Penalised regressions and data compression methods

Week 4

Daniele Bianchi[1]

whitesphd.com

[1] School of Economics and Finance
Queen Mary, University of London

Summary

This week we delve further into large dimensional linear models. In particular, we are going to explore the issues and benefits of working with high-dimensional models. Related to that, we will investigate the use of both penalised regression and dimensionality reduction methods for the analysis of credit risk.

Contents

Recap

Recap

- Structural models of credit risk:
  - Restrictions: Variables and Functional form
  - The Merton Model
  - Take model to the data, and it works well!
- Reduced-form approaches to credit risk:
  - Key: Relax variable use and functional form.
  - Functional form $\rightarrow$ Statistical learning techniques!
- Some new concepts:
  - Type I and II errors, confusion matrix.
  - ROC Curves.
  - SMOTE, Undersampling and Oversampling.
- This lecture: How do you pick the variables to predict credit-risk when there is no model to guide you?

Regularization

Regularization

- One important tool that we didn't cover in great detail in the previous lectures is Regularization.
- Regularization refers to the process of introducing additional constraints into the optimization problem in order to get a more sensible solution, and in particular to avoid the problem of overfitting.
- This is one of the most important and useful methods in the machine learning toolkit, as we will see.
- Helps to solve a common problem: How can we make progress if there are many possible predictors of an outcome variable?

Linear Regression, Prediction, and Regularization

- Consider the standard OLS regression:

$$Y_i = \beta_0 + \sum_{k=1}^{K} X_{ik}\beta_k + \varepsilon_i = X_i^{'}\beta + \varepsilon_i.$$

- We usually estimate the parameters of this regression as:

$$\arg \min_{\beta_0, \beta_1, .., \beta_K} \sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{k=1}^{K} X_{ik}\beta_k \right)^2 .$$

- Several regularization approaches can be used to modify this objective function.

- Why? To Improve prediction out of sample by reducing overfitting and only considering relevant variables.

Linear Regression, Prediction, and Regularization

- Using the linear regression is useful as an intuitive explanatory device, but less good for prediction.

- For example, when $K \geq 3$, OLS is not admissible. For more, see Brad Efron's intuitive explanation of Stein (1955), also known as Stein's paradox in statistics: **https://stanford.io/2uyQiOk**
  - In practice, this means that there are better predictors than OLS when $K \geq 3$.

- Another issue is what we might do when K is very large – for example, you will see in your credit assignment, there are over 100 possible variables but this can be much higher if we consider nonlinear transformations of these variables.
  - In medical science, and in other finance contexts, often millions.

- One possible problem when you have many explanatory variables is that you increase the chance that they are highly correlated with one another – meaning that their magnitudes may be off because of collinearity.

- Another is that it can make testing restrictions or hypotheses very difficult especially if K is high relative to N, the sample size.

- Finally there's the obvious issue of interpretation being very hard when there are numerous variables on the RHS, making it difficult to figure out which are truly important.

Regularization - Best Subsets

- More regressors into OLS $\rightarrow$ increase goodness of fit.
- But this gives rise to the usual bias-variance tradeoff.
- There are several ways to introduce regularization constraints.
- One way is to limit the set of regressors explictly (say $k \subset K$) . This is called "best subset selection"
- Computationally challenging!
- Brute force way: Fit all models with k regressors. Select amongst them using R^2 or some other metric.
- If $K = 10$, you have 1023 models to fit. More generally, there will be $\binom{K}{k}$ to fit!

Regularization - Stepwise Regressions

- A simpler approach is offered by stepwise regression. The procedure is:

1. For $k = 0, ..., K-1$
   - 1.1 Consider all models with $k+1$ regressors, i.e, those that augment the current model with one additional predictive variable.
   - 1.2 Pick the best of these models using your preferred criterion (say R^2).
   - 1.3 Either keep going until you hit a pre-specified number of regressors, or pick the model with the $k+1$ that minimizes, say, the adjusted R^2.

Hedge Fund Factors with Stepwise Regression

Agarwal and Naik (2004)

Table 4: Results with HFR Equally-Weighted Indexes

This table shows the results of the regression $R_t^i = c^i + \sum_{k=1}^{K} l_k^i F_{k,t} + u_t^i$ for the eight HFR indexes during the full sample period from January 1990 to June 2000 period. The table shows the intercept (C), statistically significant (at five percent level) slope coefficients on the various buy-and-hold and option-based risk factors and adjusted R^2 (Adj-R^2). The buy-and-hold risk factors are Russell 3000 index (RUS), lagged Russell 3000 index (LRUS)), MSCI excluding the US index (MXUS), MSCI Emerging Markets index (MEM), Fama-French Size and Book-to-Market factors (SMB & HML), Momentum factor (MOM), Salomon Brothers Government and Corporate Bond index (SBG), Salomon Brothers World Government Bond index (SBW), Lehman High Yield Composite index (LHY), Federal Reserve Bank Competitiveness-Weighted Dollar index (FRBI), Goldman Sachs Commodity index (GSCI) and the change in the default spread in basis points (DEFSPR). The option-based risk factors include the at-the-money and out-of-money call and put options on the S&P 500 Composite index ($SPC_{a/o}$ and $SPP_{a/o}$). For the two call and put option-based strategies, subscripts *a* and *o* refer to at-the-money and out-of-the-money respectively.

