Sarah Lemler

http://www.math-evry.cnrs.fr/members/slemler/welcome

Exposé des activités d'enseignement et de recherche

Activités d'enseignement

Dans ce qui suit je présente rapidement les enseignements que j'ai assurés lors de ma thèse et de mon demi-poste d'ATER à l'Université d'Évry Val d'Essonne.

Enseignements en Licence de Biologie

L'objectif de ces cours est de familiariser les étudiants avec les bases mathématiques et les outils statistiques utiles aux biologistes. De ce point de vue, les TD dans les enseignements de Licence permettent de mettre en application les notions théoriques vues en cours, à partir d'énoncés concrets inspirés d'expériences en biologie.

2011-2012 **Analyse**

formation	Licence 1 de Biologie, Université d'Évry
charge	24h de TD
contenu	[continuité, dérivation, intégration, développements limités, équations différentielles]

2011-2014 **Probabilités et statistiques - Variables aléatoires continues**

formation	Licence 2 de Biologie, Université d'Évry
charge	126h de TD (6 groupes de 19,30h et 1 groupe de 9h pendant mon demi-ATER)
contenu	[rappels de probabilités, variables aléatoires réelles, grandeurs caractéristiques, loi normale, estimation, tests d'hypothèses]

2015 **Probabilités et statistiques - Variables aléatoires discrètes**

formation	Licence 1 de Biologie
charge	24h de TD (3 groupes de 8h)
contenu	[Variables aléatoires discrètes]

2015 **Introduction aux chaînes de Markov**

formation	Licence 3 Génie Biologique et Informatique
charge	6h (3h de cours et 3h de TP)
contenu	[rappels de probabilités, chaînes de Markov]

Enseignements niveau Master 2 à l'ensiie

J'ai également été chargée d'un cours associé d'un TP de Modèle de durées en 3ème année (niveau Master 2) à l'École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (admission à partir de la banque de concours Centrale-SupElec). L'articulation Cours/TP dans cet enseignement a permis de mettre en pratique les aspects théoriques et mathématiques vus en cours pour l'analyse pertinente d'un jeu de données. Pour cet enseignement, j'ai élaboré le cours, les TD, j'ai conçu des sujets d'examens et trouvé des jeux de données réelles pour des projets que les étudiants devaient produire. J'ai fait évoluer la structure du cours année après année pour améliorer sa progressivité et son attractivité.

2012-2014 **Modèles de durées**

formation	3ème année d'école d'ingénieur (niveau Master 2)
charge	63h (10,30h de cours et 10,30h de TP sur R pendant 3 ans)
contenu	[généralités sur l'analyse de survie, analyse de données censurées, estimateur de Kaplan-Meier, modèle de Cox, sélection de variables]
réalisation	cours, sujets de TP et d'examen, projet sur un jeu de données réelles

Enseignements à l'école doctorale GAO

Enfin, j'ai été chargée d'un enseignement dans le cadre de la formation doctorale de l'École doctorale des Génomes Aux Organismes à des doctorants issus de divers laboratoires de mathématiques ou de biologie.

2015 **Bases statistiques pour la biologie - Initiation à R**

formation	Formation doctorale
charge	12h (6h de cours et 6h de TP)
contenu	[base des statistiques, initiation au logiciel R]

Activités de recherche

Contexte et cadre général

Dans mes travaux de recherche, j'ai considéré le cadre général des processus de comptage qui inclut plusieurs contextes tels que les données censurées, les processus de Poisson et les processus de Markov. Considérons donc un processus de comptage N_i et Y_i un processus prévisible à valeurs dans $[0,1]$ qui complète l'information sur les observations, et on suppose que le processus N_i satisfait un modèle à intensité multiplicative d'Aalen, de sorte que son compensateur Λ_i vérifie

$$\Lambda_i(t) = \int_0^t \lambda_0(s, \boldsymbol{Z_i}) Y_i(s) ds,$$

où λ_0 est l'intensité du processus. Dans mes travaux, je me suis principalement intéressée à la prédiction de la durée de survie, quand le nombre de covariables p est possiblement supérieur à la taille de l'échantillon n. Cette prédiction nécessite alors l'estimation de l'intensité λ_0.

