Variations of the CP

Many names of CP

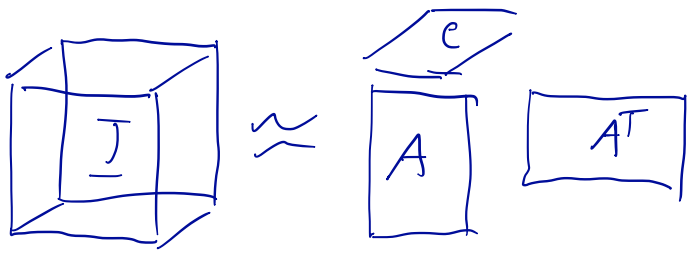
- Polyadic form of a Tensor (1927)
- PARAFAC (Parallel factors) (1970)
- CANDECOMP (Canonical decomp.) (1970)
- CP (CANDECOMP/PARAFAC) (2000)

All these are the same decomposition, often invented independently as a tensor equivalent of matrix factorizations

INDSCAL

INDSCAL (Individual Differences in Scaling) has two factor matrices and symmetric frontal slices:

$$\underline{T} \approx [\![A, A, C]\!]$$



Usually on INDSCAL we assume that the tensor is symmetric on two modes, but this is not necessary. They do have to have the same dimensions, though.

N-way INDSCAL is

$$[\![A^{(1)}, A^{(2)}, \dots, A^{(k-1)}, B, A^{(k+1)}, \dots, A^{(n-1)}, B, A^{(n+1)}, \dots, A^{(N)}]\!],$$

which is $I_1 \times I_2 \times \dots \times I_{k-1} \times J \times I_{k+1} \times \dots \times I_{n-1} \times J \times I_{n+1} \times \dots \times I_N$.

INDSCAL is usually computed by first computing a CP decomposition, and "hoping" that two factor matrices converge to similar. One can also enforce similarity either at the end, or during the update

$$\left.\begin{aligned} A &\leftarrow \mathcal{T}_{(1)} (C \odot A)^{+} \\ A &\leftarrow \mathcal{T}_{(2)} (C \odot A)^{+} \end{aligned}\right\} \text{same if } \mathcal{T} \text{ is symm. on 1st and 2nd mode}$$

$$C \leftarrow \mathcal{T}_{(3)} (A \odot A)^{+}$$

A <u>supersymmetric CP</u> will have just one factor matrix: $\mathcal{T} \approx [\![A, A, A]\!]$

Nonnegative CP

CP decomposition can be constrained to nonnegative input tensors and factor matrices to obtain a tensor version of NMF. Like with NMF, nonnegativity helps with interpretation and sometimes with sparsity of the factors.

$$\mathbb{R}_+^{I\times J\times K} \ni \underline{T} \approx [\![A, B, C]\!], \quad A \in \mathbb{R}_+^{I\times R}, \; B \in \mathbb{R}_+^{J\times R}, \; C \in \mathbb{R}_+^{K\times R}$$

The projected ALS is the simplest algorithm for NCP:

$$A \leftarrow \left[\underline{T}_{(1)}(C \odot B)^{+}\right]_+ \quad \text{etc.}$$

where

$$[x]_+ = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

As with NMF, we can also develop multiplicative update rules. Starting from random $A \in \mathbb{R}_+^{I\times R}$, $B \in \mathbb{R}_+^{J\times R}$, and $C \in \mathbb{R}_+^{K\times R}$, we update elementwise

$$a_{ir} \leftarrow a_{ir} \frac{\sum_{j,k} b_{jr} c_{kr} (t_{ijk} / q_{ijk})}{\sum_{jk} b_{jr} c_{kr}}$$

$$b_{jr} \leftarrow b_{jr} \frac{\sum_{i,k} a_{ir} c_{kr} (t_{ijk} / q_{ijk})}{\sum_{i,k} a_{ir} c_{kr}}$$

$$c_{kr} \leftarrow c_{kr} \frac{\sum_{i,j} a_{ir} b_{jr} (t_{ijk} / q_{ijk})}{\sum_{i,j} a_{ir} b_{jr}}$$

As all involved entries are nonnegative, the update rules never violate the non-negativity constraints. They can be derived from the gradients. For A, we have

$$\frac{\partial}{\partial a_{ir}} \| \mathcal{T} - [\![A, B, C]\!] \|^2 = -\sum_{j,k} (t_{ijk} / q_{ijk}) b_{jr} c_{kr}$$

and setting

$$a_{ir} \leftarrow a_{ir} - \eta_{ir} \frac{\partial}{\partial a_{ij}} \| \mathcal{T} - [\![A, B, C]\!] \|^2$$

