Co-clustering of time-dependent data

Alessandro Casa[1], Charles Bouveyron[2], Elena Erosheva[3], Giovanna Menardi[1]

casa@stat.unipd.it, charles.bouveyron@math.cnrs.fr, erosheva@uw.edu, menardi@stat.unipd.it

[1] Università degli Studi di Padova, [2] Université Côte d'Azur, [3] University of Washington

Introduction and Motivation

- Modelling multivariate time-dependent data poses peculiar challenges
- Flexible tools are needed to account for:
  - arbitrarily shaped time evolutions
  - correlation across time instants
  - correlation among different variables at the same time
- How can we extract useful information and unveil parsimonious patterns from such data?

Idea

Make use of an appropriately tuned co-clustering tool in order to summarize multivariate time-dependent data in homogeneous blocks

Co-clustering in a nutshell

- **Co-clustering** refers to those techniques aimed at jointly partitioning subjects and variables. It has been studied both from an heuristic and from a probabilistic perspective

  in the latter case the Latent Block Model (LBM) takes the lion's share
- General model definition

$$p(\mathbf{x};\theta)=\sum_{z\in Z}\sum_{w\in W}\prod_{ik}\pi_k^{z_{ik}}\prod_{jl}\rho_l^{w_{jl}}\prod_{ijkl}p(x_{ij};\theta_{kl})^{z_{ik}w_{jl}}$$

  - n and p number of subjects and variables, K and L number of row and column clusters, $\theta=(\pi_k,\rho_l,\theta_{kl})_{1<k<K,1<l<L}$ full parameters vector
  - $\mathbf{z}=(z_{ik})_{1<i<n,1<k<K}$ and $\mathbf{w}=(w_{jl})_{1<j<p,1<l<L}$ latent random variables describing subject and variable cluster memberships, $z_i\sim\mathcal{M}(1,\pi_1,\dots,\pi_K)$, $w_j\sim\mathcal{M}(1,\rho_1,\dots,\rho_L)$
  - $\mathbf{x}=(x_{ij})_{1<i<n,1<j<p}$ observed values matrix, $(x_{ij}|z_{ik}=1,w_{jl}=1)\sim p(\cdot;\theta_{kl})$

Model specification

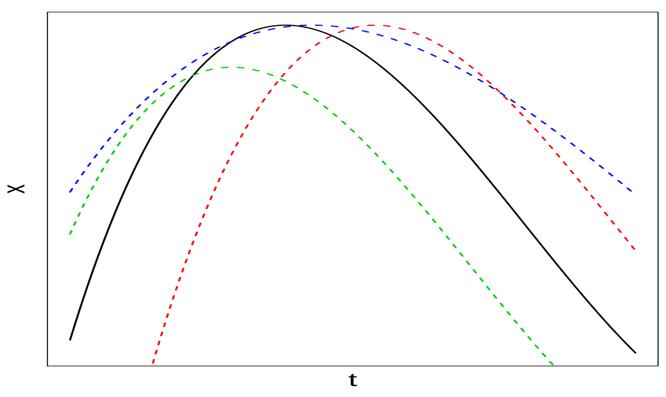
- For time data $\mathbf{x}=(x_{ij}(t_i))_{1<i<n,1<j<p}$, $t_i=1,\dots,n_i$, x_{ij} is a single curve
  $\longrightarrow p(x_{ij};\theta_{kl})$ has to be carefully chosen, respecting the nature of the data
- Here we resort to the **Shape Invariant Model (SIM)** defined as follow

$$(x_{ij}(t)|z_{ik}=1,w_{jl}=1)=\alpha_{ij,1}^{kl}+e^{\alpha_{ij,2}^{kl}}m(t-\alpha_{ij,3}^{kl};\beta_{kl})+\epsilon_{ij}(t) \qquad (1)$$

