Data enrichment for linear regression models

Aiyou Chen, Google Inc.

Art B. Owen*, Stanford University

Minghui Shi, Google Inc.

*The work reported here was done for Google, not Stanford.

Some big and small data

1) Small targeted data set S, e.g.,
   - matched to target population (e.g. NYC area shoppers)
   - panel selected by probability sample (e.g. geographic distribution)
   - high quality covariates (e.g., age, gender, education)
2) Bigger, less targeted data set B, e.g.,
   - related population (e.g. entire US), or
   - panel accepted all who opted in, or,
   - some covariates imputed from a model

Goal

Fit a model for population S

taking advantage (if possible) from data B

Data

Small sample (X_i, Y_i) $i \in S$ $|S| = n$ observations

Big sample (X_i, Y_i) $i \in B$ $|B| = N$ observations

Issues

- Focus is on S
- B might have a different X distn
- B might have a different $Y \mid X = x$ distn
- X might be measured differently in B

Main choices

1) Model for $i \in S$ only
2) Model with $i \in S \cup B$ (pooling)
3) Choose 1 or 2 based on hypothesis test
4) Shrinkage $\checkmark$

Regression setup

$$Y_i = \begin{cases} X_i\beta + \varepsilon_i & i \in S \\ X_i(\beta + \gamma) + \varepsilon_i & i \in B \end{cases}$$

$$\text{Bias}: \quad \gamma \in \mathbb{R}^d \qquad \text{Noise}: \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_S^2), \quad \mathcal{N}(0, \sigma_B^2)$$

Vector version

$$Y_S = X_S\beta + \varepsilon_S \qquad \& \qquad Y_B = X_B(\beta + \gamma) + \varepsilon_B$$

NOTES

- Google problems usually have categorical responses
- Gaussian assumption allows non-asymptotic results

Related literatures

Area	Method	One starting point
Survey sampling	Small area estimation	Rao (2003)
Chemometrics	Transfer calibration	Feudale et al. (2002)
Machine learning	Transfer learning	Cortes & Mohri (2011)
Medicine/Education	Meta-analysis	Borenstein et al. (2009)
Marketing	Data fusion	D'Orazio et al. (2006)

There are Bayesian approaches too.

Our approach is Steinian.

Data enriched regression

Minimize over β, γ:

$$\sum_{i\in S}(Y_i - X_i\beta)^2 + \sum_{i\in B}(Y_i - X_i(\beta+\gamma))^2 + \lambda P(\gamma)$$

$$= \quad \|Y_S - X_S\beta\|^2 \quad + \|Y_B - X_B(\beta+\gamma)\|^2 \quad + \lambda P(\gamma)$$

for fixed $\lambda \in [0,\infty]$ and penalty $P(\cdot)$.

Extreme cases

As $\lambda \to \infty$ we get pooling

As $\lambda \to 0$ we ignore the B data

Example penalties

$$\|\gamma\|^2 \qquad \|X_S\gamma\|^2 \qquad \|\gamma\|_1 \qquad \|X_S\gamma\|_1$$

First two are $\|X_T\gamma\|^2$, $X_T = X_S$ or $X_S = I$

As in ridge, we don't have to penalize the intercept

For large d we could/should also penalize β

Main findings

For $X_T = X_S$ or $X_T = I_d$ let

$$\hat{\beta}, \hat{\gamma} = \arg\min_{\beta,\gamma} \|Y_S - X_S\beta\|^2 + \|Y_B - X_B(\beta+\gamma)\|^2 + \lambda\|X_T\gamma\|^2$$

Our findings

1) how to compute $\hat{\beta}$ and $\hat{\gamma}$
2) fractional degrees of freedom as a function of λ
3) several ways to choose λ:
   AIC, AICc, cross-validation, bootstrap, plug-in
4) Stein-type result: using S only is inadmissible when $d \geqslant 5$ and error df $\geqslant 10$
5) Simulations validating theory

Stein results

Problem	Critical dimension	
Shrinking means to 0	$p \geqslant 3$	Stein
Shrinking means to common mean	$p \geqslant 4$	Efron & Morris
Shrinking regression coefficients	$p \geqslant 4$	Stein, Copas
Pooling two regression vectors	$p \geqslant 5$	us

Shrinking to something common seems to add 1 to the critical dimension

Our setting is a bit different. We measure loss only on S not B.

