Statistica Sinica **33** (2023), 1411-1434
doi:https://doi.org/10.5705/ss.202021.0187

ENSEMBLE SUBSET REGRESSION (ENSURE): EFFICIENT HIGH-DIMENSIONAL PREDICTION

Rong Zhu, Hua Liang and David Ruppert

Fudan University, George Washington University and Cornell University

Abstract: In high-dimensional prediction problems, we propose subsampling the predictors prior to the analysis. Specifically, we draw features using random sampling, and then fit a model and make predictions based on the sampled feature subset. This greatly reduces the dimension, storage, and computational bottlenecks. We explore this "subset regression" strategy under a linear regression framework. We propose an ensemble method that combines multiple subset regressions, called the ensemble subset regression (ENSURE) that reduces the uncertainty due to feature sampling. We provide a theoretical upper bound on the excess risk of the predictions computed in the subset regression, and provide theoretical support that the ensemble can improve the performance of the subset regression. Detailed empirical studies demonstrate that ENSURE performs well, better than methods that use all features.

Key words and phrases: Feature subset, high-dimensional, non-sparsity, random sampling, ridge regression, uniform sampling.

1. Introduction

High-dimensional problems arise in diverse scientific areas, and various statistical methods have been developed to deal with such problems. In this article, we propose a new method for predicting an outcome variable Y from the feature variables $X_1, X_2, \ldots, X_p$ measured on each of n individuals. We are interested in the high-dimensional scenario in which the dimension p is much larger than the sample size n.

However, high-dimensional prediction suffers frin the curse of dimensionality. Classic strategies use techniques to identify important variables or components in order to improve the prediction efficiency. These techniques include best subset selection methods such as the nonnegative garrote (Breiman and Spector (1992); Breiman (1995)), penalization–based methods such as the ridge regression (Hoerl (1962); Hoerl and Kennard (1970), the least absolute shrinkage and selection operator (LASSO,Tibshirani (1996), and its variants (Fan and Li (2001);

Corresponding author: Hua Liang, Statistics, George Washington University, St. NW, Washington 20052, USA. E-mail: hliang@gwu.edu.

Yuan and Lin (2006); Efron et al. (2004); Zou and Hastie (2005)), the Dantzig selector (Candes and Tao (2007)), and principal components (Bair et al. (2006)). It is well known that prediction is a fundamental concept and aim in machine learning, statistics, and other disciplines. In addition to prediction, the identification of important variables and model selection are often desirable; see Hastie, Tibshirani and Friedman (2009) for a review. Here, however, we investigate prediction rather than variable selection, because prediction is of independent interest. Recently, the prediction performance of high-dimensional modeling has garnered a substantial attention. Cook, Forzani and Rothman (2013) discusses the asymptotic characteristics of prediction in high-dimensional linear regression. Dalalyan, Hebiri and Lederer (2017) investigates the relationship between prediction performance and the correlations of the covariates for the LASSO.

In this paper, we propose a new approach to high-dimensional prediction based on a random sampling of the predictors prior to the analysis, and performing an l_2-regularized linear regression using this subset. This strategy, which we call "subset regression," uses the survey sampling principle, which states that one can obtain an accurate estimate from a subsample (a proxy for the finite-population-based estimate) taken using random sampling from a finite population. From the subset regression, we obtain an approximate prediction that acts as a proxy for the full high-dimensional prediction. Compared with methods that use all available features, the subset regression greatly reduces the computational cost, without compromising on prediction quality. We provide a theoretical upper bound on the excess risk of a prediction from a subset regression. This bound establishes the convergence rate for the prediction from the subset regression with respect to the true prediction.

More interestingly, we apply the ensemble method to the subset regression to reduce the uncertainty due to feature sampling. This is accomplished by generating multiple subsets, each of which is used to obtain a prediction, and then using the ensemble method to average the results. Our theoretical results reveal the effect on performance of applying the ensemble method, and show that the excess risk can be reduced by using this ensemble method. Empirical studies indicate that a subset regression with the ensemble method typically yields better predictions than those of methods based on all available features.

Our method is related to bagging (or its modification random forests) (Breiman (1996, 1998); Ho (1998); Breiman (2001); Brylla, Gutierrez-Osunab and Quek (2003)), which fits the same regression model many times to bootstrap-sampled versions of the training data, and averages the results. However, our method obtains multiple predictions by sampling features, rather than bootstrapping the

training data, as in the case of bagging. Here, we dhow the potential benefit of using the feature sampling ensemble for high-dimensional regression problems. This study contributes to the literature by showing that predictions using a single subset regression perform well, and that, under ensemble learning, a subset regression can improve the prediction efficiency. Empirically, our ensemble subset regression (ENSURE) often outperforms existing high-dimensional methods in terms of prediction accuracy. Our results suggest that, by combining multiple predictions, it may not be necessary to use all of the features in a high-dimensional data set.

There is a substantial body of parallel work on efficiently analyzing high-dimensional problems by using random projections, which are used to approximate a regression function in a high-dimensional linear space using projections onto a random subspace of a lower dimension. Maillard and Munos (2009) proposed the compressed least squares regression, which was later studied further by Fard et al. (2012). Guhaniyogi and Dunson (2015) proposed a Bayesian version of the compressed least squares regression. High-dimension projections are also used in other methods, such as support vector machines (Krishnan, Bhattacharyya and Hariharan (2007)) and discriminant analysis (Cannings and Samworth (2017)). However, we provide an alternative method based on feature sampling. To the best of our knowledge, this is the first work to use random sampling for high-dimensional predictions in linear models.

The remainder of the paper is organized as follows. Section 2 introduces the subset regression based on feature sampling. We provide an upper bound on the excess risk for the subset regression in Section 3. An ensemble method is proposed in Section 4. We report our simulation results in Section 5, where the performance of the subset regression is compared with that of methods that use all available features. Section 6 concludes the paper. The proofs of the theoretical results are provided in the Appendix.

Notation. Throughout this paper, for a matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{x}^{(k)}$ represents the kth column of $\mathbf{X}$, $\boldsymbol{x}_i$ represents the ith row of $\mathbf{X}$, and x_{ik} is the (i,k)th element of $\mathbf{X}$. For a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we define β_k as the kth element of $\boldsymbol{\beta}$.

2. Subset Regression Using Feature Sampling

In this section, we first introduce feature sampling, and then study the convergence properties of the feature sampling and propose a subset regression algorithm for high-dimensional predictions.

2.1. Problem setting

We study the high-dimensional prediction problem for the linear regression model

$$y_i = \mu_i + \epsilon_i, \tag{2.1}$$

where μ_i is a linear function of $\boldsymbol{x}_i$ such that

$$\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \sum_{k=1}^{p} x_{ik}\beta_k, \tag{2.2}$$

$\{\epsilon_i\}_{i=1}^n$ are the model errors with zero mean and σ^2 variance, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a parameter vector. We allow $p \gg n$ and $\boldsymbol{\beta}$ to be nonsparse. Suppose we have a finite set of training examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from model (2.1). We assume that, without loss of generality, the inputs $\{\boldsymbol{x}_i\}_{i=1}^n$ and output $\{y_i\}_{i=1}^n$ are centered. The matrix form of the model (2.1) is

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$, and $\mathbf{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$. We focus on the canonical instance of a high-dimensional prediction problem, that is, predicting $\mu_{\boldsymbol{x}} = \boldsymbol{x}^\top\boldsymbol{\beta}$ given an input $\boldsymbol{x}$, rather than estimating $\boldsymbol{\beta}$.

2.2. Feature sampling

The idea of feature sampling is to approximate the summation term $\sum_{k=1}^p x_{ik}\beta_k$ in Eqn. (2.2) using a summation based on a feature subset obtained by random sampling. Generate a *feature subset* $S \subset U = \{1, \ldots, p\}$ of size p_s ($p_s = |S|$) by *uniform sampling without replacement* from the full set of features $\{1, \ldots, p\}$, that is, draw $\mathbf{X}_s = \{\boldsymbol{x}^{(k)}, k \in S\}$ from $\mathbf{X}$. Note that we apply sampling without replacement rather than with replacement, because the latter may be less efficient (Särndal, Swensson and Wretman (2003)) and may induce additional collinearity. However, our analysis indicates that sampling with replacement is applicable, in principle, for the subset regression, and its theoretical analysis is easier than that of sampling without replacement.

