Méthodes bayésiennes pour l'inférence de réseaux de gènes.

Florence Jaffrézic

INRA, Jouy-en-Josas

Applibugs, 11 Juin 2014

Plan

Plan

Parcours

1998 : ENSAI
Ecole Nationale de la Statistique et de l'Analyse de l'Information.

Jean-Louis Foulley : Cours de Modèles Mixtes.

1998 : DEA de Statistique Mathématique de l'Université Rennes I.

Stage à l'INRA avec Jean-Louis Foulley et Christèle Robert.
"Analyse de variables ordinales par un modèle mixte à seuils hétéroscédastique."

Parcours

1998 : ASC INRA.

2001 : Thèse "Statistical models for the genetic analysis of longitudinal data", Université d'Edimbourg, encadrée par Prof R. Thompson et Prof W.G. Hill.

2002 : Chargée de Recherche (CR2) à l'INRA de Jouy-en-Josas, dans le Département de Génétique Animale.

Parcours

Plus de 10 ans de collaborations avec Jean-Louis Foulley.

Co-encadrement de deux thèses :

- Guillemette Marot (2006-2009) :
"Modélisation statistique pour la recherche de gènes différentiellement exprimés et méta-analyse."

Prix de thèse de la Société Française de Biométrie (2010), chaire d'excellence INRIA-Université de Lille.

Parcours

- Andrea Rau (2007-2010) :

"Reconstruction de réseaux de gènes."

Collaboration avec Rebecca Doerge (Purdue University, USA).
Post-doc à l'INRIA d'Orsay avec Gilles Celeux.
Recrutée CR2 INRA à partir de Oct. 2011.

Parcours

Cours international d'une semaine, donné avec Jean-Louis Foulley en 2007 (60 participants) :

"New insights into mixed model methodology with applications to genomics and biostatistics".

Parcours

Participation à différents groupes de travail :

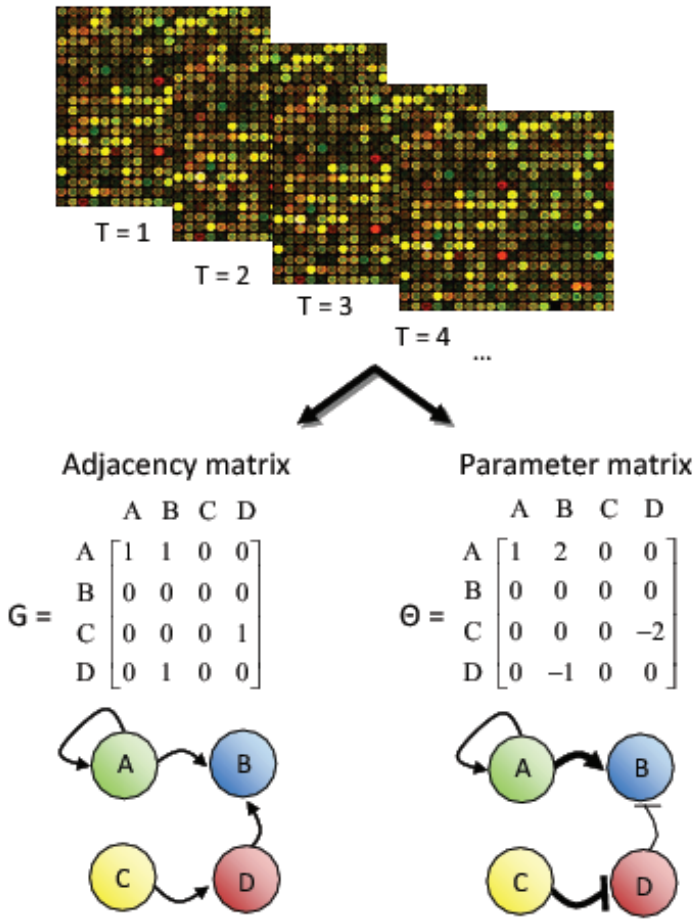
Monolix ("Modèles non linéaires à effets mixtes") : Université d'Orsay, INSERM, INRA.

