Universidade de São Paulo - Instituto de Física

Tópicos em Mecânica Estatística de Sistemas Complexos

Uma abordagem mecânico-estatística de dois tópicos de interesse em finanças, economia e sociologia.

Rafael S. Calsaverini

São Paulo
20 de dezembro de 2012

Universidade de São Paulo - Instituto de Física

Tópicos em Mecânica Estatística de Sistemas Complexos

Uma abordagem mecânico-estatística de dois tópicos de interesse em finanças, economia e sociologia.

Rafael S. Calsaverini

Tese apresentada ao Instituto de Física da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Orientador:
Prof. Dr. Nestor Felipe Caticha Alfonso

Co-Orientador:
Prof. Dr. Renato Vicente

São Paulo
20 de dezembro de 2012

Sumário

Lista de Figuras

Resumo

Neste trabalho exploramos dois tópicos de aplicação de teoria de informação e mecanica estatística a problemas de interesse em finanças, economia e sociologia. No primeiro tópico exploramos a conexão entre a teoria de dependência estatística e a teoria de informação através da teoria de cópulas. Após uma revisão do conceito de cópulas, reformulamos a definição de medidas de dependência de Renyi[**?**] usando esse conceito e mostramos que a informação mútua satisfaz todos os requisitos para ser uma medida de dependência. Em seguida mostramos uma relação entre a informação mútua e a entropia da distribuição cópula, e uma relação mais específica para a decomposição da informação mútua de distribuições elípticas em uma parte devido à dependência gaussiana linear e uma parte não-linear. As conseqüências dessas duas decomposições sobre o risco do uso de pseudo-medidas de dependências são então discutidas. Esses resultados são usados para desenvolver um método para detectar desvio de gaussianidade na dependência de séries temporais e para ajuste de cópulas t sobre dados empiricos[**?**].

No segundo tópico desenvolvemos um modelo para emergência de autoridade em sociedades humanas. Discutimos as motivações empíricas com raízes na neurociência, na primatologia e na antropologia para um modelo matemático que explique o espectro amplo de tipos de organização social humana no eixo igualitário-hierárquico. O modelo resulta da aplicação de teoria de informação sobre uma hipótese sobre os custos evolutivos envolvidos. O modelo apresenta um diagrama de fases rico, com diferentes regimes que podem ser interpretadas como correspondendo a diferentes tipos de organização social, desde igualitária até hierárquica. Os parâmetros de controle do sistema são identificados com a capacidade cognitiva da espécie em questão e as pressões ecológica e social em que o grupo está imerso.

Abstract

In this work we explore two topics of interest in the application of information theory and statistical mechanics techniques to problems in finance, economics and sociology. In the fisrt topic we study the conexion between statistical dependency theory and information theory mediated by copula theory. After a revision of the concept of copulas, we reformulate the definition of dependency measures given by Renyi [**?**] using this concepyt and show that mutual information satisfy all the requirements to be a dependency measure. We then show a relationship between mutual information and copula entropy, and a more specific decomposition of the mutual information of an elliptical distribution into its linear and non-linear parts. We evaluate the risk of using naive pseudo-dependency measures. Those results are then used to develop a method to detect deviation from gaussianity in the dependency of time series and a method to adjust t-copulas to data[**?**].

On the second topic we develop a model for the emergence of authority in early human societies. We discuss empirical motivations with roots in neuroscience, primatology and anthropology for a mathematical model able to explain the spectrum of observed types of human social organization in the egalitarian-hierarchical axis. The model results from the application of information theory on a hypothesys about the evolutive costs involved in social life. The model generates a rich phase diagram, with diferent regimes which can be interpreted as different types of societal organization, from egalitarian to hierarchical. The control parameters of the model are connected to the cognitive capacity of the species in question and ecological and social pressures.

1 | Introdução

1.1 Visão geral

Este trabalho trata de dois tópicos – uma abordagem da teoria de dependência estatística e um modelo para a origem de estruturas sociais hierárquicas – sob o ponto de vista da mecânica estatística, da teoria de informação e da inferência estatística. A adoção desse ponto de vista norteia as estratégias de modelagem matemática aqui selecionadas, e de uma certa forma, são mais essenciais ao trabalho do que os específicos tópicos em si. Dessa forma se faz necessário explicitar e esclarecer o ponto de vista adotado antes que os tópicos específicos sejam apresentados.

[1] ; and

[2]

1.2 Inferência, Probabilidades e Entropia

[3] ; ; and

Adquirir informação e tomar decisões sob incerteza – dois pontos centrais em qualquer estudo quantitativo – são os temas centrais da teoria da inferência estatística. A tradição do uso da teoria de probabilidades como ferramenta de inferência é centenária e remonta aos primeiros trabalhos sobre o conceito de probabilidades no século XVII *[more sources needed]*. A relação entre o conceito de probabilidade e os problemas de inferência ficaram ainda mais fortes com os trabalhos de **?**][1] e **?**][2], e as versões mais modernas desse paradigma[3] lançam luz sobre a natureza da física estatística e do conceito de entropia. Nessa introdução pretendemos apresentar rapidamente o paradigma de inferência segundo o método de Máxima Entropia (ME) e suas relações com a mecânica estatística, que pensamos ser a linha unificadora que dá coerência à diversidade de temas abordados nesse trabalho.

Raciocínio sobre informação completa a respeito da veracidade ou não de um conjunto de proposições pode ser representado através da tradicional álgebra booleana. Se é conhecido o valor de verdade de uma certa proposição e como ela se relaciona com outras

proposições, pode-se inferir o valor de verdade das proposições relacionadas através das regras bem definidas da álgebra de proposições. Por exemplo, se é sabido que $P_1 \Rightarrow P_2$, e há certeza de que P_1 é verdadeira, pode-se inferir imediatamente que P_2 é verdadeira. Da mesma forma, a certeza de que P_2 é falsa imediatamente implica na certeza de que P_1 é falsa. Em outras palavras, a hipótese $P_1 = V$ fornece *informação completa* a respeito de P_2, bem como a hipótese $P_2 = F$ fornece informação completa sobre P_1. Entretanto, a certeza a respeito da falsidade de P_1 não oferece conclusão alguma, dentro desse paradigma de inferência sobre informação completa, a respeito da veracidade de P_2. Não é difícil porém formular um exemplo em que a informação sobre a falsidade de P_1 fornece *alguma informação*, ainda que incompleta, sobre P_2.

Consideremos, em um exemplo simples, a hipótese de que a proposição P_1 ="vai chover" implica a proposição P_2 ="há nuvens de chuva". No *ambiente lógico* criado por essa hipótese, a observação de nuvens de chuva não leva à conclusão certa de que está chovendo, mas é uma decisão razoável carregar um guarda-chuvas ao se observar essas nuvens. De alguma forma, a observação de que há nuvens de chuva trouxe alguma informação ao observador a respeito da possibilidade de que chova. Construir um método de inferência capaz de levar em conta informação incompleta é o objetivo da teoria de probabilidades bayesiana e do método de máxima entropia.

[4] A justificativa para usar números reais vem de um argumento simples de transitividade - se a confiança na veracidade de P_1 é maior que na veracidade de P_2 e esta é maior que a confiança na veracidade de P_3, então, um requisito razoável é que a confiança em P_1 seja maior que em P_3. Dessa forma, $(P_1|Q) > (P_2|Q)$ e $(P_2|Q) > (P_3|Q)$ implica $(P_1|Q) > (P_3|Q)$. Isso é suficiente para mostrar que existe uma representação real para essas quantidades. Conseqüências interessantes de se relaxar o requesito de transitividade são discutidas em **?**]

[5]

1.2.1 *Probabilidades e Inferência*

Para derivar uma teoria coerente de inferência, devem ser estabelecidos alguns requisitos. Dada duas proposições P e Q, postulamos uma medida[4,5] $(P|Q) \in \mathbb{R}$ denominada *plausibilidade da proposição P dada a proposição Q*. A plausibilidade $(P|Q)$ representa o *grau de confiança* de que P esteja correta dada uma certa informação prévia Q. Postulamos ainda que, sempre que existam duas formas diferentes de calcular a mesma plausibilidade, o resultado deve ser idêntico. Esse requisito leva aos seguintes resultados[6,7]:

- A plausibilidade[8] de não-P dado Q é uma função monotônica e decrescente da plausibilidade de P dado Q:

$$(\bar{P}|Q) = F((P|Q)).$$

- A plausibilidade da conjunção "P_1 e P_2" $(P_1 \wedge P_2)$ dado Q é uma função das plausibilidades de P_1 dado Q e de P_2 dado $P_1 \wedge Q$:

$$(P_1 \wedge P_2|Q) = G((P_1|Q), (P_2|Q \wedge P_2)).$$

[6] Todas as outras possibilidades são consideradas em **?**] e essas são as únicas que não conduzem a resultados manifestamente inconsistentes.

[8] Os seguintes símbolos serão usados para as operações booleanas no presente capítulo:

Conjunção - $\wedge$: representa a conjunção "e" entre duas proposições: $P \wedge Q$, lido "p e q".

Disjunção - $\vee$: representa a disjunção "ou" entre duas proposições: $P \vee Q$, lido "p ou q".

Negação - $\bar{}$: representa a negação "não" de uma proposição: $\bar{P}$, lido "não-p".

Uma série de teoremas sobre a forma das funções $F(\cdot)$ e $G(\cdot,\cdot)$ podem ser demonstrados com o requisito de consistência e as regras básicas da álgebra booleana. Alguns dos principais teoremas, cujas provas se encontram no apêndice A, *Provas dos teoremas de Cox*, são:

1º teorema de Cox

Teorema 1. *Uma vez que uma representação consistente de plausibilidades $(P|Q)$ com um ordenamento bem definido foi encontrada, sempre é possível encontrar uma outra equivalente $\pi(P|Q)$ de forma que $G(u,v) = uv$, ou seja:*

$$\pi(P_1 \wedge P_2|Q) = \pi(P2|Q \wedge P_1)\pi(P_1|Q) \tag{1.1}$$

Valores limites para plausibilidades

Teorema 2. *Uma vez que uma representação consistente de plausibilidades $\tilde{\pi}(P|Q)$ que satisfaça a regra do produto, é sempre possível encontrar uma equivalente $\pi(P|Q)$ tal que:*

$$0 \leq \pi(P|Q) \leq 1 \tag{1.2}$$

de tal forma que $\pi(P|Q) = 0$ se, e somente se, P for uma proposição falsa dado Q e $\pi(P|Q) = 1$ se, e somente se, P for uma proposição verdadeira dado Q.

2º teorema de Cox

Teorema 3. *Uma vez que uma representação consistente de plausibilidades $\pi(P|Q)$ com um ordenamento bem definido foi encontrada para a qual vale a regra do produto, sempre é possível encontrar uma outra equivalente $p(P|Q)$, que também satisfaz a regra do produto, de forma que $F(u) = 1 - u$, ou seja:*

$$p(\bar{P}|Q) = 1 - p(P|Q) \tag{1.3}$$

Tomados em conjunto, esses teoremas sugerem que as regras de uma álgebra de plausibilidades deve ser idêntica às conhecidas regras da Teoria das Probabilidades. A partir de agora portanto daremos o nome "probabilidade" ao funcional $p(P|Q)$, e interpretaremos probabilidades como formas de codificar matematicamente graus de certeza a respeito de certas proposições. Quando essas proposições são afirmações sobre o valor de uma grandeza matemática, definem-se distribuições de probabilidade sobre o valor dessas variáveis:

$$P(x \in [a,b]|Q) = \int_a^b \mathrm{d}x\, p(x|Q) \tag{1.4}$$

Um modelo matemático, nesse paradigma, é portanto uma atribuição de distribuições de probabilidade para as variáveis de interesse do modelo, indicando graus de confiança sobre os valores dessas variáveis sob certas condições.

1.2.2 *Informação e Máxima Entropia*

Se modelos de inferência são atribuições de probabilidades sobre as variáveis de interesse, como é possível fazer isso a partir de informação pré-existente sobre o sistema sendo modelado, ou ainda, como é possível incorporar novas informações obtidas sobre o sistema? Eventualmente, o objetivo de realizar inferência é processar informação nova que nos é disponibilizada depois da realização de um certo experimento ou observação. No presente paradigma isso significa atualizar nossa atribuição de probabilidades. Suponha a proposição $P_{[a,b]}$ ="A variável $X \in \mathcal{X}$ tem seu valor no intervalo $[a, b]$". Suponha ainda que, inicialmente, há um certo conjunto de informações que nos levam a crer que, nas condições U, a nossa atribuição de probabilidades dever ser:

$$p(P_{[a,b]}|U) = \int_a^b p(x)\mathrm{d}x.$$

Uma vez que nova informação que nos force a mudar de opinião a respeito da atribuição de probabilidades torne-se disponível, qual deve ser a nova distribuição a ser utilizada? Se pudessemos ordenar todas as distribuições $q(x) \in \mathbb{P}$, o conjunto de todas as distribuições possíveis, em ordem de preferência como nova distribuição a ser atribuida a P_x, certamente escolheríamos a com melhor ranking de preferência. Se essa preferência for transitiva, existe um funcional $S : \mathbb{P} \to \mathbb{R}$ que representa esse ordenamento e a nova distribuição será obtida através da maximização do funcional $S[\cdot]$:

$$q^\star(x) = \underset{q(x)\in\mathbb{P}}{\arg\max}\, S[q(x)|p(x)],$$

sob quaisquer vínculos impostos pela nova informação. É possível impor requisitos plausíveis sobre $S[\cdot]$ de forma a definir um único funcional compatível com uma atualização racional de crenças (probabilidades)? De fato é possível impondo apenas um requisito: as crenças devem ser atualizadas apenas até onde requerido pela nova informação disponível. Resumidamente (detalhes podem ser obtidos em **?**]), esse princípio leva às seguintes conseqüências:

- **Localidade** - Se a nova informação diz respeito apenas a um subdomínio do espaço onde X toma valores, então a atribuição de probabilidades fora desse subdomínio não deve mudar. Isso implica que o funcional deve ser aditivo:

  $$S[q|p] = \int_{\mathcal{X}} \mathrm{d}\mu(x) F(q(x), p(x), x)$$

  onde $\mu(x)$ é uma medida de integração sobre $\mathcal{X}$.

- **Invariância por mudança de variáveis** - Uma mudança de sistema de coordenadas não deve mudar a forma do funcional S e nem a

ordem das preferências. Isso implica que:

$$S[q|p] = \int_{\mathcal{X}} \mathrm{d}\mu(x)\Phi\left(\frac{p(x)}{\mu(x)}, \frac{q(x)}{\mu(x)}\right)$$

- **Ausência de nova informação** - Se não há nova informação, não há razão para mudar de idéia e, portanto, o máximo sem vínculos de $S[\cdot|\cdot]$ deve ser a própria distribuição original $p(x)$. Isso implica que:

$$S[q|p] = \int_{\mathcal{X}} \mathrm{d}x\ p(x)\Phi\left(\frac{q(x)}{p(x)}\right)$$

- **Sistemas independentes** - Se a distribuição $p(x)$ contém informação de que dois subsistemas X_1 e X_2 são independentes, nova informação a respeito de um deles não deve afetar o outro. Esse princípio leva a uma equação funcional para $\Phi(x)$ que finalmente implica que:

$$S[q(x)|p(x)] = -\int_{\mathcal{X}} \mathrm{d}x\ q(x)\log\frac{q(x)}{p(x)} \tag{1.5}$$

Esse funcional é conhecido em Teoria de Informação[9] e é denominado "entropia relativa". Os passos acima levam à formulação do princípio de máxima entropia[10]:

[9]

[10] Os nomes atribuídos às distribuições são traduções do inglês *prior distribution* e *posterior distribution*

Princípio da Máxima Entropia. *Dada uma atribuição inicial de probabilidades sobre uma variável $X \in \mathcal{X}$ dada por $p(x)$ (distribuição a priori, ou prévia), quando nova informação se torna disponível, uma nova distribuição de probabilidades deve ser atribuida de forma a maximizar a entropia relativa entre a antiga distribuição $p(x)$ e a nova $q(x)$ (distribuição posterior), de forma a satisfazer os vínculos impostos pela nova informação.*

$$q(x) = \arg\max_{q(x)} S[q(x)|p(x)] \tag{1.6}$$

$$F_i[q(x)] = 0\ ,\ i = 1, 2, \ldots, m \tag{1.7}$$

onde $F_i[\cdot]$ são funcionais que codificam os vínculos relacionados às informações disponíveis.

