Universidade de Brasília
Departamento de Estatística

Selecionador de Características para Previsão de Indicadores Econômicos.

Rafael Costa Ramos

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

Rafael Costa Ramos

Selecionador de Características para Previsão de Indicadores Econômicos.

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

Dedico este trabalho aos meus pais, Renato e Renata, e à minha irmã, Marina.

Agradecimentos

Primeiramente à minha família, que me apoiou em cada passo da minha trajetória. Em especial aos meus pais, pela criação e por sempre terem me dado as melhores condições.

Ao meu orientador, George Freitas von Borries, pela paciência, ensinamentos e brilhante didatismo na orientação e nas disciplinas ministradas.

Aos professores da Universidade de Brasília (UnB), em especial os do Departamento de Estatística (EST), pelas contribuições com a formação intelectual minha e de cada aluno da universidade.

Aos meus amigos de dentro e fora da universidade, pelo companheirismo e apoio ao longo dos anos.

Resumo

Este trabalho propõe um algoritmo para redução de dimensionalidade e seleção de variáveis explicativas para modelos preditivos. O algoritmo consiste no agrupamento de variáveis preditoras e na seleção aleatória de preditores de cada grupo. A técnica utilizada para agrupar as variáveis foi o modelo de mistura Gaussiana (*Gaussian mixture model* - GMM). Com as variáveis selecionadas de cada grupo, o modelo preditivo aplicado é o lasso. Para avaliar estabilidade preditiva do algoritmo, a distribuição dos erros de previsão obtidos por reamostragem foi utilizada.

Os dados utilizados são de natureza macroeconômica. A primeira base dados contém dados de atividade econômica e variáveis de produção. Já a segunda base envolve dados de índices de inflação. Ambos os conjuntos de dados são superdimensionados, ou seja, com mais variáveis do que observações ($p > n$). A modelagem preditiva simula um processo de *nowcasting*, com previsões de curto prazo e atualização dos dados à medida em que novas informações ficam disponíveis.

Com os conjuntos de dados analisados, o algoritmo não obteve previsões estáveis. A instabilidade preditiva mostra que os grupos de variáveis explicativas não são homogêneos. Portanto, o agrupamento das variáveis por GMM não teve sucesso para reduzir a dimensão dos dados utilizados no trabalho.

Palavras-chaves: Modelo de Mistura de Normais; Lasso; Reamostragem; RMSE; *Nowcasting*.

Abstract

This paper proposes an algorithm for dimensionality reduction and feature selection for predictive models. The algorithm is based on clustering and random selection of variables of each cluster. The technique applied to cluster the attributes is the Gaussian mixture model (GMM), a technique that relies on mixture of normal distributions. With the variables selected, the predictive model implemented is the lasso. To evaluate the stability of the predictions, the distribution of prediction errors obtained by resampling is used.

The data utilized is from macroeconomic nature. The first dataset gathers data of economic activity and production variables. The second dataset contains data of inflation rates. Both datasets are high-dimensional, with more variables than observations ($p > n$). The predictive modelling simulates a nowcasting process, with short-term forecasts and data updates as new information becomes available.

With the datasets analysed, the algorithm did not obtain stable predictions. The predictive instability shows that the groups of features are not homogeneous. Therefore, the clustering of explanatory variables with GMM did not succeed in the task of dimensionality reduction with the data available.

Keywords: Gaussian Mixture Model; Lasso; Resampling; RMSE; Nowcasting.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

As variáveis macroeconômicas são medidas que servem como bússola para os rumos da economia de um país. Indicadores econômicos mostram a situação de questões muito importantes, como a atividade econômica e o poder de compra da população. Entre os principais índices macroeconômicos, é possível citar o produto interno bruto (PIB), inflação, juros, câmbio e desemprego.

Os índices macroeconômicos balizam as decisões de consumo e investimento dos governos, empresas e famílias. Além de decisões relacionadas com consumo, poupança e investimento, o próprio orçamento público é altamente indexado por variáveis como PIB e inflação.

Em razão da grande importância que os resultados de alguns agregados macroeconômicos possuem no conjunto da economia, a previsão desses índices se torna muito valiosa. Governos, instituições financeiras e organizações internacionais, como o Fundo Monetário Internacional (FMI), fazem previsões de índices macroeconômicos regularmente. Essas previsões são revisadas ao longo do ano, de acordo com os rumos da economia do país.

Um desafio que surge decorre do fato de que muitos índices são divulgados de maneira trimestral, como PIB e desemprego, mas existe um interesse de prever cenários mensais, ou de prazos ainda menores. É com a ideia de projeções de curto prazo que economistas do Banco Central Europeu e do *Federal Reserve* dos Estados Unidos passam a desenvolver trabalhos de *nowcasting* (GIANNONE; REICHLIN; SMALL, 2008). O termo se refere a previsões econômicas do futuro muito próximo, do presente, ou até mesmo de um passado recente.

Naturalmente, os primeiros trabalhos sobre *nowcasting* foram calcados nas técnicas da econometria. Nos últimos anos, porém, o ferramental de *machine learning* passou a ser utilizado para criar modelos de *nowcasting* de indicadores macroeconômicos. Inclusive com trabalhos de técnicos do Banco Central do Brasil (ARAUJO; GAGLIANONE, 2022).

Existe uma grande quantidade de variáveis econômicas de fácil coleta, como diversos indicadores de atividade econômica e índices de preço de uma enorme variedade de produtos. Diversos agentes, especialmente governos e bancos centrais coletam e obtém de fontes secundárias um grande volume de dados econômicos regularmente. Com uma quantidade cada vez maior de variáveis coletadas, os bancos de dados se tornam superdimensionados. Em outras palavras, os conjuntos de dados passam a apresentar mais

variáveis do que observações ($p > n$).

Ao lidar com dados superdimensionados, surgem problemas que não ocorrem em conjuntos de dados de baixa dimensão. O fenômeno é conhecido como *curse of dimensionality*, nome cunhado em Bellman, Corporation e Collection (1957). De modo geral, existem ideias e estruturas básicas de modelagem e agrupamento que funcionam em espaços de baixa dimensão, mas são distorcidas e se tornam incorretas em espaços de alta dimensão (BOUVEYRON et al., 2019). Em casos onde p é muito elevado, não é nem possível ajustar um modelo estatístico simples, como o de regressão linear (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Um desafio de lidar com dados de alta dimensão é o sobreajuste, ou *overfitting*. O sobreajuste ocorre quando o modelo é ajustado com muitas variáveis explicativas e captura o ruído presente nos dados (JAMES et al., 2021). Um modelo com *overfitting* apresentará dificuldade para generalizar seus resultados.

Para construir um modelo preditivo com dados superdimensionados, alguma seleção de variáveis explicativas se torna essencial. Uma possível alternativa seria a seleção de um número fixo de *features* mais correlacionadas com a variável resposta. Entretanto, um problema que provavelmente ocorrerá com essa abordagem é que as variáveis explicativas selecionadas devem ser altamente correlacionadas entre si. O processo descrito é conhecido como multicolinearidade (KUTNER; NACHTSHEIM; NETER, 2003).

A hipótese levantada no trabalho é que a aplicação de uma técnica de agrupamento de variáveis explicativas soluciona os problemas que ocorrem com dados de alta dimensão. Com o agrupamento, seria possível reproduzir a informação de muitas *features* em apenas alguns grupos, reduzindo a dimensão dos dados. Com *clusters* homogêneos internamente e heterogêneos externamente, a questão da multicolinearidade seria suavizada, visto que preditores selecionados de grupos diferentes devem possuir baixa correlação entre si. Com a redução de dimensão que ocorre pela técnica de seleção proposta no trabalho, os modelos seriam ajustados com menos variáveis, o que diminui a propensão ao *overfitting*.

Este trabalho busca responder se a seleção de variáveis baseadas em uma técnica de agrupamento gera previsões estáveis. O objetivo a ser alcançado pelo algoritmo proposto é reduzir a dimensão dos dados por meio de um bom processo de seleção de *features* baseado no agrupamento de variáveis. As previsões serão *nowcasts* de índices macroeconômicos com dados de alta dimensão. O resultado será analisado pela variabilidade dos erros de previsão.

2 Referencial Teórico

2.1 *Nowcasting*

O termo *nowcasting* é a aglutinação dos palavras inglesas *now* e *forecasting*, "previsão do agora"em tradução livre. O conceito começou a ser utilizado na área da meteorologia para indicar técnicas de previsão do tempo em até 6 horas (BROWNING, 1980). Os *nowcasts* meteorológicos utilizam múltiplas fontes de dados, como radares e satélites. Apesar do surgimento na meteorologia, o *nowcasting* hoje é utilizado em outras áreas.

2.1.1 *Nowcasting* em Economia

No âmbito econômico, surge a preocupação de fazer previsões de curto prazo de indicadores importantes. Em Kitchen e Monaco (2003) e Evans (2005), por exemplo, os autores desenvolvem modelos para prever o PIB dos Estados Unidos em tempo real. Os trabalhos ressaltam a importância de dados diários para a estimação do PIB, publicado apenas trimestralmente.

Em Giannone, Reichlin e Small (2008), o *nowcasting* chega ao contexto da economia, ao lidar com o desafio fazer previsões para o PIB dos Estados Unidos no trimestre corrente. Economistas do Banco Central Europeu e do *Federal Reserve*, os autores utilizaram 200 séries da economia americana publicadas em 35 lançamentos mensais. O trabalho conclui que a pontualidade com a qual novas informações ficam disponíveis são determinantes para o poder preditivo do modelo. Por essa razão, a incorporação de dados divulgados mais rapidamente, como os *surveys*, melhoram a qualidade do *nowcast* (GIANNONE; REICHLIN; SIMONELLI, 2009).

Em Giannone, Reichlin e Bańbura (2010), *nowcasting* passou a ser definido como a previsão do presente, do futuro muito próximo, ou até do passado muito próximo. O trabalho apresenta a ideia de que os *nowcasts* são obtidos por 2 componentes: a previsão anterior e as novas informações disponíveis, denominadas "*news*". Dessa forma, a magnitude da revisão preditiva depende também da pontualidade com que se obtém novas informações, e não apenas da relevância da variável preditora.

O trabalho de Giannone et al. (2013) aborda o papel das diferentes fontes de informações e sua continua incorporação nos modelos para *nowcasting*. Segundo os au-

tores, o uso de informações pontuais, como *surveys*, pode dar estimativas adiantadas da atividade econômica, enquanto dados macroeconômicos mais "sólidos"ainda não estão disponíveis. O *paper* também conclui que quanto mais próximo do fim do trimestre, melhores são as estimativas do modelo, graças ao acúmulo de dados ao longo do tempo.

2.1.2 Novas Técnicas de *Nowcasting*

Como é possível notar, os trabalhos iniciais costumam utilizar indicadores dos Estados Unidos para fazer *nowcasting*. No entanto, com o avanço da literatura sobre o tema, autores trabalharam com dados do PIB de outros países, como Indonésia (LUCIANI et al., 2018) e Japão (HAYASHI; TACHI, 2023). Além do PIB, artigos foram desenvolvidos buscando fazer *nowcasting* de outras variáveis de interesse, tal qual o hiato do produto (BERGER; MORLEY; WONG, 2023), pobreza (MAHLER; AGUILAR; NEWHOUSE, 2021), segurança alimentar (WOBCKE et al., 2022) e até gentrificação (GLAESER; KIM; LUCA, 2018).

De início, as técnicas utilizadas para *nowcasting* eram *bridge equations* (BAFFIGI; GOLINELLI; PARIGI, 2004), modelos de fatores dinâmicos (GIANNONE et al., 2013) e modelos frequências mistas, como MIDAS e VAR (KUZIN; MARCELLINO; SCHUMACHER, 2011). Esses modelos cita dos utilizam o ferramental da análise de regressão e séries temporais, técnicas usuais da econometria. Com o avanço da literatura, a modelagem de *machine learning* passa a ser empregada em *nowcasting*, como o caso do lasso (BABII; GHYSELS; STRIAUKAS, 2022). No tema da previsão do PIB em tempo real, artigos que utilizam os métodos de *machine learning* foram aplicados aos dados de países como Finlândia (FORNARO; LUOMARANTA, 2020), China (ZHANG; NI; XU, 2023), Índia (GHOSH; RANJAN, 2023), Turquia (BARLAS et al., 2024), entre outros.

Com a aplicação de modelos de *machine learning* (ML) para *nowcasting* de variáveis econômicas, os pesquisadores da área começaram a comparar os novos modelos com os tradicionais. Em Araujo e Gaglianone (2022), técnicos do Banco Central do Brasil compararam 50 modelos, tradicionais da econometria e modelos de ML, com o objetivo de fazer *nowcasting* da inflação brasileira. Com um banco de dados extenso, os autores detectam as variáveis mais importantes na previsão da inflação, pontuando que a qualidade dos dados é mais importante que quantidade. Os técnicos do Banco Central chegaram à conclusão de que os modelos de *machine learning* performam um pouco melhor que os modelos econométricos tradicionais, mas ressalvam que não há um modelo universalmente melhor.

Atualmente, várias técnicas inovadoras são empregadas na área de *nowcasting* econômico. Alguns exemplos são análise de sentimento de notícias (LUKAUSKAS et al., 2022), sensoriamento remoto (BOLIVAR, 2024), análise de texto (ZHENG et al., 2024) e até redes neurais treinadas com dados do *Google Trends* (GRYBAUSKAS et al., 2023).

2.2 Métodos de Estimação

2.2.1 Estimação por Máxima Verossimilhança (EMV)

O método da máxima verossimilhança é a técnica de estimação mais popular na estatística (CASELLA; BERGER, 2002). Leve em conta que $X_1, \ldots, X_n$ são variáveis aleatórias independentes e identicamente distribuídas (iid) com função de densidade, ou de massa, $f(x|\theta_1, \ldots, \theta_k)$, na qual $\theta_1, \ldots, \theta_k$ são parâmetros desconhecidos. A função de verossimilhança é definida como:

$$\mathcal{L}(\theta|\mathbf{x}) = \mathcal{L}(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k). \tag{2.2.1}$$

O estimador de máxima verossimilhança (EMV) para o parâmetro, ou vetor de parâmetros, desconhecido θ, dentro do espaço de parâmetros Θ, será obtido por:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta|\mathbf{x}). \tag{2.2.2}$$

Para facilitar a maximização de $\mathcal{L}(\theta|\mathbf{x})$, normalmente aplica-se o logaritmo natural na função de verossimilhança: $\log \mathcal{L}(\theta|\mathbf{x}) = \ell(\theta|\mathbf{x})$. Se a função log-verossimilhança for derivável, o ponto de máximo de $\ell(\theta|\mathbf{x})$ será obtido por:

$$\frac{\partial}{\partial \theta_i} \ell(\theta|\mathbf{x}) = 0. \tag{2.2.3}$$

Em alguns casos, é possível achar o estimador de máxima verossimilhança em uma forma fechada. Entretanto, na maior parte dos casos, são utilizadas métodos numéricos e aproximações para encontrar o EMV (EFRON; HASTIE, 2016). Uma forma muito popular de fazer a estimação é por meio do algoritmo EM.

2.2.2 Algoritmo EM

O algoritmo EM é uma técnica de estimação de máxima verossimilhança que se tornou muito popular, superando o método de Newton-Raphson e outras técnicas de substituição (WATANABE; YAMAGUCHI, 2003). Desenvolvido por Dempster, Laird e Rubin (1977) a partir de dados faltantes, o algoritmo converge para o EMV por meio de passos iterativos.

Intuitivamente, o algoritmo é baseado na ideia de substituir uma maximização de verossimilhança difícil por uma sequência de maximizações mais fáceis, cujo limite é o que resolve o problema original (CASELLA; BERGER, 2002). Na utilização do EM, dois problemas de verossimilhança são considerados: o problema dos "dados incompletos"e o dos "dados completos".

Leve em conta n observações multivariadas $(x_i,\ z_i)$, em que x é uma variável observada e z é uma variável não observada, ou latente (BOUVEYRON et al., 2019). A verossimilhança completa dos dados é dada por:

$$\mathcal{L}_C(x, z|\theta) = \prod_{i=1}^{n} f(x_i, z_i|\theta). \tag{2.2.4}$$

A verossimilhança observada dos dados, por sua vez, é obtida ao integrar em z:

$$\mathcal{L}(x|\theta) = \int \mathcal{L}_C(x, z|\theta)\ dz. \tag{2.2.5}$$

Para facilitar na maximização da verossimilhança, é utilizado $\ell_C = \log \mathcal{L}_C$. Os passos do algoritmo EM são:

- *Expecation* (E): estimar a esperança condicional de ℓ_C, dada a informação disponível e as atuais estimativas dos parâmetros.
- *Maximization* (M): determinar os parâmetros que maximizam a esperança condicional de ℓ_C do passo E.

Os passos são repetidos iterativamente até atingir a convergência numérica.

2.3 Modelos de Mistura

2.3.1 Modelo de Mistura Finita

Modelos de mistura fornecem ferramentas poderosas para análise de *clusters*, classificação e estimação de densidade. Esses modelos possuem aplicações nas mais diversas áreas, como economia, medicina, astronomia, agronomia, *marketing*, bioinformática, entre outras (SCRUCCA et al., 2023). Ao longo do trabalho, o foco será especificamente no uso de modelos de mistura para análise de *clusters*, ramo do aprendizado não supervisionado.

Como definido em Scrucca et al. (2023), uma distribuição de mistura é uma distribuição de probabilidade obtida por meio de uma combinação linear convexa de funções densidade de probabilidade. As distribuições de mistura são formadas por componentes, que são as distribuições individuais, e os pesos, ou proporções da mistura. Dada uma distribuição de mistura finita $p(\mathbf{x})$, sua forma pode ser expressa por:

$$p(\mathbf{x}) = \sum_{i=1}^{G} \pi_i f_i(\mathbf{x}|\theta_i). \tag{2.3.1}$$

Na equação 2.3.1, $f_i(\mathbf{x};\theta_i)$ são as funções de distribuição das i componentes da mistura e π_i são os pesos de cada componente.

2.3.2 Modelo de Mistura de Normais (GMM)

No caso específico do modelo de mistura de normais, ou *Gaussian mixture model* (GMM), a distribuição de mistura é formada por componentes que seguem uma distribuição normal. Da mesma forma que é expressa pela forma geral em 2.3.1, o GMM é dado por:

$$p(\mathbf{x}) = \sum_{i=1}^{K} \pi_i \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i). \tag{2.3.2}$$

Na equação 2.3.2:

- $p(\mathbf{x})$ é a distribuição de mistura de normais;
- K é a quantidade finita de componentes Gaussianas no modelo;
- π_i é o peso da i-ésima componente do modelo;

- $\mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i)$ é a função de densidade da distribuição normal das i componentes.

Uma outra forma de pensar em π_i é como a probabilidade da observação pertencer à i-ésima componente. No modelo de mistura de normais, cada componente possui uma média própria, μ_i, e sua matriz de covariância Σ_i. Na maior parte dos casos, a normal da mistura será multivariada. No caso univariado, as densidades são $\mathcal{N}(x|\mu_i, \sigma_i^2)$, na qual a i-ésima componente da mistura possui distribuição normal com média μ_i e desvio padrão σ_i (BOUVEYRON et al., 2019).

2.3.3 Estimação de Máxima Verossimilhança por Algoritmo EM

A estimação do modelo de mistura de normais é feito por meio do algoritmo EM. Com a notação utilizada em Bouveyron et al. (2019), a verossimilhança observada dos dados, ou verossimilhança da mistura, pode ser expressa no caso multivariado por:

$$\mathcal{L}_O(x|\theta) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \phi_g(x_i|\mu_i, \Sigma_i). \tag{2.3.3}$$

Na equação 2.3.3:

- $\mathcal{L}_O(x|\theta)$ é a verossimilhança da mistura;
- n é o número de observações;
- G é a quantidade de componentes da mistura;
- π_g é a probabilidade da observação pertencer a componente g;
- $\phi_g(x_i|\mu_i, \Sigma_i)$ é a densidade normal multivariada de x_i, com respeito aos parâmetros μ_i e Σ_i.

Dessa forma, a log-verossimilhança observada dos dados será:

$$\ell_O(x|\theta) = \sum_{i=1}^{n} \log\Big\{\sum_{g=1}^{G} \pi_g \phi_g(x_i|\mu_i, \Sigma_i)\Big\}. \tag{2.3.4}$$

O algoritmo de *expectation-maximization* (EM) é uma técnica de estimação de máxima verossimilhança que trata os dados como observações multivariadas de x e z. No EM, x é a variável observada e z é a não observada, ou latente. A variável latente z é definida como:

$$z_{i,g} = \begin{cases} 1 & \text{se } x_i \text{ pertence à g-ésima componente da mistura} \\ 0 & \text{caso contrário.} \end{cases} \tag{2.3.5}$$

Assume-se que os z_i são iid, cada um de acordo com uma distribuição multinomial com probabilidades $\pi_1, \ldots, \pi_g$ de serem retirados de alguma das G categorias. Também é assumido que a densidade $x_i|z_i$ é obtida por $\prod_{g=1}^{G} f_g(x_i|\theta_g)^{z_{i,g}}$. Portanto, a log-verossimilhança com os dados completos é:

$$\ell_C(\theta_g, \pi_g, z_{i,g}|x) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{i,g} \log[\pi_g f_g(x_i|\theta_g)]. \tag{2.3.6}$$

O passo E do algoritmo é dado por:

$$\hat{z}_{i,g}^{(s)} = \frac{\hat{\pi}_g^{(s-1)} f_g(x_i|\hat{\theta}_g^{(s-1)})}{\sum_{h=1}^{G} \hat{\pi}_h^{(s-1)} f_h(x_i|\hat{\theta}_h^{(s-1)})}. \tag{2.3.7}$$

No qual $\hat{z}_{i,g}^{(s)} = E[z_{i,g}|x_i, \theta_1, \ldots, \theta_G]$ é a esperança condicional de $z_{i,g}$ dado os valores estimados para os parâmetros até a $(s-1)$-ésima iteração e os dados observados. O passo M consiste na maximização de 2.3.6 em termos de π_g e θ_g com os valores obtidos para $\hat{z}_{i,g}$, no passo E.

