Universidade de Brasília
Departamento de Estatística

Modelos de regressão para previsão de curto prazo da inflação brasileira utilizando dados de alta dimensão

Pedro Henrique Lima de Menezes

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Pedro Henrique Lima de Menezes

Modelos de regressão para previsão de curto prazo da inflação brasileira utilizando dados de alta dimensão

Orientador(a): George Freitas von Borries

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Àqueles que, sem medir esforços, participaram e contribuíram para a minha trajetória até aqui, dedico este trabalho

Agradecimentos

A Deus, sempre, início e fim de todas coisas,

À minha família e à minha namorada, pelas valiosas lições, pelo apoio incondicional e pela motivação nos momentos mais difíceis,

Ao meu orientador, Prof. Dr. George von Borries, pela dedicação, prestatividade e paciência durante os cursos ministrados e na minha orientação,

Aos professores dos departamentos de Estatística e Matemática pelos ensinos que contribuíram para a minha formação,

Aos demais funcionários da UnB, pelos serviços prestados,

E aos meus amigos e colegas, por todos os bons momentos e conhecimentos que compartilhamos durante essa jornada.

Resumo

Agentes econômicos dependem e dispendem vastos recursos para a obtenção de previsões atualizadas da inflação para a definição de expectativas de mercado e tomada de decisões. Este trabalho explora diferentes modelos de regressão para a previsão semanal da inflação brasileira a partir de dados de alta dimensão. Especificamente, para previsão do Índice de Preços ao Consumidor Amplo (IPCA) são utilizados como potenciais variáveis preditoras as séries de subitens do Índice de Preços ao Consumidor do Município de São Paulo (IPC-FIPE), divulgadas semanalmente. Diversos métodos de regressão, incluindo Regressão Ridge, LASSO, Elastic Net, Regressão por Componentes Principais e Regressão por Mínimos Quadrados Parciais, Complete Subset Regressions e Best Subset Selection foram avaliados para a previsão do IPCA. Os modelos de previsão para a inflação mensal foram ajustados, separadamente aos dados de cada semana, num esquema de validação fora da amostra e então avaliados e comparados. Os resultados indicaram que as técnicas baseadas em seleção de variáveis geraram melhores previsões do que os métodos que agregam informação de todas as variáveis. Em destaque, os métodos Complete Subset Regressions e Best Subset Selection apresentaram os melhores desempenhos, o que salienta a importância da seleção de variáveis nesta aplicação.

Palavras-chave: Previsão de inflação. Dados de alta dimensão. Regressão. Aprendizado de máquinas. Validação fora da amostra. Redução de dimensão. Seleção de variáveis

Abstract

Economic agents depend on and expend extensive resources to obtain up-to-date inflation forecasts to set their expectations and for decision making. This work explores different regression models for the weekly forecast of the Brazilian inflation using high-dimensional data. Specifically, to forecast the Extended National Consumer Price Index (IPCA), hundreds of series of subitems that compose the Consumer Price Index of São Paulo City (IPC-FIPE), which is released weekly, are used as potential predictor variables. Several regression algorithms, including Ridge Regression, LASSO, Elastic Net, Principal Component Regression, Partial Least Squares Regression, Complete Subset Regressions, and Best Subset Selection, were evaluated for IPCA forecasting. The monthly inflation forecasting models were fitted separately for each week in an out-of-sample validation scheme and subsequently evaluated and compared. The results indicated that variable selection-based techniques provided better forecasts than those based on aggregation of information from all variables. In particular, the Complete Subset Regressions and Best Subset Selection methods exhibited the best performances, which highlights the importance of variable selection in this application.

Keywords: Inflation forecasting. High-dimensional data. Regression. Machine learning. Out-of-sample validation. Dimensionality reduction. Variable selection

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A inflação, fenômeno definido como o aumento geral de preços de bens e serviços na economia, é uma das preocupações centrais para agentes econômicos bem como para a elaboração de políticas econômicas, pois reflete diretamente no poder de compra da moeda. Dentre as suas causas, estão o aumento de demanda, de custos de produção ou mesmo da própria expectativa de inflação (BANCO CENTRAL DO BRASIL, 2024). Por exemplo, custo de vida, preços de aluguéis, taxas de juros, rentabilidade de diversos títulos públicos e privados mudam com a inflação. Maior inflação significa maior incerteza na economia, ou seja, há um impacto direto sobre os investimentos e crescimento do país. Deste modo, a capacidade de antecipar a inflação em tempo hábil é fundamental não só para o cumprimento das metas de inflação pelo Banco Central — através do controle da taxa de juros — como também para a formação de expectativas de mercado e tomada de decisão por parte de empresas e consumidores.

O Nowcasting é definido por Bańbura et al. (2013) como a previsão do presente, do futuro próximo e do passado recente. Esse conceito surge, em parte, da necessidade de antecipar variáveis divulgadas com atraso — como é o caso do Índice de Preços ao Consumidor Amplo (IPCA), indicador oficial de inflação no Brasil — e ao mesmo tempo, da disponibilidade crescente de dados econômicos mais atualizados. O índice mensal, produzido pelo Instituto Brasileiro de Geografia e Estatística (IBGE), normalmente é divulgado apenas no oitavo dia útil do mês subsequente ao de referência. Contudo, agentes de mercado necessitam atualizar as suas expectativas com rapidez.

Semelhantemente ao IPCA, o Índice de Preços ao Consumidor do Município de São Paulo (IPC-FIPE), da Fundação Instituto de Pesquisas Econômicas (FIPE), também mede a inflação no custo de vida, e embora sua população-alvo seja mais limitada demograficamente, esse índice é divulgado a cada semana, não só mensalmente, e em geral com apenas três dias de atraso. Não só isso, o índice é desagregado por grupos, subgrupos, itens e centenas de subitens da cesta de produtos e serviços das famílias paulistanas. Essa tempestividade e nível de detalhamento do IPC-FIPE criam a oportunidade de antever o IPCA e a inflação no país em quatro momentos dentro de um só mês.

O objetivo deste trabalho é, assim, realizar a previsão da inflação brasileira segundo mede o IPCA com base em um grande número de variáveis, sendo estas os índices que compõem o IPC-FIPE. Para lidar com essa enorme quantidade de preditores, inclusive neste caso excedendo o número de observações do conjunto de dados, serão empregados diversos algoritmos de aprendizado de máquinas. Em particular, técnicas de regularização

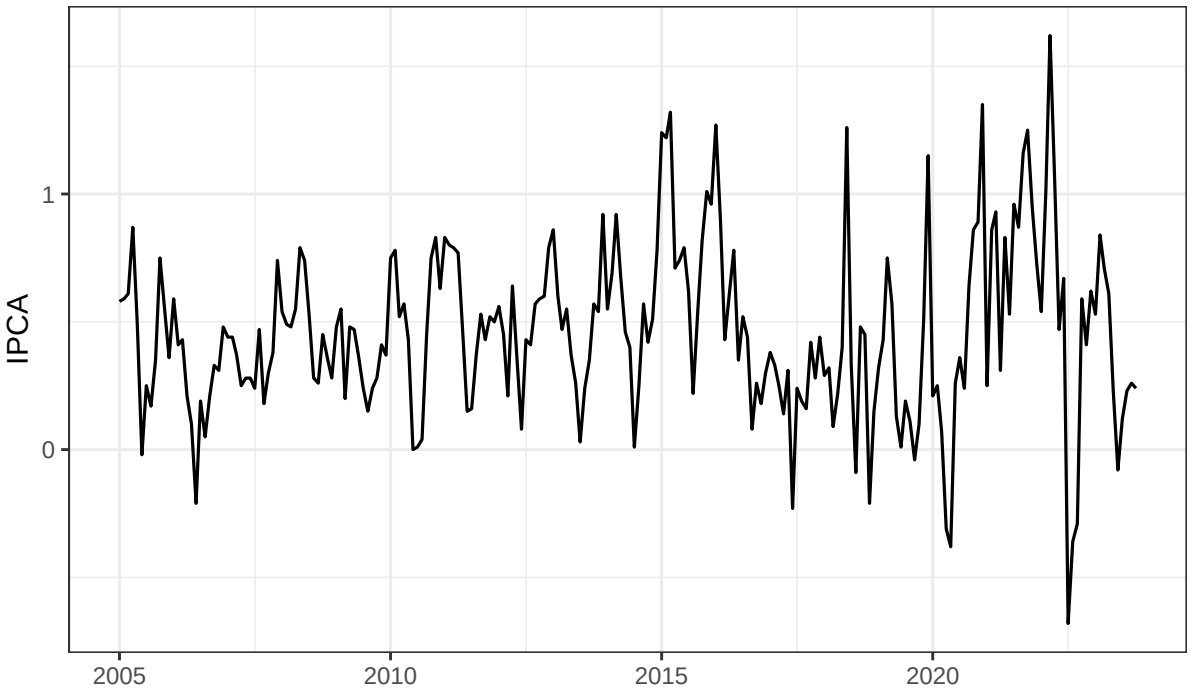
e seleção de variáveis (LASSO, Elastic Net e Best Subset Selection), de redução de dimensionalidade (Principal Component Regression e o Partial Least Squares Regression) e de comitê (Complete Subset Regressions), em que se agregam submodelos.

A estrutura do trabalho é a seguinte. No capítulo 2 são apresentados a teoria por trás dos modelos considerados para ajuste e também são discutidas e ilustradas algumas questões importantes a considerar no contexto de previsão, em particular, os problemas da multicolinearidade e do *overfitting* (sobreajuste). Estes problemas estão muito presentes em dados de alta dimensão, com muitos preditores. A validação fora da amostra pode ser útil para evitar esses problemas ao se construir modelos preditivos. Na capítulo 3 são apresentados os resultados de cada modelo em isolado, realizando-se comparações entre eles parar tentar extrair alguma interpretação das variáveis de maior peso na inflação. Por fim, na seção 4 as conclusões finais do trabalho são apresentadas.

2 Conjunto de dados e tratamento

O indicador oficial de inflação brasileira é o Índice de Preços ao Consumidor Amplo (IPCA). Seu objetivo é acompanhar a variação mensal de preços de uma cesta de produtos e serviços consumidos pelas famílias brasileiras. O índice mensal do IPCA, produzido pelo Instituto Brasileiro de Geografia e Estatística (IBGE), é calculado com base em dados coletados em estabelecimentos comerciais e de prestação de serviços, concessionárias de serviços públicos e na internet, e mede a inflação de preços em relação ao mês anterior ao de referência. Sua população-alvo é as famílias com rendimentos de 1 a 40 salários mínimos que residem em áreas urbanas das principais regiões metropolitanas do país, o Distrito Federal e alguns outros municípios abrangidos pelo Sistema Nacional de Índices de Preços ao Consumidor (SNIPC). Para cada quadrissemana, o índice mede a inflação de preços médios das quatro semanas de referência em relação às quatro semanas anteriores a essas. Os dados são coletados ao longo do mês de referência e o índice é divulgado até o dia 15 do mês subsequente, entre 8 a 14 dias após o fim da coleta.

Figura 1: Inflação mensal histórica segundo o Índice de Preços ao Consumidor Amplo (IPCA).



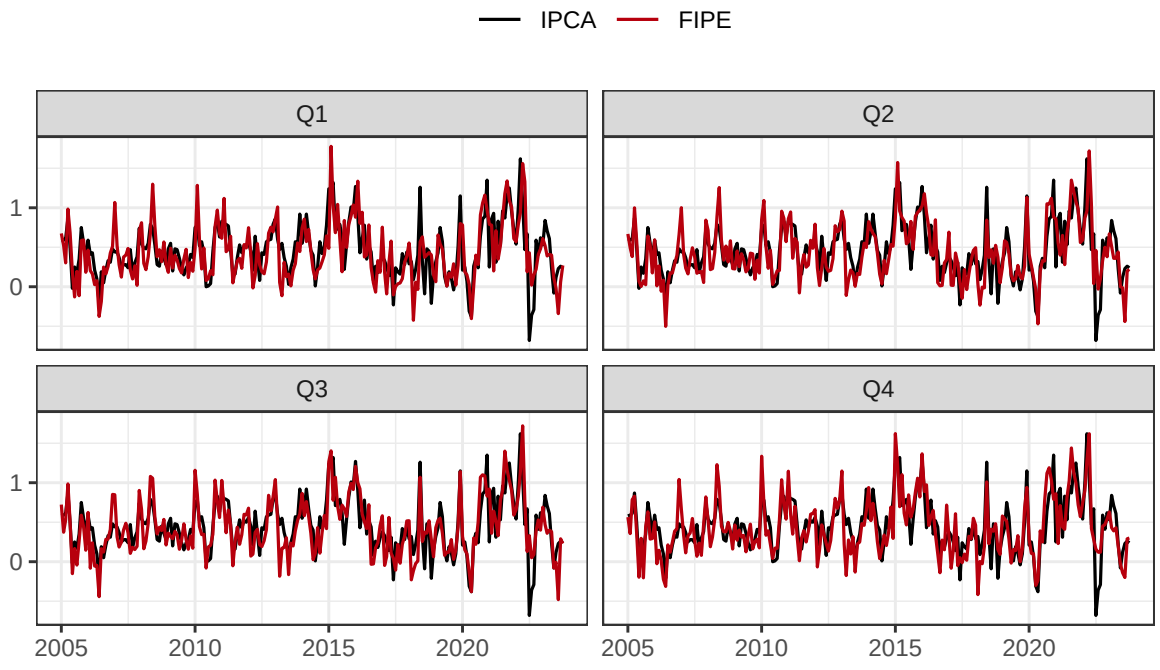
Dados de janeiro de 2005 a outubro de 2023.

