Statistics and Computing manuscript No.
(will be inserted by the editor)

A consistent and numerically efficient variable selection method for sparse Poisson regression with applications to learning and signal recovery

Sabrina Guastavino · **Federico Benvenuto**

Received: date / Accepted: date

Abstract We propose an adaptive ℓ_1-penalized estimator in the framework of Generalized Linear Models with identity-link and Poisson data, by taking advantage of a globally quadratic approximation of the Kullback-Leibler divergence. We prove that this approximation is asymptotically unbiased and that the proposed estimator has the variable selection consistency property in a deterministic matrix design framework. Moreover, we present a numerically efficient strategy for the computation of the proposed estimator, making it suitable for the analysis of massive counts datasets. We show with two numerical experiments that the method can be applied both to statistical learning and signal recovery problems.

Keywords Adaptive regularization · Lasso · Model selection · Sparse Poisson regression · Statistical learning · Image processing

Mathematics Subject Classification (2000) 62G08 · 62G20 · 62J07

1 Introduction

Variable selection for Poisson regression is a common task in both sparse signal recovery and statistical learning. In the first case the idea is to find the smallest number of elements of a suitable basis to represent an unknown signal, as for example in astronomy and medical imaging; in the second case the aim is to identify important covariates for prediction, with applications, by instance, in medicine, engineering and social sciences. In sparse signal recovery with Poisson data a lot of attention has been paid on fast and efficient optimization methods especially when the number of data is high and therefore a large scale inverse problem has to be solved. Recent improvements have been focused on acceleration of the usual proximal gradient methods requiring sophisticated optimization techniques and first order approximations of the objective function (Gu and Dogandžić 2014; Harmany et al 2010; Figueiredo and Bioucas-Dias 2010; Bonettini et al 2009). On the other hand, in statistical learning a special effort has been provided in promoting consistent variable selection and estimation. To this aim, one of the most used methods is Lasso (Tibshirani 1996) which performs sign consistent selection under the so-called Irrepresentable Condition (Zhao and Yu 2006). A major step forward in this direction was the introduction of adaptive Lasso, which guarantees variable selection consistency in the case of Generalized Linear Models (GLMs) under less restrictive statistical assumptions (Zou 2006). The idea of the adaptive approach is to introduce weights in the standard ℓ_1-penalized Lasso problem for enhancing the sparsity of the solution and it has been put forward again in different ways (Bogdan et al 2015; Candès et al 2008). For Poisson GLMs the use of data-driven adaptive weights has been recently proposed: in Jiang et al (2015) authors adapted Lasso to work with Poisson data by means of a particular choice of the adaptive weights, while in Ivanoff et al (2016) authors proposed a choice based on concentration inequalities for solving an adaptive problem arising from the Poisson GLM with the canonical log link.

In this work we propose a data-dependent global quadratic approximation of the Poisson log likelihood enabling us to formulate a simplified adaptive Lasso estimator

The research leading to these results has received funding from the European Unions Horizon2020 research and innovation programme under grant agreement no. 640216.

Sabrina Guastavino
Department of Mathematics, Universita' degli Studi di Genova
E-mail: guastavino@dima.unige.it

Federico Benvenuto
Department of Mathematics, Universita' degli Studi di Genova
E-mail: benvenuto@dima.unige.it

suitable for sparse Poisson regression. The proposed estimator has two main properties: first, despite its simplified form, it performs consistent variable selection and second, it can be computed by taking advantages of the fastest available algorithms, i.e. those developed for ℓ_1-penalized least squares regression (Friedman et al 2007, 2010; Beck and Teboulle 2009). Indeed, since it is well known that Poisson GLM solvers can suffer from convergence issues (Silva and Tenreyro 2011; Marschner and others 2011), the added value of the proposed method is also in terms of both convergence speed and stability. In this connection, path following algorithms exploiting the piecewise linear form of the regularization path can be adopted for the solution of Poisson sparse recovery and learning problems, provided that the number of relevant predictors is relatively small (Efron et al 2004). Moreover, the proposed estimator is particularly suitable to the case of large number of samples, as for large scale learning and inverse problems, making a consistent variable selection for massive counts datasets feasible. In this respect, we proved the consistency property of the variable selection as the number of data increases and the number of unknowns is fixed. Our results are based on some mild assumptions analogous to those used in Zou and Zhang (2009), which are suitable for applications with deterministic design matrices. It is worth noticing that the consistency property has different implications depending on the application: for signal recovery problems, consistency is computed against the increasing number of bins/pixels in which the signal is measured, while for statistical learning it is evaluated against the increasing number of available examples. For a detailed discussion on this topic see, for example De Vito et al (2005). Finally, the proposed approximation is based on the Poisson GLM with identity link, which is appropriate in the large majority of signal recovery problems, being physical signal formation models usually linear or, at least, linearized. However, the use of the identity link is not a limitation for statistical learning. Indeed, although the use of a Poisson GLM in real applications can be natural and well-justified, the determination of the best link function mostly concerns the predictive capabilities of the model. In the application section, we will see, with the help of a synthetic example, that using estimators based on the identity link function is appropriate for variable selection even when data have been generated by means of a model based on the log link function.

The paper is organized as follows. In Section 2 we introduce the Poisson variable selection problem with deterministic matrix design suitable for the description of many signal processing and learning applications. In Section 3 we describe the approximation of the penalized method for Poisson data and establish the model selection consistency for this method. In Section 4 we present a learning application for showing the consistency of the method with Poisson data and large scale denoising and deblurring problems for objects with sparse representation. Conclusions are drawn in Section 5. Technical details of the proofs are presented in Section 6.

2 Model description

Let us consider a Poisson random vector $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ made of independently distributed components with mean $\mu^* = (\mu_1^*, \dots, \mu_n^*)^T$, i.e.

$$Y_i \sim \text{Poisson}(\mu_i^*), \tag{1}$$

$\forall\, i \in \{1, \dots, n\}$. Suppose that the parameter μ_i^* can be expressed as

$$\mu_i^* = g^{-1}(\mathbf{X}\beta^*)_i, \tag{2}$$

$\forall\, i \in \{1, \dots, n\}$, where $g : \mathbb{R} \to \mathbb{R}$ is an invertible function, $\mathbf{X} := [\mathbf{x}_1, \dots, \mathbf{x}_p]$ is a $n \times p$ matrix with columns $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T$ with $j \in \{1, \dots, p\}$ and $\beta^* = (\beta_1^*, \dots, \beta_p^*)^T$ is a suitable vector of parameters. In statistical estimation $\mathbf{Y}$ is called the response vector, $\mathbf{X}$ is the predictor (or feature, or design) matrix, g is called the link function and equation (2) describes the GLMs (McCullagh and Nelder 1989). On the other hand, in signal recovery $\mathbf{Y}$ represents the vector of noisy measurements of a given random signal and $\mathbf{X}$ describes a linear signal formation process depending on the parameters β^*. In this paper, we assume that the true unknown vector β^* is sparse. More formally, let us denote with

$$\mathscr{A}^* := \{j \in \{1, \dots, p\} : \beta_j^* \neq 0\}$$

the set of indexes corresponding to relevant variables of the model, namely the active set, and with $|\mathscr{A}^*|$ its cardinality. We suppose that

$$q := |\mathscr{A}^*| < p\,.$$

In applications we consider q as being a substantially smaller fraction of p. Such assumption leads to the variable selection and estimation problem, i.e. to compute a model with a small number of relevant variables with good prediction capabilities (Friedman et al 2001). The standard choice for g in the statistical learning framework is the so-called canonical link function of the GLM theory, i.e. the logarithm function
$g(z) := \ln(z)$. In this way, Poisson means are equal to the exponents of linear predictors, i.e. $\mu = \exp(\mathbf{X}\beta^*)$, taking positive values only. In the case of Poisson regression with canonical link, an usual variable selection method comes from extending the Adaptive Lasso to the GLMs, suggesting the following estimator

$$\hat{\beta}^{(n)}(\text{log link}) := \arg\min_{\beta} \sum_{i=1}^n \exp(\mathbf{X}\beta)_i - Y_i(\mathbf{X}\beta)_i + \lambda \sum_{j=1}^p w_j|\beta_j|, \tag{3}$$

where $\beta \in \mathbb{R}^p$, λ is the positive regularization parameter, $\mathbf{w} = (w_j)_{j=1,\dots,p}$ is the weights vector, which has the role of weighting the contribution of the coefficients β_j.

Another possible choice for g is the identity link, i.e. $g(z) := \mathrm{id}(z)$ under the non-negativity constraint $z_i > 0$, $\forall i \in \{1,\dots,n\}$. This choice is natural in a large variety of applications, e.g. in emission tomography and in astronomical image reconstruction and deblurring, since the matrix $\mathbf{X}$ is able to describe a linear transformation which approximates the physical signal formation process (Prince and Links 2006; Starck and Murtagh 2007). In the unconventional case of Poisson GLM with identity-link the adaptive Lasso estimator can be found by minimizing the ℓ_1-penalized Kullback-Leibler divergence as follows

$$\hat{\beta}^{(n)}(\text{id link}) := \arg\min_{\beta \in \mathscr{C}} \sum_{i=1}^{n} D((\mathbf{X}\beta)_i, Y_i) + \lambda \sum_{j=1}^{p} w_j |\beta_j|, \quad (4)$$

where

$$\mathscr{C} = \{\beta \in \mathbb{R}^p : (\mathbf{X}\beta)_i > 0 \; \forall \, i \in \{1,\dots,n\}\} \quad (5)$$

is the subset of feasible β solutions and D is the Kullback-Leibler divergence, which is defined as

$$D(z,y) := y \log \frac{y}{z} + z - y, \quad (6)$$

with $z, y > 0$ and $D(z,0) := 0$. The presence of such additional constraint $(\mathbf{X}\beta)_i > 0$, $\forall \; i \in \{1,\dots,n\}$ can be a disadvantage of using the identity link. Indeed, this can result in the need for much more computationally expensive optimization methods. However, in applications the vector β^* often contains an offset parameter associated with a constant value predictor, which usually makes the quantity $\mathbf{X}\beta$ substantially larger than zero. As a consequence the solution of the problem is an interior point of the feasible solution set (5). This offset is called 'the intercept' in the statistical language and 'the background' in signal recovery.

