UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE CIÊNCIAS ECONÔMICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

EVANDRO KONZEN

PENALIZAÇÕES TIPO LASSO NA SELEÇÃO DE COVARIÁVEIS EM SÉRIES TEMPORAIS

Porto Alegre
2014

EVANDRO KONZEN

PENALIZAÇÕES TIPO LASSO NA SELEÇÃO DE COVARIÁVEIS EM SÉRIES TEMPORAIS

Dissertação submetida ao Programa de Pós-Graduação em Economia da Faculdade de de Ciências Econômicas da UFRGS, como quesito parcial para obtenção do título de Mestre em Economia, com ênfase em Economia Aplicada.

Orientador: Prof. Dr. Flávio A. Ziegelmann

Porto Alegre

2014

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Responsável: Biblioteca Gládis W. do Amaral, Faculdade de Ciências Econômicas da UFRGS

Konzen, Evandro
Penalizações tipo LASSO na seleção de covariáveis em séries temporais / Evandro Konzen. -- 2014.
47 f.

Orientador: Flávio Augusto Ziegelmann.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Ciências Econômicas, Programa de Pós-Graduação em Economia, Porto Alegre, BR-RS, 2014.

1. Séries temporais. 2. LASSO. 3. Seleção de variáveis. 4. Previsão. I. Ziegelmann, Flávio Augusto, orient. II. Título.

EVANDRO KONZEN

PENALIZAÇÕES TIPO LASSO NA SELEÇÃO DE COVARIÁVEIS EM SÉRIES TEMPORAIS

Dissertação submetida ao Programa de Pós-Graduação em Economia da Faculdade de de Ciências Econômicas da UFRGS, como quesito parcial para obtenção do título de Mestre em Economia, com ênfase em Economia Aplicada.

Aprovada em: Porto Alegre, 22 de agosto de 2014.

BANCA EXAMINADORA:

Prof. Dr. Flávio A. Ziegelmann - orientador
PPGE/UFRGS

Prof. Dr. Marcio Valk
DEST/UFRGS

Prof. Dr. Erik Alencar de Figueiredo
UFPB

Prof. Dr. Paulo de Andrade Jacinto
PUC-RS

Agradecimentos

Agradeço ao CNPq pela bolsa de estudos que permitiu a realização do curso de mestrado.

Agradeço aos professores da UFRGS, em especial do Programa de Pós-Graduação em Economia (PPGE) e do Departamento de Estatística (DEST) que sempre se dispuseram a ajudar desde a minha graduação.

Agradeço ao professor Flávio Augusto Ziegelmann, pelo apoio constante, pelos conhecimentos adquiridos e pela excelente orientação deste trabalho.

Aos meus parentes, colegas e ótimos amigos, pela consideração quando precisei estar ausente.

Aos meus pais, pela educação, carinho e compreensão.

À minha companheira Eliza, por todo seu amor, confiança e apoio constante .

Resumo

Este trabalho aplica algumas formas de penalização tipo LASSO aos coeficientes para reduzir a dimensionalidade do espaço paramétrico em séries temporais, no intuito de melhorar as previsões fora da amostra. Particularmente, o método denominado aqui como WLadaLASSO atribui diferentes pesos para cada coeficiente e para cada defasagem. Nas implementações de Monte Carlo deste trabalho, quando comparado a outros métodos de encolhimento do conjunto de coeficientes, essencialmente nos casos de pequenas amostras, o WLadaLASSO mostra superioridade na seleção das covariáveis, na estimação dos parâmetros e nas previsões. Uma aplicação a séries macroeconômicas brasileiras também mostra que tal abordagem apresenta a melhor performance de previsão do PIB brasileiro comparada a outras abordagens.

Palavras-chave: Séries temporais. LASSO. AdaLASSO. Seleção de variáveis. Previsão.

Abstract

This dissertation applies some forms of LASSO-type penalty on the coefficients to reduce the dimensionality of the parameter space in time series, in order to improve the out-of-sample forecasting. Particularly, the method named here as WLadaLASSO assigns different weights to each coefficient and lag period. In Monte Carlo implementations in this study, when compared to other shrinkage methods, essentially for small samples, the WLadaLASSO shows superiority in the covariable selection, in the parameter estimation and in forecasting. An application to Brazilian macroeconomic series also shows that this approach has the best forecasting performance of the Brazilian GDP compared to other approaches.

Keywords: Time series. LASSO. AdaLASSO. Variable selection. Forecasting.

Lista de Abreviaturas e Siglas

adaLASSO	*adaptive* LASSO
AIC	*Akaike information criterion*
ARMA	Modelo auto-regressivo de média móvel
BIC	*Bayesian information criterion*
EQM	Erro quadrático médio
EQMP	Erro quadrático médio de previsão
LASSO	*Least absolute shrinkage and selection operator*
MQO	Método de mínimos quadrados ordinários
PIB	Produto interno bruto
SQR	Soma dos quadrados dos resíduos
UMIDAS	*Unrestricted mixed data sampling*
VAM	Viés absoluto médio
WLadaLASSO	*Weighted Lag adaptive* LASSO

Sumário

1 Introdução

Modelos de alta dimensionalidade têm sido cada vez mais presentes na literatura. Sabe-se que a inclusão de um grande número de variáveis econômicas e financeiras pode contribuir para ganhos substanciais para a previsão das séries. Como argumentam Song e Bickel (2011), um problema desafiador é determinar quais variáveis e quais defasagens são relevantes, especialmente quando há uma mistura de correlação serial (dinâmica temporal), alta dimensionalidade na estrutura de dependência entre as variáveis e um tamanho de amostra pequeno (relativamente à dimensionalidade).

Conforme Fan e Lv (2010), a acurácia estatística, a capacidade de interpretação do modelo e a complexidade computacional são três pilares importantes de qualquer procedimento estatístico. Tipicamente, o número de observações n é muito superior ao número de variáveis ou parâmetros p. Nesses casos, nenhum dos três aspectos precisa ser sacrificado para se obter a eficiência dos demais. Entretanto, quando a dimensionalidade p é grande comparativamente ao tamanho da amostra n, os métodos tradicionais enfrentam alguns desafios. Dentre eles, como tornar os modelos estimados interpretáveis; como fazer com que os procedimentos estatísticos sejam robustos e computacionalmente eficientes; e como obter procedimentos com maior eficiência em termos de inferência estatística.

Além disso, em um contexto de alta dimensionalidade, quando o número de variáveis p for grande comparativamente ao tamanho da amostra n, os modelos tradicionais podem apresentar problemas de correlação espúria entre as covariáveis, que pode ser alta mesmo quando as covariáveis forem independentes e identicamente distribuídas, como mostram Fan e Lv (2008) e Fan, Guo e Hao (2012).

Uma das formas de se desafiar os problemas causados pela alta dimensionalidade é através da suposição de esparsidade do vetor de parâmetros de dimensão p, considerando que muitos de seus componentes são exatamente iguais a zero. Apesar de geralmente produzir estimativas viesadas, a suposição de esparsidade contribui na identificação das covariáveis importantes, na obtenção de um modelo mais parcimonioso, na redução da complexidade do modelo e na diminuição do custo computacional.

Dentre os estudos macroeconômicos no contexto multidimensional, tem-se como exemplos os artigos seminais de Sims (1980) e de Stock e Watson (2001), que empregam o modelo vetorial auto-regressivo (VAR) para analisar a coevolução das séries temporais macroeconômicas. Entretanto, como o número de parâmetros de um modelo VAR cresce quadraticamente com o número de variáveis, torna-se inadequada a utilização do VAR para o caso de muitas variáveis. Como os macroeconometristas podem ter o interesse de avaliar conjuntamente centenas de séries temporais, algumas alternativas de modelos começaram a surgir.

A análise fatorial tem sido empregada como uma das alternativas, como nos trabalhos de Stock e Watson (2006), Ng e Moench (2009) e Stock e Watson (2008). Como mencionam Medeiros e Mendes (2012), modelos fatoriais constituem uma boa alternativa quando a maior parte das variáveis manifesta importância no modelo, situação que os autores denominam por estrutura densa do modelo. Já quando a matriz de coeficientes é esparsa, os métodos que supõem esparsidade ganham importância. Clark e McCracken (2008) é uma das referências que discutem abordagens distintas de redução de dimensionalidade do conjunto dos coeficientes, como *Bayesian shrinkage*, no contexto de previsão de séries macroeconômicas.

Tibshirani (1996) propôs o LASSO (*least absolute shrinkage and selection operator*) no contexto de regressão linear, onde ele impõe uma penalização ao conjunto de norma L_1 dos coeficientes. Devido à natureza dessa penalização, o LASSO tende a zerar alguns coeficientes, tornando-se útil para selecionar covariáveis e para reduzir a dimensionalidade do espaço paramétrico. Tal metodologia é um exemplo das técnicas de regularização caracterizadas por Breiman (1998), que considera uma função de erro dada por

$$E = \text{medida de erro} + \lambda \cdot \text{complexidade do modelo}$$

onde a soma dos quadrados dos resíduos pode representar a medida de erro.

O segundo termo penaliza os modelos cujas complexidade e variância dos estimadores são altas, com λ representando o quão severa é a penalidade. Quando se minimiza a função de erro ao invés de apenas minimizar a medida de erro, penalizam-se modelos complexos e então reduz-se a variância dos estimadores. Se λ for muito grande, apenas

modelos muito simples são obtidos e pode ser introduzido um grande viés. Usualmente λ é otimizado via *cross-validation*.

O LASSO é uma das técnicas de regularização mais famosas, obtendo sucesso essencialmente nos casos onde existem muitos coeficientes nulos dentro do conjunto de coeficientes a serem estimados. Assim, o LASSO também se torna útil para a seleção de variáveis explicativas.

Zou (2006) investiga as propriedades *oracle* mencionadas por Fan e Li (2001) para o LASSO original, proposto por Tibshirani (1996). Zou (2006) mostra que existem casos nos quais o LASSO não é consistente na seleção de variáveis; então, propõe o adaLASSO (*adaptive LASSO*), onde a penalização ocorre com diferentes pesos para cada coeficiente, o que faz com que essa versão desfrute das propriedades *oracle*.

No contexto de séries temporais, a penalização LASSO é empregada em alguns trabalhos, como em Mol, Giannone e Reichlin (2008), em Hua (2011) e em Li (2012), enquanto a penalização adaLASSO é utilizada no *working paper* de Medeiros e Mendes (2012).

O presente trabalho analisa, principalmente através de estudos de simulação, se a penalização diferente para cada defasagem (como sugerido por Park e Sakaori (2012)) contribui para a seleção das variáveis, para a estimação dos parâmetros e para a performance de previsão fora da amostra. A abordagem foi denominada como WLadaLASSO (*Weighted Lag adaptive LASSO*). Os resultados mostram superioridade do WLadaLASSO quando comparado aos métodos de penalização Ridge, LASSO e adaLASSO, principalmente no caso de tamanhos de amostra pequenos.

