Machine Learning

Marica Valente[1]

June 25, 2024

[1]Assistant Professor, University of Innsbruck (marica.valente@uibk.ac.at)

Outline

Chapter 2: Regression-based ML (Parametric ML)

Parametric methods for HD data

J. R. Statist. Soc. B (1996)
58, *No.* 1, *pp.* 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

SUMMARY

We propose a new method for estimation in linear models. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits the stability of ridge regression. There is also an interesting relationship with recent work in adaptive function estimation by Donoho and Johnstone. The lasso idea is quite general and can be applied in a variety of statistical models: extensions to generalized regression models and tree-based models are briefly described.

Keywords: QUADRATIC PROGRAMMING; REGRESSION; SHRINKAGE; SUBSET SELECTION

Linear Regression with OLS

Consider the OLS coefficient vector $\beta = (\beta_1 \dots \beta_p)$. Let's represent the OLS model for all $i = 1, ..., n$ using matrices as $y_i = \beta' x_i + \epsilon_i$:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- y_i is the observed value of the dependent variable for observation i.
- x_{ik} is the value of the independent variable k for observation i.
- β_0 is the intercept; $\beta_0, \beta_2, \dots, \beta_p$ are the coeff. for $x_1, x_2, \dots, x_p$.
- ϵ_i represents the error term for observation i.

Find $\beta_0, \beta_1, \dots, \beta_p$ that $min_\beta \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 = (y_i - \beta' x_i)^2$.

Parametric methods for HD data

- Find a **convex criterion** that can be minimized efficiently for large p
- Take OLS criterion and fit a **constrained** problem for $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)'$ and constant $c > 0$:

$$min_\beta \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta' x_i)^2 \quad s.t. \quad \|\beta\| \leq c$$

- Put a constraint on the size of the βs (sum of their values) such that βs kept in the model if we experience a substantial decrease in RSS
- "Important" variables (leading to greater RSS reduction) will have a bigger coefficient

Parametric methods for HD data

- Find a **convex criterion** that can be minimized efficiently for large p
- Take OLS criterion and fit a **constrained** problem for $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)'$ and constant $c > 0$:

$$min_\beta \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta' x_i)^2 \quad s.t. \quad \|\beta\| \leq c$$

- Put a constraint on the size of the βs (sum of their values) such that βs kept in the model if we experience a substantial decrease in RSS
- "Important" variables (leading to greater RSS reduction) will have a bigger coefficient (rescale variables beforehand! -mean / sd)

Parametric methods for HD data

- Find a **convex criterion** that can be minimized efficiently for large p
- Take OLS criterion and fit a **constrained** problem for $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)'$ and constant $c > 0$:

$$min_\beta \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta' x_i)^2 \quad s.t. \quad \|\beta\| \leq c$$

- Put a constraint on the size of the βs (sum of their values) such that βs kept in the model if we experience a substantial decrease in RSS
- "Important" variables (leading to greater RSS reduction) will have a bigger coefficient (rescale variables beforehand! -mean / sd)
- Three types of norm regularization: l_0 norm, l_1 norm, and l_2 norm

$$\|\beta\|_0 = \sum_{j=1}^{p} 1_{\beta_j \neq 0} \ , \qquad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \ , \qquad \|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$$

- Best subset selection ($\|\beta\|_0$), LASSO ($\|\beta\|_1$), Ridge regression ($\|\beta\|_2^2$)

l_0-penalization: Best subset selection

- Consider **all possible combinations** of variables x (=subsets)
- Run a model for each subset and select subset that yields the best fit
- Computationally impractical: n. of subsets grows exponentially with x
- NP hard (non-deterministic polynomial-time)
- Requires a search of combinatorial order 2^k, where k is number of x
- With 5 predictors, evaluate models for subsets of sizes 0, 1, 2, 3, 4, 5 which corresponds to $2^5 = 32$ combinations

Alternative? Forward stepwise addition

- Start with no variables in the model, test the addition of each variable using LS criterion, add the variable whose inclusion improves the fit the most, repeat this process until none improves the model to a stat. sign. extent
- Still, computationally heavy, high risk of false positives, p-hacking (Flom and Cassell, 2007. "Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use", NESUG papers)
- **Unstable** estimator: Changing just one data point can cause a large change in the minimizer of RSS (Breiman, 1996)

$\Rightarrow$ We are **unable to locate the best model** (that's why is not used!)

l_1-penalization and selection: LASSO regression

- Substitute l_0 by a closest **convex function** - l_1 (Tibshirani, 96)
- In penalized form ($\lambda > 0$), problem for l_1 is **equivalent to**:

$$min_\beta \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta' x_i)^2 + \lambda\|\beta\|_1 \quad \text{or} \quad min_\beta (Y - X\beta)^2 + \lambda\|\beta\|_1$$

- Setting the penalty λ large enough induces the solution β to include only a subset of the original predictors
- LASSO is **model selection device**
- LASSO stands for Least Absolute Shrinkage and Selection Operator

LASSO induces variable selection

- Support of LASSO estimator is **subset** of $\{1, \ldots, p\}$ for λ large enough
- Let $A = supp(\hat{\beta})$ be LASSO active set (nonzero surviving coefficients)
- Let $s_A = sign(\hat{\beta}_A)$ be the signs of active coefficients X_A
- FOC of LASSO criterion is: $X_A'(Y - X_A\hat{\beta}_A) = \lambda s_A$
- Solution is: $\hat{\beta}_A = \frac{X_A'Y - \lambda s_A}{X_A'X_A}$, $\hat{\beta}_{-A} = 0$
- Why does LASSO induce $\hat{\beta}_{-A} = 0$?

