IIS-SQDA Notes

zxm

- high-dimensional setting, $\log p = O(n^\gamma), 0 < \gamma < 1$
- nonlinear classification
- interaction screening
- model:
  - predictor vector $Z = (Z_1, \ldots, Z_p)^T$
  - class label $\Delta \sim Binom(1, \pi)$, that is, *prior*
  $$\Pr(\Delta = 1) = \pi = \pi_1, \Pr(\Delta = 0) = 1 - \pi = \pi_2.$$
  - $z^{(k)} \sim \mathcal{N}_p(\mu_k, \Sigma_k)$, $k = 1, 2$, $z = \Delta z^{(1)} + (1 - \Delta) z^{(2)}$, that is, *likelihood*
  $$f_k(z) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(z - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right), k = 1, 2.$$
  - **Bayes rule**: to maximize *posterior*, that is
  $$z_{new} \to \arg\max_k f_k(z)\pi_k.$$
  More precisely,
  $$z_{new} \to \text{ class 1 iff } Q(z) > 0,$$
  where
  $$Q(z) = \frac{1}{2} z^T(\Omega_2 - \Omega_1)z + (\Omega_1\mu_1 - \Omega_2\mu_2)^T z + \zeta$$
  where precision matrices $\Omega_k = \Sigma_k^{-1}$, $\zeta = \frac{1}{2}\left(-\mu_1^T\Omega_1\mu_1 + \mu_2^T\Omega_2\mu_2 + \log\frac{|\Omega_1|}{|\Omega_2|}\right) + \log\frac{\pi}{1-\pi}$.
  - **LDA**: assumes $\Sigma_1 = \Sigma_2 = \Sigma$ (easily violated). Then
  $$Q(z) = (\mu_1 - \mu_2)^T \Sigma^{-1}\left(z - \frac{\mu_1 + \mu_2}{2}\right) + \log\frac{\pi}{1 - \pi}.$$
  - estimates: plug in $\pi = \frac{n_1}{n}$, $\mu_k = \hat{\mu}_k$, $\Sigma_k = \hat{\Sigma}_k$, $\Sigma = \hat{\Sigma}$ (pooled sample covariance matrix)
- Why use screening?
  1. $p > n$, LDA and QDA inapplicable due to *singularities of sample covariance matrices*
  2. noise accumulation
  3. computational cost, e.g. 1000 main effects $\Rightarrow$ 500,500 interactions
- **sparsity** assumption
- IIS-SQDA
  1. **Stage I**: IIS (Innovated Interaction Screening):
  - transforming the original p-dimensional feature vector
  2. **Stage II**: SQDA (Sparse Quadratic Discriminant Analysis):
  - further selecting important interactions and main effects, and
  - simultaneously conducting classification
  3. **Sure Screening Property**
  4. *classification error* $\leq$ *oracle classification error* + *smaller order term*
- Competitive methods
  1. *Mai et al.*, DSDA, reformulationg the LDA problem as a penalized least squares regression
  2. Penalized Logistic Regression

IIS

- Goal: to find the index set $\mathcal{I}$ of interaction variables
- Let $\Omega = \Omega_2 - \Omega_1$ and $\delta = \Omega_1 \mu_1$ and assume $\mu_2 = 0$. Then $Q(z)$ becomes

$$Q(z) = \frac{1}{2} z^T \Omega z + \delta^T z + \zeta.$$

- Decomposition of $\mathcal{I}$:

$$\begin{aligned}\mathcal{I} &= \{1 \le j \le p : Z_j Z_l \text{ is an active interaction for some } 1 \le l \le p\} \\ &= \{1 \le j \le p : \Omega_{jl} \ne 0 \text{ for some } 1 \le l \le p\} \\ &= \mathcal{A}_1 \cup \mathcal{A}_2 \\ &= \{j : (\widetilde{\Sigma}_1)_{jj} \ne 0\} \cup \{j : (\widetilde{\Sigma}_2)_{jj} \ne 0\},\end{aligned}$$

  where

$$\widetilde{\Sigma}_1 = \Omega_1 \Sigma_2 \Omega_1 - \Omega_1 = Var(\Omega_1 z^{(2)}) - Var(\Omega_1 z^{(1)})$$

  and

$$\widetilde{\Sigma}_2 = \Omega_2 - \Omega_2 \Sigma_1 \Omega_2 = Var(\Omega_2 z^{(2)}) - Var(\Omega_2 z^{(1)}).$$

