On Interpretability in Data Analytics

Emilio Carrizosa, IMUS - Instituto de Matemáticas de la Universidad de Sevilla

Nizhny Novgorod, 21st November 2020

Linear regression. Ordinary Least Squares

$$\min_{\beta} \sum_{i=1}^{m} \left(Y_i - \sum_{j=1}^{N} \beta_j X_{ij} \right)^2$$

- Convex and smooth objective, no constraints
- Necessary and sufficient optimality conditions: $\nabla = 0$
- Normal equations

Linear regression. Ordinary Least Squares

$$\min_{\beta} \sum_{i=1}^{m} \left(Y_i - \sum_{j=1}^{N} \beta_j X_{ij} \right)^2$$

- Convex and smooth objective, no constraints
- Necessary and sufficient optimality conditions: $\nabla = 0$
- Normal equations

Linear regression. Ordinary Least Squares

$$\min_{\beta} \sum_{i=1}^{m} \left(Y_i - \sum_{j=1}^{N} \beta_j X_{ij} \right)^2$$

- Convex and smooth objective, no constraints
- Necessary and sufficient optimality conditions: $\nabla = 0$
- Normal equations

Linear regression. Ordinary Least Squares

$$\min_{\beta} \sum_{i=1}^{m} \left(Y_i - \sum_{j=1}^{N} \beta_j X_{ij} \right)^2$$

- Convex and smooth objective, no constraints
- Necessary and sufficient optimality conditions: $\nabla = 0$
- Normal equations

http://statweb.stanford.edu/~tibs/lasso.html

The Lasso Page

L1-constrained fitting for statistics and data mining

$$\min_\beta \quad \sum_{i=1}^m \left(y_i - \beta^\top x_i\right)^2 \qquad \text{s.t.} \quad \|\beta\|_1 \leq C$$

$$\min_\beta \quad \sum_{i=1}^m \left(y_i - \beta^\top x_i\right)^2 + \lambda\|\beta\|_1$$

http://statweb.stanford.edu/~tibs/lasso.html

The Lasso Page

L1-constrained fitting for statistics and data mining

$$\begin{array}{ll} \min_\beta & \sum_{i=1}^m \left(y_i - \beta^\top x_i\right)^2 \\ \text{s.t.} & \|\beta\|_1 \leq C \end{array}$$

$$\min_\beta \quad \sum_{i=1}^m \left(y_i - \beta^\top x_i\right)^2 + \lambda\|\beta\|_1$$

http://statweb.stanford.edu/~tibs/lasso.html

The Lasso Page

L1-constrained fitting for statistics and data mining

$$\begin{array}{ll} \min_\beta & \sum_{i=1}^m \left(y_i - \beta^\top x_i\right)^2 \\ \text{s.t.} & \|\beta\|_1 \leq C \end{array}$$

$$\min_\beta \quad \sum_{i=1}^m \left(y_i - \beta^\top x_i\right)^2 + \lambda\|\beta\|_1$$

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]

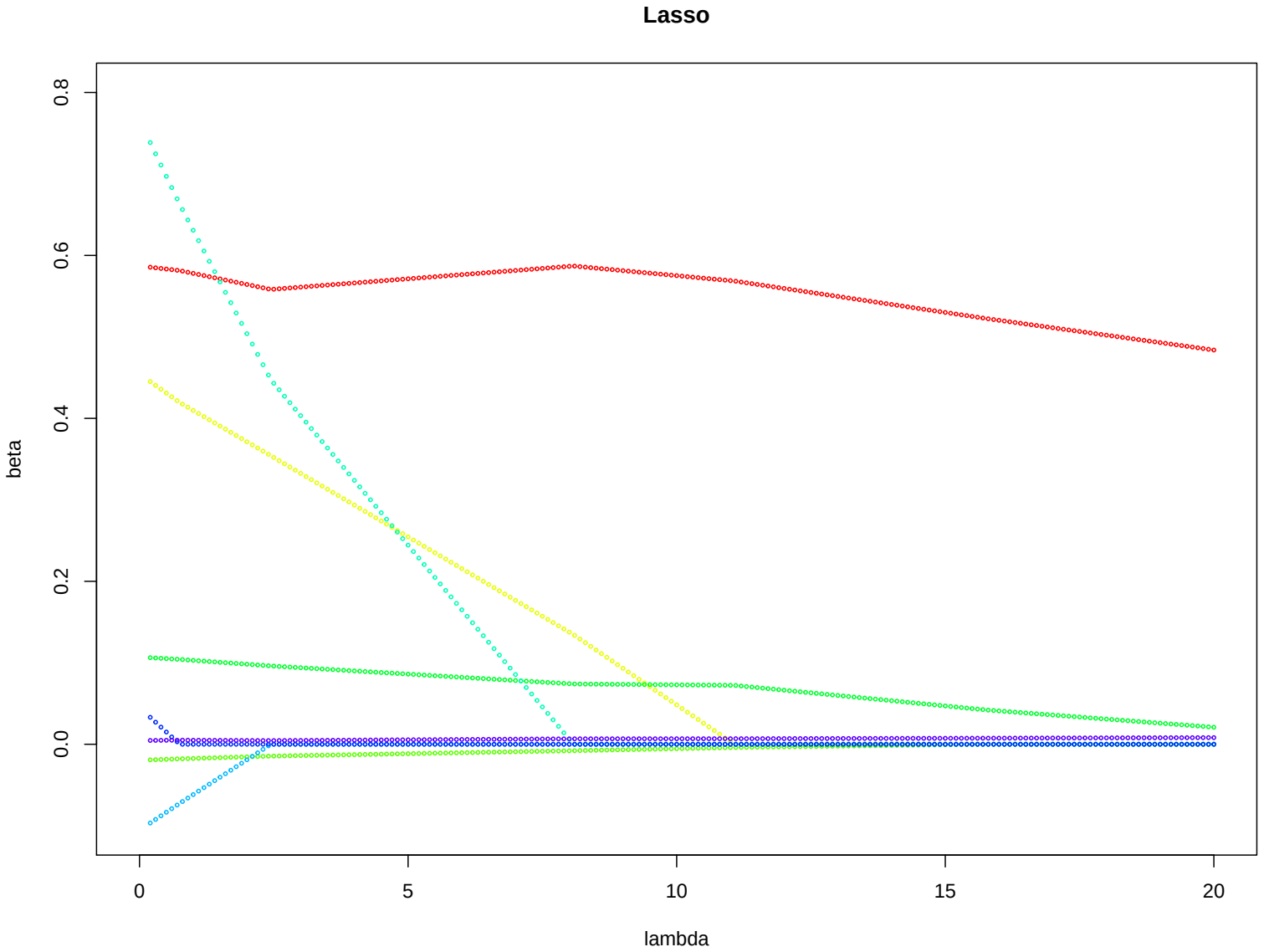
SUMMARY

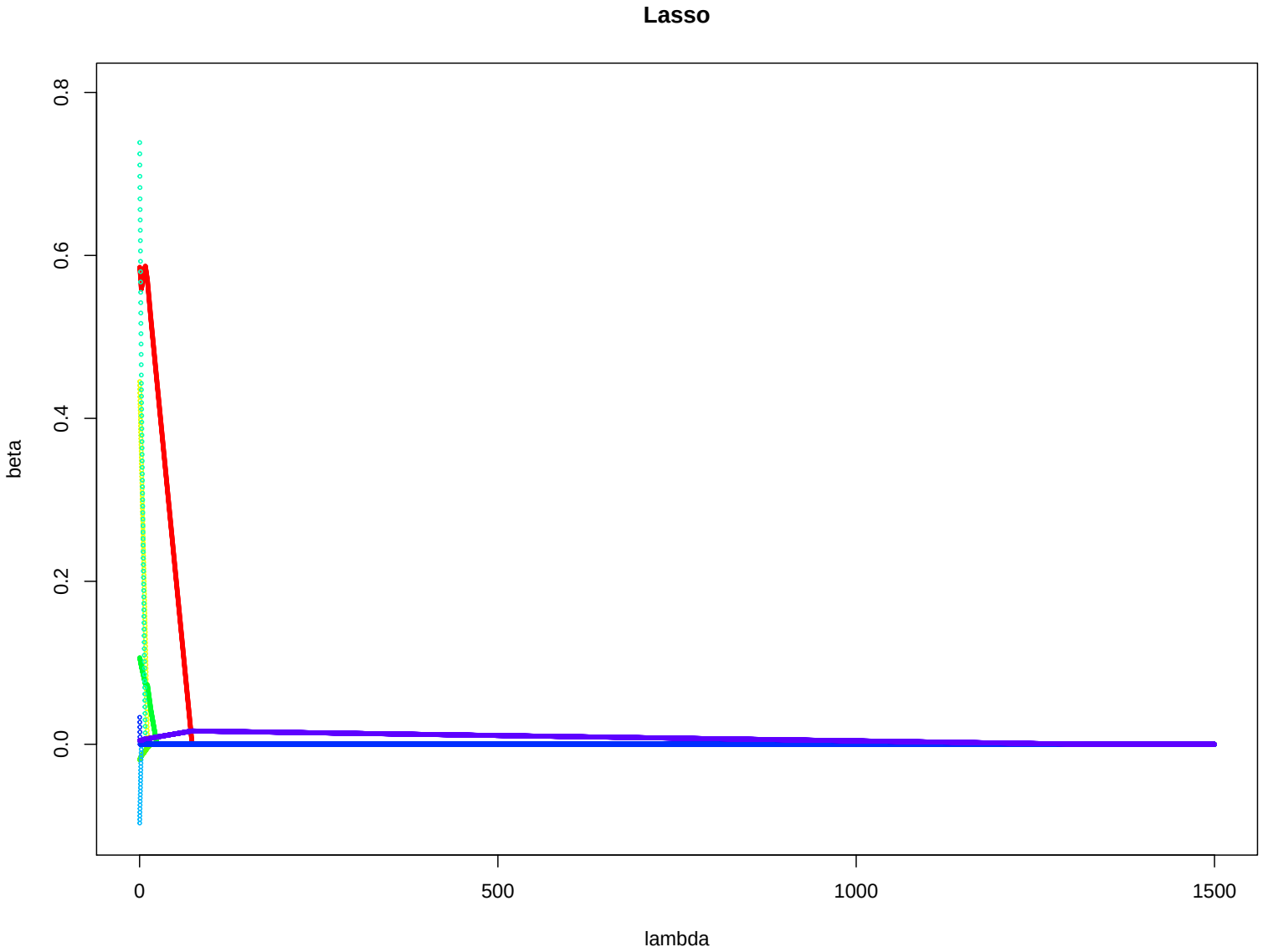
We propose a new method for estimation in linear models. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence 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

