DEST
DEPARTAMENTO DE ESTATÍSTICA

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Latent archetypes of the spatial patterns of cancer

Marcos Prates
Joint work with Mônica De Castro, Renato Assunção and Thais Menezes

UFRJ Seminar 2025

April 9, 2025

Motivation

DEST
UFMG

According to the International Agency for Research on Cancer (IARC), globally in 2020 is estimated:

- 19.3 million new cases of cancer;
- 10.0 million deaths by cancer.

Epidemiological studies are the main tools to identify risk factors that act by modifying the cell genetic code and triggering cancer.

Introduction

DEST
UFMG

- Geographically detailed atlas of cancer mortality led to the knowledge of how strongly spatially patterned is for the occurrence of cancer cases (around 1975);
- Since then, we have had the publication of many National Cancer Atlases;
- These atlases have invariably shown that almost all cancers present a large variance in risk across regions as well as considerable variation when comparing different cancers in the same region;
- Each cancer type is analyzed **individually** in search of plausible risk factors.

Introduction

DEST
UFMG

The main problem

This common practice of individually analyzing the cancer atlas maps, one at a time, represents an enormous burden on the researchers.

Besides that, it also creates scattered knowledge that lacks a connection between risk factors common to several cancers.

In the Literature

DEST
UFMG

There have been occasional exceptions to this general norm when more than one cancer type is analyzed simultaneously

- The main approach adopted in these works is the shared component model proposed by Knorr-Held and Best (2001) - which assumes that the presence of one latent risk factor is simultaneously present in the risk model of two diseases.
- Their work was extended to jointly model the variation of more than two diseases (Gómez-Rubio and Palmí-Perales, 2019; Azevedo et. al, 2021)

The applications of this model, however, have been restricted to a small number of diseases that share a clear and well-known common risk factor

Our Proposal

DEST
UFMG

- We hypothesize that the geographic variation of all/many cancers can be explained by the appropriate combination of a small number of maps representing latent geographic patterns;
- We intend to measure quantitatively the amount of the spatial risk variation in the many individual cancer maps that can be assigned to *common* latent maps;
- We want to see how these common factors combine to produce the observed patterns.

Inspired by the Greek philosopher Plato, we call these small number of latent maps of archetypes - they represent ideal forms of which real cancer maps are shadows made by combining the archetypes in different proportions and adding some degree of noise.

Our Proposal - General Idea

DEST
UFMG

1. Organize the relative risks of a large number of cancers as a matrix with rows corresponding to the spatial locations in the cancer atlas and columns representing the different cancers;
2. Use low-rank approximation techniques to decompose this matrix of relative cancer risks into a sum of spatially structured latent factors
3. As disease risks are positive variables, we were particularly interested in studying low-rank representation that we obtain by using the singular value decomposition (SVD) and non-negative matrix factorization (NMF)

We applied our method to a collection of cancer maps from different regions in the world: the USA, England, France, Australia, Spain, and Brazil.

Methodology

DEST UFMG

1. Let $\mathbf{X}$ be a $n \times q$ matrix (n is the number of regions and q is the number of cancers).
2. $\mathbf{X}_{ij} =$ the mortality or incidence risk estimate for cancer j in area i (standardized with respect to age and sex population distribution)

We approximate the matrix $\mathbf{X}$ by $\mathbf{X}_k$, a sum of few matrices $\mathbf{A}_1, \ldots, \mathbf{A}_k$, each with rank 1, as follows:

$$\mathbf{X} \approx \mathbf{X}_k = \sum_{l=1}^{k} \mathbf{A}_l = \sum_{l=1}^{k} \sigma_l \mathbf{m}_l \mathbf{v}_l^t \qquad (1)$$

Hypothesis

Our main exploratory hypothesis is the existence of $k << q$ latent maps that, duly combined, are able to retrieve approximately all q observed maps in $\mathbf{X}$.

