Pontifícia Universidade Católica
do Rio de Janeiro

Camila Rosa Epprecht

Variable selection for linear and smooth transition models via LASSO: comparisons, applications and new methodology

Tese de Doutorado

Thesis presented to the Programa de Pós-Graduação em Engenharia Elétrica of the Departamento de Engenharia Elétrica, PUC-Rio as partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica

Advisor: Prof. Álvaro de Lima Veiga Filho

Co-Advisor: Prof. Dominique Guégan

Rio de Janeiro
October 2013

Pontifícia Universidade Católica
do Rio de Janeiro

Camila Rosa Epprecht

Variable selection for linear and smooth transition models via LASSO: comparisons, applications and new methodology

Thesis presented to the Programa de Pós-Graduação em Engenharia Elétrica of the Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio, as partial fulfillment of the requirements for the degree of Doutor

Prof. Álvaro de Lima Veiga Filho
Advisor
Departamento de Engenharia Elétrica - PUC-Rio

Prof. Marcelo Medeiros
Departamento de Economia - PUC-Rio

Prof. Cristiano Fernandes
Departamento de Engenharia Elétrica - PUC-Rio

Prof. Marcelo Fernandes
School of Economics and Finance - Queen Mary University of London
São Paulo School of Economics - FGV

Prof. Joel Corrêa da Rosa
Departamento de Estatística - UFF

Prof. José Eugenio Leal
Coordinator of the Centro
Técnico Científico - PUC-Rio

Rio de Janeiro, October 17th, 2013

Camila Rosa Epprecht

The author graduated (2006) and has a Master's degree (2008) in Electrical Engineering at PUC-Rio. Her concentration area is decision support systems - statistical models and optimization. In 2011, she spent one year in the Center of Economics of University Paris 1 Panthéon-Sorbonne, Paris, doing a part of her thesis under the supervision of her co-advisor. In 2013, she obtained the French degree of PhD in Applied Mathematics at Université Paris 1 Panthéon-Sorbonne under an agreement for international co-direction of a PhD thesis.

Bibliographic data

Epprecht, Camila

Variable selection for linear and smooth transition models via LASSO: comparisons, applications and new methodology / Camila Rosa Epprecht ; advisor: Álvaro de Lima Veiga Filho ; co-advisor: Dominique Guégan– 2013.
115 f. : il. (color.) ; 30 cm

Tese (Doutorado em Engenharia Elétrica)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2013.
Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Seleção de modelos. 3. Seleção de variáveis. 4. *Autometrics.*, 5. LASSO. 6. Propriedade de oráculo. 7. Modelos de transição suave. 8. Dados genéticos. 9. CVaR. I. Veiga Filho, Álvaro de Lima. II. Guégan, Dominique. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Agradecimentos

Ao meu orientador Prof. Álvaro Veiga, pela valiosa orientação, disponibilidade, paciência, apoio, incentivo e amizade em todos os momentos. Agradeço aos inúmeros conselhos e ensinamentos que serviram como guia.

À minha co-orientadora Prof. Dominique Guégan, pelo apoio e orientação, e por me receber em seu centro de pesquisa na Université Paris 1 Panthéon-Sorbonne, Paris, França, onde fiz meu doutorado sanduiche.

Ao CNPq e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Ao Prof. Joel Corrêa da Rosa e Mayte Suárez-Fariñas da Rockefeller University, New York, USA, por disponibilizar a base de dados genéticos.

Ao Prof. Marcelo Medeiros, pela ajuda, ideias e discussões.

Ao Alexandre Moreira, Mario Souto, Joaquim Dias Garcia e Alvaro Gustavo Talavera, colegas que estiveram disponíveis para ajudar em todos os momentos, com discussões e questões técnicas.

À secretária do DEE da PUC-Rio, Alcina Portes, por toda a orientação, paciência e amizade.

