UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA

Adriana Laurindo Monteiro

Estimadores robustos e concentração de medida: uma introdução

VITÓRIA
2021

Adriana Laurindo Monteiro

Estimadores robustos e concentração de medida: uma introdução

Dissertação de mestrado apresentada ao PPGMAT como parte dos requisitos exigidos para a obtenção do título de Mestre em Matemática

Orientador: Profº Drº Fábio Júlio Valentim

VITÓRIA
2021

Estimadores Robustos e Concentração de Medida: uma Introdução

Adriana Laurindo Monteiro

Dissertação submetida ao Programa de Pós-Graduação em Matemática da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do grau de Mestre em Matemática.

Aprovada em 30 de março de 2021 por:

_____

Prof.ª Drª Fábio Júlio da Silva Valentim
Universidade Federal do Espírito Santo
Orientador

_____

Prof. Dr. Valdério Reisen
Universidade Federal do Espírito Santo
Membro Interno

_____

Profª Drª Florência Graciela Leonardi
Instituto de Matemática e Estatística da Universidade de São Paulo
Membro Externo

Universidade Federal do Espírito Santo
Vitória, março de 2021

Este documento foi assinado digitalmente por FABIO JULIO DA SILVA VALENTIM
Para verificar o original visite: https://api.lepisma.ufes.br/arquivos-assinados/169616?tipoArquivo=O

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

PROTOCOLO DE ASSINATURA

O documento acima foi assinado digitalmente com senha eletrônica através do Protocolo Web, conforme Portaria UFES nº 1.269 de 30/08/2018, por
VALDERIO ANSELMO REISEN - SIAPE 297617
Departamento de Estatística - DE/CCE
Em 31/03/2021 às 15:41

Para verificar as assinaturas e visualizar o documento original acesse o link:
https://api.lepisma.ufes.br/arquivos-assinados/167116?tipoArquivo=O

Este documento foi assinado digitalmente por VALDERIO ANSELMO REISEN
Para verificar o original visite: https://api.lepisma.ufes.br/arquivos-assinados/167116?tipoArquivo=O

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

PROTOCOLO DE ASSINATURA

O documento acima foi assinado digitalmente com senha eletrônica através do Protocolo Web, conforme Portaria UFES nº 1.269 de 30/08/2018, por
FABIO JULIO DA SILVA VALENTIM - SIAPE 2545870
Departamento de Matemática - DM/CCE
Em 07/04/2021 às 14:45

Para verificar as assinaturas e visualizar o documento original acesse o link:
https://api.lepisma.ufes.br/arquivos-assinados/169636?tipoArquivo=O

Este documento foi assinado digitalmente por FABIO JULIO DA SILVA VALENTIM
Para verificar o original visite: https://api.lepisma.ufes.br/arquivos-assinados/169636?tipoArquivo=O

Este trabalho é dedicado aos meus pais: ele que sempre me encheu com os números e ela que sempre me fez ler os bons livros!

Agradecimentos

Agradeço primeiramente ao Divino pela maravilhosa experiência de viver. Agradeço também a Ele, a minha saúde, tão preciosa em tempos pandêmicos.

Agradeço à minha mãe, a primeira a me mostrar o poder do conhecimento e da sabedoria. Aquela que me ensinou a agradecer todos os dias cada passo de uma jornada. Agradeço ao meu pai, de quem herdei grande apreço aos números e ao trabalho. Agradeço às minhas irmãs, as importantes inspirações da minha vida. Agradeço a minha companheira Débora, a pessoa que acredita em mim quando eu mesma não acredito e que me mostra todos dias a possibilidade de ser diferente.

Agradeço aos amigos. Amigos de infância, amigas do Ifes, amigos da matemática, de graduação e mestrado. Amigos de caminhada, de boteco e de festas. Agradeço a todos aqueles com quem troquei experiências, risadas e lágrimas. Muito obrigada por terem tornado os momentos difíceis um pouco mais fáceis e os momentos alegres, muito mais alegres ainda.

Agradeço a todos os professores que contribuíram com a minha formação. Agradeço em especial, àqueles que me fizeram acreditar na docência e me mostraram que a educação pode sim ser transformadora, e me inspiraram a trilhar esse caminho até aqui.

Agradeço ao Departamento de Matemática e ao Programa de Pós-Graduação em Matemática da Ufes pela excelente formação que recebi e aos seus professores que acreditaram em mim e confiaram em meu trabalho. Agradeço, em específico, ao meu orientador Fábio Júlio Valentim, com quem tive conversas fundamentais para meu desenvolvimento enquanto matemática e enquanto pessoa. Agradeço por ter me apresentado a beleza da Probabilidade.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão de bolsa de estudos durante o curso de Mestrado.

Resumo

Neste trabalho são apresentados alguns resultados importantes nas áreas de probabilidade e estatística robusta. Estudaremos funcionais estatísticos (derivadas de funcionais e sua performance) e processos estocásticos subgaussianos, além de uma aplicação em problemas de classificação, objeto da teoria de aprendizagem estatística. Os teoremas principais são um teorema central do limite para um estimador robusto da função de covariância e a Desigualdade de Dudley (Integral de entropia de Dudley).

Palavras-chave: função de covariância; estimador robusto; função de influência; Delta método funcional; probabilidade em alta dimensão; variáveis aleatórias sub-gaussianas; encadeamento; dimensão VC; teoria de aprendizagem estatística; minimização do risco empírico.

Abstract

This dissertation presents important results on probability and on robust statistics. Statistics functionals and its perfomances and derivatives, as well as subgaussian stochastics process are some of the topics. Besides that, we present an application in classification problems, a subject in statistical learning theory. The main results are a central limit theorem of a robust estimator of the covariance function and Dudley's Inequality (Dudley's entropy integral).

Key-Words: covariance function; robust estimator; influence function; functional Delta method; high dimensional probability; sub-gaussian random variables, chaining, VC dimension; statistical learning theory empirical risk minimization.

Sumário

Introdução

Este trabalho é fruto de um estudo introdutório de dois tópicos em que a estatística e matemática se encontram: performance de estimadores e probabilidade (concentração de medida). Forneceremos as principais ideias que perpassam esses dois pontos de vista a fim de constituir uma caixa de ferramentas para a investigação de problemas em estatística matemática.

O maquinário fornecido pela teoria de probabilidade estabelece duas diferentes abordagens para os problemas da área: uma abordagem assintótica, em que o objetivo é basicamente provar teoremas centrais do limite para alguns estimadores, e uma abordagem não assintótica, na qual se busca estimativas explícitas para uma quantidade fixa de variáveis.

Com a finalidade de dar uma visão geral das ideias a serem desenvolvidas, vamos descrever informalmente os princípios básicos do texto, dividido em duas grandes partes.

A parte I tem dois eixos principais: uma discussão preliminar sobre funcionais estatísticos e a análise de um estimador robusto para a função de covariância[1].

Um modo classicamente estatístico de interpretar um fenômeno é desenvolver um modelo paramétrico e a partir dele extrair informações usando *estimadores*. Em muitos casos, os modelos propostos assumem condições que nem sempre correspondem com a realidade ou os estimadores não apresentam boas performances. Para driblar esses contratempos, surgiu a Estatística Robusta. As pesquisas nesta área desenvolveram métodos que estudam, por exemplo, o comportamento assintótico dos estimadores e sua sensibilidade a certas contaminações da amostra. As Seções 1 e 2 do Capítulo 1 tratarão de conceitos que abrangem estas propriedades. Podemos citar a *função de influência* de um funcional T numa distribuição acumulada F no ponto x

$$\mathrm{IF}(x, T, F) = \lim_{t \to 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t},$$

que estuda o comportamento local dos estimadores perante a pertubações infinitesimais. Além disso, a ideia de *consistência* abordará a convergência dos estimadores.

A Estatística Robusta, cuja primeira abordagem teórica foi lançada por Huber [8] em 1964, tem fundamental importância na resolução de problemas atuais. Dado que estimadores tão simplórios como a média são suscetíveis a valores extremos, é fácil se convencer da necessidade de métodos robustos. Pesquisas recentes se concentram, por exemplo, na busca de estimadores de coeficientes de regressão, estimadores de parâmetros de escala e de localização com performances cada vez mais robustas.

A Seção 1.2, baseada em Lévy-Leduc et al. [11], se propõe a estudar um estimador robusto $\hat{\gamma}_Q$ para a função de covariância obtido em função de outro estimador, também robusto, Q_n. Boa resistência à presença de *outliers* é uma das propriedades de Q_n, como foi analisado por meio de sua função de influência e de seu *breakdown point* em Rousseeuw [13].

Ambos os estimadores quando aplicados em processos gaussianos estacionários são assintoticamente normais, e portanto consistentes, com uma taxa de convergência $\sqrt{n}$. Por esses motivos,

[1] Ao longo do texto, usaremos as expressões *função de covariância* e *de autocovariância* como sinônimos.

$\hat{\gamma}_Q$ e Q_n são, respectivamente, ótimas alternativas aos estimadores clássicos para função de covariância e aos de escala.

A parte II está baseada em duas ideias principais: concentração e supremo. O fenômeno de concentração está bem ilustrado pela Lei dos Grandes Números: dadas $X_1, \ldots, X_n, \ldots$ variáveis aleatórias i.i.d.[2] com média μ, sabemos que

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \xrightarrow{n\to\infty} 0.$$

Esse comportamento não está restrito a funções lineares de variáveis aleatórias (podemos considerar uma função Lipschitz qualquer, por exemplo). Além disso, pode-se questionar o quão próximo da média (ou de outros valores), está a soma, considerando uma quantidade fixa n. Qual o decaimento da probabilidade da cauda de uma variável X? Será que o comportamento destas variáveis pode ser descrito em comparação com outra variável, por exemplo a gaussiana? Estas e outras perguntas serão respondidas, e as definições formalizadas, ao estudarmos as *variáveis aleatórias subgaussianas* na Seção 2.1.

Perceba que o princípio da concentração está preocupado com desvios de uma função aleatória $f(X_1, \ldots, X_n)$ de sua média $\mathbb{E}\, f(X_1, \ldots, X_n)$, mas não diz nada a respeito da média em si. Alguns problemas em aplicações estão relacionados com a magnitude de $f(X_1, \ldots, X_n)$, como por exemplo, quando a função f é um supremo, isto é, $f(X_1, \ldots, X_n) = \sup_{1\leq i\leq n} X_i$. Mais geralmente, queremos tratar de um processo estocástico $(X_t)_{t\in T}$ e buscar entender

$$\sup_{t\in T} X_t.$$

A norma de matrizes ou de vetores aleatórios pode ser expressa como o supremo de um processo estocástico. O problema de minimização do risco empírico, típico em aprendizagem estatística, também pode ser analisado a partir do supremo de um processo.

Veremos que estimar a magnitude de $\sup_{t\in T} X_t$ está relacionado tanto a propriedades do processo em si quanto a propriedades do conjunto de índices T. Grosso modo, quando $(X_t)_{t\in T}$ é "minimamente bem comportado", é possível obter uma cota superior para o $\sup_{t\in T} X_t$ usando a "complexidade"do conjunto T. Daremos significados precisos às expressões em aspas quando abordarmos os *processos subgaussianos* e *número de cobertura*.

O método conhecido como *encadeamento* fará a conexão entre conceitos probabilísticos e outras noções determinísticas, como a entropia. Por meio do encadeamento obteremos a *desigualdade de Dudley* e conseguiremos uma cota superior para o valor esperado do supremo de um processo estocástico. Esses tópicos serão cobertos pela Seção 2.2.

Para as aplicações estatísticas, em especial o problema de minimizar o risco empírico, precisaremos falar do processo empírico $(X_f)_{f\in\mathscr{F}}$, definido como

$$X_f := \frac{1}{n}\sum_{i=1}^{n} \left(f(X_i) - \mathbb{E}\, f(X)\right)$$

e introduzir outras definições mais específicas acerca do conjunto T, como a dimensão de Vapnik–Chervonenkis (VC). O método de *simetrização* permite conectar a cota fornecida pela Integral de Dudley com o supremo de um processo empírico. Esses tópicos serão encontrados na Seção 2.3. Por fim, a Seção 2.4. dá uma aplicação dos resultados até então obtidos para estudar em detalhes o problema de minimizar o risco empírico no contexto de *problemas de classificação*.

[2] Variáveis aleatórias i.i.d. são variáveis independentes e identicamente distribuídas.

O apêndice contém algumas definições e resultados preliminares de variáveis aleatórias e alguns fatos sobre a distribuição normal multivariada que são usados ao longo do texto, além de uma seção sobre derivadas de funcionais. Um curso básico de graduação em teoria de probabilidade e o Capítulo 3 em van der Vaart [14] cobre esses pré-requisitos.

A autora ressalta que a maior parte do conteúdo deste trabalho foi baseada em Vershyin [16] e em Lévy-Leduc et al. [11].

Capítulo 1

Estimador Robusto

O estudo de diversas áreas do conhecimento está baseado na coleta e tratamento de dados, o que usualmente gera grande volume de informação. A pesquisa estatística almeja extrair conclusões relevantes destas pesquisas empíricas, desenvolvendo métodos que sejam computacionalmente eficientes e matematicamente consistentes. Os modelos estocásticos paramétricos, conhecidos como abordagem clássica da estatística, cumprem, em muitos casos, com esse objetivo.

Ao longo dos anos, os pesquisadores perceberam que tais modelos tinham como pressuposto algumas condições que a realidade nem sempre satisfazia por completo, e nesse ponto nasce a estatística não paramétrica. A Estatística Robusta, no entanto, se coloca no ponto de encontro das duas abordagens: usa modelos estocásticos para recolher informações e implementa procedimentos que não dependem inerentemente do modelo.

Desta seção em diante, as letras gregas Φ e φ denotarão, respectivamente, a distribuição e a densidade da gaussiana padrão $\mathscr{N}(0,1)$ e $\Phi_{\mu,\sigma}(\cdot) = \Phi(\frac{\cdot-\mu}{\sigma})$ denota a distribuição da gaussiana $\mathscr{N}(\mu,\sigma^2)$. A notação para a distribuição empírica $r \mapsto F_n(r) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq r\}}$ será F_n.

1.1 Funcionais estatísticos

Considere uma amostra aleatória, isto é, uma coleção de n observações $X_{1:n} = (X_1, \ldots, X_n)$ em que as coordenadas de $X_{1:n}$ são variáveis aleatórias independentes com distribuição F_θ, sendo $\{F_\theta \mid \theta \in \Theta\}$ um modelo paramétrico. O estudo de inferência estatística se concentra em encontrar ou estimar algumas características da população modelada por F_θ a partir da amostra aleatória. Por exemplo, pode ser de interesse o maior ou menor valor daquela amostra, sua média aritmética, a variância e etc. Estas funções são conhecidas como *estimadores*. Nesta seção discutiremos algumas propriedades dos estimadores.

Definição 1.1.1. *Identificamos a amostra $X_{1:n}$ com a distribuição empírica F_n. Chamamos de estimador*[1] *T_n qualquer função real da amostra. A notação será $T_n = T_n(X_1, \ldots, X_n) = T_n(F_n)$.*

Consideraremos estimadores que são funcionais, isto é, que assintoticamente podem ser substituídos por funcionais. Isto significa que existe um funcional T definido no conjunto de todas as distribuições para as quais T está definida tal que $T_n(X_1, \ldots, X_n)$ converge em probabilidade para $T(G)$, considerando $X_1, \ldots, X_n$ independentes e identicamente distribuídas com distribuição G.

[1] Note que um estimador é, na verdade, uma sequência de funções T_n da amostra $X_{1:n}$.

Denotaremos a convergência em probabilidade por $\to_P$:

$$T_n(X_1, \ldots, X_n) \to_P T(G).$$

1.1.1 Performance dos Estimadores

Algumas perguntas naturais logo surgem ao definir um estimador: como avaliar a performance de um estimador? Como definir precisamente qual estimador se aproxima mais da realidade que outro? Como os estimadores se comportam quando os dados possuem ruídos ou quando a amostra tem tamanho muito grande? Esta seção, que está baseada em Casella [3], traz definições que se propõem a responder estas perguntas.

Note que não basta calcular somente o seu erro direto, ou seja, não basta que $T(X) = \theta$ pois pode ser que isto nunca aconteça. De fato, considere $X \sim F$[2] e F é distribuição absolutamente contínua. Nesse caso, podemos obter[3] um funcional T tal que $\mathbb{P}(T(X) = \theta) = 0$. Podemos calcular o valor esperado do erro:

Definição 1.1.2 (Bias)**.** *O bias, ou viés, é definido por:*

$$b_T(F) = \mathbb{E}_F[T(X) - \theta],$$

sendo F a distribuição de X, o valor esperado calculado com respeito a F e o valor estimado por T é θ. Diz-se que T é não enviesado se $b_T(F) = 0$.

Podemos medir o quanto os valores de T estão concentrados em torno de θ:

Definição 1.1.3 (Mean Squared Error - MSE)**.** $mse_T(F) = \mathbb{E}_F[T(X)-\theta]^2 = b_T(F)^2 + \operatorname{Var} T(X)$

Outras maneiras de avaliar um estimador levam em conta seu comportamento assintótico, o que tem suas vantagens: quando $n \to \infty$, os cálculos se tornam mais fáceis inclusive computacionalmente. Entretanto, não é possível determinar um n exato que seja suficiente para aplicar os resultados assintóticos.

Os conceitos de convergência, como a convergência em quase todo ponto, convergência em probabilidade, convergência em L^p e convergência em distribuição, nos auxiliam a investigar as distribuições de $T_n(X)$ para n grande. Por isso, os trabalhos na área sempre são acompanhados de resultados empíricos e numéricos que confirmam as propriedades assintóticas. Uma maneira de estudar um estimador do ponto de vista assintótico é avaliar sua *consistência*.

Definição 1.1.4. *Um estimador T_n de um parâmetro θ de uma amostra $X = (X_1, \ldots, X_n)$ é consistente se $T_n(X) \to_P \theta$.*

Note que a definição se refere a uma sequência de estimadores, então o que se avalia é a consistência de uma sequência $(T_n)_n$. A consistência indica que com probabilidade alta, o estimador está arbitrariamente próximo de θ. Em outras palavras, quanto maior o tamanho de sua amostra, mais próximo está do parâmetro estimado. Esta propriedade pode ser obtida a partir de outra, a *normalidade assintótica*.

Definição 1.1.5. *Diz-se que o estimador T_n é assintoticamente normal quando $\sqrt{n}(T_n(X) - \theta)$ converge em distribuição para $\mathcal{N}(0, \sigma_\theta^2)$. A convergência em distribuição, também chamada de convergência fraca ou convergência em lei, será denotada por $\to_D$: $\sqrt{n}(T_n(X)-\theta) \to_D \mathcal{N}(0, \sigma_\theta^2)$.*

[2] Quando a variável X tem distribuição F, denotamos por $X \sim F$.
[3] Considere $T(x) = x$.

Se T_n é assintoticamente normal, então pelo Teorema A.8,

$$T_n(X) - \theta = \sqrt{n}\frac{T_n(X) - \theta}{\sqrt{n}} \to_D \lim_n \frac{1}{\sqrt{n}}\mathcal{N}(0, \sigma_\theta^2) = 0.$$

Portanto, $T_n(X) - \theta \to_P 0$, isto é, T_n é consistente[4].

1.1.2 Estimadores Robustos

A estatística robusta, como já mencionado na introdução, concentra-se no fato de que a maioria dos procedimentos estatísticos é construída a partir de meras aproximações da realidade. Um dos problemas clássicos é o problema dos *outliers*: como lidar com a presença de um dado muito distante do conjunto analisado? Queremos construir ferramentas que não ignorem o aspecto real do problema nem tampouco forneça informações contaminadas por erros.

Usando uma analogia de funções analíticas encontrada na introdução do livro Hampel et al. [7], podemos resumir a ideia básica da estatística robusta: enquanto o cálculo se utiliza das derivadas para estudar o comportamento local de uma função após uma pertubação infinitesimal num ponto, a estatística robusta usa a *função de influência* para estudar a influência sofrida por um estimador após uma pequena contaminação de uma observação. Quando há uma singularidade próxima a este ponto, a abordagem da derivada cai por terra. A função de influência corresponde à primeira derivada de um estimador ou de um teste, enquanto o *breakdown point*, outro conceito da área, mede a distância da singularidade mais próxima. Esta seção está baseada em Hampel et al. [7].

Definição 1.1.6. *A função de influência de um funcional T numa distribuição F no ponto x é o limite*

$$\mathrm{IF}(x, T, F) = \lim_{t\to 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t}, \tag{1.1}$$

onde δ_x é a distribuição de Dirac no ponto x.

Perceba que o limite acima pode ser calculado diretamente por

$$\mathrm{IF}(x, T, F) = \frac{d}{dt}\Big|_{t=0} T((1-t)F + t\delta_x). \tag{1.2}$$

Em outras palavras, a função de influência $\mathrm{IF}(x, T, F)$ é a derivada direcional de T numa distribuição F na direção δ_x. Isto significa que quando T é apropriadamente diferenciável num ponto F, podemos usar sua expansão de Taylor e substituir T, numa vizinhança de F, por uma aproximação linear.

Há alguns valores dados pela função de influência que nos permitem concluir sobre a robustez de um estimador T. O mais importante deles é o *gross error sensitivity* de T em uma distribuição F:

$$\gamma^*(T, F) = \gamma^* = \sup_x |\,\mathrm{IF}(x; T, F)|.$$

O valor γ^* dá a cota superior do comportamento assintótico causado por uma contaminação nos dados, isto é, γ^* é a pior influência que um estimador pode sofrer. Obviamente, uma propriedade desejável é que γ^* seja finito, isto é, que IF seja uma função limitada.

[4] A convergência em probabilidade para uma constante é equivalente a convergência em distribuição para esta constante. Para mais detalhes, consulte teorema 2.7 em van der Vaart [14].

Exemplo 1.1.1. *Seja* $T_n = \frac{1}{n}\sum X_i$ *a média aritmética cujo funcional correspondente é a média* $T(G) = \int u dG(u)$. *Usando* (1.2)*:*

$$\begin{aligned}
\mathrm{IF}(x, T, F) &= \frac{d}{dt}\Big|_{t=0} (T((1-t)F + t\delta_x)) = \frac{d}{dt}\Big|_{t=0} \left(\int u d[(1-t)F + t\delta_x](u) \right) \\
&= \frac{d}{dt}\Big|_{t=0} \left((1-t) \int u dF(u) + t \int u d\delta_x(u) \right) \\
&= \frac{d}{dt}\Big|_{t=0} \left(\int u dF(u) - t \int u dF(u) + tx \right) \\
&= x - \int u dF(u).
\end{aligned}$$

Note que obtemos uma função claramente ilimitada independente da distribuição avaliada, implicando $\gamma^* = \infty$. *Isto significa que mesmo um único outlier pode levar o estimador* T *a valores muito altos (ou muito baixos). A média, portanto, é mais suscetível a desvios e daí sua baixa robustez.*

O *breakdown point*, denotado por ε^*, mede a que distância o modelo está da realidade, dando uma visão qualitativa e global de robustez. Formalmente, o ε^* de uma sequência de estimadores $(T_n)_{n\geq 1}$ em F é

$$\varepsilon^* := \sup\{\varepsilon \leq 1 \mid \text{existe um compacto } K_\varepsilon \subseteq \Theta; \pi(F, G) < \varepsilon \implies G(\{T_n \in K_\varepsilon\}) \xrightarrow{n} 1\},$$

sendo $\Theta \subset \mathbb{R}$ o espaço de parâmetros e π a *distância de Prokhorov*, definida por

$$\pi(F, G) := \inf \{\, \varepsilon; F(A) \leq G(A^\varepsilon) + \varepsilon \text{ para todos os eventos } A \,\},$$

e A^ε é o conjunto dos pontos que estão a uma distância menor que ε de A.

O exemplo abaixo, fornecido sem prova, foi retirado de Hampel [6].

Exemplo 1.1.2. *A mediana possui breakdown point* $\frac{1}{2}$*, enquanto a média* α*-truncada*[5]*, definida como* $\int_\alpha^{1-\alpha} F^{-1}(t)dt/(1-2\alpha)$ *com* $0 < \alpha < \frac{1}{2}$ *tem breakdown point* α.

Grosso modo, o *breakdown point* fornece a menor fração de erros que o estimador pode ter. Intuitivamente isto significa que o valor de ε^* é o quanto dos seus dados podem ser mudados arbitrariamente antes que o estimador se torne inútil.

O estudo de estimadores busca altos valores de *breakdown point* e baixos valores de *gross error sensibility*. Uma maneira de avaliar um modelo quanto à sensibilidade a *outliers* é usar os estimadores clássicos em paralelo com estimadores robustos. Se ambos apresentarem resultados próximos, os efeitos são desprezíveis. Por outro lado, se apresentarem valores distintos, os *outliers* devem ser levados em conta.

O artigo de Rousseeuw [13] mostra que o estimador Q_n, apresentado nas seções seguintes, possui *breakdown point* de 50%, o melhor a ser obtido sob certas condições e IF suave e altamente eficiente[6] na distribuição gaussiana. Além disso, $\hat{\gamma}_Q$, que também será apresentado nas seções seguintes, tem *breakdown point* de 25% como prova Ma [12].

[5] Do inglês α-*trimmed mean*.

[6] Consulte Rousseeuw [13] para maiores detalhes.

1.2 Estimador robusto para função de covariância

Nesta seção apresentaremos dois estimadores Q_n e $\hat{\gamma}_Q$ e discutiremos suas propriedades assintóticas, assim como sua robustez. Denote por $D[0,\infty]$ o espaço das funções $f\colon [0,\infty] \to \mathbb{R}$ que são contínuas à direita e que possuem os limites à esquerda em todo ponto, conhecidas como *funções càdlàg*. O conjunto $\mathscr{M}([-\infty,\infty])$ é o conjunto das funções de distribuição acumulada definidas em $[-\infty,\infty]$ equipado com a topologia da convergência uniforme. Os resultados desta seção foram retirados do artigo de Lévy-Leduc et al. [11].

1.2.1 Estimador equivariante para função de covariância

Seguindo as linhas de Huber [9], vamos obter um estimador equivariante[7] para a covariância. Se X e Y são variáveis aleatórias com o quadrado integrável, é fácil ver que

$$\operatorname{cov}(X,Y) = \frac{1}{4ab}[\operatorname{Var}(aX+bY) - \operatorname{Var}(aX-bY)]$$

pois

$$\operatorname{Var}(aX \pm bY) = \quad a^2\,\mathbb{E}\,X^2 \pm 2ab\,\mathbb{E}\,XY + b^2\,\mathbb{E}\,Y^2 - a^2(\mathbb{E}\,X)^2 \mp 2ab\,\mathbb{E}\,X\,\mathbb{E}\,Y - b^2(\mathbb{E}\,Y)^2$$

logo

$$\operatorname{Var}(aX+bY) - \operatorname{Var}(aX-bY) = 4ab\operatorname{cov}(X,Y).$$

Quando substituirmos $\operatorname{Var}(\cdot)$ por $S^2(\cdot)$, sendo S um funcional de escala robusto tal que

$$S(aX+b) = |a|S(X),$$

isto é, S é equivariante, obtemos um estimador para a covariância de X e Y:

$$\mathrm{C_S}(X,Y) = \frac{1}{4ab}[S^2(aX+bY) - S^2(aX-bY)]. \tag{1.3}$$

Podemos normalizar S para que $S(X) = 1$ quando $X \sim \mathscr{N}(0,1)$. No caso em que (X,Y) tem distribuição normal bivariada, sabemos que $aX \pm bY \sim \mathscr{N}(\mu_\pm, \sigma_\pm^2)$ com $\mu_\pm \coloneqq a\mu_X \pm b\mu_Y$ e $\sigma_\pm^2 \coloneqq a^2\sigma_X^2 \pm 2ab\operatorname{cov}(X,Y) + b^2\sigma_Y^2$. Então a variável $\frac{aX\pm bY-\mu_\pm}{\sqrt{\sigma_\pm^2}}$ tem distribuição normal padrão[8]. Portanto, $S(\frac{aX\pm bY-\mu_\pm}{\sqrt{\sigma_\pm^2}}) = 1$. Pela propriedade de equivariância, vale

$$1 = S\Big(\frac{aX \pm bY - \mu_\pm}{\sqrt{\sigma_\pm^2}}\Big) = \frac{1}{|\sqrt{\sigma_\pm^2}|}S(aX \pm bY),$$

portanto,

$$S(aX+bY) = \sqrt{\sigma_+^2} \text{ e } S(aX-bY) = \sqrt{\sigma_-^2}.$$

Substituindo nas equações acima

$$\mathrm{C_S}(X,Y) = \frac{1}{4ab}[S^2(aX+bY) - S^2(aX-bY)] = \frac{1}{4ab}[\sqrt{\sigma_+^2}^2 - \sqrt{\sigma_-^2}^2] \tag{1.4}$$

$$= \frac{1}{4ab}[\sigma_+^2 - \sigma_-^2] = \operatorname{cov}(X,Y). \tag{1.5}$$

[7] Diz-se que S é equivariante quando $S(aX+b) = |a|S(X)$.
[8] Para maiores detalhes, consulte a seção sobre a distribuição normal multivariada no Apêndice.

Agora estamos aptos a definir Q como foi feito por Lévy-Leduc et al. [11]. Dadas duas cópias independentes X e Y com distribuição F, a probabilidade de a distância entre X e Y ser menor do que ou igual a r é medida pela integral de correlação de Grassberger-Procaccia:

$$r \mapsto U(r,F) = \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r\}} dF(x)dF(y).$$

Usando esta integral podemos definir um funcional de uma distribuição F que é proporcional ao primeiro quartil de $r \mapsto U(r,F)$:

$$Q(F_X) \coloneqq c(F_X)\inf\{r \geq 0 \mid U(r,F) \geq \frac{1}{4}\}. \tag{1.6}$$

Defina as seguintes aplicações:

$$\begin{aligned} &T_1 \colon \mathscr{M}([-\infty,\infty]) \to D[0,\infty] \\ &\quad F \mapsto T_1(F) = \Big\{r \mapsto \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r\}} dF(x)dF(y)\Big\}, \\ &T_2 \colon D[0,\infty] \to \mathbb{R} \\ &\quad U \mapsto U^{-1}(\frac{1}{4}) \end{aligned}$$

e

$$\begin{aligned} T_0 = T_2 \circ T_1 \colon \mathscr{M}([-\infty,\infty]) &\to \mathbb{R} \\ F &\mapsto U^{-1}(\frac{1}{4}). \end{aligned}$$

Denote por U^{-1} a inversa generalizada de U, a saber, $U^{-1}(x) = \inf\{r \geq 0 \mid U(r,F) \geq x\}$. Vamos definir o que é um processo gaussiano e aplicar nele o funcional Q.

Definição 1.2.1. *Um processo estocástico é uma coleção de variáveis aleatórias $(X_t)_{t\in T}$ definidas num mesmo espaço de probabilidade e indexadas por algum conjunto T. $(X_t)_{t\in T}$ é dito gaussiano quando, para qualquer subconjunto finito $T_0 \subset T$, o vetor aleatório $(X_t)_{t\in T_0}$ tem distribuição normal. Equivalentemente, $(X_t)_{t\in T}$ é gaussiano quando qualquer combinação linear finita $\sum_{t\in T_0} a_t X_t$ tem distribuição normal univariada.*

Dado um processo gaussiano $(X_i)_{i\geq 1}$ e sua distribuição empírica F_n, podemos expressar $Q(F_n) = Q_n(X_{1:n}, \Phi)$ por

$$Q_n(X_{1:n}, \Phi) = c(\Phi)T_0(F_n), \tag{1.7}$$

sendo $X_{1:n} = (X_1, \ldots, X_n)$ as n primeiras observações.

Substituindo S por Q (mostraremos mais adiante a equivariância de Q), temos a seguinte relação obtida de (1.3) e (1.5):

$$\operatorname{cov}(X,Y) = \frac{1}{4ab}[Q^2(aX+bY) - Q^2(aX-bY)].$$

Em particular, fazendo $X = Y$ e $a = b = 1$:

$$\operatorname{Var} X = \operatorname{cov}(X,X) = \frac{1}{4}[Q^2(2X) - Q^2(X-X)] = \frac{1}{4}Q^2(2X) = Q^2(X),$$

pois $Q^2(2X) = (2Q(X))^2 = 4Q^2(X)$. Resta uma definição necessária para enunciarmos o teorema principal da seção.

Definição 1.2.2. *Um processo estocástico $(X_t)_{t\in T}$ é dito estacionário quando cumpre*

i) $\mathbb{E}\,|X_t|^2 < \infty$ *para todo* $t \in T$*;*

ii) $\mathbb{E}\,X_t = m$ *para todo* $t \in T$*;*

iii) $\operatorname{cov}(X_t, X_s) = \operatorname{cov}(X_{t+k}, X_{s+k})$ *para todos* $s, t, k \in T$*.*

A noção de processo estacionário possui dois sentidos: um estrito, também conhecido como forte, e outro fraco. Trataremos aqui dos processos estacionários no sentido fraco, que é a Definição 1.2.2. Para maiores detalhes, consulte o Capítulo 1 de Brockwell [2].

O teorema principal, enunciado na Subseção 1.2.3, mostra que para um processo gaussiano estacionário com função de covariância absolutamente somável, o estimador $\hat{\gamma}_Q$ a ser definido é assintoticamente normal, e portanto consistente como já discutimos na seção anterior.

1.2.2 Lemas técnicos

Esta subseção concentra-se em demonstrar lemas necessários à prova do teorema principal. O próximo lema usa a definição do estimador Q para conseguir uma propriedade de equivariância para a função de influência aplicada num processo gaussiano.

