Confounder adjustment in large-scale linear structural models

Qingyuan Zhao

Department of Statistics, The Wharton School, University of Pennsylvania

June 19 2018, EcoStat

Based on

- Wang, J., Zhao, Q., Hastie, T., & Owen, A. B. Confounder adjustment in multiple hypothesis testing. *Annals of Statistics*, 45(5), 1863-1894, 2017.
- Song, Y., Zhao, Q. Performance evaluation in presence of latent factors. (In preparation).

Slides are available at `http://www-stat.wharton.upenn.edu/~qyzhao/`.

Setting

Multivariate linear regression

$$\underset{n\times p}{\boldsymbol{Y}} = \underset{n\times 1}{\boldsymbol{X}} \underset{p\times 1}{\boldsymbol{\alpha}}^T + \underset{n\times d}{\boldsymbol{Z}} \underset{p\times d}{\boldsymbol{\beta}}^T + \underset{n\times p}{\boldsymbol{\epsilon}}.$$

- $\boldsymbol{Y}$: "Panel data" or "transposable data". Modern datasets are often high dimensional (both $n, p \gg 1$).
- $\boldsymbol{X}$: "Primary variable", whose coefficients $\boldsymbol{\alpha}$ are of interest.
- $\boldsymbol{Z}$: "Control variables", whose coefficients $\boldsymbol{\beta}$ are not of interest (i.e. nuisance parameters).
- Noise $\boldsymbol{\epsilon} \sim \mathrm{MN}(\mathbf{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

Two examples

- Gene discovery: $\boldsymbol{Y}$ is gene expression (row: tissue; column: gene), $\boldsymbol{X}$ is the treatment.
- Mutual fund selectioin: $\boldsymbol{Y}$ is the monthly return of mutual funds (row: month; column: fund), $\boldsymbol{X}$ is the intercept, $\boldsymbol{Z}$ includes systematic risk factors.

The confounding problem

$$\underset{n\times p}{\boldsymbol{Y}} = \underset{n\times 1}{\boldsymbol{X}}\ \underset{p\times 1}{\boldsymbol{\alpha}}^T + \underset{n\times d}{\boldsymbol{Z}}\ \underset{p\times d}{\boldsymbol{\beta}}^T + \underset{n\times p}{\boldsymbol{\epsilon}}.$$

Omitted variable bias

When not all $\boldsymbol{Z}$ are known or measured, the OLS estimate of $\boldsymbol{\alpha}$ can be severely biased. To see this, suppose

$$\underset{n\times d}{\boldsymbol{Z}} = \underset{n\times 1}{\boldsymbol{X}}\ \underset{d\times 1}{\gamma}^T + \underset{n\times d}{\boldsymbol{W}},\ \text{where } \boldsymbol{W} \perp\!\!\!\perp \boldsymbol{X}.$$

Therefore $\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\gamma})^T + \boldsymbol{W}\boldsymbol{\beta}^T + \boldsymbol{\epsilon}$ and the OLS estimate of $\boldsymbol{\alpha}$ indeed converges to $\boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\gamma}$.

An illustrative example

The gender study[1]

Question: Which genes are more expressed in male/female?

A microarray experiment was conducted in this study:

- Postmortem samples from the brains of 10 individuals.
- For each individual, 3 samples from different cortices.
- Each sample is sent to 3 different labs for analysis.
- Two different microarray platforms are used by the labs.

In total, there are $10 \times 3 \times 3 = 90$ samples.

This example was first used by Gagnon-Bartsch and Speed [2] to demonstrate the importance of "removing unwanted variation" (RUV).

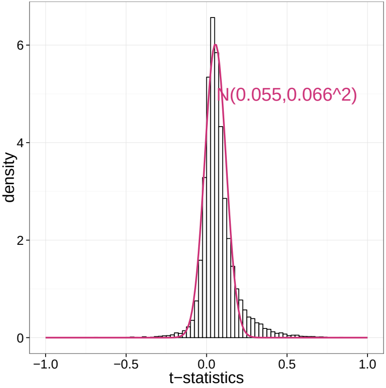
[1] Vawter, Marquis P., et al. "Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes." *Neuropsychopharmacology* 29.2 (2004).

[2] Gagnon-Bartsch, J. A., and Speed, T. P. "Using control genes to correct for unwanted variation in microarray data." *Biostatistics* 13.3 (2012).

