1
Introduction

Variable selection is an important matter in several statistical problems, for which many different approaches have been proposed. Traditionally, one can chooses the set of explanatory variables using information criteria or prior information, but the total number of models to evaluate increases exponentially as the number of candidate variables increases. One additional problem is the presence of more candidate variables than observations.

Many solutions and automatic variable selection techniques with different approaches have been proposed in the last few years. Although these approaches have been mostly applied to economic data, either for forecasting or identification of the relevant variables of the model, it can be useful for modeling and forecasting any kind of data.

In this thesis we study several aspects of the variable selection problem. First, we compare two procedures for linear regression in a simulation exercise and an application to GDP forecasting. We compare *Autometrics*, which is a general-to-specific (GETS) approach, LASSO, a shrinkage method, and the adaptive LASSO (adaLASSO). In the simulation exercise, different scenarios were contemplated and the comparison considers the predictive power (forecast out-of-sample) and the performance in the correct model selection and estimation (in-sample).

In a second part, we introduce a variable selection methodology for smooth transition regressive (STR) and autoregressive (STAR) models based on LASSO regularization. We propose a direct and a stepwise approach, that we called Group Stepwise LASSO. Both methodologies are tested with extensive simulation exercises and an application to genetic data.

Finally, we introduce a penalized least square criterion based on the LASSO ℓ_1-penalty and the CVaR (Conditional Value at Risk) of the "out-of-sample" regression errors. This is a quadratic optimization problem solved by interior point methods. In a simulation study with linear regression models, we

show that this method outperforms LASSO and adaLASSO in forecasting when the data is contaminated by outliers, showing to be a robust method of estimation and variable selection.

A more detailed introduction with bibliographical references and the study of the literature is given at the beginning of each chapter.

The thesis is organized as follows: Chapter 2 presents the comparison of *Autometrics* and LASSO methodologies, simulation results and application to GDP forecasting; Chapter 3 proposes a new methodology for variable selection in STR/STAR models, simulation results and application to genetic data; Chapter 4 introduces a robust variable selection method called LASSO-CVaR, some mathematical and simulation results; finally Chapter 5 concludes.

1.1. Contributions

The main contributions of this thesis are:

1. Comparison of two of the most applied model selection techniques in the literature for linear regression models: LASSO and *Autometrics*.

2. LASSO adaptation for variable selection in smooth transition regressive (STR) and autoregressive (STAR) models.

3. New regularization method based on LASSO and CVaR penalty, interesting for variable selection when the data is contaminated by outliers.