No caso multivariado, o algoritmo EM funciona de forma análoga. No passo E, a função f_g é substituída por uma densidade normal multivariada ϕ_g. Já no passo M, com respeito aos dados e valores obtidos no passo E, as estimativas para as médias e probabilidades são:

$$\hat{\pi}_g^{(s)} = \frac{\hat{n}_g^{(s-1)}}{n}; \quad \hat{\mu}_g^{(s)} = \frac{\sum_{i=1}^{n} \hat{z}_{i,g}^{(s-1)} x_i}{\hat{n}_g^{(s-1)}}; \quad \hat{n}_g^{(s-1)} = \sum_{i=1}^{n} \hat{z}_{i,g}^{(s-1)}. \tag{2.3.8}$$

A estimativa para a matriz de covariância $\hat{\Sigma}_g$ depende da parametrização. Celeux e Govaert (1995), por exemplo, estimaram Σ_k no passo M parametrizada por decomposição de autovalores. A forma de expressar a matriz de covariância possui um papel fundamental na determinação do tipo de modelo de mistura de normais.

2.3.4 Tipos de GMM

Bouveyron et al. (2019) explica que a grande quantidade de parâmetros que frequentemente surge em um GMM pode levar a problemas de estimação e interpretação

dos resultados. Para solucionar o problema, uma forma parcimoniosa de expressar o modelo de mistura de normais é por meio da decomposição de autovalores da matriz de covariância dos grupos:

$$\Sigma_g = \lambda_g D_g A_g D_g^T. \tag{2.3.9}$$

Na equação:

- Σ_g é a matriz de covariância dos grupos;
- λ_g é uma constante de proporcionalidade;
- D_g é a matriz de autovetores de Σ_g;
- A_g é uma matriz diagonal cujo elementos são proporcionais a Σ_g;

Cada um dos fatores de 2.3.9 é responsável por uma propriedade geométrica dos componentes da mistura. A matriz diagonal A_g é responsável pela forma do modelo. A constante de proporcionalidade λ_g é responsável pelo volume ocupado pela g-ésima componente no espaço R^d, sendo proporcional a $\lambda_g^d det(A_g)$. Já a matriz de autovalores D_g determina a orientação das componentes no espaço R^d.

Em razão das interpretações geométricas, a decomposição dos diferentes tipos de GMM se chama *Volume-Shape-Orientation* (VSO). Existem 2 modelos univariados e 14 modelos multivariados de misturas Gaussianas, nos quais as diferenças entre cada modelo são as restrições aplicadas nos fatores geométricos da decomposição. Com a aplicação das restrições, volume, forma e orientação podem ficar constantes entre as componentes da mistura, diminuindo os parâmetros do modelo e facilitando a estimação. Além dos 3 fatores do VSO, a matriz de covariância do modelo também pode ser esférica ou diagonal, de forma que fique proporcional a matriz identidade I.

Para identificar os modelos multivariados, uma sequência de 3 letras é utilizada. No caso de algum dos fatores geométricos for igual entre as componentes, a letra indicada será um "E"de *equal*. Nos modelos em que as propriedades geométricas forem variáveis entre as componentes da mistura, a letra será "V"de "varying". Na propriedade da forma e da orientação, a letra "I"indica que $A_g = I$, ou $D_g = I$, que são os casos nos quais os *clusters* são esféricos ou diagonais. Os tipos de GMM estão dispostos na Tabela 1.

Identificador	Modelo	Distribuição	Volume	Forma	Orientação
E	σ^2	Univariada	Igual		
V	σ_g^2	Univariada	Variável		
EII	λI	Esférica	Igual	Igual	Não disponível
VII	$\lambda_g I$	Esférica	Variável	Igual	Não disponível
EEI	λA	Diagonal	Igual	Igual	Alinhado ao eixo
VEI	$\lambda_g A$	Diagonal	Variável	Igual	Alinhado ao eixo
EVI	λA_g	Diagonal	Igual	Variável	Alinhado ao eixo
VVI	$\lambda_g A_g$	Diagonal	Variável	Variável	Alinhado ao eixo
EEE	Σ	Elipsoidal	Igual	Igual	Igual
VEE	$\lambda_g DAD^T$	Elipsoidal	Variável	Igual	Igual
EVE	$\lambda DA_g D^T$	Elipsoidal	Igual	Variável	Igual
EEV	$\lambda D_g A D_g^T$	Elipsoidal	Igual	Igual	Variável
VVE	$\lambda_g DA_g D^T$	Elipsoidal	Variável	Variável	Igual
EVV	$\lambda D_g A_g D_g^T$	Elipsoidal	Igual	Variável	Variável
VEV	$\lambda_g D_g A D_g^T$	Elipsoidal	Variável	Igual	Variável
VVV	Σ_g	Elipsoidal	Variável	Variável	Variável

Tabela 1: Tipos de GMM.

Quanto menos restrições forem aplicadas ao modelo, mais parâmetros devem ser estimados. Em um caso com dados de baixa dimensão e poucas componentes no GMM, a quantidade de parâmetros estimados é relativamente próxima para todos os modelos. Entretanto, com o aumento das dimensões e o número de componentes dos modelos, a quantidade de parâmetros a serem estimados cresce exponencialmente (BOUVEYRON et al., 2019). Por essa razão, para o agrupamento de dados com alta dimensão, especialmente quando $p > n$, é mais fácil ajustar GMMs esféricos ou diagonais (SCRUCCA et al., 2023).

2.3.5 Seleção de Modelos

Inicialmente pensado por Schwarz (1978) e depois desenvolvido por Kass e Raftery (1995), o critério de informação Bayesiano, ou *Bayesian information criterion* (BIC) é uma forma de selecionar GMMs. O BIC é um critério que busca modelos com valores altos de máxima verossimilhança e parcimoniosos. Para obter modelos com parcimônia, o critério penaliza o número de parâmetros, fazendo com que modelos complexos e com muitos parâmetros sejam penalizados. Para um modelo M, a equação do BIC é dada por:

$$BIC = 2\log(\hat{L}) - k\log(n). \tag{2.3.10}$$

Na equação 2.3.10:

- $\hat{L}$ é o valor estimado da máxima verossimilhança do modelo M.
- k é o número de parâmetros estimados para o modelo M;
- n é o número de observações, ou tamanho da amostra;

Um outro critério semelhante ao BIC utilizado na seleção de modelos de mistura de normais é o *integrated complete-data likelihood* (ICL), obtido por:

$$\text{ICL} = \text{BIC} - E(M). \tag{2.3.11}$$

Com $E(M)$ sendo a entropia de classificação esperada do modelo M. O ICL é o BIC penalizado pela entropia esperada do modelo. A entropia será maior quando há maior incerteza na classificação. Consequentemente, o ICL vai favorecer GMMs que conseguem separar mais claramente os *clusters* (BOUVEYRON et al., 2019). Da forma como está expresso em 2.3.10 e 2.3.11, maiores valores do BIC e do ICL serão preferíveis.

2.4 Correlação

Suponha duas variáveis aleatórias X e Y com esperança $\mathbb{E}(X)$ e $\mathbb{E}(Y)$. Segundo Morettin e Bussab (2017), a covariância entre X e Y será obtida por:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]. \tag{2.4.1}$$

Ou, de forma equivalente:

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{2.4.2}$$

Ao utilizar os desvios padrões das variáveis, σ_X e σ_Y, é possível obter a correlação entre X e Y:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{2.4.3}$$

O coeficiente de correlação dado pela 2.4.3 é uma medida de relação linear entre as variáveis X e Y. Diferente da covariância, influenciada pela escala das variáveis, o coeficiente de correlação assume valores entre -1 e 1. No caso, -1 indica uma correlação negativa perfeita, 1 indica uma correlação positiva perfeita e 0 é a ausência de correlação.

2.5 Modelagem de Regressão

2.5.1 Regressão Linear

Um modelo de regressão linear com p variáveis explicativas pode ser expresso por:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon. \tag{2.5.1}$$

Na equação 2.5.1:

- Y é a variável resposta quantitativa;
- β_0 é o intercepto;
- $X_1, \ldots, X_p$ são as variáveis explicativas;
- $\beta_1, \ldots, \beta_p$ são os coeficientes estimados;
- ε é o termo de erro.

O modelo também pode ser representado por meio da notação matricial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.5.2}$$

Na equação 2.5.2:

- $\mathbf{y}$ é o vetor de valores y_i da variável resposta;
- $\mathbf{X}$ é a matriz das variáveis explicativas;
- $\boldsymbol{\beta}$ é o vetor de parâmetros;
- $\boldsymbol{\varepsilon}$ é o termo de erro.

Para ajustar o modelo, o método utilizado é o de mínimos quadrados ordinários. O método consiste em determinar os valores estimados para os coeficientes do modelo, os parâmetros $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$, que minimizam a soma do quadrado dos resíduos (SQRes). Conforme a notação de James et al. (2021), a soma dos quadrados dos resíduos pode ser expressa por:

$$\begin{aligned} \text{SQRes} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip} \right)^2 \end{aligned} \tag{2.5.3}$$

Com a notação matricial apresentada em 2.4.3, a estimativa para os parâmetros β que minimizam a soma dos quadrados dos resíduos é dada por:

$$\text{SQRes}(\beta) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.5.4}$$

Ao derivar 2.5.5 em relação a $\boldsymbol{\beta}$ e $\boldsymbol{\beta}^T$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009), o resultado obtido é:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \tag{2.5.5}$$

Os valores encontrados em 2.5.5 são as estimativas de mínimos quadrados dos coeficientes do modelo de regressão linear.

Há muitos outros tópicos de extrema importância na análise de regressão linear que estão fora do escopo deste trabalho. O livro de Kutner, Nachtsheim e Neter (2003) aborda os múltiplos tópicos de regressão linear em maior detalhe.

2.5.2 Regressão Ridge

Em cenários com variáveis explicativas correlacionadas, fenômeno chamado de multicolinearidade, uma técnica que pode ser aplicada é o uso de uma penalização para diminuir os coeficientes do modelo de regressão. A penalização utilizada é a ℓ_2. Desenvolvido por Hoerl e Kennard (1970), o modelo de regressão que utiliza a penalização ℓ_2 é chamado de regressão *ridge*. A estimativa dos β sujeitos a ℓ_2 (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) é:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{2.5.6}$$

O parâmetro λ controla o tamanho da penalização que os coeficientes sofrem. Quanto maior λ, mais próximo de 0 os coeficientes do modelo tendem a ficar. Uma forma equivalente de ver 2.5.6 é:

$$\begin{aligned} \hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \\ \text{sujeito a} \sum_{j=1}^{p} \beta_j^2 \leq t. \end{aligned} \tag{2.5.7}$$

Há uma equivalência entre os parâmetros λ em 2.5.7 e t em 2.5.8.

Quando há multicolinearidade entre as variáveis explicativas em um modelo de regressão, as estimativas dos coeficientes não são bem determinadas e apresentam alta variância (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A regressão *ridge*, por meio da penalização ℓ_2 busca aliviar esse problema.

2.5.3 Lasso

Desenvolvido em Tibshirani (1996), o lasso é um modelo de regressão que impõe nos coeficientes uma penalização ℓ_1. Segundo Hastie, Tibshirani e Friedman (2009), as estimativas dos coeficientes no *lasso* são:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{2.5.8}$$

Outra forma de expressar o *lasso* é:

$$\begin{aligned} \hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \\ \text{sujeito a} \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned} \tag{2.5.9}$$

Uma diferença importante entre a regressão *ridge* e o *lasso* é que na *ridge* os coeficientes são diminuídos até valores próximos de 0, porém no *lasso* os coeficientes menos relevantes no modelo são realmente zerados. Esse habilidade importante faz com que o *lasso* crie modelos esparsos, promovendo uma seleção de variáveis automática (HASTIE; TIBSHIRANI; WAINWRIGHT, 2015).

Um modelo que utiliza muitas variáveis explicativas é chamado de modelo flexível (JAMES et al., 2021). Um modelo que seja flexível demais pode captar muito ruído nos dados e obter uma performance preditiva ruim quando é testado. O processo é chamado de *overfitting*, ou sobreajuste. Por essa razão, um dos pontos fortes do lasso ao remover

variáveis é o de criar modelos menos flexíveis, e portanto, menos propensos ao *overfitting*.

2.5.4 Elasticnet

Uma abordagem desenvolvida por Zou e Hastie (2005) que combina a regressão *ridge* com o *lasso* é a técnica de *elasticnet*. Na regularização por *elasticnet*, a penalização é:

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \right). \tag{2.5.10}$$

Na prática, um modelo *elasticnet* é uma mistura da regressão *ridge* com o *lasso*. A penalização dos coeficientes será uma combinação linear de ℓ_1 e ℓ_2. No caso apresentado na 2.5.11, quando $\alpha = 0$ ocorre o *lasso*, enquanto que com $\alpha = 1$ ocorre a regressão *ridge*.

2.5.5 Métricas de Avaliação

Para comparar a performance preditiva de modelos, uma técnica bastante usual é o erro quadrático médio, ou *mean squared error* (MSE). O MSE é obtido pela expressão:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{2.5.11}$$

Ao aplicar a raiz quadrada no MSE, se obtém o *root mean squared error* (RMSE), dado por:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}. \tag{2.5.12}$$

Modelos preditivos melhores costumam errar menos. Portanto, apresentam valores de MSE e RMSE mais baixos.

2.6 Técnicas de Amostragem

2.6.1 Amostragem Aleatória Simples (AAS)

De acordo com Bolfarine e Bussab (2005), a amostragem aleatória simples (AAS) é a forma mais simples de selecionar uma amostra. O processo de amostragem aleatória simples com reposição (AASc) de tamanho n é:

- Passo 1: enumerar a população: $U = \{1, \ldots, N\}$;
- Passo 2: de forma equiprovável, selecionar um elemento de U por meio de um procedimento aleatório;
- Passo 3: repor o elemento amostrado na população U e sortear o próximo elemento;
- Passo 4:repetir os Passos de 1 a 3 n vezes.

A amostra aleatória simples sem reposição (AASs) é feita por um procedimento parecido com a AASc, porém o elemento amostrado não é reposto no Passo 3. O elemento amostrado é retirado da população, de tal forma que cada elemento só pode aparecer na amostra uma única vez.

2.6.2 Amostragem Estratificada

2.7 Validação Cruzada

Para estimar o erro de teste de um determinado modelo, é possível utilizar um procedimento de reamostragem chamado de validação cruzada, ou *cross-validation* (CV). Uma possibilidade de validação cruzada consiste na separação de um conjunto de dados para ajuste do modelo e outro conjunto de dados para teste, abordagem chamada de *validation set* (JAMES et al., 2021). Uma desvantagem desse método é que os erros são muito dependentes das particularidades do conjunto de teste específico.

Para lidar com o problema que surge da abordagem de *validation set*, um método possível é a técnica de *leave-one-out cross-validation* (LOOCV). Com n observações, na abordagem de LOOCV, o modelo é ajustado com $n-1$ observações, a observação restante sendo utilizada para teste do modelo. O processo é repetido para as n observações e o RMSE obtido por meio de LOOCV é:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{RMSE}_i. \tag{2.7.1}$$

Para conjunto de dados com uma quantidade muito grande de observações, obter a estimativa LOOCV do RMSE pode ser um processo excessivamente demorado. Para contornar o problema, há uma outra abordagem de validação cruzada conhecida como *k-fold cross-validation*. Nessa abordagem, o banco de dados é separado em k grupos de tamanho igual, chamados *folds*, o modelo é ajustado em $k-1$ *folds* e testado no remanescente. Semelhante ao que é obtido em 2.7.1, a estimativa de *k-fold cross-validation* para o RMSE é:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{RMSE}_i. \tag{2.7.2}$$

A técnica de *k-fold cross-validation* é um meio termo entre a abordagem por *validation set* e LOOCV. É menos dependente da particularidade de um único conjunto de teste, ao mesmo tempo em que sua computação é menos demorada que LOOCV. Segundo James et al. (2021), normalmente é utilizado 5 ou 10 *folds*.

3 Metodologia

3.1 Conjuntos de Dados

A aplicação das técnicas propostas ao longo do trabalho será feita em conjuntos de dados de natureza macroeconômica. O primeiro conjunto de dados é composto por 556 variáveis, com uma variável representando o mês e o ano da observação, uma variável resposta e 554 variáveis explicativas. O banco de dados possui 120 observações mensais, com início em março de 2013 e final em fevereiro de 2023. Nesse conjunto de dados, a variável resposta corresponde ao Índice de Atividade Econômica do Banco Central (IBC-Br). As variáveis explicativas correspondem à variáveis de produção, obtidas por meio de diversas fontes de dados. O IBC-Br é um indicador de atividade econômica calculado pelo Banco Central do Brasil. O índice é divulgado mensalmente, calculado por meio de dados de serviços, indústria e agropecuária. Devido a sua divulgação mensal, o IBC-Br é chamado de "prévia do PIB". Com exceção da data, todas as variáveis do banco de dados estão no formato de variação percentual em relação ao mesmo mês do ano anterior, ou seja, variação interanual percentual.

O segundo conjunto de dados contém 226 observações mensais de índices de inflação, de janeiro de 2005 a outubro de 2023. A variável resposta é o índice de inflação oficial do Brasil, o Índice de Preços ao Consumidor Amplo (IPCA), calculado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). As variáveis explicativas são índices de inflação de diversas agregações. No banco de dados, estão disponíveis índices calculados pela Fundação Instituto de Pesquisas Econômicas (Fipe) e pela Fundação Getúlio Vargas (FGV). Como na base da atividade econômica, os dados de inflação também estão registrados como a variação interanual percentual. A principal diferença entre os conjuntos de dados é que na base da inflação há uma hierarquia entre as variáveis, de acordo com o nível de agregação. Um exemplo com os dados da Fipe está presente na Figura 1.

Nos índices calculados pela Fipe, há 1 índice geral, 7 itens, 24 subitens 1, 73 subitens 2 e 260 subitens 3. Já nos índices calculados pela FGV, há 1 índice geral, 7 itens, 29 subitens 1, 55 subitens 2 e 462 subitens 3. Para o trabalho, os dados utilizados serão os da Fipe, pois há um maior número de *features*. Já o nível de agregação utilizado será os subitens 3, visto que é o nível mais desagregado dos dados.

IPC-FIPE	
Código	Dados
0	TODOS OS ITENS
1	ALIMENTAÇÃO
11	GÊNEROS ALIMENTÍCIOS
1101	ARROZ E FEIJÃO
110101	ARROZ

Figura 1: Estrutura hierárquica dos dados de índices de preço.

Vale ressaltar que o índice de inflação da Fipe utiliza um sistema de cálculo de variação quadrissemanal. O período de coleta dos dados abrange oito semanas, com cada variação quadrissemanal correspondendo à divisão dos preços médios das últimas quatro semanas (referência) pelos preços médios das quatro semanas anteriores (base). O banco de dados possui os resultados da primeira a quarta quadrissemana. No trabalho, apenas a quarta quadrissemana será utilizada, pois é o período que há mais informação acumulada.

3.2 Tratamento dos Dados

A primeira etapa do tratamento dos dados será verificar a presença de *outliers* entre as observações. Caso existam observações muito discrepantes, o tratamento de winsorização (WILCOX, 2021) será utilizado. A winsorização consiste em limitar as observações acima de um certo percentil como iguais ao percentil. Para exemplificar, uma winsorização de 90% significa que os dados abaixo do 5° percentil serão igualados ao valor do 5° percentil, assim como os valores acima do 95° percentil serão igualados ao 95° percentil. A winsorização é uma técnica de limitação de valores extremos, sem descartá-los.

No caso dos dados do IBC-Br, após o tratamento dos *outliers*, o foco do tratamento será os valores faltantes. Primeiro é preciso verificar a quantidade de observações vazias no conjunto de dados. Com poucos dados faltantes, o tratamento consistirá em imputar o valor da mediana da variável na observação vazia. No caso dos dados de inflação, as variáveis com mais de 30% de valores vazios ou constantes serão removidas da base de dados.

3.3 Análise Descritiva

Após o tratamento dos dados, estatísticas descritivas dos índices de atividade econômica e inflação, como mínimo, máximo, média, mediana, os quartis e o desvio padrão serão calculadas. Gráficos do IBC-Br e do IPCA ao longo dos anos também serão empregados para mostrar a evolução das variáveis no tempo. Para avaliar o efeito do tratamento aplicado aos dados, a análise descritiva será feita antes e depois de tratar os dados.

A última etapa descritiva é analisar a correlação entre a variável resposta e as variáveis preditoras. O foco será nas variáveis mais correlacionadas com o IBC-Br e com o IPCA, verificando a magnitude das correlações em valores absolutos. A correlação entre *features* e variável resposta será utilizada em uma das etapas da modelagem preditiva.

3.4 Modelagem Preditiva

3.4.1 Lasso com a Base Completa

O primeiro modelo utilizado será o lasso com todas as variáveis explicativas disponíveis, ou seja, a base de dados completa. No caso de dados superdimensionados, como a base da atividade econômica e da inflação, a redução de dimensão e seleção de variáveis é uma tarefa importante. Isso ocorre automaticamente no lasso, pois a penalização ℓ_1 zera os coeficientes das variáveis menos importantes, selecionando os preditores mais relevantes para o modelo final.

3.4.2 Lasso com as Variáveis Mais Correlacionadas com a Resposta

O segundo grupo de modelos testados utiliza proporções das *features* mais correlacionadas com a variável resposta. Para acompanhar a evolução dos resultados à medida que mais variáveis são incluídas no modelo, as proporções de 10%, 25%, 50% e 75% das variáveis mais correlacionadas serão utilizadas. Pelo mesmo motivo de seleção automática de variáveis, o lasso também será o modelo preditivo nesta etapa.

3.4.3 Lasso com Amostra das Variáveis pelos Grupos do GMM

O principal modelo proposto no trabalho é o lasso com uma amostra aleatória estratificada de variáveis presentes nos agrupamentos. Para gerar os agrupamentos de variáveis, a técnica escolhida é o modelo de mistura de normais, ou *Gaussian mixture model* (GMM). A amostra se baseará em tomar uma proporção de variáveis de cada grupo de forma aleatória, uma amostragem estratificada pelos grupos do GMM. As proporções das variáveis de cada grupo na amostra serão 10%, 25%, 50% e 75%.

O objetivo do agrupamento é que as variáveis de um mesmo grupo sejam semelhantes entre si e diferentes das variáveis de grupos distintos. Dessa forma, seria possível ter uma boa representação da informação presente nos dados com a seleção de apenas algumas variáveis de cada *cluster*. Em outras palavras, a ideia de fazer o agrupamento das variáveis por meio do GMM é a redução de dimensionalidade, em um contexto de dados com centenas de preditores.

