A Bahadur Representation of the Linear Support Vector Machine

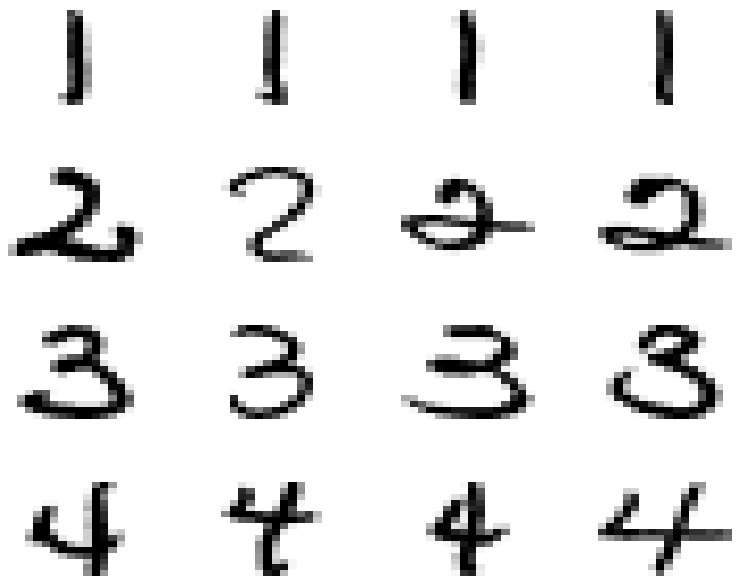
Yoonkyung Lee
Department of Statistics
The Ohio State University

October 7, 2008
Data Mining and Statistical Learning
Study Group

Outline

- Support Vector Machines
- Statistical Properties of SVM
- Main Results
  (asymptotic analysis of the linear SVM)
- An Illustrative Example
- Discussion

Applications



- Handwritten digit recognition
- Cancer diagnosis with microarray data
- Text categorization

Classification

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

$$\mathcal{L}(y, \phi(\boldsymbol{x})) = I(y \neq \phi(\boldsymbol{x}))$$

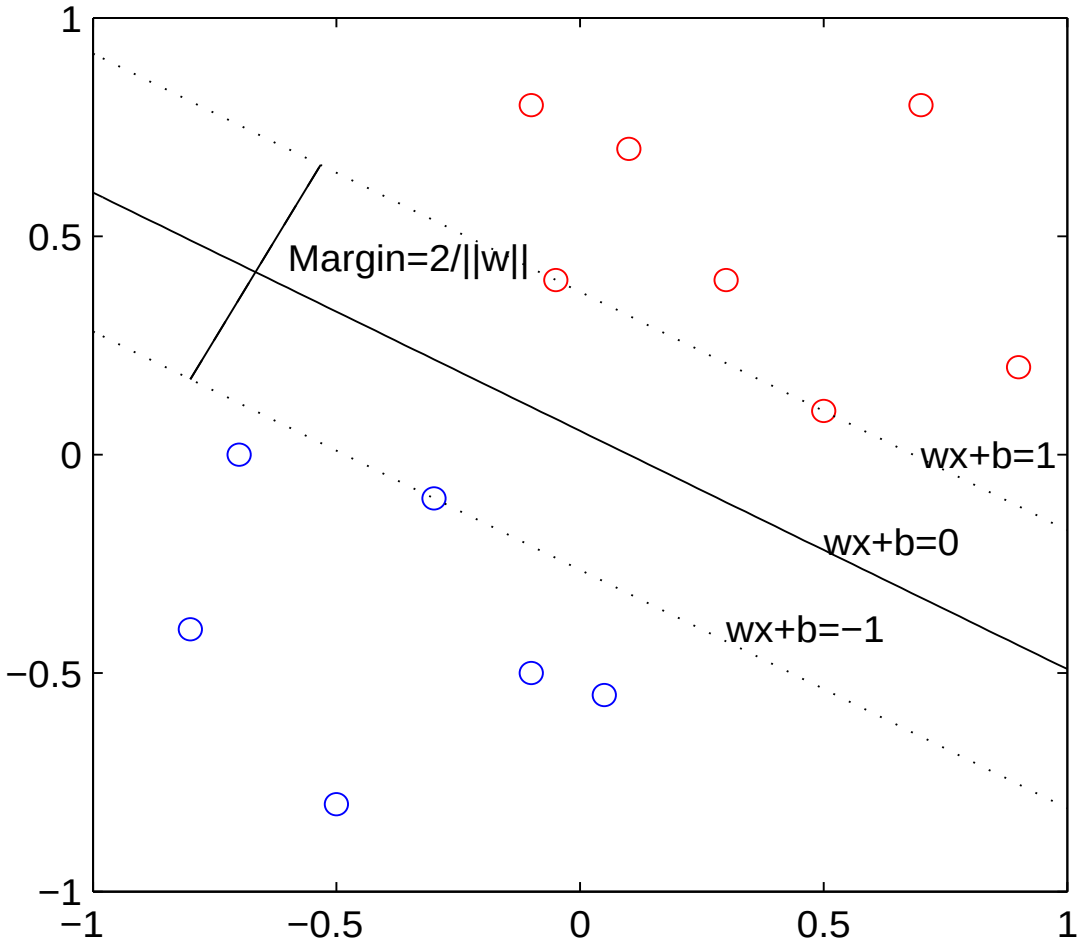
Methods of Regularization (Penalization)

