Contribution à la sélection de modèle via pénalisation Lasso en Épidémiologie

Marta Avalos Fernandez

université de BORDEAUX

Habilitation à Diriger des Recherches

UNIVERSITÉ DE BORDEAUX

École doctorale Sociétés, Politique, Santé Publique
Spécialité Santé Publique, option Biostatistique

Par Marta Ávalos Fernández

Contribution à la sélection de modèle *via* pénalisation Lasso en Épidémiologie

Contribution to model selection *via* Lasso penalization in Epidemiology

Soutenue le 11/12/2018

Membres du jury

Calle Rosingana Malu	Professeure, Université de Vic - Université Centrale de Catalogne (Vic, Espagne)	Rapporteure
Canu Stéphane	Professeur, INSA de Rouen (Rouen, FR)	Rapporteur
Delcourt Cécile	DR, INSERM (Bordeaux, FR)	Examinatrice
Lagarde Emmanuel	DR, INSERM (Bordeaux, FR)	Invité
Saracco Jérôme	Professeur, Institut Polytechnique de Bordeaux (Bordeaux, FR)	Examinateur
Thiébaut Rodolphe	Professeur, Université de Bordeaux (Bordeaux, FR)	Garant
Tubert-Bitter Pascale	DR, INSERM (Paris, FR)	Rapporteure

Table des matières

Avant-Propos

J'aurais dû vous le préciser :
Les Trolls sont sans cesse en contradiction avec eux-mêmes
Jim Henson, *The storyteller*, 1988

Ce document présente des travaux de recherche que j'ai menés après l'obtention de mon doctorat en 2004 et mon recrutement à l'Université de Bordeaux en 2005. J'ai choisi de présenter ces travaux qui se situent dans le cadre de l'apprentissage statistique à la lumière des études épidémiologiques qui les ont motivés, en mettant en évidence les contextes et les applications pratiques qui en sont faites.

Mes activités de recherche se sont déroulées initialement dans l'équipe Biostatistique dirigée par Daniel Commenges et ensuite par Hélène Jaqmin-Gadda au sein du Centre Inserm U897 "Epidémiologie et Biostatistique" dirigé par Roger Salamon. Suite à la dernière récréation en janvier 2016 du centre Inserm U1219 "Bordeaux Population Health" dirigée par Christophe Tzourio, l'équipe Biostatistique a été scindée en deux. J'ai ainsi rejoint l'équipe dirigée par Rodolphe Thiébaut "Statistiques pour la biologie systémique et la médecine translationnelle" (*Statistics In Systems biology and Translational Medicine*, SISTM) et plus particulièrement son axe "Données de grande dimension". L'équipe SISTM a une double tutelle INRIA (depuis 2014) et INSERM (depuis 2016).

L'équipe SISTM se consacre à l'élaboration de méthodes statistiques pour l'analyse intégrative des données de la médecine et de la biologie. Grâce aux progrès techniques, la recherche clinique et biologique génère des quantités très importantes de données. D'autre part, l'appariement des bases de données médico-administratives avec d'autres registres permet d'envisager la mise en œuvre de grandes études épidémiologiques, possiblement avec un suivi au cours du temps. Le défi consiste à analyser ces *Big Data* en utilisant des méthodes statistiques pour apporter des réponses appropriées aux questions posées par les épidémiologistes ou les cliniciens. La double tutelle de SISTM se traduit par des objectifs de recherche et d'application à la fois en informatique-mathématiques et en biostatistique-médecine, ce qui comprend l'épidémiologie et la recherche clinique.

Mes activités doivent beaucoup à des collaborations fructueuses avec des collègues du Centre Inserm, notamment avec les membres de l'équipe "Prévention et prise en charge des traumatismes" (*Injury epidemiology, transport, occupation*, IETO) dirigée par Emmanuel

Lagarde, aux travaux réalisés par des étudiants dans le cadre de leur stage de Master 2, notamment les étudiantes qui sont restées au sein du centre en thèse de doctorat ou en tant qu'ingénieurs statisticiennes. J'espère parvenir à mettre en évidence ce travail collectif tout au long de ce mémoire.

Ma principale contribution consiste à adapter des méthodes de l'apprentissage statistique supervisé qui sont devenues très populaires lors de la dernière décennie, les régressions pénalisées de type Lasso, à l'analyse de données issues d'études épidémiologiques. L'enjeu est de s'attaquer aux problèmes des Big Data tout en respectant les objectifs et spécificités de la discipline.

Afin de présenter mes travaux de façon cohérente, je me suis focalisée sur la recherche sur les méthodes de pénalisation. Les recherches portant sur le développement ou application d'autres méthodes statistiques ne sont pas développées ici. Plus précisément :
- Le 1er chapitre est une introduction générale dans laquelle je contextualise, motive et énonce la problématique abordée tout au long de mes recherches.
- Le 2ème chapitre est consacré à mes travaux en lien avec les études autour des traumatismes accidentels et médicaments à partir des données du système national des données de santé. Ces études ont été réalisées dans le cadre de la Plateforme académique de pharmacoépidémiologie *Drugs Systematized Assessment in real-liFe Environment* (DRUGS-SAFE), en collaboration, dans sa grande majorité, avec l'équipe IETO, mais aussi (pour une thématique non développée ici) avec les équipes Biostatistique, "Médicament et santé des populations" (PhEpi) et "Expositions vie entière, santé, vieillissement" (LEHA) du centre INSERM Bordeaux Population Health, ainsi que le groupe de recherche *Valencia Bayesian Research group* (VaBaR) du département de statistique et recherche opérationnelle de l'Université de Valencia, Espagne.
- Le 3ème chapitre est consacré à mes travaux en lien avec les études autour de données biomédicales. Tout d'abord, la prédiction de la charge virale censurée par un seuil de détection, à partir des mutations du VIH, chez des patients atteints du VIH, en collaboration avec l'équipe "VIH, hépatites virales et comorbidités : épidémiologie clinique et santé publique" (MORPH3Eus) du centre INSERM Bordeaux Population Health. Ensuite, l'automatisation de la détection des seuils d'anomalie des hémogrammes en population générale, en collaboration avec le département des laboratoires cliniques de l'école de médecine de la Pontificia Universidad Católica, Santiago, Chili (collaboration initiée avec Marcela Henríquez Henríquez lors de son séjour de recherche à l'Australian National University et de mon séjour au CSIRO, à Canberra).

Cette activité de recherche a fait l'objet de diverses publications, collaborations, encadrements, participation à des projets financés et séjours de recherche qui sont détaillés dans le curriculum vitae en fin de document.

Bibliographie

Abrahamowicz M., MacKenzie T., and Esdaile J. Time-dependent hazard ratio : modelling and hypothesis testing with application in lupus nephritis. *JASA*, 91 : 1432–9, 1996. 64

Agresti A. and Min Y. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Stat. Med.*, 23 : 65–75, 2004. 55

Akerstedt T., Bassetti C., Cirignotta F., García-Borreguero D., Gonçalves M., Horne J., Léger D., Partinen M., Penzel T., and Philip V. J. C. P. La somnolence au volant. Technical report, INSV et ASFA, 2013. 39

Al-Agha A., Faris H., Hammo B., and Al-Zoubi A. Identifying β-thalassemia carriers using a data mining approach : The case of the Gaza Strip, Palestine. *Artif Intell Med.*, 88 : 70–83, 2018. 101

Alhamzawi R. Bayesian elastic net tobit quantile regression. *Communications in Statistics-Simulation and Computation*, 45(7) : 2409–2427, 2016. 85, 94, 95

Alioum A., Jutand M., Leffondré K., Joly P., Letenneur L., Le Goff M., and Avalos M. Master sciences, technologies, santé, mention santé publique, 2017-2018 : Méthodes de régression. Technical report, Université de Bordeaux, ISPED, Bordeaux, France, 2018. 9

Amato U., Antoniadis A., and De Feis I. Additive model selection. *Stat Methods Appl*, 25 : 519–564, 2016. 28

ANSM. État des lieux de la consommation des benzodiazépines en france. Technical report, Agence nationale de sécurité du médicament et des produits de santé, 2017. 39, 40

Arlot S. and Celisse C. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4 : 40–79, 2010. 17, 30

Assoumou L., Houssaïna A., Corstagliola D., Flandre P., Standardization, and clinical relevance of HIV drug resistance testing project from the forum for collaborative HIV research. Relative contributions of baseline patient characteristics and the choice of statistical methods to the variability of genotypic resistance scores : the example of didanosine. *Journal antimicrop chemother*, 65(4) : 752–760, 2010. 81

Austin P. C. Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination. *Stat. Med.*, 27 : 3286–3300, 2008. 14

Avalos M. Model selection via the lasso in conditional logistic regression. In *Proceedings of the Second International Biometric Society Channel Network Conference*, Ghent, Belgium, 2009. 53

Avalos M. Assessing precision of lasso conditional logistic estimates in the case-crossover design. In *Proceedings of the 3rd International Biometric Society channel network conference*, Bordeaux, France, 2011a. 53

Avalos M. Sélection de variables avec lasso dans la régression logistique conditionnelle. volume RNTI-S-1, Statistique et nouvelles technologies de l'information. Hermann, 2011b. 53

Avalos M. and Pouyes H. clogitLasso : An R package for L1 penalized estimation of conditional logistic regression models. In *Proceedings of the 1ères rencontres R (in French)*, pages 99–100, Bordeaux, France, 2012. 61, 63

Avalos M., Grandvalet Y., and Ambroise C. Parsimonious additive models. *Comput. Statist. Data Anal.*, 51 : 2851–2870, 2007. 28

Avalos M., Duran-Adroher N., Thiessard F., Grandvalet Y., Orriols L., and Lagarde E. Prescription-drug-related risk in driving comparing conventional and lasso shrinkage logistic regressions. *Epidemiology*, 23 : 706–12, 2012a. 50

Avalos M., Grandvalet Y., Duran-Adroher N., Orriols L., and Lagarde E. Analysis of multiple exposures in the case-crossover design via sparse conditional likelihood. *Stat. Med.*, 31 : 2290–302, 2012b. 53

Avalos M., Grandvalet Y., Pouyes H., Orriols L., and Lagarde E. clogitLasso : An R package for high-dimensional analysis of matched case-control and case-crossover data. In *Proceedings of the Tenth International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2013)*, Nice, France, 2013. 61, 63

Avalos M., Grandvalet Y., Pouyes H., Orriols L., and Lagarde E. High–dimensional sparse matched case–control and case–crossover data : A review of recent works, description of an R tool and an illustration of the use in epidemiological studies. 10th international meeting, cibb 2013, nice, france, june 20-22, 2013, revised selected papers. In Formenti E., Tagliaferri R., and Wit E., editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, volume 8452 of *Lecture Notes in Computer Science*, pages 109–124, 2014a. 61, 63