3.4.4 Lasso com Amostra da Base Completa

A utilização de variáveis de cada agrupamento do GMM forma uma amostra estratificada, com os grupos da mistura de normais no papel dos estratos. Para comparar o resultado da seleção de variáveis por meio da amostragem estratificada, amostras aleatórias simples (AAS) das variáveis também serão testadas. Dessa forma, será possível verificar se a estratificação gerada pelo GMM contribuiu na estabilidade dos erros de previsão.

As proporções de 10%, 25%, 50% e 75% também serão utilizadas nessa etapa. É importante ressaltar que tanto na amostragem de variáveis pela estratificação dos grupos GMM, quanto na AAS de variáveis da base completa, a amostra será selecionada sem reposição. Dessa forma, a mesma variável não entrará mais de uma vez na mesma amostra.

3.4.5 Determinação da Penalização do Lasso

O hiperparâmetro de regularização λ, responsável pela penalização dos coeficientes de regressão no lasso, é determinado por validação cruzada. Dois critérios serão testados para determinar o λ. O primeiro critério é utilizar o valor da penalização que obteve o menor MSE por *cross-validation*. O segundo critério é utilizar o λ um erro padrão maior que o λ de menor MSE. A ideia do segundo critério é utilizar uma penalização maior

para gerar um modelo mais esparso e reduzir a tendência de *overfitting*. Para facilitar, ao longo do trabalho os hiperparâmetros obtidos de cada forma serão chamados de λ min MSE e λ 1 SE, sigla para *standard deviation*.

Para modelos que utilizam a base completa ou proporções das variáveis mais correlacionadas com a resposta, a técnica de validação cruzada empregada será de *leave-one-out cross-validation* (LOOCV). A justificativa é para não haver aleatoriedade nos resultados, algo que ocorre com *k-fold cross-validation*. Para os modelos que utilizam reamostragem das variáveis explicativas, a validação cruzada com 10 *folds* será utilizada por motivos de otimização de tempo de processamento computacional.

3.5 Previsões Fora da Amostra

Para avaliar os modelos propostos, os dados serão separados em um conjunto de treino e um conjunto de teste para as previsões. Para os dados do IBC-Br, os últimos 24 períodos serão utilizados para testar os modelos. Para os dados da inflação, já que há mais observações, os 36 últimos períodos serão utilizados para testar os modelos.

As previsões no conjunto de teste serão feitas somente um período (mês) adiante. Sempre que a previsão do período seguinte é feita, as observações reais, que antes estavam no conjunto de teste, são incorporadas nos dados de treinamento do modelo. O processo é repetido até o fim da janela de previsão. Essa técnica, desenvolvida por Tashman (2000), é chamada de previsões fora da amostra por origem rolante. Para simplificar, a técnica será chamada de previsão fora da amostra ao longo do trabalho.

As previsões fora da amostra, simulam o processo de *nowcasting*. As previsões são de curto-prazo e incorporam novas informações aos modelos. Como abordado previamente, a qualidade das previsões aumentam com a pontualidade (*timeliness*) que novas informações são incorporadas aos modelos (GIANNONE et al., 2013).

3.6 Reamostragem das Variáveis Explicativas

O resultado do lasso com apenas uma amostra de variáveis é dependente das particularidades daquela amostra específica, obtida aleatoriamente. Para contornar a aleatoriedade, o processo de seleção aleatória de *features* e previsão fora da amostra será repetido B vezes. No trabalho, será utilizado $B = 1000$.

A cada uma das repetições do algoritmo, as variáveis preditoras serão selecionadas

aleatoriamente com reposição. A ideia é que, com um número suficiente de repetições do processo de reamostragem de preditores, a variância dos RMSEs deve variar menos em torno da média.

Não apenas mil RMSEs serão obtidos, como também mil λ serão obtidos para o lasso. A distribuição dos λ também será analisada para cada modelo. A utilização dos λ, além dos RMSEs, servirá de auxílio na avaliação da estabilidade do algoritmo de seleção de *features*.

3.7 Estabilidade dos Erros de Previsão

Com a distribuição dos mil RMSEs obtidos por reamostragem, será possível avaliar a estabilidade dos erros de previsão. Duas técnicas serão aplicadas na análise da estabilidade dos RMSEs: o cálculo de estatísticas descritivas e gráficos de distribuição dos erros. As medidas-resumo são as de posição e dispersão, como média, mediana, quartis, desvio padrão e coeficiente de variação.

Os gráficos escolhidos para auxiliar na visualização da distribuição dos RMSEs são *boxplots* e gráficos de densidade. As densidades serão estimadas pela técnica de *kernel density estimation* (KDE). Semelhante ao histograma, o KDE é uma técnica não-paramétrica que estima a densidade diretamente dos dados por meio de suavizações (GRAMACKI, 2019).

3.8 Fluxo do Algoritmo Proposto

O fluxo do algoritmo proposto no trabalho é:

- Passo 1: Agrupar as variáveis explicativas por GMM.
- Passo 2: Selecionar uma amostra aleatória de cada grupo.
- Passo 3: Ajustar o lasso com as variáveis amostradas.
- Passo 4: Obter o λ do lasso por validação cruzada.
- Passo 5: Com o λ obtido, fazer previsão com o lasso um período a frente.
- Passo 6: Inserir a observação real nos dados de treino.
- Passo 7: Repetir o Passos 5 e 6 até o fim da janela de previsão.

- Passo 8: Calcular o RMSE fora da amostra.
- Passo 9: Repetir os Passos 2 a 8 B vezes.
- Passo 10: Analisar a distribuição dos B RMSEs fora da amostra.

4 Resultados

4.1 Resultados com Dados de Atividade Econômica

4.1.1 Análise Descritiva e Tratamento dos Dados

A trajetória do IBC-Br, de acordo com os dados disponíveis, está presente na Figura 2.

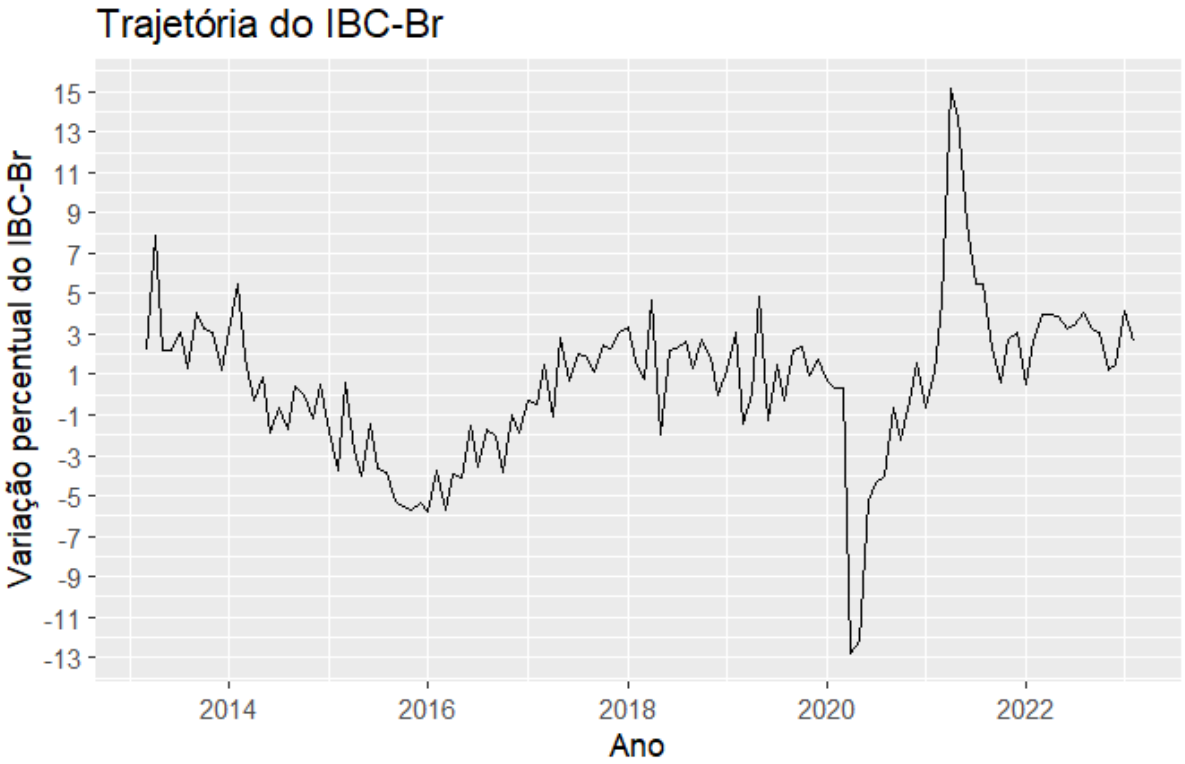

Figura 2: Trajetória da variação do IBC-Br.

As medidas-resumo dos resultados do IBC-Br estão na Tabela 2.

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
-12,785	-1,487	1,120	0,604	2,703	15,194

Tabela 2: Medidas de posição do IBC-Br.

Ao observar o a Figura 1 e os resultados presentes na Tabela 2, fica claro que há valores altamente discrepantes nos anos de 2020 e 2021. O motivo é a pandemia de Covid-19 e seu profundo impacto na economia. As observações relativas ao IBC-Br apresentam alta variabilidade, afirmação verificada pelas medidas de dispersão do IBC-Br na Tabela 3

Desvio Padrão	Coeficiente de Variação
3,855	638,019%

Tabela 3: Medidas de dispersão do IBC-Br.

É possível dizer que os valores extremos causados pela pandemia estão distorcendo os resultados e introduzindo variabilidade na série do IBC-Br. Por essa razão, os valores do IBC-Br e das variáveis explicativas passaram pelo tratamento de winsorização de 90%. Como resultado, as observações dos percentis 5 e 95 são agora o mínimo e o máximo, respectivamente. O gráfico da trajetória do IBC-Br após a winsorização está na Figura 3.

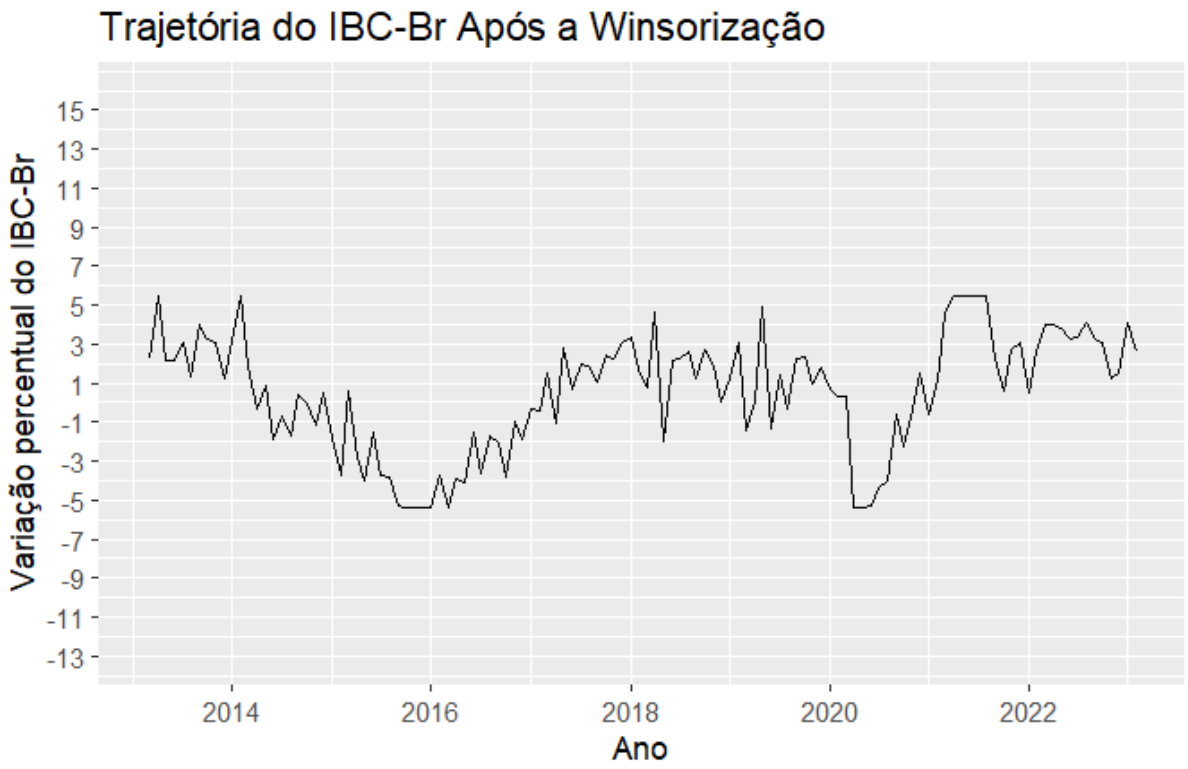

Figura 3: Gráfico da série do IBC-Br após a winsorização.

Após a winsorização, os *outliers* observados ao longo do período da pandemia foram suavizados, provocando um achatamento da curva em 2020 e 2021. As medidas-resumo após a winsorização estão na Tabela 4 e 5.

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
-5,390	-1,487	1,120	0,537	2,703	5,471

Tabela 4: Medidas de posição do IBC-Br após a winsorização.

Desvio Padrão	Coeficiente de Variação
3,006	559,801%

Tabela 5: Medidas de dispersão do IBC-Br após a winsorização.

Após a winsorização, houve uma redução nos valores de mínimo e máximo da variação do IBC-Br. Como é possível ver na Tabela 5, a variabilidade dos dados também diminuiu, apesar de ainda estar em um patamar alto.

Com os dados winsorizados, a próxima etapa do tratamento dos dados é lidar com os valores vazios. No conjunto de dados, há apenas quatro observações vazias no total, nenhuma da variável resposta. Os valores faltantes sofreram imputação pelo valor da mediana da variável.

A correlação entre preditores e variável resposta atinge valores elevados nesse banco de dados. Das 554 variáveis, mais de 80 possuem uma correlação, em módulo, de mais de 70% com o IBC-Br. No caso das 10 variáveis mais correlacionadas com o IBC-Br, a correlação está entre 85% a 88%, em módulo.

4.1.2 Resultados do Lasso com a Base Completa e as Mais Correlacionadas com o IBC-Br

Para servir de comparação com os resultados do algoritmo proposto no trabalho, cinco modelos foram utilizados. O primeiro é o lasso ajustado com a base completa, modelo que será chamado de Lasso BC. Os outros quatro modelos consistem no lasso com diferentes proporções dos preditores mais correlacionados com o IBC-Br. As proporções utilizadas foram 10%, 25%, 50% e 75%.

Os resultados de LOOCV para os cinco modelos com λ min MSE e λ 1 SE estão na Tabela 6.

Como é possível notar na Tabela 6, não necessariamente a maior quantidade inicial de variáveis no lasso resultará no modelo final com mais preditores. A quantidade de variáveis selecionadas dependerá da magnitude da penalização dos coeficientes por meio do hiperparâmetro λ.

Nas previsões fora da amostra, os modelos que utilizam as variáveis mais correlacionadas com o IBC-Br tiveram a correlação recalculada a cada previsão um período adiante. Dessa forma, a proporção de *features* mais correlacionadas com a resposta é sempre atualizada. O processo simula o *nowcasting*, onde os modelos se adaptam às novas informações disponíveis.

Modelo	λ	Variáveis Selecionadas	RMSE (LOOCV)
Lasso BC (λ min MSE)	0,025	49	0,583
Lasso BC (λ 1 SE)	0,203	17	0,628
Lasso 10% Cor (λ min MSE)	0,003	34	0,520
Lasso 10% Cor (λ 1 SE)	0,088	16	0,562
Lasso 25% Cor (λ min MSE)	0,050	24	0,566
Lasso 25% Cor (λ 1 SE)	0,140	22	0,600
Lasso 50% Cor (λ min MSE)	0,025	39	0,547
Lasso 50% Cor (λ 1 SE)	0,088	28	0,587
Lasso 75% Cor (λ min MSE)	0,025	48	0,568
Lasso 75% Cor (λ 1 SE)	0,168	23	0,614

Tabela 6: Resultados de validação cruzada dos modelos.

Os resultados dos RMSEs fora da amostra dos modelos com λ min MSE e com λ 1 SE estão na Tabela 7.

Modelo	RMSE Fora da Amostra
Lasso BC (λ min MSE)	1,133
Lasso BC (λ 1 SE)	1,074
Lasso 10% Cor (λ min MSE)	1,380
Lasso Cor 10% (λ 1 SE)	1,274
Lasso 25% Cor (λ min MSE)	1,110
Lasso Cor 25% (λ 1 SE)	1,074
Lasso 50% Cor (λ min MSE)	1,110
Lasso Cor 50% (λ 1 SE)	1,056
Lasso 75% Cor (λ min MSE)	1,112
Lasso Cor 75% (λ 1 SE)	1,059

Tabela 7: Resultados fora da amostra dos modelos.

Primeiramente, é possível conferir que a opção de utilizar o hiperparâmetro λ 1 SE foi um pouco melhor em todos os modelos testados. Ao utilizar uma regularização ℓ_1 maior, os coeficientes da regressão são mais penalizados, resultando em modelos com menos variáveis, ou seja, mais simples. Por essas razões, com os dados do IBC-Br, o critério de escolha do λ será o λ 1 SE.

Outro fator importante é que o erro de previsão é um pouco mais elevado ao se utilizar apenas 10% das variáveis mais correlacionadas. Entretanto, a partir de 25%, o RMSE parece estabilizar para as demais proporções. Isso indica que não é necessário utilizar todas as variáveis do banco de dados no modelo preditivo, sendo possível reduzir a dimensão do problema. O gráfico de previsão dos lassos com λ 1 SE está na Figura 4.

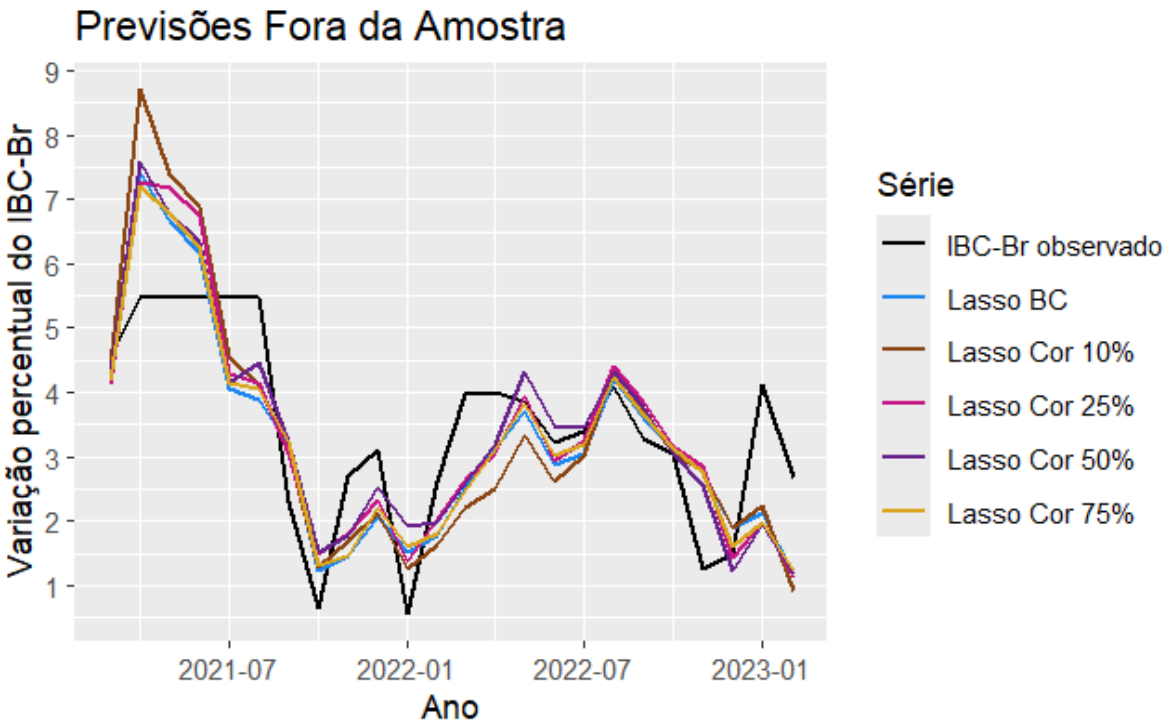

Figura 4: Gráfico de previsões dos modelos (λ 1 SE).

Pelo que é possível observar pela Figura 4, os modelos erraram mais nas previsões feitas no início e no final da janela de teste fora da amostra. Possivelmente outras formas de tratamento de dados poderiam melhorar a performance preditiva dos modelos. Um exemplo de mudança seria escolher outros limiares para a winsorização, dado o achatamento da curva no início da janela fora da amostra.

4.1.3 Ajuste e Seleção Aleatória de Variáveis do GMM

Após obter os RMSEs dos modelos que servem de comparação, o primeiro passo do algoritmo principal proposto no trabalho é ajustar o modelo de mistura de normais para agrupar as variáveis explicativas. Com as 96 observações dos dados de treino, foram ajustados diversos GMMs de 1 a 20 componentes de mistura. Os resultados dos modelos testados, de acordo com a métrica do BIC, está na Figura 5.

Pela métrica do BIC, os três melhores modelos ajustados foram o VEV com duas componentes (BIC = -352336,300), seguido pelo EEE e EEV com apenas uma componente, ambos com BIC = -353854,248. O resultado com a métrica ICL foi idêntico ao BIC. A quantidade de variáveis agrupadas em cada componente do modelo que obteve o melhor ajuste, o VEV 2 está presente na Tabela 8.

