Least Angle Regression, Lasso and Boosting

Binyi Jing

2017-03-24

Outline

1. Background: Linear Regression
2. Lasso and Least Angle Regression
   - Lasso
   - LAR
   - Connection between Lasso and LAR
3. Boosting
   - The basic of Boosting (Forward Stagewise Regression)
   - More on Boosting
4. Comments on LAR, Lasso and Boosting
   - Summary
   - Comparison
5. Extension

Least Square Problem

$$min_\beta ||y - x\beta||_2^2$$

where $y \in R^n$, $x \in R^{n \times p}$ and $\beta \in R^p$.

Pros and cons

- Closed-form solution $\beta = (x^T x)^{-1} x^T y$ when $n > p$;
- Easy to overfit;
- The solution is not well-defined when $n < p$.

Traditionally, *forward stepwise variable selection* is used to overcome some difficulties.

Geometrical interpretation of Least Square

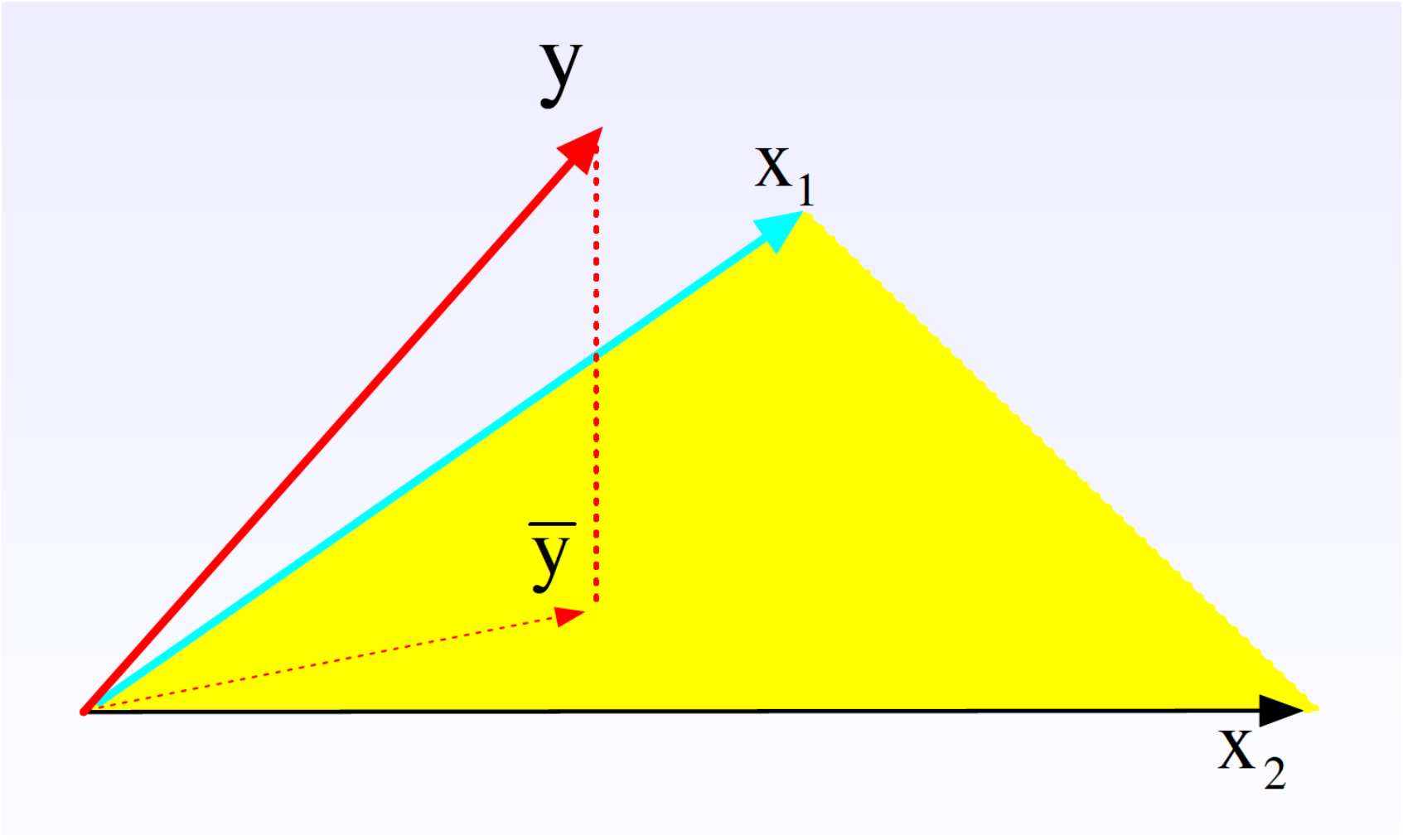

Figure 1:

Lasso vs. Ridge

Lasso problem (Tibshirani, 1996) (In signal processing, it is called Basis Pursuit. (Chen and Donoho, 1995))

$$min_{\beta}||y - x\beta||_2^2 \quad (1)$$
$$st.||\beta||_1 \leq s \quad (2)$$

Ridge problem (Hoerl, 1962)

$$min_{\beta}||y - x\beta||_2^2 \quad (3)$$
$$st.||\beta||_2 \leq s \quad (4)$$

Lasso does variable selection while Ridge does not. However, Ridge regression always has closed-form solution $\beta = (x^T x + \lambda I)^{-1} x^T y$

LARS

Prof. Bradley Efron

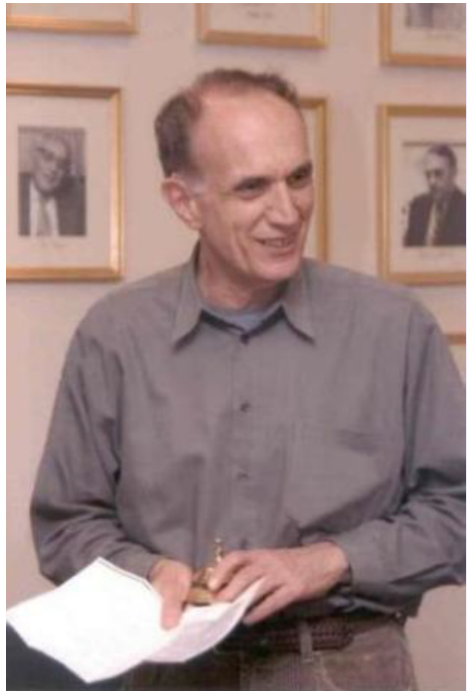

Figure: Prof. Bradley Efron: On May 29, 2007, he was awarded the National Medal of Science, the highest scientific honor by the United States, for his exceptional work in the field of Statistics (especially for his inventing of the bootstrapping methodology)

Least Angle Regression (LAR) (Efron, 2004)