Applibugs ("Applications bayésiennes utilisant le Gibbs sampling"), et Babayes.

Plan

Réseaux de gènes

- Transcritomique : Mesure de la quantité d'ARN messager de milliers de gènes simultanément.
- Objectif : Utiliser les données d'expression temporelles pour élucider les relations entre les gènes.



Approche EBDBN

Dans le cadre de la thèse d'Andrea Rau.

Inférence de réseaux de gènes à partir de données transcriptome temporelles.

Modèle étudié : Modèle espace-états ("state-space" model), s'inscrit dans le cadre des réseaux bayésiens dynamiques.

Approche EBDBN

Modèle espace-états considéré :

$$\begin{aligned} \boldsymbol{x_{t,r}} &= \boldsymbol{A x_{t-1,r}} + \boldsymbol{B y_{t-1,r}} + \boldsymbol{w_t} \\ \boldsymbol{y_{t,r}} &= \boldsymbol{C x_{t,r}} + \boldsymbol{D y_{t-1,r}} + \boldsymbol{z_t} \end{aligned}$$

$\boldsymbol{y_{t,r}}$ $(P \times 1)$: données d'expression observées pour les P gènes au temps t, pour le réplicat biologique r.

Résidus $\boldsymbol{w_t}$ et $\boldsymbol{z_t}$ supposés indépendants et normalement distribués : $\boldsymbol{w_t} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I_{K \times K}})$ et $\boldsymbol{z_t} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{V^{-1}} = diag(\boldsymbol{v})^{-1})$.

$\boldsymbol{v}$: vecteur de précisions $(P \times 1)$.

Approche EBDBN

Ce modèle permet :

- D'étudier les relations entre les observations aux temps t et $(t-1)$, comme dans un modèle autorégressif classique.

- De prendre en compte des états cachés $\boldsymbol{x_{t,r}}$ et $\boldsymbol{x_{t-1,r}}$ $(K \times 1)$: gènes ou facteurs de transcription non présents dans la liste étudiée mais ayant un rôle biologique important.

Matrice d'intérêt : $\boldsymbol{D}$ de dimension $(P \times P)$,
identifiable (Rangel et al., 2004).
Représente les relations entre les gènes observés.

Approche EBDBN

Méthode d'estimation bayésienne empirique pour estimer les paramètres :

L'algorithme se compose de trois étapes principales :

1) Choix du nombre d'états cachés K et estimation de ces états cachés par une approche filtre de Kalman.

2) Calcul des distributions a posteriori des matrices de coefficients.

3) Détermination des relations significatives entre gènes.

Approche EBDBN

1) Choix du nombre d'états cachés :

Approche basée sur la décomposition en valeurs propres de la matrice "block-Hankel" d'autocovariances entre observations (Rau et al., 2010).

Permet d'estimer directement le nombre d'états cachés.

Approche EBDBN

Une fois le nombre d'états cachés choisi, le filtrage et lissage de Kalman peuvent être utilisés pour estimer les valeurs de ces états cachés.

Les moyennes a posteriori $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{C}}$ et $\hat{\boldsymbol{D}}$ des matrices de coefficients sont utilisées dans ces équations.

Approche EBDBN

2) Calcul des distributions a posteriori.

Structure bayésienne hiérarchique.

$$
\begin{aligned}
\boldsymbol{a}_{(j)}|\boldsymbol{\alpha} &\sim \mathcal{N}(\mathbf{0}, diag(\boldsymbol{\alpha})^{-1}) \\
\boldsymbol{b}_{(j)}|\boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, diag(\boldsymbol{\beta})^{-1}) \\
\boldsymbol{c}_{(j)}|\boldsymbol{\gamma}, v_i &\sim \mathcal{N}(\mathbf{0}, v_i^{-1} diag(\boldsymbol{\gamma})^{-1}) \\
\boldsymbol{d}_{(j)}|\boldsymbol{\delta}, v_i &\sim \mathcal{N}(\mathbf{0}, v_i^{-1} diag(\boldsymbol{\delta})^{-1})
\end{aligned}
$$

$\boldsymbol{a}_{(j)}$, $\boldsymbol{b}_{(j)}$, $\boldsymbol{c}_{(j)}$ et $\boldsymbol{d}_{(j)}$ la $j$ème ligne des matrices de coefficients $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ et $\boldsymbol{D}$, respectivement.