Find $f(\boldsymbol{x}) \in \mathcal{F}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda J(f).$$

- Empirical risk + penalty
- $\mathcal{F}$: a class of candidate functions
- $J(f)$: complexity of the model f
- $\lambda > 0$: a regularization parameter
- Without the penalty $J(f)$, ill-posed problem

Maximum Margin Hyperplane



Support Vector Machines

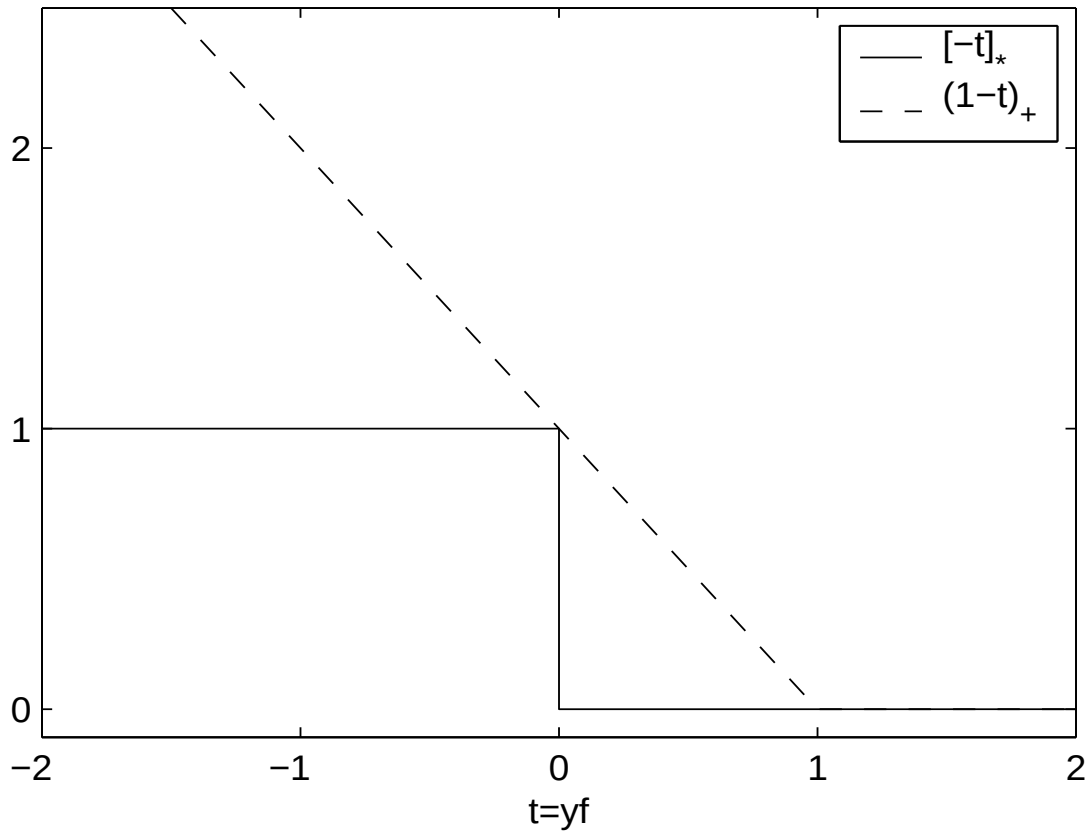
Boser, Guyon, & Vapnik (1992)
Vapnik (1995), *The Nature of Statistical Learning Theory*
A discussion paper (2006) in *Statistical Science*

- $y_i \in \{-1, 1\}$, class labels in the binary case
- Find $f \in \mathcal{F} = \{f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b \mid \boldsymbol{w} \in \mathbb{R}^d \text{ and } b \in \mathbb{R}\}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(\boldsymbol{x}_i))_+ + \lambda\|\boldsymbol{w}\|^2,$$

  where λ is a regularization parameter.
- Classification rule : $\phi(\boldsymbol{x}) = sign(f(\boldsymbol{x}))$

Hinge Loss



$(1 - yf(\boldsymbol{x}))_+$ is an upper bound of the misclassification loss function $I(y \neq \phi(\boldsymbol{x})) = [-yf(\boldsymbol{x})]_* \leq (1 - yf(\boldsymbol{x}))_+$ where $[t]_* = I(t \geq 0)$ and $(t)_+ = \max\{t, 0\}$.

