Factor-adjusted Regularized Model Selection

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Jianqing Fan, Yuan Ke, Kaizheng Wang
ORFE, Princeton University
Department of Statistics, University of Georgia
IEOR, Columbia University

Model selection challenges in financial data

Parsimonious models are desirable due to interpretability. However, most existing model selection techniques based on sparse regression are not tailored for financial applications, where the variables are cross-sectionally correlated and serially dependent. Fig 1 shows that when the variables are correlated, Lasso (Tibshirani, 1996) selects many spurious variables (true model size is 10). We propose FARMSelect to solve this problem.

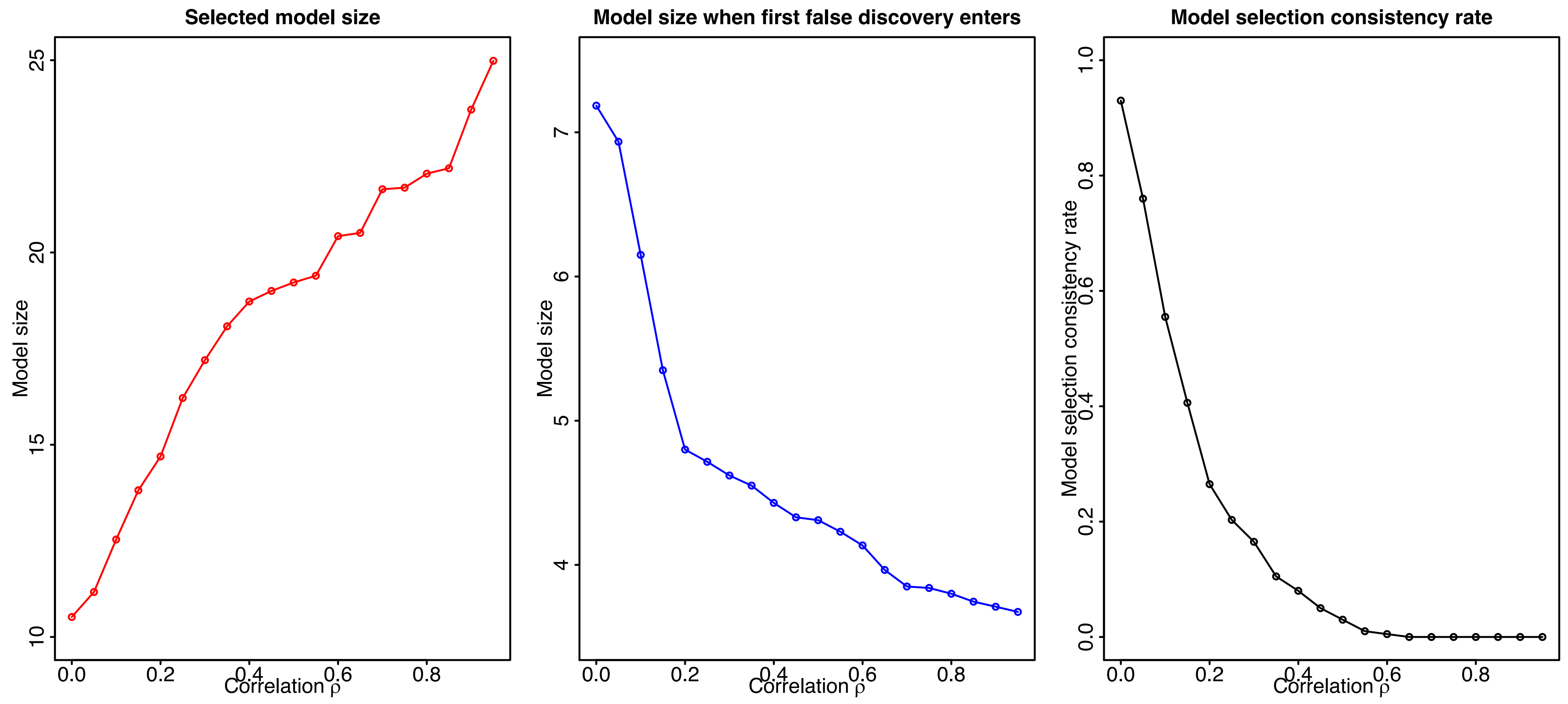

Figure 1. Inconsistent model selection in the presence of strong correlation.

Factor-Adjusted Regularized Model Selection (FARMSelect)

Linear model $y = x^T\beta^* + \varepsilon$ with response y and feature vector $x \in R^p$. The Lasso

$$\min_{\beta \in R^p} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1 \right\}$$

fails due to the dependency among variables. Consider the approximate factor model $x_i = Bf_i + u_i$ with loading matrix $B \in R^{p \times K}$, factor vector $f_i \in R^K$ and idiosyncratic component $u_i \in R^p$. Unlike x, the coordinates of u are **weakly dependent**.

Inspired by $y - f^T\,(B^T\beta^*) = u^T\beta^* + \varepsilon$, we develop the FARMSelect procedure.

1) Fit the factor model to estimate $\{f_i, u_i\}_{i=1}^n$ from $\{x_i\}_{i=1}^n$, getting $\{\hat{f}_i, \hat{u}_i\}_{i=1}^n$;
2) Estimate $\gamma^* = B^T\beta^* \in R^K$ by regressing $\{y_i\}_{i=1}^n$ against $\{x_i\}_{i=1}^n$, getting $\hat{\gamma}$;
3) Conduct Lasso using the decorrelated variables $\{u_i\}_{i=1}^n$:

$$\min_{\beta \in R^p,\, \gamma \in R^K} \left\{ \frac{1}{n} \sum_{i=1}^n [(y_i - \hat{f}_i^T \hat{\gamma}) - \hat{u}_i^T \beta]^2 + \lambda \|\beta\|_1 \right\}$$

Extension: generalized linear models (GLMs).

Real data: prediction of U.S. bond risk premia

Monthly data from Jan. 1980 to Dec. 2015 ($n = 432$)

$\{y_i\}_{i=1}^n$: U.S. bond risk premia with maturities of 2 - 5 years

$\{x_i\}_{i=1}^n$: 128 macroecon. variables in FRED-MD database (McCracken and Ng, 2016).

One month ahead rolling window prediction (window size = 120).

Comparison: FARMSelect, Lasso, Principal Component Regression (PCR)

FARMSelect: highest prediction power, most parsimonious model.

Maturity	Out-of-sample R^2			Ave. Model Size	
	FARMSelect	Lasso	PCR	FARMSelect	Lasso
2 years	0.530	0.509	0.462	5.96	6.86
3 years	0.526	0.523	0.483	5.71	7.09
4 years	0.484	0.476	0.470	5.53	6.81
5 years	0.481	0.475	0.477	5.90	6.84

Table 1. Prediction of U.S. bond risk premia

Theoretical guarantees

FARMSelect consistently identifies important variables and achieve the optimal error rate even if the data exhibit cross-sectional and serial dependency.

Main assumptions: (1) GLM and approximate factor model; (2) mixing condition.

Acknowledgments

This research was supported in part by the NSF grants DMS-1662139 and DMS-1712591; ONR grant N00014-19-1-2120; NIH grant 2R01-GM072611-13.

References

Fan, Jianqing, Ke, Yuan, and Kaizheng Wang. "Factor-adjusted regularized model selection." Journal of Econometrics (2020).

McCracken, Michael W., and Serena Ng. "FRED-MD: A monthly database for macroeconomic research." Journal of Business & Economic Statistics 34, no. 4 (2016): 574-589.

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58, no. 1 (1996): 267-288.