LARS notes (Part1) (To page9,2 end)

Abstract:

Useful and less greedy version of traditional forward selection methods.

Main property:

- Implement the Lasso, lasso Modification: Calculates all possible Lasso estimates for a given problem.
- Different version: Another modification efficiently implements forward stagewise linear regression.
- A simple approximation for the degree of freedom of a LARS estimate is available, from which we derive a C_p estimate of prediction error. thisi allws a principled choice among the range of possible LARS estimates.

  (Not quite understand the final part of the LARS goals.)

LARS relates: classic model-selection method known as "forward selection" or "forward stepwise regression."

- Forward Selection
  - Given a collection of possible predictors, select the one largest absolute correlation with the response y, say x_{j_1}, and perform simple linear regression of y on x_{j_1} , than leaves a residual vector which is orthogonal to x_{j_1}. Project the other predictors orthogonally to x_{j_1} nd repeat the selection process. After k steps this results in a set of predictors $x_{j_1}, \ldots, x_{j_k}$ that are then used in the usual way to construct a k-parameter linear model.
- Forward stagewise
- More cautious version of forwad selection-> take thousands tiny steps as it moves toward a final model.
- Original motivation for the LARS algorithm.
- LARS-Lasso-Stagewise connection is comceptually as well as computationally useful.

Model construction:

Predict response y from covariates $x_1, \cdots, x_n$.

By location and scale transformations we always assume that the covariates have been standardized to have mean 0 and unit length, and that the response has mean 0.

$$\sum_{i=1}^{n} y_i = 0 \quad \sum_{i=1}^{n} x_{ij} = 0 \quad \sum_{i=1}^{n} x_{ij}^2 = 1 \quad for\ j = 1, 2, \ldots, m \qquad (1)$$

Regression coefficients : $\widehat{\boldsymbol{\beta}}=\left(\widehat{\beta}_1,\widehat{\beta}_2,\ldots,\widehat{\beta}_m\right)'$ gives prediction vector $\hat{\mu}$

$$\hat{\boldsymbol{\mu}}=\sum_{j=1}^{m}\mathbf{x}_j\hat{\beta}_j=X\hat{\boldsymbol{\beta}}\quad[X_{n\times m}=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_m)]\tag{2}$$

Total squared error

$$S(\widehat{\boldsymbol{\beta}})=\|\mathbf{y}-\widehat{\boldsymbol{\mu}}\|^2=\sum_{i=1}^{n}(y_i-\widehat{\mu}_i)^2\tag{3}$$

L_1 norm for lasso

$$T(\widehat{\boldsymbol{\beta}})=\sum_{j=1}^{m}\left|\widehat{\beta}_j\right|\tag{4}$$

$$\textit{Lasso: minimize}\quad S(\hat{\beta})\textit{subject to}\quad T(\widehat{\beta})\le t\tag{5}$$

`Quadratic programming techniques can be used to solve (5).though we will present an easier method here, closely related to the “homotopy method” of Osborne, Presnell and Turlach (2000a)."`

Forward Stagewise Linear Regression.

- Begins with $\hat{\mu}=0$,builds up the regression function in successive small steps.
- Let $\hat{\mu}$ is the current Stagewise estimate, let $\mathbf{c}(\widehat{\boldsymbol{\mu}})$ be the vector of *current correlations*
  $\hat{\mathbf{c}}=\mathbf{c}(\widehat{\boldsymbol{\mu}})=X'(\mathbf{y}-\widehat{\boldsymbol{\mu}})$
- $\hat{c}_j$ is proportional to the correlation between covariate x_j and current residual vector. Next step is taken in the direction of the greatest current correlation,

$$\hat{j}=\operatorname{argmax}|\hat{c}_j|\quad\text{and}\quad\widehat{\boldsymbol{\mu}}\rightarrow\widehat{\boldsymbol{\mu}}+\epsilon\cdot\operatorname{sign}\left(\hat{c}_{\hat{j}}\right)\cdot\mathbf{x}_{\hat{j}}\tag{6}$$

- Need to mentioned here: ϵ is a "small" constant, "small" is important, otherwise "big" choice like $\epsilon=|\hat{c}_j|$ leads to the standard forward selection technique. this could be over greedy.

