Inference for the optimal sparse prediction set

Charles Zheng and Trevor Hastie

August 27, 2016

1 Introduction

1.1 Linear prediction with fixed design, saturated normal model

Let X be a fixed $n \times p$ design matrix, and suppose $Y \sim N(\mu, \sigma^2 I)$. Given observed X and Y, we wish to predict Y^*, an unobserved, independent draw from $N(\mu, \sigma^2 I)$. Under the *linear prediction problem*, the prediction $\hat{Y}$ takes the form $\hat{Y} = X\gamma$ for some coefficient vector $\hat{\gamma}$ to be determined in a data-dependent way. Our objective is to choose $\hat{\gamma}$ in order to minimize the prediction risk,

$$R(\gamma) = \mathbf{E}||Y^* - X\hat{\gamma}||^2.$$

1.2 Classical model selection

A classical approach to linear regression is to employ *model selection* to first select a subset $S \subset \{1, .., p\}$ of the covariates; the selected *model* M_S is the set of all vectors

$$M_S = \gamma \in \mathbb{R}^p \text{ such that } \gamma_i = 0 \text{ for all } i \notin S.$$

If we define γ_S to be the best coefficient vector in M_S, i.e.

$$\gamma_S^* = \text{argmin}_{\gamma \in M_S} R(\gamma)$$

, then the nonzero entries of γ_S^* are given by β_S,

$$\beta_S = (X_S^T X_S)^{-1} X_S^T \mu$$

where X_S is the submatrix of X obtained by taking the columns in S. Unconditionally, there exists an unbiased estimator of β_S,

$$\hat{\beta}_S = (X_S^T X_S)^{-1} X_S^T Y.$$

(Though if Y is used to select S, $\hat{\beta}_S$ is no longer unbiased conditional on the selected S.) In any case, we take $\hat{\gamma}$ to be the vector with $\hat{\gamma}_i = 0$ for all $i \notin S$, and with nonzero entries given by $\hat{\beta}_S$.

Numerous methods exist for *model selection*. Classical model selection techniques combine a search method with a model selection criterion: search methods include best-subset, forward stepwise, and backwards stepwise; while the model selection criteria include AIC and BIC.

1.3 Constrained optimal model

Let $\mathcal{S}$ be some family of subsets of $\{1, \ldots, p\}$; for instance, the set of *sparse models*

$$\mathcal{S} = \{S \subset \{1, \ldots, p\} : |S| \leq k\}$$

for some integer k.

Define the oracle risk of a subset S as $\min_{\gamma \in M_S} R(\gamma)$. That is the risk we would achieve if we knew the distribution of Y, but were constrained to the model M_S. Define the constrained optimal model as the set in $\mathcal{S}$ with the least oracle risk:

$$S^* = \text{argmin}_{S \in \mathcal{S}} \min_{\gamma \in M_S} R(\gamma)$$

Some simple algebra shows that

$$\min_{\gamma \in M_S} R(\gamma) = n\sigma^2 + ||\mu||^2 - \mu^T X_S (X_S^T X_S)^{-1} X_S \mu.$$

Therefore, defining the quantity V_S ("variance explained")

$$V_S = \mu^T X_S (X_S^T X_S)^{-1} X_S \mu,$$

we see that an equivalent definition of the constrained optimal model is the set in $\mathcal{S}$ which maximizes the variance explained,

$$S^* = \text{argmax}_{S \in \mathcal{S}} V_S.$$

In applications, one can imagine several motivations for attempting to recover an optimal constrained model.

- To find a parsimonious set of predictors, S^*.
- To achieve good prediction error in cases where at least one good sparse linear model exists.
- To infer the causal parents of Y, under the assumption that the data is generated by a sparse linear model, $\mu = X_{S^\dagger}\beta_{S^\dagger}$, where $X_{S^\dagger}$ are the causal parents of Y. Then the optimal sparse prediction set S^* coincides with $S^\dagger$ as long as one selects the correct sparsity $k = |S^\dagger|$. If $k > |S^\dagger|$, then S^* is non-unique, but all optimal prediction sets are supersets of $S^\dagger$.

1.4 Inferring the constrained optimal model

Having defined S^*, we now consider the problem of inference. A *point estimate* for S^* would be a single model $S \in \mathcal{S}$, such that S is "close" to S^* according to some metric. Popular criteria include any combination of:

- Minimizing prediction *regret*, $\mathbf{E}[V_{S^*}] - V_S$.
- Controlling *familywise error*, $\Pr[|S \setminus S^*| > 1]$.
- Controlling *false discovery proportion* (FDP), $|S \setminus S^*|/|S^*|$.
- Maximizing true discoveries, $\mathbf{E}[|S \cap S^*|]$.

For instance, one may try to minimize prediction regret while simultaneously controlling false discovery rate (expected FDP). Or one could also consider prediction regret as the only criterion.

In this work, we do not consider the point estimation problem. Rather, we consider the *set estimation* problem, that is, construct a set $\mathcal{A}$ of candidate models $\mathcal{A} \subset \mathcal{S}$, such that we ensure that the type I error probability,

$$p_e = \sup_{\mu \in \mathbb{R}^p} \Pr[S^* \notin \mathcal{A}]$$

is kept low, over all possible values of the unknown parameter $\mu \in \mathbb{R}^p$. For the time being, we assume that the noise level σ^2 is known.

