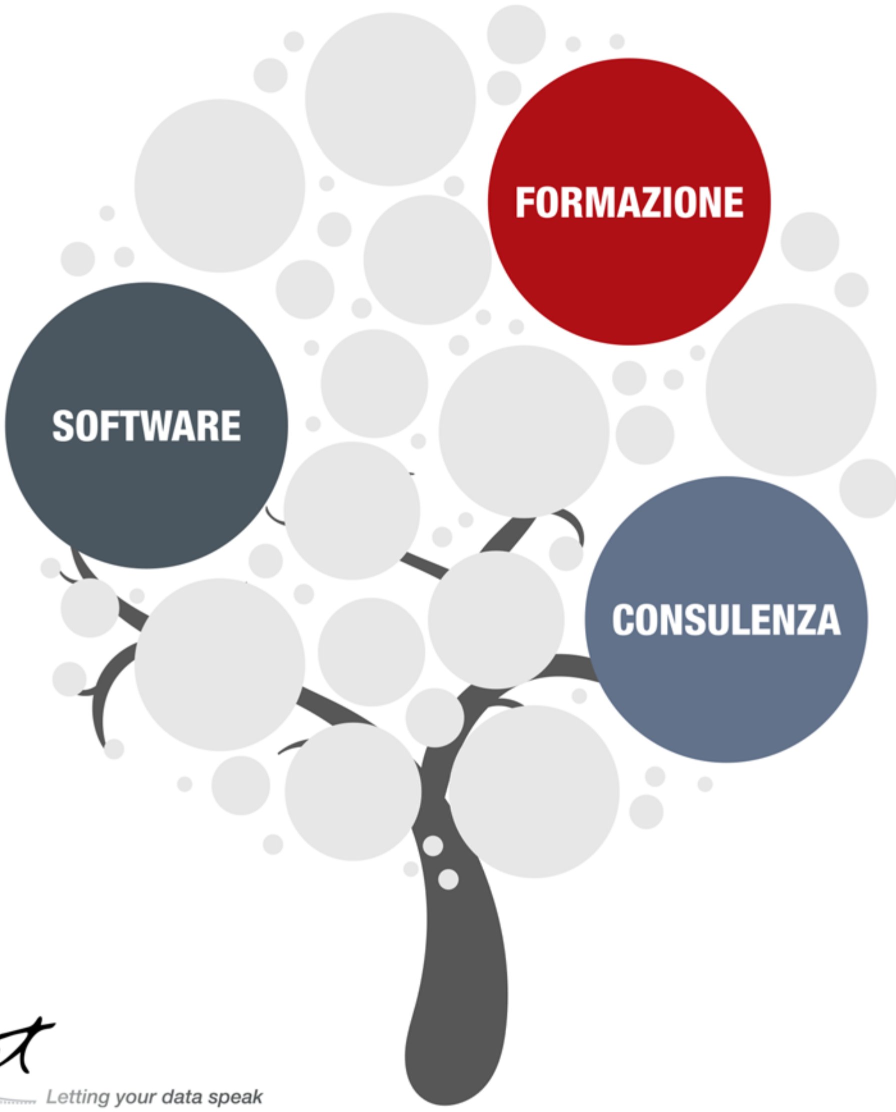
Estimation of a latent network via LASSO regression using Stata

Giovanni Cerulli IRCrES-CNR, Rome

XVI Italian Stata Users Group Meeting
Firenze, 26-27 September 2019

10.45 – 12.00 SESSION II COMMUNITY CONTRIBUTED, I



Letting your data speak

XVI Convegno Italiano degli Utenti di Stata
Firenze, 26 Settembre 2019

Estimation of a latent network via LASSO regression using Stata

Giovanni Cerulli
IRCrES-CNR

Consiglio Nazionale delle Ricerche
IRCrES
ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
INSTITUTE of RESEARCH on ECONOMIC SUSTAINABLE GROWTH

Motivation

- Estimating a **(latent) network** among *N* units characterized by *p* covariates **without any prior knowledge** about units' links

- This problem is **high-dimensional**, that is **p >> N** ==> the **LASSO** approach (*penalized regression*) is suitable for this purpose

- We implement a Stata routine to easily estimate the network using the **LASSOPACK** package in Stata

Setting the stage (with an example)

- Units ===> *N* Banks
- Features ===> set of *p* bank riskiness indicators
- No prior knowledge of the linkages
- Standard regression unsuited for high-dimensional data
- Lasso regression suitable
- Stata implementation

Setting

Starting dataset

	Risk1	Risk2	Risk3
Bank1	24	34	23
Bank2	45	76	76
Bank3	76	37	12
Bank4	25	87	87

Transposed dataset

	Bank1	Bank2	Bank3	Bank4	Bank5
Risk1	24	45	76	25	24
Risk2	34	76	37	87	34
Risk3	23	76	12	87	23

Naïve vs. LASSO networks

	Bank1	Bank2	Bank3	Bank4	Bank5
Risk1	24	45	76	25	24
Risk2	34	76	37	87	34
Risk3	23	76	12	87	23

- **Naïve network =======> BANKS' CORRELATION MATRIX**
  1. Univariate relationship (no interdependences)
  2. All cells are full (no netweok sparsness)

To overcome 1. and 2.