In Zou (2006) it has been proven that, by choosing the weights in an appropriate manner, both the estimators $\hat{\beta}^{(n)}(\text{log link})$ and $\hat{\beta}^{(n)}(\text{id link})$ perform consistent variable selection and estimation, under some mild regularity conditions where both $\mathbf{X}$ and $\mathbf{Y}$ are thought of as random variables. Now we introduce an approximation of the functional (4) which allows us to define an adaptive penalized reweighted least squares method with the property to identify the exact relevant explanatory variables when the number of observations diverges in a deterministic matrix design framework. At the same time, such an approximation overcomes the need for expensive optimization methods such as the Iterative Reweighted Least Squares (IRLS) commonly applied in the case of GLMs (Dobson and Barnett 2008).

3 Adaptive Poisson Reweighted Lasso

In this section we first present the theoretical properties of the proposed estimator and after a numerically efficient approach to compute it.

3.1 Theory

We now show a global quadratic approximation of the KL divergence and we prove that such an approximation is an asymptotically unbiased estimator of the KL divergence. Formally, we have the following

Theorem 1 *Let y be a Poisson random variable with mean θ. For any $z > 0$ such that $|z-\theta| \le c\sqrt{\theta}$ with $c > 0$ such that $\theta - c\sqrt{\theta} > 0$, we have*

$$\mathbb{E}\left(D(z,y) - \frac{1}{2}\frac{(y-z)^2}{y+1} \right) = O\left(\frac{1}{\sqrt{\theta}} \right), \quad (7)$$

as $\theta \to \infty$.

The proof of Theorem 1 is given in the Appendix. Theorem 1 implies that for all $i \in \{1,\dots,n\}$, in a neighborhood of the exact values $(\mathbf{X}\beta^*)_i$, such an approximation is more and more accurate with $(\mathbf{X}\beta^*)_i \to \infty$. This approximation calls up to the Anscombe transform (Anscombe 1948). Nonetheless, the substantial difference is that the proposed approximation (7) is globally quadratic making its numerical treatment extremely easier. In view of this term by term approximation, we can introduce a novel estimator on the basis of a positive weight vector $\mathbf{w} = \{w_j\}_{j \in \{1,\dots,p\}}$, as follows

$$\hat{\beta}^{(n)}_{(\mathbf{w},\lambda)} := \arg\min_{\beta \in \mathscr{C}} \frac{1}{2} \sum_{i=1}^{n} \frac{(Y_i - (\mathbf{X}\beta)_i)^2}{Y_i + 1} + \lambda \sum_{j=1}^{p} w_j |\beta_j|, \quad (8)$$

where λ is the regularization parameter. Therefore, functional in the r.h.s. of equation (8) is an asymptotically unbiased estimator of the functional in the r.h.s. of equation (4). In the case weights are all equal to 1, i.e. $w_j = 1$ for any j, the estimator is the minimizer of a functional that we call 'Poisson Reweighted Lasso' (PRiL). We denote it by $\hat{\beta}^{(n)}(\text{PRiL})$ and we notice that it depends on a regularization parameter that we call λ_1. We prove in the following that this choice of the weights makes $\hat{\beta}^{(n)}(\text{PRiL})$ a $\sqrt{n}$-consistent estimator, provided an appropriate asymptotics of the regularization parameter λ_1 is given. Data-dependent choices of the weights w_j in the case of Poisson problems have been recently proposed in Jiang et al (2015); Hansen et al (2015) and are based on Poisson concentration inequalities. In all cases the idea is to choose such weights in order to provide the method with the asymptotic model selection consistency

property. Inspired by the choice in Zou and Zhang (2009) for the adaptive elastic net, we introduce the following weights

$$\hat{w}_j = \frac{1}{\left(|\hat{\beta}^{(n)}(\mathrm{PRiL})_j| + \left(\frac{1}{n}\right)^{\frac{1}{\gamma}+\delta}\right)^{\gamma}}, \tag{9}$$

where γ and δ are strictly positive constants. The estimator (8) when provided with such weights is called 'APRiL' for Adaptive Poisson Reweighted Lasso and we denote it by $\hat{\beta}^{(n)}(\mathrm{APRiL})$. Now, the main goal is to prove that the $\hat{\beta}^{(n)}(\mathrm{APRiL})$ estimator has the model selection consistency property in the case of Poisson data and under some assumptions on the matrix $\mathbf{X}$. In particular, let Λ be the following $n \times n$ diagonal matrix

$$\Lambda = diag\left(\frac{1}{\sqrt{Y_1+1}}, \ldots, \frac{1}{\sqrt{Y_n+1}}\right). \tag{10}$$

We assume that:

(H1) the matrix $\mathbf{X}^T\Lambda^2\mathbf{X}$ is positive definite, and

$$\mathbb{E}\left(\left(\frac{1}{\tau_{\min}(\mathbf{X}^T\Lambda^2\mathbf{X})}\right)^4\right) \leq \frac{1}{(bn)^4} \quad \text{and}$$
$$\tau_{\max}(\mathbf{X}^T\mathbf{X}) \leq Bn$$

where $\tau_{\min}(A)$ and $\tau_{\max}(A)$ are the minimum and maximum eigenvalues of the matrix A respectively, b and B are two strictly positive constants

(H2) $\lim_{n\to+\infty} \frac{\lambda_1}{\sqrt{n}} = 0$

(H3) a) $\lim_{n\to+\infty} \lambda n^{\frac{\gamma}{2}-1} = \infty$, b) $\lim_{n\to+\infty} \lambda n^{\delta\gamma} = \infty$, c) $\lim_{n\to+\infty} \lambda n^{\delta\gamma-\frac{1}{2}} = 0$

(H4) there exists an $L > 0$ such that

$$\max_{j\in\{1,\ldots,p\}} \|\mathbf{x}_j\|_2 \leq L.$$

Assumptions in (H1) involve the matrix $\mathbf{X}$ and the random variable $\mathbf{Y}$. The hypothesis concerning $\tau_{\min}$ implies that

$$\mathbb{E}\left(\tau_{\min}\left(\frac{\mathbf{X}^T\Lambda^2\mathbf{X}}{n}\right)\right) \geq b \tag{11}$$

which calls up to the assumption used by Zou and Zhang (2009). Assumption (H2) involves the convergence rate of the regularization parameter λ_1 whereas assumptions described in (H3) involve the convergence rate of regularization parameter λ. We remark that, by choosing $\gamma > 2$ and $\delta < \frac{1}{2\gamma}$ assumptions in (H3) let the asymptotic behavior of the regularization parameter λ be zero, constant or infinity. Assumption (H4) is necessary for consistent model selection and it is automatically verified after the feature standardization/normalization procedure. In the following theorem we give a general bound of the expected error for the estimator (8).

Theorem 2 *Assuming hypothesis (H1), then it exists a constant $G < +\infty$, such that*

$$\mathbb{E}(\|\hat{\beta}^{(n)}_{(\mathbf{w},\lambda)} - \beta^*\|_2^2) \leq \frac{4\lambda^2\sqrt{\mathbb{E}\left(\left(\sum_{j=1}^p w_j^2\right)^2\right)} + pGBn}{(bn)^2}. \tag{12}$$

The proof of Theorem 2 is given in the Appendix. Such a bound takes into account that weights can be random variables. In the case weights are constants all equal to 1, the previous result boils down to the following

Corollary 1 *Assuming hypothesis (H1) then*

$$\mathbb{E}(\|\hat{\beta}^{(n)}(\mathrm{PRiL}) - \beta^*\|_2) \leq \frac{2\lambda_1\sqrt{p} + \sqrt{pGBn}}{bn}. \tag{13}$$

It is worth observing that under assumption (H2) Corollary 1 implies that $\hat{\beta}^{(n)}(\mathrm{PRiL})$ is a $\sqrt{n}$-consistent estimator. Finally, we introduce the estimated active index set

$$\hat{\mathscr{A}}^{(n)} = \{j \in \{1,\ldots,p\} : \hat{\beta}^{(n)}(\mathrm{APRiL})_j \neq 0\} \tag{14}$$

of the estimator $\hat{\beta}^{(n)}(\mathrm{APRiL})$. The model selection consistency property reads

$$\lim_{n\to+\infty} \mathbb{P}(\hat{\mathscr{A}}^{(n)} = \mathscr{A}^*) = 1\,. \tag{15}$$

Theorem 3 *Under assumptions (H1), (H2), (H3), (H4) the APRiL estimator has the model selection consistency property.*

The proof of the Theorem (given in the Appendix) is substantially based on the $\sqrt{n}$-consistency property of the estimator $\hat{\beta}^{(n)}(\mathrm{PRiL})$. This property underpins the choice of the weights defined in equation (9).

3.2 Computations

The computation of the APRiL estimator can be performed by means of the same numerically efficient algorithms developed for the solution of the Lasso problem. We propose a numerical strategy which consists of two steps. First we reweight the columns of the matrix $\mathbf{X}$ and the vector $\mathbf{Y}$ by left-multiplying by Λ defined in equation (10). Second, following the approach proposed by Zou (2006), we reweight the predictor matrix $\mathbf{X}$ for computing the adaptive solution. These two steps need the computation of the solution of two Lasso problems. In Algorithm 1 we outline the scheme of the procedure.

In many applications the presence of an offset - be it a regression intercept or a constant background signal - makes

Algorithm 1 APRiL estimator computation

1: Input: $\mathbf{X}$, $\mathbf{Y}$
2: **Data driven reweighting**. Define

$$\tilde{\mathbf{X}} := \Lambda \mathbf{X} \qquad \tilde{\mathbf{Y}} := \Lambda \mathbf{Y}$$

where Λ is defined in equation (10).
3: Compute the regularization path

$$\hat{\beta}_{\lambda_1} = \arg\min_{\beta \in \mathscr{C}} \frac{1}{2} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|$$

and select $\hat{\beta}_{\lambda_1}$ (PRiL) with λ_1 according to a cross validation process.
4: Compute the adaptive weights $\hat{\mathbf{w}}$ as in formula (9).
5: **Adaptive reweighting**. Define $\tilde{\tilde{\mathbf{X}}}$ so that

$$\tilde{\tilde{\mathbf{x}}}_j = \tilde{\mathbf{x}}_j / \hat{w}_j, \ \forall\, j \in \{1, \dots, p\}.$$

6: Compute the regularization path

$$\tilde{\hat{\beta}}_\lambda = \arg\min_{\beta \in \mathscr{C}} \frac{1}{2} \|\tilde{\mathbf{Y}} - \tilde{\tilde{\mathbf{X}}}\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

and select $\tilde{\hat{\beta}}_\lambda$ with λ according to a cross validation process.
7: Output: $\hat{\beta}_\lambda$ (APRiL) is such that

$$\hat{\beta}_\lambda(\text{APRiL})_j = (\tilde{\hat{\beta}}_\lambda)_j / \hat{w}_j \quad \forall j \in \{1, \dots, p\}.$$

the vector $\mathbf{X}\hat{\beta}_\lambda$ an interior point of the feasible set $\mathscr{C}$, i.e all its components are positive. Moreover, we notice that, unlike the functional in equation (4) which is based on the KL divergence, the proposed functional in equation (8) is meaningful for each β, even when $\mathbf{X}\beta$ has negative components. In such cases, the constraint $\mathscr{C}$ can be neglected during the optimization process and standard algorithms can be used in place of sophisticated constrained techniques. Therefore, steps 3 and 6 of the Algorithm 1 can be performed by solving the unconstrained Lasso problem. In this way, APRiL method can take advantage of numerically efficient solvers and of the piece-wise linear form of the regularization path (Efron et al 2004).