Da mesma forma, o Índice de Preços ao Consumidor do Município de São Paulo (IPC-FIPE) também é um indicador utilizado para medir a inflação no custo de vida, mas

tem como alvo especificamente as famílias da cidade de São Paulo com renda entre 1 a 10 salários mínimos. Calculado pela Fundação Instituto de Pesquisas Econômicas (FIPE), o indicador é divulgado não só mensalmente como semanalmente, desagregado por grupo, subgrupo, item e subitem da cesta de produtos, e divulgado costumeiramente em até 3 dias após a semana ou mês de referência.

O conjunto de dados completo cobre o período de janeiro de 2005 a outubro de 2023, num total de 226 observações, e é composto por, além da série do IPCA, 553 índices mensais do IPC-FIPE medindo da inflação de preços em quatro níveis crescentes de desagregação, sendo 7 grupos, 29 subgrupos, 55 itens e 462 subitens que figuram na cesta de produtos do índice geral. Contudo, para os propósitos desse trabalho foram consideradas apenas as séries de subitens [1]. Finalmente, uma vez que essas séries tiveram seu registro iniciado em momentos diferentes, algumas apenas após janeiro de 2005, foram descartadas aquelas com mais de 30% de valores faltantes, e as demais foram preenchidas com a respectiva mediana de todo o período na ausência de registro. Desse modo, após os tratamentos restaram 399 índices de subitens em cada quadrissemana.

Figura 2: Índice de Preços ao Consumidor do Município de São Paulo (IPC-FIPE) e o IPCA.



Dados de janeiro de 2005 a outubro de 2023.

[1] Não houve nenhum ganho ao utilizar a base completa de grupos, subgrupos, itens e subitens em conjunto.

Figura 3: Estrutura dos componentes e códigos do IPC-FIPE.

Grupo	**1** Habitação
Subgrupo	**11** Manutenção do domicílio
Item	**1101** Serviços de utilidade pública
Subitem	**1101001** Energia elétrica

Figura 4: Grupos de itens do IPC-FIPE.

1 Habitação
2 Alimentação
3 Transportes
4 Despesas pessoais
5 Saúde
6 Vestuário
7 Educação

3 Metodologia

Os métodos utilizados para a realização das previsões serão descritos neste capítulo. Considere y_{t+h} a variável resposta observada h períodos após o tempo (mês) t, sendo h o horizonte de previsão, que, no contexto de *nowcasting*, é curto (p.e. $h \in \{-1, 0, 1\}$), sendo negativo quando há defasagem de divulgação da variável resposta em relação às variáveis preditoras. Aqui, y_{t+h} será modelada como uma função de preditores observados no tempo $t = 1, 2, \ldots, n$, como por exemplo

$$y_{t+h} = f(\mathbf{x}_t) + \varepsilon_{t+h},$$

em que f é uma função do vetor de p preditores $\mathbf{x}_t$, e ε_{t+h} são componentes de erros com média zero. O vetor $\mathbf{x}_t$ pode incluir preditores exógenos, defasagens da variável resposta, componentes principais e etc. Na maior parte dos casos a seguir, será assumida uma estrutura de regressão linear, isto é,

$$y_{t+h} = \beta_0 + \mathbf{x}_t^\top \boldsymbol{\beta} + \varepsilon_{t+h}. \tag{1}$$

Embora os modelos de regressão não sejam os mais apropriados para lidar com dados temporais, no caso presente estamos lidando com previsões em tempo real, em particular, o escopo deste trabalho será no horizonte de previsão $h = 0$, ou seja, o interesse é antecipar o IPCA do mês com dados de menor defasagem de divulgação do próprio mês, logo, modelar a dinâmica temporal de y_t se torna menos necessário.

Defina $\hat{y}_{t+h} = f(\mathbf{x}_t; \hat{\beta}_0, \hat{\boldsymbol{\beta}})$ como a previsão de y_{t+h}, para estimativas $\hat{\beta}_0$ e $\hat{\boldsymbol{\beta}}$ dos parâmetros β_0 e $\boldsymbol{\beta}$, e seja $e_{t+h} = y_{t+h} - \hat{y}_{t+h}$ o erro de previsão. O desempenho de previsão das técnicas será avaliado pela métrica raiz do erro quadrático médio de predição ou *root mean squared error* (MSE), que no conjunto de dados completo é definido por

$$\widehat{\mathrm{RMSE}} = \sqrt{\frac{1}{n^*} \sum_{t=1}^{n^*} e_{t+h}^2},$$

sendo $n^* = n - |h|$ o número efetivo de observações no conjunto de dados, e h o horizonte de previsão.

3.1 Modelo autorregressivo

Para fins de comparação e como ponto de referência para outros modelos, defina como *benchmark* o seguinte modelo autorregressivo de ordem 1 com *drift*:

$$y_t = \beta_0 + \beta y_{t-1} + \varepsilon_t.$$

3.2 Análise de componentes principais

A análise de componentes principais (PCA - *principal component anaylsis*) é uma das técnicas mais populares no contexto de redução de dimensão, tanto para análise exploratória, quanto para eliminação de redundâncias em problemas de regressão, classificação, etc. O PCA se baseia em encontrar um conjunto de combinações lineares de $\mathbf{x}$, denominadas componentes principais (PCs), que sejam não correlacionadas — isto é, não redundantes — e que contabilizam a maior parte da variabilidade de $\mathbf{x}$.

Seja $\mathbf{X} = (\mathbf{x}_t^\top)_{t=1}^{n^*}$ a matriz das variáveis preditoras padronizadas, $\mathbf{\Sigma}$ a matriz variância-covariância de $\mathbf{X}$ e r o posto de $\mathbf{\Sigma}$. O escore para a componente principal $i = 1, \ldots, r$ no tempo t é dado pela combinação linear $c_{ti} = \mathbf{x}_t^\top \mathbf{a}_i$, em que os pesos $\mathbf{a}_1, \ldots, \mathbf{a}_r$ são obtidos sequencialmente através do problema de otimização

$$\arg\max_{\mathbf{a}_i \in S_i} \mathbf{a}_i^\top \mathbf{\Sigma} \mathbf{a}_i,$$

sendo S_i o espaço de p-vetores reais unitários ortogonais a $\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}$. Com isso, a variabilidade da i-ésima componente é dada por $\lambda_i = \mathbf{a}_i^\top \mathbf{\Sigma} \mathbf{a}_i$ e verifica-se que $\lambda_i \geq \lambda_j, \forall j > i$. Espera-se que a fração variabilidade total de $\mathbf{X}$ coberta pelas primeiras k componentes, $(\lambda_1 + \cdots + \lambda_k)/\mathrm{tr}(\mathbf{\Sigma})$, seja próxima de 1 para algum $k \ll r$. A matriz $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$ é chamada de matriz de pesos ou *loadings*, e $\mathbf{C} = \mathbf{X}\mathbf{A}$ matriz de escores.

Uma introdução moderna com interpretações geométricas e aplicações, variantes e desenvolvimentos recentes no método de componentes principais para dados de alta dimensão, dados mistos, dados incompletos, geração de componentes interpretáveis (esparsas), entre outros, pode ser encontrada no artigo de Greenacre et al. (2022).

3.3 Modelos baseados em componentes principais

3.3.1 Regressão por Componentes Principais

Um primeiro tipo de modelagem que pode vir à mente nesse tipo de caso é a regressão linear clássica. No entanto, quando o número de preditores p é grande, possivelmente excedendo o número de observações n, existe um problema de sobreajuste ou *overfitting*, que é quando um determinado modelo se ajusta muito bem a um dado conjunto de dados mas as suas previsões em uma amostra independente se mostram muito inferiores ao esperado. Além do mais, mesmo que $n^* < p$ é ainda possível que devido à presença de variáveis muito correlacionadas as estimativas em $\boldsymbol{\beta}$ sejam instáveis (com alto erro padrão), problema este conhecido como *multicolineariedade*, e como consequência geram previsões menos confiáveis para y_{t+h}.

Uma saída a esses problemas que permite utilizar a informação em todas as variáveis preditoras sem descartar alguma em particular é a regressão por componentes principais (PCR - *principal component regression*) proposta inicialmente por Massy (1965). Em síntese, o PCR consiste em ajustar um modelo de regressão como (1) colocando no lugar de $\mathbf{x}_t$ os k primeiros componentes principais de $\mathbf{x}_t$, ou seja,

$$y_{t+h} = \beta_0 + \mathbf{c}_t^{(k)}\boldsymbol{\beta} + \varepsilon_{t+h}, \tag{1}$$

em que $\mathbf{C}^{(k)} = (\mathbf{c}_t^{(k)})_{t=1}^{n^*}$ contém as primeiras k colunas da matriz de escores $\mathbf{C}$, sendo $k < \min(n^*, p)$.

3.3.2 Regressão de Mínimos Quadrados Parciais

Em contraste ao PCR, na regressão de mínimos quadrados parciais (PLSR - *partial least squares regression*) as componentes principais são geradas levando em conta a informação da variável resposta y. Mais precisamente, no PLSR com resposta univariada as componentes principais são obtidas sequencialmente a partir de combinações lineares dos preditores que maximizam a covariância com y e, novamente, sob a condição de não correlação com as componentes anteriores (GARTHWAITE, 1994).

Seja $\mathbf{y} = (y_t)_{t=1}^{n^*}$ o vetor de respostas, $\mathbf{X} = (\mathbf{x}_t^\top)_{t=1}^{n^*}$ a matriz das variáveis preditoras padronizadas. Analogamente ao PCA, no PLSR o escore para a componente principal $i = 1, \ldots, r$ no tempo t é dado pela combinação linear $c_{ti} = \mathbf{x}_t^\top \mathbf{a}_i$, em que os pesos $\mathbf{a}_1, \ldots, \mathbf{a}_r$ são obtidos sequencialmente através da maximização da covariância da

componente com a resposta, ou seja,

$$\arg\max_{\mathbf{a}_i \in S_i} \text{Cov}(\mathbf{X}\mathbf{a}_i, \mathbf{y}).$$

Então, utilizando as k componentes $\mathbf{C}^{(k)}$ no lugar de $\mathbf{X}$ no Modelo (1), como feito na Equação (1) no modelo PCR, temos o chamado modelo PLSR.

3.4 Seleção de variáveis e regularização

Ao ajustar um modelo linear, nem sempre há o interesse ou é desejável ou mesmo possível que todos os preditores disponíveis sejam inclusos no modelo. Primeiro porque é comum a presença de preditores irrelevantes, que podem ser descartados sem perdas. Segundo, porque em muitos casos, não este, é interessante manter um modelo simples pela facilidade de interpretação. Terceiro, porque há casos como o presente em que existem mais preditores disponíveis do que observações para ajuste, e isso torna impossível o ajuste tradicional por mínimos quadrados ordinários (MQO). Quarto, porque a inclusão de preditores correlacionados induz a multicolineariedade, que está relacionada à instabilidade do modelo e à incerteza nas estimativas.

3.4.1 Best Subset Selection

Um procedimento comum para decidir quais preditores incluir é o best subset selection (BSS), que envolve o ajuste de todos os modelos possíveis, e então é escolhido o melhor subconjunto de variáveis com base em um critério (AIC, BIC, C_p e R^2). Todavia isso rapidamente se torna computacionalmente inviável, posto que o número de modelos a serem testados cresce exponencialmente. Outra possibilidade computacionalmente mais eficiente é a seleção stepwise, que parte de um modelo inicial e a cada passo vai adicionando ou retirando uma variável (a depender do procedimento) até que se atinja um ponto otimal, tal que qualquer mudança pontual piore um determinado critério. Embora a seleção stepwise seja mais eficiente ela continua a ser um procedimento discreto como o BSS, e portanto ambos sofrem de instabilidade e podem entregar resultados radicalmente diferentes com a menor alteração no conjunto de dados.