Também é realizada uma aplicação para séries macroeconômicas brasileiras de diferentes frequências. O Produto Interno Bruto (PIB) é divulgado trimestralmente, enquanto muitas variáveis que podem ajudar a explicá-lo possuem frequência mensal. Com isso, utilizou-se a abordagem UMIDAS (*Unrestricted Mixed Data Sampling*), que permite o emprego de séries com frequências distintas. Tal abordagem é de difícil implementação em problemas de alta dimensionalidade do espaço paramétrico, como é o caso da aplicação deste trabalho, o que motivou o uso dos métodos de penalização aos coeficientes. Os resultados novamente mostraram a melhor performance de previsão do WLadaLASSO. Além

disso, o método selecionou variáveis indicadoras para a economia no curto prazo.

2 Seleção de variáveis

Um dos objetivos da análise de regressão é estimar os coeficientes no modelo linear

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \ , \qquad i = 1, \ldots, n. \tag{2.1}$$

Na forma vetorial, pode-se escrever o modelo como

$$y_i = \beta_0 + X_i^T \boldsymbol{\beta} + \varepsilon_i \ , \tag{2.2}$$

onde $y_i \in \mathbb{R}$ é a variável resposta, $X_i = (x_{1i}, \ldots, x_{ki})^T \in \mathbb{R}^k$ é o conjunto de preditores, $\varepsilon_i \sim N(0, \sigma^2)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ é o conjunto de parâmetros e β_0 é uma constante.

O método de mínimos quadrados ordinários (MQO), provavelmente o mais popular para estimar os parâmetros de (2.1), é baseado na minimização da soma dos quadrados dos resíduos (SQR):

$$\hat{\boldsymbol{\beta}} = \underset{\beta_0, \beta_1, \ldots, \beta_k}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 . \tag{2.3}$$

Entretanto, o método de MQO sofre com alguns problemas quando existem muitos preditores no modelo. Em primeiro lugar, as suas estimativas frequentemente têm baixo viés mas grande variância, o que prejudica a precisão das previsões. Desse modo, encolher o conjunto de coeficientes, ou forçá-los a zero, permitindo um pouco de viés para reduzir a variância das previsões, pode ajudar na precisão das previsões. Em segundo lugar, gostaria-se de determinar um conjunto menor de preditores que exiba os maiores efeitos quando objetivamos descobrir quais preditores que mais explicam a variabilidade da variável resposta. Por construção, o método de MQO não pode ser implementado quando o número de parâmetros a estimar é superior ao número de observações.

2.1 *Subset selection*

Um procedimento natural para escolher quais preditores entram no modelo (2.2) é computar todos 2^k possíveis modelos de regressão, investigando todas as combinações possíveis de preditores, e plotar a soma dos quadrados dos resíduos para cada regressão,

identificando, assim, o modelo ótimo. Também são utilizadas outras medidas para comparar modelos, como o AIC (*Akaike Information Criterion*), o BIC (*Bayesian Information Criterion*), o $\overline{R}^2$ (R^2 ajustado) e o C_p de Mallows. Como comentam Lindsey e Sheather (2010), a definição de modelo ótimo não é um consenso uniforme na literatura. O modelo ótimo é aquele que otimiza um ou mais critérios de informação, escolhidos diferentemente pelos autores.

Entretanto, se houver muitos preditores candidatos, procurar o melhor subconjunto de preditores pode exigir grandes recursos computacionais. Então, alguns procedimentos foram propostos, como o *Best-Subset Selection*, o *Forward- and Backward-Stepwise Selection* e o *Forward-Stagewise Regression*, conforme apontam Hastie, Tibshirani e Friedman (2001).

2.1.1 *Best Subset Selection*

Furnival e Wilson (1974) propõem um algoritmo eficiente para encontrar os melhores subconjuntos de preditores sem examinar todos os possíveis subconjuntos. Tal algoritmo organiza todos os possíveis modelos através de estruturas em árvore, explorando-as e descartando aqueles modelos que definitivamente não são ótimos. O procedimento encontra, para cada $k \in \{0, 1, 2, \ldots, p\}$, o subconjunto de tamanho k que fornece a menor soma dos quadrados dos resíduos. Entretanto, o procedimento é factível para um p de uma magnitude não superior a 30 ou 40, como apontam Hastie, Tibshirani e Friedman (2001).

2.1.2 *Forward Selection and Backward Elimination*

A técnica de *Forward Selection* inicia-se com nenhuma variável preditora no modelo. Para cada uma das variáveis independentes, observa-se a estatística F para concluir se tal variável entra no modelo. Se nenhuma das estatísticas F indicar a utilização de alguma variável, o algoritmo para. Caso contrário, ele adiciona a variável que tem o menor p-valor e calcula as estatísticas F novamente para as variáveis que restaram fora do modelo, e o procedimento de escolha é repetido. Assim, as variáveis são adicionadas uma a uma até que nenhuma variável produza uma estatística F significante ao entrar no modelo. Uma vez que uma variável está no modelo, ela permanecerá no mesmo.

Já a técnica de *Backward Elimination* inicia-se com um modelo que inclui todas as variáveis independentes candidatas. Em cada passo, a variável que tiver a menor contribuição no modelo (menor estatística F parcial) é excluída. As variáveis são excluídas uma a uma do modelo até que todas as variáveis restantes no modelo produzam estatísticas F significantes. Uma vez que uma variável é retirada do modelo, ela nunca mais entra.

Como aponta Miller (2002), *Backward Elimination* geralmente não é factível quando o número de variáveis é maior que o número de observações. Além disso, ela usualmente implica em um maior custo computacional que o *Forward Selection*. Se, por exemplo, tivermos 100 variáveis disponíveis e esperarmos selecionar um subconjunto de menos que 10 delas, o procedimento de *Forward Selection* pode percorrer menos que 10 etapas. Já o *Backward Elimination* deveria realizar muito mais passos para eliminar todas as variáveis irrelevantes uma por uma.

Berk (1978) afirma que ambos os procedimentos de *Forward Selection* e de *Backward Elimination* podem não selecionar o melhor subconjunto de variáveis. O artigo mostra que mesmo quando ambos os procedimentos encontram os mesmos conjuntos de variáveis, não há garantia de que não existam outros subconjuntos de variáveis que fornecem um melhor ajuste.

Outro problema desses procedimentos ocorre quando a variável resposta Y depende da diferença de dois preditores $(X_1 - X_2)$. Em Economia, por exemplo, há muitas situações nas quais alguma taxa de mudança no tempo pode ser um bom preditor para Y. Nesses casos, o *Forward Selection* pode não adicionar $(X_1 - X_2)$ ao modelo, enquanto o *Backward Elimination* pode excluir $(X_1 - X_2)$. Isso porque ambos os procedimentos avaliam se adicionam/excluem as variáveis uma por uma, sem testar se vale a pena utilizar X_1 e X_2 conjuntamente.

2.1.3 Stepwise Selection

O método de *Stepwise Selection* é uma versão modificada do método de *Forward Selection*: a diferença é que as variáveis já incluídas no modelo não necessariamente permanecem nele. Depois que uma variável é adicionada, o método de *Forward-Stepwise* investiga todas as variáveis presentes no modelo e exclui qualquer variável que não produza

mais uma estatística F significante. Assim, o procedimento tem um custo computacional maior e continua com a maioria dos problemas de seleção do *Forward Selection*.

2.1.4 *Forward-Stagewise Regression*

O procedimento de *Forward-Stagewise* inicia-se com um intercepto igual a média amostral da variável dependente e preditores centrados com coeficientes inicialmente iguais a zero. A cada passo, o algoritmo identifica qual variável é mais correlacionada com os resíduos, e então realiza uma regressão linear dos resíduos sobre essa variável escolhida. O coeficiente de inclinação obtido é usado como o coeficiente daquela variável no modelo. Esse procedimento continua até que nenhuma das variáveis tenha correlação com os resíduos. Como pode exigir muitos passos, ele tem um custo computacional grande e pode ser inviável para problemas de alta dimensionalidade.

2.2 *Shrinkage methods*

Os métodos de seleção de subconjuntos produzem um modelo que possivelmente realiza previsões melhores que o modelo completo. Entretanto, tais métodos frequentemente sofrem com grande variabilidade devido ao procedimento discreto (variáveis vão sendo retidas ou excluídas no modelo). Já os métodos de *Shrinkage* não apresentam tanta variabilidade, como apontam Hastie, Tibshirani e Friedman (2001).

2.2.1 Regressão Ridge

A regressão Ridge encolhe o conjunto de coeficientes ao impor uma penalização na soma dos quadrados dos mesmos:

$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 \qquad sujeito\ \ a \qquad \sum_{j=1}^{k} \beta_j^2 \leq t \ , \tag{2.4}$$

onde o parâmetro $t \geq 0$ controla a penalização.

Uma expressão equivalente é dada por

$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right\} , \tag{2.5}$$

onde o parâmetro $\lambda \geq 0$ é uma função do parâmetro t em (2.4) e controla o quão severa é a penalização. Quanto maior for λ, maior será a penalização. Quando $\lambda = 0$, o vetor $\hat{\boldsymbol{\beta}}^{ridge}$ será igual ao vetor de coeficientes obtidos por mínimos quadrados ordinários (MQO).

Como mostram Hastie, Tibshirani e Friedman (2001), utilizando a matriz $\mathbf{X}$ de preditores centrados, tem-se que

$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) + \lambda\, \boldsymbol{\beta}^T \boldsymbol{\beta} \right\}, \tag{2.6}$$

de onde obtém-se a solução que

$$\hat{\boldsymbol{\beta}}^{ridge} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}\,, \tag{2.7}$$

onde $\mathbf{I}$ é a matriz identidade $k \times k$.

A adição de uma constante positiva à diagonal de $\mathbf{X}^T\mathbf{X}$ em (2.7) resolve o problema da multicolineariedade perfeita. Essa foi a principal motivação do uso da regressão ridge quando foi introduzida por Hoerl e Kennard (1970).

Contudo, a regressão Ridge obtém estimativas não-nulas para todos os coeficientes, não sendo, pois, um método de seleção de variáveis.

2.2.2 *Garrote*

Em um contexto de alta dimensionalidade, a escolha das variáveis é importante para a interpretabilidade do modelo. Breiman (1995) propõe o *Garrote*, método que penaliza os coeficientes da regressão de modo que alguns deles são forçados para zero. Os coeficientes são obtidos através da seguinte minimização:

$$\hat{\boldsymbol{\beta}}^{garrote} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} c_j \hat{\beta}_j x_{ji} \right)^2 \qquad sujeito\ a \qquad c_j \geq 0 \quad e \quad \sum_{i=1}^{k} c_j \leq s\,, \tag{2.8}$$

onde $\{\hat{\beta}_j\}$ correspondem às estimativas de MQO.

A desvantagem do *Garrote* é que sua solução depende do sinal e da magnitude das estimativas de MQO. Isso porque, em regressões onde há alta correlação entre os preditores, as estimativas de MQO são prejudicadas e, consequentemente, o garrote também pode ser afetado.

