Thresholding in high dimensional statistics: an application in testing and Cosmology

Diaz Rodriguez, Jairo

UNIVERSITÉ DE GENÈVE
Section de mathématiques

FACULTÉ DES SCIENCES
Professeur Sylvain SARDY

THRESHOLDING IN HIGH DIMENSIONAL STATISTICS: AN APPLICATION IN TESTING AND COSMOLOGY

THÈSE

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention mathématiques.

par

Jairo DIAZ-RODRIGUEZ

de

Bucaramanga (COLOMBIE)

Thèse n° 5219

GENÈVE

2018

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

DOCTORAT ÈS SCIENCES, MENTION MATHEMATIQUES

Thèse de Monsieur Jairo DIAZ RODRIGUEZ

intitulée :

«Thresholding in High Dimensional Statistics: An application in Testing and Cosmology»

La Faculté des sciences, sur le préavis de Monsieur S. SARDY, professeur associé et directeur de thèse (Section de mathématiques), Madame E. CANTONI, professeure associée (Faculté d'économie et de management, Research Center for Statistics), Monsieur R. VON SACHS, professeur ordinaire (Louvain School of Statistics, Biostatistics and Actuarial Sciences, Faculté des sciences, Université catholique de Louvain, Louvain-La-Neuve, Belgique) et Monsieur D. ECKERT, docteur (Max Planck Institute for Extraterrestrial Physics, Munich, Germany), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 18 mai 2018

Thèse - 5219 -

Le Doyen

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Acknowledgments

First of all, I would like to express my gratitude to my advisor, Sylvain Sardy, for his advice, patience, comprehension, and for trusting in me.

I wish to thank Professors Eva Cantoni, Dominique Eckert and Rainer von Sachs for kindly agreeing to serve on my thesis committee. A special thank to the latter who sent me a detailed list of comments and questions that improved the quality of my thesis.

I acknowledge the Swiss National Science Foundation, the Section of Mathematics of the University of Geneva, and the Foundation for the future of Colombia COLFUTURO for supporting and funding my research.

I also wish to thank Dominique Eckert for presenting me the interesting Cosmology problem in Chapter 3, and for answering all my questions regarding the astrophysics behind it.

I would like to thank Pascaline Descloux for helping me to prepare the slides for my thesis defense, and all the discussions during our “thresholding seminars”.

More personally, I would like to say thank you again to Sylvain. Thank you very much for your advices and your guidance during all these years. Every conversation and discussion was always fruitful, and allowed me to grow during all my research. You are not just an excellent professor but also an incredible human being. Thank you for every teaching and for every coffee. You can count always with my gratitude and my friendship. You still owe me one snowshoeing and a visit to Colombia.

Thank you to my mother Luz Amparo, my father Jairo and my brother Julián in Colombia. It is not easy to be so far away from you, but I felt every day your love and your support through your words and your prayers. Mom and dad, you were my first

teachers and there is no university better than home, and no professor better than you.

Thank you to all my family in Colombia: My grandparents Pedro, Carmen and Nubia, my parents in law Gonzalo and Martha, my brothers in law Nicolás and Mateo, all my uncles, aunts and cousins. I know I was in your thoughts, thank you for remembering me. I am proud of my family and proud of my country.

Thank you to my Colombian "Eurofamily": Andres, Giovanno, Alessandra, Miguel; and thank you to my friends in Colombia: Pedro, Oscar, Silvia, Mario and Eva. There are more travels and more parties to live together. Thank you to my Geneva family: Chris, Paula, Emma, Vivian, Alex, Ericka, Hannah, Priscilla, Brian, Sarah, Joshua. It has been a blessing to be here with you.

I want to end with the most important persons in my life: Cindy and Violeta.

Cindy, there are no words to thank you. You are my best friend, my travel companion, my favorite researcher, and much more. Everything started with a young Colombian couple flying to Switzerland for a master's degree, with two suitcases and a bag full of dreams. Now we are here with two master degrees and two PhD's. Even more important, we are here with more love than ever, a bigger bag of dreams, and a beautiful baby. Thank you Cindy for being a support and an encouragement in the bad moments, and a friend to share and enjoy the good ones. You have been, you are and you will always be the love of my life.

My little and beautiful baby Violeta, I hope you can read this when you grow. You were here in my lap during the last months of my PhD. and you made me write a lot of typos, and maybe make some mistakes. Nevertheless, I want to thank you because if you wouldn't have been such a well behaved baby, I wouldn't have been able to write this thesis. Taking care of you while writing a thesis has been the most beautiful experience in my life, and I wouldn't change it, I love you.

Cindy and Violeta, a PhD. is nothing compared to the joy that you bring to my life. I am proud of both you.

Finally, I want to thank God, the reason for everything in my life.

Gracias a todos,

Jairo

Contents

Résumé

Beaucoup de problèmes du monde réel se concentrent sur l'identification de l'effet que certaines variables mesurées ont sur une réponse d'intérêt. Un point commun entre les dispositifs d'enregistrement modernes est que le nombre P de variables mesurées, qui correspondent souvent au nombre de paramètres, est grand. Il devient donc fréquent analyser les données où P dépasse la taille de l'échantillon N. Ceci est généralement appelé *statistiques de grande dimension*.

Dans cette thèse, nous commençons par examiner une famille de techniques de sélection de modèles appelées *seuillage* qui supposent que le vecteur de paramètres $\boldsymbol{\beta}$ a peu de coefficients non nuls. Cette hypothèse est appelée *sparsité*. Ces estimateurs sont indexés par un paramètre de régularisation λ. Nous passons en revue les concepts importants liés au seuillage tels que la fonction *zero thresholding function* et la *null thresholding statistique* qui sont utiles dans les chapitres 2 et 3.

Dans le chapitre 2, nous appliquons certaines propriétés des estimateurs de seuillage pour dériver une nouvelle classe de tests statistiques pour les modèles linéaires généralisés. Ces tests peuvent être utilisés si le modèle inclut plus de paramètres que d'observations ou non. Pour les modèles linéaires, les tests de seuillage reposent sur des statistiques pivotales issues des techniques de sélection de modèles. Pour les modèles linéaires généralisés, nous avons dérivé des tests qui s'appuient sur de nouvelles statistiques asymptotiquement pivotales. Un test de seuillage composite tente d'obtenir uniformément la plus grande puissance possible sous des alternatives à la fois sparses et denses. Dans une simulation, nous comparons le niveau et la puissance de ces tests sous des hypothèses alternatives sparses et denses, ainsi que l'effet de la sparsité dans la matrice de régression. Les tests de seuillage ont un meilleur contrôle du niveau nominal et une puissance plus élevée que les tests existants.

Dans le chapitre 3 nous utilisons un estimateur de seuillage pour résoudre un

problème de cosmologie qui consiste à estimer l'émissivité 3D de gas d'un amas de galaxies à partir d'une image 2D prise par un télescope. Un phénomène de floutage et des sources ponctuelles rendent ce problème inverse encore plus difficile à résoudre. Pour imposer la sparsité sur les paramètres dans l'esprit du lasso, on régularise l'estimation du maximum de vraisemblance avec deux pénalités ℓ_1 : une pour l'estimation de l'émissivité radiale et une pour la détection des sources ponctuelles. Les deux pénalités de type lasso sont choisies sur une échelle probabiliste similaire au niveau d'un test statistique. Nous quantifions également l'incertitude de l'estimation avec une approche de type bootstrap pour guider les analystes dans l'évaluation de l'importance de caractéristiques intéressantes. Nous effectuons des simulations dans lesquelles nous montrons comment notre méthodologie surpasse en termes d'erreurs quadratiques moyennes l'approche actuelle qui est en deux étapes, et comment elle a une bonne probabilité de couverture. Nous appliquons nos méthodes à cinq images réelles de télescope et discutons les résultats scientifiques.

Summary

Many real world problems focus on identifying the effect that some measured variables have on a response of interest. A shared pattern between modern recording devices is that the number of measured variables P, which often correspond to the number of parameters, is large. It is therefore getting common to analyze data where P exceeds the sample size N. This is usually called *high dimensional statistics*.

In this thesis we start by reviewing a family of model selection techniques called *thresholding* that assume the vector of parameters $\boldsymbol{\beta}$ has few non-zero coefficients. This assumption is called *sparsity*. These estimators are indexed by a regularization parameter λ. We review important concepts related to thresholding such as the *zero thresholding function* and the *null thresholding statistic* that are useful in Chapters 2 and 3.

In Chapter 2 we apply these properties of the thresholding estimators to derive a new class of statistical tests for generalized linear models. These tests can be employed whether the model includes more parameters than observations or not. For linear models, thresholding tests rely on pivotal statistics derived from model selection techniques. For generalized linear models we derived tests that rely on new asymptotically pivotal statistics. A composite thresholding test attempts to achieve uniformly most power under both sparse and dense alternatives with success. In a simulation, we compare the level and power of these tests under sparse and dense alternative hypotheses, as well as the effect of sparsity in the design matrix. The thresholding tests have a better control of the nominal level and higher power than existing tests.

In Chapter 3 we use a thresholding estimator to solve a cosmology problem that consists in recovering the 3D gas emissivity of a galaxy cluster from a 2D image taken by a telescope. Blurring and point sources make this inverse problem even harder

to solve. To enforce sparsity on the parameters in the spirit of lasso, we regularize the maximum likelihood estimation with two ℓ_1 penalties: one for the estimation of the radial emissivity and one for the detection of the point sources. The two lasso penalties are chosen on a probabilistic scale similarly to the level of a statistical test. We also quantify the uncertainty of the estimation with bootstrap to guide analysts in judging the significance of interesting features. We perform simulations in which we show how our methodology outperforms the current state-of-the-art two-step approach in terms of mean squared errors, and how it has good coverage probability. We apply our methods to five different real telescope images and discuss the scientific findings.

Prologue

This thesis is divided into three chapters. The first chapter is the introduction. It summarizes the main concepts that are common to the applications in Chapter 2 and Chapter 3. These two chapters are the basis of this thesis and present two different applications of *thresholding estimators* and the *quantile universal threshold*. It is important for the reader to notice that both the introduction and Chapter 2 share the same notation. Since the cosmology application in Chapter 3 has a more specific interpretation, we kept a different notation for the variables in this chapter to be more consistent with the astrophysics community.

This thesis is the continuation of the work done during my Master's thesis under the direction of Professor Sylvain Sardy and with Dr. Caroline Giacobino on the selection of the penalty parameter for thresholding estimators, namely the *quantile universal threshold*. We published the main article that became the inspiration for the two applications exposed in this thesis:

- C. Giacobino, S. Sardy, J. Diaz-Rodriguez, and N. Hengartner. Quantile Universal Threshold. *Electronic Journal of Statistics*, 2017.

In fact, some parts of the introduction of this thesis summarize the main points of this paper. This work leads to a consulting work with Dr. Dorothea Hug Peter from the Laboratory of Ecology and Aquatic Invertebrates of the University of Geneva, resulting in another paper:

- D. Hug Peter, S. Sardy, J. Diaz-Rodriguez, E. Castella, and V. Slaveykova. Modeling whole body trace metal concentrations in aquatic invertebrate communities: A trait-based approach. *Environmental Pollution*, 2018.

The study of the quantile universal threshold inspired the work of Chapter 2, an application in hypothesis testing based on thresholding tests, another consulting work

with Dr. Dominique Eckert from the Department of Astrophysics of the University of Geneva inspired the second main part of this thesis: the cosmology application in Chapter 3. The work of both chapters produced two papers currently in peer review process:

- S. Sardy, C. Giacobino, and J. Diaz-Rodriguez. Thresholding based tests. Sent to *Biometrika*, 2018.
- J. Diaz-Rodriguez, D. Eckert, H. Monajemi, S. Paltani, S. Sardy. Nonparametric estimation of galaxy cluster's emissivity and point source detection in astrophysics with two lasso penalties. Sent to *Annals of Applied Statistics*, 2017.

We also published a proceedings in the IEEE-Xplore database with the cosmology application:

- J. Diaz-Rodriguez and S. Sardy. A Composite Lasso Penalty With an Application in Cosmology. *IEEE Intl Conference on Computational Science and Engineering (CSE)*, 2016.

Another important contribution is the R library we developed for the quantile universal threshold and the thresholding tests in Chapter 2:

- J. Diaz-Rodriguez, S. Sardy, C. Giacobino, and N. Hengartner. qut: Quantile Universal Threshold, 2016. *URL https://CRAN.R-project.org/package=qut. R package version 1.3*.

Chapter 1

Introduction

Several real world applications involve the study of the relationship between a set of covariates and a dependent variable of interest by processing a set of measurements. In Chemometrics, Sardy [2008] deals with a problem in which $P = 315$ covariates $\mathbf{x}$ are spectrometer measurements and the responses Y are the octane level of $N = 434$ fuel samples: predicting the octane level from its cheap spectrometer measurements can save time and money in comparison with tedious and time consuming mechanical techniques. In Sociology, Kushmerick [1999] data set is a classification of $N = 2359$ possible advertisements on Internet pages based on $P = 1430$ features: the goal is to identify the most significant features in this classification problem for future predictions. In Genetics, Golub et al. [1999] study the expression of $P = 3571$ genes of $N = 72$ samples of humans with different types of acute leukemia cancer: they want to select the specific genes that determine the different types of leukemia cancer. Also in genetics, Bühlmann et al. [2014] are interested in identifying the set of genes that are significantly related to the riboflavin production rate from a population of Bacillus subtilis. They have measurements from 71 individuals and expressions from 4088 genes. The purpose of collecting such data is to relate the covariates $\mathbf{x} = (x_1, \ldots, x_P)$ to dependent variable Y. Up to model and measurement errors, it is believed that a function $\mu(\mathbf{x})$ can predict Y well.

1.1 Generalized Linear Models

Generalized Linear Models (GLMs) provide a framework to model data as discussed above. In such models the association $\mu(\mathbf{x})$ is believed to be a function of a linear

combination of the covariates. The goal is to estimate the parameters of this function based on a set of training measurements $\{(y_n, \mathbf{x}_n)\}_{n=1,\ldots,N}$. If P grows with the number N of samples, the method is nonparametric in the sense that the underlying model increases in complexity, as the sample size increases. If P is fixed, then the model is parametric. GLMs encompass normal linear models, logistic regression for binary responses, Poisson regression for count data and log-linear models for contingency tables. GLMs include many more possibilities in a class of distributions in the exponential family, taking the form

$$f_{Y_n}(y_n; \theta_n, \phi) = \exp\left(\frac{y_n\theta_n - b(\theta_n)}{a(\phi)} + c(y_n, \phi)\right), \tag{1.1}$$

where $\theta_n \in \Theta := \{\theta \in \mathbb{R} \,|\, b(\theta) < \infty\}$, $y_n \in \mathcal{Y} \subset \mathbb{R}$, $n = 1, 2, \ldots, N$, and the functions a, b and c are known. For the Gaussian distribution, for instance, $a(\phi) = \sigma^2$, $b(\theta_n) = \theta_n^2/2$, and $c(y_n, \phi) = -(y_n^2/\sigma^2 + \log(2\pi\sigma^2))/2$. Moreover, GLMs assume the mean of the dependent variable Y_n denoted by μ_n is linked to the covariates $\mathbf{x}_n$ through a linear term and a function g according to

$$\mu_n := E[Y_n \mid \mathbf{x}_n] = b'(\theta_n) \tag{1.2}$$

$$\boldsymbol{\mu} = g^{-1}(\beta_0\mathbf{1} + X\boldsymbol{\beta}), \tag{1.3}$$

where $\boldsymbol{\beta}$ is the $P \times 1$ vector of parameters and β_0 is the intercept. We note $\mathbf{y}$ and X the vector formed by y_n, and the matrix with rows $\mathbf{x}_n^T$, for $n = 1, \ldots, N$, respectively, where $\mathbf{y}$ is a realization of a random variable $\mathbf{Y}$. For instance (1.2) and (1.3) lead to the Normal linear model when g is the identity function and the distribution is Gaussian. Commonly, GLMs assume the *canonical form* (i.e $\theta_n = \beta_0 + \mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}$ in (1.2)) in which case g is called the *canonical link function*. Many applications assume the link function is the canonical one. There are also several applications in which a different link function is assumed. For binomial distribution, Bliss [1934] proposes the probit model where the probit function is used instead of the canonical logit function. Huettmann [2003] also shows alternatives to the traditional logit approach using probit and the complementary log log link function in an application of nesting location of birds. In Poisson regression, the identity link is particularly useful in epidemiology for modeling the way disease incidence is related to covariates [Benichou and Palta, 2005]. It is also important as an approximate method for fitting identity link binomial models for risk and prevalence differences [Spiegelman and Hertzmark, 2005] and it is as well highly used in x-ray astronomy to model photon counting experiments [Cash, 1979]. For Single Index Model [Sharpe, 1963], the link function is estimated

as well. We present in this thesis two applications in which we use GLMs with link function different from the canonical one. In Chapter 2 we derive statistical tests that amount to GLMs with new link functions. In Chapter 3 we show a Cosmology application where a GLM for Poisson distribution uses the identity link function.

GLMs serve several purposes depending on the application, but we highlight three of them. One is prediction, that is find models with good predictive accuracy. For instance to classify tumor subtypes in an early phase of a disease, or to predict the octane level of a fuel sample from its cheap spectrometer measurements [Sardy, 2008]. A second purpose, called model selection, is to identify the relevant variables that carry information to predict $\mathbf{y}$. For instance, identify which genes are more significant to predict a type of leukemia cancer in Golub et al. [1999]. A third purpose of fundamental interest in Statistics is testing the significance of the parameters $\boldsymbol{\beta}$. It is for instance of particular interest to test the null hypothesis

$$H_0 : \boldsymbol{\beta} = \mathbf{0}, \tag{1.4}$$

against the alternative $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$. This amounts for instance to test whether a microarray with some gene expressions carries any information to predict a certain disease through a linear model. In this thesis we use GLMs for all of this three purposes: testing in Chapter 2, and prediction and model selection in Chapter 3.

1.2 Regularizing when P is larger than N

A classical approach to estimate the coefficients $[\beta_0 \ \boldsymbol{\beta}]$ is to maximize the negative log-likelihood defined by

$$-l\left(\beta_0, \boldsymbol{\beta}\right) = C \sum_{n=1}^{N} \left[y_n \theta_n - b\left(\theta_n\right)\right], \tag{1.5}$$

with the relationship between $(\beta_0, \boldsymbol{\beta})$ and θ_n given in (1.2) and (1.3), and C a constant value. For Gaussian linear models, this equation is equivalent to the residual sum of squares $\text{RSS} = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$. In the case of testing, one of the most popular procedures for linear models is Fisher's F-test, and in GLMs the likelihood ratio or deviance tests.

A common pattern between modern recording devices is that the number of covariates P measured per sample is large. It is therefore getting common to analyze data where P exceeds the sample size N, which makes the task of Statistics more difficult, and some of the classical methods can no longer be employed. This is commonly

called *high dimensional statistics*. The maximum likelihood estimation principle falls apart for many reasons in this situations, and classical tests such as Fisher's F-test can not be used. Motivated by the seminal papers of Tikhonov [1963] (for dealing with the situation $P > N$) and James and Stein [1961], a considerable amount of literature has concentrated over the last fifty years on the estimation of the coefficients $\boldsymbol{\beta}$ by regularization techniques that aim at decreasing the variance by introducing some bias for a better prediction error. One very interesting assumption is to regularize by imposing sparsity in the coefficients. This means, the dependent variable depends on just some covariates (or few non-zero coefficients). Sparsity is natural in several applications, for instance in Genetics where only a few genes are assumed to be responsible of a disease. Sparsity assumes that only a few covariates (much less than N) compose the model to predict the response by a linear association. The goal of model selection then becomes tightly connected with the goal of testing: identify the correct covariates, or at least identify a model (not too big) that includes the correct model.

