Hidden Markov Models for daily rainfall

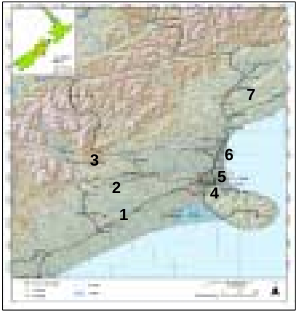
Pierre Ailliot, Université de Brest
Peter Thomson, Statistics Research Associates Ltd

Outline

Outline

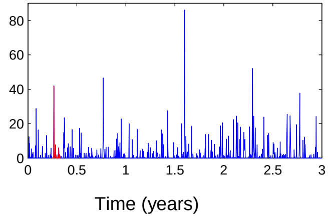
Rainfall data

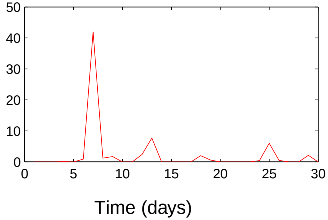
- Rainfall data in New Zealand
  - K=7 locations
  - 26 years
  - Daily rainfall
- $Y_t(k) \geq 0$: rainfall (mm) during day t at location k
- $Y_t = (Y_t(1), ..., Y_t(K))'$



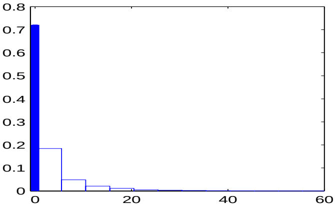
Characteristics of the data ?

- Marginal distribution (location 1)
  - Example of time series





  - Histogram

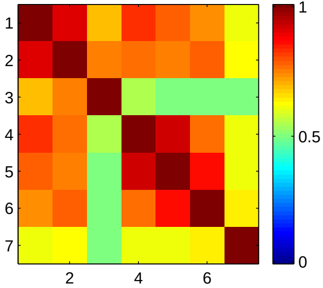


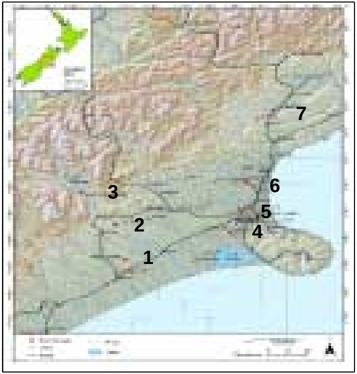
- Mixed variable
  - $Y_t(k) = 0$ if no rainfall
  - $Y_t(k) > 0$ otherwise

  Usual spatial or time series models are not appropriate !
- Heavy tails ($\kappa \approx 0.2$?)

Characteristics of the data ?

- Spatial correlation matrix

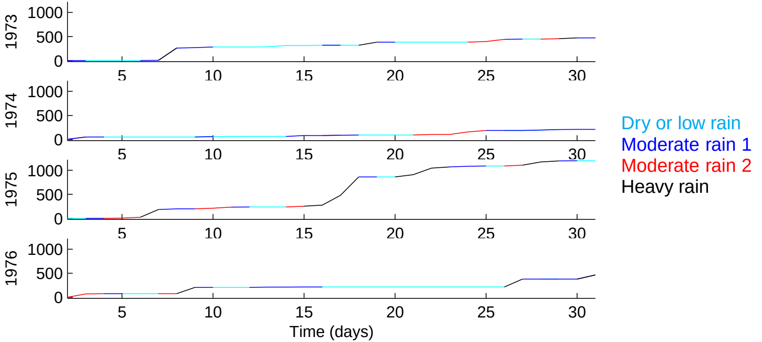




- $corr(Y_t(k), Y_t(l))$ depends on
  - the distance between location k and l
  - other covariates which create local effect and bloc structure

Characteristics of the data ?

- Temporal structure
  - Non-stationary components : seasonal, interannual(?)
    - Focus on April and neglect interannual components
  - Existence of different weather types (e.g. dry/frontal systems/convective rain...)



