Distance Weighted Discrimination

(improving SVM with SOCP)

Steve Marron
Statistics, UNC

Mike Todd
Operations Research,
Cornell

www.unc.edu/depts/statistics/postscript/papers/marron/HDD/DWD

Introduction, problem

Two class discrimination

have probability distrib. on $X \times Y$

$X \subseteq \mathbb{R}^d$, $Y = \{+1, -1\}$, unknown

have sample $\{(x_i, y_i) : i = 1, 2, \cdots, n\}$,

training set.

Want to deduce a method to predict, for a new observation (x, y) from the dist., with x known, what y is.

Example: components of x:

medical history, results of tests.

$$y = \begin{cases} +1 & \text{have certain disease} \\ -1 & \text{don't.} \end{cases}$$

Rule: $y = f(x)$

↑

constructed from training set

Linear rule: for some $w \in \mathbb{R}^d$, $\beta \in \mathbb{R}$

$$f(x) = \text{sign}(w^T x + \beta)$$

Seems substantial limitation. But can expand vector x, eg. to $(x_1, \cdots, x_d, x_1^2, x_1 x_2, \cdots, x_d^2)$, so not bad.

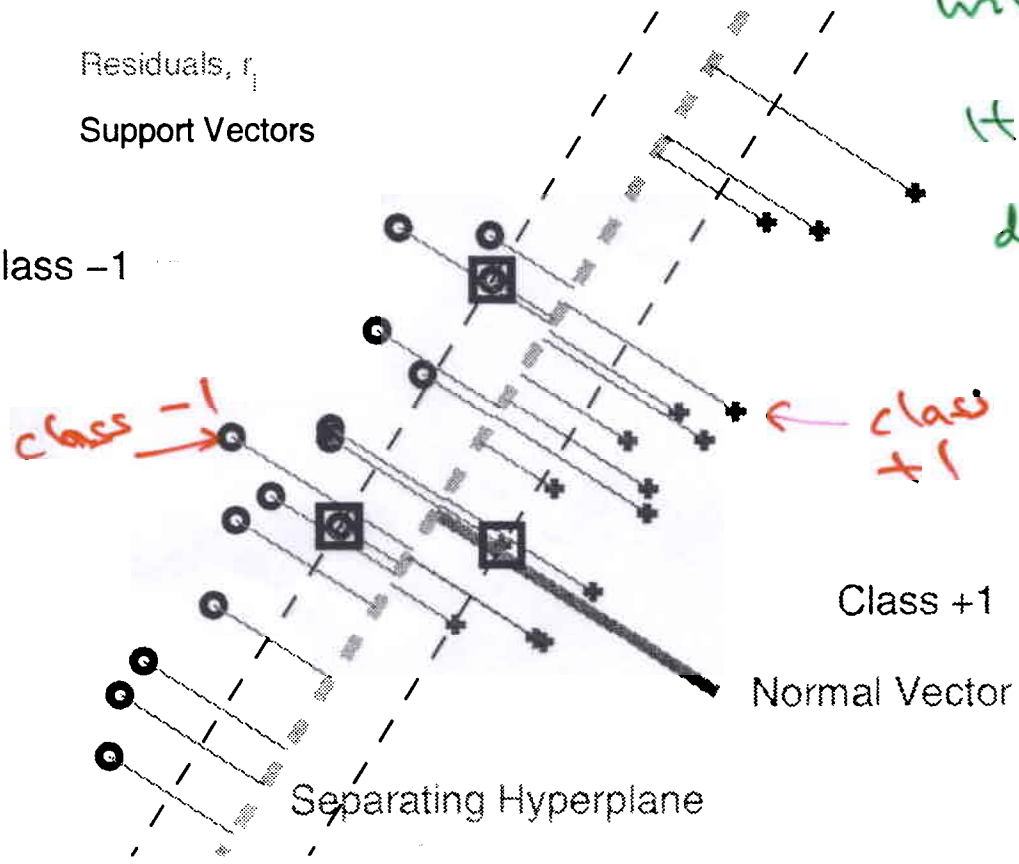
Support Vector Machine (Vapnik, 1982)

