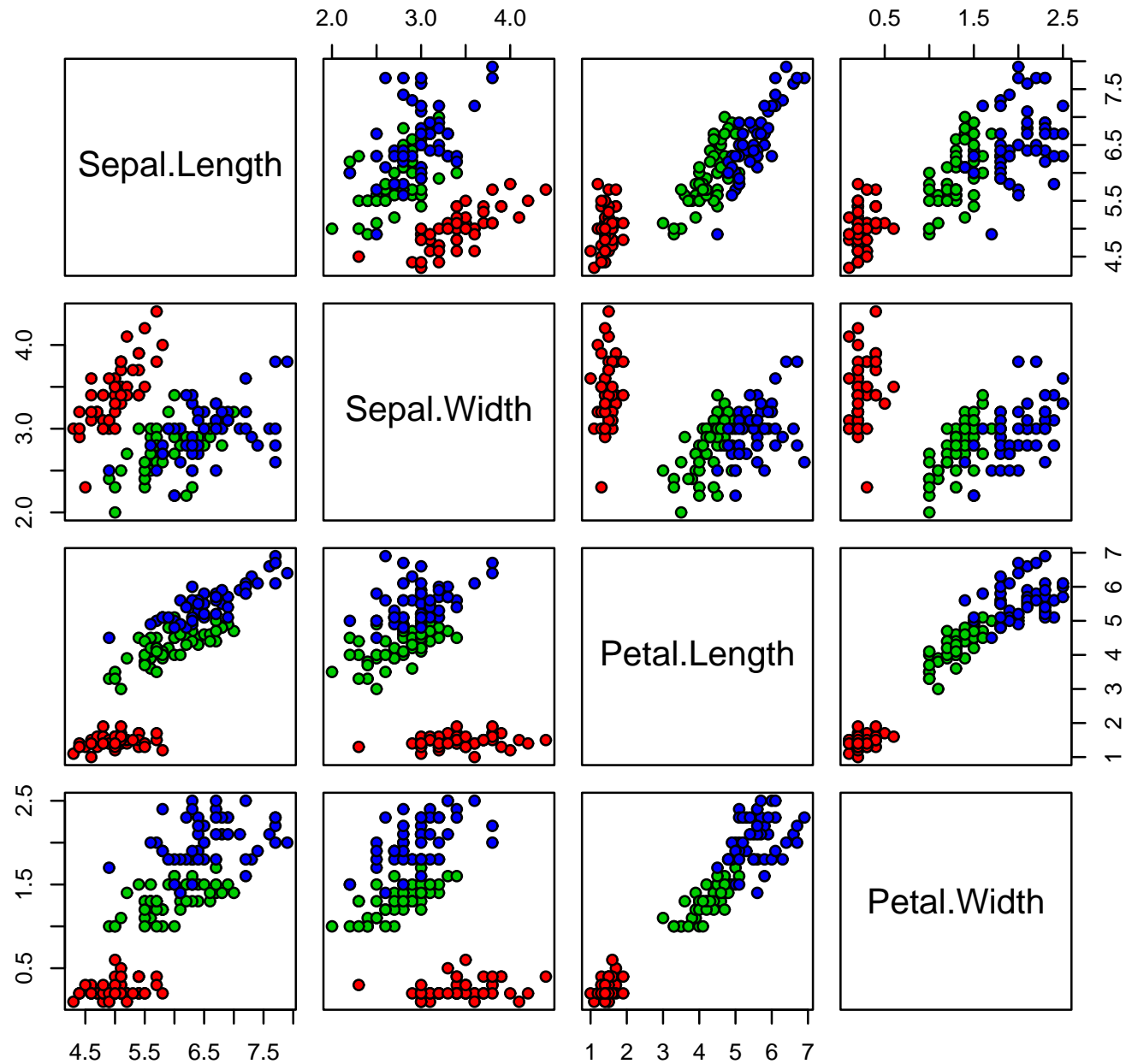


# Does Modeling Lead to More Accurate Classification?

Yoonkyung Lee\*  
Department of Statistics  
The Ohio State University  
\*joint work with Rui Wang

June 21, 2010  
ICSA Applied Statistics Symposium, Indianapolis

## Iris Data (red=setosa,green=versicolor,blue=virginica)



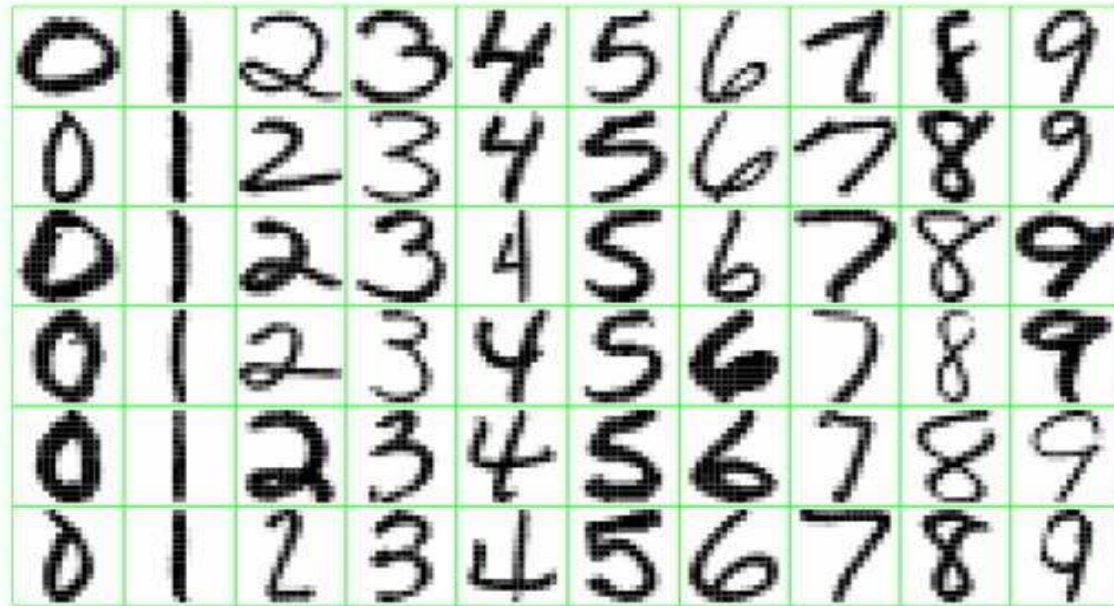


Figure: courtesy of Hastie, Tibshirani, & Friedman (2001)

- ▶ Handwritten digit recognition
- ▶ Cancer diagnosis with gene expression profiles
- ▶ Text categorization

# Classification

- ▶  $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$
- ▶  $y \in \mathcal{Y} = \{1, \dots, k\}$
- ▶ Learn a rule  $\phi : \mathbb{R}^d \rightarrow \mathcal{Y}$  from the training data  $\{(x_i, y_i), i = 1, \dots, n\}$ , where  $(x_i, y_i)$  are i.i.d. with  $P(X, Y)$ .
- ▶ The 0-1 loss function:

$$\rho(y, \phi(\mathbf{x})) = I(y \neq \phi(\mathbf{x}))$$

- ▶ The Bayes decision rule  $\phi_B$  minimizing the error rate  $R(\phi) := P(Y \neq \phi(X))$  is

$$\phi_B(\mathbf{x}) = \arg \max_k P(Y = k \mid X = \mathbf{x}).$$

# Statistical Modeling: The Two Cultures

“One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.” – Breiman (2001)

- ▶ **Model-based methods** in statistics:  
LDA, QDA, logistic regression, kernel density classification
- ▶ **Algorithmic methods** in machine learning:  
Support vector machine (SVM), boosting, decision trees, neural network
- ▶ *Less is required in pattern recognition.*  
– Devroye, Györfi and Lugosi (1996)
- ▶ *If you possess a restricted information for solving some problem, try to solve the problem directly and never solve a general problem as an intermediate step.*  
– Vapnik (1998)

# Questions

- ▶ Is modeling necessary for classification?
- ▶ Does modeling lead to more accurate classification?
- ▶ How to quantify the relative efficiency?
- ▶ How do the two approaches compare?

