Ranking and Selection with High-dimensional Covariates and General Dependence

Xiaocheng Li[†], Xiaowei Zhang[*], Zeyu Zheng[‡]

[†] Stanford University, MS&E
[*] City University of Hong Kong, MS
[‡] UC Berkeley, IEOR

Table of Contents

1. Introduction
2. High-dimensional Covariates
3. General Dependence
4. Conclusions

Introduction

Ranking and selection

k alternatives

Output of the i-th alternative: $Y_i \sim \mu_i$

$$\mathbb{E}[Y_i] = \mu_i, \quad \mathrm{Var}[Y_i] = \sigma_i^2$$

both μ_i and σ_i are unknown

Design a **procedure** to collect n_i samples of the i-th alternative and then select a alternative $\hat{i}$ as an estimate of $i^* = \arg\max_i \mu_i$

Frequentist approach

Specify where to take samples and how many samples to achieve certain statistical guarantee

Rinott (1978)

Kim and Nelson (2001)

Chick and Wu (2005)

Frazier (2014)

Fan et al. (2016)

etc.

Bayesian approach

Given a finite simulation budget, specify where to take samples sequentially to make the most of them

Chen et al. (2000): *optimal computing budget allocation*

Chick and Inoue (2001): *expected value of information*

Frazier et al. (2008): *knowledge gradient*

Chick and Frazier (2012): *Economics of selection procedures*

etc.

Covariates

Also known as contextual information or side information

They allow decisions to be made at individual level

- With covariates, we solve

$$\max_{i=1,\ldots,k} \ \mu_i(X) = \mathbb{E}[Y_i(X)|X]$$

- Without covariates, we solve

$$\max_{i=1,\ldots,k} \ \mu_i = \mathbb{E}[\mu_i(X)] = \mathbb{E}[Y_i(X)]$$

Ranking and selection with covariates

Shen, Hong, and Zhang (2017)

For the *i*-th alternative,

$$Y_i(X) = \mu_i(X) + \epsilon_i(X),$$

where $\epsilon_i(X) \sim N(0, \sigma_i^2(X))$

- homoscedasticity: $\sigma_i(X) \equiv \sigma_i$
- heteroscedasticity: $\sigma_i(X)$ depends on X

The optimal choice

$$i^*(x) := \arg\max_{1 \leq i \leq k} \{\mu_i(x)\}$$

Correct selection

Correct selection

$$\mathsf{CS}(x) := \left\{\mu_{i^*(x)}(X) - \mu_{\widehat{i^*}(x)}(X) < \delta \mid X = x\right\}$$

Conditional Probability of Correct Selection (PCS)

$$\mathsf{PCS}(x) := \mathbb{P}\left(\mu_{i^*(x)}(X) - \mu_{\widehat{i^*}(x)}(X) < \delta \mid X = x\right)$$

Unconditional PCS

$$\mathsf{PCS}_{\mathsf{E}} := \mathbb{E}[\mathsf{PCS}(X)]$$

where the expectation is taken with respect to the distribution of X

Value added by covariates

Suppose

- $X \sim N(0, \Sigma)$
- $k = 2$
- $\mu_i(X) = \mu_i + \theta_i^\intercal X \overset{D}{=} N(\mu_i, \theta_i^T \Sigma \theta_i), \quad i = 1, 2$

If $\mu_1 > \mu_2$, the conventional R&S discards X and selects alternative 1

The (unconditional) probability of incorrect selection is

$$\mathbb{P}(\mu_1(X) < \mu_2(X) - \delta) = \mathbb{P}\left(Z < \frac{\mu_2 - \mu_1 - \delta}{\theta_1^\intercal \Sigma \theta_1 + \theta_2^\intercal \Sigma \theta_2}\right),$$

which becomes large if

- μ_2 is close to μ_1
- $\|\theta_1\|$ or $\|\theta_2\|$ is large, e.g., when dimensionality is high

Key assumptions in Shen, Hong, and Zhang (2017)

Linear dependence between Y and X

Linear coefficients are estimated via least squares

- ordinary least squares (OLS) for the homoscedastic case
- generalized least squares (GLS) for the heteroschedastic case

Fixed design: design points $x_1, \ldots, x_m$ are given and fixed

- repeated samples at a given design point

Then, calculate the number of samples at each design point that is necessary to achieve a prescribed $\mathrm{PCS}_{\mathrm{E}}$

- total number of samples $O(kd/\delta^2)$

Our setup

1. Linear dependence with high-dimensional covariates
    - OLS fails unless $n \gg d$
    - a different paradigm of “large-scale” R&S than large k