Accumulated rainfall (over space and time)

- Suggests segmenting the process and using different spatio-temporal models in each bloc

Outline

Model description

- *Zucchini and Guttorp (1991)*, *Bellone et al. (2000)*
- ‘Weather types” modelled as a hidden process $S_t \in \{1...M\}$
- Time structure : HMM

Weather type (hidden)
$p(s_t|s_1^{t-1}, y_1^{t-1}) = p(s_t|s_{t-1})$

$$\cdots \rightarrow S_{t-1} \rightarrow S_t \rightarrow S_{t+1} \rightarrow \cdots$$

$$\downarrow \quad \downarrow \quad \downarrow$$

Rainfall (observed)
$p(y_t|s_1^t, y_1^{t-1}) = p(y_t|s_t)$

$$\cdots \quad Y_{t-1} \quad Y_t \quad Y_{t+1} \quad \cdots$$

  - ...Dynamics induced only by $\{S_t\}$!
- Spatial structure : conditional independence

$$p(y_t|s_t) = p(y_t(1), ..., y_t(K)|s_t) = \prod_{k=1}^{K} p(y_t(k)|s_t)$$

  - ...Spatial dependence induced only by $\{S_t\}$!

$$p(y_t(k)|s_t) = \begin{cases} 1 - \pi_k^{(s_t)} & \textit{if } y_t(k) = 0 \\ \pi_k^{(s_t)} \gamma(y_t(k); \alpha_k^{(s_t)}, \beta_k^{(s_t)}) & \textit{if } y_t(k) > 0 \end{cases}$$

  - $0 \leq \pi_k^{(s)} \leq 1$, $\alpha_k^{(s)} > 0$, $\beta_k^{(s)} > 0$

Model description

- *Zucchini and Guttorp (1991)*, *Bellone et al. (2000)*
- ‘Weather types” modelled as a hidden process $S_t \in \{1...M\}$
- Time structure : HMM

Weather type (hidden)
$p(s_t|s_1^{t-1}, y_1^{t-1}) = p(s_t|s_{t-1})$

$$\begin{array}{ccccccccc} \cdots & \rightarrow & S_{t-1} & \rightarrow & S_t & \rightarrow & S_{t+1} & \rightarrow & \cdots \\ & & \downarrow & & \downarrow & & \downarrow & & \\ \cdots & & Y_{t-1} & & Y_t & & Y_{t+1} & & \cdots \end{array}$$

Rainfall (observed)
$p(y_t|s_1^t, y_1^{t-1}) = p(y_t|s_t)$

  - ...Dynamics induced only by $\{S_t\}$!
- Spatial structure : conditional independence
- Multiplicative model

$$Y_t(k) = L_t(k)A_t(k)$$

  - $(L_t(k))_k$,$(A_t(k))_k$ independent
  - $L_t(k) \sim Ber(\pi_k^{(S_t)})$
  - $A_t(k) \sim Gam(\alpha_k^{(S_t)}, \beta_k^{(S_t)})$

Parameter estimation

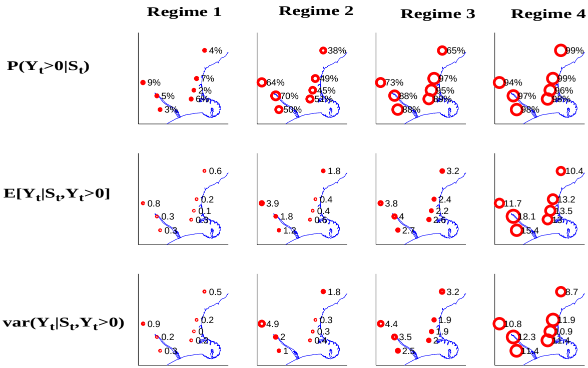
- Generalized EM algorithm
  - Numerical optimization in the M step ($K \times M$ 1D optimization)
- Model selection

M	1	2	3	4	5
BIC	17502	14523	13760	13663	13731

- Maximum likelihood estimates ($M = 4$)
  - Conditional distributions in the different regimes

Parameter estimation

- Conditional distributions in the different regimes



Parameter estimation

- Generalized EM algorithm
- Model selection

M	1	2	3	4	5
BIC	17502	14523	13760	13663	13731

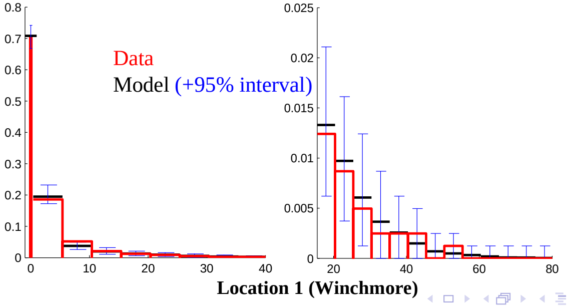
- Maximum likelihood estimates (M=4)
  - Conditional distributions in the different regimes
  - Transition matrix, stationary distribution and mean durations

