Modelling of the leaf appearance process or *phyllochron* with interval censored measurements

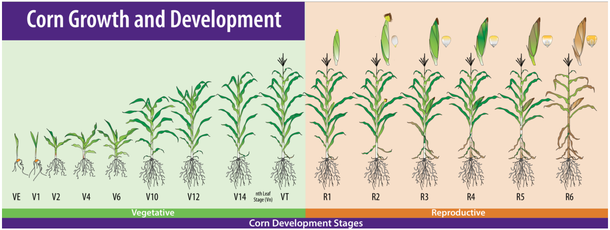
Sandra Plancade[(1)], Elodie Marchadier[(2,3)], Sylvie Huet[(1)], Adrienne Ressayre[(2)], Christine Dillmann[(2,3)]

(1) Unité MaIAGE, INRA, (2) Génétique Quantitative et Évolution - Le Moulon INRA, Université Paris-Sud, CNRS, AgroParisTech

5 december 2019

Plant growth and development

- Growth and development : synchronized processes

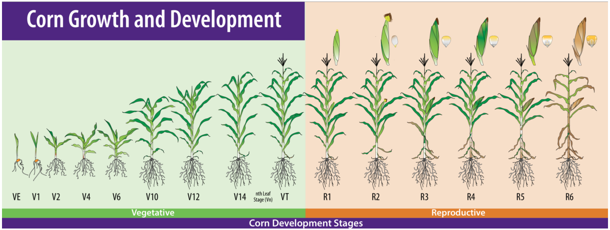


I. A. Ciampitti *et al*,

Kansas State University (2013)

Plant growth and development

- Growth and development : synchronized processes



I. A. Ciampitti *et al*,

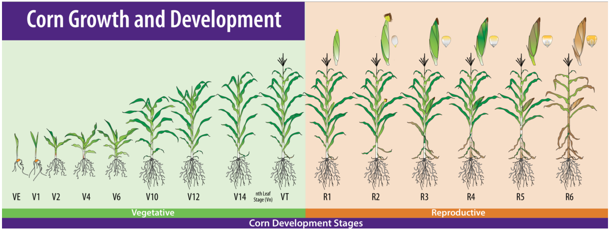
Kansas State University (2013)

- Modelling at various scales
  - Plot/field level
  - Plant level
  - Organ level
  - Cellular level

↪ Stochastic and deterministic model

Plant growth and development

- Growth and development : synchronized processes



I. A. Ciampitti *et al*,

Kansas State University (2013)

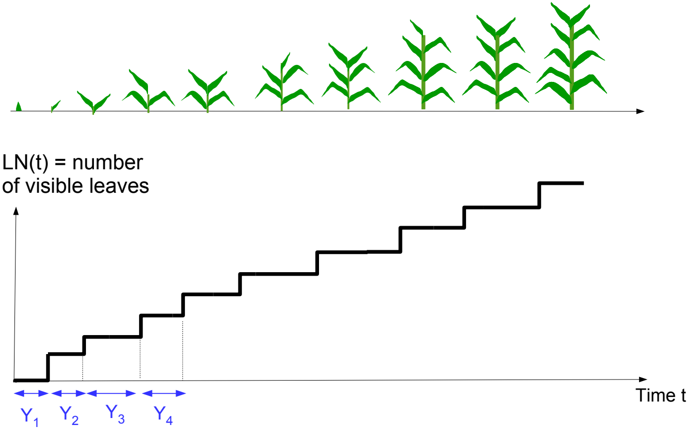
- Modelling at various scales
  - Plot/field level
  - **Plant level**
  - Organ level
  - Cellular level

↪ **Stochastic** and deterministic model

↪ Our model : **single phenotype**

Process of leaf appearance or *phyllochron*

- Phyllochron : times of leaf appearance on the plant/ interval of time between successive leaves



Phyllochron

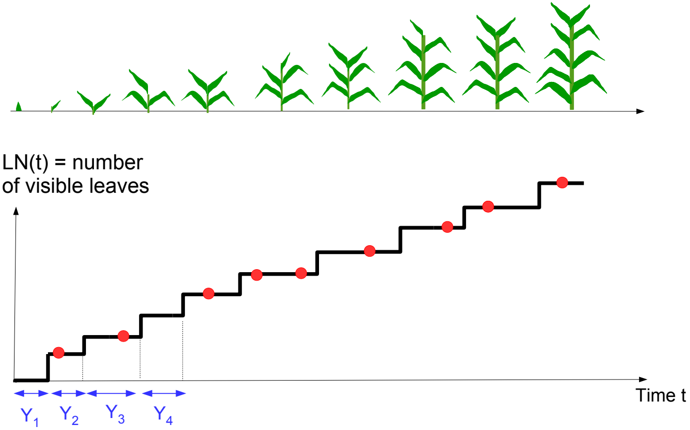
► Good indicator of plant development

► Easy measurements

► Non destructive observations

Process of leaf appearance or *phyllochron*

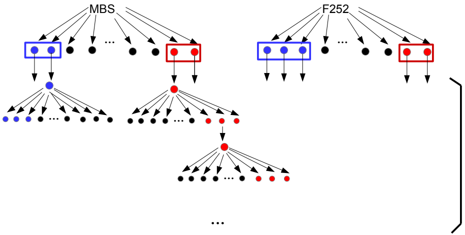
- Phyllochron : times of leaf appearance on the plant/ interval of time between successive leaves



