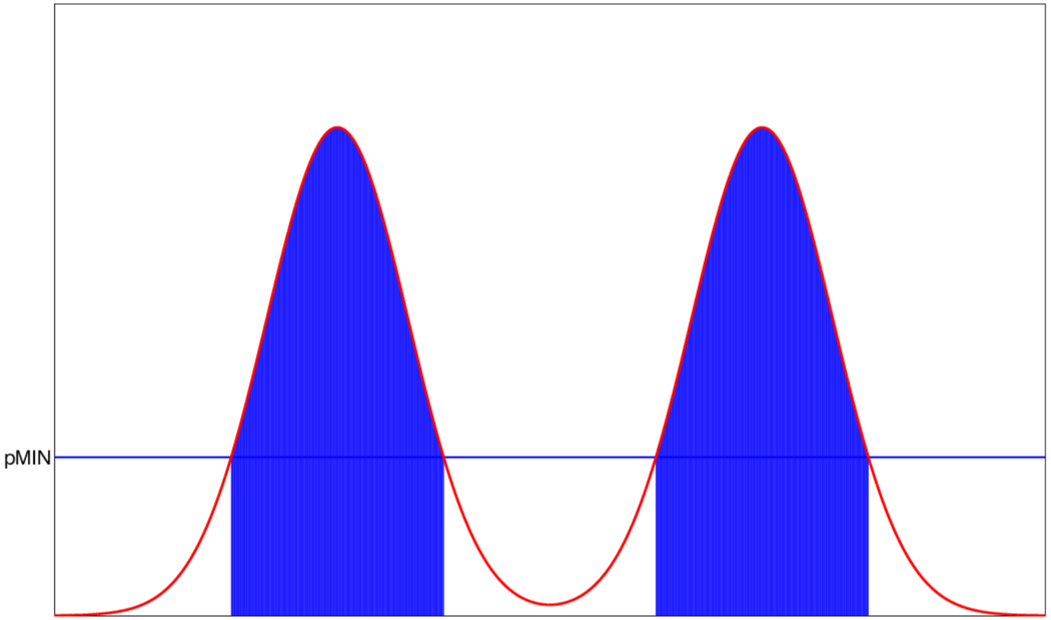
Central Interval versus Highest Posterior Density Region



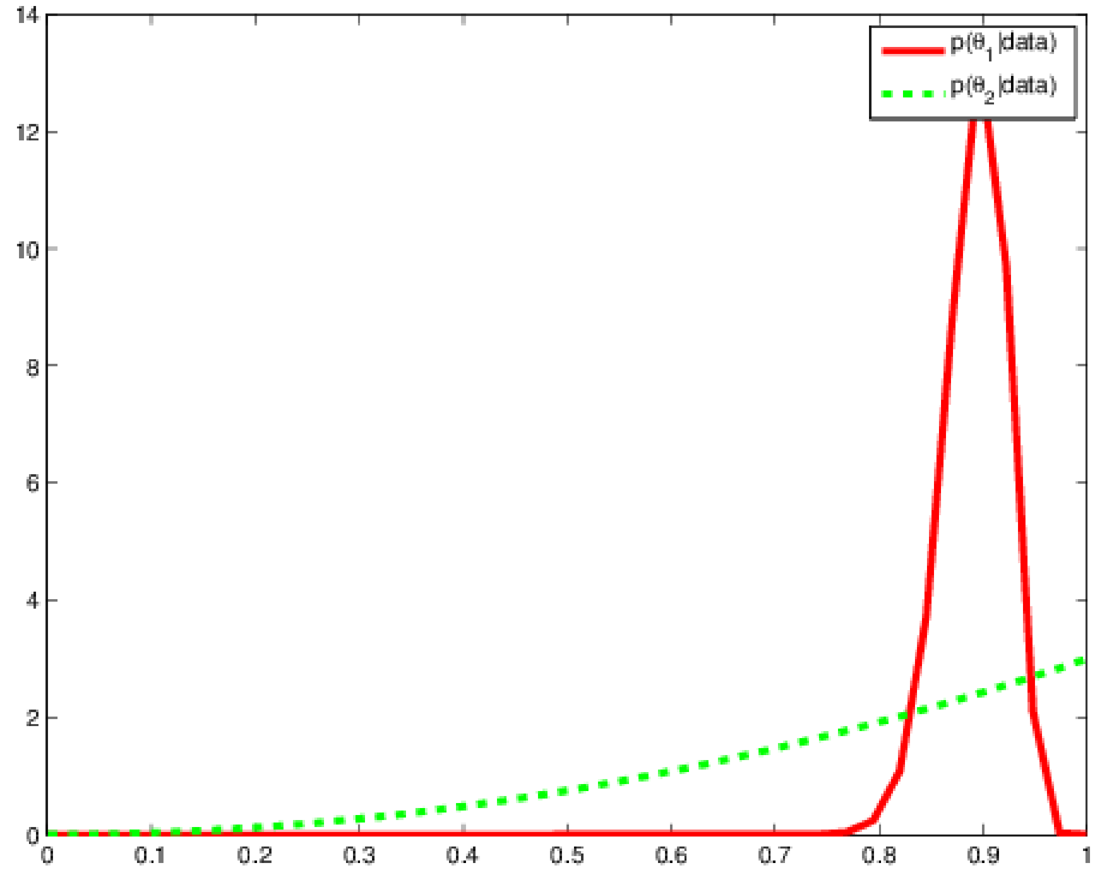
$$C_\alpha(\mathcal{D}) = (\ell, u) : P(\ell \leq \theta \leq u|\mathcal{D}) = 1 - \alpha$$