	S_t					
S_{t-1}	1	2	3	4	$\tilde{\pi}_s$	D_s
1	0.62	0.23	0.10	0.05	0.37	2.62
2	0.38	0.44	0.15	0.03	0.35	1.80
3	0.00	0.32	0.41	0.27	0.16	1.70
4	0.06	0.54	0.00	0.40	0.12	1.65

- Summary
  - **Regime 1** : dry conditions, “long” persistence
  - **Regime 2 and 3** : intermediate patterns, regional differences, higher rainfall in regime 3, short persistence
  - **Regime 4** : heavy rainfall
- Similar meteorological interpretation for other datasets

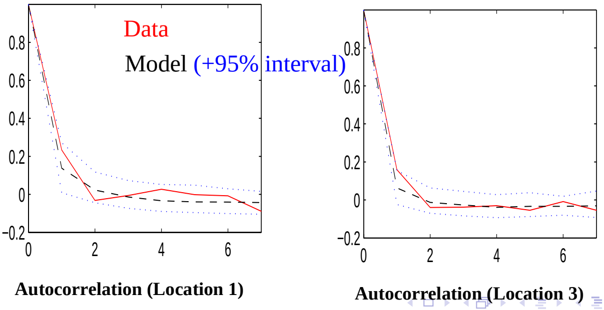
Model validation

- Motivation of this work : stochastic weather generator
  - Build models which can generate realistic weather scenarios
  - Estimate related risks (agriculture, energy production...) by simulation
- Realism of artificial sequences simulated with the model
  - Marginal distributions
    - Distributional versatility of HMM



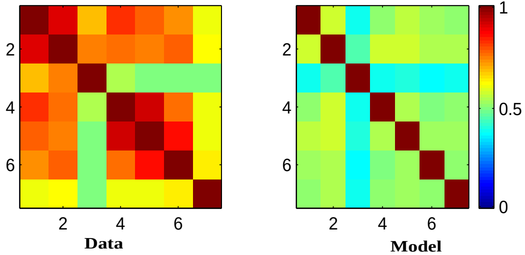
Model validation

- Motivation of this work : stochastic weather generator
- Realism of artificial sequences simulated with the model
  - Marginal distributions : ok
  - Dynamics at the different locations
    - Low correlation between successive observations



Model validation

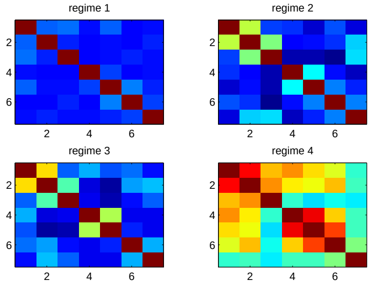
- Motivation of this work : stochastic weather generator
- Realism of artificial sequences simulated with the model
  - Marginal distributions : ok
  - Dynamics at the different locations : ok ?
  - Spatial structure : correlation underestimated



Model validation

- Motivation of this work : stochastic weather generator
- Realism of artificial sequences simulated with the model
  - Marginal distributions : ok
  - Dynamics at the different locations : ok ?
  - Spatial structure : correlation underestimated
- ... Need for a better model !
  - Existence of residual spatial structure within the weather types

Empirical correlation matrices in the different weather types (identified by the Viterbi algorithm)



Model validation

- Motivation of this work : stochastic weather generator
- Realism of artificial sequences simulated with the model
  - Marginal distributions : ok
  - Dynamics at the different locations : ok ?
  - Spatial structure : correlation underestimated
- ... Need for a better model !
  - Existence of residual spatial structure within the weather types
- Introduce spatial structure in the emission probabilities $P(Y_t|S_t = s_t)$
  - Need spatial model for mixed discrete-continuous variables
  - A first model : censored Gaussian random fields
    Ailliot P., Thompson C., Thomson P., (2009), Space time modeling of precipitation using a hidden Markov model and censored Gaussian distributions, Journal of the Royal Statistical Society, Series C (Applied Statistics). Vol. 58, no3, pp. 405-426.

