entropy

MDPI

Article

Regularization Methods Based on the L_q-Likelihood for Linear Models with Heavy-Tailed Errors

Yoshihiro Hirose [1,2]

[1] Faculty of Information Science and Technology, Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan; hirose@ist.hokudai.ac.jp

[2] Global Station for Big Data and Cybersecurity, Global Institution for Collaborative Research and Education, Hokkaido University, Hokkaido 060-0814, Japan

Received: 26 August 2020; Accepted: 13 September 2020; Published: 16 September 2020

Abstract: We propose regularization methods for linear models based on the L_q-likelihood, which is a generalization of the log-likelihood using a power function. Regularization methods are popular for the estimation in the normal linear model. However, heavy-tailed errors are also important in statistics and machine learning. We assume q-normal distributions as the errors in linear models. A q-normal distribution is heavy-tailed, which is defined using a power function, not the exponential function. We find that the proposed methods for linear models with q-normal errors coincide with the ordinary regularization methods that are applied to the normal linear model. The proposed methods can be computed using existing packages because they are penalized least squares methods. We examine the proposed methods using numerical experiments, showing that the methods perform well, even when the error is heavy-tailed. The numerical experiments also illustrate that our methods work well in model selection and generalization, especially when the error is slightly heavy-tailed.

Keywords: least absolute shrinkage and selection operator (LASSO); minimax concave penalty (MCP); power function; q-normal distribution; smoothly clipped absolute deviation (SCAD); sparse estimation

1. Introduction

We propose regularization methods based on the L_q-likelihood for linear models with heavy-tailed errors. These methods turn out to coincide with the ordinary regularization methods that are used for the normal linear model. The proposed methods work efficiently, and can be computed using existing packages.

Linear models are widely applied, and many methods have been proposed for estimation, prediction, and other purposes. For example, for estimation and variable selection in the normal linear model, the literature on sparse estimation includes the least absolute shrinkage and selection operator (LASSO) [1], smoothly clipped absolute deviation (SCAD) [2], Dantzig selector [3], and minimax concave penalty (MCP) [4]. The LASSO has been studied extensively and generalized to many models, including the generalized linear models [5]. As is well known, the regularization methods have many good properties. Many regularization methods are the penalized maximum likelihood estimators, that is, minimizing the sum of the negative log-likelihood and a penalty. The literature proposed various penalties. As described later, our regularization methods use another likelihood with existing penalties.

Because the regularization methods for the normal linear model are useful, they are sometimes used in linear models with non-normal errors. Here, popular errors include the Cauchy error and the t-distribution error, both of which are heavy-tailed errors. For example, Ref. [6] partly consider the Cauchy and t-distribution errors in their extensive experiments. These heavy-tailed

distributions are known to be q-normal distributions, which are studied in the literature on statistical mechanics [7–9]. The q-normal model is also studied in the literature on the generalized Cauchy distribution. For example, see [10–13].

In this study, we consider the problem of a linear regression with a q-normal error. We propose sparse estimation methods based on the L_q-likelihood, which is a generalization of the log-likelihood using a power function. The maximizer of the L_q-likelihood, the maximum L_q-likelihood estimator (MLqE), is investigated by [14] as an extension of the ordinary maximum likelihood estimator (MLE). Ref. [14] studies the asymptotic properties of the MLqE. However, we are interested in the regularization, not in the MLqE, because regularization estimators can be better than the MLqE. We examine the proposed methods using numerical experiments. The experiments show that our methods perform well in model selection and generalization, even when the error is heavy-tailed. Moreover, we consider the effects of the sample size, dimension and sparseness of the parameter, and value of the nonzero elements in the numerical experiments.

We also find that the proposed methods for linear models with q-normal errors coincide with the ordinary regularization methods that are applied to the normal linear model. This finding partly justifies the use of the ordinary regularization methods for linear regressions with heavy-tailed errors. Moreover, the proposed methods are penalized least squares methods, and can be efficiently computed by existing packages.

The rest of the paper is organized as follows. In Section 2, we introduce several tools, including the normal linear model, regularization methods, L_q-likelihood, and q-normal models. In Section 3, we describe the problem under consideration, that is, estimations in linear models with q-normal errors. Moreover, we propose several regularization methods based on the L_q-likelihood. In Section 4, we evaluate the proposed methods using numerical experiments. Section 5 concludes the paper.

2. Preliminaries

2.1. Normal Linear Model and Sparse Estimation

First, we introduce the normal linear model, the estimation of which is a basic problem in statistics and machine learning [15]. Furthermore, we briefly describe some well-known regularization methods.

The normal linear model is defined as follows. A response is represented by a linear combination of explanatory variables $x_1, x_2, \ldots, x_d$ as

$$y^a = \theta^0 + \sum_{i=1}^{d} x_i^a \theta^i + \varepsilon^a \qquad (a = 1, 2, \ldots, n), \tag{1}$$

where y^a is the response of the a-th sample, n is the sample size, d is the number of explanatory variables, x_i^a is the i-th explanatory variable of the a-th sample, ε^a is a normal error with mean zero and known variance, and the regression coefficient $\boldsymbol{\theta} = (\theta^0, \theta^1, \ldots, \theta^d)^\top$ is the parameter to be estimated. The normal linear model is equivalently given by

$$\boldsymbol{\mu} = X\boldsymbol{\theta},$$

where $\mu^a = \mathrm{E}[y^a]$ is the expectation of the response y^a, $\boldsymbol{\mu} = (\mu^a)$, and $X = (x_i^a)$ is a design matrix of size $n \times (d+1)$, with $x_0^a = 1$ $(a = 1, 2, \ldots, n)$. Moreover, we define a row vector $\boldsymbol{x}^a$ $(a = 1, 2, \ldots, n)$ as $\boldsymbol{x}^a = (1, x_1^a, x_2^a, \ldots, x_d^a)$, and a column vector $\boldsymbol{x}_i$ $(i = 0, 1, 2, \ldots, d)$ as $\boldsymbol{x}_i = (x_i^1, x_i^2, \ldots, x_i^n)^\top$, which results in $X = (\boldsymbol{x}^{1\top}, \boldsymbol{x}^{2\top}, \ldots, \boldsymbol{x}^{n\top})^\top = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d)$. Let $\boldsymbol{y} = (y^a)$ be the response vector of length n. We assume that each column vector $\boldsymbol{x}_i$ $(i = 1, 2, \ldots, d)$ is standardized, as follows: $\sum_{a=1}^{n} x_i^a = 0$ and $\|\boldsymbol{x}_i\| = 1$, for $i = 1, 2, \ldots, d$.

As is well known, some regularization methods for the normal linear model are formulated as an optimization problem in the form of

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2n}\|\boldsymbol{y}-X\boldsymbol{\theta}\|^2 + \rho_\lambda(\boldsymbol{\theta}) \right\}, \tag{2}$$

where $\rho_\lambda(\boldsymbol{\theta})$ is a penalty term, and $\lambda \geq 0$ is a regularization parameter. The LASSO [1] uses $\rho_\lambda(\boldsymbol{\theta}) = \lambda\|\boldsymbol{\theta}\|_1 = \lambda\sum_{i=1}^{d}|\theta^i|$. The path of the LASSO estimator when λ varies can be made by the least angle regression (LARS) algorithm [16]. The SCAD [2] uses

$$\rho_\lambda(\boldsymbol{\theta}) = \begin{cases} \displaystyle\sum_{i=1}^{d} \lambda|\theta^i| & (|\theta^i| \leq \lambda), \\ \displaystyle -\sum_{i=1}^{d} \frac{|\theta^i|^2 - 2a\lambda|\theta^i| + \lambda^2}{2(a-1)} & (\lambda < |\theta^i| \leq a\lambda), \\ \displaystyle\sum_{i=1}^{d} \frac{(a+1)\lambda^2}{2} & (a\lambda < |\theta^i|), \end{cases} \tag{3}$$

and the MCP [4] uses

$$\rho_\lambda(\boldsymbol{\theta}) = \lambda\sum_{i=1}^{d}\int_0^{|\theta^i|}\left(1-\frac{u}{\gamma\lambda}\right)_+ \mathrm{d}u, \tag{4}$$

where $a(>2)$ and $\gamma(>0)$ are tuning parameters.

The regularization problem given in (2) can be represented by

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{d+1}} \left\{ -\frac{1}{n}\log f(\boldsymbol{y}|\boldsymbol{\theta}) + \rho_\lambda(\boldsymbol{\theta}) \right\}, \tag{5}$$

where $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the probability density function of the statistical model. Note that $\log f(\boldsymbol{y}|\boldsymbol{\theta})$ is the log-likelihood.

