ECON 480-3
LECTURE 13: LASSO

Ivan A. Canay
Northwestern University

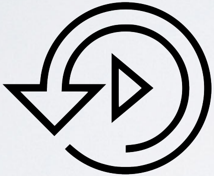
Past & Future

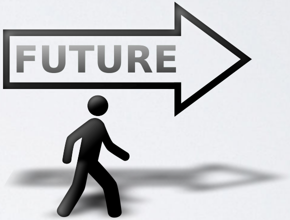
LAST CLASS

- Related to Classification Tress
- Latent Index and Identification
- Identification via Median Independence
- Parametric Models: Logit & Probit

TODAY

- Sparcity
- LASSO
- Properties
- Adaptive LASSO





- Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^k$.
- Let $\beta = (\beta_1, \ldots, \beta_k)' \in \mathbf{R}^k$ be such that

$$Y = X'\beta + U \; .$$

- **Data**: a random sample $\{(Y_i, X_i) : 1 \leqslant i \leqslant n\}$ from the distribution of (Y, X) and without loss of generality, we further assume that

$$\bar{Y}_n \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i = 0 \qquad \text{and} \qquad \hat{\sigma}^2_{n,j} \equiv \frac{1}{n} \sum_{i=1}^{n} (X_{i,j} - \bar{X}_j)^2 = 1 \; ,$$

  where $X_{i,j}$ denotes the j^{th} component of X_i.
- **Goal**: study estimation of β when k is large relative to n. That could mean that $k < n$, but not by much, or simply that $k > n$. For simplicity, we assume X and U are independent.

Sparcity

- $k > n$: the OLS estimator is not well-behaved - the $\mathbb{X}'\mathbb{X}$ matrix does not have full rank.
- The estimator is not unique and will overfit the data.
- If all explanatory variables are important in determining the outcome, it is not possible to tease out their individual effects.
- However, if the model is **sparse** then it might be possible to discriminate between the relevant and irrelevant components of X.

Definition (Sparsity)

Let $S = \{j : \beta_j \neq 0\}$ be the identity of the relevant regressors. A model is said to be sparse if $s = |S|$ is fixed as $n \to \infty$.

- **Oracle**: If we knew the identity of the relevant regressors S then we could do LS as usual.

Definition (Oracle Estimator)

The oracle estimator $\hat{\beta}_n^o$ is the infeasible estimator that is estimated by least squares using only the variables in S.

Consistency

In practice: we do not know the set S and so our goal is to estimate β and perhaps S. We do this by exploiting sparcity. Three properties are important.

Definition (Estimation Consistency)

An estimator $\hat{\beta}_n$ is estimation consistent if

$$\hat{\beta}_n \xrightarrow{P} \beta \ .$$

Definition (Model-Selection Consistency)

Let

$$\hat{S}_n = \{j : \hat{\beta}_{n,j} \neq 0\}$$

be the set of relevant covariates selected by an estimator $\hat{\beta}_n$. Then, $\hat{\beta}_n$ is model-selection consistent if

$$P\{\hat{S}_n = S\} \to 1 \text{ as } n \to \infty \ .$$

Definition (Oracle Efficiency)

An estimator $\hat{\beta}_n$ is oracle efficient if it achieves the same asymptotic variance as the oracle estimator $\hat{\beta}_n^o$.

LASSO

- **LASSO** is short for Least Absolute Shrinkage and Selection Operator and is one of the well known estimators for sparse models.

- The LASSO estimator $\hat{\beta}_n$ is defined as the solution to the following minimization problem

$$\hat{\beta}_n = \arg\min_b \left(\sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda_n \sum_{j=1}^{k} |b_j| \right) , \qquad (1)$$

  where λ_n is a scalar tuning parameter. It can be alternatively described as the solution to

$$\min_b \sum_{i=1}^{n} (Y_i - X_i' b)^2 \quad \text{subject to} \quad \sum_{j=1}^{k} |b_j| \leqslant t_n , \qquad (2)$$

  where now t_n is a scalar tuning parameter.

- LASSO corresponds to OLS with an additional term that imposes a **penalty** for non-zero coefficients.