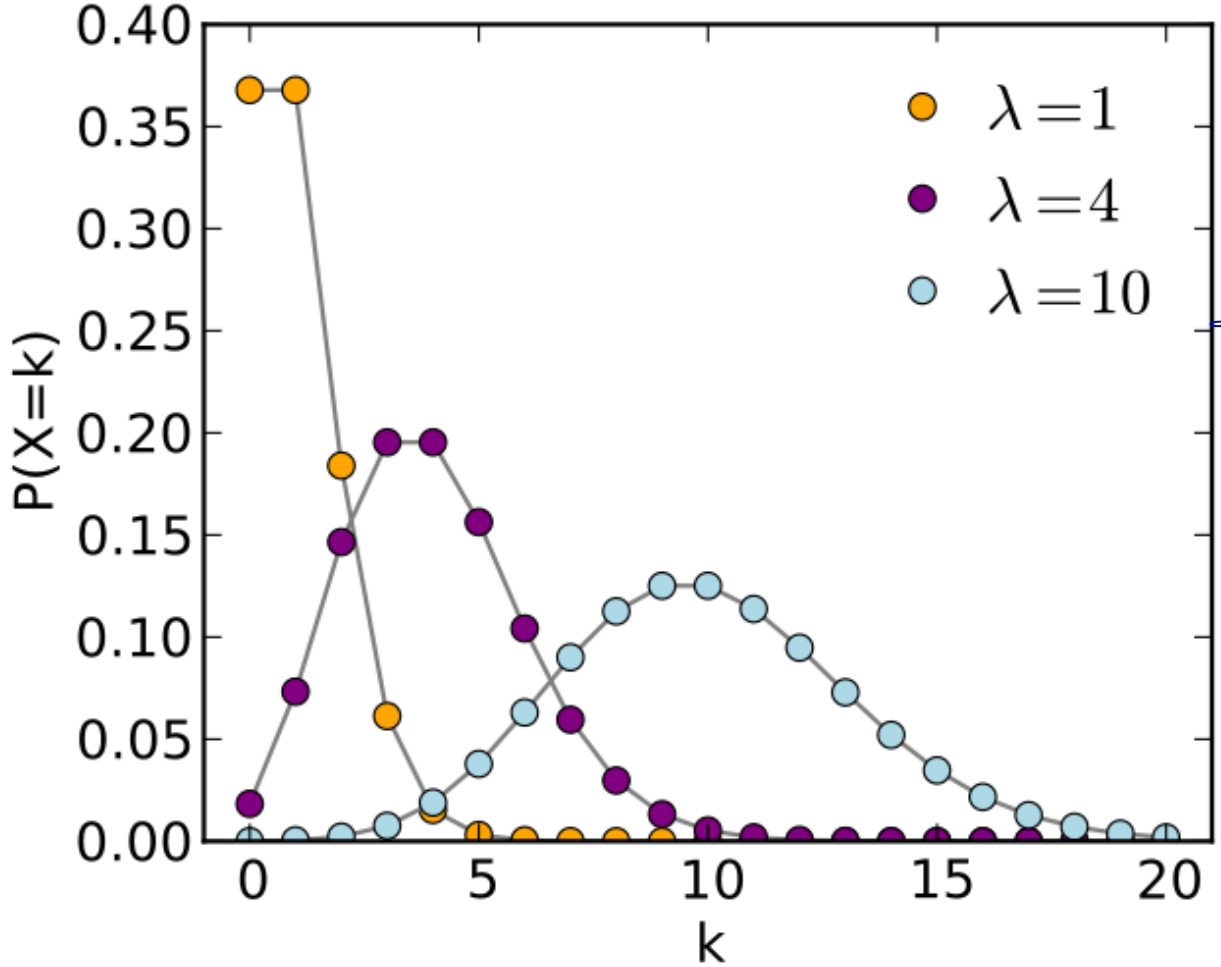
with step size $\eta_{ir} = a_{ir} / \sum_{j,k} b_{jr} c_{kr}$.

CP with alternating Poisson regression

Using the least-squares error has the implicit assumption of Gaussian noise. With certain types of data, this assumption is not very sensible. For example, in case of counting data with low counts, Gaussian noise is not a good model. (E.g. emails from i to j; students i taking courses c.) Such data is often better explained with the <u>Poisson distribution</u>. The Poisson distribution is the probability of number of events occurring in a fixed interval if they occur on known average rate and independently.