# Classification Consistency

- ▶ In the binary case ( $k = 2$ ), suppose that  $y = 1$  or  $-1$ .
- ▶ Typically obtain a discriminant function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , which induces a classifier  $\phi(x) = \text{sign}(f(x))$ , by minimizing the risk under convex surrogate loss
  - ▶ Logistic regression: deviance ( $-\log$  likelihood)
  - ▶ Support vector machine: hinge loss
  - ▶ Boosting: exponential loss
- ▶ Lin (2000), Zhang (AOS 2004), Bartlett, Jordan, and McAuliffe (JASA 2006):  
Both approaches are consistent in classification.

# Loss Functions

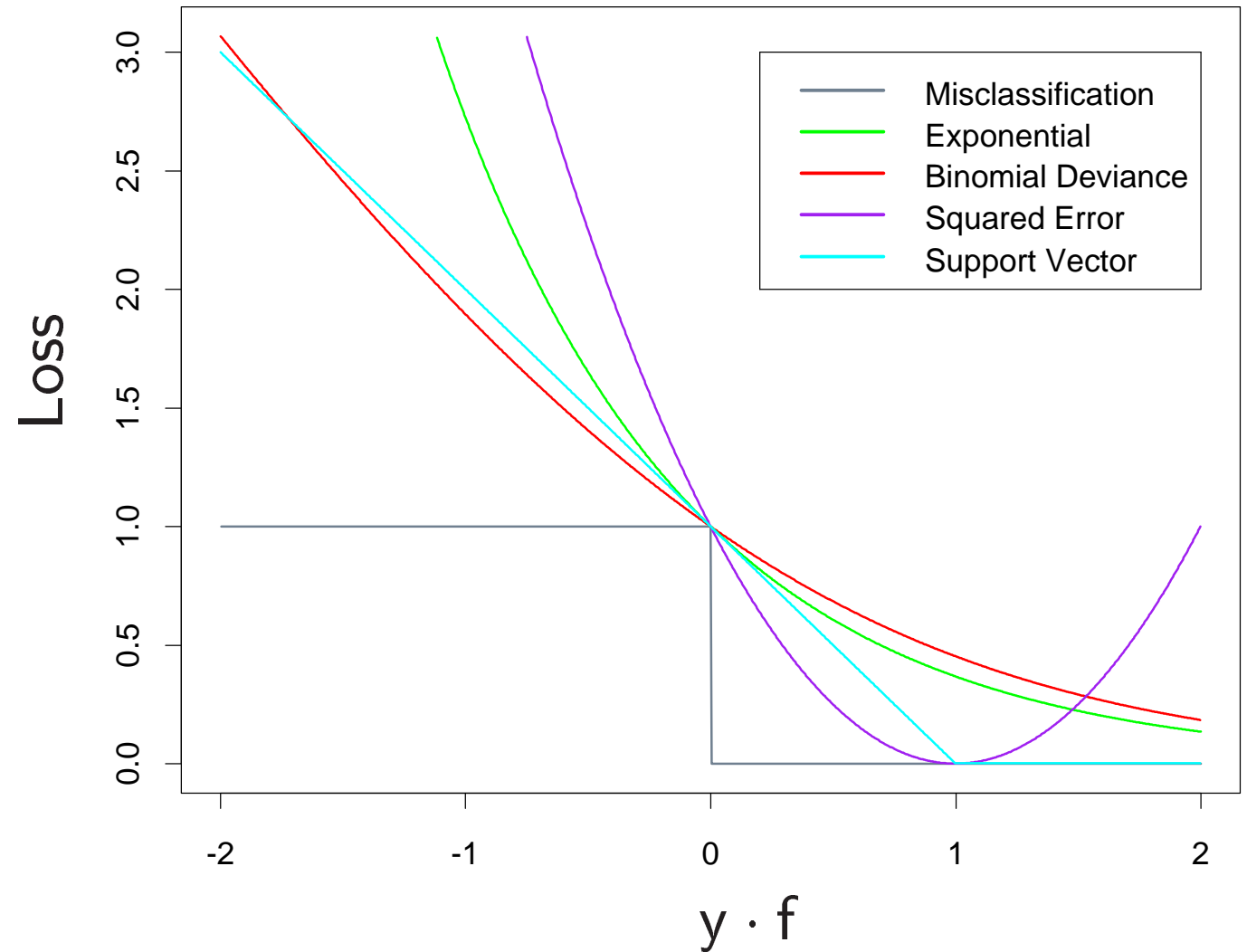


Figure: courtesy of HTF (2001)



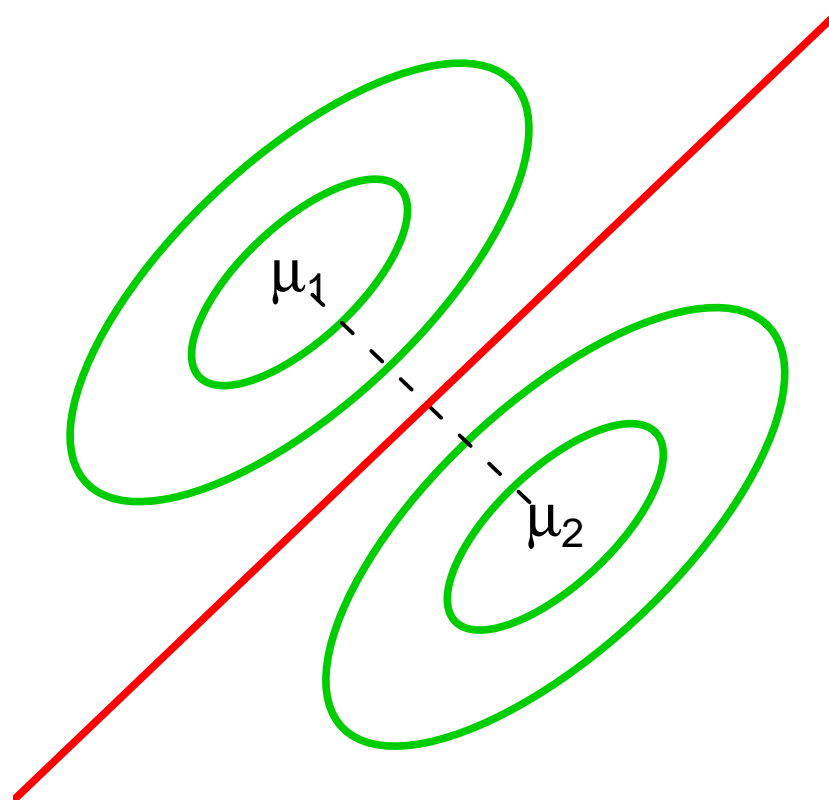
# Outline

- ▶ Efron's comparison of LDA with logistic regression
- ▶ Efficiency of algorithmic approach  
(support vector machine and boosting)
- ▶ Simulation study
- ▶ Discussion

# Normal Distribution Setting

- ▶ Two multivariate normal distributions in  $\mathbb{R}^d$  with mean vectors  $\mu_1$  and  $\mu_2$  and a common covariance matrix  $\Sigma$
- ▶  $\pi_+ = P(Y = 1)$  and  $\pi_- = P(Y = -1)$ .
- ▶ For example, when  $\pi_+ = \pi_-$ , Fisher's LDA boundary is

$$\left\{ \Sigma^{-1}(\mu_1 - \mu_2) \right\}' \left\{ \mathbf{x} - \frac{1}{2}(\mu_1 + \mu_2) \right\} = 0.$$



# Canonical LDA setting

Efron (JASA 1975), *The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis*

- ▶  $X \sim N((\Delta/2)\mathbf{e}_1, \mathbf{I})$  for  $Y = 1$  with probability  $\pi_+$   
 $X \sim N(-(\Delta/2)\mathbf{e}_1, \mathbf{I})$  for  $Y = -1$  with probability  $\pi_-$   
where  $\Delta := \{(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)\}^{1/2}$
- ▶ Fisher's linear discriminant function is

$$\ell(\mathbf{x}) = \log(\pi_+/\pi_-) + \Delta \mathbf{x}_1.$$

- ▶ Let  $\beta_0^* = \log(\pi_+/\pi_-)$ ,  $(\beta_1^*, \dots, \beta_d^*)' = \Delta \mathbf{e}_1$ , and  $\beta^* = (\beta_0^*, \dots, \beta_d^*)'$ .