Feature sampling is a standard survey sampling procedure that draws a "sample set" (subset) S of size p_s from the population $\{x_{i1}\beta_1, \ldots, x_{ip}\beta_p\}$ by *uniform sampling without replacement*. The sampling ratio is $f_s = p_s/p$. We estimate the population total in (2.2) from the subset S

$$\mu_{is} = \sum_{k \in S} f_s^{-1} x_{ik} \beta_k = f_s^{-1} \boldsymbol{x}_{is}^\top \boldsymbol{\beta}_s, \tag{2.4}$$

where $\boldsymbol{x}_{is}$ and $\boldsymbol{\beta}_s$ are subsets of $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$, respectively, corresponding to the subset S. Note that μ_{is} is inaccessible in practice because $\boldsymbol{\beta}$ is unknown. However, intuitvely, a prediction based on feature sampling should work well when μ_{is} approximates μ_i. Here, we explore this approximation from a theoretical viewpoint. It is easy to verify that $\mathrm{E}(\mu_{is}) = \mu_i$ and

$$\mathrm{Var}(\mu_{is}) = \frac{1 - f_s}{p_s} p \sum_{k=1}^{p} x_{ik}^2 \beta_k^2 < \frac{1}{f_s} \sum_{k=1}^{p} x_{ik}^2 \beta_k^2. \tag{2.5}$$

Theorem 1 follows directly from Markov's inequality.

Theorem 1. *We have that*

$$|\mu_{is} - \mu_i| \leq \rho^{-1} p_s^{-1/2} p^{1/2} \left(\sum_{k=1}^{p} x_{ik}^2 \beta_k^2 \right)^{1/2}, \tag{2.6}$$

with probability at least $1 - \rho$.

Theorem 1 indicates that the excess error of μ_{is} can be bounded by $p_s^{-1/2} p^{1/2} (\sum_{k=1}^{p} x_{ik}^2 \beta_k^2)^{1/2}$. Therefore, we define the condition that μ_i is c-compatible if there exists a constant $c \geq 0$ such that

$$\sum_{k=1}^{p} x_{ik}^2 \beta_k^2 \leq c p_s^{\delta} p^{-1} |\mu_i|^2, \text{ for some } 0 \leq \delta < 1. \tag{2.7}$$

From (2.6), as $p_s \to \infty$,

$$\frac{|\mu_{is} - \mu_i|}{|\mu_i|} = o_p(1), \tag{2.8}$$

where we assume $|\mu_i| \neq 0$. Eqn. (2.8) shows that μ_{is} can approximate μ_i well. Note that $|\mu_i|^2 \leq p \sum_{k=1}^{p} x_{ik}^2 \beta_k^2$, by the Cauchy–Schwartz inequality, and the equality holds if and only if $x_{ik}\beta_k$ are completely homogeneous. Roughly speaking, Condition (2.7) excludes some extreme sparsity cases, but the term p_s^{δ} in Condition (2.7) allows a certain degree of heterogeneity of the contribution of each feature to the prediction. Note that our excess risk bound, provided in Section 3, does not rely on Theorem 1. Instead, Theorem 1 provides a valuable hint that a subset prediction may be reasonable.

Remark 1. We perform *uniform sampling* for its computational simplicity and statistical efficiency; we verify the latter in Section 5. However, *uniform* sampling may lose statistical efficiency relative to data-driven importance sampling. It would be interesting to develop a more efficient data-based importance sampling method, because it makes sense to place more weight on those features that carry more information or are more important.

2.3. Prediction on feature subset

From Theorem 1, we know that it may make sense to replace $\boldsymbol{\mu} = \sum_{k\in U} \boldsymbol{x}^{(k)}\beta_k$ with the approximation $\widetilde{\boldsymbol{\mu}} = f_s^{-1}\sum_{k\in S}\boldsymbol{x}^{(k)}\beta_k$ to construct the least squares estimate based on the feature subset. Note that the goal of this study is prediction. Therefore, we ignore the constant f_s, which can be absorbed into β_k, and simply rewrite the approximation as $\widetilde{\boldsymbol{\mu}} = \sum_{k\in S}\boldsymbol{x}^{(k)}\beta_k$ without loss of generality. Furthermore, subset size p_s may not be much smaller, but is typically larger than the sample size n, so we add an l_2 regularized penalty to the subset regression, that is, we minimize

$$\mathrm{RSS}_{\mathrm{S}}(\boldsymbol{\beta}_s) = \left\| \boldsymbol{y} - \sum_{k\in S}\boldsymbol{x}^{(k)}\beta_k \right\|^2 + \lambda\sum_{k\in S}\beta_k^2, \tag{2.9}$$

where $\lambda > 0$ is a penalty parameter. This is a statement about the feature subset drawn using sampling without replacement.

By minimizing the function $\mathrm{RSS}_{\mathrm{S}}(\boldsymbol{\beta}_s)$ based on the feature subset, we obtain the estimator $\widehat{\boldsymbol{\beta}}_{s,\mathrm{RR}}$

$$\widehat{\boldsymbol{\beta}}_{s,\mathrm{RR}} = \left(\mathbf{X}_s^\top\mathbf{X}_s + \lambda\mathbf{I}_s\right)^{-1}\mathbf{X}_s^\top\boldsymbol{y}, \tag{2.10}$$

where we use generalized cross-validation (GCV) to choose the regularized parameter λ. From (2.4) and (2.10), the prediction of μ_i are

$$\widehat{\mu}_i = \sum_{k\in S} x_{ik}\widehat{\beta}_{s,k},$$

where $\widehat{\beta}_{s,k}$ is the kth element of $\widehat{\boldsymbol{\beta}}_{s,\mathrm{RR}}$. Thus, the approximate fitted values of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ is

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_s &= \sum_{k\in S}\boldsymbol{x}^{(k)}\widehat{\beta}_{s,k} = \mathbf{X}_s\widehat{\boldsymbol{\beta}}_{s,\mathrm{RR}} \\ &= \mathbf{X}_s\left(\mathbf{X}_s^\top\mathbf{X}_s + \lambda\mathbf{I}_s\right)^{-1}\mathbf{X}_s^\top\boldsymbol{y}.\end{aligned} \tag{2.11}$$

Our proposed subset regression is presented in Algorithm 1 below.

Algorithm 1. Subset Regression Algorithm.

- **Random Sampling Features:** Draw a feature subset using *uniform* sampling from the full features.
- **Estimation:** Solve the ridge regression problem using the feature subset, that is, minimize the function $\mathrm{RSS}_{\mathrm{S}}(\boldsymbol{\beta}_s)$ in Eqn.(2.9) to obtain the estimate $\widehat{\boldsymbol{\beta}}_{s,\mathrm{RR}}$.
- **Prediction:** Calculate the prediction values $\widehat{\boldsymbol{\mu}}$ using Eqn.(2.11) from the feature subset.

Remark 2. We use an l_2-regularization ("ridge") in our subset regression because it is simple and it performs well. We discuss performance of our subset regression with the l_2-regularization in Sections 3 and 5 from theoretical and empirical perspectives, respectively. The regularization can reduce the mean squared error by potentially allowing a slight increase in the bias, but dramatically reducing the variance. Shao and Deng (2012) theoretically investigated the consistency of the ridge estimator in a high-dimensional setting. The regression with the l_2-regularization has been shown to be effective in many applications and is remarkable in terms of its predictive performance (Frank and Friedman (1993); Malo, Libiger and Schork (2008)). Tibshirani (1996) and Fu (1998) compared predictions from a LASSO and a ridge regression, and found that the latter is competitive, even in some sparse settings.

3. An Upper Bound on Excess Risk

In this section, we provide an upper bound on the excess risk of the subset regression,

$$\mathcal{R}(\widehat{\boldsymbol{\mu}}_s) = \mathrm{E}\|\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2.$$

We re-express $\widehat{\boldsymbol{\mu}}_s$ by applying the Sherman–Morrison–Woodbury update. The classical method for a low-rank update of an inverse of a matrix is as follows: for $\mathbf{V} \in \mathbb{R}^{n\times m}$ and $\mathbf{D} \in \mathbb{R}^{m\times m}$, $(\mathbf{V}^\top\mathbf{V} + \mathbf{D})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{V}^\top(\mathbf{V}\mathbf{D}^{-1}\mathbf{V}^\top + \mathbf{I})^{-1}\mathbf{V}\mathbf{D}^{-1}$. We apply this equality to $\left(\mathbf{X}_s^\top\mathbf{X}_s + \lambda\mathbf{I}_s\right)^{-1}$, yielding the equation

$$\left(\mathbf{X}_s^\top\mathbf{X}_s + \lambda\mathbf{I}_s\right)^{-1} = \lambda^{-1}\mathbf{I}_s - \lambda^{-1}\mathbf{X}_s^\top\left(\lambda\mathbf{I}_n + \mathbf{X}_s\mathbf{X}_s^\top\right)^{-1}\mathbf{X}_s. \tag{3.1}$$

By inserting (3.1) into (2.11), we have

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu} =& (\mathbf{X}_s\mathbf{X}_s^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{X}_s\mathbf{X}_s^\top(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) - \mathbf{X}\boldsymbol{\beta} \\ =& -\lambda(\mathbf{X}_s\mathbf{X}_s^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}_s\mathbf{X}_s^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{X}_s\mathbf{X}_s^\top\boldsymbol{\epsilon}\end{aligned}$$

$$=: \boldsymbol{d}_B + \boldsymbol{d}_V. \tag{3.2}$$

The term $\boldsymbol{d}_B = -\lambda(\mathbf{X}_s\mathbf{X}_s^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{X}\boldsymbol{\beta}$ is the bias of the subset regression prediction, and $\boldsymbol{d}_V = (\mathbf{X}_s\mathbf{X}_s^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{X}_s\mathbf{X}_s^\top\boldsymbol{\epsilon}$ is the noise that determines the variance.

