Supplementary Materials

A Brief Description

Supplementary Materials contains additional discussion, numerical illustrations and mathematical details. It is organized in six sections. In Section B, we describe applicability of Randomized Independence Screening (RIS) as an independent screening algorithm. We apply RIS to four penalized likelihood methods, and compare their performance with other screening algorithms in light of the simulated datasets considered in Section 4.1 of the paper.

Section C is a supplement to Section 2 of the paper. Choices of appropriate marginal utility measure for TARP under the GLM setup are discussed in Section C.1. Sensitivity of RIS-RP and RIS-PCR to different choices of tuning parameters is discussed in Section C.2. The method of simple aggregation is compared with other ensemble leaning approaches in Section C.3.

Section D is a supplement to Section 3 of the paper. It discusses the implications of the conditions and assumptions used in the results on predictive accuracy of TARP, and interpretation of the theorems (Section D.1). Proofs of Lemma 1, Theorem 2 and Theorem 4 are in Section D.1.3, D.2 and D.3, respectively.

Section E contains additional simulation works. The detailed specifications of the competing methods, and formulae applied to find prediction intervals for the competing methods are in Section E.1. Performance of the competing methods for some other choices of p_n is shown in Section E.2. Detailed discussion of the competitive performance of the methods is in Section E.2.1.

Section F is a supplement to Section 4.3 of the paper. It contains results on three real datasets, the GTEx data, Eye data (in Section F.1) and the ultrahigh-dimensional GEUVADIS cis-eQTL data (Section F.3). Further, it contains predictive calibration of the methods when applied to the binary response datasets (see Section F.2).

The final section (Section G) contains a proofs of the Lemma 3 and 5 from the Appendix of the paper.

B RIS as a Screening Algorithm

Randomized independence screening (RIS) can be considered as a general screening algorithm when the data dimension is very large and the predictors are highly correlated, for e.g., in genome-wide association studies. It serves as a time-efficient alternative to SIS for regressors with multicollinearity.

Recall that calculation of the posterior mean of regression coefficients involves inversion of the matrix $(X_n'X_n + I)^{-1} = I - X_n'(I + X_nX_n')^{-1}X_n$ (by Woodbury matrix identity). While the inversion step has complexity at most order n^3, the matrix multiplications are of order n^2p_n. Screening algorithms reduce the complexity of the whole operation to $p_{\boldsymbol{\gamma}}^2 n + \min\{p_{\boldsymbol{\gamma}}^3, n^3\}$, by reducing the number of regressors from p_n to $p_{\boldsymbol{\gamma}} \ll p_n$. This reduction is in particular beneficial if $p_{\boldsymbol{\gamma}} < n$. However, due to multicollinearity and huge

dimensionality, screening algorithms limited to n marginally optimal predictors may not be appropriate. RIS provides the scope to access a larger list of predictors without making the computational cost much higher. We give a toy example to demonstrate that.

Example. Consider a normal linear model having regressors with marginal correlation coefficients $|r_j| \approx (1 - 1/(2n))^{j-1}$. As n is large the marginal correlations of the first $2n$ predictors exceed $\exp(-1) \approx 0.4$, and they should be included in the set of screened predictors. In RIS we take $q_j = |r_j|^\delta = (1-1/(2n))^{\delta(j-1)}$ as the inclusion probability of the j^{th} predictor. The number of predictors selected, $p_{\boldsymbol{\gamma}}$, is a random quantity following Poisson binomial distribution (Wang, 1993) with an expected value $\sum_{j=1}^{p_n}(1-1/(2n))^{\delta(j-1)} \approx n$ for $\delta = 2$, $\approx 2n/3$ for $\delta = 3$ and so on, for sufficiently large n.

The approach of generating multiple realizations compensates the loss due to randomization without greatly increasing the computational time. Consider for example a simple setup where $2n$ predictors have $|r_j| = 0.5$ and the rest have $r_j = 0$ (except possibly one with $|r_j| = 1$). While it is important to consider all $2n$ predictors, RIS randomly selects about $p_{\boldsymbol{\gamma}} = n/2^{\delta-1}$ predictors. Even after M repetitions, the computational time can still be less than considering $2n$ predictors if $\delta > (\log_2 M + 1)/2$.

Below we provide a brief overview of performance of RIS screening. We consider 4 methods, viz., RIS-LASSO, RIS-Ridge, RIS-SCAD and RIS-MCP, and compare these methods in 2 simulation schemes, viz., Schemes I and II, provided in Section 4.1. In Section 4.1, we provide the best result for SCAD and MCP among the results obtained using various packages including *SIS*, which applies ISIS screening. Similarly, for LASSO and ridge we provide the best result obtained from *glmnet* and *biglasso* package, where the later uses sequential strong rule (SSR) screening. The results for RIS-LASSO and RIS-Ridge are

obtained using *glmnet* package, and RIS-SCAD and RIS-MCP using *ncvreg* package. For methods based on penalization, we do not include the search for optimal λ under the aggregation step. We choose the best λ for a fixed choice of screened regressors. The results are summarized in Table A.

Observe that there is no visible improvement of SCAD and MCP of under RIS in *Scheme I*. The gain due to RIS screening becomes visible in *Scheme II*. Here all the 4 methods show much lower MSPE, higher average ECP and lower width under RIS screening. In particular, for Ridge the averages MSPEs are more than ~ 33 under SSR screening (see Table 1), which are reduced to ~ 13 under RIS. Width of 50% prediction interval (PI) also decrease for all the methods under RIS.

We compare the computational time of RIS-LASSO, RIS-Ridge and RIS-SCAD with that of LASSO, Ridge tuned by SSR screening (from *Biglasso* package), and SCAD tuned by ISIS screening (from *SIS* package). The results are summarized in Figure A. In terms of computational time, RIS based methods require marginally higher time for lower values of p_n due to the aggregation step. However, RIS becomes much more efficient than SSR for higher values of p_n.

C An Addition to Section 2 of the Paper

C.1 Choice of marginal utility measure for TARP in GLM:

There are many suggestions of suitable choices of marginal utility functions for screening in GLMs (see, e.g., Zheng and Agresti (2000), Fan et al. (2009), Fan and Song (2010), Kurosawa and Suzuki (2018)). These include the maximum marginal likelihood estimator

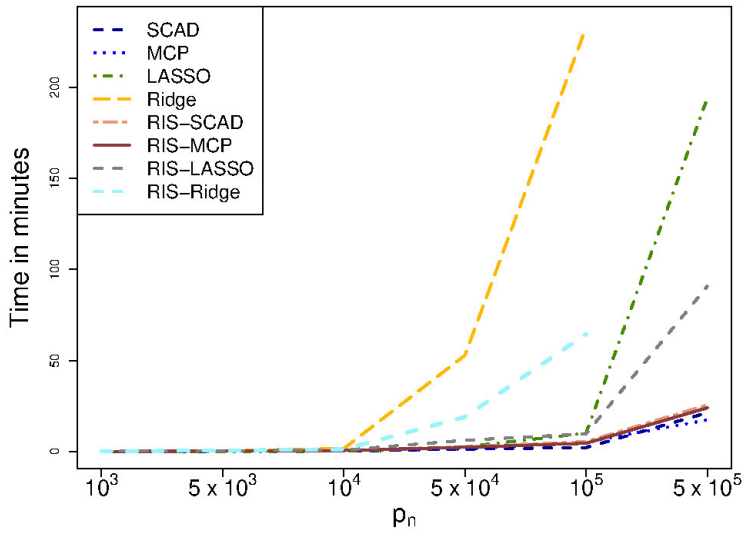


Figure A: Computational time of SCAD, MCP, LASSO and Ridge under different screening algorithms

Table A: Mean and standard deviation (sd) of mean square prediction error, empirical coverage probabilities and width of 50% prediction intervals of four RIS screened methods.

Schemes →	*I*						*II*					
(n, p_n)	$(200, 2 \times 10^3)$						$(200, 10^4)$					
Methods ↓	MSPE		ECP		Width		MSPE		ECP		Width	
RIS-SCAD	11.19	(2.17)	49.8	(6.4)	4.62	(0.40)	6.95	(1.44)	56.2	(6.0)	4.21	(0.29)
RIS-MCP	10.53	(1.83)	49.2	(5.7)	4.49	(0.33)	7.09	(1.38)	57.3	(6.4)	4.44	(0.42)
RIS-LASSO	12.79	(1.90)	49.9	(6.2)	4.82	(0.37)	7.82	(2.30)	49.5	(6.6)	3.72	(0.49)
RIS-Ridge	10.22	(1.71)	57.8	(10.3)	5.21	(1.01)	13.33	(4.48)	55.8	(6.5)	5.82	(0.39)

(MMLE), maximized marginal likelihood (MML), regression correlation coefficient (RCC), defined as the correlation between marginal estimates and observed response, etc. Theoretically, any marginal utility measure which satisfies assumptions (A1) and (A2) serves our purpose.

TARP is not sensitive to the choice of marginal utility function as long as it provides

a measure of marginal association between predictors and response. For example, for a binary response, the marginal correlation coefficient of y and a standardized predictor, x_j, is proportional to the t-test statistic for testing difference of means, and can be used as the marginal screening measure for TARP.

Usage of MMLE, MML, RCC or equivalent criteria have relatively slow computational speed due to maximization of the likelihood. Therefore, we favor the correlation coefficient, as there are no significant differences in variable selection performance compared with using the MMLE, MML or RCC (see also Saldana and Feng (2016)).

To emphasize this point we perform two simulation exercises: We consider a scheme similar to Scheme I of the paper with $p_n = 10^3$ and $n = 100$. The response variables are chosen to be binomial and Poisson, respectively, and are generated using the logit and log link. Arranging the predictors with respect to the marginal utility measure, we observe exactly the same order when using MMLE and absolute correlation. Here the MMLE is calculated via iteratively reweighted least squares (IWLS) available in the glm package of R.

Complexity for non-Gaussian Likelihood: In Step 4 (see Section 2.4), we calculate the posterior mean of the compressed regression coefficient. For non-Gaussian likelihoods, the posterior mean is not available in analytic form. We can either rely on analytic approximations like Laplace method, or use MCMC here. However, as we are dealing with m_n compressed regressors only, the computational cost is quite modest.

For example, consider the situation where the response is binary. Let y follow a Probit regression model with $P(y_i = 1) = \Phi(\mathbf{z}_i'\boldsymbol{\theta})$, where $\mathbf{z}_i$ is the i^{th} row of $Z_n = X_n R_n'$. Let y_i^* be an auxiliary random variable such that $y_i^* \sim N(\mathbf{z}_i'\boldsymbol{\theta}, 1)$ and $y_i^* > 0$ iff $y_i = 1$, $i = 1, \ldots, n$. Using Gibbs sampling, we sample from the full conditional distributions. The

full conditional of $\boldsymbol{\theta}$ given $\mathbf{y}^*$ and $\mathcal{D}^n$ is m_n-variate normal with mean $(Z_n'Z_n + I)^{-1}Z_n'\mathbf{y}^*$ and dispersion $(Z_n'Z_n + I)^{-1}$. Given $\boldsymbol{\theta}$ and $\mathbf{y}$, $\mathbf{y}^*$ is updated using a truncated normal density with mean $Z_n\boldsymbol{\theta}$ and dispersion I. The computational cost in each iteration due to matrix multiplication and inversion is at most order $O(m_n^2 n)$. In M MCMC iterations, total complexity of step 4 is $O(Mm_n^2 n)$.

C.2 Sensitivity of TARP to the tuning parameters

TARP has two tuning parameters m_n and δ, and RIS-RP has an additional parameter ψ. To show the effect of aggregation we take different values of these tuning parameters, and present the results without aggregation. Although we do not take different values of δ while aggregating, here we will show the effect of different choices of δ as well.

