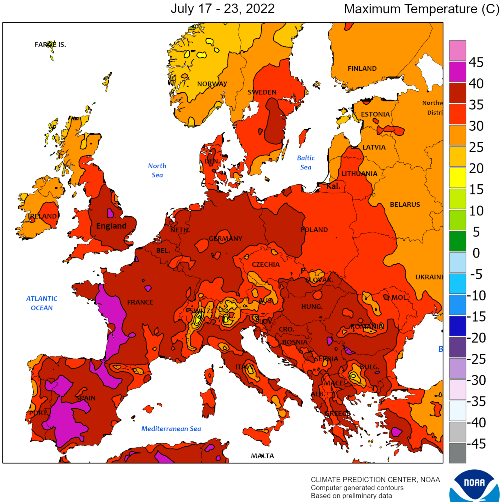
Spatio-temporal generation of precipitations using a spatially correlated Bernoulli and hidden Markov model

Caroline Cognot (EDF & MIA-PS)
Supervision : Liliane Bel (MIA-PS), Sylvie Parey (EDF)
David Métivier (INRAE - MISTEA)

Contents

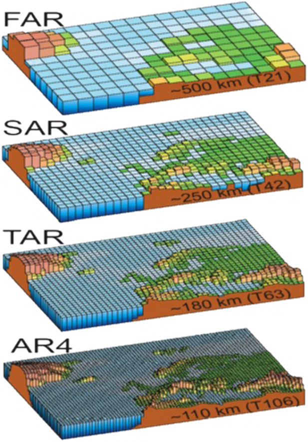
Climate change impact

- Extremes and risks: impact on agriculture, health, energy...
- Climate change: impact on frequency, intensity,...



Goal: Estimate (future) risks quantitatively
$\rightarrow$ In particular large scales extremes

Climate (physical) models

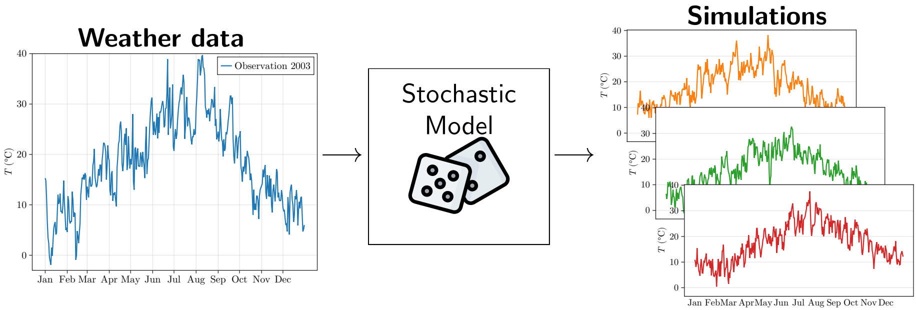


Climate models grid size

+ Used in climate projection - global and regional models
+ Complex phenomena, physical equations
+ Many variables
+ High spatiotemporal resolution

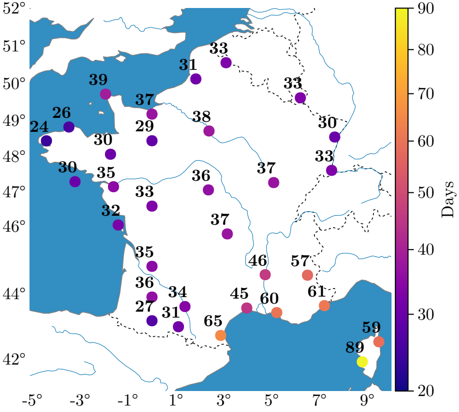
– Computationally expensive
– Not very good for extremes

Stochastic Weather Generators



- Trained on observations or climate model outputs
- Simulations: same statistical properties as observations
- Spatio-temporal resolution depends on the model and the data
- Monovariate/Multivariate e.g. (temperature and precipitation)
- Fast

Multisite Rainfall Weather Generators



$S = 37$ stations, ECA&D

Objective: Build a stochastic weather generator

- For the multisite rain occurrence $Y^{(t)} = (Y_1^{(t)}, \dots, Y_S^{(t)})$

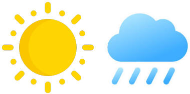
→ Reproduce the spatial-temporal structure of the data.

- In particular **large scales dry/wet episodes**

First question: Should we separate rain occurrence from rain amount ?
e.g. Benoit et al. (2018) with Censored Gaussian models
→ In this project : separate both.

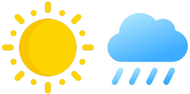
Single-site weather types model

Initial idea of Richardson 1981 : 2 weather types.



Single-site weather types model

Initial idea of Richardson 1981 : 2 weather types.



- The weather states/regimes/types Z_t are a Markov chain of order r :

$$P(Z_t|Z_{t-1},\dots,Z_{t-r}) \tag{1}$$

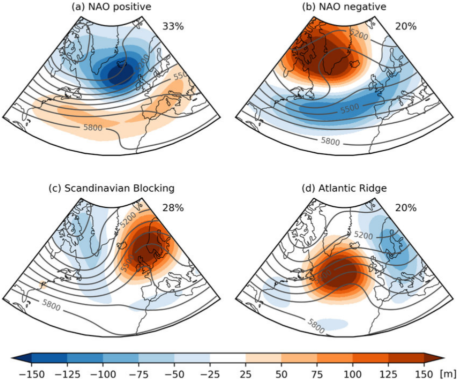
- Observed meteorological variables Y_t are generated conditionally on Z_t using **emission distributions**