Intuition

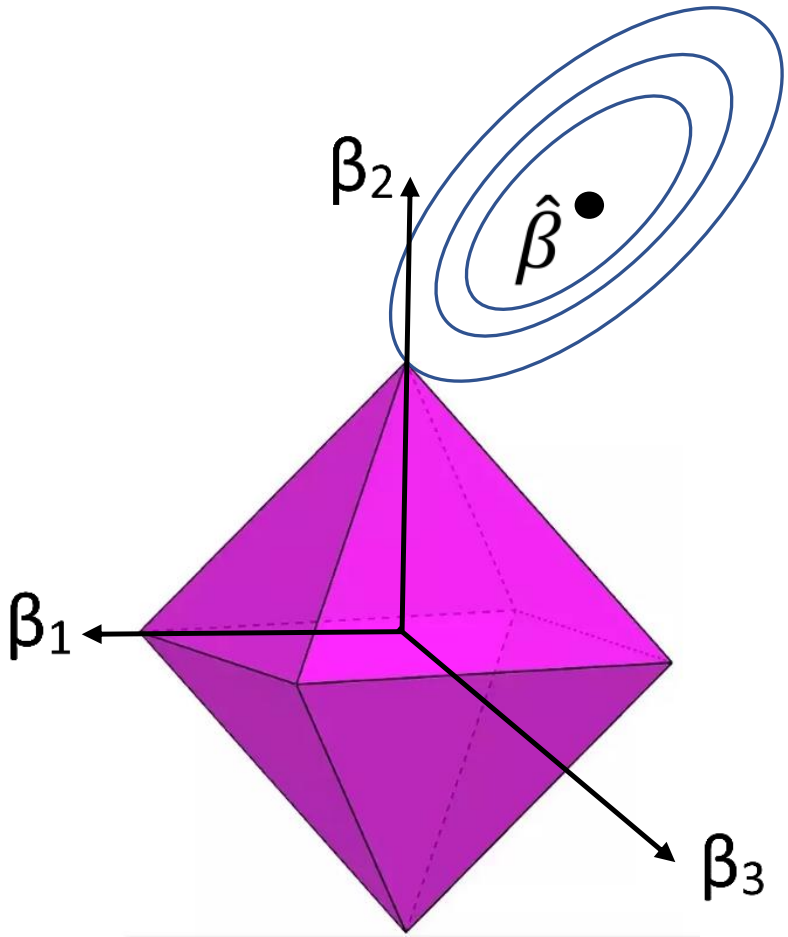
$$min_\beta \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta' x_i)^2 + \lambda\|\beta\|_1 \quad \text{or} \quad min_\beta (Y - X\beta)^2 + \lambda\|\beta\|_1$$

- If $\lambda \approx 0$, standard OLS solution (no shrinkage), all $\hat{\beta} \neq 0$ (constraint is not binding)
- If $\lambda >>> 0$, some coefficients are shrunk to zero, i.e., some $\hat{\beta} = 0$ and some $\hat{\beta} \neq 0$ (constraint becomes binding)

$\Rightarrow$ In general, LASSO does shrinkage for λ "high enough"

LASSO in 3D

- LASSO with three predictors x_1, x_2, x_3
- Diamond-shaped penalty with kinks: Some coeff $= 0$ (on axis)
- Solution is where the OLS convex criterion is tangent to the constraint



Pros and cons of LASSO

Pros: Model selection

1. The only norm regularization that offers **sparsity AND convexity**
2. Sparse solution has **few nonzero coefficients**
3. Sparse results are **easier to interpret** than nonsparse results

Cons: Shrinkage bias

1. Sparsity ass. needs to be **plausible**: Few large coefficients bounded away from zero (LASSO) vs. Gradual decay towards zero (Ridge)
2. Model selection is always **imperfect**: Theoretically impossible to distinguish coefficients that are local to zero and zero
3. LASSO regressions **bias** (shrink) non-zero coefficients **towards zero**

Shrinkage bias in LASSO

- LASSO does NOT have a UNIQUE solution (loss function not *strictly* convex)
- Among two highly correlated x_1 and x_2, LASSO drops one and typically shrinks the coefficient of the other
- $|\beta_1^{LASSO}| + ... + |\beta_j^{LASSO}|$ (sum of selected coefficients by LASSO) $<$ $|\beta_1^{OLS}| + ... + |\beta_j^{OLS}|$ (sum coefficients from OLS)
- Sometimes a single selected coefficient $|\beta_j^{LASSO}|$ can be BIGGER than $|\beta_j^{OLS}|$ because LASSO can overcompensate the shrinkage
- Depending on the sign of correlation between outcome, omitted variable, and selected variable, LASSO can inflate single coefficients

Shrinkage bias

- FOC of LASSO criterion is: $X_A'(Y - X_A\hat{\beta}_A) = \lambda s_A$
- Solution is: $\hat{\beta}_A = \frac{X_A'Y - \lambda s_A}{X_A'X_A}$, $\hat{\beta}_{-A} = 0$
- Solution is **not unique** though fitted $X_A\hat{\beta}_A$ are unique
- Shrinkage amounts to $\frac{\lambda s_A}{X_A'X_A}$

$\Rightarrow$ Can we take OLS solution on the active set and substract $\frac{\lambda s_A}{X_A'X_A}$?