1. INTRODUCTION

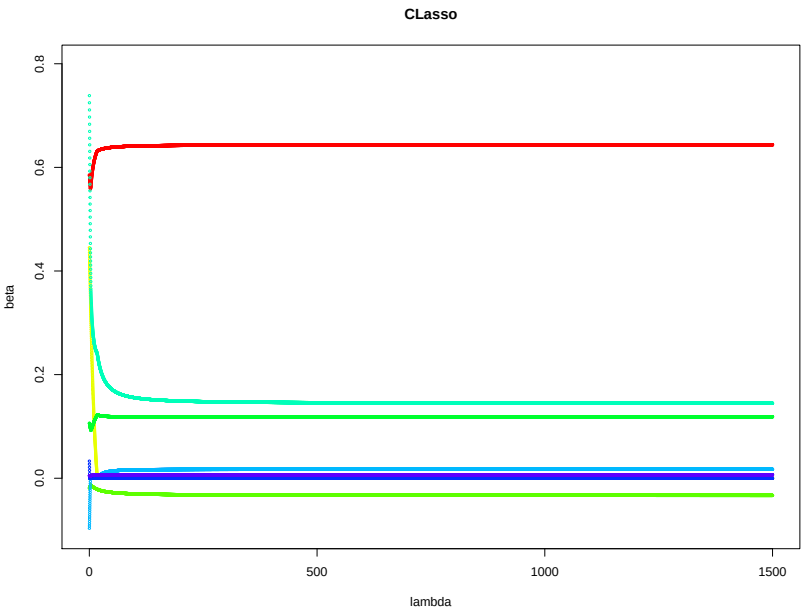
Consider the usual regression situation: we have data $(\mathbf{x}^i, y_i)$, $i = 1, 2, \ldots, N$, where $\mathbf{x}^i = (x_{i1}, \ldots, x_{ip})^T$ and y_i are the regressors and response for the ith observation. The ordinary least squares (OLS) estimates are obtained by minimizing the residual squared error. There are two reasons why the data analyst is often not satisfied with the OLS estimates. The first is *prediction accuracy*: the OLS estimates often have low bias but large variance; prediction accuracy can sometimes be improved by shrinking or setting to 0 some coefficients. By doing so we sacrifice a little bias to reduce the variance of the predicted values and hence may improve the overall prediction accuracy. The second reason is *interpretation*. With a large number of predictors, we often would like to determine a smaller subset that exhibits the strongest effects.

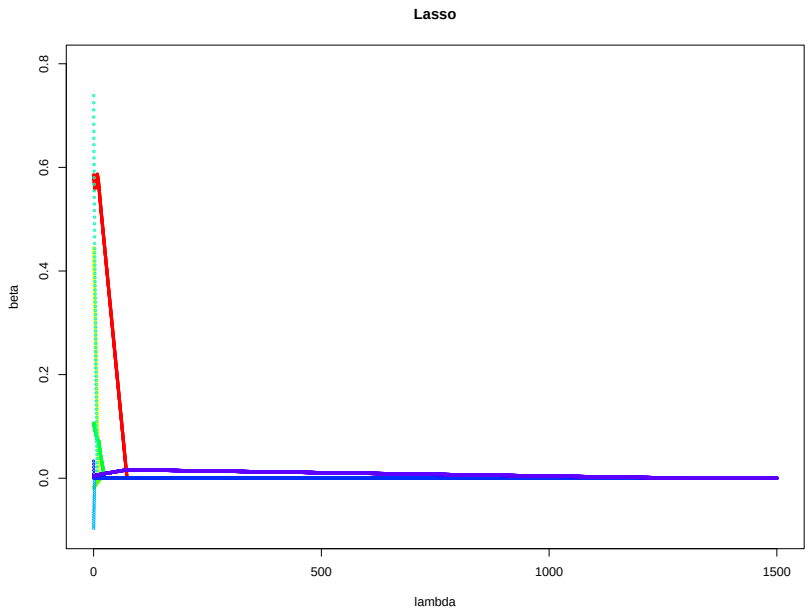

Lasso
beta
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
lambda


Lasso
beta
0.0
0.2
0.4
0.6
0.8
0
500
1000
1500
lambda

R Blanquero, E Carrizosa, P Ramírez-Cobo, MR Sillero-Denamiel
A cost-sensitive constrained Lasso
Advances in Data Analysis and Classification, 2020.

$$\begin{array}{ll} \min_{\beta} & \sum_{i=1}^{m} \left(y_i - \beta^\top x_i\right)^2 + \lambda \|\beta\|_1 \\ \text{s.t.} & \sum_{i \in \mathcal{I}_k} \left(y_i - \beta^\top x_i\right)^2 \leq f_k \qquad k = 1, 2, \ldots, K \end{array}$$


CLasso
beta
0.0
0.2
0.4
0.6
0.8
0
500
1000
1500
lambda

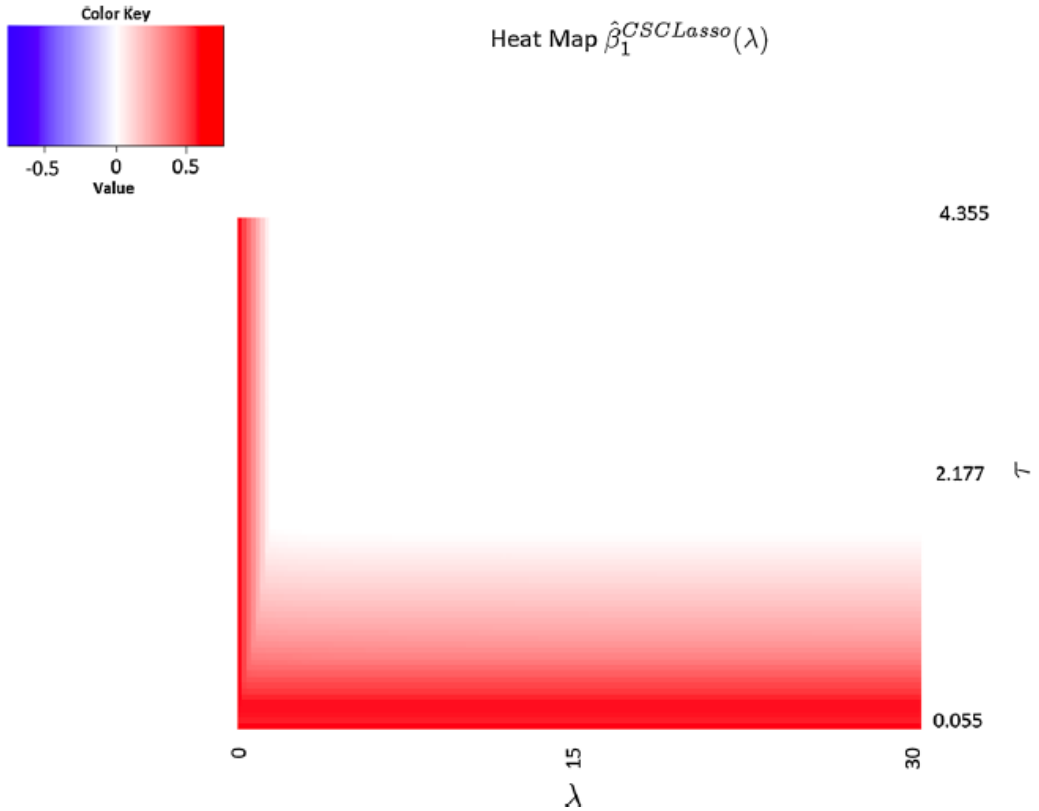

Lasso
beta
0.0
0.2
0.4
0.6
0.8
0
500
1000
1500
lambda

- $f_k^* = \min_\beta \sum_{i \in \mathcal{I}_k} \left(y_i - \beta^\top x_i\right)^2,\ k = 1, 2, \ldots, K$
-

$$\begin{array}{ll} \min_\beta & \sum_{i=1}^m \left(y_i - \beta^\top x_i\right)^2 + \lambda \|\beta\|_1 \\ \text{s.t.} & \sum_{i \in \mathcal{I}_k} \left(y_i - \beta^\top x_i\right)^2 \leq (1+\tau) f_k^* \qquad k = 1, 2, \ldots, K \end{array}$$

- $f_k^* = \min_\beta \sum_{i \in \mathcal{I}_k} \left(y_i - \beta^\top x_i\right)^2,\ k = 1, 2, \ldots, K$
-

$$\begin{array}{ll} \min_\beta & \sum_{i=1}^m \left(y_i - \beta^\top x_i\right)^2 + \lambda \|\beta\|_1 \\ \text{s.t.} & \sum_{i \in \mathcal{I}_k} \left(y_i - \beta^\top x_i\right)^2 \leq (1+\tau) f_k^* \qquad k = 1, 2, \ldots, K \end{array}$$



Making the model sparse using MINLO

$$\min_{\beta \in \mathcal{A}} \sum_{i=1}^{m} \left(Y_i - \sum_{j=1}^{N} \beta_j X_{ij} \right)^2$$

$\mathcal{A}$ modelling, among other things, which features are selected


Bertsimas and King, "An algorithmic approach to linear regression", *Operations Research*, 2015


Bertsimas, King and Mazumder, "Best subset selection via a modern optimization lens", *Annals of Statistics*, 2016


Carrizosa, Olivares-Nadal, Ramírez-Cobo, "A sparsity-controlled vector autoregressive model", *Biostatistics*, 2017.