2.2. L_q-Likelihood

The L_q-likelihood is a generalization of the log-likelihood that uses a power function instead of the logarithmic function. Let $\boldsymbol{y} = (y^1, y^2, \ldots, y^n)^\top$ be a vector of independent and identically distributed (i.i.d.) observations, and let $\boldsymbol{\theta}$ be a parameter of a statistical model. For $q > 0$ $(q \neq 1)$, the L_q-likelihood function is defined as

$$L_q(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{a=1}^{n} \log_q f(y^a|\boldsymbol{\theta}), \tag{6}$$

where $f(\cdot|\boldsymbol{\theta})$ is a probability density function of the statistical model, and

$$\log_q(u) = \frac{1}{1-q}(u^{1-q}-1) \quad (u>0)$$

is the q-logarithmic function [9]. For $q = 1$, we define

$$\log_1(u) = \log u \quad (u>0),$$

which is the ordinary logarithmic function. When $q = 1$, the L_q-likelihood is the log-likelihood.

The MLqE is defined as the estimator that maximizes the L_q-likelihood. [14] studied the asymptotic performance of the MLqE, showing that it enjoys good asymptotic properties (e.g., asymptotic normality).

2.3. q-Normal Model

Before defining the q-normal distribution [7–9], we introduce the q-exponential function. For $q > 0\,(q \neq 1)$, the q-exponential function is the inverse function of the q-logarithmic function, and is given by

$$\exp_q(u) = \{1 + (1-q)u\}^{\frac{1}{1-q}} \quad (u < -1/(1-q)).$$

For $q = 1$, the 1-exponential function is the ordinary exponential function

$$\exp_1(u) = \exp u \quad (u \in \mathbb{R}).$$

Using the q-exponential function, the q-normal model is given by

$$\begin{aligned} \mathcal{S}_q &= \{f_q(y|\,\xi,\sigma)|\xi \in \Xi, \sigma > 0\}, \\ f_q(y|\,\xi,\sigma) &= \frac{1}{Z_q} \exp_q\left\{-\frac{1}{3-q}\left(\frac{y-\xi}{\sigma}\right)^2\right\} \\ &= \frac{1}{Z_q}\left\{1 - \frac{1-q}{3-q}\left(\frac{y-\xi}{\sigma}\right)^2\right\}^{\frac{1}{1-q}}, \end{aligned}$$

where ξ is a location parameter, $\Xi \subset \mathbb{R}$ is the parameter space, and σ is a dispersion parameter. The constant Z_q is a normalizing constant.

We assume that $1 \leq q < 3$, which ensures that the sample space is the real line itself, not just part of it. Moreover, the parameter space is $\Xi = \mathbb{R}$ when $1 \leq q < 3$.

For example, the 1-normal model is the ordinary normal model. Another example is the Cauchy distribution for $q = 2$:

$$f_2(y|\mu,\sigma) = \frac{1}{\sigma B(\frac{1}{2},\frac{1}{2})}\left(1 + \frac{(y-\xi)^2}{\sigma^2}\right)^{-1},$$

where $B(\cdot,\cdot)$ is the beta function. Furthermore, the t-distribution of the degree of freedom ν is obtained for $q = 1 + 2/(\nu+1)$:

$$f_{1+2/(\nu+1)}(y|\mu,\sigma) = \frac{1}{\sqrt{\nu}\sigma B(\frac{\nu}{2},\frac{1}{2})}\left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}.$$

3. Problem and Estimation Method

3.1. Linear Model with q-Normal Error

In this subsection, we formulate our problem, that is, a linear regression with a heavy-tailed error. The errors of the Cauchy and t-distributions in linear models have been studied by researchers in the context of heavy-tailed errors [17–20]. However, they focused mainly on the least squares methods, whereas we are interested in sparse estimators. Moreover, our approach is based on the L_q-likelihood, not the ordinary log-likelihood.

We examine the problem of estimating the linear model given in (1) with i.i.d. errors from a q-normal distribution; henceforth, we refer to this as the q-normal linear model. In terms of probability distributions, we wish to estimate the parameter $\boldsymbol{\theta}$ of the q-normal linear model $\mathcal{M}_q$:

$$\begin{aligned}\mathcal{M}_q &= \{f(\cdot|\,\boldsymbol{\theta})|\,\boldsymbol{\theta} \in \mathbb{R}^{d+1}\},\\ f(\boldsymbol{y}|\,\boldsymbol{\theta}) &= \frac{1}{Z_q^n}\prod_{a=1}^{n}\exp_q\left\{-\frac{(y^a-\boldsymbol{x}^a\boldsymbol{\theta})^2}{3-q}\right\}\\ &= \frac{1}{Z_q^n}\prod_{a=1}^{n}\left\{1-\frac{1-q}{3-q}(y^a-\boldsymbol{x}^a\boldsymbol{\theta})^2\right\}^{\frac{1}{1-q}},\end{aligned} \tag{7}$$

where the dispersion parameter is assumed to be known ($\sigma = 1$). The 1-normal linear model is identical to the normal linear model, as described in Section 2.1.

3.2. L_q-Likelihood-Based Regularization Methods

We propose regularization methods based on the L_q-likelihood. For q-normal linear models, the proposed methods coincide with the original regularization methods for the normal linear model. In other words, we apply the ordinary regularization methods as if the error distribution were a normal distribution. The literature describes how to compute the proposed methods efficiently. Moreover, our method calculates the MLqE.

We define the L_q-likelihood for the q-normal linear model in (7) as (6), where $\boldsymbol{\theta}$ is the regression coefficient. Note that the components of $\boldsymbol{y}$ are not assumed to be identically distributed because their distributions are dependent on the explanatory variables.

The L_q-likelihood for the q-normal linear model is

$$\begin{aligned}L_q(\boldsymbol{\theta}|\boldsymbol{y}) &= \sum_{a=1}^{n}\log_q f(y^a|\boldsymbol{\theta})\\ &= \sum_{a=1}^{n}\log_q\left[\frac{1}{Z_q}\exp_q\left\{-\frac{(y^a-\boldsymbol{x}^a\boldsymbol{\theta})^2}{3-q}\right\}\right]\\ &= -\frac{Z_q^{q-1}}{3-q}\|\boldsymbol{y}-X\boldsymbol{\theta}\|^2 - n\log_q(Z_q),\end{aligned} \tag{8}$$

where the second term is a constant. The MLqE of the parameter $\boldsymbol{\theta}$ is defined as the maximizer of the L_q-likelihood. In the q-normal linear model, the MLqE is equal to the ordinary least square, the MLE for the normal linear model.

We propose a LASSO, SCAD, and MCP based on the L_q-likelihood by replacing the log-likelihood with the L_q-likelihood in the optimization problem in (5). That is, the L_q-likelihood-based regularization methods are given in the form of

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{d+1}}\left\{-\frac{1}{n}L_q(\boldsymbol{\theta}|\boldsymbol{y}) + \rho_\lambda(\boldsymbol{\theta})\right\}. \tag{9}$$

The penalty ρ_λ is $\lambda\|\boldsymbol{\theta}\|_1$ for the LASSO, (3) for the SCAD, and (4) for the MCP. Note that the estimator for $\lambda = 0$ is the MLqE. As a special case, the proposed methods are the ordinary regularization methods when $q = 1$.

Because of (8) and (9), for the q-normal linear models, the L_q-likelihood-based regularization methods are essentially the same as the penalized least square (2). In other words, we implicitly use the L_q-likelihood-based regularization methods when we apply the ordinary LASSO, SCAD, and MCP to data with heavy-tailed errors.

4. Numerical Experiments

In this section, we describe the results of our numerical experiments and compare the proposed methods. Here, we focus on model selection and generalization.

Our methods do not require additional implementations because the LASSO, SCAD, and MCP are already implemented in software packages. In the experiments, we use the `ncvreg` package of the software R.

4.1. Setting

The procedure for the experiments is as follows. We fix the value q of the q-normal linear model and the L_q-likelihood, the dimension d of the parameter $\boldsymbol{\theta}$, the ratio of nonzero components r_{nz} of $\boldsymbol{\theta}$, the true value θ_0 of the nonzero components of $\boldsymbol{\theta}$, and the sample size n. The value of q is selected from $1, 13/11, 3/2, 5/3, 2, 2.01, 2.1$, and 2.5, where $q = 13/11$ is the t-distribution with $\nu = 10$ degrees of freedom, $q = 3/2$ is the t-distribution with $\nu = 3$ degrees of freedom, and $q = 5/3$ is the t-distribution with $\nu = 2$ degrees of freedom. The sample size is $n = 100$ or $n = 1000$. The true parameter consists of $d \times r_{\mathrm{nz}}$ θ_0s and $d \times (1 - r_{\mathrm{nz}})$ zeros. All cases are illustrated in Table 1.

