Anim'AGAP

model4all: une initiative collaborative pour la modélisation statistique

Marie Denis et Timothée Flutre

Cirad & Inra

16/06/2016

Résumé

Un généticien, un écophysiologiste et un statisticien sont dans un bateau. Le statisticien tombera-t-il à l'eau ? Afin d'éviter un tel scénario, nous avons souhaité partager l'état actuel de notre réflexion concernant ce type de collaboration interdisciplinaire du point de vue "statisticien". Dans un premier temps, nous tenterons d'expliciter les problèmes rencontrés et les besoins identifiés. Puis, nous proposerons une initiative collaborative visant à améliorer et développer les interactions actuelles, sur le plan de la pédagogie comme de la recherche. Enfin, nous l'illustrerons sur des thématiques d'intérêt aux généticiens et écophysiologistes, ouvrant en perspective sur des approches couplant plusieurs thématiques. A l'occasion de cette présentation, nous serions heureux de pouvoir échanger avec toutes personnes à Montpellier, de l'UMR AGAP, mais bien sûr aussi au-delà.

Plan

Contexte, problèmes, besoins et motivation

Pédagogie et outils

Thèmes actuels
- Génétique quantitative
- Observations corrélées
- Réduction de dimension

Perspectives

Plan

Contexte

- Origine : participation à un groupe de travail ≪ modélisation en génétique et éco-physiologie ≫ de l'UMR AGAP réunissant des chercheurs de ces deux disciplines (Evelyne Costes, Delphine Luquet, Sébastien Tisné, ...) et des statisticiens (Laurianne Rouan, Éric Gozé)

- But : exploration et développement de modèles statistiques complexes couplant des données génétiques et éco-physiologiques en réponse à des problématiques biologiques communes

Problèmes

On s'attendait à quelque chose du style $Y = f(X)$ où :

- X et Y sont des observations potentiellement hétérogènes et de grande dimension,
- et f est une relation potentiellement très compliquée ;

et à pouvoir « jouer » avec des logiciels sur des cas simples.

Mais on a dû se faire une idée par nous même à partir de diverses ressources « papier » :

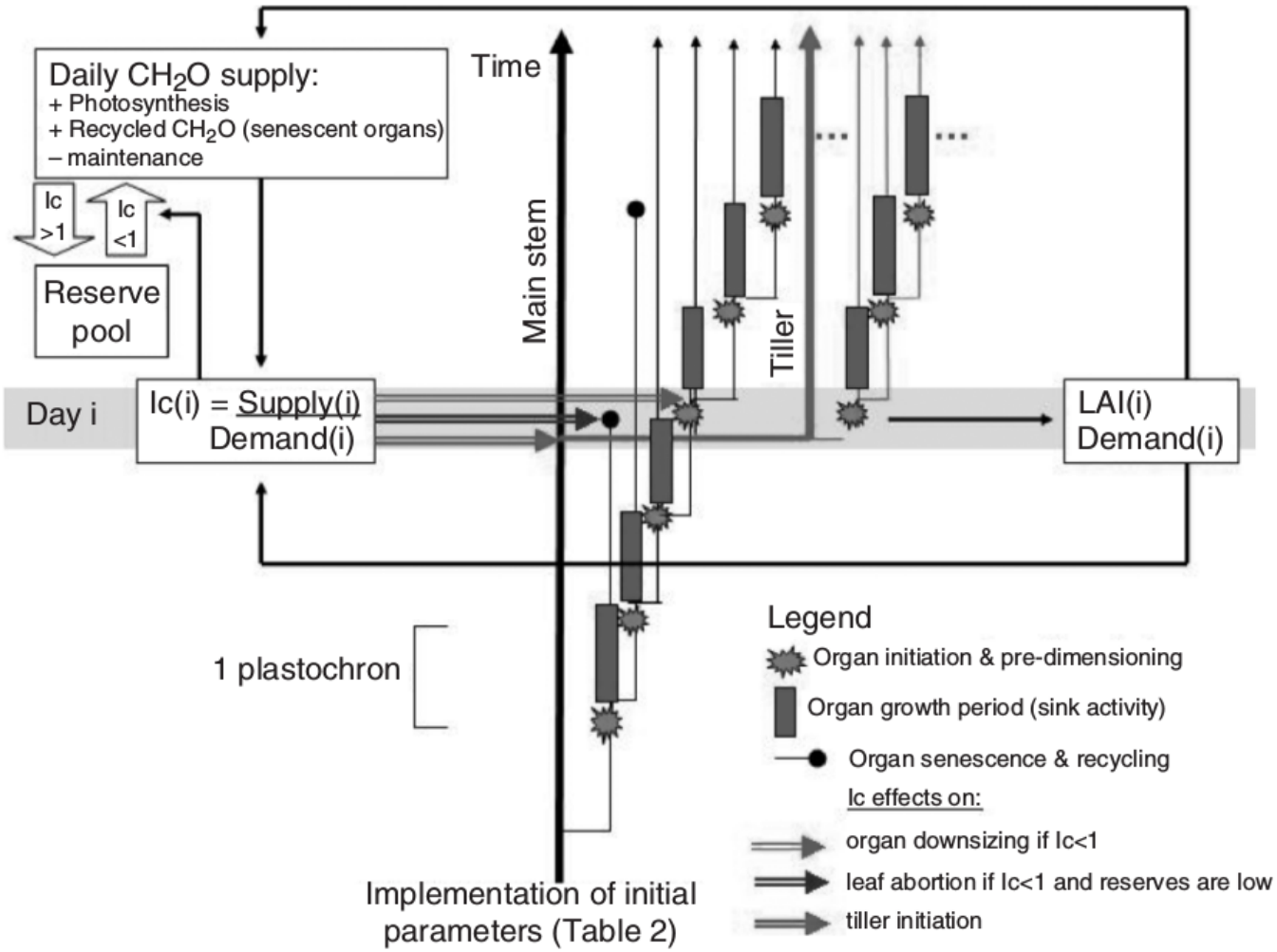
- cours de B. Pallas (mais à l'époque encore en Australie),
- articles présentant EcoMeristem, MAppleT, SUNFLO, etc.

