Contributions to sparse methods for complex data analysis

Julien Chiquet

Contributions to sparse methods for complex data analysis

Julien Chiquet

UNIVERSITÉ D'ÉVRY-VAL-D'ESSONNE
École doctorale "du Génome aux Organismes"

Habilitation à diriger les recherches
Spécialité: mathématiques appliquées

présentée par
Julien Chiquet

Contributions to Sparse Methods for Complex Data Analysis

Présentée et soutenue publiquement le 8 décembre 2015 devant le jury composé de:

M.	Alexandre D'ASPREMONT	École Normale Supérieure	(Rapporteur)
M.	Arnak DALALYAN	ENSAE/ CREST	(Rapporteur)
M.	Jean-Philippe VERT	Mines ParisTech/Institut Curie	(Rapporteur)
M	Christophe AMBROISE	Université d'Évry Val d'Essonne	(Examinateur)
M^{me}	Florence D'ALCHÉ-BUC	Télécom ParisTech	(Examinatrice)
M.	Avner BAR-HEN	Université Paris 5	(Examinateur)
M^{me}	Céline LÉVY-LEDUC	AgroParisTech	(Examinatrice)

Für Él(o)ïse

ÀMargot et Camille

FOREWORD

"Oh, I'm not a percussionist, I just like to hit things."

Tom Waits

I would like to start this thesis by a brief overview of my scientific career path.

My main background is in *applied* mathematics: I graduated from the *Université de Technologie de Compiègne* in 2003. There, I obtained a degree in computer engineering with a specialty in data mining and an MSc. in computational science. My educational background hence provided me with basics in statistical learning, mathematical modeling and numerical analysis.

From 2003 to 2007 came my first years as a novice researcher. They dealt with themes that I will not cover in this document since I consider that my PhD (obtained in 2007) and the associated scientific production do a reasonable job of summarizing this activity. Let me be a little more specific though: this time period corresponds to my MSc. internship and my PhD thesis, during which I worked under the supervision of Nikolaos LIMNIOS. Nikolaos is an expert in stochastic processes – especially of the semi-Markov kind. During this period, I acquired a reasonable understanding of Markov chains and processes, which are a fundamental toolbox in applied mathematics and statistics. I made some contributions in probabilistic modeling and developed skills linked to the implementation of such methods. I also acquired skills for answering research questions with strong practical interests. Above all, I developed a taste for projects with multidisciplinary aspects. This continued during my MSc. internship, I worked at *Gaz de France Research and Innovation* (GDF) to develop models based upon heterogeneous Markov chains to predict daily temperatures throughout the year in order to optimize pipelines for natural gas transportation (see my MSc. thesis [TS2]). This was also the first time that I encountered a computer language and environment that I found rather weird at this time as a student freshly graduated from computer science school: I was asked by my GDF supervisor Karine PERVIER to translate all my Matlab code into an R-package for a more convenient use by GDF's statisticians. Taking the R path probably led my potential developer career to a dead-end... yet I guess it was for me a new start regarding my approach to modeling, where everything starts from the data themselves. I then continued on the same themes during my PhD [TS1] supported by the French Nuclear Agency (CEA). The point was to develop a stochastic approach to describe the level of degradation of a structure operating in a possibly hazardous environment across time. The random evolution of this so-called degradation process was described by a differential system with a (semi)-Markovian

environment. Such a process is a particular case of a special process known as a piecewise deterministic Markov process. With Nikolaos and Mohamed, my two supervisors, we set forth our probabilistic framework and the associated inference methods in two journal papers [JP10, JP14]; then, we developed in [JP12] a numerical method to compute the exact reliability function associated with our framework, while another paper dealt mostly with an application in structural reliability [JP13], namely the modeling of fatigue-crack propagation; two book chapters, one summarizing the whole PhD work [BC3], and another extending the model to semi-Markovian fluctuations [BC2] were published.

The second part of my career began when I started to look for an academic position in 2007: the research covered in this manuscript goes from this point to the present.

Immediately after my PhD defense during summer 2007, I got a one-year position as a Research and Teaching assistant in Bernard Prum's Lab "Statistique et Génome", at the *Université d'Évry-Val-d'Essonne*. There I found in genomics an extremely stimulating research area: the biological questioning and the nature of the data themselves raise new challenges regarding statistical modeling, not to mention the potential for applications in fields as diverse as agronomy or cancer care. Motivated by the recent craze for network modeling in biology, I started to work on Gaussian graphical models and sparse methods with Christophe AMBROISE and Catherine MATIAS, which was quite a change in terms of research theme. Fortunately, I was reasonably equipped with the appropriate background in statistical learning. More importantly, I was greatly introduced to the subject by Christophe and Catherine and their complementary points of view.

After one year, I luckily obtained a tenured position during autumn 2008 as an Assistant Professor in the same lab. I pursued these themes and co-supervised several MSc. internships and the PhD theses of Camille HUARBONNIER and Jonathan PLASSAIS with Christophe. I also had (and still have) the good fortune to work with Yves GRANDVALET, who shares his experience in statistical learning, regularization and optimization algorithms, the latter being omnipresent in modern computational statistics. I naturally came across a large variety of problems in genomics that could advantageously be tackled with such tools. I thus chose to focus on regularization, sparse methods and related statistical learning techniques.

From late 2012 to autumn 2015, I received an invited position as an INRA[1] researcher in Stéphane ROBIN's Lab, at AgroParisTech. I have further diversified my fields of application to genetics and agronomy by elaborating more involved regularized methods to a broader class of problems. I have collaborated with Stéphane and am now co-supervizing David BAKER's Post-doctorate with Tristan MARY-HUARD, about regularization methods for genomic selection. I am also working with Marie-Laure MARTIN-MAGNIETTE and Guillem RIGAILL on network inference in plants, and we are co-supervizing Trung HUs PhD about multivariate method for high-dimensional data. I have also had other very prolific collaboration with Guillem, notably at the occasion of Pierre GUTIERREZ's MSc. More generally I have forged tight connections with many members of the lab, for both friendly and professional relationships, which will undoubtedly yield interesting work and much fun!

Finally, I would like to say a word about my direct collaborations with biologists,

[1] "Institut National de La recherche Agronomique", the French Institute for Research in Agronomy

which I will not detail in this document since the associated publications do not involve any significantly new statistical methodology. Still, they are a great opportunity to stay close to the data by following the biologists in their questioning, which quickly evolves according to the technology itself. This work is thus a great source of inspiration for more methodological research and remains essential to me. In such a context, I have had fruitful collaborations with Boulos CHALOUB on polyploïd organisms like colza and wheat in the last couple of years: [JP5, JP14], I helped for the statistical analysis of transcriptomic data to answer questions specific to polyploidy and have been participating in the co-supervision of Smahane HOCLABI 's PhD thesis and Edith LE FLOCH 's post-doctoral fellowship.

Manuscript outline. This document is organized around three chapters. The first chapter depicts the motivations for my research orientations and the related methodological choices. I wish to demonstrate that these choices are pragmatic and "data oriented". The second chapter presents my contributions to GGM and sparse network inference. The third chapter describes my contributions to regularization methods, in an attempt to account for some data features in the manner by which we shape the regularization – or the penalty term – in the models.

Remark. I use a different numbering for reference to my contributions, which are quasi exhaustively listed for completeness in a separate bibliography at the beginning of this document: I hope this will ease the reading.

I also provide an academic Curriculum Vitæ in the appendix. Its main role is to cite every colleague and student I have worked with, to whom a large part of this work is due.

Julien Chiquet, November 27, 2015

Contents

Foreword v

Scientific production 1

1 Introduction and Overview 9
1.1 A typology of complex data 11
1.1.1 Genomics data, an archetype for complex data 11
1.1.2 Data characteristics 16
1.2 Recent approaches in statistical learning 18
1.2.1 New challenges in statistical learning 18
1.2.2 Marrying statistics and optimization 24
1.3 Research overview 32
1.3.1 Themes 32
1.3.2 Organization of the manuscript 34

2 Sparse Gaussian Graphical Models for Network Inference 35
2.1 Background 37
2.1.1 Basics on Gaussian graphical models 37
2.1.2 Sparse methods for GGM inference 38
2.2 Contributions 44
2.2.1 Accounting for latent organization of networks 44
2.2.2 Accounting for sample heterogeneity 49
2.2.3 Accounting for time-course data 53
2.2.4 Accounting for multiscale data: multi-attribute GGM 55
2.3 Perspectives 60

3 Structuring Penalties to Account for Complex Data Features 63
3.1 Background 65
3.1.1 Structured regularization with penalized methods 65
3.1.2 Computational consideration 72
3.1.3 Statistical analysis 75
3.2 Contributions 79
3.2.1 The cooperative-Lasso and sign coherent groups 79
3.2.2 Structured regularization for conditional GGM 88
3.2.3 A quadratic view of sparsity 96
3.2.4 Tree reconstruction with fusion penalties 103
3.3 Perspectives 112

Bibliography 115

Curriculum Vitæ 129

SCIENTIFIC PRODUCTION

PAPERS

Preprint

[PP1] V. Brault, J. Chiquet, and C. Lévy-Leduc, *A fast approach for multiple change-point detection in two-dimensional data*, biometrika, submitted.

[PP2] J. Chiquet, T. Mary-Huard, and S. Robin, *Structured regularization for conditional Gaussian graphical models*, arXiv preprint.

[PP3] Y. Grandvalet, J. Chiquet, and C. Ambroise, *Sparsity by worst-case quadratic penalties.*, arXiv preprint.

Journal papers

[JP1] C. Bouveyron, J. Chiquet, P. Latouche, and P.-A. Mattei, *Combining a relaxed EM algorithm with Occam's razor for Bayesian variable selection in high-dimensional regression*, Journal of Multivariate Analysis, 2015, to appear.

[JP2] J. Chiquet, P. Gutierrez, and G. Rigaill, *Fast tree inference with weighted fusion penalties*, Journal of Computational and Graphical Statistics, 2015, to appear.

[JP3] T. Picchetti, J. Chiquet, M. Elati, P. Neuvial, R. Nicolle, and E. Birmelé, *A model for gene deregulation detection using expression data*, BMC Systems Biology, 2015, to appear.

[JP4] B. Chaloub, F. Denoeud, S. Liu, S. Parkin, H. Tang, W. X., J. Chiquet, and 76 more, *Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome*, Science, (6199), 2014, URL http://www.sciencemag.org/content/345/6199/950 .

[JP5] H. Chelaifa, V. Chagué, S. Chalabi, I. Mestiri, D. Arnaud, D. Deffains, Y. Lu, H. Belcram, V. Huteau, J. Chiquet, O. Coriton, J. Just, J. Jahier, and B. Chalhoub, *Prevalence of gene expression additivity in genetically stable wheat allohexaploids*, New Phytologist, 197(3):pp. 730–736, 2013, URL http://onlinelibrary.wiley.com/doi/10.1111/nph.12108/full .

[JP6] J. Chiquet, Y. Grandvalet, and C. Charbonnier, *Sparsity in sign-coherent groups of variables via the cooperative-lasso*, The Annals of Applied Statistics, 6(2):pp. 795–830, 2012, URL http://projecteuclid.org/euclid.aoas/1339419617.

[JP7] J. Chiquet, Y. Grandvalet, and C. Ambroise, *Inferring multiple graphical models*, Statistics and Computing, 21(4):pp. 537–553, 2011, URL http://dx.doi.org/10.1007/s11222-010-9191-2 .

[JP8] C. Charbonnier, J. Chiquet, and C. Ambroise, *Weighted-lasso for structured network inference from time course data*, Statistical Applications in Genomics and Molecular Biology, 9, 2010, URL http://www.bepress.com/sagmb/vol9/iss1/art15 .

[JP9] C. Ambroise, J. Chiquet, and C. Matias, *Inferring sparse Gaussian graphical models with latent structure*, Electronic Journal of Statistics, 3:pp. 205–238, 2009, URL http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ejs/1238078905 .

[JP10] J. Chiquet, N. Limnios, and M. Eid, *Piecewise deterministic Markov processes applied to fatigue crack growth modelling*, Journal of Statistical Planning and Inference, 139(5):pp. 1657–1667, 2009, URL http://dx.doi.org/10.1016/j.jspi.2008.05.034 .

[JP11] J. Chiquet, A. Smith, G. Grasseau, C. Matias, and C. Ambroise, *SIMoNe: Statistical Inference for MOdular NEtworks*, Bioinformatics, 25(3):pp. 417–418, 2009, URL http://dx.doi.org/10.1093/bioinformatics/btn637 .

[JP12] J. Chiquet and N. Limnios, *A method to compute the transition function of a piecewise deterministic Markov process*, Statistics and Probability Letters, 78(12):pp. 1397–1403, 2008, URL http://dx.doi.org/10.1016/j.spl.2007.12.016 .

[JP13] J. Chiquet, N. Limnios, and M. Eid, *Modelling and estimating stochastic dynamical systems with Markovian switching*, Reliability Engineering and System Safety, 93(12):pp. 1801–1808, 2008, URL http://dx.doi.org/10.1016/j.ress.2008.03.016 .

[JP14] J. Chiquet and N. Limnios, *Estimating stochastic dynamical systems driven by a continuous-time jump Markov process*, Methodology and Computing in Applied Probability, 8:pp. 431–447, 2006, URL http://www.springerlink.com/content/e8736480p20271l3/ .

Book chapters

[BC1] M. Jeanmougin, C. Charbonnier, M. Guedj, and J. Chiquet, *Probabilistic graphical models dedicated to applications in genetics, genomics and postgenomics*, chap. Network inference in breast cancer with Gaussian graphical models and extensions, 2014, URL http://ukcatalogue.oup.com/product/9780198709022.do .

[BC2] J. Chiquet and N. Limnios, *Stochastic Reliability and Maintenance Modeling*, vol. 9 of *Springer Series in Reliability Engineering*, chap. Dynamical systems with semi-markovian perturbations and their use in structural reliability, Springer, 2013, URL http://www.springer.com/engineering/production+engineering/book/978-1-4471-4970-5 .

[BC3] J. Chiquet and N. Limnios, *Mathematical methods in survival analysis, reliability and quality of life*, chap. Reliability of stochastic dynamical systems applied to fatigue crack growth modelling, Wiley-ISTE, 2008, URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1848210108,subjectCd-ST80.html .

[BC4] A. Vacher, C. Tamaddoni-Nezhad, S. Kamenova, N. Peyrard, L. Schwaller, J. Julien Chiquet, M. Smith, J. Vallance, Y. Moalic, R. Sabbadin, V. Fievet, B. Jakuschkin, and D. Bohan, *Advances in Ecological Research*, chap. Learning Ecological Networks from Next-Generation Sequencing Data, to appear.

Popular science

[PS1] J. Chiquet, *Statistique et génome: réseaux biologiques*, La gazette des mathématiciens, 130:pp. 76–82, 2011, URL http://smf4.emath.fr/en/Publications/Gazette/2011/130/ .

Technical reports

[R1] J. Chiquet, *Pascal : Probabilistic fracture mechanics applied safety computing ageing lwr*, Tech. Rep. SEMA/LCA / RT / 05-3459, CEA, 2005.

[R2] J. Chiquet, *Équations différentielles stochastiques appliquées à la modélisation de la fatigue des matériaux*, Tech. Rep. SEMA/LCA / RT / 05-3583, CEA, 2005.

[R3] J. Chiquet, *Vers le développement de modèles aléatoires pour le vieillissement des structures : une approche stochastique*, Tech. Rep. SEMA/LCA / RT / 04-3417, CEA, 2004.

Thesis

[TS1] J. Chiquet, *Modélisation et estimation des processus de dégradation avec application en fiabilité des structures*, Ph.D. thesis, Université de Technologie de Compiègne, 2007, http://tel.archives-ouvertes.fr/tel-00165782.

[TS2] J. Chiquet, *Estimation des températures journalières à l'aide de techniques markoviennes*, Master's thesis, Université de Technologie de Compiègne, 2003, http://stat.genopole.cnrs.fr/_media/members/jchiquet/rapportdea.pdf.

CONFERENCES

Contributed talks (international)

[CI1] J. Chiquet, P. Gutierrez, and G. Rigaill, *Weighted fusion penalties for tree inference and its oracle properties*, in Proceedings of the MLCB NIPS'14 workshop, Montréal, 2014.

[CI2] D. Laloé, F. Jaffrezic, J. Chiquet, and M. Gautier, *FLPCA: a fused-Lasso PCA-based approach to identify footprints of selection in differentiated populations from dense to SNP data: applications to human and cattle data*, in Proceedings of the International Biometric Conference, Florence, Italy, 2014.

[CI3] J. Chiquet, T. Mary-Huard, and S. Robin, *Multi-trait genomic selection via multivariate regression with structured regularization*, in Proceedings of the MLCB NIPS'13 workshop, South Lake Thaoe, 2013, URL http://ai.stanford.edu/~saram/mlcb_2013/MLCB13_submission12.pdf .

[CI4] P. Gutierrez, G. Rigaill, and J. Chiquet, *A fast homotopy algorithm for a large class of weighted classification problems*, in Proceedings of the MLCB NIPS'13 workshop, South Lake Thaoe, 2013, URL http://ai.stanford.edu/~saram/mlcb_2013/MLCB13_submission4.pdf.

[CI5] J. Chiquet, Y. Grandvalet, and C. Charbonnier, *Sparsity with sign-coherent groups of variables via the cooperative-lasso*, in Proceedings of SPARS'11, Edinburgh, 2011, URL http://www.see.ed.ac.uk/drupal/sites/default/files/spars2011/spars11.pdf .

[CI6] J. Corvol, C. Vrignaud, K. Tahiri, F. Cormier, C. Charbonnier, F. Charbonnier-Beaupel, W. Carpentier, A. Patat, E. Mascioli, Y. Chiquet, J. Grandvalet, C. Ambroise, G. Edan, and E. Zanella, *Gene expression signature in whole blood after treatment with amino acid copolymer pi-2301 in multiple sclerosis*, in European Committee for Treatment and Research in Multiple Sclerosis, 2010.

[CI7] Y. Grandvalet, J. Chiquet, and C. Ambroise, *Inferring multiple regulation networks*, in Proceedings of the MLCB NIPS'10 Workshop, Vancouver, 2010.

[CI8] J. Chiquet, N. Limnios, and M. Eid, *Reliability evaluation of a dynamical system in semi-Markovian environment*, in Proceedings of IWAP'08, Compiègne, 2008.

[CI9] J. Chiquet, C. Matias, and C. Ambroise, *Penalized maximum likelihood approach for sparse Gaussian graphical models with hidden structure*, in Proceedings of IWAP'08, Compiègne, 2008.

[CI10] J. Chiquet, N. Limnios, and M. Eid, *Modelling the reliability of degradation processes through Markov renewal theory*, in Proceedings of ESREL'07, Stavanger, 2007.

[CI11] J. Chiquet, N. Limnios, and M. Eid, *Modeling and estimating stochastic dynamical systems with Markov switching*, in Proceedings of ESREL'06, Estoril, 2006.

Contributed talks (French)

[CN1] T. Mary-Huard, J. Chiquet, A. Célisse, and M. Fuchs, *Formule exacte pour la validation croisée dans le cadre de la régression "pool-sample"*, in actes des 47 journées françaises de statistique, Rennes, 2015.

[CN2] P.-A. Mattei, P. Latouche, C. Bouveyron, and J. Chiquet, *Une relaxation continue du rasoir d'Occam pour la régression en grande dimension*, in actes des 47 journées françaises de statistique, Rennes, 2015.

[CN3] J. Chiquet, T. Mary-Huard, and S. Robin, *Inférence jointe de la structure de modèles graphiques gaussiens*, in actes des 46 journées françaises de statistique, Rennes, 2014.

[CN4] J. Plassais, J. Chiquet, A. Cervino, and C. Ambroise, *A comparison of two statistical methods combining high-throughput data to predict the level of disease activity in patients with rheumatoid arthritis*, in JOBIM'12, Rennes, 2012.

[CN5] C. Charbonnier, J. Chiquet, and C. Ambroise, *Weighted-lasso for structured network inference for time-course data*, in JOBIM'10, Montpellier, 2010.

[CN6] J. Chiquet, Y. Grandvalet, and C. Ambroise, *Inferring multiple graphical structures*, in Workshop MODGRAPHII, JOBIM'10, Montpellier, 2010.

[CN7] Y. Grandvalet, J. Chiquet, and C. Ambroise, *Inférence jointe de la structure de modèles graphiques gaussiens*, in actes de CAp'10, Clermont-Ferrand, 2010.

[CN8] J. Chiquet, C. Charbonnier, and C. Ambroise, *SIMoNe : Statistical Inference for Modular Networks*, in Workshop MODGRAPH, JOBIM'09, Nantes, 2009.

[CN9] J. Chiquet, N. Limnios, and M. Eid, *Processus markoviens de saut dans les équations différentielles stochastiques appliquées à la modélisation de la fatigue des matériaux*, in Congrès Français de Mécanique'05, Troyes, 2005.

[CN10] J. Chiquet, N. Limnios, T. Yurizin, and M. Eid, *Modèle stochastique de taille critique de fissure dans les structures soumises au vieillissement par irradiation*, in Congrès Français de Mécanique'05, Troyes, 2005.

Invited talks

[IT1] *Sparse Gaussian graphical models for biological network inference*, ISI World Statistics Congress, Hong-Kong, 2013.

[IT2] *Sparse Gaussian graphical models for biological network inference*, StatLearn'13, Bordeaux, 2013.

[IT3] *Sparsity with sign-coherent groups of variables via the cooperative-lasso*, Statistics and Modeling for Complex Data, Marne-la-Vallée, 2011.

[IT4] *Learning the structure of Bayesian networks with application in post-genomics*, International Workshop on Bayesian Networks and Applications in Post-genomics, Paris, 2010.

[IT5] *Penalized maximum likelihood approach for sparse Gaussian graphical models with hidden structure*, International Workshop on Applied Probability, Compiègne, 2008.

[IT6] *Reliability evaluation of a dynamical system in semi-Markovian environment*, International Workshop on Applied Probability, Compiègne, 2008.

[IT7] *Modelling degradation processes through a piecewise deterministic Markov process*, Mathematical Methodologies for Operational Risk, Eindhoven, 2007.

[IT8] *Modelling degradation processes through a piecewise deterministic Markov process with applications to fatigue crack growth*, Recent Advances in Stochastic Operations Research II, Nagoya, 2007.

SOFTWARE AND CODES

[SW1] J. Chiquet, SPRING: Structured selection of Primordial Relationships IN the General linear model , 2014.

https://r-forge.r-project.org/projects/spring-pkg/ .

This package fits multivariate regression models using sparse conditional Gaussian graphical modeling with Laplacian regularization.

[SW2] P. Gutierrez, G. Rigaill, and J. Chiquet, Fused-Anova, 2013.

https://r-forge.r-project.org/projects/fusedanova/ .

This package adjusts a penalized ANOVA model with Fusion penalities, i.e. a sum of weighted l1-norm on the difference of each coefficient. The fitting procedure is accompanied by a highly efficient cross-validation method.

[SW3] J. Chiquet, Quadrupen: Sparsity by Worst-Case Quadratic Penalties, 2012.

http://cran.r-project.org/web/packages/quadrupen/ .

This package fits classical sparse regression models with efficient active set algorithms by solving quadratic problems. It also provides a few methods for model selection purposes (cross-validation, stability selection).

[SW4] J. Chiquet, Scoop: Sparse Cooperative Regression, 2011.

http://stat.genopole.cnrs.fr/logiciels/scoop .

This R package fits coop-Lasso, group-Lasso and tree-group Lasso variants for linear regression and logistic regression. The cooperative-Lasso (in short, coop-Lasso) may be viewed as a modification of the group-Lasso penalty that promotes sign coherence and that allows zeros within groups.

[SW5] J. Chiquet, G. Grasseau, C. Ambroise, and C. Charbonnier, SIMoNe: Statistical Inference for MOdular NEtworks , 2010.

http://stat.genopole.cnrs.fr/logiciels/simone .

SIMoNe (Statistical Inference for MOdular NEtworks) is an R package which implements the inference of co-regulated networks based on partial correlation coefficients from either steady-state or time-course transcriptomic data. This package can deal with samples collected in different experimental conditions. In this particular case, multiple related graphs are inferred simultaneously. The underlying statistical tools enter the framework of Gaussian graphical models (GGM). Basically, the algorithm searches for a latent clustering of the network to drive the selection of edges through an adaptive l1-penalization of the model likelihood.

[SW6] S. Lèbre and J. Chiquet, G1DBN , 2008.

http://cran.r-project.org/src/contrib/Archive/G1DBN/ .

A package performing Dynamic Bayesian Network inference.

[SW7] J. Chiquet, Crack growth modeling via (semi)-Markovian switching processes, 2007.

Scilab toolbox for CEA internal use.

[SW8] J. Chiquet, Estimating daily temperatures with heterogeneous Markov Chains, 2003.

R package for internal use at Gaz de France.

[SW9] J. Chiquet, Modeling the -phage through agent-based programming 2002.

INTRODUCTION AND OVERVIEW 1

If your experiment needs statistics, you ought to have done a better experiment.

Ernest Rutherford

Contents

1.1 A typology of complex data 11
1.1.1 Genomics data, an archetype for complex data 11
1.1.2 Data characteristics 16
1.2 Recent approaches in statistical learning 18
1.2.1 New challenges in statistical learning 18
Computational issues 18
Statistical issues 20
Modeling and interpretability issues 23
1.2.2 Marrying statistics and optimization 24
What do we need? 24
Sparsity, regularization and convex optimization 24
Sparse and regularization approaches to account for complex data structure 28
1.3 Research overview 32
1.3.1 Themes 32
1.3.2 Organization of the manuscript 34

THIS introductory chapter provides motivations for my research work. I first depict informally the kind of data statisticians have to deal with in recent application problems. I build on the example of genomics, with which I am familiar, in order to extract the most striking characteristics of modern data that strongly jeopardize the common way of doing statistics. I exhibit important statistical challenges associated with such data and motivate the use of particular tools at the heart of my research preoccupations, which are at the edge of statistics, optimization and machine learning. I then briefly present the main themes of my research and set them in the landscape of the statistical learning community.

1.1 A TYPOLOGY OF COMPLEX DATA

It is now a commonplace to emphasize the irrational way data is gathered about any possible aspect of our world. The collected data sets concern both our surrounding environment (such as astrophysics, plant genomics, or telecommunication network) and ourselves (such as customer data, genomics data, social network data, or any "fancy" smart phone apps that collect information about any of our movements), with a growing contribution of personal data. It has been made possible – or caused? – by the advent of digital technologies: increasing computational and storage capacities offered the possibility of measuring new phenomena and storing the associated data in a new manner. We may think of purely digital phenomena, such as flows of information over the Internet or consumer data collected straight from the cash registers. But the new computer capacities also allow room for new technologies of acquisition and measurements, such as high-throughput technology in biology. In a way, the digital revolution creates the need for these new technologies of measurement.

In these various contexts, the common motivation for collecting more data – beyond the "because we can" – is a hope for a better understanding of the underlying processes that rule the observed systems. This hope comes from the strong faith we place in modern statistics to extract relevant information from massive data: by monitoring a huge number of features in a given context, we hope for capturing the ones truly related to the process of interest. Based on the old saying "the more data, the better", we trust statistical learning methods for this task. But the systematic gathering of data at large scales in a very exploratory fashion induces data sets with complex structures: growing databases do not necessarily simplify the statistical analysis, as the collected data endow characteristics which are hardly captured by common intuition or by classical statistical methods.

In this section, I want to enumerate the most striking characteristics of modern complex data that induce new challenges in statistical learning. To this end, I rely on the canonical example of genomics. Indeed, data that have arisen in this context are quite diverse and bring together many of the characteristics shared by modern data.

1.1.1 Genomics data, an archetype for complex data

Genomics is the field of genetics that tries to characterize and analyze the structures and the functions of the genome. This recent discipline is quickly evolving thanks to the advent of biotechnologies and high-throughput techniques. Genomics research was initially motivated mainly by some fundamental questioning related to the understanding of the underlying biological processes. Nowadays it is involved in "real world" applications such as public health (with cancer prevention and classification or computer-aided diagnosis) or agronomy (with plant genomics or marker-assisted selection for breeding enhancement) and is thus partially driven by economic stakes. Hence, there are strong expectations for scientific progress based upon genomics analyses, inducing even more data with continuously renewed technologies.

Genomics was primarily concerned with the characterization of DNA sequences – especially the human DNA sequence –. However, once whole genome-sequencing had been made possible and routinely performed for various organisms, the scope of the discipline considerably broadened. *Structural genomics*, studying the three-dimensional representation of proteins encoded by the genomes, has been facilitated. More importantly, access to the full genomic material of an individual allowed the possibility of

going beyond the static characterization of the genome, notably via the emergence of *functional genomics*: the objective is to study the dynamic of the cell and to understand the complex regulations at stake in molecular biology, from gene and RNA transcription (the "transcriptome") to protein translation (the "proteome"). It is also concerned by the way this dynamic – especially gene expression – is altered under various conditions (stress, tumor cell, gene duplication, etc.). In turn, research in functional genomics revealed that gene expression may be altered by reversible phenomena like DNA methylation which do not induce modification of the DNA sequence: this is the scope of *epigenomics*. Another important emerging research area is *metagenomics*, simultaneously studying genomes of interdependent organisms living in the same environment.

In short, genomics is interested in a growing number of biological features and evolves jointly with the biotechnologies designed to unravel the processes involving these features in the cell. New discoveries raise new questions urging for omics experiments based on refined technologies and so on. In this evolving context, there is an increasing number of protocols for acquiring data at various levels of the cell, from next-generation sequencers to a large collection of array-style technologies. The couple of examples that follow aim to illustrate the range of data sets that the statistician typically has to deal with in genomics.

Example 1.1. *Differential gene expression analysis*
With transcriptomic experiments, it is possible to evaluate the activity of the genes within a cell by measuring the quantity of mRNA produced, which we call "gene expression". Next-generation sequencing techniques can be used to measure this activity by counting the sequences of a given size – or short reads – present in a sample at a given time. These reads are then matched to a catalog of known mRNA sequences (called the transcripts) to assess their expression levels. As an example, Figure 1.1 provides the preprocessed output of an mRNA-Seq experiment on corn [134] observed in two tissues from the same biological sample, either from the plant leaf or its root: we plot the counts (resp. the negative counts) associated with the 199,047 transcripts in blue for the root (resp. in red for the leaf).

Based on several replicates of such experiments (generally just a couple!), the question addressed by differential analysis is to determine a set of transcripts the activity of which is different in two tissues, or discriminates one tissue from another.

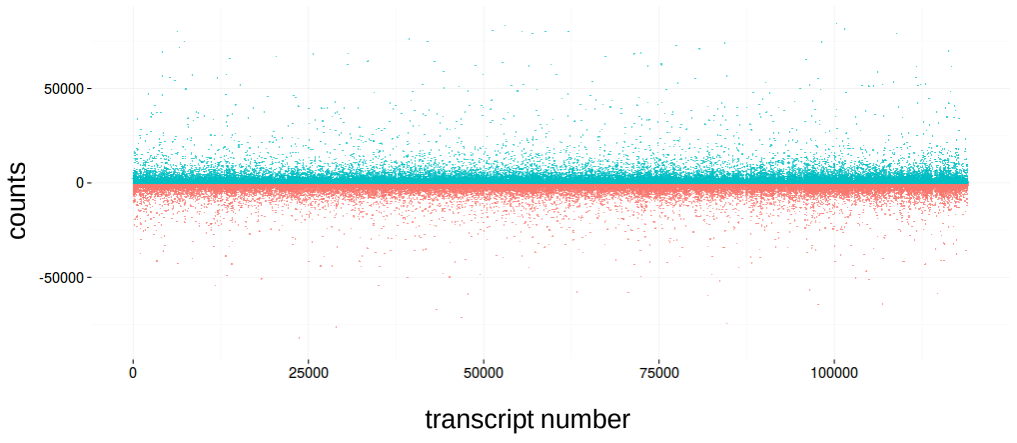


Figure 1.1 –*Gene expression data (Illumina mRNA Sequencing)*

Example 1.2. *Detecting genetic aberrations*

When a DNA sequence is suspected of being severely altered (as in tumor cells for instance), a major question is to known whether any coding regions have been affected, in which case important functions may be lost in the organism. A quick strategy is to detect gains or losses of regions along the chromosomes by comparing the ploidy level along the sequence between the suspected DNA sample and a reference sample.

To this end, array comparative genomic hybridization (aCGH) measures the copy number variations (CNV) between two genomes at a low resolution. In Figure 1.2, we represent the copy number logarithmic ratio between five DNA samples from breast cancer cell lines and a control sample from the NCI-60 data set [144]. The signal associated with each cell line is composed of approximately 44,000 points corresponding to ordered features dispatched along the genome. We use a different color for each sample.

A statistical question naturally arising is to automatically segment those signals, in order to help find the regions of the genome that are altered. This task can be performed on each single cell line independently or jointly, if those lines share some similarities (here, the kind of cancer of origin).

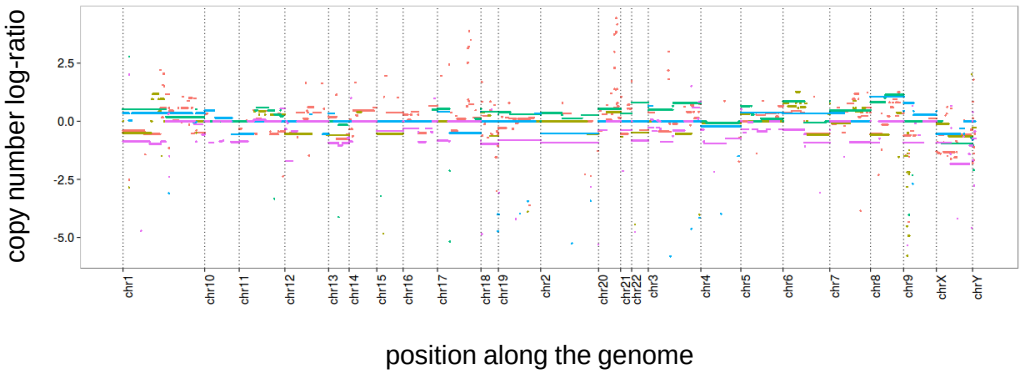


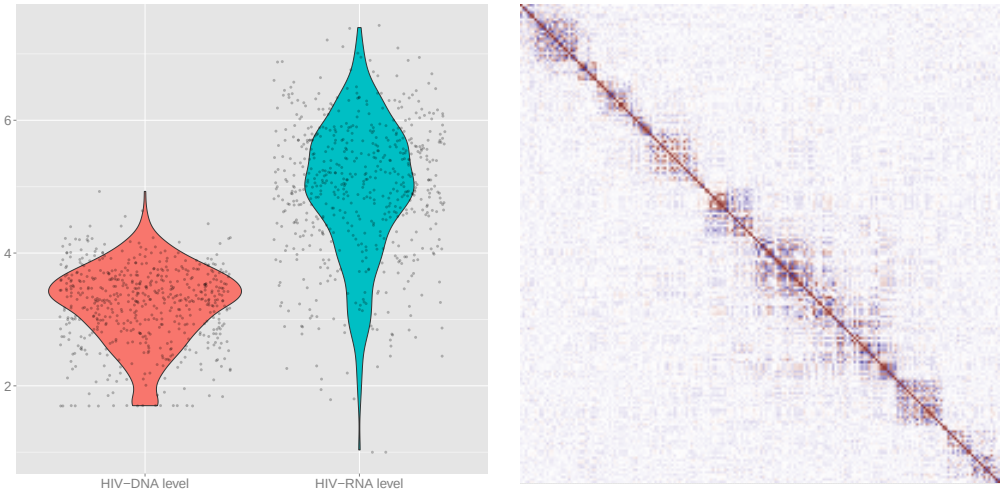
Figure 1.2 –*Chromosomal copy number changes (aCGH Agilent 44K Human array)*

Example 1.3. *Genome-wide association study and marker-assisted selection*

Small genetic variations between individuals of the same species are common and often without any effect on the phenotype, either because those variations occur on non coding parts of the genome, or because they do not affect the translation process. However, some genetic variants may be associated with some phenotypic alterations in various ways: in plant genomics for instance, these variants are exploited to select lines showing the best yields, a field known as marker-assisted selection, or genomic selection. In medical research, association studies are performed to detect variants that induce differences in a particular disease development, or alter the efficiency of some treatments.

These small variations can be assayed at the nucleotide level with SNP (single-nucleotide polymorphism) arrays, that monitor millions of genetic variants at once on predefined loci of the genome. For instance, in the GWAS (Genome-wide association study) presented in [136], SNP-genotyping has been performed on 605 HIV-infected patients in order to evaluate the influence of genetic variants on the disease progression. The latter is measured either by the HIV-DNA level or the HIV-RNA level, the distributions of which are represented on the left panel of Figure 1.3a. The objective here is to find a set of features among the 317,000 SNP which jointly explain the response variables measuring the disease progression. The right panel of Figure 1.3a represents a

small block of the empirical correlation matrix of the SNP values in the cohort, which shows interesting patterns that should be taken into account in the statistical modeling.



(a) *Indicators of disease progression*

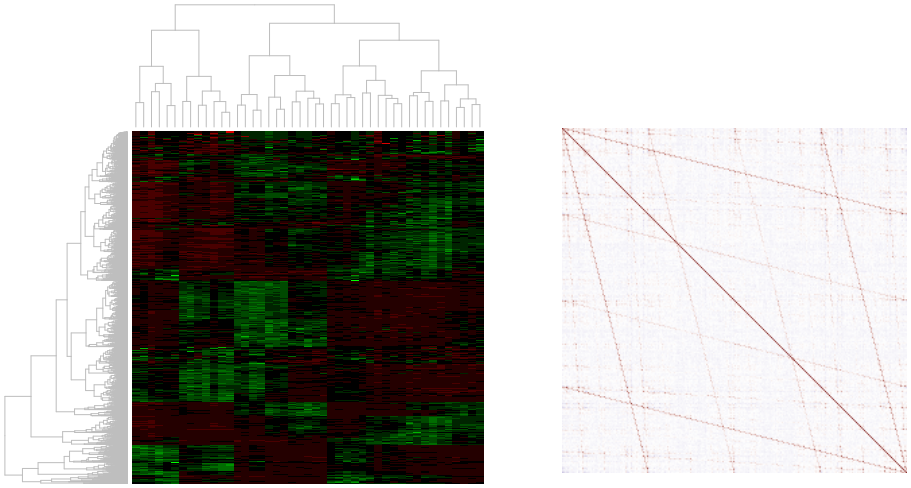
(b) *Correlation pattern restricted to the first 200 SNPs*

Figure 1.3 – *SNP-genotyping (Illumina HapMap300 array) in GWAS*

Example 1.4. *Regulatory motif discovery*

Within the cell, the gene expression is initiated by transcription factors that bind to the DNA upstream from the coding regions, called *regulatory regions*. This binding occurs when a given factor recognizes a certain (small) sequence called a *regulatory motif*. As the binding relies on chemical affinity, some degeneracy can be tolerated in the motif definition, and motifs similar but for small variations may share the same functional properties. Hence, genes hosting similar regulatory motifs will be jointly expressed under certain conditions.

In order to detect such regulatory motifs, we aim to relate the expression level of all genes across a series of conditions with the content of their respective regulatory regions in terms of motifs. Figure 1.4 provides insights into the data available to perform such a task for *Plasmodium falciparum*, a parasite infamous for causing malaria: on the left panel of Figure 1.4a, we represent gene expression profiling gathered in [19] for the approximately four thousand genes of *P. falciparum* measured across 46 conditions. A simple hierarchical clustering shows strong patterns both along the genes and the conditions. Concerning the motifs, the data are obtained by counting the occurrences of a set of candidate motifs in the regulatory regions of each gene (publicy available at http://plasmodb.org/). An example of motif counts data is illustrated on the right panel of Figure 1.4b: we plot the empirical correlation matrix between the motif counts when the set of candidates consists of 4-size motifs composed with the letters $\{A, C, G, T\}$ and classified by lexicographical order. Strong patterns appear, supporting the assumption that similar motifs have similar effects. The question is then to select motifs showing strong relationships with the expression data. However, for real application purposes, one must consider motifs with a considerably larger size (at least 11), meaning to deal with a huge number of candidates (4^{11}).



(a) *Clustering of gene expression (row: genes; column: conditions)*

(b) *Correlation pattern between motifs counts*

Figure 1.4 – *Linking regulatory region sequences to expression data (Affymetrix GeneChip array)*

Example 1.5. *Gene Regulatory Network Inference*

A synthetic view of the regulations at play between a set of genes within the cell is conveniently represented through a graph. The nodes stand for fixed genes while the edges represent interactions due to the genes and their products. The most striking interactions occur between genes encoding for particular proteins, called transcription factors, that are specifically designed in the cell to regulate other genes by binding onto their promoters.

Reconstruction of such networks is extremely informative on the gene expression machinery and has many applications: if one has at one's disposal a general scheme on how a cell operates in a given condition, we may target a given gene in the network to induce a given behavior at the cell level. In medical research, this could be a better response to a drug treatment. In plant breeding, people may target a gene resulting in a better yield or a better resistance of the plant. Hence, automatic reconstruction – or inference – of gene regulatory networks (GRN) from genomics data has been an important research theme in computational biology. To achieve this task, one would naturally rely on gene expression assays like the ones in Figures 1.1 and 1.4a. However, transcriptomic data is unable to capture the numerous regulations that may operate at various other stages of cell development: complex regulations may occur due to proteins binding together; epigenomic phenomena like DNA methylation are known to alter gene expression; and genetic alterations in certain tissues certainly have profound impact on gene expression and regulation. GRN inference is thus a challenging problem, and state-of-the art methods in statistics try to strengthen the inference process by integrating various types of data together and/or by introducing external biological information.

§

These few examples do not claim to provide a comprehensive view of genomics data. Yet, they hopefully illustrate the complexity of their typology, due to different observation scales, various technologies, continuous/discrete signals, plurality and complexity of the biological processes in place, etc.

1.1.2 Data characteristics

From the statistical point of view, the challenges arising with the analysis of genomics data are mostly the consequences of the following data characteristics.

Large databases. The most obvious feature is the size of the data: as sequencing technologies are widely evolving, the base units in 'omic' studies are getting smaller, meaning larger data-sets to cope with. In the most dreadful cases – typically metagenomics nowadays –, the preprocessed data-sets concern catalogs of transcripts of hundreds of thousands of species weighing several Terabytes. We thus have to deal with samples where the number of features ranges from a few hundred to a few billion.

More variables than individuals. A more challenging and fundamental trait of genomics data is that the sample size remains of the same order as it used to be, while the number of features per sample keeps on increasing with technological improvements. Drawing a sample (like performing a biopsy, growing a plant or breeding an animal) cannot be performed in the same systematic way as many features are measured at once with high-throughput technologies. Thus, statistics are doomed to adapt to the new paradigm of "high-dimensional data", where the number of variables may exceed the sample size by several orders of magnitude: in the couple of examples depicted above, we typically observe thousands to millions of features to be compared with only dozens to hundreds (sometimes thousands) of individuals.

Multiple sources of heterogeneity. The genomics databases are made up of data sets which are largely heterogeneous. First, we observe diversity in the types of data: in the examples above we encountered continuous variables, counts or categorical data (e.g. from SNP array); moreover some signals are originally available as images or character strings; we may also think of external biological information encoded as graphs or tree structures. Second, we observe different kinds of dependencies within the data sets depending on the relationships between the features at stake: CNV or SNP data in Examples 1.2 and 1.3 are intrinsically longitudinal because of some spatial relationships. Time dependency can also be at stake if a biological process like gene expression levels is measured across the cell cycle. Third, data may live in quite different spaces and at quite different scales due to the fundamental nature of the underlying biological processes, which operate at multiple places and times of the cell, and involving various actors. Many experimental protocols and technologies have been adapted to measure the activity of these biological actors. A consequence is that we have to cope with multiscale data. Last but not least, a larger level of noise can be observed within a given technology, especially the oldest array technique, and incoherence across platforms is likely to occur: transcriptomic experiments can be performed with sequencing technologies as in Example 1.1 or with hybridization arrays as in 1.4. While measuring the same phenomenon, multiple data sets using these two distinct technologies will not share exactly the same features, nor the same precision, nor the same resolution.

Highly structured data. The characteristics mentioned up to this point (large data size, high dimensional feature space and heterogeneity) all sound like drawbacks for statistical analysis, which may seem almost hopeless at this stage. Fortunately, genomics data – and most data arising in life science – are deeply structured: hopefully, taking

this structure into account in the statistical modeling may be sufficient to overcome the other difficulties.

