ROGERSON ALEXANDRE MARTINS

PREVISÃO DE SÉRIES TEMPORAIS COM MÁQUINAS DE SUPORTE VETORIAL

LAVRAS – MG
2023

ROGERSON ALEXANDRE MARTINS

PREVISÃO DE SÉRIES TEMPORAIS COM MÁQUINAS DE SUPORTE VETORIAL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Prof. Dr. Paulo Henrique Sales Guimarães
Orientador

LAVRAS - MG
2023

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Martins, Rogerson Alexandre.
Previsão de Séries Temporais com Máquinas de Suporte Vetorial / Rogerson Alexandre Martins. - 2023.
63 p.: il.

Orientador(a): Paulo Henrique Sales Guimarães.

Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2023.
Bibliografia.

1. Análise de Componentes Principais. 2. Análise de Componentes Independentes. 3. Análise Técnica. I. Guimarães, Paulo Henrique Sales. II. Título.

ROGERSON ALEXANDRE MARTINS

PREVISÃO DE SÉRIES TEMPORAIS COM MÁQUINAS DE SUPORTE VETORIAL

TIME SERIES FORECAST WITH VECTOR SUPPORT MACHINES

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 16 de março de 2023.

Profa. Dra. Thelma Sáfadi UFLA
Prof. Dr. Geraldo Magela da Cruz Pereira UFLA
Prof. Dr. Tiago Martins Pereira UFOP

Prof. Dr. Paulo Henrique Sales Guimarães
Orientador

LAVRAS - MG
2023

Dedico este trabalho a Deus, à minha família, aos bons amigos e àqueles que veem da oportunidade de conhecimento o traço digno de se construir sua trajetória. Obrigado, por caminharmos juntos.

AGRADECIMENTOS

Agradecer é um ato de reconhecimento, de retribuição pelo muito que se fez a quem precisava, ou se dispôs a receber/aprender. Deixo então, com grande satisfação minha gratidão: à UFLA pela oportunidade de estudo público de qualidade, aos excelentes pro- ssionais do Departamento de Estatística que me moldaram até aqui, ao Prof.º Doutor Paulo Henrique Sales Guimarães pela oportunidade singular, pela amizade e confiança de trabalho, aos avaliadores da banca de dissertação, à FAPEMIG pelo incentivo à pesquisa, à minha mãe: Maria Mágda Martins pela força e ousadia de caminhada, aos meus irmãos: Viviane, Victor, Tawane e Lorraine pela inspiração de vida, aos meus familiares em especial: à Vó Quitita, ao Vô Quirinha, Joana D'Arc Martins pela cumplicidade e afabilidade, à todos os bons amigos de tempos distantes e aos de conquista durante o curso, aos amigos-irmãos da República Consulado, ao Doutor Juliano Dantas de Menezes. Enfim, agradeço a Deus pelo dom da vida e zelo constante, pela capacidade de realização do que outrora eu não conhecia ou avistava.

Uma pequenina luz bruxuleante e muda
Como a exatidão, como a firmeza
Como a justiça, brilha
Não na distância
Aqui no meio de nós
Brilha.
(Jorge Cândido de Sena)

RESUMO

A presente dissertação utiliza a técnica de Máquina de Suporte Vetorial (SVM) combinando análise de Componentes Principais e Componentes Independentes na avaliação de séries temporais financeiras. Este assunto é de grande interesse de pesquisadores, investidores e instituições financeiras que buscam compreender o comportamento/influência na tomada de decisão no mercado de preços. Sabe-se que a combinação de análise Componentes Principais e Independentes, conjuntamente com as máquinas de suporte vetorial pode garantir melhores resultados para o contexto. Como resultados, verifica-se que os modelos PCA - SVR, ICA - SV apresentaram melhores acertabilidade quando comparado com modelos comuns, tal como o SVR simplesmente. Os resultados das métricas MAE, MSE, RMSE, R^2 corroboram com os modelos aplicados em questão.

Palavra chave: Análise de Componentes Principais. Análise de Componentes Independentes. Análise Técnica.

ABSTRACT

This dissertation uses the Support Vector Machine (SVM) technique combining Principal Components and Independent Components analysis in the evaluation of financial time series. This subject is of great interest to researchers, investors and financial institutions that seek to understand the behavior/influence on decision-making in the price market. It is known that the combination of Principal and Independent Components analysis, together with vector support machines can guarantee better results for the context. As a result, it appears that the PCA - SVR, ICA - SV models showed better accuracy when compared to common models, such as the SVR simply. The results of the MAE, MSE, RMSE, R^2 metrics corroborate the applied models in question.

keywords: Principal Component Analysis. Independent Component Analysis. Technical Analysis.

LISTA DE FIGURAS

Figura 2.1 – Representação ilustrativa das componentes de tendência e sazonalidade . . 16
Figura 2.2 – Conjunto de treinamento binário e três diferentes hipóteses 18
Figura 2.3 – Na região à direita do hiperplano, em que $f(x) > 0$, temos os pontos do tipo $+1$ representados por triângulos vazios, enquanto que na região à esquerda do hiperplano, em que $f(x) < 0$, temos os pontos do tipo -1 representados por triângulos cheios. 20
Figura 2.4 – A figura mostra alguns dos infinitos hiperplanos que dividem o conjunto de dados. 21
Figura 2.5 – Máquinas de suporte vetorial: à esquerda temos o resultado proporcionado por um *kernel* polinomial de grau 3, e à direita, para o mesmo conjunto de dados, observa-se o resultado da aplicação de um kernel radial 26
Figura 2.6 – Função de perda $\varepsilon-$insensível na linha mais espessa em vermelho. 28

Figura 3.1 – Estrutura de previsão proposta PCA. 46
Figura 3.2 – Estrutura de previsão proposta ICA. 47
Figura 3.3 – Evolução da série mensal de preços do Bradesco no período 02/01/2015 a 01/02/2023. 48
Figura 3.4 – Evolução no tempo das séries temporais dos preços Bradesco com SVR. . 48
Figura 3.5 – Evolução no tempo das séries temporais dos preços Bradesco com PCA - SVR. 49
Figura 3.6 – Evolução no tempo das séries temporais dos preços Bradesco com ICA - SVR. 50
Figura 3.7 – Evolução da série mensal de preços da Vale no período 02/01/2015 a 01/02/2023. 51
Figura 3.8 – Evolução no tempo das séries temporais dos preços Vale com SVR. 51
Figura 3.9 – Evolução no tempo das séries temporais dos preços Vale com PCA - SVR. . 52
Figura 3.10 – Evolução no tempo das séries temporais dos preços Vale com ICA - SVR. . 53

LISTA DE TABELAS

Tabela 2.1 – Organização de um conjunto de dados com n tratamentos, p variáveis e k componentes 35
Tabela 2.2 – Escores do primeiro componente para os *n* tratamentos 35
Tabela 3.1 – Resultados da pesquisa de grade para parâmetros de Kernel Radial - Bradesco. 50
Tabela 3.2 – Resultados da pesquisa de grade para parâmetros de Kernel Radial - Vale. . 53

SUMÁRIO

1 INTRODUÇÃO 12
2 REFERENCIAL TEÓRICO 14
2.1 Séries temporais 15
2.1.1 Tendência e Sazonalidade 16
2.2 Teoria de aprendizado estatístico 17
2.3 Máquina de suporte vetorial 19
2.3.1 Para fins de classificação binária: caso linearmente separável 19
2.3.2 Para fins de classificação binária: caso não linearmente separável 22
2.3.3 Para fins de classificação binária: caso não linear 25
2.3.4 Para fins de regressão 26
2.4 Análise de Componentes Principais 30
2.4.1 Matriz de dados X e de covariância S 31
2.4.2 Contribuição de cada componente principal 34
2.4.3 Interpretação de cada componente 34
2.4.4 Escores dos componentes principais 35
2.5 Análise de Componentes Independentes 35
2.5.1 Definição 36
2.5.2 Princípios básicos 37
2.5.3 Princípios para séries temporais 38
2.5.4 Hipótese de autocovariâncias diferentes 38
2.5.5 Hipótese de variâncias não estacionárias 39
2.5.6 Unificação dos princípios de separação 40
3 MATERIAIS E MÉTODOS 42
3.1 Estrutura e conceitos 42
4 CONCLUSÃO 54
REFERÊNCIAS 55
APENDICE A – 58

1 INTRODUÇÃO

Segundo Morettin (2017), o esforço de previsão de séries temporais financeiras ganhou extrema atenção de investidores individuais e institucionais, uma vez que a previsão pode influenciar a decisão por trás do investimento.

Classificar e/ou prever valores futuros de séries temporais é um assunto que abrange vários campos do conhecimento, como Economia, Medicina, Meteorologia, Agronomia entre outros. Propostas utilizando modelos estatísticos lineares e não lineares, ou ainda inteligência artificial, já foram formuladas, e o desenvolvimento de novas metodologias continua em ascensão.

Uma metodologia que podemos utilizar para predizer valores futuros de Séries Temporais consiste na Máquina de Suporte Vetorial. Esta consiste de uma técnica de Aprendizado Supervisionado de Máquinas, em que pode ser utilizada tanto para classificação, quanto para regressão.

Existem diversos métodos de análisar o comportamento de um ativo, sendo dois dos mais clássicos e populares a análise técnica, que é um conjunto de métodos matemáticos que se utilizam do histórico de preços e volume negociado para prever tendências futuras e a análise fundamentalista que se baseia nos dados econômicos das empresas ou setores em si.

A aplicação bem sucedida em vários problemas de séries temporais incentivou sua adaptação na previsão de séries temporais financeiras. Os pesquisadores usam várias abordagens de aprendizado de máquina e inteligência artificial para analisar indicadores técnicos para estudar tendências ou previsões.

A importância de fazer previsões está associada com tomar decisões. Empresas e governos fazem orçamentos, alocam recursos ou traçam políticas públicas, com base na expectiva de comportamento futuro de variáveis que afetam suas atividades, tais como consumo, inflação, arrecadação ou incidência de doenças, por exemplo.

Há vários métodos de se fazer previsões, alguns mais intuitivos e de natureza subjetiva; outros mais objetivos, com base matemática e estatística. O fato de não haver um método único e ideal de previsão, aplicável a todas as situações, deixa em aberto um amplo espaço para a pesquisa científica da aplicabilidade e eficiência de cada técnica.

Propõe-se aqui, a construção de modelagem introduzindo a análise de componentes principais (PCA do inglês *Principal Component Analysis*)) e, na mesma linha a aplicação de análise de componentes independentes (ICA do inglês *Independent Component Analysis*) via indicadores técnicos na cadeia de processamento.

O objetivo geral proposto na dissertação é:

a) Utilizar a metodologia de SVR para predição de séries temporais financeiras, com o auxílio de análise de componentes principais (PCA) e análise de componentes independentes (ICA).

Já como objetivos específicos:

a) Verificar o desempenho da metodologia de SVR na predição de séries temporais;

b) Fazer comparações entre diferentes modelos de aprendizado de máquinas mediante métricas: MAE (*Mean absolut error*), MSE (*Mean squared error*), RMSE (*Root Mean squared error*) e R^2 (*R-squared*).

Este trabalho difere neste objetivo, onde visa prever a tendência de variação do valor utilizando dados históricos, conjunto de atributos da série temporal financeira (tais como indicadores técnicos), demonstrando a capacidade de modelar em certo grau de acurácia o comportamento da tendência de variação do ativo através do SVR, de um conjunto de observações do Banco Bradesco SA (BBDC3.SA) e da Vale SA (VALE3.SA), relacionados à Bolsa de Valores do Brasil - (B3).

2 REFERENCIAL TEÓRICO

Este capítulo tem por objetivo apresentar o referencial teórico da pesquisa e fundamentar as contribuições para previsão de séries temporais com máquina de suporte vetorial, visto que existe uma grande quantidade de trabalhos que fizeram essa utilização e obtiveram bons resultados empíricos nas mais diversas áreas, o que encoraja o aprofundamento na investigação do tema.

No campo de aplicação de SVM em finanças temos o trabalho de Fan e Palaniswami (2001), os quais apresentaram um proposta para formação de portfólios por meio das máquinas de suporte vetorial. Utilizaram para isso, dados contábeis e informações sobre preços das ações das corporações de interesse negociadas na Bolsa de Valores Australiana. Os autores formularam uma proposta para construção de portfólios por meio do SVM, a qual apresentou retornos superiores a outros modelos de análise de mercado. Para cada conjunto de variáveis que compunham cada uma das categorias realizaram um análise de componentes principais e definiram o primeiro componente principal como a variável representativa do grupo de categoria financeira, essa variável foi então armazenada para a formação do portfólio de interesse.

Segundo Fernando e Oliveira (2008), as séries temporais são constituídas por observações autocorrelacionadas e como tal não podem ser permutadas entre si, contudo a suposição em relação à independência entre observações não é necessária para aplicar as técnicas de análise de componentes principais (PCA) do ponto de vista descritivo. A aplicação destas técnicas multivariadas a séries temporais permite realçar alguns resultados e interpretações que sugerem uma conexão com as inter-relações existentes entre as observações, pelo menos em termos empíricos.

O SVM foi utilizado para prever séries climáticas, por exemplo, no artigo de Sharifi e Souri (2015). Nele os autores utilizaram o método para prever um parâmetro bastante importante na área de sistemas climáticos globais, que é o vapor de água precipitável, com ele se pode fazer previsões de formações de nuvens, e previsões no curto prazo de chuva. O método do SVM foi comparado com um híbrido de estimações harmônicas e mínimos quadrados, mais uma vez se mostrando mais eficiente.

Os autores Tay e Cao (2001) utilizaram o SVM para previsões em séries financeiras, com o objetivo de analisar a factibilidade desse método utilizando Redes Neurais. O experimento mostrou um desempenho melhor do SVM em relação ao outro método estudado.

Rubio et al. (2011) estudaram alguns modelos do método de SVM aplicados em modelagem e previsão de séries temporais. Após testar diversas funções *kernel*, argumentou que o método é bastante eficiente para o que se desejou fazer no estudo, e que a função utilizada no *kernel* do SVM é de grande importância e influência na eficiência dos modelos gerados para analise de séries temporais.

Outro trabalho que tratou de investigar a eficácia do SVM em previsões de séries temporais financeiras foi o de Zhao, Ma e Yin (2012). Nesse artigo, comparou-se o método do SVM com outro chamado de *Locality Preserving Projection* (LPP). O método do SVM foi otimizado

via *Particle Swarm Optimization* (PSO), um método computacional que otimiza um problema tentando iterativamente melhorar uma solução candidata no que diz respeito a uma dada medida de qualidade. O SVM, mais uma vez, mostrou-se mais eficiente. No caso desse artigo, o autor argumenta que com a melhoria feita por meio do PSO, a previsão conseguida por ele se mostrou competitiva em relação a outras utilizadas para fins práticos no mercado financeiro.

No trabalho de Campos et al. (2021), foi analisada e feita previsões a partir do histórico de variação de preço da empresa de capital aberto com maior participação relativa no índice Bovespa, a Vale (VALE3). Tendo como objetivo a comparação das principais técnicas de predição para séries temporais no contexto de mercado financeiro foram realizadas análises qualitativas para compreender o estado da arte sobre predição de séries temporais e teorias de previsão nos mercados financeiros. Além disso, foram realizados processos de obtenção, preparação e modelagem para garantir uma padronização dos dados de entrada em cada modelo utilizado. Por fim, foi realizada uma análise comparativa dos resultados dos preditores.

Portanto, assim, pode-se afirmar que a metodologia SVM é bastante utilizada em análise de séries temporais. inclusive em artigos recentes. Os autores supracitados mostraram em seus estudos que o método do SVM é bastante eficaz na previsão e modelagem de séries temporais em diversas áreas, além disso, o método se mostra mais eficiente que diversos outros, como alguns de redes neurais ou outros mais tradicionais, como o de regressões lineares múltiplas.

2.1 Séries temporais

Uma série temporal, também denominada série histórica, é uma sequência de observações obtidas normalmente em intervalos regulares de tempo durante um período específico.

Uma série temporal $T = t_1, \ldots, t_i, t_j, \ldots, t_m$ consiste em um conjunto de m valores ordenados, $m \geq 2$, tal que se $i < j$, t_i ocorre cronologicamente antes que t_j.

Assim, uma série temporal pode ser interpretada em função das subsequências que a compõem, nesse contexto, uma subsequência é definida como:

Uma subsequência $S = t_p, \ldots, t_{p+n-1}$ consiste em um subconjunto contíguo de n valores de T com início na posição p, tal que $2 \leq n \leq m$ e $1 \leq p \leq m - n + 1$.

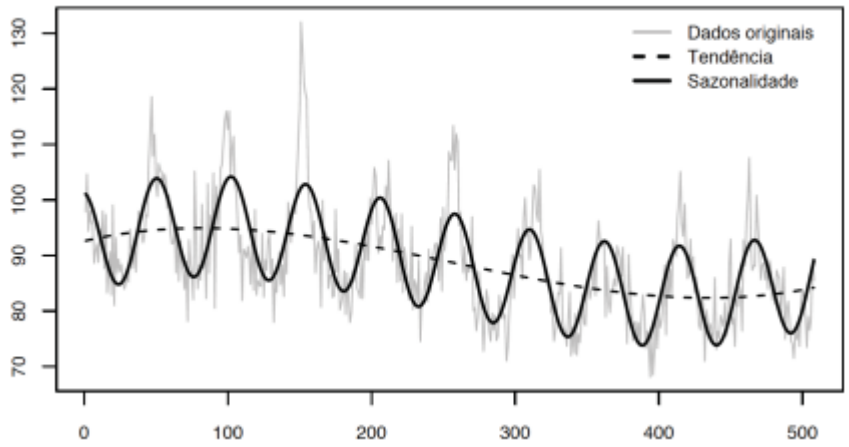
Na compreensão dos eventos que as séries temporais apresentam, a literatura destaca técnicas ou conceitos de suas decomposições dentro de um conjunto finito de componentes independentes. As principais componentes são: tendência, sazonalidade e resíduo (MORETTIN; TOLOI, 2006).

