Statistica Sinica **28** (2018), 1157-1178
doi:https://doi.org/10.5705/ss.202015.0335

VARIABLE SELECTION IN SPARSE REGRESSION WITH QUADRATIC MEASUREMENTS

Jun Fan[1], Lingchen Kong[1], Liqun Wang[2] and Naihua Xiu[1]

[1]*Beijing Jiaotong University and* [2]*University of Manitoba*

Abstract: Regularization methods for high-dimensional variable selection and estimation have been intensively studied in recent years and most of them are developed in the framework of linear regression models. However, in many problems, e.g., in compressive sensing, signal processing and imaging, the response variables are nonlinear functions of the unknown parameters. In this paper we introduce a so-called quadratic measurements regression model that extends the usual linear model. We study the ℓ_q regularized least squares method for variable selection and establish the weak oracle property of the corresponding estimator. Moreover, we derive a fixed point equation and use it to construct an efficient algorithm for numerical optimization. Numerical examples are given to demonstrate the finite sample performance of the proposed method and the efficiency of the algorithm.

Key words and phrases: ℓ_q-regularization, moderate deviation, optimization algorithm, sparsity, weak oracle property.

1. Introduction and Motivation

In the era of big data, more and more massive and high-dimensional data become available in such fields, as genome and health science, economics and finance, astronomy and physics, signal processing and imaging, etc. The large size and high dimensionality of data pose significant challenges to the traditional statistical methodologies, see, e.g., Donoho (2000) and Fan and Lv (2010) for excellent overviews. As pointed out by these authors, a common feature in high-dimensional data analysis is the sparsity of the predictors and one of the main goals is to select the most relevant variables to accurately predict a response variable of interest.

Various regularization methods have been proposed in the literature, e.g., bridge regression (Frank and Friedman (1993)), the LASSO (Tibshirani (1996)), the SCAD and other folded-concave penalties (Fan and Li (2001)), the Elastic-Net penalty (Zou and Hastie (2005)), the adaptive LASSO (Zou (2006)), the group LASSO (Yuan and Lin (2006)), the Dantzig selector (Candes and Tao

(2007)), and the MCP (Zhang (2010)). Recently, Lv and Fan (2009) pointed out that there is a distinction and close relation between the model selection problem in statistics and sparse recovery problem in compressive sensing and signal processing. Moreover, they proposed a unified approach to deal with both problems.

Most existing statistical methods for variable selection are developed in the context of sparse linear regression. On the other hand, there is a large number of problems, especially in compressive sensing, signal processing and imaging, and statistics, where the regression relationships are in nonlinear forms of unknown parameters.

Example 1. Compressive sensing has been intensively studied in the last decade and the main goal is to reconstruct sparse signals from the observations. Recently, the theory has been extended to nonlinear compressive sensing and, in particular, to the so-called quadratic compressive sensing that aims to find the sparse signal β to the problem $\min_{\beta\in\mathbb{R}^p}\|\beta\|_0$ subject to $y_i = \beta^T Z_i\beta + x_i^T\beta + \varepsilon_i, i = 1,\ldots,n$, where $\|\beta\|_0$ is the number of nonzero entries of β, $y_i, \varepsilon_i \in \mathbb{R}, x_i \in \mathbb{R}^p$ and $Z_i \in \mathbb{R}^{p\times p}$ are real matrices (vectors). For more details see, e.g., Beck and Eldar (2013), Blumensath (2013), and Ohlsson et al. (2014).

There is a special class of problems in optical imaging, where partially spatially incoherent light such as sub-wavelength optical results in a quadratic relationship between the input object β and image intensity y_i as $y_i \approx \beta^T Z_i\beta, i = 1,\ldots,n$, where Z_i is known from the mutual intensity and the impulse response function of the optical system (Shechtman et al. (2011) and Shechtman et al. (2012)).

Example 2. Phase retrieval plays an important role in X-ray crystallography, transmission electron microscopy, coherent diffractive imaging, etc. Generally speaking, the problem is to recover the lost phase information through the observed magnitudes. In particular, in the real phase retrieval problem the goal is to find $\beta \in \mathbb{R}^p$ in $y_i = \beta^T(z_i z_i^T)\beta + \varepsilon_i, i = 1,\ldots,n$, where $z_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ are observed variables and ε_i are random errors (Candes, Strohmer and Voroninski (2013), Candes, Li and Soltanolkotabi (2015), Eldar and Mendelson (2014), Lecué and Mendelson (2015), Netrapalli, Jain and Sanghavi (2015), Cai, Li and Ma (2016)).

Example 3. In wireless ad hoc and sensor networks, localization is crucial for building low-cost, low-power and multi-functional sensor networks in which direct measurements of all nodes' locations via GPS or other similar means are

not feasible (Biswas and Ye (2004), Meng, Ding and Dasgupta (2008), Wang et al. (2008)). The most important element of any localization algorithms is to measure the distances between sensors and anchors. However, the acquired data are usually imprecise because of the measurement noise and estimation errors. Suppose p-dimensional vectors $x_1, x_2, \ldots, x_n$ are the known sensor positions and $\beta \in \mathbb{R}^p$ is the signal source location that is unknown and to be determined. Then the measured distance y_i from the source to each sensor node is given by $y_i^2 = \|x_i - \beta\|_2^2 + \varepsilon_i, i = 1, \ldots, n$, where ε_i is a random error. Again, the above relation can be written as $y_i^2 - \|x_i\|_2^2 = \beta^T\beta - 2x_i^T\beta + \varepsilon_i$.

Example 4. Measurement error is ubiquitous in statistical data analysis. Wang (2003, 2004) showed that for a class of measurement error models to be identifiable and consistently estimable, at least the first two conditional moments of the response variable given the observed predictors are needed. Wang and Leblanc (2008) showed that in a general nonlinear model this second-order least squares estimator (SLSE) is asymptotically more efficient than the ordinary least squares estimator when the regression error has nonzero third moment, and the two estimators have the same asymptotic variances when the error term has symmetric distribution. In a linear model, the SLSE is derived based on the first two conditional moments $\mathbb{E}(y_i|x_i) = x_i^T\beta$ and $\mathbb{E}(y_i^2|x_i) = (x_i^T\beta)^2 + \sigma^2, i = 1, \ldots, n$, where β is the vector of regression coefficients and σ^2 is the variance of the regression error. It is easy to see that this second moment can be written as $\mathbb{E}(y_i^2|x_i) = \theta^T Z_i \theta$ with $\theta = (\beta^T, \sigma)^T$ and $Z_i = \begin{pmatrix} x_i x_i^T & 0 \\ 0 & 1 \end{pmatrix}$.

In our examples, the main goal is to recover the sparse signals in regression setups where the response variable is a quadratic function of the unknown parameters, and this not covered by linear regression models. Given their wide applications, however, the high-dimensional variable selection problem in such models has not been studied in statistical literature.

In this paper we attempt to fill in this gap. First, we introduce a so-called quadratic measurements regression (QMR) model as an extension of the usual linear model. Then we study the ℓ_q-regularized least squares (q-RLS) estimation in this model and establish its weak oracle property (Lv and Fan (2009)). Moreover, using moderate deviations we show that the estimators of the nonzero coefficients have an exponential convergence rate. To deal with the problem of numerical optimization, we derive a fixed point equation that is necessary for global optimality. This allows us to construct an iterative algorithm and to establish its

convergence. Finally, we present some numerical examples to demonstrate the efficiency of the proposed method and algorithm.

The rest of this paper is organized as follows. In Section 2 we introduce the quadratic measurements model and the q-RLS estimation. In Section 3 we discuss the weak oracle property of the q-RLS estimator using the moderate deviation technique. In Section 4, we deal with a special case of a purely quadratic measurements model that has applications in some important problems. In Section 5, we derive a fixed point equation and construct an algorithm for numerical minimization. In Section 6, we calculate some numerical examples to illustrate our proposed method and to demonstrate its finite sample performance. Discussions are given in Section 7, while technical lemmas and proofs are given in the Supplementary Material.

2. The Quadratic Measurements Model

We define the *quadratic measurements regression* (QMR) model as

$$y_i = \beta^T Z_i \beta + x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $y_i \in \mathbb{R}$ is the observed response, $x_i \in \mathbb{R}^p$ is the vector of predictors, $Z_i \in \mathbb{S}^{p\times p}$ is a symmetric design matrix, $\beta \in \mathbb{R}^p$ is the vector of unknown parameters, and $\varepsilon_i \in \mathbb{R}$ are independent and identically distributed random errors with mean 0 and variance σ^2. When $Z_i \equiv 0$, this reduces to the usual linear model

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.2}$$

In this paper we are mainly interested in the high-dimensional case where $p > n$ or $p \gg n$, although our theory applies to the case $p \leq n$ as well. Throughout the paper we assume that $\log p = o(n^\varrho)$ for some constant $\varrho \in (0, 1)$ and that there exists a constant $\delta_0 > 0$ such that $\mathbb{E} \exp(\delta_0 |\varepsilon_1|) < \infty$.