For each of $m = 1000$ trials, we create the design matrix X using the `rnorm()` function in R. The response $\boldsymbol{y}$ is generated as q-normal random variables using the `qGaussian` package. For the estimation, we apply the `ncvreg()` function to $(\boldsymbol{y}, X)$ with the default options; for example, the values of the tuning parameters are $a = 3.7$ and $\gamma = 3$.

Table 1. All cases in the experiments. Each case is studied for the values of q and n.

θ_0	$r_{\mathrm{nz}} = 0.2$		$r_{\mathrm{nz}} = 0.4$		$r_{\mathrm{nz}} = 0.6$		$r_{\mathrm{nz}} = 0.8$	
	$d = 10$	$d = 100$	$d = 10$	$d = 100$	$d = 10$	$d = 100$	$d = 10$	$d = 100$
10^0	1	5	9	13	17	21	25	29
10^1	2	6	10	14	18	22	26	30
10^2	3	7	11	15	19	23	27	31
10^3	4	8	12	16	20	24	28	32

To select one model and one estimate from a sequence of parameter estimates generated by a method, we use the AIC and BIC:

$$\mathrm{AIC} = -2 \log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + 2d', \tag{10}$$

$$\mathrm{BIC} = -2 \log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + d' \log n, \tag{11}$$

where d' is the dimension of parameters of the model under consideration. Moreover, we use other criteria based on the L_q-likelihood:

$$L_q\text{-AIC} = -2L_q p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + 2d', \tag{12}$$

$$L_q\text{-BIC} = -2L_q p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + d' \log n. \tag{13}$$

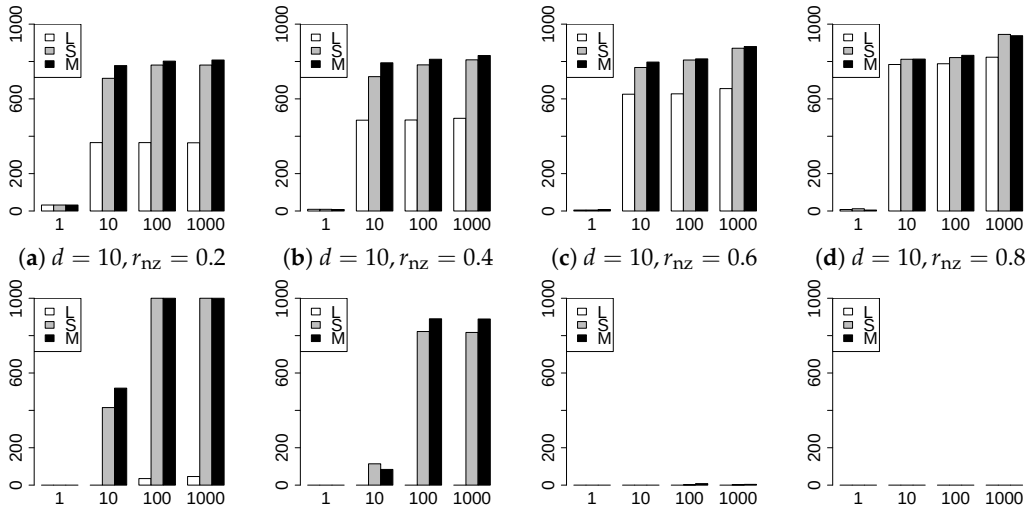
For a sequence $(\hat{\boldsymbol{\theta}}_{(k)})$ made by each of the methods, let $I_{(k)} = \{i | \, \hat{\theta}^i_{(k)} \neq 0\}$ and $\hat{\boldsymbol{\theta}}^{(k)}_{\mathrm{MLE}}$ the MLE of the model $\mathcal{M}_{(k)} = \{p(\cdot|\boldsymbol{\theta}) | \, \theta^j = 0 \ (j \notin I_{(k)})\}$. We call (10) with $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(k)}_{\mathrm{MLE}}$ AIC1, and (10) with $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{(k)}$ AIC2. Similarly, (11) with $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(k)}_{\mathrm{MLE}}$ is BIC1, and (11) with $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{(k)}$ is BIC2. The L_q-AIC and L_q-BIC are referred to in the same manner; for example, (12) with $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(k)}_{\mathrm{MLE}}$ is L_q-AIC1. Note that AIC1, BIC1, L_q-AIC1, and L_q-BIC1 are available only when the MLE exists; AIC2, BIC2, L_q-AIC2, and L_q-BIC2 are always applicable. Finally, we used cross-validation (CV) in addition to these information criteria.

4.2. Result

The results are presented in Figures 1–22, which report the best result for each method based on the various information criteria. We present the tables of the results of the numerical experiments in the Supplementary Material. In the figures, white bars represent LASSO, gray bars represent SCAD, and black bars represent MCP.

The model selection results are reported in Figures 1–14. The vertical axis indicates the number of trials (among $m = 1000$ trials) where a method selects the true model. Here, a larger value is better. The horizontal axis shows the value of θ_0.

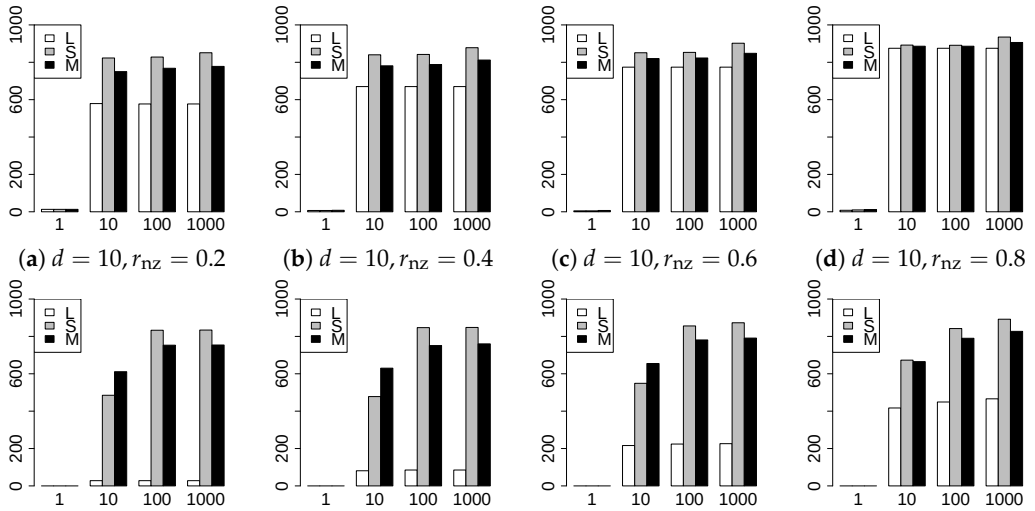
The generalization results are reported in Figures 15–22. To evaluate the generalization error of the proposed methods, we newly make $m = 1000$ independent copies $\{(\boldsymbol{y}'_1, X'_1), \ldots, (\boldsymbol{y}'_m, X'_m)\}$ in each trial. We computed the difference between $(\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_m)$ and the m predictions using each of the methods. The vertical axis indicates the average prediction error over m trials. In this case, a smaller value is better. The scaling of Figures 21 and 22 $(q = 5/3)$ is different from that of the other figures.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