Uma forma diferente diferente de atingir esse objetivo de seleção de variáveis é por meio da regularização, que trata-se de um tipo de penalização que força os coeficientes ajustados a zero. Apesar de essa penalização introduzir um viés nas estimativas, a redução

de variância pode contribuir para a melhora do modelo. Três dos principais métodos de regularização são a regressão ridge, o LASSO e o elastic net, que serão abordadas a seguir.

3.4.2 Regressão Ridge

A regressão ridge é semelhante ao ajuste por mínimos quadrados ordinários, e se baseia em minimizar a soma de quadrados dos resíduos mais uma penalização quadrática (ℓ_2) dos coeficientes exceto o intercepto. A estimativa por regressão ridge para o vetor de parâmetros $(\beta_0, \boldsymbol{\beta}^\top)^\top$ no modelo de regressão da Equação 1 é obtida a partir da solução do seguinte problema de mínimos quadrados penalizado:

$$\underset{\beta_0,\boldsymbol{\beta}}{\arg\min} \left[\sum_{t=1}^{n^*} (y_{t+h} - \beta_0 - \mathbf{x}_t^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^2 \right],$$

em que $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^2$ representa a penalização ℓ_2, sendo λ o parâmetro de regularização a ser calibrado.

3.4.3 LASSO

Desenvolvido por Tibshirani (1996), o LASSO (*Least Absolute Shrinkage and Selection Operator*) é estimado de forma semelhante à regressão ridge, trocando apenas a penalização ℓ_2 pela penalização ℓ_1, $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$, ou seja,

$$\underset{\beta_0,\boldsymbol{\beta}}{\arg\min} \left[\sum_{t=1}^{n^*} (y_{t+h} - \beta_0 - \mathbf{x}_t^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right].$$

É possível mostrar que tanto a regressão ridge quanto o LASSO podem ser reenquadrados, para um λ fixado, como o seguinte problema

$$\begin{aligned} \underset{\beta_0,\boldsymbol{\beta}}{\arg\min} \quad & \sum_{t=1}^{n^*} (y_{t+h} - \beta_0 - \mathbf{x}_t^\top \boldsymbol{\beta})^2 \\ \text{sujeito a} \quad & P(\boldsymbol{\beta}) \leq b, \end{aligned}$$

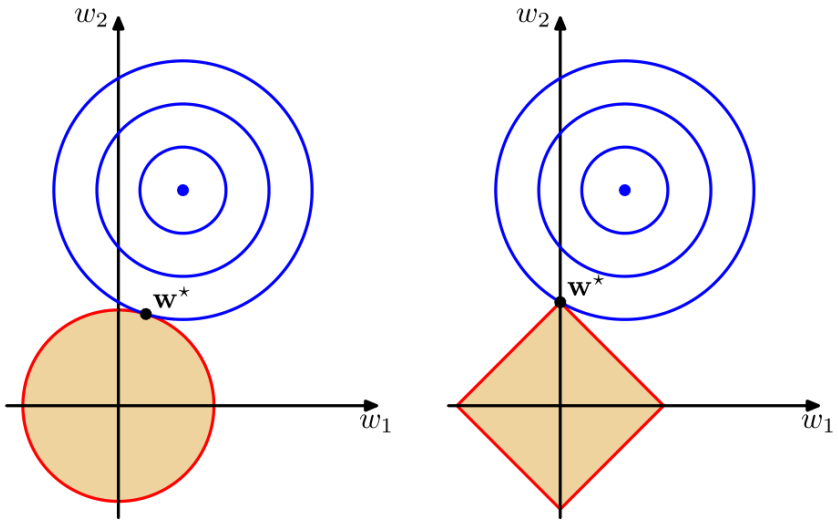
isto é, um problema de estimação por mínimos quadrados com restrição sobre o "tamanho"de $\boldsymbol{\beta}$.

Note que tanto na regressão ridge quanto no lasso, para $\lambda = 0$ as estimativas coincidem com as de MQO, desde que $p < n^*$. Veja ainda que conforme $\lambda \to \infty$, $\hat{\boldsymbol{\beta}} \to \mathbf{0}$,

o que resulta num ajuste incluindo apenas o intercepto. Assim, ajustar o parâmetro λ trata-se de encontrar um balanço entre esses dois extremos.

Como aponta Friedman, Hastie e Tibshirani (2010), e é ilustrado na Figura 5, num conjunto de preditores a penalização LASSO tende zerar os coeficientes até que sobre apenas um não-nulo, ou seja, o LASSO atua como um procedimento de seleção de variáveis. Por outro lado, a regressão ridge tende diluir os coeficientes, sem zerar nenhum. Por esse motivo, Zou e Hastie (2005) propuseram o elastic net, um híbrido que aproveita a qualidade de estabilidade da regressão ridge em casos com preditores extremamente correlacionados e capacidade de seleção de preditores do LASSO.

Figura 5: Gráfico das linhas de contorno da soma de quadrados dos erros (azul) e as restrições de regularização para a regressão ridge (esquerda) e para o LASSO (direita).



Fonte: Bishop (2006).

3.4.4 Elastic Net

A rede elástica (*elastic net*) é um procedimento que ao mesmo tempo realiza a seleção de variáveis e regularização (*shrinkage*) de coeficientes, permitindo inclusive estimar modelos lineares generalizados superparametrizados ($p > n^*$). Esse método de estimação combina a penalização ℓ_1 do LASSO e a penalização ℓ_2 da regressão ridge, e visa promover um melhor equilíbrio entre viés e variância ao reduzir coeficientes de preditores menos importantes a zero.

A estimativa por elastic net para o vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^\top)^\top$ no modelo de regressão da Equação 1 é dada pela solução do seguinte problema de mínimos quadrados

penalizado:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2n^*} \sum_{t=1}^{n^*} (y_{t+h} - \mathbf{x}_t^\top \boldsymbol{\beta})^2 + \lambda P_\alpha(\boldsymbol{\beta}_1) \right],$$

sendo $P_\alpha(\boldsymbol{\beta}_1) = \alpha\|\boldsymbol{\beta}_1\|_1 + (1-\alpha)\frac{1}{2}\|\boldsymbol{\beta}_1\|_2$ a penalização elastic net, com parâmetros de mistura $\alpha \in [0,1]$ e penalização $\lambda \geq 0$ a serem ajustados por validação cruzada, critério de informação, etc.

Observe que o elastic net com $\alpha = 1$ é equivalente ao LASSO simples e $\alpha = 0$ equivale à regressão ridge com penalização λ. Assim, se α é pouco menor que 1, o elastic net se comporta de semelhante ao LASSO, porém mais bem comportado caso existam preditores fortemente correlacionados, devido à penalização ℓ_2. Ainda, conforme λ aumenta, a penalização ℓ_1 leva a soluções cada vez mais esparsas, com menos preditores afetando as previsões $\hat{y}_{t+h}$.

Um algoritmo eficiente para ajustar de forma simultânea toda a sequência de estimativas $\{\hat{\boldsymbol{\beta}}_\lambda \colon \lambda \geq 0\}$, fixado α, é apresentado por Friedman, Hastie e Tibshirani (2010).

3.5 Comitês

Com frequência descobre-se que as previsões combinadas de vários modelos são em geral melhores que as previsões de quaisquer desses modelos em separado (TIMMERMANN, 2006). No contexto de regressão e séries temporais, por exemplo, a forma mais simples e popular de combinação é tomar como previsão a média simples das previsões de modelos individuais. Bishop (2006) mostra que, sob condições razoáveis, esse procedimento diminui o erro quadrático médio de previsão. Tais métodos de combinação de vários modelos são conhecidos na literatura como comitês de máquinas (*committee machines*). Outros métodos explorados no contexto de aprendizado de máquinas são o *bagging*, *boosting* e *stacking*, onde também são conhecidos como métodos de *ensemble*.

3.5.1 Complete Subset Regressions

Em um caso com p preditores, existem $2^{p^*} - 1$ possíveis modelos de regressão que podem ser ajustados via mínimos quadrados com $1, 2, \ldots, p^* = \min\{p, n\}$ preditores. Se todos esses modelos são ajustados para encontrar um melhor subconjunto de preditores com base em um dado critério, tem-se o chamado Best Subset Selection (BSS), contudo,

com p grande, torna-se computacionalmente inviável ajustar todos esses modelos.

A ideia do *complete subset regressions* (CSR) proposto por Elliott, Gargano e Timmermann (2013) se baseia em ajustar todos os $n_{p,k} = \binom{p}{k}$ modelos de regressão linear com exatamente k preditores, onde $k \leq p$. Assumindo sem perda de generalidade que os preditores e a resposta estão padronizados, esses modelos individuais não possuirão intercepto, de modo que a previsão por CSR fica sendo

$$\hat{y}_{t+h} = \frac{1}{n_{p,k}} \sum_{i=1}^{n_{p,k}} \mathbf{x}_t^\top \hat{\boldsymbol{\beta}}_i,$$

onde $\hat{\boldsymbol{\beta}}_i$ é um vetor de zeros, exceto nas posições que correspondem a variáveis no i-ésimo subconjunto de $\mathbf{x}_t$, onde entram as estimativas de mínimos quadrados do modelo linear.

Uma vantagem desse método é que ele se aplica a situações com mais preditores do que observações, sem necessariamente descartar algum preditor.

3.6 Pré-filtragem de preditores via LASSO

Dois dos métodos descritos anteriormente, Best Subset Selection e Complete Subset Regressions, podem não ser viáveis computacionalmente sem que antes haja uma pré-filtragem de preditores, pois necessitam de $2^{\min(p,n-1)} - 1$ e $\binom{p}{k}$ ajustes de modelos no total, respectivamente. Em particular, tendo disponíveis $p = 399$ preditores e $n = 226$ observações, torna-se impossível ajustar todos os modelos no BSS, e apenas valores muito pequenos de k podem ser experimentados no CSR.

Um procedimento muito comum para realizar essa filtragem é selecionar preditores mais correlacionados com a variável resposta. No entanto, esse procedimento ingênuo tende a introduzir multicolinearidade quando os preditores forem muito associadas entre si, prejudicando a estabilidade do modelo. Garcia, Medeiros e Vasconcelos (2017), por exemplo, faz a seleção dos chamados *targeted predictors* com base nos maiores valores absolutos na estatística t de modelos univariados (incluindo defasagens). Outra possibilidade é aproveitar variáveis selecionadas por métodos de regularização como o LASSO, que automaticamente elimina preditores redundantes.

Outra forma possivelmente mais interessante de realizar essa filtragem é aproveitando a capacidade de seleção de preditores do algoritmo LASSO. A ideia consiste em ajustar o parâmetro λ até que reste um número desejado de preditores com coeficientes não nulos, que são mantidos, sendo os demais preditores descartados da análise. Esse

método evita a inclusão de preditores muito correlacionados entre si e por consequência reduz a multicolinearidade.

3.7 Avaliação de modelos

Ao escolher um modelo de previsão particular como o "melhor", há, em última análise, um interesse em saber se esse modelo é o mais adequado, não só para prever a variável resposta de observações já conhecidas (*in-sample*), mas sim para observações futuras, ou pelo menos ainda não "vistas" (*out-of-sample*); uma ideia relacionada à capacidade de extrapolação ou generalização do modelo. Nesse sentido, modelos devem idealmente ser julgados com base em dados independentes, não utilizados para ajustá-los (*out-of-sample*).

