MACROCONFERENCE

The **Macro**Conference
Proceedings

Econometric Modeling with High-dimensional Data in Business and Economics

Sunil Sapra
Department of Economics and Statistics, California State University, Los Angeles, CA 90032, USA.

Abstract

High-dimensional data is becoming increasingly common in business and economics. Conventional econometric methods produce poor forecasts and are unreliable for statistical inference due to overfitting in such scenarios. The paper focuses on empirical applications of recently developed econometric techniques for forecasting and inference with high-dimensional data in business and economics. Shrinkage techniques, including least absolute shrinkage and selection operator (LASSO) algorithm and its variants, dimension reduction techniques, including principal components analysis, partial least squares, and best subset selection technique for linear regression models are studied. The econometric techniques employed in the paper have been extended to generalized linear models in recent years and are widely applicable to the analysis of count, binary response and duration types of data encountered in business and social sciences. Our empirical applications demonstrate that these techniques perform well for prediction as well as inference with high-dimensional data.

Keywords: Big Data; Econometric Analysis; Dimension Reduction; LASSO; Regression Models.

1. Introduction

Modern data is characterized by thousands to millions of features on each object or individual, which Giraud (2015) refers to as high-dimensional data. In more general terms, high-dimensional data is any dataset with more variables (*p*) than the number of observations points (*n*). With high-dimensional data, the conventional forecasting methods produce poor forecasts due to overfitting and the conventional techniques for statistical inference are unreliable. The purpose of this paper is to study empirical applications involving building of predictive models as well as methods for statistical inference with high-dimensional data. High-dimensional models have recently gained considerable importance in several areas of economics. For

example, the vector autoregressive (VAR) model (Sims 1980, Stock & Watson 2001) is a key technique for analyzing the joint evolution of macroeconomic time series, which can provide a large amount of structural information. Since the number of parameters grows rapidly with the size of the model, standard VAR models usually include no more than 10 variables. However, econometricians may observe hundreds of data series. In order to enrich the model information set, Bernanke et al. (2005) proposed to augment standard VAR models with estimated factors (FAVAR) to measure the effects of monetary policy. Factor analysis also plays an important role in forecastingusing large dimensional data sets (see Stock & Watson 2006, Bai & Ng 2008).

Another example of high dimensionality is large panels of home-price data. To incorporate cross-sectional effects, one may consider that the price in one county depends on several other counties, most likely its geographic neighbors. Because such correlation is unknown, initially the regression equation may include about 1,000 counties in the United States, which makes direct ordinary least-squares (OLS) estimation impossible. A common technique for reducing dimensionality is variable selection. Recently, statisticians and econometricians have developed algorithms to simultaneously select relevant variables and estimate parameters efficiently (see Fan et al. 2011 for an overview). Variable selection techniques have also been used widely in financial portfolio construction, treatment-effects models, and credit-risk models.

Estimation of volatility matrix is a high-dimensional problem in finance. To optimize the performance of a portfolio (Campbell et al. 1997, Cochrane 2005) or to manage the risk of a portfolio, asset managers need to estimate the covariance matrix or its inverse matrix of the returns of assets in the portfolio. High dimensionality here poses challenges to the estimation of matrix parameters, as small element-wise estimation errors may result in huge errors matrix-wise. In the time domain, high-frequency financial data also provide both opportunities and challenges to high-dimensional modeling in economics and finance.

This paper is organized as follows. Section 2 presents a review of literature on high-dimensional data analysis. Section 3 presents supervised and unsupervised learning techniques, including best subset selection, dimension reduction techniques, and penalty based variable selection in regression models with many parameters. Section 4 presents a variety of applications of these techniques to quantitative and qualitative data demonstrating their usefulness in prediction and inference with high-dimensional data. Section 5 concludes.

2. Literature Review

There are three key approaches to regression modeling with high-dimensional data: shrinkage methods, dimension reduction techniques, and subset selection methods. Shrinkage methods involve fitting a model involving all *p* predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) has the effect of reducing variance. Dimension reduction techniques approach involves projecting the *p* predictors into a lower-dimensional subspace by computing different linear combinations, or projections, of these variables. These projections are then used as

predictors to fit a linear regression model by least squares. Subset selection approach involves identifying a subset of the p predictors that are believed to be related to the response. A model using least squares is fitted on the reduced set of variables. There is a burgeoning literature on these supervised and unsupervised learning techniques. Nevertheless, Gareth et al. (2014) and Hastie et al. (2009) provide a thorough discussion of these approaches. The main idea underlying predictive modeling and inference with high-dimensional data is dimension reduction or regularization. The problem is that when the number of variables is equal to or greater than the number of observations, the ordinary least squares method produces a perfect fit due to overfitting-a situation in which researchers fit both the signal and the noise to the data. This results in poor forecasts and poor inference. Regularization is needed to avoid overfitting and to obtain useful out-of-sample forecasts. This is accomplished via LASSO and its variants, which result in variable selection for improved forecasts and inference with high-dimensional data.