Problèmes

$$Rdt = IR \times \sum_{levée}^{récolte} \varepsilon_b \times \varepsilon_a \times \varepsilon_c \times \Delta Rgi$$

Pallas (2011)

Problèmes



Luquet *et al* (2006)

Problèmes

- Quel est le statut des différentes variables ? (observées, non-observées)
- Sont-elles mesurées sur les mêmes unités statistiques ?
- Quels sous-modèles sont utilisés pour simuler ? pour inférer/prédire ?
- L'incertitude est-elle quantifiée ?
- Où peut-on télécharger les logiciels « domaine-spécifiques » implémentant les modèles des écophysiologistes ?

Vocabulaire différents et objectifs variés

Qu'est-ce qu'un paramètre ?

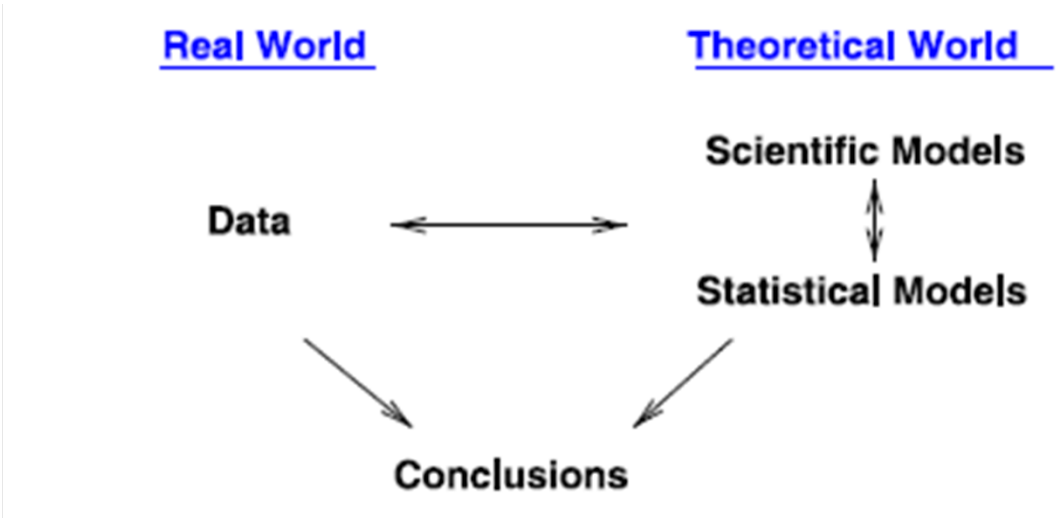
Qu'est-ce qu'un modèle ?

Qu'est-ce qu'on modélise ?

Que cherche-t-on à faire avec ce modèle ?