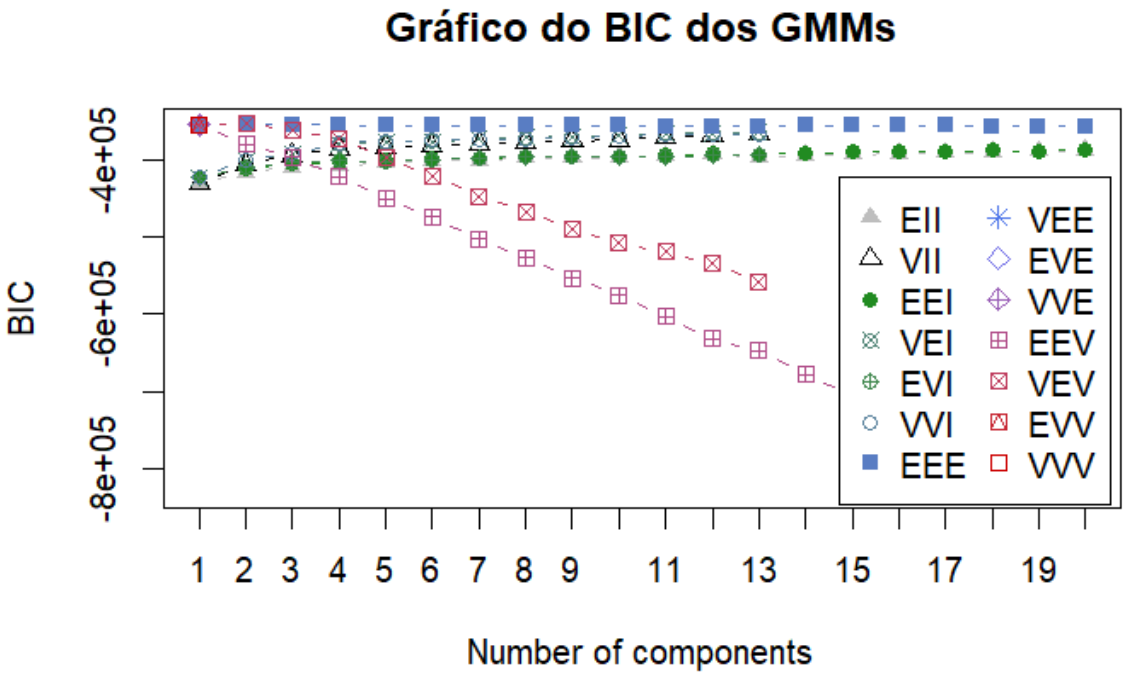

Figura 5: Gráfico do BIC dos GMMs.

Componente	Variáveis Agrupadas
1	75
2	479

Tabela 8: Agrupamento das variáveis com o modelo VEV 2.

O fato de que dois dos três modelos de melhor ajuste, de acordo com o BIC, são GMMs com apenas uma componente, pode ser um indicativo de que não há grupos bem definidos de preditores. Entretanto, é possível perceber na Figura 4 que os modelos EEE não apresentam uma queda brusca do BIC com o aumento do número de *clusters*. No intuito de achar outro GMM que acomode mais grupos, foi feito o gráfico do BIC de 1 a 20 componentes apenas do modelo EEE, presente na Figura 6.

Os resultados dispostos na Figura 6 mostram que, mesmo com o modelo EEE de maior BIC possuindo apenas uma componente, há um aumento brusco no valor da métrica quando 15 grupos são utilizados. Os resultados do agrupamento dos preditores por meio do modelo EEE 15 estão na Tabela 9.

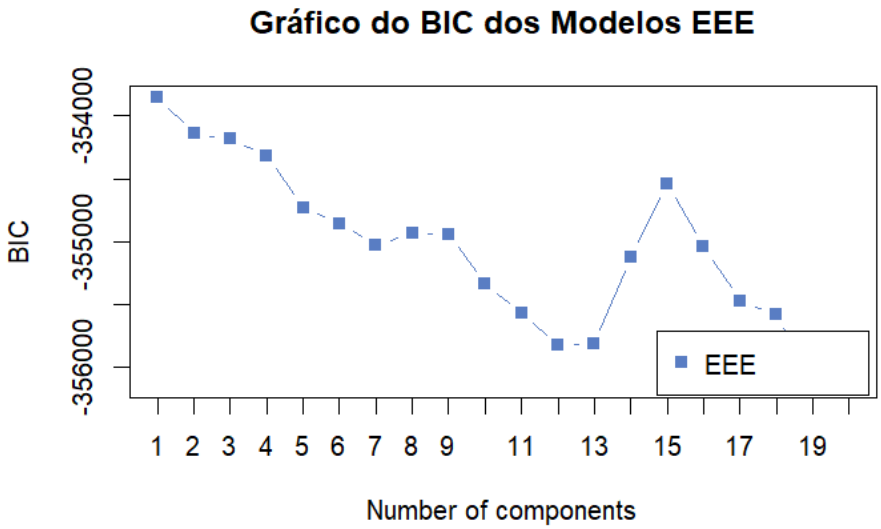

Figura 6: Gráfico do BIC dos modelos EEE.

Componente	Variáveis Agrupadas
1	9
2	26
3	11
4	14
5	99
6	191
7	45
8	5
9	4
10	31
11	72
12	23
13	1
14	1
15	22

Tabela 9: Agrupamento das variáveis com o modelo EEE 15.

Dois GMMs foram utilizados para agrupar as variáveis explicativas do banco de dados, o modelo VEV 2 e o modelo EEE 15. O VEV 2 por ter sido o modelo de melhor ajuste de acordo com o BIC e o ICL. Já o EEE 15 por possuir uma quantidade maior de grupos, e o BIC não estar muito abaixo do VEV 2.

Com os grupos do GMM, a amostra das variáveis é estratificada. A proporção de cada grupo pode resultar em números decimais. Nesse caso, a seleção das variáveis de cada grupo foi feita com arredondamento para o maior inteiro. Para *clusters* de apenas uma variável, qualquer proporção de amostra selecionará o grupo inteiro.

4.1.4 Lasso com Reamostragem de 10% das Variáveis

Nesta etapa da modelagem, o algoritmo conta com a reamostragem de variáveis explicativas. Para lidar com a variância dos resultados, dependentes de cada amostra selecionada aleatoriamente, o processo foi repetido mil vezes. No lugar de analisar apenas o valor do RMSE fora da amostra, a análise é focada na distribuição dos erros de previsão obtidos nas mil repetições do algoritmo.

Com uma amostra de variáveis explicativas na proporção de 10%, três modelos foram utilizados. Os dois primeiros consistem no lasso com uma amostra estratificada de *features* de cada grupo dos dois GMMs escolhidos, o VEV 2 e o EEE 15. O terceiro modelo, por sua vez, é o lasso com uma amostra de 10% variáveis da base de dados completa, sem estratificação pelas componentes do GMM.

O tempo de processamento foi de 135 segundos para amostra de variáveis agrupadas pelo VEV 2, 179 segundos para amostra do EEE 15 e 159 segundos para amostra da base completa. As medidas de posição dos mil RMSEs obtidos por reamostragem dos três modelos estão na Tabela 10. As medidas de dispersão estão na Tabela 11.

Modelo	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
VEV 2 (10%)	0,808	1,058	1,162	1,172	1,267	1,818
EEE 15 (10%)	0,838	1,086	1,177	1,186	1,275	1,693
Base Completa (10%)	0,810	1,063	1,157	1,176	1,275	2,001

Tabela 10: Medidas de posição dos RMSEs (amostras de 10% das variáveis).

Modelo	Desvio Padrão	Coeficiente de Variação
VEV 2 (10%)	0,152	12,950%
EEE 15 (10%)	0,140	11,772%
Base Completa (10%)	0,159	13,536%

Tabela 11: Medidas de dispersão dos RMSEs (amostras de 10% das variáveis).

A distribuição dos resultados dos RMSEs pode ser visualizada pelos *boxplots* na Figura 7.

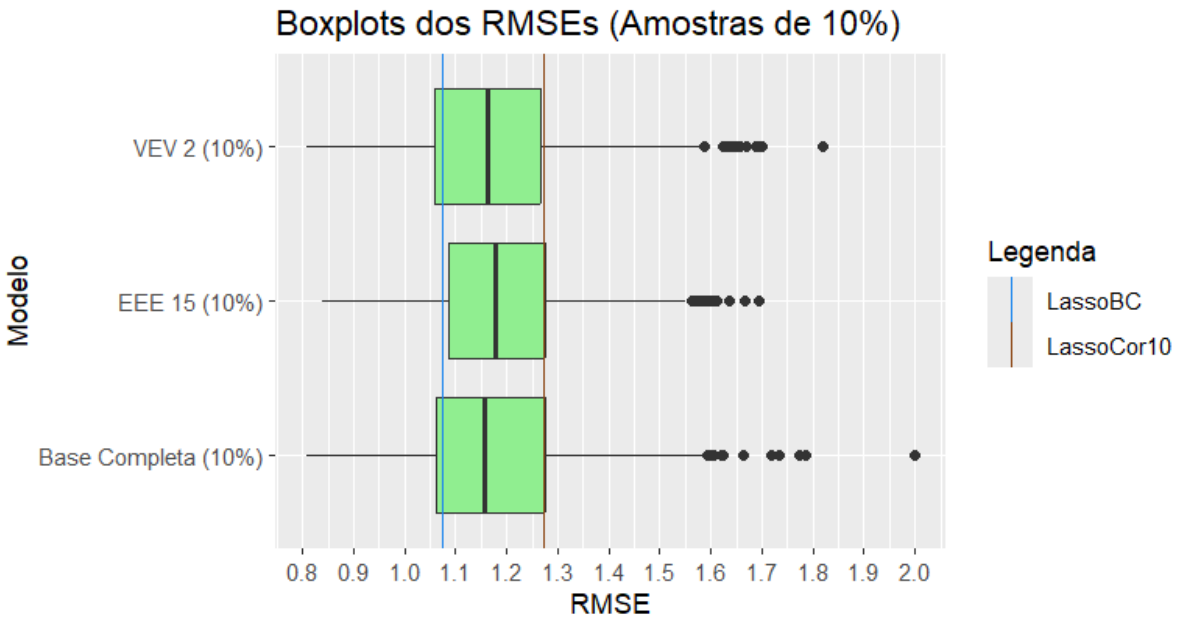

Figura 7: *Boxplots* dos RMSES (amostras de 10% das variáveis).

As densidades foram estimadas pela técnica não-paramétrica de *kernel density estimation* (KDE), podendo ser visualizadas na Figura 8. Nos gráficos, linhas verticais indicam o valor do RMSE obtido pelo lasso com a base completa, e pelo o lasso com 10% dos preditores mais correlacionados com o IBC-Br.

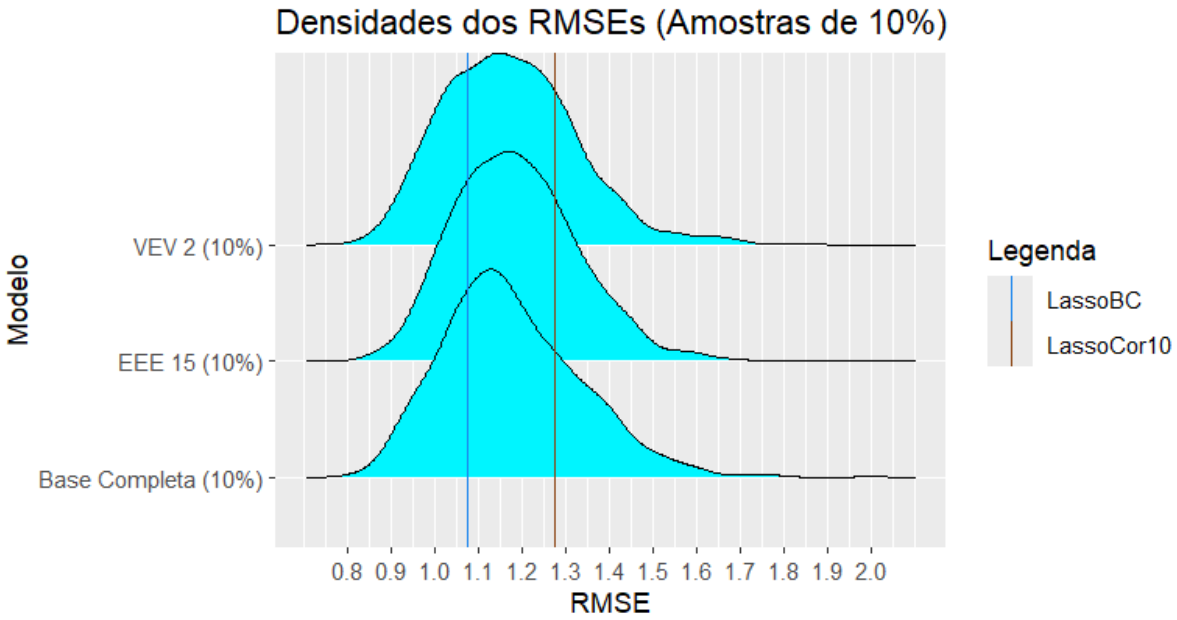

Figura 8: Densidades dos RMSEs (amostras de 10% das variáveis).

Tanto pelas tabelas, como pelos gráficos apresentados, é possível dizer que as distribuições dos RMSEs fora da amostra são muito semelhantes, mesmo com modelos que utilizaram as três técnicas diferentes de seleção de variáveis. Em média, a técnica obteve resultados um pouco piores que o lasso com todas as variáveis do banco de dados. No entanto, os resultados foram um pouco melhores que o lasso com as 10% mais correlacionadas. Há, entretanto, variabilidade nos erros de previsão.

4.1.5 Lasso com Reamostragem de 25% das Variáveis

O tempo para reamostragem das variáveis dos grupos do VEV 2 foi de 175 segundos, para o EEE 15 foi 185 segundos e para a amostra da base completa foi 181 segundos. As medidas-resumo dos RMSEs dos modelos que utilizaram uma amostra de 25% estão nas Tabelas 12 e 13. Nas Figuras 9 e 10 estão as representações das distribuições dos erros de previsão, os *boxplots* e gráficos de densidade.

Modelo	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
VEV 2 (25%)	0,782	1,001	1,062	1,072	1,134	1,451
EEE 15 (25%)	0,867	1,015	1,065	1,082	1,141	1,461
Base Completa (25%)	0,833	0,999	1,064	1,075	1,137	1,544

Tabela 12: Medidas de posição dos RMSEs (amostras de 25% das variáveis).

Modelo	Desvio Padrão	Coeficiente de Variação
VEV 2 (25%)	0,100	9,327%
EEE 15 (25%)	0,096	8,918%
Base Completa (25%)	0,101	9,374%

Tabela 13: Medidas de dispersão dos RMSEs (amostras de 25% das variáveis.

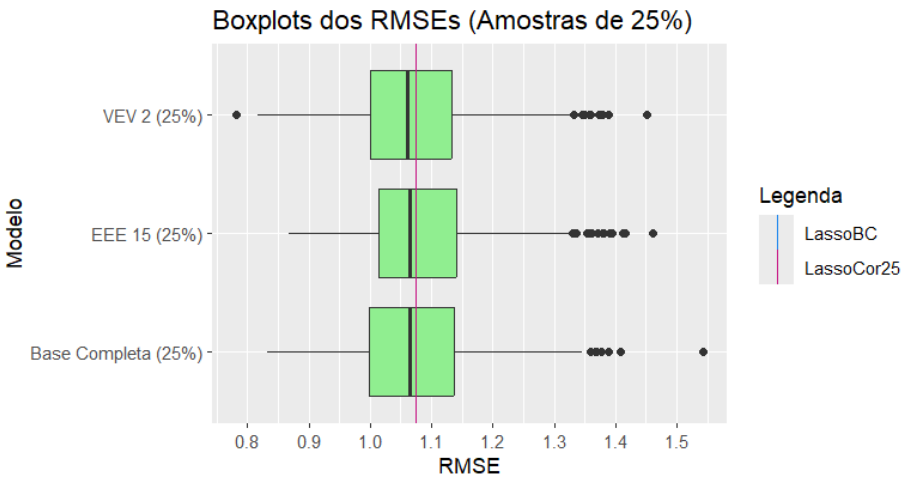

Figura 9: *Boxplots* dos RMSES (amostras de 25% das variáveis).

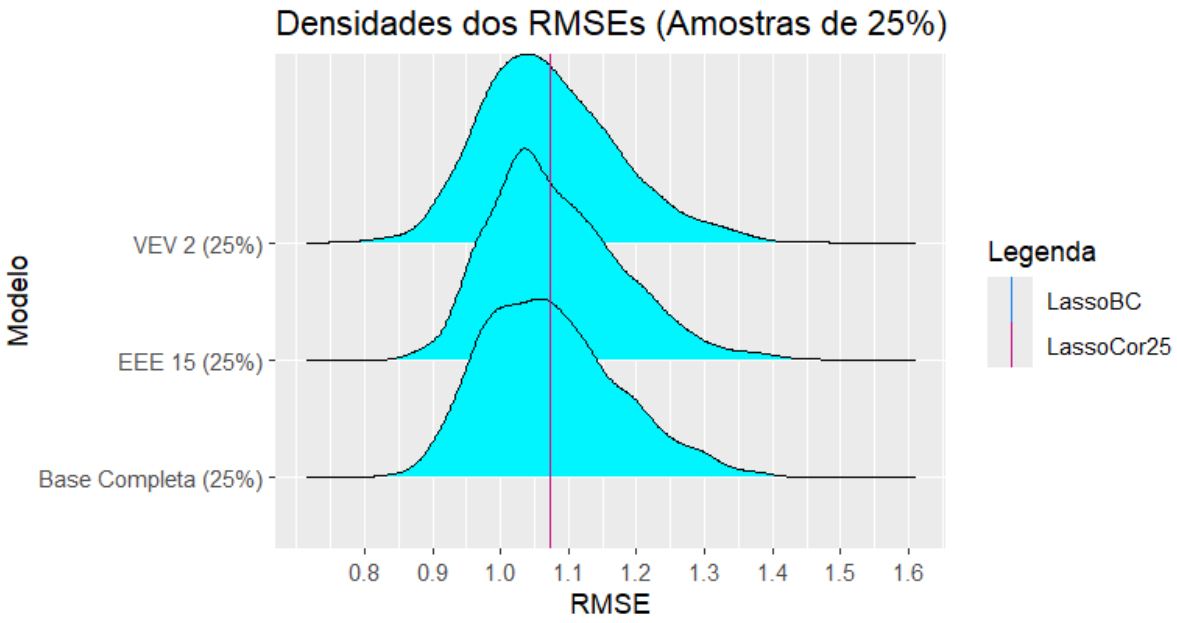

Figura 10: Densidades dos RMSEs (amostras de 25% das variáveis.

Em média, os RMSEs foram menores do que com o lasso com a amostra de 10% das variáveis, com a variabilidade dos RMSEs também um pouco menor. Da mesma forma que ocorreu com a amostra 10%, as distribuições dos RMSEs dos modelos com as três técnicas de seleção de variáveis são bem parecidas. O agrupamento das variáveis pelo GMM não parece ter feito uma diferença significativa.

4.1.6 Lasso com Reamostragem de 50% das Variáveis

Para simular as mil repetições com a amostra de 50%, o tempo gasto foi 305 segundos para amostra do VEV 2, 311 segundos para a amostra do EEE 15 e 309 segundos para a amostra da base completa. Os resultados das medidas de posição e dispersão estão na Tabela 14 e 15.

Modelo	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
VEV 2 (50%)	0,896	1,012	1,046	1,050	1,083	1,332
EEE 15 (50%)	0,876	1,017	1,048	1,054	1,087	1,269
Base Completa (50%)	0,868	1,009	1,047	1,051	1,085	1,370

Tabela 14: Medidas de posição dos RMSEs (amostras de 50% das variáveis).

Os *boxplots* e o gráfico das densidades dos RMSEs estão nas Figuras 11 e 12, respectivamente.

Modelo	Desvio Padrão	Coeficiente de Variação
VEV 2 (50%)	0,057	5,407%
EEE 15 (50%)	0,059	5,579%
Base Completa (50%)	0,060	5,692%

Tabela 15: Medidas de dispersão dos RMSEs (amostras de 50% das variáveis).

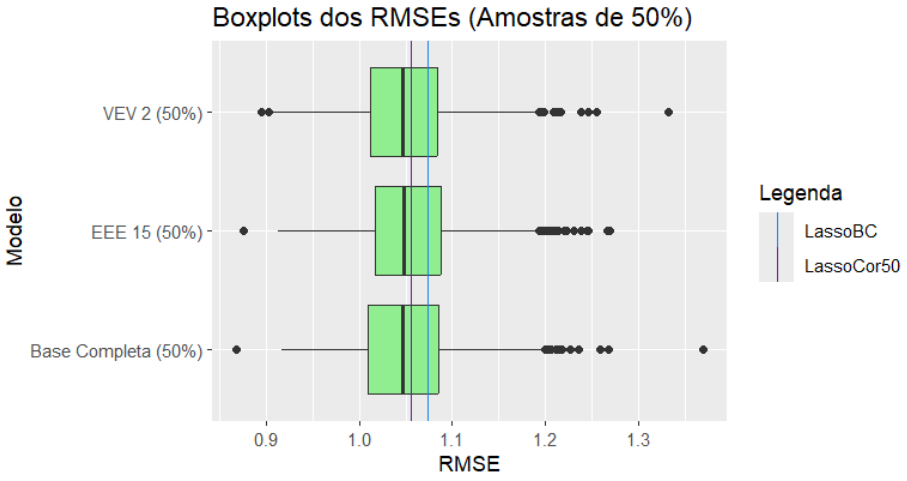

Figura 11: *Boxplots* dos RMSES (amostras de 50% das variáveis).

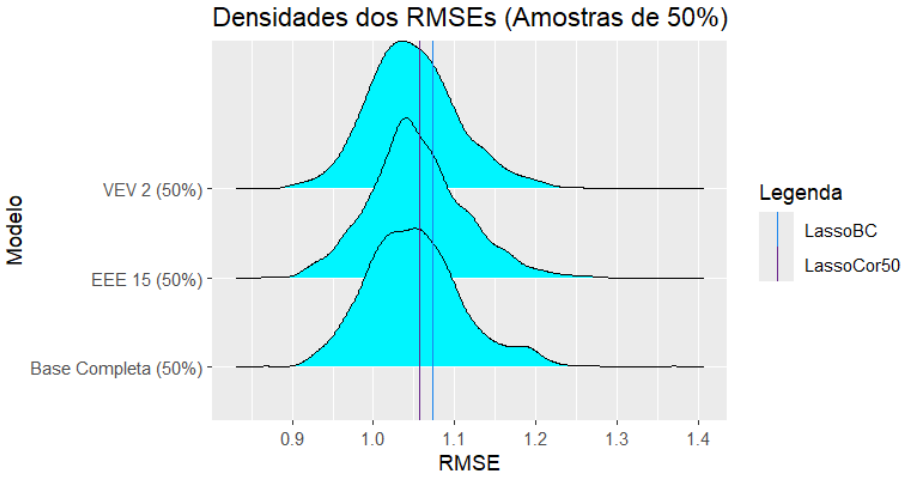

Figura 12: Densidades dos RMSEs (amostras de 50% das variáveis).