This high level of structure has various sources, some being due to the underlying biological mechanisms and the relationships between the biological actors, and some being due to the sampling scheme and the way data-sets are collected. Most of the time however, the structure is only very partially known and must be guessed from the data themselves. The series of examples above illustrates this fact:

> In Example 1.1 (differential analysis of mRNAseq colza samples), an obvious structure is due to the tissue where expression is measured (either root or leaf of the plant). With a deeper biological knowledge of the problem, however, one would know that colza is a polyploid organism. This means that some genes called homoeoalleles, sharing very similar sequences, will mostly exhibit highly correlated expression levels. This grouping defines another level of structure in the data.
>
> In Example 1.2 (chromosomal copy number changes in breast cancer), the predictors have a natural spatial structure, that is to say, their ordering along the genome. This structure is intrinsic to the segmentation problem. Another less obvious form of structure arises between the samples: some changes in the ploidy level occur simultaneously in several cell lines (e.g. on chromosome 6), in which case the segmentation would be enhanced if performed jointly.
>
> Example 1.3 (genomic selection for colza) illustrates the existence of a complex pattern of correlation between the genetic markers. This phenomenon is known as linkage disequilibrium in population genetics, which basically states that the allelic status are not independent between two loci. The most obvious reason is due to the spatial organization of the genome: close loci with given allele variants are likely to be jointly inherited. Still, other factors (population structure, mutation rate or preferential mating) influence the level of linkage disequilibrium. This explains that the correlation matrix is not defined purely block-wise but through a complex hierarchy.
>
> In Example 1.4 (regulatory motif discovery), a simple heatmap on the gene expression profiles shows a block structure both at the gene and condition levels. Structure on the conditions is likely to be connected with the nature of the considered conditions (heat stress, light stress, cell cycle, etc.). The origin of the gene structure is less clear since it is related to complex direct and indirect regulatory relationships between the genes. Finally, a strong correlation pattern arises between the predictors, measured by the occurrence of the motifs in the promoters of all genes. A part of this correlation can be explained by the similarity between the motifs, when they are equal up to a couple of letters and sorted in the lexicographical order. The correlation that remains may be due to more complex biological features, e.g. a couple of motifs related to a set of genes associated with the same biological pathway.

§

Though motivated by genomics, these characteristics are shared by many complex data sets encountered in application fields beyond biological sciences, like astronomy, imaging, signal processing or finance to cite but a few. The next section shows how these characteristics change our way of doing statistics.

1.2 RECENT APPROACHES IN STATISTICAL LEARNING

The previous section illustrates how data gathering is deeply evolving and is inducing new data characteristics to deal with. An important – though straightforward – remark is to note that most of the traditional goals of statistical learning remain basically unchanged, either in supervised learning (the goal is prediction, via classification or regression) or unsupervised learning (the goal is to unravel interesting patterns via clustering or feature extraction). Indeed, the questions we aim to answer by analyzing modern data sets as in the examples above can be cast as a classical task of statistical learning such as regression, classification, clustering, dimensionality reduction and so on. However, we cannot directly rely on the most favorite and standard methods available since they are not designed to fit data sets with the aforementioned characteristics.

This section starts by showing why traditional approaches are challenged and what their most dramatic limitations towards modern data analysis are. Then, I present the ingredients composing the methods that I develop and work with in my research, designed to answer these challenges. This path follows the recent popular trend in statistical learning which tends to marry tools from statistics and optimization.

1.2.1 New challenges in statistical learning

Various angles are possible to categorize the challenges that jeopardize the traditional way of doing statistics. Hereafter, I successively discuss the computational, the statistical and the interpretability issues. This ordering does not reflect the importance of each point; it rather mirrors how they come to the applied statistician's mind: when dealing with modern data, computational challenges come first, as in the most dramatic cases traditional methods do not even numerically apply. Even when the fit is possible, statistical issues may occur as the assumptions coming with the traditional theoretical guarantees are not fulfilled in this setting. Finally, the fits should be interpreted with caution as standard approaches are not tailored to cope with data heterogeneity nor designed to embed the structure of the problems appropriately.

Meanwhile, an important threat of modern data which connects computational, statistical and interpretability issues together is the problem of high-dimensional feature spaces. In computational biology, the couple of examples given in Section 1.1.1 illustrate that the new standard is to deal with many features p for a moderate sample size n, such that $n < p$ – or even $n \ll p$ –. We commonly speak of a *high-dimensional problem* when analyzing data entering this setting. In many other fields (e.g. signal processing, medical imaging, internet, finance) the new standard is $n \gg p$ with both n and p large, which corresponds to a class of so-called "big-data" problems. In both situations, one must deal with a large number of features at once, which has many impacts that I shall use as a common thread throughout the statements that follow.

Computational issues.

With the increasing number of features and more generally the advent of large data bases, the computational aspect is now a central question in statistics. First, the algorithms used to fit a model must show a rather low complexity regarding the number of features p. This means that many classical methods showing high statistical performance are completely out of reach due to an overly large intrinsic complexity. And second, new statistical methods should be designed to use efficiently the available computer resources (e.g., by allowing parallel computing). However, this latter point

should not overcome the former, in the sense that algorithms with low complexity in p are mandatory in order to deal with high-dimensional feature spaces.

To get better insight on the computational problems at stake, I found it interesting to adopt an optimization point of view. Indeed, most of the statistical methods can be cast as one or a series of optimization problems. Thus, studying the complexity of various classes of problems from the optimization viewpoint undoubtedly provides insights on the limitations of the statistical methods that build on them. To this intent, the classification given by Nesterov in [128] is particularly illustrative. It is reproduced with slight modifications in Table 1.1 that shows the typical operations that we can afford for a given problem size, accompanied with the memory requirement and the range of computational cost (the latter depending on a particular structure of the problem, e.g., sparsity). I added instances from omics that match the various problem scales and the learning tasks typically expected.

class of problem	dimension (# features p)	conceivable operations	computational cost	memory requirement	example in omics	expected task
small	10^0 10^2	All	p^3 p^4	10^3 (Kb)	–	–
medium	10^3 10^4	A^{-1}	p^2 p^3	10^6 (Mb)	transcriptomics	network inference
large	10^5 10^7	Ax	p p^2	10^9 (Gb)	association studies	variable selection
huge	10^8 10^{12}	$x + y$	$\log(p)$ p	10^{12} (Tb)	metagenomics	clustering

Table 1.1 –*Typical matrix algebra operations with their computational cost and memory requirement for various problem scales. A is a $p \times p$ matrix and x, y are vectors in $\mathbb{R}^p$ (source: [128]). Corresponding data regime for some problems in genomics with the desired learning tasks.*

This table suggests several comments. First, it gives clues as to the methods that can be applied depending on the situation. Consider for instance the extreme case of metagenomics where one aims to cluster billions of sequences: general agglomerative clustering algorithms in $O(n^3)$ are completely out of reach in this case. It means that some popular procedures such as UPGMA (Unweighted Pair Group Method with Arithmetic Mean) for average linkage clustering should be banned for some problem scales as it exhibits a quadratic complexity. Second, this table shows that, when possible, we must adapt the optimization procedures used to fit a given statistical method to the problem size. In the case of UPGMA, the method is defined in itself by an algorithm and there is no way to change the underlying complexity. In contrast, when a statistical model is adjusted by minimizing for instance a negative log-likelihood, many optimization procedures are available for this purpose. For instance, we may use a second-order method like Newton-Raphson which relies on the first and second derivatives of the log-likelihood. This method converges quadratically to the solution but requires the inversion of a $p \times p$ matrix at each iteration. Another possibility is to use first order methods, such as the steepest gradient descent method, which only relies on the first derivative of the log-likelihood. Such gradient methods typically have a linear convergence rate, requiring more iterations than Newton's method to meet the same precision, but only require operations like Ax in Table 1.1. Thus, statistical methods originally designed for a medium scale situation can still be applied to larger scale situations[1] if adapting the underlying fitting algorithms is possible: trading some speed of convergence, meaning more iterations, is the price to pay to scale the dimension by relying on simpler operations at each iteration of the optimization procedure.

[1] Related to this question, Nesterov's paper [128] reviews subgradient methods for huge-scale problems in Table 1.1, requiring many iterations to converge but extremely simple operations.

As discussed in the next section, other sources of motivation for using simple and highly-efficient procedures come from the statistical side. Indeed, a major concern in high-dimensional spaces is overfitting, in which case resampling and ensemble methods may be a solution, despite their additional computational cost. This again advocates for highly efficient algorithms designed to use all the computational resources available.

To conclude with this part, the computational aspect turns out to be so important in statistical learning that several authors [15, 90] advocate for criteria that take into account both statistical and numerical performance to compare statistical methods.

Statistical issues.

In general, considering large feature spaces is cumbersome since most of our intuition breaks down, especially our geometrical intuition. This is basically due to the fact that the volume of high-dimensional spaces increases exponentially fast compared to the amount of data points available, which are extremely sparse in those spaces. Thus, the sample size n of the data does not have to be smaller than the dimension p for problems to occur. This phenomenon and its various implications are often referred to as the *curse of dimensionality*. Several illustrations can be found for instance in Chapters 2 and 18 of the classical book [70] of Hastie, Tibshirani and Friedman. In a more recent effort [63], Giraud also gives many instances of this phenomenon that provide interesting insights from various points of view (geometrical, probabilistic, statistical and computational). Although I do not aim to investigate exhaustively the numerous manifestations of the curse of dimensionality, I underline here two major related problems, namely *data scarcity* and *overfitting*.

Data scarcity. In high-dimensional spaces, data points – even when n – are very isolated and are all at a similar distance from one another: points are so sparsely disseminated that the notion of neighborhood is hardly relevant. To illustrate this point, we revisit a small numerical experiment inspired by Figure 1.3 of Giraud's book [63], which is designed to show that local methods such as local regression or k-nearest-neighbor are doomed to fail in high-dimensional spaces. We consider a random vector $X \in \mathbb{R}^p$ with a normal distribution $\mathcal{N}(0_p, I_p)$ and draw a size-n random sample $(X^1, \ldots, X^n)$ with $n = 500$ and for various values of $p \in \{2, 10, 100, 1000\}$. Elementary computations show that, for all $1 \leq i < i' \leq n$,

$$\mathbb{E}\left[\|X^i - X^{i'}\|_2^2\right] = 2p, \qquad \mathrm{Var}\left[\|X^i - X^{i'}\|_2^2\right] = 8p.$$

As for Giraud in his example with uniform random variables, we also meet in the Gaussian case a pairwise square distance the mean of which grows linearly with p, while its standard deviation only grows in $\sqrt{p}$. Thus, in high-dimensional spaces when p is large, all pairs of points are at a similar distance and thus indistinguishable. Local methods, based upon the notion of neighborhood which is not relevant when p is large, will thus perform poorly. This phenomenon is illustrated on Figure 1.5, where we represent in red the (scaled) histograms of the scaled pairwise distances $\|X^i - X^{i'}\|_2 / \sqrt{2p}$ for all $1 \leq i < i' \leq n$ and $p \in \{2, 10, 100, 1000\}$.

At first glance, the most straightforward conclusion drawn from this simple experiment – and from other manifestations of the curse of dimensionality – is that separating the noise from the signal looks extremely challenging, if not impossible, in high-dimensional spaces. Hopefully, the hypothetical situation where the p features are

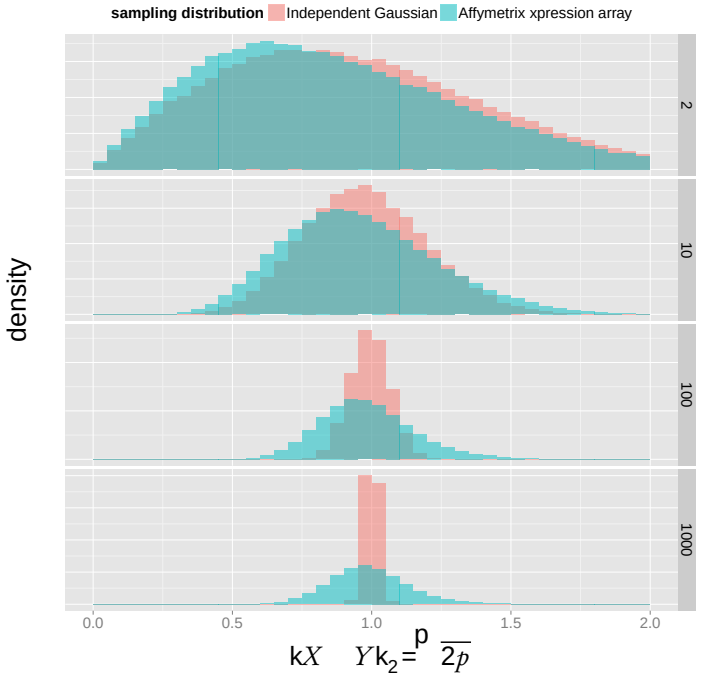


Figure 1.5 –*Empirical distributions of scaled pairwise distances between vectors in* $\mathbb{R}^p$ *sampled from* $\mathcal{N}(0_p, I_p)$ *or from the breast cancer expression data set* [67]. *Values of* p *vary in* $\{2, 10, 100, 1000\}$. *To illustrate the concentration of the distance in high-dimensional spaces for the independent Gaussian case. Pairwise distances sampled from expression data are more spread around their mean, meaning more structured data.*

independent does not fit the reality and the true underlying space where the data lie is most probably low-dimensional. To support this point, we also report in Figure 1.5 the scaled histograms (in blue) of the scaled distance when X is sampled from the breast cancer gene expression data studied in [67], where $p = 44,000$ transcripts are monitored for $n = 500$ patients. For random subsets of genes with size $p \in \{2, 10, 100, 1000\}$, the histograms look more spread out than the theoretical one, meaning that we might be able to separate those points according to their pairwise distances. This gives some hope if one has a clue about the shape of the underlying space or of some structure in the data, in other words *if we account for the structure of the problem*.

Overfitting. Overfitting affects models which are too complex with low bias but large variance. The consequence is a poor capability for generalization, meaning a large test error. This problem is greatly exacerbated in high-dimensional spaces where distinguishing noise from signal is especially challenging. Moreover, data is so scarce that adjusting a fit with the model that truly generated the observations may lead to poorer results than applying simpler models, with high bias but low variance. Let us consider a simple idealistic example in linear regression to illustrate this point: we draw $\{(x_i, y_i)\}_{i=1,\dots,n}$ with x_i sampled in the interval $[\ ,\ ]$ and we choose the "true" relationship such that $y_i = \sin(2x_i) + \mathcal{N}(0, \ ^2)$. We choose $\ $ to meet a coefficient of determination $R^2 = 0.8$. Now, suppose that we do not know the nature of the true relationship. We fit the data using polynomial regression with order p, that is

$$y_i = \ _0 + \sum_{j=1}^{p} x_i^j \ _j + \ "_i, \qquad "_i \ \mathcal{N}(0, \ ^2). \tag{1.1}$$

This toy example allows us to show on a two-dimensional fit the effect of an excessively complex model with too many features, that is, living in a high-dimensional space. The order p of the polynomial is used to control the dimension p of the feature space, meaning the model complexity. We study cases where $p \in \{1, \dots, 50\}$, *i.e.* models as simple as a regression line and as complex as a polynomial with degree 50. We consider three regimes for the training, with sample size $n \in \{10, 50, 200\}$. Whenever possible, Model (1.1) is fitted with ordinary least squares (OLS). In cases when $n < p$, (which occurs only when $p > 10$ and $n = 10$), we use ridge regression with a tiny regularization parameter in order to encounter as little bias as possible and thus stick close to an "OLS fit". Results from this experiment are summarized in Figure 1.6. The three columns correspond to the three possible regimes for the sample size. The first row shows examples of fits with p in $\{1, 5, 20\}$, for a single data set. Data points used for the training set are plotted in plain black, while a set of 1,000 points composing the (unreachable) test set appears shaded black. The second row shows the estimated generalization error with a hundred replications of the experiment conducted in the first row, using the test set for evaluating the generalization error.

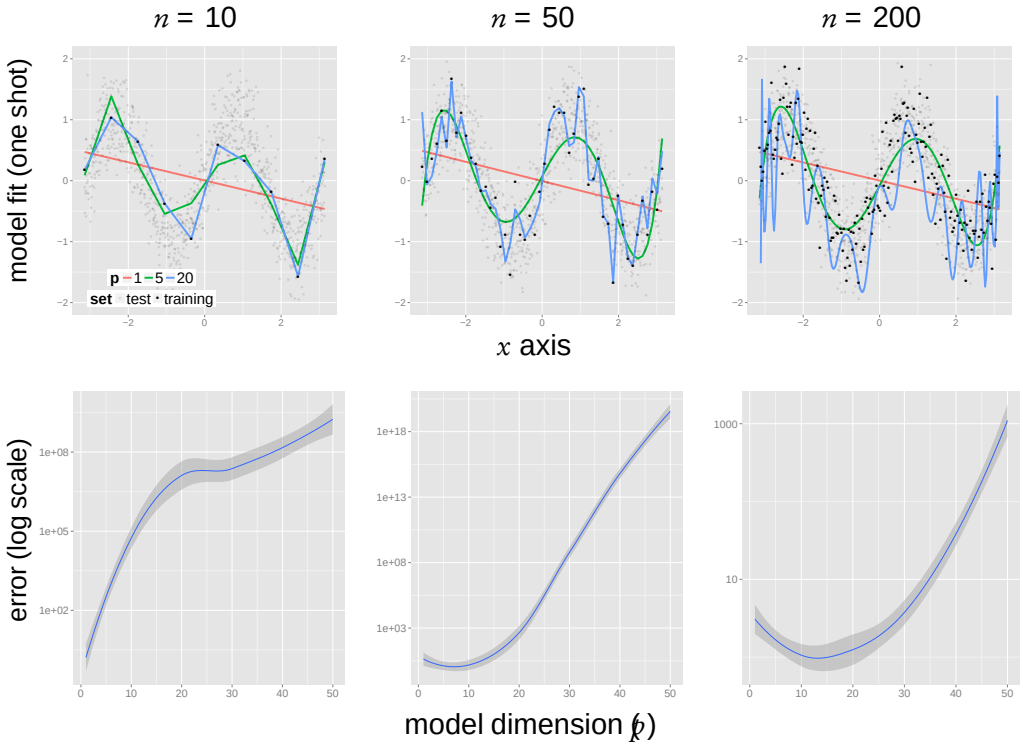


Figure 1.6 – *Overfitting is especially at play with high-dimensional data.*

What conclusion can be drawn from this experiment? First, we obviously do not need to be in the $n < p$ setup to overfit: considering large feature spaces is enough. Second, we see that this phenomenon is exacerbated when n gets smaller compared to p. More precisely, consider the case when $n = 10$: from the estimated test error (bottom left), the model which generalizes the best is the simple regression line, which is far from the one that truly generates the data (see the test set in the first row, showing a sinusoidal relationship). But there are so few data points that models lying in feature spaces with moderate dimension (e.g. $= 5$) – thus close to the true underlying generative model – already overfit and show large variance, causing a poor generalization.

This example supports the use of simple models, with potentially high bias but well controlled variance, as they seem to be sufficient to capture the main tendency of data sets that lie in high-dimensional feature spaces.

Modeling and interpretability issues.

The vast number of features is again at play regarding the modeling side of statistics, and consequently the way we interpret the fits. Typically, models involving many or ill-assorted features conflict strongly with our common sense; even if they show good predictive performance or summarize the data well, there are many application fields, notably in genomics and biology, where the interpretation of the fitted model is as important as its performance. In supervised problems, the features sought are those having a strong relationship (ideally causal) with the target response. In unsupervised problems, the objective is to unravel the underlying structure between the features themselves, which structure rules the observed system. To this end, the statistician should rely on the tools available in statistical learning for feature selection or feature extraction, the utility of which becomes even more important when the number of features grows. But again, traditional tools have to be rethought since they are not always calibrated to extract relevant information from data that live in large spaces.

To support this statement, we bring together the observations made on our two preceding numerical experiments: in Figure 1.5, we assess that, in most computational biology experiments, the dimension of the feature space that rules the underlying process is much lower than the number of features considered. The question is thus to find this underlying space, which may be done by means of feature selection or feature extraction techniques. However, in high-dimensional problems, the model which is the closest to the generative model – or to the biological process underneath – might not be the one that generalizes the best, due to the scarcity of data. This has been illustrated in Figure 1.6, where the straight line shows the smallest generalization error but is far from the true underlying model. Hence, one must be extremely careful when interpreting models fitted in a high-dimensional setup. This should especially be kept in mind in genomics where we often deal with medical data and where the temptation to interpret the inferred relationships as causal is huge.

A possible way to remedy the risk of incorrect interpretation is to adapt feature selection and extraction methods to only explore subspaces that are plausible according to the underlying biological process. In other words, we should add some constraints to the models by means of structural information that expresses our prior knowledge. This remark advocates for using statistical models that impose special structures on the features. Indeed, structure integration in the model should lead to more *interpretable models*, which is mandatory when dealing with millions of features.

§

This part has provided insights on the limitations of the traditional methods towards analyzing modern complex data. We thus hopefully have a good idea of the requirements of modern statistical methods. The next section basically justifies the general strategy constituting the backbone of all the research works presented in this thesis, in an effort to provide the community with methods fulfilling these requirements.

1.2.2 Marrying statistics and optimization

This section starts by summarizing the most desirable requirements of modern statistical methods regarding the challenges discussed so far. Then, I describe how, by bringing together tools from statistics and optimization, efficient strategies have emerged to tackle these challenges. In particular, I detail a typical strategy involving regularization and convex sparse methods, which are a central building block of my contributions.

What do we need?

Regarding the computational, statistical and interpretability issues mentioned in Section 1.2.1, an ideal method would be one fulfilling the following principles, which apply whatever the learning task (regression, classification, clustering).

1. *Favor simple models.* The use of simple models is mandatory in order to avoid overfitting, especially at stake in high-dimensional spaces. In other words, we should ban overly complex models which generalize badly when data is scarce, or at least strongly control their variance. Moreover, the use of simple models typically limits the computational burden.

2. *Favor models involving interpretable structures.* Models fitted in high-dimensional spaces should be cautiously interpreted. A possible way to limit the risk of bad interpretations is to rely on statistical models involving strong relationships between the variables, with easily absorbed representations (such as hierarchies, orderings or conditional dependencies).

3. *Perform dimension reduction.* Even when using simple models and interpretable structures of representation, the number of variables associated with the many features at hand should be controlled. Hence, we look for methods that reduce the original feature space, by performing feature extraction or feature selection jointly with the original task (prediction, classification or clustering). On top of favoring interpretability, this also controls the complexity of the models.

4. *Account for prior knowledge.* The methods should be flexible enough to allow the integration of prior information related to the underlying feature space. Hence, by biasing the feature extraction or selection processes, we hope to enhance both the interpretability of the model and the predictive performance.

5. *Favor algorithms with low/ controllable complexity.* The algorithms must show a globally low complexity regarding the dimension of the feature space. On top of this, we should favor methods the optimization of which can adapt to the problem size, achieving a tradeoff between accuracy and complexity that depends on the problem dimension (see Section 1.2.1).

Sparsity, regularization and convex optimization

In order to develop procedures that meet these prerequisites, popular methods have recently emerged in closely related fields such as statistics, machine learning and compress sensing. They all aim at revisiting standard statistical approaches from the angle of optimization, by changing the original problem from this renewed point of view. Let us attempt an outline of the strategy commonly followed by these proposals:

1. *Basic model choice*: among the possible statistical models answering a given question, favor one fulfilling the simplicity and interpretability principles.

2. *Formulation as an optimization problem*: make the criterion optimized by the method explicit, typically as a negative (log)-likelihood or a loss function.

3. *Problem modification*: alter the problem by adding/removing constraints or by modifying the fitting term (e.g., by convex relaxation). Such a modification is called *regularization*.

The objective of regularization is plural: by modifying the problem, we hope to (*i*) get a better control of the computational cost, (*ii*) account for prior knowledge, (*iii*) perform dimension reduction, and (*iv*) control the model complexity; that is, most of the principles argued in the preceding enumeration.

§

We now review three examples of application of this strategy in different contexts (regression analysis, multivariate analysis and clustering analysis) that have given birth to a wide number of papers in the past decade.

Example 1.6. *Variable selection in linear regression*
Consider the canonical regression problem with possibly many features: we aim at predicting the vector of outcomes $\mathbf{y} = (y_1, \dots, y_n)$ from the $n \times p$ data matrix $\mathbf{X}$, the jth column of which contains measurements related to the jth feature. The simplest – and highly-interpretable – conceivable model in the regression setup is the Gaussian linear model: we assume a linear relationship between the features and the outcome with an *iid* Gaussian vector of noise $\boldsymbol{\varepsilon}$ and we estimate the coefficients β_0 and $\boldsymbol{\beta}^\star$ in

$$\mathbf{y} = \beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}, \qquad \mathbb{E}\boldsymbol{\varepsilon} = 0_p, \quad \mathrm{Var}\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}_p.$$

If not assuming any special structure for the feature space – and thus for $\boldsymbol{\beta}^\star$ –, the most standard estimation strategies are ordinary least square or maximum likelihood, which both solve the same optimization problem by minimizing the residual sum of squares (RSS). When p is large, a natural assumption is to consider that only a few features explain the outcome, that is, that many entries of $\boldsymbol{\beta}^\star$ are zero. In other words, $\boldsymbol{\beta}^\star$ is sparse. Variable selection can be performed by solving the following optimization problem[2], which minimizes the RSS under the constraint that the number of non null entries in $\boldsymbol{\beta}$ equals an integer s:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimize}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \quad \text{such that} \quad \sum_{j=1}^{p} \mathbb{1}_{\{\beta_j \neq 0\}} = s, \quad s \in \{0, \dots, p\}.$$

This is a non-convex constraint optimization problem which can be solved by fitting the 2^p models such that $\boldsymbol{\beta}$ contains s non null entries. Efficient branch-and-bound algorithms allow us to avoid testing all possible models, but this combinatorial problem is intractable even for moderate p (say, $p > 30$). A widely-spread popular idea for reducing

[2] We omit the intercept for clarity.

the computational burden for large values of p is to replace the pseudo non-convex ℓ_0-norm by its closest convex surrogate, the ℓ_1-norm:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \|y - X\beta\|_2^2, \quad \text{such that} \sum_{j=1}^{p} |\beta_j| \leq s, \quad s > 0.$$

This is known as the "Lasso" in statistics [164, 20] or "basis-pursuit" in compress-sensing [28]. Note that when the sample size $n < p$, there is an infinite number of solutions to this problem, which can be further regularized by adding a second constraint based on the ℓ_2 norm [77, 198]:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \|y - X\beta\|_2^2, \quad \text{such that} \sum_{j=1}^{p} |\beta_j| \leq s_1, \sum_{j=1}^{p} \beta_j^2 \leq s_2, \quad s_1, s_2 > 0.$$

There are plenty of ways to modify this problem, especially to account for more complex structures of the underlying true feature space (a small sample [166, 72, 190, 81, 87]). These ideas are developed further in Chapter 3, especially to account for some underlying structures that rule the process observed on the fitted data.

Example 1.7. *Feature extraction in high-dimensional space*
Consider X an $n \times p$ data matrix of features where the p features live in a possibly high-dimensional space and are typically correlated. In other words, the rank of X is much smaller than p. We would like to find a small subspace with size k which is more informative than the original space, hoping for a better interpretation. This is the objective of feature extraction. The most basic – yet still fundamental and powerful – tool from multivariate analysis to perform this task is principal component analysis (PCA), which maps X to a new subspace spanned by k orthogonal (meaning uncorrelated) features while minimizing the reconstruction error of X.

More precisely, denote by T $(n \times k)$ the coordinates in the new space and β $(p \times k)$ the orthogonal linear map (a rotation) transforming the original variables into the new variables. We have $T = X\beta$, $\beta^\top \beta = I_k$ and the corresponding reconstructed X is $\hat{X}^{(k)} = T\beta^\top$. Hence, the PCA finds T and β by solving the following constraint optimization problem:

$$\underset{\beta \in \mathbb{R}^{p \times k}, T \in \mathbb{R}^{n \times k}}{\text{minimize}} \|X - T\beta^\top\|_F^2, \quad \text{such that} \quad \beta^\top \beta = I, \tag{1.2}$$

where $\|\cdot\|_F$ stands for the Frobenius norm. The solution is well known: it is obtained by performing the singular value decomposition of X and truncating the factorization to the first k largest singular values. More precisely, the SVD decomposition of X is

$$X = U\Sigma V^\top,$$

where Σ is a diagonal matrix, U and V are orthogonal matrices with respective sizes $m \times m$, $n \times m$ and $m \times m$ where $m = \min(n, p)$. The rank-k approximation $\hat{X}^{(k)}$ is obtained by restricting U and V to their first k columns and Σ to its first $k \times k$ block. The respective sizes of the restricted matrices $\tilde{U}$, $\tilde{V}$ and $\tilde{\Sigma}$ are $n \times k$, $p \times k$ and $k \times k$, and the solution to (1.2) is

$$\hat{X}^{(k)} = T\beta^\top, \quad \text{with } T = \tilde{U}\tilde{\Sigma}, \quad \beta = \tilde{V}.$$

Now, consider that the initial number of features p is very large: PCA basically loses its appealing interpretability property, since the map Φ performs a linear combination of all the p features: in the same vein as in the sparse regression example 1.6, selecting among the set of features during the operation of feature extraction would produce a highly interpretable model, by performing the feature extraction only on a small set of variables. A possibility is to modify the optimization problem by directly working on the matrix decomposition as follows:

$$\underset{U \in \mathbb{R}^{n\times k}, V \in \mathbb{R}^{p\times k}, \mathrm{diag}(\Lambda)\in\mathbb{R}^{k}}{\text{minimize}} \quad \|X - U\Lambda V^\top\|_F^2, \quad \text{s.t. } V^\top V = I_k, \quad U^\top U = I_k, \quad \|v^j\|_1 \leq c_j,$$

where v^j is the jth column of V and $\|\cdot\|_1$ stands for the ℓ_1 norm. Resolution of this problem – and some variants – continues further in many papers related to *sparse principal component analysis* [88, 199, 38, 184].

Example 1.8. *Clustering analysis in high-dimensional space*

Our last example concerns the ubiquitous problem of clustering. A popular strategy is agglomerative clustering: starting from one data point per cluster, it successively merges clusters that are the closest according to a given distance, until all clusters merge together. This procedure is extremely appealing from the interpretation viewpoint since it produces a dendrogram, that is, a tree defining a hierarchy on the data points.

Concretely, suppose we are given an $n \times p$ data matrix X where we want to cluster the n points when p remains small but n is (possibly very) large.[3] In this case, classical agglomerative methods will fail due to an excessive algorithmic complexity. The question is, how can we regularize this problem in order to scale the dimension?

Contrary to the problems treated in the two preceding examples, hierarchical clustering is not defined by a statistical model, but by a heuristic ("merging close clusters"). Thus, in order to apply our strategy consisting in modifying the optimization problem at hand, we must find what criterion HCA tries to optimize. The answer to this question was given by [75]: for a given level in the hierarchy, tuned by a real parameter c, HCA tries to minimize the reconstruction error between the data X and the coefficient matrix Λ which encodes the clustering. To merge the coefficients – and thus cluster individuals –, the total number of different rows in Λ is constrained, *i.e.*,

$$\underset{\Lambda \in \mathbb{R}^{n\times p}}{\text{minimize}} \|X - \Lambda\|_F^2, \quad \text{such that} \sum_{i>j} \mathbb{1}_{\{\lambda_j \neq \lambda_i\}} \leq c.$$

Hence, for $c = n(n-1)/2$, all rows are different, $\Lambda = X$ and we are at the very bottom of the hierarchy. With $c < n(n-1)/2 - 1$, we force two rows to merge, thus performing the first step of HCA and so on. This optimization problem is combinatorial but [75] proposed a series of relaxed versions, in which cases the problem turns out to be convex:

$$\underset{\Lambda \in \mathbb{R}^{n\times p}}{\text{minimize}} \|X - \Lambda\|_F^2, \quad \text{such that} \sum_{i>j} \|\lambda_i - \lambda_j\| \leq c,$$

where $\|\cdot\|$ can be any ℓ_p norm. Several recent extensions to this work have been proposed ever since [81, 138, 30]. In particular, I have been working [JP2] on a weighted version of this problem that ensures good statistical properties, along with an $n \log n$ algorithm that scales to huge-dimensional spaces.

[3] Such problems occur when many features are observed for a small number of individuals, and when the features are to be clustered. In this case, the features become the observation. We "invert" the notation n and p compared to what was said up to now to remain coherent with classical notation in clustering.

These examples hopefully demonstrate that marrying statistics and optimization can lead to successful approaches in statistical learning tailored to analyzing modern data sets. At first glance, this framework is powerful because it combines the good properties of well-known statistical models and the computational power of (convex) optimization. Still, it is more than a mere reformulation of a statistical problem into an optimization problem. It also provides methods with great flexibility for modeling the problem at hand. Specifically, I think that the most sensitive point in the aforementioned strategy lies in the third point,*i.e.*, *Problem modification*: indeed, one should modify the problem in order to find a good balance between computational and statistical performance, but also to achieve a better modeling of the problem. This latter point is especially important in application fields such as genomics, where integrating the structure of the problem can have dramatic effects on the performance and on the interpretability of the fit. The next paragraph illustrates this point, as accounting for the structure of the data within the framework of regularization and sparse methods characterizes most of my contributions.

Sparse and regularization approaches to account for complex data structure

Regularized problems that we consider can be cast as the following general constraint optimization problem

$$\underset{\theta \in S}{\text{minimize}}\ f(\theta; \text{data}), \quad \text{such that} \quad \Omega(\theta) \leq c, \tag{1.3}$$

where θ is the set of parameters of interest living in the space S. The set Ω describes the *feasible-set*, forcing the parameters for living in a subspace that we deem "adequate". If we choose f a convex function and Ω and S some (possibly non-smooth) convex sets, things get easier both on the computational and statistical sides, as we have many tools from the optimization literature [18] at our disposal. Thus, "convexification" is the typical modification done to the original problem in order to perform regularization. Still, as argued in the previous paragraph, computational and statistical motivations should not minimize the interest towards a better modeling of the problem at hand. We illustrate this on the idealistic Picture 1.7: we have a two-dimensional set of parameters θ, a convex function f to minimize and a feasible set. Of course, the choice of Ω is

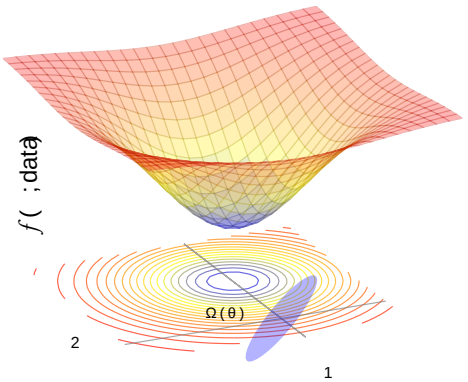


Figure 1.7 – *Idealistic two-dimensional constrained convex problem*

very important: it may be chosen in order to simply regularize the problem, meaning giving the problem a solution – as in the original motivation of ridge regression –. But

it can also be used to integrate structural information. Often, this information can be recovered from the data themselves. Most of my research in genomics has been driven by such motivation, that is, developing regularization techniques or statistical methods that account for an underlying structure of the data in the fit. The simple, basic example that follows advocates for such a choice.

A toy example advocating for structured regularization. This numeric illustration builds on the beginning of Chapter 18 of *The Elements of Statistical Learning* (second edition) [70], dedicated to high-dimensional problems. The authors design a numerical example to show that "simple" regularization (exemplified by ridge regression) fails to recover the true interesting parameters in the model when the number of features p gets too large compared to the sample size n, although more regularization helps in improving predictive performance, lack of information may mislead the method in finding the relevant features. The idea to improve these methods is to find the appropriate shape of the regularization, *i.e.* of in Figure 1.7, by relying on contextual knowledge. In the following, I revisit their example by introducing the presence of structure between the predictors and show how to simply build a regularizer that improves both interpretability and predictive performance.

We choose a simple block diagonal setting as a structure between the predictors that mimics the correlation structure typically found between SNP, just as in the right panel of Figure 1.3b, Example 1.3. We split the p features into five groups with respective sizes $\{p=4, p=8, p=4, p=8, p=4\}$. We may represent this structure by an undirected graph G on the features the adjacency matrix of which is block diagonal with zeros on its diagonal. Under this assumption, we generate data $\{(x_i, y_i)\}_{i=1}^n$ with the linear model

$$y = X\beta^\star + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0_n, I_n \sigma^2)$$

such that the true underlying relationship between the response and the predictors follows the previously mentioned grouping pattern:

$$\beta^\star = \Big(\underbrace{0.25, \ldots, 0.25}_{p=4 \text{ times}}, \underbrace{1, \ldots, 1}_{p=8 \text{ times}}, \underbrace{0.25, \ldots, 0.25}_{p=4 \text{ times}}, \underbrace{1, \ldots, 1}_{p=8 \text{ times}}, \underbrace{0.25, \ldots, 0.25}_{p=4 \text{ times}}\Big). \tag{1.4}$$

We sample the predictors from a Gaussian multivariate distribution $x_i \sim \mathcal{N}(0_p, \Sigma)$ where Σ is defined blockwise with the same pattern as $\beta^\star$, with inner group correlations such that

$$\Sigma = \quad \text{with} \quad \Sigma_{ij} = \begin{cases} 1 & i = j, \\ .25 & i, j \in \text{blocks}\{1,3,5\}, i \neq j, \\ .75 & i, j \in \text{blocks}\{2,4\}, i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

We investigate cases where $p = 16, 192, 2048$ with $n = 200$ for the learning set and 1000 points in the test set that we keep to estimate the prediction error. We set the value of σ in order to meet a coefficient of determination $R^2 \approx 0.8$.

The regularized method that we consider is a "structured" version of ridge regression, fitted by solving the optimization problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \|y - X\beta\|_2^2, \quad \text{such that } \beta^\top L \beta \leq c, \tag{1.5}$$

where L is chosen to account for our knowledge on the features. This is a special case of Problem 1.3, such that changing L typically changes the shape of Ω in Figure 1.7.

When L is positive semi-definite, the problem is convex and is equivalently stated in its Lagrangian form, which we also call the "penalized version" of the problem. Indeed, the regularization term is often called a penalty term in this case. Here the solution is derived analytically and we have

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|y - X\beta\|_2^2 + \beta^\top L \beta \right\} = (X^\top X + L)^{-1} X^\top y. \tag{1.6}$$

This analytic expression gives insight into how this simple structured regularizer biases the estimator, by performing a mixture between the empirical covariance $X^\top X$ and our prior knowledge L, like in Bayesian regression with the posterior mean [1.18]. Choosing $L = I$ leads to the usual ridge estimator. More generally, assume that the contextual knowledge about the relationships between the features can be described through a weighted graph encoded in a weighted adjacency matrix $W = (w_{ij})_{i,j=1,\dots,p}$, with $w_{ij} \geq 0$. This prior information can be integrated by means of the combinatorial graph Laplacian (see [32]). Denoting by $\deg_i = \sum_k w_{ik}$ the degree of node i, the Laplacian matrix $L = (l_{ij})_{i,j=1,\dots,p}$ is defined by

$$l_{ij} = \begin{cases} \deg_i & \text{if } i = j, \\ -w_{ij} & \text{otherwise.} \end{cases}$$

Expanding the penalty term, we see that we encourage regression coefficients corresponding to connected features to be the same:

$$\beta^\top L \beta = \sum_{i,j} w_{ij} (\beta_i - \beta_j)^2.$$

Regarding the fitting cost, ridge regression is also appealing since it can be computed at the cost of a single singular value decomposition. In the structured version, we also have to factorize L, but it is straightforward to show that

$$\hat{\beta}^{\text{ridge}} = (X^\top X + L)^{-1} X^\top y = L^{-1/2} V (D^2 + I)^{-1} D U^\top y \tag{1.7}$$

where $XL^{-1/2} = UDV^\top$ and $L^{-1/2}$ is understood in the matrix sense.[4] Of course, $(D^2 + I)^{-1} D$ is diagonal and can be computed for a series of $\lambda > 0$.

Finally, we may compute the effective degrees of freedom of the generalized ridge fit, in the sense proposed by Efron and Hastie for regression. Effective degrees of freedom is a far more interpretable quantity for evaluating model complexity than is the λ parameter. For a linear smoother (as in ridge regression), it can be computed as the trace of the "hat" matrix, the computation of which further simplifies thanks to the SVD:

$$\text{df}(\hat{y}^{\text{ridge}}) = \text{Tr}\left[X (X^\top X + L)^{-1} X^\top \right] = \sum_{i=1}^{p} \frac{d_{ii}^2}{d_{ii}^2 + \lambda}. \tag{1.8}$$

Let us now comment the numerical results displayed in Figure 1.8. We investigate for the three values of p ($p \ll n$, $p \approx n$, $p \gg n$) the behavior of three ridge estimators

[4] For some special graphs, $L^{-1/2}$ can be computed analytically. Otherwise, a Cholesky decompostion may be used.

on 100 replicated simulations where we evaluate the prediction error with the test set, on a large grid of λ values. Error has been normalized according to the Bayes error σ^2 so that the minimum achievable is always 1. The three ridge variants are: i) the standard one (with no structure), ii) the one embedding the perfect structure (provided by the graph with the exact block diagonal pattern) and finally iii) one which integrates a structure inferred on the data set. To this end, we perform a hierarchical clustering (complete linkage) on a distance based on the empirical correlation matrix (1 minus the absolute value of the empirical correlation). We cut the dendrogram in order to obtain 5 groups, which is the correct number of groups in the simulated block diagonal structure. We use these blocks to build the graph and the associated graph Laplacian to integrate the structural information inferred from the data.

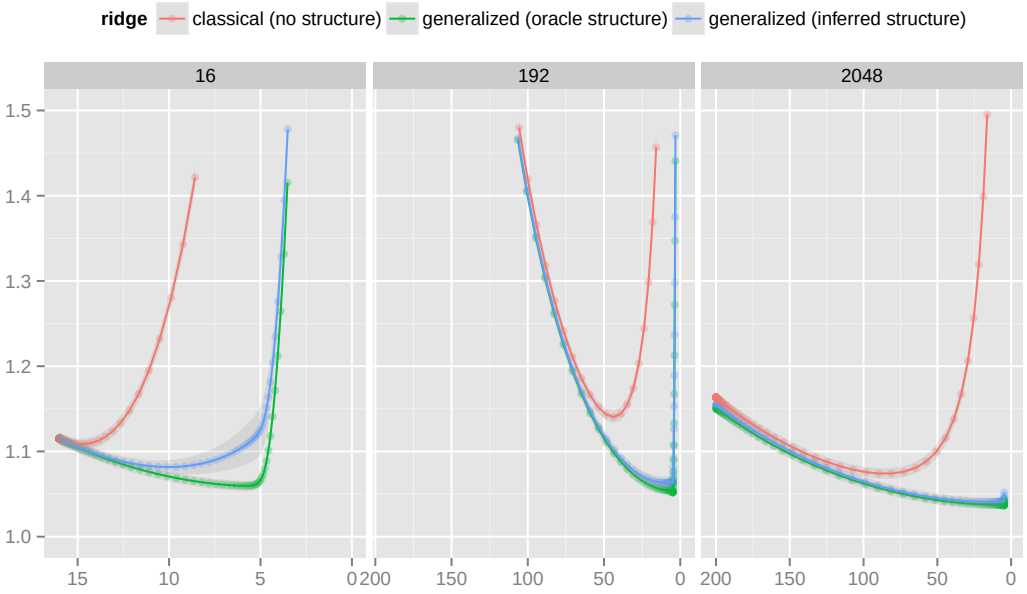


Figure 1.8 –*Toy example: structuring ridge regression leads to efficient regularization (correct model, lower generalization error).*

Several comments can be made: first, the standard ridge regression outperforms the OLS even in the $p \gg n$ and $p \simeq n$ cases, because it takes into account the strong correlations between the features. Still, the optimal model in terms of prediction error never corresponds to the true model, that is, the one having 5 degrees of freedom (corresponding to the 5 groups of features), and standard ridge regression overestimates the model complexity, because it requires many coefficients to maintain a relatively low error. Conversely, when the regularization is guided by structural information, not only do we lower the prediction error, but we also find an optimal model close to the true complexity. This holds true even when the structure is inferred with a straightforward method from the data themselves. Though somewhat ideal, this simple numerical example advocates for integrating the structure within the regularization process as soon as possible when dealing with high-dimensional data. This can be done within a framework where statistics meet constraint optimization, which offers great facilities for enhancing computational and modeling aspects of the standard statistical procedures.

1.3 RESEARCH OVERVIEW

To conclude this introductory chapter, let me bring together a series of research themes I am concerned with. Of course, the research questions addressed there are motivated by issues related to the modern data setting depicted in Section 1.1 and by the methodological tools mentioned in Section 1.2. There is a large and growing literature on these themes and many brilliant research teams are working on these topics. At this stage of the manuscript, I only provide references to my own papers as my goal is to set my contributions in the present landscape of statistical learning. References to related works come in the next chapters where my contributions are detailed.

1.3.1 Themes

Interpretable models. Due to the combinatorial explosion of the possible relationships between variables in high-dimensional spaces, I work with simple and highly interpretable models to represent links between variables.

(Gaussian) Graphical models (GGM) are among them: they depict the conditional dependency structure of a random vector by means of a graph. As conditional dependence is a good statistical modeling of direct relationships between variables, and graphs are convenient for interpretation, GGM are extremely popular. I worked with such models in both unsupervised and supervised frameworks. In the unsupervised setting, I used them to describe strong interactions between biological actors in the cell, e.g. in regulatory networks [JP7, JP8, JP9] and more generally to infer (partial) covariance structures in high-dimensional spaces. In the supervised setting, I used a conditional form of GGM to revisit multivariate regression and distinguish weak from strong interactions between responses and predictive variables in various application fields [PP2].

Hierarchies (or trees) are another convenient tool to depict interpretable structures between variables, because they summarize well the relationships between the variables and exhibit potential clusters. They are especially attractive when the number of variables is large. However, their reconstruction may be cumbersome in situations where the number of variables is very large. I worked on such an issue in a recent work [JP12].

In short, I favor models involving structures of relationships simple enough to exhibit the most important trends governing the data.

Modeling data heterogeneity. Accounting for the heterogeneity of data in modeling is another side of my research.