Carrizosa, Olivares-Nadal, Ramírez-Cobo. "Integer constraints for enhancing interpretability in linear regression", *SORT, 2020.*

Making the model sparse using MINLO

$$\min_{\beta \in \mathcal{A}} \sum_{i=1}^{m} \left(Y_i - \sum_{j=1}^{N} \beta_j X_{ij} \right)^2$$

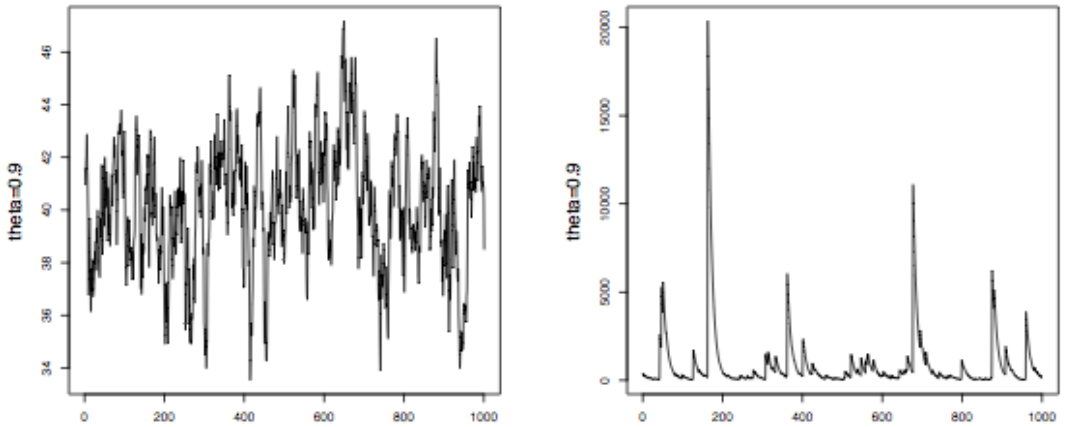
$\mathcal{A}$ modelling, among other things, which features are selected


Bertsimas and King, "An algorithmic approach to linear regression", *Operations Research*, 2015


Bertsimas, King and Mazumder, "Best subset selection via a modern optimization lens", *Annals of Statistics*, 2016


Carrizosa, Olivares-Nadal, Ramírez-Cobo, "A sparsity-controlled vector autoregressive model", *Biostatistics*, 2017.


Carrizosa, Olivares-Nadal, Ramírez-Cobo. "Integer constraints for enhancing interpretability in linear regression", *SORT, 2020.*

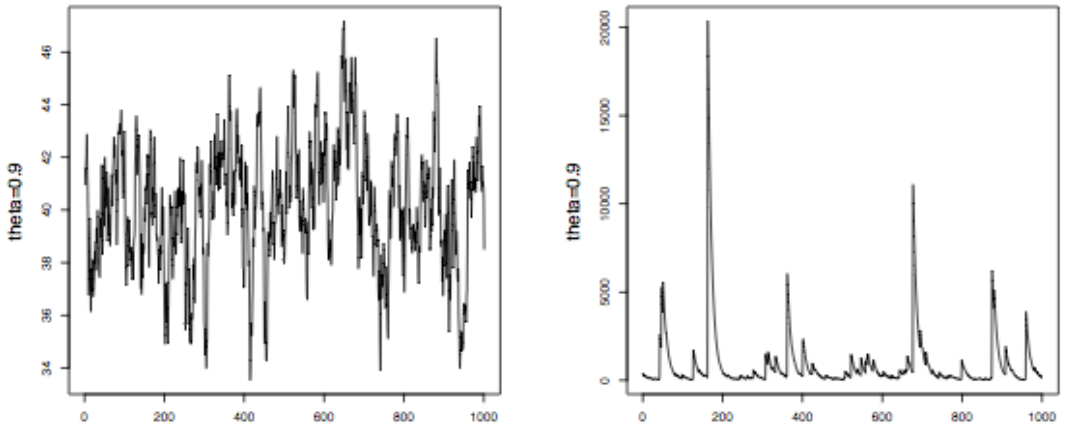


VAR(p), Vector AutoRegressive Model of order p

$$X_{i,t} = c^i + \sum_{j=1}^{N}\sum_{k=1}^{p} \beta_{jk}^i X_{j,t-k} + e_t^i \qquad t \geq 0 \qquad i = 1, 2, \ldots, N$$

OLS estimation of VAR(p) coefficients

$$\min_{c,A} \sum_{i=1}^{N}\sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N}\sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2$$



VAR(p), Vector AutoRegressive Model of order p

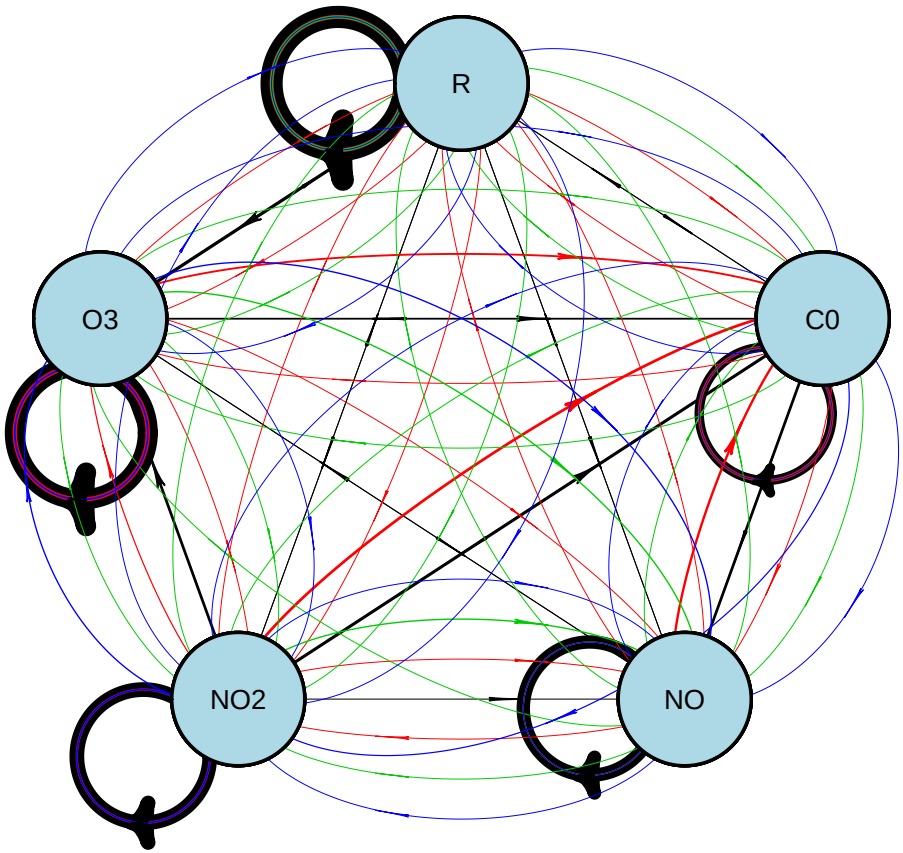
$$X_{i,t} = c^i + \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t-k} + e_t^i \qquad t \geq 0 \qquad i = 1, 2, \ldots, N$$

OLS estimation of VAR(p) coefficients

$$\min_{c,A} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2$$

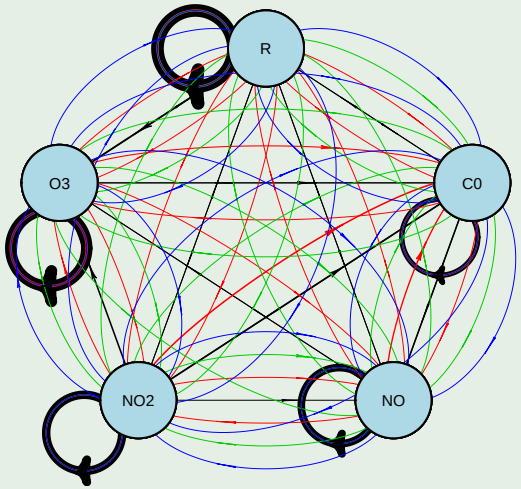
An example: Air pollution data set

MSE = 1



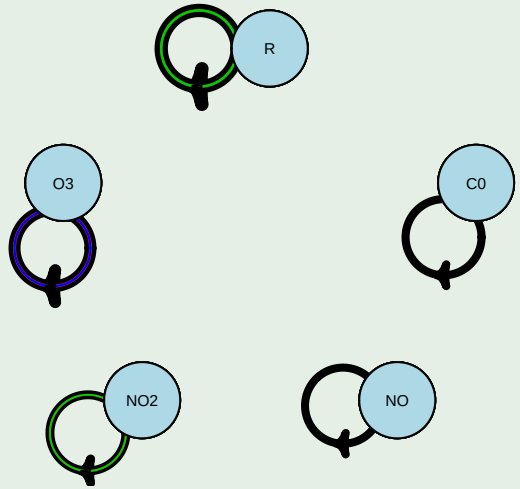
air pollution. OLS

MSE = 1



air pollution. SVAR

MSE = 1.08



OLS

$$\min_{c,A} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2$$

Group Lasso

$$\min_{c,A} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2 + \sum_{i=1}^{N} \lambda^i \left(\sum_{r=1}^{R} \|\boldsymbol{\beta}_{\boldsymbol{j(r)}}^{\boldsymbol{i}}\|_2 \right)$$

OLS

$$\min_{c,A} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2$$

Lasso

$$\min_{c,A} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2 + \sum_{i=1}^{N} \lambda^i \left(\sum_{j=1}^{N} \sum_{k=1}^{p} |\beta_{jk}^i| \right)$$

Group Lasso

$$\min_{c,A} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2 + \sum_{i=1}^{N} \lambda^i \left(\sum_{r=1}^{R} \|\beta_{j(r)}^i\|_2 \right)$$

OLS

$$\min_{c,A} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2$$

Lasso

$$\min_{c,A} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2 + \sum_{i=1}^{N} \lambda^i \left(\sum_{j=1}^{N} \sum_{k=1}^{p} |\beta_{jk}^i| \right)$$

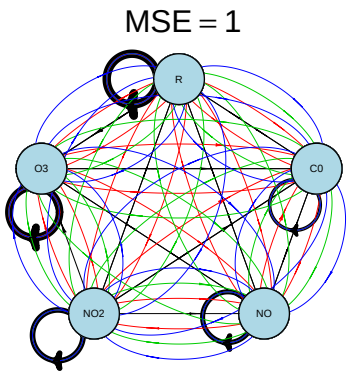
Group Lasso

$$\min_{c,A} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2 + \sum_{i=1}^{N} \lambda^i \left(\sum_{r=1}^{R} \|\boldsymbol{\beta}_{\boldsymbol{j(r)}}^{\boldsymbol{i}}\|_2 \right)$$

SC-VAR (VAR with Sparsity Controlled)

Instead of playing with the tuning parameters λ^i ...

