The Adaptive Dantzig Selector for Variable Selection and Estimation

Lee Dicker and Xihong Lin

Department of Biostatistics

Harvard School of Public Health

May 15, 2009

The Problem

Variable selection and estimation: Select predictors of a given outcome from a collection of potential predictors and estimate their effects.

$$y = X\beta^* + \epsilon,$$

Goal: Let $T^* = \{j;\ \beta_j^* \neq 0\}$. Identify T^* and $\beta_{T^*}^* = \{\beta_j^*\}_{j \in T^*}$.

- Penalized likelihood methods have been studied extensively over the past several years. Estimates for β^* are obtained by minimizing the negative log-likelihood plus a penalty term.
  - Non-smooth penalty term $\Rightarrow$ variable selection.
- The Dantzig selector (Candes and Tao, 2007) is an ℓ^1-minimization method for variable selection and estimation based on the normal score equations.

Score Equations $\longrightarrow$ Estimating Equations

The Dantzig Selector

$$\begin{array}{ll} \text{minimize} & ||\beta||_1 \\ \text{subject to} & ||X'(y - X\beta)||_\infty \leq \lambda. \end{array} \qquad \text{(DS)}$$

Our results for DS include large sample asymptotics.

- $\lambda/n \to 0 \Rightarrow \hat{\beta}(\text{DS}) \xrightarrow{P} \beta^*$.
- $\lambda/\sqrt{n} \to 0 \Rightarrow \hat{\beta}(\text{DS})$ asymptotically equivalent to $\hat{\beta}(\text{OLS})$.
- $\limsup_{n\to\infty} \lambda/\sqrt{n} > 0 \Rightarrow \hat{\beta}(\text{DS})$ is not asymptotically normal.
- DS is *not* in general consistent for model selection.
- Results are similar to those for LASSO (Knight and Fu, 2000)

The Adaptive Dantzig Selector

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^{n} w_j|\beta_j| \\ \text{subject to} & |X_j'(y - X\beta)| \leq w_j\lambda, \ j = 1, ..., p, \end{array} \qquad \text{(ADS)}$$

where $w_1, .., w_p$ are non-negative, data-dependent weights.

- Double-weighting scheme for ADS.
- In principle, one should choose w_j that are inversely related to the magnitude of $|\beta_j^*|$. The weights, $w_1, ..., w_p$, should
  - heavily penalize non-zero β_j when we suspect $\beta_j^* = 0$

    and
  - encourage solving the j-th scoring equation when we suspect $\beta_j^* \neq 0$.
- Reasonable weights include $w_j = |\hat{\beta}_j(\text{OLS})|^{-\gamma}$.

ADS v. Adaptive Lasso

(lasso) (alasso)

$$\min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1 \quad \leftrightarrow \quad \min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|$$

- Let $W = \text{diag}(w_1, ..., w_p)$. Alasso is equivalent to an instance of lasso where we
  - replace X with XW^{-1} and
  - replace β with $W\beta$.
- ADS is equivalent to an instance of DS where the same change of variables is made.
- DS and lasso are similar (Meinshausen et al, 2007), (James et al, 2009), (Bickel et al, 2008), etc. $\longrightarrow$ ADS and alasso are similar.

Asymptotics for ADS

- If w_j are chosen wisely and λ follows the appropriate rate, then

  (i) ADS is consistent for model selection.

  (ii) $\hat{\beta}(\text{ADS})$ is asymptotically equivalent to $(X'_{T^*}X_{T^*})^{-1}X'_{T^*}y$, the OLS estimator for β^* based on the true model.

- If $w_j = |\hat{\beta}(\text{OLS})|^{-1}$ and λ is chosen by BIC, then (i) and (ii) hold.

- ADS outperforms DS in our (moderate p) simulation study, in terms of model selection properties and prediction/squared error.

Concluding Remarks

- Large sample asymptotic results for the Dantzig selector further confirm the close relationship between the Dantzig selector and lasso.
- The adaptive Dantzig selector has advantages over the Dantzig selector.
- More complex, non-linear extensions of DS are possible.

$$\begin{aligned} \text{lasso} &\longrightarrow \text{PL-problem with penalty } p_\lambda \\ \text{Dantzig selector} &\longrightarrow p_\lambda\text{-Dantzig selector} \end{aligned}$$

- Extensions to generalized linear models and longitudinal data are in progress.