Number of rows N << Number of columns p

- **LASSO network =======> BANKS' HIGH-DIMENSIONAL REGRESSION MATRIX**
  1. Multivariate relationship (yes to «interdependences»)
  2. Some network relationships (yes to «network sparsness»)

Consiglio Nazionale delle Ricerche
IRCrES
ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
INSTITUTE of RESEARCH on ECONOMIC SUSTAINABLE GROWTH

The LASSO regression

Consiglio Nazionale delle Ricerche
IRCrES
ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
INSTITUTE of RESEARCH on ECONOMIC SUSTAINABLE GROWTH

Motivation for LASSO regression

Consider the standard linear model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

- The true specification is unknown
- Which regressors are "really" important?
- Including too many regressors leads to **overfitting**:
  good in-sample fit (high R^2), but bad *out-of-sample* prediction
- Including too few regressors leads to **omitted variable bias**

→ How to find the **correct model specification**?

Consiglio Nazionale delle Ricerche
IRCrES
ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
INSTITUTE of RESEARCH on ECONOMIC SUSTAINABLE GROWTH

Model selection (I)

Model selection is a general econometric issue, but becomes even more challenging when the data are **high-dimensional**:

Data are **high-dimensional** if p is close to or larger than n

- If $p > n$, the linear model is not identified
- If $p = n$, perfect fit. The model is meaningless
- If $p < n$ but large, overfitting is likely

Model selection (II)

Model selection

The standard approach for model selection in econometrics is (arguably) hypothesis testing.

- test biases in many procedures (wrong **test size** or **low test power**)
- if *p* is large, inference may be problematic
- excessive reliance on coefficients' statistical significance, at expenses of on their magnitude

Example: Cross-country regressions, where we have only small number of countries, but thousands of macro variables.

Consiglio Nazionale delle Ricerche
IRCrES
ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
INSTITUTE of RESEARCH on ECONOMIC SUSTAINABLE GROWTH

Model selection (III)

Consider again the linear model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

Only s out of p regressors belong to the model, i.e.:

$$s := \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\} \ll n$$

In other words: most of the true coefficients β_j are actually zero. But we don't know which ones are zeros and which ones aren't.

The LASSO solution (I)

The **LASSO** (Least Absolute Shrinkage and Selection Operator), minimizes this objective function:

β minimizing →

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

→ Penalization term based on L1 norm

The **parameter λ** is called **shrinkage penalty**, as it has the effect of *shrinking* some of the estimates to exactly zero

Consiglio Nazionale delle Ricerche
IRCrES
ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
INSTITUTE of RESEARCH on ECONOMIC SUSTAINABLE GROWTH

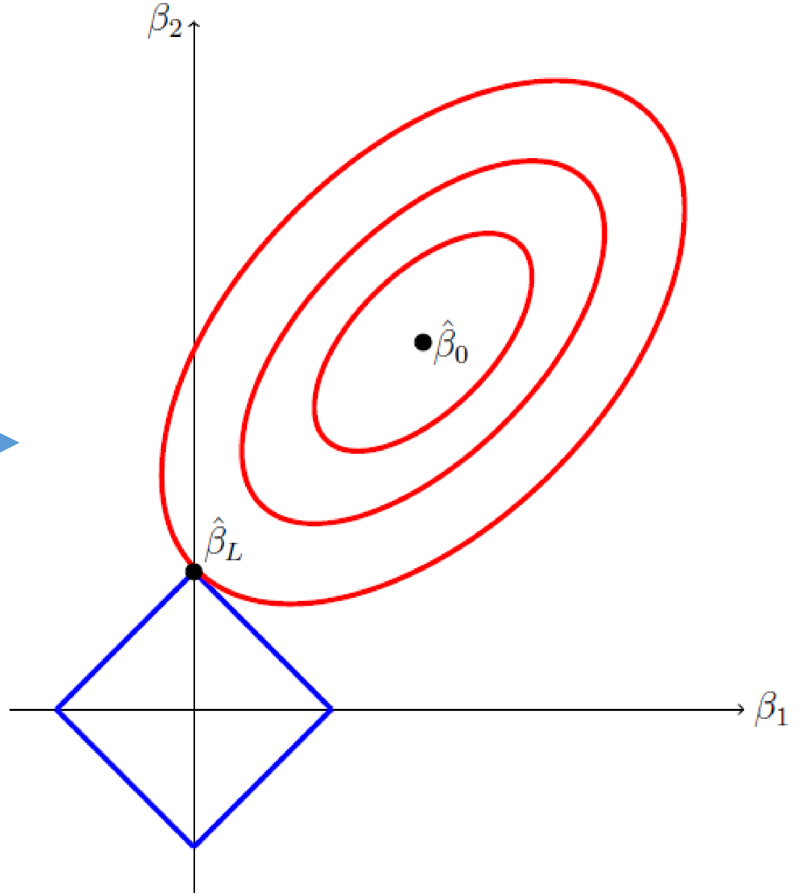
The LASSO solution (II)