Nesse contexto, cada uma das observações t_i que compõem uma determinada série temporal T, podem estar influenciadas por uma ou mais dessas componentes. Portanto, em grande parte dos problemas da área de séries temporais, não é possível identificar diretamente a atuação dessas componentes na série temporal, de modo que somente podem ser extraídas e compreendidas por meio da aplicação de técnicas específicas de decomposição (CHEN; DAVIS; BROCKWELL, 1996). Uma vez que as componentes sazonal e tendência possuem uma re-

lação forte, e uma delas pode afetar os métodos de análises, é possível separar uma da outra (CHEN; DAVIS; BROCKWELL, 1996); (MORETTIN; TOLOI, 2006).

De um modo resumido são destacadas as definições e fundamentos das componentes conforme é mostrado na Figura 2.1:

Figura 2.1 – Representação ilustrativa das componentes de tendência e sazonalidade



Fonte:Ferrero (2009)

2.1.1 Tendência e Sazonalidade

O movimento dominante em uma série temporal é chamado de tendência. Ela exerce influência nas observações por longos instantes, chega a alterar o nível médio da série. Na Figura 2.1 a tendência é representada pela linha tracejada, nesse caso a série do fenômeno é observada em cor cinza. Ainda nesse contexto, as séries apresentam comportamentos diferentes de tendência onde estão baseados os métodos de identificação da mesma componente (EHLERS, 2007).

A sazonalidade é um comportamento que se repete em diferentes instantes de acordo com alguma característica. Ela apresenta oscilação ao longo da tendência (FERRERO, 2009). A identificação desta componente é um processo importante no que tange à análise de séries temporais, a sua presença permite a descoberta de informações relevantes, enquanto que na sua ausência a remoção pode afetar significativamente outras características da série (MALETZKE, 2009). Na Figura 2.1 a componente é representada por uma linha contínua de cor preta.

Na análise de uma série temporal, primeiramente deseja-se modelar o fenômeno estudado para, a partir daí, descrever o comportamento da série, fazer estimativas e, por último, avaliar quais os fatores que influenciaram o comportamento da série, buscando definir relações de causa e efeito entre duas ou mais séries. Para tanto, há um conjunto de técnicas estatísticas disponíveis que dependem do modelo definido (ou estimado para a série), bem como do tipo de série analisada e do objetivo do trabalho.

Segundo Latorre e Cardoso (2001), para análise de tendências, podem se ajustar modelos de regressão polinomial baseados na série inteira ou em vizinhança de um determinado ponto. Isso também pode ser realizado com funções matemáticas.

Define-se como um fenômeno sazonal aquele que ocorre regularmente em períodos fixos de tempo e, se existir sazonalidade dita determinística na série, podem-se utilizar modelos de regressão que incorporem funções do tipo seno ou cosseno à variável tempo.

Os modelos auto-regressivos formam outra classe de modelos. Na análise do comportamento de uma série histórica livre de tendência e de sazonalidade podem ser utilizados modelos auto-regressivos (AR) ou que incorporem médias móveis (MA). Quando há tendência, utilizam-se os modelos auto-regressivos integrados de médias móveis (ARIMA) e, para incorporar o componente de sazonalidade, utilizam-se os modelos SARIMA.

Por último há os modelos lineares generalizados. Neste grupo de modelos estatísticos, a variável resposta é comumente um processo de contagem e as variáveis independentes são variáveis candidatas a explicar o comportamento da série ao longo do tempo. Estes modelos são indicados quando as variáveis em estudo não têm aderência à distribuição normal, principalmente pelo fato de serem processos de contagem.

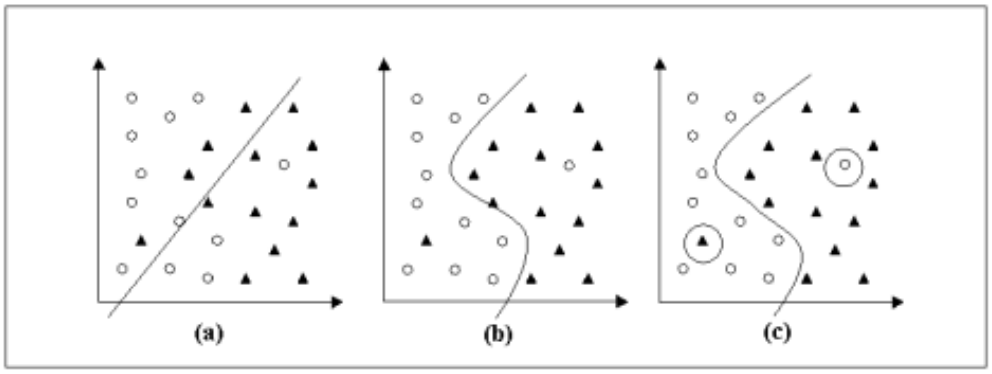
Segundo Sicsu e Dana (2017), estes modelos compõem um grupo de distribuições de probabilidades conhecido como família exponencial de distribuições que englobam diversas funções aditivas, como a regressão linear, de Poisson, logística, log-linear etc. Os modelos aditivos generalizados são uma extensão desta classe de modelos, nos quais cada variável independente analisada não entra no modelo com o seu valor, mas sim, adotando uma função não paramétrica de forma não especificada, estimada a partir de curvas de alisamento.

2.2 Teoria de aprendizado estatístico

Seja f um classificador e F o conjunto de todos os classificadores que um determinado algoritmo de aprendizado de máquina pode gerar. Esse algoritmo, durante o processo de aprendizado, utiliza um conjunto de treinamento T, composto de n pares (x_i, y_i), para gerar um classificador particular $\hat{f} \in F$.

Considere, por exemplo, o conjunto de treinamento da Figura 2.2. O objetivo do processo de aprendizado é encontrar um classificador que separe os dados das classes "círculo"e "triângulo". As funções ou hipóteses consideradas são ilustradas na figura por meio das bordas, também denominadas fronteiras de decisão, traçadas entre as classes.

Figura 2.2 – Conjunto de treinamento binário e três diferentes hipóteses



Fonte: Próprio autor (2023)

Na Figura 2.2(c) tem-se uma hipótese que classifica corretamente todos os exemplos do conjunto de treinamento, incluindo dois possíveis ruídos. Por ser muito específica para o conjunto de treinamento, essa função apresenta elevada suscetibilidade a cometer erros quando confrontada com novos dados. Esse caso representa a ocorrência de um superajustamento do modelo aos dados de treinamento.

Outro classificador poderia desconsiderar pontos pertencentes a classes opostas que estejam muito próximos entre si. A ilustração Figura 2.2(a) representa essa alternativa. A nova hipótese considerada, porém, comete muitos erros, mesmo para casos que podem ser considerados simples. Tem-se assim a ocorrência de um subajustamento, pois o classificador não é capaz de se ajustar mesmo aos exemplos de treinamento.

Um meio termo entre as duas funções descritas é representado na Figura 2.2(b). Esse classificador tem complexidade intermediária e classifica corretamente grande parte dos dados, sem fixar demasiadamente em qualquer ponto individual.

A Teoria de Aprendizado Estatístico estabelece condições matemáticas que auxiliam na escolha de um classificador particular $\hat{f}$ a partir de um conjunto de dados de treinamento. Essas condições levam em consideração o desempenho do classificador no conjunto de treinamento e sua complexidade, com objetivo de obter um bom desempenho também para novos dados do mesmo domínio.

As máquinas de suporte vetorial são embasadas pela teoria de aprendizado estatístico, desenvolvido Chervonenkis e Vapnik (1971). Essa teoria estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu. As SVMs têm sido amplamente utilizados em qualquer tipo de problema de aprendizagem, principalmente em problemas de classificação, mas também em outros problemas como agrupamento (BEN-HUR et al., 2001) ou regressão (SMOLA; SCHÖLKOPF, 2004).

2.3 Máquina de suporte vetorial

Segundo Lorena e Carvalho (2007), uma máquina suporte vetorial, pode ser utilizada para o reconhecimento de padrões M-dimensionais entre duas classes distintas. O classificador SVM utiliza um hiperplano, obtido a partir dos chamados vetores de suporte, para separar o universo M-dimensional em duas regiões, cada uma associada a uma classe.

A máquina de suporte vetorial (SVM) é uma técnica para análise de dados que se baseia na teoria do aprendizado estatístico (CHERVONENKIS; VAPNIK, 1971); (VAPNIK; CHERVONENKIS, 1982); (CORTES; VAPNIK, 1995). Ela propicia aplicações, por exemplo, em análise multivariada, análise de regressão (JUDGE et al., 1988) e análise de séries temporais (MORETTIN; TOLOI, 2006).

Diversos trabalhos na área de estatística têm mostrado resultados interessantes no que se refere ao SVM. O método mostra-se bastante competitivo em relação aos métodos estatísticos em geral. Visto isso, detalharemos os aspectos essenciais sobre o SVM, abordando sua formulação para fins de classificação e de regressão. Além disso, a sua aplicabilidade em regressão permite sua utilização para a obtenção de previsões em análise de séries temporais.

2.3.1 Para fins de classificação binária: caso linearmente separável

Seja $\mathbf{X}$ uma matriz de tamanho $n \times p$ formada por n observações de treinamento em um espaço de dimensão p, ou seja

$$x_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \cdots \\ x_{1p} \end{pmatrix}, ..., x_n = \begin{pmatrix} x_{n1} \\ x_{n2} \\ \cdots \\ x_{np} \end{pmatrix}$$

Considere que a i-ésima linha da matriz $\mathbf{X}$ possa ser classificada em dois tipos, conforme a variável $y_i \in \{-1, +1\}$. Considere ainda um vetor de observações de teste $x^* = (x_1^*, ..., x_p^*)^T$, cujos elementos são separados de acordo com os tipos -1 e $+1$ produzidos pelo classificador obtido com base nos dados de treinamento. Para a construção desse classificador, os dados são separados por meio de um hiperplano de forma que

$$\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} > 0 \quad \text{se} \quad y_i = 1,$$

e

$$\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} < 0 \quad \text{se} \quad y_i = -1.$$

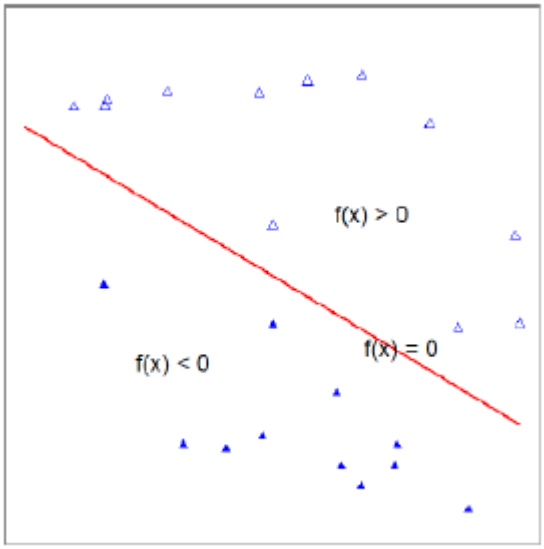
Assim, define-se o hirpeplano

$$f(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

como critério para classificar o tipo da observação de teste x^* com base no sinal da função $f(x^*)$.

Para ilustrar, a Figura 2.3 mostra um hiperplano de um espaço bidimensional, isto é, uma reta $f(x) = 0$ que separa o conjunto de dados em duas partes conforme o sinal de $f(x)$.

Figura 2.3 – Na região à direita do hiperplano, em que $f(x) > 0$, temos os pontos do tipo $+1$ representados por triângulos vazios, enquanto que na região à esquerda do hiperplano, em que $f(x) < 0$, temos os pontos do tipo -1 representados por triângulos cheios.



Fonte: Próprio autor (2023)

A magnitude de $f(x')$ indica o quão distante a observação $x' = (x'_1, ..., x'_p)^T$ está do hiperplano. Para mostrarmos isso, vamos lembrar que se a e b forem dois vetores em

$$\mathbb{R}^p$$

, formando um ângulo θ entre eles conforme ilustra a Figura 2.3(a), então temos que $\langle a, b\rangle = \|a\| \times \|b\| \times \cos\theta$.

$$f(x) \quad = \quad \beta_0 + \sum_{j=1}^{p} \beta_j x_j = \beta_0 + \langle \beta, x\rangle,$$

em que $\beta = (\beta_1, ..., \beta_p)^T$ e $x = (x_1, ..., x_p)^T$. Nesse caso, β é um vetor perpendicular ao hiperplano no ponto x, pois $\langle \beta, x\rangle = 0$, e o vetor do ponto x' ao ponto x é

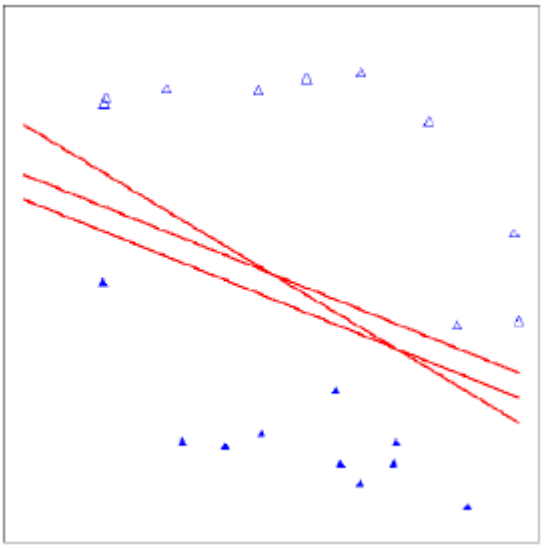
$$w = \begin{pmatrix} x'_1 - x_1 \\ x'_2 - x_2 \\ \cdots \\ x'_p - x_p \end{pmatrix}.$$

Logo, a distância $\|d\|$ do ponto x' ao hiperplano é a projeção escalar do vetor w sobre a normal β, ou seja

$$\begin{aligned} \|d\| &= \|P_{w\beta}\| = \frac{\langle w, \beta \rangle}{\|\beta\|} = \frac{\langle \beta, x' \rangle - \langle \beta, x \rangle}{\|\beta\|} \\ &= \frac{\langle \beta, x' \rangle - \beta_0}{\|\beta\|} = \frac{f(x')}{\|\beta\|}. \end{aligned}$$

A distância de um ponto qualquer x' ao hiperplano indica, de certa forma, uma medida de precisão na classificação feita. Isso porque à medida que x' se afasta do hiperplano, maior será a certeza acerca de sua classificação. Em contraste, a incerteza tende a aumentar à medida que x' se aproxima do hiperplano. Por desejarmos a maior separabilidade possível, tem-se como objetivo maximizar a magnitude de $f(x')$.

Figura 2.4 – A figura mostra alguns dos infinitos hiperplanos que dividem o conjunto de dados.



Fonte: Próprio autor (2023)

Se os dados de treinamento forem linearmente separáveis, existem $M > 0$, β_0 e β tais que

$$y_i(\beta_0 + \langle \beta, x_i \rangle) \geq M,$$

para todo $i = 1,...,n$. Portanto, a margem que separa as duas classes corresponde à soma da distância do ponto x_i' (com $y_i' = +1$) e do ponto x_i'' (com $y_i'' = -1$) mais próximos do hiperplano $f(x) = 0$ que os separa (Figura 2.4). As observações x tais que $y_s(\beta_0 + \langle \beta, x_s \rangle) = M$ são chamadas de suporte vetorial.

Considerando a distância entre os dois pontos de suporte vetorial $x \in \mathbb{R}^p$ (do tipo $y' = +1$) e $x'' \in \mathbb{R}^p$ (do tipo $y'' = -1$) mais próximos do hiperplano que separa o conjunto de dados, desenvolve-se um raciocínio parecido com o da Figura 2.3.

Ou seja, calculamos a distância entre os dois pontos, a multiplicamos por um vetor unitário perpendicular ao hiperplano, e obtemos a projeção cuja medida representará a margem.

Se chamarmos essa distância de $\|D\|$, o melhor hiperplano será o que a maximize, proporcionando a maior separação possível entre os dados de tipos diferentes.

Lembrando que os pontos x de suporte vetorial são tais que $y_s(\beta_0 + \langle \beta, x_s \rangle) = M$, a distância $\|D\|$ (ou margem) que queremos maximizar é dada por

$$\begin{aligned} \|D\| &= (x' - x'')\frac{\beta}{\|\beta\|} = (\langle \beta, x' \rangle - \langle \beta, x'' \rangle)\frac{1}{\|\beta\|} \\ &= [(M - \beta_0) - (-M - \beta_0)]\frac{1}{\|\beta\|} = \frac{2M}{\|\beta\|}. \end{aligned}$$

Observe que maximizar $\frac{2M}{\|\beta\|}$ é equivalente a minimizar seu inverso multiplicativo $\frac{\|\beta\|}{2M}$. Além disso, Vapnik (1995) definiu que se $M = 1$ (hiperplano canônico) existe outra forma de se encontrar o hiperplano ótimo, sem perda de generalidade, achando os β_i de modo a

$$\min_{\beta} \frac{\|\beta\|^2}{2},$$

em que $\beta = (\beta_1, ..., \beta_p)^T$, sujeito a

$$y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) > 0, \quad \forall i = 1, ..., n,$$

Uma vez que se obtém os β_i que otimizam a função anterior, a função classificadora do SVM é dada por

$$g(x) = \frac{f(x)}{|f(x)|} = -(f(x)). \tag{2.1}$$

2.3.2 Para fins de classificação binária: caso não linearmente separável

No caso de não podermos separar os dados linearmente, o conceito de hiperplano que divide perfeitamente as duas classes será modificado para o que "quase separa" as classes. Assim é preciso considerar uma margem mais flexível (*soft margin*). A generalização do classifica-

dor de máxima margem para o caso não separável é conhecida como classificador de suporte vetorial.