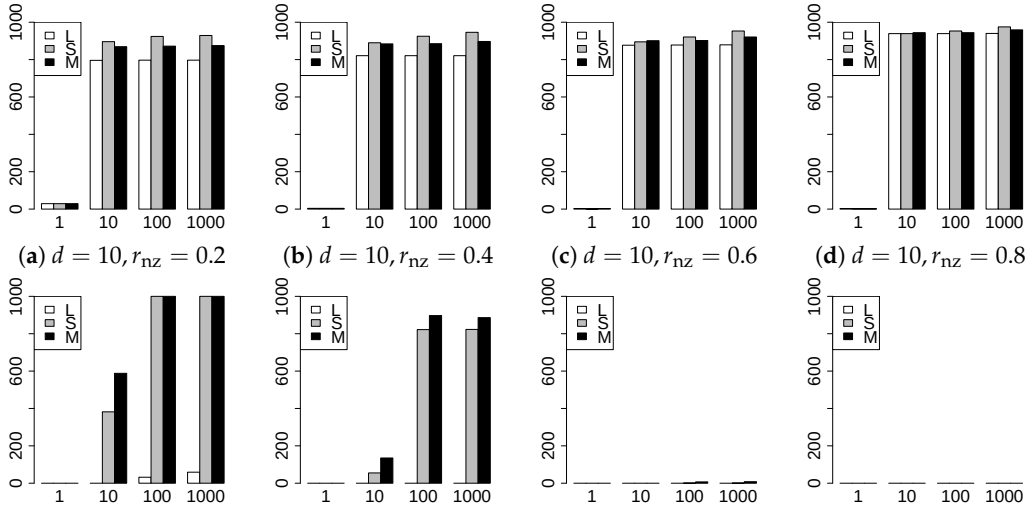
Figure 1. Model selection for $q = 1, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

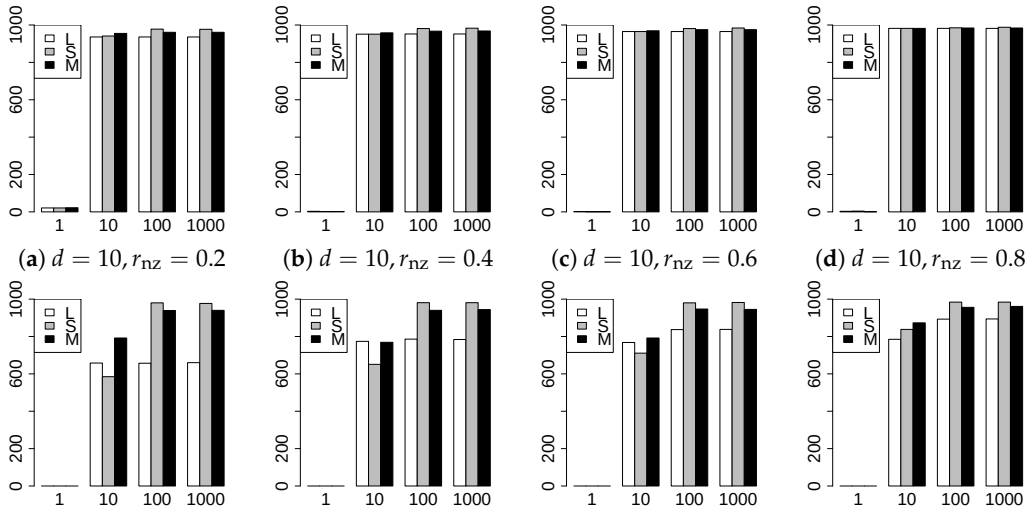
Figure 2. Model selection for $q = 1, n = 1000$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

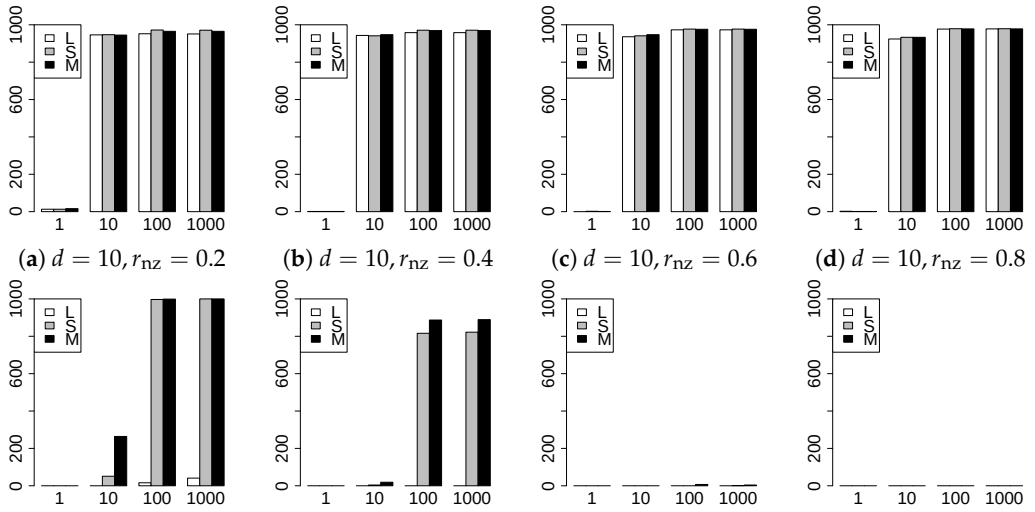
Figure 3. Model selection for $q = 13/11, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

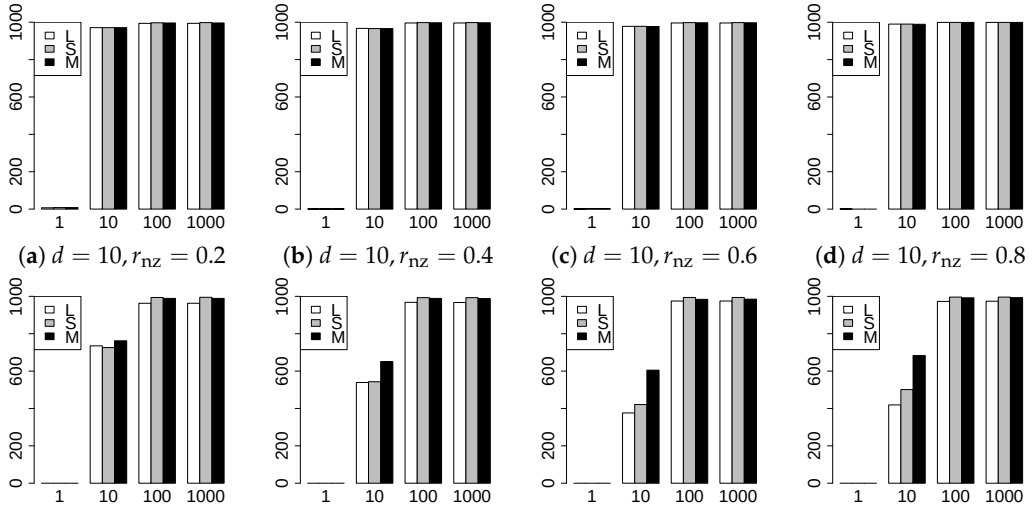
Figure 4. Model selection for $q = 13/11, n = 1000$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

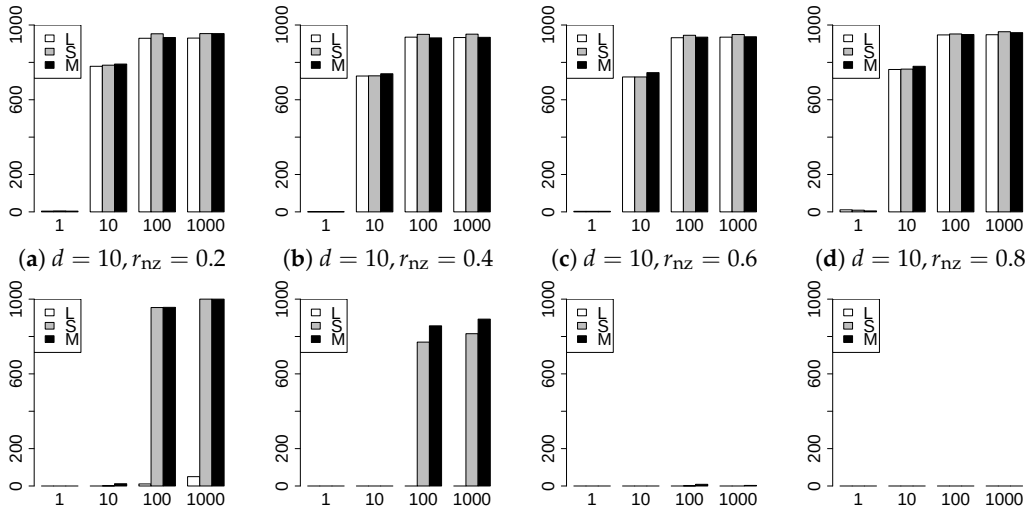
Figure 5. Model selection for $q = 3/2, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