Avalos M., Orriols L., Pouyes H., Grandvalet Y., Thiessard F., and Lagarde E. Variable selection on large case-crossover data : Application to a registry–based study of prescription drugs and road–traffic crashes. *Pharmacoepidemiology and drug safety*, 23, 2014b. 31, 49, 55, 56, 67

Avalos M., Nock R., Ong C., Rouar J., and Sun K. Representation learning of compositional data. In *NIPS 2018*, Montréal, Canada, 2018a. 111

Avalos M., Pouyes H., Grandvalet Y., Orriols L., and Lagarde E. Sparse conditional logistic regression for analyzing large-scale matched data from epidemiological studies : a simple algorithm. *BMC Bioinformatics*, 16(6) : S1, 2015. 56

Avalos M., Pouyes H., Kwemou M., and Xu B. *clogitLasso : Sparse Conditional Logistic Regression for Matched Studies*, 2018b. R package version 1.1. 61

Bach F. Bolasso : Model consistent lasso estimation through the bootstrap. In A. McCallum S. R., editor, *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, Helsinki, Finland, 2008. 23, 25, 27, 53

Ballarín E., Ferrer P., Sabaté M., and Ibáñez L. Drug consumption databases in europe - country profile. Technical report, PROTECT project, Barcelona, 2015. 34

Barnes P., McFadden S., Machin S., and Simson E. The international consensus group for hematology review : suggested criteria for action following automated CBC and WBC differential analysis. *Lab Hematol.*, 11 : 83–90, 2005. 97, 101, 106, 107, 108

Barré-Sinoussi F., Chermann J., Rey F., Nugeyre M., Chamaret S., Gruest J., Dauguet C., Axler-Blin C., Vézinet-Brun F., Rouzioux C., Rozenbaum W., and Montagnier L. Isolation of a t-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (aids). *Science*, (220) : 868–71, 1983. 74

Bartolucci F. On the conditional logistic estimator in two-arm experimental studies with non-compliance and before-after binary outcomes. *Stat. Med.*, 29 : 1411–29, 2010. 55

Beer J. C., Aizenstein H. J., Anderson S. J., and Krafty R. T. Incorporating prior information with fused sparse group lasso : Application to prediction of clinical measures from neuroimages. Technical report, arxiv :1801.06594v3, 2018. 28

Beerenwinkel N., Montazeri H., Schuhmacher H., Knupfer P., von Wyl V., Furrer H., Battegay M., Hirschel B., Cavassini M., Vernazza P., Bernasconi E., Yerly S., Böni J., Klimkait T., Cellerai C., Günthard H. F., and Study T. S. H. C. The individualized genetic barrier predicts treatment response in a large cohort of HIV-1 infected patients. *PLoS Computational Biology*, 9(8) : 1–11, 2013. 81, 82, 83

Belloni A. and Chernozhukov V. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19 : 521–547, 2013. 52

Belloni A. and Chernozhukov V. L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1) : 82–130, 2011. 85

Bezin J., Duong M., Lassalle R., Droz C., Pariente A., Blin P., and Moore N. The national healthcare system claims databases in france, sniiram and egb : Powerful tools for pharmacoepidemiology. *Pharmacoepidemiol Drug Saf.*, 26 : 954-962, 2017. 34, 35

Bien J., Taylor J., and Tibshirani R. A lasso for hierarchical interactions. *Ann. Statist.*, 41 : 1111–1141, 2013. 51

Bihl P.-A. Optimisation des règles de déclenchement d'une revue microscopique du frottis sanguin en laboratoire de ville . *Annales de Biologie Clinique*, 76(2) : 157–64, 2018. 100

Binder H., Sauerbrei W., and Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates : a simulation study with continuous response. *Stat Med.*, 6 : 2262–2277, 2013. 14

Bleakley K. and Vert J. The group fused lasso for multiple change-point detection. Technical report, arXiv :1106.4199v1, 2011. 26

Branco P., Torgo L., and Ribeiro R. A survey of predictive modeling on imbalanced domains. 49, 2016. 102

Breiman L. Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24 : 2350–2383, 1996. 14, 17, 22

Breiman L. Statistical modeling : The two cultures (with comments and a rejoinder by the author). *Statist. Sci.*, 16(3) : 199–231, 2001. 11, 13

Brun-Vézinet F. and Commission "Résistance du VIH-1 aux antirétroviraux". Résistance du VIH-1 aux antirétroviraux. In Morlat P., editor, *Prise en charge médicale des personnes vivant avec le VIH - Recommandations du groupe d'experts sous la direction du Pr Philippe Morlat et sous l'égide du CNS et de l'ANRS*. 2016. 75

Buckley J. and James I. Linear regression with censored data. *Biometrika*, 66 : 429–436, 1979. 80, 85

Buhlmann P. Causal statistical inference in high dimensions. *Mathematical Methods of Operations Research*, 77(3) : 357–370, 2013. 19

Bunea F. and Barbu A. Dimension reduction and variable selection in case control studies via regularized likelihood optimization. *Electron. J. Statist.*, 3 : 1257–1287, 2009. 23, 24, 25, 29

Bunea F., She Y., Ombao H., Gongvatana A., Devlin K., and Cohen R. Penalized least squares regression methods and applications to neuroimaging. *Neuroimage*, 55 : 1519–1527, 2011. 18, 31, 53

Bureau A. T. S. Monograph 14 – Male pedestrian fatalities. Technical report, 2003. 40

Bursac Z., Heath Gauss C., Keith Williams D., and Hosmer D. Purposeful selection of variables in logistic regression. *Source Code Biol Med.*, pages 3–17, 2008. 11

Cai T., Huang J., and Tian L. Regularized estimation for the accelerated failure time model. *Biometrics*, 65 : 394–404, 2009. 82, 86, 93

Candes E. J. and Plan Y. Near-ideal model selection by L1 minimization. Technical report, Caltech, USA, 2007. 23

Candes E. J., Wakin M., and Boyd S. Enhancing sparsity by reweighted l1 minimization. *J Fourier Anal Appl*, 14 : 877–905, 2008. 57

Cembrowski G. S., Smith B., and Tung D. Rationale for using insensitive quality control rules for today's hematology analyzers. *International Journal of Laboratory Hematology*, 32 : 606–615, 2010. 100

Chambaz A. A. H. and van der Laan M. Special issue on data-adaptive statistical inference. *The International Journal of Biostatistics*, 12(1) : 1–1, 2016. 19

Chang C., Wu E., Chen C., Wu K., Liang H., Chau Y., Wu C., Lin K., and Tsai H. Psychotropic drugs and risk of motor vehicle accidents : a population-based case-control study. *Br J Clin Pharmacol.*, 75 : 1125–33, 2013. 34

Chatterjee A. and Lahiri S. N. Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138 : 4497–4509, 2010. 52, 53

Chatterjee A. and Lahiri S. N. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106 : 608–625, 2011. 25, 30, 53

Chawla N. V., Bowyer K. W., Hall L. O., and Kegelmeyer W. P. SMOTE : Synthetic minority over-sampling technique. *Journal of Articial Intelligent Research*, 16 : 321–357, 2002. 103

Chen T., Zeng D., and Wang Y. Multiple kernel learning with random effects for predicting longitudinal outcomes and data integration. *Biometrics*, 71(4) : 918–28, 2015. 111

Chouldechova A. and Hastie T. Generalized additive model selection. Technical report, arXiv :506.03850v2, 2015. 28, 104

Chung M., Long Q., and Johnson B. A. A tutorial on rank-based coefficient estimation for censored data in small-and large-scale problems. *Statistics and computing*, 23(5) : 601–614, 2013. 81

Chêne G. and Savès M. Master sciences, technologies, santé, mention santé publique, 2017-2018 : Introduction à l'épidémiologie. Technical report, Université de Bordeaux, ISPED, Bordeaux, France, 2018. 9

Comar S. R., Malvezzi M., and Pasquini R. Are the review criteria for automated complete blood counts of the International Society of Laboratory Hematology suitable for all hematology laboratories ? . *Revista Brasileira de Hematologia e Hemoterapia*, 36(3) : 219–225, 2014. 99, 100, 106, 107, 108

Comar S. R., Malvezzi M., and Pasquini R. Evaluation of criteria of manual blood smear review following automated complete blood counts in a large university hospital . *Revista Brasileira de Hematologia e Hemoterapia*, 39(4) : 306–317, 2017. 100, 106

Commenges D., Jacqmin-Gadda H., Proust C., and Guedj J. A newton-like algorithm for likelihood maximization the robust-variance scoring algorithm. Technical report, arxiv :math/0610402v2, 2006. 93

Commenges D. and Jacqmin-Gadda H. *Modèles biostatistiques pour l'épidémiologie.* de Boeck, Bruxelles, Belgique, 2015. 80

Corcoran C., Mehta C., Patel N., and Senchaudhuri P. Computational tools for exact conditional logistic regression. *Stat. Med.*, 20 : 2723–39, 2001. 55

Cozzi-Lepri A. Initiatives for developing and comparing genotype interpretation systems : external validation of existing rule-based interpretation systems for abacavir against virological response. *HIV medicine*, 9(1) : 27–40, 2008. 90

Cozzi-Lepri A., Prosperi M. C. F., Kjær J., Dunn D., Paredes R., Sabin C. A., Lundgren J. D., Phillips A. N., Pillay D., for the EuroSIDA, and the United Kingdom CHIC/United Kingdom HDRD Studies. Can linear regression modeling help clinicians in the interpretation of genotypic resistance data ? an application to derive a lopinavir-score. *PLoS one*, 6(11) : 1–9, 2011. 81, 94

Cui W., Wu W., Wang X., Wang G., Hao Y.-Y., Chen Y., Luo D., Shou W.-L., Zhang S., Xiang X.-F., Si Y.-Z., Chen Q., Cai H., Li T., Shen H., Shang K., and Zhang Y.-Q. Development of the personalized criteria for microscopic review following four different series of hematology analyzer in a chinese large scale hospital. *Chinese medical journal*, 123 : 3231–7, 2010. 100

Dabis F. Contrôler durablement l'épidémie VIH en France. Éditorial. *Bull Epidémiol Hebd.*, (18) : 346–7, 2017. 15, 47, 48

D'Angelo G. M., Rao D. C., and Gu C. C. Combining least absolute shrinkage and selection operator (lasso) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *BMC Proceedings*, 3 : S62, 2009. 53

