Lasso Regularization Paths for NARMAX Models via Coordinate Descent

Antônio H. Ribeiro, Luis A. Aguirre

Universidade Federal de Minas Gerais (UFMG), Brazil

American Control Conference, June 29, 2018
Milwaukee, U.S.

Problem Statement

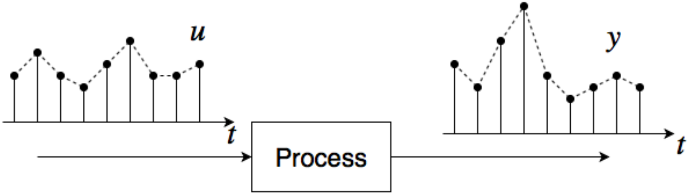

Figure: The system identification problem.

Prediction Error Methods Framework

Cost Function

$$V(\boldsymbol{\theta}) = \sum_k \Big\| \overbrace{y[k]}^{\text{observed}} - \underbrace{\hat{y}_{\boldsymbol{\theta}}[k]}_{\text{predicted}} \Big\|^2.$$

Linear-in-the-Parameters Model

Linear-in-the-parameter models:

$$\hat{y}_{\boldsymbol{\theta}}[k] = \sum_i \theta_i \cdot \overbrace{x_i(y[k-1], u[k-1])}^{\text{basis functions}},$$

Ordinary least-squares formulation:

$$\min_{\boldsymbol{\theta}} \sum_k \|y[k] - \hat{y}_{\boldsymbol{\theta}}[k]\|^2 \Rightarrow \min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$$

L_1 penalty

The Lasso

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1,$$

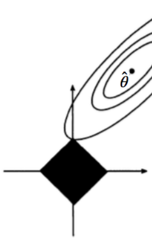

Figure: Lasso interpretation (Tibshirani, 1996).

Tibshirani, R. (1996).
Regression shrinkage and selection via the LASSO.
Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288.

Literature Review

Solving Lasso Problem

- Quadratic Programming;

Tibshirani, R. (1996).
Regression shrinkage and selection via the LASSO.
Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288.

Literature Review

Solving Lasso Problem

- Quadratic Programming;
- LARS (Least Angle Regression) algorithm;

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004).
Least angle regression.
The Annals of Statistics, 32(2):407–499.

Literature Review

Solving Lasso Problem

- Quadratic Programming;
- LARS (Least Angle Regression) algorithm;
- *Coordinate Descent*;

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007).
Pathwise coordinate optimization.
The Annals of Applied Statistics, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2009).
Glmnet: Lasso and elastic-net regularized generalized linear models.
R package version, 1(4).

Friedman, J., Hastie, T., and Tibshirani, R. (2010).
Regularization paths for generalized linear models via coordinate descent.
Journal of statistical software, 33(1):1.

Coordinate Descent Algorithm

One-at-a-time coordinate optimization:

$$\theta_j \leftarrow \arg_{\theta_j} \min \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1,$$

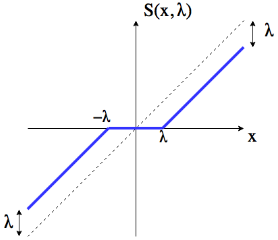

Figure: Soft threshold operator

Coordinate Descent Algorithm

One-at-a-time coordinate optimization:

$$\theta_j \leftarrow \frac{1}{\|\mathbf{x}_j\|^2} S\left(\left(\overbrace{(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})}^{\mathbf{r}} + \mathbf{x}_j\theta_j\right)^T \mathbf{x}_j;\ \lambda\right),$$

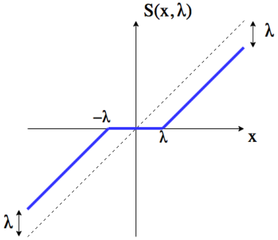

Figure: Soft threshold operator

Coordinate Descent Algorithm

Optimization Problem

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1,$$

Repeat:

1. $\theta_j \leftarrow \frac{1}{\|\mathbf{x}_j\|^2} S\left((\mathbf{r} + \mathbf{x}_j\theta_j)^T\mathbf{x}_j;\ \lambda\right)$
2. Update $\mathbf{r} = (\mathbf{y} - \mathbf{X}\theta)$
3. Next j.