The main point:

LARS is a stylized version of the stagewise procedure that uses a simple mathematical formula to accelerate the computations.

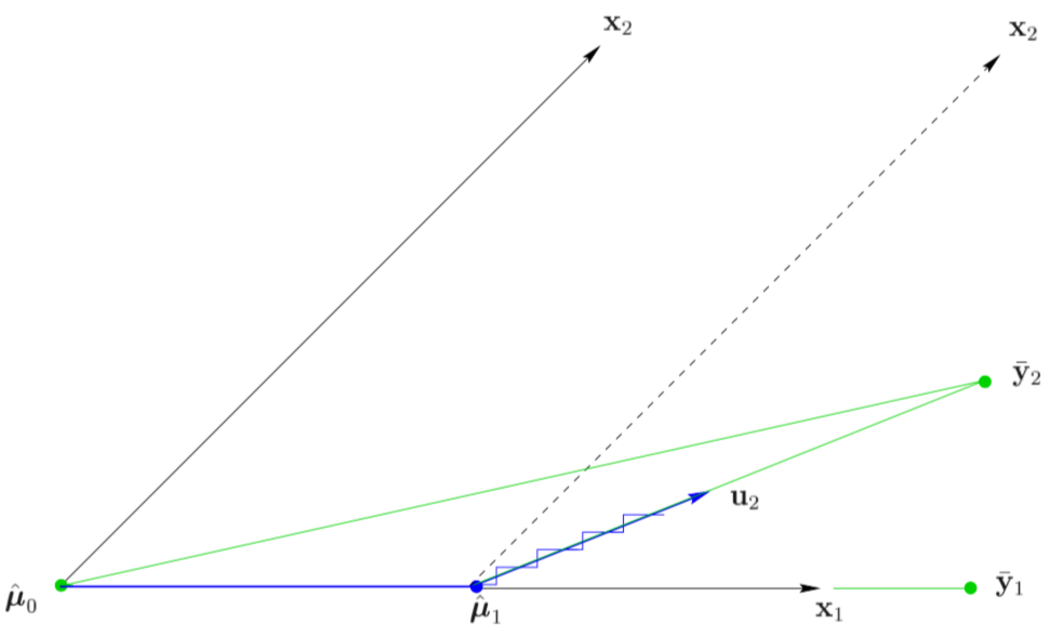

Figure 2. The LARS algorithm in the case of $m = 2$ covariates; $\bar{\mathbf{y}}_2$ is the projection of $\mathbf{y}$ into $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2)$. Beginning at $\widehat{\boldsymbol{\mu}}_o = \mathbf{0}$, the residual vector $\bar{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}_o$ has greater correlation with $\mathbf{x}_1$ than $\mathbf{x}_2$; the next LARS estimate is $\widehat{\boldsymbol{\mu}}_1 = \widehat{\boldsymbol{\mu}}_o + \widehat{\gamma}_1 \mathbf{x}_1$, where $\widehat{\gamma}_1$ is chosen such that $\bar{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}_1$ bisects the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$; then $\widehat{\boldsymbol{\mu}}_2 = \widehat{\boldsymbol{\mu}}_1 + \widehat{\gamma}_2 \mathbf{u}_2$ when $\mathbf{u}_2$ is the unit bisector; $\widehat{\boldsymbol{\mu}}_2 = \bar{\mathbf{y}}_2$ in the case $m = 2$. The staircase indicates a typical Stagewise path. Here LARS gives the Stagewise track as $\epsilon \to 0$, but a modification is necessary to guarantee agreement in higher dimensions, see Section 3.2.