- No, $\frac{X_A'Y}{X_A'X_A} \neq \frac{X'Y}{X'X} = \hat{\beta}_A^{OLS}$ if X_A not orthogonal (correlated X_A)

$\Rightarrow$ $\hat{\beta}_A > \hat{\beta}_A^{OLS}$ (in magnitude) possible for some active regressors

$\Rightarrow$ However, $\|\hat{\beta}_A\|_1 < \|\hat{\beta}_A^{OLS}\|_1$: l_1-norm of LASSO (sum of all $|\hat{\beta}_A|$) always smaller than OLS on the active set (Tibshirani, 1996)

Ridge regression

- Ridge estimator **replaces** l_1 **with** l_2 **penalty** (Hoerl and Kennard, 88)
- Selection property of LASSO induced by non-smooth l_1 penalty
- Since l_2 penalty is smooth, Ridge estimator does **not select variables**
- Ridge reg keeps all predictors: is **continuous shrinkage method**
- Original motivation: Use λ to make problem non-singular even if $X'X$ is non-invertible due to high correlation (Hoerl and Kennard, 70)
- In penalized form ($\lambda > 0$), problem for l_2 is **equivalent to**:

$$min_\beta \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta' x_i)^2 + \lambda \|\beta\|_2^2 \quad \text{or} \quad min_\beta (Y - X\beta)^2 + \lambda \|\beta\|_2^2$$

- Solution is **never sparse**: all $\hat{\beta} \neq 0$ (no model selection)

Orthonormal X in Ridge regression

- If X is orthonormal, then $X'X = I_p$, and a closed form property exists
- Let $\hat{\beta}^{OLS}$ denote the LS solution for our orthonormal $X = (X_1, \ldots, X_p)$:

$$\hat{\beta}^{RIDGE} = \frac{1}{1 + 2\lambda I_p} \hat{\beta}^{OLS}$$

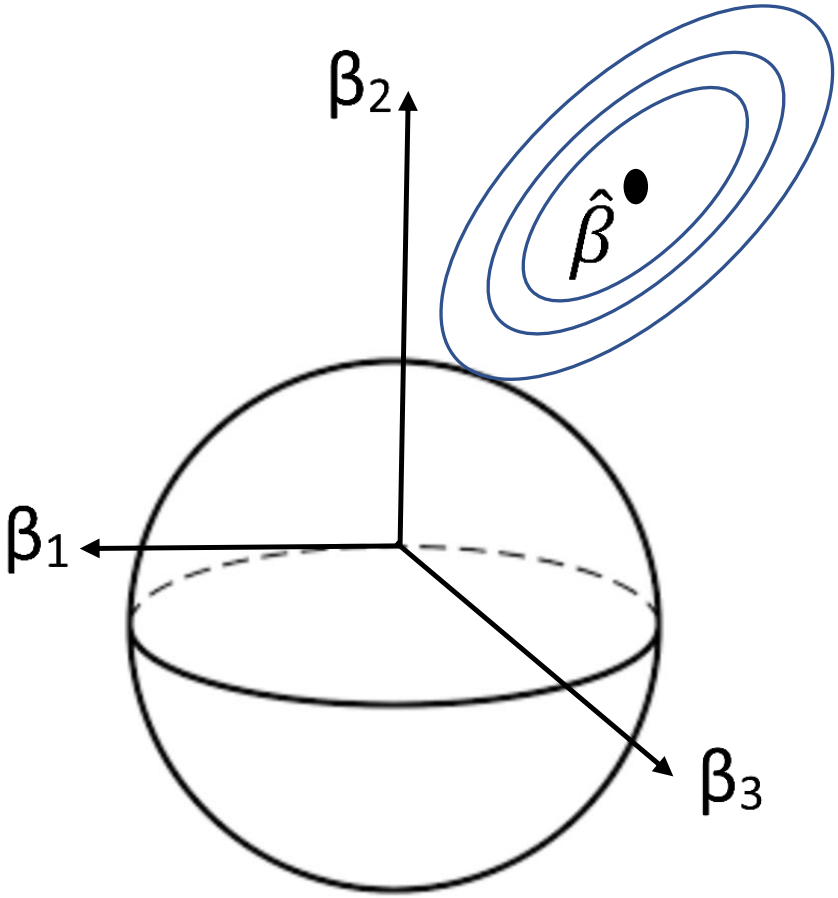
- All OLS coefficients are penalized in the same way

Correlated X in Ridge regression

- If two variables are highly correlated, Ridge will tend to estimate similar coefficients vs. LASSO will send one to zero
- Two variables x, z centered and scaled (to mean zero, variance one)
- Since x, z can approximately substitute each other in predicting Y, many linear combination of x, z will be good predictors
- For example: $(0.2x + 0.8z)$ or $(0.3x + 0.7z)$ or $(0.5x + 0.5z)$
- Ridge penalty is: $\beta_1^2 + \beta_2^2 = \{0.68, 0.58, 0.5\}$
- LASSO penalty is: $|\beta_1| + |\beta_2| = 1$ in all three cases
- Lowest Ridge penalty (less binding given λ) is 0.5 corresponding to assigning equal weight (coefficient) to highly correlated variables
- How is the penalization split among the two variables? In such a way to keep the total effect constant

Ridge in 3D

- Sphere-shaped penalty with no kinks: Coeff. $\to 0$ but always $\neq 0$
- Solution is where the OLS convex criterion is tangent to the constraint



LASSO and Ridge penalty in higher dimensions

1. LASSO penalty spikier, volume increases with mass far from zero (in the spikes)

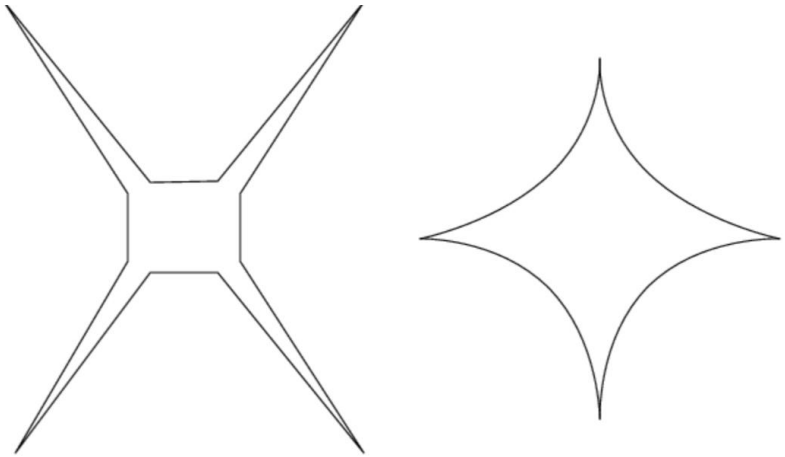
2. Ridge and **the strange geometry of HD spheres**: when adding dimensions, the p-dimensional sphere has still to satisfy its equation $\beta_1^2 + \cdots + \beta_p^2 = c$ with $c \geq 0$

$\Rightarrow$ Equation satisfied if each coordinate (coefficient) gets a smaller share of c

$\Rightarrow$ HD spheres get spiky and concentrate (shrink) around the horizontal axis (equator)

$\Rightarrow$ The volume of the unit p-dimensional sphere goes to 0 as p increases!