Uma vez que o funcional $S[q|p]$ é convexo, se os vínculos forem também convexos o máximo será interior e único, e pode ser encontrado pela técnica de multiplicadores de Lagrange, resolvendo para $q(x)$ a condição variacional de primeira ordem:

$$\frac{\delta}{\delta q(x)}\left(S[q|p] - \sum_i \lambda_i F_i[q]\right) = 0 \tag{1.8}$$

A entropia relativa na eq.(1.5) também pode ser ligada a outro conceito corrente em teoria de informação e geometria de distribuições de probabilidade, denominado divergência de Kullback-Leibler[11]:

[11]

$$D[q(x)|p(x)] = \int_{\mathcal{X}} \mathrm{d}x\; q(x) \log \frac{q(x)}{p(x)} = -S[q(x)|p(x)]. \tag{1.9}$$

A divergência de Kullback-Leibler apresenta as propriedades de uma pré-métrica, ou seja: $D[q(x)|p(x)] \geq 0$ e $D[q(x)|p(x)] = 0$ se, e somente se, $q(x) = p(x)$, no sentido de igualdade de distribuições. Entretanto, por não ser simétrica e não satisfazer a desigualdade do triângulo, $D[q|p]$ não oferece uma estrutura métrica para o conjunto de distribuições de probabilidades. Se restrita a uma família paramétrica $\mathcal{P}_\theta$ de distribuições parametrizadas por um certo conjunto de parâmetros θ (denotaremos por $p(x|\theta)$), a quantidade:

$$D[p(x|\theta + \mathrm{d}\theta), p(x|\theta)] = \frac{1}{2} \sum g_{ij} \mathrm{d}\theta_i \mathrm{d}\theta_j + \mathrm{O}(\mathrm{d}\theta^3), \tag{1.10}$$

onde $g_{ij} = \left\langle \partial \log p(x|\theta)/\partial\theta_i \; \partial \log p(x|\theta)/\partial\theta_j \right\rangle$ é a chamada métrica de Fisher-Rao, que provê uma estrutura métrica[12] ao conjunto $\mathcal{P}_\theta$. Nessa linguagem, o princípio de máxima entropia pode ser entendido como um princípio de "mínima distância" – a distribuição posterior deve ser tão próxima da distribuição *a priori* quanto permitido pelos vínculos impostos pela nova informação.

[12]

1.2.3 *Informação e Vínculos - atualização Bayesiana*

Um caso específico de aplicação do princípio de Máxima Entropia é o da ajuste de parâmetros teóricos de um modelo quando novos dados empíricos são coletados. Suponha que um par de variáveis X e Θ são considerados em um modelo M. A variável X é experimentalmente mensurável e a variável Θ é um parâmetro teórico do modelo. O modelo oferece informação prévia sob a forma de (1) uma distribuição *a priori* dos valores possíveis do parâmetro Θ, dada por $p(\theta|M)$ e (2) uma distribuição condicional *a priori* $p(x|\theta, M)$ que indica, dado um valor do parâmetro $\Theta = \theta$, os possíveis resultados para X. Eventualmente o valor de X é medido, com resultado $X = x_0$. Como devemos atualizar nossa atribuição de probabilidades? O princípio da máxima entropia indica que a distribuição posterior $q(x, \theta)$ é aquela que maximiza o funcional da eq.(1.5) sob o vínculo de que conhecemos o valor de X. Ou seja, o vínculo é dado por[13]:

[13] Note que essa equação implementa, na verdade, um número infinito de vínculos – um para cada valor de x.

$$\int q(x, \theta) d\theta = q(x) = \delta(x - x_0). \tag{1.11}$$

A distribuição *a priori* sobre x e θ é dada por $p(x, \theta|M) = p(x|\theta, M)p(\theta|M)$. Finalmente, a minimização é dada por:

$$\frac{\delta}{\delta q}\left[S[q|p] + \lambda\left(\int \mathrm{d}x\mathrm{d}\theta\; q(x,\theta) - 1\right) + \int \mathrm{d}x\; \beta(x)\left(\int \mathrm{d}\theta q(x,\theta) - \delta(x - x_0)\right)\right] = 0$$

onde λ e $\beta(x)$ são multiplicadores de Lagrange que implementam, respectivamente, o vínculo de normalização de $q(x,\theta)$ e os vínculos impostos pela eq.(1.11). Executando essa diferenciação funcional e isolando $q(x,\theta)$, obtemos[14]:

[14] O segundo passo é uma aplicação a definição de distribuição condicional

$$q(x,\theta|M) = p(x,\theta|M)\frac{e^{\beta(x)}}{Z} = p(\theta|x,M)p(x|M)\frac{e^{\beta(x)}}{Z} \tag{1.12}$$

onde Z é uma constante de normalização. Impondo o vínculo eq.(1.11), finalmente obtém-se:

$$\begin{aligned} q(x,\theta|M) &= q(x|M)q(\theta|x,M) = \delta(x-x_0)p(\theta|x,M) \\ &\Rightarrow q(\theta|x_0,M) = p(\theta|x_0,M) \end{aligned}$$

Ou seja, as distribuições condicionais de θ devem ser iguais antes e depois de receber a informação, pois apenas informação a respeito de x – informação marginal, que diz respeito apenas à distribuição marginal de x – foi recebida. Isso é conseqüência do princípio de mínima atualização usado na dedução do princípio de máxima entropia: apenas se deve atualizar as probabilidades quando isso é imposto pela nova informação recebida. Essa equação pode ser reescrita como:

$$q(\theta|x_0,M) = \frac{p(x_0|\theta,M)p(\theta,M)}{p(x_0|M)} \tag{1.13}$$

e esse é o teorema de Bayes como entendido em inferência bayesiana – como um princípio de atualização de graus de confiança quando um nova informação está disponível. A distribuição $q(\theta|x_0,M)$ incorpora informações a respeito do modelo original e do fato de que a variável X foi medida e vale x_0.

1.2.4 *Informação e vínculos - distribuições de Gibbs*

Outro tipo de possível informação que pode ser recebido é a respeito do valor esperado de uma certa função de x:

$$\langle E(x)\rangle = E_0.$$

[15] Esse raciocínio é reminiscente do encontrado em **?**], onde a distribuição de probabilidades para um conjunto de partículas é encontrada maximizando a entropia sob vínculos associados a quantidades conservadas microscopicamente.

Nesse caso, a maximização da entropia será:

$$\begin{aligned} &\frac{\delta}{\delta q}\left[S[q|p] - \lambda\left(\int \mathrm{d}x\, q(x) - 1\right) - \beta\left(\int \mathrm{d}x\, q(x)E(x) - E_0\right)\right] = 0 \\ &= -\log\left(\frac{q(x)}{p(x)}\right) - 1 - \lambda - \beta E(x) = 0 \end{aligned}$$

[16]

e obtém-se finalmente:

$$q(x) = \frac{1}{Z}p(x)e^{-\beta E(x)} \tag{1.14}$$

[17]

Essa distribuição faz parte da classe de distribuições gibbsianas, comuns em mecânica estatística[15,16].

Uma forma alternativa dessa visão pode ser encontrada[17] em **?**].

2 | *Informação mútua, teoria de cópulas e dependência estatística*

2.1 *Dependência Estatística e medidas de dependência*

O CONCEITO DE DEPENDÊNCIA ESTATÍSTICA é central à teoria de probabilidades. Fazer hipóteses a respeito da dependência estatística entre as variáveis de interesse em um modelo é comum em diversas áreas – da física à análise financeira. Não é óbvio, entretanto, como expressar esse conceito de maneira precisa. A formalização precisa desse conceito é um dos objetivos desse capítulo. De maneira informal e grosseira, dependência estatística diz respeito a quanta *informação* se obtém a respeito de uma variável quando o valor de outra variável é conhecido. Os dois casos extremos podem ser mais facilmente entendidos em primeira análise: o caso de completa independência e o caso de completa dependência. Duas variáveis são ditas independentes se sua distribuição conjunta pode ser fatorada em um produto[1]:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y). \tag{2.1}$$

[1] Manteremos o foco de nossa atenção em distribuições bivariadas, uma vez que a generalização é imediata.

De maneira similar, pode-se dizer que duas variáveis são independentes quando a distribuição condicional $p(x|y)$ é idêntica à distribuição marginal de X - $p_X(x)$. Nessa situação, o conhecimento do valor da variável Y não fornece qualquer informação a respeito da variável X. Duas variáveis são ditas completamente dependentes quando uma pode ser escrita como função monotônica da outra:

$$x = F(y). \tag{2.2}$$

Nesse caso, o conhecimento de uma das variáveis determina completamente o valor de outra. Ou seja $p(x|y) = \delta(x - F(y))$, com $F(\cdot)$ uma função monotônica. Pode-se, dessa maneira, tentar introduzir alguma forma concreta de medir a dependência estatística em um parâmetro que possa ser estimado e usado para caracterizar a dependência entre variáveis de forma mais concreta.

Um parâmetro comumente usado para esse fim é a chamada correlação linear[2]

$$\hat{\rho}_{XY} = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{2.3}$$

A correlação linear $\hat{\rho}$ é um número real no intervalo $[-1, 1]$, simétrica com relação à transposição de X e Y que sempre se anula quando duas variáveis são independentes. É certamente a medida mais popular de dependência utilizada em todo tipo de análise estatística. Entretanto, seria mais preciso dizer que a correlação linear, como a nomenclatura aqui empregada sugere, é apenas uma medida de dependência linear. Há diversas falhas dessa medida em capturar a completa dependência entre duas variáveis. Em particular é fácil notar que é possível obter duas variavéis com correlação linear nula em que, entretanto, haja forte dependência entre ambas. Um exemplo simples é o par de variáveis definido por:

$$Y = f(X) + \epsilon \tag{2.4}$$

em que X e ϵ sejam independentes e tenham distribuições simétricas em torno da origem e $f(\cdot)$ seja qualquer função par. Nesse caso temos:

$$\mathrm{Cov}(X,Y) = E[Xf(X)] + E[X]E[\epsilon] - E[X]E[f(X)]$$

Note que se a distribuição de X é uma função par, então o valor esperado de qualquer função ímpar de X é nulo, o que anula a expressão acima. Dessa forma temos $\hat{\rho}_{XY} = 0$. Entretanto, caso a distribuição de ϵ seja bastante concentrada em torno da origem, o conhecimento de X pode fornecer informação quase completa a respeito de Y. Essa informação não é capturada pela correlação linear. O epíteto "linear", usado nesse trabalho para descrever a correlação, pode ser melhor compreendido se notarmos uma propriedade interessante da correlação: ela é invariante por mudanças de escala lineares nas variáveis X e Y. Uma reparametrização do tipo:

$$X' = \alpha_X X + \beta_X$$
$$Y' = \alpha_Y Y + \beta_Y$$

não muda o valor da correlação linear: $\hat{\rho}_{X'Y'} = \hat{\rho}_{XY}$. Entretanto uma mudança mais geral de escala não preserva essa propriedade. Uma nova variável:

$$Y' = g(Y)$$

com $g(\cdot)$ monotônica, a princípio contém exatamente a mesma informação a respeito de X que a antiga variável Y. Entretanto não se garante que a correlação se mantenha invariante. Seria esperado, além disso, que a afirmação de que a correlação entre duas variáveis é máxima em módulo fosse uma indicação de que a dependência entre as duas variáveis é máxima. Entretanto, isso não é garantido.

[2] A rigor, o módulo da correlação linear. A correlação linear mede, além de dependência, concordância, ou seja, quanto duas variáveis reais apresentam variação coordenada de seus sinais. Essa informação extra não é captada apenas pelo conceito de dependência.

2.1.1 *O que se deseja de uma medida de dependência?*

A DIGRESSÃO ACIMA acerca da natureza da correlação linear imediatamente suscita a pergunta: que parâmetros são boas medidas de dependência e quais são suas características? Pode-se enumerar uma série de desejos a respeito dessas medidas que se baseiem na noção intuitiva de dependência como o conteúdo de informação de uma variável a respeito de outra. Explicitamente, esperamos que:

1. uma boa medida de dependência seja um funcional $R : P_2 \to \mathbb{R}$ da distribuição conjunta de probabilidades, bem definido para qualquer par de variáveis aleatórias X e Y;
2. o funcional seja invariante sob permutação das variávies X e Y: $R(X, Y) = R(Y, X)$;
3. o funcional seja nulo *se, e somente se* as variáveis forem estritamente independentes;
4. o funcional atinja um valor máximo *se, e somente se* as variáveis apresentem dependência máxima, ou seja, sejam funções monotônicas uma da outra;
5. o funcional seja invariante por escolhas de novas variáveis $(X, Y) \to (U = u(X), V = v(Y))$ desde que nenhuma informação seja perdida, ou seja, desde que $u(\cdot)$ e $v(\cdot)$ sejam funções bijetoras;
6. o funcional seja uma função monotônica e crescente do módulo da correlação linear para o caso de distribuições conjuntas gaussianas.

Essa série de requisitos, essencialmente[3] enumerados pela primeira vez [4] por **?**] , não são suficientes para escolher um parâmetro único e são satisfeitos por uma grande variedade de diferentes parâmetros usados em estatística. Como exemplo podemos citar o τ de Kendall. Dados dois pares, (X_1, Y_1) e (X_2, Y_2), de pontos sorteados independentemente da distribuição $p_{XY}(x, y)$, o τ de Kendall é dado por:

$$\begin{aligned}\tau_{XY} &= \text{Prob}\left\{(X1 - X2)(Y1 - Y2) > 0\right\} - \text{Prob}\left\{(X1 - X2)(Y1 - Y2) < 0\right\} \\ &= 4\int F(x, y)dF(x, y) - 1,\end{aligned}$$

onde $F(x, y)$ é a distribuição cumulativa de X e Y. Outra medida que satisfaz esses requisitos é a correlação de postos de Spearman dada por:

$$\rho^S_{XY} = 12\int \left(F(x, y) - F_X(x)F_Y(y)\right) dF_X(x)dF_Y(y), \tag{2.5}$$

onde $F_X(x)$ e $F_Y(y)$ são as distribuições cumulativas marginais de X e Y respectivamente.

[3] Renyi exigia ainda que a medida tomasse valores no conjunto $[0, 1]$, o que dispensamos, uma vez que é sempre possível transpor uma medida no intervalo $[0, \infty]$ para esse conjunto através de uma função monotônica. Além disso há requisitos de continuidade e convergência para sequências convergentes de distribuições.

[4] ; and

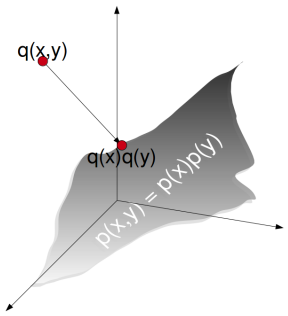


Figura 2.1 – Representação pictórica da projeção de uma distribuição em uma certa família de distribuições convenientes. O espaço representado na figura corresponde de forma pictórica ao espaço formado por todas as distribuições de probabilidade. A superfície corresponde a um sub-espaço, nesse caso, o sub-espaço de distribuições fatoráveis, correspondentes a variáveis independentes. A projeção de uma distribuição qualquer $q(x, y)$ sobre esse sub-espaço através da minimização da divergência de Kullback-Leibler resulta na distribuição dada pelo produto das distribuições marginais $q(x)q(y)$.

2.2 *Informação Mútua*

Do ponto de vista de teoria de informação a dependência mútua de um conjunto de variáveis pode ser quantificada através da mínima "distância na variedade estatística" (divergência de Kullback-Leibler) entre a distribuição conjunta dessas variáveis e a "sub-variedade" de distribuições independentes (veja figura 2.1) Isso pode ser escrito da forma:

$$I(X_1, \ldots, X_n) = \min_{\{\phi_i(x_i)\}} \int p(x_1, \ldots, x_n) \log \frac{p(x_1, \ldots, x_n)}{\prod_{i=1}^n \phi_i(x_i)} \prod_{i=1}^n \mathrm{d}x_i \quad (2.6)$$

Fazendo a derivada funcional da expressão a ser minimizada com relação às distribuições indeterminadas $\phi_k(x)$ obtemos a condição de extremo, com o vínculo de que as distribuições $\phi_j(x)$ sejam normalizadas:

$$\phi_k(x_k) = \int p(x_1, x_2, \ldots, x_n) \prod_{i \neq k} \mathrm{d}x_i = p_{X_k}(x_k) \quad (2.7)$$

Ou seja: a distribuição $\phi_j(x_j)$ devem ser a distribuição marginal da variável X_j, e podemos reescrever:

$$I(X_1, \ldots, X_n) = \int p(x_1, \ldots, x_n) \log \frac{p(x_1, \ldots, x_n)}{\prod_{i=1}^n p_{X_i}(x_i)} \prod_{i=1}^n \mathrm{d}x_i \quad (2.8)$$

Esse funcional que mede a dependência mútua entre grupos de variáveis é denominado Informação Mútua ou Correlação Total[5]. Interpretações para a o funcional podem ser obtidas notando que:

$$I(X_1, \ldots, X_n) = \sum_{i=1}^n H(X_i) - H(X_1, \ldots, X_n) \quad (2.9)$$

$$= H(X_j) - H(X_j | X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n) \quad (2.10)$$

Dessa forma, pode-se interpretar $I(X_1, \ldots, X_n)$ como a redução na incerteza a respeito da variável X_j proporcionada pelo conhecimento das variáveis $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n$. A informação mútua para duas variáveis, dada por:

$$I(X, Y) = \int \mathrm{d}x\, \mathrm{d}y\; p(x, y) \log \frac{p(x, y)}{p_X(x), p_Y(y))}, \quad (2.11)$$

satisfaz todos os critérios discutidos na seção anterior para ser uma boa medida de dependência.

[5] Alguns autores reservam o nome Informação Mútua para o caso de duas variáveis

2.3 *Cópulas*

A elucidação do conceito de dependência eventualmente se choca com a noção de cópula[6]. Um teorema devido a Sklar[7] permite

[6]

[7]

separar a distribuição cumulativa conjunta - $F_{X,Y}(x,y)$ - de duas variáveis em duas partes: (a) as distribuições cumulativas marginais de cada variável - $F_X(x)$ e $F_Y(y)$, que trazem informação idiossincrática a respeito de cada uma das variáveis (b) e uma função cópula C(u,v), que traz informação sobre a dependência entre as variáves. De maneira geral o teorema de Sklar pode ser enunciado da seguinte forma:

Teorema 4 (Teorema de Sklar)**.** *Para toda distribuição cumulativa conjunta contínua de duas variáveis $F_{XY}(x,y)$, com distribuições cumulativas $F_X(x)$ e $F_Y(y)$, existe uma função cópula única $C(u,v)$ tal que:*

$$F_{XY}(x,y) = C(F_X(x), F_Y(y)). \tag{2.12}$$

Similarmente, dadas quaisquer duas distribuições cumulativas univariadas $F_X(x)$ e $F_Y(y)$ e uma função cópula $C(u,v)$, é possível construir uma distribuição conjunta dada por $F_{XY}(x,y) = C(F_X(x), F_Y(y))$.

