Statistica Sinica **32** (2022), 181-207
doi:https://doi.org/10.5705/ss.202019.0315

ELASTIC-NET REGULARIZED HIGH-DIMENSIONAL NEGATIVE BINOMIAL REGRESSION: CONSISTENCY AND WEAK SIGNAL DETECTION

Huiming Zhang[1,2] and Jinzhu Jia[1]

[1]*Peking University and* [2]*University of Macau*

Abstract: We study a sparse negative binomial regression (NBR) for count data by showing the non-asymptotic advantages of using the elastic-net estimator. Two types of oracle inequalities are derived for the NBR's elastic-net estimates by using the Compatibility Factor Condition and the Stabil Condition. The second type of oracle inequality is for the random design and can be extended to many $\ell_1+\ell_2$ regularized M-estimations, with the corresponding empirical process having stochastic Lipschitz properties. We derive the concentration inequality for the suprema empirical processes for the weighted sum of negative binomial variables to show some high–probability events. We apply the method by showing the sign consistency, provided that the nonzero components in the true sparse vector are larger than a proper choice of the weakest signal detection threshold. In the second application, we show the grouping effect inequality with high probability. Third, under some assumptions for a design matrix, we can recover the true variable set with a high probability if the weakest signal detection threshold is large than the turning parameter up to a known constant. Lastly, we briefly discuss the de-biased elastic-net estimator, and numerical studies are given to support the proposal.

Key words and phrases: De-biased elastic-net, empirical processes, high-dimensional count data regressions, oracle inequalities, sign consistency, stochastic Lipschitz condition.

1. Introduction

In this study, we focus on regression problems involving count data (sometimes called categorical data). The responses are denoted as $\{Y_i\}_{i=1}^n$, each of which follows a univariate discrete distribution. Here, the covariates $\{\boldsymbol{X}_i := (x_{i1},\ldots,x_{ip})^T\}_{i=1}^n \in \mathbb{R}^p$ are supposed to be a deterministic or random variable. If they are random, we can deal with the model by conditioning on design matrix $\mathbf{X} := (\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)^T$. The conditional expectation of $Y_i|\boldsymbol{X}_i^T$ is related to $\boldsymbol{X}_i^T\boldsymbol{\beta}^*$ after a transformation using a link function, where $\boldsymbol{\beta}^* = (\beta_1^*,\ldots,\beta_p^*)^T$ is the unknown true coefficient vector. The Poisson regression is a well-known

Corresponding author: Jinzhu Jia, School of Public Health and Center for Statistical Science, Peking University, 100191, P.R. China. E-mai: jzjia@math.pku.edu.cn.

example. Covariates in a count data regression may take discrete or continuous values. Here, important examples includes logistic regression, Poisson regression and negative binomial regression (NBR), amony others. There are many monographs on statistical models for counting data; see for example, Hilbe (2011) and Tutz (2011).

A commonly used regression model for count data is the Poisson generalized linear model, particularly in the economic, social, and biological sciences, see Tutz (2011). A Poisson regression considers that the response variables Y_i's are nonnegative integers that follow the Poisson distribution, i.e. $P(Y_i = y_i|\lambda_i) = (\lambda_i^{y_i}/y_i!)e^{-\lambda_i}$ for $i = 1, 2, \ldots, n$, where the expectation of Y_i is $\lambda_i := \mathrm{E}(Y_i)$. We require that the positive parameter λ_i be related to a linear combination of p covariate variables. Specifically, the Poisson regression assumes the logarithmic link function $\eta(\lambda_i) =: \log \lambda_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^*$. Owing to the nature of the Poisson distribution, the variance is equal to the expectation: $\mathrm{E}(Y_i \,|\, \boldsymbol{X}_i) = \mathrm{Var}(Y_i \,|\, \boldsymbol{X}_i) = \lambda_i$, called *equidispersion*.

However, in practice, we often encounter *overdispersion*. In this case, the variance of count data is greater than the mean comparing to Poisson count data. For example in RNA-Seq gene expression data, the negative binomial (NB) distribution provides a good choice for modeling a set of count variables and related high-dimensional sets of quantitative or binary variables are of interest, that is $p \gg n$. As evidence of overdispersion, in real data, the variance of the response variable is greater than its mean; see Rauschenberger et al. (2016) and Qiu, Chen and Nettleton (2018). To test whether the variance of count data is greater than the expectation, Cameron and Trivedi (1990) proposed the Cameron–Trivedi test:

$$\mathrm{H}_0\colon \mathrm{Var}(Y_i \,|\, \boldsymbol{X}_i) = \mathrm{E}(Y_i \,|\, \boldsymbol{X}_i) =: \mu_i \quad \text{vs.} \quad \mathrm{H}_1\colon \mathrm{Var}(Y_i \,|\, \boldsymbol{X}_i) = \mu_i + \alpha g(\mu_i),$$

where $g(\mu_i) = \mu_i$ or $g(\mu_i) = \mu_i^2$, and the constant α is the value to be tested. Therefore, the hypothesis test is alternatively written as $\mathrm{H}_0\colon\ \alpha = 0$ vs. $\mathrm{H}_1\colon\ \alpha \neq 0$. For $\alpha \neq 0$, the count data is overdispersed if $\alpha > 0$, and it is underdispersed if $\alpha < 0$. Here the *underdispersion* means that the variance of the data is less than the mean, which suggests that a binomial regression (see Section 3.3.2 of Tutz (2011)) or a COM-Poisson regression (see Sellers and Shmueli (2008)) should be suitable. More details on the overdispersion test can be found in Chapter 7 of Hilbe (2011).

When testing for overdispersion, we have to correct the hypothetical distributions and select a flexible distribution, such as some two-parameter models. A

suggested overdispersed distribution is the negative binomial (NB) distribution that is a particular case of the discrete compound Poisson (DCP) family. NB also belongs to the class of infinitely divisible distribution. For more detailed NB and DCP distributions properties, please refer to Section 5.9.3 of Johnson, Kemp and Kotz (2005) and Zhang, Liu and Li (2014).

In low– and fixed–dimensional regressions with $p < n$, researcher often use the maximum likelihood estimator (MLE) of the regression coefficients. Here, we employ the *average negative log-likelihood function* of the NBR (i.e. a convex empirical process indexed by n):

$$\ell_n(\boldsymbol{\beta}) := -\frac{1}{n}\sum_{i=1}^{n}[Y_i\boldsymbol{X}_i^T\boldsymbol{\beta} - (\theta + Y_i)\log(\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})],\ \boldsymbol{\beta} \in \mathbb{R}^p;$$

see Section 2.1. Here, $\ell_n(\boldsymbol{\beta})$ is also termed the empirical NBR loss function in the field of machine learning point. If θ is given (or treated as a tuning parameter), the NBR actually belongs to the class of generalized linear models (GLMs) with noncanonical links. It should be noted that the coefficient of Y_i in the log-likelihood of a common GLM with a canonical link function is linear in $\boldsymbol{X}_i^T\boldsymbol{\beta}$, whereas the coefficient of Y_i in the log-likelihood of the NBR is nonlinear in $\boldsymbol{X}_i^T\boldsymbol{\beta}$ owing to the noncanonical link function.

In a high-dimensional setting, a powerful tool for remedying the MLE is to add the penalty function to the $\ell_n(\boldsymbol{\beta})$ to get the penalized (regularized) likelihood estimator. Here, we study the elastic-net regularized MLE defined as follows.

Definition 1. (Elastic-net method of NBR) For the empirical NB loss function $\ell_n(\boldsymbol{\beta})$, let $\lambda_1, \lambda_2 > 0$ be tuning parameters. Then, the elastic-net estimates are defined as

$$\hat{\boldsymbol{\beta}} =: \hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\{\ell_n(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\}, \tag{1.1}$$

where $\|\boldsymbol{\beta}\|_q := (\sum_{i=1}^{p}|\beta_i|^q)^{1/q}$ is the l_q-norm of β, for $1 \le q < \infty$.

In the section below, we usually denote $\hat{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$, for simplicity.

Chapter 3 of Tutz (2011) begins with three important criteria for penalized estimation methods for sparse coefficient vectors:

1°. *Existence of unique estimates: this is where MLEs often fail;*

2°. *Prediction accuracy: a model should yield a decent prediction of the outcome;*

3°. *Sparseness and interpretation: a parsimonious model that contains the strongest effects is easier to interpret than a big model with hardly any structure.*

For 3°, as the penalty function, we study the elastic-net estimate because it enjoys the advantages of both the Lasso and the Ridge, see Zou and Hastie (2005). The Lasso can only select one variable in a group of highly related variables, whereas the elastic-net can choose more than one, which we called a grouping effect. For 1° and 2°, we concentrate on the nonasymptotic oracle inequalities of the elastic-net penalized MLE in NB regression because asymptotic distribution of the high-dimensional penalized estimator is usually not available. Essentially, deriving oracle inequalities is a powerful mathematical skill that gives deep insight into an estimator's nonasymptotic fluctuation compared to that of an ideal unknown parameter (the oracle). Wang et al. (2016) compared the NBR and Poisson regression models based on the elastic-net, MCP-net, and SCAD-net penalty functions by using hospitalization days in hospitalized pediatric cardiac surgery and the associated covariates for variable selection analysis. Massaro (2016) constructed the elastic-net penalized NBR to analyze overdispersed count data: time-to-death (in days). Here, the elastic-net selects the genes' functional characteristics that increase or decrease the survival time in the high-dimensional scenario, as $p \gg n$. In practice, the covariates are usually corrupted because they contain unavoidable measurement errors. Sørensen et al. (2018) suggested that elastic-net penalty (or generalized elastic-net penalty with higher-order terms, such as cubic, quadratic terms, etc.) can decorrupt the corrupted covariates in high-dimensional GLMs, by choosing the second tuning parameter in the elastic-net.

Contributions:

- For GLMs, Bunea (2008) investigated the oracle inequalities in the setting of logistic and linear regression models for the elastic-net penalization schemes under the Stabil Condition. By extending the proofs from Bunea (2008), Blazere, Loubes and Gamboa (2014) derived oracle inequalities for GLMs with canonical link functions that do not contain the NBR. The empirical processes technique is used by Blazere, Loubes and Gamboa (2014) to get the oracle inequalities for elastic-net in GLMs; however, their assumption of GLMs does not contain the NBR. Even under a fixed design, the Hessian matrix of the NB log-likelihood contains random responses. This complex phenomenon is substantially different from the canonical link GLMs. Additional treatments for the concentration of a random Hessian matrix are needed. To show the KKT-like event with high probability, we propose a new concentration inequality for the superma of multiplier NB empirical processes.