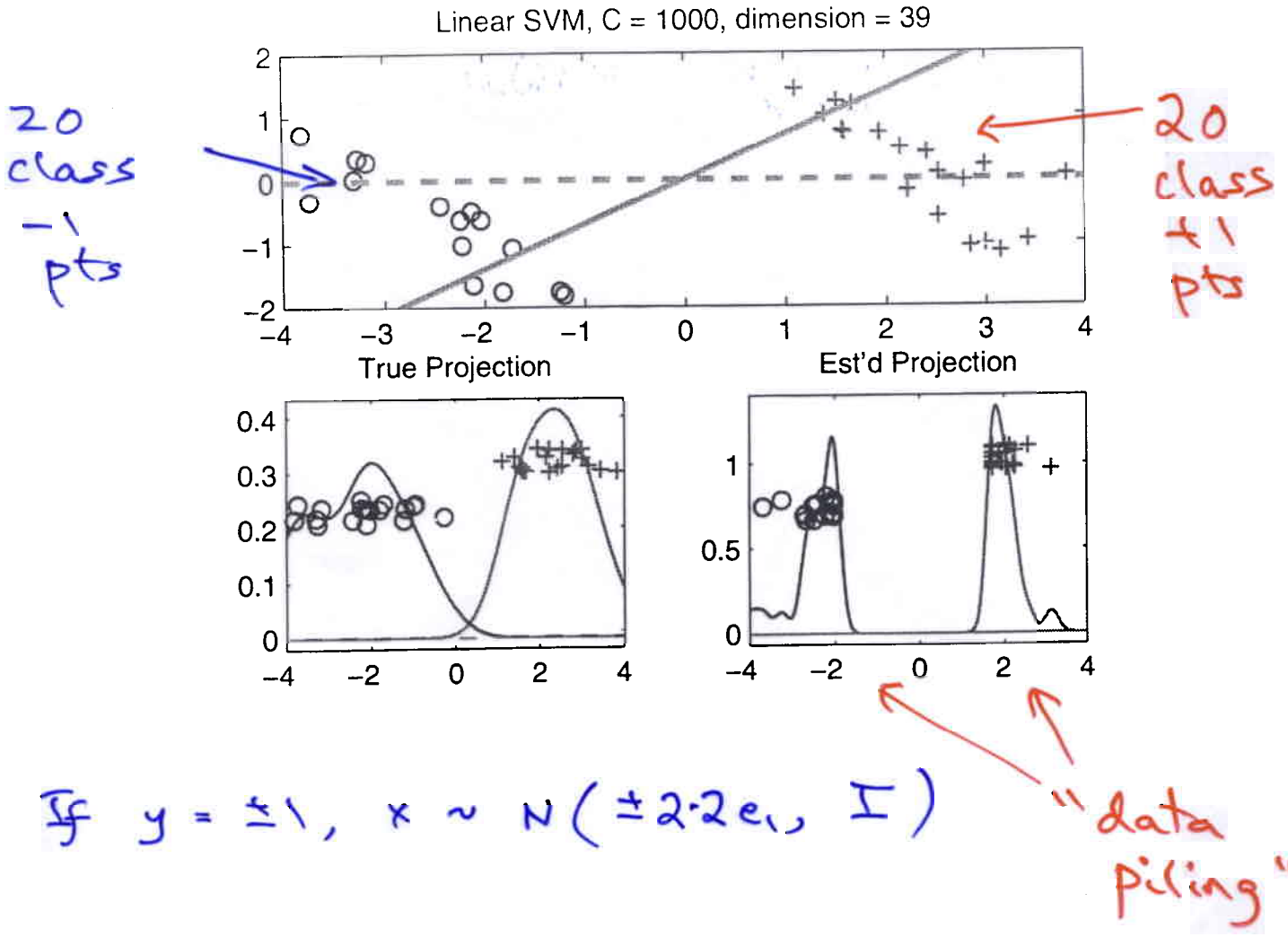
What's wrong with this ??