Essa mudança consiste no seguinte problema de maximização

$$\max_{\beta_0,\ldots,\beta_p,\xi_1,\ldots,\xi_n} M,$$

sujeito a

$$\sum_{j=1}^{p} \beta_j^2 = 1, \tag{2.2}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq M(1-\xi_i), \quad i = 1,\ldots,n, \tag{2.3}$$

$$\sum_{i=1}^{n} \xi_i \leq C, \quad \xi_i \geq 0, \tag{2.4}$$

em que C é um parâmetro de ajuste, e $\xi_1,\ldots,\xi_n$ são as variáveis de folga. A inclusão dessas variáveis de folga permite que alguns pontos sejam classificados incorretamente. Se $\xi_i = 0$ a observação i seria classificada corretamente, estando no lado correto da margem. Se $0 < \xi_i < 1$ a observação i também seria classificada corretamente, mas estando no lado errado da margem. E, se $\xi_i > 1$, ela seria classificada incorretamente, estando do lado errado do hiperplano. Se os valores dos parâmetros de folga forem demasiadamente elevados, haverá muitos pontos incorretamente classificados (ou seja, um ponto x_k incorretamente classificado será aquele tal que $y_k(\beta_0 + \langle \beta, x_k \rangle) \leq 0$).

Assim, a restrição dada por 2.4 proporciona um limite superior para o erro de classificação gerado pelo classificador da equação 2.1. Para minimizar a equação 2.4 e, ao mesmo tempo, maximizar a margem, a formulação SVM para o caso de dados não separáveis consiste em

$$\max_{\beta_0,\ldots,\beta_p,\xi_1,\ldots,\xi_n} \frac{1}{n}\sum_{i=1}^{n} \xi_i + \frac{\lambda \|\beta\|^2}{2}, \tag{2.5}$$

sujeito a $\xi_i + y_i(\beta_0 + \langle \beta, x_i \rangle) \geq 1$, (considerando $M = 1$, sem perda de generalidade), em que $\lambda > 0$ é um parâmetro de ajuste que controla o balanceamento entre o erro e a margem.

Dada uma constante a, considere a notação $[a]_+ = \max\{a, 0\} = \min(\xi_i$ sujeito a $\xi_i \geq 0$ e $\xi_i \geq a)$ com base nela, 2.5 pode ser expressa como

$$\frac{1}{n}\sum_{i=1}^{n} [1 - y_i(\beta_0 + \langle \beta, x_i \rangle)] + \frac{\lambda \|\beta\|^2}{2},$$

ou

$$\frac{1}{n}\sum_{i=1}^{n} [1 - y_i f(x_i)]_+ \frac{\lambda \|\beta\|^2}{2}.$$

Nessa expressão, o termo

$$L(f(x_i), y_i) = [1 - y_i f(x_i)]_+$$

denomina-se função de perda, e $f(x_i) = y_i(\beta_0 + \langle \beta, x \rangle)$ é a margem do ponto (x_i, y_i).

Considerando $M = 1$, note que a restrição dada pela Equação 2.3 pode ser reescrita como

$$\xi_i - 1 + y_i(\beta_0 + \langle \beta, x_i \rangle) \geq 0. \tag{2.6}$$

Assim, introduzindo dois conjuntos de multiplicadores de Lagrange, um para a restrição dada pela Equação 2.6 e outro para a restrição $\xi_i \geq 0$, temos o Lagrangeano

$$l_p(\beta_0, \beta, \xi, \alpha, \gamma) = \sum_{i=1}^{n} \xi_i + \frac{n\lambda \|\beta\|^2}{2} + \sum_{i=1}^{n} \alpha_i [1 - y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) - \xi_i] - \sum_{i=1}^{n} \gamma_i \xi_i. \tag{2.7}$$

Derivando a Equação 2.7 com respeito a β, β_0 e ξ_i, obtém-se

$$\frac{\partial l_p}{\partial \beta} = n\lambda\beta - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \iff \beta = \frac{1}{n\lambda} \sum_{i=1}^{n} \alpha_i y_i x_i, \tag{2.8}$$

$$\frac{\partial l_p}{\partial \beta_0} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \iff \sum_{i=1}^{n} \alpha_i y_i = 0, \tag{2.9}$$

$$\frac{\partial l_p}{\partial \xi_i} = 1 - \alpha_i - \gamma_i = 0 \quad = \quad \gamma_i = 1 - \alpha_i, \tag{2.10}$$

para $\alpha_i \geq 0$, $\gamma_i \geq 0$ e $i = 1, ..., n$. Simplificando l_p de acordo com esses resultados, obtemos o problema dual

$$\begin{aligned}
& \sum_{i=1}^{n} \xi_i + \frac{n\lambda \|\beta\|^2}{2} + \sum_{i=1}^{n} \alpha_i [1 - y_i(\beta_0 + \langle \beta, x_i \rangle) - \xi_i] - \sum_{i=1}^{n} \gamma_i \xi_i = \\
= & \sum_{i=1}^{n} (1 - \alpha_i - \gamma_i)\xi_i + \frac{n\lambda \|\beta\|^2}{2} + \sum_{i=1}^{n} \alpha_i y_i - \sum_{i=1}^{n} \alpha_i y_i \langle \beta, x_i \rangle \\
= & \sum_{i=1}^{n} \alpha_i + \frac{n\lambda \|\beta\|^2}{2} - \sum_{i=1}^{n} \alpha_i y_i \sum_{j=1}^{n} \beta_j x_{ij} \\
= & \sum_{i=1}^{n} \frac{n\lambda}{2(n\lambda)^2} \left\langle \sum_{i=1}^{n} \alpha_i y_i x_i, \sum_{j=1}^{n} \alpha_j y_j x_j \right\rangle \\
= & \sum_{i=1}^{n} \alpha_i + \frac{1}{2n\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \frac{1}{n\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\
= & \sum_{i=1}^{n} \alpha_i - \frac{1}{2n\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle,
\end{aligned} \tag{2.11}$$

ou seja,

$$\min_{\alpha_1,\ldots,\alpha_n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2n\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle, \tag{2.12}$$

sujeito a $0 \leq \alpha_i \leq 1$ e $\sum_{i=1}^{n} \alpha_i y_i = 0$ para $i, j = 1, \ldots, n$.

Computacionalmente chega-se à conclusão de que a otimização que se deseja realizar para os classificadores de suporte vetorial se resumem a produtos internos entre as observações, e não com as próprias observações.

Também se observa que o classificador de suporte vetorial linear pode ser escrito como

$$f(x) = \beta_0 + \langle \beta, x \rangle + \beta_0 + \left\langle \frac{1}{n\lambda} \sum_{i=1}^{n} \alpha_i y_i x_i, x \right\rangle = \beta_0 + \frac{1}{n\lambda} \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle, \tag{2.13}$$

em que se tem n parâmetros α_i, um para cada observação de treinamento.

O problema dual revela algumas características interessantes do SVM. Em primeiro lugar, como ele requer n variáveis, o tamanho da amostra representa o fator dominante da complexidade do problema em vez do número p de preditores. Em segundo lugar, a solução $\hat{\alpha}$ depende apenas dos produtos internos dos dados de treinamento $\langle x_i, x_j \rangle$ de 2.12. Em terceiro lugar, uma vez que $\hat{\alpha}$ satisfaz o critério $0 \leq \hat{\alpha} \leq 1$ e a condição de equilíbrio $\sum_{i=1}^{n} \hat{\alpha}_i y_i = 0$, tem-se o valor normal ao hiperplano ótimo, dado por $\hat{\beta} = \frac{1}{n\lambda} \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i = 0$. E em quarto lugar, condicionando os pontos x_i' tais que $1 - y_i(\hat{\beta}_0 + \langle \hat{\beta}, x_i' \rangle) = 0$, tem-se $\hat{\beta}_0 = y_i' - \frac{1}{n\lambda} \sum_{i=1}^{n} \hat{\alpha}_i y_i \langle x_i, x_i' \rangle$.

Observe que, se $y_i(\hat{\beta}_0 + \langle \hat{\beta}, x_i \rangle) > 1$, então temos $\xi_i = 0$ (restrição da Equação **??**) e, por isso, $\alpha_i = 0$. Por outro lado, se $\hat{\alpha}_i > 0$, o ponto constitui o suporte vetorial S, de modo que podemos reescrever

$$f(x) = \beta_0 \sum_{i \in S} \alpha_i \langle x_i, x \rangle. \tag{2.14}$$

2.3.3 Para fins de classificação binária: caso não linear

Uma extensão direta do SVM para o caso não linear consiste em substituir o produto interno $\langle x_i, x_j \rangle$ do espaço euclidiano por um outro espaço característico, em que se consideram transformações de $x \in \mathbb{R}^p$ para um espaço de dimensão superior.

Seja $\Phi(x) = (\phi_1(x), \ldots, \phi_n(x))^T$ um vetor de transformações de x. Substituindo-se $\langle x_i, x_j \rangle$ por $\langle \phi(x_i), \phi(x_j) \rangle$, tem-se de imediato uma extensão do SVM linear.

Para essa generalização não é necessário especificar a função Φ, sendo suficiente especificar uma função $K(x, q)$ que resulta do produto interno $\langle \Phi(x), \Phi(q) \rangle$. Tendo em mãos uma função K, a função de discriminação não linear pode ser escrita como

$$\hat{f}(x) = \frac{1}{n\lambda} \sum_{i=1}^{n} \hat{\alpha}_i y_i K(x_i, x) + \hat{\beta}_0. \tag{2.15}$$

Desse modo, a forma da curva que proporciona a classificação é determinada por K.

Essa função, que se denomina *kernel*, deve ser uma função simétrica e não negativa. Entre os kernels mais usuais, encontram-se o linear,

$$K(x_i, x_l) = \langle x_i, x_l \rangle, \tag{2.16}$$

o polinomial de grau d

$$K(x_i, x_l) = \gamma(C_0 + \langle x_i, x_l \rangle)^d, \tag{2.17}$$

o radial (ou gaussiano),

$$K(x_i, x_l) = \exp(-\gamma \|x_i - x_l\|^2), \tag{2.18}$$

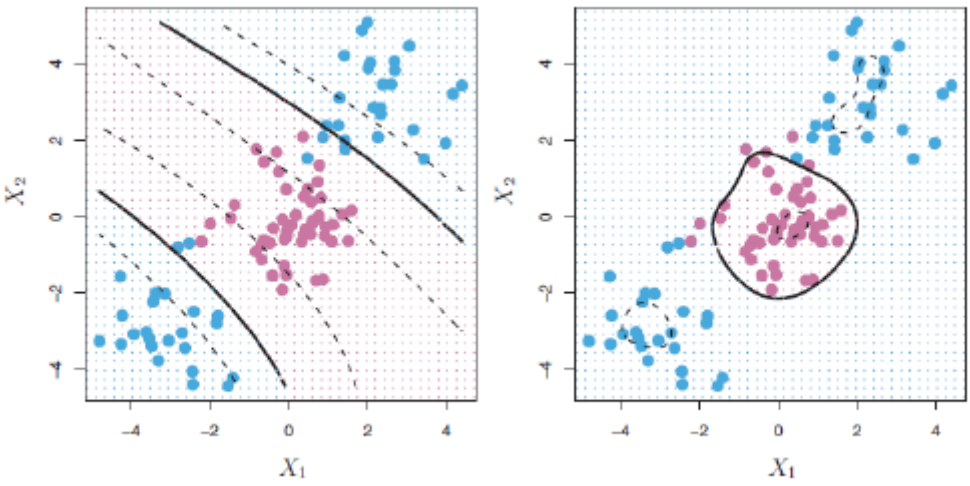
e o sigmóide (ou sigmoidal),

$$K(x_i, x_l) = \tanh(C_0 + \gamma \langle x_i, x_l \rangle), \tag{2.19}$$

em que C_0 é uma constante e $\tanh(x)$ é a função tangente hiperbólica de x.

A Figura 3.2, retirada de James et al. (2013), ilustra a aplicação do *kernel* polinomial e do *kernel* radial para a classificação do conjunto de dados.

Figura 2.5 – Máquinas de suporte vetorial: à esquerda temos o resultado proporcionado por um *kernel* polinomial de grau 3, e à direita, para o mesmo conjunto de dados, observa-se o resultado da aplicação de um kernel radial



James et al. (2013).

2.3.4 Para fins de regressão

Nas seções anteriores, a máquina de suporte vetorial (SVM) foi apresentada como uma ferramenta para classificação. Esta seção trata da regressão por suporte vetorial (SVR), em que $y_i \in \mathbb{R}$ é uma variável resposta e os elementos $x_{i1}, ..., x_{ip}$ são as variáveis regressoras. Do ponto

de vista estatístico, a regressão do tipo SVR é uma técnica não paramétrica, em que se obtém um ajuste para a resposta y com base no princípio da máxima margem.

Considere inicialmente a função linear

$$f(x_i) = \beta_0 + \langle \beta, x \rangle, \tag{2.20}$$

em que $f(x_i)$ seja tal que minimize a norma

$$\frac{\|\beta\|^2}{2} = \frac{\langle \beta, \beta \rangle}{2}, \tag{2.21}$$

sujeita a uma condição do tipo

$$|y_i - f(x_i)| \leq \varepsilon, \quad \forall i = 1, ..., n, \tag{2.22}$$

A equação 2.22 é conhecida como L1 ou $\varepsilon-$insensível, e significa que o problema de otimização de 2.21 se restringe aos resíduos absolutos inferiores a ε. No entanto, é possível que nem todos os pontos possam atender a essa restrição. Para acomodar tais pontos, são introduzidas variáveis de folga ξ_i e ξ_i^* para cada ponto. Isso é similar à ideia de *Margem Suave* em SVM para fins de classificação. Agora, a função a ser maximizada é:

$$\frac{\|\beta\|^2}{2} + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{2.23}$$

sujeito às restrições

$$y_i - f(x_i) \leq \varepsilon + \xi_i, \tag{2.24}$$

e

$$f(x_i - y_i) \leq \varepsilon + \xi_i^*, \tag{2.25}$$

em que $\xi_i \geq 0$, $\xi_i^* \geq 0$ e $\forall i = 1, ..., n$. A constante $C > 0$ permite controlar a penalidade imposta às observações que se encontram fora da margem ε.

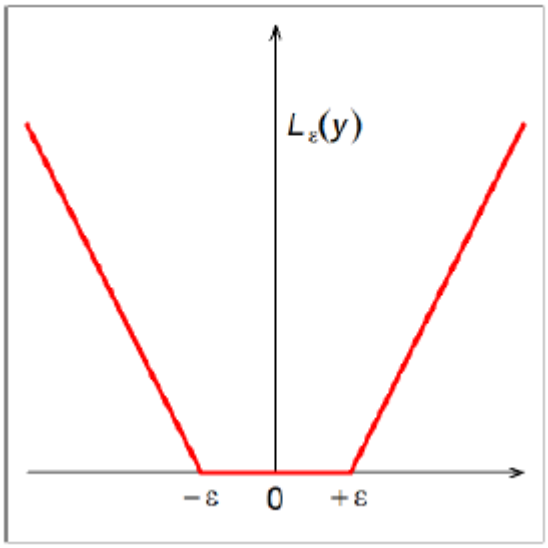
A função de perda ε-insensível definida em (Vapnik, 1995) é dada por

$$L_\varepsilon(y) = \begin{cases} 0, & \text{para } |f(x) - y| < \varepsilon, \\ |f(x) - y| < \varepsilon, & \text{caso contrário.} \end{cases} \tag{2.26}$$

Essa função de perda ignora os erros inferiores a ε, tratando-os como nulos. A Figura 2.6 mostra a forma gráfica da função de perda $\varepsilon-$insensível.

Com base na função objetivo primal $\|\beta\|^2/2$, e suas restrições, o Lagrangeano L é obtido mediante utilização de quatro multiplicadores de Lagrange: η_i, η_i^*, α_i e α_i^*, de modo que:

Figura 2.6 – Função de perda $\varepsilon-$insensível na linha mais espessa em vermelho.



Fonte: Próprio autor (2023)

$$\begin{aligned} L &= \frac{\|\beta\|^2}{2} + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ &- \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i - y_i + \langle\beta, x_i\rangle + \beta_0) - \sum_{i=1}^{n}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle\beta, x_i\rangle - \beta_0), \end{aligned} \tag{2.27}$$

em que $\eta_i \geq 0$, $\eta_i^* \geq 0$, $\alpha_i \geq 0$ e $\alpha_i^* \geq 0$.

A otimização de L requer

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0, \tag{2.28}$$

$$\frac{\partial L}{\partial \beta} = \beta - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i = 0, \tag{2.29}$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i = 0, \tag{2.30}$$

e

$$\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0. \tag{2.31}$$

Substituindo as equações 2.28, 2.29, 2.30 e 2.31 em 2.27, obtém-se o problema dual. Da Equação 2.29, temos

$$\beta = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*),$$

de modo que a função de regressão possa ser escrita como

$$\begin{aligned} f(x) &= \beta_0 + \langle \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) x_i, x \rangle \\ &= \beta_0 + \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle. \end{aligned} \tag{2.32}$$

Agora, a solução para o problema primal dos SVM será dada por

$$\max \left[-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i (y_i - \xi) - \alpha_i^* (y_i - \xi) \right], \tag{2.33}$$

restrito a

$$0 \leq \alpha_i, \quad \alpha_i^* \leq C, \quad \forall i = 1, ..., n,$$

e

$$\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0,$$

em que C é um valor pré-especificado.

Para obter β_0, consideram-se as condições

$$\alpha_i (\varepsilon + \xi_i - y_i + \langle \beta, x_i \rangle, \beta_0) = 0, \tag{2.34}$$

$$\alpha_i^* (\varepsilon + \xi_i^* - y_i + \langle \beta, x_i \rangle, \beta_0) = 0, \tag{2.35}$$

$$(C - \alpha_i) \xi_i = 0, \tag{2.36}$$

e

$$(C - \alpha_i^*) \xi_i^* = 0, \tag{2.37}$$

em que o produto entre as restrições e os seus respectivos multiplicadores de Lagrange anula-se (KARUSH, 2014; KUHN; TUCKER, 1951). Se $0 < \alpha_i < C$, então $\xi_i = 0$ e $\varepsilon - y_i \langle \beta, x_i \rangle + \beta_0 = 0$. Analogamente, se $0 < \alpha_k^* < C$, temos $\xi_k^* = 0$ e $\varepsilon + y_k - \langle \beta, x_k \rangle - \beta_0 = 0$.

