Learning Heteroscedastic Models by Convex Programming under Group Sparsity

Arnak S. Dalalyan ARNAK.DALALYAN@ENSAE.FR
ENSAE-CREST-GENES

Mohamed Hebiri MOHAMED.HEBIRI@UNIV-MLV.FR
LAMA, Université Paris Est

Katia Meziani MEZIANI@CEREMADE.DAUPHINE.FR
CEREMADE, Université Paris Dauphine

Joseph Salmon JOSEPH.SALMON@TELECOM-PARISTECH.FR
Institut Mines-Télécom ; Télécom ParisTech ; CNRS LTCI

Abstract

Popular sparse estimation methods based on ℓ_1-relaxation, such as the Lasso and the Dantzig selector, require the knowledge of the variance of the noise in order to properly tune the regularization parameter. This constitutes a major obstacle in applying these methods in several frameworks—such as time series, random fields, inverse problems—for which the noise is rarely homoscedastic and its level is hard to know in advance. In this paper, we propose a new approach to the joint estimation of the conditional mean and the conditional variance in a high-dimensional (auto-) regression setting. An attractive feature of the proposed estimator is that it is efficiently computable even for very large scale problems by solving a second-order cone program (SOCP). We present theoretical analysis and numerical results assessing the performance of the proposed procedure.

1. Introduction

Over the last fifteen years, sparse estimation methods based on ℓ_1-relaxation, among which the Lasso (Tibshirani, 1996) and the Dantzig selector (Candès and Tao, 2007) are the most famous examples, have become a popular tool for estimating high dimensional linear models. So far, their wider use in several fields of applications (*e.g.*, finance and econometrics) has been constrained by the difficulty of adapting to heteroscedasticity, *i.e.*, when the noise level varies across the components of the signal.

Let $\mathcal{T}$ be a finite set of cardinality T. For every $t \in \mathcal{T}$ we observe a sequence $(\boldsymbol{x}_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$ obeying:

$$y_t = \mathsf{b}^*(\boldsymbol{x}_t) + \mathsf{s}^*(\boldsymbol{x}_t)\xi_t, \tag{1}$$

where $\mathsf{b}^* : \mathbb{R}^d \to \mathbb{R}$ and $\mathsf{s}^{*2} : \mathbb{R}^d \to \mathbb{R}_+$ are respectively the unknown conditional mean and conditional variance[1] of y_t given $\boldsymbol{x}_t$. Then, the errors ξ_t satisfy $\mathbf{E}[\xi_t|\boldsymbol{x}_t] = 0$ and $\mathbf{Var}[\xi_t|\boldsymbol{x}_t] = 1$. Depending on the targeted applications, elements of $\mathcal{T}$ may be time instances (financial engineering), pixels or voxels (image and video processing) or spatial coordinates (astronomy, communication networks).

In this general formulation, the problem of estimating unknown functions b^* and s^* is ill-posed: the dimensionality of unknowns is too large as compared to the number of equations T, therefore, the model is unidentifiable. To cope with this issue, the parameters $(\mathsf{b}^*, \mathsf{s}^*)$ are often constrained to belong to low dimensional spaces. For instance, a common assumption is that for some given dictionary $\mathsf{f}_1, \ldots, \mathsf{f}_p$ of functions from $\mathbb{R}^d$ to $\mathbb{R}$ and for an unknown vector $(\boldsymbol{\beta}^*, \sigma^*) \in \mathbb{R}^p \times \mathbb{R}$, the relations $\mathsf{b}^*(\boldsymbol{x}) = [\mathsf{f}_1(\boldsymbol{x}), \ldots, \mathsf{f}_p(\boldsymbol{x})]\boldsymbol{\beta}^*$ and $\mathsf{s}^*(\boldsymbol{x}) \equiv \sigma^*$ hold for every $\boldsymbol{x}$. Even for very large values of p, much larger than the sample size T, such a model can be efficiently learned in the sparsity scenario using recently introduced scaled versions

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

[1] This formulation of the problem includes "time-dependent" mean and variance, *i.e.*, the case of $\mathbf{E}[y_t|\boldsymbol{x}_t] = \mathsf{b}_t^*(\boldsymbol{x}_t)$ and $\mathbf{Var}[y_t|\boldsymbol{x}_t] = \mathsf{s}_t^*(\boldsymbol{x}_t)$, since it is sufficient then to consider as explanatory variable $[t; \boldsymbol{x}_t^\top]^\top$ instead of $\boldsymbol{x}_t$.

of ℓ_1-relaxations: the square-root Lasso (Antoniadis, 2010; Belloni et al., 2011; Sun and Zhang, 2012; Gautier and Tsybakov, 2011), the scaled Lasso (Städler et al., 2010) and the scaled Dantzig selector (Dalalyan and Chen, 2012). These methods are tailored to the context of a fixed noise level across observations (homoscedasticity), which reduces their attractiveness for applications in the aforementioned fields. In the present work, we propose a new method of estimation for model (1) that has the appealing properties of requiring neither homoscedasticity nor any prior knowledge of the noise level. The only restriction we impose is that the variance function s^{*2} is of reduced dimensionality, which in our terms means that its inverse $1/\mathsf{s}^*$ is of a linear parametric form.

Our contributions and related work We propose a principled approach to the problem of joint estimation of the conditional mean function b^* and the conditional variance s^{*2}, which boils down to a second-order cone programming (SOCP) problem. We refer to our procedure as the Scaled Heteroscedastic Dantzig selector (ScHeDs) since it can be seen an extension of the Dantzig selector to the case of heteroscedastic noise and group sparsity. Note that so far, inference under group-sparsity pioneered by (Yuan and Lin, 2006; Lin and Zhang, 2006), has only focused on the simple case of known and constant noise level both in the early references (Nardi and Rinaldo, 2008; Bach, 2008; Chesneau and Hebiri, 2008; Meier et al., 2009), and in the more recent ones (Lounici et al., 2011; Huang et al., 2012). In this work we provide a theoretical analysis and some numerical experiments assessing the quality of the proposed ScHeDs procedure.

More recently, regression estimation under the combination of sparsity and heteroscedasticity was addressed by (Daye et al., 2012; Wagener and Dette, 2012; Kolar and Sharpnack, 2012). Because of the inherent nonconvexity of the penalized (pseudo-)log-likelihood considered in these works, the methods proposed therein do not estimate the conditional mean and the variance in a joint manner. They rather rely on iterative estimation of those quantities: they alternate between the two variables, estimating one while keeping the other one fixed. Furthermore, the theoretical results of these papers are asymptotic. In contrast, we propose a method that estimates the conditional mean and the variance by solving a jointly convex minimization problem and derive nonasymptotic risk bounds for the proposed estimators.