Outline

Model description

Hidden weather type
Markov chain

$$\dots \rightarrow S_{t-1} \rightarrow S_t \rightarrow S_{t+1} \rightarrow \dots$$

$$\downarrow \quad \downarrow \quad \downarrow$$

Partially observed Gaussian RV
$P[W_t|S_t = s_t] \sim \mathcal{N}\left(m^{(s_t)}, \Sigma^{(s_t)}\right)$

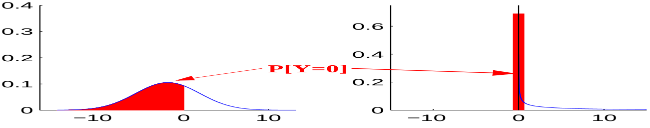
$$\dots \quad W_{t-1} \quad W_t \quad W_{t+1} \quad \dots$$

$$\downarrow \quad \downarrow \quad \downarrow$$

Observed precipitation

$$Y_t(k) = \begin{cases} 0 & \text{if } W_t(k) \leq 0 \\ W_t(k)^{\beta^{(s)}(k)} & \text{if } W_t(k) > 0 \end{cases}$$

$$\dots \quad Y_{t-1} \quad Y_t \quad Y_{t+1} \quad \dots$$



pdf of $W \sim N(-1.88, 3.78)$ pdf of $= max(W, 0)^{1.81}$

Model description

- Spatial information can be included in the covariance matrices
  - **Model C0 :** $\Sigma^{(s)}(i,i) = (\sigma_i^{(s)})^2$
  - **Model C1 :** $\Sigma^{(s)}(i,j) = \sigma_i^{(s)}\sigma_j^{(s)} \exp(-\lambda^{(s)} d(z_i, z_j))$
  - **Model C2 :** $\Sigma^{(s)}(i,j) = \sigma_i^{(s)}\sigma_j^{(s)} \kappa(\lambda_i^{(s)}, \lambda_j^{(s)}) \exp(-\kappa(\lambda_i^{(s)}, \lambda_j^{(s)})\sqrt{\lambda_i^{(s)}\lambda_j^{(s)}} d(z_i, z_j))$
    with $\kappa^2(x,y) = 2\sqrt{x^2y^2}/(x^2+y^2)$

	AIC					BIC				
M	1	2	3	4	5	1	2	3	4	5
C0	17403	14445	13639	13398	13289	17501	14651	13963	13849	13875
Cγ	17404	14317	13436	13213	13144	17502	14523	13760	13663	13731
C1	13092	12770	12697	12616	12623	13196	**12985**	13035	13085	13233
C2	12995	12741	12600	**12506**	12509	13127	13013	13022	13089	13260
C$*$	12904	12643	12640	12674	12611	13101	13046	13259	13519	13690

Parameter estimation

- Monte Carlo EM algorithm
- Need to compute the following smoothing probabilities for the M-step
  - W_t^- : vector of censored components (dry locations) at time t

$$\gamma_t(s) = p(S_t = s | y_1^T; \hat{\theta}_n), \qquad \gamma_t(s,s') = p(S_{t-1} = s, S_t = s' | y_1^T; \hat{\theta}_n) \tag{1}$$

$$E(W_t^- | S_t = s, y_t; \hat{\theta}_n), \qquad E(W_t^- (W_t^-)' | S_t = s, y_t; \hat{\theta}_n) \tag{2}$$

- Several algorithms can be used in the E-step
  - Generic algorithms can be used : Gibbs sampler, particle filter,...
  - More efficient to use the specific structure of the model
    - Computing (2) requires computing integrals of the form (if $W_t^- = (W_t(1), \ldots, W_t(d))$)

$$\int_{-\infty}^{0} \ldots \int_{-\infty}^{0} w(k)\phi(w; m^{(s)}, \Sigma^{(s)})dw(1)\ldots dw(d) \tag{3}$$

$$\int_{-\infty}^{0} \ldots \int_{-\infty}^{0} w(k)w(k')\phi(w; m^{(s)}, \Sigma^{(s)})dw(1)\ldots dw(d) \tag{4}$$