$$P(Y_t|Z_t) \tag{2}$$

Separates the complexity of the model into several categories

Spatial weather types

Weather types derived from atmospheric circulation or predefined

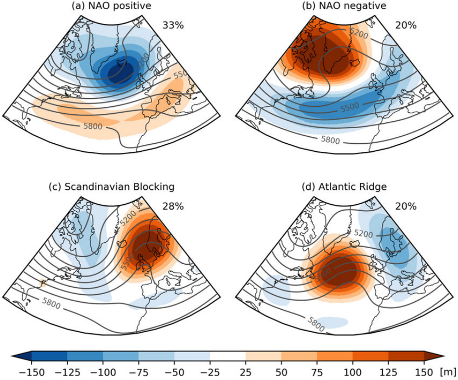


Atlantic weather types

→ Not data-driven

Spatial weather types

Weather types derived from atmospheric circulation or predefined

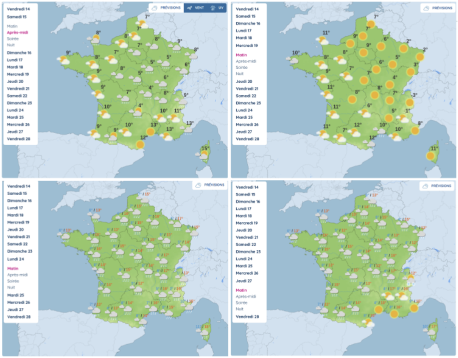


Atlantic weather types

→ Not data-driven

Hidden Markov chain

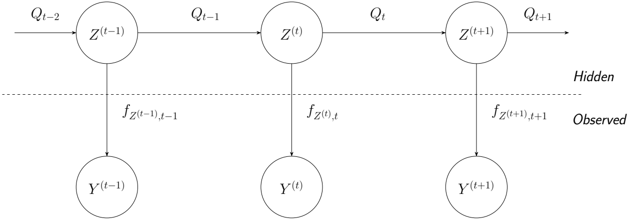
→ Weather types learned from the data



→**This work**

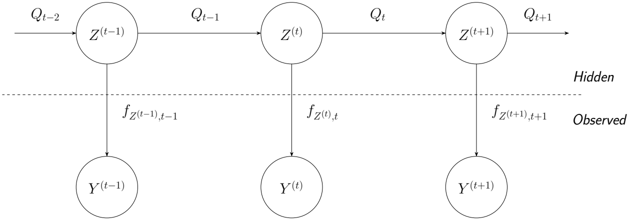
Multisite HMM based model

Rain occurrence $Y^{(t)} = (Y_1^{(t)}, \dots, Y_S^{(t)}) \in \{0,1\}^S$ - Unobserved weather type $Z^{(t)} \in \{1, \dots, K\}$



Multisite HMM based model

Rain occurrence $Y^{(t)} = (Y_1^{(t)}, \dots, Y_S^{(t)}) \in \{0, 1\}^S$ - Unobserved weather type $Z^{(t)} \in \{1, \dots, K\}$



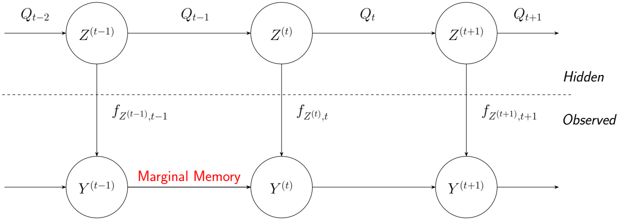
Conditional independence (Zucchini et al. 1991, Gobet et al. 2024)

$$\mathbb{P}\left(Y = y \mid Z = z^{(t)}\right) = f_{z^{(t)},t}(y) = \prod_{s=1}^{S} \left(y_s \lambda_{Z^{(t)},t,s} + (1 - y_s)(1 - \lambda_{Z^{(t)},t,s})\right)$$

With $\lambda_{k,t,i} = \mathbb{P}(Y_i^{(t)} = 1 | Z^{(t)} = k)$ for $i \in 1, \dots, \mathcal{S}$

Multisite HMM based model

Rain occurrence $Y^{(t)} = (Y_1^{(t)}, \dots, Y_S^{(t)}) \in \{0,1\}^S$ - Unobserved weather type $Z^{(t)} \in \{1, \dots, K\}$



Conditional independence (Zucchini et al. 1991, Gobet et al. 2024)

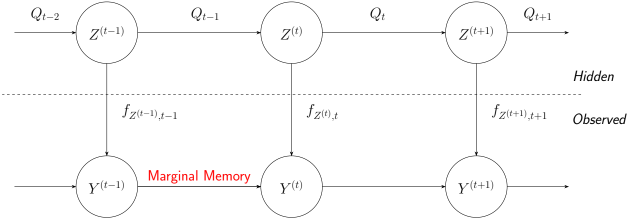
$$\mathbb{P}\left(Y = y \mid Z = z^{(t)}, Y^{(t-1)} = y^{t-1}\right) = f_{z^{(t)},t}(y) =$$

$$\prod_{s=1}^{S}\left(y_s \lambda_{Z^{(t)},t,s,y_s^{t-1}} + (1-y_s)(1-\lambda_{Z^{(t)},t,s,y_s^{t-1}})\right)$$

With $\lambda_{k,t,i,y_i^{t-1}} = P(Y_i^{(t)} = 1 | Z^{(t)} = k, Y_i^{(t-1)} = y_i^{t-1})$ for $i \in 1, \dots, \mathcal{S}$

Multisite HMM based model

Rain occurrence $Y^{(t)} = (Y_1^{(t)}, \dots, Y_S^{(t)}) \in \{0,1\}^S$ - Unobserved weather type $Z^{(t)} \in \{1, \dots, K\}$



Conditional independence (Zucchini et al. 1991, Gobet et al. 2024)

– Stations must be "far apart enough": 10 stations in the paper

\+ Correlations between stations are captured by the weather types Z

Large scale weather types

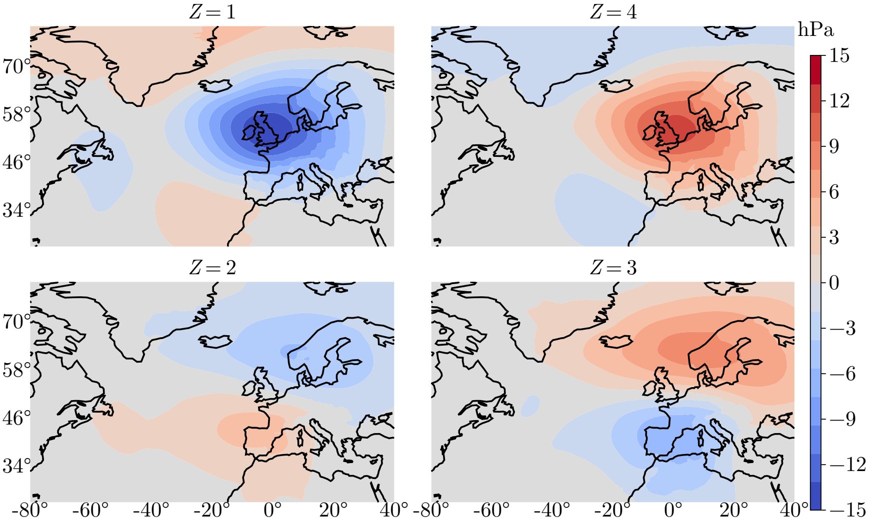
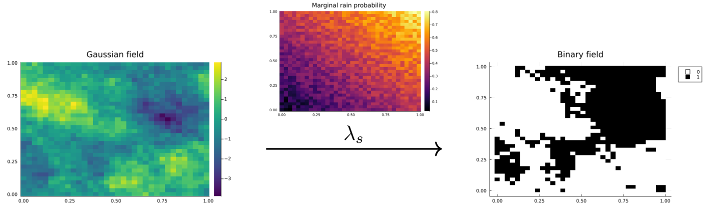