- **Penalty term**: shrinks the estimated coefficients towards zero and this gives us model selection, albeit at the cost of introducing bias in the estimated coefficients.

Penalty Function

- **LASSO**: estimated coefficients can be **exactly** 0 for a given n.
- The form of the penalty function is important for selection, which does not occur under OLS or other penalty functions (e.g., ridge regression).
- **Intuition**: consider penalty functions of the form $\sum_{j=1}^{k} |b_j|^{\gamma}$.
- **If** $\gamma > 1$: the objective function is continuously differentiable at all points. The first order condition with respect to $\beta_{n,j}$ would be

$$2 \sum_{i=1}^{n} (Y_i - X_i' \beta) X_{i,j} = \lambda_n \gamma |\beta_j|^{\gamma-1} \text{sign}(\beta_j) \ .$$

Suppose $\beta_j = 0$. Then, $\hat{\beta}_{n,j} = 0$ **iff**

$$0 = \sum_{i=1}^{n} (Y_i - X_i' \hat{\beta}_n) X_{i,j} = \sum_{i=1}^{n} (U_i - X_i' (\hat{\beta}_n - \beta)) X_{i,j} \ .$$

If U is continuously distributed, this holds with **probability 0** and model selection **does not occur**.

Sub-Gradient

If $\gamma \leqslant 1$: the penalty function is not differentiable at 0. In this case, Karush-Kuhn-Tucker conditions are expressed in terms of the **subgradient**.

Definition (Sub-gradient & Sub-differential)

The scalar $g \in \mathbf{R}$ is a sub-gradient of $f(x) : \mathbf{R} \to \mathbf{R}$ at point x if $f(z) \geqslant f(x) + g \cdot (z - x)$ for all $z \in \mathbf{R}$.
The set of sub-gradients of $f(\cdot)$ at x, denoted by $\partial f(x)$, is the **sub-differential** of $f(\cdot)$ at x.

- **LASSO**: we need the sub-differential of the absolute value $f(x) = |x|$.
- For $x < 0$ the sub-gradient is uniquely given by $\partial f(x) = \{-1\}$ (for $x > 0$ it is $\partial f(x) = \{1\}$).
- At $x = 0$ the sub-differential is defined by the inequality $|z| \geqslant gz$ for all z, which holds for $g \in [-1, 1]$. Thus $\partial f(0) = [-1, 1]$.

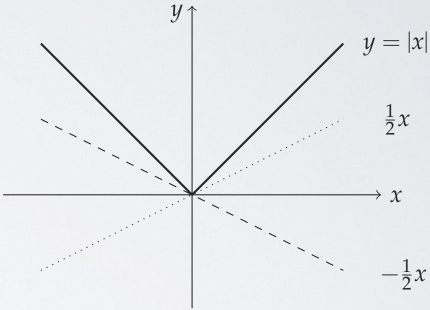


Figure: Two sub-gradients of $f(x) = |x|$ at $x = 0$

Exact Zeros

- For **non-differentiable functions**, the Karush-Kuhn-Tucker theorem states that a point minimizes the objective function **iff** 0 is in the sub-differential.

- Applying this to our problem gives

$$2\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_n)X_{i,j} = \lambda_n \,\mathsf{sign}(\hat{\beta}_{n,j}) \quad \text{if} \quad \hat{\beta}_{n,j} \neq 0$$

and

$$-\lambda_n \leqslant 2\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_n)X_{i,j} \leqslant \lambda_n \quad \text{if} \quad \hat{\beta}_{n,j} = 0\,.$$

- This inequality is attained with **positive probability** even when U is continuously distributed.

- Model selection is therefore possible when the penalty function has a cusp at 0.

Graphical Intuition

- The difference between using a penalty with $\gamma = 1$ (LASSO) and $\gamma = 2$ (Ridge) in the constraint problem in (2) is illustrated in Figure 2 for the simple case where $k = 2$.