Dao C. N., Patel P., Overton E. T., Rhame F., Pals S. L., Johnson C., Bush T., Brooks J. T., and Study to Understand the Natural History of HIV and AIDS in the Era of Effective Therapy (SUN) Investigators. Low vitamin D among HIV-infected adults : prevalence of and risk factors for low vitamin D levels in a cohort of HIV-infected adults and comparison to prevalence among adults in the US general population. *Clinical Infectious Diseases*, 52(3) : 396–405, 2011. 94

Dartigues J., Gagnon M., Barberger-Gateau P., Letenneur L., Commenges D., Sauvel C., Michel P., and Salamon R. The paquid epidemiological program on brain ageing. *Neuroepidemiology*, 11 : S14–S18, 1992. 71

Dassanayake T., Michie P., Carter G., and A. J. Effects of benzodiazepines, antidepressants and opioids on driving : a systematic review and meta-analysis of epidemiological and experimental evidence. *Drug Saf.*, 34 : 125–56, 2011. 39

Datta S., Le-Rademacher J., and Datta S. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and Lasso. *Biometrics*, 63 : 259–271, 2007. 81

Daubechies I., DeVore R., Fornasier M., and Gunturk C. Iteratively reweighted least squares minimization for sparse recovery. *Comm. Pure Appl. Math.*, 63 : 1–38, 2010. 57

Del Greco M. F., Pattaro C., Minelli C., and Thompson J. R. Bayesian analysis of censored response data in family-based genetic association studies. *Biometrical Journal*, 58(5) : 1039–1053, 2016. 80, 81

Delaney J. and Suissa S. The case-crossover study design in pharmacoepidemiology. *Stat Methods Med Res*, 18 : 53–65, 2009. 65

Demidenko E. Sample size determination for logistic regression revisited. *Stat Med.*, 26 : 3385–97, 2007. 11

Dezeure R., Buhlmann P., Meier L., and Meinshausen N. High-dimensional inference : Confidence intervals, p-values and R-software hdi. *Statistical Science*, 30 : 533–558, 2015. 53

Dinse G., Jusko A., Ho L., Annam K., Graubard B., Hertz-Picciotto I., Miller F., Gillespie B., and Weinberg C. Accomodating measurements below a limit of detection : A novel application of Cox regression. *American Journal of Epidemiology*, 179(8) : 1018–1024, 2014. 80, 85, 94

DiRienzo A. G. Parsimonious covariate selection with censored outcomes. *Biometrics*, 72 : 452–462, 2016. 81, 86

Dmochowski J. P., Sajda P., and Parra L. C. Maximum likelihood in cost-sensitive learning : Model specification, approximations, and upper bounds. *Journal of Machine Learning Research*, 11 : 3313–3332, 2010. 103, 104, 105

Doerken S., Mockenhaupt M., Naldi L., Schumacher M., and Sekula P. The case-crossover design via penalized regression. *BMC Medical Research Methodology*, 16(1) : 103, 2016. 65

Dolley S. Big data's role in precision public health. *Front Public Health.*, 6 : 68, 2018. 15

Drummer O. H., Gerostamoulos J., Batziris H., Chu M., Caplehorn J., Robertson M. D., and Swann P. The involvement of drugs in drivers of motor vehicles killed in australian road traffic crashes. *Accid Anal Prev*, 36(2) : 239–48, 2004. 39

Dunkler D., Plischke M., Leffondré K., and Heinze G. Augmented backward elimination : A pragmatic and purposeful way to develop statistical models. *PLoS ONE*, 9 : e113677, 2014. 12

Efron B., Hastie T., Johnstone I., and Tibshirani R. Least angle regression. *Ann. Statist.*, 32 : 407–499, 2004. 56, 60, 61, 62

Eilers P. H., Röder E., Savelkoul H. F., and van Wijk R. G. Quantile regression for the statistical analysis of immunological data with many non-detects. *BMC Immunology*, 13 : 13–37, 2012. 80, 84

Elkan C. The foundations of cost-sensitive learning. In *Proc. Int. Joint. Conf. Artificial Intelligence (IJCAI'01)*, volume 16, pages 973–978, 2001. 103

Engeland A., Skurtveit S., and Morland J. Risk of road traffic accidents associated with the prescription of drugs : a registry-based cohort study. *Ann Epidemiol*, 17(8) : 597–602, 2007. 34, 39

Engin Y. Z., Turhan K., and Örem A. Prediction of some complete blood count parameters' values using boosted regression tree. In *2012 International Symposium on Innovations in Intelligent Systems and Applications*, pages 1–4, 2012. 101

Engler D. and Li Y. Survival analysis with high-dimensional covariates : An application in microarray studies. *Statistical Applications in Genetics and Molecular Biology*, 8 : Article 14, 2009. 61

European Monitoring Centre for Drugs and Drug Addiction. Drug use, impaired driving and traffic accidents, second edition. In *EMCDDA Insights 16*. Publications Office of the European Union, Luxembourg, 2014. 38

European Monitoring Centre for Drugs and Drug Addiction. Cannabis and driving. questions and answers for policymaking. In *EMCDDA*. Publications Office of the European Union, Luxembourg, 2018. 38

EuroSafe. Injuries in the european union, summary on injury statistics 2012-2014. Report, 2016. 37

Farchione D. and Kabaila P. Variable-width confidence intervals in gaussian regression and penalized maximum likelihood estimators. Technical report, Department of Mathematics and Statistics, La Trobe University, Australia, 2010. 53

Fautrel B. Base de données du sniiram : l'ouverture de la boîte de pandore. *La Lettre du Rhumatologue*, 439 : 4–7, 2018. 35

Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*, 27 : 861–874, 2006. 106, 107

Fernandez A., Garcia S., Herrera F., and Chawla N. V. SMOTE for learning from imbalanced data : Progress and challenges, marking the 15-year anniversary. *Journal of Articial Intelligence Research*, 61, 2018. 103

Flandre P., Marcelin A., Pavie J., Shmidely N., Wirden M., Lada O., Bernard M., Molina J., and Calvez V. Comparison of tests and procedures to build clinically relevant genotypic scores : application to the Jaguar study. *Antivir Ther.*, 10(4) : 479–87, 2005. 76

Fouret N. Prise en compte de la censure à gauche dans la modélisation de données de grande dimension. application à l'analyse de l'effet des mutations vih sur la réponse virologique aux thérapies antirétrovirales. Technical report, Université Blaise Pascal, Clermont-Ferrand, France, 2013. 93

Fouret N., Avalos M., Wittkop L., Thiébaut R., and Commenges D. Prise en compte de la censure à gauche dans la modélisation de données de grande dimension. In *46èmes Journées de Statistique*, Rennes, France, 2014. 82, 93

Fournier J. P., Wilchesky M., Patenaude V., and Suissa S. Concurrent use of benzodiazepines and antidepressants and the risk of motor vehicle accident in older drivers : A nested case-control study. *Neurol Ther*, 4(1) : 39–51, 2015. 34, 39

Friedman J., Hastie T., and Tibshirani R. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 : 1–22, 2010. 57, 61, 62, 63, 89, 105

Froom P., Havis R., and M. B. The rate of manual peripheral blood smear reviews in outpatients. *Clin Chem Lab Med.*, 47 : 1401–5, 2009. 100

Fu P., Hughes J., Zeng G., Hanook S., Orem J., Mwanda O., and Remick S. A comparative investigation of methods for longitudinal data with limits of detection through a case study. *Statistical Methods in Medical Research*, 25(1) : 153–166, 2016. 80

Gange S. and Golub E. From smallpox to big data : The next 100 years of epidemiologic methods. *Am J Epidemiol.*, 183 : 423–6, 2016. 15

Geneviève F., Galoisy A., and Mercier-Bataille D. Revue microscopique du frottis sanguin : Propositions du groupe francophone d'hématologie cellulaire (gfhc). *Feuil Biol*, 317 : 7–16, 2014. 100

Genuer R., Capitaine L., and Thiébaut R. Forêts aléatoires pour données longitudinales de grande dimension, application a l'analyse de données omiques répétées. In *50èmes Journées de Statistique*, Paris, France, 2018. 111

Gibson J., Hubbard R., Smith C., Tata L., Britton J., and A.W.Fogarty. Use of self-controlled analytical techniques to assess the association between use of prescription medications and the risk of motor vehicle crashes. *Am J Epidemiol*, 169 : 761–768, 2009. 34, 64

Gillespie B. W., Chen Q., Reichert H., Franzblau A., Hedgeman E., Lepkowski J., Adriaens P., Demond A., Luksemburg W., and Garabrant D. H. Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. *Epidemiology*, 21 : 64–70, 2010. 80, 85

Gleit A. Estimation for small normal data sets with detection limits. *Environmental Science and Technology*, 19(12) : 1201–1206, 1985. 86

Goeman J. J. L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52 : 70—–84, 2010. 61, 62, 63

Gomes T., Redelmeier D. A., Juurlink D. N., Dhalla I. A., Camacho X., and Mamdani M. M. Opioid dose and risk of road trauma in canada : a population-based study. *JAMA Intern Med*, 173(3) : 196–201, 2013. 39

Goycoolea M., Quiroga T., Maulén R., Uribe R., and Parada J. Algoritmo para la liberación de revisión microscópica manual del hemograma en pacientes ambulatorios. In *XVIII Congreso Chileno de Hematología y VIII Congreso de Medicina Transfusional*, Coquimbo, Chile, 2012. 100, 106, 107, 108

Green P. J. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. Roy. Statist. Soc. Ser. B*, 46 : 149–192, 1984. 56

Greenland S. Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators. *Biostatistics*, 1 : 113–22, 2000. 55

Greenland S. Invited commentary : Variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol.*, 167 : 523–9, 2008. 11

Greenland S. Modeling and variable selection in epidemiologic analysis. *Am J Public Health.*, 79 : 340–9, 1989. 13

Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology (Cambridge, Mass.)*, 6 : 450–454, 1995. 14

Greenland S. and Pearce N. Statistical foundations for model-based adjustments. *Annu Rev Public Health.*, 18 : 89–108, 2015. 11, 17, 19, 50

Greenland S. and Pearl J. Causal diagrams. In Lovric M., editor, *International Encyclopedia of Statistical Science*, pages 208–216, Part 3, DOI : 10.1007/978–3–642–04898–2$_1$62.2011. 12

Gui J. and Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21 : 3001–3008, 2005. 61

Guichet E. *Etude des résistances du VIH-1 au traitement antirétroviral et amélioration du suivi virologique des patients vivant avec le VIH dans les pays du Sud.* PhD thesis, Université Montpellier, 2016. 75