Besoins : sensibilisation des chercheurs à la modélisation *statistique*

Kass (2011) : besoin d'un cadre logique qui permet de mettre en relation le « monde réel » et le « monde théorique »



"Careful consideration of the connection between models and data is a core component of both the art of statistical practice and the science of statistical methodology."

Besoins : sensibilisation des chercheurs à la modélisation *statistique*

- Besoin d'introduire des objets mathématiques avec des hypothèses théoriques pour que le monde théorique puisse correspondre « raisonnablement bien » au monde réel (et donc à la question portant sur les données)

$\Rightarrow$ Besoin de comprendre l'écriture d'un modèle

Besoins : améliorer la communication et l'accès aux différentes « ressources statistiques »

- Besoin de communiquer sur les méthodes d'inférence existantes, les critères de comparaison, les implémentations logicielles, ...
- Besoin d'avoir accès à ces différentes ressources de façon « optimisée » (i.e. via des outils informatiques adaptés)

Motivation

Elaborer une stratégie collaborative ≪ la plus simple possible ≫ :

- faciliter les échanges efficaces entre statisticiens, et entre biologistes et statisticiens ;
- donner accès aux différentes approches d'inférences statistiques de façon pédagogique ;
- favoriser la réutilisation des ≪ briques ≫ entre chercheurs pour éviter de réinventer la roue ;
- partager son questionnement quotidien de statisticiens en explorant les limites rencontrées et possibilités de développement méthodologique.

Plan

Particularité de l'analyse statistique de données

Lindley (2000) : "statistics is the study of uncertainty"

- toutes les disciplines scientifiques sont donc concernées par les statistiques
- tout statisticien appliqué est donc de fait obligé de s'impliquer dans la formation de ses collaborateurs

Particularité de l'analyse statistique de données

Kass (2011) : "it makes more sense to place in the center of our logical framework the match or mismatch of theoretical assumptions with the real world of data"

- s'assurer au début d'une collaboration que la démarche de modélisation statistique est partagée
- introduire ces composants de base, sans trop rentrer dans les considérations techniques

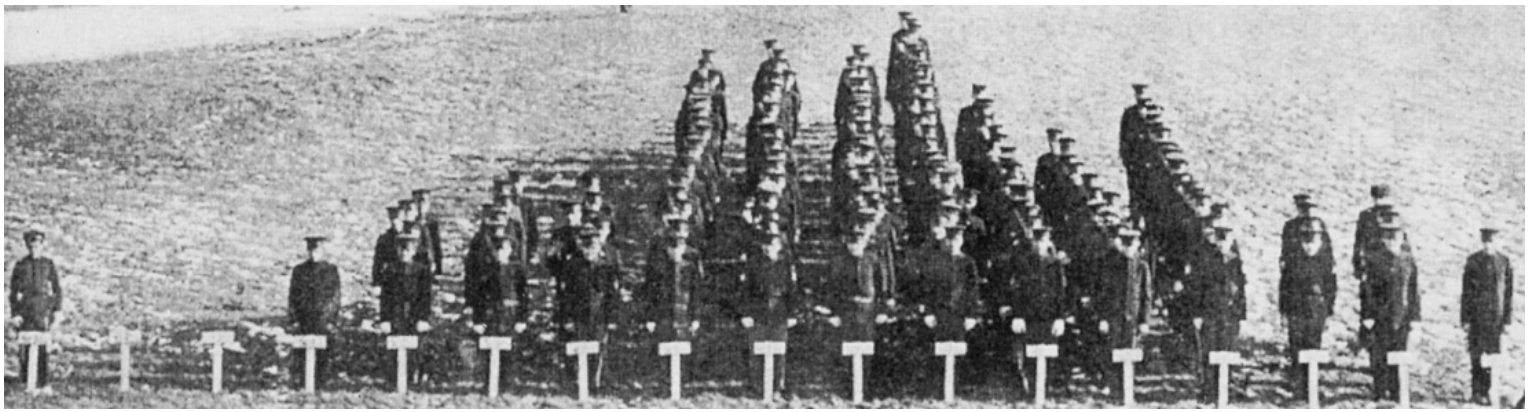
Particularité de l'analyse statistique de données

Gelman and Shalizi (2011) : "we build a model out of available parts and drive it as far as it can take us, and then a little farther. When the model breaks down, we take it apart, figure out what went wrong, and tinker with it, or else try a radically new design"