A PRÓPRIA FUNÇÃO CÓPULA é uma legítima distribuição cumulativa conjunta, associada às variáveis $U = F_X(X)$ e $V = F_Y(Y)$:

$$F_{UV}(u,v) = C(u,v)$$

Dessa forma, podemos também definir a densidade de cópula, a densidade de probabilidade das variávies U e V:

$$p_{UV}(u,v) = c(u,v) = \frac{\partial^2 C}{\partial u \partial v}$$

Essa definição implica que a densidade de probabilidade conjunta das variáveis X e Y é dada por:

$$p_{XY}(x,y) = c(F_X(x), F_Y(y)) p_X(x) p_Y(y) \tag{2.13}$$

dá para explicar isso melhor O teorema de Sklar permite dividir a informação contida na distribuição conjunta em duas partes: a parte que diz respeito às propriedades de cada uma das variáveis, dada pelas distribuições marginais, e a parte que diz respeito à dependência estatística entre as duas variáveis.

2.3.1 Exemplos

COMO EXEMPLOS DE FUNÇÕES CÓPULA temos as cópulas associadas às distribuições multivariadas mais comumente utilizadas. Qualquer família de distribuições multivariadas com um conjunto de parâmetros θ dada por $p(x,y|\theta)$ define uma função cópula dada por:

$$C(u,v|\theta) = \int_{-\infty}^{F_X^{-1}(u)} \int_{-\infty}^{F_X^{-1}(v)} \mathrm{d}x\mathrm{d}y p(x,y|\theta) \tag{2.14}$$

A cópula mais comumente empregada em todo tipo de análise estatística é a cópula normal ou gaussiana que, argumentaremos mais adiante, postula a mínima dependência linear entre duas variáveis:

$$N_\rho(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \mathrm{d}u\mathrm{d}v \; e^{-\frac{u^2+v^2-2uv\rho}{2(1-\rho^2)}} \tag{2.15}$$

onde $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp -u^2/2\mathrm{d}u$ é a distribuição cumulativa normal padronizada. A família de cópulas normais depende apenas de um parâmetro $\rho \in [-1,1]$ que, *no caso em que sejam inseridas marginais gaussianas para formar uma distribuição conjunta*, é igual à correlação entre as variáveis assim distribuidas. É importante notar que, para quaisquer outras marginais, a correlação poderá uma função do parâmetro ρ e de quaisquer outros parâmetros dessas distribuições marginais. Essa família contém a cópula de variáveis independentes $C(u,v) = uv$ quando $\rho = 0$.

Uma cópula ligeiramente mais complicada que a normal é a cópula associada à distribuição t de Student que, além da dependência linear descrita pelo parâmetro ρ, apresenta também dependência nas caudas da distribuição. A distribuição t de Student bivariada padrão[8] é dada por:

$$p_T(x,y \mid \rho,\nu) = \frac{\Gamma(1+\frac{\nu}{2})}{\Gamma(\frac{\nu}{2})\pi\nu\sqrt{1-\rho^2}} \left[1 + \frac{q_\rho(x,y)}{\nu}\right]^{-(1+\frac{\nu}{2})}, \tag{2.16}$$

onde $q_\rho(x,y) = \frac{1}{1-\rho^2}\left[x^2 + y^2 - 2\rho xy\right]$. As marginais dessa distribuição são distribuições t univariadas:

$$p(x_i|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{x_i^2}{\nu}\right)^{-(\frac{\nu+1}{2})} \tag{2.17}$$

No limite $\nu \to \infty$ essa distribuição se reduz à distribuição normal bivariada padronizada, com parâmetro de correlação linear ρ, médias nulas e variâncias unitárias. Para ν finito as marginais adquirem caudas pesadas e a dependência entre as variáveis ganham uma componente além da correlação linear - adicionando peso nas caudas da cópula. A cópula t pode ser obtida introduzindo essa distribuição na eq. (2.14):

$$C_T(u,v|\nu,\rho) = \frac{\Gamma(1+\frac{\nu}{2})}{\Gamma(\frac{\nu}{2})\pi\nu\sqrt{1-\rho^2}} \int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(u)} \mathrm{d}x\mathrm{d}y \left[1 + \frac{q_\rho(x,y)}{\nu}\right]^{-\frac{\nu+2}{2}} \tag{2.18}$$

onde $t_\nu(x)$ é a distribuição cumulativa associada à distribuição univariada em (2.17).

Uma terceira família de cópulas que convém citar são as cópu-

[8] Médias não-nulas e variância diferente da unitária podem ser trivialmente acrescentadas através de translações e mudanças de escala. Uma vez que essas transformações só afetam as marginais e mantém a cópula invariate - são trasformações inversíveis, coordenada a coordenada - não afetam a dependência

las arquimedianas, que podem ser escritas como:

$$C(u,v|\Psi(\cdot)) = \Psi^{-1}\left(\Psi(u) + \Psi(v)\right) \tag{2.19}$$

parametrizadas por uma função ϕ. Essas funções cópula existem desde que: $\Psi(1) = 0$, $\lim_{x\to 0}\Psi(x) = \infty$, $\Psi'(x) < 0$ e $\Psi''(x) > 0$. Essa família contém a cópula de variáveis independentes quando $\Psi(x) = -\log(x)$.

2.3.2 *Dependência extrema - limites de Frechet-Hoeffding*

A definição de cópula permite tornar mais preciso o conceito de dependência extrema. As equações (2.1) e (2.13) em conjunto nos permitem concluir que, para duas variáveis independentes:

$$c(u,v) = 1, \tag{2.20}$$
$$C(u,v) = uv, \tag{2.21}$$

Para o caso de dependência completa é possível mostrar que[9] toda cópula está limitada por duas funções que representam dependência máxima, denominadas limites de Frechet-Hoeffding. Essas funções são:

[9]

$$W(u,v) = \max(0, u+v-1) \tag{2.22}$$
$$M(u,v) = \min(u,v) \tag{2.23}$$

Essas duas funções são elas próprias cópulas e limitam por cima e por baixo todas as outras cópulas possíveis:

$$W(u,v) \leq C(u,v) \leq M(u,v) \tag{2.24}$$

para qualquer possível cópula $C(u,v)$. As densidades de cópula associadas a essas duas funções evidenciam que casos descrevem:

$$w(u,v) = \delta(u+v) \tag{2.25}$$
$$m(u,v) = \delta(u-v). \tag{2.26}$$

Inserindo marginais F(x) e G(y) quaisquer, nota-se o tipo de distribuições conjuntas que essas cópulas geram: ambas descrevem duas variáveis com dependência monotônica - crescente no caso de M e decrescente no caso de W.

2.3.3 *Medidas de dependência revisitadas*

Os "axiomas" de Renyi a respeito de medidas de dependência podem ser revisitados e tornados mais precisos com o conceito de cópula

e cópulas extremas em mãos. Os itens 1, 2 e 5 ficam imediatamente satisfeitos se a medida de dependência em questão for funcional apenas da cópula e não das distribuições marginais. Além disso, os itens 3 e 4 podem ser reescritos em termos das cópulas extremas e da cópula independente. Podemos reescrever então esses requisitos da seguinte forma:

- Uma boa medida de dependência entre duas variáveis X e Y é um funcional $\mathcal{F} : C_2 \to \mathcal{R}$ que leva funções cópula $C_{XY}(\cdot, \cdot)$ em números reais e independe das distribuições marginais;
- atinge um valor mínimo, que será arbitráriamente escolhido como zero, se, e somente se, $C_{XY}(u, v) = uv$;
- atinge um valor máximo quando $C_{XY}(u, v) = W(u, v)$ ou $C_{XY}(u, v) = M(u, v)$.
- para $C_{XY}(u, v) = N_\rho(u, v)$, o funcional é um função monotônica crescente do parâmetro ρ.

A correlação linear falha em dois itens: é possível representar a correlação linear como função da cópula, mas não é possível eliminar sua dependência com as marginais:

$$\hat{\rho}_{X,Y} = \frac{1}{\sigma_X \sigma_Y} \int \int C(u, v) \mathrm{d}F_X^{-1}(u) \mathrm{d}F_X^{-1}(v).$$

Essa expressão depende das marginais explicitamente nas medidas de integração e implicitamente nas variâncias σ_i. Além disso, a correlação não atinge seus valor extremo sempre que a cópula escolhida como uma das cópulas de Frechet-Hoeffding - o valor assumido nesse caso depende das marginais específicas. Outras medidas apresentadas acima, o tau de Kendall (τ) e a correlação de postos de Spearman (ρ^S) podem ser facilmente escritas de forma a satisfazer todas os critérios acima:

$$\tau = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1 \tag{2.27}$$

$$\rho^S = 12 \int_0^1 \int_0^1 \left[C(u, v) - uv \right] du dv \tag{2.28}$$

O último requisito pode ser verificado notando-se que, para o caso de cópulas normais, essas expressões se reduzem a[10]:

$$\tau = \frac{2}{\pi} \arcsin(\rho), \tag{2.29}$$

$$\rho^S = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right). \tag{2.30}$$

Além disso, a Informação Mútua, como mostraremos na próxima seção, também satisfaz os requisitos acima.

[10] A expressão para o tau de Kendall vale para toda distribuição elíptica, incluindo a distribuição t de Student

2.4 *Entropia de Cópula*

PARA ESCREVER a informação mútua como um funcional da cópula basta recorrer à definição da densidade de cópula e a expressão da distribuição conjunta em termos desta, na eq. (2.13) que reproduzimos abaixo:

$$p_{XY}(x,y) = c(F_X(x), F_Y(y))p_X(x)p_Y(y).$$

Usando essa expressão na definição de informação mútua eq. (2.11):

$$I(X,Y) = \int \mathrm{d}x\,\mathrm{d}y\; p(x,y)\log\frac{p(x,y)}{p_X(x),p_Y(y))},$$

temos:

$$\begin{aligned} I(X,Y) &= \int p_X(x)\mathrm{d}x\, p_Y(y)\mathrm{d}y\; c(F_X(x),F_Y(y))\log\left[c(F_X(x),F_Y(y)\right] \\ &= \int \mathrm{d}u\mathrm{d}v\; c(u,v)\log c(u,v) \end{aligned}$$

e portanto:

$$I(X,Y) = \int\int \mathrm{d}u\mathrm{d}v\; c(u,v)\log c(u,v) = -S[c] \geq 0 \tag{2.31}$$

onde $S[c]$ é a entropia associada à distribuição conjunta $c(u,v)$, daqui por diante denominada *entropia de cópula* [11]. Esse resultado oferece ainda mais uma interpretação aos múltiplos significados da informação mútua - que ressoa diretamente com a definição de dependência apresentada no início deste capítulo - é o conteúdo de informação associado à dependência entre duas variáveis. A combinação desse resultado com a eq. (2.9) permite escrever[12]:

[11] ; and

[12] A expressão acima pode ser imediatamente generalizada para um número arbitrário de variáveis

$$H(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i) + H_{\mathrm{cop}} \tag{2.32}$$

decompondo então a entropia, o conteúdo informacional da distribuição conjunta, em parcelas devidas às características de cada uma das variáveis e uma parcela correspondente ao acoplamento entre essas variáveis. Eq. (2.31) também oferece um conveniente princípio para encontrar cópulas informacionalmente neutras segundo o princípio de máxima entropia[13]: a cópula menos informativa é a que postula a menor informação mútua possível entre as variáveis.

[13]

O ÚLTIMO PASSO para mostrar que a informação mútua satisfaz todos os critérios para ser uma boa medida de dependência é notar que, no caso de cópulas gaussianas:

$$I(X_1, X_2) = -\frac{1}{2}\log\left(1-\rho^2\right) \tag{2.33}$$

2.5 *"Excesso" de informação mútua*

2.5.1 *Correlação Linear vs. parâmetro*

A DISTRIBUIÇÃO DE MÁXIMA ENTROPIA que satisfaz vínculos de que correlação, médias, e variâncias de um par de variáveis X e Y assumam certos valores é a distribuição normal. Usando a decomposição da eq. (2.32), temos, uma vez que as entropias das marginais dependem apenas das variâncias e a informação mútua apenas da correlação[14]:

$$H(X_1, X_2) = h(\sigma_1) + h(\sigma_2) + \frac{1}{2}\log\left(1 - \hat{\rho}^2\right)$$

onde $h(\sigma)$ é a entropia de uma distribuição normal univariada com variância σ^2. Se essa é a maior possível entropia dada a correlação e variâncias, e uma vez que a informação mútua independe das variâncias[15], então o valor:

$$I_0(\hat{\rho}) = -\frac{1}{2}\log\left(1 - \hat{\rho}^2\right) \tag{2.34}$$

é um limite inferior para a informação mútua de qualquer par de variáveis que tenham correlação $\hat{\rho}$:

$$I_{XY} \geq I_0(\hat{\rho}_{XY}). \tag{2.35}$$

Dessa forma, representando formalmente em um plano I *vs.* $\hat{\rho}$ as possíveis distribuições conjuntas com cópula gaussiana[16], temos a figura 2.2.

NESSA FIGURA temos duas distribuições destacadas. Uma delas possui marginais gaussianas, sendo portanto uma distribuição normal bivariada e deve estar localizada sobre a curva $I_0(\hat{\rho})$, representada pela linha escura tracejada. Para essa distribuição o parâmetro ρ, uma boa medida de dependência se restrito a cópulas gaussianas, é exatamente igual à correlação linear $\hat{\rho}$. Se as marginais forem trocadas por marginais não-gaussianas, a informação mútua, independente das marginais, deve permanecer a mesma. Entretanto a correlação, como argumentado acima, deve mudar com a troca de marginais. Uma vez que uma mudança para valores maiores do módulo da correlação linear violaria a condição $I \geq I_0(\hat{\rho})$, a única alternativa é que o módulo da correlação diminua. Dessa forma, teríamos uma distribuição conjunta que tem cópula gaussiana com parâmetro ρ_0 e correlação linear $\hat{\rho} < \rho_0$. Uma tentativa de identificar o parâmetro ρ da cópula gaussiana com a correlação linear levaria a atribuir ao par de variáveis uma dependência menor - talvez muito menor - do que a real. Em muitas aplicações isso pode ser perigoso. Em particular em aplicações financeiras o risco

[14] Note a importância de separar conceitualmente o parâmetro correlação, que identifica uma certa distribuição na família de distribuições normais, do funcional são homônimos, o qual estamos chamado de "correlação linear" e que está definido para qualquer distribuição.

[15] Pois independe das marginais.

[16] Essa representação não é única. Está sendo empregada apenas como ilustração. Essas duas grandezas não formam um bom sistema de coordenadas para a variedade de distribuições com cópula gaussiana.

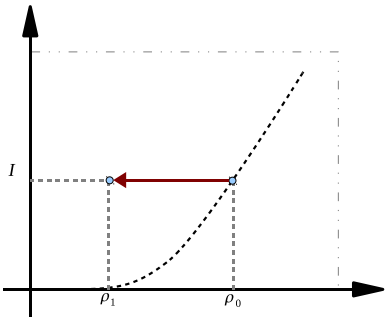


Figura 2.2 – A correlação linear é subestimada no caso de marginais não-Gaussianas. Se ambas as marginais são Gaussianas, a distribuição conjunta está localizada sobre o limite inferior para a informação mútua. Uma mudança nas marginais mantendo fixa a cópula, preserva a informação mútua, entretanto a correlação estimada deve se deslocar para valores de menor módulo.

de uma operação pode ser subestimado por se subestimar a frequencia de co-ocorrência de eventos negativos. É notório que séries temporais financeiras apresentam distribuições marginais com caudas pesadas - e portanto distantes de uma gaussiana. O uso, bastante difundido[17], de cópulas gaussianas para estimativa de risco combinado com o uso de estimadores baseados na correlação linear nessas condições pode significar que o risco de uma posição é substancialmente maior do que o estimado. Um dos requisitos para uma boa medida de dependência é que, para o caso de cópulas gaussianas, a medida em questão seja função monotônica e crescente do módulo do parâmetro ρ. Dessa forma, qualquer outra medida de dependência é mais adequada que a correlação linear para estimar a dependência de fato entre duas variáveis com dependência gaussiana.

[17]

2.5.2 *Informação mútua para cópulas elípticas*

O TAU DE KENDALL é um estimador ainda mais completo para ρ: ele é uma função monotônica do parâmetro ρ para qualquer cópula pertencente à família elíptica - da qual a gaussiana é um caso particular. Uma distribuição é dita elíptica sempre que:

$$\left\langle e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \right\rangle = e^{i\boldsymbol{\mu}^T\boldsymbol{k}}\psi\left(i\frac{\boldsymbol{k}^T\Sigma\boldsymbol{k}}{2}\right), \tag{2.36}$$

onde Σ é a matriz de covariância e $\boldsymbol{\mu}$ o vetor de valores esperados de $\boldsymbol{x}$. Uma distribuição elíptica padronizada é aquela em que as médias são nulas e todas as variâncias unitárias, de modo que a matriz de covariâncias é igual à matriz de correlações - $\Sigma_{ij} = \rho_{ij}$. Denotaremos essa família de distribuições por $p(\boldsymbol{x}|\Sigma, \psi(\cdot))$. No caso de $\Sigma_{ij} = \delta_{ij}$, a distribuição é invariante por rotações no vetor $\boldsymbol{x}$ e é dita uma *distribuição esférica* - na família $p(\boldsymbol{x}|\psi(\cdot))$.