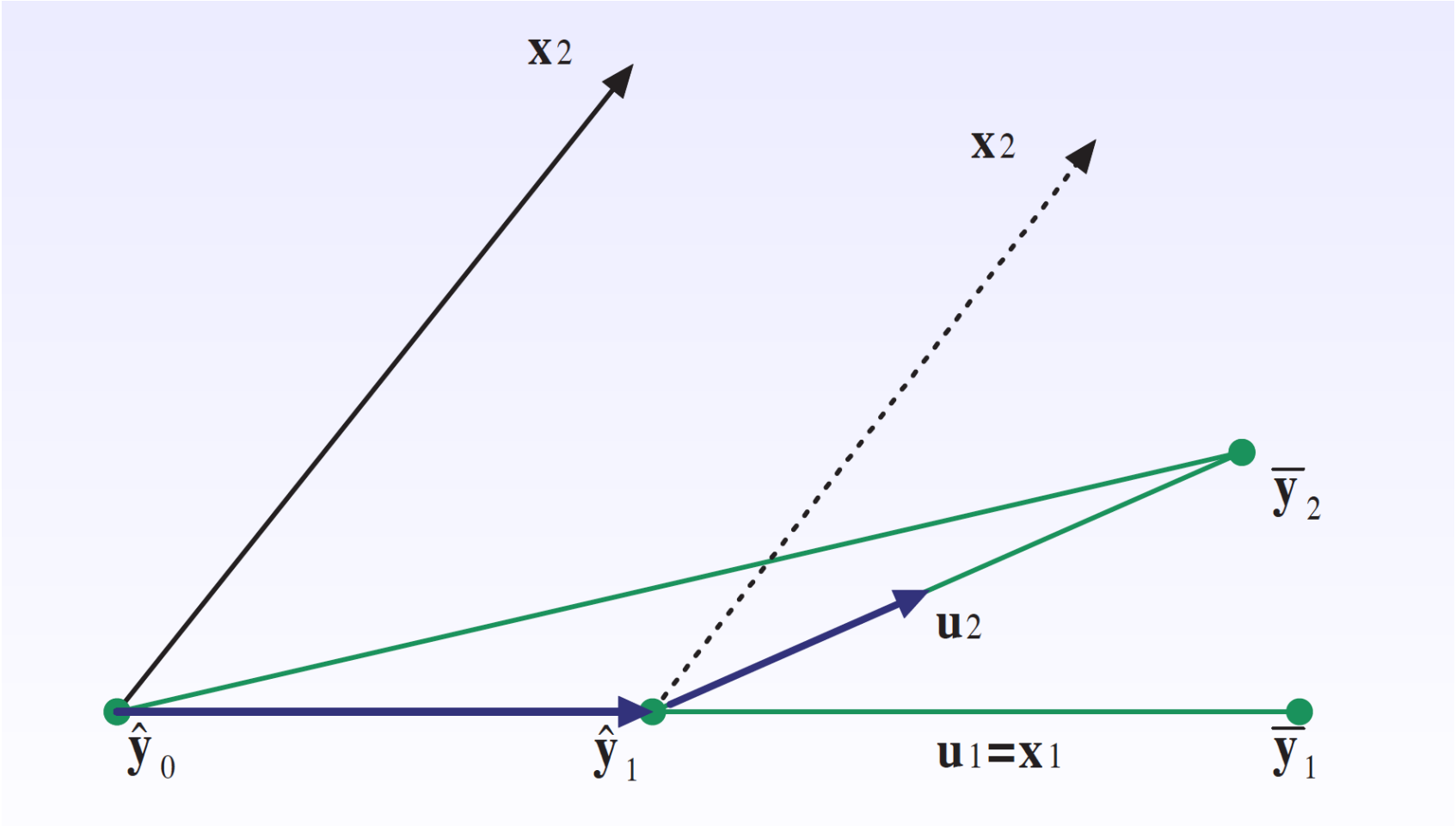

Figure 2: Geometric interpretation of LAR

Least Angle Regression (LAR)

- Start with $r = y - mean(y)$, $\beta_1, \beta_2, \ldots, \beta_p = 0$. Assume x_j standardized.
- Find predictor x_j most correlated with r (i.e.$j = argmax_{i \in \{1,\ldots,p\}} |x_i^T r|$).
- Increase β_j in the direction of $sign(corr(r, x_j))$ until some other competitor x_k has as much correlation with current residual as does x_j.
- Move (β_j, β_k) in the joint least squares direction for (x_j, x_k) until some other competitor x_l has as much correlation with the current residual.
- Continue in this way until all predictors have been entered. Stop when $corr(r, x_j) = 0, \forall j$, i.e. OLS solution.

Diabetes Data: Lasso and LAR

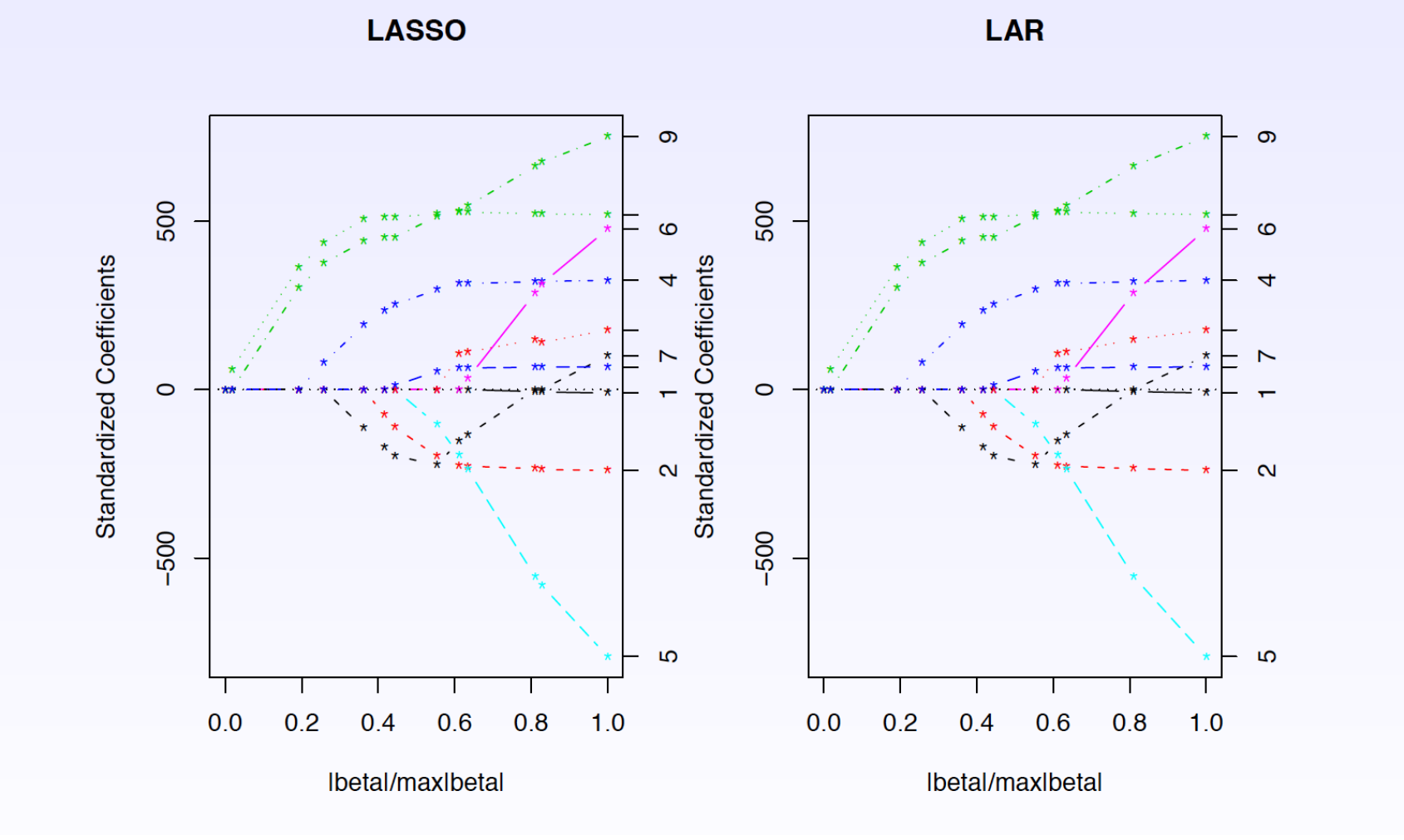

Figure 3: Diabetes Data: Lasso and LAR. They are very similar.

Why are Lasso and LAR so similar?