- van de Geer (2008) mainly studied the oracle inequalities for high-dimensional GLMs with Lipschitz loss functions. However, the loss of NBR is not Lipschitz owing to the unbounded responses. To handle the non-Lipschitz loss, we have to ensure the stochastic Lipschitz property (see Chi (2010)) of the NB loss with high probability. Thus we derive oracle inequalities for elastic-net estimates for the NBR under the Compatibility Factor Condition or Stabil Condition, which differs from the conditions in van de Geer (2008).

- Apart from the ℓ_1 consistency, few studies focus the sign consistent (Zhao and Yu (2006)) of the elastic-net type estimators, see Jia and Yu (2010) for the linear model, and Yu (2010) for the Cox model. Based on the bounded covariates assumption, we study the sign consistency of an elastic-net regularized NBR without using the *Irrepresentable Condition* in Zhao and Yu (2006).

We examine the theoretical properties of the elastic-net methods for a sparse estimator in the NBR within the framework of the nonasymptotic theory. Section 2.1 and Section 2.2 present a review of the NBR and KKT conditions. In Section 2.3 and 2.4, we show that two types of oracle inequalities can be derived for ℓ_1 estimation and prediction error bound under the assumption of the Compatibility Factor Condition or Stabil Condition with measurement errors. The remaining sections are byproducts of our proposed oracle inequalities. We establish a uniform bound for the grouping effect in Section 3.1. To obtain the sign consistency in Section 3.2, we require a uniform signal strength in order to detect coefficients larger than a constant multiplied by the tuning parameter of the ℓ_1 penalty. Using the weakest signal condition, in Section 3.3, we find that the probability of correct inclusion for all true variables in the selected set and the probability of corrected subset selection are high. We discuss de-biased elastic-net regularized M-estimators for low-dimensional parameters in Section 3.4. All proofs of the main theorems, lemmas, and propositions are given in Appendix S1, and the assisted lemmas are presented in Appendix S2. Simulation studies are provided in Appendix S3.

2. High-Dimensional NBR

In the following two subsections, we review the negative binomial GLMs and the corresponding mathematical optimization problems.

2.1. NBR

The probability mass function of the negative binomial distribution random variable is $p_y =: P(Y = y) = (\Gamma(y+\theta)/\Gamma(\theta)y!)(1-p)^\theta p^y, (p \in (0,1), y \in \mathbb{N})$. The expectation and variance of the NB distribution are $\theta p/(1-p)$ and $\theta p/(1-p)^2$, respectively. If θ is a positive integer, it is called a Pascal distribution. This special case of the NB is modeled as the number of failures $Y = y$ before the θ-th success in repeated mutually independent Bernoulli trials (with success probability $1-p$). Here, θ is a positive integer or real number.

In the regression setting, one type of NBR assumes that the count data response obeys the NB distribution (denoted as $Y \sim \mathrm{NB}(\mu_i, \theta)$) with overdispersion:

$$P(Y_i = y_i|\boldsymbol{X}_i) =: f(y_i, \theta, \mu_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(\theta)y_i!}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{y_i}\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}, \ (i = 1, 2, \ldots, n)$$

Here, $\mathrm{E}(Y_i\,|\boldsymbol{X}_i) = \mu_i$ and $\mathrm{Var}(Y_i\,|\boldsymbol{X}_i) = \mu_i + \mu_i^2/\theta$. The θ is a qualification of the level of overdispersion that underlies a count data set. Furthermore, θ is assumed as the known dispersion parameter which can be estimated (see Section 8 of Hilbe (2011)). When the mean parameter μ_i and the covariates are linked by $\log\mu_i = \boldsymbol{X}_i^T\boldsymbol{\beta}^*$, we have an NBR. When $\theta \to +\infty$, $\mathrm{Var}(Y_i\,|\boldsymbol{X}_i) \to \mu_i = \mathrm{E}(Y_i\,|\boldsymbol{X}_i)$. Thus, the Poisson regression is a limiting case of the NBR when the dispersion parameter tends to infinite. Because overdispersion occurs in real data, the NBR can be more powerful and interpretable than a Poisson regression.

The log-likelihood function of the NB responses is:

$$\begin{aligned} L(\boldsymbol{Y};\boldsymbol{\beta}) &= \log\left[\prod_{i=1}^{n} f(Y_i, \theta, \mu_i)\right] = \sum_{i=1}^{n}\log\left\{\frac{\Gamma(\theta+Y_i)}{\Gamma(\theta)Y_i!}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{Y_i}\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\right\} \\ &= \sum_{i=1}^{n}\{\log\Gamma(\theta+Y_i) + Y_i\log\mu_i + \theta\log\theta - \log\Gamma(\theta) - \log Y_i! - (\theta+Y_i)\log(\theta+\mu_i)\} \\ &= c_0 + \sum_{i=1}^{n}[Y_i\boldsymbol{X}_i^T\boldsymbol{\beta} - (\theta+Y_i)\log(\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})], \text{ with a constant } c_0. \end{aligned}$$

Then, take the derivative of the vector $\boldsymbol{\beta}$. Let $\partial L(\boldsymbol{Y};\boldsymbol{\beta})/\partial\boldsymbol{\beta} := \{\partial L(\boldsymbol{Y};\boldsymbol{\beta})/\partial\beta_1, \ldots, \partial L(\boldsymbol{Y};\boldsymbol{\beta})/\partial\beta_p\}^T$. We get the *score function*

$$\dot{\ell}_n(\boldsymbol{\beta}) := -\frac{1}{n}\frac{\partial L(\boldsymbol{Y};\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\theta\left[\frac{\theta+Y_i}{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}} - 1\right] = -\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i(Y_i - e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})\theta}{\theta+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}. \tag{2.1}$$

By setting $\dot{\ell}_n(\boldsymbol{\beta}) = 0$, we obtain the solution $\hat{\boldsymbol{\beta}}_{mle}$. The second derivative is calculated as the *Hessian matrix* $\ddot{\ell}_n(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^T \theta(\theta + Y_i) e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}/(\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})^2$, which is semi-negative, such that $\hat{\beta}_{mle}$ makes the likelihood function take the maximum value globally.

2.2. KKT conditions

Let $g(\boldsymbol{\beta})$ be a nonnegative convex function with $g(\mathbf{0}) = \mathbf{0}$, and λ_1 and λ_2 be positive turning parameters. Yu (2010) considered a penalized likelihood for the convex loss function $\ell(\boldsymbol{\beta})$,

$$F(\boldsymbol{\beta}; \lambda_1, \lambda_2) = \ell_n(\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 g(\boldsymbol{\beta})$$

as the generalized Lasso-type convex penalty (GLCP). The GLCP estimator for the general log-likelihood is $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2) = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} F(\boldsymbol{\beta}; \lambda_1, \lambda_2)$. By the subderivative technique in the optimization function, the corresponding Karush–Kuh–Tucker(KKT) conditions of GLCP estimator are

$$\begin{cases} \dot{\ell}_{n,j}(\hat{\boldsymbol{\beta}}) + \lambda_2 \dot{g}_j(\hat{\boldsymbol{\beta}}) = -\lambda_1 \operatorname{sign}(\hat{\beta}_j) & \text{if } \hat{\beta}_j \neq 0 \\ |\dot{\ell}_{n,j}(\hat{\boldsymbol{\beta}}) + \lambda_2 \dot{g}_j(\hat{\boldsymbol{\beta}})| \leq \lambda_1 & \text{if } \hat{\beta}_j = 0 \end{cases} \tag{2.2}$$

See page 68 of Bühlmann and van de Geer (2011)). Thus, in the NBR, the KKT conditions for the non-zero (or zero) elastic-net estimate is

Lemma 1. *(Necessary and Sufficient Condition). Let $k \in \{1, 2, \ldots, p\}$ and $\lambda_2 > 0$. Then, a necessary and sufficient condition for elastic-net estimates of the NBR to be a solution of* (1.1) *is*

1. $\hat{\beta}_k = \hat{\beta}_k \neq 0$ *if* $(1/n)\sum_{i=1}^n x_{ik}\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - Y_i)/(\theta + e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}}) = [\operatorname{sign}\hat{\beta}_k](\lambda_1 + 2\lambda_2|\hat{\beta}_k|)$.
2. $\hat{\beta}_k = 0$ *if* $\left|(1/n)\sum_{i=1}^n x_{ik}\theta(e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}} - Y_i)/(\theta + e^{\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}})\right| \leq \lambda_1$.

Zhou (2013) gave an elementary proof of KKT conditions for the elastic-net penalized optimization problem in a linear regression. Note that the KKT conditions are a standard result of sub-differentiation techniques. The prerequisite $\lambda_2 > 0$ in Lemma 1 is indispensable. The reason is that we need $\lambda_2 > 0$, such that $F(\hat{\boldsymbol{\beta}} + \varepsilon\mathbf{e}_k; \lambda_1, \lambda_2) - F(\hat{\boldsymbol{\beta}}; \lambda_1, \lambda_2) > 0$ where $\{\mathbf{e}_k\}_{k=1}^p$ are unit coordinate vectors, see Appendix S2. Then $\hat{\boldsymbol{\beta}}$ is the unique local minimum. The KKT conditions are crucial for all sections below.

2.3. ℓ_q-estimation error using a compatibility factor

This section presents the sparse estimator for a high-dimensional NBR by using the fact that the elastic-net estimator is asymptotically close to the true parameter under some suitable regularity conditions.

For fixed designs $\{\boldsymbol{X}_i\}_{i=1}^n$, let $\boldsymbol{\beta}^*$ be the vector of true coefficients, which satisfies

$$\mathrm{E}Y_i = e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}. \tag{2.3}$$

In some sense, we can never really know the expectation of the negative log-likelihood, because $\boldsymbol{\beta}^*$ is the unknown parameter in the functional estimating equation $\boldsymbol{X}_i^T\boldsymbol{\beta}^* = \log(\mathrm{E}Y_i)$.