Approche EBDBN

Jeu de paramètres : $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{V}\}$.

Jeu d'hyperparamètres : $\boldsymbol{\psi} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}\}$.

$\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_K\}$, $\boldsymbol{\beta} = \{\beta_1, ..., \beta_P\}$, $\boldsymbol{\gamma} = \{\gamma_1, ..., \gamma_K\}$,
$\boldsymbol{\delta} = \{\delta_1, ..., \delta_P\}$, $j = (1, ..., K)$ et $i = (1, ..., P)$.

Approche EBDBN

Comme on suppose des distributions Gaussiennes, on peut écrire explicitement la vraisemblance jointe :

$$p(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{V}, \boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{A}|\alpha)p(\boldsymbol{B}|\beta)p(\boldsymbol{V})p(\boldsymbol{C}|\boldsymbol{V}, \gamma)p(\boldsymbol{D}|\boldsymbol{V}, \delta)$$

$$p(\boldsymbol{x_0}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) \prod_{t=1}^{T} p(\boldsymbol{x_t}|\boldsymbol{x_{t-1}}, \boldsymbol{y_{t-1}}, \boldsymbol{A}, \boldsymbol{B}) p(\boldsymbol{y_t}|\boldsymbol{x_t}, \boldsymbol{y_{t-1}}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{V})$$

On note $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{V})$ le vecteur des données complètes et $\boldsymbol{w} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ le vecteur des données manquantes.

Approche EBDBN

Sachant les estimations courantes des états cachés ($\hat{\boldsymbol{x}}$) et des hyper-paramètres ($\hat{\boldsymbol{\psi}}$), on peut obtenir un nouvel estimateur ponctuel des hyper-paramètres $\boldsymbol{\psi}$ grâce à un algorithme EM (Expectation-Maximization) (Dempster et al., 1977).

Les distributions a posteriori des paramètres $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ et $\boldsymbol{D}$ peuvent ensuite être obtenues analytiquement car elles suivent toutes des distributions Gaussiennes.

Algorithme EBDBN

1) Choix de la dimension des états cachés. On initialise les valeurs des états cachés $\boldsymbol{x}$ et des hyper-paramètres ψ.

2) On utilise l'algorithme EM afin de mettre à jour l'estimation de ψ (cadre bayésien empirique, Carlin et Louis (2008)).

3) On obtient analytiquement les distributions a posteriori des paramètres $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ et $\boldsymbol{D}$.

Algorithme EBDBN

4) On utilise leurs moyennes a posteriori dans le filtre de Kalman pour mettre à jour les valeurs des états cachés $\boldsymbol{x}$.

5) On itère ensuite cet algorithme jusqu'à convergence et on base la construction du réseau de gènes sur la distribution a posteriori obtenue pour la matrice $\boldsymbol{D}$.

Algorithme EBDBN

Détection des arcs significatifs :

Comme les distributions a posteriori des éléments de $\boldsymbol{D}$ sont Gaussiennes, on peut calculer les Z-scores pour chaque arc, utiliser les seuils standards.

Les arcs dont les distributions sont très supérieures à zéro sont considérés comme des activations et ceux dont les distributions sont très inférieures à 0 comme des inhibitions.

Etude de simulation

Comparaison avec deux autres approches :

1) VBSSM (Variational Bayes State Space Model) (Beal et al., 2005), modèle espace-états, mais méthode d'estimation différente.

2) VAR (Vector Autoregressive model) (Opgen-Rhein et Strimmer, 2007), simple modèle autorégressif, sans états cachés.