Consider Lasso problem

$$min_{\beta}||y - x\beta||_2^2 \tag{5}$$
$$st.||\beta||_1 \leq s \tag{6}$$

Because the l_1 norm is the non-differentiable, we do the following: Let $\beta = \beta_+ - \beta_-$, Lasso problem becomes

$$min_{\beta_0,\beta_j^+,\beta_j^-} \sum_{i=1}^{n}(y_i - \beta_0 - [\sum_{j=1}^{p} x_{ij}\beta_j^+ - \sum_{j=1}^{p} x_{ij}\beta_j^-])^2 \tag{7}$$
$$st.\beta_j^+ \geq 0, \beta_j^- \geq 0, \forall j; \sum_{j=1}^{p} \beta_j^+ + \beta_j^- \leq s \tag{8}$$

KKT conditions for Lasso

The Lagrangian is

$\sum_{i=1}^{n}(y_i - \beta_0 - [\sum_{j=1}^{p} x_{ij}\beta_j^+ - \sum_{j=1}^{p} x_{ij}\beta_j^-])^2 + \lambda \sum_{j=1}^{p}(\beta_j^+ + \beta_j^-) - \sum_{j=1}^{p} \lambda_j^+ \beta_j^+ - \sum_{j=1}^{p} \lambda_j^- \beta_j^-$

The KKT conditions are:

$$-x_j^T r + \lambda - \lambda_j^+ = 0 \quad (9)$$

$$x_j^T r + \lambda - \lambda_j^- = 0 \quad (10)$$

$$\lambda_j^+ \beta_j^+ = 0 \quad (11)$$

$$\lambda_j^- \beta_j^- = 0 \quad (12)$$

where $r = y - \beta_0 1 - [\sum_{j=1}^{p} x_j^+ \beta_j^+ - \sum_{j=1}^{p} x_j^- \beta_j^-]$

KKT conditions for Lasso

- If $\lambda = 0$, then $x_j^T r = 0, \forall j$, and the solution corresponds to the unrestricted least-square fit.

$$\begin{aligned} \beta_j^+ > 0, \lambda > 0 &\implies \lambda_j^+ = 0 && (13) \\ &\implies x_j^T r = \lambda && (14) \\ &\implies \lambda_j^- > 0 && (15) \\ &\implies \beta_j^- = 0 && (16) \end{aligned}$$

- Likewise $\beta_j^- > 0, \lambda > 0 \implies \beta_j^+ = 0$. Hence 2 and 3 give the intuitive result that only one of the pair (β_j^+, β_j^-) can be positive at any time.
- $|x_j^T r| \leq \lambda$.
- If $\beta_j^+ > 0$, then $x_j^T r = \lambda$ or if $\beta_j^- > 0$, then $-x_j^T r = \lambda$

Diabetes Data: Lasso and LAR (revisited)

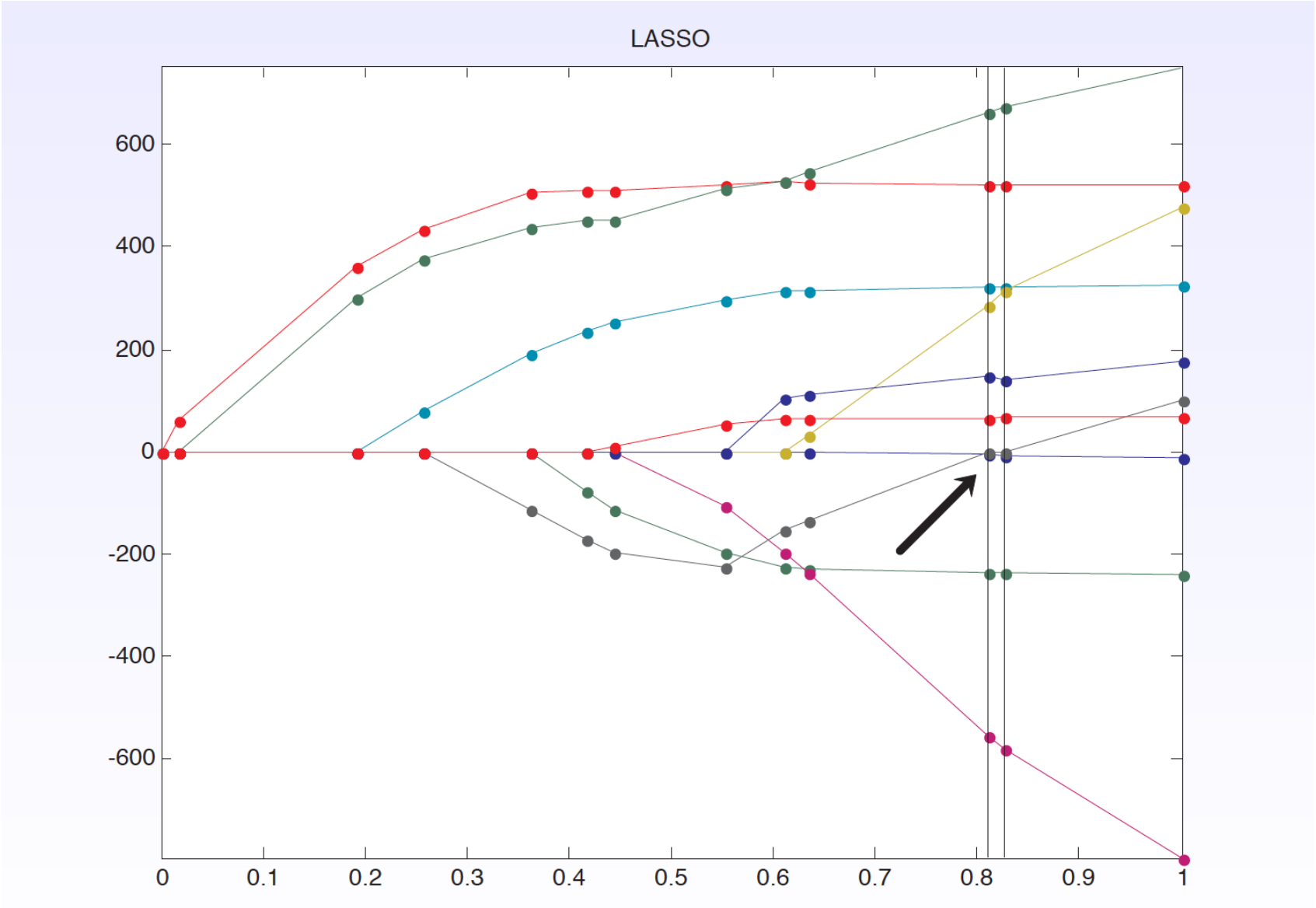

Figure 4:

The characteristic of the Lasso path

Definition

- Lasso path is given by $\beta(\lambda)$, where $\beta(\lambda)$ satisfies the KKT conditions.
- Define A be the active set, i.e. $A = j : \beta_j^+ > 0 || \beta_j^- > 0$.
- $\beta(\lambda_0)$ and $\beta(\lambda_1)$ are two points on the lasso path for the same A, and $\lambda_1 - \lambda_0 = \delta$, where δ is a small number.

We are going to show $\beta(\lambda_1) - \beta(\lambda_0)$ lies on the direction $(X_A^T X_A)^{-1} X_A^T r$, where $r = y - X_A \beta(\lambda_0)$. (According to the KKT conditions, $X_A^T r = \lambda_0 1$.)

The characteristic of the Lasso path

Define $\beta_A(\lambda)$ to be the corresponding coefficients at λ, where $\lambda \in [\lambda_0, \lambda_1]$. Deduction 5 of KKT conditions

$$\Longrightarrow X_A^T(y - X_A\beta_A(\lambda)) = \lambda 1 \tag{17}$$

$$\Longrightarrow X_A^T X_A(\beta_A(\lambda_1) - \beta_A(\lambda_0)) = \delta 1 \tag{18}$$

$$\Longleftrightarrow \beta_A(\lambda_1) - \beta_A(\lambda_0) = \delta(X_A^T X_A)^{-1} 1 \tag{19}$$

Since $r = y - X_A\beta_A(\lambda_0)$ and $X_A^T r = \lambda_0 1$

$$\beta_A(\lambda_1) - \beta_A(\lambda_0) = \frac{\delta}{\lambda_0}(X_A^T X_A)^{-1} X_A^T r \tag{20}$$

The characteristic of the Lasso path

Theorem

Let $\beta^0 \in R^{2p}$ be a point on the Lasso path in the expanded-variable space ($X = [x, -x]$), and let A be the active set of variables achieving the maximal correlation with the current residual $r = y - X\beta^0$. The Lasso coefficients move in a direction given by the coefficients of the least squares fit of X_A on r. Only the coefficients in A change, and this fixed direction is pursued until the first of the following events occurs:

- *a variable not in A attains the maximal correlation and joins A;*
- *The coefficient of a variable in the active set reaches 0, at which point it leaves A;*
- *the residuals match those of the unrestricted least square fit. when 1 or 2 occur, the direction is recomputed.*

Connection between Lasso and LAR

Lasso can be thought of as restricted versions of LAR.