# Excess Error

- ▶ For a linear discriminant method  $\hat{\ell}$  with coefficient vector  $\hat{\beta}_n$ , if  $\sqrt{n}(\hat{\beta}_n - \beta^*) \rightarrow N(0, \Sigma_\beta)$ , the expected increased error rate of  $\hat{\ell}$ ,  $E(R(\hat{\ell}) - R(\phi_B))$

$$= \frac{\pi_+ \phi(D_1)}{2\Delta n} \left[ \sigma_{00} - \frac{2\beta_0^*}{\Delta} \sigma_{01} + \frac{\beta_0^{*2}}{\Delta^2} \sigma_{11} + \sigma_{22} + \cdots + \sigma_{dd} \right] + o\left(\frac{1}{n}\right),$$

where  $D_1 = \Delta/2 + (1/\Delta) \log(\pi_+/\pi_-)$ .

- ▶ In particular, when  $\pi_+ = \pi_-$ ,

$$E(R(\hat{\ell}) - R(\phi_B)) = \frac{\phi(\Delta/2)}{4\Delta n} \left[ \sigma_{00} + \sigma_{22} + \cdots + \sigma_{dd} \right] + o\left(\frac{1}{n}\right).$$

# Relative Efficiency

- ▶ Efron (1975) studied the **Asymptotic Relative Efficiency** (ARE) of logistic regression (LR) to normal discrimination (LDA) defined as

$$\lim_{n \rightarrow \infty} \frac{E(R(\hat{\ell}_{LDA}) - R(\phi_B))}{E(R(\hat{\ell}_{LR}) - R(\phi_B))}.$$

- ▶ Logistic regression is shown to be between one half and two thirds as effective as normal discrimination typically.

# General Framework for Comparison

- ▶ Identify the limiting distribution of  $\hat{\beta}_n$  for other classification procedures (SVM, boosting, etc.) under the canonical LDA setting:

$$\hat{\beta}_n = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n \rho(y_i, \mathbf{x}_i; \beta)$$

- ▶ Need large sample theory for M-estimators.
- ▶ Find the excess error of each method and compute the efficiency relative to LDA.

# M-estimator Asymptotics

- ▶ Pollard (ET 1991), Hjort and Pollard (1993), Geyer (AOS 1994), Knight and Fu (AOS 2000), Rocha, Wang and Yu (2009)
- ▶ Convexity of the loss  $\rho$  is the key.
- ▶ Let  $L(\beta) := E\rho(Y, X; \beta)$ ,  $\beta^* := \arg \min L(\beta)$ ,

$$H(\beta) := \frac{L(\beta)}{\partial\beta\partial\beta'}, \text{ and } G(\beta) := E \left( \frac{\partial\rho(Y, X; \beta)}{\partial\beta} \right) \left( \frac{\partial\rho(Y, X; \beta)}{\partial\beta} \right)'$$

- ▶ Under some regularity conditions,

$$\sqrt{n}(\hat{\beta}_n - \beta^*) \rightarrow N(0, H(\beta^*)^{-1} G(\beta^*) H(\beta^*)^{-1})$$

in distribution.

# Linear SVM

Koo, Lee, Kim, and Park (JMLR 2008), *A Bahadur Representation of the Linear Support Vector Machine*

- ▶ With  $\beta := (\beta_0, w')'$ ,  $\ell(\mathbf{x}; \beta) = w' \mathbf{x} + \beta_0$
- ▶  $\hat{\beta}_{\lambda, n} = \arg \min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^n (1 - y_i \ell(\mathbf{x}_i; \beta))_+ + \lambda \|w\|^2 \right\}$
- ▶ Under the canonical LDA setting with  $\pi_+ = \pi_-$ , for  $\lambda = o(n^{-1/2})$ ,

$$\sqrt{n} (\hat{\beta}_{\lambda, n} - \beta_{SVM}^*) \rightarrow N(0, \Sigma_{\beta_{SVM}^*}),$$

where  $\beta_{SVM}^* = \frac{2}{\Delta(2a^* + \Delta)} \beta_{LDA}^*$  and

$a^*$  is a constant such that  $\phi(a^*)/\Phi(a^*) = \Delta/2$ .

- ▶ If  $\pi_+ \neq \pi_-$ ,  $\hat{w}_n \propto w_{LDA}^*$  but  $\hat{\beta}_0$  is inconsistent.



# Relative Efficiency of SVM to LDA

Under the canonical LDA setting with  $\pi_+ = \pi_- = 0.5$ , the ARE of the linear SVM to LDA is

$$Eff = \frac{2}{\Delta} \left(1 + \frac{\Delta^2}{4}\right) \phi(a^*).$$

$\Delta$	$R(\phi_B)$	$a^*$	<b>SVM</b>	LR
2.0	0.1587	-0.3026	0.7622	0.899
2.5	0.1056	-0.6466	0.6636	0.786
3.0	0.0668	-0.9685	0.5408	0.641
3.5	0.0401	-1.2756	0.4105	0.486
4.0	0.0228	-1.5718	0.2899	0.343

# Boosting

- ▶  $\hat{\beta}_n = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n \exp(-y_i \ell(\mathbf{x}_i; \beta))$

- ▶ Under the canonical LDA setting with  $\pi_+ = \pi_-$ ,

$$\sqrt{n} (\hat{\beta}_n - \beta_{boost}^*) \rightarrow N(0, \Sigma_{\beta_{boost}^*}),$$

where  $\beta_{boost}^* = \frac{1}{2} \beta_{LDA}^*$ .

- ▶ In general,  $\hat{\beta}_n$  is a consistent estimator of  $(1/2) \beta_{LDA}^*$ .

# Relative Efficiency of Boosting to LDA

Under the canonical LDA setting with  $\pi_+ = \pi_- = 0.5$ , the ARE of Boosting to LDA is

$$Eff = \frac{1 + \Delta^2/4}{\exp(\Delta^2/4)}.$$

$\Delta$	$R(\phi_B)$	<b>Boosting</b>	SVM	LR
2.0	0.1587	0.7358	0.7622	0.899
2.5	0.1056	0.5371	0.6636	0.786
3.0	0.0668	0.3425	0.5408	0.641
3.5	0.0401	0.1900	0.4105	0.486
4.0	0.0228	0.0916	0.2899	0.343

# Smooth SVM

Lee and Mangasarian (2001), *SSVM: A Smooth Support Vector Machine*

- ▶  $\hat{\beta}_{\lambda,n} = \arg \min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^n (1 - y_i \ell(\mathbf{x}_i; \beta))_+^2 + \lambda \|\mathbf{w}\|^2 \right\}$
- ▶ Under the canonical LDA setting with  $\pi_+ = \pi_-$ , for  $\lambda = o(n^{-1/2})$ ,

$$\sqrt{n} (\hat{\beta}_{\lambda,n} - \beta_{SSVM}^*) \rightarrow N(0, \Sigma_{\beta_{SSVM}^*}),$$

where  $\beta_{SSVM}^* = \frac{2}{\Delta(2a^* + \Delta)} \beta_{LDA}^*$  and

$a^*$  is a constant such that  $\{a^* \Phi(a^*) + \phi(a^*)\} \Delta = 2\Phi(a^*)$ .