以下是重复论文的无聊描述，参考价值有限，但是因为内容已经高度概括化所以不方便删除，以下是用中文复述一遍主要思想。下午浪费半天一个原因就是因为没注意仔细$\overline{y_2}$ 是y在$\mathcal{L}(X)$上的投影，也就是说，如果我们只考虑(x_1, x_2) ，那么$\overline{y}_2$ 就是目标的response.（而且在第一遍看的时候当初理解了这个事实！但是忘了orz）。

$\overline{y}_2$是目标response, 此时， $\overline{y}_2$和x_1 的correlation比 x_2 大，于是$\hat{\mu}_1$ 朝 x_1 更新，然后呢，此时的residual $\overline{y}_2 - \hat{\mu}_1$ 继续对x_1和x_2 找correlation. 慢慢找，直到如上图$\hat{\mu}_1$那个点,这时，correlation相等了，那么就停在这里，进行方向切换。而新的方向则是x_1和x_2 的角平分线方向。而LARS的计算过程可以通过一步直接跳到这个需要换方向的点，可以极大的降低计算消耗。（是否能跳过去这里存疑）

Each step adding one covariate to the model, after k steps just k of the $\hat{\beta}_j$'s are non-zero. The figure2 showed the m=2 covariates, $X = (x_1, x_2)$. In this case the current correlation $\hat{\mathbf{c}} = \mathbf{c}(\widehat{\boldsymbol{\mu}}) = X'(\mathbf{y} - \widehat{\boldsymbol{\mu}})$ depend only on the projection $\overline{y}_2$ of y into the linear space $\mathcal{L}(X)$ spanned by x_1 and x_2 .

- The algorithm begins at $\hat{\mu}_0 = 0$, Figure 2 has $\overline{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}_o$ making a smaller angle with $x_1$than x_2 . In this case, the new correlation can be expressed as

$$\mathbf{c}(\widehat{\boldsymbol{\mu}}) = X'(\mathbf{y} - \widehat{\boldsymbol{\mu}}) = X'\left(\overline{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}\right)$$

Then, choose x_1 as the direction.

$$\widehat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_o + \widehat{\gamma}_1 \mathbf{x}_1 \tag{7}$$

- Stagewise would choose $\hat{\gamma}_1$ equal to some value ϵ , than repeat many times, or make $\hat{\gamma}_1$ larger enough to make $\hat{\mu}_1$ equal $\overline{y}_1$, the projection of y into $\mathcal{L}(\mathbf{x}_1)$.

  LARS uses an intermediate value of $\hat{\gamma}_1$, the value that makes $\overline{\mathbf{y}}_2 - \widehat{\boldsymbol{\mu}}$, equally correlated with x_1 and x_2 ; that is, $\overline{y}_2 - \hat{\mu}_1$ bisects the angle between x_1 and x_2 , so $c_1(\hat{\mu}_1) = c_2(\hat{\mu}_1)$.

- u_2 be the unit vector lying along the bisector. The next LARS estimate is

$$\widehat{\boldsymbol{\mu}}_2 = \widehat{\boldsymbol{\mu}}_1 + \widehat{\gamma}_2 \mathbf{u}_2 \tag{8}$$

  With $\hat{\gamma}_2$ chosen to make $\hat{\mu}_2 = \overline{y}_2$ in the case m=2. With $m > 2$ covariates , $\hat{\gamma}_2$ would be smaller, leading to another change of direction.

- LARS is motivated by the fact that it is easy to calculate the step size $\hat{\gamma}_1, \hat{\gamma}_2, \ldots$, short-circuiting the small Stagewise steps.

We assume that the covariate vector $x_1, x_2, \ldots, x_m$ are linearly independent.

