STATS305B – Lecture 11

Jonathon Taylor
Scribed by Michael Howes

02/09/22

Contents

1 Stationary conditions for the LASSO

The lecture will primarily be about stationary conditions for the LASSO. As we will see, these conditions do not allow us to directly solve the LASSO, we still need an iterative method like coordinate descent to do that. The stationary conditions are still useful. They let us describe what a typical LASSO solution looks like. This can be used provide guarantees that the LASSO will select the correct coordinates or be close to the true solution. The stationary conditions can also be used to check that the iterative algorithm has indeed converged to the minimizer.

1.1 Sub-gradients

The stationary conditions for the LASSO are a bit tricky since the LASSO penalty is not smooth. If we have a smooth, convex objective function $f(\beta)$ and a smooth, convex penalty $\mathcal{P}(\beta)$, then we have the first order conditions,

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} f(\beta) + \mathcal{P}(\beta) \iff \nabla f(\widehat{\beta}) + \nabla \mathcal{P}(\widehat{\beta}) = 0.$$

If the penalty $\mathcal{P}$ is convex but not necessarily smooth, then the first order conditions become

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} f(\beta) + \mathcal{P}(\beta) \iff \nabla f(\widehat{\beta}) + \widehat{u} = 0 \quad \text{for some } \widehat{u} \in \partial \mathcal{P}(\widehat{\beta}),$$

where $\partial \mathcal{P}(\widehat{\beta})$ is the *set of sub-gradients* of $\mathcal{P}$ at $\widehat{\beta}$. The set of sub-gradients is defined as follows

Definition 1. Let $\mathcal{P} : \mathbb{R}^p \to \mathbb{R}$ be convex. A vector $u \in \mathbb{R}^p$ is a *sub-gradient* of $\mathcal{P}$ at $\beta_0 \in \mathbb{R}^p$ if for all $\beta \in \mathbb{R}^p$,

$$\mathcal{P}(\beta) \geq \mathcal{P}(\beta_0) + u^T(\beta - \beta_0).$$

The set of all sub-gradients at β_0 is denoted by $\partial \mathcal{P}(\beta_0)$.

Proposition 1. *Suppose that* $\mathcal{P}(\beta) = \|\beta\|_q = \left(\sum_{j=1}^p |\beta|_j^q\right)^{1/q}$, *for some* $q \in [1, \infty]$. *Choose* $d \in [1, \infty]$ *so that* $\frac{1}{q} + \frac{1}{d} = 1$. *Then,*

$$\partial \mathcal{P}(\beta_0) = \{u \in \mathbb{R}^p : \|u\|_d \leq 1 \text{ and } u^T\beta = \|\beta_0\|_q\}.$$

Furthermore, if $\mathcal{P}(\beta) = \lambda \|\beta\|_q$, *then*

$$\partial \mathcal{P}(\beta_0) = \lambda \partial(\|\cdot\|_d)(\beta_0).$$

Proof. First suppose that $\|u\|_d \leq 1$ and $u^T\beta = \|\beta_0\|_q$. Let $\beta \in \mathbb{R}^p$ be given, then by Hölder's inequality

$$\begin{aligned}
\mathcal{P}(\beta_0) + u^T(\beta - \beta_0) &= \|\beta_0\|_q + u^T\beta - u^T\beta_0 \\
&= \|\beta_0\|_q + u^T\beta - \|\beta_0\|_q \\
&= u^T\beta \\
&\leq \|u\|_d \|\beta\|_q \\
&\leq \|\beta\|_q \\
&= \mathcal{P}(\beta).
\end{aligned}$$