3 Métodos de penalização de norma L_1

Assim como na regressão Ridge (equação (2.4)) ocorre uma penalização da soma dos quadrados dos coeficientes, existem métodos alternativos que impõem penalização na soma dos valores absolutos dos coeficientes. Alguns de tais métodos serão descritos nas seções deste capítulo.

3.1 LASSO

O LASSO (*least absolute shrinkage and selection operator*), proposto por Tibshirani (1996), é outro método de encolhimento do conjunto de coeficientes. Da mesma forma que o *Garrote*, o LASSO tem o objetivo de estimar um modelo que produza previsões com pequena variância e que determine o conjunto de preditores que mais explicam a variável resposta.

Tibshirani (1996) argumenta que as duas técnicas usualmente empregadas para melhorar as estimativas de MQO, *Subset Selection* e regressão Ridge, têm desvantagens. Os procedimentos de *Subset Selection* produzem modelos de fácil interpretação, mas cujo processos de escolha das variáveis apresenta grande variabilidade por serem processos discretos. A regressão Ridge, por sua vez, possui menos variabilidade, encolhe os coeficientes da regressão mas permanece com todos os preditores no modelo.

Como na regressão usual, representada pelas equações (2.1) e (2.2), consideramos que os y_is são condicionalmente independentes dado os x_{ki}s. Assumimos que os x_{ki}s são padronizados tal que $\sum_i x_{ki} = 0$ e $\sum_i x_{ki}^2/n = 1$.

As estimativas LASSO são obtidas através da minimização dos quadrados dos resíduos sujeita uma penalização de norma L_1 dos coeficientes:

$$\hat{\boldsymbol{\beta}}^{LASSO} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 \qquad sujeito\ \ a \qquad \sum_{j=1}^{k} |\beta_j| \leq t \ , \tag{3.1}$$

onde o parâmetro de ajuste $t \geq 0$ controla a penalização. Para todo t, a solução para β_0 é $\hat{\beta}_0 = \overline{y}$. Então é possível assumir, sem perda de generalidade, que $\overline{y} = 0$ e então omitir β_0.

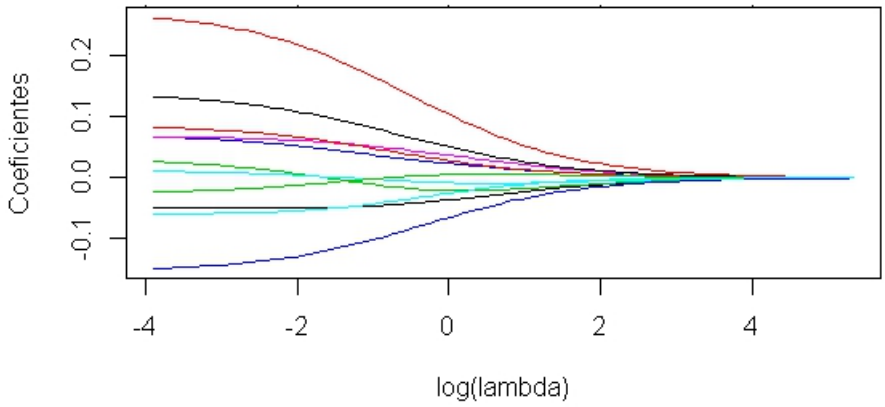
O parâmetro de ajuste $t \geq 0$ em (3.1) controla o quanto de penalização é aplicado ao conjunto de coeficientes. Sendo $\{\hat{\beta}_j^0\}_{1 \leq j \leq k}$ o conjunto de coeficientes de MQO e $t_0 = \sum |\hat{\beta}_j^0|$, os valores $t < t_0$ vão causar um encolhimento dos coeficientes em direção a 0, e alguns coeficientes podem ser exatamente iguais a 0. Já valores $t \geq t_0$ vão resultar em estimativas LASSO iguais às estimativas de MQO.

Usando o Lagrangiano, (3.1) tem expressão equivalente dada por

$$\hat{\boldsymbol{\beta}}^{LASSO} = \underset{\beta_0, \beta_1, \ldots, \beta_k}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right\}, \tag{3.2}$$

onde o parâmetro $\lambda \geq 0$ é uma função do parâmetro t. Quanto maior for λ, maior será a penalização aos coeficientes; quando $\lambda = 0$, as estimativas LASSO serão iguais às estimativas de MQO. A Figura 1 ilustra um exemplo de como λ impacta as estimativas dos coeficientes LASSO.

Figura 1 – Valores dos coeficientes $\hat{\boldsymbol{\beta}}_{LASSO}$ em função de $log(\lambda)$



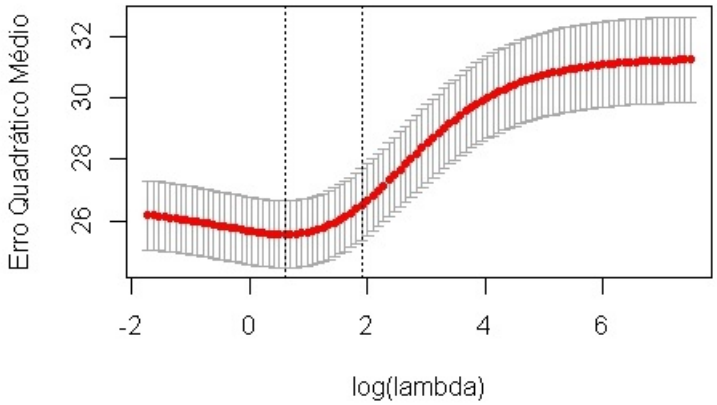
Fonte: Elaborado pelo autor (2014).

O valor do parâmetro λ é escolhido via *K-fold cross-validation*. Nessa técnica, a amostra original é particionada aleatoriamente em K subamostras de tamanhos iguais. Uma das K subamostras é removida e o modelo é estimado com as $K-1$ subamostras restantes. Com o modelo estimado, realizam-se previsões fora da amostra que são comparadas aos dados originais da subamostra anteriormente removida, obtendo-se erros de previsão. O erro quadrático médio de previsão (EQMP) é a medida empregada para avaliar a qualidade do ajuste naquela subamostra. Realiza-se o mesmo procedimento para cada uma

das K subamostras, obtendo-se K resultados de EQMP. A média desses K resultados formam uma medida de EQMP única que mensura a qualidade de ajuste do modelo. O λ ótimo será aquele que minimiza essa medida de erro.

Como retratado na Figura 2, poderia ser utilizado o λ que produz um ajuste com o menor EQMP (correspondente à 1ª linha vertical) ou o maior valor de λ tal que o erro está afastado a no máximo um desvio-padrão do ponto de EQMP mínimo (correspondente à 2ª linha vertical). Neste trabalho, λ será aquele que produz um ajuste com menor EQMP.

Figura 2 – Erro quadrático médio do ajuste em função de $log(\lambda)$



Fonte: Elaborado pelo autor (2014).

Tibshirani (2013) mostra as condições suficientes para a unicidade da solução do LASSO empregando as condições de Karush-Kuhn-Tucker.

O LASSO tem menor variabilidade que os métodos de *Subset Selection*. Além disso, encolhe alguns coeficientes e forçam outros para zero, mantendo as boas características dos procedimentos de *Subset Selection* e da regressão Ridge. Ainda, o LASSO realiza a escolha das variáveis e a estimação dos coeficientes simultaneamente.

Consistência do LASSO na seleção de variáveis

A fim de usar o LASSO como critério de seleção, é importante saber se a solução esparsa obtida pelo método representa bem o modelo verdadeiro. Uma das propriedades desejáveis é que o método seja consistente na seleção das variáveis do modelo; ou seja, que identifique o modelo verdadeiro quando $n \to \infty$.

Para estudar a consistência do LASSO na seleção de modelos, Zhao e Yu (2006) consideram dois problemas: i) se existe uma quantidade determinística de regularização que fornece a consistência na seleção; e ii) se para cada amostra existe uma quantidade correta de regularização que seleciona o modelo verdadeiro. Seus resultados mostram que existe uma condição que denominam por "Condição Irrepresentável" que é quase necessária e suficiente para ambos os tipos de consistência. Seus resultados são para modelos lineares com k fixo e com k crescendo com n.

Seja o modelo de regressão linear

$$Y_n = \mathbf{X_n}\beta^n + \varepsilon_\mathbf{n} \ , \tag{3.3}$$

onde $\varepsilon_\mathbf{n} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ é um vetor de variáveis aleatórias i.i.d. com média 0 e variância σ^2. Y_n é uma variável resposta $n \times 1$ e $\mathbf{X_n} = (X_1^n, \ldots, X_k^n)^T$ é a matriz de preditores $n \times k$, onde $X_i^n = (x_{i,1}, \ldots, x_{i,n})$, para $i = 1, \ldots, k$. β^n é o vetor de coeficientes $k \times 1$. Diferentemente do tradicional cenário em que k é fixo, os dados e os parâmetros do modelo são indexados por n para permitir que variem quando n cresce.

As estimativas LASSO $\beta^n = (\hat{\beta}_1^n, \ldots, \hat{\beta}_k^n)^T$ são definidas por

$$\beta^n(\lambda) = \underset{\beta}{\operatorname{argmin}} ||Y_n - \mathbf{X_n}\beta^n||_2^2 + \lambda||\beta||_1 \ , \tag{3.4}$$

onde $||\cdot||_2^2$ denota a norma L_2 de um vetor, ou seja, a soma dos quadrados dos valores dos componentes do vetor; e $||\cdot||_1$ denota a norma L_1 de um vetor, ou seja, a soma dos valores absolutos dos componentes do vetor.

Diz-se que há consistência na seleção do modelo correto quando

$$P\Big(\{i : \hat{\beta}_i^n \neq 0\} = \{i : \beta_i^n \neq 0\}\Big) \to 1 \quad , \qquad \text{as} \quad n \to \infty. \tag{3.5}$$

Conforme Zhao e Yu (2006), a consistência na seleção do modelo é equivalente à consistência do sinal. As definições a seguir são baseadas no artigo de Zhao e Yu (2006).
Definição 1: Uma estimativa $\hat{\beta}^n$ é igual em sinal a β^n do modelo verdadeiro (que se escreve por $\hat{\beta}^n = {}_s\beta^n$) se e somente se

$$sign(\hat{\beta}^n) = sign(\beta^n) \ , \tag{3.6}$$

onde $sign(\cdot)$ assume valor 1 para valores positivos, -1 para valores negativos e zero para zero.

Definem-se dois tipos de consistências em sinal para o LASSO dependendo de como a quantidade de regularização é determinada.