HDLSS setting

$d >> 1$, $n << d$

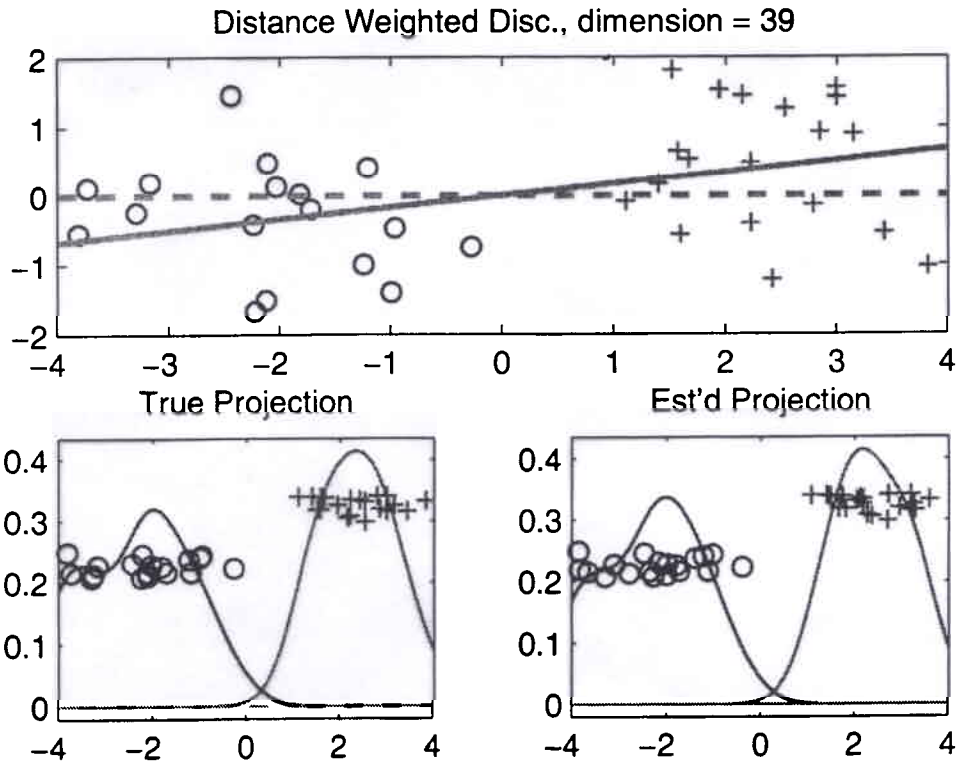
Sensitivity





If $y = \pm 1$, $x \sim N(\pm 2.2 e_1, I)$

"data piling"



Optimization Formulations

Have $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$, $i = 1, \dots, n$.

Let $X = [x_1 \mid \cdots \mid x_n]$, $y = (y_i) \in \mathbb{R}^n$,

$Y = \mathrm{Diag}(y) \in \mathbb{R}^{n \times n}$ $\quad (X \in \mathbb{R}^{d \times n})$.

Given $w \in \mathbb{R}^d$, $\beta \in \mathbb{R}$, residuals

$$\bar{r}_i := y_i (x_i^T w + \beta_i) \quad \text{or}$$

$$\bar{r} := Y(X^T w + \beta e) = YX^T w + \beta y.$$

Want all components of $\bar{r}$ "large and pos."