Lema 1.2.1. *Considere* $(X_i)_{i\geq 1}$ *um processo gaussiano de média* μ *e variância* $\sigma^2 > 0$*. Então, para todo* $x \in \mathbb{R}$*,*

$$\operatorname{IF}(x, Q, \Phi_{\mu,\sigma}) = \sigma \operatorname{IF}(\frac{x-\mu}{\sigma}, Q, \Phi). \tag{1.8}$$

Demonstração. Pela definição de função de influência, temos:

$$\begin{aligned}
\operatorname{IF}(x, Q, \Phi_{\mu,\sigma}) &= \frac{d}{dt}\Big(Q((1-t)\Phi_{\mu,\sigma}(u) + t\delta_x(u))\Big)\big|_{t=0} \\
&= \frac{d}{dt}\Big(Q((1-t)\Phi(\frac{u-\mu}{\sigma}) + t\delta_{\frac{x-\mu}{\sigma}}(\frac{u-\mu}{\sigma}))\Big)\big|_{t=0} \\
&= \frac{d}{dt}\Big(\sigma Q((1-t)\Phi(u) + t\delta_{\frac{x-\mu}{\sigma}}(u)\Big)\big|_{t=0} = \sigma \operatorname{IF}(\frac{x-\mu}{\sigma}, Q, \Phi)
\end{aligned}$$

onde no último passo usamos a seguinte propriedade do estimador Q:

$$Q(F(\frac{\cdot-\mu}{\sigma})) = \sigma Q(F(\cdot)).$$

Provar esta propriedade é simples, supondo a existência e a Riemann-integrabilidade da função F' :

$$\begin{aligned}
T_1(F(\frac{\cdot-\mu}{\sigma})).[r] = U(r, F(\frac{\cdot-\mu}{\sigma})) &= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}} dF(\frac{x-\mu}{\sigma}) dF(\frac{y-\mu}{\sigma}) \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}} \frac{1}{\sigma}F'(\frac{x-\mu}{\sigma})\frac{1}{\sigma}F'(\frac{y-\mu}{\sigma}) dx dy \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|\tilde{x}\sigma+\mu-(\tilde{y}\sigma+\mu)|\leq r|\}} \frac{1}{\sigma}F'(\tilde{x})\frac{1}{\sigma}F'(\tilde{y})\sigma d\tilde{x}\sigma d\tilde{y} \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|\tilde{x}-\tilde{y}|\leq \frac{r}{\sigma}|\}} dF(\tilde{x}) dF(\tilde{y}) = U(\frac{r}{\sigma}, F(\cdot)) \\
&= T_1(F(\cdot)).[\frac{r}{\sigma}]
\end{aligned}$$

Perceba que as substituições $\tilde{x} = \frac{x-\mu}{\sigma}$ e $\tilde{y} = \frac{y-\mu}{\sigma}$ foram usadas no segundo passo. Como $\sigma \geq 0$ e usando $k = \frac{r}{\sigma}$, temos:

$$\begin{aligned} Q(F(\frac{\cdot - \mu}{\sigma})) = c(F)\inf\{r \geq 0 \mid U(r, F(\frac{\cdot - \mu}{\sigma})) \geq \frac{1}{4}\} &= c(F)\inf\{r \geq 0 \mid U(\frac{r}{\sigma}, F(\cdot)) \geq \frac{1}{4}\} \\ &= c(F)\inf\{\sigma k \geq 0 \mid U(k, F(\cdot)) \geq \frac{1}{4}\} \\ &= c(F)\sigma\inf\{k \geq 0 \mid U(k, F(\cdot)) \geq \frac{1}{4}\} \\ &= \sigma Q(F(\cdot)). \end{aligned}$$

□

Note que esse lema demonstra exatamente a propriedade de equivariância do estimador Q. De fato, $Q(F(\frac{\cdot-\mu}{\sigma})) = \sigma Q(F(\cdot))$ é equivalente, supondo que X tem distribuição acumulada F, a

$$Q(\sigma X + \mu) = \sigma Q(X).$$

O resultado acima será usado para obter uma expressão explícita da função de influência, o que permitirá conhecer a expansão assintótica de Q_n. Antes disso, esclareceremos mais uma notação: dada uma sequência de variáveis aleatórias $(X_n)_{n\in\mathbb{N}}$ convergindo em probabilidade para zero, usaremos o símbolo "pequeno oh"estocástico:

$$X_n = o_P(1) \text{ significa } X_n \to_P 0.$$

Lema 1.2.2. *Seja $(X_i)_{i\geq 1}$ um processo gaussiano estacionário de média μ e variância $\sigma^2 > 0$. Assuma que existe uma sequência não-decrescente (a_n) tal que $a_n(F_n - \Phi_{\mu,\sigma})$ converge fracamente em $(D[0,\infty], ||.||_\infty)$. Então, $Q_n(X_{1:n}, \Phi)$ definido acima tem a seguinte expansão assintótica:*

$$a_n(Q_n(X_{1:n}, \Phi) - \sigma) = \frac{a_n}{n}\sum_{i=1}^{n} \mathrm{IF}(X_i, Q, \Phi_{\mu,\sigma}) + o_P(1), \tag{1.9}$$

onde, para todo $x \in \mathbb{R}$,

$$\mathrm{IF}(x, Q, \Phi_{\mu,\sigma}) = \sigma\,\mathrm{IF}(\frac{x-\mu}{\sigma}, Q, \Phi), \quad e \tag{1.10}$$

$$\mathrm{IF}(X, Q, \Phi) = c(\Phi)\frac{\frac{1}{4} - \Phi(x + \frac{1}{c(\Phi)}) + \Phi(x - \frac{1}{c(\Phi)})}{\int_{\mathbb{R}} \varphi(y)\varphi(y + \frac{1}{c(\Phi)})dy}. \tag{1.11}$$

Demonstração. Denote por F a distribuição cumulativa $\Phi_{\mu,\sigma}$ de X_1. A convergência fraca de $a_n(F_n - F)$ em $(D[0,\infty], ||.||_\infty)$ é suficiente para aplicarmos o Delta Método Funcional[9] e obtermos a expansão assintótica (1.9). Para tanto, provaremos que a aplicação composta $T_0 = T_2 \circ T_1$ é Hadamard diferenciável e a Hadamard-diferencial correspondente está definida e é contínua em todo o espaço das funções *càdlàg*. Mostraremos apenas a diferenciabilidade do funcional T_1, já que a diferenciabilidade de T_2 segue do Lema 21.3 em van der Vaart [14]. Temos:

$$DT_2(F).g = -\frac{g(\xi_p)}{F'(\xi_p)}$$

sendo $F(\xi_p) = p$. Considere g uma função *càdlàg* e $(g_t)_t$ uma sequência de funções *càdlàg* de variação limitada tal que $||g_t - g||_\infty \to 0$ quando $t \to 0$. Para qualquer $r \geq 0$, temos:

[9] Discutiremos o Delta Método Funcional no Apêndice.

$$\begin{aligned}\frac{T_1(F+tg_t)[r]-T_1(F)[r]}{t} &= \frac{1}{t}\Big(\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}}dF(x)dF(y) + \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}}dF(x)dtg_t(y)\\ &+ \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}}dtg_t(x)dF(y) + \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}}dtg_t(x)dtg_t(y)\\ &- \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}}dF(x)dF(y)\Big)\\ &= 2\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}}dF(x)dg_t(y) + t\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}}dg_t(x)dg_t(y).\end{aligned}$$

Como

$$\begin{aligned}&|\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}}dF(x)dg_t(y) - \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}}dF(x)dg(y)|\\ &\qquad = |\int_{\mathbb{R}} g_t(x+r)-g_t(x-r)dF(x) - \int_{\mathbb{R}} g(x+r)-g(x-r)dF(x)|\\ &\qquad \leq |\int_{\mathbb{R}} g_t(x+r)-g(x+r)dF(x)| + |\int_{\mathbb{R}} g(x-r)-g_t(x-r)dF(x)|\\ &\qquad \leq 2||g_t-g||_\infty \to 0,\end{aligned}$$

quando $t\to 0$, a Hadamard-diferencial de T_1 em g é dada por:

$$(DT_1(F).g)(r) = 2\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\leq r|\}}dF(x)dg(y) = \int_{\mathbb{R}} g(x+r)-g(x-r)dF(x).$$

Então, pela regra da cadeia, Teorema B.3, obtemos a Hadamard diferenciabilidade de T_0 com a seguinte diferencial:

$$DT_0(F).g = \frac{-(DT_1(F).g)(T_0(F))}{(T_1(F)'[T_0(F)])} = \frac{-2\int_{\mathbb{R}} g(x+T_0(F))-g(x-T_0(F))dF(x)}{(T_1(F)'[T_0(F)])}. \tag{1.12}$$

Vamos calcular o denominador desta derivada. A expressão de $T_1(F)[r]$ pode ser escrita como

$$\int_{\mathbb{R}}(F(x+r)-F(x-r))dF(x) = \int_{\mathbb{R}}(F(x+r)-F(x-r))F'(x)dx.$$

Disto,

$$\begin{aligned}T_1(F)'[r] &= \frac{\partial}{\partial r}\left(\int_{\mathbb{R}}[F(x+r)-F(x-r)]F'(x)dx\right) = \int_{\mathbb{R}}\frac{\partial}{\partial r}\Big([F(x+r)-F(x-r)]F'(x)\Big)dx\\ &= \int_{\mathbb{R}}\Big[\frac{\partial}{\partial r}\Big(F(x+r)-F(x-r)\Big)F'(x) + (F(x+r)-F(x-r))\frac{\partial F'(x)}{\partial r}\Big]dx\\ &= \int_{\mathbb{R}}\Big[F'(x+r)+F'(x-r))F'(x)\Big]dx = \int_{\mathbb{R}}\Big(\frac{\varphi(\frac{x+r-\mu}{\sigma})}{\sigma}+\frac{\varphi(\frac{x-r-\mu}{\sigma})}{\sigma}\Big)\frac{\varphi(\frac{x-\mu}{\sigma})}{\sigma}dx.\end{aligned}$$

Com a substituição $y=\frac{x-\mu}{\sigma}$ temos $\sigma dy = dx$, donde vem

$$\begin{aligned}T_1(F)'[r] &= \int_{\mathbb{R}}\Big(\frac{\varphi(\frac{x+r-\mu}{\sigma})}{\sigma}+\frac{\varphi(\frac{x-r-\mu}{\sigma})}{\sigma}\Big)\frac{\varphi(\frac{x-\mu}{\sigma})}{\sigma}dx = \int_{\mathbb{R}}\Big(\frac{\varphi(y+\frac{r}{\sigma})}{\sigma}+\frac{\varphi(y-\frac{r}{\sigma})}{\sigma}\Big)\frac{\varphi(y)}{\sigma}\sigma dy\\ &= \frac{1}{\sigma}\Big[\int_{\mathbb{R}}\varphi(y+\frac{r}{\sigma})\varphi(y)dy + \int_{\mathbb{R}}\varphi(y-\frac{r}{\sigma})\varphi(y)dy\Big]\\ &= \frac{2}{\sigma}\int_{\mathbb{R}}\varphi(y+\frac{r}{\sigma})\varphi(y)dy.\end{aligned}$$

O último passo decorre da substituição $y = x + \frac{r}{\sigma}$:

$$\int_{\mathbb{R}} \varphi(y - \frac{r}{\sigma})\varphi(y)dy = \int_{\mathbb{R}} \varphi(x)\varphi(x + \frac{r}{\sigma})dx. \tag{1.13}$$

Finalmente, a equação (1.12) fica:

$$\begin{aligned} DT_0(F).g &= \frac{-2\int_{\mathbb{R}} g(x + T_0(F)) - g(x - T_0(F))dF(x)}{\frac{2}{\sigma}\int_{\mathbb{R}} \varphi(y + \frac{T_0(F)}{\sigma})\varphi(y)dy} \\ &= \frac{-\sigma\int_{\mathbb{R}} g(x + T_0(F)) - g(x - T_0(F))dF(x)}{\int_{\mathbb{R}} \varphi(y + \frac{T_0(F)}{\sigma})\varphi(y)dy}. \end{aligned}$$

Substituindo $\sigma = Q(F) = c(\Phi)T_0(F)$, temos:

$$DT_0(F).g = \frac{-\sigma\int_{\mathbb{R}} g(x + \frac{\sigma}{c(\Phi)}) - g(x - \frac{\sigma}{c(\Phi)})dF(x)}{\int_{\mathbb{R}} \varphi(y + \frac{1}{c(\Phi)})\varphi(y)dy}. \tag{1.14}$$

Agora podemos calcular a função de influência do funcional Q em F no ponto x usando a Hadamard-diferenciabilidade de Q:

$$\mathrm{IF}(X_i, Q, F) = DQ_F(\delta_{X_i} - F),$$

isto é, vamos calcular

$$\mathrm{IF}(X_i, Q, \Phi) = DQ_\Phi(\delta_{X_i} - \Phi) = Dc(\Phi)T_0(\Phi)(\delta_{X_i} - \Phi) = c(\Phi)DT_0(\Phi)(\delta_{X_i} - \Phi).$$

Usando as expressões anteriores, $DT_0(\Phi)(\delta_{X_i} - \Phi)$ é dado por

$$\frac{-1}{\int_{\mathbb{R}} \varphi(y + \frac{1}{c(\Phi)})\varphi(y)dy}\int_{\mathbb{R}} (\delta_{X_i} - \Phi)(x + \frac{1}{c(\Phi)}) - (\delta_{X_i} - \Phi)(x - \frac{1}{c(\Phi)}))d\Phi(x). \tag{1.15}$$

Calcularemos a integral explicitamente:

$$\begin{aligned} &\int_{\mathbb{R}} (\delta_{X_i} - \Phi)(x + \frac{1}{c(\Phi)})) - (\delta_{X_i} - \Phi)(x - \frac{1}{c(\Phi)}))d\Phi(x) \\ &\qquad = \int_{\mathbb{R}} \delta_{X_i}(x + \frac{1}{c(\Phi)})) - \delta_{X_i}(x - \frac{1}{c(\Phi)}))d\Phi(x) \\ &\qquad + \int_{\mathbb{R}} \Phi(x - \frac{1}{c(\Phi)})) - \Phi(x + \frac{1}{c(\Phi)}))d\Phi(x) \\ &\qquad = \Phi(X_i + \frac{1}{c(\Phi)}) - \Phi(X_i - \frac{1}{c(\Phi)}) - \frac{1}{4}, \end{aligned}$$

onde no último passo usamos

$$\int_{\mathbb{R}} \Phi(x - \frac{1}{c(\Phi)})) - \Phi(x + \frac{1}{c(\Phi)}))d\Phi(x) = \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y| \le T_0(\Phi)|\}}d\Phi(x)d\Phi(y) = T_1(\Phi)[T_0(\Phi)] = \frac{1}{4},$$

pois $T_0(F) = U^{-1}(\frac{1}{4})$ logo $T_1(F)[T_0(F)] = U(T_0(F), F) = U(U^{-1}(\frac{1}{4}), F) = \frac{1}{4}$. Tudo isto resulta em

$$DT_0(\Phi)(\delta_{X_i} - \Phi) = \frac{\frac{1}{4} - \Phi(X_i + \frac{1}{c(\Phi)}) + \Phi(X_i - \frac{1}{c(\Phi)})}{\int_{\mathbb{R}} \varphi(x + \frac{1}{c(\Phi)})\varphi(x)dx}. \tag{1.16}$$

Portanto,

$$\mathrm{IF}(X_i, Q, \Phi_{\mu,\sigma}) = \sigma\,\mathrm{IF}(\frac{X_i-\mu}{\sigma}, Q, \Phi) = \sigma c(\Phi)\frac{\frac{1}{4} - \Phi(\frac{X_i-\mu}{\sigma} + \frac{1}{c(\Phi)}) + \Phi(\frac{X_i-\mu}{\sigma} - \frac{1}{c(\Phi)})}{\int_{\mathbb{R}} \phi(y)\phi(y+\frac{1}{c(\Phi)})dy}$$

implicando

$$\begin{aligned}\frac{a_n}{n}\sum_{i=1}^{n}\mathrm{IF}(X_i, Q, \Phi_{\mu,\sigma}) &= \tfrac{a_n}{n}\textstyle\sum_{i=1}^{n}\sigma c(\Phi)\frac{\frac{1}{4}-\Phi(\frac{X_i-\mu}{\sigma}+\frac{1}{c(\Phi)})+\Phi(\frac{X_i-\mu}{\sigma}-\frac{1}{c(\Phi)})}{\int_{\mathbb{R}}\phi(y)\phi(y+\frac{1}{c(\Phi)})dy}\\ &= a_n\sigma c(\Phi)\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\frac{1}{4}-\Phi(\frac{X_i-\mu}{\sigma}+\frac{1}{c(\Phi)})+\Phi(\frac{X_i-\mu}{\sigma}-\frac{1}{c(\Phi)})}{\int_{\mathbb{R}}\phi(y)\phi(y+\frac{1}{c(\Phi)})dy}\\ &= c(\Phi)a_n\sigma\frac{\frac{1}{4}-\frac{1}{n}\sum_{i=1}^{n}\Phi(\frac{X_i-\mu}{\sigma}+\frac{1}{c(\Phi)})-\Phi(\frac{X_i-\mu}{\sigma}-\frac{1}{c(\Phi)})}{\int_{\mathbb{R}}\phi(y)\phi(y+\frac{1}{c(\Phi)})dy}. \qquad (1.17)\end{aligned}$$

Agora, vamos aplicar o Delta Método para funcionais, Teorema B.1 do Apêndice, da onde vem

$$r_n(\phi(T_n) - \phi(\theta)) = D\phi(\theta)(r_n(T_n - \theta)) + o_P(1).$$

Colocando $a_n = r_n, \phi = T_0, T_n = F_n, \theta = F$ ficamos:

$$a_n(T_0(F_n) - T_0(F)) = DT_0(F)(a_n(F_n - F)) + o_P(1).$$

Sabemos que $Q(F_n) = Q_n(X_{1:n}, \Phi) = c(\Phi)T_0(F_n)$ e $\sigma = Q(F) = c(\Phi)T_0(F)$, logo

$$\begin{aligned}a_n(\frac{Q_n(X_{1:n},\Phi)}{c(\Phi)} - \frac{\sigma}{c(\Phi)}) &= DT_0(F)(a_n(F_n - F)) + o_P(1).\\ \Rightarrow a_n(Q_n(X_{1:n},\Phi) - \sigma) &= c(\Phi)DT_0(F)(a_n(F_n - F)) + o_P(1).\end{aligned}$$

Substituindo $g = a_n(F_n - F)$ em (1.14), encontramos uma expressão para $DT_0(F).a_n(F_n - F)$:

$$\begin{aligned}DT_0(F).a_n(F_n - F) = \frac{-\sigma a_n}{\int_{\mathbb{R}}\varphi(y+\frac{1}{c(\Phi)})\varphi(y)dy}\Big[&\int_{\mathbb{R}} F_n(x+\frac{\sigma}{c(\Phi)}) - F_n(x-\frac{\sigma}{c(\Phi)})dF(x)\\ &- \int_{\mathbb{R}} F(x+\frac{\sigma}{c(\Phi)}) - F(x-\frac{\sigma}{c(\Phi)})dF(x)\Big].\end{aligned}$$

A última integral é facilmente calculada:

$$\int_{\mathbb{R}} F(x+T_0(F)) - F(x-T_0(F))dF(x) = \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y|\le T_0(F)|\}}dF(x)dF(y) = T_1(F)[T_0(F)] = \frac{1}{4},$$

pelo mesmo raciocínio já anteriormente citado. Para calcular a primeira, note que $F_n(x + T_0(F)) - F_n(x - T_0(F)) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{x-T_0(F)\le X_i\le x+T_0(F)\}} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i-T_0(F)\le x\le X_i+T_0(F)\}}$, donde segue

$$\begin{aligned}\int_{\mathbb{R}} F_n(x+T_0(F)) - F_n(x-T_0(F))dF(x) &= \int_{\mathbb{R}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i-T_0(F)\le x\le X_i+T_0(F)\}}dF(x)\\ &= \frac{1}{n}\sum_{i=1}^{n}F(X_i+\frac{\sigma}{c(\Phi)}) - F(X_i-\frac{\sigma}{c(\Phi)})\\ &= \frac{1}{n}\sum_{i=1}^{n}\Phi(\frac{X_i-\mu}{\sigma}+\frac{1}{c(\Phi)}) - \Phi(\frac{X_i-\mu}{\sigma}-\frac{1}{c(\Phi)}).\end{aligned}$$

Concluímos que $DT_0(F).a_n(F_n - F)$ é dado por

$$\frac{\sigma a_n}{\int_{\mathbb{R}} \varphi(y + \frac{1}{c(\Phi)})\varphi(y)dy}\Big[\frac{1}{4} - \frac{1}{n}\sum_{i=1}^{n} \Phi(\frac{X_i - \mu}{\sigma} + \frac{1}{c(\Phi)}) - \Phi(\frac{X_i - \mu}{\sigma} - \frac{1}{c(\Phi)})\Big].$$

Pela equação (1.17), finalizamos a demonstração:

$$\frac{a_n}{n}\sum_{i=1}^{n} \mathrm{IF}(X_i, Q, \Phi_{\mu,\sigma}) = c(\Phi)DT_0(F).a_n(F_n - F).$$

□

Antes de chegar no resultado principal, nos falta provar alguns lemas técnicos.

Lema 1.2.3. *Seja X uma variável aleatória gaussiana padrão. A função de influência definida em (1.11) tem as seguintes propriedades:*

$$\mathbb{E}[\mathrm{IF}(X, Q, \Phi)] = 0, \tag{1.18}$$
$$\mathbb{E}[X\,\mathrm{IF}(X, Q, \Phi)] = 0, \tag{1.19}$$
$$\mathbb{E}[X^2\,\mathrm{IF}(X, Q, \Phi)] = \frac{1}{2\sqrt{\pi}\beta}\exp(\frac{-1}{4c^2}) \neq 0. \tag{1.20}$$

sendo Φ a distribuição cumulativa da gaussiana padrão, $c = c(\Phi)$ como definido em (1.7) e $\beta = \int \varphi(y)\varphi(y + \frac{1}{c})dy$.

Demonstração. Perceba que $\mathrm{IF}(X, Q, \Phi)$ é uma função da variável X, portanto calcular sua esperança é calcular a integral abaixo:

$$\begin{aligned}\mathbb{E}[\mathrm{IF}(X, Q, \Phi)] &= \int \mathrm{IF}(X, Q, \Phi)d\Phi(x) = \int c\frac{\frac{1}{4} - \Phi(x + \frac{1}{c}) + \Phi(x - \frac{1}{c})}{\int_{\mathbb{R}} \phi(y)\phi(y + \frac{1}{c})dy}d\Phi(x) \\ &= \frac{c}{\beta}\Big[\frac{1}{4} - \int \Phi(x + \frac{1}{c}) - \Phi(x - \frac{1}{c})d\Phi(x)\Big]\end{aligned}$$

Logo, basta mostrar que $\int \Phi(x + \frac{1}{c}) - \Phi(x - \frac{1}{c})d\Phi(x) = \frac{1}{4}$:

$$\int \Phi(x + \frac{1}{c}) - \Phi(x - \frac{1}{c})d\Phi(x) = \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{|x-y| \leq \frac{1}{c}\}}d\Phi(x)d\Phi(y) = T_1(\Phi)[\frac{1}{c}] = \frac{1}{4}.$$

Analogamente, vê-se que

$$\begin{aligned}\mathbb{E}[X\,\mathrm{IF}(X, Q, \Phi)] &= \frac{c}{\beta}\Big[\frac{\mathbb{E}(X)}{4} - \int x\Phi(x + \frac{1}{c}) - x\Phi(x - \frac{1}{c})d\Phi(x)\Big] \\ &= \frac{c}{\beta}\Big[\int x\Phi(x + \frac{1}{c}) - x\Phi(x - \frac{1}{c})d\Phi(x)\Big]\end{aligned}$$

Portanto, basta ver que $\int x\Phi(x+\frac{1}{c}) - x\Phi(x-\frac{1}{c})d\Phi(x) = 0$:

$$\begin{aligned}
\int x\Phi(x+\frac{1}{c})d\Phi(x) &= \int x\mathbb{P}(X \le x+\frac{1}{c})d\Phi(x) = \int x\mathbb{P}(-X \le x+\frac{1}{c})d\Phi(x) \\
&= \int x\mathbb{P}(X \ge -x-\frac{1}{c})d\Phi(x) = \int x(1-\mathbb{P}(X \le -x-\frac{1}{c}))d\Phi(x) \\
&= \int x(1-\Phi(-x-\frac{1}{c})d\Phi(x) = \int xd\Phi(x) - \int x\Phi(-x-\frac{1}{c})d\Phi(x) \\
&= -\int_{-\infty}^{+\infty} x\Phi(-x-\frac{1}{c})d\Phi(x) = \int_{-\infty}^{+\infty} -x\Phi(-x-\frac{1}{c})\varphi(x)dx = \\
&= \int_{+\infty}^{-\infty} y\Phi(y-\frac{1}{c})\varphi(-y)(-1)dy = \int_{-\infty}^{+\infty} y\Phi(y-\frac{1}{c})\varphi(y)dy \\
&= \mathbb{E}[X\Phi(X-\frac{1}{c})]
\end{aligned}$$

usando a simetria da variável X.
Finalmente, usaremos integração por partes para calcular (1.20):

$$\begin{aligned}
\mathbb{E}[X^2\,\mathrm{IF}(X,Q,\Phi)] &= \frac{c}{\beta}\Big[\frac{\mathbb{E}(X^2)}{4} - \int x^2(\Phi(x+\frac{1}{c}) - \Phi(x-\frac{1}{c}))d\Phi(x)\Big] \\
&= \frac{c}{\beta}\Big[\frac{1}{4} - \int x^2(\Phi(x+\frac{1}{c}) - \Phi(x-\frac{1}{c}))\varphi(x)dx\Big] \\
&= \frac{c}{\beta}\Big[\frac{1}{4} - \int x^2(\int_{-\infty}^{x+\frac{1}{c}} \varphi(t)dt - \int_{-\infty}^{x-\frac{1}{c}} \varphi(t)dt)\varphi(x)dx\Big] \\
&= \frac{c}{\beta}\Big[\frac{1}{4} - \int x^2(\int_{x-\frac{1}{c}}^{x+\frac{1}{c}} \varphi(t)dt)\varphi(x)dx\Big] \\
&= \frac{c}{\beta}\Big[\frac{1}{4} - \int \Big(\int_{y-\frac{1}{c}}^{y+\frac{1}{c}} x^2\varphi(x)dx\Big)\varphi(y)dy\Big]
\end{aligned}$$

e substituindo $u = x$ e $x\varphi(x) = dv$, temos:

$$\int_{y-\frac{1}{c}}^{y+\frac{1}{c}} x^2\varphi(x)dx = uv\Big|_{y-\frac{1}{c}}^{y+\frac{1}{c}} - \int_{y-\frac{1}{c}}^{y+\frac{1}{c}} vdu$$

mas

$$\begin{aligned}
v\Big|_{y-\frac{1}{c}}^{y+\frac{1}{c}} = \int_{y-\frac{1}{c}}^{y+\frac{1}{c}} x\varphi(x)dx &= \int_{y-\frac{1}{c}}^{y+\frac{1}{c}} x\frac{1}{\sqrt{2\pi}}\exp(\frac{-x^2}{2})dx = \int_{(y-\frac{1}{c})^2}^{(y+\frac{1}{c})^2} \frac{1}{\sqrt{2\pi}}\exp(\frac{-u}{2})\frac{du}{2} \\
&= \frac{1}{\sqrt{2\pi}}(-2)\frac{1}{2}\exp(\frac{-u}{2})\Big|_{(y-\frac{1}{c})^2}^{(y+\frac{1}{c})^2} = -[\varphi(y+\frac{1}{c}) - \varphi(y-\frac{1}{c})]
\end{aligned}$$

resultando em

$$\begin{aligned}
\int_{y-\frac{1}{c}}^{y+\frac{1}{c}} x^2\varphi(x)dx &= x(-\varphi(x))\Big|_{y-\frac{1}{c}}^{y+\frac{1}{c}} - \int_{y-\frac{1}{c}}^{y+\frac{1}{c}} -\varphi(x)dx \\
&= -(y+\frac{1}{c})\varphi(y+\frac{1}{c}) + (y-\frac{1}{c})\varphi(y-\frac{1}{c}) + \int_{y-\frac{1}{c}}^{y+\frac{1}{c}} \varphi(x)dx.
\end{aligned}$$

Finalmente, a integral $-\int_{\mathbb{R}}\Big[\int_{y-\frac{1}{c}}^{y+\frac{1}{c}} x^2\varphi(x)dx\Big]\varphi(y)dy$ fica

$$\begin{aligned}
&\int_{\mathbb{R}}\Big[(y+\frac{1}{c})\varphi(y+\frac{1}{c})-(y-\frac{1}{c})\varphi(y-\frac{1}{c})-\int_{y-\frac{1}{c}}^{y+\frac{1}{c}}\varphi(x)dx\Big]\varphi(y)dy\\
&=\int_{\mathbb{R}}\Big[(y+\frac{1}{c})\varphi(y+\frac{1}{c})-(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy-\int_{\mathbb{R}}\Big[\int_{y-\frac{1}{c}}^{y+\frac{1}{c}}\varphi(x)dx\Big]\varphi(y)dy\\
&=\int_{\mathbb{R}}\Big[(y+\frac{1}{c})\varphi(y+\frac{1}{c})-(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy-\frac{1}{4}
\end{aligned}$$

pois

$$\int_{\mathbb{R}}\Big[\int_{y-\frac{1}{c}}^{y+\frac{1}{c}}\varphi(x)dx\Big]\varphi(y)dy=\int_{\mathbb{R}}\Big[\Phi(x+\frac{1}{c})-\Phi(x-\frac{1}{c})\Big]\varphi(x)dx=T_1(\Phi)[\frac{1}{c}]=\frac{1}{4}$$

e

$$\begin{aligned}
&\int_{\mathbb{R}}\Big[(y+\frac{1}{c})\varphi(y+\frac{1}{c})-(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy\\
&=\int_{\mathbb{R}}\Big[-(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy+\int_{\mathbb{R}}\Big[(y+\frac{1}{c})\varphi(y+\frac{1}{c})\Big]\varphi(y)dy\\
&=\int_{\mathbb{R}}\Big[-(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy+\int_{\mathbb{R}}\Big[-(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy\\
&=-2\int_{\mathbb{R}}\Big[(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy\\
&=\frac{1}{2c\sqrt{\pi}}\exp(\frac{-1}{4c^2})
\end{aligned}$$

sendo o penúltimo passo dado pela substituição $y=-x$:

$$\begin{aligned}
\int_{\mathbb{R}}\Big[(x+\frac{1}{c})\varphi(x+\frac{1}{c})\Big]\varphi(x)dx&=\int_{+\infty}^{-\infty}\Big[(-y+\frac{1}{c})\varphi(-y+\frac{1}{c})\Big]\varphi(-y)(-1)dy\\
&=\int_{+\infty}^{-\infty}\Big[(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy\\
&=-\int_{-\infty}^{+\infty}\Big[(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy.
\end{aligned}$$

O cálculo da última integral se resume em duas outras integrais:

$$-2\int_{\mathbb{R}}\Big[(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy=\quad 2\int_{\mathbb{R}}\Big[\frac{1}{c}\varphi(y-\frac{1}{c})\Big]\varphi(y)dy-2\int_{\mathbb{R}}\Big[y\varphi(y-\frac{1}{c})\Big]\varphi(y)dy.$$

Segue o cálculo da primeira:

$$
\begin{aligned}
2\int_{\mathbb{R}} \Big[\frac{1}{c}\varphi(y-\frac{1}{c})\Big]\varphi(y)dy &= 2\int_{\mathbb{R}} \frac{1}{c}\frac{1}{\sqrt{2\pi}}\exp(\frac{-(y-\frac{1}{c})^2}{2})\frac{1}{\sqrt{2\pi}}\exp(\frac{-y^2}{2})dy \\
&= \frac{2}{c\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp(\frac{-(y^2-\frac{2y}{c}+\frac{1}{c^2})-y^2}{2})dy \\
&= \frac{2}{c\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp(\frac{-2(y^2-\frac{y}{c}+\frac{1}{2c^2})}{2}dy \\
&= \frac{2}{c\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp(\frac{-2[(y-\frac{1}{2c})^2+\frac{1}{4c^2}]}{2}dy \\
&= \exp(\frac{-1}{4c^2})\frac{2}{c\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp(\frac{-[\sqrt{2}(y-\frac{1}{2c})]^2}{2}dy \\
&= \exp(\frac{-1}{4c^2})\frac{2}{c\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp\frac{-u^2}{2}\frac{du}{\sqrt{2}} \\
&= \exp(\frac{-1}{4c^2})\frac{1}{c\sqrt{\pi}}.
\end{aligned}
$$

Para calcular a segunda integral,

$$
\begin{aligned}
-2\int_{\mathbb{R}} \Big[y\varphi(y-\frac{1}{c})\Big]\varphi(y)dy &= -2\int_{\mathbb{R}} y\frac{1}{\sqrt{2\pi}}\exp(\frac{-(y-\frac{1}{c})^2}{2})\frac{1}{\sqrt{2\pi}}\exp(\frac{-y^2}{2})dy \\
&= \frac{-2}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} y\exp(\frac{-(y-\frac{1}{c})^2}{2})\exp(\frac{-y^2}{2})dy \\
&= \frac{-2}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} y\exp(\frac{-2[(y-\frac{1}{2c})^2+\frac{1}{4c^2}]}{2}dy \\
&= \frac{-2}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\exp(\frac{-1}{4c^2})\int_{\mathbb{R}}(\frac{1}{2c}-\frac{1}{2}(\frac{1}{c}-2y))\exp(-(y-\frac{1}{2c})^2)dy \\
&= \frac{-2}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\exp(\frac{-1}{4c^2})\frac{1}{2c}\int_{\mathbb{R}}\exp(-(y-\frac{1}{2c})^2)dy \\
&+ \frac{-2}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\exp(\frac{-1}{4c^2})\frac{1}{2}\int_{\mathbb{R}}(2y-\frac{1}{c})\exp(-(y-\frac{1}{2c})^2)dy.
\end{aligned}
$$

Note que escrevemos $y = \frac{1}{2c} - \frac{1}{2}(\frac{1}{c} - 2y)$ e chegamos numa soma de integrais. A segunda integral é nula, fato que pode ser observado de duas maneiras: a substituição $u = (y - \frac{1}{2c})^2$ implica $du = dy(2y - \frac{1}{c})$ e torna os limites de integração iguais; o integrando é uma translação horizontal

para direita de $\frac{1}{2c}$ da função ímpar $f(x) = 2x\exp(-x^2)$. Para calcular o que resta,

$$\begin{aligned}
&= \frac{-2}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}}\exp(\frac{-1}{4c^2})\frac{1}{2c}\int_{\mathbb{R}}\exp(\frac{-2(y-\frac{1}{2c})^2}{2})dy \\
&= \frac{-2}{\sqrt{2\pi}}\exp(\frac{-1}{4c^2})\frac{1}{2c}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp(\frac{-1}{2}[\sqrt{2}(y-\frac{1}{2c})]^2)dy \\
&= \frac{-2}{\sqrt{2\pi}}\exp(\frac{-1}{4c^2})\frac{1}{2c}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp(\frac{-1}{2}[u]^2)\frac{du}{\sqrt{2}} \\
&= \frac{-1}{\sqrt{\pi}}\exp(\frac{-1}{4c^2})\frac{1}{2c} \\
&= \frac{-1}{2c\sqrt{\pi}}\exp(\frac{-1}{4c^2}).
\end{aligned}$$

Por fim,

$$-2\int_{\mathbb{R}}\Big[(y-\frac{1}{c})\varphi(y-\frac{1}{c})\Big]\varphi(y)dy = \frac{1}{2c\sqrt{\pi}}\exp(\frac{-1}{4c^2}),$$

como queríamos demonstrar. □

Lema 1.2.4. *Seja (X,Y) um vetor gaussiano padrão tal que $\mathrm{cov}(X,Y)=0$ e sejam Φ_+ e Φ_-, respectivamente, a distribuição acumulada de $X+Y$ e $X-Y$. A função de influência ψ, definida para todo x e y em $\mathbb{R}$ por*

$$\psi(x,y) = \frac{1}{2}[Q(\Phi_+)\,\mathrm{IF}(x+y,Q,\Phi_+) - Q(\Phi_-)\,\mathrm{IF}(x-y,Q,\Phi_-)]$$

satisfaz as seguintes propriedades:

$$\begin{aligned}
\mathbb{E}[\psi(X,Y)] &= 0, &(1.21)\\
\mathbb{E}[X\psi(X,Y)] = \mathbb{E}[Y\psi(X,Y)] &= 0, &(1.22)\\
\mathbb{E}[XY\psi(X,Y)] &\neq 0. &(1.23)
\end{aligned}$$

Demonstração. Sabemos que $\mathrm{cov}(X,Y)=0$ é suficiente para que X e Y sejam independentes. E, se X e Y são independentes com $X \sim N(\mu_1,\sigma_1^2)$ e $Y \sim N(\mu_2,\sigma_2^2)$, então $c_1X + c_2Y \sim N(c_1\mu_1+c_1\mu_2, c_1^2\sigma_1^2+c_2^2\sigma_2^2)$. Portanto, $\frac{X\pm Y}{\sqrt{\mathrm{Var}(X\pm Y)}}$ tem distribuição normal padrão. Esta variável será denotada por U. Usando $Q(\Phi_\pm) = \sqrt{\mathrm{Var}(X\pm Y)}$ e a equação (1.10), temos:

$$\mathrm{IF}(X\pm Y,Q,\Phi_\pm) = Q(\Phi_\pm)\,\mathrm{IF}(\frac{X\pm Y-0}{Q(\Phi_\pm)},Q,\Phi) = Q(\Phi_\pm)\,\mathrm{IF}(U,Q,\Phi). \tag{1.24}$$

Portanto,

$$\begin{aligned}
\mathbb{E}[\psi(X,Y)] &= \mathbb{E}\left[\frac{1}{2}[Q(\Phi_+)\,\mathrm{IF}(X+Y,Q,\Phi_+) - Q(\Phi_-)\,\mathrm{IF}(X-Y,Q,\Phi_-)]\right] \\
&= \frac{1}{2}Q(\Phi_+)^2\,\mathbb{E}\,\mathrm{IF}(U,Q,\Phi_+) - \frac{1}{2}Q(\Phi_-)^2\,\mathbb{E}\,\mathrm{IF}(U,Q,\Phi_-) \\
&= \frac{1}{2}[Q(\Phi_+)^2 - Q(\Phi_-)^2]\,\mathbb{E}\,\mathrm{IF}(U,Q,\Phi) = 0.
\end{aligned}$$

Perceba que $\frac{1}{2}[Q(\Phi_+)^2 - Q(\Phi_-)^2] = 2\,\mathrm{cov}(X,Y) = 0$. Além disso, $\mathbb{E}\,\mathrm{IF}(U,Q,\Phi) = 0$ pela equação (1.18).