Methodology

DEST
UFMG

To obtain the approximation, we use the singular value decomposition theorem (SVD) that guarantees, for $\mathbf{X}$ with $n > q$, the existence of matrices $\mathbf{M}$ and $\mathbf{V}$, of dimensions $n \times q$ and $q \times q$, respectively, such that:

SVD

$$\mathbf{X} = \mathbf{MSV}^t = \mathbf{MW}$$

Methodology

DEST UFMG

SVD

$$\mathbf{X} = \mathbf{MSV}^t = \mathbf{MW}$$

- $\mathbf{S}$ is a $q \times q$ diagonal matrix with elements $\sigma_l \geq 0$;
- The q columns $\mathbf{m}_l$ of matrix $\mathbf{M}$ are the q orthonormal eigenvectors of the symmetric and semi-definite positive matrix $\mathbf{XX}^t$ associated with its q largest eigenvalues.
- The q columns $\mathbf{v}_i$ of matrix $\mathbf{V}$ are the q orthonormal eigenvectors associated with the symmetric and semi-definite positive matrix $\mathbf{X}^t\mathbf{X}$
- The σ_i diagonal elements of the matrix $\mathbf{S}$ are the singular values, which coincide with the q eigenvalues of $\mathbf{X}^t\mathbf{X}$.

Methodology

DEST UFMG

The SVD approximation implies that the j-th column $\mathbf{x}_j$ of the matrix $\mathbf{X}$, associated with the j-th cancer, can be approximately written as

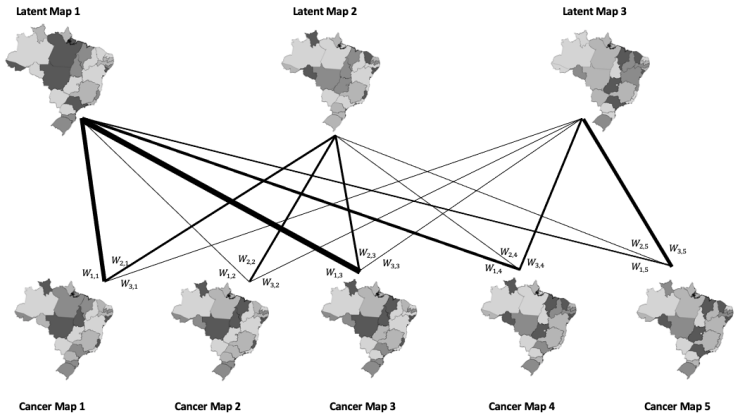
$$\mathbf{x}_j \approx \mathbf{W}_{1,j} \times \mathbf{m}_1 + \ldots + \mathbf{W}_{k,j} \times \mathbf{m}_k \tag{2}$$

The geographic pattern $\mathbf{x}_j$ of the j-th cancer is approximately equal to a linear combination of the first k geographic patterns embedded in the vectors $\mathbf{m}_1, \ldots, \mathbf{m}_k$ - called *latent factors* or *latent archetypes*

Interpretation

There are only k different latent maps or archetypes and each of the p cancers is obtained approximately as a cancer-specific linear combination of these few latent archetypes plus a small random noise.

Methodology - Demonstration



Preparing the Database

DEST
UFMG

- Different cancers have widely different values for the mean level risk in a given map as well as their variation range across areas.
- Different scales and high variability in the data can make similar spatial behavior to be identified as distinct.
- It is necessary to reduce such variability in the data and to use comparable scales between maps.

Standardized Mortality (or incidence) Rate (SMR)

The SMR is a measure of relative risk varying around 1 and hence we can deal with cancers having widely different incidence levels.

Preparing the Database - SMR

DEST UFMG

- The expected numbers are usually calculated using the global empirical rates r_k with k indexing the age-sex classes.

$$e_i = \sum_k \pi_{ik} r_k = \sum_k \pi_{ik} \left(\frac{n_{+k}}{\sum_l \pi_{lk}} \right)$$

where π_{ik} denotes the number of individuals at risk or population size in the k-th category of in the i-th region and $n_{+k} = \sum_l n_{lk}$.

SMR

The SMR in area i is given by the ratio $\text{SMR}_i = y_i / e_i$.

Preparing the Database - SMR

DEST
UFMG

Problem

A major problem with SMR_i is its large variance when the i-th area population is small.

This extreme variation is not due to the risk variation and it should be smoothed out. We avoid methods that require untestable assumptions or impose prior knowledge that may strongly affect the final risk estimates.

Solution

We decide to resort to a highly effective but also simple risk smoothing method, the empirical Bayes proposed by Marshall (1991).

Preparing the Database - SMR

$$\hat{\theta}_i = w_i \text{SMR}_i + (1 - w_i) m_i$$

- m_i is an estimate of $E_\theta(\theta_i)$
- $0 \leq w_i \leq 1$ is an estimate of the variance ratio $\mathbb{V}_\theta(\theta_i)/\mathbb{V}_y(\text{SMR}_i)$ between the underlying relative risk variation and the unconditional variance of the SMRs.