Event Arbitrage		**Restructuring**		**Event Driven**		**Relative Value Arbitrage**		**Convertible Arbitrage**		**Equity Hedge**		**Equity Non-Hedge**		**Short Selling**	
Factors	l	**Factors**	l	**Factors**	l	**Factors**	l	**Factors**	l	**Factors**	l	**Factors**	l	**Factors**	l
C	0.04	C	0.43	C	0.20	C	0.38	C	0.24	C	0.99	C	0.56	C	-0.07
SPP_o	-0.92	SPP_o	-0.63	SPP_o	-0.94	SPP_o	-0.64	SPP_a	-0.27	RUS	0.41	RUS	0.75	SPC_o	-1.38
SMB	0.15	SMB	0.24	SMB	0.31	MOM	-0.08	LRUS	0.10	SMB	0.33	SMB	0.58	RUS	-0.69
HML	0.08	HML	0.12	HML	0.12	SMB	0.17	SMB	0.05	HML	-0.08	MEM	0.05	SMB	-0.77
		LRUS	0.06	RUS	0.17	HML	0.08	MEM	0.03	GSCI	0.08			HML	0.40
		LHY	0.13	MEM	0.06	MXUS	0.04	SBG	0.16						
		FRBI	0.27												
		MEM	0.09												
Adj-R^2	44.04	Adj-R^2	65.57	Adj-R^2	73.38	Adj-R^2	52.17	Adj-R^2	40.51	Adj-R^2	72.53	Adj-R^2	91.63	Adj-R^2	82.02

Regularization - Best Subsets

- Can also approach this using backward stepwise regression. The procedure begins by estimating the full model with K regressors. Then:

1. For $k = K, K-1, ..., 1$

   1.1 Consider all models with $k-1$ regressors, i.e, those that delete one regressor at a time from the full model.

   1.2 Pick the best of these models using your preferred criterion (say R^2).

   1.3 Either keep going until you hit a pre-specified number of regressors, or pick the model with the k that minimizes, say, the adjusted R^2.

Regularization - Best Subsets

- Note that both forward and backward stepwise regression are substantially less computationally intensive than best subset selection overall.

- Also worth noting that backward stepwise regression cannot be used if $K > N$, the sample size.

- Forward stepwise is the only solution in this case.

- Now let's move across to understanding two other important techniques for regularization that are commonly applied in big data applications. These are Ridge Regression and the LASSO.

Ridge regression

LASSO in Returns Forecasting

Chinco, Clark-Joseph, Ye (2019)