Para a segunda propriedade, note primeiramente que

$$\mathbb{E}[X\psi(X,Y)] = \frac{1}{2}\Big[\mathbb{E}[(X+Y)\psi(X,Y)] + \mathbb{E}[(X-Y)\psi(X,Y)]\Big].$$

Se $U = \frac{X+Y}{Q(\Phi_+)}$ e $V = \frac{X-Y}{Q(\Phi_-)}$, então novamente pela equação (1.10):

$$\mathrm{IF}(X+Y,Q,\Phi_+) = Q(\Phi_+)\,\mathrm{IF}(\frac{X+Y-0}{Q(\Phi_+)},Q,\Phi) = Q(\Phi_+)\,\mathrm{IF}(U,Q,\Phi), \tag{1.25}$$

$$\text{e } \mathrm{IF}(X-Y,Q,\Phi_-) = Q(\Phi_-)\,\mathrm{IF}(\frac{X-Y-0}{Q(\Phi_-)},Q,\Phi) = Q(\Phi_-)\,\mathrm{IF}(V,Q,\Phi), \tag{1.26}$$

donde segue

$$\begin{aligned}
\mathbb{E}[(X+Y)\psi(X,Y)] &= \frac{1}{2}\,\mathbb{E}\left[(X+Y)[Q(\Phi_+)^2\,\mathrm{IF}(U,Q,\Phi) - Q(\Phi_-)^2\,\mathrm{IF}(V,Q,\Phi)]\right] \\
&= \frac{1}{2}\,\mathbb{E}\left[UQ(\Phi_+)^3\,\mathrm{IF}(U,Q,\Phi) - UQ(\Phi_+)Q(\Phi_-)^2\,\mathrm{IF}(V,Q,\Phi)\right] \\
&= \frac{Q(\Phi_+)^3}{2}\,\mathbb{E}\left[U\,\mathrm{IF}(U,Q,\Phi)\right] - \frac{Q(\Phi_+)Q(\Phi_-)^2}{2}\,\mathbb{E}\left[U\,\mathrm{IF}(V,Q,\Phi)\right] \\
&= -\frac{Q(\Phi_+)Q(\Phi_-)^2}{2}\,\mathbb{E}\,U\,\mathbb{E}\,\mathrm{IF}(V,Q,\Phi) = 0.
\end{aligned}$$

No último passo usamos a equação (1.19) e a independência entre U e V. Analogamente, prova-se que $\mathbb{E}[(X-Y)\psi(X,Y)] = 0$, o que conclui a demonstração da segunda propriedade. Para a última, considere a seguinte relação:

$$4XY = (X+Y)^2 - (X-Y)^2,$$

a qual implica

$$\begin{aligned}
\mathbb{E}[XY\psi(X,Y)] &= \mathbb{E}\left[\frac{(X+Y)^2-(X-Y)^2}{8}[Q(\Phi_+)^2\,\mathrm{IF}(U,Q,\Phi_+) - Q(\Phi_-)^2\,\mathrm{IF}(V,Q,\Phi_-)]\right] \\
&= \frac{1}{8}\,\mathbb{E}\left[U^2Q(\Phi_+)^4\,\mathrm{IF}(U,Q,\Phi_+)\right] + \frac{1}{8}\,\mathbb{E}\left[V^2Q(\Phi_-)^4\,\mathrm{IF}(V,Q,\Phi_-)\right] \\
&- \frac{1}{8}\,\mathbb{E}\left[U^2Q(\Phi_+)^2Q(\Phi_-)^2\,\mathrm{IF}(V,Q,\Phi_-)\right] - \frac{1}{8}\,\mathbb{E}\left[V^2Q(\Phi_-)^2Q(\Phi_+)^2\,\mathrm{IF}(U,Q,\Phi_+)\right] \\
&= \frac{Q(\Phi_+)^4}{8}\,\mathbb{E}\left[U^2\,\mathrm{IF}(U,Q,\Phi_+)\right] + \frac{Q(\Phi_-)^4}{8}\,\mathbb{E}\left[V^2\,\mathrm{IF}(V,Q,\Phi_-)\right] \\
&- \frac{Q(\Phi_+)^2Q(\Phi_-)^2}{8}\,\mathbb{E}\,U^2\,\mathbb{E}\left[\mathrm{IF}(V,Q,\Phi_-)\right] - \frac{Q(\Phi_-)^2Q(\Phi_+)^2}{8}\,\mathbb{E}\,V^2\,\mathbb{E}\left[\mathrm{IF}(U,Q,\Phi_+)\right] \\
&= \frac{Q(\Phi_+)^4}{8}\,\mathbb{E}\left[U^2\,\mathrm{IF}(U,Q,\Phi_+)\right] + \frac{Q(\Phi_-)^4}{8}\,\mathbb{E}\left[V^2\,\mathrm{IF}(V,Q,\Phi_-)\right] \neq 0,
\end{aligned}$$

onde usamos as equações (1.20) e (1.18) a independência de U e V. □

Perceba que o fato fundamental ao longo da demonstração desse lema foi a independência entre as variáveis $X+Y$ e $X-Y$. Por esse motivo, podemos substituir a hipótese de $\mathrm{cov}(X,Y) = 0$ pela estacionaridade do processo gaussiano e ainda será possível provar as mesmas três propriedades da função ψ.

Lema 1.2.5. *Sejam $(X_i)_{i\geq 1}$ um processo gaussiano estacionário de média zero com função de covariância $\gamma(h) = \mathbb{E}(X_1 X_{h+1})$ e Φ_+ e Φ_-, respectivamente, a distribuição acumulada de $X_i + X_{i+h}$ e $X_i - X_{i+h}$. A função de influência ψ, definida para todo x e y em $\mathbb{R}$ por*

$$\psi(x,y) = \frac{1}{2}[Q(\Phi_+)\,\mathrm{IF}(x+y, Q, \Phi_+) - Q(\Phi_-)\,\mathrm{IF}(x-y, Q, \Phi_-)]$$

satisfaz as seguintes propriedades:

$$\begin{aligned} \mathbb{E}[\psi(X_i, X_{i+h})] &= 0, && (1.27)\\ \mathbb{E}[X_i\psi(X_i, X_{i+h})] = \mathbb{E}[X_{i+h}\psi(X_i, X_{i+h})] &= 0, && (1.28)\\ \mathbb{E}[X_i X_{i+h}\psi(X_i, X_{i+h})] &\neq 0. && (1.29) \end{aligned}$$

Demonstração. Começaremos a prova mostrando que $X_i + X_{i+h}$ e $X_i - X_{i+h}$ são variáveis aleatórias independentes. Sabemos que as coordenadas de um vetor gaussiano multivariado são independentes se, e somente se, são não-correlacionadas. Temos:

$$\begin{aligned} \mathrm{cov}(X_i + X_{i+h}, X_i - X_{i+h}) &= \mathbb{E}[(X_i + X_{i+h})(X_i - X_{i+h})] - \mathbb{E}(X_i + X_{i+h})\,\mathbb{E}(X_i - X_{i+h})\\ &= \mathbb{E}[X_i^2 - X_{i+h}^2] = \mathrm{Var}\,X_i - \mathrm{Var}\,X_{i+h}\\ &= \mathrm{cov}(X_i, X_i) - \mathrm{cov}(X_{i+h}, X_{i+h}) = \gamma(0) - \gamma(0) = 0, \end{aligned}$$

onde no último passo usamos a estacionaridade do processo. Resta provar que $\mathbf{X} = (X_i + X_{i+h}, X_i - X_{i+h})$ é normal bivariado[10]. Para isto, considere um vetor real $\mathbf{a}$ qualquer:

$$\begin{aligned} \mathbf{a}^T\mathbf{X} &= a_1\,(X_i + X_{i+h}) + a_2\,(X_i - X_{i+h})\\ &= (a_1 + a_2)X_i + (a_1 - a_2)X_{i+h} = (a_1 + a_2, a_1 - a_2)^T\,(X_i, X_{i+h}). \end{aligned}$$

O vetor (X_i, X_{i+h}) é normal bivariado já que o processo é gaussiano. Logo, qualquer combinação linear de suas coordenadas tem distribuição normal. Concluímos que $X_i + X_{i+h}$ e $X_i - X_{i+h}$ são independentes. Além disso, para concluir que $X_i + X_{i+h}$ e $X_i - X_{i+h}$ ambas tem distribuição normal, basta tomar $a_1 = 1$ e $a_2 = 0$ no primeiro caso e $a_1 = 0$ e $a_2 = 1$ no segundo caso.

Note que $\frac{X_i \pm X_{i+h}}{\sqrt{\mathrm{Var}(X_i \pm X_{i+h})}}$ tem distribuição normal padrão. Esta variável será denotada por U. Usando $Q(\Phi_\pm) = \sqrt{\mathrm{Var}(X \pm Y)}$ e a equação (1.10), vale:

$$\mathrm{IF}(X_i \pm X_{i+h}, Q, \Phi_\pm) = Q(\Phi_\pm)\,\mathrm{IF}\Big(\frac{X_i \pm X_{i+h} - 0}{Q(\Phi_\pm)}, Q, \Phi\Big) = Q(\Phi_\pm)\,\mathrm{IF}(U, Q, \Phi). \qquad (1.30)$$

Logo, temos:

$$\begin{aligned} \mathbb{E}[\psi(X_i, X_{i+h})] &= \mathbb{E}\left[\frac{1}{2}[Q(\Phi_+)\,\mathrm{IF}(X_i + X_{i+h}, Q, \Phi_+) - Q(\Phi_-)\,\mathrm{IF}(X_i - X_{i+h}, Q, \Phi_-)]\right]\\ &= \frac{1}{2}Q(\Phi_+)^2\,\mathbb{E}\,\mathrm{IF}(U, Q, \Phi_+) - \frac{1}{2}Q(\Phi_-)^2\,\mathbb{E}\,\mathrm{IF}(U, Q, \Phi_-)\\ &= \frac{1}{2}[Q(\Phi_+)^2 - Q(\Phi_-)^2]\,\mathbb{E}\,\mathrm{IF}(U, Q, \Phi) = 0. \end{aligned}$$

Perceba que $\frac{1}{2}[Q(\Phi_+)^2 - Q(\Phi_-)^2] = 2\,\mathrm{cov}(X_i, X_{i+h}) = \gamma(h)$. Quando as variáveis são independentes, esse termo é automaticamente nulo. Caso contrário, recorremos à equação (1.18).

[10] Veja no Apêndice maiores detalhes sobre esta definição.

Para a segunda propriedade, note primeiramente que

$$\mathbb{E}[X_i\psi(X_i, X_{i+h})] = \frac{1}{2}\Big[\,\mathbb{E}[(X_i + X_{i+h})\psi(X_i, X_{i+h})] + \mathbb{E}[(X_i - X_{i+h})\psi(X_i, X_{i+h})]\Big]. \quad (1.31)$$

Se $U = \frac{X_i+X_{i+h}}{Q(\Phi_+)}$ e $V = \frac{X_i-X_{i+h}}{Q(\Phi_-)}$, então pela equação (1.10):

$$\mathrm{IF}(X_i + X_{i+h}, Q, \Phi_+) = Q(\Phi_+)\,\mathrm{IF}(\frac{X_i + X_{i+h} - 0}{Q(\Phi_+)}, Q, \Phi) = Q(\Phi_+)\,\mathrm{IF}(U, Q, \Phi)$$
$$\text{e } \mathrm{IF}(X_i - X_{i+h}, Q, \Phi_-) = Q(\Phi_-)\,\mathrm{IF}(\frac{X_i - X_{i+h} - 0}{Q(\Phi_-)}, Q, \Phi) = Q(\Phi_-)\,\mathrm{IF}(V, Q, \Phi).$$

Portanto,

$$\begin{aligned}
\mathbb{E}[(X_i + X_{i+h})\psi(X, Y)] &= \frac{1}{2}\,\mathbb{E}\left[(X_i + X_{i+h})Q(\Phi_+)^2\,\mathrm{IF}(U, Q, \Phi) - (X_i + X_{i+h})Q(\Phi_-)^2\,\mathrm{IF}(V, Q, \Phi)\right] \\
&= \frac{1}{2}\,\mathbb{E}\left[UQ(\Phi_+)^3\,\mathrm{IF}(U, Q, \Phi) - UQ(\Phi_+)Q(\Phi_-)^2\,\mathrm{IF}(V, Q, \Phi)\right] \\
&= \frac{Q(\Phi_+)^3}{2}\,\mathbb{E}\left[U\,\mathrm{IF}(U, Q, \Phi)\right] - \frac{Q(\Phi_+)Q(\Phi_-)^2}{2}\,\mathbb{E}\left[U\,\mathrm{IF}(V, Q, \Phi)\right] \\
&= -\frac{Q(\Phi_+)Q(\Phi_-)^2}{2}\,\mathbb{E}\,U\,\mathbb{E}\,\mathrm{IF}(V, Q, \Phi) = 0.
\end{aligned}$$

No último passo usamos a equação (1.19) e a independência entre U e V. Analogamente, prova-se que $\mathbb{E}[(X_i - X_{i+h})\psi(X_i, X_{i+h})] = 0$, concluindo a demonstração da segunda propriedade. Para a última, considere a seguinte relação:

$$4X_iX_{i+h} = (X_i + X_{i+h})^2 - (X_i - X_{i+h})^2,$$

donde vem

$$\begin{aligned}
\mathbb{E}[X_iX_{i+h}\psi(X_i, X_{i+h})] &= \mathbb{E}\left[\frac{(X_i + X_{i+h})^2 - (X_i - X_{i+h})^2}{8}[Q(\Phi_+)^2\,\mathrm{IF}(U, Q, \Phi) - Q(\Phi_-)^2\,\mathrm{IF}(V, Q, \Phi)]\right] \\
&= \frac{1}{8}\,\mathbb{E}\left[U^2Q(\Phi_+)^4\,\mathrm{IF}(U, Q, \Phi)\right] + \frac{1}{8}\,\mathbb{E}\left[V^2Q(\Phi_-)^4\,\mathrm{IF}(V, Q, \Phi)\right] \\
&- \frac{1}{8}\,\mathbb{E}\left[U^2Q(\Phi_+)^2Q(\Phi_-)^2\,\mathrm{IF}(V, Q, \Phi)\right] - \frac{1}{8}\,\mathbb{E}\left[V^2Q(\Phi_-)^2Q(\Phi_+)^2\,\mathrm{IF}(U, Q, \Phi)\right] \\
&= \frac{Q(\Phi_+)^4}{8}\,\mathbb{E}\left[U^2\,\mathrm{IF}(U, Q, \Phi)\right] + \frac{Q(\Phi_-)^4}{8}\,\mathbb{E}\left[V^2\,\mathrm{IF}(V, Q, \Phi)\right] \\
&- \frac{Q(\Phi_+)^2Q(\Phi_-)^2}{8}\,\mathbb{E}\,U^2\,\mathbb{E}\,\mathrm{IF}(V, Q, \Phi) - \frac{Q(\Phi_-)^2Q(\Phi_+)^2}{8}\,\mathbb{E}\,V^2\,\mathbb{E}\,\mathrm{IF}(U, Q, \Phi) \\
&= \frac{Q(\Phi_+)^4}{8}\,\mathbb{E}\left[U^2\,\mathrm{IF}(U, Q, \Phi)\right] + \frac{Q(\Phi_-)^4}{8}\,\mathbb{E}\left[V^2\,\mathrm{IF}(V, Q, \Phi)\right] \neq 0,
\end{aligned}$$

onde usamos as equações (1.20) e (1.18) e a independência entre U e V. □

1.2.3 Prova do teorema principal

Lembre-se que, no caso de um processo gaussiano, vale

$$\begin{aligned}
\gamma(h) = \mathrm{cov}(X_i, X_{i+h}) &= \frac{1}{4}[\mathrm{Var}(X_i + X_{i+h}) - \mathrm{Var}(X_i + X_{i+h})] \\
&= \frac{1}{4}[Q^2(X_i + X_{i+h}) - Q^2(X_i + X_{i+h})]. \quad (1.32)
\end{aligned}$$

E quando aplicamos o funcional Q na distribuição empírica, obtemos um estimador robusto para a função de covariância e podemos finalmente definir $\hat{\gamma}_Q$:

$$\hat{\gamma}_Q(h, X_{1:n}, \Phi) \coloneqq \frac{1}{4}\Big\{Q^2_{n-h}(X_{1:n-h} + X_{1+h:n}, \Phi) - Q^2_{n-h}(X_{1:n-h} - X_{1+h:n}, \Phi)\Big\},$$

cuja distribuição assintótica é dada pelo próximo teorema.

Teorema 1.2.1. *Seja $(X_i)_{i\geq 1}$ um processo gaussiano estacionário de média zero com função de covariância $\gamma(h) = \mathbb{E}(X_1 X_{h+1})$ satisfazendo:*

$$\sum_{h\geq 1} |\gamma(h)| < \infty. \tag{A1}$$

Seja h um inteiro não-negativo. Então, o estimador de autocovariância $\hat{\gamma}_Q(h, X_{1:n}, \Phi)$ satisfaz o seguinte teorema central do limite:

$$\sqrt{n}(\hat{\gamma}_Q(h, X_{1:n}, \Phi) - \gamma(h)) \to_d \mathscr{N}(0, \check{\sigma}^2_h),$$

onde

$$\check{\sigma}^2_h = \mathbb{E}[\psi^2(X_1, X_{1+h})] + 2\sum_{k\geq 1} \mathbb{E}[\psi(X_1, X_{1+h})\psi(X_{k+1}, X_{k+1+h})], \tag{1.33}$$

e a função ψ é definida por

$$\psi\colon (x,y) \mapsto \left\{ (\gamma(0)+\gamma(h))\,\mathrm{IF}\left(\frac{x+y}{\sqrt{2(\gamma(0)+\gamma(h))}}, Q, \Phi\right) - (\gamma(0)-\gamma(h))\,\mathrm{IF}\left(\frac{x-y}{\sqrt{2(\gamma(0)+\gamma(h))}}, Q, \Phi\right) \right\} \tag{1.34}$$

e a função IF *definida em* (1.11).

Demonstração. Sejam $\Phi_{\sigma,+}$ $\Phi_{\sigma,-}$ a distribuição de $(X_i + X_{i+h})_{i\geq 1}$ e $(X_i - X_{i+h})_{i\geq 1}$ respectivamente. Denote por $F_{+,n-h}$ e $F_{-,n-h}$ a distribuição empírica de $(X_i + X_{i+h})_{i\geq 1}$ e $(X_i - X_{i+h})_{i\geq 1}$ respectivamente. Como $(X_i)_{i\geq 1}$ satisfaz (A1), o mesmo ocorre para $(X_i + X_{i+h})_{i\geq 1}$ e $(X_i - X_{i+h})_{i\geq 1}$. Usando o teorema de Csörgő [5], obtemos a convergência fraca de $\sqrt{n-h}(F_{+,n-h} - \Phi_{\sigma,+})$ para um processo gaussiano no espaço das funções *càdlàg* com a topologia da convergência uniforme. O mesmo ocorre para $\sqrt{n-h}(F_{+,n-h} - \Phi_{\sigma,+})$. Consequentemente, temos a expansão assintótica (1.9) para $Q_{n-h}(X_{1:n-h} + X_{1+h:n}, \Phi)$ e $Q_{n-h}(X_{1:n-h} - X_{1+h:n}, \Phi)$ com $a_n = \sqrt{n-h}$:

$$\sqrt{n-h}(Q_{n-h}(X_{1:n-h} \pm X_{1+h:n}, \Phi) - Q(\Phi_{\sigma,\pm})) = \frac{\sqrt{n-h}}{n}\sum_{i=1}^{n-h} \mathrm{IF}(X_i \pm X_{i+h}, Q, \Phi_{\sigma,\pm}) + o_P(1).$$

Como $\frac{\sqrt{n-h}}{n} \to \frac{1}{\sqrt{n-h}}$ quando $n \to \infty$, reescrevemos as equações acima:

$$\sqrt{n-h}(Q_{n-h}(X_{1:n-h} \pm X_{1+h:n}, \Phi) - Q(\Phi_{\sigma,\pm})) = \frac{1}{\sqrt{n-h}}\sum_{i=1}^{n-h} \mathrm{IF}(X_i \pm X_{i+h}, Q, \Phi_{\sigma,\pm}) + o_P(1).$$

Agora, vamos aplicar o delta método, Teorema B.1 para a função $\phi(x) = x^2, \theta = Q(\Phi_{\sigma,\pm})$ e $T_n = Q^2_{n-h}(X_{1:n-h} \pm X_{1+h:n}, \Phi)$, donde vem:

$$\sqrt{n-h}(Q^2_{n-h}(X_{1:n-h} \pm X_{1+h:n}, \Phi) - Q^2(\Phi_{\sigma,\pm})) = \frac{2Q(\Phi_{\sigma,\pm})}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i \pm X_{i+h}, Q, \Phi_{\sigma,\pm}) + o_P(1).$$

Subtraindo as duas equações, resultamos em:

$$\sqrt{n-h}(Q^2_{n-h}(X_{1:n-h} + X_{1+h:n}, \Phi) - Q^2(\Phi_{\sigma,+})) - \sqrt{n-h}(Q^2_{n-h}(X_{1:n-h} - X_{1+h:n}, \Phi) + Q^2(\Phi_{\sigma,-})) =$$
$$\frac{2Q(\Phi_{\sigma,+})}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i + X_{i+h}, Q, \Phi_{\sigma,+}) + o_P(1) - \frac{2Q(\Phi_{\sigma,-})}{\sqrt{n-h}} \sum_{i=1}^{n-h} \mathrm{IF}(X_i - X_{i+h}, Q, \Phi_{\sigma,-}) - o_P(1)$$

logo

$$\sqrt{n-h}\Big[Q^2_{n-h}(X_{1:n-h} + X_{1+h:n}, \Phi) - Q^2_{n-h}(X_{1:n-h} - X_{1+h:n}, \Phi) + Q^2(\Phi_{\sigma,-}) - Q^2(\Phi_{\sigma,+})\Big] =$$
$$\frac{2}{\sqrt{n-h}}\Big[Q(\Phi_{\sigma,+}) \sum_{i=1}^{n-h} \mathrm{IF}(X_i + X_{i+h}, Q, \Phi_{\sigma,+}) - Q(\Phi_{\sigma,-}) \sum_{i=1}^{n-h} \mathrm{IF}(X_i - X_{i+h}, Q, \Phi_{\sigma,-})\Big] + o_P(1).$$

Substituindo a expressão de $\hat{\gamma}_Q$, obtemos:

$$\sqrt{n-h}\Big[4\hat{\gamma}_Q(h, X_{1:n}, \Phi) - \Big(Q^2(\Phi_{\sigma,+}) - Q^2(\Phi_{\sigma,-})\Big)\Big] = \frac{2}{\sqrt{n-h}}\Big[Q(\Phi_{\sigma,+}) \sum_{i=1}^{n-h} \mathrm{IF}(X_i + X_{i+h}, Q, \Phi_{\sigma,+})$$
$$- Q(\Phi_{\sigma,-}) \sum_{i=1}^{n-h} \mathrm{IF}(X_i - X_{i+h}, Q, \Phi_{\sigma,-})\Big] + o_P(1).$$

Pondo

$$\psi(x, y) = \frac{1}{2}\Big\{Q(\Phi_{\sigma,+})\,\mathrm{IF}(x + y, Q, \Phi_{\sigma,+}) - Q(\Phi_{\sigma,-})\,\mathrm{IF}(x - y, Q, \Phi_{\sigma,-})\}$$

fica:

$$\sqrt{n-h}\Big[\hat{\gamma}_Q(h, X_{1:n}, \Phi) - \frac{1}{4}\Big(Q^2(\Phi_{\sigma,+}) - Q^2(\Phi_{\sigma,-})\Big)\Big] = \frac{1}{\sqrt{n-h}} \sum_{i=1}^{n-h} \psi(X_i, X_{i+h}) + o_P(1). \quad (1.35)$$

Agora falta chegar na expressão (1.34). Usando (1.10):

$$\mathrm{IF}(x \pm y, Q, \Phi_{\sigma,\pm}) = \sqrt{2\gamma(0) \pm 2\gamma(h)}\,\mathrm{IF}(\frac{x \pm y}{\sqrt{2\gamma(0) \pm 2\gamma(h)}}, Q, \Phi)$$

pois

$$\mathrm{Var}(X_i \pm X_{i+h}) = \mathbb{E}(X_i^2 \pm 2X_iX_{i+h} + X_{i+h}^2) = 2\,\mathbb{E}\,X_i^2 \pm 2\,\mathbb{E}\,X_iX_{i+h} = 2\gamma(0) \pm 2\gamma(h).$$

Como $Q(\Phi_{\sigma,\pm}) = \sqrt{\mathrm{Var}(X_i \pm X_{i+h})}$, temos:

$$\psi(x,y) = \frac{1}{2}\Big\{\sqrt{2\gamma(0)+2\gamma(h)}\sqrt{2\gamma(0)+2\gamma(h)}\,\mathrm{IF}\left(\frac{x+y}{\sqrt{2\gamma(0)+2\gamma(h)}}, Q, \Phi\right) -$$
$$- \sqrt{2\gamma(0)-2\gamma(h)}\sqrt{2\gamma(0)-2\gamma(h)}\,\mathrm{IF}\left(\frac{x-y}{\sqrt{2\gamma(0)-2\gamma(h)}}, Q, \Phi\right)\Big\}$$
$$\psi(x,y) = (\gamma(0)+\gamma(h))\,\mathrm{IF}\left(\frac{x+y}{\sqrt{2\gamma(0)+2\gamma(h)}}, Q, \Phi\right) - (\gamma(0)-\gamma(h))\,\mathrm{IF}\left(\frac{x-y}{\sqrt{2\gamma(0)-2\gamma(h)}}, Q, \Phi\right),$$

como queríamos.

Já sabemos que $\frac{1}{4}\Big(Q^2(\Phi_{\sigma,+}) - Q^2(\Phi_{\sigma,-})\Big) = \mathbb{E}(X_1X_{1+h}) = \gamma(h)$ pela equação (1.32), logo a equação (1.35) fica:

$$\sqrt{n-h}\Big[\hat{\gamma}_Q(h, X_{1:n}, \Phi) - \gamma(h)\Big] = \frac{1}{\sqrt{n-h}}\sum_{i=1}^{n-h}\psi(X_i, X_{i+h}) + o_P(1).$$

Agora, resta mostrar um teorema central do limite para $\frac{1}{\sqrt{n-h}}\sum_{i=1}^{n-h}\psi(X_i, X_{i+h})$. Antes disso, precisamos da definição de posto Hermitiano de uma função mensurável f. Para mais detalhes, consultar Arcones [1].

Definição 1.2.3. *Seja $f\colon \mathbb{R}^d \to \mathbb{R}$ função mensurável e X um vetor gaussiano. Então se f possui segundo momento finito, o posto Hermitiano de f com respeito a X é*

$$\mathrm{rank}(f) := \inf\{\tau \mid \exists \text{ polinômio } P \text{ de grau } \tau \text{ com } \mathbb{E}[(f(X) - \mathbb{E}\,f(X))P(X)] \neq 0]\}.$$

Pelo Lema 1.2.5, usando $\psi = f$, temos:

$$\mathbb{E}[(f(X,Y) - \mathbb{E}\,f(X,Y))P(X,Y)] = \mathbb{E}[\psi(X,Y)XY] \neq 0.$$

Tomando $X = Y$, vemos que $\tau = 2$. Agora, note que (A1) implica $\lim_{h\to\infty}|\gamma(h)| = 0$ ou seja, existe h_0 tal que $|\gamma(h)| < 1$ para $h \geq h_0$. Portanto, para $\varepsilon > 0$ arbitrário,

$$\begin{aligned}\sum_{h\geq 1}|\gamma(h)|^2 &= \sum_{h\geq 1}^{h_0}|\gamma(h)|^2 + \sum_{h\geq h_0+1}|\gamma(h)|^2 \\ &\leq \sum_{h\geq 1}^{h_0}|\gamma(h)|^2 + \sum_{h\geq h_0+1}|\gamma(h)| \leq \varepsilon.\end{aligned}$$

Ademais, a estacionaridade da sequência $(X_i)_i$ torna a função de covariância uma função par:

$$\gamma(-h) = \mathbb{E}(X_tX_{t-h}) = \mathbb{E}(X_tX_{t+h}) = \gamma(h).$$

Portanto, temos a convergência de $\sum_{h\geq 1}|\gamma(h)|^2$ também para os índices negativos. Concluímos finalmente que

$$\sum_{h=-\infty}^{+\infty}|\gamma(h)|^2 < \infty.$$

Esta é exatamente a condição necessária para obter o resultado do teorema 4 em Arcones [1]:

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(f(X_j)-\mathbb{E}\,f(X_j)) \to_D N(0,\sigma^2), \tag{1.36}$$

onde

$$\sigma^2 \coloneqq \mathbb{E}[(f(X_1)-\mathbb{E}\,f(X_1))^2] + 2\sum_{k=1}^{\infty}\mathbb{E}[\Big(f(X_1)-\mathbb{E}\,f(X_1))(f(X_{1+k})-\mathbb{E}\,f(X_{1+k}))\Big].$$

Substituindo f por ψ e usando o fato de $\mathbb{E}\,\psi(X,Y)=0$, a expressão acima fica:

$$\sigma^2 \coloneqq \mathbb{E}\,\psi^2(X_1,X_{1+h}) + 2\sum_{k\geq 1}\mathbb{E}\left[\psi(X_1,X_{1+h})\psi(X_{1+k},X_{1+k+h})\right]. \tag{1.37}$$

Acabamos de mostrar que

$$\frac{1}{\sqrt{n-h}}\sum_{i=1}^{n-h}\psi(X_i,X_{i+h}) \to_D \mathcal{N}(0,\sigma^2) \tag{1.38}$$

sendo σ^2 dado em (1.37). Isto implica

$$\sqrt{n-h}\,(\hat{\gamma}_Q(h,X_{1:n},\Phi)-\gamma(h)) \to_D \mathcal{N}(0,\sigma^2), \tag{1.39}$$

finalizando a demonstração do Teorema 1.2.1. □

1.3 Conclusão

O estimador Q_n aqui apresentado (proposto por Rousseeuw [13]) possui boa resistência a presença de *outliers*, como foi analisado por meio de sua função de influência e o *breakdown point*. Devido a estas boas propriedades, o estimador $\hat{\gamma}_Q$ para função de covariância, foi construído a partir de Q_n. O Lema 1.2.2 mostrou a expansão assintótica de Q_n que é usada para provar um teorema central do limite para esse estimador, consulte Teorema 1 em Lévy-Leduc et al. [11]. O Teorema 1.2.1 é um teorema central do limite para $\hat{\gamma}_Q$. Ambos os estimadores quando aplicados em processos Gaussianos estacionários são assintoticamente normais, e portanto consistentes, com uma taxa de convergência $\sqrt{n}$. Por esses motivos, $\hat{\gamma}_Q$ e Q_n são, respectivamente, boas alternativas aos estimadores clássico para função de covariância e aos de escala.