Estimation: just like ridge regression

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = (\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\mathcal{Y}$$

$$\mathcal{X} = \begin{pmatrix} X_S & 0 \\ X_B & X_B \\ 0 & \lambda^{1/2} X_T \end{pmatrix} \qquad \mathcal{Y} = \begin{pmatrix} Y_S \\ Y_B \\ 0 \end{pmatrix}$$

E.g. $X_T = X_S$ or $X_T = I_d$

Degrees of freedom

For a matrix $W_\lambda \in \mathbb{R}^{d\times d}$ we get

$$\hat\beta = W_\lambda \hat\beta_S + (I - W_\lambda)\hat\beta_B$$
$$\hat\beta_S = (X_S^\mathsf{T} X_S)^{-1} X_S^\mathsf{T} Y_S$$
$$\hat\beta_B = (X_B^\mathsf{T} X_B)^{-1} X_B^\mathsf{T} Y_B$$

Ye & Efron df

$$\mathrm{df}(\lambda) \equiv \frac{1}{\sigma_S^2}\sum_{i\in S}\mathrm{Cov}(Y_i, \hat Y_i) = \mathrm{tr}(W_\lambda)$$

DF continued

Take special case penalty $P(\gamma) = \|X_S\gamma\|^2$ (i.e., $X_T = X_S$)

After some algebra

Let $\nu_1, \nu_2, \dots, \nu_d$ be eigenvalues of

$$(X_S^\mathsf{T} X_S)^{1/2}(X_B^\mathsf{T} X_B)^{-1}(X_S^\mathsf{T} X_S)^{1/2}$$

Then

$$\mathrm{df}(\lambda) = \sum_{j=1}^{d} \frac{1+\lambda\nu_j}{1+\lambda+\lambda\nu_j}$$

Upshot

- Easy to find λ for desired df
- df$(0) = d$
- df(∞) can be < 1

Picking λ

AIC: minimize $n\log\hat\sigma_S^2(\lambda)+n\Bigl(1+\frac{2\mathrm{df}(\lambda)}{n}\Bigr)$

AICc: minimize $n\log\hat\sigma_S^2(\lambda)+n\Bigl(\frac{1+\mathrm{df}(\lambda)/n}{1-(\mathrm{df}(\lambda)+2)/n}\Bigr)$ Hurvich & Tsai (1989)

Plug in

Derive optimal λ as if we knew γ, σ_S and σ_B

$$\lambda_{\mathrm{orcl}}(\gamma,\sigma_S,\sigma_B)$$

Plug in estimates

$$\hat\lambda=\lambda_{\mathrm{orcl}}(\hat\gamma,\hat\sigma_S,\hat\sigma_B)$$

Bias-corrected plug-in:

adjust for bias, eg $\mathbb{E}(\hat\gamma^{\mathsf{T}}\hat\gamma)\ne\gamma^{\mathsf{T}}\gamma$ for $\hat\gamma=\hat\beta_B-\hat\beta_S$

Sample reuse

Bootstrap: re-sample both S and B

Cross-validation: K-fold split of S, retain all of B

Special case: location

$$X_S = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n \quad \text{and} \quad X_B = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^N$$

Rewrite model as

$$Y_i = \begin{cases} \mu + \varepsilon_i, & i \in S \\ \mu + \delta + \varepsilon_i, & i \in B \end{cases}$$

Minimize $\sum_{i \in S}(Y_i - \mu)^2 + \sum_{i \in B}(Y_i - \mu - \delta)^2 + \lambda\delta^2$

