BAYESIAN MIXTURE MODEL FOR ENVIRONMENTAL APPLICATION

Bayesian Statistics Project 2022/23

P. Bogani, P. Botta, S. Caresana, R. Carrara, G. Corbo, L. Mainini

Tutoring Team: R. Argiento, S. Legramanti, L. Paci

Politecnico di Milano, Mathematical Engineering

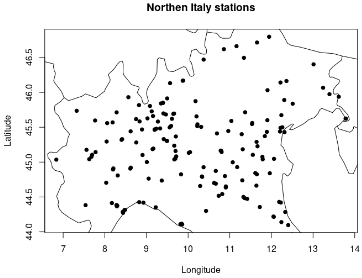
February 14, 2023

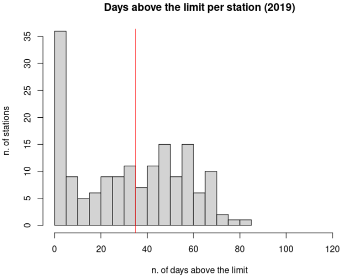
Introduction

- The World Health Organization considers air pollution a major global environmental risk to human health.
- Only in the EU in 2020, a total of 238,000 premature deaths were linked to exposure to particulate matter.

Objective

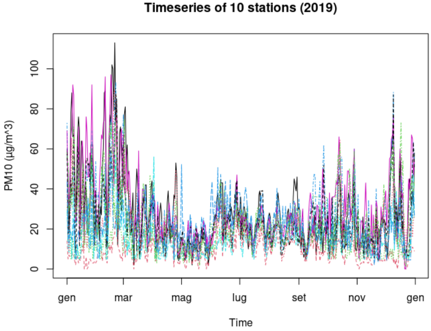
Clustering time series of PM10 concentration





Data

The data under consideration are collected from 162 monitoring stations between 2013 and 2019 in Northern Italy (European Environmental Agency).



We define y_{it} as the concentration of PM10 in the station i at time t, and $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})^\top$, for $i = 1, \ldots, n$.

Starting model

$$\mathbf{y}_i = \mathbf{Z}\boldsymbol{\alpha}_i + \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, \ldots, n$$

$$\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})^\top \overset{\text{iid}}{\sim} \mathrm{N}_T\left(\mathbf{0}, \sigma^2_{\epsilon_i}\mathbf{I}\right)$$

$$\theta_{it} = \rho\theta_{i,t-1} + \nu_{it} \quad \text{with } \nu_{it} \sim \mathrm{N}\left(0, \sigma^2_\theta\right)$$

$$\boldsymbol{\theta}_i \sim \mathrm{N}_T(\mathbf{0}, \mathbf{R}(\rho)) \quad \text{with } R_{ls} = \sigma^2_\theta \rho^{|l-s|}$$

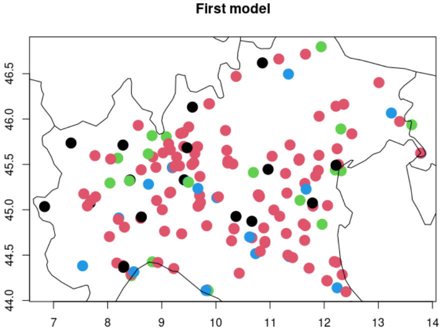
We adopt a generalization of the Dirichlet process (Poisson-Dirichlet) as prior associated to the distributions of the coefficients used for clustering $\boldsymbol{\gamma}_i = (\boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$

$$\boldsymbol{\gamma}_i \mid G \overset{\text{iid}}{\sim} G, \text{ for } i = 1, \ldots, n \text{ with } G \sim \mathcal{PD}\left(a, b, G_0\right)$$

The almost-certain discretization of the Dirichlet Process induces a clustering among the data, grouping the observations with the same latent variables $\boldsymbol{\gamma}_i$ sampled from the DP, the so-called **ties**.

Results

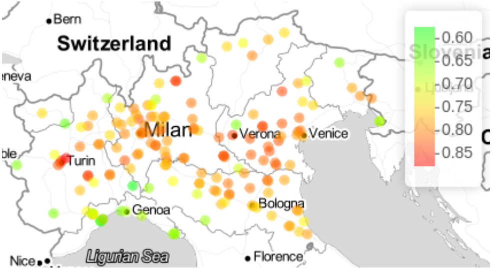
The clusters obtained with this model are:

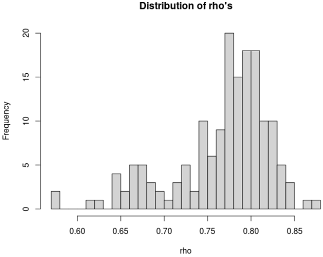


We could have expected a poor outcome because there isn't much difference in level or trend between the stations, as can be seen in the matplot of the PM10 concentration in Northern Italy.