Notation We use boldface letters to denote vectors and matrices. For an integer $d > 0$, we set $[d] = \{1, \dots, d\}$. If $\boldsymbol{v} \in \mathbb{R}^d$ and $J \subset [d]$, then $\boldsymbol{v}_J$ denotes the sub-vector of $\boldsymbol{v}$ obtained by removing all the coordinates having indexes outside J. If $J = \{j\}$, we write $\boldsymbol{v}_J = v_j$. The ℓ_q-norms of $\boldsymbol{v}$ are defined by:

$$|\boldsymbol{v}|_0 = \sum_{j=1}^d \mathbf{1}(v_j \neq 0), \quad |\boldsymbol{v}|_\infty = \max_{j\in\{1,\dots,d\}} |v_j|,$$
$$|\boldsymbol{v}|_q^q = \sum_{j=1}^d |v_j|^q, \ 1 \leq q < \infty.$$

For a matrix $\mathbf{A}$, $\boldsymbol{A}_{i,:}$ and $\boldsymbol{A}_{:,j}$ stand respectively for its i-th row and its j-th column. For a vector $\boldsymbol{Y} = [y_1, \dots, y_T]^\top \in \mathbb{R}^T$, we define $\mathrm{diag}(\boldsymbol{Y})$ as the $T \times T$ diagonal matrix having the entries of $\boldsymbol{Y}$ on its main diagonal.

2. Background and assumptions

We start by reparameterizing the problem as follows:

$$\mathsf{r}^*(\boldsymbol{x}) = 1/\mathsf{s}^*(\boldsymbol{x}), \qquad \mathsf{f}^*(\boldsymbol{x}) = \mathsf{b}^*(\boldsymbol{x})/\mathsf{s}^*(\boldsymbol{x}). \tag{2}$$

Clearly, under the condition that s^* is bounded away from zero, the mapping $(\mathsf{s}^*, \mathsf{b}^*) \mapsto (\mathsf{r}^*, \mathsf{f}^*)$ is bijective. As shown later, learning the pair $(\mathsf{r}^*, \mathsf{f}^*)$ appears to be more convenient than learning the original mean-variance pair, in the sense that it can be performed by solving a convex problem.

We now introduce two assumptions underlying our approach. The first one is a group sparsity assumption on the underlying function f^*. It states that there exists a given dictionary of functions $\mathsf{f}_1, \dots, \mathsf{f}_p$ from $\mathbb{R}^d$ to $\mathbb{R}$ such that f^* is well approximated by a linear combination $\sum_{j=1}^p \phi_j^* \mathsf{f}_j$ with a (fixed) group-sparse vector $\boldsymbol{\phi}^* = [\phi_1^*, \dots, \phi_p^*]^\top$. The precise formulation is:

Assumption (A1) We denote by $\mathbf{X}$ the $T \times p$ matrix having $[\mathsf{f}_1(\boldsymbol{x}_t), \dots, \mathsf{f}_p(\boldsymbol{x}_t)]$ as t-th row. Then, for a given partition $G_1, \dots, G_K$ of $\{1, \dots, p\}$, there is a vector $\boldsymbol{\phi}^* \in \mathbb{R}^p$ such that $[\mathsf{f}^*(\boldsymbol{x}_1), \dots, \mathsf{f}^*(\boldsymbol{x}_T)]^\top \approx \mathbf{X}\boldsymbol{\phi}^*$ and $\mathrm{Card}(\{k : |\boldsymbol{\phi}^*_{G_k}|_2 \neq 0\}) \ll K$.

Assumption (A1) is a restriction on f^* only; the function r^* does not appear in its formulation. Let us describe two practical situations which fit into the framework delineated by Assumption (A1), some other examples can be found in (Lounici et al., 2011; Mairal et al., 2011; Huang et al., 2012).

Sparse linear model with qualitative covariates Consider the case of linear regression with a large number of covariates, an important portion of which are qualitative. Each qualitative covariate having m modalities is then transformed into a group of m binary quantitative covariates. Therefore, the irrelevance of one qualitative covariate implies the irrele-

vance of a group of quantitative covariates, leading to the group-sparsity condition.

Sparse additive model (Ravikumar et al., 2009; Koltchinskii and Yuan, 2010; Raskutti et al., 2012) If f^* is a nonlinear function of a moderately large number of quantitative covariates, then—to alleviate the curse of dimensionality—a sparse additive model is often considered for fitting the response. This means that f^* is assumed to be of the simple form $\mathsf{f}_1^*(x_1)+\ldots+\mathsf{f}_d^*(x_d)$, with most functions f_j^* being identically equal to zero. Projecting each of these functions onto a fixed number of elements of a basis, $\mathsf{f}_j^*(x) \approx \sum_{\ell=1}^{K_j} \phi_{\ell,j}\psi_\ell(x)$, we get a linear formulation in terms of the unknown vector $\boldsymbol{\phi} = (\phi_{\ell,j})$. The sparsity of the additive representation implies the group-sparsity of the vector $\boldsymbol{\phi}$.

Our second assumption requires that there is a linear space of dimension q, much smaller than the sample size T, that contains the function r^*. More precisely:

Assumption (A2) For q given functions $\mathsf{r}_1, \ldots, \mathsf{r}_q$ mapping $\mathbb{R}^d$ into $\mathbb{R}_+$, there is a vector $\boldsymbol{\alpha} \in \mathbb{R}^q$ such that $\mathsf{r}^*(\boldsymbol{x}) = \sum_{\ell=1}^q \alpha_\ell \mathsf{r}_\ell(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{R}^d$.

Here are two examples of functions r^* satisfying this assumption.

Blockwise homoscedastic noise In time series modeling, one can assume that the variance of the innovations varies smoothly over time, and, therefore, can be well approximated by a piecewise constant function. This situation also arises in image processing where neighboring pixels are often corrupted by noise of similar magnitude. This corresponds to choosing a partition of $\mathcal{T}$ into q cells and to defining each r_ℓ as the indicator function of one cell of the partition.

Periodic noise-level In meteorology or image processing, observations may be contaminated by a periodic noise. In meteorology, this can be caused by seasonal variations, whereas in image processing, this may occur if the imaging system is subject to electronic disturbance of repeating nature. Periodic noise can be handled by (A2) stating that r^* belongs to the linear span of a few trigonometric functions.