Get $\hat{\mu} = \omega \bar{Y}_S + (1 - \omega)\bar{Y}_B$

For $\omega = \frac{1+\lambda/N}{1+\lambda/N+\lambda/n}$

Location model

$$\hat{\mu} = \omega \bar{Y}_S + (1-\omega)\bar{Y}_B$$

We find

$$\omega_{\mathrm{orcl}} = \frac{\delta^2 + \sigma_B^2/N}{\delta^2 + \sigma_B^2/N + \sigma_S^2/n} \tag{1}$$

Judged by mean square error

Oracle's effective sample size is

$$n + \frac{\sigma_S^2}{\delta^2}$$

as $N \to \infty$ for fixed n.

Using the big sample adds information (does not multiply it).

Simulation (location)

$$X_i \sim \mathcal{N}(0,1), \qquad i \in S, \qquad n = 100$$

$$X_i \sim \mathcal{N}(\delta,1), \qquad i \in B, \qquad N = 1000$$

Relative bias

$$\delta_* = \frac{|\delta|}{\sigma_S/\sqrt{n}} = \sqrt{n}|\delta|$$

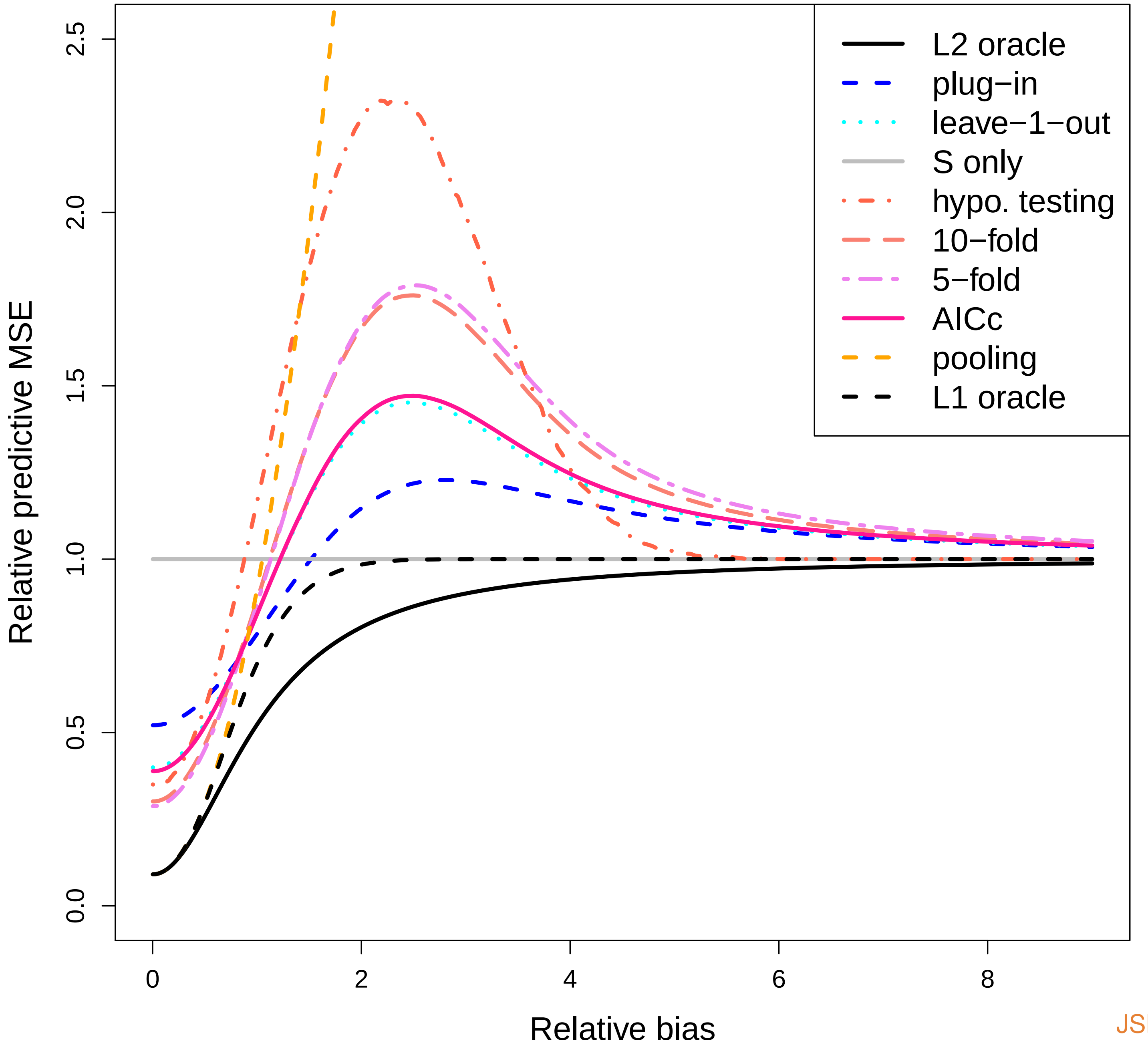
Relative MSE

$$\frac{(\hat{\mu}(\hat{\omega}) - \mu)^2}{\sigma_S^2/n}$$ equals 1 for $\bar{Y}_S$

repeat 10,000 times

NB: $\bar{Y}_S$ is admissible (Stein)

Simulation results



Simulation (regression)

Small sample, for $i \in S$

$$Y_i = X_i\beta + \varepsilon_i, \quad \text{(WLOG } \beta = 0\text{)}$$
$$X_i = (1, Z_i) \quad \text{(ie has intercept)}$$
$$Z_i \sim \mathcal{N}(0, C_S)$$
$$C_S \sim \text{Wishart}(I, d-1, d-1)$$

Big sample, for $i \in B$

$$Y_i = X_i\gamma + \varepsilon_i, \quad \gamma \text{ uniform on } d\text{-sphere}$$
$$X_i = (1, Z_i)$$
$$Z_i \sim \mathcal{N}(0, C_B)$$
$$C_B \sim \text{Wishart}(I, d-1, d-1)$$