1.3 Thresholding

1.3.1 Thresholding for point estimation

The concept of *sparsity* allows to introduce a special class of regularization techniques called *thresholding* in the sense that:

- they assume that the true parameter $\boldsymbol{\beta}$ is *sparse*, meaning

$$\mathcal{S} := \{q \in \{1, \dots, Q\} : \beta_q \neq 0\} \tag{1.6}$$

  has small cardinality.

- result in an estimated support

$$\hat{\mathcal{S}}_\lambda := \{q \in \{1, \dots, Q\} : \hat{\beta}_{\lambda,q} \neq 0\} \tag{1.7}$$

  whose cardinality is governed by the choice of a threshold parameter $\lambda \geq 0$.

Thresholding techniques are employed in various settings such as linear regression [Donoho and Johnstone, 1994, Tibshirani, 1996], Generalized Linear Models [Park and Hastie, 2007], low-rank matrix estimation [Mazumder et al., 2010, Cai et al.,

2010], density estimation [Donoho et al., 1996, Sardy and Tseng, 2010], linear inverse problems [Donoho, 1995], compressed sensing [Donoho, 2006, Candès and Romberg, 2007] and time series [Neto et al., 2012]. This thesis uses thresholding techniques in two different settings: testing in Chapter 2 and solving a Cosmology linear inverse problem in Chapter 3. A famous example of *thresholding estimator* in linear regression is lasso [Tibshirani, 1996] which calculates

$$[\hat{\beta}_0 \ \hat{\boldsymbol{\beta}}_\lambda] \in \underset{[\beta_0 \ \boldsymbol{\beta}]\in\mathbb{R}^{P+1}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta} - \beta_0\mathbf{1}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{1.8}$$

for a given $\lambda > 0$. Inspired by lasso, Belloni et al. [2011] proposed the square root lasso by substituting the quadratic loss by its square root:

$$[\hat{\beta}_0 \ \hat{\boldsymbol{\beta}}_\lambda] \in \underset{[\beta_0 \ \boldsymbol{\beta}]\in\mathbb{R}^{P+1}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta} - \beta_0\mathbf{1}\|_2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{1.9}$$

The advantage of square root lasso over lasso is that it is pivotal in the sense that it neither relies on the knowledge of the standard deviation σ or nor does it need to pre-estimate it. The extension of lasso to Generalized Linear Models replaces the quadratic loss by the negative log-likelihood, up to constants [Park and Hastie, 2007]:

$$[\hat{\beta}_0 \ \hat{\boldsymbol{\beta}}_\lambda] \in \underset{[\beta_0 \ \boldsymbol{\beta}]\in\mathbb{R}^{P+1}}{\operatorname{argmin}} \sum_{n=1}^{N} \left[y_n\theta_n - b\left(\theta_n\right)\right] + \lambda\|\boldsymbol{\beta}\|_1. \tag{1.10}$$

Notice that we used the symbol "$\in$" instead of "$=$" in this three equations to point out that uniqueness of the solution is not guaranteed. These estimators force sparsity in $\boldsymbol{\beta}$ and the amount of sparsity is controlled by λ. Assuming group sparsity, group lasso [Yuan and Lin, 2006] and its extension to Generalized Linear Models, as well as the group square root lasso [Bunea et al., 2014], replace the penalty term in (1.8), (1.10) and (1.9) by $\lambda\sum_{k=1}^{M}\|\boldsymbol{\beta}_{G_k}\|_2$, where $\{G_1, \cdots, G_M\}$ is a partition of $\{1, \cdots, P\}$. Many other thresholding estimators have been proposed for linear models: best subset selection, least absolute deviation (LAD) lasso [Wang et al., 2007], Dantzig selector [Candès and Tao, 2007], Subbotin lasso [Sardy, 2009], smoothly clipped absolute deviation (SCAD) [Fan and Peng, 2004], minimax concave penalty (MCP) [Zhang, 2010], smooth lasso [Sardy, 2012], among others, some of which can also be also extended to GLMs. The challenge of these estimators lies in identifying the active subset $\mathcal{S}$ (1.6), which amounts to selecting basis coefficients in wavelet denoising, or in some cancer research applications for instance, identifying what genes are responsible for cancer. The selection of the penalty parameter λ is crucial: a too

large λ results in a simplistic model missing important features whereas a too small λ leads to a model including many features outside the true model. A typical goal is to determine λ such that with high probability the selected model $\hat{\mathcal{S}}_\lambda$ satisfies

$$\hat{\mathcal{S}}_\lambda \supseteq \mathcal{S}, \tag{1.11}$$

along with few false detections $\{q : \hat{\beta}_{\lambda,q} \neq 0,\ \beta_q = 0\}$. This property is called *variable screening*. For a suitably chosen λ, certain estimators allow variable screening. The optimal threshold for model identification often differs from the threshold aimed at prediction optimality [Yang, 2005, Leng et al., 2006, Meinshausen and Bühlmann, 2006, Zou, 2006]. Several methodologies have been proposed to select this parameter λ, for instance the Quantile Universal Threshold [Giacobino et al., 2017], or the more classical approaches that consist in minimizing a criterion like cross-validation, AIC [Akaike, 1998], BIC [Schwarz, 1978] or Stein unbiased risk estimation (SURE) [Stein, 1981], among others.

1.3.2 Thresholding for testing linear models

A key property shared by a class of estimators is to set the estimated parameters to zero for a sufficiently large but finite threshold λ, this is $\hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y}) = \mathbf{0}$ for a certain value of λ large enough. The smallest λ that satisfies this property is a function $\lambda_0(\mathbf{Y})$ of the data called *zero thresholding function* [Giacobino et al., 2017], as formalized in the following definition.

Definition 1.1 *A thresholding estimator $\hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y})$ admits a* zero-thresholding function *$\lambda_0(\mathbf{Y})$ if*

$$\hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y}) = \mathbf{0} \quad \Leftrightarrow \quad \lambda \geq \lambda_0(\mathbf{Y}) \quad \textit{almost everywhere}. \tag{1.12}$$

Some thresholding estimators have this property, and for some of them the *zero-thresholding function* has a closed form expression. For instance, the *zero-thresholding function* for lasso (1.8) is

$$\lambda_0(\mathbf{y}) = \|X^{\mathrm{T}}(\mathbf{y} - \bar{y}\mathbf{1})\|_\infty, \tag{1.13}$$

and for square root lasso (1.9),

$$\lambda_0(\mathbf{y}) = \frac{\|X^{\mathrm{T}}(\mathbf{y} - \bar{y}\mathbf{1})\|_\infty}{\|\mathbf{y} - \bar{y}\mathbf{1}\|_2}. \tag{1.14}$$

By observing the equivalence (1.12) in Definition 1.1 between setting all coefficients in a thresholding estimator to zero, and the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$, Sardy et al. [2018] define a *thresholding test*.

Definition 1.2 *Let $\hat{\boldsymbol{\beta}}_\lambda(\mathbf{Y})$ be a thresholding estimator of the linear model* (1.2) *and* (1.3)*, with zero thresholding function $\lambda_0(\mathbf{Y})$. Letting the* null-thresholding statistic*.*

$$\Lambda_0 := \lambda_0(\mathbf{Y}_0) \quad \textit{with} \quad \mathbf{Y}_0 =_d \mathbf{Y} \quad \textit{under} \quad H_0 : \boldsymbol{\beta} = \mathbf{0}, \tag{1.15}$$

then a test function of the form

$$\phi(\mathbf{y}) = \begin{cases} 1 & \text{if } \lambda_0(\mathbf{y}) > \lambda_\alpha \\ 0 & \text{otherwise} \end{cases}, \tag{1.16}$$

defines a thresholding test*, where $\lambda_\alpha = F^{-1}_{\Lambda_0}(1-\alpha)$ is a* test-threshold *of level α.*

This is called a *thresholding test* since it is based on a thresholding estimator through its zero-thresholding function $\lambda_0(\mathbf{Y})$. The test-threshold λ_α can be evaluated for instance by Monte Carlo simulation. For example, using the lasso and its zero thresholding function (1.13), the test-threshold can be evaluated simulating M vectors $\mathbf{y}_0^{(1)}, \ldots, \mathbf{y}_0^{(M)}$ from $\mathbf{Y}_0$ under H_0, calculating the corresponding $\lambda^{(m)} = \|X^{\mathrm{T}}(\mathbf{y}_0^{(m)} - \bar{y}_0^{(m)}\mathbf{1})\|_\infty$ for $m = 1, \ldots, M$ and taking the upper α-quantile. The larger M the more precision on λ_α. One easily sees that test (1.16) has the desired level by choosing $\lambda_\alpha = F^{-1}_{\Lambda_0}(1-\alpha)$ where F_{Λ_0} is the distribution of $\Lambda_0 = \|X^{\mathrm{T}}(\mathbf{Y}_0 - \bar{Y}_0\mathbf{1})\|_\infty$ and $\mathbf{Y}_0 =_d \mathbf{Y}$ under H_0. Notice that in linear models, $\mathbf{Y}_0 = \beta_0 + \sigma\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(0,1)$, then $\mathbf{Y}_0 - \bar{Y}_0\mathbf{1} = (I - \frac{1}{n}M)\mathbf{Y} = \sigma(I - \frac{1}{n}O)\boldsymbol{\epsilon}$ and $\|\mathbf{Y}_0 - \bar{Y}_0\mathbf{1}\|_2 = \sigma\|(I - \frac{1}{n}O)\boldsymbol{\epsilon}\|_2$, where O is a $N \times N$ matrix with all entries equal to one. Therefore the statistic $\|X^{\mathrm{T}}(\mathbf{y} - \bar{y}\mathbf{1})\|_\infty$ is pivotal under H_0 with respect to the intercept β_0, but is not pivotal with respect to σ. On the contrary, the null thresholding statistic for square root lasso $\|X^{\mathrm{T}}(\mathbf{y} - \bar{y}\mathbf{1})\|_\infty / \|\mathbf{y} - \bar{y}\mathbf{1}\|_2$ is pivotal under H_0 to both nuisance parameters β_0 and σ. This property is useful to obtain the desired level α regardless of the underlying unknown nuisance parameters. Notice that thresholding tests can be used in linear models even if $P > N$, where classical Fisher's F-test can not be used. Moreover, Section 2.3 shows that they are more powerful under sparse alternatives. Classical likelihood ratio tests also fail when P larger than N. For GLMs, Giacobino et al. [2017] show that the zero thresholding function for (1.10) using the canonical link function is the same as that of lasso (1.13). This statistic is not pivotal with respect to the intercept parameter β_0 in GLMs, since the variance of $\mathbf{Y}_0 - \bar{Y}_0\mathbf{1}$ depends on β_0 in some exponential distributions (i.e Poisson and binomial). In Chapter 2 we extend thresholding tests by deriving test statistics that are asymptotically pivotal.

1.4 Overview of this thesis

The challenging problem of testing GLMs when $P > N$ is addressed in Chapter 2. Thresholding tests in linear models have the advantage over classical tests like Fisher's F-test that they can be employed even if $P > N$ through pivotal test statistics. In GLMs, classical testing methods, such as likelihood ratio and deviance tests, also fail when $P > N$ and have a poor control of the level, even when P is large but not higher than N. Inspired by the good results for linear models, we extend thresholding tests to GLMs by deriving new thresholding test statistics. We prove that our test statistics are asymptotically pivotal to all nuisance parameters under the null model. We show through simulations that our method better controls the level of the test and has better power than classical and more modern techniques.

In Chapter 3 we consider an inverse problem in Cosmology. The goal is to estimate two cosmological objects, an emissivity function and a matrix of point sources, from an X-ray emission image of photon counts. We model data as a GLM for Poisson distribution with identity link function. For the estimation of the parameters, we consider the lasso for GLM with two penalty parameters, one for the emissivity and other for the point sources. We derive the zero thresholding function for this estimator, and we use it to choose the two penalties based on the quantile universal threshold [Giacobino et al., 2017]. This optimization problem is high dimensional and non differentiable. Commonly used methods to choose λ such as cross validation are computationally inefficient and have no straightforward interpretation for the astronomers, therefore the advantage of using the quantile universal threshold. Our estimator outperforms the state-of-the-art methodology currently employed in Cosmology, in terms of mean square error and coverage probability in simulated data. We show the performance of our methodology with five real X-ray emission images.

Chapter 2

Testing in Generalized Linear Models

This chapter proposes tests based on new asymptotic pivotal statistics for Generalized Linear Models, and is organized as follows. First, Section 2.1 presents the settings of the problem and shows the motivation to our approach. Section 2.2 considers asymptotic tests for Generalized Linear Models. Section 2.2.1 proposes a new asymptotic pivot, while Section 2.2.2 shows its connection with the zero thresholding function for a specific link function. Section 2.2.3 illustrates our methodology with an example and Section 2.2.4 shows how to combine several tests in a single one and describes the composite $\oplus$-test between lasso and group lasso. Since our test statistics are asymptotically pivotal and do not have exact level with finite samples, Section 2.2.5 describes two methodologies that allow to obtain tests closer to the nominal level. Section 2.3 presents the results of a simulation study to observe the behavior of our tests. To compare the new thresholding tests to existing tests, Section 2.3.1 performs power analyses in low- and high-dimensional settings for Gaussian, binomial and Poisson data. Section 2.3.2 gives an insight into the asymptotic of our tests and the effect of sparsity in matrix X. Inspired in metagenomics data, with an X matrix with up to 70% of zero entries, Section 2.3.3 presents a power analysis to compare the nominal level tests from Section 2.2.5. Section 2.4 concludes by giving recommendations on what test to use. The research is reproducible and codes are available in the `qut` package in `R` [Diaz-Rodriguez et al., 2016].

2.1 Motivation

The primary goal in this chapter is testing $\boldsymbol{\beta}$ based on lasso for GLMs. In linear models with $P < N$, Fisher's F-test is widely applied and based on the statistic

$$\frac{(\mathrm{RSS}_{H_0} - \mathrm{RSS})/R}{\mathrm{RSS}/(N-P)} \sim F_{R,N-P} \tag{2.1}$$

that is pivotal under $H_0 : A\boldsymbol{\beta} = \mathbf{c}$, where A is an $R \times P$ full row rank matrix, and RSS_{H_0} and RSS are the residual sum of squares under the null model and the full models, respectively. We contend that one drawback of the F-test is that it is based on an indirect measure of the coefficients $\boldsymbol{\beta}$ through the predictive measure of $\mathbf{Y}$ that is RSS. Arias-Castro et al. [2011] show the F-test is suboptimal and sometimes powerless when testing against a sparse alternative, that is, when only a few coefficients are different from zero. A test based on a direct measure of the coefficients shall bring more power, as we see with thresholding tests. Another drawback is that the F-test requires $P < N$ for the second degree of freedom to be positive and for the rank of X to be smaller than the length of the response vector $\mathbf{Y}$, otherwise the estimation of variance (the denominator in (2.1)) gives zero.

In Generalized Linear Models [Nelder and Wedderburn, 1972], testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is also difficult because the model is saturated when $P \geq N$. In the standard setting with $P < N$ fixed, ϕ known, letting $L(\beta_0, \boldsymbol{\beta})$ be the likelihood function, the likelihood ratio test relies on the asymptotic distribution

$$-2\log\frac{\sup_{\boldsymbol{\beta}} L(\beta_0, \boldsymbol{\beta})}{L(\beta_0, \mathbf{0})} \to_d \chi^2_P \tag{2.2}$$

under H_0 as N tends to infinity, provided the model satisfies the conditions for asymptotic normality of maximum likelihood estimation [Wilks, 1938]. But asymptotic convergence is slow when P is large and fails in high-dimension $P \geq N$, which motivated Goeman et al. [2011], Guo and Chen [2016], Sur et al. [2017] to propose tests based on other asymptotic distributions. In the Gaussian case, the F distribution of (2.1) converges to the χ^2_P distribution when N gets large for a fixed $R = P$.

The situation $P \geq N$ is difficult in testing but is well addressed in model selection. In this chapter we exploit the ability of model selection of lasso in (1.10) to cope with $P \geq N$ to provide new solutions to testing in Generalized Linear Models. As mentioned before, lasso can be employed whether $P < N$ or not, and is a model selection technique in the sense that the solution $\hat{\boldsymbol{\beta}}_\lambda$ in (1.10) is sparse. Based on

these two properties we develop new tests of the form (1.16) that continue to hold when $P \geq N$, and that have good level and power properties.

2.2 Generalized Linear Models

2.2.1 Asymptotic pivotal thresholding statistic

We assume each component of the response $\mathbf{Y}$ has a distribution in the exponential family in (1.1) and follows the Generalized Linear Model (1.2) and (1.3). To test $H_0 : A\boldsymbol{\beta} = \mathbf{c}$ in linear models, Sardy et al. [2018] derived thresholding tests based on affine lasso estimator:

$$[\hat{\beta}_0 \ \hat{\boldsymbol{\beta}}_\lambda] \in \underset{[\beta_0 \ \boldsymbol{\beta}] \in \mathbb{R}^{P+1}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y} - X\boldsymbol{\beta} - \beta_0\mathbf{1}\|_2^2 + \lambda\|A\boldsymbol{\beta} - \mathbf{c}\|_1. \tag{2.3}$$

For the sake of simplicity we consider testing here the linear null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$. They derived pivotal test statistics of the form $\Lambda_0 = \|X^{\mathrm{T}}(\mathbf{Y} - \bar{Y}\mathbf{1})\|/\hat{\sigma}$, where $\hat{\sigma}/\sigma$ is pivotal under H_0. A natural extension to Generalized Linear Models is to consider test statistics of the form $\Lambda_0 = \|X^{\mathrm{T}}(\mathbf{Y} - \bar{Y}\mathbf{1})\|/D(\mathbf{Y})$ with a denominator $D(\mathbf{Y})$ that makes the statistic Λ_0 asymptotically pivotal. The aims of these new tests are a tighter control of the level of the test when P is large or possibly larger than N, and to achieve higher power than the existing tests. Indeed, most tests are based on the likelihood ratio statistic (2.2) which asymptotic chi-squared distribution can be a poor approximation when P is large and fails when P is larger than N.

The following theorem leads to a new asymptotic pivot for Generalized Linear Models.

Theorem 2.1 *Let $\mathbf{Y} = (Y_1, \ldots, Y_N)$ with i.i.d. entries with $E[Y_n] = \mu$ and finite variance ξ, for $n = 1, \ldots, N$. Let X be an $N \times P$ random matrix of N vectors of non-degenerate covariance $\Sigma \in \mathbb{R}^{P\times P}$, with $\mathbf{Y}$ independent of X. Consider the test statistic*

$$T(\mathbf{Y}) = \frac{\|X^{\mathrm{T}}(\mathbf{Y} - \bar{Y}\mathbf{1})\|}{\sqrt{N\hat{\xi}}}. \tag{2.4}$$

Assuming $\hat{\xi} \to_p \xi$, then $T(\mathbf{Y}) \to_d \|\mathbf{W}\|$, where $\mathbf{W} \sim \mathrm{N}(0, \Sigma)$.