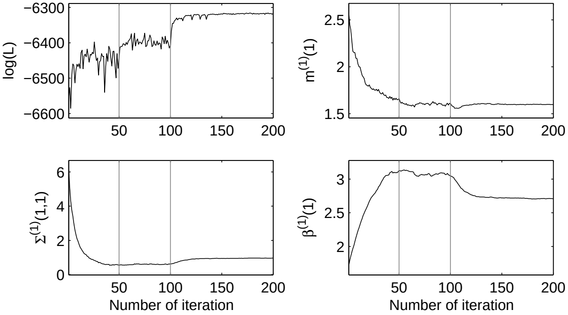
    - Emission probabilities $p(y_t|s_t)$ depend on integrals of the form

$$\int_{-\infty}^{0} \ldots \int_{-\infty}^{0} \phi(w; m^{(s)}, \Sigma^{(s)})dw(1)\ldots dw(d) \tag{5}$$

    - Monte-Carlo integration for (3),(4) and (5) and forward-backward algorithm for (1)

Parameter estimation

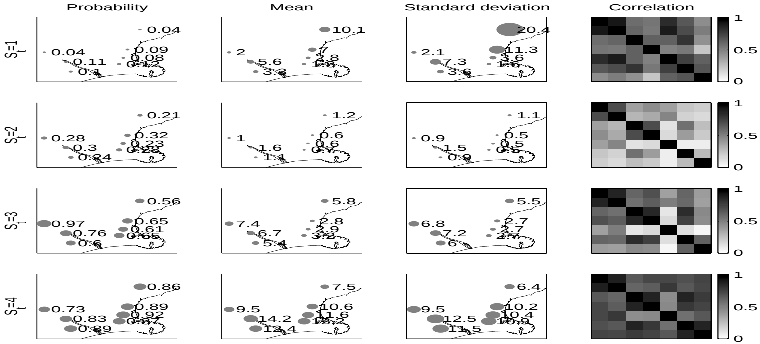
- Sample size for the Monte-Carlo approximations increases progressively
  - 100 for iterations $n \leq 50$
  - 500 for $50 < n \leq 100$
  - n^2 for $n > 100$



- CPU time : 140 minutes when $M = 4$

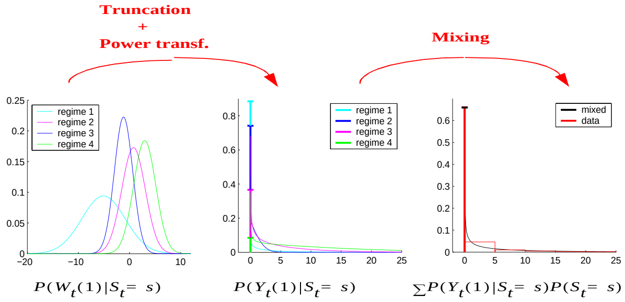
Parameter estimation

- Conditional distributions in the different regimes
  - Similar interpretation that for the previous model !



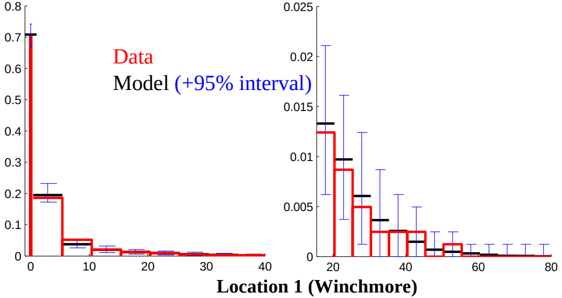
Model validation

- Realism of artificial sequences simulated with the model
  - Marginal distributions



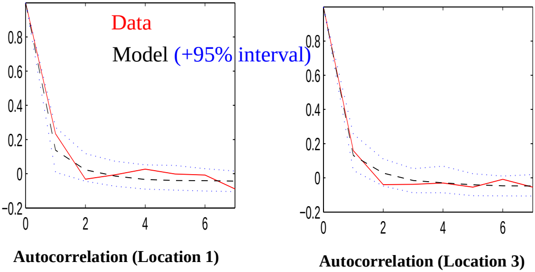
Model validation

- Realism of artificial sequences simulated with the model
  - Marginal distributions : ok



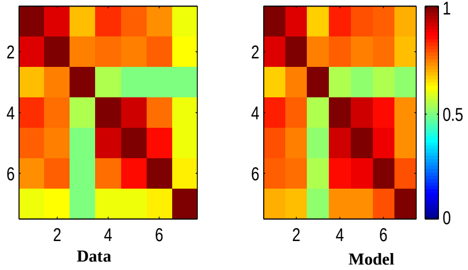
Model validation

- Realism of artificial sequences simulated with the model
  - Marginal distributions : ok
  - Dynamics at the different locations : ok ?