Dans le cas particulier du modèle de Cox, l'intensité est définie par

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \exp(\boldsymbol{\beta_0^T Z_i}), \tag{1}$$

où le risque de base $\alpha_0 : \mathbb{R}^+ \to \mathbb{R}^+$ et le paramètre de régression $\boldsymbol{\beta_0} \in \mathbb{R}^p$ sont inconnus. Dans mes travaux, je me suis intéressée à l'estimation de l'intensité dans le cas général et dans le cas particulier du modèle de Cox.

Mes contributions

[Estimation de l'intensité générale en grande dimension]

Dans un premier travail, je me suis intéressée à l'estimation globale de l'intensité $\lambda_0 : \mathbb{R}^+ \times \mathbb{R}^p \to \mathbb{R}^+$ sur laquelle je n'ai imposé aucune forme et j'ai considéré des covariables en grande dimension, c'est-à-dire lorsque $p > n$. J'ai approché λ_0 par une fonction de la forme

$$\lambda(t, \boldsymbol{Z_i}) = \alpha(t) \exp(f(\boldsymbol{Z_i})), \quad i = 1, ..., n,$$

J'ai cherché à estimer λ_0 par le meilleur modèle de Cox étant donné deux dictionnaires de fonctions. Le premier dictionnaire est utilisé pour construire une approximation du logarithme du risque de base et le second pour approximer le risque relatif. J'ai proposé une procédure Lasso pondéré, spécifique à la grande dimension, appliqué à la vraisemblance empirique pour estimer simultanément les deux paramètres inconnus du meilleur modèle de Cox approximant l'intensité. J'ai établi des inégalités oracles non-asymptotiques pour l'estimateur obtenu en divergence de Kullback empirique, qui est la fonction de perte la plus appropriée à notre procédure, ainsi qu'en norme empirique pondérée. Mes résultats reposent sur une inégalité de Bernstein pour les martingales à sauts et sur des propriétés des fonctions self-concordantes. Ce travail fait l'objet d'un proceeding publié dans *ESAIM* en 2014 [1] et d'un article accepté aux *Annales de l'Institut Henri Poincaré* en 2014 [2].

[Estimation dans le modèle de Cox en grande dimension]

Dans une deuxième partie de mon travail, j'ai considéré le modèle de Cox (1) pour des covariables en grande dimension. Je me suis intéressée à l'estimation des deux paramètres. Pour cela, j'ai procédé en deux étapes. La grande dimension étant portée par le vecteur de covariables, j'ai estimé dans une première étape le paramètre de régression à l'aide d'une procédure Lasso, spécifique à la grande dimension. Puis dans une seconde étape, j'ai proposé

deux procédures d'estimation non-paramétriques du risque de base α_0. La première procédure utilise la sélection de modèles appliquée à un critère type moindres carrés qui fait intervenir l'estimateur Lasso de $\boldsymbol{\beta_0}$ obtenu à la première étape. La deuxième procédure est basée sur un estimateur à noyau du risque de base, qui dépend aussi de l'estimateur Lasso de $\boldsymbol{\beta_0}$, avec un choix adaptatif de la fenêtre par la méthode de Goldenshluger et Lepski. Pour les deux estimateurs de α_0 ainsi obtenus, j'ai établi les premières inégalités oracles non-asymptotiques. Ce sont aussi les premiers résultats énoncés dans le cadre de covariables en grande dimension, qui permettent donc de mesurer l'influence de la grande dimension sur l'estimation du risque de base. J'ai ensuite effectué un gros travail de simulation : j'ai généré des données simulées dans le cadre de données censurées, j'ai considéré différentes procédures d'estimation de $\boldsymbol{\beta_0}$ et j'ai implémenté les procédures d'estimation de α_0 afin de vérifier leurs performances pratiques. J'ai enfin appliqué ces procédures à un jeu de données réelles sur le cancer du sein. Ces résultats font l'objet de deux articles co-écrits avec A. Guilloux et M-L. Taupin, un sur la sélection de modèles [3] soumis et un sur l'estimateur à noyau avec une fenêtre sélectionnée par la méthode de Goldenshluger et Lepski [4], qui sera soumis prochainement.