Coordinate Descent Algorithm

Optimization Problem

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1,$$

Repeat:

1. $\theta_j \leftarrow \frac{1}{\|\mathbf{x}_j\|^2} S\left((\mathbf{r} + \mathbf{x}_j\theta_j)^T\mathbf{x}_j;\ \lambda\right) \quad \rightarrow \quad \mathcal{O}(N)$
2. Update $\mathbf{r} = (\mathbf{y} - \mathbf{X}\theta) \quad \rightarrow \quad \mathcal{O}(N)$
3. Next j.

NARMAX model

Assuming that:

$$\begin{aligned} r[k] &= y[k] - \hat{y}_{\boldsymbol{\theta}}[k] \\ \hat{y}_{\boldsymbol{\theta}}[k] &= \sum_{i=1}^{p} \theta_i \cdot x_i(\underbrace{y[k-1], u[k-1]}_{\text{measured values}}, \underbrace{r[k-1]}_{\text{noise term}}). \end{aligned}$$

Estimated parameter:

$$\hat{\boldsymbol{\theta}} = \arg_{\boldsymbol{\theta}} \min \|\mathbf{y} - \mathbf{X}_{(\mathbf{y},\mathbf{u},\mathbf{r})}\boldsymbol{\theta}\|_2^2.$$

Extended Least Squares

Optimization Problem

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}_{(\mathbf{y},\mathbf{u},\mathbf{r})}\boldsymbol{\theta}\|_2^2,$$

Repeat:

1. $\hat{\boldsymbol{\theta}}^{(i+1)} \leftarrow \arg_{\boldsymbol{\theta}} \min \left\|\mathbf{y} - \mathbf{X}_{(\mathbf{y},\mathbf{u},\mathbf{r}^{(i)})}\boldsymbol{\theta}\right\|^2$
2. $\hat{\mathbf{r}}^{(i+1)} \leftarrow \mathbf{y} - \mathbf{X}_{(\mathbf{y},\mathbf{u},\mathbf{r}^{(i)})}\boldsymbol{\theta}^{(i+1)}$
3. $i \leftarrow i + 1.$

Coordinate Descent Algorithm (Revisited)

Optimization Problem

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}_{(\mathbf{y},\mathbf{u},\mathbf{r})}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1,$$

Repeat:

1. Update $\mathbf{x}_j$ if it depends on $\mathbf{r}$
2. $\theta_j^+ \leftarrow \frac{1}{\|\mathbf{x}_j\|^2} S\left((\mathbf{r} + \mathbf{x}_j\theta_j)^T\mathbf{x}_j;\ \lambda\right)$
3. Update $\mathbf{r} = (\mathbf{y} - \mathbf{X}\theta)$
4. Next j.

Coordinate Descent Algorithm (Revisited)

Optimization Problem

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}_{(\mathbf{y},\mathbf{u},\mathbf{r})}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1,$$

Repeat:

1. Update $\mathbf{x}_j$ if it depends on $\mathbf{r}$ $\rightarrow$ $\mathcal{O}(N)$
2. $\theta_j^+ \leftarrow \frac{1}{\|\mathbf{x}_j\|^2} S\left((\mathbf{r} + \mathbf{x}_j\theta_j)^T\mathbf{x}_j;\ \lambda\right)$ $\rightarrow$ $\mathcal{O}(N)$
3. Update $\mathbf{r} = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$ $\rightarrow$ $\mathcal{O}(N)$
4. Next j.

Example I

- The dataset was generated from the linear system:

$$y[k] = 0.5y[k-1] - 0.5u[k-1] + 0.5v[k-1] + v[k].$$

Example I

- The dataset was generated from the linear system:

$$y[k] = 0.5y[k-1] - 0.5u[k-1] + 0.5v[k-1] + v[k].$$

- We try to fit the following linear model to the training data (30 regressors):

$$y[k] = \sum_{i=1}^{10} \theta_i y[k-i] + \sum_{i=1}^{10} \theta_{(i+10)} u[k-i] + \sum_{i=1}^{10} \theta_{(i+20)} r[k-i].$$