In the following, we investigate the performance of the subset regression as $\lambda \to 0$, that is,

$$\begin{aligned}
&\boldsymbol{d}_B \to -[\mathbf{I}_n - \mathbf{X}_s(\mathbf{X}_s^\top\mathbf{X}_s)^+\mathbf{X}_s^\top]\mathbf{X}\boldsymbol{\beta}, \text{ and} \\
&\boldsymbol{d}_V \to \mathbf{X}_s(\mathbf{X}_s^\top\mathbf{X}_s)^+\mathbf{X}_s^\top\boldsymbol{\epsilon}.
\end{aligned}$$

Note that the bias term of the ridge regression based on all available features goes to zero as $\lambda \to 0$, but the bias $\boldsymbol{d}_B$ of our subset regression does not. Thus, compared with the ridge method that uses all features, the error caused by the bias in the subset regression is larger. On the other hand, the error caused by the variance in the subset regression may decrease because the dimension decreases from p to p_s.

Based on the decomposition of Eqn.(3.2), we are now ready to present our main theoretical result, that is, an upper bound on the excess risk of $\hat{\boldsymbol{\mu}}_s$.

Theorem 2. *Assume that* $\mathbf{X}$ *is full row rank and denote* $d_1 \geq \cdots \geq d_n > 0$ *as the ordered nonzero eigenvalues of* $p^{-1}\mathbf{X}\mathbf{X}^\top$. *Define* $a = \max\{\|\boldsymbol{x}^{(k)}\|^2\}_{k=1}^p$. *Let* $1 > \rho, \kappa > 0$. *A feature subset of size* p_s *is drawn randomly from all* p *features by uniform sampling without replacement. If*

$$p_s > \left(\ln\frac{n}{\rho}\right)\frac{\kappa/3 + ad_n^{-1}}{\kappa^2/2}, \tag{3.3}$$

then

$$\mathcal{R}(\widehat{\boldsymbol{\mu}}_s) \leq \frac{1}{(1-\kappa)^2}\left[\sum_{i=1}^n \frac{\lambda^2\mu_i^2}{(pd_i+\lambda)^2} + \sigma^2\sum_{i=1}^n \frac{pd_i}{pd_i+\lambda}\right],$$

with probability at least $1 - \rho$.

Now, we examine the bound on $\mathcal{R}(\widehat{\boldsymbol{\mu}}_s)$ in Theorem 2. The bound has three terms: the risk from the model approximation, represented by $\sum_{i=1}^n \lambda^2\mu_i^2 / (pd_i+\lambda)^2$, the risk from the noise, represented by $\sigma^2\sum_{i=1}^n pd_i/(pd_i+\lambda)$, and the risk from the subsampling, represented by $(1-\kappa)^{-2}$. The first two terms are the risk from the ridge regression based on all available features. When λ goes to zero, the first term becomes small, and the second term becomes large. The third term $(1-\kappa)^{-2}$ controls the extra risk from the subsampling

step. Therefore, Theorem 2 provides a theoretical foundation for the subset regression showing that the extra risk can be small when the feature subset size p_s is sufficiently large. This property indicates that the computationally cheap algorithm is also statistically efficient for high-dimensional predictions. Next, we discuss the requirement on p_s in Eqn. (3.3). Assuming each element in $\mathbf{X}$ is bounded, we have $a = \max\{\|\boldsymbol{x}^{(k)}\|^2\}_{k=1}^p = O(n)$. Assuming that each $\boldsymbol{x}_i$ is independent and identically distributed (i.i.d.), $p^{-1}\mathbf{X}\mathbf{X}^\top$ is close to a diagonal matrix that has the same diagonal element. This implies that d_i and d_i^{-1} are constants of order $O(1)$. Thus, in Eqn. (3.3), given constants κ and ρ, $(\ln(n/\rho))(\kappa/3 + ad_n^{-1})/(\kappa^2/2) = O(n \log n)$, implying that the requirement on p_s can still be much smaller than p, given that $n \ll p$ in our high-dimensional setting. As a result, the feature subset size p_s that satisfies Eqn.(3.3) can be dramatically smaller than p of the original features.

To choose p_s, we provide a cross-validation strategy, and examine its numerical performance in the empirical studies below. Specifically, equation (3.3) describes the theoretical relationship between the feature subset size and the approximation accuracy, that is, $p_s = O(n \log n)$. We show empirically using simulated and real data sets in Section 5 that the risk based on the subset regression procedure can approach that based on the full model as p_s increases. Therefore, we suggest choosing p_s using $p_s = cn$, where c is some integer that can be chosen using validation.

Remark 3. Theorem 2 does not require the result in Theorem 1, that is, we do not require the term $\sum_{k=1}^p \|\boldsymbol{x}^{(k)}\|^2\beta_k^2$ to be as small as it is in Theorem 1. The condition in Theorem 2 is that the subset size p_s needs to be sufficiently large enough relative to the sample size n.

Remark 4. Although we choose simple uniform sampling, we can also use unequal sampling dependent on $\mathbf{X}$ for the proposed subset regression and the excess risk analysis. Specifically, assume that we sample the subset S by unequal sampling without replacement from the full features proportional to the sampling probabilities $\{\pi_1, \ldots, \pi_p\}$, such that $\sum_{k=1}^p \pi_k = 1$. The objective function Eqn.(2.9) is replaced by

$$\mathrm{RSS}_{\mathrm{S}}^{\boldsymbol{\pi}}(\boldsymbol{\beta}_s) = \left\| \boldsymbol{y} - \sum_{k\in S} \frac{1}{\pi_k} \boldsymbol{x}^{(k)} \beta_k \right\|^2 + \lambda \sum_{k\in S} \frac{1}{\pi_k} \beta_k^2.$$

Following the process of the subset regression, we get the approximate prediction of $\boldsymbol{\mu}$,

$$\widehat{\boldsymbol{\mu}}_s = \mathbf{X}_s \left(\mathbf{X}_s^\top \mathbf{X}_s + \lambda \boldsymbol{\Phi}_s \right)^{-1} \mathbf{X}_s^\top \boldsymbol{y},$$

where $\boldsymbol{\Phi}_s$ is the submatrix of $\boldsymbol{\Phi} = \operatorname{diag}\{\pi_1, \ldots, \pi_p\}$ corresponding to the subset S. Moreover, Theorem 2 still holds when $p_s > (\ln(n/\rho))(\kappa/3 + Ad_n^{-1})/(\kappa^2/2)$, where $A = \max\{(p\pi_k)^{-1}\|\boldsymbol{x}^{(k)}\|^2\}_{k=1}^p$. We supply the proofs in Appendix A.

4. Subset-Ensemble Prediction

In this section, we show that a mixture of subset regression predictions, which we call subset-ensemble prediction, yields a more accurate prediction than that of a single subset regression. This result is of great practical significance, given the prevalence of distributed computing frameworks in large-scale learning problems. The ensemble algorithm naturally fits a distributed computing environment in which the computational cost is roughly the same as that of the standard subset regression. The algorithm is very simple and easily implemented. Assume there are T predictions from the subset regression. Then, treat each prediction as though from an expert, and combine T experts to obtain an improved prediction. We present the ensemble process in Algorithm 2 below.

Algorithm 2. Ensemble Subset Regression Algorithm.