Definição 2: O LASSO é fortemente consistente em sinal se existe $\lambda_n = f(n)$, isto é, uma função de n e independente de Y_n ou $\mathbf{X_n}$ tal que

$$\lim_{n\to+\infty} P\Big(\hat{\beta}^n(\lambda_n) = {}_s\beta^n\Big) = 1\ . \tag{3.7}$$

Definição 3: O LASSO é geralmente consistente em sinal se

$$\lim_{n\to+\infty} P\Big(\exists\lambda \geq 0\,,\ \hat{\beta}^n(\lambda) = {}_s\beta^n\Big) = 1\ . \tag{3.8}$$

Consistência forte em sinal significa que se pode usar um λ pré-determinado para obter um modelo de seleção consistente. Consistência geral significa que para uma realização aleatória existe uma quantidade correta de regularização que seleciona o verdadeiro modelo. Zhao e Yu (2006) mostram que os dois tipos de consistência são quase equivalentes a uma determinada condição. Primeiramente deve-se definir algumas notações para definir essa condição.

Sem perda de generalidade, assume-se que $\beta^n = (\beta_1^n, \ldots, \beta_q^n, \beta_{q+1}^n, \ldots, \beta_k^n)^T$, onde $\beta_j^n \neq 0$ para $j = 1, \ldots, q$ e $\beta_j^n = 0$ para $j = q+1, \ldots, k$. Seja $\beta_{(1)}^n = (\beta_1^n, \ldots, \beta_1^q)^T$ e $\beta_{(2)}^n = (\beta_{q+1}^n, \ldots, \beta_k^q)^T$. Então escreve-se $\mathbf{X_n}(1)$ e $\mathbf{X_n}(2)$ como sendo as primeiras q e as últimas $k-q$ colunas de $\mathbf{X_n}$, respectivamente, e seja $C^n = \frac{1}{n}\mathbf{X_n}^T\mathbf{X_n}$.

Ao fazer $C_{11}^n = \frac{1}{n}\mathbf{X_n}(1)^T\mathbf{X_n}(1)$, $C_{22}^n = \frac{1}{n}\mathbf{X_n}(2)^T\mathbf{X_n}(2)$, $C_{12}^n = \frac{1}{n}\mathbf{X_n}(1)^T\mathbf{X_n}(2)$ e $C_{21}^n = \frac{1}{n}\mathbf{X_n}(2)^T\mathbf{X_n}(1)$, C^n pode ser expresso em uma matriz de blocos:

$$C^n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix}. \tag{3.9}$$

Assumindo que C_{11}^n é invertível, definem-se as seguintes Condições Irrepresentáveis.

Condição Irrepresentável Forte. Existe um vetor positivo e constante η tal que

$$\left|C_{21}^n (C_{11}^n)^{-1} sign(\beta_{(1)}^n)\right| \leq \mathbf{1} - \eta\ , \tag{3.10}$$

onde $\mathbf{1}$ é um vetor $(k-q)\times 1$.

Condição Irrepresentável Fraca.

$$\left|C_{21}^{n}(C_{11}^{n})^{-1}sign(\beta_{(1)}^{n})\right| < \mathbf{1} \ . \tag{3.11}$$

Zou (2006), de forma semelhante, conclui pela existência de uma condição necessária para a consistência do LASSO na seleção de variáveis.

3.2 adaLASSO

Ao observar que pode haver situações nas quais o LASSO não é consistente na seleção de variáveis, Zou (2006) propõe o LASSO adaptativo (adaLASSO), onde emprega diferentes pesos para diferentes coeficientes:

$$\hat{\boldsymbol{\beta}}^{adaLASSO} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} \omega_j |\beta_j| \right\} \ , \tag{3.12}$$

onde $\omega_j = |\hat{\beta}_j^{ridge}|^{-\tau}$, $\tau > 0$.

Os pesos individuais ω_j servem para ajudar na seleção das variáveis relevantes. Uma variável relevante x_j tende a possuir um coeficiente $\hat{\beta}_j^{ridge}$ grande, o que faz diminuir o peso ω_j atribuído ao coeficiente daquela variável; caso contrário, se a variável x_j for irrelevante, o coeficiente $\hat{\beta}_j^{ridge}$ tende a ser pequeno e implicará em um peso ω_j grande. Desse modo, penalizam-se mais os coeficientes das variáveis que aparentam ser pouco importantes.

Os pesos ω_j também podem ser obtidos a partir de estimativas OLS; entretanto, essa maneira se restringiria ao caso onde $n > k + 1$.

Seguindo sua notação, $A = \{j : \beta_j \neq 0\}$ representa o conjunto verdadeiro de coeficientes não-nulos. Por sua vez, $A_n^* = \{j : \hat{\beta}_j^{*(n)} \neq 0\}$ representa o conjunto de coeficientes não-nulos estimados por (3.12), onde λ_n varia com o tamanho da amostra n. Zou (2006) mostra que, com a utilização de pesos ω_j apropriados, o adaLASSO possui as propriedades *oracle*.

Teorema 1 (Zou (2006)): Suponha que $\lambda_n/\sqrt{n} \to 0$ e que $\lambda_n n^{(\tau-1)/2} \to \infty$. Então, o adaLASSO satisfaz:

1. A consistência na seleção de variáveis: $\lim_{n\to\infty} P(A_n^* = A) = 1$.

2. A normalidade assintótica: $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_A^{*(n)} - \boldsymbol{\beta}_A^*\right) \rightarrow_d N\left(\mathbf{0}, \sigma^2 \times \boldsymbol{C}_{11}^{-1}\right)$.

Dessa forma, além de selecionar corretamente as variáveis relevantes quando o tamanho da amostra aumenta, o adaLASSO possui estimativas dos coeficientes não-nulos que assintoticamente seguem a mesma distribuição que os estimadores de OLS quando este for estimado apenas com as variáveis relevantes.

3.3 WLadaLASSO

Empregando o adaLASSO em séries temporais, cada uma das defasagens das variáveis, inseridas como preditores candidatos, tem seu coeficiente penalizado apenas de acordo com o tamanho de sua estimativa Ridge (ou MQO). É de se indagar se a penalização menor para *lags* menos distantes contribui para a previsão das séries, uma vez que comumente uma informação mais recente é mais importante do que uma antiga.

Park e Sakaori (2012) propõem algumas alternativas de penalização para diferentes *lags*. Semelhantemente a uma delas, o presente trabalho propõe o adaLASSO com *lags* ponderados, denominado aqui por WLadaLASSO (*Weighted Lag adaptive LASSO*):

$$\hat{\boldsymbol{\beta}}^{WLadaLASSO} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} \omega_j |\beta_j| \right\} , \tag{3.13}$$

onde $\omega_j = \left(|\hat{\beta}_j^{ridge}| \; \alpha(1-\alpha)^l \right)^{-\tau}$, $\tau > 0$, $0 < \alpha < 1$, e l representa a ordem de defasagem.

4 Simulação

Este e o próximo capítulo mostram a performance do WLadaLASSO comparado a outros métodos de penalização, através de algumas simulações e de uma aplicação a dados reais.

Todas as implementações foram realizadas no software livre R, onde o processo de estimação das equações (3.2), (3.12) e (3.13) contou com o auxílio da função *glmnet* para otimizar os parâmetros β_j e λ, onde se utilizou a técnica de *10-fold cross-validation*. O parâmetro τ foi fixado em 1.

No WLadaLASSO, para cada α pertencente ao intervalo $[0{,}1; 0{,}2; \ldots; 0{,}9]$, a técnica de *10-fold cross-validation* foi efetuada para encontrar o λ ótimo observando a menor medida de erro. O valor escolhido para α foi aquele que gerou a menor medida de erro dentre os nove procedimentos de *10-fold cross-validation* realizados.

4.1 Estudo 1

Através de simulações de Monte Carlo com 1.000 replicações, foram simuladas 10 séries temporais independentes que seguem um processo AR(1) da forma $x_{i,t} = 0.3x_{i,t-1} + u_{i,t}$, onde $u_{i,t} \sim N(0{,}1)$, $i = 1, \ldots, 10$.

Considerou-se o seguinte processo gerador de dados:

$$\begin{aligned} y_t = \quad & 0.8y_{t-1} + 0.6x_{1,t-1} + 0.3x_{1,t-2} - 0.5x_{2,t-1} - 0.2x_{2,t-2} + 0.4x_{3,t-1} + 0.3x_{3,t-2} + \\ & +0.4x_{4,t-1} - 0.3x_{5,t-1} + 0.2x_{6,t-1} + \varepsilon_t \ , \qquad t = 1, 2, \ldots, T, \end{aligned} \tag{4.1}$$

onde $\varepsilon_t \sim N(0{,}\sigma^2)$.

Os métodos Ridge, LASSO, adaLASSO e WLadaLASSO foram empregados, utilizando-se 10 defasagens de y e 10 defasagens de $x_j, j = 1, \ldots, 10$, totalizando 110 preditores candidatos.

Para comparar as estimativas dos coeficientes, foram utilizadas as medidas Erro Quadrático Médio (EQM) e Viés Absoluto Médio (VAM) dos estimadores através de simulações de

Monte Carlo com 1.000 replicações:

$$EQM = \frac{1}{1000k} \sum_{i=1}^{1000} \sum_{j=1}^{k} \left(\hat{\beta}_j - \beta_j \right)^2 , \tag{4.2}$$

$$VAM = \frac{1}{1000k} \sum_{i=1}^{1000} \sum_{j=1}^{k} \left| \hat{\beta}_j - \beta_j \right| . \tag{4.3}$$

As últimas 10 observações das séries simuladas foram retiradas para avaliar as previsões fora da amostra de um passo à frente . A previsão de y_t foi realizada com base em todo o conjunto de informação existente até a data $t-1$. Foram analisados dois cenários para o processo gerador de dados: com $\sigma = 0{,}1$ e com $\sigma = 1$.

Cenário onde $\sigma = 0{,}1$

A Tabela 1, inspirada em uma tabela do trabalho de Medeiros e Mendes (2012), mostra várias estatísticas acerca da seleção das variáveis do modelo simulado, divididas em painéis, para diferentes tamanhos de amostra. O painel (a) mostra a fração de replicações nas quais o modelo foi corretamente selecionado: todas as variáveis relevantes foram incluídas e todas as variáveis irrelevantes foram excluídas do modelo final; (b) mostra a fração de replicações onde todas as variáveis relevantes foram incluídas; (c) apresenta a fração de regressores relevantes incluídos; (d) apresenta a fração de regressores irrelevantes excluídos; e (e) mostra o número médio de regressores incluídos.

A Tabela 1 mostra que, para grandes amostras, os três métodos identificaram bem o modelo verdadeiro, que possui 10 variáveis relevantes e 100 irrelevantes.

Já para amostras pequenas, a Tabela 1 revela superioridade do WLadaLASSO na expulsão de variáveis irrelevantes do modelo. Para $T = 50$, em média o LASSO e o ada-LASSO incluíram 3,7 e 5,1 covariáveis irrelevantes, respectivamente, enquanto o WLada-LASSO incluiu 0,8.

Os três métodos tiveram performance similar na identificação das variáveis relevantes. Observa-se, na Tabela 2, a fração de replicações nas quais as covariáveis relevantes foram identificadas como relevantes pelo WLadaLASSO. Percebe-se que as covariáveis com coeficientes maiores são mais facilmente detectadas.