A simple association test

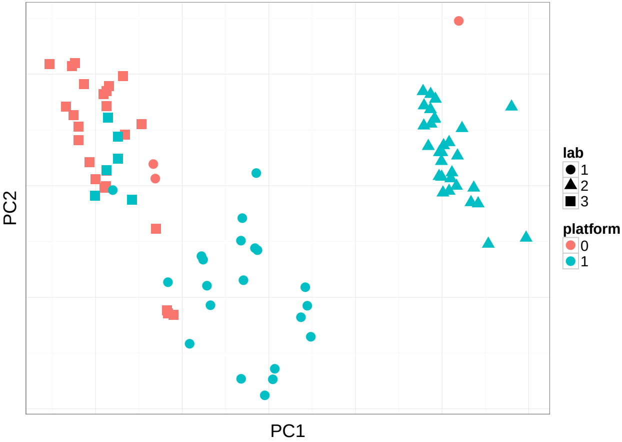
- Regress each column of $\boldsymbol{Y}$ (gene) on $\boldsymbol{X}$.
- In `R`, run `summary(lm(Y~X))`.
- Equivalent to a two-sample t-test with equal variance.

Histogram of t-statistics: skewed and underdispersed



What happened?

Plot of largest principle components



Our solution in a nutshell

Recall that (for simplicity, assume $\boldsymbol{Z}$ is entirely unobserved)

$$\underset{n\times p}{\boldsymbol{Y}} = \underset{n\times 1}{\boldsymbol{X}} \underset{p\times 1}{\boldsymbol{\alpha}}^T + \underset{n\times d}{\boldsymbol{Z}} \underset{p\times d}{\boldsymbol{\beta}}^T + \underset{n\times p}{\boldsymbol{\epsilon}}, \quad \underset{n\times d}{\boldsymbol{Z}} = \underset{n\times 1}{\boldsymbol{X}} \underset{d\times 1}{\boldsymbol{\gamma}}^T + \underset{n\times d}{\boldsymbol{W}}$$

$$\Downarrow$$

$$\boldsymbol{Y} = \boldsymbol{X}(\underbrace{\boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\gamma}}_{\boldsymbol{\tau}})^T + \boldsymbol{W}\boldsymbol{\beta}^T + \boldsymbol{\epsilon}.$$

Confounder adjusted testing and estimation (CATE)

1. OLS using the observed regressors:

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \approx \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\gamma}, \ \boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y} \approx \boldsymbol{W}\boldsymbol{\beta}^T + \boldsymbol{\epsilon}.$$

2. Factor analysis of $\boldsymbol{R}$ $\Rightarrow$ loading matrix $\hat{\boldsymbol{\beta}}$.
3. Path analysis: $\underset{p\times 1}{\hat{\boldsymbol{\tau}}} \approx \underset{p\times 1}{\boldsymbol{\alpha}} + \underset{p\times d}{\hat{\boldsymbol{\beta}}} \underset{d\times 1}{\boldsymbol{\gamma}}$.

Problem: the third step is not going to work because it has $(p+d)$ parameters but only p equations, i.e. $\boldsymbol{\alpha}$ **is not identified**.

Identification

Path analysis equation:

$$\underset{p\times 1}{\boldsymbol{\tau}} \approx \underset{p\times 1}{\boldsymbol{\alpha}} + \underset{p\times d}{\boldsymbol{\beta}} \; \underset{d\times 1}{\boldsymbol{\gamma}}.$$

- $\boldsymbol{\tau}$ and (the column space of) $\boldsymbol{\beta}$ can be identified from data.
- $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ cannot be identified from data. In other words, different values of $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ may correspond to the same distribution of the observed data.
- Solution to non-identifiability: put additional restrictions.

Proposition

Suppose $\boldsymbol{\Gamma}$ can be identified from the factor analysis. Then $\boldsymbol{\beta}$ is identifiable under either of the two following conditions:

1. **Negative control:** $\boldsymbol{\alpha}_{\mathcal{C}} = \mathbf{0}$ **for a known set** $\mathcal{C}$ such that $|\mathcal{C}| \geq d$ and $\mathrm{rank}(\boldsymbol{\beta}_{\mathcal{C}}) = d$.
2. **Sparsity:** $\|\boldsymbol{\alpha}\|_0 \leq \lfloor (p-d)/2 \rfloor$, and

$$\mathrm{rank}(\boldsymbol{\beta}_{\mathcal{C}}) = d, \; \forall \mathcal{C} \subset \{1, \ldots, p\} \text{ such that } |\mathcal{C}| = d.$$

Estimation under sparsity

Is sparsity reasonable?

