Statistical Inference 2 Summer Term 2025

Lecture: Prof. Dr. Göran Kauermann
Exercises: Sergio Buttazzo, Jan Anders
Tutorial: Eugen Gorich

Exercise Sheet 7

Exercise 7.1 - Log-Linear Models

A survey administered to statistics students gathered data on how they like to spend their free time in a typical week. In particular, for each student, we have information on the following variables:

$$M = \text{Number of movies watched}$$

$$B = \text{Number of books read}$$

$$V = \text{Hours spent playing videogames}$$

with $M \leq 7$, $B \leq 2$, and $V \leq 20$. We are interested in studying the behaviour of the students with respect to these three aspects.

(a) Is it possible to sketch a contingency table for this type of data? If yes, outline how, if not, explain why.

(b) Suppose we want to jointly model the probability of a student watching a certain number of movies, reading a certain number of books and playing a certain amount of hours of videogames in a given week. Specify a log-linear model for this joint probability, assuming independence between the three variables. Explain the structure of the model and what all variables mean.

(c) A sociologist comes into the picture, and tells you that it is likely for some of these quantities to depend on each another. In particular, the number of books read is likely to have some interdependence with both the hours of videogames played and the number of movies watched. How can we incorporate these dependencies into the model defined in point b)?

(d) Given the assumptions made in point c), are the variables M and V independent? Are they conditionally independent? Sketch a dependency graph of the three variables.

(e) Propose a hierarchical sequence of log-linear models for the joint distribution of the variables M, B, and V, starting from the complete independence model and ending with the saturated model containing all interaction effects. For each model, specify which interaction terms are included and compute the number of free parameters. Given the information in this exercise, which model would you consider most appropriate for analysis, and why?

Exercise 7.2 - Multivariate Normal Distribution

Suppose we have a system X consisting of temperature (T, in °C), humidity (H, in %), air quality index (AQI, European Air Quality Index), and energy consumption for air conditioning (EC, in Watts). We are particularly interested in the expected energy consumption of our system in summer, i.e. the conditional distribution of energy consumption given high values of temperature, humidity, and (for the sake of this exercise) air quality index.

(a) Define and draw the graphical model associated with the system. More specifically, construct a graph where temperature, humidity, air quality index, and energy consumption are represented as nodes. The edges should capture the conditional dependencies between the variables based on the following domain knowledge:

$$X \sim N\left(\begin{pmatrix}\mu_T\\ \mu_H\\ \mu_{AQI}\\ \mu_{EC}\end{pmatrix}, \begin{pmatrix}\sigma_T^2 & \sigma_{T,H} & \sigma_{T,AQI} & \sigma_{T,EC}\\ \sigma_{H,T} & \sigma_H^2 & \sigma_{H,AQI} & \sigma_{H,EC}\\ \sigma_{AQI,T} & \sigma_{AQI,H} & \sigma_{AQI}^2 & \sigma_{AQI,EC}\\ \sigma_{EC,T} & \sigma_{EC,H} & \sigma_{EC,AQI} & \sigma_{EC}^2\end{pmatrix}\right)$$

$$= N\left(\begin{pmatrix}18\\ 55\\ 50\\ 400\end{pmatrix}, \begin{pmatrix}40 & -10 & -2 & 200\\ -10 & 40 & 13 & -5\\ -2 & 13 & 50 & 5\\ 200 & -5 & 5 & 1450\end{pmatrix}\right)$$

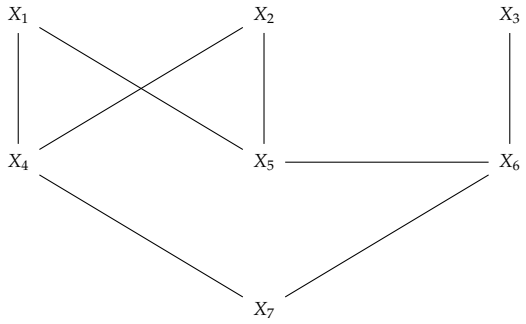
(b) Since it is cumbersome to derive the conditional distribution of interest analytically, we now want to obtain it by simulation. Using the Metropolis-Hastings algorithm, generate a sample from the **full (unconditional) joint distribution** in R. You may want to use the package `mvtnorm` to sample from and calculate the density of a multivariate normal.
Ensure that your Markov chain has reached a stable distribution for the full joint distribution, and then use it to obtain an approximate i.i.d. sample of the conditional distribution of the energy consumption given $T > 25$, $H > 50$ and $AQI > 50$. Estimate the mean and variance of the energy consumption under these constraints.

(c) Are there any modifications or enhancements you would suggest to improve the performance or efficiency of the Metropolis-Hastings algorithm here? In other words, is there a faster way to obtain a sample from the conditional distribution?

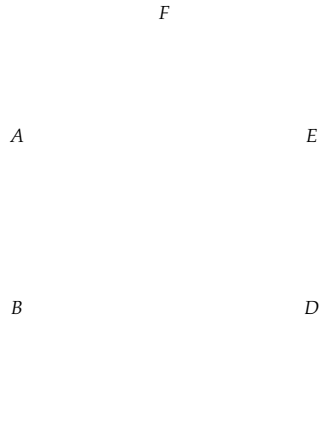
Exercise 7.3 - MV Normal and Graphical Models

(a) Let $\epsilon = (\epsilon_1, ..., \epsilon_n)$ follow an n-dimensional multivariate normal distribution and let $\text{Cov}(\epsilon)$ be a diagonal matrix. Show that $\epsilon_1, ..., \epsilon_n$ are independent. Is this still true if $Cov(\epsilon)$ is not a diagonal matrix?

(b) Below, you are given a Directed Acyclic Graph (DAG) depicting the causal dependence between seven random variables. For each variable $X_i,\ i = 1, ..., n$, state the variables that it is independent of. Additionally, factorise the joint distribution $f_{X_1,X_2,X_3,X_4,X_5,X_6,X_7}(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ of this model, i.e. define it as the product of multiple conditional distributions.



(c) Given the conditional independencies next to the empty graph below, fill in the (undirected) edges, assuming all variables to be dependent unless stated otherwise.
Hint: Start with a full graph, then remove edges one by one.



$$A \perp\!\!\!\perp \{C, D, E, F\} \mid B$$
$$B \perp\!\!\!\perp \{C, F\} \mid \{D, E\}$$
$$C \perp\!\!\!\perp \{A, B, F\} \mid D$$
$$F \perp\!\!\!\perp \{A, B, E\} \mid \{C, D\}$$