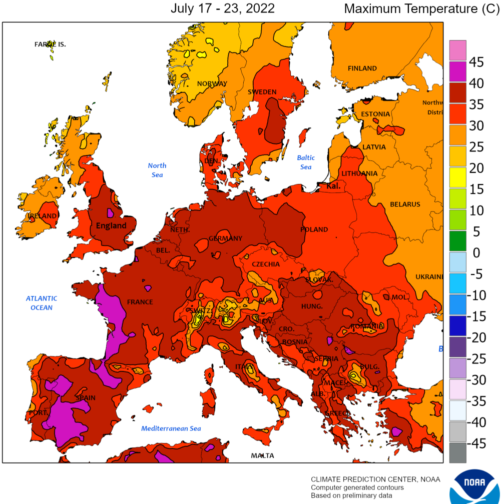
Spatio-temporal generation of precipitations using a spatially correlated Bernoulli and hidden Markov model

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

Contents

Climate change impact

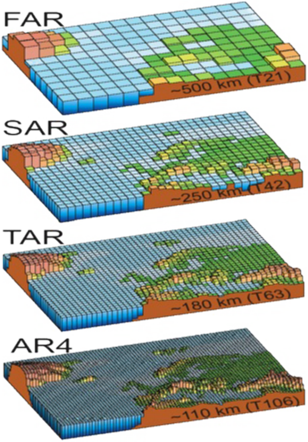
- Extremes and risks: impact on agriculture, health, energy...
- Climate change: impact on frequency, intensity,...



Goal: Estimate (future) risks quantitatively

$\rightarrow$ In particular large scales extremes

Climate (physical) models

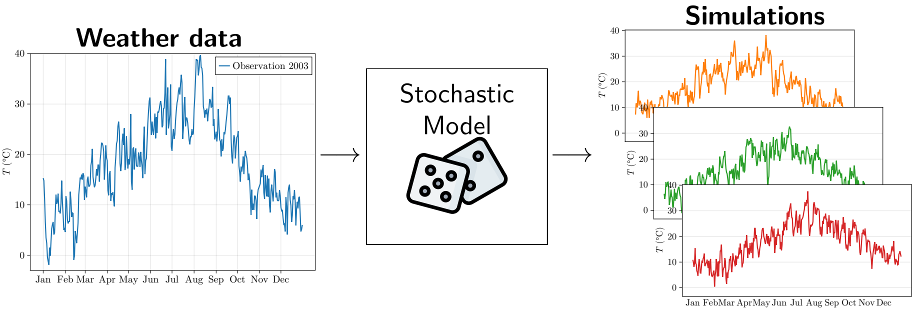


Climate models grid size

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

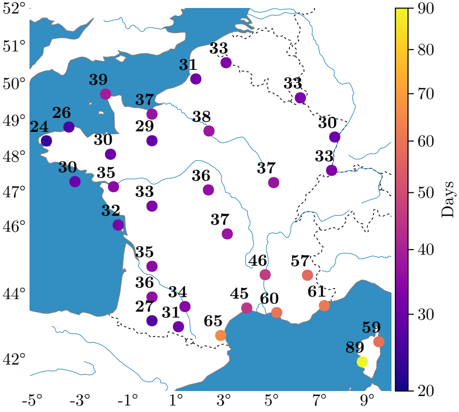
– Computationally expensive
– Not very good for extremes

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 $Y^{(t)} = (Y_1^{(t)}, \ldots, Y_S^{(t)})$

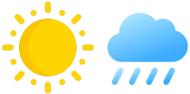
→ Reproduce the spatial-temporal structure of the data.