4 Simulation studies

In this section we show two applications of the proposed adaptive method. In the first one, we apply it to some statistical learning test problems and in the second one, we show that it can be successfully applied to wavelet-based Poisson denoising and deblurring. One of the main difference between these applications is that in the first case the model (or the link function) is not known while in the second case it is a linear operator representing the signal formation process. This leads us to make a performance comparison between our method and the adaptive technique for GLMs with Poisson data in the statistical learning application, and to check the performance of the proposed method in the sparse signal recovery one.

4.1 Statistical learning application

We present a synthetic variable selection problem in order to compare the proposed method with the adaptive GLM for Poisson data (Zou 2006). The main goal of this synthetic experiment is to assess the variable selection performance of the proposed method as the number of samples increases and its computational advantages when the number of samples reaches the order of million. It is worth observing that in statistical learning regression methods are based on a given data model (equation (2)), i.e. on a particular choice of the link function. The standard method based on GLM theory uses the log-link function, which is the canonical choice for Poisson data, while the proposed method is based on the identity-link. Therefore, in order to perform a comprehensive comparison of the two methods, we consider two sets of data generated according to the log-link and the identity-link function based model, respectively. We are interested in evaluating the performance of the APRiL method and the Adaptive GLM (AGLM) by applying them to both datasets. In particular, these two datasets are generated according to the following assumptions. We fix $p = 15$ and $q = |\mathscr{A}^*| = 5$. We construct the $n \times p$ predictor matrix $\mathbf{X}$ for $n = 125, 250, 500$, so that each of its columns is extracted by a p-dimensional normal multivariate distribution with zero mean and covariance Σ with $\Sigma_{jr} = \rho^{|j-r|}$, for $j, r \in \{1, \dots, p\}$. We assume $\rho = 0.5$ and $\rho = 0.75$. We consider the following two cases:

1. Log-link dataset. We generate the data $\mathbf{Y}$ by using log-link function as follows

$$Y_i = \text{Poisson}(\beta_0^* \exp((\mathbf{X}\beta^*)_i)), \ \forall\, i \in \{1, \dots, n\}, \qquad (16)$$

where $\beta^* = (0.7, -0.5, 0.3, -0.4, 0.6, \mathbf{0}_{p-5})^T$ is the true coefficient vector and β_0^* is a suitable constant intercept.
2. Identity-link dataset. We generate $\mathbf{Y}$ by using the identity-link function as follows

$$Y_i = \text{Poisson}((\mathbf{X}\beta^{**})_i + \beta_0^{**}), \ \forall\, i \in \{1, \dots, n\}, \qquad (17)$$

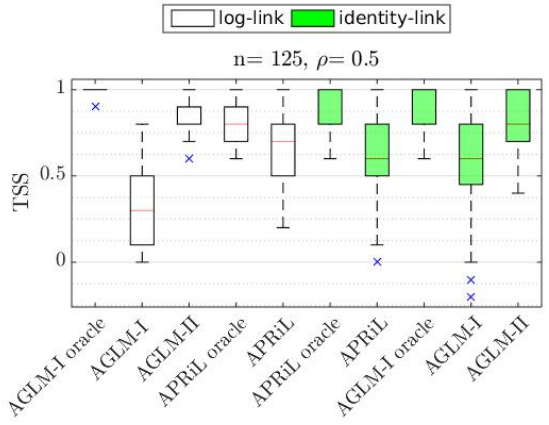
where $\beta^{**} = (e^{0.7}, e^{-0.5}, e^{0.3}, e^{-0.4}, e^{0.6}, \mathbf{0}_{p-5})^T$ is the true coefficient vector and β_0^{**} is a suitable constant intercept.

In the second case we select the intercept in order to make each component of the vector $(\mathbf{X}\beta^{**})_i + \beta_0^{**}$ positive. In the first case we tune the intercept value so that data generated in the first case has about the same signal to noise ratio of the data generated in the second case. Moreover, for each problem, we generate 100 realizations of Poisson data

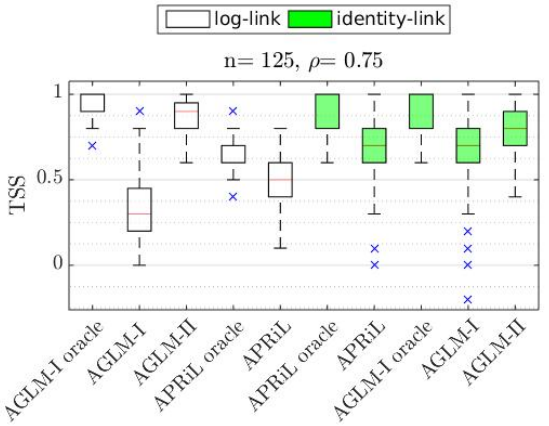
Table 1 Mean Square Error values obtained by averaging over 100 replicates the results provided by AGLM and APRiL method for each problem.

	log-link dataset			**identity-link dataset**		
n	125	250	500	125	250	500
	$\rho = 0.5$					
AGLM	$4.1_{(\pm 2.1)}\ 10^{-4}$	$2.1_{(\pm 1.2)}\ 10^{-4}$	$8_{(\pm 5)}\ 10^{-5}$	$6.2_{(\pm 0.1)}\ 10^{-1}$	$6.2_{(\pm 0.1)}\ 10^{-1}$	$6.2_{(\pm 0.1)}\ 10^{-1}$
APRiL	$1.5_{(\pm 0.6)}$	$1.8_{(\pm 0.6)}$	$4.4_{(\pm 0.5)}$	$2.4_{(\pm 0.5)}\ 10^{-1}$	$2.1_{(\pm 0.4)}\ 10^{-1}$	$1.9_{(\pm 0.2)}\ 10^{-1}$
	$\rho = 0.75$					
AGLM	$7.3_{(\pm 4.5)}\ 10^{-4}$	$4.1_{(\pm 2.2)}\ 10^{-4}$	$1.4_{(\pm 0.7)}\ 10^{-4}$	$6.3_{(\pm 0.1)}\ 10^{-1}$	$6.3_{(\pm 0.1)}\ 10^{-1}$	$6.3_{(\pm 0.03)}\ 10^{-1}$
APRiL	$2.7_{(\pm 1.2)}$	$5.6_{(\pm 1)}$	$8.3_{(\pm 1.2)}$	$2.7_{(\pm 0.7)}\ 10^{-1}$	$2.2_{(\pm 0.5)}\ 10^{-1}$	$2.1_{(\pm 0.4)}\ 10^{-1}$

and therefore we obtained 600 estimation problems ($\#n = 3$ and $\#\rho = 2$). For each one of these problems we perform regression by means of the APRiL and the AGLM methods.



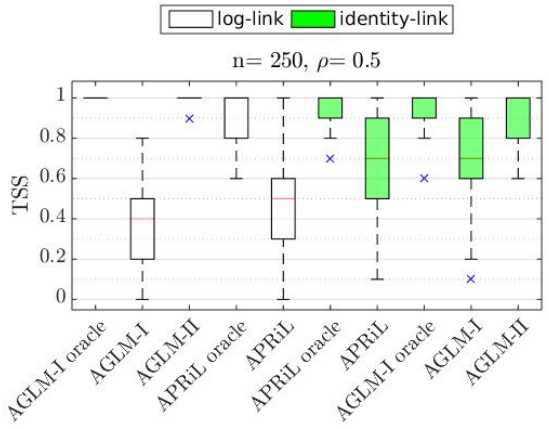
(a) case $\rho = 0.5$



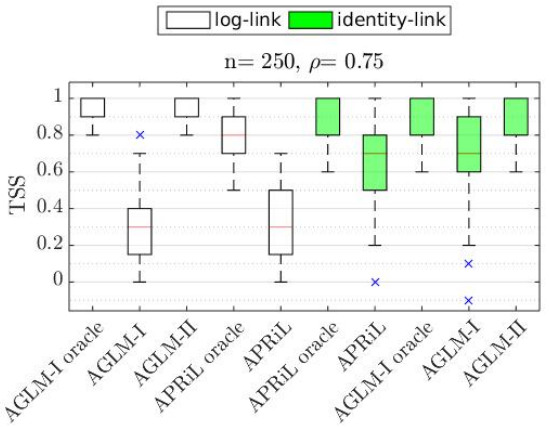
(b) case $\rho = 0.75$

Fig. 1 Comparing distributions of TSS, fixing number of samples equal to $n = 125$.

The APRiL weights are parametrized according with the assumptions in Theorem 3. In particular we use $\hat{w}_j$ as defined in equation (9), and we fix constants $\gamma = 3$ and



(a) case $\rho = 0.5$



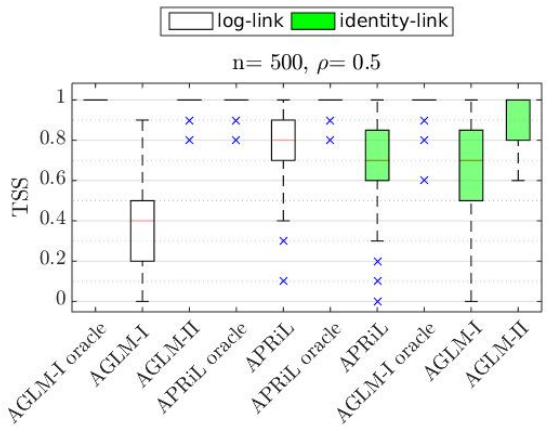
(b) case $\rho = 0.75$

Fig. 2 Comparing distributions of TSS, fixing number of samples equal to $n = 250$.