For $\mathcal{A}$ a subset of the indices $\{1, 2, \ldots, m\}$, define the matix

$$X_{\mathcal{A}} = (\cdots s_j \mathrm{x}_j \cdots)_{j \in \mathcal{A}} \tag{9}$$

when the signs s_j equal ± 1 . Let

$$\mathcal{G}_{\mathcal{A}} = X_{\mathcal{A}}' X_{\mathcal{A}} \quad \text{and} \quad A_{\mathcal{A}} = \left(1_{\mathcal{A}}' \mathcal{G}_{\mathcal{A}}^{-1} 1_{\mathcal{A}}\right)^{-\frac{1}{2}} \tag{10}$$

$1_{\mathcal{A}}$ being a vector of 1's of length equaling $|\mathcal{A}|$ the size of $\mathcal{A}$. The

$$\text{equiangular vector: } u_{\mathcal{A}} = X_{\mathcal{A}} w_{\mathcal{A}} \quad \text{where} \quad w_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} 1_{\mathcal{A}} \tag{11}$$

is the unit vector making equal angles, less than 90° , with the column of $X_{\mathcal{A}}$,

$$X_{\mathcal{A}}' \mathbf{u}_{\mathcal{A}} = A_{\mathcal{A}} 1_{\mathcal{A}} \quad \text{and} \quad \|\mathbf{u}_{\mathcal{A}}\|^2 = 1 \tag{12}$$

We saw the previous part in a negative direction. First, we look the final part (12), which should be satisfied as the equal angular .

We need find a vector $u_{\mathcal{A}}$ satisfied $X'u = a\mathbf{1}$, now note that $1 = X'X(X'X)^{-1}1$. That is , we have X' in the left hand side. so $u_{\mathcal{A}}$ have a candidate $X(X'X)^{-1}\mathbf{1}$, then what we need to do is just standardize it.

$$
\begin{aligned}
u &= \frac{X(X'X)^{-1}\mathbf{1}}{\sqrt{\mathbf{1}'(X'X)^{-1}X'X(X'X)^{-1}\mathbf{1}}} && (13)\\
&= \frac{X(X'X)^{-1}\mathbf{1}}{\sqrt{\mathbf{1}'(X'X)^{-1}\mathbf{1}}} && (14)\\
&= \frac{XG^{-1}\mathbf{1}}{\sqrt{\mathbf{1}'G^{-1}\mathbf{1}}} && (15)\\
&= XG^{-1}\mathbf{1}\times A && (16)\\
&= XA^{*}G^{-1}\mathbf{1} && (17)\\
&= Xw && (18)
\end{aligned}
$$

This is how the equal-angular vector is constructed.

So the problem becomes to a nother one, why the form of vector making equal angles?

$$X'_{\mathcal{A}}\mathbf{u}_{\mathcal{A}} = A_{\mathcal{A}}1_{\mathcal{A}} \tag{19}$$

That is, what is this formula mean?

Equal angular is equals to which angular????

That is, the $cos(\theta)$ between any subset $x_{i_{\mathcal{A}}}$ and $cos < u, x_{i_{\mathcal{A}}} >$ = $\frac{x'_{i_{\mathcal{A}}}u_{\mathcal{A}}}{||x'_{i_{\mathcal{A}}}u_{\mathcal{A}}||}$. In this case,

$X'_{\mathcal{A}}\mathbf{u}_{\mathcal{A}}$ is the vector about $x'_{i_{\mathcal{A}}}u_{\mathcal{A}}$, then if the angular is equal, cos also should equal. However, $X'_{\mathcal{A}}$projection to the direction in $u_{\mathcal{A}}$, then right hand side is the length of projection times 1. ($A_{\mathcal{A}}$ is a value rather a matrix.)

The describtion upon is so confusing!!!! need more clear idea about it!!!!

Then is the "fullly" describe about the LARS.

- Start with $\hat{\mu}_0 = 0$ and build up $\hat{\mu}$ by steps.