SVM in General

Find $f(\boldsymbol{x}) = b + h(\boldsymbol{x})$ with $h \in \mathcal{H}_K$ (RKHS) minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(\boldsymbol{x}_i))_+ + \lambda \|h\|^2_{\mathcal{H}_K}.$$

- Linear SVM:
  $\mathcal{H}_K = \{h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} \mid \boldsymbol{w} \in \mathbb{R}^d\}$ with
  i) $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}'$
  ii) $\|h\|^2_{\mathcal{H}_K} = \|\boldsymbol{w}^\top \boldsymbol{x}\|^2_{\mathcal{H}_K} = \|\boldsymbol{w}\|^2$
- Nonlinear SVM: $K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top \boldsymbol{x}')^d$,
  $\exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\sigma^2)$, ...

Statistical Properties of SVM

- Fisher consistency (Lin, DM & KD 2002)

$$\arg_f \min \; E[(1 - Yf(X))_+ | X = x] = sign(p(x) - 1/2)$$

  where $p(x) = P(Y = 1 | X = x)$
- SVM approximates the Bayes decision rule

$$\phi_B(x) = sign(p(x) - 1/2).$$

- Bayes risk consistency (Zhang, AOS 2004, Steinwart, IEEE IT 2005, Bartlett et al., JASA 2006)

$$R(\hat{f}_{SVM}) \to R(\phi_B) \text{ in prob.}$$

  under universal approximation condition on $\mathcal{H}_K$
- Rate of convergence (Steinwart et al., AOS 2007)

Main Questions

- Recursive Feature Elimination (Guyon et al., ML 2002): backward elimination of variables based on the fitted coefficients of the linear SVM

- What is the statistical behavior of the coefficients?
- What determines their variances?
- Study asymptotic properties of the coefficients of the linear SVM.

Something New, Old, and Borrowed

- The hinge loss:
  not everywhere differentiable,
  no closed form expression for the solution
- Useful link:
  $sign(p(x) - 1/2)$, the population minimizer w.r.t. the hinge loss is the median of Y at x.
- Asymptotics for least absolute deviation (LAD) estimators (Pollard, ET 1991)
- Convexity of the loss

Preliminaries

- (X, Y): a pair of random variables with $X \in \mathcal{X} \subset \mathbb{R}^d$ and $Y \in \{1, -1\}$
- $P(Y = 1) = \pi_+$ and $P(Y = -1) = \pi_-$
- Let f and g be the densities of X given $Y = 1$ and -1.
- With $\widetilde{x} = (1, x_1, \ldots, x_d)^\top$ and $\beta = (b, w^\top)^\top$,

$$h(x; \beta) = w^\top x + b = \widetilde{x}^\top \beta$$

- $\widehat{\beta}_{\lambda,n} = \arg\min_\beta \frac{1}{n} \sum_{i=1}^{n} (1 - y_i h(x_i; \beta))_+ + \lambda \|w\|^2$

Population Version

- $L(\beta) = \mathbb{E}\Big[1 - Yh(X;\beta)\Big]_+$
- $\beta^* = \arg\min_\beta L(\beta)$
- The gradient of $L(\beta)$:

$$S(\beta) = -\mathbb{E}\Big(\psi(1 - Yh(X;\beta))Y\widetilde{X}\Big)$$

  where $\psi(t) = I(t \geq 0)$
- The Hessian matrix of $L(\beta)$:

$$H(\beta) = \mathbb{E}\Big(\delta(1 - Yh(X;\beta))\widetilde{X}\widetilde{X}^\top\Big)$$

  where δ is the Dirac delta function.

More on $H(\beta)$

- The Hessian matrix of $L(\beta)$:

$$H(\beta) = \mathbb{E}\Big(\delta(1 - Yh(X;\beta))\widetilde{X}\widetilde{X}^\top\Big)$$

$$\begin{aligned} H_{j,k}(\beta) &= \mathbb{E}\Big(\delta(1 - Yh(X;\beta))X_jX_k\Big) \quad \text{for} \quad 0 \le j,k \le d \\ &= \pi_+ \int_{\mathcal{X}} \delta(1 - b - w^\top x)x_jx_kf(x)dx \\ &+ \pi_- \int_{\mathcal{X}} \delta(1 + b + w^\top x)x_jx_kg(x)dx. \end{aligned}$$

- For a function s on $\mathcal{X}$, define the Radon transform $\mathcal{R}s$ of s for $p \in \mathbb{R}$ and $\xi \in \mathbb{R}^d$ as

$$(\mathcal{R}s)(p,\xi) = \int_{\mathcal{X}} \delta(p - \xi^\top x)s(x)dx.$$

  (the integral of s over hyperplanes $\xi^\top x = p$)
- $H_{j,k}(\beta) = \pi_+(\mathcal{R}f_{j,k})(1 - b, w) + \pi_-(\mathcal{R}g_{j,k})(1 + b, -w)$, where $f_{j,k}(x) = x_jx_kf(x)$ and $g_{j,k}(x) = x_jx_kg(x)$.