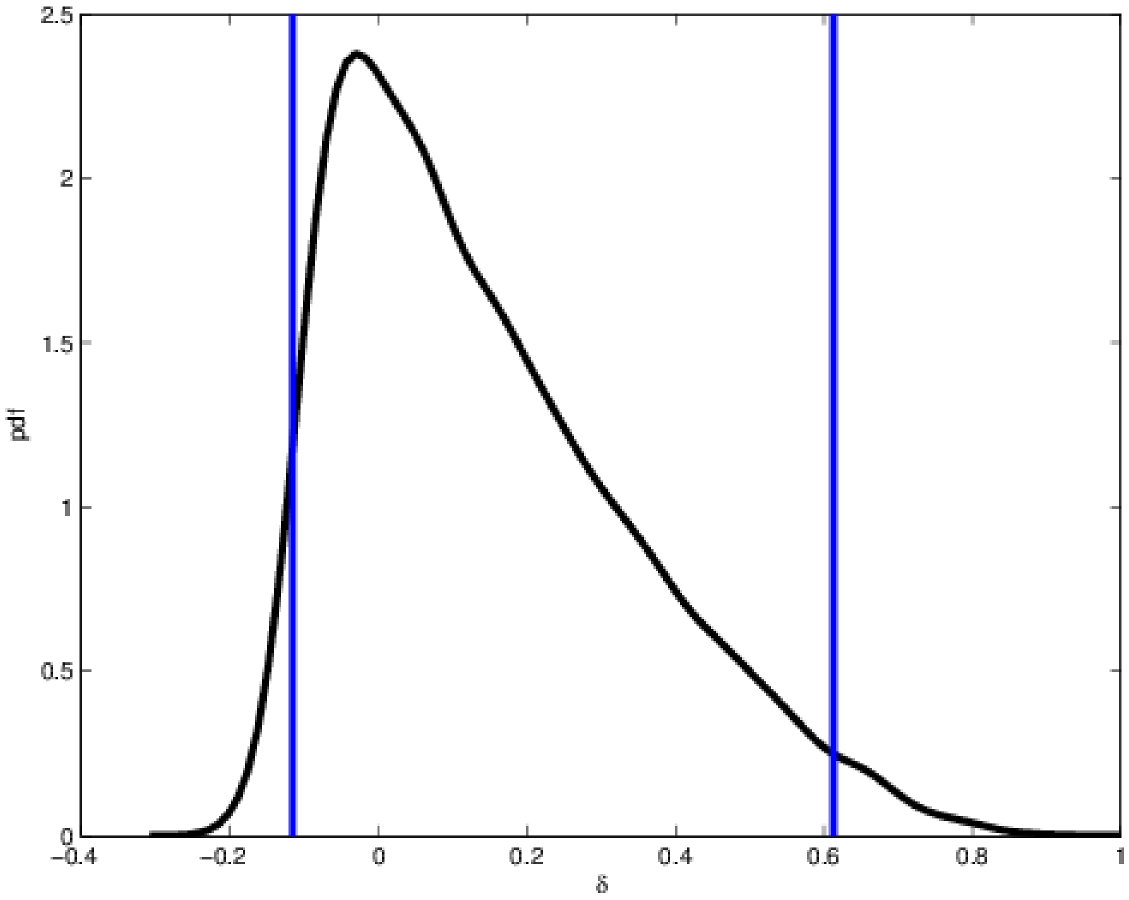
Central Interval versus Highest Posterior Density Region For a MultiModal Distribution