The results presented here correspond to *Scheme I* of the paper with $p_n = 2 \times 10^3$. We consider 100 samples and computed the mean square prediction error (MSPE), empirical coverage probability (ECP) and width of 50% prediction intervals (PI). For each of the samples, we predicted y_{new} without using the aggregation step. Four choices of each of the tuning parameters are taken. When one tuning parameter varies, the others are kept fixed at $m_n = 100$, $\psi = 0.25$ and $\delta = 2$. Table B shows the results.

From Table B the following statements can be made:

(i) m_n is the number of linear combinations (principal components) of screened regressors considered for prediction in RIS-RP (RIS-PCR). TARP tends to perform better for lower values of m_n. This behavior is theoretically supported, as the conditions for predictive accuracy include $m_n \log p_n < n\varepsilon_n^2/4$ for all sufficiently large n (see Theorems 1, 2). So smaller values of m_n imply higher rate of convergence.

(ii) ψ controls the density of zeros in the random projection matrix in RIS-RP. Variation

Table B: Mean and sd of MSPE, ECP and width of 50% PIs over 100 simulations for different choices of tuning parameters of RIS-RP and RIS-PCR.

RIS-RP											
m_n	MSPE	ECP	Width	ψ	MSPE	ECP	Width	δ	MSPE	ECP	Width
40	$16.13_{2.23}$	$37.7_{5.6}$	$3.95_{0.26}$	0.1	$17.74_{3.00}$	$29.3_{5.0}$	$3.16_{0.24}$	0.5	$23.35_{3.94}$	$34.0_{5.3}$	$4.19_{0.20}$
80	$17.22_{2.73}$	$31.3_{5.3}$	$3.37_{0.29}$	0.2	$17.53_{2.78}$	$30.5_{5.2}$	$3.22_{0.28}$	1.5	$20.05_{3.16}$	$30.2_{4.9}$	$3.47_{0.22}$
120	$19.48_{4.00}$	$27.0_{5.4}$	$3.03_{0.28}$	0.3	$18.23_{2.80}$	$29.6_{5.0}$	$3.18_{0.25}$	2.5	$15.62_{2.45}$	$31.0_{5.1}$	$3.13_{0.26}$
160	$21.27_{4.34}$	$25.4_{5.5}$	$2.93_{0.33}$	0.4	$18.19_{2.88}$	$29.4_{5.1}$	$3.21_{0.22}$	3.5	$14.23_{2.19}$	$35.8_{5.6}$	$3.45_{0.30}$

RIS-PCR							
m_n	MSPE	ECP	Width	δ	MSPE	ECP	Width
40	$13.20_{1.91}$	$30.7_{5.7}$	$2.90_{0.24}$	0.5	$12.32_{1.77}$	$23.6_{4.6}$	$2.11_{0.16}$
80	$13.85_{2.18}$	$28.1_{5.0}$	$2.70_{0.28}$	1.5	$13.36_{1.75}$	$24.3_{4.9}$	$2.28_{0.20}$
120	$15.60_{2.79}$	$26.9_{5.7}$	$2.67_{0.40}$	2.5	$15.56_{2.29}$	$30.5_{5.2}$	$3.07_{0.32}$
180	$17.67_{2.82}$	$26.7_{5.8}$	$2.85_{0.36}$	3.5	$14.23_{2.09}$	$36.7_{6.0}$	$3.49_{0.30}$

of ψ does not seem to affect RIS-RP much. In fact we could take much sparser choices of the random matrix as described in Remark 1 of the paper.

(iii) δ controls the number of screened predictors in the RIS step. While RIS-RP tends to improve performance for higher values of δ, RIS-PCR tends to deteriorate as δ increases. The methods BCR and PCR correspond to cases with $\delta = 0$ of RIS-RP and RIS-PCR, respectively. The differences in MSPEs of BCR and PCR in *Scheme I* also support this observation.

(iv) RIS-RP seems to gain more advantage due to the aggregation step than RIS-PCR, which is again intuitive, as RIS-RP relies on the random projection matrix unlike RIS-PCR.

C.3 Comparison of simple aggregation over other ensemble learning approaches

We compare three approaches of aggregation, viz., simple averaging, cross-validation and model averaging, with respect to computational complexity and performance on simulated datasets.

Complexity of RIS-RP. Recall the steps of RIS-RP in Section 2.4: (i) Screening with complexity $O(p_n)$, (ii) random matrix generation and matrix post-multiplication with total complexity $O(np_{\boldsymbol{\gamma}}m_n)$ where $p_{\boldsymbol{\gamma}}$ is number of selected regressors in the RIS step, and (iii) calculation of Bayes estimate with complexity $O(m_n^2 n)$ as $m_n < n$.

Complexity of RIS-PCR. The second step of RIS-RP is replaced by SVD of $X_{\boldsymbol{\gamma}}$ in RIS-PCR, which involves the complexity of $O\left(np_{\boldsymbol{\gamma}} \min\{n, p_{\boldsymbol{\gamma}}\}\right)$, followed by a multiplication step of complexity $O(np_{\boldsymbol{\gamma}}m_n)$.

Aggregating over different choices of tuning parameters. Note that the first step of screening is not repeated over the steps of aggregation.

Model Averaging: Suppose we consider N different choices of $\{m_n, \psi, \boldsymbol{\gamma}, R_{\boldsymbol{\gamma}}\}$. For each of the l choices we have a model $\mathcal{M}_l : y \sim f\left(y|\mathbf{x}, m_{n,l}, \psi_l, \boldsymbol{\gamma}_l, R_{\boldsymbol{\gamma}_l,l}\right)$ and a corresponding estimate of $\mathbf{y}_{new}$ given X_{new}, say $\hat{\mathbf{y}}_{new,l}$, where $l \in \{1, 2, \ldots, N\}$. The method of model averaging puts forward the expected value of $\hat{\mathbf{y}}_{new,l}$ as an estimate of $\mathbf{y}_{new}$ as

$$\hat{\mathbf{y}}_{new} = \textstyle\sum_{l=1}^{N} \hat{\mathbf{y}}_{new,l} P\left(\mathcal{M}_l|\mathcal{D}^n\right)$$

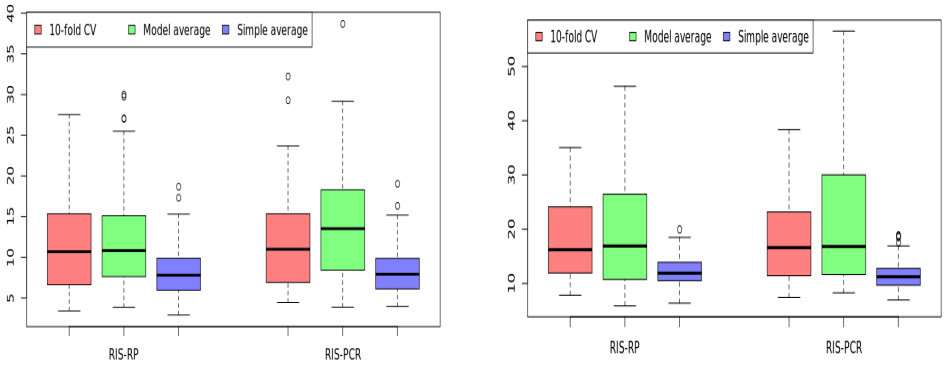
where $P\left(\mathcal{M}_l|\mathcal{D}^n\right)$ is the posterior probability of $\mathcal{M}_l$. For normal linear models, as well as for non-Gaussian GLMs, the posterior probability, $P\left(\mathcal{M}_l|\mathcal{D}^n\right)$, requires calculation of $|Z_n'Z_n + I|$ in addition to the components required to calculate $\hat{\mathbf{y}}_{new,l}$, which is of order $O(m_n^3)$. Therefore, for model averaging the complexity of step (ii) is increased by a term

of $O(m_n^3)$, and the steps (ii) and (iii) are multiplied N times each.

K-fold Cross Validation (CV): Like model averaging, for K-fold CV we consider N different choices of $\{m_n, \psi, \boldsymbol{\gamma}, R_{\boldsymbol{\gamma}}\}$. For each of these choices, we split the training dataset into K equal parts and obtain an estimate of $\mathbf{y}_{new}$, $\hat{\mathbf{y}}_{new,l}$, using $K-1$ parts. This estimate is then validated based on the remaining unused part, and a MSPE is obtained. The combined MSPE for the l^{th} model is obtained by aggregating the K MSPEs. Finally that model is considered which yields minimum MSPE. Clearly K-fold CV requires N repetitions of step (ii) and KN repetitions of step (iii), although the last step now has complexity $m_n^2(n/K + m_n)$.

Simple aggregation: This method adds the least computational complexity to the method. If we consider N different models $\mathcal{M}_l$, then the steps (ii) and (iii) are repeated N times.

Performance of different methods of aggregation in simulated datasets. We compare different methods of aggregation under *Scheme II* (see Section 4.1 of the paper). We consider $n = 100$ and 3 choices of p_n, viz., $p_n = 10^3, 5 \times 10^3$ and 10^4 to see the effect of increments of dimension. From Figures B(a)-C(b), observe that simple averaging has better and more stable performance than model averaging and cross validation. Difference in performance increases as p_n increases. Model averaging tends to be more affected by increment of dimension. Simple aggregation also requires less time to compute, and the difference is significant for RIS-PCR.



(a) $(n, p_n) = (100, 10^3)$. (b) $(n, p_n) = (100, 5 \times 10^3)$.

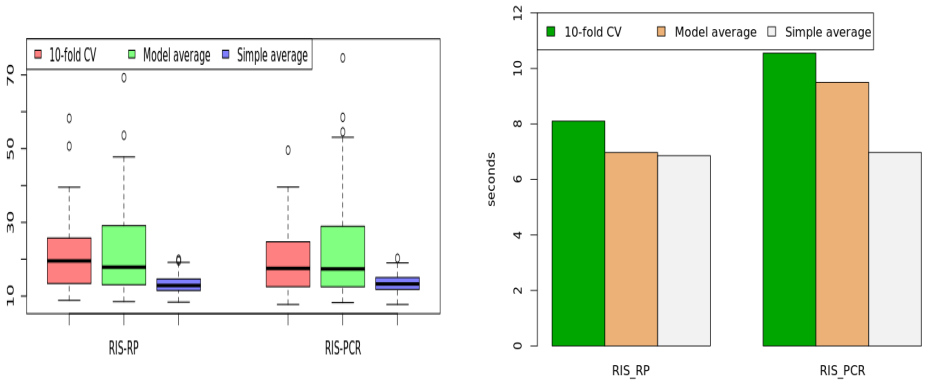
Figure B: Box-plot of MSPEs in *Scheme II*

D An Addition to Section 3 of the Paper

D.1 Implications of the assumptions and conditions required to prove the theorems

The following two remarks discuss implications of Assumptions (A3) and (A3′), and that of the conditions (i)-(iii) in Theorems 1 and 2, respectively.