Tabela 1 – Estatísticas descritivas da seleção do modelo ($\sigma = 0{,}1$)

T	50	100	200	500	1000	2000
	(a) Fração de variáveis corretamente identificadas					
LASSO	0,909	0,958	0,977	0,992	0,995	1
adaLASSO	0,891	0,94	0,982	0,993	0,997	0,999
WLadaLASSO	0,937	0,956	0,980	0,993	0,997	0,999
	(b) Verdadeiro modelo incluído					
LASSO	0,038	0,341	0,558	0,809	0,920	0,987
adaLASSO	0	0,162	0,393	0,541	0,711	0,888
WLadaLASSO	0,045	0,173	0,416	0,603	0,761	0,927
	(c) Fração de variáveis relevantes incluídas					
LASSO	0,368	0,638	0,821	0,956	0,988	0,999
adaLASSO	0,320	0,606	0,812	0,924	0,964	0,988
WLadaLASSO	0,384	0,550	0,782	0,921	0,967	0,992
	(d) Fração de variáveis irrelevantes excluídas					
LASSO	0,963	0,990	0,992	0,996	0,996	1
adaLASSO	0,949	0,973	0,999	1	1	1
WLadaLASSO	0,992	0,997	1	1	1	1
	(e) Número de variáveis incluídas					
LASSO	7,338	7,359	9,008	9,966	10,330	10,007
adaLASSO	8,335	8,731	8,251	9,264	9,649	9,886
WLadaLASSO	4,615	5,831	7,843	9,209	9,667	9,917

Fonte: Elaborado pelo autor (2014).

Tabela 2 – Fração de replicações onde os coeficientes não-nulos foram identificados como não-nulos pelo WLadaLASSO ($\sigma = 0{,}1$)

	Valor verdadeiro	$T = 50$	$T = 100$	$T = 200$	$T = 500$	$T = 1000$	$T = 2000$
β_{01}	0,8	0,999	1	1	1	1	1
β_{11}	0,6	0,515	0,758	0,955	0,998	1	1
β_{12}	0,3	0,247	0,437	0,813	0,978	1	1
β_{21}	-0,5	0,425	0,675	0,923	0,993	1	1
β_{22}	-0,2	0,160	0,341	0,588	0,822	0,920	0,973
β_{31}	0,4	0,392	0,625	0,859	0,983	1	1
β_{32}	0,3	0,218	0,407	0,762	0,958	0,997	0,998
β_{41}	0,4	0,364	0,541	0,803	0,961	0,998	1
β_{51}	-0,3	0,286	0,430	0,657	0,866	0,957	0,998
β_{61}	0,2	0,235	0,281	0,457	0,647	0,794	0,948

Fonte: Elaborado pelo autor (2014).

A Tabela 3 apresenta as medidas de erro do conjunto dos parâmetros estimados, calculadas por (4.2) e por (4.3). Observa-se que as regressões com penalização de norma L_1 forneceram melhores estimativas que a regressão Ridge. Apesar de elas terem sido semelhantes entre si para amostras maiores, o WLadaLASSO se destacou para a amostra de tamanho $T = 50$.

A Figura 3 da pág. 34 mostra o histograma suavizado (via *Kernel* gaussiano) dos 1000 valores estimados para β_{11}, o coeficiente de $x_{1,t-1}$. Os gráficos mostram que as estimativas WLadaLASSO foram melhores para todos os tamanhos de amostra.

Tabela 3 – Estatísticas descritivas da estimação dos parâmetros ($\sigma = 0{,}1$)

T	50	100	200	500	1000	2000
			Erro Quadrático Médio			
Ridge	0,0159	0,0147	0,0096	0,0089	0,0089	0,0090
LASSO	0,0101	0,0058	0,0033	0,0014	0,0007	0,0003
adaLASSO	0,0142	0,0062	0,0031	0,0013	0,0007	0,0003
WLadaLASSO	0,0092	0,0060	0,0031	0,0013	0,0007	0,0003
			Viés Absoluto Médio			
Ridge	0,0546	0,0527	0,0554	0,0483	0,0461	0,0448
LASSO	0,0291	0,0187	0,0132	0,0085	0,0060	0,0040
adaLASSO	0,0370	0,0198	0,0122	0,0075	0,0052	0,0034
WLadaLASSO	0,0256	0,0188	0,0125	0,0078	0,0055	0,0035

Fonte: Elaborado pelo autor (2014).

A Tabela 4 mostra o Erro Quadrático Médio e o Erro Absoluto Médio da diferença entre as estimativas WLadaLASSO e de MQO quando este for estimado com a utilização das variáveis selecionadas pelo WLadaLASSO. A simulação mostra que as estimativas se aproximam assintoticamente, o que é uma das propriedades *oracle*.

Tabela 4 – Comparação das estimativas WLadaLASSO e MQO ($\sigma = 0{,}1$)

T	50	100	200	500	1000	2000
EQM da diferença	0,002477	0,001724	0,001455	0,000877	0,000544	0,000280
EAM da diferença	0,006719	0,006412	0,007125	0,006072	0,004792	0,003361

Fonte: Elaborado pelo autor (2014).

Os resultados das previsões de um passo à frente estão na Tabela 5. O WLadaLASSO mostrou a melhor performance de previsão, sobretudo para $T = 50$. A Figura 4 da página 35 mostra o histograma suavizado (via *Kernel* gaussiano) dos erros de previsão obtidos.

Tabela 5 – Estatísticas descritivas das previsões ($\sigma = 0{,}1$)

T	50	100	200	500	1000	2000
	Média do Erros Quadráticos Médios					
Ridge	6,9089	6,0815	1,0440	0,2958	0,1419	0,0646
LASSO	2,7802	1,2791	0,5724	0,2147	0,1179	0,0526
adaLASSO	3,8760	1,3539	0,5807	0,2171	0,1169	0,0532
WLadaLASSO	2,5803	1,2875	0,5644	0,2077	0,1114	0,0500
	Mediana do Erros Quadráticos Médios					
Ridge	5,2577	4,3612	0,4520	0,1103	0,0586	0,0358
LASSO	1,7406	0,5578	0,2081	0,0847	0,0489	0,0257
adaLASSO	2,5941	0,6365	0,2173	0,0878	0,0496	0,0260
WLadaLASSO	1,4721	0,5746	0,2162	0,0838	0,0488	0,0243
	Média dos Erros Absolutos Médios					
Ridge	2,1084	1,9676	0,7185	0,3680	0,2620	0,1876
LASSO	1,2520	0,7576	0,4897	0,3083	0,2311	0,1619
adaLASSO	1,5171	0,8054	0,4950	0,3109	0,2309	0,1623
WLadaLASSO	1,1702	0,7723	0,4856	0,3032	0,2258	0,1578
	Mediana dos Erros Absolutos Médios					
Ridge	1,9643	1,7969	0,5668	0,2823	0,2013	0,1583
LASSO	1,1024	0,6287	0,3754	0,2402	0,1812	0,1323
adaLASSO	1,3362	0,6602	0,3824	0,2420	0,1820	0,1304
WLadaLASSO	1,0106	0,6242	0,3796	0,2364	0,1804	0,1272

Fonte: Elaborado pelo autor (2014).

Cenário onde $\sigma = 1$

O cenário com $\sigma = 1$ apresenta, na Tabela 6, que o LASSO mostrou muito maior dificuldade na expulsão das variáveis irrelevantes mesmo no caso de amostras grandes. Os demais resultados desse cenário foram bastante similares ao caso onde $\sigma = 0{,}1$.

Tabela 6 – Estatísticas descritivas da seleção do modelo ($\sigma = 1$)

T	50	100	200	500	1000	2000
	(a) Fração de variáveis corretamente identificadas					
LASSO	0,900	0,904	0,897	0,881	0,859	0,845
adaLASSO	0,886	0,911	0,929	0,935	0,936	0,977
WLadaLASSO	0,934	0,952	0,956	0,967	0,974	0,992
	(b) Verdadeiro modelo incluído					
LASSO	0,005	0,365	0,658	0,906	0,976	0,998
adaLASSO	0	0,154	0,564	0,793	0,928	0,986
WLadaLASSO	0,024	0,209	0,564	0,808	0,938	0,990
	(c) Fração de variáveis relevantes incluídas					
LASSO	0,335	0,660	0,854	0,975	0,997	1
adaLASSO	0,306	0,621	0,849	0,962	0,990	0,998
WLadaLASSO	0,382	0,591	0,822	0,956	0,990	0,999
	(d) Fração de variáveis irrelevantes excluídas					
LASSO	0,956	0,928	0,901	0,872	0,845	0,829
adaLASSO	0,944	0,940	0,937	0,932	0,931	0,975
WLadaLASSO	0,989	0,988	0,969	0,968	0,972	0,991
	(e) Número de variáveis incluídas					
LASSO	7,728	13,758	18,403	22,528	25,428	27,054
adaLASSO	8,678	12,200	14,810	16,385	16,797	12,502
WLadaLASSO	4,959	7,090	11,310	12,801	12,722	10,855

Fonte: Elaborado pelo autor (2014).

Tabela 7 – Fração de replicações onde os coeficientes não-nulos foram identificados como não-nulos pelo WLadaLASSO ($\sigma = 1$)

	Valor verdadeiro	$T = 50$	$T = 100$	$T = 200$	$T = 500$	$T = 1000$	$T = 2000$
β_{01}	0,8	0,998	1	1	1	1	1
β_{11}	0,6	0,520	0,774	0,961	0,999	1	1
β_{12}	0,3	0,249	0,485	0,837	0,984	1	1
β_{21}	-0,5	0,438	0,706	0,926	0,993	1	1
β_{22}	-0,2	0,149	0,406	0,661	0,897	0,978	0,996
β_{31}	0,4	0,385	0,664	0,871	0,987	1	1
β_{32}	0,3	0,197	0,475	0,786	0,969	0,999	1
β_{41}	0,4	0,373	0,574	0,833	0,975	0,999	1
β_{51}	-0,3	0,278	0,486	0,737	0,928	0,986	1
β_{61}	0,2	0,235	0,343	0,609	0,83	0,943	0,991

Fonte: Elaborado pelo autor (2014).

Tabela 8 – Estatísticas descritivas da estimação dos parâmetros ($\sigma = 1$)

T	50	100	200	500	1000	2000
			Erro Quadrático Médio			
Ridge	0,0159	0,0147	0,0095	0,0070	0,0062	0,0058
LASSO	0,0109	0,0057	0,0027	0,0008	0,0002	0,0001
adaLASSO	0,0156	0,0062	0,0026	0,0007	0,0002	$\simeq 0$
WLadaLASSO	0,0095	0,0056	0,0025	0,0007	0,0002	$\simeq 0$
			Viés Absoluto Médio			
Ridge	0,0556	0,0535	0,0604	0,0466	0,0403	0,0367
LASSO	0,0315	0,0205	0,0130	0,0066	0,0038	0,0022
adaLASSO	0,0397	0,0217	0,0120	0,0057	0,0030	0,0013
WLadaLASSO	0,0267	0,0182	0,0112	0,0052	0,0026	0,0012

Fonte: Elaborado pelo autor (2014).