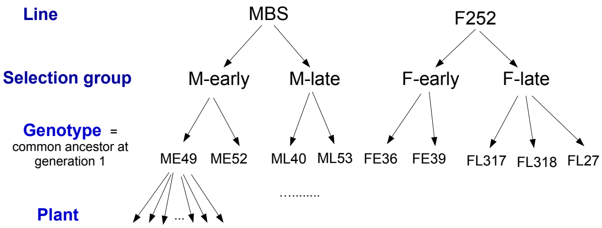
Phyllochron

- ► Good indicator of plant development
- ► Easy measurements
- ► Non destructive observations

- Divergent selection of early/late flowering plants



⇒ hierarchical grouping



Thermal time (TT) approach

- Idea : main driving factor of phyllochron is temperature

Thermal time (TT) approach

- Idea : main driving factor of phyllochron is temperature
- TT = accumulation of temperature weighted by temperature efficiency
  ↪ time rescaling

Thermal time (TT) approach

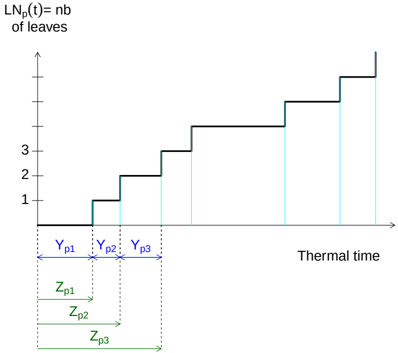
- Idea : main driving factor of phyllochron is temperature
- TT = accumulation of temperature weighted by temperature efficiency
  $\hookrightarrow$ time rescaling
- Linear phyllochron : $\text{Leaf Number}_p(t) = \alpha_p TT_t + \varepsilon_{p,t}$
  - Inference of α_p by linear regression
  - F-test/mixed model on (α_p) to test genotype/environment effect
  - Linear asssumption : questionable

Thermal time (TT) approach

- Idea : main driving factor of phyllochron is temperature
- TT = accumulation of temperature weighted by temperature efficiency
  $\hookrightarrow$ time rescaling
- Linear phyllochron : Leaf $\text{Number}_p(t) = \alpha_p TT_t + \varepsilon_{p,t}$
  - Inference of α_p by linear regression
  - F-test/mixed model on (α_p) to test genotype/environment effect
  - Linear asssumption : questionable
- More flexible models : $LN_p(t) = f(TT_t) + \varepsilon_{p,t}, \quad \varepsilon_{p,t} \quad i.i.d.$
  - bi/tri-linear, splines
  - Descriptive analysis
  - Statistical analysis (confidence interval, tests)
    $\hookrightarrow$ auto-correlation $\Rightarrow$ bias

Independent waiting times model

Consider a given genotype and a given year.

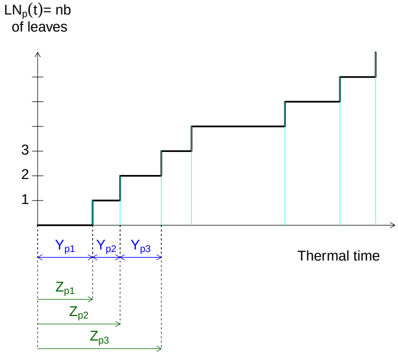


- Realistic biological assumption : $(Y_{p,f})_{f=1,\dots,F}$ independent

$$\begin{cases} \{LN_p(t)\}_{t>0} & \text{number of leaves} \\ (Z_{p,f})_f & \text{Times of leaf appearance} \\ (Y_{p,f})_f & \text{Interval between leaves} \end{cases}$$

Independent waiting times model

Consider a given genotype and a given year.



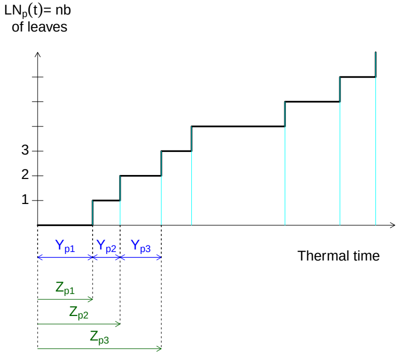
- Realistic biological assumption : $(Y_{p,f})_{f=1,\ldots,F}$ independent
- Flexible modelling :

$$Y_{p,f} = \mu_f + \varepsilon_{p,f}$$

$$\left\{ \begin{array}{ll} \{LN_p(t)\}_{t>0} & \text{number of leaves} \\ (Z_{p,f})_f & \text{Times of leaf appearance} \\ (Y_{p,f})_f & \text{Interval between leaves} \end{array} \right.$$

Independent waiting times model

Consider a given genotype and a given year.



- Realistic biological assumption : $(Y_{p,f})_{f=1,\dots,F}$ independent
- Flexible modelling :

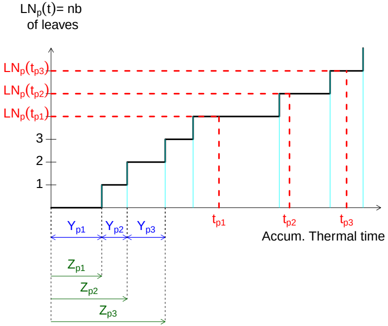
$$Y_{p,f} = \mu_f + \varepsilon_{p,f}$$

- More generally

$$Y_{p,f} \sim \mathcal{D}(\theta_f)$$

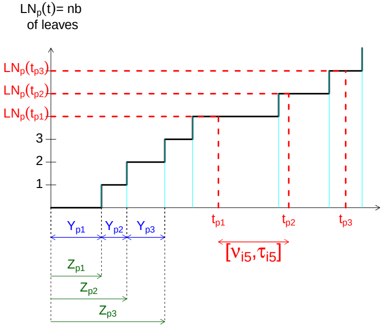
$$\begin{cases} \{LN_p(t)\}_{t>0} & \text{number of leaves} \\ (Z_{p,f})_f & \text{Times of leaf appearance} \\ (Y_{p,f})_f & \text{Interval between leaves} \end{cases}$$