É POSSÍVEL VER UMA VARIÁVEL DISTRIBUIDA com respeito a uma distribuição elíptica $\boldsymbol{x} \sim p(\boldsymbol{x}|\Sigma, \psi(\cdot))$ como uma transformação linear de variáveis distribuidas com relação à distribuição esférica com o mesma função $\psi(\cdot)$:

$$x_i = \sum_j A_{ij} y_j$$

Onde $\Sigma = A^T A$ e $\boldsymbol{y} \sim p(\boldsymbol{y}|\psi(\cdot))$. Cópulas elípticas são as cópulas associadas a essas distribuições e têm como parâmetros as correlações de pares - ρ_{ij} - e a função $\psi(\cdot)$. A eq.(2.36) é uma transformação de Fourier, que sempre pode ser invertida para obter a distribuição conjunta. No caso de distribuições elípticas padrão, temos:

$$p(\boldsymbol{x}|\Sigma, \psi(\cdot)) = \int \frac{\mathrm{d}^N k}{(2\pi)^N} e^{-i\boldsymbol{k}^T\boldsymbol{x}}\psi\left(i\frac{\boldsymbol{k}^T\Sigma\boldsymbol{k}}{2}\right)$$

podemos introduzir uma função delta e obter:

$$p(\boldsymbol{x}|\Sigma,\psi(\cdot)) = \int \mathrm{d}u \int \frac{\mathrm{d}^N k}{(2\pi)^N} e^{-i\boldsymbol{k}^T\boldsymbol{x}} \psi(u)\, \delta\left(u - i\frac{\boldsymbol{k}^T\Sigma\boldsymbol{k}}{2}\right)$$

e usando a representação integral da função delta:

$$\begin{aligned} p(\boldsymbol{x}|\Sigma,\psi(\cdot)) &= \int \frac{\mathrm{d}u\mathrm{d}\hat{u}}{2\pi} \int \frac{\mathrm{d}^N k}{(2\pi)^N} \psi(u) \exp\left[i\hat{u}\left(u - i\frac{\boldsymbol{k}^T\Sigma\boldsymbol{k}}{2}\right) - i\boldsymbol{k}^T\boldsymbol{x}\right] \\ &= \int \frac{\mathrm{d}u\mathrm{d}\hat{u}}{2\pi} e^{i\hat{u}u}\psi(u) \int \frac{\mathrm{d}^N k}{(2\pi)^N} \exp\left[\hat{u}\frac{\boldsymbol{k}^T\Sigma\boldsymbol{k}}{2} - i\boldsymbol{k}^T\boldsymbol{x}\right] \end{aligned}$$

A integral gaussiana sobre $\boldsymbol{k}$ pode ser feita e temos:

$$\begin{aligned} p(\boldsymbol{x}|\Sigma,\psi(\cdot)) &= \int \frac{\mathrm{d}u\mathrm{d}\hat{u}}{2\pi} e^{i\hat{u}u}\psi(u)\, N(x|\hat{u}\Sigma) \\ &= \int \mathrm{d}u\; p(u) N(x|u\Sigma) = \end{aligned}$$

onde $N(x|\Sigma)$ é a distribuição normal padronizada com matriz de correlação Σ e $p(u) = \int \mathrm{d}u e^{ivu}\psi(v)$ é uma certa função ligada à transformada de Fourier de $\psi(u)$. Essa representação para as distribuições elípticas pode ser entendida da seguinte forma: $p(\boldsymbol{x}|\Sigma,\psi(\cdot))$ é a distribuição resultante quando se toma $\boldsymbol{x}$ de uma gaussiana de matriz de correlação $u\Sigma$ em que u é sorteado de acordo com uma distribuição $p(u)$. Dessa forma podemos, de forma alternativa, parametrizar a família elíptica pela distribuição $p(u)$ - $p(\boldsymbol{x}|\Sigma,p(\cdot))$. Também podemos escrever de forma mais explícita:

$$p(\boldsymbol{x}|\Sigma,p(\cdot)) = \frac{1}{\sqrt{(2\pi)^N|\Sigma|}} \int \mathrm{d}^N u\; \frac{1}{u^{N/2}} p(u) e^{-\frac{1}{2u}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}} \tag{2.37}$$

$$= \frac{1}{\sqrt{(2\pi)^N|\Sigma|}} \Psi_N\left(-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right) \tag{2.38}$$

onde definimos: $\Psi_j(q) = \int \mathrm{d}^j u\; \frac{1}{u^{j/2}} p(u) e^{-\frac{q}{u}}$. As marginais de uma distribuição elíptica podem ser facilmente calculadas notando-se que as marginais da normal padronizada são distribuições normais padronizadas sobre 1 variavel. Dessa forma:

$$p_j(x_j) = \frac{1}{\sqrt{2\pi}} \int \mathrm{d}u\; \frac{1}{\sqrt{u}} p(u) e^{-\frac{x_j^2}{2u}} = \frac{1}{\sqrt{2\pi}} \Psi_1\left(\frac{x_j^2}{2}\right) \tag{2.39}$$

Finalmente de posse dessas das eqs. (2.38) e (2.39) podemos mostrar a seguinte proposição.

Proposição 1 (Decomposição da informação mútua de uma cópula elíptica)**.** *Se $C(u,v|\Sigma,p(\cdot))$ é uma cópula elíptica com matriz de correlação Σ, então a informação mútua associada pode ser decomposta na forma:*

$$I(\Sigma,\psi(\cdot)) = I_0(\Sigma) + I[p(\cdot)], \tag{2.40}$$

onde $I_0(\Sigma) = -\frac{1}{2}\log\Sigma$ *é a informação mútua de uma cópula gaussiana com matriz de correlação* Σ *e* $I[p(\cdot)]$ *é um funcional de* $p(u)$ *que é igual à informação mútua da distribuição esférica correspondente e independe da matriz de correlação.*

Demonstração. Para mostrar essa proposição recorremos à eq.(2.9):

$$I = \sum_i^N H[X_j] - H[\boldsymbol{x}] = NH[X_1] - H[\boldsymbol{x}]$$

onde a segunda igualdade é obtida notando que as marginais $p_j(\cdot)$ são todas idênticas. A primeira parcela já é um funcional de p(u) que independe da matriz de correlação que pode ser escrito como:

$$\begin{aligned} NH[X_1] &= N\int \mathrm{d}x \frac{1}{\sqrt{2\pi}}\Psi_1\left(\frac{x^2}{2}\right)\log\left[\frac{1}{\sqrt{2\pi}}\Psi_1\left(\frac{x^2}{2}\right)\right] \\ &= \int \mathrm{d}^N x \frac{1}{\sqrt{(2\pi)^N}}\Psi_N\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\log\left[\frac{1}{\sqrt{(2\pi)^N}}\prod_j\Psi_1\left(\frac{x_j^2}{2}\right)\right] \end{aligned}$$

A segunda parcela pode ser explicitamente escrita como:

$$H[\boldsymbol{x}] = -\int \mathrm{d}^N x \frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\Psi_N\left(-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right)\log\left[\frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\Psi_N\left(-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right)\right]$$

A inversa da matriz de correlação é simétrica e quadrada, e portanto sempre pode ser escrita como $\Sigma^{-1} = U^T\Lambda U$, onde U é uma matriz unitária e $\Lambda_{ij} = \delta_{ij}\lambda_j$ é uma matriz diagonal dos autovalores de Σ^{-1}. Fazendo a mudança de variáveis $\boldsymbol{y} = U\boldsymbol{x}$ temos:

$$H[\boldsymbol{x}] = -\int \mathrm{d}^N y \frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\Psi_N\left(-\frac{1}{2}\boldsymbol{y}^T\Lambda\boldsymbol{y}\right)\log\left[\frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\Psi_N\left(-\frac{1}{2}\boldsymbol{y}^T\Lambda\boldsymbol{y}\right)\right]$$

A matriz diagonal Λ pode ser escrita como $\Lambda = D^T D$, onde $D_{ij} = \delta_{ij}\sqrt{\lambda_j}$. Fazendo a mudança de variáveis[18] $\boldsymbol{x} = D\boldsymbol{y}$, temos:

$$H[\boldsymbol{x}] = -\int \mathrm{d}^N x \frac{1}{\sqrt{(2\pi)^N}}\Psi_N\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\log\left[\frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\Psi_N\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\right]$$

Uma vez que $|D| = \prod_i\sqrt{\lambda_i} = \frac{1}{\sqrt{|\Sigma|}}$. O termo que contém $|\Sigma|$ dentro do logaritmo pode ser removido da integral[19] e ficamos com:

$$H[\boldsymbol{x}] = -\int \mathrm{d}^N x \frac{1}{\sqrt{(2\pi)^N}}\Psi_N\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\log\left[\frac{1}{\sqrt{(2\pi)^N}}\Psi_N\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\right] + \frac{1}{2}\log|\Sigma|$$

[18] Vamos renomear a nova variável de integração como $\boldsymbol{x}$ novamente por conveniência

[19] Note que a normalização de p($\boldsymbol{x}$) exige que $\int \mathrm{d}^N x\psi_N(\boldsymbol{x}^T\boldsymbol{x}) = \sqrt{(2\pi)^N}$

e finalmente podemos escrever[20]:

$$I = I[p(u)] - \frac{1}{2}\log|\Sigma| \tag{2.41}$$

onde:

$$I[p(u)] = \int \mathrm{d}^N x \frac{1}{\sqrt{(2\pi)^N}} \Psi_N\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right) \log\left[\frac{\Psi_N\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)}{\prod_j \Psi_1\left(\frac{x_j^2}{2}\right)}\right] \tag{2.42}$$

□

Essa proposição nos permite escrever a informação mútua em duas parcelas - uma dependente da estrutura linear de dependência, relacionada à matriz de correlações, e uma parcela que contém informações sobre dependências não-lineares entre as variáveis. Uma vez que a informação mútua é sempre positiva, esse resultado nos permite ainda escrever uma versão mais forte da desigualdade eq.(2.35) para o caso de cópulas elípticas:

$$I_{XY} \geq I_0(\Sigma_{XY}) \tag{2.43}$$

onde, nesse caso, Σ não é apenas a matriz de correlações lineares, mas o conjunto de parâmetros que identifica unicamente uma certa distribuição dentro de uma sub-família de cópulas elípticas com mesma função $\psi(\cdot)$.

2.5.3 *Excesso de Informação Mútua*

As inequações eq.(2.35) e eq.(2.43) permitem concluir que a cópula gaussiana é a cópula de menor entropia dado o vínculo de dependências lineares representados por $\hat{\rho}$ ou Σ. Em outras palavras, a cópula gaussiana é a cópula que assume que uma variável tem a menor quantidade possível de informação a respeito de outra que ainda explica a parte linear da dependência. Além disso, a cópula gaussiana possui apenas a parte linear da informação mútua, e portanto representa uma dependência apenas linear entre as variáveis. O uso dessa cópula portanto representa uma hipótese implícita de dependência linear e mínima entre as variáveis. Caso essa hipótese falhe, um termo adicional deve surgir na informação mútua, que diz respeito à dependência não-linear. Esse termo será chamado "excesso" de informação mútua - entenda-se excesso com relação à cópula linear. A observação de um excesso de informação mútua permite criar um diagnóstico de "gaussianidade" da dependência entre duas variáveis. Para tal é necessário ser capaz de estimar o parâmetro ρ da distribuição e a informação mútua. Para estimar ρ, empregaremos o tau de Kendall -

[20] No caso bivariado essa expressão se reduz à expressão 2.34

uma medida que independe das marginais e permite estimar:

$$\rho = \sin\left(\frac{\pi\tau}{2}\right). \tag{2.44}$$

O tau de Kendall pode ser estimado empiricamente com um algoritmo simples: dado um conjunto de pontos $\{\boldsymbol{x}_\mu \sim p(\boldsymbol{x})|\mu = 1, 2, \ldots, P\}$, temos:

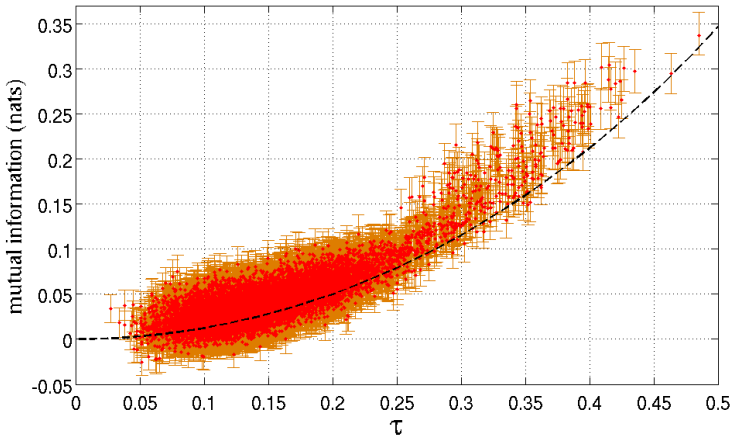
$$\tilde{\tau}[X_i, X_j] = \binom{N}{2}^{-1} \sum_{\mu<\nu} \operatorname{sign}(x_\mu^i - x_\nu^i)\operatorname{sign}(x_\mu^j - x_\nu^j) \tag{2.45}$$

onde a soma é feita sobre todos os pares (μ, ν). Um método para estimação da informação mútua foi recentemente publicado por **?**], baseado em estatísticas de k-vizinhos. Medindo I e τ é possível diagnosticar um eventual "excesso" de informação mútua que, caso presente, indica que a dependência não é mínima e tem uma componente não-linear. Nesses casos a correlação linear não pode ser usada como medida de dependência. Como exemplo desse diagnóstico, mostramos na figura 2.3 estimativas para essas quantidades para séries temporais de uma seleção de 150 das 500 ações que compõe o índice S&P500, um índice de ações de alta capitalização negociadas em bolsas da NYSE Euronext e da NASDAQ OMX definido e mantido pela Standard & Poor's. As barras de erro para a informação mútua, que representam um intervalo de confiança de 90%, foram calculadas usando o método de bootstrap, repetindo a estimativa do algoritmo KSG[21] para diversas amostragens com repetição dos dados. O tau de Kendall foi também estimado usando procedimentos padrão como o da eq.(2.45). Observa-se que uma boa quantidade de pontos apresentam, dentro do intervalo de confiança, um valor para informação mútua não compatível com uma cópula gaussiana. Isso sugere que técnicas de avaliação e administração de risco baseadas no uso de cópulas gaussianas podem subestimar de forma substancial a dependência entre duas ações e o grau de co-movimento que elas apresentam. Na figura 2.4 por exemplo, mostramos alguns pares de ações com correlações muito pequenas ($\rho < 0.1$) e que no entanto apresentam informação mútua compativel dependência muito maior do que a capturada por essa medida linear.

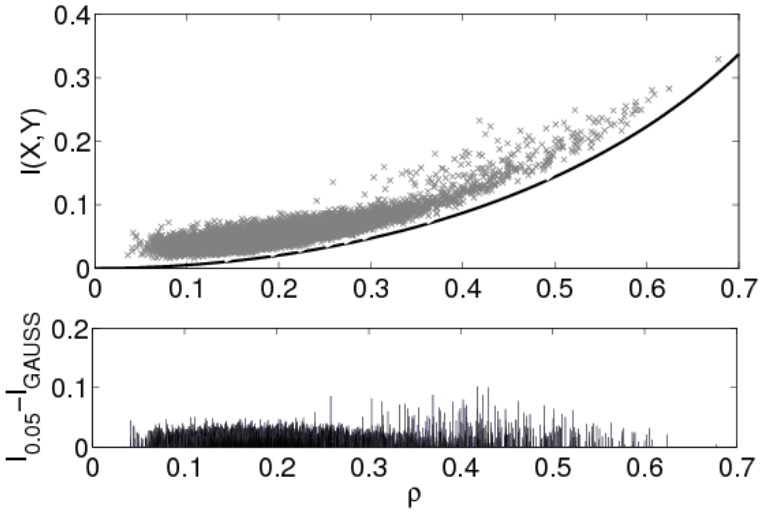
[21]

2.5.4 *Ajuste empírico de cópulas via informação mútua*

A determinação de uma particular cópula para realizar essas medidas de risco poderia ser uma alternativa ao uso cego da cópula gaussiana. Uma possível forma de realizar essa determinação é minimizar a divergência de Kullback-Leibler entre a cópula empírica e uma família $\mathcal{C}_\theta$ de cópulas parametrizadas por um conjunto de parâmetros

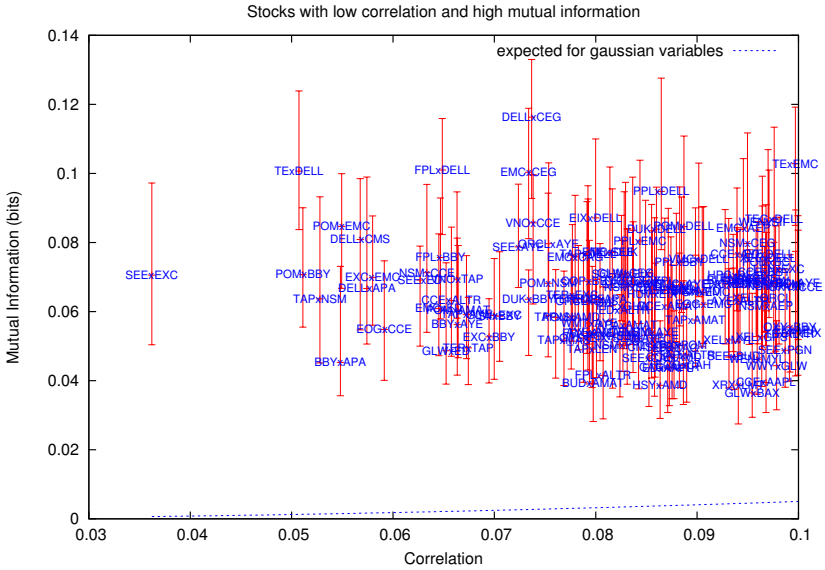


(a) Todos os pares de ações.