In high-dimensions, we are interested in the sparse estimates defined in (1.1) by adding elastic-net penalty. For the true coefficient vector $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$, let $H = \{j : \beta_j^* \neq 0, j = 1, \ldots, p\}$ and $H^c = \{j : \beta_j^* = 0, j = 1, \ldots, p\}$ be the nonzero and zero components, respectively. Let $d_H^* = |H|$ be the number of nonzero coefficients in $\boldsymbol{\beta}^*$, i.e. the support of $\boldsymbol{\beta}^*$. For any $\boldsymbol{b} \in \mathbb{R}^p$ and index set $H \in \{1, 2, \ldots, p\}$, define the sub-vector indexed by H as $\boldsymbol{b}_H = (\ldots, \tilde{b}_j, \ldots)^T \in \mathbb{R}^p$, with $\tilde{b}_j = b_j$ if $j \in H$, and $\tilde{b}_j = 0$ if $j \notin H$. In the MLE theory, we know that the Kullback–Leibler (K–L) divergence measures how one probability distribution is different from another, based on a quasi-distance of two log-likelihoods. Similarly, in order to measure the derivative discrepancy between two penalized log-likelihood function w.r.t. the parameters, the *symmetric Bregman (SB) divergence* between $\ell(\boldsymbol{\beta}_1)$ and $\ell(\boldsymbol{\beta}_2)$ is

$$D_g^s(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^T[\dot{\ell}_n(\boldsymbol{\beta}_1) - \dot{\ell}_n(\boldsymbol{\beta}_2) + \lambda_2(\dot{g}(\boldsymbol{\beta}_1) - \dot{g}(\boldsymbol{\beta}_2))],\ \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p.$$

If $g = 0$, the symmetric Bregman divergence is $D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T[\dot{\ell}_n(\hat{\boldsymbol{\beta}}) - \dot{\ell}_n(\boldsymbol{\beta})]$. In this case, the symmetric Bregman divergence is a type of generalized quadratic distance (Mahalanobis distance), which can been viewed as a symmetric extension of the K–L divergence. See Nielsen and Nock (2009) and Huang et al. (2013) for more details about SB divergence. Because $g(\boldsymbol{\beta})$ is a nonnegative convex function, we have the inequality: $D_g^s(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \geq D^s(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$.

The key to derive the oracle inequalities also depends on the behavior of the Hessian matrix of the NBR: $\ddot{\ell}_n(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^T$, where $\tilde{\boldsymbol{X}}_i := \boldsymbol{X}_i(\theta(\theta+Y_i)e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}/((\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})^2))^{1/2}$ is the *curvature-scaled design*.

In the fixed design linear model $\mathrm{E}\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^*$ with $\mathrm{Var}\boldsymbol{Y} = \mathbf{I}_p\sigma^2$, it can be

shown that, with probability greater than $1-\delta_n$,

$$\|\hat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta}^*\|_2 \leq \sigma\sqrt{\frac{p}{n}}\cdot\left[\delta_n\lambda_{\min}\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)\right]^{-1/2}. \tag{2.4}$$

for the ordinary least square (OLS) estimator $\hat{\boldsymbol{\beta}}_{LS}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$, see Section 8.1 of Zhang and Chen (2021). In an increasing dimension $p=p(n)$, it is well–known that the *Gram matrix* is $(1/n)\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$ (i.e., the correlation matrix between the covariates), which is singular when $p>n$. The positivity assumption of the $\lambda_{\min}((1/n)\mathbf{X}^T\mathbf{X})$ is crucial to obtain optimal convergence under $p<\infty$. In the sparse high-dimensional linear model via Lasso, to obtain the oracle inequality with the fast and optimal rate as discussed in Bickel, Ritov and Tsybakov (2009), the following versions of the restricted minimal eigenvalue is usually needed under sparse cone set (2.5).

Let the *sparse cone set* be

$$\mathrm{S}(s,H) := \{\boldsymbol{b}\in\mathbb{R}^p : ||\boldsymbol{b}_{H^c}||_1 \leq s||\boldsymbol{b}_H||_1\}, \quad (s\in\mathbb{R}^+). \tag{2.5}$$

The *compatibility factor* (denoted by $C(s,H,\boldsymbol{\Sigma})$; see van de Geer (2007)) of a $p\times p$ nonnegative-definite matrix $\boldsymbol{\Sigma}$ is defined by

$$C^2(s,H,\boldsymbol{\Sigma}) := \inf_{0\neq\boldsymbol{b}\in\mathrm{S}(s,H)}\frac{d_H^*(\boldsymbol{b}^T\boldsymbol{\Sigma}\boldsymbol{b})}{\|\boldsymbol{b}_H\|_1^2} > 0, \quad (s\in\mathbb{R}^+). \tag{2.6}$$

To derive the ℓ_q-loss ($q>1$) oracle inequalities for the target coefficient vectors, we require the concept of *weak cone invertibility factors* (weak CIF; see (53) of Ye and Zhang (2010)),

$$C_q(s,H,\boldsymbol{\Sigma}) := \inf_{0\neq\boldsymbol{b}\in\mathrm{S}(s,H)}\frac{{d_H^*}^{1/q}(\boldsymbol{b}^T\boldsymbol{\Sigma}\boldsymbol{b})}{||\boldsymbol{b}_H||_1\cdot||\boldsymbol{b}||_q} > 0, \quad (s\in\mathbb{R}^+). \tag{2.7}$$

This constant generalizes the compatibility factor, and is close to the restricted eigenvalue; see Bickel, Ritov and Tsybakov (2009).

From the results in Ye and Zhang (2010) and Huang et al. (2013), we know that the positivity assumptions of compatibility factor and the weak CIF can achieve sharper upper bounds for the oracle inequalities because both are bigger than the restricted eigenvalue:

$$\mathrm{Re}(s,H,\boldsymbol{\Sigma}) := \inf_{0\neq\boldsymbol{b}\in\mathrm{S}(s,H)}\frac{\boldsymbol{b}^T\boldsymbol{\Sigma}\boldsymbol{b}}{\|\boldsymbol{b}\|_2^2} \leq \inf_{0\neq\boldsymbol{b}\in\mathrm{S}(s,H)}\frac{d_H^*(\boldsymbol{b}^T\boldsymbol{\Sigma}\boldsymbol{b})}{\|\boldsymbol{b}_H\|_1^2} = C^2(s,H,\boldsymbol{\Sigma}), \ (s\in\mathbb{R}^+),$$

due to $\|\boldsymbol{b}_H\|_1 \leq {d_H^*}^{1/2}\|\boldsymbol{b}\|_2$.

Using the definitions of SB divergence with $\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_2 = \boldsymbol{\beta}^*$, let $z^* := \|\dot{\ell}_n(\boldsymbol{\beta}^*) + \lambda_2 \dot{g}(\boldsymbol{\beta}^*)\|_\infty$ and $\Delta := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. We now provide the lower and upper bounds for the symmetric Bregman divergence.

Lemma 2 (Theorem 1 in Yu (2010))**.** *For the GLCP estimation, we have*

$$(\lambda_1 - z^*)||\Delta_{H^c}||_1 \leq D_g^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + (\lambda_1 - z^*)||\Delta_{H^c}||_1 \leq (\lambda_1 + z^*)||\Delta_H||_1. \quad (2.8)$$

If $z^* \leq ((\zeta - 1)/(\zeta + 1))\lambda_1$, for some $\zeta > 1$, the inequality (2.8) imply

$$\frac{2\lambda_1}{\zeta + 1}||\Delta_{H^c}||_1 \leq D_{\mathrm{g}}^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + \frac{2\lambda_1}{\zeta + 1}||\Delta_{H^c}||_1 \leq \frac{2\zeta\lambda_1}{\zeta + 1}||\Delta_H||_1, \quad (2.9)$$

from $\lambda_1 - z^* \geq 2\lambda_1/(\zeta + 1)$ and $\lambda_1 + z^* \leq 2\zeta\lambda_1/(\zeta + 1)$. By (2.9), we have

$$||\Delta_{H^c}||_1 \leq \zeta\|\Delta_H\|_1. \quad (2.10)$$

Hence we conclude that in the event

$$\mathcal{K}_\lambda := \left\{ z^* = \|\dot{\ell}_n(\boldsymbol{\beta}^*) + \lambda_2 \dot{g}(\boldsymbol{\beta}^*)\|_\infty \leq \frac{\zeta - 1}{\zeta + 1}\lambda_1 \right\},$$

the error of estimate $\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathrm{S}(\zeta, H)$. Then assumptions $C^2(s, H, \boldsymbol{\Sigma}) > 0$ and $C_q(s, H, \boldsymbol{\Sigma}) > 0$ for the Hessian matrix $\boldsymbol{\Sigma} = \ddot{\ell}_n(\boldsymbol{\beta}^*)$ are indispensable assumptions for deriving the targeted oracle inequalities from the optimization (1.1) and the expected version (2.3). Some additional regularity conditions are required.

- (C.1): Assume bounded covariates,

$$\max\{|x_{ij}|\,; 1 \leq i \leq n, 1 \leq j \leq p\} = L < \infty.$$

- (C.2): Based on the covairates $\{\boldsymbol{X}_i\}_{i=1}^n$, we assume identifiability condition that $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfies

$$\boldsymbol{X}_i^T(\boldsymbol{\beta} + \boldsymbol{\delta}) = \boldsymbol{X}_i^T\boldsymbol{\beta} \text{ implies } \boldsymbol{X}_i^T\boldsymbol{\delta} = 0 \text{ for } \boldsymbol{\delta} \in \mathbb{R}^p.$$

- (C.3): Suppose that $||\boldsymbol{\beta}^*||_1 \leq B$.

The bounded covariates in C.1 are a common assumption in GLMs (see Example 5.40 of van der Vaart (1998)); it may be achieved by performing a bounded and monotone transformation of the covariates in the real data. The identifiability condition C.2 and the compact parameter space C.3 are common assumptions

for obtaining the consistency for a general M-estimation; see Section 5.5 and the remark after Theorem 5.9 in van der Vaart (1998). Recently, Weißbach and Radloff (2020) showed the consistency of the NBR with fixed covariates, under the assumption that all possible parameters and the regressor are in some compact spaces.

First, we present the nonasymptotic upper bounds for the elastic-net regularized NBR in the following two theorems.