- voir la modélisation statistique comme un processus sur un temps long, et non le statisticien comme un simple calculateur de $p - values$
- il n'y a pas qu'un seul modèle pertinent, il y a une diversité d'alternatives à chaque étape du processus
- il faut au minimum être conscient de la liste des alternatives, sans pour autant avoir besoin de toutes les connaître dans les moindres détails soi-même

Observer de la variation $\rightarrow$ utiliser une variable aléatoire

Exemple de la taille chez l'homme (Blakeslee, 1914) :

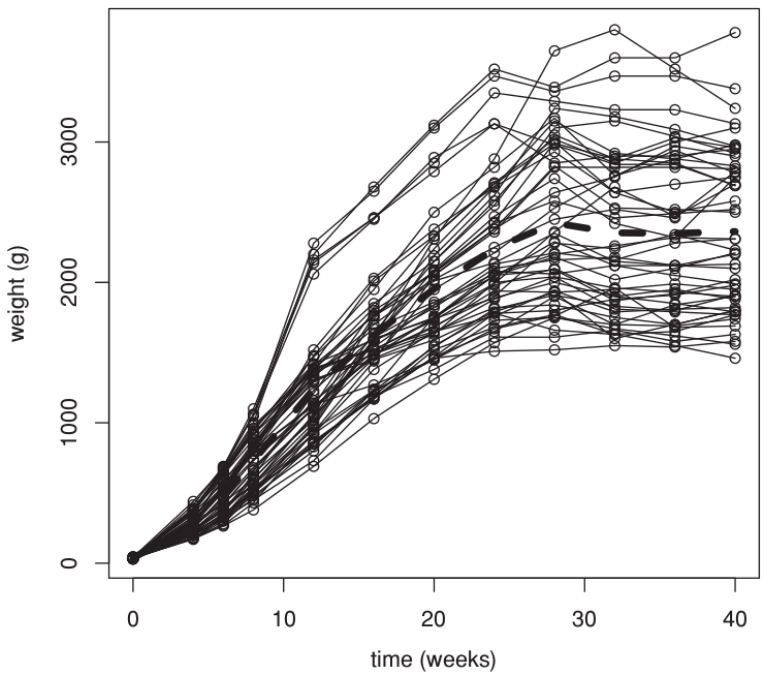


Y : variable aléatoire continue de dimension 1

- y_i : réalisation de Y pour l'individu i

Observer de la variation $\rightarrow$ utiliser une variable aléatoire

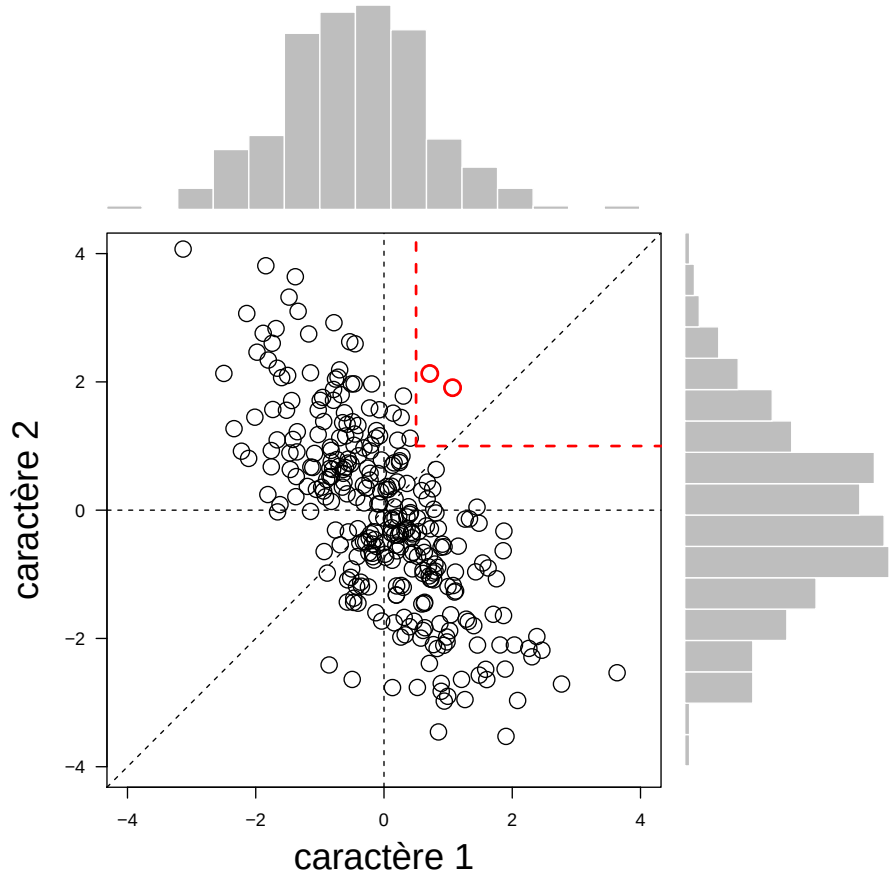
Exemple du poids des poulets (Donnet et coll., 2010) :



