UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE GRADUAÇÃO EM ESTATÍSTICA

Natália Viviane Silva Reis

APLICAÇÕES DA DISTRIBUIÇÃO T^2 DE HOTELLING EM RECONHECIMENTO DE ASSINATURAS

Recife

2025

Natália Viviane Silva Reis

APLICAÇÕES DA DISTRIBUIÇÃO T^2 DE HOTELLING EM RECONHECIMENTO DE ASSINATURAS

Trabalho de Conclusão de Curso apresentado à Universidade Federal de Pernambuco como parte das exigências para obtenção do título de bacharel em Estatística

Orientador (a): Prof° Dr° Manoel Raimundo de Sena Jr.

Recife

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Reis, Natália Viviane Silva .
Aplicações da distribuição T^2 de Hotelling em reconhecimento de assinaturas / Natália Viviane Silva Reis. - Recife, 2024.
31p.

Orientador(a): Manoel Raimundo de Sena Júnior
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Estatística - Bacharelado, 2024.
Inclui referências.

1. Reconhecimento de assinaturas. 2. Distribuição T2 de Hotelling. I. Sena Júnior, Manoel Raimundo de. (Orientação). II. Título.

310 CDD (22.ed.)

NATÁLIA VIVIANE SILVA REIS

APLICAÇÕES DA DISTRIBUIÇÃO T^2 DE HOTELLING EM RECONHECIMENTO DE ASSINATURAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovado em: 09/10/2024

BANCA EXAMINADORA

_____

Prof. Dr. Manoel Raimundo de Sena Júnior (Orientador)

Universidade Federal de Pernambuco

_____

Prof. Dr. Abraão David Costa do Nascimento (Examinador Interno)

Universidade Federal de Pernambuco

_____

Profa. Dra. Audrey Helen Mariz de Aquino Cysneiros (Examinador Interno)

Universidade Federal de Pernambuco

Dedico este trabalho a Deus, aos meus pais, Anacleto e Ildaci, pela formação do meu caráter, pelo incentivo e apoio constantes. Aos meus irmãos e todos que me ajudaram ao longo desta caminhada, contribuindo de alguma forma para conclusão desse projeto.

AGRADECIMENTOS

Agradeço em primeiro lugar, a Deus, por ter permitido que eu tivesse saúde e determinação para não desanimar, fazendo com que meus objetivos fossem alcançados, durante todos os meus anos de estudos.

À minha família, por toda a minha formação pessoal e que me incentivaram nos momentos difíceis, especialmente aos meus pais, Anacleto e Ildaci que sempre batalharam incansavelmente por mim e por meus irmãos.

Ao professor Manoel, por ter sido meu orientador e ter desempenhado tal função com dedicação, pela ajuda e pela paciência com a qual guiaram o meu aprendizado.

Por fim, a todos aqueles que contribuíram, de alguma forma, para a realização deste trabalho.

"Penso que cumprir a vida
Seja simplesmente
Compreender a marcha
E ir tocando em frente".
(Almir Sater/ Renato
Teixeira, 1990)

RESUMO

O presente trabalho tem como principal objetivo realizar a análise entre assinaturas verdadeiras e falsas, através de métodos estatísticos com a aplicação da distribuição T^2 Hotelling. Inicialmente, realizou-se os ajustes dos dados com o uso da transformada Spline e do método DTW (Dynamic Time Warping) com isso obtendo médias. Em seguida, calculou-se a média entre as assinaturas que serão utilizadas como referência, para construção do teste da distribuição T^2 de Hotelling. Por fim, é levada em consideração a construção da regra de decisão com a identificação dos pontos críticos. Com o objetivo de mostrar a não rejeição de uma assinatura verdadeira e rejeição de uma assinatura falsa.

Palavras-chaves: Assinaturas, distância, distribuição T^2 de Hotelling, DTW, falsa, spline, verdadeira.

ABSTRACT

The main objective of this work is to carry out the analysis between true and false signatures, using statistical methods using the T^2 Hotelling distribution. Initially, data adjustments were made using the Spline transform and the DTW (Dynamic Time Warping) method, obtaining averages. Then, the average of the signatures that will be used as a reference was calculated to construct the Hotelling T^2 distribution test. Finally, the construction of the decision rule is taken into account with the identification of critical points. With the aim of showing the non-rejection of a true signature and rejection of a false signature.

Keywords: Signatures, distance, Hotelling T^2 distribution, DTW, false, spline, true.

LISTA DE SÍMBOLOS

β	Beta - representa uma Variável Aleatória com distribuição F;
Σ	Sigma - matriz de covariância;
μ	Mu - (é um número real) representa a média da distribuição normal;
$\top$	Transposto;
$\sim$	Til - usado para denotar que uma variável "tem uma determinada distribuição de probabilidade";
α	Alfa - representa uma Variável Aleatória segue com distribuição T^2 de Hotelling;

SUMÁRIO

1 INTRODUÇÃO

No mundo, o uso de assinaturas são comumente usadas como forma de aprovar autenticação de documentos ou para provar o recebimento de encomendas. Verificar uma assinatura na maioria das situações não ocorre, pois levaria muito tempo e esforço para verificar manualmente as assinaturas (JAIN; GRIESS; CONNELL, 2002).

Com o avanço da tecnologia, houve a necessidade de estabelecer um padrão, para o controlar de uma forma mais segura, o processo de autenticação de assinaturas. Apresentando um padrão, em que possibilitará classificar se pertence ou não a aquele padrão predeterminado. Os sistemas de reconhecimentos de padrões é uma técnica em que se permite extrair informações úteis para a tomada de decisões (HEINEN; OSÓRIO, 2005).

Um sistema para o reconhecimento é através da captura de assinaturas, onde pode ser digitalizada em pontos e igualmente espaçada no tempo. Levando em conta a velocidade, caneta, inclinação da mão, papéis diferentes. Considere duas curvas onde deseja-se medir a distância entre elas para saber se pertencem ao mesmo grupo. Uma das formas que podemos fazer isso é através da suavização delas, acreditando que elas são capturadas por meio da digitalização do plano cartesiano medido no tempo. Cada ponto da curva é uma trinca (t, x, y)onde para cada instante de tempo temos um ponto (x, y) no espaço bidimensional.

Através da suavização da curva pela transformação Splines (pacote incluso do $softR$) e comparando ponto a ponto podemos medir essa distância. Como cada curva tem um nó de pontos distintos faz-se necessário o uso da técnica, conhecida como "Time Warping", ou "Dynamic Time Warping- DTW" usada para colocar todas no número fixo de pontos (pacote incluso do $softR$). Em seguida, obter medidas de distâncias entre ela e uma nova, supostamente desconhecidas, e comparadas, ponto a ponto, através de sua distância.