Vejamos agora o caso não linear

$$f(x_i) = \beta_0 + \langle \beta, \Phi(x_i) \rangle, \tag{2.38}$$

em que $\Phi(x) = [\phi_i(x), ..., \phi_p(x)]^T$, $\beta = (\beta_1, ..., \beta_p)^T$. A transformação Φ é uma função de transferência não linear que mapeia a entrada x do espaço $\mathbb{R}^p$ para o espaço característico desejado. Com base na forma expandida $f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + \beta_0$ podemos escrever 2.38 como

$$\begin{aligned} f(x_i) &= \beta_0 + \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle \phi(x_i), \phi(x) \rangle \\ &= \beta_0 + \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x), \end{aligned} \tag{2.39}$$

em que $K(x_i, x)$ é a função *kernel*. Assim, é suficiente especificar a função K em vez de explicitar a transformação Φ. Por analogia aos resultados anteriores, isso nos remete ao seguinte problema de otimização, maximizar

$$\max \left[-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) - \varepsilon \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) + \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*) \right], \tag{2.40}$$

$$\text{sujeito a} \quad \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0, \quad 0 \leq \alpha_i \leq C, 0 \leq \alpha_i \leq C.$$

Na prática, os valores ε e C são determinados com base em experimentos computacionais.

2.4 Análise de Componentes Principais

A análise de componentes principais é uma técnica da estatística multivariada que consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão denominadas de componentes principais.

Os componentes principais apresentam propriedades importantes: cada componente principal é uma combinação linear de todas as variáveis originais, sendo independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados.

A análise de componentes principais é associada à ideia de redução de massa de dados, com menor perda possível da informação. Procura-se redistribuir a variação observada nos eixos originais de forma a se obter um conjunto de eixos ortogonais não correlacionados. Esta técnica pode ser utilizada para geração de índices e agrupamento de indivíduos. A análise agrupa os indivíduos de acordo com sua variação, isto é, os indivíduos são agrupados segundo suas variâncias, ou seja, segundo seu comportamento dentro da população, representado pela variação do conjunto de características que define o indivíduo.

Segundo Regazzi (2000), apesar das técnicas de análise multivariada terem sido desenvolvidas para resolver problemas específicos, principalmente de Biologia e Psicologia, podem ser também utilizadas para resolver outros tipos de problemas em diversas áreas do conhecimento. A análise de componentes principais é a técnica mais conhecida, contudo é importante ter uma visão conjunta de todas ou quase todas as técnicas da estatística multivariada para resolver a maioria dos problema práticos.

2.4.1 Matriz de dados X e de covariância S

Considere a situação em que observamos p características de n indivíduos de uma população π. As características observadas são representadas pelas variáveis $X_1, X_2, X_3, ..., X_p$. A matriz de dados é de ordem $n \times p$ é denominada de matriz X.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}$$

A estrutura de interdependência entre as variáveis da matriz de dados é representada pela matriz de covariância S ou pela matriz de correlação R. O entendimento dessa estrutura por meio das variáveis $X_1, X_2, ..., X_p$, pode ser na prática uma coisa complicada. Assim, o objetivo da análise de componentes principais é transformar essa estrutura complicada, representada pelas variáveis $X_1, X_2, ..., X_p$, em uma outra estrutura representada pelas variáveis $Y_1, Y_2, ..., Y_p$ não correlacionadas e com variâncias ordenadas, para que seja possível comparar os indivíduos usando apenas as variáveis Y_i's que apresentam maior variância. A solução é dada a partir da matriz de covariância S ou da matriz de correlação R.

A partir da matriz X de dados de ordem $n \times p$ podemos fazer uma estimativa da matriz de covariância Σ da população π que representaremos por S. A matriz S é simétrica e de ordem $p \times p$.

$$S = \begin{bmatrix} \text{Var}(x_1) & \text{Cov}(x_1x_2) & \text{Cov}(x_1x_3) & \cdots & \text{Cov}(x_1x_p) \\ \text{Cov}(x_2x_1) & \text{Var}(x_2) & \text{Cov}(x_2x_3) & \cdots & \text{Cov}(x_2x_p) \\ \text{Cov}(x_3x_1) & \text{Cov}(x_3x_2) & \text{Var}(x_3) & \cdots & \text{Cov}(x_3x_p) \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ \text{Cov}(x_px_1) & \text{Cov}(x_px_2) & \text{Cov}(x_px_3) & \cdots & \text{Var}(x_p) \end{bmatrix}.$$

Geralmente as características são obervadas em unidades de medidas diferentes entre si, e neste caso, segundo Regazzi (2000) é conveniente padronizar as variáveis X_j $(i = 1, 2, 3, ..., p)$. A padronização pode ser feita com média zero e variância um, ou com variância um e média qualquer.

1. Padronização com média zero e variância unitária

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s(x_j)}, \quad i = 1, 2, ..., n \quad \text{e} \quad j = 1, 2, ..., p.$$

2. Padronização com variância e média qualquer

$$z_{ij} = \frac{x_{ij}}{s(x_j)}, \quad i = 1,2,...,n \quad \text{e} \quad j = 1,2,...,p,$$

em que, $\bar{X}_j$ e $s(x_j)$ são, respectivamente, a estimativa da média e o desvio padrão da característica j:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

e

$$s(x_j) = \sqrt{\text{Var}(x_j)}, \quad j = 1,2,...,p$$

$$\sqrt{\text{Var}(x_j)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

Após a padronização obtemos uma nova matriz de dados Z:

$$Z = \begin{bmatrix} z_{11} & z_{12} & z_{13} & \cdots & z_{1p} \\ z_{21} & z_{22} & z_{23} & \cdots & z_{2p} \\ z_{31} & z_{32} & z_{33} & \cdots & z_{3p} \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ z_{n1} & z_{n2} & z_{n3} & \cdots & z_{np} \end{bmatrix}$$

A matriz Z das variáveis padronizadas z_j é igual a matriz de correlação da matriz de dados X. Para determinar os componentes principais normalmente partimos da matriz de correlação R. É importante observar que o resultado encontrado para a análise a partir da matriz S pode ser diferente do resultado encontrado a partir da matriz R. A recomendação é que a padronização só ser feita quando as unidades de medidas das características observadas não forem as mesmas.

3. Determinação dos componentes principais

Os componentes principais são determinados resolvendo-se a equação característica da matriz S ou R, isto é:

$$\det(R - \lambda \mathbf{I}) = 0 \quad \text{ou} \quad |R - \lambda \mathbf{I}| = 0$$

em que R é dada por

$$R = \begin{bmatrix} 1 & r(x_1x_2) & r(x_1x_2) & \cdots & r(x_1x_p) \\ r(x_2x_1) & 1 & r(x_2x_3) & \cdots & r(x_2x_p) \\ r(x_3x_1) & r(x_3x_2) & 1 & \cdots & r(x_3x_p) \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ r(x_px_1) & r(x_px_2) & r(x_px_3) & \cdots & 1 \end{bmatrix}$$

e $\mathbf{I}$ a matriz identidade.

Se a matriz R for de posto completo igual a p, isto é, não apresentar nenhuma coluna que seja combinação linear de outra, a equação $|R-\lambda\mathbf{I}|=0$ terá p raízes chamadas de autovalores ou raízes características da matriz R. Na montagem da matriz de dados X é importante observar que o valor de n (indivíduos, tratamentos, genótipos, etc.) dever ser pelo menos igual a $p+1$.

Sejam $\lambda_1, \lambda_2, ..., \lambda_p$ as raízes da equação característica da matriz R ou S, então:

$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_p.$$

Para cada autovalor λ_i existe um autovetor $\mathbf{a}_i$

$$\mathbf{a}_i = \begin{bmatrix} a_{i1} \\ a_{i2} \\ \cdots \\ a_{ip} \end{bmatrix}$$

Os autovetores $\mathbf{a}_i$ são normalizados, isto é, a soma dos quadrados dos coeficientes é igual a 1, e ainda são ortogonais entre si. Devido a isso apresentam as seguintes propriedades:

$$\sum_{j=1}^{p} a_{ij}^2 = 1, \quad (\tilde{a}_i' \cdot \tilde{a}_j = 1)$$

e

$$\sum_{j=1}^{p} a_{ij} \cdot a_{kj} = 0 \quad (\tilde{a}_i' \cdot \tilde{a}_k = 1 \text{ para } i \neq k).$$

Sendo $\mathbf{a}_i$ o autovetor correspondente ao autovalor λ_i, então o i-ésimo componente principal é dado por:

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p.$$

Os componentes principais apresentam as seguintes propriedades:

1. A variância do componente principal Y_i é igual ao valor do autovalor λ_i.

$$\text{Var}(Y_i) = \lambda_i$$

2. O primeiro componente é o que apresenta maior variância e assim por diante:

$$\text{Var}(Y_1) > \text{Var}(Y_2) > \cdots > \text{Var}(Y_p)$$

3. O total de variância das variáveis originais é igual ao somatório dos autovalores que é igual ao total de variância dos componentes principais:

$$\sum \text{Var}(X_i) = \sum \lambda_i = \sum \text{Var}(Y_i).$$

4. Os componentes principais não são correlacionados entre si:

$$\text{Cov}(Y_i, Y_j) = 0.$$

2.4.2 Contribuição de cada componente principal

A contribuição C_i de cada componenete principal Y_i é expressa em porcentagem. É calculada dividindo-se a variância de Y_i pela varância total. Representa a proporção de variância total explicada pelo componenete principal Y_i.

$$C_i = 100\frac{\text{Var}(Y_i)}{\sum_{i=1}^{p}\text{Var}(Y_i)} = 100\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i} = 100\frac{\lambda_i}{\text{traço}(S)}.$$

A importância de um componente principal é avaliada por meio de sua contribuição, isto é, pela proporção de variância total explicada pelo componente. A soma dos primeiros k autovalores representa a proporção de informação retida na redução de p para k dimensões. Com essa informação podemos decidir quantas componentes vamos usar na análise, isto é, quantos componentes serão utilizados para diferenciar os indivíduos.

Não existe um modelo estatístico que ajude nesta decisão. Segundo Regazzi (2000) para aplicações em diversas áreas do conhecimento o número de componentes utilizados têm sido aquele que acumula, por exemplo,70% ou mais de proporção da variância total.

$$100\frac{\text{Var}(Y_1) + \cdots + \text{Var}(Y_k)}{\sum_{i=1}^{k}\text{Var}(Y_i)} \geq 70\% \quad \text{onde} \quad k < p.$$

2.4.3 Interpretação de cada componente

Esta análise é feita verificando-se o grau de influência que cada variável X_j tem sobre o componente Y_i. O grau de influência é dado pela correlação entre cada X_j e o componente Y_i que está sendo interpretado. Por exemplo, a correlação entre X_j e Y_1 é:

$$\text{Corr}(X_j, Y_1) = r_{Xj} \cdot Y_1 = a_{1j} \cdot \frac{\sqrt{\text{Var}(Y_1)}}{\sqrt{\text{Var}(X_j)}} = \sqrt{\lambda_j} \cdot \frac{a_{1j}}{\sqrt{\text{Var}(X_j)}}.$$

Para comparar a influência de $X_1, X_2, ..., X_p$ sobre Y_1 análisamos o peso ou *loading* de cada variável sobre o componente Y_1. O peso de cada variável sobre um determinado componente é dado por:

$$w_1 = \frac{a_{11}}{\sqrt{\text{Var}(X_1)}}, w_2 = \frac{a_{12}}{\sqrt{\text{Var}(X_2)}}, \cdots, w_p = \frac{a_{1p}}{\sqrt{\text{Var}(X_p)}},$$

sendo w_1 o peso de X_1.

2.4.4 Escores dos componentes principais

Os escores são os valores dos componentes principais. Após a redução de p para k dimensões, os k componentes principais serão os novos índivíduos e toda análise é feita utilizando-se os escores desses componentes. Na Tabela 2.1 é exemplificado a organização de um conjunto de dados composto por n tratamentos, p variáveis e k componentes principais.

Tabela 2.1 – Organização de um conjunto de dados com n tratamentos, p variáveis e k componentes

Tratamentos	Variáveis				Escores dos componentes principais			
(Indivíduos)	X_1	X_2	...	X_p	X_1	X_2	...	X_p
1	X_{11}	X_{12}	...	X_{1p}	Y_{11}	X_{12}	...	X_{1p}
2	X_{21}	X_{22}	...	X_{2p}	Y_{21}	X_{22}	...	X_{2p}
...	...	...	$\ddots$	...	...	...	$\ddots$	...
n	X_{n1}	X_{n2}	...	X_{np}	Y_{n1}	X_{n2}	...	X_{np}

Fonte: Próprio autor (2023).

Assim temos,

Tabela 2.2 – Escores do primeiro componente para os n tratamentos

Tratamento	Primeiro componente principal
1	$Y_{11} = a_{11}X_{11} + a_{12}X_{12} + \cdots + a_{1p}X_{1p}$
2	$Y_{21} = a_{11}X_{21} + a_{12}X_{22} + \cdots + a_{1p}X_{2p}$
...	...
N	$Y_{n1} = a_{n1}X_{n1} + a_{n2}X_{n2} + \cdots + a_{np}X_{np}$

Fonte: Próprio autor (2023).

2.5 Análise de Componentes Independentes

A análise de componentes independentes (ICA) é uma técnica capaz de revelar fatores escondidos em conjuntos de sinais (MORETTIN; SINGER, 2021). ICA define um modelo gerador para os dados observados, que são assumidos serem misturas de variáveis ocultas e

desconhecidas. As variáveis latentes são assumidas mutuamente independentes, e são chamadas de componentes independentes ou fontes dos dados observados (LATHAUWER; MOOR; VANDEWALLE, 2000).

As raízes de ICA vêm dos trabalhos de Darmois (DARMOIS, 1953) na década de 50 e (ORD, 1975) na década de 70, caracterizando variáveis aleatórias em estruturas lineares. Os trabalhos pioneiros em análise de componentes independentes foram desenvolvidos por Herault e Jutten (1986) na década de 80 e, na década de 90, Comon (1994) formalizou e desenvolveu a teoria básica de análise de componentes independentes concentrando os trabalhos nas condições de existência, unicidade e indeterminações da estimação. Durante a década de 90 e até os dias de hoje, diversas aplicações têm sido propostas nos mais variados contextos e bons resultados têm sido emonstrados.

2.5.1 Definição

Vamos assumir a hipótese de que os dados consistem de m variáveis aleatórias conjuntamente observadas T vezes. Assim, denotaremos os dados por $\mathbf{x}_j(t)$, onde $j = 1,2,...,m$ e $t = 1,2,...,T$. A formulação geral para o problema seria encontrar uma função, mapeando-se o espaço m-dimensional para o espaço n-dimensional, de maneira que as variáveis transformadas fornecessem as informações escondidas no espaço original. Ou seja, as variáveis transformadas deveriam ser os componentes implícitos que descrevessem a estrutura essencial dos dados. É esperado que estes componentes correspondam a alguma causa física envolvida no processo de geração dos dados. Cada componente $(\mathbf{y}_i(t))$ pode ser expresso como uma combinação linear das variáveis observadas $(\mathbf{x}_j(t))$.

$$\mathbf{y}_i(t) = \sum_j b_{ij}\mathbf{x}_j(t) \tag{2.41}$$

onde, $i = 1,2,...,n$ e $j = 1,2,...,m$ e b_{ij} são os coeficientes que definem a representação. O problema pode ser resolvido encontrando-se os coeficientes b_{ij} . Com o auxílio da álgebra linear, a transformação linear pode ser representada por:

$$\begin{bmatrix} \mathbf{y}_1(t) \\ \mathbf{y}_2(t) \\ \vdots \\ \mathbf{y}_n(t) \end{bmatrix} = \mathbf{B} \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \\ \vdots \\ \mathbf{x}_m(t) \end{bmatrix} \tag{2.42}$$

Agora, podemos determinar a matiz $\mathbf{B}$ através das propriedades estatísticas dos componentes transformados $\mathbf{y}_i(t)$, tais como descorrelação não-linear, máxima não-gaussiana e descorrelação no tempo (LATHAUWER; MOOR; VANDEWALLE, 2000).

2.5.2 Princípios básicos

Podemos considerar ICA como um passo além da simples descorrelação linear. De fato, a descorrelação linear (ou branqueamento) é utilizada como um pré-processamento para ICA.

Um vetor $\mathbf{z}(t) = (\mathbf{z}_1(t), \mathbf{z}_2(t), ..., \mathbf{z})_n)^T$ de média zero é dito branqueado se os elementos $\mathbf{z}_i(t)$ são descorrelacionados entre si e têm variância unitária. Em termos da matriz de covariância, isso significa:

$$E\{\mathbf{z}(t)\mathbf{z}^T(t)\} = \mathbf{I} \tag{2.43}$$

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

Para se obter as variáveis branqueadas $\mathbf{z}(t)$, aplica-se uma transformação $\mathbf{V}$ na variável observada $\mathbf{x}(t)$:

$$\mathbf{z}(t) = \mathbf{V}\mathbf{x}(t) \tag{2.44}$$

Para solucionar o problema, consideremos a matriz $\mathbf{R} = (\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n)$, cujas colunas são autovetores de norma unitária da matriz de covariância $\mathbf{C}^x = E\{\mathbf{x}\mathbf{x}^T\}$ (Esperança), e $\mathbf{D} = diag(d_1, d_2, ..., d_n)$ a matriz diagonal de autovalores de $\mathbf{C}^x$. A decomposição em autovetores e autovalores dá-nos a matriz de branqueamento:

$$\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^T. \tag{2.45}$$

Após o branqueamento, basicamente dois princípios norteiam ACI para acessar as fontes independentes: descorrelação não linear e máxima não-gaussianidade. A descorrelação não-linear encontra uma matriz separação de maneira que, para qualquer $i \neq j$, os componentes $\mathbf{y}_i(t)$ e $\mathbf{y}_j(t)$ são descorrelacionados e os componentes transformados $g(\mathbf{y}_i(t))$ e $h(\mathbf{y}_j(t))$ são também descorrelacionados, onde $g(\cdot)$ e $h(\cdot)$ são funções não lineares apropriadas.