Y : variable aléatoire continue de dimension 2

- y_{it} : réalisation de Y pour le poulet i au temps t

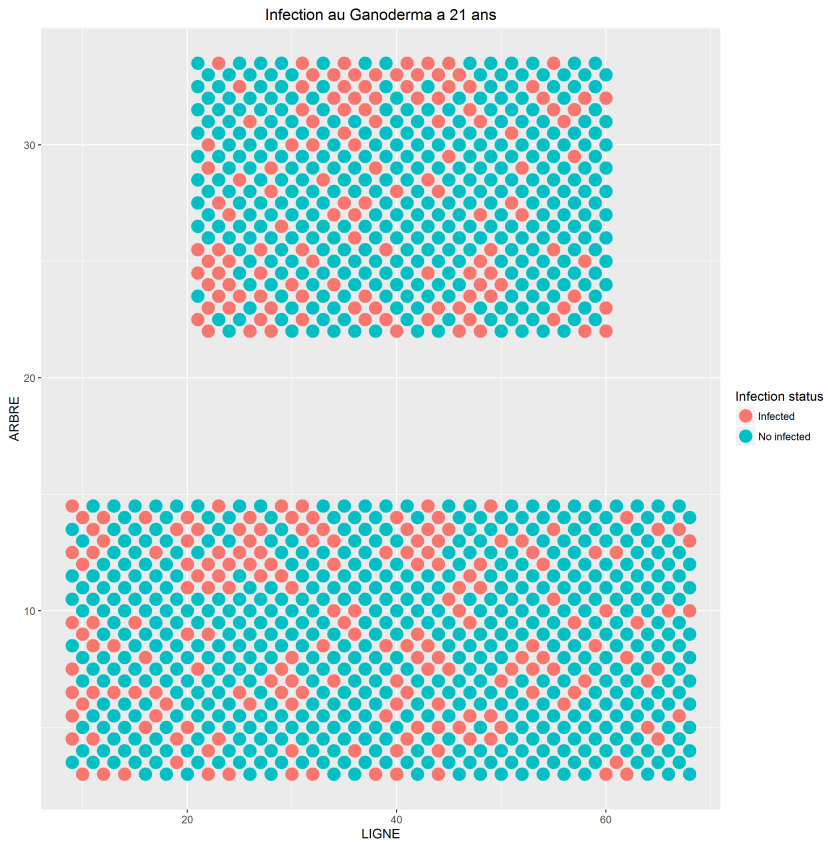
Observer de la variation $\rightarrow$ utiliser une variable aléatoire



Bien sûr, à chaque fois qu'une nouvelle dimension apparaît, de la co-variation apparaît aussi...

Observer de la variation $\rightarrow$ utiliser une variable aléatoire

Exemple de propagation de l'infection au *Ganoderma* :



Y : variable aléatoire discrète de dimension 3

- y_{its} : réalisation de Y pour l'arbre i le mois t à la position $s = (x, y)$

Observer de la variation $\rightarrow$ utiliser une variable aléatoire

Kass (2011) :

> *Introducing mathematical objects called random variables is an abstraction, but it can be an extraordinarily useful abstraction whenever the theoretical world of random variables is aligned well with the real world of the data.*
>
> *When we use a statistical model to make a statistical inference we implicitly assert that the variation exhibited by data is captured reasonably well by the statistical model, so that the theoretical world corresponds reasonably well to the real world.*

Relier toutes les variables en un modèle statistique

Ecrire une distribution de probabilité conjointe pour toutes les variables du modèle : $p_1(y, \theta)$

- y : variable observée (donnée)
- θ : variable non-observable (paramètre, variable latente)

Autres variables :

- X : variable explicative (covariable ; non-aléatoire)
- $\tilde{y}$: variable observable (prédiction)

Relier toutes les variables en un modèle statistique

Ecrire une distribution de probabilité conjointe pour toutes les variables du modèle : $p_1(y, \theta)$

- y : variable observée (donnée)
- θ : variable non-observable (paramètre, variable latente)

Autres variables :

- X : variable explicative (covariable ; non-aléatoire)
- $\tilde{y}$: variable observable (prédiction)

Quel que soit le paradigme dans lequel on se place, il y a toujours des choix « subjectifs » à faire, plus ou moins interdépendants, concernant les distributions (p), les paramètres (θ) et leur nombre (P), les facteurs à mesurer (y) et leur nombre (N), etc.