Proof. Let $M = \mathbf{1}\boldsymbol{\mu}_X^{\mathrm{T}}$ be the matrix of size $N \times P$ with $\boldsymbol{\mu}_X = E(\mathbf{X})$ and $\mathbf{X}$ is the random vector generating covariates, and let $E(Y_n) = \mu$ for $n = 1, \ldots, N$. Notice that

$$X^{\mathrm{T}}(\mathbf{Y} - \bar{Y}\mathbf{1}) \ = \ (X - M)^{\mathrm{T}}(\mathbf{Y} - \mu\mathbf{1}) + (X - M)^{\mathrm{T}}(\mu\mathbf{1} - \bar{Y}\mathbf{1}). \tag{2.5}$$

On the one hand $(X-M)^{\mathrm{T}}(\mathbf{Y}-\mu\mathbf{1}) = \sum_{n=1}^{N}\mathbf{Z}_n$ with $\mathbf{Z}_n = (Y_n-\mu)(\mathbf{X}_n-\boldsymbol{\mu}_X) \in R^P$, where $\mathbf{X}_n^{\mathrm{T}}$ is the n-th (random) row of the matrix of covariates X for $n = 1,\ldots,N$. The first two moments are $E[\mathbf{Z}_n] = \mathbf{0}$ and $\mathrm{var}(\mathbf{Z}_n) = \xi\Sigma$. The central limit theorem states that $\sum_{n=1}^{N}\mathbf{Z}_n/\sqrt{N} \to_d \mathrm{N}(\mathbf{0},\xi\Sigma)$. On the other hand $(X-M)^{\mathrm{T}}(\mu\mathbf{1}-\bar{Y}\mathbf{1}) = (\mu-\bar{Y})\sum_{n=1}^{N}\mathbf{W}_n$ with $\mathbf{W}_n = (\mathbf{X}_n-\boldsymbol{\mu}_X)$. The first two moments are $E[\mathbf{W}_n] = \mathbf{0}$ and $\mathrm{var}(\mathbf{W}_n) = \Sigma$. Combining central limit theorem, law of large numbers and Slutsky's lemma, we have that

$$\frac{(\mu-\bar{Y})(X-M)^{\mathrm{T}}\mathbf{1}}{\sqrt{N}} \xrightarrow{p} 0. \tag{2.6}$$

Combining (2.5), the consistency of $\hat{\xi}$ and (2.6) with Slutsky's lemma leads to

$$\frac{X^{\mathrm{T}}(\mathbf{Y}-\bar{Y}\mathbf{1})}{\sqrt{N\hat{\xi}}} \xrightarrow{d} \mathrm{N}(\mathbf{0},\Sigma).$$

Finally, any norm being a continuous map, we have the desired result. ■

In GLM (1.2) and (1.3) under $H_0 : \boldsymbol{\beta} = \mathbf{0}$, the entries of $\mathbf{Y}$ are i.i.d with $E(Y_n) = g^{-1}(\beta_0)$ and have finite variance, and $\mathbf{Y}$ is independent of X, so Theorem 2.1 implies that $T(\mathbf{Y})$ is asymptotically pivotal under the null hypothesis for GLMs. Notice that the theorem is valid for any norm. This is particularly interesting since the sup-norm is used for the lasso, and the l_2-norm is used for group lasso. Therefore Theorem 2.1 implies a test based on $T(\mathbf{Y})$ can be employed for GLMs, and a critical value asymptotically independent of the nuisance parameter β_0 can be obtained by Monte Carlo simulation as discussed in Section 1.3.2. For the Poisson distribution $\hat{\xi} = \bar{Y}$ is a consistent estimate of the variance under the null; likewise with $\hat{\xi} = \bar{Y}(1-\bar{Y})$ for the Bernoulli distribution. Section 2.3.1 shows that the test has a good level even for finite N and large P, and that the test has high power also when P is larger than N.

2.2.2 Connection with the zero-thresholding function

The statistic $T(\mathbf{Y})$ is the zero-thresholding function $\lambda_0(\mathbf{Y})$ of lasso for Generalized Linear Models [Park and Hastie, 2007] for certain link functions. When employing the canonical link, Giacobino et al. [2017] show that the zero-thresholding function of the estimator is the numerator of $T(\mathbf{Y})$ in (2.4), which is not asymptotically pivotal. The following theorem states a condition on the link function for lasso to have $T(\mathbf{Y})$ as a zero-thresholding function.

Theorem 2.2 *Let* $\mathbf{Y}$ *be a random vector with a distribution in the exponential family with variance function* $V(\mu)$ *and known* ϕ*, and let* X *a matrix of predictors such that* $E[\mathbf{Y}] = g^{-1}(\beta_0\mathbf{1}+X\boldsymbol{\beta})$*, where* g *is the link function. If* $h = g^{-1}$ *satisfies that the negative log-likelihood of* $\mathbf{Y}$ *is convex and that* $\{h'(\beta_0)\}^2 = V(h(\beta_0))$*, then*

$$\lambda_0(\mathbf{Y}) = \frac{\|X^T\left(\mathbf{Y} - \bar{Y}\mathbf{1}\right)\|_\infty}{\sqrt{NV(\bar{Y})a(\phi)}}, \tag{2.7}$$

is (up to a constant) the zero-thresholding statistic *of lasso for Generalized Linear Models. Here,* a *is the known function in the exponential distribution* (1.1).

Proof. Assuming ϕ is known and for a fixed λ, Park and Hastie [2007] estimate β_0 and $\boldsymbol{\beta}$ by minimizing the penalized likelihood

$$\mathrm{PL}_\lambda(\beta_0, \boldsymbol{\beta}) = -\sum_{n=1}^{N}\left(\frac{Y_n\theta_n - b(\theta_n)}{a(\phi)}\right) + \lambda\|\boldsymbol{\beta}\|_1. \tag{2.8}$$

By properties of the exponential family, we have $E(Y_n) = b'(\theta_n) = h(\beta_0 + \mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta})$ and $\mathrm{var}(Y_n) = b''(\theta_n)a(\phi)$. Consequently

$$\left\{\begin{array}{lcl} \frac{\partial\theta_n}{\partial\beta_0} & = & \frac{h'(\beta_0+\mathbf{x}_n^T\boldsymbol{\beta})}{b''(\theta_n)} \\ \nabla_{\boldsymbol{\beta}}\theta_n & = & \frac{\mathbf{x}_n h'(\beta_0+\mathbf{x}_n^T\boldsymbol{\beta})}{b''(\theta_n)}. \end{array}\right.$$

By assumption PL_λ is convex, so the point $(\hat{\beta}_0, \mathbf{0})$ belongs to the minimum set of PL_λ if and only if $\mathbf{0}$ is a subgradient of PL_λ at $(\beta_0, \boldsymbol{\beta}) = (\hat{\beta}_0, \mathbf{0})$. This is equivalent to

$$\left\{\begin{array}{lcl} \frac{\partial\mathrm{PL}_\lambda}{\partial\beta_0} & = & \sum_{n=1}^{N}\left(\frac{y_n\frac{\partial\theta_n}{\partial\beta_0} - b'(\theta_n)\frac{\partial\theta_n}{\partial\beta_0}}{a(\phi)}\right) = 0 \\ \nabla_{\boldsymbol{\beta}}\mathrm{PL}_\lambda & = & \sum_{n=1}^{N}\left(\frac{y_n\nabla_{\boldsymbol{\beta}}\theta_n - b'(\theta_n)\nabla_{\boldsymbol{\beta}}\theta_n}{a(\phi)}\right) + \lambda[-1,1]^P \ni \mathbf{0}. \end{array}\right.$$

Since at $\boldsymbol{\beta} = \mathbf{0}$ we have $b'(\theta_n) = h(\beta_0)$, $\frac{\partial\theta_n}{\partial\beta_0} = \frac{h'(\beta_0)}{b''(\theta_n)}$ and $\nabla_{\boldsymbol{\beta}}\theta_n = \frac{\mathbf{x}_n h'(\beta_0)}{b''(\theta_n)}$, this is also equivalent to

$$\left\{\begin{array}{lcl} \sum_{n=1}^{N}\left(\frac{h'(\beta_0)}{b''(\theta_n)a(\phi)}\left(y_n - h(\beta_0)\right)\right) & = & 0 \\ \sum_{n=1}^{N}\left(\frac{\mathbf{x}_n h'(\beta_0)}{b''(\theta_n)a(\phi)}\left(y_n - h(\beta_0)\right)\right) + \lambda[-1,1]^P & \ni & \mathbf{0}. \end{array}\right.$$

A solution exists if and only if $h(\beta_0) = \bar{y}$ and λ at least as large as

$$\lambda_0(\mathbf{y}) = \left\|\frac{h'(\beta_0)X^T(\mathbf{y} - \bar{y}\mathbf{1})}{V(h(\beta_0))a(\phi)}\right\|_\infty,$$

where V is the variance function such that $V(h(\beta_0)) = b''(\theta)$.

So if $|h'(\beta_0)| = \sqrt{V(h(\beta_0))}$, we obtain the desired zero-thresholding function $\lambda_0(\mathbf{y})$ up to the constant $\sqrt{Na(\phi)}$. ■

In particular, one can obtain the specific link functions for GLMs in Gaussian, binomial and Poisson distributions for which theorem 2.2 applies. The following corollary describes it.

Corollary 2.3 *Let $h(x) = x$, $h(x) = x^2/4$ for $x \geq 0$, and $h(x) = (\sin(x) + 1)/2$ for $x \in [-\pi/2, \pi/2]$, be the inverse link function in Generalized Linear Models corresponding to Gaussian, Poisson and binomial distributions, respectively. The zero-thresholding function of the lasso estimator of Park and Hastie [2007] is (up to a constant) the asymptotically pivotal test statistic* (2.4) *with $\hat{\xi} = \hat{\sigma}^2$, $\hat{\xi} = \bar{Y}$ and $\hat{\xi} = \bar{Y}(1 - \bar{Y})$, respectively.*

Proof. Using Theorem 2.2 it suffices to satisfy the equality $\{h'(\beta_0)\}^2 = V(h(\beta_0))$. Notice that $V(x)$ equals 1, x and $x(1 - x)$ for Gaussian, binomial and Poisson distributions, respectively [Nelder and Wedderburn, 1972]. For each distribution is easy to check that the result holds. ■

Notice that these new links, $h^{-1}(y) = 2\sqrt{y}$ for Poisson and $h^{-1}(y) = \sin^{-1}(2y - 1)$ for binomial, are reminiscent of Anscombe's transforms $A(y) = \sqrt{y + 3/8}$ and $A(y) = \sin^{-1}\sqrt{(8y + 3)/14}$, respectively [Anscombe, 1948]. With corollary 2.3, we get the asymptotically pivotal test statistic (2.4) for Gaussian, Poisson and binomial, respectively:

$$\lambda_0(\mathbf{Y}) = \frac{\|X^T\left(\mathbf{Y} - \bar{Y}\mathbf{1}\right)\|}{\sqrt{N\hat{\sigma}^2}}, \tag{2.9}$$

$$\lambda_0(\mathbf{Y}) = \frac{\|X^T\left(\mathbf{Y} - \bar{Y}\mathbf{1}\right)\|}{\sqrt{N\bar{Y}}}, \tag{2.10}$$

$$\lambda_0(\mathbf{Y}) = \frac{\|X^T\left(\mathbf{Y} - \bar{Y}\mathbf{1}\right)\|}{\sqrt{N\bar{Y}(1 - \bar{Y})}}. \tag{2.11}$$

By taking $\hat{\sigma} = \|Y - \bar{Y}\mathbf{1}\|_2$ we have a pivotal lasso test that corresponds to the square root lasso. Notice that by taking the sup-norm or the l_2-norm in the numerator of (2.9) we have the corresponding test statistics for lasso or group lasso, respectively. Table 2.1 shows the summary of the test statistics that we use. The test-threshold λ_α can be obtained by Monte Carlo simulations as is described in Section 2.2.3.

Table 2.1: Test statistics for Gaussian, Poisson and binomial distributions, with the corresponding inverse link functions $h(x)$ and Anscombe's transforms $A(y)$.

Family	$\lambda_0(\mathbf{Y})$	$\hat{\xi}$	$h(x)$	$h^{-1}(y)$	$A(y)$
Gaussian	$\frac{\left\|X^T\left(\mathbf{Y}-\bar{Y}\mathbf{1}\right)\right\|}{\sqrt{N}\left\|Y-\bar{Y}\mathbf{1}\right\|_2}$	$\|Y-\bar{Y}\mathbf{1}\|_2$	x	y	y
Poisson	$\frac{\left\|X^T\left(\mathbf{Y}-\bar{Y}\mathbf{1}\right)\right\|}{\sqrt{N\bar{Y}}}$	$\bar{Y}$	$x^2/4 \quad x \geq 0$	$2\sqrt{y}$	$\sqrt{y+3/8}$
Binomial	$\frac{\left\|X^T\left(\mathbf{Y}-\bar{Y}\mathbf{1}\right)\right\|}{\sqrt{N\bar{Y}(1-\bar{Y})}}$	$\bar{Y}(1-\bar{Y})$	$(\sin(x)+1)/2$ $x \in [-\pi/2, \pi/2]$	$\sin^{-1}(2y-1)$	$\sin^{-1}\sqrt{(8y+3)/14}$

2.2.3 Illustrative example

In order to show how our methodology works in practice, we describe here an illustrative example. Say we are interested in testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$, from a given Poisson data $(\mathbf{y}, X)$. The procedure is as follows.

1. Identify the corresponding Poisson test-statistic in Table 2.1 and choose the desired norm, here the sup-norm to test with lasso penalty:
$$\lambda_0(\mathbf{Y}) = \frac{\left\|X^T\left(\mathbf{Y}-\bar{Y}\mathbf{1}\right)\right\|_\infty}{\sqrt{N\bar{Y}}}.$$

2. Obtain the empirical distribution of Λ_0 by simulating M vectors $\mathbf{y}_0^{(1)}, \ldots, \mathbf{y}_0^{(M)}$ from $\mathbf{Y}_0 \overset{\text{i.i.d}}{\sim} \text{Poisson}(g^{-1}(\beta_0\mathbf{1}))$ under H_0, and calculating $\lambda^{(m)} = \frac{\left\|X^T\left(\mathbf{y}_0^{(m)}-\bar{\mathbf{y}}_0^{(m)}\mathbf{1}\right)\right\|_\infty}{\sqrt{N\bar{\mathbf{y}}_0^{(m)}}}$ for $m = 1, \ldots, M$.

3. Calculate the test-threshold λ_α by taking the upper α-quantile of the empirical distribution of Λ_0.

4. Test the data with the corresponding thresholding-test (1.16):
$$\phi(\mathbf{y}) = \begin{cases} 1 & \text{if } \frac{\left\|X^T(\mathbf{y}-\bar{y}\mathbf{1})\right\|_\infty}{\sqrt{N\bar{y}}} > \lambda_\alpha \\ 0 & \text{otherwise} \end{cases}.$$

To obtain the empirical distribution of Λ_0 in step 2 it is necessary to choose a value for β_0. Since the statistic is asymptotically pivotal, any value of β_0 can be chosen to simulate $\mathbf{y}_0^{(m)}$, for instance $\beta_0 = 0$. This choice is in particular convenient for binomial data, since for any value of N, Λ_0 has the following property.

Property 2.1 *Let $\lambda_0(\mathbf{Y}_{\beta_0})$ be the distribution of the test-statistic* (2.10) *for binomial i.i.d data $\mathbf{Y}_{\beta_0}$ and $E(\mathbf{Y}_{\beta_0}) = g^{-1}(\beta_0)\mathbf{1}$ with g the canonical link. The distribution $\lambda_0(\mathbf{Y}_{\beta_0})$ is symmetric with respect to β_0, this is*

$$\lambda_0(\mathbf{Y}_{\beta_0}) = \lambda_0(\mathbf{Y}_{-\beta_0}) \tag{2.12}$$

Proof. Notice that $\lambda_0(\mathbf{Y}_{\beta_0}) = \frac{\left\|X^T\left(\mathbf{Y}_{\beta_0}-\bar{Y}_{\beta_0}\mathbf{1}\right)\right\|}{\sqrt{N\bar{Y}_{\beta_0}(1-\bar{Y}_{\beta_0})}} = \frac{\left\|X^T\left((\mathbf{1}-\mathbf{Y}_{\beta_0})-(1-\bar{Y}_{\beta_0})\mathbf{1}\right)\right\|}{\sqrt{N(1-\bar{Y}_{\beta_0})(1-(1-\bar{Y}_{\beta_0}))}}$. Since $1 - Y_{\beta_0} \sim Y_{-\beta_0}$, we have $\lambda_0(\mathbf{Y}_{\beta_0}) = \frac{\left\|X^T\left(\mathbf{Y}_{-\beta_0}-\bar{Y}_{-\beta_0}\mathbf{1}\right)\right\|}{\sqrt{N\bar{Y}_{-\beta_0}(1-\bar{Y}_{-\beta_0})}}$. ■

2.2.4 Combining tests and the composite $\oplus$-test

Suppose that test $\phi^{(1)}$ based on a first thresholding estimator has level α and good power properties for a type of alternative hypothesis, and that test $\phi^{(2)}$ based on a second thresholding estimator has level α and good power properties for another type of alternative hypothesis. It is reasonable to wish a single test ϕ that has level α and that is almost as powerful as the best of both tests regardless of the type of alternative hypothesis. We propose the following way to combine both tests. Let $\lambda_0^{(i)}$ and $\lambda_\alpha^{(i)}$ be the zero-thresholding function and test-threshold of test $\phi^{(i)}$ for $i \in \{1, 2\}$. The composite null-thresholding statistic

$$\Lambda_0 = \max\left(\frac{\lambda_0^{(1)}(\mathbf{Y}_0)}{\lambda_\alpha^{(1)}}, \frac{\lambda_0^{(2)}(\mathbf{Y}_0)}{\lambda_\alpha^{(2)}}\right). \tag{2.13}$$

can be employed to develop a single test of level α. The standardization by either $\lambda_\alpha^{(1)}$ or $\lambda_\alpha^{(2)}$ ensures both individual test statistics within (2.13) possess the same rejection region $[1, \infty]$.

Arias-Castro et al. [2011] conclude that a test based on lasso is powerful under sparse alternatives and powerless under dense alternatives, while Fisher's or group lasso tests behave the other way around. Based on (2.13), Sardy et al. [2018] proposes the composite $\oplus$-test that combines the test based on lasso ("+" character symbolizes the coordinate-wise nature of lasso) and the test based on group lasso ("$\circ$" character symbolizes the ℓ_2-ball of group lasso's penalty). We extend this test to the generalized linear model scenario by combining the test based on lasso for GLMs (sup-norm of statistics in Table 2.1) and the test based on group lasso for GLMs (l_2-norm of statistics in Table 2.1). The goal of this test is to be nearly as powerful as the best test between lasso's and group lasso's tests, which we investigate in Section 2.3.1.

2.2.5 Parametric and non-parametric pivotal statistics

Nowadays it is of great interest in scientific community to perform studies on a certain type of data, called metagenomics, which is the study of genetic material recovered directly from environmental samples. A specific characteristic of most of these data is that their X matrix is sparse, with up to 70% of its entries equal to zero. Theorem 2.1 shows the convergence of $\lambda_0(\mathbf{Y})$ as N increases. This means that in finite samples, the level of the test becomes more sensitive to the selection of β_0 if N is small. Moreover, since the asymptotic convergence relies in the Central Limit Theorem, if matrix X is sparse the convergence is slower. Step 2 in the illustrative example in Section 2.2.3 relies on this convergence to obtain the level by Monte Carlo simulation, therefore in this particular case the level may be far from its nominal level. To cope with this problem, we propose here two methodologies to perform the Monte Carlo simulation and obtain the empirical distribution of Λ_0, leading to tests with the nominal level. One is based on estimating β_0 under the null model. Under the null hypothesis the maximum likelihood estimator of the intercept is $\hat{\beta}_0 = g(\bar{y})$. We propose to perform the Monte Carlo simulations to obtain Λ_0 with this choice of β_0. According to Theorem 2.1 it is also asymptotically pivotal. The main difference is that for finite samples the level is closer to its nominal value, regardless of the true value of β_0 or the structure of matrix X. The second proposal is to perform the Monte Carlo simulation by bootstrapping $\mathbf{y}$, in the spirit of permutation tests. This approach is non-parametric and also has a level closer to the nominal one for finite samples. Clearly both methodologies are pivotal under the null hypothesis regardless of X. We perform some simulations in Section 2.3.3 and we obtain some power plots to observe the general behavior of these new approaches for the selection of β_0.