The LASSO estimator can also be written as

$$\hat{\beta}_L = \arg\min \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| < \tau.$$

EXAMPLE:

- $p = 2$.
- Blue diamond is the constraint region $|\beta_1| + |\beta_2| < \tau$.
- $\hat{\beta}_0$ is the OLS estimate.
- $\hat{\beta}_L$ is the LASSO estimate.
- Red lines are RSS contour lines.
- $\hat{\beta}_{1,L} = 0$ implying that the LASSO omits regressor 1 from the model.



The RIDGE estimator

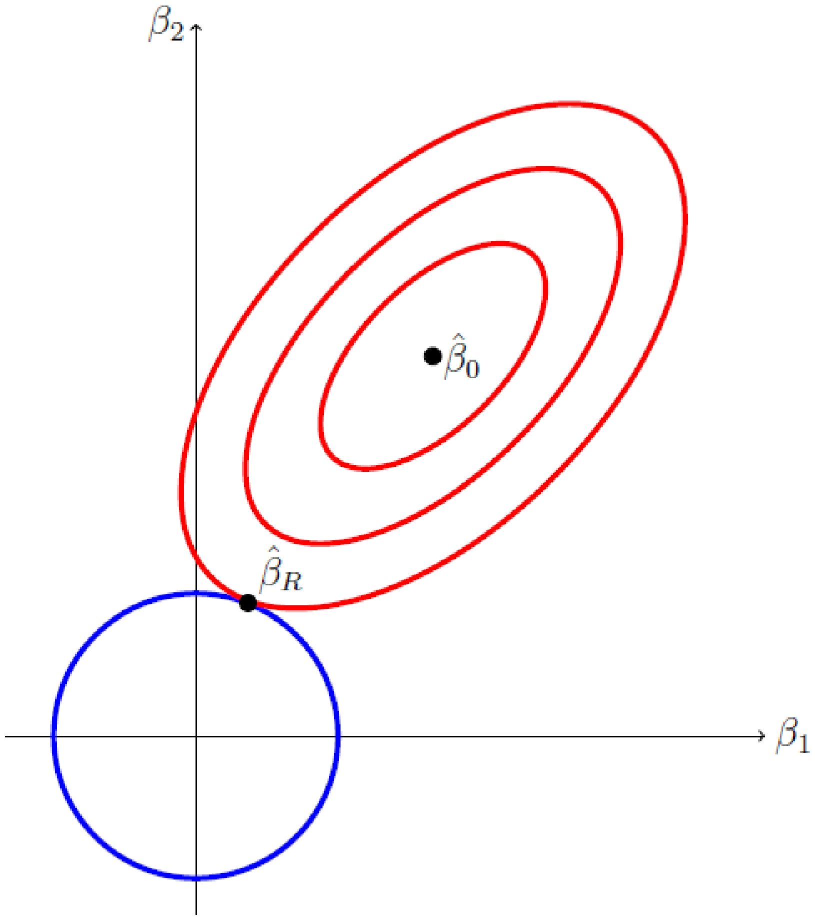
For comparison, the Ridge estimator is

$$\hat{\beta}_R = \arg\min \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \beta)^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 < \tau.$$