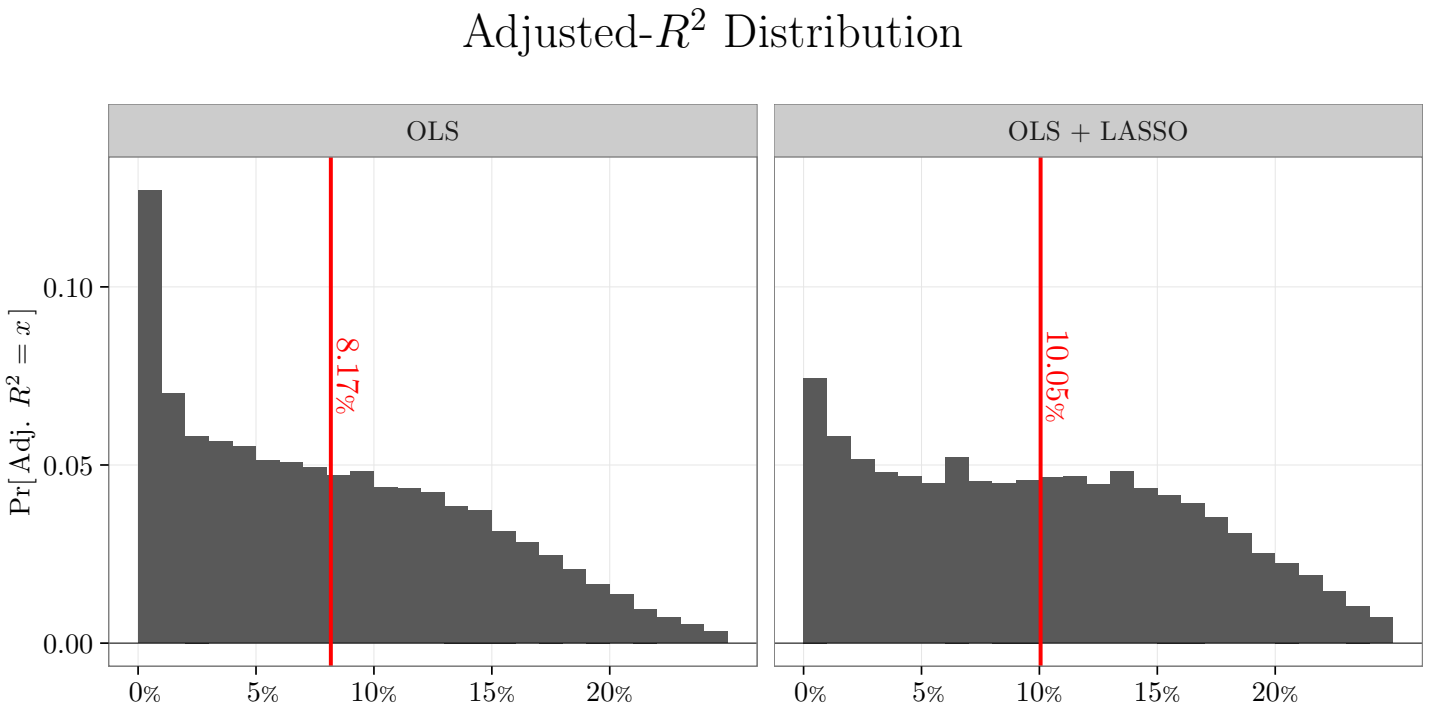


Figure 4: *Distribution of adjusted R^2s from the forecasting regressions in Equations (4) and (8). Black bars: Probability that the adjusted R^2 from a single out-of-sample forecasting regression falls within a 1%-point interval. Red vertical line: Average adjusted R^2 from these regressions corresponding to the point estimates in the bottom row of Table 1. Left panel: Out-of-sample prediction made using OLS as in Equation (4). Right panel: Out-of-sample predictions made using both OLS and the LASSO as in Equation (8). Reads: "Including the LASSO's return forecast increases out-of-sample predictive power by $^{10.05}/_{8.17} - 1 = 23\%$ relative to the benchmark OLS model."*

Ridge Regression

- The Ridge Regression estimates parameters by minimizing the following objective function:

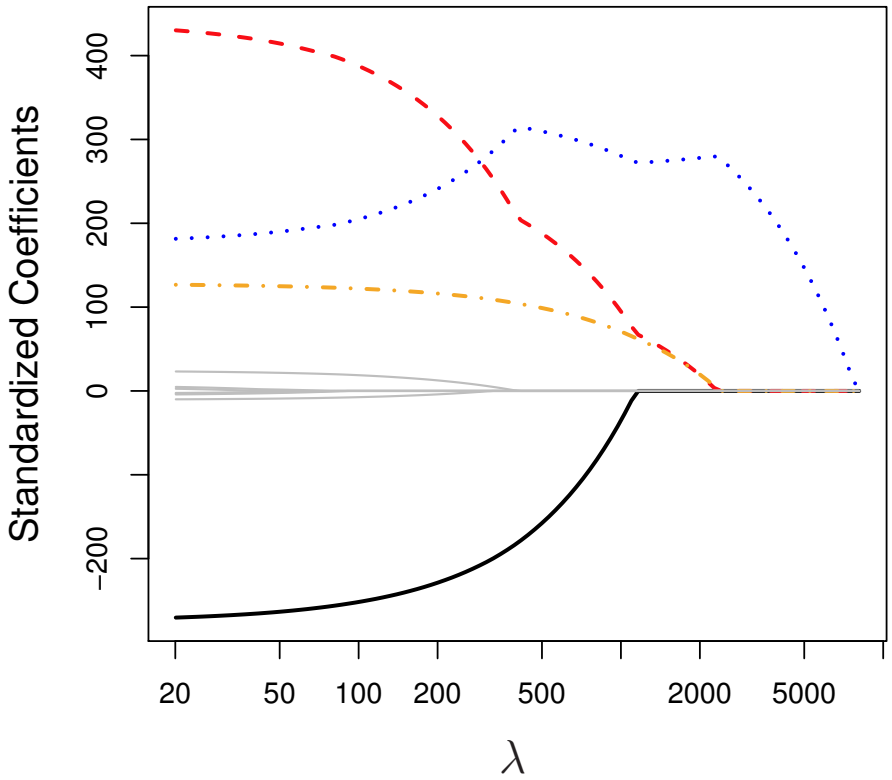
$$\begin{aligned} & \sum_{i=1}^{N}\left(Y_i-\beta_0-\sum_{k=1}^{K}X_{ik}\beta_k\right)^2+\lambda\sum_{k=1}^{K}\beta_k^2 \\ = \quad & RSS+\lambda\sum_{k=1}^{K}\beta_k^2. \\ \hat{\beta}_{Ridge} \quad = \quad & (X'X+\lambda I_K)^{-1}(X'Y) \end{aligned}$$