3. Methodology

Shrinkage Methods: Penalty-based Variable Selection in Regression Models with Many Parameters: LASSO and Ridge Regression

Shrinkage methods fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero. Hastie et al. (2009) explain that such a constraint can improve the fit since shrinking the coefficient estimates can significantly reduce their variance without significantly increasing the bias. The two best-known techniques for shrinking the regression coefficients towards zero are ridge regression and the LASSO. These methods are discussed next.

Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge Regression

We focus on the linear model

$y = X\beta + \varepsilon$ and assume for simplicity that the columns of X have $l^2 -$ norm 1.

The LASSO estimator is defined as

$\hat{\beta}^{LASSO} = argmin_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \beta_j x_{ij}\right)^2$ subject to $\sum_{j=1}^{k} |\beta_j| \leq t$

Where $t \geq 0$ is a tuning parameter that controls the shrinkage that is being applied to the estimates.

The LASSO optimization problem can be written in the equivalent Lagrangian form

$$\hat{\beta}^{LASSO} = argmin_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j|.$$

Similarly, the ridge estimator is defined as

$$\hat{\beta}^{RIDGE} = argmin_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{k} \|\beta\|^2.$$

Dimension Reduction Techniques

This approach involves projecting the *p* predictors into a *M*-dimensional subspace, where *M* <*p*. This is achieved by computing *M* different linear combinations, or projections, of the variables. Then these *M* projections are used as predictors to fit a linear regression model by least squares.

Best Subset Selection Algorithm

Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.

Step 1: For k =1,2.,,,.p:

(a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

(b) Pick the best model among these $\binom{p}{k}$ models and call it M_k. The best model is defined as the model with the smallest residual sum of squares (RSS), or largest R^2.

Step 2: Select the best model among $M_0, M_1,\ldots,M_p$ using cross-validated prediction errors, C_p, AIC, BIC, or adjusted R^2.

4. Empirical Applications

Application of LASSO and Ridge Regression to Unemployment Data for 50 States in the US

We analyze monthly seasonally adjusted unemployment rates over the period January 1976 through August 2010 for the 50 US states taken from Ledolter (2013). As in Ledolter (2013), we treat this data set as a multivariate (50-variate) time series spanning 416 periods. Vector autoregressive (VAR) models are commonly used to predict vector time series. For VARs, the observations (unemployment rates) of a certain state (say the first state) on the previous (lagged) unemployment rates of that state, as well as those of all other states-a huge increase in the dimensions of the data. The order of the VAR tells us how many lags to consider.

LASSO Coefficient Estimates: The following is only a partial list of coefficients to illustrate that a large number of coefficients have been shrunk to zero via LASSO resulting in variable selection.

(Intercept)	X1	X2	X3	X4
0.0364485503	0.0000000000	0.9452388992	0.0000000000	0.00000000

X5	X6	X7	X8	X9
0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000

X10	X11	X12	X13	X14
0.0000000000	0.0000000000	0.0214941756	0.0000000000	0.0060662062

X15	X16	X17	X18	X19
0.0	0.00000000	0.00000000	0.00000000	0.00000000

X20	X21	X22	X23	X24
0.0011793817	0.0000000000	0.0190052216	0.0160142689	0.00000000

X25	X26	X27	X28	X29
0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000

Ridge Regression Coefficient Estimates: The following is only a partial list of coefficients to illustrate that a very small number of coefficients have been shrunk to zero via Ridge Regression resulting in almost no variable selection.

(Intercept)	X1	X2	X3	X4
6.3539991478	0.0000000000	0.0002160775	0.0002697094	0.0002385934

X5	X6	X7	X8	X9
0.0003351591	0.0001967763	0.0002980288	0.0001548183	0.0001997160

X10	X11	X12	X13	X14
0.0002153148	0.0002902588	0.0001322214	0.0003286162	0.0002627515

X15	X16
0.0002157248	0.0003220925

Note that the ridge regression has shrunk just one coefficient to zero unlike LASSO, which shrank several coefficients to zero. Thus, as far as variable selection is concerned, LASSO outperforms the ridge regression in this application.