Tabela 9 – Comparação das estimativas WLadaLASSO e MQO ($\sigma = 1$)

T	50	100	200	500	1000	2000
EQM da diferença	0,00255	0,00161	0,00106	0,00042	0,00014	0,00003
EAM da diferença	0,00709	0,00648	0,00566	0,00327	0,00165	0,00069

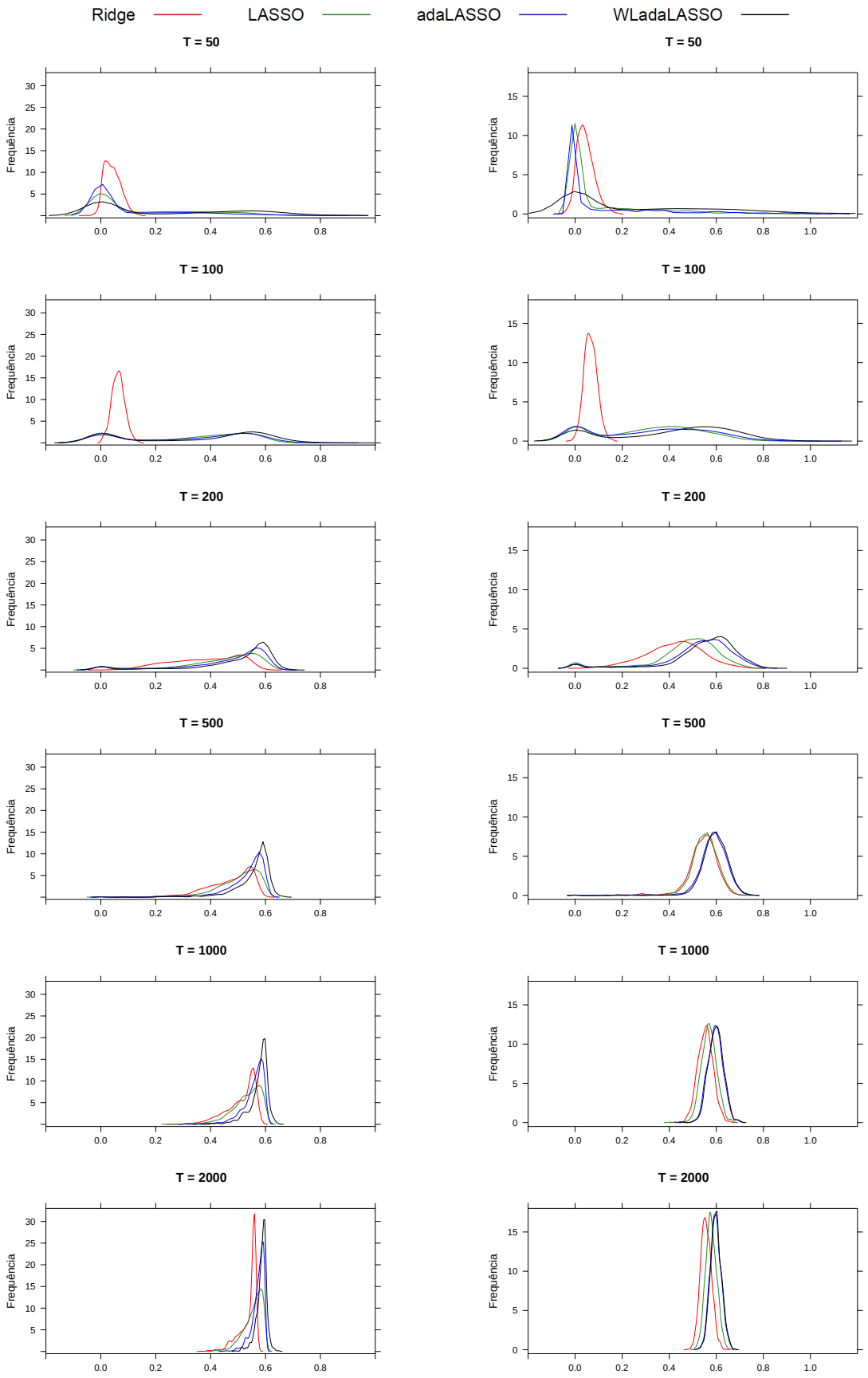
Fonte: Elaborado pelo autor (2014).

Tabela 10 – Estatísticas descritivas das previsões ($\sigma = 1$)

T	50	100	200	500	1000	2000
	Média do Erros Quadráticos Médios					
Ridge	7,4249	6,5281	1,3333	0,5267	0,3889	0,3433
LASSO	3,2239	1,4365	0,7267	0,363	0,2867	0,2569
adaLASSO	4,5595	1,5438	0,7277	0,3601	0,2832	0,2550
WLadaLASSO	2,8517	1,4106	0,7110	0,355	0,2815	0,2550
	Mediana do Erros Quadráticos Médios					
Ridge	5,5856	4,5400	0,8110	0,4082	0,3422	0,3038
LASSO	2,1879	0,7286	0,4060	0,2803	0,2566	0,2367
adaLASSO	3,2509	0,8761	0,3839	0,2724	0,2527	0,2363
WLadaLASSO	1,7422	0,7212	0,3703	0,2671	0,2504	0,2349
	Média dos Erros Absolutos Médios					
Ridge	2,1884	2,0310	0,8656	0,5639	0,4958	0,4688
LASSO	1,3826	0,8702	0,6164	0,4643	0,4251	0,4039
adaLASSO	1,6603	0,9119	0,6151	0,4622	0,4226	0,4025
WLadaLASSO	1,2668	0,8532	0,6045	0,4590	0,4215	0,4025
	Mediana dos Erros Absolutos Médios					
Ridge	2,0198	1,8460	0,7500	0,5275	0,4771	0,4533
LASSO	1,2320	0,7038	0,5131	0,4349	0,4122	0,3961
adaLASSO	1,5146	0,7655	0,5089	0,4284	0,4099	0,3948
WLadaLASSO	1,0922	0,6898	0,4984	0,4271	0,4081	0,3951

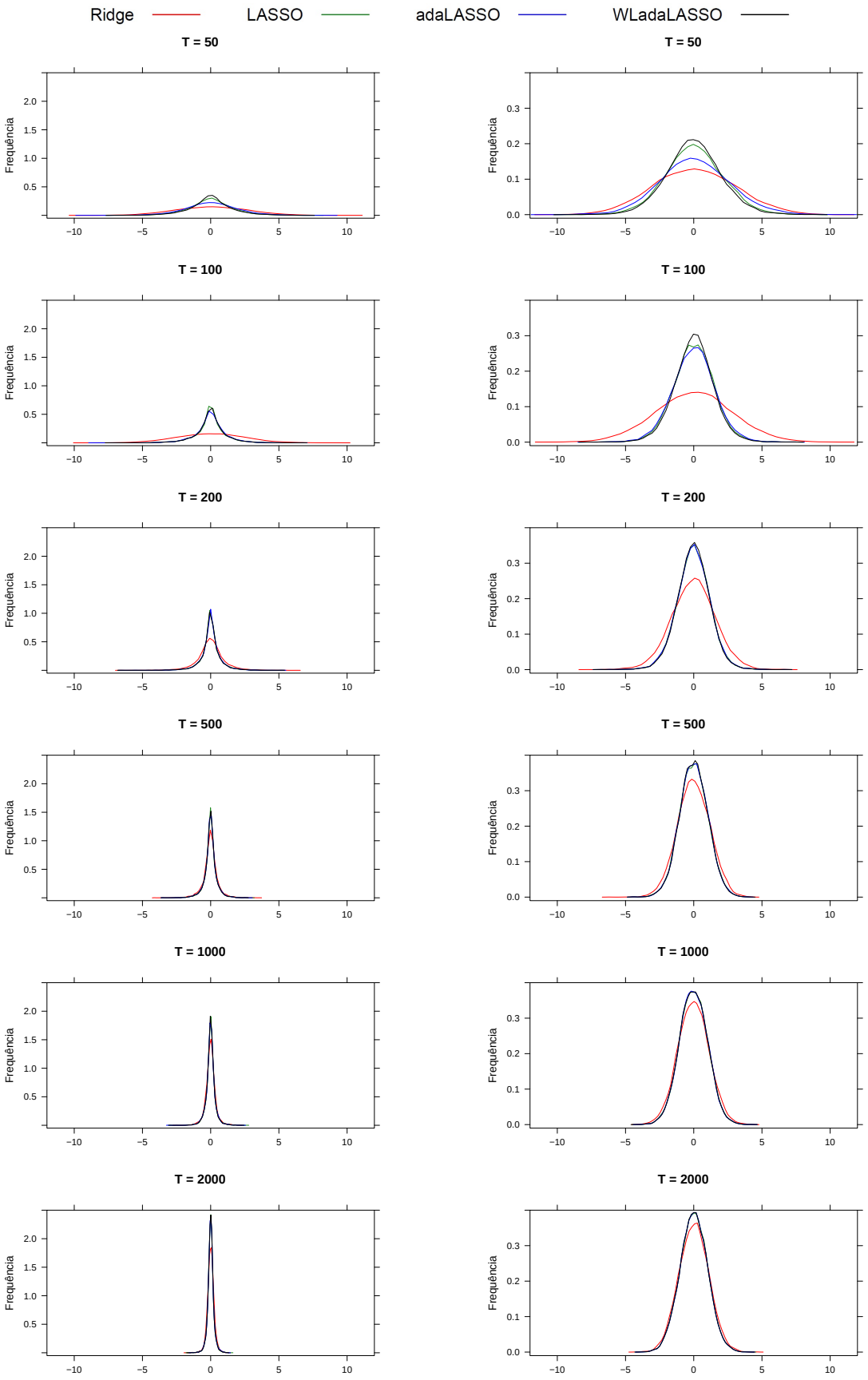
Fonte: Elaborado pelo autor (2014).

Figura 3 – Distribuição de $\hat{\beta}_{11}$ (estimador para $\beta_{11} = 0{,}6$). Modelo com $\sigma = 0{,}1$ (à esquerda) e com $\sigma = 1$ (à direita)



Fonte: Elaborado pelo autor (2014).

Figura 4 – Distribuição dos erros de previsão. Modelo com $\sigma = 0{,}1$ (à esquerda) e com $\sigma = 1$ (à direita)



Fonte: Elaborado pelo autor (2014).

4.2 Estudo 2

Novamente, via simulações de Monte Carlo com 1.000 replicações, foram simuladas 10 séries temporais independentes que seguem um processo AR(1) da forma $x_{i,t} = \phi_x x_{i,t-1} + u_{i,t}$, onde $u_{i,t} \sim N(0,1)$, $i = 1, \ldots, 10$.