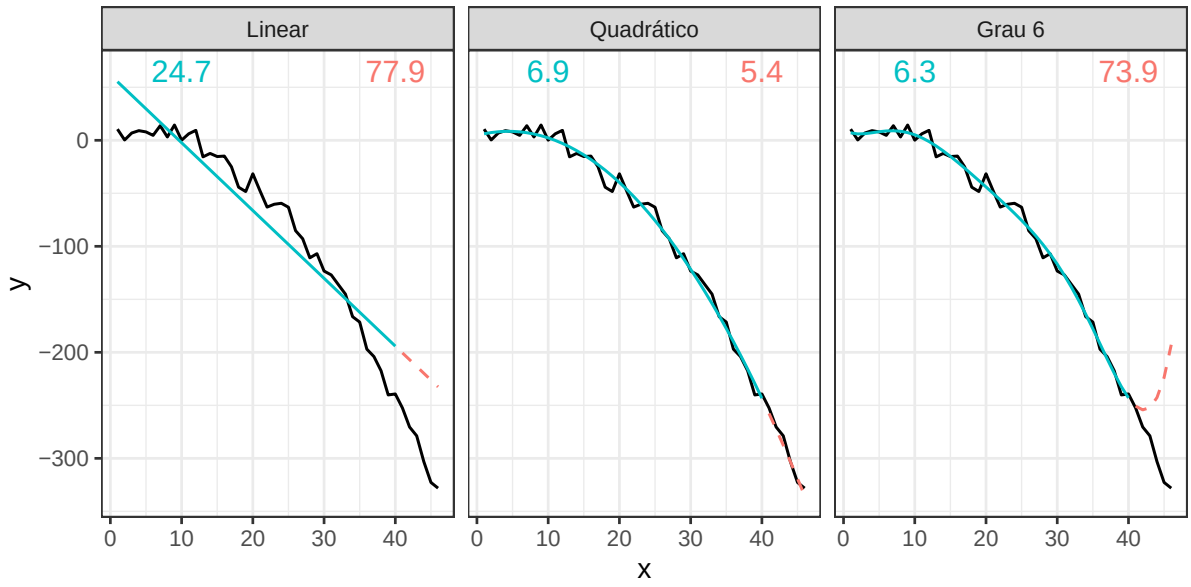
3.7.1 Sobreajuste e subajuste

Para ilustrar o ponto anterior, a Figura 6 mostra alguns modelos de diferentes complexidades ajustados a dados simulados. Foram gerados aleatoriamente 46 valores de y como função quadrática de x, mais um erro aleatório, $y = 0.5 + 2x + 0.2x^2 + \varepsilon$, sendo $x_1 = 1$, $x_2 = 2$, ..., $x_{46} = 46$ e erros $\varepsilon \sim \mathcal{N}(0, 6^2)$ independentes. Três modelos de regressão polinomial para y sobre x são então ajustados com base nas primeiras 40 observações, sendo um linear, um quadrático e um de grau 6. As seis observações restantes são utilizadas para teste.

Nota-se que o modelo linear se ajusta mal aos dados, tanto para as observações de treino quanto de teste, caracterizando subajuste ou *underfit*. O terceiro modelo, por outro lado, apresenta sobreajuste ou *overfit*, uma vez que a princípio parece bem ajustado, mas é incapaz de generalizar para novas observações. Já o segundo, o modelo quadrático teoricamente correto, consegue generalizar a relação entre y e x, mesmo não se ajustando melhor que polinômio de grau 6 aos dados de treinamento. Isso ocorre pois, apesar de modelos necessitarem de certa flexibilidade para se ajustar adequadamente aos dados, muita flexibilidade permite aos modelos se ajustar a perturbações irrelevantes que não se replicam a novas observações, o que gera instabilidade e maior incerteza nas previsões. Existe, portanto, um interesse em obter modelos que façam boas previsões e ao mesmo tempo a necessidade de que estas previsões sejam estáveis, um problema conhecido como equilíbrio entre viés e variância ou, na literatura estrangeira, *bias-variance trade-off*. No entanto, o exemplo deixa claro que julgar a qualidade de um modelo com base nos mesmos

dados aos quais ele foi ajustado pode não ser a melhor opção.

Figura 6: Exemplo de *underfitting* e *overfitting* em modelos de regressão.



A linha preta representa os valores observados, enquanto os valores ajustados pelos modelos polinomiais estão destacados para o conjunto de treinamento (azul) e de teste (vermelho). Os números na parte superior representam os respectivos RMSEs.

3.7.2 Validação cruzada

Postos esses problemas, surge a questão de como encontrar, num conjunto de modelos candidatos, o modelo mais adequado para fazer previsões, isto é, aquele que potencialmente tem o menor erro de previsão médio. Uma ideia inicial é a do conjunto de validação, onde dispõe-se de observações à parte sobre as quais os modelos candidatos ajustados devem ser testados e comparados entre si. No entanto, com frequência tal conjunto não está disponível. Outras técnicas como a validação cruzada tentam superar isso ajustando modelos e os testando múltiplas vezes em porções de dados diferentes, sempre avaliando os modelos em dados à parte do treino.

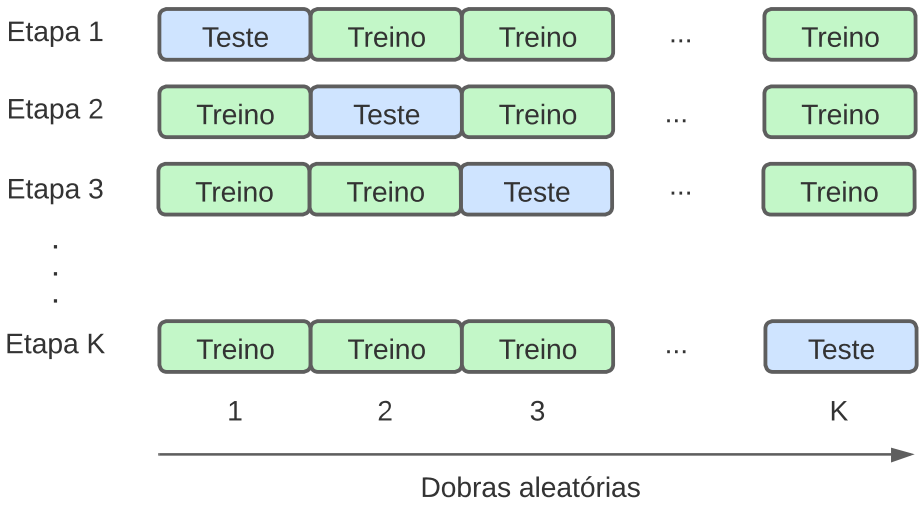
Uma das mais populares técnicas de validação cruzada é a validação K-*fold*, em que se divide os dados aleatoriamente em K partes (dobras) de tamanhos aproximadamente iguais e cada modelo é ajustado K vezes. Na etapa i, os modelos são avaliados tendo os dados da dobra i como teste após ajustados com os dados das demais $(K-1)$ dobras. Mais precisamente, seja $\kappa\colon \{1,2,\ldots,n^*\} \mapsto \{1,2,\cdots,K\}$ uma função que indica o mapeamento aleatório de cada observação $t = 1,\ldots,n^*$ às dobras $1,2,\ldots,K$. A

estimativa de validação cruzada para o erro de previsão fica definida como

$$\mathrm{CV}_{(K)}(\hat{f}) = \frac{1}{n^*}\sum_{t=1}^{n^*} L(y_{t+h}, \hat{y}_{t+h}^{-\kappa(t)}),$$

onde $\hat{y}_{t+h}^{-\kappa(t)}$ é a previsão de $y_{t+h} \mid \mathbf{x}_t$ gerada pelo modelo em questão ajustado ao conjunto de dados sem a dobra $\kappa(t)$, e $L(y, \hat{y})$ é uma métrica de erro de previsão, como o erro quadrático do MSEP. Embora em geral se utilize $K = 5$ ou 10 dobras, é possível ter até n^* dobras, caso esse chamado *leave-one-out*, já que se retira uma observação para teste em cada partição.

Figura 7: Ilustração da validação cruzada *K-fold*.



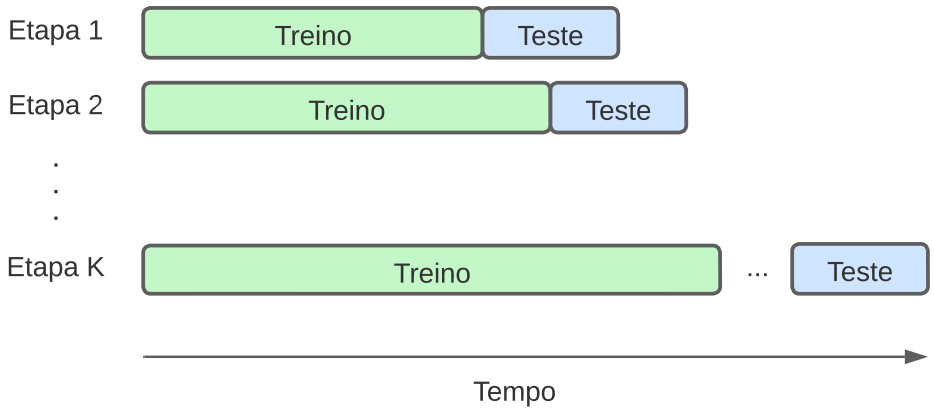
Todavia, esses procedimentos anteriores deixam de ser os mais adequados para a validação em modelagem preditiva de séries temporais uma vez que a aleatorização ignora a temporalidade dos dados. Outro problema é que os modelos são ajustados em dados à frente daqueles em que estes estão sendo avaliados, o que além de ser pouco lógico gera um viés de antecipação (*look-ahead bias*) na seleção dos modelos. Embora esse primeiro problema possa ser resolvido abrindo mão da aleatorização e mantendo a ordem natural dos dados, o viés de antecipação permanece na validação *K-fold*.

3.7.3 Validação por origem deslizante

Um outro procedimento de validação mais apropriado para modelagem de dados temporais é o de origem deslizante (TASHMAN, 2000), ilustrado na Figura 8, no qual os modelos são avaliados sempre à frente do tempo para o qual foram ajustados, evitando

o viés de antecipação. Nesse esquema, os conjuntos de treino e teste contíguos avançam no tempo, sendo que estes últimos se mantêm em um tamanho fixo, o qual é chamado de origem deslizante. Na literatura de economia, esse esquema também é chamado de "fora da amostra", e tipicamente são usados conjuntos de teste compostos de apenas uma observação. Opcionalmente, pode-se limitar também o tamanho dos conjuntos de treino, de modo que as observações mais iniciais vão sendo descartadas conforme se avança no tempo.

Figura 8: Ilustração da validação por origem deslizante.



Para o caso específico de validação por origem deslizante utilizando as K observações finais para teste, uma a uma, o erro de previsão fica definido como:

$$\mathrm{CV}_{\mathrm{OD}}(\hat{f}) = \frac{1}{K} \sum_{t=n^*-K+1}^{n^*} L(y_{t+h}, \hat{y}_{t+h}^{-(t+1:)}),$$

onde $\hat{y}_{t+h}^{-(t+1:)}$ é a previsão de $y_{t+h} \mid \mathbf{x}_t$ gerada pelo modelo em questão ajustado ao conjunto de dados sem as observações $t+1, t+2, \ldots, n$.

4 Resultados

Esta seção discutirá os resultados encontrados no ajuste dos modelos descritos anteriormente. Separadamente, para cada quadrissemana da base de subitens do IPC-FIPE, foram ajustados os mesmos modelos para o IPCA. Deste modo, não é necessária a modelagem da dinâmica temporal das variáveis preditoras. A calibração dos hiperparâmetros (número de componentes, nível de regularização, etc.) foi realizada visando minimizar a raiz quadrada do erro quadrático médio (RMSE) fora da amostra na validação por origem deslizante. O período de teste foi os últimos 36 meses da base de dados, ou seja, novembro de 2020 a outubro de 2023.

A Tabela 1 reúne os RMSEs fora da amostra para os modelos nas quatro quadrissemanas. Os modelos com menor RMSE em cada quadrissemana estão destacados em negrito. De modo geral, como esperado, observa-se que as previsões realizadas com base no IPC-FIPE da quarta quadrissemana têm menor erro e portanto são melhores que as da terceira quadrissemana, e assim por diante. Ou seja, quanto mais recentes forem as coletas, melhores são as previsões do IPCA para o mês.

Em destaque, o modelo best subset selection (BSS) foi melhor para todas as quadrissemanas, seguido pelo complete subset regressions (CSR), que trouxe resultados próximos. Os modelos LASSO e Elastic Net entregaram resultados bastante similares entre si, portanto, por simplicidade, neste caso faria mais sentido utilizar o LASSO em vez do Elastic Net, posto que o primeiro possui apenas um parâmetro ajustável. No entanto, nenhum desses dois últimos desempenhou tão bem quanto o CSR e o BSS. Ainda, observa-se que os modelos de regressão por componentes principais (PCR) e por mínimos quadrados parciais (PLSR) foram os piores para quase todas as quadrissemanas, com exceção da primeira, em que o LASSO e o Elastic Net estiveram ainda abaixo. Por fim, a regressão ridge e o *benchmark* AR foram os dois piores modelos em quase todas as quadrissemanas.