# Relative Efficiency of SSVM to LDA

Under the canonical LDA setting with  $\pi_+ = \pi_- = 0.5$ , the ARE of the Smooth SVM to LDA is

$$Eff = \frac{(4 + \Delta^2)\Phi(a^*)}{\Delta(2a^* + \Delta)}.$$

$\Delta$	$R(\phi_B)$	$a^*$	<b>SSVM</b>	SVM	LR
2.0	0.1587	0.4811	0.9247	0.7622	0.899
2.5	0.1056	0.0058	0.8200	0.6636	0.786
3.0	0.0668	-0.4073	0.6779	0.5408	0.641
3.5	0.0401	-0.7821	0.5206	0.4105	0.486
4.0	0.0228	-1.1312	0.3712	0.2899	0.343

# Possible Explanation for Increased Efficiency

Hastie, Tibshirani, and Friedman (2001),  
*Elements of Statistical Learning*

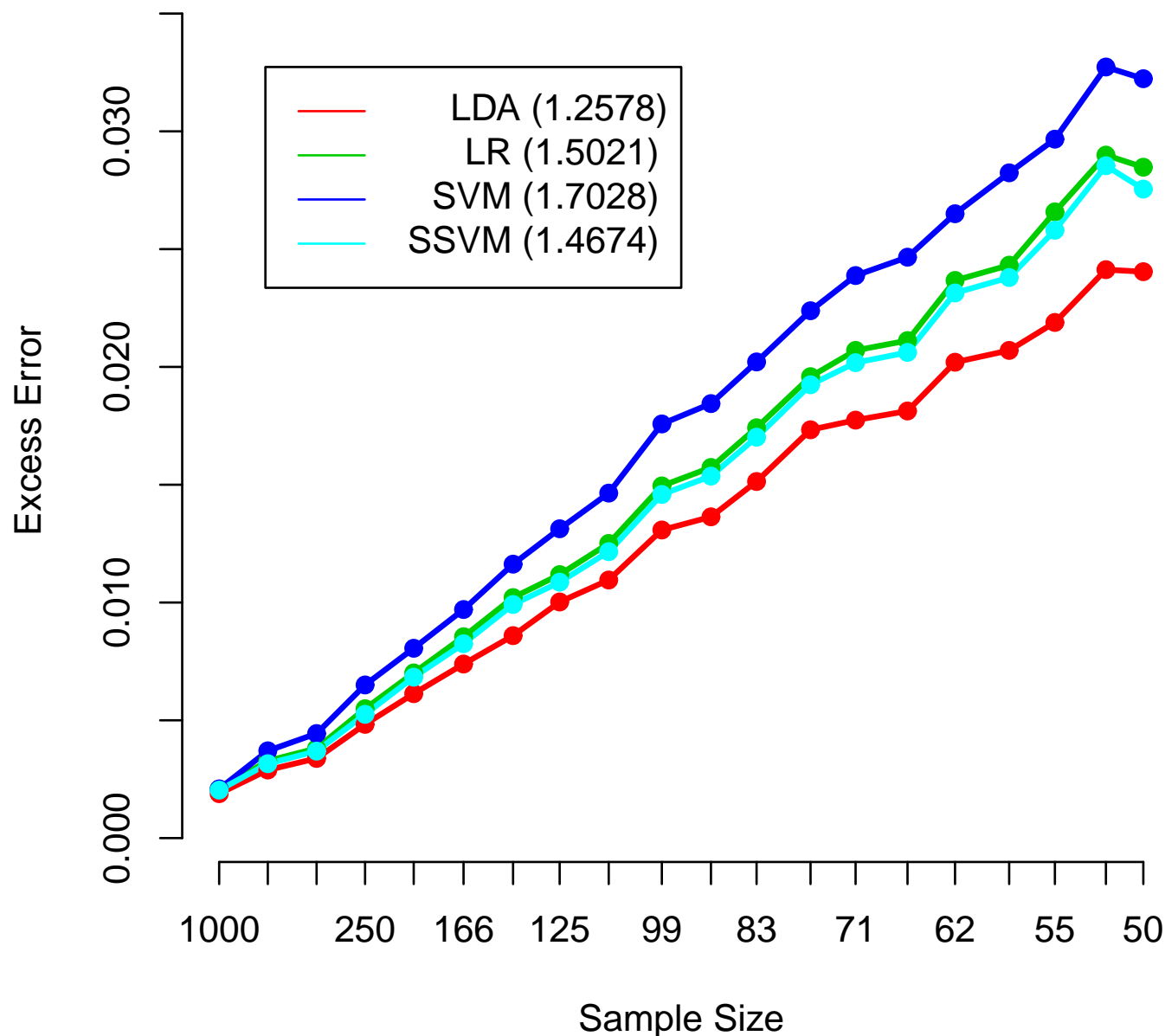
- ▶ There is a close connection between Fisher's LDA and regression approach to classification with class indicators:

$$\min \sum_{i=1}^n (y_i - \beta_0 - w'x_i)^2$$

- ▶ The least squares coefficient is identical up to a scalar multiple to the LDA coefficient:

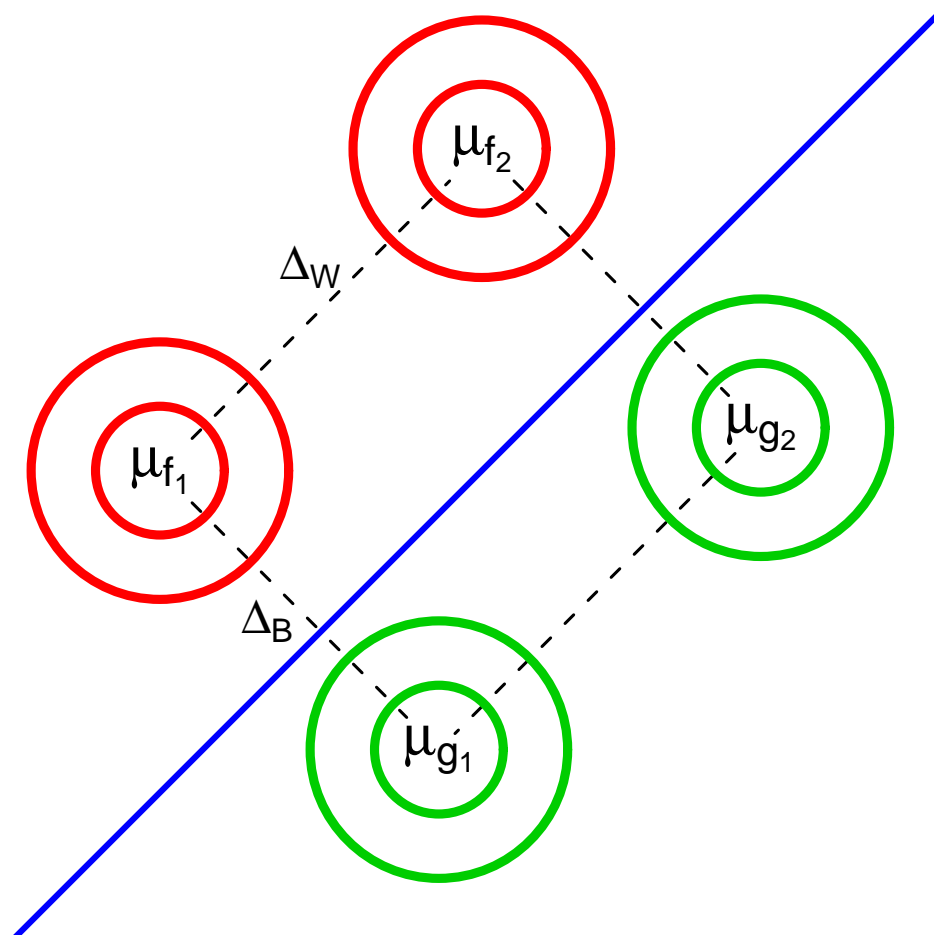
$$\hat{w} \propto \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$$

# Finite-Sample Excess Error



**Figure:**  $\Delta = 2$ ,  $d = 5$ ,  $R(\phi_B) = 0.1587$ , and  $\pi_+ = \pi_-$ . The results are based on 1000 replicates.