Etude de simulation

Données simulées soit avec un processus autorégressif et en générant des états cachés, soit en utilisant les paramètres d'un jeu de données réel (Zak et al., 2003).

Plusieurs critères utilisés pour comparer ces différentes approches : aire sous la courbe ROC, sensibilité et spécificité.

Etude de simulation

Résultats :

Performances assez faibles pour toutes les méthodes quand peu de temps de mesure (5) et peu de réplicats par temps (5), même lorsque seulement 50 gènes sont analysés.

Performances considérablement améliorées lorsque le nombre de temps augmente à 10 ou 15, même pour 5 réplicats.

Pour un nombre de lames donné : préférable d'augmenter le nombre de temps de mesure plutôt que le nombre de réplicats biologiques par temps.

Etude de simulation

Approche VAR : moins performante que les autres méthodes.

Modèle EBDBN sans états cachés : moins performant que le modèle espace-états complet pour les petits jeux de données (T=5), mais aussi bon quand le nombre de temps et de réplicats augmente.

Etude de simulation

Approche VBSSM : un peu meilleure que les autres méthodes pour les différents critères considérés, surtout pour de petits jeux de données.

Elle nécessite cependant de grands temps de calcul.

Approche EBDBN sans états cachés : bon compromis entre performance et temps de calcul.

Méthode ABC-Net

Méthode EBDN : approche bayésienne empirique.

Fortes hypothèses de normalité.

Inférence par une approche ABC (Approximate Bayesian Computing).

Pas d'hypothèse restrictive sur les distributions des paramètres.

Seule condition : pouvoir simuler facilement des données selon le modèle proposé.

Méthode ABC-Net

Extension de l'algorithme ABC-MCMC (Marjoram, 2003) pour l'inférence de réseaux à partir de données d'expression temporelles.

$\mathbf{y_t} = (y_{1t}, ..., y_{Pt})'$ représente le vecteur des données observées pour P gènes au temps t.

La structure du réseau de gènes associé est déterminée par la matrice $\mathbf{\Theta} = (\theta_{ij})$ de dimension $(P \times P)$.

Méthode ABC-Net

θ_{ij} représente la relation entre le gène j au temps $(t-1)$ et le gène i au temps t.

Une valeur $\theta_{ij} > 0$ signifie que le gène j active le gène i, $\theta_{ij} < 0$ le gène j réprime le gène i et $\theta_{ij} = 0$ si le gène j ne régule pas le gène i.

Méthode ABC-Net

Objectif : obtenir un échantillon de valeurs simulées dans la distribution a posteriori approchée $f(\boldsymbol{\Theta}|\rho(\boldsymbol{Y}^*, \boldsymbol{Y}) \leq \epsilon)$.

Algorithme Métropolis-Hastings.

Paramètres $\boldsymbol{\Theta}^*$ sont proposés dans une distribution instrumentale $q(.|.)$ et sont utilisés pour simuler des données $\boldsymbol{Y}^*$ d'après le modèle $f(.|\boldsymbol{\Theta}^*)$.

Méthode ABC-Net

Les données simulées et observées sont comparées grâce à une fonction de distance $\rho(.)$ et un paramètre de tolérance ϵ.

Les paramètres proposés sont acceptés avec la probabilité :

$$\alpha = \min\left(1, \frac{\pi(\mathbf{\Theta}^*)q(\mathbf{\Theta}^{(r)}|\mathbf{\Theta}^*)}{\pi(\mathbf{\Theta}^{(r)})q(\mathbf{\Theta}^*|\mathbf{\Theta}^{(r)})}\mathbf{1}(\rho(\boldsymbol{Y}^*, \boldsymbol{Y}) < \epsilon)\right)$$

où $\mathbf{1}(.)$ est une fonction indicatrice et $\pi(.)$ est la distribution a priori de $\mathbf{\Theta}$.