Figure: Pressure map $\Delta P = \mathbb{E}\left(P \mid Z=k\right) - \mathbb{E}\left(P\right)$ for winter months

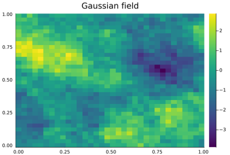
How to generate binary correlated variables

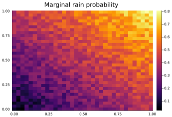


$X \sim \mathcal{N}(0, C_\theta)$: latent

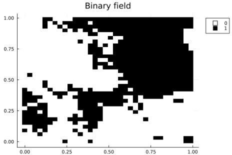
$$\forall s, Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$$

How to generate binary correlated variables





$\xrightarrow{\lambda_s}$

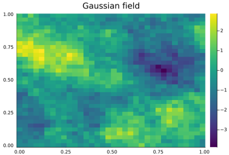


$X \sim \mathcal{N}(0, C_\theta)$: latent

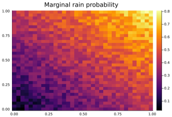
$$\forall s, Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$$

Parameters : $\lambda_i = \mathbb{P}(Y_i = 1)$ for $i \in 1, \ldots, \mathcal{S}$ and θ the parameters of the latent covariance.

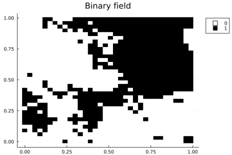
How to generate binary correlated variables



$X \sim \mathcal{N}(0, C_\theta)$: latent



$$\xrightarrow{\lambda_s}$$



$$\forall s, Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$$

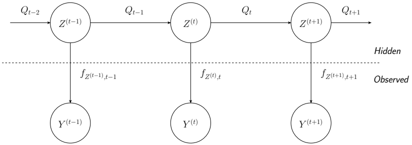
Parameters : $\lambda_i = \mathbb{P}(Y_i = 1)$ for $i \in 1, \dots, \mathcal{S}$ and θ the parameters of the latent covariance.

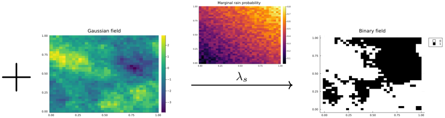
Associated distribution

$$f(y, \theta) = \int_{a_1}^{b_1} \cdots \int_{a_\mathcal{S}}^{b_\mathcal{S}} f_X((x_1, ..., x_\mathcal{S}))d(x_1, ..., x_\mathcal{S})$$

With $a_i = -\infty$ if $Y_i = 1, \Phi^{-1}(\lambda_i)$ else, $b_i = \infty$ if $Y_i = 0, \Phi^{-1}(\lambda_i)$ else

Multisite mixed model

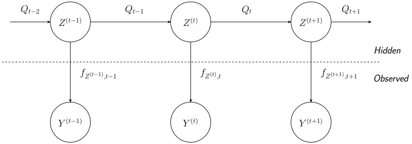


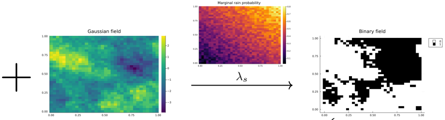


$X \sim \mathcal{N}(0, C_\theta)$: latent

$$\forall s, Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$$

Multisite mixed model





$X \sim \mathcal{N}(0, C_\theta)$: latent

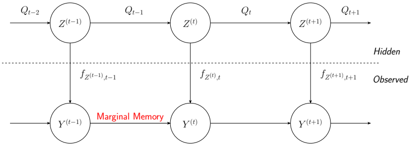
$$\forall s, Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$$

Spatial conditional dependence

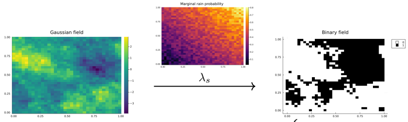
- Covariance function : $C_{\theta_{k,t}}(h)$ with distance h between stations for state K
- $\lambda_{k,t,i} = \mathbb{P}(Y_i^{(t)} = 1 | Z^{(t)} = k)$ for $i \in 1, \dots, \mathcal{S}$
- $\mathbb{P}\big(Y^{(t)} = y \mid Z^{(t)} = z\big) = \displaystyle\int_{a_1}^{b_1} \cdots \int_{a_S}^{b_S} f_X^{(\theta_z, t)}(x_1, \dots, x_S)\, \mathrm{d}x_1 \cdots \mathrm{d}x_S$

  With $a_s = -\infty$ if $y_s = 1, \Phi^{-1}(\lambda_{z,t,s})$ else, $b_i = \infty$ if $y_i = 0, \Phi^{-1}(\lambda_{z,t,s})$ else

Multisite mixed model



$+$



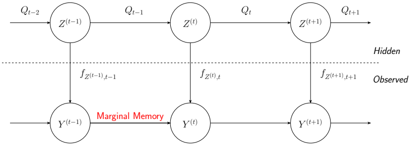
$X \sim \mathcal{N}(0, C_\theta)$: latent

$\forall s, Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$

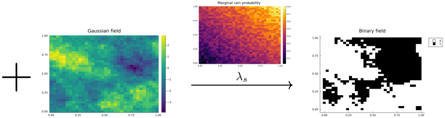
Spatial conditional dependence

- Covariance function $C_{\theta_{k,t}}(h)$ with distance h between stations for state K
- $\lambda_{k,t,i,y_i^{t-1}} = \mathbb{P}(Y_i^{(t)} = 1 | Z^{(t)} = k, Y_i^{(t-1)} = y_i^{t-1})$ for $i \in 1, \dots, \mathcal{S}$
- $\mathbb{P}\big(Y^{(t)} = y \mid Z^{(t)} = z, Y_i^{t-1} = y_i^{t-1}\big) =$
  $$\int_{a_1}^{b_1} \cdots \int_{a_S}^{b_S} f_X^{(\theta_{z,t})}(x_1, \dots, x_S)\, \mathrm{d}x_1 \cdots \mathrm{d}x_S$$
  With $a_s = -\infty$ if $y_s = 1, \Phi^{-1}(\lambda_{z,t,s,y_i^{t-1}})$ else, $b_i = \infty$ if $y_i = 0, \Phi^{-1}(\lambda_{z,t,s,y_i^{t-1}})$ else

Multisite mixed model



$+$



$X \sim \mathcal{N}(0, C_\theta)$: latent

$$\forall s, Y_s = \begin{cases} 1 & \text{if } X_s \leq \Phi^{-1}(\lambda_s) \\ 0 & \text{else} \end{cases}$$

Spatial conditional dependence

- $+$ Spatial structure $\rightarrow$ less parameters than $O(\mathcal{S}^2)$
- $+$ No restriction on the stations distance
- $\pm$ Correlations between stations are captured by the weather types Z and C_θ
- $-$ The emission is more complicated !