A Teoria da Estimação e a Teoria da Informação fornecem muitos métodos clássicos para a escolha e estimação das não linearidades $g(\cdot)$ e $h(\cdot)$, tais como a máxima semelhança e a informação mútua (COVER, 1999).

Já o princípio da máxima não gaussiana busca no teorema do limite central os fundamentos para a separação de fontes independentes. No teorema, a soma de variáveis não gaussianas é mais próxima de uma gaussiana do que as variáveis originais. Assim, este princípio de separação busca encontrar máximos locais de não-gaussianas de uma combinação linear. Dessa forma, cada máximo local revela um componente independente. Para se medir a não gaussiana, métodos tais como a assimetria, curtose e cumulantes de ordem elevada são utilizados.

Ainda, em sua formulação básica, o modelo de ICA assume as hipóteses (i) das fontes serem estatisticamente independentes entre si, (ii) com distribuições não gaussianas e (iii) o número de fontes independentes ser igual ao número de misturas observadas. Com base nestas

hipóteses, em geral, os algoritmos utilizam estatística de ordem superior para estimar a matriz de separação.

2.5.3 Princípios para séries temporais

Os sinais mistura podem ser variáveis ordenadas ao invés de variáveis aleatórias. Isto contrasta com a formulação básica de ICA, na qual a sequência das amostras não tem ordem particular. Se os componentes independentes (CIs) são, digamos, séries temporais, eles podem conter muito mais estrutura do que simples variáveis aleatórias. A informação adicional pode tornar possível a estimação dos modelos por meio de estatística de segunda ordem e deixar de assumir a hipótese de fontes não-gaussianas.

Para sinais com estrutura temporal, as t realizações do processo representam a sequência de tempo. Além disso, a estimação do modelo baseia-se em hipóteses alternativas às hipótese de não gaussiana apresentada em ICA básico: assumir que os CIs têm autocovariâncias diferentes ou assumir que as variâncias dos CIs são não estacionárias. No entanto, apesar de não utilizar a estrutura temporal das séries, frequentemente a formulação básica também pode ser aplicada em séries temporais.

2.5.4 Hipótese de autocovariâncias diferentes

No caso de assumir a hipótese de autocovariâncias diferentes para cada CI, a forma da estrutura temporal é dada pelas autocovariâncias de cada sinal, $\text{cov}(\mathbf{x}_i(t), \mathbf{x}_i(t-\tau))$, e a covariância entre dois sinais, $\text{cov}(\mathbf{x}_i(t), \mathbf{x}_j(t-\tau))$. Assim, após a retirada da média de $\mathbf{x}(t)$, as estatísticas necessárias para obter os componentes independentes podem ser agrupadas na Matriz Covariância deslocada no tempo:

$$\mathbf{C}_\tau^x = E\{\mathbf{x}(t)\mathbf{x}^T(t-\tau)\} \tag{2.46}$$

onde, $\mathbf{x}(t)$ é a matriz contendo os sinais mistura e $\mathbf{x}^T(t-\tau)$ é a matriz transposta contendo os sinais mistura com atraso τ.

Aqui o ponto chave é que a informação de segunda ordem pode ser usada no lugar da informação de ordem superior para se obter as fontes independentes (PAJUNEN, 1998), (PAJUNEN, 2000). Assim, devemos encontrar a matriz-separação $\mathbf{B}$, além da covariância instantânea $(\tau = 0)$, as covariâncias defasadas $(\tau > 0)$ serem zero:

$$E\{\mathbf{y}_i(t)\mathbf{y}_j^T(t-\tau)\} = 0 \tag{2.47}$$

A motivação para se igualar a zero todas as covariâncias defasadas é o fato desta característica ser própria da independência. Para melhor compreender esta separação, consideremos

apenas uma matriz de covariância atrasada ($\tau = 1$). Após retirar a média e branquear $\mathbf{x}(t)$, tem-se $\mathbf{z}(t)$ e chega-se a matriz de separação ortogonal $\mathbf{W}$:

$$\mathbf{y}(t) = \mathbf{W}\mathbf{z}(t) \tag{2.48}$$

$$\mathbf{y}(t-\tau) = \mathbf{W}\mathbf{z}(t-\tau) \tag{2.49}$$

Pela linearidade e ortogonalidade, pode-se escrever a matriz covariância atrasada dos sinais branqueados (LATHAUWER; MOOR; VANDEWALLE, 2000):

$$\bar{\mathbf{C}}_\tau^z = \mathbf{W}^T \bar{\mathbf{C}}_\tau^y \mathbf{W} \tag{2.50}$$

onde

$$\bar{\mathbf{C}}_\tau^z = \frac{1}{2}[\mathbf{C}_\tau^z + (\mathbf{C}_\tau^z)^T] \tag{2.51}$$

Observa-se que $\bar{\mathbf{C}}_\tau^y$ é diagonal devido à independência dos vetores de $\mathbf{y}(t)$. O que a equação mostra é que $\mathbf{W}$ deve fazer parte da decomposição de autovalores de $\bar{\mathbf{C}}_\tau^z$ (LATHAUWER; MOOR; VANDEWALLE, 2000).

Estendendo o raciocínio para vários atrasos, é suficiente que apenas a covariância de um deles seja diferente das demais. Assim, a escolha deles não seria tão problemática. Em princípio, utilizando vários atrasos no tempo, deseja-se diagonalizar simultaneamente todas as matrizes-covariância correspondentes. No entanto, a diagonalização exata é pouco provável, o que nos leva a formular um indicador para o grau de diagonalização:

$$off(\mathbf{M}) = \sum_{i \neq j} m_{ij}^2 \tag{2.52}$$

onde $\mathbf{M} \in \{\mathbf{C}_1^z, \mathbf{C}_2^z, ..., \mathbf{C}_\tau^z\}$.

2.5.5 Hipótese de variâncias não estacionárias

A hipótese anterior pode não ser eficiente quando as componentes independentes têm autocovariâncias iguais (espectro de potência idêntico). Neste caso, uma alternativa é assumir a hipótese de variâncias não estacionárias dos CIs (MATSUOKA; OHOYA; KAWAMOTO, 1995). Assume-se também que a variância se modifica lentamente no tempo. Observa-se que este pressuposto independe das hipóteses mencionadas nas seções anteriores. Dada a hipótese de variâncias não estacionárias, pode-se chegar aos componentes independentes através da análise das autocorrelações locais ou através da análise dos cumulantes cruzados dos CIs.

i) Autocorrelações locais

Se encontrarmos a matriz $\mathbf{B}$ que produza $\mathbf{y}(t) \in \mathbf{R}^n$ descorrelacionado a cada instante do tempo, tem-se a independência (MATSUOKA; OHOYA; KAWAMOTO, 1995). Note que, como não é estacionária, a covariância de $\mathbf{y}(t)$ depende do atraso. Assim, se os componentes serem descorrelacionados a cada instante, haverá uma condição muito mais forte que o simples branqueamento.

ii) Cumulantes cruzados

Um segundo método, baseado na interpretação das variâncias não estacionárias, é através de cumulantes cruzados de ordem superior. Podemos medir a variância não estacionária do sinal $\mathbf{y}(t)$ usando uma medida baseada na correlação temporal das energias:

$$E\{\mathbf{y}_i^1(t)\mathbf{y}_j^2(t-\tau)\}. \tag{2.53}$$

Por uma questão de simplificação matemática, utilizam-se cumulantes. A autocorrelação não-linear é interpretada através do cumulante cruzado de 4ª ordem, correspondente às correlações de energias:

$$cum(\mathbf{y}(t),\mathbf{y}(t),\mathbf{y}(t-\tau),\mathbf{y}(t-\tau)). \tag{2.54}$$

2.5.6 Unificação dos princípios de separação

Os princípios de separação abordados foram unificados em (PAJUNEN, 1998),(PAJUNEN, 2000), com base no conceito da complexidade de Kolmogoroff (RISSANEN, 1978),(RISSANEN, 1983). Define-se complexidade de Kolmogoroff de uma "string" $\mathbf{z}(t)$ como a descrição mínima de seu comprimento; ou seja, a quantidade mínima de código (bits) necessária para descrever esta variável. Dessa forma, pode se medir a quantidade de estrutura de um sinal $\mathbf{y}(t)$ pela quantidade de compressão possível na codificação do sinal.

Dado o sinal $\mathbf{z}(t)$, o grau de incerteza desta variável pode ser medido pela entropia:

$$H(\mathbf{z}(t)) = -E\{\log p(\mathbf{z}(t))\} \tag{2.55}$$

e, dada uma transformação $\mathbf{y}(t) = \mathbf{W}\mathbf{z}(t)$, a entropia da transformação fica

$$H(\mathbf{y}(t)) - H(\mathbf{y}(t)) + \log(\det \mathbf{W}) \tag{2.56}$$

Definindo-se a informação mútua entre as variáveis transformadas $\mathbf{y}_i(t)$

$$IM(\mathbf{y}_1(t),\mathbf{y}_2(t),...,\mathbf{y}_n(t)) = \sum_i H(\mathbf{y}_i(t)) - H(\mathbf{y}(t)) \tag{2.57}$$

Chega-se

$$IM(\mathbf{y}_1(t),\mathbf{y}_2(t),...,\mathbf{y}_n(t)) = \sum_i H(\mathbf{y}_i(t)) - H(\mathbf{z}(t)) - \log|\det \mathbf{W}|. \tag{2.58}$$

No entanto, dado que a variável $\mathbf{z}(t)$ é conhecida a priori, $H(\mathbf{z}(t)) = 0$.

A entropia pode ser interpretada como o comprimento médio ótimo do código, o que leva à função objetivo:

$$IM(\mathbf{y}_1(t),\mathbf{y}_2(t),...,\mathbf{y}_n(t)) = \sum_i K(\mathbf{b}_i\mathbf{z}(t)) - \log|\det \mathbf{W}|, \tag{2.59}$$

onde $K(\mathbf{b}_i\mathbf{z}(t))$ é a Complexidade de Komolgoroff.

Avaliar a complexidade de Komolgorff dos sinais significa avaliar a correlação entre as amostras - uma vez que códigos mínimos referem-se a variáveis mais correlacionadas. Por outro lado, minimizar a informação mútua equivale a procurar estruturas nos dados - assim como o princípio da não-gaussiana utilizado em ACI básico. Dessa forma, a complexidade de Komolgorff pode ser interpretada como uma ferramenta que unifica os princípios de correlação temporal e os princípios de máxima não-gaussiana.

3 MATERIAIS E MÉTODOS

A seção apresenta estrutura e conceitos qua abordam a dissertação, permeando os assuntos e técnicas utilizadas para o desenvolvimento do trabalho.

3.1 Estrutura e conceitos

A dissertação utiliza dados reais de séries temporais financeiras, obtidas no Yahoo Finance, especificamente do Banco Bradesco SA (BBDC3.SA) e da Vale SA (VALE3.SA), acessados no dia 06 de fevereiro de 2023 às 14h22min, cujo período histórico está compreendido de 02/01/2015 a 01/02/2023, e podem ser coletados a partir do endereço (https://finance.yahoo.com/).

A análise técnica é uma abordagem popular para estudar os padrões e movimentos do mercado de capitais. Os resultados da análise técnica podem ser uma previsão de curto ou longo prazo baseada em padrões recorrentes; no entanto, a abordagem aqui trazida assume que os preços das ações movem-se em tendências, e que as informações que afetam os preços entram no mercado por um período de tempo finito, não instantaneamente (KAUFMAN, 2011). Os indicadores técnicos usados nesta análise são calculados a partir dos dados históricos de negociação.

O primeiro passo importante no desenvolvimento de um modelo de previsão baseado em SVR é a extração de recursos (transformando os recursos originais em novos) e a seleção de recursos (escolhendo o conjunto de recursos mais influente).

O uso da análise de componentes prinicipais (PCA) é um método de extração de recursos amplamente aplicado na estrutura de SVR. Com ela, vetores de entrada de alta dimensão são transformados em componentes principais não correlacionados(CP) calculando os autovetores da matriz de covariância das entradas originais. Já, a técnica de análise de componentes independentes (ICA) é uma técnica de processamento de sinal que foi originalmente desenvolvida para separação cega de fontes. Seu objetivo é obter componentes estatisticamente indenpendentes (CI) dos vetores transformados.

A abordagem aqui, expressa um modelo de previsão baseado em SVR, desenvolvido integrando PCA e ICA para aumentar a precisão da previsão para os preços das ações, pois mesmo um pequena melhoria desse desempenho pode ter uma influência significativa nas decisões de investimento.

Considerando o fato de que a análise técnica desempenha um papel importante na previsão, ela foi conduzida para calcular indicadores técnicos como recursos de entrada. Em seguida, o PCA é usada para extrair os componentes influentes dos recursos de entrada que são filtrados para transformar a entrada de alta dimensão em recursos de baixa dimensão.

Por outro lado, o ICA é aplicado para converter os recursos reduzidos em componentes independentes. O SVR finalmente usa as variáveis de entrada de baixa dimensão filtradas e transformadas para construir o modelo de previsão e prever os preços das ações.

Ao lado das técnicas de análise de *candles* [1], os indicadores técnicos são largamente utilizados por investidores e operadores de mercado no mundo todo, apresentamos no Quadro 3.1 os abordados na dissertação, bem como outros que poderão ser consultados no Apêndice A, para fins de conhecimento.

Quadro 3.1 - Indicadores técnicos utilizados na pesquisa (Continua)

1. Média móvel simples - MMS.	Segundo Queji e Caetano (2011), é utilizada para identificar a tendência que o ativo está enfrentando. Ela corresponde ao movimento dos candles (ou seja, à oscilação do preço)	MMS $= \frac{P_1+P_2+\cdots+P_n}{n}$, onde P_i é qualquer quantidade de períodos que se queira calcular.
2. Média móvel exponencial - MME.	Segundo Queji e Caetano (2011), bastante parecida com a média móvel simples, a exponencial só se distingue por um fator: ela valoriza mais os preços mais recentes. Significa que ela se aproxima mais dos candles do que a média móvel simples.	MME(n) = (Preço $\times$ K) + (MME$(n-1) \times (1-K)$, onde $K = 2/(n+1)$
3. Média móvel exponencial dupla - MMED.	Desenvolvida por Patrick Mulloy com o objetivo de reduzir o atraso e aumentar a capacidade de resposta. Essa média móvel de ação rápida permite que os traders identifiquem as mudanças de tendência rapidamente, resultando em melhores entradas nas tendências recém-formadas. O indicador é obviamente baseado na média móvel exponencial (MME), mas segue o preço mais de perto.	MMED $= 2 \times$ MME$_n$ $-$ MME(MME$_n$), onde n é o número de perídos anteriores.
4. Média móvel ponderada - MMP.	É um indicador de análise técnica que determina a direção da tendência. Ela fornece sinais para operar, atribuindo um peso maior a pontos de dados recentes, e menor aos pontos de dados passados.	MMP $= \frac{P_1 \times n + P_2 \times (n-1) + \cdots + P_n}{n(n-1)/2}$, onde n é p período de tempo e P_i é o i-ésimo preço.

[1] Recebe este nome devido ao fato de seus elementos de representação dos preços terem a aparência de uma vela.