Penalization term based on L2-norm

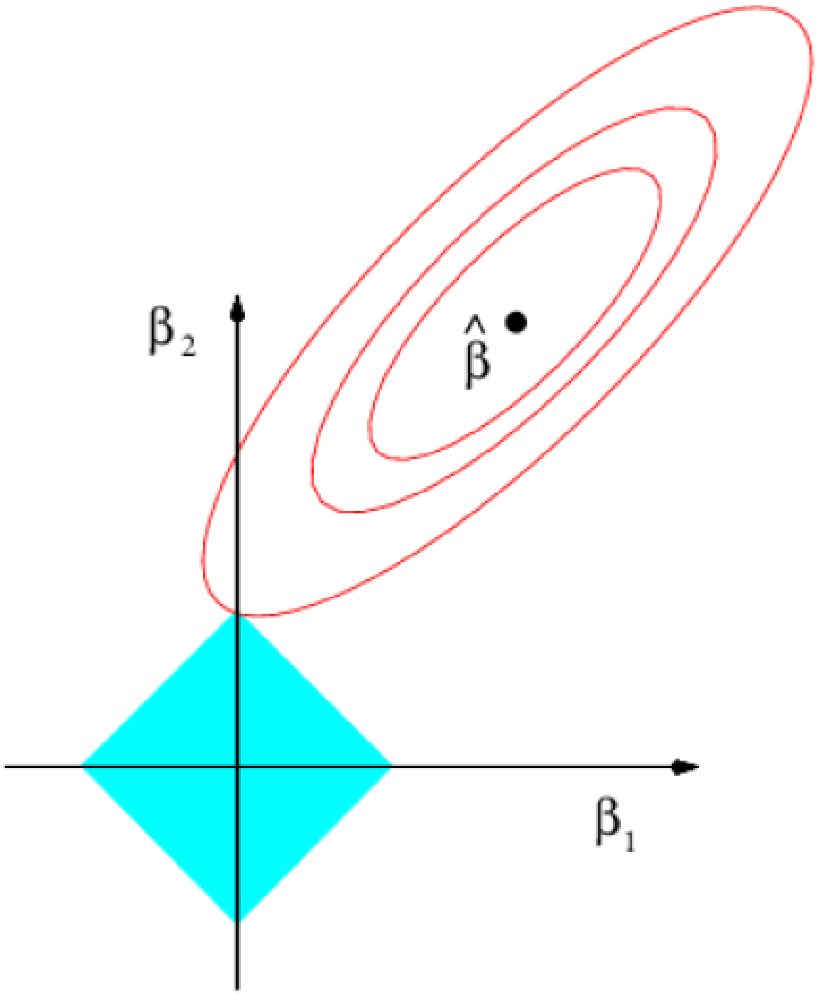
EXAMPLE:

- $p = 2$.
- Blue circle is the constraint region $\beta_1^2 + \beta_2^2 < \tau$.
- $\hat{\beta}_0$ is the OLS estimate.
- $\hat{\beta}_R$ is the Ridge estimate.
- Red lines are RSS contour lines.
- $\hat{\beta}_{1,L} \neq 0$ and $\hat{\beta}_{2,L} \neq 0$. Both regressors are included.