1. KKT 5: If $\beta_j^+ > 0$, then $x_j^T r = \lambda$ or if $b_j^- > 0$, then $-x_j^T r = \lambda$. (Lasso has this constrain while LAR does not.)
2. LARS-uses least squares directions in the active set of variables;
3. Lasso)uses least square directions; if a variable crosses zero, it is removed from the active set.

Boosting for linear regression (I) (Friedman, 2000)

Algorithm

1. Start with $r = y - mean(y)$, $\beta_1, \beta_2, ...\beta_p = 0$;
2. Find the predictor x_j most correlated with r (i.e. $j = argmax_{i \in \{1,...,p\}} |x_i^T r|$);
3. Update $\beta_j \leftarrow \beta_j + \delta_j$, where $\delta_j = \epsilon \Delta sign(corr(r, x_j))$;
4. Set $r \leftarrow r - \delta_j x_j$ and repeat steps 2 and 3 until no predictor has any correlation with r.

Boosting for linear regression (II)

To get rid of sign, we can rewrite the algorithm in the expanded space $X = [x, -x]$.

Algorithm

1. Start with $r = y - mean(y)$, $\beta_1, \beta_2, ...\beta_{2p} = 0$;
2. Find the predictor x_j most correlated with r (i.e. $j = argmax_{i \in \{1,...,2p\}} X_i^T r$);
3. Update $\beta_j \leftarrow \beta_j + \epsilon_j$;
4. Set $r \leftarrow r - \epsilon X_j$ and repeat steps 2 and 3 until no predictor has any correlation with r.

What is Boosting doing?

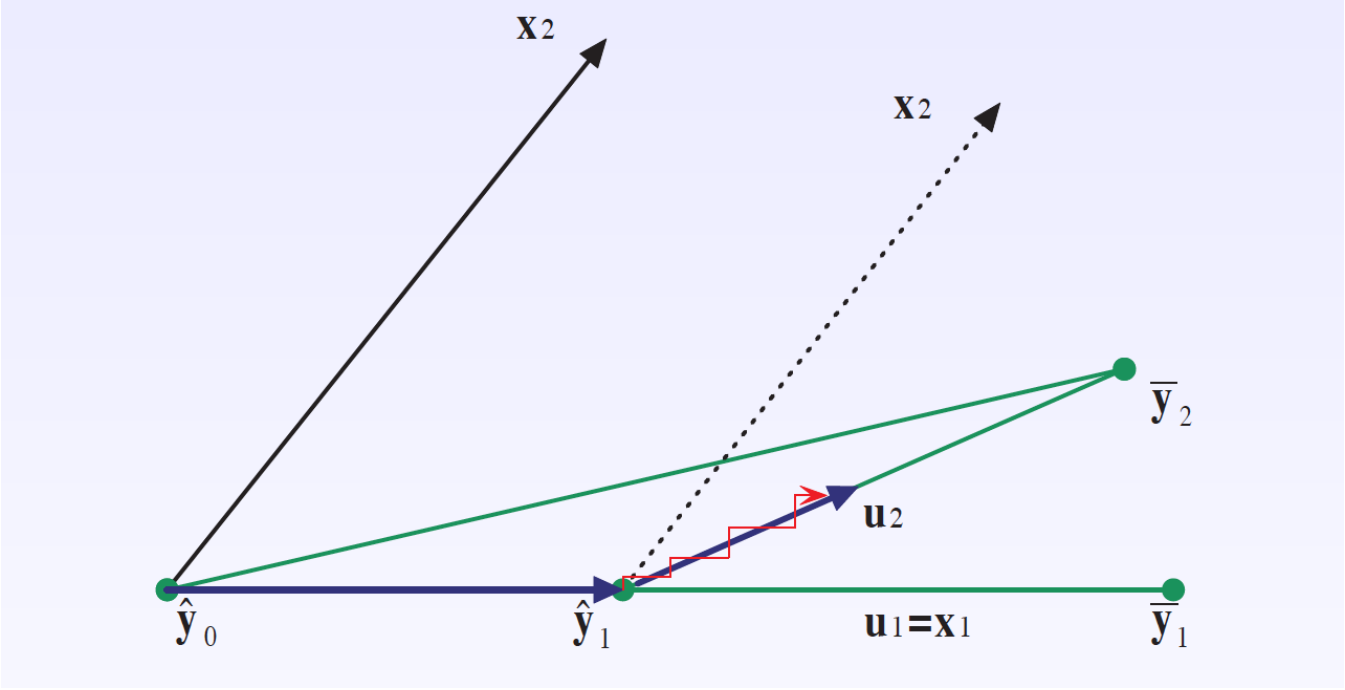

Figure 5: Figure: An illustration of Boosting based on LAR

- At each step, it selects the variable having largest correlation with the residuals, and moves its coefficient by ϵ.
- There may be a set A of variables competing for this maximal correlation.

What are the natural constrains of Boosting?

Suppose there are successive N updates. For each j, it takes N_j updates (i.e. $\sum_j N_j = N$). Define $\rho_j = N_j/N$ and thus $\sum_{j\in A} \rho_j = 1$.

1. The change of the coefficient of the variable X_j in A is $\epsilon N_j = \epsilon N \rho_j$, which must be positive $\forall j \in A$.
2. Decrease the residual sum-of-squares as fast as possible.

Consider the optimization problem

$$min_\rho \frac{1}{2}||r - \varepsilon X_A \rho_A||^2 s.t. \rho_j \geq 0; \sum_{j\in A} \rho_j = 1 \qquad (21)$$

where $\varepsilon = N\epsilon$.

KKT condition for Boosting

The Lagrangian is

$$L(\rho, \gamma, \lambda) = \frac{1}{2}||r - \varepsilon X_A \rho_A||^2 - \sum_j \gamma_j \rho_j + \lambda(\sum_j \rho_j - 1) \tag{22}$$

with KKT conditions

$$-\varepsilon X_A^T (r - \varepsilon X_A \rho_A) - \gamma + \lambda 1 = 0 \tag{23}$$
$$\lambda_j \geq 0 \tag{24}$$
$$\rho_j \geq 0 \tag{25}$$
$$\lambda_j \rho_j = 0 \tag{26}$$
$$\sum_j \rho_j = 1 \tag{27}$$

Note that $\rho_j \geq 0 \implies \gamma_j = 0$. This shows that the correlations with the residual remain equal. This also implies the relationship between LAR and Boosting.

Boosting and non-negative least square

$$min_\rho \frac{1}{2}||r - X_A\theta_A||^2 s.t. \theta_j \geq 0 \tag{28}$$

Let θ^* be the solution. Then $\rho^* = \frac{\theta^*}{||\theta^*||_1}$ solve the optimization problem of Boosting:

$$min_\rho \frac{1}{2}||r - X_A\theta_A||^2 s.t. \sum_{j \in A} \rho_j = 1 \tag{29}$$

This can be done by checking the KKT conditions.

Boosting uses non-negative least squares directions in the active set.

Boosting path

Theorem

Let $\beta^0 \in R^{2p}$ be a point on the Lasso path in the expanded-variable space ($X = [x, -x]$), and let A be the active set of variables achieving the maximal correlation with the current residual $r = y - X\beta^0$. The Boosting coefficients move in a direction given by the coefficients of the non-negative least squares fit of X_A on r. Only the coefficients in A change, and this fixed direction is pursued until the first of the following events occurs:

- *a variable not in A attains the maximal correlation and joins A;*
- *The coefficient of a variable in the active set reaches 0, at which point it leaves A; This is only for Lasso. For Boosting, the coefficients should be nondecreasing.*
- *the residuals match those of the unrestricted least square fit. when 1 occurs, the direction is recomputed.*

summary: LAR, Lasso and Boosting

LAR uses least squares directions in the active set of variables.

Lasso uses least square directions; if a variable crosses zero, it is removed from the active set.

Boosting uses non-negative least squares directions in the active set.

From another perspective,

Boosting successive differences of β_j agree in sign with the current correlation $c_j = x_j^T r$. (Step 3: $\beta_j \leftarrow \beta_j + \delta_j$, where $\delta_j = \epsilon \Delta sign(corr(r, x_j)))$

Lasso β_j agrees in sign with c_j. (KKT condition 5)

LAR no sign restrictions.

What are their performances ?