- 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 talk only rainfall occurrence

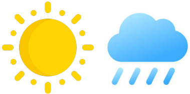
Single-site weather types model

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 :

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

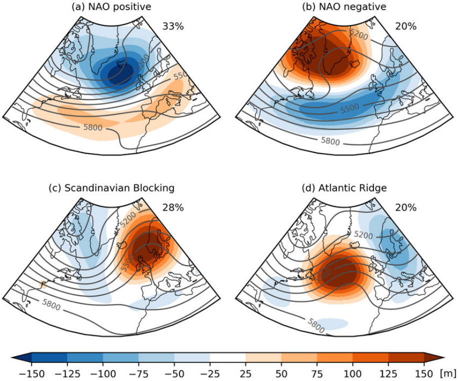
- Observed meteorological variables Y_t are generated conditionally on Z_t :

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

Separates the complexity of the model into several categories
Latent models are very common in Machine Learning

Spatial weather types

Weather types derived from atmospheric circulation or predefined

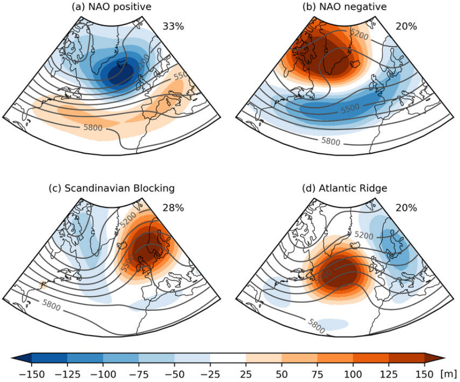


Atlantic weather types

→ Not data-centered

Spatial weather types

Weather types derived from atmospheric circulation or predefined

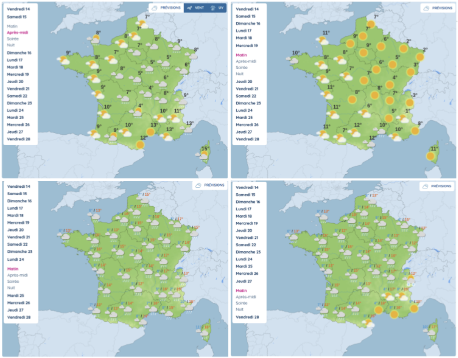


Atlantic weather types

→ Not data-centered

Hidden Markov chain

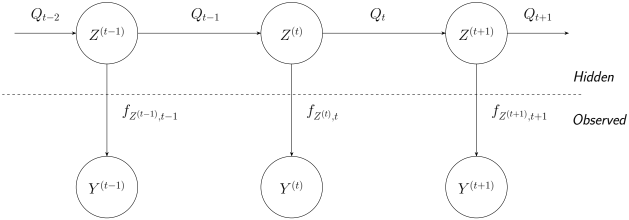
→ Weather types learned from the data



→ **This work**

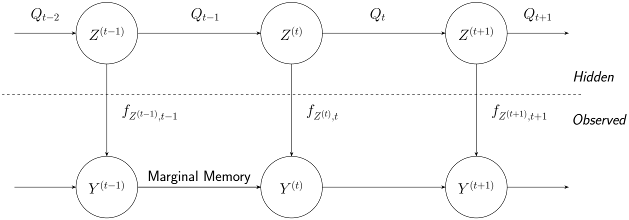
Multisite HMM based model

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 $Z^{(t)} \in \{1, \dots, K\}$

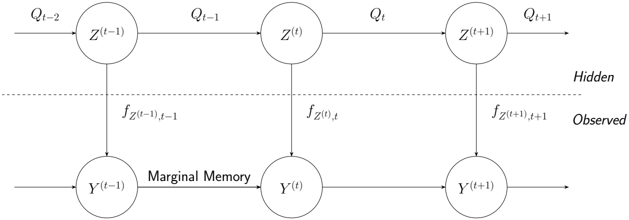


Conditional independence (Zucchini et al. 1991, Gobet et al. 2024)