Capítulo 2

Processos subgaussianos e Aprendizagem estatística

Neste capítulo estudaremos uma classe especial de processos estocásticos e falaremos um pouco sobre como esses podem ser usados em diversas aplicações no mundo científico.

2.1 Variáveis Aleatórias Subgaussianas e a Integral de Dudley

O estudo das chamadas *desigualdades de concentração* busca compreender o quão próximas, ou o quão distantes, as variáveis aleatórias estão de certos valores. Podemos nos perguntar, por exemplo, se uma variável aleatória X está concentrada em torno de sua média μ. Um resultado clássico nesse sentido é a Lei dos Grandes Números: sob determinadas hipóteses, a média aritmética de variáveis aleatórias independentes se concentra em torno do valor esperado.

Esse é um comportamento do tipo assintótico: sabemos que para uma quantidade suficientemente grande, a soma de variáveis aleatórias apresenta uma certa característica. Um questionamento razoável, portanto, é pensar acerca de um número fixo de variáveis. Em algumas aplicações que daremos adiante, esse número é na verdade o tamanho de uma amostra. Nesse contexto, queremos encontrar cotas superiores ou inferiores para a probabilidade da cauda:

$$\mathbb{P}(|X - \mu| > t) \leq \text{ alguma coisa pequena.}$$

Sobre esse comportamento, conseguimos subdividir as variáveis aleatórias em duas grandes classes: *subgaussianas* e *subexponenciais*. Em cada uma delas, estamos comparando o decaimento da variável com o decaimento gaussiano e o decaimento exponencial, respectivamente. Trataremos aqui somente das subgaussianas. Os resultados, exemplos e definições em sua maioria foram retirados do livro Vershyin [16].

2.1.1 Variáveis Aleatórias Subgaussianas

É bem sabido a importância da distribuição normal nos estudos de estatística e probabilidade, portanto, nada mais natural que estudar o comportamento da cauda gaussiana. Após alguns cálculos usando a densidade e a simetria desta variável, vemos que se $X \sim \mathcal{N}(0, 1)$, então

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-\frac{t^2}{2}) \text{ para todo } t \geq 0.$$

A variável subgaussiana é aquela que tem a probabilidade da cauda com decaimento super-exponencial semelhante ao gaussiano, motivando a seguinte definição:

Definição 2.1.1. *Uma variável aleatória X é subgaussiana se satisfaz a seguinte desigualdade de concentração para alguma constante $C > 0$:*

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-Ct^2) \text{ para todo } t \geq 0.$$

Como veremos mais adiante, esta ampla classe de variáveis é digna de especial atenção, já que ela contém, por exemplo, as gaussianas, Bernoulli e todas as distribuições limitadas. Em diversas aplicações, inclusive nas que serão estudadas neste texto, podemos usar as desigualdades de concentração subgaussiana para obter importantes resultados.

Outra maneira de estudar o decaimento da cauda de uma variável aleatória X é por meio de sua norma subgaussiana, denotada por $\|X\|_{\psi_2}$ e definida como:

$$\|X\|_{\psi_2} \coloneqq \inf_{r>0} \left\{ \mathbb{E}\exp(\frac{X^2}{r^2}) \leq 2 \right\}. \tag{2.1}$$

Antes de mostrar como a norma subgaussiana se relaciona com a Definição 2.1.1, vamos ver que (2.1) é de fato uma norma no espaço $L_{\psi_2} = L_{\psi_2}(\Omega, \Sigma, \mathbb{P})$ que consiste das variáveis aleatórias definidas no espaço de probabilidade $(\Omega, \Sigma, \mathbb{P})$ cuja norma subgaussiana é finita, isto é,

$$L_{\psi_2} = \{X \text{ é variável aleatória } ; \|X\|_{\psi_2} < \infty\}.$$

Proposição 2.1.1. *$\|\cdot\|_{\psi_2}$ é uma norma no espaço L_{ψ_2}.*

Demonstração. Primeiro, vejamos que (2.1) está bem definido. Tome $X \in L_{\psi_2}$ e $\lambda > 0$ tal que $\mathbb{E}\exp(\frac{X^2}{\lambda^2}) < \infty$. Como a função $\psi_2(x) \coloneqq \exp(x^2) - 1$ é crescente, a sequência $(\psi_2(\frac{X}{\lambda \cdot n}))_{n\in\mathbb{N}}$ é decrescente. A consequência disto é a integrabilidade de $\psi_2(\frac{X}{\lambda \cdot n})$ para todo $n \geq 1$, pois

$$\psi_2(\frac{X}{\lambda \cdot n}) \leq \psi_2(\frac{X}{\lambda}).$$

Aplicando o Teorema da Convergência Dominada,

$$\mathbb{E}\,\psi_2(\frac{X}{\lambda \cdot n}) \to \mathbb{E}\lim \psi_2(\frac{X}{\lambda \cdot n}) = 0,$$

pois ψ_2 é contínua em zero e $\psi_2(0) = 0$. Portanto, existe $n_0 \in \mathbb{N}$ tal que para todo $n \geq n_0$ vale

$$\mathbb{E}\,\psi_2(\frac{X}{\lambda \cdot n}) \leq 1.$$

Para que $\|\cdot\|_{\psi_2}$ seja norma, deve cumprir as seguintes propriedades para X e Y em L_{ψ_2}:

(N1) $\|X\|_{\psi_2} \geq 0$ e $\|X\|_{\psi_2} = 0$ se, e somente se $X \equiv 0$;

(N2) $\|\alpha X\|_{\psi_2} = |\alpha|\|X\|_{\psi_2}$ para todo $\alpha \in \mathbb{R}$;

(N3) $\|X + Y\|_{\psi_2} \leq \|X\|_{\psi_2} + \|Y\|_{\psi_2}$

A norma $\|\cdot\|_{\psi_2}$ é um número não-negativo direto pela definição. Se considerarmos $X \equiv 0$, então $\psi_2(X) = 0$ e portanto $\mathbb{E}\,\psi_2(X) = 0$, resultando em $\|X\|_{\psi_2} = 0$. Se por outro lado, $\|X\|_{\psi_2} = 0$

então $\mathbb{P}(|X| > \varepsilon) = 0$ para todo $\varepsilon > 0$ implicando $X = 0$ em quase todo ponto. De fato, a desigualdade de Markov fornece para todo $\varepsilon > 0$:

$$\mathbb{P}(|X| > \varepsilon) = \mathbb{P}\left(\frac{|X|}{\lambda} > \frac{\varepsilon}{\lambda}\right) = \mathbb{P}\left(\psi_2(\frac{|X|}{\lambda}) > \psi_2(\frac{\varepsilon}{\lambda})\right) \leq \frac{\mathbb{E}\,\psi_2(\frac{|X|}{\lambda})}{\psi_2(\frac{\varepsilon}{\lambda})}.$$

Fazendo $\lambda \to \|X\|_{\psi_2}$, obtemos

$$\mathbb{P}(|X| > \varepsilon) \leq \frac{1}{\psi_2(\frac{\varepsilon}{\|X\|_{\psi_2}})},$$

finalizando a demonstração do item (N1). O próximo item é facilmente obtido pelas propriedades de ínfimo:

$$\begin{aligned}
\|\alpha X\|_{\psi_2} &= \inf_{r>0}\left\{\,\mathbb{E}\exp(\frac{\alpha^2 X^2}{r^2}) \leq 2\,\right\} \\
&= \inf_{r>0}\left\{\,\mathbb{E}\exp(\frac{X^2}{\frac{r^2}{\alpha^2}}) \leq 2\,\right\} \\
&= \inf_{r>0}\left\{\,\mathbb{E}\exp(\frac{X^2}{u^2}) \leq 2\,\right\}, \quad (u = \frac{r^2}{\alpha^2}) \\
&= \inf_{|\alpha|\cdot u>0}\left\{\,\mathbb{E}\exp(\frac{X^2}{u^2}) \leq 2\,\right\} \\
&= |\alpha| \inf_{u>0}\left\{\,\mathbb{E}\exp(\frac{X^2}{u^2}) \leq 2\,\right\} = |\alpha|\|X\|_{\psi_2}.
\end{aligned}$$

Resta a desigualdade triangular. Sejam $\lambda_X, \lambda_Y > 0$ tais que $\mathbb{E}\,\psi_2(\frac{X}{\lambda_X}) \leq 1$ e $\mathbb{E}\,\psi_2(\frac{Y}{\lambda_Y}) \leq 1$. Temos:

$$\begin{aligned}
\mathbb{E}\,\psi_2\left(\frac{|X+Y|}{\lambda_X+\lambda_Y}\right) &\leq \mathbb{E}\,\psi_2\left(\frac{|X|+|Y|}{\lambda_X+\lambda_Y}\right) = \mathbb{E}\,\psi_2\left(\frac{\lambda_X}{\lambda_X+\lambda_Y}\cdot\frac{|X|}{\lambda_X} + \frac{\lambda_Y}{\lambda_X+\lambda_Y}\cdot\frac{|Y|}{\lambda_X}\right) \\
&\leq \mathbb{E}\left(\frac{\lambda_X}{\lambda_X+\lambda_Y}\cdot\psi_2(\frac{|X|}{\lambda_X}) + \frac{\lambda_Y}{\lambda_X+\lambda_Y}\cdot\psi_2(\frac{|Y|}{\lambda_Y})\right), \quad \text{pela convexidade de } \psi_2 \\
&\leq \frac{\lambda_X}{\lambda_X+\lambda_Y}\,\mathbb{E}\,\psi_2(\frac{|X|}{\lambda_X}) + \frac{\lambda_Y}{\lambda_X+\lambda_Y}\,\mathbb{E}\,\psi_2(\frac{|Y|}{\lambda_Y}) \\
&\leq \frac{\lambda_X}{\lambda_X+\lambda_Y} + \frac{\lambda_Y}{\lambda_X+\lambda_Y} = 1.
\end{aligned}$$

Isto mostra que $|X+Y| \in L_{\psi_2}$ e que $\|X+Y\|_{\psi_2} \leq \lambda_X + \lambda_Y$. Tomando o ínfimo em λ_X e λ_Y, obtemos a subaditividade da norma $\|\cdot\|_{\psi_2}$ finalizando a proposição. □

Podemos obter algumas cotas para os momentos e para a função geradora de momentos ao examinar o comportamento da cauda, resultando na seguinte proposição:

Proposição 2.1.2. *Seja X variável aleatória. Existem constantes absolutas*[1] *e positivas C_1, C_2 e C_3 tais que as seguintes afirmações são equivalentes:*

a) $\|X\|_{\psi_2} \leq C_1$.

b) A cauda de X satisfaz

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-C_2 t^2) \text{ para todo } t \geq 0. \tag{S1}$$

[1] Constantes absolutas são constantes numéricas independentes das hipóteses em questão.

c) Os momentos de X satisfazem

$$\|X\|_{L^p} = (\mathbb{E}\,|X|^p)^{\frac{1}{p}} \leq C_3\sqrt{p} \text{ para todo } p \geq 1. \tag{S2}$$

Demonstração. a) $\Rightarrow$ b) Seja $t \geq 0$ fixo. Considerando o item a) e a Desigualdade de Markov, temos:

$$\begin{aligned}\mathbb{P}\{|X| \geq t\} = \mathbb{P}\{\exp\left(\frac{|X|^2}{C_1^2}\right) \geq \exp\left(\frac{t^2}{C_1^2}\right)\} &\leq \exp\left(-\frac{t^2}{C_1^2}\right)\mathbb{E}\exp\left(\frac{|X|^2}{C_1^2}\right)\\ &\leq 2\exp\left(-\frac{t^2}{C_1^2}\right).\end{aligned}$$

b) $\Rightarrow$ c) Considerando a desigualdade[2] $(\frac{p}{2})! \leq (\frac{p}{2})^{\frac{p}{2}}$ e a definição da função gama, $\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)dt$ temos:

$$\begin{aligned}\|X\|_{L^p}^p = \mathbb{E}\,|X|^p = \int_0^\infty \mathbb{P}\{|X|^p \geq t\}dt &= \int_0^\infty \mathbb{P}\{|X| \geq u\}pu^{p-1}du, \quad (t^{\frac{1}{p}} = u)\\ &\leq \int_0^\infty 2\exp\left(-C_2u^2\right)pu^{p-1}du\\ &= 2p\int_0^\infty \exp\left(-C_2u^2\right)u^{p-1}du.\end{aligned}$$

Nesse ponto usaremos a substituição $u = t^{\frac{1}{2}}C_2$:

$$\begin{aligned}&= 2p\int_0^\infty \exp\left(-t\right)(t^{\frac{1}{2}}C_2)^{p-1}\frac{1}{2}t^{-\frac{1}{2}}C_2dt\\ &= 2pC_2C_2^{p-1}\frac{1}{2}\int_0^\infty \exp\left(-t\right)(t^{\frac{1}{2}})^{p-1}t^{-\frac{1}{2}}dt\\ &= pC_2^p\int_0^\infty \exp\left(-t\right)t^{\frac{p}{2}-1}dt = pC_2^p\Gamma(\frac{p}{2})\\ &= 2C_2^p\frac{p}{2}\Gamma(\frac{p}{2}) = 2C_2^p(\frac{p}{2})!\\ &\leq 2C_2^p(\frac{p}{2})^{\frac{p}{2}} \leq 2C_2^pp^{\frac{p}{2}}.\end{aligned}$$

Tirando a raiz p-ésima, obtemos:

$$\|X\|_{L^p} \leq 2^{\frac{1}{p}}C_2\sqrt{p}.$$

c) $\Rightarrow$ a) Para que $\|X\|\|_{\psi_2} < \infty$, basta mostrar a existência de algum $\lambda > 0$ que torne $\mathbb{E}\exp\left(\frac{X^2}{\lambda^2}\right) \leq 2$. Considerando a expansão em série de Taylor da função exponencial e o Teorema da Convergência Monótona, temos:

$$\mathbb{E}\exp\left(\frac{X^2}{\lambda^2}\right) = \mathbb{E}\left[\sum_{p=0}^\infty \frac{X^{2p}}{\lambda^{2p}}\frac{1}{p!}\right] = \sum_{p=0}^\infty \frac{\mathbb{E}[X^{2p}]}{\lambda^{2p}}\frac{1}{p!}.$$

[2] Consulte a Proposição D.3 no Apêndice.

Por hipótese, $\mathbb{E}[X^{2p}] \leq C_3^{2p}(2p)^{\frac{2p}{2}} = C_3^{2p}(2p)^p$, resultando em:

$$\begin{aligned}
\mathbb{E}\exp\left(\frac{X^2}{\lambda^2}\right) &\leq \sum_{p=0}^{\infty} \frac{C_3^{2p}(2p)^p}{\lambda^{2p}} \frac{1}{p!} \\
&= \sum_{p=0}^{\infty} \left(\frac{C_3(\sqrt{2p})}{\lambda}\right)^{2p} \frac{1}{p!} \\
&\leq \sum_{p=0}^{\infty} \left(\frac{C_3(\sqrt{2p})}{\lambda}\right)^{2p} \frac{e^p}{p^p} \\
&= \sum_{p=0}^{\infty} \left(\frac{C_3(\sqrt{2p})(\sqrt{e})}{\lambda\sqrt{p}}\right)^{2p} \\
&= \sum_{p=0}^{\infty} \left(\frac{C_3(\sqrt{2e})}{\lambda}\right)^{2p} = \frac{1}{1-\left(\frac{C_3(\sqrt{2e})}{\lambda}\right)^2}.
\end{aligned}$$

Usamos que $\frac{1}{p!} \leq \frac{e^p}{p^p}$[3] e na última igualdade estamos considerando $\left|\left(\frac{C_3(\sqrt{2e})}{\lambda}\right)^2\right| \leq 1$. Basta tomar $\lambda > 0$ que torna a série obtida acima menor do que ou igual a 2. □

Quando a variável X está centrada, isto é, quando $\mathbb{E}\,X = 0$, há outra equivalência adicional para a subgaussianidade, resultando em nova proposição:

Proposição 2.1.3. *Se $\mathbb{E}\,X = 0$, então o item seguinte é equivalente aos itens da Proposição 2.1.2.*

d) Existe constante positiva C_4 tal que a função geradora de momentos de X satisfaz

$$\mathbb{E}\exp(\lambda X) \leq \exp(C_4\lambda^2) \text{ para todo } \lambda \in \mathbb{R}. \tag{S3}$$

Demonstração. *d)* $\Rightarrow$ *b)* Pela desigualdade de Markov, segue

$$\begin{aligned}
\mathbb{P}\{X \geq t\} = \mathbb{P}\{\exp(\lambda X) \geq \exp(\lambda t)\} &\leq \exp(-\lambda t)\,\mathbb{E}\,(\exp(\lambda X)) \\
&\leq \exp(-\lambda t)\exp(C_4\lambda^2) \\
&= \exp\left(-\lambda t + C_4\lambda^2\right).
\end{aligned}$$

Otimizando em λ e portanto escolhendo $\lambda = \frac{t}{2C_4}$, obtemos

$$\mathbb{P}\{X \geq t\} \leq \exp\left(-\frac{t^2}{2C_4} + C_4\frac{t^2}{2^2C_4^2}\right) = \exp\left(-\frac{t^2}{4C_4}\right).$$

Com raciocínio análogo obtemos a mesma cota para $\mathbb{P}\{-X \geq t\}$, resultando em

$$\mathbb{P}\{|X| \geq t\} \leq 2\exp\left(-\frac{t^2}{4C_4}\right).$$

c) $\Rightarrow$ *d)* Cálculos similares aos feitos na implicação *c)* $\Rightarrow$ *a)* da proposição anterior e a desigualdade[4] $\frac{1}{1-x} \leq \exp(2x)$, válida para $x \in [0, \frac{1}{2}]$ demonstram que o item *c)* também resulta em:

[3] Consulte a Proposição D.4 no Apêndice.
[4] Consulte a Proposição D.2 no Apêndice.

$\mathbb{E}\exp(\lambda^2 X^2) \leq \exp(\lambda^2)$, se $|\lambda| \leq 1$. Para maiores detalhes, consulte a página 26 em Vershyin [16]. Além disso, considere a desigualdade[5] válida para todo $x \in \mathbb{R}$: $\exp x \leq x + \exp x^2$. Como $\mathbb{E}\, X = 0$, vale

$$\mathbb{E}\exp(\lambda X) \leq \mathbb{E}\,\lambda X + \exp(\lambda^2 X^2) = \mathbb{E}\exp(\lambda^2 X^2) \leq \exp(\lambda^2), \quad \text{se } |\lambda| \leq 1.$$

Para outros valores de λ, considere nova desigualdade[6] numérica $2\lambda x \leq \lambda^2 + x^2$, que é válida para todo x e para todo λ. Segue:

$$\mathbb{E}\exp(\lambda X) \leq \exp(\frac{\lambda^2}{2})\,\mathbb{E}\exp(\frac{X^2}{2}) \leq \exp(\frac{\lambda^2}{2}) \cdot \exp\frac{1}{2} \leq \exp(\lambda^2), \quad \text{já que } |\lambda| \geq 1.$$

□

Podemos enunciar as duas últimas proposições em termos da norma subgaussiana, uma vez que, a menos de fatores constantes, o número $\|X\|_{\psi_2}$ é o ínfimo de todas as constantes positivas que torna cada uma das desigualdades válidas. Resumimos esse fato no seguinte corolário.

Corolário 2.1.1. *Toda variável aleatória X subgaussiana satisfaz as seguintes desigualdades com constantes absolutas $C, c > 0$:*

a) $\mathbb{E}\exp(X^2/\|X\|_{\psi_2}^2) \leq 2$;

b) $\mathbb{P}(|X| \geq t) \leq 2\exp(-ct^2/\|X\|_{\psi_2})$ *para todo* $t \geq 0$;

c) $\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p}$ *para todo* $p \geq 1$;

d) Se $\mathbb{E}\, X = 0$, *então* $\mathbb{E}\exp(\lambda X) \leq \exp(C\lambda^2\|X\|_{\psi_2}^2)$ *para todo* $\lambda \in \mathbb{R}$.

Estamos aptos a entender alguns exemplos.

Exemplo 2.1.1. *(Gaussiana) Considere a variável aleatória com distribuição gaussiana padrão $X \sim \mathcal{N}(0, \sigma^2)$. Então*

$$\|X\|_{\psi_2} \leq C \cdot \sigma.$$

De fato, considere um número real $t > 0$:

$$\begin{aligned}
\mathbb{E}\exp\left(\frac{X^2}{t^2}\right) &= \int_{-\infty}^{+\infty} \exp\left(\frac{x^2}{t^2}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp(-x^2(\frac{1}{2\sigma^2} - \frac{1}{t^2}))dx = \\
&= \frac{1}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(-\frac{x^2}{2u^2})dx \\
&= \frac{u}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(-\frac{z^2}{2})dz \\
&= \frac{u}{\sigma},
\end{aligned}$$

onde usamos a substituição $u = \frac{1}{\sqrt{2}\sqrt{\frac{1}{2\sigma^2} - \frac{1}{t^2}}}$. *Para que* $\frac{u}{\sigma} \leq 2$, *basta tomar*

$$\frac{2\sqrt{2}}{\sqrt{3}} \cdot \sigma \leq t.$$

[5] Consulte a Proposição D.1 no Apêndice.
[6] Consulte a Proposição D.5 no Apêndice.

Exemplo 2.1.2. *(Bernoulli simétrica) Seja X a variável aleatória com distribuição de Bernoulli simétrica, ou distribuição de Rademacher, isto é, $\mathbb{P}\{X=1\} = \mathbb{P}\{X=-1\} = \frac{1}{2}$. Então*

$$\|X\|_{\psi_2} = \frac{1}{\sqrt{\log 2}}.$$

De fato, considere um número real $t > 0$:

$$\mathbb{E}\exp\left(\frac{X^2}{t^2}\right) = \mathbb{E}\sum_{k=0}^{\infty}\frac{X^{2k}}{t^{2k}\cdot k!} = \mathbb{E}\sum_{k=0}^{\infty}\frac{|X|^{2k}}{t^{2k}\cdot k!} = \mathbb{E}\sum_{k=0}^{\infty}\frac{1^k}{t^{2k}\cdot k!} = \exp\left(\frac{1}{t^2}\right).$$

Para que $\exp\left(\frac{1}{t^2}\right) \leq 2$, *basta tomar*

$$\frac{1}{\sqrt{\log 2}} \leq t.$$

Exemplo 2.1.3. *(Limitada) Qualquer variável aleatória X limitada é subgaussiana com*

$$\|X\|_{\psi_2} \leq \frac{1}{\sqrt{\log 2}}\|X\|_{\infty}.$$

Podemos usar o mesmo raciocínio do exemplo anterior, considerando um número real $t > 0$ e $M > 0$ tal que $|X| \leq M$ em quase todo ponto:

$$\mathbb{E}\exp\left(\frac{X^2}{t^2}\right) = \mathbb{E}\sum_{k=0}^{\infty}\frac{X^{2k}}{t^{2k}\cdot k!} = \mathbb{E}\sum_{k=0}^{\infty}\frac{|X|^{2k}}{t^{2k}\cdot k!} \leq \mathbb{E}\sum_{k=0}^{\infty}\frac{M^{2k}}{t^{2k}\cdot k!} = \exp\left(\frac{M^2}{t^2}\right).$$

Para que $\exp\left(\frac{M^2}{t^2}\right) \leq 2$, *basta tomar*

$$\frac{M}{\sqrt{\log 2}} \leq t.$$

Dada uma sequência de variáveis subgaussianas não necessariamente independentes, é possível cotar o valor esperado do máximo dessas variáveis em termos da norma subgaussiana.

Lema 2.1.1. *(Valor esperado do máximo de variáveis subgaussianas)*
Seja $(X_i)_{i\in\mathbb{N}}$ sequência de variáveis subgaussianas não necessariamente independentes. Então existe constante $K > 0$ tal que:

$$\mathbb{E}\max_{i\leq N}|X_i| \leq K\max_{i\in\mathbb{N}}\|X_i\|_{\psi_2}\sqrt{\log N}.$$

Demonstração. Da Desigualdade de Jensen, temos para $t > 0$ e $N \geq 2$ fixo:

$$\begin{aligned}\exp t\,\mathbb{E}\max_{i\leq N}|X_i| &\leq \mathbb{E}\exp t\max_{i\leq N}|X_i| \\ &= \mathbb{E}\max_{i\leq N}\exp t|X_i| \leq \mathbb{E}\sum_{i=1}^{N}\exp t|X_i|.\end{aligned}$$

Para cada $i \in \mathbb{N}$ a subgaussianidade das variáveis garante a existência de constante $C_i > 0$ tal que $\mathbb{E}\exp t|X_i| \leq 2\exp C_i\cdot t^2\|X_i\|_{\psi_2}^2$, resultando em

$$\exp t\,\mathbb{E}\max_{i\leq N}|X_i| \leq N\cdot 2\exp(C\cdot t^2\max_{i\leq N}\|X_i\|_{\psi_2}^2).$$

Aplicando log em ambos os lados da desigualdade, obtemos:

$$t\,\mathbb{E}\max_{i\leq N}|X_i| \leq \log 2N + C\cdot t^2(\max_{i\leq N}\|X_i\|_{\psi_2})^2.$$

E isto implica

$$\mathbb{E}\max_{i\leq N}|X_i| \leq \frac{\log 2N}{t} + C\cdot t(\max_{i\leq N}\|X_i\|_{\psi_2})^2 \leq \frac{\log 2N}{t} + C\cdot t(\max_{i\in\mathbb{N}}\|X_i\|_{\psi_2})^2.$$

Para $t>0$, a função $t\mapsto \frac{A}{t}+Bt$ sendo $A,B>0$ atinge seu mínimo em $t=\sqrt{\frac{A}{B}}$. Portanto, escolhendo $t=\frac{\sqrt{\log 2N}}{\sqrt{C}\cdot\max_{i\in\mathbb{N}}\|X_i\|_{\psi_2}}$ ficamos com:

$$\begin{aligned}\mathbb{E}\max_{i\leq N}|X_i| &\leq \log 2N\cdot\frac{\sqrt{C}\cdot\max_{i\in\mathbb{N}}\|X_i\|_{\psi_2}}{\sqrt{\log 2N}} + C\cdot(\max_{i\in\mathbb{N}}\|X_i\|_{\psi_2})^2\cdot\frac{\sqrt{\log 2N}}{\sqrt{C}\cdot\max_{i\in\mathbb{N}}\|X_i\|_{\psi_2}}\\ &= \sqrt{\log 2N}\cdot\sqrt{C}\cdot\max_{i\in\mathbb{N}}\|X_i\|_{\psi_2} + \sqrt{C}\cdot\max_{i\in\mathbb{N}}\|X_i\|_{\psi_2}\cdot\sqrt{\log 2N}\\ &= K\max_{i\in\mathbb{N}}\|X_i\|_{\psi_2}\sqrt{\log N}.\end{aligned}$$

□

2.2 Supremo de Processos subgaussianos

Um processo estocástico $(X_t)_{t\in T}$ é uma coleção de variáveis aleatórias definidas no mesmo espaço de probabilidade e indexadas por $t\in T$, sendo T um conjunto arbitrário. Cada processo pode ser associado a um espaço pseudométrico a partir de seus *incrementos* $X_t - X_s$, com $s,t\in T$. Podemos definir, por exemplo, quando X_t e X_s são quadrado integráveis,

$$d(t,s) \coloneqq \|X_s - X_t\|_{L^2} = (\mathbb{E}(X_s-X_t)^2)^{\frac{1}{2}}, \quad s,t\in T.$$

Como a distância entre dois elementos $s,t\in T$ distintos pode ser nula, a definição acima torna (T,d) um espaço pseudométrico.

Muitos problemas em diversas áreas do conhecimento são estudados a partir de variáveis aleatórias indexadas num conjunto arbitrário T. Em alguns casos é possível inferir algumas propriedades do processo a partir das propriedades de T, e vice-versa. Uma pergunta bastante razoável nesse contexto é como mensurar

$$\mathbb{E}\sup_{t\in T} X_t.$$

Definição 2.2.1. *Dado um espaço métrico (T,d) e um subconjunto $K\subset T$, uma $\varepsilon-$rede de K é um conjunto $\{x_1,\ldots,x_n\}=N\subset K$ tal que para cada $y\in K$, existe $x_i\in N$ tal que $d(x_i,y)\leq\varepsilon$. Em outras palavras, N é $\varepsilon-$rede de K quando $K\subset\cup_{x_i\in N}^n\overline{B}(x_i,\varepsilon)$. O número de cobertura $N(K,d,\varepsilon)$ é a menor cardinalidade de uma $\varepsilon-$rede de K, isto é,*

$$N(K,d,\varepsilon) = \inf\{\,\#S \mid S\subset K \text{ é } \varepsilon\text{-rede de } K\,\}.$$

Quando não é possível definir uma $\varepsilon-$ rede em T, considere $N(T,d,\varepsilon)=\infty$. Estamos interessados nos espaços métricos em que $N(T,d,\varepsilon)<\infty$ para todo $\varepsilon>0$, os espaços *totalmente limitados*. Note que o número de cobertura é não-crescente com relação ao raio. Portanto,

$\varepsilon_1 \leq \varepsilon_2$ implica $N(T, d, \varepsilon_2) \leq N(T, d, \varepsilon_1)$. Tipicamente, o número de cobertura diverge quando $\varepsilon \to 0^+$. O número $\log N(T, d, \varepsilon)$ é conhecido como *entropia métrica* do conjunto T com respeito a d.

Como ressaltou Wainwright [17], apesar de entropia métrica ser um conceito determinístico puramente relacionado ao conjunto T, a partir dele podemos obter propriedades importantes de um processo estocástico indexado em T. A integral de Dudley é uma ilustração clássica de como a estrutura de T influencia o comportamento de um processo. Com ela obteremos uma cota superior do supremo de $(X_t)_{t\in T}$.

2.2.1 Processos subgaussianos

Almejando investigar o comportamento de processos estocásticos, definiremos agora um processo subgaussiano.

Definição 2.2.2. *Um processo estocástico $(X_t)_{t\in T}$ será chamado de subgaussiano com respeito a uma pseudométrica d em T quando existir constante $K \geq 0$ tal que*

$$\|X_s - X_t\|_{\psi_2} \leq K d(t, s) \quad \text{para todos } t, s \in T.$$

Por exemplo, se $(X_t)_{t\in T}$ é um processo gaussiano de média zero, então $(X_t)_{t\in T}$ possui incrementos subgaussianos. De fato, quando $\mathbb{E}\, X_t = 0$, vale:

$$\|X_t\|_{L^2} = (\mathbb{E}\,|X_t|^2)^{\frac{1}{2}} = (\mathbb{E}\, X_t^2)^{\frac{1}{2}} = (\mathbb{E}(X_t - \mathbb{E}\, X_t)^2)^{\frac{1}{2}} = \sqrt{\operatorname{Var} X}.$$

Além disso, sabemos que se $X_t \sim N(0, \sigma_t^2)$ então existe constante $C_t > 0$ tal que $\|X_t\|_{\psi_2} \leq C_t \sigma_t$. Por fim, se colocarmos em T a pseudométrica $d(t, s) \coloneqq \|X_s - X_t\|_{L^2}$, teremos

$$\|X_s - X_t\|_{\psi_2} \leq C_{s-t}\sigma_{s-t} = C_{s-t}\|X_s - X_t\|_{L^2} = C_{s-t} d(t, s).$$

2.2.2 Encadeamento e a Integral de Dudley

Voltemos ao problema do supremo de $(X_t)_{t\in T}$. Quando T tem cardinalidade finita, uma cota da união[7] é o suficiente. Por outro lado, se T não é finito, é preciso um método mais refinado, que é a conhecida técnica do *encadeamento*. A ideia é basicamente discretizar o conjunto T: a partir de um ponto inicial, vamos tomar uma cadeia de $\varepsilon-$redes cada vez mais refinadas e reduzir o problema de calcular o supremo a calcular o máximo num conjunto finito.

Seja $t \in T$. Considere T_ε uma ε-rede de T. Então existe $\pi(t) \in T_\varepsilon$ tal que $d(t, \pi(t)) \leq \varepsilon$. Podemos escrever:

$$\begin{aligned} X_t = X_t - X_{\pi(t)} + X_{\pi(t)} \implies & \sup_{t\in T} X_t \leq \sup_{t\in T}\left(X_t - X_{\pi(t)}\right) + \sup_{t\in T} X_{\pi(t)} \\ \implies & \mathbb{E}\sup_{t\in T} X_t \leq \mathbb{E}\sup_{t\in T}\left(X_t - X_{\pi(t)}\right) + \mathbb{E}\sup_{t\in T} X_{\pi(t)}. \end{aligned}$$

Para limitar a segunda parcela, basta usar uma cota da união sobre todos os elementos da rede. O problema se instala ao tentarmos cotar a primeira, e a solução dada pelo encadeamento é decompor o supremo numa soma finita de máximos sobre conjuntos sucessivamente refinados. Com esta técnica, obteremos a *Integral de entropia de Dudley*. Daqui em diante, considere T um espaço pseudométrico dotado da métrica d que torna o processo em questão subgaussiano. Considere $\operatorname{diam} T = \sup_{x,y\in T} d(x, y)$.