Regularity Conditions

(A1) The densities f and g are continuous and have finite second moments.

(A2) There exists $B(x_0, \delta_0)$ such that $f(x) > C_1$ and $g(x) > C_1$ for every $x \in B(x_0, \delta_0)$.

(A3) For some $1 \leq i^* \leq d$,

$$\int_{\mathcal{X}} \{x_{i^*} \geq G_{i^*}^-\} x_{i^*} g(x) dx < \int_{\mathcal{X}} \{x_{i^*} \leq F_{i^*}^+\} x_{i^*} f(x) dx$$

$$\text{or} \int_{\mathcal{X}} \{x_{i^*} \leq G_{i^*}^+\} x_{i^*} g(x) dx > \int_{\mathcal{X}} \{x_{i^*} \geq F_{i^*}^-\} x_{i^*} f(x) dx.$$

(when $\pi_+ = \pi_-$, it says that the means are different.)

(A4) Let $M^+ = \{x \in \mathcal{X} \mid \widetilde{x}^\top \beta^* = 1\}$ and $M^- = \{x \in \mathcal{X} \mid \widetilde{x}^\top \beta^* = -1\}$. There exist two subsets of M^+ and M^- on which the class densities f and g are bounded away from zero.

Bahadur Representation

- Bahadur (1966), *A Note on Quantiles in Large Samples*
- A statistical estimator is approximated by a sum of independent variables with a higher-order remainder.
- Let $\xi = F^{-1}(p)$ be the pth quantile of distribution F. For $X_1, \ldots, X_n \sim iid\ F$, the sample pth quantile is

$$\xi + \left[\sum_{i=1}^{n} I(X_i > \xi) - n(1-p) \right] / nf(\xi) + R_n,$$

where $f(x) = F'(x)$.

Bahadur-type Representation of the Linear SVM

Theorem

Suppose that (A1)-(A4) are met. For $\lambda = o(n^{-1/2})$, we have

$$\sqrt{n}\,(\widehat{\beta}_{\lambda,n} - \beta^*) = -\frac{1}{\sqrt{n}} H(\beta^*)^{-1} \sum_{i=1}^{n} \psi(1 - Y_i h(X_i; \beta^*)) Y_i \widetilde{X}_i + o_{\mathbb{P}}(1).$$

Recall that $H(\beta^*) = \mathbb{E}\Big(\delta(1 - Yh(X; \beta^*))\widetilde{X}\widetilde{X}^\top\Big)$ and $\psi(t) = I(t \geq 0)$.

Asymptotic Normality of $\widehat{\beta}_{\lambda,n}$

Theorem

Suppose (A1)-(A4) are satisfied. For $\lambda = o(n^{-1/2})$,

$$\sqrt{n}\,(\widehat{\beta}_{\lambda,n} - \beta^*) \to N\Big(0, H(\beta^*)^{-1}G(\beta^*)H(\beta^*)^{-1}\Big)$$

in distribution, where

$$G(\beta) = \mathbb{E}\Big(\psi(1 - Yh(X;\beta))\widetilde{X}\widetilde{X}^\top\Big).$$

Corollary

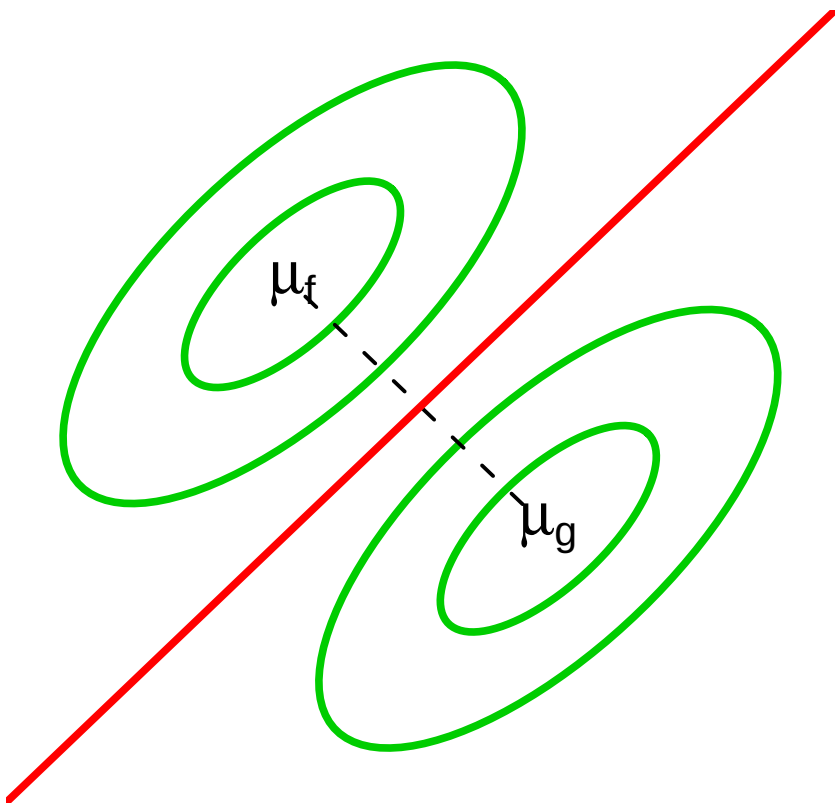
Under the same conditions as in Theorem,

$$\sqrt{n}\Big(h(x;\widehat{\beta}_{\lambda,n}) - h(x;\beta^*)\Big) \to N\Big(0, \tilde{x}^\top H(\beta^*)^{-1}G(\beta^*)H(\beta^*)^{-1}\tilde{x}\Big)$$

in distribution.