Example I

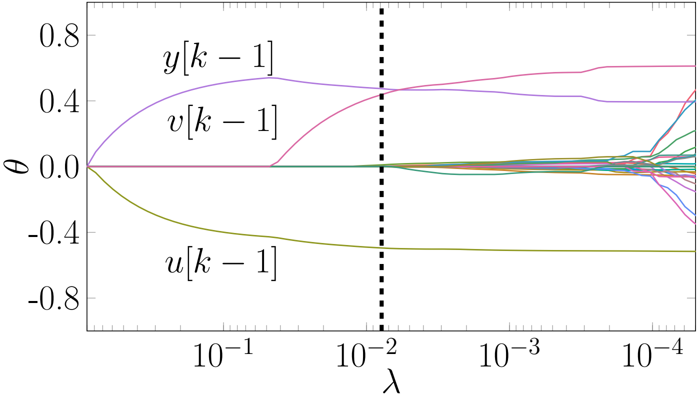

Figure: Estimated parameter vector $\boldsymbol{\theta}$ as a function of λ. Estimated system: $y[k] = 0.48y[k-1] - 0.50u[k-1] + 0.44v[k-1]$.

Example II

The dataset was generated from the nonlinear system (Chen, et. al., 1990):

$$\begin{aligned} y[k] &= (0.8 - 0.5\exp(-y[k-1]^2)y[k-1] + u[k-1] - \\ &\quad (0.3 + 0.9\exp(-y[k-1]^2)y[k-2] + 0.2u[k-2] + \\ &\quad 0.1u[k-1]u[k-2] + 0.1v[k-1] + 0.3v[k-2] + v[k], \end{aligned}$$

And, we fit a polynomial model with degree 2 and 44 regressors to it.

S. Chen, S. A. Billings, and P. M. Grant (1990).
Non-linear system identification using neural networks
International Journal of Control, vol. 51, no. 6, pp. 1191–1214, 1990.

Example II

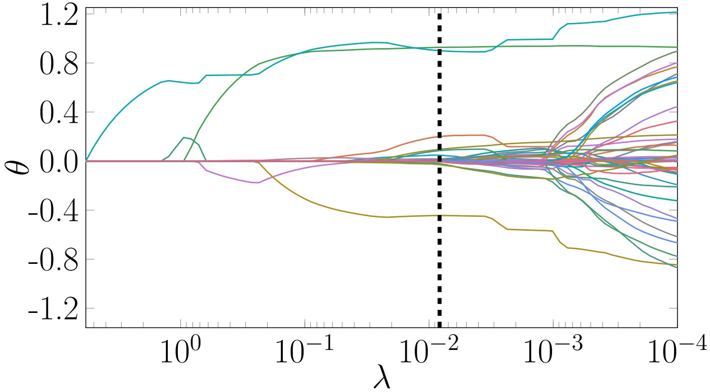

Figure: Estimated parameter vector $\boldsymbol{\theta}$ as a function of λ. For this optimal λ the mean absolute error in the validation set is 1.03 and the model includes the regressors $y[k-1]$, $u[k-1]$, $y[k-3]$, $y[k-2]$, $u[k-2]$, $r[k-1]$, $r[k-2]$, $y[k-1]y[k-2]$, $u[k-1]u[k-2]$, $y[k-3]r[k-1]$, $y[k-2]u[k-2]$.

Related Work

H. Wang, G. Li, and C.-L. Tsai (2007).
Regression Coefficient and Autoregressive Order Shrinkage and Selection Via the Lasso.
Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 69, no. 1, pp. 63–78, 2007.

Y. J. Yoon, C. Park, and T. Lee (2013).
Penalized regression models with autoregressive error terms.
Journal of Statistical Computation and Simulation, vol. 83, no. 9, pp. 1756–1772, Sep. 2013.

Conclusion

1. Timmings;

Conclusion

1. Timmings;
2. Convergence;

Conclusion

1. Timmings;
2. Convergence;
3. Scaling;

Conclusion

1. Timmings;
2. Convergence;
3. Scaling;
4. *Elastic net*;

Acknoledgments

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

CAPES

The implementation is available at:
https://github.com/antonior92/NarmaxLasso.jl