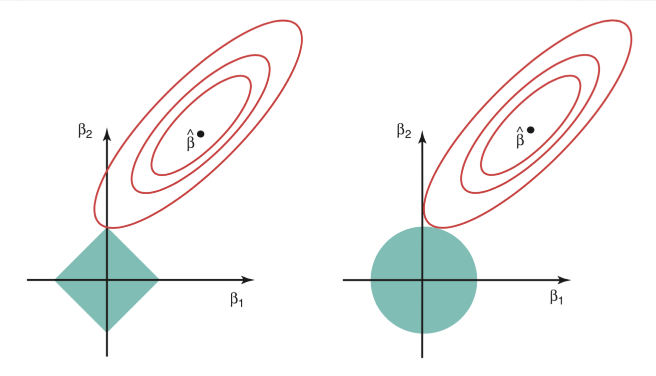


FIGURE: Constrained problem in (2) when $k = 2$: $\gamma = 1$ (left panel) and $\gamma = 2$ (right panel).

QUESTIONS?

Irrepresentable Condition

- For ease of exposition, we only discuss the case where k as fixed as $n \to \infty$.

- **WLOG**: S consists of the first s variables and partition X into $X = (X_1', X_2')'$ where X_1 are the first s explanatory variables. Partition the variance-covariance matrix of X accordingly,

$$\Sigma = E[XX'] = \begin{pmatrix} E[X_1 X_1'] & E[X_1 X_2'] \\ E[X_2 X_1'] & E[X_2 X_2'] \end{pmatrix}.$$

Assumption (Irrepresentable Condition)

$$\|E[X_2 X_1']E[X_1 X_1']^{-1} \cdot \text{sign}(\beta_1, \ldots, \beta_s)\|_\infty \leqslant 1 - \eta \quad \textit{for some } \eta > 0\,.$$

- **Note**: when the sign of β is unknown we require this to hold for all possible signs, i.e.,

$$\|E[X_1 X_1']^{-1} E[X_1 X_2{}']\|_\infty \leqslant 1 - \eta\,.$$

- **Interpretation**: the regression coefficients of the irrelevant variables on the relevant variables must all be less than 1 , i.e., the former are “irrepresentable” by the latter.

LASSO is Model Selection Consistent

Theorem (Zhao and Yu (2006))

Suppose k and s are fixed and that $\{X_i : 1 \leqslant i \leqslant n\}$ and $\{U_i : 1 \leqslant i \leqslant n\}$ are i.i.d. and mutually independent. Let X have finite second moments, and U have mean 0 and variance σ^2. Suppose also that the irrepresentable condition holds and that

$$\frac{\lambda_n}{n} \to 0 \quad \text{and} \quad \frac{\lambda_n}{n^{\frac{1+c}{2}}} \to \infty \quad \text{for} \quad 0 \leqslant c < 1 \,.$$

Then LASSO is **model-selection consistent**.

Discussion

- The irrepresentable condition is a **restrictive** condition.
- When this condition fails and $\lambda_n/\sqrt{n} \to \lambda^* > 0$, it can be shown that LASSO selects **too many** variables (i.e., it selects a model of bounded size that **contains** all variables in S).
- **Intuition**: if the relevant and irrelevant variables are highly correlated, we can't **discriminate** between them.
- Knight and Fu (2000) showed that the LASSO estimator is **asymptotically normal** when

  $$\lambda_n/\sqrt{n} \to \lambda^* \geqslant 0$$

  but that the nonzero parameters are estimated with asymptotic bias if $\lambda^* > 0$.
- If $\lambda^* = 0$, LASSO has the same limiting distribution as the LS estimator and so is not oracle efficient.
- **Note**: $\lambda_n/\sqrt{n} \to \lambda^* \geqslant 0$ is **at conflict** with $\lambda_n/n^{\frac{1+c}{2}} \to \infty$ and so LASSO **cannot be both** model selection consistent and asymptotically normal (hence oracle efficient) at the same time.
- **Goal**: penalize small coefficients a lot and large coefficients very little or not at all. This could be done by using weights or by changing the penalty function.

Adaptive LASSO

Definition (Adaptive LASSO)

The adaptive LASSO is the estimator $\tilde{\beta}_n$ that arises from the following **two steps**.

1. Estimate β using ordinary LASSO,

$$\hat{\beta}_n = \arg\min_b \left(\sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda_{1,n} \sum_{j=1}^{k} |b_j| \right),$$

   where $\lambda_{1,n}/\sqrt{n} \to \lambda^* > 0$.