Model validation

- Realism of artificial sequences simulated with the model
  - Marginal distributions : ok
  - Dynamics at the different locations : ok ?
  - Spatial structure : ok



Model validation

- Realism of artificial sequences simulated with the model
  - Marginal distributions : ok
  - Dynamics at the different locations : ok ?
  - Spatial structure : ok
- Limitations ?
  - Heavy computation (MCEM)
    - Problematic for networks with more rainfall stations
    - CPU time increases with the number of dry days/locations !
    - We would like to make it even more complicated to include more dynamics !
  - Physical explanation for the censoring is missing
- Look for another spatial model for mixed variable such that
  - Quicker EM recursions
    - E-step : avoid Monte-Carlo simulations
    - M-step : allow numerical optimization
  - Flexibility
    - Realistic uni/multivariate distribution (margins with heavy tail ?)
    - Correlation with block structure and possibility to include spatial information
  - Interpretability
    - Structural/hierarchical model

Outline

Model description : general structure

Hidden regional weather type $S_t \in \{1, ..., M\}$	$\dots$	$\rightarrow$	S_{t-1} $\downarrow$	$\rightarrow$	S_t $\downarrow$	$\rightarrow$	S_{t+1} $\downarrow$	$\rightarrow$ $\dots$
Hidden regional rainfall index $I_t > 0$	$\dots$		I_{t-1} $\swarrow \searrow$		I_t $\swarrow \searrow$		I_{t+1} $\swarrow \searrow$	$\dots$
Local occurrence/amount $L_t(k) \in \{0,1\}$ and $A_t(k) > 0$	$\dots$		L_{t-1} A_{t-1} $\searrow \swarrow$		L_t A_t $\searrow \swarrow$		L_{t+1} A_{t+1} $\searrow \swarrow$	$\dots$
Observed local rainfall $Y_t(k) = L_t(k)A_t(k)$	$\dots$		Y_{t-1}		Y_t		Y_{t+1}	$\dots$

- $I_t > 0$ is supposed to summarize what governs rainfall at the regional scale
  - Both probability rainfall and amount expected to increase at each location with I
  - Common to all locations : realistic only for small scale networks ?
- "Downscaling" ($I_t \rightarrow L_t$ and $I_t \rightarrow A_t$) models local effects
- I_t creates dependence between A_t and L_t
- Can we find parametrizations such that $p(y_t|s_t)$ is analytical ?

Model for the positive field A_t (amount)

- Conditional Inverse Gamma distribution for the regional index

$$P(1/I_t|S_t = s_t) \sim Gam(\gamma^{(s_t)}, \delta^{(s_t)})$$

  - Technical result : if $1/I \sim Gam(\gamma, \delta)$ then $E\left[I^{-\alpha} exp\left(-\frac{\beta}{I}\right)\right] = \frac{\Gamma(\alpha+\gamma)}{\delta^{\alpha}\Gamma(\gamma)}(1+\frac{\beta}{\delta})^{-\alpha-\gamma}$
  - Useful to compound with Gamma distributions
  - Permit to integrate the effect of I_t and avoid Monte Carlo simulations

- Conditional independent Gamma distribution for the positive amounts

$$p(a_t(1), .., a_t(K)|i_t, s_t) = \prod_{k=1}^{K} p(a_t(k)|i_t, s_t)$$

$$P(A_t(k)|I_t = i_t, S_t = s_t) \sim Gam\left(\alpha_k^{(s_t)}, \beta_k^{(s_t)} i_t\right)$$

- Can be written as a multiplicative model

$$A_t(k) = I_t J_t(k)$$

  - $J_t(k) \sim Gam(\alpha_k^{(S_t)}, \beta_k^{(S_t)})$ (mutually independent and independent of I_t)
  - I_t represents the regional effect and J_t the local effects
  - Model with independent Gamma distributions is a limit case ($\gamma^{(s_t)} = \frac{1}{\delta^{(s_t)}}$ and $\delta^{(s_t)} \rightarrow 0$)
  - Identifiability constraint : $\delta^{(s_t)} = 1$

Model for the positive field A_t (amount)