A new approach

- Our idea was to develop a clustering technique based on the **persistence**.
- We decided to cluster our data on a higher hierarchical level, implementing a Dirichlet process as prior for the distribution of the parameters of the auto-regressive model $\gamma_i = (\rho_i, \sigma_i^2)$.





$\boldsymbol{\rho}$'s obtained by fitting an AR(1) on PM10 concentrations

Proposed model

$$\mathbf{y}_i = \mathbf{Z}\boldsymbol{\alpha}_i + \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, \ldots, n$$
$$\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})^\top \overset{\text{iid}}{\sim} \mathrm{N}_T\left(\mathbf{0}, \sigma^2_{\epsilon_i}\mathbf{I}\right)$$
$$\theta_{it} = \rho_i \theta_{i,t-1} + \nu_{it} \quad \text{with } \nu_{it} \sim \mathrm{N}\left(0, \sigma_i^2\right)$$
$$\boldsymbol{\theta}_i \sim \mathrm{N}_T(\mathbf{0}, \mathbf{R}(\rho_i)) \quad \text{with } R_{ls} = \sigma_i^2 \rho_i^{|l-s|}$$

We adopt a Dirichlet process as prior associated to the distribution of the coefficients used for clustering $\gamma_i = (\rho_i, \sigma_i^2)$

$$\gamma_i \mid P \overset{\text{iid}}{\sim} P, \text{ for } i = 1, \ldots, n \text{ with } P \sim \mathrm{Dir}\left(a^P, P_0\right)$$
$$p_0\left(\gamma_i\right) = p_0(\sigma_i^2) \times p_0(\rho_i) = \mathrm{IGa}(a, b) \times \mathrm{Beta}(c, d),$$

and the following prior distributions:

$$\boldsymbol{\alpha}_i \overset{\text{iid}}{\sim} \mathrm{N}_p\left(\mathbf{0}, \boldsymbol{\Sigma}_\alpha\right) \quad \text{with } \boldsymbol{\Sigma}_\alpha = \mathrm{diag}(\sigma_{\alpha_1}, \ldots, \sigma_{\alpha_p})$$
$$\sigma^2_{\epsilon_i} \sim \mathrm{IGa}\left(c_0^\epsilon, c_1^\epsilon\right), \quad \sigma^2_{\alpha_k} \sim \mathrm{IGa}\left(c_0^\alpha, c_1^\alpha\right)$$

Full conditionals

$$f\left(\boldsymbol{\alpha}_i \mid \mathbf{y}, \sigma^2_{\epsilon_i}, \boldsymbol{\Sigma}_\alpha\right) = \mathrm{N}_p\left(\boldsymbol{\alpha}_i \mid \boldsymbol{\mu}_\alpha, \mathbf{V}_\alpha\right) \tag{1}$$

for $i = 1, \ldots, n$, where $\boldsymbol{\mu}_\alpha = \mathbf{V}_\alpha \mathbf{Z}' \mathbf{W}_i^{-1} \mathbf{y}_i$ and $\mathbf{V}_\alpha = \left(\mathbf{Z}'\mathbf{W}_i^{-1}\mathbf{Z} + \boldsymbol{\Sigma}_\alpha^{-1}\right)^{-1}$ with $\mathbf{W}_i = \sigma^2_{\epsilon_i}\mathbf{I} + \mathbf{R}(\rho_i)$ of dimensions $T \times T$.

$$f\left(\sigma^2_{\epsilon_i} \mid \mathbf{y},\ -\ \right) = \mathrm{IGa}\left(\sigma^2_{\epsilon_i} \mid c_0^\epsilon + \frac{T}{2}, c_1^\epsilon + \frac{1}{2}\mathbf{M}_i'\mathbf{M}_i\right), \tag{2}$$

where $\mathbf{M}_i = \mathbf{y}_i - \mathbf{Z}\boldsymbol{\alpha}_i - \boldsymbol{\theta}_i$, for $i = 1, \ldots, n$.

$$f\left(\sigma^2_{\alpha_j} \mid \mathbf{y},\ -\ \right) = \mathrm{IGa}\left(\sigma^2_{\alpha_j} \mid c_0^\alpha + \frac{n}{2}, c_1^\alpha + \frac{1}{2}\sum_{i=1}^{n}\alpha_{ij}^2\right), \tag{3}$$

for $j = 1, 2, \ldots, p$.

$$f\left(\boldsymbol{\theta}_i \mid \mathbf{y}, \sigma^2_{\epsilon_i}, \mathbf{R}(\rho_i)\right) = \mathrm{N}_T\left(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}_\theta, \mathbf{S}_\theta\right) \tag{4}$$

for $i = 1, \ldots, n$, where $\mathbf{S}_\theta = \left(\left(\sigma^2_{\epsilon_i}\mathbf{I}\right)^{-1} + \mathbf{R}(\rho_i)^{-1}\right)^{-1}$ and
$\boldsymbol{\mu}_\theta = \mathbf{S}_\theta\left(\sigma^2_{\epsilon_i}\mathbf{I}\right)^{-1}(\mathbf{y}_i - \mathbf{Z}\boldsymbol{\alpha}_i)$.