2. Let $\hat{S}_1 = \{j : \hat{\beta}_n \neq 0\}$ be the set of selected covariates from the first step. Estimate β by

$$\tilde{\beta}_n = \arg\min_b \left(\sum_{i=1}^{n} (Y_i - \sum_{j \in \hat{S}_1} X_{i,j} b_j)^2 + \lambda_{2,n} \sum_{j \in \hat{S}_1} |\hat{\beta}_{n,j}|^{-1} |b_j| \right),$$

   where $\lambda_{2,n}/\sqrt{n} \to 0$ and $\lambda_{2,n} \to \infty$.

Note: Adaptive LASSO imposes a penalty in the second step that is inversely proportional to the magnitude of the estimated coefficient in the first step.

Theorem (Zou (2006))

Suppose $\{X_i : 1 \leqslant i \leqslant n\}$ and $\{U_i : 1 \leqslant i \leqslant n\}$ are i.i.d. and mutually independent. Let X have finite second moments, and U have mean 0 and variance σ^2. The adaptive LASSO is **model selection consistent** *and* **oracle efficient**, *i.e.,*

$$\sqrt{n}(\tilde{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma^2 E(X_1 X_1')^{-1}) .$$

- **Oracle efficiency**: note that the asymptotic variance is the same we would have achieved had we known the set S and performed OLS on it. The rates of $\lambda_{1,n}$ and $\lambda_{2,n}$ are important for this result.
- To see why the adaptive LASSO is model selection consistent and oracle efficient, consider the following argument.
- Recall that $\beta_1, \ldots, \beta_s \neq 0$ and $\beta_{s+1}, \ldots, \beta_k = 0$.
- Suppose that $\hat{\beta}_n$ has r non-zero components asymptotically (the first r components wlog).
- Without the irrepresentable condition, the LASSO includes too many variables, so that $s \leqslant r \leqslant k$.

Informal Argument

Let $u = \sqrt{n}(b - \beta)$ where b is any $r \times 1$ vector. Let $\tilde{\beta}_n$ be the adaptive LASSO estimator.

$$\sqrt{n}(\tilde{\beta}_n - \beta) = \arg\min_u \sum_{i=1}^{n} \left(U_i - \frac{1}{\sqrt{n}} \sum_{j=1}^{r} X_{i,j} u_j \right)^2 + \lambda_{2,n} \sum_{j=1}^{r} |\hat{\beta}_{n,j}|^{-1} (|\beta_j + \frac{1}{\sqrt{n}} u_j| - |\beta_j|) \; .$$

CASE 1: $\beta_j = 0$

Informal Argument

Let $u = \sqrt{n}(b - \beta)$ where b is any $r \times 1$ vector. Let $\tilde{\beta}_n$ be the adaptive LASSO estimator.

$$\sqrt{n}(\tilde{\beta}_n - \beta) = \arg\min_u \sum_{i=1}^{n} \left(U_i - \frac{1}{\sqrt{n}} \sum_{j=1}^{r} X_{i,j} u_j \right)^2 + \lambda_{2,n} \sum_{j=1}^{r} |\hat{\beta}_{n,j}|^{-1} \left(|\beta_j + \frac{1}{\sqrt{n}} u_j| - |\beta_j| \right) .$$

CASE 2: $\beta_j \neq 0$

QUESTIONS?

Penalties for Model Selection Consistency

- Another way to achieve a model-selection consistent estimator is to use a penalty function that is **strictly concave** (as a function of $|b_j|$) and has a **cusp at the origin**.

- LASSO is essentially OLS with an L^1 **penalty** term. As such, it belongs to the larger class of **Penalized Least Squares** estimators:

$$\hat{\beta}_n^{PLS}(\lambda) = \arg\min_b \left(\sum_{i=1}^{n} (Y_i - X_i' b)^2 + \sum_{j=1}^{k} p_\lambda(|b_j|) \right).$$

- LASSO corresponds to the case where $p_\lambda(|v|) = \lambda|v|$, but such a penalty is not strictly concave and so model selection consistency generally does not occur.