O monitoramento e controle desses processos exigem que tenha características determinadas, onde envolvem correlações entre as variáveis (KIELING et al., 2012). Uma regra para proximidade delas, por exemplo, se pelo menos 95% desses pontos estiverem dentro de um limite, limite este estabelecido pela distribuição T^2 de Hotelling, considerando verdadeira.

A autenticação das assinaturas neste trabalho, deve-se ressaltar a comparação pela utilização de métodos estatísticos, levando em conta que não é uma técnica infalível ao analisá-las, assim espera-se que esse processo tenha uma taxa de resposta adequada.

2 OBJETIVOS

2.1 OBJETIVO GERAL

O objetivo geral é realizar uma análise conforme as entradas das assinaturas, fazendo uma simples leitura das assinaturas com a utilização de uma aproximação linear. Levando em conta a distância entre os pontos. Obtido pelo produto interno do vetor de desvios e obtendo um nível de confiança para o vetor de desvio.

2.2 OBJETIVO ESPECÍFICO

Gerar modelos, aplicar os métodos à análise de dados e investigar com o uso da:

- Modelar e suavizar as curvas através do uso do spline e do dynamic time warping;
- Encontrar a distância média das assinaturas padrões;
- Calcular os α^* usando a distribuição T^2 de Hotelling;
- Classificá-las a um nível de 95% de confiança, devendo ou não rejeitar uma assinatura;
- Apresentar um resultado da taxa de erro comparando os resultados.

3 METODOLOGIA

A realização desse trabalho tem como base analisar e determinar os objetivos de comparação das curvas. Nesta seção, descreveremos técnicas e métodos utilizados, isso proporcionará uma visão transparente de como foram aplicadas.

3.1 DISTRIBUIÇÃO NORMAL MULTIVARIADA E DISTRIBUIÇÃO DE WISHART

Considere $\mathbf{X}$ um vetor p-variável com média μ e uma matriz de covariância Σ, distribuído segundo uma distribuição normal. Notação $\mathbf{X} \sim N_p(\mu, \Sigma)$.

Considere agora as funções quadráticas do tipo $\mathbf{X}^\top \mathbf{C}\mathbf{X}$, onde $\mathbf{C}$ é uma matriz simétrica e idempotente, e $\mathbf{M}$ uma matriz de dimensão $p \times p$, isto é $\mathbf{M}_{(p\times p)}$, escrita da forma $\mathbf{M} = \mathbf{X}^\top \mathbf{X}$, onde $\mathbf{X}$ é uma matriz $m \times p$, ou seja , $\mathbf{X}_{(m\times p)}$, em que cada linha tem distribuição $N_p(0, \Sigma)$, $\mathbf{X}$ também é chamado de matriz de dados normais, nesse caso $\mathbf{M}$ tem distribuição de Wishart com parâmetros Σ e m. Notação $\mathbf{M} \sim W_p(\Sigma, m)$. Observe que se $\Sigma = I_p$ temos que a distribuição está da forma Wishart padrão. Nota-se ainda que se $\mathbf{M} \sim W_p(\Sigma, m)$ em que m é o número de grau de liberdade da matriz $\mathbf{M}$, fazendo $\mathbf{M}^* = \Sigma^{-1/2}\mathbf{M}\Sigma^{-1/2} = \Sigma^{-1/2}\mathbf{X}^\top\mathbf{X}\Sigma^{-1/2} = \mathbf{Y}^\top\mathbf{Y}$, em que $\mathbf{Y}$ é uma matriz de dados que tem distribuição $N(0, I)$, assim que $\mathbf{M}^*$ tem distribuição Wishart padrão.(MARDIA; KENT; BIBBY, 1979)

3.2 A DISTRIBUIÇÃO T^2 DE HOTELLING

A distribuição T^2 de Hotelling é obtida pelo produto interno do vetor de desvios, obtido pela diferença entre o vetor de médias amostrais e o vetor de médias populacionais, multiplicada pelos números de graus de liberdade da matriz com distribuição Wishart, em termo de notação $\alpha \sim t^2(p, m)$, isto é, $\alpha = md^\top \mathbf{M}^{-1} d$ tendo $d \sim N_p(0, I)$ e $\mathbf{M} \sim W_p(I, m)$, então diremos que α tem distribuição T^2 de Hotteling com parâmetros p e m.

Se $\mathbf{X}$ e $\mathbf{M}$ são independentes e fazendo $\mathbf{X} \sim N_p(\mu, \Sigma)$, $d = \Sigma^{-1/2}(x - \mu)$, $M = nS$ e $\mathbf{M}^* = \Sigma^{-1/2}\mathbf{M}\Sigma^{-1/2}$ onde $S = \frac{1}{n}\mathbf{X}^\top H\mathbf{X}$ e $H = I - \frac{1}{n}J$, além de $J = \mathbf{1}\mathbf{1}^\top$ sendo $\mathbf{1} = [1, 1, ..., 1]^\top$, temos que $d \sim N_p(0, I)$ e $\mathbf{M}^* \sim W_p(I, n-1)$, podemos observar:

$$S_u = \frac{1}{n-1}\mathbf{X}^\top H\mathbf{X}$$

$$= \frac{1}{n-1}\frac{n}{n}\mathbf{X}^\top H\mathbf{X}$$
$$= \frac{n}{n-1}\frac{1}{n}\mathbf{X}^\top H\mathbf{X}$$
$$S_u = \frac{n}{n-1}S \Rightarrow S = \frac{n-1}{n}S_u$$

Nota-se que $\alpha \sim T^2(p,m)$, e fazendo as mesmas suposições para as distribuições de $\mathbf{X}$ e $\mathbf{M}$, respectivamente $N_p(\mu,\Sigma)$ e $M = nS \sim W_p(\Sigma,m)$, podemos reescrever d^* e $\mathbf{M}^*$ da forma padrão, com $d^* = \sqrt{n}\Sigma^{-1/2}(\overline{\mathbf{x}}-\mu)$, $M^* = \Sigma^{-1/2}\mathbf{M}\Sigma^{-1/2}$ e $m = n-1$ assim:

$$\alpha = md^{*\top}\mathbf{M}^{*-1}d^*$$
$$= (n-1)\sqrt{n}\left(\Sigma^{-1/2}(\overline{\mathbf{x}}-\mu)\right)^\top\left(\Sigma^{-1/2}\mathbf{M}\Sigma^{-1/2}\right)^{-1}\sqrt{n}\left(\Sigma^{-1/2}(\overline{\mathbf{x}}-\mu)\right)$$
$$= (n-1)\sqrt{n}\left(\Sigma^{-1/2}(\overline{\mathbf{x}}-\mu)\right)^\top\left(\Sigma^{-1/2}(nS)\Sigma^{-1/2}\right)^{-1}\sqrt{n}\left(\Sigma^{-1/2}(\overline{\mathbf{x}}-\mu)\right)$$
$$= (n-1)\left(\overline{\mathbf{x}}-\mu\right)^\top S^{-1}(\overline{\mathbf{x}}-\mu)$$

Isto é, α tem distribuição $T^2(p,n-1)$.