A fim de comparar as previsões dos métodos para diferentes modelos, consideraram-se os seguintes processos geradores de dados:

$$w_t = \quad 0{,}3w_{t-1} - 0{,}2x_{1,t-1} + 0{,}1x_{2,t-1} + 0{,}1x_{3,t-1} + \varepsilon_t \ , \quad \text{onde } \varepsilon_t \sim N(0,\sigma^2) \ , \tag{4.4}$$

$$y_t = \quad 0{,}6y_{t-1} + \sum_{k=1}^{5} 0{,}2x_{k,t-1} + \sum_{k=6}^{10}(-0{,}2)x_{k,t-1} + \varepsilon_t \ , \quad \text{onde } \varepsilon_t \sim N(0,\sigma^2) \tag{4.5}$$

e

$$\begin{aligned} z_t = \quad & 0{,}8z_{t-1} + 0{,}6x_{1,t-1} + 0{,}3x_{1,t-2} - 0{,}5x_{2,t-1} - 0{,}2x_{2,t-2} + 0{,}4x_{3,t-1} + 0{,}3x_{3,t-2} + \\ & + 0{,}4x_{4,t-1} - 0{,}3x_{5,t-1} + 0{,}2x_{6,t-1} + \varepsilon_t \ , \quad \text{onde } \varepsilon_t \sim N(0,\sigma^2) \ . \end{aligned} \tag{4.6}$$

Assim, testaram-se os casos de: poucas variáveis relevantes, onde apenas a primeira defasagem era relevante (w_t); todas as variáveis relevantes, onde apenas a primeira defasagem era relevante (y_t); e algumas variáveis relevantes, onde ordens de defasagens diferentes eram relevantes (z_t).

Os métodos Ridge, LASSO, adaLASSO e WLadaLASSO foram utilizados, empregando-se 10 defasagens da variável dependente e 10 defasagens de $x_j, j = 1, \ldots, 10$, totalizando 110 preditores candidatos.

Cada um dos modelos foi gerado sob dois cenários: com $\phi_x = 0{,}1$ e $\phi_x = 0{,}5$. Isso para observar se a dependência entre os preditores candidatos tem influência na performance preditiva dos métodos. Todos os cenários foram gerados utilizando $\sigma = 0{,}1$.

Analisando as Tabelas 11 e 12, novamente observa-se que o método Ridge teve a pior performance de previsão. Os demais métodos realizaram previsões com performance semelhante para $T = 200$ e $T = 1000$, mas novamente o WLadaLASSO se destacou para a amostra menor.

Tabela 11 – Média dos Erros Quadráticos Médios de Previsão

Modelo	Dependência	Método	$T = 50$	$T = 200$	$T = 1000$
w	$\phi_x = 0{,}1$	Ridge	0,0713	0,0203	0,0115
		LASSO	0,0269	0,0111	0,0101
		adaLASSO	0,0279	0,0113	0,0100
		WLadaLASSO	0,0177	0,0111	0,0099
	$\phi_x = 0{,}5$	Ridge	0,1019	0,0210	0,0116
		LASSO	0,0384	0,0114	0,0102
		adaLASSO	0,0371	0,0116	0,0101
		WLadaLASSO	0,0258	0,0115	0,0101
y	$\phi_x = 0{,}1$	Ridge	0,5850	0,0548	0,0148
		LASSO	0,4203	0,0227	0,0108
		adaLASSO	0,4175	0,0224	0,0106
		WLadaLASSO	0,1896	0,0223	0,0106
	$\phi_x = 0{,}5$	Ridge	1,1755	0,1222	0,0219
		LASSO	0,6422	0,0583	0,0157
		adaLASSO	0,6681	0,0603	0,0156
		WLadaLASSO	0,4091	0,0602	0,0156
z	$\phi_x = 0{,}1$	Ridge	4,6399	0,6472	0,0852
		LASSO	1,8904	0,3615	0,0675
		adaLASSO	2,2462	0,3665	0,0678
		WLadaLASSO	1,6370	0,3553	0,0643
	$\phi_x = 0{,}5$	Ridge	12,3918	1,8482	0,2417
		LASSO	5,3397	1,0195	0,2088
		adaLASSO	9,9637	1,0575	0,2081
		WLadaLASSO	5,6459	1,0160	0,1989

Fonte: Elaborado pelo autor (2014).

Tabela 12 – Mediana dos Erros Quadráticos Médios de Previsão

Modelo	Dependência	Método	$T = 50$	$T = 200$	$T = 1000$
w	$\phi_x = 0{,}1$	Ridge	0,0654	0,0185	0,0106
		LASSO	0,0224	0,0105	0,0094
		adaLASSO	0,0229	0,0103	0,0093
		WLadaLASSO	0,0149	0,0103	0,0093
	$\phi_x = 0{,}5$	Ridge	0,0882	0,0190	0,0108
		LASSO	0,0291	0,0106	0,0096
		adaLASSO	0,0292	0,0108	0,0096
		WLadaLASSO	0,0182	0,0105	0,0095
y	$\phi_x = 0{,}1$	Ridge	0,4939	0,0396	0,0134
		LASSO	0,3502	0,0153	0,0098
		adaLASSO	0,3511	0,0143	0,0096
		WLadaLASSO	0,1376	0,0139	0,0096
	$\phi_x = 0{,}5$	Ridge	0,9278	0,0680	0,0170
		LASSO	0,5024	0,0236	0,0116
		adaLASSO	0,5300	0,0234	0,0116
		WLadaLASSO	0,2961	0,0231	0,0116
z	$\phi_x = 0{,}1$	Ridge	3,5591	0,3136	0,0441
		LASSO	1,2845	0,1486	0,0323
		adaLASSO	1,5492	0,1486	0,0320
		WLadaLASSO	0,9717	0,1452	0,0309
	$\phi_x = 0{,}5$	Ridge	8,0825	0,8174	0,0914
		LASSO	2,7800	0,3707	0,0826
		adaLASSO	6,4376	0,3742	0,0793
		WLadaLASSO	2,6000	0,3507	0,0757

Fonte: Elaborado pelo autor (2014).

5 Aplicação

As previsões do crescimento econômico de um país são de grande importância para auxiliar os agentes na tomada de decisões. Este trabalho busca realizar previsões do PIB brasileiro com base no seu próprio passado e no de outras variáveis macroeconômicas. O primeiro desafio a enfrentar é que o PIB é divulgado trimestralmente, enquanto a maioria das outras variáveis macroeconômicas são divulgadas mensalmente. O segundo desafio é como obter boa performance de previsão do PIB utilizando muitas covariáveis e poucas observações disponíveis.

Para trabalhar com dados de diferentes frequências, pode-se utilizar a regressão UMIDAS (*Unrestricted Mixed Data Sampling*), que estima um parâmetro para cada defasagem de cada covariável. Essa metodologia foi aplicada em Foroni, Marcellino e Schumacher (2013) mas só denominada como UMIDAS em Koenig, Dolmas e Piger (2003). Uma aplicação de previsão do PIB brasileiro é realizada em Ziegelmann e Zuanazzi (no prelo 2014).

Seja $x_{j,t}^{(m)}$ a j-ésima variável de alta frequência, y_t a variável de baixa frequência e m o número de vezes em que $x_{j,t}^{(m)}$ ocorre para cada valor de y_t em cada período t. Seja L o operador defasagem que atua da forma $L^{\frac{(p-1)}{m}} x_t^{(m)} = x_{t-\frac{(p-1)}{m}}^{(m)}$. A regressão UMIDAS estima a seguinte equação:

$$y_t = \beta_0 + \sum_{d=1}^{D} \phi_d y_{t-d} + \sum_{j=1}^{k} \sum_{p=p_0}^{P} \beta_{j,p} L^{\frac{(p-1)}{m}} x_{j,t}^{(m)} + \varepsilon_t \ , \tag{5.1}$$

onde $p_0 \geq 2$, D é a ordem de defasagem da variável dependente e P é a ordem de defasagem das variáveis independentes.

Quando a variável resposta tiver frequência trimestral e as variáveis explicativas possuírem frequência mensal ($m = 3$) e deseja-se obter previsões fora da amostra, utiliza-se $p_0 = 2$ para prever 1/3 passo à frente, $p_0 = 3$ para prever 2/3 passo à frente e $p_0 = 4$ para prever 1 passo à frente.

Assim, a equação (5.1) torna-se útil para realizar previsões do PIB trimestral brasileiro tendo covariáveis mensais. O maior problema da abordagem UMIDAS é que o número

de parâmetros pode ser muito grande comparativamente ao tamanho da amostra, o que motiva o uso de métodos que penalizam os coeficientes. Assim, o presente trabalho realiza uma aplicação das penalizações Ridge, LASSO, adaLASSO e WLadaLASSO aos coeficientes da equação (5.1), no intuito de reduzir a dimensionalidade do espaço paramétrico e de obter melhores previsões.

As variáveis macroeconômicas escolhidas foram aquelas consideradas candidatas potenciais para prever o PIB brasileiro, que têm frequência mensal e que cujas observações coletadas tiveram início antes da primeira observação do PIB brasileiro. Chauvet e Silva (2004) apontam várias razões econômicas pelas quais algumas variáveis antecipam ciclos econômicos brasileiros, justificando o uso de informações do passado para prever o comportamento da economia.

Utilizou-se a série da variação dessazonalizada do PIB do 2º trimestre de 1996 ao 4º trimestre de 2012. Ao todo, foram empregados outros 16 regressores mensais (séries correspondentes ao mesmo período): variação do índice Bovespa, variação do índice Dow Jones, variação da taxa de câmbio em reais perante o dólar (PTAX), variação do rendimento dos Certificados de Depósito Interbancário (CDI), variação dessazonalizada da produção industrial (geral e separada em extrativa e de transformação), variação em dólar do preço do barril de petróleo, variação dessazonalizada em dólar do valor total das exportações e os retornos das ações CMIG4, PETR4, BBDC4, BBAS3, AMBV4, USIM5 e VALE5.

Trabalhou-se com a taxa de variação de todas as séries, de modo que todas elas são estacionárias de acordo com os testes de raiz unitária ADF (*Augmented Dickey–Fuller*) e PP (*Phillips–Perron*). Todas as séries foram normalizadas para possuírem média zero e desvio padrão unitário. Assim, os modelos foram implementados sem intercepto.

As últimas 24 observações trimestrais foram retiradas para avaliar as previsões fora da amostra. A previsão de y_t foi realizada com base em todo o conjunto de informação existente até a data $t-1$ (previsão 1 passo à frente), data $t-2/3$ (previsão 2/3 passo à frente) e data $t-1/3$ (previsão 1/3 passo à frente). O número de defasagens escolhido foi o que resultou na menor medida de erro no procedimento de *10-fold cross-validation*.

As regressões Ridge, LASSO, adaLASSO e WLadaLASSO foram implementadas empregando a matriz de preditores da equação (5.1), incorporando a ideia da regressão UMIDAS. As previsões obtidas por tais métodos foram comparadas com as previsões do benchmark modelo auto-regressivo de média móvel (ARMA), cujas ordens de defasagem foram escolhidas através do AIC (*Akaike Information Criterion*).