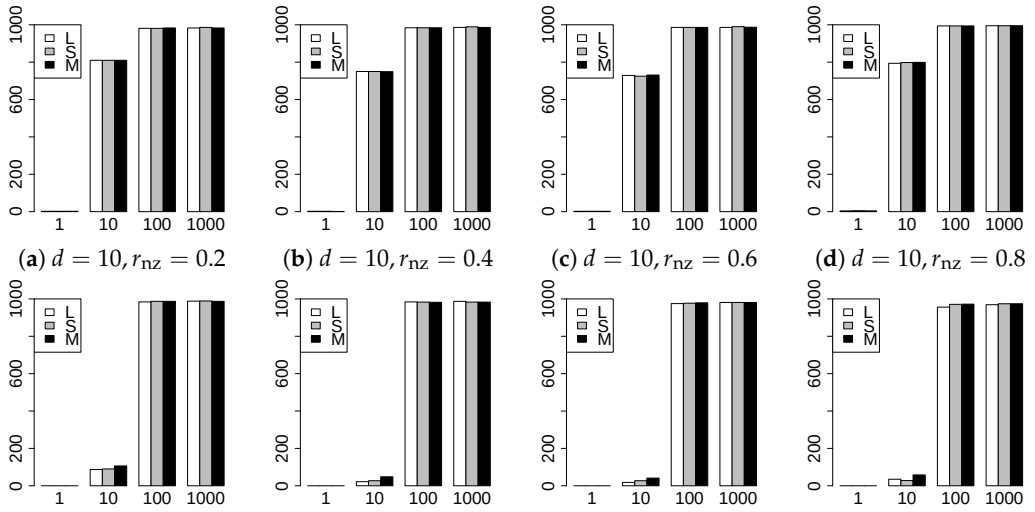
Figure 6. Model selection for $q = 3/2, n = 1000$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

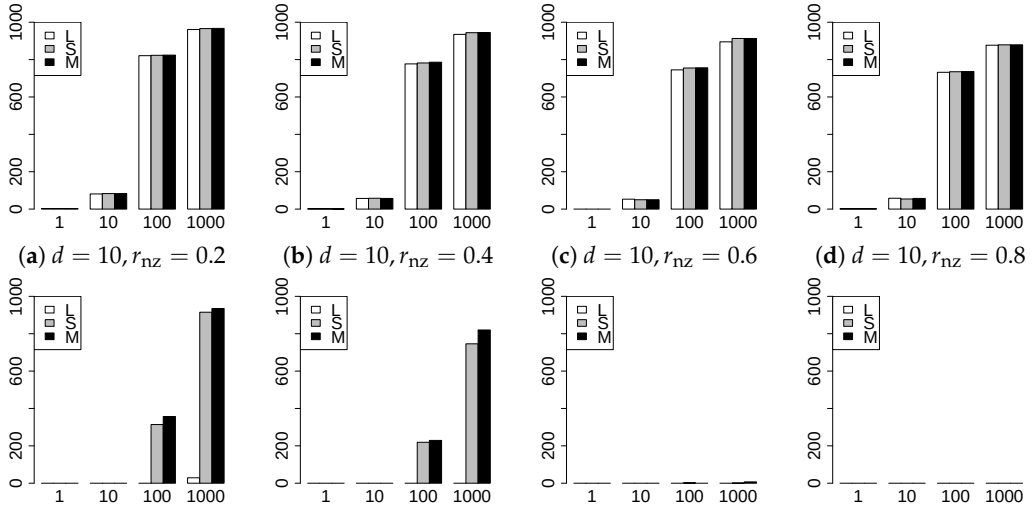
Figure 7. Model selection for $q = 5/3, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

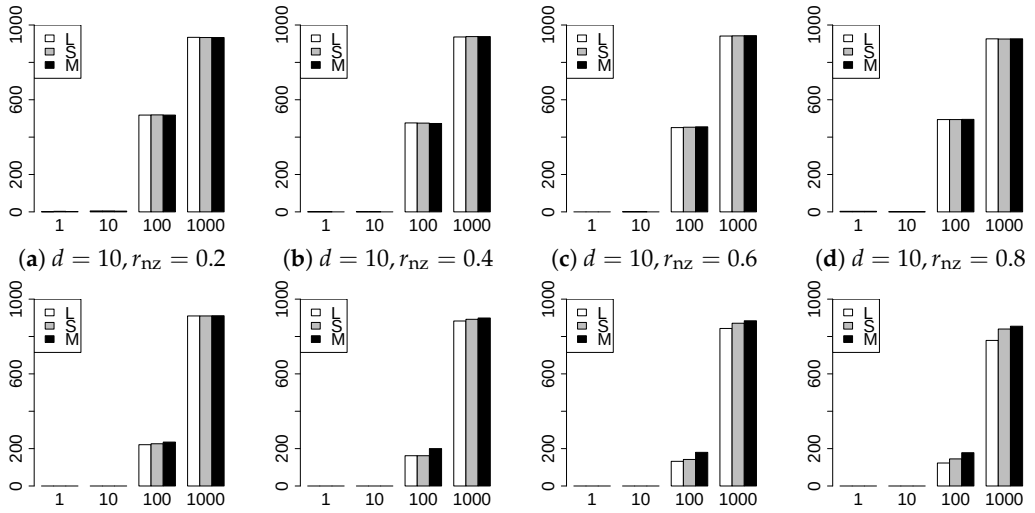
Figure 8. Model selection for $q = 5/3, n = 1000$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

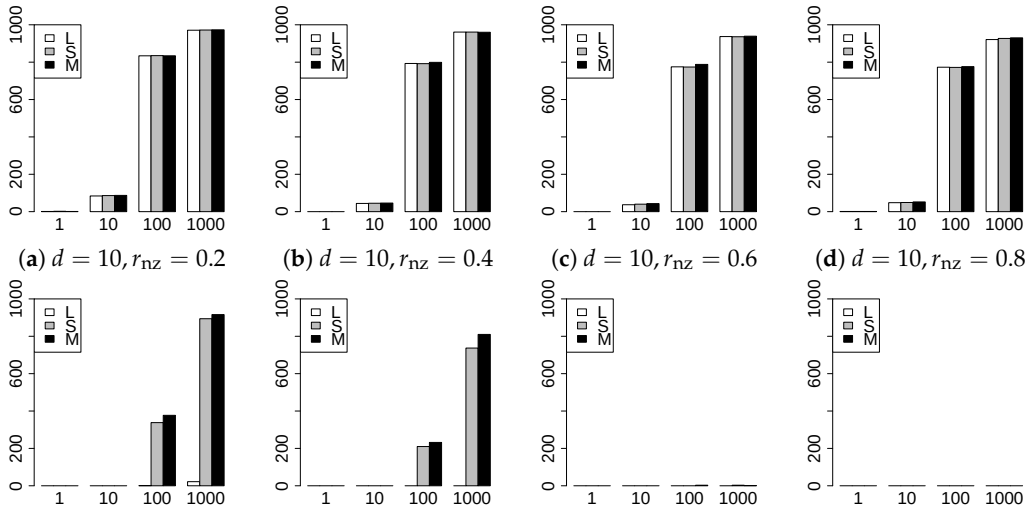
Figure 9. Model selection for $q = 2, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

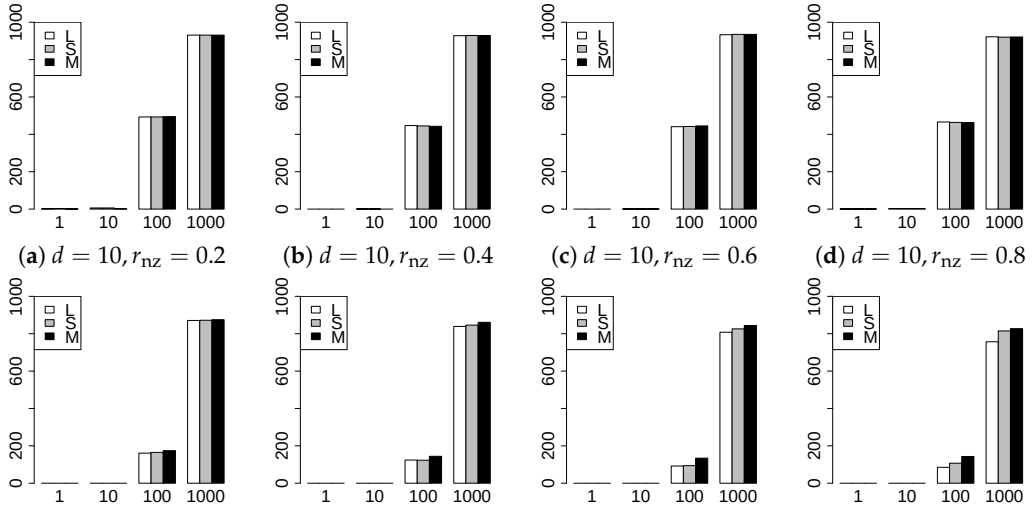
Figure 10. Model selection for $q = 2, n = 1000$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