An Illustrative Example

- Two multivariate normal distributions in $\mathbb{R}^d$ with mean vectors μ_f and μ_g and a common covariance matrix Σ
- $\pi_+ = \pi_- = 1/2$.
- What is the relation between the Bayes decision boundary and the optimal hyperplane by the SVM, $h\left(x; \beta^*\right) = 0$?



Example

- The Bayes decision boundary (Fisher's LDA):

$$\left\{\Sigma^{-1}(\mu_f - \mu_g)\right\}^\top \left\{x - \frac{1}{2}(\mu_f + \mu_g)\right\} = 0.$$

- The hyperplane determined by the SVM :

$$\widetilde{x}^\top \beta^* = 0 \text{ with } S(\beta^*) = 0.$$

- β^* balances two classes within the margin

$$\mathbb{E}\Big(\psi(1 - Yh(X;\beta^*))Y\widetilde{X}\Big) = 0$$

$$\begin{aligned} P(h(X;\beta^*) \leq 1 | Y = 1) &= P(h(X;\beta^*) \geq -1 | Y = -1) \\ \mathbb{E}\Big(I\{h(X;\beta^*) \leq 1\}X_j | Y = 1\Big) &= \mathbb{E}\Big(I\{h(X;\beta^*) \geq -1\}X_j | Y = -1\Big) \end{aligned}$$