Travail de recherche en cours

J'ai commencé à travailler, en collaboration avec M. Alaya et A. Guilloux [5], sur une généralisation classique et très utilisée en pratique, qui consiste à considérer que le paramètre de régression du modèle de Cox dépend du temps. L'idée est d'estimer chacune des coordonnées de $\boldsymbol{\beta_0}(t)$ par des histogrammes pénalisés par leurs variations totales.

Publications et communications

Publications

Publication dans une conférence avec actes

2014 [1] Chazottes, J.R., Cuny, C., Dedecker, J., Fan, X., Lemler, S. *Limit theorems and inequalities via martingale methods*, ESAIM: Proceedings (MAS), vol.44, 177-196.*

Publication dans un journal international

2014 [2] Lemler, S. *Oracle inequalities for the Lasso in the high-dimensional Aalen multiplicative intensity model*, accepté pour la publication aux Annales de l'Institut Henri Poincaré.*

Publication soumise dans un journal international

2015 [3] Guilloux, A., Lemler, S., Taupin, M.L. *Adaptive estimation of the baseline hazard function in the Cox model by model selection, with high-dimensional covariates*, soumis, https://hal.archives-ouvertes.fr/hal-01120683*

Travaux en fin de rédaction

2015 [4] Guilloux, A., Lemler, S., Taupin, M.L. *Adaptive kernel estimation of the baseline function in the Cox model, with high-dimensional covariates.**

(*) Travaux qui seront soumis en cas d'audition.

Travail en cours de rédaction

[5] Guilloux, A., Alaya, M., Lemler, S. *High-dimensional Aalen and Cox model with change-points*

Séjour à l'étranger

24-26 avril 2015 Invitée dans l'équipe Statistics and Probability Theory, Department of Mathematical Sciences, University of Copenhagen, Danemark

Communications orales

Conférences internationales

1-7 sept. 2013 StatMathAppli 2013, Fréjus
7-20 juillet 2013 43ème école d'été de Saint-Flour
9-10 janvier 2012 The Statistical Analysis of Multi-outcome data, Université Pierre et Marie Curie

Conférences nationales

6-11 avril 2014	Onzième colloque Jeunes Probabilistes et Statisticiens, à Forges-les-eaux
29-31 août 2012	Journées MAS à Clermont-Ferrand

Séminaires

25 mars 2015	Séminaire de statistique à l'Université de Copenhague
17 mars 2015	Séminaire de statistique à l'ENSAI, à Rennes
9 mars 2015	Séminaire de statistique et imagerie à l'Université Paris Dauphine
3 mars 2015	Séminaire Statistique IMT à Toulouse
6 février 2015	Séminaire de statistiques à Rennes 2
19 mai 2014	Séminaire Modélisation statistique à Strasbourg
24 mars 2014	Séminaire de probabilités et statistiques à Montpellier
28 janvier 2014	Séminaire TEST, Telecom ParisTech
27 novembre 2012	Séminaire de Statistiques de Strasbourg, IRMA
14 février 2012	Statistics for Systems Biology, Evry
24 janvier 2012	Groupe de Travail des Thésards, LSTA, Université Pierre et Marie Curie, Paris

Activités administratives et responsabilités collectives

Membre élue au conseil du Laboratoire Statistique et Génome pour l'année 2012-2013
Activité de review pour Journal of Multivariate Analysis
Animatrice MATh.en.JEANS au Lycée du Parc des Loges, Évry