Comparing Two Proportions



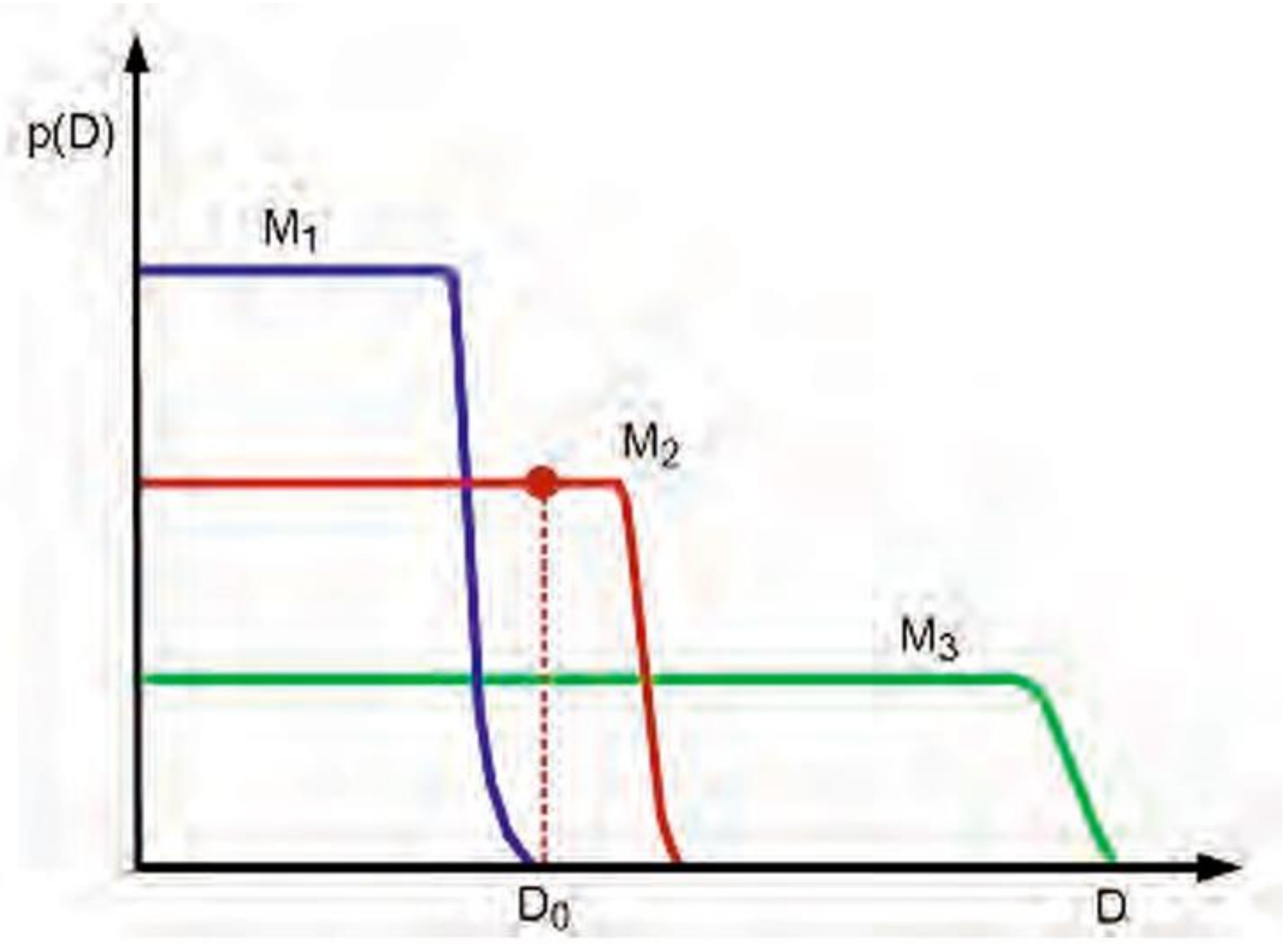


$\mathrm{Beta}(91, 11)$ versus $\mathrm{Beta}(3, 1)$

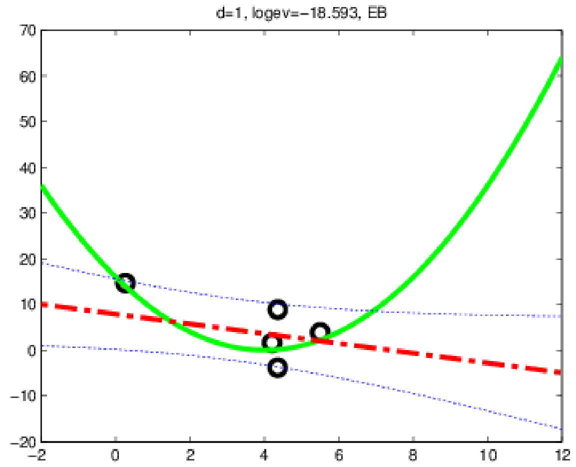
$p(\theta_1 > \theta_2 | \mathcal{D})$

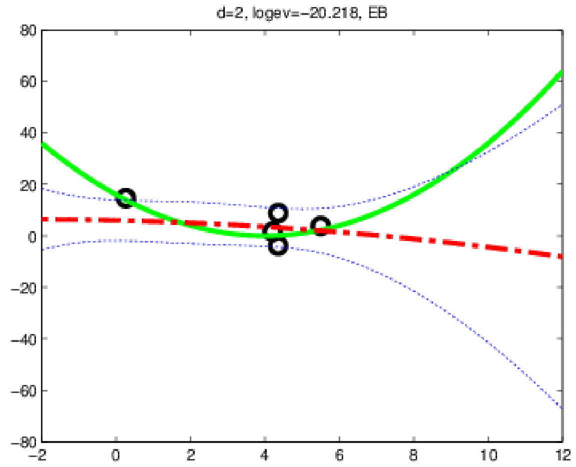
Bayesian Occam's Razor

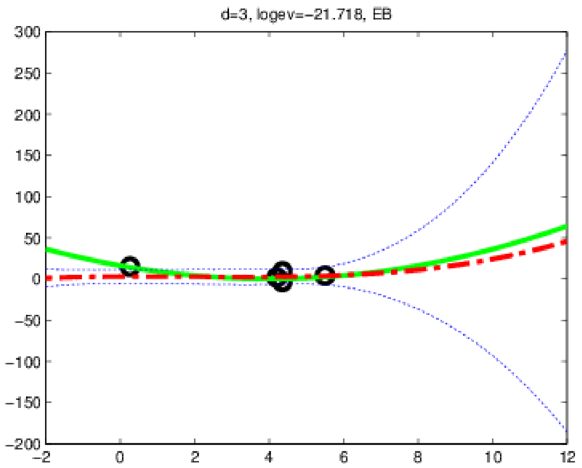
$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m\in\mathcal{M}} p(m,\mathcal{D})}$$

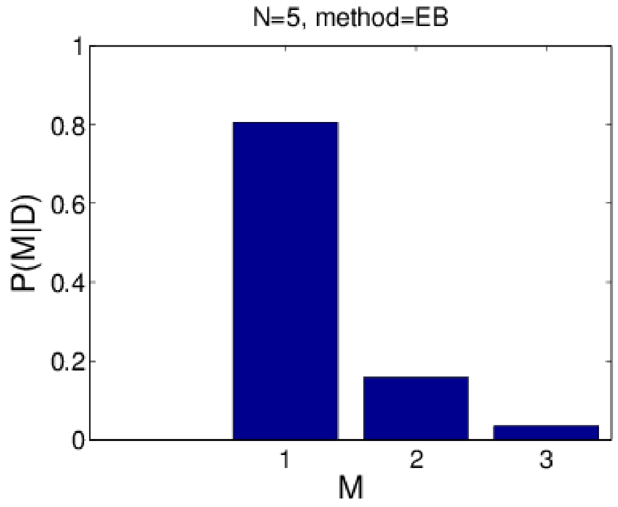


Model Selection Example: n = 5

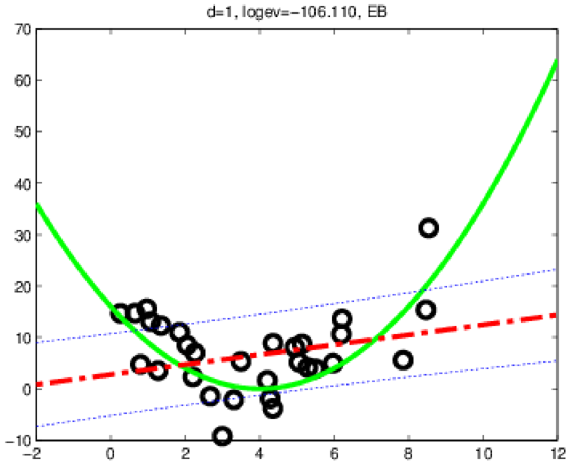


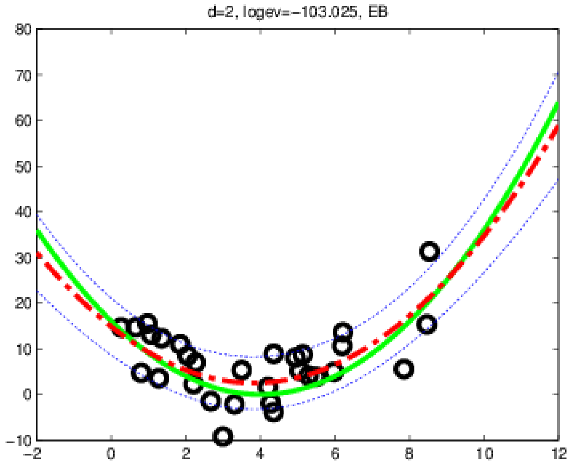


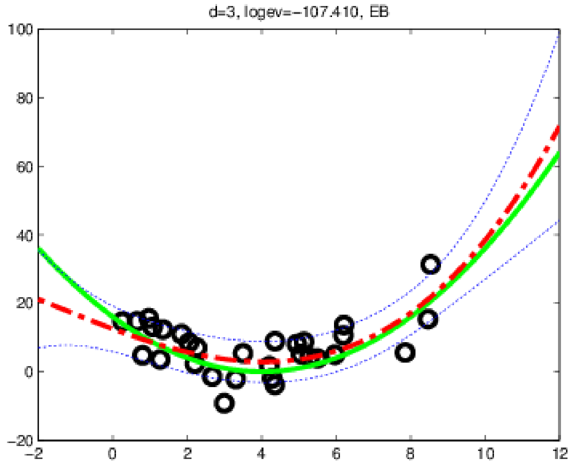


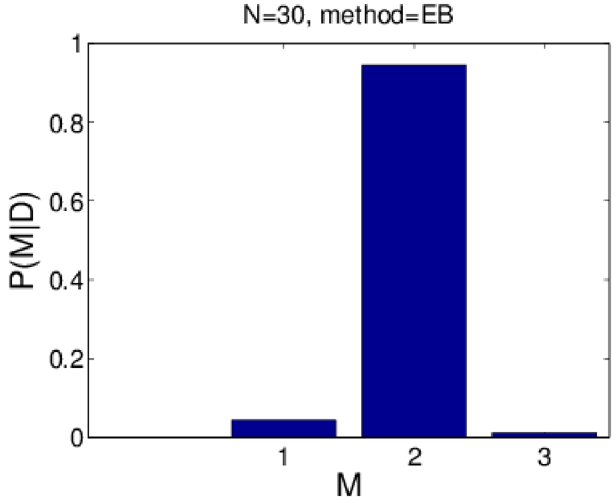


Model Selection Example: n = 30









Marginal Likelihood for Beta-Binomial Model

$$p(\mathcal{D}) = \binom{N}{N_1} \frac{B(a+N_1, b+N_0)}{B(a,b)}$$

We're able to simply add exponents for the probability of success for the prior and likelihood. Ditto for the probability of failure for the prior and likelihood.