O RMSE médio dos modelos que utilizaram a amostra de 50% foi bem próximo ao RMSE médio obtido pelos modelos que utilizaram a amostra de 25%. No entanto, a variabilidade diminuiu, resultado que pode ser observado pelos valores do coeficiente de variação na Tabela 15.

4.1.7 Lasso com Reamostragem de 75% das Variáveis

O tempo gasto para simular as mil repetições foi de 487 segundos para amostra do VEV 2, 488 segundos para amostra do EEE 15 e 466 segundos para a amostra da base completa. As medidas-resumo estão presentes na Tabela 16 e 17.

Modelo	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
VEV 2 (75%)	0,945	1,033	1,056	1,059	1,081	1,244
EEE 15 (75%)	0,920	1,038	1,061	1,063	1,089	1,223
Base Completa (75%)	0,909	1,035	1,057	1,063	1,084	1,283

Tabela 16: Medidas de posição dos RMSEs (amostras de 75% das variáveis).

Modelo	Desvio Padrão	Coeficiente de Variação
VEV 2 (75%)	0,039	3,712%
EEE 15 (75%)	0,040	3,748%
Base Completa (75%)	0,044	4,127%

Tabela 17: Medidas de dispersão dos RMSEs (amostras de 75% das variáveis).

Os *boxplots* e gráficos de densidade estão nas Figuras 13 e 14.

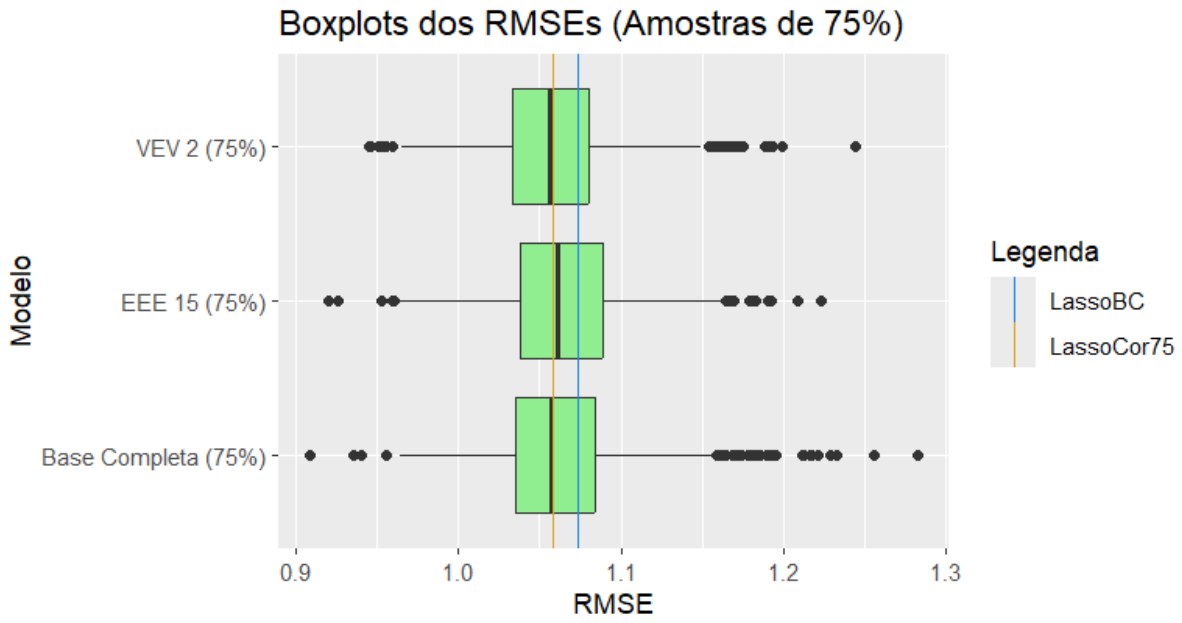

Figura 13: *Boxplots* dos RMSES (amostras de 75% das variáveis).

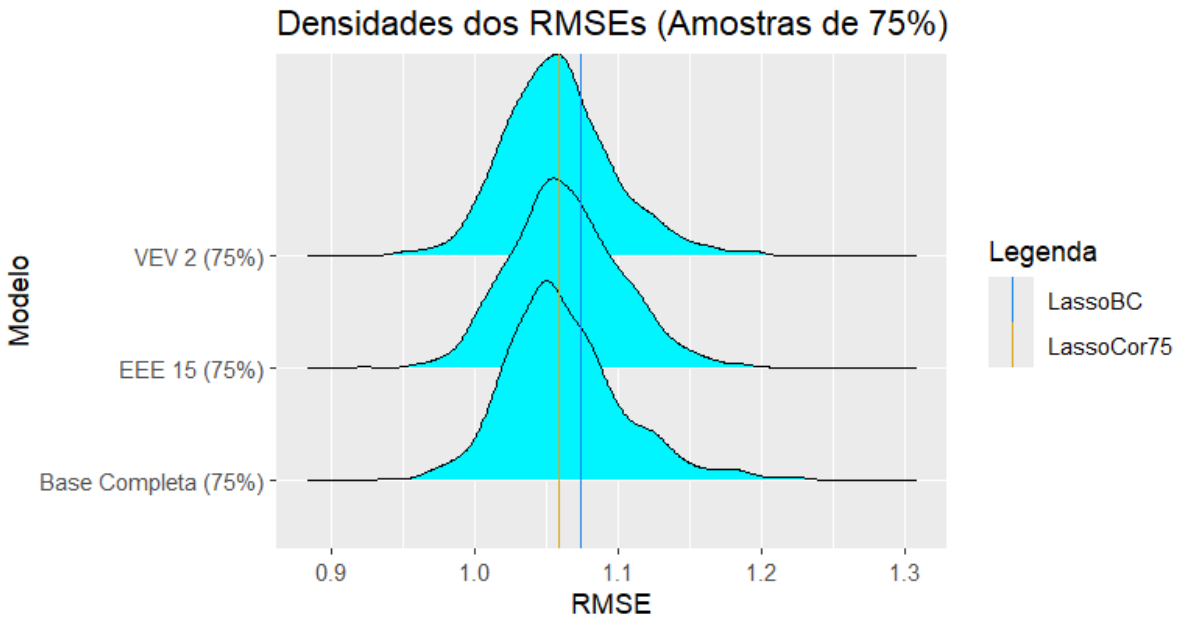

Figura 14: Densidades dos RMSEs (amostras de 75% das variáveis).

Os resultados com a amostra de 75% das variáveis foram parecidos com os obtidos com a amostra de 50%. No entanto, a variabilidade dos RMSEs diminuíram com o aumento da amostra.

4.1.8 Resultados Gerais

Utilizando os lassos que selecionaram proporcionalmente de cada um dos grupos do modelo de mistura VEV 2, os *boxplots* e densidades dos erros de previsão fora da amostra estão dispostos nas Figuras 15 e 16.

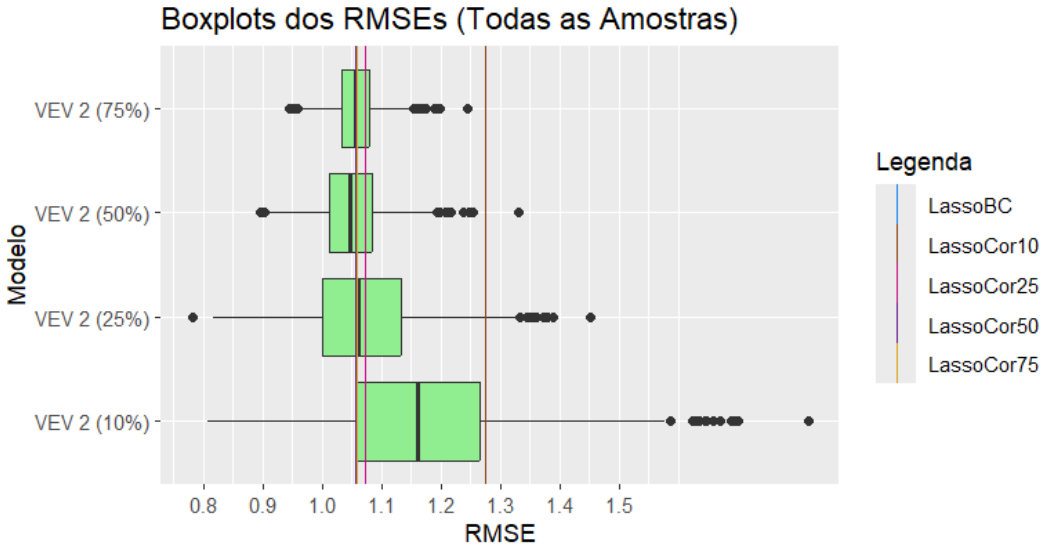

Figura 15: *Boxplots* dos RMSES (todas as amostras das variáveis).

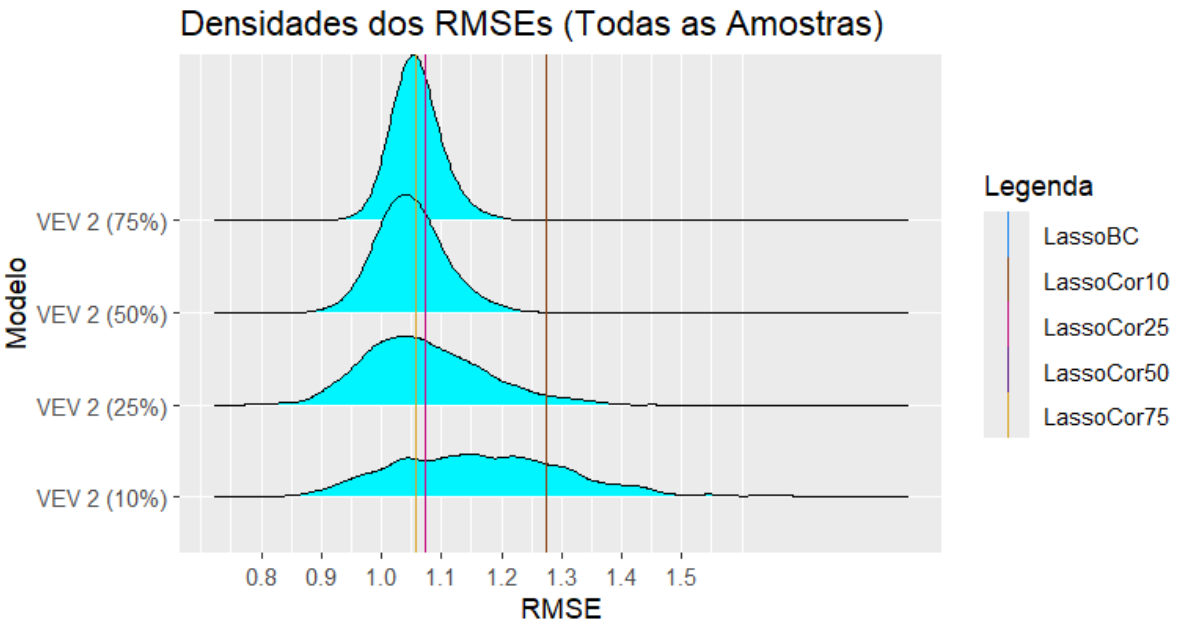

Figura 16: Densidades dos RMSES (todas as amostras das variáveis).

Pelo que foi possível observar, a partir de uma amostra de 25% de cada *cluster* do GMM, os RMSEs médio e mediano não mudaram muito. Nesses casos, os erros ficaram próximos dos obtidos pelo lasso com a base completa e com as *features* mais correlacionadas. Por outro lado, a variabilidade dos RMSEs diminuiu à medida em que as amostras aumentaram.

Os λ obtidos a cada repetição do processo de reamostragem podem ser visualizados pelos *boxplots* da Figura 17.

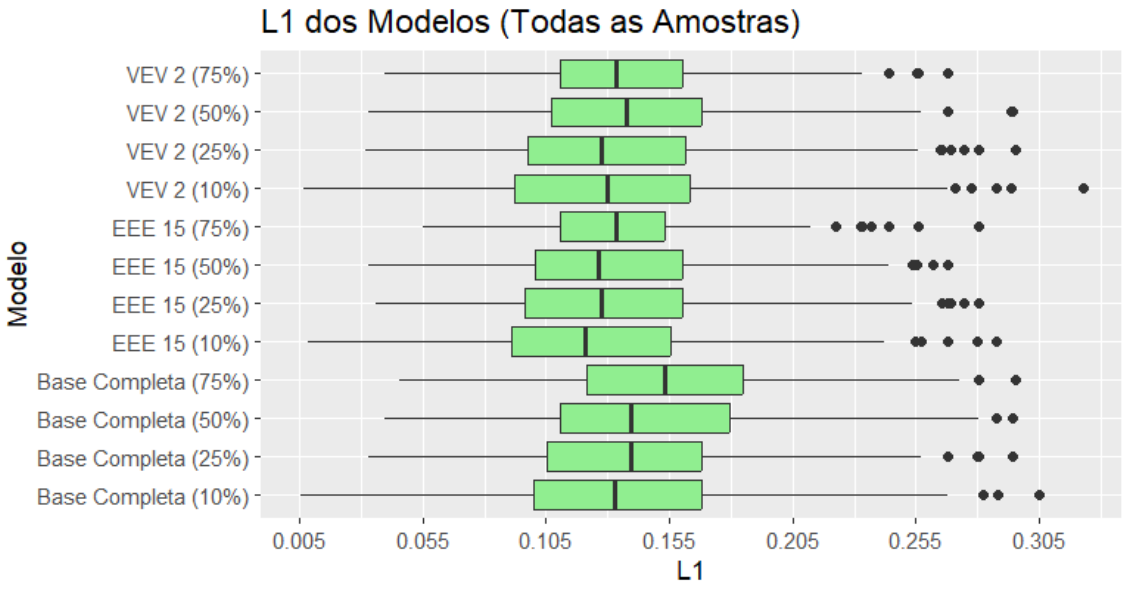

Figura 17: λ dos modelos (todas as amostras das variáveis).

Os λ obtidos por reamostragem não são muito informativos. A Figura 16 indica que os valores dos hiperparâmetros não mudaram muito de acordo com a técnica de seleção de *features*, nem com o tamanho da amostra de variáveis. A penalização ℓ_1 ficou em média em torno dos 0,110 e 0,150 para todos os modelos, independente da forma de seleção de variáveis.

4.1.9 Resultados com Novo Tratamento

Como é possível perceber pela Figura 2, a winsorização fez com que os valores do IBC-Br ficassem achatados nos períodos de 2020 e 2021 que sofreram o tratamento. Por essa razão, um novo tratamento de *outliers* foi testado. O tratamento consiste em decompor a série em tendência, sazonalidade e resto, remover as duas primeiras partes e encontrar os *outliers* no resto (HYNDMAN; ATHANASOPOULOS, 2021). Os valores extremos encontrados são substituídos por valores linearmente interpolados utilizando as observações vizinhas. No R, a função que implementa a técnica é chamada *tsoutliers*.

A trajetória do IBC-Br após esse novo tratamento pode ser visualizada na Figura 18. Como é possível perceber, a série tratada não apresenta o achatamento presente na Figura 3.

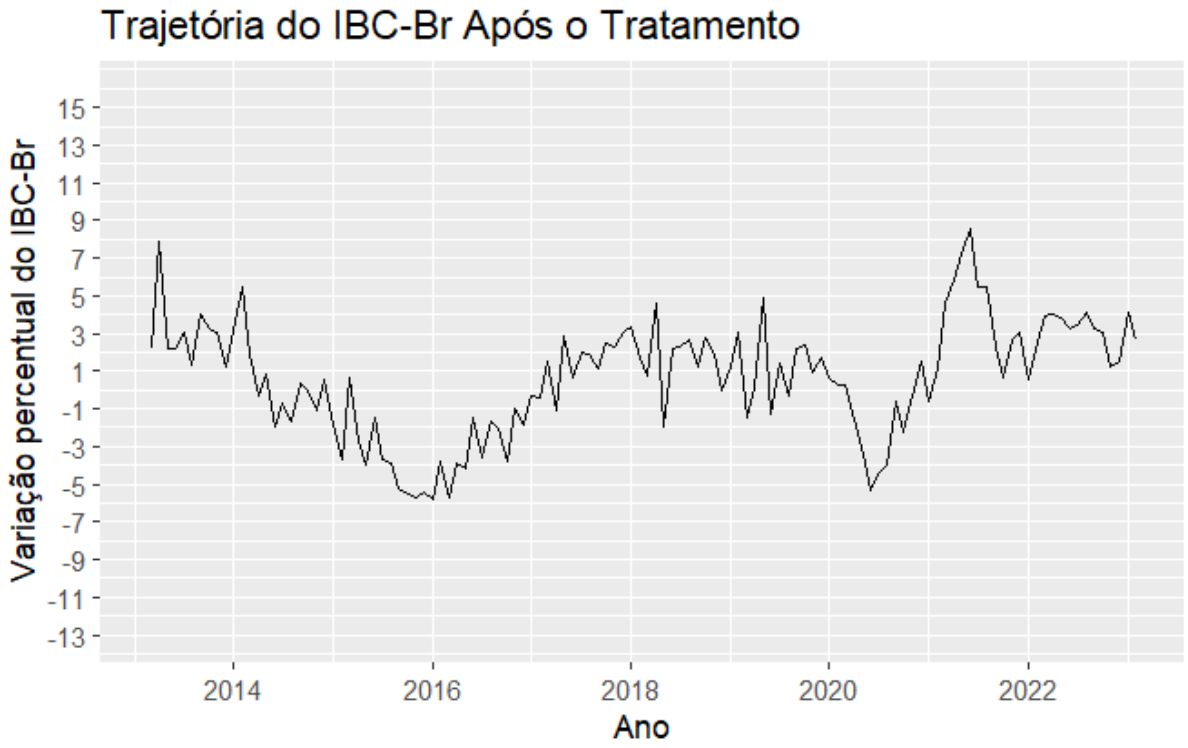

Figura 18: Gráfico da série do IBC-Br após o novo tratamento.

Com o novo tratamento, os resultados dos lassos com a base completa e as variáveis mais correlacionadas com o IBC-Br estão dispostos na Tabela 18.

Modelo	RMSE Fora da Amostra
Lasso BC (λ min MSE)	0,900
Lasso BC (λ 1 SE)	0,803
Lasso 10% Cor (λ min MSE)	1,074
Lasso Cor 10% (λ 1 SE)	1,024
Lasso 25% Cor (λ min MSE)	0,955
Lasso Cor 25% (λ 1 SE)	0,867
Lasso 50% Cor (λ min MSE)	0,946
Lasso Cor 50% (λ 1 SE)	0,936
Lasso 75% Cor (λ min MSE)	0,869
Lasso Cor 75% (λ 1 SE)	0,835

Tabela 18: Resultados fora da amostra dos modelos após o novo tratamento da base.

Em comparação com os resultados da Tabela 7, há uma melhora significativa nas previsões fora da amostra, com o melhor modelo apresentando um RMSE de 0,803. Como ocorreu com os modelos previamente testados, os melhores resultados foram obtidos com a utilização do λ 1 SE. A trajetória das previsões dos múltiplos modelos pode ser vista na Figura 19.

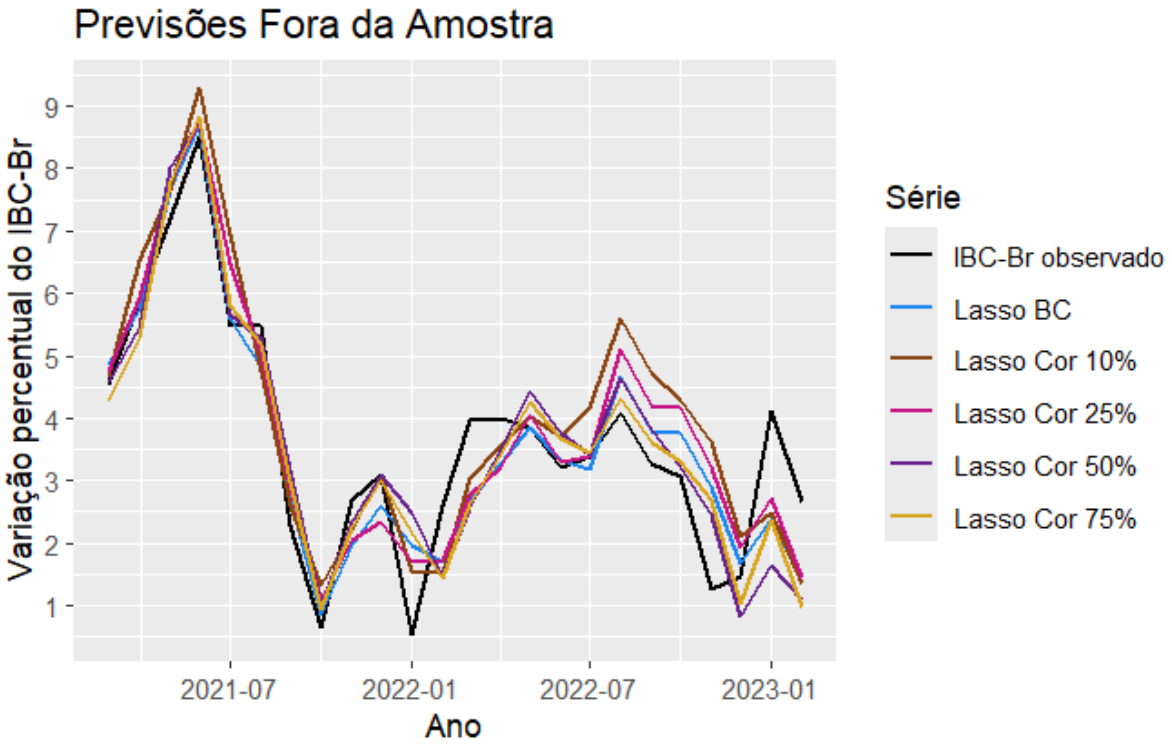

Figura 19: Gráfico de previsões dos modelos (λ 1 SE).

As previsões dos modelos com o novo tratamento acompanham melhor os valores reais do IBC-Br no início da janela de previsão fora da amostra. Com o tratamento da winsorização, a trajetória do IBC-Br no início da janela fora da amostra ficou achatada, com os modelos não prevendo bem, o que é possível ver na Figura 4.

Com o novo tratamento, o melhor agrupamento de variáveis, de acordo com o BIC, foi um GMM do tipo VEV 2, semelhante ao que foi obtido anteriormente. Os *boxplots* comparativos dos RMSEs obtidos para os dados com ambos os tratamentos estão na Figura 20.