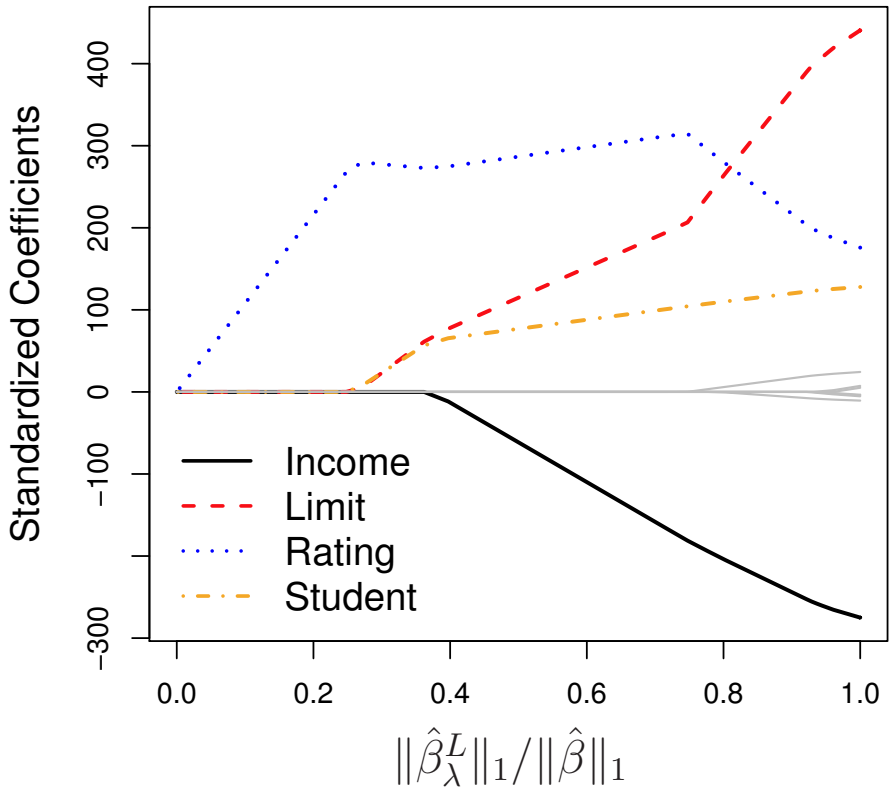
Ridge Regression

- Here, $\lambda \geq 0$ is a tuning parameter to be estimated separately (More later).
- As with OLS, ridge regression minimizes the residual sum of squares.

- But it also does one additional thing: The second term, $\lambda \sum_k \beta_k^2$ is the *shrinkage penalty*.
- When β_k is small (close to zero), it shrinks the estimate to zero.
- When $\lambda = 0$, it is the same as the OLS. As $\lambda \to \infty$, then the coefficient estimates will approach zero.
- How does one choose λ? More on this later.

An Application to Credit Data





- $\hat{\beta}$ vector of OLS estimates, and $\hat{\beta}_\lambda^R$ the ridge regression estimates. $||x||$ is the ℓ_2 norm: distance of β from zero.
- x-axis for the second plot measures the amount of shrinkage towards zero relative to the OLS estimates.

Ridge Regression: Data processing Note

- In standard OLS, if we multiply one of the predictor variables by a constant c, the coefficient estimate simply scales by $\frac{1}{c}$. This means that for any scaled predictor, $X_k\hat{\beta}_k$ remains the same.

- However, because of the penalty function, this is not the case for the Ridge Regression. So it is best to first standardize the predictors before estimating (turn them into mean zero, variance one by dividing by in-sample standard deviation).

- Note also that the Ridge Regression will smoothly shrink the parameters to zero, i.e., if regressors are orthogonal to one another and normalized as above, all β_k coefficients will be shrunk by a factor of $\frac{1}{1+\lambda}$.

The LASSO

The LASSO

- Best subset selection is problematic: Computationally infeasible.
- Ridge Regression is computationally feasible, but: All predictors are generally selected (though shrunk).
- The LASSO does not have this drawback. The estimator is the solution to the optimization problem:

$$\begin{aligned} & \sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{k=1}^{K} X_{ik} \beta_k \right)^2 + \lambda \sum_{k=1}^{K} |\beta_k| \\ = \quad & RSS + \lambda \sum_{k=1}^{K} |\beta_k| \, . \end{aligned}$$

- The estimator yields sparse models, which only involve a subset of variables.

The LASSO

- Ridge regression: β_k^2 penalty. (ℓ_2 penalty)
- LASSO regression: $|\beta_k|$ penalty. (ℓ_1 penalty)
- ℓ_1 penalty forces coefficients to be exactly equal to zero when the parameter λ is sufficiently large.
- Like the best subset selection approach, LASSO does *variable selection* for you!
- Easier to interpret than ridge regressions.