BIC Approximation to Log Marginal Likelihood

BIC: Bayesian Information Criterion

$$\mathrm{BIC} \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \frac{\mathrm{dof}(\hat{\boldsymbol{\theta}})}{2} \log N \approx \log p(\mathcal{D})$$

For linear regression …

$$\log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) = -\frac{N}{2} \log(2\pi\hat{\sigma}^2) - \frac{N}{2}$$

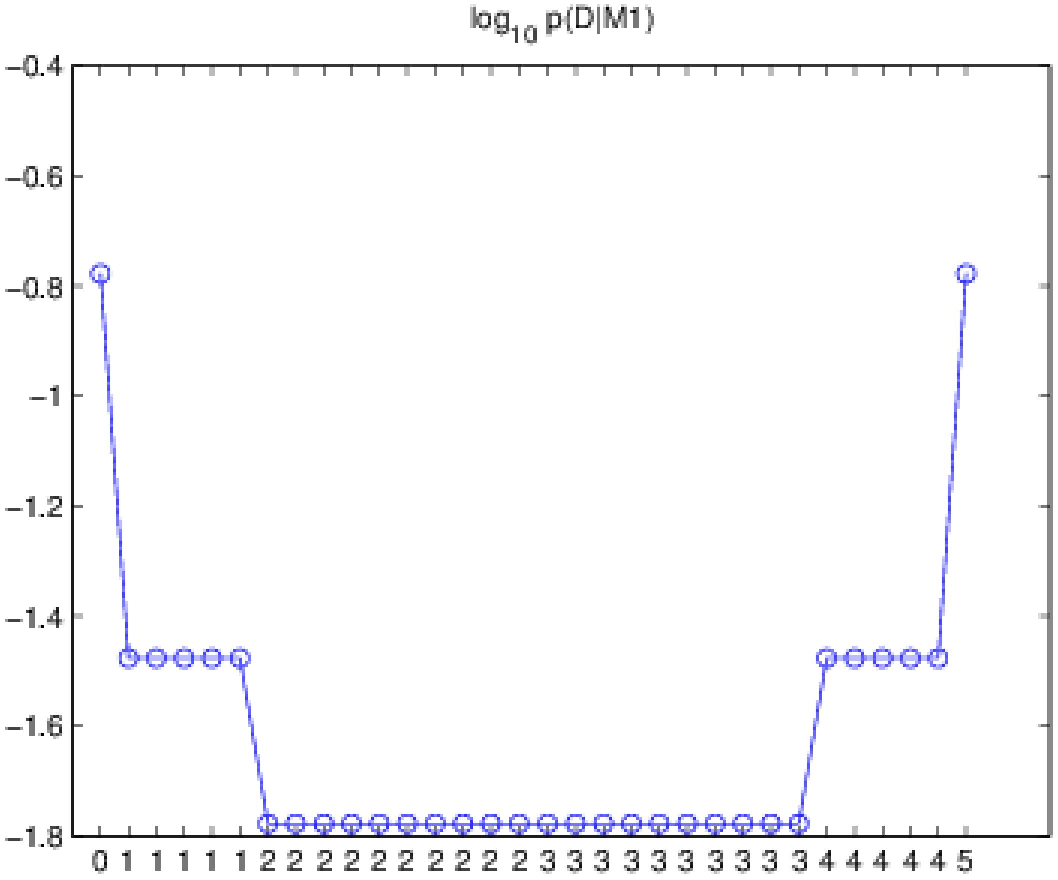
$$\mathrm{BIC} = -\frac{N}{2} \log(\hat{\sigma}^2) - \frac{D}{2} \log(N)$$