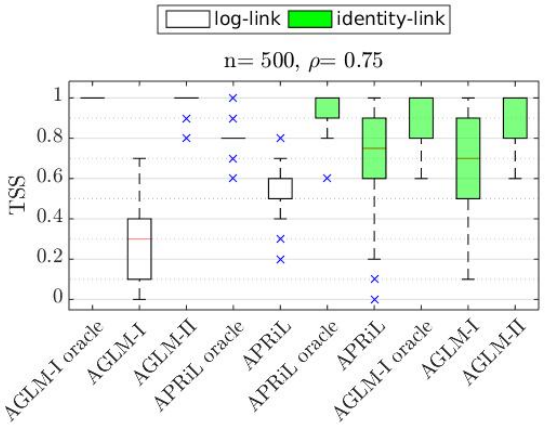
$\delta = \frac{1}{8}$. For what concerns AGLM defined in equation (3) we fix the weights as

$$\hat{w}_j = \frac{1}{|\hat{\beta}(\text{MLE})_j|^{\bar{\gamma}}} \ \forall\ j \in \{1, \dots, p\}\ , \tag{18}$$

where $\hat{\beta}(\text{MLE})$ is the maximum likelihood estimate in Poisson log-linear regression model and $\bar{\gamma}$ is a positive constant



(a) case $\rho = 0.5$



(b) case $\rho = 0.75$

Fig. 3 Comparing distributions of TSS, fixing number of samples equal to $n = 500$.

(Zou 2006). For computing the solution of these optimization problems we use the *glmnet* MATLAB package Friedman et al (2010). Moreover we select the regularization parameter by means of the 10-fold Cross Validation (CV) implemented in the same package. We use the mean squared error (MSE) for measuring the estimation accuracy of each 10-fold cross-validated APRiL and AGLM solution. In Table 1 we show MSE values for both algorithms. It is evident that the algorithm based on the same model by which data have been generated achieves a lower MSE. In other words, the AGLM method performs better when applied to the log-link dataset and the APRiL when applied to the identity-link dataset.

Moreover, we compare the variable selection performance of the AGLM and APRiL methods by computing the confusion matrix which represents matches and mismatches between predicted active variables and exact ones. On the basis of the components of the confusion matrix, i.e. false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN), we compute the True Skill Score (TSS) which is defined as the balance between the true positive rate and the false alarm rate, i.e.

$$\mathrm{TSS} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}\,, \tag{19}$$

and ranges from -1 to 1. The optimal variable selection is obtained when the TSS is 1 and a direct consequence of Theorem 3 is that the TSS value provided by the APRiL estimator converges to one in probability as n goes to infinity. For having a broader picture, for each method, in addition to the 10-fold cross-validated solution, we compute the solution which maximizes the TSS value along the regularization path, and we refer to it as the oracle solution. Oracle solutions allow us to make a performance assessment of the algorithms independently of the choice of the regularization parameter. Each box-plot in Figures 1, 2 and 3 shows the TSS distribution obtained by applying the algorithm written in the x-axis label to one hundred replicates of $\mathbf{Y}$. In each figure, from left to right the odd box-plots show the TSS value provided by the oracle solution while the even ones show the TSS value provided by the cross validated solution. The first four box-plots refer to the AGLM and APRiL algorithms applied to the log-link dataset (equation (16)), whereas the second four box-plots refer to the algorithms applied to the identity-link dataset (equation (17)).

Some comments about variable selection results:

1. The TSS provided by oracle AGLM solutions is larger than the one provided by the oracle APRiL solutions in all the experiments we performed. This can be explained by the fact that AGLM method is based on the maximization of the Poisson likelihood, which is the actual distribution used for generating data. Oracle solutions provided by the APRiL method, which is based on an approximation of the Poisson log-likelihood, do not achieve the same performance.
2. The use of CV procedure for finding the regularization parameter reduces the performance of the variable selection so that it does not seem to be an efficient method in the case of small and moderately sized samples. However, for large scale problems the regularization path is more stable and the CV selects a solution closer to the oracle one (Martinez et al 2011). In general, TSS distributions corresponding to cross validated solutions are over-dispersed and for each problem among the 100 replicates we can find a variable selection with a very low TSS value. Moreover CV behaves differently across algorithms. The striking fact is that the cross validated APRiL solution tends to produce a better variable selection than the cross validated AGLM one, overall in the case of smaller sized samples and log-link dataset.

Obtained results have been proven to be robust by varying the number of folds in the cross validation analysis and

Table 2 Computation mean time in seconds.

Link	Method	$n = 10^4$	$n = 5\ 10^4$	$n = 10^5$	$n = 5\ 10^5$	$n = 10^6$
		time in s				
log	AGLM	$1.6\ 10^{-2}$	$7.5\ 10^{-2}$	$1.5\ 10^{-1}$	$8.3\ 10^{-1}$	2.0
	APRiL	$9.6\ 10^{-4}$	$5.0\ 10^{-3}$	$1.0\ 10^{-2}$	$6.1\ 10^{-2}$	$1.3\ 10^{-1}$
identity	AGLM	$1.4\ 10^{-2}$	$6.5\ 10^{-2}$	$1.3\ 10^{-1}$	$7.3\ 10^{-1}$	1.8
	APRiL	$9.7\ 10^{-4}$	$5.1\ 10^{-3}$	$1.1\ 10^{-2}$	$6.2\ 10^{-2}$	$1.3\ 10^{-1}$

the definition of the adaptive weights. In this regard, we replicated the experiments introduced above by using the 5-fold cross validation and by choosing the adaptive weights of the AGLM method in a way analogous to the one described in equation (9) obtaining similar outcomes.

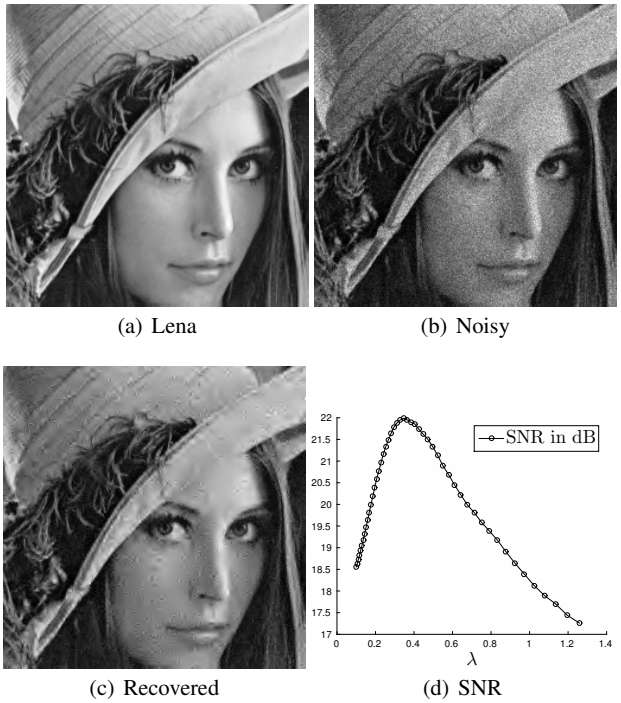
Finally, we check the numerical efficiency of the two algorithms. Following the above described setup, for each method we estimate the required CPU time for computing a solution of the problem having fixed the regularization parameter λ, for $\rho = 0.5$ and $n = 10^4, 5\ 10^4, 10^5, 5\ 10^5, 10^6$. In Table 2 we show the computational time by reporting the mean time in seconds to compute a solution of the regularization path. From Table 2 the benefit in terms of computational efficiency provided by the use of the APRiL method with respect to the AGLM method is evident. Indeed, in each case the computational cost is shrunk by a factor of about 15. In addition, another advantage of the proposed method is that it does not suffer of convergence issues which are instead well-known in the case of the Poisson regression (Marschner and others 2011; Silva and Tenreyro 2011).

4.2 Sparse signal recovery application

We present two simulated experiments in sparse signal recovery: the first is an example of image denoising and the second is an example of image deblurring. Formally, these problems are described by equation (17) where $\mathbf{X} := \Omega\Psi$ where Ω represents the convolution with a given point-spread-function and Ψ is the standard synthesis operator which decomposes a given image f on an orthogonal wavelet basis $\{\psi_j\}_{j\in\{1,\dots,p\}}$. The image to recover is characterized by coefficients denoted by $(\beta_j^*)_{j\in\{1,\dots,p\}}$, i.e.

$$f^* = \sum_{j=1}^{p} \beta_j^* \psi_j. \tag{20}$$

In both cases we consider 256×256 images leading to large scale inverse problems with size $n = 65536$. For the denoising application we generate a compressed version of the 'lena' image by thresholding its coefficients in the wavelet basis and we use the resulting image as the 'true' image to recover. The true image is then represented by 17368 non-zero coefficients in the wavelet basis (about 74% of sparsity) with a Relative Square Error (RSE) of about 0.001% with respect to the original image. In this case the operator Ω is the identity. For the deblurring application we used a medical image and we performed the above described procedure for obtaining a 'true' image represented by 10005 non-zero coefficients (about 85% of sparsity) corresponding to a RSE value of about 0.003% with respect to its original version. The convolution kernel of the operator Ω is a Gaussian function with $\sigma = 1.5$. We apply APRiL method to both problems. Thanks to its particular form, we can solve optimization problems by using an iterative forward-backward splitting algorithm: we perform a gradient step with step-size $\tau = 1.5$ and then we apply the soft thresholding operator in the wavelet domain. Iteration stops when convergence is reached. The numerical optimization has been performed by using the MATLAB Numerical Tours Peyré (2011).



(a) Lena (b) Noisy

(c) Recovered (d) SNR

Fig. 4 Image denoising application: (a) true object, (b) noisy image, (c) recovered image with APRiL method, (d) SNR as a function of λ.

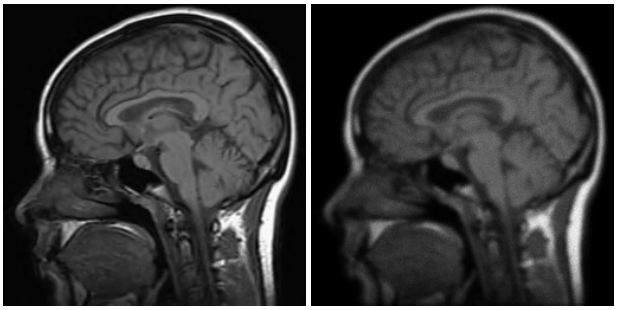
In both examples we select the regularization parameter in order to maximize the Signal-to-Noise Ratio (SNR).

Table 3 Recovery performance results for image denoising and deblurring applications, respectively.