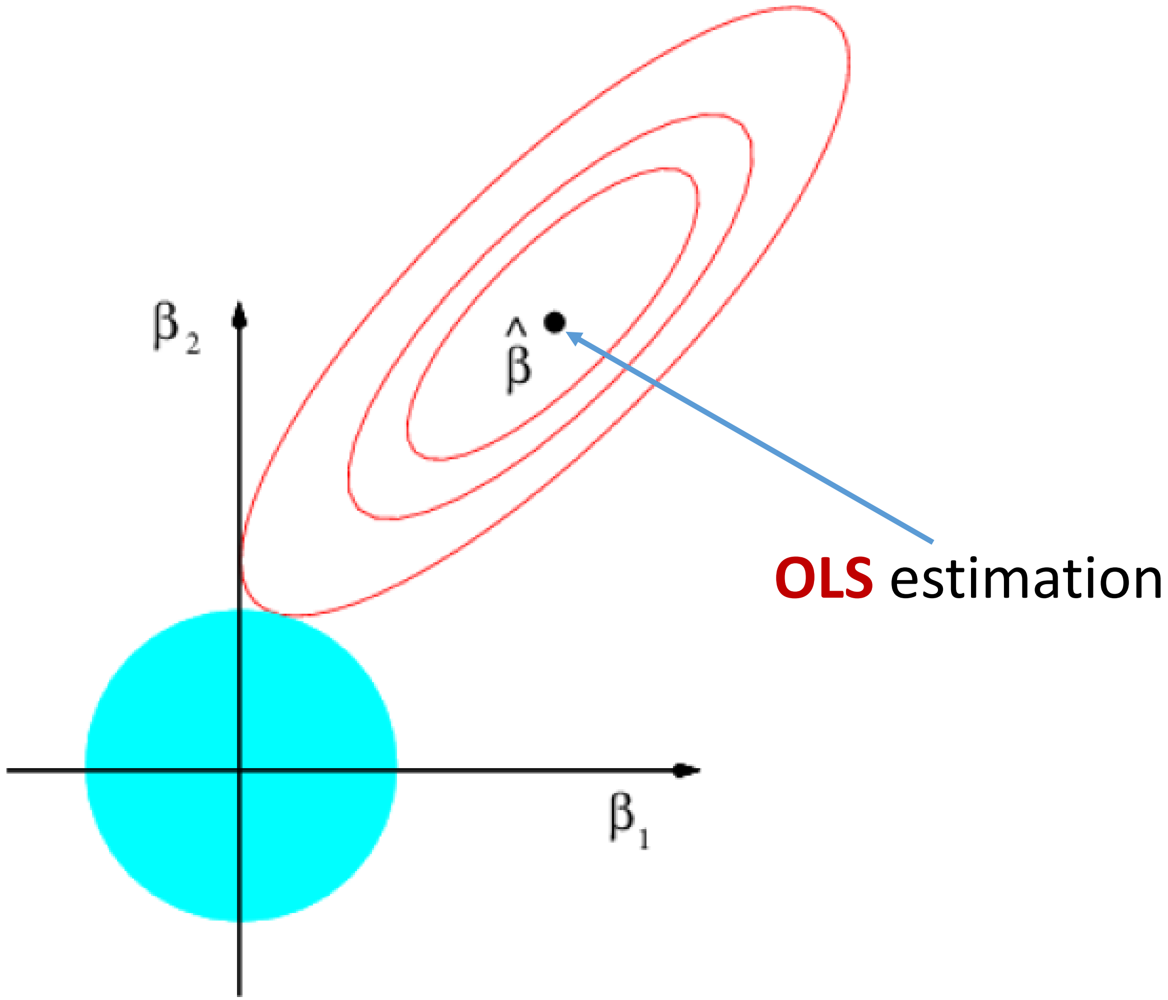


Consiglio Nazionale delle Ricerche
IRCrES
ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
INSTITUTE of RESEARCH on ECONOMIC SUSTAINABLE GROWTH

Comparison among OLS, LASSO, RIDGE estimates



LASSO estimation



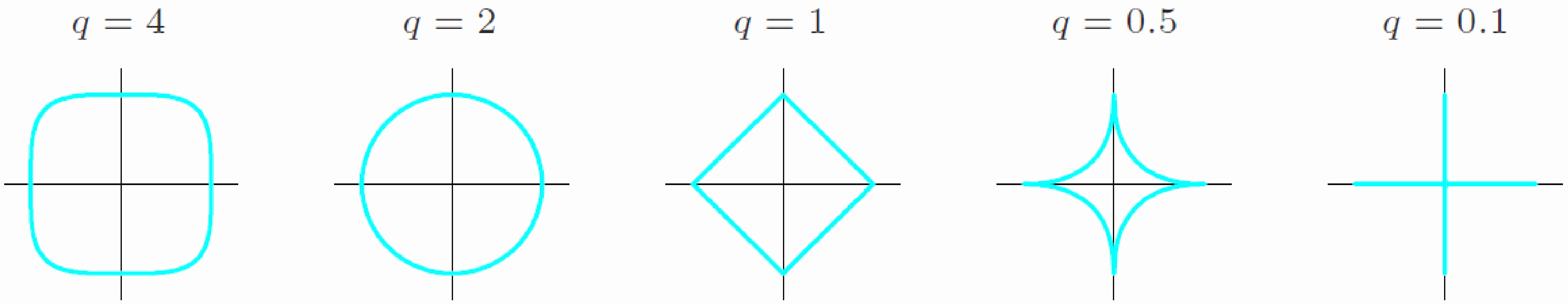
RIDGE estimation

Consiglio Nazionale delle Ricerche
IRCrES
ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
INSTITUTE of RESEARCH on ECONOMIC SUSTAINABLE GROWTH