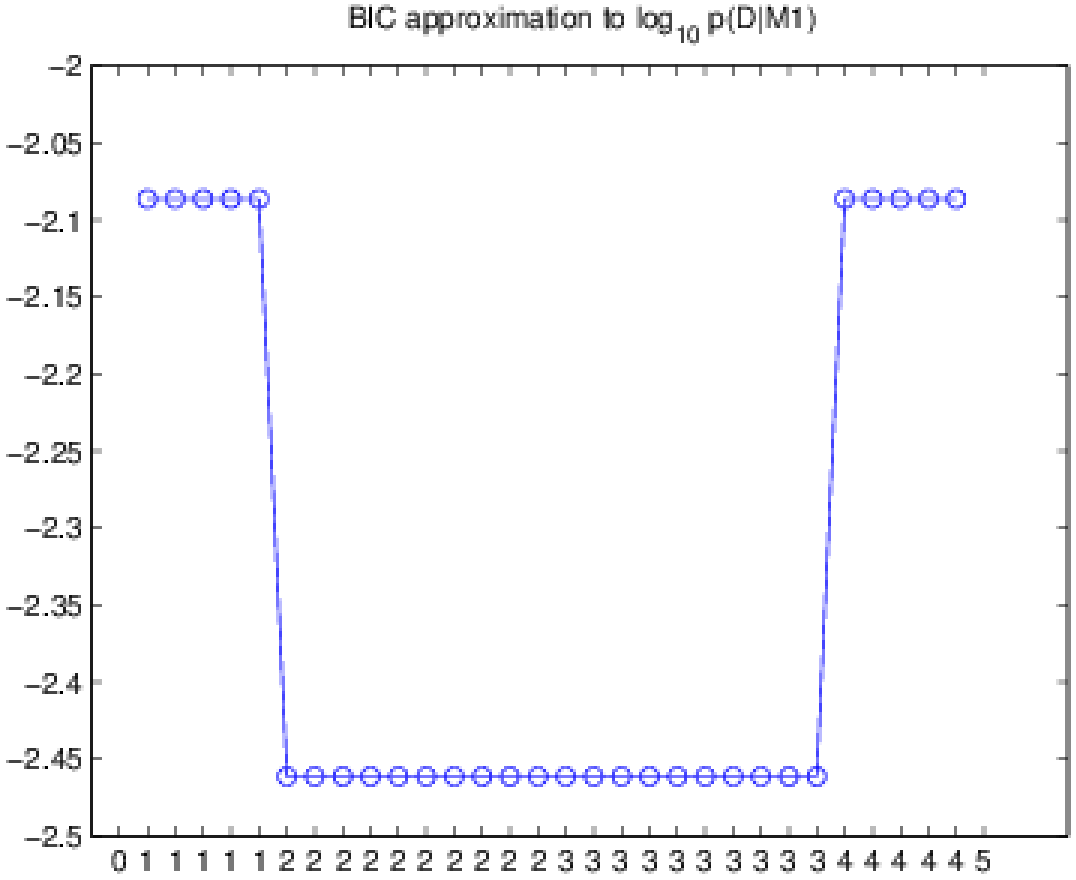
Bayes Factor

$$BF_{1,0} \triangleq \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \frac{p(M_1|\mathcal{D})}{p(M_0|\mathcal{D})} \Big/ \frac{p(M_1)}{p(M_0)}$$

Bayes factor $BF(1,0)$	Interpretation
$BF < \frac{1}{100}$	Decisive evidence for M_0
$BF < \frac{1}{10}$	Strong evidence for M_0
$\frac{1}{10} < BF < \frac{1}{3}$	Moderate evidence for M_0
$\frac{1}{3} < BF < 1$	Weak evidence for M_0
$1 < BF < 3$	Weak evidence for M_1
$3 < BF < 10$	Moderate evidence for M_1
$BF > 10$	Strong evidence for M_1
$BF > 100$	Decisive evidence for M_1

Marginal Likelihood for 5 Coin Tosses





$$p(\mathcal{D}|M_1) = \int p(\mathcal{D}|\theta)p(\theta)d\theta = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)}$$

Jeffreys Prior for Bernoulli and Multinoulli

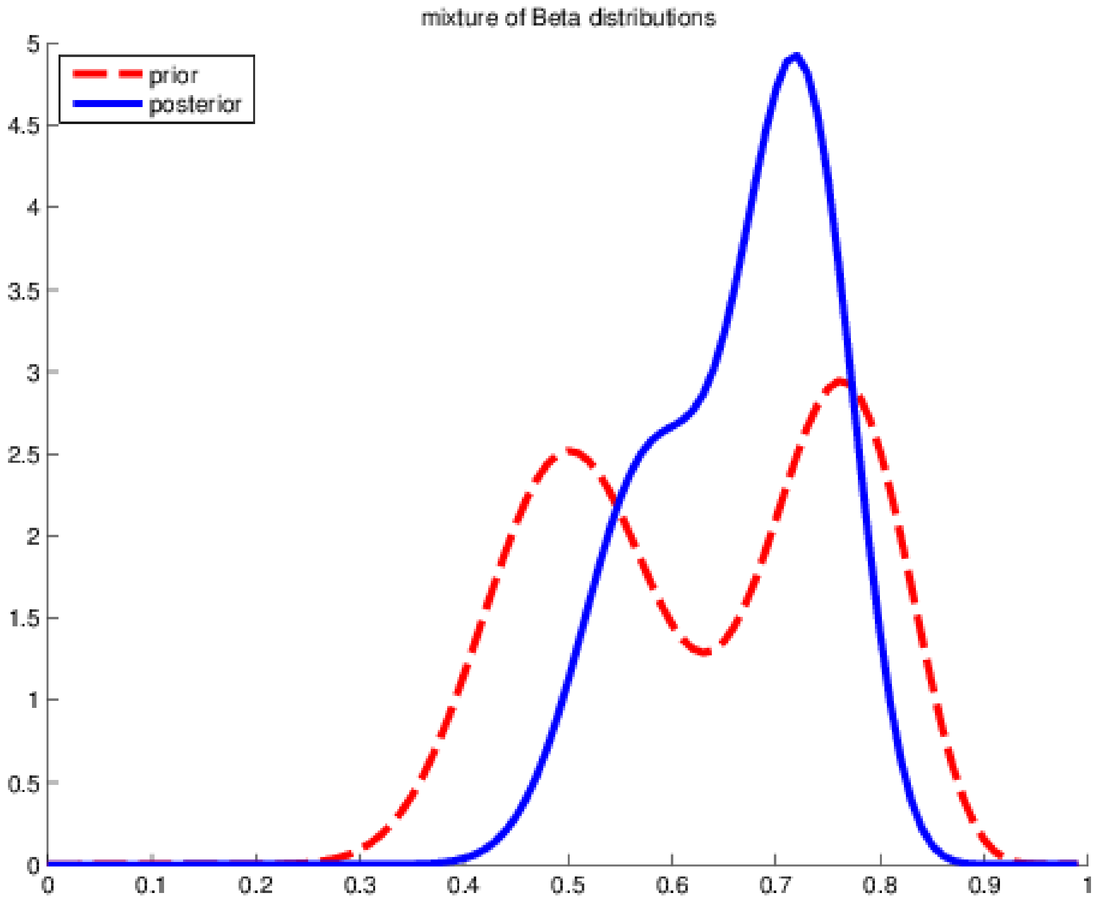
- Bernoulli

$$p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} = \frac{1}{\sqrt{\theta(1-\theta)}} \propto \mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$$

- Multinoulli

$$p(\boldsymbol{\theta}) \propto \mathrm{Dir}(\frac{1}{2}, \ldots, \frac{1}{2})$$