Interval censored observations



- Observations $(LN_p(t_{p,1}), \ldots, LN_p(t_{p,N_p}))$

Interval censored observations



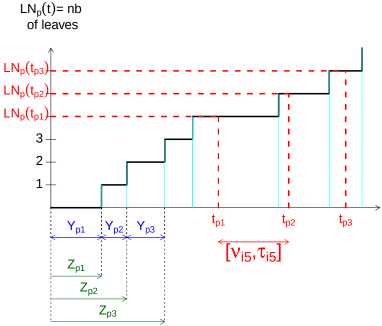
- Observations $(LN_p(t_{p,1}), \ldots, LN_p(t_{p,N_p}))$

  $\Leftrightarrow Z_{p,f} \in (\nu_{p,f}, \tau_{p,f}],\ \ f = 1, \ldots, F_p$

  with

  - $\nu_{p,f}$ = last observation time before appearance of leaf f (or 0)
  - $\tau_{p,f}$ = first observation time after appearance of leaf f (or $+\infty$)

Interval censored observations



- Observations $(LN_p(t_{p,1}), \dots, LN_p(t_{p,N_p}))$

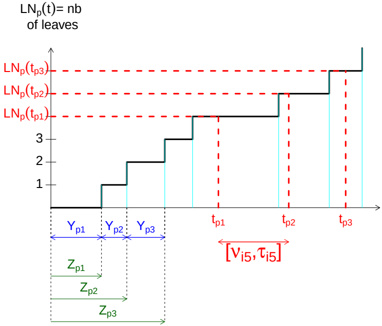
  $\Leftrightarrow Z_{p,f} \in (\nu_{p,f}, \tau_{p,f}], \ f = 1, \dots, F_p$

  with

  - $\nu_{p,f}$ = last observation time before appearance of leaf f (or 0)
  - $\tau_{p,f}$ = first observation time after appearance of leaf f (or $+\infty$)

- Rk : Observations $\not\Leftrightarrow Y_{p,f} \in (\nu'_{p,f}, \tau'_{p,f}]$

Interval censored observations



- Observations $(LN_p(t_{p,1}), \ldots, LN_p(t_{p,N_p}))$

  $\Leftrightarrow Z_{p,f} \in (\nu_{p,f}, \tau_{p,f}],\ f = 1, \ldots, F_p$

  with

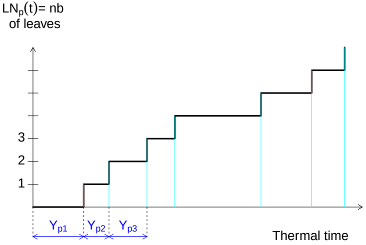
  - $\nu_{p,f}$ = last observation time before appearance of leaf f (or 0)
  - $\tau_{p,f}$ = first observation time after appearance of leaf f (or $+\infty$)
- Rk : Observations $\not\Leftrightarrow Y_{p,f} \in (\nu'_{p,f}, \tau'_{p,f}]$

Issue : estimation of the model from discrete measurements/ interval censored observations

The model

Consider a given genotype and a given year, for each plant p

$$\mathbf{Y}_p = (Y_{p,f})_{f=1,\ldots,F_p} \sim \mathcal{N}_{F_p}(\mu, D), \quad D = \mathrm{diag}(\sigma^2)$$

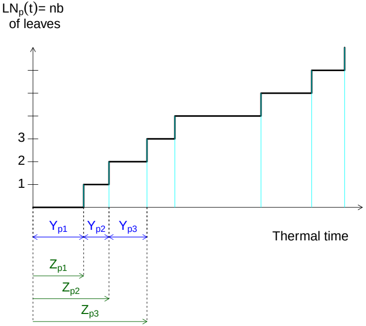


- $Y_{p,f} = \mu_f + \varepsilon_{p,f}, \quad \varepsilon_{p,f} \sim \mathcal{N}(0, \sigma_f^2)$
- Normality assumption for computing reasons

The model

Consider a given genotype and a given year, for each plant p

$$\mathbf{Y}_p = (Y_{p,f})_{f=1,\dots,F_p} \sim \mathcal{N}_{F_p}(\mu, D), \quad D = \text{diag}(\sigma^2)$$



- $Y_{p,f} = \mu_f + \varepsilon_{p,f}, \quad \varepsilon_{p,f} \sim \mathcal{N}(0, \sigma_f^2)$
- Normality assumption for computing reasons

$\Rightarrow$ Leaf appearance times $\mathbf{Z}_p \sim \mathcal{N}(\overline{\mu}, \Sigma))$ with Σ not diagonal

Monte Carlo EM algorithm

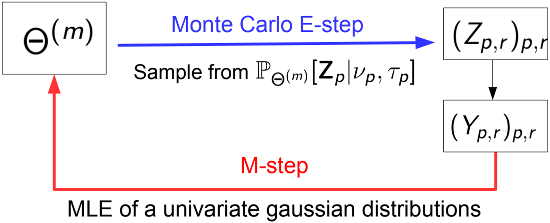
- Latent variables :

$$(\mathbf{Y}_p)_p \Leftrightarrow (\mathbf{Z}_p)_p$$

- Observed variables :