Remark 1. *If the matrix $X'_{\boldsymbol{\gamma}}X_{\boldsymbol{\gamma}}$ has rank less than m_n, then $\alpha_n = 1$ by Perseval's identity. Suppose rank of $X'_{\boldsymbol{\gamma}}X_{\boldsymbol{\gamma}}$, say $r_n(\leq n)$, is bigger than m_n. Then the row space of $X_{\boldsymbol{\gamma}}$, or that of $X'_{\boldsymbol{\gamma}}X_{\boldsymbol{\gamma}}$, is spanned by a set of r_n basis vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{r_n}$. Therefore, any data point $\mathbf{x}$ can be written as a linear combination of these r_n vectors as $\mathbf{x} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_{r_n}\mathbf{v}_{r_n}$, where $a_1, \ldots, a_{r_n}$ are constants not all equal to zero. As the vectors $\mathbf{v}_j$ are orthonormal, $\mathbf{v}'_j\mathbf{x} = a_j$ for all $j = 1, \ldots, r_n$, which in turn implies that $\mathbf{x}'\mathbf{x} = \sum_{j=1}^{r_n} a_j^2$. Also, note that*



(a) $(n, p_n) = (100, 10^4)$. (b) Time in sec., $p_n = 10^4$

Figure C: Box-plot of MSPEs and multiple bar-chart of computational time in *Scheme II*.

the first m_n *among these* r_n *vectors constitute* $V_{\boldsymbol{\gamma}}'$*, which implies* $\|V_{\boldsymbol{\gamma}}'\mathbf{x}\|^2 = \sum_{j=1}^{m_n} a_j^2$*. Thus* $\|V_{\boldsymbol{\gamma}}\mathbf{x}\|^2/\|\mathbf{x}\|^2 = \sum_{j=1}^{m_n} a_j^2 / \sum_{j=1}^{r_n} a_j^2$*, and magnitude of the ratio depends on the part of* $\mathbf{x}$ *explained by the last few principal component directions. The lower bounds* $\alpha_n \sim (n\varepsilon_n^2)^{-1}$ *(in (A3)) or* $\alpha_n \sim (n\varepsilon_n^2)^{-1+b}$ *for some* $b > 0$ *(in (A3′)) are reasonable in view of many real data scenarios where most of the variation is explained by the first few principal components.*

Remark 2. *The conditions (i)-(iii) in Theorems 1 and 2 are related to the sizes of* p_n*,* m_n *and* k_n *in comparison with* $n\varepsilon_n^2$*. A sufficient condition for (i) is* $m_n \log n < n\varepsilon_n^2/4$*, providing an upper bound on the dimension of the subspace,* m_n*. Condition (ii) restricts the permissible number of regressors,* p_n*, and the number of possible models of each dimension. If there is a strict ordering in the marginal utilities* $|r_{x_j,y}|$*, so that* $k_n \leq \kappa$ *for some large number* κ*, then the condition reduces to* $\log p_n < n\varepsilon_n^2/4$*. To illustrate that the condition (iii) tends to be weak, consider distributions of* y *corresponding to Bernoulli, Poisson and*

normal. For these cases, the quantity $D(h^*)$ is at most of order $O(h^*)$. Therefore, condition (iii) does not impose much additional restriction over (i)-(ii), except $m_n \log p_n < n\varepsilon_n^2/4$, which induces a stronger upper-bound to m_n.

D.1.1 Explanation of the Assumption (A2)

We choose predictors based on marginal utility. Suppose the absolute marginal utilities, q_j, are normalized to range $[0, 1]$. Then the "inclusion probability" of a model $\boldsymbol{\gamma}$ is $q(\boldsymbol{\gamma}) = \prod_{j=1}^{p_n} q_j^{\gamma_j}(1-q_j)^{(1-\gamma_j)}$.

Assumption (A2) effectively limits inclusion of models with small $q(\boldsymbol{\gamma})$, without adding any restriction to the model size. The class of models considered in (A2), $\mathcal{A}_n = \cup_l \mathcal{M}_l$, contains the top $p_n^{k_n}$ models (ordered w.r.t. marginal inclusion probabilities) of all dimension. Assumption (A2) makes $\mathcal{A}_n$ the effective model space.

Note that all models of dimension $l < k_n$ and $l > (p_n - k_n)$ belong to $\mathcal{A}_n$. Further models considered in SIS, i.e., the models (of any dimension) containing predictors with highest marginal utility, are included in $\mathcal{A}_n$. However, as SIS based methods consider only one selected model for further analysis, sparsity is necessary condition for SIS. But sparsity is not a necessary condition for (A2). Below we illustrate this with some simple examples.

Example 1: Strong Sparsity. Consider the situation where there exists a few active regressors. Under high signal-to-noise ratio and partial orthogonality condition, Fan and Song (2010) show that the marginal regression parameter of inactive regressors will be zero, and those for the active regressors will exceed a threshold in such situations. They also show that the marginal maximum likelihood estimates (MMLE) will be close to their population counterparts almost surely. Therefore, it is likely that most of the regressors

have q_j close to zero, and a few regressors have q_j bigger than a threshold.

We consider a simple scenario with $k_n = 1$, $p_n = \exp\{n^s\}$ with $0 < s < 1$. Thus, p_n highest probability models of each dimension are included in $\mathcal{M}_l$.

Case 1: Suppose $p_n - m_n$ covariates have normalized utility $q_j \leq c_1 n^{s_1} p_n^{-(1+\nu)}$, and the remaining m_n have higher utility $q_j > (1 - c_2 n^{s_2} p_n^{-\nu})$, for some constants $c_1, c_2, s_1, s_2, \nu > 0$, and $m_n = O(n)$. In such cases the probability of the m_n-dimensional model with m_n important predictors (i.e., predictors having strong marginal association with y) is bigger than

$$\left(1 - \frac{c_2 n^{s_2}}{p_n^{\nu}}\right)^{m_n} \left(1 - \frac{c_1 n^{s_1}}{p_n^{1+\nu}}\right)^{p_n - m_n} \to 1, \quad \text{as } n \to \infty.$$

Further, the rate of convergence is approximately $1 - \exp\{-\nu n^s\}$. This probability is greater than $1 - \exp\{-n\varepsilon_n^2/4\}$ if $\nu = 1$ (see conditions (i)-(iii) of Theorems 1, 2). However, this is only one among $p_n^2 + 1$ models we consider in (A2). Therefore, we expect (A2) to hold for some smaller choices of ν in this scenario.

Case 2. Next we slightly generalize the situation, where m_n regressors have q_j at least $1 - c_1 n^{s_1} p_n^{-\nu}$, $p_n - 2m_n$ regressors have q_j at most $c_2 n^{s_2} p_n^{-(1+\nu)}$, and m_n regressors having intermediate utilities. Let $m_n = O(n^r)$ where $r < s$. In such cases let us consider the probability, $\mathcal{P}$, of all the models having m_n important regressors (i.e., regressors with $q_j > (1 - c_1 n^{s_1} p_n^{-\nu})$) included, and $p_n - 2m_n$ unimportant regressors (i.e., regressors with $q_j < c_2 n^{s_2} p_n^{-(1+\nu)}$) excluded. Note that there can be at most 2^{m_n} such models, and as $\binom{m_n}{l} < p_n$ for all $l = 1, \ldots, m_n$, all these models are included in $\mathcal{A}_n$. Here also we get

$$\mathcal{P} \geq \left(1 - \frac{c_1 n^{s_1}}{p_n^{\nu}}\right)^{m_n} \left(1 - \frac{c_2 n^{s_2}}{p_n^{1+\nu}}\right)^{p_n - 2m_n} \approx 1 - \exp\{-\nu n^s\},$$

which implies (A2) holds at least for $\nu = 1$.

Example 2: Dense cases. Now we consider some dense cases. Keeping the choices of (k_n, p_n) as before, we consider the case where $p_n/2 - m_n$ covariates have $q_j \leq c_1 n^{s_1} p_n^{-(1+\nu)}$ and $p_n/2 - m_n$ covariates have $q_j \geq (1 - c_2 n^{s_2} p_n^{-(1+\nu)})$ for some constants $s_1, s_2, c_1, c_2, \nu > 0$. The remaining $2m_n$ covariates have intermediate marginal utilities. Here we also consider $m_n = O(n^r)$ with $r < s$. As before, it is easy to see that the probability, $\mathcal{P}$, of 2^{2m_n} models including all the important covariates and excluding all the unimportant covariates is at least $1 - \exp\{-\nu n^s\}$.

Further, observe that in each of the situations in Example 1, if we replace the marginal utility, q_j, of each regressor by $1 - q_j$, then (A2) holds. For e.g., if $p_n - m_n$ covariates have $q_j \geq 1 - c_1 n^{s_1} p_n^{-(1+\nu)}$ and the remaining m_n have $q_j \leq c_2 n^{s_2} p_n^{-\nu}$, a situation opposite to *Case 1* arises. In this case if we calculate the probability, $\mathcal{P}$, of the model with $p_n - m_n$ important covariates, it can similarly be shown that $\mathcal{P} > 1 - \exp\{-\nu n^s\}$.

Discussion. It is difficult to check (A2) in situations other than extremely dense and sparse cases, as calculation of the probability $P(\{\boldsymbol{\gamma} : \boldsymbol{\gamma} \in \mathcal{A}_n\})$ is not trivial. However, (A2) holds if some of the covariates have sufficiently large marginal utility relative to the others. Situations where (A2) does not hold include the case where all the regressors have q_js uniformly spread in the interval $[0, 1]$. A convenient way of informally checking (A2) is to draw a histogram of the normalized utility measures. If it sufficiently deviates from uniform on $[0, 1]$, then (A2) holds.

Further recall that, if r_j is the marginal utility measure (e.g., correlation coefficient) of the j^{th} covariate, then we define q_j as the normalized value of $|r_j|^\delta$. As δ becomes larger, the distribution of q_js deviates more from Uniform$(0, 1)$. Choosing a suitable δ, we can control q_js as well to satisfy (A2).

D.1.2 Interpretation of the theorems

In Theorems 1 and 2 we provide a rate of closeness of the true density f_0 and the estimated density f of y under the posterior distribution. These convergence results describe "often closeness" between f and f_0 (see Jiang (2007)). These results imply existence of point estimates of f_0 that have the convergence rate ε_n in a frequentist sense. Such a point estimate can be obtained by finding the center of an ε_n-ball with high posterior probability, or by posterior expectation (see Ghosal et al. (2000)).

In Theorems 1 and 2, we have shown that the predictive density $f(y|\boldsymbol{\theta}, R_{\boldsymbol{\gamma}}, \mathbf{x}_i)$, with $\boldsymbol{\theta}$ drawn from $\pi(\boldsymbol{\theta}|\mathcal{D}^n, R_{\boldsymbol{\gamma}})$ and $\boldsymbol{\gamma} \in \mathcal{A}_n$ ($\mathcal{A}_n = \cup_l \mathcal{M}_l$, as in Assumption (A2)), concentrates around the true predictive density under f_0 in the above sense. In order to find a point estimate for prediction, we consider the posterior mean of $\boldsymbol{\theta}$, and simply average over multiple realizations of $\boldsymbol{\gamma}$. To argue that the proposed point estimate is a "good" estimate (i.e., it lies inside ε_n-Hellinger balls containing f_0), it is enough to show that each realization of $\boldsymbol{\gamma}$ considered with inclusion probability $q(\boldsymbol{\gamma})$ is in $\mathcal{A}_n$, which is evident under the assumption (A2).

D.1.3 Proof of Lemma 1

Proof of Lemma 1(a). By Serfling (1980, Theorem 1.8.E), $\|\mathbf{x}\|^2/p_n = \sum_i x_i^2/p_n \to \sum_j E\left(x_j^2\right)/p_n$ almost surely if $cov\left(x_i^2, x_j^2\right) \leq \rho^*_{|i-j|}\sqrt{var\left(x_i^2\right) var\left(x_j^2\right)}$ with $\sum_j var\left(x_j^2\right)(\log j)^2/j^2 < \infty$, and $\sum_j \rho_j^* < \infty$. Here $E(x_j^2) = 1$, $var(x_j^2) = 2$ for all j, and $cov(x_i^2, x_j^2) = 2\sigma_{i,j}^2$. Therefore by (B1), $\|\mathbf{x}\|^2/p_n \to 1$ almost surely.