- Identify the many dimensions of sparsity
- Find the best solution satisfying the sparsity requirements



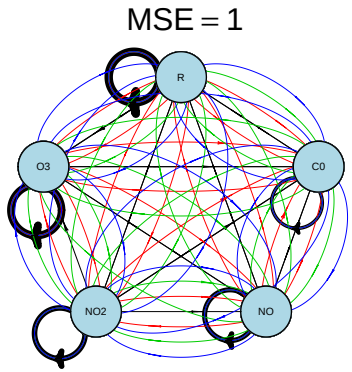
The (many) dimensions of sparsity

- Number of arrows in the graph ($\leq V_A$)
- Number of arrows pointing to feature i ($\leq V_T^i$)
- Number of features pointing to feature i ($\leq V_S^i$)
- Number of lags back allowed for variable i ($\leq V_{S\beta}^i$)
- Minimum size of the coefficients ($\geq \epsilon_j^i$)

SC-VAR (VAR with Sparsity Controlled)

Instead of playing with the tuning parameters λ^i ...

- Identify the many dimensions of sparsity
- Find the best solution satisfying the sparsity requirements



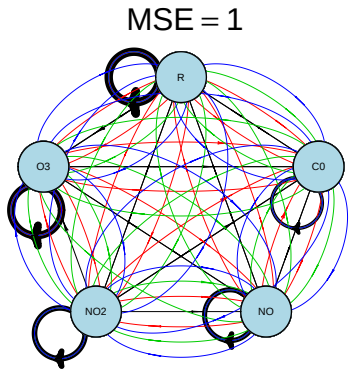
The (many) dimensions of sparsity

- Number of arrows in the graph ($\leq V_A$)
- Number of arrows pointing to feature i ($\leq V_T^i$)
- Number of features pointing to feature i ($\leq V_S^i$)
- Number of lags back allowed for variable i ($\leq V_{S\beta}^i$)
- Minimum size of the coefficients ($\geq \epsilon_j^i$)

SC-VAR (VAR with Sparsity Controlled)

Instead of playing with the tuning parameters λ^i ...

- Identify the many dimensions of sparsity
- Find the best solution satisfying the sparsity requirements



The (many) dimensions of sparsity

- Number of arrows in the graph ($\leq V_A$)
- Number of arrows pointing to feature i ($\leq V_T^i$)
- Number of features pointing to feature i ($\leq V_S^i$)
- Number of lags back allowed for variable i ($\leq V_{S\beta}^i$)
- Minimum size of the coefficients ($\geq \epsilon_j^i$)

Objective function: OLS

$$\min_{c,A,\Delta,\Gamma,} \sum_{i=1}^{N} \sum_{t=p}^{T} \left(X_{i,t+1} - c^i - \sum_{j=1}^{N} \sum_{k=1}^{p} \beta_{jk}^i X_{j,t+1-k} \right)^2$$

Variables

$$\beta_{jk}^i \neq 0 \text{ allowed?} \rightarrow \delta_{jk}^i = \begin{cases} 1 & \text{if } \beta_{jk}^i \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{any } \beta_{jk}^i \neq 0 \text{ allowed ? } (i,j : \text{fixed}) \rightarrow \gamma_j^i = \begin{cases} 1 & \text{if } \beta_{jk}^i \neq 0 \text{ for some } k \\ 0 & \text{otherwise} \end{cases}$$

$$\delta_{jk}^i \leq \gamma_j^i \qquad \forall k \in K, j, i \in I$$

SC-VAR: constraints

Number of arrows in the graph (V_A)

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{p}\delta_{jk}^{i} \leq V_A$$

SC-VAR: constraints

Number of arrows pointing to feature i (V_T^i)

$$\sum_{j=1}^{N}\sum_{k=1}^{p}\delta_{jk}^{i} \leq V_T^i \qquad \forall i \in I$$

SC-VAR: constraints

Number of features pointing to feature i (V_S^i)

$$\sum_{j=1}^{N} \gamma_j^i \leq V_S^i \qquad \forall i \in I$$

SC-VAR: constraints

Number of lags back allowed for variable i ($V^i_{S\beta}$)

$$\sum_{k=1}^{p} \delta^i_{jk} \leq V^i_{S\beta} \qquad \forall j, i \in I$$

SC-VAR: constraints

Minimum size of the coefficients (ϵ_j^i)

$$M\delta_{jk}^i \geq |\beta_{jk}^i| \geq \epsilon_j^i \delta_{jk}^i \qquad \forall k \in K, j, i \in I$$

SC-VAR: constraints (All together now)

$$\delta^i_{jk} \leq \gamma^i_j \qquad \forall k \in K, j, i \in I$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{p} \delta^i_{jk} \leq V_A$$

$$\sum_{j=1}^{N}\sum_{k=1}^{p} \delta^i_{jk} \leq V^i_T \qquad \forall i \in I$$

$$\sum_{j=1}^{N} \gamma^i_j \leq V^i_S \qquad \forall i \in I$$

$$\sum_{k=1}^{p} \delta^i_{jk} \leq V^i_{S\beta} \qquad \forall j, i \in I$$

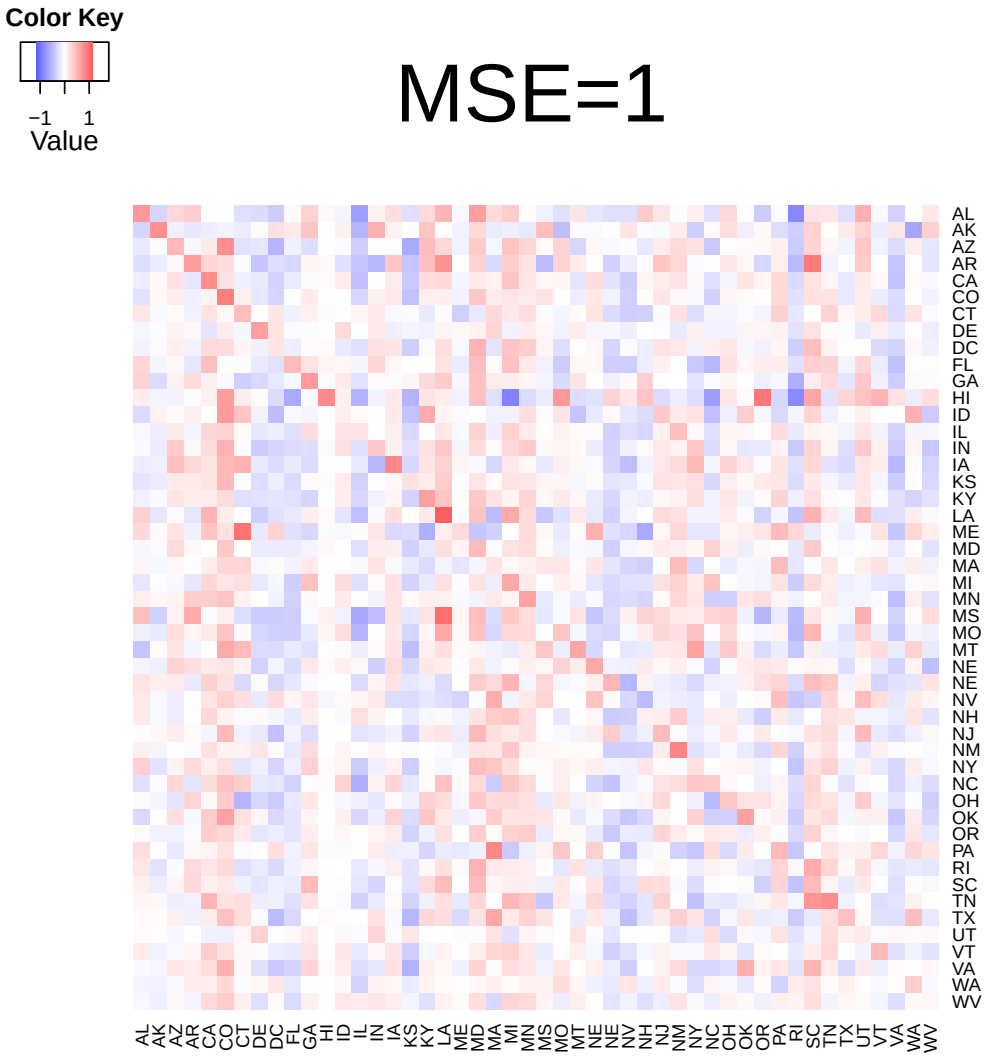
$$\beta^i_{jk} \geq \epsilon^i_j \nu^{i+}_{jk} - (1 - \nu^{i+}_{jk})M \qquad \forall k \in K, j, i \in I$$

$$\beta^i_{jk} \leq -\epsilon^i_j \nu^{i-}_{jk} + (1 - \nu^{i-}_{jk})M \qquad \forall k \in K, j, i \in I$$