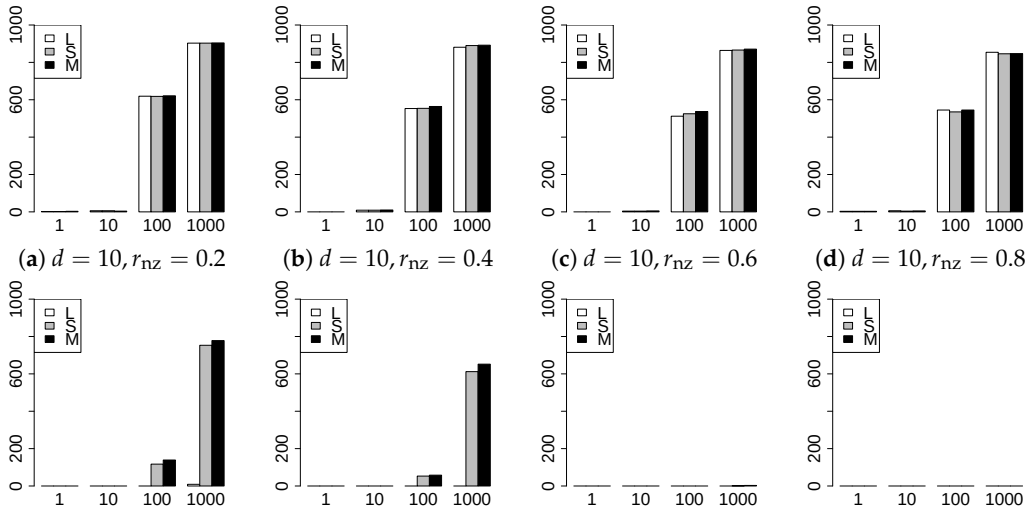
Figure 11. Model selection for $q = 2.01, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

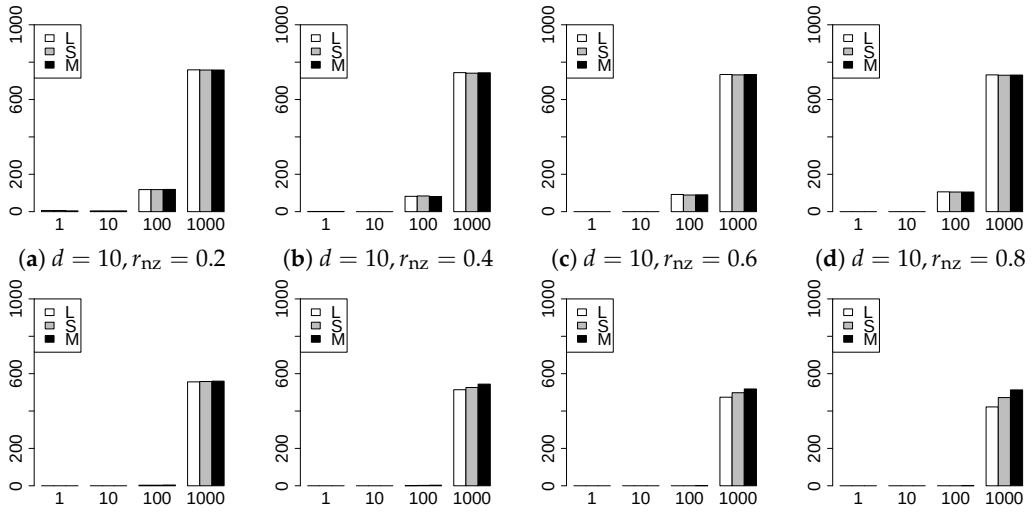
Figure 12. Model selection for $q = 2.01, n = 1000$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

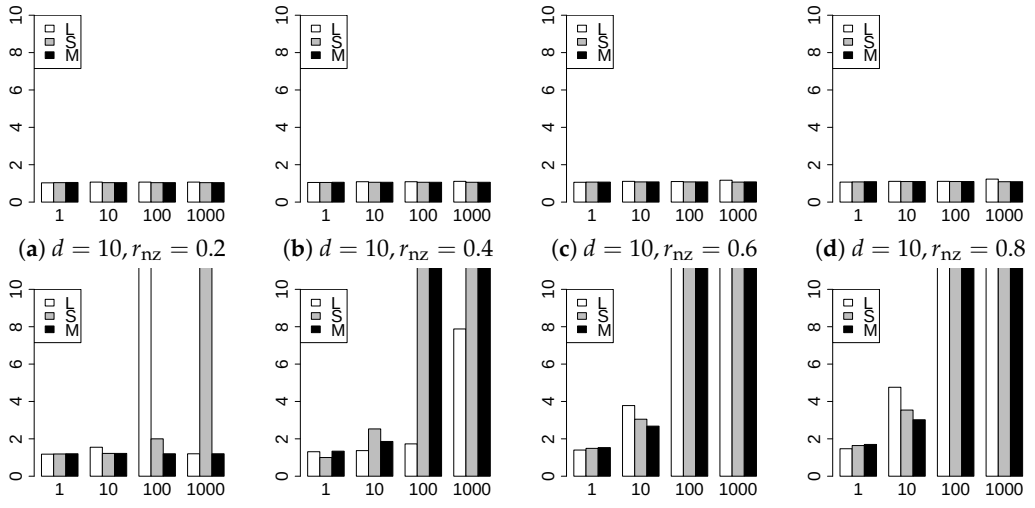
Figure 13. Model selection for $q = 2.1, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

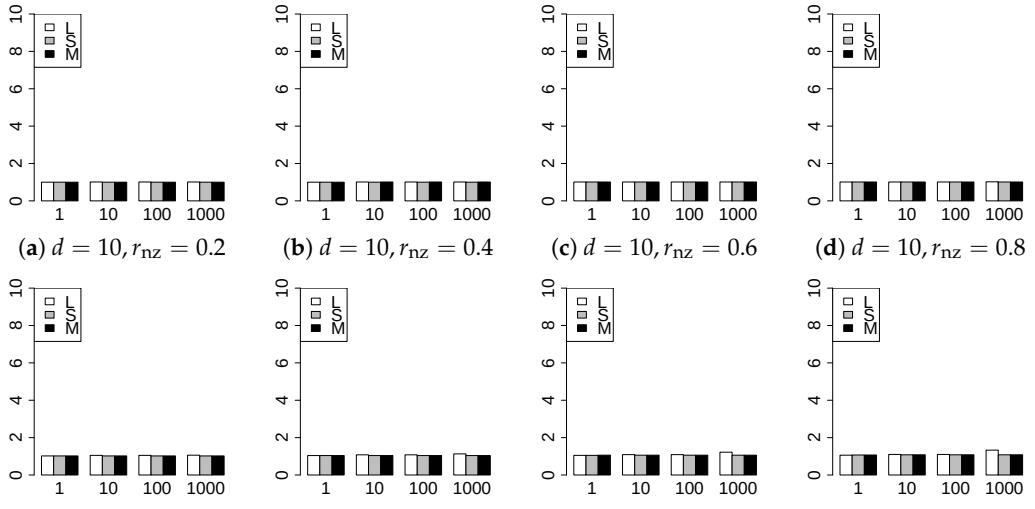
Figure 14. Model selection for $q = 2.1, n = 1000$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

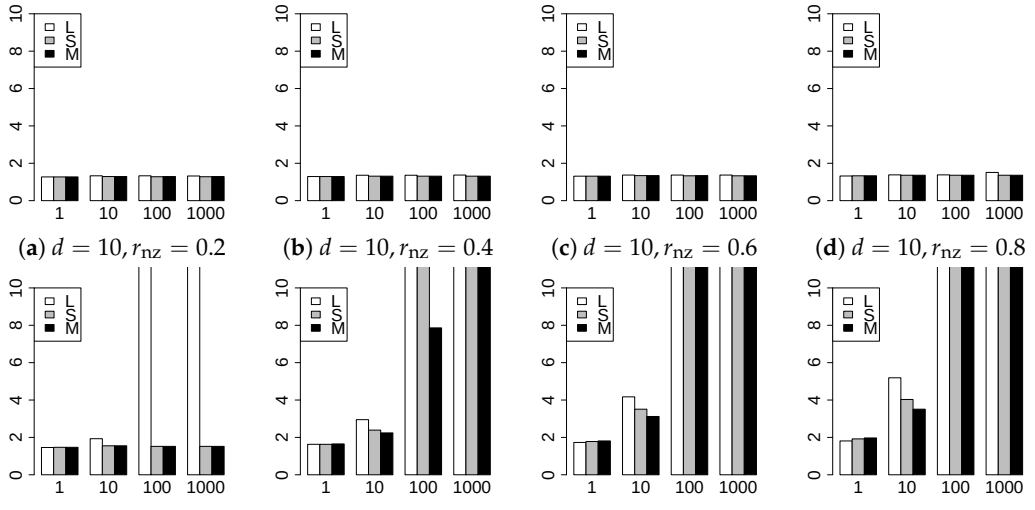
Figure 15. Generalization error for $q = 1, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