Figure: LASSO vs. Ridge penalty in HD



Elastic-net regression

- Elastic-net estimator involves **both l_1 and l_2 penalty**
- LASSO does not care which variable is selected among two correlated
- Elastic net encourages a **grouping effect**, where strongly correlated predictors tend to be in or out of the model together
- Ridge penalty removes the limitation on the n. of selected variables
- In penalized form ($\lambda, \mu \geq 0$), problem is **equivalent to**:

$$min_\beta (Y - X\beta)^2 + \lambda \|\beta\|_2^2 + \mu \|\beta\|_1$$

- Elastic Net solution converges either to Ridge or LASSO solution depending on relative size of λ, μ

Zou and Hastie, J. R. Statist. Soc., 2005

Elastic net: Heuristics

- Finding a set of highly correlated variables is relevant in many settings, e.g., gene selection
- **Grouping effect** = Using elastic net, coefficients of a group of highly correlated variables tend to be equal
- Ridge penalty pushes many of the correlated features toward each other: In the extreme situation where some variables are exactly identical, coefficients should be identical
- Elastic Net penalty function is **strictly convex** (unique solution!)
- Pairs of correlated variables: **LASSO** kicks one out, **Ridge** assigns similar coeffs, **Elastic Net** assigns similar coeffs to selected variables but also kicks out groups of irrelevant correlated variables

Improving elastic net performance

- Yet, elastic net performs poorly in many settings when resulting estimator is far from pure LASSO or Ridge
- Why? Two-step procedure: First find Ridge penalty level, and then do LASSO-type shrinkage
- Double shrinkage does not help to reduce the variances much and introduces extra bias
- Rescale elastic net estimator by $1 + \lambda$, i.e., $(1 + \lambda)\hat{\beta}$
- Motivation for the $(1 + \lambda)$-rescaling comes from a decomposition of the Ridge operator (Zou and Hastie, 2005)

$\Rightarrow$ Rescaling the naive estimator preserves the variable selection property and undoes shrinkage bias

$\Rightarrow$ **Rescaled Elastic Net** often outperforms LASSO and Ridge regression

Bridge regression

- Is a **generalization** of both the LASSO and Ridge regression
- In penalized form ($\lambda, q > 0$), problem is **equivalent to**:

$$min_{\beta}(Y - X\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

- **Adaptively** selects q the penalty order and level λ from data
- For $q \in (1; 2)$: Bridge $\sim$ Elastic net, but Bridge is NOT sparse
- For $q \in (0; 1]$: Bridge estimators produce sparse models

Frank and Friedman, 1993; Fu, 1998

Geometry of l_q norm in 2D

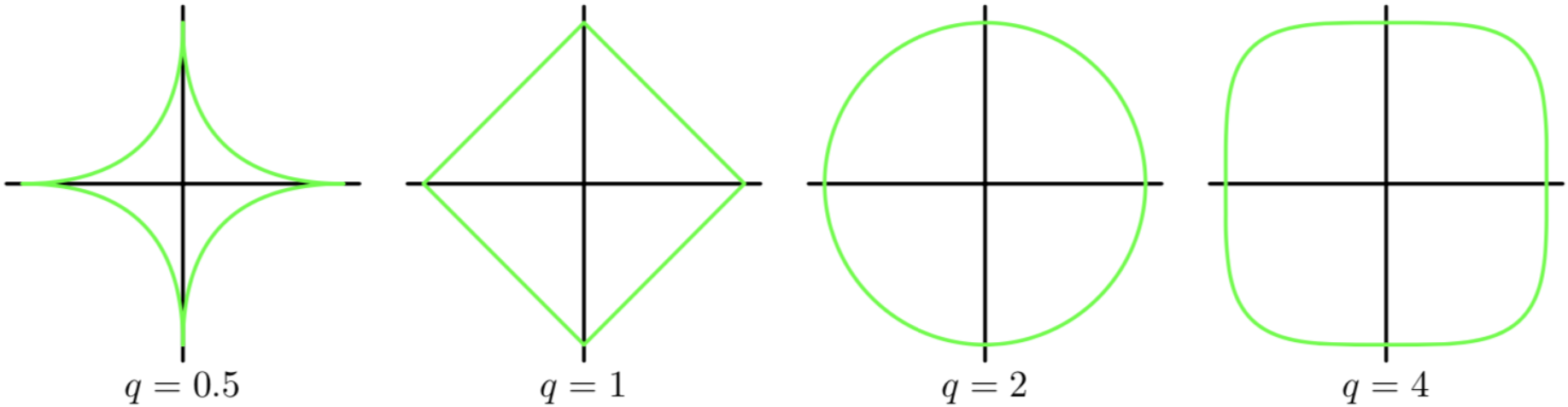


Figure: Penalties of Bridge (sparse), LASSO, Ridge, Bridge (non-sparse) regressions.