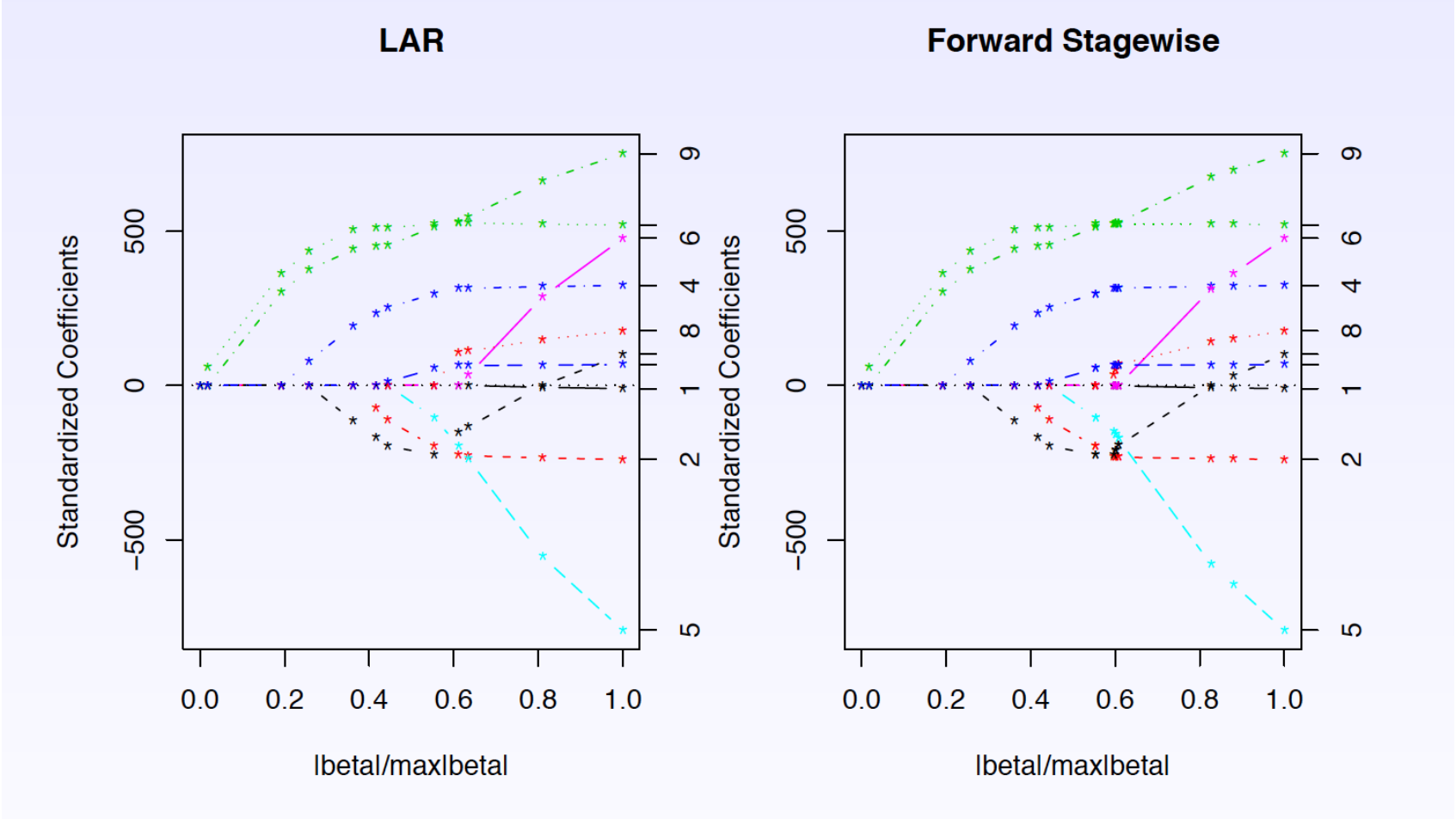

Figure 6: For low dimensional problems, their performances are almost the same (e.g., Diabetes data). How about high dimensional problems?