Parameterization : periodic parameters

Parameters

- $\lambda_{k,t,s}$: rain probability at site s, time t, for state k
- $\rho_{k,t}$: latent spatial covariance parameter in exponential covariance model
- $Q_t(k,l), \pi(k)$: transition probabilities and initial probabilities.

Parameterization : periodic parameters

Parameters

- $\lambda_{k,t,s}$: rain probability at site s, time t, for state k
- $\rho_{k,t}$: latent spatial covariance parameter in exponential covariance model
- $Q_t(k,l), \pi(k)$: transition probabilities and initial probabilities.

$$P_\theta(t) = c_0^\theta + \sum_{j=1}^{degP} \left(c_{2j-1}^\theta \cos(2\pi jt/T) + c_{2j}^\theta \sin(2\pi jt/T) \right)$$

Parameterization : periodic parameters

Parameters

- $\lambda_{k,t,s}$: rain probability at site s, time t, for state k
- $\rho_{k,t}$: latent spatial covariance parameter in exponential covariance model
- $Q_t(k,l), \pi(k)$: transition probabilities and initial probabilities.

$$P_\theta(t) = c_0^\theta + \sum_{j=1}^{degP} \left(c_{2j-1}^\theta \cos(2\pi jt/T) + c_{2j}^\theta \sin(2\pi jt/T) \right)$$

$$\lambda_{k,t,s} = \frac{1}{1 + \exp(P^{\lambda_{k,t,s}}(t))}$$

$$\rho_{k,t} = \exp\left(P^{\rho_{k,t}}(t)\right)$$

$$Q_t(k,l) = \frac{e^{P^{Q_t(k,l)}(t)}}{1 + \sum_{l=1}^{K-1} e^{P^{Q_t(k,l)}(t)}} \; \forall\, l < K, \quad Q_t(k,K) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{P^{Q_t(k,l)}(t)}}$$

Parameterization : periodic parameters

Parameters

- $\lambda_{k,t,s}$: rain probability at site s, time t, for state k
- $\rho_{k,t}$: latent spatial covariance parameter in exponential covariance model
- $Q_t(k,l), \pi(k)$: transition probabilities and initial probabilities.

Hyperparameters

K : number of classes
$degP$: degree of seasonality polynomial

Adding the Rain intensity

For now : we get $Z^{(t)}$ the state, $(Y_1^{(t)}, ..., Y_{\mathcal{S}}^{(t)})$ the occurence of rain.

Adding the Rain intensity

For now : we get $Z^{(t)}$ the state, $(Y_1^{(t)}, ..., Y_{\mathcal{S}}^{(t)})$ the occurence of rain.

What about the amount of rain $(R_1^{(t)}, ..., R_{\mathcal{S}}^{(t)})$?

Adding the Rain intensity

For now : we get $Z^{(t)}$ the state, $(Y_1^{(t)}, ..., Y_{\mathcal{S}}^{(t)})$ the occurence of rain.

What about the amount of rain $(R_1^{(t)}, ..., R_{\mathcal{S}}^{(t)})$?

Model for each location s using EGPD (Naveau et al 2016) :

$$R_s = \sigma_s H_{\xi_s}^{-1}\left[G^{-1}(U_s)\right]$$

where $U_s \sim \mathcal{U}([0,1])$, $G(v) = v^{\kappa_s}$ and H_{ξ_s} : $GPD(1, \xi_s)$.

Adding the Rain intensity

For now : we get $Z^{(t)}$ the state, $(Y_1^{(t)}, ..., Y_{\mathcal{S}}^{(t)})$ the occurence of rain.

What about the amount of rain $(R_1^{(t)}, ..., R_{\mathcal{S}}^{(t)})$?

Model for each location s using EGPD (Naveau et al 2016) :

$$R_s = \sigma_s H_{\xi_s}^{-1}\left[G^{-1}(U_s)\right]$$

where $U_s \sim \mathcal{U}([0,1])$, $G(v) = v^{\kappa_s}$ and $H_{\xi_s} : GPD(1, \xi_s)$.

Make it correlated between all sites ?

Adding the Rain intensity

For now : we get $Z^{(t)}$ the state, $(Y_1^{(t)}, ..., Y_{\mathcal{S}}^{(t)})$ the occurence of rain.

What about the amount of rain $(R_1^{(t)}, ..., R_{\mathcal{S}}^{(t)})$?

Model for each location s using EGPD (Naveau et al 2016) :

$$R_s = \sigma_s H_{\xi_s}^{-1} \left[G^{-1}(U_s) \right]$$

where $U_s \sim \mathcal{U}([0,1])$, $G(v) = v^{\kappa_s}$ and H_{ξ_s} : $GPD(1, \xi_s)$.

Make it correlated between all sites ?

$$V_s = \Phi^{-1}(U_s) = \Phi^{-1} \left[G_{\kappa_s} \left[H_{\xi_s} \left(\frac{R_s}{\sigma_s} \right) \right] \right] \sim \mathcal{N}(0, C_\theta)$$

Idea :

1. Estimate EGPD parameters $\sigma_{k,\text{season},s}, \xi_{k,\text{season},s}, \kappa_{k,\text{season},s}$
2. Transform R_s to V_s
3. Find covariance parameters $\theta_{k,\text{season}}$

HMM : Maximum likelihood estimation

- Hidden states $\rightarrow$ Expectation-Maximization (EM) algorithm
- Seasonal parameters $\theta(t)$

!! High dimensional integrals

HMM : Maximum likelihood estimation

- Hidden states $\rightarrow$ Expectation-Maximization (EM) algorithm
- Seasonal parameters $\theta(t)$

!! High dimensional integrals

$\rightarrow$ Tricks make the problem computable

Example: Maximize composite pairwise likelihood $\sum_{\text{pairs}i,j} w_{ij} \log L((y_i, y_j); \theta)$ instead of full likelihood $\log L((y_1, \ldots, y_{\mathcal{S}}); \theta)$ during the **M** step of the EM algorithm.

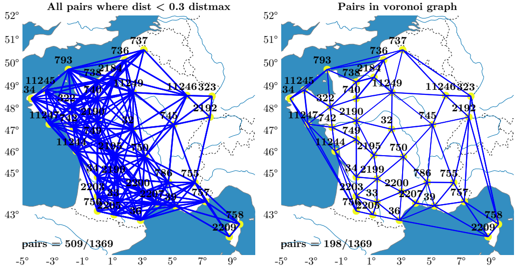
HMM : Maximum likelihood estimation

- Hidden states → Expectation-Maximization (EM) algorithm
- Seasonal parameters $\theta(t)$

!! High dimensional integrals

→ Tricks make the problem computable

Example: Maximize composite pairwise likelihood $\sum_{\text{pairs}i,j} w_{ij} \log L((y_i, y_j); \theta)$ instead of full likelihood $\log L((y_1, \ldots, y_S); \theta)$ during the **M** step of the EM algorithm.