There are essentially three methods in the literature providing estimators of $(\mathsf{b}^*, \mathsf{s}^*)$ in a context close to the one described above. All of them assume that s^* is constant and equal to σ^* and $[\mathsf{b}^*(\boldsymbol{x}_1), \ldots, \mathsf{b}^*(\boldsymbol{x}_T)]^\top = \mathbf{X}\boldsymbol{\beta}^*$ with some sparse vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$. The first method, termed the scaled Lasso (Städler et al., 2010), suggests to recover $(\boldsymbol{\beta}^*, \sigma^*)$ by computing a solution $(\widehat{\boldsymbol{\beta}}^{\text{Sc-L}}, \widehat{\sigma}^{\text{Sc-L}})$ to the optimization problem

$$\min_{\boldsymbol{\beta},\sigma} \Big\{ T\log(\sigma) + \frac{|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}|_2^2}{2\sigma^2} + \frac{\lambda}{\sigma}\sum_{j=1}^p |\boldsymbol{X}_{:,j}|_2 |\beta_j| \Big\}, \quad (3)$$

where $\lambda > 0$ is a scale-free tuning parameter controlling the trade-off between data fitting and sparsity level. After a change of variables, this can be cast as a convex program. Hence, it is possible to find the global minimum relatively efficiently even for large p.

A second method for joint estimation of $\boldsymbol{\beta}^*$ and σ^* by convex programming, the Square-Root Lasso (Antoniadis, 2010; Belloni et al., 2011), estimates $\boldsymbol{\beta}^*$ by $\widehat{\boldsymbol{\beta}}^{\text{SqR-L}}$ which solves

$$\min_{\boldsymbol{\beta}} \Big\{ \big|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\big|_2 + \lambda \sum_{j=1}^p |\boldsymbol{X}_{:,j}|_2 |\beta_j| \Big\} \quad (4)$$

and then defines $\widehat{\sigma}^{\text{SqR-L}} = \frac{1}{\sqrt{T}} \big|\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{SqR-L}}\big|_2$ as an estimator of σ^*. Both in theory and in practice, these two methods perform quite similarly (Sun and Zhang, 2012).

A third method, termed scaled Dantzig selector, was studied by (Dalalyan and Chen, 2012) under a more general type of sparsity assumption (called fused or indirect sparsity). Inspired by these works, we propose a new procedure for joint estimation of the conditional mean and the conditional variance in the context of heteroscedasticity and group-sparsity.

3. Definition of the procedure

Our methodology originates from the penalized log-likelihood minimization. Assuming errors ξ_t are i.i.d. Gaussian $\mathcal{N}(0,1)$ and setting $\mathsf{f}(\boldsymbol{x}) = \sum_{j=1}^p \phi_j \mathsf{f}_j(\boldsymbol{x})$, the penalized log-likelihood used for defining the group-Lasso estimator is (up to summands independent of (f, r)):

$$\mathrm{PL}(\mathsf{f},\mathsf{r}) = \sum_{t\in\mathcal{T}} \Big\{ -\log(\mathsf{r}(\boldsymbol{x}_t)) + \frac{1}{2}\big(\mathsf{r}(\boldsymbol{x}_t)y_t - \boldsymbol{X}_{t,:}\boldsymbol{\phi}\big)^2 \Big\} + \sum_{k=1}^K \lambda_k \Big| \sum\nolimits_{j\in G_k} \boldsymbol{X}_{:,j}\phi_j \Big|_2, \quad (5)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K) \in \mathbb{R}_+^K$ is a tuning parameter. A first strategy for estimating $(\mathsf{f}^*, \mathsf{r}^*)$ is to minimize $\mathrm{PL}(\mathsf{f},\mathsf{r})$ with respect to $\boldsymbol{\phi} \in \mathbb{R}^p$ and $\mathsf{r} \in \{\mathsf{g} : \mathbb{R}^d \to \mathbb{R} : \mathsf{g}(\boldsymbol{x}) \geq 0, \text{ for almost all } \boldsymbol{x} \in \mathbb{R}^d\}$. In view of assumption (A2), we can replace r by $\sum_{\ell=1}^q \alpha_\ell \mathsf{r}_\ell$ with an unknown q-vector $\boldsymbol{\alpha}$.

If we introduce the $T \times q$ matrix $\mathbf{R}$ having as generic entry $\mathsf{r}_\ell(\boldsymbol{x}_t)$, (5) translates into a convex program with

respect to the $p+q$ dimensional parameter $(\boldsymbol{\phi}, \boldsymbol{\alpha}) \in \mathbb{R}^p \times \mathbb{R}^q$, in which the cost function is:

$$\mathrm{PL}(\boldsymbol{\phi}, \boldsymbol{\alpha}) = \sum_{t \in \mathcal{T}} \Big\{ -\log(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}) + \frac{1}{2}\big(y_t \boldsymbol{R}_{t,:}\boldsymbol{\alpha} - \boldsymbol{X}_{t,:}\boldsymbol{\phi}\big)^2 \Big\} + \sum_{k=1}^{K} \lambda_k \big| \boldsymbol{X}_{:,G_k} \boldsymbol{\phi}_{G_k} \big|_2, \quad (6)$$

and the constraint $\min_t \boldsymbol{R}_{t,:}\boldsymbol{\alpha} \geq 0$ should be imposed to guarantee that the logarithm is well defined. This is a convex optimization problem, but it does not fit well the framework under which the convergence guarantees of the state-of-the-art optimization algorithms are established. Indeed, it is usually required that the smooth components of the cost function have Lipschitz-smooth derivative, which is not the case for (6) because of the presence of the logarithmic terms. One can circumvent this drawback by smoothing these terms[2], but we opted for another solution that relies on an argument introduced in (Candès and Tao, 2007) for justifying the Dantzig selector. Let $\boldsymbol{\Pi}_{G_k} = \mathbf{X}_{:,G_k}(\mathbf{X}_{:,G_k}^\top \mathbf{X}_{:,G_k})^+ \mathbf{X}_{:,G_k}^\top$ be the orthogonal projector onto the range of $\mathbf{X}_{:,G_k}$ in $\mathbb{R}^{\mathcal{T}}$.