- Some alternative penalty functions include that have the desire property are: **Bridge, Smoothly Clipped Absolute Deviation (SCAD), and Minimax Concave.**

Penalties

Alternative **penalty functions** that have the desire property:

1. **Bridge**: $p_\lambda(|v|) = \lambda|v|^\gamma$ for $0 < \gamma < 1$

2. **SCAD**: for $a > 2$,

$$p'_\lambda(|v|) = \lambda\left[I\left\{|v| \leqslant \frac{\lambda}{n}\right\} + \frac{(a\lambda/n - |v|)_+}{(a-1)\lambda/n} I\left\{|v| > \frac{\lambda}{n}\right\}\right]$$

3. **Minimax Concave**: for $a > 0$,

$$p_\lambda(|v|) = \lambda \int_0^{|v|} \left(1 - \frac{nx}{a\lambda}\right)_+ dx$$

where $(x)_+ = \max\{0, x\}$.

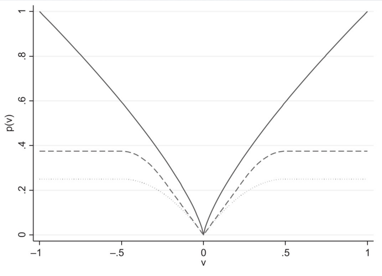


FIGURE: Bridge penalty (solid line), SCAD penalty (dashed line) and minimax concave penalty (dotted line)

Choosing lambda

- **Model selection consistency** imposes constraints on the growth rate of λ_n.
- λ_n for the ordinary LASSO is often chosen by Q-fold cross validation.

Cross Validation

Let Q be some integer and $n = Qn_q$

1. Partition the sample into the sets $I_1, \ldots, I_Q$ each with n_q members.
2. For each $1 \leqslant q \leqslant Q$, perform LASSO on all **but** the observations in I_q to obtain $\hat{\beta}_{n,-q}(\lambda)$.
3. Calculate the squared prediction error of $\hat{\beta}_{n,-q}(\lambda)$ on the set I_q:

$$\Gamma_q(\lambda) = \sum_{i \in I_q} (Y_i - X_i' \hat{\beta}_{n,-q}(\lambda))^2 .$$

4. Doing so for each q, find **total error** for each λ: $\Gamma(\lambda) = \sum_{q=1}^{Q} \Gamma_q(\lambda)$.

- We define the **cross validated λ** as:

$$\hat{\lambda}_n^{CV} = \arg\min_{\lambda} \Gamma(\lambda) .$$

LASSO WITH CV

- There exist **few results** about the properties of the LASSO when λ_n is chosen via cross-validation.
- **Recent paper**: Chetverikov et al (2020, annals) show that in a model where k is allowed to depend on n, and assuming $U_i|X_i$ is Gaussian, it follows that

$$\|\hat{\beta}_n - \beta\|_{2,n} \leqslant Q \cdot ((|S| \log k)/n)^{1/2} \log^{7/8}(kn)$$

  holds with high probability, where $\|b - \beta\|_{2,n} = (\frac{1}{n}\sum_{i=1}^{n}(X_i'b)^2)^{1/2}$ is the prediction norm.
- $((|S| \log k)/n)^{1/2}$ is the **fastest convergence rate** possible so that cross-validated LASSO is nearly optimal.
- Not known if the $\log^{7/8}(kn)$ term can be dropped.

Remarks

- There are **other ways** to choose λ_n
- **Example**: Minimize the Bayesian Information Criterion where

$$\hat{\sigma}^2(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_n(\lambda))^2 \quad \text{and} \quad BIC(\lambda) = \log\left(\hat{\sigma}^2(\lambda)\right) + |\hat{S}_n(\lambda)|C_n\frac{\log(n)}{n}$$

  where C_n is an arbitrary sequence that tends to ∞.
- Under some conditions, choosing λ_n to minimize $BIC(\lambda)$ leads to model selection consistency when U is normally distributed.
- Today we focused on the framework that keeps k fixed even as $n \to \infty$. There exist many extensions to the stated theorems that are valid in cases where $k_n = O(n^a)$ or even $k_n = O(e^n)$.
- Many **packages** exist for LASSO estimation: `lassopack` in Stata and `glmnet` or `parcor` in R.
- Joel will teach an entire quarter on the LASSO in 481-1 next year!

THE END!