Bridge models in action

- Used for prediction with irregularities in data availability
- Link ("bridge") high-frequency variables, e.g. financial data or Google Trends, to low-frequency ones, e.g. the quarterly GDP growth
- Provide **nowcasting** of current and short-term developments in advance of the release of the low-frequency variable
- E.g.: Nowcasting GDP growth is an important task to inform decision makers about the current state of the economy
- "Bridge Models to Forecast the Euro Area GDP", Baffigi et al. (2004)
- "Short-term Forecasts of Euro Area Real GDP Growth. An Assessment of Real-Time Performance Based on Vintage Data", Diron

LASSO vs. Ridge vs. Bridge vs. Elastic net

- Tibshirani (1996) and Fu (1998) compared the prediction performance of LASSO, Ridge and Bridge regression and found that none of them uniformly dominates the other two
- LASSO/Elastic net parsimonious and interpretable
- Ridge outperforms LASSO when X highly correlated (Tibshirani, 1996)
- Elastic net outperforms LASSO when variables have high pairwise correlation (Zou and Hastie, 2005)
- Elastic net outperforms LASSO especially when $p >> n$
- Bridge has no theoretical justification (e.g. is (non)sparsity plausible?)

Implementation in R

R Packages

- LASSO: hdm (Chernozhukov et al.), glmnet (Friedman et al.)
- Ridge: glmnet
- Elastic net: glmnet, elasticnet (Zou and Hastie)
- Bridge: grpreg (Breheni and Zeng), bridge function in Iqa (Ulbricht)
- R applications: `https://www.pluralsight.com/guides/linear-lasso-and-ridge-regression-with-r`
- R codes for k-fold CV of the penalty term: `https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net`

Appendix

Orthonormal setting

- Assume X orthonormal, so $X'X = I_p$ (uncorrelated regressors)
- Solution becomes: $\hat{\beta}_A = X'_A Y - \lambda s_A \ , \quad \hat{\beta}_{-A} = 0$

Three cases:

1. $\hat{\beta}_A > 0$ $(s_A > 0)$ enforces $X'_A Y > 0$ and $X'_A Y > \lambda$
2. $\hat{\beta}_A < 0$ $(s_A < 0)$ enforces $X'_A Y < 0$ and $|X'_A Y| > \lambda$
3. $\hat{\beta}_{-A} = 0$ when $|X'Y| < \lambda$

$\Rightarrow$ LASSO does shrinkage for $|X'Y| < \lambda$

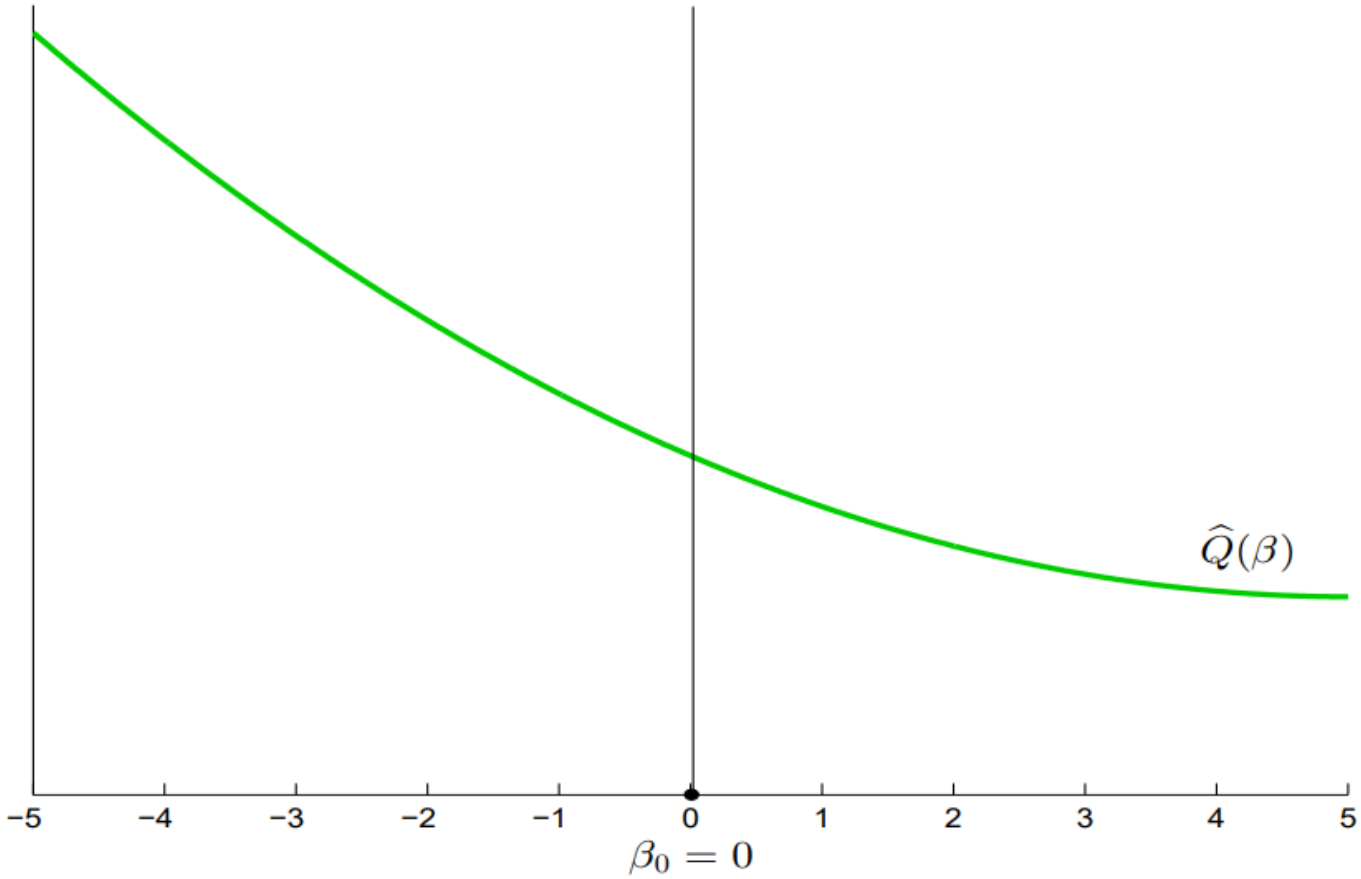
The geometry of LASSO

- Define $\hat{Q}(\beta)$ the LS criterion (squared-error loss, RSS)
- Assume the model $Y = \beta_0' X + \epsilon$, but true $\beta_0 = 0$ (irrelevant)
- How do we penalize LS criterion to shrink β_0 to zero?