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

Multisite HMM based model

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



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

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

– 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

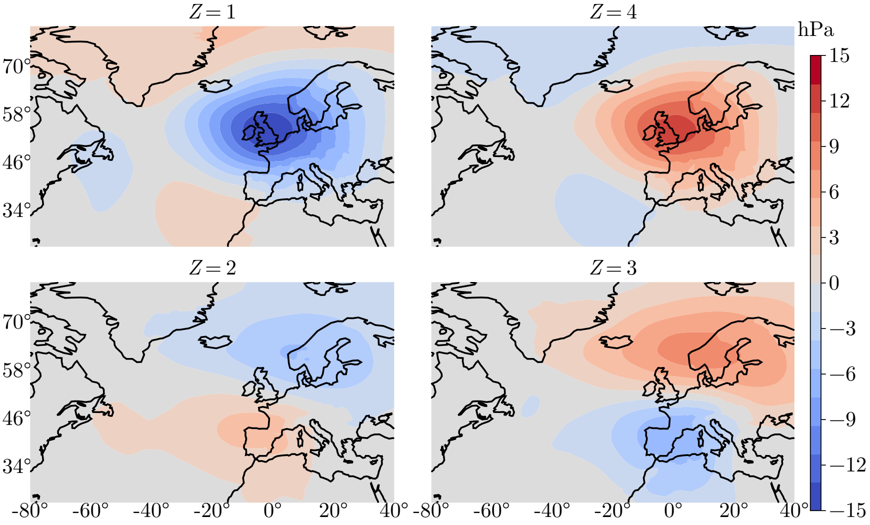


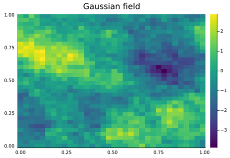
Figure: Pressure map $\Delta P = \mathbb{E}\left(P \mid Z = k\right) - \mathbb{E}\left(P\right)$ for winter months

Multisite Pair correlation model

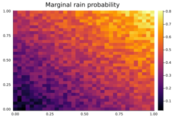
- At each station: Y_s follows a high order Markov chain $\mathbb{P}\left(Y_s^{(t)} \mid Y_s^{(t-1)}, \dots\right)$

Multisite Pair correlation model

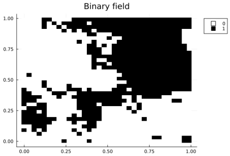
- At each station: Y_s follows a high order Markov chain $\mathbb{P}\left(Y_s^{(t)} \mid Y_s^{(t-1)}, \dots\right)$
- Multisite:



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



$$\xrightarrow{\lambda_s}$$



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

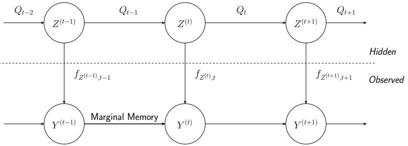
Each correlation coefficient of Σ has to be estimated

– A lot of parameters to estimate $O(\mathcal{S}^2)$ vs $O(K^2)$ for the HMM

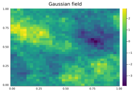
+ They can be estimated separately

+ Not restriction on the stations distance

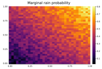
Multisite mixed model



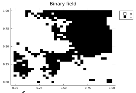
$+$



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

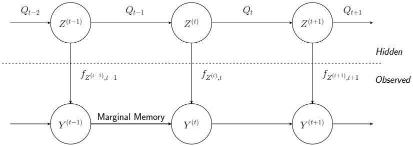


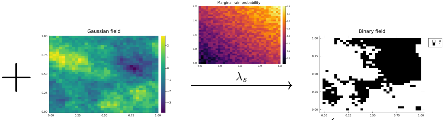
$\xrightarrow{\lambda_s}$



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

Multisite mixed model





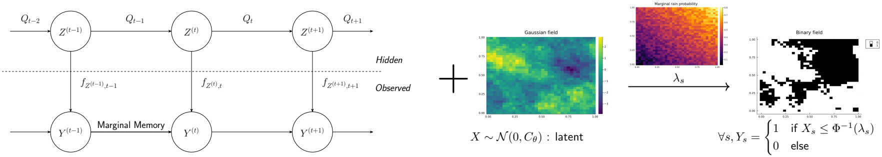
$X \sim \mathcal{N}(0, C_\theta)$: latent