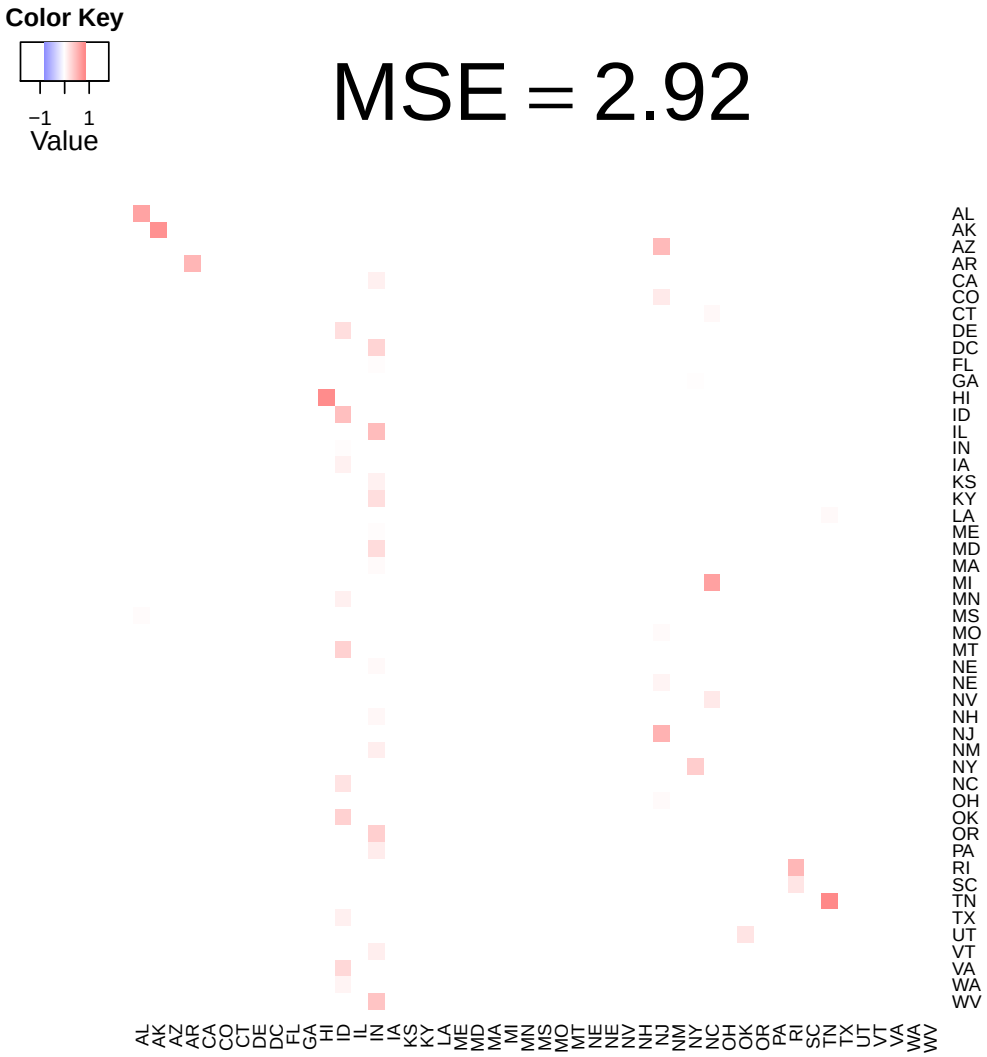
$$\delta^i_{jk} \leq \nu^{i+}_{jk} + \nu^{i-}_{jk} \leq 1 \qquad \forall k \in K, j, i \in I$$

$$\nu^{i+}_{jk}, \nu^{i-}_{jk} \in \{0, 1\} \qquad \forall k \in K, j, i \in I$$

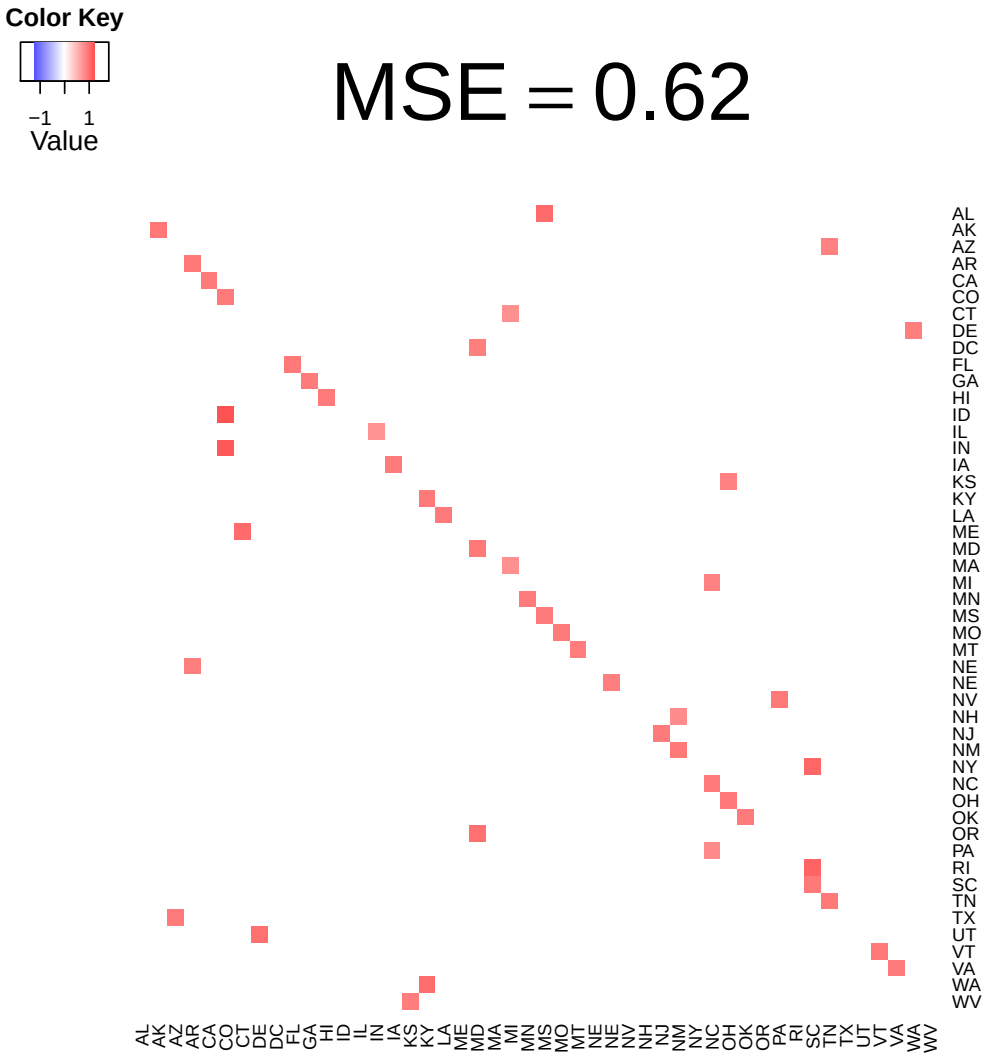
Google flu data set. VAR



Google flu data set. Lasso ($V_T = 1$)



Google flu data set. SC-VAR ($V_T = 1$, $\epsilon = 0.2$)



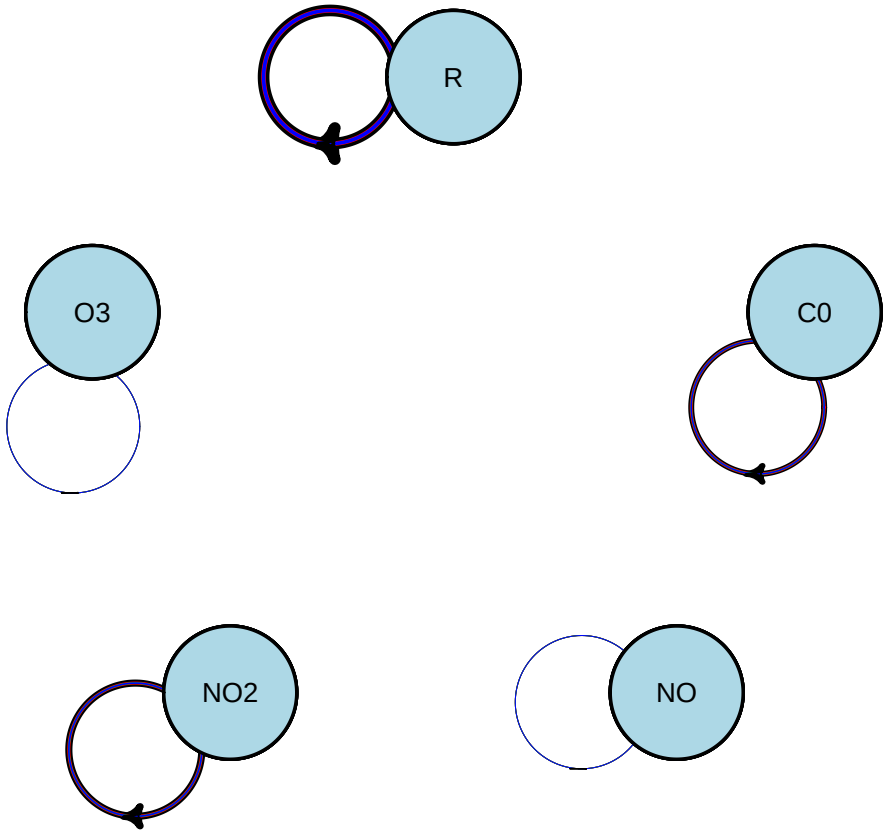
SC-VAR

SC-VAR: MINLP, with

- Quadratic convex objective function
- Linear constraints
- Some binary variables
- Solvable to optimality with benchmark solvers; good solutions with short time limit