In the context of unsupervised learning and GGM inference, I included [JP9, JP8] a latent modeling of the target GGM by assuming a stochastic block model on the unobserved conditional graph to enhance the inference of its structure. In the same framework, I proposed in [JP7] a method for inferring multiple GGM by accounting for similarities and differences between several samples, in order to account for their heterogeneity.

In supervised learning, natural tools to account for data heterogeneity are multivariate regression and its mixed model counterpart, where we typically describe the structure of the sample and of the signal *via* a complex covariance structure. The multivariate regularization scheme that I proposed in [PP2], motivated by applications in genetics, fits in such a framework by including population heterogeneity.

A more complicated question occurs when one considers heterogeneous data in the sense of data collected from multiple platforms, in order to perform *horizontal* data integration. An unpublished proposal is made in Section 2.2.4 in the GGM setting again, by introducing a multi-attribute framework along with some inference procedures.

Finally, I have recently been working on a model to better understand some of the heterogenenity at stake in cancer data and tumor cells. To this end, we develop [JP16] a statistical methodology to identify the misregulated genes given a reference network and gene expression data. The objective is then to characterize cancer subtypes according to these misregulations, some of which are out-of-reach of classical techniques such as differential analysis.

Structured and sparse methods, prior integration. To perform inference and selection of the important links between variables in order to estimate and reconstruct the target structures in the models (hierarchies, graphical models, and so on), I rely on sparse and regularization methods, which are ubiquitous in the statistical learning community.

I proposed some new regularization schemes in the regression framework when the set of parameters is endowed with a group structure, in the same vein as the group-Lasso [JP6, JP7]. I also worked on sparse methods for network inference that bias the reconstructed graph toward an underlying topology [JP9, JP7]. I more recently worked on regularized methods for multivariate regression allowing prior integration for various problems in genomics [PP2]. In an ongoing work [JP11], I am working on a formulation at the edge of Bayesian regression and frequentist formalism for variable selection.

In my opinion, structured sparsity and methods for prior integration are the natural inference tools coming with the interpretable statistical models adapted to high-dimensional data, which explains that they are at the heart of my research.

Convex methods and efficient algorithms. Closely related to sparse and regularization methods is the development of efficient algorithms. Statisticians working with modern data sets cannot be unaware of the state-of-the-art optimization procedures.

A natural way to achieve low complexity is to rely on convex methods. This framework is adopted in most of my works. There are two options to achieve convexity when designing a new (sparse) inference procedure: the first possibility is to directly choose a statistical model with a convex loss and then "craft" the regularization accordingly, as I did in [JP6, JP7, PP2]. A second option is to rely on a classical estimator defined by a non-convex criterion and find an appropriate convex surrogate. By convexifying an optimization problem, we hope for both a drastic decrease in the computational cost and an indirect regularization which provides the model with finer interpretation and better performance than does the original criterion. I explore this option for convex clustering and (M)ANOVA in [JP2].

Even in the convex case, dedicated implementations are often needed since efficiency of an optimization procedure strongly depends on the data regime or on the structure of the estimator. I address these questions by discussing trade-offs between accuracy and performance in [PP3]. Recently I have also been working on a fast implementation of the LARS for detecting change-points in two-dimensional data [RP1].

Of course, the need for efficiency is ubiquitous, beyond convex methods. In this perspective, I am involved in a research project that aims at developing efficient re-

sampling procedures for high-dimensional data, by deriving for instance closed-form formula for cross-validation. A preliminary work has been presented in [GN1].

As a more general comment about the computational side of my work, I try to maintain implementations by providing the community with R/C++ packages [SW5, SW4, SW3, SW1, SW2]. This is also a manner to promote reproducible research.

Statistical analysis. With sparse models, sparse regularized learning methods and new data settings, new tools for relevant statistical analysis are needed.

The first question is to characterize the estimators arising from regularized and convex methods, and to provide statistical guarantees in the high-dimensional setup. Many authors at the edge of the statistical community and of the machine learning community tried to tackle these issues, working on prediction performance, estimation performance and support recovery for sparse estimation. I studied such properties for the newly-defined penalized method for linear regression known as "cooperative-Lasso" in [JP6]. More recently, I worked in [JP2] on a penalized version of the ANOVA where support recovery is at stake, and where good statistical and computational properties are exhibited with techniques from high-dimensional statistics and convex analysis.

Another question is to develop measures of performance that bring all indicators together to simultaneously characterize numerical performance and statistical accuracy. I humbly and partially explored such questions in [PP3].

1.3.2 Organization of the manuscript

This series of research themes is disseminated in the contributions composing this thesis. I chose to organize them into two chapters that can be read almost independently.

The first one is dedicated to my work on Gaussian graphical models and network reconstruction: I develop several inference procedures to reconstruct the conditional structure associated with these models, motivated by the characteristics of genomics data.

The second chapter presents a series of sparse and regularization methods that embed in various possible ways the structural information related to fields of application, which are diverse but where life science has a place of choice.

As sparse methods are at the heart of my preoccupation and since I work on amending those regularizing/penalizing terms according to the data themselves, I chose to entitle this manuscript *Contributions to Sparse Methods for Complex data analysis*.

2 Sparse Gaussian Graphical Models for Network Inference

"Look! A trickle of water running through some dirt! I'd say our afternoon just got booked solid!"

Bill Watterson

Contents

2.1 Background 37
2.1.1 Basics on Gaussian graphical models 37
2.1.2 Sparse methods for GGM inference 38
2.2 Contributions 44
2.2.1 Accounting for latent organization of networks 44
2.2.2 Accounting for sample heterogeneity 49
2.2.3 Accounting for time-course data 53
2.2.4 Accounting for multiscale data: multi-attribute GGM 55
2.3 Perspectives 60

This chapter proposes an overview of my contributions to Gaussian Graphical Models (GGM), in terms of modeling and more importantly in terms of inference of the conditional structure associated with such models. After a quick outline of the existing literature and an introduction to the most popular methods in the community, I present my contribution to this field. These contributions were motivated by the need to account for some special features characterizing genomics data and biological networks.

2.1 BACKGROUND

Gaussian Graphical Models (GGMs) [101, 180] are a very convenient tool for describing the patterns at play in complex data sets. Indeed, through the notion of partial correlation, they provide a well-studied framework for spotting direct relationships between variables, and thus reveal the latent structure in a way that can be easily interpreted. Application areas are very broad and include for instance gene regulatory network inference in biology (using gene expression data) as well as spectroscopy, climate studies, functional magnetic resonance imaging, etc. Estimation of GGMs in a sparse, high-dimensional setting has thus received much attention recently. This section provides an overview of this hot and competitive research field of statistical learning. I mainly focus on the state-of-the-art ℓ_1-regularization methods and their most recent striking variants, insisting on their computational and statistical properties. This provides the reader with the necessary material to approach the second section of this chapter dedicated to my personal contributions to this field.

2.1.1 Basics on Gaussian graphical models

Let $V = \{1, \ldots, p\}$ be a set of fixed vertices and $X = (X_1, \ldots, X_p)^\top$ a random vector describing a signal over this set. The vector $X \in \mathbb{R}^p$ is assumed to be multivariate Gaussian with unknown mean and unknown covariance matrix $\Sigma = (\sigma_{ij})_{(i,j)\in V^2}$. No loss of generality is involved when centering X, so we may assume that $X \sim \mathcal{N}(0_p, \Sigma)$. The covariance matrix Σ, equal to $\mathbb{E}(XX^\top)$ under the assumption that X is centered, belongs to the set $\mathbb{S}_p^+$ of positive definite symmetric matrices of size $p \times p$.

Graph of conditional dependencies. GGMs endow Gaussian random vectors with a graphical representation G of their *conditional dependency structure*: two variables i and j are linked by an undirected edge (i, j) if, conditional on all other variables indexed by $V \setminus \{i, j\}$, random variables X_i and X_j remain or become dependent. Thanks to the Gaussian assumption, conditional independence actually boils down to a zero conditional covariance $\mathrm{cov}(X_i, X_j | X_{V \setminus \{i,j\}})$, or equivalently to a zero partial correlation which we denote by ρ_{ij}, the latter being a normalized expression of the former.

Concretely, the inference of a GGM is based upon a classical result originally emphasized in [42] stating that partial correlations ρ_{ij} are actually proportional to the corresponding entries in the *inverse* of the covariance matrix $\Sigma^{-1} = \Theta$, also known as the *concentration matrix*. More precisely, we have

$$\rho_{ij} = -\tilde{\theta}_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\,\theta_{jj}}}, \qquad \theta_{ii} = \mathrm{Var}(X_i | X_{V \setminus i})^{-1}, \tag{2.1}$$

thus Θ directly describes the conditional dependency structure of X. Indeed, after a simple rescaling, Θ can be interpreted as the adjacency matrix of an undirected weighted graph representing the partial covariance (or correlation) structure between variables $X_1, \ldots, X_p$. Formally, we denote by $G = (V, E)$ this graph, the edges of which are characterized by

$$(i, j) \in E \iff \theta_{ij} \neq 0, \quad \forall (i, j) \in V^2 \text{ such that } i \neq j.$$

In words, G has no self-loop and contains all edges (i, j) such that θ_{ij} is nonzero. Therefore recovering nonzero entries of Θ is equivalent to inferring the graph of con-

ditional dependencies $\mathcal{G}$, and the correct identification of nonzero entries is the main issue in this framework.

Maximum Likelihood inference. GGMs fall into the family of exponential models for which the whole range of classical statistical tools applies. As soon as the sample size n is greater than the number p of variables, the likelihood admits a unique maximum over $\mathcal{S}_p^+$, defining a maximum likelihood estimator (MLE): suppose we observe a sample $X^1, \ldots, X^n$ composed of n i.i.d. copies of X, stored row-wise once centered in a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ such that $(X^i)^\top$ is the ith row of $\mathbf{X}$. The empirical covariance matrix is denoted by $\mathbf{S}_n = \mathbf{X}^\top\mathbf{X}/n$. Maximizing the likelihood is equivalent to

$$\hat{\Theta}^{\text{mle}} = \arg\max_{\Theta \in \mathcal{S}_p^+} \log\det(\Theta) - \text{Tr}(\mathbf{S}_n\Theta). \tag{2.2}$$

When $n > p$, Problem (2.2) admits a unique solution equal to $\mathbf{S}_n$. The scaled empirical covariance matrix $\mathbf{S}_n$ follows a Wishart distribution while its inverse $\mathbf{S}_n^{-1}$ follows an inverse Wishart distribution with computable parameters.

There are two major limitations with the MLE regarding the objective of graph reconstruction by recovering the pattern of zeroes in Θ. First, it provides an estimate of the saturated graph: all variables are connected to each other; second, we need n to be larger than p to be able to even define this estimator, which is rarely the case in genomics. In any case, the need for regularization and feature selection is huge. A natural assumption is that the true set of direct relationships between the variables remains small, that is, the true underlying graph is sparse (say, of the order p rather than the order of p^2). Sparsity makes estimation feasible in the case where $n < p$ since we can concentrate on sparse or shrinkage estimators with fewer degrees of freedom than in the original problem. Henceforth, the question of selecting the correct set of edges in the graph is treated as a question of variable selection.

High-dimensional inference of GGM. The different methods for the inference of sparse GGMs in high-dimensional settings fall into roughly three categories. The first contains constraint-based methods, performing statistical tests [124, 45, 46, 93, 181]. However, they either suffer from the excessive computational burden [124, 181] or strong assumptions [45, 46] that correspond to regimes never attained in real situations. The second of these categories is composed of Bayesian approaches, see for instance [43, 89, 141, 151]. However, constructing priors on the set of concentration matrices is not a trivial task and the use of MCMC procedures limits the range of applications to moderate-sized networks. The third category contains regularized estimators, which add a penalty term to the likelihood in order to reduce the complexity or degrees of freedom of the estimator and more generally regularize the problem: throughout this chapter, I focus on methods of this kind. More precisely, I focus on ℓ_1-regularized procedures, which are freed from any test procedure – and thus multiple testing issues – since they directly perform estimation and selection of the most significant edges by zeroing entries in the estimator of Θ. The remainder of this section is dedicated to a quick review of the state-of-the-art methods of this kind.

2.1.2 Sparse methods for GGM inference

The idea underlying sparse methods for GGM is the same as for the Lasso in linear regression (see Example 1.6, Section 1.2.2): it basically uses ℓ_1 regularization as a convex

surrogate of the ideal but computationally intensive ℓ_0-regularized problem:

$$\arg\max_{\Theta \in \mathbb{S}_p^+} \log\det(\Theta) - \mathrm{Tr}(S_n\Theta) - \rho\|\Theta\|_0. \tag{2.3}$$

Problem (2.3) achieves a trade-off between the maximization of the likelihood and the sparsity of the graph within a single optimization problem. The penalty term can also be interpreted as a log prior on the coefficients in a Bayesian perspective. BIC or AIC criteria are special cases of such ℓ_0 regularized problems, except that the maximization is made upon a restricted subset of candidates $\{\tilde{\Theta}_1, \ldots, \tilde{\Theta}_m\}$ and the choice of ρ is fixed ($\log(n)$ for BIC and 1 for AIC). Actually solving (2.3) would require the exploration of all possible 2^{p^2} graphs. On the contrary, by preserving the convexity of the optimization problem, ℓ_1-regularization paves the way to fast algorithms. For the price of a little bias on all the coefficients, we get to shrink some coefficients to exactly 0, operating selection and estimation in one single step as hoped in Problem (2.3).

Graphical-Lasso. The criterion optimized by the graphical-Lasso was simultaneously proposed in [19] and [7]. It corresponds to the estimator obtained by fitting the ℓ_1-penalized Gaussian log-likelihood, *i.e.* the tightest convex relaxation of (2.3):

$$\hat{\Theta}^{\text{glasso}} = \arg\max_{\Theta \in \mathbb{S}_p^+} \log\det(\Theta) - \mathrm{Tr}(S_n\Theta) - \rho\|\Theta\|_1. \tag{2.4}$$

In this regularized problem, the ℓ_1-norm drives some coefficients of Θ to zero. The non-negative parameter ρ tunes the global amount of sparsity: the larger ρ, the fewer edges in the graph. A large enough penalty level produces an empty graph. As ρ decreases towards zero, the estimated graph tends towards the saturated graph and the estimated concentration matrix tends towards the usual MLE (2.2). By construction, this approach guarantees a well-behaved estimator of the concentration matrix *i.e.* sparse, symmetric and positive-definite, which is a great advantage of this method.

Ever since Criterion (2.4) was proposed, many efforts have been dedicated to developing efficient algorithms for its optimization. In the original proposal of [7], it is shown that solving for one row of matrix Θ in (2.4) while keeping other rows fixed boils down to a Lasso problem. The global problem is solved by cycling over the matrix rows until convergence. Thus, if one considers that L passes over the whole matrix are needed to reach convergence, a rough estimation of the overall cost is of the order of Lp (cost for solving for one row). With a block-coordinate update each iteration over a row has $O(p^3)$ complexity and their implementation is $O(Lp^4)$ for L sweeps over the whole matrix $\hat{\Theta}$. In [7] again, a rigorous analysis is conducted in Nesterov's framework [129] showing that the complexity for a single ρ reaches $O(p^{4.5}/\varepsilon)$ where ε is the desired accuracy of the final estimate.

The *Graphical-Lasso* algorithm of [57] follows the same line but builds on a coordinate descent algorithm to solve each underlying Lasso problem. While no precise complexity analysis is possible with these methods, empirical results tend to show that this algorithm is faster than the original proposal of [7]. Additional insights on the convergence of the graphical-Lasso are provided in [112, 120], simultaneously with [182], showing how to take advantage of the problem sparsity by decomposing (2.4) into block diagonal problems depending on ρ: this considerably reduces the computational burden in practice. Implementations of the graphical-Lasso algorithm are available in the R-packages glasso, huge [196], or simone [SW5, JP11]. The most recent notable

efforts related to the optimization of (2.4) are due to [78, 79] and the QUIC (then BIG&QUIC) algorithm, a quadratic approximation which allows (2.4) to be solved up to $p = 1,000,000$ with a super-linear rate of convergence and with bounded memory. The R-package quic implements the first version of this algorithm.

On the statistical side, the most striking results are due to [142]: they show that selection consistency of the estimator defined by (2.4) – that is, recovery of the true underlying graphical structure –, is met in the sub-Gaussian case when, for an appropriate choice of λ, the sample size n is of the same order as $O(d^2 \log(p))$, where d is the highest degree in the target graph. Additional conditions on the empirical covariance between relevant and irrelevant features are required, known as the "irrepresentability conditions" in the Lasso case. Such statistical results are important since they provide insights on the "data" situations where such methods may either be successful or completely hopeless. More on this is discussed in [172]. For instance, this should prevent blindly applying the graphical-Lasso in situations where the sample size n is too small compared to p. Similarly, when the presence of hub nodes with high degree is suspected, the estimated graph should be interpreted with care.

Neighborhood selection. This approach, proposed in [122], determines the graph of conditional dependencies by solving a series of p independent Lasso problems, successively estimating the neighborhoods of each variable and then applying a final reconciliation step as post-treatment to recover a symmetric adjacency matrix. Concretely, a given column X_j of the data matrix is "explained" by the remaining columns $X_{\setminus j}$ corresponding to the remaining variables: the set $\mathrm{ne}(j)$ of neighbors of variable j in the graph G is estimated by the support of the vector solving

$$\hat{\beta}_j = \arg\min_{\beta \in \mathbb{R}^{p-1}} \frac{1}{2n} \| X_j - X_{\setminus j} \beta \|_2^2 + \lambda \| \beta \|_1. \tag{2.5}$$

Indeed, if each row of X is drawn from a multivariate Gaussian $\mathcal{N}(0, \Sigma^{-1})$, then the best linear approximation of X_j by $X_{\setminus j}$ is given by

$$X_j = \sum_{k \in \mathrm{ne}(j)} \beta_{jk} X_k = \sum_{k \in \mathrm{ne}(j)} -\frac{\Sigma_{jk}}{\Sigma_{jj}} X_k, \tag{2.6}$$

thus coefficients β_j and column Σ_j – once its diagonal elements are removed – share the same support. By support, we mean the set of nonzero coefficients. Adjusting (2.5) for each $j = 1, \ldots, p$ allow us to reconstruct the full graph G. Because the neighborhoods of the p variable are selected separately, a post symmetrization must be applied to manage inconsistencies between edge selection: [122] suggests AND or OR rules.

Let us fill the gap with Criterion (2.4). First, note that the p regression problem can be rewritten as a unique matrix problem, where B contains p vectors $\beta_j, j = 1, \ldots, p$:

$$\hat{B}^{\mathrm{ns}} = \arg\min_{B \in \mathbb{R}^{p \times p}, \mathrm{diag}(B) = 0_p} \frac{1}{2} \mathrm{Tr}(B^\top S_n B) - \mathrm{Tr}(B^\top S_n) + \lambda \| B \|_1. \tag{2.7}$$

In fact, it can be shown [145, JP9, 143] that the optimization problem (2.7) corresponds to the minimization of a penalized, negative *pseudo*-likelihood: the joint distribution of X is approximated by the product of the p distributions of the p variables conditional

on the other ones, that is

$$\log P(X; \quad) = \sum_{j=1}^{p} \sum_{i=1}^{n} \log P(X_j^i | X_{\mathrm{n}j}^i; \quad_j).$$

This pseudo-likelihood is based upon the (false) assumption that conditional distributions are independent. Moreover, all variables are assumed to share the same variance in this formulation. Building on these remarks, [145] amend criterion (2.7) by the adjunction of an additional symmetry constraint, and introduce additional parameters to account for different variances between the variables.

Concerning the computational aspect, this approach has very efficient implementation as it basically boils down to solving p Lasso problems. Suppose for instance that the target neighborhood size k is per variable: fitting the whole solution path of a Lasso problem using the Lars algorithm can be done in $O(mpk)$ complexity [5]. This must be multiplied by p for the whole network, yet we underline that a parallel implementation is straightforward in this case. This makes this approach quite competitive, especially when coupled with additional bootstrap or resampling techniques [123].

On the statistical side, neighborhood selection has been reported to be sometimes empirically more accurate in terms of edge detection than is the graphical-Lasso [174, 145] on certain types of data. This is somewhat supported by the statistical analysis of [142], who show that under the classical irrepresentability conditions for the Lasso [195, 122] and for an appropriate choice of , neighborhood selection achieves selection consistency with high probability when the sample size is of the order of $O(d \log(p))$ with d the maximal degree of the target graph. This is to be compared with the $O(d^2 \log(p))$ required by the graphical-Lasso (even if the corresponding "irrepresentability conditions" are not strictly comparable). A rough explanation for this difference on the asymptotic is that the graphical-Lasso intends to estimate the concentration matrix on top of selecting the nonzero entries, while neighborhood selection focuses on the selection problem.

Constrained ℓ_1-minimization for inverse matrix estimation (CLIME). The CLIME estimator has been proposed by [22] and is designed to avoid the cumbersome "irrepresentability conditions" required for the Graphical-Lasso and the neighborhood selection approaches, while providing statistical guarantees on the support recovery.

The definition of CLIME builds on the remark that the solution to (2.4) must verify the following first order optimality condition – or subgradient equations:

$$b^{\text{glasso}} \quad ^1 \quad S_n = \quad , \quad \text{with} \quad _{ij} \quad \begin{cases} = \operatorname{sign}(b_{ij}^{\text{glasso}}) & \text{if } b_{ij}^{\text{glasso}} \neq 0, \\ 2 [\ 1, 1] & \text{otherwise.} \end{cases}$$

This suggests the optimization problem

$$\underset{2S_p^+}{\text{minimize}} \| \|_1, \quad \text{s.c.} \quad \| \quad ^1 \quad S_n \|_1 \quad ,$$

which is too hard to solve. Removing the positive-definite requirement and multiplying the constraint by , we encounter the problem solved by CLIME:

$$b^{\text{clime}} = \underset{2M_{pp}}{\arg\min} \| \|_1, \quad \text{s.c.} \quad \| I_p \quad S_n \|_1 \quad . \qquad (2.8)$$

This estimator is not necessarily symmetric and a post-treatment is required as for neighborhood selection. But it can also be easily distributed for each column of $_j$, which requires the resolution of a linear program of complexity $O(p^2k)$, with k the targeted number of neighbors per variable. This is slightly more demanding than neighborhood-selection but remains extremely competitive.

On the statistical side, the CLIME estimator achieves selection consistency at a rate comparable to that of the Graphical-Lasso [22], and is better in its adaptive (weighted) version. Its great advantage is that no particular assumption like a irrepresentability condition – which can never be established in practice – is required for the data matrix X. This method is distributed via the R-package fastclime, and an implementation [177] is reported to solve for problems with millions of features.

Sparse PArtial Correlation Estimation (SPACE). In [136], the gap is completely filled between linear regression, Gaussian graphical model and neighborhood selection with a method that directly penalizes the partial correlations within the linear model. Indeed, by combining firstly Relationship 2.1 between the partial correlations and the concentration matrix, and secondly, Relationship 2.6 between the coefficients in linear regression and concentration matrix, one has

$$X_j = \sum_{k \in \text{ne}(j)} {}_{jk} X_k + " = \sum_{k \in \text{ne}(j)} {}_{jk} \sqrt{\frac{kk}{jj}} X_k + ",$$

which suggests the following optimization problem

$$b^{\text{space}}, \text{diag}( ) = \underset{2 R^{p(p-1)}, \text{diag}( )}{\arg\min} \frac{1}{2} \sum_{j=1}^{p} ! _j \left\| X_j - \sum_{k=1}^{p} {}_{jk} \sqrt{\frac{kk}{jj}} X_k \right\|_2^2 + \| \|_1, \tag{2.9}$$

where is a vector containing all the pairwise partial correlations, diag contains the diagonal elements of , that is to say, the partial covariances of all the variables, and finally $! _j$ are some positive (given) weights.

Although the optimization of (2.9) is more demanding than is neighborhood selection, the problem is jointly convex in (diag( ),). When diag( ) is fixed, the problem has the same complexity as does neighborhood selection, and the authors claim that only a couple of iterations alternating over each of the two parameters (diag( ),) are needed for convergence. It thus remains a lot more efficient than the graphical-Lasso. On top of that, the method intrinsically imposes symmetry over the partial correlations . In short, it embeds the computational advantage of neighborhood selection while estimating the conditional variance as in the graphical-Lasso. It is available in the R-package space. Further refinements and statistical analyses have been recently proposed in [92].

Model selection issues. Up to this point, we have completely avoided the fundamental model selection issue, that is, the choice of the tuning parameter , which is at play in all the sparse methods mentioned thus far. The first possibility is to rely on information criteria of the form

$$\text{IC} = -2\text{loglik}(b ; X) + \text{pen}(\text{df}(b)),$$

where "pen" is a function penalizing the model complexity, described by df, the degrees of freedom of the current estimator. We meet the AIC by choosing pen( ) = 2

and the BIC by choosing pen$(n) = \log(n)x$. However, AIC and BIC are based upon assumptions which are not suited to high-dimensional settings (see [65]). Moreover, the notion of degrees of freedom for sparse methods has to be specified, not to mention that one has to adapt these criteria to the case of GGMs. An example of a criterion meeting these prerequisites is the extended BIC for sparse GGMs [53]

$$\text{EBIC}_\gamma(\hat{b}) = -2\text{loglik}(\hat{b}; X) + |E|(\log(n) + 4\gamma\log(p)), \qquad (2.10)$$

where the function df is equal to $|E|$, the total number of edges in the inferred graph. The parameter $\gamma \in [0, 1]$ is used to adjust the tendency of the usual BIC – recovered for $\gamma = 0$ – to choose overly dense graphs in the high-dimensional setting. Further justification can be found in [53]. A competing approach, designed to compare a family of GGM – possibly inferred with different methods –, is GGMSelect [64, 62].

Another possibility is to rely on resampling/subsampling procedures to select a set of edges which are robust to small variations of the sample. The most popular approach is the *Stability Selection* procedure proposed in [123], also related to the bootstrapped procedure of [3]. A similar approach, called StaRS (Stability approach to Regularization Selection) is developed specifically in the context of GGM in [109]. The basic idea is as follows: for a given range of the tuning parameter $\Gamma = [\gamma_{\min}, \gamma_{\max}]$, the same method is fitted on many subsamples (with or without replacement) with size, say, $n/2$. The idea is then to construct a score indexed on Γ that measures stability – or instability – of the selected variables. The selected edges are those matching a given score, for which the probability of false discovery is controlled. This requires an additional threshold in place of a choice of γ, but the authors in [123, 109] claim that such a threshold is typically much less sensitive than is the tuning parameter γ. An application of such resampling techniques to the inference of biological networks has been pursued with success in [71], advocating for the use of stability methods on real problems.

A final possibility — that remains somewhat confidential while writing these lines — is to rely on sparse procedures which are less sensitive to γ: among these, we may cite the "scaled-Lasso" [159] for linear regression, adapted to the context of network inference in a neighborhood-selection-like fashion in [160].

Extensions towards non Gaussian settings. As hopefully illustrated throughout this section, sparse GGM is a mature and well controlled framework, with solid contributions both on the statistical and the computational sides. There is also expanding innovative literature tending to broaden the applicability of GGMs, especially to overcome the Gaussian assumption. Indeed, particularly in genomics, there is a growing interest for the multivariate modeling of discrete random vectors, as sequencing techniques provide us with count data. In this perspective, some attempts were made for a Poisson version of the above techniques: in [1] for instance, the neighborhood selection approach is extended to a sparse generalized linear model setup; still, interpretability of the inferred network is questionable, as a null partial correlation does not mean conditional dependency in the non-Gaussian case. In a recent paper [186], a review of existing Poisson graphical models is provided, where the notion of conditional dependency is more carefully specified.

Finally, there is much interest for pretreatment methods which change the original data into more "Gaussian" data via simple transformations. Hence, we can still take advantage of the well-controlled sparse GGM framework. A successful work based on Gaussian copulas is the nonparanormal distribution developed in [108]. It is imple-

mented within the R-package huge, at a negligible cost compared to that of the inference process itself.

2.2 CONTRIBUTIONS

This section proposes an overview of my contributions to the framework of sparse GGM for network inference. Although these extensions apply to a broad class of application fields where sparse covariance estimation is involved, I readily acknowledge that I have mainly drawn inspiration from genomics problems to motivate those extensions. The main guideline for these contributions is to account for some kind of structure or characteristics possessed by genomics data as discussed in Chapter 1.

The first contribution is developed in Section 2.2.1. It corresponds to the journal paper [JP9] written with Catherine Matias and Christophe Ambroise, and implemented in an R-package initially described in [JP11]. It addresses the introduction of a possible special organization of the network itself to drive the reconstruction process. Indeed, while sparsity is necessary to solve the problem when few observations are available, biasing the estimation of the network towards a given topology can help us find the correct graph in a more robust way, by preventing the algorithm from looking for solutions in regions where the correct graph is less likely to reside.

The second contribution (Section 2.2.2) emerged from a collaboration with Yves Grandvalet and Christophe Ambroise [JP7]. It addresses the problem of sample heterogeneity which typically occurs when several assays are performed in different experimental conditions that potentially affect the regulations, but are still merged together to perform network inference as data is very scarce. We remedy heterogeneity among sample experiments by estimating multiple GGMs, each of which matches different modalities of the same set of variables, which correspond here to the different experimental conditions. This idea, coupled with the integration of biological knowledge, was further explored for application in cancer with Marine Jeanmougin and Camille Charbonnier [BC1].

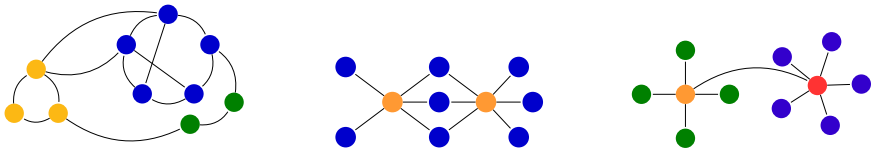
In Section 2.2.3, I describe adaptations of the two preceding GGM extensions to time-course data, relying on a VAR(1) modeling. This corresponds to the first part of Camille Charbonnier's PhD thesis published in the journal paper [JP8].

Finally, a deeper generalization of GGM comes by integrating multiple types of data measured from diverse platforms, what is sometimes referred to as *horizontal* integration: not only does this means a better treatment of the heterogeneity of the data, but it also makes the network reconstruction more robust. An option is proposed in Section 2.2.4 which corresponds to an unpublished work started with Eric Kolazcyk that unfortunately remains unfinished as a similar proposal scooped our own work.

Remark. The choice of not overly detailing neither the technical side of the algorithm nor the statistical properties of the corresponding penalty-based approaches is deliberate, as this chapter is more intended to illustrate how biological data motivated modification of the criteria at hand with sparse GGM. Computational and statistical properties of the original sparse methods used for GGM inference are addressed in more detail and in a broader context in Chapter 3.

2.2.1 Accounting for latent organization of networks

The major originality of the method developed in this work lies in the fact that it searches for a latent modular representation of the GGM to drive the sparse inference of



(a) Community (or affiliation) (b) Star (or hub) (c) Mixed community and star

Figure 2.1 –*Examples of typical network structures.*

its conditional structure. Indeed, modularity and more generally heterogeneity are an important property of gene regulatory networks, see [82]. Typical network structures which can be expected in biological networks are illustrated in Figure 2.1. For instance, the so-called "hubs" in Figure b) are highly connected biological features, showing a different behavior from that of the rest of the graph.

Providing the network with a latent structure. To describe network heterogeneity, we adopt the Stochastic Block Model (SBM) framework which provides mixture models for random graphs. This model has been reinvented many times in the literature and a non exhaustive bibliography should include [55, 157, 39]. An SBM can be stated as follows: vertices of $\mathcal{V}$ are distributed among a set $\mathcal{Q} = \{1, \ldots, Q\}$ of hidden clusters that model the latent structure of the graph $\mathcal{G}$. For any vertex i, the indicator variable Z_{iq} is equal to 1 if $i \in q$ and 0 otherwise, hence describing which cluster the vertex i belongs to. A vertex is assumed to belong to one cluster only, thus the random vector $Z_i = (Z_{i1}, \ldots, Z_{iQ})$ follows a multinomial distribution such that $Z_i \sim \mathcal{M}(1, \alpha)$, where $\alpha = (\alpha_1, \ldots, \alpha_Q)$ is a vector of cluster proportions, such that $\sum_q \alpha_q = 1$. The connectivity matrix $\beta = (\beta_{q\ell})_{q,\ell \in \mathcal{Q}}$, describes $\mathbb{P}(i \sim j | i \in q, j \in \ell)$, that is, how variables from each cluster connect to each other. If working with a valuated network – as will be the case in the following – we shall use a density function in place of the matrix β to define the distribution of the value of the edge according to the class to which the node belongs, that is, a set of density functions $f_q(\cdot)$. Various choices are possible, as depicted in [117].

This framework embraces a large variety of network topologies. In that respect, Figure 2.1 illustrates only a small subset of all possible structures. In the context of gene regulatory networks, the SBM is able to capture functional modules in the spirit of community structures but also other main topological properties of biological networks like star-shaped models isolating transcription factors. The set of parameters (α, β) describing the latent structure Z could either be considered as prior knowledge, motivated by biological expert knowledge or bibliographical references, or inferred directly from the data. We will see how to deal with the second option in the following, but we first consider Z as known, to show how it can be easily integrated into the network reconstruction process with sparse methods.

High-dimensional inference driven by latent network structure. The idea developed in our work [JP9] is to refine the regularization in penalized criteria such as (2.4), (2.7), (2.8) or (2.9) by adding entrywise *structure adaptive* penalty parameters. Consider for instance the penalized likelihood framework (2.4): the different level of the penalty

parameters for each entry $_{ij}$ should be driven by the latent structure Z, leading to

$$\hat{}_{,Z} = \arg\max_{2S_p^+} \log\det( ) \quad \mathrm{Tr}(\hat{S}) \quad kP^Z ? \quad k_{,1}, \tag{2.11}$$

where P^Z is a $p \quad p$ matrix of penalty terms the entries of which depend on Z and ? denotes the term-by-term product. The penalty term decomposes into a common part tuning the overall amount of sparsity of the graph and a new structured part P^Z used to tune the strength of prior information, which will encourage the edge structure to adopt more or less strongly the prior structure Z. Indeed, we wish to penalize the elements of the concentration matrix according to the clusters to which the variables belong. For instance, let us imagine a graph endowed with a community structure as in Figure 2.1 a): if two variables belong to the same community, we wish to lower the penalty acting on the corresponding entry in the concentration matrix. Conversely, we want to increase the penalty on entries corresponding to variables belonging to different communities with low connectivity probability, in order to shrink the estimated partial correlation to zero. When Z and the associated parameters are known, various penalty values can be defined as decreasing functions of the connectivity matrix . Suppose variables i and j are assigned to clusters q and $‘$, then an efficient penalty weight for edge $_{ij}$ is $P^Z_{ij} = 1 \quad _{q‘}$. We explored in [BC1] such an option for analyzing breast cancer data, where prior knowledge based upon existing cancer signatures and pathway analysis have been integrated in P^Z to drive the network reconstruction.

However, a fully integrated statistical model is desirable to recover both Z and simultaneously: an option that we explored in [JPS1] is to rely on an EM-like strategy. The main lines of this approach are depicted in the following.

A global EM-strategy for sparse GGM with latent structure. Our idea is to reach an EM strategy such that the E-step corresponds to the inference of the latent structure Z, while the M-step corresponds to the resolution of (2.11), that is to say, inference of the network G.

To this end, we must put a probabilistic model on in order to write the complete likelihood of the model. We thus extend the clustering of vertices to the concentration matrix . Accordingly, both the existence and the weight of the edges, described by the off-diagonal elements of , will depend on the cluster to which each vertex belongs. Conditional on the events i 2 q and j 2 $‘$ where $q, ‘$ are clusters chosen from Q, we provide each $_{ij}$ with a prior Laplace distribution denoted for all i 6 j by $f_{q‘}$ and identified by its scaling parameter $_{q‘} > 0$, that is[1]

$$_{ij} \mid Z_{iq}Z_{j‘} = 1 \quad f_{q‘}(; _{q‘}) \quad \text{with} \quad f_{q‘}(; _{q‘}) = \frac{1}{2 _{q‘}} \exp \quad \frac{| |}{_{q‘}}.$$

It will be noticed that in this formulation the variables $_{ij}$ are assumed to be independent, conditional on the clusters to which the vertices belong to. Moreover, we are considering only undirected graphs, so we may assume that $= _{‘q}$.

The reason for choosing a Laplace distribution is that $_1$ penalties may be interpreted as Laplace priors on the parameters: this interpretation enables us to embed the sparse $_1$-procedure into an EM-algorithm. Indeed, under the modeling specified, we have the following result:

[1] For technical reasons, we also assume a distribution on diagonal elements, namely $_{ii}$ $f_0(; _0)$, where the parameter $_0$ is fixed and not estimated.

Proposition. *The complete likelihood can be written as*

$$\log P(X, \alpha, Z) = \frac{n}{2}(\log\det(\Theta) - \mathrm{Tr}(S_n\Theta)) - P^Z \cdot \Big[\sum_{\substack{i,j\in P,\, i\neq j \\ q,\ell\in Q}} Z_{iq}Z_{j\ell}\log(2\alpha_{q\ell}) + \sum_{i\in P,\, q\in Q} Z_{iq}\log\alpha_q \Big] + c, \quad (2.12)$$

where c is a constant term and P^Z is defined by

$$P^Z_{ij} = \begin{cases} 1-\alpha_{q\ell}, & \text{if } i\neq j, i\in q, j\in\ell \\ 1-\alpha_0, & \text{otherwise.} \end{cases} \quad (2.13)$$

In the classical EM framework developed by [41] inferring the parameters spread over a latent structure Z would make use of the conditional expectation:

$$Q(\cdot \mid \cdot^{(m)}) = E\big[\log P(X, \alpha, Z) \mid X; \cdot^{(m)}\big] = \sum_{Z\in\mathcal{Z}} P(Z \mid \cdot^{(m)}) \log P(X, \alpha, Z), \quad (2.14)$$

where $\cdot^{(m)}$ is the estimation of $\cdot$ from the previous step of the algorithm.

E-step. The usual EM strategy would be to alternate an E-step computing the conditional expectation (2.14) with an M-step maximizing this quantity over the parameter of interest $\cdot$. Unfortunately, no closed form of $Q(\cdot \mid \cdot^{(m)})$ can be formulated in the present case. The technical difficulty lies in the complex dependency structure contained in the model. Indeed, $P(Z \mid \cdot)$ cannot be factorized, as argued in [39]. This makes the direct calculation of $Q(\cdot \mid \cdot^{(m)})$ impossible. To tackle this problem we use a variational approach [83], which has been further investigated for SBM in [117]. In this framework, the conditional distribution of the latent variables $P(Z \mid \cdot^{(m)})$ is approximated by a more convenient distribution denoted by $R_m(Z)$, which is chosen carefully in order to be tractable. Hence, our EM-like algorithm deals with the following approximation of the conditional expectation (2.14)

$$E_{R_m}\{\log P(X, \alpha, Z)\} = \sum_{Z\in\mathcal{Z}} R_m(Z)\log P(X, \alpha, Z). \quad (2.15)$$

Though not detailed here (see more in our original paper [99]), this variational approach enables us to compute the conditional expectation (2.15) by providing an estimation $\hat{Q}(\cdot)$, where $\tau_{iq} = \hat{P}(i \in q)$, defined for all $i \in P, q \in Q$, are the variational parameters estimating the latent structure Z. Note that we can derive analytic expressions for estimating the parameters α and $\alpha_{q\ell}$, though details are omitted here.

M-step. Now, we wish to infer the concentration matrix Θ, assuming α is known. This is the aim of the M-step of our EM–like strategy, that deals with the maximization problem $\arg\max_{\Theta\succ 0} \hat{Q}(\Theta)$. Conditional on the estimated structure τ, we can compute the maximum *a posteriori* estimate of Θ defined as follows

$$\hat{\Theta} = \arg\max_{\Theta\succ 0} \log P(\Theta \mid X, \tau) = \arg\max_{\Theta\succ 0} \log P(\Theta, X \mid \tau). \quad (2.16)$$

Using Expression 2.12 of the complete likelihood and the equality $E_R(Z_{iq}Z_{j\ell}) = \tau_{iq}\tau_{j\ell}$, it is a simple matter to rewrite the problem as the ℓ_1-penalized optimization problem

$$\hat{\Theta} = \arg\max_{\Theta\succ 0} \big[\log\det(\Theta) - \mathrm{Tr}(S\Theta) - \|P^Z \cdot \Theta\|_1\big]. \quad (2.17)$$

There are various algorithms to optimize (2.17). In [JP9, JP11], we rely on the approach originally developed in [57] for the Graphical-Lasso, briefly depicted in Section 2.1.2. The main bottleneck here is to efficiently solve a weighted-Lasso Problem. We used a modified coordinate descent approach [55, 56] to optimize the latter. More details on such computational tools are provided in paper [JP9] and in Chapter 3, Section 3.1.2.

Finally, note that we start our EM algorithm by choosing initial values of and using a classification algorithm such as spectral clustering [136]. The number of classes is chosen using the ICL (integrated complete likelihood) criterion [41] and remains fixed throughout the EM steps. The q is a user parameter, which may typically be cross-validated.

Illustrative example on a breast cancer data set. We tested our procedure on a gene expression data set provided by [73] and concerning 133 patients with stage III breast cancer. The patients were treated with chemotherapy prior to surgery. Patient response to the treatment is classified as either a pathologic complete response (pCR) or a residual disease (not-pCR). A multigene predictor for treatment response is developed in [126] on this data set, consisting in a set of 26 genes having a high predictive value. We show on Figure 2.2 the network reconstructed with our method, exhibiting a modular structure in 3 classes, and a number of edges of the same order as 26 after cross-validation of .

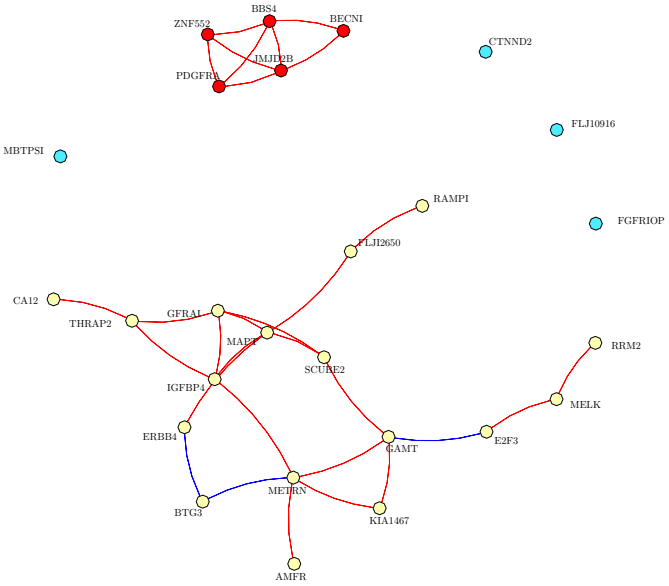


Figure 2.2 – *Breast cancer data set of* [*73*]*: inferred graphs on the signature proposed by* [*126*]*. Red (respectively blue) edges correspond to positive (respective negative) partial correlations.*

Some final comments on structured GGM inference. This method has been implemented in the R-package , of which I am the principal developer. In the full length paper [JP9], a simulation study characterizes the situations where our proposal outperforms its competitors. We also propose an alternative choice for the to bound the probability of misclassifying a couple of nodes from class . Following our work, a Bayesian algorithm [119] was proposed for implementing such a model; a paper very close to our work but that does not seem to be aware of it [186]. Recent works sharing similar ideas on an underlying network organization are [25, 163, 65].

2.2.2 Accounting for sample heterogeneity

In order to deal with the data scarcity in genomics, it is a common practice in GGM-based inference methods to merge different experimental conditions from wetlab data as in [149, 169]. This process increases the number of observations available for inferring interactions. However, GGMs assume that the observed data form an independent and identically distributed sample. In the aforementioned paradigm, assuming that the merged data are drawn from a single Gaussian component is obviously wrong and is likely to have detrimental effects on the estimation process. In the journal paper [97], I present with my co-authors a series of sparse inference methods that propose to remedy this problem by estimating multiple GGMs.

GGM inference in a multi-task framework. From a statistical viewpoint, we have n observations belonging to C different sub-populations (or "tasks"), hence with different distributions. Assuming that each sample was drawn independently from a Gaussian distribution, we set

$$X^{(c)} \quad \mathcal{N} \quad (0_p, \quad ^{(c)}),$$

where the C samples may be processed separately, following any approach described in Section 2.1.2: denote by L ($^{(c)}; S_n^{(c)}$) the data-fitting term in condition c with the corresponding concentration matrix and empirical covariance matrix; can be the multivariate Gaussian log-likelihood as in (2.4), the pseudo-likelihood as in (2.7) or the losses arising in CLIME (2.8) or SPACE (2.9). In the case of the graphical-Lasso, optimizing the C problems separately can be expressed as the unique optimization problem

$$\arg\max_{\{ \ ^{(c)}_{ij} : i \in j \}_{c=1}^{C}} \sum_{c=1}^{C} L \ (\ ^{(c)}; S_n^{(c)}) \quad k \ ^{(c)} k_1 \ . \tag{2.18}$$

Note that it is sensible to apply the same penalty parameter for all samples, provided that the C samples have similar sizes and originate from similar distributions, in particular regarding scaling and sparseness.