In compressive sensing and signal processing the main goal is to identify and estimate the smallest possible number of nonzero coefficients. Thus we consider the problem of estimating unknown parameters of model (2.1) under the sparsity constraint $\|\beta\|_0 \leq s$, where $s < n$ is a certain integer. And accordingly, we study the ℓ_q-regularized least squares (q-RLS) problem

$$\min_{\beta\in\mathbb{R}^p} L_n(\beta) := \ell_n(\beta) + \lambda_n \|\beta\|_q^q, \tag{2.3}$$

where $\ell_n(\beta) = \sum_{i=1}^n (y_i - \beta^T Z_i \beta - x_i^T \beta)^2, \lambda_n > 0$ and $q \in (0, 1)$. The ℓ_q-regularization has been widely used in compressive sensing. Compared to ℓ_1-regularization, this method tends to produce precise signal reconstruction with

fewer measurements (Chartrand (2007)), and increases the robustness to noise and image non-sparsity (Saab, Chartrand and Yilmaz (2008)). Moreover, Krishnan and Fergus (2009) demonstrated very high efficiency of $\ell_{1/2}$ and $\ell_{2/3}$ regularization in image deconvolution.

A minimizer $\hat{\beta}$ of the optimization problem (2.3) is called q-RLS estimator and it is a generalization of the bridge estimator in linear models (Frank and Friedman (1993)). It is well-known that the bridge estimator has various desirable properties including sparsity and consistency (Knight and Fu (2000), Huang, Horowitz and Ma (2008)). A natural question is whether the q-RLS solution of (2.3) continues to enjoy these properties in the more general model. To answer this question, we study the moderate deviation (MD) of $\hat{\beta}$ which gives the rate of convergence to β at a slower rate than $n^{-1/2}$ (Kallenberg (1983)).

Although we are mainly interested in variable selection problem, our results on identifiability and numerical optimization algorithm apply also to the case $q \geq 1$. Our consistency results for selection and estimation hold only for the case where $q \in (0, 1)$; this is not surprising given that it is a well-known fact in linear models (Fan and Li (2001), Zou (2006)).

Throughout the paper we use the following notation. For any d-dimensional vector $v = (v_1, \ldots, v_d)^T$, let $|v| = (|v_1|, \ldots, |v_d|)^T, v^2 = (v_1^2, \ldots, v_d^2)^T, \|v\|_2 = (\sum_{i=1}^d v_i^2)^{1/2}$, $\|v\|_1 = \sum_{i=1}^d |v_i|$, and $\|v\|_\infty = \max\{|v_1|, \ldots, |v_d|\}$. For any set $\Gamma \subseteq \{1, \ldots, d\}$, denote its cardinality by $|\Gamma|$ and $\Gamma^c = \{1, \ldots, d\}/\Gamma$. For any $n \times d$ matrix $A = [a_{ij}]$, let$\|A\|_F = \sqrt{\{\sum_{i=1}^n \sum_{j=1}^d a_{ij}^2\}}$ and $|A|_\infty = \max_{1 \leq i,j \leq d} |a_{ij}|$. Denote by A_Γ the sub-matrix of A consisting of its columns associated with index set $\Gamma \subseteq \{1, \ldots, d\}$, $A^{\Gamma'}$ the sub-matrix of A consisting of its rows indexed by $\Gamma' \subseteq \{1, \ldots, n\}$ and by $A^{\Gamma'\Gamma}$ the sub-matrix consisting of the rows and columns of A indexed by Γ' and Γ respectively. We use the notation v_Γ for a column or a row vector v. Denote by $e_{d,j}$ the jth column of the $d \times d$ identity matrix I_d.

3. Weak Oracle Property

In this section we discuss the moderate deviation and consistency of the q-RLS estimators. Let β^* be the true parameter value of model (2.1) and $\Gamma^* = \text{supp}(\beta^*) := \{j : e_{p,j}^T \beta^* \neq 0, j = 1, \ldots, p\}$. Without loss of generality, let $|\Gamma^*| = s < n$. Let $X = (x_1, \ldots, x_n)^T$, where $x_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, \ldots, n$. Then following Huang, Horowitz and Ma (2008), we assume that there exist constants $0 < \underline{c} \leq \overline{c} < \infty$ such that

$$\underline{c} \leq \min\{|e_{p,j}^T \beta^*|, j \in \Gamma^*\} \leq \max\{|e_{p,j}^T \beta^*|, j \in \Gamma^*\} \leq \overline{c}.$$

Following the literature (e.g., Zou and Hastie (2005), Huang, Horowitz and Ma (2008), Fan, Fan and Barut (2014)), the data are assumed to be standardized so that

$$\sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij} = 0,\ \max\left(\sum_{i=1}^{n} x_{ij}^2, \sum_{i=1}^{n} |Z_i|_\infty^2\right) = n,\ j = 1, \ldots, p. \tag{3.1}$$

In the linear model, the third equality above reduces to $\sum_{i=1}^{n} x_{ij}^2 = n$.

3.1. Identifiability of β^*

For the sparse linear model, Donoho and Elad (2003) introduced the concept of spark and showed that the uniqueness of β^* can be characterized by spark(X) which is defined as the minimum number of linearly dependent columns of the design matrix X. Another way to express this property is via the s-regularity of X, any s columns of X are linearly independent. Indeed, X is s-regular if and only if spark$(X) \geq s + 1$, (Beck and Eldar (2013)). Further, in the linear model, $-X$ is the Jacobian matrix of the residual function $R(\beta) = y - X\beta$, where $y = (y_1, \ldots, y_n)^T$. Correspondingly, under model (2.1) the residual function is $R(\beta) = \big(R_1(\beta), \ldots, R_n(\beta)\big)^T$ with $R_i(\beta) = y_i - \beta^T Z_i \beta - x_i^T \beta$ and hence the Jacobian is $\big(-2Z_1\beta - x_1, \ldots, -2Z_n\beta - x_n\big)^T$.

Definition 1. *The affine transform $\mathcal{A}(\beta) = \big(Z_1\beta + x_1, \ldots, Z_n\beta + x_n\big)^T$ is said to be uniformly s-regular, if $\mathcal{A}(\beta)_\Gamma$ has full column rank for any $\Gamma \subseteq \{1, \ldots, p\}$ with $|\Gamma| = s$ and $\beta \in \mathbb{R}^p$ with supp$(\beta) \subseteq \Gamma$.*

Obviously, the uniform s-regularity of $\mathcal{A}(\beta)$ implies the s-regularity of X. It is straightforward to verify that $\mathcal{A}(\beta)$ is uniformly s-regular if and only if the submatrix $\mathcal{A}_\Gamma(\beta_1) = \big(Z_1^{\Gamma\Gamma}\beta_1, \ldots, Z_n^{\Gamma\Gamma}\beta_1\big)^T + X_\Gamma$ has full column rank for any index set $\Gamma \subseteq \{1, \ldots, p\}$ with $|\Gamma| = s$ and $\beta_1 \in \mathbb{R}^s$.

In the linear model, we have $\mathcal{A}_\Gamma(\beta_1) = X_\Gamma$ since $Z_i \equiv 0$, and therefore the uniform s-regularity of $\mathcal{A}(\beta)$ reduces to the s-regularity of X. On the other hand, if $Z_i \equiv I_p$ as in Example 3, then $\mathcal{A}(\beta)$ is uniformly s-regular provided $\sum_{i=1}^{n} x_i = 0$ and X is s-regular.

Proposition 1. *If $\bar{y}_i = \beta^{*T} Z_i \beta^* + x_i^T \beta^*$, $i = 1, \ldots, n$, the system of equations $\beta^T Z_i \beta + x_i^T \beta = \bar{y}_i$, $i = 1, \ldots, n$, has a unique solution β^* satisfying $\|\beta^*\|_0 \leq s$ if $\mathcal{A}(\beta)$ is uniformly $2s$-regular.*

3.2. Moderate deviation and consistency

Strong convexity is the standard condition for the existence of unique so-

lution to a convex optimization problem. When the objective function is twice differentiable, an equivalent condition is that the Hessian is uniformly positive definite. To establish the consistency of an M-estimator in high-dimension, Negahban et al. (2012) introduced the concept of the restricted strong convexity when the objective function is strongly convex on a certain set. To achieve the accuracy of a greedy method for the sparsity-constrained optimization problem, Bahmani, Raj and Boufounos (2013) used stable restricted Hessian to characterize the curvature of the loss function over the sparse subspaces that can be bounded locally from above and below such that the corresponding bounds have the same order. However, the calculation of the exact Hessian of our model is costly. The transform $\mathcal{A}(\beta)$ has a special structure that allows us to not only use the Jacobian to obtain the gradient $\nabla \ell_n(\beta) = -2\mathcal{A}(2\beta)^T R(\beta)$, but also to employ it to approximate the Hessian near β^*. We need some conditions.