Generalized Shrinkage Regression

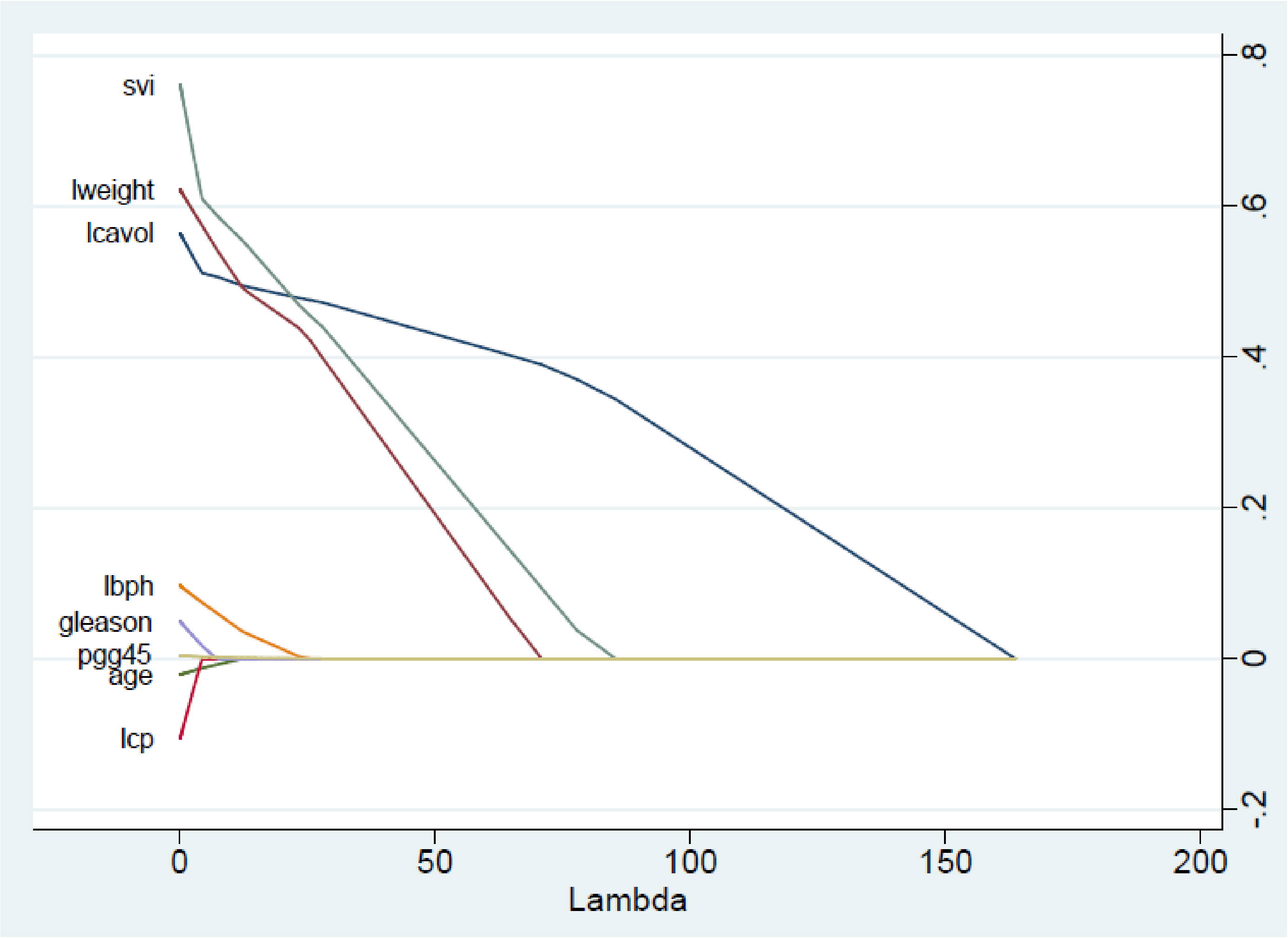
The Lasso and Ridge regression are specific cases of the **Generalized Shrinkage Regression** we obtain by varying q into this equation (thus obtaining **different coefficients contours**):