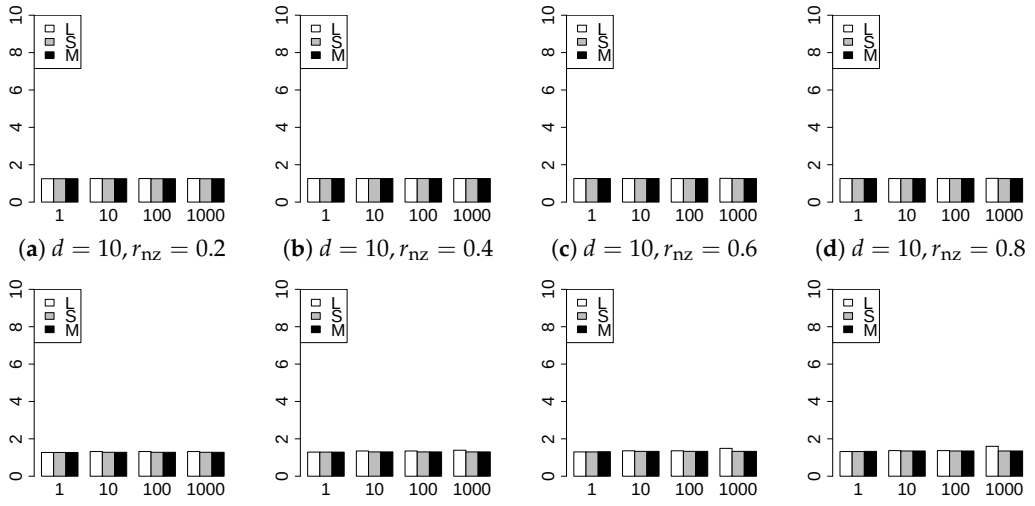
Figure 16. Generalization error for $q = 1, n = 1000$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

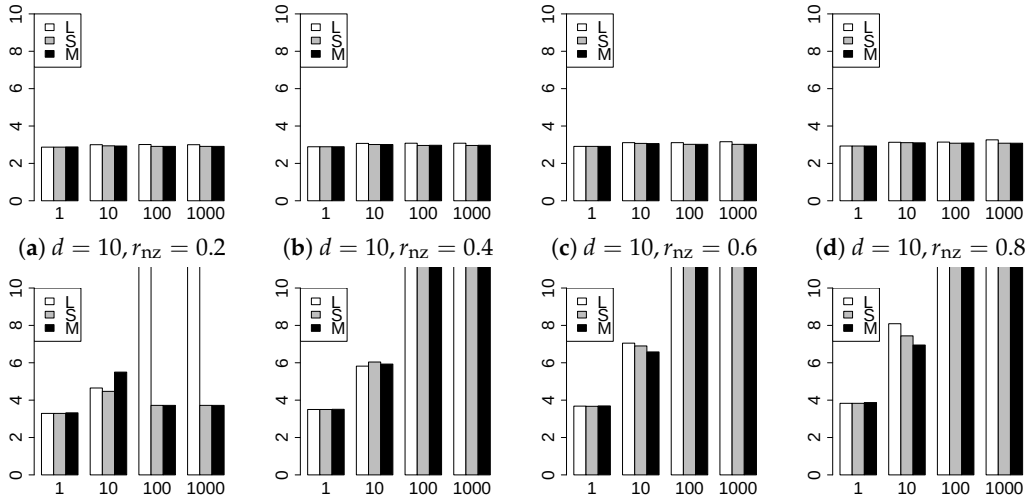
Figure 17. Generalization error for $q = 13/11, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

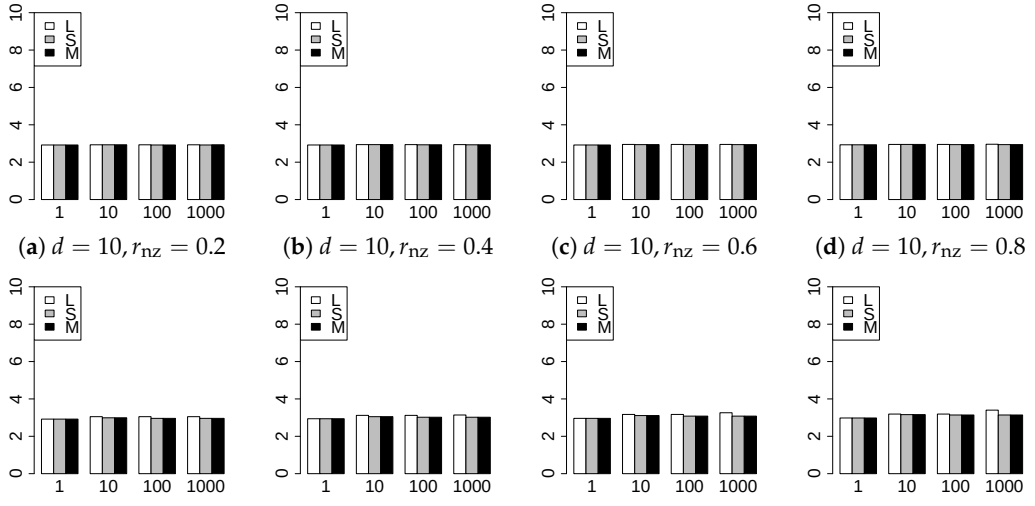
Figure 18. Generalization error for $q = 13/11, n = 1000$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

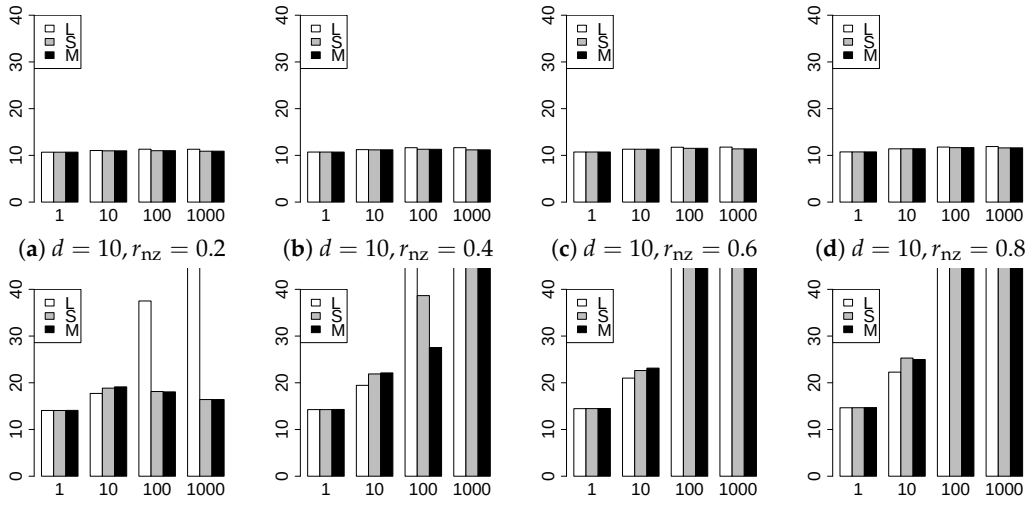
Figure 19. Generalization error for $q = 3/2, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