Thus, $u \in \partial \mathcal{P}(\beta_0)$. Conversely, suppose that $u \in \partial \mathcal{P}(\beta_0)$. Let $\beta = \beta_0 + v$ where $v_j = \text{sign}(u_j)|u_j|^{q/d}$. Then

$$\begin{aligned}
\|u\|_d \|\beta - \beta_0\|_q &= \left(\sum_{j=1}^p |u|^q\right)^{1/q} \left(\sum_{j=1}^p |u|^q\right)^{1/d} \\
&= \sum_{j=1}^p |u_j|^q \\
&= \sum_{j=1}^p |u_j|^{q(1/q+1/d)} \\
&= \sum_{j=1}^p u_j \,\text{sign}(u_j)|u_j|^{q/d} \\
&= u^T(\beta - \beta_0)
\end{aligned}$$

Since $u \in \partial \mathcal{P}(\beta_0)$ we have

$$\|\beta\|_q \geq \|\beta_0\|_q + u^T(\beta - \beta_0).$$

Which implies that $\|u\|_d \|\beta - \beta_0\|_q \leq \|\beta\|_q - \|\beta_0\|_q$. Thus, either $u = 0$ or

$$\|u\|_d \leq \frac{\|\beta\|_q - \|\beta_0\|_q}{\|\beta - \beta_0\|_q} \leq 1.$$

Thus, in either case $\|u\|_d \leq 1$. It remains to show that $u^T\beta_0 = \|\beta\|_q$. By Hölder's inequality we know that $u^T\beta_0 \leq \|\beta\|_q$. For the other direction note that if $\beta = 0$, then we have

$$0 \geq \|\beta_0\|_q + u^T(-\beta_0) = \|\beta_0\|_q - u^T\beta_0.$$

Showing that $u^T\beta_0 \geq \|\beta_0\|_q$. For the additional comment, note that for any convex function $\mathcal{P}$,

$$\begin{aligned}
u \in \partial(\lambda\mathcal{P})(\beta_0) &\iff \lambda\mathcal{P}(\beta) \geq \lambda\mathcal{P}(\beta_0) + u^T(\beta - \beta_0) \quad \text{for all } \beta \\
&\iff \mathcal{P}(\beta) \geq \mathcal{P}(\beta_0) + \lambda^{-1}u^T(\beta - \beta_0) \quad \text{for all } \beta \\
&\iff \lambda^{-1}u \in \partial\mathcal{P}(\beta) \\
&\iff u \in \lambda\partial\mathcal{P}(\beta).
\end{aligned}$$

□

Corollary 1. *Let $\mathcal{P}(\beta) = \lambda \|\beta\|_1^1$, then $v \in \partial\mathcal{P}(\beta)$ if and only if $v = \lambda u$ where $\|u\|_\infty \leq 1$ and $u_j = \text{sign}(\beta_j)$ for all j such that $\beta_j \neq 0$.*

1.2 LASSO solutions

Consider using the LASSO for linear regression. The objective function is thus,

$$\widehat{\beta}_\lambda = \underset{\beta}{\text{argmin}} \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 .$$

Note that

$$\nabla_\beta \frac{1}{2} \|Y - X\beta\|_2^2 = X^T(Y - X\beta).$$

Thus, the sub-gradient condition is

$$-X^T(Y - X\widehat{\beta}) + \lambda\widehat{u} = 0,$$

where $\widehat{\mu} \in \partial(\|\cdot\|_1)(\widehat{\beta})$, meaning that

$$\widehat{u}_j \in \begin{cases} \{1\} & \text{if } \widehat{\beta}_j > 0, \\ [-1, 1] & \text{if } \widehat{\beta}_j = 0, \\ \{-1\} & \text{if } \widehat{\beta}_j < 0. \end{cases}$$