- **Repeat:** Repeat the following steps T times:
  - **(a) Draw a feature subset:** A feature subset of size p_s is sampled from the full features;
  - **(b) Make a single prediction:** A prediction from subset regression has output $\hat{\boldsymbol{\mu}}_{s,t}$;
- **Combine:** Majority vote using the equation.

$$\hat{\boldsymbol{\mu}}^{\text{ensure}} = T^{-1} \sum_{t=1}^{T} \hat{\boldsymbol{\mu}}_{s,t}, \tag{4.1}$$

From Algorithm 2, we get the risk of $\hat{\boldsymbol{\mu}}^{\text{ensure}}$,

$$\mathcal{R}(\widehat{\boldsymbol{\mu}}^{\text{ensure}}) = \mathrm{E}\|\widehat{\boldsymbol{\mu}}^{\text{ensure}} - \boldsymbol{\mu}\|_2^2 = \mathrm{E}\left\| \frac{1}{T} \sum_{t=1}^{T} (\hat{\boldsymbol{\mu}}_{s,t} - \boldsymbol{\mu}) \right\|_2^2.$$

We compare the risks $\mathcal{R}(\widehat{\boldsymbol{\mu}}^{\text{ensure}})$ and $\mathcal{R}(\widehat{\boldsymbol{\mu}}_s)$ in the following theorem.

Theorem 3. *Denote* $\sigma_B^2 = E\|\boldsymbol{d}_B\|^2 - \|E(\boldsymbol{d}_B)\|^2$ *and* $\sigma_V^2 = E(\|\boldsymbol{d}_V\|^2|\boldsymbol{y}) - \|E(\boldsymbol{d}_V|\boldsymbol{y})\|^2$.

$$\mathcal{R}(\widehat{\boldsymbol{\mu}}_s) - \mathcal{R}(\widehat{\boldsymbol{\mu}}_s^{ensure}) = \left(1 - \frac{1}{T}\right)\{\sigma_B^2 + E_m(\sigma_V^2)\}, \tag{4.2}$$

where "$E_m()$*" refers to the expectation over the model* (2.1).

Theorem 3 tells us an interesting observation about the subset-ensemble method. The risk performance depends on three factors: σ_B^2 from $\boldsymbol{d}_B$, σ_V^2 from $\boldsymbol{d}_V$, and the ensemble number T. Here, σ_B^2 denotes the variability of $\|\boldsymbol{d}_B\|$ from the feature sampling, and σ_V^2 denotes, fixing the responses $\boldsymbol{y}$, the variability of $\|\boldsymbol{d}_V\|$ from the feature sampling. After introducing the ensemble, the excess risk decreases by $(1 - 1/T)\{\sigma_B^2 + \mathrm{E}_m(\sigma_V^2)\}$. In other words, the ensemble algorithm improves the performance by reducing the risk from the variabilities of $\|\boldsymbol{d}_B\|$ and $\|\boldsymbol{d}_V\|$. From a computational viewpoint, although the subset-ensemble prediction requires T times more time than the single subset regression, the ensemble version can be easily parallelized, because all T experts can be computed simultaneously. Thus, for a cluster of T machines, the running time complexity of a subset-ensemble prediction is nearly equal to that of a single subset regression.

Choice of T. Theorem 3 shows that the benefit of the ensemble narrows as T increases. Empirical studies also verify this observation. In practice, it is enough to set $T = 10$ such that Algorithm 2 has a sufficiently nice prediction.

5. Numerical Experiments

In this section, we conduct detailed experiments to assess the performance of the subset regression. We compare ENSURE with several representative high-dimensional methods: the ridge regression (ridge), LASSO, sure independence screening (SIS), and random forests (RF). We implement the LASSO, SIS, and RF using the R packages `glmnet` (Friedman, Hastie and Tibshirani (2010)), `SIS` (Fan and Lv (2008)), and `randomForest` (Breiman (2001)), respectively. The default settings are used in the implementation of these methods. In the implementation of SIS, we use the SCAD penalty (default) and tune the regularization parameter using the BIC. The penalizing parameters for the subset regression and ridge regression are selected using GCV.

In Section 5.1, we report the results of extensive simulation experiments. In Section 5.2, we apply ENSURE to two real data sets. In both sets of examples, the proposed method outperforms the methods based on all available features.

5.1. Simulation studies

We generated data from the linear model $y_i = \mu_i + \epsilon_i$ with $\mu_i = 1 + \boldsymbol{x}_i^\top \boldsymbol{\beta}$, for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$. For evaluation purposes, we consider two design matrix settings: (1) autoregressive correlation $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{ij} = \rho^{|i-j|}$ (referred as to the AR data set); and (2) an input matrix from a heavier tail, that is, $\boldsymbol{x}_i$ is drawn independently from a t_2 distribution (referred as to the T2 data set). Specifically, $\boldsymbol{x}_i = \boldsymbol{z}_i/\sqrt{u}$, where $\boldsymbol{z}_i \sim N(0, \boldsymbol{\Sigma})$ independent of $2u \sim \chi_2^2$. To control the signal-to-noise ratio, we define $\sigma^2 = n^{-1}\sum_{i=1}^n (\mu_i - \bar{\mu})^2/r$, where $\bar{\mu} = n^{-1}\sum_{i=1}^n \mu_i$, and set r to three for all experiments. For $\boldsymbol{\beta}$, we consider the following scenarios:

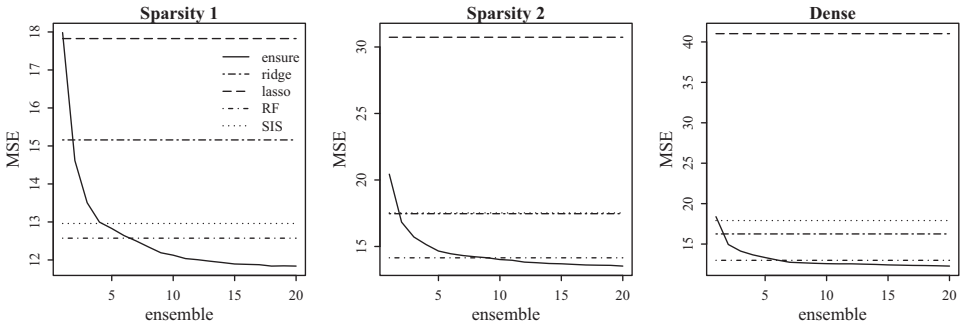
(A) The first 10 regression coefficients are drawn uniformly between -1 to 1, and the rest are zero;

(B) The first 100 regression coefficients are drawn uniformly between -0.5 to 0.5, and the rest are zero;

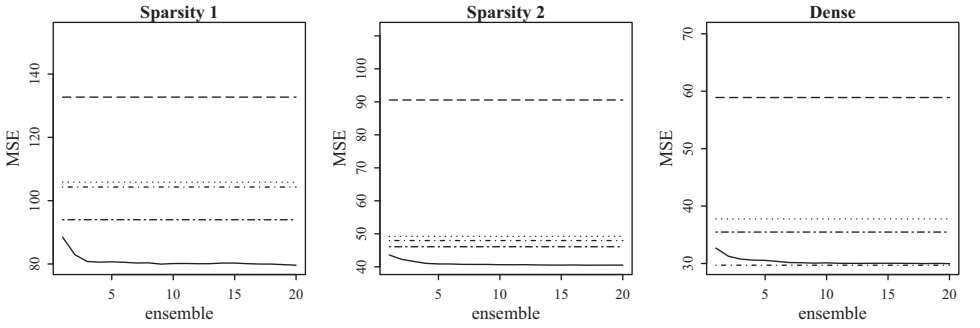
(C) Regression coefficients are drawn uniformly between -0.1 to 0.1.

We normalize $\boldsymbol{\beta}$ so that $\|\boldsymbol{\beta}\| = 1$. The third scenario is referred to as the *Dense* case, and scenario A (*Sparsity 1*) is much sparser than scenario B (*Sparsity 2*). The sample size $n = 50$, the data dimension is $p = 10{,}000$, and $\rho = 0.5$ for the main experiments. We also consider other settings of (n, p) and ρ to further our understanding and for comparison purposes. Because the true μ_i is known in the simulation, we measure the performance with MSE $= \sum_{i=1}^n (\hat{\mu}_i - \mu_i)^2$, where $\hat{\mu}_i$ are the predictions from the different methods.