Theorem 1. *Let $C(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*))$ and $C_q(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*))$ be the compatibility factor and the weak cone invertibility factor, respectively, defined above. Define $\tau := L(\zeta+1)d^*\lambda_1/(2[C(\zeta,H)]^2) \leq (1/2)e^{-1}$. Assume that (*C.1*), (*C.2*), and the event $\mathcal{K}_\lambda$ hold. Then, we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{e^{2a_\tau}(\zeta+1)d_H^*\lambda_1}{2C^2(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*))} \quad \textit{and} \quad \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q \leq \frac{2e^{2a_\tau}\zeta {d_H^*}^{1/q}\lambda_1}{(\zeta+1)\, C_q(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*))}, \tag{2.11}$$

where $a_\tau \leq 1/2$ is the smaller solution of the equation $ae^{-2a} = \tau$.

On the one hand, the Theorem 1 contains basic oracle inequalities conditioning on the random event, which needs further refinements. What remains to be done is to focus the probability upper bound of event $\mathcal{K}_\lambda$. With assumption (C.3), we have $z^* \leq \|\dot{\ell}_n(\boldsymbol{\beta}^*)\|_\infty + 2\lambda_2 B$. Our aim of proof is to have

$$P(\mathcal{K}_\lambda^c) \leq P\left(\|\dot{\ell}_n(\boldsymbol{\beta}^*)\|_\infty \geq \frac{\zeta-1}{\zeta+1}\lambda_1 - 2\lambda_2 B\right) \to 0 \ \text{ as } n, p \to \infty, \tag{2.12}$$

provided that λ_2 is sufficient small.

To bound $\|\dot{\ell}_n(\boldsymbol{\beta}^*)\|_\infty$, all we need is to apply some concentration inequalities in terms of the NB empirical processes (2.1), that is the sum of independent weighted centralized NB random variables. Because the dispersion parameter θ is known, the NB random variables $\{Y_i\}_{i=1}^n$ belong to the exponential family

$$f(y_i; \eta_i) \propto \exp\{y_i\eta_i - \psi(\eta_i)\} \text{ with } \eta_i := \boldsymbol{X}_i^T\boldsymbol{\beta}^* + \log(\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}) \in \Theta, \tag{2.13}$$

where Θ is the compact parameter space. Thus, under fixed design, the sub-Gaussian concentration inequalities for the non-random weighted sum of exponential family random variables with compact parameter space is applicable; see Lemma 6.1 in Rigollet (2012) or Proposition 3.2 in Zhang and Chen (2021) with more discussion.

On the other hand, the Compatibility Factor and weak CIR we employ in this section are random constants. They contains the Hessian matrix of the true

coefficient vector, and thus encapsulate the random quantities $\{Y_i\}_{i=1}^n$. Note that deriving the lower bound for these random quantities decreases the probability that oracle inequalities are true, but the loss is negligible in the next theorem. Next, we successfully show using the NB concentration inequality that a reasonable non-random lower bounds of the compatibility factor (or the weak CIR) makes sure that the upper bounds are constants with high probability. Thus the rigorous convergence rate of $\hat{\boldsymbol{\beta}}$ is well established. Note that Yu, Bradic and Samworth (2021) directly assume that the inverse of compatibility factor of $\ddot{\ell}(\boldsymbol{\beta}^*)$ for the Cox model is $O_p(1)$, which they call it "a high-level condition". The Hessian matrix of the Cox model is also a random element.

Two events for truncating the compatibility factor and the weak CIR, is defined by

$$\mathcal{E}_c := \{C^2(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*)) > C_t^2(\zeta, H)\} \text{ and } \mathcal{E}_w := \{C_q(\zeta, H, \ddot{\ell}_n(\boldsymbol{\beta}^*)) > C_{qu}(\zeta, H)\},$$

where $C_t^2(\zeta, H)$ and $C_{qu}(\zeta, H)$ are nonrandom constants defined in the proof for certain constants $t, u > 0$.

Theorem 2. *Under the assumptions of Theorem* 1*, we further assume (*C.2*). Let B_1 be the constant satisfying $C_{\xi,B_1} := (\zeta - 1)/(\zeta + 1) - 2B_1 > 0$. Let $\lambda_1 = (C_{LB}L/C_{\xi,B_1})\sqrt{2r \log p/n}$, where $C_{LB}^2 := e^{LB} + e^{2LB}/\theta$ is a variance-depending constant and $r > 1$ is a constant. Put $\lambda_2 = B_1\lambda_1/B$. Under the event $\mathcal{K} \cap \mathcal{E}_c$ (or $\mathcal{K} \cap \mathcal{E}_w$), we have:*

$$\begin{aligned} & P\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le \frac{e^{2a_\tau}(\zeta+1)d_H^*\lambda_1}{2C_t^2(\zeta, H)}\right) \\ & \ge 1 - \frac{2}{p^{r-1}} - 2p^2 e^{-nt^2/(2[d_H^* C_{LB}(1+\varsigma)L^2]^2)} \qquad (2.14) \\ & or\ P\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_q \le \frac{2e^{2a_\tau}\zeta {d_H^*}^{1/q}\lambda_1}{(\zeta+1)\, C_{qu}(\zeta, H)}\right) \\ & \ge 1 - \frac{2}{p^{r-1}} - 2p^2 e^{-nu^2/(2[d_H^* C_{LB}(1+\varsigma)L^2]^2)}. \qquad (2.15) \end{aligned}$$

If we presume the condition $d_H^* = O(1)$ in Theorem 2, which implies that the error bound is of order $\sqrt{\log p/n}$, the elastic-net estimates have ℓ_1-consistency property when the dimension of covariates increases with order $e^{o(n)}$. The MLE has the convergence rate $1/\sqrt{n}$. Nevertheless, in high-dimensional condition, we have to magnify $\sqrt{\log p}$ to the convergence rate of MLE. If we assume $d_H^* = o(\sqrt{n/\log p})$, that is $p = e^{o(n/d_H^*)}$, then $d_H^*\lambda = o(1)$ which implies the consistency property. If we consider random designs, the story is different. Our purpose in

next section is to present an approach that avoids the random upper bound for the ℓ_1 or ℓ_2 estimation error, and provides the oracle inequality for the squared prediction error.

2.4. The prediction error under a random design

In this section, we focus on the prediction error. We assume that the $n \times p$ design matrix $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ is random. In our applications, the test data set is a new design $\mathbf{X}^*$, which is an independent copy of $\mathbf{X}$. Thus it requires the randomness assumption of the design matrix. We aim to predict the response Y_{n+1} using the new random covariates $\boldsymbol{X}_{n+1}$ by resorting to elastic-net estimator $\hat{\boldsymbol{\beta}}$ to estimate the unknown Y_{n+1}.

Here $\boldsymbol{Y} \in \mathbb{R}^n$ contains n independent (ind.) responses $\{Y_i\}_{i=1}^n$. Thus the covariates and responses are considered pairs of random vectors $(\mathbf{X}, \boldsymbol{Y})$. When $\{\boldsymbol{X}_i\}_{i=1}^n$ is degenerately distributed, it reduces to a fixed design, and hence the result here also holds for a fixed design. Through this paper, we denote the element in the design matrix $\{x_{ij}\}$ as fixed design, and $\{X_{ij}\}$ as random design. The conditional distribution of a single observation $Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i$ is assumed to be conditional NB distributed with $\mathrm{E}(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) = e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}$.

Let $\boldsymbol{\beta}^*$ be the vector of true coefficients, which is defined by the minimizer

$$\boldsymbol{\beta}^* = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathrm{E}l(Y, \boldsymbol{X}, \boldsymbol{\beta}), \tag{2.16}$$

where $l(Y, \boldsymbol{X}, \boldsymbol{\beta}) = Y\boldsymbol{X}^T\boldsymbol{\beta} - (\theta + Y)\log(\theta + e^{\boldsymbol{X}^T\boldsymbol{\beta}})$ is the NB loss.

To derive nonasymptotical bounds for the ℓ_1 estimation and square prediction error, we focus on the empirical process for any possible $\boldsymbol{\beta}$ [on the NB loss function in (2.16) with random $\mathbf{X}$],

$$\mathbb{P}_n l(\boldsymbol{X}, Y, \boldsymbol{\beta}) := -\frac{1}{n}\sum_{i=1}^n [Y_i\boldsymbol{X}_i^T\boldsymbol{\beta} - (\theta + Y_i)\log(\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}})],$$

where $\mathbb{P}_n$ is the empirical measure of the samples $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n \overset{\text{ind.}}{\sim} (\boldsymbol{X}, Y)$.

The concentration and fluctuation of the empirical process are crucial to evaluating the consistent properties of the estimates. The proof oracle inequalities in this section consists 3 steps, including: 1. Checking $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ be in cone set by using definition of penalized estimation and KKT-like conditions; 2. Verifying the high probability of KKT-like conditions; 3. Deriving the oracle inequalities from restricted eigenvalue condition with some elementary inequalities. For simplicity, we use symbols for the empirical process in this section. We need some

assumptions, such that $\hat{\boldsymbol{\beta}}$ is consistent.