The Poisson distribution has a one parameter, the <u>rate</u> $\lambda > 0$. The probability mass function is $f(k;\lambda) = \lambda^k e^{-\lambda}/k!$ and if $X \sim \text{Poisson}(\lambda)$, then $\mathbb{E}[X] = \text{Var}[X] = \lambda$.

The pmf of a Poisson



Source: Wikipedia

Notice that with larger values of λ the distribution starts resembling the Gaussian distribution.

To model the data, we assume that the elements t_{ijk} are i.i.d. Poisson distributed. We assume that the parameters come from a parameter

tensor $\underline{\Lambda}$ that has a low-rank structure. That is, we assume that there exists A, B, and C such that

$$t_{ijk} \sim \text{Poisson}\left(\sum_r a_{ir} b_{jr} c_{kr}\right)$$

for all i, j, k. We aim to find the most likely A, B, and C given the data tensor $\underline{T}$ and the rank R. Hence, the quality is measured using the <u>log-likelihood</u>

$$\log L(t_{ijk}; A, B, C) = t_{ijk} \ln(\lambda_{ijk}) - \lambda_{ijk} - \ln(t_{ijk}!),$$

where $\lambda_{ijk} = [\![A, B, C]\!]_{ijk}$ is the parameter. As $-\ln(t_{ijk}!)$ is constant, we can <u>minimize</u> the <u>negative</u> log-likelihood

$$-\log L \propto \sum_i \sum_j \sum_k \left(\lambda_{ijk} - t_{ijk} \log(\lambda_{ijk})\right).$$

This function is in fact the KL-divergence between the data and the model.

We agree that $0 \cdot \log(x) = 0$ for all $x \geq 0$. But if $\lambda_{ijk} = 0$ and $t_{ijk} > 0$, we have

modelled an observation with a zero probability and we have $t_{ijk}\log(\lambda_{ijk}) = -\infty$. To prevent this to never happen, we require that if $t_{ijk} > 0$ then $\lambda_{ijk} > 0$, too.

To solve this model, called CP-APR, we first write it open:

$$L(A) = A(C \odot B)^T - T_{(1)} * \log(A(C \odot B)^T).$$

We can have similar equations for the other modes, as well. Optimizing for A in $L(A)$ is not trivial, though. We can use the following update rule

$$A \leftarrow A * \left(T_{(1)} \oslash \left(A(C \odot B)^T\right)\right)(C \odot B)^T,$$

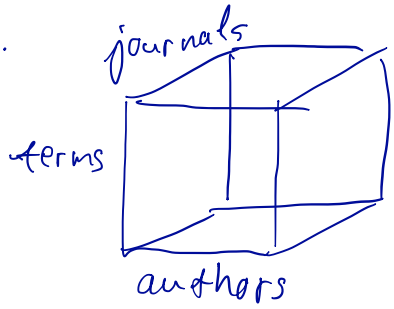
where $\oslash$ is the element-wise division.

We can run this multiplicative update rule a few times before optimizing for the next factor matrix.

Applications of CP and its variations

Applications of CP-APR

CP-APR is designed to model counting data. Consider a terms-authors-journals tensor $\underline{T}$, where t_{ijk} is the count how many times term i was used by author j in the titels of articles published in journal k.



Considering the factor matrices of CP-APR for this data, we notice that if we have a rank-1 component $\vec{a}_r \circ \vec{b}_r \circ \vec{c}_r$ with high

values, that component corresponds to some "topic" used by some authors in some journals. For example, from a $4952 \times 6955 \times 11$ tensor of data from SIAM journals, Chi & Kolda (2012) obtained the results in the next page.

Applications of CP

TOPHITS

TOPHITS is a link-based way to rank web pages similarly to HITS. The data is $\underline{T} = (t_{ijk})$, where t_{ijk} is the number of times page i links to page j using term k in the anchor text. Computing the CP (or CP-APR) behaves similarly to HITS: each rank-1 component is one topic, A and B give the authority

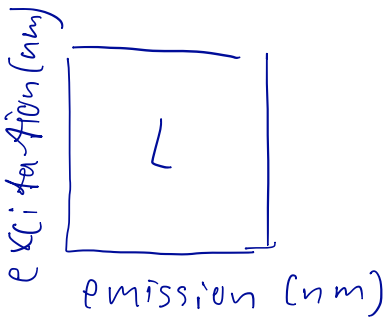
and hub scores, respectively. Matrix C gives weights for each term in each topic.

