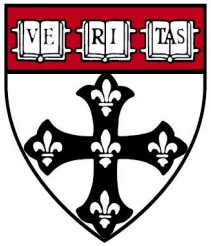


Sparse Distance Weighted Discrimination Method

Lingsong Zhang and Xihong Lin
Harvard School of Public Health

Statistics Workshop at Yale University
Innovation and Inventiveness in Statistics Methodologies
May 14-17, 2009

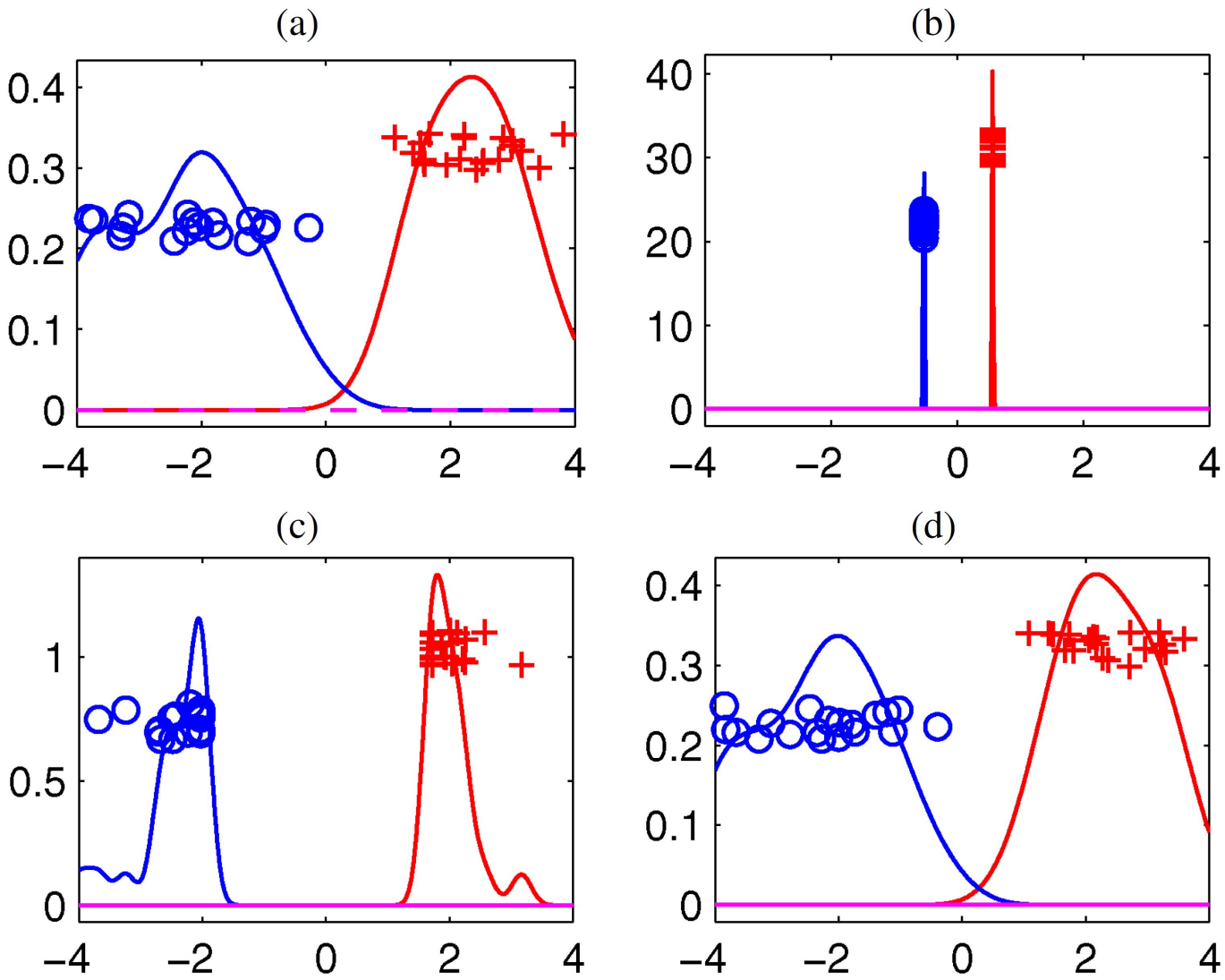
Distance Weighted Discrimination

- Support Vector Machine (SVM) is widely used in machine learning and statistical applications.
  - Pros: Fisher consistency, robustness of model specification
  - Cons: Data piling when p is close or larger than n

- Distance Weighted Discrimination (DWD) was proposed to overcome the undesired data piling feature (Marron et al, 2007).

  - Maximize the smallest distance (to the separating hyperplane) $\longrightarrow$ Minimize the average inverse distance (to the separating hyperplane)

Data piling - Example from Marron et al (2007)



Motivation of Sparse DWD

- Pros of DWD method
  - Fisher consistent when n large (our result)
  - Solve one type of overfitting: avoid data piling when p is large
- Cons of DWD method
  - Too many noisy variables included
  - less interpretability and less power
- Small n and large p classification problem is challenging

Consider a p-dimensional two-group classification ($Y = +1$, or -1) training data set, with sample size (n_1, n_2), and all p variables are independent. For all the variables X_i, the cases and the controls are from the same distribution. When $n = n_1 + n_2$ is fixed and $p \to \infty$, there exists (w, β), that

$$\lim_{p \to \infty} P(y_i(x_i^T w - \beta) > 0, \forall i) = 1$$

Hall et al (2005) provides similar results from geometry view points.

Our work

- We propose sparse DWD methods
  - Simultaneously estimate the classifier and remove some noise variables.
  - Incorporate a sparse penalty
    - LASSO (Tibshirani, 1996), L_1 penalty
    - Adaptive LASSO (Zou, 2006), Weighted L_1 penalty
    - SCAD penalty (Fan and Li, 2001).
- We show that the DWD and SDWD methods are Fisher consistent
- We develop an oracle theory for sparse classifier. Both the adaptive lasso and scad penalties enjoy oracle property, but the lasso does not.
- See examples and simulations in the poster.