Performance of our method. We evaluate the performance of ENSURE. We set the ensemble number T equal to 20 and the subset size to $p_s = cn$, where c is five-fold cross-validated among $\{0.5, 1, 2, 4, 6, 10\}$ based on only one feature subset sampled from the full features. We plot the results for AR and T2 in Figure 1. From the plots, we have several observations. (a) The MSE decreases as the number of ensembles, T, grows. Combining the ensemble, the subset regression even outperforms the regression based on the full features remarkably. However, effect of increasing T becomes weak after $T > 10$, suggesting, that we should choose $T = 10$ in practice. (b) For the AR and T2 settings, the ensemble subset prediction outperforms the methods based on the full features, except for Dense of T2. Note that combining 10 predictions based on a subset of size $p_s = 500$ uses at most 5,000 features, but it achieves much better performance than that of the methods based on all features. (c) Comparing the Sparsity 1, sparsity 2, and



(a) AR



(b) T2

Figure 1. The performance of the ensemble method for the AR and T2 settings: the MSE against the ensemble number. From left to right: Sparsity 1, sparsity 2, and Dense. From top to bottom: AR and T2. The dashed lines correspond to the ridge, LASSO, RF, and SIS.

Dense cases yields a useful observation. Figure 1 tells us that ENSURE performs well across all sparsity scenarios. It attains relatively better performance as the coefficients become sparser. This is a striking observation. When the coefficients are very sparse, the probability of obtaining the important variables using a subset regression is tiny. Thus, when our aim is prediction, identifying the important variables may not be very necessary, because of the curse of dimensionality and possible correlation among the variables. Therefore, for high-dimensional prediction, ENSURE provides statistical benefits and is computationally simple.

Effect of the p_s choice. We verify the performance of the ensemble subset prediction under various $p_s \in \{50, 100, 200, 300, 500\}$, and report the results in Figure 2. The size p_s has some effect on the performance. The case of $p_s = 200$ performs best for Sparsity 1, Sparsity 2, and Dense. Therefore, the ensemble subset prediction may require that p_s be not too large. More importantly, we

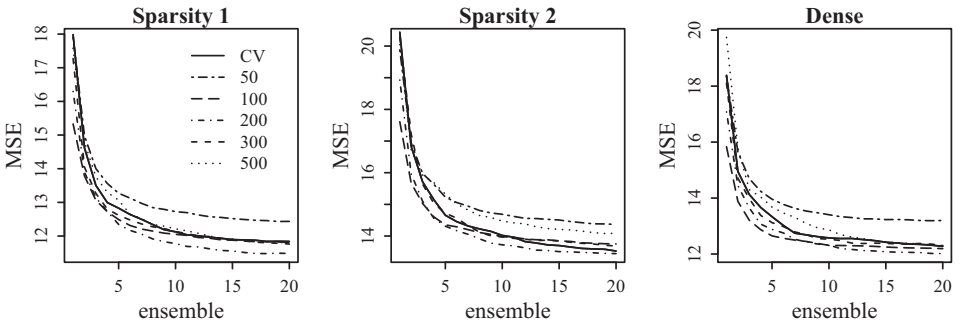


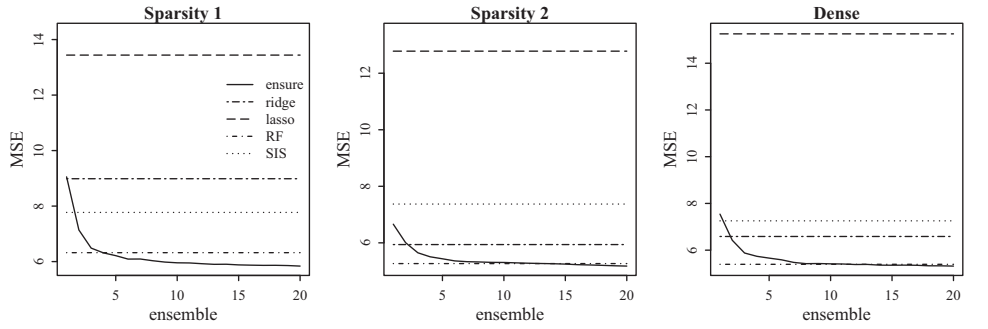
Figure 2. The performance of the ensemble method for the AR setting. Various $p_s \in \{50, 100, 200, 300, 500\}$ and the p_s choice by CV are considered.

check the effect of the p_s choice using five-fold CV. In general, the CV choice does not attain the best performance, but it performs well across all sparsity scenarios.

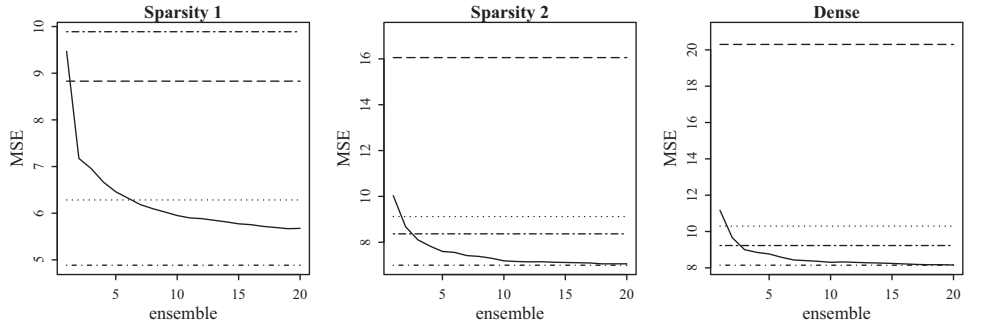
Effect of correlation of features. We next check how well ENSURE performs when the correlation among the features varies. We consider different ρ in the AR setting. To study the effect of varying ρ, we set $\rho = 0, 0.8$. Figure 4 shows that our approach's good performance does not rely on correlations among the features, making it suitable for a variety of applications.

Effect of (p, n). We next consider different settings of (p, n), specifically, $(p, n) = (1000, 20)$ and $(500, 20)$, and report the results in Figure 3. Our method works well under both settings. This results shows the robustness of our method to (p, n). By comparing $(p, n) = (1000, 20)$ and $(p, n) = (500, 20)$, we see that our approach works better when p is bigger with respect to n. This suggests that ENSURE may lose its advantage when p is not much larger than n. This makes sense, because the subset regression has a requirement on p_s with respect to n, as shown in Theorem 2.

Computational time comparison. Finally, we compare the methods in terms of their computational costs. In Table 1, we report the computational cost of ENSURE under three settings $(n, p) = (20, 10\text{K}), (50, 10\text{K}), (50, 50\text{K})$, and $(100, 100\text{K})$. To show the computational advantage of ENSURE, we also report the computation times for five other methods. The ridge regression is solved using a QR decomposition. The LASSO is implemented using the R package "glmnet" (Friedman, Hastie and Tibshirani (2010)), RF is implemented using the R package "randomForest" (Breiman (2001)), and SIS is implemented using the R package "SIS" (Fan and Lv (2008)). In all cases, we use the default parameters. The computation is performed on a computer with a 3 GHz Intel i7



(a) $(p, n) = (1000, 20)$



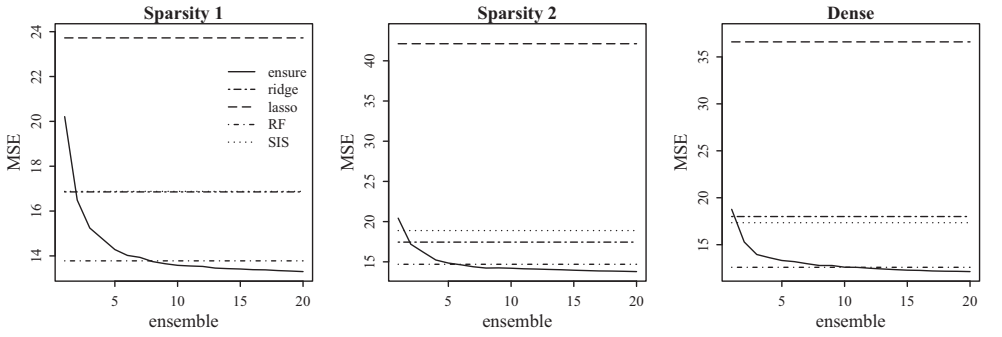
(b) $(p, n) = (500, 20)$

Figure 3. The performance of the ensemble method under various (p, n). The input $\mathbf{X}$ is from AR under $\rho = 0$

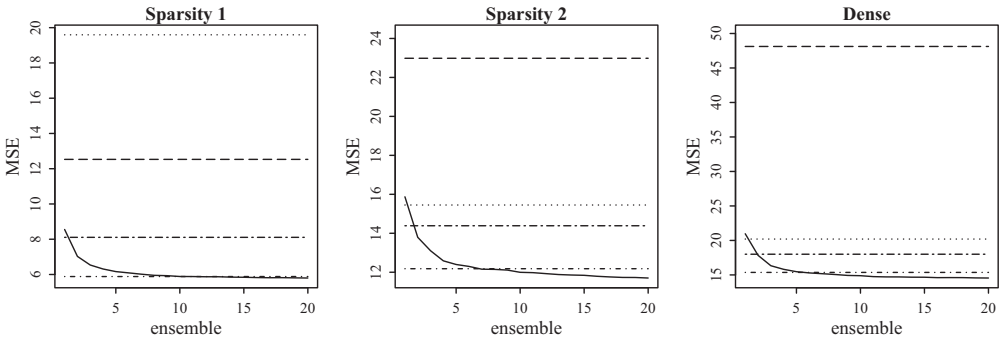
processor, 8 GB memory, and OS X operation system. From Table 1, we see that when p is relatively large compared to n, then ENSURE requires a far shorter computation time than those of the methods based on the full features. In particular, the computational burden of RF and SIS increases much faster as n or p becomes large. In contrast, ENSURE greatly reduces the computational cost, for the following reasons: (i) the subset regression and the CV procedure are calculated using a single ridge regression from the feature subset; (ii) simple random sampling costs little; and (iii) the proposed ensemble procedure is a linear function of the ensemble number.