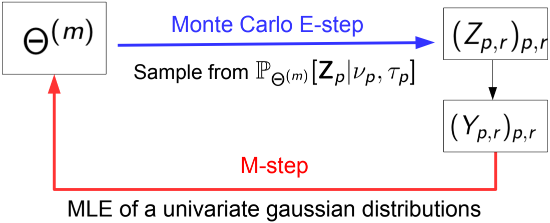
$$((\nu_{p,f}, \tau_{p,f})_f)_p$$

- $\Theta = (\mu, \sigma)$



Monte Carlo EM algorithm

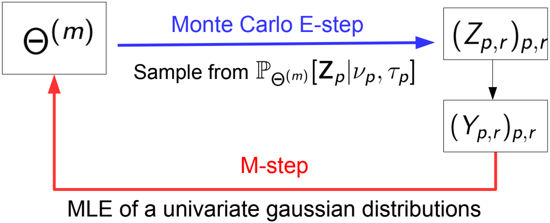
- Latent variables :
$$(\mathbf{Y}_p)_p \Leftrightarrow (\mathbf{Z}_p)_p$$
- Observed variables :
$$((\nu_{p,f}, \tau_{p,f})_f)_p$$
- $\Theta = (\mu, \sigma)$



- Sample from truncated multivariate distribution $\mathbb{P}[\mathbf{Z}_p|\mathbf{Z}_p \in [\nu_p, \tau_p)]$
  ↪ Basic rejection methods fail when dimension increases

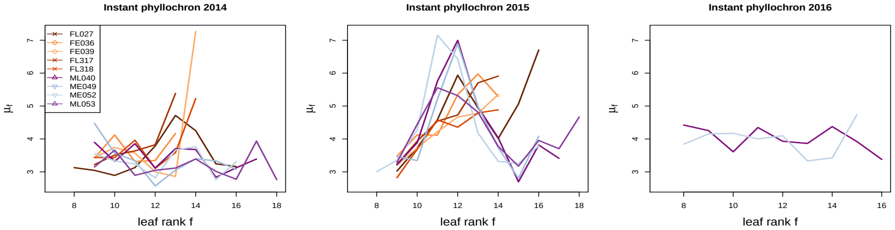
Monte Carlo EM algorithm

- Latent variables :
  $$(\mathbf{Y}_p)_p \Leftrightarrow (\mathbf{Z}_p)_p$$
- Observed variables :
  $$((\nu_{p,f}, \tau_{p,f})_f)_p$$
- $\Theta = (\mu, \sigma)$



- Sample from truncated multivariate distribution $\mathbb{P}[\mathbf{Z}_p|\mathbf{Z}_p \in [\nu_p, \tau_p)]$
  $\hookrightarrow$ Basic rejection methods fail when dimension increases
- Various methods for gaussian distribution
  $\hookrightarrow$ Package `truncatedNormal`, Botev (2016)

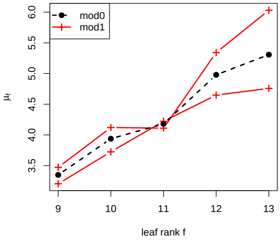
Implementation on ITEMAIZE data



- Test genotypic group effect
- Test assumption of *linear phyllochron* i.e. $\mu_f = constant$
- Imput of climatic variables

Genotypic group effects

- Model comparison



- (mod1) $(\mu_f, \sigma_f)_{f=f^{\mathbf{min}},\dots,f^{\mathbf{max}}}$ depend on genotype
- (mod0) $(\mu_f, \sigma_f)_{f=f^{\mathbf{min}},\dots,f^{\mathbf{max}}}$ same for all genotypes in the selection group

- Criteria based on likelihood : AIC, χ^2-likelihood ratio test
  - Almost all tests are significant (some strongly)
  - Permutation test on one comparison

    $\hookrightarrow$ Results reliable even slightly over-estimated

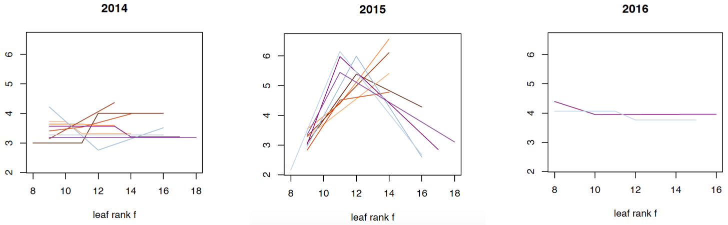
Parametric submodels (Linear phyllochron ?)

Parametric submodels (Linear phyllochron ?)

- Various parametric models for $f \mapsto \mu_f$:
  - constant
  - linear
  - piecewise constant
  - piecewise linear

Parametric submodels (Linear phyllochron ?)

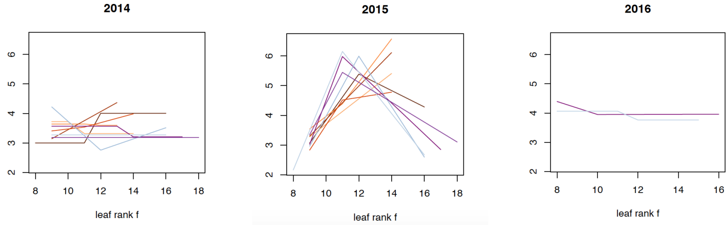
- Various parametric models for $f \mapsto \mu_f$:
  - constant
  - linear
  - piecewise constant
  - piecewise linear
- Constant model (H0) vs each other model (H1) (χ^2-likelihood ratio test /AIC)



Parametric submodels (Linear phyllochron ?)