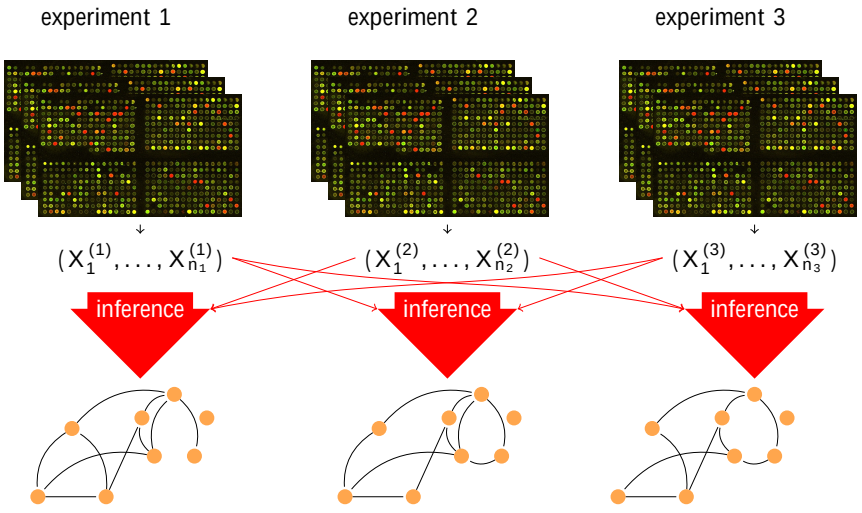


Figure 2.3 – *Multi-task learning framework*

Problem (2.18) ignores the relationships between regulation networks. When sub-population networks are assumed to share a large common core of edges and only differ by a small subset of edges, the multi-task learning framework presented in Figure 2.3 is well adapted, especially for small sample sizes. First, sharing information may considerably improve estimation accuracy. Second, keeping the opportunity to identify differences between the networks is the key to understanding the regulatory system up to its sub-population variations. Starting from problem (2.18), coupling the estimation of $\Theta^{(1)}, \ldots, \Theta^{(C)}$ may be achieved by either modifying the data-fitting term or the penalizer. These two options result respectively in the graphical intertwined-Lasso and the graphical cooperative-Lasso presented below.

Intertwined inference. In the maximum *a posteriori* framework, the estimation of a concentration matrix can be biased towards a specific value $(S_n^0)^{-1}$. From a practical viewpoint, this is usually done by considering a conjugate prior on, that is, a Wishart distribution $\mathcal{W}((S_n^0)^{-1}, n)$. The MAP estimate is then computed as if we had observed additional observations of empirical covariance matrix S_n^0.

Here, we would like to bias each estimation problem towards the same concentration matrix, the value of which is unknown. An empirical Bayes solution would be to set $S_n^0 = \bar{S}_n$, where $\bar{S}_n$ is the weighted average of the C empirical covariance matrices. As in the maximum likelihood framework, this approach would lead to a full concentration matrix. Hence, we consider here a penalized criterion, which does not exactly fit the penalized maximum likelihood nor the MAP frameworks, but which will perform the desired coupling between the estimates of $\Theta^{(1)}, \ldots, \Theta^{(C)}$ while pursuing the original sparseness goal.

Formally, let $n_1, \ldots, n_C$ be the sizes of the respective samples, the empirical covariance matrices of which are denoted by $S_n^{(1)}, \ldots, S_n^{(C)}$. Also denoting $n = \sum n_c$, we consider the following problem:

$$\max_{\{\Theta^{(c)}: i \neq j\} \succeq 0} \sum_{c=1}^{C} L(\Theta^{(c)}; \tilde{S}_n^{(c)}) - \lambda \|\Theta^{(c)}\|_1, \qquad (2.19)$$

where $\tilde{S}_n^{(c)} = \alpha S_n^{(c)} + (1-\alpha)\bar{S}_n$ and $\bar{S}_n = n^{-1} \sum_{c=1}^{C} n_t S_n^{(c)}$. As this criterion amounts to considering that we observed a blend of the actual data for task c and data from the other tasks, we will refer to this approach as intertwined estimation.

The idea is reminiscent of the compromise between linear discriminant analysis and its quadratic counterpart performed by the regularized discriminant analysis [58]. Although the tools are similar, the primary goals differ: [58] aims at getting control on the number of effective parameters, while we want to bias empirical distributions towards a common model. The additional tuning parameter α can typically be chosen by cross-validation.

Cooperative inference. The second approach consists in devising penalties that encourage similar sparsity patterns across tasks.

This kind of setting has received much attention in the statistics and machine learning communities, mainly on variants and applications of the Group-Lasso proposed in [190], which has already inspired some multi-task learning strategies as in [121, 2, 4, 112], but had never been considered for learning graph models when we proposed this work. We briefly describe how group-Lasso may be used for inferring multiple graphs before introducing a slightly more complex penalty, the *cooperative*-Lasso,

which was inspired by the application to biological interactions, but which should be relevant in many other applications.

As in the single task case, sparsity of the concentration matrices is obtained in ℓ_1-norm. An additional constraint imposes the similarity between the two concentration matrices. Each interaction is considered as a group.

The group-Lasso penalty is based upon a mixed norm (see Section 3.1.1) which encourages sparse solutions with respect to groups, where groups form a pre-defined partition of variables. The partition acts at the edge level, by grouping each partial correlation coefficient across the C conditions. It is therefore useful to define vectors $\{\rho_{ij}^{(c)}\}_{i \neq j} \in \mathbb{R}^C$ containing all partial correlations between genes i and j across the C conditions. Such a penalty will favor graphs $G^{(1)}, \ldots, G^{(C)}$ with common regulations, not necessarily with the same strength but present or absent together across the conditions. The graphical group-Lasso learning problem designed to infer multiple GGMs is then

$$\underset{\{\rho_{ij}^{(c)}: i \neq j\}_{c=1}^{C}}{\arg\max} \sum_{c=1}^{C} \mathcal{L}(\Theta^{(c)}; S_n^{(c)}) - \sum_{i \neq j} \|\rho_{ij}\|_2. \tag{2.20}$$

Although this formalization expresses some of our expectations regarding the commonalities between tasks, it is not really satisfying here since we aim at inferring the support of the solution (that is, the set of non-zero entries of $\Theta^{(c)}$). To enable the inference of different networks (c, c'), we must have some (i, j) such that $\rho_{ij}^{(c)} = 0$ and $\rho_{ij}^{(c')} \neq 0$. This event occurs with probability zero with the group-Lasso, where variables enter or leave the support group-wise. However, we may cure this problem by considering a regularization term that better suits our needs. Namely, when the graphs represent the regulation networks of the same set of molecules across experimental conditions, we expect a stronger similarity pattern than the one expressed in (2.20). Specifically, the co-regulation encompasses up-regulation and down-regulation and the type of regulation is not likely to be inverted across assays: in terms of partial correlations, sign swaps are very unlikely. This additional constraint is formalized in the following cooperative-Lasso learning problem (2.21):

$$\underset{\{\rho_{ij}^{(c)}: i \neq j\}_{c=1}^{C}}{\arg\max} \sum_{c=1}^{C} \mathcal{L}(\Theta^{(c)}; S_n^{(c)}) - \sum_{i \neq j} \|\rho_{ij}^{+}\|_2 + \|\rho_{ij}^{-}\|_2, \tag{2.21}$$

where $u^+ = \max(0, u)$ and $u^- = \max(0, -u)$.

Figure 2.4 illustrates the construction of the two grouped penalties. The group-Lasso switches on or off all edges between variables i and j across all conditions while the cooperative-Lasso disconnects the activations of up- and down-regulations. In this way, the cooperative-Lasso allows for instance the activation of an up-regulation in a subset of conditions while this regulation disappears in the remaining conditions. More insights come on the cooperative-Lasso in Section 3.2.1 of Chapter 3, where we present our work [JPV] about its statistical properties in the linear regression framework as well as an efficient optimization strategy.

Illustrative example on a breast cancer data set. We now revisit the breast cancer data set of [73, 126] analyzed previously in Section 2.2.1. This time, we would like to account for heterogeneity of the samples by splitting the $n = 133$ patients into the two subsamples "pCR" and "not-pCR" with respective sizes 34 and 99, to produce

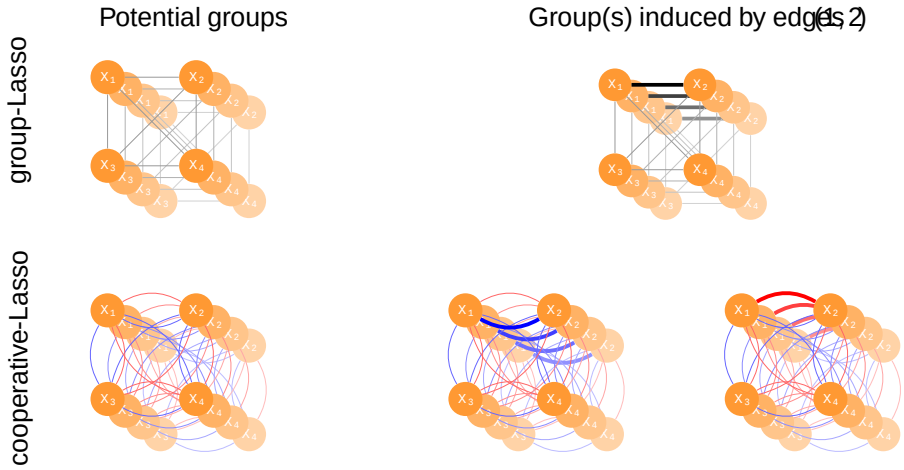


Figure 2.4 –*Grouping edges without or with sign-effect for* $C = 4$ *conditions*

two graphs – one per each type of patient. We learn the two networks jointly with our cooperative-Lasso procedure. We couple this multiple network learning approach with the procedure presented in Section 2.2.1 that accounts for the network heterogeneity, where the latent structure that drives the inference is estimated on the intersection of the two networks. For comparison purposes, we fix the penalty level so that we obtain the same number of edges as in Figure 2.2, where the samples are all merged together. Results of the inference are in Figure 2.5, where we highlight the differences between the networks by pointing out three edges which may distinguish one condition from another.

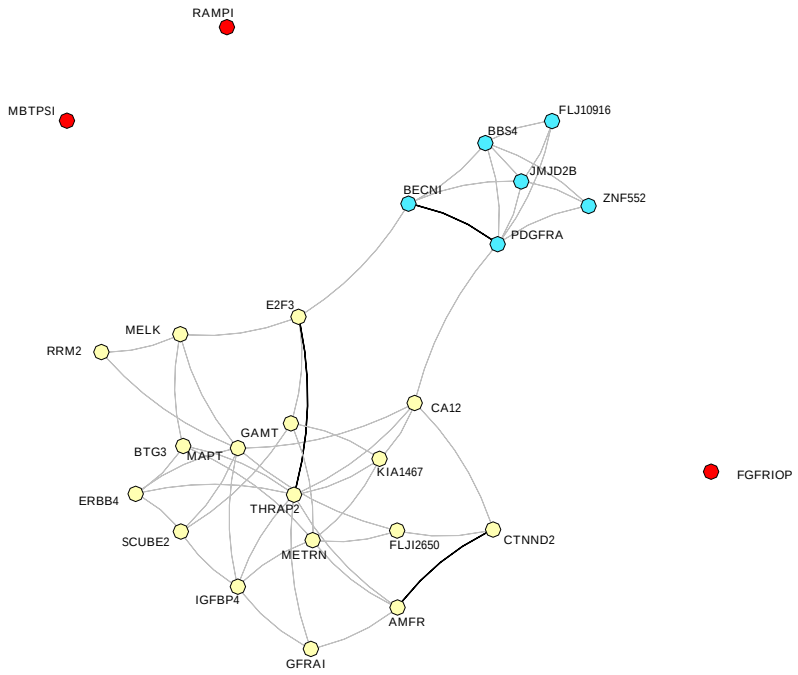


Figure 2.5 –*Breast cancer data set of* [73]*: multiple networks inferred on the signature proposed by* [126] *by splitting the samples into pCR/ not-pCR: gray edges are common to the two conditions, while black edges are specific to no-pCR case.*

Some final comments on multiple GGM inference. We omit in this part the optimization of problems (2.19),(2.20) and (2.21). More details are given in Chapter 3, Sections 3.1.2 and 3.2.1. Basically, they decouple such that the major technical part is to solve a non-smooth convex regularized problem. The original paper [JP7] also includes an application to the inference of the T-cell signaling pathway, and a numerical study comparing the intertwined, the coop- and the group-Lasso for multiple network inference with the baselines that either merge all the data together or treat each condition independently. This shows the range of applicability of the proposals.

Also note that the small numerical example depicted above is purely illustrative: in a book chapter [BC1], I develop with co-authors a complete application to a large cohort of patients with breast cancer. There, we infer jointly two networks corresponding to the patients' ER status ("ER+" or "ER-"). Indeed, ER is a hormone receptor the activity of which is highly correlated to the efficiency of chemotherapy: we identified interesting differences in the network characterizing each condition. In the continuity of this work, a chapter of Camille Charbonnier's PhD thesis is dedicated to the design of test procedures for comparing two networks, in order to decipher whether the edges identified by our procedure as different between tasks are statistically significant: more is developed in a paper in revision at this stage, written by Camille in collaboration with N. Verzelen and F. Villers [26]. Finally, note that all the multiple GGM inference methods described in this section were integrated by myself into the R-package simone.

In the literature, related works [68, 37] followed and proposed approaches close to ours. A recent paper [173] builds on our multi-task framework in order to construct a consensus network between multiple conditions.

2.2.3 Accounting for time-course data

In this section, we briefly show how the statistical model underlying the preceding sparse GGM methods can be simply amended when observations are gathered over time, which can be seen as another source of heterogeneity or structure in the data.

The problem of network inference from data gathered over time is typically motivated by applications in genomics, where it is common to perform time course experiments for expression data. In this case, most learning strategies rely on first-order vector auto-regressive (VAR1) models [137], the inference of which should be amended to deal with the high-dimensional setting: [132] depicts a shrinkage estimate while in [102] statistical tests on limited-order partial correlations are performed to select significant edges. In [154], the VAR(1) setup is dealt with by combining ideas from two major developments of the Lasso, thus defining the Recursive elastic-net. During Camille Charbonnier's Master Thesis, we extended our work on sparse GGM with latent structure (Section 2.2.1) to the VAR(1) setup. The multi-task framework of Section 2.2.2 also straightforwardly follows. An application with this setup was published in a journal paper [JP8]. It also gave us the opportunity for a collaboration with a medical lab on Parkinson disease [CI6]. We found it interesting to quickly review this VAR(1) setting as the conditional graph has a different status in this case, and interpretation of the network inferred differs as well.

Graph of conditional dependencies in the VAR(1) modeling. Suppose that the dynamics of $(X^0, \ldots, X^T)$ at regular time points are represented by a first order vector autoregressive model VAR(1) as in equation (2.22). Each measurement $X^t \in \mathbb{R}^p$ is a

vector containing the observations of the p variables at time t.

$$X^t = X^{t-1}A + \varepsilon^t, \quad \text{for all } t \geq 1, \tag{2.22}$$

where $A = (A_{ij})_{i,j \in P}$ is a $p \times p$ matrix governing the dynamics of the observation over time. Variations from these dynamics are captured by the white Gaussian process $\{\varepsilon^t\}_{t=1,\ldots,T}$, namely, $\varepsilon^t \sim \mathcal{N}(0, D)$ where D is a diagonal matrix such that $D_{ii} = \sigma_i^2$ and $\mathrm{Cov}(\varepsilon^t, \varepsilon^s) = 1_{\{s=t\}} D$ for all $s, t > 0$. Moreover, $X^0 \sim \mathcal{N}(0, \Sigma_0)$. Also assume that $\mathrm{Cov}(X^t, \varepsilon^s) = 0$ for all $s > t$: hence, X^t is obviously a first-order Markov process homogeneous in time, which means that the regulatory structure is assumed constant over time.

In this setting, matrix A plays the role of the concentration matrix Ω in the i.i.d. framework presented in the previous sections. Indeed, we have $X^t | X^{t-1} \sim \mathcal{N}(X^{t-1}A, D)$ and each entry A_{ij} is proportional to the partial correlation coefficient between variables X_i^t and X_j^{t-1}, that is to say between the observation of variable i at time t and the observation of j at the previous time point, with respect to all other variables at time $t-1$, as expressed in

$$A_{ij} = \frac{\mathrm{Cov}\left(X_i^t, X_j^{t-1} \middle| X_{\setminus j}^{t-1}\right)}{\mathrm{Var}\left(X_j^{t-1} \middle| X_{\setminus j}^{t-1}\right)}.$$

Compared to the i.i.d. setting, nonzero entries of A code for a directed graph describing the conditional dependencies between the elements of P. An edge from i to j is added to the graph if, conditional on all variables except variable j at time $t-1$, the covariance between X_i^t and X_j^{t-1} is nonzero. Inferring A is again equivalent to reconstructing the graph of conditional dependencies. However, there are two main differences between this dynamic version of partial correlation and the notion of partial correlation expressed in the previous section. First, the conditioning is made upon all observations from the previous time-point, therefore self-loops are allowed. Second, the correlation considered between two variables is asymmetric: we consider the correlation between the past of i and the present of j, leading naturally to an asymmetric matrix of partial correlations and a directed graph of conditional dependencies.

The Penalized Likelihood. Similarly to (2.11) in the i.i.d. settings, our aim is to induce a structure adaptive penalty for A. For a fixed structure Z, if $S = X_{0:(n-1)}^{\top} X_{1:n} / n$ denotes the across time empirical covariance matrix, the the estimate $\hat{A}^{Z,\lambda}$ in a weighted penalized likelihood framework is the solution of

$$\hat{A}^{Z,\lambda} = \arg\max_{A \in \mathbb{R}^{p \times p}} \mathrm{Tr}(AV^{\top}) - \frac{1}{2}\mathrm{Tr}(A^{\top} S_n A) - \lambda \| P^Z \star A \|_1. \tag{2.23}$$

The structure adaptive penalty matrix P^Z can be adjusted as in the i.i.d. setting either by statistical inference of the latent structure or prior knowledge. Since the network is now directed, this structure needs to take the direction of edges into account. Note that the generalization of (2.23) to the multi-task criteria like (2.19),(2.20) and (2.21) is straightforward. All of them have been implemented in the `simone` R-package.

Some comments on the VAR(1) modeling. In this model, regulations are assumed to be constant over time. Therefore, it is suited to drawing a picture of short-term regulation dynamics based upon measurements taken at close time points and over a short period of time. Models taking into account possible evolutions of the network over time and better suited for life cycle data sets were for instance developed [108] with a Bayesian viewpoint, or with a fused-Lasso like penalty [97].

2.2.4 Accounting for multiscale data: multi-attribute GGM

We now place ourselves in the situation where, for our collection of features, we observe not one but several attributes. The question at hand remains the same, that is to say, unraveling strong interactions between these features according to the observation of their attributes. Such networks are known as "association networks", which are systems of interacting elements, where a link between two different elements indicates a sufficient level of similarity between element attributes. In this section, we are interested in reconstructing such networks based upon n observations of a set of K attributes of the p elements composing the vertices of the network. To this end, we propose a natural generalization of sparse GGM to sparse *multi-attribute* GGMs.

Remark. This work was planned for submission as a research paper with Christophe Ambroise and Eric Kolazyck when we came across an independent work [96] on the arXiv that proposes nearly the same approach. We somewhat gave up this project in its original form, which I choose to include in this manuscript as it brings interesting and renewed questions on GGMs.

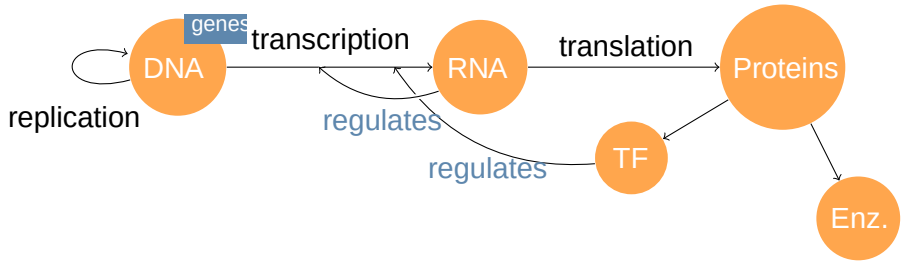


Figure 2.6 –*Basic example of a multi-attribute network in genomics: activity of a gene can be measured at the transcriptomic and proteomic levels, and gene regulation affected accordingly*

Why multi-attribute networks? The need for multi-attribute networks is relevant in many application fields, but seems particularly applicable in genomics. Indeed, with the plurality of emerging technologies and sequencing techniques, it is possible to record many signals related to the same set of biological features at various scales or locations of the cell. Consider for instance the simplifying – still hopefully didactic – central dogma of molecular biology, sketched in Figure 2.6: basically, expression of a gene encoding for a protein can be measured either at the transcriptome level, in terms of its quantity of mRNA, or at the protein level, in terms of the concentration of the associated protein. Still, different technologies are used to measure either the transcriptome or the proteome, typically, microarray or sequencing technology for gene expression levels and cytometric or spectrometric experiments for protein concentrations. Although these signals are very heterogeneous (different levels of noise, count vs. continuous data, etc.), they do share commonality as they undergo common

biological processes. We then put an edge in the network if it is supported in both spaces (gene and protein spaces). Our hope is that molecular profiles combined on the same set of biological samples can be *synergistic*, in order to identify a "consensus" and hopefully more robust network.

Multi-attribute GGM. Let $P = \{1,\dots,p\}$ be a set of variables of interest, each of them having some K attributes. Consider the random vector $X = (X_1,\dots,X_p)^\ddot{u}$ such as $X_i = (X_{i1},\dots,X_{iK})^\ddot{u} \in \mathbb{R}^K$ for $i \in P$. The vector $X \in \mathbb{R}^{pK}$ describes the K recorded signals for the p features. We assume that X is a multivariate centered Gaussian vector, that is, $X \sim N(0, \Sigma)$, with covariance and concentration matrices defined block-wise

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1p} \\ \vdots & \ddots & \vdots \\ \Sigma_{p1} & \cdots & \Sigma_{pp} \end{bmatrix}, \quad \Omega = \begin{bmatrix} \Omega_{11} & \cdots & \Omega_{1p} \\ \vdots & \ddots & \vdots \\ \Omega_{p1} & \cdots & \Omega_{pp} \end{bmatrix}, \quad \Sigma_{ij}, \Omega_{ij} \in \mathcal{M}_{K,K}, \forall (i,j) \in P^2,$$

where $\mathcal{M}_{a,b}$ is the set of real-valued matrices with a rows, b columns. Such a multi-attribute framework has been studied in [91] with a reconstruction method based upon canonical correlations in order to test dependencies between pairs (i,j) at the attribute level using covariance. Here, we propose to rely on partial correlations in a multivariate framework rather than (canonical) correlations to describe relationships between the features, and thus extend GGM to a multi-attribute framework. The objective is to define a "canonical" version of partial correlations. In our setting, the target network $G = (P, E)$ is defined as the multivariate analog of the conditional graph for univariate GGM, that is

$$(i,j) \notin E, \quad \Omega_{ij} = 0_{KK}, \quad \forall i \neq j. \tag{2.24}$$

In words, there is no edge between two variables i and j when their attributes are all conditionally independent.

A multivariate version of neighborhood selection. Our idea for performing sparse multi-attribute GGM inference is to define a multivariate analog of the neighborhood selection approach [122] (see Section 2.1.2, Equations (2.5) and (2.7)). Indeed, it seems to be the most natural and convenient setup toward multivariate generalization. Nevertheless, we think that the graphical-Lasso (2.4), CLIME (2.8) or SPACE (2.9) settings may have a close equivalent multi-attribute version.

To this end, we look at the multivariate analog of equation (2.6): in a multivariate linear regression setup, it is a matter of straightforward algebra to see that the conditional distribution of $X_j \in \mathbb{R}^K$ on the other variables is

$$X_j \,|\, X_{\setminus j} = x \sim N\left(\Omega_{jj}^{-1} \Omega_{j\setminus j} x, \Omega_{jj}^{-1} \right).$$

Equivalently, letting $B_j^T = \Omega_{jj}^{-1} \Omega_{j\setminus j}$, one has

$$X_j \,|\, X_{\setminus j} = B_j^T X_{\setminus j} + \varepsilon_j, \quad \varepsilon_j \sim N(0, \Omega_{jj}^{-1}), \quad \varepsilon_j \perp X,$$

where $B_j \in \mathcal{M}_{(p-1)K,K}$ is defined block-wise

$$B_j = \begin{bmatrix} B_j^{(1)} \\ \hline \vdots \\ \hline B_j^{(p-1)} \end{bmatrix} = \Omega_{jj}^{-1} \begin{bmatrix} \Omega_j^{(1)} \\ \hline \vdots \\ \hline \Omega_j^{(p-1)} \end{bmatrix},$$

and where each $B_j^{(i)}$ is a $K \times K$ matrix which links attributes of variables (i, j). We see that recovering the support of B_j block-wise is equivalent to reconstructing the network defined in (2.24). Estimation of B_j is thus typically achieved through sparse methods. To this end, we consider an i.i.d. sample $\{X^n\}_{n=1}^N$ of X such that each attribute is observed n times for the p variable, each X^n being a pK-size row vector staked in a $\mathbb{M}_{n,pK}$ data matrix $\mathbf{X}$, so that $\mathbf{X}_j \in \mathbb{M}_{n,K}$ is a real-valued $n \times K$ block matrix containing the data related to the jth variable:

$$\mathbf{X} = \begin{bmatrix} x^1 \\ \vdots \\ x^N \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \dots & \mathbf{X}^p \end{bmatrix} = \left[\begin{array}{cc|c|cc} X_1^{11} & X_1^{1K} & \dots & X_1^{p1} & X_1^{pK} \\ \vdots & \vdots & \dots & & \\ X_n^{11} & X_n^{1K} & \dots & X_n^{1K} & X_n^{pK} \end{array}\right].$$

Using these notations, a direct generalization of the neighborhood selection is to predict for each $j = 1, \dots, p$ the data block $\mathbf{X}_j$ by regressing on $\mathbf{X}_{\backslash j}$. In matrix form, this can be written as the optimization problem

$$\arg\min_{\mathbf{B}_j \in \mathbb{R}^{pK \times K}} J(\mathbf{B}_j), \qquad J(\mathbf{B}_j) = \frac{1}{2n} \left\| \mathbf{X}_j - \mathbf{X}_{\backslash j} \mathbf{B}_j \right\|_F^2 + \Omega(\mathbf{B}_j), \tag{2.25}$$

where $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$ is the Frobenius norm of matrix A and Ω is a penalty which constrains $\mathbf{B}_j$ block-wise.

Choosing a penalizer. Various choices for Ω in (2.25) seem relevant: by simply setting $\Omega_0(A) = \sum_{i,j} |A_{i,j}|$, we just encourage sparsity among the $\mathbf{B}_i$ and thus do not couple the attributes. A clever choice would be to activate a set of attributes all together: hence, the group is defined by all the K attributes between variables i and j, therefore the penalizer turns to a group-Lasso like penalty

$$\Omega_1(\mathbf{B}_j) = \sum_{i \in \mathcal{P}_{\backslash j}} \|\mathbf{B}_j^{(i)}\|_F, \tag{2.26}$$

in which case convex analysis and subdifferential calculus [Section 1.8] can be used to show that a $\mathbf{B}_i$ is optimal for Problem (2.25) if and only if

$$\begin{cases} \forall i : \mathbf{B}_j^{(i)} \neq 0, & S_{ij} + \frac{1}{\|\mathbf{B}_j^{(i)}\|_F} I \, S_{ij} = \mathbf{B}_j^{(i)}, \\ \forall i : \mathbf{B}_j^{(i)} = 0_{KK}, & \|S_{ij}\|_F \end{cases} \tag{2.27}$$

where $S_{ij} \in \mathbb{M}_{KK}$ is a $K \times K$ block in the empirical covariance matrix $S_n = n^{-1} \mathbf{X}^\top \mathbf{X}$, which shows the same block-wise decomposition as for Ω_1. This paves the way for an optimization algorithm like block-coordinate descent which we implemented, although we omit details here.

At the time we were working on this model, another idea that we had in mind — although we did not push too far — was to propose a penalty based upon the nuclear norm $\|A\|_\star = \sum_j \sigma_j$, where $(\sigma_1, \dots, \sigma_p)$ is the vector of singular values of A. This somewhat penalizes the rank of a matrix, which would be desirable for matrix $\mathbf{B}_{ij}$ when

many attributes are shared between (i, j). We thus might define a penalty on $$ in place of B_j, with something like

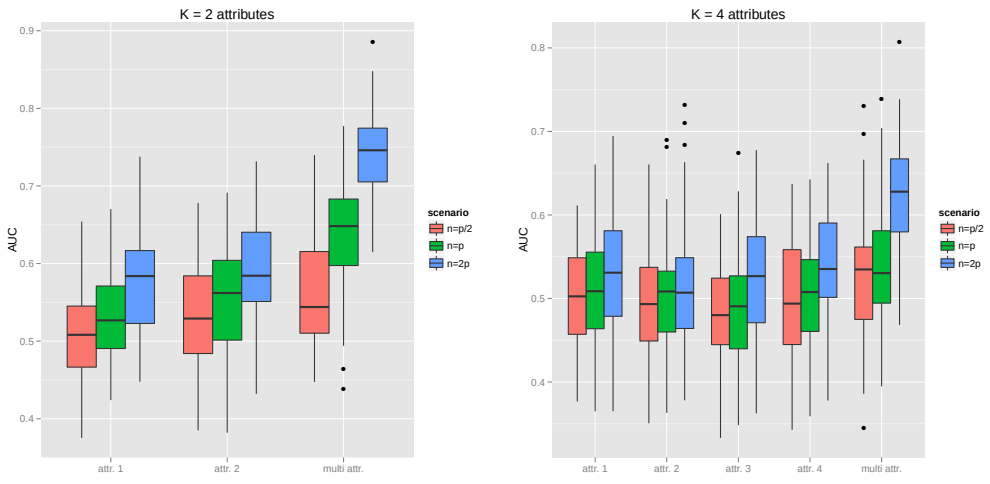
$$\Omega_1(B_j) = \sum_{i \in P n j} k_{ij} k_2. \tag{2.28}$$

However, this idea remains only at the feasible stage for now.

Numerical study. We propose a simple simulation to illustrate the interest of using multi-attribute networks and the efficiency of our proposal. The simulations are set up as follows:

1. Draw a random undirected network with p nodes from the Erdös-Renyi model;
2. Expand the associated adjacency matrix to multivariate space with
$$A = (A + I) \quad I_{K \ K};$$
3. Compute a positive definite approximation of A by replacing null and negative eigenvalues by a small constant;
4. Control the difficulty of the problem with > 0 such that $= + I$;
5. Draw an i.i.d. sample X of $X \ N \ 0, \ ^1$.

We choose small networks with $p = 20$, with 20 edges on average and vary from $p=2$ to $2p$. We consider cases where the number of attributes is $K = 2$ or $K = 4$. We either apply the usual neighborhood selection procedure on each dimension separately, or its multi-attribute counterpart with group-like penalty (2.26) on the multivariate data. We compute the AUC for each method and replicate the experiment 50 times. On Figure 2.7, it is clear that aggregation improves upon single-attribute methods.



(a) $K = 2$ (b) $K = 4$

Figure 2.7 –*Simple simulation study for the multi-attribute network inference problem: the multivariate procedure improves over the univariate procedures in every situation when networks are close for each attribute.*

Illustration: Gene / Protein regulatory network inference. As an illustration, we applied our sparse multi-attribute GGM approach to the NCI-60 cancer line data set. This data set consists in molecular profiles on a panel of 60 diverse human cancer cell lines. We use both protein and gene profiling experiments. For the former, we have samples for 92 antibodies from reverse-phase lysate arrays (RPLA); for the latter, expression is measured for 9,000 RNA with Human Genome U95 affymetrix. A *consensus set* composed of 91 protein and the corresponding gene profiles is retained for the $n = 60$ samples.

We infer a sparse GGM on each attribute (gene and protein), separately to start with, and then on its multi-attribute version. We do this on a large grid of the tuning parameter and thus have three families of networks indexed by their number of edges. Figure 2.8 demonstrates that our sparse multi-attribute method capture the characteristics of both univariate networks, as the Jaccard similarity index is high between each uni-attribute network and the multi-attribute network, while it remains low when comparing uni-attribute networks together. This tends to prove that this multi-attribute version proposes a consensus version of the interactions at hand in the cell, and one which is hopefully more robust to noise and small misregulations.

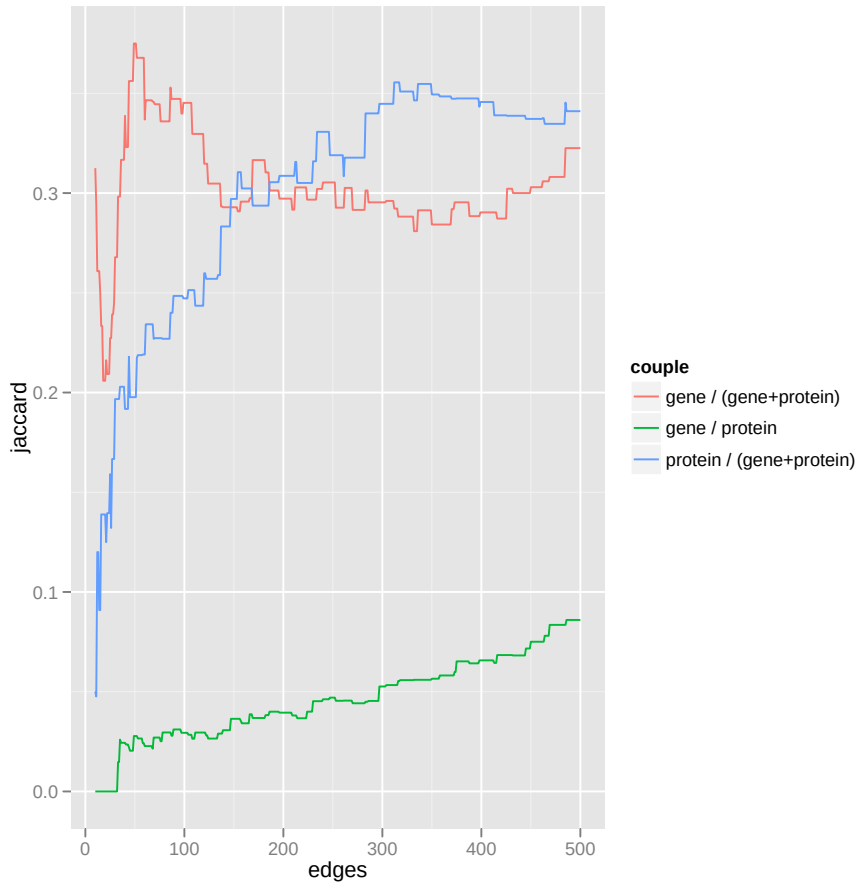


Figure 2.8 –*Jaccard's similarity index* $J(A,B) = \frac{|A \setminus B|}{|A[B|}$*: multi-attribute network shares a high Jaccard index with both uni-attribute networks.*

2.3 PERSPECTIVES

Network inference based upon sparse methods has been a hot topic in statistical learning since the middle of the 2000s. We have seen that the original baseline methods published between 2006 and 2011 have a well-controlled statistical framework and are supported by good computational tools. However, they do have some limitations, especially regarding the treatment of heterogeneous data. This motivated the emergence of a second wave of works the common objective of which is to overcome this restriction. A typical example that has received much attention recently is the multiple network inference framework. My contributions belong to this second wave and I think that I have brought some significant proposals for accounting for various sources of heterogeneity and structures that characterize modern complex data, where genomics data hold a place of choice. Several important issues remain to be addressed, however. Among these, an emblematic and recurring question, which is typical in unsupervised problems, is the difficulty for evaluating the level of trust that we have in the reconstructed network, with the exception of some idealistic cases that do not fit any real data situations. This suggests questions concerning both statistical and application perspectives.

On the statistical side, we naturally ask for tools to evaluate the *significance* of the inferred network. In other words, we would like to be able to test the significance of an edge, or to provide it with a p-value: statistical inference for sparse methods is an important matter and several works recently tried to tackle this issue [110, 86, 12]. A related question is the comparison of two networks, which can be stated as a test problem, either at the global level (are these two networks different?) or the local level (is this edge significantly different between these two networks?). This question is partially addressed in [26]. Another statistical issue is the question of characterizing the robustness of the proposed estimators, which seems especially challenging for sparse estimators and high dimensional data, where the notion of outlier is hardly relevant. In the sparse GGM framework, interesting preliminary work on influential function is proposed in [8] and seems promising.

Several questions also arise on the application side, as sparse GGM fail to reconstruct real regulatory networks at a low false positive rate, when only based upon transcriptomic data, even in a data regime supported by the theory. More disturbingly, recent results from the DREAM challenge [115] show that no method among (Dynamic) Bayesian Network, GGM, Random forest or Mutual Information based methods clearly dominates the others. It really questions the utility of such models, since typically the reconstructed network represents statistical interactions which probably do not have the same meaning as the biological interactions expected by the biologist.

In my opinion, even if the reconstructed networks are not straightforwardly interpretable in terms of biological mechanisms and interactions, they still potentially capture some important statistical features which can be used to strengthen the discovery of other biological processes that rule the cell. In this perspective, I am continuing to work on sparse GGM approaches in genomics with two guidelines:

1. perform more data integration and handle various sources of heterogeneity,
2. couple network inference with the estimation of other biological features, with

the hope of enhancing the estimation of the latter, and maybe of the former as well.

I briefly depict some on-going works and ideas in this vein in the following couple of paragraphs.

Enhancing network reconstruction by embedded transcription factor elucidation. Transcription factors (TF) are proteins which regulate gene expression. Keeping a good knowledge of TF is crucial in order to decode complex biological mechanisms which control gene expression. It could help to robustly drive the reconstruction of regulatory networks, that is, infer direct inhibition and activation relationships between a set of genes. This is already done in the most successful sparse network inference methods[71], where the candidate edges are only considered from TF to target genes. Still, this information is not always available, or very partial: a sound approach would be to jointly recover the candidate genes for TF and their potential links (or edges) with respect to other target genes. In [PP2], I am developing with Stéphane Robin and Tristan Mary-Huard a general purpose sparse multivariate procedure integrating various kinds of prior. We illustrate, among other applications, how to unravel *regulatory motifs* associated with the genes encoding those TF, which could provide weights to drive the network inference in a second step.

I am also working with Marie-Laure Martin-Magniette on validating such methods on well-controlled data for the model plant *Arabidopsis thaliana*, for which many sources of data and prior biological knowledge are available. By these means, we hope to evaluate the relevance of GGM for modeling different kinds of biological processes and interactions.

Coupling differential analysis and network inference. The topic of Trung Ha's PhD. thesis, which I am co-supervising, concerns the multiple network learning framework of Section 2.2.2, but with an additional assumption on the means of the Gaussian vectors: we consider several related Gaussian vectors $X^{(c)} \sim \mathcal{N}(\mu^{(c)}, \Sigma^{(c)})$, for $c = 1, \ldots, C$; our objective is to estimate both $\mu^{(c)}$ and $\Sigma^{(c)}$, assuming that the means *and* the covariance matrices respectively share some commonalities across tasks. To this end, we rely on a sparse multivariate convex criterion, where we encourage similarities between the vectors of means via a fused-Lasso penalty and the penalties described in Section 2.2.2 to encourage similarity at the covariance level.

On the application side in genomics, the goal of this project is to take advantage of the unobserved correlations between genes (described by the covariance matrices $\Sigma^{(c)}$) to enhance differential analysis (performed on the vectors of means $\mu^{(c)}$): existing works proposed to rely on known gene networks (e.g. via pathways) to drive the classification, either by including such information by means of the graph Laplacian[140], or by including the graph structure directly within the testing procedure to gain power[84]. On the other hand, knowing which genes are differentially expressed could be precious information for network inference purpose. For instance, the genes the expression of which are strongly correlated could be co-regulated and are hence more likely to belong to the same cluster in the inferred graph, as shown in [JP9].

Enumeration of Perturbation Scenarios in Biological Networks. This following path of research corresponds to a project submitted during early 2015 for an INSERM

grant with Étienne Birmelé, Pierre Neuvial, Mohamed Elati, Sophie Lèbre and collaborators from the Curie Institute. In the methodological part of this project, our idea is to explore near-optimal solutions of a network inference procedure to check whether those solutions are more relevant from the biological point of view. In a way, it bypasses the inference problem by instead questioning two antagonist goals which are often interchanged: biological interpretability and prediction accuracy.

The goal is to study from a mathematical viewpoint a perturbation model that integrates an inferred normal regulatory network with heterogeneous data from tumor cells to elucidate:i) How a small number of changes to the network alters the function of the network; ii) which regulators explain the observed alterations with respect to the normal behavior, and which processes are influenced by these driver regulators. As an application, through an established collaboration with the team of molecular Oncology of the Curie Institute, we will study two types of data for regulatory network alterations in tumorous bladder cells: (1) transcriptional gene regulation networks, with alterations taking into account expression and copy number variation data; (2) alternative splicing regulation networks, with alterations on exon array data. These applications should lead to the identification of new putative oncogenes and tumor suppressor genes associated with pathways specifically altered in tumors.

At this stage, we develop in a preliminary work [KJP3] a statistical methodology to identify misregulated genes given a reference network and gene expression data. The learning relies on a message-passing algorithm coupled to a sparse GGM method tailored to account for groups of co-activators and co-inhibitors in the reference network.

3 STRUCTURING PENALTIES TO ACCOUNT FOR COMPLEX DATA FEATURES

"All models are wrong, some are useful."

Guillem Rigaill

Contents

3.1 Background 65
3.1.1 Structured regularization with penalized methods 65
3.1.2 Computational consideration 72
3.1.3 Statistical analysis 75
3.2 Contributions 79
3.2.1 The cooperative-Lasso and sign coherent groups 79
3.2.2 Structured regularization for conditional GGM 88
3.2.3 A quadratic view of sparsity 96
3.2.4 Tree reconstruction with fusion penalties 103
3.3 Perspectives 112

THIS chapter is dedicated to my works on structured sparse methods. After a brief overview of the basic computational and statistical tools related to sparse regularization, I present four of my contributions to this field. I wish to demonstrate the diversity of these contributions, some being related to algorithmic and computational considerations, while some others are focused on the statistical properties of the methods. All emerged from motivations anchored in applications.

3.1 BACKGROUND

Section 1.2.2 of Chapter 1 detailed various motivations for relying on sparse and regularization approaches to analyze modern complex data. One of the most important features of these methods is their capability to provide structured estimators. By forcing the estimator to be endowed with a particular structure through regularization, we hope for a model that shows better statistical performance and that is more suitable for interpretation. Ever since such regularization methods have gained in popularity, a tremendous number of variants have emerged, "crafted" to account for various kinds of structures in the targeted set of parameters. This structure depends on the prior knowledge at our disposal (such as a natural grouping of the features, information about their spatial organization, or more simply sparsity). This section provides the reader with a brief overview of penalty-based approaches for structured regularization, using convex norms and for different kinds of structures. It also gives the basic pointers to the bibliographical references detailing the associated computational and statistical issues. Section 3.2 then presents my contributions to this field.

3.1.1 Structured regularization with penalized methods

Recall the framework presented in Section 1.2.2, that is, convex constrained optimization problems with the form

$$\underset{\in S}{\text{minimize}}\ f(\ ;\text{data}),\quad \text{such that}\quad (\ )\quad c, \tag{3.1}$$

where is the set of parameters of interest living in space; f is a convex function describing how well the model indexed by fits the data; is a convex set describing the constraint imposed on the parameters. Problem (3.1) can be equivalently[1] stated in an unconstrained Lagrangian form [18]

$$\underset{\in S}{\text{minimize}}\ f(\ ;\text{data}) + \quad (\ ). \tag{3.2}$$

In this latter form, we typically refer to ( ) as the "penalty" or "regularization term"[2], the amount of which is controlled by the positive tuning parameter. Technically speaking, regularization has various virtues: it generally aims at i) guaranteeing the existence of a solution, ii) guaranteeing the uniqueness of the solution (in the case of a strongly convex problem), iii) preventing lack of stability of the solution. In other words, it typically turns an ill-posed problem into a well-posed problem. In a statistical modeling perspective, the choice of also controls the behavior of our estimator, that is, of the solution to (3.2). Such a strategy typically reduces the variance by introducing a little bias.

§

In this section, I present regularization approaches where are convex sets obtained with a combination of ' -norms. We show how one may "play" with norms to obtain various structural behaviors for the estimated coefficients, such as controlling their size, performing selection, unraveling grouping structures or accounting for spatial organization, thus providing the estimator with some special structure.

[1] Problems are equivalent in the sense that there is a couple(,) such that the two solutions coincide.

[2] Strictly speaking, regularization can also be achieved by modifying the fitting term f (e.g., by replacing f by a surrogate function facilitating the optimization).

The bridge family and ‘ norms: Ridge, Lasso and others. We start by recalling the basic regularization effects induced by the use of standard norms. We illustrate our point by considering the case where the setof parameters is described by a vector with p real entries,*i.e.*, $\in \mathbb{R}^p$. A simple way to regularize is by controlling its ‘ -norm, defined by

$$\| \| = \left(\sum_{j=1}^{p} | _j | \right)^{1/}.$$

This idea is pursued in the linear regression framework with*ridge regression* [54, 59]. The range of -norms for > 0 defines the family of bridge estimators, obtained by regularization of ordinary least squares:

$$\hat{}_{,} = \arg\min_{\in \mathbb{R}^p} f(;X,y) + \| \| , \quad \text{with} \quad f(;X,y) = \frac{1}{2}\|y \quad X \|_2^2, \qquad (3.3)$$

wherey $\in \mathbb{R}^n$ is a vector of outcomes predicted by a linear combination of the columns of the $n \times p$ matrix of predictorsX. Although such a penalization is applicable to other loss functions beyond the quadratic loss, we rely on this important example to illustrate the structural nature of the regularization induced by -norm penalizations.