Condition 1 (Uniformly Stable Restricted Jacobian).

(a) For any $\Gamma \subseteq \{1, \ldots, p\}$ with $|\Gamma| = s$ and $\beta \in \mathbb{R}^p$ satisfying $\operatorname{supp}(\beta) \subseteq \Gamma$, there exists a positive constant c_1 that bounds all eigenvalues of $n^{-1}\big(\big(\mathcal{A}(\beta)_\Gamma\big)^T \mathcal{A}(\beta)_\Gamma\big)$ from below.

(b) For any $\Gamma \subseteq \{1, \ldots, p\}$ with $|\Gamma| = s$ and $\beta \in \mathbb{R}^p$ satisfying $\operatorname{supp}(\beta) \subseteq \Gamma$ and $\|\beta\| \le \Big(2\overline{c} + 3\sqrt{(\sigma^2+1)/c_1}\Big)\sqrt{s}$, there exists a positive constant c_2 that bounds all eigenvalues of $n^{-1}\big(\big(\mathcal{A}(\beta)_\Gamma\big)^T \mathcal{A}(\beta)_\Gamma\big)$ from above.

It is easy to see that (a) and (b) are respectively equivalent to the following.

(a′) For any $\Gamma \subseteq \{1, \ldots, p\}$ with $|\Gamma| = s$ and $\beta_1 \in \mathbb{R}^s$, there exists a positive constant c_1 that bounds all eigenvalues of $n^{-1}\big(\mathcal{A}_\Gamma(\beta_1)^T \mathcal{A}_\Gamma(\beta_1)\big)$ from below.

(b′) For any $\Gamma \subseteq \{1, \ldots, p\}$ with $|\Gamma| = s$ and $\beta_1 \in S := \{u \in \mathbb{R}^s : \|u\| \le \big(2\overline{c} + 3\sqrt{(\sigma^2+1)/c_1}\big)\sqrt{s}\}$, there exists a positive constant c_2 that bounds all eigenvalues of $n^{-1}\big(\mathcal{A}_\Gamma(\beta_1)^T\ \mathcal{A}_\Gamma(\beta_1)\big)$ from above.

In the linear model (2.2), Condition 1 reduces to the first assumption of Condition 2 in Fan, Fan and Barut (2014). For the general case, (a′) is similar to the restricted strong convexity in Negahban et al. (2012). Indeed, the minimization problem (2.3) is derived from the original optimization problem $\min_{\beta \in \mathbb{R}^p} \ell_n(\beta)$ subject to $\|\beta\|_0 \le s$. So, we first consider the unconstrained optimization problem

$$\min_{\beta_1 \in \mathbb{R}^s} \frac{1}{n}\tilde{\ell}_n(\beta_1) := \frac{1}{n}\sum_{i=1}^{n} \big(y_i - \beta_1^T Z_i^{\Gamma^*\Gamma^*}\beta_1 - x_{i\Gamma^*}^T \beta_1\big)^2 \tag{3.2}$$

that is clearly non-convex and may not have a unique solution in general. However, one can calculate the Hessian matrix of the objective function $(1/n)\tilde{\ell}_n(\beta_1)$ at $\beta^*_{\Gamma^*}$ as $\nabla^2\big\{(1/n)\tilde{\ell}_n(\beta^*_{\Gamma^*})\big\} = (2/n)\big\{\mathcal{A}_{\Gamma^*}(2\beta^*_{\Gamma^*})^T\mathcal{A}_{\Gamma^*}(2\beta^*_{\Gamma^*})\big\} - (4/n)\sum_{i=1}^n \varepsilon_i Z_i^{\Gamma^*\Gamma^*}$.

Since $\|Z_i^{\Gamma^*\Gamma^*}\|_F^2 \le s^2|Z_i|_\infty$, the third equality of (3.1) implies that $\sum_{i=1}^n \|Z_i\|_F^2 \le ns^2$. Further, it follows from Chebyshev's inequality and $s = o(\sqrt{n})$ that $n^{-1}\|\sum_{i=1}^n \varepsilon_i Z_i^{\Gamma^*\Gamma^*}\|_F \xrightarrow{\mathbb{P}} 0$, as $n \to \infty$. Hence Condition 1 (a′) ensures that the Hessian matrix $\nabla^2\big(n^{-1}\tilde{\ell}_n(\beta^*_{\Gamma^*})\big)$ is strictly positive definite and therefore (3.2) has an unique solution in a neighborhood of $\beta^*_{\Gamma^*}$ with probability approaching one. It follows that the minimization problem (2.3) may have a unique solution in a neighborhood of β^*, as in Negahban et al. (2012). Moreover, (a′) implies that $\mathcal{A}_\Gamma(\beta_1)$ has full column rank for any Γ with $|\Gamma| = s$ and therefore $\mathcal{A}(\beta)$ is uniformly s-regular.

Further, (b′) is similar to the upper bound of the stable restricted Hessian. In particular, if s is finite, then (b′) implies that the curvature of the loss function has upper bounds at locations that are within a neighbourhood of the origin. From the proof in Appendix A, one can see that (b′) ensures a more accurate convergent rate.

Condition 2 (Asymptotic Property of Design Matrix). Let $\kappa_{1n} = |X|_\infty$ and $\kappa_{2n} = \max_{1\le i\le n}|Z_i|_\infty$ be such that, as $n \to \infty$,

$$\frac{\kappa_{1n}\sqrt{s}}{\sqrt{n}} \to 0, \qquad \frac{\kappa_{2n}s^{3/2}}{\sqrt{n}} \to 0. \tag{3.3}$$

The first convergence in (3.3) is the same as in Fan, Fan and Barut (2014, Condition 2). The second convergence in (3.3) and (3.5), below, are required to deal with the quadratic term in the low-dimensional space $\mathbb{R}^s$.

Condition 3 (Partial Orthogonality). For any $\Gamma \subseteq \{1,\dots,p\}$ with $|\Gamma| = s$, there exists a positive constant c_0 such that

$$\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n x_{i\Gamma}\otimes x_{i\Gamma^c}\right|_\infty \le c_0, \quad \frac{1}{\sqrt{n}}\left(\left|\sum_{i=1}^n x_{i\Gamma}\otimes Z_i^{\Gamma\Gamma^c}\right|_\infty + \left|\sum_{i=1}^n x_{i\Gamma^c}\otimes Z_i^{\Gamma\Gamma}\right|_\infty\right) \le c_0,$$
$$\frac{1}{\sqrt{n}}\left(\left|\sum_{i=1}^n Z_i^{\Gamma\Gamma}\otimes Z_i^{\Gamma^c\Gamma}\right|_\infty + \left|\sum_{i=1}^n Z_i^{\Gamma\Gamma}\otimes Z_i^{\Gamma^c\Gamma^c}\right|_\infty\right) \le c_0,$$

where $\otimes$ is the Kronecker product.

In the linear model (2.2), Condition 3 coincides with the partial orthogonality condition of Huang, Horowitz and Ma (2008) that $n^{-1/2}|\sum_{i=1}^n x_{ij}\, x_{ik}|_\infty \le c_0$ for any $j \in \Gamma, k \in \Gamma^c$.

Condition 4 (Asymptotic Property of Tuning Parameter). Let $\lambda_n \ge \sigma\underline{c}^{1-q}\sqrt{n\log p}$ be such that, as $n \to \infty$,

$$\frac{\sqrt{n^q}s^{3-q}(\log n)^{2-q}}{\lambda_n} \to 0, \quad \frac{\lambda_n s^{(4-q)/\{2(1-q)\}}}{n} \to 0, \tag{3.4}$$

$$\frac{\lambda_n \kappa_{1n} s \log n}{n} \to 0, \quad \frac{\lambda_n \kappa_{2n} s^2 \log n}{n} \to 0. \tag{3.5}$$

The inequality here is equivalent to $\lambda_n > 2\sqrt{(1+C)n \log p}$ for some positive constant C, which is used in Fan, Fan and Barut (2014). The first convergence is similar to the first one in their condition (4.4), and the first convergence in (3.5) is similar to their second convergence in (4.4). The first convergence of this condition is trivial when s is finite and the inequality in Condition 2 holds. The second convergence implies that the penalty parameter λ_n is $o(n)$ if s is finite. If, for example, $\lambda_n = n^\delta$ for a positive constant δ, then Condition 2 implies that $\delta \in (1/2, 1)$ and $\log p = o(n^{(2\delta-1)})$. Thus, Condition 2 imposes a range for the penalty parameter with respect to the sample size n and dimension p. It is easy to verify that Condition 2 also implies that $s = o(\sqrt{n})$ as needed to approximate the Hessian through the Jacobian.