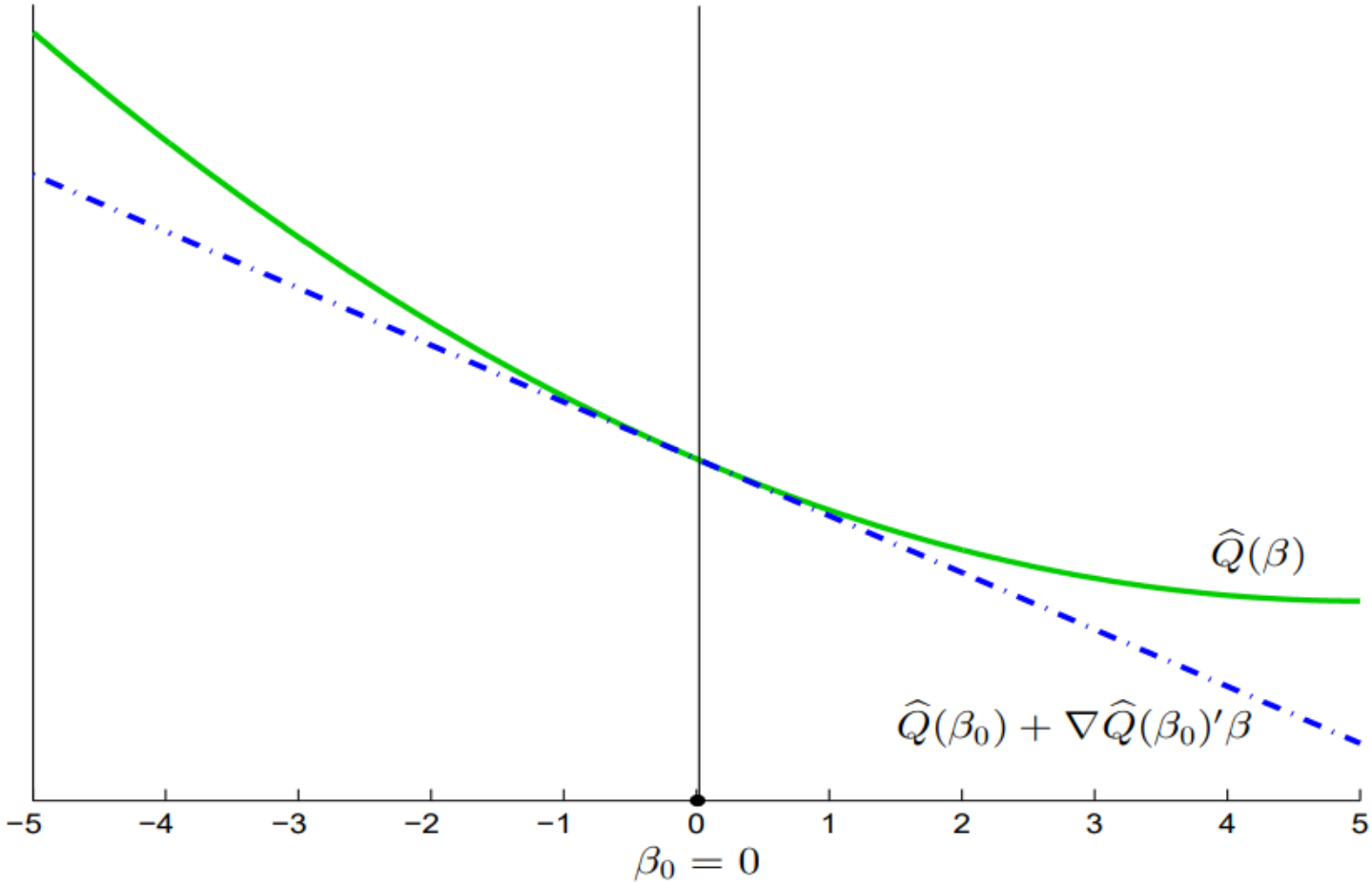
Chernozhukov, V. and Hansen, C. (2013). "Econometrics of High-Dimensional Sparse Models", NBER Seminar