May be impossible to even make all pos. Introduce $\xi \in \mathbb{R}^n_+$ errors, to give perturbed residuals

$$r := \bar{r} + \xi = YX^T w + \beta y + \xi.$$

SVM: "$\max_{w,\beta} \min_i r_i$" $\quad \|r^{-1}\|_\infty$

DWD: "$\min_{w,\beta} \sum_i 1/r_i$" $\quad \|r^{-1}\|_1$

Support vector machine

First, assume no errors.

$$\max_{\delta,w,\beta} \delta, \quad \bar{r} = YX^T w + \beta y, \quad \bar{r} \geq \delta e, \quad \tfrac{1}{2} w^T w \leq \tfrac{1}{2}$$

$$\Updownarrow$$

$$\min_{w,\beta} \tfrac{1}{2} w^T w, \quad YX^T w + \beta y \geq e.$$

Now allow perturbations, penalized via L_1 norm

$$(P_{SVM}) \qquad \min_{w,\beta,\xi} \tfrac{1}{2} w^T w + Ce^T \xi$$

$$YX^T w + \beta y + \xi \geq e,$$

$$\xi \geq 0.$$

QP with dual

$$(D_{SVM}) \qquad \max_{\alpha} -\tfrac{1}{2} \alpha^T Y X^T X Y \alpha + e^T \alpha$$

$$y^T \alpha = 0$$

$$0 \leq \alpha \leq Ce.$$

Optimality Conditions:

$$e_+^T \alpha_+ = e_-^T \alpha_-$$

$$\left(X_+ \alpha_+ - X_- \alpha_-\right) = X^T Y \alpha = w \quad , \quad y^T \alpha = 0$$

$$s = Y X^T w + \beta y + \xi - e \geq 0, \ \alpha \geq 0 \ , \quad s^T \alpha = 0$$

$$Ce - \alpha \geq 0 \ , \quad \xi \geq 0 \ , \quad (Ce - \alpha)^T \xi = 0 .$$

$$X = [X_+ \ , \ X_-] \ , \ y = \begin{pmatrix} 1 \\ \vdots \\ -1 \end{pmatrix} , \ \alpha = \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix}$$

Geometric interpretation:

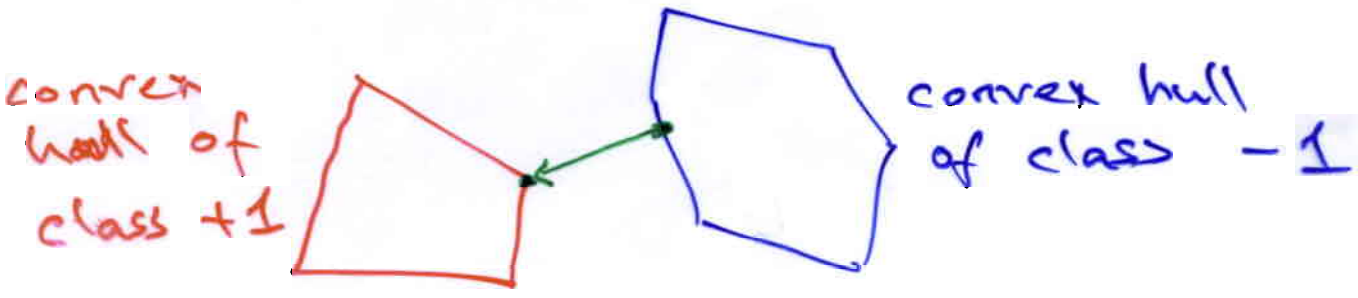
Assume optimal $\alpha < Ce$.

Let $\alpha = \gamma \hat{\alpha}$, $e_+^T \hat{\alpha}_+ = e_-^T \hat{\alpha}_- = 1$,

so $w = \gamma \left(X_+ \hat{\alpha}_+ - X_- \hat{\alpha}_-\right)$

By optimizing over γ for fixed $\hat{\alpha}$,

$$(D_{SVM}) \equiv \max \ 2 / \| X_+ \hat{\alpha}_+ - X_- \hat{\alpha}_- \|$$



Distance weighted discrimination

Now minimize sum of reciprocals of perturbed residuals, + penalty term:

$$\min_{r,w,\beta,\xi} \quad \sum_i \frac{1}{r_i} + Ce^T\xi$$

$$r = YX^Tw + \beta y + \xi \geq 0$$

$$\tfrac{1}{2} w^Tw \leq \gamma_2, \quad \xi \geq 0.$$

(Could consider $\sum f(r_i)$, f smooth, convex, $\to +\infty$ as arg $\downarrow 0$.)