Gulati G., Song J., Florea A., and Gong J. Purpose and criteria for blood smear scan, blood smear examination, and blood smear review. *Ann Lab Med.*, 33(1) : 1–7, 2013. 100

Guyon I. and Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning*, 3 : 1157–1182, 2003. 10

Hall P., Lee E., and Park B. Bootstrap-based penalty choice for the lasso, achieving oracle performance. *Statistica Sinica*, 19 : 449–471, 2009. 25

Hans C. Model uncertainty and variable selection in bayesian lasso regression. *Stat Comput*, 20 : 221–229, 2010. 53

Hao N. and Zhang H. H. Interaction screening for ultra-high dimensional data. *J. Am. Stat. Assoc.*, 109 : 1285–1301, 2014. 52

Haris A., Witten D., and Simon N. Convex modeling of interactions with strong heredity. *J. Comput. Graph. Stat.*, 25 : 981–1004, 2016. 52

Hastie T. J. and Tibshirani R. J. *Generalized additive models.* London : Chapman & Hall, 1990. 14, 28

Hastie T. J., Tibshirani R. J., and J. F. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, New York, 2001. 14, 16

Hastie T., Tibshirani R., and Wainwright M. *Statistical Learning with Sparsity : The Lasso and Generalizations.* Chapman & Hall/CRC, 2015. 23, 24, 27

He H. and Garcia E. A. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9) : 1263–1284, 2009. 102, 103, 104

Hebiri M. *Quelques questions de sélection de variables autour de l'estimateur LASSO.* PhD thesis, Université Paris-Diderot - Paris VII, France, 2009. 26

Heinze G. and Puhr R. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Stat. Med.*, 29 : 770–777, 2010. 55

Heinze G. and Dunkler D. Five myths about variable selection. *Transplant International*, 30, 2016. 14, 16

Heinze G., Wallisch C., and Dunkler D. Variable selection - a review and recommendations for the practicing statistician. *Biometrical Journal*, 60 : 431–449, 2018. 14, 17

Helmbold D. P. and Long P. M. On the necessity of irrelevant variables. *Journal of Machine Learning Research*, 13(Jul) : 2145–2170, 2012. 18

Helsel D. R. More than obvious : Better methods for interpreting nondetect data. *Environmental Science & Technology*, 39(20) : 419A–423A, 2005. 80, 81, 82, 84

Henriquez M., Avalos M., and Quiroga T. Reducing the number of manual film reviews in hematology laboratories : improving the consensus algorithm by machine learning. Technical report, Université de Bordeaux, France, 2018. Working paper. 107

Hewett P. and Ganser G. H. A comparison of several methods for analyzing censored data. *The Annals of Occupational Hygiene*, 51 : 611–632, 2007. 81, 84, 93

Hirsch M. S., Günthard H. F., Schapiro J. M., Vézinet F. B., Clotet B., Hammer S. M., Johnson V. A., Kuritzkes D. R., Mellors J. W., Pillay D., et al. Antiretroviral drug resistance testing in adult HIV-1 infection : 2008 recommendations of an International AIDS Society-USA panel. *Clinical Infectious Diseases*, 47(2) : 266–285, 2008. 75

Hoerl A. E. and Kennard R. W. Ridge regression : applications to nonorthogonal problems. *Technometrics*, 12(1) : 69–82, 1970. 21

Hofstra L. M., Sauvageot N., Albert J., Alexiev I., Garcia F., Struck D., Van de Vijver D. A., Åsjö B., Beshkov D., Coughlan S., et al. Transmission of HIV drug resistance and the predicted effect on current first-line regimens in europe. *Clinical infectious diseases*, 62(5) : 655–663, 2016. 75

Holder Y., Peden M., Krug E., Lund J., Gururaj G., and Kobusingye O. Injury surveillance guidelines. Report, World Health Organization, 2001. 36

Hornbrook M., Goshen R., Choman E., M O.-R., Kinar Y., Liles E., and Rust K. Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. *Dig Dis Sci.*, 62 : 2719–2727, 2017. 101

Hosmer D. W. and Lemeshow S. *Applied logistic regression (Wiley Series in probability and statistics)*. Wiley-Interscience Publication, 2 edition, 2000. 11, 12

Huang H. Controlling the false discoveries in lasso. *Biometrics*, 73 : 1102–1110, 2017. 30

Huang J. and Zhang C. Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research*, 13 : 1839–1864, 2012. 23

Huang J., Ma S., and Zhang C. The iterated lasso for high–dimensional logistic regression. Technical report, The University of Iowa, USA, 2008. No. 392. 23

Huang J., Ma S., and Xie H. Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62 : 813–820, 2006. 81

Huang X., Pan W., Park S., Han X., Miller L. W., and Hall J. Modeling the relationship between LVAD support time and gene expression changes in the human heart by penalized partial least squares. *Bioinformatics*, 20(6) : 888–894, 2004. 82

Hughes J. P. Mixed effects models with censored data with application to HIV RNA levels. *Biometrics*, 55 : 625–629, 1999. 80, 81

Hur M., Cho J., Kim H., Hong M., Moon H., Yun Y., and Kim J. Optimization of laboratory workflow in clinical hematology laboratory with reduced manual slide review : comparison between Sysmex XE-2100 and ABX Pentra DX120. *Int J Lab Hematol.*, 33(4) : 434–40, 2011. 100

IHME. Global burden of disease study 2010. In *Global Burden of Disease Study 2010 (GBD 2010) Results by Cause 1990-2010.* Institute for Health Metrics and Evaluation. Seattle, United States, 2012. 38

IMS. IMS Health. Enquête Permanente sur la Prescription Médicale (EPPM). 2005-2015. Technical report, IMS Health, Danbury CT, 2015. 45

Iobagiu C., Nehar D., Denis I., de Saint-Trivier A., and Boyer M. Vers les objectifs analytiques pertinents pour les paramètres de l'hémogramme. *Ann Biol Clin*, 72 : 705–14, 2014. 100

Ishwaran H., Kogalur U. B., Blackstone E. H., and Lauer M. S. Random survival forests. *The Annals of Applied Statistics*, 2 : 841–860, 2008. 81

Iyidogan P. and Anderson K. S. Current perspectives on HIV-1 antiretroviral drug resistance. *Viruses*, 6(10) : 4095–4139, 2014. 75

Jacob L., Obozinski G., and Vert J. P. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning, ACM*, pages 433–440, 2009. 27

Jacqmin-Gadda H., Thiébaut R., Chêne G., and Commenges D. Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics*, 1(4) : 355–368, 2000. 80, 81

Janes H., Sheppard L., and Lumley T. Overlap bias in the case-crossover design, with application to air pollution exposures. *Stat. Med.*, 24 : 285–300, 2005. 64

Janzing D. and Schölkopf B. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2017. 19

Janzing D. and Schölkopf B. Detecting non-causal artifacts in multivariate linear regression models. In Dy J. and Krause A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2245–2253, Stockholmsmässan, Stockholm Sweden, 2018. PMLR. 19

Jaro M. Probabilistic linkage of large public health data files. *Stat Med.*, 14 : 491–8, 1995. 42, 43

Jia J. and Yu B. On model selection consistency of the elastic net when p»n. *Statistica Sinica*, 20(2) : 595–611, 2010. 26

Johnson B. A. Rank-based estimation in the 1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data. *Biostatistics*, 10 : 659–666, 2009a. 81

Johnson B. A. Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70 : 351–370, 2008. 81, 82, 86, 89, 93

Johnson B. A. On Lasso for censored data. *Electronic Journal of Statistics*, 3 : 485–506, 2009b. 81, 82, 86, 88, 93, 94

Johnson B. A., Long Q., and Chung M. On path restoration for censored outcomes. *Biometrics*, 67 : 1379–1388, 2011. 86

Johnson V. A., Brun-Vézinet F., Clotet B., Gunthard H., Kuritzkes D. R., Pillay D., Schapiro J. M., and Richman D. D. Update of the drug resistance mutations in HIV-1 : December 2009. *Top HIV Med*, 17(5) : 138–145, 2009. 90, 91, 92, 95

Jörnsten R., Abenius T., Kling T., Schmidt L., Johansson E., Nordling T., Nordlander B., Sander C., Gennemark P., Funa K., Nilsson B., Lindahl L., and Nelander S. Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Molecular Systems Biology*, 7(486), 2011. 57, 62

Juditsky A. and Nemirovski A. On verifiable sufficient conditions for sparse signal recovery via L1 minimization. *Math. Programming*, 127 : 57–88, 2011. 23

Kafatos G., Andrews N., McConway K. J., and Farrington P. Regression models for censored serological data. *Journal of medical microbiology*, 62(Pt 1) : 93–100, 2013. 81, 84, 93

Kaplan R. M., Chambers D. A., and Glasgow R. E. Big data and large sample size : A cautionary note on the potential for bias. *Clinical and Translational Science*, 7 : 342–6, 2014. 15

Karjalainen K., Blencowe T., and Lillsunde P. Substance use and social, health and safety-related factors among fatally injured drivers. *Accid Anal Prev*, 45 : 731–6, 2012. 34

Keerthi S. and Shevade S. A fast tracking algorithm for generalized lars/lasso. *IEEE Transactions on Neural Networks*, 18(6) : 1826–1830, 2007. 60

Keller J. and Rice K. Selecting shrinkage parameters for effect estimation : The multi-ethnic study of atherosclerosis. *Am J Epidemiol.*, 187 : 358–365, 2018. 17

Khoury M. Planning for the future of epidemiology in the era of big data and precision medicine. *Am J Epidemiol.*, 182 : 977–9, 2015. 15

Kim Y., Kwon S., and Choi H. Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13 : 1037–1057, 2012. 30

Knight K. and Fu W. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28 : 1356–1378, 2000. 23, 52

Lagarde E. Traumatismes : les enjeux de santé publique. *Questions de santé publique - IReSP*, 23 : 1–4, 2013. 37, 38

Lang G. Éléments pour une histoire du “numéro de sécurité sociale”. *Revue "Statistique et société"*, 6(1) : 37–45, 2018. 35

Lasbeur L. and Thélot B. Mortalité par accident de la vie courante en france métropolitaine, 2000-2012. *Bull Epidemiol Hebd*, 1 : 2–12, 2017. 37

Laumon B., Gadegbeku B., Martin J. L., Biecheler M. B., and Group S. A. M. Cannabis intoxication and fatal road crashes in france : population based case-control study. *BMJ*, 331 (7529) : 1371, 2005. 38, 43, 44

Laurin C., Boomsma D., and G. L. The use of vector bootstrapping to improve variable selection precision in lasso models. *Statistical applications in genetics and molecular biology*, 15 : 305–320, 2016. 25