Réaliser l'inférence statistique

Gelman et coll. (2014) : "statistical inference draws conclusions, from numerical data, about quantities that are not observed"

En bref, on utilise des données $\mathcal{D} = \{y_1, \ldots, y_N\}$ pour :

- estimer des paramètres $\Theta = \{\theta_1, \ldots, \theta_P\}$;
- prédire de nouvelles données $\{\tilde{y}_1, \ldots, \tilde{y}_{N'}\}$.

Réaliser l'inférence statistique

Gelman et coll. (2014) : "statistical inference draws conclusions, from numerical data, about quantities that are not observed"

En bref, on utilise des données $\mathcal{D} = \{y_1, \ldots, y_N\}$ pour :

- estimer des paramètres $\Theta = \{\theta_1, \ldots, \theta_P\}$;
- prédire de nouvelles données $\{\tilde{y}_1, \ldots, \tilde{y}_{N'}\}$.

Les jeux de données hétérogènes et de grande taille mènent naturellement à des modèles hiérarchiques ("borrow strength") au sein desquels on peut viser à :

- sélectionner un sous-ensemble de paramètres ;
- réduire la dimension des données ;
- comparer, selectionner et moyenner plusieurs sous-modèles similairement plausibles, etc.

L'approche générale

En trois étapes :

- simulation : simuler des données en fonction d'un modèle théorique donné
- inférence : inférer avec une ou plusieurs méthodes d'inférence et un ou plusieurs logiciels en ajustant des modèles statistiques qui peuvent être différents du modèle utilisé précédemment
- évaluation : estimation des paramètres, qualité d'ajustement du modèle, capacité de prédiction du modèle, temps de calcul,...

Simulation : une démarche pédagogique

Avec les simulations, nous connaissons et contrôlons la « vérité », ce qui permet :

- d'étudier l'adéquation des hypothèses faites
- de tester le développement de modèles statistiques complexes
- de se familiariser avec la démarche d'inférence statistique
- de tester une implémentation
- de ne pas être bloqué par le non-accès aux données
- ...

Inférence : notion de vraisemblance

De nombreuses méthodes d'inférence existent ; l'objectif ici est d'introduire ces méthodes globalement, et de manière pédagogique et pragmatique.

La « vraisemblance », fonction des paramètres :

- notion-clé pour l'estimation par maximum de vraisemblance et dans l'approche Bayésienne
- intervient dans la distribution de probabilité conjointe :

$$p_1(y, \theta) = p_2(\theta) \times p_3(y|\theta)$$

$$p_1(y, \theta) = p_2(\theta) \times \underbrace{L(\theta; y)}_{\textit{vraisemblance}}$$

- permet de comparer des modèles entre eux

Inférence : méthodes explorées

- Estimateur du maximum de vraisemblance (ML)
- Estimateur de maximum de vraisemblance restreint (REML)
- Méthodes de Monte Carlo : par chaînes de Markov (MCMC : algorithmes de Gibbs et de Metropolis-Hastings, Hamiltonian Monte Carlo), Sequential Monte Carlo (SMC), etc
- Approximations bayesiennes : Integrated Nested Laplace Approximation (INLA), variational Bayes (VB), expectation-propagation (EP), etc
- ...

Inférence : choix d'une méthode

Différents questionnements peuvent amener à choisir différentes méthodes d'inférence :

- quel logiciel me permet d'implémenter le modèle statistique que j'ai choisi ?
- quelle méthode d'inférence me permet d'analyser des jeux de données très volumineux ?
- quelle approche me permet d'intégrer de la connaissance issue de travaux antérieurs ?
- la capacité mémoire de ma machine est-elle suffisante ? quel langage, quel logiciel sont le plus adaptés à mes données, à mon modèle ?
- ...