- (H.1): All variables $\boldsymbol{X}_i$ are bounded: there exists a constant $L > 0$, such that $|||\mathbf{X}|||_\infty := \sup_{1\le i\le\infty} \|\boldsymbol{X}_i\|_\infty \le L$ a.s.
- (H.2): Assume that $||\boldsymbol{\beta}^*||_1 \le B$.
- (H.3): There exists a large constant M_0, such that $\hat{\boldsymbol{\beta}}$ is in the ℓ_1 ball:

$$\hat{\boldsymbol{\beta}} \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le M_0\}.$$

- (H.4): Let $\theta > 1$. The negative log-density of n independent NB responses $\psi(\boldsymbol{y}) := -\log p_{\boldsymbol{Y}}(\boldsymbol{y})$, for $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, satisfies the *strongly midpoint log-convex* properties for some $\gamma > 0$,

$$\psi(\boldsymbol{x})+\psi(\boldsymbol{y})-\psi\left(\left\lceil\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\right\rceil\right)-\psi\left(\left\lfloor\frac{1}{2}\boldsymbol{x}+\frac{1}{2}\boldsymbol{y}\right\rfloor\right) \ge \frac{\gamma}{4}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}^n. \tag{2.17}$$

Remark 1. (H.1) and (H.2) are mentioned in Blazere, Loubes and Gamboa (2014), and (H.3) is a high technique condition owing to the noncanonical link GLMs. The constraint in the optimization is equivalent to $\alpha\|\boldsymbol{\beta}\|_1+(1-\alpha)\|\boldsymbol{\beta}\|_2^2 \le t$, with unknown $\alpha \in [0,1]$ and $t \in \mathbb{R}$ leading to $\|\hat{\boldsymbol{\beta}}\|_1 \le M_0$ if we suppose that $t/\alpha \le M_0$. There is a constant $K > 0$, such that $\max_{1\le i\le n}\left|\boldsymbol{X}_i^T\boldsymbol{\beta}^*\right| \le K$ a.s., for all n. A convex function F is called strongly convex if the Hessian matrix of F has a (uniformly) lower bounded eigenvalue. While examining exponential families in high dimensions, Kakade et al. (2010) assumed that continuous exponential families (2.13) have strongly convex log-likelihood function with η_i in a sufficiently small neighborhood. For a fixed dimensional MLE, Balabdaoui et al. (2013) show that the discrete log-concave maximum likelihood estimator is strongly consistent under some settings. Our assumption (H.4) is a condition that ensures that the suprema of the multiplier empirical processes of n independent responses have sub-Gaussian concentration phenomena in (S1.19), which can be alternatively be checked by the tail inequality for suprema of empirical processes corresponding to classes of unbounded functions (Adamczak (2008)). For the case of a fixed design in Section 2.3, we do not require (H.4) in order to derive the oracle inequalities.

In this section, we give sharp bounds for ℓ_1 estimation and squared prediction errors for NBR models by looking for a weaker condition that is analogous to the restricted eigenvalue (RE) condition proposed by Bickel, Ritov and Tsybakov (2009), and the weak CIF and compatibility factor conditions presented in Section

3.2. Here, we borrow a condition from the Stabil Condition introduced by Bunea (2008) for ℓ_1 and $\ell_1 + \ell_2$ penalized logistic regressions.

For $c, \varepsilon > 0$, we define the *fluctuated cone set* for some bias vector $\boldsymbol{b}$ as

$$\mathrm{V}(c, \varepsilon, H) := \{\boldsymbol{b} \in \mathbb{R}^p : ||\boldsymbol{b}_{H^c}||_1 \leq c||\boldsymbol{b}_H||_1 + \varepsilon\}, \tag{2.18}$$

which is a fluctuated (or measurement error) version of the cone set $\mathrm{S}(s, H) := \{\boldsymbol{b} \in \mathbb{R}^p : ||\boldsymbol{b}_{H^c}||_1 \leq s||\boldsymbol{b}_H||_1\}$ mentioned in (2.5).

We substitute $\boldsymbol{b} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ into the proof. For real data, let $\hat{\boldsymbol{\beta}}$ be the estimator based on the true covariates, and let $\hat{\boldsymbol{\beta}}_{me}$ be the estimator from covariates with a measurement error. Note that under the cone condition $||\boldsymbol{b}_{H^c}||_1 \leq c||\boldsymbol{b}_H||_1$, for $\boldsymbol{b} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, we get

$$\begin{aligned} ||(\hat{\boldsymbol{\beta}}_{me} - \boldsymbol{\beta}^*)_{H^c}||_1 - ||(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{me})_{H^c}||_1 &\leq ||(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{H^c}||_1 \leq c||(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_H||_1 \\ &\leq c||(\hat{\boldsymbol{\beta}}_{me} - \boldsymbol{\beta}^*)_H||_1 + c||(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{me})_H||_1. \end{aligned}$$

Then,

$$||\boldsymbol{b}^{me}_{H^c}||_1 \leq c||\boldsymbol{b}^{me}_H||_1 + \varepsilon \quad \text{for } \boldsymbol{b}^{me} := \hat{\boldsymbol{\beta}}_{me} - \boldsymbol{\beta}^*,$$

where $\varepsilon = c||(\hat{\boldsymbol{\beta}}_{me} - \boldsymbol{\beta}^*)_H||_1 + ||(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{me})_{H^c}||_1$. This argument indicates that the fluctuated cone set quantifies the level of the measurement error if $\hat{\boldsymbol{\beta}}_{me}$ is misspecified as $\hat{\boldsymbol{\beta}}$.

On the fluctuated cone set, we assume that the $p \times p$ the expected empirical covariance matrix $\boldsymbol{\Sigma} = \mathrm{E}\boldsymbol{X}\boldsymbol{X}^T$ fulfills the Stabil Condition as below. The Stabil Condition for matrix $\boldsymbol{\Sigma}$ avoids the random Hessian matrix in the Compatibility Factor Condition and the weak CIF Condition. However, there is no free lunch. The proposed oracle inequalities in this section require (H.4), which serves for the tail inequality for the suprema of NB empirical processes.

Definition 2. (Stabil Condition with measurement error). For given $c, \varepsilon > 0$, the matrix $\boldsymbol{\Sigma}$ satisfies the Stabil Condition $S(c, \varepsilon, k)$ if there exists $0 < k < 1$, such that

$$\mathbf{b}^T\boldsymbol{\Sigma}\mathbf{b} \geq k||\mathbf{b}_H||_2^2 - \varepsilon$$

for any $\mathbf{b} \in V(c, \varepsilon, H)$. Here, the restriction $0 < k < 1$ can be attained by scaling $\boldsymbol{X}$.

Let $l_1(\boldsymbol{\beta}) := l_1(\boldsymbol{\beta}, \boldsymbol{X}, Y) := -Y[\boldsymbol{X}^T\boldsymbol{\beta} - \log(\theta + \exp\{\boldsymbol{X}^T\boldsymbol{\beta}\})]$, which is a linear function of the response, and let $l_2(\boldsymbol{\beta}) := l_2(\boldsymbol{\beta}, \boldsymbol{X}) := \theta\log(\theta + \exp\{\boldsymbol{X}^T\boldsymbol{\beta}\})$, which is free of the response. The NB loss function $l(\boldsymbol{\beta}, \boldsymbol{X}, Y) = l_1(\boldsymbol{\beta}, \boldsymbol{X}, Y) + l_2(\boldsymbol{\beta}, \boldsymbol{X})$ is thus decomposed into two parts. Let $\mathbb{P}l(\boldsymbol{\beta}) := \mathrm{E}l(\boldsymbol{\beta}, \boldsymbol{X}, Y)$ be the expected risk

function, where the expectation is under the randomness of $(\boldsymbol{X}, Y)$. We prefer the centralized empirical loss $(\mathbb{P}_n - \mathbb{P})\, l(\boldsymbol{\beta})$, which represents the fluctuation between the expected and the sample loss, rather than the loss itself. We break down the empirical process into two parts:

$$(\mathbb{P}_n - \mathbb{P})\, l(\boldsymbol{\beta}) = (\mathbb{P}_n - \mathbb{P})\, l_1(\boldsymbol{\beta}) + (\mathbb{P}_n - \mathbb{P})\, l_2(\boldsymbol{\beta}). \tag{2.19}$$

In the following, we give upper bounds for the first and second parts of the empirical process: $(\mathbb{P}_n - \mathbb{P})(l_m(\boldsymbol{\beta}^*) - l_m(\hat{\boldsymbol{\beta}}))$, for $m = 1, 2$. We show that $(\mathbb{P}_n - \mathbb{P})(l_m(\boldsymbol{\beta}^*) - l_m(\hat{\boldsymbol{\beta}}))$ has *stochastic Lipschitz properties* (see Chi (2010)) with respect to $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$. Let the ℓ_1 ball be $\mathcal{S}_{M_0}(\boldsymbol{\beta}^*) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le M_0\}$, which is referred as the *local set*. Then,

Proposition 1. *Let the centered responses be* $\{Y_i^c := Y_i - \mathrm{E}Y_i\}_{i=1}^n$ *and,* (H.1)*–*(H.4) *are satisfied. If* $\lambda_1 \ge 4L(2\tilde{C}_{LB} + A\sqrt{2\gamma})\sqrt{2\log 2p/n}$,*(*$A \ge 1, \tilde{C}_{LB}^2 := e^{LB} + (1+\theta)e^{2LB}/\theta$*), define the event* $\mathcal{A}$ *for the suprema of the multiplier empirical processes as*

$$\mathcal{A} := \left\{ \sup_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)} \left| \frac{1}{n} \sum_{i=1}^n \frac{Y_i^c \theta \boldsymbol{X}_i^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*)}{(\theta + \exp\{\boldsymbol{X}_i^T \boldsymbol{\beta}_2\}) \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\|_1} \right| \le \frac{\lambda_1}{4} \right\}.$$

Then, we have $P(\mathcal{A}) \ge 1 - (2p)^{-A^2}$. *Moreover,*

$$P\left\{ (\mathbb{P}_n - \mathbb{P})(l_1(\boldsymbol{\beta}^*) - l_1(\hat{\boldsymbol{\beta}})) \le \frac{\lambda_1}{4} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \right\} \ge 1 - (2p)^{-A^2}.$$

This proposition indicates that the discrepancy between the first part of the empirical process and its expectation is bounded from above by the tuning parameter multiplied by the ℓ_2 norm of the difference between the estimated vector and the target vector. The $\lambda_1/4$ can be seen as a Lipschitz constant of the first part of the centralized empirical process.

Similarly to $\mathcal{A}$ as a KKT-like condition, we provide a crucial lemma to bound the second part of the empirical process with responses. Let $\nu_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*) := (\mathbb{P}_n - \mathbb{P})\,(l_2(\boldsymbol{\beta}^*) - l_2(\boldsymbol{\beta}))/(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 + \varepsilon_n)$ be the normalized second part of the empirical process, which is a random variable indexed by $\boldsymbol{\beta}$. Then we define the *local stochastic Lipschitz constant* for a certain $M > 0$,

$$Z_M(\boldsymbol{\beta}^*) := \sup_{\boldsymbol{\beta} \in \mathcal{S}_M(\boldsymbol{\beta}^*)} |\nu_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*)|, \text{ and a random event } \mathcal{B} := \left\{ Z_M(\boldsymbol{\beta}^*) \le \frac{\lambda_1}{4} \right\},$$

where we bound the local stochastic Lipschitz constant using the rescaled tuning

parameter $\lambda_1/4$. Moreover, by definition, we have

$$|\nu_n(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)| \leq \sup_{\mathcal{S}_M(\boldsymbol{\beta}^*)} |\nu_n(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)| \leq \frac{\lambda_1}{4},$$

which gives following bound,

$$|(\mathbb{P}_n - \mathbb{P})(l_2(\hat{\boldsymbol{\beta}}) - l_2(\boldsymbol{\beta}^*))| \leq \frac{\lambda_1}{4}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \varepsilon_n) \text{ on } \mathcal{B}, \tag{2.20}$$

provided that $\hat{\boldsymbol{\beta}} \in \mathcal{S}_M(\boldsymbol{\beta}^*)$.