# A Mixture of Two Gaussian Distributions



**Figure:**  $\Delta_W$  and  $\Delta_B$  indicate the mean difference between two Gaussian components within each class and the mean difference between two classes.



# As $\Delta_W$ Varies

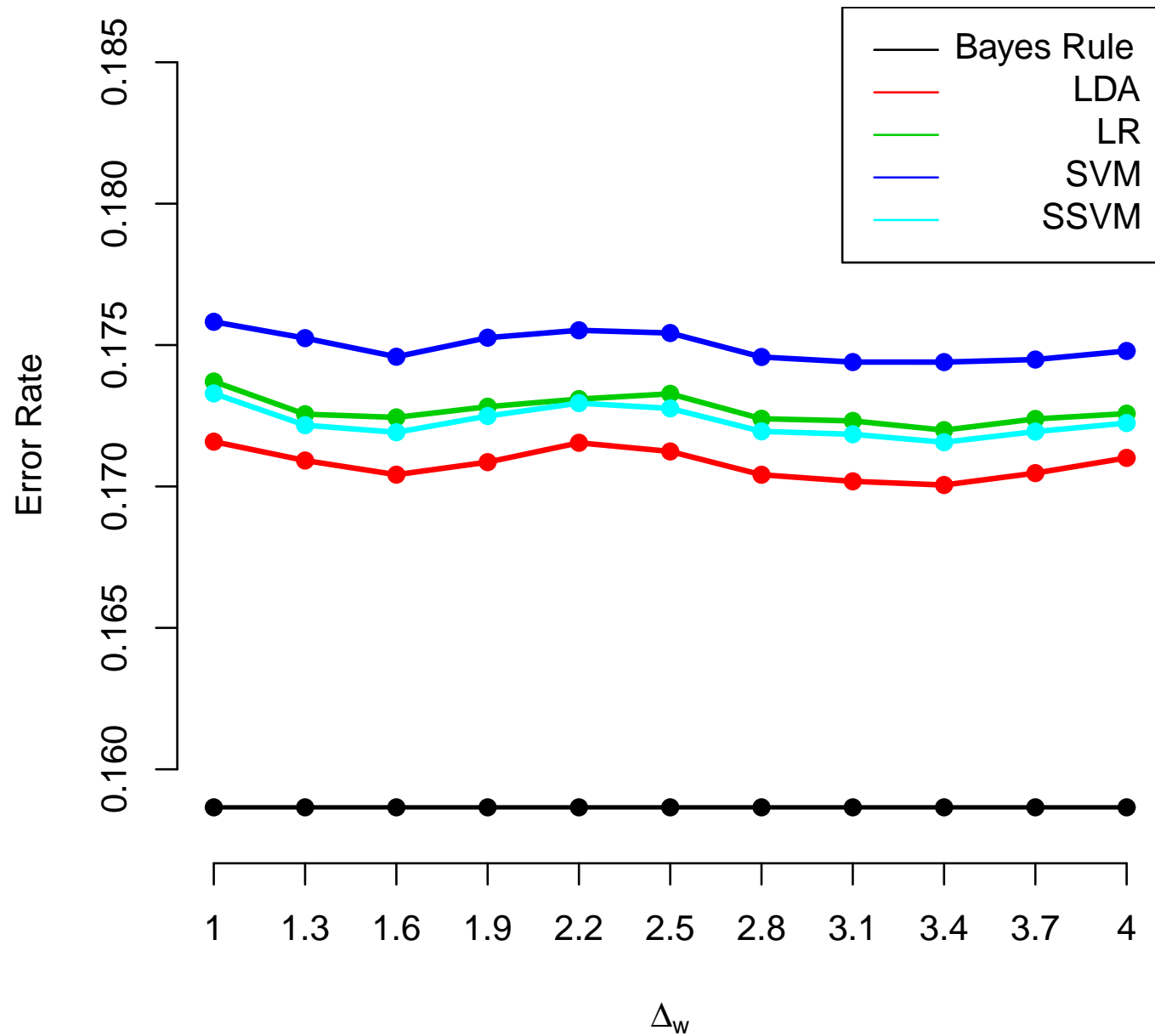


Figure:  $\Delta_B = 2$ ,  $d = 5$ ,  $\pi_+ = \pi_-$ ,  $\pi_1 = \pi_2$ , and  $n = 100$