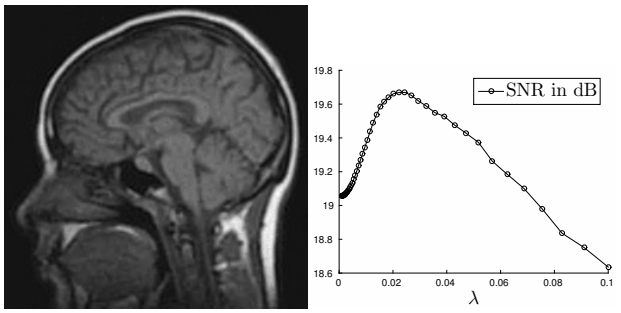
Example	RSE	SNR in dB	PSNR in dB	confusion matrix	
denoising	0.05 %	21.99	28.52	TP = 5668	FN = 11700
				FP = 2257	TN = 45911
deblurring	0.08 %	19.67	31.66	TP = 5420	FN = 4585
				FP = 2684	TN = 52847

Figure 4 and Figure 5 show the results in the case of denoising and deblurring problems, respectively: for each example we show the true image, the noisy image, the best recovered image with APRiL method and the SNR of the recovered images as a function of the regularization parameter. In Table 3 we show the following performance values: the RSE, the SNR and the Peak SNR. In addition, in such Table we provide for each problem the confusion matrix showing how many wavelet coefficients have been correctly recovered.

In both imaging applications we can notice a high number of TN, whereas a relatively small number of FP and a quite large number of FN.



(a) MRI (b) Blurred and noisy



(c) Recovered (d) SNR

Fig. 5 Image deblurring application: (a) true object, (b) blurred and noisy image, (c) recovered image with APRiL method, (d) SNR as a function λ.

However most of such incorrectly estimated coefficients have very small absolute value: indeed they do not significantly contribute to the signal formation.

5 Conclusions

In this paper we proposed to use a globally quadratic approximation of the Kullback-Leibler divergence for performing adaptive ℓ_1-penalized Poisson regression. The gain of this approach, called APRiL, is to perform consistent model selection alongside a reduced computational cost deriving from the quadratic approximation. Indeed, APRiL enjoys the computational advantages of standard Lasso by means of a simple reweighting of the input matrix and data. On the other hand, we proved that the method provides consistent model selection under some unrestrictive assumptions on the regularization parameter λ as a function of the number of samples n. Although we proved that the model selection is ensured in a wide asymptotic range ($\lambda^{(n)}$ can have infinite or finite limit value, even zero), we showed in simulations that the selection of λ by Cross Validation is not effective in the case of small sized samples. We also showed that this method is efficient in the case of very large sized dataset, as in the case of signal processing. A data-dependent choice of the regularization parameter in according with Theorem 3 is an intriguing problem and will be object of future work.

6 Appendix: proofs

In order to prove Theorem 1, we start by proving the following

Lemma 1 *Let y be a Poisson random variable with mean θ. Let $z > 0$ be such that $|z-\theta| \leq c\sqrt{\theta}$, where c is a positive constant smaller than $\sqrt{\theta}$. Then*

$$\mathbb{E}\left(D(z,y) - \frac{1}{2}\frac{(y-z)^2}{z}\right) = O\left(\frac{1}{\theta}\right), \text{ as } \theta \to \infty. \tag{21}$$

Proof Following Zanella et al (2013) we obtain

$$D(z,k) = \frac{1}{2}\frac{(k-z)^2}{z} - \frac{1}{6}\frac{(k-z)^3}{z^2} + \frac{1}{3}\frac{(k-z)^4}{z^3} + kR_3\left(\frac{k-z}{z}\right),$$

where $k \in \mathbb{N}$ and R_3 is defined as follows

$$R_3(\xi) = \int_0^{\xi} \frac{(t-\xi)^3}{(1+t)^4} dt,$$

where $\xi \geq -1$. By computing the moments of the Poisson random variable y centered in z we obtain

$$\mathbb{E}\left(D(z,y) - \frac{1}{2}\frac{(y-z)^2}{z}\right) = -\frac{1}{6}\frac{\theta - 3\theta r - r^3}{z^2} + \frac{1}{3}\frac{3\theta^2 + 6\theta r^2 - 4\theta r + \theta + r^4}{z^3} + \mathscr{E}(\theta), \tag{22}$$

where $r := z - \theta$ and

$$\mathscr{E}(\theta) = \mathbb{E}\left(yR_3\left(\frac{y-z}{z}\right)\right) = \sum_{k=1}^{\infty} \frac{e^{-\theta}\theta^k}{k!} kR_3\left(\frac{k-z}{z}\right).$$

To conclude we now prove that $\mathscr{E}(\theta) = O(\frac{1}{\theta})$. Following the idea of the proof given in Zanella et al (2013), we split the series into two parts: in the first ranging k between 0 and $[\frac{z}{2}]$ and in the second one $k > [\frac{z}{2}] + 1$, where $[\chi]$ denotes the integer part of χ. We observe that for k from 1 to $[\frac{z}{2}]$, or equivalently $\xi \in (-1, -\frac{1}{2}]$, then

$$(1+\xi)|R_3(\xi)| \leq \frac{1}{e}. \tag{23}$$

Since $\frac{\theta^s}{s!} = \frac{\theta^s}{\Gamma(s+1)}$ is monotonically increasing for $0 \leq s \leq [\frac{\theta + c\sqrt{\theta}}{2}]$, using equation (23) and the Stirling formula we obtain

$$\begin{aligned}
&\left|\sum_{k=1}^{[\frac{z}{2}]} \frac{e^{-\theta}\theta^k}{k!} kR_3\left(\frac{k-z}{z}\right)\right| \leq \frac{1}{e}\sum_{k=1}^{[\frac{z}{2}]} ze^{-\theta}\frac{\theta^k}{k!} \\
&\leq \frac{1}{e}\sum_{k=1}^{[\frac{\theta+c\sqrt{\theta}}{2}]} ze^{-\theta}\frac{\theta^k}{k!} \leq \frac{1}{e}\left[\frac{\theta + c\sqrt{\theta}}{2}\right] e^{-\theta}\frac{\theta^{[\frac{\theta+c\sqrt{\theta}}{2}]}}{[\frac{\theta+c\sqrt{\theta}}{2}]!} z \\
&\leq \frac{e^{-\theta - 1 + [\frac{\theta+c\sqrt{\theta}}{2}]}}{\sqrt{2\pi}} \left(\frac{\theta}{[\frac{\theta+c\sqrt{\theta}}{2}]}\right)^{[\frac{\theta+c\sqrt{\theta}}{2}]} \left[\frac{\theta + c\sqrt{\theta}}{2}\right]^{\frac{1}{2}} z \\
&\leq \frac{e^{-\frac{1}{2}\theta\left(1 - c\theta^{-\frac{1}{2}} - v\left(\log\left(\frac{2}{v - 2\theta^{-1}}\right)\right)\right)}(\theta + c\sqrt{\theta})^{\frac{3}{2}}}{2\sqrt{\pi}} \\
&=: M(\theta),
\end{aligned} \tag{24}$$

where $v := 1 + c\theta^{-\frac{1}{2}}$. Then $M(\theta) \to 0$ exponentially as $\theta \to \infty$. Now we consider $k \leq [\frac{z}{2}] + 1$, or equivalently $\xi > -\frac{1}{2}$. Since

$$|R_3(\xi)| \leq 4\xi^4, \tag{25}$$

we obtain

$$\begin{aligned}
&\left|\sum_{k=[\frac{z}{2}]+1}^{\infty} \frac{e^{-\theta}\theta^k}{k!} kR_3\left(\frac{k-z}{z}\right)\right| \\
&\leq 4\sum_{k=0}^{\infty} \frac{e^{-\theta}\theta^k}{k!} k\left(\frac{k-z}{z}\right)^4 \\
&= 4\frac{3\theta^3 + 6\theta^2 r^2 - 16\theta^2 r + 11\theta^2 + (r-1)^4}{z^4} \\
&\leq \frac{(3+6c^2)\theta^3 + 16c\theta^2\sqrt{\theta} + 11\theta^2 + (c\sqrt{\theta}+1)^4}{(\theta - c\sqrt{\theta})^4} \\
&= O\left(\frac{1}{\theta}\right).
\end{aligned}$$

Proof (Proof of Theorem 1) By the triangular inequality we have

$$\begin{aligned}
&\left|\mathbb{E}\left(D(z,y) - \frac{1}{2}\frac{(y-z)^2}{y+1}\right)\right| \\
&\leq \left|\mathbb{E}\left(D(z,y) - \frac{1}{2}\frac{(y-z)^2}{z}\right)\right| \\
&+ \left|\mathbb{E}\left(\frac{1}{2}\frac{(y-z)^2}{z} - \frac{1}{2}\frac{(y-z)^2}{y+1}\right)\right|.
\end{aligned} \tag{26}$$

Then, to get the thesis, thanks to Lemma 1, it is sufficient to prove that

$$\mathbb{E}\left(\frac{(y-z)^2}{z} - \frac{(y-z)^2}{y+1}\right) = O\left(\frac{1}{\sqrt{\theta}}\right), \text{ as } \theta \to \infty. \tag{27}$$

By straightforward computations and using that $|z - \theta| \leq c\sqrt{\theta}$ we get the following

$$\begin{aligned}
&\left|\mathbb{E}\left(\frac{(y-z)^2}{z} - \frac{(y-z)^2}{y+1}\right)\right| \\
&\leq e^{-\theta}\left|\frac{(z+1)^2}{\theta}\right| + \left|\frac{(\theta - z)^3}{z\theta}\right| \\
&+ \left|\frac{(\theta - z)^2}{z\theta}\right| + \left|\frac{3(\theta - z)}{\theta} - \frac{1}{\theta}\right| \\
&\leq e^{-\theta}\frac{(\theta + c\sqrt{\theta} + 1)^2}{\theta} + \frac{c^3}{\sqrt{\theta} - c} \\
&+ \frac{c^2}{\theta - c\sqrt{\theta}} + \frac{3c}{\sqrt{\theta}} + \frac{1}{\theta} = O\left(\frac{1}{\sqrt{\theta}}\right).
\end{aligned}$$

To prove Theorem 2 we need some preliminary results. We start by defining

$$\varepsilon := \mathbf{Y} - \mathbf{X}\beta^*. \tag{28}$$

We observe that the components ε_i are independent random variables with zero mean and $\mathrm{Var}(\varepsilon_i) = (\mathbf{X}\beta^*)_i$, for all $i \in \{1, \ldots, n\}$. Hereafter, for easy of notation we suppress the superscript (n) from the estimators.