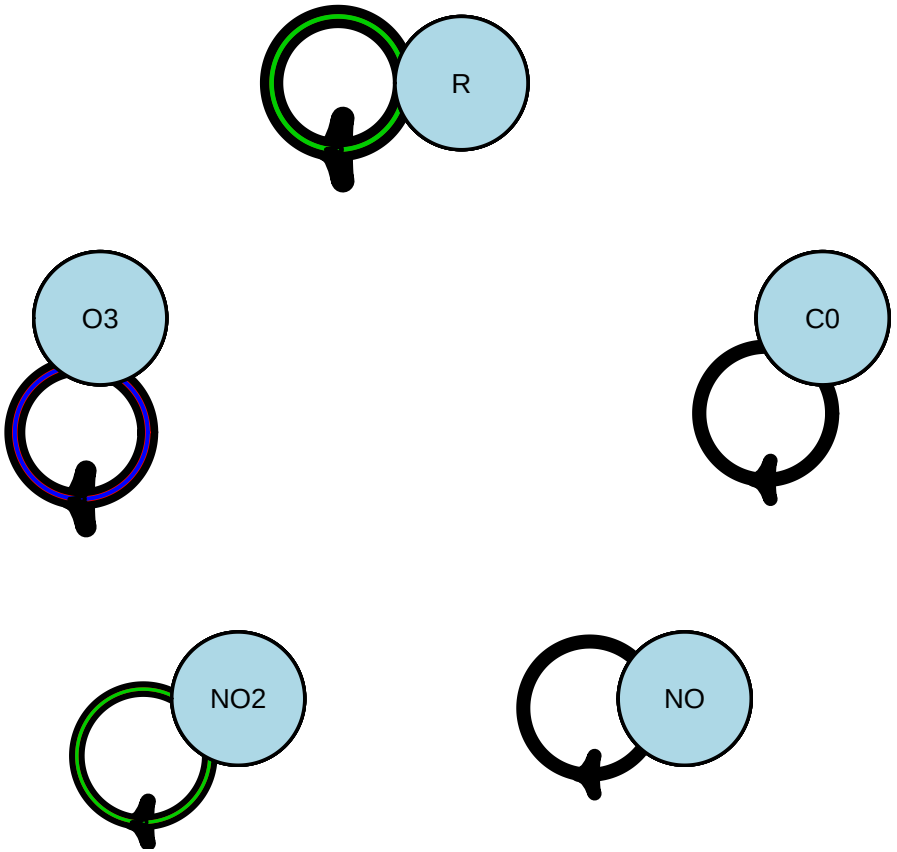
Air pollution data set. Group Lasso (1 causal feature)

$$\text{MSE} = 2.91$$



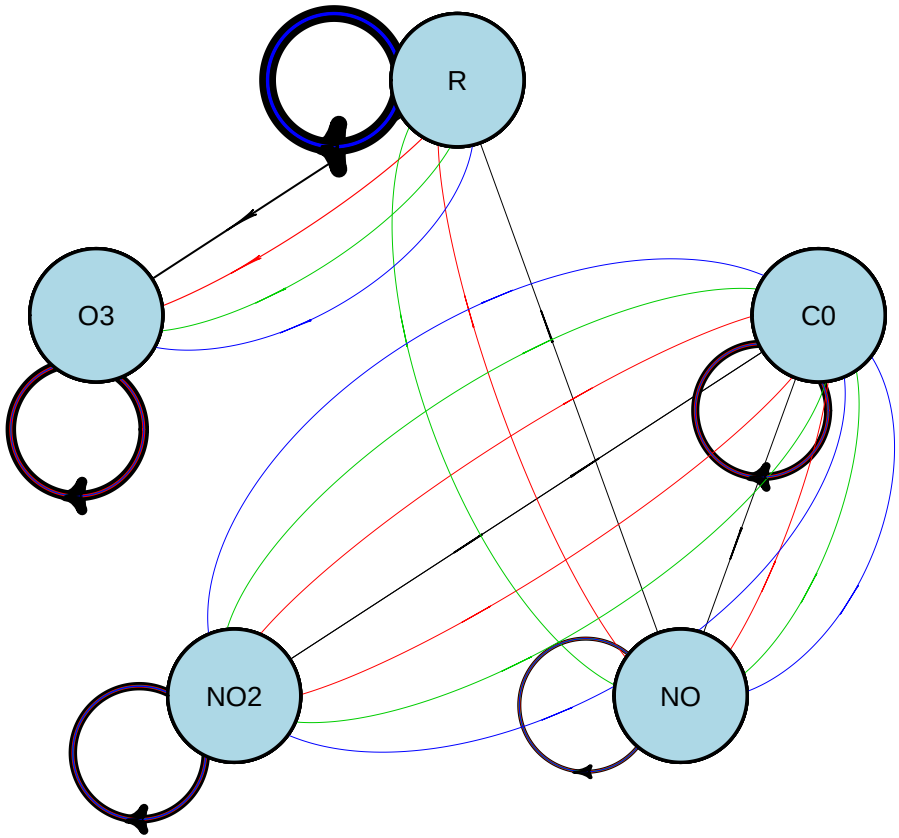
Air pollution data set. SC-VAR (1 causal feature)

$$\mathrm{MSE} = 1.08$$



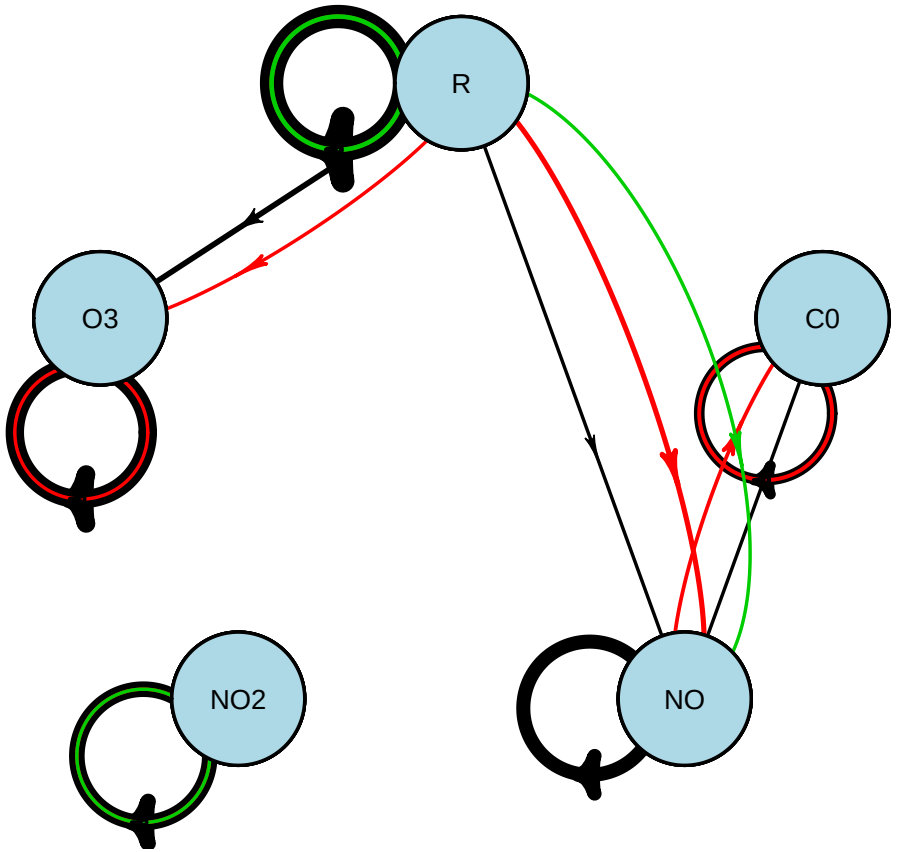
Air pollution data set. Group Lasso (2 causal features)

$$\text{MSE} = 1.4$$

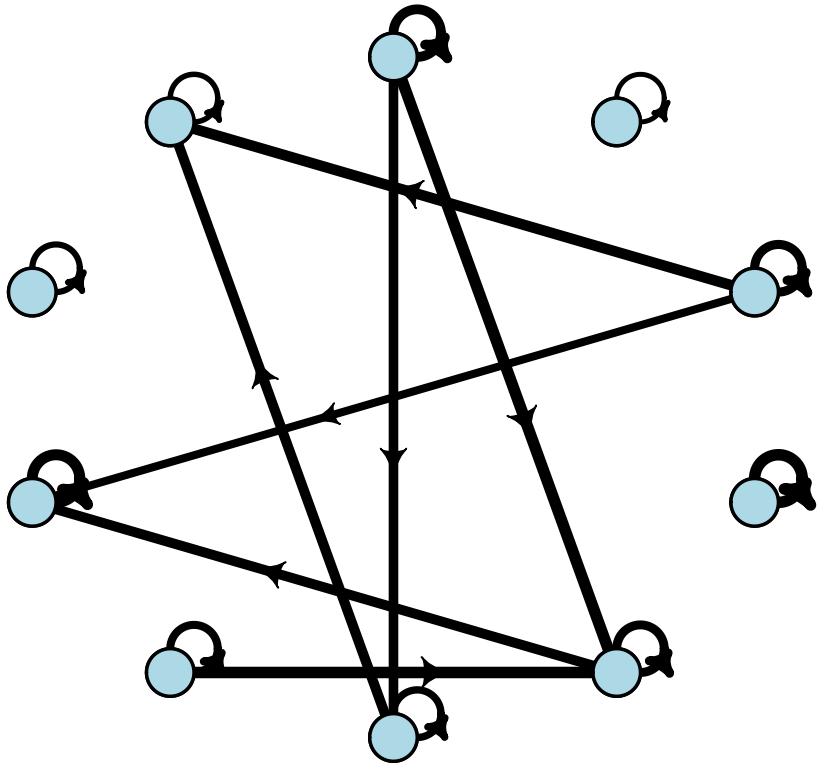


Air pollution data set. SC-VAR (2 causal features)