We can rewrite the sub-gradient condition as

$$X^T(Y - X\widehat{\beta}) = \lambda\widehat{u}. \tag{1}$$

Suppose that $\widehat{\beta}$ satisfies the above equation. Let E be the set of *active states*. That is, E contain the indices for which $\widehat{\beta}_j \neq 0$. Let $\widehat{\beta}_E$ be the vector in $\mathbb{R}^{|E|}$ with the non-zero entries of $\widehat{\beta}$. Define X_E be the matrix containing the columns $X[\cdot, j]$ for which $j \in E$. Define X_{-E} to be the matrix with columns $X[\cdot, j]$ for $j \notin E$. Finally, let $s_E = \text{sign}(\widehat{\beta}_E)$. Since $\widehat{\beta}_j = 0$ for $j \notin E$, we have

$$X\widehat{\beta} = X_E\widehat{\beta}_E.$$

Thus, (1) is equivalent to,

$$X^T(Y - X_E\widehat{\beta}_E) = \lambda\widehat{u}.$$

The equations (1) can be written as two blocks of equations. We have the equations for the active states:

$$X_E^T(Y - X_E\widehat{\beta}_E) = \lambda\widehat{u}_E = \lambda s_E, \tag{2}$$

since if $\widehat{\beta}_j \neq 0$, then $\widehat{u}_j = \text{sign}(\widehat{\beta}_j)$. Let $\bar{\beta}_E$ be the OLS solution which design matrix X_E. That is,

$$\bar{\beta}_E = (X_E^T X_E)^{-1} X_E^T Y.$$

We know that $X_E^T Y = (X_E^T X_E)\bar{\beta}_E$, by the stationary conditions for least squares. Therefore,

$$X_E^T X_E(\bar{\beta}_E - \widehat{\beta}_E) = X_E^T(Y - X_E\widehat{\eta}_E) = \lambda s_E.$$

Thus, we have

$$\widehat{\beta}_E = \bar{\beta}_E - \lambda(X_E^T X_E)^{-1} s_E. \tag{3}$$

We also have the conditions for the inactive states. These are

$$X_{-E}(Y - X_E\widehat{\beta}_E) = \lambda\widehat{u}_{-E}.$$

We know that for $j \notin E$, $\widehat{u}_j$ can be any number in $[-1, 1]$. Thus, the inactive conditions can be rewritten as

$$\left\| X_{-E}^T (Y - X_E \widehat{\beta}_E) \right\|_\infty \leq \lambda. \tag{4}$$

By equation (3), we have

$$\begin{aligned} \left\| X_{-E}^T (Y - X_E \widehat{\beta}_E) \right\|_\infty &= \left\| X_{-E}^T (Y - X_E (\bar{\beta}_E - \lambda (X_E^T X_E)^{-1} s_E)) \right\|_\infty \\ &= \left\| X_{-E}^T (Y - H_E Y) + \lambda X_{-E}^T X_E (X_E^T X_E)^{-1} s_E \right\|_\infty, \end{aligned}$$

where $H_E = X_E (X_E^T X_E)^{-1} X_E^T$ is the hat matrix for the states in the active block E. By the triangle inequality,

$$\begin{aligned} &\left\| X_{-E}^T (Y - H_E Y) + \lambda X_{-E}^T X_E (X_E^T X_E)^{-1} s_E \right\|_\infty \\ &\leq \left\| X_{-E}^T (Y - H_E Y) \right\|_\infty + \lambda \left\| X_{-E}^T X_E (X_E^T X_E)^{-1} s_E \right\|_\infty. \end{aligned}$$

Therefore, if $norm X_{-E}^T (Y - H_E Y)_\infty \leq \lambda \left(1 - \left\| X_{-E}^T X_E (X_E^T X_E)^{-1} s_E \right\|_\infty\right)$, then the inactive conditions (4) hold. What we have derived so far can be used to describe when the LASSO selects the correct coefficients with high probability. Suppose that $Y = X\beta^* + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let A be the set indices for the non-zero indices of β^* and let s_A be the signs of β^*. Suppose that the following hold

1. $\kappa = \left\| X_{-A}^T X_A (X_A X_A)^{-1} s_A \right\|_\infty < 1$.
2. $\lambda > 0$ is taken to be so large so that,

$$\left\| X_{-A}^T (I - H_A) Y \right\|_\infty \leq (1 - \kappa)\lambda,$$

   with high probability.