- Various parametric models for $f \mapsto \mu_f$:
  - constant
  - linear
  - piecewise constant
  - piecewise linear

- Constant model (H0) vs each other model (H1) (χ^2-likelihood ratio test /AIC)



$\hookrightarrow$ Constant μ selected only once

Input of climatic variables

Input of climatic variables

- Regression of 7 climatic variable on μ : for each year and genotype

  (1) Estimation of μ

  (2) Regression of climatic variables on μ + variable selection

Input of climatic variables

- Regression of 7 climatic variable on μ : for each year and genotype

  (1) Estimation of μ

  (2) Regression of climatic variables on μ + variable selection

  $\hookrightarrow$ No account for error of the estimate of μ

Input of climatic variables

- Regression of 7 climatic variable on μ : for each year and genotype

  (1) Estimation of μ

  (2) Regression of climatic variables on μ + variable selection

  $\hookrightarrow$ No account for error of the estimate of μ

- Results : in progress

Strengths of the model

- Less biased modelling

Classic approach

Leaf number :

$$LN(t) = f(TT_t) + \varepsilon_t$$

with $(\varepsilon_t)_t$ independent

$\hookrightarrow$ No account for auto-correlation

Our model

Interval between successive leaves

$$Y_f \sim \mathcal{D}(\theta_f)$$

with $(Y_f)_f$ independent.

$\hookrightarrow$ Independence : realistic

- More flexible modelling
- Allows to
  - Test conditions/genotypic effects
  - Select parametric models for μ
  - Evaluate climate effect

Weaknesses of the model

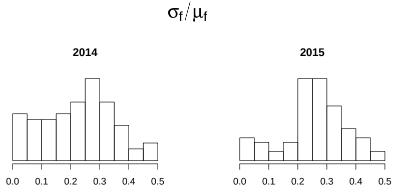
- Large number of parameters
  - Price to pay for flexible model
  - Option : equal variance parameters within line

Weaknesses of the model

- Large number of parameters
  - Price to pay for flexible model
  - Option : equal variance parameters within line
- Model comparison via likelihood : less flexible than mixed models

Weaknesses of the model

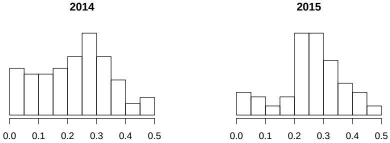
- Large number of parameters
  - Price to pay for flexible model
  - Option : equal variance parameters within line
- Model comparison via likelihood : less flexible than mixed models
- Plant level variation : gaussian distribution for a time-to-event



Weaknesses of the model

- Large number of parameters
  - Price to pay for flexible model
  - Option : equal variance parameters within line
- Model comparison via likelihood : less flexible than mixed models
- Plant level variation : gaussian distribution for a time-to-event

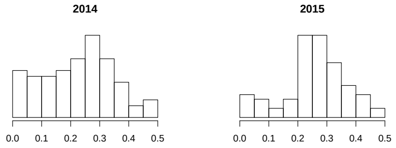
σ_f/μ_f



Weaknesses of the model

- Large number of parameters
  - Price to pay for flexible model
  - Option : equal variance parameters within line
- Model comparison via likelihood : less flexible than mixed models
- Plant level variation : gaussian distribution for a time-to-event

σ_f/μ_f

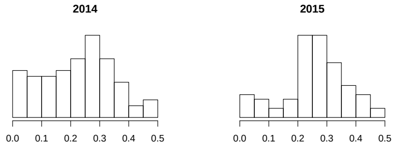


- Longitudinal covariates
  - Regression on the phyllochron parameters μ
  - $\neq$ include longitudinale covariate in the model

Weaknesses of the model

- Large number of parameters
  - Price to pay for flexible model
  - Option : equal variance parameters within line
- Model comparison via likelihood : less flexible than mixed models
- Plant level variation : gaussian distribution for a time-to-event

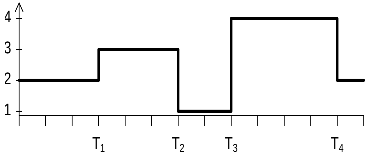
σ_f/μ_f



- Longitudinal covariates
  - Regression on the phyllochron parameters μ
  - $\neq$ include longitudinale covariate in the model

(Discrete time) Semi-Markov models

$(S_t)_{t\in\mathbb{N}^*}$ a discrete time process with $S_t \in \{1,\dots,J\}$

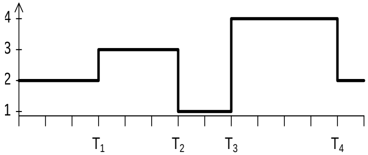


- Semi Markov property : State and sojourn duration depends only on previous state and sojourn duration

$$\mathbb{P}[S_{[t+1:t+d]} = i | (S_{t'})_{t'\leq t}] = \mathbb{P}[S_{[t+1:t+d]} = i | S_{[t-d'+1:t]} = j] := a_{(i,d)(j,d')}$$

(Discrete time) Semi-Markov models

$(S_t)_{t\in\mathbb{N}^*}$ a discrete time process with $S_t \in \{1,\dots,J\}$



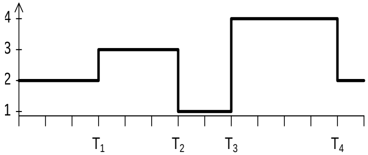
- Semi Markov property : State and sojourn duration depends only on previous state and sojourn duration

$$\mathbb{P}[S_{[t+1:t+d]} = i | (S_{t'})_{t'\leq t}] = \mathbb{P}[S_{[t+1:t+d]} = i | S_{[t-d'+1:t]} = j] := a_{(i,d)(j,d')}$$

with $\{S_{[t+1:t']} = i\} = \{S_{t+1} = \dots = S_{t'} = i, S_t \neq i, S_{t'+1} \neq i\}$