According to the following lemma, in the event $\mathcal{A} \bigcap \mathcal{B}$, the estimator $\hat{\boldsymbol{\beta}}$ lies in a known neighborhood of the true coefficient vector $\boldsymbol{\beta}^*$.

Lemma 3. *Under (*H.2*), let* $8B\lambda_2 + 4M = \lambda_1$*, we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 16||\boldsymbol{\beta}^*|| + 2\varepsilon_n \quad \textit{on } \mathcal{A} \bigcap \mathcal{B}.$$

The proof of Lemma 3 relies on the optimization (1.1) and the definition of the minimizer $\boldsymbol{\beta}^*$ from the expected loss (2.16). By Lemma 3, on the event $\mathcal{A} \bigcap \mathcal{B}$, we immediately get $\hat{\boldsymbol{\beta}} \in \mathcal{S}_{16B+2\varepsilon_n}(\boldsymbol{\beta}^*)$. Note that we assume that $\hat{\boldsymbol{\beta}} \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$, for some finite $M_0 > M = 16B + 2\varepsilon_n$ in (H.3). That is Lemma 3 sharpens $\hat{\boldsymbol{\beta}}$ in the ℓ_1-ball $\mathcal{S}_M(\boldsymbol{\beta}^*)$, whereas $\hat{\boldsymbol{\beta}}$ is originally assumed in the ℓ_1 ball $\mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$. Therefore, the following probability analysis of the event $\mathcal{A} \bigcap \mathcal{B}$ is indispensable. The event $\mathcal{A} \cap \mathcal{B}$ associated with the empirical loss functions plays an important role in deriving the oracle inequalities for general loss functions, because we could bound the ℓ_1 estimation error conditioning on event $\mathcal{A} \bigcap \mathcal{B}$. We now give the result that the event $\mathcal{A} \cap \mathcal{B}$ occurs with a high probability.

Proposition 2. *Let* $M = 16B + 2\varepsilon_n$*. Suppose* $\hat{\boldsymbol{\beta}} \in \mathcal{S}_{M_0}(\boldsymbol{\beta}^*)$*, for* $\infty > M_0 > M$*, and that (*H.1*)-(*H.4*) hold. If*

$$\lambda_1 \geq \max\left(\frac{20\theta AML}{M + \varepsilon_n}\sqrt{\frac{2\log 2p}{n}}, 4L(2\tilde{C}_{LB} + A\sqrt{2\gamma})\sqrt{\frac{2\log 2p}{n}}\right), \ A \geq 1, \tag{2.21}$$

then $P(\mathcal{A} \cap \mathcal{B}) \geq 1 - 2(2p)^{-A^2}$.

The proof of Theorem 3 is based on some lemmas in Appendix S1, which show that the event $\mathcal{A} \cap \mathcal{B}$ holds with a high probability. Judging from the above probability analysis, we can formulate the main result of this section that gives bounds for the estimation and prediction error because the target model is sparse, and $\log p$ is tiny compared to n. In particular, the oracle inequality of the estimation error is useful in the following sections.

Theorem 3. *Assume condition* $S(3.5, \varepsilon_n, k)$ *and (*H1*)–(*H4*) hold. Let* λ_1 *be chosen by* (2.21) *and* $\lambda_2 \leq \lambda_1/8B$. *Then, under the event* $\mathcal{A} \cap \mathcal{B}$, *we have* $P(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathrm{V}(3.5, \varepsilon_n/2, H)) \geq 1 - 2(2p)^{-A^2}$ *and*

$$P\left\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{2.25^2 \lambda_1 d_H^*}{ak + 2\lambda_2} + \left(1 + \frac{a}{\lambda_1}\right)\varepsilon_n\right\} \geq 1 - 2(2p)^{-A^2}. \tag{2.22}$$

Moreover, let the test data $(\boldsymbol{X}^*, Y^*)$ *be an independent copy of the training data* $(\boldsymbol{X}, Y)$, *and denote* $\mathrm{E}^*(\cdot) := \mathrm{E}(\cdot|\boldsymbol{X}^*)$. *Conditioning on the event* $\mathcal{A} \cap \mathcal{B}$, *the squared prediction error is*

$$\mathrm{E}^*[\boldsymbol{X}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 \leq \frac{17.71875 d_H^* \lambda_1^2}{a(ak + 2\lambda_2)} + \left(\frac{4\lambda_1}{a} + 3.5\right)\varepsilon_n, \tag{2.23}$$

where $a := \min_{\{|x| \leq LM+K, |y| \leq K\}}\{(1/2)\theta \mathrm{e}^x(\mathrm{e}^y + \theta)/[\theta + \mathrm{e}^x]^2\} > 0$.

Comparing with the upper bounds under the Compatibility Factor Condition in Section 2.3, in much the same fashion, we observe that when $d^* = O(1)$, the number of covariates increases by as much as $o(\exp(n))$. Then, the bound on the estimation error is $o(1)$, and the elastic-net estimator ensures the consistent property. Theorem 3 is also an improvement over Lemma 3 from a big neighborhood of $\boldsymbol{\beta}^*$ to the desired small neighborhood of $\boldsymbol{\beta}^*$.

Remark 2. Discussion of the measurement error ε_n when $d_H^* < \infty$:

- 1. If $\varepsilon_n = o(\sqrt{\log p/n})$, then $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq O(\sqrt{\log p/n})$, $\mathrm{E}^*[\boldsymbol{X}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 \leq O(\log p/n)$;
- 2. If $\varepsilon_n = O(\sqrt{\log p/n})$, then $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq O(1)$, but $\mathrm{E}^*[\boldsymbol{X}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^2 \leq O(\sqrt{\log p/n})$;

More typical examples for ε_n are $1/n$ or even zero. Under the restricted condition $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in V(3.5, \varepsilon_n/2, H)$, Case 2 tells us that if the order of fluctuations ε_n is sightly lower than the order of the tuning parameter, elastic-net with $\lambda_2 \leq \lambda_1/8B$ guarantees that the squared prediction error is asymptotically zero, with a lower rate $O(\sqrt{\log p/n})$.

3. Applications of the Oracles Results

We now examine the non-asymptotic and asymptotic results. In this section, the applications are derived from oracle inequalities about the ℓ_1 estimation error, and we assume that the design matrix is fixed, for simplicity.

3.1. Grouping effect from oracle inequality

Zou and Hastie (2005) show that the elastic-net has a grouping effect that asserts that strongly correlated predictors tend to be in or out of the model together when the coefficients have the same sign. Zhou (2013) proves that the grouping effect of the elastic-net estimates holds without the assumption of the sign. Yu (2010) derives the asymptotical result of the grouping effect for elastic-net estimates of the Cox models. Based on the oracle inequalities we put forward, we provide an asymptotical version of the grouping effect inequality as $p, n \to \infty$ for the fixed design case.

Theorem 4. *Under the assumption of Theorem 2 with $d_H^* < \infty$, suppose that the covariates (nonrandom) are standardized as*

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, \ \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \ \ \textit{for} \ \ j = 1, 2, \ldots, p. \tag{3.1}$$

Denote $\rho_{kl} = (1/n)\sum_{i=1}^{n} x_{ik}x_{il}$ as the correlation coefficient. For any constant $E_s > 0$, with probability at least $1 - 2/p^{r-1} - 2p^2 e^{-nt^2/(2[d_H^* C_{LB}(1+\zeta)L^2]^2)} - \sigma_n^2/nE_s^2$,

(i). $|\hat{\beta}_k - \hat{\beta}_l|^2 \le (1 - \rho_{kl})[Ke^{2LM}O(1) + (1/\lambda_2^2)(E_s + \mu_s)]$;

(ii). If the asymptotic correlation between two random predictors is asymptotically up to one, that is $\rho_{kl} = 1 - o(\lambda_2^2)$, with $\lambda_2^2 = O(\log p/n) \to 0$, we have

$$|\hat{\beta}_k - \hat{\beta}_l| \le \sqrt{o_p(1)[\lambda_2^2 e^{2LM}O(1) + (E + \mu)]}.$$

This grouping effect oracle inequality asserts that if ρ_{kl} tends to one with a high probability, the elastic-net is able to select covariates $k, l \in \{1, 2, \ldots, p\}$ together. Combined with the Lasso sparse estimation, the $\ell_1 + \ell_2$ penalty enables strongly correlated predictors to be in or out simultaneously. In addition to the sparse estimation, intuitively, highly related covariates should have similar regression coefficients, but the Lasso cannot select them simultaneously.

3.2. Sign consistency

Sign consistency indicates whether an estimate is good, relating tp the estimated sign of the coefficient. A few researchers have studied the sign consistency property of the elastic-net. One condition for sign consistency is the *Irrepresentable Condition* (IC). Zhao and Yu (2006) explore the IC for the sign consistency of a linear regression under a Lasso penalty. Moreover, the model selection consistency of elastic-net is studied by Jia and Yu (2010), following

Zhao and Yu (2006). Along the same line, for the elastic-net penalized Cox model, Yu (2010) investigates the selection consistency. Here, the basic idea is that the KKT condition is a necessary and sufficient condition for the global minimizer of the target function. We focus on the elastic-net penalized NBR model's selection consistency based on some reasonable assumptions in a similar fashion. It is interesting to see that under the bounded covariates assumption, we do not need the IC, which is assumed in Yu (2010) and Lv et al. (2018). We rely only on the assumptions in Theorem 2.

Uniform Signal Strength Condition.

$$\beta_* := \min_{j\in H} |\beta_j^*| \geq \frac{e^{2a_\tau}(\zeta+1)d_H^*\lambda_1}{2C^2(\zeta,H)},$$

with $\lambda_1 = O(\sqrt{\log p/n}), B\lambda_2 = B_1\lambda_1$.