The proof of the following is given in the Supplementary Material.

Theorem 1 (Moderate Deviation)**.** *Under model* (2.1)*, if Condition* 1-4 *hold, then there exists a strict local minimizer* $\hat{\beta} = \big(\hat{\beta}^T_{\Gamma^*}, \hat{\beta}^T_{\Gamma^{*c}}\big)^T$ *of* (2.3) *and a positive constant* $C_0 < \min\{1/(8\sigma^2), 1/(2c_2\sigma^2), c_1^2/(8c_2\sigma^2)\}$ *such that*

$$\mathbb{P}\big(\hat{\beta}_{\Gamma^{*c}} = 0\big) \geq 1 - \exp\big(-C_0 a_n^2\big), \tag{3.6}$$

$$\mathbb{P}\left(\|\hat{\beta}_{\Gamma^*} - \beta^*_{\Gamma^*}\|_2 \leq r_n\right) \geq 1 - \exp\big(-C_0 a_n^2\big), \tag{3.7}$$

where

$$\hat{\beta}_{\Gamma^*} \in \operatorname{argmin}_{\beta_1 \in \mathbb{R}^s} \tilde{L}_n(\beta_1) := \sum_{i=1}^n \left(y_i - \beta_1^T Z_i^{\Gamma^*\Gamma^*} \beta_1 - x_{i\Gamma^*}^T \beta_1\right)^2 + \lambda_n \|\beta_1\|_q^q,$$

$r_n = \big(a_n/\sqrt{n} + (2\underline{c}^{q-1}\lambda_n\sqrt{s})/(c_1 n)\big)$*, and* $\{a_n\}$ *is a sequence of positive numbers such that, as* $n \to \infty$*,*

$$\frac{a_n}{\sqrt{s \log n}} \to \infty, \tag{3.8}$$

$$\frac{a_n \kappa_{1n}\sqrt{s}}{\sqrt{n}} \to 0, \qquad \frac{a_n \kappa_{2n} s^{3/2}}{\sqrt{n}} \to 0, \tag{3.9}$$

$$a_n \left(\frac{n^{q/2} s^{(4-q)/2}}{\lambda_n}\right)^{1/(2-q)} \to 0. \tag{3.10}$$

Note that since $\max(\kappa_{1n}^2, \kappa_{2n}^2) \geq 1$, condition (3.9) implies $a_n\sqrt{s/n} \to 0$. Again, if s is finite and $\lambda_n = n^\delta$ for some $\delta \in (1/2, 1)$, then conditions (3.8)-(3.10) simplify to

$$\frac{a_n}{\sqrt{\log n}} \to \infty, \quad \frac{a_n \kappa_{1n}}{\sqrt{n}} \to 0, \qquad \frac{a_n \kappa_{2n}}{\sqrt{n}} \to 0, \quad \frac{a_n}{n^{(2\delta-q)/\{2(2-q)\}}} \to 0.$$

It follows that $\{a_n\}$ tends to infinity faster than $\log n$ but slower than $n^{(2\delta-q)/(2(2-q))} = o(\sqrt{n})$. This differs from the case of the linear model with fixed dimension $p \ll n$, where only $a_n\kappa_{1n}/\sqrt{n} \to 0$ is required to establish the MD of M-estimators (Fan (2012), Fan, Yan and Xiu (2014)). We assume (3.8)-(3.10) to cover the case of $p \gg n$.

By inequality (3.6) the q-RLS estimator correctly selects nonzero variables with probability approaching one exponentially. It follows from (3.7) that the estimators of nonzero variables are consistent with an exponential rate of convergence. Theorem 1 also implies that the tail probability decreases exponentially with rate a_n^2, as the tail probability of the Gaussian.

Theorem 1 gives general results on the MD. By taking $a_n = \sqrt{s}\log n$, we obtain the familiar forms of convergence rate.

Theorem 2 (Weak Oracle Property)**.** *Under model* (2.1)*, if Conditions* 1-4 *hold, then there exists a strict local minimizer* $\hat{\beta} = \big(\hat{\beta}_{\Gamma^*}^T, \hat{\beta}_{\Gamma^{*c}}^T\big)^T$ *of* (2.3) *such that, for sufficiently large* n*,*

$$\mathbb{P}\big(\hat{\beta}_{\Gamma^{*c}} = 0\big) \geq 1 - n^{-C_0 s\log n}, \tag{3.11}$$

$$\mathbb{P}\left(\|\hat{\beta}_{\Gamma^*} - \beta^*_{\Gamma^*}\|_2 \leq \frac{\sqrt{s}\log n}{\sqrt{n}} + \frac{2\underline{c}^{q-1}\lambda_n\sqrt{s}}{c_1 n}\right) \geq 1 - n^{-C_0 s\log n}. \tag{3.12}$$

In particular, when $Z_i \equiv 0$, Conditions 1-4 reduce to similar conditions of Huang, Horowitz and Ma (2008) and Fan, Fan and Barut (2014) for the linear model (2.2).

Corollary 1. *Under* (2.2)*, the results of Theorem* 2 *hold, provided*
(1) *for each* $\Gamma \subseteq \{1,\ldots,p\}$ *with* $|\Gamma| = s$*, the eigenvalues of* $1/nX_\Gamma^T X_\Gamma$ *are bounded from below and above by some positive constants* c_1 *and* c_2 *respectively;*
(2) $\kappa_{1n}\sqrt{s}/\sqrt{n} \to 0$*, as* $n \to \infty$*;*
(3) *for each* $\Gamma \subseteq \{1,\ldots,p\}$ *with* $|\Gamma| = s$*, there exists a positive constant* c_0 *such that* $n^{-1/2}|\sum_{i=1}^n x_{ij}x_{ik}|_\infty \leq c_0,\ \forall j \in \Gamma, k \in \Gamma^c$;
(4) $\lambda_n \geq \sigma\underline{c}^{1-q}\sqrt{n\log p}$ *and* $\lambda_n^{-1}\sqrt{n^q}s^{3-q}(\log n)^{2-q} \to 0, n^{-1}\lambda_n s^{(4-q)/(2(1-q))} \to 0, n^{-1}\lambda_n\kappa_{1n}s\log n \to 0$*, as* $n \to \infty$.

Remark 1. To deal with the case $p > n$, Huang, Horowitz and Ma (2008) showed that the marginal bridge estimators satisfy $\mathbb{P}\big(\hat{\beta}_{\Gamma^{*c}} = 0\big) \to 1$ and $\mathbb{P}\big(e_{p,j}^T\hat{\beta} \neq 0,\ j \in \Gamma^*\big) \to 1$. Here we provide the rate of this convergence. The result (3.12) is slightly different from Theorem 2 in Fan, Fan and Barut (2014) that has

$$\mathbb{P}\left\{\|\hat{\beta}_{\Gamma^*} - \beta^*_{\Gamma^*}\|_2 \leq \gamma_0\left(\frac{\sqrt{s\log n}}{\sqrt{n}} + \frac{\lambda_n\|d_0\|_2}{n}\right)\right\} \geq 1 - O(n^{-cs}),$$

where γ_0 and c are two positive constants and d_0 is a s-dimensional vector of nonnegative weight. To find the constant c, we use the number $\sqrt{\log n}$ to dominate the constant γ_0, which results in the lower consistent rate. To compensate this loss, the right hand side of (3.12) tends to one at a faster rate.

4. Purely Quadratic Model

In this section we consider the purely quadratic measurements model

$$y_i = \beta^T Z_i \beta + \varepsilon_i, \quad i = 1, \dots, n. \tag{4.1}$$

As demonstrated in Example 2, this covers the phase retrieval model where $Z_i = z_i z_i^T$. As this model differs from the general model (2.1), some theoretical conditions and results in the previous sections need to be modified.