Reformulate as SOCP. Use second-order Lorentz cone

$$S_{m+1} := \{(\psi; u) \in \mathbb{R}^{m+1} : \psi \geq \|u\|\}.$$

Write $r_i = \rho_i - \sigma_i$, where $\rho_i = \dfrac{r_i + \frac{1}{r_i}}{2}$,

$\sigma_i = \dfrac{\frac{1}{r_i} - r_i}{2}$. Then $\rho_i + \sigma_i = \frac{1}{r_i}$, and

$\rho_i^2 - \sigma_i^2 = 1$, or $(\rho_i; \sigma_i; 1) \in S_3$. Also,

$(1, w) \in S_{d+1}$.

$$\min_{\omega, w, \beta, \xi, \rho, \sigma, \tau} \quad Ce^T\xi + e^T\rho + e^T\sigma$$

(P_{DWD})

$$\begin{aligned} Y X^T w + \beta z + \xi - \rho + \sigma &= 0 \\ \omega &= 1 \\ \tau &= e \end{aligned}$$

$$(\omega, w) \in S_{d+1}, \ \xi \geq 0, \ (\rho_i, \sigma_i, \tau_i) \in S_3, \text{ all } i.$$

SOCTs have duals: get

$$\max_{\alpha, \eta, \gamma, \pi, \rho, \chi, \lambda, \mu, \nu} \quad \eta + e^T\gamma$$

(D_{DWD})

$$\begin{aligned} \eta + \pi &= 0 \\ \eta + \rho &= 0 \\ X Y \alpha &= 0 \\ y^T \alpha &= 0 \\ \alpha + \chi &= Ce \\ -\alpha + \lambda &= e \\ \alpha + \mu &= e \\ \gamma + \nu &= 0 \end{aligned}$$

$$(\pi; \rho) \in S_{d+1}, \ \chi \geq 0, \ (\lambda_i; \mu_i; \nu_i) \in S_3 \text{ all } i.$$

Can greatly simplify: equivalent to

$$\max_{\alpha} \; -\|XY\alpha\| + 2e^T\sqrt{\alpha}, \quad y^T\alpha = 0, \; 0 \le \alpha \le Ce.$$

Can check (P_{DWD}) and (D_{DWD}) have strictly feasible solutions, so have optimal soln's and no duality gap.

Optimality conditions:

Want feasibility + orthog:

$$(\pi;\rho)^T(w;w) = x^T\xi = (\lambda_i;\mu_i;\nu_i)^T(\rho_i;\tau_i;\tau_i) = 0$$

Note: for SOC: $(\psi;u)^T(\phi;v) = 0$ iff

$(\psi;u) = 0$ or $(\phi;v) = 0$ or $\|u\| = \psi, \|v\| = \phi, u\alpha - v$.

$$YX^Tw + \beta y + \gamma - \rho + \tau = 0, \quad y^T\alpha = 0$$

$$\alpha \ge 0, \; \alpha \le Ce, \; \xi \ge 0, \; (Ce - \alpha)^T\xi = 0,$$

Either $XY\alpha = 0$ and $\|w\| \le 1$ or

$$w = \frac{XY\alpha}{\|XY\alpha\|} = \frac{X_+\alpha_+ - X_-\alpha_-}{\|X_+\alpha_+ - X_-\alpha_-\|}$$

$$\rho_i = \frac{\alpha_i + 1}{2\sqrt{\alpha_i}}, \quad \tau_i = \frac{\alpha_i - 1}{2\sqrt{\alpha}} \quad \text{all } i.$$

Geometric interpretation:

Again, assume all optimal $\alpha < Ce$.

Write $\alpha = \gamma \hat{\alpha}$, $e_+^T \hat{\alpha}_+ = e_-^T \hat{\alpha}_- = 1$.