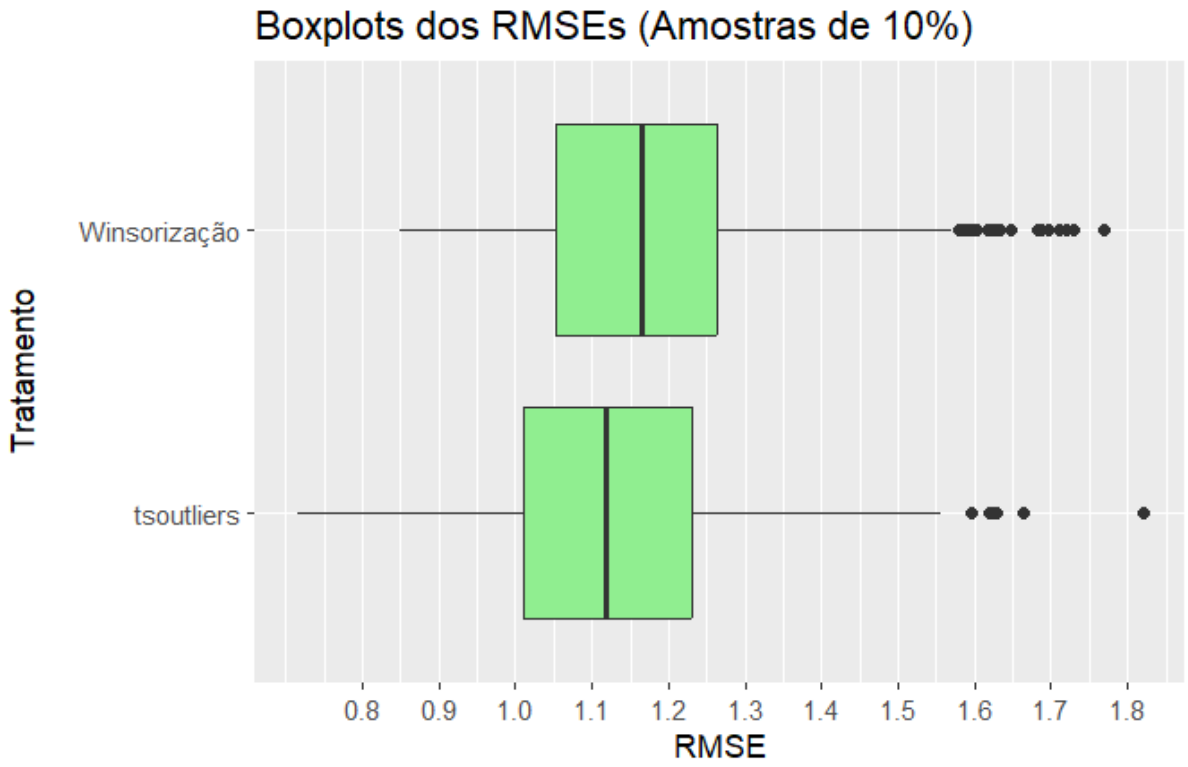

Figura 20: *Boxplots* dos RMSES com ambos os tratamentos (amostras de 10%).

Com a amostra de 10% das variáveis preditoras, o RMSE mediano do tratamento por meio do novo tratamento foi menor. Entretanto, a estabilidade desses RMSEs melhorou com a mudança de tratamento de *outliers*. A coeficiente de variação dos RMSEs com os dados winsorizados foi de 13,055%, enquanto que com o novo tratamento foi de 14,493%. Com o novo tratamento, o padrão de RMSEs menores, mas variabilidade um pouco maior também foi observado com as demais proporções de amostra de *features*.

Em resumo, o resultado mostra que a mudança de tratamento dos dados melhorou consideravelmente a qualidade preditiva dos modelos. Entretanto, o novo tratamento por si só não foi capaz de gerar previsões mais estáveis com o algoritmo proposto no trabalho.

4.2 Resultados com Dados de Inflação

4.2.1 Análise Descritiva e Tratamento dos Dados

Os dados utilizados nesta etapa do trabalho foram os índices de inflação calculados pela Fipe. A variável resposta é o IPCA, índice de inflação oficial do Brasil, calculado pelo IBGE. A trajetória da variação da inflação, ao longo dos anos está na Figura 21.

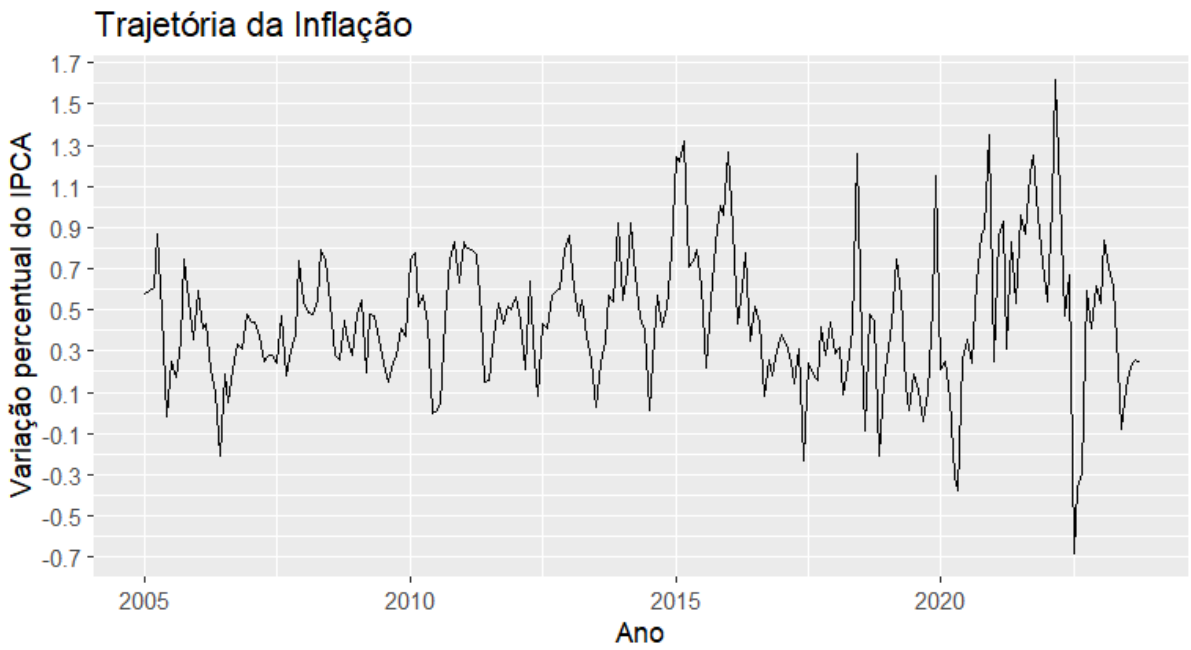

Figura 21: Trajetória da variação da inflação.

Na base de dados da variação da inflação, parte do tratamento consistiu em remover variáveis que possuem mais de 30% de valores faltantes e remover variáveis com valores constantes. Por fim, nas variáveis abaixo de 30% dos valores vazios, a imputação foi feita pela mediana da variável. Após esse tratamento, a quantidade de variáveis explicativas diminuiu de 554 para 490. Ao selecionar apenas o menor nível de agregação dos dados, sobram 399 *features*.

Assim como aconteceu com os valores da variação do IBC-Br, o momento de maior variabilidade do índice de inflação também é no período de 2020 a 2022, durante a pandemia de Covid-19. Entretanto, os valores extremos parecem menos influentes na inflação do que no IBC-Br. As medidas-resumo do IPCA estão nas Tabelas 19 e 20.

O coeficiente de variação do IPCA é de 73,394%, um resultado muito inferior ao que foi obtido no banco de dados de atividade econômica. Pelo fato de haver uma variabilidade menor, não foi utilizada uma técnica de tratamento de *outliers*.

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
-0,680	0,250	0,440	0,457	0,620	1,620

Tabela 19: Medidas de posição da inflação.

Desvio Padrão	Coeficiente de Variação
0,336	73,394%

Tabela 20: Medidas de dispesão da inflação.

A magnitude da correlação entre as variáveis explicativas e o IPCA é bem diferente do que foi observado nos dados de atividade econômica. No banco de dados anterior, as correlações entre preditores e variável resposta chegavam a mais de 85%. Por outro lado, nos dados da inflação, a variável mais correlacionada com o IPCA possui apenas 54% de correlação. As outras 10 *features* mais correlacionadas com o IPCA estão entre 30% e 42%. Os resultados indicam uma estrutura de correlação entre *features* e variável resposta consideravelmente mais fraca do que a observada nos dados de atividade econômica.

4.2.2 Resultados do Lasso com a Base Completa e as Mais Correlacionadas com o IPCA

Como há mais observações no banco de dados da inflação do que no do IBC-Br, a janela de previsão fora da amostra também foi maior. Com os dados da inflação, foram utilizados os últimos 36 meses para o teste fora da amostra.

Ao aplicar as mesmas técnicas utilizadas com os dados de atividade econômica, os resultados de validação cruzada e fora da amostra estão nas Tabelas 21 e 22, respectivamente.

Modelo	λ	Variáveis Selecionadas	RMSE (LOOCV)
Lasso BC (λ min MSE)	0,008	97	0,192
Lasso BC (λ 1 SE)	0,028	39	0,203
Lasso 10% Cor (λ min MSE)	0,007	26	0,186
Lasso 10% Cor (λ 1 SE)	0,022	24	0,196
Lasso 25% Cor (λ min MSE)	0,010	46	0,181
Lasso 25% Cor (λ 1 SE)	0,020	38	0,190
Lasso 50% Cor (λ min MSE)	0,006	88	0,187
Lasso 50% Cor (λ 1 SE)	0,023	47	0,198
Lasso 75% Cor (λ min MSE)	0,009	83	0,189
Lasso 75% Cor (λ 1 SE)	0,027	42	0,201

Tabela 21: Resultados de validação cruzada dos modelos.

A utilização da regularização λ min MSE obteve resultados melhores no RMSE

Modelo	RMSE Fora da Amostra
Lasso BC (λ min MSE)	0,271
Lasso BC (λ 1 SE)	0,291
Lasso 10% Cor (λ min MSE)	0,261
Lasso 10% Cor (λ 1 SE)	0,283
Lasso 25% Cor (λ min MSE)	0,273
Lasso 25% Cor (λ 1 SE)	0,284
Lasso 50% Cor (λ min MSE)	0,266
Lasso 50% Cor (λ 1 SE)	0,283
Lasso 75% Cor (λ min MSE)	0,264
Lasso 75% Cor (λ 1 SE)	0,289

Tabela 22: Resultados fora da amostra dos modelos.

fora da amostra, o que justifica sua utilização na etapa de reamostragem. Esse resultado é o oposto do que ocorreu com a base do IBC-Br, onde os modelos com λ 1 SE obtiveram melhores erros de previsão fora da amostra. Outro ponto importante é que o RMSE das previsões parece ter estabilizado a partir de 10% das variáveis mais correlacionadas com a resposta. Nos dados da variação do IBC-Br, por outro lado, o RMSE só estabilizou a partir de 25% dos preditores mais correlacionados.

O gráfico das previsões fora da amostra está na Figura 22.

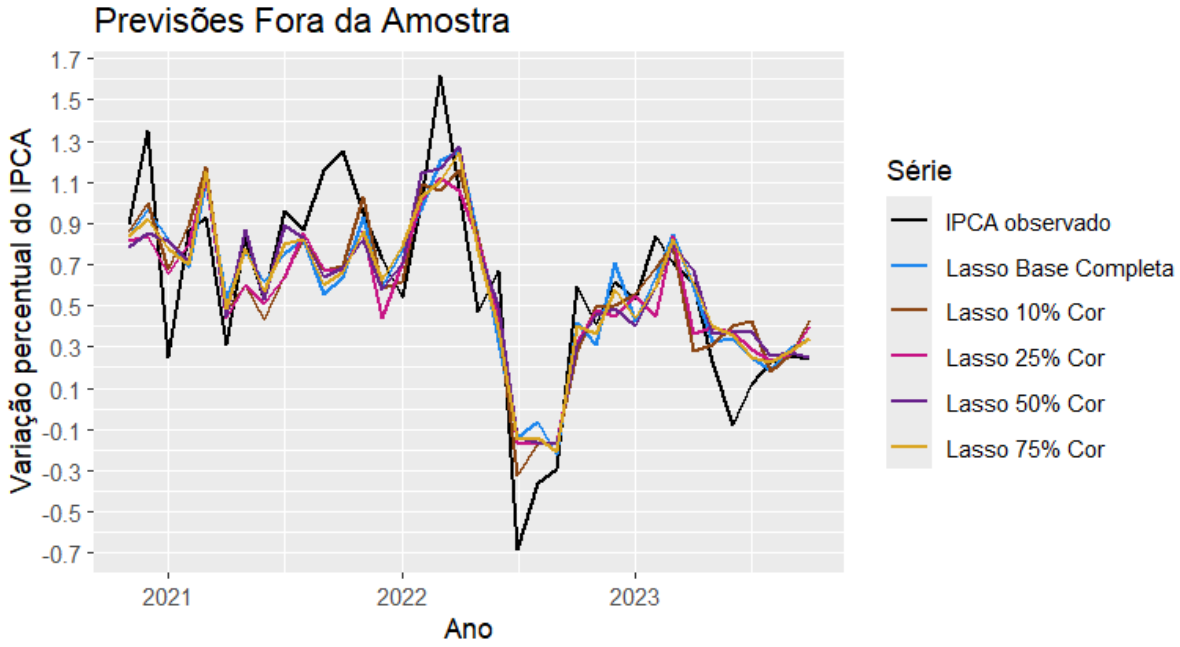

Figura 22: Gráfico de previsões dos modelos com λ min MSE.

Pelo que é possível perceber pela Figura 22, os maiores erros de previsão parecem ocorrem em momentos onde há aumentos e quedas bruscas do IPCA, como no ano de 2022. O modelos também apresentam menor variabilidade que o IPCA observado no

início da janela fora da amostra, ao longo do ano de 2021.

4.2.3 Ajuste e Seleção Aleatória de Variáveis do GMM

Ao ajustar o modelo de mistura de normais com os dados de treinamento para agrupar as variáveis explicativas, o resultado pela métrica BIC foi idêntica à métrica ICL. O resultado do BIC dos modelos de 1 a 20 componentes está na Figura 23.

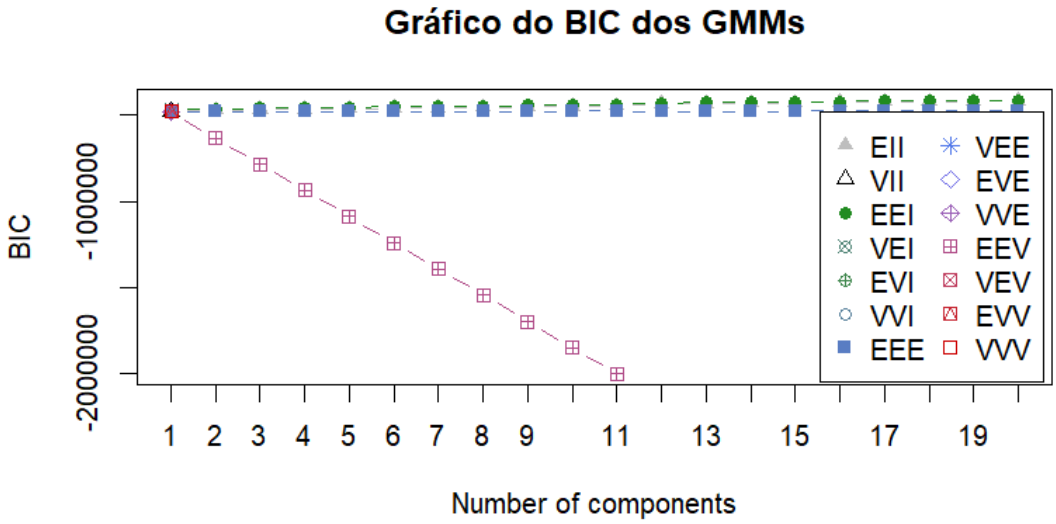

Figura 23: Gráfico do BIC dos GMMs.

O tipo de GMM que se ajustou melhor, de acordo com o BIC, foi o modelo EEI. O gráfico de BIC apenas do modelo EEI, de 1 a 20 componentes, está na Figura 24.

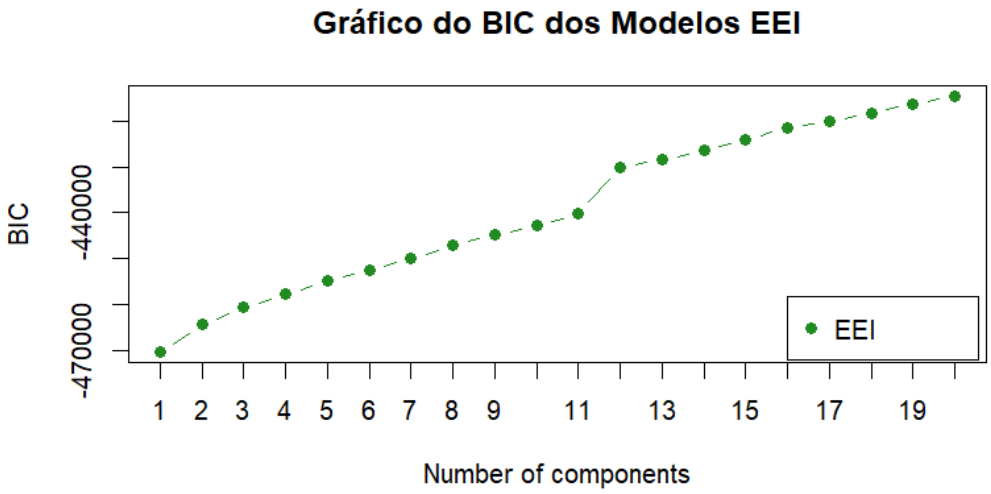

Figura 24: Gráfico do BIC dos GMMs.

Como é possível perceber pela Figura 24, há um aumento brusco do BIC com 12 componentes da mistura. Por esse motivo, o modelo utilizado será o EEI com 12 componentes. Os resultados do agrupamento do GMM escolhido está na Tabela 23

Componente	Variáveis Agrupadas
1	375
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	14

Tabela 23: Agrupamento das variáveis com o modelo EEI 12.

A proporção amostrada das variáveis será apenas para o grupo 1. Para os demais grupos, todas as variáveis serão utilizadas no modelo, visto que são agrupamentos de apenas uma variável, ou 14 no caso do último grupo.

Cada uma das 14 variáveis do grupo 12 são legumes ou verduras, todas presentes no subitem 1 dos produtos *in natura*. Esse resultado mostra que uma técnica de agrupamento estatística foi capaz de encontrar um padrão real presente nos dados.

4.2.4 Lasso com Reamostragem de 10% das Variáveis

O tempo de processamento foi 151 segundos com a amostra das variáveis dos grupos do GMM e 112 segundos para AAS das variáveis da base completa. As medidas-resumo dos modelos com amostra de 10% estão nas Tabelas 24 e 25.

Modelo	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
EEI 12 (10%)	0,240	0,388	0,408	0,397	0,421	0,468
Base Completa (10%)	0,268	0,415	0,438	0,430	0,460	0,520

Tabela 24: Medidas de posição dos RMSEs (amostras de 10% das variáveis).

Modelo	Desvio Padrão	Coeficiente de Variação
EEI 12 (10%)	0,040	10,080%
Base Completa (10%)	0,046	10,617%

Tabela 25: Medidas de dispersão dos RMSEs (amostras de 10% das variáveis).

Ao observar os resultados da média e mediana dos RMSEs fora da amostra, já é possível notar que os valores estão bem distantes dos que foram atingidos com a base completa ou com as proporções de variáveis mais correlacionadas com a inflação. O *boxplot* e o gráfico de densidade dos RMSEs estão nas Figuras 25 e 26.

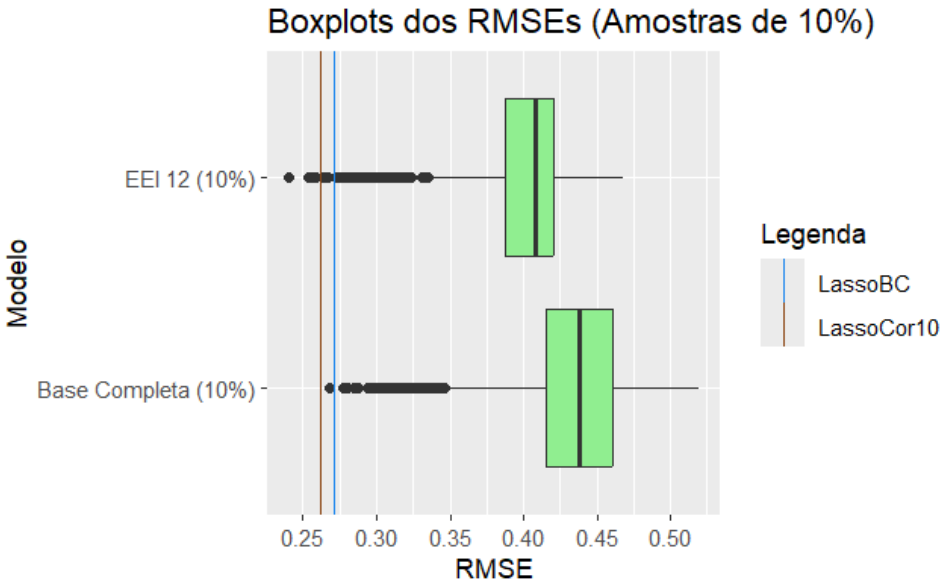

Figura 25: *Boxplots* dos RMSES (amostras de 10% das variáveis).