Tabela 1: Raiz quadrada do erro quadrático médio de previsão.

Modelo	**Q1**	**Q2**	**Q3**	**Q4**
AR	0,409	0,409	0,409	0,409
Ridge	0,375	0,367	0,341	0,322
LASSO	0,379	0,342	0,299	0,270
Elastic Net	0,377	0,339	0,299	0,270
PCR	0,352	0,354	0,336	0,322
PLSR	0,346	0,345	0,326	0,311
CSR	0,307	0,291	0,267	0,265
BSS	**0,300**	**0,284**	**0,254**	**0,241**

Raiz quadrada do erro quadrático médio de previsão (RMSE) na validação fora da amostra dos modelos ajustados aos dados do IPC-FIPE das quadrissemanas 1, 2, 3 e 4 do mês de referência. Os valores em negrito destacam os melhores modelos em cada quadrissemana.

Figura 9: Índice de Preços ao Consumidor Amplo (IPCA) e previsões fora da amostra de todos os modelos ajustados.

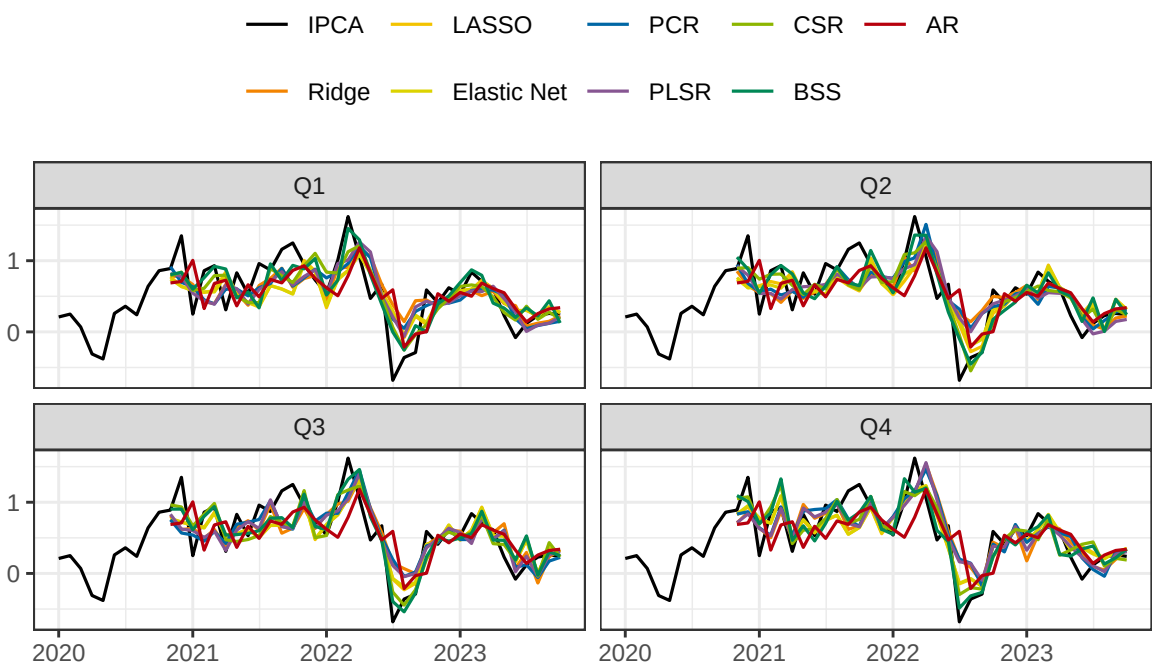
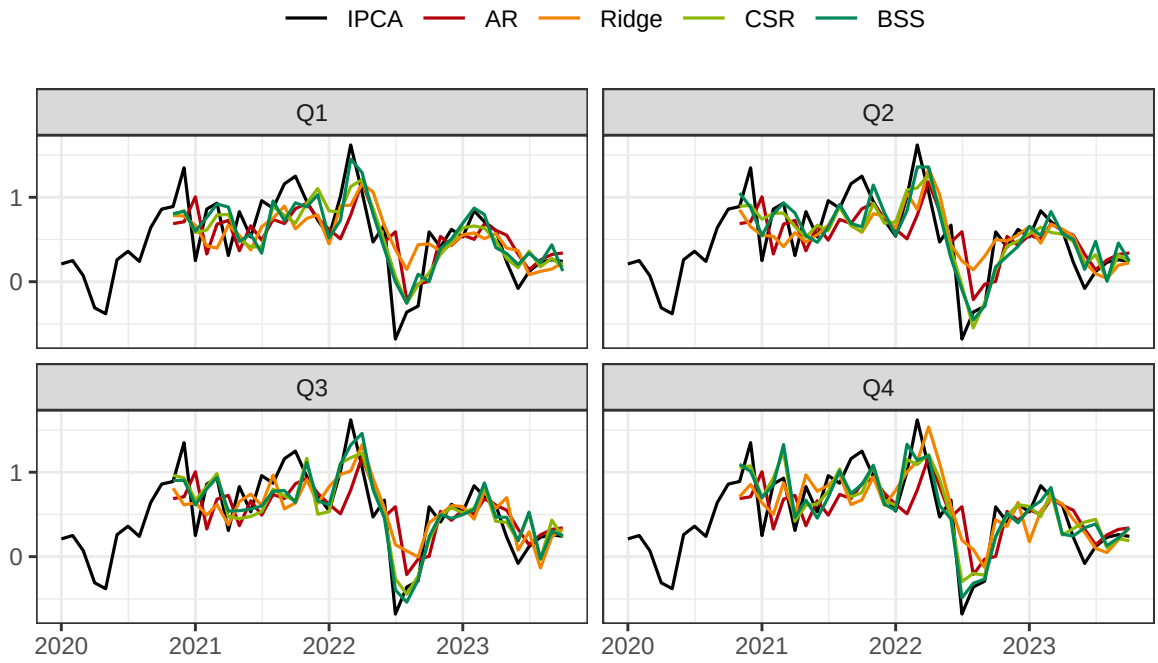


Figura 10: Índice de Preços ao Consumidor Amplo (IPCA) e previsões fora da amostra dos dois piores e dois melhores modelos ajustados.



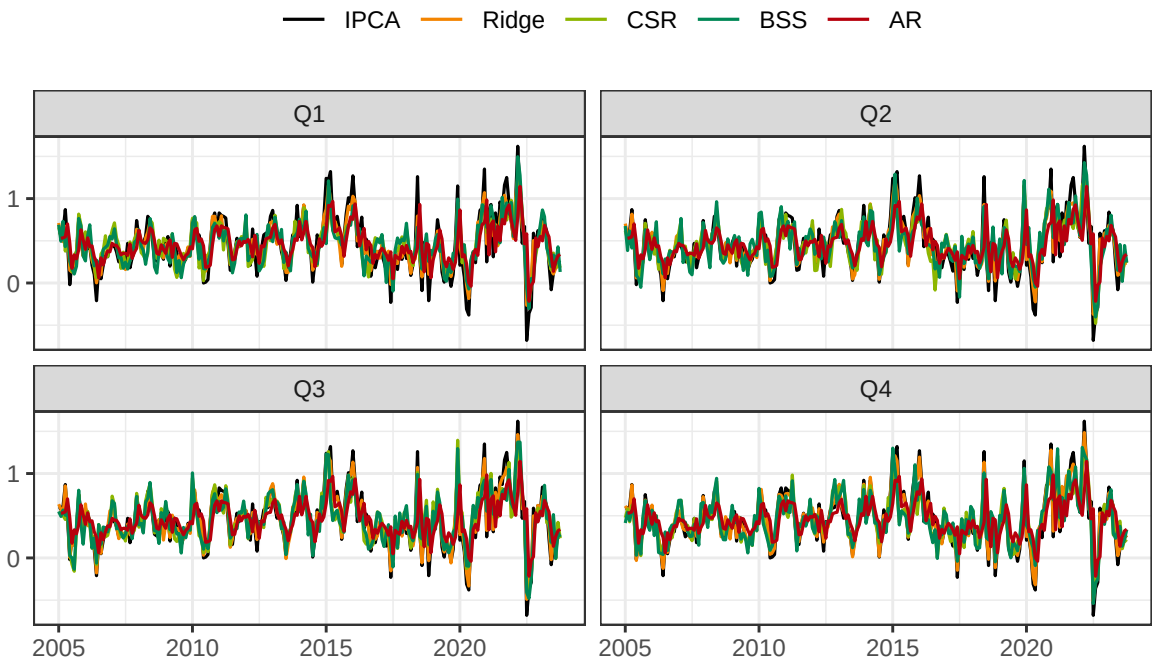
Uma vez que foram identificados os melhores modelos em cada quadrissemana, estes foram reajustados, com os mesmos hiper-parâmetros, à base de dados completa, compreendida pelas 226 observações do período de janeiro de 2005 a outubro de 2023. Se observa um grande contraste da qualidade do ajuste fora e dentro da amostra olhando as Figuras 10 e 11. Nota-se também pela Tabela 2 que o erro de previsão fora da amostra é muito maior que o erro de ajuste dentro da amostra para a regressão Ridge, que teve o pior desempenho em três das quatro quadrissemanas, em claro exemplo do fenômeno de *overfitting*. Por outro lado, o modelo *benchmark* AR, teve o pior desempenho tanto olhando para o ajuste dentro da amostra quanto no fora da amostra configurando, assim, um caso de *underfitting*. Essas observações destacam a importância de se estimar o erro de previsão utilizando dados fora do conjunto de ajuste, e como a validação fora da amostra pode ser útil para evitar modelos com esse comportamento indesejado.

Tabela 2: Raiz quadrada do erro quadrático médio.

Modelo	**Q1**	**Q2**	**Q3**	**Q4**
AR	0.271	0.271	0.271	0.271
Ridge	**0.129**	**0.106**	**0.069**	**0.054**
LASSO	0.195	0.164	0.127	0.121
Elastic Net	0.199	0.154	0.128	0.114
PCR	0.174	0.134	0.106	0.086
PLSR	0.148	0.171	0.083	0.077
CSR	0.226	0.202	0.180	0.185
BSS	0.221	0.198	0.178	0.172

Raiz quadrada do erro quadrático médio (RMSE) no ajuste na amostra completa dos modelos ajustados aos dados do IPC-FIPE das quadrissemanas 1, 2, 3 e 4 do mês de referência. Os valores em negrito destacam os melhores modelos em cada quadrissemana.

Figura 11: Índice de Preços ao Consumidor Amplo (IPCA) e valores ajustados na amostra completa dos dois melhores modelos ajustados e a regressão ridge.



4.1 Regressão Ridge e LASSO

As Figuras 12 e 13 apresentam o RMSE da regressão ridge e do LASSO na validação fora da amostra considerando diferentes valores de penalização (λ). Se percebe que para otimização do erro de validação os modelos ajustados à primeira quadrissemana

requerem maiores níveis de regularização do que para a segunda quadrissemana, e assim por diante. Posto ainda que a regularização diminui a quantidade de variáveis no modelo no caso do LASSO ou reduz a influência dessas variáveis no caso da regressão ridge, isso quer dizer que a inflação de alguns subitens deixa de ser relevante para mais longos prazos, o que pode explicar por que as previsões na primeira quadrissemana são menos voláteis que as da quarta, como se vê nas Figuras 9 e 10.

Figura 12: Curvas de RMSE fora da amostra do modelo de regressão ridge em função do nível de regularização (λ).