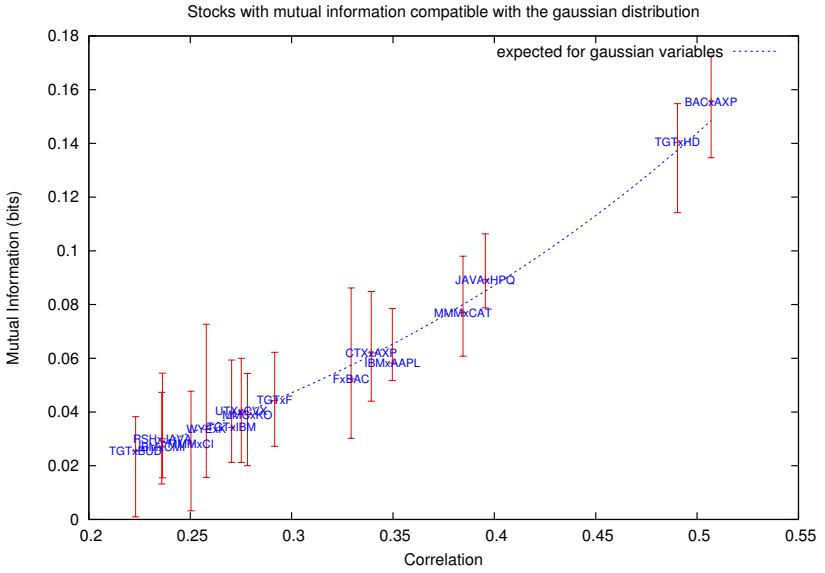


(b) Seleção dos pares cujo limite inferior da barra de erro é maior que a informação mútua gaussiana - excesso de informação não nulo com 90% de confiança (as barras de erro foram removidas para melhor visualização).

Figura 2.3 – Estimativas para a informação mútua usando o algoritmo KSG contra o tau de Kendall (ou correlação medida via tau de Kendall) para pares de séries temporais de log-retornos diários $\log \frac{P_{\text{close}}}{P_{\text{open}}}$ (onde P_{close} e P_{open} são, respectivamente, preços dos ativos na abertura e fechamento diários do mercado) para 150 das ações mais negociadas que compõe o índice *S&P*500, no período de 02/01/1990 a 16/09/2008 (aproximadamente 4700 pontos por série). As barras de erro representam intervalos de confiança de 90% determinados segundo o procedimento de Bootstrap. Note que, nesse intervalo de confiança, um grande número de pares apresentam um excesso de informação mútua não-nulo com respeito à cópula gaussiana. As linhas tracejadas indicam o limite gaussiano para a informação mútua.



(a) Seleção de pares com baixa correlação e grande excesso de informação mútua



(b) Seleção dos pares compatíveis com uma distribuição gaussiana com 90% de confiança.

Figura 2.4 – Seleção de pares de ações com baixa correlação e grande excesso de informação mútua e pares compatíveis com uma distribuição gaussiana. Modelos modelos de risco baseados em medidas de correlação linear devem ser adequados para o segundo grupo e devem falhar seriamente para o primeiro.

θ:

$$D[C(\boldsymbol{u})||C(\boldsymbol{u}|\theta)] = \int \mathrm{d}^N u \; c(\boldsymbol{u}) \log \frac{c(\boldsymbol{u})}{c(\boldsymbol{u}|\theta)} \tag{2.46}$$

A minimização desse funcional com respeito a θ fornece uma possível cópula $c(\boldsymbol{u}|\theta^\star)$ que é a mais próxima possível da cópula real dentro dessa família. Essa expressão pode ser manipulada da seguinte forma:

$$D[\cdot|\cdot] = \int \mathrm{d}^N u \; c(\boldsymbol{u}) \log c(\boldsymbol{u}) - \int \mathrm{d}^N u \; c(\boldsymbol{u}) \log c(\boldsymbol{u}|\theta) \tag{2.47}$$

$$= I - L_\infty(\theta) \tag{2.48}$$

onde o primeiro termo é o negativo da entropia de cópula, igual à informação mútua, como mostrado anteriormente, e o segundo termo é o valor assintótico da log-verossimilhança quando o número de amostras é grande:

$$L_N(\theta) = \frac{1}{N} \sum_{\mu=1} N \log c(\boldsymbol{u}_\mu|\theta)$$

Uma vez que $D[|] \geq 0$ e $I \geq 0$, então minimizar a divergência de Kullback-Leibler, cujo menor valor possível é nulo, é equivalente a maximizar a log-verossimilhança com a informação mútua como limite. Se a log-verossimilhança fosse conhecida analíticamente isso poderia ser feito de maneira imediata resolvendo a equação:

$$L_\infty(\theta) = I$$

Para um I determinado empiricamente a partir dos dados,numericamente se necessário. Não há garantia alguma, no entanto de que há solução. Apenas haverá solução para essa equação se a própria cópula original fizer parte da família $\mathcal{C}_\theta$. Nesse caso a solução é única. Além disso não é possível conhecer L_∞ analiticamente sem conhecer a cópula e uma aproximação é necessária. Supondo que a família de cópulas $\mathcal{C}_\theta$ é suficientemente próxima da cópula original, podemos substituir $c((u))$ pela própria $c(\boldsymbol{u}|\theta)$ na integral e aproximar $L_\infty(\theta)$ pela informação mútua de $c(\boldsymbol{u}|\theta)$. Dessar forma ficamos com uma espécie de método de "correspondência de momentos" (moment matching): deve ser escolhida na família $\mathcal{C}_\theta$ a cópula que tem mesma informação mútua que a empiricamente obtida:

$$I(\theta) = I.$$

Novamente, não há garantia de solução, e agora nem mesmo da unicidade da solução. Mas a cópula escolhida certamente será capaz de descrever uma estrutura de dependência mais complexa do que a descrita pela cópula gaussiana. Se a família escolhida for um subconjunto da famíla de cópulas elípticas, o procedimento pode ser ainda melhorado. Parte do conjunto de parâmetros θ corresponde às correlações Σ. Escrevendo $I(\theta) = I_0(\Sigma) + \Delta I(\theta')$ pode-se já eliminar a parte linear da informação mútua usando o tau de Kendall para calcular Σ e ajustar o excesso de informação mútua ao medido empiricamente.

2.5.5 *Cópula t*

Como exemplo desse procedimento vamos escolher uma subfamilia das cópulas elípticas, as cópulas t. Essas cópulas, como discutido anteriormente, são as cópulas que se originam da distribuição t de Student. Quando se permite que o parâmetro ν seja contínuo, essa é, em essência, a mesma distribuição obtida pela maximização da chamada entropia de Tsallis, e essa distribuição e sua cópula associada têm recebido certa atenção na literatura de análise financeira por seu bom ajuste empirico a dados de diversas naturezas[22]. O cálculo do excesso de informação mútua da distribuição t de Student se resume a calcular a eq.(2.42) - a informação mútua da distribuição t de Student esférica:

[22] ; and

$$p(\mathbf{t} \mid \hat{\Sigma}, \nu) = \frac{1}{Z_N(\nu)} \left[1 + \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}{\nu}\right]^{-\frac{\nu+N}{2}} \tag{2.49}$$

onde a normalização é dada por:

$$Z_N(\nu) = \int \mathrm{d}^N x \ \left[1 + \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}{\nu}\right]^{-\frac{\nu+N}{2}} = \frac{B(\frac{\nu}{2}, \frac{N}{2})\sqrt{(\pi\nu)^N}}{\Gamma(\frac{N}{2})} \tag{2.50}$$

Uma vez que todas as marginais são idênticas[23] a informação mútua se reduz a:

[23] Todas são iguais à distribuição t de Student em uma dimensão com o mesmo parâmetro ν

$$I(\nu) = NH_1(\nu) - H_N(\nu), \tag{2.51}$$

onde $H_n(\nu)$ é a entropia de uma distribuição de student n-dimensional. Para calcular H_n note que:

$$\begin{aligned}
H_n(\nu) &= -\int \mathrm{d}^n x \frac{1}{Z_n(\nu)} \left(1 + \frac{\boldsymbol{x}^T\boldsymbol{x}}{\nu}\right)^{-\frac{\nu+n}{2}} \left[-\frac{\nu+n}{2}\log\left(1 + \frac{\boldsymbol{x}^T\boldsymbol{x}}{\nu}\right) - \log Z_n(\nu)\right] \\
&= \log Z_n(\nu) + \frac{\nu+n}{2Z_n(\nu)} \int \mathrm{d}^n x \left(1 + \frac{\boldsymbol{x}^T\boldsymbol{x}}{\nu}\right)^{-\frac{\nu+n}{2}} \log\left(1 + \frac{\boldsymbol{x}^T\boldsymbol{x}}{\nu}\right) \\
&= \log Z_n(\nu) + \frac{(\nu+n)\Omega_n}{2Z_n(\nu)} \int_0^\infty \mathrm{d}r r^{n-1} \left(1 + \frac{r^2}{\nu}\right)^{-\frac{\nu+n}{2}} \log\left(1 + \frac{r^2}{\nu}\right) \\
&= \log Z_n(\nu) + \frac{(\nu+n)\Omega_n\sqrt{\nu^n}}{2Z_n(\nu)} \int_0^\infty \mathrm{d}u u^{n-1} \left(1 + u^2\right)^{-\frac{\nu+n}{2}} \log\left(1 + u^2\right) \\
&= \log Z_n(\nu) + \frac{(\nu+n)\Omega_n\sqrt{\nu^n}}{2Z_n(\nu)} R_n(\nu)
\end{aligned}$$

Onde $\Omega_n = \frac{2\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)}$ é a área de uma esfera unitaria em n dimensões e

$$R_n(\nu) = \int_0^\infty \mathrm{d}u u^{n-1} \left(1 + u^2\right)^{-\frac{\nu+n}{2}} \log\left(1 + u^2\right).$$

Essa integral pode ser feita com o auxílio de um truque similar ao truque de réplicas comum em mecânica estatística. Se notarmos que $\log(x) = \lim_{r\to 0} \frac{\partial}{\partial r} x^r$ podemos escrever[24]:

[24] Pois:

$$\int_0^\infty \mathrm{d}u u^{n-1} \left(1 + u^2\right)^{-\alpha} = B\left(\alpha - \frac{n}{2}, \frac{n}{2}\right),$$

$$\frac{\partial}{\partial x} B(x, y) = -B(x, y)(\psi(x+y) - \psi(x))$$

$$R_n(\nu) = \lim_{r\to 0} \frac{\partial}{\partial r} B\left(\frac{\nu}{2} - r, \frac{n}{2}\right) = -B\left(\frac{\nu}{2}, \frac{n}{2}\right)\left[\psi\left(\frac{\nu+n}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right]$$

onde $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ é a função beta e $\psi(x)$ é a função digamma, substituindo esse resultado na expressão original e escrevendo todos os termos explicitamente temos:

$$H_n(\nu) = \log\left[\frac{\sqrt{(\pi\nu)^n} B\left(\frac{\nu}{2}, \frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\right] + \frac{\nu+n}{2}\left[\psi\left(\frac{\nu+n}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right] \tag{2.52}$$

Para o caso $n = 1$ isso se reduz a:

$$H_1(\nu) = \log\left[\sqrt{\nu} B\left(\frac{\nu}{2}, \frac{1}{2}\right)\right] + \frac{\nu+1}{2}\left[\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right] \tag{2.53}$$

E a informação mútua $I = NH_1(\nu) - H_N(\nu)$ finalmente pode ser escrita:

$$I = \log\left\{\frac{\left[B\left(\frac{\nu}{2}, \frac{1}{2}\right)\right]^N \Gamma\left(\frac{N}{2}\right)}{\pi^{\frac{N}{2}} B\left(\frac{\nu}{2}, \frac{N}{2}\right)}\right\} - \frac{\nu(N-1)}{2}\psi\left(\frac{\nu}{2}\right) + \frac{N(\nu+1)}{2}\psi\left(\frac{\nu+1}{2}\right) - \frac{\nu+N}{2}\psi\left(\frac{\nu+N}{2}\right).$$

Para $N = 2$ isso se reduz a:

$$I(\nu) = 2\log\left(\sqrt{\frac{\nu}{2\pi}} B\left(\frac{\nu}{2}, \frac{1}{2}\right)\right) - \frac{2+\nu}{\nu} + (1+\nu)\left[\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right] \tag{2.54}$$

Finalmente, a expressão acima pode ser usada, segundo o método discutido na seção 2.5.4, para ajustar cópulas t a dados empíricos. Como ilustração, a figura 2.5 apresenta uma série de simulações de ajuste com dados sorteados de uma cópula t com diversos valores de correlação e ν conhecidos. O excesso de informação mútua é estimado usando o algoritmo KSG [25] e o tau de Kendall e plotado em função do ν conhecido. A linha cheia corresponde à eq.(2.54). Esse gráfico mostra que, exceto por um ponto que não pode ser recuperado[26], é possível recuperar cópulas t a partir de dados experimentais através desse procedimento.

[25]

[26] Para um valor muito pequeno de ν, para o qual a distribuição t começa a apresentar diversas patologias, como variância infinita.

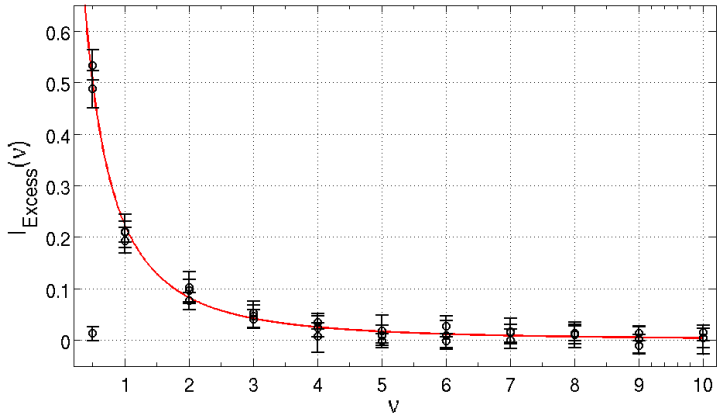


Figura 2.5 – Excesso de informação mútua na cópula t. $I(\nu)$, como dado na eq.(2.54). Círculos mostram estimativas para 20 amostragens de pontos de uma cópula t usando o método de "moment matching" para diversos valores de correlação e ν.

2.6 *Conclusões*

A literatura em teoria de informação e teoria de cópulas e dependência estatística - ambas com décadas de existência - se desenvolveram em relativo isolamento, com apenas pontos muito recentes de contato. Neste trabalho tentamos discutir as consequencias de alguns desses pontos e conexões entre os dois tópicos. A teoria de cópulas pode ser usada para decompor as distribuições conjuntas em flutuações idiossincráticas das marginais de cada variável e flutuações devidas ao acoplamento entre as variáveis. A conexão com a teoria de informação permite levar adiante essa decomposição e isolar contribuições à informação total contida na distribuição em partes relativas às marginais, ao acoplamento de natureza linear e acoplamentos de mais alta ordem. Essa decomposição oferece testes simples a respeito da linearidade e "gaussianidade" do acoplamento e também sugere um método de ajuste de cópulas baseados no ajuste da informação mútua. Essa abordagem também clarifica os perigos do uso

da correlação linear como medida de dependência em séries financeiras para, por exemplo, estimativas de riscos de contratos complexos e otimização de carteiras, pois essa medida é fadada a subestimar a dependência em séries em que flutuações não-gaussianas são esperadas. Finalmente, pensamos que uma conexão entre essas duas áreas - teoria de informação e teoria de dependência estatística - pode ser útil em fornecer conceitos e técnicas novas para o estudo de sistemas complexos.

3 | *Um Modelo para emergência de autoridade em sociedades humanas.*

3.1 Introdução

Os GRANDES PRIMATAS, em particular os humanos, apresentam vidas sociais intensas. Atividades sociais, formação de coalizões, cooperação para realização de diversas tarefas, guerras, compartilhamento de alimentos, disputas por liderança *et cetera*, situações ubíquas em agrupamentos humanos, não são entretanto limitadas a essa espécie mas são pervasivas em todas as espécies desse grupo[*more sources needed*]. Estudos em chimpanzés e bonobos[1] [*more sources needed*], por exemplo, mostram que a variedade de suas experiências sociais são comparáveis apenas à da espécie humana.

[1]

Entretanto a natureza dessa experiência social pode ser bem diferente entre humanos e outras espécies aparentadas. Enquanto a maioria dos grandes primatas vivem em sociedades hierárquicas, marcadas por uma grande concentração dos usos de recursos energéticos e reprodutivos por parte de poucos membros do grupo, os humanos se destacam pela variabilidade de suas experiências sociais nesse espectro. Certos grupos humanos[2] apresentam organização fortemente centralizada e hierárquica, com concentração de uso de recursos e riqueza. Outros grupos apresentam sociedades basicamente equalitárias, com compartilhamento de recursos e alimentos, ausência de distinções de status, autoridades ou concentração de riqueza. [*more sources needed*]

[2] Dizer aqui que nos preocupamos com a evolução dos humanos e portanto vamos nos fixar em hunter-gatherers e sociedades simples.[*this must be better explained*]

Nossa intenção é desenvolver um modelo matemático para a formação de estruturas sociais baseado em recentes observações nos campos da arqueologia, primatologia e neurociências.