Properties of the model

- Joint pdf can be integrated analytically over I_t (required for quick E-step)

$$p(a_t(1),\ldots,a_t(K)|s_t) = \int p(a_t(1),\ldots,a_t(K)|i_t,s_t)p(i_t|s_t)di_t$$

$$= \frac{\Gamma(\gamma^{(s_t)} + \sum_{k=1}^K \alpha_k^{(s_t)})}{\Gamma(\gamma^{(s_t)}) \prod_{k=1}^K \Gamma(\alpha_k^{(s_t)}) \prod_{k=1}^K \beta_k^{(s_t)}} \frac{\prod_{k=1}^K \left(\frac{a_t(k)}{\beta_k^{(s_t)}}\right)^{\alpha_k^{(s_t)}-1}}{\left(1+\sum_{k=1}^K \frac{a_t(k)}{\beta_k^{(s_t)}}\right)^{\gamma^{(s_t)}+\sum_{k=1}^K \alpha_k^{(s_t)}}}$$

- Marginal pdf : beta distribution of the second kind

$$p(a_t(k)|s_t) = \frac{1}{B(\gamma^{(s_t)},\alpha_k^{(s_t)})\beta_k^{(s_t)}} \frac{\left(\frac{a_t(k)}{\beta_k^{(s_t)}}\right)^{\alpha_k^{(s_t)}-1}}{\left(1+\frac{a_k}{\beta_k^{(s_t)}}\right)^{\gamma^{(s_t)}+\alpha_k^{(s_t)}}}$$

  - Gamma distribution as a limit case
  - Heavy tail : $E[A_t(k)^p|s_t] < +\infty$ iif $p < \gamma^{(s_t)}$

Model for the positive field A_t (amount)

Properties of the model

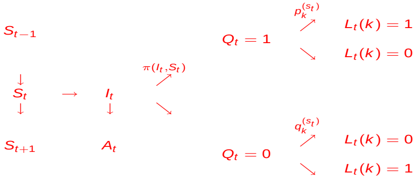
- Joint pdf can be integrated analytically over I_t (required to avoid MCEM)
- Marginal pdf : beta distribution of the second kind (heavy tail)
- Correlation matrix ($\gamma^{(s_t)} > 2$ and $k \neq l$) :

$$corr(A_t(k), A_t(l)|s_t) = \left(1 + \frac{\gamma^{(s_t)}-1}{\alpha_k^{(s_t)}}\right)^{-1/2} \left(1 + \frac{\gamma^{(s_t)}-1}{\alpha_l^{(s_t)}}\right)^{-1/2}$$

  - The spatial dependence comes from the regional index I_t
  - $corr(A_t(k), I_t) =$
    $$\left(\frac{\gamma^{(s_t)}+1}{\alpha_k^{(s_t)}}\right)^{-1/2} \begin{array}{lll} \nearrow 0 & \text{if } \gamma^{(s_t)}/\alpha_k^{(s_t)} \to +\infty & \text{Local dominates} \\ \searrow 1 & \text{if } \gamma^{(s_t)}/\alpha_k^{(s_t)} \to 0 & \text{Regional dominates} \end{array}$$
  - $corr(A_t(k), A_t(l)) \approx 1$ if regional conditions dominates at location k AND l
  - $corr(A_t(k), A_t(l)) \approx 0$ if local conditions dominates at location k OR l
  - Possible to get a correlation matrix with positive coefficients and one bloc of strongly correlated locations
  - We also would like to include geographic information (distance,...)...

Model for the binary field L_t (occurrence)

- Introduce a "regional occurrence" process $Q_t \in \{0,1\}$
  - $Q_t = 0$: mainly dry at the regional scale
  - $Q_t = 1$: mainly wet at the regional scale



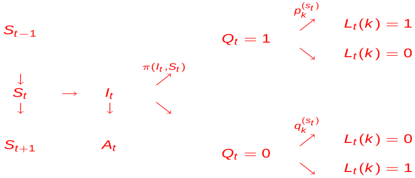
- Joint distribution : analytic expression
  - If $\pi(i_t, s_t) = P[Q_t = 1|s_t, i_t] = \exp\left(-\frac{\theta^{(s_t)}}{i_t} - \phi^{(s_t)}\right)$
  - then $P[Q_t = 1|s_t] = \int P[Q_t = 1|s_t, i_t]p(i_t|s_t)di_t = \frac{exp(-\phi^{(s_t)})}{(1+\theta^{(s_t)})\gamma^{(s_t)}}$
  - and we get analytic expression for the joint distribution (integrated over I)

$$p(l_t(1), \ldots, l_t(K)|s_t) = P[Q_t = 0|s_t] \prod_{k=1}^{K} \left(q_k^{(s_t)}\right)^{1-l_t(k)} \left(1 - q_k^{(s_t)}\right)^{l_t(k)}$$

$$+ P[Q_t = 1|s_t] \prod_{k=1}^{K} \left(p_k^{(s_t)}\right)^{l_t(k)} \left(1 - p_k^{(s_t)}\right)^{1-l_t(k)}$$