Fluorescence Excitation-Emission Analysis

Fluorescence spectroscopy is a method used to analyze organic compounds. It is based on pointing a beam of UV light to the compound. The light excites electrons in molecules, and each excited electron releases a photon, which can be observed.

Different materials react differently to different wavelengths, emitting photons of different wavelengths. A <u>fluorescence landscape</u> of a compound is a rank-1 matrix that maps the

exciter's wavelength to the emitted photon's wavelength. Identifying the landscapes lets us identify the compounds.

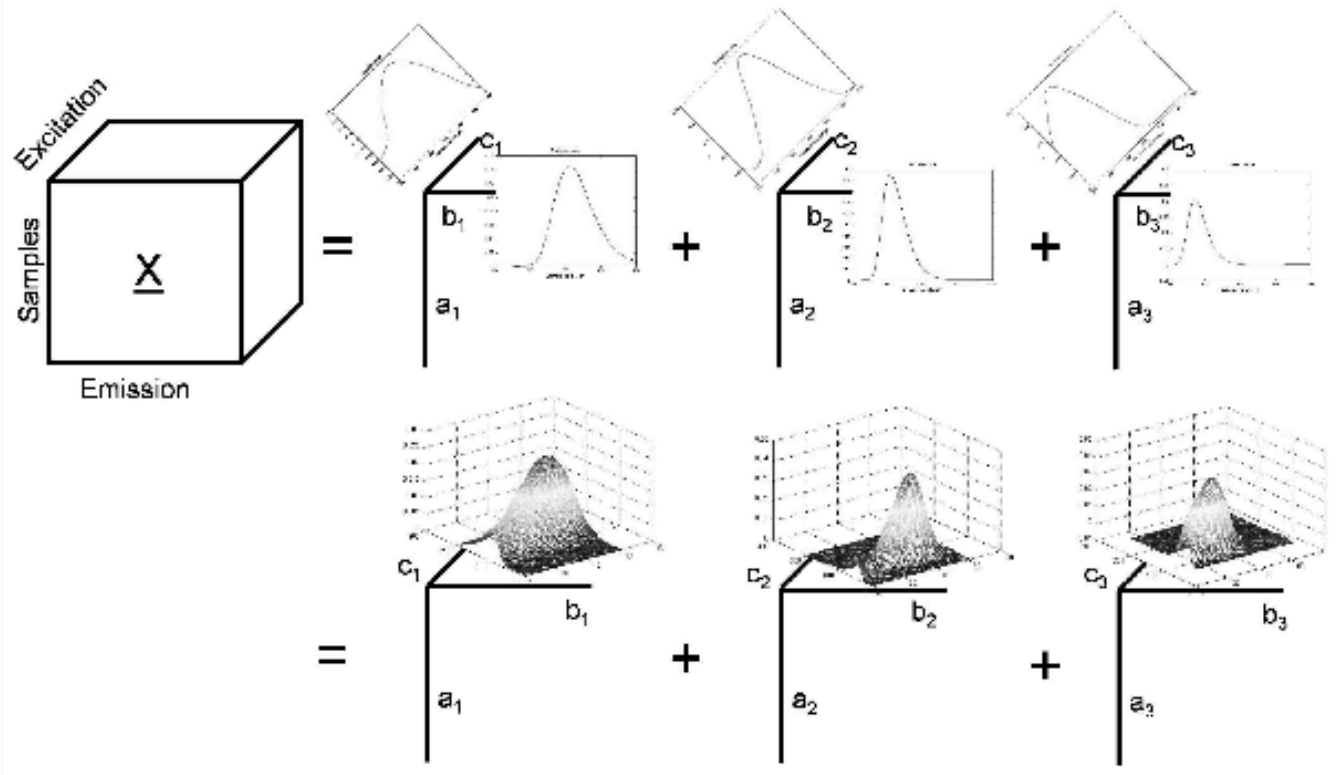


l_{ij} = intensity of excitation in wavelength j when emitted with wavelength i.

Our data is a samples-by-emission wavelength-by-excitation wavelength tensor. We assume that each sample is a linear combination of the sampled combounds' landscapes. If we do a CP on this tensor to get A, B, and C, we can interpret the factors as follows:

$\vec{b}_r \vec{c}_r^{\,T}$ is the rth landscape

$\vec{a}_r$ gives the abundancies of compound r in different samples



Example of fluorescence analysis from Acar & Yener (2009). Bottom row shows three landscapes.

