Algoritmos para Estimação de Modelos de Gráficos

Rodrigo Ribeiro Santos de Carvalho
Orientadora: Florencia Graciela Leonardi
Universidade de São Paulo

Introdução

O modelo gráfico probabilístico é um modelo que expressa as dependências condicionais de variáveis aleatórias através de um grafo. O campo aleatório de Markov é a forma de grafo não-direcionado deste modelo.

Um problema de pesquisa atual é a reconstrução, a partir de amostras, do campo aleatório de Markov. Nesse trabalho, estudamos dois algoritmos presentes na literatura e realizamos simulações para verificar empiricamente a consistência. E por fim, aplicamos um dos algoritmos em uma situação com dados reais.

Definições básicas

Sejam $G = (V, E)$ um grafo não-direcionado e $\{X_v : v \in V\}$ uma família de variáveis aleatórias indexadas pelo conjunto de vértices V e assumindo valores em um conjunto A finito. Dizemos que $\{X_v : v \in V\}$ é um *campo aleatório de Markov* em relação ao grafo G se para cada $v \in V$ existir um conjunto $ne(v) \subseteq V \setminus \{v\}$ tal que para todo $W \subseteq V \setminus \{v\} \cup ne(v)$,

$$\mathbb{P}(X_v = a_v | X_{ne(v)} = a_{ne(v)}) = \mathbb{P}(X_v = a_v | X_{ne(v)} = a_{ne(v)}, X_W = a_W)$$

O conjunto $ne(v)$ é chamado de vizinhança Markoviana de v.

Seja $X_V^{(1:n)}$ uma amostra independente de tamanho n de um campo aleatório de Markov $\{X_v : v \in V\}$. x_v^i denotará a i-ésima observação da variável X_v. Sejam $v \in V$ e $W \subseteq V \setminus \{v\}$. O operador $N(a_v, a_W)$ conta o número de vezes que o evento $\{x_v^i = a_v, x_W^i = a_W\}$ acontence na amostra e é definido por

$$N(a_v, a_W) = \sum_{i=1}^{n} = \mathbf{1}\{x_v^i = a_v, x_W^i = a_W\}$$

Algoritmo de máxima verossimilhança penalizada

Este algoritmo estima cada vizinhança $ne(v)$ do grafo através do cálculo de máxima log-verossimilhança penalizada por um fator que depende diretamente do tamanho da vizinhança candidata. O seu estimador é, para $c > 0$

$$\widehat{ne}(v) = \underset{W \subseteq V \setminus \{v\}}{\arg\max} \{ \sum_{a_v \in A} \sum_{a_W \in A^W} N(a_v, a_W) \log \widehat{p}(a_v, a_W) - c|A|^{|W|} \log n \}$$

onde $\widehat{p}(a_v, a_W) = \widehat{\mathbb{P}}(X_v = a_v | X_W = a_W) = \dfrac{N(a_v, a_W)}{N(a_W)}$ é o estimador de máxima verossimilhança probabilidade condicional $\mathbb{P}(X_v = a_v | X_W = a_W)$ e

$$N(a_W) = \sum_{a_v \in A} N(a_v, a_W)$$

.

Teorema: Para qualquer $v \in V$ e $c > 0$, temos que $\widehat{ne}(v) \to ne(v)$ quase certamente, quando $n \to \infty$

Algoritmo de Chow-Liu

Em casos em que queremos estimar um grafo de dependências que seja uma árvore há um algoritmo quadrático. Funciona da seguinte forma:

1. Para todo par X_v e X_w de vértices, calculamos a *medida de informação mútua* estimada:

$$\widehat{I}(X_v, X_w) = \sum_{(a_v, a_w) \in A^2} \frac{N(a_v, a_w)}{n} [\log \frac{N(a_v, a_w)}{N(a_v)(N_w)} + \log n]$$

2. Aplique um algoritmo de árvore geradora máxima onde cada aresta (v, w) tem peso $\widehat{I}(X_v, X_w)$

Teorema: O algoritmo de Chow-Liu estima uma distribuição de segunda ordem que mais se aproxima da distribuição original em relação à divergência de Kullback-Leibler.

Simulações

Para simular, criamos um gerador de grafos e árvores, de distribuição e de amostras aleatórios. Aqui apresentamos duas simulações, uma para cada algoritmo. Nelas medimos o erro de subestimação, de sobrestimação e total. No caso do algoritmo de máxima verossimilhança penalizada, avaliamos para tamanho de amostra de 10 até 3500.

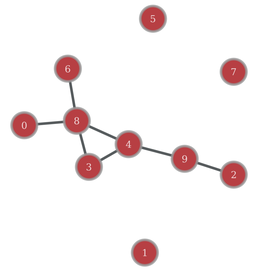

Figura 1: Grafo de 10 vértices usado para gerar amostras e avaliar o algoritmo de máxima verossilhança penalizada.