Proof of Lemma 1(b) Let $Y_n = \|\mathbf{x}\|^2/p_n$. This part is proved noticing that for each $A_n = \{\omega : Y_n(\omega) \to 1\}$ implies $\left\{\omega : Y_n(\omega)/(n\varepsilon_n^2)^b \to 0\right\}$ for any $b > 0$ as $n\varepsilon_n^2 \to \infty$. □

D.2 Proof of Theorem 2

Proof. The outline of the proof of Theorem 2 closely follows the arguments given in the proof of Theorem 1. Therefore we only present those parts of the proof which are different. As in Theorem 1, we show that $P\left(\mathcal{B}_n^c|\mathcal{A}_n\right) > 1 - 2e^{-n\varepsilon_n^2/4}$ by checking the three conditions of Lemma 2.

The proof of Condition (a) is the same as for Theorem 1, except for the places involving the projection matrix $R_{\boldsymbol{\gamma}}$. Observe that given a dataset $\mathcal{D}^n$ and other tuning parameters we fix a particular projection matrix $R_{\boldsymbol{\gamma}}$. The only property of $R_{\boldsymbol{\gamma}}$ needed to prove condition (a) is $\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2 \leq m_n p_n$ for sufficiently large n. To show this we use that fact that $R_{\boldsymbol{\gamma}}$ is a matrix with orthonormal row vectors, and $R_{\boldsymbol{\gamma}}R_{\boldsymbol{\gamma}}'$ has only one eigenvalue 1 with algebraic and geometric multiplicity m_n. Therefore, 1 must be an eigenvalue of $R_{\boldsymbol{\gamma}}'R_{\boldsymbol{\gamma}}$ with algebraic multiplicity at least m_n. As the later matrix has only m_n non-zero eigenvalues, this implies that highest eigenvalue of $R_{\boldsymbol{\gamma}}'R_{\boldsymbol{\gamma}}$ is 1. Thus, $\|R_{\boldsymbol{\gamma}}\mathbf{x}\| \leq \|\mathbf{x}_{\boldsymbol{\gamma}}\| \leq \sqrt{p_n}$.

Therefore the choice of ϵ required to ensure $d(f_u, f_v) \leq \varepsilon_n$ is

$$\epsilon = \varepsilon_n^2 / \left\{\sqrt{m_n p_n}\, \sup_{|h|\leq c_n\sqrt{m_n p_n}} |a'(h)| \sup_{|h|\leq c_n\sqrt{m_n p_n}} \left(|b'(h)|/|a'(h)|\right)\right\},$$

and as before we can show that

$$N(\varepsilon_n, \mathcal{P}_n) \leq c p_n^{k_n+1} \left(\frac{1}{\varepsilon_n^2} D(c_n\sqrt{m_n p_n}) + 1\right)^{m_n},$$

where $D(R)$ is as defined in Theorem 1. By using the assumptions in Theorem 2 condition (a) follows.

The proof of Condition (b) depends only on the prior assigned on $\boldsymbol{\theta}$, and therefore remains the same under the settings of Theorem 2.

The proof of Condition (c) differs from that of Theorem 1 in showing $P(|(R_{\boldsymbol{\gamma}}\mathbf{x})'\boldsymbol{\theta} - \mathbf{x}'\boldsymbol{\beta}_0| < \Delta_n) > \exp\{-n\varepsilon^2/4\}$ for some constant Δ_n. To see this consider a positive constant

Δ_n. As before, from Lemma 4 we have

$$\begin{aligned} P(|(R_{\boldsymbol{\gamma}}\mathbf{x})'\boldsymbol{\theta} - \mathbf{x}'\boldsymbol{\beta}_0| < \Delta_n) &\geq E_{\boldsymbol{\gamma}}\left[\exp\left\{-\frac{(\mathbf{x}'\boldsymbol{\beta}_0)^2 + \Delta_n^2}{\sigma_\theta^2\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2}\right\}\frac{2^4\Delta^4}{\sigma_\theta^2\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2}\right] \\ &\geq E_{\boldsymbol{\gamma}}\left[\exp\left\{-\frac{(\mathbf{x}'\boldsymbol{\beta}_0)^2 + \Delta_n^2}{\sigma_\theta^2\alpha_n\|\mathbf{x}_{\boldsymbol{\gamma}}\|^2}\right\}\frac{2^4\Delta^4}{\sigma_\theta^2\|\mathbf{x}_{\boldsymbol{\gamma}}\|^2}\right] \\ &= \frac{2^4\Delta_n^4}{(\mathbf{x}'\boldsymbol{\beta}_0)^2 + \Delta_n^2}E_{\boldsymbol{\gamma}}\left\{\frac{Z_{\boldsymbol{\gamma}}}{p_n}\exp\left(-\frac{Z_{\boldsymbol{\gamma}}}{\alpha_n p_n}\right)\right\}, \end{aligned} \tag{D.1}$$

where $Z_{\boldsymbol{\gamma}} = \{(\mathbf{x}'\boldsymbol{\beta}_0)^2 + \Delta_n^2\} / \{\sigma_\theta^2\|\mathbf{x}_{\boldsymbol{\gamma}}\|^2/p_n\}$, and α_n is as in (A3). From part (b) of Lemma 3, and continuous mapping theorem $Z_{\boldsymbol{\gamma}} - z_n \xrightarrow{p} 0$ in $\boldsymbol{\gamma}$ where $z_n = \{(\mathbf{x}'\boldsymbol{\beta}_0)^2 + \Delta_n^2\} / (\sigma_\theta^2 c\alpha_\delta)$ $> \Delta_n^2/(\sigma_\theta^2 c\alpha_\delta)$. For some positive random variable Z and non-random positive numbers p, a and b, as before we can show that

$$E\left(\frac{Z}{p}\exp\left\{-\frac{Z}{\alpha p}\right\}\right) \geq aP\left(\frac{ap}{b} < Z < -\alpha p\log(ab)\right). \tag{D.2}$$

Replacing Z by $Z_{\boldsymbol{\gamma}}$, p by p_n, α by α_n and taking $a = \Delta_n^2\exp\{-n\varepsilon_n^2/3\}/(\sigma_\theta^2 c\alpha_\delta)$, and $b = p_n\exp\{-n\varepsilon_n^2/3\}$ we get $-\alpha_n p_n\log(ab) = -\alpha_n p_n\log\left[\Delta_n^2 p_n\exp\{-2n\varepsilon_n^2/3\}/(\sigma_\theta^2 c\alpha_\delta)\right] \sim$ $2p_n\log\left(\Delta_n^2 p_n/(\sigma_\theta^2 c\alpha_\delta)\right)/3 > p_n/2$ for sufficiently large n and $ap/b = \Delta_n^2/(\sigma_\theta^2 c\alpha_\delta)$. Therefore the expression in (D.2) is greater than

$$\frac{\Delta_n^2}{\sigma_\theta^2 c\alpha_\delta}e^{-n\varepsilon_n^2/3}P\left(\frac{\Delta_n^2}{\sigma_\theta^2 c\alpha_\delta} \leq Z_{\boldsymbol{\gamma}} \leq \frac{p_n}{2}\right).$$

Note that $(\mathbf{x}'\boldsymbol{\beta}_0)^2 < \sum_{j=1}^{p_n}|\beta_{0,j}| < K$, and the probability involved in the above expression can be shown to be bigger than some positive constant p for sufficiently large n. Using these facts along with equation (D.1), we have $P(|(R_{\boldsymbol{\gamma}}\mathbf{x})'\boldsymbol{\theta} - \mathbf{x}'\boldsymbol{\beta}_0| < \Delta_n) > \exp\{-n\varepsilon_n^2/4\}$. Choosing $\Delta_n = \varepsilon_n^2/(4M)$ condition (c) follows. □

D.3 Proof of Theorem 4

Proof. As in the proof of Theorem 3, we will only prove the condition (c) of Lemma 2. The proof of condition (c) closely follows that of Theorem 3. Here also we write $d_{t=1}(f, f_0) = E_{\mathbf{x}}\left[\{(R_{\boldsymbol{\gamma}}\mathbf{x})'\boldsymbol{\theta} - \mathbf{x}'\boldsymbol{\beta}_0\}\, g\left(u^*\right)\right]$, split it into 2 parts as in equation (8) of the paper, and argue that in order to prove (c) it is sufficient to show equation (9) of the paper.

The first part of equation (9) is essentially same as the proof of part (c) in Theorem 2. The only place require attention is the proof of claim that the following expression is no less than $\exp\{-n\varepsilon_n^2/4 + 2\log 2\}$,

$$\frac{\Delta_n^2}{\sigma_\theta^2 c\alpha_\delta} e^{-n\varepsilon_n^2/3} P\left(\frac{\Delta_n^2}{\sigma_\theta^2 c\alpha_\delta} \leq Z_{\boldsymbol{\gamma}} \leq -\alpha_n p_n \log\left(\frac{\Delta_n^2 p_n}{\sigma_\theta^2 c\alpha_\delta} e^{-2n\varepsilon_n^2/3}\right)\right). \tag{D.3}$$

where $Z_{\boldsymbol{\gamma}} - z_n \to 0$ where $z_n = \{(\mathbf{x}'\boldsymbol{\beta}_0)^2 + \Delta_n^2\} / (\sigma_\theta^2 c\alpha_\delta)$ in probability in $\boldsymbol{\gamma}$. The right hand side within the above probability is bigger than $\alpha_n p_n n\varepsilon_n^2/2 \geq p_n \left(n\varepsilon_n^2\right)^b /2$ for some $b > 0$ by assumption (A3′). Note further that $(\mathbf{x}'\boldsymbol{\beta}_0)^2 / \left\{p_n (n\varepsilon_n^2)^b\right\} < \|\mathbf{x}\|^2 \|\boldsymbol{\beta}_0\|^2 / \left\{p_n (n\varepsilon_n^2)^b\right\} \to 0$ almost surely in $\mathbf{x}$ by Lemma 1(b). Therefore $\Delta_n^2/(\sigma_\theta^2 c\alpha_\delta) < z_n < p_n \left(n\varepsilon_n^2\right)^b /2$, almost surely in $\mathbf{x}$, and the probability involved in (D.3) is bigger than some positive constant p for sufficiently large n. Using these facts and choosing Δ_n as in the proof of Theorem 2, we can show that the expression in (D.3) is bigger than $\exp\{-n\varepsilon_n^2/4 + 2\log 2\}$. This completes the first part of (9).

We prove the second part of (9) in the same manner as in Theorem 3. Consider the same set D_n, such that $\pi\left((\boldsymbol{\theta}, \boldsymbol{\gamma}) \in D_n | \mathcal{A}_n\right) \geq \exp\{-n\varepsilon_n^2/4 + 2\log 2\}$, i.e., we consider any $\boldsymbol{\gamma} \in \cup_l \mathcal{M}_l$ (see assumption (A2)) and any $\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq \sigma_\theta \sqrt{3n}\varepsilon_n/\sqrt{2}$.