- Oracle-assisted IIS.
  - data points $\{(z_i^T, \Delta_i)\}_{i=1}^n$
  - $n_1 = \sum_{i=1}^n \Delta_i$, $n_2 = n - n_1$
  - data matrix $Z = (z_1, \ldots, z_n)^T$
  - transformed data matrix $\widetilde{Z} = Z\Omega_1$, $\check{Z} = Z\Omega_2$
  - test statistics

$$\widetilde{D}_j = \log \widetilde{\sigma}_j^2 - \sum_{k=1}^2 \frac{n_k}{n} \log \left[\left(\widetilde{\sigma}_j^{(k)} \right)^2 \right]$$

$$\check{D}_j = \log \check{\sigma}_j^2 - \sum_{k=1}^2 \frac{n_k}{n} \log \left[\left(\check{\sigma}_j^{(k)} \right)^2 \right]$$

    for $j = 1, \ldots, p$, where
    - $\widetilde{\sigma}_j^2$: pooled sample variance for $\widetilde{Z}_j$
    - $\check{\sigma}_j^2$: pooled sample variance for $\check{Z}_j$
    - $\left(\widetilde{\sigma}_j^{(k)} \right)^2$: with-in sample variance for $\widetilde{Z}_j$
    - $\left(\check{\sigma}_j^{(k)} \right)^2$: with-in sample variance for $\check{Z}_j$
  - denote

$$\widehat{\mathcal{A}}_1 = \{j : \widetilde{D}_j \text{ greater than some threshold in } (cn^{-\kappa}, 2cn^{-\kappa})\}$$
$$\widehat{\mathcal{A}}_2 = \{j : \check{D}_j \text{ greater than some threshold in } (cn^{-\kappa}, 2cn^{-\kappa})\}$$

    where $0 < 2\kappa < 1 - \gamma$. Then with probability tending to 1,

$$\widehat{\mathcal{A}}_k = \mathcal{A}_k.$$

- IIS with unknown precision matrices.
  - replace Ω_k with *acceptable estimator* $\widehat{\Omega}_k$
  - denote

$$\widehat{\mathcal{A}}_1 = \{j : \widetilde{D}_j \text{ greater than some threshold in } (cn^{-\kappa} + T_{n,p}, 2cn^{-\kappa} - T_{n,p})\}$$
$$\widehat{\mathcal{A}}_2 = \{j : \check{D}_j \text{ greater than some threshold in } (cn^{-\kappa} + T_{n,p}, 2cn^{-\kappa} - T_{n,p})\}$$

    where $T_{n,p} = o(n^{-\kappa})$. Then with probability tending to 1,

$$\widehat{\mathcal{A}}_k = \mathcal{A}_k.$$

SQDA

- $d = |\widehat{\mathcal{I}}| = |\widehat{\mathcal{A}}_1 \cup \widehat{\mathcal{A}}_2|$, cardinality
- $\tilde{p} = 1 + p + C_d^2 + d = 1 + p + \frac{d(d+1)}{2}$
- augmented feature $X = (1, Z_1, \ldots, Z_p, Z_{j_1}^2, \ldots, Z_{j_d}^2, Z_{j_1}Z_{j_2}, \ldots)^T \in \mathbb{R}^{\tilde{p}}$, $j_1, \ldots, j_d \in \widehat{\mathcal{I}}$
- logistic regression model:
$$\log \frac{\Pr(\Delta = 1|x)}{\Pr(\Delta = 0|x)} = X^T\theta$$
Thus,
$$\Pr(\Delta = 1|x) = \frac{\exp(X^T\theta)}{1 + \exp(X^T\theta)} = 1 - \Pr(\Delta = 0|x).$$
- negative log-likelihood function (logistic loss function):
$$\begin{aligned}\ell_n(\theta) &= -\log \prod_{i=1}^n \frac{\exp(x_i^T\theta)^{\Delta_i}}{1 + \exp(x_i^T\theta)} \\ &= \sum_{i=1}^n \left[-\Delta_i x_i^T\theta + \log\left(1 + \exp(x_i^T\theta)\right)\right]\end{aligned}$$
- penalty function: elastic net penalty
$$pen(\theta) = \lambda_1\|\theta\|_1 + \lambda_2\|\theta\|_2^2$$
- regularization problem:
$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^{\tilde{p}}} \left[\frac{1}{n}\ell_n(\theta) + pen(\theta)\right]$$
- classification: LDA or QDA (based on whether interaction terms survived from screening and variable selection)