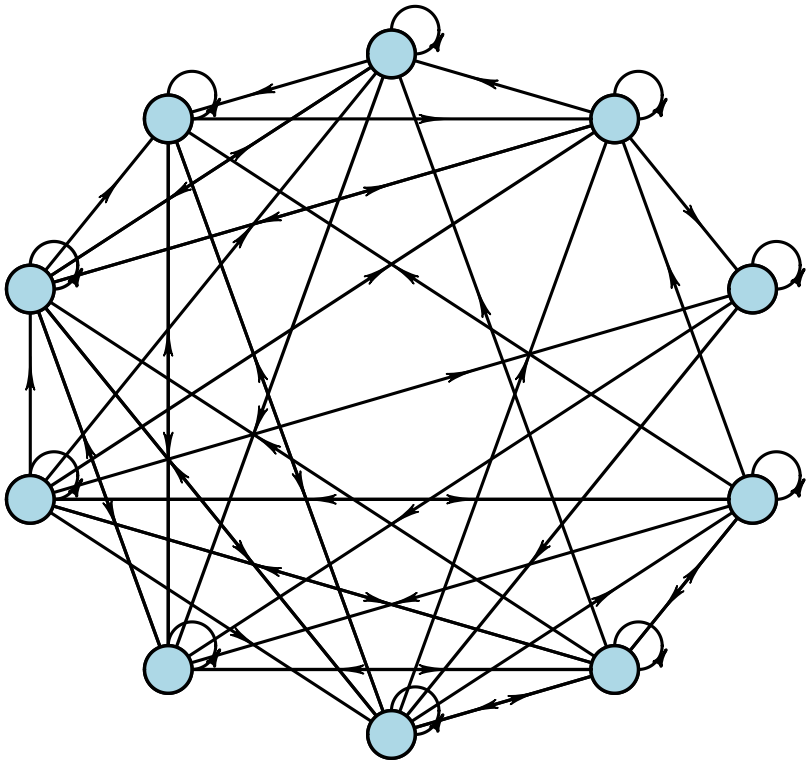
MSE = 1.04



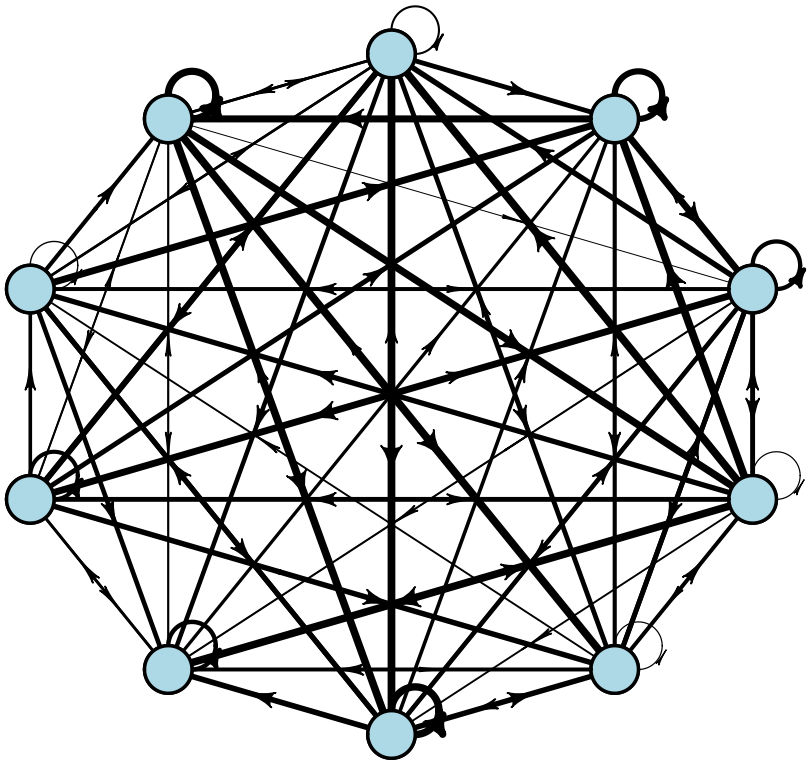
Synthetic data. Density 10%



Synthetic data. Density 50%



Synthetic data. Density 90%



VAR(1)

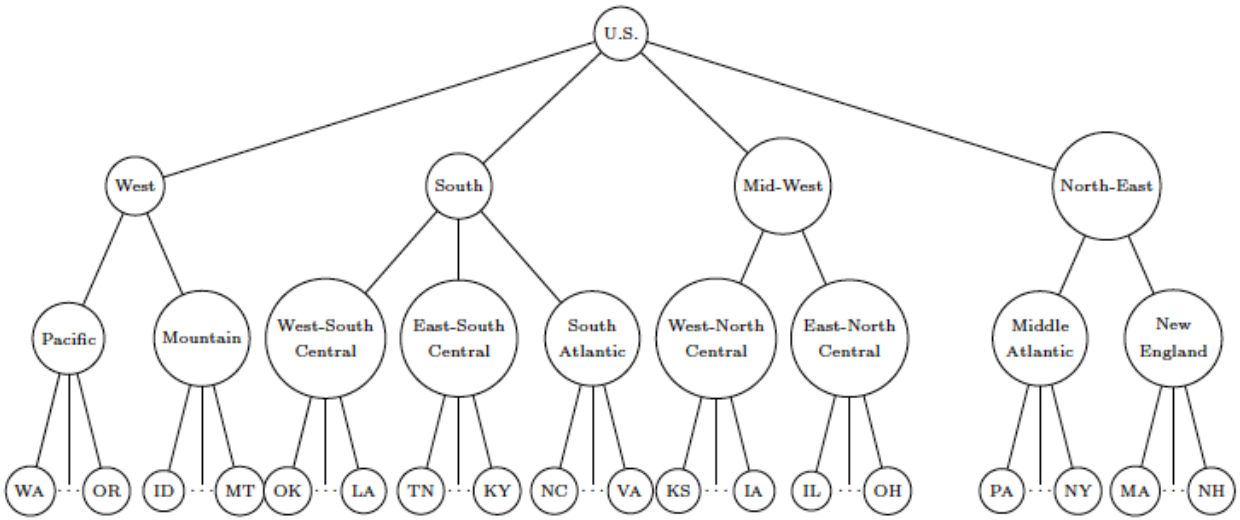
	Density 10%		Density 50%		Density 90%	
	Lasso	SC-VAR	Lasso	SC-VAR	Lasso	SC-VAR
$V_T = 1$	1.09	1.04	1.23	1.15	1.23	1.17
$V_T = 2$	1.03	1.00	1.15	1.07	1.20	1.11
$V_T = 3$	1.01	0.99	1.09	1.03	1.15	1.08
$V_T = 4$	0.99	0.99	1.05	1.02	1.12	1.05
$V_T = 5$	0.99	0.99	1.03	1.01	1.09	1.04
$V_T = 6$	1.00	0.99	1.01	1.01	1.06	1.04
$V_T = 7$	1.00	0.99	1.00	1.01	1.03	1.03
$V_T = 8$	1.00	0.99	1.00	1.01	1.02	1.03
$V_T = 9$	1.00	0.99	1.00	1.01	1.00	1.03
$V_T = 10$	1.00	0.99	1.00	1.01	1.00	1.03

VAR(2)

	Density 10%		Density 50%		Density 90%	
	GLasso	SC-VAR	GLasso	SC-VAR	GLasso	SC-VAR
$V_S = 1$	1.13	0.95	1.11	1.01	1.12	1.02
$V_S = 2$	1.15	0.90	1.11	0.96	1.13	0.99
$V_S = 3$	1.16	0.90	1.14	0.90	1.14	0.94
$V_S = 4$	1.17	0.90	1.17	0.89	1.15	0.91
$V_S = 5$	1.18	0.90	1.20	0.87	1.18	0.89
$V_S = 6$	1.19	0.90	1.24	0.87	1.21	0.88
$V_S = 7$	1.20	0.90	1.28	0.87	1.25	0.87
$V_S = 8$	1.20	0.90	1.30	0.87	1.31	0.87
$V_S = 9$	1.21	0.90	1.33	0.87	1.40	0.87
$V_S = 10$	1.24	0.90	1.36	0.87	1.48	0.87

Hierarchical linear regression


E Carrizosa, LH Mortensen, D Romero Morales, MR Sillero-Denamiel
On linear regression models with hierarchical categorical variables
https://www.researchgate.net/publication/341042405


U.S.
West
South
Mid-West
North-East
Pacific
Mountain
West-South Central
East-South Central
South Atlantic
West-North Central
East-North Central
Middle Atlantic
New England
WA
OR
ID
MT
OK
LA
TN
KY
NC
VA
KS
IA
IL
OH
PA
NY
MA
NH

The model

$$(\mathcal{T}_j)_{j\in\mathcal{J}} \qquad y = \beta'_0 + \sum_{j'\in\mathcal{J}'} \beta'_{j'} x'_{j'} + \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{jl}$$

Goodness of fit of tree model $(\mathcal{T}_j)_{j\in\mathcal{J}}$

$$MSE((\mathcal{T}_j)_{j\in\mathcal{J}}) : \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta'_0 - \sum_{j'\in\mathcal{J}'} \beta'_{j'} x'_{ij'} - \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{ijl}\right)^2$$

Complexity of tree model $(\mathcal{T}_j)_{j\in\mathcal{J}}$

$$C((\mathcal{T}_j)_{j\in\mathcal{J}}) : \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} c_{jl}$$

$$\min_{(\mathcal{T}_j)_{j\in\mathcal{J}}} \left(MSE((\mathcal{T}_j)_{j\in\mathcal{J}}), C((\mathcal{T}_j)_{j\in\mathcal{J}})\right)$$

The model

$$(\mathcal{T}_j)_{j\in\mathcal{J}} \qquad y = \beta'_0 + \sum_{j'\in\mathcal{J}'} \beta'_{j'} x'_{j'} + \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{jl}$$

Goodness of fit of tree model $(\mathcal{T}_j)_{j\in\mathcal{J}}$

$$MSE((\mathcal{T}_j)_{j\in\mathcal{J}}) : \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta'_0 - \sum_{j'\in\mathcal{J}'} \beta'_{j'} x'_{ij'} - \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{ijl}\right)^2$$

Complexity of tree model $(\mathcal{T}_j)_{j\in\mathcal{J}}$

$$C((\mathcal{T}_j)_{j\in\mathcal{J}}) : \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} c_{jl}$$

$$\min_{(\mathcal{T}_j)_{j\in\mathcal{J}}} \left(MSE((\mathcal{T}_j)_{j\in\mathcal{J}}), C((\mathcal{T}_j)_{j\in\mathcal{J}})\right)$$