Adapting Neal's algorithm 8 ...

We can now update c_i by sampling from its conditional distribution given θ_i and the parameters of all existing and empty clusters. Specifically,
For $i = 1, \dots, n$, let m_i the number of distinct c_i for $j \neq i$ and $h = m_i + n_{\text{aux}}$. If $c_i = c_j$ for some $j \neq i$, draw values independently from P_0 for those γ_c^* for which $m_i < c \leq h$. If $c_i \neq c_j$ for all $j \neq i$, let c_i have the label $m_i + 1$, and draw values independently from P_0 for those γ_c^* for which $m_i + 1 < c \leq h$. Draw a new value for c_i from $\{1, \dots, h\}$ using the following probabilities:

$$\mathbb{P}\left[c_i = c \mid \boldsymbol{c}_{-i}, \boldsymbol{\theta}_i, \gamma_1^*, \dots, \gamma_h^*\right] = \begin{cases} b \cdot \frac{n_{c,i}^*}{n-1+a^p} \cdot f\left(\boldsymbol{\theta}_i \mid \gamma_{c_i}^*, -\right), & \text{for } 1 \leq c \leq m_i \\ b \cdot \frac{a^p/n_{\text{aux}}}{n-1+a^p} \cdot f\left(\boldsymbol{\theta}_i \mid \gamma_{c_i}^*, -\right), & \text{for } m_i < c \leq h \end{cases}$$

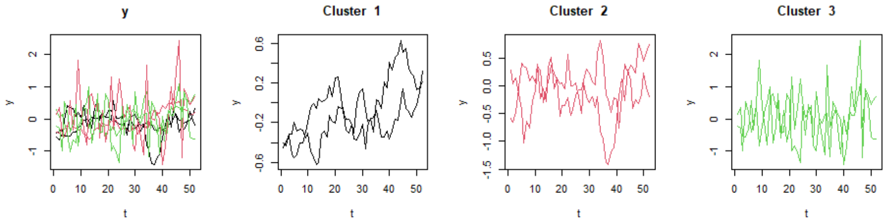
where $n_{c,i}^*$ is the number of c_j for $j \neq i$ that are equal to c, and b is the appropriate normalizing constant. The values of $\gamma_{c_i}^*$ where $m_i < c \leq h$ are drawn independently from P_0.

Algorithm scheme

- Construction of the design matrices
- Initialization of the parameters
- Beginning of **Gibbs sampling**
  1. Sample α from its full conditional
  2. Sample θ from its full conditional
  3. Sample $\gamma = \left(\rho, \sigma^2\right)$'s
  4. Sample σ_ϵ^2 from its full conditional
  5. Sample σ_α^2 from its full conditional
- End of Gibbs sampling
- Determination of the "best" cluster configuration

Synthetic data

To verify the correctness of the algorithm we set some random values for the unknown parameters and we generated θ_{it} and $\boldsymbol{y}_i$ accordingly.

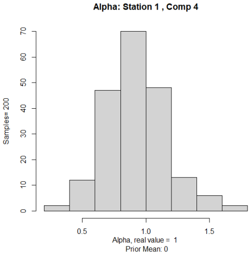


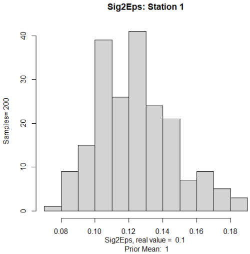
The above data are obtained simulating 6 stations equally divided in three clusters. $\alpha_i = 0$ for each i, and $\sigma_{\epsilon_i} = 0.0001$ for each i (i.e. low noise)

1. The first cluster has $\rho = 0.9$ and $\sigma = 0.1$
2. The second cluster has $\rho = 0.9$ and $\sigma = 0.5$
3. The third cluster has $\rho = 0.1$ and $\sigma = 0.5$

Synthetic data

To verify the correctness of the posteriors we used a larger synthetic dataset with 45 stations and three clusters (with the same ρ_i and σ_i as before). This time, $\sigma_{\epsilon_i} = 0.1$ and $(\boldsymbol{\alpha}_i)_4 = 1$ for each i.