2.3 Simulation study

2.3.1 Comparative power analysis

To illustrate how thresholding tests in Generalized Linear Models compare with classical and more contemporary tests in terms of power, we consider the class of alternative hypotheses

$$H_1^{s,\theta} : \boldsymbol{\beta} = \theta \cdot \pi((\underbrace{\pm 1, \ldots, \pm 1}_{s}, \underbrace{0, \ldots, 0}_{P-s})^{\mathrm{T}}), \tag{2.14}$$

indexed by $s \in \{0, 1, \ldots, P\}$ and $\theta \in \mathbb{R}$: s controls the amount of sparsity and θ controls the signal-to-noise ratio. Here $\pi(\mathbf{u})$ performs a random permutation of the vector $\mathbf{u}$. The sign of the coefficients $\beta_p = \pm\theta$ are random and equiprobable for $p = 1, \ldots, s$. We say that the alternative hypothesis is sparse when s is small and dense when s is large with respect to P. We estimate by Monte Carlo simulation power functions as a function of the two parameters (s, θ) indexing the alternative $H_1^{s,\theta}$-hypotheses (2.14). Three X matrices with dimension $N = 100$ and $P \in \{10, 40, 1000\}$ are generated according to the Monte Carlo simulation of Guo and Chen [2016]. We simulated Gaussian, binomial and Poisson data generated with the canonical link according to linear model (1.2) and (1.3), and add an intercept $\beta_0 = -2$. Five tests are compared in all cases: three thresholding tests (lasso, group lasso, composite lasso), the test of Guo and Chen [2016], and Fisher's F-test (Gaussian) or likelihood ratio test (non-Gaussian) when $P < N$. The test of Guo and Chen [2016] is based on the convergence

$$\frac{(N-1)\sum_{i\neq j}^{N}\left((Y_i-\bar{Y})(Y_j-\bar{Y})X_i^{\mathrm{T}}X_j\right)}{\sqrt{2\sum_{i\neq j}^{N}\left((Y_i-\bar{Y})^2(Y_j-\bar{Y})^2(X_i^{\mathrm{T}}X_j)^2\right)}} \to_d N(0,1),$$

under H_0 and some assumptions. In the binomial case, the rescaled χ^2 method of Sur et al. [2017] is also compared. This method replaces the χ^2 convergence of the likelihood ratio test by an $\alpha(\kappa)\chi^2$ convergence with $\kappa = P/N$ to fix the error in the level of the likelihood ratio test when P is not negligible to N. For the thresholding tests, we simulate with $\beta_0 = 0$ to obtain the test-threshold. The effect of the selection of β_0 and the asymptotic behavior of Λ_0 is better discussed in Section 2.3.2.

Figures 2.1 and 2.2 plot the power functions for sparse and dense alternative hypotheses, respectively. First, second and third row correspond to Gaussian, binomial and Poisson simulations, respectively, in both plots. Interesting behaviors can be observed. First, comparing lasso and group lasso tests on sparse and dense situations, one sees that lasso is more powerful when the alternative hypothesis is sparse; on the contrary, when the alternative is dense, group lasso is more powerful. This corroborates the results of Arias-Castro et al. [2011]. Second, the composite $\oplus$-test of Section 2.2.4 has power close to the most powerful test between lasso and group lasso. Third, Fisher's test, like group lasso's test, is better in the dense case than the sparse case, while the likelihood ratio test performs poorly due to the poor χ^2 approximation when P is large. The test of Guo and Chen [2016] (HDGLM in the plot) is slightly off in terms of level and its power is not as good as that of the $\oplus$-test except

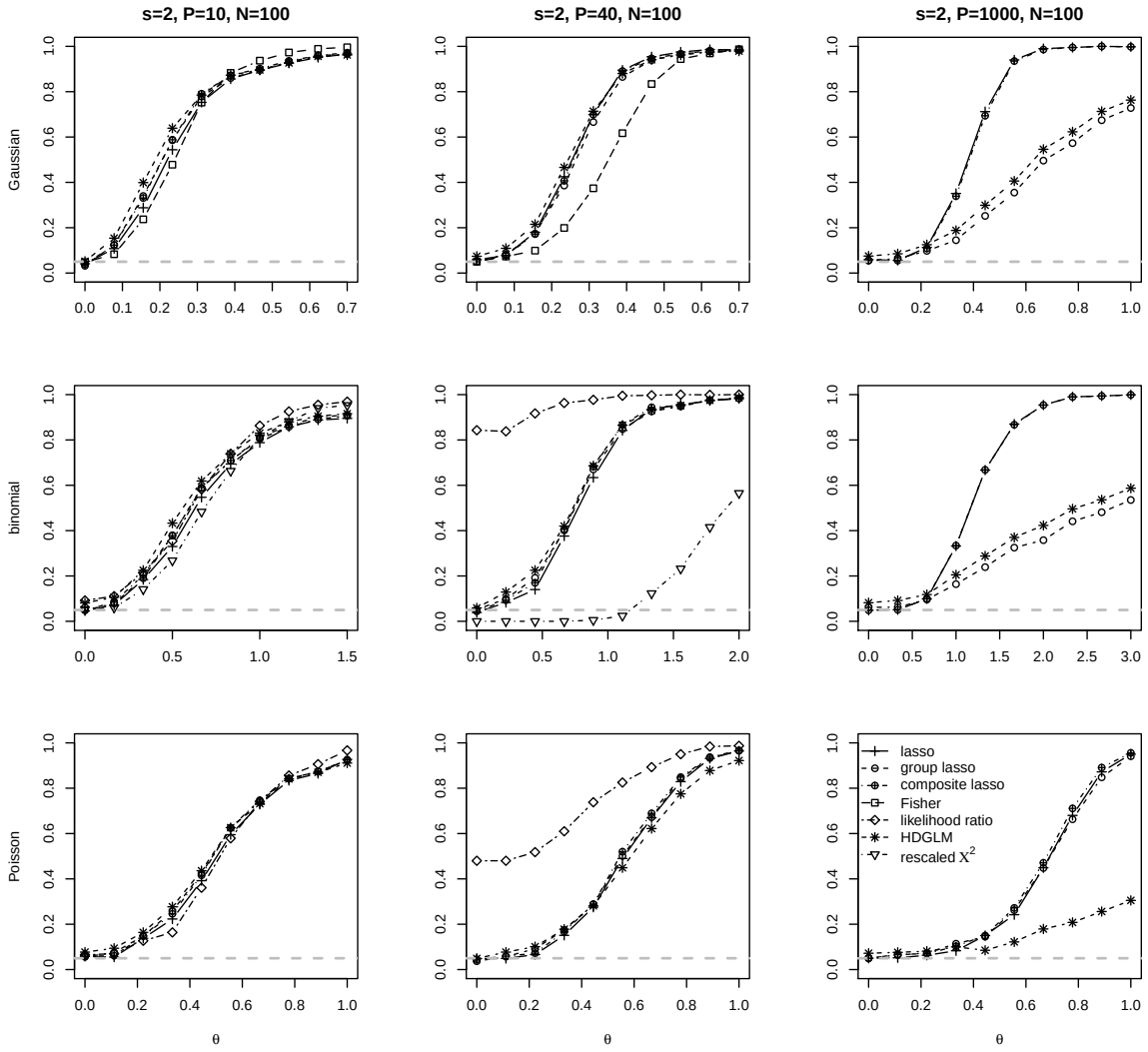

Figure 2.1: Power functions estimated by Monte Carlo simulation for sparse alternative hypotheses.

in the dense case when $P = 1000$.

Sometimes some tests appear to have higher power, but it is important to observe that, at $\theta = 0$ on the power plot, their level is larger than $\alpha = 0.05$. Figure 2.3 plots the empirical levels achieved by the tests in all cases. Clearly, the thresholding tests have the best control on the level. Next comes the HDGLM method of Guo and Chen [2016] with a slight bias. Likelihood ratio test has a poor control of the level with two values outside the range $[0, 0.1]$ of the plot (not shown here).

Overall the composite $\oplus$-test is best in terms of power and in respecting the nominal level.

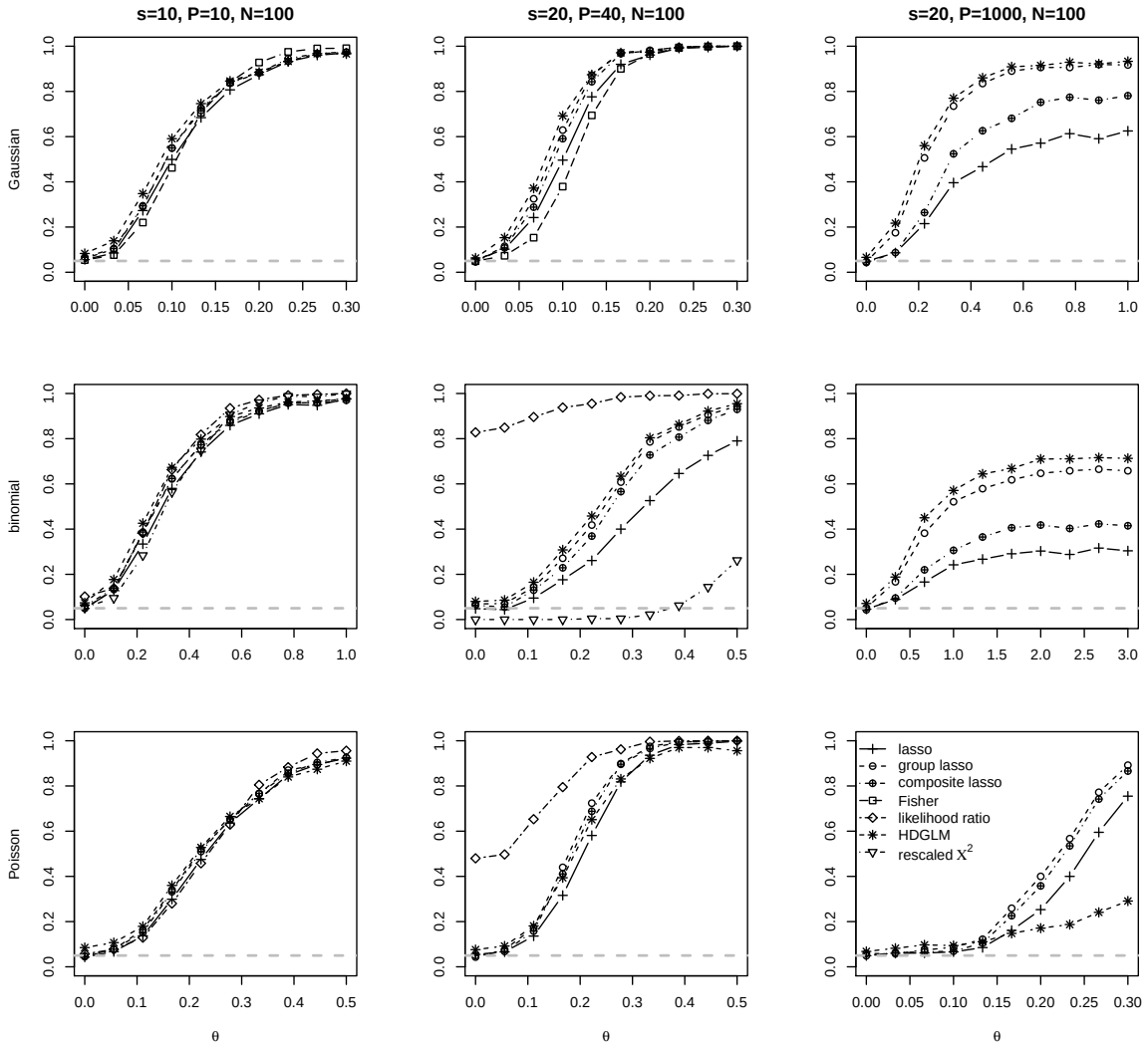

Figure 2.2: Power functions estimated by Monte Carlo simulation for dense alternative hypotheses.

2.3.2 Selection of β_0 and asymptotic behavior of $\lambda_0(\mathbf{Y})$

We perform a simulation to observe the effect of selection of β_0 and the asymptotic convergence for several values of the intercept and sparsity levels in X. We simulate six matrices X according to Monte Carlo simulation of Guo and Chen [2016]. Three of size $N = 100, P = 100$, and three of size $N = 1000, P = 100$. We take the three matrices of size $N = 100, P = 1000$ and we randomly set to zero 0%, 45% and 90% of its entries, respectively. We do the same with the three matrices of size $N = 1000, P = 100$. We simulated binomial data with $\beta_0 \in \{0, 1, 2\}$. Figure 2.4 plots the empirical densities of Λ_0 for the six matrices X. In each plot, the densities for $\beta_0 \in \{0, 1, 2\}$ are

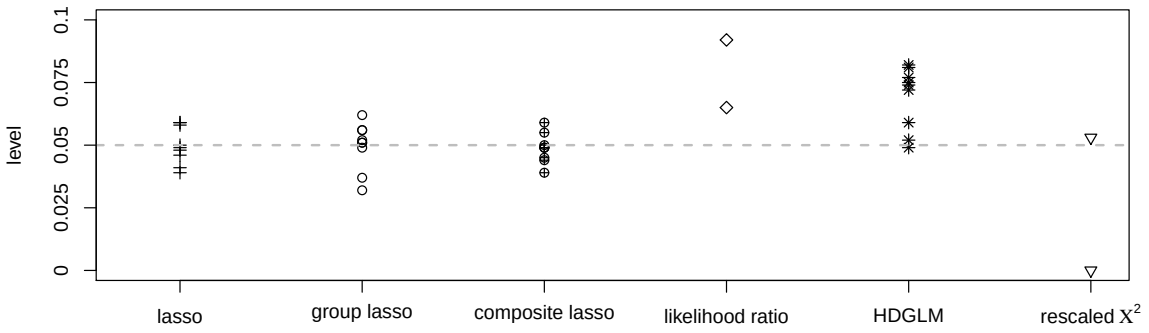

Figure 2.3: Empirical levels achieved by the tests for the nine scenarios (Gaussian, binomial and Poisson and $P \in \{10, 40, 1000\}$). The values are plotted in the range $[0, 0.1]$ around the nominal level $\alpha = 0.05$ (dotted line).

plotted in black, red, and green, respectively, and their corresponding test-thresholds appear as a vertical dotted line. First and second row correspond to matrices with $N = 100$, and $N = 1000$, respectively. First, second, and third column correspond to matrices with 0%, 45% and 90% levels of sparsity, respectively.

As expected, we see that as sparsity in X increases, the densities for the three values of β_0 differ more. When N increases, the difference between them decreases.

2.3.3 Power analysis under different sparsity levels in X

We are interested in studying data consisting in measurements of gut microbiota i.e. abundances of bacteria species in the gut. A specific characteristic of these data is that its matrix X is sparse with up to 70% of zero entries. Due to disclosure prohibitions regarding the project in which we are currently working on, we are not allowed to share any results concerning this particular data. Nevertheless we mimic the structure of such matrix by simulating one matrix of size $N = 100, P = 200$ according to Monte Carlo simulation of Guo and Chen [2016], and choosing uniformly at random 70% of its entries and setting them to zero. We estimate by Monte Carlo simulation power functions as in Section 2.3.1. We simulated binomial data generated with the canonical link according to linear model (1.2) and (1.3), and add an intercept $\beta_0 \in \{0, 1, 2\}$. We compare the same tests for binomial data as in Section 2.3.1. We also compared the two methodologies described in Section 2.2.5 with their three lasso, group lasso and composite lasso versions. We call here lassohat and permalasso

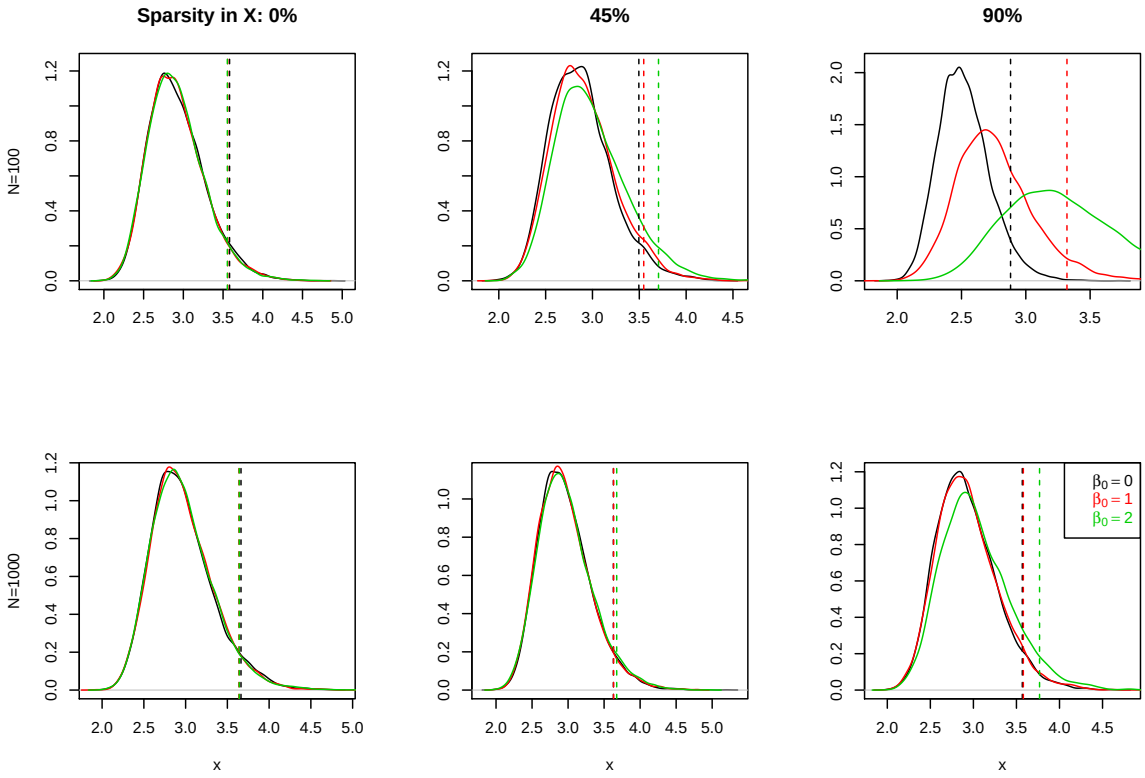

Figure 2.4: Empirical densities of Λ_0 with $\beta_0 \in \{0, 1, 2\}$ in black, red and green curves, respectively, and their corresponding test-thresholds λ_α in vertical dotted lines. First and second row correspond to matrices with $N = 100$, and $N = 1000$, respectively, with $P = 100$. First, second, and third column correspond to matrices with 0%, 45% and 90% levels of sparsity, respectively.

the methodologies estimating β_0 by $\hat{\beta}_0 = g^{-1}(\bar{y})$ and bootstrapping y, respectively. Finally we also compared the permanova test [Anderson, 2001], widely used for this kind of data in current applications.