Advantage

The w_i weight is inversely proportional to the region's population so that the risk estimate is shrunk towards the global estimate in the case of small populations.

Number of Maps

DEST
UFMG

The selection of the number k of latent maps capable of explaining the spatial variability of the diseases is a key aspect of our proposal.

A common practice to select k is by plotting the singular values σ_i of $\mathbf{X}$ in decreasing order, and looking for an elbow shape to identify that integer.

A more principled way was proposed by Gavish and Donoho (2014) who study a threshold to recover noisy data from singular values in random matrices.

Their method is asymptotically robust to unknown rank and noise being the optimal choice for thresholding when performing hard truncation to approximate a low-rank matrix under the asymptotic mean squared error.

Number of Maps - Optimal Threshold

The original solution presented by the authors is computationally demanding. However, they show a third-order polynomial approximation for the optimal threshold:

Approximation

$\tau = \omega(\beta) \times \sigma_{med}$

- $\omega(\beta) = 0.56\beta^3 - 0.95\beta^2 + 1.82\beta + 1.43$, with a maximum error of 0.02 and $\beta = \min(m/n, n/m)$.
- σ_{med} is the median of the singular values of the observed data matrix $\mathbf{Y}$.

This is the metric that will be considered to find the total number of latent maps. Therefore, k is the number of singular values of $\mathbf{X}$ that are greater or equal to τ.

Approximation Measure

DEST UFMG

- We define a measure that expresses the amount of variability on the real maps that are explained by the proposed approach;
- The goal is to use the linear combination of k latent maps to capture the actual variability of the observed one.

Considering $\mathbf{X}_k$ as the approximation for the relative risk matrix $\mathbf{X}$ using k latent maps, we measure the approximation quality with:

$$VE(k) = \frac{\text{tr}(\mathbf{X}_k'\mathbf{H}\mathbf{X}_k)}{\text{tr}(\mathbf{X}'\mathbf{H}\mathbf{X})} \times 100 \tag{3}$$

$\mathbf{H} = \mathbf{I}_n - 1/n\mathbf{1}_n\mathbf{1}_n'$ is the centering matrix.

We measure the percentage of the total variance in $\mathbf{X}$ that is explained by $\mathbf{X}_k$.

Simulation Study

DEST
UFMG

The derivation of the optimal threshold τ is based on several assumptions that do not hold in our application such as the Gaussian noise and independent data assumptions.

We considered a simulation study to check the capacity of τ to detect the number of relevant latent maps (k) and how the metric $VE(k)$ can recover the explainable variation in the data.

The method assumes an unknown but unique noise variance σ^2 for the entire matrix $\mathbf{X}$ so before finding τ, all columns in $\mathbf{X}$ are standardized.

Simulation Study

DEST UFMG

To generate realistic simulated data, we used the SVD to the dataset $\mathbf{X}$ from Brazil composed of 557 regions and 30 cancers.

The SVD decomposition produced $\mathbf{X} = \mathbf{M}\,\mathbf{S}\,\mathbf{V}^t$.

Holding fixed the obtained matrix $\mathbf{V}$ and the estimated maps in $\mathbf{M}$ as the ground truth latent maps, we altered the singular values in $\mathbf{S}$ and add some noise to build simulated $\tilde{\mathbf{X}}$ matrices.

Simulation Study

DEST UFMG

We set the first k_T singular values as positive values and made the remaining ones equal to zero taking k_T equal to $2, 4$, or 7.

We considered three different scenarios for the k positive singular values:

1. the same positive singular values for all maps: $\sigma = \sigma_1 = \ldots = \sigma_{k_T}$;
2. linearly decaying singular values: $\sigma_1, \sigma_2 = (k_T - 1)\sigma_1/k_T, \ldots, \sigma_{k_T} = \sigma_1/k_T$;
3. the same singular values as the ones obtained in the Brazilian data decomposition.

Simulation Study

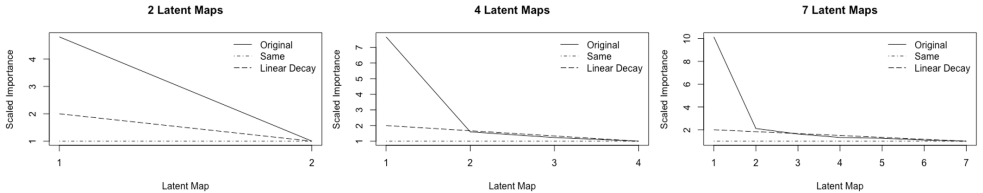


Figure: The singular values (or latent maps' relevance) for each scenario. The horizontal axis presents the number of latent maps while the vertical axis contains the singular value associated with each latent map.