4.1. Identifiability of β^*

The absence of the linear term in model (4.1) makes it unidentifiable because obviously β^* and $-\beta^*$ are indistinguishable from the observed data. In the phase retrieval literature, e.g., Balan, Casazza and Edidin (2006) and Ohlsson and Eldar (2014), this problem is treated by identifying $\pm\beta$ for any $\beta \in \mathbb{R}^p$. Without loss of generality, we assume that the first nonzero element of β^* is positive.

For the phase retrieval problem, Balan, Casazza and Edidin (2006) and Bandeira et al. (2014) introduce the complement property which is necessary and sufficient for identifiability. For the sparse regression, Ohlsson and Eldar (2014) propose the more general concept of s-complement property. In the phase retrieval model where $Z_i = z_i z_i^T$, the s-complement property of $\{z_i\}$ means that either $\{z_i^\Gamma\}_{i\in\mathbb{N}}$ or $\{z_i^\Gamma\}_{i\in\mathbb{N}^c}$ span $\mathbb{R}^s$ for every subset $\mathbb{N} \subseteq \{1, \dots, n\}$ and $\Gamma \subseteq \{1, \dots, p\}$ with $|\Gamma| = s$. Here, the identifiability of β^* in (2.1) is guaranteed by the uniform s-regularity of the affine transform $\mathcal{A}(\beta)$. In model (4.1), the residual function $R(\beta) = \big(R_1(\beta), \dots, R_n(\beta)\big)^T$ with $R_i(\beta) = y_i - \beta^T Z_i \beta$ has Jacobian matrices $(-2Z_1\beta, \dots, -2Z_n\beta\big)^T$.

Definition 2. *The linear transform $\mathcal{B}(\beta) = (Z_1\beta, \dots, Z_n\beta\big)^T$ is uniformly s-regular if $\mathcal{B}(\beta)_\Gamma$ has full column rank for any $\Gamma \subseteq \{1, \dots, p\}$ with $|\Gamma| = s$, and $\beta \in \mathbb{R}^p/\{0\}$ with supp$(\beta) \subseteq \Gamma$.*

If $Z_i = z_i z_i^T$, then the uniform s-regularity of $\mathcal{B}(\beta)$ is equivalent to the s-complement property of $\{z_i\}$. Further, it is straightforward to verify that $\mathcal{B}(\beta)$ is uniformly s-regular if and only if the submatrix $\mathcal{B}_\Gamma(\beta_1) = (Z_1^{\Gamma\Gamma}\beta_1, \dots, Z_n^{\Gamma\Gamma}\beta_1\big)^T$ has full column rank for any $\Gamma \subseteq \{1, \dots, p\}$ with $|\Gamma| = s$ and $\beta_1 \in \mathbb{R}^s/\{0\}$.

The proof of the following is analogous to that of Theorem 4 in Ohlsson and Eldar (2014) and is therefore omitted.

Proposition 2. *Under* (4.1), β^* *is the unique solution satisfying* $\|\beta^*\|_0 \le s$ *if* $\mathcal{B}(\beta)$ *is uniformly* $2s$*-regular.*

4.2. Weak oracle property

To drive the MD and consistency results under model (4.1), we modify Conditions 1-4 in Section 3.2 as follows.

Condition 1′ (Uniformly Stable Restricted Jacobian).

(a) For any $\Gamma \subseteq \{1,\dots,p\}$ with $|\Gamma| = s$ and $\beta_1 \in S_1 := \big\{u \in \mathbb{R}^s : |\{j : |e_{s,j}^T u| \ge \underline{c}/2\}| \ge s - [s/2]\big\}$, there exists a positive constant c_1 that bounds all eigenvalues of $n^{-1}\big(\mathcal{B}_\Gamma(\beta_1)^T\mathcal{B}_\Gamma(\beta_1)\big)$ from below.

(b) For any $\Gamma \subseteq \{1,\dots,p\}$ with $|\Gamma| = s$ and $\beta_1 \in S$, there exists a positive constant c_2 that bounds all eigenvalues of $n^{-1}\big(\mathcal{B}_\Gamma(\beta_1)^T\ \mathcal{B}_\Gamma(\beta_1)\big)$ from above.

Condition 2′ (Asymptotic Property of Design Matrix). As $n \to \infty$, $n^{-1/2}\kappa_{2n}s^{3/2} \to 0$.

Condition 3′ (Partial Orthogonality). There exists a positive constant c_0 such that

$$n^{-1/2}\left(\left|\sum_{i=1}^n Z_i^{\Gamma\Gamma}\otimes Z_i^{\Gamma^c\Gamma}\right|_\infty + \left|\sum_{i=1}^n Z_i^{\Gamma\Gamma}\otimes Z_i^{\Gamma^c\Gamma^c}\right|_\infty\right) \le c_0.$$

Condition 4′ (Asymptotic Property of Tuning Parameter). Let $\lambda_n \ge \sigma\underline{c}^{1-q}\sqrt{n\log p}$ be such that, as $n \to \infty$,

$$\frac{\sqrt{n^q}s^{3-q}(\log n)^{2-q}}{\lambda_n} \to 0, \quad \frac{\lambda_n s^{(4-q)/\{2(1-q)\}}}{n} \to 0 \ \text{ and } \ \frac{\lambda_n\kappa_{2n}s^2\log n}{n} \to 0.$$

For the phase retrieval model, since $\mathcal{B}_\Gamma(\beta_1)^T\mathcal{B}_\Gamma(\beta_1) = \sum_{i=1}^n \big(z_{i\Gamma}^T\beta_1\big)^2 z_{i\Gamma}z_{i\Gamma}^T$, Condition 1′ implies that at $\beta_{\Gamma^*}^*$,

$$c_1\|u\|^2 \le \frac{1}{n}\sum_{i=1}^n \big(z_{i\Gamma}^T\beta_{\Gamma^*}^*\big)^2(z_{i\Gamma}^T u)^2 \le c_2\|u\|^2, \qquad \forall u \in \mathbb{R}^s/\{0\}.$$

This is similar to Corollary 7.6 of Candes, Li and Soltanolkotabi (2015).

Theorem 3 (Moderate Deviation)**.** *Under model* (4.1)*, if Conditions* 1′-4′ *hold, then there exists a strict local minimizer* $\hat\beta = \big(\hat\beta_{\Gamma^*}^T, \hat\beta_{\Gamma^{*c}}^T\big)^T$ *of* (2.3) *such that* (3.6) *and* (3.7) *hold with* $\{a_n\}$ *satisfying* (3.8), (3.10) *and the second condition in* (3.9)*, where*

$$\hat\beta_{\Gamma^*} \in \operatorname{argmin}_{\beta_1\in\mathbb{R}^s}\tilde L_n(\beta_1) := \sum_{i=1}^n\Big(y_i - \beta_1^T Z_i^{\Gamma^*\Gamma^*}\beta_1\Big)^2 + \lambda_n\|\beta_1\|_q^q.$$

Theorem 4. *(Weak Oracle Property). Under model* (4.1) *and Conditions* 1′-4′*, there exists a strict local minimizer* $\hat{\beta} = \big(\hat{\beta}^T_{\Gamma^*}, \hat{\beta}^T_{\Gamma^{*c}}\big)^T$ *of* (2.3) *such that* (3.11) *and* (3.12) *hold.*

For the phase retrieval problem, Candes, Strohmer and Voroninski (2013) used convex relaxation to construct a consistent estimator of the matrix $\beta^*(\beta^*)^T$ but not for β^*. The consistency of β^* was studied by Eldar and Mendelson (2014) and Lecué and Mendelson (2015). We obtain the following weak oracle property of β^* as a consequence of Theorem 4.

Corollary 2. *Under model* (4.1) *with* $Z_i = z_i z_i^T, i = 1, 2, \ldots, n$*, the result of Theorem* 4 *holds if Conditions* 1′-4′ *hold.*

5. Optimization Algorithm

The numerical computation of the q-RLS estimator as the solution of (2.3) is an important and challenging issue, since the $\ell_q(0 < q < 1)$-regularization is a nonconvex, nonsmooth, and non-Lipschitz optimization problem. Recently, this type of problems has attracted much attention in the field of optimization, including developing optimality conditions and computational algorithms, see, e.g., Xu et al. (2012), Chen, Niu and Yuan (2013), Lu (2014) and references therein. In this section, we propose an algorithm for the minimization problem (2.3). Since n and p are given, to simplify notation we omit the subscript n of $\ell_n(\beta)$ and λ_n so that (2.3) is written as

$$\min_{\beta\in\mathbb{R}^p} L(\beta) := \ell(\beta) + \lambda\|\beta\|_q^q, \tag{5.1}$$

where $\lambda > 0$. We start by considering the simple minimization problem

$$\min_{u\in\mathbb{R}} \varphi_t(u) := \frac{1}{2}(u-t)^2 + \lambda|u|^q, \tag{5.2}$$

where $t \in \mathbb{R}, \lambda > 0$, and $q \in (0, 1)$. For this problem Chen, Xiu and Peng (2014) show that there exists an implicit function $h_{\lambda,q}(\cdot)$ such that the minimizer $\hat{u}$ of (5.2) satisfies $\hat{u} = h_{\lambda,q}(t)$. In particular, for $q = 1/2$, Xu et al. (2012) give an explicit expression $h_{\lambda,1/2}(t) = 2/3t\big[1 + \cos\big\{2\pi/3 - 2/3\phi_\lambda(t)\big\}\big]$ with $\phi_\lambda(t) = \arccos\big(\lambda/4((|t|)/3)^{-3/2}\big)$.