Figure 2.5 plots the power functions for different sparsity levels in X, sparse and dense alternatives, and $\beta_0 \in \{0, 1, 2\}$. First and second column correspond to sparse alternative ($s = 2$), with matrix fully dense X and X with 70% of zero entries, respectively. Third and fourth column correspond to dense alternative ($s = 10$), with matrix fully dense X and X with 70% of zero entries, respectively. First, second and third row correspond to $\beta_0 = 0$, $\beta_0 = 1$ and $\beta_0 = 2$, respectively. For simplicity we just plot the composite lasso in all thresholding tests (lasso, lassohat, permalasso), since the conclusions regarding lasso and group lasso are similar to the ones in Section 2.3.1

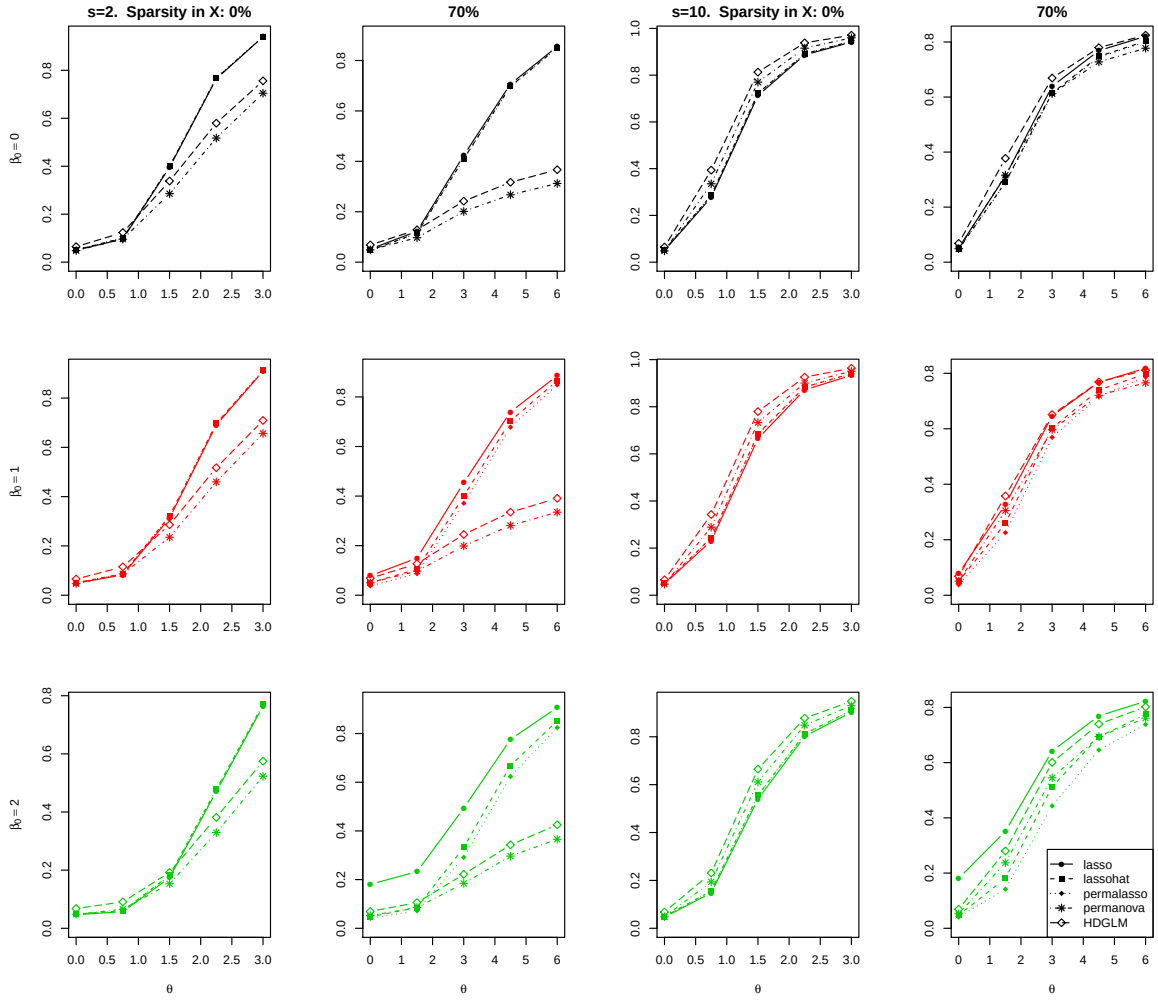

Figure 2.5: Power functions estimated by Monte Carlo simulation for two matrices X of size $N = 100$, $P = 200$ with two different sparsity levels, sparse and dense alternatives, and $\beta_0 \in \{0, 1, 2\}$.

(in the plot we just call lasso the composite version of test in this section).

Results are interesting. First, we see that for $\beta_0 = 0$, lasso, lassohat and permalasso have a similar behavior. This is also the case when X is not sparse. As expected, in this case lasso is still close to its nominal level regardless of β_0, but when β_0 increases and X is sparse, this is not longer the case. In the sparse X scenario, lassohat and permalasso still are close to the nominal level. Second, permanova and HDGLM have a similar performance: good power in dense alternative and bad power in sparse alternative. In dense alternatives they are the best, closely followed by lassohat and permalasso. In sparse alternative these tests are the worst, and are affected when X is sparse.

Overall, lassohat and permalasso have the best performance in the sense that their level is closer to the nominal level and their power is high.

2.4 Discussion

Thresholding tests have good control of the nominal level of the test and have high power for distributions in the exponential family regardless of the relative size of N and P. We can always use our approach whether $N > P$ or not. When the alternative hypothesis is dense, one should use group lasso since it has more power than lasso. On the contrary, when the alternative hypothesis is sparse, choosing lasso leads to more power than with group lasso's and Fisher's tests. When no a priori dense or sparsity assumption can be made on the likely alternative hypothesis, the composite lasso $\oplus$-test should be used. When matrix X is sparse, methodologies in Section 2.2.5 are preferred to better control the level. Our approach can be extended to null hypothesis of the form $H_0 : A\boldsymbol{\beta} = \mathbf{c}$ with future work in the extension of affine lasso to GLMs.

Chapter 3

Estimation of galaxy cluster's emissivity in astrophysics

This chapter proposes a new methodology to estimate the 3D gas emissivity of a galaxy cluster from a 2D image taken by a telescope. The image involves artifacts such as blurring and sensitivity of the telescope, as well as the presence of point sources that are behind the galaxy cluster in the universe. We start in Section 3.1 describing the general problem. In Section 3.1.1 we explain the astrophysical meaning of the images taken by the telescope. Section 3.1.2 describes the XMM-mission telescope from where most of the images were taken from, and Section 3.1.3 explains the current state-of-the-art method. Then in Section 3.2 we describe our approach to the problem. Section 3.2.1 details all the features involved in the image, assumptions, artifacts, point sources, etc. Section 3.2.2 describes our modeling of the problem and the parameters to estimate. Section 3.2.3 shows how to deal with the asymmetry of the galaxy clusters. In Section 3.3 we model the data as a Poisson generalized linear model with identity link, penalized by two parameters, one corresponding to the basis functions of the profile, and other to the point sources. Section 3.3.1 derives the zero thresholding function for our model, and shows how to choose the penalty parameters. Section 3.3.2 explains our approach to obtain uncertainty quantification by bootstrapping the image. To test our methodology we do some numerical experiments in Section 3.4, first with simulated data in Section 3.4.1 and then with five real images in Section 3.4.2, with an explanation of our findings in Section 3.4.3. Finally we give some conclusions in Section 3.5. In Section 3.6 we provide information regarding the reproducibility of our work.

3.1 Motivation

3.1.1 Emissivity of astrophysical sources

Several types of astrophysical sources originate from the radiative processes occurring in an "optically thin" environment, that is, a situation in which a photon has a low probability of interacting with the surrounding material and can escape the source freely. Such a situation occurs when the mean density of material in the source is very low. Examples of such astronomical sources include galaxies (where the observed light is the sum of the light emitted by all stars), the coronae of the Sun and other convective stars, cocoons of expanding material after supernova explosions (*supernova remnants*) and galaxy groups and clusters (which are filled with a hot ($10^7 - 10^8$ Kelvin) low-density plasma that constitutes the majority of the ordinary matter of large-scale structures in the Universe). In case the source is optically thin, the electromagnetic radiation I in a given direction is the integral of the intrinsic emissivity of the source over the source volume,

$$I = \frac{1}{4\pi D^2} \int_V \varepsilon \, dV, \tag{3.1}$$

where the emissivity ε is the energy emitted by the source in electromagnetic radiation and D is the source distance. The three-dimensional distribution of the emissivity is of interest as it provides valuable information on the physical properties of the emitting material (e.g., density, temperature, metallicity).

In the case of galaxy clusters, the emitting plasma is so hot that these structures radiate predominantly in X-rays [Sarazin, 1988]. Current X-ray telescopes like *XMM-Newton* and *Chandra* are able to detect the emission from the plasma and make detailed maps of the distribution of hot gas in galaxy clusters, which are extremely useful to understand the formation and evolution of structures in the Universe [Kravtsov and Borgani, 2012], study the overall matter content and the missing mass ("dark matter") problem [Clowe et al., 2006], and constrain the cosmological parameters governing the evolution of the Universe as a whole [Allen et al., 2011]. In most cases, X-ray images of galaxy clusters show round, azimuthally symmetric morphologies indicating that the geometry of these structures is nearly spherical. The observed emissivity decreases radially from the center of the source to its outermost border [Eckert et al., 2012]. Assuming spherical symmetry, (3.1) can be written explicitly as

a function of projected distance s to the cluster center,

$$I(s) \propto \int \varepsilon(r)\, dz \quad \text{with} \quad r^2 = s^2 + z^2, \tag{3.2}$$

where r is the three-dimensional distance to the cluster center, $I(s)$ is the observed azimuthally-averaged brightness profile, and the integral is performed along the line of sight z. While $\varepsilon(r)$ can in principle be evaluated directly from the observed emission by solving the integral (3.2), in practice the problem is rendered complicated by the presence of noise in the original data, as for instance with the XMM-Newton telescope described below. Indeed, as for many inverse problems the convolution smooths small-scale fluctuations, thus the inverse transformation has the opposite effect and the noise can be greatly amplified [see Lucy, 1974, 1994]. This effect is particularly important in the low signal-to-noise regime.

3.1.2 The *XMM-Newton* mission

The *XMM-Newton* space telescope [Jansen et al., 2001] is a cornerstone mission of the European Space Agency. It was put in orbit on December 10, 1999 by an Ariane 5 launcher and it remains to this day the largest X-ray telescope ever operated. The spacecraft is made of three co-aligned X-ray telescopes that observe the sky simultaneously. At the focal point of the three telescopes are located two instrument, the European Photon Imaging Camera (EPIC) and the Reflection Grating Spectrometer (RGS). The bottom image of Figure 3.1 is an image of the galaxy cluster Abell 2142 recorded by the *XMM-Newton* observatory [Tchernin et al., 2016]. The data were acquired in 2012 (PI: Eckert) as part of the *XMM-Newton* guest observer program, in which astronomers are invited to propose suitable targets to be observed by the spacecraft and provide a detailed scientific justification for their program.

EPIC [Turner et al., 2001] consists of three high-sensitivity cameras which cover a field of view of 30 arcmin diameter roughly equivalent to the size of the full moon. The cameras are made of 600×600 pixels organized in 8 individual chips which record the time, energy and position of incoming X-ray photons, resulting in an image like on the bottom of Figure 3.1. The sensitivity of the instrument is maximal for sources precisely aligned with the axis of the telescopes (the aim point) and gradually declines for sources located slightly offset from the optical axis. The angular resolution of the telescope is 6 arcsec at the aim point and it degrades to 15 arcsec at the edge of the field of view. Astrophysical sources with an apparent size smaller than the angular

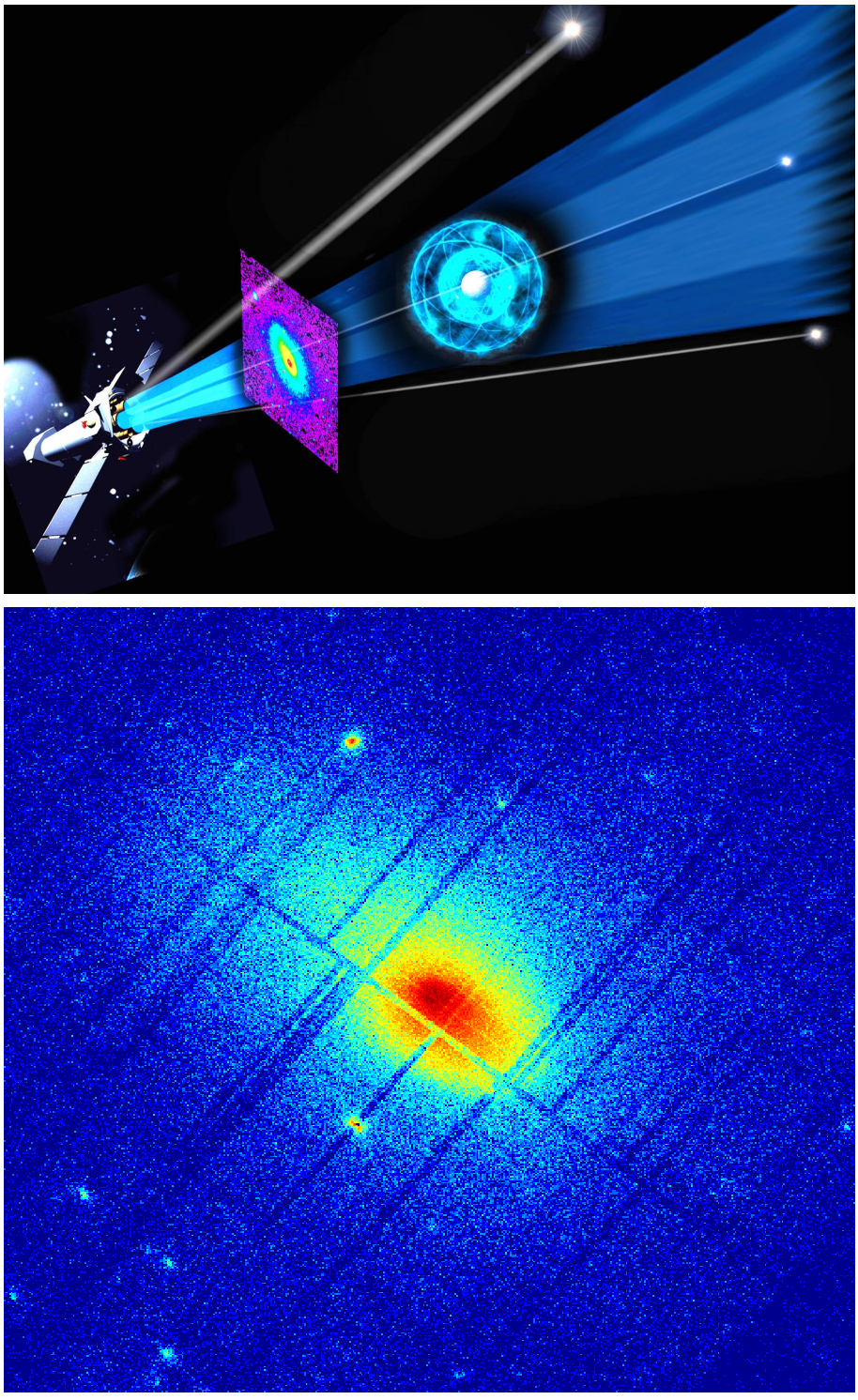

Figure 3.1: Upper: schematically view of a telescope, the image taken by it, a galaxy cluster and two point sources. Lower: real image taken by the XMM-Newton telescope.

resolution of the instrument thus appear blurred with a typical size and shape that is known from the characteristics of the telescopes. Similarly, the degradation of the sensitivity of the instrument with off-axis angle has been extensively calibrated and follows a known pattern that needs to be taken into account to recover the true flux radiated by a source.

Apparent on the image of Figure 3.1 are bright spots called point sources. The vast majority of these sources are active galactic nuclei, which originate from material falling onto a supermassive black hole located at the center of a galaxy. Since they are not originated from the galaxy cluster under study, the estimation of emissivity should be robust to potential point sources.

3.1.3 State-of-the-art "onion peeling" deprojection

Traditionally, the main approach used to solve (3.2) has been by inverting directly the convolution [e.g. Fabian et al., 1981, Kriss et al., 1983]. Consider a segmentation of the radius r from the center of the image of Figure 3.1 to the corner of the image into P intervals $[r_i, r_{i+1})$ for $i = 1, \ldots, P$. Within the region encompassed between projected radii r_i and r_{i+1}, the counts are averaged to give an estimate $\hat{I}_i$ of the quantity of radiation received. This amounts to discretizing (3.2) such that the projection kernel reduces to an upper-triangular convolution matrix V, where the matrix element $V_{i,j}$ correspond to the volume of the spherical shell j projected along the line of sight of annulus i [Kriss et al., 1983]. The averaged counts $\hat{I}_i$ are related to the intrinsic 3D emissivity in the spherical shell between r_i and r_{i+1} as

$$\hat{I}_i = \sum_{j=1}^{P} V_{i,j}\varepsilon_j + \mathrm{error}_i, \quad i = 1, \ldots, P \tag{3.3}$$

where P is the number of spherical shells. If P grows with the number of pixels of the image, the method is nonparametric in the sense that the underlying model increases complexity (here, the emissivity $\epsilon(r)$ is assumed piecewise constant on $[r_i, r_{i+1})$ for $i = 1, \ldots, P$) as the number of pixels increases. If P is fixed, the method is parametric. Since the projection matrix $\mathbf{V}$ is upper triangular, the deprojected profile can be evaluated starting from the outermost shell (where projection effects are assumed to be negligible) and then solving (3.3) iteratively when proceeding inwards (hence the nickname of "onion peeling").

This method has the advantage of being nonparametric, if the level of discretization P is large and grows with the number of pixels of the image. Nonparametric

methods do not lead to strong biases caused by a wrongly specified low dimensional model, at the cost of more variance however. This method suffers from severe drawbacks. As already discussed in the introduction, this method is very sensitive to measurement uncertainties, since small variations in the projected profile can be greatly magnified; therefore, the resulting profile is generally not smooth. Moreover, the propagation of statistical fluctuations can result in unphysical negative emissivities. This method also requires that the position of contaminating point sources be estimated in a first step, so as to mask the corresponding areas prior to applying the algorithm.

To alleviate these issues, many variants of the direct deprojection technique exist, including a correction for edge effects [McLaughlin, 1999], spectral information [Nulsen and Bohringer, 1995, Pizzolato et al., 2003], or emission-weighted volumes [Morandi et al., 2007]. However, from the point of view of the mathematical treatment these procedures are similar.

In summary, the current method is a two step method (identify, mask the point sources, and then estimate the emissivity) that does not model well the stochastic nature of the data and that propagates errors from the outskirt of the galaxy cluster (large radius) to the center of the cluster.

3.2 A nonparametric Poisson linear inverse model

3.2.1 Astrophysical and instrumental features

The salient features of the astrophysical data described above can be summarized as follows:

- *Feature 1: Presence of point sources.* Many bright spots are observed on the image. They are the so-called point sources, that is, sources with an angular size that is much smaller than the angular resolution of the telescope. Their location is unknown.

- *Feature 2: Brightness of point sources.* Although point sources are expected to be much smaller than the size of a pixel, their apparent size is much larger. This is due to the finite precision of the alignment of the telescope, which induces a blurring effect that has been well studied and can be considered as known.

- *Feature 3: Telescope artifacts.* There are artifacts in the form of lines that are due

to the poor sensitivity of the telescope at the connection between the various chips.

- *Feature 4: Approximate spherical symmetry.* Near its center, the image has a region of high intensity: it is the center of a galaxy cluster where the gas density is high. The emissivity decreases sharply towards the outskirts, implying that the gas density drops radially. The overall shape is nearly spherically symmetric, exception made of the point sources.