Na prática vamos estimar μ por $\overline{\mathbf{x}}$, isso irá provocar uma mudança na padronização do d por d^*. Fazendo $d^* = \sqrt{\frac{n}{n+1}}\Sigma^{-1/2}(\mathbf{x}-\overline{\mathbf{x}})$ terá distribuição normal padrão, pois sabe-se que:

- $E(\mathbf{x}-\overline{\mathbf{x}}) = E(\mathbf{x}) - E(\overline{\mathbf{x}}) = 0$;
- $V(\mathbf{x}-\overline{\mathbf{x}}) = V(\mathbf{x}) + V(\overline{\mathbf{x}}) - 2Cov(\mathbf{x},\overline{\mathbf{x}})$;
- $V(\mathbf{x}) = \Sigma$;
- $V(\overline{\mathbf{x}}) = \frac{1}{n}V(\Sigma)$;
- $Cov(\mathbf{x},\overline{\mathbf{x}}) = 0$, pois $\mathbf{x}$ e $\overline{\mathbf{x}}$ são independentes devido ao fato de x ser uma nova observação e não entra no calculo de $\overline{\mathbf{x}}$.

Logo

$$E(d^*) = E\left(\sqrt{\frac{n}{n+1}}\Sigma^{-1/2}(\mathbf{x}-\overline{\mathbf{x}})\right)$$

$$= \sqrt{\frac{n}{n+1}}\Sigma^{-1/2}E(\mathbf{x}-\overline{\mathbf{x}})\Sigma^{-1/2} = 0$$

$$V(d^*) = V\left(\sqrt{\frac{n}{n+1}}\Sigma^{-1/2}(\mathbf{x}-\overline{\mathbf{x}})\right) = \frac{n}{n+1}\Sigma^{-1/2}V(\mathbf{x}-\overline{\mathbf{x}})\Sigma^{-1/2}$$

$$= \frac{n}{n+1}\Sigma^{-1/2}\left[V(\mathbf{x}) + V(\overline{\mathbf{x}}) - 2Cov(\mathbf{x}, \overline{\mathbf{x}})\right]\Sigma^{-1/2}$$

$$= \frac{n}{n+1}\Sigma^{-1/2}\left[V(\mathbf{x}) + V(\overline{\mathbf{x}}) - 0\right]\Sigma^{-1/2}$$

$$= \frac{n}{n+1}\Sigma^{-1/2}\left[\Sigma + \frac{1}{n}\Sigma\right]\Sigma^{-1/2}$$

$$= \frac{n}{n+1}\Sigma^{-1/2}\left[\left(\frac{n+1}{n}\right)\Sigma\right]\Sigma^{-1/2}$$

$$= \frac{n}{n+1}\Sigma^{-1/2}\left[\Sigma^{1/2}\left(\frac{n+1}{n}\right)\Sigma^{1/2}\right]\Sigma^{-1/2}$$

$$= \Sigma^{-1/2}\left[\Sigma^{1/2}\Sigma^{1/2}\right]\Sigma^{-1/2}$$

$$= \Sigma^{-1/2}\Sigma^{1/2}\Sigma^{1/2}\Sigma^{-1/2} = I$$

Aqui $E(\mathbf{M}^*) = E\left(\Sigma^{-1/2}\mathbf{M}\Sigma^{-1/2}\right) = E\left(\Sigma^{-1/2}nS\Sigma^{-1/2}\right) = (n-1)I$.
Se $\mathbf{M} \sim W_p(\Sigma, m)$, então $E(\mathbf{M}) = m\Sigma$. Agora se $M = nS : \mathbf{M} \sim W_p(\Sigma, n-1)$, assim $E(nS) = (n-1)\Sigma$. Sendo assim:

$$\alpha = md^{*\top}\mathbf{M}^{*-1}d^*$$
$$= (n-1)\left(\sqrt{\frac{n}{n+1}}\Sigma^{-1/2}(x - \overline{\mathbf{x}})\right)^\top \left(\Sigma^{-1/2}M\Sigma^{-1/2}\right)^{-1}\left(\sqrt{\frac{n}{n+1}}\Sigma^{-1/2}(x - \overline{\mathbf{x}})\right)$$
$$= (n-1)\left(\sqrt{\frac{n}{n+1}}\Sigma^{-1/2}(x - \overline{\mathbf{x}})\right)^\top \left(\Sigma^{-1/2}nS\Sigma^{-1/2}\right)^{-1}\left(\sqrt{\frac{n}{n+1}}\Sigma^{-1/2}(x - \overline{\mathbf{x}})\right)$$
$$= \left(\frac{n-1}{n+1}\right)(x - \overline{\mathbf{x}})^\top S^{-1}(x - \overline{\mathbf{x}}) \sim t^2(p, n-1)$$

Usando o teorema da relação entre as distribuições de T^2 e F, podemos encontrar um ponto crítico para a utilização de α.

Teorema : Se α tem distribuição $T^2(p, m)$ então

$$\alpha^* = \frac{\alpha(m-p+1)}{mp} \sim F_{p,m-p+1}$$

Prova : A ideia da prova deste Teorema é usar a identidade

$$\alpha = md^\top M^{-1}d = (md^\top M^{-1}d)\frac{d^\top d}{d^\top d}$$

$$= m\left(\frac{d^\top M^{-1}d}{d^\top d}\right)\left(\frac{d^\top d}{1}\right) = m\frac{d^\top d/1}{d^\top d/d^\top M^{-1}d}$$

$$= \frac{md^\top d\left(\frac{p}{p}\right)}{\frac{d^\top d}{d^\top M^{-1}d}\left(\frac{m-p+1}{m-p+1}\right)} = \frac{mp\left(d^\top d/p\right)}{m-p+1\left(\frac{d^\top d}{d^\top M^{-1}d}\right)}$$

$$= \frac{mp}{m-p+1}\beta,$$

tendo $\beta \sim F_{p,m-p+1}$.

Como $d \sim N(0, I)$. Cada elemento de d tem distribuição $N(0, 1)$.

