Fundamental of Statistical Learning

October 22-23 2025

Classwork-00 | The Drill

Question 1

The file `Country-data.csv` contains information on different socio-economic and health factors that determine the overall development of countries around the world.

Tomorrow I'll select three of them, and your task will be to propose a probabilistic model for one of them (of your choice), supporting your claim with relevant comments + plots/summaries in `R`.

Question 2

Remark: on this question simply do your best, meaning that you can also come to the opposite conclusion for lack of time. Just comment whatever result you got.

The `mixing-trick` has been used in a variety of contexts for a variety of reasons: from extending the expressiveness of a base model (also deep networks) to capture the variability of natural images.

Here you'll focus on the latter, and in particular on the paper `Scale Mixtures of Gaussians and the Statistics of Natural Images`.

The question is: if you want to empirically check (or defend) the claim in this paper, what would you do?

For sure you need a suitable dataset (and `Kaggle` might help here), and for sure you need to know what does it mean to *wavelet transform* an image. `Section 21.4` of the **purple book** might be useful here, but to simplify at a minimum, you "just" need to know the following (see also the technical Appendix):

1. A (digital) image can be thought as bivariate function $f(x, y)$ sampled on a regular grid.
2. Under some technical conditions, it can be proved that large classes of (sampled) functions can be represented as *linear combinations* of simpler, fixed, basis functions (or atoms).
3. Wavelets are very famous (and successful) basis functions able to capture local features in pixel-space and frequency. They also come with a very fast algorithm (the *Discrete Wavelet Transform* or `DWT`) that calculates the coefficient of the linear combination (the *wavelet coefficients* of the paper) in linear time.
4. The paper is about the statistical properties of these coefficients.
5. In `R`, the package `jpeg` can be used to import an image (see `?readJPEG`) and the package `waveslim` can be used for the `DWT` (see `?dwt.2d`). There are different types of implemented wavelet families. You can work with the `haar` no problem. For technical reason, the `DWT` works best when the number of pixels is a power of two. In case it is not, you can always *pad* with zeros the original image (stored as a matrix).

Bonus Question (optional)

Today we gave a first look at the `ELBO`, and we said it is a way to relax a complex model selection model into a larger but simpler one to handle. In the end, we got that the gap was given by

$$\log p_\theta(\texttt{obs}) - \mathcal{F}(\theta, q) = -\mathbb{E}_q\left[\log \frac{p_\theta(\texttt{obs} \mid \texttt{hid})}{q(\texttt{hid})}\right] = \mathrm{KL}\big(q(\texttt{hid}) \,\big\|\, p_\theta(\texttt{obs} \mid \texttt{hid})\big)$$

where $\mathcal{F}(\theta, q) = \mathbb{E}_q\left[\log \frac{p_\theta(\texttt{obs},\texttt{hid})}{q(\texttt{hid})}\right]$ is our lower bound driven by $q(\cdot)$, the arbitrary distribution over the latent/hidden variable we designed and picked.

For any fixed θ, the best choice (when available) for $q(\cdot)$ is the posterior distribution $p_\theta(\texttt{obs} \mid \texttt{hid})$. But what happen when we pick another, possibly simpler one?

You can try to numerically investigate this in a setup where you know everything, for example, when $p_\theta(\texttt{obs})$ is a `Beta-Binomial` and we know exactly how to represent it as a mixture of Binomial (…actually, you also know the posterior!).

Appendix: Wavelet Decomposition of a Bivariate Function

The 1D Case

For a one-dimensional function $f(x) \in L^2(\mathbb{R})$, the wavelet decomposition is

$$f(x) = \sum_k c_{J,k}\,\phi_{J,k}(x) + \sum_{j=J}^{\infty}\sum_k d_{j,k}\,\psi_{j,k}(x),$$

where

$$\phi_{J,k}(x) = 2^{J/2}\phi(2^J x - k), \qquad \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k),$$

are, respectively, the *scaling* and the *wavelet* basis functions. The coefficients are obtained by projection:

$$c_{J,k} = \langle f, \phi_{J,k}\rangle, \qquad d_{j,k} = \langle f, \psi_{j,k}\rangle.$$

The 2D (Bivariate) Case

For a bivariate function (e.g., an image) $f(x,y) \in L^2(\mathbb{R}^2)$, the wavelet basis is typically constructed from tensor products of the 1D scaling and wavelet functions.

$$\begin{aligned}
\phi_{j,k_1,k_2}(x,y) &= 2^j\,\phi(2^j x - k_1)\,\phi(2^j y - k_2),\\
\psi^{(H)}_{j,k_1,k_2}(x,y) &= 2^j\,\psi(2^j x - k_1)\,\phi(2^j y - k_2),\\
\psi^{(V)}_{j,k_1,k_2}(x,y) &= 2^j\,\phi(2^j x - k_1)\,\psi(2^j y - k_2),\\
\psi^{(D)}_{j,k_1,k_2}(x,y) &= 2^j\,\psi(2^j x - k_1)\,\psi(2^j y - k_2).
\end{aligned}$$

Here:

- (H): horizontal detail (high frequency in x);
- (V): vertical detail (high frequency in y);
- (D): diagonal detail (high frequency in both directions).

The two-dimensional wavelet expansion of $f(x,y)$ is then:

$$\boxed{\begin{aligned}
f(x,y) = &\sum_{k_1,k_2} c_{J,k_1,k_2}\,\phi_{J,k_1,k_2}(x,y)\\
&+ \sum_{j=J}^{\infty}\sum_{k_1,k_2}\Big(d^{(H)}_{j,k_1,k_2}\,\psi^{(H)}_{j,k_1,k_2}(x,y) + d^{(V)}_{j,k_1,k_2}\,\psi^{(V)}_{j,k_1,k_2}(x,y) + d^{(D)}_{j,k_1,k_2}\,\psi^{(D)}_{j,k_1,k_2}(x,y)\Big).
\end{aligned}}$$

The coefficients are computed via inner products:

$$\begin{aligned}
c_{J,k_1,k_2} &= \iint f(x,y)\,\phi_{J,k_1,k_2}(x,y)\,dx\,dy,\\
d^{(H)}_{j,k_1,k_2} &= \iint f(x,y)\,\psi^{(H)}_{j,k_1,k_2}(x,y)\,dx\,dy,\\
d^{(V)}_{j,k_1,k_2} &= \iint f(x,y)\,\psi^{(V)}_{j,k_1,k_2}(x,y)\,dx\,dy,\\
d^{(D)}_{j,k_1,k_2} &= \iint f(x,y)\,\psi^{(D)}_{j,k_1,k_2}(x,y)\,dx\,dy.
\end{aligned}$$

We can interpret these components as follows:

- c_{J,k_1,k_2}: approximation (low-pass) coefficients;
- $d^{(H)}, d^{(V)}, d^{(D)}$: detail (high-pass) coefficients capturing horizontal, vertical, and diagonal structures;
- In discrete form, these coefficients are efficiently computed using separable filter banks and downsampling.