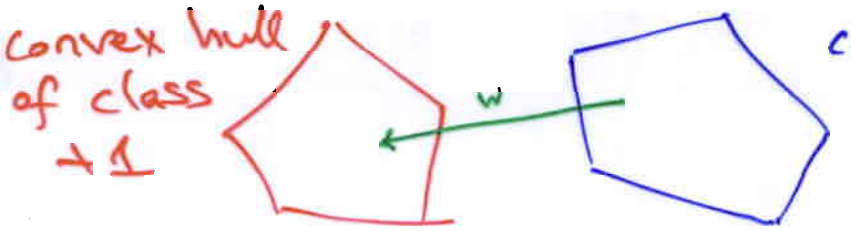
Then (D_{SVD}) equiv to

$$\max_{\gamma,\, \hat{\alpha}} \; -\gamma \|XY\hat{\alpha}\| + 2\sqrt{\gamma}\, e^T \sqrt{\hat{\alpha}}, \quad \hat{\alpha} \ldots$$

For fixed $\hat{\alpha}$, choose $\gamma = \left(e^T \sqrt{\hat{\alpha}} \,/\, \|XY\hat{\alpha}\|\right)^2$,

$$\Rightarrow \max \frac{\left(e_+^T \sqrt{\hat{\alpha}_+} + e_-^T \sqrt{\hat{\alpha}_-}\right)^2}{\| X_+ \hat{\alpha}_+ - X_- \hat{\alpha}_- \|} \cdot \quad \begin{array}{l} e_+^T \hat{\alpha}_+ = 1, \\ e_-^T \hat{\alpha}_- = 1, \\ \hat{\alpha}_+, \hat{\alpha}_- \geq 0. \end{array}$$

So we want to minimize the norm between a convex combination of the class $+1$ and the class -1 points, but scaled down by the square of the sum of the square roots of all convex weights (this forces all $\alpha_i > 0$).



Dimension reduction

(P_{DWD}) has $2n+1$ eq. constraints,

$d + 2 + 4n$ variables (S_{d+1}).

Bad in HDLSS case, $d >> n$.

Cure: $X = QR$, $Q \in \mathbb{R}^{d \times n}$, orthog cols

$R \in \mathbb{R}^{n \times n}$, upper triang.

Replace $YX^T w$ ... $(w; w) \in S^{d+1}$

with $YR^T \bar{w}$ ... $(w; \bar{w}) \in S^{n+1}$.

"Correspondence" of feasible solutions via $\bar{w} = Q^T w$. Solve smaller problem to get optimal $\bar{w}$, and $w = Q\bar{w}$ soln to original problem.

A similar idea works to treat the nonlinear case via a kernel function, similar to with the SVM.