Theorem 5. *There exists a function* $h_{\lambda,q}(\cdot)$ *and a constant* $r > 0$*, such that any minimizer* $\hat{\beta}$ *of problem* (5.1) *satisfies*

$$\hat{\beta} = \mathcal{H}_{\lambda\tau,q}\big\{\hat{\beta} - \tau\nabla\ell(\hat{\beta})\big\} \tag{5.3}$$

for any $\tau \in \big(0, \min\{G_r^{-1}, 1\}\big)$*, where* $G_r = \sup_{\beta\in B_r}\|\nabla^2\ell(\beta)\|_2$*,* $B_r = \{\beta \in \mathbb{R}^p :$

$\|\beta\|_2 \le r\}$, *and* $\mathcal{H}_{\lambda,q}(u) = \big(h_{\lambda,q}(u_1), \ldots, h_{\lambda,q}(u_p)\big)^T$ *for* $u = (u_1, \ldots, u_p)^T \in \mathbb{R}^p$.

Remark 2. The result of Theorem 5 remains true for any function ℓ that is bounded from below, twice continuously differentiable, and for which $\lim_{\|x\|\to\infty} \ell(\beta) = \infty$. An appropriate algorithm here can be derived similarly to that below.

Remark 3. In general, the ℓ_q minimization problem $\min_{\beta\in\mathbb{R}^p} f(\beta) + \lambda\|\beta\|_q^q$ with $\lambda > 0,\ q \in (0,1)$ has been well studied in the optimization literature and efficient algorithms have been proposed for $f(\beta) = \|X\beta - y\|^2$. For example, Chen, Xu and Ye (2010) derived lower bounds for nonzero entries of the local minimizer and presented a hybrid orthogonal matching pursuit-smoothing gradient method, while Xu et al. (2012) provided a globally necessary optimality condition for the case $q = 1/2$ and proposed an efficient iterative algorithm. More recently, the general ℓ_q problem has been studied by Chen, Niu and Yuan (2013), who proposed a smoothing trust region Newton method for solving a class of non-Lipschitz optimization problems. Lu (2014) studied iterative reweighted methods for a smooth and bounded (from below) function f with an L_f-Lipschitz continuous gradient satisfying $\|\nabla f(\beta) - \nabla f(\beta')\| \le L_f\|\beta - \beta'\|$. Bian, Chen and Ye (2015) proposed interior point algorithms for solving a class of non-Lipschitz nonconvex optimization problems with nonnegative bounded constraints. In these works the solution sequence of the algorithm converges to a stationary point derived from the Karush-Kuhn-Tucker conditions.

Based on (5.3), we propose a fixed point iterative algorithm (FPIA).

Algorithm 1

Step 0. Given $\lambda > 0, \epsilon \ge 0, \gamma, \alpha \in (0,1), \delta > 0$, choose an arbitrary β^0 and set $k = 0$.

Step 1. (a) Compute $\nabla\ell(\beta^k)$ from $\nabla\ell(\beta) = 2\sum_{i=1}^m (\beta^T Z_i\beta + x_i^T\beta - y_i)(2Z_i\beta + x_i)$;
(b) Compute $\beta^{k+1} = \mathcal{H}_{\lambda\tau_k,q}(\beta^k - \tau_k\nabla\ell(\beta^k))$ with $\tau_k = \gamma\alpha^{j_k}$ and j_k the smallest nonnegative integer such that

$$L(\beta^k) - L(\beta^{k+1}) \ge \frac{\delta}{2}\|\beta^k - \beta^{k+1}\|_2^2. \tag{5.4}$$

Step 2. Stop if $\|\beta^{k+1} - \beta^k\|_2 \le \epsilon\max\{1, \|\beta^k\|_2\}$. Otherwise, replace k by $k+1$ and go to Step 1.

An important step here is to evaluate the operator $\mathcal{H}_{\lambda,q}(\cdot)$. It has an explicit expression when $q = 1/2$. For more general $q \in (0,1)$, by Lemma 6 in the Supplementary Material, there exists a constant $t^* > 0$ such that $h_{\lambda,q}(t) > 0$, $h_{\lambda,q}(t) - t + \lambda q h_{\lambda,q}(t)^{q-1} = 0$, and $1 + \lambda q(q-1)h_{\lambda,q}(t)^{q-2} > 0$, for $t > t^*$; and

$h_{\lambda,q}(t) < 0$, $h_{\lambda,q}(t) - t - \lambda q|h_{\lambda,q}(t)|^{q-1} = 0$ and $1 + \lambda q(q-1)\ |h_{\lambda,q}(t)|^{q-2} > 0$, for $t < -t^*$. Hence one can use the function `fsolve` in Matlab to get the desired solution at each iteration.

Another step is the computation of step length τ_k, which represents a tradeoff between the speed of reduction of the objective function L and search time for the optimal length. According to Theorem 5, the ideal choice of τ_k depends on the maximum eigenvalue of the Hessian $\nabla^2\ell(\beta^k)$ at kth iteration, which is expensive to calculate. A more practical strategy is to perform an inexact line search to identify a step length that achieves adequate reduction in L. One such technique is the so-called Armijo-type line search that is adopted in our proposed algorithm. In our context this method requires finding the smallest nonnegative integer j_k such that (5.4) holds. That this can be done successfully is assured by Lemmas 8 and 9 in the Supplementary Material. We also verify the convergence property of the FPIA by Theorem 1 in the Supplementary Material.

Remark 4. Xu et al. (2012) studied a q-regularized least square method with $q = 1/2$ in a linear model and proposed several strategies for choosing the optimal regularization parameter λ besides cross validation. Analogous to their method we can derive the range of the optimal regularization parameter in our problem as $\hat{\lambda} \in [\sqrt{96}/(9\tau)|[B_\tau(\hat{\beta})]_{s+1}|^{3/2}, \sqrt{96}/(9\tau)|[B_\tau(\hat{\beta})]_s|^{3/2})$ where $B_\tau(\beta) = \beta - \tau\nabla\ell(\beta)$ and $|[B_\tau(\hat{\beta})]_k|$ is the kth largest component of $B_\tau(\hat{\beta})$ in magnitude for each $k = 1, \ldots, p$. Xu et al. (2012) suggest that $\hat{\lambda} = \sqrt{96}/(9\tau)|[B_\tau(\hat{\beta})]_{s+1}|^{3/2}$ is a reliable choice with an approximation such as $\hat{\beta} \approx \beta^k$. They recommend this strategy for s-sparsity problems and cross validation for more general problems.

Our algorithm can also be used to compute the q-RLS estimator for $q \geq 1$. Indeed, similar to Lemma 6 in the Supplementary Material, we can show that there exists a unique function $h_{\lambda,q}(t)$ such that the global minimizer of problem (5.2) is $\hat{u} = h_{\lambda,q}(t)$. In particular, we can obtain the explicit expressions of this function for $q = 1, 2/3, 2$ as $h_{\lambda,1}(t) = \max(0, t-\lambda) - \max(0, -t-\lambda,)$, $h_{\lambda,3/2}(t) = (\sqrt{9/16\lambda^2 + |t|} - 3/4\lambda)^2\text{sign}(t)$ and $h_{\lambda,2}(t) = t/(1+2\lambda)$.

6. Numerical Examples

In this section we calculate two examples to illustrate the proposed approach and demonstrate the finite sample performance of the q-RLS estimator. The first example is the second-order least squares method described in Example 4, and the second is the quadratic equations problem considered by Beck and Eldar (2013). In a phase diagram study Xu et al. (2012a) point out that the

ℓ_q-regularization method yields sparser solutions with smaller value of q in the range $[1/2, 1)$, while there is no significant difference for $q \in (0, 1/2]$. In view of these findings, we use $q = 1/2$ in both examples. In addition, following the literature we use 5-fold cross validation to choose the parameter λ. In each simulation 100 Monte Carlo samples were generated and in each case the true value β^* was generated randomly with s nonzero components standard normal. The numerical optimization is done using FPIA with iteration stopping criterion

$$\frac{\|\beta^{k+1} - \beta^k\|}{\max\{1, \|\beta^{k+1}\|\}} \le 10^{-6},$$

or the maximum iterative time of 5,000s is reached.