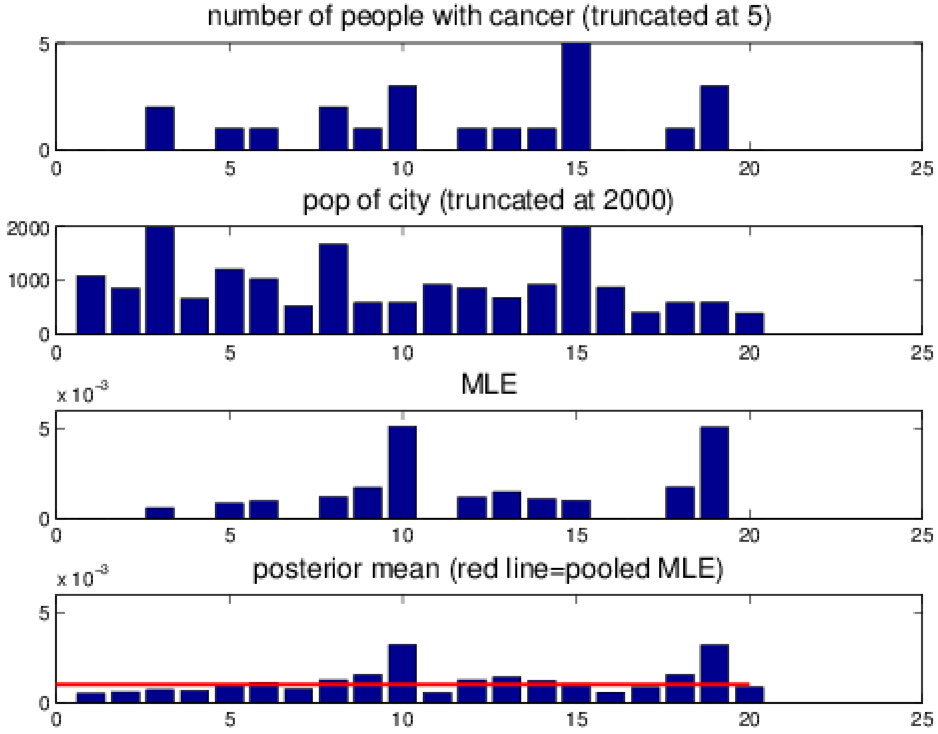
Mixture of Two Beta Distributions

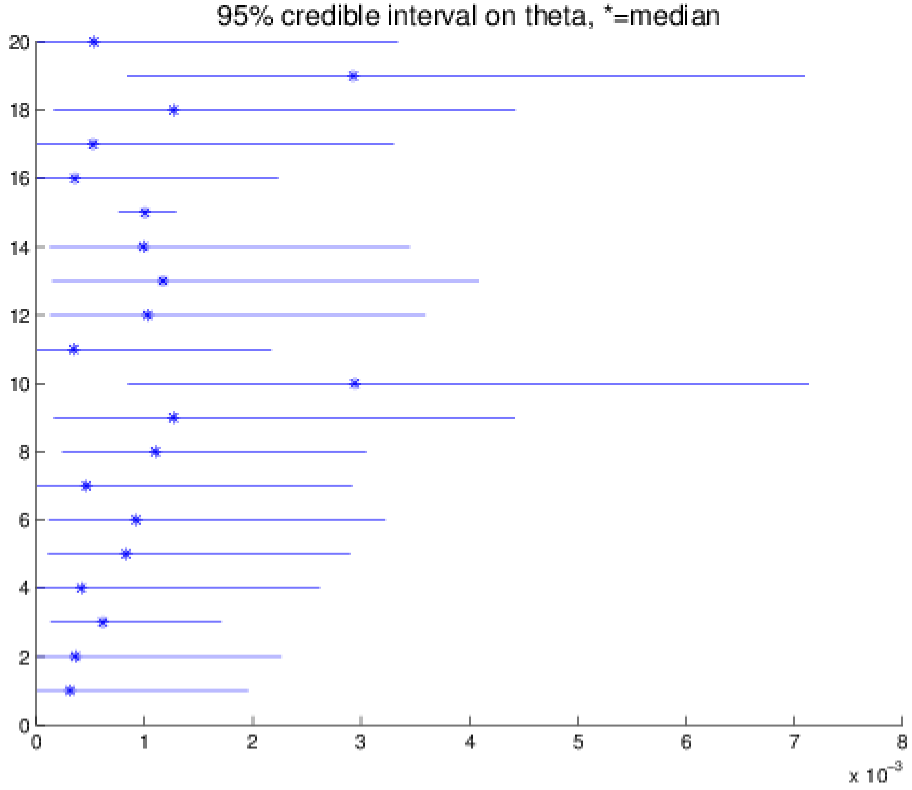


$$p(\theta) = 0.5 \text{ Beta}(\theta|20, 20) + 0.5 \text{ Beta}(\theta|30, 10)$$

Modeling Cancer Rates

Beta-Binomial Example





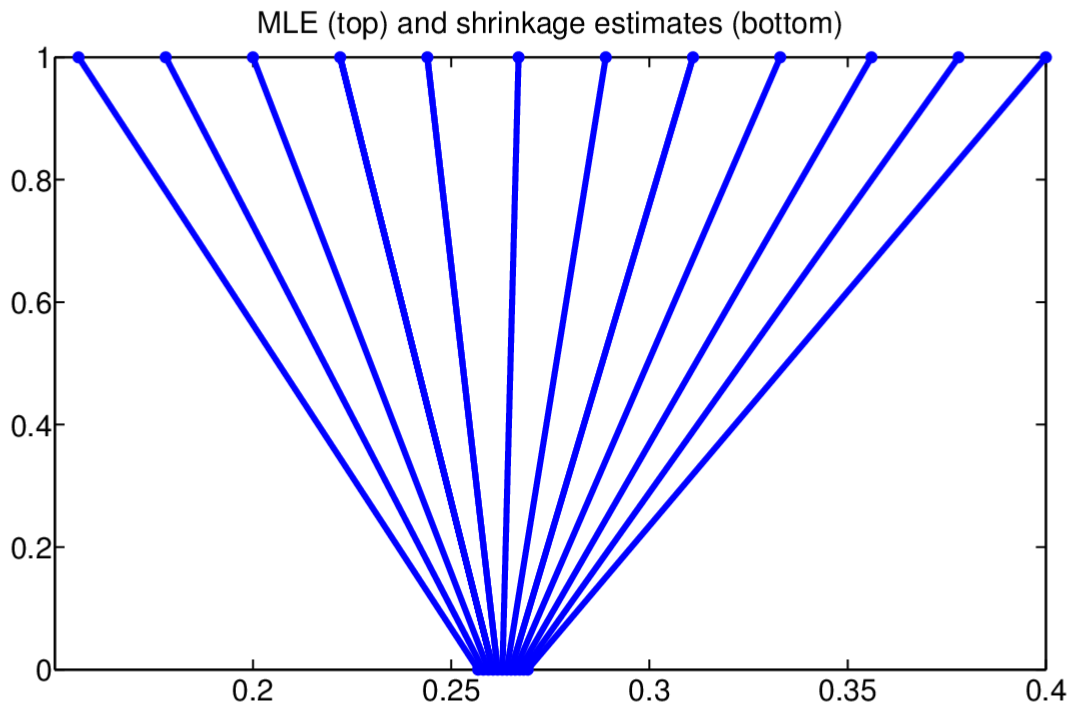
$$\eta \rightarrow \theta \rightarrow \mathcal{D}$$