5.2. Two real data sets

We now analyze two real data sets. The first is the gene microarray RMA (Scheetz et al. (2006)), which consists of gene expression levels of 31,041 genes obtained from 120 rats, that is, $(n, p) = (120, 31041)$. The target variable is the TRIM32 gene. The other data set is PUL (Ziyatdinov et al. (2015)), from



(a) $\rho = 0$



(b) $\rho = 0.8$

Figure 4. The performance of the ensemble method for AR under $\rho = 0$ and 0.8.

Table 1. Computational cost (seconds), including user and system time for ENSURE, ($T = 1, 5$, and 10) and four methods that use all available features: ridge, LASSO, RF, and SIS. The results are average values of 20 runs.

(n, p)	time	T (ensure)			ridge	LASSO	RF	SIS
		1	5	10				
(20, 10K)	User time	0.04	0.05	0.06	0.12	0.48	3.35	8.15
	System time	0.00	0.00	0.00	0.01	0.15	0.02	0.13
(50, 10K)	User time	0.17	0.27	0.33	0.27	0.79	15.0	19.43
	System time	0.01	0.03	0.03	0.02	0.20	0.07	0.52
(50, 50K)	User time	0.22	0.31	0.38	1.45	3.99	91.68	91.02
	System time	0.03	0.07	0.07	0.07	0.94	0.66	2.53
(100, 100K)	User time	1.48	1.86	2.40	8.04	14.87	529.4	284.5
	System time	0.34	0.35	0.39	0.42	1.41	4.07	17.78

the UCI machine learning repository, that contains 58 time series acquired from 16 chemical sensors under a gas flow modulation. The sensors were exposed to gaseous binary mixtures of acetone and ethanol at different concentrations. We

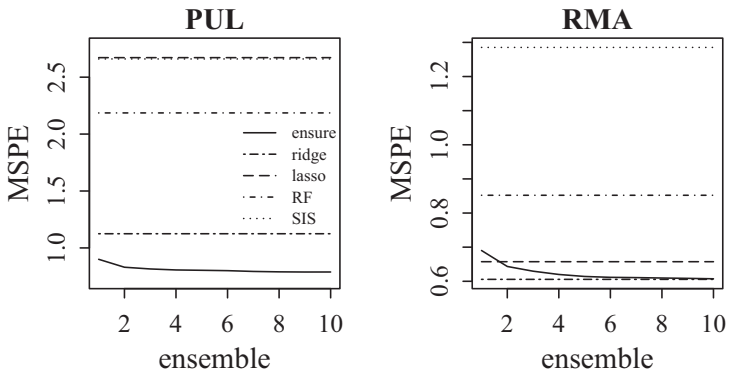


Figure 5. The performance of the ensemble method for the data sets RMA and PUL.

use the first chemical sensor, and the response variable is acetone concentration. There are 7,500 features, so $(n, p) = (58, 7500)$.

Similarly to the analysis on simulated data sets, we compare ENSURE with methods based on the full features. Because we do not know the true parameters, we calculate the prediction performance using test data. Both data sets were divided into two parts, training and testing data sets, by randomly selecting half-and-half observations. The results of 500 replicated experiments are summarized in Figure 5, where "MSPE" is the prediction error for the test data set defined as $\text{MSPE} = \text{E}\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2$, and the values are arithmetic means of 200 replicated experiments.

For RMA, the ENSURE method outperms the high-dimensional methods based on the full features. For PUL, ENSURE obviously beats the methods based on the full features. These results are consistent with our observations based on the synthetic data sets. They show that ENSURE may be a better option for high-dimensional prediction than using methods based on all available features.

6. Conclusion

The analysis of high-dimensional data has attracted a lot of attention. The prediction performance and computational cost are problematic when predictors os high dimensional which is not unusual. We have proposed a novel and easily implemented algorithm for high-dimensional prediction. We perform a subset regression based on a feature subset that is chosen by sampling without replacement uniformly. Moreover, we apply the ensemble method to the subset regression. We have provided a theoretical justification for this procedure by means of an excess risk analysis. We have conducted intensive experiments to demonstrate that the

procedure exhibits has promising performance.

There are three issues to address in future research. First, it is still unclear why ENSURE does work in many empirical studies, such as the synthetic and real data sets investigated here. Theorem 1 provides a bound without considering the estimation process of the coefficients, and Theorem 2 provides an upper bound on the excess risk using the randomness from sampling. However, the mechanism behind these remains unaddressed. Understanding this mechanism can help us know when and where the subset regression performs well for high-dimensional prediction.

Second, we construct our subset regression by l_2-regularized least squares for its theoretical convenience. Constructing alternative subset regression functions by using other high-dimensional methods, for example, l_1-regularization or other related methods, deserves additional research.

Lastly, we use the simplest sampling, *uniform* sampling, owing to the advantages of computational cost and theoretical convenience in terms of approximation accuracy. However, from an approximation accuracy viewpoint, there may be more efficient sampling methods than *uniform* sampling, because the features of the data sets may be heterogeneous. Therefore, developing an efficient data-driven sampling method is worthy of investigation.

Acknowledgments

The authors thank the two reviewers for their insightful comments and suggestions. Zhu's work was partially supported by the National Natural Science Foundation of China, grants 11871459 and 71532013, and by the Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01) and 111 Project (No.B18015).

Appendices

A. Proofs of the Upper Bound of the Subset Regression

Here we provide the excess risk bound under the general sampling framework rather than just uniform sampling; that is, we assume that the subset S is drawn by sampling from features $\{1, \ldots, p\}$ proportional to the probabilities $\{\pi_1, \ldots, \pi_p\}$. Let $\mathbf{\Phi} = \text{diag}\{\pi_k\}_{k=1}^p$, and $\mathbf{\Phi}_s$ is the partition of $\mathbf{\Phi}$ corresponding to the subset S. Note uniform sampling is a specific case, i.e., $\pi_k = 1/p$, so Theorem 2 will be proved by setting $\pi_k = 1/p$.

A.1. Two lemmas

Lemma A.1. *Let* $\mathbf{M} = \mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_n$ *and* $\mathbf{\Delta} = \mathbf{X}\mathbf{X}^\top - \mathbf{X}_s\mathbf{\Phi}_s^{-1}\mathbf{X}_s^\top$. *Under the event that*

$$\lambda_{\max}(\mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2}) \leq \kappa \tag{A.1}$$

holds, we have

$$\|\boldsymbol{d}_B\|^2 \leq (1-\kappa)^{-2}\lambda^2\|\mathbf{M}^{-1}\mathbf{X}\boldsymbol{\beta}\|_2^2,$$
$$\|\boldsymbol{d}_V\|^2 \leq (1-\kappa)^{-2}\|\mathbf{M}^{-1}\mathbf{X}\mathbf{X}^\top\boldsymbol{\epsilon}\|^2.$$

Proof. Let $\mathbf{M}_s = \mathbf{X}_s\mathbf{\Phi}_s^{-1}\mathbf{X}_s^\top + \lambda\mathbf{I}_n$. Firstly, we investigate the bias term. Since $\mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2} = \mathbf{I}_n - \mathbf{M}^{-1/2}\mathbf{M}_s\mathbf{M}^{-1/2}$,

$$\mathbf{M}^{1/2}\mathbf{M}_s^{-1}\mathbf{M}^{1/2} = (\mathbf{I}_n - \mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2})^{-1}. \tag{A.2}$$