À equipe do suporte do DEE da PUC-Rio, Luis Fernando, Isnard e Danilo, que sempre estiveram disponíveis e dispostos a resolver qualquer problema técnico.

Ao meu pai, por todo o seu apoio e amor, discussões e orientações técnicas.

À minha mãe por todo o seu apoio moral, companhia e amor.

Aos meus amigos no Brasil e aos muitos que fiz em Paris.

Ao meu namorado pela paciência, apoio e amor.

Abstract

Epprecht, Camila Rosa; Veiga Filho, Álvaro de Lima. **Variable selection for linear and smooth transition models via LASSO: comparisons, applications and new methodology**. Rio de Janeiro, 2013. 115p. PhD. Thesis - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Variable selection in statistical models is an important problem, for which many different solutions have been proposed. Traditionally, one can choose 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. In this thesis we study several aspects of the variable selection problem. First, we compare two procedures for linear regression: *Autometrics*, which is a general-to-specific (GETS) approach based on statistical tests, and LASSO, a shrinkage method. Different scenarios were contemplated for the comparison in a simulation experiment, varying the sample size, the number of relevant variables and the number of candidate variables. In a real data application, we compare the methods for GDP forecasting. In a second part, we introduce a variable selection methodology for smooth transition regressive (STR) and autoregressive (STAR) models based on LASSO regularization. We present a direct and a stepwise approach. Both methods 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 in a linear regression framework, we show that the proposed method outperforms the LASSO when the data is contaminated by outliers, showing to be a robust method of estimation and variable selection.

Keywords

Model selection; variable selection; *Autometrics*; LASSO; adaLASSO; oracle property; smooth transition models; interactions; genetic data; CVaR

Resumo

Epprecht, Camila Rosa; Veiga Filho, Álvaro de Lima. **Seleção de variáveis para modelos lineares e de transição suave via LASSO: comparações, aplicações e nova metodologia.** Rio de Janeiro, 2013. 115p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

A seleção de variáveis em modelos estatísticos é um problema importante, para o qual diferentes soluções foram propostas. Tradicionalmente, pode-se escolher o conjunto de variáveis explicativas usando critérios de informação ou informação à priori, mas o número total de modelos a serem estimados cresce exponencialmente a medida que o número de variáveis candidatas aumenta. Um problema adicional é a presença de mais variáveis candidatas que observações. Nesta tese nós estudamos diversos aspectos do problema de seleção de variáveis. No Capítulo 2, comparamos duas metodologias para regressão linear: *Autometrics*, que é uma abordagem geral para específico (GETS) baseada em testes estatísticos, e LASSO, um método de regularização. Diferentes cenários foram contemplados para a comparação no experimento de simulação, variando o tamanho da amostra, o número de variáveis relevantes e o número de variáveis candidatas. Em uma aplicação a dados reais, os métodos foram comparados para a previsão do PIB dos EUA. No Capítulo 3, introduzimos uma metodologia para seleção de variáveis em modelos regressivos e autoregressivos de transição suave (STR e STAR) baseada na regularização do LASSO. Apresentamos uma abordagem direta e uma escalonada (*stepwise*). Ambos os métodos foram testados com exercícios de simulação exaustivos e uma aplicação a dados genéticos. Finalmente, no Capítulo 4, propomos um critério de mínimos quadrados penalizado baseado na penalidade ℓ_1 do LASSO e no CVaR (*Conditional Value at Risk*) dos erros da regressão *out-of-sample*. Este é um problema de otimização quadrática resolvido pelo método de pontos interiores. Em um estudo de simulação usando modelos de regressão linear, mostra-se que o método proposto apresenta performance superior a do LASSO quando os dados são contaminados por *outliers*, mostrando ser um método robusto de estimação e seleção de variáveis.

Palavras-chave

Seleção de modelos; seleção de variáveis; *Autometrics*; LASSO; adaLASSO; propriedade de oráculo; modelos de transição suave; interações; dados genéticos; CVaR

Sumário