Florian Rohart

Docteur en Statistiques Appliquées

10-12 Port St-Sauveur
31000 Toulouse, FRANCE
+33632968916
florian.rohart@gmail.com
http://florian.rohart.free.fr
28/11/1985 – Célibataire

Expériences professionnelles

2012– **Assistant Temporaire Enseignement-Recherche**, INSA Toulouse.

2009 – 2012 **Thèse**, *Prédiction phénotypique et sélection de variables en grande dimension dans les modèles linéaires et linéaires mixtes.*
Directrices de thèse : Béatrice Laurent-Bonneau, INSA Toulouse et Magali San-Cristobal, INRA Toulouse

Mars–Juillet 2009 **Stage de Master 2 Recherche**, *Zone de confiance et analyse de sensibilité*, réalisé à ACTIA-Toulouse, entreprise de diagnostic automobile, le stage a porté sur la détection de panne à partir de données recueillies sur un véhicule.

Formation

2008 – 2009 **Master 2 Recherche Mathématiques Appliquées en Aléatoire**, *Mention Assez Bien*, Formation en Processus Stochastiques : équations stochastiques, comportement en temps long ; et en Statistiques avec différents outils : ACP, SVM, Méthodes de seuillage, formule de Rice, méthode des records, Méthode Delta, M- et Z-estimateurs,. . . .
Université Paul Sabatier, Toulouse

2006 – 2008 **Master 1 et Licence de Mathématiques Fondamentales**, *Mention Assez Bien* ($10^{éme}/61$ *et* $9^{éme}/76$), Université Paul Sabatier, Toulouse.

2003 – 2006 **Classes Préparatoires aux Grandes Ecoles MPSI/MP**.
lycée Georges Clémenceau, Reims

Juin 2003 **Baccalauréat Scientifique**, *Mention Bien*, Lycée Gay-Lussac, Chauny.

Publications

2012 **Fixed Effects Selection in High Dimensional Linear Mixed Models**, Rohart F., B. Laurent and M. SanCristobal, *Soumis*.

Phenotypic Prediction Based on Metabolomic Data on the Growing Pig from three main European Breeds., Rohart F., Paris A., Laurent B., Canlet C., Molina J., Mercat M.J., Tribout T., Muller N., Iannucelli N., Villa-Vialaneix N., Liaubet L., Milan D. and San Cristobal M., *Accepté pour publication dans Journal of Animal Science*.

2011 **Multiple Hypotheses Testing For Variable Selection.**, Rohart F., *Soumis*.

Communications

Rohart F. Variable Selection in High Dimensional Linear Mixed Model through ℓ^1 Penalization. Dixième Colloque *"Jeunes Probabilistes et Satisticiens"*, 16-20 Avril 2012, CIRM Marseille, France

Rohart F. Multiple Hypotheses Testing For Variable Selection.Travail présenté à *"Statistical Methods for Post-Genomic Data"* les 26-27 Janvier 2012 à Lyon, à *"Statistiques Mathématiques et Applications"* 28 Aout - 2 Septembre 2011 à Fréjus, ainsi qu'au *"Séminaire des Doctorants du Département de Génétique*

Animale de l'INRA" 5-6 Avril 2011 à Limoges

Rohart F., C. Canlet, N. Villa-Vialaneix, J. Molina, D. Milan, B. Laurent, A. Paris and M. SanCristobal (2011) Feature selection in 1H NMR-based metabolomic profiles enables the prediction of growth phenotypes in various pig breeds and highlight few metabolites involved in this complex trait. *5èmes Journées Scientifiques du Réseau Français de Métabolomique et Fluxomique*, 23 au 25 mai 2011, Jussieu, France

Rohart F., N. Villa-Vialaneix, A. Paris, C. Canlet, J. Molina, D. Milan, B. Laurent and M. SanCristobal (2010) Phenotypic prediction based on metabolomic data: lasso vs Bolasso, primary data vs wavelet data. *9th World Congress on Genetics Applied to Livestock Production*, 1-6 Août 2010, Leipzig, Germany.

Organisations

Membre du comité d'organisation des Journées Statistiques du Sud, 20-22 Juin 2012, Toulouse, France.

Membre du comité d'organisation du quinzième Séminaire des Doctorants du Département de Génétique Animale, 26-27 Mars 2012, Ile d'Oléron, France.

Encadrements-Enseignements

2012 **Intervenant dans la formation "Statistiques pour l'analyse de données post-génomiques haut-débit"**, *Formation Génotoul/Interbio - Génopole Toulouse Midi-Pyrénées*, Formation qui s'adresse à des utilisateurs des statistiques ne possédant pas de formation particulière en mathématiques mais souhaitant acquérir une bonne autonomie dans l'analyse de leurs données.

2011 **Co-encadrant d'un projet (3mois) puis d'un stage (3mois) de Master 1**, *Analyse de données transcriptomique*, INSA Toulouse, INRA Toulouse.