Model for the binary field L_t (occurrence)

- Introduce a "regional occurrence" process $Q_t \in \{0, 1\}$



- Joint distribution : analytic expression
- Special cases :
  - If $p_k^{(s_t)} = 1 - q_k^{(s_t)}$ then $L_t(k)$ is independent of Q_t and $L_t(l)$ for $l \neq k$
  - If $p_k^{(s_t)} = q_k^{(s_t)} = 1$ then $L_t(k) = Q_t$: allow one bloc of strongly correlated locations
- We also would like to include geographic information (distance,...)...

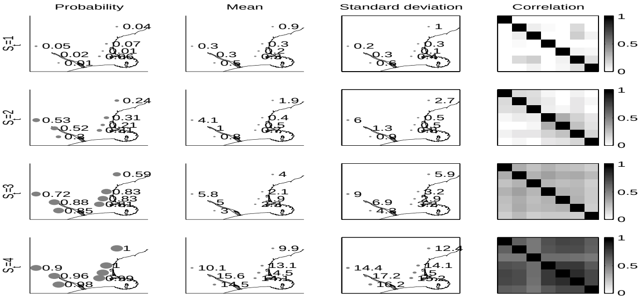
Parameter estimation

- Generalized EM algorithm
  - E-step : usual forward-backward algorithm
    - Analytical expressions for $p(y_t|s_t)$
  - M-step : numerical optimization (M optimizations in $4K+3$-dimensional spaces)
    - Quasi-Newton with increasing accuracy
    - May be improved ? Need to look more precisely at $Q\left(\theta, \theta^{(k)}\right)$
  - Starting point : model with independent Gamma distributions
  - CPU time : 40 minutes when $M = 4$
- Model selection

	BIC				
M	1	2	3	4	5
Independent	17502	14523	13760	**13663**	13731
Regional index	13775	13372	**13279**	13439	13587

Parameter estimation

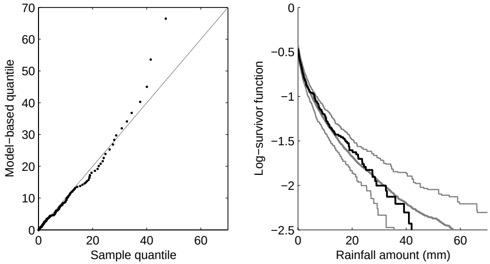
- Maximum likelihood estimates ($M = 4$)
  - Similar interpretation that for the previous models



  - Heavy tail distributions : $\hat{\gamma}^{(1)} = 3.54$, $\hat{\gamma}^{(4)} = 2.35$

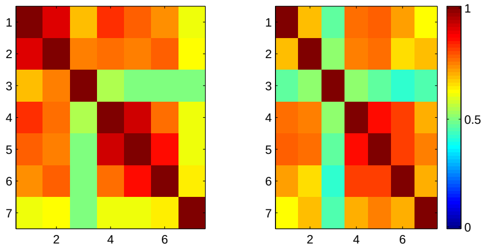
Model validation

- Realism of artificial sequences simulated with the model
  - Marginal distributions : ok, check tail



Model validation

- Realism of artificial sequences simulated with the model
  - Marginal distributions : ok
  - Dynamics at the different locations : ok ?
  - Spatial structure : ok



Outline

Conclusion

- HMMs provide a flexible framework for modelling meteorological processes
- Spatial model for mixed variables are needed when modelling daily rainfall
- Censored Gaussian distributions permit a good description of rainfall properties but lead to heavy computation
- Compounding Gamma/Inverse Gamma distributions leads to a tractable model
  - Seems to be able to reproduce the properties of the data at a regional scale
  - Perspectives
    - Include dynamics in the regimes
    - Combine different regional models to model rainfall at a bigger spatial scale