Next note that, the quantity $\{(R_{\boldsymbol{\gamma}}\mathbf{x})'\boldsymbol{\theta} - \mathbf{x}'\boldsymbol{\beta}_0\} \leq \|R_{\boldsymbol{\gamma}}\mathbf{x}\| \|\boldsymbol{\theta}\| + |\mathbf{x}'\boldsymbol{\beta}_0| \leq (\|\boldsymbol{\theta}\| + K) \|\mathbf{x}\|$,

as R_γ is row-orthogonal. Therefore, we consider

$$\max\left\{\|R_\gamma \mathbf{x}\|\|\boldsymbol{\theta}\|, |\mathbf{x}'\boldsymbol{\beta}_0|, |(R_\gamma \mathbf{x})'\boldsymbol{\theta} - \mathbf{x}'\boldsymbol{\beta}_0|\right\} \leq c\sqrt{n\varepsilon_n^2}\|\mathbf{x}\|,$$

for a suitable constant $c > 0$. Further by assumption (B3), $|g(u)| \leq \exp\{c_0 u\}$ for some fixed $c_0 > 0$. Thus,

$$E_{\mathbf{x}}\left[\left\{(R_\gamma \mathbf{x})'\boldsymbol{\theta} - \mathbf{x}'\boldsymbol{\beta}_0\right\} g\left(u^*\right)\right| \|\mathbf{x}\| > \sqrt{3}p_n\right] \leq E_{\mathbf{x}}\left[\exp\left\{c\sqrt{n\varepsilon_n^2}\|\mathbf{x}\|\right\}\middle| \|\mathbf{x}\| > \sqrt{3}p_n\right],$$

for a suitable constant $c > 0$. Finally as in Theorem 3 we observe that

$$\begin{aligned}
&E_{\mathbf{x}}\left[\exp\left\{c\sqrt{n\varepsilon_n^2}\|\mathbf{x}\|\right\}\middle| \|\mathbf{x}\| > \sqrt{3}p_n\right] \\
&\quad \leq \exp\left[-\frac{3c_1 p_n^2}{c_0 l_n}\left\{\left(1 - \frac{c\, l_n\sqrt{n\varepsilon_n^2}}{\sqrt{3}p_n}\right)^2 + \frac{c\, l_n^2 n\varepsilon_n^2}{p_n^2} + \frac{c l_n}{p_n}\log\left(\frac{c_0 l_n}{c_2}\right) + \frac{c}{p_n}\log(p_n)\right\}\right] \\
&\qquad\qquad \times P\left(\|\mathbf{x}\| > \sqrt{3}p_n | \mathbf{x} \sim N(\mathbf{0}, c_0 c_2^{-1} l_n\right).
\end{aligned}$$

Noting that $\max\left\{l_n \log(l_n), l_n\sqrt{n\varepsilon_n^2 m_n}\right\} = o(p_n)$ we have

$$E_{\mathbf{x}}\left[\left\{(R_\gamma \mathbf{x})'\boldsymbol{\theta} - \mathbf{x}'\boldsymbol{\beta}_0\right\} g\left(u^*\right) | A_{p_n}^c\right] P\left(A_{p_n}^c\right) \leq \exp\{-cp_n\} \leq \varepsilon_n^2/20.$$

This proves the second part of (9), and the following the same procedure as in Theorem 3 the proof is completed. □

E An Addition to Section 4.1 of the Paper

E.1 Details of specifications of the tuning parameters of the competitors

SCAD and MCP are calculated using three packages, viz., *SIS*, *ncvreg* and *ncpen* (Kim et al., 2018). For one-step SCAD (1-SCAD) we use a R-code provided by the authors of Fan et al. (2014). LASSO, ridge and elastic net (EN) are calculated using two packages,

glmnet and *biglasso* (Zeng and Breheny, 2017). In each case, the best among all the results is provided. As we are interested in prediction problem, AIC tuning parameter selector is chosen (see Zhang et al. (2010)). For EN the parameter α is set to 0.5. The tuning parameter λ should converge to 0 at a certain rate. Fan et al. (2014) has chosen $\lambda = \sqrt{p_n/n}$ for practical purpose. We consider 200 equidistant points in the range $[0.0005, 2]$ for λ. Additionally, we have considered the default data-adaptive range of λ provided in the respective packages for all the penalization methods. The best result (calibrated in terms of average MSPE) among these two is provided.

SPCR and RPCR are performed using *PMA* and *rsvd* packages in R, respectively. To estimate PC scores, we rely on approximate SVD using *fast.svd* in the *corpcor* package. For BCR, we average over 100 different random projection matrices with varying m_n values within the range $[2 \log p_n, 3n/4]$. We use the *qr* function in R to apply QR factorization in place of Gram-Schmidt orthogonalization of the random matrix, which is computationally prohibitive for large p_n. For BASAD we use *basad* R-package with the default settings which includes choosing the initial estimate for number of active covariates by Gibbs sampling, and number of burn-in and estimation iterations as 1000 each. For SSLASSO we use *SSLASSO* R-package with $\lambda_1 = 1$ and λ_0 chosen from the interval $[1, 100]$ with grid increments of 1.

Methods used to find prediction intervals (PIs) of the competing methods For TARP and BCR, PIs are obtained from quantiles of the posterior predictive distribution of $\mathbf{y}_{new}$ given $\mathcal{D}^n, X_{new}$. For normal-linear model the $100(1-\alpha)\%$ PI of the frequentist methods can be obtained as

$$\hat{y}_{new} \pm t_{\alpha/2, n-k-1} \sqrt{\hat{\sigma}^2 \left(1 + \mathbf{x}'_{new} (X'_{[k]} X_{[k]})^{-1} \mathbf{x}_{new}\right)},$$

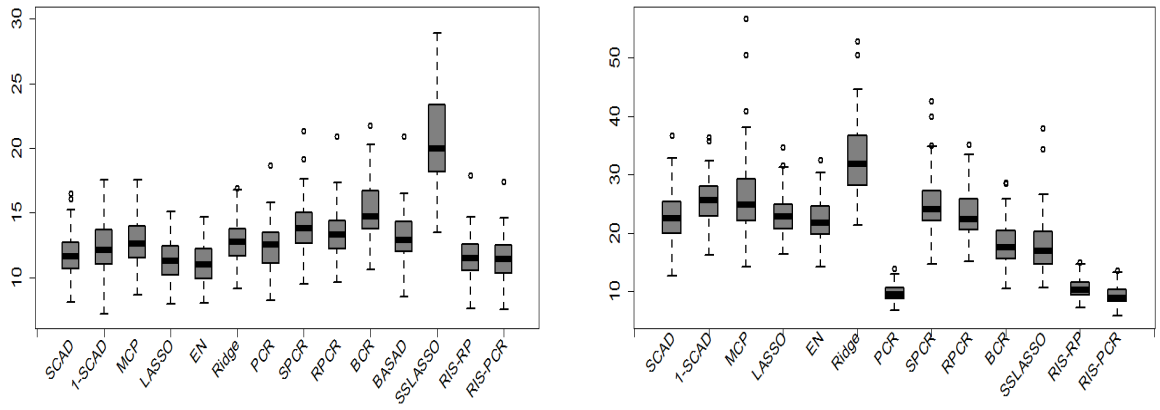
where $\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-k-1)$, k is the number of predictors used to predict y, $X_{[k]}$ is the design matrix with k selected predictors. We used this formula for all other methods except the penalized likelihood methods. The penalization methods are subject to over-fitting, and consequently the resulting $\hat{\sigma}$ can be very close to zero. Some possible solutions to this problem are provided in Lei et al. (2018), Steinberger and Leeb (2016). The former takes an approach of conformal prediction, while the later takes a simple approach of estimating PI based on leave-one-out residuals. It considers the inter-quartile range of the leave-one-out residuals as the 50% PI. We consider the later approach as calculation of the conformal prediction interval (provided in *conformalInference* R-package) takes prohibitive time for higher values of p_n.

E.2 Additional simulation results for higher values of p_n

Here we present the performance of the different methods with respect to mean square prediction error (MSPE), and empirical coverage probability (ECP) and the width of 50% prediction interval (PI) for larger choices of p_n in each scheme considered in Section 4.1 of the paper. The relative performance of the methods in terms of MSPE for different simulation schemes are shown in Figures D(a)-F(b), and that in terms of ECP and width of 50% PI are presented in Table C.

E.2.1 Summary of comparative performance of the competing methods in schemes I-IV.

All the methods except SSLASSO yield reasonable performance in terms of MSPE in *Scheme I* (see Figures 1(a) and D(a)). Further SSLASSO has the lowest coverage among all the methods in *Scheme I*. Among the other methods, the best overall performance in terms



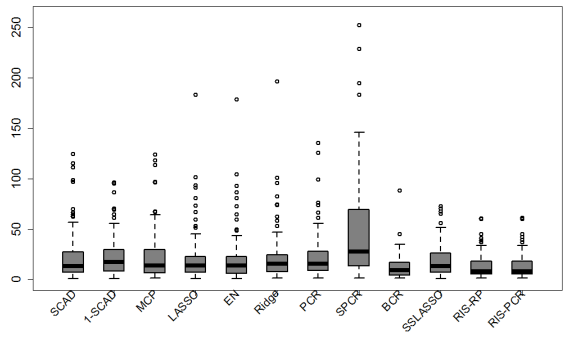
(a) *Scheme I*, $p_n = 3 \times 10^3$.

(b) *Scheme II*, $p_n = 2 \times 10^4$.

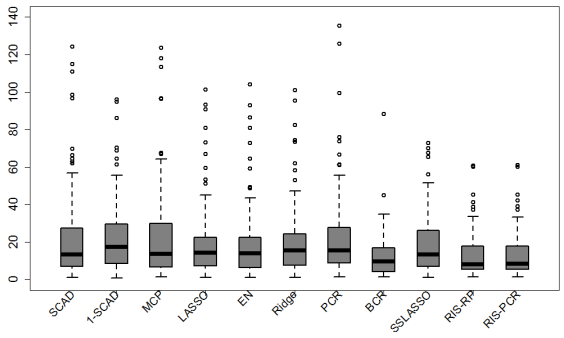
Figure D: Box-plot of MSPEs for the competing methods in Schemes I and II.

of MSPE are by EN, RIS-PCR and RIS-RP. However, TARP has lower empirical coverage than most of the methods. All the other methods have nearly comparable performance in terms of MSPE and ECP. Finally PCR has average ECP and width comparable to RIS-RP. However, it has the highest variance of both ECP and width reflecting lack of stability. 1-SCAD also have high variance of ECP and width in *Scheme I*.

In *Scheme II* the best overall performance is by PCR and RIS-PCR (see Figures 1(b) and D(b)). However, RIS-PCR has lower average ECP than most of the methods, and PCR also has low coverage and highest variance of ECP compared to all the methods. RIS-RP closely follows the former two methods in terms of MSPE, and also yield higher coverage probability. SSLASSO and BCR have somewhat better results than others in terms of MSPE. However, like in *Scheme I*, here also SSLASSO has the lowest ECP among all the competing methods. BCR has the highest coverage among all the competing methods. Average MSPE of all the penalization methods, SPCR and RPCR are on the higher side,

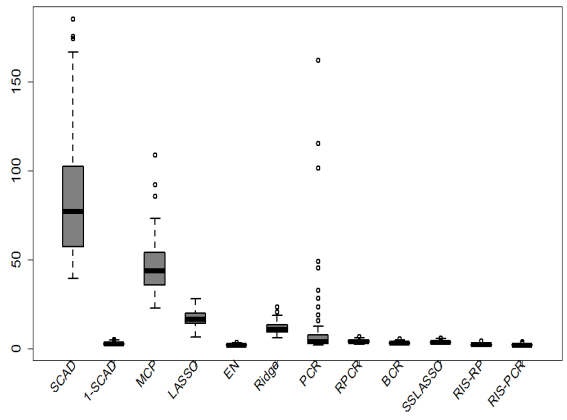


(a) MSPE of all the methods.

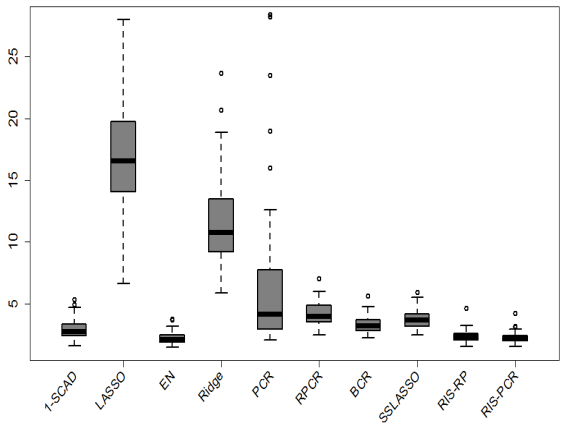


(b) MSPE of selected methods.

Figure E: Box-plot of MSPEs for $p_n = 10^4$ in *Scheme III*.