Composite likelihood, Varin, Reid, and Firth 2011

For $Y \sim f_\theta$ of high dimension m where $L(y, \theta)$: complicated to compute :

Definition

For $A_1, \ldots, A_K$ marginal or conditional events with $L_k(\theta; y) \propto f(y \in A_k; \theta)$, and $w_k \geq 0$ weights

$$L_C(\theta; y) = \prod_{k=1}^{K} L_k(\theta; y)^{w_k},$$

L_k are easier to compute ! !

Composite likelihood, Varin, Reid, and Firth 2011

For $Y \sim f_\theta$ of high dimension m where $L(y, \theta)$: complicated to compute :

Definition

For $A_1, \ldots, A_K$ marginal or conditional events with $L_k(\theta; y) \propto f(y \in A_k; \theta)$, and $w_k \geq 0$ weights

$$L_C(\theta; y) = \prod_{k=1}^{K} L_k(\theta; y)^{w_k},$$

L_k are easier to compute ! !

Convergence

For n iid realizations of Y, $L_C(\theta; (y^1 ... y^n)) = \prod_{t=1}^{n} L_C(\theta; y^t)$

$$\sqrt{n}(\hat{\theta}_{CL} - \theta) \xrightarrow{d} N_p(0, G^{-1}(\theta)). \tag{3}$$

with G the Godambe matrix, $G = HJ^{-1}H$

Maximizing the composite likelihood gives asymptotically the true parameter.

EM algorithm : Implementation

Bottleneck : computing multivariate normal integrals of type

$$Pr(X_1 < x_1, \ldots, X_{\mathcal{S}} < x_{\mathcal{S}}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_{\mathcal{S}}} f_X((x'_1, ..., x'_{\mathcal{S}}))d(x'_1, ..., x'_{\mathcal{S}})$$

- Dimension $\mathcal{S}$ (in E step): Quasi-Monte-Carlo, julia package MvNormalCDF. Best method, from Genz 1992.

EM algorithm : Implementation

julia

Bottleneck : computing multivariate normal integrals of type

$$Pr(X_1 < x_1, \ldots, X_{\mathcal{S}} < x_{\mathcal{S}}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_{\mathcal{S}}} f_X((x_1', ..., x_{\mathcal{S}}'))d(x_1', ..., x_{\mathcal{S}}')$$

- Dimension $\mathcal{S}$ (in E step): Quasi-Monte-Carlo, julia package MvNormalCDF. Best method, from Genz 1992.
- Other options : approximations, unfortunately bad for highly correlated model.
- Dimension 2 (in M step) : Expression from Tsay and Ke 2023

Model selection

Integrated Complete Likelihood

$$ICL(K, \mathsf{deg}_P) = L(Y, Z, \theta) - \frac{\log(n)}{2}|\theta|$$

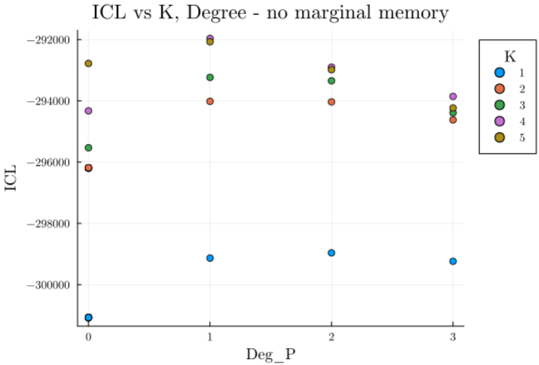
With Z the real hidden states... but they are hidden !

Model selection

Integrated Complete Likelihood

$$\widehat{ICL}(K, \text{deg}_P) = L(Y, \widehat{Z}, \theta) - \frac{\log(N)}{2}|\theta|$$

With $\widehat{Z}$ the MAP sequence (Viterbi).

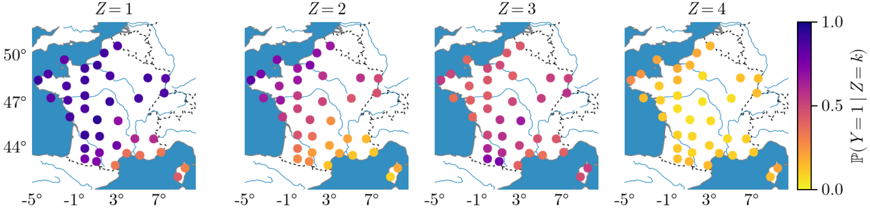


Parameters fitted from real data

Rain probability

$\lambda_{k,t,s}$: probability of rain in state k, at date t, station s.

Mean rain probability for each state :

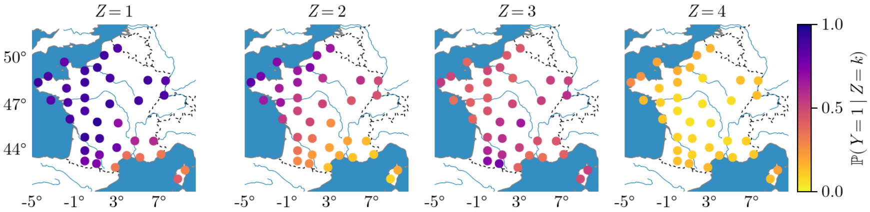


Parameters fitted from real data

Rain probability

$\lambda_{k,t,s}$: probability of rain in state k, at date t, station s.

Mean rain probability for each state :

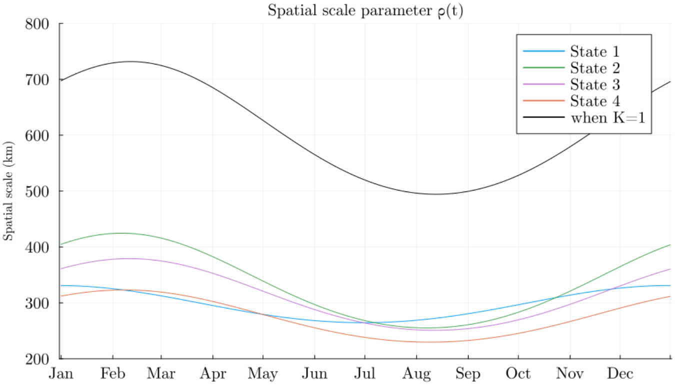


- North/ South expected difference
- Separate states.