Intuition for the LASSO

- Useful intuition for the LASSO, when the RHS variables in the LASSO regression are uncorrelated and have unit variance:
  - If $\hat{\beta}_k^{ols}$ is the OLS estimator, and $\hat{\beta}_k^{LASSO}$ the corresponding LASSO estimator, then:

$$\hat{\beta}_k^{LASSO} = \text{sign}(\hat{\beta}_k^{ols})[\max(0, \left|\hat{\beta}_k^{ols}\right| - \lambda)]$$

  - If OLS coefficient is estimated large (assume positive), then LASSO delivers similar estimate, since $\hat{\beta}_k^{LASSO} = \hat{\beta}_k^{ols} - \lambda \approx \hat{\beta}_k^{ols}$.
  - If the OLS coefficient is estimated small relative to λ, then $\hat{\beta}_k^{LASSO} = 0$.

Intuition for the LASSO

Chinco, Clark-Joseph, Ye (2019)

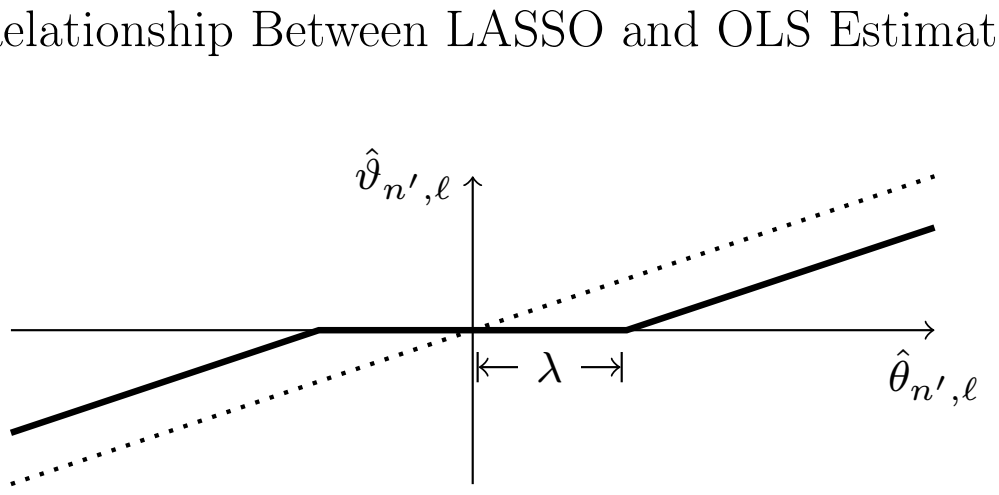


Figure 2: *x-axis: OLS-regression coefficient in an infinite sample. y-axis: Penalized-regression coefficient from the LASSO. Dotted: $x = y$ line. Reads: "If an OLS regression would have estimated a small coefficient value given enough data, $|\hat{\theta}_{n',\ell}| < \lambda$, then the LASSO will set $\hat{\vartheta}_{n',\ell} = 0$."*

LASSO in Finance

- As you will see, the LASSO is an incredibly useful tool in a wide variety of contexts. You will use it in at least two cases in this course:
  - Credit scoring (your assignment!).
  - Returns forecasting.

- But we will also consider how the LASSO helps with causal inference and not just prediction, later in the course.

Data compression methods

Transforming predictors

- So far, we have reduced the number of variables by using all the predictors, and then work with the estimated coefficients.
- Now, we will see one approach that transforms the predictors before we run various methods to extract the prediction.
- $X_1, ..., X_k$ are our original predictors. Suppose we can extract a smaller set of $Z_1, ..., Z_m$ where $m < k$, such that:

$$Z_m = \sum_{j=1}^{k} \phi_{jm} X_j$$

- This is just the linear combination of our original k predictors.
- We reduce the dimension by capturing all the variation in the data in fewer variables extracted from a large set of predictors.
- One approach to do this is the Principal Components Analysis.

Principal Components Analysis

- Suppose you have an nxk data matrix, X.
- n $=$ number of observations, and k is the number of predictors.
- We want to capture most of the *variation* in this data in fewer number of variables than K.
- Let us suppose we want M variables than K. Let's say we want only 5, i.e, $m = 5$.
- These M predictors should explain as much variation of all N observations as possible.
- The total variation is $\sum_{n=1}^{N} \sigma_n^2$.
- Let $\widehat{\Omega} = \widehat{Var}\,(X)$ be the $(N \times N)$ sample covariance matrix.

Intuition

- Often predictors are correlated with each other and contain the "same" information.
- Principal Components reduces the dimensionality of the problem by only looking for the unique information set in the data.
- All m variables extracted from k variables are orthogonal to each other.
    - They are uncorrelated with each other.
- Very powerful tool, can reduce dimensionality by many factors in finance.