  - $\alpha_{ij}^{kl}=(\alpha_{ij,1}^{kl},\alpha_{ij,2}^{kl},\alpha_{ij,3}^{kl})\sim\mathcal{N}_3(\mu_{kl}^{\alpha},\Sigma_{kl}^{\alpha})$ vector of cell and block-specific random parameters
  - $m(\cdot)$ common shape function $\longrightarrow$ a spline function is used where β_{kl} are the corresponding parameters
  - $\epsilon_{ij}(t)\sim\mathcal{N}(0,\sigma_{kl}^2)$ measurement error
- Given the specification $\theta_{kl}=(\mu_{kl}^{\alpha},\Sigma_{kl}^{\alpha},\sigma_{kl}^2,\beta_{kl})$

Main idea

curves belonging to the same block arise as random shift and scale transformations of a common mean shape function



Estimation procedure - 1

- We aim at maximizing wrt θ the *complete-data log-likelihood*

$$\ell_c(\theta,\mathbf{z},\mathbf{w})=\sum_{ik}z_{ik}\log\pi_k+\sum_{jl}w_{jl}\log\rho_l+\sum_{ijkl}z_{ik}w_{jl}\log p(x_{ij};\theta_{kl}) \qquad (2)$$

- Double missing data structure prevents the use of the EM algorithm
  $\longrightarrow$ SEM-Gibbs overcomes arising issues by mean of a stochastic E-step
- Model (1) does not lead to a closed-form expression for the marginal likelihood $p(x_{ij};\theta_{kl})$ here defined as

$$p(x_{ij};\theta_{kl})=\int p(x_{ij}|\alpha_{ij}^{kl};\theta_{kl})p(\alpha_{ij}^{kl};\theta_{kl})d\alpha_{ij}^{kl} \qquad (3)$$

- A **Marginalized SEM-Gibbs algorithm** is proposed

Estimation procedure - 2

- Starting from initial values $\theta^{(0)}$ and $\mathbf{w}^{(0)}$ we alternate:
  - Marg-step: obtain marginal cell distribution via Monte Carlo integration

$$p(x_{ij};\theta_{kl}^{(q)})\simeq\frac{1}{M}\sum_{m=1}^{M}p(x_{ij}|\alpha_{ij}^{kl,(m)};\theta_{kl}^{(q)}), \qquad (4)$$

    with $\alpha_{ij}^{kl,(1)},\dots,\alpha_{ij}^{kl,(M)}$ drawn from $\mathcal{N}(\mu_{kl}^{\alpha,(q)},\Sigma_{kl}^{\alpha,(q)})$
  - SE-step: generate row and column partition as $z_i\sim\mathcal{M}(1,\tilde{z}_{i1},\dots,\tilde{z}_{iK})$, $\forall\,1<i<n$ and $w_j\sim\mathcal{M}(1,\tilde{w}_{j1},\dots,\tilde{w}_{jL})$ $\forall\,1<j<p$ as

