Mathematical Engineering A.Y. 2022-23
Bayesian Statistics(052499) - Prof. Alessandra Guglielmi
Tutoring team: Raffaele Argiento, Sirio Legramanti, Lucia Paci

# BAYESIAN MIXTURE MODEL FOR ENVIRONMENTAL APPLICATION

Trascription of the model
29th January 2023

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

$$\mathbf{y}_i = \mathbf{Z}\boldsymbol{\alpha}_i + \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i \quad \text{for } i = 1, 2, \ldots n,$$
$$\boldsymbol{\epsilon}_i' = (\epsilon_{i1}, \ldots, \epsilon_{iT}) \sim \mathrm{N}_T\left(\mathbf{0}, \sigma_{\epsilon_i}^2 \mathbf{I}\right),$$
$$\theta_{it} = \rho_i \theta_{i,t-1} + \nu_{it} \quad \text{with } \nu_{i1} \sim \mathrm{N}\left(0, \sigma_i^2\right),$$
$$\gamma_i = (\sigma_i^2, \rho_i),$$
$$\boldsymbol{\theta}_i \sim \mathrm{N}_T(\mathbf{0}, \mathbf{R}(\gamma_i)) \quad \text{with } R_{ls}(\gamma_i) = \sigma_i^2 \rho_i^{|l-s|} \text{ and } l, s = 1, \ldots, T.$$

## 1 Priors distributions

$$\boldsymbol{\alpha}_i \overset{\text{iid}}{\sim} \mathrm{N}_p\left(\mathbf{0}, \boldsymbol{\Sigma}_\alpha\right) \quad \text{with } \boldsymbol{\Sigma}_\alpha = \mathrm{diag}(\sigma_{\alpha_1}, \ldots, \sigma_{\alpha_p}),$$
$$\sigma_{\epsilon_i}^2 \sim \mathrm{IGa}\left(c_0^\epsilon, c_1^\epsilon\right), \quad \sigma_{\alpha_k}^2 \sim \mathrm{IGa}\left(c_0^\alpha, c_1^\alpha\right),$$
$$\gamma_i \mid P \overset{\text{iid}}{\sim} P, \text{ for } i = 1, \ldots, n \text{ with } P \sim \mathrm{Dir}\left(a^P, P_0\right),$$
$$p_0\left(\gamma_i\right) = p_0(\sigma_i^2) \times p_0(\rho_i)$$
$$p_0(\sigma_i^2) = \mathrm{IGa}(a, b) \qquad p_0(\rho_i) = \mathrm{Beta}_{[-1,1]}(c, d),$$

for $i = 1, \ldots, n$, where $\mathrm{Beta}_{[-1,1]}(c, d)$ is a beta distribution with domain in $[-1, 1]$ and hyper-parameters $c$ and $d$.

## 2 Posterior distributions

In the following computations, we'll make use of some properties related to multivariate Gaussian distribution that we report here.

**Theorem 2.1.** *Given a marginal Gaussian distribution for* $\mathbf{x}$ *and a conditional Gaussian distributton for* $\mathbf{y}$ *given* $\mathbf{x}$ *in the form*

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right),$$
$$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right),$$

*the marginal distribution of* $\mathbf{y}$ *and the conditional distribution of* $\mathbf{x}$ *given* $\mathbf{y}$ *are given by*

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\right), \tag{1}$$
$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\Sigma}\left\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}\right), \tag{2}$$

*where*

$$\boldsymbol{\Sigma} = (\Lambda + A^T L A)^{-1}.$$

A proof of the previous properties can be found in Appendix B(Pag 91-92 of [1]).

If we let $\boldsymbol{\alpha}' = (\boldsymbol{\alpha}'_1, \ldots, \boldsymbol{\alpha}'_n)$, $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_n)$, and $\boldsymbol{\sigma}'_\epsilon = \left(\sigma^2_{\epsilon_1}, \ldots, \sigma^2_{\epsilon_n}\right)$ then the likelihood function is given by

$$f\left(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\sigma}_\epsilon\right) = \prod_{i=1}^{n} \mathrm{N}_T\left(\mathbf{y}_i \mid \mathbf{Z}\boldsymbol{\alpha}_i + \boldsymbol{\theta}_i, \sigma^2_{\epsilon_i}\mathbf{I}\right).$$

With this notation, we denote that $\mathbf{y}_i \mid \boldsymbol{\alpha}_i, \boldsymbol{\theta}_i, \boldsymbol{\sigma}_{\epsilon_i}$ are distributed as an $T$-variate multivariate normal distribution with mean vector $\mathbf{Z}\boldsymbol{\alpha}_i + \boldsymbol{\theta}_i$ and variance-covariance matrix $\sigma^2_{\epsilon_i}\mathbf{I}$. We will continue using this notation from now on.