[7] Desigualdade de Boole.

Teorema 2.2.1. *(Integral de Dudley discreta) Sejam T um conjunto enumerável e $(X_t)_{t\in T}$ processo subgaussiano de média zero. Então existe constante $C > 0$ tal que*

$$\mathbb{E}\sup_{t\in T} X_t \leq CK\sum_{i=1}^{\infty}\frac{\operatorname{diam} T}{2^i}\sqrt{\log N(T,d,\frac{\operatorname{diam} T}{2^i})}, \tag{2.2}$$

sendo K a constante da Definição 2.2.2.

Demonstração. Para cada $j \in \mathbb{N}$, considere uma cadeia de ε_j-redes pondo $\varepsilon_j = \frac{\operatorname{diam}(T)}{2^j}$ e denote por $|T_j| = N(T,d,\varepsilon_j)$. Dado um ponto $t \in T$, defina $\pi_j(t) \in T_j$ tal que $d(t,\pi_j(t)) = d(t,T_j)$. Chame $t_0 \coloneqq \pi_0(t)$ e $T_0 = \{\, t_0 \,\}$ com $\varepsilon_0 = \operatorname{diam}(T)$. Note que $|T_j|$ é não-decrescente em j, pois uma ε_j-rede com raio menor contém mais elementos. Portanto, $|T_{j-1}| \leq |T_j|$ para todo $j \in \mathbb{N}$.

O refinamento dessa cadeia de redes tem como consequência fundamental a seguinte convergência: $X_{\pi_j(t)} \to X_t$ quase certamente. O primeiro passo para demonstrar esta convergência é notar que

$$\|X_{\pi_j(t)} - X_t\|_{\psi_2} \leq Kd(t,\pi_j(t)) \leq K\frac{\operatorname{diam}(T)}{2^j} \xrightarrow{j\to\infty} 0.$$

Além disso, sabemos que para todo $p \geq 1$ vale $\|X\|_{L^p} \leq C\sqrt{p}\|X\|_{\psi_2}$ pelo item $c)$ do Corolário 2.1.1. Portanto, a desigualdade acima implica $X_{\pi_j(t)} \to X_t$ em L^2, o que por sua vez implica $X_{\pi_j(t)} \to X_t$ em probabilidade. A convergência em probabilidade implica a convergência $X_{\pi_{j_k}(t)} \to X_t$ quase certamente para alguma subsequência $j_1 < j_2 < \ldots < j_k < \ldots$. Usando a soma telescópica $\sum_{i=1}^{j_k} X_{\pi_i(t)} - X_{\pi_{i-1}(t)} = X_{\pi_{j_k}(t)} - X_{\pi_0(t)} = X_{\pi_{j_k}(t)} - X_{t_0}$, podemos escrever

$$\begin{aligned}
X_t &= \lim_k X_{\pi_{j_k}(t)} + X_{t_0} - X_{t_0} = X_{t_0} + \lim_k X_{\pi_{j_k}(t)} - X_{\pi_0(t)} \\
&= X_{t_0} + \lim_k \sum_{i=1}^{j_k} X_{\pi_i(t)} - X_{\pi_{i-1}(t)} \\
&\leq X_{t_0} + \lim_k \sum_{i=1}^{j_k} |X_{\pi_i(t)} - X_{\pi_{i-1}(t)}| \\
&\leq X_{t_0} + \sum_{i=1}^{\infty} |X_{\pi_i(t)} - X_{\pi_{i-1}(t)}| \\
&\leq X_{t_0} + \sum_{i=1}^{\infty} \frac{|X_{\pi_i(t)} - X_{\pi_{i-1}(t)}|}{d(\pi_i(t),\pi_{i-1}(t))} d(\pi_i(t),\pi_{i-1}(t)).
\end{aligned} \tag{2.3}$$

Pela desigualdade triangular,

$$d(\pi_i(t),\pi_{i-1}(t)) \leq d(\pi_i(t),t) + d(t,\pi_{i-1}(t)) \leq \frac{\operatorname{diam}(T)}{2^i} + \frac{\operatorname{diam}(T)}{2^{i-1}} = \frac{3\operatorname{diam}(T)}{2^i}.$$

Como T é enumerável, podemos passar a desigualdade (2.3) ao supremo sobre os possíveis elementos $t_i \in T_i$ e $t_{i-1} \in T_{i-1}$:

$$\begin{aligned}
X_t &\leq X_{t_0} + \sum_{i=1}^{\infty} \sup_{\substack{t_i\in T_i \\ t_{i-1}\in T_{i-1}}} \frac{|X_{t_i} - X_{t_{i-1}}|}{d(t_i,t_{i-1})} d(\pi_i(t),\pi_{i-1}(t)) \\
&\leq X_{t_0} + \sum_{i=1}^{\infty} \sup_{\substack{t_i\in T_i \\ t_{i-1}\in T_{i-1}}} \frac{|X_{t_i} - X_{t_{i-1}}|}{d(t_i,t_{i-1})} \frac{3\operatorname{diam}(T)}{2^i}.
\end{aligned}$$

Defina a variável $S_i \coloneqq \frac{|X_{t_i} - X_{t_{i-1}}|}{d(t_i, t_{i-1})}$. O Lema 2.1.1 fornece

$$\mathbb{E} \max_{1 \le i \le N} |S_i| \le C \max_{1 \le i \le N} \|S_i\|_{\psi_2} \sqrt{\log N}.$$

Note que $\|S_i\|_{\psi_2} = \|\frac{|X_{t_i} - X_{t_{i-1}}|}{d(t_i, t_{i-1})}\|_{\psi_2} \le \frac{K d(t_i, t_{i-1})}{d(t_i, t_{i-1})} \le K$ e como $|T_{i-1}| \le |T_i|$, temos $|T_{i-1}| \cdot |T_i| \le |T_i|^2$ e ficamos com:

$$\mathbb{E} \max_{\substack{t_i \in T_i \\ t_{i-1} \in T_{i-1}}} |S_i| \le CK \sqrt{\log |T_i|^2} = CK \sqrt{2 \log N(T, d, \varepsilon_i)}.$$

Dessa forma, o resultado é a seguinte desigualdade válida em quase todo o ponto:

$$\sup_{t \in T} X_t \le X_{t_0} + \sum_{i=1}^{\infty} \sup_{\substack{t_i \in T_i \\ t_{i-1} \in T_{i-1}}} S_i \frac{3 \operatorname{diam}(T)}{2^i}.$$

Passando ao valor esperado e usando a hipótese de média zero, obtemos

$$\begin{aligned} \mathbb{E} \sup_{t \in T} X_t &\le \mathbb{E}\, X_{t_0} + \sum_{i=1}^{\infty} \mathbb{E} \sup_{\substack{t_i \in T_i \\ t_{i-1} \in T_{i-1}}} S_i \frac{3 \operatorname{diam}(T)}{2^i} \\ &\le \sum_{i=1}^{\infty} \frac{3 \operatorname{diam}(T)}{2^i} CK \sqrt{2 \log N(T, d, \varepsilon_i)}. \end{aligned}$$

Obtemos então a desigualdade de Dudley discreta:

$$\mathbb{E} \sup_{t \in T} X_t \le CK \sum_{i=1}^{\infty} \frac{\operatorname{diam}(T)}{2^i} \sqrt{\log N(T, d, \frac{\operatorname{diam}(T)}{2^i})}. \tag{2.4}$$

□

Provada a versão discreta da desigualdade de Dudley, resta um passo para obter o resultado original como veremos no próximo teorema.

Teorema 2.2.2. *(Integral de Dudley) Sob as mesmas hipóteses do Teorema 2.2.1, existe uma constante $C > 0$ tal que*

$$\mathbb{E} \sup_{t \in T} X_t \le C \int_0^{\infty} \sqrt{\log N(T, d, \varepsilon)} d\varepsilon. \tag{2.5}$$

Demonstração. Basta escrever a série em (2.2) como uma integral. Veja primeiramente que $\frac{\operatorname{diam} T}{2^i} = 2 \int_{\frac{\operatorname{diam} T}{2^{i+1}}}^{\frac{\operatorname{diam} T}{2^i}} d\varepsilon$ e então

$$\begin{aligned} \sum_{i=1}^{\infty} \frac{\operatorname{diam}(T)}{2^i} \sqrt{\log N(T, d, \frac{\operatorname{diam}(T)}{2^i})} &= \sum_{i=1}^{\infty} 2 \int_{\frac{\operatorname{diam} T}{2^{i+1}}}^{\frac{\operatorname{diam} T}{2^i}} d\varepsilon \sqrt{\log N(T, d, \frac{\operatorname{diam}(T)}{2^i})} \\ &= 2 \sum_{i=1}^{\infty} \int_{\frac{\operatorname{diam} T}{2^{i+1}}}^{\frac{\operatorname{diam} T}{2^i}} \sqrt{\log N(T, d, \frac{\operatorname{diam}(T)}{2^i})} d\varepsilon. \end{aligned}$$

Para $\varepsilon \in [\frac{\operatorname{diam} T}{2^{i+1}}, \frac{\operatorname{diam} T}{2^i}]$, vale $N(T,d,\frac{\operatorname{diam}(T)}{2^i}) \leq N(T,d,\varepsilon)$ e portanto, $\log N(T,d,\frac{\operatorname{diam}(T)}{2^i}) \leq \log N(T,d,\varepsilon)$. Por fim, temos:

$$2\sum_{i=1}^{\infty}\int_{\frac{\operatorname{diam} T}{2^{i+1}}}^{\frac{\operatorname{diam} T}{2^i}} \sqrt{\log N(T,d,\frac{\operatorname{diam}(T)}{2^i})}d\varepsilon \leq 2\sum_{i=1}^{\infty}\int_{\frac{\operatorname{diam} T}{2^{i+1}}}^{\frac{\operatorname{diam} T}{2^i}} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon.$$

Podemos usar Teorema da Convergência Monótona para mostrar que

$$2\sum_{i=1}^{\infty}\int_{\frac{\operatorname{diam} T}{2^{i+1}}}^{\frac{\operatorname{diam} T}{2^i}} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon = 2\int_0^{\frac{\operatorname{diam} T}{2}} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon.$$

A integral acima é limitada por:

$$\begin{aligned}
&\leq 2\int_0^{\frac{\operatorname{diam} T}{2}} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon + 2\int_{\frac{\operatorname{diam} T}{2}}^{\operatorname{diam} T} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon \\
&= 2\int_0^{\operatorname{diam} T} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon \\
&= 2\int_0^{\infty} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon.
\end{aligned}$$

□

Observação 2.2.1. *Embora a integral de Dudley seja formalmente sobre* $[0,+\infty]$, *ela coincide com a mesma integral sobre* $[0,\operatorname{diam}(T)]$. *De fato, se* $\varepsilon > \operatorname{diam}(T)$, *qualquer* $\{x\} \in T$ *é* $\varepsilon-$*rede. Nesse caso, teríamos* $N(T,d,\varepsilon) = 1$ *implicando* $\log N(T,d,\varepsilon) = 0$.

Observação 2.2.2. *O cálculo feito em (2.3) mostra que na verdade provamos*[8]*:*

$$\mathbb{E}\sup_{t\in T}|X_t - X_{t_0}| \leq C\int_0^{\infty} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon. \tag{2.6}$$

Aplicando a desigualdade triangular obtemos uma cota para o supremo dos incrementos:

$$\mathbb{E}\sup_{t,s\in T}|X_t - X_s| \leq C\int_0^{\infty} \sqrt{\log N(T,d,\varepsilon)}d\varepsilon. \tag{2.7}$$

Usando a definição de *processos separáveis*, poderíamos dar um enunciado ainda mais geral para o Teorema 2.2, considerando um conjunto T não necessariamente enumerável.

Definição 2.2.3. *Um processo estocástico* $(X_t)_{t\in T}$ *é separável se existir um conjunto enumerável* $T_0 \subset T$ *tal que*

$$X_t = \lim_{\substack{s\to t \\ s\in T_0}} X_s \quad \text{para todo } t \in T \, q.t.p,$$

isto é, existe uma sequência $(s_n)_n \subset T_0$ $s_n \to t$ *tal que* $X_{s_n} \to X_t$.

Note que esta definição é de certa forma intrínseca ao encadeamento, já que a ideia principal é a convergência $X_t = \lim_k X_{\pi_k(t)}$ para todo $t \in T$. O caso em que T é não-enumerável e o processo não é separável não será tratado, uma vez que o supremo $\sup_{t\in T} X_t$ de uma família

[8] Basta ver que $|X_t - X_{t_0}| \leq \sum_{j=1}^{\infty} |X_{\pi_j(t)} - X_{\pi_{j-1}(t)}|$ e repetir toda a demonstração do teorema.

não-enumerável de funções não necessariamente é mensurável. Sob a condição de separabilidade, temos

$$\sup_{t\in T} X_t = \sup_{t\in T_0} X_t \text{ q.t.p}$$

e o problema da mensurabilidade não existe, já que supremo de uma família enumerável de funções mensuráveis é mensurável num espaço de probabilidade completo.

Portanto, a prova do Teorema 2.2 usando a separabilidade ficaria: Seja $T' \subset T$ enumerável tal que $\sup_{t\in T} X_t = \sup_{t\in T_0} X_t$ em quase todo ponto. Denote por T_k os primeiros k elementos de T' em uma ordem arbitrária. Então, pelo Teorema da Convergência Monótona,

$$\mathbb{E}\sup_{t\in T} X_t = \mathbb{E}\sup_{t\in T'} X_t = \sup_{k\geq 1}\mathbb{E}\sup_{t\in T_k} X_t.$$

Aplicando o resultado já provado em cada máximo finito, basta usar que $N(T_k, d, \varepsilon) \leq N(T, d, \varepsilon)$ e teremos o resultado geral. A referência para a definição de processos separáveis e a generalização da Integral de Dudley para esses casos é van Handel [15].

Exemplo 2.2.1. *Vamos ver um caso em que $\mathbb{E}\sup_{t\in T} X_t < +\infty$ mas a integral de Dudley diverge. Considere $T = \mathbb{N}$ e $X_i \sim N(0, a_i^2)$ independentes com $a_i \to 0$. Já sabemos que para distribuições gaussianas, existe constante $C > 0$ tal que*

$$\|X_i - X_j\|_{\psi_2} \leq C\sqrt{a_i^2 + a_j^2} = C\|X_i - X_j\|_{L_2}.$$

Escolhendo a pseudo-métrica $d(t,s) = \|X_t - X_s\|_{L_2}$, segue que o processo $(X_i)_{i\in\mathbb{N}}$ é subgaussiano.

Afirmação 1. *Se $T' \subset \mathbb{N}$ é $\varepsilon-$rede de T, então todo $i \in \mathbb{N}$ com $a_i > \varepsilon$ está em T'. De fato,*

$$d(i,j) = \|X_i - X_j\|_{L_2} = \sqrt{a_i^2 + a_j^2} > \sqrt{\varepsilon^2 + a_j^2} \geq \varepsilon.$$

Isto implica que o índice i está $\varepsilon-$distante de todo índice $j \in \mathbb{N}$. Como T' é $\varepsilon-$rede de T, a única maneira de cobrir i é tê-lo como um centro das bolas que cobrem T. Portanto, $i \in T'$. A convergência $a_i \to 0$ faz com que exista somente um número finito de a_i com $a_i > \varepsilon$ para todo ε. Desta maneira, obtemos uma cota inferior para o número de cobertura:

$$\#\{\, i \in \mathbb{N} \mid a_i > \varepsilon \,\} \leq N(\mathbb{N}, d, \varepsilon).$$

Afirmação 2. *Dado $\varepsilon > 0$, seja $i_0 \in \mathbb{N}$ o índice tal que $a_{i_0} \leq \frac{\varepsilon}{2}$. O conjunto $\left\{\, i \in \mathbb{N}; a_i > \frac{\varepsilon}{2} \,\right\} \cup \{\, a_{i_0} \,\}$ é uma ε-rede de T. De fato, para $y \in \mathbb{N}$ há duas possibilidades: $a_y > \frac{\varepsilon}{2}$, o que força y ser um dos centros das bolas que cobrem T ou $a_y \leq \frac{\varepsilon}{2}$, donde teremos:*

$$d(y, i_0) = \sqrt{a_y^2 + a_{i_0}^2} \leq \sqrt{\frac{\varepsilon^2}{2^2} + \frac{\varepsilon^2}{2^2}} = \frac{\varepsilon}{\sqrt{2}} \leq \varepsilon.$$

Esta afirmação implica

$$N(\mathbb{N}, d, \varepsilon) \leq \#\{\, i \in \mathbb{N} \mid a_i > \varepsilon \,\} + 1 < \infty. \tag{i}$$

Tome $a_i = \sqrt{\frac{1}{\log i + 10}}$. Temos

$$\mathbb{P}(|X_i| > \lambda) \leq 2\exp\left(\frac{\lambda^2}{2}(\log i + 10)\right) = 2\exp\left(\frac{\lambda^2}{2}\log i + 5\lambda^2\right) = i^{\frac{\lambda^2}{2}}\exp(5\lambda^2). \tag{2.8}$$

Se $\lambda \geq 2$, então podemos usar a cota da união para obter:

$$\mathbb{P}(\sup_i |X_i| > \lambda) \leq \sum_i \frac{1}{i^{\frac{\lambda^2}{2}}} \exp(5\lambda^2) < \infty. \tag{2.9}$$

Como $\mathbb{E}\sup_t X_t = \int_0^\infty \mathbb{P}(\sup_t X_t > \lambda)d\lambda$, segue que esta integral e, portanto, $\mathbb{E}\sup_t X_t$ são finitos. Vamos agora mostrar que o número de entropia diverge. Se $\sqrt{\frac{1}{\log i+10}} = a_i > \varepsilon$, então

$$\frac{1}{\log i + 10} > \varepsilon^2 \Leftrightarrow \log i < \frac{1}{\varepsilon^2} - 10 \Leftrightarrow i < \exp\left(\frac{1}{\varepsilon^2} - 10\right).$$

Logo, por (i), temos:

$$N(\mathbb{N}, d, \varepsilon) \leq \exp\left(\frac{1}{\varepsilon^2} - 10\right) \implies \log N(\mathbb{N}, d, \varepsilon) \leq \frac{1}{\varepsilon^2} - 10 \leq \frac{1}{\varepsilon^2}.$$

Combinando estas informações na integral de Dudley discreta, obtemos:

$$\sum_{i=1}^{\infty} \frac{\operatorname{diam} T}{2^i} \sqrt{\log N(T, d, \frac{\operatorname{diam} T}{2^i})} \leq \sum_{i=1}^{\infty} \frac{\operatorname{diam} T}{2^i} \frac{1}{\varepsilon} = \sum_{i=1}^{\infty} \frac{\operatorname{diam} T}{2^i} \frac{2^i}{\operatorname{diam} T} = +\infty.$$

2.2.3 Lei dos Grandes Números Uniforme

Uma importante aplicação da desigualdade de Dudley provada na seção anterior está no estudo dos chamados *processos empíricos*. Como definiremos adiante, os processos empíricos são processos estocásticos indexados por uma certa família de funções. Neste ponto, é possível entender o objetivo desta teoria: passar de resultados que lidam com uma sequência fixa de variáveis aleatórias para obter um comportamento comum em toda uma coleção de variáveis.

Definição 2.2.4. *(Processo Empírico[9]) Seja $\mathscr{F}$ uma classe de funções $f\colon \Omega \to \mathbb{R}$, e (Ω, Σ, μ) espaço de probabilidade. Considere X um ponto aleatório em Ω cuja distribuição é μ e $X_1, \ldots, X_n$ cópias independentes de X. O processo $(X_f)_{f\in\mathscr{F}}$ dado por*

$$X_f \coloneqq \frac{1}{n}\sum_{i=1}^{n}(fX_i - \mathbb{E}\, fX)$$

é chamado processo empírico indexado por $\mathscr{F}$.

Observação 2.2.3. *Alguns livros, por exemplo van der Vaart [14], chamam de processo empírico o processo*

$$X_f \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(fX_i - \mathbb{E}\, fX).$$

O conceito central da Lei dos Grandes Números Uniforme enunciada abaixo é obter um resultado de convergência[10] para toda uma coleção de variáveis aleatórias, enquanto a Lei dos Grandes Números clássica garante a convergência somente para uma função fixa.

[9] Por vezes, aboliremos os parênteses e usaremos a notação $f(X_i) = fX_i$.
[10] Use esse resultado unido a desigualdade de Markov para obter a convergência em probabilidade.

Teorema 2.2.3. *Sejam $X, X_1, \ldots, X_n$ variáveis aleatórias i.i.d tomando valores em $[0,1]$. Então*

$$\mathbb{E} \sup_{f \in \mathscr{F}} |\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)| \leq \frac{C}{\sqrt{n}},$$

sendo $\mathscr{F} \coloneqq \{\, f \colon [0,1] \to \mathbb{R} \mid \| f\|_{Lip} \leq 1 \,\}$ e $\| \cdot \|_{Lip}$ a norma Lipschitz.[11]

Demonstração. Vamos usar a integral de Dudley. O primeiro passo é checar se os incrementos do processo $(Z_f)_{f \in \mathscr{F}}$, com $Z_f \coloneqq \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)$, são subgaussianos. Temos:

$$\begin{aligned}
|Z_f - Z_h| &= |\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X) - \frac{1}{n} \sum_{i=1}^{n} h(X_i) - \mathbb{E}\, h(X)| \\
&= |\frac{1}{n} \sum_{i=1}^{n} f(X_i) - h(X_i) - (\mathbb{E}\, f(X) - \mathbb{E}\, h(X))| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - h(X_i) - (\mathbb{E}\, f(X) - \mathbb{E}\, h(X))| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - h(X_i)| + |\mathbb{E}\, f(X) - \mathbb{E}\, h(X)| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} 2\|f - h\|_\infty = 2\|f - h\|_\infty.
\end{aligned}$$

Como vimos no Exemplo 2.1.3, segue:

$$\|Z_f - Z_h\|_{\psi_2} \leq C_1 \|Z_f - Z_h\|_\infty \leq C_2 \|f - h\|_\infty.$$

O próximo passo é aplicar a integral de Dudley, lembrando que $\operatorname{diam}(\mathscr{F}) \leq 2$:

$$\mathbb{E} \sup_{f \in \mathscr{F}} |X_f| = \mathbb{E} \sup_{f \in \mathscr{F}} |X_f - X_0| \leq C \cdot \frac{1}{\sqrt{n}} \int_0^2 \sqrt{(\log N(\mathscr{F}, \| \cdot \|_\infty, \varepsilon)} d\varepsilon. \tag{2.10}$$

Aqui estamos considerando que a função $f \equiv 0$ pertence a classe $\mathscr{F}$. É possível mostrar que (para mais detalhes consulte o exercício 8.2.6 em Vershyin [16])

$$N(\mathscr{F}, \| \cdot \|_\infty, \varepsilon) \leq \left(\frac{C}{\varepsilon}\right)^{\frac{C}{\varepsilon}},$$

implicando em

$$\mathbb{E} \sup_{f \in \mathscr{F}} |X_f| \leq C \cdot \frac{1}{\sqrt{n}} \int_0^2 \sqrt{\log \left(\frac{C}{\varepsilon}\right)^{\frac{C}{\varepsilon}}} d\varepsilon < C_2 \cdot \frac{1}{\sqrt{n}}.$$

□

[11] Dados (A, d_A) e (B, d_B) espaços métricos e uma função $f \colon A \to B$, definimos $\| \cdot \|_{Lip} \coloneqq \inf_{L \in \mathbb{R}} \{\, d_B(f(u), f(v)) \leq L d_A(u, v);$ para todos $u, v \in A \,\}$.

2.3 Processo empírico via Dimensão VC

A Lei dos Grandes Números Uniforme garantiu uma cota para controlar os processos empíricos sobre uma classe de funções Lipschitz. O que almejamos nesta seção é encontrar um resultado similar para uma família arbitrária de funções Booleanas, o Teorema 2.3.3. Para isto, introduziremos um conceito fundamental em aprendizagem estatística, a dimensão VC, que foi criado por V. Vapnik e A. Chervonenkis na década de 1970. Posteriormente, falaremos sobre a simetrização, uma técnica recorrente no estudo de processos empíricos.

2.3.1 Dimensão VC

Uma função f é dita Booleana quando assume valores binários, isto é, quando seu contradomínio é um conjunto com dois elementos, usualmente $\{0,1\}$. A dimensão VC - *Vapnik-Chervonenkis dimension* - é uma maneira de medir a complexidade de classes de funções Booleanas. Além disso, ela se relaciona com o número de cobertura do conjunto e por meio da Integral de Dudley, fornece informações sobre processos empíricos.

Definição 2.3.1. *(Dimensão VC) Considere $\mathscr{F}$ uma classe de funções Booleanas definidas em Ω. Um subconjunto $\Lambda \subseteq \Omega$ é estilhaçado por $\mathscr{F}$ se qualquer função $g\colon \Lambda \to \{0,1\}$ puder ser obtida pondo $g = f\big|_\Lambda$ para alguma $f \in \mathscr{F}$. A dimensão VC da classe $\mathscr{F}$ é definida como o número* $\operatorname{vc}(\mathscr{F}) = \max\{\, \#\Lambda \mid \Lambda \subseteq \Omega \text{ é estilhaçado} \,\}$.

Exemplo 2.3.1. *(Intervalos) Seja $\mathscr{F} = \{\, \mathbb{1}_{[a,b]} \mid a, b \in \mathbb{R}, a \leq b. \,\}$ Afirmamos que $\operatorname{vc}(\mathscr{F}) = 2$. Seja $\Lambda = \{\, 3, 5 \,\}$. A cada subconjuntos de Λ podemos associar uma lista de $0's$ e $1's$:*

$$\begin{aligned} \emptyset &\iff 00 \iff g_1 \\ \{3\} &\iff 10 \iff g_2 \\ \{5\} &\iff 01 \iff g_3 \\ \{\, 3, 5 \,\} &\iff 11 \iff g_4. \end{aligned}$$

Isto significa que toda função $g : \Lambda \to \{0,1\}$ pode ser vista como uma dupla de 0 e 1. Com isto em mente, é fácil construir funções $f \in \mathscr{F}$ tais que $g_i = f\big|_\Lambda$. De fato, tomando $f_1 = \mathbb{1}_{[2,4]}$, $f_2 = \mathbb{1}_{[4,6]}$, $f_3 = \mathbb{1}_{[2,6]}$, e $f_4 = \mathbb{1}_{[5,7]}$, temos:

$$g_1 = f_4\big|_\Lambda, g_2 = f_1\big|_\Lambda, g_3 = f_2\big|_\Lambda \text{ e } g_4 = f_3\big|_\Lambda.$$

Agora, veremos que nenhum conjunto $\{\, x_1, x_2, x_3 \,\}$ é estilhaçado por $\mathscr{F}$, o que implica $\operatorname{vc}(\mathscr{F}) = 2$. Sem perda de generalidade, podemos assumir $x_1 \leq x_2 \leq x_3$. Os subconjuntos de $\{\, x_1, x_2, x_3 \,\}$ são $\{\, \emptyset, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}, \{\, x_1, x_2, x_3 \,\} \,\}$, gerando as funções

$$\begin{aligned} \emptyset &\iff 000 \iff g_1 \\ \{x_1\} &\iff 100 \iff g_2 \\ \{x_2\} &\iff 010 \iff g_3 \\ \{x_3\} &\iff 001 \iff g_4. \\ \{x_1, x_2\} &\iff 110 \iff g_5. \\ \{x_1, x_3\} &\iff 101 \iff g_6. \\ \{x_2, x_3\} &\iff 011 \iff g_7. \\ \{\, x_1, x_2, x_3 \,\} &\iff 111 \iff g_8. \end{aligned}$$

É fácil ver que não existe $f \in \mathscr{F}$ tal que $g_6 = f\big|_\Lambda$, pois todo intervalo que contém x_1 e x_3 deve conter x_2, já que $x_1 \leq x_2 \leq x_3$.

Podemos dar uma definição equivalente de conjunto estilhaçado por uma família de conjuntos e definir a dimensão VC de classes de conjuntos:

Definição 2.3.2. *(Conjunto Estilhaçado) Considere $\mathscr{B}$ uma classe de conjuntos. Diz-se que um conjunto $A = \{ x_1, \ldots, x_n \}$ é estilhaçado por $\mathscr{B}$ se para toda n-upla $b = (b_1, \ldots, b_n) \in \{ 0, 1 \}^n$, existir $C \in \mathscr{B}$ tal que*

$$(\mathbb{1}_{\{x_1 \in C\}}, \ldots, \mathbb{1}_{\{x_n \in C\}}) = b,$$

ou, equivalentemente, se

$$\left\{ (\mathbb{1}_{\{x_1 \in C\}}, \ldots, \mathbb{1}_{\{x_n \in C\}}) \mid C \in \mathscr{B} \right\} = \{ 0, 1 \}^n .$$

A dimensão VC de B, denotada por $\mathrm{vc}(B)$, é a maior cardinalidade de um conjunto estilhaçado por esta família, isto é, $\mathrm{vc}(B) = \max \{ \#A \mid A \text{ é estilhaçado} \}$.

Dada a natural correspondência entre funções Booleanas e listas de 0's e 1's, isto é, entre conjuntos e funções indicadoras de conjuntos, vemos que a dimensão VC de uma classe de conjuntos é a dimensão VC da classe de funções indicadoras desses conjuntos. Uma função Booleana $f \colon \Omega \to \{0, 1\}$ determina o subconjunto $\Omega_0 = \{ x \in \Omega \mid f(x) = 1 \}$ e o subconjunto gera a função Booleana $f = \mathbb{1}_{\Omega_0}$.

Em outras palavras, dizemos que $A = \{ x_1, \ldots, x_n \} \subset \Omega$ é estilhaçado pela família de funções Booleanas $\mathscr{F}$ se é estilhaçado pela família de conjuntos $\{ C_f \mid f \in \mathscr{F} \}$, sendo $C_f = \{ x \in \Omega \mid f(x) = 1 \}$.

Exemplo 2.3.2. *(Circunferências em $\mathbb{R}^2$) Seja $\mathscr{F} = \left\{ \mathbb{1}_C \mid C \text{ é uma circunferência em } \mathbb{R}^2 \right\}$. Afirmamos que $\mathrm{vc}(\mathscr{F}) = 3$. O conjunto $A = \{ (1,0), (-1,0), (0,1) \}$ é estilhaçado por $\mathscr{F}$. De fato, sejam $C_1 = B((2,0),1)$, $C_2 = B((-2,0),1)$, $C_3 = B((0,2),1)$, $C_4 = B((1,1),1)$, $C_5 = B((-1,1),1)$, $C_6 = B((0,-1),\sqrt{2})$, $C_7 = B((0,0),1)$ e $C_8 = B((0,0),\frac{1}{2})$, então $\{(1,0)\} = C_1 \cap A, \{(-1,0)\} = C_2 \cap A, \{(0,1)\} = C_3 \cap A, \{(1,0),(0,1)\} = C_4 \cap A, \{(1,0),(-1,0)\} = C_5 \cap A, \{(1,0),(-1,0)\} = C_6 \cap A, \{ (1,0),(-1,0),(0,1) \} = C_7 \cap A$ e $\emptyset = C_8 \cap A,$. Veja a figura 2.1:*

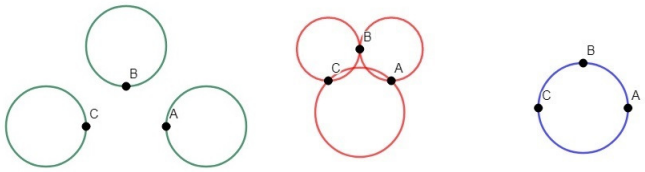

Figura 2.1: As circunferências em verde contém os subconjuntos unitários, as vermelhas contém os subconjuntos com dois pontos e a circunferência azul contém os três pontos.

Qualquer conjunto $\{x_1, x_2, x_3, x_4\} \subset \mathbb{R}^2$ *não pode ser estilhaçado por* $\mathscr{F}$. *Para provar isto, vamos analisar dois casos. Considere primeiramente que* $\{x_1, x_2, x_3\}$ *são colineares. Neste caso, o conjunto não pode ser estilhaçado pelo mesmo raciocínio do Exemplo 2.3.1. Quando* $\{x_1, x_2, x_3\}$ *não estão arranjados de forma colinear, sabemos por argumento geométricos*[12] *que existe uma, e somente uma, circunferência* C_1 *que contém esses três pontos. A* $\{x_4\}$ *resta duas situações possíveis: pertencer ou não a* C_1. *Se* $x_4 \in C_1$, *então não existe circunferência que contenha* $\{x_1, x_2, x_3\}$ *e não contenha* $\{x_4\}$. *Se* $x_4 \notin C_1$, *não existe outra circunferência que contenha* $\{x_1, x_2, x_3, x_4\}$. *Portanto, o conjunto não pode ser estilhaçado.*

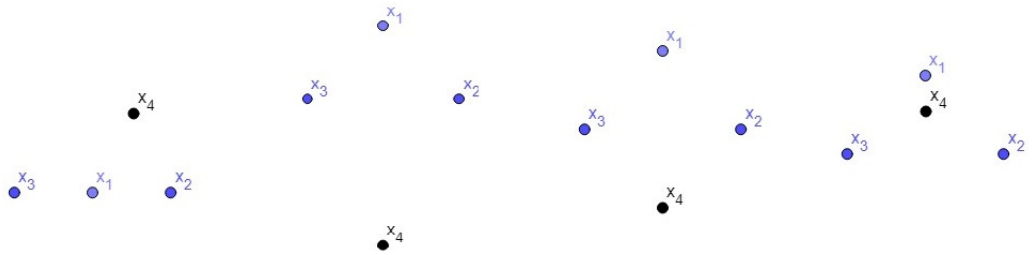

Figura 2.2: Da esquerda para direita: três pontos colineares; três pontos não-colineares em que um deles ficaria exterior a circunferência; três pontos não-colineares de tal forma que não há circunferência contendo somente três; três pontos não-colineares em que um deles ficaria interior a circunferência.