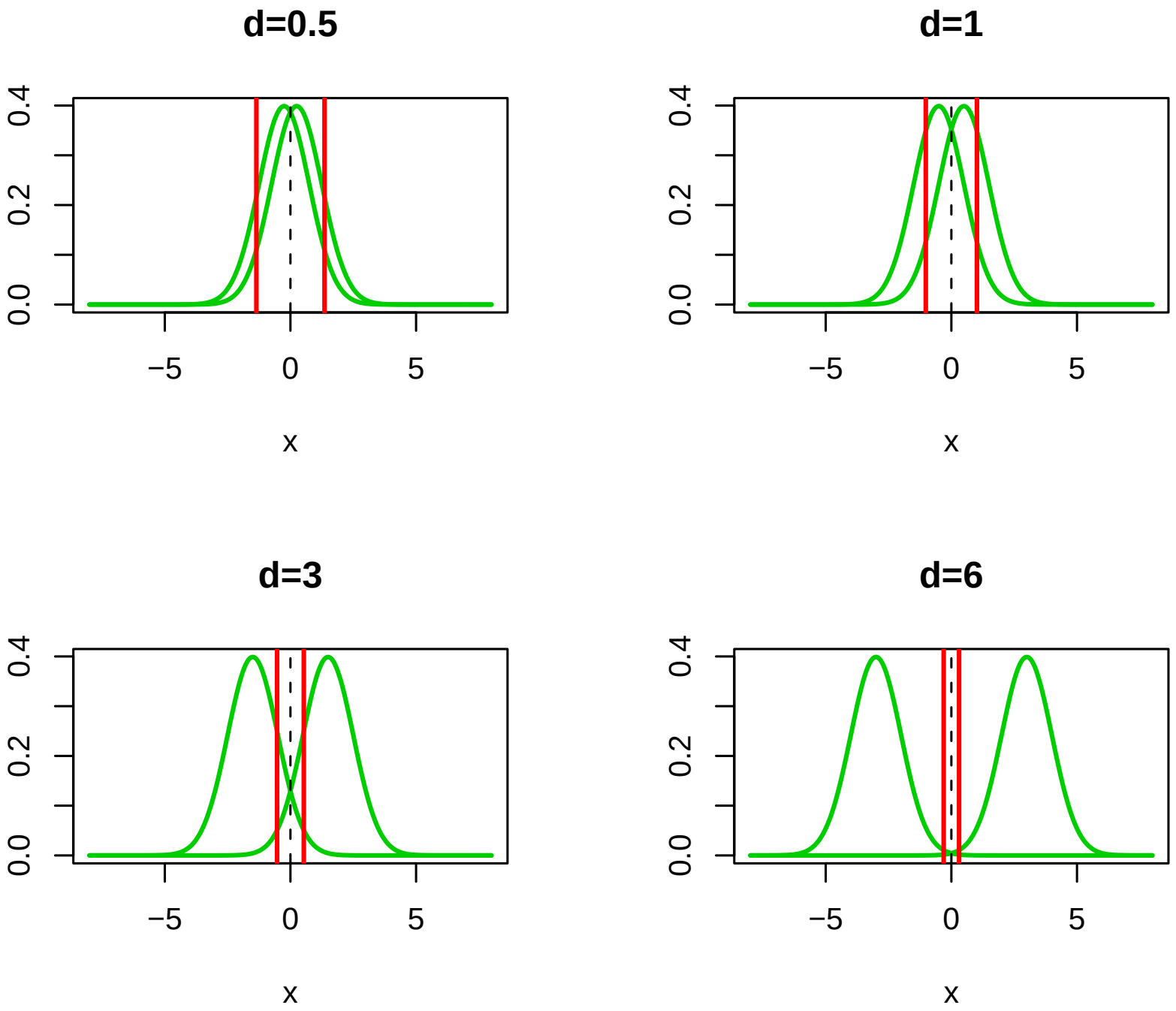
Example

- Direct calculation shows that

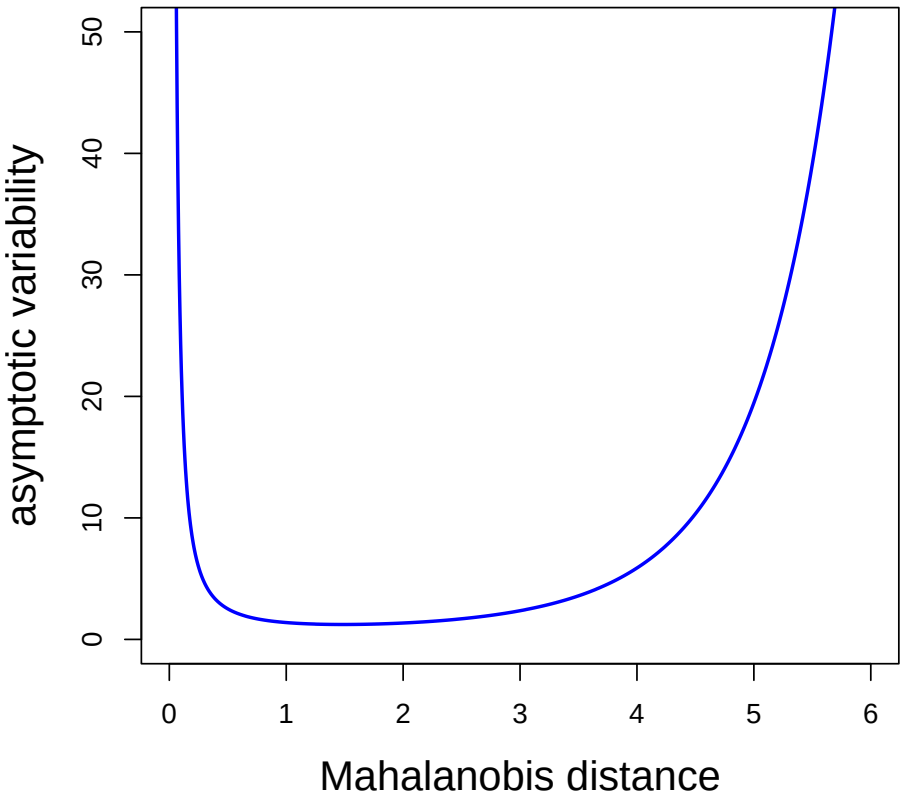
$$\beta^* = C(d_\Sigma(\mu_f, \mu_g))\begin{bmatrix} -\frac{1}{2}(\mu_f + \mu_g)^\top \\ I_d \end{bmatrix}\Sigma^{-1}(\mu_f - \mu_g),$$

  where $d_\Sigma(\mu_f, \mu_g)$ is the Mahalanobis distance between the two distributions.
- The linear SVM is equivalent to Fisher's LDA.
- The assumptions (A1)-(A4) are satisfied. So, the main theorem applies.
- Consider $d = 1$, $\mu_f + \mu_g = 0$, $\sigma = 1$, and $d_\Sigma(\mu_f, \mu_g) = |\mu_f - \mu_g|$.