Definition 3.1. *Let $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ be a vector of tuning parameters. We call the Scaled Heteroscedastic Dantzig selector (ScHeDs) the pair $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}})$, where $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{v}})$ is a minimizer w.r.t. $(\boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{v}) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}_+^T$ of the cost function*

$$\sum\nolimits_{k=1}^{K} \lambda_k \big| \mathbf{X}_{:,G_k} \boldsymbol{\phi}_{G_k} \big|_2$$

subject to the constraints

$$\Big| \boldsymbol{\Pi}_{G_k} \big(\mathrm{diag}(\boldsymbol{Y}) \mathbf{R} \boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\phi} \big) \Big|_2 \leq \lambda_k, \ \forall k \in [K]; \quad (7)$$

$$\mathbf{R}^\top \boldsymbol{v} \leq \mathbf{R}^\top \mathrm{diag}(\boldsymbol{Y})(\mathrm{diag}(\boldsymbol{Y})\mathbf{R}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\phi}); \quad (8)$$

$$1/v_t \leq \boldsymbol{R}_{t,:}\boldsymbol{\alpha}; \ \forall t \in \mathcal{T}. \quad (9)$$

Constraints (7)-(9) are obtained as convex relaxations of the first-order conditions corresponding to minimizing (6). In fact, Eq. (7) is a standard relaxation for the condition $\mathbf{0} \in \partial_{\boldsymbol{\phi}} \mathrm{PL}(\boldsymbol{\phi}, \boldsymbol{\alpha})$, whereas constraints (8) and (9) are convex relaxations of the equation $\partial_{\boldsymbol{\alpha}} \mathrm{PL}(\boldsymbol{\phi}, \boldsymbol{\alpha}) = \mathbf{0}$. Further details on this point are provided in the supplementary material. At this stage and before presenting theoretical guarantees on the statistical performance of the ScHeDs, we state a result telling us the estimator we introduced is meaningful.

Theorem 3.2. *The ScHeDs is always well defined in the sense that the feasible set of the corresponding optimization problem is not empty: it contains the minimizer of (6). Furthermore, the ScHeDs can be computed by any SOCP solver.*

[2] This will result in introducing new parameters the tuning of which may increase the difficulty of the problem.

The proof is placed in the supplementary material. As we will see later, thanks to this theorem, we carried out two implementations of the ScHeDs based on an interior point algorithm and an optimal first-order proximal method.

4. Comments on the procedure

Tuning parameters One apparent drawback of the ScHeDs is the large number of tuning parameters. Fortunately, some theoretical results provided in the supplementary material suggest to choose $\lambda_k = \lambda_0 \sqrt{r_k}$, where $\lambda_0 > 0$ is a one-dimensional tuning parameter and $r_k = \mathrm{rank}(\mathbf{X}_{:,G_k})$. In particular, when all the predictors within each group are linearly independent, then one may choose $\boldsymbol{\lambda}$ proportional to the vector $(\sqrt{\mathrm{Card}(G_1)}, \ldots, \sqrt{\mathrm{Card}(G_K)})$.

Additional constraints In many practical situations one can add some additional constraints to the aforementioned optimization problem without leaving the SOCP framework. For example, if the response y is bounded by some known constant L_y, then it is natural to look for conditional mean and conditional variance bounded respectively by L_y and L_y^2. This amounts to introducing the (linearizable) constraints $|\boldsymbol{X}_{t,:}\boldsymbol{\phi}| \leq L_y \boldsymbol{R}_{t,:}\boldsymbol{\alpha}$ and $\boldsymbol{R}_{t,:}\boldsymbol{\alpha} \geq 1/L_y$ for every $t \in \mathcal{T}$.

Bias correction It is well known that the Lasso and the Dantzig selector estimate the nonzero coefficients of the regression vector with a bias toward zero. It was also remarked in (Sun and Zhang, 2010), that the estimator of the noise level provided by the scaled Lasso is systematically over-estimating the true noise level. Our experiments showed the same shortcomings for the ScHeDs. To attenuate these effects, we propose a two-step procedure that applies the ScHeDs with the penalties $\lambda_k = \lambda_0 \sqrt{r_k}$ at the first step and discards from $\mathbf{X}$ the columns that correspond to vanishing coefficients of $\widehat{\boldsymbol{\phi}}$. At the second step, the ScHeDs is applied with the new matrix $\mathbf{X}$ and with $\boldsymbol{\lambda} = 0$.

Gaussian assumption Although the proposed algorithm takes its roots from the log-likelihood of the Gaussian regression, it is by no means necessary that the noise distribution should be Gaussian. In the case of deterministic design $\boldsymbol{x}_t$, it is sufficient to assume that the noise distribution is sub-Gaussian. For random i.i.d. design, arguments similar to those of (Belloni et al., 2011; Gautier and Tsybakov, 2011) can be applied to show oracle inequalities for even more general noise distributions.

Equivariance Given the historical data $(y_{1:T}, \boldsymbol{x}_{1:T})$ of the response and the covariates, let us denote by $\widehat{y}_{T+1}(y_{1:T}) = \big[\sum_{\ell=1}^{p} \widehat{\phi}_j \, \mathsf{f}_j(\boldsymbol{x}_{T+1})\big] / \big[\sum_{\ell=1}^{q} \widehat{\alpha}_\ell \, \mathsf{r}_\ell(\boldsymbol{x}_{T+1})\big]$ the prediction provided by the ScHeDs for a new observation $\boldsymbol{x}_{T+1}$. This prediction is equivariant with respect to scale change in the following sense. If all the response values $y_1, \ldots, y_T$ are multiplied by some constant c, then it can easily be proved that the new prediction can be deduced from the previous one by merely multiplying it by c: $\widehat{y}_{T+1}(cy_{1:T}) = c\widehat{y}_{T+1}(y_{1:T})$.

Most papers dealing with group-sparsity (Lounici et al., 2011; Liu et al., 2010; Huang and Zhang, 2010) use penalties of the form $\sum_k |\mathbf{D}_k \boldsymbol{\phi}_{G_k}|_2$ with some diagonal matrices $\mathbf{D}_k$. In general, this differs from the penalty we use since in our case $\mathbf{D}_k = (\mathbf{X}_{:,G_k}^\top \mathbf{X}_{:,G_k})^{1/2}$ is not necessarily diagonal. Our choice has the advantage of being equivariant w.r.t. (invertible) linear transformations of predictors within groups.

Interestingly, this difference in the penalty definition has an impact on the calibration of the parameters λ_k: while the recommended choice is $\lambda_k^2 \propto \mathrm{Card}(G_k)$ when diagonal matrices[3] $\mathbf{D}_k$ are used, it is $\lambda_k^2 \propto \mathrm{rank}(\mathbf{X}_{:,G_k})$ for the ScHeDs. Thus, the penalty chosen for the ScHeDs is slightly smaller than that of the usual group-Lasso, which also leads to a tighter risk bound.

5. Risk bounds

We present a finite sample risk bound showing that, under some assumptions, the risk of our procedure is of the same order of magnitude as the risk of a procedure based on the complete knowledge of the noise-level.