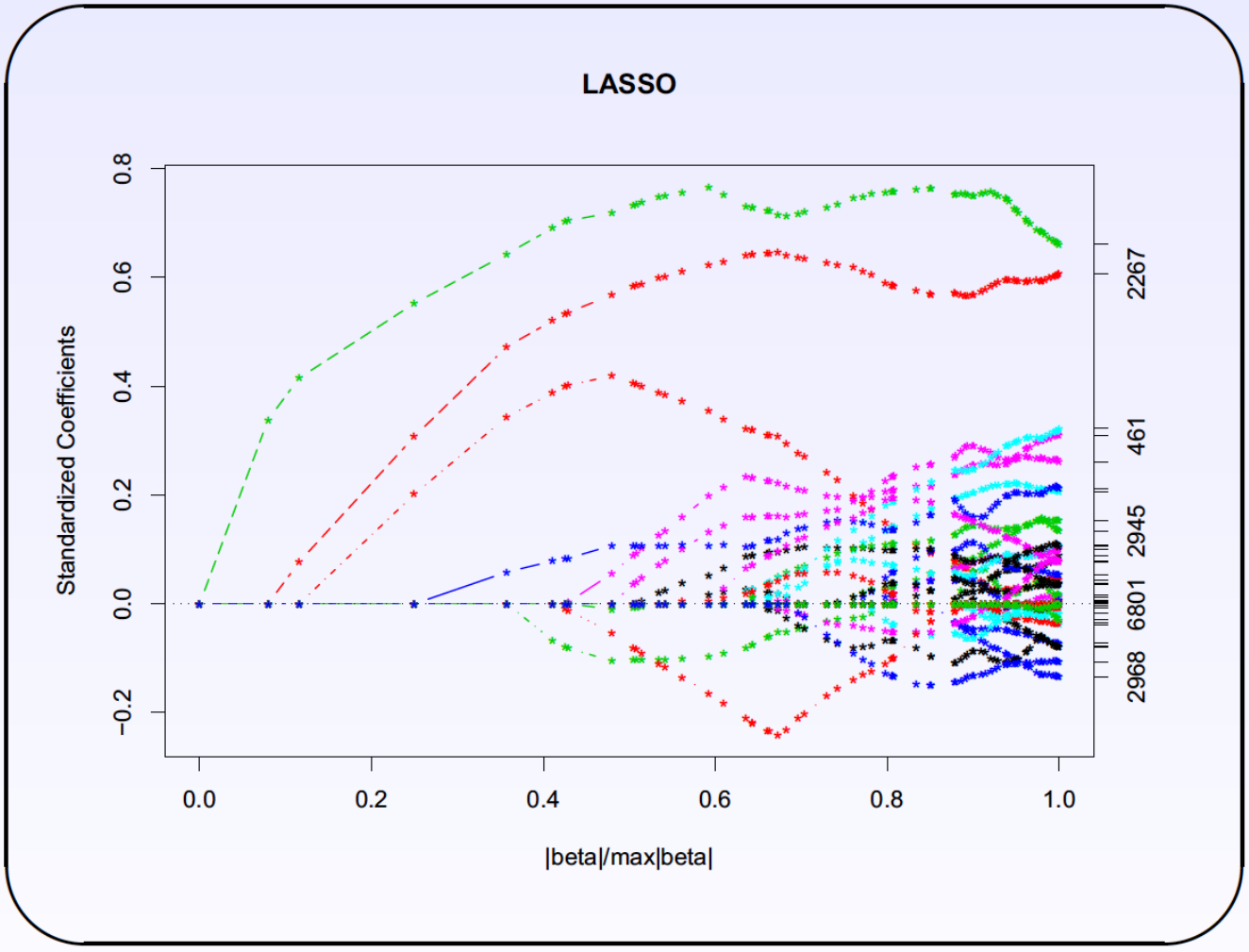

Figure 7: Lasso for Leukemia data: copy from Prof. Hastie's talk

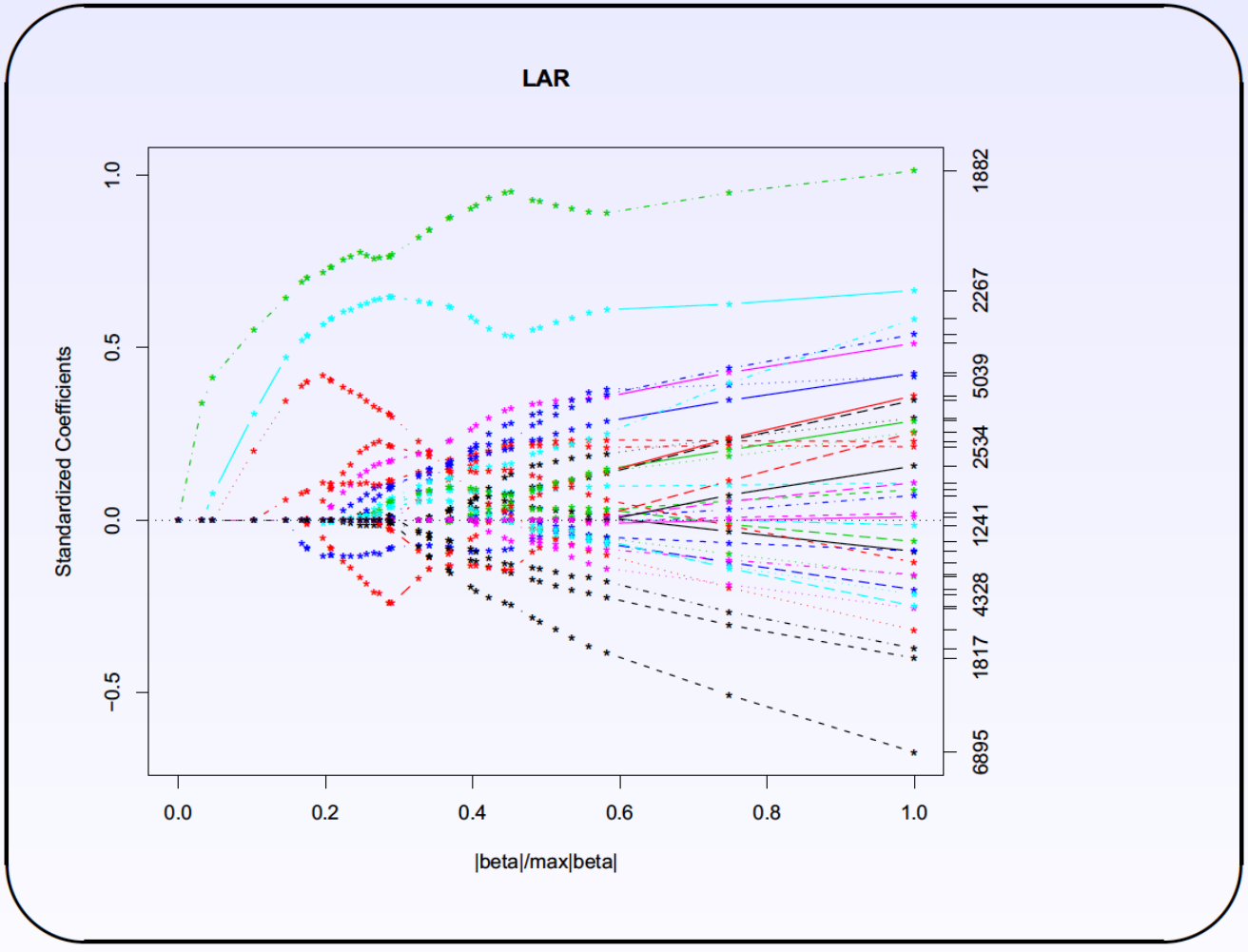

Figure 8: LAR for Leukemia data: copy from Prof. Hastie's talk

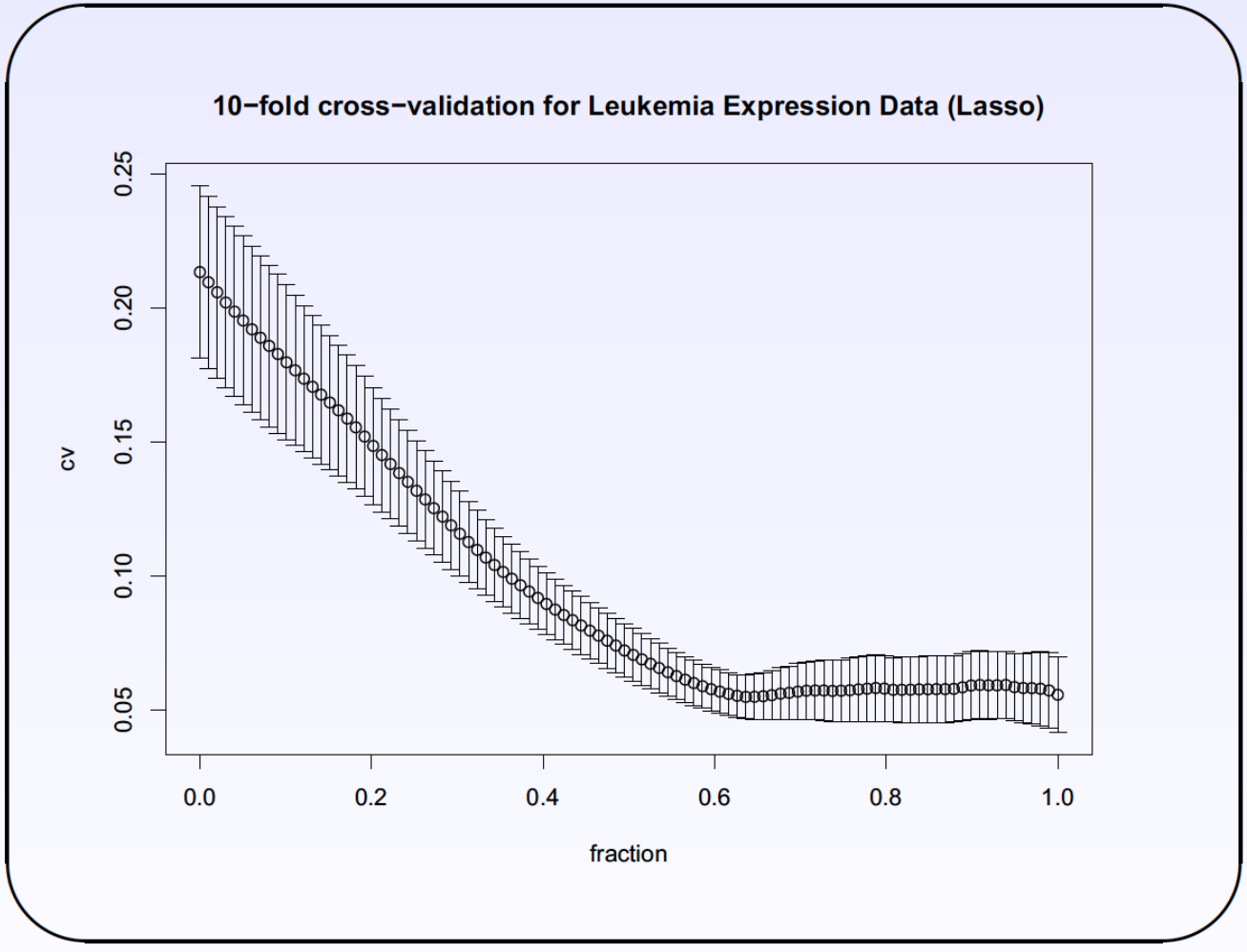

Figure 9: Lasso for Leukemia data: copy from Prof. Hastie's talk

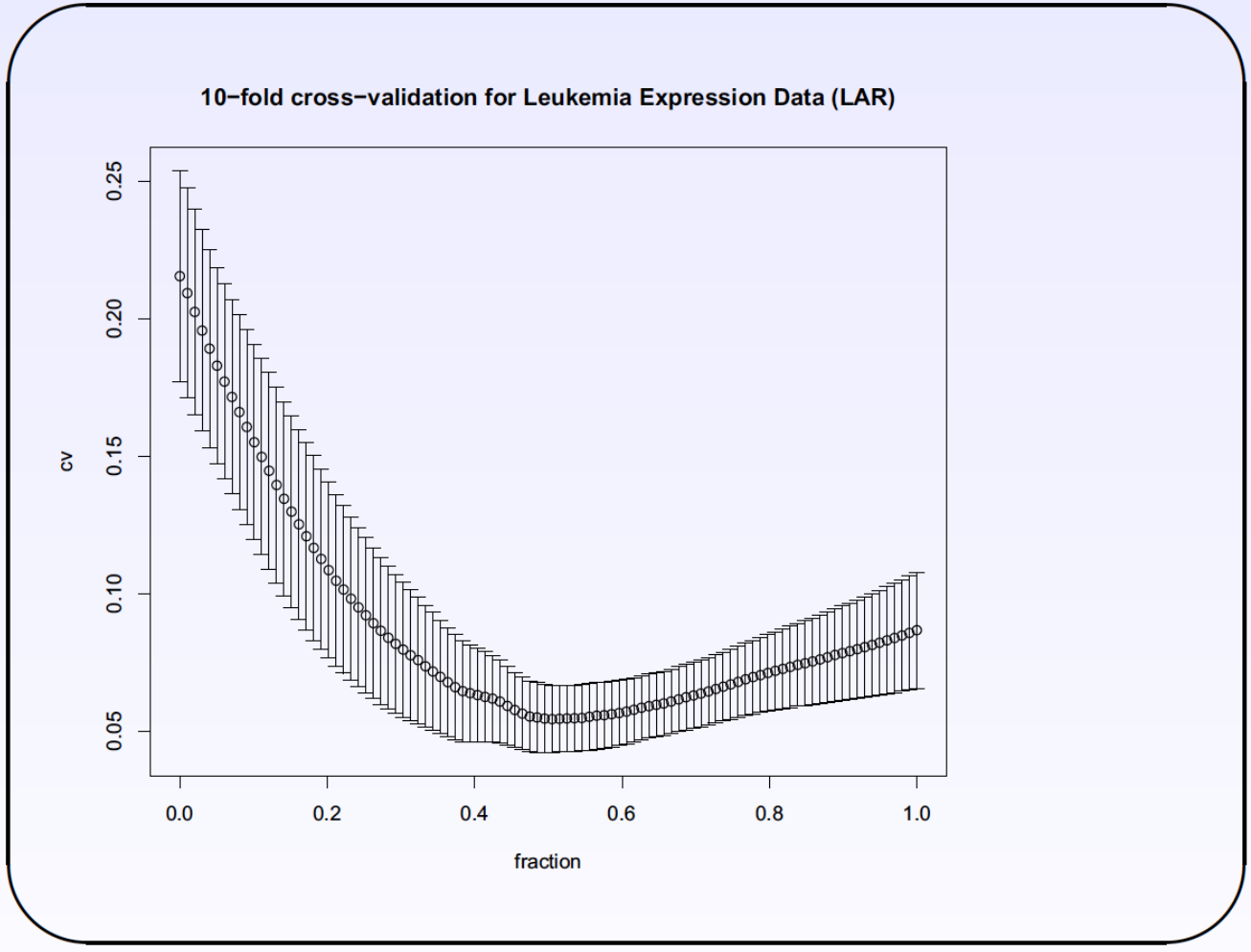

Figure 10: LAR for Leukemia data: copy from Prof. Hastie's talk

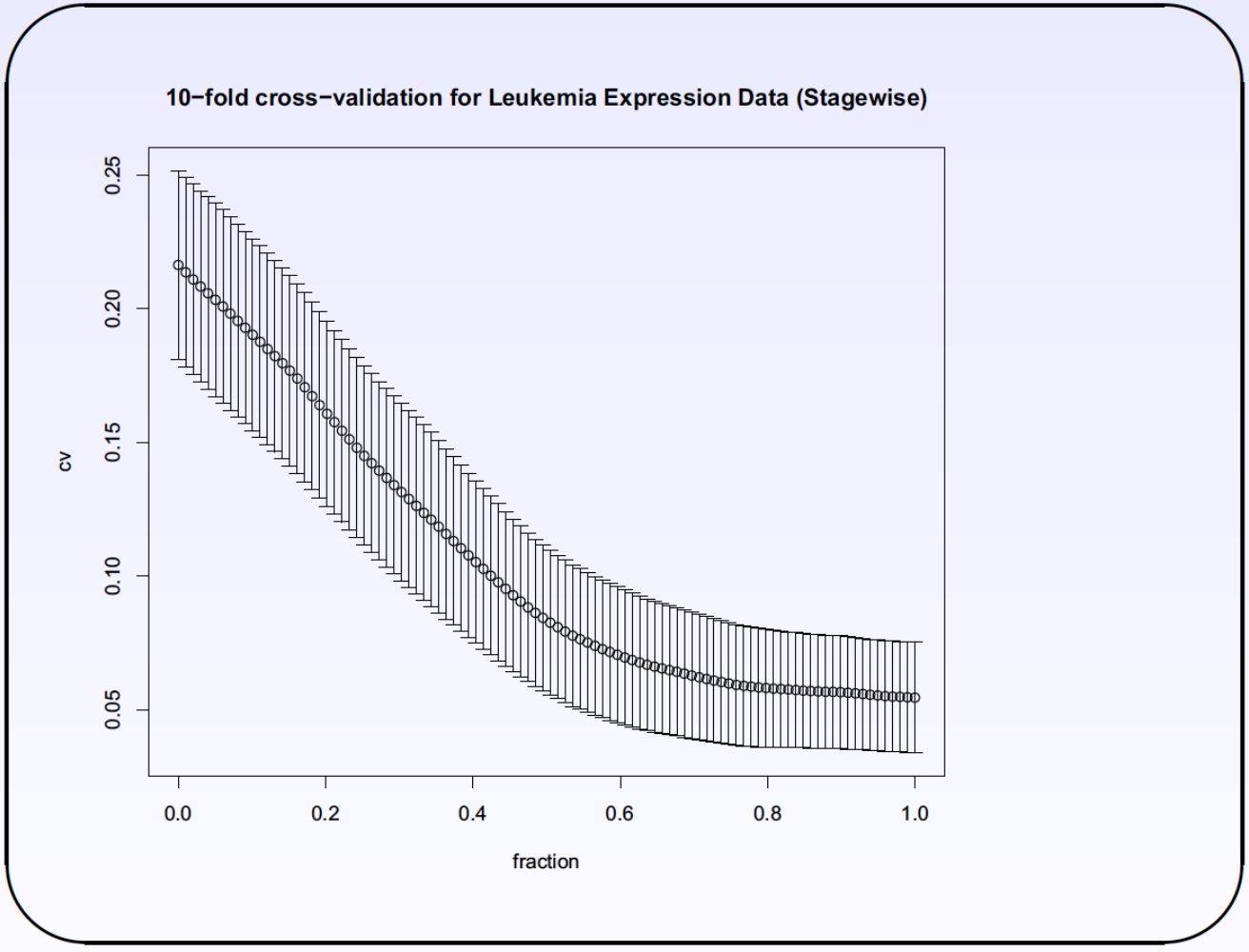

Figure 11: Boosting for Leukemia data: copy from Prof. Hastie's talk

Comments

- The speed of Overfitting: LAR $>$ Lasso $>$ Boosting
- The smoothness of the Path: Boosting $>$ Lasso and LAR
- The minimum CV error: They are almost the same.

Boosting is preferable to Lasso and LAR in the problems with large numbers of correlated predictors because of its stability.

More examples which support this comment can be found in the paper:

T. Hastie, Forward stagewise regression and the monotone lasso. Electronic Journal of Statistics, 2007.

Boosting with CART

Instead of linear predictors, CART can be used in Boosting.

1. Start with function $F(x) = 0$ and residual $r = y$
2. Fit a CART regression tree to r giving $f(x)$
3. Set $F(x) \leftarrow F(x) + \epsilon f(x), r \leftarrow r - \epsilon f(x)$ and repeat step 2 many times

Loss function

- Boosting is easily extended to some other convex loss functions $L(y, F(x))$, e.g., Logistical loss function. The residual r will be taken place by the negative gradient of the loss function. This is so-called "gradient boosting".
  - Least square loss:
    $L = \frac{1}{2}\sum_i (y_i - f(x_i))^2 \rightarrow -\frac{\partial L}{\partial f(x_i)} = y_i - f(x_i)$
  - Logistical loss: $L = \frac{1}{2}\sum_i (y_i f(x_i) - log(1 + f(x_i)))^2 \rightarrow$
    $-\frac{\partial L}{\partial f(x_i)} = y_i - \frac{1}{1 + exp(f(x_i))}$
  - Exponential loss, Poisson loss, ...
- Lasso can be extended to some other convex loss functions as well, and solved by convex optimization.