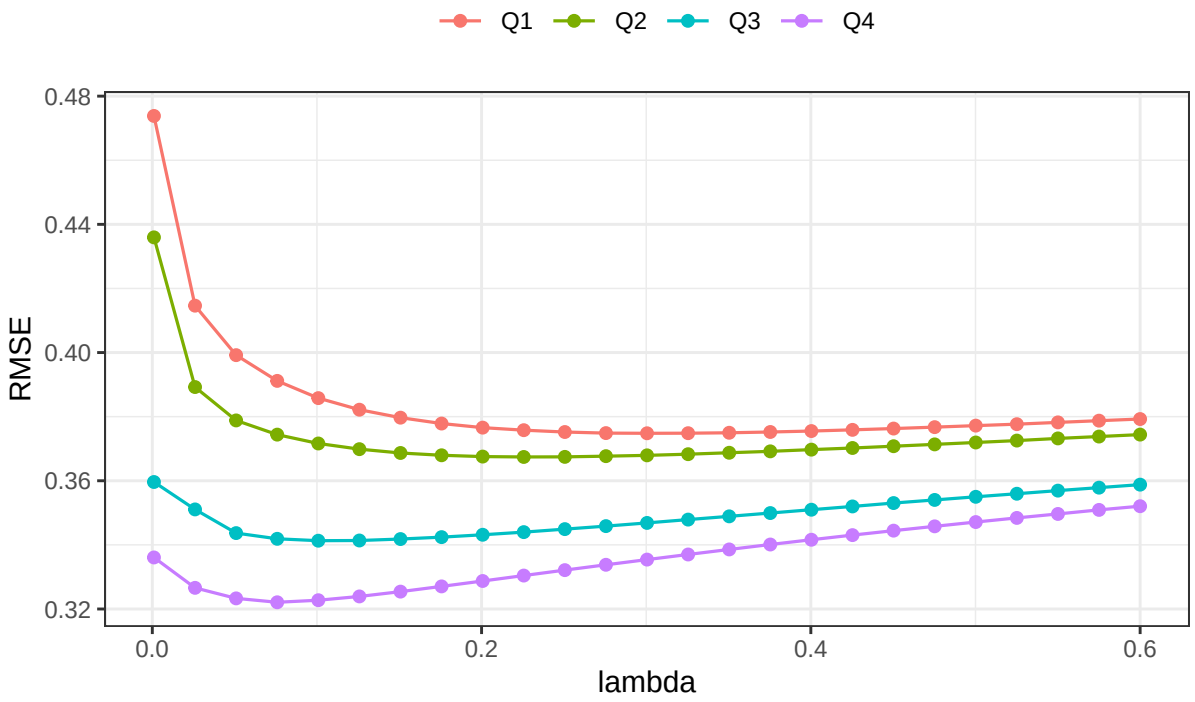
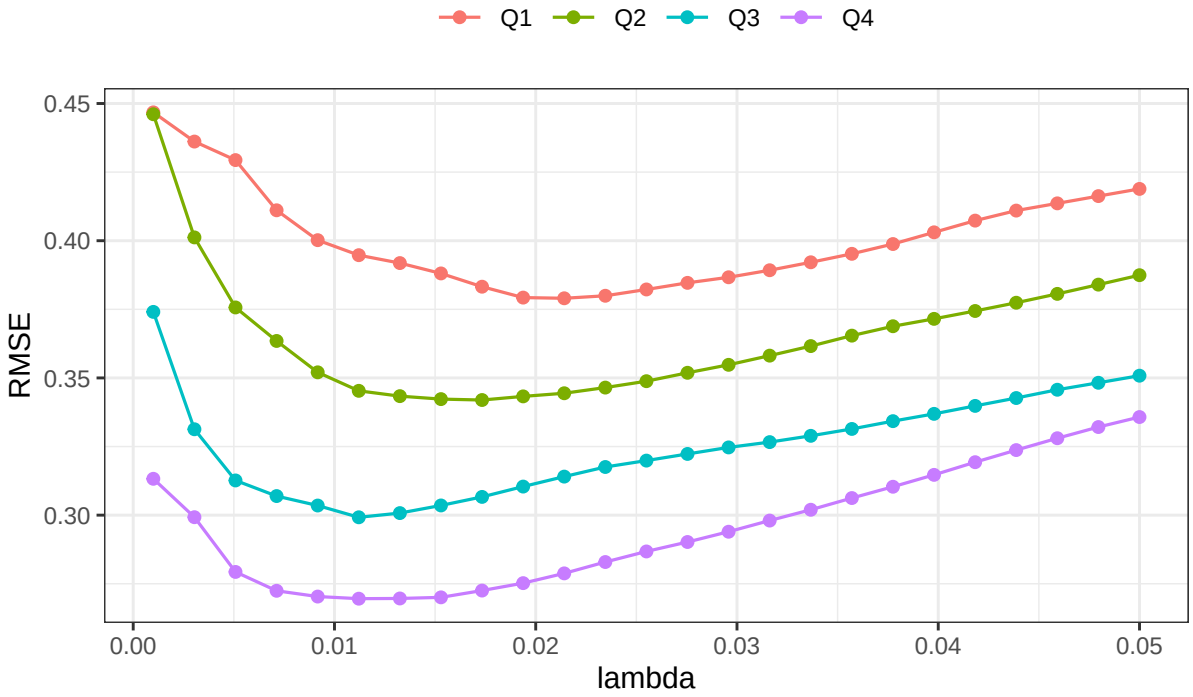


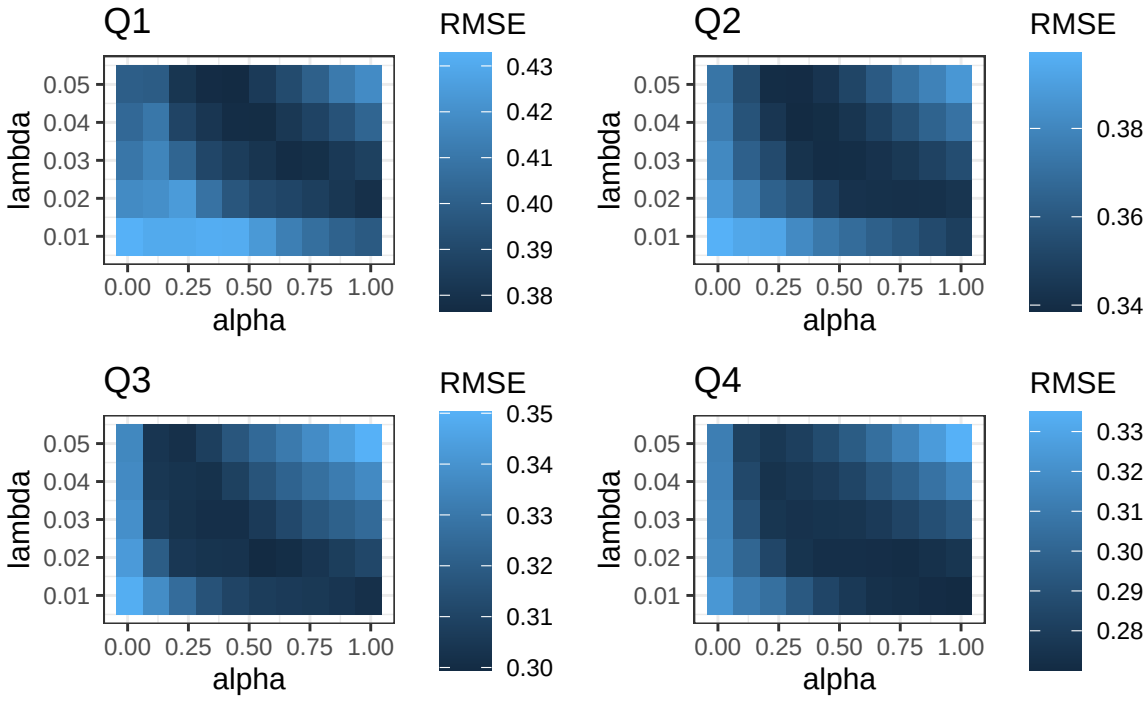
Figura 13: Curvas de RMSE fora da amostra do modelo LASSO segundo a regularização (λ).



4.2 Elastic Net

Os quadros da Figura 14 mostram o erro de previsão fora da amostra para o modelo elastic net ajustado a diferentes combinações dos parâmetros de mistura (α) e de regularização (λ). Nota-se que para os dados coletados nas últimas quadrissemanas do mês valores de α próximos a 1, que fazem o elastic net ser mais semelhante ao LASSO puro, trazem os melhores resultados, além de a penalização λ ser mais branda. Para as primeiras quadrissemanas α é mais próximo de 0.5 e a penalização λ é mais severa, muito embora ajustes com λ próximos de 1 não se distanciem tanto. Esses resultados acerca da maior penalização em quadrissemanas mais antigas reafirmam a ideia de que alguns subitens percam importância para previsões de mais longo prazo.

Figura 14: RMSE fora da amostra do modelo elastic net ajustado a diferentes combinações de parâmetros alpha (α) e lambda (λ).



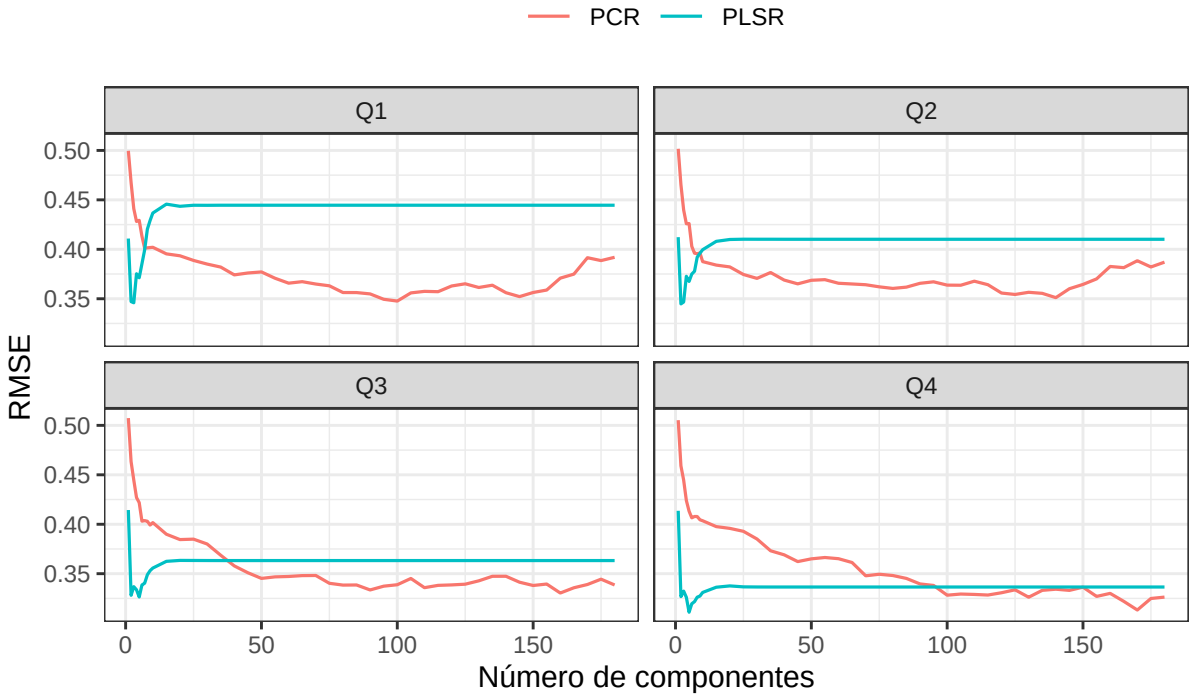
4.3 Principal Component Regression e Partial Least Squares

O gráfico da Figura 15 mostra como os erros RMSE de validação fora da amostra mudam segundo o número de componentes para os modelos regressão por componentes principais (PCR - *Principal Component Regression*) e para a regressão por mínimos quadrados parciais (PLSR - *Partial Least Squares Regression*). Nota-se, antes de tudo, que o PCR demonstra ganhos preditivos mesmo para números muito grandes (mais de 100) componentes principais, e a partir de um ponto o RMSE volta a aumentar, enquanto o erro do PLSR se estabiliza. No entanto, o PLSR se demonstra muito mais eficiente que o PCR uma vez que aproveita melhor a informação dos preditores utilizando um número muito mais limitado de componentes principais, e mesmo para um número alto de componentes o PCR jamais ultrapassa o PLSR. Além disso, o número ótimo de componentes vai diminuindo da quarta para a primeira quadrissemana, indicando que as previsões de mais longo prazo são afetadas por menos fatores.

Esses fatos mostram como pode ser significativo o ganho ao extrair componentes de forma supervisionada pelo PLSR, levando em conta também a variação conjunta de $\mathbf{x}$ com y, em vez da variação de $\mathbf{x}$ apenas, como feito pelo PCR. Isso é verdade pois nem toda a variabilidade de $\mathbf{x}$ se converte em informação sobre y. E de fato, para a quarta

quadrissemana, embora as cinco componentes do PLSR cubram apenas 12% da variação em $\mathbf{x}$, elas são suficientes para explicar 95% da variação em y. Para o PCR, apesar de 95% da variação em $\mathbf{x}$ ser explicada, apenas 92% de y é capturada por 170 componentes principais.