Regularization paths are a common visualization tool to gain insight into the effect of the parameters and . Figure 3.1 represents such paths for the bridge estimator, *i.e.*, f $\hat{}_{,}$, > 0g for a couple of striking values of , with data drawn as in Example (1.4) of Chapter 1.

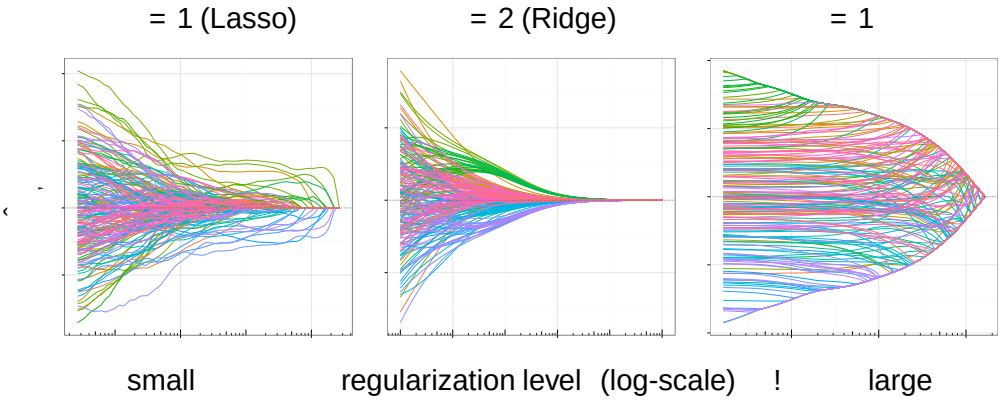


Figure 3.1 –*Regularization paths for the bridge estimators fitted with* R*package* quadrupen

From this figure, the most obvious effect of -regularization is to control the size of the coefficients: a sufficiently large value ofzeroes all the coefficients, while $\hat{}_{,}$ tends to the OLS estimate when ! 0. However, control of the size is achieved in various way depending on: with = 1 , all coefficients lie in a given convex envelop with a given magnitude; Ridge regression [77] (= 2) tends to group correlated features together along the paths, while the Lasso (= 1) has the great capability of activating the most relevant coefficients one after the other. More insights into the effect of on the induced regularization is gained thanks to simple geometrical arguments, and by considering the constrained formulation of (3.3), as follows:

$$\hat{}_{,} = \arg\min_{\in \mathbb{R}^p} f(;X,y) \quad \text{such that} \quad \| \| \quad c. \qquad (3.4)$$

In this equivalent formulation of bridge regression, the constraint induced by the ℓ_γ-norm defines a feasible set which is nothing more than the corresponding ball of radius $c^{1/\gamma}$ in $\mathbb{R}^p$ where c constrains the volume of the set. Figure 3.2 represents such sets for various values of γ when $\beta \in \mathbb{R}^2$ and $c = 1$.

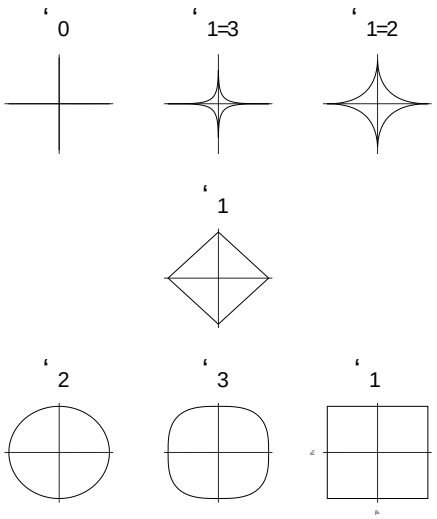


Figure 3.2 – *Contour of the feasible sets defined by* $\|\beta\|_\gamma \leq 1$ *for various values of* γ *when* $\beta \in \mathbb{R}^2$.

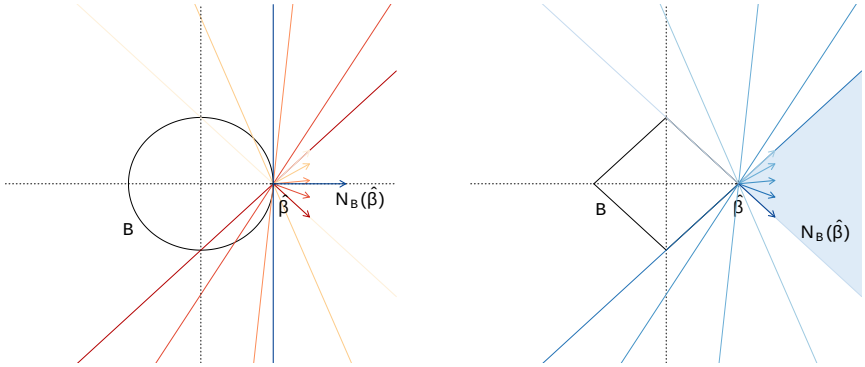
How does singularity induce sparsity? A simple argument from constrained optimization shows that a solution to Problem (3.4) is necessarily on the boundary of the sets drawn in Figure 3.2, as long as the unconstrained OLS solution does not belong to these feasible sets. In other words, the solution to (3.4) corresponds to the projection of the OLS solution onto the ball of radius $c^{1/\gamma}$. With this geometrical viewpoint in mind, we better understand how the various natures of the feasible sets in Figure 3.2 affect the regularization induced on the coefficients: in a variable selection perspective, the ℓ_0 pseudo-norm $\|\beta\|_0 = \text{card}\{j : \beta_j \neq 0\}$ defines the set of models over which we would like to optimize, that is, all models with a given number of nonzero entries, corresponding to as many predictors involved. However, the resolution of this combinatorial problem is prohibitive even for moderate values of p (say, 30). The same computational argument applies for non-convex norms (*a.k.a.* *quasi*-norms), that is, ℓ_γ-norms in the range $0 < \gamma < 1$: the presence of singularities at the boundary induces variable selection by exactly zeroing some coefficients; still, non-convexity means problems which are hard to optimize. On the contrary, convex norms in the range of $\gamma \geq 1$ smoothly control the size of the coefficients, but absence of singularity prevents them from achieving variable selection. At the interface of pseudo-norms ($\gamma \leq 1$, performing variable selection) and norms ($\gamma \geq 1$, enjoying convexity), the only candidate that remains is the ℓ_1-norm, a.k.a. the Lasso [164] in the linear regression framework.

The ability of the ℓ_1-norm to promote sparsity is clearly stated by considering the first-order optimality conditions for (3.4), stating that a point $\hat{\beta} \in \mathbb{R}^p$ is optimal if and only if the least square derivative $\nabla f(\hat{\beta}; y, X) = -X^\top(y - X\hat{\beta})$ defines a supporting hyperplane to the feasible set at $\hat{\beta}$. In other words, the opposite of the least square derivative must belong to the normal cone $\mathcal{N}_B$ to the feasible set at $\hat{\beta}$, where the normal cone to a convex set $\mathcal{C}$ at point x_0 is defined by $\{y \in \mathbb{R}^p, \langle y, x - x_0 \rangle \leq 0, \forall x \in \mathcal{C}\}$.

Thereby, for every in the feasible set, the least square derivative must satisfy

$$h\ddot{u}y \quad X\hat{\ }, \hat{\ } \quad i \quad 0.$$

Figure 1.3 pictures unit balls in R^2 for ℓ_1 and ℓ_2 balls, along with their normal cones at $(1,0)$. If we think of the least square derivative as a continuous random variable (as a function of "), then it will almost never fall into the normal cone to the ℓ_2 ball at $(1,0)$, which is degenerated into a single half-line of zero Lebesgue mass. On the contrary, there is a non negligeable probability for it to fall into the normal cone to the ℓ_1 ball at $(1,0)$, thanks to the singularity. In other words, contrary to the ℓ_2-norm which is differentiable on R^p, the ℓ_1-norm favors the selection of its points of singularity, which are interestingly located on the axis, shrinking some coefficients to 0.



(a) *Optimal point* $(1,0)$ *on the* ℓ_2 *ball* (b) *Optimal point* $(1,0)$ *on the* ℓ_1 *ball*

Figure 3.3 – *Geometry of sparsity and optimality*

Accounting for prior knowledge with group-norms and mixed-norms. Being now equipped with the basic norms, we would like to blend them in order to introduce a wide variety of structures depicting different types of prior information that can be extracted from external sources of knowledge. This paragraph addresses the case where such information can be described by a group structure on the variables $P = f1,\ldots,pg$ that we denote by G in general. We assume that G contains K elements such that $G = fG_1,\ldots,G_Kg$. Each group G_k is non empty, that is, it contains at least one element from P. Depending on additional assumptions on G and its elements, this structure corresponds to a partition, a hierarchy or an ordering on P.

Now that the set of variables is endowed with a known group structure, the idea is to impose a regularization scheme which is faithful to this structure. A natural way to achieve this goal is to control the coefficients at the group level. This can be done by measuring the volume of the coefficients by means of norms decomposing at the group level, or *mixed-norms*, which are defined in general as follows:

$$k \quad k_{,} = \sum_{k=1}^{K} ! \quad \sum_{j2G_k} j \quad {}_j j \quad ! = \sum_{k=1}^{K} ! \quad {}_k k \quad {}_{G_k} k \;, \qquad (3.5)$$

where $_{G_k} = (\ _j, j \ 2\ G_k) \ 2\ R^{jG_kj}$ is the vector of coefficients restricted to the elements of G_k. The weights $f! \ _k, k = 1\ldots,Kg$ are used to adjust the regularization to

each group, typically by accounting for the number of elements that belong to it. In (3.5), groups are penalized according to an ℓ_1-norm, while elements within a group are penalized according to an ℓ_γ-norm. Choices of γ and λ for a particular problem should be guided by the same kind of consideration as in the bridge case treated above, depending on the desired behavior at the group or at the coefficient level.

We can find many instances of (3.5) in the literature. We only provide quick comments and important references as a starting point for the interested reader. In this perspective, Figure 3.4 represents the feasible sets in $\mathbb{R}^3$ for a series of couples (γ, λ), when the grouping structure splits $\{1, 2, 3\}$ into a partition such that $\mathcal{G} = \{\{1, 2\}, \{3\}\}$

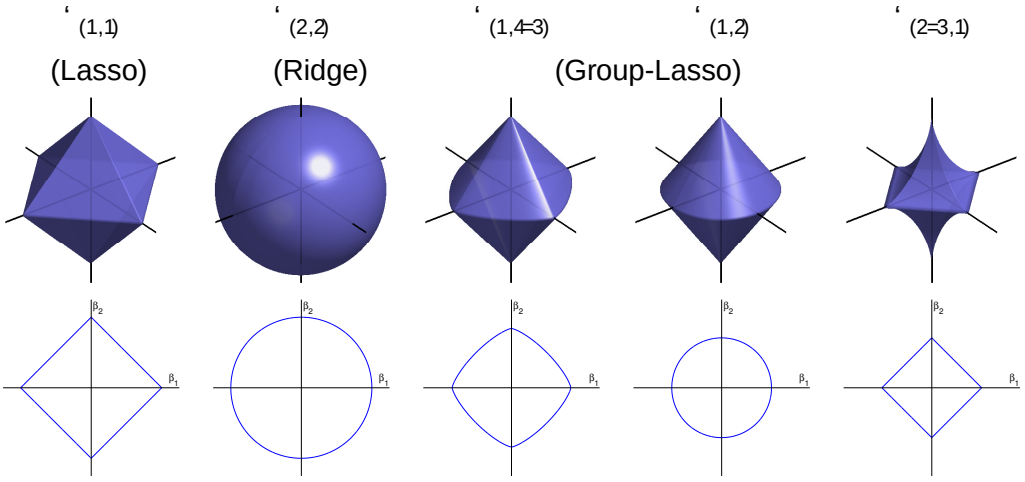


Figure 3.4 – *Feasible sets defined by the mix norms* $\|\cdot\|_{\lambda,\gamma}$ *for various couples* (γ, λ)*, with two groups* $G_1 = \{1, 2\}$ *(first plane) and* $G_2 = \{3\}$ *(vertical axis). (source:* [161]*, thanks to M. Szafranski).*

Choosing $\gamma = \lambda$ masks any grouping effect and we recover the corresponding norm, as exemplified for Lasso and Ridge regularization. On the right, we represent an example of (γ, λ) leading to a non convex set. The two remaining examples where $\gamma = 1$ are instances of the popular "group-Lasso", imposing sparsity at the group-level:

$$\|\beta\|_{\lambda,1} = \sum_{k=1}^{K} \omega_k \left(\sum_{j \in G_k} |\beta_j|^\lambda \right)^{1/\lambda} = \sum_{k=1}^{K} \omega_k \|\beta_{G_k}\|_\lambda . \tag{3.6}$$

The original group-Lasso [186, 190] corresponds to the case where $\lambda = 2$. A complete study for varying values of λ can be found in [175]. Group-Lasso is the first instance of "structured" sparsity. It gave birth to a number of works trying to generalize to a wider class of structures. In particular, working with a structure $\mathcal{G}$ which is a partition is too restrictive to integrate the vast sources of prior information that we can find in fields of application such as signal processing, imaging, genomics, etc. In this perspective, an overlapping group-Lasso was developed [85]. Some analyses of penalties where the grouping structure $\mathcal{G}$ defines a hierarchy on the variables is proposed [161, 162, 87]. The popular sparse group-Lasso [155] is just one instance of these. Structures more general than hierarchies, *i.e.* orderings, depicted by a direct acyclic graph, are used in the composite absolute penalties [194]. For a general study about sparsity with mixed-norms, the reader may refer to [99].

Accounting for prior knowledge with fusion and graph penalties. In the previous paragraph, we have seen how to introduce information about a known grouping structure between variables. However, such accurate knowledge is not always available, and we may have only partial or "smooth" information on the potential relationships. A typical example is prior spatial information: in a segmentation problem, we expect the successive points composing the signal to be mostly the same except for a few points, showing brutal changes corresponding to jumps in the signal. This idea is exploited in [69] by means of the total variation (TV) penalty, which penalizes the differences between the successive entries of a vector by $\Omega_{TV}(\beta) = \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|$. This idea generalizes to more complex relationships as follows: suppose that proximity between the variables can be depicted by a weighted graph $\mathcal{G} = (V, E, W)$ with vertices $V = \{1, \ldots, p\}$ and edges E weighted by values $W = \{\omega_{ij}, (i,j) \in E\}$. The generalized TV penalty on the graph $\mathcal{G}$ is defined by

$$\sum_{(i,j)\in E} \omega_{ij} |\beta_i - \beta_j| = \|D\beta\|_1, \tag{3.7}$$

with D a $|E| \times p$ matrix encoding the pairwise differences between the variables. In the standard TV-penalty encouraging similarity between neighbors, $\mathcal{G}$ is a chain graph with edges $E = \{(1,2), (2,3), \ldots, (p-1, p)\}$ and D a $(p-1) \times p$ bidiagonal matrix such that $D_{ii} = 1$ and $D_{i(i+1)} = -1$. Penalty (3.7) has also been referred to as the generalized Lasso [167]. We call it a "fusion" penalty, as it fuses pairs of coefficients β_i and β_j into the same value. When a smoother effect is desired, one may rely on a ℓ_2 version of this approach:

$$\sum_{(i,j)\in E} \omega_{ij} (\beta_i - \beta_j)^2 = \beta^T D^T D \beta = \beta^T L \beta = \|\beta\|_L^2. \tag{3.8}$$

Penalty (3.8) can be seen as a generalized version of Ridge regression: when no prior on the relationships between the variables is available, we set $L = I$ and encounter the usual ridge penalty. Meanwhile, L has a nice interpretation in spectral graph theory [32], as it is the combinatorial Laplacian of $\mathcal{G}$. This penalty has been used for instance in [140] to integrate prior information in classification problems for genomics.

These two kinds of fusion penalties are convenient for encoding a wide range of structural information. They have been especially popular when coupled with an additional norm to induce, for instance, sparsity on top of the structural prior defined by G. The corresponding regularization terms are written as a mixture of two penalties:

$$\|\beta\| + (1-\gamma)\|\beta\|_L^2 \tag{3.9a}$$

$$\|\beta\| + (1-\gamma)\|D\beta\|_1, \tag{3.9b}$$

with $\gamma \in [0, 1]$. Figure 3.5 represents possible options derived from (3.9a), (3.9b).

Consider the case where $\gamma = 1$, *i.e.* where sparsity is promoted on top of structured regularization of the parameters. First, in (3.9a), ℓ_1 and ℓ_2-norms are mixed, which is known as the Elastic-net [198]. In its standard version, no structure is introduced, that is, $L = I$. Still, the ℓ_2 norm tends to group correlated variables, thus the Elastic-net is less sparse than is the Lasso, the latter tending to select a single candidate among a set of correlated variables. Note that the "structured" version of the Elastic-net, defined for general L, has been rediscovered many times in the literature. Possible references are [105, 156, 80]

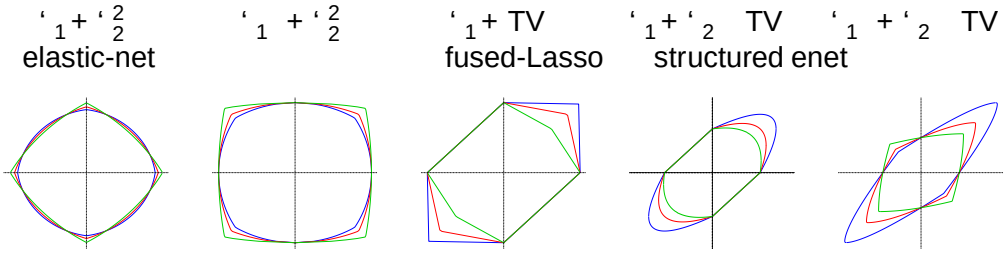


Figure 3.5 –*Some mixtures of penalties promoting diverse structures between the variables*

Second, with (3.9b), we couple two ℓ_1 penalties: when the structuring regularization boils down to the TV-penalty, we encounter the popular Fused-Lasso [166] which promotes piecewise-constant coefficients while zeroing most of the segments of the signal. The generalized version was studied [76]. To my knowledge, cases where $\alpha \neq 1$ are absent from the literature. I implemented cases where $\alpha = 1$ in the R-package quadrupen, as the ℓ_1 has the nice property of being dual to the ℓ_1-norm and may be used in future work on its own.

In order to illustrate how including structural information can have a dramatic effect on interpretability, let us consider a couple of regularization paths for some mixtures of penalties. To this end, we rely on the same linear regression problem (1.4) as for the bridge regularization paths from Figure 3.1. Figure 3.6 represents the paths obtained by mixing the ℓ_2-fusion penalty (3.8) with a simple chain graph (encouraging a smooth similarity between neighbors) to other norms: on the left panel, this norm is used by itself; on the middle panel, it is mixed with the ℓ_1-norm, thus performing structured sparsity with a generalized form of the elastic-net; on the right panel, it is mixed with an ℓ_1 -norm. Each color corresponds to a group in the simulation setup. While the groups are unknown to the regularization procedure, they are recovered with the three variants. These results have to be compared with paths from Figure 3.1, where no special structure is encouraged by the regularization, and where the groups are completely lost.

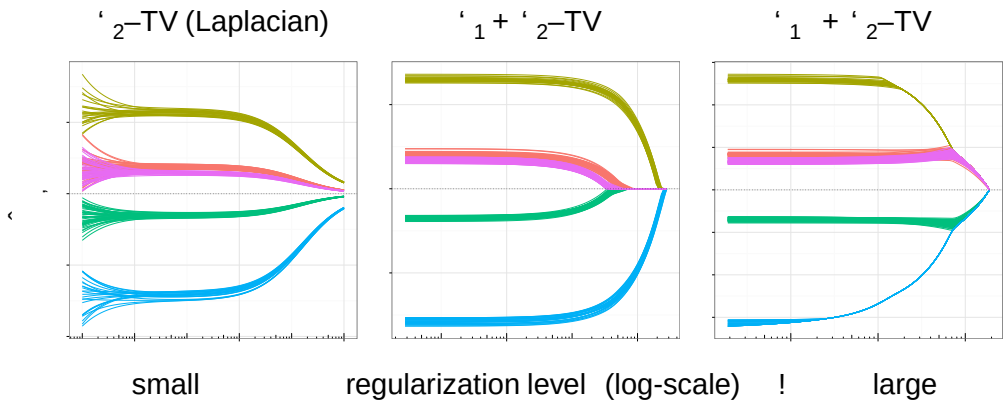


Figure 3.6 –*Regularization paths for the structured estimator fitted with the* R*-package* quadrupen

3.1.2 Computational consideration

The great success of penalized approaches comes from various reasons addressed in Section 1.2.2. Among these reasons, the existence of computationally efficient fitting procedures, largely due to tools from convex optimization, plays a central role. The literature developing new optimization tools to improve on the existent computational methods is thus extremely abundant. Although I do not target exhaustiveness, this section aims at insisting on the most important principles to solve structured regularization problems like (3.2), in cases where f is a convex differentiable function and Ω a convex norm with singularities. In such a situation, various general purpose convex optimization solvers could be used (see [118]). However, they do not exploit the structure of the regularization problem – especially sparsity – and thus are sub-optimal in terms of computational efficiency.

When f is convex and smooth, the techniques are relatively standard and in some cases, analytic solutions are available. This is the case for the generalized ridge regression, *i.e.* when $f(\beta, X, y) = \|y - X\beta\|_2^2$ and Ω is given by (3.8). The whole regularization path can be obtained at the cost of a single SVD decomposition, as stated in Equation (1.7), Section 1.2.2. Note that computing the SVD of an $n \times p$ matrix can be prohibitive when the problem dimensions increase. But in a high-dimensional setup when $p \gg n$, one only needs the truncated version of the SVD up to n singular values. When both n and p become large, approximate versions can be used by randomly sampling the columns of X (see e.g. [185]).

More demanding techniques are required when Ω is convex but non smooth, that is, when it involves singularities due to ℓ_1 or ℓ_1-norms. As my contributions and applications in genomics almost always concern variable or feature selection at some point, I mainly rely on algorithms and optimization procedures taking advantage of the sparsity of the problem. The next couple of paragraphs are thus concerned with general strategies that can be adopted to dramatically improve the computational efficiency in situations involving sparsity.

For a more detailed introduction to optimization methods suited for convex problems involving sparsity, [15] is of great help to the statistician as it covers many tools known from the optimization literature (general convex solver, proximal methods, coordinate descent, basics on duality, etc.). Since [15] was published, other techniques known from the optimization community like stochastic gradient methods [125], the Frank-Wolfe algorithm [100] or ADMM (Alternating Direction Method of Multipliers) [17] have gained in popularity and have been adapted to the resolution of sparse problems. In my opinion, a researcher working in the field of computational statistics should be aware of such techniques.

Active-set algorithms. These algorithms take advantage of the sparsity of the solution by solving a series of small linear systems, the size of which is incrementally increased/decreased. This approach was originally developed [134] for the Lasso in a linear regression framework. The same idea was then pursued for the group-Lasso in [147]. A more general description of working set algorithms – a close variant where the set of "active" variables can only grow – is provided in [15]. We present here similar ideas for the active set framework. A general active set method for solving Problem (3.2) in the presence of sparsity-inducing norms is sketched in Algorithm 1.

Algorithm 1: General purpose active-set algorithm

S0 **Initialization**

$A \leftarrow \{j : \beta_j^0 \neq 0\}$ // Start with a feasible β^0 // Recover the active-set

S1 **Update active variables** A

$\beta_{A^c} = 0$

$\beta_A \leftarrow \arg\min_{\beta \in \mathbb{R}^{|A|}} f(\beta_A) + \lambda\Omega(\beta_A)$ // (Small) subproblem resolution

S2 **Update active set** A **by monitoring the optimality conditions**

if $\Omega^*(\nabla f(\beta)) > \lambda$ *and* $\exists j \in A : \beta_j = 0$ then
  $A \leftarrow A \setminus \{j\}$ // Downgrade j
else if $\Omega^*(\nabla f(\beta)) > \lambda$ then
  find $j^* : \Omega^*(\nabla f(\beta)) > \lambda$, $A \leftarrow A \cup \{j^*\}$ // Upgrade j^*
else
  Stop and return // Optimality is reached

In words, this algorithm starts from a feasible sparse initial guess β^0 and then basically iterates over two steps:

1. Step 1 solves Problem (3.2) with respect to β_A, the subset of "active" variables, currently identified as being nonzero. At this stage the current feasible set is restricted to the orthants where the gradient of $f(\beta)$ has no discontinuities: the problem is smooth and can be solved with any convex optimization procedure.

2. Step 2 assesses the validity of the set A by checking the optimality conditions. They are monitored thanks to $\Omega^*(\nabla f(\beta))$, where Ω^* is the dual norm[3] of Ω. For the general group-Lasso (3.6) (including the Lasso), the dual norm is

$$\Omega^*(\beta) = \max_{k=1,\dots,K} \|\beta_k\|_{q_0}, \text{ where } q' \text{ is such that } \frac{1}{q} + \frac{1}{q_0} = 1.$$

   There are two possibilities: if elements from β_A have been zeroed during step 1 *and* optimality conditions are met, the corresponding variables are removed from A. Otherwise, if optimality conditions are violated, we add an element j to A (it can be a group of variables), by picking the one that most violates these conditions. This simple strategy has been observed to require few changes in the active-set. When no such violation exists, the current solution is optimal.

In the first step, only small convex problems need to be solved. In a high-dimensional setup where $n \ll p$ and where the underlying feature space is typically sparse, only a few activations/deactivations are required. Anyway, high-dimensional statistics [63] tells us that few guarantees can be obtained for estimators lying in large spaces when $n \ll p$ – and even when $n \gg p$ –. Thus, activating too many variables in such procedures would only produce estimators that do not make any statistical sense.

[3] See Proposition 1.6 in [15] to see how dual norms are related to the duality gap, and thus helps in monitoring the optimality conditions.

To solve each small problem, we can rely on various optimization methods (see the review [150] for the Lasso). In my personal implementation (package `pepadrupen` [SW3] or `scoop` [SW4]), I rely either on versatile and robust first order methods like proximal [129, 9] and coordinate descent [56, 59, 171] methods, or on more involved second order methods like BFGS methods with box constraints. The final choice depends on the problem size and on the level of accuracy required. I discuss this point in detail in Section 3.2.3, which summarizes the working paper [PP3].

Finally, note that to compute a series of solutions along the regularization path for convex problem (3.2), we simply choose a series of penalties $\lambda_1 = \lambda_{\max} > \cdots > \lambda_\ell > \cdots > \lambda_L = \lambda_{\min} \geq 0$ such that $\hat{\beta}(\lambda_{\max}) = 0$ and then use the usual warm start strategy, where the feasible initial guess for $\hat{\beta}(\lambda_\ell)$ is initialized with $\hat{\beta}(\lambda_{\ell-1})$.

Homotopy algorithms and piecewise linear regularization paths. In some situations, active-set algorithms can be made even more efficient when computed on a series of values. This is due to a special property of the associated regularization path $\{\hat{\beta}(\lambda) : \lambda \geq 0\}$, namely, when this function is a piecewise linear function of λ. In this case, we can detect events corresponding to the activation or deactivation of a variable and compute the exact values of λ associated with these events. The procedure following the whole path of solutions along these values is called a homotopy algorithm.

The homotopy algorithm associated with the Lasso for linear regression was first proposed in [133]. Then, more insight was gained in the famous LARS paper [48], which put forward stagewise regression and the Lasso all together in a unifying framework. More generally, [146] gives sufficient and necessary conditions for the existence of such a property for the family of penalized Problems (3.2). Basically, the function f must be a piecewise quadratic function of β, and the regularization term Ω must build on the ℓ_1 and/or the ℓ_∞ norms. Relying on these ideas, a path algorithm for the generalized fused-Lasso problem is proposed in [76]. The OSCAR penalty defined in [14] could also theoretically come with an accompanying homotopy algorithm, although it has never been implemented to my knowledge.

Finally, note that there is in general no guarantee for the number of steps to be small in a homotopy algorithm. In fact, [114] exhibits pathological cases for the LARS algorithm where the number of kinks in the piecewise linear path grows exponentially with the number of variables. Such cases are extremely unlikely on real data, however. Note that for a version of the fused-Lasso, we show in Section 3.2.4 of this chapter that choosing appropriate weights in the penalty brings guarantees on the number of steps.

Screening rules. Another possibility to gain speed and scalability is to discard some features which are guaranteed to never enter the model, by means of simple rules. In the linear regression framework, these rules are typically based on the marginal correlations between the response variable and the predictor variables. In the $n \ll p$ setup, the goal would be to reduce the initial number of predictors p to an order similar to n by a fast and efficient method. Such methods, when they are guaranteed to keep all the important variables, have been referred to as "sure screening" [51]. More generally, one refers to them as a "screening rule".

To my knowledge, [49] was the first to bring a proposal for the Lasso in the linear regression setup. Their "Safe rule" discards the j-th variable if

$$|x_j^T y| < \lambda - \|x_j\|_2 \|y\|_2 \frac{\lambda_{\max} - \lambda}{\lambda_{\max}},$$

where $_{max}$ is the smallest value of the tuning parameter discarding all variables, while $< _{max}$ is the current value for $\hat{}$. Of course, this simple rule is more likely to be violated when gets smaller. To overcome this limitation, [165] proposes a "recursive" variant of this rule called the "Strong rule":

$$jx_j^{\ddot{u}}(y \quad X\hat{}_{k\ 1})j < 2 \ _k \quad _{k\ 1}.$$

Such a recursive approach can be typically used to maintain a subset of "potentially active" variables, when fitting a regularization path over a grid. However, the strong rule is not safe and a relevant variable may be discarded.

These rules generalize to other loss functions (for instance, logistic regression) and other Lasso-type penalties (e.g., group-Lasso, Elastic-net and so on). Other approaches coming from optimization consider dual problems to build screening rules for the Lasso family [178]. Then, it is almost straightforward to adapt these rules to a penalty based upon $_1$ regularization. When the initial number of features is very large, such a rule could really save time by limiting the computational work requiring storage of objects with large size in the RAM.

3.1.3 Statistical analysis

Theoretical analysis of sparse and shrinkage estimators in a high-dimensional setting is a relatively new field in statistics. A consequence is that a huge number of works has been published in the past decade and many statistical tools have emerged to assess the properties of sparse estimators. In penalized problems like (3.2), the desired properties can be measured in terms of *i*) prediction capability, *ii*) quality of the parameter estimation, and *iii*) capability of unraveling the true non-null model (*i.e.*, support recovery of the true parameters in case of sparse methods). Although we look for an estimator showing good performance on these criteria both in low and high-dimensional settings, support recovery is more typical of sparse estimators and high-dimensional statistics. These questions have been addressed by researchers from different communities (*e.g.* statistics, machine learning, signal processing), with closely related tools, and this is not the place to provide an exhaustive bibliography. A sound statistical synthesis of these tools is the book [121]. The recent piece of work [163] provides a slightly different point of view, at the interface of machine learning and statistics.

In this part, I present a couple of fundamental notions on these questions to helping the understanding of the upcoming contribution section 3.2. The discussion leans on Camille Charbonnier's PhD thesis, which I co-supervised between 2009 and 2012.

Basic statistical assumptions for sparse estimators: the Lasso-case. Although most of the principles developed in this part apply to many sparse estimators, the discussion focuses for illustrative purposes on the estimator defined by the Lasso:

$$\arg\min_{2R^p} \frac{1}{n} ky \quad X \quad k_2^2 + \ _n k \quad k_1,$$

where the dependency on the sample size is explicitly stated for the purpose of asymptotic analysis. The true underlying model is $= X^T \ ^? + "$, with $" \quad N \ (0, \ ^2)$.

An exhaustive summary of the various assumptions required to guarantee estimation and selection properties of the Lasso is given in [21]. Figure 3.7 provides a simplified version. It highlights the distinction between *irrepresentability conditions* required

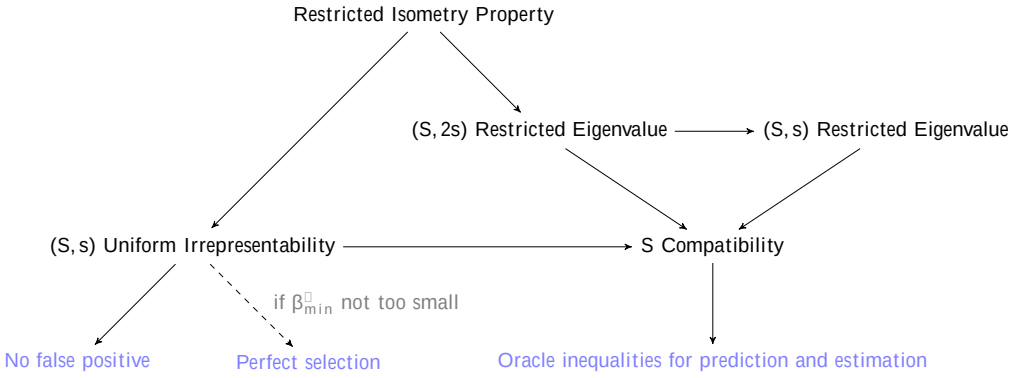


Figure 3.7 – *Summary of causal links between main assumptions required to prove estimation and selection properties of the Lasso (thanks to Camille Charbonnier; source:* [21]*).*

for selection consistency and milder *restricted eigenvalue assumptions* required for estimation and prediction oracle inequalities.[4] The former has notably been proved necessary for selection properties, and the latter, in its "compatibility" formulation, is the weaker assumption required to obtain at least estimation and prediction consistencies.

Irrepresentability condition. The irrepresentable condition, also known as mutual incoherence condition in the community of signal processing, appears simultaneously in a large body of work as a sufficient and necessary condition for selection properties of ℓ_1 regularized least squares ([105] in statistics, [44] and [170] in the field of signal processing, while [122] defines the equivalent assumption of neighborhood stability for sparse GGM inference). The main assumption remains the same in both its deterministic and Gaussian random design forms. We recall the former here, where A is the subset of relevant covariates and A^c its complementary subset.

Definition (Irrepresentable condition for the Lasso under deterministic design). *Consider a fixed design stored in an $n \times p$ matrix X. There exists $\eta > 0$ such that:*

$$\|X_{A^c}^T X_A (X_A^T X_A)^{-1} sign(\beta^\star_A)\|_1 \leq 1 - \eta .$$

Parameter η is referred to as the incoherence parameter of exact recovery coefficient. This condition stems from the primal-dual witness construction clearly formulated in [176] and guarantees that no irrelevant covariate can be included in the model on top of relevant covariates. Intuitively, these conditions measure in terms of correlation how close irrelevant covariates are to relevant covariates in such a way that least squares could be misguided into including those irrelevant covariates, hence the regression term of irrelevant covariates onto relevant ones $(X_A^T X_A)^{-1} X_A^T X_{A^c}$ coupled with the sign of $\beta^\star_A$. Indeed, a high-correlation between relevant and irrelevant covariates only presents a risk if it is of the same sign as the true coefficient.

Earliest works based upon the irrepresentable condition required an asymptotic framework. A probabilistic approach was introduced in [176] to work at fixed n.

[4] The restricted isometry property (introduced by [23] in the field of compress-sensing) is one of the main assumptions usually used to prove consistency results, but we will not dwell on that one since both previous assumptions are weaker.

Restricted Eigenvalue assumption. Relaxing the objective of perfect support recovery, other good properties can be demonstrated for the Lasso under weaker conditions. In fact, it can adapt itself to the true sparsity level in order to perform at minimax rates up to a logarithmic factor in terms of estimation and prediction, under weaker conditions called restricted eigenvalue conditions.

Definition (Restricted Eigenvalue assumption) *Consider a given amount of sparsity* $s \leq p$. *There exists* $\kappa(s) > 0$ *such that:*

$$\min_{S \subset \{1,\dots,p\}, |S| \leq s} \min_{\|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1} \frac{\|X\delta\|_2}{\sqrt{n}\|\delta_S\|_2} \geq \kappa(s). \tag{3.10}$$

The basic intuition is as follows: in classical statistical terms, sharp estimation and prediction properties are met when the Fisher information is large enough so that an estimation gap $\delta = \beta - \hat{\beta}$ induces a difference in likelihood of at least $\|\delta\|_2^2$. Analytically, we derive a second-order Taylor series expansion near β^* in the direction δ, to observe that this strong convexity assumption amounts to uniformly lower bounding the eigenvalues of the Hessian matrix in the neighborhood of the true parameter β^*:

$$\|y - X\hat{\beta}\|_n^2 - \|y - X\beta^*\|_n^2 = \frac{2}{n}\langle X^T(y - X\beta^*), \delta\rangle + \|X\delta\|_n^2 + o(\|\delta\|_2^2).$$

Of course, a uniform lower bound is too strong in high-dimensional settings. Therefore on top of considering reduced size matrices, we focus on a restricted neighborhood, which is the cone $\{\delta \in \mathbb{R}^p, \|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1\}$, where we know the Lasso error term $\delta = \beta^* - \hat{\beta}$ to reside, hence the denomination *restricted eigenvalue*.

The consequence of the restriction to the cone is that there is no guarantee that the solution will be unique. However, with large probability, all solutions are concentrated within the same ℓ_2 or ℓ_1 ball around β^*. Moreover, under supplementary assumptions on the miminal nonzero value, estimation or prediction bounds can be completed by thresholding steps in order to provide model selection guarantees.

This assumption is the weakest assumption possible except for a slight modification which consists in changing the $\|\delta_S\|_2$ at the denominator into a $\|\delta_S\|_1$ (we hence obtain the "compatibility assumption"). However, we lose the eigenvalue interpretation. We refer to [127] for a generalization of this assumption to address regularized M-estimators under a larger spectrum of sparsity assumptions $\mathcal{O}$, and to [33] for a non-parametric setup.

Other notable results. Statistical results relying on these two assumptions are very important in order to understand the behavior of sparse estimators in a high-dimensional setting. However, we can never verify them on real data. To avoid such assumptions on the design matrix (in particular the strong irrepresentability condition for support recovery), *adaptive* versions of sparse estimators, notably for the Lasso [197] were introduced. They basically weight the penalty related to each covariate according to a previous estimator supposedly asymptotically consistent, like the OLS. Although consistent estimators are not available, two-step procedures which build the weights in the first step and scale the penalty associated to each covariate accordingly in the second step show very good performance: they are state-of-the-art versions of the Lasso and widely used in practice.

More practically, recent works [179, 135] try to "precondition" the data in order to be more in agreement with the above conditions and enhance Lasso performance.

Finally, we underline that much effort has recently been put into providing sparse methods with a complete framework for statistical inference, that is, into performing statistical tests in a high-dimension setting (see [124] and [110]).

Model selection and parameter tuning. We close this section by addressing the most important issue of penalized problems, both from the theoretical and applied viewpoint, namely, the choice of an appropriate amount of regularization: although many solutions are available, none is universally better than the others.

In the two preceding paragraphs, λ_n was purposely considered as given. Actually, the questions related to correct estimation or support recovery are conditioned by an "appropriate" choice of λ, which typically determines the size of the model in the case of sparsity-inducing norms. Hence model selection, which amounts here to choosing the tuning parameters, is a recurrent issue for sparse methods.

Trial values $\Lambda = \{\lambda_{\min}, \ldots, \lambda_{\max}\}$ define the set of models we have to choose from along the regularization path[5]. We aim at picking either the model with minimum prediction error, or the one closest to the true model. These two perspectives generally do not lead to the same model choice: when looking for the model minimizing the prediction error, K-fold cross-validation is the recommended option [74] despite its additional computational cost. As an alternative, the penalized criterion developed in [65] addresses the problem of selecting the estimator with the smallest Euclidean risk among any family of estimators in the linear regression setup. In particular, this criterion is valid under high-dimensional settings and is proved to satisfy non-asymptotic risk bounds under no assumptions on the true model. Still, it requires the resolution of a convex problem, the cost of which may be similar to cross-validation.

For a choice of the tuning parameters more suited to the selection of the true model, information criteria provide a fast way to perform model selection, as an alternative to Breiman's "1-SE" rule for CV, which picks the sparsest model within one standard error of the minimum. For penalized methods, the general form of an information criterion is expressed as a function of the log-likelihood and the effective degrees of freedom of the fit. However, we have to give some sense to the notion of degrees of freedom associated with regularized estimators. This question is resolved in [170] for estimators defining a smoother, like ridge regression, in which case the degrees of freedom equal the trace of the hat matrix (see Expression (1.8) in Section 1.2.2 and example therein). With sparse methods however, the problem is different as the estimator does not have a closed form. This question is relatively well treated for Lasso-style problems: [200] introduces degrees of freedom, coming with AIC, BIC and Mallows's Cp criteria for the Lasso. Then, works on the generalized-Lasso [167, 168] provide degrees of freedom for a wider class of Lasso problems, in light of constrained optimization.

Rather than applying BIC or AIC, new information criteria have been proposed more suited to the high-dimensional setup. A notable example is the *extended*-BIC [27] that we already mentioned in the sparse GGM framework (2.10):

$$\text{EBIC}(\lambda) = -\log \ell(\hat{\theta}_\lambda, X, y) + |A_\lambda| \frac{\log n}{2} + \log \binom{p}{|A_\lambda|},$$

with A_λ the current support of $\hat{\theta}_\lambda$. This criterion comes from the addition of a uniform prior on the set of models tested along the regularization path: starting with p variables, each model with size s is given a prior probability of $(p+1)^{-1}\binom{p}{s}^{-1}$.

[5] If several regularization parameters are at stake, we work on multi-dimensional grids.

3.2 CONTRIBUTIONS

The reader is now equipped with the standard background to discuss my contributions to penalty-based approaches. These contributions are driven by two main guidelines: first, shape the regularization to account for a particular feature of the data or the problem at hand; second, focus and develop methods suited to high-dimensional problems.

Section 3.2.1 presents the paper [JPG] written with Yves Grandvalet and Camille Charbonnier. This work thoroughly studies the cooperative-Lasso in the linear regression framework, a penalty that we initially introduced for multiple network reconstruction [JP7]. The coop-Lasso is a refinement of the group-Lasso that promotes sign-coherence within groups. It also allows for sparsity within groups while not suffering from an additional tuning parameter as does the sparse group-Lasso. Algorithms for the coop-Lasso and related methods are distributed in the `sgcoop` R-package [SW4].

Section 3.2.2 corresponds to a contribution that saw the light of day during the beginning of my stay in Stéphane Robin's lab in late 2012. With Stéphane and Tristan Mary-Huard, we wanted a general regularized multivariate regression framework suitable for various applications in genomics and genetics. At the end of the day, our method accounts for unknown correlation between the responses and allows the integration of smooth prior information to drive the selection of the most relevant predictors. At this stage, this work was published in proceedings [GI3]; the journal version [PP2] is still under review. The R-package `spring` [SW1] implements our proposal.

In Section 3.2.3, the third contribution focuses on computational and optimization aspects of sparse methods. We propose a unifying view of a large family of methods mixing the ℓ_2-norm and either the ℓ_1 or the ℓ_∞ norm, by representing the corresponding feasible sets as the intersection of quadratic sets. This has a connection with homotopy algorithms. At this stage, a tech report [PP3] is available and the method is implemented for the generalized Elastic-net and what we called "bounded regression" ($\ell_2 + \ell_\infty$) in the R/C++-package `quadrupen` [SW3].

In Section 3.2.4, we present a work that arose from a collaboration with Guillem Rigaill and Pierre Gutierrez [JP2]. It studies a version of the (M)ANOVA regularized by means of weighted fusion penalties. When using the ℓ_1-norm for fusing the parameters, this method can be used to reconstruct hierarchies at very large scales. More technically, we introduce weights ensuring a piecewise-linear regularization path, with no split events, and an estimator that asymptotically enjoys oracle properties for support recovery. It is implemented in R/C++-package `fusedanova` [SW2].

3.2.1 The cooperative-Lasso and sign coherent groups

This work addresses the problems of estimation and inference of parameters when a group structure among parameters is known. Compared to the group-Lasso, we assume a stronger assumption: groups should not only reveal the sparsity pattern, but should also be relevant for sign patterns: all coefficients within a group should be *sign coherent*, that is, should either be null, non-positive or non-negative. This desideratum often arises when the groups gather redundant or consonant variables (a usual outcome when groups are defined from clusters of correlated variables). To perform this sign-coherent grouped variable selection, we propose a novel penalty that we call the cooperative-Lasso, in short the *coop-Lasso*. The coop-Lasso is amenable to the selection of patterns that cannot be achieved with the group-Lasso. This ability, which can be observed for finite samples, also leads to consistency results under mildest assumptions.

Cooperative-Lasso: definition and optimality conditions. In our original paper [JPV], the coop-Lasso was defined as an analogue of the ℓ_2 group-Lasso, when the group structure $\mathcal{G}$ defines a partition of the set $\mathcal{P} = \{1, \dots, p\}$ and when these groups are sign-coherent. However, we may relax this assumption and generalize to a broader class of structure, like a hierarchy, and define the coop-Lasso as an analogue to the ℓ_1 group-Lasso (3.6). The corresponding family of regularization norms is as follows.