3.2 *Evidências empiricas*

3.2.1 *"U-shaped evolution" e dinâmica da organização social em primatas pré-humanos*

O registro arqueológico revela uma dinâmica temporal na organização social dos humanos através da pré-história. Os humanos, descendendo de primatas com provável organização social hierárquica, similar às dos grandes primatas mais próximos - chimpanzés, bonobos e gorilas - passaram por um período de grupos equalitários, sem autoridade central, com baixa densidade populacional. No neolítico houve uma transição para grupos fortemente hierárquicos, conforme a densidade populacional aumenta após a revolução agrícola. Esse quadro é ilustrado pela figura 3.1. Evidências etnográficas também apontam para uma relação entre o tamanho dos grupos de humanos caçadores-coletores e suas formas de organização social[3]. Grupos pequenos de humanos tendem a apresentar organização equalitária, sem concentração de poder. Grupos maiores tendem a apresentar organizações hierárquicas, concentração de poder e hereditariedade de poder.

[3]

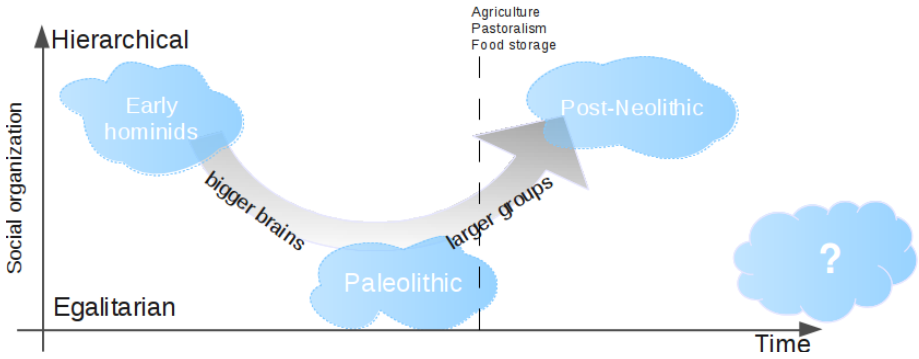


Figura 3.1 – Ilustração da história da organização social dos humanos e primatas pré-humanos.

[4] ; ; ; ; and

3.2.2 *Evolução do cérebro primata e a Teoria Maquiavélica*

Diversos trabalhos[4] relacionam o tamanho relativo de regiões do cérebro de diversas espécies de primatas a medidas relacionadas com a capacidade social da espécie, como tamanho dos grupos em que vivem, o tamanho de coalizões, número médio de indivíduos que interagem diretamente, etc. O que é típicamente encontrado é ilustrado na figura 3.2. Essa figura mostra um gráfico do tamanho médio do grupo em função da razão média entre o volume do neo-córtex - região do cérebro dedicada ao planejamento, raciocínio, linguagem, entre outras funções cognitivas de ordem superior - e o volume total do cérebro para diversas espécies de primatas. O gráfico sugere uma relação do

tipo lei de potência entre as duas grandezas, similar a encontradas em diversas outras comparações desse tipo. Em essência, essa relação sugere que a capacidade cognitiva dos primatas está intimamente relacionada a sua necessidade de dar conta de interações sociais cada vez mais complexas e sugere um cenário em que o rápido crescimento na importância relativa do neo-córtex é uma resposta a uma pressão seletiva associada a essa necessidade de interação social.

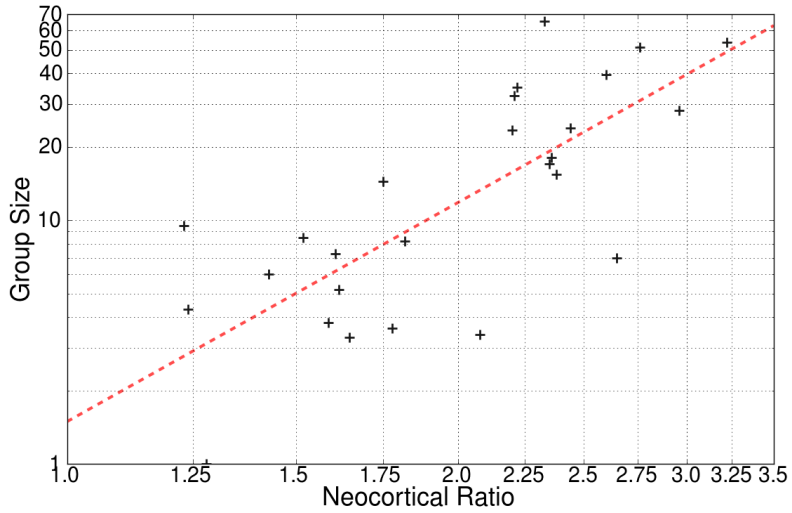


Figura 3.2 – Gráfico em escala dilogaritimica do tamanho médio do grupo em função da razão média entre o volume do neo-córtex e o volume total do cérebro para diversas espécies de primatas. Dados disponíveis em [**?**].

3.2.3 *Dados etnoráficos (?)*

3.3 *Um modelo mecanico-estatístico baseado em agentes*

3.3.1 *Descrição dos agentes*

O objetivo dessa seção é descrever um modelo estatístico para a formação de estrutura social que seja tratável por técnicas comuns à mecânica estatística e teoria de informação e compatível com as informações experimentais descritas acima. O primeiro passo na descrição do modelo consiste na elaboração de uma dinâmica de comportamento para um conjunto de agentes hipotéticos, que será a dinâmica temporal microscópica que dará origem a um modelo mecânico-estatístico.

Considere um grupo de n agentes dotados de certa capacidade cognitiva limitada e engajados em atividades sociais. Cada agente carrega um registro mental da informação que possui a respeito das relações sociais entre os membros do seu grupo. Essa informação está

relacionada a como cada par de outros agentes do grupo se relaciona socialmente. Essa informação deve responder perguntas como:

- qual a possibilidade de esse par de individuos serem adversários em uma disputa ou aliados em uma coalizão?
- com que frequencia cooperam em uma atividade conjunta?
- como compartilham seus recursos um com o outro?
- etc...

Cada agente adquire essa informação através de mecanismos diversos: através da história social do grupo, baseado no comportamento pregresso dos agentes; através de mecanismos de aprendizado social como fofoca (*gossip*) em que a comunicação com outros agentes permite que o agente aprenda sobre experiências de outros; etc. Uma vez adquirida, essa informação é critica para subsidiar decisões sociais a serem tomadas pelo agente: com que grupo de agentes fazer uma coalizão, quando esperar cooperação de um certo indivíduo, etc. Erros em decisões podem custar recursos e posição social, e influenciar negativamente a capacidade reprodutiva do agente. Portanto, espera-se que uma espécie de agentes que tenha surgido por evolução via seleção natural tenha mecanismos cognitivos adequados para tentar minimizar esses erros em algum sentido.

[5] ; and

A AQUISIÇÃO DESSE TIPO DE INFORMAÇÃO SOCIAL é uma atividade cognitivamente custosa. A capacidade limitada de processar essas informações implica que para mantê-las atualizadas, o indivíduo precisa desviar recursos que poderiam ser aplicados em outras atividades - coleta de alimentos, construção de abrigos, etc. Há evidências [5] [*more sources needed*]de que o rastreamento de relações sociais demanda um considerável tempo dos individuos adultos em tribos de chimpanzés e humanos. Conforme se aumenta o tamanho do grupo, esse custo cresce com o número de ligações sociais possíveis e, portanto, quadraticamente com o número de indivíduos. Isso pode pode tornar a estratégia de adquirir e manter informações sobre todas as ligações sociais possíveis no grupo pouco adaptativa. Pode ser preferível ao agente nesse caso obter apenas informação sobre certas ligações sociais importantes e fiar-se em heurísticas para inferir as outras relações sociais (regras como "amigo do amigo é amigo", etc...).

DESSA FORMA, O MODELO CONSISTIRÁ dos seguintes elementos: muitos *agentes* que interagem entre si, cada um carregando uma *representação mental da estrutura social* do grupo a que pertence e individualmente

tentando *minimizar custos* associados a carregar esass informações sociais. Abaixo discutiremos uma representação matemática de cada um desses elementos.

3.3.2 *Variáveis dinâmicas - grafos sociais*

A REALIZAÇÃO MATEMÁTICA da representação mental da estrutura social que cada agente carrega será feita através de grafos. A cada agente está associado um grafo cujos nós representam todos os indivíduos do grupo e arestas representam as relações sociais sobre as quais o agente possui informação. A informação será considerada binária: o agente pode ter certeza sobre a relação social entre dois individuos, havendo portanto uma aresta ligada entre os nós correspondentes de seu grafo, ou não tem nenhuma informação direta sobre ela, caso em que não haverá uma aresta enter os nós correspondentes. As arestas do grafo portanto podem estar apenas ligadas ou desligadas, sem estados intermediários. As arestas do grafo são entidades dinâmicas, que podem ser criadas quando o agente adquire informação sobre uma relação social anteriormente desconhecida, ou destruídas quando o agente, por alguma razão, desiste de continuar mantendo aquela informação. Dados os custos, que serão discutidos abaixo, o agente deverá decidir quais informações valem a pena ser guardadas ou não. Um agente que possui informação completa sobre todas as relações sociais do grupo tem um grafo totalmente conectado, como o representado na figura 3.3.

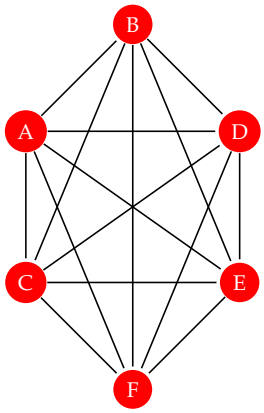


Figura 3.3 – Exemplo de grafo social - um grafo completamente conectado. Um agente com essa estratégia despende recursos para conhecer todas as relações sociais do grupo. Um grafo como esse possui $\frac{1}{2}n(n-1)$ arestas, onde n é o número de agentes.

QUANDO UMA ARESTA É FALTANTE no grafo carregado por um agente, a informação social correspondente à essa aresta é incompleta, e o agente deve recorrer a heuristicas para determinar quaisquer informações necessárias. Para tal, vamos considerar que o grafo deve ser conexo - deve ser possível, para todos os pares de nós, encontrar um *caminho* de arestas ligadas conectando os dois nós. Dessa forma, sempre é possível a um agente determinar alguma informação indireta entre dois nós desconexos, através de uma heuristica que utilize as outras arestas conhecidas. Um exemplo é o grafo da figura 3.4, que representa um grafo do tipo estrela.

Nesse grafo, não é possível conhecer diretamente todas as relações sociais pois apenas $n-1$ das $\frac{1}{2}n(n-1)$ arestas possíveis está presente. Mas todas as relações sociais podem ser indiretamente determinadas por heurísticas sobre caminhos de comprimento 2 ("amigo do amigo é amigo", etc...). Os grafos mais esparsos possíveis que ainda são conexos possuem $n-1$ arestas, no mínimo.

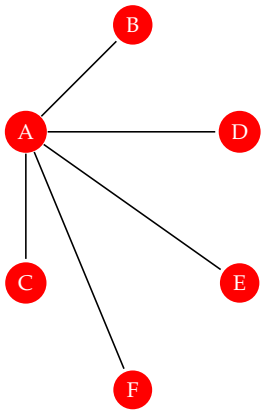


Figura 3.4 – Exemplo de grafo social - um grafo estrela. Um agente com essa estratégia despende recursos para conhecer apenas as relações envolvendo um certo indivíduo central (o nó A na figura). As outras relações são determinadas através de regras heurísticas. Esse grafo possui $n-1$ arestas.

3.3.3 *Custos*

Em nosso modelo existem custos associados à manutenção de um certo grafo de informações sociais. Há dois tipos de custos:

1. O custo cognitivo de adquirir e manter informação dos diferentes pares de individuos. Como discutido anteriormente, se o agente tem recursos cognitivos limitados, manter essas informações é custoso. Se assumirmos que o investimento de recursos para obter informações sobre cada aresta do grafo é constante, o custo cognitivo total por agente deverá ser proporcional ao número de arestas do grafo:

$$H_{\text{cognitivo}} \propto n_e \tag{3.1}$$

   onde n_e é o número de arestas (*edges*).

2. O custo social de falhar em determinar corretamente a relação entre dois indivíduos. As heuristicas utilizadas pelo agente quando ele não possui informação direta sobre uma certa relação social podem falhar e, nesse caso, o agente pode inferir erroneamente a relação social entre dois agentes. Ao tomar decisões baseadas nessa avaliação errônea, o agente incorre em custos - o agente pode avaliar incorretamente em que lado de uma disputa um individuo vai se posicionar, falhar em reconhecer uma coalizão em formação, etc, e ter prejuízos reais com uma situação social inesperada. As heuristicas que se valem de relações conhecidas para inferir relações desconhecidas serão tão mais confiáveis quanto menor o caminho a ser percorrido no grafo entre os dois nós em questão. Quanto mais longos os trajetos a serem percorridos, maior é a probabilidade de erro. Portanto, o custo social esperado deve ser proporcional à distância geodésica média entre os nós do grafo:

$$H_{\text{social}} \propto \frac{2}{n(n-1)} \sum_{i<j} L_{ij} \tag{3.2}$$

   onde L_{ij} é a distância geodésica entre os nós i e j, e a soma é realizada sobre todos pares de agentes.

Assim, para cada agente, o custo total de se manter uma certa representação mental da rede social do grupo é dado, portanto, por:

$$H = \frac{n_e}{\alpha} + \bar{L} \tag{3.3}$$

Onde $\bar{L}$ é a distancia geodésica média do grafo, α é uma constante associada à importância relativa entre o custo cognitivo e o custo social

(quanto maior α, menos importante é o custo cognitivo). Note que tipicamente n_e escala como uma fração do número total de arestas possíveis $n(n-1)/2$ e que $\bar{L}$ escala tipicamente com $\log(n)$ para grafos aleatórios. É, portanto, interessante, reescrever (3.3), a menos de uma constante multiplicativa, como:

$$H = \phi_e + a\bar{L} \tag{3.4}$$

onde $\phi_e = \frac{n_e}{n(n-1)/2}$ é a fração de arestas ocupadas e $a = \frac{2\alpha}{n(n-1)}$ é uma constante de acoplamento normalizada.

3.3.4 *Dinâmica para agentes isolados - máxima entropia*

3.3.5 *Minimização do custo*

Para que um modelo possa ser estabelecido, não basta a expressão para o custo, mas uma descrição de que ações o agente deverá tomar com base em sua aferição do custo. A estratégia do agente é definida por uma certa escolha de arestas a investir. Em uma primeira abordagem, atribuimos ao agente a tendência a usar a estratégia que minimiza o custo total. Então deveriamos procurar pelo grafo definido por:

$$G_\star(\alpha) = \arg\min_G H(G, \alpha) \tag{3.5}$$

Se $\alpha \gg 1$, então as limitações cognitivas são menos e menos importantes. O grafo que minimiza o custo é o grafo que minimiza a distância geodésica média - um grafo totalmente conectado como o da figura 3.3. Se $\alpha \ll 1$, então as limitações cognitivas se tornam mais e mais importantes, e o grafo ótimo é aquele que minimiza o número de arestas, enquanto ainda mantendo uma distância média finita - um grafo em forma de estrela [6] como o da figura 3.4. Para valores intermediários de α, o grafo ótimo possui configurações intermediárias entre esses dois extremos.

[6] Estritamente para $\alpha = 0$ o grafo ótimo é o grafo sem aresta alguma. Não há prescrição canônica para a distância geodésica média de um grafo sem arestas. Adotaremos aqui a convenção de que se não é possível desenhar um caminho entre dois nós (em outras palavras: dois nós que pertencem a diferentes componentes do grafo), então a distância entre eles é infinita. Isso restringe nossa análise apenas a grafos conexos. Mesmo entre os grafos conexos, a estrela não é o único mínimo do custo acima para α estritamente zero - o grafo caminho também é possível. Entretanto, para o grafo caminho, $\bar{L}$ é proporcional a N, e portanto, para $\alpha \to +0$ o único mínimo existente quando $\alpha > 0$ tem grafo estrela como limite.