Recall that the model introduced in the foregoing sections can be rewritten in its matrix form

$$\mathrm{diag}(\boldsymbol{Y})\mathbf{R}\boldsymbol{\alpha}^* = \mathbf{X}\boldsymbol{\phi}^* + \boldsymbol{\xi}, \tag{10}$$

with $\xi_1, \ldots, \xi_T$ i.i.d. zero mean random variables. To state the theoretical results providing guarantees on the accuracy of the ScHeDs estimator $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}})$, we need some notation and assumptions.

For $\boldsymbol{\phi}^* \in \mathbb{R}^p$, we define the set of relevant groups $\mathcal{K}^*$ and the sparsity index s^* by $\mathcal{K}^* = \big\{k : \big|\boldsymbol{\phi}^*_{G_k}\big|_1 \neq 0\big\}$, $s^* = \sum_{k \in \mathcal{K}^*} r_k$, Note that these quantities depend on $\boldsymbol{\phi}^*$. To establish tight risk bounds, we need the following assumption on the Gram matrix $\mathbf{X}^\top\mathbf{X}$, termed Group-Restricted Eigenvalues (GRE).

Assumption GRE(N, κ)**:** For every $\mathcal{K} \subset [p]$ of cardinality not larger than N and for every $\boldsymbol{\delta} \in \mathbb{R}^p$ satisfying

$$\sum\nolimits_{\mathcal{K}^c} \lambda_k \big|\mathbf{X}_{:,G_k}\boldsymbol{\delta}_{G_k}\big|_2 \leq \sum\nolimits_{\mathcal{K}} \lambda_k \big|\mathbf{X}_{:,G_k}\boldsymbol{\delta}_{G_k}\big|_2, \tag{11}$$

it holds that $\big|\mathbf{X}\boldsymbol{\delta}\big|_2^2 \geq \kappa^2 \sum_{k\in\mathcal{K}} \big|\mathbf{X}_{:,G_k}\boldsymbol{\delta}_{G_k}\big|_2^2$.

We also set

$$C_1 = \max_{\ell=1,\ldots,q} \frac{1}{T} \sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}^2 (\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*)^2}{(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2}\,, \tag{12}$$

$$C_2 = \max_{\ell=1,\ldots,q} \frac{1}{T} \sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}^2}{(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2}\,, \tag{13}$$

$$C_3 = \min_{\ell=1,\ldots,q} \frac{1}{T} \sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)}, \tag{14}$$

and define $C_4 = (\sqrt{C_2} + \sqrt{2C_1})/C_3$.

To establish nonasymptotic risk bounds in the heteroscedastic regression model with sparsity assumption, we first tried to adapt the standard techniques (Candès and Tao, 2007; Bickel et al., 2009) used in the case of known noise-level. The result, presented in Theorem 5.1 below, is not satisfactory, since it provides a risk bound for estimating $\boldsymbol{\phi}^*$ that involves the risk of estimating $\boldsymbol{\alpha}^*$. Nevertheless, we opted for stating this result since it provides guidance for choosing the parameters λ_k and also because it constitutes an important ingredient of the proof of our main result stated in Theorem 5.2 below.

Theorem 5.1. *Consider model (10) with deterministic matrices* $\mathbf{X}$ *and* $\mathbf{R}$*. Assume that the distribution of* $\boldsymbol{\xi}$ *is Gaussian with zero mean and an identity covariance matrix and that Assumption GRE*(K^*, κ) *is fulfilled with* $K^* = \mathrm{Card}(\mathcal{K}^*)$*. Let* $\varepsilon \in (0,1)$ *be a tolerance level and set*

$$\lambda_k = 2\big(r_k + 2\sqrt{r_k \log(K/\varepsilon)} + 2\log(K/\varepsilon)\big)^{1/2}.$$

If $T \geq 8C_4 \log(\frac{2q}{\varepsilon})$ *then, with probability at least* $1-2\varepsilon$*,*

$$\begin{aligned}\big|\mathbf{X}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)\big|_2 \leq\; & C_4\sqrt{(8/T)\log(2q/\varepsilon)}(2|\mathbf{X}\boldsymbol{\phi}^*|_2 + |\boldsymbol{\xi}|_2) \\ & + \frac{8}{\kappa}\sqrt{2s^* + 3K^*\log(K/\varepsilon)} \\ & + |\mathrm{diag}(\boldsymbol{Y})\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)|_2. \end{aligned} \tag{15}$$

In order to gain understanding on the theoretical limits delineated by the previous theorem, let us give more details on the order of magnitude of the three terms appearing in (15). First, one should keep in mind that the correct normalization of the error consists in dividing $\big|\mathbf{X}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)\big|_2$ by $\sqrt{T}$. Assuming that the function b^* is bounded and using standard tail bounds on the χ^2_T distribution, we can see that the first term in the

[3] Even if the matrices $\mathbf{D}_k$ are not diagonal and are chosen exactly as in our case, recent references like (Simon and Tibshirani, 2012) suggest to use $\lambda_k^2 \propto \mathrm{Card}(G_k)$ without theoretical support.

	ScHeDs						Square-root Lasso					
	$\lvert\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\rvert_2$		$\lvert\widehat{s}-s^*\rvert$		$10\lvert\widehat{\sigma}-\sigma^*\rvert$		$\lvert\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\rvert_2$		$\lvert\widehat{s}-s^*\rvert$		$10\lvert\widehat{\sigma}-\sigma^*\rvert$	
$(T,\ p,\ s^*,\ \sigma^*)$	Ave	StD	Ave	StD	Ave	StD	Ave	StD	Ave	StD	Ave	StD
$(100,\ 100, 2, 1.0)$	.13	.07	.02	.14	.53	.40	.16	.11	.19	.44	.56	.42
$(100,\ 100, 5, 1.0)$	.28	.24	.08	.32	.76	.78	.25	.13	.19	.44	.66	.57
$(200,\ 100, 5, 0.5)$	.08	.03	.00	.00	.20	.16	.09	.03	.20	.46	.22	.16
$(200,\ 100, 5, 1.0)$	.15	.05	.01	.09	.40	.30	.17	.07	.20	.44	.42	.31
$(200,\ 500, 8, 0.5)$	.10	.03	.00	.04	.23	.16	.11	.03	.17	.40	.24	.17
$(200,\ 500, 8, 1.0)$	.21	.13	.02	.17	.50	.58	.22	.08	.19	.43	.46	.38
$(200, 1000, 5, 1.0)$	.15	.05	.01	.08	.40	.31	.17	.07	.17	.40	.42	.33

Table 1. Performance of the (bias corrected) ScHeDs compared with the (bias corrected) Square-root Lasso on a synthetic dataset. The average values and the standard deviations of the quantities $|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|_2$, $|\widehat{s}-s^*|$ and $10|\widehat{\sigma}-\sigma^*|$ over 500 trials are reported. They represent respectively the accuracy in estimating the regression vector, the number of relevant covariates and the level of noise.

right-hand side of (15) is negligible w.r.t. the second one. Thus if we ignore for a moment the third term, Theorem 5.1 tells us that the normalized squared error $T^{-1}\big|\mathbf{X}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*)\big|_2^2$ of estimating $\boldsymbol{\phi}^*$ by $\widehat{\boldsymbol{\phi}}$ is of the order of s^*/T, up to logarithmic terms. This is the (optimal) fast rate of estimating an s^*-sparse signal with T observations in linear regression.