Forecasting bond excess returns

Ludvigson and Ng (2009)

Model: $rx_{t+1}^{(n)} = \beta_0 + \beta_1' \widehat{F}_t + \beta_2 CP_t + \epsilon_{t+1}$,

		$\widehat{F}_{1t}$	$\widehat{F}_{1t}^3$	$\widehat{F}_{2t}$	$\widehat{F}_{3t}$	$\widehat{F}_{4t}$	$\widehat{F}_{8t}$	CP_t	$F5_t$	$F6_t$	$\bar{R}^2$
$rx_{t+1}^{(2)}$	(a)							**0.45**			0.31
								(8.90)			
	(b)	**−0.93**	**0.06**	**−0.40**	**0.18**	**−0.33**	**0.35**				0.26
		(−5.19)	(2.78)	(−3.10)	(2.24)	(−2.94)	(4.35)				
	(c)	**−0.74**	**0.05**	0.08	**0.24**	**−0.24**	**0.24**	**0.41**			0.45
		(−4.48)	(2.70)	(0.71)	(3.84)	(−2.51)	(2.70)	(5.22)			
	(d)	**−0.93**	**0.06**		0.18	**−0.33**	**0.35**				0.22
		(−4.96)	(2.87)		(1.87)	(−2.65)	(3.83)				
	(e)	**−0.75**	**0.05**		**0.24**	**−0.25**	**0.24**	**0.40**			0.45
		(−4.71)	(2.71)		(3.85)	(−2.61)	(2.89)	(5.89)			
	(f)								**0.54**		0.22
									(5.52)		
	(g)									**0.50**	0.26
										(6.78)	
	(h)							**0.39**	**0.43**		0.44
								(6.0)	(5.78)		

Forecasting bond excess returns

Ludvigson and Ng (2009)