Méthode ABC-Net

Sous certaines conditions de régularité (Marjoram, 2003), on peut montrer que la distribution stationnaire de la chaîne est $\pi(\boldsymbol{\Theta}|\rho(\boldsymbol{Y}^*, \boldsymbol{Y}) \leq \epsilon)$.

Si le paramètre ϵ est suffisamment petit, cette distribution sera une bonne approximation de la distribution a posteriori recherchée $\pi(\boldsymbol{\Theta}|\boldsymbol{Y})$.

Méthode ABC-Net

Modification de l'algorithme pour l'inférence de réseaux de gènes :

1) Méthode de simulation des données $\boldsymbol{Y}^*$ selon une structure de réseau connue définie par une matrice $\boldsymbol{\Theta}^*$.

2) Choix des distributions instrumentale $q(.|.)$ et a priori $\pi(.)$.

Méthode ABC-Net

Simulation des données : modèle autorégressif multivarié d'ordre 1 (VAR(1)), comme proposé par Beal et al. (2005) ; Opgen-Rhein et Strimmer (2007).

$\boldsymbol{y_1^* = y_1}$ et pour $t = 2, ..., T$, $\boldsymbol{y_t^* = \Theta^* y_{t-1}}$.

Remarque : Prédictions au temps t, $\boldsymbol{y_t^*}$: calculées à partir des valeurs observées $\boldsymbol{y_{t-1}}$ et non des données simulées $\boldsymbol{y_{t-1}^*}$.

Méthode ABC-Net

D'autres modèles plus complexes pourraient être considérés pour simuler les données : autorégressif d'ordre 2, modèle non-linéaire ou à équations différentielles.

La distribution instrumentale $q(.|.)$ définit la transition entre la structure courante du réseau et la nouvelle structure proposée.

Pour le choix de cette fonction, on introduit la matrice $\boldsymbol{G}$ de dimension $(P \times P)$ définie par : $G_{ij} = 1$ si le gène j régule le gène i et 0 sinon.
$G_{ij} = 0 \Leftrightarrow \theta_{ij} = 0$ et $G_{ij} = 1 \Leftrightarrow \theta_{ij} \neq 0$.

Méthode ABC-Net

Procédure en deux étapes pour proposer les nouvelles valeurs $\boldsymbol{G}^*$ et $\boldsymbol{\Theta}^*$.

1) Trois des déplacements de base (Husmeier et al., 2005) appliqués à la matrice courante $\boldsymbol{G}^{(r)}$: ajouter une interaction (changer un 0 en 1), enlever une interaction (changer un 1 en 0), changer la direction d'une interaction (si $G_{ij} = 1$ et $G_{ji} = 0$, on inverse ces valeurs).

Probabilité de transition pour la première étape est :
$q(\boldsymbol{G}^* | \boldsymbol{G}^{(r)}) = 1/N(\boldsymbol{G}^{(r)})$,
$N(\boldsymbol{G})$: taille du voisinage de la matrice $\boldsymbol{G}$.

Méthode ABC-Net

2) Distribution instrumentale de $\boldsymbol{\Theta}$ sachant la valeur courante $\boldsymbol{\Theta}^{(r)}$ et la matrice $\boldsymbol{G}^*$:

$$q(\theta_{ij}|\theta_{ij}^{(r)}, G_{ij}^*) \sim \mathcal{N}(\theta_{ij}^{(r)}, \sigma_\theta), \text{ si } G_{ij}^* \neq 0$$

et 0 sinon.

Le paramètre σ_θ peut être choisi de façon à avoir un taux d'acceptation entre 15% et 50%, comme recommandé par Gilks et al. (1996).

Méthode ABC-Net

Choix des distributions a priori pour les matrices $\boldsymbol{G}$ et $\boldsymbol{\Theta}$.

A priori non informatif uniforme pour $\pi(\boldsymbol{G})$ avec une contrainte sur le nombre de gènes régulateurs.

A priori uniforme pour $\pi(\boldsymbol{\Theta}|\boldsymbol{G})$ avec une contrainte sur les bornes de l'intervalle en prenant les valeurs -2 et 2, qui représentent déjà des forts effets d'activation et de répression.