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \quad q \geq 0$$



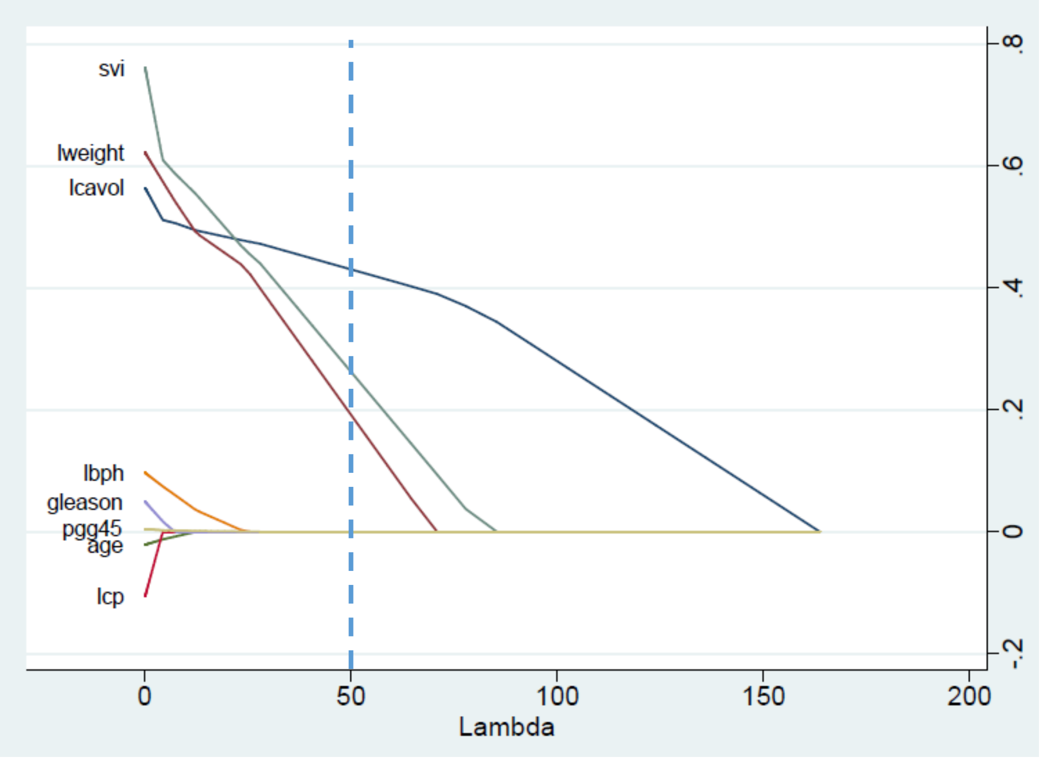
Contours of constant value of $\sum_j |\beta_j|^q$ *for given values of* q

The LASSO solution path

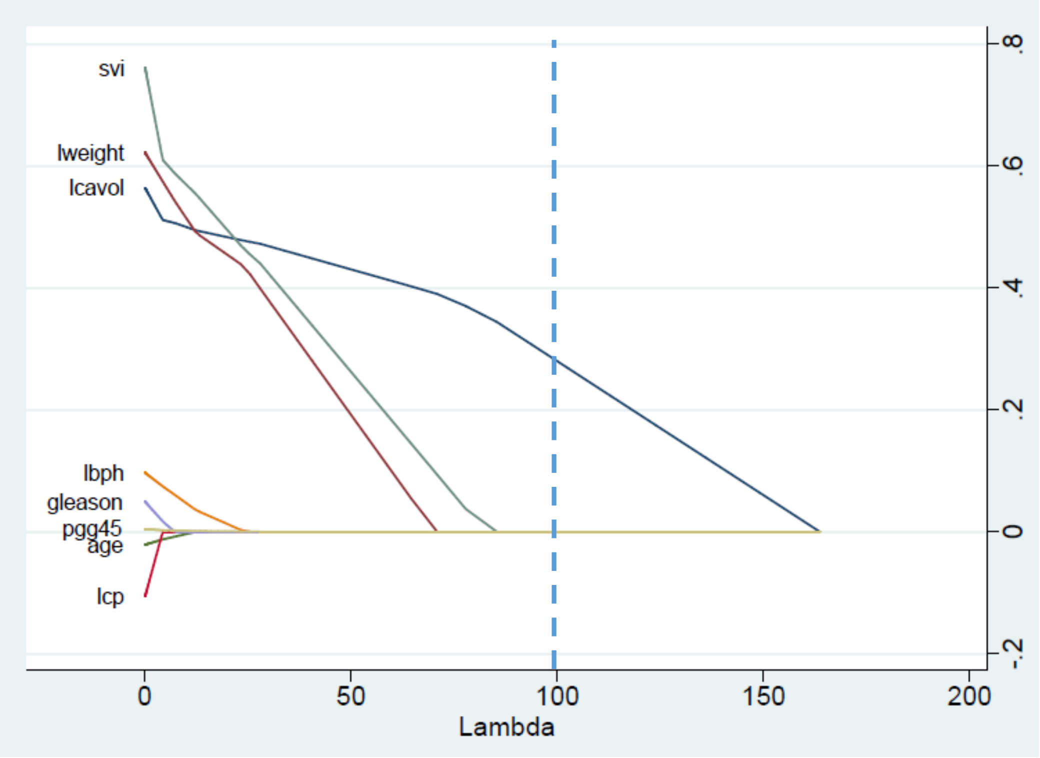


The LASSO coefficient path is a continuous and piecewise linear function of λ, with changes in slope where variables enter/leave the active set

The LASSO solution via different values of λ



For $\lambda = 50$, 3 out of 8 covariates are left



For $\lambda = 100$, 1 out of 8 covariates are left

NOTE:

- If $\lambda = 0$, all 8 covariates are left in the model $==>$ **OLS** solution
- If $\lambda \to \infty$, no variables are left in the model

Consiglio Nazionale delle Ricerche
IRCrES
ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
INSTITUTE of RESEARCH on ECONOMIC SUSTAINABLE GROWTH

How do we select λ ?