LASSO in 2D

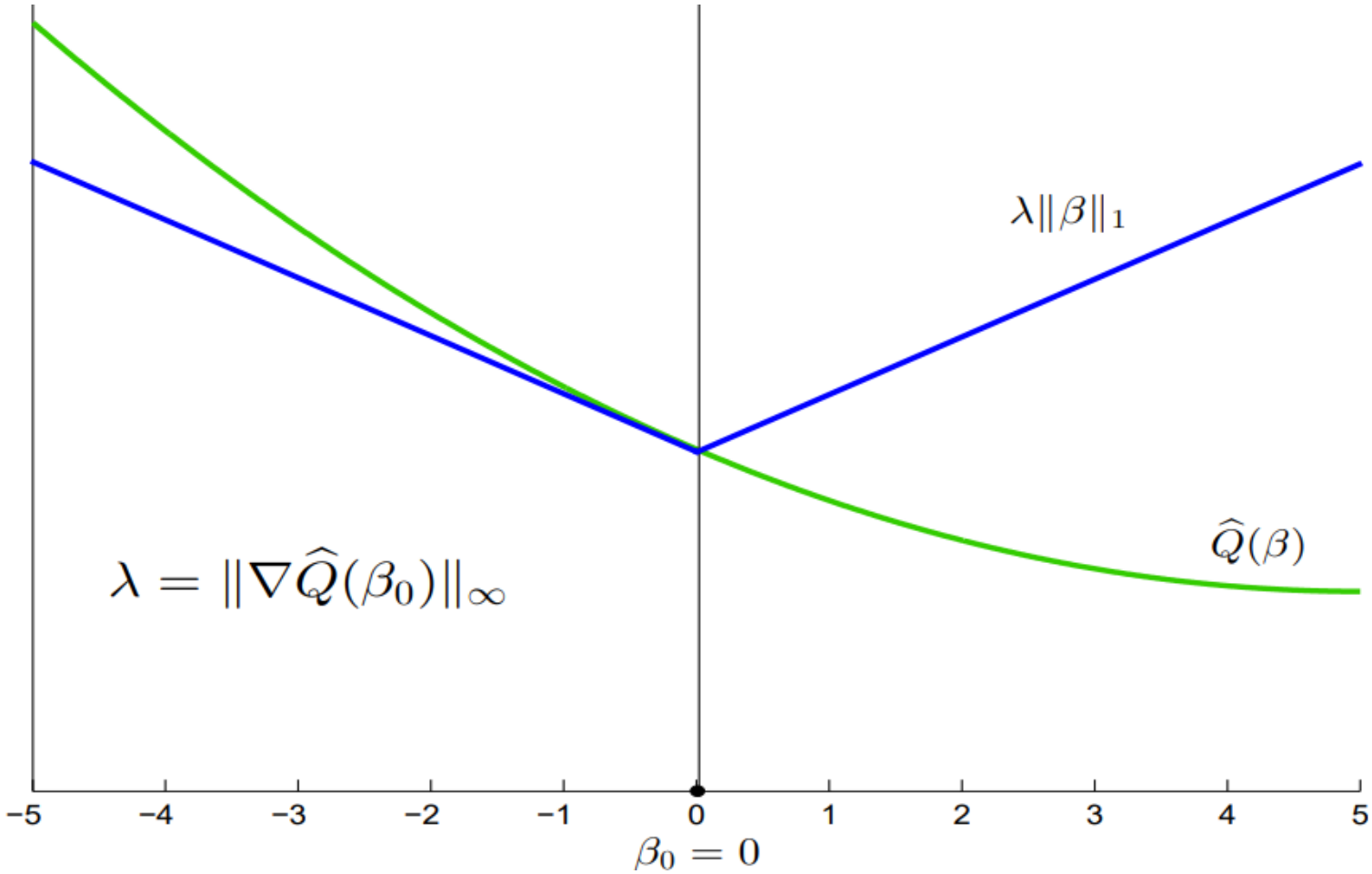
- Draw a 2D slice of the LS criterion $\hat{Q}(\beta)$



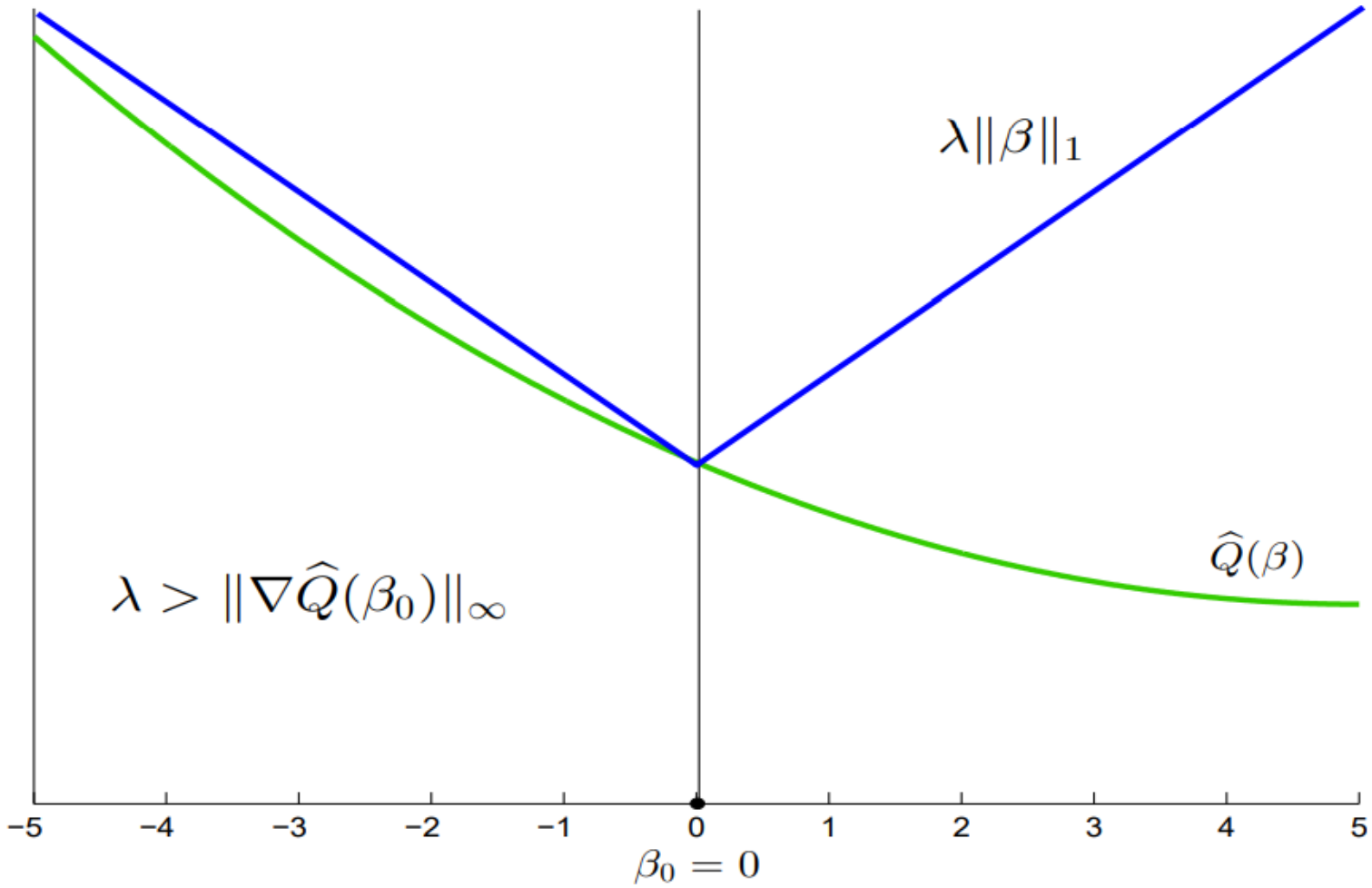
- Define the supporting hyperplane (2D) of the LS criterion at β_0