Definition (Cooperative norms) *Let* $v = (v_1, \dots, v_p)^\top \in \mathbb{R}^p$ *and* p_k *denote the cardinality of group* k. *We define* $v_G \in \mathbb{R}^{p_k}$ *as the vector* $(v_j)_{j \in G}$. *The* $\ell_{\gamma,\pi}$ *coop-norm of* v *is defined as a sign-adaptive mixed* $\ell_{\gamma,\pi}$ *norm by*

$$\|v\|_{coop,\gamma,\pi} = \left(\sum_{k=1}^{K} \omega_k \left(\|v_{G_k}^+\|_\pi^\gamma + \|v_{G_k}^-\|_\pi^\gamma \right)^{1/\gamma} \right),$$

where $\omega_k > 0$ *are fixed weights used to adapt the penalty to each group. In particular,*

$$\|v\|_{coop,1,2} = \sum_{k=1}^{K} \omega_k \left(\|v_{G_k}^+\|_2 + \|v_{G_k}^-\|_2 \right) \quad < \quad \|v\|_{coop},$$
$$\|v\|_{coop,1,\infty} = \max_{k=1,..K} \omega_k \max \left(\|v_{G_k}^+\|_2, \|v_{G_k}^-\|_2 \right) \quad < \quad \|v\|_{coop^\star},$$

where $\|\cdot\|_{coop^\star}$ *is the dual norm of* $\|\cdot\|_{coop}$ *that is, for* $v, w \in \mathbb{R}^p$ *with grouping* $\mathcal{G}$, *one has*

$$|v^\top w| \leq \|v\|_{coop} \|w\|_{coop^\star}.$$

With the definition of the cooperative-norms clarified, we turn to the associated optimization problem, in the particular case of linear-regression:

$$\hat{\beta}^{coop} = \arg\min_{\beta \in \mathbb{R}^p} f(\beta, X, y) + \lambda \|\beta\|_{coop}, \quad \text{with } f(\beta, X, y) = \frac{1}{2} \|y - X\beta\|^2 . \quad (3.11)$$

Coop-Lasso sparsity patterns. We now illustrate the sparsity patterns with which this estimator is able to deal with. Problem (3.11) is a combination of a convex and differentiable function and a convex but non differentiable norm. Thus β is a global minimum if and only if $-\nabla f(\beta; X, y)$ belongs to the subdifferential of $\lambda \|\cdot\|_{coop}$ at β:

$$-\nabla f(\beta; X, y) \in \partial \lambda \|\beta\|_{coop}. \quad (3.12)$$

As discussed in Section 3.1, points with singularities have a non-zero probability of being selected as optimal: if singularities are placed at particular points of interest, there is an increased probability for this support to be selected. We illustrate this phenomenon by rephrasing 3.11 in terms of constrained least squares, minimizing the sum of squares under the constraint that $\|\beta\|_{coop} < c$. Under this formulation, β is optimal iif

$$-\nabla f(\beta; X, y) \in N_B(\beta), \quad (3.13)$$

that is to say, the score vector needs to belong to the normal cone to the feasible set $B = \{\beta \in \mathbb{R}^p, \|\beta\|_{coop} \leq c\}$ at the optimum. In more "geometrical" words, Equation (3.13) implies that the solution to 3.11 corresponds to the orthogonal projection of the OLS estimate onto a coop norm ball of a certain radius. Coefficients are set at zero when level curves of the loss hit the ball at singularities, as illustrated in Figure 3.8.

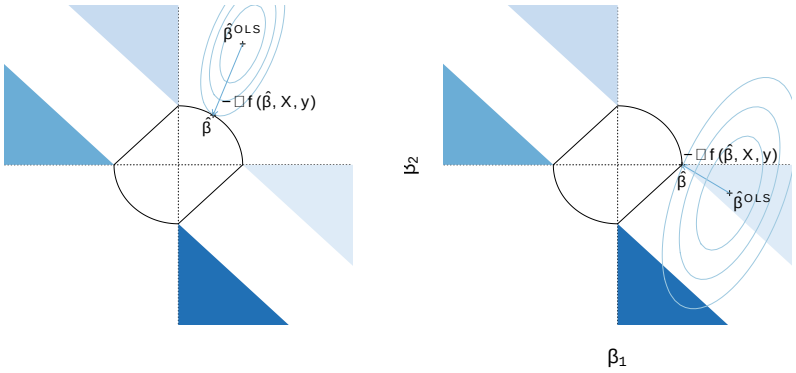


Figure 3.8 – *Projection of the OLS estimate on the coop norm ball of radius s in 2D, with one group of size 2* $\{\beta_1, \beta_2\}$. *On the left panel, projection hits on the* $\mathbb{R}^+ \quad \mathbb{R}^+$ *quadrant: all variables are included. On the right panel, projection hits on the* $\mathbb{R}^+ \quad \mathbb{R}$ *quadrant:* $\hat{\beta}_2$ *can be set at 0. Normal cones* N_B *are represented in color (figures by Camille Charbonnier).*

Now, how does the coop-Lasso compare to its closest cousins, the group- and sparse group-Lasso? Illustrations of the three regularization norms are given in Figure 3.9 for a vector $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^\top$ with two groups $G_1 = \{1, 2\}$ and $G_2 = \{3, 4\}$. Several views of the unit ball are given for each norm, which represents the sets of feasible solutions in the constrained formulation equivalent to (3.11),

First, consider the group-Lasso: the first row illustrates that when β_4 is null its group companion β_3 may also be exactly zero (corners on the boundary at $\beta_3 = 0$); the second row shows that this event is improbable when β_4 differs from zero (smooth boundary at $\beta_3 = 0$). The second and third columns display the same type of relationships within G_1 between β_2 and β_1, due to the symmetries of the unit ball. The last column displays ℓ_2 balls, showing that once a group is activated, so are all its members.

Now, consider the sparse group-norm: the combination of the group and Lasso penalties has uniformly shrunk the feasible set towards the ℓ_1 unit ball, creating new edges that provide a chance to zero any parameter in any situation, with an elastic-net-like penalty within and between groups. The comparison of the last two columns illustrates that the differentiation between the within-group and between-group penalties is less marked than for the group-Lasso.

Finally, consider the coop-norm: compared to the group-norm, there are additional discontinuities resulting in new edges on the 3-D plots. While the sparse group-Lasso edges were created by a uniform shrinkage towards the ℓ_1 ball, the coop-Lasso new edges result from slicing the group-Lasso unit ball, depriving sign-incoherent orthants of some of the group-Lasso feasible solutions ($\|\beta\|_{coop} > \|\beta\|_{group}$ in these regions). In general, there are fewer new edges than with the sparse group-Lasso, since the new opportunities to zero some coefficients are limited to the case where the group-Lasso would have allowed a solution with opposite signs within a group. The crucial difference is the loss of the axial symmetry when some variables are non-zero: decoupling the positive and negative parts of the regression coefficients favors solutions where signs match within a group. Slicing of the unit group-norm ball does not affect the positive and negative orthants, but large areas corresponding to sign mismatches have been peeled off, as best seen on the last column, which also illustrates the strong differentiation between within-group and between-group penalties.

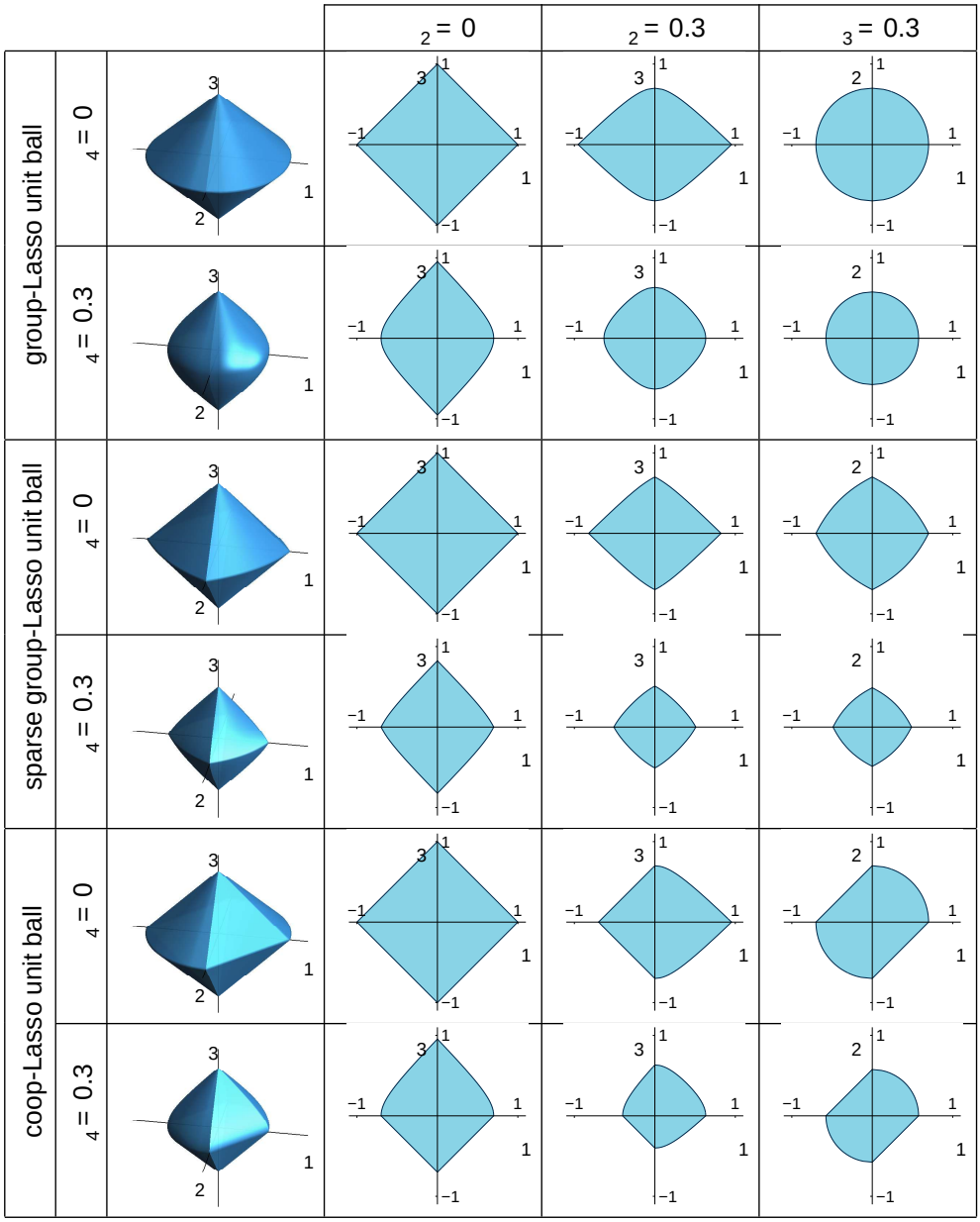


Figure 3.9 –*Feasible sets for the coop-Lasso, group-Lasso and sparse group-Lasso penalties. First column: cuts through* $(\beta_1, \beta_2, \beta_3)$ *at* $\beta_4 = 0$ *and* $\beta_4 = 0.3$*.* (β_1, β_2) *span the horizontal plane and* β_3 *is on the vertical axis; second and third columns: cuts through* (β_1, β_3) *at various values of* (β_2, β_4)*; last column: cuts through* (β_1, β_2) *at various values of* (β_3, β_4)*.*

Optimality conditions. We now state explicitly the optimality conditions for (3.12). They directly result from subdifferential calculus, combined with Equation (3.12) and the definition of the dual coop norm. They are an essential building block of the algorithm we propose to compute the coop-Lasso estimate and provide an important basis for showing the consistency results. Note that we provide the optimality conditions in a different (though equivalent) form than in the original paper [JPV].

Theorem 1 (Optimality conditions for the cooperative-Lasso) *The vector* $\hat{\beta}$ *is optimal for Problem* (3.11) *if and only if* $z = X^T(y - X\hat{\beta})/n\lambda$ *belongs to the subdifferential of the coop-norm associated with groups* $\{G_k\}_{k=1}^K$ *at* $\hat{\beta}$*, characterized by indicator function of the coop-dual norm* $\|\cdot\|_{coop^?}$*. The vector* $\hat{\beta}$ *is optimal if and only if, for every group* G_k*,*

$$\max(\|z^+_{G_k}\|_2, \|z_{G_k}\|_2) \leq 1.$$

In particular,

$$z_j = \varepsilon_j \|\varphi'_j(\hat{\beta}_{G_k})\|^{-1}, \quad \forall j \in G_k \text{ s.t. } \hat{\beta}_j \neq 0,$$

where $\{\varphi'_j\}_{j=1}^p$*, returns the componentwise positive or negative part of a vector according to the sign of its* j*th element, that is,* $\forall k \in \{1,\dots,K\}$, $\forall j \in G_k$, $\forall v \in \mathbb{R}^{p_k}$,

$$\varphi'_j(v) = (\mathrm{sign}(v_j)v)^+. \tag{3.14}$$

Note here an important distinction compared to the group-Lasso, where the optimality conditions are expressed solely according to the groups [JPV]. Indeed, a strong consequence of Theorem 1 is that if both positive and negative coefficients are activated within the same group, then no other coefficients can be shrunk to zero in that group. Hence, while the sparsity pattern of the solution is strongly constrained by the predefined group structure in the group-Lasso, deviations from this structure are possible for the coop-Lasso. The asymptotic analysis that follows confirms that exact support recovery is possible even when the support of $\beta^?$ cannot be expressed as a simple union of groups, provided that the groups intersecting the true support are sign-coherent.

Consistency analysis. We provide two types of results, based upon best achievable results for the Lasso. First, we derive selection properties in an asymptotic linear regression framework, with an irrepresentable condition. Second, we prove estimation and prediction sparsity oracle inequalities, valid non-asymptotically, based upon a restricted eigenvalue assumption. While the former asymptotic results belong to paper [JPV], the latter is unpublished work belonging to Camille Charbonnier's PhD thesis.

Asymptotic properties for support recovery. Here we concentrate on the estimation of the support of $\beta^?$. Our proof technique is drawn from the works on the Lasso [Zha06] and the group-Lasso [Bac08]. In this type of analysis, some assumptions on the joint distribution of (X, Y) are required to guarantee the convergence of empirical covariances. For the sake of simplicity, we keep assuming that data are centered so that we have zero mean random variables and $Q = \mathbb{E}[XX^T]$ is the covariance matrix of X.

(A1) X and Y have finite 4th order moments: $\mathbb{E}\|X\|^4 < \infty$, $\mathbb{E}|Y|^4 < \infty$.

(A2) The covariance matrix $Q = \mathbb{E}[XX^T] \in \mathbb{R}^{p \times p}$ is invertible.

In addition to these standard technical assumptions, we need a more specific one, substantially avoiding situations where the coop-Lasso will almost never recover the true support A. In the sequel, A denotes the true support of β^*, while A_k denotes the intersection between the support and group G_k.

(A3) All sign-incoherent groups are included in the true support: $\forall k \in \{1,\dots,K\}$, if $\|(\beta^*_{G_k})^+\|_k > 0$ and $\|(\beta^*_{G_k})^-\|_k > 0$, then $\forall j \in G_k$, $\beta^*_j \neq 0$.

This latter assumption is less stringent than the one required for the group-Lasso since it does not require that each group of variables be either included in or excluded from the support. For the coop-Lasso, sign-coherent groups may intersect the support.

The suitable variants of the irrepresentable conditions for the coop-Lasso result in two assumptions: a general one, on the magnitude of correlations between relevant and irrelevant variables, and a more specific one for groups which intersect the support, on the sign of correlations. These conditions will be expressed in a compact vectorial form using the diagonal weighting matrix $D(\beta)$ such that,

$$\forall k \in \{1,\dots,K\},\ \forall j \in A_k(\beta),\ (D(\beta))_{jj} = \omega_k \|\beta_{G_k}\|_k^{1-\gamma}. \tag{3.15}$$

(A4) For every group G_k including at least one null coefficient (that is, such that $\beta^*_j = 0$ for some $j \in G_k$ or equivalently $A_k^c \neq \emptyset$), there exists $\eta > 0$ such that

$$\frac{1}{\omega_k}\left\| A_k^c A_A \left(A_A^1 D(\beta^*_A)\right) \beta^*_A \right\|_{\text{coop}^*} \leq 1 - \eta. \tag{3.16}$$

(A5) For every group G_k intersecting the support and including either positive or negative coefficients, let ε_k be the sign of these coefficients ($\varepsilon_k = 1$ if $\|(\beta^*_{G_k})^+\|_k > 0$ and $\varepsilon_k = -1$ if $\|(\beta^*_{G_k})^-\|_k > 0$), the following inequalities should hold:

$$\varepsilon_k A_k^c A_A \left(A_A^1 D(\beta^*_A)\right) \beta^*_A \leq 0, \tag{3.17}$$

where $\leq$ denotes componentwise inequality.

Theorem 2. *If assumptions (A1-5) are satisfied, the coop-Lasso estimator is asymptotically unbiased and has the property of exact support recovery:*

$$\hat{\beta}^{\text{coop}}_n \xrightarrow{P} \beta^* \quad \textit{and} \quad P\left(A(\hat{\beta}^{\text{coop}}_n) = A\right) \to 1, \tag{3.18}$$

for every sequence λ_n such that $\lambda_n = \lambda_0 n^{-\gamma}$, $\gamma \in (0, 1/2)$.

Compared to the group-Lasso, the consistency of support recovery for the coop-Lasso primarily differs regarding possible intersection (besides inclusion and exclusion) between groups and support. This additional flexibility applies to every sign-coherent group. Even if the support is the union of groups, when all groups are sign-coherent, the coop-Lasso still has an edge on group-Lasso since the irrepresentable condition (3.16) is weaker. Indeed, the norm in (3.16) is dominated by the ℓ_2-norm used for the group-Lasso. This difference can have remarkable outcomes, as illustrated on the following numerical example: we generate data from an ordinary regression model with $\beta^* = (1, 1, 1, 1, 0, 0, 0, 0)$, equipped with the group structure

$\{G_k\}_{k=1}^4 = \{\{1,2\},\{3,4\},\{5,6\},\{7,8\}\}$. The vector X is generated as a centered Gaussian random vector with a covariance matrix chosen so that the irrepresentable conditions hold for the coop-Lasso, but not for group-Lasso, which, we recall, are more demanding for the current situation, with sign-coherent groups. The random error follows a centered Gaussian distribution with standard deviation $= 0.1$, inducing a very high signal to noise ratio ($R^2 = 0.99$ on average), so that asymptotics provide a realistic view of the finite sample situation. We generated 1000 samples of size $n = 20$ from the described model, computed the corresponding 1000 regularization paths for the group-Lasso, sparse group-Lasso, and coop-Lasso. Figure 3.10 reports the 50% coverage intervals (lower and upper quartiles) along the regularization paths. In this setup, the sparse group-Lasso behaves like the group-Lasso, leading to nearly identical graphs.

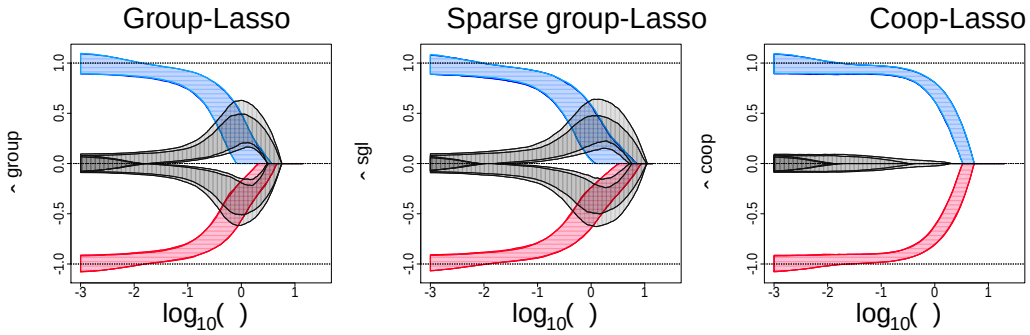


Figure 3.10 – *50% coverage intervals for the group (left), sparse group (center), and (right) Lasso estimated coefficients along regularization paths: coefficients from the support of $\beta^?$ are marked by colored horizontal stripes and the other ones by gray vertical stripes.*

Estimation is difficult in this small sample problem: the two versions of the group-Lasso, which first select the wrong covariates, never reach the situation where they would have a decisive advantage over OLS, while the coop-Lasso immediately selects the right covariates, the coefficients of which steadily dominate the irrelevant ones.

Remark. In order to adapt Theorem 2 to the high-dimensional setting, one would need to add technical assumptions guaranteeing the existence of concentration inequalities, however assumptions required on the design to obtain exact support recovery would remain the same as in (A4) and (A5). Since the latter are the only assumptions that would differ between [131] and the coop-Lasso, there is no major interest in rewriting Theorem 2 according to those new developments.

Non-asymptotic properties for estimation and prediction. We now derive non-asymptotic oracle inequalities based upon restricted eigenvalue assumptions, which is roughly the same as for the Lasso (3.10), except that the cone on which the assumption relies is defined by the cooperative-norm, and the sparsity considered is a group sparsity.

Assumption 3.1 (Restricted eigenvalue) *There exists* $\kappa(s) > 0$ *such that:*

$$\min \left\{ \frac{\|Xu\|_2}{\sqrt{n}\|u_A\|_2} : |A| \leq s, u \in \mathbb{R}^p, \|u_{A^c}\|_{coop} \leq 3\|u_A\|_{coop} \right\} \geq \kappa(s).$$

In order to explicit the bounds and order of probability in our oracle inequalities, we restrict ourselves to the case where all groups share the same size m.

Theorem 3 (Oracle inequalities with groups of equal sizes). *Under Assumption 3.1 and considering that the data matrix has been scaled so that all diagonal elements of* $X^{\ddagger}X/n$ *are equal to 1, for a choice of* λ *equal to* $A\sigma\frac{\sqrt{m}+\sqrt{\log K}}{\sqrt{n}}$, *then with probability larger than* $1-2/K^2$, *any solution to Problem 3.11 with groups of equal sizes* m *satisfies the following prediction and estimation oracle inequalities:*

$$\|X(\beta^\star - \hat{\beta}^{coop})\|_n^2 \leq \frac{32}{(s^-)^2}\frac{s^+(m+\log K+2\sqrt{m\log K})}{n}\lambda^2, \tag{3.19}$$

$$\|\beta^\star - \hat{\beta}^{coop}\|_{coop} \leq \frac{32}{(s^-)^2}\frac{s^+(\sqrt{m}+\sqrt{\log K})}{\sqrt{n}}\lambda, \tag{3.20}$$

$$\|\beta^\star - \hat{\beta}^{coop}\|_2 \leq \frac{32}{(s^-)^2}\frac{s^+(\sqrt{m}+\sqrt{\log K})}{\sqrt{n}}\lambda. \tag{3.21}$$

Remark. We need to restrict ourselves to groups of equal size because the upper bound on the probability of the event $\lambda_n \geq 2\|X^{\ddagger}\varepsilon\|_\infty/n$ relies on tail bounds of the maximum of K χ^2 distributions. If all groups share the same size, we can use a union bound on the tails of K independent χ^2 with similar degrees of freedom. Otherwise, each of the K χ^2 distributions has its own degree of freedom, which makes it impossible to bound explicitly the probability of the intersection, unless we use a very raw upper bound.

Besides, there is no improvement compared to the group-Lasso oracle inequalities because we cannot exploit the advantages of the cooperative-norm on two fronts: first, the probability of event A uses an upper bound of the dual cooperative-norm by the dual group-norm, because the dual coop-norm leads to 2^p distributions of unknown degrees of freedom which we cannot control explicitly; second, what appears is actually a rate of $2s$, twice the group-sparsity: we cannot count the number of activated signed-groups instead. Indeed, following the terms of [127] the cooperative-norm is only decomposable to group-sparse subsets, not to signed quadrants: we can write $\|\cdot\|_{coop} = \|\cdot\|_{coop} + \|\cdot\|_{coop}$ for every $\in M$ and $\in M^\perp$ for subsets $M = \{x \in \mathbb{R}^p, \forall k \in S^c, x_G = 0\}$ defined by the activation of a subset of groups, but not by the activation of a subset of *signed* subgroups. Recent developments [113] could help improve the results, by allowing us to work with group-specific penalties λ_j.

Application to monotonicity of responses to ordinal covariates. This section illustrates the applicability of the coop-Lasso on categorical and continuous covariates, which may be widely applied to ordered categorical variables.[6] When not treated as numerical, ordinal variables are often coded by a set of variables that code differences between levels. Several types of coding have been developed in the ANOVA setting, with relatively little impact in the regression setting, where the so-called dummy coding is intensively used. Indeed, least squares fits are not sensitive to coding choices provided there is a one-to-one mapping from one to the other, so that coding only matters regarding the direct interpretation of regression coefficients. However, coding evidently affects the solution in penalized regression, and we will here use specific coding to penalize targeted variations. In order to build a monotonicity-based penalty, we

[6] An application specific to microarray data appears in the original paper but omitted here, as our point is to demonstrate the versatility of regularized methods to a wide class of problems beyond genomics.

simply use contrasts that compare two adjacent levels. An example of these contrasts is displayed in Table 3.1, with the corresponding coding, known as backward difference coding, which is simply obtained by solving a linear system [152]. Irrespective of the

Table 3.1 – *Contrasts and coding for comparing the adjacent levels of a covariate with 4 levels.*

Level	Contrasts			Codings		
0	-1	0	0	-3/4	-1/2	-1/4
1	1	-1	0	1/4	-1/2	-1/4
2	0	1	-1	1/4	1/2	-1/4
3	0	0	1	1/4	1/2	3/4

coding, group penalties act as a selection tool for factors, at variable level [106]. On top of this, the sparse group penalty presents the ability to discard a level. With difference coding, some increments between adjacent levels may be set at zero, that is, levels may be fused [61]. With the coop-Lasso, increments are urged to be sign-coherent, thereby favoring monotonicity. As a side effect, level fusion may also be obtained.

Experimental setup. We illustrate the approach on the Statlog "German Credit" data set (available at the UCI repository) which gathers information about people classified as low or high credit risks. This binary response requires an appropriate model, such as logistic regression. We adapt (3.11) and the accompanying optimization methods in this perspective. All quantitative variables are used for the analysis, but we focus here on the regression coefficients of four variables, encoded as integers or nominal in the Statlog project, which seem better interpreted as ordered nominal, namely: history , with 4 levels describing the ability to pay back credit in the past and now; savings , with 4 levels giving the balance of the savings account in currency intervals; employment, with 5 levels reporting the duration of the present employment in year intervals; and job , with 4 levels representing an employment qualification scale.

Results. The performance of the three methods are identical, either in terms of deviance or classification error and omitted here. The regression coefficients differ however, as shown in Figure 3.11, displaying the paths of sparse group- and coop-Lasso. Recall that we only represent the ordinal covariates history , savings , employment and job . Each coefficient represents the increment between two adjacent levels, with positive and negative values resulting in an increase and decrease respectively. Monotonicity with respect to all levels is reached if all the values corresponding to a factor are nonnegative or nonpositive. We also provide an alternative view of the coop-Lasso path, with the overall effects corresponding to levels, obtained by summing up the increments.

The solutions differ regarding monotonicity, which is never observed along the group-Lasso regularization path (not shown here). The sparse group-Lasso paths have long sign-coherent sections. These sections extend further with the coop-Lasso. The sparse group and the coop-Lasso set some increments to zero, leading to the fusion of adjacent levels that should be welcomed regarding interpretation. The solutions tend to agree on these fusions on long sections of the paths, with some additional fusions of the sparse group-Lasso when slight monotonic solutions are provided by the coop-Lasso (see employment, levels 2 and 3, and savings , levels 1 and 2). These fusions are perceived more directly on the coop-Lasso path of effects, displayed in the bottom right of Figure 3.11, where the effect of each level is displayed directly.

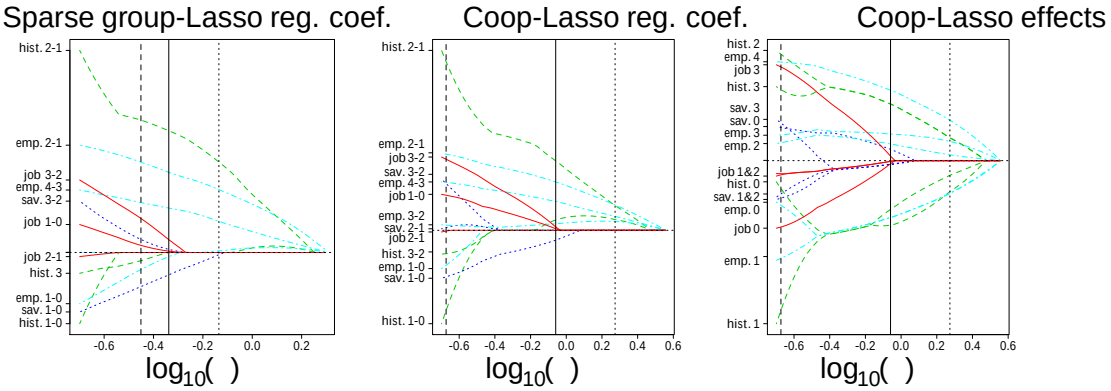


Figure 3.11 – *Regularization paths for 4 ordinal covariates (history, savings, job and employment) for the coop and sparse group Lasso on the contrast coefficients obtained from backward difference coding. The transcription of contrasts to levels is also displayed for the coop-Lasso. The vertical lines mark the model selected by CV on the validation set, for different criteria: deviance (plain), misclassification rate (dashed), and weighted misclassification error (dotted).*

Some final comments on the cooperative-Lasso. We did not detail the implementation here, as it roughly follows the active-set strategy 1 of Section 3.1.2, where optimality conditions a group activation are handled with the dual coop-norm. In the R-package scoop [SW4], the cooperative Lasso is implemented for linear and logistic regressions, when the structure $\mathcal{G}$ defines either a partition or a hierarchy.

In the full-length paper [JPV1], a thorough simulation study is conducted comparing the performance of the coop-Lasso to that of the competing methods. We derive an estimator of the degrees of freedom for a coop-Lasso fit for linear regression, which paves the way for using penalized criteria for model selection. An application in genomics for sign-coherent groups of genes is thoroughly explored, and I supervised two MSc. in this direction. I am also currently working on the hierarchical version of the coop-Lasso in a sparse biological network inference perspective, when the predictors are spread in groups of co-activators or co-inhibitors regulating other genes [JPS3].

Finally, the cooperative-Lasso has been used by other researchers [LS] for enforcing monotonicity in regression with splines.

3.2.2 Structured regularization for conditional GGM

This contribution is at the crossroads of Chapters 2 and 3: first, it entails tools from GGM in order to describe direct relationships between predictors and responses in multivariate regression, in an effort to propose more interpretable models. Second, we rely on a sparse method where variable selection is driven by structural information thanks to a graph Laplacian penalty suited to the multivariate framework. At the end of the day, our structured sparse estimator of the regression coefficients is able to discriminate coefficients having direct effects on the responses from those being induced by spurious correlation between the responses themselves.

Model setup. Compared to its univariate counterpart, the multivariate linear regression model aims to predict some q responses from a set of p predictors, relying on a training data set $(x_i, y_i)_{i=1,\dots,n}$:

$$y_i = B^T x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, R), \quad \forall i = 1, \dots, n. \tag{3.22}$$

The $p \times q$ matrix of regression coefficients B and the $q \times q$ covariance matrix R of the Gaussian noise ε_i are unknown. Model (3.22) has been studied in [116] in the low dimensional case where both ordinary and generalized least squares estimators of B coincide and do not depend on R. These approaches boil down to performing q independent regressions, each column B_j describing the weights associating the p predictors with the jth response. In the $n < p$ setup however, these estimators are not defined and we need to regularize the problem in some way. To this end, we aim a general multivariate regression framework with three purposes: i) We want to account for the dependency structure between the outputs, if it exists. That is to say, we want to integrate the estimation of R in the inference process. ii) We want to have the possibility of integrating some prior information about the predictors. iii) In order to improve interpretability, we pay prior attention to the direct links between predictors and responses. We reach these three goals by relying on *conditional* Gaussian graphical models (cGGM), a recent proposal that has emerged in the literature [158, 193]. It extends to the multivariate case the links existing between the linear model, partial correlations and GGM, existing between the linear model, partial correlations and GGM, as depicted for instance by [122], [136] and then [183]. We then propose a multivariate regularization scheme for this model that draws inspiration from existing works for penalized multivariate regression with known [94, 95] and unknown [148, 189, 104] residual variance R. Our sparse Laplacian penalty draw inspiration from [139, 106, 156, 72, 111].

Conditional Gaussian graphical model. The statistical framework of cGGM arises from a different parametrization of (3.22). To the best of our knowledge, this was first underlined by [158]. It amounts to investigating the joint probability distribution of (x_i, y_i) in the Gaussian case, with the following block-wise decomposition of the covariance matrix Σ and its inverse $\Lambda = \Sigma^{-1}$:

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{pmatrix}. \tag{3.23}$$

Back to the distribution of y_i conditional on x_i, multivariate Gaussian analysis shows that, for centered x_i and y_i, $y_i | x_i \sim \mathcal{N}\left(-\Lambda_{yy}^{-1}\Lambda_{yx}x_i, \Lambda_{yy}^{-1}\right)$. This model, associated with the full sample $\{(x_i, y_i)\}_{i=1,\dots,n}$, can be written in a matrix form by stacking in rows first the observations of the responses, and then the observations of the predictors, in two data matrices Y and X with respective sizes $n \times q$ and $n \times p$:

$$Y = X\,\Sigma_{xy}\Sigma_{yy}^{-1} + \varepsilon, \quad \mathrm{vec}(\varepsilon) \sim \mathcal{N}\left(0_{nq}, I_n \otimes \Lambda_{yy}^{-1}\right), \tag{3.24}$$

where $\mathrm{vec}(A) = (A_1^T \dots A_p^T)^T$. Introducing the empirical matrices of covariance $S_{yy} = n^{-1}\sum_{i=1}^n y_i y_i^T$, $S_{xx} = n^{-1}\sum_{i=1}^n x_i x_i^T$, and $S_{yx} = n^{-1}\sum_{i=1}^n y_i x_i^T$, the log-likelihood of (3.24) – a conditional likelihood regarding (3.23) – is written

$$\begin{aligned} \frac{2}{n}\log L(\Sigma_{xy}, \Lambda_{yy}) = &-\log|\Lambda_{yy}| + \mathrm{Tr}\left(S_{yy}\Lambda_{yy}\right) \\ &+ 2\,\mathrm{Tr}\left(S_{xy}\Lambda_{yx}\right) + \mathrm{Tr}\left(\Lambda_{yx}S_{xx}\Sigma_{xy}\Lambda_{yy}^{-1}\right) + \mathrm{cst}. \end{aligned} \tag{3.25}$$

We notice by comparing the cGGM (3.24) to the multivariate regression model (3.22) that $\Lambda_{yy}^{-1} = R$ and $B = \Sigma_{xy}\Lambda_{yy}^{-1}$. This alternative parametrization shows two important differences. First, the negative log-likelihood (3.25) is jointly convex (in $\Sigma_{xy}, \Lambda_{yy}$)

(see [193]). Minimization problems involving (3.25) are thus amenable to a global solution, which facilitates both optimization and theoretical analysis. Second, it unveils new interpretations for the relationships between input and output variables: Θ_{xy} describes the *direct* relationships between predictors and responses, the support of which we are seeking in order to select relevant interactions. On the other hand, $\mathbf{B}$ entails *both direct and indirect* influences, possibly due to some strong correlations between the responses, described by the covariance matrix $\mathbf{R}$ (or equivalently its inverse Θ_{yy}). To provide insights on cGGM, Figure 3.12 illustrates the relationships between $\mathbf{B}$, Θ_{xy} and $\mathbf{R}$ in two simple scenarios where $p = 40$ and $q = 5$. Scenarios *a*) and *b*) are discriminated by the presence of a strong structure among the predictors. The important point to grasp at this stage is how strong correlations between outcomes can completely "mask" the direct links in the regression coefficients: the stronger the correlation in $\mathbf{R}$, the less possible it is to distinguish in $\mathbf{B}$ the non-zero coefficients of Θ_{xy}.

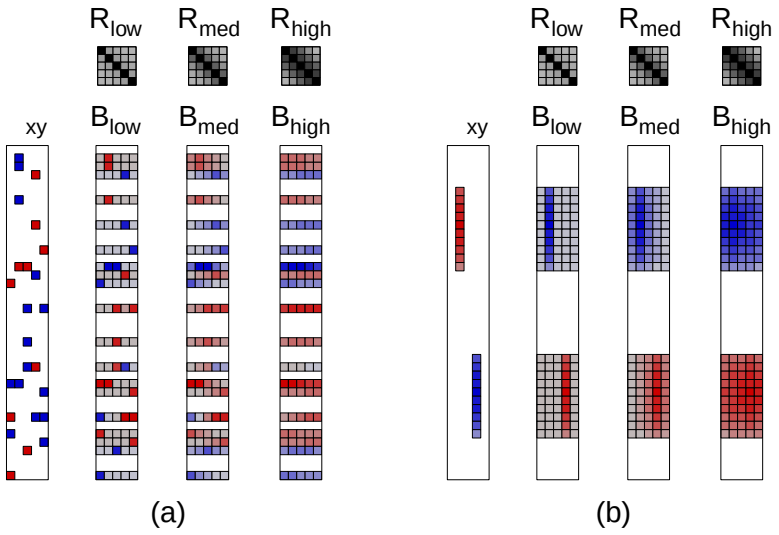


Figure 3.12 –*Toy examples to illustrate the relationships between* $\mathbf{B}$, Θ_{xy} *and* $\mathbf{R}$*: on panel a*), *a situation with no structure among the predictors; on panel b*), *a strong neighborhood structure. For each panel, we see the effect of stronger correlations in* $\mathbf{R}$ *on masking the direct links in* $\mathbf{B}$.

Structured regularization with underlying sparsity. Our regularization scheme starts by considering some structural prior information about the relationships between the coefficients. We are typically thinking of a situation where similar inputs are expected to have similar direct relationships with the outputs. The right panel of Figure 3.12 represents an illustration of such a situation, where there exists an extreme neighborhood structure between the predictors. This depicts a pattern that acts along the rows of $\mathbf{B}$ or Θ_{xy} as substantiated by the following Bayesian point of view.

Bayesian interpretation. Suppose that the similarities can be encoded into a matrix $\mathbf{L}$. Our aim is to account for this information when learning the coefficients. The Bayesian framework provides a convenient setup for defining how structural information should be accounted for. In the single output case [118], a conjugate prior for β would be $\mathcal{N}(0, \mathbf{L}^{-1})$. Combined with the covariance between the outputs, this gives

$$\mathrm{vec}(\mathbf{B}) \sim \mathcal{N}(0, \mathbf{R} \otimes \mathbf{L}^{-1}),$$

where $\otimes$ is the Kronecker product. By properties of the vec operator, it is stated as

$$\mathrm{vec}(\Theta_{xy}) \sim \mathcal{N}(0, R^{-1} \otimes L^{-1})$$

for the direct links. Choosing such a prior results in

$$-\log P(\Theta_{xy} | L, R) = \frac{1}{2} \mathrm{Tr}\, \Theta_{xy}^T L \Theta_{xy} R + \text{cst}.$$

Criterion. Through this argument, we propose a criterion with two terms to regularize the conditional negative log-likelihood (3.25): first, a smooth trace term relying on the available structural information L; and second, an ℓ_1 norm that encourages sparsity among the direct links. We write the criterion as a function of $(\Theta_{xy}, \Theta_{yy})$ rather than (Θ_{xy}, R), although equivalent in terms of estimation, since the former leads to a convex formulation. The optimization problem turns to the minimization of

$$J(\Theta_{xy}, \Theta_{yy}) = \frac{1}{n} \log L(\Theta_{xy}, \Theta_{yy}) + \frac{\lambda_2}{2} \mathrm{Tr}\, \Theta_{yx} L \Theta_{xy} \Theta_{yy}^{-1} + \lambda_1 \|\Theta_{xy}\|_1.$$

Optimization. In the classical framework of parametrization (3.22), alternate strategies where optimization is successively performed over Θ_{xy} and Θ_{yy} have been proposed [148, 189]. These strategies come with no guarantee of convergence to the global optimum since the objective is only bi-convex. In the cGGM framework [158, 193], the optimized criterion is jointly convex yet no convergence result is provided regarding the optimization procedure proposed by the authors. Here we also consider the alternate strategy for which theoretical guarantees are provided by the following theorem:

Theorem 4. *Let $n \geq q$. Criterion* (3.2.2) *is jointly convex in* $(\Theta_{xy}, \Theta_{yy})$. *Moreover, the alternate optimization*

$$\hat{\Theta}_{yy}^{(k+1)} = \arg\min_{\Theta_{yy} \succ 0} J_{\lambda_1 \lambda_2}(\hat{\Theta}_{xy}^{(k)}, \Theta_{yy}), \tag{3.26a}$$

$$\hat{\Theta}_{xy}^{(k+1)} = \arg\min_{\Theta_{xy}} J_{\lambda_1 \lambda_2}(\Theta_{xy}, \hat{\Theta}_{yy}^{(k+1)}). \tag{3.26b}$$

leads to the optimal solution.

We skip the proof, which relies on the fact that efficient procedures exist to solve the two sub-problems (3.26a) and (3.26b): we derive an analytic form for the former, requiring a single SVD. The latter problem can be recast to a generalized Elastic-net problem. Details are given in the original version of the paper [PP2].

Because our procedure relies on alternating optimization, it is difficult to give either a global rate of convergence or a complexity bound. Nevertheless, the complexity of each iteration is easy to derive, since it amounts to two well-known problems: the main computational cost in (3.26a) is due to the SVD of a $q \times q$ matrix, which costs $O(q^3)$. Concerning (3.26b), it amounts to the resolution of an Elastic-net problem with p variables and $n \geq q$ samples. If the final number of nonzero entries in $\hat{\Theta}_{xy}$ is k, a good implementation with Cholesky update/downdate is roughly in $O(npq^2k)$ (see, e.g. [5]). Since we typically assume that $p \geq n \geq q$, the global cost of a single iteration of the alternating scheme is thus $O(npq^2k)$, and we can theoretically treat problems with large p when k remains moderate.

Model selection. When looking for the best model in terms of variable selection, penalized criteria provide a credible alternative to cross-validation. A general form is expressed as a function of the likelihood (3.25) and the effective degrees of freedom:

$$-2\log L(\hat{\Theta}_{xy}^{\lambda_1,\lambda_2}, \hat{\Theta}_{yy}^{\lambda_1,\lambda_2}) + \text{pen}\ \widehat{\text{df}}_{\lambda_1,\lambda_2}. \tag{3.27}$$

Setting pen$= 2$ or $\log(n)$ leads to AIC or BIC. For the practical evaluation of (3.27), we use the definition of [147] for the degrees of freedom and rely on the work of [168] to derive the following Proposition, the proof of which is detailed in our paper [PPP2].

Proposition. *An unbiased estimator of* df_{λ_1,λ_2} *for our fitting procedure is*

$$\widehat{\text{df}}_{\lambda_1,\lambda_2} = \text{card}(A) - \lambda_2 \text{tr}\left((\hat{R} \otimes L)_{AA}\,(\hat{R} \otimes (S_{xx} + \lambda_2 L))_{AA}^{-1}\right),$$

where $A = \{j : \text{vec}\,\hat{\Theta}_{xy}^{\lambda_1,\lambda_2} \neq 0\}$ *is the set of nonzero entries in* $\hat{\Theta}_{xy}^{\lambda_1,\lambda_2}$.

Application Studies. In this section the flexibility of our proposal is illustrated by investigating two multivariate problems in genetics and genomics corresponding to Examples 1.3 and 1.4. We insist on the construction of the structuring matrix L.

Multi-trait Genomic Selection in Brassica napus. Genomic selection is aimed at predicting one or several phenotypes based on the information of genetic markers (see Example 1.3, Chapter One). To this end, regularization methods such as ridge or Lasso regression or their Bayesian counterparts have been proposed [46]. Still, in most studies only single trait genomic selection is performed, neglecting correlations between phenotypes. Moreover, little attention has been devoted to the development of regularization methods including prior genetic knowledge. We consider the *Brassica napus* dataset described in [52] and [98]. Data consists in $n = 103$ double-haploid lines derived from 2 parent cultivars, "Stellar" and "Major", on which $p = 300$ genetic markers and $q = 8$ traits (responses) were recorded. Each marker is a covariate with $x_i^j = 0$ if line i has the Stellar allele at marker j, and $x_i^j = 1$ otherwise. Traits included are percent winter survival for 1992, 1993, 1994, 1997 and 1999 (surv92, surv93, surv94, surv97, surv99, respectively), and days to flowering after no vernalization (flower0), 4 weeks vernalization (flower4) or 8 weeks vernalization (flower8).

Structure specification. In a biparental line population, correlation between 2 markers depends on their genetic distance defined in terms of recombination fraction. As a consequence, one expects adjacent markers on the sequence to be correlated, yielding similar direct relationships with the phenotypic traits. Noting d_{12} the genetic distance between markers M_1 and M_2, one has $\text{cor}(M_1, M_2) = \rho^{d_{12}}$, where $\rho = .98$[7]. The covariance matrix L^{-1} can hence be defined as $L_{ij}^{-1} = \rho^{d_{ij}}$. Moreover, assuming recombination events are independent between M_1 and M_2 on the one hand, and M_2 and M_3 on the other hand, one has $d_{13} = d_{12} + d_{23}$ and matrix L^{-1} exhibits an inhomogeneous AR(1) profile. As a consequence, L is tridiagonal with general elements

$$w_{i,i} = \frac{1 - \rho^{2d_{i-1,i}+2d_{i,i+1}}}{(1-\rho^{2d_{i-1,i}})(1-\rho^{2d_{i,i+1}})}, \qquad w_{i,i+1} = \frac{-\rho^{d_{i,i+1}}}{1-\rho^{2d_{i,i+1}}}, \qquad w_{i,j} = 0 \text{ if } |i-j| > 1.$$

[7] This value directly arises from the definition of the genetic distance itself.