To evaluate the selection and estimation accuracy of our method, we calculated the mean squared error (MSE) which is the average of $\|\hat{\beta} - \beta^*\|_2^2$; the false positive (FP) which is the number of zero coefficients incorrectly identified as nonzero; the false negative (FN) which is the number of nonzero coefficients incorrectly identified as zero. We also report the rate of successful recovery (SR) using the criterion $\hat{\Gamma} = \Gamma^*$ and $\|\hat{\beta} - \beta^*\|_2^2 \le 2.5 \times 10^{-5}$, where $\hat{\Gamma} = \{j : \hat{\beta}_j \ne 0\}$ and $\Gamma^* = \{j : \beta_j^* \ne 0\}$.

Example 1. Second-order Least Square Method

We applied the second-order least squares method described in Example 4 to the variable selection problem in (2.2). It is known that in low-dimensional set-ups the SLS estimator is asymptotically more efficient than the ordinary least squares estimator when the error distribution is asymmetric. Therefore it is interesting to see if this robustness property carries over to high-dimensional regularized estimation. In particular, we considered the q-regularized second-order least squares (q-RSLS) problem

$$\min_{\theta} \sum_{i=1}^{n} \rho_i(\theta)^T W_i \rho_i(\theta) + \lambda \|\beta\|_q^q,$$

where $\theta = (\beta^T, \sigma^2)^T$, $\rho_i(\theta) = \{y_i - x_i^T\beta, y_i^2 - (x_i^T\beta)^2 - \sigma^2\}^T$ and W_i is a 2×2 nonnegative definite weight matrix. Here the objective function becomes that of the q-regularized least squares (q-RLS) method if the weight is taken to be $W_i = \text{diag}(1, 0)$. To simplify computation, we used the weight $W_i = \begin{pmatrix} 0.75 & 0.1 \\ 0.1 & 0.25 \end{pmatrix}$ that is not necessarily optimal according to Wang and Leblanc (2008).

We considered five error distributions $\text{logN}(0, 0.1^2) - e^{-0.005}$, $(\mathcal{X}^2(5) - 5)/100$, $0.01 * t$, U$[-0.1, 0.1]$ and N$(0, 0.1^2)$. In each case, we took dimension $p = 400$

Table 1. Selection and estimation results of Example 1.

error	method	FP		FN		MSE
		mean	se	mean	se	
e_1	q-RSLS	0.12	0.04	0.00	0.00	3.41e-05
	q-RLS	0.27	0.05	0.00	0.00	1.38e-04
e_2	q-RSLS	0.12	0.04	0.00	0.00	2.91e-05
	q-RLS	0.21	0.05	0.00	0.00	9.34e-05
e_3	q-RSLS	0.09	0.03	0.00	0.00	1.32e-05
	q-RLS	0.22	0.05	0.00	0.00	9.51e-05
e_4	q-RSLS	0.09	0.03	0.00	0.00	3.34e-05
	q-RLS	0.29	0.05	0.00	0.00	1.64e-04
e_5	q-RSLS	0.11	0.03	0.00	0.00	2.14e-05
	q-RLS	0.24	0.05	0.00	0.00	1.30e-04
Noiseless	q-RSLS	0.10	0.03	0.00	0.00	3.80e-05
	q-RLS	0.19	0.04	0.00	0.00	1.02e-04

Table 2. Rates of successful recovery of Example 1.

method \ error	e_1	e_2	e_3	e_4	e_5	Noiseless
q-RSLS	0.62	0.78	0.88	0.52	0.86	0.56
q-RLS	0.08	0.15	0.13	0.06	0.12	0.10

with sparsity $s = 8$ and sample size $n = 200$.

The results in Table 1 show that q-RSLS and q-RLS perform well in identifying zero coefficients; this is expected for ℓ_q-regularized methods with $q = 1/2$. Although both methods have fairly low FP values, the values of q-RLS is about 3 times higher than that of the q-RSLS. Moreover, The MSE of the q-RSLS estimator is about three times smaller than that of the q-RLS estimator. The results in Table 2 show clearly that q-RSLS has much higher rate of SR than q-RLS does, and this is true not only for the skewed error distributions, such as log-normal and Chi-square, but also for normal or uniform distributions.

Example 2. Quadratic Measurements

We considered (4.1) with $\varepsilon_i \sim N(0, \sigma^2)$. A noise-free version of this model was considered by Beck and Eldar (2013). For the sake of comparison we set $\sigma = 0.01$ and generated matrices as $Z_i = z_i z_i^T$, $i = 1, 2, \ldots, m$ with vectors $z_i \in \mathbb{R}^p$ from the standard normal. We considered $n = 80$, $p = 120$ with various sparsity $s = 3, 4, \ldots, 10$. For comparison, we calculated the q-RLS estimator for $q = 1/2, 1, 3/2, 2$.

Table 3. Selection and estimation results of Example 2.

$\|\beta^*\|_0$	method	FP		FN		MSE	SR
		mean	se	mean	se		
3	$q = 1/2$	3.95	0.64	0.36	0.09	1.56e-03	0.57
	$q = 1$	64.36	5.84	1.07	0.14	1.47e-01	0.00
	$q = 3/2$	117.00	0.00	0.00	0.00	2.04e-01	0.00
4	$q = 1/2$	3.73	0.65	0.09	0.04	6.98e-04	0.62
	$q = 1$	62.64	5.79	1.63	0.20	2.97e-01	0.00
	$q = 3/2$	116.00	0.00	0.00	0.00	2.15e-01	0.00
5	$q = 1/2$	4.66	0.71	0.05	0.02	4.97e-05	0.61
	$q = 1$	76.75	5.38	1.55	0.23	3.08e-01	0.00
	$q = 3/2$	115.00	0.00	0.00	0.00	3.18e-01	0.00
6	$q = 1/2$	5.99	0.88	0.04	0.02	4.75e-05	0.58
	$q = 1$	81.88	5.07	1.50	0.26	2.60e-01	0.00
	$q = 3/2$	114.00	0.00	0.00	0.00	4.36e-01	0.00
7	$q = 1/2$	4.70	0.84	0.07	0.03	3.37e-05	0.63
	$q = 1$	83.32	4.91	1.01	0.30	3.27e-01	0.00
	$q = 3/2$	113.00	0.00	0.00	0.00	5.76e-01	0.00
8	$q = 1/2$	3.76	0.77	0.32	0.14	5.22e-02	0.67
	$q = 1$	87.54	4.37	1.28	0.30	2.78e-01	0.00
	$q = 3/2$	111.99	0.01	0.00	0.00	8.02e-01	0.00
9	$q = 1/2$	4.01	0.97	0.34	0.16	4.92e-02	0.73
	$q = 1$	86.05	4.38	1.53	0.34	3.30e-01	0.00
	$q = 3/2$	111.00	0.00	0.00	0.00	6.35e-01	0.00
10	$q = 1/2$	5.46	0.46	0.11	0.03	2.68e-02	0.58
	$q = 1$	84.69	4.22	1.50	0.36	3.56e-01	0.00
	$q = 3/2$	110.00	0.00	0.00	0.00	6.57e-01	0.00

The results are given in Table 3, with the results for $q = 2$ omitted since they are very similar to those for $q = 3/2$. They show clearly that the FP values with $q = 1/2$ is much lower than the other cases. In particular, the FP values with $q = 3/2, 2$ are the same as the number of true nonzero coefficients, which means that no variable selection was performed.

The MSE and SR are both very small; this demonstrates that the q-RLS with $q = 1/2$ is efficient and stable in variable selection and estimation. Compared to the results in Beck and Eldar (2013), our SR rates are lower when $s = 3, 4$ but significantly higher when $s = 5, 6, 7, 8, 9, 10$.