A Tabela 13 mostra superioridade do WLadaLASSO na performance preditiva, essencialmente nas previsões de 1/3 passo à frente. Os testes de Diebold e Mariano (2002) para a comparação das previsões apontaram que a diferença entre as previsões do ARMA e do WLadaLASSO, de 2/3 passo à frente e de 1/3 passo à frente, foi quase estatisticamente significante a 10%.

Tabela 13 – Estatísticas descritivas das previsões do PIB brasileiro

	1 passo à frente	2/3 passo à frente	1/3 passo à frente
	Erro Quadrático Médio de Previsão		
ARMA	1,131213	1,131213	1,131213
Ridge	0,847527	0,812628	0,788010
LASSO	0,899071	0,927329	0,653488
adaLASSO	0,841299	0,580865	0,47756
WLadaLASSO	0,818903	0,487315	0,38988
	Erro Absoluto Médio de Previsão		
ARMA	0,718676	0,718676	0,718676
Ridge	0,678062	0,649818	0,631775
LASSO	0,748338	0,671024	0,568424
adaLASSO	0,742196	0,561927	0,479257
WLadaLASSO	0,650831	0,485628	0,444672

Fonte: Elaborado pelo autor (2014).

Portanto, a abordagem WLadaLASSO, ao identificar quais variáveis possuem maior poder preditivo e ao reduzir o problema da alta dimensionalidade do espaço paramétrico, conseguiu obter previsões melhores para o PIB brasileiro. A Tabela 14 mostra as covariáveis selecionadas (e os respectivos *lags*) pelo WLadaLASSO para prever o PIB bra-

sileiro do 4º trimestre de 2012, utilizando o conjunto de informação até o 3º trimestre de 2012.

Tabela 14 – Covariáveis selecionadas pelo WLadaLASSO para prever o PIB brasileiro do 4º trimestre de 2012, utilizando o conjunto de informação até o 3º trimestre de 2012

Covariável	*Lags* selecionados
Retorno da ação BBAS3	2
Retorno da ação BBDC4	1
Retorno da ação USIM5	1,2
Retorno da ação VALE5	2
Variação do índice Dow Jones	4
Variação dessazonalizada da indústria de transformação	1,2
Variação do rendimento do CDI mensal	1
Variação dessazonalizada das exportações	3
Variação em dólar do preço do barril de petróleo	2

Fonte: Elaborado pelo autor (2014).

As covariáveis selecionadas pelo WLadaLASSO podem ser consideradas como indicadores antecedentes do PIB, uma vez que a mudança das mesmas hoje pode indicar o comportamento da economia no curto prazo, o que contribui na previsão das próximas observações do PIB.

a **Retornos das ações e índice Dow Jones:** O aumento (queda) do preço das ações geralmente começa a subir (cair) antes da economia como um todo, uma vez que reflete a expectativa dos investidores.

b **Variação dessazonalizada da indústria de transformação e variação dessazonalizada das exportações:** A variação da produção industrial e das exportações podem sinalizar o futuro crescimento ou queda da economia, visto que ambos são justamente componentes do PIB.

c **Variação do rendimento do CDI mensal:** A variação do rendimento do CDI é um dos principais indicadores de taxas de juros utilizados pelo mercado no Brasil.

Assim, é um indicador da condução da política monetária, que interfere no volume de investimento.

d **Variação em dólar do preço do barril de petróleo:** A alta do preço do petróleo pode apontar um aumento do valor das importações no curto prazo, além de contribuir para o aumento da inflação doméstica.

6 Conclusão

Este trabalho tem o objetivo de investigar como a penalização do conjunto de coeficientes pode contribuir para a performance de previsão de séries temporais. Os métodos que penalizam os coeficientes são de extrema importância na redução da dimensionalidade de problemas macroeconômicos, cujas séries temporais são muitas e possuem poucas observações.

São aplicados os métodos Ridge, LASSO e adaLASSO, que surgiram no contexto de regressão linear mas que estão cada vez mais presentes em estudos de séries temporais. Observando que uma informação mais recente tende a contribuir mais na previsão das séries, este trabalho utiliza o WLadaLASSO, que penaliza mais os *lags* mais distantes e que é inspirado no trabalho de Park e Sakaori (2012).

Os estudos de simulação apontam que, essencialmente no caso de amostras pequenas, o WLadaLASSO supera outros métodos de penalização na seleção das covariáveis, na estimação dos parâmetros e na previsão fora da amostra.

Além disso, a aplicação a dados macroeconômicos mostra que o WLadaLASSO novamente apresenta melhor performance de previsão para o PIB brasileiro, ao selecionar as variáveis que podem representar indicadores da economia no curto prazo.

Estudos futuros serão necessários para investigar as propriedades teóricas dessa abordagem, assim como identificar a quais situações o seu uso é mais apropriado.

Referências

BERK, K. N. Comparing subset regression procedures. **Technometrics**, Taylor & Francis Group, Alexandria, v. 20, n. 1, p. 1–6, 1978.

BREIMAN, L. Better subset selection using the non-negative garrote. **Technometrics**, Taylor & Francis, Alexandria, v. 37, n. 4, p. 738–754, 1995.

BREIMAN, L. Bias-variance, regularization, instability and stabilization. **NATO ASI series**. Series F: computer and system sciences, Plenum, [S.l.], p. 27–56, 1998.

CHAUVET, M.; SILVA, J. A. B. d. **Indicadores antecedentes de recessões brasileiras**. XXVI Encontro Brasileiro de Econometria. João Pessoa, v. 10, p. 11–12, 2004.

CLARK, T. E.; MCCRACKEN, M. W. Forecasting with small macroeconomic vars in the presence of instabilities. **Frontiers of Economics and Globalization**, Emerald Group Publishing Limited, Bingley, v. 3, p. 93–147, 2008.

DIEBOLD, F. X.; MARIANO, R. S. Comparing predictive accuracy. **Journal of Business & economic statistics**, [S.l.], v. 20, n. 1, 2002.

FAN, J.; GUO, S.; HAO, N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. **Journal of the Royal Statistical Society**: Series B (Statistical Methodology), Wiley Online Library, Londres, v. 74, n. 1, p. 37–65, 2012.

FAN, J.; LI, R. Variable selection via nonconcave penalized likelihood and its oracle properties. **Journal of the American Statistical Association**, Taylor & Francis, Alexandria, v. 96, n. 456, p. 1348–1360, 2001.

FAN, J.; LV, J. Sure independence screening for ultrahigh dimensional feature space. **Journal of the Royal Statistical Society**: Series B (Statistical Methodology), Wiley Online Library, Londres, v. 70, n. 5, p. 849–911, 2008.

FAN, J.; LV, J. A selective overview of variable selection in high dimensional feature space. **Statistica Sinica**, NIH Public Access, Taipei, v. 20, n. 1, p. 101, 2010.

FORONI, C.; MARCELLINO, M.; SCHUMACHER, C. Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials. **Journal of the Royal Statistical Society**: Series A (Statistics in Society), Wiley Online Library, Londres, 2013.

FURNIVAL, G. M.; WILSON, R. W. Regressions by leaps and bounds. **Technometrics**, Taylor & Francis, Alexandria, v. 16, n. 4, p. 499–511, 1974.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. J. H. The elements of statistical learning. Nova Iorque: Springer New York, 2001.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**, Taylor & Francis Group, Alexandria, v. 12, n. 1, p. 55–67, 1970.

HUA, Y. Macroeconomic forecasting using large vector auto regressive model - master thesis. **Centre for Applied Statistics and Economics** - Humboldt-Universität zu Berlin, Berlim, 2011.

KOENIG, E. F.; DOLMAS, S.; PIGER, J. The use and abuse of real-time data in economic forecasting. **Review of Economics and Statistics**, MIT Press, Cambridge, v. 85, n. 3, p. 618–628, 2003.

LI, J. Monetary policy analysis based on lasso-assisted vector autoregression (lavar). Available at SSRN 2017877, 2012. Disponível em: ⟨http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2017877⟩. Acesso em: 25 dez. 2013.

LINDSEY, C.; SHEATHER, S. Variable selection in linear regression. **Stata Journal**, College Station, v. 10, n. 4, p. 650, 2010.

MEDEIROS, M. C.; MENDES, E. F. Estimating High-Dimensional Time Series Models. [S.l.], 2012.

MILLER, A. **Subset selection in regression**. CRC Press, Boca Raton, 2002.

MOL, C. D.; GIANNONE, D.; REICHLIN, L. Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? **Journal of Econometrics**, Elsevier, Amsterdam, v. 146, n. 2, p. 318–328, 2008.

NG, S.; MOENCH, E. A factor analysis of housing market dynamics in the us and the regions. manuscript, **Columbia University**, Nova Iorque, 2009.

PARK, H.; SAKAORI, F. Lag weighted lasso for time series model. **Computational Statistics**, Springer, Berlim, p. 1–12, 2012.

SIMS, C. A. Macroeconomics and reality. **Econometrica**: Journal of the Econometric Society, Hoboken, p. 1–48, 1980. Disponível em: ⟨http://ideas.repec.org/a/ecm/emetrp/v48y1980i1p1-48.html⟩. Acesso em: 11 abr. 2013.

SONG, S.; BICKEL, P. J. Large vector auto regressions. arXiv preprint arXiv:1106.3915, 2011. Disponível em: ⟨http://arxiv.org/abs/1106.3915⟩. Acesso em: 18 jul. 2013.

STOCK, J. H.; WATSON, M. W. Vector autoregressions. **The Journal of Economic Perspectives**, Nashville, v. 15, n. 4, p. 101–115, 2001.

STOCK, J. H.; WATSON, M. W. Forecasting with many predictors. **Handbook of economic forecasting**, Elsevier, Amsterdam, v. 1, p. 515–554, 2006.

STOCK, J. H.; WATSON, M. W. The evolution of national and regional factors in us housing construction. **Princeton University**, Princeton, 2008.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society**. Series B (Methodological), Londres, p. 267–288, 1996.

TIBSHIRANI, R. J. The lasso problem and uniqueness. **Electronic Journal of Statistics**, Institute of Mathematical Statistics, v. 7, p. 1456–1490, 2013. Disponível em: ⟨http://arxiv.org/abs/1206.0313⟩. Acesso em: 17 jan. 2014.

ZHAO, P.; YU, B. On model selection consistency of lasso. **The Journal of Machine Learning Research**, Cambridge, v. 7, p. 2541–2563, 2006.

ZIEGELMANN, F. A.; ZUANAZZI, P. T. Previsões para o crescimento do pib trimestral brasileiro com séries financeiras e econômicas mensais: uma aplicação de MIDAS. **Economia Aplicada**, Ribeirão Preto, no prelo 2014.

ZOU, H. The adaptive lasso and its oracle properties. **Journal of the American statistical association**, Taylor & Francis, Alexandria, v. 101, n. 476, p. 1418–1429, 2006.