Méthode ABC-Net

L'algorithme ABC-Net fournit des échantillons dépendants tirés dans la distribution $f(\boldsymbol{\Theta}, \boldsymbol{G} | \rho(\boldsymbol{Y}^*, \boldsymbol{Y}) \leq \epsilon)$.

Les tirages consécutifs étant assez corrélés, on ne garde que une itération sur 50. Comme suggéré par Geyer (1992), on prend une période de burn-in entre 1% et 2% du nombre total d'itérations n.

En pratique, 10 chaînes indépendantes de 10^6 itérations sont réalisées et la convergence est vérifiée par la statistique R de Gelman et Rubin (1992).

Etude de simulation

Plusieurs aspects de l'algorithme ABC-Net sont étudiés :

1) La fonction de distance ρ et le paramètre ϵ.

2) La sensibilité au choix des distributions a priori.

3) L'adéquation du modèle autorégressif VAR(1) utilisé pour simuler les données dans le cas d'un processus biologique plus complexe.

Méthode ABC-Net

Distance Euclidienne est bien adaptée pour cet algorithme.

Paramètre ϵ : compromis entre une valeur très faible afin d'obtenir une bonne approximation de la distribution a posteriori, et une valeur suffisamment grande pour garder un temps de calcul raisonnable.

$\Rightarrow$ Utiliser une approche adaptative pendant la période de burn-in en prenant une séquence décroissante de valeurs pour ϵ (Ratmann et al., 2007).

Méthode ABC-Net

Choix des bornes a priori pour la fonction $\pi(\boldsymbol{\Theta}|\boldsymbol{G})$: distributions a posteriori approchées plus diffuses quand intervalle plus large (par exemple (-10,10)) qu'avec un intervalle de (-2,2).

Quelque soit le choix de ces bornes, certaines distributions restent diffuses alors que d'autres présentent un pic plus étroit.

$\Rightarrow$ Certaines interactions entre gènes sont donc faciles à inférer, même avec un a priori très peu informatif, alors que d'autres restent difficiles à estimer même avec un intervalle a priori plus petit.

Méthode ABC-Net

Choix du simulateur VAR(1) : permet une première estimation des interactions entre gènes quand aucune information biologique n'est connue sur la dynamique du réseau.

Quand d'autres modèles (d'ordre 2 ou non linéaires) sont connus pour mieux représenter le phénomène biologique étudié, il est préférable de les utiliser pour simuler les données dans l'algorithme ABC-Net.

Méthode ABC-Net

Application à un jeu de données réelles sur le "S.O.S. DNA repair system" chez la bactérie *Escherichia coli* avec 8 gènes et 50 temps de mesure.

Pour le moment, le temps de calcul nécessaire pour l'approche ABC-Net limite son utilisation à l'inférence de réseaux à peu de gènes.

Amélioration possible de cet algorithme grâce à l'utilisation de nouvelles techniques de simulation proposées telles que Monte Carlo séquentiel (Del Moral et al., 2006 ; Robert, 2010).

Plan

Perspectives

Nouvelles technologies :

Données de séquençage haut débit RNASeq.

Permet d'avoir des mesures d'expression de tout le génome.

Même pour des gènes non encore identifiés.

Information plus complète et plus précise (niveau gène, transcrit, exon).

Données de comptage.

Données RNASeq

Thèse de Mélina Gallopin, en co-encadrement avec Gilles Celeux (INRIA, Orsay).

Inférence de réseaux de gènes partir de données RNA-seq.

Modèle de Poisson sur-dispersé.

Estimation d'effets causaux

Estimation d'effets causaux à partir de données d'expression et de données d'intervention (*knock-outs*).

Stage de Master de Gilles Monneret en co-encadrement avec Grégory Nuel (UPMC, Collège de France).

Remerciements

Merci à Jean-Louis pour toutes ces années de collaborations scientifiques très enrichissantes !