Transformação de base e compressão de dados

Pablo Goulart Silva

Universidade Federal de Minas Gerais
pgoulart@dcc.ufmg.br

1 Introdução

Um sinal frequentemente possui mais dados do que necessário para transmitir a informação que ele contém. Dada uma maneira de identificar a informação desnecessária torna-se possível comprimir a mensagem pelo descarte do supérfluo.

O objetivo da compressão de dados é descrever um sinal representado por uma cadeia binária de caracteres longa W em uma cadeia binária mais curta U para transmissão através de um canal C. Espera-se que a informação contida em W possa ser recuperada a partir de U algoritmicamente. Em compressão de dados *com perdas* é permitido transmitir dados através de C assumindo-se que o sinal W não poderá ser integralmente recuperado a partir de U pois durante a transformação $W \to U$ parte da informação contida em W será descartada. A partir dessa definição, um questionamento se torna óbvio: como selecionar as informações relevantes de W que permita descartar em U parte da informação contida em W sem que U perca o sentido.

Alguns métodos permitem caracterizar sinais de modo que seja possível classificar a contribuição de suas 'componentes' na informação global. Esses métodos, conhecidos como transformadas, são comumente aplicados em sinais físicos.

Nesse trabalho abordaremos dois domínios de transformadas clássicos na literatura e como eles permitem a seleção de informações relevantes: são eles o domínio da *frequência* e o domínio das *variâncias*. A abordagem desse trabalho será pragmática ao invés de teórica, com intuito de revelar a motivação de efetuar uma transformação *linear* para compressão dos dados.

2 Transformadas

A representação de um sinal W obtida a partir da observação pode, à primeira vista, não revelar nenhuma característica relevante dele. Entretanto, uma boa mudança de base pode não apenas simplificar a análise mas também permitir a caracterização de W de forma que decisões sejam tomadas observando-se a nova 'forma' deste sinal.

Dado um conjunto de dados representados por um vetor m-dimensional em um espaço ortonormal[1], a transformada de um sinal é uma mudança de base dada pela combinação linear da base original [5].

[1] Isto é, onde cada eixo da base é perpendicular em relação aos demais.

A transformação de base é realizada na esperança de que o dado tenha sua dinâmica revelada no novo sistema de coordenadas. Geralmente a base é definida como resposta à pergunta: Existe alguma base que seja uma combinação linear da base original que melhor representa a base de dados?

3 Domínio da frequência

A transformada de um sinal para o domínio da frequência é uma técnica aplicadas sobre funções no domínio do tempo que sejam compostas por sinais de frequências especiais. Sinais clássicos de aplicação são: imagens, áudio, biomédicos, econometria, etc.

Algumas técnicas, como o JPEG, utilizam transformadas na frequência para descartar informações desses sinais em algoritmos de compressão. A separação entre frequências relevantes e não-relevantes é subjetiva do problema sob estudo, e à ação de eliminar frequências dá-se o nome de *quantização*, que provê o mecanismo de compressão com perdas.

As transformadas de *Fourier* e seus casos particulares *Cosseno* e *Seno* permitem que o sinal seja caracterizado observando as contribuições de cada componente de frequência[2].

3.1 Transformada de *Fourier*

A Transformada Discreta de *Fourier* (DFT) descreve o sinal original através das amplitudes de cada uma de suas componentes de frequência[3]:

$$W_v = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} h(k) e^{-2\pi i v k/N}, v = 0, ..., N-1 \tag{1}$$

As transformadas no domínio da frequência exigem que o sinal seja periódico. Dado um sinal W observado durante um período $T > 0$ unidades de tempo, utiliza-se a extensão de período T de W para aplicação da técnica. Não é possível obter, na prática, o valor $W(t)$ $\forall t$ no intervalo de tempo contínuo $0 \leq t \leq T$. Portanto, W é amostrado em intervalos discretos t_n, tal que $0 \leq t_n \leq T$ e $0, T/N, ..., (N-1)T/N$.

[2] Utilizaremos ao longo das seguintes seções os termos Transformada de Fourier, Cosseno e Seno sendo que, formalmente, estaremos descrevendo as Transformadas *Discretas* de Fourier, Cosseno e Seno, por serem transformadas aplicadas sobre sinais amostrados em intervalos discretos no tempo.

[3] A DFT de um sinal é um mapeamento $t \in \mathbb{R} \to n \in \mathbb{C}$ sendo, portanto, um sinal complexo.

3.2 Transformada do Cosseno

A Transformada Discreta do Cosseno (DCT) é obtida a partir da Transformada de *Fourier* utilizando-se a identidade de Euler $2\cos\theta = e^{i\theta} + e^{-i\theta}$. Portanto:

$$\mathbf{x}(v) = \sum_{k=0}^{N-1} x(k)C(v)\cos\frac{(2k+1)v\pi}{2N}, v = 0, ..., N-1 \tag{2}$$

Para $C(0) = \sqrt{1/N}$ e $C(k) = \sqrt{2/N}$, se $k \neq 0$.