(a) MSPE of all the methods.



(b) MSPE of selected methods.

Figure F: Box-plot of MSPEs for $p_n = 2 \times 10^4$ in *Scheme IV*.

Table C: Mean and standard deviation (in bracket) of empirical coverage probabilities (ECP) and width of 50% prediction intervals for 13 competing methods.

Schemes →	*I*		*II*		*III*		*IV*	
(n, p_n)	$(200, 2 \times 10^3)$		$(200, 10^4)$		$(200, 5 \times 10^3)$		$(200, 10^4)$	
Methods ↓	ECP	Width	ECP	Width	ECP	Width	ECP	Width
SCAD	49.8 (6.0)	4.58 (0.42)	50.2 (8.3)	6.51 (0.88)	11.3 (10.1)	1.37 (0.08)	48.1 (6.8)	11.74 (2.30)
1-SCAD	49.4 (11.0)	4.77 (1.07)	46.3 (9.6)	6.29 (1.36)	10.8 (10.6)	1.42 (0.21)	49.2 (9.6)	2.23 (0.42)
MCP	51.7 (6.7)	4.92 (0.62)	49.1 (9.9)	6.84 (1.22)	11.7 (10.8)	1.39 (0.19)	50.9 (6.7)	9.2 (1.24)
LASSO	47.9 (6.9)	4.80 (0.50)	47.4 (8.5)	6.36 (1.05)	10.6 (11.0)	1.38 (0.12)	47.3 (6.7)	5.09 (0.60)
EN	48.9 (6.5)	4.47 (0.39)	48.0 (9.0)	6.31 (0.95)	10.4 (11.0)	1.38 (0.12)	58.2 (7.8)	2.40 (0.33)
Ridge	47.2 (6.2)	4.65 (0.42)	48.6 (7.3)	7.55 (0.92)	10.3 (9.9)	1.38 (0.12)	52.8 (7.5)	4.89 (0.66)
PCR	41.3 (28.1)	4.45 (3.70)	37.8 (24.5)	3.34 (2.41)	11.0 (9.5)	1.37 (0.08)	51.7 (32.1)	4.45 (3.59)
SPCR	50.4 (5.3)	5.03 (0.24)	49.5 (5.4)	6.58 (0.50)	9.4 (7.5)	1.38 (0.08)	49.8 (5.5)	49.05 (13.68)
RPCR	50.2 (5.5)	4.91 (0.22)	47.2 (5.2)	5.98 (0.41)	**	**	49.0 (5.3)	2.67 (0.28)
BCR	52.3 (5.5)	5.45 (0.26)	54.6 (6.1)	6.25 (0.41)	26.5 (13.2)	2.42 (0.69)	44.8 (5.4)	2.16 (0.14)
BASAD	50.3 (5.4)	4.95 (0.22)	*	*	*	*	*	*
SSLASSO	18.7 (3.9)	2.15 (0.12)	22.0 (4.4)	2.28 (0.13)	11.1 (9.3)	1.38 (0.09)	33.3 (4.0)	1.64 (0.10)
RIS-RP	39.3 (4.9)	3.53 (0.15)	47.0 (5.4)	4.03 (0.12)	67.6 (18.6)	8.16 (4.60)	64.5 (5.1)	2.84 (0.18)
RIS-PCR	28.4 (5.0)	2.50 (0.22)	29.4 (4.8)	2.26 (0.16)	21.9 (8.7)	2.22 (0.85)	39.6 (4.5)	1.55 (0.12)

* BASAD requires prohibitive computational time for $p_n = 10^4$, and hence removed from comparison
** RPCR produces extremely high MSPE and width of PI for *Scheme III*, and hence removed from comparison

and among these methods Ridge and MCP have the worst performance in terms of MSPE. MCP and Ridge fail to perform well in terms of MSPE in *Scheme II*. Finally, we skip BASAD as it requires prohibitive computational time for $p_n \sim 10^4$.

In *Scheme III* (see Figures 2(a)-2(b) and E(a)-E(b)), RPCR has unrealistically high MSPE, therefore we skip it from comparison. BASAD shows the worst performance in terms of MSPE among the other methods. SPCR also has larger, occasionally extremely high, MSPE values compared to other methods. Among the other methods the box-plot of MSPE is most stable for 3 methods, viz., RIS-RP, RIS-PCR and BCR. All the other methods show similar performance in terms of MSPE. All of them have some large outlying MSPE values indicating lack of robustness in the presence of outliers. In terms of ECP as well, all the methods except RIS-RP have low coverage probabilities. BCR and RIS-PCR

also have better ECP and width than others. All other methods have comparable average ECP and width, although BASAD has the highest width of PI among all.

In *Scheme IV* (see Figures 3(a)-3(b) and F(a)-F(b)), SPCR has the worst overall performance. SCAD and MCP also show poor performance in terms of MSPE compared to others. Among the other methods LASSO has the worst performance in terms of MSPE, followed by Ridge and PCR. MSPE of PCR frequently becomes extremely large indicating instability. The other methods have comparable performance in terms of MSPE, although TARP and EN have the best MSPE results. In terms of ECP, all the methods except SS-LASSO has comparable ECP. However, RIS-RP has highest average ECP, and RIS-PCR has average ECP nearly 40%. As in *Scheme I*, PCR has highest variance of ECP in *Scheme IV* as well, followed by 1-SCAD. Except SPCR, SCAD and MCP, all the other methods have reasonable average width of 50% PI.

E.3 Summary of performance of the methods with respect to computational time

When p_n is below 10^4 all the methods except BASAD require comparable computational time (see Figures 4 and G). BASAD takes more than an hour for $p_n = 10^4$ and the code crashes $p_n \geq 5 \times 10^4$. Computation of SSLASSO also takes prohibitive time when $p_n = 5 \times 10^5$, and it is much higher than all other methods (except BASAD) for $p_n \geq 5 \times 10^4$. 1-SCAD takes much longer time among the other methods. The computational time is more than 15 minutes for $p_n = 10^5$, and more than an hour for $p_n = 5 \times 10^5$ for 1-SCAD. SCAD, BCR, RIS-PCR and EN require comparable computational time up to $p_n = 10^5$. For $p_n = 5 \times 10^5$, EN requires highest time, followed by RIS-PCR. SPCR, LASSO and RIS-

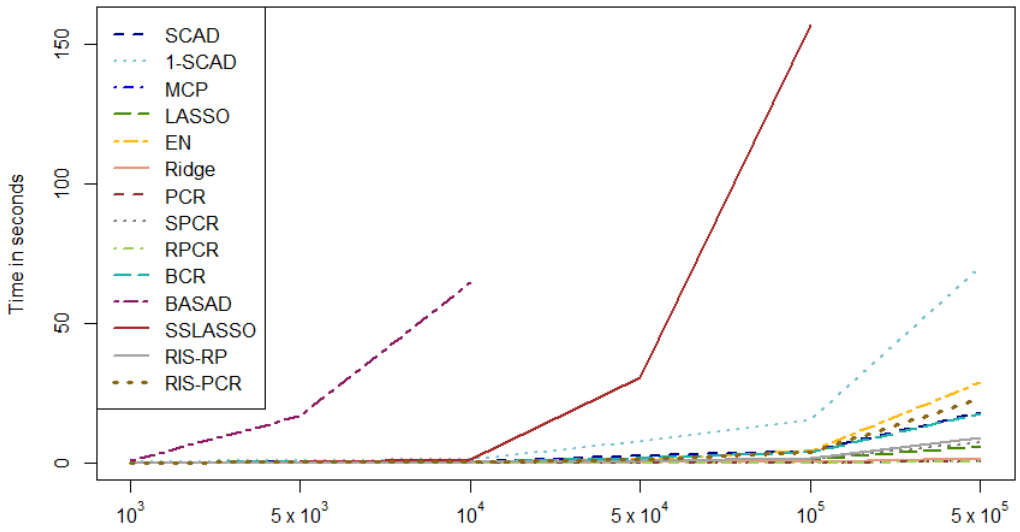


Figure G: Time required by different methods to predict y as p_n grows.

RP require comparable time throughout, RIS-RP requiring maximum time and LASSO requiring minimum time to compute among these three. The remaining 4 methods, MCP, Ridge, PCR and RPCR have best overall performance. Among these methods, Ridge takes highest time to compute, less than 1.5 minutes for 5×10^5.

The increment of computational time of RIS-PCR is due to the computation of exact SVD of the screened design matrix X_γ. However, this can be reduced if one uses some approximation of the SVD.

Table D: Mean and standard deviation (in braces) of percentage of misclassification and area under the ROC curve for *GTEx* dataset.

Methods	SCAD	1-SCAD	MCP	LASSO	EN	Ridge	PCR	SPCR	RPCR	BCR	RIS-RP	RIS-PCR
Misclassification Rate (in %)	0.00 (0.00)	34.83 (0.00)	0.00 (0.00)	58.81 (22.39)	00.42 (0.10)	0.32 (0.19)	0.07 (0.14)	3.26 (3.03)	0.16 (0.19)	13.59 (2.00)	0.42 (0.18)	0.50 (0.32)
Area under ROC curve	1.000 (.000)	0.500 (.000)	1.000 (.000)	0.570 (.168)	0.999 (.001)	0.998 (.001)	0.999 (.001)	0.966 (.033)	0.999 (.001)	0.877 (.041)	1.000 (.000)	0.996 (.002)

F An Addition to the Section 4.3 of the Paper

F.1 Analysis of GTEx and Eye datasets

GTex Dataset To understand the functional consequences of genetic variation, Consortium et al. (2015) presented an analysis of RNA sequencing data from 1641 samples across 43 tissues from 175 individuals, generated as part of the pilot phase of the Genotype-Tissue Expression (GTEx) project. We selected RNA-seq data on two normal tissues, viz., Artery-Aorta and Artery-Tibial. The dataset contains RNA-seq expressions on 36115 ($= p_n$) genes and 556 ($= n$) samples, among which 224 are from Artery-Aorta, and 332 are from Artery-Tibial. A training set of 100 samples from each of the tissue types is considered, and the remaining 446 samples are used as test set. Table D provides the average and standard deviation (sd) of percentages of misclassification, and those for the area under the ROC curve over 100 random subsets of the same size for the competing methods.

Eye Dataset The Eye dataset consists of gene expressions for 200 ($= p_n$) gene probes from the microarray experiments of mammalian-eye tissue samples of 120 ($= n$) rats (see Scheetz et al. (2006)). The response variable is the expression level of the TRIM32 gene. We consider 100 sample points as the training set, and the remaining 20 samples as the

test set. The eye dataset has continuous response, and therefore we evaluate the methods by MSPE and empirical coverage probabilities (ECP) of 50% prediction intervals (PI) as in Section 4.1. As variation in the expression levels of the TRIM32 gene is very small (the range is 1.37), we multiply the MSPEs of different methods by 10 to increase the variability. Table E provides the mean and sd of MSPEs, ECPs of 50% PIs, and widths of the PIs over 100 different training and test sets selected from the dataset, for the competing methods.

Results: For the Golub data set, both the lowest misclassification rate and the highest area under ROC curve are achieved by RIS-RP, which is closely followed by RIS-PCR. TARP based methods attain lower sd than other methods as well. PCR and Ridge also yield reasonable performance with average misclassification rate lower than 9% and area under ROC more than 0.9. RPCR, LASSO, EN, SCAD and MCP produce average misclassification rates of at least 10%, with area under the ROC about 0.9. BCR possesses high misclassification rate (about 19%), although area under ROC is more than 0.8. Finally, either the MSPE, nor the area under ROC curve, is satisfactory for SPCR and 1-SCAD.