Quadro 3.1 - Indicadores técnicos utilizados na pesquisa (Continua)

5. Média móvel ponderada por volume elástico - MMPVE.	É uma medida estatística que usa o volume para definir o período da média móvel. Ele incorpora informações de volume de maneira natural e lógica. O eVWMA pode ser visto como uma aproximação do preço médio pago por ação. A capacidade de "Usar Volume Médio"como seu período de volume torna este indicador independente de símbolo e independente de período de tempo. Isso permite que o uso alterne o período de tempo e o símbolo sem precisar alterar o período do volume.	MMPVE $= \frac{(n-v)\times \text{MMPVE anterior} + v\times p}{n}$, onde p é o preço e v o volume de ações negociadas.
6. Média móvel ponderada por volume - MMPV.	Utiliza o volume pesando preços com base na quantidade de operações em um determinado período de tempo. Os usuários podem definir o período, a fonte e o deslocamento. Os preços com forte quantidade de operações ganham mais peso do que os preços com baixa quantidade de operações.	$\text{MMPV} = \frac{\sum(\text{Volume} \times \text{Preço})}{\sum \text{Volume}}$
7. Média móvel exponencial de atraso zero - MMEAZ.	Objetivo é eliminar o atraso inerente associado a todas as tendências seguintes indicadores que calculam a média de um preço ao longo do tempo.	$\text{Lag} = \frac{n-1}{2}$, $\text{MMEData} = \text{Data} + (\text{Data} - \text{Data(Lag days ago)})$, MMEAZ = MME(MMEData,n)
8. Média móvel Hull - MMH.	É um pouco mais recentes em comparação às médias móveis tradicionais na análise técnica. Na década de 1990, o trader australiano Alan Hull desenvolveu a média móvel que leva seu nome com a finalidade de ser uma média mais rápida e menos propensa a falsos rompimentos no comparativo às médias tradicionais do mercado à época. Ela foi criada para resolver a demora de resposta aos indicadores de preço, mas sem perder a suavidade da curva apontada pelos indicadores.	$\text{MMH} = \text{MMP}(2 \times \text{MMP}(n/2) - \text{MMP}(n)), \sqrt{n}))$

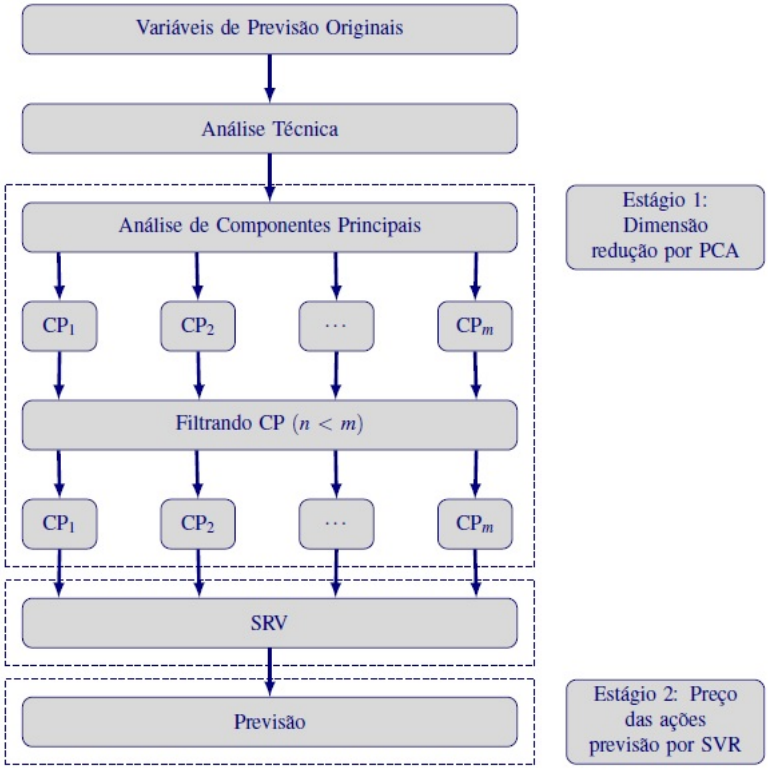
Quadro 3.1 - Indicadores técnicos utilizados na pesquisa (Conclusão)

9. Média móvel de Arnaud Legoux - MMAL.	Projetada para abordar dois problemas, geralmente identificados em diferentes tipos de Média Móvel: suavidade e capacidade de resposta. Ao usar, digamos, uma média móvel simples, você pode notar que quanto mais suave ela é, mais tempo ela leva para fornecer um sinal. Pode ser até que, quando o sinal for entregue, o movimento pelo qual você estava esperando já tenha terminado. Por outro lado, uma Média Móvel de curto prazo, embora seja mais responsiva, pode parecer muito variável. Portanto, ao usar uma média móvel tradicional, você precisa escolher entre capacidade de resposta e suavidade. A Média Móvel de Arnaud Legoux (do inglês Arnaud Legoux Moving Average, ALMA) foi criada com o propósito de resolver exatamente esse problema.	MMAL $= \frac{1}{\text{NORM}} \sum_{i=1}^{n} \text{source}(i) \times e^{-\frac{(i-offset)^2}{\sigma^2}}$,onde $offset$ é um multiplicador que determina o alinhamento do preço (por definição é igual a 0.85) e σ é o valor por definição igual a 6, valores menores tornarão o indicador ALMA menos reactivo a alterações de preços.
10. Média móvel convergente e divergente - MMCD.	Um dos indicadores mais utilizados por traders. Criada em 1960 por Gerald Appel, a principal finalidade é monitorar a aceleração ou a desaceleração das tendências de preços das ações.	MMCD $=$ Período 12 MME $-$ Período 26 MME
11. Índice de Força Relativa - IFR.	O IFR é um oscilador de momentum que mede a velocidade e a mudança dos movimentos de preços. Geralmente, o IFR é considerado superavaliado quando está acima de 70 e subavaliado quando abaixo de 30.	IFR $= 100 - \frac{100}{1+\frac{\text{Média da Mudança de Preços para cima}}{\text{Média da Mudança de Preços para baixo}}}$

Fonte:Próprio autor(2023).

As análises estatísticas foram realizadas por meio do software R, segundo (R Development Core Team, 2023) para manipulação, organização e tratamento dos dados.

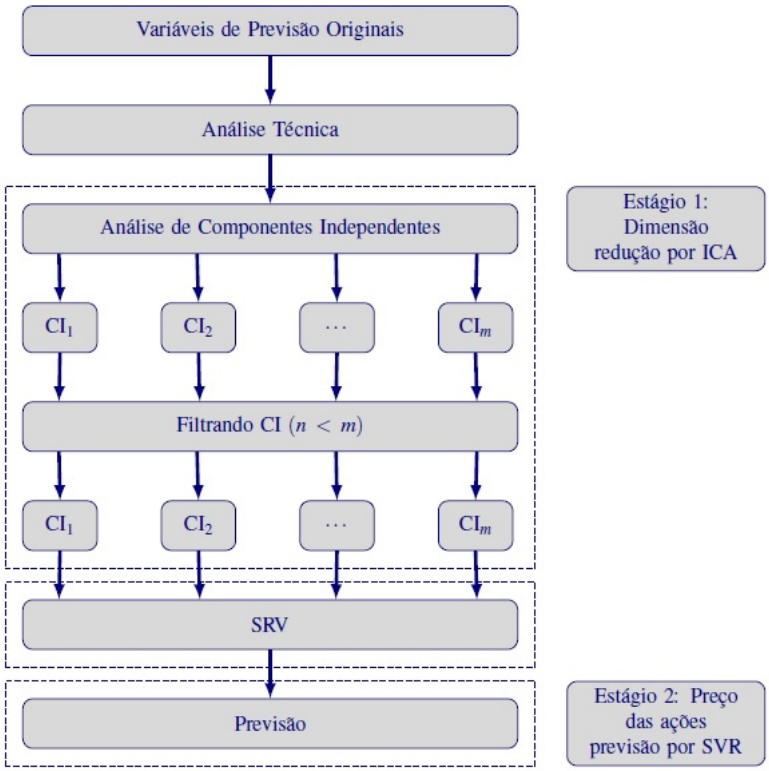
Figura 3.1 – Estrutura de previsão proposta PCA.



Fonte: Próprio autor (2023)

Optou-se por uma estrutura de análise que norteasse os objetivos da pesquisa. Através delas, é possível compreender o esqueleto de toda modelagem trabalhada.

Figura 3.2 – Estrutura de previsão proposta ICA.



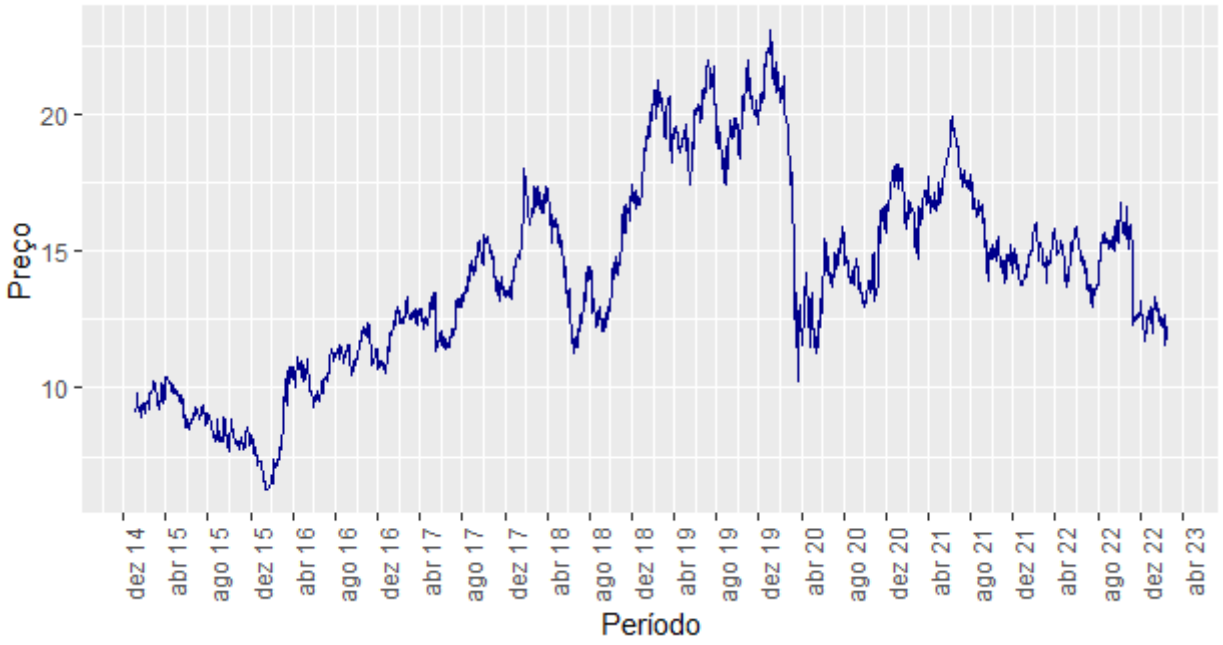
Fonte: Próprio autor (2023)

A robustez dos métodos avaliados é definida com 90% de tamanho de amostra de treinamento e 10% de tamanho de amostra para teste em todos os modelos aplicados. Justifica-se, por se tratar de um período histórico longo de séries temporais.

O PCA transforma vetores de entrada de alta dimensão em componentes principais não correlacionados (CPs) calculando os autovetores da matriz de covariância das entradas originais. Novamente, o ruído latente que reside nos dados de séries temporais financeiras muitas vezes leva a um ajuste excessivo ou insuficiente e, portanto, prejudica o desempenho do sistema de previsão.

O *Kernel RBF* - (função radial básico) é utilizado nas análises, por ser mais popular devido à sua semelhança com o algoritmo K-Nearest Neighborhood (K-NN). Ele tem as vantagens do K-NN e supera o problema da complexidade do espaço, pois as máquinas de suporte vetorial do *kernel RBF* precisam apenas armazenar os vetores de suporte durante o treinamento e não todo o conjunto de dados. A Figura 3.3 apresenta o comportamento da série temporal Bradesco ao longo do período em análise, observa-se uma tendência de alta na primeira parte da série e na sequência uma mudança de regime nesses preços, durante todo o ano de 2019, embora a última parte da série tenha voltado ao patamar a esta mudança de regime.

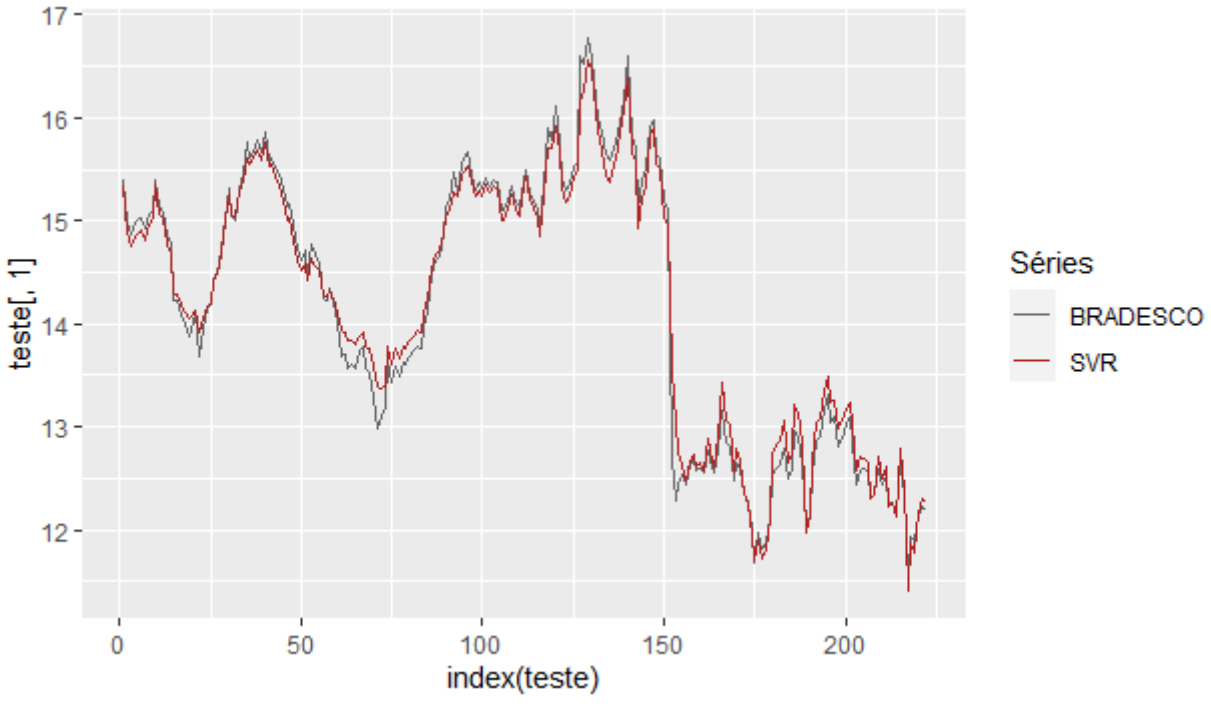
Figura 3.3 – Evolução da série mensal de preços do Bradesco no período 02/01/2015 a 01/02/2023.



Fonte: Próprio autor (2023)

A Figura 3.4 mostra as previsões por parte do conjunto de teste do conjunto de dados.

Figura 3.4 – Evolução no tempo das séries temporais dos preços Bradesco com SVR.

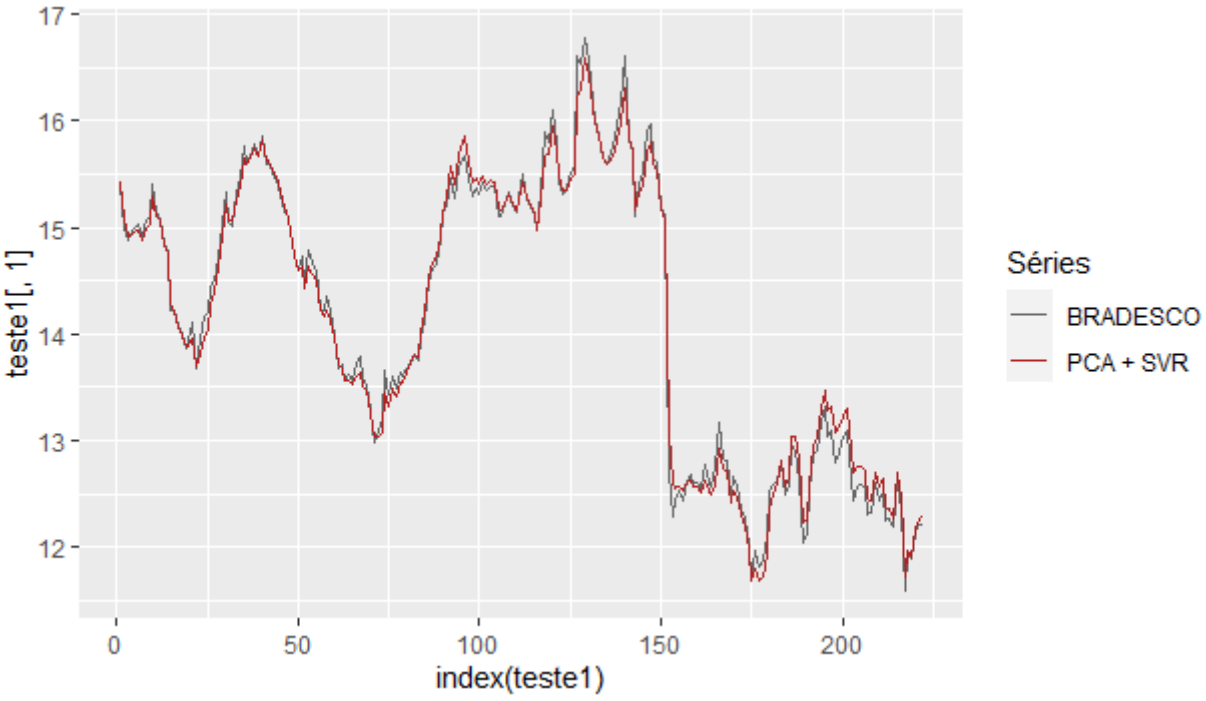


Fonte: Próprio autor (2023)

Percebe-se que os dados de teste se aproximam bem da série original.

A Figura 3.5 apresenta resultados referente à aplicação do PCA - SVR, nota-se que a aplicação do PCA seguida do SVR, ajusta-se bem aos conjunto de dados teste preditos. Adiante, verifica-se RMSE menor, na Tabela 3.1.

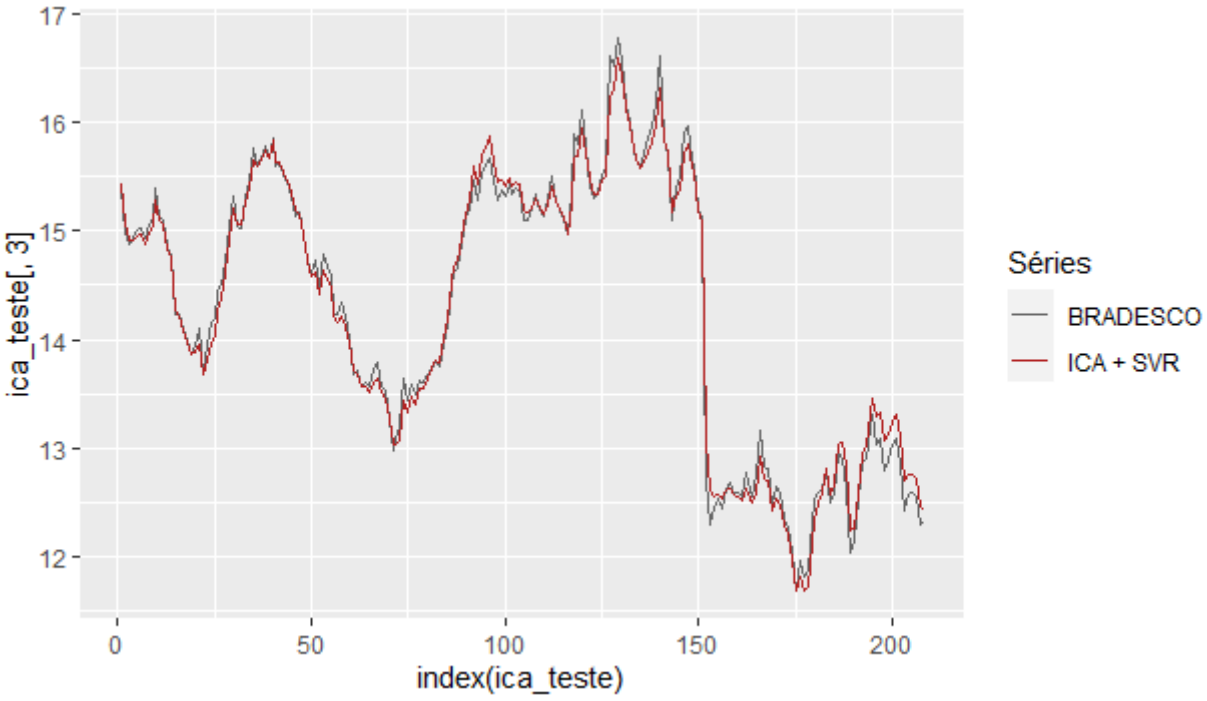
Figura 3.5 – Evolução no tempo das séries temporais dos preços Bradesco com PCA - SVR.