Podemos relaxar o vínculo de estrita minimização e propor o seguinte modelo: o agente decide sua estratégia através de uma dinamica interna que ocorre em seu cérebro, ditada por regras que levam em conta o custo total. Podemos então associar probabilidades às estratégias de acordo com nossa expectativa de que grafos devem surgir dessa dinâmica. Se não conhecemos detalhes da dinâmica, mas temos informação de que o valor esperado do custo total é uma variável importante, o procedimento bayesiano adequado é associar ao grafo uma distribuição de máxima entropia, restringida pelo valor do custo total. A distribuição resultante é a distribuição de Gibbs:

$$p(G) = \frac{q(G)}{Z} e^{-\beta H(G)} \tag{3.6}$$

onde β, uma espécie de inverso de temperatura, regula a importância relativa dos efeitos que o custo $H(G)$ e outros efeitos negligenciados pelo modelo, e $q(G)$ é a atribuição de probabilidades a priori. Assumiremos probabilidades uniformes a priori, com a restrição de que o grafo seja conexo (veja nota [6]).A atribuição de probabilidades da equação (3.6) torna o problema de determinar as propriedades macroscópicas desse sistema em um problema de mecânica estatística, cuja variável dinâmica é a matriz de adjacências do grafo G:

$$M_{ij} = \begin{cases} 1 & \text{se } i \text{ e } j \text{ estão ligados por uma aresta} \\ 0 & \text{outro caso} \end{cases} \tag{3.7}$$

3.3.6 *Interpretação dos parâmetros α e β*

O parâmetro α (ou sua versão normalizada, a) regula a importância relativa entre os dois custos - social e cognitivo. O valor de α é regulado pela capacidade cognitiva dos agentes: quanto maior a capacidade de realizar cálculos sociais, menor é a importância do custo cognitivo, maior é o valor de α. No presente trabalho α será, portanto, interpretado como uma medida da capacidade cognitiva dos agentes. O parâmetro normalizado a é, portanto, a fração da capacidade cognitiva que deveria ser gasta para aprender todas as relações sociais do grupo.

O parâmetro β regula a escala em que flutuações no valor do custo total são toleradas. Para β grande, variações no custo total acima do custo mínimo são muito pouco prováveis. Para valores maiores de β as mesmas flutuações apresentam probabilidades maiores. Dessa forma, β regula o quão prováveis são configurações que consistem de flutuações em torno das configurações ótimas discutidas na seção *Minimização do custo*. Flutuações em torno do custo mínimo representam um dispêndio extra de energia e recursos que poderiam ser gastos em outras atividades, portanto β pode ser parcialmente entendido como uma variável ecológica - escassez de recursos implica em menor possibilidade de permitir grandes flutuações do custo total. Além disso, pressões de pares pode causar o mesmo tipo de efeito de intolerância a erros de minimização do custo ***[more sources needed]***. Assim, β é uma variável externa, que regula a intensidade da pressão para otimização do custo total, que pode ter origem social ou ecológica.

3.3.7 *Resultados numéricos para agentes isolados*

Simulações de Monte Carlo desse modelo foram feitas usando o algoritmo de Metrópolis. Partindo de um grafo inicial aleatório (sorteado do ensemble de Erdos-Renyi, com fração de arestas $\frac{1}{2}$), a cada

passo do algoritmo é proposta uma mudança em uma aresta do grafo - adicionando uma aresta faltante ou removendo uma aresta existente. Caso a mudança não quebre a conectividade do grafo, ela será aceita com probabilidade dada pelo fator de Gibbs:

$$e^{-\beta(H(G')-H(G))}. \tag{3.8}$$

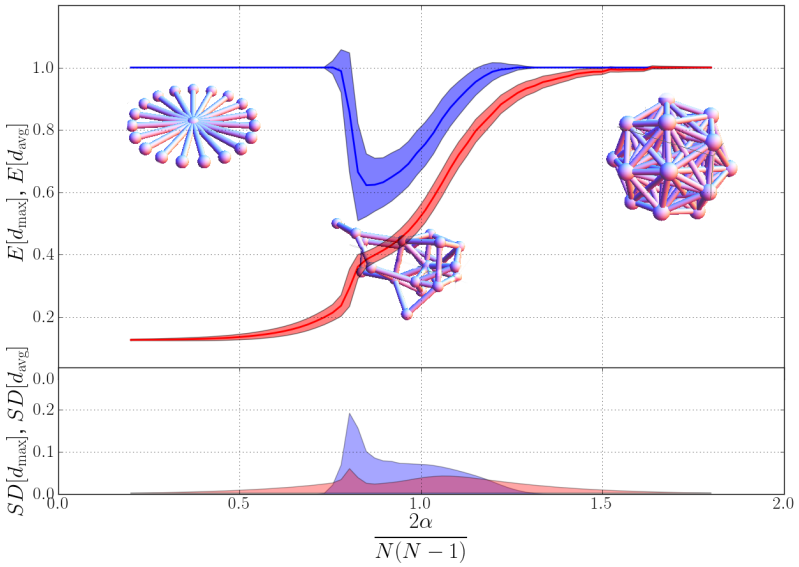


Figura 3.5 – Corte do diagrama de fases apresentando o valor esperado obtido via Monte Carlo dos parâmetros de ordem normalizados $d_{\max}$ e d_{avg} em função de α, com temperatura e número de agentes fixo, bem como seus desvios padrão. Sobreposto ao gráfico se observam exemplos de arquiteturas do grafo sorteadas da distribuição de equilibrio na região correspondente do diagrama.

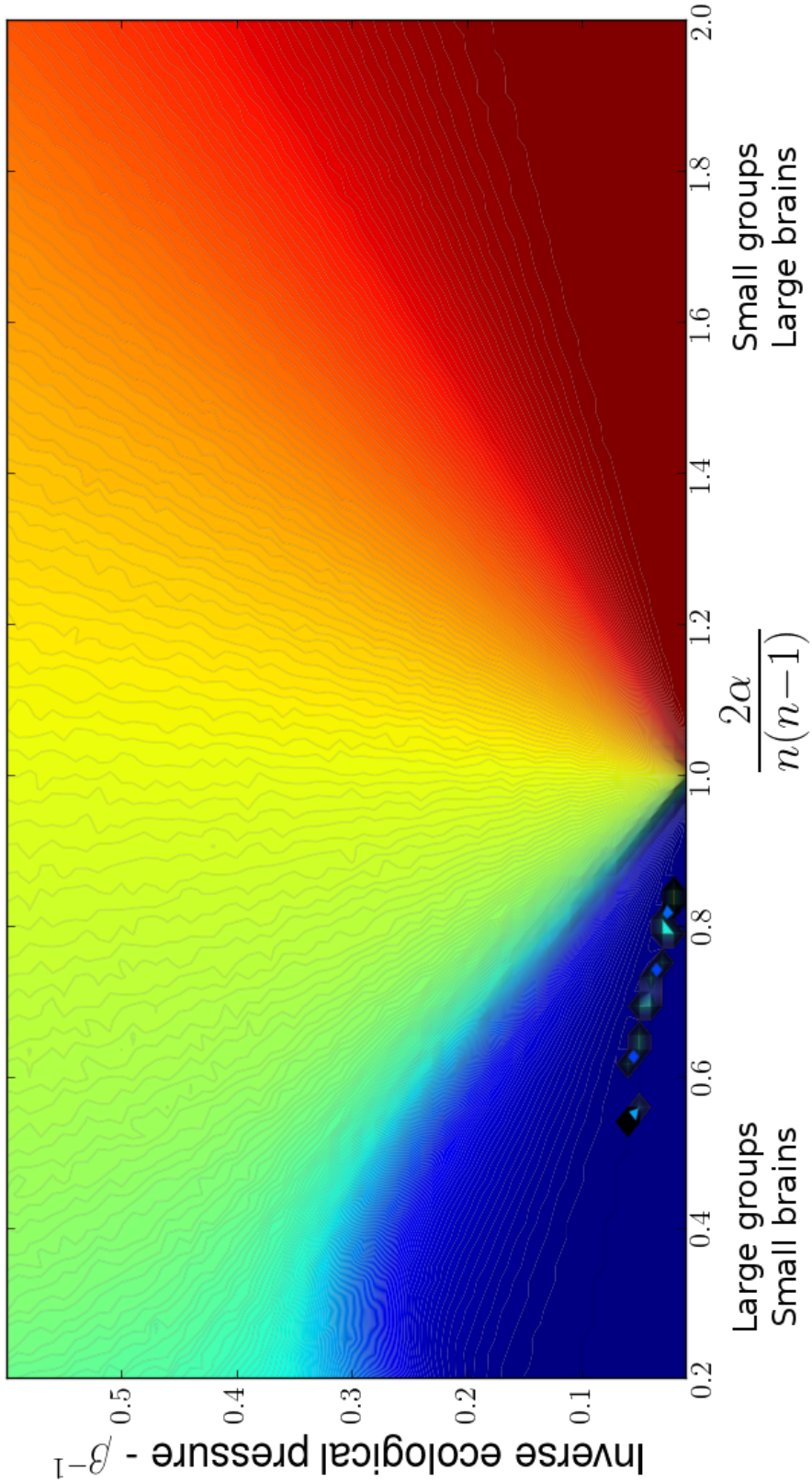
Na figura 3.5 temos um corte do diagrama de fase desse modelo com temperatura e número de agentes constante, variando o parâmetro α. Seja d_i o grau do i-ésimo nó do grafo de um agente escolhido ao acaso. No painel superior destacamos dois parâmetros de ordem: $d_{\max} = \frac{1}{N-1}\max(d_1, d_2, \ldots, d_N)$, em azul, é proporcional ao grau do nó mais bem conectado do grafo, e $d_{\text{avg}} = \frac{1}{N(N-1)}\sum_{i=1}^{N} d_i$, em vermelho, é proporcional ao grau médio de todos os nós do grafo. Uma vez que o maior possível grau para um nó é $N-1$, esses parâmetros foram normalizados para que o valor máximo seja 1. As linhas sólidas correspondem aos valores esperados:

$$E\left[d_{\max}\right] = \sum_G P(G|\alpha, n, \beta) d_{\max}(G) \tag{3.9}$$

$$E\left[d_{\text{avg}}\right] = \sum_G P(G|\alpha, n, \beta) d_{\text{avg}}(G) \tag{3.10}$$

obtida através da simulação de Monte Carlo do modelo, enquanto o sombreado ao redor da linha representa o desvio padrão obtido da mesma forma. No painel inferior apresentamos novamente o desvio

Figura 3.6 – Diagrama de fases apresentando a razão $\frac{d_{\text{avg}}}{d_{\text{max}}}$ em função de α e temperatura, para um número fixo de agentes.



padrão para melhor visualização. Sobrepostas aos gráficos estão figuras representativas de grafos sorteados da distribuição de equilíbrio em pontos correspondentes do diagrama de fases.

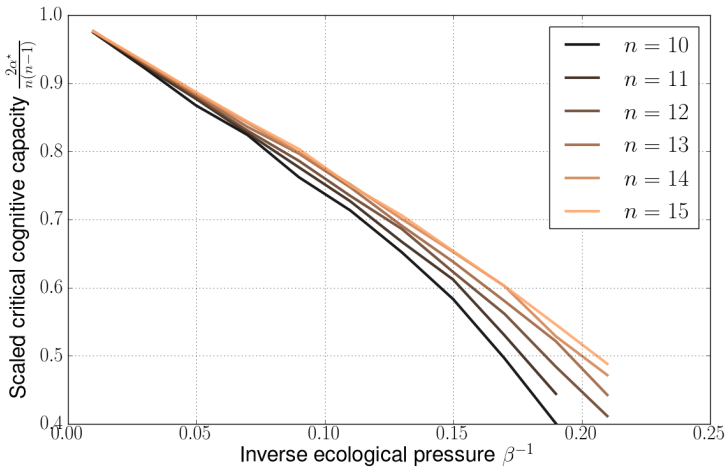


Figura 3.7 – Valor crítico do parâmetro a em função da temperatura para diferentes tamanhos do sistema.

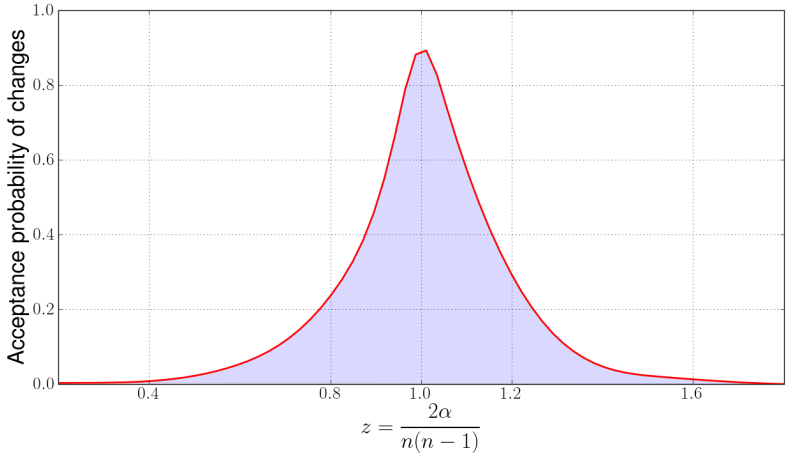


Figura 3.8 – Taxa de aceitação do algoritmo de Monte Carlo - fração das propostas de mudanças no microestado do sistema que foram aceitas com probabilidade dada pelo fator de Gibbs, sampleada com $\beta = 1.0$

Na figura3.6 temos um diagrama de fases completo variando α e a temperatura para um número fixo de agentes. A variável descrita no mapa de cores é a razão $\frac{d_{\text{avg}}}{d_{\text{max}}}$. Esse diagrama mostra uma linha de transição de fases entre a região azul escura - a região em que a organização do grafo é fortemente centralizada, com nós periféricos

pouco conectados, e uma região em que a razão $\frac{d_{\text{avg}}}{d_{\text{max}}}$ é menos extrema. Acima de uma temperatura crítica essa fase não é mais observada. A região vermelho escura corresponde à fase totalmente conectada, ou situações bem próximas disso. Nessa região não há grandes saltos nos parâmetros de ordem, que mudam continuamente com a temperatura e a. A linha de transição de fase pode ser observada para diferentes valores da temperatura na figura 3.7. Note que para $\beta^{-1} \to 0$, temos $\alpha^\star = 1$.

3.3.8 Interpretação parcial dos resultados

O diagrama de fases apresenta três regimes. Para valores altos de $a = \frac{2\alpha}{N(N-1)}$, ou seja, alta capacidade cognitiva ou grupos com poucos agentes, todos os nós apresentam praticamente a mesma conectividade. O grafo é simétrico, com conectividade densa e bem próximo de totalmente conexo. A variação de $E\left[d_{\text{max}}\right]$ e $E\left[d_{\text{avg}}\right]$ com a nesse regime é compatível com a de um ensemble de Erdös-Renyi com alta conectividade, o que indica que as arestas são mais ou menos aleatóriamente distribuidas, de forma totalmente simétrica.

Para valores intermediários de a, existe um nó com conectividade ligeiramente maior, mas existem flutuações grandes. A taxa de aceitação do algoritmo de monte carlo é alta (ver: figura 3.8). Nesse regime, o grafo é momentaneamente não-simétrico, mas estatísticamente qualquer nó pode ocupar a posição de conectividade maior e alterações desse nó central são frequentes. Há um pico nos desvios padrão dos parâmetros de ordem para um certo valor do parâmetro de controle $a^\star$, indicando um possível ponto crítico.

Para valores mais baixos de a, a simetria é espontaneamente quebrada e apenas um nó ocupa uma posição central. Esse nó está conectado a todos os outros, que estão quase que exclusivamente conectados a ele. O grafo se torna uma estrela e não há flutuação observável na conectividade do nó central. Neste regime, a representação mental da rede de relacionamenos sociais construídas pelo agente é assimétrica e existe um único nó que serve como proxy para todas as relações sociais do grupo. Na representação mental do agente em questão, o status social de cada um dos outros nós é definido por como ele se relaciona a esse nó central.

Na figura 3.9 pode-se ver como o diagrama de fases varia com o número de agentes. Quanto maior o número de agentes, maiores as variações dos parametros de ordem durante a possível transição de fase. Na figura 3.10 podemos ver uma simulação com finite size scaling em um gráfico do inverso do desvio padrão de $E\left[d_{\text{max}}\right]$ no ponto crítico $\alpha^\star$ contra o tamanho do sistema, mostrando que essa grandeza diverge para $N \to \infty$.

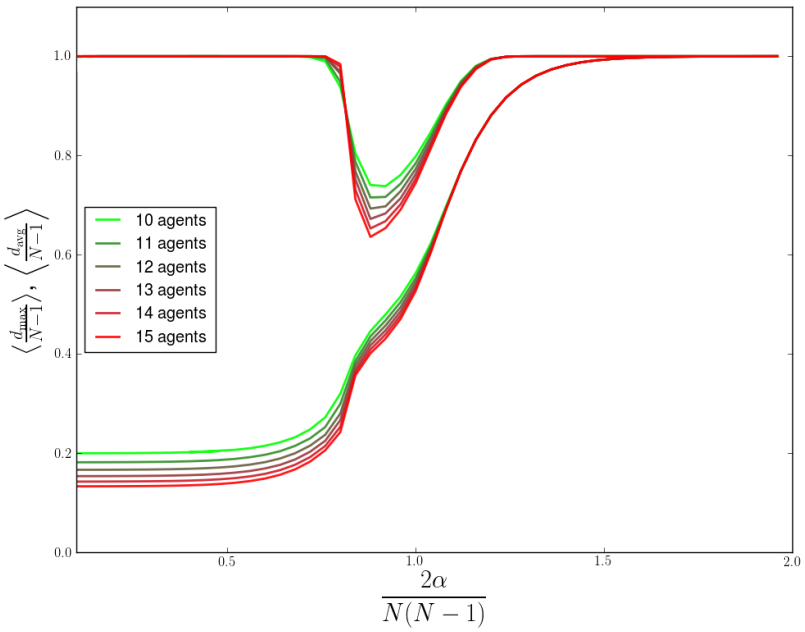


Figura 3.9 – Corte do diagrama de fase para vários tamanhos do grupo.