Simulation Study

DEST UFMG

We also added i.i.d. random noise following a $N(0, \phi^2)$ normal distribution to inject controlled randomness in the simulated maps.

The noise variance ϕ^2 was selected to make the generated cancer maps with a certain predefined proportion p of its total variance due to the latent maps rather than the noise.

We set spatial signal levels p for the explained variability as $\{0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.75, 0.90\}$.

Simulation Study

DEST UFMG

To summarize, denote by $\tilde{\mathbf{M}}$ the matrix $\mathbf{M}$ with their k_T-th first columns given by $\tilde{\mathbf{m}}_j = \mathbf{m}_j + \varepsilon_j, j = 1, \ldots, k_T$, with ε_j given by n i.i.d. $N(0, \phi_j^2)$ random variables with $\phi_j^2 = (1-p)\mathbb{V}(\mathbf{m}_j)/p$.

Simulated Matrix

$$\tilde{\mathbf{X}} = \tilde{\mathbf{M}}\, \tilde{\mathbf{S}}_k\, \mathbf{V}^t$$

For each of the 72 scenarios combinations, were generated 1000 data sets. In each one of them, the number of latent maps was selected according to the proposed threshold, and the explained variability percentage $VE(k)$ was obtained.

The same threshold was applied using the NMF Frobenius algorithm to check if it provides adequate results as in the SVD method.

Simulation Study - Equal Weights

In this scenario, we let the k_T latent maps contribute equally to the spatial signal.

Explained Variability	2 Latent maps k	4 Latent maps k		7 Latent maps k				
	2	3	4	3	4	5	6	7
25%	100%	100%	0%	77.9%	22.1%	0%	0%	0%
30%	100%	100%	0%	0.9%	99.1%	0%	0%	0%
35%	100%	100%	0%	0%	95.4%	4.6%	0%	0%
40%	100%	100%	0%	0%	18.8%	81.0%	0.2%	0%
45%	100%	99.99%	0.01%	0%	0%	65.6%	34.4%	0%
50%	100%	94.8%	5.2%	0%	0%	0.02%	99.80%	0%
75%	100%	0%	100%	0%	0%	0%	82.2%	17.8%
90%	100%	0%	100%	0%	0%	0%	0%	100%

Simulation Study - Equal Weights

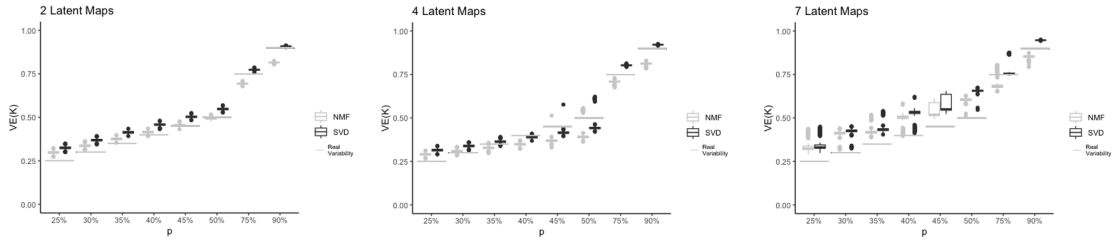


Figure: Box-plots with the explained variability for each scenario. The boxes outlined in dark gray are the results for SVD while the light gray boxes represent the NMF.

Simulation Study - Linear Weights

DEST UFMG

Under this simulation, the latent maps have linearly decaying weights.