  Suppose current estimate is $\hat{\mu}_{\mathcal{A}}$, that $\hat{c} = X'(y - \hat{\mu}_{\mathcal{A}})$, is the current correlations. $\mathcal{A}$ is active set which indices corresponding covariates with the greatest absolute current correlations

  $$\widehat{C} = \max_i\{|\hat{c}_j|\} \quad \text{and} \quad \mathcal{A} = \left\{j : |\hat{c}_j| = \widehat{C}\right\} \tag{20}$$

  Letting

  $$s_j = \text{sign}\{\hat{c}_j\} \quad \text{for} \quad j \in \mathcal{A} \tag{21}$$

  we compute $X_{\mathcal{A}}, A_{\mathcal{A}}$ and $u_{\mathcal{A}}$ we showed previous, the equal angle trisector and the inner product vector.

  $$\mathbf{a} \equiv X'\mathbf{u}_{\mathcal{A}} \tag{22}$$

Then the next step of the LARS algorithm updates $\hat{\mu}_{\mathcal{A}}$, say to

$$\widehat{\boldsymbol{\mu}}_{\mathcal{A}_+} = \widehat{\boldsymbol{\mu}}_{\mathcal{A}} + \widehat{\gamma}\mathbf{u}_{\mathcal{A}} \tag{23}$$

where

$$\widehat{\gamma} = \min_{j \in \mathcal{A}^c}{}^{+} \left\{ \frac{\widehat{C} - \hat{c}_j}{A_{\mathcal{A}} - a_j}, \frac{\widehat{C} + \hat{c}_j}{A_{\mathcal{A}} + a_j} \right\} \tag{24}$$

- 还是老样子从$\hat{\mu}_0 = 0$ 开始，所以还是两件事，计算当前的correlation $\hat{c} = X'(y - \hat{\mu}_{\mathcal{A}})$. 其中 $\mathcal{A}$ 是一个indices set包括了每一步最大相关系数。这里有点奇怪。特别是$\hat{C}$ 的定义形式。

Emmmm，是不是可以这么解释，因为开始转向的时候，$c_1 = c_2$所以此时往u_2的方向走并不会影响c_1和c_2 但是会影响c_3。 所以等走走走,c_3 又会减小减小。好像不太对，residual的变化。。。。如果按figure2的话，residual对x_1 的角度一直在增大，角度增大导致correlation会减小，因为x_1 的γ 在增大，x_1 方向的解释越来越多，而同时，对x_2 的角度一直在减小,也就是cos在增大，correlation在增大，一减一增两个过程直到这两个correlation相等

然后回到这个过程，也就是在找了一段时间之后，有记录: $\mathcal{A}$ 。注意correlation的定义式：$\hat{c} = X'(y - \hat{\mu}_{\mathcal{A}})$ X 是$n \times p$，$p \times n \times n \times 1 = p \times 1$，$\hat{c}$是此时的模型的已建模部分 $\hat{\mu}_{\mathcal{A}}$ 的residual关于 X的 correlation。

按角平分线的思路,这时候在第i 维的变化是主要的，$i-1$维上应该都是角平分线所以是一致的?

但是$\mathcal{A}$ 这样定义感觉很奇怪啊。 $\mathcal{A}$里面只包含了correlation是最大的那几个，也就是说一直走一直走，走到有新的correlation能加进来，那再改变方向，重新计算$u_{\mathcal{A}}$，否则就一直按这个方向走。比如说回到figure2的图例，在一开始，最大的correlation只有x_1这个方向，走一小步，x_2 方向的correlation还是小于x_1 ，所以$\mathcal{A}$还是只有{1}。直到residual 的correlation到x_1和x_2 一致，那么就有了两个同时达到max的correlation，那就得重新计算$u_{\mathcal{A}}$.

那这么理解就没问题了，于是下一步是通过这些东西计算步长.