3. And,

$$\operatorname{sign}(\beta^* - \lambda (X_A^T X_A)^{-1} \operatorname{sign}(\beta_A^*)) = s_A.$$

Then, we have $E = A$ and $\operatorname{sign}(\widehat{\beta}_\lambda) = \operatorname{sign}(\beta_A^*)$ with high probability. The first condition holds if the columns of X are well conditioned and close to orthogonal. The next two conditions say something about the strength of the signal in the directions A versus the strength of the signal in the directions $-A$. Since the ∞-norm is a maximum, we expect that if $\operatorname{diag}(X^T X) = n$, then

$$\left\| X_{-A}^T (I - H_A) Y \right\| \approx \sqrt{n} \sqrt{2 \log(p)}.$$

Thus, taking λ to be of the order $\sqrt{n}\sqrt{2\log(p)}$ is commonly used in theory of the LASSO papers.

2 Log-linear models

Recall that a glm with a Poisson response and the canonical link function is called a log-linear model. We will see how such models can be used to model contingency tables.

2.1 Two-way tables

Suppose we have an $I \times J$ table. We can model the cell counts via

$$N_{ij} \sim \mathsf{Poisson}(\mu_{ij}),$$

subject to certain constraints on μ_{ij}. The *independence model* has

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y.$$

The *saturated model* has

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}.$$

In the independence model we typically set $\lambda_I^X = \lambda_J^Y = 0$ to make our model identifiable. For the saturated model we set

$$\lambda_I^X = \lambda_J^Y = \lambda_{iJ}^{XY} = \lambda_{Ij}^{XY} = 0,$$

to make our parameters identifiable. A common theme we will see in log linear models is that the form of the regression model relates to assumptions about conditional dependence.

2.2 Three-way tables

Suppose we introduce an additional variable Z taking values $1 \leq k \leq K$. We now have models of the form,

$$N_{ijk} \sim \mathsf{Poisson}(\mu_{ijk}).$$

For a three-way table, the *saturated model* is

$$\begin{aligned}\log(\mu_{ijk}) = \lambda &+ \lambda_i^X + \lambda_j^Y + \lambda_k^Z \\ &+ \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \\ &+ \lambda_{ijk}^{XYZ}.\end{aligned}$$

To make this model identifiable, we set parameters equal to 0 whenever $i = I, j = J$ or $k = K$. The *homogeneous association* model introduces the constraint $\lambda_{ijk}^{XYZ} = 0$ for all i, j, k. That is,

$$\begin{aligned}\log(\mu_{ijk}) = \lambda &+ \lambda_i^X + \lambda_j^Y + \lambda_k^Z \\ &+ \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.\end{aligned}$$

2.3 m-way tables

It is easy to see how the saturated model would extend to the case when we have m variables. Suppose the variables are $(Y_1, Y_2, \ldots, Y_m)$ each taking values $1 \leq j_m \leq J_m$. The homogeneous association model has only first and second order effects. That is,

$$\log(\mu_{j_1,\ldots,j_m}) = \lambda + \sum_{i=1}^{m} \lambda_{j_i}^{Y_i} + \sum_{i=1}^{m} \sum_{l=i+1}^{m} \lambda_{j_i,j_l}^{Y_i,Y_l}.$$

2.4 Homogeneous association models

Homogeneous association models are useful for testing conditional independence. For instance one test of $X \perp\!\!\!\perp Y | Z$ in a three-way table can be represented as

$$H_0 : \lambda_{ij}^X Y \equiv 0, \lambda_{ijk}^{XYZ} \equiv 0 \quad \text{vs} \quad H_1 : \lambda_{ijk}^{XYZ} \equiv 0,$$

where $\lambda_{ij}^{XY} \equiv 0$ means $\lambda_{ij}^{XY} = 0$ for all i, j. We could test such a model with a likelihood ratio test. Homogeneous association models are also useful since they can be represented by graphs. Each variable has a vertex and there is an edge between variables Y_i and Y_l if and only if $\lambda_{j_i,j_l}^{Y_i,Y_l} \neq 0$ for some j_i, j_l. Conditional independence information can easily be read off from such graphs.