Lemma 2 *There exists a constant $G < +\infty$ such that*

$$\mathbb{E}\left(\|\mathbf{X}^T\Lambda^2\varepsilon\|_2^4\right) \leq p^2G^2(\tau_{\max}(\boldsymbol{X}^T\boldsymbol{X}))^2. \tag{29}$$

Proof (Proof of Lemma 2) We compute the term in the l.h.s. in equation (29). We have

$$\mathbb{E}\left(\|\mathbf{X}^T\Lambda^2\varepsilon\|_2^4\right) = \mathbb{E}((\mathbf{D}^T\mathbf{X}\mathbf{X}^T\mathbf{D})^2) \tag{30}$$

with $\mathbf{D} := \Lambda^2\varepsilon$. For the Singular Value Decomposition we can write

$$\mathbf{X}\mathbf{X}^T = \mathbf{U}^T\Sigma\mathbf{U}\,,$$

where $\mathbf{U}$ is an orthogonal matrix and Σ a diagonal matrix containing the eigenvalues of $\mathbf{X}^T\mathbf{X}$. We define We define $\mathbf{H} := \mathbf{U}\mathbf{D} = \mathbf{U}\Lambda^2\varepsilon$. The i-th component of $\mathbf{H}$ is given by

$$H_i = \sum_{l=1}^{n} u_{il}\frac{\varepsilon_l}{Y_l+1}\,,$$

where u_{il} represents the (i,l)-entry of the matrix $\mathbf{U}$. Since $\frac{\varepsilon_l}{Y_l+1}$ takes values between $-(\mathbf{X}\beta^*)_l$ and 1, we have that each component H_i takes values in a compact subset $[R,S]$. Therefore, as $\mathbf{H} \in [R,S]^n$, the quadratic form $(\mathbf{H}^T\Sigma\mathbf{H})^2$ admits a maximum, i.e. there exists an $\mathbf{H}^*$ such that $(\mathbf{H}^T\Sigma^2\mathbf{H})^2 \leq ((\mathbf{H}^*)^T\Sigma^2\mathbf{H}^*)^2$. Then

$$\begin{aligned}\mathbb{E}\left(\|\mathbf{X}^T\Lambda^2\varepsilon\|_2^4\right) &= \mathbb{E}((\mathbf{H}^T\Sigma\mathbf{H})^2) \leq ((\mathbf{H}^*)^T\Sigma^2\mathbf{H}^*)^2 \\ &\leq p^2G^2(\tau_{\max}(\mathbf{X}^T\mathbf{X}))^2,\end{aligned} \tag{31}$$

with

$$G := \max_{\substack{i\in\{1,\dots,n\}:\\ \Sigma_{ii}\neq 0}} (H_i^*)^2. \tag{32}$$

Corollary 2 *Under assumption (H1) there exists a constant $G < +\infty$ such that we have the following bound*

$$\mathbb{E}(\|\hat{\beta}(\text{PRLS}) - \beta^*\|_2^2) \leq \frac{pGBn}{(bn)^2}, \tag{33}$$

where $\hat{\beta}(\text{PRLS})$ is the reweighted least square estimator defined as follows

$$\hat{\beta}(\text{PRLS}) = \arg\min_{\beta} \frac{1}{2}\|\Lambda(\mathbf{Y} - \mathbf{X}\beta)\|_2^2\,. \tag{34}$$

Proof (Proof of Corollary 2) By using optimality conditions of problem in equation (34) and the definition of ε we have

$$\hat{\beta}(\text{PRLS}) - \beta^* = (\mathbf{X}^T\Lambda^2\mathbf{X})^{-1}(\mathbf{X}^T\Lambda^2\varepsilon)\,. \tag{35}$$

Then, by using the Cauchy-Schwartz inequality we obtain

$$\begin{aligned}&\mathbb{E}(\|\hat{\beta}(\text{PRLS}) - \beta^*\|_2^2) \\ &\leq \sqrt{\mathbb{E}(\|(\mathbf{X}^T\Lambda^2\mathbf{X})^{-1}\|_2^4)\mathbb{E}(\|\mathbf{X}^T\Lambda^2\varepsilon\|_2^4)},\end{aligned} \tag{36}$$

where

$$\|(\mathbf{X}^T\Lambda^2\mathbf{X})^{-1}\|_2^4 = \frac{1}{(\tau_{\min}(\mathbf{X}^T\Lambda^2\mathbf{X}))^4}\,. \tag{37}$$

By assumption (H1) and Lemma 2 we have the thesis.

Proof (Proof of Theorem 2) We want to prove the bound in equation (12). From Corollary 2, since

$$\begin{aligned}\mathbb{E}(\|\hat{\beta}_{(\mathbf{w},\lambda)} - \beta^*\|_2^2) &\leq \mathbb{E}(\|\hat{\beta}_{(\mathbf{w},\lambda)} - \hat{\beta}(\text{PRLS})\|_2^2) \\ &+ \mathbb{E}(\|\hat{\beta}(\text{PRLS}) - \beta^*\|_2^2),\end{aligned} \tag{38}$$

we have to establish a bound for the first term of the r.h.s. of (38). In order to do so, we follow similar arguments as in the proof of Theorem 3.1 by Zou and Zhang (2009). By definition of $\hat{\beta}_{(\mathbf{w},\lambda)}$, the following inequality applies

$$\begin{aligned}&\frac{1}{2}\|\Lambda(\mathbf{Y} - \mathbf{X}\hat{\beta}_{(\mathbf{w},\lambda)})\|_2^2 - \frac{1}{2}\|\Lambda(\mathbf{Y} - \mathbf{X}\hat{\beta}(\text{PRLS}))\|_2^2 \\ &\leq \lambda\sum_{j=1}^{p} w_j(|\hat{\beta}(\text{PRLS})_j| - |(\hat{\beta}_{(\mathbf{w},\lambda)})_j|).\end{aligned} \tag{39}$$

From the optimality conditions of the optimization problem in equation (34), we have

$$\begin{aligned}&\frac{1}{2}\|\Lambda(\mathbf{Y} - \mathbf{X}\hat{\beta}_{(\mathbf{w},\lambda)})\|_2^2 - \frac{1}{2}\|\Lambda(\mathbf{Y} - \mathbf{X}\hat{\beta}(\text{PRLS}))\|_2^2 \\ &= \frac{1}{2}(\hat{\beta}_{(\mathbf{w},\lambda)} - \hat{\beta}(\text{PRLS}))^T\mathbf{X}^T\Lambda^2\mathbf{X}(\hat{\beta}_{(\mathbf{w},\lambda)} - \hat{\beta}(\text{PRLS})),\end{aligned} \tag{40}$$

and we notice that

$$\begin{aligned}&\tau_{\min}(\mathbf{X}^T\Lambda^2\mathbf{X})\|\hat{\beta}_{(\mathbf{w},\lambda)} - \hat{\beta}(\text{PRLS})\|_2^2 \\ &\leq (\hat{\beta}_{(\mathbf{w},\lambda)} - \hat{\beta}(\text{PRLS}))^T\mathbf{X}^T\Lambda^2\mathbf{X}(\hat{\beta}_{(\mathbf{w},\lambda)} - \hat{\beta}(\text{PRLS}))\end{aligned} \tag{41}$$

and

$$\begin{aligned}&\sum_{j=1}^{p} w_j(|\hat{\beta}(\text{PRLS})_j| - |(\hat{\beta}_{(\mathbf{w},\lambda)})_j|) \\ &\leq \sqrt{\sum_{j=1}^{p} w_j^2}\|\hat{\beta}(\text{PRLS}) - \hat{\beta}_{(\mathbf{w},\lambda)}\|_2.\end{aligned} \tag{42}$$

Using (39), (40), (41) and (42) we obtain

$$\|\hat{\beta}_{(\mathbf{w},\lambda)} - \hat{\beta}(\text{PRLS})\|_2 \leq \frac{2\lambda\sqrt{\sum_{j=1}^{p} w_j^2}}{\tau_{\min}(\mathbf{X}^T\Lambda^2\mathbf{X})}, \tag{43}$$

and finally, the Cauchy Schwartz inequality and assumption (H1) lead to

$$\mathbb{E}(\|\hat{\beta}_{(\mathbf{w},\lambda)} - \hat{\beta}(\text{PRLS})\|_2^2) \leq \frac{4\lambda^2\sqrt{\mathbb{E}\left(\left(\sum_{j=1}^{p} w_j^2\right)^2\right)}}{(bn)^2}. \tag{44}$$

The thesis follows from equations (38), (44) and Corollary 2.

Proof (Proof of Theorem 3) For brevity we denote the APRiL estimator by $\hat{\beta}$. To prove the model selection consistency we prove that for $n \to +\infty$

$$\mathbb{P}(\forall j \in (\mathscr{A}^*)^C, \hat{\beta}_j = 0) \longrightarrow 1 \tag{45}$$

and

$$\mathbb{P}(\forall j \in \mathscr{A}^*, |\hat{\beta}_j| > 0) \longrightarrow 1\,. \tag{46}$$

We now prove equation (45). The functional defined in equation (8) is convex and not differentiable and $\mathscr{C}$ is convex set. Then the solution $\hat{\beta}$ satisfies the (KKT) optimality conditions:

- $(\mathbf{X}\hat{\beta})_i \geq 0 \; \forall \, i \in \{1,\dots,n\}$;
- $v_i \geq 0 \; \forall \, i \in \{1,\dots,n\}$;
- $v_i(\mathbf{X}\hat{\beta})_i = 0 \; \forall \, i \in \{1,\dots,n\}$;
- if $\hat{\beta}_j \neq 0$

$$-\mathbf{x}_j^T \Lambda^2(\mathbf{Y} - \mathbf{X}\hat{\beta}) + \lambda w_j sgn(\hat{\beta}_j) - \mathbf{x}_j^T v = 0; \tag{47}$$

- if $\hat{\beta}_j = 0$

$$-\mathbf{x}_j^T \Lambda^2(\mathbf{Y} - \mathbf{X}\hat{\beta}) + \lambda w_j s_j - \mathbf{x}_j^T v = 0, \tag{48}$$

with $s_j \in [-1,1]$.