For the first (resp. last) marker, the distance $d_{1,i}$ (resp. $d_{i,i+1}$) is infinite.

Results. In the full-length paper [PP2], predictive performance is estimated by repetitive splits of the data into test and training sets, in which case SPRING provides the smallest error for half of the traits compared with state-of-the-art penalized multivariate regression procedures. Here, we want to insist on model parameter interpretation. Hence, a picture of the between-response covariance matrix estimated with SPRING is given in Figure 3.13. It reflects the correlation between the traits, which are either explained by an unexplored part of the genotype, by the environment or by some interaction between the two. The residuals of the flowering times exhibit strong correlations, whereas correlations between the survival rates are weak. It also shows that the survival traits have a larger residual variability than do the flowering traits, suggesting a higher sensitivity to environmental conditions. We then turn to the effects of each marker on

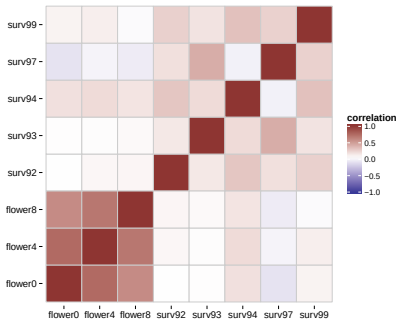


Figure 3.13 – *Brassica study: residual covariance estimation*

the different traits. The left panels of Figure 3.14 give both the regression coefficients (top) and the direct effects (bottom). The gray zones correspond to chromosomes 2, 8 and 10, respectively. The exact location of the markers within these chromosomes is displayed in the right panels, where the size of the dots reflects the absolute value of the regression coefficients (top) and of the direct effects (bottom). The interest of consid-

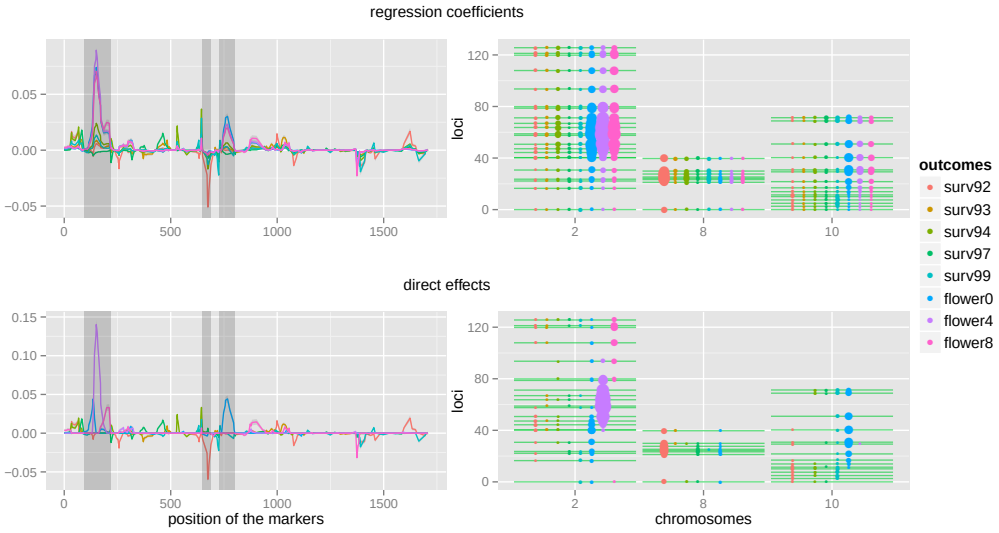


Figure 3.14 – *Brassica Study: direct and indirect genetic effects of the markers on the traits estimated by SPRING (better seen in color).*

ering direct effects rather than regression coefficients appears clearly here, if one looks for example at chromosome 2. Three large overlapping regions are observed in the coefficient plot, for each flowering trait. A straightforward interpretation would suggest that the corresponding region controls the general flowering process. The direct effect plot allows one to go deeper and shows that these three responses are actually controlled by three separate sub-regions within this chromosome. The confusion in the coefficient plot only results from the strong correlations observed among the three flowering traits.

Selecting regulatory motifs from multiple microarrays. We are interested in the detection of regulatory motifs (see Example 1.4), the presence of which controls the gene expression profile. To this aim we try to establish a relationship between the expression level of all genes across a series of conditions with the content of their respective regulatory regions in terms of motifs. In this context, we expect (i) the set of influential motifs to be small for each condition, (ii) the influential motifs for a given condition to be degenerate versions of each other, and (iii) the expression under similar conditions to be controlled by the same motifs.

We rely on the series of microarray experiments conducted on yeast cells (*Saccharomyces cerevisae*) in [60]. Among these assays, we consider 12 time-course experiments profiling $n = 5883$ genes under various environmental changes as listed in Table 3.2. These expression sets form 12 potential response matrices Y, the column number of which corresponds to the number of time points.

Experiment	*# time point*	*# motifs selected*		
		$k = 7$	$k = 8$	$k = 9$
Heat shock	8	30	68	43
Shift from 37°C to 25°C	5	3	11	33
Mild heat shock	4	24	13	23
Response to H_2O_2	10	15	10	21
Menadione exposure	9	16	1	7
DDT exposure 1	8	15	10	30
DDT exposure 2	7	11	33	21
Diamide treatment	8	45	25	35
Hyperosmotic shock	7	36	24	15
Hypo-osmotic shock	5	20	8	29
Amino-acid starvation	5	47	30	39
Diauxic shift	7	16	14	20
total number of unique motifs inferred		87	82	72

Table 3.2 –*Time-course data from* [60] *considered for regulatory motif discovery*

For the predictors, we consider all motifs with length k formed with the four nucleotides, that is $\mathcal{M}_k = \{A, C, G, T\}^k$. There are $p = |\mathcal{M}_k| = 4^k$ such motifs. Unless otherwise stated, the motifs in $\mathcal{M}$ are lined up in lexicographical order, *i.e.* when $k = 2$, $AA, AC, AG, AT, CA, CC, \ldots$ and so on. The $n \times p$ matrix of predictors X is filled such that X_{ij} equals the occurrence count of motif j in the regulatory region of gene i.

Structure specification. As we expect influential motifs for a given condition to be de-

generate versions of each other, we first measure the similarity between any two motifs from $\mathcal{M}_k$ with the Hamming defined as $\text{dist}(a, b) = \text{card}\{i : a_i \neq b_i\}$, for all $a, b \in \mathcal{M}_k$. For a fixed value of interest $0 \leq \ell \leq k$, we define the *ℓ-distance matrix*

$$\mathbf{D}^{k,\ell} = (d^{k,\ell}_{ab})_{a,b \in \mathcal{M}_k}, \quad d^{k,\ell}_{ab} = \begin{cases} 1 & \text{if } \text{dist}(a, b) \leq \ell \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbf{D}^{k,\ell}$ can be viewed as the adjacency matrix of a graph where the nodes are the motifs and where an edge is present when the 2 motifs are at a Hamming distance less or equal to ℓ. We finally use the Laplacian of this graph in place of the structuring matrix $\mathbf{L}$.

Results. We apply our methodology for candidate motifs from $\mathcal{M}_7$, $\mathcal{M}_8$ and $\mathcal{M}_9$, which results in three lists of putative motifs having a direct effect on gene expression. Due to the very large number of potential predictors that comes with a sparser matrix $\mathbf{X}$ when k increases, we first perform a screening step that keeps the 5,000 motifs with the highest marginal correlations with $\mathbf{Y}$. Second, SPRING is applied to each of the twelve time-course experiments described in Table 3.2. The selection of (γ_1, γ_2) is performed on a grid using the BIC (3.27). At the end of the day, the three lists corresponding to $k = 7, 8, 9$ include respectively 87, 82 and 72 motifs, for which at least one coefficient in the associated row $\beta_{xy}(j, \cdot)$ was found non-null for some of the twelve experiments.

To assess the relevance of the selected motifs, we compared them with the MotifDB patterns available in Bioconductor [153], where known transcription factor binding sites are recorded. There are 453 such reference motifs with size varying from 5 to 23 nucleotides. Consider the case of $k = 7$ for instance: among the 87 SPRING motifs, 62 match one MotifDB pattern each and 25 are clustered into 11 MotifDB patterns as depicted in Table 3.3.

CTAAGCCAC	GCATGTGAA	CATGTAATT	TGAAACA
TAGCCCC	CCATATG	TGTAAAT	TTAGACC
GCGCCCC	TTGTGAG	TGTATAT	TAAAAAG
TGATCGGCGCCGCACGACTGA	GACTGGTT	GATCGTATGATA	ACGCGAAAA
GTATAAC	GCTGGTT	ATCATAT	AACGAAA
GCGCCGT	GCTGGTG	TTGGTAT	ACGAAAA
CCATACATCAC	ATTGACCTGGTC	GACTAGATATATATATTCGAT	
CATAGAC	TCGACTT	ATATATT	
ATATCAC	CGACTTG	CATATAT	
	CCAGCTT	ATATATG	
		ATATATA	

Table 3.3 –*Comparison of SPRING-selected motifs with* MotifDB *patterns. Each cell corresponds to a* MotifDB *pattern (top) compared to a set of aligned SPRING motifs with size 7 (down).*

As seen in this table, the clusters of motifs selected correspond to sets of variants of the same pattern. In this example, the ability of SPRING to use domain-specific definitions of the structure between the predictors oriented the regression problem to account for motif degeneracy and helped in selecting motifs that are consistent, known binding sites.

3.2.3 A quadratic view of sparsity

In this work, we propose a unifying view of a wide family of regularized linear regression problems that includes the Lasso, the (generalized) Elastic-net, the ℓ_1-ℓ_∞ group-lasso and the fused Lasso signal approximator, among others. The main idea is to represent the feasible set associated with these regularization schemes as the intersection of simple quadratic sets. On top of providing a new viewpoint on the aforementioned penalization methods, our approach results in a unified optimization strategy. A description of a general-purpose active-set algorithm derived from this formalism is given. We also demonstrate that our solver is highly efficient compared to existing algorithms and popular implementations.

Rationale. We consider the usual linear model where the response variable Y is related to the p predictor variables $X = (X_1, \ldots, X_p)$: $Y = X^T \beta^\star + \varepsilon$, where $\beta^\star$ is a *sparse* vector of unknown parameter, and ε is a perturbation variable. The estimation of $\beta^\star$ is based on training data consisting in an $n \times p$ matrix $\mathbf{X}$ and a response vector $\mathbf{y} \in \mathbb{R}^n$.

Our rationale follows a general robust regularized regression approach:

$$\min_{\beta \in \mathbb{R}^p} \max_{\Delta \in \mathcal{D}} \|\mathbf{y} - (\mathbf{X} - \Delta)\beta\|_2^2 + \lambda \|\beta\|_2^2 , \tag{3.28}$$

where $\mathcal{D}$ describes an uncertainty set for the regression parameters and Δ acts as a spurious *adversary* over the true $\beta^\star$. This problem could be fully motivated in a robust regression framework (just as we showed in [PP3]), where the spurious vector of coefficients Δ is interpreted as an adversarial noise. Here, we rely instead on a simpler – though equivalent – geometrical point of view. By considering sound definitions of the apposite uncertainty set $\mathcal{D}$, we use Problem (3.28) as a starting point to recover known sparse penalties. Such an approach provides an original interpretation of these existing penalties and may also inspire new ones specifically tailored for a given purpose. Indeed, it is sometimes easier to formalize spurious effects through the definition of $\mathcal{D}$, than beneficial ones through the direct definition of a penalty on β.

Assumptions on the spurious regression coefficients. We now proceed by proposing several options for $\mathcal{D}$, each one entailing an equivalence with a well-known sparse regression method. All our examples follow the same pattern: assuming a given regularity on the regression coefficients $\beta^\star$, we consider the adversarial dual assumption on the spurious coefficients. When the initial regularity conditions on $\beta^\star$ are expressed by ℓ_1 or ℓ_∞ norms, this process results in uncertainty sets $\mathcal{D}$ which are very easy to manage when solving Problem (3.28) since they can be defined as the convex hulls of a finite number of possible perturbations.

Elastic-net. As a first option, let us consider the regularity assumption stating that the ℓ_1-norm of $\beta^\star$ should be small:

$$\mathcal{H}_{\beta^\star}^{\text{Lasso}} = \left\{ \beta \in \mathbb{R}^p : \|\beta - \beta^\star\|_1 \leq \beta^{\star T} \beta^\star \right\}.$$

The dual assumption is that the ℓ_1-norm of Δ should be controlled, say:

$$\mathcal{D}^{\ell_1} = \left\{ \Delta \in \mathbb{R}^p : \sup_{\beta \in \mathcal{H}_{\beta^\star}^{\ell_1}} \Delta^T \beta \leq 1 \right\} = \left\{ \Delta \in \mathbb{R}^p : \|\Delta - \beta^\star\|_1 \leq \beta^{\star T} \beta^\star \right\} = \text{conv}\left\{ \beta^\star, \beta^{\star p} \right\},$$

where = 1= and conv denotes convex hull, so that Problem (3.28) becomes

$$\min_{2R^p} \max_{2D} jjX \quad yjj_2^2 + \quad jj \quad jj_2^2$$
$$(\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad)$$
$$= \min_{2R^p} \quad jjX \quad yjj_2^2 + \quad jj \quad jj_2^2 + \quad \max_{2f \quad , \quad g^p} \quad ü \quad + jj \quad jj_2^2$$
$$= \min_{2R^p} f \quad jjX \quad yjj_2^2 + \quad jj \quad jj_2^2 + \quad k \quad k_1 + c$$
$$, \quad \min_{2R^p} \frac{1}{2} jjX \quad yjj_2^2 + \quad _1 k \quad k_1 + \frac{2}{2} jj \quad jj_2^2 \quad . \quad (3.29)$$

which is recognized as an Elastic-net problem. When is null, we recover Ridge regression, and when the magnitude of grows, the problem approaches the Lasso. A 2D pictorial illustration of this evolution is given in Figure 3.15, where the shape of the uncertainty set D is the convex hull of the points located at (t ,)ü, which are identified by the cross markers. Then, the sublevel set max $_{2D}$ jj jj$_2^2$ tg is simply defined as the intersection of the four sublevel sets: jj jj$_2^2$ tg for = (,)ü, which are Euclidean balls centered at these values.

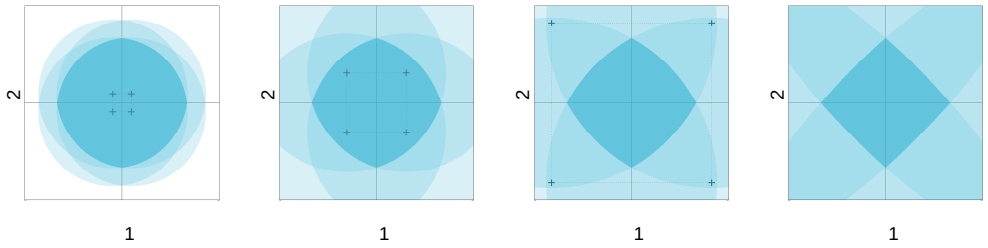


Figure 3.15 – *Sublevel sets for Elastic-net (darker colored patches). Each set is defined as the intersection of the Euclidean balls (lighter color patches) the centers of which are represented by crosses.*

‘ $_1$=‘ $_1$ group-Lasso. For this option, we first consider the one-group situation, in which case the ‘ $_1$=‘ $_1$ group-Lasso boils down to $_1$ regularization. The regularity assumption now states that the $_1$-norm of ? should be small:

$$H \; ‘_{\frac{1}{?}} = f \quad 2 R^p : jj \quad jj_1 \quad .$$

The dual assumption is that the $_1$-norm of should be controlled:

$$D \; ‘_1 = \left\{ \quad 2 R^p : \sup_{2H \; ‘_{\frac{1}{?}}} ü \quad 1 \right\}$$
$$= f \quad 2 R^p : jj \quad jj_1 \quad = \text{conv} f \; e_1^p, \ldots, \; e_p^p, \quad e_1^p, \ldots, \quad e_p^p \; ,$$

where = 1= and e_j^p is the jth element of the canonical basis of R^p, that is $e_{jj^0} = 1$ if $j = j^0$ and $e_{jj^0} = 0$ otherwise. Then, Problem (3.28) becomes:

$$\min_{2R^p} \max_{2D \; ‘_1} \; jjX \quad yjj_2^2 + \quad jj \quad jj_2^2 \; , \; \min_{2R^p} jjX \quad yjj_2^2 + 2 \quad jj \quad jj_1 + \quad jj \quad jj_2^2 .$$

Now, consider the more general situation where a group structure $\mathcal{G} = \{G_k\}_{k=1}^K$ is given with cardinality p_k for group k. We examine the regularity assumption stating that the $\ell_{1,1}$ mixed-norm of $\beta^\star$ (that is, its groupwise ℓ_1-norm) should be small:

$$H_{1,1} = \left\{ \beta \in \mathbb{R}^p : \sum_{k=1}^K \|\beta_{G_k}\|_1 \right\}.$$

The dual assumption is that the groupwise ℓ_∞-norm of $\beta^\star$ should be controlled:

$$D_{1,1} = \left\{ \beta \in \mathbb{R}^p : \sup_{\beta \in H_{1,1}} \beta^T \beta \leq 1 \right\} = \left\{ \beta \in \mathbb{R}^p : \max_{k \in \{1,\dots,K\}} \|\beta_{G_k}\|_1 \right\}$$
$$= \operatorname{conv}\left(\left\{ e_1^{p_1}, \dots, e_{p_1}^{p_1}, e_1^{p_1}, \dots, e_{p_1}^{p_1}, \dots \right. \right.$$
$$\left. \left. e_1^{p_K}, \dots, e_{p_K}^{p_K}, e_1^{p_K}, \dots, e_{p_K}^{p_K} \right\} \right),$$

so that Problem (3.28) becomes:

$$\min_{\beta \in \mathbb{R}^p} \|X\beta - y\|_2^2 + \lambda \sum_{k=1}^K \|\beta_{G_k}\|_1 + \mu \|\beta\|_2^2.$$

Notice that the limiting cases of this penalty are two classical problems: ridge regression and the $\ell_{1,1}$ group-Lasso. A 2D pictorial illustration of this evolution is given in Figure 3.16, where the shape of the uncertainty set D is the convex hull of the points located on the axes at ± 1, which are identified by the cross markers. Then, the sublevel set $\{\beta : \max_{2D} \|\beta\|_1^2 + \|\beta\|_2^2 \leq t\}$ is simply defined as the intersection of the four sublevel sets $\{\beta : \|\beta - \beta^\star\|_1^2 + \|\beta\|_2^2 \leq t\}$ for $\beta^\star = \pm e_1^2$ and $\beta^\star = \pm e_1^2$, which are Euclidean balls centered at these $\beta^\star$ values.

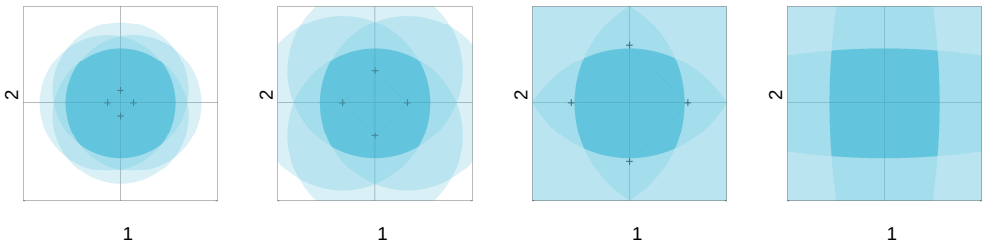


Figure 3.16 – *Sublevel sets for the $\ell_{1,1}$ group-Lasso (darker colored patches). Each set is defined as the intersection of the Euclidean balls (lighter color patches) the centers of which are represented by crosses.*

Remark (Other problems entering this framework.) Though we omit the details, the assumptions on $\beta^\star$ and D can be written for a series of other penalized problems the feasible sets of which are represented in Figure 3.17 with their quadratic representation.

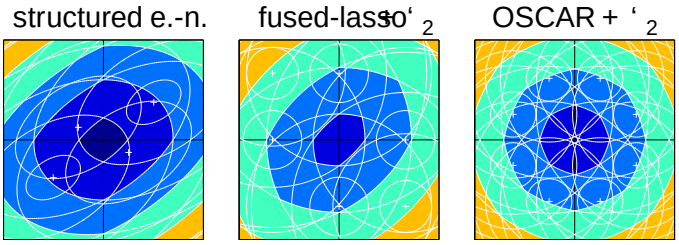


Figure 3.17 – *Additional penalty shapes built from quadratic functions.*

Algorithm. Our derivation for the above problems suggests a unified processing based on the iterative resolution of quadratic problems. This algorithm enters the general active set strategy 1 depicted in Section 3.1.2. We state it more specifically in light of the adversarial formulation in Algorithm 2.

Algorithm 2: Active set algorithm under the angle of adversarial formulation

S0 Initialization

$^0, A$ $f_j : {}^0_j \neq 0$ // Start with a feasible , recover A

$= \arg\max_{g \in D} jj \quad gjj_2^2$ // Pick a worst admissible

S1 Update active variables $_A$

$_A \quad X^{\ddot{u}}_A X_A + I_{jAj}{}^1 X^{\ddot{u}}_A y + \quad _A$ // Solve quadratic problem

S2 Verify coherence of $_A$ with the updated $_A$

if $jj \quad _A \quad _A jj_2^2 < \max_{g \in D} jj \quad _A \quad g_A jj_2^2$ then // if $_A$ is not worst-case

$\quad _A \quad {}^{old}_A + (\quad _A \quad {}^{old}_A)$ // Find $_A$-coherent solution

S3 Update active set A

$g_j \quad \min_{2D} x^{\ddot{u}}_j (X_A \quad _A \quad y) + (\quad _j \quad _j) \quad j = 1, \ldots, p$ // worst gradient

if $9 j \in A : \quad _j = 0$ *and* $g_j = 0$ then

$\quad A \quad A \setminus f \quad j g$ // Downgrade j

else if $\max_{j \in A^c} g_j \neq 0$ then

$\quad$ find $j^? \quad \arg\max_{j \in A^c} g_j , \quad A \quad A \cup [f \quad j^? g$ // Upgrade $j^?$

else

$\quad$ Stop and return // Optimality is reached

Starting from a sparse initial guess, Algorithm 2 iterates the following three steps:

1. the first step solves Problem (3.28) considering that A, the set of "active" variables, is correct. This penalized least squares problem is defined from X_A, which is the submatrix of X comprising all rows and the columns indexed by A and $_A$, which is set at its current most adversarial values.[8]

2. the second step updates $_A$ (and $_A$) if necessary, so that $_A$ is indeed (one of) the most adversarial values of the current $_A$. This is checked for the penalized

[8] When several $_A$ are equally unfavorable to $_A$, we use gradient information to pick the worst one among those when $_A$ moves along the steepest descent direction.

problems considered above, where $\mathcal{D}$ is a convex polytope the vertices of which (that is, extreme ν-values) are associated with a cone of coherent ν-values.

3. the last step updates the active set $\mathcal{A}$. It relies on the "worst-case gradient" with respect to ν, where ν is chosen so as to minimize infinitesimal improvements of the current solution. Picking the right ν is easy for the problems we considered. When no violation of the optimality conditions exists, the solution is optimal, since, at this stage, the problem is solved exactly within the active set $\mathcal{A}$.

Note that the structure is essentially identical to that of the homotopy algorithm of [134] or the Lars [48] for the Lasso, but that it applies to any penalty that can be decomposed as in Problem (3.28). Our viewpoint is also radically different, as the global non-smooth problem is dealt with via subdifferentials by [134], whereas we rely on the maximum of smooth functions.

Numerical experiments. This section compares the performance of our algorithm to its state-of-the-art competitors from an optimization viewpoint, where efficiency is assessed by accuracy and speed: accuracy is the difference between the optimum of the objective function and its value at the solution returned by the algorithm; speed is the computing time required for returning this solution. Obviously, the timing of two algorithms/ packages has to be compared at similar precision requirements, which are rather crude in machine learning, far from machine precision [116].

Comparing optimization strategies. We compare here the performance of three state-of-the-art optimization strategies implemented in our own computational framework: accelerated proximal method [9], coordinate descent [56], and our algorithm, that will respectively be named hereafter proximal, coordinate and quadratic. Our implementations estimate the solution to Elastic-net problem

$$J^{\text{enet}}_{\lambda_1,\lambda_2}(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \frac{\lambda_2}{2}\|\beta\|_2^2, \tag{3.30}$$

which is strictly convex when $\lambda_2 > 0$ and thus admits a unique solution even if $n < p$.

The three implementations are embedded in the same active set routine, which approximately solves the optimization problem with respect to a limited number of variables as in Algorithm 2. They only differ regarding the inner optimization problem with respect to the current active variables, which is performed by an accelerated proximal gradient method for proximal, by coordinate descent for coordinate, and by the resolution of the worst-case quadratic problem for quadratic. We followed the practical recommendations of [15] for accelerating the proximal and coordinate descent implementations, and we used the same halting condition for the three implementations, based on the approximate satisfaction of the first-order optimality conditions:

$$\max_{j \in \{1,\dots,p\}} \left| x_j^{\top}\left(y - X\hat{\beta} + \lambda_2\hat{\beta}\right) \right| < \lambda_1 + \varepsilon, \tag{3.31}$$

where the threshold ε was set at $\varepsilon = 10^{-2}$ in our simulations.[9] Finally, the active set algorithm is itself wrapped in a warm-start routine, where the approximate solution to $J^{\text{enet}}_{\lambda_1,\lambda_2}$ is used as the starting point for the resolution of $J^{\text{enet}}_{\lambda'_1,\lambda_2}$ for $\lambda'_1 < \lambda_1$.

[9] The rather loose threshold is favorable to coordinate and proximal, which reach the threshold, while quadratic ends up with a much smaller value, due to the exact resolution, up to machine precision, of the inner quadratic problem.

Our benchmark considers small-scale linear regression problems, with size $p = 100$, and the nine situations stemming from the choice of following three parameters: i) low, medium and high levels of correlation between predictors ($\rho \in \{0.1, 0.4, 0.8\}$), ii) low, medium and high-dimensional setting ($n/p \in \{2, 1, 0.5\}$), iii) low, medium and high levels of sparsity ($s/p \in \{10\%, 30\%, 60\%\}$). Each solver computes the elastic net for the tuning parameters λ_1 and λ_2 on a 2D-grid of 50×50 values, and their running times have been averaged over 100 runs. All results are qualitatively similar re-

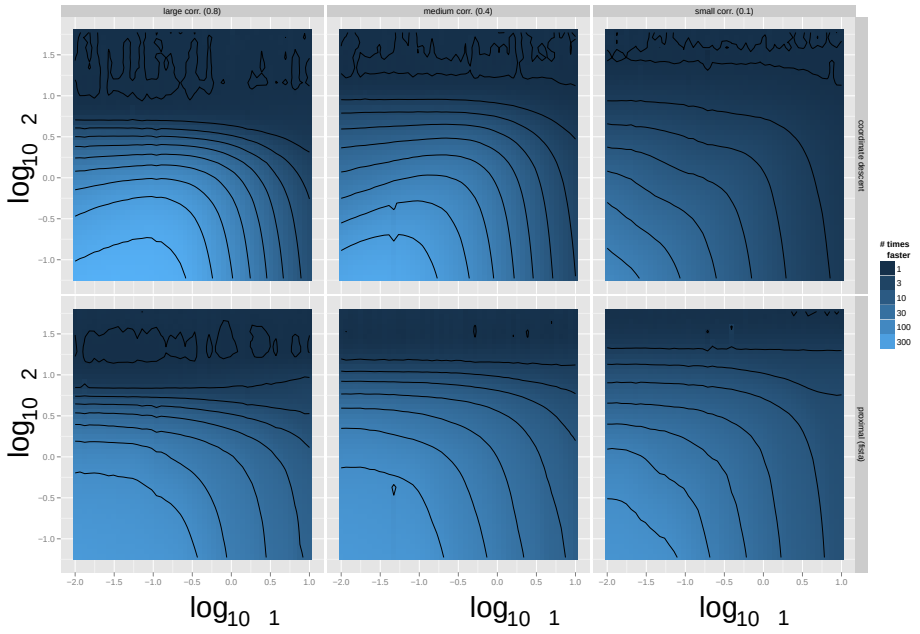


Figure 3.18 – *Log-ratio of computation times for* coordinate *(top) and* proximal *(bottom) versus* quadratic *, for high, medium and low variable correlation (left, center and right respectively).*

garding the dimension and sparsity settings. Figure 3.18 displays the high-dimensional case ($p = 2n$) with a medium level of sparsity ($s = 30$) for the three levels of correlation. Each map represents the log-ratio between the timing of either coordinate or proximal versus quadratic , according to (λ_1, λ_2) for a given correlation level. Dark regions with a value of 1 indicate identical running times while lighter regions with a value of 10 indicate that quadratic is 10 times faster. Figure 3.18 illustrates that quadratic outperforms both coordinate and proximal , by running much faster in most cases, even reaching 300-fold speed increases. The largest gains are observed for (λ_1, λ_2) small for which the problem is ill-conditioned, including many active variables, resulting in a huge slowdown of the first-order methods. As the penalty parameters increase, smaller gains are observed, especially when λ_2, attached to the quadratic penalty, reaches high values for which all problems are well-conditioned, and where the elastic net is leaning towards univariate soft thresholding, in which case all algorithms behave similarly.

Link between accuracy and prediction performance. When the "irrepresentable condition" holds, the Lasso should select the true model consistently. However, even when this rather restrictive condition is fulfilled, perfect support recovery obviously requires numerical accuracy: rough estimates may speed up the procedure, but whatever optimization strategy is used, stopping an algorithm is likely to prevent either the removal of all irrelevant coefficients or the insertion of all relevant ones. The support of the solution may then be far from the optimal one.

We advocate here that our solver is competitive in computation time when support recovery matters, that is, when a high level of accuracy is needed, in small (a few hundred variables) and medium sized problems (a few thousand). As an illustration, we generate 100 data sets under the linear model with a coefficient of determination R^2 0.8, a high level of correlation between predictors (0.8) and a medium level of sparsity ($=p$ = 30%). The number of variables is kept low (p = 100) and the difficulty is tuned by the $n=p$ ratio. For each data set, we generate a large test set (say, 10n) to evaluate the quality of the prediction without depending on any sampling fluctuation. We compare the Lasso solutions computed by our solver to the ones returned by glmnet with various levels of accuracy[10]. Figure 3.19 reports performance. As ex-

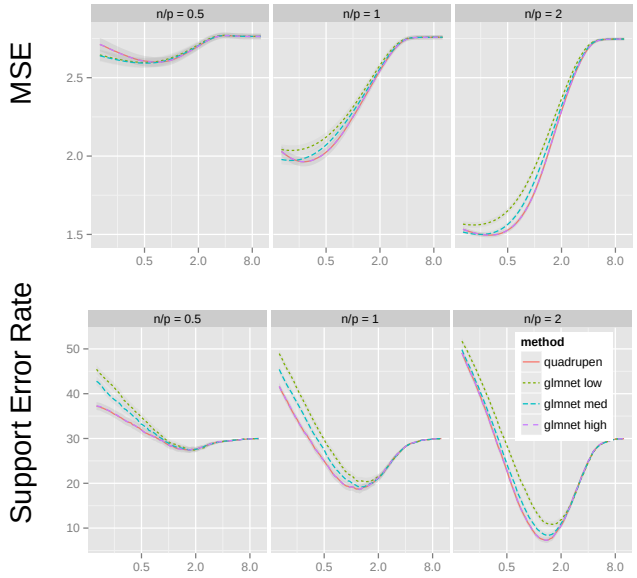


Figure 3.19 –*Test performance according to the penalty parameter for the Lasso returned by* quadrupen *and* glmnet *for various levels of accuracy and sample sizes (*$n=p$ 2 f1=2, 12g*).*

pected, the curves show that selecting variables and searching for the best prediction are two different problems. The selection problem (bottom of Figure 3.19) always requires a sparser model. But despite this obvious difference, the more accurate the solution returned by the algorithm, the better the performance for any level of penalty and for both performance measures. Now focusing on glmnet performance, the better the ac-

methods	quadrupen	glmnet (low)	glmnet (med)	glmnet (high)
timing (ms)	8	7	8	64
accuracy (dist. to opt.)	5.9 10 ^{14}	7.2 10^{0}	6.04 10^{0}	1.47 10 ^{2}

Table 3.4 –*Median timings and solution accuracies*

curacy, the smaller the MSE and the support error rate, but the slower the algorithm becomes. With high precision, the performance differences become negligible between our approach and glmnet running. However, Table 3.4 illustrates that high accuracy is achieved at a high computational cost: to be at par with quadrupen with respect to test performance, glmnet is about ten times slower.

[10] This is done via the thresh argument of the glmnet procedure. In our experiments, low, med and high level of accuracy for glmnet respectively correspond to thresh set to 1e-1, 1e-4, and 1e-9.

3.2.4 Tree reconstruction with fusion penalties

Given a data set with many features observed in a large number of conditions, it is desirable to fuse and aggregate conditions which are similar to ease the interpretation and extract the main characteristics of the data. We present [JP2] a multidimensional weighted fusion penalty to address this question when the number of conditions is large. If the fusion penalty is encoded by a mixed ℓ_1/ℓ_q norm, uniform weights lead to a path of solutions which is a tree, very suitable for interpretability. Moreover, when the ℓ_q is the ℓ_1 norm, the path is piecewise linear and we derive a homotopy algorithm to uncover exactly the whole tree structure. For weighted-ℓ_1 fusion penalties, we demonstrate that distance-decreasing weights lead to balanced tree structures. For a subclass of these weights that we call "exponentially adaptive", we derive a $O(n \log(n))$ homotopy algorithm and we prove an asymptotic oracle property.

Model setup. Consider y_{ij} the observation of a continuous random variable that describes the intensity of the jth feature in condition i, with $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$. The p-dimensional vector $y_i = (y_{i1}, \dots, y_{ip})$ encompasses the data related to condition i across the p features. We are given a partition with K groups as prior knowledge that is depicted by the indexing function $\beta : \{1, \dots, n\} \to \{1, \dots, K\}$. In words, β indicates the group to which condition i is allocated *a priori*. The number of elements in group k is denoted by $n_k = \text{card}\{i : \beta(i) = k\}$, such that $\sum_k n_k = n$.

We consider the convexified Lagrangian formulation of hierarchical clustering suggested by [75], which we adapt to the (M)ANOVA framework:

$$\arg\min_{\beta \in \mathbb{R}^{Kp}} \frac{1}{2} \sum_{i=1}^{n} \left\| y_i - \beta_{(i)} \right\|_2^2 + \sum_{k,\, k' : k < k'} w_{k,k'} \left\| \beta_k - \beta_{k'} \right\|_q . \tag{3.32}$$

In general, $w_{k,k'}$ are positive, symmetric weights over all pairs of groups. The penalty is a weighted ℓ_1/ℓ_q mixed-norm as in the ℓ_1/ℓ_q group-Lasso (3.6), acting on the pairwise differences between vectors $\{\beta_k\}_{k=1}^K$. It is designed to encourage elements of β to "fuse" group-wise, as done element-wise in the fused-Lasso signal approximator [56]. The level of fusion is tuned by two parameters: the global level of penalty γ and the group specific weights $w_{k,k'}$, the choice of which is of the highest importance. It conditions i) the ability of the method to infer an interpretable structure between the conditions, ii) the existence of fast algorithms to fit the parameters for various values of γ and iii) the existence of statistical guarantees for the estimator. The main objective of this paper is to study classes of weights that reach these three goals simultaneously.

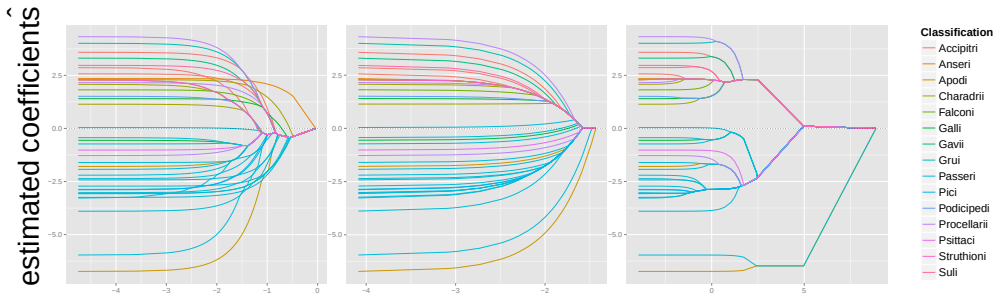
Although a part of our full-length paper [JP2] provides results for general ℓ_1/ℓ_q norms, here we stick to the case where the fusion penalty, the ℓ_q-norm, is the ℓ_1-norm. Hence, the problem decouples across the p dimensions, in which case our contributions are the following:

> We introduce distance-decreasing weights for which we prove that the path is a tree. From an interpretation point of view, this family of weights is particularly interesting as it leads to *balanced* tree structures.
>
> we introduce exponentially adaptive weights that enter the family of distance-decreasing weights. They enjoy asymptotic oracle properties that guarantee selection of the true underlying structure for a large scale of possible

We provide a general homotopy algorithm for (3.32) $_q$ ' $_1$, whatever the choice of $w_{k'}$. On a single feature, the initialization for unspecified weights is in $O(K^2)$ and the homotopy itself is in $O(K \log(K))$. However, we propose a faster initialization procedure for exponentially adaptive weights such that the whole complexity for p features is in $O(pK \log(K))$ – or $O(pn \log(n))$ in the clustering framework.

When the number K of prior groups is smaller than n (*e.g.*, in the ANOVA settings, when there are replicates per condition/group), a natural cross-validation (CV) error can be defined. In this case, we develop a fast procedure that takes advantage of the DAG (directed acyclic graph) structure of the path of solutions along . This approach has a lower complexity than does the standard CV.

Motivating example in phylogeny. As a simple motivating example, we consider a univariate problem in phylogeny. We want to reconstruct a tree between many species based on some simple features (like the height, or the weight of individuals). Ideally, this tree should resemble the known phylogeny. We illustrate this task on the "Animal Ageing Longevity Database"[11], which provides various features for many animal species. Here, we consider classifying bird species based on their birth weight. The known phylogeny groups these $n = 184$ individuals into 40 bird families, themselves grouped into 15 orders. We reconstruct the tree based on the weights and check whether it matches the orders and the family classification. Recovered solution paths of (3.32) are plotted in Figure 3.20 for (a) the Cas-ANOVA weights [13] ; (b) the "default" Clusterpath weights [75]; and (c) our own weights that we call "fused-ANOVA" weights. On the left panel, the Cas-ANOVA path includes many splits which make interpretation rather difficult. On the middle panel, default Clusterpath weights, as expected, provide a tree structure. Still, the structure of this tree is unbalanced and thus not fully satisfactory in the sense that small groups often fuse with very large ones. Specifically, the Clusterpath tree does not capture the simple fact that there are visibly three groups corresponding to light, medium or heavy birds. Conversely, the fused-ANOVA tree in the right panel is more balanced and clearly exhibits these three groups. Furthermore, it is in better agreement with the known phylogenetic classification, improving the rand index by 5% compared to ClusterPath.



a) Cas-ANOVA *b*) Clusterpath *c*) Fused-ANOVA

Figure 3.20 – *Reconstructed phylogenetic trees for various weighting schemes. Families classified in the same order share the same color.*

[11] publicly available at http://genomics.senescence.info/species/

Distance-decreasing weights guaranteeing no split for ℓ_1 fusion penalties. By choosing the ℓ_q-norm to be the ℓ_1-norm in (3.32), the p-dimensional problem splits into p univariate problems. We then recover a consensus, multidimensional classification by first inferring p independent trees (one per dimension) and then aggregate those p trees by considering the same penalty value. Thus, without loss of generality, we restrict the discussion to the following ℓ_1 univariate problem:

$$\underset{\in \mathbb{R}^K}{\text{minimize}} \frac{1}{2} \sum_{k=1}^{K} n_k (\bar{y}_k -)^2 + \sum_{k,\ell:k<\ell} w_{k\ell} | - |. \tag{3.33}$$

Although the choice $w_{k\ell} = n_k n_\ell$ can be shown to prevent splits (see the full-length paper [JP2]), it will typically lead to fusion events occurring very late (that is, for large), even between groups having close empirical means. This corresponds to an unbalanced tree structure between the condition, which is hardly interpretable. Intuitively, distance-decreasing weights should ensure that close neighbors fuse quickly. Here, we indeed demonstrate that for such weights there is no split.

Theorem 5. *The path of solutions of* (3.33) *does not contain splits with weights*

$$w_{k\ell} = n_k n_\ell f(|\bar{y}_k - \bar{y}_\ell|),$$

where $f(\cdot)$ *is a decreasing positive function.*

Schematically, the proof (detailed in [JP2]) is based on two ingredients:

1. Using geometrical arguments, we show that absence of splits is equivalent to preservation of the order along the path, i.e., $\bar{y}_k \geq \bar{y}_\ell \Rightarrow \hat{}_k(\cdot) \geq \hat{}_\ell(\cdot)$;
2. By considering a problem that is dual to (3.33) as in [167] for the generalized Lasso, we show that distance-decreasing weights preserve the order.

Fast homotopy algorithm for ℓ_1 weighted penalties. In this paragraph, we consider algorithmic issues when ℓ_q is the ℓ_1-norm. As in the previous paragraph, we restrict the discussion to univariate Problem (3.33) and thus give the numerical complexity in the case $p = 1$. For a p-dimensional problem, we aggregate the p univariate trees by considering the same values of for all trees.

An algorithm for general weights and its limitations. Problem (3.33) can be solved for general weights $w_{k\ell}$ by the homotopy algorithm proposed in [76] for the generalized fused-Lasso. This is also the solution retained in the clustering framework by [75]. A schematic view of this algorithm adapted to (3.33) is depicted in 3.

Algorithm 3: Homotopy algorithm for the generalized fused-Lasso

Input : data, weights and initial groups $\{y_i, w_{k\ell}, g$
Initialization for $= 0$
Initialize $_k$ parameters (equal to the empirical means $\bar{y}_k$)
Initialize the list of possible next events (only fusion at this stage)
while *all groups are not fused* do
  Find the next event (having the smallest), it can be a split or a fusion
  Update $_k$ parameters accordingly
  Update the list of possible next events (fusion and split)
Output : DAG of fusion and split events and associated values of the parameters

This procedure for general weights has two major flaws that may have detrimental effects on its computational performance:

By piecewise-linearity of the solution path, the total number of iterations (that is, the total number of events before all the groups have fused) is bounded. However, by rewriting (3.33) as a Lasso problem – which only requires straightforward algebra – we may exhibit pathological cases where there are $(3^K + 1)/2$ linear segments in the path of solutions [114], a complexity that we cannot afford even for a moderate number of conditions K.

While detecting fusion events in Algorithm 3 may be cheap since it roughly only requires calculation of the slopes $\partial \mathcal{L}_k(\beta)/\partial \beta$, checking for the possibility of split events boils down to maximum-flow problems, the resolution of which may clearly be a bottleneck at large scale [76].

To circumvent these limitations, we consider using the family of distance-decreasing weights introduced above to prevent splits and lead to a balanced tree. In this case the total number of events is exactly $K - 1$, which is the number of iterations required to fuse K groups into 1, assuming that there cannot be a fusion of more than two groups at once. Regarding the maximum-flow problems, they are completely absent from the algorithm. Still, we have to take into account the cost of detecting successive fusion events and for updating the coefficients $\beta_k(t)$ along the $K - 1$ steps. In the next paragraph, we propose a solution inducing a global complexity of $O(K \log(K))$ for a given choice of weights belonging to the family of distance-decreasing weights.