Explained Variability	2 Latent maps k	4 Latent maps k		7 Latent maps k				
	2	3	4	3	4	5	6	7
25%	100%	100%	0%	100%	0%	0%	0%	0%
30%	100%	100%	0%	84%	16%	0%	0%	0%
35%	100%	99.4%	0.6%	4.1%	95.9%	0%	0%	0%
40%	100%	71.4%	28.6%	0%	99.5%	0.5%	0%	0%
45%	100%	3.9%	96.1%	0%	61.3%	38.7%	0%	0%
50%	100%	0%	100%	0%	0.4%	82.2%	17.4%	0%
75%	100%	0%	100%	0%	0%	0%	0%	100%
90%	100%	0%	100%	0%	0%	0%	0%	100%

Simulation Study - Linear Weights

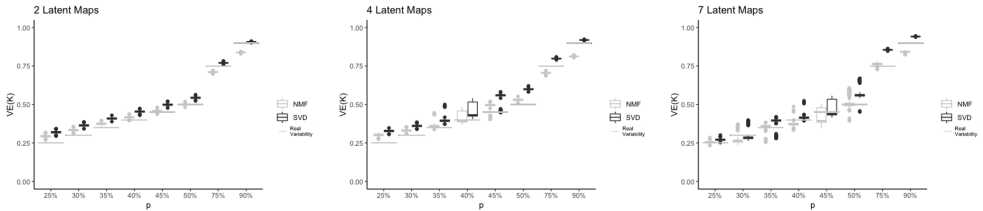


Figure: Linearly decaying weights: Box-plot with the explained variability for each scenario. The boxes outlined in dark gray are the results for SVD while the light gray boxes represent the NMF.

Simulation Study - Original Weights

Using the original weights found in the SVD decomposition of the Brazilian dataset $\mathbf{X}$

Explained Variability	2 Latent Maps k		4 Latent Maps k				7 Latent Maps k						
	1	2	1	2	3	4	1	2	3	4	5	6	7
25%	43%	57%	97.10%	2.90%	0%	0%	99.8%	0.2%	0%	0%	0%	0%	0%
30%	0.01%	99.9%	23.1%	76.9%	100%	0%	80.3%	19.7%	0%	0%	0%	0%	0%
35%	0%	100%	0.01%	95.1%	5%	0%	5.2%	93.7%	1.1%	0%	0%	0%	0%
40%	0%	100%	0%	21.7%	72.9%	5.4%	0%	51.5%	48.5%	0%	0%	0%	0%
45%	0%	100%	0%	0%	16.0%	84.0%	0%	0.7%	95.3%	4.0%	0%	0%	0%
50%	0%	100%	0%	0%	0%	100%	0%	0%	32.9%	67.0%	0.1%	0%	0%
75%	0%	100%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0.3%	99.7%
90%	0%	100%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	100%

Simulation Study - Original Weights

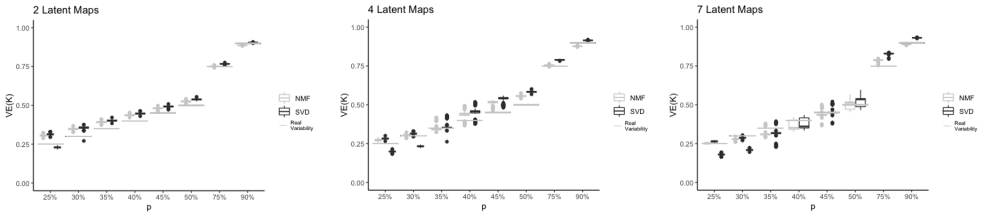


Figure: Original weights: Box-plot with the explained variability for each scenario.

Simulation Study - Conclusions

DEST UFMG

- The total number of latent maps selected by the method satisfactorily recovered the spatial signal when the explained variability is above 40%, regardless of correctly recovering the number of latent maps;
- Although the method is designed to choose the number of maps for the SVD, it is reasonable to use it in the NMF as the explained variability is close to the correct value.
- This study is the first empirical evidence that the number of selected latent maps, k, can also be applied to NMF methods.

Real Data Analysis

DEST UFMG

We consider cancer databases from 6 countries: Brazil, the United States, Spain, England, Australia, and France.

For Brazil, the USA, and Spain, mortality data were provided.

- The USA's data was obtained using the `SEER*Stat` software from 1999 to 2017 divided by US 3107 counties.
- As for Brazilian data, it was collected from DataSUS with the reports from 2000 to 2018 in the level of micro-regions, totaling 557 areas.
- Spain's data was gathered through the Spain statistical office, containing information by province from 2010 to 2018, with a total number of 47 provinces.

Real Data Analysis

DEST UFMG

For England, Australia, and France, only incidence data is provided.

- For England, the data is available at the Office for National Statistics and it is divided by the 195 "National Health Service Regions" with incidences from 2001 to 2017.
- The Australian data was obtained at the Australian Institute of Health and Welfare with information accessible from 2006 to 2010 accounting for the 87 "Statistical Areas Level 4".
- France data was reported by the "Observatoire Géodes" at the department level (94 in total) with data from 2007 to 2016.