Incorporate disease models into Boosting

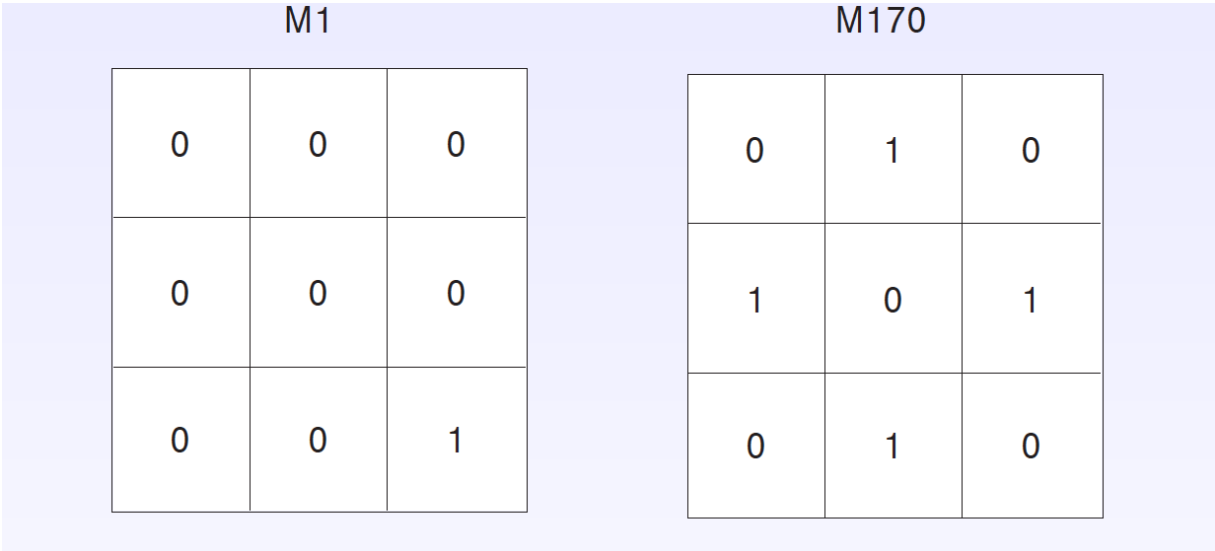

Figure 12: Figure: epistasis models

Take the place of CART by epistasis models.

1. Start with function $F(x) = 0$ and residual $r = y$
2. Fit a disease model to r giving $f(x)$
3. Set $F(x) \leftarrow F(x) + \epsilon f(x), r \leftarrow r - \epsilon f(x)$ and repeat step 2 many times

The advantages of incorporate disease models into boosting for the SNP problem

1. We will not worry about the marginal effects (i.e., main effects) as in MegaSNPHunter. Of course we pay more computation efforts to choose a disease model.
2. We can answer the question formally: Whether these disease models are realistic?
3. We can handle heterogeneity in the data naturally since we employ additive model in Boosting process. (Additive models can approximate heterogeneity model well.)
4. It's computationally feasible for one chromosome (about 10,000 SNPs). There is no need for worrying about the memory.
5. It overfits very slowly because of the nature of Boosting.

A simple experiments

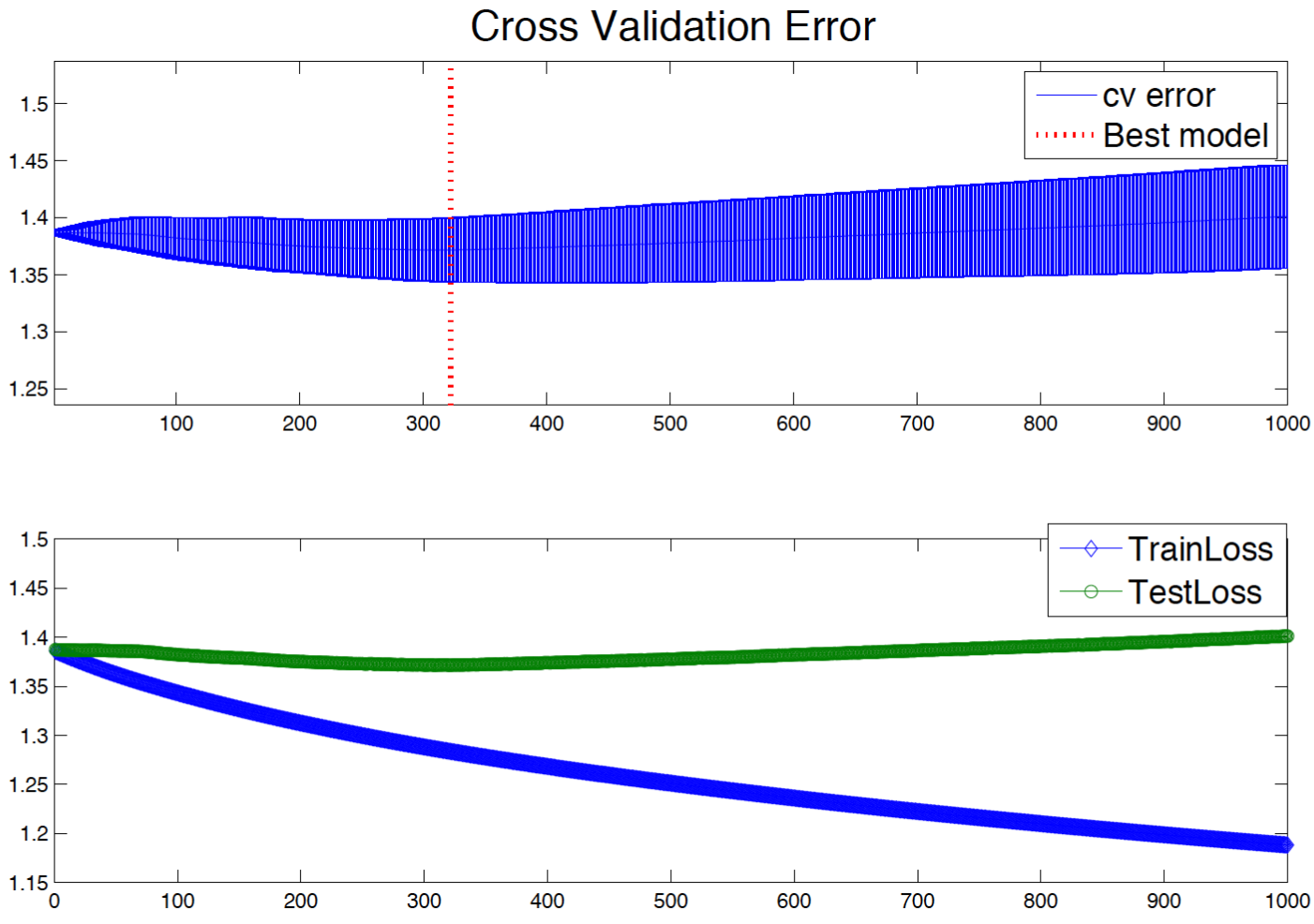

Figure 13:

References on the epistasis models:

- Wen T. Li etal, A Complete Enumeration and Classification of Two-Locus DiseaseModels. Human Heredity, 2000.
- David M. Evans etal, Two-Stage Two-Locus Models in Genome-Wide Association. PLOS Genetics. 2006, Sep.

Strongly recommended references on Lasso, LAR and Boosting:

- R. Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Series B-Methodological, vol.58 no.1, pp.267-288, 1996.
- J. Friedman. Additive Logistic Regression: A Statistical View of Boosting. Annals of Statistics (With Discussion), vol. 28, no. 2, pp.337-407, 2000.
- J. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.
- B. Efron. Least Angle Regression. Annals of Statistics vol.32, no.2, pp. 407ı499, 2004.
- T. Hastie. The Elements of statistical learning, (2nd), 2009.