The model

$$(\mathcal{T}_j)_{j\in\mathcal{J}} \qquad y = \beta'_0 + \sum_{j'\in\mathcal{J}'} \beta'_{j'} x'_{j'} + \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{jl}$$

Goodness of fit of tree model $(\mathcal{T}_j)_{j\in\mathcal{J}}$

$$MSE((\mathcal{T}_j)_{j\in\mathcal{J}}) : \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta'_0 - \sum_{j'\in\mathcal{J}'} \beta'_{j'} x'_{ij'} - \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{ijl}\right)^2$$

Complexity of tree model $(\mathcal{T}_j)_{j\in\mathcal{J}}$

$$C((\mathcal{T}_j)_{j\in\mathcal{J}}) : \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} c_{jl}$$

$$\min_{(\mathcal{T}_j)_{j\in\mathcal{J}}} \left(MSE((\mathcal{T}_j)_{j\in\mathcal{J}}), C((\mathcal{T}_j)_{j\in\mathcal{J}})\right)$$

The model

$$(\mathcal{T}_j)_{j\in\mathcal{J}} \qquad y = \beta'_0 + \sum_{j'\in\mathcal{J}'} \beta'_{j'} x'_{j'} + \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{jl}$$

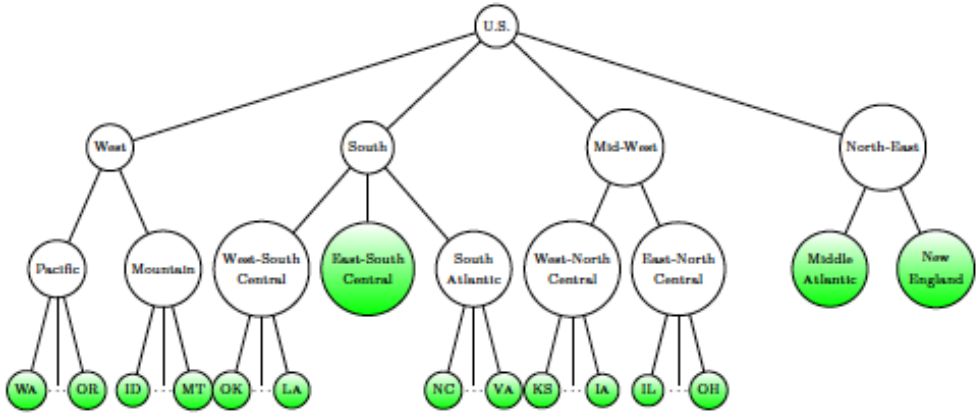
Goodness of fit of tree model $(\mathcal{T}_j)_{j\in\mathcal{J}}$

$$MSE((\mathcal{T}_j)_{j\in\mathcal{J}}) : \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta'_0 - \sum_{j'\in\mathcal{J}'} \beta'_{j'} x'_{ij'} - \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{ijl}\right)^2$$

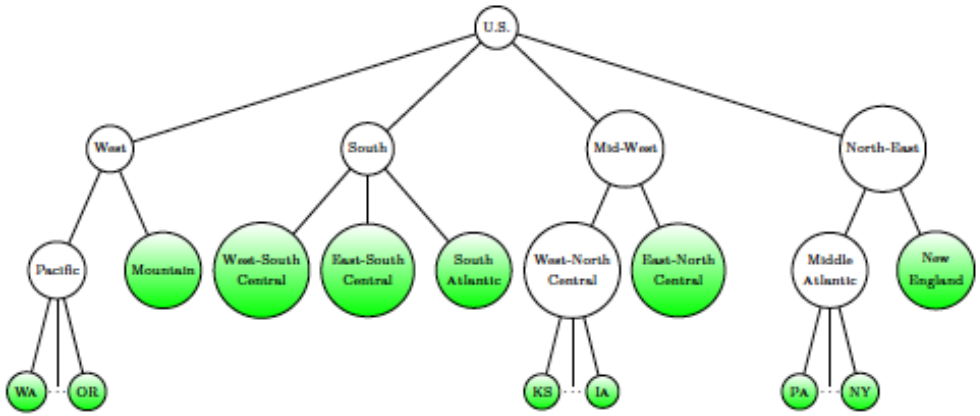
Complexity of tree model $(\mathcal{T}_j)_{j\in\mathcal{J}}$

$$C((\mathcal{T}_j)_{j\in\mathcal{J}}) : \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(\mathcal{T}_j)} c_{jl}$$

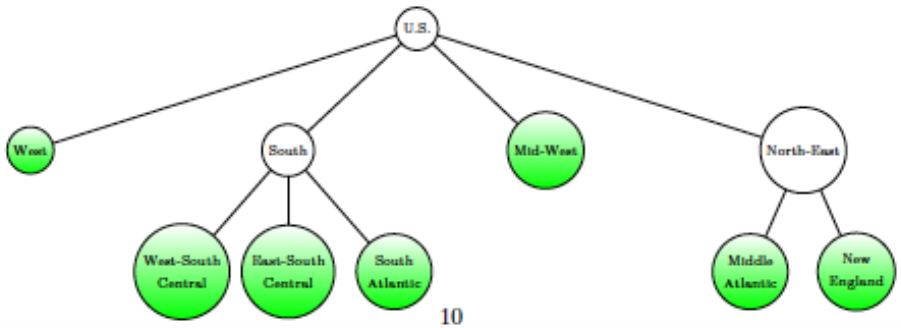
$$\min_{(\mathcal{T}_j)_{j\in\mathcal{J}}} \left(MSE((\mathcal{T}_j)_{j\in\mathcal{J}}), C((\mathcal{T}_j)_{j\in\mathcal{J}})\right)$$



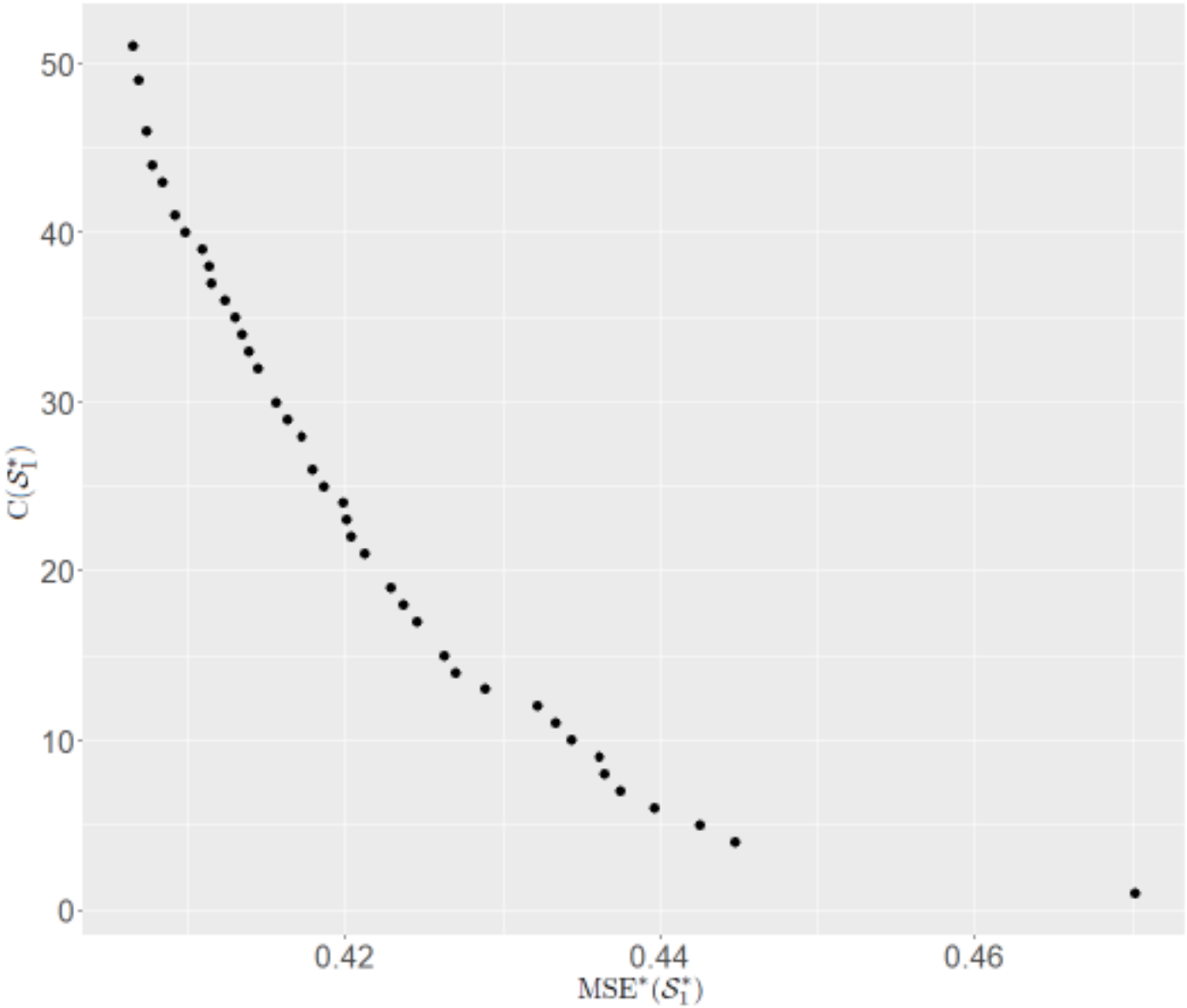
(a) $\mathcal{S}_1^*$ when $f = 0.409$ and $\mathrm{C}(\mathcal{S}_1^*) = 41$



(b) $\mathcal{S}_1^*$ when $f = 0.422$ and $\mathrm{C}(\mathcal{S}_1^*) = 21$



(c) $\mathcal{S}_1^*$ when $f = 0.438$ and $\mathrm{C}(\mathcal{S}_1^*) = 7$


50
40
30
20
10
0
$C(\mathcal{S}_1^*)$
0.42
0.44
0.46
$MSE^*(\mathcal{S}_1^*)$