2. General dependence between Y and X
    - linear dependence may fail even for low-dimensional covariates

High-dimensional Covariates

Assumptions

1. Linear dependence with homoscedastic errors

$$Y_i(X) = X^\intercal \beta_i + \epsilon_i,$$

   where $\epsilon_i \sim N(0, \sigma_i^2)$

2. Sparsity: the number of non-zeros in β_i is bounded by a known constant s_0, i.e. $\|\beta_i\|_0 \leq s_0$
   - s_0 is the number of "significant" covariates
3. Design space is bounded by an L_1 ball with radius B

LASSO: linear regression with L_1 penalty

Let $\mathcal{X} \in \mathbb{R}^{n \times d}$ be the design matrix

$$\widehat{\beta}_{\mathrm{OLS}} = \arg\min_{\beta} \ \frac{1}{n}\|Y - \mathcal{X}\beta\|_2^2$$

$\widehat{\beta}_{\mathrm{OLS}} = (\mathcal{X}^\intercal \mathcal{X})^{-1}\mathcal{X}^\intercal Y$ works well only if $n \gg d$

LASSO is a classic variable selection method

$$\widehat{\beta} = \arg\min_{\beta} \ \frac{1}{n}\|Y - \mathcal{X}\beta\|_2^2 + \lambda\|\beta\|_1$$

The regularization parameter λ is crucial for bias-variance trade-off

- often determined via cross-validation
- its order relative to n determines convergence rate of $\widehat{\beta}$

Convergence rate of LASSO (Bühlmann and van de Geer 2011)

Fix $t > 0$. Let $\widehat{\sigma}$ be an estimator of σ. Let

$$\lambda := 4\widehat{\sigma}\sqrt{\frac{t^2 + 2\log d}{n}}$$

Then,

$$\mathbb{P}\left(\|\hat{\beta} - \beta^0\|_1 \leq \frac{4\lambda^2 s_0}{\phi_0^2}\right) \geq 1 - \alpha,$$

where ϕ_0 is the *restricted eigenvalue* of $\frac{1}{n}(\mathcal{X}^\intercal\mathcal{X})$ and

$$\alpha = 2\exp\left(-\frac{t^2}{2}\right) + \mathbb{P}\left(\frac{\widehat{\sigma}^2}{\sigma^2} \leq 1\right).$$

- Set $\widehat{\sigma} = c \times$ sample s.d., with $c \geq 1$ to be determined
- $\frac{(n-1)\widehat{\sigma}^2}{c\sigma^2} \sim \chi^2_{n-1}$

Two-stage procedure

Setup. Set $t = \sqrt{\frac{1}{2} \log \frac{6k}{\alpha}}$

First stage.

- Generate a *random* Gaussian design matrix of size $n_0 \times d$
- Simulate $Y_i(X)$ for X being each row of the design matrix and each i
- Construct estimator $\widehat{\sigma}$ by choosing c so that $\mathbb{P}\left(\frac{\widehat{\sigma_i^2}}{\sigma_i^2} \leq 1\right) \geq 1 - \frac{\alpha}{6k}$

Second stage.

- Set $n_i = \max\{n_0, 128 B s_0 \widehat{\sigma}_i^2 (t^2 + 2 \log d)/\delta\}$
- Generate a random design matrix of size $(n_i - n_0) \times d$ and simulate responses
- Compute the LASSO estimator $\widehat{\beta}_i$ with $\lambda_i = 4\widehat{\sigma}_i \sqrt{\frac{t^2 + 2 \log d}{n}}$

Selection. $\widehat{i^*}(x) = \arg\max_i (x^\intercal \widehat{\beta}_i)$

Asymptotic analysis

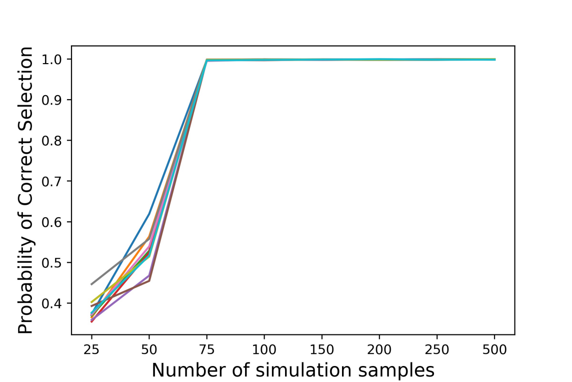
The two-stage procedure guarantees $\mathsf{PCS}_{\mathsf{E}} \geq 1 - \alpha$

The total number of samples is $O(k \log(d)/\delta^2)$ as opposed to $O(kd/\delta^2)$

Numerical experiments

$k = 3$, $d = 1000$

Sparsity $s_0 = 10$



OLS-based procedure in Shen et al. (2017) would require at least 1000 samples

General Dependence

Empirical risk minimization

For the *i*-th alternative,

$$Y_i(X) = \mu_i(X) + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma_i^2)$