Fonte: Próprio autor (2023)

A Figura 3.6 mostra resultados referente a aplicação do ICA - SVR, nota-se que a aplicação da ICA seguida do SVR, ajusta-se com menor aproximação aos conjunto de dados teste preditos, adiante quando se olha para o R^2 do PCA - SVR 3.1 veremos o quão o modelo ajusta-se aos dados originais.

Figura 3.6 – Evolução no tempo das séries temporais dos preços Bradesco com ICA - SVR.



Fonte: Próprio autor (2023)

A Tabela 3.1 refere-se aos resultados da pesquisa para os modelos de previsão propostos. Como resultado, observa-se que o PCA - SVR obteve melhor desempenho para a série de dados Bradesco, pontuando métricas de desempenho melhores avaliadas, quando comparando os modelos de previsão propostos.

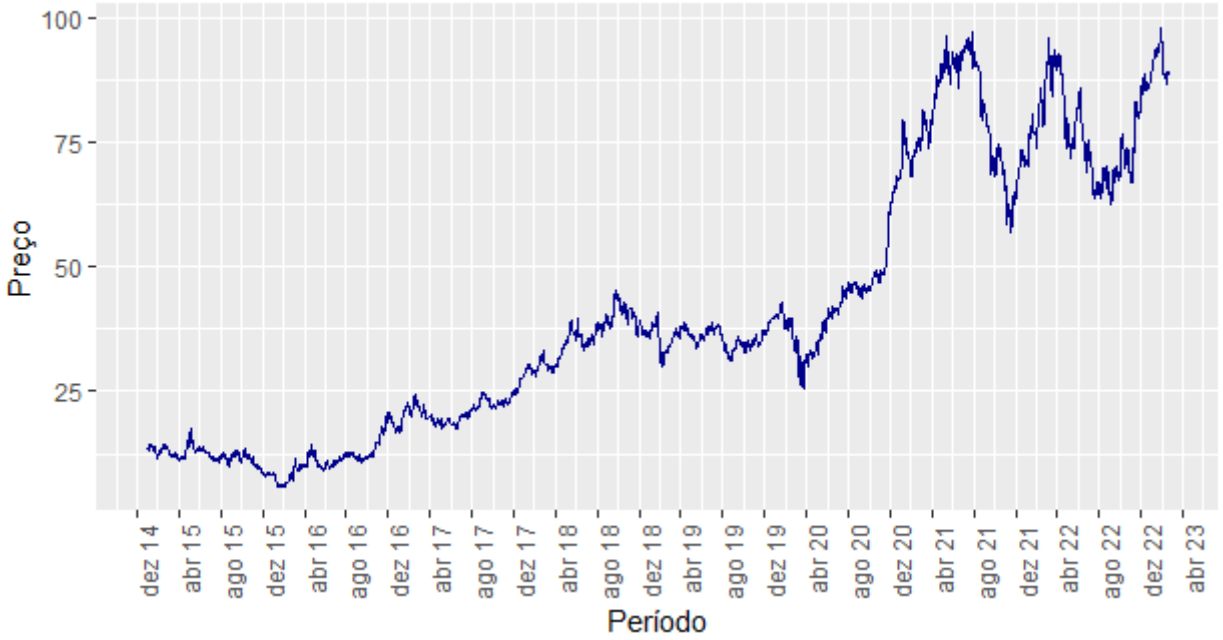
Tabela 3.1 – Resultados da pesquisa de grade para parâmetros de Kernel Radial - Bradesco.

Modelos de previsão	Métricas de desempenho			
	MAE	MSE	RMSE	R^2
SVR	0,12629	0,02564	0,16014	0,98129
PCA - SVR	0,08858	0,01317	0,11476	0,99161
ICA -SVR	0,08960	0,01340	0,11577	0,99148

Fonte: Próprio autor (2023).

A Figura 3.7 apresenta o comportamento da série temporal Vale ao longo do período em análise, observa-se uma volatilidade razoalvelmente similiar, exceto a partir de outubro de 2020, quando a mesma apresentou picos de alta e baixa com amplitudes maiores.

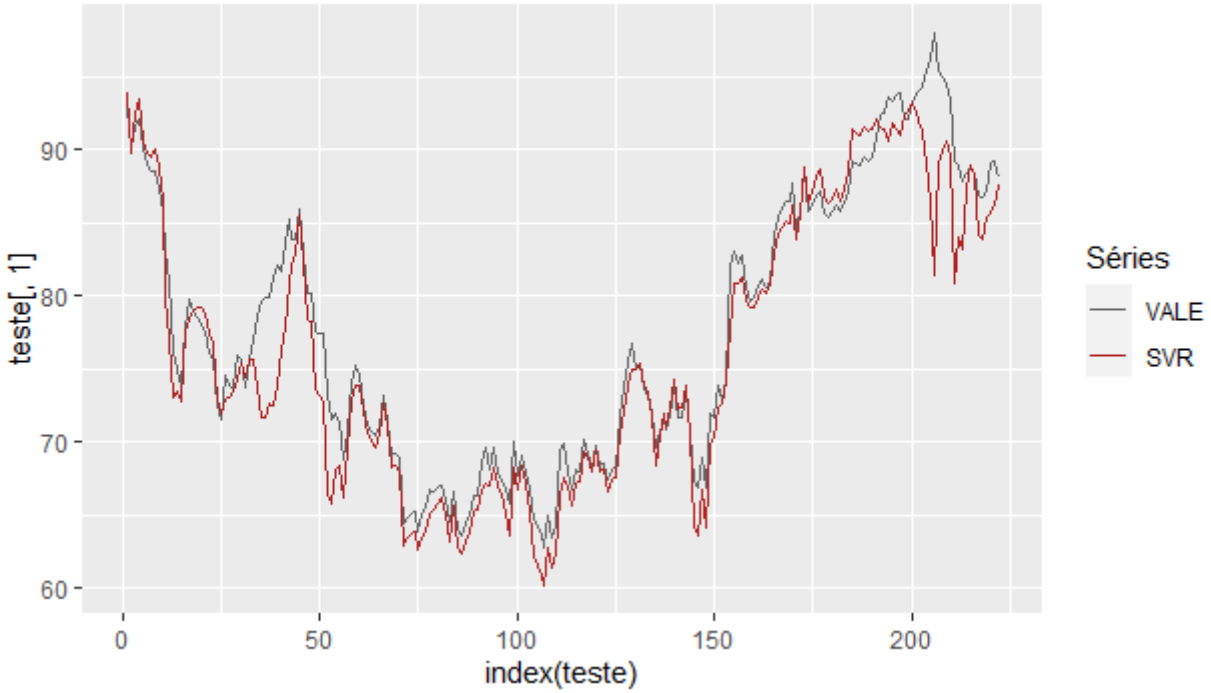
Figura 3.7 – Evolução da série mensal de preços da Vale no período 02/01/2015 a 01/02/2023.



Fonte: Próprio autor (2023)

A Figura 3.8 mostra as previsões por parte do conjunto de teste a série histórica da Vale.

Figura 3.8 – Evolução no tempo das séries temporais dos preços Vale com SVR.



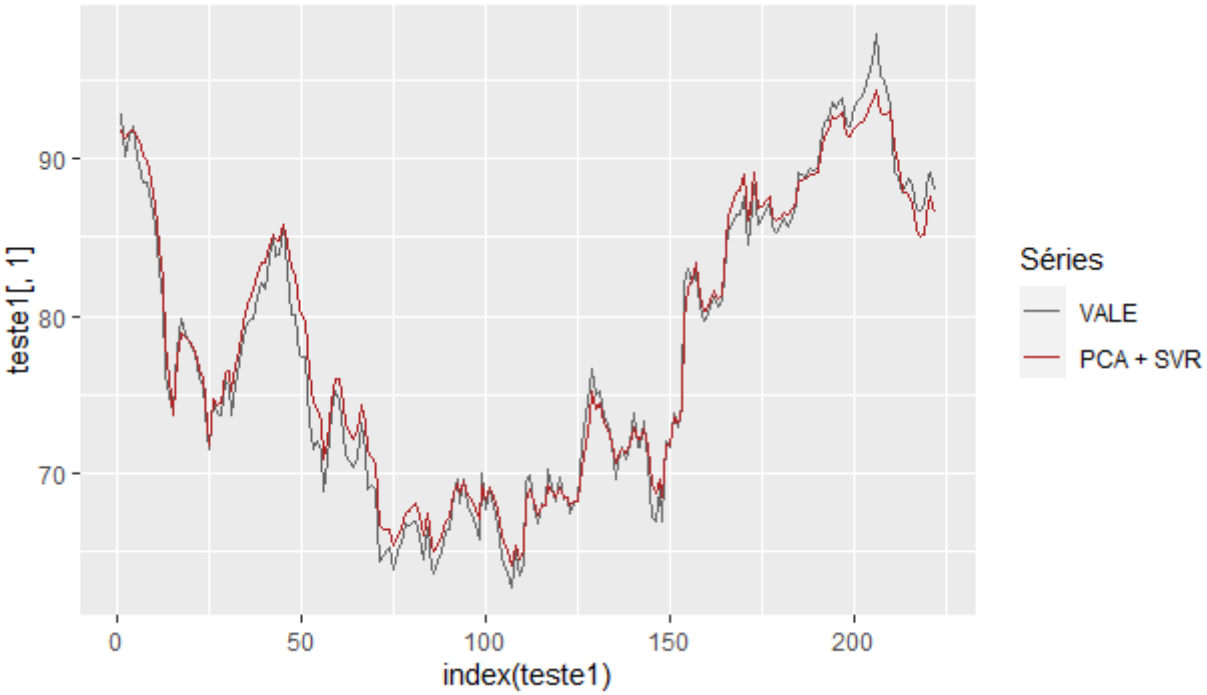
Fonte: Próprio autor (2023)

Observa-se que o SVR, apresenta um desempenho de maior qualidade no ínicio da série, tendo uma sequência de queda de aproximação na sequência da mesma, até seu fim de teste.

Nesse sentido, múltiplos pontos podem ser onde todos os indicadores capturaram o mesmo tipo de informação, mas o SVR não a capta (em particular porque os indicadores, apesar do mesmo tipo de informação, podem ter movimentos opostos), produzindo um RMSE superior, vide Tabela 3.2.

A Figura 3.9 apresenta resultados referente a aplicação do PCA - SVR por parte do conjunto de teste do conjunto de dados, observa-se que o ajuste está alinhado desde no ínicio da série, apresentando uma pequena variação entre as 50 e 150 observações.

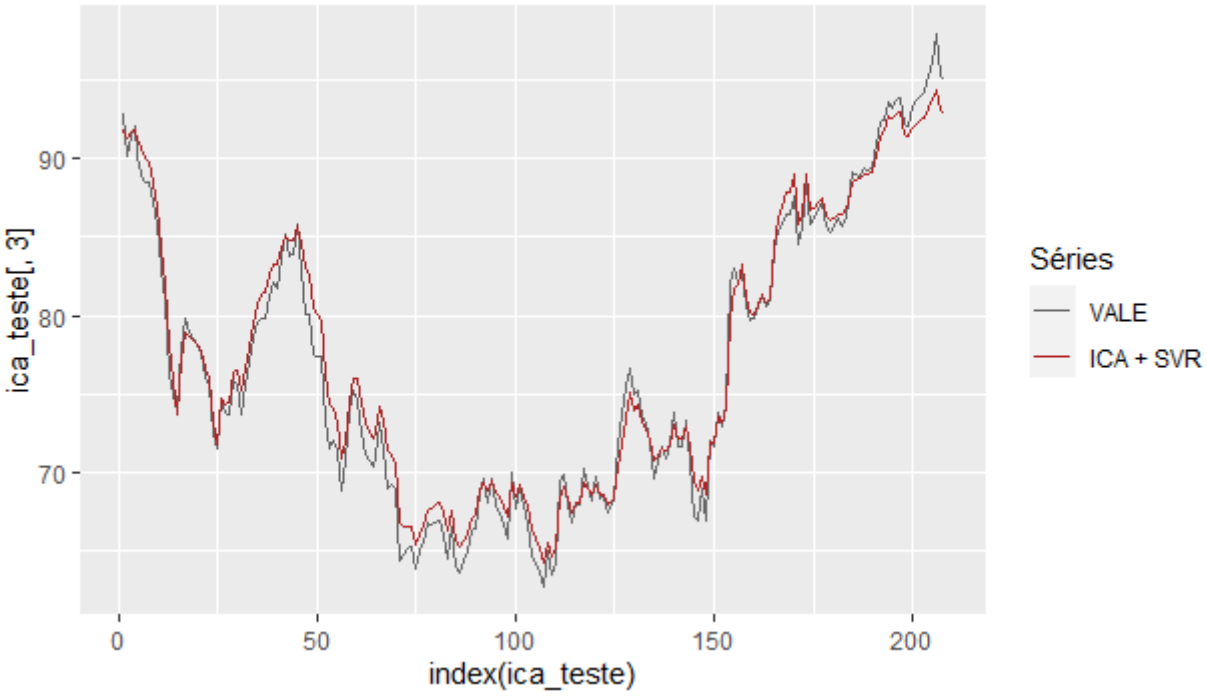
Figura 3.9 – Evolução no tempo das séries temporais dos preços Vale com PCA - SVR.



Fonte: Próprio autor (2023)

A Figura 3.10 apresenta resultados referente a aplicação do ICA - SVR por parte do conjunto de teste do conjunto de dados, nota-se que o modelo apresenta melhor ajuste que os anteriores, a se confirmar pelo R^2 da Tabela 3.2.

Figura 3.10 – Evolução no tempo das séries temporais dos preços Vale com ICA - SVR.



Fonte: Próprio autor (2023)

Confirma-se assim, que as variáveis dependentes são melhores explicadas pela variáveis independentes no modelo de previsão, onde 98% dos movimentos da série podem ser explicados pelos componentes independentes, utilizados no modelo.

A Tabela 3.2 refere-se aos resultados da pesquisa para os modelos de previsão propostos.

Tabela 3.2 – Resultados da pesquisa de grade para parâmetros de Kernel Radial - Vale.

Modelos de previsão	Métricas de desempenho			
	MAE	MSE	RMSE	R^2
SVR	1,74417	7,04953	2,65509	0,92077
PCA - SVR	1,00606	1,52463	1,23475	0,98057
ICA -SVR	0,99746	1,49310	0,98079	0,98079

Fonte: Próprio autor (2023)

As métricas de desempenho ICA - SVR, apresentaram melhores resultados, quando comparado ao outros modelos, isso é, produz claramente menor erros de previsão do que outras abordagens.

Utilizou-se ainda, a implementação do algoritmo FastICA (pacote de análise presente no software R, para realizar Análise de Componentes Independentes (ICA). Sob este modelo generativo, os 'sinais' medidos em X tenderão a ser 'mais gaussianos' do que os componentes da fonte (em S) devido ao Teorema do Limite Central. Assim, para extrair as fontes/componentes independentes buscamos uma matriz de desmistura W que maximize a não gaussianidade das fontes.

4 CONCLUSÃO

A dissertação mostra como diferentes modelos de previsão associados ao SVR podem ser construídos em meio aos mais comumente usados na análise de dados financeiros. Testamos por meio de indicadores técnicos, para duas séries temporais (Bradesco e Vale), a máquina de suporte vetorial, especificamente SVR para fazer a previsão de preço e avaliar a importância desta funcionalidade por meio de PCA e ICA.

Os resultados mostram que para a série Bradesco o PCA - SVR, bem como o ICA - SVR tiverão melhores desempenhos, ambos com resultados de métricas bem próximos; em contrapartida para a série Vale o ICA - SVR com melhor destaque.

Nesse sentido, podemos acreditar que a representação por meio de um modelo baseado unicamente em SVR, tem menor desempenho quando não atrelado a outros indicadores de dados financeiros ou técnicas estatísticas de redução ou investigação de dimensionalidade dos dados. A influência das características do PCA e ICA, podem verificar, pelo menos nos casos estudados, que a representação da dissertação, onde a modelagem obteve bons resultados de previsão aplicando-as em sequência afim de elevar o desempenho do SVR na previsão do preço das ações. Como trabalho futuro, destacamos a possível aplicação do PCA - ICA - SVR, conjuntamente, para verificar a melhoria e o comportamento de séries financeiras considerando as métricas e indicadores técnicos necessários.

REFERÊNCIAS

BEN-HUR, A. et al. Support vector clustering. **Journal of machine learning research**, v. 2, n. Dec, p. 125–137, 2001.

CAMPOS, B. A. R. d. M. et al. Análise comparativa de técnicas para a previsão de séries temporias no contexto de mercado financeiros. Florianópolis, SC, 2021.

CHEN, C.; DAVIS, R. A.; BROCKWELL, P. J. Order determination for multivariate autoregressive processes using resampling methods. **Journal of multivariate analysis**, Elsevier, v. 57, n. 2, p. 175–190, 1996.