Parameters fitted from real data

Spatial scale $\rho_{k,t}$

latent $X^{(t)}|Z^{(t)} = k \sim \mathcal{N}(0, C_{\rho_{k,t}})$ where $C_{\rho_{k,t}}(h) = \exp(-h/\rho_{k,t})$ $\rho_{k,t}$: spatial scale in exponential covariance model (latent variable)



How to evaluate seasonal large scales properties

- Seasonal effects
  Proposal : we use periodic, time-varying parameters.
- Quantify long, dry/wet episodes
- Quantify spatially large episodes

Spatio-temporal indicator to evaluate models :

Rain Occurence Ratio

$$\mathrm{ROR}(t) = \frac{\sum_{s \in \mathcal{S}} Y_s^{(t)}}{|\mathcal{S}|}$$

Spatiotemporal evaluation

$$\text{Rain Occurrence Rate (ROR)} = \frac{\sum_{s \in \mathcal{S}} Y_s^{(t)}}{|\mathcal{S}|}$$

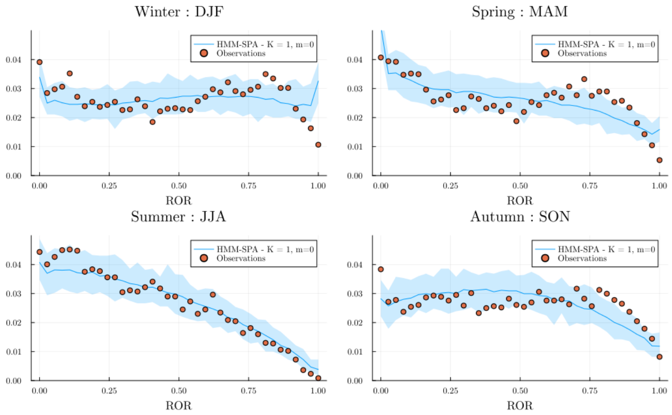


Figure: Distribution of ROR for the model with K = 1, no marginal memory

Spatiotemporal evaluation

$$\text{Rain Occurrence Rate (ROR)} = \frac{\sum_{s\in\mathcal{S}} Y_s^{(t)}}{|\mathcal{S}|}$$

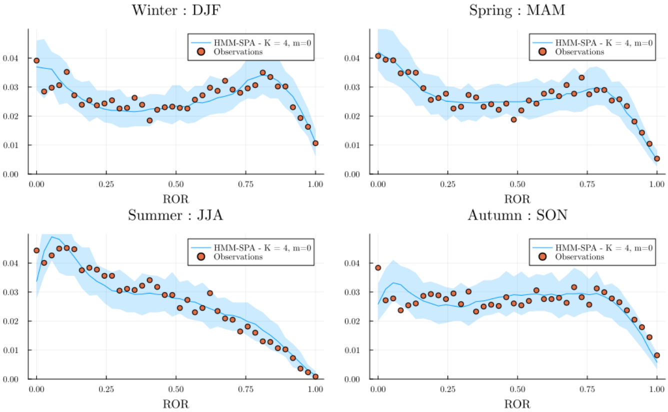


Figure: Distribution of ROR for the model with K = 4, no marginal memory

Spatiotemporal evaluation

$$\text{Rain Occurrence Rate (ROR)} = \frac{\sum_{s\in\mathcal{S}} Y_s^{(t)}}{|\mathcal{S}|}$$

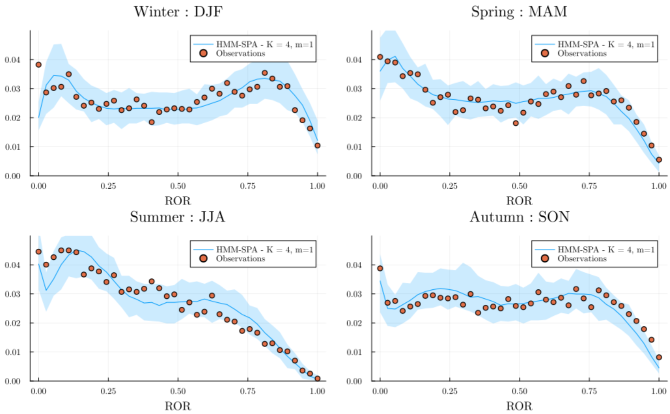


Figure: Distribution of ROR for the model with $K = 4$, with marginal memory

Dry spells

Dry spell of Rain Occurrence Rate (ROR)

$$P_\ell = \mathbb{P}(ROR_t < 0.2, ..., ROR_{t+\ell} < 0.2)$$

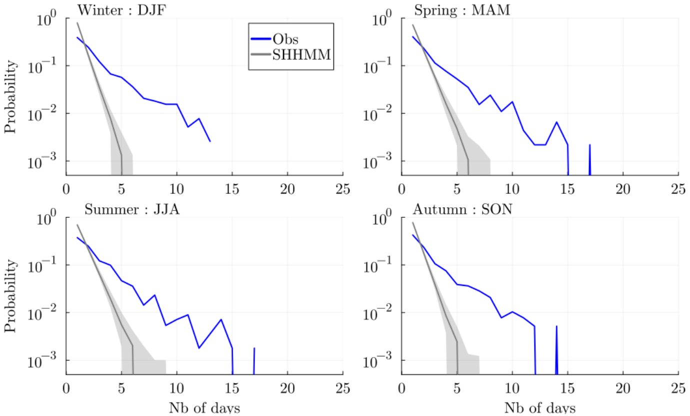


Figure: Distribution of ROR dry spells for the model with K = 1, no marginal memory

Dry spells

Dry spell of Rain Occurrence Rate (ROR)

$$P_\ell = \mathbb{P}(ROR_t < 0.2, ..., ROR_{t+\ell} < 0.2)$$

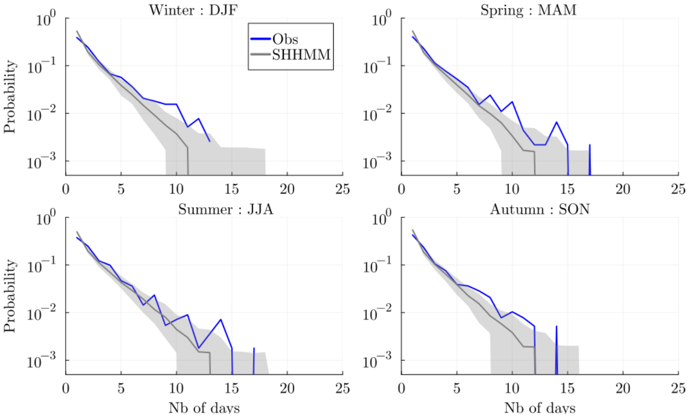


Figure: Distribution of ROR dry spells for the model with K $=$ 4, no marginal memory

