Modèles mixtes en génétique animale : sélection de variables par optimisation combinatoire

Julie Hamon, Clarisse Dhaenens, Gaël Even, Julien Jacques

Modèles mixtes en génétique animale : sélection de variables par optimisation combinatoire

Julie Hamon [1,2,3], Clarisse Dhaenens [1,2], Gaël Even [3] & Julien Jacques [1,4]

[1] *Inria Lille - Nord Europe - julie.hamon@inria.fr*
[2] *LIFL / Université de Lille 1 - clarisse.dhaenens@lifl.fr*
[3] *Gènes Diffusion Douai - g.even@genesdiffusion.com*
[4] *Laboratoire Paul Painlevé / CNRS & Université de Lille 1 - julien.jacques@inria.fr*

Résumé. En sélection génomique animale, un des enjeux consiste à identifier un sous-ensemble de marqueurs génomiques explicatifs pour un trait d'intérêt quantitatif. La spécificité des études animales nécessite l'utilisation de modèles mixtes, du fait des liens de parenté entre individus. Nous proposons d'effectuer, dans ce cadre, une sélection des marqueurs d'intérêt à l'aide de méthodes d'optimisation combinatoire.
Mots-clés. Modèle mixte, optimisation combinatoire, sélection de variables, sélection génomique.

Abstract. In the context of genomic selection in animal breeding, an objectif consists in identifying a subset of explicative markers for a quantitative trait under study. In order to take into account familial relationships between individuals, we need to use mixed models. We propose in this context to select interesting markers using combinatorial optimization.
Keywords. Mixed model, combinatorial optimization, feature selection, genomic selection.

1 Introduction

La sélection génomique consiste à évaluer les animaux à partir de leur ADN, basé sur un grand nombre de marqueurs couvrant l'ensemble du génome. Les marqueurs que nous étudions ici sont des variations de la séquence d'ADN sur une seule paire de bases, nommées SNPs (*Single nucleotide polymorphisms*). Un objectif important dans ce domaine est d'établir un modèle prédictif de traits d'intérêt (quantité de lait, qualité de la viande), utilisant l'information génomique. Étant donné le grand nombre de SNPs disponibles dans ces études (environ 54 000 SNPs pour les études de sélection animales), le challenge consiste à identifier un nombre réduit de SNPs, capable de prédire au mieux, pour un nouvel animal, la valeur du trait d'intérêt.

Dans un premier travail (Hamon, 2012), nous avons montré que l'utilisation d'algorithmes d'optimisation combinatoire, comme outil de sélection de variables en régression

linéaire, pouvait s'avérer être une alternative intéressante aux techniques usuelles. En effet, même si la complexité algorithmique induite peut être plus importante que pour des techniques usuelles de type lasso (Tibshirani, 1994), il s'avère que les capacités d'exploration des techniques d'optimisation combinatoire permettent parfois d'identifier des sous-ensembles de variables particulièrement pertinents. Nous proposons dans ce papier d'utiliser ces techniques dans le cadre de l'utilisation de modèles mixtes, permettant de prendre en compte les relations familiales importantes qui existent en génomique animale.

La section 2 présente le modèle ainsi que les techniques classiquement utilisées en génomique animale, puis notre approche est proposée en section 3. Des résultats expérimentaux illustrent l'intérêt de l'utilisation des techniques d'optimisation combinatoire (section 4), tandis que les perspectives de ce travail sont présentées dans la section 5.

2 Modélisation du problème

L'objectif de ce travail est d'établir un modèle prédictif pour un trait quantitatif (phénotype), tel que la production de lait ou la qualité de la viande, sur la base de marqueurs génétiques (ici des SNPs). Dans les données dont nous disposons, nous savons qu'il existe entre les animaux des relations de parenté, et ces dernières sont connues. Nous sommes donc face à un problème de régression dans lequel les individus ne sont pas indépendants, et nous souhaitons introduire un paramètre modélisant ces relations de parenté. Pour ce faire, nous utilisons un modèle mixte dans lequel les effets fixes sont les effets dûs aux SNPs et les effets aléatoires permettent de tenir compte des relations entre animaux. Ce que nous pouvons décrire par :