Application of Dimension Reduction Techniques to Fuel Efficiency Data

A data set on the fuel efficiency of 38 cars from Ledohlter (2013) is used. The fuel efficiency measured in GPM (gallons per 100 miles) as a function of the weight of the car in 1000 lbs), cubic displacement (in cubic inches), number of cylinders, horsepower, acceleration (in seconds from 0 to 60 mph), and engine size (V type and straight (coded as 1)). As in Ledohlter (2013), GPM is analyzed instead of the usual EPA fuel efficiency measure MPG since the reciprocal transformation GPM = 100/MPG leads to approximate linear relationships between the response and the predictors.

We perform principal components regression and partial least squares regression on test data and evaluate its performance on test data.

Principal Components Regression Results

Data: X dimension: 19 6
Y dimension: 19 1
Fit method: svdpc
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
CV	1.267	0.7246	0.6670	0.4183	0.4063	0.3101	0.2651
adjCV	1.267	0.7200	0.6605	0.4109	0.3978	0.3029	0.2598

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
X	72.52	92.22	97.74	99.39	99.89	100.00
GPM	71.14	78.96	93.73	95.43	97.49	97.91

MSE of Prediction = 0.2532946

Partial Least Squares Regression Results

PARTIAL LEAST SQUARES USING TEST DATA ONLY

Data: X dimension: 19 6
Y dimension: 19 1
Fit method: kernelpls
Number of components considered: 6
VALIDATION: RMSEP
Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
CV	1.267	0.6446	0.5068	0.3296	0.3198	0.2686	0.2825
adjCV	1.267	0.6421	0.5043	0.3258	0.3142	0.2642	0.2762

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
X	72.01	89.44	97.65	99.18	99.88	100.00
GPM	77.17	89.11	95.27	96.94	97.71	97.91

MSE of Prediction = 0.2077569

With both principal components and partial least squares regression, 4 components capture almost all of the variance. These results show that partial least squares, a supervised learning technique outperforms the principal components regression, an unsupervised learning technique on test data in terms of mean squared error of prediction for out-of-sample forecasts.

Application of Best Subset Selection Method to Fuel Efficiency Data

We now apply the best subset selection method to the fuel efficiency data of the previous section. The dependent variable is GPM and the predictors are WT, DIS, NC, HP, ACC, and ET.

The following results display the best models selected for various model sizes. The best two-variable model contains WT and DIS only, the best five-variable model contains WT, DIS, NC, HP, and ET, and so on. An asterisk means a given variable is included in the model.

- 1 subset of each size **up to** 6
- Selection Algorithm: exhaustive

BEST SUBSETS

	WT	DIS	NC	HP	ACC	ET
1	"*"	" "	" "	" "	" "	" "
2	"*"	"*"	" "	" "	" "	" "
3	" "	" "	"*"	"*"	" "	"*"
4	" "	" "	"*"	"*"	"*"	"*"
5	"*"	"*"	"*"	"*"	" "	"*"
6	"*"	"*"	"*"	"*"	"*"	"*"

5. Conclusions

This paper has presented applications of some recently developed techniques for sparse high-dimensional modeling in business and economics. We focused on best subset selection methods, dimension reduction methods, and regularization methods including LASSO and Ridge regression. Most of the methods studied in the paper are supervised learning methods with a response variable and predictors. All of the applications to high-dimensional data presented illustrate various approaches to reducing dimensionality and mitigating the problems of collinearity for high-dimensional linear regression and can be extended to generalized linear models (GLMs) and hazard regressions. These methods look very promising for improving forecasts and inference in ultra-high dimensional settings.

REFERENCES

Bai J, Ng S. 2008. Large dimensional factor analysis. Found. Trends *Econometrica* 3(2):89–163

Bernanke B, Boivin J, Eliasz PS. 2005. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *Quarterly Journal of Economics* 120 (1): 387–422

Campbell J, Lo A, MacKinlay C. 1997. *The Econometrics of Financial Markets*. Princeton, NJ: Princeton Univ. Press

Cochrane, J. 2001, *Asset Pricing*. Princeton, NJ: Princeton University Press, 2001

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least angle regression. *Annals of Statistics* 32:407–99

Jianqing Fan, Jinchi Lv, and Lei Qi 2011. Sparse High-Dimensional Models in Economics, *Annual Review of Economics* 3:291–317

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference*, and Prediction, Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2014), *An Introduction to Statistical Learning with Applications in R*, Springer.

Ledolter, Johannes (2013), *Data Mining and Business Analytics with R*, Wiley: New York

Sims CA. 1980. Macroeconomics and reality. *Econometrica* 48(1):1–48

Stock, J and Watson, M. 2002 Forecasting using principal components from a large number of predictors. *J Am Stat Assoc* 97: 1167–79.