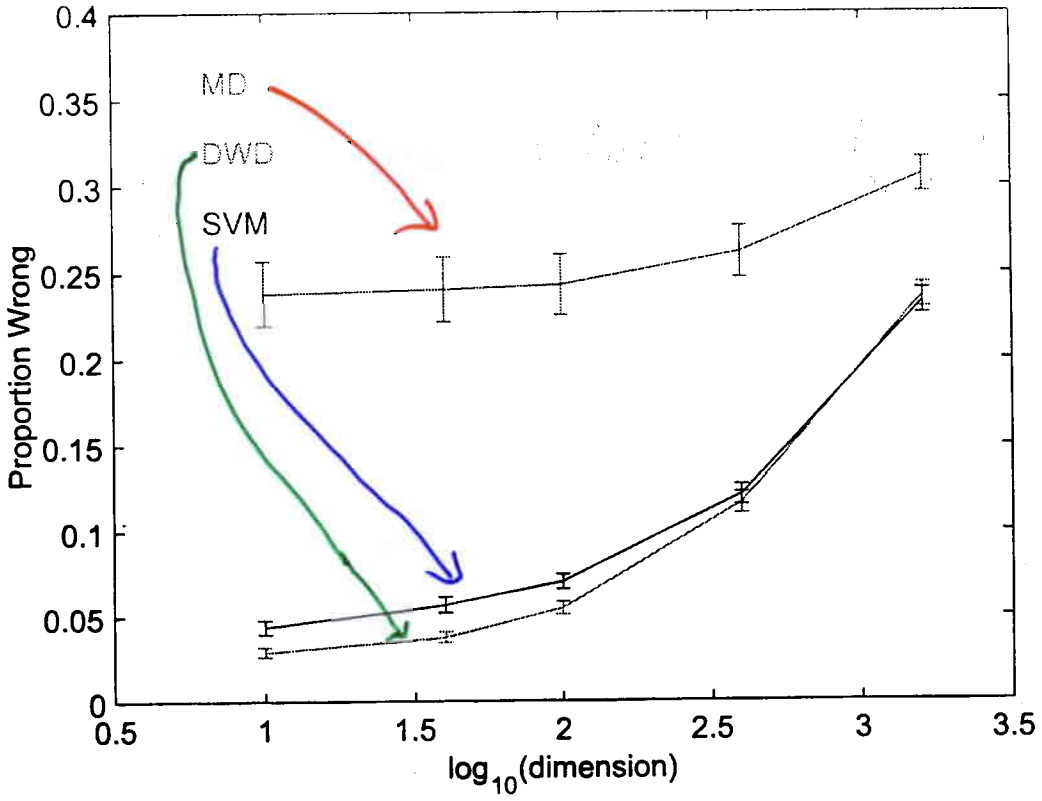


FIGURE 7: *Simulation comparison, for the "wobble" distribution. This is a case where DWD gives superior performance to MD and SVM.*