2009–2012 **Enseignant de Travaux Dirigés de Mathématiques de Licence 1 à Licence 3**, *Modélisation Statistique, Probabilités et Statistiques, Calcul Différentiel et Calcul Intégral,...*, INSA Toulouse .

Compétences

Langues

Français	**Courant**
Anglais	**lu, écrit, parlé**
Espagnol :	**lu, écrit**

Informatique

Système	Windows, Linux, Mac
Logiciels	Maîtrise de R, Matlab, Maple, LaTeX,...
Language	Notions de C/C++, apprentissage java

Centre d'intérêt

Sport	Football en club, tennis, randonnée (Mare e Monti, Corse, Sept 2012)
Photo	Photographe amateur *http://florian.rohart.free.fr/Photos*
Activités culturelles	Cinéma -amateur de vieux films tels Metropolis, Forbidden Planet-, théâtre, opéra, lecture en langue anglaise ou française de roman policier, thriller, fantastique -The Farseer Trilogy-
Musique	Apprentissage de la guitare
Autre	Chef dans un groupe scout pendant trois ans

Résumé du travail de recherche

Les travaux de recherche ont été effectués dans le cadre de la thèse "Prédiction phénotypique et sélection de variables en grande dimension dans les modèles linéaires et linéaires mixtes". Cette thèse possède un **double encadrement** INRA/INSA, cette double étiquette permet de toujours avoir en tête la relation entre statistiques théoriques et statistiques appliquées.
Les deux objectifs principaux étaient de prédire un phénotype d'intérêt économique ainsi que d'expliciter des marqueurs biologiques, le tout à partir de données métabolomiques. La méthode la plus classique répondant à ce double objectif est la méthode Lasso (Tibshirani, 1996) : pénalisation ℓ^1 d'un critère des moindres carrés qui permet de faire de la sélection de variables dans un modèle linéaire. Il est important de noter que **jamais une étude aussi approfondie sur autant d'animaux n'avait vu le jour avant que je n'étudie ces données.** Ces travaux ont fait l'objet du papier Rohart et al. (2012b) dans lequel l'analyse est étendue sur tous les phénotypes disponibles et dans lequel on donne une interprétation biologique des résultats obtenus. **On montre que le métabolome a un réel pouvoir prédictif de certains phénotypes de production.**
Cependant, l'utilisation de la méthode Lasso soulève plusieurs problèmes en pratique (stabilité, choix de la pénalité). J'ai développé des nouvelles méthodes de sélection de variables dans un modèle linéaire $Y = X\beta + \epsilon$, dans le cas de la grande dimension (le nombre de variables explicatives p est plus grand que le nombre d'observations n). Ces méthodes sont des procédures séquentielles de tests multiples basées sur une procédure développée par Baraud et al. (2003). **J'ai développé une méthode pour la sélection ordonnée, et une pour la sélection non ordonnée**. Ces deux méthodes sont expliquées dans Rohart (2011), elles donnent de très bons résultats de simulations et **elles possèdent des résultats théoriques non asymptotiques valable en grande dimension.** Les simulations ont été effectuées à l'aide du R-package *mht* que j'ai développé (http://cran.r-project.org/). La méthode a aussi été utilisée pour construire des réseaux, et donne de bons résultats en simulations.
Afin d'inclure les effets de l'environnement et des liens de parenté entre les animaux comme des effets aléatoires, la sélection de variables dans un modèle linéaire mixte a été envisagé. Peu de méthodes existantes permettent de sélectionner des variables dans un modèle linéaire mixte, la plus performante est la méthode lmmLasso (Schelldorfer et al., 2011), qui est une pénalisation de la log-vraisemblance du modèle marginal $Y = X\beta + \epsilon$ où $\epsilon \sim \mathcal{N}(0, V)$. Cette méthode étant longue en temps de calcul sur les données de l'INRA, **J'ai développé une nouvelle méthode de sélection de variables dans un modèle mixte de grande dimension** plus rapide, cf. (Rohart et al., 2012a). Basé sur un algorithme de type multicycle ECM, la méthode est implémenté dans le R-package *MMS* (http://cran.r-project.org/).

Références

Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptative test of linear hypotheses by model selection. *The Annals of Statistics*, 31(1):225–251.

Rohart, F. (2011). Multiple hypotheses testing for variable selection.

Rohart, F., Laurent, B., and SanCristobal, M. (2012a). Fixed effects selection in high dimensional linear mixed model.

Rohart, F., Paris, A., Laurent, B., Canlet, C., Molina, J., Mercat, M., Tribout, T., Muller, N., Iannucelli, N., Villa-Vialaneix, N., Liaubet, L., Milan, D., and San Cristobal, M. (2012b). Phenotypic prediction based on metabolomic data on the growing pig from three main european breeds.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38:197–214.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B 58(1):267–288.