As  $\Delta_B$  Varies

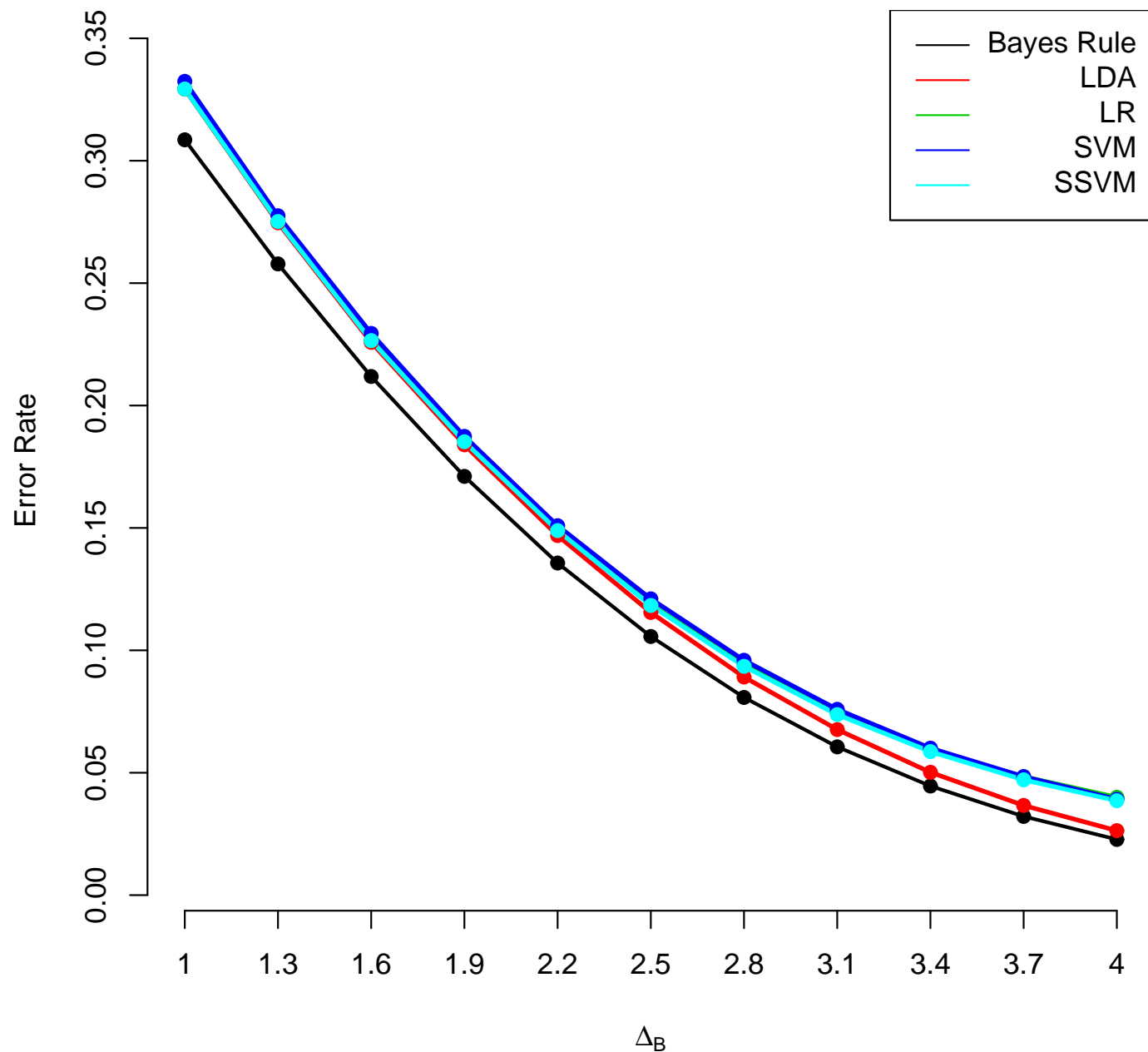


Figure:  $\Delta_W = 1$ ,  $d = 5$ ,  $\pi_+ = \pi_-$ ,  $\pi_1 = \pi_2$ , and  $n = 100$

As  $\Delta_B$  Varies

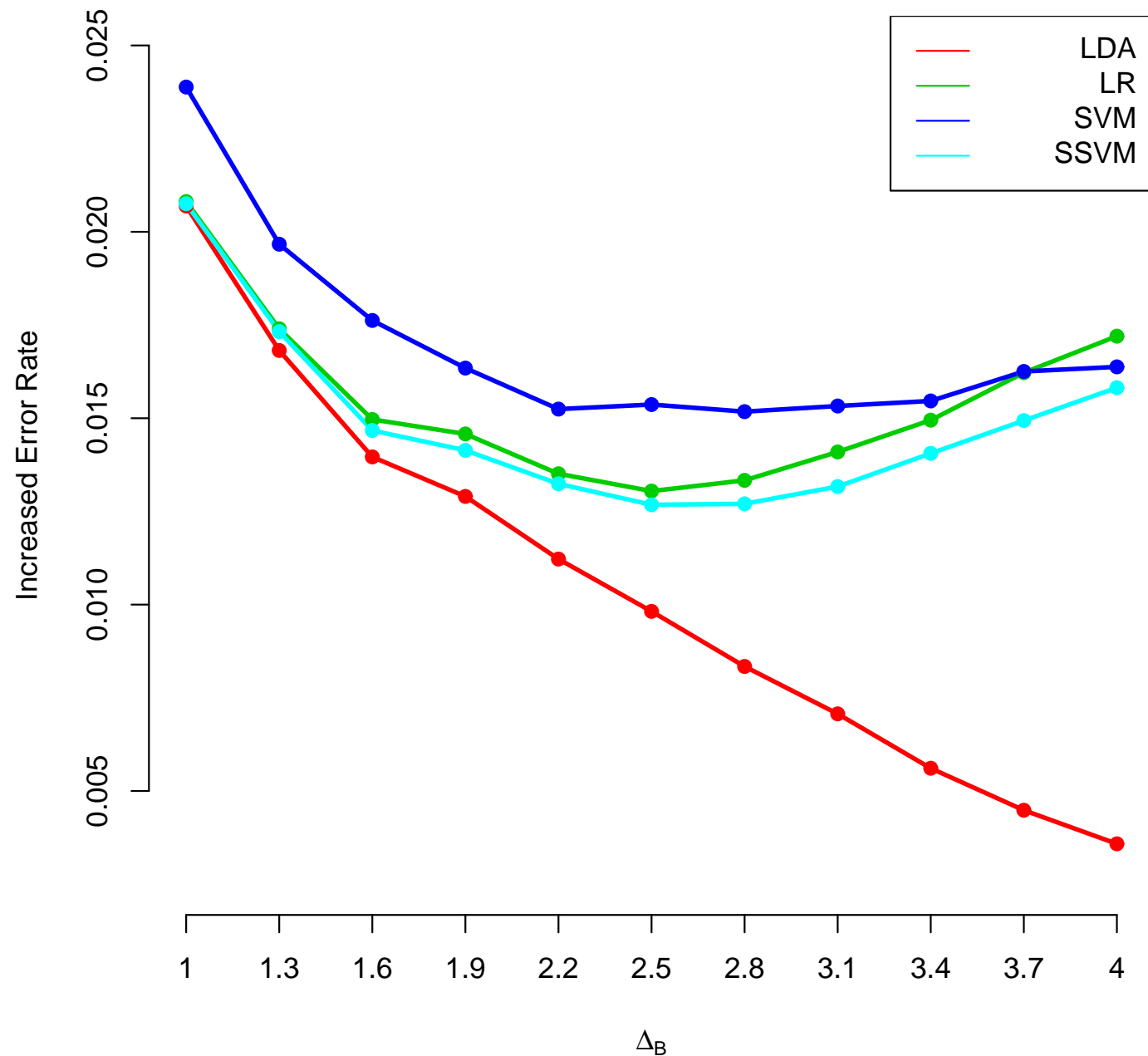


Figure:  $\Delta_W = 1$ ,  $d = 5$ ,  $\pi_+ = \pi_-$ ,  $\pi_1 = \pi_2$ , and  $n = 100$