Dry spells

Dry spell of Rain Occurrence Rate (ROR)

$$P_\ell = \mathbb{P}(ROR_t < 0.2, ..., ROR_{t+\ell} < 0.2)$$

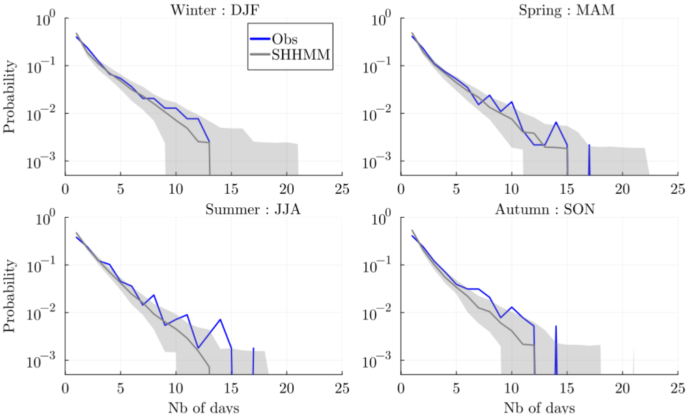


Figure: Distribution of ROR dry spells for the model with K = 4, with marginal memory

Conclusion

What has been done

- Seasonal Multisite rain occurrence model with hidden weather regimes
- No restriction on the distance between stations
- Evaluation with the spatiotemporal indicator ROR

Conclusion

What has been done

- Seasonal Multisite rain occurrence model with hidden weather regimes
- No restriction on the distance between stations
- Evaluation with the spatiotemporal indicator ROR

What is left to do

- Finalize the model selection, analyze the new weather regimes Z
- Add rainfall amounts $+$ other weather variables

Conclusion

What has been done

- Seasonal Multisite rain occurrence model with hidden weather regimes
- No restriction on the distance between stations
- Evaluation with the spatiotemporal indicator ROR

What is left to do

- Finalize the model selection, analyze the new weather regimes Z
- Add rainfall amounts + other weather variables

Other works

- High resolution large scale temperature model

Potential computational improvement : Composite Likelihood EM (CLEM) algorithm

!! Issue in the EM : E-step still require $\mathcal{S}-$dimensional integrals

Potential computational improvement : Composite Likelihood EM (CLEM) algorithm

!! Issue in the EM : E-step still require $\mathcal{S}-$dimensional integrals

- Gao and Song 2011 : use the pairwise likelihood as objective !

classic EM : $R(\theta, \theta^{(q)}) = E_{Z \sim p(;Y, \theta^{(q)})}(log(L(\theta, Z, Y))|Y^{obs})$

Potential computational improvement : Composite Likelihood EM (CLEM) algorithm

!! Issue in the EM : E-step still require $\mathcal{S}-$dimensional integrals

- Gao and Song 2011 : use the pairwise likelihood as objective !

classic EM : $R(\theta, \theta^{(q)}) = E_{Z \sim p(;Y,\theta^{(q)})}(log(L(\theta, Z, Y))|Y^{obs})$
CLEM : $R_C(\theta, \theta^{(q)}) = \sum_{a \in A} E_{Z \sim p(;Y,\theta^{(q)})}(log(L(\theta, Z_a, Y_a))|Y_a^{obs})$

Potential computational improvement : Composite Likelihood EM (CLEM) algorithm

!! Issue in the EM : E-step still require $\mathcal{S}-$dimensional integrals

- Gao and Song 2011 : use the pairwise likelihood as objective !

classic EM : $R(\theta, \theta^{(q)}) = E_{Z\sim p(;Y,\theta^{(q)})}(log(L(\theta, Z, Y))|Y^{obs})$
CLEM : $R_C(\theta, \theta^{(q)}) = \sum_{a\in A} E_{Z\sim p(;Y,\theta^{(q)})}(log(L(\theta, Z_a, Y_a))|Y_a^{obs})$
For us : CLEM = having one forward-backward for each pair and to combine the probabilities during the M-step.

\+ no $\mathcal{S}$-dimensional integral

\- As many FB as pairs instead of just 1

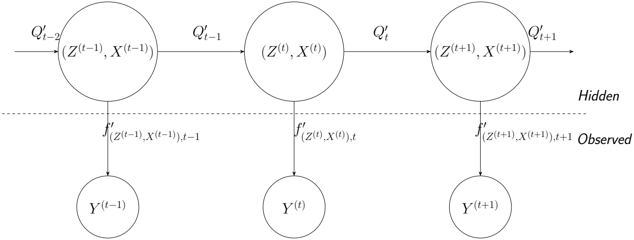
? Less efficiency

Alternative modelization

Unobserved : weather type and latent field $(Z^{(t)}, X^{(t)}) \in \{1, \ldots, K\} \times \mathbb{R}^{\mathcal{S}}$
Rain occurence $Y_s^{(t)} | (X_s^{(t)}, Z^{(t)}) \sim Bernoulli(\Phi(X_s^{(t)} + \sqrt{2}\Phi^{-1}(\lambda_{Z^{(t)},t,s})))$

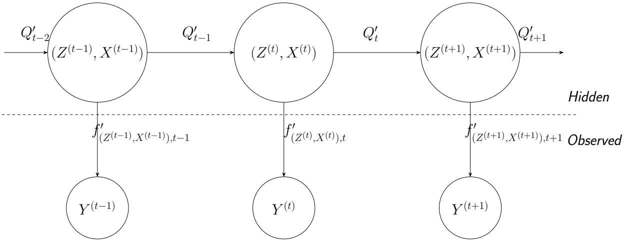
Alternative modelization

Unobserved : weather type and latent field $(Z^{(t)}, X^{(t)}) \in \{1, \dots, K\} \times \mathbb{R}^{\mathcal{S}}$
Rain occurence $Y_s^{(t)} | (X_s^{(t)}, Z^{(t)}) \sim Bernoulli(\Phi(X_s^{(t)} + \sqrt{2}\Phi^{-1}(\lambda_{Z^{(t)},t,s})))$



Alternative modelization

Unobserved : weather type and latent field $(Z^{(t)}, X^{(t)}) \in \{1, \dots, K\} \times \mathbb{R}^{\mathcal{S}}$
Rain occurence $Y_s^{(t)} | (X_s^{(t)}, Z^{(t)}) \sim Bernoulli(\Phi(X_s^{(t)} + \sqrt{2}\Phi^{-1}(\lambda_{Z^{(t)},t,s})))$



