Fundamental of Statistical Learning

November 19-20 2025

Classwork-01

Question 1

The file `CRSPday.csv` contains information on the returns for specific stocks (e.g. `IBM`) and the `CRSP` index.

Tomorrow I'll select two of them, and your task will be to describe (at least qualitatively) a joint (bivariate) probabilistic model, supporting your claim with relevant comments + plots/summaries in `R`.

Question 2

Remark: on this question simply do your best and comment whatever result you get.

The `CLT` *as-we-know -it* is eminently about random **variables**, not **vectors**. Nevetheless, to get its multivariate version (also in our notes), we can resort to the following neat result:

Cramer-Wold Device: let $\{\boldsymbol{Y}_n\}_{n>0}$ be a sequence of p-dimensional random vector, and $\boldsymbol{Y}$ a target p-dimensional random vector. Then $\boldsymbol{Y}_n \overset{\text{d}}{\to} \boldsymbol{Y}$ if, for **all** norm-1 vectors $\boldsymbol{\gamma}$, we get

$$\langle \boldsymbol{\gamma}, \boldsymbol{Y}_n \rangle = \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Y}_n \xrightarrow{\text{d}} \langle \boldsymbol{\gamma}, \boldsymbol{Y} \rangle = \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Y}.$$

Nice... but, this result, after some smart mathematical "massaging", can also be turned upside down as follows:

1. let $\boldsymbol{X} \sim F_{\boldsymbol{X}}$ and $\boldsymbol{Y} \sim F_{\boldsymbol{Y}}$ be two p-dimensional random vectors whose distributions $F_{\boldsymbol{X}}$ and $F_{\boldsymbol{Y}}$ are **different**;
2. let $\boldsymbol{\gamma} \sim F_{\boldsymbol{\gamma}}$ be another "fully" continuous **random** vector **independent** from the other two;
3. then the **distributions** of the projections $\boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Y}$ and $\boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{X}$ of $\boldsymbol{X}$ and $\boldsymbol{Y}$ on the one-dimensional subspace generated by $\boldsymbol{\gamma}$ will **differ** (with probability 1).

Taking into account that the distribution of the projections coincide if $\boldsymbol{X} \overset{\text{d}}{=} \boldsymbol{Y}$, we now have a way to check (for now qualitatively) if the distribution behind some *multivariate* dataset follow some specific *multivariate* model we have in mind.

Example: if $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\} \overset{\text{IID}}{\sim} F_{\boldsymbol{X}}$, and $\boldsymbol{Y}$ is uniformly distributed over the p-dimensional hypercube $[0,1]^p$, we could generate one (or more) vector(s) $\boldsymbol{\gamma}_0$ from a convenient $F_{\boldsymbol{\gamma}}$, and then compare the distribution of the projections $\{\boldsymbol{\gamma}_0^{\mathsf{T}} \boldsymbol{X}_1, \dots, \boldsymbol{\gamma}_0^{\mathsf{T}} \boldsymbol{X}_n\}$ with the distribution – exact or approximated by simulation from $F_{\boldsymbol{Y}}$ – of $\boldsymbol{\gamma}_0^{\mathsf{T}} \boldsymbol{Y}$. Since these are realizations of random variables (not vectors), the comparison can be easily done visually (with histograms or similar graphics) and numerically (upon selecting a suitable distance between distributions).

Hint: play around with this idea by picking a small p (like 2 or 3) and simple *multivariate* distributions for $F_{\boldsymbol{X}}$ and $F_{\boldsymbol{\gamma}}$. Notice that the latter may well have independent components.

Bonus Question (optional)

We said that there's an entire industry devoted to forge "correlation measures". Often interesting, sometimes not so much... but still published in Science! As one of my prof once said:

"*This is a travesty! It has bad power and other bad properties and they were not even aware of the enormous statistical literature on the topic. A good example of non-statisticians inventing statistics and then getting it in Science!*"

For example, what happens to this correlation measure $\text{MIC}(X, Y)$ when we take $X \sim \text{Unif}(-1, 1)$, $Z \sim \text{Unif}(-1, 1)$ independent of X, and $Y = Z$ if X and Z have the same sign, else $Y = -Z$? Mah... explore... somehow...