v is the n-dimensional vector whose components are the Lagrangian multipliers. From equation (47), the event

$$\left\{\forall j \in (\mathscr{A}^*)^C, \hat{\beta}_j = 0\right\}$$

is the same of

$$\left\{|\mathbf{x}_j^T(\Lambda^2(\mathbf{Y} - \mathbf{X}_{\mathscr{A}^*}\hat{\beta}_{\mathscr{A}^*}) + v)| \leq \lambda \hat{w}_j, \forall j \in (\mathscr{A}^*)^C\right\}, \tag{49}$$

where $\mathbf{X}_{\mathscr{A}^*}$ is the matrix constituted by the columns $\mathbf{x}_j$ with $j \in \mathscr{A}^*$, and $\hat{\beta}_{\mathscr{A}^*}$ is the vector constituted by the components $\hat{\beta}_j$ with $j \in \mathscr{A}^*$. Equation (45) is equivalent to $\mathbb{P}\left(\exists j \in (\mathscr{A}^*)^C, \left|\mathbf{x}_j^T(\Lambda^2(\mathbf{Y} - \mathbf{X}_{\mathscr{A}^*}\hat{\beta}_{\mathscr{A}^*}) + v)\right| > \lambda \hat{w}_j\right) \to 0$ for $n \to +\infty$. We set

$$\hat{S}_j := |\hat{\beta}(\mathrm{PRiL})_j| + \left(\frac{1}{n}\right)^{\frac{1}{\gamma}+\delta},$$

$$\hat{\eta} := \min_{j \in \mathscr{A}^*} \hat{S}_j,$$

$$\eta := \min_{j \in \mathscr{A}^*} |\beta_j^*| + \left(\frac{1}{n}\right)^{\frac{1}{\gamma}+\delta},$$

and

$$\hat{E}_j := \left|\mathbf{x}_j^T(\Lambda^2(\mathbf{Y} - \mathbf{X}_{\mathscr{A}^*}\hat{\beta}_{\mathscr{A}^*}) + v)\right|.$$

Then

$$\begin{aligned}
&\mathbb{P}\left(\exists j \in (\mathscr{A}^*)^C \hat{E}_j > \lambda \hat{w}_j\right) \\
&\leq \textstyle\sum_{j \in (\mathscr{A}^*)^C} \mathbb{P}\left(\hat{E}_j > \lambda \hat{w}_j, \hat{\eta} > \frac{\eta}{2}, \hat{S}_j \leq \left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}\right) \\
&\quad + \textstyle\sum_{j \in (\mathscr{A}^*)^C} \mathbb{P}\left(\hat{S}_j > \left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}\right) + \mathbb{P}\left(\hat{\eta} \leq \frac{\eta}{2}\right).
\end{aligned} \tag{50}$$

The idea is to determine three bounds M_1, M_2 and M_3 depending on n, such that

$$\mathbb{P}\left(\hat{\eta} \leq \frac{\eta}{2}\right) \leq M_1, \tag{51}$$

$$\sum_{j \in (\mathscr{A}^*)^C} \mathbb{P}\left(\hat{S}_j > \left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}\right) \leq M_2, \tag{52}$$

$$\sum_{j \in (\mathscr{A}^*)^C} \mathbb{P}\left(\hat{E}_j > \lambda \hat{w}_j, \hat{\eta} > \frac{\eta}{2}, \hat{S}_j \leq \left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}\right) \leq M_3. \tag{53}$$

and M_1, M_2 and M_3 go to 0 for $n \to +\infty$. Let us start with the determination of bound M_1. Using Corollary 1, it follows that

$$\begin{aligned}
\mathbb{P}\left(\hat{\eta} \leq \frac{\eta}{2}\right) &\leq \mathbb{P}\left(\|\hat{\beta}(\mathrm{PRiL}) - \beta^*\|_2 \geq \frac{\eta}{2}\right) \\
&\leq \frac{2}{\eta}\left(\frac{2\lambda_1\sqrt{p} + \sqrt{pGBn}}{bn}\right) =: M_1.
\end{aligned}$$

For the determination of bound M_2, we use again Corollary 1. We have that

$$\begin{aligned}
&\sum_{j \in (\mathscr{A}^*)^C} \mathbb{P}\left(\hat{S}_j > \left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}\right) \\
&\leq \frac{\mathbb{E}\left(\|\hat{\beta}(\mathrm{PRiL}) - \beta^*\|_2\right) + p\left(\frac{1}{n}\right)^{\frac{1}{\gamma}+\delta}}{\left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}} \\
&\leq \frac{1}{(\frac{\lambda}{n})^{\frac{1}{\gamma}}}\left(\frac{2\lambda_1\sqrt{p} + \sqrt{pGBn}}{bn} + \frac{p}{n^{\frac{1}{\gamma}+\delta}}\right) =: M_2.
\end{aligned}$$

Finally, for the determination of bound M_3, we write

$$\begin{aligned}
&\textstyle\sum_{j \in (\mathscr{A}^*)^C} \mathbb{P}\left(\hat{E}_j > \lambda \hat{w}_j, \hat{\eta} > \frac{\eta}{2}, \hat{S}_j \leq \left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}\right) \\
&\leq 2\frac{\mathbb{E}\left(\sum_{j \in (\mathscr{A}^*)^C} \hat{E}_j \mathbf{1}_{\left\{\hat{\eta} > \frac{\eta}{2}\right\}}\right)}{n}.
\end{aligned}$$

By definition of ε, we have

$$\begin{aligned}
&\textstyle\sum_{j \in (\mathscr{A}^*)^C} \hat{E}_j \\
&= \textstyle\sum_{j \in (\mathscr{A}^*)^C} \left|\mathbf{x}_j^T \Lambda^2 \mathbf{X}_{\mathscr{A}^*}(\beta^*_{\mathscr{A}^*} - \hat{\beta}_{\mathscr{A}^*}) + \mathbf{x}_j^T \Lambda^2 \varepsilon + \mathbf{x}_j^T v\right| \\
&\leq \textstyle\sum_{j \in (\mathscr{A}^*)^C} \|\mathbf{x}_j^T \Lambda\|_2 \sqrt{\tau_{\max}(\mathbf{X}^T \Lambda^2 \mathbf{X})} \|\beta^*_{\mathscr{A}^*} - \hat{\beta}_{\mathscr{A}^*}\|_2 \\
&\quad + \textstyle\sum_{j \in (\mathscr{A}^*)^C} \left|\mathbf{x}_j^T v\right| + \sum_{j \in (\mathscr{A}^*)^C} \|\mathbf{x}_j^T \Lambda\|_2 \|\Lambda \varepsilon\|_2.
\end{aligned} \tag{54}$$

By using assumption (H4), we get

$$\sum_{j \in (\mathscr{A}^*)^C} \|\mathbf{x}_j^T \Lambda\|_2 \leq p \max_{j=1,\dots,p} \|\mathbf{x}_j\|_2 \leq pL, \tag{55}$$

and

$$\mathbb{E}(\|\Lambda\varepsilon\|_2) = \mathbb{E}\left(\sqrt{\sum_{i=1}^{n}\left(\frac{\varepsilon_i}{\sqrt{Y_i+1}}\right)^2}\right) \leq \sqrt{2n} \tag{56}$$

where we used that $\mathbb{E}\left(\frac{\varepsilon_i^2}{Y_i+1}\right) \leq 2$ for all $i \in \{1,\dots,n\}$. Following the idea of the proof of Lemma 2, in particular the calculus which leads (33), we have

$$\|\hat{\beta}_{\mathscr{A}^*} - \hat{\beta}(\mathrm{PRLS})_{\mathscr{A}^*}\|_2 \leq \frac{2\lambda\sqrt{p}\frac{1}{\hat{\eta}^\gamma}}{\tau_{\min}(\mathbf{X}^T_{\mathscr{A}^*}\Lambda^2\mathbf{X}_{\mathscr{A}^*})}. \tag{57}$$

Thanks to the Cauchy-Schwartz inequality, equations (54), (55), (56), (57) and hypothesis (H1) we obtain the following bound

$$\begin{aligned}
&\mathbb{E}\left(\sum_{j\in(\mathscr{A}^*)^C}\hat{E}_j\mathbf{1}_{\left\{\hat{\eta}>\frac{\eta}{2}\right\}}\right)\\
&\leq pL\sqrt{\mathbb{E}\left(\tau_{\max}(\mathbf{X}^T\Lambda^2\mathbf{X})\right)\mathbb{E}\left(\|\beta^*_{\mathscr{A}^*}-\hat{\beta}_{\mathscr{A}^*}\|_2^2\mathbf{1}_{\left\{\hat{\eta}>\frac{\eta}{2}\right\}}\right)}\\
&+\mathbb{E}\left(\sum_{j\in(\mathscr{A}^*)^C}\left|\mathbf{x}_j^T\nu\right|\mathbf{1}_{\left\{\hat{\eta}>\frac{\eta}{2}\right\}}\right)+pL\mathbb{E}(\|\Lambda\varepsilon\|_2)\\
&\leq pL\sqrt{Bn}\left(\frac{2\lambda\sqrt{p}\left(\frac{\eta}{2}\right)^{-\gamma}+\sqrt{pGBn}}{bn}\right)\\
&+\mathbb{E}\left(\sum_{j\in(\mathscr{A}^*)^C}\left|\mathbf{x}_j^T\nu\right|\mathbf{1}_{\left\{\hat{\eta}>\frac{\eta}{2}\right\}}\right)+pL\sqrt{2n}.
\end{aligned} \tag{58}$$

From optimality conditions in equations (48) and (47) it follows that

$$|\mathbf{x}_j^T\nu| \leq |\mathbf{x}_j^T\Lambda^2(\mathbf{Y}-\mathbf{X}\hat{\beta})|+\lambda w_j,$$

so we have

$$\begin{aligned}
&\mathbb{E}\left(\sum_{j\in(\mathscr{A}^*)^C}\left|\mathbf{x}_j^T\nu\right|\mathbf{1}_{\left\{\hat{\eta}>\frac{\eta}{2}\right\}}\right)\\
&\leq \mathbb{E}\left(\sum_{j\in\mathscr{A}^*}|\mathbf{x}_j^T\Lambda^2\varepsilon|\right)\\
&+\mathbb{E}\left(\sum_{j\in\mathscr{A}^*}|\mathbf{x}_j^T\Lambda^2\mathbf{X}(\beta^*-\hat{\beta})|\mathbf{1}_{\left\{\hat{\eta}>\frac{\eta}{2}\right\}}\right)\\
&+\lambda\mathbb{E}\left(\sum_{j\in\mathscr{A}^*}\hat{w}_j\mathbf{1}_{\left\{\hat{\eta}>\frac{\eta}{2}\right\}}\right)\\
&\leq pL\sqrt{2n}+pL\sqrt{Bn}\left(\frac{2\lambda\sqrt{p}n^{1+\gamma\delta}+\sqrt{pGBn}}{bn}\right)+\lambda p\left(\frac{2}{\eta}\right)^\gamma
\end{aligned}$$

where we have used the following bound

$$\begin{aligned}
&\mathbb{E}\left(\left(\sum_{j=1}^p\hat{w}_j^2\right)^2\right)\\
&=\mathbb{E}\left(\left(\sum_{j=1}^p\frac{1}{\left(|\hat{\beta}(\mathrm{PRiL})_j|+\left(\frac{1}{n}\right)^{\frac{1}{\gamma}+\delta}\right)^{2\gamma}}\right)^2\right)\\
&\leq p^2n^{4(1+\delta\gamma)}.
\end{aligned} \tag{59}$$

Then, we obtain

$$\begin{aligned}
&\sum_{j\in(\mathscr{A}^*)^C}\mathbb{P}\left(\hat{E}_j>\lambda\hat{w}_j,\hat{\eta}>\frac{\eta}{2},\hat{S}_j\leq\left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}\right)\\
&\leq\frac{4pL}{n}\left(\sqrt{2n}+B\sqrt{pG}+\sqrt{Bnp}\frac{\lambda\left(\frac{2}{\eta}\right)^\gamma+\lambda n^{1+\delta\gamma}}{bn}+\lambda\frac{2^{\gamma-1}}{\eta^\gamma L}\right)\\
&=: M_3.
\end{aligned}$$

Now we prove that M_1, M_2 and M_3 go to 0 for $n\to+\infty$.