Assim $d^\top d \sim X_p^2$

$$d^\top d = \sum_{i=1}^{p} d_i^2$$

Resta mostra que, dado d

$$\frac{d^\top d}{d^\top M^{-1}d} \sim X^2_{(m-p+1)}$$

Esta parte pode ser encontrado na literatura, como por exemplo: (MARDIA; KENT; BIBBY, 1979)

3.2.1 Modelo usado para distribuição T^2 de Hotelling

Utilizando o α com distribuição $T^2(p, n-1)$ e que m é substituído por $(n-1)$ então

$$\alpha^* \sim F_{(p,m-p+1)}$$

Note que, $\alpha\left(\frac{(n-1)-p+1}{(n-1)p}\right) = \alpha\left(\frac{n-p}{np-p}\right)$ tem distribuição $F(p, n-p)$.

Isto é $\left(\frac{n-p}{np-p}\right)\left(\frac{n-1}{n+1}\right)(\mathbf{x}-\overline{\mathbf{x}})^\top S^{-1}(\mathbf{x}-\overline{\mathbf{x}})$ tem distribuição $F(p, n-p)$, logo para testar se cada observação está próxima da média, podemos usar essa estatística e verificar se 95% desses pontos estão no limite para determinar a sua autenticidade.

Para a eficiência do teste, utilizamos o F calculado como critério de rejeição ou não rejeição comparado com F tabelado, como mostra o quadro abaixo.

Quadro - Critério do teste

se	logo	então
F calculado $\geq F$ tabelado	o teste é **significativo** ao nível de significância $(0,05)$ considerado.	**Rejeitamos a hipótese nula** (H_0).
F calculado $< F$ tabelado	o teste é **não significativo** ao nível de significância $(0,05)$ considerado.	**Não rejeitamos a hipótese nula** (H_0) .

Fonte: Elaborada pela autora (2024)

3.3 SPLINES

A curva de spline tem uma ideia inicial bem simples nas leituras das assinaturas pelo qual se utiliza o $soft\ R$, em que se manipula uma aproximação linear para que ao invés de modelar um conjunto de observações ele determine os pontos de dados, escolhendo-os pontos distintos no intervalo de observações os "nós"e assim retornando uma lista de pontos onde possa modelar as curvas complexas por polinômios simples que se unem automaticamente de forma suave.

Existem três tipo de modelos:

- Splines de primeiro ordem (ou linear) é definida por uma base de pontos dados, através dos nós, que fazem um ligação dos conjunto de polinômios de grau um (para um conjunto de pontos ordenados). Podendo ser generalizada de um forma em que irar produzir curvas polinomiais suaves e de graus maiores (LYCHE; MORKEN, 2008);

- Splines cúbicas (ou Hermite) são denotadas através das de menor ordem nas quais a sua descontinuidade nos nós são suaves o suficiente que não são capazes de serem vistas a olho nu. A flexibilidade dada pelos nós fornecidos, produz curvas suaves de forma que podemos construir curvas de vários formatos. Dividida em duas categorias: Restrita ou também conhecido como spline natural (as caudas são modeladas por funções lineares) e a Irrestrita (as caudas não são modeladas por funções lineares)(LYCHE; MORKEN, 2008);

- Splines natural (ou spline cúbica restrita) como foi visto na spline cúbicas ela utiliza a suposição de que a funções são lineares além das fronteiras(LYCHE; MORKEN, 2008).

Nesse trabalho utilizaremos os splines cúbicos.

3.4 DYNAMIC TIME WARPING (DTW)

O Dynamic Time Warping (DTW) é uma técnica que compara e alinha duas sequências (dependentes do tempo), que podem ser sinais discretos (séries temporais), geralmente são sequências de características amostradas em pontos equidistantes no tempo.

DTW é baseado em programação dinâmica, onde encontrar padrões medições entre eventos com diferentes ritmos, que podem variar com o tempo. A idéia básica é bem simples pois não requer modelos complexos para a sua utilização, precisando apenas de amostras para serem comparadas. (SOUZA; PANTOJA; SOUZA, 2015)

Dadas duas séries temporais que são representadas por sequências. Se as sequências estão tomando valores em um determinado espaço (tempo), então para comparar duas sequências diferentes X e Y é preciso usar a medida de distância local e organizando todos os pontos da sequência (SENIN, 2008).

O algoritmo tem como principal vantagem a sua simplicidade, tendo como único requisito amostras das classes a serem comparadas (SOUZA; PANTOJA; SOUZA, 2015).

No trabalho a seguir, será utilizada as assinaturas do mesmo autor, para que seja montada e calculada o DTW.

3.5 TAXA DO ERRO

Existem dois tipos de erro, chamados de erro do tipo I (ETI) e do tipo II (ETII). O ETI é aquele que consiste em rejeitar a hipótese nula (H_0) quando é verdadeira e o ETII não rejeita H_0 sendo ela falsa (falso positivo), e com isso criando um intervalo de confiança usando a distribuição T^2 de Hotelling, para avaliar se devemos rejeitar ou não rejeitar uma assinatura, com 95% de confiança. Note que faremos os testes em cada ponto da assinatura, assim não rejeitamos aquelas que tenham pelo menos 95% dos pontos dentro dos limites.

Quadro - Resumo dos tipos de erros

	H_0 é verdadeira	H_0 é falsa
Rejeitar H_0	Erro tipo I (rejeitar H_0 verdadeiro)	Decisão Correta
Não rejeitar H_0	Decisão Correta	Erro tipo II (não rejeitar H_0 falsa)

Fonte: Elaborada pela autora (2024)

A taxa de erro será dada por

$$TE_I = \frac{\text{número de rejeição}}{\text{Total de comparações}} \ , \ TE_{II} = \frac{\text{número de não rejeição}}{\text{Total de comparações}}$$

em que TE_I é a taxa de erro de rejeição entre o total comparado e TE_{II} é a taxa de erro de não há rejeição entre o total comparado.

São portanto:

$$P(\text{ETI}) = P(\text{rejeitar } H_0 \mid H_0 \text{ é verdadeiro})$$

$$P(\text{ETII}) = P(\text{não rejeitar } H_0 \mid H_0 \text{ é falsa})$$

3.6 CONSTRUÇÃO DO MODELO

No estudo deste trabalho foi criado um modelo para construção de uma série contendo um conjunto de trincas. Com o auxílio do $softR$ aplicou-se algumas funções para se estabelecer uma didática mais prática para essa construção.

Estabelecendo valores para cada uma das variáveis da trinca (Z, X, Y), com o uso da função $function()$, definido assim alguns critérios para criar essas curvas:

- t: será utilizado para gera uma amostra de elementos da normal padrão, com o uso da função $rnorm()$ $(0, 1)$ e também como entrada da função $function()$;
- Para X e Y fixou o primeiro elemento da amostra o valor de 0,9;

- Para Z fixou o primeiro elemento da amostra o valor de 0, e os outros valores não negativos, por isso , o uso $abs()$ e representando o tempo percorrido na construção da curva;

Com os conjuntos de dados criados, nomeamos cada um deles. Logo depois, foram convertendas em dataframe com o uso do pacote $data.frame()$, e salvas por meio do $write.table()$

Agora com os conjuntos de curvas criadas e salvas, partiremos para algumas análises, como: conferir o comprimento máximo, o tempo mínimo e tempo máximo das séries (curvas). Sendo assim, criando o número de pontos para a suavização das séries, onde será sempre o dobro de seu comprimento máximo. Com isso, construindo o modelo spline com suas curvas mais suaves.