Applications of INDSCAL

INDSCAL was proposed for analysing psychological data (Carroll & Chang, 1970). In particular, they wanted to analyze how similar or dissimilar people find things, and what are the underlying factors for these similarities. Their hypothesis is that there are a small number of (latent) reasons, but that the different reasons have different effects

on different people. Hence they collected data where 20 people rated the similarity of a number of countries. Factorizing this with INDSCAL gave the latent similarity matrices $\vec{a}_i \vec{a}_i^T$, and a weight matrix C.

developed vs communist

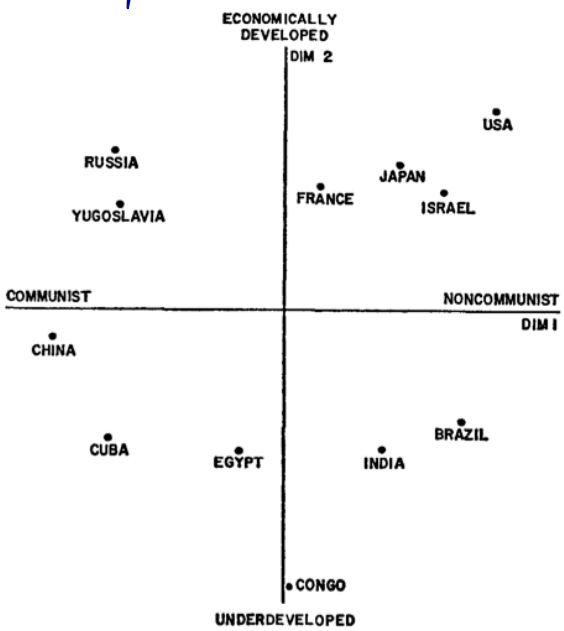


FIGURE 11

west vs communist

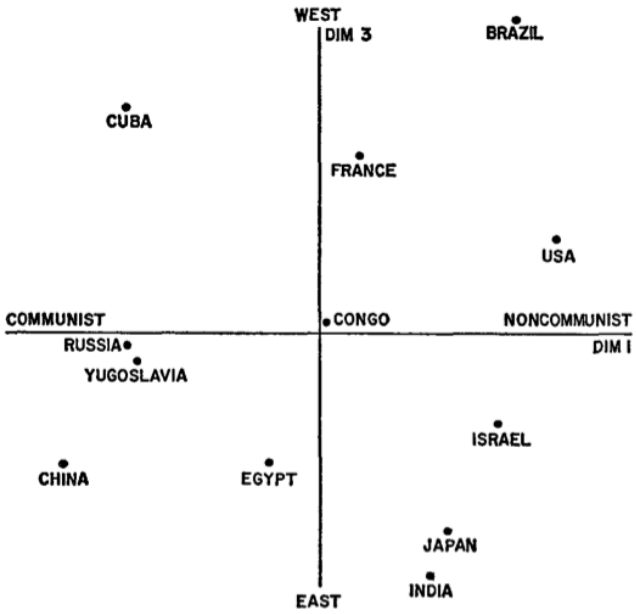


FIGURE 12

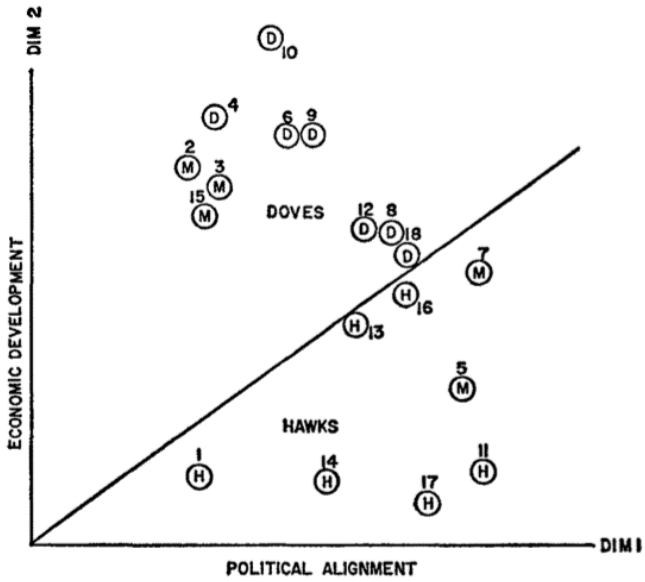


FIGURE 13

Doves and hawks and moderates (D, H, and M) in $\vec{a}_1 \vec{a}_2^T$ scatter plot.