For the GTEx dataset, perfect classification is achieved by SCAD and MCP. These methods along with RIS-RP also have the highest area under the ROC curve. PCR, RPCR, EN, Ridge, RIS-RP and RIS-PCR also yield satisfactory results, having less than 0.5% average misclassification rate and more than 99% area under the ROC curve. SPCR yield reasonable performance with an average MSPE of less than 4%. BCR attains 13.3% average misclassification rate, with the area under the ROC curve almost 0.9. Finally LASSO and 1-SCAD fail to show any discriminatory power with average MSPE more than 34%.

RPCR, RIS-PCR, RIS-RP, SPCR, LASSO and EN yield excellent performance in terms

Table E: Mean and standard deviation (in braces) of mean square prediction errors, empirical coverage probabilities and widths of 50% prediction interval of 13 competing methods in Eye dataset.

Methods	SCAD	1-SCAD	MCP	LASSO	EN	Ridge	PCR	SPCR	RPCR	BCR	BASAD	SS-LASSO	RIS-RP	RIS-PCR
MSPE	11.66 (4.07)	18.76 (17.15)	11.66 (4.06)	8.89 (3.37)	8.90 (3.12)	10.21 (2.70)	13.84 (3.94)	8.65 (3.08)	7.67 (3.30)	10.01 (4.04)	68.09 (26.42)	11.52 (7.81)	8.54 (3.09)	8.29 (2.99)
ECP	0.502 (.138)	0.459 (.132)	0.502 (.138)	0.574 (.144)	0.573 (.122)	0.537 (.115)	0.423 (.325)	0.508 (.123)	0.522 (.114)	0.564 (.117)	0.541 (.148)	0.467 (.143)	0.598 (.101)	0.507 (.107)
Width	1.208 (.057)	1.340 (0.018)	1.208 (.057)	3.05 (11.524)	1.41 (.189)	1.489 (.184)	1.884 (1.61)	1.202 (.079)	1.055 (.049)	1.249 (.056)	1.600 (.288)	1.069 (.042)	1.341 (.038)	1.056 (.036)

of MSPE in the eye data with an average MSPE of less than 1 (see Table E). All of these methods show stable performance in terms of ECP. However, LASSO has much higher width than all other methods, with exceptionally large variance. BCR, Ridge, SSLASSO, SCAD and MCP have similar overall performance. In terms of MSPE, BCR outperforms the other three methods. PCR and 1-SCAD are not quite as good in terms of either measures. 1-SCAD also show high variability in MSPE results. Finally, performance of BASAD is worst in terms of MSPE, although it yields comparable results with respect to the other measures.

F.2 Predictive calibration in binary response datasets

Apart from measuring the misclassification rates and the area under ROC curve, we validate TARP in terms of it's ability to quantify uncertainly in real datasets with binary responses. To this end, we partition the interval $[0, 1]$ into ten equal sub-intervals, viz., $[0, 0.1)$, $[0.1, 0.2)$ and so on, and classify the test data points $(\mathbf{x}, y)_{i,new}$ to the k^{th} class if predictive probability of $y_{i,new}$ falls in that class. Next, we consider the squared difference of the empirical proportion of $y_{i,new} = 1$ among the data points classified in a given

interval with the middle point of the interval, and consider the mean of these squared differences (*MSD*) of all the intervals. If a method is well calibrated, then the *MSD* would be small. The following table (Table F) shows means and standard deviations of *MSD*s of the competing methods for Golub and GTEx datasets.

Table F indicates that TARP based methods, PCR, RPCR and BCR perform relatively well compared to the others in both the datasets. Among these methods, RIS-PCR and RIS-RP have lowest MSD for Golub and GTEx data, respectively. SCAD and MCP has lower MSD in GTEx dataset, but they fail to perform well in Golub dataset. SPCR is deficient in terms of MSD for both the datasets. Finally LASSO, Ridge, EN and 1-SCAD have worst performance in terms of MSD among all the methods.

F.3 The GEUVADIS cis-eQTL dataset

We conclude this section by illustrating the TARP approach on a massive dataset. The GEUVADIS cis-eQTL dataset (Lappalainen et al., 2013) is publicly available at `http://www.ebi.ac.uk/Tools/geuvadis-das/`. This dataset consists of messenger RNA and microRNA on lymphoblastoid cell line (LCL) samples from 462 individuals provided by the 1000 Genomes Project along with roughly 38 million SNPs. E2F2 plays a key role in the control of the cell cycle. Hence, as in Chen and Dunson (2017), we choose the gene E2F2 (Ensemble ID: ENSG00000000003) as the response. A total of 8.2 million ($= p_n$) SNPs are preselected as candidate predictors on the basis of having at least 30 non-zero expressions. The total number of subjects included in the dataset is about 450 ($= n$). The genotype of each SNP is coded as 0, 1 or 2 corresponding to the number of copies of the minor allele.

TARP is applied on this dataset. We consider four different training sample sizes, viz., $n_t = 200, 250,$ 300 and 350, and test sample size 100 in each case. As p_n is huge, we

Table F: Mean and standard deviation (in braces) of mean square differences (MSD) of empirical and predictive probabilities.

Methods→ Dataset ↓	SACD	1-SCAD	MCP	LASSO	EN	Ridge	PCR	SPCR	RPCR	BCR	RIS-RP	RIS-PCR
Golub	4.454 (.000)	4.454 (.000)	4.469 (.032)	4.454 (.000)	4.454 (.000)	4.454 (.000)	2.589 (.012)	3.429 (.073)	2.587 (.017)	2.886 (.032)	2.611 (.045)	2.555 (.044)
GTEx	2.784 (.000)	5.000 (.000)	2.784 (.000)	4.652 (.000)	4.652 (.000)	4.652 (.000)	2.784 (.000)	3.216 (.102)	2.784 (.000)	2.873 (.007)	2.782 (.001)	2.783 (.001)

Table G: MSPE, ECP and width of PI (in order) obtained by RIS-RP and RIS-PCR for three values of δ and different training sample sizes (n_t).

	RIS-RP								
	$\delta = 2$			$\delta = 5$			$\delta = 8$		
n_t	MSPE	ECP	Width	MSPE	ECP	Width	MSPE	ECP	Width
200	0.800	0.39	1.059	0.872	0.42	0.983	0.855	0.34	0.928
250	0.852	0.39	1.102	0.920	0.42	1.023	0.921	0.35	1.013
300	0.860	0.36	1.126	0.855	0.44	1.075	0.866	0.36	1.069
350	0.778	0.45	1.210	0.779	0.48	1.221	0.829	0.46	1.219
	RIS-PCR								
	$\delta = 2$			$\delta = 5$			$\delta = 8$		
n_t	MSPE	ECP	Width	MSPE	ECP	Width	MSPE	ECP	Width
200	0.834	0.06	0.177	0.838	0.12	0.192	0.831	0.10	0.252
250	0.858	0.14	0.355	0.882	0.12	0.289	0.896	0.19	0.420
300	0.845	0.14	0.399	0.867	0.20	0.511	0.865	0.20	0.487
350	0.757	0.35	0.893	0.786	0.36	0.886	0.826	0.41	0.984

applied three different values of δ, namely, $2, 5$ and 8, to analyze the effect of a conservative screening. The recommended choice of δ lies within $(5, 6)$ when $p_n = 8.2 \times 10^6$ and $n \in [200, 400]$. To perform SVD for RIS-PCR, we use *fast.svd* instead of the usual *svd* to cope with the massive number of regressors. Table G provides the MSPE, the ECP of 50% PI and width of the PI, obtained by two different variants of TARP.

Results: The MSPEs of RIS-RP and RIS-PCR are comparable for all the choices on n. However, RIS-RP yields much better empirical coverage probabilities than RIS-PCR,

especially when $n \leq 300$. The three choices of δ yield comparable results in terms of all the measures in general. For RIS-RP, $\delta = 5$ results in higher ECP and for RIS-PCR higher ECP is obtained using $\delta = 8$. Moreover, the choice $\delta = 8$ makes both the procedures much faster compared to other choices of δ. When the training sample is 350, $\delta = 2, 5$ and 8 select about $290800, 12600$ and 7960 variables, respectively, on an average in the screening stage out of 8.2×10^6 variables. In view of the results in this massive dimensional dataset, it seems reasonable to use a higher value of δ for filtering out noisy regressors, and computational convenience.

G Mathematical Details

Proof of Lemma 3

Proof of part a. Consider the conditional expectation and variance of $\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2$ given $(\boldsymbol{\gamma}, \mathbf{x})$ as follows: $E\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right) = m_n\|\mathbf{x}_\gamma\|^2$

$$var\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right) = m_n\|\mathbf{x}_\gamma\|^4\left[1 + \left\{(2\psi)^{-1} - 2\right\}\sum_{j=1}^{p_{\boldsymbol{\gamma}}} x_{\gamma,j}^4/\|\mathbf{x}_\gamma\|^4\right],$$

where $\mathbf{x}_{\boldsymbol{\gamma}}$ includes the regressors j for which $\gamma_j = 1$. The details is given in the proof of Result 1 below. Next consider the conditional expectation of $\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2$ given $\mathbf{x}$ is given by

$$E_{\boldsymbol{\gamma}}E\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right) = m_n E_{\boldsymbol{\gamma}}\left(\sum_j x_j^2 I(\gamma_j = 1)\right) = c\ m_n \sum_j x_j^2 |r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta, \tag{G.1}$$

where $c > 0$ is the proportionality constant. Also the conditional variance of $\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2$ given $\mathbf{x}$ is given by $var_{\boldsymbol{\gamma}}\left\{E\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right)\right\} + E_{\boldsymbol{\gamma}}\left\{var\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right)\right\}$. Considering both the terms of

the above expression separately we get

$$\begin{aligned} var_{\boldsymbol{\gamma}}\left\{E\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right)\right\} &= var_{\boldsymbol{\gamma}}\left(m_n\sum_j x_j^2 I(\gamma_j=1)\right) \\ &= c\, m_n^2\sum_j x_j^4|r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta\left(1-c|r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta\right) \le c\, m_n^2 p_n, \end{aligned} \tag{G.2}$$

as given $\mathbf{x}$, γ_js are independent, each $|x_j|\le 1$, and $q_j = c|r_j|^\delta < 1$. Again

$$\begin{aligned} E_{\boldsymbol{\gamma}}\left\{var\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right)\right\} &= E_{\boldsymbol{\gamma}}\left[m_n\|\mathbf{x}_\gamma\|^4\left\{1+\left(\frac{1}{2\psi}-2\right)\frac{\sum_{j=1}^{p_{\boldsymbol{\gamma}}}x_{\gamma,j}^4}{\|\mathbf{x}_\gamma\|^4}\right\}\right] \\ &\le c\, m_n E_{\boldsymbol{\gamma}}\left[\|\mathbf{x}_\gamma\|^4\right] \le c\, m_n E_{\boldsymbol{\gamma}}\left[\|\mathbf{x}\|^4\right] \le c\, m_n\, p_n^2 \end{aligned} \tag{G.3}$$

for some constant c, as $\sum_{j=1}^{p_{\boldsymbol{\gamma}}} x_{\gamma,j}^4 < \|\mathbf{x}_\gamma\|^4$.

Therefore, from (G.1), (G.2) and (G.3) it can be shown that the expectation of $\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2/(m_n p_n)$ converges to $c\alpha_\delta$, and variance of the same converges to 0, as $p_n\to\infty$ and $m_n\to\infty$.