(Discrete time) Semi-Markov models

$(S_t)_{t\in\mathbb{N}^*}$ a discrete time process with $S_t \in \{1,\dots,J\}$



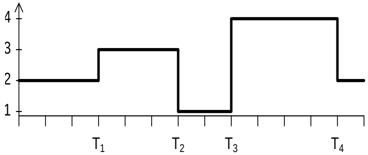
- Semi Markov property : State and sojourn duration depends only on previous state and sojourn duration

$$\mathbb{P}[S_{[t+1:t+d]} = i | (S_{t'})_{t'\leq t}] = \mathbb{P}[S_{[t+1:t+d]} = i | S_{[t-d'+1:t]} = j] := a_{(i,d)(j,d')}$$

with $\{S_{[t+1:t']} = i\} = \{S_{t+1} = \dots = S_{t'} = i, S_t \neq i, S_{t'+1} \neq i\}$

(Discrete time) Semi-Markov models

$(S_t)_{t \in \mathbb{N}^*}$ a discrete time process with $S_t \in \{1, \dots, J\}$



- Semi Markov property : State and sojourn duration depends only on previous state and sojourn duration

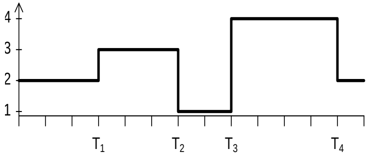
$$\mathbb{P}[S_{[t+1:t+d]} = i | (S_{t'})_{t' \leq t}] = \mathbb{P}[S_{[t+1:t+d]} = i | S_{[t-d'+1:t]} = j] := a_{(i,d)(j,d')}$$

with $\{S_{[t+1:t']} = i\} = \{S_{t+1} = \cdots = S_{t'} = i, S_t \neq i, S_{t'+1} \neq i\}$

- *Unidirectional* SMM = transition only from i to $i+1$

(Discrete time) Semi-Markov models

$(S_t)_{t\in\mathbb{N}^*}$ a discrete time process with $S_t \in \{1, \ldots, J\}$



- Semi Markov property : State and sojourn duration depends only on previous state and sojourn duration

$$\mathbb{P}[S_{[t+1:t+d]} = i | (S_{t'})_{t'\leq t}] = \mathbb{P}[S_{[t+1:t+d]} = i | S_{[t-d'+1:t]} = j] := a_{(i,d)(j,d')}$$

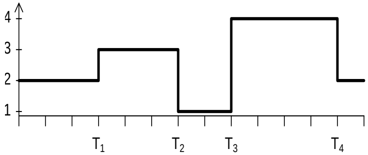
with $\{S_{[t+1:t']} = i\} = \{S_{t+1} = \cdots = S_{t'} = i, S_t \neq i, S_{t'+1} \neq i\}$

- *Unidirectional* SMM $=$ transition only from i to $i+1$
- Phyllochron : sojourn time independent of time in previous state :

$$a_{(i,d)(j,d')} = \mathbb{1}_{i=j+1} f_i(d) \quad \text{with} \quad f_i(d) = \mathbb{P}[Y_i = d].$$

(Discrete time) Semi-Markov models

$(S_t)_{t\in\mathbb{N}^*}$ a discrete time process with $S_t \in \{1, \dots, J\}$



- Semi Markov property : State and sojourn duration depends only on previous state and sojourn duration

$$\mathbb{P}[S_{[t+1:t+d]} = i | (S_{t'})_{t' \leq t}] = \mathbb{P}[S_{[t+1:t+d]} = i | S_{[t-d'+1:t]} = j] := a_{(i,d)(j,d')}$$

with $\{S_{[t+1:t']} = i\} = \{S_{t+1} = \dots = S_{t'} = i, S_t \neq i, S_{t'+1} \neq i\}$

- *Unidirectional* SMM = transition only from i to $i+1$
- Phyllochron : sojourn time independent of time in previous state :

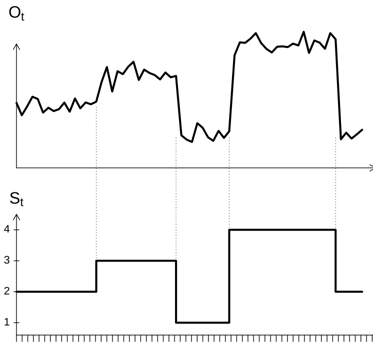
$$a_{(i,d)(j,d')} = 1\!\!1_{i=j+1} f_i(d) \quad \text{with} \quad f_i(d) = \mathbb{P}[Y_i = d].$$

$\hookrightarrow$ If f_i geometric : Poisson point process

SMM with interval censoring : analogy with HSMM

Hidden SMM

- $(S_t)_{t=1,\ldots,T_{\max}}$ unobserved SMM



SMM with interval censoring : analogy with HSMM

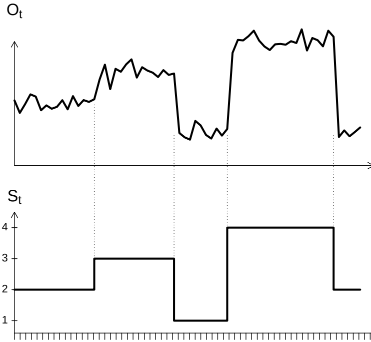
Hidden SMM

- $(S_t)_{t=1,\ldots,T_{\max}}$ unobserved SMM
- $(O_t)_{t=1,\ldots,T_{\max}}$ observed process with

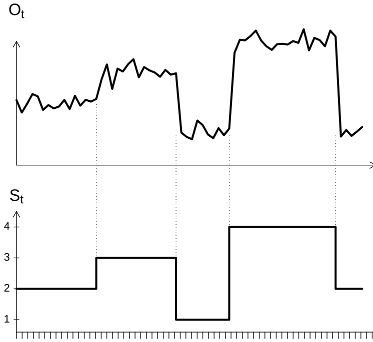
$$\mathbb{P}[O_t = x | (S_{t'})_{t'}, (O_{t'})_{t' \neq t}]$$

$$= \mathbb{P}[O_t = x | S_t = j] = b_j(x)$$

$\hookrightarrow b_j(\cdot) =$ emission proba in state j.