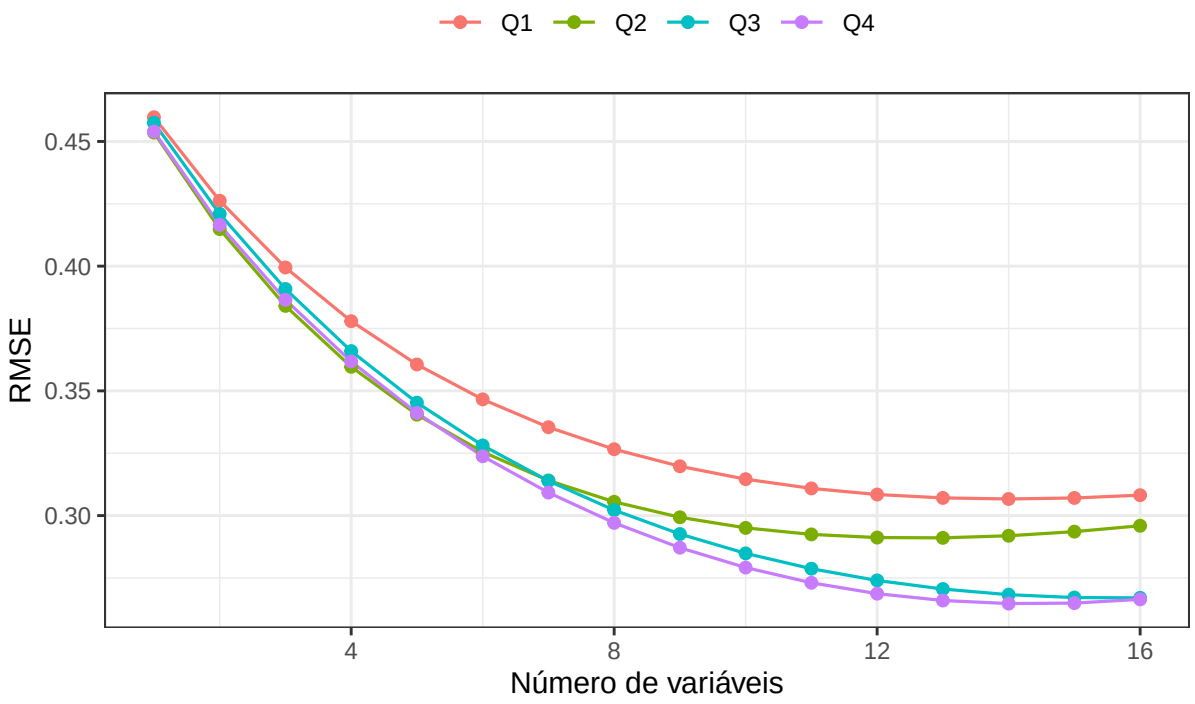
Figura 15: Curvas de RMSE fora da amostra para os modelos PCR e PLSR em função do número de componentes principais retidas.



4.4 Complete Subset Regressions

A Figura 16 apresenta o RMSE fora da amostra para as quadrissemanas considerando diferentes números de variáveis em cada modelo do CSR com pré-filtragem de 16 subitens preditores. Nota-se uma curva com formato de parábola e muito regular, possivelmente por conta de envolver uma ponderação de muitos modelos. Se observa um ganho consistente de poder preditivo conforme aumenta-se o número de variáveis até certo ponto. E, de maneira interessante, os resultados para as duas últimas quadrissemanas são muito próximos.

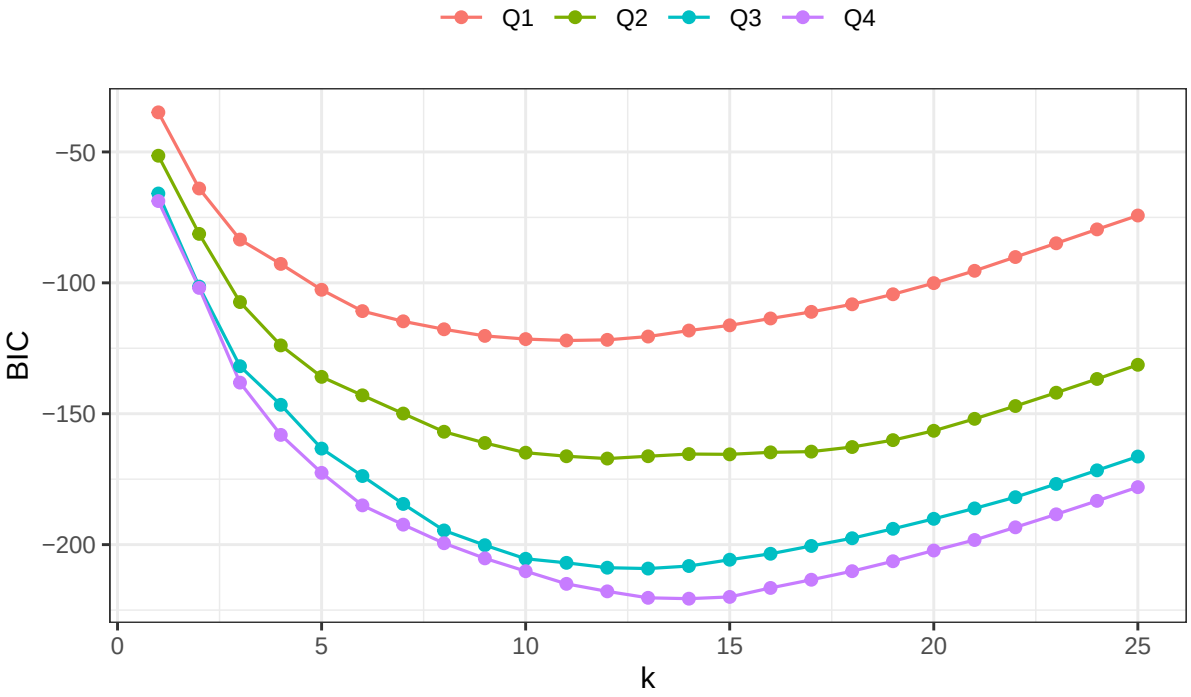
Figura 16: Curvas de RMSE fora da amostra para o CSR em função do número de variáveis em cada submodelo do agregado.



4.5 Best Subset Selection

Da mesma forma que o complete subset regressions requer o ajuste de diversos modelos para combiná-los numa média, o best subset selection necessita desses vários ajustes para escolher um melhor modelo entre eles. No entanto, como não é viável validar cada um dos milhares de subconjuntos de preditores, a escolha do melhor subconjunto foi feita considerando o Critério de Informação Bayesiano (BIC) no ajuste da amostra completa. A Figura 17 apresenta o melhor BIC entre cada conjunto de modelos com k variáveis. Aqui tem-se um comportamento semelhante a uma parábola como no CSR. Note que o número ótimo de variáveis para todas as quadrissemanas entre 10 e 14, com as quadrissemanas 1 e 2 necessitando de menos variáveis que as quadrissemanas 3 e 4.

Figura 17: Menor BIC entre modelos com k variáveis no best subset selection.



Ajustes na amostra completa.

4.6 Variáveis mais importantes

Nesta seção serão apresentadas as variáveis mais "importantes"em cada modelo, cada qual com seu critério de importância. Como base de comparação, a Tabela 3 apresenta as variáveis mais correlacionadas com o IPCA. Apesar de não haver nenhum subitem particularmente muito correlacionado com o IPCA, observa-se uma grande consistência de subitens entre as quadrissemanas, em especial a Gasolina e o Repolho, que são os mais importantes para todas as semanas. De fato, a Gasolina como componente associada ao custo de transporte da maioria dos produtos da cesta naturalmente tem seu preço repercutido de forma imediata nos preços dos produtos e consequentemente na inflação geral. Com exceção dos combustíveis e de alguns componentes da Habitação, todos os demais itens mais importantes são parte de Alimentação, o que faz todo sentido, já que tanto o IPCA quanto o IPC-FIPE buscam mensurar a inflação na cesta de produtos das famílias, onde a alimentação costuma ter maior peso.

Os subitens mais importantes no LASSO e Elastic Net, nesse caso consideradas as variáveis a mais perdurarem no modelo conforme se aumenta a penalização/regularização, mostrados nas Tabelas 4 e 5, foram muito similares entre si e com os subitens mais relacio-

nadas ao IPCA, também abarcando itens dos Transportes, Alimentação e Habitação. Por outro lado, de maneira interessante, para o modelo PCR os subitens de maior destaque, isto é, aqueles com maior peso na primeira componente principal, foram aqueles do grupo de Educação, embora também figuraram itens da Alimentação, como mostra a Tabela 6. Já o PLSR, representado na Tabela 7, foi mais similar ao LASSO e ao Elastic Net quanto à importância da Alimentação, mas incluiu assim como o PCR subitens da Educação para a previsão até a terceira quadrissemana. Por fim, para o BSS, modelo com melhor desempenho, houve uma maior variabilidade de grupos de itens, contendo inclusive itens de Despesas Pessoais.

Tabela 3: Correlação linear de Pearson entre o IPCA e os dez subitens do IPC-FIPE mais relacionados ao índice em cada quadrissemana.

Q1	**Q2**	**Q3**	**Q4**
Gasolina (0.43)	Gasolina (0.49)	Gasolina (0.54)	Gasolina (0.54)
Repolho (0.41)	Repolho (0.42)	Repolho (0.44)	Repolho (0.42)
Cenoura (0.36)	Couve (0.38)	Etanol (0.39)	Etanol (0.39)
Alface (0.36)	Etanol (0.36)	Couve (0.38)	Refeição (0.37)
Coentro (0.35)	Alface (0.35)	Peças de iluminação (0.38)	Couve (0.36)
Couve (0.35)	Cenoura (0.35)	Pão Alface (0.34)	Peças de iluminação (0.34)
Etanol (0.32)	Peças de iluminação (0.34)	Cenoura (0.33)	Brócolis (0.34)
Salsa/cebolinha (0.32)	Brócolis (0.33)	Refeição (0.33)	Linguiça (0.33)
Brócolis (0.32)	Coentro (0.32)	Salsa/cebolinha (0.32)	Alimentos embalados (0.32)
Escarola (0.32)	Salsa/cebolinha (0.32)	Batata (0.32)	Alface (0.32)

Grupos de itens: Alimentação, Transportes e Habitação.

Tabela 4: Dez principais subitens do IPC-FIPE segundo o modelo LASSO.

Q1	**Q2**	**Q3**	**Q4**
Gasolina	Gasolina	Gasolina	Gasolina
Repolho	Repolho	Repolho	Repolho
Cenoura	Peças de iluminação	Peças de iluminação	Refeição
Alcatra	Refeição	Refeição	Peças de iluminação
Toalha de rosto	Coxão mole	Linguiça	Linguiça
Outros pães	Cenoura	Batata	Alimentos embalados
Contrafilé	Toalha de rosto	Outros pães	Alimentos congelados
Laranja	Couve	Contrafilé	Pão francês
Peças de iluminação	Presunto	Pão francês	Batata
Refeição	Contrafilé	Coxão mole	Outros pães

A primeira variável, mais importante, é a última a sair do LASSO ao se aumentar a penalização/regularização, e assim por diante. Grupos de itens: Alimentação, Transportes e Habitação e Artigos de residência.

Tabela 5: Dez principais subitens do IPC-FIPE segundo o modelo Elastic Net.

Q1	**Q2**	**Q3**	**Q4**
Gasolina	Gasolina	Gasolina	Gasolina
Repolho	Repolho	Repolho	Repolho
Cenoura	Couve	Peças de iluminação	Refeição
Alface	Peças de iluminação	Refeição	Peças de iluminação
Coentro	Refeição	Linguiça	Linguiça
Toalha de rosto	Coxão mole	Couve	Alimentos embalados
Outros pães	Cenoura	Batata	Alimentos congelados
Alcatra	Contrafilé	Outros pães	Batata
Contrafilé	Toalha de rosto	Contrafilé	Pão francês
Laranja	Alcatra	Pão francês	Outros pães

A primeira variável, mais importante, é a última a sair do LASSO ao se aumentar a penalização/regularização, e assim por diante. Grupos de itens: Alimentação, Transportes e Habitação e Artigos de residência.

Tabela 6: Subitens do IPC-FIPE com maior peso (*loading*) na primeira componente principal no PCR.

Q1	**Q2**	**Q3**	**Q4**
Curso de idiomas	Curso de idiomas	Curso de idiomas	Curso de idiomas
Curso pré-vestibular	Ensino fundamental	Berçário/maternal	Berçário/maternal
Berçário/maternal	Ensino médio	Curso pré-vestibular	Curso pré-vestibular
Ensino médio	Berçário/maternal	Ensino fundamental	Ensino médio
Ensino fundamental	Educação infantil	Ensino médio	Ensino fundamental
Educação infantil	Curso pré-vestibular	Escarola	Educação infantil
Ensino superior	Ensino superior	Agrião	Ensino superior
Beterraba	Couve	Alface	Escarola
Alface	Agrião	Educação infantil	Alface
Curso de informática	Alface	Couve	Agrião

Grupos de itens: Alimentação e Educação.