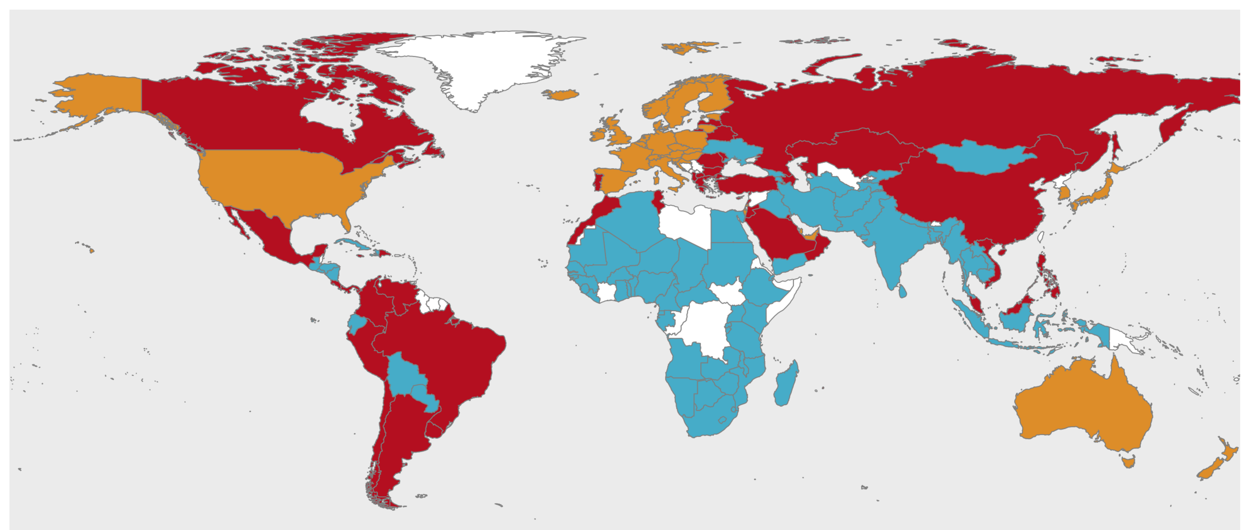
$$\tilde{z}_{ik}=\frac{\pi_k^{(q)}p_k(\mathbf{x}_i|\mathbf{w}^{(q)};\theta^{(q)})}{\sum_{k'}\pi_{k'}^{(q)}p_{k'}(\mathbf{x}_i|\mathbf{w}^{(q)};\theta^{(q)})} \qquad \tilde{w}_{jl}=\frac{\rho_l^{(q)}p_l(\mathbf{x}_j|\mathbf{z}^{(q+1)};\theta^{(q)})}{\sum_{l'}\rho_{l'}^{(q)}p_{l'}(\mathbf{x}_j|\mathbf{z}^{(q+1)};\theta^{(q)})},$$

    $\mathbf{x}_i$, $\mathbf{x}_j$ are the *i*th row and *j*th column, $p_k(\mathbf{x}_i|\mathbf{w}^{(q)};\theta^{(q)})=\prod_{jl}p(x_{ij};\theta_{kl}^{(q)})^{w_{jl}^{(q)}}$ and $p_l(\mathbf{x}_j|\mathbf{z}^{(q+1)};\theta^{(q)})=\prod_{ik}p(x_{ij};\theta_{kl}^{(q)})^{z_{ik}^{(q+1)}}$ with $p(x_{ij};\theta_{kl}^{(q)})$ defined as in (4).
  - M-step: estimate $\theta^{(q+1)}$ conditionally on $\mathbf{z}^{(q+1)}$ and $\mathbf{w}^{(q+1)}$.
    Mixture proportions updated as $\pi_k^{(q+1)}=\frac{1}{n}\sum_i z_{ik}^{(q+1)}$ and $\rho_l^{(q+1)}=\frac{1}{p}\sum_j w_{jl}^{(q+1)}$.
    Block-specific parameters $\theta_{kl}^{(q+1)}$ obtained maximizing an approximate version of the marginal likelihood (3)