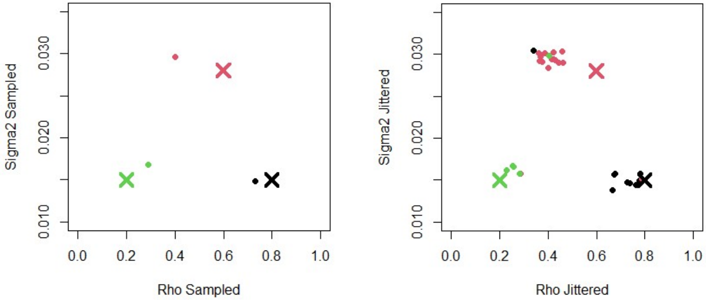




The real value of $(\boldsymbol{\alpha}_1)_4$ is 1. The distribution of the samples is slightly shifted to the left due to the prior (mean) equals to 0. Similarly, the real value of σ_{ϵ_1} is 0.1 but the distribution of the samples is slightly shifted to the right due to the prior (mean) equals to 1.

Synthetic data

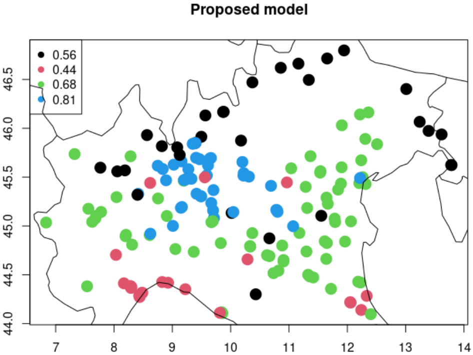
Finally, we applied the algorithm on a synthetic dataset with 3 clusters with dimensions 9, 15, 5 and respectively ρ: 0.8, 0.6, 0.2 and σ^2: 0.015, 0.028, 0.015.



In the figure on the right, each observation is represented with a jittered value of ρ and σ^2, just for visual purpose.

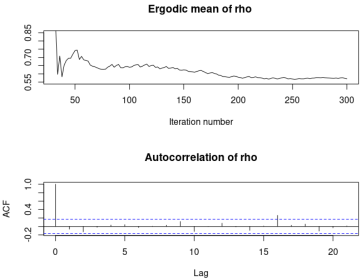
Results

We proceeded to apply the new proposed model to the PM10 concentration data. We used the `salso` algorithm developed by Dahl which implements a greedy, stochastic search given the number of desired clusters, in our case 4 and chosen the partition minimizing the posterior of the VI loss function.

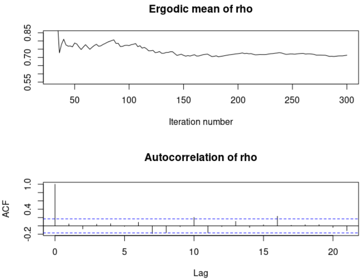


Results

We've then decided to compare the convergence of the ergodic mean of ρ of two stations belonging to different clusters.



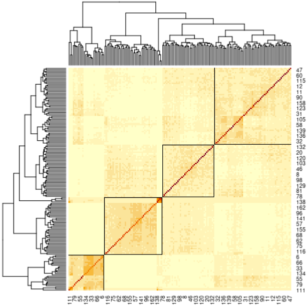
Station of the "red" cluster

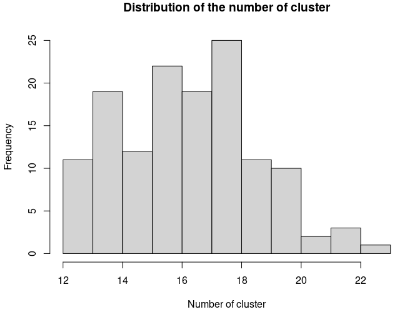


Station of the "blue" cluster

Results

MCMC clustering information can be summarized by counting the number of times (iterations) that two parameters belong to the same cluster, constructing a similarity matrix.





Conclusions and developments

Conclusions

- Northern Italy does not have pollution levels that far apart, and the first model does not provide particular insights.
- Cluster's results show that the persistence is often higher in the proximity of urban areas and dry zones while it is lower in areas with more vegetation or breezy regions (for instance closer to sea).

Further developments

- Try other non-parametric prior for the distribution of γ or introduce the spatial component to obtain better-distributed clusters in the space.
- Deeply tune the hyper-parameters or introduce an acceleration step (in the case of different priors) to reduce the number of clusters obtained by the algorithm.

References

- Nieto-Barajas, L. E., Contreras-Cristán, A. (2014). A Bayesian nonparametric approach for time series clustering. Bayesian Analysis, 9(1), 147-170.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. Journal of computational and graphical statistics, 9(2), 249-265.
- Wade, S., Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion).
- Lau, J. W., Green, P. J. (2007). Bayesian model-based clustering procedures. Journal of Computational and Graphical Statistics, 16(3), 526-558.
- Dahl, D. B., Johnson, D. J., Müller, P. (2022). Search algorithms and loss functions for Bayesian clustering. Journal of Computational and Graphical Statistics, 31(4), 1189-1201.