The Australian data contains information for 22 most common cancers while the France data has only 9 types. All the other ones have the 30 more common cancers.

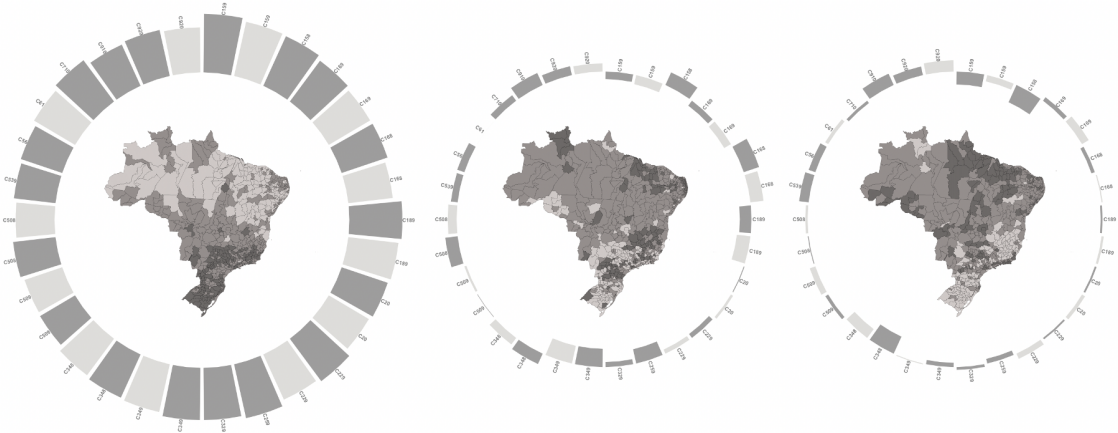
Real Data Analysis - SVD Results

DEST UFMG

Country	# Cancers q	Latent Maps k	$VE(k)$	% Reduction
Brazil	30	6	78.37	80.00
USA	30	4	63.03	86.67
England	30	7	84.21	76.67
Spain	30	3	54.81	90.00
Australia	22	4	72.27	81.82
France	9	2	67.45	77.78

Table: Table containing the summary results for each data set. Column 2 represents the total number of maps, while columns 3 and 4 represent the selected number of maps k in the approximation and the percentage of explained variability, respectively. The last column shows the percentage of reduction on the total number of maps when using the k latent maps.

Real Data Analysis - SVD Results



Real Data Analysis - SVD Results

DEST
UFMG

- The first latent map reflects the bad quality of the mortality information system in Brazil, showing a positive correlation of cancers death and socioeconomic coniditions.
- The recorded cancer rates reflect an entangled mix of actual mortality risks and the quality of the registration system.

Real Data Analysis - SVD Results

DEST UFMG

- The contribution of this second latent factor to the rates is to positively boost the cancers composed of male and female bronchus or lung, stomach, and colon cancers, female breast, pancreas, esophageal cancers, and acute lymphoblastic leukemia.
- Thus it predominantly emphasizes lifestyle factors such as smoking, alcohol consumption, and diet, whereas the second group includes cancers that may be influenced by viral infections and genetic factors.

Real Data Analysis - SVD Results

DEST UFMG

- For latent factor 3, positive values seem to be associated with acute lymphoblastic leukemia, acute myeloid leukemia, cervix, stomach, and ovary cancers, while negative values are associated with esophagus, bronchi, lungs, and pancreas cancers.
- We were unable to find a simple explanation for the spatial pattern of this latent factor.
- Interpreting this factor is challenging because it is heavily influenced by the misregistration of causes of death in Brazil. The inaccuracy in recording specific cancer types can obscure the true associations, making it difficult to draw clear conclusions about the underlying risk factors.

Real Data Analysis - SVD versus Confirmatory Analysis

DEST UFMG

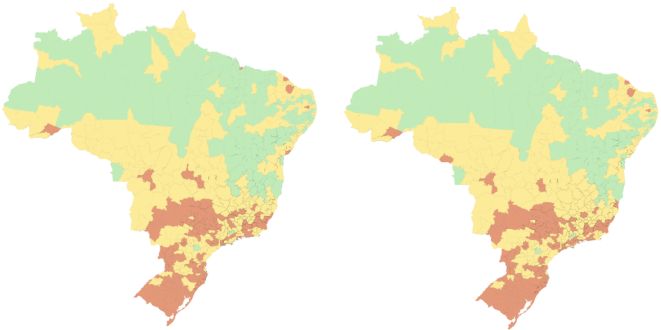
To check our proposal capacity to bring valid insights over the problem. We compare our output with the shared component model (Knorr-Held and Best, 2001; Held et al., 2005) fitting the most common 4 cancers in Brazil.