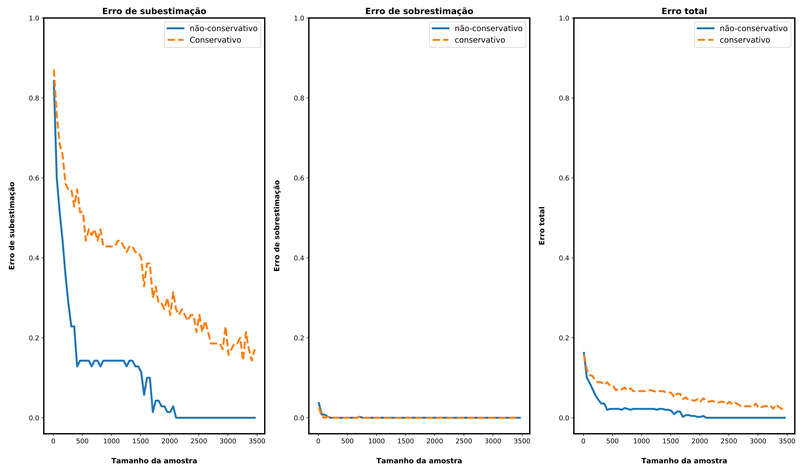

Figura 2: Erros de subestimação, de sobrestimaçao e total do grafo da Figura 1

Para avaliar o algoritmo de Chow-Liu geramos uma árvore e simulamos para tamanho de 10 até 5000.

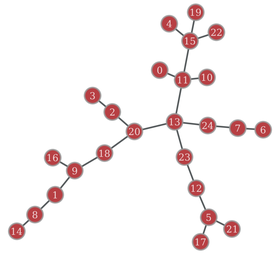

Figura 3: Árvore de 25 vértices usado para gerar amostras e avaliar o algoritmo de Chow-Liu.

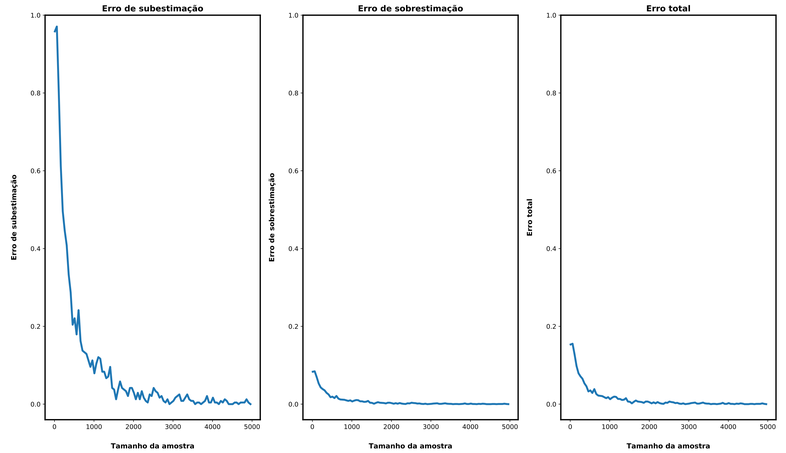

Figura 4: Erros de subestimação, de sobrestimação e total do grafo da Figura 3

Aplicação

Usamos dados de bolsas de valores de 6 países para estimar o grafo de dependências condicionais entre os países. Coletamos índices da bolsa de Bovespa - Brasil, NASDAQ - EUA, FTSE 100 - Reino Unido, CAC 40 - França, Nifty 50 - Índia e Nikkei 225 - Japão a partir do dia 05 de Janeiro de 2001 até 22 de outubro de 2018.

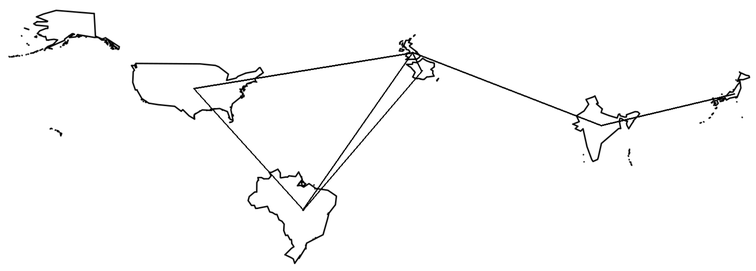

Figura 5: Grafo estimado das 6 bolsas de valores

O grafo estimado mostra que as conexões entre as bolsas de valores estão relaciondas com a aproximação geográfica dos países.

Agredecimento

Agradecemos ao CNPq pelo apoio financeiro durante o desenvolvimento deste trabalho.

IME-USP

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico