POLITECNICO
MILANO 1863

A Novel Spatio-Temporal Model for Bayesian Source Apportionment

Bayesian Statistics Project

G.Aurina, P.Bettanin, A.Calanchi, R.Scorza, C.Strino, F.Volpicelli

Tutor: Michela Frigeri February 14, 2024

Objective: Develop a **Bayesian spatio-temporal model** for **source apportionment** of particulate matter pollution.

- Addresses key challenges: **temporal dependence** and **spatial variability**.
- Provides probabilistic estimates of pollution source contributions.

Method: Using **Monte Carlo Markov Chain (MCMC)** sampling.

- Implementing **Turing.jl** for efficient posterior inference.
- Build a **customized** MCMC.

Source Contribution Decomposition

Observed concentration of pollutant C at site s_i on day t :

$$y^c(s_i,t) \mid \mu^c(s_i,t), \sigma_c^2 \overset{ind}{\sim} \mathcal{N}\left(\mu^c(s_i,t), \sigma_c^2\right) \quad i = 1,\dots,N; \quad c = 1,\dots,C \quad [1]$$

where:

$$\mu^c(s_i,t) = \sum_{k=1}^{K} h_k^c g_k(s_i,t) \quad [4]$$

- $h_k|\alpha_0 \overset{iid}{\sim} \text{Dirichlet}(\alpha_0,\dots,\alpha_0) \quad k = 1,\dots,K$
- $g_k(s_i,t) = e^{\gamma_{ki}} f(t - \tau_i)$
- $\sigma_c^2 \mid a, b \overset{iid}{\sim} \text{InvGamma}(a,b) \quad c = 1,\dots,C$

We define the following parameters and variables:

- $i = 1,\dots,N$: monitoring sites.
- $k = 1,\dots,K$: pollution sources.
- $c = 1,\dots,C$: pollutants.
- $t = 1,\dots,T$: time steps.

Spatial Dependence [2]

For $k = 1, \ldots, K$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, we define:

$$g_k(s_i, t) = e^{\gamma_{ki}} f_k(t - \tau_i)$$
$$\gamma_k \mid \beta_k, \phi_k \overset{ind}{\sim} \mathcal{N}_N(X\beta_k, \Sigma_{\gamma_k})$$

where:

- $\beta_k \overset{iid}{\sim} \mathcal{N}_p(\mathbf{0}, I_p)$
- $\Sigma_{\gamma_k}[i,j] = \exp\left\{-\frac{\phi_k^2}{2}||s_i - s_j||^2\right\}$
  - $\phi_k \overset{iid}{\sim} \text{Gamma}(a_\phi, b_\phi)$

γ_k represents the spatial variation of source intensity across monitoring sites, modeled via a Gaussian process with squared exponential covariance.

Temporal Dependence [5][3]

For $k = 1, \dots, K$, $i = 1, \dots, N$, $t = 1, \dots, T$, we define:

$$g_k(s_i, t) = e^{\gamma_{ki}} \cdot f_k(t - \tau_i)$$

- $f_k \mid \rho_k \overset{ind}{\sim} \mathcal{N}_T\left(\mathbf{0}, \Sigma_{f_k}\right)$
- $\Sigma_{f_k}[m, n] = \exp\left\{-\frac{\rho_k^2}{2}(t_m - t_n)^2\right\}$
  - $\rho_k \overset{iid}{\sim} \text{Gamma}(a_\rho, b_\rho)$

The term τ_i models site-specific temporal shifts, capturing delayed effects of pollution sources:

$$\tau \sim \mathcal{N}_N\left(0, \sigma_\tau^2\left(I_n - \frac{1}{n+1}\mathbf{1}\mathbf{1}'\right)\right)$$

We generate a synthetic dataset that mimics real-world PM10 pollution trends, simulating spatial and temporal variations across monitoring sites.
The goal is to verify whether MCMC inference (with Turing support) can successfully recover the underlying generative structure, validating our modeling approach.

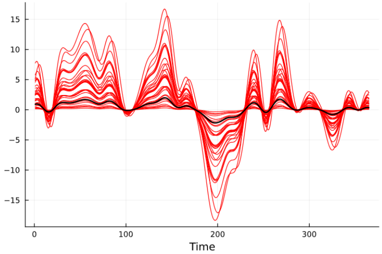

Figure: Simulated data $k = 1$

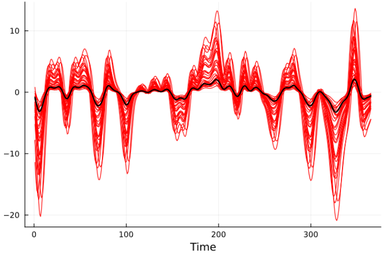

Figure: Simulated data $k = 2$

Construction of the model in Turing.jl

Turing.jl is a general-purpose probabilistic programming language (PPL) written entirely in Julia. It offers a streamlined approach to Bayesian inference, requiring users to define only the probabilistic model, including priors and likelihood, within a function.
We experimented with different sampling methods already implemented in Turing.jl but obtained poorer results compared to our customized MCMC approach.

Possible reason: Turing.jl is a relatively new framework in Julia, and sampling from the proposed deep hierarchical model may present challenges due to its complexity.

Customized MCMC: Full Conditional of g_k

Given the joint distribution:

$$\mathcal{L}\left(f_k, g_k(s_i)\right) = \begin{pmatrix} f_k(1) \\ \vdots \\ f_k(T) \\ a_i f_k(1-\tau_i) \\ \vdots \\ a_i f_k(T-\tau_i) \end{pmatrix} \sim \mathcal{N}_{2T}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \Sigma_{f_k} & a_i \Sigma_{ki} \\ a_i \Sigma_{ki}^T & a_i^2 \Sigma_{f_k} \end{bmatrix}\right).$$

where $a_i = e^{\gamma_{ki}}$.

This representation shows that $g_k(s_i)$ follows a Gaussian process evaluated at new points, effectively performing temporal kriging over τ_i.

Customized MCMC: Full Conditional of g_k

This allows us to derive the conditional distribution:

$$g_k(s_i) \mid f_k, \gamma_{ki}, \tau_i, \rho_k \sim \mathcal{N}_T(\mu_{g_{ki}}, \Sigma_{g_{ki}})$$

where the mean and covariance are given by:

- $\mu_{g_{ki}} = e^{\gamma_{ki}} \Sigma_{ki} \Sigma_{f_k}^{-1} f_k$
- $\Sigma_{g_{ki}} = e^{2\gamma_{ki}} (\Sigma_{f_k} - \Sigma_{k_i} \Sigma_{f_k}^{-1} \Sigma_{k_i}^T + \epsilon I)$

The covariance matrices are defined as:

$$\Sigma_{k_i}[m, n] = \exp\left\{ -\frac{\rho_k^2}{2} (t_m - \tau_i - t_n)^2 \right\}$$

$$\Sigma_{f_k}[m, n] = \exp\left\{ -\frac{\rho_k^2}{2} (t_m - t_n)^2 \right\}$$

- **Full Conditional of f_k:**

$$f_k | G_k, \gamma_k, \tau, \rho_k \sim \mathcal{N}(\mu_{f\text{-post}}, \Sigma_{f\text{-post}})$$

- **Full Conditional of γ_k:**

$$\gamma_{ki} \mid \gamma_k^{(-i)}, \tau, f_k, g_k(s_i) \propto \mathcal{L}(g_k(s_i) \mid \gamma_{k_i}, \tau_i, f_k)\mathcal{L}(\gamma_{ki} \mid \gamma_k^{(-i)}).$$

$$\gamma_{ki} \mid \gamma_k^{(-i)} \sim \mathcal{N}\Big(X_i^\top \beta_k + \Sigma_{\gamma_k}[i, -i] \left(\Sigma_{\gamma_k}[-i, -i]\right)^{-1} (\gamma_k^{(-i)} - X_{[-i,:]}\beta_k), \\ \Sigma_{\gamma_k}[i, i] - \Sigma_{\gamma_k}[i, -i] \left(\Sigma_{\gamma_k}[-i, -i]\right)^{-1} \Sigma_{\gamma_k}[-i, i]\Big).$$

- **Full Conditional of** β_k:

$$\beta_k \mid \gamma_k, \phi_k \sim \mathcal{N}_P \left((X^T \Sigma_{\gamma_k}^{-1} X + I_p)^{-T} X^T \Sigma_{\gamma_k}^{-T} \gamma_k, (X^T \Sigma_{\gamma_k}^{-1} X + I_p)^{-1} \right)$$

- **Full Conditional of** τ:

$$\tau | \gamma_k, f_k, g_k(s_i) \propto \prod_{i=1}^{N} \mathcal{L}(g_k(s_i) | \gamma_{ki}, \tau_i, f_k) \pi(\tau)$$

- **Full Conditional of ϕ_k:**

$$\phi_k|\gamma_k, \beta_k \propto \mathcal{L}(\gamma_k|\beta_k, e^{\theta_k})\pi_{\phi_k}(e^{\theta_k})e^{\theta_k}$$

we sample from the transformation $\theta_k = \log(\phi_k)$, *considering a proposal* $\mathcal{N}(\theta^{\text{old}}, \varepsilon)$

- **Full Conditional of ρ_k:**

$$\rho_k|\gamma_k, \tau, f_k, g_k(s_i) \propto \prod_{i=1}^{N} \mathcal{L}(g_k(s_i)|\gamma_{ki}, \tau_i, f_k)\mathcal{L}(f_k|\rho_k)\pi_{\rho_k}(e^{\log\rho_k})e^{\log\rho_k}$$

MCMC Algorithm

Initialize $\gamma_k^{(0)}, f_k^{(0)}, \beta_k^{(0)}, \phi_k^{(0)}, \tau^{(0)}, \rho_k^{(0)}$

for *iter* $\in 1$: *num.iterations* **do**

  Update $f_k^{(\text{iter})} | \gamma_k^{(\text{iter}-1)}, \tau^{(\text{iter}-1)}, \rho_k^{(\text{iter}-1)}, G_k^{(\text{iter-1})}$ **(closed-form)**

  Update $\beta_k^{(\text{iter})} | \gamma_k^{(\text{iter}-1)}, \phi_k^{(\text{iter}-1)}$ **(closed-form)**

  Update $\phi_k^{(\text{iter})} | \gamma_k^{(\text{iter}-1)}, \beta_k^{(\text{iter})}$

  for $i \in 1 : N$ **do**

    Update $\gamma_{ki}^{(\text{iter})} | \gamma_{k(-i)}^{(\text{iter})}, \tau_i^{(\text{iter}-1)}, f_k^{(\text{iter})}, \beta_k^{(\text{iter})}, \phi_k^{(\text{iter})}$ **(MH on single marginal)**

  end

  Update $\tau^{(\text{iter})} | \gamma_k^{(\text{iter})}, f_k^{(\text{iter})}, g_k^{(\text{iter-1})}$

  Update $\rho_k^{(\text{iter})} | \gamma_k^{(\text{iter})}, \tau^{(\text{iter})}, f_k^{(\text{iter})}, g_k^{(\text{iter-1})}$

end

Initialize $\gamma_k^{(0)}, f_k^{(0)}, \beta_k^{(0)}, \phi_k^{(0)}, \tau^{(0)}, \rho_k^{(0)}$

for *iter* $\in 1 : num.iterations$ **do**

  Update $f_k^{(\text{iter})} | \gamma_k^{(\text{iter}-1)}, \tau^{(\text{iter}-1)}, \rho_k^{(\text{iter}-1)}, G_k^{(\text{iter-1})}$ **(closed-form)**

  Update $\beta_k^{(\text{iter})} | \gamma_k^{(\text{iter}-1)}, \phi_k^{(\text{iter}-1)}$ **(closed-form)**

  Update $\phi_k^{(\text{iter})} | \gamma_k^{(\text{iter}-1)}, \beta_k^{(\text{iter})}$

  for $i \in 1 : N$ **do**

    Update $\gamma_{ki}^{(\text{iter})} | \gamma_{k(-i)}^{(\text{iter})}, \tau_i^{(\text{iter}-1)}, f_k^{(\text{iter})}, \beta_k^{(\text{iter})}, \phi_k^{(\text{iter})}$ **(MH on single marginal)**

  end

  Update $\tau^{(\text{iter})} | \gamma_k^{(\text{iter})}, f_k^{(\text{iter})}, g_k^{(\text{iter-1})}$

  Update $\rho_k^{(\text{iter})} | \gamma_k^{(\text{iter})}, \tau^{(\text{iter})}, f_k^{(\text{iter})}, g_k^{(\text{iter-1})}$

end

MCMC Algorithm

Initialize $\gamma_k^{(0)}, f_k^{(0)}, \beta_k^{(0)}, \phi_k^{(0)}, \tau^{(0)}, \rho_k^{(0)}$

for *iter* $\in 1$: *num.iterations* **do**

  Update $f_k^{(\text{iter})} | \gamma_k^{(\text{iter}-1)}, \tau^{(\text{iter}-1)}, \rho_k^{(\text{iter}-1)}, G_k^{(\text{iter-1})}$ **(closed-form)**

  Update $\beta_k^{(\text{iter})} | \gamma_k^{(\text{iter}-1)}, \phi_k^{(\text{iter}-1)}$ **(closed-form)**

  Update $\phi_k^{(\text{iter})} | \gamma_k^{(\text{iter}-1)}, \beta_k^{(\text{iter})}$

  for $i \in 1 : N$ **do**

    Update $\gamma_{ki}^{(\text{iter})} | \gamma_{k(-i)}^{(\text{iter})}, \tau_i^{(\text{iter}-1)}, f_k^{(\text{iter})}, \beta_k^{(\text{iter})}, \phi_k^{(\text{iter})}$ **(MH on single marginal)**

  end

  Update $\tau^{(\text{iter})} | \gamma_k^{(\text{iter})}, f_k^{(\text{iter})}, g_k^{(\text{iter-1})}$

  Update $\rho_k^{(\text{iter})} | \gamma_k^{(\text{iter})}, \tau^{(\text{iter})}, f_k^{(\text{iter})}, g_k^{(\text{iter-1})}$

end

Posterior Inference - $g_k(s_i, t)$

Our MCMC of 11000 iterations manages to retrieve the simulated data:

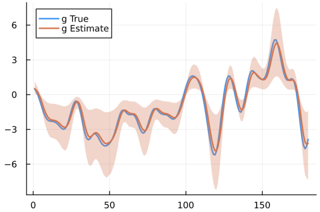

Figure: Retrieved g_k for site 1

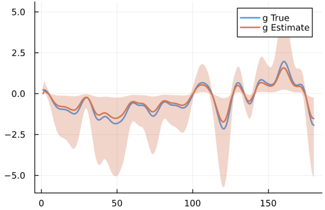

Figure: Retrieved g_k for site 2

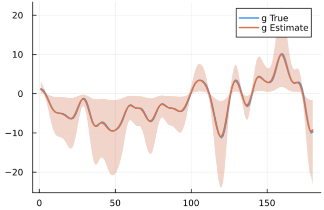

Figure: Retrieved g_k for site 3

Posterior Inference - f_k

We partially succeed in retrieving the second hierarchical level of our model.
Up to a constant, we successfully recover f_k. In particular, by standardizing both the true function $f_{k_{\text{true}}}$ and the estimated function $f_{k_{\text{estimate}}}$, we observe a clear alignment between them.

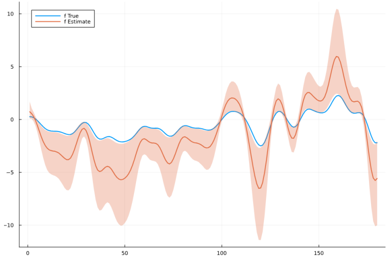

Figure: Retrieved original f_k

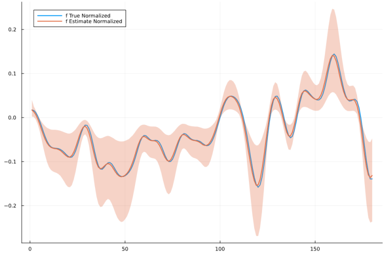

Figure: Retrieved normalized f_k

Posterior Inference - γ_k

On the other hand, due to the normalization applied to f_k, we must adjust γ_k accordingly. Specifically, we correct γ_k using the transformation: $\gamma_{k_{\text{adjusted}}} = \ln(\|f_k\|) + \gamma_k$. This adjustment ensures consistency between the model's inference and the generative process.

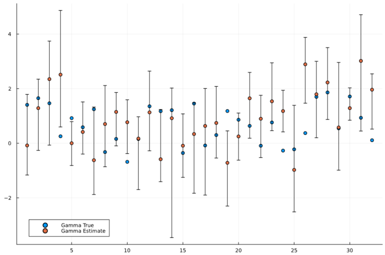

Figure: Retrieved original γ_k

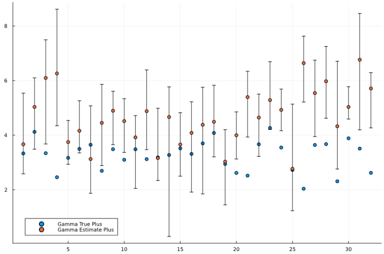

Figure: Retrieved adjusted γ_k

Convergence Diagnostic - β_k

Thanks to the closed-form solution for β_k, the trace plot exhibits well-mixed "fat caterpillar" behavior, indicating good sampling. Moreover, we successfully recover the true value of each component within the 0.95 credible interval, except for β_0, which is the intercept, so we do not care too much.

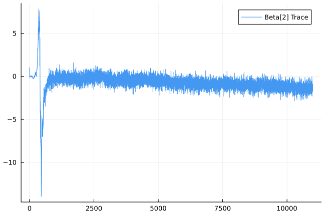

Figure: Traceplot of β_1

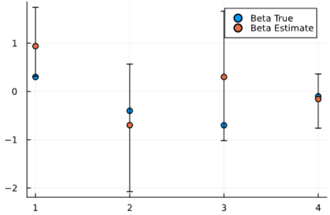

Figure: Estimated β's

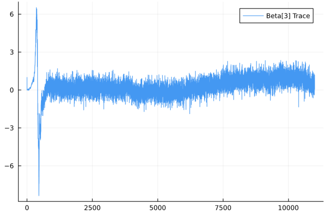

Figure: Traceplot of β_2

Convergence Diagnostic - Autocorrelation

To further assess the mixing quality of the chain, we examine the autocorrelation plots. Ideally, these plots should show a decreasing pattern as the lag increases, indicating good mixing. Given the high number of iterations, we apply thinning to improve clarity and reduce autocorrelation bias.

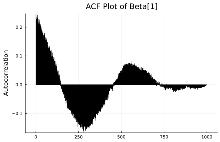

Figure: ACF of β_0

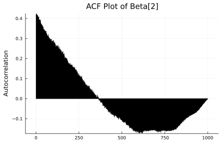

Figure: ACF of β_1

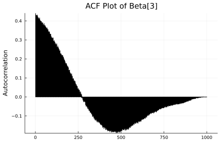

Figure: ACF of β_2

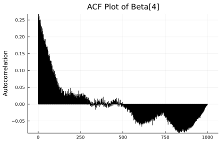

Figure: ACF of β_3

Conclusions - Key findings

In summary, we successfully achieved the following:

- Starting from the Vannucci's model for random phase amplitude Gaussian process in R language, we have implemented a complex pipeline in Julia to generate data as similar as possible to the real one, using a complex and rigorous Bayesian hierarchical model.
- After having done a lot of math calculations to get all the full conditionals, in particular to seek for possible closed form, we developed a customized MCMC algorithm for a deep, more-level-organized model and extended its efficiency over time.
- We retrieved the target functions g_k, representing the first level of the model, and second-level elements such as β_k and f.

Our approach was able to efficiently estimate the parameters, demonstrating the potential for handling intricate spatial-temporal models with hierarchical structures.

Conclusions - Future directions

To further refine the model and improve its performance, the following steps are proposed:

- **Implement shrinkage techniques**: Introducing shrinkage relaxes the constraint of a fixed number of pollutant sources K, making the model more flexible in automatically selecting the relevant sources.
- **Extend MCMC to the y variable**: Incorporate the observed pollutant levels y into the MCMC framework to enhance inference at the data level. This involves applying a transformation to ensure the data remains positive, improving model stability and interpretability.
- **Test the model on real ARPA data**: Applying the model to real-world data will provide valuable insights into its practical applicability and robustness.

Thank You for The Attention!

Presented by:
G. Aurina, **P. Bettanin**, **A. Calanchi**,
R. Scorza, **C. Strino**, **F. Volpicelli**

Full conditional of β_k

- Model: $\gamma_k \mid \beta_k, X, \phi_k \overset{ind}{\sim} \mathcal{N}_N(X\beta_k, \Sigma_{\gamma_k})$
- Marginal prior: $\beta_k \overset{iid}{\sim} \mathcal{N}_P(0, I_p)$
- Full conditional: $\beta_k \mid \gamma_k, X, \phi_k \overset{ind}{\sim} \mathcal{N}_P(m, S^{-1})$
- $S = X^T\Sigma_{\gamma_k}^{-1}X + I_p,\ m = (X^T\Sigma_{\gamma_k}^{-1}X + I_p)^{-1}X^T\Sigma_{\gamma_k}^{-1}\gamma_k$

Explanation: The posterior of β_k is a closed-form normal distribution, computed from the likelihood and the prior. This is efficient to compute and does not require sampling.

Full conditional of ϕ_k (MH)

- Transformation: $\theta_k = \log(\phi_k)$
- Proposal: $q(\theta^*|\theta^{\text{old}}) = \mathcal{N}(\theta^{\text{old}}, \varepsilon)$
- Target: $\mathcal{L}(\gamma_k|\beta_k, X, e^{\theta_k})\pi_{\Phi_k}(e^{\theta_k})e^{\theta_k}$

Explanation: The posterior of Φ_k is not available in closed form. We apply Metropolis-Hastings within Gibbs steps to sample from the posterior by using a log-transformation and a Gaussian proposal distribution.

Full conditional of τ (MH)

- Target: $\mathcal{L}(\tau|\gamma_k, f_k) \propto \prod_{i=1}^{N} \mathcal{L}(g_k(s_i)|\gamma_{ki}, \tau_i, f_k)\pi(\tau)$

Explanation: The full conditional of τ requires Metropolis-Hastings because it does not have a closed-form solution. We sample τ iteratively using the likelihood and prior.

Full conditional of ρ_k (MH)

- Transformation: $\theta = \log(\rho_k)$
- Target: $\mathcal{L}(\rho_k | \gamma_k, \tau, f_k) \propto \mathcal{L}(g_k(s_i) | \gamma_{ki}, \tau_i, f_k) \pi_\rho(e^\theta) e^\theta \mathcal{L}(f_k)$

Explanation: Similar to Φ_k, the full conditional of ρ_k is transformed into θ to make it more amenable to sampling via Metropolis-Hastings.

Full conditional of g_k

- $g_k(s_i, t) \mid f_k, \gamma_{k_i}, \tau_i \overset{ind}{\sim} \mathcal{N}_T(\mu_g, \Sigma_g)$
- Mean: $\mu_g = e^{\gamma_{k_i}} \Sigma_{k_i} \Sigma_{f_k}^{-1} f_k$
- Covariance: $\Sigma_g = e^{2\gamma_{k_i}} (\Sigma_{f_k} - \Sigma_{k_i} \Sigma_{f_k}^{-1} \Sigma_{k_i}^T + \epsilon I)$
- $\Sigma_{k_i}[m, n] = \exp\left\{ -\frac{\rho_k^2}{2} (t_m - \tau_i - t_n)^2 \right\}$

Explanation: The full conditional of g_k is available in closed form as a multivariate normal distribution, with mean and covariance dependent on the parameters f_k and γ_{k_i}.

Full conditional of f_k

- $f_k | \mathcal{G}_k, \gamma_k, \tau, \rho_k \overset{ind}{\sim} \mathcal{N}(\mu_{f\text{-post}}, \Sigma_{f\text{-post}})$

$$\Sigma_{f_k\text{-post}} = (\Sigma_{f_k}^{-1} + \sum_{i=1}^{n} (\Sigma_{k_i} \Sigma_{f_k}^{-1})^T \Sigma_{g_{ki}}^{-1} \Sigma_{k_i} \Sigma_{f_k}^{-1})^{-1} \tag{1}$$

$$\mu_{f_k\text{-post}} = \left(\Sigma_{f_k}^{-1} + \sum_{i=1}^{n} \left(\Sigma_{k_i} \Sigma_{f_k}^{-1} \right)^T \Sigma_{g_{ki}}^{-1} \Sigma_{k_i} \Sigma_{f_k}^{-1} \right)^{-1} \sum_{i=1}^{n} G_k[i,:]^T \Sigma_{g_{ki}}^{-1} \Sigma_{k_i} \Sigma_{f_k}^{-1} \tag{2}$$

Explanation: The full conditional of f_k is a closed-form normal distribution, computed from the likelihood and prior, based on the observed data $\mathcal{G}_k$ and other parameters.

Full conditional of γ_k (MH)

- Full conditional of γ_k not available in closed form
- Use marginalization and conditional sampling: $\gamma_{k_i} \mid \gamma_k^{(-i)} \sim \mathcal{N}(\mu_{k_i}, \Sigma_{k_i})$
- Metropolis-Hastings target function for γ_{k_i}: $\mathcal{L}(g_k(s_i) \mid \gamma_{k_i}, \tau_i, f_k)\mathcal{L}(\gamma_{k_i} \mid \gamma_k^{(-i)})$
- Computational advantage: Reduce dimensionality by sampling components of γ_k iteratively

Explanation: The full conditional of γ_k cannot be computed in closed form, so we apply Metropolis-Hastings to sample from its distribution, iteratively updating each component of γ_k.

Motivation: In our model, the number of pollutant sources K is fixed. By introducing a shrinkage mechanism, we allow the model to select K by controlling the variance of each source via a shrinkage parameter η_k.

$$g_k(s_i, t) = \gamma_{ki} \cdot f_k(t - \tau_i)$$

- $f_k \overset{\text{ind}}{\sim} \mathcal{N}_T^+\left(\mathbf{0}, \frac{1}{\eta_k}\Sigma_f\right), \quad \Sigma_f(t', t) = \exp\left(-\frac{\rho^2}{2}(t' - t)^2\right)$
- $\eta_k = \prod_{l=1}^{k} \delta_l, \quad \delta_1 \sim \text{Gamma}(a_1, 1), \quad \delta_l \sim \text{Gamma}(a_2, 1)$ for $l \geq 2$
- $\rho \overset{iid}{\sim} \text{Gamma}(a_\rho, b_\rho)$

This mechanism get rid of redundant sources by retaining only those with sufficiently large variance.

Marginal Priors and Proposals of our parameters

Marginal Prior for τ :

$$\tau \sim \mathcal{N}_N \left(0, \sigma_\tau^2 \left(I_n - \frac{1}{n+1}\mathbf{1}\mathbf{1}'\right)\right), \tag{3}$$

$\sigma_\tau = \sqrt{3}$

Proposal for τ :

$$\tau \sim \mathcal{N}_N \left(0, \sigma_\tau^2 \left(I_n\right)\right), \tag{4}$$

$\sigma_\tau = 0.01$

Bibliography

[1] Oliver Baerenbold et al. *A dependent Bayesian Dirichlet process model for source apportionment of particle number size distribution*. 2022.

[2] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

[3] A. Bhattacharya and D. B. Dunson. *Sparse Bayesian infinite factor models*. 2011.

[4] Jenna R. Krall and Howard H. Chang. *Statistical methods for source apportionment*. 2019.

[5] Dustin Pluta and Marina Vannucci. *Improved data quality and statistical power of trial-level event-related potentials with Bayesian random-shift Gaussian processes*. 2024.