# As Dimension $d$ Varies

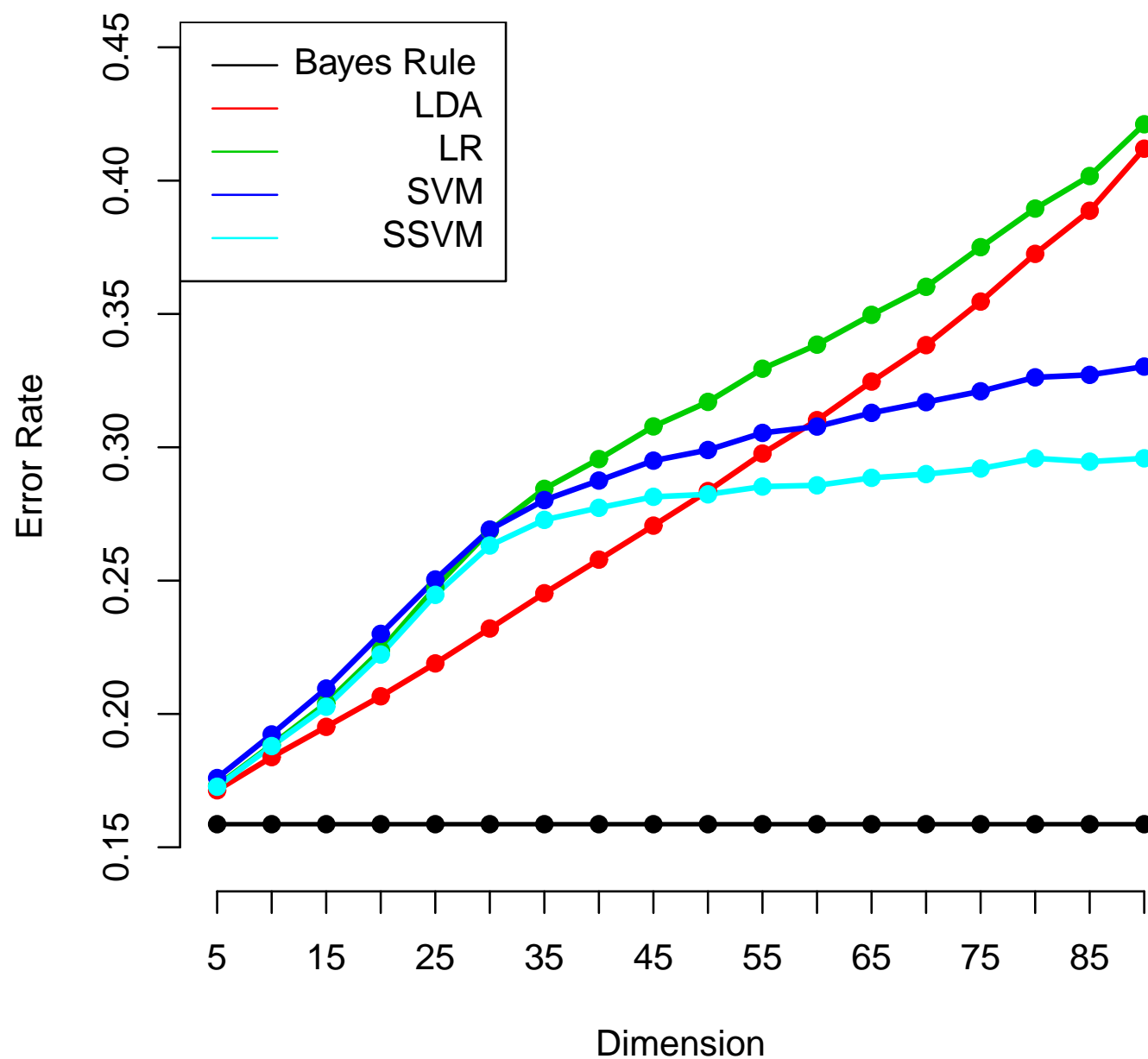


Figure:  $\Delta_W = 1$ ,  $\Delta_B = 2$ ,  $\pi_+ = \pi_-$ ,  $\pi_1 = \pi_2$ , and  $n = 100$

# QDA Setting

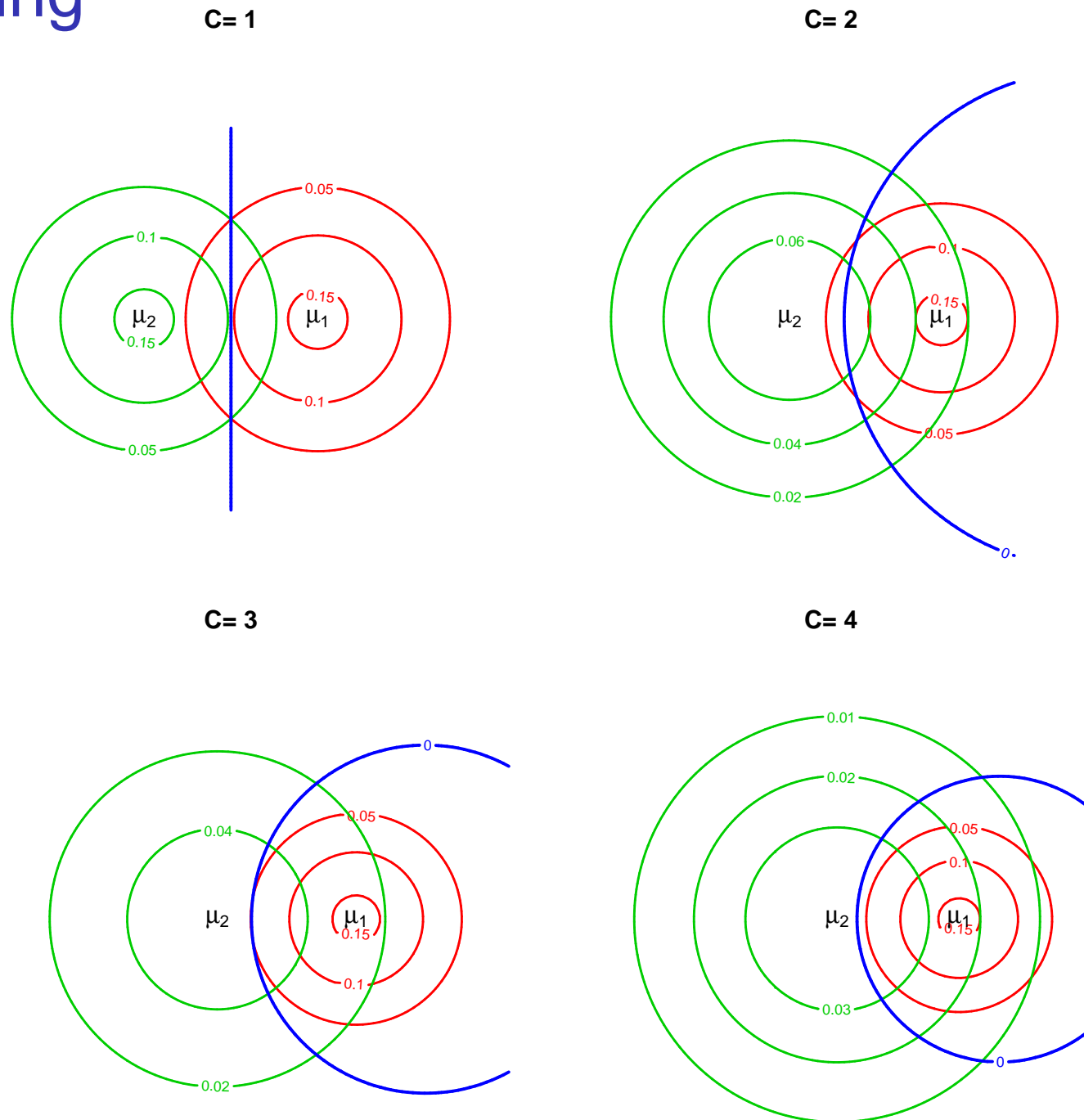


Figure:  $X|Y = 1 \sim N(\mu_1, \Sigma)$  and  $X|Y = -1 \sim N(\mu_2, C\Sigma)$

# Decomposition of Error

- For a rule  $\phi \in \mathcal{F}$ ,

$$R(\phi) - R(\phi_B) = \underbrace{R(\phi) - R(\phi_{\mathcal{F}})}_{\text{'estimation' error}} + \underbrace{R(\phi_{\mathcal{F}}) - R(\phi_B)}_{\text{approximation error}}$$

where  $\phi_{\mathcal{F}} = \arg \min_{\phi \in \mathcal{F}} R(\phi)$ .

- When a method  $M$  is used to choose  $\phi$  from  $\mathcal{F}$ ,

$$R(\phi) - R(\phi_{\mathcal{F}}) = \underbrace{R(\phi) - R(\phi_M)}_{\text{M-specific est.error}} + \underbrace{R(\phi_M) - R(\phi_{\mathcal{F}})}_{\text{M-specific approx.error}}$$

where  $\phi_M$  is the method-specific limiting rule within  $\mathcal{F}$ .