- *Feature 5: Random counts.* Each pixel is a random count of X-rays during a time of exposure.

3.2.2 Model

To account for these specificities, we propose the following model. Considering the telescope first, each image pixel indexed by (x, y) is modeled as

$$Y_{x,y} \sim \text{Poisson}(\mu_{x,y}) \quad \text{for} \quad x = 1, \dots, N \quad \text{and} \quad y = 1, \dots, N, \tag{3.4}$$

where $\mu_{x,y}$ reflects the integral of the intrinsic emissivity of the cosmos. Without the presence of any cosmological background, the XMM telescope has its own electronic noise with small and known mean counts $e_{x,y} \geq 0$. In other words, without any cosmological object facing the telescope, we have $\mu_{x,y} = e_{x,y}$, which can be seen as a known offset.

Considering now the cosmos, each pixel faces a region of the cosmos along a line going from zero (the captor) to infinity. Some lines go through the galaxy cluster, some go through a point source, other go through both. Calling $\epsilon(x, y, z) \geq 0$ the emissivity of the galaxy cluster along that line and $S_{x,y} \geq 0$ a potential point source, the integral of the cosmos emissivity along that line is

$$I_{x,y} = \int_0^\infty \epsilon(x, y, z)dz + S_{x,y} \quad \text{for} \quad x = 1, \dots, N \quad \text{and} \quad y = 1, \dots, N. \tag{3.5}$$

Moreover, owing to the rare existence of point sources (see Feature 1), S is a sparse $N \times N$ matrix.

The connection between $\mu_{x,y}$ and $I_{x,y}$ depends on the characteristics of the telescope. The blurring effect (Feature 2) is known through the so-called point spread function of the telescope. Likewise the sensitivity of the telescope (Feature 3) is known. As a result, the Poisson intensity in (3.4) is modeled as

$$\mu_{x,y} = g(e_{x,y} + (B(E \circ \boldsymbol{I}))_{x,y}), \tag{3.6}$$

where g is the identity function, B is the known blurring operator, E is the known $N \times N$ matrix representing the sensitivity of the telescope at each pixel, and $\circ$ is the notation for the Hadamard product between two matrices. We pause here to make an important remark for statisticians. The Poisson counts (3.4) are linked to the unknown parameters (3.5) through a linear model, which belongs to the class of nonparametric generalized linear models [Nelder and Wedderburn, 1972, Green and Silverman, 1994, Wood, 2017]. In Statistics, the canonical inverse link function $g = \exp$ is often used since it necessarily leads to positive Poisson intensities $\mu_{x,y}$ in (3.6). Physical considerations forces g in (3.6) to be the identity function in this astrophysics application of GLMs, however. A consequence is that standard codes used for estimation do not apply (see Section 3.3.1).

The unknown objects are the gas emissivity $\epsilon(x, y, z)$ as well as the location and intensities of the point sources S. An assumption is needed to estimate the three-dimensional gas density function because the problem is unidentifiable in its current form. Indeed, an infinite number of 3D-functions have the same 2D projection, that is, one cannot recover $\epsilon(x, y, z)$ from $\int \epsilon(x, y, z)dz$. Feature 4 states that a good approximation of the shape of the galaxy cluster is that it is spherical, that is, $\epsilon(x, y, z) = \epsilon_R(r)$ with $r = \sqrt{x^2 + y^2 + z^2}$ is radial. Invariance by rotation makes the problem simpler since the emissivity is known through a univariate function $\epsilon_R(r)$ of the distance r to the center must be estimated. The association is moreover linear since the integral in (3.5) becomes

$$\int_0^\infty \epsilon(x, y, z)dz = 2 \int_{\sqrt{x^2+y^2}}^\infty \frac{r\epsilon_R(r)}{r^2 - x^2 - y^2} dr =: (A\epsilon_R)(x, y), \tag{3.7}$$

where A is called the Abel transform.

The final assumption we make is that ϵ_R has a sparse representation on basis functions ϕ_p:

$$\epsilon_R(r) = \gamma_0 + \sum_{p=1}^{P} \gamma_p \phi_p(r), \tag{3.8}$$

where P is the number of basis functions used. Future telescopes will be more precise in many ways, including their number N^2 of pixels that tends to increase. The asymptotic relevant to our estimation procedure is defined as the limit as N tends to infinity, for fixed exposure time and window size of the observed scene. If the number P of basis function ϕ_p in (3.8) grows with N, the method is called nonparametric: it becomes more flexible as the image gets more precise. Here we choose P of the order of N, more precisely $P = 2^{\lfloor \log_2(N) \rfloor}$. Such method can fit the underlying emissivity

better than a parametric method, provided the many coefficients $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_P)$ can be well estimated from the data. Our method is nonparametric by choosing P of the order N (more details below). Our choice of basis functions ϕ_p is important and partly based on prior knowledge. Cosmologists expect a decreasing function from the center of the galaxy cluster to its outskirt. So we use a generalization of the so-called *King*'s functions

$$\phi_p(r) = (1 + (r/\rho)^2)^{-\beta}, \quad \rho \in \{\rho_1, \ldots, \rho_I\}, \, \beta \in \{\beta_1, \ldots, \beta_J\} \tag{3.9}$$

parametrized by $p = (\rho, \beta)$ [Eckert et al., 2016]. A grid of (ρ, β) lead to $P/2$ such functions. To allow more flexibility and discover galaxy clusters with singularities, we also use $P/2$ orthonormal wavelets defined on equispaced radii [Daubechies, 1992, Donoho and Johnstone, 1994]. By default we choose Daubechies wavelets of order 8 for estimation of the emissivity for their smoothness; other wavelets are considered for uncertainty quantification in Section 3.3.2. Another possible (nonlinear) approach proposed by a reviewer would consist in estimating ρ and β, but we prefer the more conventional linear approach of defining a large set of basis functions.

The emissivity function $\epsilon_R(r)$ defined on $\mathbb{R}^+$ typically has a peak at zero and decreases (often monotonically) to zero as r gets large. Because there is a big discrepancy between the left boundary, typically a peak, and the right boundary, typically flat, wavelets used in (3.8) will have difficulties even with wavelet boundary corrections. Various boundary schemes have been proposed, the simplest one assumes periodicity, which is clearly violated here. We overcome this difficulty by splitting the original image into two half-images going through the center of the galaxy cluster, for instance the left image and the right image. Each half faces half of the galaxy cluster. Let us call ϵ_R^{left} and $\epsilon_R^{\text{right}}$ the corresponding emissivities. If the galaxy cluster is exactly spherical then $\epsilon_R^{\text{left}}(r) = \epsilon_R^{\text{right}}(r)$ for all $r \geq 0$, otherwise they share the same value at $r = 0$ and both tend to zero when the radius r is large. Hence the double emissivity function

$$\epsilon_R^{\text{left}\cup\text{right}}(r) = \epsilon_R^{\text{left}}(-r) \cdot 1(r < 0) + \epsilon_R^{\text{right}}(r) \cdot 1(r \geq 0) \tag{3.10}$$

defined for negative radii (left part of the galaxy cluster) and for positive radii (right part) can be well represented as a linear combination of wavelets with periodic boundaries. Plotting both left and right estimated emissivities can reveal asymmetry in the cluster, or can be averaged to provide the cosmologist with a single emissivity curve. Note that instead of splitting the image into a left and right sectors, one could also split into more sectors where the sphericity assumption seems to better hold.

Putting all components together leads to the following linear model for the Poisson parameters:

$$\mu_{x,y} = e_{x,y} + (B(E \circ (A(\gamma_0 \mathbf{1} + \Phi\boldsymbol{\gamma}) + \mathbf{s})))_{x,y}, \tag{3.11}$$

where Φ is $N \times P$ matrix of discretized basis function ϕ_p (namely, $\Phi_{n,p} = \phi_p(r_n)$ for a grid of radii $r_1, \ldots, r_N$), and the unknown parameters are γ_0 for the intercept vector $\mathbf{1}$ (vector of ones), the sparse N-vector $\boldsymbol{\gamma}$ of the linear expansion (3.8) and the sparse $N \times N$-matrix S of potential point sources put in vector form $\mathbf{s}$. This is a linear inverse problem in the sense that the unknown quantities are indirectly observed through the linear operators.

3.2.3 Taking asymmetry into account

Feature 4 states that galaxy clusters are only approximately symmetric. In practice, some galaxy clusters are strongly asymmetric (see for instance the images of Figure 3.4). In an attempt to be robust to asymmetry, we model the emission ϵ_R into a left and right parts in (3.10), as discussed at the end of the previous section. The potential left/right asymmetry of the emissivity is reflected in (3.11) in the way the matrix Φ is built. The $\mathbf{s}$-term in (3.11) originally introduced to detect point sources also helps take into account asymmetry by combining the point sources with the residuals due to a lack of symmetry. Looking at the image $\hat{s}$ may reveal spatial features pointing towards asymmetry in the observed galaxy cluster, as we do with five cosmology images in Figure 3.4.

3.3 Estimation with two sparsity constraints

3.3.1 Estimation of emissivity

Based on Feature 5, the Poisson negative log-likelihood

$$-l(\gamma_0, \boldsymbol{\gamma}, S; \mathbf{y}) \;\; = \sum_{(x,y)\in\{1,\ldots,N\}^2} \mu_{x,y} - Y_{x,y} \log \mu_{x,y} \tag{3.12}$$

is a natural measure of goodness-of-fit of the counts data to the linear model for $\mu_{x,y}$ (3.11). This model is a generalized linear model (GLM) for Poisson noise with identity link. Note that the log-term in (3.12) prevents the estimated Poisson intensities from being negative.

The number $1 + 2^{\lfloor \log_2(N) \rfloor} + N^2$ of parameters $(\gamma_0, \boldsymbol{\gamma}, \mathbf{s})$ exceeds the number of observations N^2, so that regularization is needed. Owing to the sparse representation of the univariate gas density on its basis functions and to the rare existence of point sources, we regularize the likelihood by enforcing sparsity on the estimation of $\boldsymbol{\gamma}$ and $\mathbf{s}$ with two ℓ_1 penalties

$$(\hat{\gamma}_0, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{s}})_{\lambda_1,\lambda_2} = \arg\min_{\gamma_0,\boldsymbol{\gamma},\mathbf{s}} -l(\gamma_0, \boldsymbol{\gamma}, \mathbf{s}; \mathbf{y}) + \lambda_1 \|\boldsymbol{\gamma}\|_1 + \lambda_2 \|\mathbf{s}\|_1 \tag{3.13}$$

in the spirit of lasso [Tibshirani, 1996, Sardy et al., 2004] and glmnet [Park and Hastie, 2007]. We rely on FISTA [Beck and Teboulle, 2009] to solve the high dimensional and non-differentiable optimization problem for given hyperparameters (λ_1, λ_2). This gradient-based algorithm can solve general convex program by successively solving quadratic approximation to the cost function by means of the soft-thresholding function. It has the advantage over `glmnet` to handle the identity link function and positivity constraints on the King's coefficients, and does no require building and storing a very large matrix. Our current implementation of FISTA algorithm is in `MATLAB`.

The selection of the regularization parameters (λ_1, λ_2) is a key issue. Performing cross validation on a 2D-grid would be computationally intensive and would require segmenting the image into sub-images. The empirical Bayes approach is another possible avenue that would entail calculating some multivariate integrals. Theoretical results on lasso [Bühlmann and van de Geer, 2011] and wavelet smoothing with the universal threshold of Donoho and Johnstone [1994] show that, for good prediction and model selection, the threshold should have the property to reproduce the true signal with a probability tending to one asymptotically (i.e., as the size of the image N tends to infinity) when the true signal is the constant function. Such a choice of λ has remarkable asymptotic near minimax properties when the function to estimate is not null (e.g., existence of a galaxy cluster with some emissivity and existence of point sources), in particular for an emissivity function $\epsilon_R(r)$ that belongs to a Besov space [Donoho et al., 1995]. From a practical point of view, this property means that a radial emissivity with sharp changes like a peak or an abrupt change in first derivative can be well recovered by the procedure. And the asymptotic property means that as future telescopes will increase their pixel resolution (i.e., the number of pixels N^2 tending to infinity), the universal threshold leads to an emissivity estimation that is near minimax (i.e., it minimizes the worst estimation for emissivity functions in a Besov space).

The quantile universal threshold is the extension of the universal threshold to other noise distributions, models and estimators [Giacobino et al., 2017]. It is defined as the upper quantile of the null thresholding statistic of a thresholding estimator. As defined in Section 1.3.2, the null thresholding statistics requires the zero thresholding function of the corresponding thresholding estimator. The derivation of the quantile universal threshold for (3.13) requires new results because the link function g in (3.6) is not the canonical one, and because two penalties are involved. First we derive the zero-thresholding function for (3.13).

Property 3.1 *Given an image* $\mathbf{y}$*, the smallest* λ_1 *and* λ_2 *that jointly set* $(\hat{\boldsymbol{\gamma}}, \hat{s})_{\lambda_1,\lambda_2}$ *in* (3.13) *to zero is given by the zero-thresholding function*

$$\lambda(\mathbf{y}) = (\lambda_1(\mathbf{y}), \lambda_2(\mathbf{y})) := \begin{cases} \left(\|X_1^{\mathrm{T}} \left(\frac{\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda(\hat{\gamma}_0)}{\hat{\boldsymbol{\mu}}_\lambda(\hat{\gamma}_0)} \right) \|_\infty, \|X_2^{\mathrm{T}} \left(\frac{\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda(\hat{\gamma}_0)}{\hat{\boldsymbol{\mu}}_\lambda(\hat{\gamma}_0)} \right) \|_\infty \right) & \text{if } \mathbf{y} \in \mathcal{D} \\ (+\infty, +\infty) & \text{otherwise} \end{cases}, \tag{3.14}$$

where $\hat{\boldsymbol{\mu}}_\lambda(\hat{\gamma}_0) = \mathbf{e} + \mathbf{x}_0\hat{\gamma}_0$*,* $\mathbf{x}_0 = BE \circ A\mathbf{1}$*,* $X_1 = BE \circ A\Phi$*,* $X_2 = BE \circ A$ *and* $\mathcal{D} = \{\mathbf{y} : \exists \hat{\gamma}_0 \in \mathbb{R} \text{ satisfying } \hat{\mathbf{x}}_0^{\mathrm{T}}\mathbf{1} = \mathbf{x}_0^{\mathrm{T}}(\mathbf{y}/(\mathbf{e} + \mathbf{x}_0\hat{\gamma}_0)) \text{ and } \mathbf{e} + \mathbf{x}_0\hat{\gamma}_0 > \mathbf{0}\}$.

Proof. The KKT conditions for (3.13) at $\boldsymbol{\gamma} = \mathbf{0}$ and $\mathbf{s} = \mathbf{0}$ are

$$\begin{aligned} \partial/\partial\gamma_0 : \ \mathbf{x}_0^{\mathrm{T}} \left(\frac{\boldsymbol{\mu} - \mathbf{y}}{\boldsymbol{\mu}} \right) &= 0 \\ \nabla_{\boldsymbol{\gamma}} : \ X_1^{\mathrm{T}} \left(\frac{\boldsymbol{\mu} - \mathbf{y}}{\boldsymbol{\mu}} \right) &\in \lambda_1 \mathcal{B}^\infty \\ \nabla_{\mathbf{s}} : \ X_2^{\mathrm{T}} \left(\frac{\boldsymbol{\mu} - \mathbf{y}}{\boldsymbol{\mu}} \right) &\in \lambda_2 \mathcal{B}^\infty \end{aligned}$$

where $\mathcal{B}^\infty$ is the ℓ_∞-unit ball and $\mu = \mathbf{e} + \mathbf{x_0}\gamma_0$. The first equation has a solution provided $\mathbf{y} \in \mathcal{D} = \{\mathbf{y} : \exists \hat{\gamma}_0 \in \mathbb{R} \text{ satisfying } \hat{\mathbf{x}}_0^{\mathrm{T}}\mathbf{1} = \mathbf{x}_0^{\mathrm{T}}(\mathbf{y}/(\mathbf{e} + \mathbf{x}_0\hat{\gamma}_0)) \text{ and } \mathbf{e} + \mathbf{x}_0\hat{\gamma}_0 > \mathbf{0}\}$, and the smallest λ_i allowing this system to have a solution are $\lambda_i = \|X_i^{\mathrm{T}} \left(\frac{\boldsymbol{\mu} - \mathbf{y}}{\boldsymbol{\mu}} \right) \|_\infty$ for $i \in \{1, 2\}$ ■

Second we define the corresponding null-thresholding statistic.

Definition 3.1 *The null-thresholding statistic* Λ_0 *for* $(\hat{\boldsymbol{\gamma}}, \hat{s})_{\lambda_1,\lambda_2}$ *in* (3.13) *is*

$$\Lambda_0 = (\Lambda_0^{(1)}, \Lambda_0^{(2)}) := (\lambda_1(\mathbf{Y}_0), \lambda_2(\mathbf{Y}_0)) \quad \text{with} \quad \mathbf{Y}_0 \sim \text{Poisson}(\mathbf{e} + \mathbf{x_0}\gamma_0). \tag{3.15}$$

Note that $\mathbf{Y}_0$ has mean $\mathbf{e} + \mathbf{x_0}\gamma_0$, that is, the zero-scene assumes zero emissivity (i.e., $\boldsymbol{\gamma} = \mathbf{0}$) and no point source (i.e., $\mathbf{s} = \mathbf{0}$). The goal of our selected hyperparameters $(\lambda_1^{\mathrm{QUT}}, \lambda_2^{\mathrm{QUT}})$ is to reproduce this zero-scene with high probability. This is achieved with the third step by taking marginal quantiles of the null-thresholding statistic.

Definition 3.2 *The quantile universal thresholds* $(\lambda_1^{\mathrm{QUT}}, \lambda_2^{\mathrm{QUT}})$ *are the upper* α_1*-quantile of* $\Lambda_0^{(1)}$ *for* λ_1 *and the upper* α_2*-quantile of* $\Lambda_0^{(2)}$ *for* λ_2.

The quantile universal thresholds has the following desired property.

Property 3.2 *With* $(\lambda_1^{\mathrm{QUT}}, \lambda_2^{\mathrm{QUT}})$*, the estimator* (3.13) *reproduces the zero-scene with probability at least* $1 - \alpha_1 - \alpha_2$ *since* $\mathbb{P}((\hat{\boldsymbol{\gamma}}, \hat{\mathbf{s}})_{\lambda_1^{\mathrm{QUT}}, \lambda_2^{\mathrm{QUT}}} = (\mathbf{0}, \mathbf{0}); \boldsymbol{\gamma} = \mathbf{0}, \mathbf{s} = \mathbf{0}) \geq 1 - \alpha_1 - \alpha_2$.

Proof. From Property 3.1 $\mathbb{P}((\hat{\boldsymbol{\gamma}}, \hat{\mathbf{s}})_{\lambda_1^{\mathrm{QUT}}, \lambda_2^{\mathrm{QUT}}} = (\mathbf{0}, \mathbf{0}); \boldsymbol{\gamma} = \mathbf{0}, \mathbf{s} = \mathbf{0}) = \mathbb{P}(\lambda_1^{\mathrm{QUT}} > \lambda_1(\mathbf{Y}_0) \cap \lambda_2^{\mathrm{QUT}} > \lambda_2(\mathbf{Y}_0)) = 1 - \mathbb{P}(\lambda_1^{\mathrm{QUT}} \leq \lambda_1(\mathbf{Y}_0) \cup \lambda_2^{\mathrm{QUT}} \leq \lambda_2(\mathbf{Y}_0)) \geq 1 - \alpha_1 - \alpha_2$. ■

In practice, the choice of α_1 and α_2 can be guided by the following considerations. Since the former is linked to the estimation of the emissivity function ϵ_R, we choose $\alpha_1 = 1/\sqrt{\pi \log P}$ as for the universal threshold of Donoho and Johnstone [1994] in the Gaussian case. The latter is linked to the identification of the point sources, so we recommend for instance $\alpha_2 = 1/N^2$ to control the false discovery rate at level α_2 in the weak sense: with $\alpha_2 = 1/N^2$, the average number of falsely detected point sources is one per image when no point sources are present.

