Título: O Uso de métricas estatísticas em dados reais

Resumo

(Colocar depois)

1 Justificativa

(Colocar Depois)

2 Objetivos

(Colocar Depois)

3 Metodologia

3.1 Teoria da Informação

É consenso que o desenvolvimento e o progresso precisam ocorrer de maneira sustentável, mantendo as condições desejáveis e evitando, assim, mudanças catastróficas na condição do sistema. É possível que mudanças de regime resultem em danos significativos e, até mesmo, irrecuperáveis (Brock & Carpenter, 2006).

Como o objetivo de detectar tais mudanças de regime, surgiu a Teoria da Informação que, dentro de um contexto multidisciplinar, desenvolve e aplica métricas qualitativas e quantitativas para avaliação de mudanças de regime. Tal tópico tornou-se de grande importância para análise de sustentabilidade e, embora não exista um consenso sobre medidas universais, pesquisadores continuam a estudar métricas capazes de detectar e mensurar mudanças de comportamento de sistemas complexos. Como discutido em (Fisk, 2010), o desenvolvimento sustentável de sistemas humanos envolve a avaliação de uma gama de componentes, incluindo fontes sociais, ambientais e econômicas. Neste cenário, a Teoria da Informação tem se mostrado muito promissora devido à sua capacidade de avaliar a estrutura, complexidade, estabilidade e diversidade do ecossistema (Mayer et al, 2006).

As metodologias apresentadas a seguir abordam conceitos provenientes da estatística. Em particular, a Informação de Fisher é uma medida de ordem dos dados, sendo capaz de monitorar as variáveis de um sistema e capturar as mudanças de regime (Fath el al, 2003).

3.2 Informação de Fisher

A Informação de Fisher (IF) foi criada pelo estatístico Ronald Fisher, em 1922. Sua motivação inicial foi mensurar a quantidade de informação do parâmetro θ de uma distribuição f, dado um conjunto de medições, independentemente, dessa mesma distribuição.

Note que se f tem um pico acentuado quando varia-se o parâmetro θ, o conjunto de dados X fornece muita informação sobre tal parâmetro. Por outro lado, se f é plana e esparsa, então muitas amostras de X são necessárias para estimar o valor de θ. Isto sugere o estudo de alguma métrica em que se leve em conta a variância do parâmetro θ (Ly et al,2017).

A Informação de Fisher observada é dada por:

$$I = \left(\frac{\partial}{\partial \theta} \ln f(x|\theta) \right)^2 \tag{1}$$

onde $f(x|\theta)$ é a função densidade de probabilidade da variável, dado o parâmetro θ. Valores altos da Informação de Fisher em são obtidos quando f tem um pico acentuado ao variar-se θ e indicam que o conjunto de dados fornece grande informação sobre este parâmetro. Já valores baixos no caso em que f é plana e esparsa e indicam que os dados possuem pouca informação sobre o parâmetro de interesse.

Sob outro ângulo, pode-se mensurar a quantidade de informação do parâmetro θ levando-se em conta a função Score $S(\theta)$ da distribuição:

$$S(\theta) = \frac{\partial}{\partial \theta} \ln(f(x|\theta)^2) \tag{2}$$

Da mesma forma que na Equação (1), esta função, em suma, nos diz o quanto distribuição f é sensível a uma mudança pequena sobre o parâmetro θ.

Parei aqui! Novos comentários...

A idéia de R. Fisher para quantificar nossa informação foi pegar o valor médio da curvatura da nossa função score, que também pode ser pensado como a variância da nossa função score (Pawitan, 2006). Tal análise nos gera a informação de fisher esperada, denotada por $\mathcal{I}$.

$$\mathcal{I} = E\left[\frac{\partial}{\partial \theta} \ln(f(x|\theta)^2) \right] \tag{3}$$

Para relacionar os dois tipos de informação de fisher, basta observar que $\mathcal{I}$ é o valor médio de I. Logo, se fossemos explicar, I seria uma função, que nos diria a curvatura da função score em um determinado θ e consequentemente sua informação e $\mathcal{I}$ nos diria essa curvatura média, e consequentemente, sua

informação média (Pawitan, 2006). Ademais, por consequência da nossa definição, $\mathcal{I}$ é independente da quantidade de dados que coletamos, podendo ser usado em várias coletas diferentes. I por sua vez é bastante dependente do conjunto de dados, variando de coleta em coleta.

$$\mathcal{I} = E(I) \tag{4}$$

Ademais, por meio da desigualdade de Cramer-Rao, a informação de fisher se relaciona com o erro que cometemos na nossa estimativa $\hat{\theta}$ por meio de

$$e\mathcal{I} \geq 1 \tag{5}$$

Ou seja, se tivermos uma informação de fisher muito grande, necessariamente teremos que ter um erro muito baixo. Então na estatística de inferência se é muito interessante tentar maximizar nossa informação de Fisher para um determinado parâmetro como meio de diminuir nosso erro cometido (Pawitan, 2006). Ademais, algumas considerações é que o valor da informação de fisher é sempre positiva e maior ou igual a zero e não é afetada por uma mudança de coordenadas, no sentido de mudarmos para direita ou esquerda nossa distribuição (Frieden, 2010).