Tabela 7: Subitens do IPC-FIPE com maior peso (*loading*) na primeira componente principal no PLS.

Q1	**Q2**	**Q3**	**Q4**
Couve	Couve	Ensino fundamental	Sabonete
Alface	Alface	Ensino médio	Couve
Escarola	Escarola	Couve	Refeição
Salsa/cebolinha	Agrião	Alface	Alface
Agrião	Coentro	Ensino superior	Brócolis
Coentro	Salsa/cebolinha	Educação infantil	Pão de forma
Brócolis	Repolho	Curso pré-vestibular	Linguiça
Repolho	Ensino superior	Salsa/cebolinha	Repolho
Ensino médio	Educação infantil	Refeição	Salsa/cebolinha
Curso pré-vestibular	Ensino fundamental	Repolho	Macarrão

Grupos de itens: Alimentação e Educação.

Tabela 8: Subitens do IPC-FIPE selecionados pelo BSS.

Q1	**Q2**	**Q3**	**Q4**
Gasolina	Gasolina	Gasolina	Gasolina
Repolho	Repolho	Repolho	Repolho
Outros pães	Refeição	Peças de iluminação	Refeição
Automóvel usado	Coxão mole	Refeição	Peças de iluminação
Maçã	Cama	Batata	Pão francês
Coxão mole	Batata	Pão francês	Alimentos congelados
Refrigerante	Pão francês	Coxão mole	Batata
Cama	Transporte escolar	Talher	Ovos
Malha/agasalho feminino	Açúcar	Transporte escolar	Transporte escolar
Acessórios de veículo	Perfume/colônia	Amaciante p/ roupa	Sofá
Transporte escolar	Toalha de banho	Leite aromatizado	Café em pó
	Acessórios de veículo	Presunto	Coxão duro
		Toalha de banho	Xampu

Grupos de itens: Alimentação, Transportes, Habitação e Artigos de residência e Despesas Pessoais.

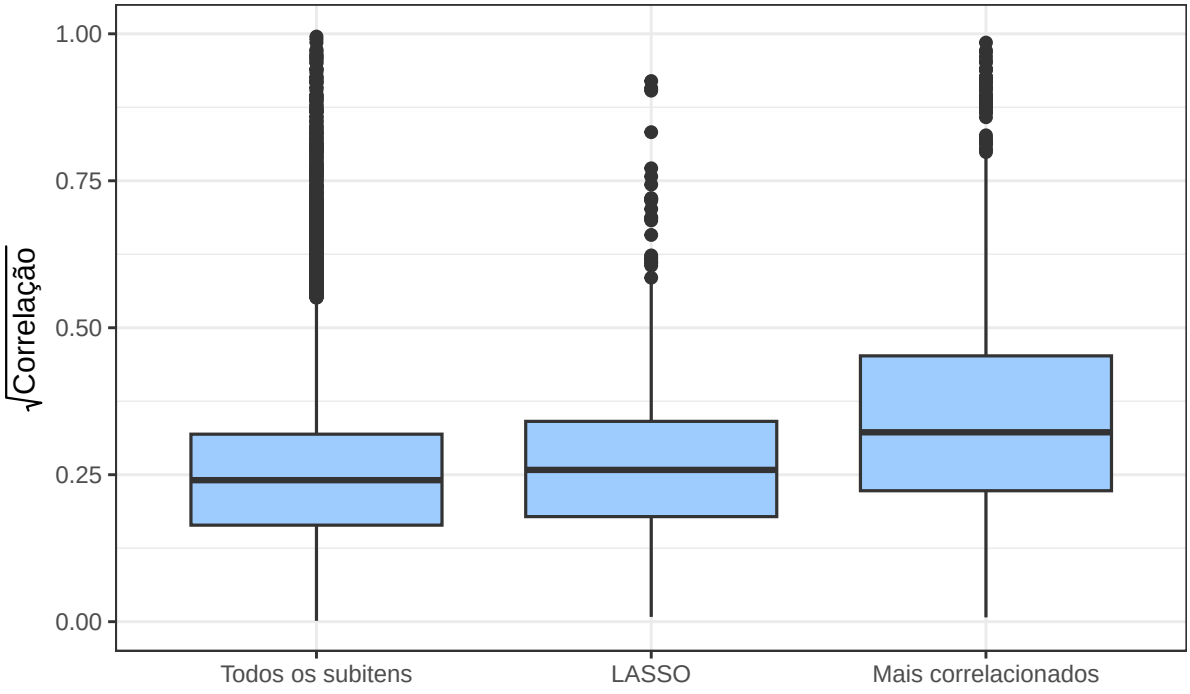
5 Discussão

5.1 Seleção de preditores por correlação

Como já comentado, um procedimento muito comum para escolher variáveis para compor um modelo de regressão é selecionar as variáveis mais correlacionadas com a

variável resposta que se deseja prever. Contudo, esse procedimento tende a produzir modelos ineficientes e que sofrem de multicolinearidade, uma vez que essas variáveis mais correlacionadas com a resposta tendem também a ser mais correlacionadas entre si. Esse ponto é ilustrado para o caso presente através da Figura 18. É observado que os subitens do IPC-FIPE mais correlacionadas com o IPCA são em geral mutuamente mais correlacionadas do que no caso das variáveis selecionadas pelo LASSO. Por sua vez, o LASSO seleciona subitens cuja distribuição de correlação entre pares é similar à observada entre todos os pares de subitens no geral, mostrando que o LASSO é muito mais eficiente na seleção de preditores do que o critério de correlação por si só.

Figura 18: Distribuição da raiz da correlação linear de Pearson entre os pares de: subitens, quarenta principais subitens do IPC-FIPE no LASSO e quarenta subitens mais correlacionados com o IPCA.



5.2 Continuação

Algumas limitações deste trabalho abrem margem para continuação e expansão do projeto nas seguintes direções:

- Estender o conjunto de preditores com outras fontes de dados, inclusive de maior frequência que a semanal, como: oferta monetária, desemprego, taxas de câmbio, dados financeiros, bancários e fiscais, importações e exportações, expectativa de inflação, tal como fizeram Garcia, Medeiros e Vasconcelos (2017), Araujo e Gaglianone

(2023) e Garnitskiy (2020) em seus trabalhos;

- Considerar outros *benchmarks* tradicionais como o modelo de passeio aleatório, ARMA e as projeções do Focus;
- Considerar a combinação de modelos através outras técnicas de comitês como o Model Confidence Set (SAMUELS; SEKKEL, 2017) e ensembles como o *bagging*, *boosting* e *stacking*;
- Estudar o ganho da previsão do IPCA desagregada por grupo de itens que compõem a cesta de produtos;
- Realizar o ajuste sazonal das variáveis, de maneira que os modelos de regressão sejam mais apropriados para o contexto de previsão.

6 Conclusão

Este trabalho apresentou resultados de previsão semanal do Índice de Preços ao Consumidor Amplo (IPCA) usando indicadores do Índice de Preços ao Consumidor do Município de São Paulo (IPC-FIPE) quadrissemanal. Para tanto, foram empregadas diversas técnicas para dados de alta dimensão, ou seja, com número de preditores comparável ou superior ao número de observações. Cada modelo considerado teve seus hiper-parâmetros ajustados em um esquema de origem deslizante para estimação da Raiz do Erro Quadrático Médio (RMSE) fora da amostra, tendo sido a primeira previsão no mês de novembro de 2020 e a última em outubro de 2023. Os modelos selecionados foram então avaliados seguindo o mesmo esquema.

Em resumo, os resultados encontrados mostraram que: (1) mesmo mais flexível, o Elastic Net pode trazer resultados semelhantes ao LASSO quando a dimensão dos dados é alta; (2) sendo supervisionado pela variável resposta, o PLSR é capaz de gerar componentes principais mais úteis e com maior concentração de informação do que o comumente utilizado PCR; (3) os modelos que agregam informação de todas as centenas de preditores — Regressão Ridge, LASSO, Elastic Net, PCR e PLSR — foram piores em todas as quadrissemanas em relação aos modelos que filtram preditores; (4) a filtragem de preditores irrelevantes feita pelo LASSO ou Elastic Net ainda pode ser insuficiente, e possivelmente por esse motivo a redução adicional com o BSS numa segunda etapa melhorou os resultados; (5) o Best Subset Selection trouxe os melhores resultados nas quatro quadrissemanas, seguido pelo Complete Subset Regressions; (6) todos os modelos superaram o modelo *benchmark* autorregressivo; (7) de modo geral, os modelos para previsões de maior prazo tendem a depender de menos variáveis ou componentes, e assim geram previsões menos voláteis; (8) o critério de selecionar preditores com base em correlação dificilmente é uma boa estratégia, pois induz à redundância e se torna ineficiente; (9) a validação fora da amostra é uma importante aliada na avaliação, seleção e comparação de modelos, posto que critérios *in-sample* são limitados.

Referências

ARAUJO, G. S.; GAGLIANONE, W. P. Machine learning methods for inflation forecasting in brazil: New contenders versus classical models. *Latin American Journal of Central Banking*, Elsevier, v. 4, n. 2, p. 100087, 2023.

BAŃBURA, M. et al. Now-casting and the real-time data flow. In: *Handbook of economic forecasting*. [S.l.]: Elsevier, 2013. v. 2, p. 195–237.

BANCO CENTRAL DO BRASIL. *O que é inflação?* 2024. ⟨https://www.bcb.gov.br/controleinflacao/oqueinflacao⟩. Acesso em: 11 jun. 2024.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006. (Information Science and Statistics). ISBN 978-0-387-31073-2.

ELLIOTT, G.; GARGANO, A.; TIMMERMANN, A. Complete subset regressions. *Journal of Econometrics*, v. 177, n. 2, p. 357–373, dez. 2013. ISSN 03044076.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, v. 33, n. 1, 2010. ISSN 1548-7660.

GARCIA, M. G.; MEDEIROS, M. C.; VASCONCELOS, G. F. Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting*, v. 33, n. 3, p. 679–693, jul. 2017. ISSN 01692070.

GARNITSKIY, L. *Nowcasting brazilian inflation with machine learning*. Dissertação (Mestrado) — Fundação Getulio Vargas, Escola Brasileira de Economia e Finanças, Rio de Janeiro, 2020.

GARTHWAITE, P. H. An interpretation of partial least squares. *Journal of the American Statistical Association*, Taylor & Francis, v. 89, n. 425, p. 122–127, 1994.

GREENACRE, M. et al. Principal component analysis. *Nature Reviews Methods Primers*, Nature Publishing Group UK London, v. 2, n. 1, p. 100, 2022.

MASSY, W. F. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, Taylor & Francis, v. 60, n. 309, p. 234–256, 1965.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023. Disponível em: ⟨https://www.R-project.org/⟩.

SAMUELS, J. D.; SEKKEL, R. M. Model Confidence Sets and forecast combination. *International Journal of Forecasting*, v. 33, n. 1, p. 48–60, jan. 2017. ISSN 01692070.

TASHMAN, L. J. Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, Elsevier, v. 16, n. 4, p. 437–450, 2000.

TIBSHIRANI, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 58, n. 1, p. 267–288, jan. 1996. ISSN 0035-9246, 2517-6161.

TIMMERMANN, A. Forecast combinations. *Handbook of economic forecasting*, Elsevier, v. 1, p. 135–196, 2006.

ZOU, H.; HASTIE, T. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, v. 67, n. 2, p. 301–320, abr. 2005. ISSN 1369-7412, 1467-9868.

Apêndice

A Código

Todos os cálculos e análises apresentadas neste trabalho foram realizados por meio da linguagem `R` 4.3.2 (R Core Team, 2023). Em particular, foram utilizados os pacotes `pls` para os modelos Principal Component Regression e Partial Least Squares Regression, `glmnet` para a Regressão Ridge, LASSO e Elastic Net, `leaps` para o Best Subset Selection, e `ggplot2` para a confecção das representações gráficas. Os demais algoritmos de previsão explorados neste trabalho, bem como os procedimentos de validação fora da amostra — que podem, inclusive, ser facilmente adaptados para outros modelos além destes — foram implementados pelo próprio autor. Detalhes adicionais sobre a implementação do projeto podem ser obtidos com o autor através do endereço eletrônico pedrohenrique6180@gmail.com.