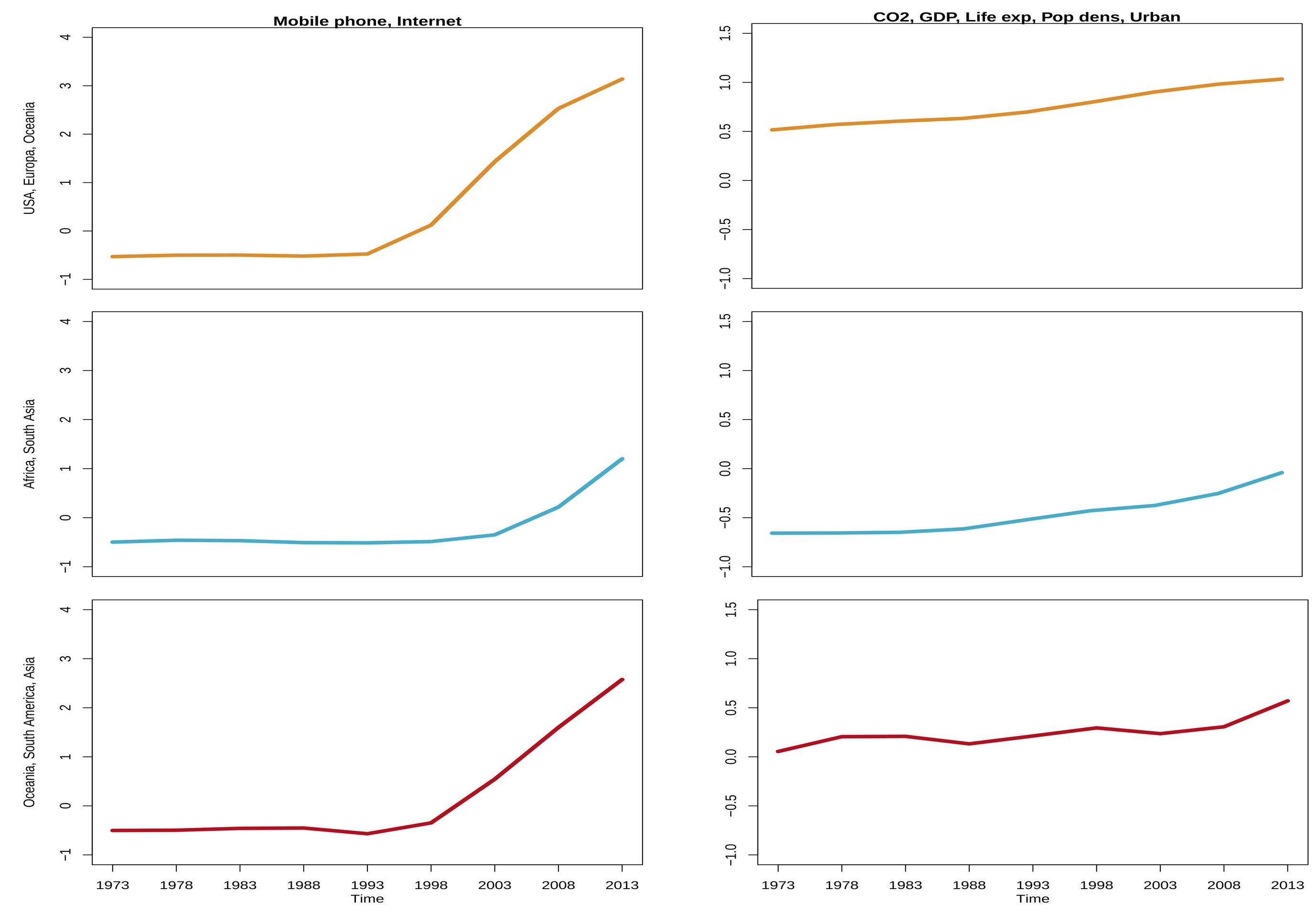
Data and Results

- **Data**: $n=147$ countries, $p=7$ socio-economical, technological and environmental related variables standardized and measured over 9 times
- **Results**:

  Row clusters



  Column clusters
  - Cluster 1: Mobile phone usage, Internet usage
  - Cluster 2: CO2 emissions, GDP per capita, Population density, Life expectancy, Urbanization



Remarks and Discussion

Pros

- Highly flexible, it allows to consider arbitrarily complex evolutions in time and to take into account of different dependence structures
- By excluding some of the random parameters in (1), we encompass different concepts of cluster

Cons

- Highly flexible (?)
- Cumbersome estimation procedure

References

- Bouveyron C., Bozzi L., Jacques J. and Jollois F.X. (2018). *The functional latent block model for the co-clustering of electricity consumption curves.* JRSS-C, 67(4): 897-915.
- Govaert G. and Nadif M. (2013). *Co-clustering: models, algorithms and applications.* John Wiley & Sons.
- Lindstrom, M.J. (1995). *Self-modelling with random shift and scale parameters and a free-knot spline shape function.* Statistics in Medicine, 14(18): 2009-2021.