Quando vamos tratar de sistemas dinâmicos, sua grande maioria é dependente do tempo, sendo então necessário uma forma de inclusão na nossa métrica (Gonzalez-Mejia et al., 2015). Incluindo um termo para o tempo, nossa equação fica:

$$\mathcal{I} = \int_{-\infty}^{\infty} f(x|\theta)(\frac{\partial}{\partial\theta}\ln f(x|\theta))^2(\frac{\mathrm{d}x}{\mathrm{d}t})^{-1}\,\mathrm{d}x \tag{6}$$

É muito difícil ter as funções necessárias para a informação de fisher, onde para muitos datasets reais, temos informações faltando, com um certo grau de imprecisão, dentre outros problemas (Gonzalez-Mejia et al., 2015). Se desenvolveu então métodos para ser calculado essa informação para que seja aplicável em situações reais. O primeiro é a utilização da forma continua da equação, estimando-se os valores necessários e O segundo é por meio de um 'empacotamento' dos dados.

3.3 Formas de cálculo de FI

Para o método contínuo, utilizando algumas condições de regularização podemos definir qualquer sistema como seguindo uma certa trajetória em um espaço n-dimensional, ligado à quantidade de variáveis. Utilizando-se um pouco de física, esse método manipula nossa equação original para utilizar da velocidade e aceleração tangencial do nosso sistema, utilizando seu cálculo como meio de avaliar a informação de fisher (Mayer et al., 2006).

$$\int_{-\infty}^{\infty} \frac{(\mathrm{d}x/\mathrm{d}x)^2}{(\mathrm{d}^2x/\mathrm{d}x^2)^4}\,\mathrm{d}x \tag{7}$$

Como nossos pontos de dados são discretos, precisaremos estimar e aproximar nossos termos de derivadas de primeira e segunda ordem. Para isso, podemos usar métodos como a série de Taylor, onde tenta-se encaixar a maior quantia de dados possíveis (Gonzalez-Mejia et al., 2015). Mas claro que ainda está suscetível a erros, principalmente a segunda derivada, a qual é muito sensível à barulho nos dados e pode causar diversos problemas caso isso não seja remediado.

Existem algumas formas mais comuns que são usadas nisso, dentre elas está tentar achar uma função de melhor encaixe para os dados e a partir dela usar métodos analíticos. Mas claro que esse método também sofre de problemas, que talvez nossos dados sofram um comportamento que uma curva não consiga aproximar tão bem e problemas de generalização, onde se quisermos estender nossas análise, para talvez prever o que pode acontecer, a generalização da nossa curva pode não ser boa (Gonzalez-Mejia et al., 2015).

Outra forma é fazermos o calculo para um período T suficientemente grande de tal forma que esses barulhos não interfiram na nossa estimativa (Mayer et al., 2006). Mas claro que, sofre de problemas, onde perdemos detalhes dos nossos dados. Pode não ser problemático, mas caso seja uma mudança muito pequena que influencie, a análise não seria útil.

Uma das grandes necessidades desse método é que se precisa calcular a informação de fisher para no mínimo um ciclo do sistema, ou seja, precisaria estimar o quanto um ciclo representaria na escala do tempo. Porém, para sistemas complexos, como os reais, isso é muito difícil, portanto foi desenvolvido uma estimativa que se aproxima muito do ciclo real, já que o valor exato não é possível ser calculado, com certas limitações (Rawlings et al., 2020). A não utilização de um período exato, ou próximo, nos gera um gráfico de fisher com muita oscilação e com uma interpretação muito difícil de ser feita.

Dado essas problemáticas, também existe o método discreto do cálculo da nossa Informação da Fisher (Karunanithi et al., 2008). Porém, invés de usarmos nossa fórmula já descrita, usamos a forma de amplitude dela, de forma que fazendo uma substituição de funções, $p(s) = \sqrt{g(s)}$, nossa equação fica:

$$I = 4 \int_{-\infty}^{\infty} \frac{\mathrm{d}q(s)}{\mathrm{d}t}^2 \mathrm{d}s \tag{8}$$