$$\mathrm{y} = \beta X + Zu + \epsilon, \tag{1}$$

où y est le trait étudié ($y_i \in \mathbb{R}$), X est la matrice des SNPs, Z est la matrice d'incidence des effets aléatoires, et $\epsilon \sim \mathcal{N}(0, R)$. Nous choisissons $x_{ij} \in \{0, 1, 2\}$, x_{ij} représentant alors le nombre d'allèles variants (Usai (2009)).
L'objectif est d'estimer le vecteur de paramètres β et de prédire la variable aléatoire $u \sim N(0, G)$. Henderson (1950) montre que les meilleurs estimateurs linéaires (resp. prédicteurs) sans biais de β (resp. de u) sont donnés par la résolution du système suivant :

$$\begin{pmatrix} X^t R^{-1} X & X^t R^{-1} Z \\ Z^t R^{-1} X & Z^t R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X^t R^{-1} y \\ Z^t R^{-1} y \end{pmatrix} \tag{2}$$

Les problèmes typiques auxquels nous nous intéressons ont entre 1 000 et 3 000 animaux pour environ 40 000 SNPs. Dans de telles situations, le système d'équation (2) ne peut être résolu. Pour répondre à ce problème, différentes méthodes sont utilisées en sélection génomique :

- dans une grande partie des travaux (Ogutu *et al.*, 2012), les relations familiales ne sont pas prises en compte, et des techniques de régressions pénalisées sont directement utilisées. Parmis ces techniques, ridge (Hoerl *et al.* (1970)) et lasso (Tibshirani (1994)) introduisent une pénalité dans le critère des moindre carrés :

$$\beta = argmin\{RSS(\beta) + \lambda||\beta||\} \tag{3}$$

où $||.||$ est la norme L^1 pour le lasso et L^2 pour ridge. Ridge permet en particulier de traiter les variables corrélées mais ne conduit pas à sélectionner des variables à l'instar du lasso. Pour combiner les avantages de ces deux méthodes, Zou et Hastie (2005) ont proposé une méthode de régularisation appelée elastic net (EN) combinant les pénalités lasso et ridge.

- Les méthodes de type GBLUP (VanRaden, 2008) passent outre le problème de la grande dimension en considèrant les effets marqueurs comme des effets aléatoires, de variance constante.

- Les méthodes bayésiennes (Meuwissen *et al.*, (2001)) font l'hypothèse d'une connaissance *a priori* sur la distribution des marqueurs. Par exemple, BayesA considère que tous les marqueurs ont une variance différente, alors que BayesB considère que certains marqueurs ont un effet nul. Ces méthodes sont plus complexes à mettre en œuvre.

L'approche que nous proposons, afin de prendre en compte les relations familiales et des effets marqueurs fixes, consiste à utiliser des techniques d'optimisation combinatoire pour sélectionner les variables d'intérêt. Ces techniques ont déjà montré leur efficacité dans le cadre du modèle linéaire (généralisé) appliqué à d'autres problématiques. Par exemple, dans un contexte de marketing, Meiri (2006) utilise un algorithme de recuit simulé combiné à l'utilisation du critère AIC (Akaike, 1974), tandis que Kapetanios (2007) applique un algorithme génétique couplé à un critère BIC (Schwarz, 1978). Dans un cadre bayésien, Hans *et al.* (2007) utilisent une recherche locale itérée dans laquelle les modèles sont évalués sur la base de leurs probabilités *a posteriori*, et ce dans un contexte d'analyse d'expression de gènes.

3 Approche proposée

Dans cette section nous présentons l'approche que nous proposons pour prédire un trait quantitatif en utilisant un nombre réduit de variables sélectionnées à partir d'un grand ensemble de variables possibles. Cette approche est basée sur un modèle mixte dans lequel une sélection de variables est réalisée à l'aide d'une recherche locale itérée.