Classical frequentist error control requires that the procedure satisfies $p_e \leq \alpha$ for some level $\alpha \in [0, 1]$. In such a case, $\mathcal{A}$ is considered a *valid* confidence set. From a practical standpoint, a useful confidence set would

not only be valid (or approximately valid) but also be small in size. After all, taking $\mathcal{A} = \mathcal{S}$ gives a trivially valid confidence set, but this is of no use.

Alternatively, a less stringent condition is to require *consistency* in some asymptotic regime: that is, for a sequence of problems with parameters $X^{(i)}, \mu^{(i)}$ for $i = 1, \ldots$, we have

$$\lim_{i \to \infty} p_e^{(i)} \to 1$$

where $p_e^{(i)}$ is the type I error for the ith problem in the sequence. Again, after we establish a proposed procedure as *consistent*, we would further hope to show (perhaps empirically) that has additional practical properties such as produce small confidence sets and have good finite-sample performance.

[Where has this problem been studied?]

2 A first attempt

A first attempt to the problem follows from the following idea. First of all, define the estimated variance explained as

$$\hat{V}_S = Y^T X_S (X_S X_S)^{-1} X_S^T Y - \sigma^2 |S| = Y^T P_{X_S} Y - \sigma^2 r_S$$

where $P_{X_S} = X_S (X_S X_S)^{-1} X_S^T$ and r_S is the rank of X_S.

We subtract $\sigma^2 r_S$ in order to apply Mallow's C_p correction.

2.1 Claim

We claim that we can construct a level-α confidence set $\mathcal{A}$ for S^* in the following way.

1. Let $\tilde{V}$ be the maximal observed value of $\hat{V}_S$:

$$\tilde{V} = \max_{S \in \mathcal{S}} \hat{V}_S.$$

2. We find a data-dependent threshold value T, with the property that

$$\Pr[\hat{V}_{S^*} < \tilde{V} - T] \leq \alpha.$$

3. We construct $\mathcal{A}$ as

$$\mathcal{A} = \{S \in \mathcal{S} : \hat{V}_S \geq \tilde{V} - T\}.$$

2.2 Derivation

Write $Y = \mu + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I)$. Then

$$\hat{V}_S = Y^T P_{X_S} Y = V_S + 2\mu^T P_{X_S}\epsilon + \epsilon^T P_{X_S}\epsilon - \sigma^2 r_S.$$

Since $\mathbf{E}[\epsilon^T P_{X_S}\epsilon] = \sigma^2 r_S$, we get

$$\mathbf{E}[\hat{V}_S] = V_S,$$

i.e. $\hat{V}_S$ is unbiased for V_S.

Let us define

$$\Delta V_S = \hat{V}_S - V_S = 2\mu^T P_{X_S}\epsilon + \epsilon^T P_{X_S}\epsilon - \sigma^2 r_s.$$

By Cauchy-Schwarz inequality,

$$|\mu^T P_{X_S}\epsilon| \leq \|P_{X_S}\mu\|\|P_{X_S}\epsilon\| \leq \|\mu\|\|P_{X_S}\epsilon\|$$

Therefore,

$$\Delta V_S \leq \|P_{X_S}\epsilon\|(2\|\mu\| + \|P_{X_S}\epsilon\|) - \sigma^2 r_S.$$

And furthermore, letting $A = \max_{S\in\mathcal{S}} \|P_{X_S}\epsilon\|$, we have

$$\max_{S\in\mathcal{S}} \Delta V_S \leq A(2\|\mu\| + A) - \sigma^2 r_{min}.$$

where $r_{min} = \min_{S\in\mathcal{S}} r_S$.

A is random, but we can calculate its $\frac{\alpha}{3}$th upper quantile, κ, i.e. we can find κ such that

$$\Pr[A > \kappa] \leq \frac{\alpha}{3}.$$

We can compute κ given X, $\mathcal{S}$, and σ^2. More details are provided in section 3 on computation.

Meanwhile, since $||Y||^2$ is a scaled noncentral chi-squared distribution with noncentrality $\|\mu\|^2$, n degrees of freedom, and variance parameter σ^2, we can obtain a $\alpha/3$ lower confidence bound for $\|\mu\|$. Let γ denote the lower confidence bound:

$$\sup_{\mu} \Pr[\gamma(Y) \geq \|\mu\|^2] \leq \frac{\alpha}{3}.$$

Therefore,

$$\max_{S\in\mathcal{S}} \Delta V_S \leq \kappa(2\gamma + \kappa).$$

Meanwhile, we can find a $\alpha/3$ lower confidence bound for V_{S^*}, conditional on $\|\mu\|^2 \leq \gamma$ and using the fact that $\hat{V}_{S^*} + \sigma^2 r_{S^*}$ has a noncentral chi-squared distribution with noncentrality $\mu^T P_{X_{S^*}} \mu$, n degrees of freedom and variance parameter σ^2. Therefore, we construct D such that

$$\Pr[\Delta_{V_{S^*}} + \sigma^2 r_{S^*} < D | \|\mu^2\| < \gamma] \leq \frac{\alpha}{3}.$$

Thus, defining

$$T = \kappa(2\gamma + \kappa) + (r_{max} - r_{min})\sigma^2 - D$$

where $r_{max} = \max_{S \in \mathcal{S}} r_S$, we have the needed property

$$\Pr[\hat{V}_{S^*} < \max_{S \in \mathcal{S}} V_S - T] \leq \alpha.$$