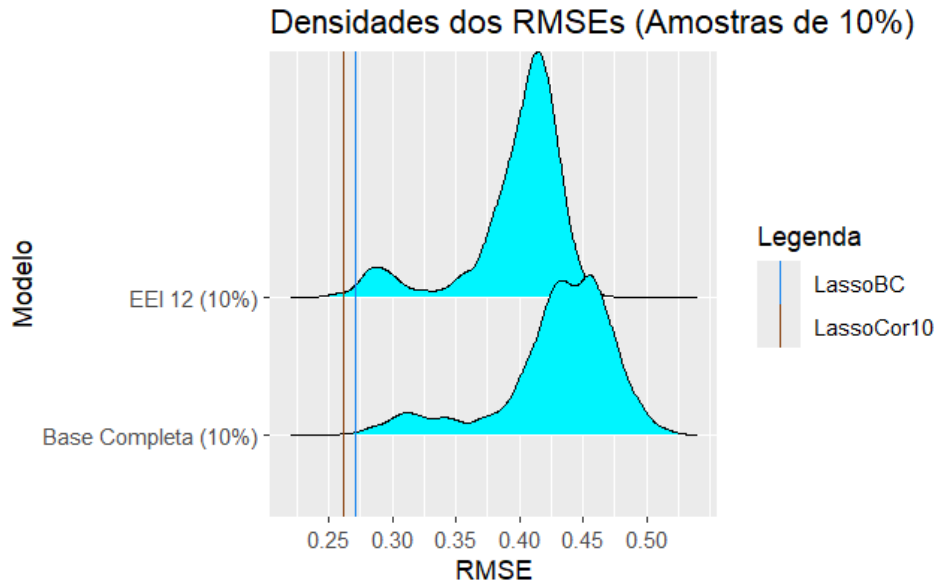

Figura 26: Densidades dos RMSEs (amostras de 10% das variáveis).

Os resultados para a amostra de 10%, tanto dos grupos do GMM quanto da base completa, estão muito diferentes do que foi observado com os dados do IBC-Br. A distribuição dos RMSEs se assemelha com uma distribuição normal bimodal. A concentração à direita é bem maior, distante dos RMSEs obtidos ao utilizar todas as variáveis, ou com

as 10% mais correlacionadas com o IPCA.

4.2.5 Lasso com Reamostragem de 25% das Variáveis

Com a proporção de 25%, o tempo gasto para as mil repetições foi de 327 segundos com as variáveis dos grupos do GMM e 263 segundos com a amostra de variáveis da base completa. As estatísticas descritivas estão nas Tabelas 26 e 27.

Modelo	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
EEI 12 (25%)	0,243	0,336	0,382	0,363	0,399	0,444
Base Completa (25%)	0,247	0,357	0,395	0,381	0,418	0,480

Tabela 26: Medidas de posição dos RMSEs (amostras de 25% das variáveis).

Modelo	Desvio Padrão	Coeficiente de Variação
EEI 12 (25%)	0,050	13,700%
Base Completa (25%)	0,052	13,647%

Tabela 27: Medidas de dispersão dos RMSEs (amostras de 25% das variáveis).

O *boxplot* e gráfico de densidade estão nas Figuras 23 e 24.

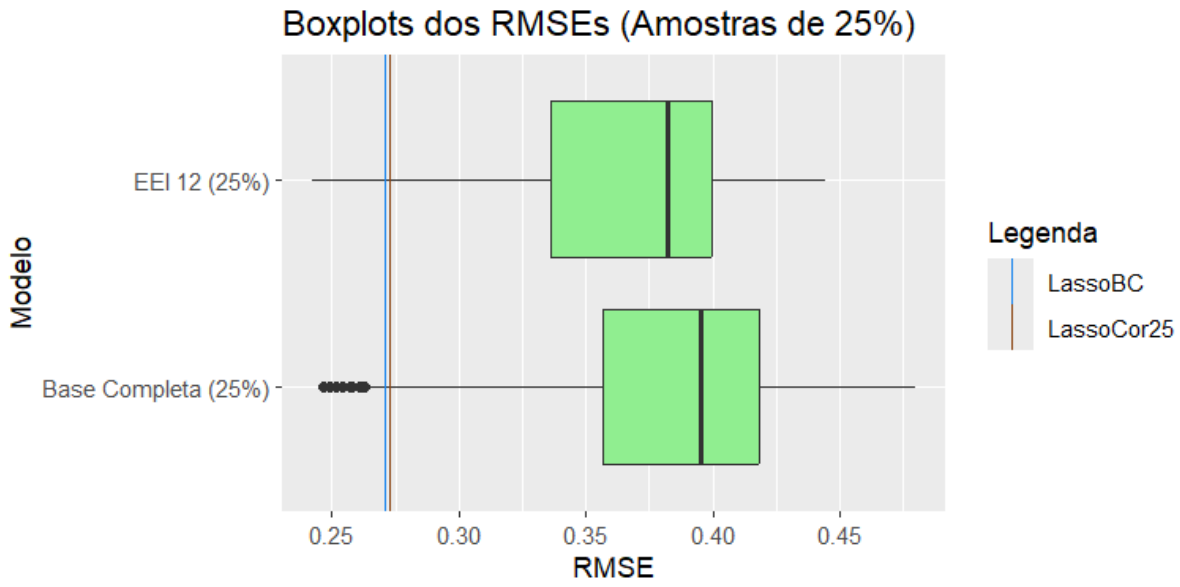

Figura 27: *Boxplots* dos RMSES (amostras de 25% das variáveis).

O formato da distribuição continua bimodal. Entretanto, o aumento da amostra fez com que a concentração à esquerda crescesse. A média e mediana dos RMSEs do lasso que utilizou o agrupamento por GMM na seleção de variáveis está um pouco abaixo do que selecionou variáveis aleatoriamente da base completa.

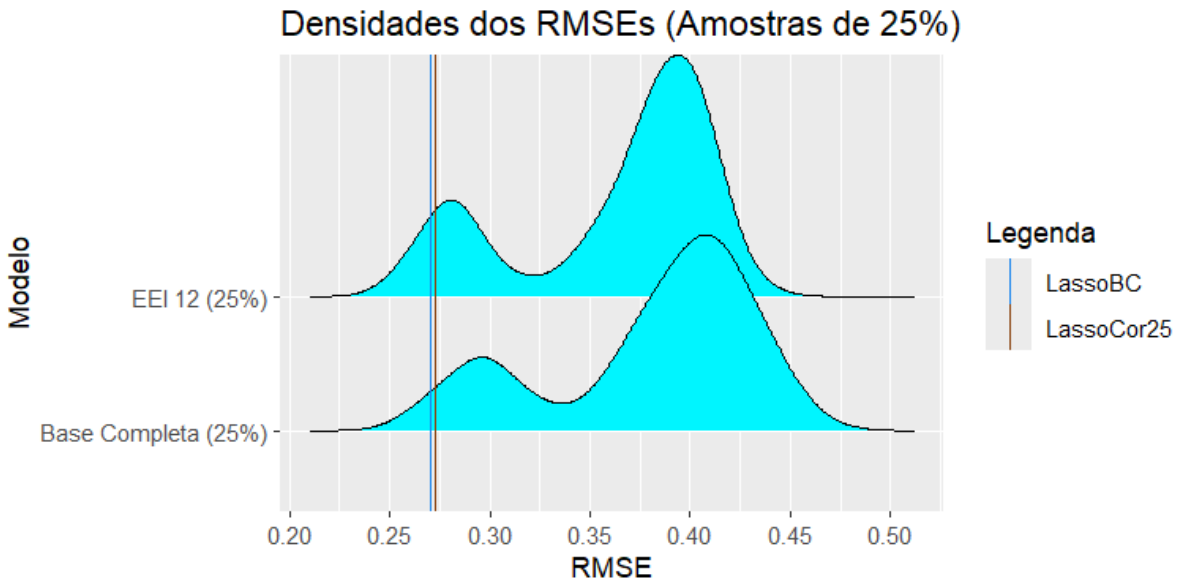

Figura 28: Densidades dos RMSEs (amostra de 25% das variáveis).

4.2.6 Lasso com Reamostragem de 50% das Variáveis

Com a amostra de 50% das variáveis, o tempo foi 452 segundos com a amostra do GMM e 477 segundos com a AAS da base completa. Com essa proporção de amostra, houve uma mudança no comportamento dos RMSEs fora da amostra. As Tabelas 28 e 29 contém as medidas de posição e variabilidade.

Modelo	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
EEI 12 (50%)	0,235	0,275	0,331	0,322	0,366	0,416
Base Completa (50%)	0,244	0,282	0,338	0,330	0,377	0,438

Tabela 28: Medidas de posição dos RMSEs (amostras de 50% das variáveis).

Modelo	Desvio Padrão	Coeficiente de Variação
EEI 12 (50%)	0,048	15,066%
Base Completa (50%)	0,050	15,272%

Tabela 29: Medidas de dispersão dos RMSEs (amostras de 50% das variáveis).

As medidas-resumo indicam uma queda significativa na média dos RMSEs, quando comparados com os valores obtidos para as amostras menores de preditores. A distribuição dos RMSEs pode ser visualizada nas Figuras 29 e 30.

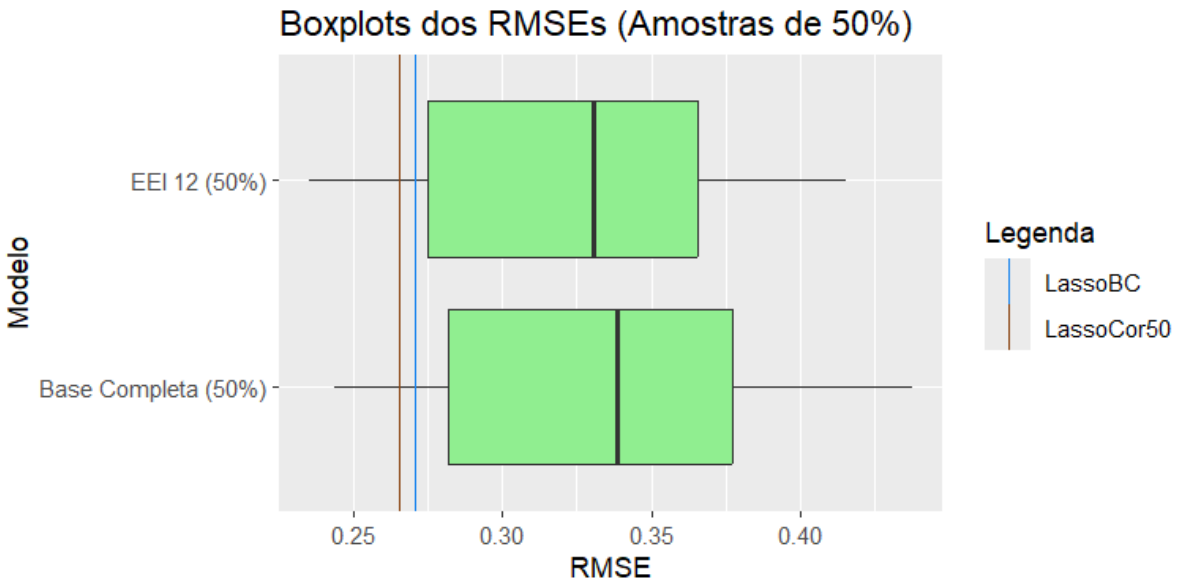

Figura 29: *Boxplots* dos RMSES (amostras de 50% das variáveis).

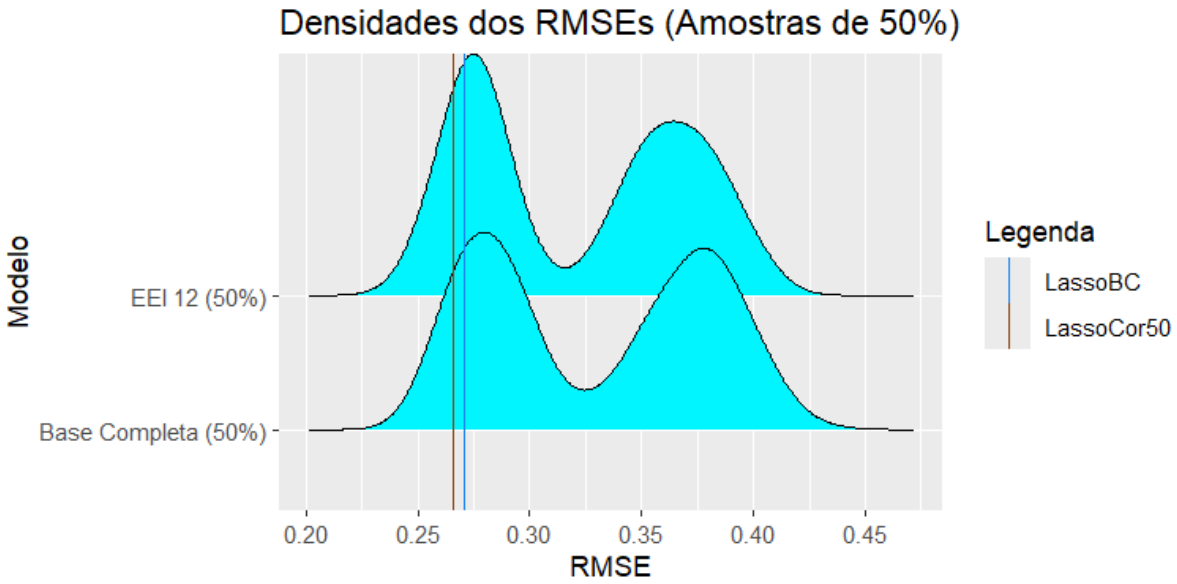

Figura 30: Densidades dos RMSEs (amostras de 50% das variáveis).

As Figuras 25 e 26 mostram que houve uma espécie de inversão entre as duas modas da distribuição dos RMSEs. Com a amostra de 50%, a concentração à esquerda passa a ser maior que a concentração à direita. Esse resultado indica uma melhora da capacidade preditiva com o aumento da amostra de preditores.

4.2.7 Lasso com Reamostragem de 75% das Variáveis

O tempo com a maior amostra de *features* foi de 647 segundos utilizando o GMM e 664 segundos para AAS de variáveis preditoras da base completa. Os resultados descritivos estão nas Tabelas 30 e 31.

Modelo	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
EEI 12 (75%)	0,247	0,268	0,276	0,293	0,334	0,390
Base Completa (75%)	0,247	0,270	0,279	0,296	0,333	0,404

Tabela 30: Medidas de posição dos RMSEs (amostras de 75% das variáveis).

Modelo	Desvio Padrão	Coeficiente de Variação
EEI 12 (75%)	0,037	12,614%
Base Completa (75%)	0,037	12,481%

Tabela 31: Medidas de dispersão dos RMSEs (amostras de 75% das variáveis).

O erro de previsão médio parece ter diminuído com o aumento da amostra. Os *boxplots* e gráficos de densidade estão na Figura 31 e 32.

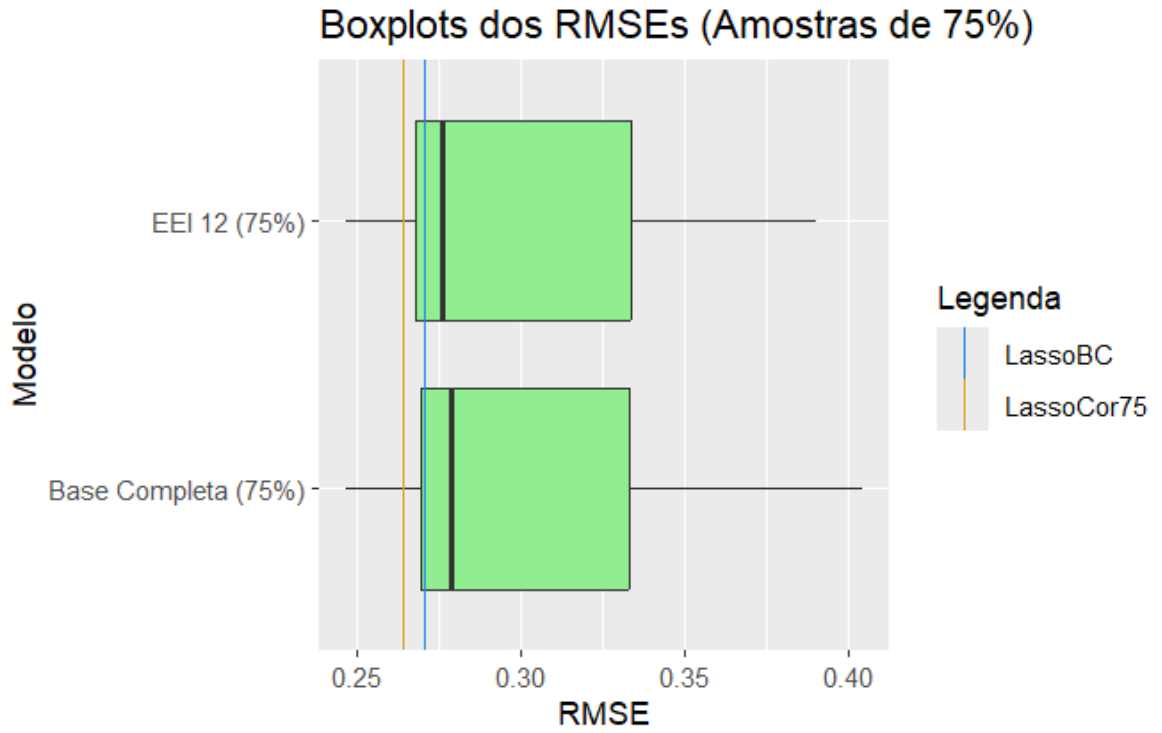

Figura 31: *Boxplots* dos RMSES (amostra de 75% das variáveis).

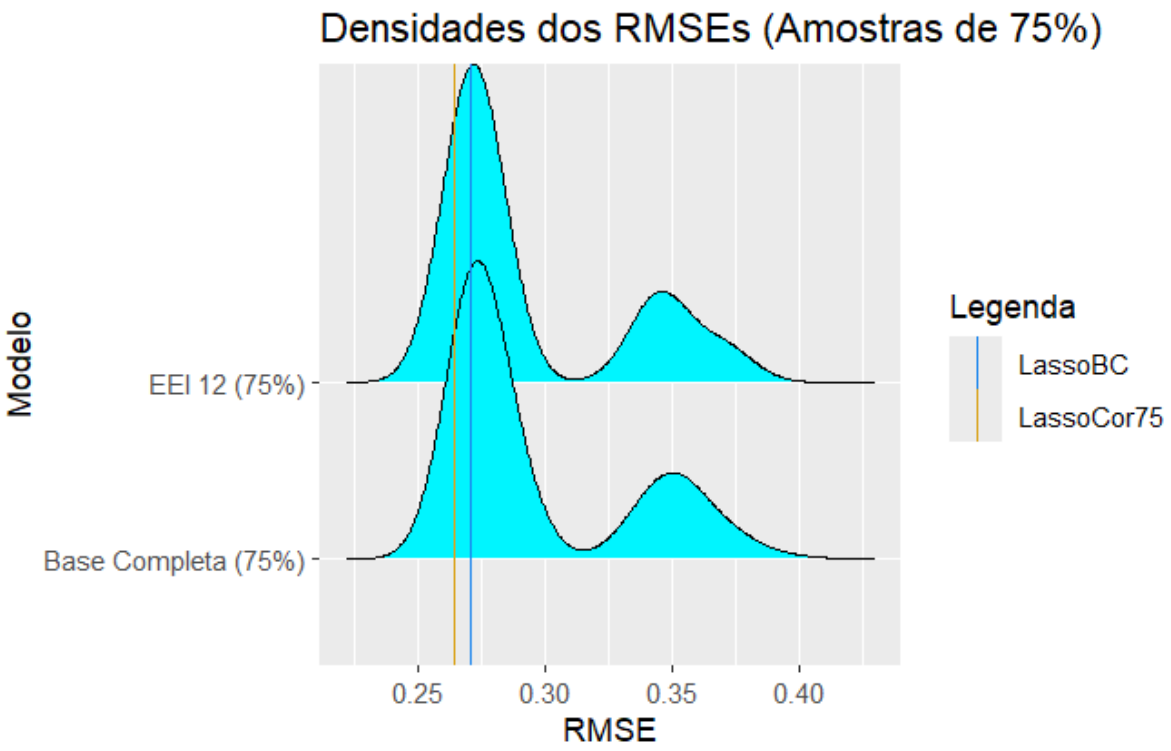

Figura 32: Densidades dos RMSEs (amostra de 75% das variáveis).

Com o maior tamanho da amostra, há uma inversão do formato da distribuição dos RMSEs em relação ao observado para as amostras menores, de 10% e 25%. Com a maior amostra, a distribuição bimodal possui a concentração à esquerda consideravelmente maior que a concentração à direita.

As distribuições dos RMSEs dos modelos que utilizaram o agrupamento do GMM, e os que amostraram variáveis da base completa estão muito semelhantes.

4.2.8 Resultados Gerais com os Dados de Inflação

Os *boxplots* e densidades dos modelos para todas as proporções de amostra de variáveis estão nas Figuras 33 e 34.

É evidente a inversão do formato das distribuições dos RMSEs com o aumento da amostra. A moda à esquerda fica maior com o aumento da amostra, e o contrário acontece com a moda à direita. Portanto, com o aumento das variáveis amostradas, o RMSE se aproxima do obtido pelo lasso com todas as variáveis. As distribuições dos RMSEs dos modelos que utilizaram o agrupamento do GMM e os que selecionaram variáveis da base completa foram semelhantes para todas as proporções de amostras de variáveis explicativas.

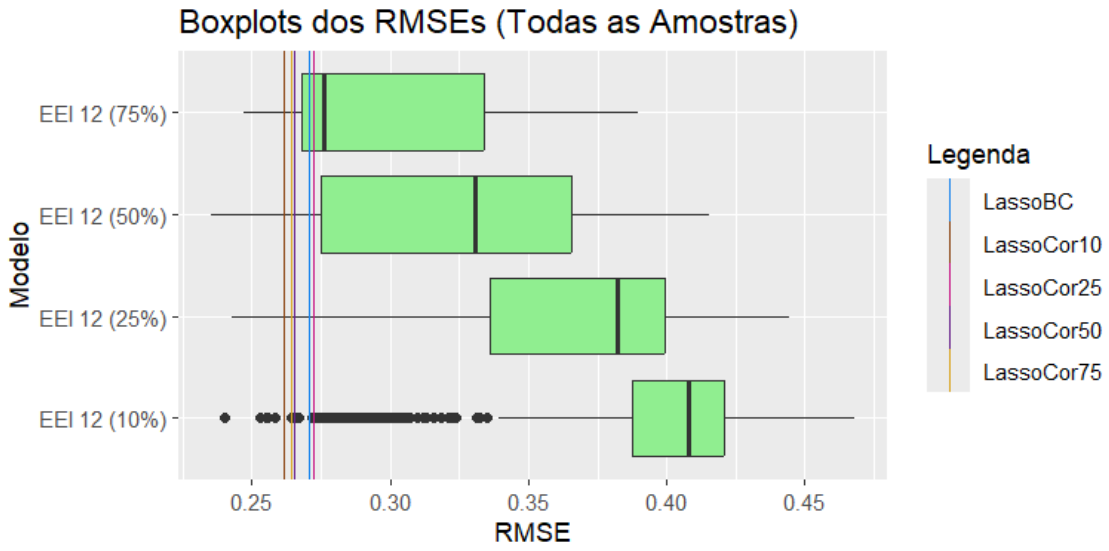

Figura 33: *Boxplots* dos RMSES (todas as amostras das variáveis).