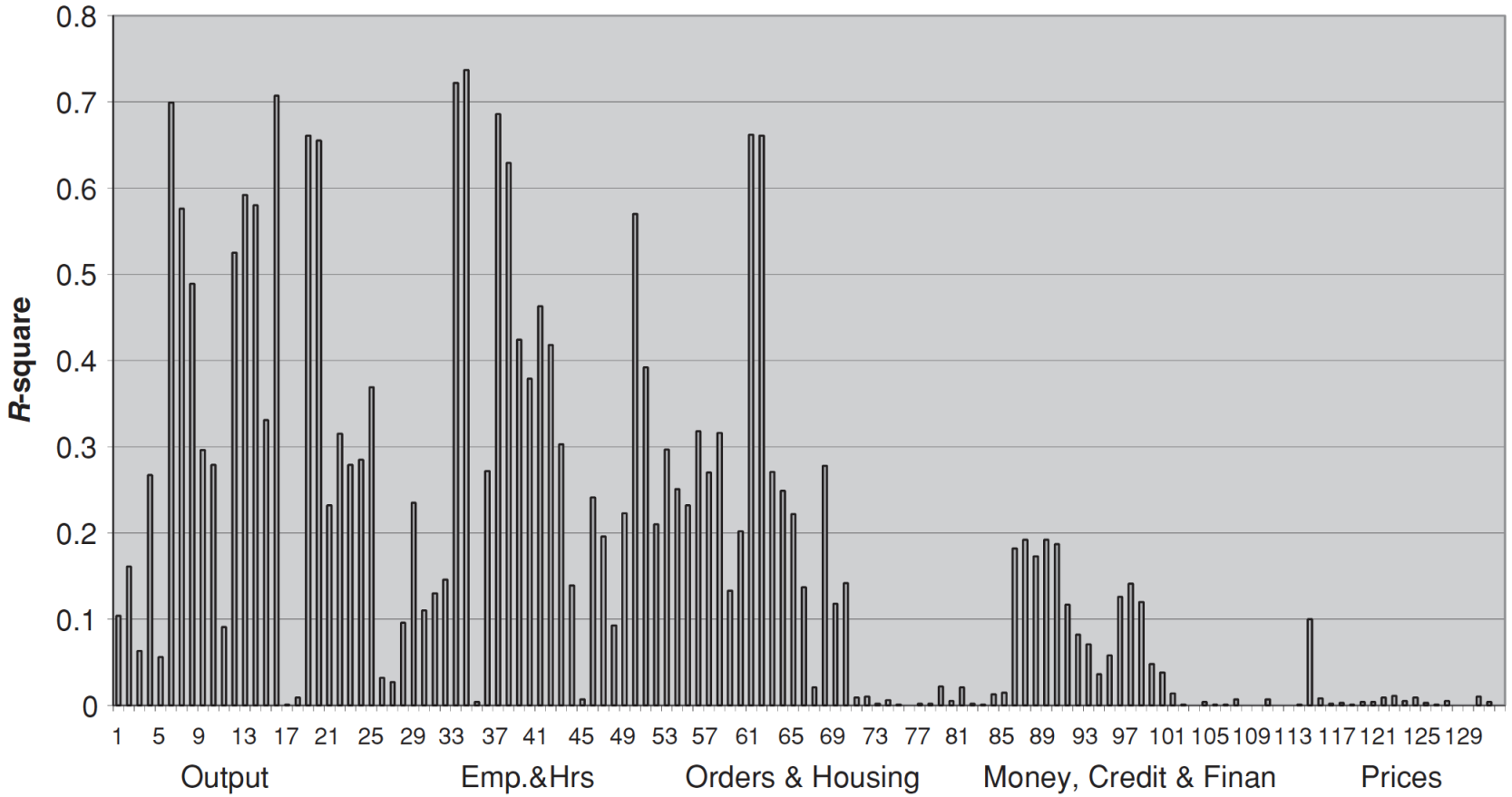


Figure 1
Marginal *R*-squares for F_1
Note: Chart shows the *R*-square from regressing the series number given on the *x*-axis onto F_1. See the Appendix for a description of the numbered series. The factors are estimated using data from 1964:1 to 2003:12.

Forecasting bond excess returns

Ludvigson and Ng (2009)

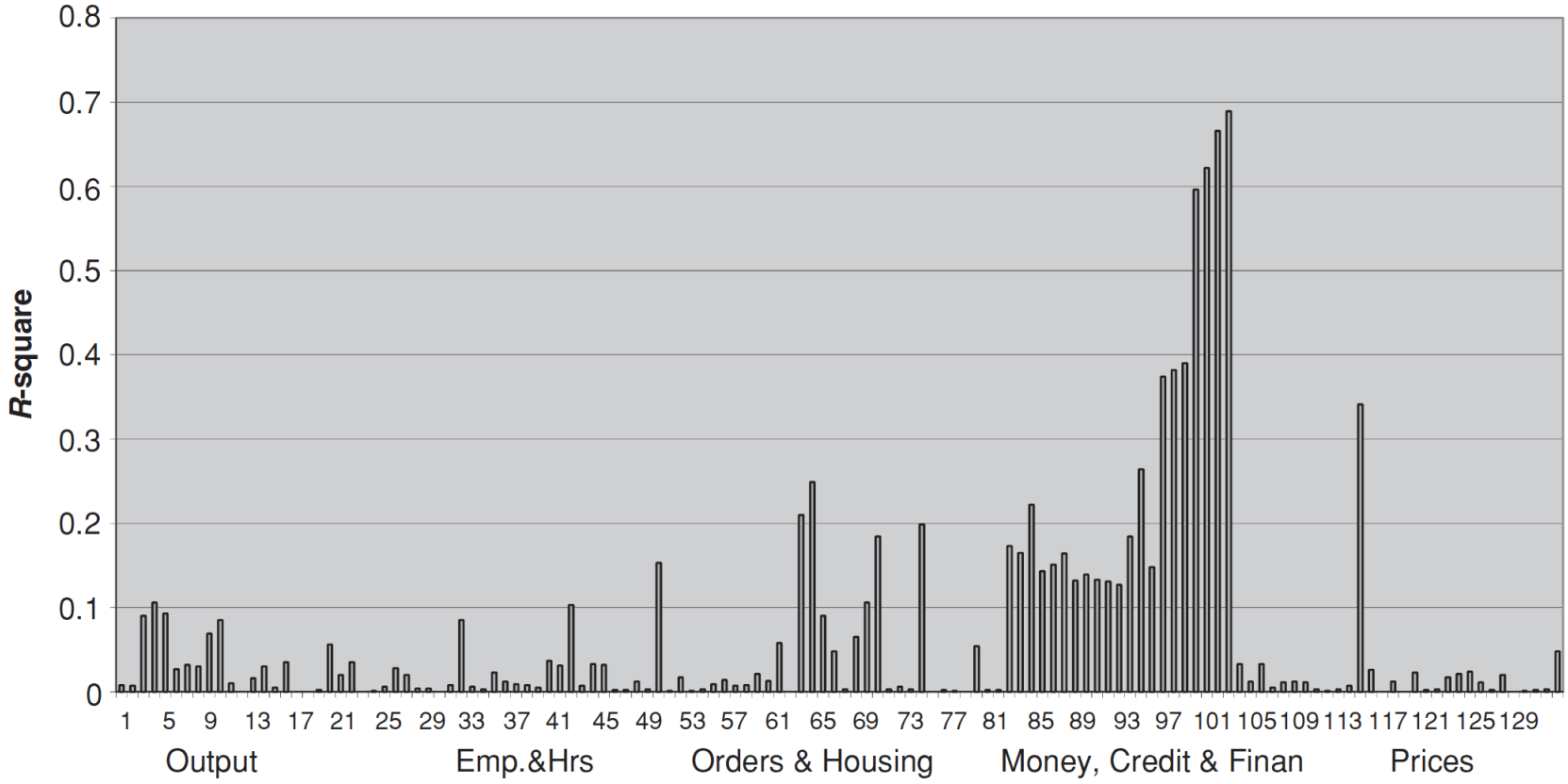


Figure 2
Marginal *R*-squares for F_2
Note: See Figure 1.