Evaluation : méthodes et critères à disposition

Ils sont nombreux et ils dépendent des objectifs :

- Méthodes de rééchantillonage : bootstrap, jacknife, permutations, validation croisée, etc
- Mesures de qualité d'un modèle : Akaike information criterion (AIC), Bayesian information criterion (BIC), Deviance information criterion (DIC), Watanabe-Akaike information criterion (WAIC), Bayes factor, R^2, erreur quadratique moyenne (MSE), intervalles de confiance, intervalles de crédibilité, scores basés sur la distribution prédictive a posteriori, erreur standard de Monte Carlo, etc

Evaluation : méthodes et critères à disposition

- Qualité de prédiction : MSEP, scores basés sur la distribution prédictive, validation croisée
- Critère d'évaluation de la convergence des chaînes de Markov : $\hat{R}$ ou le diagnostique de Gelman-Rubin critère principal nécessitant plusieurs chaînes (implémenté dans "coda" package) (Gelman and Shirley, 2011)
- Temps d'exécution

Outils actuellement utilisés par model4all

Performants, libres, adaptés aux biologistes :

- language de haut-niveau : R (GPL-2)
- documents dynamiques : rmarkdown (GPL3)
- interface performante : RStudio (AGPL3)

Collaboration efficace :

- gestion distribuée de versions : git (GPL2)

Communauté d'utilisateurs et contributeurs :

- hébergement facilitant les contributions : GitHub (gratuit)
- contributions sous licences CC BY-SA et AGPL3

Plan

Plan

Bref résumé du thème « génétique qauntitative »

- Contexte : décomposer la variance phénotypique à partir d'un "grand" nombre de génotypes (ex. 200) avec un "petit" nombre de phénotypes (ex. 1-5) et quelques répétitions (ex. 3 années) ; prédire les phénotypes
- Simulation : génotypes via le coalescent ; phénotypes via des modèles mixtes ("modèle animal", prédiction génomique)
- Inférence : rrBLUP, lme4, MCMCglmm, rjags, INLA, rstan, BGLR, np
- Evaluation : estimation, temps de calcul

Plan

Bref résumé du thème « observations corrélées »

- Contexte : prendre en compte des dépendances/dynamiques entre les observations (individuelles, spatiales, temporelles, spatiales et temporelles, etc)
- Simulations : simulation présentée ci-dessous, vecteur autoregressif d'ordre 1, modèle dynamique univarié d'ordre 1
- Inference : INLA, nlme, MCMCglmm, rstan
- Evaluation : estimation, prédiction, temps de calcul

Contexte

Type d'observations rencontré dans divers contextes :

- caractères mesurés à différents temps : croissance d'un arbre, d'un animal, séries temporelles, ...
- individus exposés à différents niveaux d'un même traitement : même plante exposée à différentes doses de produit

Étape 1 : simulation

Étape 1 : simulation

On a simulé selon le modèle suivant (modèle S) :

$$y_{it} = \mu + a_i + \varepsilon_{it}$$

avec $i \in \{1, \ldots, N\}$, $t \in \{1, \ldots, T\}$, $a_i \sim \mathcal{N}(0, \sigma_a^2)$ et

$$\varepsilon_{i1} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right) \qquad |\rho| < 1 \; \forall i$$
$$\varepsilon_{it} = \rho\, \varepsilon_{i(t-1)} + s_{it} \qquad s_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2(1-\rho^2)) \;\; , t = 2, \ldots, T$$

- données : $\mathcal{D} = \{y_{it}\}_{i=1,\ldots,N;t=1,\ldots,T}$
- paramètres : $\Theta = \{\mu, \rho, \sigma_\varepsilon^2, \sigma_a^2\}$
- vraisemblance : $\mathcal{L}(\Theta) = f(\mathcal{D}|\Theta)$

Essai avec $N = 200$, $T = 50$, $\mu = 5$, $\rho = 0.7$, $\sigma_\varepsilon^2 = 4$, $\sigma_a^2 = 9$.

Étape 2 : choix de la méthode d'inférence

Ce que l'on observe :

- une variable continue mesurée plusieurs fois sur un même individu
- une dépendance entre ces mesures faites sur un même individus
- nombre de données faible

Ce dont on a besoin :