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

Spatial conditional dependence

- Covariance function $\Sigma = C_\theta(h)$ of distance h between stations

Multisite mixed model



Spatial conditional dependence

- Covariance function $\Sigma = C_\theta(h)$ of distance h between stations
- $$\mathbb{P}\left(Y = y \mid Z = z^{(t)}\right) = \int_{a_1}^{b_1} \cdots \int_{a_S}^{b_S} f_X^{(\theta)}((x_1, ..., x_S))\, \mathrm{d}(x_1, ..., x_S)$$
  With $a_s = -\infty$ if $Y_s = 1, \Phi^{-1}(\lambda_s)$ else, $b_i = \infty$ if $Y_i = 0, \Phi^{-1}(\lambda_i)$ else
  $\rightarrow$ Looks like CDF of multivariate Gaussian...

+ Spatial structure $\rightarrow$ less parameters than $O(\mathcal{S}^2)$
+ Not restriction on the stations distance

− They **cannot** be estimated independently

± Correlations between stations are captured by the weather types Z and C_θ

Maximum likelihood estimation

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

!! High dimensional integrals

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 pairwise likelihood instead of full likelihood during the **M** step of the EM algorithm.

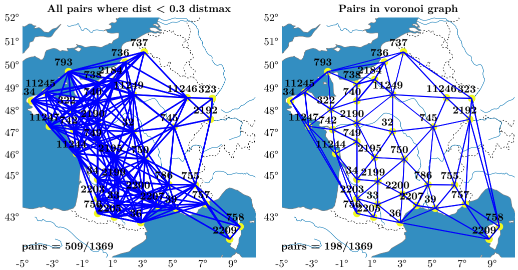
Maximum likelihood estimation

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

!! High dimensional integrals

$\rightarrow$ Tricks make the problem computable

Example: Maximize pairwise likelihood instead of full likelihood during the **M** step of the EM algorithm.