Not always, but acceptable in our examples:

- In genomics screening, most genes are probably unrelated.
- Most mutual funds likely have no "alpha" (otherwise they will be quickly identified by the investors)[3]

Estimation via robust regression in CATE

Using a robust loss function $\rho(\cdot)$ (such as Huber's), solve

$$\hat{\gamma} = \arg\min_{\gamma} \sum_{j=1}^{p} \rho\left(\frac{\hat{\tau}_j - \hat{\boldsymbol{\beta}}_j^T \boldsymbol{\gamma}}{\hat{\sigma}_j}\right),$$

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\tau}} - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\gamma}}.$$

This is similar to solving a penalized regression in outlier detection:[4]

$$(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\alpha}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \left\|\hat{\boldsymbol{\tau}} - \boldsymbol{\alpha} - \hat{\boldsymbol{\beta}}\boldsymbol{\gamma}\right\|_{\hat{\boldsymbol{\Sigma}}}^2 + P_\rho(\boldsymbol{\alpha})$$

---

[3] Berk, J. B., & Green, R. C. (2004). "Mutual fund flows and performance in rational markets." *Journal of Political Economy*, 112(6).

[4] She, Y., & Owen, A. B. (2011). "Outlier detection using nonconvex penalized regression." *JASA*, 106.

Some theoretical guarantees

Theorem

When $n, p \to \infty$, if the factor analysis estimates[5] *of $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ are uniformly consistent, the robust loss function ρ is "nice", we have for a fixed j,*

1. *$\hat{\alpha}_j$ is consistent if $\|\boldsymbol{\beta}\|_1/p \to 0$;*
2. *$\hat{\alpha}_j$ is asymptotically normal and has "oracle efficiency" if $\|\boldsymbol{\beta}\|_1\sqrt{n}/p \to 0$.*

- "Oracle efficiency" means it has the same variance as the OLS estimator that observes the latent factors $\boldsymbol{Z}$.

[5] Bai, J., & Li, K. (2012). Statistical analysis of factor models of high dimension. *Annals of Statistics*, 40(1).

Mutual fund example

Dataset

Mutual fund returns from 1984—2015, obtained from Center for Research in Security Prices (CRSP).

Factor model

In finance, it is common to fit a linear model to the returns

$$\underbrace{Y_{tj} - r_t}_{Excess\ \ return} = \underbrace{\alpha_j}_{"Skill"\ \ of\ \ manager} + \underbrace{\boldsymbol{\beta}_j^T \boldsymbol{Z}_t}_{systematic\ \ risk} + \underbrace{\epsilon_{tj}}_{idiosyncratic\ \ risk} .$$

People have discovered many systematic risk factors $\boldsymbol{Z}$ over the years:

- Market-average: this is the Capital Asset Pricing Model (CAPM).
- Stock caps and book-to-market ratio[6].
- Momentum[7].
- ......

[6] Fama, E. F., & French, K. R. (1993). "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics*, 33(1).

[7] Carhart, M. M. (1997). "On persistence in mutual fund performance." *Journal of Finance*, 52(1).

Mutual fund selection by CAPM

A recent study[8] shows that

- Most investors use **CAPM-alpha** to select mutual funds.
- More sophisticated investors adjust for more risk factors.

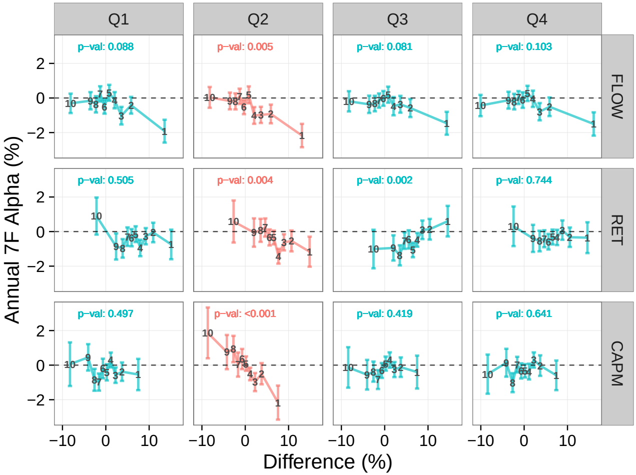
Is CAPM-alpha a good indicator for future performance?