Assume $d_H^* < \infty$. Zhang and Zhang (2014) points out that the selection consistency theory characteristically necessitates a *uniform signal strength condition* (or *beta-min condition*) that the smallest nonzero regression coefficients $\beta_* := \min\{|\beta_j| : j \in H\}$ should be greater in size than a thresholded level $O(\sqrt{\log p/n})$. When β_* is less than the level, the presence of weak signals cannot be detected by statistical inferences procedures.

Theorem 5. *Suppose that the uniform signal strength condition and the assumptions of Theorem* 2 *hold. Let* $\lambda_1 = O(\sqrt{\log p/n}), d_H^* < \infty$. *Then, for* $\sqrt{\log p/n} = o(1)$ *and a suitable tuning parameter* r *in Theorem 2, we have the following sign consistency:*

$$\lim_{n,p\to\infty} P(\text{sign}\hat{\boldsymbol{\beta}} = \text{sign}\boldsymbol{\beta}^*) = 1. \tag{3.2}$$

3.3. Honest variable selection and detection of weak signals

As a particular case of the random design in Section 2.4, we focus on the fixed design in this section, where the $\{\boldsymbol{X}_i\}_{i=1}^n$ is deterministic.

Recall that $\hat{H} := \{j : \hat{\beta}_j \neq 0\}$; thus $\hat{H}$ is an estimator of the true variable set $H := \{j : \beta_j \neq 0\}$ (or the set of positives). Let δ_1, δ_2 be constants such that $P(\hat{H} \not\subset H) \leq \delta_1, P(H \not\subset \hat{H}) \leq \delta_2$. Then we have $P(H \neq \hat{H}) \leq P(\hat{H} \not\subset H) + P(H \not\subset \hat{H}) \leq \delta_1 + \delta_2$. If we treat H as the null hypothesis, $P(\hat{H} \not\subset H)$ is often called the false positive rate in the language of ROC curves (or type I error in statistical hypothesis testing; the estimate is $\hat{H}$ but it makes the decision $\hat{H} \subset H^c$); $P(H \not\subset \hat{H})$ is often called the false negative rate (or type II error).

Thus, the probability of correct subset selection under some random events W (the assumptions hold with probability $P(W)$) is

$$P(H = \hat{H}) \geq P(W) - \delta_1 - \delta_2. \tag{3.3}$$

From the ℓ_1 estimation error obtained in Theorem 3, we easily bound the false negative rate $P(H \not\subset \hat{H})$ in Proposition 3. However, the upper bound of the false positive rate $P(\hat{H} \not\subset H)$ cannot be obtained directly, additional assumptions on the covariates correlation are required.

Proposition 3. *Let $\delta \in (0,1)$ be a fixed number, and let the assumption of Theorem 3 be satisfied. The weakest and strongest signal meet the condition: $B_0 := 2.25^2\lambda_1 d_H^*/(ak + 2\lambda_2) + (1 + a/\lambda_1)\varepsilon_n \leq \min_{j\in H}|\beta_j^*| \leq B$. If $p = \exp\{(1/(A^2-1))\log(2^{1-A^2})/\delta\}$, with $A > 1$, then*

$$P(H \subset \hat{H}) \geq P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq B_0) \geq 1 - \frac{\delta}{p}.$$

Note that the lower bound we have derived may be too large in some settings. For example, this may occur if d_H^* is as large as $\lambda_1 d_H^* = O(1)$ and $\min_{j\in H}|\beta_j^*| \geq 2.25^2 O(1)/ak + 2\lambda_2 =: D$, where D is also a moderately large constant compared with the strongest signal threshold B. Then, we can only detect a few parts of the overall signals. To deal with this problem, we use a new approach (inspired by Section 3.1.2 in Bunea (2008)) to find a constant-free weakest signal detection threshold that relies only on the tuning parameter λ_1. Under some mild conditions on the design matrix, we show that the lower bounds can be sharpen considerably.

First, we assume that the covariates are centered and standardized as in (3.1). This crucial method of processing covariates is also employed when studying the grouping effect in Section 3.1. Second, let $\rho_{kl} = (1/n)\sum_{i=1}^n X_{ik}X_{il}$, for $k, l \in \{1, 2, \ldots, p\}$ be the correlation constants between covarates k and l. For a constant $h \in (0,1)$, we have the following condition.

Identifiable Condition: $\max_{k,l\in H, k\neq l}|\rho_{kl}| \leq h/\theta d_H^*$, $(\theta/n)\sum_{i=1}^n X_{ik}^2 = 1$.

This assumption of a maximal correlation constant of two distinct covariates on the true set H measures the dependence structure using a constant h in the whole predictor. A lower h indicates a higher degree of separation, which makes it easier to detect weak signals. Bunea (2008) explained the intuition as follows:" If the signal is very weak and the true variables are highly correlated with one another and with the rest, one cannot hope to recover the true model with high

probability". Interestingly, the grouping effect states that the elastic-net is able to simultaneously estimate highly correlated true variables, and this grouping effect is valid without the premise that the signal is enough strong. If both signals are faint under the level of the detection bounds, then the elastic-net estimates are both zero, and the grouping effect is also true.

Additionally, we require two conditions because we have to build some connections between $P(H \not\subset \hat{H}), P(\hat{H} \not\subset H)$ and the ℓ_1-estimation error in Theorem 3. Let a_i (b_i) be the intermediate point between $\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_i^T\boldsymbol{\beta}^*$, by the first-order Taylor expansion of the function $f(t) = (e^t/\theta + e^t)$ ($g(t) = 1/(\theta + e^t)$), and $L_1, L_2 \in [1, \infty)$. By (H.1)–(H.3), it leads to for all i,

$$\begin{aligned} |a_i| \text{ or } |b_i| &\le |\boldsymbol{X}_i^{*T}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| + |\boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| \\ &\le |\boldsymbol{X}_i^{*T}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*| + |\boldsymbol{X}_i^{T}\boldsymbol{\beta}^*| \le L(M+B). \end{aligned}$$

Next, we pose some weighted correlation conditions (WCC):

Weighted Correlation Condition (1):

$$\sup_{\substack{k,j\in H,\\ |a_i|\le L(M+B)}} \frac{1}{n}\left(\left|\sum_{i=1}^{n} X_{ij}X_{ik}\frac{\theta^2 e^{a_i}}{(\theta+e^{a_i})^2}\right| \vee \left|\sum_{i=1}^{n}\theta X_{ij}X_{ik}\left(1-\frac{\theta e^{a_i}}{(\theta+e^{a_i})^2}\right)\right|\right) \le \frac{hL_1}{d_H^*}.$$

Weighted Correlation Condition (2) holds with a high probability:

$$P\left(\sup_{\substack{k,j\in H,\\ |b_i|\le L(M+B)}} \left|\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ik}X_{ij}Y_i\cdot\theta^2 e^{b_i}}{(\theta+e^{b_i})^2}\right| \le \frac{hL_2}{d_H^*}\right) = 1-\varepsilon_{n,p},$$

where $\varepsilon_{n,p}$ is a constant satisfying $\lim_{n,p\to\infty}\varepsilon_{n,p} = 0$.

By (H.1) and (H.2), a_i, b_i are uniformly bounded random variables, and are viewed as ignorable constants in an asymptotic analysis, as are $\theta e^{a_i}/(\theta + e^{a_i})^2$ and $(1-\theta e^{a_i}/(\theta + e^{a_i})^2)$. We can check WCC(2) using a similar approach to that if the concentration phenomenon for the suprema of the multiplier empirical processes. The conditions above can be obtained by taking a linear transformation of the covariates, that is, by scaling the covariates. WCC(1) is a technical condition used by Bunea (2008) for the case of a logistic regression. This assumption means that the maximum weighted-correlation version of ρ_{kl} ($k \ne l$) is less than $hL_1/\theta d_H^*$. However, the NBR is more complex than a logistic regression since its Hessian matrix depends on random responses; thus WCC(2) should be assumed with a high probability.

We now have the following constant-free weakest signal detection threshold

for correct subset selection.

Theorem 6. *If the assumptions in Theorem* 3 *hold with* $\varepsilon_n = 0$*, under the identifiable condition, WCC(*1,2*) with* $h \le a + 2\lambda_2/(20.25L_i + 8a) \wedge 1/8$*, for* $i = 1, 2$*. Let* $p = \exp\{1/(1-A^2)\log(2^{A^2-1}\delta)\}$*,*

$$P(H = \hat{H}) \ge 1 - 2\left(1 + \frac{d_H^*}{p}\right)\delta - 2pe^{-n\lambda_1^2/32C_{LB}^2L^2} - \varepsilon_{n,p},$$

provided that the minimal signal condition $\min_{j\in H}|\beta_j^*| \ge 2\lambda_1$ *is satisfied.*

3.4. De-biased elastic-net and confidence interval

Introduced by Zhang and Zhang (2014), the de-biased Lasso was further studied in van de Geer et al. (2014) and Janková and van de Geer (2016) within some generalized linear models. Following the the de-biasing idea, we deal with the de-biased estimator $\hat{\boldsymbol{b}} =: \hat{\boldsymbol{\beta}} - \hat{\Theta}\dot{\ell}(\hat{\boldsymbol{\beta}})$, which is asymptotically normal, based on the established oracle inequality in Section 2. Let $\hat{\boldsymbol{\beta}}$ be defined as in optimization problem (1.1). Let $\hat{\Theta}$ be an approximated estimator of the inverse of the Hessian $-\ddot{\ell}(\boldsymbol{\beta}^*)$(e.g., the CLIME or nodewise Lasso estimator for the estimated Hessian matrix). If $\dot{\ell}(\hat{\boldsymbol{\beta}})$ is continuously differentiable, by Taylor's expansion of vector-valued functions, we have

$$\begin{aligned}
\dot{\ell}(\boldsymbol{\beta}^*) &= \dot{\ell}(\hat{\boldsymbol{\beta}}) - \ddot{\ell}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - r(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2) \\
&= \ddot{\ell}(\boldsymbol{\beta}^*)[\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} - \ddot{\ell}(\boldsymbol{\beta}^*)^{-1}\dot{\ell}(\hat{\boldsymbol{\beta}})] - r(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2) \\
&= \ddot{\ell}(\boldsymbol{\beta}^*)[\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} + \hat{\Theta}\dot{\ell}(\hat{\boldsymbol{\beta}})] - \ddot{\ell}(\boldsymbol{\beta}^*)[\ddot{\ell}(\boldsymbol{\beta}^*)^{-1} + \hat{\Theta}]\dot{\ell}(\hat{\boldsymbol{\beta}}) - r(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2) \\
&=: \ddot{\ell}(\boldsymbol{\beta}^*)[\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} + \hat{\Theta}\dot{\ell}(\hat{\boldsymbol{\beta}})] + R_n,
\end{aligned}$$

where $r(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2) = o_p(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2)$ is a vector-valued function.