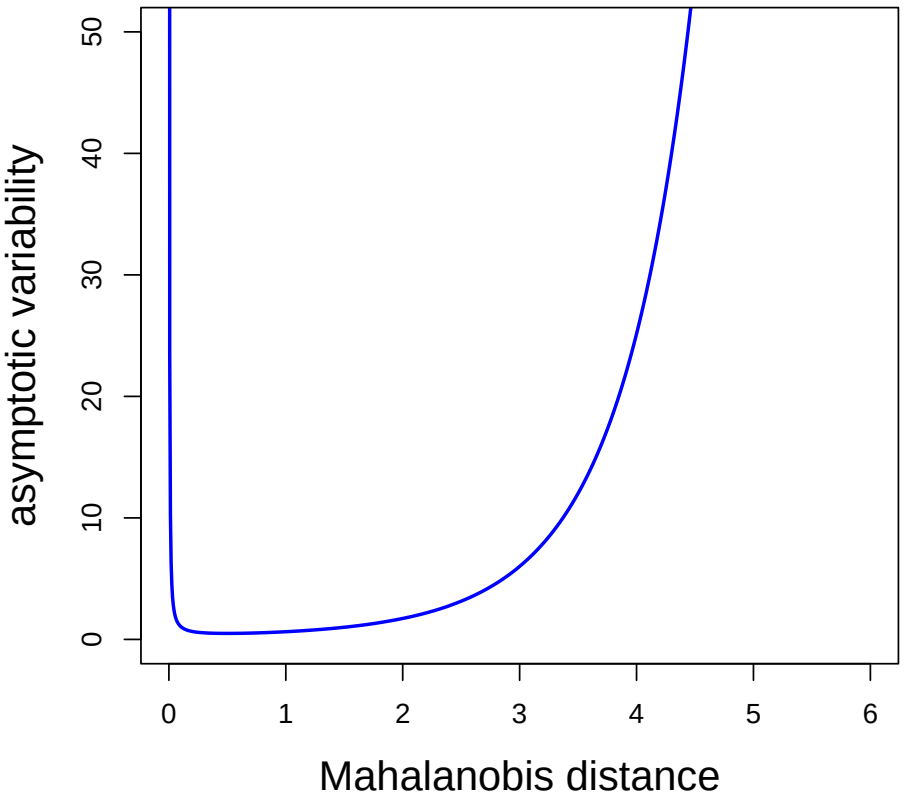
Distance and Margins



Asymptotic Variance



(a) Intercept



(b) Slope

Figure: The asymptotic variabilities of the intercept and the slope for the optimal hyperplane as a function of the Mahalanobis distance.

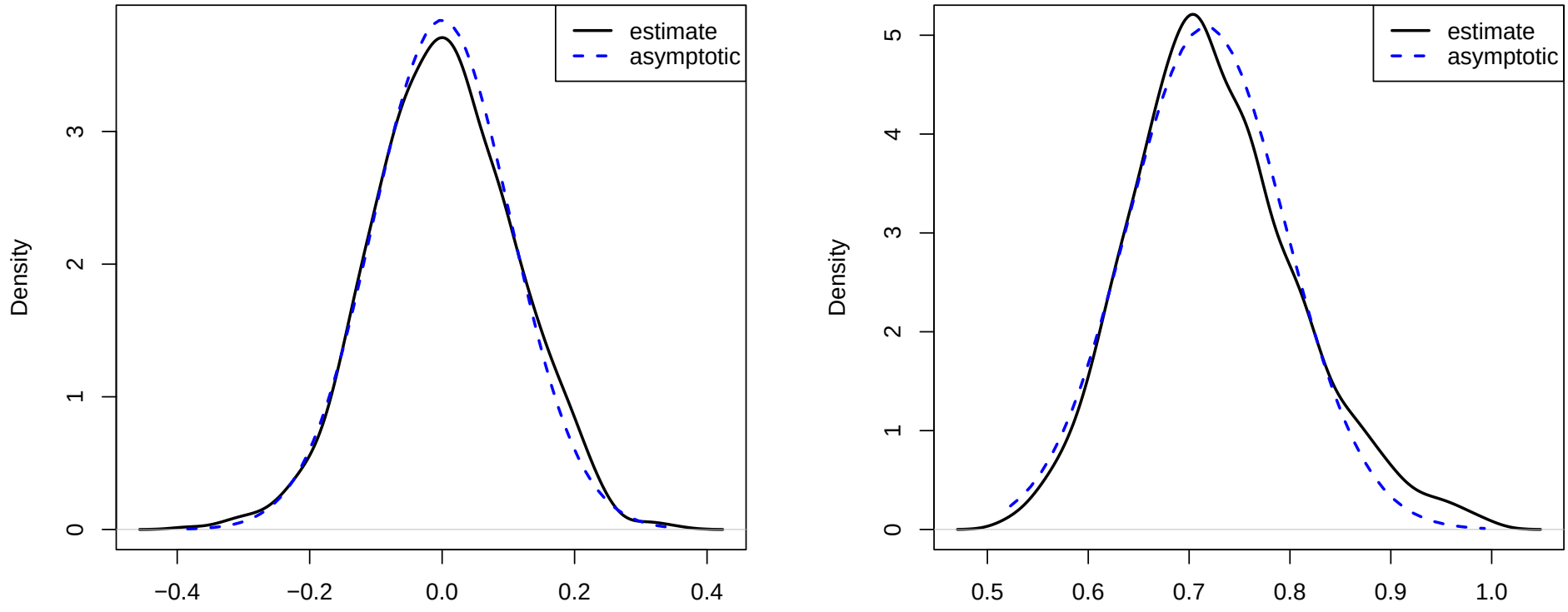
Bivariate Normal Example

- $\mu_f = (1,1)^\top$, $\mu_g = (-1,-1)^\top$ and $\Sigma = diag(1,1)$
- $d_\Sigma(\mu_f, \mu_g) = 2\sqrt{2}$ and the Bayes error rate is 0.07865.
- Find $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \arg\min_\beta \frac{1}{n}\sum_{i=1}^{n}(1 - y_i\, \widetilde{x}_i^\top \beta)_+$