Exemplo 2.3.3. *(Semi-espaços em* $\mathbb{R}^n$*) Seja* $\mathscr{F}$ *a classe de todos os semi-espaços em* $\mathbb{R}^n$. *Então* $\mathrm{vc}(\mathscr{F}) = n + 1$. *Consulte o exemplo 4.21 do livro Wainwright [17].*

Usando a dimensão VC de uma classe de funções Booleanas, podemos encontrar uma cota para o número de cobertura $N(\mathscr{F}, L^2(\mu), \varepsilon)$, sendo μ medida de probabilidade em Ω e a norma $L^2(\mu)$ definida abaixo:

$$d(f, g) = \|f - g\|_{L^2(\mu)} = \left(\int_\Omega |f - g|^2 d\mu \right)^{\frac{1}{2}}, \quad f, g \in \mathscr{F}.$$

Teorema 2.3.1. *(Número de cobertura via dimensão VC) Seja* $\mathscr{F}$ *uma classe de funções Booleanas num espaço de probabilidade* (Ω, Σ, μ). *Então, para todo* $\varepsilon \in (0, 1)$, *temos*

$$N(\mathscr{F}, L^2(\mu), \varepsilon) \leq \left(\frac{2}{\varepsilon} \right)^{C\,\mathrm{vc}(\mathscr{F})}$$

Demonstração. Ver Teorema 8.3.18 em Vershyin [16]. □

2.3.2 Simetrização de processos empíricos

Os processos empíricos e os processos gaussianos estão intimamente relacionados pelo Teorema Central do Limite.

[12] Basta traçar o único triângulo que passa pelos três pontos fixados e o encontro de suas mediatrizes é o circuncentro. Como ele é equidistante de $\{x_1, x_2, x_3\}$, está garantida a existência e unicidade da circunferência em questão.

Teorema 2.3.2. *Para $f_1, \ldots, f_k \in \mathscr{F}$, temos*

$$\sqrt{n}(X_{f_1}, \ldots, X_{f_k}) \to_D (Z(f_1), \ldots, Z(f_k)),$$

quando $n \to \infty$ sendo $((Z(f))_{f\in\mathscr{F}}$ processo gaussiano com $\operatorname{cov}[Z(f), Z(g)] = \operatorname{cov}_m u[f, g]$.

Por causa desse resultado, espera-se que para n suficientemente grande, o processo empírico tenha um comportamento do tipo gaussiano. Porém, para n fixo, isto não é verdade. Apesar disso, conseguimos mostrar que com a métrica $d(f, g) := \|f - g\|_\infty$, o processo $(X_f)_{f\in\mathscr{F}}$ é subgaussiano.

Em alguns casos, esta métrica pode ser muito maior que a métrica $L^2(\mu)$ definida anteriormente, o que resultada na perda de eficiência em controlar o processo empírico comparado ao gaussiano. O exemplo seguinte, retirado de van Handel [15], ilustra esse fato.

Exemplo 2.3.4. *Sejam $X_1, \ldots, X_n, \ldots$ variáveis aleatórias independentes e identicamente distribuídas com distribuição μ. Para todo $x \in \mathbb{R}$, é consequência direta da Lei dos Grandes Números que a função de distribuição empírica $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[X_i\leq x]} \to F(x) = \mu(-\infty, x]$ em quase todo ponto. É possível mostrar que a convergência é na verdade, uniforme:*

$$\|F_n - F\|_\infty \xrightarrow{n\to\infty} 0 \quad q.t.p.$$

Para entender o fenômeno, vamos estudar o supremo desse processo

$$\sup_{f\in\mathscr{F}} X_f$$

sobre a classe $\mathscr{F} := \{\mathbb{1}_{(-\infty,x]} \mid x \in \mathbb{R}\}$. Quando $x_1 < x_2$, então

$$\|\mathbb{1}_{(-\infty,x_1]} - \mathbb{1}_{(-\infty,x_2]}\|_\infty = 1.$$

Então toda função em $\mathscr{F}$ está a uma distância 1 de outra função na classe, fazendo com que $N(\mathscr{F}, \|\cdot\|_\infty, \varepsilon) = \infty$ para qualquer $0 < \varepsilon < 1$.

O exemplo anterior mostra como o argumento de encadeamento poderá falhar usando a métrica $\|\cdot\|_\infty$. Por outro lado, alguns cálculos mostram que a métrica $\|f - g\|_{L^2(\mu)}$ torna o número de cobertura $N(\mathscr{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon)$ pequeno nesta classe.

Toda esta análise preliminar foi para motivar os mecanismos que usaremos para controlar o supremo de processos empíricos. Precisamos de nova ferramenta que capture seu comportamento gaussiano e o método de *simetrização* cumpre esse papel. Antes de defini-lo, vamos dar uma justificativa informal sobre seu funcionamento. Se quisermos entender por que a simetrização nos ajuda nesse problema, precisamos antes entender o funcionamento do Teorema Central do Limite. A discussão seguinte foi traduzida do início do capítulo 7 em van Handel [15].

Fixe f uma função limitada e considere a soma $\sum_{i=1}^n (fX_i - \mathbb{E}\, fX)$. Já que cada parcela tem ordem 1, a soma pode se tornar tão grande quanto n no pior caso. Porém, o Teorema Central do Limite mostra que a soma é somente de ordem $\sqrt{n}$ em probabilidade! A razão é a seguinte: para que a soma tenha ordem n, a maioria das parcelas deve ter o mesmo sinal para que suas contribuições sejam somadas. Acontece que os termos são independentes e centrados, e dificilmente terão o mesmo sinal. Tipicamente, há vários termos de sinais opostos que se cancelam. Daí a redução de $O(n)$ para $O(\sqrt{n})$.

O cancelamento dos sinais é o mecanismo chave para o TCL[13]: é o efeito agregado dos sinais aleatórios que leva ao comportamento gaussiano. As outras características da distribuição μ

[13] Teorema Central do Limite.

afetam o limite somente para determinar sua variância. Isso sugere que para obter o comportamento gaussiano dos processos empíricos, devemos de alguma forma isolar os sinais de maneira a capturar somente a "parte gaussiana"dos processos empíricos. A simetrização torna isto possível, e por conseguinte, estaremos aptos a aplicar as ferramentas já desenvolvidas. Antes de obter o resultado de simetrização, precisaremos de dois lemas que relacionam funções convexas e supremo de processos usando a Desigualdade de Jensen.

Lema 2.3.1. *Seja $\mathscr{F}$ uma família arbitrária de funções. A função $X \mapsto F(X) = \sup_{f\in T} |fX|$ é convexa, para X uma variável aleatória.*

Demonstração. De fato, considerando $t \in (0,1)$ e duas variáveis aleatórias $X \neq Y$, a convexidade segue da desigualdade triangular e das propriedades do supremo:

$$\begin{aligned} F(tX + (1-t)Y) &= \sup_{f\in\mathscr{F}} |tfX + (1-t)fY| \leq \sup_{f\in\mathscr{F}} t|fX| + (1-t)|fY| \\ &\leq \sup_{f\in\mathscr{F}} t|fX| + \sup_{f\in\mathscr{F}} (1-t)|fY| = t \sup_{f\in\mathscr{F}} |fX| + (1-t) \sup_{f\in\mathscr{F}} |fY| \\ &= tF(X) + (1-t)F(Y). \end{aligned}$$

□

Lema 2.3.2. *Seja $\mathscr{F}$ uma classe de funções $f\colon \Omega \to \mathbb{R}$, sendo (Ω, Σ, μ) espaço de probabilidade. Considere $X_1, \ldots, X_n$ variáveis aleatórias independentes em Ω e $\epsilon_1, \ldots, \epsilon_n$ variáveis aleatórias independentes com distribuição de Bernoulli simétrica que também são independentes de $X_1, \ldots, X_n$. Se $\mathbb{E}\, fX_i = 0$ para todo $i = 1, \ldots, n$ e para toda função $f \in \mathscr{F}$, então*

$$\mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^n fX_i| \leq 2\, \mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^n \epsilon_i fX_i|.$$

Demonstração. Sejam $X_1', \ldots, X_n'$ cópias independentes de $X_1, \ldots, X_n$. Então $X_1 - X_1', \ldots, X_n - X_n'$, são independentes, simétricas e tem a mesma distribuição[14] de $\epsilon_1(X_1 - X_1'), \ldots, \epsilon_n(X_n - X_n')$. Note que

$$\sup_{f\in\mathscr{F}} |\sum_{i=1}^n fX_i| = \sup_{f\in\mathscr{F}} |\sum_{i=1}^n fX_i - \mathbb{E}\, fX_i'|,$$

e para cada $\omega_0 \in \Omega$ fixo, vale

$$\sup_{f\in\mathscr{F}} |\sum_{i=1}^n f(X_i(\omega_0)) - \mathbb{E}\, fX_i'| = \sup_{f\in\mathscr{F}} |\,\mathbb{E} \left(\sum_{i=1}^n f(X_i(\omega_0)) - fX_i' \right)|. \tag{2.11}$$

Pela Desigualdade de Jensen aplicada a (2.11),

$$\sup_{f\in\mathscr{F}} |\sum_{i=1}^n f(X_i(\omega_0)) - fX_i'| \leq \mathbb{E} \left(\sup_{f\in\mathscr{F}} |\sum_{i=1}^n f(X_i(\omega_0)) - fX_i'| \right).$$

[14] Para demonstrar esse fato, basta olhar para a função características destas variáveis.

Como isto é válido para $\omega_0 \in \Omega$ arbitrário, obtemos

$$\begin{aligned}
\mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^{n} fX_i| &= \mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^{n} fX_i - \mathbb{E}\, fX_i'| \\
&\leq \mathbb{E}\left(\mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^{n} fX_i - fX_i'|\right) \\
&= \mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^{n} (fX_i - fX_i')| \\
&= \mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^{n} \epsilon_i(fX_i - fX_i'| = \mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^{n} \epsilon_i fX_i - \sum_{i=1}^{n} \epsilon_i fX_i'| \\
&\leq \mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^{n} \epsilon_i fX_i| + \mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^{n} \epsilon_i fX_i'| \\
&\leq 2\,\mathbb{E} \sup_{f\in\mathscr{F}} |\sum_{i=1}^{n} \epsilon_i fX_i|.
\end{aligned}$$

□

Estamos prontos para entender a simetrização e finalmente aplicá-la no estudo de valor esperado do supremo de um processo empírico.

Proposição 2.3.1. *Seja $\mathscr{F}$ uma classe de funções $f\colon \Omega \to \mathbb{R}$, sendo (Ω, Σ, μ) espaço de probabilidade. Considere X um ponto aleatório em Ω cuja distribuição é μ e $X_1, \ldots, X_n$ cópias independentes de X. Então*

$$\mathbb{E} \sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^{n} (fX_i - \mathbb{E}\, fX)| \leq 2\,\mathbb{E} \sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^{n} \epsilon_i fX_i|,$$

onde $\epsilon_1, \ldots, \epsilon_n$ são variáveis aleatórias independentes com distribuição de Bernoulli simétrica que também são independentes de $X_1, \ldots, X_n$.

Demonstração. Aplicando o Lema 2.3.2 para $\frac{1}{n}\sum_{i=1}^{n}(fX_i - \mathbb{E}\, fX)$ obtemos o que queríamos:

$$\mathbb{E} \sup_{f\in T} |\frac{1}{n}\sum_{i=1}^{n} (fX_i - \mathbb{E}\, fX)| \leq 2\,\mathbb{E} \sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^{n} \epsilon_i fX_i|.$$

□

Por fim, enunciaremos o resultado principal desta seção.

Teorema 2.3.3. *Seja $\mathscr{F}$ uma classe de funções Booleanas num espaço de probabilidade (Ω, Σ, μ) com $1 \leq \mathrm{vc}(\mathscr{F}) < \infty$ tal que a função $f_0 \equiv 0$ pertença a $\mathscr{F}$. Considere $X, X_1, \ldots, X_n$ pontos aleatórios e independentes em Ω cuja distribuição é μ. Então*

$$\mathbb{E} \sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^{n} (fX_i - \mathbb{E}\, fX)| \leq C\sqrt{\frac{\mathrm{vc}(\mathscr{F})}{n}}.$$

Demonstração. Pelo Lema 2.3.1,

$$\mathbb{E}\sup_{f\in\mathscr{F}}|\frac{1}{n}\sum_{i=1}^{n}(fX_i-\mathbb{E}\,fX)|\leq\frac{2}{\sqrt{n}}\,\mathbb{E}\sup_{f\in\mathscr{F}}|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i fX_i|.$$

A ideia desta demonstração é condicionar nas variáveis $X_{1:n}=(X_1,\ldots,X_n)$, deixando toda a aleatoriedade em ϵ_i e posteriormente, usar a Integral de Dudley para cotar o processo de média zero $(Z_f)_{f\in\mathscr{F}}$, sendo $Z_f\coloneqq\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i fX_i$.

Para tanto, precisamos checar a subgaussianidade dos incrementos de Z_f. Vamos usar dois fatos: $\|\epsilon_i\|_{\psi_2}=\frac{1}{\sqrt{\log 2}}$, do Exemplo 2.1.3 e a Proposição 2.6.1 em Vershyin [16] que fornece $\|\sum_{i=1}^{n}X_i\|_{\psi_2}^2\leq C_1\sum_{i=1}^{n}\|X_i\|_{\psi_2}^2$. Temos:

$$\|Z_f-Z_g\|_{\psi_2}=\frac{1}{\sqrt{n}}\|\sum_{i=1}^{n}\epsilon_i(fX_i-gX_i)\|_{\psi_2}\leq\frac{\sqrt{C_1}}{\sqrt{\log 2}}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}(fX_i-gX_i)^2\right)^{\frac{1}{2}}.$$

Uma maneira de ler a última expressão é ver que podemos definir uma medida μ_n, que é uniforme em $\{\,X_1,\ldots,X_n\,\}\subset\Omega$. Como esta medida é suportada em $\{\,X_1,\ldots,X_n\,\}$, a integral é na verdade uma soma finita:

$$d(f,g)\coloneqq\|f-g\|_{L^2(\mu_n)}=\int_{\{\,X_1,\ldots,X_n\,\}}(f(x)-g(x))^2d\mu_n=\left(\frac{1}{n}\sum_{i=1}^{n}(fX_i-gX_i)^2\right)^{\frac{1}{2}}.$$

Agora usamos a Integral de Dudley na forma (2.6) condicionalmente em $X_{1:n}$:

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}}|Z_f-Z_{f_0}\mid X_{1:n}|\right]\leq C_2\int_0^{\operatorname{diam}(\mathscr{F})}\sqrt{\log N(\mathscr{F},L^2(\mu_n),\varepsilon)}d\varepsilon.$$

Perceba que $\mathbb{E}\sup_{f\in\mathscr{F}}|Z_f-Z_{f_0}\mid X_{1:n}=\alpha_i|=\mathbb{E}\sup_{f\in\mathscr{F}}|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(\alpha_i)|$. Além disso, $\mathbb{E}[(\sup_{f\in\mathscr{F}}|Z_f-Z_{f_0}|)\mid X_{1:n}=\alpha_i]=\mathbb{E}[(\sup_{f\in\mathscr{F}}|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(\alpha_i)|)]$. Portanto,

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}}|Z_f-Z_{f_0}\mid X_{1:n}|\right]=\mathbb{E}\left[(\sup_{f\in\mathscr{F}}|Z_f-Z_{f_0}|)\mid X_{1:n}\right].$$

Substituindo na Integral de Dudley, ficamos com:

$$\mathbb{E}\left[(\sup_{f\in\mathscr{F}}|Z_f-Z_{f_0}|)\mid X_{1:n}\right]\leq C_2\int_0^{\operatorname{diam}(\mathscr{F})}\sqrt{\log N(\mathscr{F},L^2(\mu_n),\varepsilon)}d\varepsilon.$$

Note que esta desigualdade está comparando variáveis aleatórias ambas dependentes de X_i, e portanto tomamos o valor esperado dos dois lados para obter:

$$\mathbb{E}\left[\mathbb{E}(\sup_{f\in\mathscr{F}}|Z_f-Z_{f_0}|)\mid X_{1:n}\right]\leq\mathbb{E}\left[C_2\int_0^{\operatorname{diam}(\mathscr{F})}\sqrt{\log N(\mathscr{F},L^2(\mu_n),\varepsilon)}d\varepsilon\right]$$

e pelas propriedades de esperança condicional o lado esquerdo fica:

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}}|Z_f-Z_{f_0}|\right]\leq\mathbb{E}\left[C_2\int_0^{\operatorname{diam}(\mathscr{F})}\sqrt{\log N(\mathscr{F},L^2(\mu_n),\varepsilon)}d\varepsilon\right].$$

Agora vamos nos concentrar em cotar o lado direito da desigualdade. É fácil ver que $\operatorname{diam}(\mathscr{F}) \leq 2$ pois $\operatorname{diam}(\mathscr{F}) = \sup_{f,g\in\mathscr{F}} \|f-g\|_{L^2(\mu_n)} \leq 2$ pois f e g são funções Booleanas, e portanto cotadas por 1. O Teorema 2.3.1 fornece uma cota uniforme para esta variável, já que podemos cotar o integrando por

$$\log N(\mathscr{F}, L^2(\mu_n), \varepsilon) \leq C_3 \cdot \operatorname{vc}(\mathscr{F}) \log \frac{2}{\varepsilon}.$$

Reunindo os cálculos na integral, temos a integral de $\sqrt{\log \frac{2}{\varepsilon}}$, que é cotada por uma constante absoluta D. Isto dá

$$\begin{aligned}
\mathbb{E} \sup_{f\in\mathscr{F}} |Z_f - Z_{f_0}| &\leq C_2 \,\mathbb{E} \int_0^2 \sqrt{\log N(\mathscr{F}, L^2(\mu_n), \varepsilon)}\, d\varepsilon \\
&\leq C_2 \,\mathbb{E} \int_0^2 \sqrt{C_3 \cdot \operatorname{vc}(\mathscr{F}) \log \frac{2}{\varepsilon}}\, d\varepsilon \\
&\leq C_2 \cdot \sqrt{C_3 \cdot \operatorname{vc}(\mathscr{F})}\, \mathbb{E} \int_0^2 \sqrt{\log \frac{2}{\varepsilon}}\, d\varepsilon \\
&\leq C_2 \cdot \sqrt{C_3} \sqrt{\operatorname{vc}(\mathscr{F})} \cdot D.
\end{aligned}$$

Note que $|Z_f - Z_{f_0}| = |Z_f|$ pois

$$Z_{f_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f_0 X_i = \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i \cdot 0 = 0,$$

e isto implica $\mathbb{E} \sup_{f\in\mathscr{F}} |Z_f - Z_{f_0}| = \mathbb{E} \sup_{f\in\mathscr{F}} |Z_f|$ Por fim, temos:

$$\begin{aligned}
\mathbb{E} \sup_{f\in\mathscr{F}} |\frac{1}{n} \sum_{i=1}^n (f X_i - \mathbb{E}\, f X)| &\leq \frac{2}{\sqrt{n}}\, \mathbb{E} \sup_{f\in\mathscr{F}} |Z_f| \\
&\leq \frac{2}{\sqrt{n}} \cdot C_2 \cdot \sqrt{C_3} \sqrt{\operatorname{vc}(\mathscr{F})} \cdot D \\
&= C \sqrt{\frac{\operatorname{vc}(\mathscr{F})}{n}}.
\end{aligned}$$

□

Observação 2.3.1. *Adicionar a função $f_0 \equiv 0$ a família $\mathscr{F}$ não causa aumento significativo em* $\operatorname{vc}(\mathscr{F})$. *De fato, suponha que* $\operatorname{vc}(\mathscr{F}) = k$. *Usando a analogia entre funções Booleanas e listas de 0's e 1's, isto significa que qualquer k-upla de 0's e 1's pode ser realizada por uma função em $\mathscr{F}$ e que existe alguma $(k+1)$-upla que não pode ser realizada. Caso esta $(k+1)$-upla contenha 1 em alguma posição, adicionar $f_0 \equiv 0$ a família $\mathscr{F}$ não influencia em* $\operatorname{vc}(\mathscr{F})$, *pois $f_0 \equiv 0$ só realiza as listas nulas. Portanto, para que f_0 contribua no aumento de* $\operatorname{vc}(\mathscr{F})$, *uma $(k+1)$-upla que contenha somente zeros não pode ser estilhaçada por $\mathscr{F}$. Do contrário, teríamos* $\operatorname{vc}(\mathscr{F}) = k+1$.

Então suponha que f_0 contribua no aumento de $\operatorname{vc}(\mathscr{F})$. *Afirmamos que qualquer $(k+2)$-upla não pode ser realizada por $\mathscr{F} \cup \{f_0\}$. Se supusermos por contradição que qualquer $(k+2)$-upla é realizada por $\mathscr{F} \cup \{f_0\}$, a lista com 0 até a posição k+1 e 1 na posição k+2 é, em particular, realizada por $\mathscr{F} \cup \{f_0\}$. Está aí a contradição, pois isto significa que $\mathscr{F}$ estilhaça a $(k+1)$-upla formada somente por 0. Concluímos que* $\operatorname{vc}(\mathscr{F} \cup \{f_0\}) \leq \operatorname{vc}(\mathscr{F}) + 1$.

Uma pergunta razoável que pode surgir durante a prova desse teorema é por que condicionar a $X_{1:n}$? Deixar a aleatoriedade nas variáveis ϵ é uma vantagem pois além destas variáveis serem simétricas, o cálculo da norma ψ_2 é muito mais simples. Além disso, ao explorar o lema de simetrização, o problema de cotar o supremo de um processo empírico qualquer foi reduzido a controlar o valor esperado $\mathbb{E}\sup_{f\in\mathscr{F}}|Z_f|$.

Um caso particular do Teorema 2.3.3 é o Teorema de Glivenko-Cantelli, que fala sobre a convergência uniforme da distribuição empírica para a distribuição cumulativa.

Teorema 2.3.4. *(Glivenko-Cantelli) Sejam $X_1,\dots,X_n$ variáveis aleatórias independentes com a mesma distribuição cumulativa F. Então*

$$\mathbb{E}\,\|F_n - F\|_\infty = \mathbb{E}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| \leq \frac{C}{\sqrt{n}.}$$

Demonstração. Considere $\mathscr{F} \coloneqq \{\mathbb{1}_{(-\infty,x]} \mid x \in \mathbb{R}\}$ e μ a distribuição de X_i, isto é, $\mu(A) = \mathbb{P}\{X \in A\}$ para cada boreliano $A \subset \mathbb{R}$. Pelo Exemplo 2.3.1, $\mathrm{vc}(\mathscr{F}) \leq 2$. □

2.4 Teoria de Aprendizagem Estatística

A teoria de aprendizagem estatística almeja fornecer predições baseadas em dados. Em forte contraste com a estatística clássica, quando em geral os esforços estão concentrados em estimadores, grande parte dos problemas em aprendizagem estatística está focada em predição. Há uma diferença sutil em cada uma das abordagens: enquanto os estimadores usam dados para estimar um certo parâmetro, por exemplo a média, os *preditores*[15] são funções que usam os dados para "adivinhar"um valor aleatório que não faz parte da sua amostra a priori.

Um problema típico de predição pode ser formulado da seguinte maneira: sejam X e Y variáveis aleatórias com distribuição conjunta P. Somente X é conhecida, e o objetivo é predizer o valor de Y com base na observação X. A premissa importante aqui é que os detalhes sobre a distribuição P ou são vagos ou são inexistentes, restando apenas de informação uma sequência de n observações independentes $(X_1,Y_1),\dots(X_n,Y_n)$ geradas por P. Algumas perguntas razoáveis logo surgem: qual o tamanho mínimo da amostra para que se tenha uma boa predição? Como medir a acurácia desta predição?

Podemos citar algumas aplicações de aprendizagem estatística em outras áreas do conhecimento: predizer o valor de uma ação em seis meses, predizer se um paciente que está doente ou não, estimar a quantidade de glicose no sangue de uma pessoa, identificar o risco de certas doenças como o câncer e etc.

Os resultados, exemplos e definições foram retirados do livro Vershyin [16].

2.4.1 Problemas de Classificação

Os problemas que nos concentraremos aqui são os chamados *problemas de classificação binária*. A formulação matemática é simples: considere uma função Booleana $T\colon \Omega \to \mathbb{R}$ e suponha que T seja desconhecida. Queremos usar uma amostra finita $X_1,\dots,X_n \in \Omega$ gerada de forma independente a partir de uma distribuição $\mathbb{P}$ em Ω, para encontrar uma boa predição de $T(X)$, sendo $X \in \Omega$ um ponto aleatório. Chamamos de *função alvo* a função T e de *dados de treinamento*[16] a amostra

$$(X_i, T(X_i)), \quad i = 1,\dots,n.$$

[15] Do inglês, *predictors*.
[16] Do inglês, *training data*.

Quando T é uma função Booleana, o conjunto Ω fica dividido em duas classes. Usando ferramentas como a dimensão VC e a integral de entropia de Dudley, é possível dar uma cota superior ao valor esperado do risco empírico e fornecer o tamanho necessário de uma amostra para que esse valor esperado tenha risco excessivo cotado por ε arbitrário.

Exemplo 2.4.1. *Considere um estudo de saúde em uma amostra de n pacientes. Os parâmetros a serem coletados, por exemplo, pressão sanguínea, peso corporal, doenças cardíacas e etc, formam um vetor $X_i \in \mathbb{R}^d$. Suponha que sabemos quando esses pacientes possuem diabetes, informação que é traduzida pondo $T(X_i) \in \{0,1\}$ (1=diabético, 0=não diabético). O objetivo é usar os dados de treinamento para obter uma função alvo $T\colon \mathbb{R}^d \to \{0,1\}$ que forneça o diagnóstico de diabetes para qualquer pessoa.*

2.4.2 Risco e complexidade

Uma solução para o problema de aprendizagem pode ser expressa como uma função $f\colon \Omega \to \mathbb{R}$. O objetivo, é claro, é que f seja o mais próxima da função alvo T possível, o que é atingido minimizando o *risco*

$$R(f) \coloneqq \mathbb{E}\left(f(X) - T(X)\right)^2.$$

Nesta definição, X é uma variável aleatória com a mesma distribuição $\mathbb{P}$ da amostra $X_1, \ldots, X_n \in \Omega$. Como T e f são Booleanas, a expressão acima é, na verdade, a probabilidade de uma classificação errada:

$$\begin{aligned} R(f) = \mathbb{E}\left(f(X) - T(X)\right)^2 &= 0 \cdot \mathbb{P}\left\{(f(X) - T(X))^2 = 0\right\} + 1 \cdot \mathbb{P}\left\{(f(X) - T(X))^2 = 1\right\} \\ &= \mathbb{P}\left\{f(X) \neq T(X)\right\}. \end{aligned}$$

A quantidade de dados necessária dependerá da complexidade do problema. Pode ser que a função T seja muito dependente X, ou não. Além disso, uma boa escolha de $\mathscr{F}$, a qual chamamos *espaço de hipóteses*, é determinante para encontrar a solução. Nossa intuição nos leva a pensar que quanto maior o espaço de hipóteses, mais chances de encontrar uma solução ótima. Ou, no extremo oposto, escolher $\mathscr{F}$ mais restrito demandará uma amostra de tamanho menor.

Ocorre que o cenário ideal é um equilíbrio das duas situações: não podemos tomar $\mathscr{F}$ tão pequena para não correr o risco de subestimar a complexidade do caso, tampouco queremos $\mathscr{F}$ muito ampla a ponto de a solução ser afetada por ruídos. A figura abaixo, retirada do livro Vershyin [16], ilustra esse paradigma.

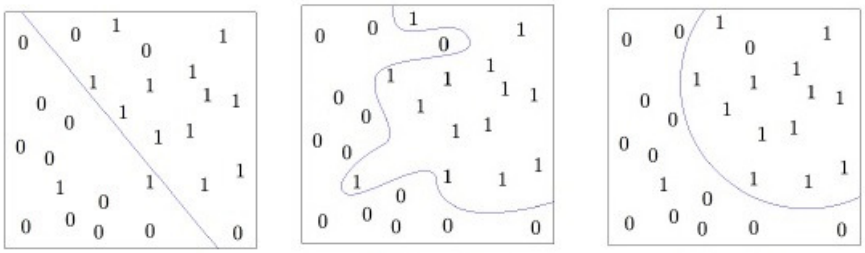

Figura 2.3: Subajuste, sobreajuste e ajuste correto, respectivamente.

2.4.3 Risco Empírico via Dimensão VC

Idealmente, queremos achar uma função f^* no espaço de hipótese $\mathscr{F}$ que minimiza o risco $R(f) = \mathbb{E}\left(f(X) - T(X)\right)^2$:

$$f^* = \arg\min_{f \in \mathscr{F}} R(f).$$

Pode ser que a função T pertença a $\mathscr{F}$, e nesse caso, o risco é zero. Ocorre que não é possível calcular o risco e portanto, f^* somente a partir dos dados de treinamento. O que esses dados fornecem é uma estimativa dos valores de $R(f)$ e de f^*.

Definição 2.4.1. *O risco empírico para uma função $f\colon \Omega \to \mathbb{R}$ é definido como*

$$R_n(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - T(X_i)\right)^2.$$

Denotamos por f_n^* a função no espaço de hipóteses $\mathscr{F}$ que minimiza o risco empírico:

$$f_n^* = \arg\min_{f \in \mathscr{F}} R_n(f).$$

Verifica-se que podemos produzir um excesso de risco quando estamos usando uma amostra finita de tamanho n. Quão grande pode ser a diferença

$$R_n(f_n^*) - R(f^*)?$$

O próximo teorema se debruça sobre essa questão. Antes disso, provaremos um lema necessário para sua demonstração.