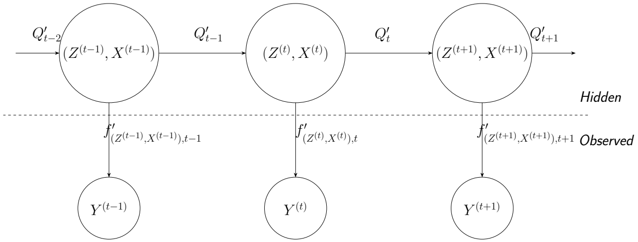
$$Q'_t((z^{(t+1)}, x^{(t+1)}), (z^{(t)}, x^{(t)})) = Q_t(z^{(t+1)}, z^{(t)}) \times f_{z^{(t+1)},t+1}(x^{(t+1)})$$

$$f'_{(z^{(t)},x^{(t)}),t}(y^{(t)}) = \prod_{s=1}^{\mathcal{S}} \left[y_s^{(t)} \Phi\big(\xi_{t,s}\big) + (1 - y_s^{(t)})\big(1 - \Phi\big(\xi_{t,s}\big)\big) \right],$$

where $\xi_{t,s} = x_s^{(t)} + \sqrt{2}\,\Phi^{-1}(\lambda_{z^{(t)},t,s})$

Alternative modelization

Unobserved : weather type and latent field $(Z^{(t)}, X^{(t)}) \in \{1, \dots, K\} \times \mathbb{R}^{\mathcal{S}}$
Rain occurence $Y_s^{(t)} | (X_s^{(t)}, Z^{(t)}) \sim Bernoulli(\Phi(X_s^{(t)} + \sqrt{2}\Phi^{-1}(\lambda_{Z^{(t)},t,s})))$



+ Same interpretation for the parameters
+ Much simpler M-step : conditional independence of stations
- No longer a discrete HMM : E-step bears all the difficulty

The EM algorithm -Dempster, Laird, and Rubin 1977

- Suited for latent models : observed $Y_1, ..., Y_N$, latent $Z_1, ..., Z_N$, parameter θ
- Objective : find $\hat{\theta} \in \mathrm{argmax}(\log(L_\theta(y)))$
- Algorithm :
  - θ_0 initial.
  - At each step q :
    - E step: Compute $R(\theta, \theta^{(q)}) = E_{Z \sim p(;Y, \theta^{(q)})}(log(L(\theta, Z, Y))|Y)$
    - M step : find $\theta^{(q+1)} \in \mathrm{argmax}\, R(\theta, \theta^{(q)})$

The EM algorithm -Dempster, Laird, and Rubin 1977

$$R(\theta, \theta^{(q)}) = \Sigma_{k=1}^{K} \Sigma_{t=1}^{n} \pi_{t|n}^{(q)}(k) \log(L(y_1^{(t)}, \dots, y_d^{(t)}; \theta_{t,k}))$$
$$+ \Sigma_{k=1}^{K} \pi_{1|n}^{(q)}(k) \log(\pi(k))$$
$$+ \Sigma_{k,l=1}^{K} \Sigma_{t=1}^{n-1} \pi_{t,t+1|n}^{(q)}(k,l) \log(Q_t(k,l))$$

E step : classic Forward-Backward Baum-Welch algorithm, to compute the $\pi_{t|n}^{(q)}(k)$, $\pi_{t,t+1|n}^{(q)}(k,l)$ where

$$\pi_{t|n}^{(q)}(k) = \mathbb{P}_{\theta^{(q)}} \left[Z^{(t)} = k | \left(Y^{(1)}, \dots, Y^{(n)} \right) = \left(y^{(1)}, \dots, y^{(n)} \right) \right]$$

$$\pi_{t,t+1|n}^{(q)}(k,\ell) = \mathbb{P}_{\theta^{(q)}} \left[Z^{(t)} = k, Z^{(t+1)} = \ell | \left(Y^{(1)}, \dots, Y^{(n)} \right) = \left(y^{(1)}, \dots, y^{(n)} \right) \right]$$

The EM algorithm -Dempster, Laird, and Rubin 1977

$$\begin{aligned}
\theta^{(q+1)} &= \text{argmax}\left(R(\theta, \theta^{(q)})\right) \\
&= \text{argmax}(\Sigma_{k=1}^{K}\Sigma_{t=1}^{n}\pi_{t|n}^{(q)}(k)\log(L(y_1^{(t)},\dots,y_d^{(t)};\theta_{t,k})) && = R_1 \\
&+ \Sigma_{k=1}^{K}\pi_{1|n}^{(q)}(k)\log(\pi(k)) && = R_2 \\
&+ \Sigma_{k,l=1}^{K}\Sigma_{t=1}^{n-1}\pi_{t,t+1|n}^{(q)}(k,l)\log(Q_t(k,l))) && = R_3
\end{aligned}$$

M step : maximise R_1, R_2, R_3 (parameters are independent)
Issue for R_1 : maximize a high-dimensional integral !

How to maximise a sum of type $\Sigma_{t=1}^{n} w_t \log(L(y_1^{(t)},...,y_d^{(t)};\theta_t))$?
$=$ weighted likelihood

The EM algorithm -Dempster, Laird, and Rubin 1977

$$
\begin{aligned}
\theta^{(q+1)} &= \text{argmax}\left(R(\theta,\theta^{(q)})\right) \\
&= \text{argmax}(\Sigma_{k=1}^{K}\Sigma_{t=1}^{n}\pi_{t|n}^{(q)}(k)\log(L(y_1^{(t)},\dots,y_d^{(t)};\theta_{t,k})) && = R_1 \\
&+ \Sigma_{k=1}^{K}\pi_{1|n}^{(q)}(k)\log(\pi(k)) && = R_2 \\
&+ \Sigma_{k,l=1}^{K}\Sigma_{t=1}^{n-1}\pi_{t,t+1|n}^{(q)}(k,l)\log(Q_t(k,l))) && = R_3
\end{aligned}
$$

M step : maximise R_1, R_2, R_3 (parameters are independent)
Issue for R_1 : maximize a high-dimensional integral !

How to maximise a sum of type $\Sigma_{t=1}^{n} w_t \log(L(y_1^{(t)}, ..., y_d^{(t)}; \theta_t))$?
$=$ weighted likelihood

Solution : composite likelihood

Pairwise likelihood : back to my case

- For many iid samples, maximizing $\Sigma_{t=1}^{n} \log(L(y_1^{(t)}, ..., y_d^{(t)}; \theta_t))$ or $\Sigma_{t=1}^{n} \Sigma_{i,j} w_{ij} \log(L(y_i^{(t)}, y_j^{(t)}; \theta_t))$ give same estimate.
- We have dependent data in time with weights from the E step
- Want to maximise $\Sigma_{t=1}^{n} w_t \log(L(y_1^{(t)}, ..., y_d^{(t)}; \theta_t))$
- Replace by $\Sigma_{t=1}^{n} \Sigma_{i,j} w_{ij} w_t \log(L(y_i^{(t)}, y_j^{(t)}; \theta_t))$
- Hope for the best.