- un modèle linéaire mixte

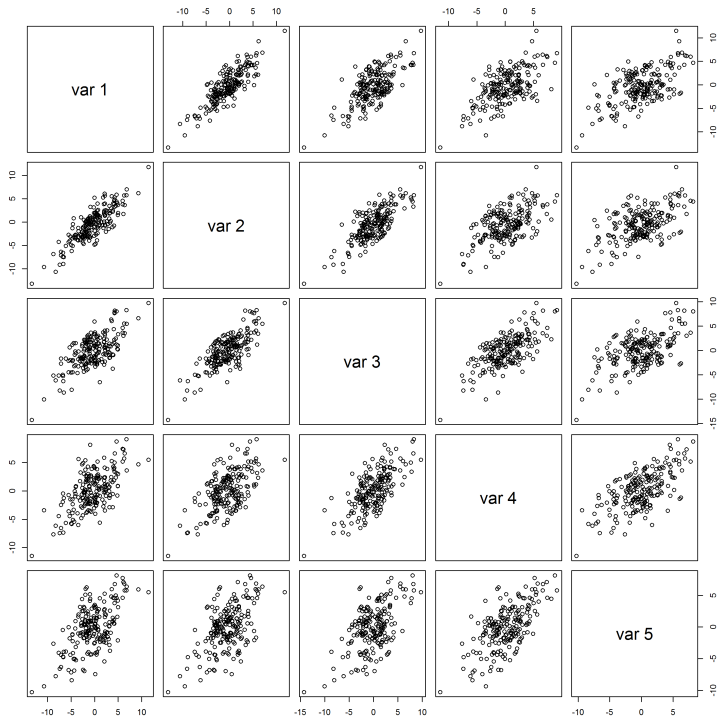
Étape 2 : choix de la méthode d'inférence

Plusieurs méthodes d'inférence possibles.

- package nlme du logiciel R $\Rightarrow$ simple d'utilisation, beaucoup de tutoriels
- mise en œuvre d'un modèle linéaire mixte avec un effet aléatoire « individus » (modèle I1)

Étape 2 : choix de la méthode d'inférence

Représentation des résidus en fonction du temps :



- besoin de prendre en compte une structure de dépendance
- comment ? obtient-on un meilleur ajustement ?

$\Rightarrow$ modélisation de la matrice de variance-covariance des erreurs via une structure auto-regressive d'ordre 1 (modèle I2)

Étape 3 : évaluation des deux modèles

Quels outils pour comparer les deux modèles d'inférence ?

- test de rapport de vraisemblance, AIC, BIC :

	df	AIC	BIC	logLik	Test	L.Ratio	p-value
modèle I1	3.00	42161.23	42182.86	-21077.62			
modèle I2	4.00	35925.85	35954.69	-17958.92	1 vs 2	6237.38	< 0.0001

- estimations et intervalles de confiance à 95% :

	μ	σ_a	ρ	σ_ε
modèle S	5	3	0.7	2
modèle I1	4.827(4.4-5.255)	3.074(2.784-3.394)		1.896(1.87-1.923)
modèle I2	4.805(4.379-5.231)	3.005(2.711-3.33)	0.709(0.693-0.724)	1.99(1.938-2.044)

Pour aller plus loin

J'ai des connaissances a priori sur mon paramètre ρ que je souhaite intégrer au modèle, par exemple de la structuration spatiale.

Quelle(s) méthode(s)/logiciel(s) d'inférence utilisés ?

- STAN (MCMC), INLA, car possibilité d'intégrer ce niveau supplémentaire d'information et pas de limite calculatoire...

Plan

Bref résumé du thème « réduction de dimension »

- Contexte : réduire la dimension d'un jeu de données observé dans le temps avec l'ACP Bayésienne (ex. extraire les temps-clés d'un phénomène agronomique)
- Simulations : simulation de fonctions temporelles en utilisant un processus gaussien
- Inference : rjags, rstan
- Evaluation : estimation avec écart-type de Monte Carlo (MCSE), temps de calcul

Plan

(Auto-)formation

- outils (R, rmarkdown, RStudio, git, GitHub)
- notions de vraisemblance (fiveMinuteStats)
- méthodes d'inférence (...)
- logiciels disponibles (lme4, INLA, JAGS, STAN, PyMC3, etc)

Débuter un nouveau thème

- combiner les deux premiers thèmes

$$y_{it} = \mu + a_i + \varepsilon_{it}$$

thème 2 avec $a_i \sim \mathcal{N}(0, \sigma_a^2)$, mais avec $a \sim \mathcal{N}(0, \sigma_a^2 A)$ si couplé avec le thème 1, où A est une matrice d'apparentement

- combiner des simulations type « écophy » (ex. via OpenAlea) et génétique (ex. contribution des « breeding values »), puis inférence avec modèle stat (vraisemblablement très simplifié par rapport au modèle de simul) le plus convenable possible

Oyez oyez ;-)

Merci pour votre attention...

... et n'hésitez pas à contribuer : le projet ne marchera que si vous participez, toute bonne volonté est donc plus que bienvenue !

https ://github.com/timflutre/model4all