Criando matrizes, onde serão armazenadas os valores das splines para cada uma das variáveis. Notado que elas não estavam partindo do ponto $(0, 0)$, então reordenamos elas para assim dar início a realização dos testes, com a comparação do α^* com o uso da distribuição T^2 de Hotelling.

Na elaboração do teste, iremos definir o tamanho da amostra 'n', com isso construir uma matriz de médias e covariâncias, que será usada para a modelagem do α^*. Lembro que α^* será da forma apresentada na seção 3.2.1

Com os valores dos α^* para cada um dos pontos das curvas já determinados, analisaremos quantos deles serão maiores que o ponto crítico. Sendo assim, para cada comparação do α^* com o ponto crítico, se o seu valor for maior guardamos na matriz até que todos sejam comparados, e só assim partindo para a próxima curva até análise de todas. Com isso teremos uma matriz que irá mostrar o total de pontos rejeitados em cada uma das curvas para aquele determinado ponto crítico. Com essa matriz, faremos a seguinte comparação: se a curva tiver mais que 5% dos pontos maiores que o ponto crítico, rejeitamos essa curva.

3.7 CONJUNTO DE DADOS

3.7.1 Construção das curvas simuladas

Utilizando o que foi descrito na seção 3.6, com o objetivo de criar 200 conjuntos de trincas simuladas de amostras aleatórias.

Sendo cada umas delas formadas por 3 colunas $(X1, X2, X3)$:

- A primeira foi criada como forma de demarcação do tempo que se leva para a sua construção;
- A segunda como a largura de marcação dos pontos simulados;
- A terceira como a altura de marcação dos pontos simulados.

E reajustando-as todas elas para a origem, ponto $(0, 0)$, depois que ambas as técnicas forem aplicadas as curvas simuladas.

3.7.2 Banco de dados

Utilização de 75 conjuntos de dados de assinaturas reais, obtidos pelo Departamento de Comunicações da Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas (Decom - FEEC UNICAMP, Brasil). Sendo 50 assinaturas verdadeiras e 25 assinaturas falsas, também cada uma formada por 3 colunas $(V1, V2, V3)$, sendo V1 o tempo em que a assinatura foi escrita, V2 a largura em que o ponto foi marcado e V3 a altura em que o ponto foi marcado, respectivamente.

4 RESULTADOS

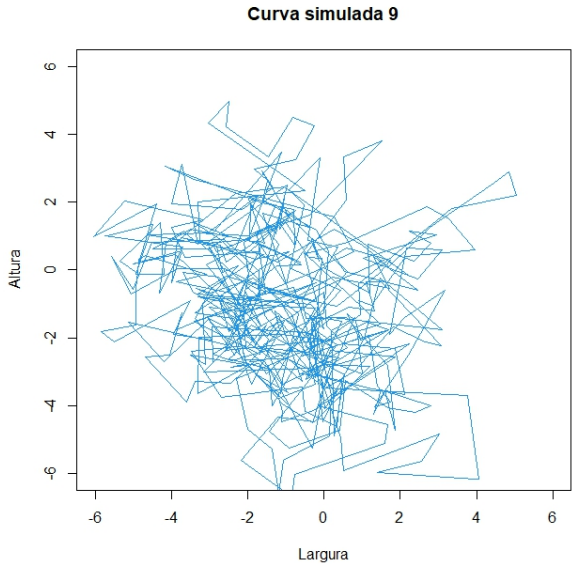
4.1 RESULTADOS DAS CURVAS SIMULADAS

4.1.1 Criar e suavizar as curvas

Dentre as 200 curvas simuladas, foram separados em dois grupos de 100. O primeiro grupo representam os de referência e o segundo grupo o de teste.

O grupo 1 representa as médias para construção dos ajustes, onde foram usados 6 tamanhos $(5, 6, 7, 8, 9, 10)$ para a base das distâncias médias e aí serem comparadas ponto a ponto com cada uma do grupo 2. Entre as 10 primeiras usadas para obtenção das médias a curva de maior tamanho é a curva simulada 9, decidido por conveniência como modo de representação.

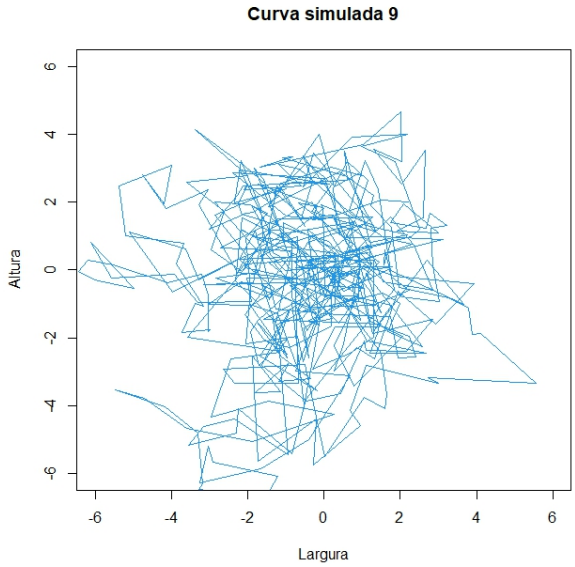
Figura 1 – 9 Curvas Simuladas do grupo 1



Fonte: Elaborada pela autora (2024)

Comparando os grupos, podemos ver que as curvas geradas são distintas e com tamanhos de pontos diferentes, ver Figura 1 e a Figura 2.

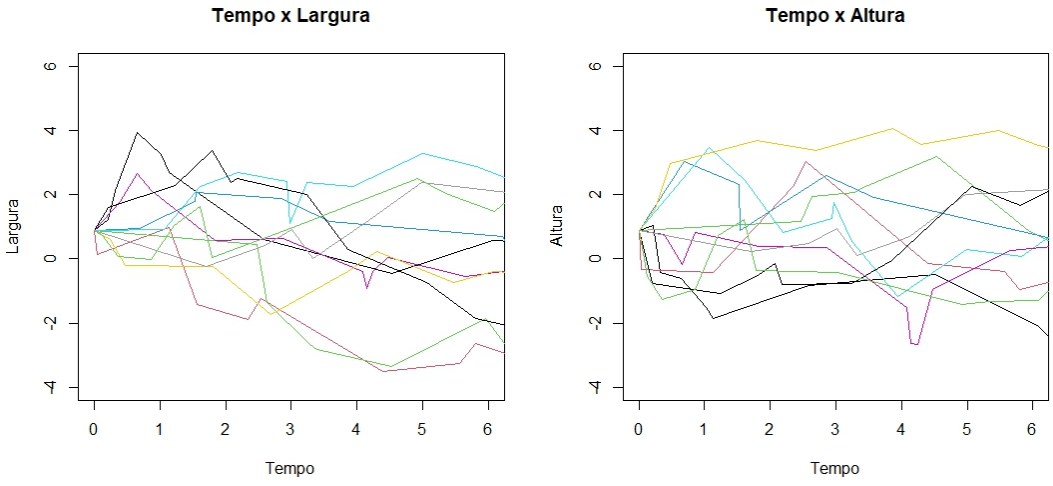
Figura 2 – 9 Curvas Simuladas do grupo 2



Fonte: Elaborada pela autora (2024)

Levando em conta as variáveis largura e altura podemos ver o comportamento de ambas com o decorrer do tempo. Conforme a Figura 3 foram usadas 10 curvas daquelas simuladas para se ter a representação do comportamento delas com tempo.

Figura 3 – Comparação entre as variáveis



Fonte: Elaborada pela autora (2024)

As técnicas de DTW e Spline foram aplicadas para as curvas simuladas modelando e suavizando cada grupo de conjunto e deixando todas com o tamanho de pontos da maior simulada. No estudo escolheu-se 992 pontos. Reajustando as matrizes splines para origem agora utilizaremos a distribuição T^2 de Hotelling, para fazer as 992 comparações.

4.1.2 Não rejeição ou rejeição

Com as técnicas já aplicadas aos modelos simulados, podemos supor que $\alpha^* \sim F_(p, n-p)$ e para justificar a não rejeição, temos que aceitar pelo menos 95% dos pontos de cada uma simuladas, ou seja, 942 dos pontos devem estar abaixo do ponto crítico do teste.

Resumindo o critério do teste:

$$F_{calc} < F_{tab}(5\%) - \text{não rejeitamos} H_0$$

$$F_{tab} \geq F_{calc}(5\%) - \text{rejeitamos} H_0$$

4.1.3 Construção do ponto crítico

Devido a suposição de normalidade ter sido violada, surge a necessidade de se criar os pontos críticos, visto que na prática os pontos de sua assinatura não tem distribuição normal bivariada.

Assim buscamos os pontos críticos entre 1 e 25 e aplicando para cada tamanho dentre as 100 do grupo 2 para termos uma rejeição de 5%.

Tabela 1 – Ponto crítico para 5%

n	Ponto crítico
5	19,5
6	14,5
7	14
8	12,5
9	12,5
10	10,5

Fonte: Elaborada pela autora (2024)

Ao calcular o número de rejeição para 6%, 7%, 8%, 9% e 10% teremos novos pontos críticos para o mesmo tamanho 'n'.

Tabela 2 – Pontos de rejeições

n	6%	7%	8%	9%	10%
5	19,5	19	18,5	18,5	18,5
6	14,5	14	13,5	13,5	13
7	13	12	11,5	11,5	11,5
8	12,5	12	11,5	11	10,5
9	12	11,5	11,5	11	10,5
10	10,5	10,5	10,5	10,5	10

Fonte: Elaborada pela autora (2024)

Em resumo, dependendo do tamanho da amostra, podemos obter pontos críticos empíricos, de tal forma a se ter um determinado número de pontos que ultrapassam o limite (ultrapassa o ponto empírico). Por exemplo, se tomarmos uma amostra de 6 assinaturas do grupo de referência, para construir os parâmetros $\overline{x}$ e S, de cada ponto, devemos estabelecer 14 como ponto crítico e assim 7% das assinaturas são rejeitadas, isto é:

$$TE_I = \frac{\text{número de rejeição}}{\text{Total de comparações}} = \frac{7}{100} = 0,07 = 7\%$$

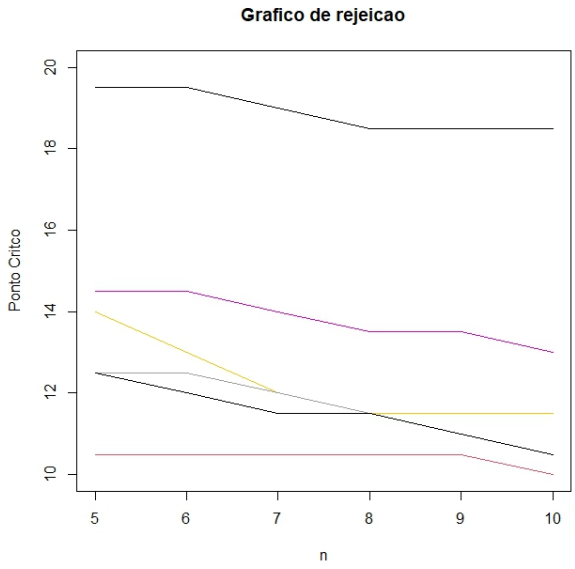
Pois 95% dos pontos teste (α^*) são menores que o ponto crítico 14, ou seja, existe menos que 942 pontos menores que 14.

Quadro 1 – 5% dos pontos de cada assinaturas são maiores que o ponto crítico 14

0	40	18	36	19	26	34	41	21	36	39	25	22	28	15	**52**	22	36	24	27
15	28	30	29	22	19	**67**	27	28	47	30	43	30	48	44	22	41	43	30	24
27	30	38	44	32	23	28	25	31	14	30	14	38	35	27	36	38	19	30	5
47	21	39	26	15	31	27	24	13	**80**	22	23	25	43	41	43	38	**54**	**50**	**80**
8	17	22	15	27	47	47	**50**	15	18	26	23	28	47	37	26	41	31	40	31

Fonte: Elaborada pela autora (2024)

Sendo assim comparando o α^*, notamos no Quadro 1 que 7 delas 100 serão rejeitadas, pois tem que 50 pontos são maiores que a ponto critico 14, ou seja, mais de 5% dos pontos são rejeitados.



Fonte: Elaborada pela autora (2024)

O gráfico acima representa os pontos críticos com o aumento do tamanho amostra, usado como referência a distribuição de Wishart e assim podendo estabelecer uma medida na qual é usada para construção da distribuição T^2 de Hotelling.

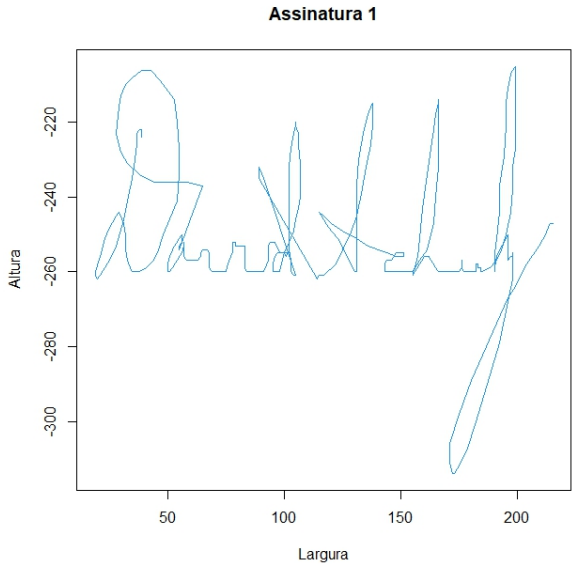
Mostrando como as curvas se comportam com o aumento do ponto crítico para uma determinado tamanho de amostras. Onde quanto maior o tamanho da amostra menor será o número de rejeição das curvas.

4.2 RESULTADO DO CASO REAL

4.2.1 Suavização das curvas Reais

Nos casos real no conjunto de 75 assinaturas entre verdadeiras e falsas, foram utilizadas 5 das 50 verdadeiras. A assinatura 1 foi escolhida como referência sem pretensões como mostrado logo abaixo.

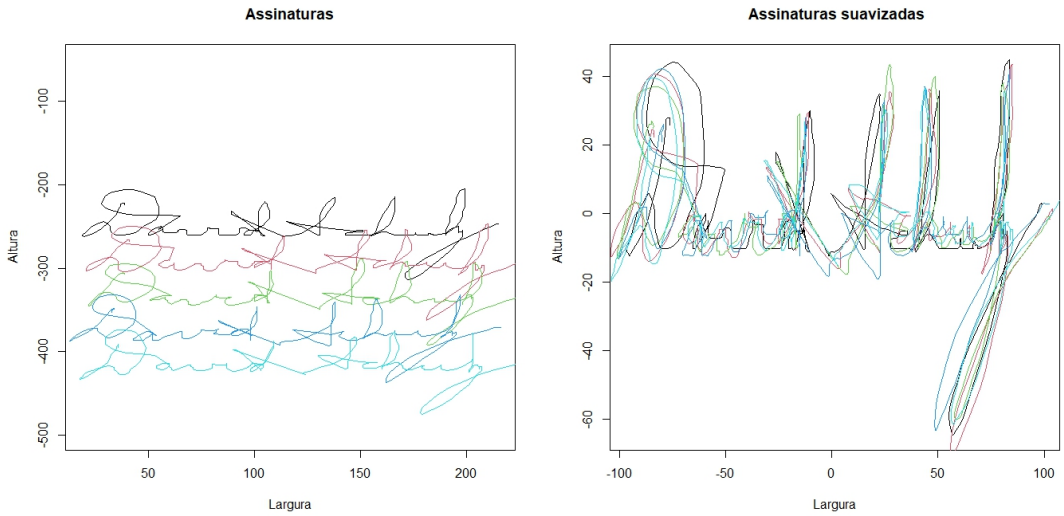
Figura 4 – Assinatura 1 das verdadeiras do caso real



Fonte: Elaborada pela autora (2024)

Aplicando spline as assinaturas podem apresentar curvas mais suaves. Porém como temos alguns fatores que podem alterar a escrita, nota-se que as assinaturas estão em origem distintas e apresentando uma diferença na sua localização, como mostrado na Figura 5.

Figura 5 – Assinatura caso real

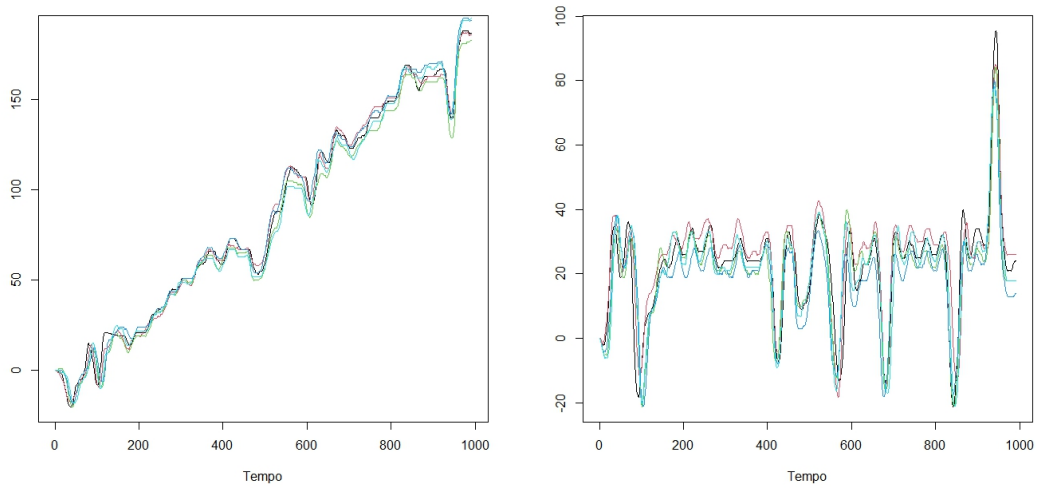


Fonte: Elaborada pela autora (2024)

Visualmente podemos notar que as assinaturas não estão no mesmo ponto de partida com isso ajustaremos elas da mesma forma que foi realizado nas simuladas.

Com a suavização e ambas partindo agora da origem zero .

Figura 6 – Assinaturas partindo do ponto $(0, 0)$



Fonte: Elaborada pela autora (2024)

4.2.2 Rejeição ou não rejeição das assinaturas reais

Com os ajustes e alinhamentos das assinaturas com as distâncias médias, podemos partir para o usando pontos críticos, em que também serão de 1 até 25 com o mesmo tamanho de 'n'. Quando se estabelece o tamanho de 'n' para a construção das médias, o teste de comparação utilizaremos o total menos o 'n' para a distribuição T^2 de Hotelling.

Desta forma temos um número de pontos que serão rejeitados para cada uma das assinaturas. Assim podendo construir a Tabela 3 que representará os pontos críticos, em que terão 5%, 6%, 7%, 8%, 9% e 10% de rejeição.