Hidden SMM



- $(S_t)_{t=1,\ldots,T_{\max}}$ unobserved SMM
- $(O_t)_{t=1,\ldots,T_{\max}}$ observed process with

$$\mathbb{P}[O_t = x | (S_{t'})_{t'}, (O_{t'})_{t' \neq t}]$$

$$= \mathbb{P}[O_t = x | S_t = j] = b_j(x)$$

$\hookrightarrow b_j(\cdot) =$ emission proba in state j.

- Forward/backward algo : complexity

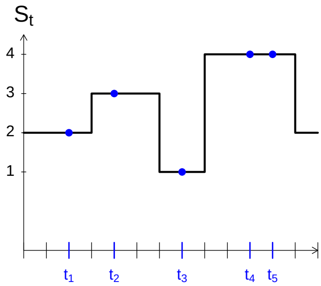
$$O((J^2 + JD^2)T_{\max}) \quad \text{with} \quad \begin{cases} J & \text{number of states} \\ D & \text{maximum sojourn time} \\ T_{\max} & \text{total time} \end{cases}$$

- For unidirectional HSMM, complexity $= O(JDT_{\max})$

SMM with interval censoring : analogy with HSMM

- Interval censoring :

$$\begin{cases} (S_t)_{t=1,dots,T_{\max}} & \text{SMM} \\ (t_1,\ldots,t_N) & \text{monitoring times} \\ (t_i, S_{t_i})_{i=1,\ldots,N} & \text{Observations} \end{cases}$$

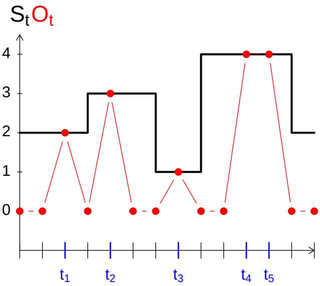


- Interval censoring :

$$\left\{ \begin{array}{ll} (S_t)_{t=1,dots,T_{\max}} & \text{SMM} \\ (t_1, \ldots, t_N) & \text{monitoring times} \\ (t_i, S_{t_i})_{i=1,\ldots,N} & \text{Observations} \end{array} \right.$$

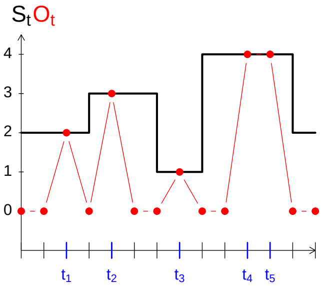
- $(O_t)_t$ observed process (determ. given $(S_t)_t$)

$$O_t = \left\{ \begin{array}{lll} S_t & \text{if} & t \in \{t_1, \ldots, t_N\} \\ 0 & & \text{otherwise} \end{array} \right.$$



SMM with interval censoring : analogy with HSMM

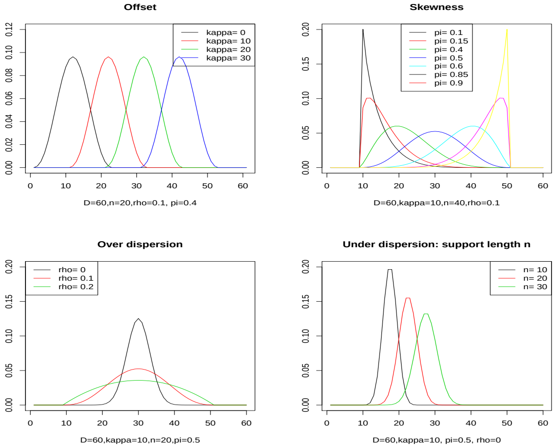
- Interval censoring :
$$\left\{ \begin{array}{ll} (S_t)_{t=1,dots,T_{\max}} & \text{SMM} \\ (t_1, \ldots, t_N) & \text{monitoring times} \\ (t_i, S_{t_i})_{i=1,\ldots,N} & \text{Observations} \end{array} \right.$$
- $(O_t)_t$ observed process (determ. given $(S_t)_t$)
$$O_t = \left\{ \begin{array}{lll} S_t & \text{if} & t \in \{t_1, \ldots, t_N\} \\ 0 & & \text{otherwise} \end{array} \right.$$