Lema 2.4.1. *(Excesso de risco via desvios uniformes) A seguinte desigualdade é válida pontualmente:*

$$R(f_n^*) - R(f^*) \leq 2\sup_{f \in \mathscr{F}} |R_n(f) - R(f)|.$$

Demonstração. Seja $\varepsilon \coloneqq \sup_{f\in\mathscr{F}} |R_n(f) - R(f)|$. Como $f_n^* \in \mathscr{F}$, temos $R(f_n^*) - R_n(f_n^*) \leq \varepsilon$, o que implica

$$\begin{aligned} R(f_n^*) &\leq R_n(f_n^*) + \varepsilon \\ &\leq R_n(f^*) + \varepsilon, \end{aligned}$$

pois $f_n^* = \arg\min R_n$. Isto resulta em

$$\begin{aligned} R(f_n^*) - R(f^*) &\leq R_n(f^*) - R(f^*) + \varepsilon \\ &\leq 2\varepsilon. \end{aligned}$$

□

Teorema 2.4.1. *(Excesso de risco via dimensão VC) Assuma que a função alvo T é uma função Booleana e o espaço de hipóteses $\mathscr{F}$ é uma classe de funções Booleanas com dimensão VC finita e $\mathrm{vc}(\mathscr{F}) \geq 1$. Então existe constante $C > 0$ tal que*

$$\mathbb{E}\, R(f_n^*) \leq R(f^*) + C \cdot \sqrt{\frac{\mathrm{vc}(\mathscr{F})}{n}}. \tag{2.12}$$

Demonstração. É suficiente mostrar que

$$\mathbb{E}\sup_{f\in\mathscr{F}}|R_n(f)-R(f)|\leq C\cdot\sqrt{\frac{\mathrm{vc}(\mathscr{F})}{n}}.$$

De fato, pelo Lema 2.4.1 vale

$$\begin{aligned}\mathbb{E}\,R(f_n^*)&\leq\mathbb{E}\,R(f^*)+2\,\mathbb{E}\sup_{f\in\mathscr{F}}|R_n(f)-R(f)|\\&=R(f^*)+2\,\mathbb{E}\sup_{f\in\mathscr{F}}|R_n(f)-R(f)|,\end{aligned}$$

donde seguirá a desigualdade (2.12). Para chegar a desigualdade acima, lembre-se que $R(f_n^*)$ é uma função das variáveis aleatórias X_i, enquanto $R(f^*)$ é um valor determinístico. Portanto, basta tomar o valor esperado na desigualdade dada pelo lema anterior. Usando as definições de risco empírico e o verdadeiro risco, temos:

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathscr{F}}|R_n(f)-R(f)|&=\mathbb{E}\sup_{f\in\mathscr{F}}|\frac{1}{n}\sum_{i=1}^{n}\left(f(X_i)-T(X_i)\right)^2-\mathbb{E}\left(f(X)-T(X)\right)^2|\\&=\mathbb{E}\sup_{f\in\mathscr{F}}|\frac{1}{n}\sum_{i=1}^{n}\left[(f-T)X_i\right]^2-\mathbb{E}\left[(f-T)(X)\right]^2|\\&=:\mathbb{E}\sup_{f\in\mathscr{F}}|\frac{1}{n}\sum_{i=1}^{n}(f-T)^2(X_i)-\mathbb{E}(f-T)^2(X)|\\&=\mathbb{E}\sup_{l\in\mathscr{L}}|\frac{1}{n}\sum_{i=1}^{n}l(X_i)-\mathbb{E}\,l(X)|,\end{aligned}$$

sendo $\mathscr{L}:=\left\{\,(f-T)^2\mid f\in\mathscr{F}\,\right\}$. Para não cair numa cota que depende da dimensão VC da classe $\mathscr{L}$, não vamos aplicar diretamente o Teorema 2.3.3. No entanto, podemos repetir o roteiro de sua demonstração (simetrização e integral de Dudley) e chegar em

$$\mathbb{E}\sup_{l\in\mathscr{L}}|\frac{1}{n}\sum_{i=1}^{n}l(X_i)-\mathbb{E}\,l(X)|\leq\frac{1}{\sqrt{n}}\,\mathbb{E}\int_0^{\mathrm{diam}(\mathscr{L})}\sqrt{\log N(\mathscr{L},L^2(\mu_n),\varepsilon)}d\varepsilon.$$

O Lema 2.4.2 mostrará que os números de cobertura de $\mathscr{F}$ e $\mathscr{L}$ estão relacionados:

$$N(\mathscr{L},L^2(\mu_n),\varepsilon)\leq N(\mathscr{F},L^2(\mu_n),\frac{\varepsilon}{4}).$$

Após uma mudança de variáveis, chegamos em

$$\mathbb{E}\sup_{l\in\mathscr{L}}|\frac{1}{n}\sum_{i=1}^{n}l(X_i)-\mathbb{E}\,l(X)|\leq\frac{4}{\sqrt{n}}\,\mathbb{E}\int_0^{2}\sqrt{\log N(\mathscr{F},L^2(\mu_n),\varepsilon)}d\varepsilon.$$

Novamente usamos o Teorema 2.3.1 e o mesmo raciocínio do final da demonstração do Teorema 2.3.3, provando a existência de uma constante absoluta C tal que

$$\mathbb{E}\sup_{l\in\mathscr{L}}|\frac{1}{n}\sum_{i=1}^{n}l(X_i)-\mathbb{E}\,l(X)|\quad\leq\quad C\sqrt{\frac{\mathrm{vc}(\mathscr{F})}{n}}.\tag{2.13}$$

□

Lema 2.4.2. *(Número de cobertura de $\mathscr{F}$ e de $\mathscr{L}$)*

Seja T função Booleana e $\mathscr{F}$ uma classe de funções Booleanas. Seja $\mathscr{L}$ a família induzida por $\mathscr{F}$, $\mathscr{L} \coloneqq \left\{ (f-T)^2 \mid f \in \mathscr{F} \right\}$. Então

$$N(\mathscr{L}, L^2(\mu_n), \varepsilon) \leq N(\mathscr{F}, L^2(\mu_n), \frac{\varepsilon}{4}).$$

Demonstração. Seja $\{ f_j \}_{j=1}^N$ uma $\frac{\varepsilon}{4}-$cobertura de $\mathscr{F}$. Vamos mostrar que $\left\{ (f_j - T)^2 \right\}_{j=1}^N$ é uma $\varepsilon-$cobertura de $\mathscr{L}$. Já sabemos que $d(f,g) \coloneqq \|f-g\|_{L^2(\mu_n)} = \left(\frac{1}{n}\sum_{i=1}^n (fX_i - gX_i)^2\right)^{\frac{1}{2}}$, logo tomando $g \in \mathscr{L}$, com $g = (h-T)^2$ e $g \in \mathscr{F}$, considere $f_i \in \mathscr{F}$ tal que $d(f_i, h) \leq \frac{\varepsilon}{4}$. Vamos calcular $d(g, (f_i - T)^2)$:

$$\begin{aligned}
d(g,(f_i-T)^2) &= \left(\frac{1}{n}\sum_{i=1}^n (gX_i - (f_i-T)^2X_i)^2\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n}\sum_{i=1}^n ((h-T)^2X_i - (f_i-T)^2X_i)^2\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n}\sum_{i=1}^n ((hX_i - TX_i)^2 - (f_iX_i - TX_i)^2)^2\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n}\sum_{i=1}^n ((h^2X_i - 2hX_iTX_i + T^2X_i) - (f_i^2X_i - 2f_iX_iTX_i + T^2X_i))^2\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n}\sum_{i=1}^n (h^2X_i - f_i^2X_i - 2hX_iTX_i + 2f_iX_iTX_i)^2\right)^{\frac{1}{2}} \\
&= \|h^2 - f_i^2 - (2hT - 2f_iT)\|_{L^2(\mu_n)} \\
&\leq \|h^2 - f_i^2\|_{L^2(\mu_n)} + 2\|hT - f_iT\|_{L^2(\mu_n)} \text{ (desigualdade triangular)} \\
&\leq \|h - f_i\|_{L^2(\mu_n)} + 2\|h - f_i\|_{L^2(\mu_n)} \qquad (2.14) \\
&= 3d(f_i,h) \leq \frac{3\varepsilon}{4}, \text{ (por hipótese)} \qquad (2.15)
\end{aligned}$$

implicando $d(g,(f_i-T)^2) \leq \frac{3\varepsilon}{4} \leq \varepsilon$. Note que na desigualdade (2.14) usamos o fato de h e f_i serem Booleanas, implicando $h^2 = h$ e $f_i^2 = f_i$ e também $\|hT - f_iT\|_{L^2(\mu_n)} \leq \|h - f_i\|_{L^2(\mu_n)}$, (basta analisar os casos em que $T=0$ e $T=1$). □

Observação 2.4.1. *Podemos provar na verdade um fato muito melhor que o anterior:*

$$N(\mathscr{L}, L^2(\mu_n), \varepsilon) = N(\mathscr{F}, L^2(\mu_n), \varepsilon).$$

De fato, note que se f é Booleana, então $f = |f| = f^2$. Disto,

$$\begin{aligned}
d((h-T)^2,(f_i-T)^2) &= \|(h-T)^2 - (f_i-T)^2\|_{L^2(\mu_n)} = \||h-T| - |f_i-T|\|_{L^2(\mu_n)} \qquad (2.16) \\
&\leq \|\|h - T - f_i + T|\|_{L^2(\mu_n)} \\
&= \|\|h-f_i|\|_{L^2(\mu_n)} \\
&= \|h-f_i\|_{L^2(\mu_n)} = d(h,f_i).
\end{aligned}$$

Perceba que em (2.16) usamos $|a|-|b| \leq |a-b|$. Como a norma $L^2(\mu_n)$ é dada por uma integral, isto implica a desigualdade.

O que provamos com isto é que uma $\varepsilon-$cobertura $\{ f_j \}_{j=1}^N$ de $\mathscr{F}$ gera a $\varepsilon-$cobertura de $\mathscr{L}$, a saber, $\{ (f_j - T)^2 \}_{j=1}^N$. Portanto, os números de cobertura coincidem.

O Teorema 2.4.1 fornece justamente a quantidade que queríamos: o tamanho da amostra necessário para que tenhamos o excesso de risco controlado. É suficiente tomar uma amostra de tamanho proporcional a dimensão VC da classe de hipótese. De fato, dado $\varepsilon > 0$,

$$C\sqrt{\frac{\operatorname{vc}(\mathscr{F})}{n}} \leq \varepsilon \iff \varepsilon^{-2} \operatorname{vc}(\mathscr{F}) \leq n \cdot C.$$

Neste ponto, fica explícita a importância da família $\mathscr{F}$: se quisermos que o algoritmo de aprendizagem funcione minimamente bem, precisamos escolher o espaço de hipóteses com dimensão VC finita.

Uma vez encontrada a solução f_n^* de um certo problema de classificação, o Teorema 2.4.1 torna possível mensurar o quão confiável é a predição dada por f_n^*. Nestes casos, o risco é dado pela probabilidade de a solução classificar erroneamente.

Exemplo 2.4.2. *Considere o seguinte problema de classificação: estamos tentando obter a função $T\colon \mathbb{R}^2 \to \{ 0, 1 \}$ com os dados de treinamento $X_1, \ldots, X_n \in \Omega$, gerados de forma independente a partir de uma distribuição $\mathbb{P}$ em $\mathbb{R}^2$. Antes de escolher $\mathscr{F}$, pense que não podemos tomar uma classe tão grande (para evitar o sobreajuste) e nem tão pequena (para evitar o subajuste). Em outras palavras, estamos buscando uma classe de curvas não tão complicadas e nem tão triviais quanto uma reta. Escolher $\mathscr{F} = \{ \mathbb{1}_C \mid C$ é uma circunferência em $\mathbb{R}^2 \}$ parece razoável, portanto. Já sabemos do Exemplo 2.3.2 que $\operatorname{vc}(\mathscr{F}) = 3$.*

Apesar de ser possível calcular o risco empírico

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - T(X_i))^2$$

para qualquer função f definida no plano, estamos interessados no seu mínimo sobre $\mathscr{F}$. Considere que a solução desse problema de aprendizagem é

$$f_n^* = \arg\min_{f \in \mathscr{F}} R_n(f).$$

Calculando $f_n^*(x)$ para algum ponto $x \in \mathbb{R}^2$, saberemos a classificação de qualquer x que não esteja no conjunto dos pontos de treinamento. Para avaliar nossa predição, basta aplicar o Teorema 2.4.1 e teremos

$$\mathbb{E}\, R(f_n^*) \leq R(f^*) + C \cdot \sqrt{\frac{\operatorname{vc}(\mathscr{F})}{n}} = R(f^*) + C \cdot \sqrt{\frac{3}{n}}.$$

Concluímos que, em média, a solução encontrada f_n^* dá a classificação correta dentro de um erro de ordem $\frac{1}{\sqrt{n}}$. Ou seja, a probabilidade de f_n^* dar a classificação correta é quase a mesma da melhor função f^* (o melhor círculo) no espaço de hipótese.

2.5 Conclusão

Usando algumas técnicas de concentração de medida relacionadas na Subseção 2.1.1, vimos que a classe de processos subgaussianos é ampla e rica em propriedades. Por meio de dois conceitos determínisticos, o número de cobertura e a dimensão VC de um conjunto, foi possível analisar propriedades de um processo estocástico. Com o primeiro obtivemos uma cota superior para o

supremo de um processo subgaussiano $(X_t)_{t\in T}$ explícita na Integral de Dudley, Teorema 2.2.2. Podemos destacar o encadeamento como a ideia fundamental para a obtenção desse resultado. O segundo conceito, juntamente com a técnica de simetrização e a própria Integral de Dudley, foram os ingredientes principais para chegar ao Teorema 2.3.3, uma cota superior para o supremo de um processo empírico. Esse, por sua vez, foi o objeto matemático que permitiu investigar alguns problemas reais, como o exemplo dado em problemas de classificação na Subseção 2.4.1.

Apêndice

A Preliminares em Variáveis Aleatórias

Esta seção contém os principais resultados e definições de um curso de probabilidade introdutório. Para maiores detalhes e demonstrações, consulte Chung [4] e James [10].

A.1 Definições básicas

Fixado um espaço de probabilidade $(\Omega, \mathscr{F}, \mathbb{P})$, uma variável aleatória X é uma função mensurável definida num conjunto arbitrário Ω, $X\colon \Omega \to \mathbb{R}$. Vamos listar algumas quantidades e funções associadas a variável X :

Definição A.1. *(Função de distribuição ou distribuição acumulada) A função de distribuição da variável aleatória X, representada por F_X, é definida por*

$$F_X(x) = \mathbb{P}(X \leq x), x \in \mathbb{R}. \tag{17}$$

Definição A.2. *(Média, esperança ou valor esperado) Seja X uma variável aleatória e F sua função de distribuição. A esperança[17] de X é definida[18] por*

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x dF(x). \tag{18}$$

Definição A.3. *(Variância) Seja X uma variável aleatória e F sua função de distribuição. A variância de X é definida por*

$$\mathbb{E}(X - \mathbb{E}\,X)^2 = \int_{-\infty}^{+\infty} (x - \mathbb{E}\,X)^2 dF(x), \tag{19}$$

quando a integral existir.

Definição A.4. *(Covariância) Dadas duas variáveis aleatórias X e Y, a covariância é definida por*

$$\operatorname{cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y)\right],$$

quando o lado direito desta igualdade existir.

[17] Em alguns momentos, abolimos o uso de parênteses para não carregar a notação e simplesmente escrevemos $\mathbb{E}\,X$.

[18] Para que o valor esperado esteja bem definido, é necessário que o lado direito da equação convirja absolutamente. Um exemplo clássico em que isto não ocorre é a distribuição de Cauchy.

Definição A.5. *(Função Característica) A função característica de uma variável X com distribuição F é definida por*

$$\varphi_X(t) = \int_{-\infty}^{+\infty} \exp(ixt) dF(x) = \mathbb{E}\exp(itX), t \in \mathbb{R}.$$

Definição A.6. *(Função Geradora de Momentos) A função geradora de momentos de uma variável X é definida por*

$$M_X(t) = \mathbb{E}\exp(tX), \quad t \in \mathbb{R},$$

quando o lado direito desta igualdade existir.

Definição A.7. *(p-ésimo momento) Para $p > 0$, definimos, quando existirem as integrais, o p-ésimo momento por $\mathbb{E}\,X^p$ e o p-ésimo momento absoluto por $\mathbb{E}\,|X|^p$. Tomando a raiz p-ésima, obtemos a norma L^p:*

$$\|X\|_{L^p} = (\mathbb{E}\,|X|^p)^{\frac{1}{p}}.$$

Quando $p = \infty$ temos o supremo essencial de X :

$$\|X\|_{L^\infty} = \inf\{\, a \in \mathbb{R} \mid |\, X| \leq a \quad q.t.p\,\}$$

A.2 Desigualdades e teoremas clássicos

Teorema A.1. *(Desigualdade de Jensen) Seja X variável aleatória tal que $\mathbb{E}\,|X|, \mathbb{E}\,|\varphi(X)| < \infty$ e $\varphi\colon \mathbb{R} \to \mathbb{R}$ função convexa*[19]*. Então*

$$\varphi(\mathbb{E}\,X) \leq \mathbb{E}\,\varphi(X).$$

Teorema A.2. *(Desigualdade de Minkowski) Para $p \in [1, \infty]$ e variáveis aleatórias $X, Y \in L^p$,*

$$\|X + Y\|_{L^p} \leq \|X\|_{L^p} + \|Y\|_{L^p}.$$

Teorema A.3. *(Desigualdade de Hölder) Para $p, q \in [1, \infty]$ com $\frac{1}{p} + \frac{1}{q} = 1$ e $X \in L^p$ e $Y \in L^q$,*

$$|\,\mathbb{E}\,XY| \leq \|X\|_{L^p}\|Y\|_{L^q}$$

Teorema A.4. *Para toda variável aleatória X, vale*

$$\mathbb{E}\,X = \int_0^\infty \mathbb{P}\{X > t\}dt - \int_{-\infty}^0 \mathbb{P}\{X < t\}dt.$$

Teorema A.5. *Seja X variável aleatória e $p \in (0, \infty)$. Então*

$$\mathbb{E}\,|X|^p = \int_0^\infty pt^{p-1}\mathbb{P}\{X > t\}dt,$$

sempre que o lado direito da igualdade for finito.

Teorema A.6. *(Desigualdade de Markov) Seja X variável aleatória não-negativa e $t > 0$. Então*

$$\mathbb{P}\{X > t\} \leq \frac{\mathbb{E}\,X}{t}.$$

[19] Uma função φ é convexa quando $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$ para todo $\lambda \in [0,1]$ e para todos x e y no domínio de φ.

Teorema A.7. *(Desigualdade de Chebyshev) Sejam φ uma função crescente e estritamente positiva em $(0,\infty)$ com $\varphi(u) = \varphi(-u)$ e X variável aleatória tal que $\mathbb{E}\,\varphi(X) < \infty$. Então para todo $u > 0$, vale:*

$$\mathbb{P}(|X| \geq u) \leq \frac{\mathbb{E}\,\varphi(X)}{\varphi(u)}.$$

Teorema A.8. *(Teorema de Slutsky) Sejam X_n, X e Y variáveis ou vetores aleatórios. Se $X_n \to_D X$ e $Y_n \to_D c$ sendo c uma constante, então*

a) $X_n + Y_n \to_D X + c$;

b) $Y_n X_n \to_D cX$;

c) $\frac{X_n}{Y_n} \to_D \frac{X}{c}$, quando $c \neq 0$.

A.3 Teoremas Limite

Os próximos teoremas são resultados clássicos em Probabilidade. Aqui estão enunciados em suas versões mais básicas, para hipóteses mais gerais consulte James [10] e Chung [4].

Teorema A.9. *(Lei fraca dos grandes números) Sejam $X_1, X_2, \ldots$ variáveis aleatórias independentes, identicamente distribuídas e integráveis com média μ. Então*

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to_P \mu.$$

Teorema A.10. *(Lei forte dos grandes números) Sejam $X_1, X_2, \ldots$ variáveis aleatórias independentes, identicamente distribuídas e integráveis com média μ. Então*

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu \text{ quase certamente.}$$

Teorema A.11. *(Teorema Central do Limite) Sejam $X_1, X_2, \ldots$ variáveis aleatórias independentes, identicamente distribuídas com média μ e variância $\sigma^2 < \infty$. Se $S_n = X_1 + X_2 + \ldots + X_n$, então*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to_D \mathcal{N}(0,1).$$

Teorema A.12. *(Teorema Central do Limite - caso multivariado) Sejam $X_1, X_2, \ldots$ vetores aleatórios em $\mathbb{R}^k$ independentes, identicamente distribuídas com vetor de média μ e matriz de covariância Σ. Então,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \to_D N_k(0, \Sigma).$$

B Derivadas de funcionais

Esta seção contém definições e resultados necessários para estudar o comportamento local de estimadores perante a pertubações infinitesimais. Para tanto, definiremos a seguir diferentes conceitos de diferenciabilidade. O capítulo 20 em van der Vaart [14] contém todos detalhes aqui omitidos.

Definição B.1. *Uma aplicação $T\colon \mathbb{D} \mapsto \mathbb{E}$ entre espaços normados é dita Gateaux diferenciável em $F \in \mathbb{D}$ se para todo h fixo existir um elemento $T'_F(h) \in \mathbb{E}$ tal que*

$$T(F + th) - T(F) = tT'_F(h) + o(t), \quad t \downarrow 0,$$

ou, equivalentemente,

$$T'_F(h) = \lim_{t \to 0} \frac{T(F + th) - T(F)}{t}.$$

A diferenciabilidade de Gateaux é conhecida como derivada direcional e sua definição também inclui a linearidade e continuidade da aplicação $T'_F\colon \mathbb{D} \mapsto \mathbb{E}$.

Observação B.1. *Decorre diretamente da definição que a função de influência de um funcional T aplicado numa distribuição F no ponto x é a derivada de Gateaux de T em F na direção $\delta_x - F$:*

$$\mathrm{IF}(x, T, F) = T'_F(\delta_x - F).$$

Um conceito mais forte de diferenciabilidade, necessário para aplicar o Delta Método Funcional, é a derivada de Hadamard.

Definição B.2. *Uma aplicação $T\colon \mathbb{D}_T \subset \mathbb{D} \mapsto \mathbb{E}$ tal que $F \in \mathbb{D}_T$ é dita Hadamard diferenciável em F se existir uma aplicação linear e contínua, $T'_F\colon \mathbb{D} \mapsto \mathbb{E}$ tal que*

$$T(F + th_t) - T(F) = tT'_F(h) + o(t), \quad t \downarrow 0, \text{ para toda sequência } (h_t)_t \text{ tal que } h_t \to h, \quad (20)$$

ou, equivalentemente,

$$T'_F(h) = \lim_{t \to 0} \frac{T(F + th_t) - T(F)}{t}, \quad \text{para toda sequência } (h_t)_t \text{ tal que } h_t \to h.$$

Mais precisamente, para toda $h_t \to h$ tal que $F + th$ pertence ao domínio de T para $t > 0$ pequeno.

A diferenciabilidade de Hadamard é conhecida como diferenciabilidade compacta, já que sua definição é equivalente ao limite (20) com h em subconjuntos compactos de $\mathbb{D}$.

Note que o valor de ambas as derivadas é o mesmo, os conceitos diferem apenas no detalhe de que a Hadamard diferenciabilidade permite variar as direções h_t com t enquanto a derivada de Gateaux mantém a direção fixa.

A definição de Hadamard diferenciabilidade requer que $\phi'_\theta\colon \mathbb{D} \mapsto \mathbb{E}$ exista em todo o conjunto $\mathbb{D}$. Quando esse não é o caso e ϕ'_θ existe somente em $\mathbb{D}_0 \subset \mathbb{D}$, com as sequências $h_t \to h$ convergindo para limites $h \in \mathbb{D}_0$, então ϕ é dita Hadamard diferenciável tangencialmente a esse conjunto.

B.1 Delta Método

Considere T_n um estimador do parâmetro θ e ϕ uma função dada. Se estivermos interessados em $\phi(\theta)$, é razoável usarmos $\phi(T_n)$ como estimativa para $\phi(\theta)$. O Delta Método é uma técnica simples que almeja usar as propriedades assintóticas de T_n para obter as de $\phi(T_n)$, e com isso estudar a distribuição limite de $\sqrt{n}(T_n - \theta)$. Em resumo, usaremos a expansão de Taylor para aproximar um vetor aleatório da forma $\phi(T_n)$ por um polinômio. A versão para funcionais desta ferramenta, o Delta Método Funcional, é ingrediente imprescindível para a prova do teorema principal do Capítulo 1.

Teorema B.1 (Delta Método). *Sejam $\phi\colon \mathbb{D}_\phi \subset \mathbb{R}^k \mapsto \mathbb{R}^m$ diferenciável em θ e $T_n\colon \Omega_n \mapsto \mathbb{D}_\phi$ aplicações tomando valores no domínio de ϕ e tais que $r_n(T_n - \theta) \to_D T$ para alguma sequência de números $r_n \to \infty$. Então*

$$r_n(\phi(T_n) - \phi(\theta)) \to_D \phi'_\theta(T)$$

e, além disso, também vale

$$r_n(\phi(T_n) - \phi(\theta)) - \phi'_\theta(r_n(T_n - \theta)) = o_P(1).$$

Demonstração. Consulte Teorema 3.1 em van der Vaart [14]. □

Teorema B.2 (Delta Método Funcional). *Sejam $\mathbb{D}$ e $\mathbb{E}$ espaços normados e $\phi\colon \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ Hadamard diferenciável em θ tangencialmente a $\mathbb{D}_0$. Sejam $T_n\colon \Omega_n \mapsto \mathbb{D}_\phi$ aplicações tais que $r_n(T_n - \theta) \to_D T$ para alguma sequência de números $r_n \to \infty$ e um elemento aleatório T que toma valores em $\mathbb{D}_0$. Então*

$$r_n(\phi(T_n) - \phi(\theta)) \to_D \phi'_\theta(T)$$

e, se ϕ'_θ está definida e é contínua em todo o espaço $\mathbb{D}$, também vale

$$r_n(\phi(T_n) - \phi(\theta)) = \phi'_\theta(r_n(T_n - \theta)) + o_P(1).$$

$\mathbb{D}_0$ é o subconjunto de $\mathbb{D}$ onde ϕ'_θ existe.

Demonstração. Consulte Teorema 20.8 em van der Vaart [14]. □

Teorema B.3 (Regra da cadeia). *Sejam $\mathbb{D}$, $\mathbb{E}$ e $\mathbb{F}$ espaços normados e $\mathbb{D}_\phi \subset \mathbb{D}$ e $\mathbb{E}_\psi \subset \mathbb{E}$. Defina $\phi\colon \mathbb{D}_\phi \mapsto \mathbb{E}_\psi$ e $\psi\colon \mathbb{E}_\psi \mapsto \mathbb{F}$. Seja ϕ Hadamard diferenciável em θ tangencialmente a $\mathbb{D}_0$ e ψ Hadamard diferenciável em $\phi(\theta)$ tangencialmente a $\phi'_\theta(\mathbb{D}_0)$. Então $\psi \circ \phi\colon \mathbb{D}_\psi \mapsto \mathbb{F}$ é Hadamard diferenciável em θ tangencialmente a $\mathbb{D}_0$ com derivada*

$$\psi'_{\phi(\theta)} \circ \psi'_\theta.$$

Demonstração. Consulte Teorema 20.9 em van der Vaart [14]. □

C Distribuição normal Multivariada

Os resultados, definições e exemplos desta seção foram baseados no livro James [10].

C.1 Definições equivalentes

Um vetor aleatório n-dimensional $\mathbf{X} = (X_1, \ldots, X_n)^T$ é um vetor coluna de variáveis aleatórias. Quando estas são integráveis, definimos o vetor das médias

$$\mu_X = \mathbb{E}\,\mathbf{X} = (\mathbb{E}\,X_1, \ldots, \mathbb{E}\,X_n)^T.$$

Denote por $\mathbf{X}^T$ o transposto do vetor $\mathbf{X}$. Do mesmo modo, podemos definir o vetor linha das médias de um vetor aleatório.

Dados $\mathbf{X}$ e $\mathbf{Y}$ dois vetores aleatórios com $\mathbb{E}\,X_i^2 < \infty$, $i \in [n]$[20] e $\mathbb{E}\,Y_i^2 < \infty$, $i \in [m]$, definimos a matriz de covariância de $\mathbf{X}$ e $\mathbf{Y}$:

$$\begin{aligned}\Sigma_{\mathbf{XY}} \coloneqq \operatorname{cov}(\mathbf{X}, \mathbf{Y}) &= \mathbb{E}(\mathbf{X} - \mu_X)(\mathbf{Y} - \mu_Y)^T = \mathbb{E}(\mathbf{X} - \mu_X)(\mathbf{Y}^T - \mu_Y^T) = \\ &= \mathbb{E}[\mathbf{X}\mathbf{Y}^T - \mu_X\mathbf{Y}^T - \mathbf{X}\mu_Y^T + \mu_X\mu_Y^T] \\ &= \mathbb{E}[\mathbf{X}\mathbf{Y}^T - \mu_X\mu_Y^T].\end{aligned}$$

[20] Denotamos por $[n] = \{1, \ldots, n\}$.

Segue direto desta definição que a matriz de covariância é simétrica[21]. Quando $\mathbf{X} = \mathbf{Y}$, temos outra propriedade dada pelo teorema abaixo.

Teorema C.1. *A matriz de covariância* $\Sigma_{\boldsymbol{X}} = \operatorname{cov}(\boldsymbol{X}, \boldsymbol{X}) = \mathbb{E}\,\boldsymbol{X}\boldsymbol{X}^T - \mathbb{E}\,\boldsymbol{X}(\mathbb{E}\,\boldsymbol{X})^T$ *é não-negativa definida.*

Demonstração. Seja $\mathbf{v}$ um vetor n-dimensional arbitrário. Pelas propriedades[22] do produto interno, temos:

$$\begin{aligned}\mathbf{v}^T\Sigma_{\mathbf{X}}\mathbf{v} &= \mathbf{v}^T\left(\mathbb{E}\,\mathbf{X}\mathbf{X}^T - \mathbb{E}\,\mathbf{X}(\mathbb{E}\,\mathbf{X})^T\right)\mathbf{v} = \mathbb{E}\left(\mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}\right) - \mathbf{v}^T\,\mathbb{E}\,\mathbf{X}(\mathbb{E}\,\mathbf{X})^T\mathbf{v} \\ &= \mathbb{E}\left(\mathbf{v}^T\mathbf{X}\right)^2 - \left(\mathbf{v}^T\,\mathbb{E}\,\mathbf{X}\right)^2 = \operatorname{Var}\mathbf{v}^T\mathbf{X} \geq 0.\end{aligned}$$

□

Uma das ferramentas mais importantes e úteis relacionadas a uma variável aleatória é sua função característica, definida em A.5. Também podemos defini-la para um vetor aleatório $\mathbf{X}$ e um vetor $\mathbf{t}$ arbitrário, ambos de mesma dimensão:

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\exp i\mathbf{t}^T\mathbf{X}.$$

Assim como no caso unidimensional, a função característica de um vetor determina univocamente sua distribuição. Usando esse fato, obtemos o seguinte teorema que relaciona a função característica de um vetor com as funções características de suas coordenadas.

Teorema C.2. *Dado um vetor aleatório* $\boldsymbol{X} = (X_1, \ldots, X_n)$*, então as coordenadas* X_i *são independentes se, e somente se,*

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \prod_{i=1}^{n}\varphi_{X_i}(t_i).$$

Demonstração. Suponha que as coordenadas X_i sejam independentes. Pela definição,

$$\begin{aligned}\varphi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E}\exp(i\mathbf{t}^T\mathbf{X}) = \mathbb{E}\exp(i\sum_{i=1}^{n}X_it_i) = \mathbb{E}\prod_{i=1}^{n}\exp(iX_it_i) \\ &= \prod_{i=1}^{n}\mathbb{E}\exp(iX_it_i) = \prod_{i=1}^{n}\varphi_{X_i}(t_i).\end{aligned}$$

Por outro lado, suponha $\varphi_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^{n}\varphi_{X_i}(t_i)$. Sabemos pelo Teorema da Unicidade em James [10], p. 239, que a função característica de um vetor é única. Se as coordenadas de $\mathbf{X}$ não fossem independentes, então a função característica do vetor $\mathbf{X}$ seria diferente do produto, o que é um absurdo por hipótese. Logo, são independentes. □

O que daremos a seguir é a extensão vetorial da definição de uma variável aleatória normal. Já sabemos que em Probabilidade e Estatística esta distribuição é das mais importantes, e por isso, nada mais natural que entender como se dá a distribuição gaussiana em mais de uma dimensão.

Definição C.1. *Um vetor aleatório* $\boldsymbol{X}$ *tem distribuição normal multivariada não-degenerada se possui densidade conjunta:*

$$f_X(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n|\det V|}}\exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^TV^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right\} \tag{21}$$

para algum vetor $\boldsymbol{\mu}$ *e alguma matriz* V *positiva-definida.*

[21] Basta notar que suas entradas são dadas por $(\Sigma_{\mathbf{XY}})_{ij} = \operatorname{cov}(X_i, Y_j) = \operatorname{cov}(Y_j, X_i) = (\Sigma_{\mathbf{XY}})_{ji}$.
[22] Dados dois vetores u e v, $\langle u, v\rangle = u^Tv = v^Tu$, logo $\langle u, v\rangle^2 = u^Tvv^Tu$.

O teorema seguinte mostra como esta definição pode ser substituída por outras equivalentes.

Teorema C.3. *São equivalentes:*

a) $\boldsymbol{X}$ *tem distribuição normal multivariada não-degenerada.*

b) Existem matriz D *e um vetor* $\boldsymbol{\mu}$ *tais que*

$$\boldsymbol{X} = D\,\boldsymbol{W} + \boldsymbol{\mu}, \tag{22}$$

sendo $\boldsymbol{W}$ *um vetor com coordenadas independentes normais padrão.*

c) Seja $\boldsymbol{X}$ *vetor aleatório de média* $\boldsymbol{\mu}$ *e matriz de covariância* Σ*. A variável aleatória unidimensional* $\boldsymbol{a}^T\boldsymbol{X}$ *tem distribuição normal, isto é,*

$$\boldsymbol{a}^T\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\Sigma\boldsymbol{a}), \text{ para todo vetor } \boldsymbol{a}. \tag{23}$$

d) A função característica de $\boldsymbol{X}$ *é dada por:*

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}\right), \tag{24}$$

sendo $\boldsymbol{\mu}$ *o vetor média de* $\boldsymbol{X}$ *e* Σ *sua matriz de covariância.*

Um fato interessante (ou inusitado) é que de acordo com a Definição 22, o vetor $\mathbf{X}$ identicamente nulo tem distribuição normal multivariada, bastando pegar a matriz D com todas as entradas iguais a zero. O caso unidimensional, isto é, a variável aleatória $X \equiv 0$, será portanto, considerada uma normal degenerada.

Demonstração. *b)* $\Leftrightarrow$ *c)* Seja $\mathbf{X}$ vetor aleatório satisfazendo (22). Para qualquer vetor $\mathbf{a}$,

$$\mathbf{a}^T\mathbf{X} = \mathbf{a}^T(D\mathbf{W} + \mu) = \mathbf{a}^T(D\mathbf{W}) + \mathbf{a}^T\mu = (D^T\mathbf{a})^T\mathbf{W} + \mathbf{a}^T\mu.$$

Concluímos que $\mathbf{a}^T\mathbf{X}$ é uma combinação linear de normais independentes[23], logo uma variável aleatória normal pelo Teorema C.5.

Agora suponha que o vetor $\mathbf{X}$ cumpra (23). Vamos mostrar que $\mathbf{X} = D\mathbf{W} + \mu$ para alguma matriz D e algum vetor real $\boldsymbol{\mu}$, sendo $\mathbf{W} = (W_1, W_2, \ldots, W_n)$ e $W_i \sim \mathcal{N}(0,1)$ independentes. Seja $\Sigma_{\mathbf{X}}$ a matriz de covariância de $\mathbf{X}$, $\boldsymbol{\mu} = \mathbb{E}\,\mathbf{X}$ e $D^2 = \Sigma_{\mathbf{X}} = V$. Supondo a existência de V^{-1}, tome o vetor $\mathbf{W} = D^{-1}(\mathbf{X} - \boldsymbol{\mu})$. A existência de V^{-1} garante a existência de D^{-1}. Para qualquer vetor $\mathbf{s}$, $\mathbf{s}^T\mathbf{W}$ é uma combinação linear de $\mathbf{X}$ somada a uma constante, portanto normal por hipótese. De fato,

$$\mathbf{s}^T\mathbf{W} = \mathbf{s}^T D^{-1}\mathbf{X} - \mathbf{s}^T D^{-1}\boldsymbol{\mu} = (D^{-1^T}\mathbf{s})^T\mathbf{X} - (D^{-1^T}\mathbf{s})^T\boldsymbol{\mu}.$$

Além disso, $\mathbb{E}\,\mathbf{W} = D^{-1}(\mathbb{E}\,\mathbf{X} - \boldsymbol{\mu}) = 0$ implicando

$$\begin{aligned}
\operatorname{cov}(\mathbf{W},\mathbf{W}) &= \mathbb{E}\,\mathbf{W}\mathbf{W}^T - \mathbb{E}\,\mathbf{W}(\mathbb{E}\,\mathbf{W})^T = \mathbb{E}\,\mathbf{W}\mathbf{W}^T \\
&= \mathbb{E}(D^{-1}(\mathbf{X}-\boldsymbol{\mu})(D^{-1}(\mathbf{X}-\mu))^T = \mathbb{E}\,D^{-1}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T D^{-1^T} \\
&= D^{-1^2}\,\mathbb{E}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T \\
&= D^{-1^2}\Sigma_X = D^{2^{-1}}D^2 = I_n.
\end{aligned}$$

[23] Somar uma constante a uma variável aleatória apenas multiplica sua distribuição por uma constante (pense na função característica).