A DCT é útil quando deseja-se atenuar componentes de alta frequência de W, dado que $\cos(W)$ é uma extensão par em torno do eixo y em $k = (N-1)/2$.[4]

JPEG Em 1980, um comitê de especialistas em imagens (*Joint Photographic Experts Group* ou JPEG) criou uma especificação para compressão de imagens com e sem perdas. O método de compressão com perdas (*lossy* JPEG) baseia-se na ideia de aproximação local cujo detector de detalhes utilizado é o DCT. Blocos de dimensão 8×8 são transformadas utilizando-se DCT e cada bloco é *quantizado* por um método que suprime altas frequências. A saída do quantizador é comprimida utilizando um método de compressão sem perdas. Essa etapa utiliza o método de *Huffman* ou *Codificação Aritmética*. [3]

$$\mathcal{W} \xrightarrow{Transformação} \mathcal{T}_{\mathcal{W}} \xrightarrow{Quantização} \mathcal{Q}_{\mathcal{W}} \xrightarrow{Compressão} \mathcal{U}$$

$$\mathcal{U} \xrightarrow{Descompressão} \mathcal{Q}_{\mathcal{W}'} \xrightarrow{Dequantização} \mathcal{T}_{\mathcal{W}'} \xrightarrow{T.Inversa} \mathcal{W}'$$

3.3 Transformada do Seno

A Transformada do Seno (DST) é dada pela utilização da Identidade de Euler $2i\sin\theta = e^{i\theta} - e^{-i\theta}$. A função *Seno* é utilizada para evidenciar componentes de alta frequência do sinal[5]:

$$\mathbf{x}(v) = \sqrt{\frac{2}{N+1}} \sum_{k=0}^{N-1} x(k)\sin\frac{\pi(k+1)(v+1)}{N+1}, v = 0, ..., N-1 \tag{3}$$

[4] Essa é uma das motivações da utilização da DCT, pois ela não introduz componentes artificiais de alta frequência como DFT e DST. Isso evidencia componentes de alta frequência originais do sinal, além de ser um mapeamento $\mathbb{R} \to \mathbb{R}$.

[4] https://github.com/pablosistemas

[5] Analogamente a DCT, a identidade de Euler para o Seno é um mapeamento $\mathbb{R} \to \mathbb{C}$, sendo um valor complexo puro. Aliada à extensão ímpar do sinal, a DST amplia os efeitos de componentes de alta frequência no sinal.

4 Análise de Componentes Principais

Para o estudo de fenômenos desconhecidos a abordagem comum é monitorar diversas métricas (deslocamento, velocidade, aceleração, tensão, corrente, etc.) para identificar o modelo físico que o representa. Cada métrica colhida nesse processo representa uma dimensão do fenômeno sob análise.

Entretanto quando um conjunto de dados é altamente dimensional a aplicação de algumas técnicas algorítmicas torna-se proibitiva. Para solucionar esse problema é necessário uma maneira de selecionar as dimensões mais relevantes e definir uma estrutura enxuta do fenômeno sob estudo [1]. Esse princípio é bem explicado pelo Princípio da Navalha de *Occam*, que defende a utilização do modelo mais simples que explica os dados sob análise [4].

A Análise de Componentes Principais (PCA) é uma técnica não paramétrica utilizada em análise de dados para seleção de características [2]. A seleção das dimensões principais de um conjunto de dados realizada pelo PCA é motivada por duas razões: utilização do dado com menor relação sinal-ruído (*Signal Noise Ratio* ou SNR) e descarte de redundâncias[6] pela escolha de dimensões não-correlacionadas. O SNR pode ser definido como:

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}} \tag{4}$$

Ambas as motivações se apoiam sobre a matriz covariância das dimensões dos dados originais [1][7]. Portanto, as premissas para a aplicação do PCA (bem como o entendimento quando essa técnica não produz os resultados desejados) podem ser descritas por [5]:

- Linearidade[8].
- Média e variância explicam a distribuição de probabilidade dos dados[9].
- Importância das dimensões de maior variância na dinâmica do sistema.
- Os componentes principais selecionados pela técnica formam uma base ortonormal.

[6] A remoção de dados redundantes é um passo essencial na identificação de sistemas caixa-preta. Ela evita diversos problemas de mal-condicionamento numérico, além de permitir escolher o tempo de amostragem para o sistema sob identificação em cenários que a frequência de *Nyquist* deste não é conhecida. Em [1], capítulo 12, o autor aborda técnicas de covariância lineares e não-lineares para escolha da frequência de amostragem de um sistema desconhecido.

[7] Em [1], capítulo 4, o autor explica de maneira prática o efeito *averaging* das técnicas baseadas em covariância na obtenção de dados com maior SNR. Essa técnica motiva a utilização de sinais de amplitudes maiores na identificação de sistemas caixa-preta. O intuito é atenuar o efeito do ruído do sinal capturado nos experimentos.

[8] Mudanças não lineares podem ser aplicadas sobre os dados antes da aplicação do PCA, como estudado pelos *Kernel PCA*.

[9] Cenário comum no mundo real graças ao Teorema do Limite Central.

Referências

1. L. A. Aguirre. *Introdução à identificação de sistemas–Técnicas lineares e não-lineares aplicadas a sistemas reais*. Editora UFMG, 2004.
2. A. C. Frery and T. Perciano. *Introduction to Image Processing Using R: learning by examples*. Springer Science & Business Media, 2013.
3. P. D. Johnson Jr, G. A. Harris, and D. Hankerson. *Introduction to information theory and data compression*. Chapman and Hall/CRC, 2003.
4. D. J. MacKay and D. J. Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
5. J. Shlens. A tutorial on principal component analysis; http://www. cs. princeton. edu/picasso/mats. *PCA-Tutorial-Intuition_jp. pdf*, 2003.