Proof of part b. Observing that $E_{\boldsymbol{\gamma}}\left(\|\mathbf{x}_{\boldsymbol{\gamma}}\|^2\right) = c\sum_j x_j^2|r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta$ and

$$var_{\boldsymbol{\gamma}}\left(\|\mathbf{x}_{\boldsymbol{\gamma}}\|^2\right) = c\sum_j x_j^4|r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta\left(1-c|r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta\right) \le p_n.$$

Therefore it can be shown that the expectation of $\|\mathbf{x}_{\boldsymbol{\gamma}}\|^2/p_n$ converges to the limit $c\alpha_\delta$, and variance of the same converges to 0. □

Proof of Lemma 5

Proof of the statement of Lemma 3 (a). Recall Result 1. Under assumption (A1) we have

$$\frac{1}{m_n p_n}E_{\boldsymbol{\gamma}}E\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|\right) \to \alpha_\delta,$$

given $\mathbf{x}$ for α_δ as in (A1). To see that the variance $var\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|\right) = o(m_n^2 p_n^2)$, observe that

$$var_{\boldsymbol{\gamma}}\left\{E\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right)\right\} = c\ m_n^2 \sum_j x_j^4 |r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta \left(1 - c|r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta\right) = o\left(m_n^2 p_n^2\right), \tag{G.4}$$

almost surely. To verify the last statement note that by Serfling (1980, Theorem 1.8.E), $\sum_i x_i^4/p_n \to \sum_j E\left(x_j^4\right)/p_n$ almost surely if $cov\left(x_i^4, x_j^4\right) \leq \rho^*_{|i-j|}\sqrt{var\left(x_i^4\right) var\left(x_j^4\right)}$ with $\sum_j \rho_j^* < \infty$, and $\sum_j var\left(x_j^4\right)(\log j)^2/j^2 < \infty$. Here $E\left(x_j^4\right) = 3$, $var\left(x_j^4\right) = 96$ for all j, and $cov\left(x_i^4, x_j^4\right) = 24\sigma_{i,j}^2\left(\sigma_{i,j}^2 + 3\right)$. Thus it is easy to see that strong law of large numbers (SLLN) holds for $\sum_j x_j^4/p_n$ by assumption (B1), and therefore (G.4) holds almost surely.

Similarly, $E_{\boldsymbol{\gamma}}\left\{var\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right)\right\} \leq c\ m_n E_{\boldsymbol{\gamma}}\left[\|\mathbf{x}_\gamma\|^4\right] \leq c\ m_n\|\mathbf{x}\|^4 \leq o\left(m_n^2 p_n^2\right)$ almost surely. To prove the last statement, we argue as before that $\|\mathbf{x}\|^2/p_n \to 1$ almost surely. Here $E(x_j^2) = 1$, $var(x_j^2) = 2$ for all j, and $cov(x_i^2, x_j^2) = 2\sigma_{i,j}^2$. Therefore by (B1) SLLN holds for $\|\mathbf{x}\|^2/p_n$, and therefore $\|\mathbf{x}\|^4/p_n^2$ is bounded almost surely. As $m_n \to \infty$ the above statement holds. Therefore the statement of Lemma 3(a) holds.

Proof of the statement of Lemma 3 b. Observe that $E_{\boldsymbol{\gamma}}\left(\|\mathbf{x}_{\boldsymbol{\gamma}}\|^2\right) = c\sum_j x_j^2 |r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta$ and

$$var_{\boldsymbol{\gamma}}\left(\|\mathbf{x}_{\boldsymbol{\gamma}}\|^2\right) = c\sum_j x_j^4 |r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta \left(1 - c|r_{\mathbf{x}_j,\mathbf{y}_n}|^\delta\right) \leq c\sum_j x_j^4.$$

Thus the expectation of $\|\mathbf{x}_{\boldsymbol{\gamma}}\|^2/p_n$ converges to the limit $c\alpha_\delta$, and variance of the same converges to 0 almost surely as $p_n \to \infty$. This completes the proof. □

Result 1. *Consider a random matrix $R_{\boldsymbol{\gamma}}$ which depends on another random vector $\boldsymbol{\gamma}$ distributed as in (2). Then the conditional distribution of $R_{\boldsymbol{\gamma}}$ satisfies the following:*

a. $E\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right) = m_n\|\mathbf{x}_\gamma\|^2$, *and*

b. $var\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right) = m_n\|\mathbf{x}_\gamma\|^4\left[1+\{(2\psi)^{-1}-2\}\sum_{j=1}^{p_\gamma} x_{\gamma,j}^4/\|\mathbf{x}_\gamma\|^4\right]$.

Proof of part a. Observe that

$$\begin{aligned}\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2 &= \left\|\left(\sum_j r_{1,j}\gamma_j x_j, \sum_j r_{2,j}\gamma_j x_j, \ldots, \sum_j r_{m_n,j}\gamma_j x_j\right)'\right\|^2 \\ &= \left(\sum_j r_{1,j}\gamma_j x_j\right)^2 + \left(\sum_j r_{2,j}\gamma_j x_j\right)^2 + \ldots + \left(\sum_j r_{m_n,j}\gamma_j x_j\right)^2. \quad \text{(G.5)}\end{aligned}$$

Now $E\left(\sum_j r_{i,j}\gamma_j x_j\right)^2 = E\left\{\sum_j r_{i,j}^2\gamma_j x_j^2 + \sum_{j\neq j'} r_{i,j}r_{i,j'}\gamma_j\gamma_{j'}x_j x_{j'}\right\} = \sum_j \gamma_j x_j^2 = \|\mathbf{x}_\gamma\|^2$, as $E(r_{i,j}^2) = 1$ and $E(r_{i,j}r_{i,j'}) = 0$ as $i = 1, 2, \ldots, m_n$, $j, j' = 1, 2, \ldots, p_n$, and $j \neq j'$.

Proof of part b. From (G.5) we have

$$\begin{aligned}var\left(\|R_{\boldsymbol{\gamma}}\mathbf{x}\|^2|\boldsymbol{\gamma}\right) &= var\left\{\sum_i\left(\sum_j r_{i,j}\gamma_j x_j\right)^2\right\} = \sum_i var\left(\sum_j r_{i,j}\gamma_j x_j\right)^2 \\ &\quad + \sum_{i\neq i'} cov\left\{\left(\sum_j r_{i,j}\gamma_j x_j\right)^2, \left(\sum_j r_{i',j}\gamma_j x_j\right)^2\right\}. \quad \text{(G.6)}\end{aligned}$$

We will consider each term of (G.6) one by one. Consider the first term. Note that

$$\begin{aligned}var\left(\sum_j r_{i,j}\gamma_j x_j\right)^2 &= var\left\{\sum_j r_{i,j}^2\gamma_j x_j^2 + \sum_{j\neq k} r_{i,j}r_{i,j'}\gamma_j\gamma_k x_j x_{j'}\right\} \\ &= var\left\{\sum_j r_{i,j}^2\gamma_j x_j^2\right\} + var\left\{\sum_{j\neq j'} r_{i,j}r_{i,j'}\gamma_j\gamma_k x_j x_{j'}\right\} \\ &\quad + cov\left\{\sum_j r_{i,j}^2\gamma_j x_j^2, \sum_{j\neq j'} r_{i,j}r_{i,j'}\gamma_j\gamma_{j'}x_j x_{j'}\right\}.\end{aligned}$$

Consider the first term in (G.6).

$$\begin{aligned} var\left\{\sum_j r_{i,j}^2\gamma_j x_j^2\right\} &= \sum_j var\left(r_{i,j}^2\gamma_j x_j^2\right) + \sum_{j\neq j'} cov\left(r_{i,j}^2\gamma_j x_j^2, r_{i,j'}^2\gamma_{j'}x_{j'}^2\right) \\ &= \sum_j \gamma_j x_j^4 var\left(r_{i,j}^2\right) + \sum_{j\neq j'}\gamma_j x_j^2\gamma_{j'}x_{j'}^2 cov\left(r_{i,j}^2, r_{i,j'}^2\right) \\ &= \sum_j \gamma_j x_j^4\left\{E\left(r_{i,j}^4\right) - E^2\left(r_{i,j}^2\right)\right\} = \left(\frac{1}{2\psi}-1\right)\sum_j \gamma_j x_j^4, \end{aligned}$$

as $E\left(r_{i,j}^4\right) = (2\psi)^{-1}$. Again,

$$\begin{aligned} var\left\{\sum_{j\neq j'} r_{i,j}r_{i,j'}\gamma_j\gamma_k x_j x_{j'}\right\} = E\left(\sum_{j\neq j'} r_{i,j}r_{i,j'}\gamma_j\gamma_k x_j x_{j'}\right)^2 &= \sum_{j\neq j'}\gamma_j\gamma_k x_j^2 x_{j'}^2 E\left(r_{i,j}^2 r_{i,j'}^2\right) \\ + \sum_{\substack{(j,j')\neq(k,k') \\ j\neq j', k\neq k'}} \gamma_j\gamma_k\gamma_{j'}\gamma_{k'}x_j^2x_{j'}^2x_k^2x_{k'}^2 E\left(r_{i,j}r_{i,j'}r_{i,k}r_{i,k'}\right) &= \sum_{j\neq j'}\gamma_j\gamma_k x_j^2 x_{j'}^2 \end{aligned}$$

as the other term will be zero. Next

$$cov\left\{\sum_j r_{i,j}^2\gamma_j x_j^2, \sum_{j\neq j'} r_{i,j}r_{i,j'}\gamma_j\gamma_{j'}x_j x_{j'}\right\} = \sum_j\sum_{k\neq k'}\gamma_j x_j^2, \gamma_k\gamma_{k'}x_k x_{k'} cov\left(r_{i,j}^2, r_{i,k}r_{i,k'}\right) = 0.$$

Therefore the first term in (G.6) is

$$\sum_i var\left(\sum_j r_{i,j}\gamma_j x_j\right)^2 = \left(\frac{1}{2\psi}-2\right)\sum_j\gamma_j x_j^4 + \left(\sum_j \gamma_j x_j^2\right)^2. \tag{G.7}$$

The last term in (G.6), $cov\left\{\left(\sum_j r_{i,j}\gamma_j x_j\right)^2, \left(\sum_j r_{i',j}\gamma_j x_j\right)^2\right\} = 0$. This is because the $\left(\sum_j r_{i,j}\gamma_j x_j\right)^2$ depends on the i^{th} row of the random matrix R for a fixed i, and $\left(\sum_j r_{i',j}\gamma_j x_j\right)^2$ depends on a fixed $i' \neq i$. Therefore these two terms are independent,

hence uncorrelated. Combining the above result and (G.7) the proof follows. □

References

Chen, Y. and Dunson, D. B. (2017). Modular Bayes screening for high-dimensional predictors. *ArXiv preprint arXiv:1703.09906*.

Consortium, G. et al. (2015). The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.*, 10(Sep):2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, 38(6):3567–3604.

Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.*, 42(3):819–849.

Ghosal, S., Ghosh, J. K., Van Der Vaart, A. W., et al. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities. *Ann. Statist.*, 35(4):1487–1511.

Kim, D., Lee, S., and Kwon, S. (2018). A unified algorithm for the non-convex penalized estimation: The ncpen package. *arXiv preprint arXiv:1811.05061*.

Kurosawa, T. and Suzuki, N. (2018). Asymptotic distribution of regression correlation coefficient for Poisson regression model. *Comm. Statist. Theory Methods*, 47(1):166–180.

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *J. Amer. Statist. Assoc.*, 113(523):1094–1111.

Saldana, D. F. and Feng, Y. (2016). Sis: An r package for sure independence screening in ultrahigh dimensional statistical models. *J. Stat. Softw.*

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natl. Acad. Sci. U.S.A.*, 103(39):14429–14434.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Steinberger, L. and Leeb, H. (2016). Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*.

Wang, Y. H. (1993). On the number of successes in independent trials. *Statist. Sinica*, 3(2):295–312.

Zeng, Y. and Breheny, P. (2017). The biglasso package: A memory-and computation-efficient solver for lasso model fitting with big data in R. *ArXiv preprint arXiv:1701.05936*.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.*, 105(489):312–323. With supplementary material available online.

Zheng, B. and Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. *Stat. Med.*, 19(13):1771–1781.