Lê Cao K.-A., Boitard S., and Besse P. Sparse pls discriminant analysis : Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12 : 253, 2011. 31

Lecca P., Re A., Ihekwaba A. E., Mura I., and Nguyen T.-P. *Computational Systems Biology : Inference and Modelling*. Woodhead Publishing, Limited, 1st edition, 2016. 10

LeDell E., Petersen M., and van der Laan M. Computationally efficient confidence intervals for cross-validated area under the roc curve estimates. *Electronic Journal of Statistics*, 9 : 1583–1607, 2015. 30

Lee J. D., Sun D. L., Sun Y., and Taylor J. E. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44 : 907–927, 2016. 53

Lee M., Kong L., and Weissfeld L. Multiple imputation for left-censored biomarker data based on Gibbs sampling method. *Statistics in Medicine*, 31 : 1838—-1848, 2012. 80

Lee P. H. Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification ? *J Epidemiol.*, 24 : 161–167, 2014a. 12

Lee P. H. Should we adjust for a confounder if empirical and theoretical criteria yield contradictory results ? a simulation study. *Scientific Reports*, 4, 2014b. 12

Lee S.-I., , Lee H., Abbeel P., and Ng A. Efficient L1-regularized logistic regression. In *Proceedings of the 21th National Conference on Artificial Intelligence (AAAI)*, 2006. 56

Legifrance. Arrêté du 27 mars 2007 relatif aux conditions d’élaboration des statistiques relatives aux accidents corporels de la circulation. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000274974. 37

Leng C., Lin Y., and Wahba G. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16 : 1273–1284, 2006. 18

Leon A., Perez I., Ruiz-Mateos E., Benito J. M., Leal M., Lopez-Galindez C., Rallon N., Alcami J., Lopez-Aldeguer J., Viciana P., et al. Rate and predictors of progression in elite and viremic HIV-1 controllers. *AIDS*, 30(8) : 1209–1220, 2016. 94

Li S., Hsu L., Peng J., and Wang P. Bootstrap inference for network construction with an application to a breast cancer microarray study. *The annals of applied statistics*, 7 : 391–417, 2013. 30

Li Y., Nan B., and Zhu J. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71 : 354–63, 2015. 104

Lim M. and Hastie T. Learning interactions via hierarchical group-lasso regularization. *J. Comput. Graph. Stat.*, 24 : 627–654, 2015. 52

Liu K., Markovic J., and Tibshirani R. More powerful post-selection inference, with application to the lasso. Technical report, arXiv :1801.09037v2, 2018. 53

Liu X., Wang Z., and Wu Y. Group variable selection and estimation in the tobit censored response model. *Computational Statistics & Data Analysis*, 60 : 80–89, 2013. 85, 94, 95

Lockhart R., Taylor J., Tibshirani R., and R. T. A significance test for the lasso. *Annals of statistics*, 42 : 413–468, 2014. 53

Lokhorst J. The lasso and generalised linear models. Honors project, Department of Statistics, The University of Adelaide, South Australia, Australia, 1999. 56

Loo B. P. and Tsui K. L. Pedestrian injuries in an ageing society : insights from hospital trauma registry. *J Trauma*, 66(4) : 1196–201, 2009. 46

Loth M. *Active Set Algorithms for the LASSO. (Algorithmes d'Ensemble Actif pour le LASSO).* PhD thesis, Lille University of Science and Technology, France, 2011. 23

Luxcey A. Consommation de médicaments et risque d'accident chez les piétons. Stage de Master 2 Santé Publique, spécialité Épidémiologie. Technical report, Université de Bordeaux, France, 2014. Encadrement : E. Lagarde, L. Orriols et M. Avalos, équipe "Prévention des Traumatismes", INSERM U897. 65

Lynn H. S. Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine*, 20 : 33–45, 2001. 80, 81

Ma L., Wong W. H., and Owen A. B. A sparse transmission disequilibrium test for haplotypes based on bradley-terry graphs. *Human Heredity*, 73 : 52–61, 2012. 56

Maalouf M. and Siddiqi M. Weighted logistic regression for large-scale imbalanced and rare events data. 59, 2014. 104

Maclure M. 'Why me ?' versus 'why now'–differences between operational hypotheses in case-control versus case-crossover studies. *Pharmacoepidemiology and drug safety*, 16 : 850–853, 2007. 49

Maclure M. The case-crossover design : A method for studying transient effects on the risk of acute event. *Am J Epidemiol.*, 133 : 144–153, 1991. 48

Maclure M. and Mittleman M. Should we use a case-crossover design ? *Annu. Rev. Public Health*, 21 : 193–221, 2000. 48

Maldonado G. and Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol.*, 138 : 923–936, 1993. 12

Marchetti-Bowick M., Yin J., Howrylak J., and Xing E. A time-varying group sparse additive model for genome-wide association studies of dynamic complex traits. *Bioinformatics*, 32 : 2903–2910, 2016. 28

Marks G., Gardner L. I., Craw J., Giordano T. P., Mugavero M. J., Keruly J. C., Wilson T. E., Metsch L. R., Drainoni M.-L., and Malitz F. The spectrum of engagement in HIV care : do more than 19% of HIV-infected persons in the US have undetectable viral load ? *Clinical infectious diseases*, 53(11) : 1168–1169, 2011. 94

Marschner I., Betensky R., DeGruttola V., Hammer S., and Kuritzkes D. Clinical trials using HIV-1 RNA-based primary endpoints : Statistical analysis and potential biase. *J Acquir Immune Defic Syndr Hum Retrovirol*, 20(3) : 220–227, 1999. 80, 86

Meier L., van de Geer S., and Buehlmann P. High-dimensional additive modeling. annals of statistics. *Annals of Statistics*, 37 : 3779–3821, 2009. 104

Meinshausen N. and Bühlmann P. Stability selection. *J. Roy. Statist. Soc. Ser. B*, 72 : 417–473, 2010. 25

Mickey R. and Greenland S. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol.*, 129 : 125–137, 1993. 12

Mittleman M. A. and Mostofsky E. Exchangeability in the case-crossover design. *Int J Epidemiol*, 43(5) : 1645–55, 2014. 49

Mittleman M., Maclure M., and Robins J. Control sampling strategies for case-crossover studies : An assessment of relative efficiency. *Am J Epidemiol*, 142 : 91–98, 1995. 64

Monárrez-Espino J., Laflamme L., Rausch C., Elling B., and Möller J. New opioid analgesic use and the risk of injurious single-vehicle crashes in drivers aged 50-80 years : A population-based matched case-control study. *Age Ageing*, 45 : 628–34, 2016. 34

Mooney S., Westreich D., and El-Sayed A. Commentary : Epidemiology in the era of big data. *Epidemiology*, 26 : 390–4, 2015. 15

Morvan M. L. and Vert J.-P. Whinter : A working set algorithm for high-dimensional sparse second order interaction models. Technical report, arXiv :1802.05980v1, 2018. 52

Moulis G., Lapeyre-Mestre M., Palmaro A., Pugnet G., Montastruc J., and Sailler L. French health insurance databases : What interest for medical research ? *Rev Med Interne*, 36 : 411–417, 2015. 35

Müller P. and van de Geer S. Censored linear model in high dimensions. *TEST*, pages 75–92, 2015. 81, 85, 94, 95

Nie L., Chu H., Liu C., Cole S. R., Vexler A., and Schisterman E. F. Linear regression with an independent variable subject to a detection limit. *Epidemiology*, 21 : 17–24, 2010. 80, 81

Noize P., Bazin F., Dufouil C., Lechevallier-Michel N., Ancelin M. L., Dartigues J. F., Tzourio C., Moore N., and Fourrier-Reglat A. Comparison of health insurance claims and patient interviews in assessing drug use : data from the three-city (3c) study. *Pharmacoepidemiol Drug Saf*, 18(4) : 310–9, 2009. 46

Novis D., Walsh M., Wilkinson D., St Louis M., and Ben-Ezra J. Laboratory productivity and the rate of manual peripheralblood smear review : a college of american pathologists q-probes study of 95,141 complete blood count determinations performed in 263 institutions. *Arch Pathol Lab Med.*, 130 : 596–601, 2006. 100

Née M., Avalos M., Orriols L., and Lagarde E. Étude de l'association entre consommation médicamenteuse et risque d'accident de la route. In *Groupe Biopharmacie et Santé de la SFdS, Journée annuelle*, Paris, France, 2014. 64, 71

Née M., Avalos M., Orriols L., and Lagarde E. Impact of unmeasured covariates on bias and statistical power in health administrative databases : a simulation study. In *XVth Spanish Biometric Conference and the Vth Ibéro-American Biometric Meeting*, Bilbao, Spain, 2015. 64, 71

Née M., Avalos M., Luxcey A., Contrand B., Salmi L.-R., Fourrier-Réglat A., Gadegbeku B., Lagarde E., and Orriols L. Prescription medicine use by pedestrians and the risk of injurious road traffic crashes : A case-crossover study. *PLoS Medicine*, 14 : e1002347, 2017. 56, 65, 66

Née M. Consommation médicamenteuse et risque d'accident de la route : exploration par simulation de schémas d'études épidémiologiques applicables à partir des données médico-administratives. Stage de Master 2 Santé Publique, spécialité Biostatistique. Technical report, Université de Bordeaux, France, 2014. Encadrement : M. Avalos et L. Orriols, équipe INRIA SISTM. 64, 71

ONISR. La sécurité routière en france, bilan de l'accidentalité de l'année 2016. Technical report, Observatoire National Interministériel de Sécurité Routière, Paris, 2017. Consulté le 4 juillet 2018. 37, 38, 39

Orriols L., Salmi L., Philip P., Moore N., Delorme B., Castot B., and Lagarde E. The impact of medicinal drugs on traffic safety : A systematic review of epidemiological studies. *Pharmacoepidemiol Drug Saf*, 18 : 647–658, 2009. 39

Orriols L., Delorme B., Gadegbeku B., Tricotel A., Contrand B., Laumon B., Salmi L., and Lagarde E. Prescription medicines and the risk of road traffic crashes : A French registry-based study. *PLoS Med*, 7(11) : e1000366, 2010. 34, 40, 55

Orriols L., Philip P., Moore N., Castot A., Gadegbeku B., Delorme B., Mallaret M., Lagarde E., and Group C. R. Benzodiazepine-like hypnotics and the associated risk of road traffic accidents. *Clin Pharmacol Ther*, 89(4) : 595–601, 2011. 39

Orriols L., Queinec R., Philip P., Gadegbeku B., Delorme B., Moore N., Suissa S., Lagarde E., and Group C. R. Risk of injurious road traffic crash after prescription of antidepressants. *J Clin Psychiatry*, 73(8) : 1088–94, 2012. 39