CHERVONENKIS, A.; VAPNIK, V. Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data(average risk minimization based on empirical data, showing relationship of problem to uniform convergence of averages toward expectation value). **Automation and Remote Control**, v. 32, p. 207–217, 1971.

COMON, P. Independent component analysis, a new concept? **Signal processing**, Elsevier, v. 36, n. 3, p. 287–314, 1994.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, n. 3, p. 273–297, 1995.

COVER, T. M. **Elements of information theory**. [S.l.]: John Wiley & Sons, 1999.

DARMOIS, G. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. **Revue de l'Institut international de statistique**, JSTOR, p. 2–8, 1953.

EHLERS, R. S. Análise de séries temporais. **Laboratório de Estatística e Geoinformação. Universidade Federal do Paraná**, v. 1, p. 1–118, 2007.

FAN, A.; PALANISWAMI, M. Stock selection using support vector machines. In: IEEE. **IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)**. [S.l.], 2001. v. 3, p. 1793–1798.

FERNANDO, S.; OLIVEIRA, I. Estudo de séries temporais na análise em componentes principais e na análise em componentes independentes. In: **XVI Congresso Anual da Sociedade Portuguesa de Estatística**. [S.l.: s.n.], 2008.

FERRERO, C. A. **Algoritmo kNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia**. Tese (Doutorado) — Universidade de São Paulo, 2009.

HERAULT, J.; JUTTEN, C. Space or time adaptive signal processing by neural network models. In: AMERICAN INSTITUTE OF PHYSICS. **AIP conference proceedings**. [S.l.], 1986. v. 151, n. 1, p. 206–211.

JAMES, G. et al. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112.

JUDGE, G. et al. **Principles of Econometrics**. [S.l.]: John Wiley and Sons, New York, 1988.

KARUSH, W. Minima of functions of several variables with inequalities as side conditions. In: **Traces and Emergence of Nonlinear Programming**. [S.l.]: Springer, 2014. p. 217–245.

KAUFMAN, P. J. **Alpha trading: profitable strategies that remove directional risk**. [S.l.]: John Wiley & Sons, 2011.

KUHN, H.; TUCKER, A. Nonlinear programming in proceedings of 2nd berkeley symposium (pp. 481–492). **Berkeley: University of California Press.[Google Scholar]**, 1951.

LATHAUWER, L. D.; MOOR, B. D.; VANDEWALLE, J. An introduction to independent component analysis. **Journal of Chemometrics: A Journal of the Chemometrics Society**, Wiley Online Library, v. 14, n. 3, p. 123–149, 2000.

LATORRE, M. d. R. D. d. O.; CARDOSO, M. R. A. Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos. **Revista Brasileira de Epidemiologia**, SciELO Brasil, v. 4, p. 145–152, 2001.

LORENA, A. C.; CARVALHO, A. C. D. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007.

MALETZKE, A. G. **Uma metodologia para extração de conhecimento em séries temporais por meio da identificação de motifs e da extração de características**. Tese (Doutorado) — Universidade de São Paulo, 2009.

MATSUOKA, K.; OHOYA, M.; KAWAMOTO, M. A neural net for blind separation of nonstationary signals. **Neural networks**, Elsevier, v. 8, n. 3, p. 411–419, 1995.

MORETTIN, P.; TOLOI, C. Análise de séries temporais—2ª edição revista e ampliada. **ABE–Projeto Fisher, Editora Edgar Blücher**, 2006.

MORETTIN, P. A. **Econometria financeira: um curso em séries temporais financeiras**. [S.l.]: Editora Blucher, 2017.

MORETTIN, P. A.; SINGER, J. M. Estatıstica e ciência de dados. **Texto Preliminar, IME-USP**, 2021.

ORD, J. **Characterization problems in mathematical statistics**. [S.l.]: Wiley Online Library, 1975.

PAJUNEN, P. Blind source separation using algorithmic information theory. **Neurocomputing**, Elsevier, v. 22, n. 1-3, p. 35–48, 1998.

PAJUNEN, P. Extensions of linear independent component analysis: Neural and information theoretic methods. 2000.

QUEJI, L. M.; CAETANO, V. H. S. Operando na bolsa de valores de são paulo com a utilização da análise gráfica: setup da média móvel exponencial de 9 períodos. **Revista ADMPG**, v. 4, n. 2, 2011.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

REGAZZI, A. J. Análise multivariada, notas de aula inf 766. **Departamento de Informática da Universidade Federal de Viçosa**, v. 2, 2000.

RISSANEN, J. Modeling by shortest data description. **Automatica**, Elsevier, v. 14, n. 5, p. 465–471, 1978.

RISSANEN, J. A universal prior for integers and estimation by minimum description length. **The Annals of statistics**, Institute of Mathematical Statistics, v. 11, n. 2, p. 416–431, 1983.

RUBIO, G. et al. A heuristic method for parameter selection in ls-svm: Application to time series prediction. **International Journal of Forecasting**, Elsevier, v. 27, n. 3, p. 725–739, 2011.

SHARIFI, M.; SOURI, A. A hybrid ls-he and ls-svm model to predict time series of precipitable water vapor derived from gps measurements. **Arabian Journal of Geosciences**, Springer, v. 8, n. 9, p. 7257–7272, 2015.

SICSU, A. L.; DANA, S. **Estatística aplicada**. [S.l.]: Saraiva Educação SA, 2017.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and computing**, Springer, v. 14, n. 3, p. 199–222, 2004.

TAY, F. E.; CAO, L. Application of support vector machines in financial time series forecasting. **omega**, Elsevier, v. 29, n. 4, p. 309–317, 2001.

VAPNIK, V. N.; CHERVONENKIS, A. Y. Necessary and sufficient conditions for the uniform convergence of means to their expectations. **Theory of Probability & Its Applications**, SIAM, v. 26, n. 3, p. 532–553, 1982.

ZHAO, X.; MA, Z.; YIN, M. Using support vector machine and evolutionary profiles to predict antifreeze protein sequences. **International journal of molecular sciences**, Molecular Diversity Preservation International (MDPI), v. 13, n. 2, p. 2196–2207, 2012.

APÊNDICE A –

Apresentamos abaixo, alguns indicadores técnicos, também utilizados no mercado financeiro como forma de avaliar a movimentação de preços. Os mesmos, servem para mostrar a quantidade existente e traz seus seguintes conceitos.

1. Taxa de Mudança - ROC. O indicador de taxa de mudança (ROC) é um oscilador de momentum. Calcula a variação percentual no preço entre os períodos. O ROC leva o preço atual e compara-o com um período anterior "n" de preço (definido pelo usuário). O valor calculado é então plotado e flutua acima e abaixo de uma linha zero. Um analista técnico pode usar a taxa de mudança (ROC) para identificação de tendências e identificar condições de sobrecompra e sobrevenda.

$$\text{ROC} = \frac{\text{Preço de fechamento hoje} - \text{Preço de fechamento } n \text{ períodos atrás}}{\text{Preço de fechamento } n \text{ períodos atrás}} \times 100. \qquad (1)$$

2. Momentum. Os indicadores de momento nos ajudam a analisar a taxa de velocidade na qual os preços das ações caem ou sobem. A fórmula para este indicador compara o preço de fechamento mais recente com o preço de fechamento anterior de qualquer período de tempo.

$$\text{Momentum} = \frac{\text{Preço atual}}{\text{Preço no período anterior}} \times 100 \qquad (2)$$

O indicador de momento é mostrado como uma única linha abaixo do gráfico de preço do que na linha de preço ou nas barras.

3. Média exponencial tripla - TRIX. É um indicador de momento usado por traders técnicos que mostra a mudança percentual em uma média móvel que foi suavizada exponencialmente três vezes. A suavização tripla das médias móveis foi projetada para filtrar movimentos de preços considerados insignificantes ou sem importância. O TRIX também é implementado por traders técnicos para produzir sinais de natureza semelhante à divergência de convergência da média móvel (MACD).

$$\text{TRIX} = 3 \times \text{EMA} - 3 \times \text{EMA}(\text{EMA}) + \text{EMA}(\text{EMA}(\text{EMA})). \qquad (3)$$

4. Oscilador estocástico (K). O oscilador estocástico foi elaborado pelo americano George C. Lane na década de 1950.

$$K = \frac{\text{Preço atual} - \text{menor preço selecionado}}{\text{Maior preço selecionado} - \text{Menor preço selecionado}} \times 100 \qquad (4)$$

Trata-se de um indicador desenvolvido para mostrar a relação entre o preço de fechamento de uma ação e suas máximas e mínimas durante certo período de tempo.

Os investidores o utilizam para tentar prever os movimentos nos preços de mercado. Até hoje o estocástico é usado para tentar definir as direções dos preços, além de antever padrões de alta e de baixa para se identificar reversões e rompimentos.

5. On balance volume - OBV. É um indicador técnico de tendência, que usa o fluxo de volume para prever mudanças no preço das ações. Joseph Granville desenvolveu essa métrica pela primeira vez no seu livro de 1963, Granville's New Key to Stock Market Profits.

Se o preço de fechamento de hoje é maior do que o de ontem, então:

$$\text{OBV} = \text{OBV de ontem} + \text{Volume de hoje} \tag{5}$$

Se o preço de fechamento de hoje é menor do que o de ontem, então:

$$\text{OBV} = \text{OBV de ontem} - \text{Volume de hoje} \tag{6}$$

Se o preço de fechamento de hoje é igual do que o de ontem, então:

$$\text{OBV} = \text{OBV de ontem.} \tag{7}$$

O OBV mostra o sentimento dos investidores, podendo prever um movimento de alta ou baixa.

Comparar o movimento dos preços com o OBV gera sinais mais confiáveis do apenas utilizandoos histogramas de volume verde ou vermelho comumente encontrados na parte inferior dos gráficos de preços.

6. Índice direcional médio - ADX. É um indicador de análise técnica, usado para saber se os preços estão em tendência ou faixa e para medir a força da tendência.

$$\text{ADX} = (\text{ADX a priori} \times 13) + \text{DX atual}/14 \tag{8}$$

Do tipo oscilador não direcional, ou seja, quantifica a força de uma tendência independentemente de sua direção. Em espanhol ADX significa índice de direção média. É comum usá-lo em conjunto com Indicadores de Movimento Direcional (DMI), que nos mostram a tendência de mercado prevalecente.

7. Indicador SAR Parabólico - SAR. O indicador SAR parabólico , desenvolvido por J. Wells Wilder, é usado por traders para determinar a direção da tendência e possíveis reversões de preço. O indicador usa um método trailing stop e reverso chamado "SAR", ou stop and reverse, para identificar pontos de saída e entrada adequados . Os comerciantes também se referem ao indicador como o stop e reverso parabólicos, SAR parabólico ou PSAR. Para calcular o SAR Parabólico de hoje, você precisará conhecer o preço mais extremo (EP), o fator de aceleração (AF), bem como o PSAR mais recente. Você também precisará determinar se existe atualmente uma tendência de alta ou uma tendência de baixa.

$$\text{PSAR} = \text{Prior PSAR} + \text{Prior AF}(\text{Prior EP} - \text{Prior PSAR}) \tag{9}$$

para tendências acima;

$$\text{PSAR} = \text{Prior PSAR} - \text{Prior AF}(\text{Prior PSAR} - \text{Prior EP}) \tag{10}$$

para tendências abaixo.

O indicador SAR parabólico aparece em um gráfico como uma série de pontos, acima ou abaixo do preço de um ativo, dependendo da direção em que o preço está se movendo. Um ponto é colocado abaixo do preço quando está em tendência de alta e acima do preço quando está em tendência de queda.

8. Money Flow Index - MFI . O Money Flow Index (MFI) é um oscilador técnico que usa dados de preço e volume para identificar sinais de sobrecompra ou sobrevenda em um ativo. Também pode ser usado para detectar divergências que alertam para uma mudança de tendência no preço. O oscilador se move entre 0 e 100. Ao contrário dos osciladores convencionais, como o Índice de Força Relativa (RSI), o Índice de Fluxo de Dinheiro incorpora dados de preço e volume, em vez de apenas preço. Por esse motivo, alguns analistas chamam o MFI de RSI ponderado por volume.

$$\text{Razão fluxo dinheiro} = 100 - \frac{100}{1 + \text{Razão fluxo dinheiro}} \tag{11}$$

em que:

$$\text{Razão fluxo dinheiro} = \frac{\text{14 fluxo positivo de período dinheiro}}{\text{14 fluxo negativo de período dinheiro}} \text{Fluxo dinheiro bruto} = \text{Típico preco} * \text{Volume}$$

9. Kelter Channels. Criado com o intuito de aproveitar as oportunidades criadas pela volatilidade dos preços, Keltner Channels é um indicador bastante utilizado, em especial no mercado americano. Esta técnica guarda semelhanças com as conhecidas Bandas de Bollinger , uma vez que consiste em duas linhas flutuantes calculadas a partir de um valor central médio.

A técnica foi desenvolvida pelo operador de grãos de Chicago Chester W. Keltner. O autor a apresentou ao mundo em seu livro de 1960, *How to Make Money in Commodities?*. Desde então surgiram diferentes variações para os sinais, a maioria empregando médias móveis exponenciais (não usadas originalmente pelo autor). Ainda assim, todas essas modificações são agrupadas na literatura especializada sob a denominação de Keltner Channels.

10. Bollinger Bands. É uma ferramenta de negociação técnica criada por John Bollinger no início dos anos 80. Eles surgiram da necessidade de bandas de negociação adaptáveis e da observação de que a volatilidade era dinâmica, não estática como se acreditava na época.

As Bandas de Bollinger podem ser usadas na maioria dos períodos de tempo, desde períodos de muito curto prazo até por hora, diariamente, semanalmente ou mensalmente.

As bandas de Bollinger respondem a uma pergunta: os preços são altos ou baixos em uma base relativa? Por definição, o preço é alto na banda superior e o preço é baixo na banda

inferior. Esse pedaço de informação é incrivelmente valioso. É ainda mais poderoso se combinado com outras ferramentas, como outros indicadores para confirmação.

11. Volatilidade de Chaikin - CHV. É um indicador que foi desenvolvido por Marc Chaikin. Ele mede a volatilidade olhando as máximas e mínimas dos preços em um período de tempo específico que pode ser configurado. O cálculo exato envolve uma Média Móvel Exponencial (EMA), mas em geral: quanto mais o intervalo entre máximas e mínimas se alarga, maior volatilidade dos preços. O indicador pode ser usado de várias maneiras, quer para prever uma potencial mudança de tendência (se a volatilidade diminui por um longo período de tempo) ou para avaliar o risco (se a volatilidade aumenta de repente, apontando para traders nervosos). O CHV é frequentemente utilizado em combinação com outros sinais e técnicas de análise.

12. Ehler's Correlation Trend Indicator. Indicador de tendência de correlação é um estudo que estima a direção atual e a força de uma tendência. Ele pode ser usado para detectar surtos de tendências ou exaustão. O Indicador de Tendência de Correlação usa a correlação de Spearman ao estimar quão próximo o comportamento dos preços de fechamento se correlaciona com uma linha reta de inclinação positiva. Isso significa que:

- 1: Valores próximos a +1,0 significam condições de tendência de alta.
- 2: Valores próximos a -1,0 significam condições de tendência de baixa.
- 3: Valores em torno de 0,0 significam condições de tendência lateral.

13. Donchian Channel - DC. São utilizados na análise técnica para medir a volatilidade do mercado. Assim como o indicador criado por Bollinger, também é composto por bandas. Além de medir a volatilidade de um mercado, os Canais de Donchian são utilizados para identificar possíveis quebras e condições de sobrecompra/sobrevenda quando o preço atinge a banda superior ou inferior que podem ser negociadas.

14. Oscilador intermediário DV - DVI.É um oscilador de momento muito suave que também pode ser usado como um indicador de tendência. Criado por David Varadi. O DVI combina retornos suavizados em diferentes janelas de tempo e o número relativo de dias de alta versus de baixa (estiramento) em diferentes janelas de tempo.

15. Arms' Ease of Movement Value - EMV. O indicador de facilidade de movimento (EOM ou EMV) de Richard Arms é um estudo técnico que tenta quantificar uma combinação de informações de momento e volume em um valor.

A intenção é usar esse valor para discernir se os preços podem subir, ou cair, com pouca resistência no movimento direcional.

Teoricamente, se os preços se moverem facilmente, eles continuarão a fazê-lo por um período de tempo que pode ser negociado de forma eficaz.

O indicador Ease of Movement mostra a relação entre preço e volume e é frequentemente usado para avaliar a força de uma tendência subjacente.

A facilidade de movimento calcula a facilidade com que um preço pode subir ou descer, com base no momento.

O cálculo subtrai o preço médio de ontem do preço médio de hoje e divide a diferença pelo volume.

16. Média Móvel Múltipla do Guppy - GMMA. É um indicador técnico que visa antecipar uma possível quebra no preço de um ativo. O termo recebe o nome de Daryl Guppy, um colunista financeiro australiano e autor de livros que desenvolveu o conceito em seu livro, "Trading Tactics".

O GMMA usa a média móvel exponencial (EMA) para capturar a diferença entre preço e valor em uma ação. Uma convergência desses fatores está associada a uma mudança significativa de tendência. Guppy sustenta que o GMMA não é um indicador de atraso, mas um aviso prévio de uma mudança em desenvolvimento no preço e no valor.

17. Volatilidade. Os indicadores baseados em volatilidade são valiosas ferramentas de análise técnica que analisam as mudanças nos preços de mercado durante um período de tempo especificado. Quanto mais rápido os preços mudam, maior a volatilidade. Quanto mais lentos os preços mudam, menor a volatilidade. Pode ser medido e calculado com base nos preços históricos e pode ser usado para identificação de tendências. Também normalmente indica se um mercado está sobrecomprado ou sobrevendido (o preço significa que é injustificadamente alto ou injustificadamente baixo), o que pode significar uma paralisação ou reversão da tendência.

Exemplos de tais indicadores são Média de Amplitude de Variação(ATR), as muito populares e fáceis de usar Bandas de Bollinger (BB), Canais de Donchian e Canais de Keltner (KC).