$M_3\to 0$

because $\frac{\sqrt{n}\lambda\left(\frac{2}{\eta}\right)^\gamma}{n^2}\longrightarrow 0$, $\frac{\sqrt{n}\lambda n^{1+\delta\gamma}}{n^2}\longrightarrow 0$ and $\frac{\lambda}{n}\left(\frac{2}{\eta}\right)^\gamma\longrightarrow 0$ for $n\to+\infty$, for the assumption (H3 c) and for the positivity of constants γ and δ;

$$M_2=\frac{1}{\left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}}\left(\frac{2\lambda_1\sqrt{p}+\sqrt{pGBn}}{bn}+\frac{p}{n^{\frac{1}{\gamma}+\delta}}\right)\to 0$$

because $\frac{\lambda_1}{\left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}n}\longrightarrow 0$ for $n\to+\infty$ for assumptions (H2) and (H3 a), $\frac{\sqrt{n}}{n\left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}}=\frac{1}{\left(\lambda n^{\frac{\gamma}{2}-1}\right)^{\frac{1}{\gamma}}}\longrightarrow 0$ for $n\to+\infty$ for the assumption (H3 a), and at last $\frac{1}{\left(\frac{\lambda}{n}\right)^{\frac{1}{\gamma}}n^{\frac{1}{\gamma}+\delta}}\longrightarrow 0$ for $n\to+\infty$ for the assumption (H3 b);

$$M_1=\frac{2}{\eta}\left(\frac{2\lambda_1\sqrt{p}+\sqrt{pGBn}}{bn}\right)\to 0$$

because $\frac{\lambda_1}{n\eta}\longrightarrow 0$ for $n\to+\infty$, for the assumption (H2) and the definition of η, and $\frac{\sqrt{n}}{n\eta}=O(\frac{1}{\sqrt{n}})$ for $n\to+\infty$.

Now we prove equation (46). It is sufficient to show that

$$\mathbb{P}\left(\min_{j\in\mathscr{A}^*}|\hat{\beta}_j|>0\right)\longrightarrow 1,\quad n\to+\infty.$$

By equation (57) we have

$$\min_{j\in\mathscr{A}^*}|\hat{\beta}_j|>\min_{j\in\mathscr{A}^*}|\hat{\beta}(\mathrm{PRLS})_j|-\frac{2\lambda\sqrt{p}\hat{\eta}^{-\gamma}}{\tau_{\min}(\mathbf{X}^T\Lambda^2\mathbf{X})}, \tag{60}$$

where

$$\min_{j\in\mathscr{A}^*}|\hat{\beta}(\mathrm{PRLS})_j|\geq\min_{j\in\mathscr{A}^*}|\beta^*_j|-\|\beta^*_{\mathscr{A}^*}-\hat{\beta}(\mathrm{PRLS})_{\mathscr{A}^*}\|_2. \tag{61}$$

Since $\min_{j\in\mathscr{A}^*}|\beta^*_j|>0$, to conclude we show that $\|\beta^*_{\mathscr{A}^*}-\hat{\beta}(\mathrm{PRLS})_{\mathscr{A}^*}\|_2$ and $\frac{2\lambda\sqrt{p}\hat{\eta}^{-\gamma}}{\tau_{\min}(\mathbf{X}^T\Lambda^2\mathbf{X})}$ go to 0 in probability. Equation (33) implies that the second term in the r.h.s of

(61) goes to zero. Moreover, for the second term in equation (60) we have that, given $M > 0$

$$\begin{aligned} &\mathbb{P}\left(\frac{\lambda}{\hat{\eta}^{\gamma}\tau_{\min}(\mathbf{X}^T\Lambda^2\mathbf{X})} > M\right) \qquad (62)\\ &\leq \mathbb{P}\left(\frac{\lambda}{\hat{\eta}^{\gamma}\tau_{\min}(\mathbf{X}^T\Lambda^2\mathbf{X})} > M, \{\hat{\eta} > \tfrac{\eta}{2}\}\right) + \mathbb{P}\left(\hat{\eta} \leq \tfrac{\eta}{2}\right)\\ &\leq \frac{\lambda}{M}\sqrt{\mathbb{E}\left(\left(\frac{1}{\tau_{\min}(\mathbf{X}^T\Lambda^2\mathbf{X})}\right)^2\right)\mathbb{E}\left(\frac{1}{\hat{\eta}^{2\gamma}}\mathbf{1}_{\{\hat{\eta}>\frac{\eta}{2}\}}\right)} + M_1\\ &\leq \frac{\lambda}{bnM}\left(\frac{2}{\eta}\right)^{\gamma} + M_1 \longrightarrow 0 \text{ for } n \to +\infty \end{aligned}$$

as $M_1 \to 0$ for $n \to +\infty$ and $\frac{\lambda}{n}\left(\frac{2}{\eta}\right)^{\gamma} \to 0$ for $n \to +\infty$ thanks to assumption (H3 c). This proves equation (46) and concludes the proof.

References

Anscombe FJ (1948) The Transformation of Poisson, Binomial and Negative-Binomial Data. Biometrika 35(3/4):246–254, DOI 10.2307/2332343

Beck A, Teboulle M (2009) A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. SIAM Journal on Imaging Sciences 2(1):183–202, DOI 10.1137/080716542

Bogdan M, van den Berg E, Sabatti C, Su W, Candès EJ (2015) SLOPE-adaptive variable selection via convex optimization. The annals of applied statistics 9(3):1103

Bonettini S, Benvenuto F, Zanella R, Zanni L, Bertero M (2009) Gradient projection approaches for optimization problems in image deblurring and denoising. In: 2009 17th European Signal Processing Conference, pp 1384–1388

Candès EJ, Wakin MB, Boyd SP (2008) Enhancing Sparsity by Reweighted $\ell 1$ Minimization. Journal of Fourier Analysis and Applications 14(5-6):877–905, DOI 10.1007/s00041-008-9045-x

De Vito E, Rosasco L, Caponnetto A, Giovannini UD, Odone F (2005) Learning from examples as an inverse problem. Journal of Machine Learning Research 6(May):883–904

Dobson AJ, Barnett A (2008) An introduction to generalized linear models. CRC press

Efron B, Hastie T, Johnstone I, Tibshirani R, others (2004) Least angle regression. The Annals of statistics 32(2):407–499

Figueiredo MAT, Bioucas-Dias JM (2010) Restoration of Poissonian Images Using Alternating Direction Optimization. IEEE Transactions on Image Processing 19(12):3133–3145, DOI 10.1109/TIP.2010.2053941

Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning, vol 1. Springer series in statistics Springer, Berlin

Friedman J, Hastie T, Hfling H, Tibshirani R, others (2007) Pathwise coordinate optimization. The Annals of Applied Statistics 1(2):302–332

Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33(1), DOI 10.18637/jss.v033.i01

Gu R, Dogandžić A (2014) A fast proximal gradient algorithm for reconstructing nonnegative signals with sparse transform coefficients. In: Signals, Systems and Computers, 2014 48th Asilomar Conference on, IEEE, pp 1662–1667

Hansen NR, Reynaud-Bouret P, Rivoirard V, others (2015) Lasso and probabilistic inequalities for multivariate point processes. Bernoulli 21(1):83–143

Harmany ZT, Marcia RF, Willett RM (2010) SPIRAL out of convexity: Sparsity-regularized algorithms for photon-limited imaging. In: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, pp 75,330R–75,330R

Ivanoff S, Picard F, Rivoirard V (2016) Adaptive Lasso and group-Lasso for functional Poisson regression. Journal of Machine Learning Research 17(55):1–46

Jiang X, Reynaud-Bouret P, Rivoirard V, Sansonnet L, Willett R (2015) A data-dependent weighted LASSO under Poisson noise. arXiv preprint arXiv:150908892

Marschner IC, others (2011) glm2: fitting generalized linear models with convergence problems. The R journal 3(2):12–15

Martinez JG, Carroll RJ, Mller S, Sampson JN, Chatterjee N (2011) Empirical Performance of Cross-Validation With Oracle Methods in a Genomics Context. The American Statistician 65(4):223–228

McCullagh P, Nelder JA (1989) Generalized Linear Models, Second Edition

Peyré G (2011) The Numerical Tours of Signal Processing. Computing in Science & Engineering 13(4):94–97, DOI 10.1109/MCSE.2011.71

Prince JL, Links JM (2006) Medical imaging signals and systems. Pearson Prentice Hall Upper Saddle River, New Jersey

Silva J, Tenreyro S (2011) Poisson: some convergence issues. Stata journal 11(2):207–212

Starck JL, Murtagh F (2007) Astronomical image and data analysis. Springer Science & Business Media

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) pp 267–288

Zanella R, Boccacci P, Zanni L, Bertero M (2013) Corrigendum: efficient gradient projection methods for edge-preserving removal of poisson noise. Inverse Problems 29(11):119,501

Zhao P, Yu B (2006) On Model Selection Consistency of Lasso. J Mach Learn Res 7:2541–2563

Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American statistical association 101(476):1418–1429

Zou H, Zhang HH (2009) On the adaptive elastic-net with a diverging number of parameters. Annals of statistics 37(4):1733