Orriols L., Wilchesky M., Lagarde E., and Suissa S. Prescription of antidepressants and the risk of road traffic crash in the elderly : a case-crossover study. *Br J Clin Pharmacol*, 76(5) : 810–5, 2013. 39

Orriols L. *Santé et insécurité routière : influence de la consommation de médicaments (étude CESIR-A)*. PhD thesis, Université Bordeaux, 2010. 42

Osborne M. R., Presnell B., and Turlach B. A. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9 : 319–337, 2000. 24, 52

Palur K. and Arakeri S. U. Effectiveness of the International Consensus Group criteria for manual peripheral smear review. *Indian Journal of Pathology and Microbiology*, 61(3) : 360–365, 2018. 100

Pariente A., Dartigues J., Benichou J., Letenneur L., Moore N., and Fourrier-Réglat A. Benzodiazepines and injurious falls in community dwelling elders. *Drugs Aging*, 25 : 61–70, 2008. 71

Park M. and Casella G. The bayesian lasso. *J. Am. Stat. Assoc.*, 103 : 681–686, 2008. 53

Park M. and Hastie T. l_1-regularization path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B*, 69 : 659–677, 2007. 53, 60, 61, 62, 63

Paxton W., Coombs R., McElrath M., Keefer M., Hughes J., Sinangil F., Chernoff D., Demeter L., B B. W., and Corey L. Longitudinal analysis of quantitative virologic measures in human immunodeficiency virus-infected subjects with $>$ or $=$ 400 CD4 lymphocytes : implications for applying measurements to individual patients. National Institute of Allergy and Infectious Diseases AIDS Vaccine Evaluation Group. *Journal of Infectious Disease*, 175(2) : 247–254, 1997. 80

Peduzzi P., Concato J., Kemper E., Holford T., and AR. F. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*, 49 : 1373–9, 1996. 11

Percival D. Theoretical properties of the overlapping groups lasso. *Electronic Journal of Statistics*, 6, 2011. 27

Peter Wu C.-S. and Zubovic Y. A large-scale monte carlo study of the Buckley-James estimator with censored data. *Journal of Statistical Computation and Simulation*, 51(2-4) : 97–119, 1995. 86

Petersen A., Witten D., and Simon N. Fused lasso additive model. *Journal of Computational and Graphical Statistics*, 25 : 1005–1025, 2016a. 26, 28, 104

Petersen M. L., LeDell E., Schwab J., Sarovar V., Gross R., Reynolds N., Haberer J. E., Goggin K., Golin C., Arnsten J., Rosen M., Remien R., Etoori D., Wilson I., Simoni J., Erlen J. A., van der Laan M., Liu H., and Bangsberg D. R. Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective HIV RNA monitoring. *Journal of Acquired Immune Deficiency Syndromes*, 69 : 109–118, 2016b. 95

Pötscher B. Confidence sets based on sparse estimators are necessarily large. *Sankhya*, 71 : 1–18, 2009. 53

Pötscher B. and Schneider U. Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electron. J. Stat.*, 4 : 334–360, 2010. 53

Powell J. L. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25 : 303–325, 1984. 80, 85

Powell J. L. Censored regression quantiles. *Journal of Econometrics*, 32 : 143–155, 1986. 80, 85

Pratumvinit B., Wongkrajang P., Reesukumal K., Klinbua C., and Niamjoy P. Validation and optimization of criteria for manual smear review following automated blood cell analysis in a large university hospital. *Archives of Pathology & Laboratory Medicine*, 137(3) : 408–414, 2013. 100

Qian J., Payabvash S., Kemmling A., Lev M., Schwamm L., and Betensky R. Variable selection and prediction using a nested, matched case-control study : Application to hospital acquired pneumonia in stroke patients. *Biometrics*, 70, 2014. 56

Quantin C., Fassa M., Coatrieux G., Riandey B., Trouessin G., and Allaert F. Linking anonymous databases for national and international multicenter epidemiological studies : a cryptographic algorithm. *Rev Epidemiol Sante Publique*, 57 : 33–9, 2009. 42

R Core Team. R : A language and environment for statistical computing. Vienna, Austria : R Foundation for Statistical Computing. 2017.

Rabinowitz M., Myers L., Banjevic M., Chan A., Sweetkind-Singer J., Haberer J., McCann K., and Wolkowicz R. Accurate prediction of HIV-1 drug response from the reverse transcriptase and protease amino acid sequences using sparse models created by convex optimization. *Bioinformatics*, 22(5) : 541–549, 2006. 81

Radchenko P. and James G. Variable selection using adaptive nonlinear interaction structures in high dimensions. *J. Am. Stat. Assoc.*, 105 : 1541–1553, 2010. 51

Rakotomamonjy A. Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms. *Signal Processing*, 91 : 1505–1526, 2011. 27

Rao C. R. and Wu Y. On model selection. In Lahiri P., editor, *Model Selection*, Lecture Notes - Monograph Series, pages 1–64, Beachwood, OH, 2001. Institute of Mathematical Statistics. 14

Reid S. and Tibshirani R. Regularization paths for conditional logistic regression : The clogitL1 package. *Journal of Statistical Software*, 58(12) : 1–23, 2014. 62, 63

Reid S., Tibshirani R., and Friedman J. A study of error variance estimation in lasso regression. *Statistica Sinica*, 30 : 35–67, 2016. 53

Rhee S., Taylor J., Wadhera G., Ben-Hur A., Brutlag D., and Shafer R. Genotypic predictos of human immunodeficiency cirus type 1 drug resistance. *Proc Natl Acad Sci U S A*, 103(46) : 17355–17360, 2006. 81

Robertson M. D. and Drummer O. H. Responsibility analysis : a methodology to study the effects of drugs in driving. *Accid Anal Prev*, 26(2) : 243–7, 1994. 43, 44

Rohart F., Villa-Vialaneix N., Paris A., Laurent B., and SanCristobal M. Phenotypic prediction based on metabolomic data : Lasso vs bolasso, primary data vs wavelet data. In *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production (WCGALP)*, Leipzig, Germany, 2010. 31

Rojas C. and Wahlberg B. On change point detection using the fused lasso method. Technical report, arXiv :1401.5408v1, 2014. 26

Rojas Castro M. Bayesian modeling strategies for risk analysis of home leisure and sport injuries (hlis). Technical report, Universitat de Valencia, Spain, Valencia, Spain, 2017. 70

Rojas Castro M., Travanca M., Avalos M., Conesa D., L O., and E L. MAVIE-Lab sports : a mHealth for injury prevention and risk management in sport. In *Proceeding of the 8th International Digital Health Conference*, Lyon, France, 2018. 70

Roohi S. Implementation of international slide review criteria for improving the efficiency of the haematology laboratory. *Apollo Medicine*, 7 : 286–288, 2010. 100

Ross M., Wei W., and Ohno-Machado L. Big data and the electronic health record. *Yearb Med Inform.*, 9 : 97–104, 2014. 15

Rosset S. and Zhu J. Piecewise linear regularized solution paths. *Ann. Statist.*, 35(3) : 1012–1030, 2007. 17, 60

Sabourin J., Valdar W., and Nobel A. A permutation approach for selecting the penalty parameter in penalized model selection. *Biometrics*, 71 : 1185–94, 2015. 30

Saeys Y., Inza I., and Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23 : 2507–17, 2007. 16

Samiullah Jatoi M., Panhwar A., Memon M. S., Baloch J. A., and Saddar S. Mining complete blood count reports for disease discovery. In *IJCSNS International Journal of Computer Science and Network Security*, volume 18, pages 121–7, 2018. 101

Sarbaz M., Pournik O., Ghalichi L., Kimiafar K., and AR. R. Designing a human t-lymphotropic virus type 1 (htlv-i) diagnostic model using the complete blood count. *Iran J Basic Med Sci.*, 16 : 247–51, 2013. 101

Sardy S. On the practice of rescaling covariates. *International Statistical Review*, 76 : 285–297, 2008. 51

Sartori S. *Penalized regression : Bootstrap confidence intervals and variable selection for high-dimensional data sets*. PhD thesis, Raleigh, NC, 2011. 53

Sauerbrei W., Royston P., and Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat. Med.*, 26 : 5512–28, 2007. 13

Schindlerova K. Prediction consistency of lasso regression does not need normal errors. *British Journal of Mathematics and Computer Science*, 19 : 1–7, 2016. 23

Schmidt M., Fung G., and Rosales R. Fast optimization methods for L1 regularization : A comparative study and two new approaches. In *European Conference on Machine Learning (ECML)*, pages 286–297, 2007. 56

Schneeweiss S., Eddings W., Glynn R., Patorno E., Rassen J., and Franklin J. Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. *Epidemiology*, 28 : 237–248, 2017. 17

Segal M. Microarray gene expression data with linked survival phenotypes : Diffuse large-B-cell lymphoma revisited. *Biostatistics*, 7 : 268–285, 2006. 61

Seiter P. Prise en compte de la censure à gauche dans le modèle pénalisé : analyse de l'effet des mutations vih sur la réponse virologique aux thérapies antirétrovirales. Technical report, Université de Strasbourg, Strasbourg, France, 2010. Encadrement : M. Avalos, équipe Biostatistique, INSERM U897. 93

SFDS, Didier E., Thalabard J.-C., Fieschi M., Bar-Hen A., Gissot C., Polton D., Gléau J.-P. L., Bossi J., Zins M., Goldberg M., Briatte F., and Goëta S. Données de santé : données sensibles. In *Revue "Statistique et société"*, volume 2. Société française de statistique, Paris, 2014. 35

SFDS, Didier E., Belorgey N., Le Gléau J., Zins M., Goldberg M., Benabdeslem K., Biernacki C., Lebbah M., Chambaz A., and Drouet I. Deux débats sur les données. In *Revue "Statistique et société"*, volume 3. Société française de statistique, Paris, 2015. 35

Shafer R. W. and Schapiro J. M. HIV-1 drug resistance mutations : an updated framework for the second decade of HAART. *AIDS reviews*, 10(2) : 67, 2008. 90, 91, 92, 95

Shah R. D. Modelling interactions in high-dimensional data with backtracking. *J. Mach. Learn. Res.*, 17 : 1–31, 2016. 52

She Y. Thresholding-based iterative selection procedures for model selection and shrinkage. *The Electronic Journal of Statistics*, 3 : 384–415, 2009. 25

Shi W., Lee K., and Wahba G. Detecting disease-causing genes by lasso-patternsearch algorithm. *BMC Proceedings*, 1 (Suppl 1) : S60, 2007. 29