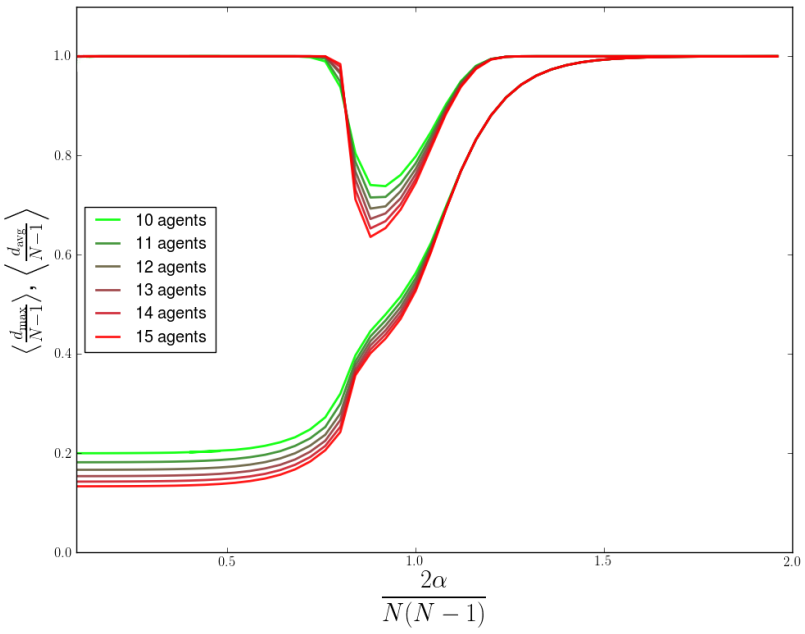


Figura 3.10 – Finite size scaling do modelo.

3.3.9 *Dinâmica para muitos agentes e resultados numéricos*

As figuras acima tratam de propriedades independentes da interação entre os agentes. Essa interação, como dito anteriormente, será introduzidas na forma de aprendizado social ("fofoca" ou *gossip*). Durante a simulação de Monte Carlo, duas possíveis fontes serão consideradas para a proposta de uma nova aresta no passo de Metropolis:

- Com probabilidade $1 - g$, um novo valor para a aresta (i, j) do agente k será sorteado ao acaso,
- Com probabilidade g, um novo valor para a aresta (i, j) do agente k será copiado da aresta (i, j) de um outro agente l sorteado ao acaso.

Essa proposta de novo valor de aresta será aceita com probabilidade proporcional ao fator de Gibbs (3.8). Esse procedimento visa imitar o aprendizado social observado em humanos [7]. Essa escolha de interação não altera os diagramas de fase já mostrados no capítulo anterior, mas introduz correlação entre os grafos de diferentes agentes.

[7]

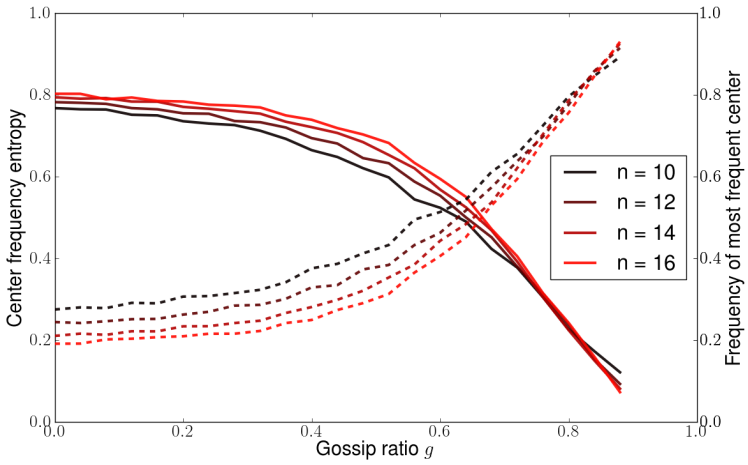


Figura 3.11 – Parâmetros de ordem associados à correlação entre grafos de diferentes agentes, calculados na fase em que os grafos apresentam estrutura de estrela. As curvas tracejadas correspondem à frequência do nó central mais frequente. As linhas tracejadas correspondem à entropia da distribuição de centros.

Na figura 3.11 são exibidas duas grandezas que quantificam a correlação entre os grafos na fase estrela. Vamos denotar por c_i o label que identifica o nó central do grafo do i-ésimo agente. Para um certo número N de agentes temos então o conjunto $\{c_1, c_2, \ldots, c_N\}$. Seja a variável aleatória C definida como um valor sorteado ao acaso desse conjunto e seja:

$$p(c) = \text{Prob}\{C = c\}$$

a sua distribuição de probabilidades. O primeiro parâmetro de ordem, correspondente às linhas tracejadas, é dada por $E[\max_c p(c)]$, ou seja,

o fração do número de agentes que possuem como nó central o nó que mais vezes aparece como nó central. Isso corresponde de forma grosseira a que fração dos agentes tem grafos estrela com o mesmo nó ocupando o centro da estrela. A segunda variável é, a menos de uma constante multiplicativa, simplesmente a entropia da distribuição de c: $S(c) = -\sum_c p(c) \log p(c)$. Ambas as grandezas são calculadas para grafos em forma de estrela, em função da probabilidade de encontro entre dois agentes dada por g, para valores fixos de temperatura, variando-se o número de agentes. O resultado mostra que, para baixos valores de g, a probabilidade de que um certo nó seja o centro de um agente tomado ao acaso é aproximadamente uniforme, e nenhum dos nós domina como centro de uma fração substancial de grafos. Para valores maiores de g, os grafos estrela tendem a se correlacionar e o mesmo nó pode ser central em uma grande fração de agentes. Dessa forma, é possível que no regime em que o grafo é uma estrela, o mesmo agente sirva como proxy para as relações sociais de todo o grupo para uma substancial maioria dos agentes.

3.4 Sumarização e interpretação dos Resultados

Construímos um modelo para a organização social de uma sociedade de agentes que tentam representar mentalmente sua estrutura social. A representação mental tem um custo cognitivo, derivado da limitação cognitiva do agente, e um custo social, derivado da necessidade de se navegar corretamente as relações sociais. Esses custos levam a um modelo mecanico-estatístico que possui um diagrama de fases com alguns regimes interessantes, controlados pelos parâmetros a, que regula a capacidade cognitiva do agente e/ou o tamanho do grupo, g que controla a intensidade da interação social e β, que controla a tolerância a flutuações no custo total do agente. As fases observadas são:

3.4.1 Grupos pequenos e/ou alta capacidade cognitiva

Para a grande, ou seja, grupos de tamanho pequeno ou agentes com grande capacidade cognitiva, os agentes possuem modelos mentais do panorama social do seu grupo em que nenhum agente em particular ocupa uma posição central. Em outras palavras, a representação mental das redes sociais nesse grupo são todas simétricas e nenhum agente se destaca. Nessa fase, em que nenhum agente se destaca como referência social para os outros, pode-se invocar a teoria da dominância reversa. Em situações em que a representação mental dos agentes é simétrica, tentativas de dominação hierárquica encontram resistência. A ausência de um indivíduo específico com suficiente capital

social para dominar o grupo implica em uma estrutura igualitária. [*more sources needed*][*this must be better explained*]

3.4.2 *Grupos de tamanho intermediário e/ou capacidade cognitiva intermediária*

Para valores intermediários de a, os agentes possuem representações fluidas do panorama social de seu grupo, com flutuações grandes. Há uma pequena concentração de conectividade, que pode rapidamente flutuar entre um ou outro nó temporariamente central. Estatísticamente, o modelo ainda é simétrico. [*this must be better explained*]

3.4.3 *Grupos grandes e/ou capacidade cognitiva menor*

Quando a é pequeno, ou seja, para grupos grandes e/ou agentes com menor capacidade cognitiva, existe uma quebra de simetria: as representações mentais da rede social dos agentes é centralizada e congelada em um grafo em forma de estrela. Cada um dos agentes possui uma representação mental assimétrica da rede social. O nó central do grafo de um agente se torna a única referência para todos os cálculos sociais a serem realizado por ele e todas as suas decisões em jogos sociais são tomadas levando em conta a natureza da relação dos outros nós com esse nó central.

Quando g é pequeno (baixo aprendizado via *"gossip"*), entretanto, os nós centrais de cada agente são aleatórios e, de certa forma, apesar de haver uma quebra de simetria na representação mental que cada agente faz do grupo, a situação global ainda é simétrica. Para valores maiores de g, a maioria dos agentes possui o mesmo modelo mental: um grafo em forma de estrela, centrado em torno do mesmo agente específico. O grupo usa as conexões desse mesmo agente central como informação mais relevante na tomada de decisões em jogos sociais. Esse agente central está posicionado de forma privilegiada na solução de dilemas sociais, formação de coalizões e outras atividades sociais do grupo. Fazendo a hipótese de que o capital social derivado dessa posição quebra a simetria do resultado de jogos sociais de maneira vantajosa ao agente central, pode-se esperar que esse agente atinja um certo grau de proeminência ou dominância. A simetria assumida na teoria da dominância reversa é quebrada, pois existe um agente com vantagens sociais claras.

4 | Conclusão e Observações Finais

4.1 ...

4.2 ...

4.3 ...

A Provas dos teoremas de Cox

Este apêndice contém demonstrações dos teoremas exibidos na seção 1.2.1, *Probabilidades e Inferência*.

A.1 Primeiro teorema de Cox e a regra do produto

[*this must be better explained*]

Por exemplo, a conjunção booleana ($\wedge$) é uma operação associativa, ou seja:

$$P_1 \wedge (P_2 \wedge P_3) = (P_1 \wedge P_2) \wedge P_3.$$

Isso implica também na associatividade da função $G(u, v)$, ou seja:

$$G(G(u, v), w) = G(u, G(v, w)).$$

Esse vínculo é satisfeito por infinitas possíveis funções $G(u, v)$, porém todas elas [1] têm a forma:

[1] ; and

$$G(u, v) = g^{-1}(g(u)g(v))$$

com $g(\cdot)$ uma função monotônica. Sendo $g(u)$ monotônica, pode-se redefinir a atribuição de números reais às plausibilidades para $g(P|Q)$ sem perder o ordenamento de proposições segundo suas plausibilidades. Ao fazer isso podemos enunciar o primeiro teorema de Cox: A equação (1.1) é reminiscente da regra do produto da Teoria das Probabilidades, o que indica o objetivo do presente raciocínio.

A.2 Valores extremos

[*this must be better explained*]

Sejam[2] P_T e P_F os valores associados à plausibilidade regraduada $\pi(\cdot|\cdot)$ de eventos sabidamente verdadeiros ou falsos, respectivamente[3]. Se P é sabido verdadeiro, então a plausibilidade de que P e Q sejam simultaneamente verdadeiros é exatamente a plausibilidade de apenas Q ser verdadeiro, ou seja, $\pi(P \wedge Q|P) = \pi(Q|P)$. Mas, pela regra do produto:

[2] Das palavras inglesas *"true"* e *"false"*, respectivamente.

[3] Requisitos de consistência exigem que sejam iguais para quaisquer proposições falsas ou verdadeiras

$$\pi(P \wedge Q|P) = \pi(P|P)\pi(Q|P \wedge P) = P_T\pi(Q|P)$$

Dessa forma, $P_T \pi(Q|P) = \pi(Q|P)$, para quaisquer Q e P, o que implica em $P_T = 1$. Da mesma forma, a plausibilidade de que simultaneamente P e $\bar{P}$ sejam verdadeiros, dada uma proposição Q qualquer, deve ser P_F, pois é uma contradição. Mas, pela regra do produto:

$$\pi(P \wedge \bar{P}|Q) = \pi(P|Q \wedge \bar{P})\pi(\bar{P}|Q)$$

Independentemente de Q, $\pi(P|Q \wedge \bar{P})$ deve ser também igual a P_F e, assim, $P_F = P_F \pi(\bar{P}|Q)$, para quaisquer P e Q. Duas soluções são possíveis: $P_F = 0$ ou $P_F = \infty$. Uma vez que quaisquer das soluções para P_F pode ser mapeada na outra por uma regraduação monotônica, pode-se arbitrariamente escolher $P_F = 0$, e assim limitar valores de $\pi(P|Q)$ no intervalo $[0, 1]$.

A.3 Teorema de Bayes

[*this must be better explained*]

Uma conseqüência imediata da regra do produto segue da seguinte observação. Uma vez que a conjunção $P \wedge Q$ é simétrica:

$$\pi(P_1 \wedge P_2|Q) = \pi(P_2 \wedge P_1|Q)$$

Aplicando a regra do produto em ambos os membros da equação acima, temos:

$$\pi(P2|Q \wedge P_1)\pi(P_1|Q) = \pi(P1|Q \wedge P_2)\pi(P_2|Q)$$

e portanto:

$$\pi(P2|Q \wedge P_1) = \frac{\pi(P1|Q \wedge P_2)\pi(P_2|Q)}{\pi(P_1|Q)} \tag{A.1}$$

que é similar ao teorema de Bayes da Teoria de Probabilidades.

A.4 Regra da soma

[*this must be better explained*]

Considere as proposições P, S e $Q = \overline{P \wedge S}$. Note, em primeiro lugar, que:

$$P \wedge \bar{Q} = P \wedge (P \wedge S) = P \wedge S = \bar{Q}.$$

Note ainda que:

$$\overline{\bar{P} \wedge Q} = P \vee \bar{Q} = P \vee (P \wedge S) = P,$$

e portanto $\bar{P} \wedge Q = \bar{P}$ e $P \wedge \bar{Q} = \bar{Q}$.

Considere a seguir a plausibilidade regraduada dada por $\pi(P \wedge Q|R) = \pi(P|R)\pi(Q|P \wedge R)$. Note que a função $F(\cdot)$ deve ser idempotente, uma vez que $\bar{\bar{P}} = P$ implica que $F(F(u)) = u$. Portanto,

$$\pi(Q|S) = F(\pi(\bar{Q}|S))$$

para qualquer S. Assim, usando a regra do produto:

$$\pi(P \wedge Q|R) = \pi(P|R)F\left(\pi(\bar{Q}|P \wedge R)\right) \tag{A.2}$$

$$= \pi(P|R)F\left(\frac{\pi(\bar{Q} \wedge P|R)}{\pi(P|R)}\right). \tag{A.3}$$

Mas a mesma operação pode ser feita em outra ordem – uma vez que a conjunção $P \wedge Q$ é simétrica na troca de P por Q e, portanto:

$$\pi(P \wedge Q|R) = \pi(P|R)F\left(\frac{\pi(\bar{Q} \wedge P|R)}{\pi(P|R)}\right) = \pi(Q|R)F\left(\frac{\pi(\bar{P} \wedge Q|R)}{\pi(Q|R)}\right)$$

para quaisquer P, Q e R. Em particular, deve valer para o caso particular em que $\bar{Q} = P \wedge S$. Nesse caso, é possível mostrar, através das regras da álgebra booleana, que $P \wedge \bar{Q} = \bar{Q}$ e que $\bar{P} \wedge Q = \bar{P}$[4]. Finalmente isso significa que:

$$uF\left(\frac{v}{u}\right) = vF\left(\frac{u}{v}\right)$$

Novamente, há infinitas soluções $F(\cdot)$ para esse vínculo, mas todas elas satisfazem[5]:

$$f(u)^\alpha + u^\alpha = 1, \tag{A.4}$$

o que, regraduando as plausibilidades novamente por uma transformação monotônica $p(P|Q) = \pi(P|Q)^\alpha$, que preserva os valores $P_F = 0$ e $P_V = 1$, permite enunciar o segundo teorema de Cox.

Teorema 5 (2º teorema de regraduação de Cox)**.** *Uma vez que uma representação consistente de plausibilidades $\pi(P|Q)$ com um ordenamento bem definido foi encontrada para a qual vale a regra do produto, sempre é possível encontrar uma outra equivalente $p(P|Q)$ tal que:*

$$p(P|Q) + p(\bar{P}|Q) = 1 \tag{A.5}$$

Essa é denominada a regra da soma. A partir das regras da soma e do produto todas as outras regras comuns da teoria de probabilidades podem ser facilmente derivadas. Por exemplo a regra de Bayes vem diretamente da regra do produto através da identificação:

$$p(Q|R \wedge P)P(Q|R) = p(P \wedge Q|R) = p(P|R \wedge Q)P(Q|R). \tag{A.6}$$

A forma mais geral da regra da soma:

$$p(Q|R) + p(P|R) = p(Q \vee P|R) + p(Q \wedge P|R) \tag{A.7}$$

também pode ser deduzida dessas duas regras. Isso leva à conclusão de que, se há uma teoria consistente de inferência sobre informação incompleta representada por números reais, essa teoria deve ser idêntica à teoria das probabilidades. Por essa razão, chamaremos o funcional $p(\cdot|\cdot)$ de probabilidade, daqui por diante.

[4] Basta notar que:

$$P \wedge \bar{Q} = P \wedge (P \wedge S) = (P \wedge P) \wedge S = P \wedge S = \bar{Q}$$

Para simplificar o outro caso, note que $\bar{P} \wedge \bar{Q} = \bar{P} \wedge (P \wedge S)$ e é falso, pois é uma contradição. Mas $\bar{P} \wedge \bar{Q} = \overline{P \vee Q}$ ($\vee$ representando a disjunção "OU"), o que implica que $P \vee Q$ é verdadeiro. uma vez que $A \wedge V = A$, para qualquer A, temos $\bar{P} \wedge (P \vee Q) = \bar{P}$. Mas:

$$\bar{P} = \bar{P} \wedge (P \vee Q) = (\bar{P} \wedge P) \vee (\bar{P} \wedge Q) = F \vee (\bar{P} \wedge Q) = \bar{P} \wedge Q$$

onde usamos o fato que $F \vee A = A$ para qualquer A e que $A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$.

[5] ; and