The shared component model is defined as:

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\mu_{ij}, 1/\tau_j),\ i = 1\ldots, n \text{ and } j = 1, \ldots, d, \\ \mu_{ij} &= \alpha_j + \psi_i \delta_j \end{aligned}$$

where τ_j is the disease specific precision, $\psi = (\psi_1, \ldots, \psi_n)^\top$ is the shared component that captures the spatial structure common to the cancers, and, follow an ICAR prior. Finally, the term δ_j is the weight term for each disease j.

Real Data Analysis - SVD versus Confirmatory Analysis



The map on the left is the latent map selected from the SVD approach. The map on right is the shared component map.

Conclusions

DEST
UFMG

- The main point is that the k latent maps used on the approximation are responsible to explain most of the variability of the data and the rest can be attributed to random noise.
- In practice, the latent maps can be viewed as the base for the data generation and should be the focus of specialized studies that seeks to understand how some diseases are spread on the map and create strategies to tackle this problem and reduce the number of cancer incidence or deaths.

Conclusions

DEST
UFMG

- The simulation study showed that the chosen strategy to find the total number of maps proposed by Gavish and Donoho (2014) is able to accurately recover the corrected number of maps used to generate the data when moderate levels of noise are present.
- When the random noise is responsible for most of the data variability (75%), it is hard to detect the correct number of maps but this is expected since most of the data were random.
- Besides that, although the method was created for SVD, our empirical results show that they can be used for the NMF algorithms too.

Conclusions

DEST UFMG

- An application on multiple real databases showed that most of the data variability could be explained considering a much smaller number of latent maps.
- The reductions in the total number of maps to be considered were between 76% and 90%.
- A comparison between the SVD and the NMF showed that the two methods have similar performance once the total amount of explained variability by the maps is similar. Extension to other type os decomposition methods can be used (Wang and Carvalho, 2023).
- The SVD execution is considerably faster and hence, more suitable for very large maps or large numbers of diseases.
- A comparison between the SVD and the shared component approach showed indistinguishable spatial conclusion.

Conclusion

DEST
UFMG

Main Conclusions

- The SVD approximation is not intended to substitute the individual maps. Rather, it aims at organizing the map collection.
- Due to the highly specialized nature of cancer research, studies generally address only one specific cancer type at a time. However, these studies did not explore multiple correlations among different cancers as we have done here.
- It is worthwhile exploring the possibility of routinely using other cancer incidences to estimate a given cancer incidence when reliable rates are difficult to obtain due to the small number of expected cases.

References

Menezes, T. P. and Prates, M. O. and Assunção, R. and de Castro, M. S. M.. Latent Archetypes of the Spatial Patterns of Cancer. *Statistics in Medicine*, 43, 5115–5137, 2024.

Knorr-Held, L. and Best, N. G. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2001; 164(1), 73 – 85.

Marshall, R. J.. Mapping disease and mortality rates using empirical Bayes estimators. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1991; 40(2), 283 – 294.

Gavish, M. and Donoho, D. L.. The optimal hard threshold for singular values is $4\backslash\sqrt{3}$. *IEEE Transactions on Information Theory*, 2014; 60(8), 5040 – 5053.

Gómez-Rubio V. and Palmí-Perales F.. Multivariate posterior inference for spatial models with the integrated nested Laplace approximation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2019; 68(1), 199–215.

Azevedo, D. R. M. and Prates, M. O. and Bandyopadhyay, D.. MSPOCK: Alleviating spatial confounding in multivariate disease mapping models. *Journal of Agricultural, Biological and Environmental Statistics*, 2021; 26, 464–491.

Held, L. and Natário, I. and Fenton, S. E. and Rue, H. and Becker, N.. Towards joint disease mapping. *Statistical methods in medical research*, 2005; 14(1), 61 – 82.

Wang, L. and Carvalho, L.. Deviance Matrix Factorization. *Electronic Journal of Statistics*, 2023; 17(2), 3762–6810.

Any questions?