Weights with an $O(K \log(K))$ implementation. First we need to define the next time a fusion event is going to happen. We proceed mainly as in [76] for the one-dimensional fused-Lasso, except that the initial ordering is not defined by the neighborhood between the coefficients, but by the ordering of the empirical means $\bar{y}_k$. And thanks to the property of the distance-decreasing weights, this ordering remains the same throughout the algorithm, which allows us to compute the path in $K \log K$ operations. Here are some details. At the initialization step $t_0 = 0$ and the next fusion time is

$$t(\beta) = \arg\min_{t_{k\ell}(\beta) > t_0} t_{k\ell}, \quad t_{k\ell}(\beta) = t_0 + \left(\beta_k(t_0) - \beta_\ell(t_0)\right) \left(\frac{\partial \mathcal{L}_k}{\partial \beta}(\beta_0) - \frac{\partial \mathcal{L}_\ell}{\partial \beta}(\beta_0)\right)^{-1}. \quad (3.34)$$

In words, it is the smallest value of t among all the values such that two coefficients fuse. The main cost in (3.34) is due to the calculation of the slopes $\partial \mathcal{L}_k / \partial \beta$ at $\beta_0 = 0$. Note that $\beta_k(0) = \bar{y}_k$, and by means of the subdifferential equations for the fused-Lasso we can show that

$$\frac{\partial \mathcal{L}_k}{\partial \beta}(0) = \frac{1}{n_k} \sum_{\ell \neq k} w_{k\ell} \, \mathrm{sign}(\bar{y}_k - \bar{y}_\ell). \quad (3.35)$$

For general weights $w_{k\ell}$, computing these slopes for all k requires $O(K^2)$ operations and is the limiting factor of the algorithm. However, we provide an $O(K \log(K))$ procedure for a special case of our distance-decreasing weights that we call "exponentially adaptive weights" because of their statistical properties (as will be seen below):

$$w_{k\ell} = n_k n_\ell \exp\left(-\gamma \sqrt{\bar{n}} |\bar{y}_k - \bar{y}_\ell|\right), \quad \gamma > 0, \quad (3.36)$$

for a positive constant. The key idea to achieve $K\log(K)$ complexity with these weights is that each slope can be computed as the sum of two terms, for which there exist simple recurrence formulas: first, we order the $\bar{y}_k$ in decreasing order, which can be done in $K\log(K)$ operations. Assuming this is done, we get

$$\frac{\partial}{\partial}_k(0) = \sum n_\cdot \text{signs}(\bar{y}_k - \bar{y}_\cdot) \exp\left(-\sqrt{n_\cdot n_k}|\bar{y}_k - \bar{y}_\cdot|\right)$$
$$= \sum_{\cdot \le k} n_\cdot \exp\left(-\sqrt{n_\cdot n_k}(\bar{y}_\cdot - \bar{y}_k)\right) - \sum_{\cdot > k} n_\cdot \exp\left(-\sqrt{n_\cdot n_k}(\bar{y}_k - \bar{y}_\cdot)\right)$$
$$= \exp\left(-\sqrt{n_k}\bar{y}_k\right) \underbrace{\sum_{\cdot \le k} n_\cdot \exp\left(\sqrt{n_\cdot}\bar{y}_\cdot\right)}_{L_k} - \exp\left(\sqrt{n_k}\bar{y}_k\right) \underbrace{\sum_{\cdot > k} n_\cdot \exp\left(-\sqrt{n_\cdot}\bar{y}_\cdot\right)}_{R_k}.$$

The recurrence formulas are $R_{k+1} = R_k + n_k \exp\{-\sqrt{n_k}\bar{y}_k\}$ and $L_{k-1} = L_k + n_k \exp\{\sqrt{n_k}\bar{y}_k\}$. Thus, the initial slopes (3.35) and the first fusion time can be computed in $O(K\log(K))$.

Then, for each of the $K-1$ steps of the algorithm, we only need to update the two slopes and the two coefficients which are currently fusing: this requires a constant number of operations. Concerning the next fusion time however, the new minimum among the updated $t_{k'}(\cdot)$ is found in $\log(K)$ if stored in an appropriate structure. This way we can reach $O(K\log(K))$ for the global complexity.

As a final remark, note that we use the same storage solution – namely a binary tree – as did [76] for the one-dimensional fused-Lasso. By this means, we maintain the memory requirement at a low level that only grows linearly in K.

An embedded cross-validation procedure. When the number K of prior groups is smaller than n (*e.g.*, in the ANOVA settings), a natural cross-validation (CV) error can be defined, in order to choose an appropriate value of and thus provide the user with a fixed classification between the initial conditions. Although CV is often incriminated for being time-consuming, it is possible in this case to rely on the tree structure of the solution – or DAG in the case where split is allowed in the algorithm – to enhance the performance. Indeed, we can first build a tree using a training set (in which all prior groups are present) and then assess its performance by measuring its ability to predict the remaining individuals of the test set for any given value of Here, we perform the CV on a predefined grid of L values of because the fusion times will be different for every new training set and it would be memory-intensive to store the CV-error for every one of those fusion times. More details are provided in [JP2].

It is difficult to assess exactly the gain brought by using the tree structure for computing the CV error in general. Indeed, it depends on the tree itself, the length of its branches, its height and so on. Assuming a binary balanced tree of height $\log(K)$ with branches of equal length and an equally spaced grid of we can show that the complexity is in $O(LK/\log(K))$. If some groups fused rapidly (as with the fused-ANOVA weights), the gain could be even greater. In practice (see Figure 3.21.c), we often see a ten-fold difference between our CV procedure and a naive implementation.

Timing. We implemented both the general and the without-split version of Algorithm 3 in C++ embedded in an R-package called fusedanova distributed on R-forge. It contains a wide family of weights which are not mentioned in this paper due to space requirements. Figure 3.21 illustrates the rather good performance of our algorithm and implementation through three numerical experiments:

a) In the left panel, we illustrate the capability of our method to treat large scale problems extremely fast: we generate a size n vector such that $y_i \sim \mathcal{N}(0,1)$ and assume $n = K$, meaning one condition per group[12]. We vary n from 10^2 to 10^8 and record the corresponding timing in seconds. We apply our method with the exponentially adaptive weights and average over 10 trials. As can be seen, we can reconstruct a tree on $n = 10^6$ observations in about 10 seconds.

b) The middle panel illustrates the gain in run-time due to the fact that we no longer have to check for splits in the homotopy algorithm using a maximum-flow solver. We generate data as in the preceding experiment but with K conditions each containing $n_k = 20$ replicates. When $K = 10^3$, the gain in seconds brought by not checking for the possibility of splits is of almost 2 orders of magnitude.

c) The right panel shows the performance of our embedded CV procedure compared to the naive implementation, with the same settings as in the previous experiment.

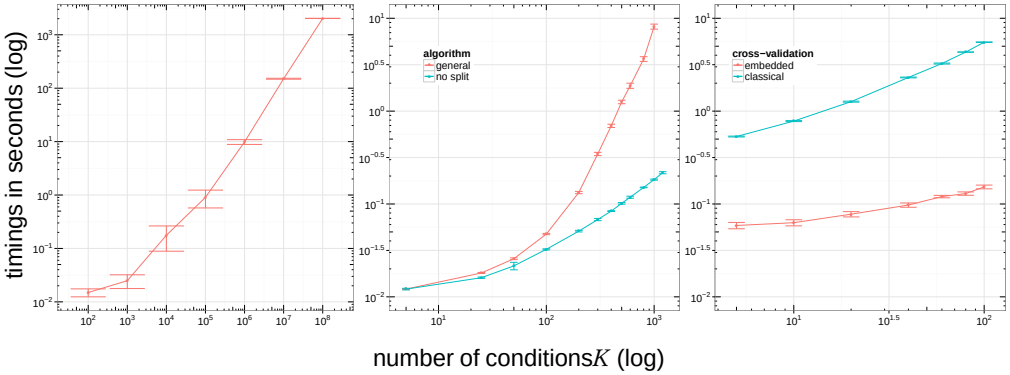


Figure 3.21 – *timing experiments: a) time in seconds as a function of the number of conditions K; b) timing comparison for generall without-split algorithm; and c) timing comparison for naivel embedded cross-validation.*

Statistical guarantees. To discuss the asymptotic properties of our exponentially adaptive weights (3.36), we shall consider the following univariate[13] ANOVA model

$$y_i = \beta^\star_{(i)} + \varepsilon_i, \quad \text{s.t.} \quad \mathrm{E}(\varepsilon_i) = 0, \ \mathrm{Var}(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, n, \tag{3.37}$$

where $\beta^\star = (\beta^\star_1, \ldots, \beta^\star_K)$ is the true vector of parameters and ε_i are iid residuals. The correct structure between the coefficients – or classification – is denoted by $\mathcal{A}^\star = \{(k,l) : \beta^\star_k = \beta^\star_l\}$. A usual technical assumption is to consider designs the associated gram matrices of which converge to positive definite matrices. In the one-way ANOVA settings, we just need to assume that when $n \to \infty$, then $n_k = n\,\omega_k \to \infty$ for all $k = 1, \ldots, K$. We denote by D the corresponding asymptotic covariance matrix which is a K-diagonal matrix with diagonal entries equal to $\omega_1, \ldots, \omega_K$. For the purpose of asymptotic analysis, we consider the problem

$$\hat{\beta}^{(n)} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^K} \frac{1}{2} \sum_{k=1}^{K} n_k (\bar{y}_k - \beta_k)^2 + \sum_{k \neq l} w^{(n)}_{k,l} |\beta_k - \beta_l|, \tag{3.38}$$

[12] There is no underlying clustering in this setting since our point is to compare run-times here.

[13] We numerically study the multidimensional case at the end of this section.

which is just a rewriting of (3.33) where the dependency on n is stated explicitly for the estimator and the tuning parameter. Similarly, $\hat{A}_n = \mathcal{A}(\hat{k,\ell}) : \hat{\theta}^{(n)}_k = \hat{\theta}^{(n)}_\ell$.

Exponentially adaptive weights and the fused-ANOVA. In this paragraph, we study the exponentially adaptive weights (3.36). In the context of the penalized ANOVA problem (3.38), we denote these weights by $w^{FA}_{k\ell}$ and call the associated estimator the *fused-ANOVA*. We show that they enjoy some "oracle properties" in the sense of [50], that is, both i) correct model identification (recovering the true classification $A^?$) and ii) optimal estimation rate $\sqrt{n}$. These weights are adaptive as in the adaptive-Lasso of [197]: it is known that raw ℓ_1 methods like the Lasso do not enjoy the aforementioned oracle properties, yet this can be fixed by choosing judicious weights that depend on an estimator of $\theta^?$ which is asymptotically $\sqrt{n}$-consistent – like the ordinary least squares, which equals $(\bar{y}_1, \ldots, \bar{y}_K)$ in the case at hand. Here we are interested in the differences between the entries of $\hat{\theta}$, thus the quantity $\sqrt{n}|\bar{y}_k - \bar{y}_\ell|$ seems quite natural in (3.36).

Another possible choice for those weights is given by [113], who consider Problem (3.38) with additional constraints on the α_k's – that must sum to zero – and the following weights, that we refer to as the *Cas-ANOVA* weights:

$$w^{CA}_{k\ell} = \frac{\sqrt{n_k + n_\ell}}{|\bar{y}_k - \bar{y}_\ell|}. \tag{3.39}$$

We now proceed to the Theorem stating the required conditions on γ_n for the fused-ANOVA to enjoy the oracle properties.

Theorem 6 (Oracle properties) *Suppose that* $\gamma_n n^{3/2} \exp(\sqrt{n}) \to 0$ *and* $\gamma_n n^{3/2} \to \infty$ *when* $n \to \infty$. *Then the fused-ANOVA enjoys asymptotic normality and consistency for recovering the true classification, i.e.,*

$$\sqrt{n}(\hat{\theta}^{(n)} - \theta^?) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1}) \quad \text{and} \quad \mathbb{P}(\hat{A}_n = A^?) \to 1 \text{ when } n \to \infty.$$

Remark (On the exponentially adaptive weights) The key idea behind this theorem is that when n goes to infinity, then $w^{FA}_{k\ell} = \sqrt{n}$ goes to infinity if $(k,\ell) \in A^?$ and to zero exponentially fast if $(k,\ell) \notin A^?$. This is due to the joint effect of the $\sqrt{n}$-consistency of the $\bar{y}_k$ and of the exponential. This is to be compared with Cas-ANOVA weights, where, when $n \to \infty$, $w^{CA}_{k\ell} = \sqrt{n}$ goes to infinity if $(k,\ell) \in A^?$, but only to a constant if $(k,\ell) \notin A^?$.

Remark (On the range of γ_n). Theorem 6 is true for a large range of γ_n values. In particular it is true for a constant γ_n. Asymptotically all groups belonging to the same class fuse almost immediately (i.e., for small values of ℓ of the order $n^{3/2} \exp(\sqrt{n})$) and the groups belonging to different classes fuse for very large ℓ, i.e., of the order $n^{3/2}$.

Numerical illustration in the univariate case. We generate data from model (3.37) as follows, for K the number of prior groups and n being fixed: the true vector $\theta^?$ is composed of K entries drawn randomly from $\{1, 2, 3\}$, such that the correct structure $A^?$ is always composed of 3 groups. Then, the initial group sizes z_k are drawn from a multinomial $\mathcal{M}(n, (p_1, \ldots, p_K))$ with $p_k = 1/K$ for all $k = 1, \ldots, K$, such that the n_k are approximately balanced. Finally, we let $\varepsilon_i \sim \mathcal{N}(0, 1)$.

We compare the capability of three weighting schemes to recover the true grouping $A^?$, namely the fused-ANOVA weights, the Cas-ANOVA weights, and the so-called *default weights* corresponding to $w_{k\ell} = n_k n_\ell$, which are not adaptive but produce a path of solutions that contains no split. Such weights correspond to the Clusterpath weights adapted to the ANOVA setup. We use our own code for each method. Typically, the computational burden required by Cas-ANOVA is huge, compared to that of the other two procedures as the path of solutions may contain splits. Qualitatively, the difference would be as in Figure 3.21, middle panel. Thus, we typically force the algorithm not to split when using the Cas-ANOVA weights.

We generate data as specified below, and for each procedure we check whether there exists at least one for which the correct structure is identified along the path of solutions. The probability of true support recovery is evaluated by replicating this experiment a large number of times . To investigate the asymptotic behavior of each method, we vary n from 50 to 1,000 and consider two scenarios for the initial number of groups K. First, K is fixed at 10 such that the number of elements in each group grows with n. In the second scenario, K grows with n through the relationship $K = 2.5 \log(n)$. The results are reported on Figure 3.22, with the first (resp. the second) scenario on the left (resp. the right) panel. The results confirm Theorem 6. The two adaptive procedures, Cas-ANOVA, and to a greater extent, fused-ANOVA, dominate the non-adaptive weights. As expected, fused-ANOVA always dominate Cas-ANOVA, as experienced in other scenarios (e.g., $K = C\sqrt{n}$) not reported here to save space.

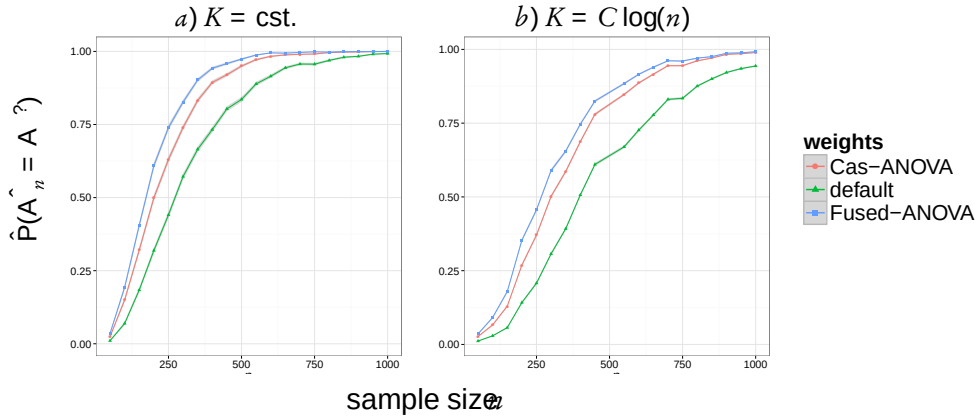


Figure 3.22 – *Univariate case: estimated probability of consistency as a function of the sample size n, for various weights and in two scenarios: the number of initial groups K is either a) fixed to a constant (10) or b) increases in $C \log(n)$ with $C = 2.5$. The true number of groups in $A^?$ is 3.*

Numerical illustration in the bivariate case. Theorem 6 characterizes the asymptotic of the fused-ANOVA estimators when considering one dimension at a time. Concerning the multidimensional setting, there are two situations. In the first one, there exists a dimension such that all the true groups are different, i.e., $\forall k \neq \ell, \mu^?_{kj} \neq \mu^?_{\ell j}$. In this case, our theorem guarantees that, using this particular dimension, the recovered classification will converge to the true one. In the second situation, there exists no dimension such that the true groups are all different. In that case, we have no theoretical guarantee to support the fused-ANOVA weights. It is nonetheless possible to aggregate the classification obtained in each dimension to a consensus classification.

For a given , two individuals k and ‘ are in the same multidimensional cluster if they have been fused on every dimension.

In order to evaluate empirically the performance of the aggregation step, we consider a two-dimensional classification problem with three classes and two scenarios. Each *prior* group is drawn from one of three classes. In the first scenario, the three classes have different means on the first dimension and the same mean on the second dimension. The mean vectors are $(1, 1.5)$; $(2, 1.5)$; $(3, 1.5)$, as in the top left panel of Figure 3.23. In the second scenario, both dimensions are informative: the first dimension separates classes $f1, 2g$ from $f3g$ while the second dimension separates class $f3g$ from $f1, 2g$. The mean vectors are $(1, 1)$; $(1, 2)$; $(2, 1)$, as in the top right panel of Figure 3.23). We increase the difficulty in each scenario by adding a Gaussian noise with increasing standard deviation . Results in Figure 3.23 correspond to the estimated probability of true classification recovery along the path, averaged over 2,000 runs.

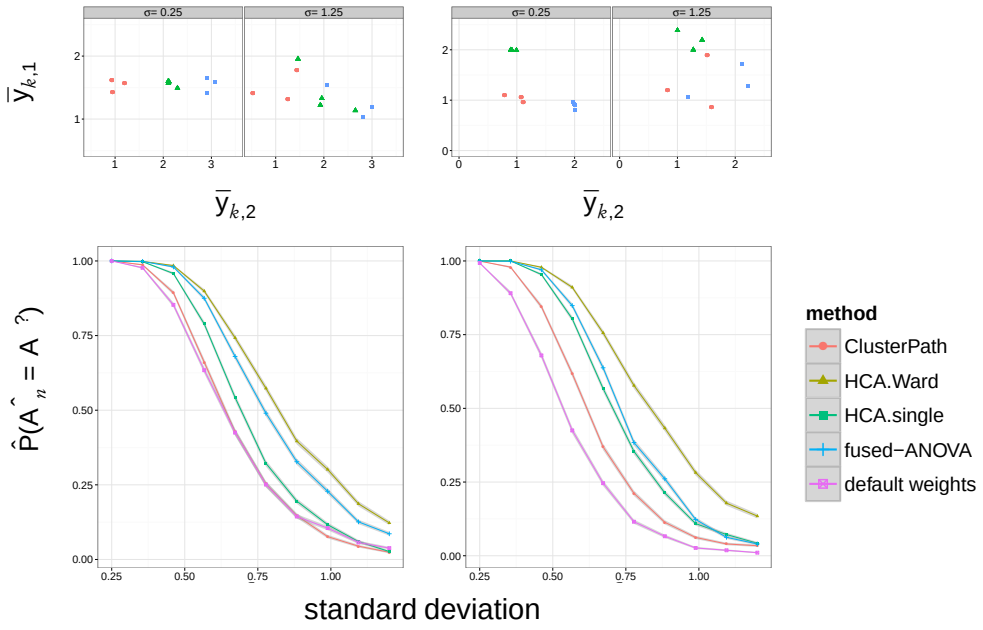


Figure 3.23 – *Bivariate example: estimated probability of consistency as a function of the noise standard deviation , for various clustering methods. The initial number of groups K fixed to a constant (10). The true number of groups in* $A^?$ *is 3.*

In both scenarios, the fused-ANOVA weights with aggregation outperform the multidimensional ‘ $_2$-Clusterpath as well as the single linkage hierarchical clustering. The Ward hierarchical clustering shows better performance but at a much higher computational cost.

Some final comments. Our full-length paper [JP2] contains more thorough application in phylogeny where it is shown that our method outperforms Ward hierarchical clustering, in the sense that it reconstructs a tree structure in better adequacy with the true phylogeny.

At the moment, we are developing an efficient algorithm for performing the aggregation of the many trees reconstructed independently for the multidimensional case in order to apply fused-ANOVA to metagenomics, where one needs to reconstruct hierarchy for billions of features.

3.3 PERSPECTIVES

Since I have been working on genomic data, I have almost always been relying on embedded sparse methods, the most famous and emblematic of which is the Lasso. In the past decade, it has become our best ally for performing simultaneously estimation and variable selection in a high-dimensional setting. This success is undoubtedly also due to the powerful computational methods that fit the Lasso such as the LARS and other homotopy algorithms, coordinate-wise descent strategies or proximal methods. As a consequence, ever since the Lasso was published in 1996, an outstanding number of variations around sparse methods have been published: ten years ago, we could have spoken of the Lasso "zoo"; now, it has turned out to be a jungle.

However, sparse methods still remain perfectible. Their most important limitation is probably their lack of stability when used as a variable selection operator. Such an instability is due to several points, two of which retain my attention: first, there is no universal method for calibrating the amount of regularization. And second, sparse methods are highly sensitive to small changes in the data, as they are mostly applied in a high-dimensional setting.

Instability and lack of robustness are exacerbated in genomics and other complex data, due to diverse reasons that we have discussed in this manuscript such as high-dimensional feature spaces, high level of noise or multiple levels of heterogeneity. Also, the presence of structure in the data, which can be a precious ally when it is well characterized and integrated as prior information, can have detrimental effects since strong correlations are known to mislead sparse methods from selecting the relevant features. And finally, a great source of confusion comes from the fact that these methods are often used to perform variable selection, hoping for biological interpretability of the selected predictors, while they are mostly designed to select variables doing a good job for prediction purposes.

To alleviate these limitations, possible research paths that I wish to follow are:

1. to keep on introducing priors/constraints *via* structured sparsity. However, this does not imply the design of a new method for each problem considered! I rather think that ℓ_1 has become an intrinsic part of what is now "mainstream" statistics. It is thus natural to integrate regularization in most of the statistical methods that we know.

2. to address the question of statistical inference in order to properly endow sparse estimators with the notion of statistical significance. Several works have recently tried to tackle this issue [110, 86, 12] but remain largely imperfect.

3. to revisit robust statistics for high-dimensional data: first attempts have been made to equip sparse fitting procedures against the effect of outliers and adapt standard robust statistics to the high-dimensional setup [29, 34, 35]. However, the notion of outlier is hard to define when data is scarce [87].

4. to keep on focusing on methods that allow for efficient algorithms. In fact, efficiency should be kept in mind when designing a regularization method, as our objective as an applied statistician is – of course – to scale real data situations.

Here is some of my ongoing work related to these perspectives:

Marker assisted selection. The contribution [PP2] on structured regularization for conditional Gaussian graphical models has been initially motivated by application in agronomy and genomic selection (or marker assisted selection). In this context, I am co-supervizing, with Tristan Mary-Huard, David Baker's Post-doctorate, which intends to extend our proposal [PP2] to a more involved regularization procedure motivated by data characteristics which are typical problems in agronomy. An important objective of the post-doc is to generalize existing regularized multivariate regression approaches to the context of multi-task learning, where the learning task has to be jointly performed on several inhomogeneous training populations. Indeed, the population structure is very strong in agronomy because individuals are obtained by multiple crossing of the same parents, for both plant breeding and animal breeding.

Another aspect of this post-doc is more algorithmic: regularization methods used by the genetic community often rely on Bayesian formulation, the underlying optimization of which is sometimes close, nay equivalent, to their frequentist counterpart (think for instance about ridge regression). Hence, we suggest reconsidering the most widespread Bayesian models in genetics in light of their frequentist, penalized formulations, in order to speed up the whole process.

"Spiny" regression: take advantage of both frequentist and Bayesian interpretations. In the context of Pierre-Alexandre Mattei's PhD thesis, supervised by Charles Bouveyron and Pierre Latouche, I am involved in a Bayesian method for variable selection in high-dimensional linear regression [JP1]. The method builds on a generative model that uses a spike-and-slab-like prior distribution obtained by multiplying a deterministic binary vector, which describes the sparsity of the problem, with a random Gaussian parameter vector. The originality of the work is to consider inference through relaxing the model and using a type-II log-likelihood maximization based on an EM algorithm, thus providing both fine estimation from the Bayesian formulation and fast frequentist algorithms.

Two-dimensional segmentation with fast Lasso like approaches. With Céline Lévy-Leduc and Vincent Brault, in the context of Vincent's Post-Doctorate, we are working on a novel approach for estimating the location of block boundaries (change-points) in a random matrix consisting of a block-wise constant matrix observed in white noise. Our method consists in rephrasing this task as a variable selection issue. We use a penalized least-squares criterion with a ℓ_1-type penalty for dealing with this problem. This problem arises from Hi-C genome-wide interaction data [LL07], a recent technique in genomics allowing the assessment of chromosome conformation across the entire genome. Hi-C data can be represented by large square matrices of similarity across all the positions along the genome. These matrices are very sparse and typically exhibit block-wise structures which are of interest to the biologist. At the end of the day, we hope to apply our method to Hi-C data with fine resolution, up to the nucleotide level. A paper is currently under review related to this work [JP1].

BIBLIOGRAPHY

[1] G.I. Allen and Z. Liu. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Mach. Learn.*, 73(3):243–272, 2008.

[3] F. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM, 2008.

[4] F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

[5] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

[6] S. Bakin. *Adaptive regression and model selection in data mining problems*. PhD thesis, Australian National University, Canberra, 1999.

[7] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008.

[8] A. Bar-Hen and J.-M. Poggi. Influential observations in a graphical model. In *Proceedings of Journées Françaises de la Statistique*, 2014.

[9] A. Beck and M. Teboulle. Fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, 2:183–202, 2009.

[10] L.C. Bergersen, K. Tharmaratnam, and I.K. Glad. Monotone splines lasso. *Comput. Stat. Data Anal.*, 77:336–351, 2014.

[11] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, 2000.

[12] M. Bogdan, E. van den Berg, W. Su, and E. Candes. Statistical estimation and testing via the sorted l1 norm. *arXiv preprint arXiv:1310.1969*, 2013.

[13] H.D. Bondell and B.J. Reich. Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, 65(1):169–177, 2008.

[14] H.D. Bondell and B.J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123, 2008.

[15] L. Bottou and O. Bousquet. Learning using large datasets. *Mining Massive DataSets for Security*, 3, 2008.

[16] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends R in Machine Learning*, 3(1):1–122, 2011.

[18] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, third edition, 2006.

[19] Z. Bozdech, M. Llinás, B.L. Pulliam, E.D. Wong, J. Zhu, and J.L. DeRisi. The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS biology*, 1(1):e5, 2003.

[20] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[21] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[22] T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106:594–607, 2011.

[23] E. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[24] R. Castelo and A. Roverato. A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *J. Mach. Learn. Res.*, 7:2621–2650, 2006.

[25] S. Celik, B. Logsdon, and S.-I. Lee. Efficient dimensionality reduction for high-dimensional network estimation. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1953–1961, 2014.

[26] C. Charbonnier, N. Verzelen, and F. Villers. A global homogeneity test for high-dimensional linear regression. *arXiv preprint arXiv:1308.3568*, 2013.

[27] J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771, 2008.

[28] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159 (electronic), 2001. Reprinted from SIAM J. Sci. Comput. 20 (1998), no. 1, 33–61 (electronic) [MR1639094 (99h:94013)]

[29] Y. Chen, C. Caramanis, and S. Mannor. Robust high dimensional sparse regression and matching pursuit. *arXiv preprint arXiv:1301.2725*, 2013.

[30] Eric C Chi, Genevera I Allen, and Richard G Baraniuk. Convex biclustering. *arXiv preprint arXiv:1408.0856*, 2014.

[31] Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *J. Comput. Graph. Statist.*, (just-accepted):00–00, 2014.

[32] F.R.K. Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[33] L. Comminges and A. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Stat.*, 40(5):2667–2696, 2012.

[34] A. Dalalyan and Y. Chen. Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1259–1267, 2012.

[35] A. Dalalyan and R. Keriven. l_1-penalized robust estimation for a class of inverse problems arising in multiview geometry. In *Advances in Neural Information Processing Systems (NIPS)*, pages 441–449, 2009.

[36] C. Dalmasso, W. Carpentier, L. Meyer, C. Rouzioux, C. Goujard, M.-L. Chaix, O. Lambotte, V. Avettand-Fenoel, S. Le Clerc, L. De Senneville, et al. Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS genome wide association 01 study. *PloS one*, 3(12):e3907, 2008.

[37] P. Danaher, P. Wang, and D.M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc. B*, 76(2):373–397, 2014.

[38] A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, 9:1269–1294, 2008.

[39] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, 2008.

[40] G. de los Campos, J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus. Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 2012.

[41] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39(1):1–38, 1977. With discussion.

[42] A.P. Dempster. Covariance selection. *Biometrics, Special Multivariate Issue*.

[43] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, 90(1):196–212, 2004.

[44] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52:6–18, 2006.

[45] M. Drton and M.D. Perlman. Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, 22:430, 2007.

[46] M. Drton and M.D. Perlman. A SINful approach to Gaussian graphical model selection. *J. Statist. Plann. Inference*, 138(4):1179–1200, 2008.

[47] B. Efron. The estimation of prediction error: Covariance penalties and cross-validation (with discussion). *J. Amer. Statist. Assoc.*, 99:619–642, 2004.

[48] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

[49] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*, 2010.

[50] J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *JASA*, 96(456):1348–1360, 2001.

[51] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B*, 70(5):849–911, 2008.

[52] M.E. Ferreira, J. Satagopan, B.S. Yandell, P.H. Williams, and T.C. Osborn. Mapping loci controlling vernalization requirement and flowering time in brassica napus. *Theor. Appl. Genet.*, 90:727–732, 1995.

[53] R. Foygel and M. Drton. Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2020–2028, 2010.

[54] L.E. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[55] O. Frank and F. Harary. Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association*, 77(380):835–840, 1982.

[56] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

[57] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[58] J.H. Friedman. Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, 84(405):165–175, 1989.

[59] W.J. Fu. Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.*, 7(3):397–416, 1998.

[60] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, and P.O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12):4241–4257, 2000.

[61] J. Gertheiss and G. Tutz. Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, 4(4):2150–2180, 2010.

[62] C. Giraud. Estimation of Gaussian graphs by model selection. *Electronic Journal of Statistics*, 2:542–563, 2008.

[63] C. Giraud. *High-Dimensional Statistics*. CRC Monographs on Statistics & Applied Probability. Chapman & Hall, 2014.

[64] C. Giraud, S. Huet, and N. Verzelen. Graph selection with GGMselect. *Stat. Appl. Genet. Mol. Biol.*, 11(3):1–50, 2012.

[65] C. Giraud, S. Huet, and N. Verzelen. High-dimensional regression with unknown variance. *Statist. Sci.*, 27(4):500–518, 2012.

[66] M. Grechkin, M. Fazel, D. Witten, and S.-I. Lee. Pathway graphical lasso. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[67] M. Guedj, L Marisa, A de Reynies, B Orsetti, R Schiappa, F Bibeau, G MacGrogan, F Lerebours, P Finetti, M Longy, P Bertheau, F Bertrand, F Bonnet, A L Martin, J P Feugeas, I Bieche, J Lehmann-Che, R Lidereau, D Birnbaum, F Bertucci, H de The, and C Theillet. A refined molecular taxonomy of breast cancer. *Oncogene*, advanced publication online:1–11, 2011.

[68] J. Guo, E Levina, G. Michailidis, and J. Zhu. Biometrika. *Joint estimation of multiple graphical models*, 2011.

[69] Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.*, 105(492), 2010.

[70] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[71] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):145, 2012.

[72] M. Hebiri and S. van de Geer. The smooth-lasso and other $\ell_1+\ell_2$ penalized methods. *Electron. J. Stat.*, 5:1184–1226, 2011.

[73] K.R. Hess, K. Anderson, W.F. Symmans, V. Valero, N. Ibrahim, J.A. Mejia, D. Booser, R.L. Theriault, U. Buzdar, P.J. Dempsey, R. Rouzier, N. Sneige, J.S. Ross, T. Vidaurre, H.L. Gómez, G.N. Hortobagyi, and L. Pustzai. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, 24(26):4236–4244, 2006.

[74] T. Hesterberg, N. M. Choi, L. Meier, and C. Fraley. Least angle and ℓ_1 penalized regression: A review. *Statistics Surveys*, 2:61–93, 2008.

[75] T. Hocking, J.-P. Vert, F. Bach, and A. Joulin. Clusterpath: an algorithm for clustering using convex fusion penalties. In *Proceedings of the 28th ICML*, pages 745–752, 2011.

[76] H. Hoefling. A path algorithm for the fused lasso signal approximator. *J. Comput. Graph. Statist.*, 19(4):984–1006, 2010.

[77] A. Hoerl and R.W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[78] C.-J. Hsieh, M.A. Sustik, I.S. Dhillon, and P. Ravikumar. Quic: quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.*, 15(1):2911–2947, 2014.

[79] C.-J. Hsieh, M.A. Sustik, I.S. Dhillon, P. K Ravikumar, and R. Poldrack. Big & quic: Sparse inverse covariance estimation for a million variables. In *vances in Neural Information Processing Systems (NIPS)*, pages 3165–3173, 2013.

[80] J. Huang, S. Ma, H. Li, and C.-H. Zhang. The sparse Laplacian shrinkage estimator for high-dimensional regression. *Ann. Stat.*, 39(4):2021, 2011.

[81] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 417–424, 2009.

[82] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31(4):370–377, 2002.

[83] T. Jaakkola. *Advanced mean field methods: theory and practice*, chapter Tutorial on variational approximation methods. Neural Information Processing Series. MIT Press, Cambridge, MA, 2001.

[84] L. Jacob, P. Neuvial, and S. Dudoit. More power via graph-structured tests for differential expression of gene networks. *Ann. Appl. Stat.*, 6(2):561–600, 2012.

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

[86] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15(1):2869–2909, 2014.

[87] R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.*, 12:2777–2824, 2011.

[88] I.T. Jolliffe, N.T. Trendafilov, and M. Uddin. A modified principal component technique based on the lasso. *J. Comput. Graph. Statist.*, 12(3):531–547, 2003.

[89] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.*, 20(4):388–400, 2005.

[90] M. I Jordan. On statistics, computation and scalability. *Bernoulli*, 19(4):1378–1390, 2013.

[91] N. Katenka and E.D. Kolaczyk. Inference and characterization of multi-attribute networks with application to computational biology. *Ann. Appl. Stat.*, 6(3):1068–1094, 2012.

[92] K. Khare, S.-Y. Oh, and B. Rajaratnam. A convex pseudo-likelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. R. Statist. Soc. B*, 2014.

[93] H. Kiiveri. Multivariate analysis of microarray data: differential expression and differential connection. *BMC Bioinformatics*, 12(1):42, 2011.

[94] S. Kim and E.P. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genetics*, 5(8), 2009.

[95] S. Kim and E.P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the 27th International Conference on Machine Learning*, pages 543–550, 2010.

[96] M. Kolar, H. Liu, and E.P. Xing. Graph estimation from multi-attribute data. *J. Mach. Learn. Res.*, 15(1):1713–1750, 2014.

[97] M. K. Kolar, A. A. Le Song, and E. P. Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123, 2010.

[98] C. Kole, C.E. Thorman, B.H. Karlsson, J.P. Palta, P. Gaffney, B.S. Yandell, and T.C. Osborn. Comparative mapping of loci controlling winter survival and related traits in oilseed brassica rapa and B. napus. *Mol. Breed.*, 1:201–210, 2002.

[99] M. Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303–324, 2009.

[100] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. *arXiv preprint arXiv:1207.4747*, 2012.

[101] S.L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. Clarendon Press, New York, 1996. Oxford Science Publications.

[102] S. Lèbre. Inferring dynamic genetic networks with low order independencies. *Statistical Applications in Genetics and Molecular Biology*, 8(1):9, 2009.

[103] S. Lèbre, J. Becq, F. Devaux, M. P. H. Stumpf, and G. Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, 4(130):1–16, 2010.

[104] W. Lee and Y. Liu. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *J. Multivar. Anal.*, 111:241–255, 2012.

[105] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[106] C. Li and H. Li. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.*, 4(3):1498–1516, 2010.

[107] E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, and M.O. Dorschner. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[108] H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

[109] H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in neural information processing systems (NIPS)*, pages 1432–1440, 2010.

[110] R. Lockhart, J. Taylor, R.J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Ann. Statist.*, 42(2):413, 2014.

[111] A. Lorbert, D. Eis, V. Kostina, D.M. Blei, and P.J. Ramadge. Exploiting covariate similarity in sparse regression via the pairwise elastic net. In Yee W. Teh and D. M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)*, volume 9, pages 477–484, 2010.

[112] K. Lounici, M. Pontil, A.B. Tsybakov, and S. van de Geer. Sparsity for multi-task learning. In *Conference On Learning Theory*, 2009.

[113] K. Lounici, M. Pontil, A.B. Tsybakov, and S. van de Geer. Oracle inequalities and optimal inference under group sparsity. *Ann. Stat.*, 39:2164–2204, 2011.

[114] J. Mairal and B. Yu. Complexity analysis of the lasso regularization path. In *Proceedings of the 29th ICML*, 2012.

[115] D. Marbach, J.C. Costello, R. Küffner, N.M. Vega, R.J. Prill, D.M. Camacho, K.R. Allison, M. Kellis, J.J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

[116] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[117] S. Mariadassou, M.and Robin and C. Vacher. Uncovering latent structure in valued graphs: a variational approach. *Ann. Appl. Stat.*, pages 715–742, 2010.

[118] J.-M. Marin and C.P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag: New-York, 2007.

[119] B. Marlin, M. Schmidt, and K. Murphy. Group sparse priors for covariance estimation. In *Uncertainty in Artificial Intelligence*, 2009.

[120] R. Mazumder and T. Hastie. The graphical lasso: New insights and alternatives. *Electronic journal of statistics*, 6:2125.

[121] L. Meier, S. Van De Geer, and P. Bühlmann. The group Lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71, 2008.

[122] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[123] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society, Series B*, 72:417–473, 2010.

[124] N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104:1671–1681, 2009.

[125] E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 451–459, 2011.

[126] R. Natowicz, R. Incitti, E.G. Horta, B. Charles, P. Guinot, K. Yan, C. Coutant, F. André, and R. Pusztai, L. Rouzier. Prediction of the outcome of a preoperative chemotherapy in breast cancer using DNA probes that provide information on both complete and incomplete responses. *BMC Bioinformatics*, 9(149), 2008.

[127] S. Negahban, B. Yu, M. Wainwright, and P. Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1348–1356, 2009.

[128] Y. Nesterov. Subgradient methods for huge-scale optimization problems. *Mathematical Programming*, pages 1–23.

[129] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[130] A.Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS 14*, 2002.

[131] G. Obozinski, M.J. Wainwright, and M.I. Jordan. Support union recovery in high-dimensional multivariate regression. *Ann. Statist.*, 39(1):1–47, 2011.

[132] R. Opgen-Rhein and K. Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive model. *BMC Bioinformatics*, 8, 2007.

[133] M.R. Osborne, B. Presnell, and B.A. Turlach. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–403, 2000.

[134] M.R. Osborne, B. Presnell, and B.A. Turlach. On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2):319–337, 2000.

[135] D. Paul, E. Bair, T. Hastie, and R. Tibshirani. Preconditioning for feature selection and regression in high-dimensional problems. *Ann. Statist.*, pages 1595–1618, 2008.

[136] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.*, 104(486):735–746, 2009.

[137] Bruno-Edouard Perrin, Liva Ralaivola, Aurelien Mazurie, Samuele Bottani, Jacques Mallet, and Florence d'Alche Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19(suppl 2):ii138–ii148, 2003.

[138] P. Radchenko and G. Mukherjee. Consistent clustering using l_1 fusion penalty. *arXiv preprint arXiv:1412.0753*, 2014.

[139] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8(35), 2007.

[140] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert. Classification of microarray data using gene networks. *BMC bioinformatics*, 8(1):35, 2007.

[141] A. Rau, F. Jaffrézic, J.-L. Foulley, and R.W. Doerge. Reverse engineering gene regulatory networks using approximate Bayesian computation. *Statistics and Computing*, 22(6):1257–1271, 2012.

[142] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[143] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Stat.*, 38:1287–1319, 2010.

[144] W.C. Reinhold, M. Sunshine, H. Liu, S. Varma, K.W. Kohn, J. Morris, J. Doroshow, and Y. Pommier. Cellminer: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the nci-60 cell line set. *Cancer research*, 72(14):3499–3511, 2012.

[145] G.V. Rocha, P. Zhao, and B. Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE), 2008.

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

[147] V. Roth and B. Fischer. The group-Lasso for generalized linear models: Uniqueness of solutions and efficient algorithms. In *ICML '08: Proceedings of the 25th international conference on Machine Learning*, pages 848–855, 2008.

[148] A. Rothman, E. Levina, and J. Zhu. Sparse multivariate regression with covariance estimation. *J. Comp. Graph. Stat.*, 19(4):947–962, 2010.

[149] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, 4(1), 2005.

[150] M. Schmidt. Least squares optimization with l1-norm regularization, 2005.

[151] L. Schwaller, S. Robin, and M. Stumpf. Bayesian inference of graphical model structures using trees. *arXiv preprint arXiv:1504.02723*, 2015.

[152] R.C. Serlin and J.R. Levin. Teaching how to derive directly interpretable coding schemes for multiple regression analysis. *Journal of Educational Statistics*, 10(3):223–238, 1985.

[153] P. Shannon. *MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs*, 2013. R package version 1.4.0.

[154] T. Shimamura, S. Imoto, R. Yamaguchi, A. Fujita, M. Nagasaki, and S. Miyano. Recursive regularization for inferring gene networks from time-course gene expression profiles. *BMC Systems Biology*, 3(41), 2009.

[155] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Comput. Graph. Statist.*, 22(2):231–245, 2013.

[156] M. Slawski, W. zu Castell, and G. Tutz. Feature selection guided by structural information. *Ann. Appl. Stat.*, 4:1056–1080, 2010.

[157] T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

[158] K.A Sohn and S. Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. *JMLR*, W&CP(22):1081–1089, 2012.

[159] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, page ass043, 2012.

[160] T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.*, 14(1):3385–3418, 2013.

[161] M. Szafranski, Y. Grandvalet, and P. Morizet-Mahoudeaux. Hierarchical penalization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[162] M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. Composite kernel learning. *Machine Learning*, 79(1-2):73–103, 2010.

[163] K.M. Tan, P. London, K. Mohan, S.-I. Lee, M. Fazel, and D. Witten. Learning graphical models with hubs. *J. Mach. Learn. Res.*, 15(1):3297–3331, 2014.

[164] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

[165] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R.J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. R. Statist. Soc. B*, 74(2):245–266, 2012.

[166] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, 67:91–108, 2005.

[167] R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *Ann. Statist.*, 39(3):1335–1371, 2011.

[168] R.J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *Electron. J. Stat.*, 2012.

[169] H. Toh and K. Horimoto. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics*, 18:287–297, 2002.

[170] J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52:1030–1051, 2006.

[171] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

[172] N. Verzelen. Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. *Elec. Journal Stat.*, 6:38–90, 2012.

[173] N. Villa-Vialaneix, M. Vignes, N. Viguerie, and M. San Cristobal. Inferring networks from multiple samples with concensus LASSO. *Quality Technology and Quantitative Management*, 11(1):39–60, 2014.

[174] F. Villers, B. Schaeffer, C. Bertin, and S. Huet. Assessing the validity domains of graphical Gaussian models in order to infer relationships among components of complex biological systems. *Stat. Appl. Genet. Mol. Biol.*, 7(2), 2008.

[175] J. Vogt and V. Roth. A complete analysis of the l_1, p group-lasso. *ArXiv preprint arXiv:1206.4632*, 2012.

[176] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55, 2009.

[177] H. Wang, A. Banerjee, C.-J. Hsieh, P. K Ravikumar, and I.S. Dhillon. Large scale distributed sparse precision estimation. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 584–592. Curran Associates, Inc., 2013.

[178] J. Wang, J. Zhou, P. Wonka, and J. Ye. Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1070–1078, 2013.

[179] F. L Wauthier, N. Jojic, and M.I. Jordan. A comparative framework for preconditioned lasso algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1061–1069, 2013.

[180] J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley, 1990.

[181] A. Wille and P. Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Stat. Appl. Genet. Mol. Biol.*, 5(1), 2006.

[182] D.M. Witten, J.H. Friedman, and N. Simon. New insights and faster computations for the graphical lasso. *J. Comput. Graph. Statist.*, 20(4):892–900, 2011.

[183] D.M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high -dimensional problems. *J. R. Statist. Soc. B*, 71(3):615–636, 2009.

[184] D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

[185] R. Witten and E. Candes. Randomized algorithms for low-rank matrix factorizations: sharp performance bounds. *Algorithmica*, pages 1–18, 2013.

[186] Eleanor Wong, Awate Suyash, and Fletcher Thomas. Adaptive sparsity in Gaussian graphical models. In *In Proceedings of the 30th International Conference on Machine Learning*, pages 311–319. 2013.

[187] H. Xu, C. Caramanis, and S. Mannor. Outlier-robust pca: The high-dimensional case. *Information Theory, IEEE Transactions on*, 59(1):546–572, 2013.

[188] E. Yang, P. Ravikumar, G.I. Allen, and Z. Liu. On poisson graphical models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1718–1726, 2013.

[189] J. Yin and H. Li. A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.*, 5:2630–2650, 2011.

[190] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68(1):49–67, 2006.

[191] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[192] M. Yuan and Y. Lin. On the non-negative garrotte estimator. *J. R. Statist. Soc. B*, 69(2):143–161, 2007.

[193] X.-T. Yuan and T. Zhang. Partial Gaussian graphical model estimation. *Information Theory, IEEE Transactions on*, 60(3):1673–1687, 2014.

[194] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Stat.*, 37(6):3468–3497, 2009.

[195] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

[196] T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. *huge: High-dimensional Undirected Graph Estimation*, 2014. R package version 1.2.6.

[197] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

[198] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67(2):301–320, 2005.

[199] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[200] H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Stat.*, 35(5):2173–2192, 2007.