Shmueli G. To explain or to predict ? *Statistical Science*, 25 : 289–310, 2010. 17, 109

Shows J. H., Lu W., and Zhang H. H. Sparse estimation and inference for censored median regression. *Journal of Statistical Planning and Inference*, 140, 2010. 81, 85, 94, 95

Sigrist F. and Stahel W. A. Using the censored gamma distribution for modeling fractional response variables with an application to loss given default. *ASTIN Bulletin : The Journal of the International Actuarial Association*, 41(02) : 673–710, 2011. 84

Silenou B., Avalos M., Pariente A., and Jacqmin-Gadda H. Adjustment for unobserved confounders in health administrative databases. In *Proceeding of the 32nd International Conference on Pharmacoepidemiology & Therapeutic Risk Management*, Dublin, Ireland, 2016. 71

Simon N. and Tibshirani R. A permutation approach to testing marginal interactions in many dimensions. Technical report, Stanford University, CA, USA, 2012. 51

Simon N., Friedman J., Hastie T., and R. T. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39 : 1–13, 2011. 61, 63

Simon N., Friedman J., Hastie T., and Tibshirani R. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2) : 231–245, 2013. 28

Smith A. D. A. C., Heron J., Mishra G., Gilthorpe M. S., Ben-Shlomo Y., and Tilling K. Model selection of the effect of binary exposures over the life course. *Epidemiology*, 26 : 719–726, 2015. 68

Smith A. D. A. C., Hardy R., Heron J., Joinson C. J., Lawlor D. A., Macdonald-Wallis C., and Tilling K. A structured approach to hypotheses involving continuous exposures over the life course. *International Journal of Epidemiology*, 45 : 1271–1279, 2016. 68

Sohn I., Kim J., Jung S.-H., and C. P. Gradient lasso for Cox proportional hazards model. *Bioinformatics*, 25 : 1775–1781, 2009. 61

Soret P., Avalos M., Wittkop L., Thiébaut R., and Commenges D. Lasso pour données censurées à gauche : une comparaison par simulation d'algorithmes proposés dans la littérature. In *47èmes Journées de Statistique*, Lille, France, 2015. 82, 94

Soret P., Avalos M., Wittkop L., Commenges D., and Thiébaut R. Lasso-regularization for left-censored outcome and high-dimensional predictors. Technical report, Université de Bordeaux, 2017a. 94

Soret P., Avalos M., Wittkop L., Commenges D., and Thiébaut R. Lasso regularization for left-censored outcome and high-dimensional predictors. *Submitted*, 2018a. 82, 93, 94

Soret P. *Régression pénalisée de type Lasso pour l'analyse de données cliniques de grande dimension : application à la charge virale du VIH censurée à gauche, aux données compositionnelles du microbiote et à l'expression génique longitudinale.* PhD thesis, Université de Bordeaux, 2018. 76, 82

Soret P. and Avalos M. Données longitudinales en grande dimension : état des lieux des packages R. In *Troisièmes rencontres R*, Montpellier, France, 2014a. 111

Soret P. and Avalos M. Méthodes d'apprentissage statistique pour des données longitudinales : une revue systématique. In *GdR Statistique et Santé*, Toulouse, France, 2014b. 111

Soret P., Avalos M., Ong C. S., and Thiebaut R. High-dimensional compositional microbiota data : state-of-the-art of methods and software implementations. In *2017 - GDR " Statistiques et santé "*, Bordeaux, France, 2017b. 111

Soret P., Avalos M. F., Delhaes L., and Thiébaut R. A simulation framework of high-dimensional phylogenetic microbiota data. 29th International Biometric Conference, 2018b. Poster. 111

Sperrin M. and Jaki T. Direct effects testing : A two-stage procedure to test for effect size and variable importance for correlated binary predictors and a binary response. *Stat. Med.*, 29 : 2544–2556, 2010. 53

Suissa S. The case-time-control design. *Epidemiology*, 6 : 248–53, 1995. 64

Sun H. and Wang S. Penalized logistic regression for high-dimensional DNA methylation data analysis with case-control studies. *Bioinformatics*, 28 : 1368–1375, 2012. 63

Sun H. and Wang S. Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data. *Stat. Med.*, 32 : 2127–39, 2013. 56, 62, 63

Sun J. X., Sinha S., Wang S., and Maiti T. Bias reduction in conditional logistic regression. *Stat. Med.*, 30 : 348–355, 2011. 55

Sutton M., Thiébaut R., and Liquet B. Sparse partial least squares with group and subgroup structure. *Stat Med*, 37(23) : 3338–3356, 2018. 69, 111

Suzumura S., Nakagawa K., Umezu Y., Tsuda K., and Takeuchi I. Selective inference for sparse high-order interaction models. In *Proceedings of the 34th Int. Conf. Mach. Learn. - ICML '17*, volume 70, pages 3338–3347, 2017. 52

Szafranski M. *Pénalités hiérarchiques pour l'ntégration de connaissances dans les modèles statistiques.* PhD thesis, Université de Technologie de Compiègne, France, 2008. 27

Sánchez-Navarro L., Castro-Castro M. J., Dot-Bach D., and Fuentes-Arderiu X. Estimation of alert and change limits of haematological quantities and its application in the plausibility control. *EJIFCC*, 25 : 115–127, 2014. 100, 106

Taylor J. and Tibshirani R. Post-selection inference for l1-penalized likelihood models. *The Canadian Journal of Statistics*, 46 : 41–61, 2018. 53

The 3C Study Group. Vascular factors and risk of dementia : design of the three-city study and baseline characteristics of the study population. *Neuroepidemiology*, 22 : 316–325, 2003. 71

Thiébaut R., Hejblum B. P., and Richert L. L'analyse des « big data » en recherche clinique. *Revue d'Epidémiologie et de Santé Publique*, 62 : 1–4, 2014. 73

Tian Z., Zhang H., and Kuang R. Sparse group selection on fused lasso components for identifying group-specific dna copy number variations. pages 665–74. 28

Tibshirani R. and Wang P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9 : 18–29, 2008. 26

Tibshirani R., Saunders M., Rosset S., Zhu J., and Kneight K. Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc. Ser. B*, 67 : 91–108, 2005. 26

Tibshirani R., Taylor J., Lockhart R., Tibshirani R., Fithian W., Lee J., Sun Y., Sun D., Choi Y., G'Sell M., Wager S., and Chouldechova A. Recent advances in post-selection statistical inference. In *Breiman lecture, NIPS 2015*, Montréal, Canada, 2015. 52

Tibshirani R. J. and Taylor J. Degrees of freedom in lasso problems. *Ann. Statist.*, 40 : 1198–1232, 2012. 30

Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 16, 22, 29, 52, 56, 83

Tibshirani R. The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16 : 385–395, 1997. 29, 52, 61, 81

Tobin J. Estimation of relationships for limited dependent variables. *Econometrica*, 26 : 24–36, 1958. 80, 83

Toh S. and Platt R. Is size the next big thing in epidemiology ? *Epidemiology*, 24 : 349–51, 2013. 15

Travanca M. Prédiction des accidents de la vie courante à partir de facteurs environnementaux et comportementaux : comparaison de méthodes d'apprentissage statistique adaptées aux données de l'observatoire mavie. Stage de Master 2 d'Ingénierie Mathématique à Toulouse (IMAT). Technical report, Université Paul Sabatier, Toulouse III, France, 2015. Encadrement : M. Avalos et L. Orriols, équipe IETO de l'INSERM U1219. 69

Trifirò G., Sultana J., and Bate A. From big data to smart data for pharmacovigilance : The role of healthcare databases and other emerging sources. *Drug Safety*, 41 : 143–149, 2017. 15

Trouessin G. and Allaert F. FOIN : a nominative information occultation function. *Stud Health Technol Inform.*, 43 : 196–200, 1997. 42

Tutz G. and Binder H. Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics*, 51 : 961–971, 2006. 104

Ueki M. A note on automatic variable selection using smooth-threshold estimating equations. *Biometrika*, pages 1005–1011, 2009. 81, 93

Uh H.-W., Hartgers F. C., Yazdanbakhsh M., and Houwing-Duistermaat J. J. Evaluation of regression methods when immunological measurements are constrained by detection limits. *BMC Immunology*, 9(1) : 59, 2008. 81, 84, 93

UNAIDS. UNAIDS data 2018. Technical report, The UNAIDS Reference Group on Estimates, Modelling and Projections, Geneva, Switzerland, 2018. 74, 75

Vaez A., van der Most P. J., Prins B. P., Snieder H., van den Heuvel E., Alizadeh B. Z., and Nolte I. M. lodgwas : a software package for genome-wide association analysis of biomarkers with a limit of detection. *Bioinformatics*, 32(10) : 1552, 2016. 94

Van de Geer S. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36 : 614–645, 2008. 23

Van der Burgh H. K., Schmidt R., Westeneng H.-J., de Reus M. A., van den Berg L. H., and van den Heuvel M. P. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. *NeuroImage : Clinical*, 13, 2017. 82

Van der Laan M., Dudoit S., and Keles S. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3 : Art. 4, 2004. 30

Van Helden P. Data-driven hypotheses. *EMBO Reports*, 14 : 104, 2013. 11, 13

Van Houwelingen H. C., Bruinsma T., Hart A. A. M., van't Veer L. J., and Wessels L. F. A. Cross-validated Cox regression on microarray gene expression data. *Stat. Med.*, 25 : 3201–16, 2006. 30

Vansteelandt S., Bekaert M., and Claeskens G. On model selection and model misspecification in causal inference. *Stat Methods Med Res*, 21 : 7–30, 2012. 12

Vittinghoff E. and McCulloch C. E. Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*, 165(6) : 710–718, 2007. 11

Vulliet-Tavernier S. Protection des données personnelles et recherche : constats et perspectives d'évolution. *Revue "Statistique et société"*, 6(1) : 31–36, 2018. 35

Wainwright M. J. Sharp thresholds for noisy and high-dimensional recovery of sparsity using L1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55 : 2183, 2009. 23

Waldron L., Pintilie M., Tsao M.-S., Shepherd F., Huttenhower C., and Jurisica I. Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics*, 27 : 3399–3406, 2011. 31

Walter S. and Tiemeier H. Variable selection : Current practice in epidemiological studies. *Eur J Epidemiol*, 24 : 733–736, 2009. 17, 50

Wang H. J., Zhu Z., and Zhou J. Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, 37(6B) : 3841–3866, 2009a. 80

Wang H. J., Zhou J., and Li Y. Variable selection for censored quantile regression. *Statistica Sinica*, 23(1) : 145–167, 2013. 81, 85, 94, 95