3.3.2 Uncertainty quantification

Estimation of the emissivity is of little value for astrophysical purposes without uncertainty quantification, which helps judging whether a feature is significant or not. The estimation of the emissivity ϵ_R is conditional on the observed (random) image, the location of the center of the galaxy cluster and the choice of the bases in the expansion (3.8). To quantify the uncertainty related to randomness of the image, we first segment the image into boxes of size 4×4. Each pixel in each box faces approximately the same region of the universe, therefore assuming the emissivity

is approximately constant in each box, the 16 pixels can be seen as being approximately i.i.d., and therefore can be bootstrapped within each box [Efron and Tibshirani, 1993]. Since the location of the center of the cluster is prone to uncertainty, we also randomly select the center of the image. To prevent conditioning the estimation on a particular type of wavelets in the expansion (3.8) so as to avoid artifacts, we also choose randomly one type of wavelets out of nine (three Daubechies, three symmlets, three coiflets [Daubechies, 1992]). The procedure is repeated M times and pointwise $(1-\alpha)$-quantiles of these estimated emissivity curves provide a measure of uncertainty.

3.4 Numerical experiments

3.4.1 Simulated data

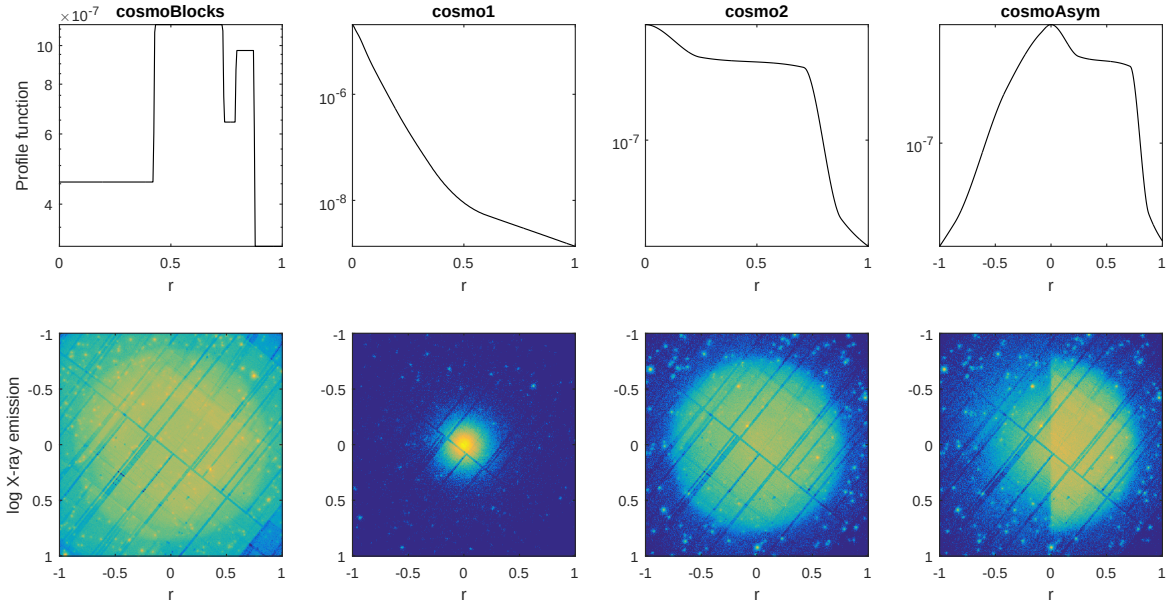

Figure 3.2: Three different simulation profiles (top row) with a corresponding simulated galaxy cluster images (bottom row).

We simulate galaxy clusters according to model (3.11) with known constant background $e_{x,y} = 10^{-4}$, known sensitivity matrix E and blurring operator B correspond-

ing to the point spread function

$$\mathrm{psf}(r; r_0, a) = \left(1 + \left(\frac{r}{r_0}\right)^2\right)^{-a} \tag{3.16}$$

of the XMM telescope ($a = 1.449$ and $r_0 = 2.2364$ pixels). The simulations are based on four profile functions $\epsilon_R(r)$. The first three defined for $r \in [0, 1]$ assume symmetry: `cosmoBlocks` is a cropped version of the well known standard function `blocks` used in signal processing [Donoho and Johnstone, 1994] and although not expected to describe a galaxy cluster, it allows to show the flexibility of our procedure; `cosmo1` and `cosmo2` are typical symmetric profiles according to cosmologists; the `cosmoAsym` profile defined for $r \in [-1, 1]$ has some strong asymmetry to illustrate how to estimate left and right emissivities. These four functions are publicly available, as discussed in Section 3.6.

For each test profile, we simulate galaxy cluster images of size $N \times N$ for $N \in \{128, 256, 512\}$ and repeat the Monte Carlo $M \in \{96, 48, 24\}$ times, respectively, to estimate the mean squared error. We consider two scenarios: without and with point sources to quantify the robustness of the methods to the presence of point sources. To make the simulation realistic, we simulate a total of N points sources that follow integral source flux distribution power law with index $\alpha = 1.82$ with maximum boundary flux $S_{\max} = 500\bar{I}$ and $\bar{I}$ is the average intrinsic emissivity of the source without point sources [Moretti et al., 2003, eq. (2) and (4)].

We compare our estimator (QUT-lasso) to the state-of-the-art method used by cosmologists (SA) described in Section 3.1.3. Recall that the SA method is a two step method: first estimate the location of potential point sources, then perform the deprojection. We help the SA method by being oracle in the first step: since we are doing a simulation, we know where the point sources are and provide this information through the sensitivity matrix E in that $E_{x,y} = 0$ when pixel (x, y) has a point source. For the first three emissivities that are symmetric, the estimation is the average between the left and right emissivities in (3.10).

Table 3.1 reports the estimated mean square error between $\log \hat{\epsilon}_R$ and $\log \epsilon_R$ for each simulation in column F. The other three columns break the estimated mean square error into three even regions: inner (i), middle (m) and outer (o). Table 3.1 shows that QUT-lasso performs much better than the state-of-the-art method, without and with point sources, whether in the heart or to the outskirt of the galaxy cluster. A region of particular interest to cosmologists is the outer shell of the galaxy cluster

Table 3.1: Results of Monte Carlo simulation for the mean squared errors.

		MSE of the log-profile (*100)							
		cosmoBlocks				cosmo1			
M		F	i	m	o	F	i	m	o
Without									
128	QUT-lasso	8.5	14	4	6.5	6.1	6.3	3	7.9
	SA	*21*	*41*	*5.4*	*15*	*33*	*55*	*34*	*6*
256	QUT-lasso	3.8	1.9	4.6	4.4	1.1	1.7	0.66	0.77
	SA	*15*	*26*	*3.8*	*14*	*8.6*	*17*	*7.5*	*2.5*
512	QUT-lasso	2.5	2.2	3.2	2	0.24	0.53	0.085	0.2
	SA	*12*	*20*	*2.9*	*14*	*2.7*	*4.7*	*2.3*	*1.1*
With									
128	QUT-lasso	24	55	11	6.7	17	5.9	15	29
	SA	*32*	*59*	*8.2*	*25*	*63*	*56*	*49*	*76*
256	QUT-lasso	29	77	5.5	6.7	4	1.6	4.2	4.4
	SA	*20*	*39*	*4.8*	*17*	*15*	*15*	*11*	*17*
512	QUT-lasso	4.5	7.4	3.6	2.2	1	0.76	0.58	1.3
	SA	*15*	*24*	*3.3*	*16*	*4*	*4.7*	*3.2*	*3.5*

		cosmo2				cosmoAsym			
M		F	i	m	o	F	i	m	o
Without									
128	QUT-lasso	7.2	3.6	0.36	17	5.4	1.3	1.4	13
	SA	*53*	*74*	*0.71*	*79*	*42*	*70*	*3.1*	*49*
256	QUT-lasso	2.8	3.3	0.073	4.8	0.99	0.97	0.13	1.8
	SA	*29*	*37*	*0.21*	*51*	*26*	*50*	*1.4*	*25*
512	QUT-lasso	0.95	1.7	0.11	0.94	0.71	1.1	0.098	0.91
	SA	*15*	*30*	*0.057*	*14*	*19*	*47*	*0.5*	*8.4*
With									
128	QUT-lasso	19	7.5	1.9	41	13	5.4	2.2	28
	SA	*60*	*67*	*0.72*	*109*	*49*	*67*	*5.2*	*74*
256	QUT-lasso	2.8	4.1	0.28	3.9	1.4	1.3	0.39	2.5
	SA	*33*	*40*	*0.16*	*60*	*32*	*54*	*2*	*34*
512	QUT-lasso	1.2	1.5	0.12	1.6	1	0.93	0.18	1.9
	SA	*17*	*35*	*0.054*	*17*	*20*	*50*	*0.68*	*10*

where QUT-lasso outperforms the state-of-the-art method in the presence of point sources. As expected, we observe that QUT-lasso is robust to point sources thank to the s-term in (3.11), and MSE decreases for the asymptotic we considered, namely when the number N^2 of pixels increases.

In the particular case of `cosmoBlocks` for $N = 256$ with point sources in Table 3.1, SA is better than QUT-lasso. This is since `cosmoBlocks` is the only test function that does not have a peak in the inner region. By adding point sources, QUT-lasso sometimes does not estimate a point source located in the center of the image as a point source, but as part of the emissivity. Currently we help SA method with the mask of the true locations of point sources. If we were not to help SA method then it would also suffer from this problem. This is fixed increasing the image size since the size of a point source becomes smaller in relation to the center of the image and it becomes easier to identify it, as we see when $N = 512$.

We also investigate the coverage probability of the bootstrap-based uncertainty quantification method of Section 3.3.2 with a Monte Carlo simulation summarized in Table 3.2. Aiming at the target nominal value of 95% pointwise coverage probability, QUT-lasso achieves a reasonable coverage especially in the absence of point sources. The SA method reaches a much lower coverage probability. We conclude that this approach provides the astrophysicist with a reasonable guidance to judge the significance of interesting features. To visually illustrate the benefit of pointwise coverage with QUT-lasso, Figure 3.3 shows typical pointwise measures of uncertainty for the four test signals. We observe that as the sample size increases from $N = 256$ to $N = 512$, QUT-lasso gets better estimation of the emissivity and narrower pointwise confidence intervals. The SA method shows some strong bias and wider intervals.

QUT-lasso outperforms SA in the two areas of great interest to cosmologists, the inner and outer regions, both in terms of MSE and coverage probability. Figure 3.3 reveals the typical respective behaviors of both methods. In particular, SA tends to overestimate emissivity in the outskirt and pointwise confidence regions with SA are too wide in the inner region, showing great sensitivity of the method to small perturbation of the data. In particular, cases where coverage percentage in the inner part is almost 100% for SA are due to wide coverage regions, and it should not mislead to good properties of SA. In summary, this Monte Carlo simulation shows great improvement with our method in comparison with the SA method.

Table 3.2: Results of Monte Carlo for pointwise coverage percentage at the 95% nominal level.

		Coverage percentage							
		cosmoBlocks				cosmo1			
M		F	i	m	o	F	i	m	o
Without									
256	QUT-lasso	82	100	68	81	99	97	100	100
	SA	*37*	*35*	*54*	*22*	*64*	*11*	*77*	*100*
512	QUT-lasso	73	65	76	78	99	98	100	100
	SA	*9*	*17*	*5*	*4*	*59*	*17*	*70*	*89*
With									
256	QUT-lasso	73	64	84	72	89	88	81	100
	SA	*49*	*50*	*61*	*38*	*39*	*26*	*81*	*11*
512	QUT-lasso	66	47	73	78	77	94	100	39
	SA	*27*	*57*	*18*	*6*	*37*	*21*	*79*	*12*

		cosmo2				cosmoAsym			
M		F	i	m	o	F	i	m	o
Without									
256	QUT-lasso	79	38	100	100	82	62	88	96
	SA	*62*	*100*	*77*	*13*	*43*	*87*	*29*	*15*
512	QUT-lasso	71	36	77	100	81	61	84	98
	SA	*46*	*100*	*30*	*10*	*31*	*86*	*6*	*2*
With									
256	QUT-lasso	89	71	100	97	62	62	40	84
	SA	*78*	*100*	*100*	*36*	*41*	*79*	*32*	*15*
512	QUT-lasso	71	29	94	91	56	33	69	66
	SA	*71*	*95*	*100*	*19*	*41*	*68*	*47*	*8*

3.4.2 Real data

We now apply QUT-lasso and the SA method to five telescope images shown on the first column of Figure 3.4. These images cover a wide range of cases encountered in the observed population of galaxy clusters. In addition to our test case Abell 2142,

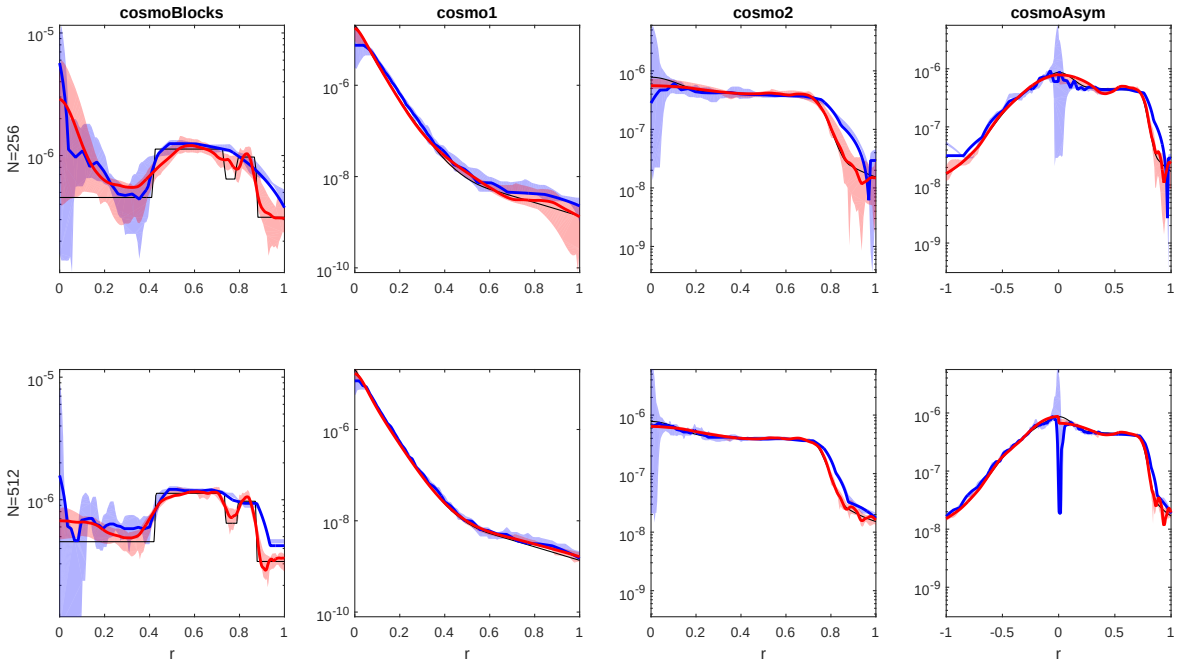

Figure 3.3: Example of an estimated emissivity and its pointwise confidence intervals obtained by bootstrap for images of size $N \times N$ with $N = 256$ (top) and $N = 512$ (bottom). The true emissivity is in black, the SA method is in blue and QUT-lasso is in red.

we present the case of Abell 3667, which features a very sharp contact discontinuity East of the cluster core (Vikhlinin et al. [2001]). This particular case allows us to test our reconstruction method in the case of abrupt changes in the X-ray brightness. The fourth row of Figure 3.4 shows the case of the "Bullet cluster" 1E 0657-56 (Markevitch et al. [2002]), the prototypical merging cluster where a high-velocity subcluster (the "Bullet") has gone through the main cluster. Finally, the last row shows an image of the Perseus cluster, where outflows originating from the supermassive black hole located at the center of the cluster are interacting with the hot gas around, thus creating bubbles and cavities within the surrounding medium (Fabian et al. [2000]). The second column of Figure 3.4 plots the matrix $\hat{S}$, the combination of estimated point sources and residuals due to asymmetry. The third column of Figure 3.4 plots $\hat{\epsilon}_R$, the emissivity estimated either with the SA method in blue or with our proposal in red. Along with the point estimate, we provide the measure of uncertainty discussed in Section 3.3.2.

In the first two rows, we compare the *XMM-Newton* data for our test cluster Abell 2142 with the data acquired for the same target by the *Chandra* spacecraft. The an-

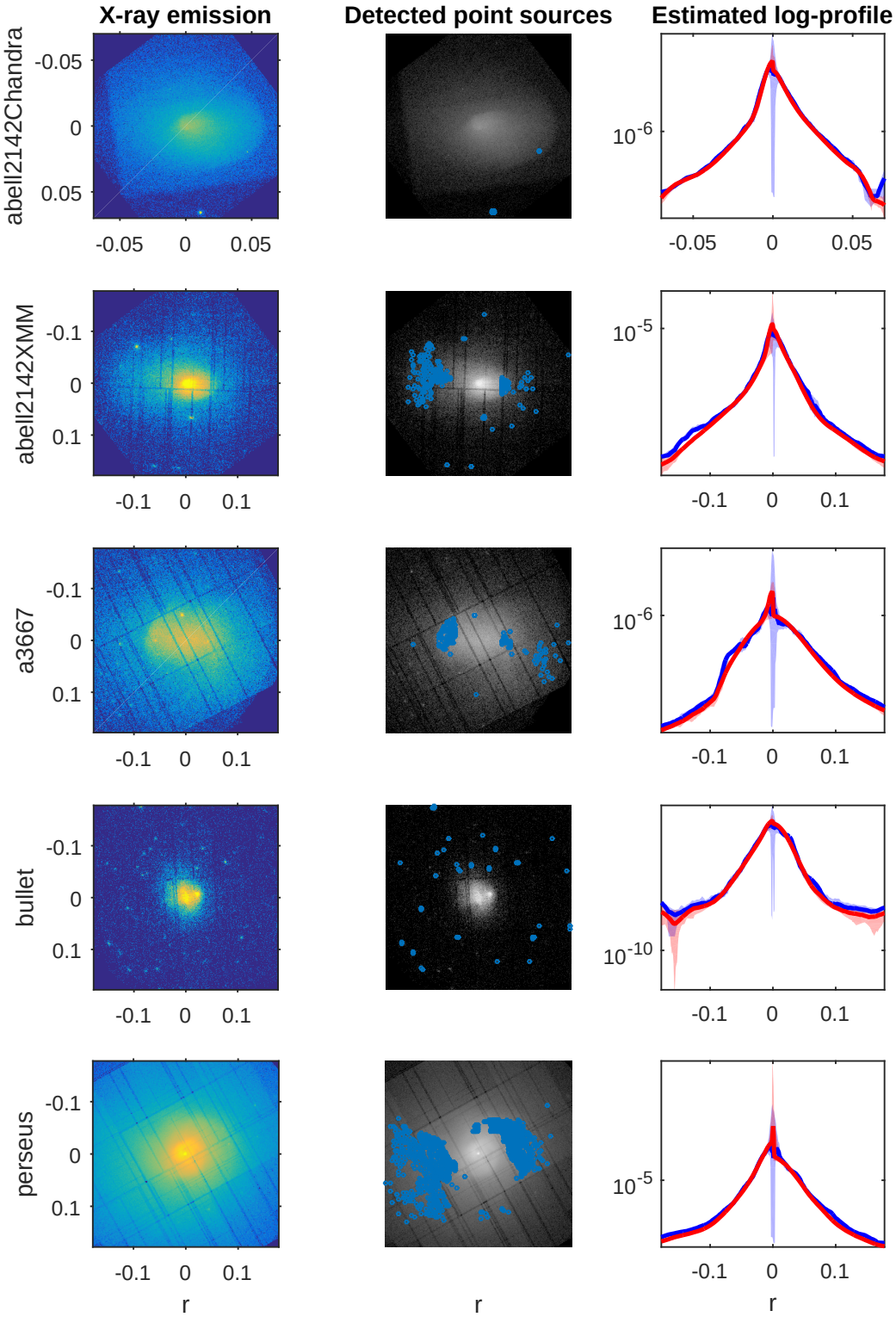

Figure 3.4: Real data results.

gular resolution of *Chandra* (0.5 arcsec) is far superior to that of *XMM-Newton* (8 arcsec), which allows us to observe narrow features with more detail. The better spatial resolution of *Chandra* allows to better sample the shape of the emissivity in the innermost regions, whereas in *XMM-Newton* the peak is smeared out by the point spread function of the telescope. Conversely, *XMM-Newton* is a bigger telescope and covers a wider sky area, making it more sensitive than *Chandra* to detect faint X-ray emission. Taking a closer look at the results, Figure 3.5 plots the four emissivities estimated by the two methods based on the two telescopes in the interval $(-0.06, 0.06)$ and $(-0.16, 0.16)$ for *Chandra* and *XMM-Newton*, respectively. We find an excellent agreement between the results obtained with the two independent telescopes.