Forecasting bond excess returns

Ludvigson and Ng (2009)

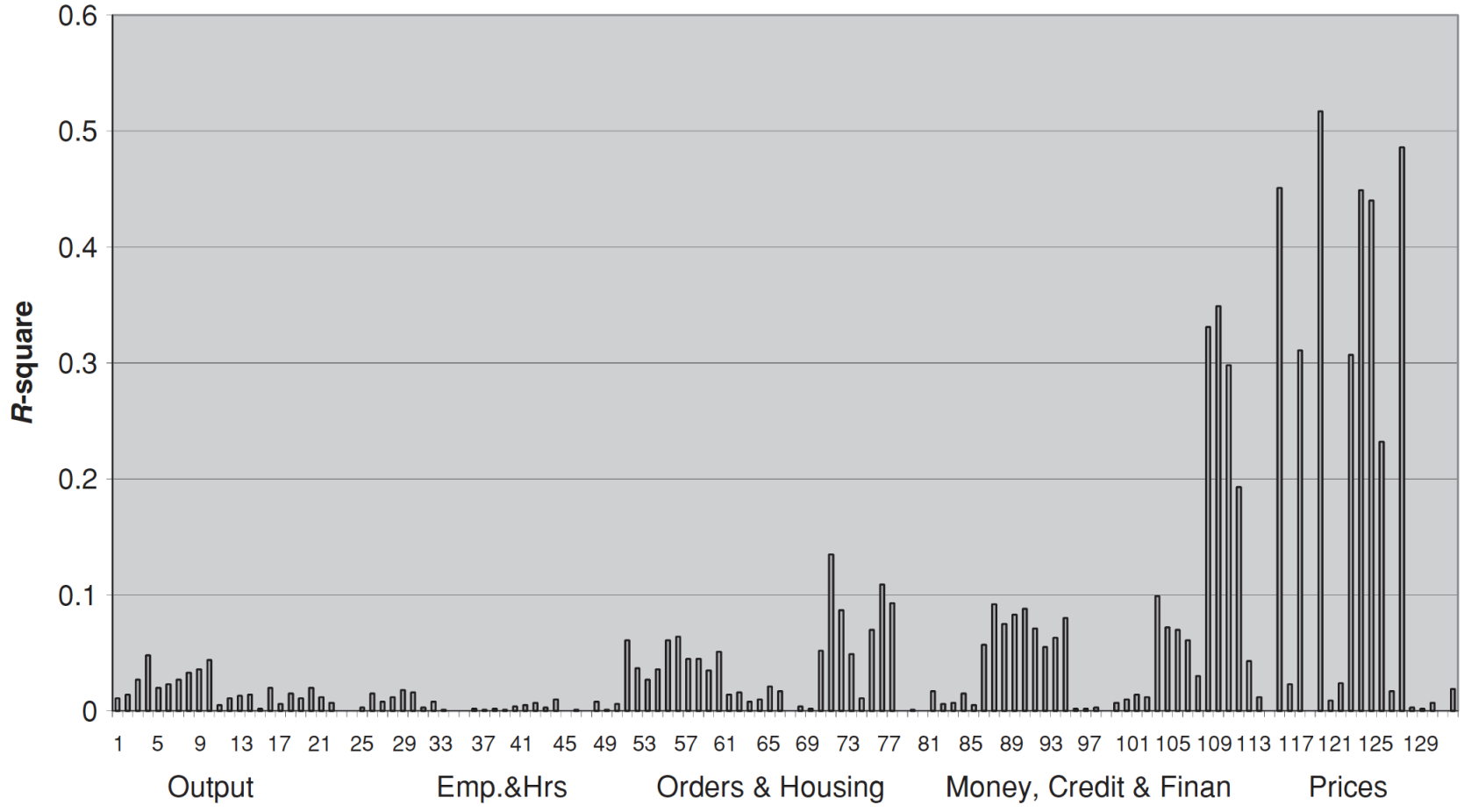


Figure 3
Marginal *R*-squares for F_3
Note: See Figure 1.

Wrap-up

- Regularization as a tool for variable selection.
- Ridge Regressions and LASSO as a tool for variable selection.
- Introduced Data compression Methods: Principal Components Analysis.
- Next lecture:
  1. Tree Models.
  2. Cross-validation.
  3. Equity Return Prediction and Asset Management.

References

1. Agarwal, Vikas, and Narayan Y. Naik. "Risks and portfolio decisions involving hedge funds." The Review of Financial Studies 17.1 (2004): 63-98.
2. Chinco, Alex, Adam D. Clark-Joseph, and Mao Ye. "Sparse signals in the cross-section of returns." The Journal of Finance 74.1 (2019): 449-492.
3. Ludvigson, Sydney C., and Serena Ng. "Macro factors in bond risk premia." The Review of Financial Studies 22.12 (2009): 5027-5067.