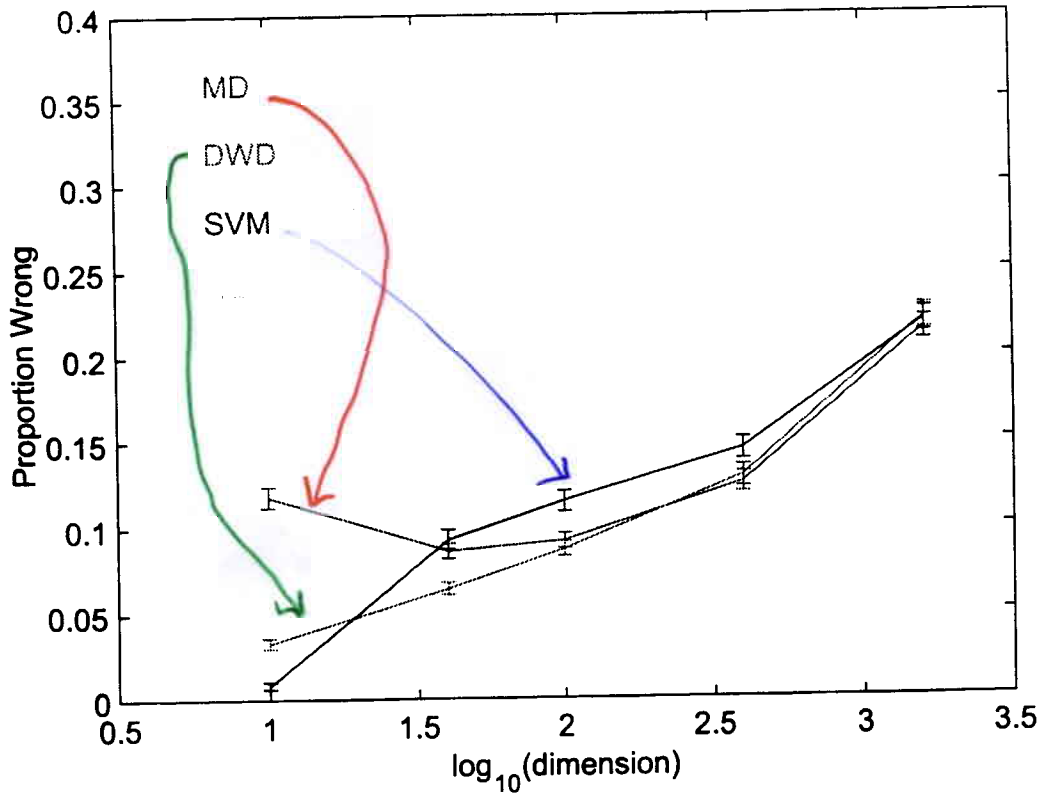


FIGURE 8: *Simulation comparison, for the "nested sphere" distribution. This case shows a fair overall summary, because each method is best for some d, and DWD tends to be near whichever method is best.*

Group 1 Luminal cancer vs. other cancer types and normals: A first rough classification suggested by clustering of the data in Perou et al. ([12]). Tested using $n_+ = 47$ and $n_- = 38$ training cases, and 51 test cases.

Group 2 Luminal A vs. Luminal B&C: an important distinction that was linked to survival rate in Perou et al. ([12]). Tested using $n_+ = 35$ and $n_- = 15$ training cases, and 21 test cases.

Group 3 Normal vs. Erb & Basal cancer types. Tested using $n_+ = 13$ and $n_- = 25$ training cases, and 30 test cases.

Group 4 Erb vs. Basal cancer types. Tested using $n_+ = 11$ and $n_- = 14$ training cases, and 21 test cases.

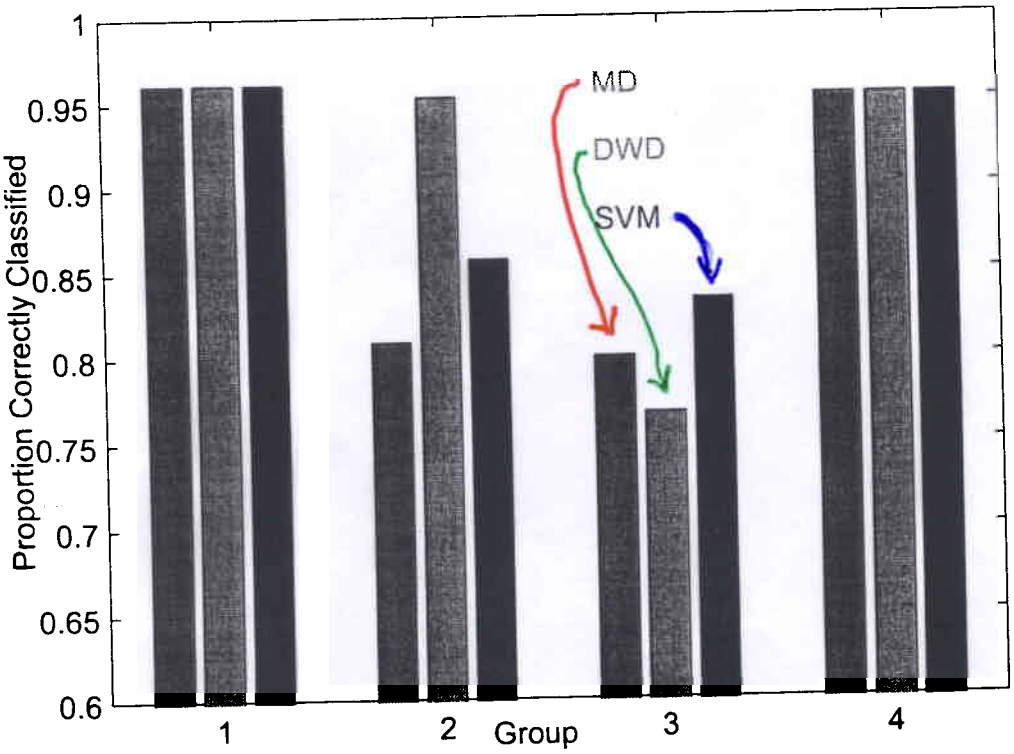


FIGURE 9: *Graphical summary of correct classification rates for gene expression data.*