Onde faremos uma substituição por quantidades discretas, onde nossa integral se torna um somatório e nossas quantias $\mathrm{d}s, \mathrm{d}t$ se tornam quantias reais e finitas, que dada certas condições, nossa aproximação fica

$$I \approx 4 \sum_{i}^{n} [q_i - q_{i+1}]^2 \tag{9}$$

A estratégia desse método é tentar reduzir as variáveis de estados agrupando-as quando obedecem uma certa regularização. Para cada variável medida, existe um erro em sua medição. Se a diferença entre a mesma variável, em diferentes tempos, for menor que o erro para essa variável, consideramos que São diferentes observações para o mesmo valor e o agrupamos. Agora, se o mesmo valor

observado em diferentes variáveis foi observado, ou seja, se diferenciam por menos que a incerteza, agrupando-as em um grupo (Karunanithi et al., 2008). Cada variável representaria um ponto num hiperplano que contenha todas as variáveis. Portanto, para cada variável agrupada, faremos um volume de tal forma que a probabilidade de observamos um estado seria proporcional à quantidade de pontos que esse volume estaria contendo. Dessa forma, conseguimos estimar cada probabilidade de um estado ser observado e calcular a informação de fisher.

Porém, esse método tem o problema de necessitarmos dessa incerteza para cada variável. Uma forma de fazer isso é observar um sistema que já esteja estável e contenha as mesmas variáveis e usarmos a variação dessas variáveis como incerteza para nossos cálculos. Caso isso não seja possível, teremos que achar o estado mais estável do nosso sistema em estudo e calcular o desvio padrão para cada variável e assumi-lo como incerteza. (Gonzalez-Mejia et al., 2015,Karunanithi et al., 2008)

Posteriormente, se calcula a evolução da função densidade de probabilidade com o tempo, dividindo o período em janelas de tempo e calcula-se a probabilidade de cada estado e a forma discreta de amplitude da nossa informação de fisher. Tal método também sofre do problema de necessitar do período, mas é bastante útil no sentido de conseguir lidar com variações e incertezas nos nossos dados.

3.4 Aplicações no Real

O uso da informação de Fisher vem sendo estudado a um certo tempo com resultados bastantes promissores. No artigo seminal publicado em 2006 por Audrey L. Mayer, et al. Eles propõem uma forma de uso e aplicações da Informação de Fisher. Por meio o uso da formula (7), fazendo-se estimativas numéricas para a velocidade e aceleração, conseguiram com sucesso o calculo para diversos ecossistemas a qual se tiveram mudança de regime.

A ideia da analise para essa informação é analisar a variação da Informação de Fisher ao longo do tempo, de forma que se o sistema está em um ciclo estavel, seu valor não se alteraria ao longo do tempo. Porem, caso se tenha uma variação, ou seja, o sistema saísse do seu ciclo, classificaríamos como um flip ecologico. Foram feitas algumas analises em sistemas que por meio de outros métodos foi-se identificado essas mudannças ecologicas. Portanto, a Informação de Fisher, precisaria ter uma mudança, positiva ou negativa, na faixa dessas mudanças. Para o ecossistema do norte do pacifico, na região proxima do estreito de bering, se tiveram duas mudanças ecologicas, uma entre 1976 e 1977 e outra em 1989. Utilizando-se uma base de dados com 100 variaveis, com a seleção de cerca de 70, foram-se feitas ás analises e estimativas e observou-se uma clara mudança do valor da Informação de Fisher coicidindo com os flips ecologicos ja estabelecidos por outras metodologias.

Ademais, foi feita a analise sobre o clima global, onde analisou-se niveis de gases, como o dioxido de carbono e metano, que ficaram presos no gelo da antartica ao longo dos milhares de anos. A quantia deses gases se relacionam

diretamente com o clima mundial daquela epoca, de forma que servem como um bom indicador de mudancas climaticas em escala global. Mudancas globais sao vistas especialmente entre climas majoritariamente frios e majoritariamente quentes, portanto, foi evidenciado umamudanca da Informacao de fisher entre esses periodos.

Por fim, a ultima analise que foi feita foram nos climas Saharianos, onde ambos tiveram umamudanca similar entre climas aridos e umidos. Por meio de coleta de sedimentos de lagos e oceanos, foram coletados dados como porcentagem de poeira terrestre no leito oceanico Pesquisamr mais variaveis que foram analisadas e por meio delas foi feito o calculo da informacao. As mudancas observadas se relacionam as mudancas entre regimes e estao de acordo com estimativas de mudancas nesse sistema por meio de outras metodologias.

Apesar do aparente sucesso dessa metrica, foram ressaltados algumas limitacoes. Por mais que se tenha a mudança, nao foi muito bem definido se a variacao foi negativa ou positiva seria de significancia para a deteminacao do regime. Ademais, a dificuldade da estimativa do periodo foi um fator limitante quanto o calculo sendo que ja era sabido previamente um valor para esse periodo, que caso queira analisar para novos ecossistemas, esse valor nao seria conhecido. Parei aqui

Entropia de Shannon: (Pesquisar e colocar depois) Outra Métrica: (Pesquisar outras possíveis métricas)