Wang L. and Michoel T. Controlling false discoveries in bayesian gene networks with lasso regression p-values. Technical report, arXiv :1701.07011v2, 2018. 53

Wang S., Nan B., Zhou N., and Zhu J. Hierarchically penalized Cox regression with grouped variables. *Biometrika*, 96 : 307–322, 2009b. 61

Wang S., Linkletter C., Maclure M., Dore D., Mor V., Buka S., and Wellenius G. Future cases as present controls to adjust for exposure trend bias in case-only studies. *Epidemiology*, 22 : 568–74, 2011. 64

Wang S., Nan B., Zhu J., and Beer D. G. Doubly penalized Buckley-James method for survival data with high-dimensional covariates. *Biometrics*, 64(1) : 132–140, 2008. 81, 86

Wang Y.-G., Zhao Y., and Fu L. The Buckley–James estimator and induced smoothing. *Australian & New Zealand Journal of Statistics*, 58(2) : 211–225, 2016a. 86, 89, 93

Wang Y., Chen T., and Zeng D. Support vector hazards machine : A counting process framework for learning risk scores for censored outcomes. *Journal of Machine Learning Research*, 17(167) : 1–37, 2016b. 81

Wang Z., Wu Y., and Zhao L. A LASSO-type approach to variable selection and estimation for censored regression model. *Chinese Journal of Applied Probability and Statistics*, 26(1) : 66–80, 2010. 81

Wang Z. and Wang C. Buckley-James boosting for survival analysis with high-dimensional biomarker data. *Statistical Applications in Genetics and Molecular Biology*, 9(1) : 24, 2010. 81, 82, 86, 89, 93

Wang Z., Wang M. Z., and Suggests T. Package `bujar` : Buckley-James regression for survival data with high-dimensional covariates. 2015. 89

Wei F. and Huang J. Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16 : 1369–1384, 2010. 27

Weiss G. M., McCarthy K., and Zabar B. Cost-sensitive learning vs. sampling : Which is best for handling unbalanced classes with unequal error costs ? In *DMIN*, pages 37–41, 2007. 104

Wensing A. M., Calvez V., Günthard H. F., Johnson V. A., Paredes R., Pillay D., Shafer R. W., and Richman D. D. 2017 update of the drug resistance mutations in HIV-1. *Topics in antiviral medicine*, 24(4) : 132, 2017. 75, 77, 78

WHO. Injuries and violence : the facts. Report, World Health Organization, Geneva, 2014. 36, 37

WHO. Global status report on road safety 2015. Report, World Health Organization, Geneva, 2015. 37

Wiegand R. E., Rose C. E., and Karon J. M. Comparison of models for analyzing two-group, cross-sectional data with a gaussian outcome subject to a detection limit. *Statistical Methods in Medical Research*, pages 2733–2749, 2016. 80, 81, 82, 84, 93, 94

Wipf D. and Nagarajan S. Iterative reweighted l1 and l2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing (Special Issue on Compressive Sensing)*, 4 : 317–329, 2010. 57

Wittkop L., Commenges D., Pellegrin I., Breilh D., Neau D., Lacoste D., Pellegrin J., Chêne G., Dabis F., and Thiébaut R. Alternative methods to analyse the impact of HIV mutations on virological response to antiviral therapy. *BMC Medical Research Methodology*, pages 8–68, 2008. 75, 81

Wittkop L., Günthard H., de Wolf F., Dunn D., Cozzi-Lepri A., de Luca A., Kücherer C., Obel N., von Wyl V., Masquelier B., Stephan C., Torti C., Antinori A., Garcia F., Judd A., Porter K., Thiébaut R., Castro H., van Sighem A., Colin C., Kjaer J., Lundgren J., Paredes R., Pozniak A., Clotet B., philipps A., Pillay D., Chêne G., and study group E.-C. Effects of transmitted drug resistance on virological and immunological response to initial combination antiretroviral therapy for HIV (euro-coord-chain joint project) : a european multicohort study. *The Lancet infectious diseases*, 11(5) : 363–371, 2011. 75

Wittkop L. *Analyse statistique de l'impact des mutations génotypiques du VIH-1 sur la réponse virologique au traitement antirétroviral.* PhD thesis, Université Victor Segalen Bordeaux 2, 2010. 76

Woo H.-Y., Shin S.-Y., Park H., Kim Y. J., Kim H.-J., Kyung Lee Y., Chae S., Hwan Chang Y., Rak Choi J., Han K., Ran Cho S., and Chul Kwon K. Current status and proposal of a guideline for manual slide review of automated complete blood cell count and white blood cell dfferential. *The Korean journal of laboratory medicine*, 30 : 559–66, 2010. 100

Xu B., Avalos M., and Lagarde E. Analysis of high-dimensional longitudinal data from the french health-administrative databases using machine learning methods : the example of cesir's project in injury epidemiology. In *4th International Conference on Big Data and Information Analytics (BigDIA 2018)*, Houston, USA, 2018. 69

Xue X., Xie X., and Strickler H. D. A censored quantile regression approach for the analysis of time to event data. *Statistical Methods in Medical Research*, 27(3) : 955–965, 2018. 85, 94, 95

Yang Y. Can the strengths of aic and bic be shared ? A conflict between model identification and regression estimation. *Biometrika*, 92 : 937–950, 2005. 18

Yang Y. Comparing learning methods for classification. *Statist. Sinica*, 16 : 635–657, 2006. 18, 30

Yang Y. Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, 35 : 2450–2473, 2007. 30

Ye J. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93 : 120–131, 1998. 30

Young R. Cell phone use and crash risk : evidence for positive bias. *Epidemiology*, 23(1) : 116–8, 2012. 50

Yu Z. and Deng L. Pseudosibship methods in the case-parents design. *Stat. Med.*, 30 : 3236–3251, 2011. 55

Yuan G., Chang K., Hsieh C.-J., and Lin C. A comparison of optimization methods and software for large-scale L1-regularized linear classification. *Journal of Machine Learning Research*, 11 : 3183–3234, 2010. 56, 62

Yuan L., Liu J., and Ye J. Efficient methods for overlapping group lasso. In Shawe-Taylor J., Zemel R. S., Bartlett P. L., Pereira F., and Weinberger K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 352–360. Curran Associates, Inc., 2011. 27

Yuan M. and Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68 : 49–67, 2006. 27

Yue Y. R. and Hong H. G. Bayesian tobit quantile regression model for medical expenditure panel survey data. *Statistical Modelling*, 12(4) : 323–346, 2012. 85, 94, 95

Zeng Y. and Breheny P. Overlapping group logistic regression with applications to genetic pathway selection. *Cancer Informatics*, 15 : 179–187, 2016. 27

Zhanfeng W., Yaohua W., and Lincheng Z. A lasso-type approach to variable selection and estimation for censored regression model. *Chinese Journal of Applied Probability and Statistics*, 26(1) : 66–80, 2010. 85, 94, 95

Zhang H. H. and Lin Y. Component selection and smoothing for noparametric regression in exponential families. *Statistica Sinica*, 16 : 1021–1041, 2006. 104

Zhang T. Some sharp performance bounds for least squares regression with L1 regularization. *Ann. Statist.*, 37 : 2109–2114, 2009. 23

Zhang Y., Ray S., and Gurj W. On the consistency of feature selection with lasso for non-linear targets. In *33rd International Conference on Machine Learning, ICML 2016*, volume 1, pages 322–330. International Machine Learning Society (IMLS), 2016. 23

Zhao P. and Yu B. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7 : 2541–2563, 2006. 23

Zhao Q. *Topics in Causal and High Dimensional Inference*. PhD thesis, Stanford University, 2016. 19

Zhao S. D., Lee D., and Li Y. The Dantzig selector for censored linear regression models. *Statistica Sinica*, 24(1) : 251–268, 2014. 81, 86

Zhao T. and Liu H. Sparse additive machine. In Lawrence N. D. and Girolami M., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 1435–1443, 2012. 104

Zhou J., Liu J., Narayan V. A., and Ye J. Modeling disease progression via fused sparse group lasso. page 1095–1103. 28

Zhou X. and Liu G. LAD-lasso variable selection for doubly censored median regression models. *Communications in Statistics-Theory and Methods*, 45(12) : 3658–3667, 2013. 85, 94, 95

Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101 : 1418–1429, 2006. 25, 52

Zou H., Hastie T., and Tibshirani R. On the degrees of freedom of the lasso. *Ann. Statist.*, 35 : 2173–2192, 2007. 30

Zou H. and Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67 : 301–320, 2005. 25

Résumé

Mes travaux portent principalement sur le développement, l'adaptation, l'implémentation et l'application de méthodes statistiques de sélection de modèle. Ma principale contribution consiste à adapter des méthodes de l'apprentissage statistique supervisé qui sont devenues très populaires lors de la dernière décennie, les régressions pénalisées de type Lasso, à l'analyse de données issues d'études épidémiologiques. L'enjeu est de s'attaquer aux problèmes des données volumineuses (*Big Data*) tout en respectant les objectifs et spécificités de la discipline. Le volume important se réfère ici au fait que le nombre d'observations et/ou le nombre de variables est bien plus important que celui qui était classique dans le domaine, sans exclure le cas où le nombre de variables est supérieur au nombre d'observations (données de grande dimension).

Le contexte de la pratique épidémiologique est en plein changement avec les évolutions technologiques et la conséquente disponibilité croissante des Big Data. Le Système National des Données de Santé (SNDS), regroupant les principales bases de données de santé publique existantes en France, constitue un exemple de Big Data en santé. Le données "omiques" (génomiques, transcriptomiques, protéomiques, métabolomiques, microbiomiques, mycobiomiques, viromiques,...) issues des avancées des techniques de séquençage à haut débit constituent un autre exemple de Big Data en santé. Enfin, les mesures de l'*exposome* (par opposition aux facteurs génétiques), qui désigne en épidémiologie l'ensemble des expositions environnementales que subit un individu au long de sa vie peut également constituer une source de Big Data.

Ce document s'articule autour de trois chapitres. Il résume mon activité de recherche depuis 2005, soit depuis mon recrutement à l'Université de Bordeaux après ma thèse. Le premier chapitre est une introduction générale dans laquelle je contextualise, motive et énonce la problématique abordée tout au long de mes recherches. Le deuxième chapitre est consacré à mes travaux en lien avec les études sur les traumatismes accidentels et expositions médicamenteuses à partir des données du SNDS. Le troisième chapitre est consacré à mes travaux en lien avec des études biomédicales : la prédiction de la charge virale censurée par un seuil de détection à partir des mutations du VIH, d'une part, et l'automatisation de la détection des seuils d'anomalie des hémogrammes en population générale, d'autre part.