Include $\sqrt{n}\hat{\Theta}$ in the equation above if $\sqrt{n}R_n = o_p(1)$. Then

$$\sqrt{n}(\hat{\boldsymbol{b}} - \boldsymbol{\beta}^*) \approx \hat{\boldsymbol{\Theta}}[\sqrt{n}R_n - \sqrt{n}\dot{\ell}(\boldsymbol{\beta}^*)] \xrightarrow{d} N(0, \hat{\boldsymbol{\Theta}}\boldsymbol{\Sigma}\hat{\boldsymbol{\Theta}}^T)$$

where the notation $\approx$ means asymptotic equivalence under some regular conditions. Here, $\boldsymbol{\Sigma}$ is the asymptotic variance of $\sqrt{n}\dot{\ell}(\boldsymbol{\beta}^*)$, where $\mathrm{Var}\dot{\ell}(\boldsymbol{\beta}^*) = (1/n)\sum_{i=1}^n \theta e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*}/(\theta + e^{\boldsymbol{X}_i^T\boldsymbol{\beta}^*})\boldsymbol{X}_i\boldsymbol{X}_i^T$. We can subsitute in a consistent estimator for $\boldsymbol{\Sigma}$ in the high-dimensional case.

The asymptotic confidence level of $1-\alpha$ for β_j^* is then given by

$$\big[\hat{b}_j - c(\alpha, n, \sigma), \hat{b}_j + c(\alpha, n, \sigma)\big], \quad c(\alpha, n, \sigma) := \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{(\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Theta}}^T)_{j,j}}{n}},$$

where $\Phi(\cdot)$ denotes the c.d.f. of $N(0,1)$.

By the KKT conditions in Lemma 1, the de-biased elastic-net estimator is expressed as

$$\hat{\boldsymbol{b}} = \hat{\boldsymbol{\beta}} - \hat{\Theta}\dot{\ell}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}(\mathrm{I}_p - 2\lambda_2\hat{\Theta}) - \hat{\Theta}\lambda_1 \mathrm{sign}(\hat{\boldsymbol{\beta}}).$$

A theoretical analysis of the de-biased elastic-net estimator (includeing precision matrix estimation, confidence interval, and hypothesis testing) is beyond the scope of the this study, please refer to the proofs in Janková and van de Geer (2016) for some additional details. A simulation study for the de-biased elastic-net is presented in Appendix S4, showing that the de-biased elastic-net has less bias than that of the de-biased Lasso. In the simulation, it is important to estimate the nuisance parameter θ and estimate the inverse of the Hessian.

4. Conclusions

We study sparse high-dimensional NBR problems using several consistency results, such as prediction or ℓ_q-estimation error bounds. NBRs are widely used in modeling count data. We show that under a few conditions, the elastic-net estimator has oracle properties, which means that when the sample size is large enough, our sparse estimator is very close to the true parameter if the tuning parameters are properly chosen. We also show the sign consistency property under the beta-min condition. We discuss the detection of weak signals, and give a constant-free weakest signal threshold for correct subset selection under some correlation conditions of the covariates. The asymptotic normality of the de-biased elastic-net estimator is also discussed, although doing so further is beyond the scope of this study. These results provide a theoretical understanding of the proposed sparse estimator and provide practical guidance for the use of the elastic-net estimator.

Note that the oracles inequalities in Sections 2.4 and 3 can be extended to many ℓ_1 or $\ell_1 + \ell_2$ regularized M-estimation regressions with the corresponding empirical process (2.19) having stochastic Lipschitz properties as presented in Proposition 1. For example, the analysis of the stochastic Lipschitz properties of the average negative log-likelihood empirical process can be employed to elastic-net or Lasso penalized COM-Poisson regressions (see Sellers and Shmueli (2008)). As shown in the simulation, the two-step estimation of $\hat{\theta}$ is not well behave. Like the misspecified models in Example 5.25 of van der Vaart (1998), θ, which is a nuisance parameter, is not an important estimate in the consistency results. It would be interesting and important to find a better estimator of θ in the further research, because θ is a crucial quantization when constructing confidence

interval.

Supplementary Material

All proofs and simulation results are in the Supplementary Material.

Acknowledgments

We are grateful for the kind assistance of Xiaoxu Wu. The authors would like to thank the anonymous referees for their valuable comments. The authors also thank Prof. Cun-Hui Zhang, Prof. Fang Yao and Dr. Sheng Fu for their helpful discussions. This work was partially supported by the National Science Foundation of China (11571021).

References

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability* **13**, 1000–1034.

Balabdaoui, F., Jankowski, H., Rufibach, K. and Pavlides, M. (2013). Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 769–790.

Bickel, P. J., Ritov, Y. A. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

Blazere, M., Loubes, J. M. and Gamboa, F. (2014). Oracle inequalities for a group Lasso procedure applied to generalized linear models in high dimension. *IEEE Transactions on Information Theory* **60**, 2303–2318.

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

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1+\ \ell_2$ penalization. *Electronic Journal of Statistics* **2**, 1153–1194.

Cameron, A. C. and Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics* **46**, 347–364.

Chi, Z. (2010). A local stochastic Lipschitz condition with application to Lasso for high dimensional generalized linear models. *arXiv preprint arXiv:1009.1052*.

Hilbe, J. M. (2011). *Negative Binomial Regression*. 2nd Edition. Cambridge University Press.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C. H. (2013). Oracle inequalities for the Lasso in the Cox model. *Annals of Statistics* **41**, 1142–1165.

Janková, J. and van de Geer, S. (2016). Confidence regions for high-dimensional generalized linear models under sparsity. *arXiv:1610.01353*.

Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Univariate Discrete Distributions*. 3rd Edition. Wiley.

Jia, J. and Yu, B. (2010). On model selection consistency of the Elastic Net when $p \gg n$. *Statistica Sinica* **20**, 595–611.

Kakade, S., Shamir, O., Sindharan, K. and Tewari, A. (2010). Learning exponential families in high-dimensions: Strong convexity and sparsity. In *International Conference on Artificial*

Intelligence and Statistics, 381–388.

Lv, S., You, M., Lin, H., Lian, H. and Huang, J. (2018). On the sign consistency of the Lasso for the high-dimensional Cox model. *Journal of Multivariate Analysis* **167**, 79–96.

Massaro, T. J. (2016). *Variable Selection via Penalized Regression and the Genetic Algorithm Using Information Complexity, with Applications for High-Dimensional-Omics Data*. PhD Dissertations, University of Tennessee.

Nielsen, F. and Nock, R. (2009). Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory* **55**, 2882–2904.

Qiu, Y., Chen, S. X. and Nettleton, D. (2018). Detecting rare and faint signals via thresholding maximum likelihood estimators. *The Annals of Statistics* **46**, 895–923.

Rauschenberger, A., Jonker, M. A., van de Wiel, M. A. and Menezes, R. X. (2016). Testing for association between RNA-Seq and high-dimensional data. *BMC Bioinformatics* **17**, 118.

Rigollet, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics* **40**, 639–665.

Sørensen, Ø., Hellton, K. H., Frigessi, A. and Thoresen, M. (2018). Covariate selection in high-dimensional generalized linear models with measurement error. *Journal of Computational and Graphical Statistics* **27**, 739–749.

Sellers, K. F. and Shmueli, G. (2008). A flexible regression model for count data. *The Annals of Applied Statistics* **4**, 943–961.

Tutz, G. (2011). *Regression for Categorical Data*. Cambridge University Press.

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

van de Geer, S. A. (2007). The deterministic Lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics* **36**, 614–645.

van de Geer, S. A., Bühlmann, P., Ritov, Y. A. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42**, 1166–1202.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. 3rd Edition. Cambridge University Press.

Wang, Z., Ma, S., Zappitelli, M., Parikh, C., Wang, C. Y. and Devarajan, P. (2016). Penalized count data regression with application to hospital stay after pediatric cardiac surgery. *Statistical Methods in Medical Research* **25**, 2685–2703.

Weißbach, R. and Radloff, L. (2020). Consistency for the negative binomial regression with fixed covariate. *Metrika* **83**, 627–641.

Ye, F. and Zhang, C. H. (2010). Rate Minimaxity of the Lasso and Dantzig Selector for the l_q Loss in l_r Balls. *Journal of Machine Learning Research* **11**, 3519–3540.

Yu, Y. (2010). High-dimensional Variable Selection in Cox Model with Generalized Lasso-type Convex Penalty. `https://warwick.ac.uk/fac/sci/statistics/staff/academic-research/yu/Cox_generalized_convex.pdf`

Yu, Y., Bradic, J. and Samworth, R. J. (2021). Confidence intervals for high-dimensional Cox models. *Statistica Sinica* **31**, 243–267.

Zhang, C. H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 217–242.

Zhang, H., Liu, Y. and Li, B. (2014). Notes on discrete compound Poisson model with applications to risk theory. *Insurance: Mathematics and Economics* **59**, 325–336.

Zhang, H. and Chen, S. X. (2021). Concentration inequalities for statistical inference. *Communications in Mathematical Research* **37**, 1–85.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research* **7**, 2541–2563.

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

Zhou, D. X. (2013). On grouping effect of elastic net. *Statistics & Probability Letters* **83**, 2108–2112.

Huiming Zhang

School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, P.R. China.

Department of Mathematics, Faculty of Science and Technology, University of Macau, Taipa, Macau, P.R. China.

E-mail: zhanghuiming@pku.edu.cn

Jinzhu Jia

School of Public Health and Center for Statistical Science, Peking University, Beijing 100191, P.R. China.

E-mail: jzjia@math.pku.edu.cn

(Received May 2018; accepted June 2020)