Using Property 1 of Theorem 2.1:

$$f\left(\boldsymbol{y}_i \mid \boldsymbol{\alpha_i}, \sigma^2_{\epsilon_i}, \boldsymbol{\Sigma}_\alpha, \gamma_i\right) = \mathrm{N}_p\left(\boldsymbol{y}_i \mid Z\boldsymbol{\alpha}_i, \mathbf{W}_i\right), \tag{3}$$

with matrices $\mathbf{W}_i = \sigma^2_{\epsilon_i}\mathbf{I} + \mathbf{R}(\gamma_i)$.
The conditional posterior distribution of $\boldsymbol{\alpha}_i$ becomes

$$f\left(\boldsymbol{\alpha}_i \mid \mathbf{y}, \sigma^2_{\epsilon_i}, \boldsymbol{\Sigma}_\alpha\right) = \mathrm{N}_p\left(\boldsymbol{\alpha}_i \mid \boldsymbol{\mu}_\alpha, \mathbf{V}_\alpha\right), \tag{4}$$

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 matrices

$$\mathbf{W}_i = \sigma^2_{\epsilon_i}\mathbf{I} + \mathbf{R}(\gamma_i),$$

of dimensions $T \times T$.
The conditional posterior distribution for the variances $\sigma^2_{\epsilon_i}, i = 1, \ldots n$, and $\sigma^2_{\alpha_k}, k = 1, \ldots, p$ given the data and the rest of the parameters are all conditionally conjugate. The conditional posterior distribution for $\sigma^2_{\epsilon_i}$ has the form

$$f\left(\sigma^2_{\epsilon_i} \mid \mathbf{y},\ \text{rest}\ \right) = \text{IGa}\left(\sigma^2_{\epsilon_i} \mid c_0^\epsilon + \frac{T}{2}, c_1^\epsilon + \frac{1}{2}\mathbf{M}_i'\mathbf{M}_i\right), \tag{5}$$

where $\mathbf{M}_i = \mathbf{y}_i - \mathbf{Z}\boldsymbol{\alpha}_i - \boldsymbol{\theta}_i$, for $i = 1, \ldots, n$.
The conditional posterior distribution for $\sigma^2_{\alpha_j}$ has the form

$$f\left(\sigma^2_{\alpha_j} \mid \mathbf{y},\ \text{rest}\ \right) = \text{IGa}\left(\sigma^2_{\alpha_j} \mid c_0^\alpha + \frac{n}{2}, c_1^\alpha + \frac{1}{2}\sum_{i=1}^{n} \alpha^2_{ij}\right), \tag{6}$$

for $j = 1, 2, \ldots, p$.
The conditional posterior distribution for $\boldsymbol{\theta_i}$ has the form

$$f\left(\boldsymbol{\theta}_i \mid \mathbf{y}, \sigma^2_{\epsilon_i}, \mathbf{R}(\gamma_i)\right) = \text{N}_T\left(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}_\theta, \mathbf{S}_\theta\right), \tag{7}$$

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

Fixing $i$, we recall that $\boldsymbol{\gamma}_{-i}$ denotes the set of all $\gamma$ 's excluding the $i^{th}$ element. $\gamma^*_{j,i}$'s denote the unique values in $\boldsymbol{\gamma}_{-i}$, each occurring with frequency $n^*_{j,i}$, with $j = 1, \ldots, m_i$, where $m_i$ represents the number of unique values in $\boldsymbol{\gamma}_{-i}$.
The posterior distribution for $\gamma$ is characterized by the formulation of algorithm 8 by Neal [2], using $n_{\text{aux}}$ auxiliary variables to prevent the fact that the prior are not conjugate. *da scrivere meglio sta parte*
We can rewrite the model as

$$\begin{aligned} y_i \mid c_i, \boldsymbol{\gamma} &\sim f\left(\gamma_{c_i}\right) \\ c_i \mid \boldsymbol{p} &\sim \text{Discrete}\left(p_1, \ldots, p_K\right) \\ \gamma_c &\sim P_0 \\ \boldsymbol{p} &\sim \text{Dirichlet}(a^p/K, \ldots, a^p/K). \end{aligned}$$

Here, $c_i$ indicates which "latent class" is associated with observation $y_i$. The mixing proportions for the classes, $\boldsymbol{p} = (p_1, \ldots, p_K)$, are given a symmetric Dirichlet prior, with concentration parameter written as $a^p/K$, so that it approaches zero as $K$ goes to infinity. $f\left(\gamma_{c_i}\right)$ represents the density function:

$$f\left(\mathbf{y_i} \mid \gamma_{c_i}, \boldsymbol{\alpha}_i, \sigma_{\epsilon_i}\right) = \text{N}_T\left(\mathbf{y}_i \mid \mathbf{Z}\boldsymbol{\alpha}_i, \sigma^2_{\epsilon_i}\mathbf{I} + \mathbf{R}(\gamma_{c_i})\right),$$

due to the theorem 2.1.
We can now update $c_i$ by sampling from its conditional distribution given $y_i$ and the parameters of all existing and empty clusters. Specifically,
For $i = 1, \ldots, n$, 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 $\gamma_c^*$ 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 $\gamma_c^*$ for which $m_i + 1 < c \leq h$.
Draw a new value for $c_i$ from $\{1, \ldots, h\}$ using the following probabilities:

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

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 $P_0$.

## References


[1] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[2] Radford M Neal. 'Markov chain sampling methods for Dirichlet process mixture models'. In: *Journal of computational and graphical statistics* 9.2 (2000), pp. 249–265.