From (A.2),

$$\begin{aligned}\|\boldsymbol{d}_B\|^2 =& \lambda^2\|\mathbf{M}_s^{-1}\mathbf{X}\boldsymbol{\beta}\|_2^2 = \lambda^2\|\mathbf{M}^{-1/2}\mathbf{M}^{1/2}\mathbf{M}_s^{-1}\mathbf{M}^{1/2}\mathbf{M}^{1/2}\mathbf{M}^{-1}\mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \leq& \lambda^2\left[\lambda_{\max}(\mathbf{M}^{1/2}\mathbf{M}_s^{-1}\mathbf{M}^{1/2})\right]^2\|\mathbf{M}^{-1}\mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \leq& \lambda^2\left[1-\lambda_{\max}(\mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2})\right]^{-2}\|\mathbf{M}^{-1}\mathbf{X}\boldsymbol{\beta}\|_2^2. \end{aligned}\tag{A.3}$$

If $\lambda_{\max}(\mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2}) \leq \kappa$, so from (A.3) we get

$$\|\boldsymbol{d}_B\|^2 \leq (1-\kappa)^{-2}\lambda^2\|\mathbf{M}^{-1}\mathbf{X}\boldsymbol{\beta}\|_2^2.$$

Secondly, we investigate the variance term. Denote $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$. From (A.2), we thus have

$$\begin{aligned}&\|\boldsymbol{d}_V\|^2 \\ &= \left\|\mathbf{M}_s^{-1}(\mathbf{M}_s - \lambda\mathbf{I})\boldsymbol{\epsilon}\right\|^2 = \left\|\mathbf{M}^{-1/2}\mathbf{M}^{1/2}\mathbf{M}_s^{-1}\mathbf{M}^{1/2}\mathbf{M}^{-1/2}(\mathbf{M}_s - \lambda\mathbf{I})\boldsymbol{\epsilon}\right\|^2 \\ &\leq \left[1-\lambda_{\max}(\mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2})\right]^{-2}\left[\lambda_{\max}(\mathbf{K}^{-1/2}(\mathbf{K}-\mathbf{\Delta})\mathbf{K}^{-1/2})\right]^2\|\mathbf{M}^{-1}\mathbf{K}\boldsymbol{\epsilon}\|^2 \\ &\leq \left[1-\lambda_{\max}(\mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2})\right]^{-2}\|\mathbf{M}^{-1}\mathbf{K}\boldsymbol{\epsilon}\|^2.\end{aligned}$$

Thus, we have that, if $\lambda_{\max}(\mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2}) \leq \kappa$,

$$\|\boldsymbol{d}_V\|^2 \leq (1-\kappa)^{-2}\|\mathbf{M}^{-1}\mathbf{K}\boldsymbol{\epsilon}\|^2.$$

Lemma A.2. *Denote $d_1 \geq d_2 \geq \cdots \geq d_n > 0$ as the eigenvalues of $p^{-1}\mathbf{X}\mathbf{X}^\top$. Let $A = \max\{(p\pi_k)^{-1}\|\boldsymbol{x}^{(k)}\|^2\}_{k=1}^p$. We have, for all $t > 0$,*

$$Pr\left(\lambda_{\max}(\mathbf{M}^{-1/2}\boldsymbol{\Delta}\mathbf{M}^{-1/2}) > t\right) \leq n \exp\left\{\frac{-p_s t^2/2}{d_n^{-1}A + t/3}\right\}. \tag{A.4}$$

Proof. Motivated by Bach (2013), we study the probability bound under the sampling with replacement, then apply the theorem in (Gross and Nesme (2010)) to get the bound under the sampling without replacement. Let $\boldsymbol{\Delta}_{\text{with}}$, which has the same expression as $\boldsymbol{\Delta}$, be obtained by sampling independently p features *with* replacement. We thus have

$$\begin{aligned}\mathbf{M}^{-1/2}\boldsymbol{\Delta}_{\text{with}}\mathbf{M}^{-1/2} &= \sum_{j=1}^{p_s}\frac{1}{p_s}\left(\sum_{k=1}^{p}\boldsymbol{x}^{(k)}\mathbf{M}^{-1}(\boldsymbol{x}^{(k)})^\top - \sum_{k=1}^{p}\frac{I_{kj}}{\pi_k}\boldsymbol{x}^{(k)}\mathbf{M}^{-1}(\boldsymbol{x}^{(k)})^\top\right)\\ &=: \sum_{j=1}^{p_s}\mathbf{M}_j,\end{aligned}$$

where I_{kj} is a random variable such that $\Pr(I_{kj} = 1) = \pi_i$ for the k-feature during the jth draw. From the process of sampling with replacement, random matrices series $\{\mathbf{M}_j\}_{j=1}^{p_s}$ are independently distributed with $\mathrm{E}(\mathbf{M}_j) = 0$ and

$$\mathrm{E}(\mathbf{M}_j^2) = \frac{1}{p_s^2}\left[\sum_{k=1}^{p}\frac{1}{\pi_k}\{\boldsymbol{x}^{(k)}\mathbf{M}^{-1}(\boldsymbol{x}^{(k)})^\top\}^2 - (\mathbf{M}^{-1/2}\mathbf{X}\mathbf{X}^\top\mathbf{M}^{-1/2})^2\right]. \tag{A.5}$$

Denote $d_1 \geq d_2 \geq \cdots \geq d_n > 0$ as the eigenvalues of $p^{-1}\mathbf{X}\mathbf{X}^\top$. Let $A = \max\{(p\pi_k)^{-1}\|\boldsymbol{x}^{(k)}\|^2\}_{k=1}^p$. From (A.5), we have that

$$\begin{aligned}\lambda_{\max}\left(\sum_{j=1}^{p_s}\mathrm{E}(\mathbf{M}_j^2)\right) &= \frac{1}{p_s}\lambda_{\max}\left(A\mathbf{M}^{-1}\mathbf{X}\mathbf{X}^\top\mathbf{M}^{-1} - (\mathbf{M}^{-1/2}\mathbf{X}\mathbf{X}^\top\mathbf{M}^{-1/2})^2\right)\\ &\leq p_s^{-1}A\lambda_{\max}\left(p\mathbf{M}^{-1}\mathbf{X}\mathbf{X}^\top\mathbf{M}^{-1}\right) = p_s^{-1}A\frac{d_n}{(d_n + p^{-1}\lambda)^2}\\ &< p_s^{-1}Ad_n^{-1}.\end{aligned} \tag{A.6}$$

On the other hand, from the matrix norm inequality

$$\lambda_{\max}(\mathbf{M}_j) \leq \frac{1}{p_s}\lambda_{\max}\left(\mathbf{M}^{-1/2}\mathbf{X}\mathbf{X}^\top\mathbf{M}^{-1/2}\right) = \frac{1}{p_s}\frac{d_1}{d_1 + p^{-1}\lambda} < p_s^{-1}. \tag{A.7}$$

From (A.6) and (A.7), we apply the matrix Bernstein inequality of Tropp (2012)

into $\mathbf{\Delta}_{\text{with}}$ to obtain its probability bound:

$$Pr(\mathbf{M}^{-1/2}\mathbf{\Delta}_{\text{with}}\mathbf{M}^{-1/2} > t) \leq n \exp\left(\frac{-p_s t^2/2}{(d_n^{-1}A + t/3)}\right). \tag{A.8}$$

By Gross and Nesme (2010), we have that for all $g \in \mathbb{R}$

$$\mathrm{E}[\mathrm{tr}\{\exp(g\mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2})\}] \leq \mathrm{E}[\mathrm{tr}\{\exp(g\mathbf{M}^{-1/2}\mathbf{\Delta}_{\text{with}}\mathbf{M}^{-1/2})\}]. \tag{A.9}$$

Combing (A.8) and (A.9) leads to the desired result.