- Algorithm forward/backward : very similar to HSMM

Application to phyllochron : flexible distribution

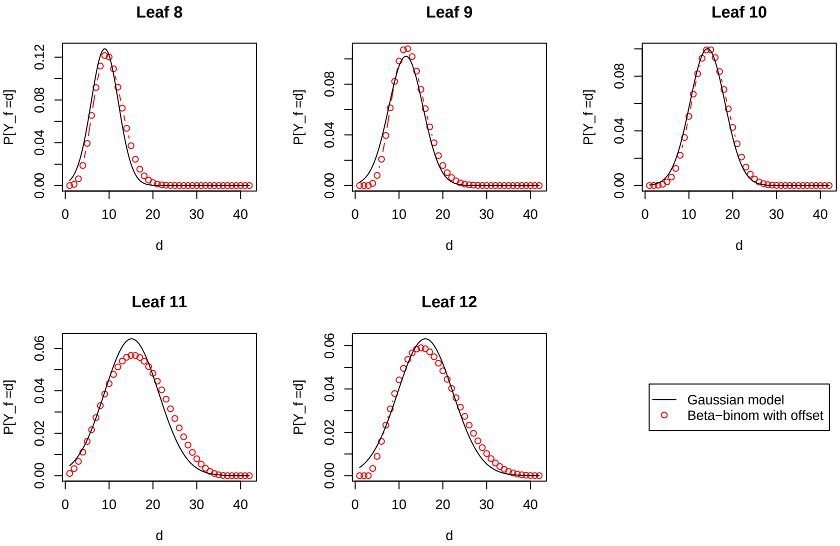
- Support $[\![1, D]\!]$
- Beta-binomiale with offset : $\kappa + \mathcal{BB}(\text{size} = n, \text{prob} = \pi, \rho)$, $n + \kappa \leq D$



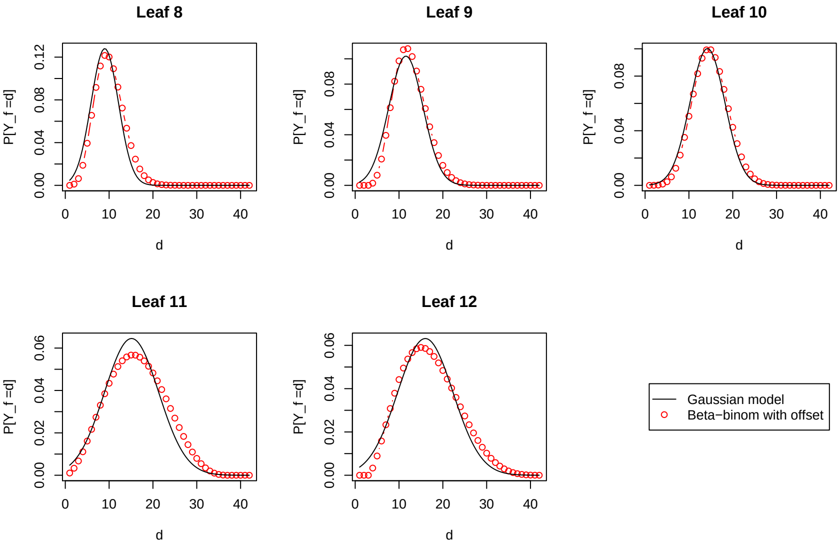
Flexible (unimodal)

- Mode =0 / $\neq$ 0
- Skewness
- Over/under dispersion

Ex : results for group F-late



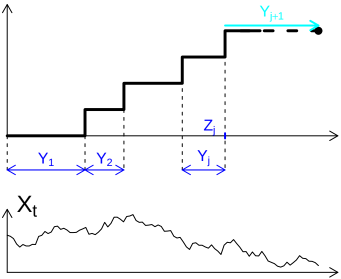
Ex : results for group F-late



⇒ Gaussian assumption is reasonable

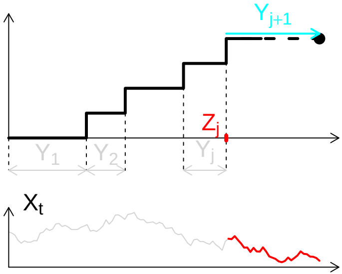
Longitudinal covariate in unidirectional SMM with interval censoring

- $\mathbf{X} = (X_t)_{t=1:T_{\max}}$ longitudinal covariate observed at all times



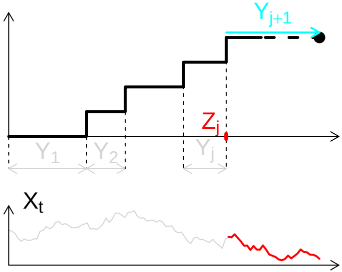
Longitudinal covariate in unidirectional SMM with interval censoring

- $\mathbf{X} = (X_t)_{t=1:T_{\max}}$ longitudinal covariate observed at all times
- $\mathbb{P}[Y_{j+1} = d | Y_1, \ldots, Y_j, \mathbf{X}] = \mathbb{P}[Y_{j+1} = d | X_{t+1:t+d}] = a_{j,d}(t)$



Longitudinal covariate in unidirectional SMM with interval censoring

- $\mathbf{X} = (X_t)_{t=1:T_{\max}}$ longitudinal covariate observed at all times
- $\mathbb{P}[Y_{j+1} = d | Y_1, \ldots, Y_j, \mathbf{X}] = \mathbb{P}[Y_{j+1} = d | X_{t+1:t+d}] = a_{j,d}(t)$



Exple. Discrete time Cox model

Perspectives

- Improve numerical implementation (general SMM w. interval censoring)

  $\hookrightarrow$ Issues similar to HSMM

- Penalized likelihood
  - Flexible distribution (splines, etc)
  - Several longitudinal covariates

- Multi-chain SMM

  $\hookrightarrow$ Application : floral transition

Conclusion

Phyllochron modelling

- Flexible average phyllochron

  ↪ Enlighten temporal trends

- Downstream analysis : less tractable than classic linear phyllochron
- Stochastic process *vs* regression models
  - Other fields e.g. plant pathogens (Nemis *et al* 2013)

Semi-Markov model with interval censoring

- More flexible framework for phyllochron
  - Any parametric distribution
  - Longitudinal covariate (unidirectional SMM)
- General interest
  - SMM with interval censoring : almost not adressed in literature
  - Application in various fields (disease progression...)
- Algorithm analogous to Hidden SMM
  - Adapt methods for HSMM to interval censored SMM