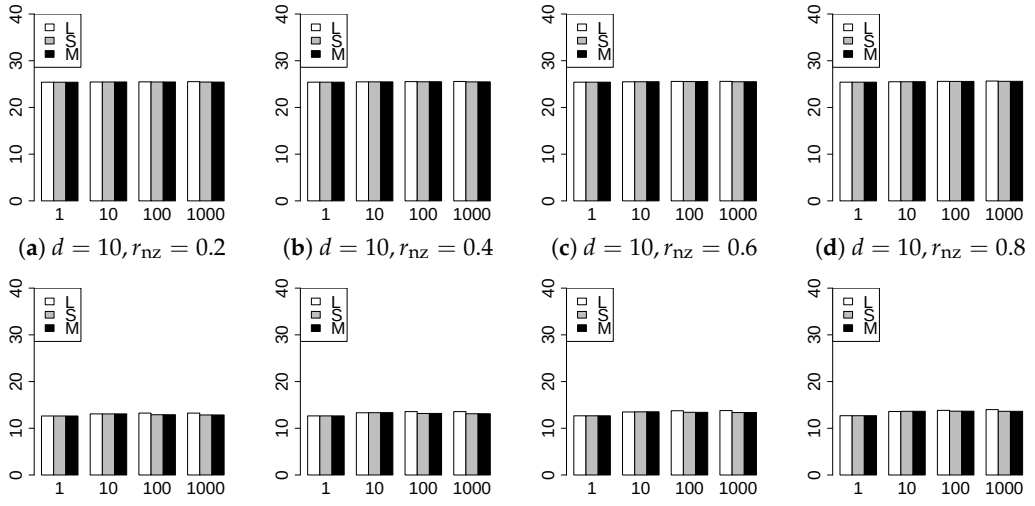
Figure 20. Generalization error for $q = 3/2, n = 1000$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

Figure 21. Generalization error for $q = 5/3, n = 100$.



(**a**) $d = 10, r_{\mathrm{nz}} = 0.2$ (**b**) $d = 10, r_{\mathrm{nz}} = 0.4$ (**c**) $d = 10, r_{\mathrm{nz}} = 0.6$ (**d**) $d = 10, r_{\mathrm{nz}} = 0.8$

(**e**) $d = 100, r_{\mathrm{nz}} = 0.2$ (**f**) $d = 100, r_{\mathrm{nz}} = 0.4$ (**g**) $d = 100, r_{\mathrm{nz}} = 0.6$ (**h**) $d = 100, r_{\mathrm{nz}} = 0.8$

Figure 22. Generalization error for $q = 5/3, n = 1000$.

Our first concern is whether the proposed methods work well. The results for $q = 1$ can be regarded as a reference for the other values of q. The figures show that the proposed methods work well in both model selection and generalization, especially for $q < 2$. The methods also perform well in terms of model selection for $q = 2, 2.01$, and 2.1. However, they perform poorly for $q = 2.5$ in terms of model selection and for $q \geq 2$ in terms of generalization. As anticipated, a large q makes the problem difficult.

Second, we evaluate the performance of the proposed methods, finding that the MCP performs best in most cases. In a few cases, the MCP performed similarly to or slightly worse than the other methods. For model selection, the cases with $q = 1, n = 1000$ and large θ_0 are exceptions. Furthermore, the LASSO performed worse than the SCAD and MCP.

Third, we consider the effect of r_{nz}, θ_0, d, and n, in addition to q. The cases with large r_{nz} and/or small θ_0 are difficult. Moreover, a large d makes the problems difficult. However, if we have a small q ($1 \leq q < 2$), large θ_0 ($\theta_0 = 10^2, 10^3$) and small r_{nz}, the problems with large d can be easier than those with small d. Furthermore, a small n makes the problems difficult in a similar manner to a large d. These observations imply that, for $1 \leq q < 2$, small-sample problems can be easier than large-sample problems if r_{nz} is small and θ_0 is large.

Fourth, the choice of information criterion changes the methods' performance. In terms of model selection, BIC2 was mostly the best for many values of q. For $3/2 \leq q \leq 2.1$, BIC1 was a little better than BIC2 if BIC1 was available. For $q = 1$ and $13/11$, BIC2 was better than BIC1. AIC1 and AIC2 were as good as BICs for $2 \leq q \leq 2.1$. Moreover, the L_q-BIC1 and -BIC2 were best only for $q = 3/2$, when BIC1 and BIC2 performed just as well. Overall, the L_q-information criteria performed poorly.

Furthermore, in terms of generalization, BIC2 was mostly the best. AIC2 was as good as BIC2, whereas AIC2 was sometimes a little worse than BIC2. The information criteria using the L_q-likelihood were poor for $q = 13/11$. For $q = 1, 3/2$, and $5/3$, the L_q-information criteria worked as well as the ordinary criteria and CV, except for some cases. The performance of CV was mostly good, but was occasionally very poor.

In summary, using an appropriate criterion, the proposed methods perform well for linear models with slightly heavy-tailed errors ($1 \leq q < 2$). Moreover, the proposed methods work in terms of model selection, even if the error is heavy-tailed ($2 \leq q < 2.5$). Overall, we recommend using the MCP and BIC2.

5. Conclusions

We proposed regularization methods for q-normal linear models based on the L_q-likelihood. The proposed methods coincide with the ordinary regularization methods. Our methods perform well for slightly heavy-tailed errors ($1 \leq q < 2$) in terms of model selection and generalization. Moreover, they work well in terms of model selection for heavy-tailed errors ($2 \leq q < 2.5$). A theoretical analysis of the proposed methods is left to future work.

Supplementary Materials: The following are available online at http://www.mdpi.com/1099-4300/22/9/1036/s1, Tables S1–S34: The results of the numerical experiments.

Funding: This work was partly supported by JSPS KAKENHI Grant Number JP18K18008 and JST CREST Grant Number JPMJCR1763.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. B* **1996**, *58*, 267–288. [CrossRef]
2. Fan, J.; Li, R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
3. Candes, E.; Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **2007**, *36*, 2313–2351. [CrossRef]
4. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
5. Park, M.Y.; Hastie, T. L_1-Regularization Path Algorithm for Generalized Linear Models. *J. R. Stat. Soc. B* **2007**, *69*, 659–677. [CrossRef]
6. Ahmed, S.E.; Kim, H.; Yildirim, G.; Yüzbasi, B. High-Dimensional Regression Under Correlated Design: An Extensive Simulation Study. In *International Workshop on Matrices and Statistics*; Springer: Cham, Switzerland, 2016; pp. 145–175.
7. Furuichi, S. On the maximum entropy principle and the minimization of the Fisher information in Tsallis statistics. *J. Math. Phys.* **2009**, *50*, 013303. [CrossRef]
8. Prato, D.; Tsallis, C. Nonextensive foundation of Lévy distributions. *Phys. Rev. E* **2000**, *60*, 2398–2401. [CrossRef] [PubMed]
9. Tsallis, C. *Introduction to Nonextensive Statistical Mechanics*; Springer: New York, NY, USA, 2009.
10. Alzaatreh, A.; Lee, C.; Famoye, F.; Ghosh, I. The Generalized Cauchy Family of Distributions with Applications. *J. Stat. Distrib. Appl.* **2016**, *3*, 12. [CrossRef]
11. Bassiou, N.; Kotropoulos, C.; Koliopoulou, E. Symmetric α-Stable Sparse Linear Regression for Musical Audio Denoising. In Proceedings of the 8th International Symposium on Image and Signal Processing and Analysis (ISPA 2013), Trieste, Italy, 4–6 September 2013; pp. 382–387.

12. Carrillo, R.E.; Aysal, T.C.; Barner, K.E. Generalized Cauchy Distribution Based Robust Estimation. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2008, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 3389–3392.
13. Carrillo, R.E.; Aysal, T.C.; Barner, K.E. A Generalized Cauchy Distribution Framework for Problems Requiring Robust Behavior. *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 1–19. [CrossRef]
14. Ferrari, D.; Yang, Y. Maximum L_q-Likelihood Estimation. *Ann. Stat.* **2010**, *38*, 753–783. [CrossRef]
15. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009.
16. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least Angle Regression. *Ann. Stat.* **2004**, *32*, 407–499.
17. Hinich, M.J.; Talwar, P.P. A Simple Method for Robust Regression. *J. Am. Stat. Assoc.* **1975**, *70*, 113–119. [CrossRef]
18. Holland, P.W.; Welsch, R.E. Robust Regression Using Iteratively Reweighted Least-Squares. *Commun. Stat. Theory Methods* **1977**, *6*, 813–827. [CrossRef]
19. Kadiyala, K.R.; Murthy, K.S.R. Estimation of regression equation with Cauchy disturbances. *Can. J. Stat.* **1977**, *5*, 111–120. [CrossRef]
20. Smith, V.K. Least squares regression with Cauchy errors. *Oxf. Bull. Econ. Stat.* **1973**, *35*, 223–231. [CrossRef]