Estimate $\mu_i(\cdot)$ via *empirical risk minimization*:

$$\widehat{\mu}_i = \arg\min_{f \in \mathcal{F}} \sum_{l=1}^{n} L(Y_{il}, f(X_{il}))$$

where $L(\cdot, \cdot)$ is the loss function and $\mathcal{F}$ is a class of candidate functions

- $\mathcal{F}$ can be collection of linear functions, decision trees, or neural networks

Assumptions

1. The loss function $L(\cdot, \cdot)$ takes value in $[B_l, B_u]$ on the support of response-covariate pair (Y, X).
2. We can generate X from its distribution $\mathbb{P}_X$

Results from statistical learning

Rademacher Complexity

$$R_n(\mathcal{C}) := E\left[\sup_{f\in\mathcal{C}} \frac{1}{n}\sum_{j=1}^{n} U_j f(Z_j)\right]$$

where $Z_1, ..., Z_n$ are drawn i.i.d from p^* and $U_1, ..., U_n$ are i.i.d. uniform distribution over $\{-1, 1\}$.

Proposition
Define $\mathcal{L} = \{(x, y) \to (y - f(x))^2 : f \in \mathcal{F}\}$ as the loss class. Then,

$$\mathbb{E}_{\mathbb{P}_X}[(\widehat{\mu}_n(X) - \mu(X))^2] \leq 4R_n(\mathcal{L}) + \sqrt{\frac{2\log(2/\eta)}{n}} \cdot (B_u - B_l)$$

with probability $1 - \eta$.

Single-stage Procedure

Setup.

- Set $\eta = \frac{\alpha\delta^2}{4k}$ and $n_0 = \frac{8k^2 \log(2/\eta)(B_u - B_l)^2}{\delta^2}$
- Choose n_0' such that $R_n(\mathcal{L}) < \frac{\alpha\delta^2}{8k}$ for all $n > n_0'$ and set $N = \max\{n_0, n_0'\}$

Sampling.

- Generate N samples X_{il} from $\mathbb{P}_X$ for each alternative $i = 1, ..., k$, $l = 1, ..., N$ and simulate the responses
- Estimate $\widehat{f}_i$ as the empirical risk minimizer:

$$\widehat{\mu}_i = \arg\min_{f \in \mathcal{F}} \sum_{l=1}^{N} (Y_{il} - f(X_{il}))^2$$

Selection. $\widehat{i^*}(x) = \arg\max_i \widehat{\mu}_i(x)$

Statistical guarantee

The single-stage procedure achieves $\mathsf{PCS}_{\mathsf{E}} \geq 1 - \alpha$

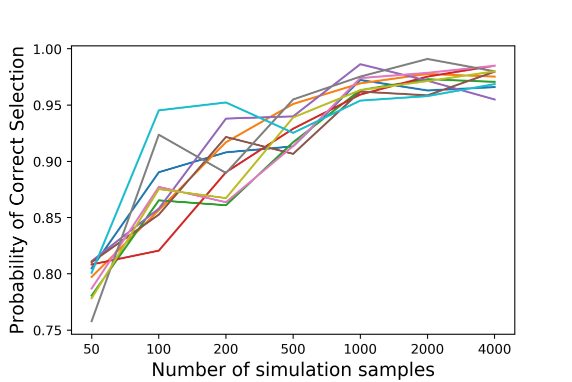
No need for two-stage sampling because the error bound based on Rademacher complexity does not depend on the variance σ^2

Tend to be conservative, nonetheless

Numerical experiments

$k = 3$

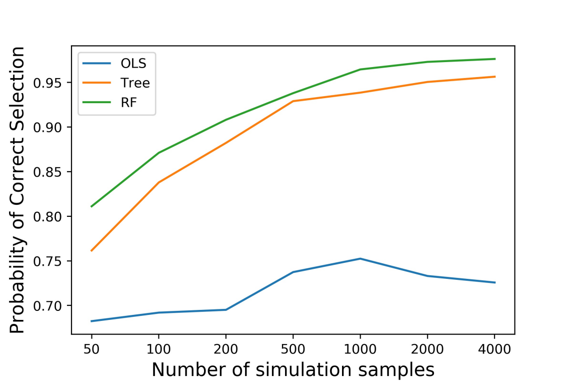
True functions are random forests



Much more samples are needed than the linear case because of the larger function space

Model misspecification

The function space $\mathcal{F}$ is misspecified to be collection of linear functions or decision trees



Conclusions

Conclusions

Importance of covariates

Combine statistical learning with ranking and selection

- handle high-dimensionality with sparsity
- handle general dependence with empirical risk minimization