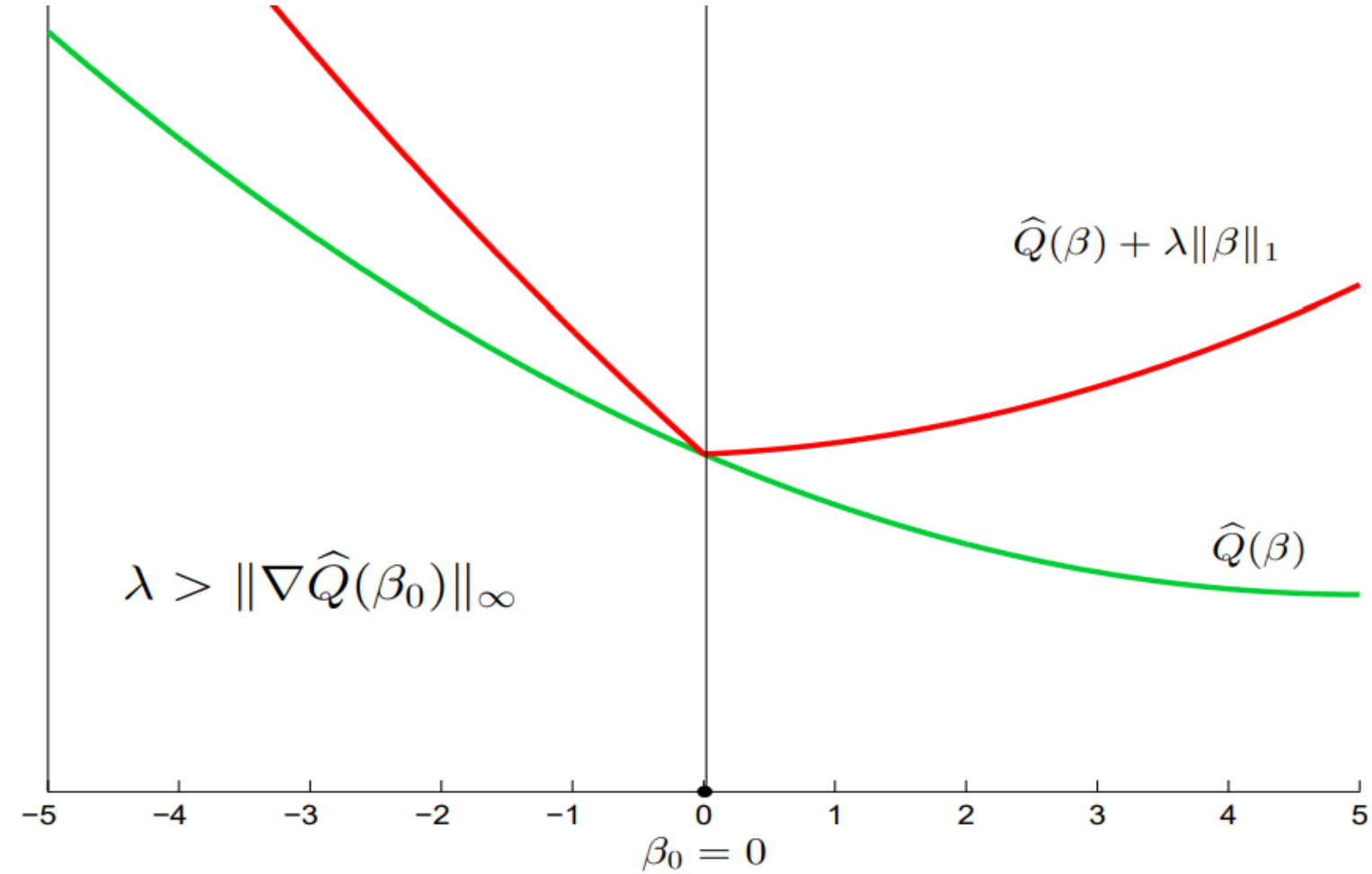
- Bend the hyperplane to define the penalty function
- Bended hyperplane at β_0 has function $\lambda\|\beta\|_1$ with slope λ defined as the maximum norm of subgradient of LS criterion



- Select λ to dominate the norm for all β_0 values
- Penalty function has to be steeper than supporting hyperplane



- Sum up the LS function and the penalty function (= constrained LS)
- $\hat{Q}(\beta) + \lambda\|\beta\|_1$ (in red) **is uniquely minimized at zero ($\beta_0 = 0$)**



LASSO: Heuristics

- Due to **kink** in the penalty, LASSO "zeros out" irrelevant regressors
- When $p > 1$, penalty function has a diamond-shaped geometry
- Increasing λ makes the penalty function "spikier"
  $\Rightarrow$ More likely to hit the LS criterion on coefficient axis
  $\Rightarrow$ Heavier model selection with more coefficients shrunk to zero
- Data-driven choice of λ s.t. $P(\lambda > \|\nabla \hat{Q}(\beta_0)\|_\infty) \to 1$
  (Belloni et al., Econometrica, 2012; Belloni and Chern., Ann. Statist, 2011)