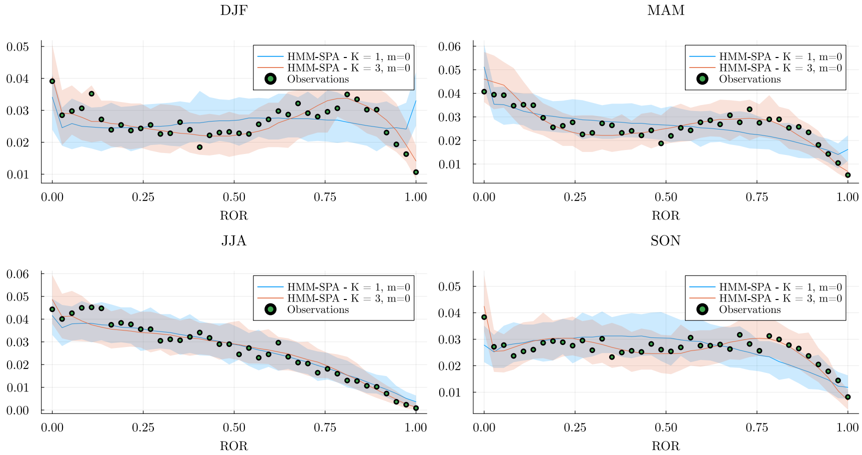


Spatiotemporal evaluation

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

Spatiotemporal evaluation

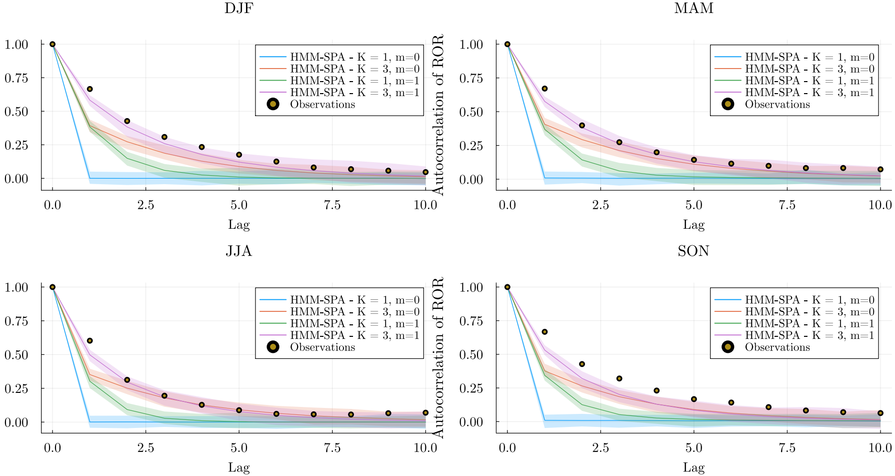
$$\text{Rain Occurrence Rate (ROR)} = \frac{\sum_{s\in\mathcal{S}} Y_s^{(t)}}{|\mathcal{S}|}$$



$K = 1$: no hidden states, only pairwise correlations

Autocorrelation

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



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
- Multivariate SWG with applications for agronomy

Julia package
`StochasticWeatherGenerators.jl`

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
- Multivariate SWG with applications for agronomy

Julia package
`StochasticWeatherGenerators.jl`

Thank you for your attention!

Parameterization : periodic 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

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

$$Q_t(k,l) = \frac{\exp(P_{c_{k,l}}(t))}{1+\sum_{l=1}^{K-1}\exp(P_{c_{k,l}}(t))} \text{ for } l < K,\ Q_t(k,K) = \frac{1}{1+\sum_{l=1}^{K-1}\exp(P_{c_{k,l}}$$

$$\lambda_{k,t,s,y_s^{(t-1)}} = \frac{1}{1+\sum_{l=1}^{K-1}\exp(P_{c_{k,s,y_s^{(t-1)}}}(t))}$$

$$\rho_{k,t} = \exp\left(P_{c_k(t)}\right)$$

Case rain probability= $\lambda_{k,t,s}$: no marginal local memory. Can replace by $\lambda_{k,t,s,y_s^{(t-1)}}$ = marginal local memory.

Seasonal large scales properties

- Seasonal effects
  Proposal : Periodic, time-varying parameters.
- Long, dry/wet episodes
- Spatially large episodes

Spatio-temporal indicator to evaluate models :

Rain Occurence Ratio

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

Temporal univariate series : distribution, autocorrelation function,...

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

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

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

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

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

The EM algorithm -Dempster, Laird, and Rubin 1977

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

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

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

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

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 $w_k \geq 0$ some 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 $w_k \geq 0$ some 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, $L_C(\theta; (y^1 ... y^n)) = \prod_{t=1}^{n} L(\theta; y^t)$

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

with G the Godambe matrix.

So, asymptotically, maximizing the composite likelihood gives the true parameter.

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.

Implementation

julia Choices : how to approximate integral of the multivariate normal distribution ?

- Dimension S : Quasi-Monte-Carlo, julia package MvNormalCDF `https://github.com/PharmCat/MvNormalCDF.jl`.
- Other options : approximations, unfortunately bad for highly correlated model.
- Dimension 2 (pairwise maximization) : Expression from Tsay and Ke 2023 - translated from `https://github.com/david-cortes/approxcdf` in Julia by David Métivier.

Implementation

julia Choices : how to approximate integral of the multivariate normal distribution ?

- Dimension S : Quasi-Monte-Carlo, julia package MvNormalCDF `https://github.com/PharmCat/MvNormalCDF.jl`.
- Other options : approximations, unfortunately bad for highly correlated model.
- Dimension 2 (pairwise maximization) : Expression from Tsay and Ke 2023 - translated from `https://github.com/david-cortes/approxcdf` in Julia by David Métivier.

If you have any clever idea for computing or maxizing integral

$$Pr(X_1 < x_1, \dots, X_d < x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_S} f_X((x'_1, ..., x'_S)) d(x'_1, ..., x'_S)$$

Please tell!