An empirical exercise:

- In the beginning of every quarter, we use data in the past five years to compute their **cash flow**, **average returns**, and **CAPM-alpha**.
- For each metric, funds are then divided into **10 groups**.
- We evaluate the performance of each group in the next year.

[8] Barber, B. M., Huang, X., & Odean, T. (2016). "Which factors matter to investors? Evidence from mutual fund flows." *Review of Financial Studies*, 29(10)

Failure of CAPM-alpha



- Mutual funds with **higher** cash flow/return/CAPM-alpha have **worse** performance in the future.
- The phenomenon is not just "regression to the mean", but a complete reversal between past and future.

A possible explanation

Mutual funds also load on other risk factors.

Scenario 1: "Lucky" funds

1. When the other risk factors generated positive returns in the training period, the CAPM-alpha looks high.
2. High CAPM-alpha attracts investment.
3. Difficult to find investment opportunities $\Rightarrow$ bad future performance.

Scenario 2: "Unlucky" funds

1. When the other risk factors generated negative returns in the training period, the CAPM-alpha looks low.
2. Low CAPM-alpha repels investment.
3. Easier to invest $\Rightarrow$ good future performance.

Mutual fund selection by CATE

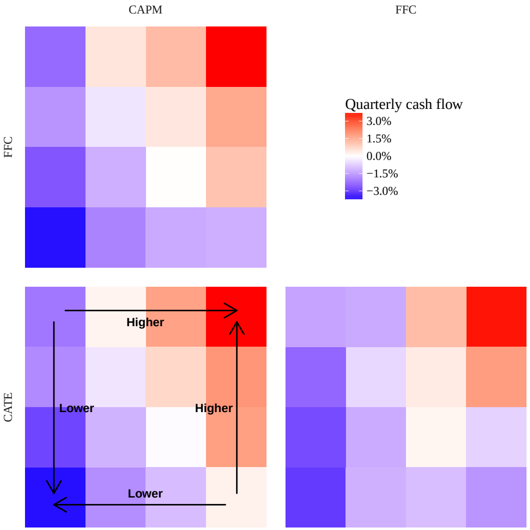
Better measurements of skill

- FFC-alpha: Use Fama-French-Carhart four factor model as $\boldsymbol{Z}$.
- CATE-alpha: In addition to FFC, use 3 latent factors

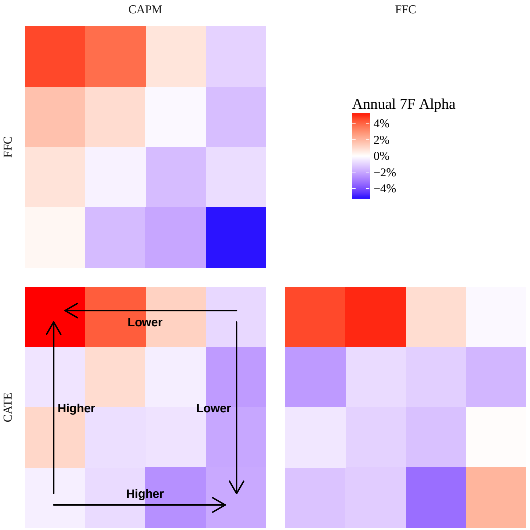
Another empirical exercise

- In the beginning of every quarter, we use data in the past five years to compute their **CAPM-alpha**, **FFC-alpha** and **CATE-alpha**.
- For each metric, funds are then divided into **4 groups**.
- For every two skill measurements, we examine the cash flow and the future return of the 4×4 grid.

High CAPM-alpha attracts investment



Reversal in future performance



Take-away messages

- We proposed a method to remove confounding bias (omitted variable bias) in multivariate linear regression.
- The key for identification and estimation is sparsity.
- Two applications were given:
  1. Remove batch effects in genomics screening;
  2. Estimate mutual fund skill in finance.
- The persistence of mutual fund performance depends on:
  - Whether the manager truly has skill (can be estimated by CATE);
  - Whether the investors have discovered it (usually using the incorrect CAPM).