# Approximation Error

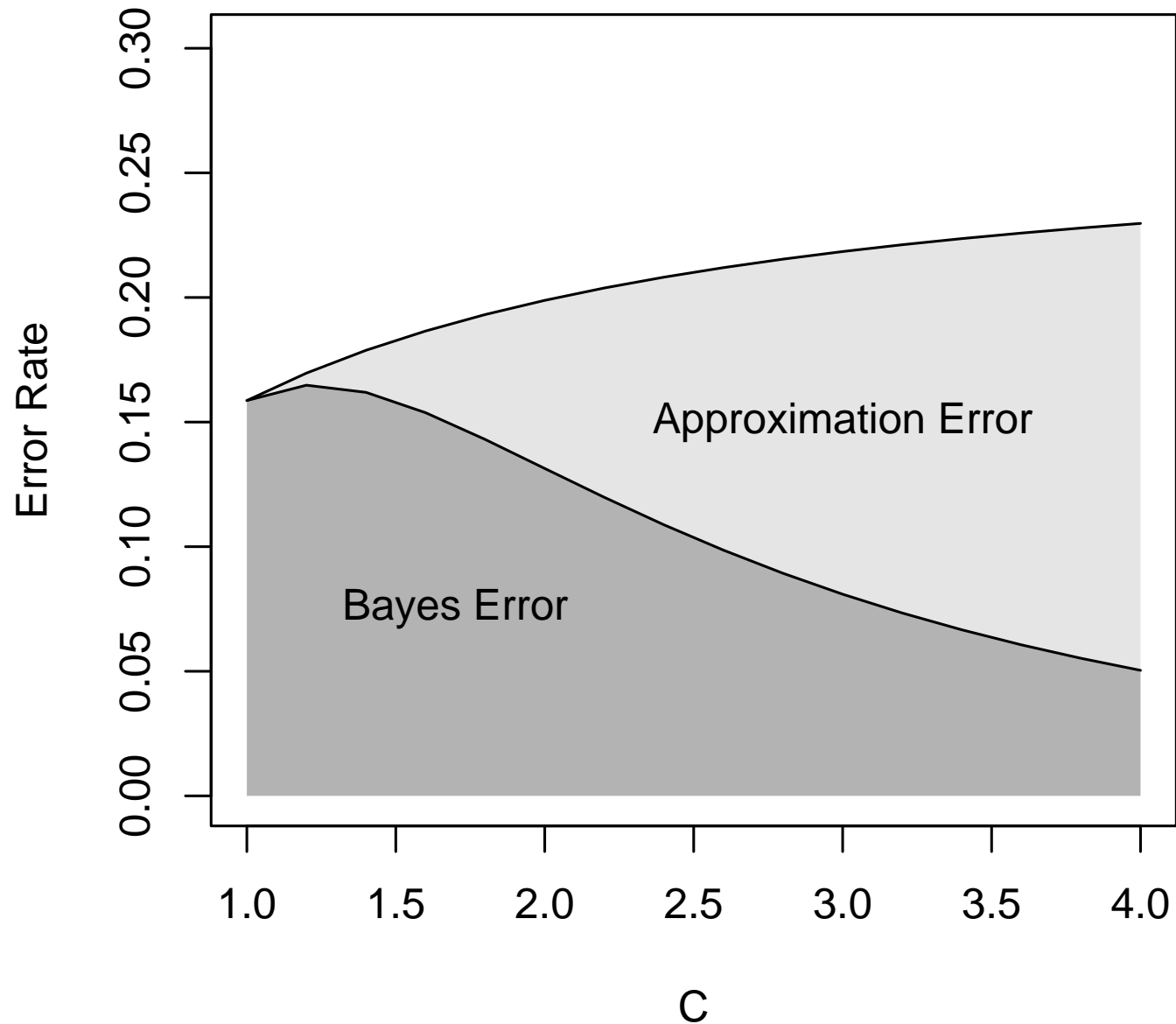
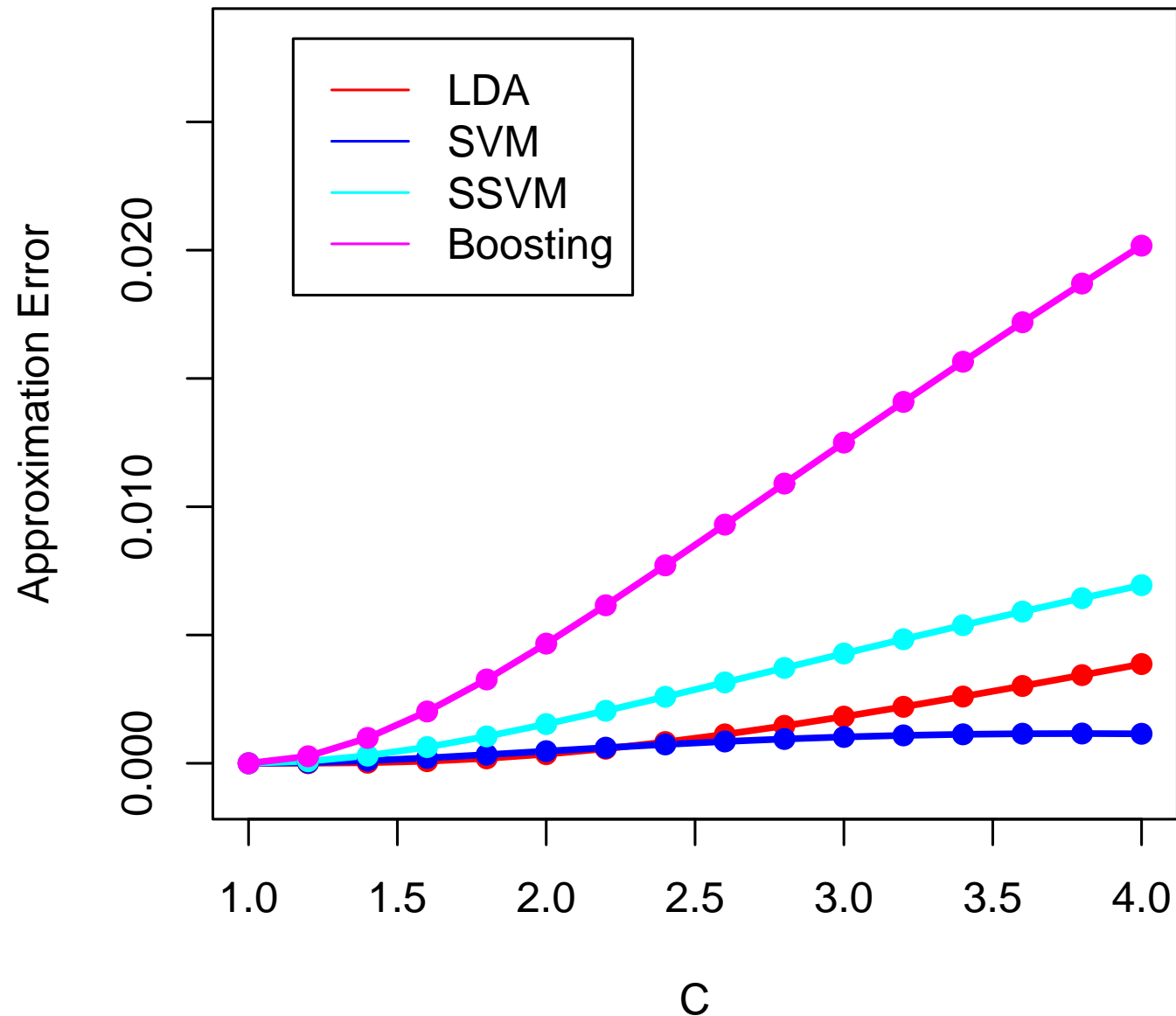


Figure: QDA setting with  $\Delta = 2$ ,  $\Sigma = I$ ,  $d = 10$ , and  $\pi_+ = \pi_-$

# Method-Specific Approximation Error



**Figure:** Method-specific approximation error of linear classifiers in the QDA setting



# Extensions

- ▶ For high dimensional data, study double asymptotics where  $d$  also grows with  $n$ .
- ▶ Compare methods in a regularization framework.
- ▶ Investigate consistency and relative efficiency under other models.
- ▶ Consider potential mis-specification of a model.  
*All models are wrong, but some are useful* – Box
- ▶ Compare methods in terms of robustness.

# Concluding Remarks

- ▶ Compared modeling approach with algorithmic approach in the efficiency of reducing error rates.
- ▶ Under the normal setting, modeling leads to more efficient use of data.
  - Linear SVM is shown to be between 40% and 67% as effective as LDA when the Bayes error rate is between 4% and 10%.
- ▶ A loss function plays an important role in determining the efficiency of the corresponding procedure.
  - Squared hinge loss could yield more effective procedure than logistic regression.
- ▶ The theoretical comparisons can be extended in many directions.