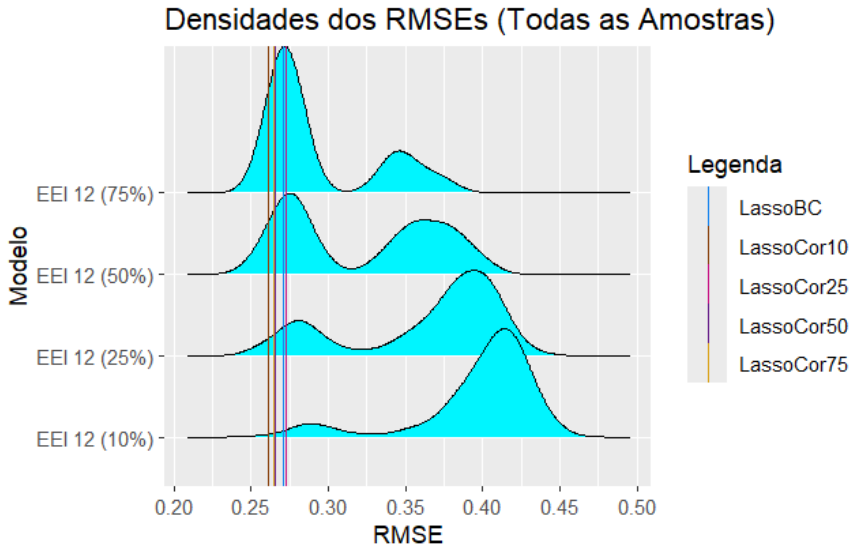

Figura 34: Densidades dos RMSEs (todas as amostras das variáveis).

Os valores λ obtidos pelo processo de reamostragem estão dispostos nos *boxplots* da Figura 35. Assim como nos resultados de atividade econômica, não houve uma grande diferença dos valores da penalização do lasso para os diferentes modelos e seleções de variáveis. Apenas o lasso com uma amostra de 10% da base completa apresentou um valor muito discrepante. Os valores dos λ ficaram em torno de 0,015 para todos os modelos, independente da forma de seleção de variáveis.

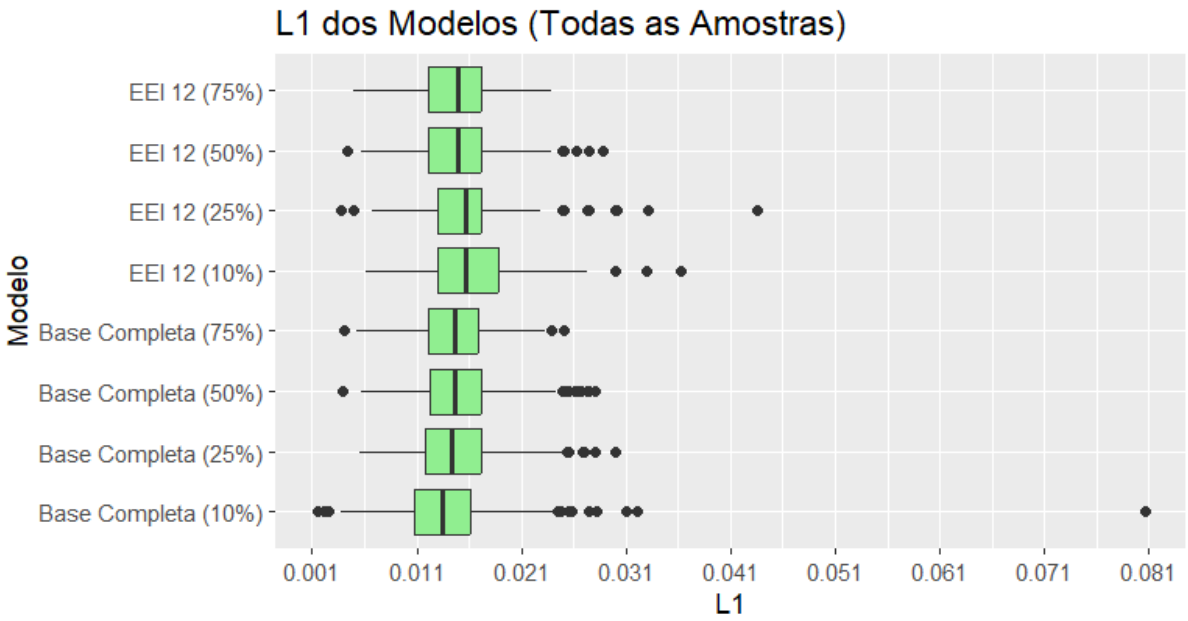

Figura 35: λ dos modelos (todas as amostras das variáveis).

5 Conclusão

Nos resultados do banco de dados de atividade econômica, houve uma aproximação na média dos RMSEs a partir da proporção de amostra de 25%. Isso também ocorreu tanto com a proporção de variáveis mais correlacionadas com a resposta, como com a seleção de variáveis aleatoriamente da base completa. Apesar dos erros de previsão se aproximarem na média a partir de amostras de 25%, amostras maiores aumentam a estabilidade dos erros de previsão. Portanto, com a base do IBC-Br, o algoritmo proposto não foi capaz de reduzir a dimensão dos dados, visto que foi necessário muitas variáveis para obtenção de resultados mais estáveis do RMSE. Também é possível dizer que a utilização do agrupamento das *features* não parece ter tido efeito na estabilidade dos RMSEs, dado que os resultados ficaram muito parecidos com a seleção aleatória de variáveis do banco de dados completo.

Na base da inflação, por outro lado, os modelos que utilizaram algum tipo de procedimento de amostragem de variáveis obtiveram RMSEs consistentemente mais altos que os dos modelos feitos para comparação. A distribuição dos erros de previsão foi bimodal, com duas modas bem aparentes. Assim como ocorreu nos dados de atividade econômica, selecionar os preditores por meio das componentes da mistura de normais não parece ter surtido efeito na estabilidade dos RMSEs. Entretanto, ao levar em conta que a utilização das variáveis dos *clusters* menores não melhorou as previsões dos modelos, conclui-se que essas *features* são ruído e podem ser descartadas.

Ficou evidente que o resultado da aplicação do algoritmo de seleção de variáveis por meio do GMM é determinado pelas particularidades de cada conjunto de dados. As bases de dados apresentaram padrões diferentes de fatores importantes, como valores extremos, estrutura de agrupamento e correlação entre variáveis. Até o tempo de processamento da reamostragem foi diferente, com os modelos que utilizaram dados de inflação demorando um pouco mais. Da mesma forma, os resultados dos erros de previsão também apresentaram padrões diferentes.

No caso dos dados de atividade econômica, os modelos que utilizaram a seleção de variáveis dos grupos do GMM obtiveram RMSEs próximos dos lasso com todas as variáveis, ou com os preditores mais correlacionados. Entretanto, a variabilidade dos RMSEs obtidos por reamostragem faz com que a técnica possa gerar resultados de previsão ruins. Portanto, utilizar a base de dados completa, ou algumas das *features* mais correlacionadas, é mais recomendado do que selecionar preditores do GMM, pois o erro de previsão é próximo e não há variabilidade.

Para os dados dos índices de preço, o problema da aplicação do algoritmo proposto para seleção de variáveis foi pior. Não apenas há variabilidade dos RMSEs, como também a média dos erros foi pior do que para os modelos com a base completa e os preditores mais correlacionados. No caso de amostras menores de variáveis preditoras, a diferença dos RMSEs obtidos entre as técnicas testadas foi ainda mais elevada, como é possível notar pelas Figuras 33 e 34. De forma geral, a seleção aleatória de variáveis gerou modelos com previsões piores e com alta variabilidade, tanto ao utilizar os grupos do GMM, quanto pela AAS da base completa.

O presente trabalho utilizou a técnica de seleção de *features* por meio das componentes do GMM para reduzir a dimensão dos dados. Entretanto, os dados utilizados no trabalho não apresentaram grupos bem definidos de variáveis preditoras. Ao agrupar os preditores pela mistura de normais, os modelos resultantes apresentaram dois problemas: poucas componentes ou componentes com poucas variáveis. Desse modo, a técnica não atingiu resultados estáveis nos dados utilizados.

Levando em conta a área de aplicação utilizada no trabalho, é possível que a estabilidade preditiva do algoritmo tenha sido prejudicada pela instabilidade macroeconômica dos últimos anos. Dessa forma, a utilização de mais observações dos dados e a mudança de período da janela de fora da amostra podem ser medidas que melhorem o funcionamento do algoritmo. Com mais dados para ajustar os modelos e com as previsões fora da amostra não coincidindo com o período da pandemia de Covid-19, os resultados poderiam ser mais estáveis.

A utilização de um agrupamento de preditores que não seja baseado em modelos estatísticos, e sim nos aspectos econômicos das variáveis também poderia melhorar o funcionamento do algoritmo. Para os dados de atividade econômica, um exemplo seria agrupar as variáveis entre as de serviços, indústrias e agropecuária. Para a base de inflação, uma opção seria utilizar a hierarquia de itens e subitens para fazer um agrupamento de variáveis.

Para trabalhos futuros, também seria pertinente testar o algoritmo com bases de dados que possuam agrupamento mais bem definidos de preditores, com mais grupos contendo mais *features*. Possivelmente, com grupos de preditores mais homogêneos internamente e heterogêneos externamente, os RMSEs dos modelos preditivos podem ser mais estáveis ao utilizar a seleção aleatória de variáveis de cada grupo. Nesse caso, a técnica obteria melhores resultados para reduzir a dimensão dos dados e lidar com problemas comuns de modelagem preditiva, como o *overfitting* e a multicolinearidade.

Referências

ARAUJO, G. S.; GAGLIANONE, W. P. *Machine Learning Methods for Inflation Forecasting in Brazil: new contenders versus classical models*. [S.l.], 2022. Disponível em: ⟨https://ideas.repec.org/p/bcb/wpaper/561.html⟩.

BABII, A.; GHYSELS, E.; STRIAUKAS, J. Machine Learning Time Series Regressions With an Application to Nowcasting. *Journal of Business & Economic Statistics*, v. 40, n. 3, p. 1094–1106, June 2022. Disponível em: ⟨https://ideas.repec.org/a/taf/jnlbes/v40y2022i3p1094-1106.html⟩.

BAFFIGI, A.; GOLINELLI, R.; PARIGI, G. Bridge models to forecast the euro area GDP. *International Journal of Forecasting*, v. 20, n. 3, p. 447–460, 2004. Disponível em: ⟨https://ideas.repec.org/a/eee/intfor/v20y2004i3p447-460.html⟩.

BARLAS, A. B. et al. Big data financial transactions and GDP nowcasting: The case of Turkey. *Journal of Forecasting*, v. 43, n. 2, p. 227–248, March 2024. Disponível em: ⟨https://ideas.repec.org/a/wly/jforec/v43y2024i2p227-248.html⟩.

BELLMAN, R.; CORPORATION, R.; COLLECTION, K. M. R. *Dynamic Programming*. Princeton University Press, 1957. (Rand Corporation research study). ISBN 9780691079516. Disponível em: ⟨https://books.google.com.br/books?id=wdtoPwAACAAJ⟩.

BERGER, T.; MORLEY, J.; WONG, B. Nowcasting the output gap. *Journal of Econometrics*, v. 232, n. 1, p. 18–34, 2023. Disponível em: ⟨https://ideas.repec.org/a/eee/econom/v232y2023i1p18-34.html⟩.

BOLFARINE, H.; BUSSAB, W. de O. *Elementos de Amostragem*. Blucher, 2005. ISBN 9788521214991. Disponível em: ⟨https://books.google.com.br/books?id=okniDwAAQBAJ⟩.

BOLIVAR, O. Gdp nowcasting: A machine learning and remote sensing data-based approach for bolivia. *Latin American Journal of Central Banking*, v. 5, n. 3, p. 100126, 2024. ISSN 2666-1438. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S2666143824000085⟩.

BOUVEYRON, C. et al. *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge University Press, 2019. (Cambridge Series in Statistical and Probabilistic Mathematics). ISBN 9781108494205. Disponível em: ⟨https://books.google.com.br/books?id=ldGoDwAAQBAJ⟩.

BROWNING, K. Review lecture: Local weather forecasting. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 371, p. 179–211, 06 1980.

CASELLA, G.; BERGER, R. *Statistical Inference*. Thomson Learning, 2002. (Duxbury advanced series in statistics and decision sciences). ISBN 9780534243128. Disponível em: ⟨https://books.google.com.br/books?id=0x_vAAAAMAAJ⟩.

CELEUX, G.; GOVAERT, G. Gaussian parsimonious clustering models. *Pattern Recognition*, v. 28, n. 5, p. 781–793, 1995. ISSN 0031-3203. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/0031320394001256⟩.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, v. 39, p. 1–38, 1977. Disponível em: ⟨http://web.mit.edu/6.435/www/Dempster77.pdf⟩.

EFRON, B.; HASTIE, T. *Computer Age Statistical Inference*. Cambridge University Press, 2016. (Institute of Mathematical Statistics Monographs). ISBN 9781107149892. Disponível em: ⟨https://books.google.com.br/books?id=Sj1yDAAAQBAJ⟩.

EVANS, M. D. D. Where Are We Now? Real-Time Estimates of the Macroeconomy. *International Journal of Central Banking*, v. 1, n. 2, September 2005. Disponível em: ⟨https://ideas.repec.org/a/ijc/ijcjou/y2005q3a4.html⟩.

FORNARO, P.; LUOMARANTA, H. Nowcasting Finnish real economic activity: a machine learning approach. *Empirical Economics*, v. 58, n. 1, p. 55–71, January 2020. Disponível em: ⟨https://ideas.repec.org/a/spr/empeco/v58y2020i1d10.1007_s00181-019-01809-y.html⟩.

GHOSH, S.; RANJAN, A. A Machine Learning Approach To Gdp Nowcasting: An Emerging Market Experience. *Bulletin of Monetary Economics and Banking*, v. 26, n. Special I, p. 33–54, 2023. Disponível em: ⟨https://ideas.repec.org/a/idn/journl/v26y2023ispdp33-54.html⟩.

GIANNONE, D.; REICHLIN, L.; BAńBURA, M. *Nowcasting*. [S.l.], 2010. Disponível em: ⟨https://ideas.repec.org/p/ecb/ecbwps/20101275.html⟩.

GIANNONE, D. et al. *Now-casting and the real-time data flow*. [S.l.], 2013. Disponível em: ⟨https://ideas.repec.org/p/ecb/ecbwps/20131564.html⟩.

GIANNONE, D.; REICHLIN, L.; SIMONELLI, S. Nowcasting euro area economic activity in real time: The role of confidence indicators. *National Institute Economic Review*, v. 210, 12 2009.

GIANNONE, D.; REICHLIN, L.; SMALL, D. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, v. 55, n. 4, p. 665–676, 2008. ISSN 0304-3932. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0304393208000652⟩.

GLAESER, E. L.; KIM, H.; LUCA, M. Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change. *AEA Papers and Proceedings*, v. 108, p. 77–82, May 2018. Disponível em: ⟨https://ideas.repec.org/a/aea/apandp/v108y2018p77-82.html⟩.

GRAMACKI, A. *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Springer International Publishing, 2019. (Studies in Big Data). ISBN 9783319890944. Disponível em: ⟨https://books.google.com.br/books?id=slc_wAEACAAJ⟩.

GRYBAUSKAS, A. et al. Nowcasting Unemployment Using Neural Networks and Multi-Dimensional Google Trends Data. *Economies*, v. 11, n. 5, p. 1–23, April 2023. Disponível em: ⟨https://ideas.repec.org/a/gam/jecomi/v11y2023i5p130-d1132215.html⟩.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer New

York, 2009. (Springer Series in Statistics). ISBN 9780387848587. Disponível em: ⟨https://books.google.com.br/books?id=tVIjmNS3Ob8C⟩.

HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015. (ISSN). ISBN 9781498712170. Disponível em: ⟨https://books.google.com.br/books?id=f-A_CQAAQBAJ⟩.

HAYASHI, F.; TACHI, Y. Nowcasting Japan's GDP. *Empirical Economics*, v. 64, n. 4, p. 1699–1735, April 2023. Disponível em: ⟨https://ideas.repec.org/a/spr/empeco/v64y2023i4d10.1007_s00181-022-02301-w.html⟩.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, Taylor & Francis Group, v. 12, n. 1, p. 55–67, 1970. ISSN 0040-1706.

HYNDMAN, R.; ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. OTexts, 2021. ISBN 9780987507136. Disponível em: ⟨https://books.google.com.br/books?id=gZB-zgEACAAJ⟩.

JAMES, G. et al. *An Introduction to Statistical Learning: with Applications in R*. Springer US, 2021. (Springer Texts in Statistics). ISBN 9781071614181. Disponível em: ⟨https://books.google.com.br/books?id=5dQ6EAAAQBAJ⟩.

KASS, R. E.; RAFTERY, A. E. Bayes factors. *Journal of the American Statistical Association*, Taylor & Francis, v. 90, n. 430, p. 773–795, 1995. Disponível em: ⟨https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572⟩.

KITCHEN, J.; MONACO, R. The u.s. treasury staff's real-time gdp forecast system. *Business Economics*, v. 38, 01 2003.

KUTNER, M.; NACHTSHEIM, C.; NETER, J. *Applied Linear Regression Models*. McGraw-Hill Higher Education, 2003. (The McGraw-Hill/Irwin Series Operations and Decision Sciences). ISBN 9780072955675. Disponível em: ⟨https://books.google.com.br/books?id=0nAMAAAACAAJ⟩.

KUZIN, V.; MARCELLINO, M.; SCHUMACHER, C. MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting*, v. 27, n. 2, p. 529–542, April 2011. Disponível em: ⟨https://ideas.repec.org/a/eee/intfor/v27yi2p529-542.html⟩.

LUCIANI, M. et al. Nowcasting Indonesia. *Empirical Economics*, v. 55, n. 2, p. 597–619, September 2018. Disponível em: ⟨https://ideas.repec.org/a/spr/empeco/v55y2018i2d10.1007_s00181-017-1288-4.html⟩.

LUKAUSKAS, M. et al. Economic Activity Forecasting Based on the Sentiment Analysis of News. *Mathematics*, v. 10, n. 19, p. 1–22, September 2022. Disponível em: ⟨https://ideas.repec.org/a/gam/jmathe/v10y2022i19p3461-d922541.html⟩.

MAHLER, D. G.; AGUILAR, R. A. C.; NEWHOUSE, D. L. *Nowcasting Global Poverty*. [S.l.], 2021. Disponível em: ⟨https://ideas.repec.org/p/wbk/wbrwps/9860.html⟩.

MORETTIN, P.; BUSSAB, W. *ESTATÍSTICA BÁSICA*. Editora Saraiva, 2017. ISBN 9788502207172. Disponível em: ⟨https://books.google.com.br/books?id=vDhnDwAAQBAJ⟩.

Schwarz, G. Estimating the Dimension of a Model. *Annals of Statistics*, v. 6, n. 2, p. 461–464, jul. 1978.

SCRUCCA, L. et al. *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. CRC Press, 2023. (Chapman & Hall/CRC The R Series). ISBN 9781000868340. Disponível em: ⟨https://books.google.com.br/books?id=APm0EAAAQBAJ⟩.

TASHMAN, L. Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, v. 16, p. 437–450, 10 2000.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the royal statistical society series b-methodological*, v. 58, p. 267–288, 1996. Disponível em: ⟨https://api.semanticscholar.org/CorpusID:16162039⟩.

WATANABE, M.; YAMAGUCHI, K. *The EM Algorithm and Related Statistical Models*. CRC Press, 2003. (Statistics: A Series of Textbooks and Monographs). ISBN 9780203913055. Disponível em: ⟨https://books.google.com.br/books?id=LVQhHgKXfrUC⟩.

WILCOX, R. *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier Science, 2021. (Statistical Modeling and Decision Science). ISBN 9780128200988. Disponível em: ⟨https://books.google.com.br/books?id=HH1bzgEACAAJ⟩.

WOBCKE, W. et al. Nowcasting for hunger relief: a study of promise and perils. *Information Technology for Development*, v. 29, p. 1–21, 08 2022.

ZHANG, Q.; NI, H.; XU, H. Nowcasting Chinese GDP in a data-rich environment: Lessons from machine learning algorithms. *Economic Modelling*, v. 122, n. C, 2023. Disponível em: ⟨https://ideas.repec.org/a/eee/ecmode/v122y2023ics0264999323000160.html⟩.

ZHENG, T. et al. Words or numbers? macroeconomic nowcasting with textual and macroeconomic data. *International Journal of Forecasting*, v. 40, n. 2, p. 746–761, 2024. ISSN 0169-2070. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S016920702300050X⟩.

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, 03 2005. ISSN 1369-7412. Disponível em: ⟨https://doi.org/10.1111/j.1467-9868.2005.00503.x⟩.

Apêndice

A implementação computacional do trabalho foi feita na linguagem R. Para o lasso e os modelos de mistura de normais, foram utilizados os pacotes `glmnet` e `mclust`, respectivamente. Na etapa que requer reamostragem, o procedimento foi feito em paralelo para otimização do tempo de processamento. Para a paralelização, os pacotes `foreach`, `doParallel` e `doRNG` foram utilizadas. Para os gráficos, foram utilizados os pacotes `ggplot2` e `ggridges`. O pacote `Metrics` foi empregado para o cálculo do RMSE. Para cronometrar o tempo gasto nos processos, foi utilizado o pacote `tictoc`. Adicionalmente, os pacotes `dplyr`, `tidyr`, `purrr`, `stingr` e `reshape2` auxiliaram no tratamento e manipulação dos dados. A função *tsoutliers* do `forecast` foi utilizada para tratar os valores extremos na base de dados de atividade econômica.

Os códigos em R para reproduzir os resultados desse trabalho podem ser obtidos com o autor. Para obtenção dos códigos e informações adicionais, o *e-mail* do autor é rafa.costa.ramos@gmail.com.