To complete the theoretical analysis, we need a bound on the error of estimating the parameter $\boldsymbol{\alpha}^*$. This is done in the following theorem.

Theorem 5.2. *Let all the conditions of Theorem 5.1 be fulfilled. Let q and T be two integers such that $1 \le q \le T$ and let $\varepsilon \in (0, 1/5)$. Assume that for some constant $\widehat{D}_1 \ge 1$ the inequality $\max_{t\in\mathcal{T}} \frac{\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*} \le \widehat{D}_1$ holds true and denote $D_{T,\varepsilon} = \widehat{D}_1(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 5\log(2T/\varepsilon))$. Then, on an event of probability at least $1-5\varepsilon$, the following inequality is true:*

$$\big|\mathbf{X}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*)\big|_2 \le 4(C_4+1)D_{T,\varepsilon}^{3/2}\sqrt{2q\log(2q/\varepsilon)} + \frac{8D_{T,\varepsilon}}{\kappa}\sqrt{2s^*+3K^*\log(K/\varepsilon)}. \quad (16)$$

Furthermore, on the same event,

$$\frac{\big|\mathbf{R}(\boldsymbol{\alpha}^*-\widehat{\boldsymbol{\alpha}})\big|_2}{\widehat{D}_1^{1/2}|\mathbf{R}\boldsymbol{\alpha}^*|_\infty} \le 4(C_4+2)D_{T,\varepsilon}^{3/2}\sqrt{2q\log(2q/\varepsilon)} + \frac{8D_{T,\varepsilon}}{\kappa}\sqrt{2s^*+3K^*\log(K/\varepsilon)}. \quad (17)$$

The first important feature of this result is that it provides fast rates of convergence for the ScHeDs estimator. This compares favorably with the analogous result in (Kolar and Sharpnack, 2012), where asymptotic bounds are presented under the stringent condition that the local minimum to which the procedure converges coincides with the global one. The joint convexity in $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$ of our minimization problem allows us to avoid such an assumption without any loss in the quality of prediction.

One potential weakness of the risk bounds of Theorem 5.2 is the presence of the quantity $\widehat{D}_1$, which controls, roughly speaking, the ℓ_∞ norm of the vector $\mathbf{R}\widehat{\boldsymbol{\alpha}}$. One way to circumvent this drawback is to add the constraint $\max_t \boldsymbol{R}_{t,:}\boldsymbol{\alpha} \le \mu^*$ to those presented in (7)-(9), for some tuning parameter μ^*. In this case, the optimization problem remains an SOCP and in all the previous results one can replace the random term $\widehat{D}_1$ by μ^*/μ_*, where μ_* is a lower bound on the elements of the vector $\mathbf{R}\boldsymbol{\alpha}^*$. This being said, we hope that with more sophisticated arguments one can deduce the boundedness of $\widehat{D}_1$ by some deterministic constant without adding new constraints to the ScHeDs.

One may also wonder how restrictive the assumptions (12)-(14) are and in which kind of contexts they are expected to be satisfied. At a heuristic level, one may remark that the expressions in (12)-(14) are all empirical means: for instance, $(1/T)\sum_t r_{t\ell}^2/(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2 = (1/T)\sum_t \mathsf{r}_\ell(\boldsymbol{x}_t)^2/\mathsf{r}^*(\boldsymbol{x}_t)^2$. Assuming that the time series $\{\boldsymbol{x}_t\}$ is stationary or periodic, these empirical means will converge to some expectations. Therefore, under these types of assumptions, (12)-(14) are boundedness assumptions on some integral functionals of r^*, f^* and r_ℓ's. In particular, if r_ℓ's are bounded and bounded away from 0, f^* is bounded and r^* is bounded away from zero, then the finiteness of the constant C_4 is straightforward.

To close this section, let us emphasize that the GRE condition is sufficient for getting fast rates for the performance of the ScHeDs measured in prediction loss, but is by no means necessary for the consistency. In other terms, even if the GRE condition fails, the ScHeDs still provides provably accurate estimates that converge at a slower rate. This slow rate is, roughly speaking, of the order $[T^{-1}(s^*+K^*\log K)]^{1/4}$ instead of $[T^{-1}(s^*+K^*\log K)]^{1/2}$.

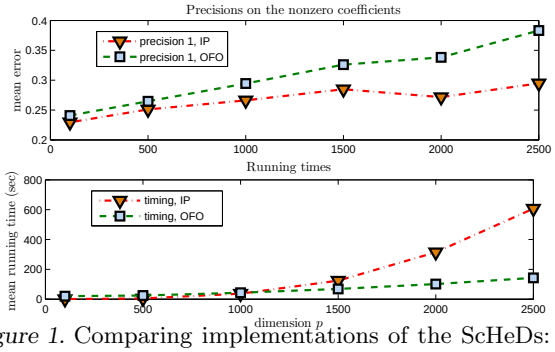

Figure 1. Comparing implementations of the ScHeDs: interior point (IP) *vs.* optimal first-order (OFO) method. We used the experiment described in Section 6.2 with $T = 200$, $s^* = 3$, $\sigma = 0.5$. Top: square-root of the MSE on the nonzero coefficients of $\boldsymbol{\beta}^*$. Bottom: running times.

6. Experiments

To assess the estimation accuracy of our method and to compare it with the state-of-the-art alternatives, we performed an experiment on a synthetic dataset. Then, the prediction ability of the procedure is evaluated on a real-world dataset containing the temperatures in Paris over several years.