The optimal λ minimizes the out-of-sample (or test) mean prediction error

Three methods for doing this:

- **Data-driven**

  Based on **cross-validation**, i.e. re-sample the data and find λ that optimizes out-of-sample prediction. Implemented in Stata via `cvlasso`

- **Rigorous penalization**

  Belloni *et al.* (2012, *Econometrica*) develop theory and feasible algorithms for finding the optimal λ under heteroskedastic and non-Gaussian errors. Implemented in Stata via `rlasso`.

- **Information criteria**

  Select the value of λ that minimizes information criterion (AIC, AICc, BIC or EBIC). Implemented in Stata via `lasso2`

Network estimation via LASSO

- 4 Banks
- 3 Risk indicators

$$Bank_1 = A_{12} \cdot Bank_2 + A_{13} \cdot Bank_3 + A_{14} \cdot Bank_4 + e_1$$
$$Bank_2 = A_{21} \cdot Bank_1 + A_{23} \cdot Bank_3 + A_{24} \cdot Bank_4 + e_2$$
$$Bank_3 = A_{31} \cdot Bank_1 + A_{33} \cdot Bank_3 + A_{34} \cdot Bank_4 + e_3$$
$$Bank_4 = A_{41} \cdot Bank_1 + A_{42} \cdot Bank_2 + A_{43} \cdot Bank_3 + e_4$$

LASSO NETWORK

$$\begin{pmatrix} - & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & - & A_{34} \\ A_{41} & A_{42} & A_{43} & - \end{pmatrix}$$

Consiglio Nazionale delle Ricerche
IRCrES
ISTITUTO di RICERCA sulla CRESCITA ECONOMICA SOSTENIBILE
INSTITUTE of RESEARCH on ECONOMIC SUSTAINABLE GROWTH

Simulating a network estimation via `lassonet`

```
**************************************************************
* Simulation of a network estimation
**************************************************************
preserve
global N_nodes=15    // number of nodes
global p_features=3  // number of geatures by node
* Estimate the network
lassonet , nodes($N_nodes) features($p_features) seed(1010)
save network , replace
restore
use network , clear
**************************************************************
```

Simulating a network estimation via `lassonet`

`lassonet` ouptut

	names	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15
1	x1	0	.07	0	0	-.13	-.42	.25	0	0	-.27	0	0	0	.45	0
2	x2	.13	0	0	.48	0	-.12	.35	0	0	0	0	.31	0	0	0
3	x3	0	0	0	0	.02	0	0	0	0	.16	-.29	0	.1	-.06	-.36
4	x4	.04	.15	0	0	0	-.07	.11	0	0	0	0	.08	0	0	0
5	x5	-.36	0	.08	0	0	.33	0	0	0	.81	0	0	0	-.9	0
6	x6	-.08	-.13	0	0	.11	0	-.22	0	0	.04	0	0	0	-.1	0
7	x7	.03	.14	0	.52	0	-.14	0	0	0	0	0	.13	0	0	0
8	x8	0	0	.11	0	0	0	0	0	3.85	0	0	.17	2.25	0	0
9	x9	0	0	.13	.02	0	0	0	.1	0	0	-.11	.04	.12	0	-.09
10	x10	-.03	0	.26	0	.07	0	0	0	0	0	-.34	0	0	-.2	-.27
11	x11	0	0	-.15	0	0	0	0	0	0	-.22	0	0	0	.02	.16
12	x12	0	.38	0	1.31	0	-.19	.49	.04	.04	0	0	0	0	0	0
13	x13	0	0	.04	0	0	0	0	.1	.62	0	-.15	.01	0	0	-.03
14	x14	.12	0	-.05	0	-.09	-.12	0	0	0	-.29	.12	0	0	0	.06
15	x15	0	0	-.24	0	0	0	0	0	0	-.07	.31	0	-.02	.19	0

Conclusions

- **Machine Learning** is revolutionizing statistical applications
- The possibility to estimate a **network without knowing nodes' linkages** was unimaginable just few years ago
- We proposed a model and **Stata implementation** to this purpose via the `lassonet` command
- We foresee to shortly provide the community with a **Stata command** to use with any possible dataset