Estimates	Sample size n			Optimal coefficients
	100	200	500	
$\hat{\beta}_0$	0.0006	-0.0013	0.0022	0
$\hat{\beta}_1$	0.7709	0.7450	0.7254	0.7169
$\hat{\beta}_2$	0.7749	0.7459	0.7283	0.7169

Table: Averages of estimated optimal coefficients over 1000 replicates.

Sampling Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$



(a) $\hat{\beta}_0$ (b) $\hat{\beta}_1$

Figure: Estimated sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

Type I error rates

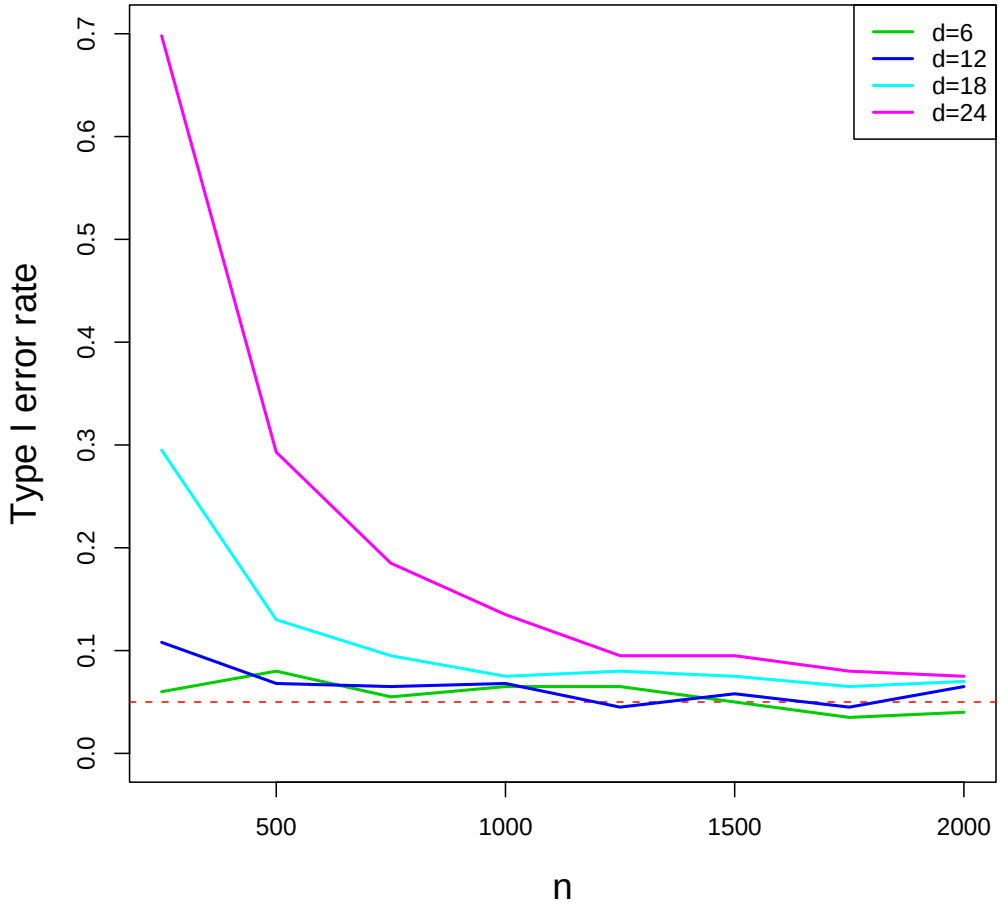


Figure: The median values of the type I error rates in variable selection when $\mu_f = (\mathbf{1}_{d/2}, \mathbf{0}_{d/2})^\top$, $\mu_g = \mathbf{0}_d^\top$, and $\Sigma = I_d$

Concluding Remarks

- Examine asymptotic properties of the coefficients of variables in the linear SVM.
- Establish Bahadur type representation of the coefficients.
- How the margins of the optimal hyperplane and the underlying probability distribution characterize their statistical behavior
- Variable selection for the SVM in the framework of hypothesis testing
- For practical applications, need consistent estimators of $G(\beta^*)$ and $H(\beta^*)$.
- Extension of the SVM asymptotics to the nonlinear case
- Explore a different scenario where d also grows with n.

Reference

- *A Bahadur Representation of the Linear Support Vector Machine*, Koo, J.-Y., Lee, Y., Kim, Y., and Park, C., *Journal of Machine Learning Research* (2008).

  Available at www.stat.osu.edu/∼yklee or http://www.jmlr.org/.