To see the effectiveness of our numerical algorithm FPIA, we also ran the simulations with $n = 3p/4$, $s = 0.05p$, and $p = 100,\ 200,\ 300, 400,\ 500$. The results in Table 4 show that, as the dimension increases, the FP and FN, as well as MSE, remain fairly low and stable. In all cases, the rates of successful recovery

Table 4. The successful recoveries of Example 2.

p	$\|\beta^*\|_0$	FP		FN		MSE	SR
		mean	se	mean	se		
100	5	2.99	0.60	0.12	0.07	1.90e-03	0.73
200	10	3.40	0.80	0.05	0.02	2.49e-05	0.86
300	15	9.50	1.20	0.09	0.03	5.17e-04	0.53
400	20	11.34	1.43	0.11	0.05	5.26e-04	0.53
500	25	13.07	2.56	0.07	0.03	5.45e-04	0.51

are over 50% and reach 86% when $p = 200$.

7. Discussion

Compared to the linear model, the quadratic measurements model is more complex and therefore it is harder to obtain the MD rate. Under some further assumptions, it is possible to establish more accurate results. Another open question is the asymptotic normality of the q-RLS estimator for model (2.1), which deserves further research.

We have studied the generalized bridge estimator because of the simplicity and tractability of numerical optimization. We focused on the ℓ_q regularization with $q < 1$, mainly because in phase retrieval and compressive sensing the primary goal is to find the smallest set of predictors and the ℓ_q method with $q < 1$ helps to achieve this goal. Our identification results and numerical optimization algorithm apply when $q \geq 1$. Of course in such cases the consistency results do not hold generally as in linear models. It is also interesting to investigate the SCAD and other regularization methods in quadratic measurements models. Our model (2.1) can be viewed as a special case of the partially linear index model $y = \sum_{j=1}^{d} f_j(\beta^T w_j) + x^T\beta + \varepsilon$. While it is interesting to study the regularization estimation problem in this model, the theory and method are much more complicated.

Supplementary Materials

The supplementary file covers technical lemmas and proofs.

Acknowledgment

The authors thank the Editor, an associate editor and two anonymous reviewers for their comments and suggestions that helped to improve the previous version of this paper. Fan, Kong and Xiu's research was supported by

the National Natural Science Foundation of China (NSFC) (No. 11431002 and 11671029), while Wang's research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2016-06002). The work was also supported by the 111 Project of China (No. B16002).

References

Bahmani, S., Raj, B. and Boufounos, P. T. (2013). Greedy sparsity-constrained optimization. *J. Mach. Learn. Res.* **14**, 807-841.

Balan, R., Casazza, P. and Edidin, D. (2006). On signal reconstruction without phase. *Appl. Comput. Harmon. Anal.* **20**, 345-356.

Bandeira, A. S., Cahill, J., Mixon, D. G. and Nelson, A. A. (2014). Saving phase: Injectivity and stability for phase retrieval. *Appl. Comput. Harmon. Anal.* **37**, 106-125.

Beck, A. and Eldar, Y. C. (2013). Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM J. Optim.* **23**, 1480-1509.

Bian, W., Chen, X. and Ye Y. (2015). Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Math. Program.* **149**, 301-327.

Biswas, P. and Ye, Y. (2004). Semidefinite programming for ad hoc wireless sensor network localization. In *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*, 46-54, Berkeley, CA.

Blumensath, T. (2013). Compressed sensing with nonlinear observations and related nonlinear optimization problems. *IEEE Trans. Inform. Theory* **59**, 3466-3474.

Cai, T., Li, X. and Ma, Z. (2016). Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *Ann. Statist.* **44**, 2221-2251.

Candes, E., Strohmer, T. and Voroninski, V. (2013). Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.* **66**, 1241-1274.

Candes, E., Li, X. and Soltanolkotabi, M. (2015). Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Trans. Inform. Theory* **61**, 1985-2007.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313-2351.

Chartrand, R. (2007). Exact reconstruction of sparse signals via nonconvex minimizaion. *IEEE Signal Process. Lett.* **14**, 707-710.

Chen, X., Niu, L. and Yuan, Y. (2013). Optimality conditions and smoothing trust region Newton method for non-Lipschitz optimization. *SIAM J. Optim.* **23**, 1528-1552.

Chen, X., Xu, F. and Ye, Y. (2010). Lower bound theory of nonzero entries in solutions of $\ell_2 - \ell_p$ minimization. *SIAM J. Sci. Comput.* **32**, 2832-2852.

Chen, Y., Xiu, N. and Peng, D. (2014). Global solutions of non-Lipschitz $S_2 - S_p$ minimization over positive semidefinite cone. *Optim. Lett.* **8**, 2053-2064.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1-32.

Donoho, D. L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization. *Proceedings of the National Academy of Sciences* **100**, 2197-2202.

Eldar, Y. C. and Mendelson, S. (2014). Phase retrieval: Stability and recovery guarantees. *Appl. Comput. Harmon. Anal.* **36**, 473-494.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20**, 101-148.

Fan, J., Fan, Y. and Barut, E. (2014). Adaptive robust variable selection. *Ann. Statist.* **42**, 324-351.

Fan, J. (2012). Moderate deviations for m-estimators in linear models with ϕ-mixing errors. *Acta Math. Sin. (Engl. Ser.)* **28**, 1275-1294.

Fan, J., Yan, A. and Xiu, N. (2014). Asymptotic properties for m-estimators in linear models with dependent random errors. *J. Stat. Plan. Infer.* **148**, 49-66.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemonmetrics regression tools (with discussion). *Technometrics* **35**, 109-148.

Huang, J., Horowitz, J. L. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587-613.

Kallenberg, W. C. M. (1983). On moderate deviation theory in estimation. *Ann. Statist.* **11**, 498-504.

Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356-1378.

Krishnan, D. and Fergus, R. (2009). Fast image deconvolution using hyper-Laplacian priors. In *Adavances in Neural Information Processing Systems 22*, 1033-1041.

Lecué, G. and Mendelson, S. (2015). Minimax rate of convergence and the performance of empirical risk minimization in phase recovery. *Electron. J. Probab.* **20**(57), 1-29.

Lu, Z. (2014). Iterative reweighted minimization methods for l_p regularized unconstrained nonlinear programming. *Math. Program.* **147**, 277-307.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498-3528.

Meng, C., Ding, Z. and Dasgupta, S. (2008). A semidefinite programming approach to source localization in wireless sensor networks. *IEEE Signal Processing Letters* **15**, 253-256.

Netrapalli, P., Jain, P. and Sanghavi, S. (2015). Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing* **63**, 4814-4826.

Negahban, S. N., Ravikumar, M., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27**, 538-557.

Ohlsson, H. and Eldar, Y. C. (2014). On conditions for uniqueness in sparse phase retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1841-1845.

Ohlsson, H., Yang, A. Y., Dong, R., Verhaegen, M. and Sastry, S. S. (2014) Quadratic basis pursuit. *Regularization, Optimization, Kernels, and Support Vector Machines*, 195, CRC Press, Boca Raton.

Saab, R., Chartrand, R. and Yilmaz, O. (2008). Stable sparse approximations via nonconvex optimization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3885-3888.

Shechtman, Y., Eldar, Y. C., Szameit, A. and Segev, M. (2011). Sparsity-based sub-wavelength imaging with partially spatially incoherent light via quadratic compressed sensing. *Optics Express* **19**, 14807-14822.

Shechtman, Y., Szameit, A., Bullkich, E. et al. (2012). Sparsity-based single-shot sub-wavelength coherent diffractive imaging. *Nature Materials* **11**, 455-459.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267-288.

Wang, L. (2003). Estimation of nonlinear models with Berkson measurement error models. *Statist. Sinica* **13**, 1201-1210.

Wang, L. (2004). Estimation of nonlinear models with Berkson measurement errors. *Ann. Statist.* **32**, 2343-2775.

Wang, L. and Leblanc A. (2008). Second-order nonlinear least squares estimation. *Ann. Inst. Stat. Math.* **60**, 883-900.

Wang, Z., Zheng, S., Boyd, S. and Ye, Y. (2008). Further relaxations of the SDP approach to sensor network localization. *SIAM J. Optim.* **19**, 655-671.

Xu, Z., Guo, H., Wang, Y. and Zhang H. (2012a). Representative of $L_{1/2}$ regularization among $L_q(0 < q \leq 1)$ regularizations: an experimental study based on phase diagram. *Acta Autom. Sinica* **38**, 1225-1228.

Xu, Z., Chang X., Xu, F. and Zhang, H. (2012b). $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. Neural networks and learning systems. *IEEE Transactions on Neural Networks and Learning Systems* **23**, 1013-1027.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 49-67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894-942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301-320.

Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, P. R. China.

E-mail: fanjunmath@hotmail.com

Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, P. R. China.

E-mail: konglchen@126.com

Department of Statistics, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada.

E-mail: Liqun.wang@umanitoba.ca

Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, P. R. China.

E-mail: nhxiu@bjtu.edu.cn

(Received September 2015; accepted February 2017)