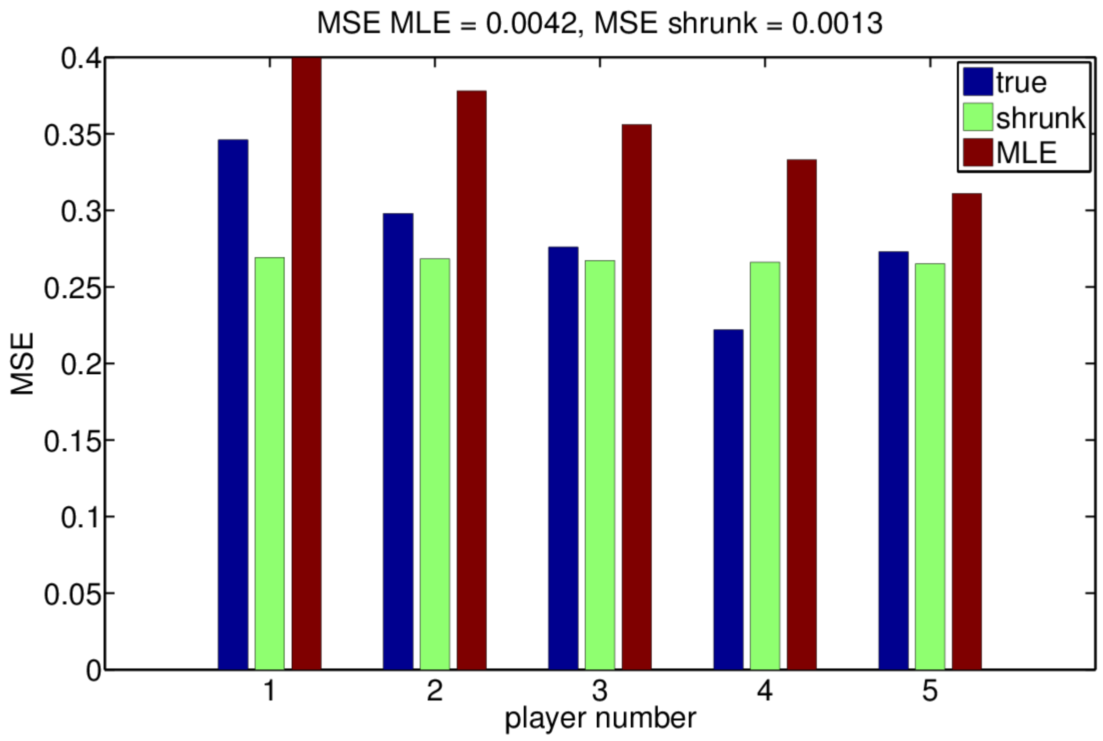
eta -> theta -> data ☺

hierarchical: we're using another model to "shrink" the posterior rates toward the pooled MLE