3.1 Recherche locale itérée

Comme indiqué dans la publication de Corne *et al.* (2012), l'utilisation des méthodes d'optimisation est un moyen efficace de traiter les problèmes de sélection de variables. En effet, pour aborder les problèmes ayant un grand nombre de variables, les métaheuristiques (recherche locales, algorithmes évolutionnaires, ...) ont prouvé leur efficacité. Dans notre contexte de sélection de variables en régression, nous proposons d'utiliser une recherche locale itérée (ILS - Iterated Local Search), dans laquelle une solution est évaluée par des critères classiques d'évaluation de modèle de régression.
L'ILS est basée sur une succession de recherches locales et de perturbations (cf. Talbi (2009) pour une revue de ces techniques). Partant d'une solution initiale donnée, codée dans ce contexte sous forme de vecteur binaire, on applique une méthode de descente. Quand un optimum local est atteint, il est perturbé et la recherche locale repart de cette solution perturbée, jusqu'à ce que le critère d'arrêt de l'algorithme soit atteint. La solution initiale, qui correspond à un premier sous ensemble de variables, est générée aléatoirement tout en étant guidée par la corrélation de chaque variable avec le trait étudié. Nous avons choisi l'opérateur bit-flip pour le voisinage et un flip de plusieurs variables pour la perturbation. L'algorithme s'arrête lorsqu'un nombre maximum d'itérations est atteint.
De part leur popularité, certains auteurs comme Kapetanios (2007) ou Meiri *et al.* (2006) proposent d'utiliser un recuit simulé ou un algorithme génétique pour traiter le problème de sélection de variables. Cependant, l'ILS est plus simple, ce qui permet d'évaluer indépendamment la pertinence de l'approche combinatoire et de la fonction d'évaluation, et donne des résultats tout aussi intéressants que les méthodes citées précédemment.
Pour évaluer la qualité d'une solution (un sous ensemble de variables sélectionnées), nous calculons l'erreur de prédiction moyenne du modèle mixte défini sur cet ensemble de variables, comme expliqué dans la partie suivante.

3.2 Évaluation d'une solution

L'objectif des méthodes d'optimisation est d'explorer efficacement des espaces de recherche de grande dimension définis dans notre problème par tous les sous-ensembles possibles de variables. Pour cela, un critère d'évaluation (fitness) capable d'associer à chaque solution une mesure de qualité est utilisé ; il représente, in fine, l'objectif. Dans notre contexte, si le but est d'identifier le meilleur sous ensemble de variables, qui est donc celui qui générera le meilleur modèle prédictif, évaluer la qualité d'un tel sous ensemble n'est pas trivial et peut être discuté. En particulier, une difficulté bien connue est d'être capable d'évaluer la qualité du modèle sur des données indépendantes de celles ayant été utilisées pour l'estimation des paramètres du modèle (échantillon de validation). Dans un précédent travail (Hamon *et al.* (2012)), plusieurs critères d'évaluation ont été comparés (AIC, BIC, validation croisée), et nous avons retenu la validation croisée 3-fold.

4 Expérimentations et résultats

Pour évaluer la qualité de la méthode proposée, nous la testons dans un premier temps sur données simulées. Nous simulons donc des données suivant l'équation (1) où la matrice de variance-covariance des erreurs est $R = \sigma_e^2 I$ avec $\sigma_e^2 = 1$ et celle des effets aléatoires est $G = \sigma_u^2 A$ avec $\sigma_u^2 = 9$ et A est la matrice représentant les relations familiales entre animaux, construite à partir du pedigree. Les individus de X sont simulés suivant une loi normale de moyenne zéro et de matrice de variance-covariance A.
Un échantillon de 500 individus, répartis en 350 d'apprentissage et 150 de validation, est généré. Cinq jeux de simulation sont testés. La Figure 1 compare, en terme d'erreur moyenne de prédiction (MSEP), 3 méthodes : l'algorithme proposé avec évaluation par modèle mixte (MM), une méthode similaire proposée dans un précédent travail (Hamon *et al.* (2012)) qui s'appuie sur une évaluation par régression multiple (RM - qui ne prends pas en compte les relations familiales) et elastic net (EN), considérée comme une méthode de référence en sélection génomique (qui englobe ridge et lasso). Les deux premières méthodes étant stochastiques, sur chaque jeu, 10 exécutions sont lancées. L'erreur de prédiction moyenne a été normalisée par rapport à l'erreur de la méthode elastic net.

