Laboratory of biological data mining

Giacomo Fantoni
Telegram: @GiacomoFantoni
Github: https://github.com/giacThePhantom/BioDataMining

November 19, 2021

Contents

Chapter 1

Introduction

Data mining and data analysis are not the same thing: the former deals with data that pre-exists with respect to the research question, whereas the latter deals with data that is generated and collected in order to answer some research question.

1.1 Type of data

Data is canonically represented via tables. It implies the existence of a mapping between facts of the world and symbols. The measurement is a mapping between facts in the world and elements of sets equipped with some mathematical structure.

1.1.1 Categorical data

In categorical data the set is finite and it has no structure. The only operation that can be done on it is to distinguish the elements of the set.

1.1.2 Ordered categorical data

In ordered categorical data the set has an order: its element are in a mathematical relationship with the properties of an order:

- Reflexivity.
- Antisymmetry.
- Transitivity.

1.1.3 Discrete data

In discrete data the set is a subset of the natural numbers. Operations permitted are sum and difference.

1.1.4 Continuous data

In continuous data the set is an interval of the real numbers. Some scales have an absolute zero. The operations permitted are sum and difference. Division makes sense only if the scale is absolute.

1.2 Metadata

Metadata is data about the data. For example it contains informations about its origin, the method, time, format, source and owner.

1.3 Interventional data and observational data

Interventional data and observational data are distinguished on the basis of the intervention on the system. The former is generated by experimental procedures and the latter by pure observation/

Chapter 2

PC algorithm

2.1 Estimating high-dimensional directed acyclic graphs with the PC-algorithm

2.1.1 Introduction

Graphical models are a popular probabilistic tool to analyse and visualize conditional independence relationships between random variables. The major building blocks of these models are nodes (the random variables) and edges (conditional dependence). The structure of conditional independence among the random variables can be explored using the Markov properties. The estimation of a DAG from data is difficult due to the enormous size of the space of DAGs. The PC-algorithm is an alternative to greedy or structurally restricted approaches. It starts from a complete, undirected graph and deletes recursively edges based on conditional independence decisions. This yields an undirected graph that can be partially directed and further extended to represent the underlying DAG. This algorithm runs in the worst case in exponential time, but if the true underlying DAG is sparse, it is reduced to a polynomial runtime. Here there is a focus on estimating the equivalence class and the skeleton of DAGs corresopnding to multivariate Gaussian distributions in high-dimensional context, or where the number of nodes p may be much larger than the sample size n.

2.1.2 Finding the equivalence class of a DAG

Let $G = (V, E)$ a graph. The set of vertices V corresponds to the components of a random vector $\vec{X} \in \mathbb{R}^p$. A probability distribution P on $\mathbb{R}^p$ is said to be faithful with respect to G if conditional independences of the distribution can be inferred from d-separation in G. Consider a random vector $\vec{X} \sim P$. Faithfulness of P with respect to G means that for any $i, j \in V$ with $i \neq j$ and any set $s \subseteq V$:

$$\vec{X}^{(i)} \wedge \vec{X}^{(j)} \text{ are conditionally independent given } \{\vec{X}^{(r)} | r \in s\} \Leftrightarrow i \wedge y \text{ are d-separated by } s$$

Faithfulness is ruling out some classes of probability distributions. The skeleton of a DAG is the undirected graph obtained from G by substituting undirected edges for directed ones. A v-structure in a DAG is an ordered triple of node such that $(i, j) \in G \wedge (k, j) \in G \wedge (i, k) \notin G$. Two DAGs are equivalent is and only is they have the same skeleton and v-structures. If P is faithful with

respect to a DAG G there is an edge between node i and j in the skeleton of DAG G is and only if $\forall s \subseteq V \backslash \{i,j\}, \vec{X}^{(i)} \wedge \vec{X}^{(j)}$ are conditionally dependent given $\{\vec{X}^{(r)}, r \in s\}$. If P is faithful with respect to a DAG G the skeleton of the DAG is a subset to the conditional independence graph corresponding to P. Every edge in the skeleton indicates some strong dependence which cannot be explained by accounting for other variables.

2.1.2.1 PC-algorithm for finding the skeleton

The PC-algorithm betters from a naive strategies that checks for conditional independences given all subsets.

2.1.2.1.1 Population version In the population version of the PC-algorithm perfect knowledge about all necessary conditional independence relations is assumed available.

: $PC_{pop}(V)$

```
l = -1
C = C̃
repeat
    l += 1
    repeat
        Select a new ordered pair of nodes i, j that are adjacent in C such that
         |adj(C, i)\{j}|gel
        repeat
            Choose new k ⊆ adj(C, i)\{j} with |k| = l
            if i ∧ j are conditionally independent given k then
                Delete edge i, j
                Denote this new graph by C
                Save k in S(i, j) and S(j, i)
        until edge i, j is deleted or all k ⊆ adj(C, i)\{j} with |k| = l have been chosen
    until all ordered pairs of adjacent variables i and j such that |adj(C, i)\{j} ≥ l and
     k ⊆ adj(C, i)\{j} with |k| = l have been tested for conditional independence
until for each ordered pair of adjacend nodes i, j : |adj(C, i)\{j}| < l
return Estimated skeleton C, separation sets S
```

The maximal value of l is denoted m_{reach} and depends on the underlying distribution. Considering a DAG G and assume that the distribution P is faithful to G. Denote the maximal number of neighbours by $q = \max_{1 \leq j \leq p} |adj(G,j)|$. Then the PC_{pop} algorithm construct the true skeleton of the DAG. Moreover $m_{reach} \in \{q-1, q\}$.