- Back to paper understanding mode

Define:

$$\boldsymbol{\mu}(\gamma) = \widehat{\boldsymbol{\mu}}_{\mathcal{A}} + \gamma\mathbf{u}_{\mathcal{A}} \tag{25}$$

for $\gamma > 0$, so that the current correlation

$$c_j(\gamma) = x_j'(y - \mu(\gamma)) = \hat{c}_j - \gamma\alpha_j. \tag{26}$$

for $j \in \mathcal{A}$, around the equation (20), the definition of correlation, max correlation, direction , yield

$$|c_j(\gamma)| = \widehat{C} - \gamma A_{\mathcal{A}} \tag{27}$$

也就是当前的j的correlation，是最大correlation减去步长乘以和投影向量的长度有关的一个玩意($A_{\mathcal{A}}$)。注意，这里$j \in \mathcal{A}$ 。也就和之前描述的，因为走的是角平分线，所以这帮已经active的variable同进退。

然后下一步该考虑的就是不在$\mathcal{A}$里的j。

For $j \in \mathcal{A}^c$, the two formula upon shows that $c_j(\gamma)$ equals the maximal value at $\gamma = (\hat{C} - \hat{c}_j)/(A_{\mathcal{A}} - a_j)$. Likewise $-c_j(\gamma)$, the current correlation for the reversed covariate $-x_j$, achieves maximality at $(\hat{C} + \hat{c}_j)/(A_{\mathcal{A}} + a_j)$.

Therefore γ in (24), is the *smallest positive value of γ such that some new index $\hat{j}$ joins the active set;* $\hat{j}$ is the smallest positive value of γ such that some new index $\hat{j}$ joins the active set; $\hat{j}$ is the minimizing index in (24) , the foot length of every j in $\mathcal{A}^c$. the new active set is $\mathcal{A} \cup \{\hat{j}\}$, the new maximum absolute correlation is $\widehat{C}_+ = \widehat{C} - \hat{\gamma} A_A$.

The figure 10 shows the LARS in diabetes data. 10 iterations for procedure from start to end. The join order or LARS is same as Lasso. However, tracks of $\hat{\beta}_j$ are nearly but not exactly as either the LASSO or Stagewise tracks.

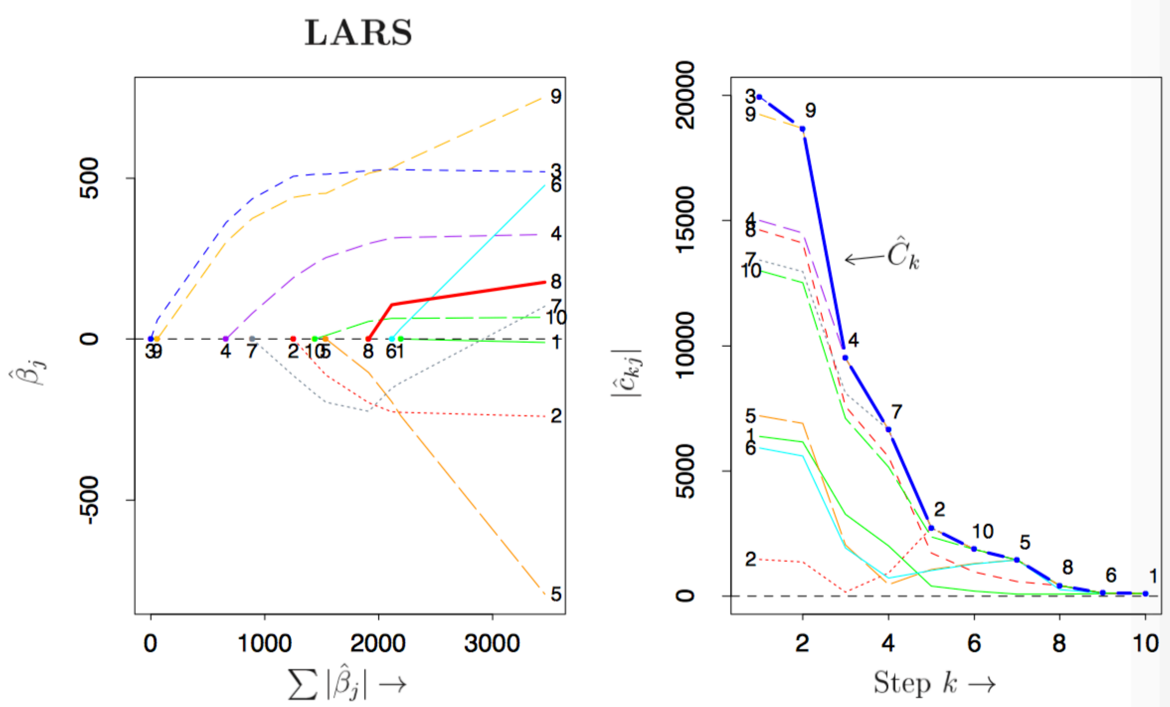

Figure 3. LARS analysis of the diabetes study. *Left*: estimates of regression coefficients $\widehat{\beta}_j$, $j = 1, 2, \ldots 10$; plotted versus $\Sigma|\widehat{\beta}_j|$; plot is slightly different than either Lasso or Stagewise, Figure 1. *Right*: Absolute current correlations as function of LARS step; variables enter active set (2.9) in order $3, 9, 4, 7, \ldots, 1$; heavy curve shows maximum current correlation $\widehat{C}_k$ declining with k.

The right panel shows the absolute current correlation goes done with the LARS step k.

$$\widehat{C}_k = \max\{|\hat{c}_{kj}|\} = \widehat{C}_{k-1} - \widehat{\gamma}_{k-1} A_{k-1} \qquad (28)$$

Declines with k.

Relation between LARS and OLS.

Suppose LARS has just completed step k-1, giving $\hat{\mu}_{k-1}$ and is embarking upon step k.The active set $\mathcal{A}_k$ will have k members, giving $X_k, \mathcal{G}_k, A_k$ and u_k . Similarly, let $\overline{y}_k$ indicate the projection of y into $\mathcal{L}\left(X_k\right)$, which, since $\widehat{\boldsymbol{\mu}}_{k-1} \in \mathcal{L}\left(X_{k-1}\right)$, is

$$\overline{\mathbf{y}}_k=\widehat{\boldsymbol{\mu}}_{k-1}+X_k \mathcal{G}_k^{-1} X_k^{\prime}\left(\mathbf{y}-\widehat{\boldsymbol{\mu}}_{k-1}\right)=\widehat{\boldsymbol{\mu}}_{k-1}+\frac{\hat{C}_k}{A_k} \mathbf{u}_k \tag{29}$$

因为等角性质和Ak里面的东西在correlation上同进退，所以有

$$X_k^{\prime}\left(\mathbf{y}-\widehat{\boldsymbol{\mu}}_{k-1}\right)=\widehat{C}_k 1_{\mathcal{A}} \tag{30}$$

Since u_k is a unit vector, (29) ，$\overline{y}_k$则有 $\overline{\mathbf{y}}_k-\widehat{\boldsymbol{\mu}}_{k-1}$ 有长度

$$\overline{\gamma}_k \equiv \frac{\widehat{C}_k}{A_k} \tag{31}$$

和update的那个公式进行比较，则LARS的估计$\hat{\mu}_k$ 在 $\hat{\mu}_{k-1}$到$\overline{y}_k$ 的延长线上.

$$\widehat{\boldsymbol{\mu}}_k-\widehat{\boldsymbol{\mu}}_{k-1}=\frac{\hat{\gamma}_k}{\overline{\gamma}_k}\left(\overline{\mathbf{y}}_k-\widehat{\boldsymbol{\mu}}_{k-1}\right) \tag{32}$$

可以发现一个问题, $\hat{\gamma}_k$ 总是比 $\overline{\gamma}_k$小，所以 LARS estimates always approaching but never reaching the OLS estimates $\overline{y}_k$.

有一个情况例外，如果LARS包含了所有的covariates，然后。。。。反正就和OLS等了。

因为一步到位的性质，LARS算起来特别快。

以上的计算都没有好好看，但是大概意思结论和过程都不复杂所以先放着吧。如果有用再拿起来看。

因为typero打了那么多字好卡，所以暂时把文本分割。