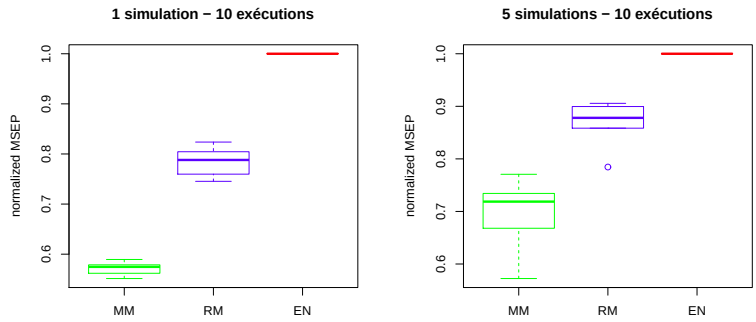

Figure 1: Comparaison des 3 méthodes

Cette simulation illustre l'intérêt de la prise en compte de la dépendance entre individus, les résultats de MM étant naturellement meilleur que RM et EN.

5 Conclusion

Dans cette étude nous abordons le problème de sélection de marqueurs explicatifs pour un trait quantitatif dans le contexte de sélection animale. Nous sommes face à un problème de régression en grande dimension ($n \ll p$) dans lequel nous devons intégrer une spécifité des données animales : les relations familiales entre animaux. Nous proposons de combiner une approche d'optimisation combinatoire avec un modèle mixte pour traiter ce problème.

Les premiers résultats sur simulation illustrent l'importance de prendre en compte les relations entre individus. Des résultats sur données réelles de la société Gènes Diffusion, entreprise spécialisée en génétique animale, seront présentés lors de la conférence.

Bibliographie

[1] H. Akaike (1974), *A new look at the statistical model identification.*, IEEE Transactions on Automatic Control, vol. 19, n. 6, pp. 716–723.
[2] D. Corne, C. Dhaenens, and L. Jourdan (2012), *Synergies between operations research and data mining: The emerging use of multi-objective approaches.*, European Journal of Operational Research, vol. 221, n. 3, pp. 469–479.
[3] J. Hamon, C. Dhaenens, G. Even and J. Jacques (2012), *Feature selection for high dimensional regression using local search and statistical criteria.*, International Conference on Metaheuristics and Nature Inspired Computing.
[4] C. Hans, A. Dobra and M. West (2007), *Shotgun stochastic search for "large p" regression.*, Journal of the American Statistical Association, vol. 102, n. 478, pp. 507–516.
[5] A. E. Hoerl and R. Kennard (1970), *Ridge regression: biaised estimation for nonorthogonal problems.*, Technometrics, vol. 12, pp. 55–67.
[6] G. Kapetanios (2007), *Variable selection in regression models using nonstandard optimisation of information criteria.*, Computational Statistics & Data Analysis, vol. 52, pp. 4–15.
[7] R. Meiri, J. Zahavi (2006), *Using simulated annealing to optimize the feature selection problem in marketing applications.*, European Journal of Operational Research, vol. 171, pp. 842–858.
[8] T. H. E. Meuwissen, B. J. Hayes and M. E. Goddard, *Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.*, Genetics Society of America, 2001.
[9] J. O. Ogutu, T. Schulz-Streeck, and H.-P. Piepho (2012), *Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions.*, BMC Proceedings, vol. 6, Suppl 2, S10.
[10] G. E. Schwarz (1978), *Estimating the dimension of a model.*, Annals of Statistics, vol. 6, n. 2, pp. 461–464.
[11] E.-G. Talbi (2009), editor. *Metaheuristics.*, John Wiley and Sons, Inc..
[12] R. Tibshirani (1994), *Regression Shrinkage and Selection Via the Lasso.*, Journal of the Royal Statistical Society, Series B, vol. 58, pp. 267–288.
[13] M. G. Usai, M. E. Goddard and B. J. Hayes (2009), *LASSO with cross-validation for genomic selection.*, Genet Res (Camb)., vol. 91, n. 6, pp. 427–36.
[14] P. M. VanRaden, *Efficient Methods to Compute Genomic Predictions.*, Journal of Dairy Science, vol. 91, n. 11, p. 4414–4423, nov.2008.
[15] H. Zou and T. Hastie (2005), *Regularization and Variable Selection via the Elastic Net.*, Journal of the Royal Statistical Society, Series B, vol. 67, Part 2, pp. 301–320.