6.1. Implementation

To effectively compute the ScHeDs estimator we rely on Theorem 3.2 that reduces the computation to solving a second-order cone program. To this end, we implemented a primal-dual interior point method using the SeDuMi package (Sturm, 1999) of Matlab as well as several optimal first-order methods (Nesterov, 1983; Auslender and Teboulle, 2006; Beck and Teboulle, 2009) using the TFOCS (Becker et al., 2011). We intend to make our code publicly available if the paper is accepted. Each of these implementations has its strengths and limitations. The interior point method provides a highly accurate solution for moderately large datasets (Fig. 1, top), but this accuracy is achieved at the expense of increased computational complexity (Fig. 1, bottom). Although less accurate, optimal first-order methods have cheaper iterations and can deal with very large scale datasets (see Table 2). All the experiments were conducted on an Intel(R) Xeon(R) CPU @2.80GHz.

6.2. Synthetic data

In order to be able to compare our approach to other state-of-the-art algorithms, we place ourselves in a setting of homoscedastic noise with known ground truth. We randomly generate a matrix $\mathbf{X} \in \mathbb{R}^{T\times p}$ with i.i.d. standard Gaussian entries and a standard Gaussian noise vector $\boldsymbol{\xi} \in \mathbb{R}^T$ independent of $\mathbf{X}$. The noise variance is defined by $\sigma_t \equiv \sigma^*$ with varying values $\sigma^* > 0$.

p	200	400	600	800	1000
IP (sec/iter)	0.14	0.70	2.15	4.68	9.46
OFO (100*sec/iter)	0.91	1.07	1.33	1.64	1.91

Table 2. Comparing implementations of the ScHeDs: interior point (IP) *vs.* optimal first-order (OFO) method. We report the time per iteration (in seconds) for varying p in the experiment described in Section 6.2 with $T = 200$, $s^* = 2$, $\sigma = 0.1$. Note that the iterations of the OFO are very cheap and their complexity increases linearly in p.

We set $\boldsymbol{\beta}^0 = [\mathbf{1}_{s^*},\ \mathbf{0}_{p-s^*}]^\top$ and define $\boldsymbol{\phi}^* = \boldsymbol{\beta}^*/\sigma^*$, where $\boldsymbol{\beta}^*$ is obtained by randomly permuting the entries of $\boldsymbol{\beta}^0$. Finally, we set $\boldsymbol{Y} = \sigma^*(\mathbf{X}\boldsymbol{\phi}^* + \boldsymbol{\xi})$.

Seven different settings depending on the values of (T, p, s^*, σ^*) are considered. In each setting the experiment is repeated 500 times; the average errors of estimation of $\boldsymbol{\beta}^*$, s^* and σ^* for our procedure and for the Square-root Lasso are reported in Table 1 along with the standard deviations. For both procedures, the universal choice of tuning parameter $\lambda = \sqrt{2\log(p)}$ is used (after properly normalizing the columns of $\mathbf{X}$) and a second step consisting in bias correction is applied (*cf.* (Sun and Zhang, 2012) and the discussion in Section 4 on bias correction). Here, we did not use any group structure so the penalty is merely proportional to the ℓ_1-norm of $\boldsymbol{\beta}$. One can observe that the ScHeDs is competitive with the Square-root Lasso, especially for performing variable selection. Indeed, in all considered settings the ScHeDs outperforms the Square-root Lasso in estimating s^*.

6.3. Application to the prediction of the temperature in Paris

For experimental validation on a real-world dataset, we have used data on the daily temperature in Paris from 2003 to 2008. It was produced by the National Climatic Data Center (NCDC), (Asheville, NC, USA) and is publicly available at `ftp://ftp.ncdc.noaa.gov/pub/data/gsod/`. Performing good predictions for these data is a challenging task since, as shown in Fig. 2, the observations look like white noise. The dataset contains the daily average temperatures, as well as some other measurements like wind speed, maximal and minimal temperatures, *etc.*

We selected as response variable y_t the difference of temperatures between two successive days. The goal was to predict the temperature of the next daybased on historical data. We selected as covariates $\boldsymbol{x}_t$ the time t, the increments of temperature over past 7 days, the maximal intraday variation of the temperature over past 7 days and the wind speed of the day before. Including the intercept, this resulted in a 17 dimensional

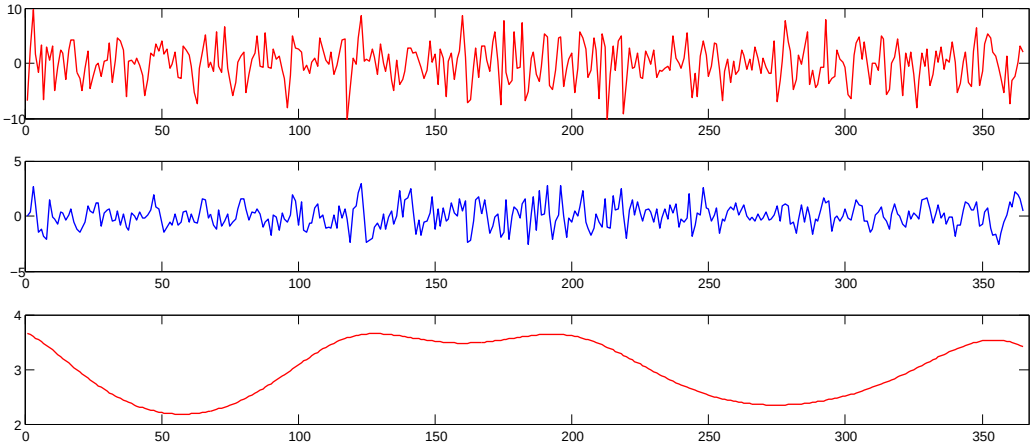

Figure 2. Top row: increments of temperatures (in Fahrenheit) from one day to the next observed in Paris in 2008. Middle row: predictions provided by our ScHeDs procedure; we observe that the sign is often predicted correctly. Bottom row: estimated noise level.

vector $\boldsymbol{x}_t$. Based on it, we created 136 groups of functions f, each group containing 16 elements. Thus, the dimension of $\boldsymbol{\phi}^*$ was $136 \times 16 = 2176$. We chose $q = 11$ with functions r_ℓ depending on time t only. The precise definitions of f_j and r_ℓ are presented below.