A.2. Proof of theorem 2

From the expression of $\hat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}$ in Eqn.(3.2) of Section 3, we have that

$$\mathrm{E}\|\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2 = \mathrm{E}\|\boldsymbol{d}_B\|^2 + \mathrm{E}\|\boldsymbol{d}_V\|^2. \tag{A.10}$$

Write the event that for $0 < \kappa < 1$,

$$\mathcal{E} = \{\lambda_{\max}(\mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2}) < \kappa\}.$$

So if the event $\mathcal{E}$ holds, then from Lemma A.1

$$\begin{aligned}\mathrm{E}\|\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2 \leq& (1-\kappa)^{-2}(\lambda^2\|\mathbf{M}^{-1}\mathbf{X}\boldsymbol{\beta}\|_2^2 + \mathrm{E}\|\mathbf{M}^{-1}\mathbf{X}\mathbf{X}^\top\boldsymbol{\epsilon}\|^2) \\ =& (1-\kappa)^{-2}[\lambda^2\|(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}\boldsymbol{\mu}\|^2 + \sigma^2\|\mathbf{I}_n - \lambda(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}\|_F^2].\end{aligned} \tag{A.11}$$

Notice that

$$\|(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}\boldsymbol{\mu}\|^2 = \sum_{i=1}^{n} \frac{\mu_i^2}{(pd_i + \lambda)^2}, \tag{A.12}$$

$$\|\mathbf{I}_n - \lambda(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}\|_F^2 = \sum_{i=1}^{n} \frac{pd_i}{pd_i + \lambda}, \tag{A.13}$$

On the other hand, from Lemma A.2 if $p_s > (\ln n/\rho)(\kappa/3 + d_n^{-1}A)/(\kappa^2/2)$, then

$$\Pr(\mathcal{E}) \geq 1 - \rho. \tag{A.14}$$

Combining (A.11), (A.12), (A.14) and Theorem 1, we have that

$$\mathrm{E}\|\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2 \leq (1-\kappa)^{-2}\left[\lambda^2\sum_{i=1}^{n}\frac{\mu_i^2}{(pd_i+\lambda)^2} + \sigma^2\sum_{i=1}^{n}\frac{pd_i}{pd_i+\lambda}\right] \tag{A.15}$$

with probability at least $1 - \rho$. Therefore, Theorem 2 is proved.

B. Proof of Theorem 3

From the process of the subset-ensemble prediction, we have that

$$\|\hat{\boldsymbol{\mu}}^{\text{ensure}} - \boldsymbol{\mu}\| = \left\| \frac{1}{T} \sum_{t=1}^{T} \hat{\boldsymbol{\mu}}_{s,t} - \boldsymbol{\mu} \right\| = \left\| \frac{1}{T} \sum_{t=1}^{T} (\hat{\boldsymbol{\mu}}_{s,t} - \boldsymbol{\mu}) \right\|.$$

It follows that

$$\begin{aligned}
E\|\hat{\boldsymbol{\mu}}^{\text{ensure}} - \boldsymbol{\mu}\|^2 &= E\left\| \frac{1}{T} \sum_{t=1}^{T} (\hat{\boldsymbol{\mu}}_{s,t} - \boldsymbol{\mu}) \right\|^2 \\
&= \frac{1}{T^2} \left(\sum_{t_1 \neq t_2} E(\hat{\boldsymbol{\mu}}_{s,t_1} - \boldsymbol{\mu})^\top (\hat{\boldsymbol{\mu}}_{s,t_2} - \boldsymbol{\mu}) + \sum_{t=1}^{T} E\|\hat{\boldsymbol{\mu}}_{s,t} - \boldsymbol{\mu}\|^2 \right) \\
&= \frac{1}{T^2} \left\{ \sum_{t_1 \neq t_2} E_m \left([E(\hat{\boldsymbol{\mu}}_{s,t_1} - \boldsymbol{\mu}|\boldsymbol{y})]^\top [E(\hat{\boldsymbol{\mu}}_{s,t_2} - \boldsymbol{\mu}|\boldsymbol{y})] \right) + \sum_{t=1}^{T} E\|\hat{\boldsymbol{\mu}}_{s,t} - \boldsymbol{\mu}\|^2 \right\} \\
&= \left(1 - \frac{1}{T}\right) \left(\|\mathrm{E}(\boldsymbol{d}_B)\|^2 + \mathrm{E}_m \|\mathrm{E}(\boldsymbol{d}_V|\boldsymbol{y})\|^2 \right) + \frac{1}{T} (\mathrm{E}\|\boldsymbol{d}_B\|^2 + \mathrm{E}\|\boldsymbol{d}_V\|^2) \qquad \text{(B.1)}
\end{aligned}$$

where the 3rd equality is from that each feature subset is independently and repeatedly in our ensemble process, $\hat{\boldsymbol{\mu}}_{s,t} : t = 1, \ldots, T$ can be considered as i.i.d. random vectors.

$$\begin{aligned}
&E\|\widehat{\boldsymbol{\mu}}_s - \boldsymbol{\mu}\|^2 - E\|\widehat{\boldsymbol{\mu}}_s^{\text{ensure}} - \boldsymbol{\mu}\|^2 \\
&= \left(1 - \frac{1}{T}\right) \left\{ \mathrm{E}\|\boldsymbol{d}_B\|^2 + \mathrm{E}\|\boldsymbol{d}_V\|^2 - \|\mathrm{E}(\boldsymbol{d}_B)\|^2 - \mathrm{E}\|\mathrm{E}(\boldsymbol{d}_V|\boldsymbol{y})\|^2 \right\} \\
&= \left(1 - \frac{1}{T}\right) \left[\mathrm{E}\|\boldsymbol{d}_B\|^2 - \|\mathrm{E}(\boldsymbol{d}_B)\|^2 + \mathrm{E}_m\{\mathrm{E}(\|\boldsymbol{d}_V\|^2|\boldsymbol{y})\} - \mathrm{E}\|\mathrm{E}(\boldsymbol{d}_V|\boldsymbol{y})\|^2 \right] \\
&= \left(1 - \frac{1}{T}\right) \{\sigma_B^2 + \mathrm{E}_m(\sigma_V^2)\}.
\end{aligned}$$

Thus, Theorem 3 is proved.

References

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the 26th Annual Conference on Learning Theory* **30**, 185–209.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* **101**, 119–137.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **26**, 123–140.

Breiman, L. (1998). Arcing classifiers (with discussion). *The Annals of Statistics* **26**, 801–849.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. *Interntional Statistical Review* **60**, 291–319.

Brylla, R., Gutierrez-Osunab, R. and Quek, F. (2003). Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition* **36**, 1291–1302.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35**, 2313–2351.

Cannings, T. and Samworth, R. (2017). Random projection ensemble classification. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **79**, 959–1035.

Cook, R., Forzani, L. and Rothman, A. (2013). Predicition in abundant high-dimensional linear regression. *Electronic Journal of Statistics* **7**, 3059–3088.

Dalalyan, A., Hebiri, M. and Lederer, J. (2017). On the prediction performance of the Lasso. *Bernoulli* **23**, 552–581.

Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **70**, 849–911.

Fard, M., Grinberg, Y., Pineau, J. and Precup, D. (2012). Compressed least-squares regression on sparse space. In *Proceedings of the 26th AAAI conference on Artificial Intelligence*. AAAI Press, Toronto.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–135.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Fu, W. (1998). Penalized regression: The bridge versus the Lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

Gross, D. and Nesme, V. (2010). Note on sampling without replacing from a finite collection of matrices. *arXiv:1001.2738*.

Guhaniyogi, R. and Dunson, D. (2015). Bayesian compressed regression. *Journal of the American Statistical Association* **110**, 1500–1513.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd Edition. Springer, New York.

Ho, T. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 832–844.

Hoerl, A. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress* **58**, 54–59.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Krishnan, S., Bhattacharyya, C. and Hariharan, R. (2007). A randomized algorithm for large scale support learning. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems*.

Maillard, O. and Munos, R. (2009). Compressed least-squares regression. In *Advances in Neural Information Processing Systems* **22**.

Malo, N., Libiger, O. and Schork, N. (2008). Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *The American Journal of Human Genetics* **82**, 375–385.

Särndal, C., Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer, New York.

Scheetz, T., Kim, K., Swiderski, R., Philp, A. R., Braun, T., Knudtson, K. et al. (2006). Regulation of gene expression in the Mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences* **103**, 14429–14434.

Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics* **40**, 812–831.

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

Tropp, J. A. (2012). User-friendly tools for random matrices: An introduction. In *Advances in Neural Information Processing Systems*.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **68**, 49–67.

Ziyatdinov, A., Fonollosa, J., Fernandez, L., Gutierrez-Galvez, A., Marco, S. and Perera, A. (2015). Bioinspired early detection through gas flow modulation in chemo-sensory systems. *Sensors and Actuators B: Chemical* **206**, 538–547.

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

Rong Zhu

Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai 200433, China.

E-mail: rongzhu@fudan.edu.cn

Hua Liang

Department of Statistics, George Washington University, Washington, DC 20052, USA.

E-mail: hliang@gwu.edu

David Ruppert

Department of Statistics and Data Science, School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853, USA.

E-mail: dr24@cornell.edu

(Received May 2021; accepted February 2022)