2.1.2.1.2 Sample version For finite sample there is a need to estimate conditional independencies. Assuming faithful models (conditional independence relations correspond to d-separations) in the Gaussian case conditional independences can be inferred from partial correlations. Assume that distribution P of the random vector $\vec{X}$ is a multivariate normal. For $i \neq \in \{1, \ldots, p\}, k \subseteq \{1, \ldots, p\} \backslash \{i,j\}$, denote $\rho_{i,j|k}$ the partial correlation between $\vec{X}^{(i)}$ and $\vec{X}^{(j)}$ given $\{\vec{X}^{(r)}, r \in k\}$ then $\rho_{i,j|k} = 0$ if and only if $\vec{X}^{(i)}$ and $\vec{X}^{(j)}$ are conditionally independent given $\{\vec{X}^{(r)}, r \in k\}$. The partial correlations can be estimated and the sample partial correlation $\hat{\rho}_{i,j|k}$ can be calculated, for some $h \in k$:

$$\rho_{i,j|k} = \frac{\rho_{i,j|k\setminus h} - \rho_{i,h|k\setminus h}\rho_{j,h|k\setminus h}}{\sqrt{(1-\rho^2_{i,h|k\setminus h})(1-\rho^2_{j,h|k\setminus h})}}$$

For testing whether a partial correlation is zero or not Fisher's z-transform is applied:

$$Z(i,j|k) = \frac{1}{2}\log(\frac{1=\hat{\rho}_{i,j|k}}{1-\hat{\rho}_{i,j|k}})$$

The null hypothesis is rejected $H_o(i,j|k) : \rho_{i,j|k} = 0$ against the two sided alternative $H_A(i,j|K) : \rho_{i,j|k} \neq 0$ if $\sqrt{n-|k|-3}Z(i,j|k) > \Phi^{-1}(1-\frac{\alpha}{2})$, where Φ denotes the cdf of $\mathcal{N}(0,1)$. The sample version of the PC-algorithm is the same of the population version, with the if statement replaced by $\sqrt{n-|k|-3}Z(i,j|k) > \Phi^{-1}(1-\frac{\alpha}{2})$. This algorithm yields a data-dependent $\hat{m}_{reach,n}$. The only tuning parameter is α, the significance level for testing partial correlations. This algorithm is asymptotically consistent even if p is much larger than n but the DAG is sparse.

2.1.2.2 Extending the skeleton to the equivalence class

While finding the skeleton the separations sets that made edges drop out were recorded by S. This is essential for extending the skeleton to the equivalence class.

: `Extending_to_CPDAG(`G_{skel}, S`)`

for all *pairs of non adjacent* i, j *with common neighbour* k **do**
  if $k \notin S(i,j)$ **then**
    Replace $i - k - j \in G_{skel}$ with $i \to k \leftarrow j$

In the resulting PDAG try to orient as many undirected edges as possible by repeated application of:

R1 orient $j - k$ into $j \to k$ whenever there is $i \to j$ such that i and k are non-adjacent.

R2 orient $i - j$ into $i \to j$ whenever there is a chain $i \to k \to j$.

R3 orient $i - j$ into $i \to j$ whenever there are two chains $i - k \to j$ and $i - l \to k$ such that k and l are non-adjacent.

R4 orient $i - j$ into $i \to j$ whenever there are two chains $i - k \to l$ and $k \to l \to j$ such that k and l are non-adjacent

return CPDAG G

2.2 A computing system for discovering causal relationships among human genes to improve drug repositioning

The genehome project aims to expand gene networks using transcriptomic datasets. For human data the objective is to provide a public resource to navigate and combine results by expanding each single human transcript. The platform hosted the NES^2RA algorithm. Starting from a local gene network LGN based on previous biological knowledge, its expansion consists in a set of genes and a list of interactions which describe putative causal relationships with the genes in the LGN. The expansion is calculated on observational gene expression data organized in a coherent normalized data matrix. To overcome the problem of unique elaborations OneGenE has been developed. The list of gene expansions is calculated for each gene in an organism by systematically running single-gene NES^2RA expansions with fixed parameters and then combine them afterwards to simulate LGN expansions. The expansions of the gene networks is based on the transcriptomic dataset provided

by the FANTOM project. Their data comes from sequencing of RNA extracted from different samples of human tissues and cell lines and contains expression profiles of 201802 gene isoforms. Drug repositioning is an alternative approach for the discovery of new therapeutic opportunities for already approved medicines. This method, which relies on previous knowledge, speeds up the approval procedure of the drug regulators and can represent a valuable approach.