O fato de $\mathbb{E}\,\mathbf{W} = 0$ também implica $\mathbb{E}\,\mathbf{s}^T\mathbf{W} = 0$ para todo vetor $\mathbf{s}$ pela linearidade do valor esperado, resultando em

$$\operatorname{Var}\mathbf{s}^T\mathbf{W} = \mathbb{E}(\mathbf{s}^T\mathbf{W})^2 - (\mathbb{E}\,\mathbf{s}^T\mathbf{W})^2 = \mathbb{E}\,\mathbf{s}^T\mathbf{W}\mathbf{W}^T\mathbf{s} = \mathbf{s}^T\mathbf{s}\,\mathbb{E}\,\mathbf{W}\mathbf{W}^T = \mathbf{s}^T\mathbf{s}.$$

Portanto, a função característica da variável $\mathbf{s}^T\mathbf{W}$ é dada por:

$$\varphi_{\mathbf{s}^T\mathbf{W}}(t) = \exp(-\frac{t^2\mathbf{s}^T\mathbf{s}}{2}). \tag{25}$$

Mas $\varphi_{\mathbf{s}^T\mathbf{W}}(1) = \varphi_{\mathbf{W}}(s)$, donde concluímos que $\mathbf{W}$ é um vetor normal multivariado com média zero, matriz de covariância I_n identidade e de coordenadas independentes. Para verificar a independência das coordenadas, use o Teorema C.2.

Agora, vejamos o caso em que não existe V^{-1}, isto é, V é singular. Assuma que $\boldsymbol{\mu} = 0$. Como V age em espaços de mesma dimensão ($\mathbb{R}^n$), a não existência da inversa de V implica a não injetividade (e não sobrejetividade) de V. Logo, existe algum $\mathbf{a} \neq 0$ tal que $V\mathbf{a} = 0$, o que resulta em $\mathbf{a}^T V\mathbf{a} = 0$. Note que

$$\mathbf{a}^T V\mathbf{a} = \mathbf{a}^T\Sigma_{\mathbf{X}}\mathbf{a} = \mathbf{a}^T\,\mathbb{E}\,\mathbf{X}\mathbf{X}^T\mathbf{a} = \mathbb{E}\,\mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{a} = \mathbb{E}\,\mathbf{a}^T\mathbf{X}\mathbf{a}^T\mathbf{X} = \mathbb{E}(\mathbf{a}^T\mathbf{X})^2, \tag{26}$$

o que significa que $\mathbf{a}^T\mathbf{X} = 0$ com probabilidade 1[24]. Consequentemente, alguma componente do vetor $\mathbf{X}$ é uma combinação determinística linear das componentes restantes. Podemos assumir que X_n é combinação linear de $(X_1, \ldots, X_{n-1})$, a menos de possível rearranjo das componentes de $\mathbf{X}$. Se a matriz de covariância de $(X_1, \ldots, X_{n-1})$ também não é invertível, repetimos o argumento até que eventualmente uma matriz não-degenerada seja obtida. Neste ponto, o vetor $\mathbf{X}$ é tal que $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ sendo $\Sigma_{\mathbf{Y}}$ não-negativa definida e $\mathbf{Z} = A\mathbf{Y}$ para alguma matriz A com probabilidade 1. O vetor $\mathbf{Y}$ satisfaz a (22) pelo que já foi provado acima e também a definição (23) pois escolhendo o vetor $\mathbf{a} = (1, 0)$, temos por hipótese que $\mathbf{a}^T\mathbf{X} = \mathbf{Y}$ tem distribuição normal, logo qualquer transformação linear de $\mathbf{Y}$ também terá distribuição normal. Se $\mathbf{Y}$ é um vetor de dimensão k, considere D a matriz $k \times k$ tal que $\mathbf{Y} = D\mathbf{W}$ e $\overline{\mathbf{W}}$ um vetor de $n - k$ normais padrão independentes. Por fim, podemos escrever:

$$\mathbf{X} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix} = \begin{bmatrix} D & 0 \\ AD & 0 \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ \overline{\mathbf{W}} \end{bmatrix}, \tag{27}$$

demonstrando que $\mathbf{X}$ satisfaz a definição (22). Caso uma matriz não singular nunca seja obtida nesse processo, teremos $\mathbf{X} = 0$ o que também satisfaz a definição (22) com $D = 0$. Esse é o caso de vetor normal multivariado mais degenerado possível.

b) $\Leftrightarrow$ *a*) Seja $\mathbf{X}$ vetor aleatório satisfazendo (22). Sabemos que a independência das variáveis implica que sua densidade conjunta é dada pelo produto das densidades marginais:

$$\begin{aligned} f_{\mathbf{W}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x_i^2}{2}) &= \frac{1}{\sqrt{(2\pi)^n}} \exp(-\frac{1}{2}\sum_{i=1}^{n} x_i^2) \\ &= \frac{1}{\sqrt{(2\pi)^n}} \exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x}) \end{aligned}$$

Supondo a existência de D^{-1}[25], podemos usar o método jacobiano[26] para obter a densidade de

[24] Use a propriedade da integral de Lebesgue.

[25] Lembre-se que ${D^{-1}}^T = {D^T}^{-1}$ e $(DD^T)^{-1} = {D^{-1}}^T D^{-1}$.

[26] Consulte James [10] para maiores detalhes.

$\mathbf{X}$:

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{W}}\left(D^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)|\det D^{-1}| = \frac{1}{\sqrt{(2\pi)^n|\det D|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T {D^{-1}}^T D^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$
$$= \frac{1}{\sqrt{(2\pi)^n|\det V|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T (DD^T)^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

No último passo usamos $V = DD^T$, o que implica

$$\sqrt{|\det V|} = \sqrt{|\det DD^T|} = \sqrt{|\det D \det D^T|} = \sqrt{|\det D|^2} = |\det D|.$$

Caso não exista a inversa da matriz D, não existe densidade conjunta.

Vamos mostrar a outra direção $a) \Rightarrow b)$. Como V é positiva definida, existe[27] matriz D também simétrica positiva tal que $D^2 = V$. A matriz D é, portanto, invertível. De fato, que $\det V = \det D^2 = (\det D)^2 > 0$. Então tome o vetor $\mathbf{W} = D^{-1}(\mathbf{X}-\boldsymbol{\mu})$. Já mostramos que $\mathbb{E}\,\mathbf{W} = 0$ e $\operatorname{cov}(\mathbf{W},\mathbf{W}) = I_n$. Note que cada coordenada W_i do vetor $\mathbf{W}$ é uma combinação linear das variáveis normais X_i:

$$W_i = \mathbf{e}_i^T\mathbf{W} = \mathbf{e}_i^T D^{-1}(\mathbf{X}-\boldsymbol{\mu}) = \mathbf{e}_i^T D^{-1}\mathbf{X} - \mathbf{e}_i^T\boldsymbol{\mu} = ({D^{-1}}^T\mathbf{e}_i)^T\mathbf{X} - \mathbf{e}_i^T\mu, \tag{28}$$

sendo $\mathbf{e}_i$ o vetor canônico correspondente. Para aplicar o Teorema C.6 e concluir que W_i tem distribuição normal, falta ver que X_i são normais unidimensionais. Para tanto, segue o cálculo da densidade marginal, onde usamos o fato de a matriz V ser diagonal e denotamos por λ_i seus autovalores. Note que podemos transformar a integral múltipla em $\mathbb{R}^{n-1}$ em $n-1$ integrais iteradas pois o integrando é C^∞ (todas as derivadas parciais existem e são contínuas, logo podemos integrar em qualquer ordem). A densidade resultante é de uma variável normal:

$$\begin{aligned}
f_{X_i}(x_i) &= \int_{\mathbb{R}^{n-1}} \frac{1}{\sqrt{(2\pi)^n|\det V|}}\exp\Big\{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T V^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\Big\}dx_1\cdots\hat{dx_i}\cdots dx_n \\
&= \int_{\mathbb{R}^{n-1}} \frac{1}{\sqrt{(2\pi)^n|\det V|}}\exp\Big\{-\frac{\sum_{k,j}V_{kj}(x_k-\boldsymbol{\mu}_k)(x_j-\boldsymbol{\mu}_j)}{2}\Big\}dx_1\cdots\hat{dx_i}\cdots dx_n \\
&= \frac{1}{\sqrt{(2\pi)\lambda_i}}\exp\Big\{-\frac{V_{ii}(x_i-\boldsymbol{\mu}_i)^2}{2}\Big\}\frac{1}{\sqrt{(2\pi)|\lambda_1|}}\int_{\mathbb{R}} exp\Big\{-\frac{V_{11}(x_1-\boldsymbol{\mu}_1)^2}{2}\Big\}dx_1\cdots \\
&\cdots\frac{1}{\sqrt{(2\pi)|\lambda_n|}}\int_{\mathbb{R}} exp\Big\{-\frac{V_{nn}(x_n-\boldsymbol{\mu}_n)^2}{2}\Big\}dx_n \\
&= \frac{1}{\sqrt{(2\pi)\lambda_i}}\exp\Big\{-\frac{V_{ii}(x_i-\boldsymbol{\mu}_i)^2}{2}\Big\}.
\end{aligned}$$

Resta mostrar a independência das coordenadas do vetor $\mathbf{W}$. Vamos calcular a densidade conjunta de $\mathbf{W}$ usando o método jacobiano, donde segue que $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{W}}(D^{-1}(\mathbf{x}-\boldsymbol{\mu}))|\det D^{-1}|$ e portanto,

$$\frac{1}{|\det D^{-1}|}f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{W}}\left(D^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Denotando $y = D^{-1}(\mathbf{x}-\boldsymbol{\mu})$,

$$f_{\mathbf{W}}(\mathbf{y}) = |\det D| f_{\mathbf{X}}\left(D\mathbf{y}+\boldsymbol{\mu}\right),$$

[27] Consulte o Teorema 13.8 do livro *Álgebra Linear* de Elon Lages Lima.

logo

$$
\begin{aligned}
f_{\mathbf{W}}(\mathbf{y}) &= |\det D| \frac{1}{\sqrt{(2\pi)^n |\det V|}} \exp\left\{ -\frac{(D\mathbf{y}+\boldsymbol{\mu}-\boldsymbol{\mu})^T V^{-1}(D\mathbf{y}+\boldsymbol{\mu}-\boldsymbol{\mu})}{2} \right\} \\
&= |\det D| \frac{1}{\sqrt{(2\pi)^n |(\det D)^2|}} \exp\left\{ -\frac{(D\mathbf{y})^T D^{2^{-1}}(D\mathbf{y})}{2} \right\} \\
&= \frac{1}{\sqrt{(2\pi)^n}} \exp\left\{ -\frac{(\mathbf{y}^T D^T D^{-1^2}(D\mathbf{y})}{2} \right\} \\
&= \frac{1}{\sqrt{(2\pi)^n}} \exp\left\{ -\frac{\mathbf{y}^T\mathbf{y}}{2} \right\} \\
&= \frac{1}{\sqrt{(2\pi)}} \exp\left\{ -\frac{y_1^2}{2} \right\} \cdots \frac{1}{\sqrt{(2\pi)}} \exp\left\{ -\frac{y_n^2}{2} \right\}.
\end{aligned}
$$

que é o produto de densidades de variáveis normais padrão. Portanto, as coordenas de $\mathbf{W}$ são independentes.

b) $\Leftrightarrow$ *d*) Seja $\mathbf{X}$ vetor aleatório satisfazendo (22). Vamos calcular sua função característica usando o Teorema C.2:

$$
\begin{aligned}
\varphi_{\mathbf{X}}(\mathbf{t}) &= \mathbb{E} \exp\left(i\mathbf{t}^T(D\mathbf{W}+\boldsymbol{\mu})\right) = \exp(i\mathbf{t}^T\boldsymbol{\mu})\, \mathbb{E} \exp(i\mathbf{t}^T D\mathbf{W}) \\
&= \exp(i\mathbf{t}^T\boldsymbol{\mu}) \varphi_{\mathbf{W}}(D^T\mathbf{t}) = \exp(i\mathbf{t}^T\boldsymbol{\mu}) \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(D^T\mathbf{t})_i)^2 \right) \\
&= \exp(i\mathbf{t}^T\boldsymbol{\mu}) \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (D^T\mathbf{t})_i)^2 \right) \\
&= \exp(i\mathbf{t}^T\boldsymbol{\mu}) \exp\left(-\frac{1}{2} (D^T\mathbf{t})^T (D^T\mathbf{t}) \right) \\
&= \exp\left(i\mathbf{t}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T DD^T\mathbf{t} \right).
\end{aligned}
$$

Resta mostrar que $\mathbb{E}\,\mathbf{X} = \boldsymbol{\mu}$ e $\Sigma = DD^T$. O primeiro segue direto da linearidade da esperança e de $\mathbb{E}\,\mathbf{W} = \mathbf{0}$. Para o cálculo da matriz de covariância, lembre-se que já calculamos $\mathbb{E}\,\mathbf{W}\mathbf{W}^T = Id$:

$$
\begin{aligned}
\operatorname{cov}(\mathbf{X},\mathbf{X}) &= \mathbb{E}[(D\mathbf{W}+\boldsymbol{\mu})(D\mathbf{W}+\boldsymbol{\mu})^T - \boldsymbol{\mu}\boldsymbol{\mu}^T] \\
&= \mathbb{E}[D\mathbf{W}\mathbf{W}^T D^T + D\mathbf{W}\boldsymbol{\mu}^T + \boldsymbol{\mu}\mathbf{W}^T D^T + \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T] \\
&= D\,\mathbb{E}\,\mathbf{W}\mathbf{W}^T D^T + D\,\mathbb{E}\,\mathbf{W}\boldsymbol{\mu}^T + \boldsymbol{\mu}\,\mathbb{E}\,\mathbf{W}^T D^T = D\,\mathbb{E}\,\mathbf{W}\mathbf{W}^T D^T = DD^T.
\end{aligned}
$$

Por outro lado, considere $\mathbf{X}$ um vetor aleatório satisfazendo (24). Considere a diagonalização da matriz Σ $\Sigma = UDU^{-1}$ sendo U matriz ortogonal e D a matriz diagonal com os autovalores de Σ. Defina o vetor $\mathbf{W} = A^{-1}(\mathbf{X}-\boldsymbol{\mu})$ com $A = U\sqrt{D}U^{-1}$. Note que $A^2 = (U\sqrt{D}U^{-1})(U\sqrt{D}U^{-1}) =$

$UDU^{-1} = \Sigma$. Isto implica, usando algumas propriedades de matrizes, [28] o que segue abaixo:

$$\begin{aligned}
\mathbb{E}\,\mathbf{W} &= \mathbb{E}[A^{-1}(\mathbf{X}-\boldsymbol{\mu})] = A^{-1}(\mathbb{E}\,\mathbf{X}-\boldsymbol{\mu}) = 0\\
\operatorname{cov}(\mathbf{W},\mathbf{W}) &= \mathbb{E}\,A^{-1}(\mathbf{X}-\boldsymbol{\mu})[A^{-1}(\mathbf{X}-\boldsymbol{\mu})]^T\\
&= \mathbb{E}\,A^{-1}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T A^{-1^T} = A^{-1}A^{-1^T}\Sigma = A^{-1}A^{T^{-1}}\Sigma\\
&= A^{-1}A^{-1}\Sigma = A^{2^{-1}}\Sigma = Id.
\end{aligned}$$

Para ver que $\mathbf{W}$ é um vetor de coordenadas independentes normais padrão, usaremos o Teorema C.2 e $A^T = A$. Temos:

$$\begin{aligned}
\varphi_{\mathbf{W}}(\mathbf{t}) &= \mathbb{E}\exp(i\mathbf{t}^T A^{-1}(\mathbf{X}-\boldsymbol{\mu}))\\
&= \exp(-i\mathbf{t}^T A^{-1}\boldsymbol{\mu})\,\mathbb{E}\exp(i(A^{-1^T}\mathbf{t})^T\mathbf{X})\\
&= \exp(-i\mathbf{t}^T A^{-1}\boldsymbol{\mu})\,\mathbb{E}\exp(i(A^{-1}\mathbf{t})^T\mathbf{X})\\
&= \exp(-i\mathbf{t}^T A^{-1}\boldsymbol{\mu})\varphi_{\mathbf{X}}(A^{-1}\mathbf{t})\\
&= \exp(-i\mathbf{t}^T A^{-1}\boldsymbol{\mu})\exp\left(i(A^{-1}\mathbf{t})^T\boldsymbol{\mu} - \frac{1}{2}(A^{-1}\mathbf{t})^T\Sigma A^{-1}\mathbf{t}\right)\\
&= \exp\left(-i(A^{-1}\mathbf{t})^T\boldsymbol{\mu} + i(A^{-1}\mathbf{t})^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T A^{-1}A\mathbf{t}\right)\\
&= \exp\left(-\frac{1}{2}\mathbf{t}^T\mathbf{t}\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^{n} t_i^2\right) = \prod_{i=1}^{n}\exp(-\frac{t_i^2}{2})
\end{aligned}$$

c) $\Leftrightarrow$ *d*) Seja $\mathbf{X}$ um vetor aleatório satisfazendo (23). Calculando a função característica da variável $\mathbf{a}^T\mathbf{X}$ no ponto $t = 1$:

$$\varphi_{\mathbf{a}^T\mathbf{X}}(1) = \mathbb{E}\exp(i\mathbf{a}^T\mathbf{X}).$$

Sabendo que $\mathbf{a}^T\mathbf{X}$ é normal, a equação acima é igual a $\exp\left(i\boldsymbol{\mu} - \frac{1}{2}\sigma^2\right)$, sendo μ e σ^2 respectivamente a média e a variância de $\mathbf{a}^T\mathbf{X}$. Denote por $\mu_{\mathbf{X}}$ o vetor $\mathbb{E}\,\mathbf{X} = (\mu_{X_1},\ldots,\mu_{X_n})$, $\sigma_i^2 = \operatorname{Var} X_i$ e Σ_X a matriz de covariância de $\mathbf{X}$. Pelos mesmos cálculos do Teorema C.6, temos $\mu = \mathbf{a}^T\mu_{\mathbf{X}}$ e $\sigma^2 = \mathbf{a}^T\Sigma_{\mathbf{X}}\mathbf{a}$. Portanto,

$$\begin{aligned}
\varphi_{\mathbf{X}}(\mathbf{a}) = \mathbb{E}\exp(i\mathbf{a}^T\mathbf{X}) = \varphi_{\mathbf{a}^T\mathbf{X}}(1) &= \exp\left(i\mu - \frac{1}{2}\sigma^2\right)\\
&= \exp\left(i\mathbf{a}^T\mu_{\mathbf{X}} - \frac{1}{2}\mathbf{a}^T\Sigma_{\mathbf{X}}\mathbf{a}\right),
\end{aligned}$$

como queríamos demonstrar. Para provar a outra direção, considere um vetor real $\mathbf{a}$:

$$\varphi_{\mathbf{a}^T\mathbf{X}}(t) = \mathbb{E}\exp(it\mathbf{a}^T\mathbf{X}) = \mathbb{E}\exp(i(t\mathbf{a})^T\mathbf{X}) = \varphi_{\mathbf{X}}(t\mathbf{a}).$$

Supondo que $\mathbf{X}$ tenha a função característica $\varphi_{\mathbf{X}}(\mathbf{a}) = \exp\left(i\mathbf{a}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{a}^T\Sigma\mathbf{a}\right)$ sendo $\boldsymbol{\mu}$ o vetor da média de $\mathbf{X}$ e Σ sua matriz de covariância, segue:

$$\varphi_{\mathbf{a}^T\mathbf{X}}(t) = \varphi_{\mathbf{X}}(t\mathbf{a}) = \exp\left(it\mathbf{a}^T\boldsymbol{\mu} - \frac{1}{2}t\mathbf{a}^T\Sigma t\mathbf{a}\right) = \exp\left(it\mathbf{a}^T\boldsymbol{\mu} - \frac{1}{2}t^2\mathbf{a}^T\Sigma\mathbf{a}\right),$$

a função característica de uma variável normal com média $\mathbf{a}^T\boldsymbol{\mu}$ e variância $\mathbf{a}^T\Sigma\mathbf{a}$.

[28] $A = U\sqrt{D}U^{-1} \Rightarrow A^T = U^{-1^T}\sqrt{D}^T U^T = U\sqrt{D}U^{-1} = A$.

$a) \Leftrightarrow d)$ Esta equivalência é demonstrada mais facilmente usando a Definição 22, isto é, fazendo $a) \Rightarrow b)$ seguido de $b) \Rightarrow d)$ e posteriormente, $(4) \Rightarrow b)$ e $b) \Rightarrow a)$. Lidar com a expressão da densidade torna as contas intragáveis.

$c) \Leftrightarrow a)$ Seja $\mathbf{X}$ vetor aleatório satisfazendo a Definição C.1. Vamos calcular a densidade marginal da variável X_i para obter a distribuição de $\mathbf{a}^T\mathbf{X}$. Note que o fato da matriz V ser positiva definida implica $|\det V| = \prod_{i=1}^n \lambda_i$, sendo λ_i os autovalores de V. Como já calculamos as marginas na prova da implicação $b) \Leftrightarrow a)$, segue

$$f_{X_i}(x_i) = \frac{1}{\sqrt{(2\pi)\lambda_i}} \exp\Big\{ -\frac{V_{ii}(x_i-\mu_i)^2}{2} \Big\}.$$

Agora, basta usar o Teorema C.6 para concluirmos que $\mathbf{a}^T\mathbf{X}$ tem distribuição normal.

Para a outra direção, faça $c) \Rightarrow b)$ e depois $b) \Rightarrow a)$. □

C.2 Teoremas úteis

Teorema C.4. *Sejam $\boldsymbol{X} = (X_1, \ldots, X_n)$ um vetor normal multivariado. Então as coordenadas de $\boldsymbol{X}$ são independentes se, e somente se, a matriz de covariância de $\boldsymbol{X}$ é diagonal, isto é, se $\operatorname{cov}(X_i, X_j) = 0$ para todos os índices i, j.*

Demonstração. Se as coordenadas são independentes, então é direto que $\operatorname{cov}(X_i, X_j) = 0$ para todos os índices i, j. Suponha então que a matriz de covariância de $\mathbf{X}$ é diagonal. Pela definição (24),

$$\begin{aligned}\varphi_{\mathbf{X}}(\mathbf{t}) &= \exp\left(i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right) = \exp\left(i\sum t_i\mu_i - \frac{1}{2}\sum t_i^2\sigma_i^2\right)\\ &= \prod \exp\left(it_i\mu_i - \frac{1}{2}t_i^2\sigma_i^2\right) = \varphi_{X_i}(t_i).\end{aligned}$$

Pelo Teorema C.2, as coordenadas de $\mathbf{X}$ são independentes. □

Teorema C.5. *Sejam $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ n variáveis aleatórias independentes e $\boldsymbol{t} \in \mathbb{R}^n$. Então $\sum_{i=1}^n t_iX_i \sim \mathcal{N}(\sum_{i=1}^n t_i\mu_i, \sum_{i=1}^n t_i^2\sigma_i^2)$.*

Demonstração. Vamos calcular a função característica da variável $\sum t_iX_i$:

$$\begin{aligned}\varphi_{\sum t_iX_i}(z) &= \mathbb{E}\exp\left(iz\sum t_iX_i\right) = \mathbb{E}\prod_{i=1}^n \exp(izt_iX_i) = \prod_{i=1}^n \mathbb{E}\exp(izt_iX_i)\\ &= \prod_{i=1}^n \varphi_{X_i}(zt_i) = \prod_{i=1}^n \exp\left(i\mu_i zt_i - \frac{1}{2}(zt_i)^2\sigma_i^2\right)\\ &= \exp\left(iz\sum_{i=1}^n \mu_it_i - \frac{1}{2}z^2\sum_{i=1}^n t_i^2\sigma_i^2\right).\end{aligned}$$

Esta é a função característica de uma variável normal com média $\sum_{i=1}^n t_i\mu_i$ e variância $\sum_{i=1}^n t_i^2\sigma_i^2$, como queríamos demonstrar. □

Um resultado parecido e igualmente útil é quando as coordenadas X_i formam um vetor normal multivariado.

Teorema C.6. *Seja* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *um vetor normal multivariado com* $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. *Então* $\sum_{i=1}^n t_i X_i \sim N(\sum_{i=1}^n t_i\mu_i, \boldsymbol{t}^T\Sigma\boldsymbol{t})$.

Demonstração. Usando a definição (23) de vetor normal multivariado, o resultado é direto pois $\sum t_i X_i = \mathbf{t}^T\mathbf{X}$. Resta calcular a média, obtida facilmente:

$$\mathbb{E}\,\mathbf{t}^T\mathbf{X} = \mathbf{t}^T\,\mathbb{E}\,\mathbf{X} = \sum t_i\mu_i.$$

A variância pode ser calculada diretamente[29]:

$$\begin{aligned}\operatorname{Var}\mathbf{t}^T\mathbf{X} &= \mathbb{E}(\mathbf{t}^T\mathbf{X})^2 - (\mathbb{E}\,\mathbf{t}^T\mathbf{X})^2\\ &= \mathbb{E}(\mathbf{t}^T\mathbf{X}\mathbf{X}^T\mathbf{t}) - (\mathbf{t}^T\,\mathbb{E}\,\mathbf{X})^2 = \mathbf{t}^T\,\mathbb{E}\,\mathbf{X}\mathbf{X}^T\mathbf{t} - \mathbf{t}^T\,\mathbb{E}\,\mathbf{X}(\mathbb{E}\,\mathbf{X})^T\mathbf{t}\\ &= \mathbf{t}^T\left(\mathbb{E}\,\mathbf{X}\mathbf{X}^T - \mathbb{E}\,\mathbf{X}\,\mathbb{E}\,\mathbf{X}^T\right)\mathbf{t} = \mathbf{t}^T\Sigma_{\mathbf{X}}\mathbf{t},\end{aligned}$$

ou podemos usar a função característica:

$$\begin{aligned}\varphi_{\sum t_iX_i}(u) = \mathbb{E}(iu\sum t_iX_i) = \varphi_{\mathbf{X}}(u\mathbf{t}) &= \exp\left(i(u\mathbf{t})^T\mu - \frac{1}{2}(u\mathbf{t})^T\Sigma u\mathbf{t}\right)\\ &= \exp\left(iu\mathbf{t}^T\mu - \frac{1}{2}u^2\mathbf{t}^T\Sigma\mathbf{t}\right)\end{aligned}$$

que é a função característica de uma variável normal com média $\mathbf{t}^T\mu = \sum_{i=1}^n t_i\mu_i$ e variância $\mathbf{t}^T\Sigma\mathbf{t}$.

Quando $n = 2$, temos:

$$\begin{aligned}\mathbf{t}^T\Sigma_{\mathbf{X}}\mathbf{t} &= \sum_{i,j=1}^{2} t_i\operatorname{cov}(X_i, X_j)t_j = \sum_{j=1}^{2} t_1\operatorname{cov}(X_1, X_j)t_j + t_2\operatorname{cov}(X_2, X_j)t_j\\ &= t_1\operatorname{cov}(X_1, X_1)t_1 + t_2\operatorname{cov}(X_2, X_1)t_1 + t_1\operatorname{cov}(X_1, X_2)t_2 + t_2\operatorname{cov}(X_2, X_2)t_2\\ &= t_1^2\operatorname{Var}X_1 + 2t_1t_2\operatorname{cov}(X_1, X_2) + t_2^2\operatorname{Var}X_2. \qquad (29)\end{aligned}$$

□

Observação C.1. *O Teorema C.5 poderia ter sido obtido do Teorema C.6, já que um vetor de variáveis normais independentes possui densidade conjunta, a saber, o produto das densidades marginais.*

D Desigualdades

Aqui daremos a prova das desigualdades usadas ao longo do texto.

Proposição D.1. *Para todo* $x \in \mathbb{R}$, *vale*

$$\exp(x) \leq x + \exp(x^2).$$

[29] Novamente usamos a simetria do produto interno $\mathbf{u}^T\mathbf{v} = \sum_{i=1}^n u_iv_i = \sum_{i=1}^n v_iu_i = \mathbf{v}^T\mathbf{u}$.

Demonstração. Usaremos as derivadas da função $f(x) = x + \exp(x^2) - \exp(x)$. Como $f'(x) = 1 + 2x \cdot \exp(x^2) - \exp(x)$,

$$f''(x) = 0 + 2 \cdot \exp(x^2) + 2x \cdot 2x \exp(x^2) - \exp(x) = (4x^2 + 2)\exp(x^2) - \exp(x).$$

Note que $x = 0$ é um ponto crítico:

$$f'(0) = 1 + 2 \cdot 0 \exp(0^2) - \exp(0) = 0,$$

e que

$$f''(0) = (4 \cdot 0^2 + 2)\exp(0^2) - \exp(0) = 1 > 0,$$

o que torna $x = 0$ um ponto de mínimo local. Resta ver que $\lim_{x\to\infty} f(x) = \lim_{x\to-\infty} f(x) = +\infty$ e podemos concluir que $x = 0$ um ponto de mínimo global. Portanto, para todo $x \in \mathbb{R}$, segue:

$$0 = f(0) \leq f(x) = x + \exp(x^2) - \exp(x) \implies \exp(x) \leq x + \exp(x^2).$$

□

Proposição D.2. *Para todo $x \in [0, \frac{1}{2}]$ vale*

$$\frac{1}{1-x} \leq \exp(2x).$$

Demonstração. Considere a função $h(x) = \exp(2x) - x \cdot \exp(2x) - 1$. Bastar mostrar que $h(x) \geq 0$ para $x \in [0, \frac{1}{2}]$, pois

$$\exp(2x) - x \cdot \exp(2x) - 1 = h(x) \geq 0 \implies 1 \leq \exp(2x) - x \cdot \exp(2x) = (1-x)\exp(2x) \implies \frac{1}{1-x} \leq \exp(2x).$$

Temos:

$$h(0) = \exp(2 \cdot 0) - 0 \cdot \exp(2 \cdot 0) - 1 = 0$$

e

$$h(\frac{1}{2}) = \exp(2 \cdot \frac{1}{2}) - \frac{1}{2} \cdot \exp(2 \cdot \frac{1}{2}) - 1 = \frac{1}{2}\mathrm{e} - 1 > 0.$$

Além disso, a derivada de h é positiva no intervalo em questão

$$h'(x) = 2\exp(2x) - (1 \cdot \exp(2x) + x2\exp(2x)) = (1 + 2x)\exp(2x) > 0.$$

Portanto, h é função crescente e obtemos o resultado:

$$0 = h(0) \leq h(x) \text{ para } x \in [0, \frac{1}{2}].$$

□

Proposição D.3. *Para todo $n \in \mathbb{N}$, vale*

$$n! \leq n^n.$$

Demonstração. Provaremos por indução em n. Para $n = 1$, é direto:

$$1 = 1! \leq 1^1 = 1.$$

Suponha que para $n = k > 1$ seja verdade

$$k! \leq k^k.$$

Então,

$$\begin{aligned}(k+1)! = (k+1) \cdot k! &\leq (k+1) \cdot k^k \\ &\leq (k+1) \cdot (k+1)^k \\ &\leq (k+1) \cdot (k+1)^{k+1},\end{aligned}$$

como queríamos demonstrar. □

Proposição D.4. *Para todo $p \in \mathbb{N}$, vale*

$$\frac{p^p}{p!} \leq \exp(p).$$

Demonstração. A aproximação de Stirling $p! \sim \sqrt{2\pi p}(\frac{p}{\mathrm{e}})^p$ fornece

$$\sqrt{2\pi} p^{p+\frac{1}{2}} \exp(-p) \leq p! \leq \mathrm{e} p^{p+\frac{1}{2}} \exp(-p),$$

o que implica

$$\sqrt{2\pi p}\frac{p^p}{p!} \leq \exp(p),$$

e como $\frac{p^p}{p!} \leq \sqrt{2\pi p}\frac{p^p}{p!}$, segue o resultado. Outra maneira simples de provar a desigualdade é usar a expansão em série da função exponencial. Basta ignorar todos os termos da série exceto o p-ésimo:

$$\begin{aligned}\exp(p) = \sum_{n=1}^{\infty} \frac{p^n}{n!} &= \frac{p}{1} + \frac{p^2}{2!} + \ldots + \frac{p^p}{p!} + \ldots \\ &\geq \frac{p^p}{p!},\end{aligned}$$

finalizando o resultado. □

Proposição D.5. *Para todos $x, \lambda \in \mathbb{R}$ vale*

$$2\lambda x \leq \lambda^2 + x^2.$$

Demonstração. Basta notar que $f(x) = \lambda^2 + x^2 - 2\lambda x = (\lambda - x)^2 \geq 0$ para todo $x \in \mathbb{R}$. □

Referências Bibliográficas

[1] Arcones, M.
1994. Limit theorems for nonlinear functionals of a stationary gaussian sequence of vectors. *Annals of Probability*, 22:2242–74.

[2] Brockwell, P. J. e Davis, R. A.
1991. *Time Series: Theory and Methods*. Springer.

[3] Casella, G. e Berger, R. L.
2002. *Statistical Inference*. Thomson Learning.

[4] Chung, K. L.
2006. *A Course in Probability Theory*. Academic Press.

[5] Csörgő, S. e Mielniczuk, J.
1996. The empirical process of a short-range dependent stationary sequence under gaussian subordination. *Probability Theory Related Fields*, 104:15–25.

[6] Hampel, F. R.
1971. A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 46:1887–1896.

[7] Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel
1986. *Robust Statistics: The Approach Based on Influence Functions.* John Wiley.

[8] Huber, P. J.
1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics.*, 35:73–101.

[9] Huber, P. J.
1981. *Robust Statistics*. John Wiley & Sons Inc.

[10] James, B. R.
1981. *Probabilidade: um curso em nível intermediário.* Projeto Euclides.

[11] Lévy-Leduc, C., H. Boistardb, M. E, T. M. S., and V. A. Reisen
2011. Robust Estimation of the Scale and of the Autocovariance Function of Gaussian Shortand Long-range Dependent Processes. *Journal of Time-series Analysis*, 32:135–156.

[12] Ma, Y. e Genton, M.
2000. Highly robust estimation of the autocovariance function. *Journal of Time-series Analysis*, 21:663–84.

[13] Rousseeuw, P. e Croux, C.
1993. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–83.

[14] van der Vaart, A. W.
1998. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

[15] van Handel, R.
. *Probability in high dimension, Lecture notes*. https://web.math.princeton.edu/~rvan/APC550.pdf, Acesso em 16.02.2021.

[16] Vershyin, R.
2019. *High-Dimensional Probability - An Introduction with Applications in Data Science*. University of California, Irvine.

[17] Wainwright, M. J.
2019. *High-Dimensional Statistics - A Non-Asymptotic Viewpoint*. Cambridge University Press.