Tabela 3 – Pontos de rejeições

n	5%	6%	7%	8%	9%	10%
5	18,5	17,5	17,5	16,5	16,5	16,5
6	19,5	18	18	15	15	15
7	19,5	17	16,5	16,5	14,5	14,5
8	17	16	16	16	15,5	15,5
9	16,5	16,5	16	16	15,5	15,5
10	13,5	13,5	12,5	12,5	11,5	11,5

Fonte: Elaborada pela autora (2024)

Assim para cada uma das suposições, o α^* será rejeitado ou não para um determinado ponto crítico. Se o $(\alpha^* <$ Ponto crítico$)$, ou seja, $F_{calc} < F_{tab}(5\%)$. Não rejeitamos H_0 se o total de pontos for superior há 95%, a assinatura não será rejeitada podendo afirma que a assinatura é verdadeira.

Agora por exemplo, levando em conta um tamanho amostral 8, teremos 42 assinaturas verdadeiras para testar se rejeitamos ou não o α^* para o ponto crítico 16. Note que para uma assinatura ser considerada verdadeira não pode ter mais que 50 pontos maiores que 16, ou seja, não pode haver mais que 5% dos pontos maiores que 16.

No quadro abaixo podemos ver quantos pontos foram rejeitados em cada uma das assinaturas reais para o ponto crítico 16.

Quadro 2 – Número de pontos que ultrapassam o ponto crítico 16 dentro das 42 assinaturas

1	8	**52**	0	4	11	25
41	5	33	3	0	3	2
4	33	**51**	8	8	0	35
60	3	13	8	13	6	43
23	31	29	4	44	13	22
4	5	0	15	46	21	6

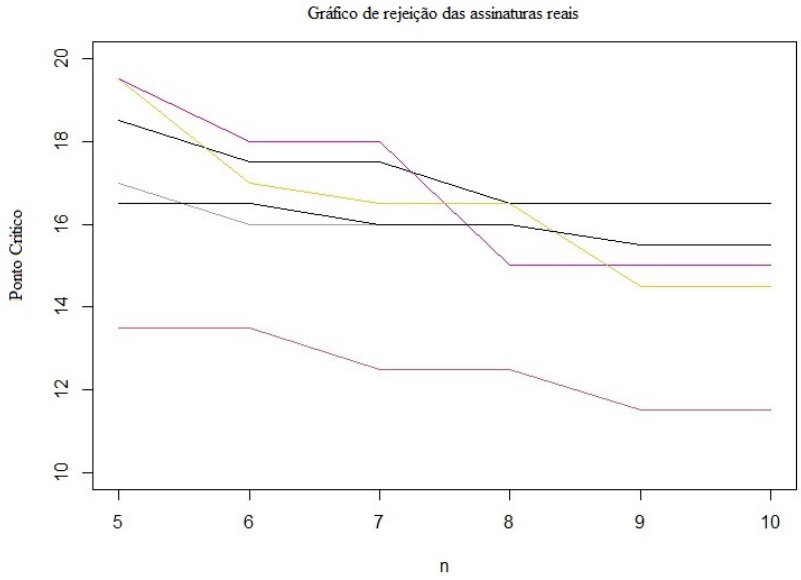
Fonte: Elaborada pela autora (2024)

Observe que foram encontradas 3 das 42 assinaturas, apresentaram mais que 50 pontos acima do ponto crítico 16, ou aproximadamente 7% das assinaturas tem mais que 5% dos pontos maiores, como mostrado no Quadro 2.

$$TE_I = \frac{\text{número de rejeição}}{\text{Total de comparações}} = \frac{3}{42} \approx 0.07 \approx 7\%$$

Constatou também que para as 42 assinaturas que foram testadas para o ponto crítico 16 e tamanho amostral (n) 8, temos uma rejeição de 6 a 8 por cento. Sendo assim, aproximadamente 3 das assinaturas das 42 serão rejeitadas, pois o $softR$ faz uma aproximação.

O gráfico podemos ver o comportamento das assinaturas de acordo com o tamanho da amostra e o ponto crítico. Onde quanto maior o tamanho da amostra menor será a número de rejeição das assinaturas.



Fonte: Elaborada pela autora (2024)

Com a análise dessas assinaturas, foi possível notar o controle de falsificação de assinaturas, mesmo com a necessidade de criar os pontos críticos. Mostrado que o processo se encontra sob controle estatístico, para abordagem que foi utilizada neste trabalho.

5 CONSIDERAÇÕES FINAIS

Conforme resultados obtido, o trabalho mostrou como obter pontos críticos empíricos para adequação do sistema de reconhecimento. Dependendo do grau de segurança do sistema, podemos estabelecer níveis de confiança adequados para cada situação. Assim, se o sistema tiver a preocupação com uma segurança mais rígida podemos adotar pontos críticos mais altos.

Para o erro do tipo II, a taxa foi nula para todos os pontos críticos estudados, pois não existiu a possibilidade de não rejeitar uma assinatura falsa para o α^*, mesmo com o aumento do ponto crítico não há indicativos de não rejeição de uma assinatura falsa. Indicando que as assinaturas não são realmente autênticas e provando que existe um padrão entre elas.

Ainda assim, temos que necessitamos destacar que não é inevitável que tenhamos assinaturas falsas dentre as que não foram rejeitadas, lembrando então que mesmo com a eficiência do uso da distribuição T^2 de Hotelling é importante destacar o uso de outras técnicas para a análise das assinaturas.

Garantindo que a utilização da distribuição T^2 de Hotelling para análise estatística é uma importante técnica para garantir a veracidade das curvas ou assinaturas. Concluindo assim que os resultados obtidos nesse trabalho, podem assegurar a validade de assinaturas.

REFERÊNCIAS

HEINEN, M. R.; OSÓRIO, F. S. Autenticação de assinaturas usando algoritmos de aprendizado de máquina",. *Anais da V ENIA*, 2005.

JAIN, A. K.; GRIESS, F. D.; CONNELL, S. D. On-line signature verification. *Pattern recognition*, Elsevier, v. 35, n. 12, p. 2963–2972, 2002.

KIELING, A. C.; SANTOS, V. S. dos; PINO, G. G. del; SANSONE, J. L. Estudo do gráfico t2 de hotelling no controle estatístico do processo multivariado. In: *VII Congresso Nacional de Engenharia Mecânica*. [S.l.: s.n.], 2012.

LYCHE, T.; MORKEN, K. Spline methods draft. *Department of Informatics, Center of Mathematics for Applications, University of Oslo, Oslo*, 2008.

MARDIA, K.; KENT, J.; BIBBY, J. *Multivariate Analysis*. Academic Press, 1979. (Probability and Mathematical Statistics : a series of monographs and textbooks). ISBN 9780124712508. Disponível em: <https://books.google.com.br/books?id=bxjvAAAAMAAJ>.

SENIN, P. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, v. 855, n. 1-23, p. 40, 2008.

SOUZA, C. F.; PANTOJA, C. E.; SOUZA, F. C. Verificação de assinaturas offline utilizando dynamic time warping. *Universidade de Brasília*, 2015.