Modeling Batting Averages

Gaussian-Gaussian Example





empirical: we're using data to estimate priors

Loss Functions

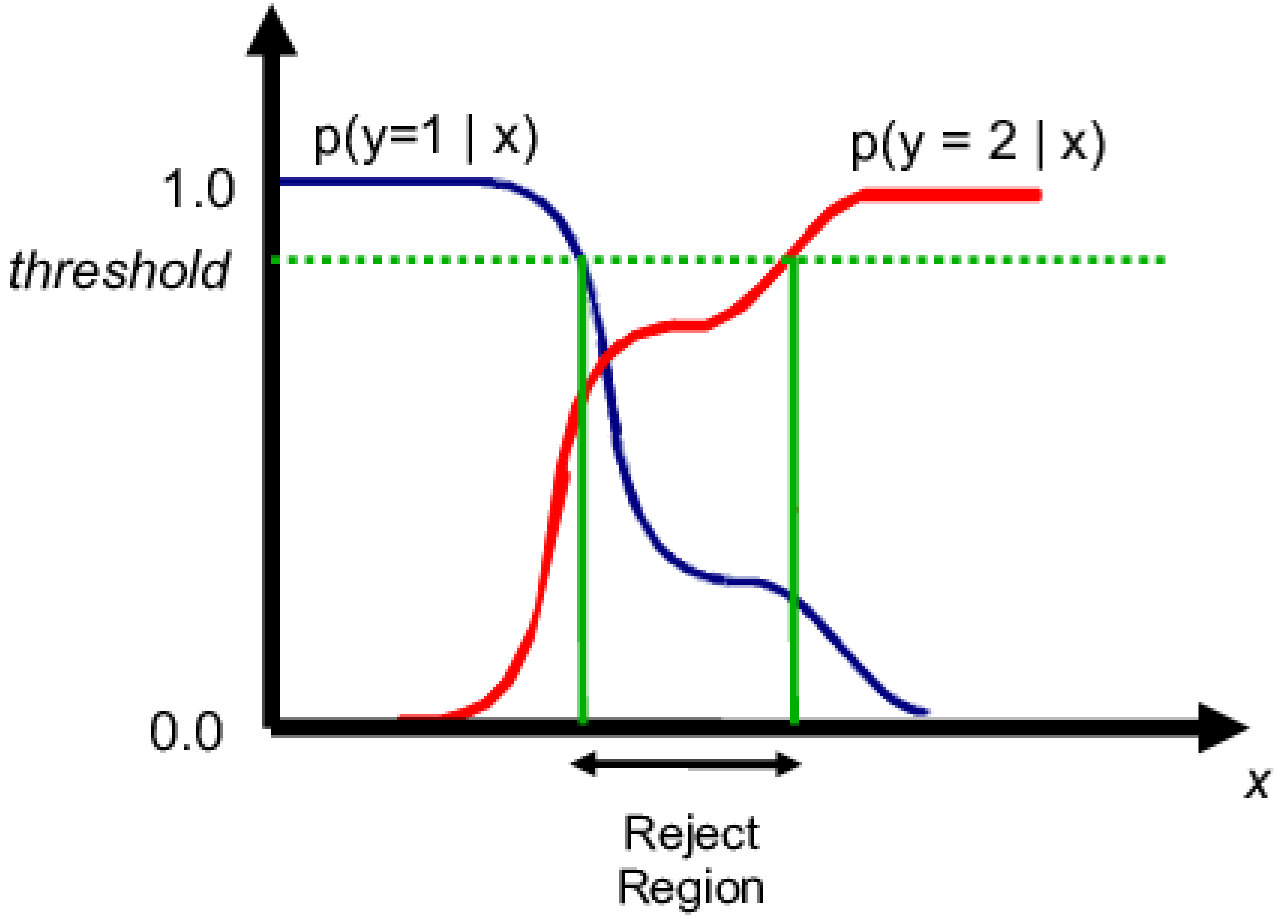
0-1 Loss

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	0	1
$y = 0$	1	0

Weighted Loss

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	0	L_{FN}
$y = 0$	L_{FP}	0

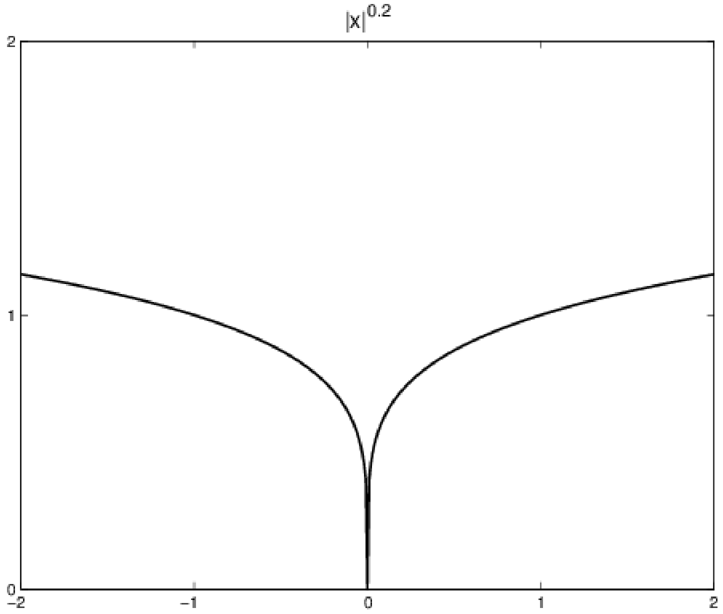
Example of Reject Region



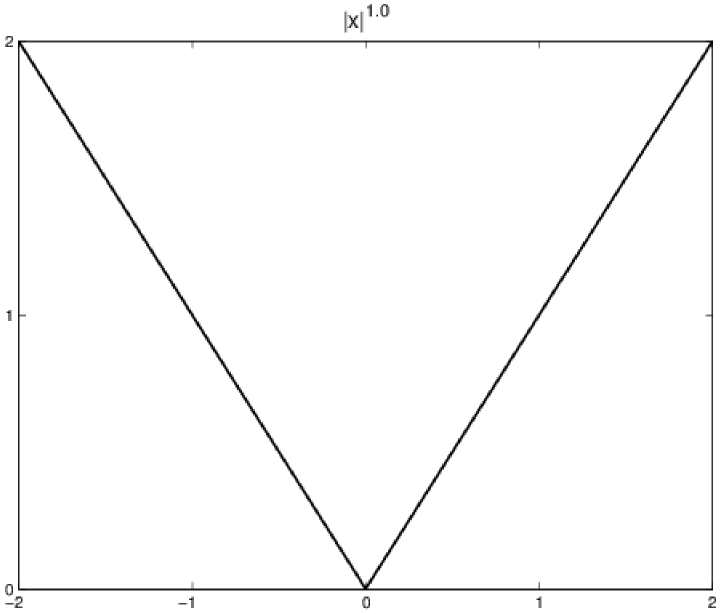
$$L(y=j, a=i) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, \ldots, C\} \\ \lambda_r & \text{if } i = C+1 \\ \lambda_s & \text{otherwise} \end{cases}$$

Absolute Loss

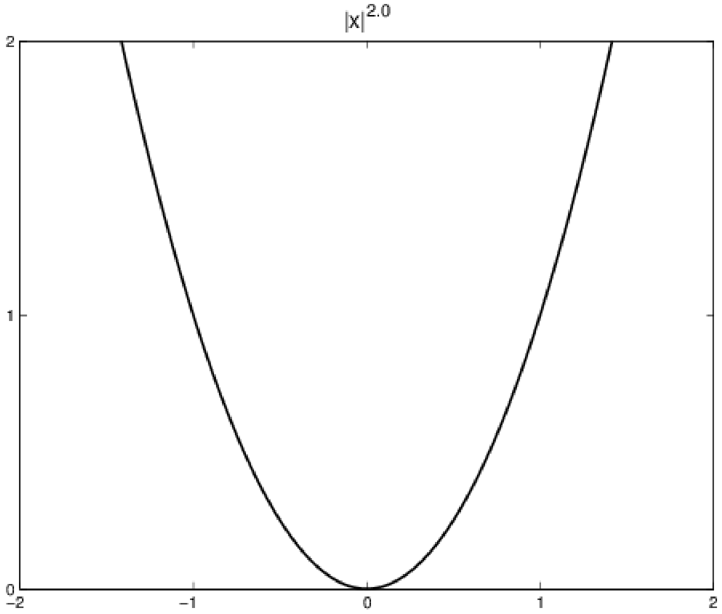
q=.2:
heavier penalty for
small deviations



q=1:
linear penalty



q=2:
heavier penalty for
larger deviations



$$L(y,a) = |y-a|^q$$

Supervised Learning

$$L(\boldsymbol{\theta}, \delta) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|\boldsymbol{\theta})}\left[\ell(y, \delta(\mathbf{x})\right] = \sum_{\mathbf{x}} \sum_{y} L(y, \delta(\mathbf{x}))p(\mathbf{x}, y|\boldsymbol{\theta})$$

Loss matrix with unequal weights …

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	0	L_{FN}
$y = 0$	L_{FP}	0

Confusion Matrix

		Truth		
		1	0	Σ
Estimate	1	TP	FP	$\hat{N}_+ = TP + FP$
	0	FN	TN	$\hat{N}_- = FN + TN$
	Σ	$N_+ = TP + FN$	$N_- = FP + TN$	$N = TP + FP + FN + TN$

Rates Derived from Confusion Matrix

- Rates normalized by Actual counts

	$y=1$	$y=0$
$\hat{y}=1$	TP/N_+=TPR=sensitivity=recall	FP/N_-=FPR=type I
$\hat{y}=0$	FN/N_+=FNR=miss rate=type II	TN/N_-=TNR=specifity

- Rates normalized by Predicted counts

	$y=1$	$y=0$
$\hat{y}=1$	$TP/\hat{N}_+$=precision=PPV	$FP/\hat{N}_+$=FDP
$\hat{y}=0$	$FN/\hat{N}_-$	$TN/\hat{N}_-$=NPV

F_1 Score

- Harmonic mean of Precision (P) and Recall (R)

$$F_1 \triangleq \frac{2}{1/P + 1/R} = \frac{2PR}{R+P}$$

$$F_1 = \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i}$$

Micro versus Macro Averaging

- Macro-average: unweighted
- Micro-average: weighted

Class 1

	$y=1$	$y=0$
$\hat{y}=1$	10	10
$\hat{y}=0$	10	970

Class 2

	$y=1$	$y=0$
$\hat{y}=1$	90	10
$\hat{y}=0$	10	890

Pooled

	$y=1$	$y=0$
$\hat{y}=1$	100	20
$\hat{y}=0$	20	1860

Receiver Operating Characteristic (ROC) Curve versus Precision Recall (PR) Curve