To specify $\mathbf{X}$, we need to define the functions f_j generating its columns. We denote by $\boldsymbol{u}_t$ the subvector of $\boldsymbol{x}_t$ obtained by removing the time t. Thus, $\boldsymbol{u}_t$ is a 16-dimensional vector. Using this vector $\boldsymbol{u}_t \in \mathbb{R}^{16}$, we define all the second-order monomes: $\chi_{i,i'}(\boldsymbol{u}_t) = u_t^{(i)} u_t^{(i')}$ with $i \leq i'$. We look for fitting the unknown function f^* by a second-order polynomial in $\boldsymbol{u}_t$ with coefficients varying in time. To this end, we set $\psi_1(t) = 1$, $\psi_\ell(t) = t^{1/(\ell-1)}$, for $\ell = 2, 3, 4$ and

$$\psi_\ell(t) = \cos(2\pi(\ell-4)t/365); \qquad \ell = 5, \ldots, 10;$$
$$\psi_\ell(t) = \sin(2\pi(\ell-10)t/365); \qquad \ell = 11, \ldots, 16.$$

Once these functions $\chi_{i,i'}$ and ψ_ℓ defined, we denote by f_j the functions of the form $\psi_\ell(t)\chi_{i,i'}(\boldsymbol{u}_t)$. In other terms, we compute the tensor product of these two sets of functions, which leads to a set of functions $\{\mathsf{f}_j\}$ of cardinality $16 \times 16 \times 17/2 = 2176$. These functions are split into 136 groups of 16 functions, each group defined by $G_{i,i'} = \{\psi_\ell(t) \times \chi_{i,i'}(\boldsymbol{u}_t) : \ell = 1, \ldots, 16\}$.

We defined $\mathbf{R}$ as a $T \times 11$ matrix, each of its eleven columns was obtained by applying some function r_ℓ to the covariate $\boldsymbol{x}_t$ for $t = 1, \ldots, T$. The functions r_ℓ were chosen as follows: $\mathsf{r}_1(\boldsymbol{x}_t) = 1$, $\mathsf{r}_2(\boldsymbol{x}_t) = t$, $\mathsf{r}_3(\boldsymbol{x}_t) = 1/(t + 2 \times 365)^{\frac{1}{2}}$ and

$$\mathsf{r}_\ell(\boldsymbol{x}_t) = 1 + \cos(2\pi(\ell-3)t/365); \qquad \ell = 4, \ldots, 7;$$
$$\mathsf{r}_\ell(\boldsymbol{x}_t) = 1 + \cos(2\pi(\ell-7)t/365); \qquad \ell = 8, \ldots, 11.$$

Note that these definitions of $\mathbf{X}$ and $\mathbf{R}$ are somewhat arbitrary. Presumably, better results in terms of prediction would be achieved by combining this purely statistical approach with some expert advice.

We used the temperatures from 2003 to 2007 for training (2172 values) and those of 2008 (366 values) for testing. Applying our procedure allowed us to reduce the dimensionality of $\boldsymbol{\phi}$ from 2176 to 26. The result of the prediction for the increments of temperatures in 2008 is depicted in Fig. 2. The most important point is that in 62% of the cases the sign of the increments is predicted correctly. It is also interesting to look at the estimated variance: it suggests that the oscillation of the temperature during the period between May and July is significantly higher than in March, September and October. Interestingly, when we apply a Kolmogorov-Smirnov test to the residuals $y_t \boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}} - \boldsymbol{X}_{t,:}\widehat{\boldsymbol{\phi}}$ for t belonging to the testing set, the null hypothesis of Gaussianity is not rejected and the p value is 0.72.

7. Conclusion

We have introduced a new procedure, the ScHeDs, that allows us to simultaneously estimate the conditional mean and the conditional variance functions in the model of regression with heteroscedastic noise. The ScHeDs relies on minimizing a group-sparsity promoting norm under some constraints corresponding to suitably relaxed first-order conditions for maximum penalized likelihood estimation. We have proposed several implementations of the ScHeDs based on various algorithms of second-order cone programming. We have tested our procedure on synthetic and real world datasets and have observed that it is competitive with the state-of-the-art algorithms, while being applicable in a much more general framework. Theoretical guarantees for this procedure have also been proved.

References

Anestis Antoniadis. Comments on: ℓ_1-penalization for mixture regression models. *TEST*, 19(2):257–258, 2010.

Alfred Auslender and Marc Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16(3):697–725 (electronic), 2006.

Francis Bach. Consistency of the Group Lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

Stephen Becker, Emmanuel J. Candès, and Michael C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.*, 3(3):165–218, 2011.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

Emmanuel Candès and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

Christophe Chesneau and Mohamed Hebiri. Some theoretical results on the grouped variables Lasso. *Math. Methods Statist.*, 17(4):317–326, 2008.

Arnak Dalalyan and Yin Chen. Fused sparsity and robust estimation for linear models with unknown variance. In *NIPS*, pages 1268–1276. 2012.

John Daye, Jinbo Chen, and Hongzhe Li. High-dimensional heteroscedastic regression with an application to eQTL data analysis. *Biometrics*, 68(1):316–326, 2012.

Eric Gautier and Alexandre Tsybakov. High-dimensional instrumental variables regression and confidence sets. Technical Report arxiv:1105.2454, September 2011.

Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high dimensional models. *Statist. Sci.*, 27(4):481–499, 2012.

Junzhou Huang and Tong Zhang. The benefit of group sparsity. *Ann. Statist.*, 38(4):1978–2004, 2010.

Mladen Kolar and James Sharpnack. Variance function estimation in high-dimensions. In *Proceedings of the ICML-12*, pages 1447–1454, 2012.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *Ann. Statist.*, 38(6):3660–3695, 2010.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, 34(5):2272–2297, 2006.

Han Liu, Jian Zhang, Xiaoye Jiang, and Jun Liu. The group Dantzig selector. *J. Mach. Learn. Res. - Proc. Track*, 9:461–468, 2010.

Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4): 2164–2204, 2011.

Julien Mairal, Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Convex and network flow optimization for structured sparsity. *J. Mach. Learn. Res.*, 12: 2681–2720, 2011.

Lukas Meier, Sara van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B): 3779–3821, 2009.

Yuval Nardi and Alessandro Rinaldo. On the asymptotic properties of the group Lasso estimator for linear models. *Electron. J. Stat.*, 2:605–633, 2008.

Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427, 2012.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(5):1009–1030, 2009.

Noah Simon and Robert Tibshirani. Standardization and the Group Lasso penalty. *Stat. Sin.*, 22(3):983–1001, 2012.

Nicolas Städler, Peter Bühlmann, and Sara van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19 (2):209–256, 2010.

Jos F. Sturm. Using sedumi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999.

Tingni Sun and Cun-Hui Zhang. Comments on: ℓ_1-penalization for mixture regression models. *TEST*, 19 (2):270–275, 2010.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

Jens Wagener and Holger Dette. Bridge estimators and the adaptive Lasso under heteroscedasticity. *Mathematical Methods of Statistics*, 21:109–126, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.