The tests performed on three different cases (A3667, Bullet Cluster, Perseus) demonstrate the stability of our method when applied to more complex cases where the gas distribution can deviate substantially from spherical symmetry. In the case of A3667, we find clear differences between the reconstructions obtained in different sectors, as shown in the center-right panel of Fig. 3.4. The profile in the South-West direction appears regular, whereas the profile reconstructed in the North-East direction shows an abrupt drop consistent with the presence of the sharp cold front discovered by Vikhlinin et al. [2001]. We recover a smooth profile as well for the case of the merging Bullet cluster and for the case of the Perseus cluster, where cavities and shocks injected by the central active galactic nucleus are present.

3.4.3 Summary of empirical findings

As shown in Table 3.1, our method outperforms the current state-of-the-art method by providing results that are typically closer to the true value by a factor of three to five on average. Thanks to the use of wavelets in the linear expansion (3.8), QUT-lasso adapts to local features of the emissivity. Moreover our method does not require an *a priori* knowledge of the position of contaminating point sources, but proposes, in a single step, an estimation of the emissivity robust to the presence of point sources. Table 3.2 and Figure 3.3 also shows good coverage by the bootstrap-based pointwise uncertainty quantification. The new QUT-lasso approach provides better estimation in the inner and outer regions which are particularly relevant to astrophysicists. Overestimation of emissivity in the outer region seems to be corrected with QUT-lasso.

Overestimation in the outer region may also occur with the state-of-the-art method

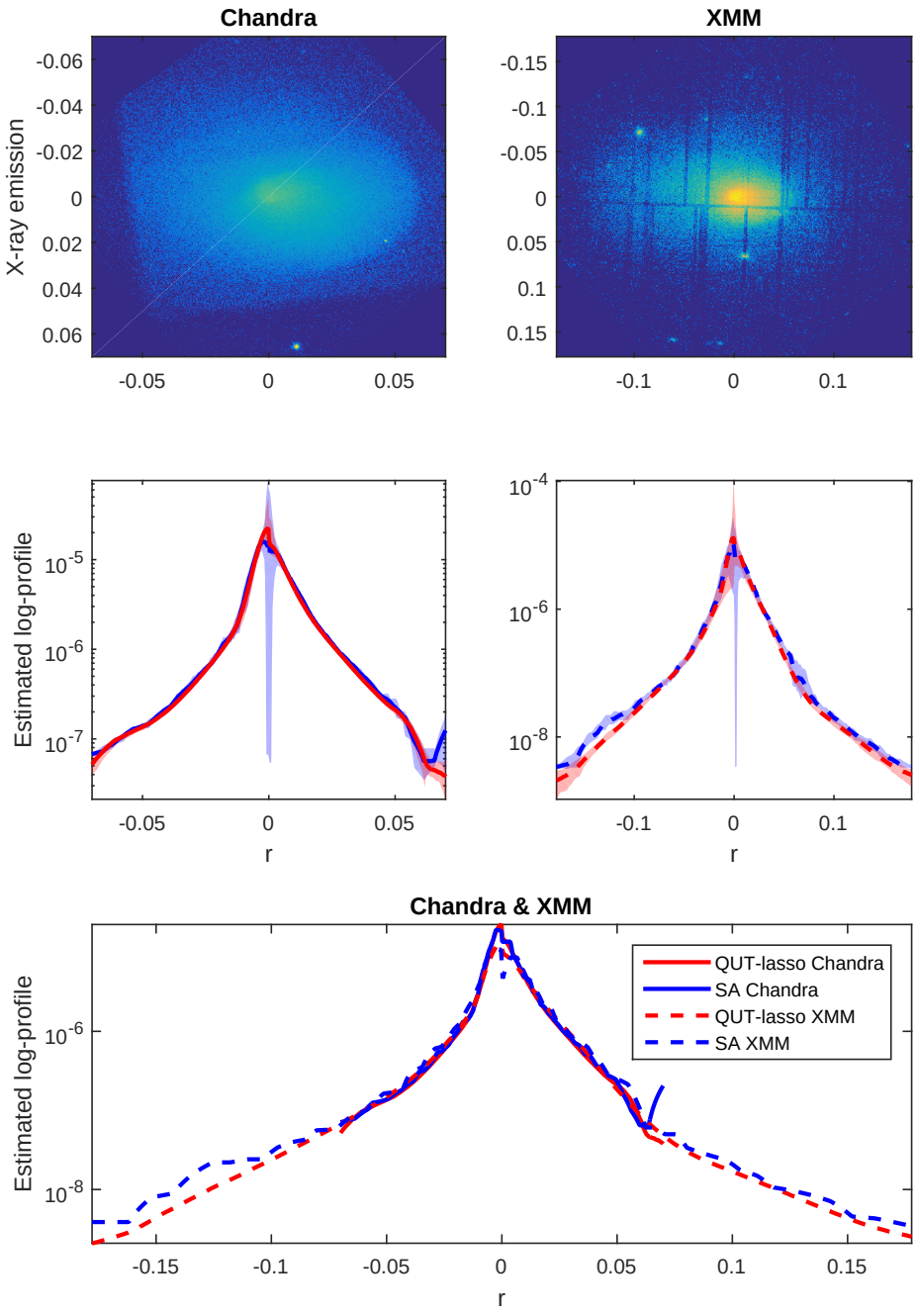

Figure 3.5: Real data results. Top: pictures taken by two telescopes of same galaxy cluster: Chandra (high resolution) and XMM (high sensitivity). Middle: Estimated emissivities by our method (continuous line) and state-of-the-art (dotted line). Bottom: all four estimates on the same plot.

on the five telescope images of Figure 3.5 where the blue curve (SA method) is always above the red one (QUT-lasso). Given that the true emissivity is unknown here, we cannot make a quantitative assessment of this statement, but if it is really the case (as the Monte Carlo simulations suggest) then the SA approach currently used by astrophysicists may mislead findings and interpretation in the outskirt of galaxy clusters. Likewise, the behavior the SA approach in the inner region shows high instability which renders poor estimation. QUT-lasso shows more stability in that region as well, as expected from the numerical experiments performed on simulated data. Application to the Chandra and XMM-Newton telescopes shows on Figure 3.5 good agreement between the profiles reconstructed with QUT-lasso and the standard method, yet with a smoother profile recovered by QUT-lasso.

3.5 Conclusions

In this chapter, we have presented a novel technique to reconstruct the three dimensional properties of an "optically thin" astrophysical source from two-dimensional observations including the presence of background, unrelated point sources and Poisson noise. This method is based on Poisson GLM with identity link and a lasso-type regularization with two regularization parameters that are selected with the quantile universal threshold (QUT). The proposed method is fully automatic and superior by far to the methods that are commonly used in astrophysics. The linear model for the emissivity curve is based on an expansion on basis functions which include wavelets. This makes the QUT-lasso method particularly flexible to discover galaxy clusters with unusual shapes.

Future applications to real data will allow us to reconstruct accurately the three-dimensional gas density profiles in galaxy clusters, which can be used to study the astrophysical properties of the plasma in clusters of galaxies, estimate cosmological parameters, and measure the gravitational field in massive structures to set constraints on dark matter and modified gravity.

3.6 Reproducible research

The code and data that generated the figures in this article may be found online in the following at `http://www.unige.ch/math/folks/sardy/astroRepository`

List of Tables

List of Figures

Bibliography

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

S. W. Allen, A. E. Evrard, and A. B. Mantz. Cosmological Parameters from Observations of Galaxy Clusters. *Annual Review of Astronomy and Astrophysics*, 49:409–470, 2011.

M. Anderson. A new method for non-parametric multivariate analysis of variance. 26:32 – 46, 02 2001.

F. J. Anscombe. The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35(3/4):246–254, 1948.

E. Arias-Castro, E. J. Candès, and Y. Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39 (5):2533–2556, 10 2011.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

J. Benichou and M. Palta. *Handbook of epidemiology*, chapter Risks, Measures of Association and Impact. Springer-Verlag, 2005.

C. I. Bliss. The method of probits. *Science*, 79(2037):38–39, 1934.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg, 2011.

P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 2014.

F. Bunea, J. Lederer, and Y. She. The group square-root lasso: theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60(2):1313–1325, 2014.

J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

E. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.

E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

W. Cash. Parameter estimation in astronomy through application of the likelihood ratio. *The Astrophysical Journal*, 228:939–947, mar 1979.

D. Clowe, M. Brada, A. H. Gonzalez, M. Markevitch, S. W. Randall, C. Jones, and D. Zaritsky. A Direct Empirical Proof of the Existence of Dark Matter. *The Astrophysical Journal Letters*, 648:L109–L113, 2006.

I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

J. Diaz-Rodriguez, S. Sardy, C. Giacobino, and N. Hengartner. *qut: Quantile Universal Threshold*, 2016. URL `https://CRAN.R-project.org/package=qut`. R package version 1.3.

D. L. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and Computational Harmonic Analysis*, 2(2):101–126, 1995.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B*, 57(2):301–369, 1995.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24:508–539, 1996.

D. Eckert, F. Vazza, S. Ettori, S. Molendi, D. Nagai, E. T. Lau, M. Roncarelli, M. Rossetti, S. L. Snowden, and F. Gastaldello. The gas distribution in the outer regions of galaxy clusters. *Astronomy and Astrophysics*, 541:A57, 2012.

D. Eckert, S. Ettori, J. Coupon, F. Gastaldello, M. Pierre, J.-B. Melin, A. M. C. Le Brun, I. G. McCarthy, C. Adami, L. Chiappetti, L. Faccioli, P. Giles, S. Lavoie, J. P. Lefèvre, M. Lieu, A. Mantz, B. Maughan, S. McGee, F. Pacaud, S. Paltani, T. Sadibekova, G. P. Smith, and F. Ziparo. The XXL Survey. XIII. Baryon content of the bright cluster sample. *Astronomy and Astrophysics*, 592:A12, 2016.

B. Efron and R. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, London; New York, 1993.

A. C. Fabian, E. M. Hu, L. L. Cowie, and J. Grindlay. The distribution and morphology of X-ray-emitting gas in the core of the Perseus cluster. *Astrophysical Journal*, 248: 47–54, 1981.

A. C. Fabian, J. S. Sanders, S. Ettori, G. B. Taylor, S. W. Allen, C. S. Crawford, K. Iwasawa, R. M. Johnstone, and P. M. Ogle. Chandra imaging of the complex x-ray core of the perseus cluster. *Monthly Notices of the Royal Astronomical Society*, 318(4): L65–L68, 2000.

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

C. Giacobino, S. Sardy, J. Diaz-Rodriguez, and N. Hengartner. Quantile universal threshold. *Electronic Journal of Statistics*, 11:4701–4722, 2017.

J. J. Goeman, H. C. van Houwelingen, and L. Finos. Testing against a high-dimensional alternative in the generalized linear model: asymptotic type i error control. *Biometrika*, 98(2):381–390, 2011.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman amd Hall, London; New York, 1994.

B. Guo and S. X. Chen. Tests for high dimensional generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1079–1102, 2016.

Julia Huettmann, Falkand Linke. Assessment of different link functions for modeling binary data to derive sound inferences and predictions. In Marina L.and Tan Chih Jeng Kennethand L'Ecuyer Pierre Kumar, Vipinand Gavrilova, editor, *Computational Science and Its Applications — ICCSA 2003*, pages 43–48, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, California, 1961. University of California Press.

F. Jansen, D. Lumb, B. Altieri, J. Clavel, M. Ehle, C. Erd, C. Gabriel, M. Guainazzi, P. Gondoin, R. Much, R. Munoz, M. Santos, N. Schartel, D. Texier, and G. Vacanti. XMM-Newton observatory. I. The spacecraft and operations. *Astronomy and Astrophysics*, 365:L1–L6, 2001.

A. V. Kravtsov and S. Borgani. Formation of Galaxy Clusters. *Annual Review of Astronomy and Astrophysics*, 50:353–409, 2012.

G. A. Kriss, D. F. Cioffi, and C. R. Canizares. The X-ray emitting gas in poor clusters with central dominant galaxies. *Astrophysical Journal*, 272:439–448, 1983.

N. Kushmerick. Learning to remove internet advertisements. In *Proceedings of the third international conference on Autonomous Agents*, pages 175–181. ACM, 1999.

C. Leng, Y. Lin, and G. Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284, 2006.

L. B. Lucy. An iterative technique for the rectification of observed distributions. *Astronomical Journal*, 79:745–754, 1974.

L. B. Lucy. Optimum strategies for inverse problems in statistical astronomy. *Astronomy and Astrophysics*, 289:983–994, 1994.

M. Markevitch, A. H. Gonzalez, L. David, A. Vikhlinin, S. Murray, W. Forman, C. Jones, and W. Tucker. A Textbook Example of a Bow Shock in the Merging Galaxy Cluster 1E 0657-56. *The Astrophysical Journal Letters*, 567:L27–L31, mar 2002.

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

D. E. McLaughlin. The Efficiency of Globular Cluster Formation. *Astronomical Journal*, 117:2398–2427, 1999.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

A. Morandi, S. Ettori, and L. Moscardini. X-ray and Sunyaev-Zel'dovich scaling relations in galaxy clusters. *Monthly notices of the royal astronomical society*, 379: 518–534, 2007.

A. Moretti, S. Campana, D. Lazzati, and G. Tagliaferri. The Resolved Fraction of the Cosmic X-Ray Background. *The Astrophysical Journal*, 588:696–703, may 2003.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A*, 135(3):370–384, 1972.

D. Neto, S. Sardy, and P. Tseng. ℓ_1-penalized likelihood smoothing and segmentation of volatility processes allowing for abrupt changes. *Journal of Computational and Graphical Statistics*, 21(1):217–233, 2012.

P. E. J. Nulsen and H. Bohringer. A ROSAT determination of the mass of the central Virgo Cluster. *Monthly notices of the royal astronomical society*, 274:1093–1106, 1995.

M. Y. Park and T. Hastie. L_1-regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B*, 69(4):659–677, 2007.

F. Pizzolato, S. Molendi, S. Ghizzardi, and S. De Grandi. Smaug: A New Technique for the Deprojection of Galaxy Clusters. *Astrophysical Journal*, 592:62–78, 2003.

C. L. Sarazin. *X-ray emission from clusters of galaxies*. Cambridge Astrophysics Series, Cambridge: Cambridge University Press, 1988.

S. Sardy. On the practice of rescaling covariates. *International Statistical Review*, 76 (2):285–297, 2008.

S. Sardy. Adaptive posterior mode estimation of a sparse sequence for model selection. *Scandinavian Journal of Statistics*, 36(4):577–601, 2009.

S. Sardy. Smooth blockwise iterative thresholding: a smooth fixed point estimator based on the likelihood's block gradient. *Journal of the American Statistical Association*, 107(498):800–813, 2012.

S. Sardy and P. Tseng. Density estimation by total variation penalized likelihood driven by the sparsity ℓ_1 information criterion. *Scandinavian Journal of Statistics*, 37(2):321–337, 2010.

S. Sardy, A. Antoniadis, and P. Tseng. Automatic smoothing with wavelets for a wide class of distributions. *Journal of Computational and Graphical Statistics*, 13(2): 399–421, 2004.

S. Sardy, C. Giacobino, and J. Diaz-Rodriguez. Thresholding tests. *arXiv:1708.02908v2*, 2018.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

W. Sharpe. A simplified model for portfolio analysis. 9:277–293, 01 1963.

Do. Spiegelman and E. Hertzmark. Easy sas calculations for risk or prevalence ratios and differences. 162:199–200, 09 2005.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

P. Sur, Y. Chen, and E. J. Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. arXiv:1706.01191v1, 2017.

C. Tchernin, D. Eckert, S. Ettori, E. Pointecouteau, S. Paltani, S. Molendi, G. Hurier, F. Gastaldello, E. T. Lau, D. Nagai, M. Roncarelli, and M. Rossetti. The XMM Cluster Outskirts Project (X-COP): Physical conditions to the virial radius of Abell 2142. *Astronomy and Astrophysics in press*, 2016.

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

A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 4(4):1035–1038, 1963.

M. J. L. Turner, A. Abbey, M. Arnaud, M. Balasini, M. Barbera, E. Belsole, P. J. Bennie, J. P. Bernard, G. F. Bignami, M. Boer, U. Briel, I. Butler, C. Cara, C. Chabaud, R. Cole, A. Collura, M. Conte, A. Cros, M. Denby, P. Dhez, G. Di Coco, J. Dowson, P. Ferrando, S. Ghizzardi, F. Gianotti, C. V. Goodall, L. Gretton, R. G. Griffiths, O. Hainaut, J. F. Hochedez, A. D. Holland, E. Jourdain, E. Kendziorra, A. Lagostina, R. Laine, N. La Palombara, M. Lortholary, D. Lumb, P. Marty, S. Molendi, C. Pigot, E. Poindron, K. A. Pounds, J. N. Reeves, C. Reppin, R. Rothenflug, P. Salvetat, J. L. Sauvageot, D. Schmitt, S. Sembay, A. D. T. Short, J. Spragg, J. Stephen, L. Strüder, A. Tiengo, M. Trifoglio, J. Trümper, S. Vercellone, L. Vigroux, G. Villa, M. J. Ward, S. Whitehead, and E. Zonca. The European Photon Imaging Camera on XMM-Newton: The MOS cameras. *Astronomy and Astrophysics*, 365:L27–L35, 2001.

A. Vikhlinin, M. Markevitch, and S. Murray. A moving cold front in the intergalactic medium of a3667. 551:160–171, 04 2001.

H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the LAD-lasso. *Journal of Business & Economic Statistics*, 25(3): 347–355, 2007.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Statistics*, 9:60–62, 1938.

S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman amd Hall/CRC, London; New York, 2017.

Y. Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2006.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.