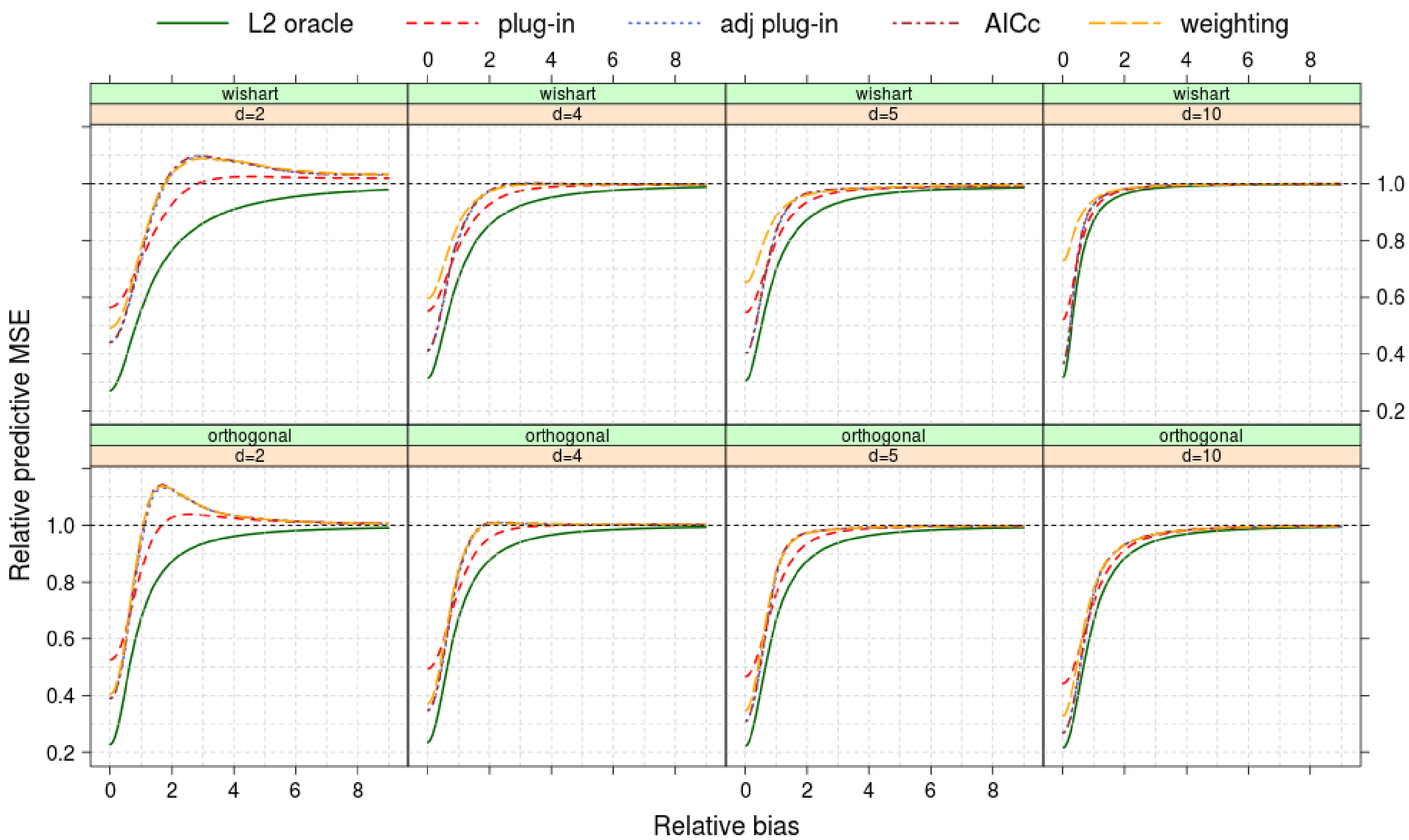
Second scenario

$$C_S = I + d u_S u_S^\mathsf{T} \qquad C_B = I + d u_B u_B^\mathsf{T} \qquad u_S, u_B \text{ uniform on } d-1\text{-sphere}$$

Sample sizes

$$n = 1000 \qquad N = 10{,}000$$

Regression results



Bias adjusted plug-in $\doteq$ AICc seem best overall (esp lower left & high dim)

Weighting is $\omega\hat\beta_S + (1-\omega)\hat\beta_B$ with plug-in ω

Inadmissibility of S only

1) Get the oracle's λ assuming $X_S^\mathsf{T} X_S = n\Sigma$ and $X_B^\mathsf{T} X_B = N\Sigma$
2) Plug-in estimates $\hat\gamma$, $\hat\sigma_S$, $\hat\sigma_B$ to pick λ
3) Resulting estimate makes $\hat\beta_S$ inadmissible
4) Even if assumption 1) is wrong

Inadmissibility ctd.

If $X_S^\mathsf{T} X_B = n\Sigma$ and $X_B^\mathsf{T} X_B = N\Sigma$, then

$$\hat\beta = \omega\hat\beta_S + (1-\omega)\hat\beta_B, \qquad 0 \leqslant \omega \leqslant 1$$

$$\omega_{\mathrm{orcl}} = \frac{\gamma^\mathsf{T}\Sigma\gamma + d\sigma_B^2/N}{\gamma^\mathsf{T}\Sigma\gamma + d\sigma_B^2/N + d\sigma_S^2/n}$$

Plug-in estimates

$$\hat\gamma = \hat\beta_B - \hat\beta_S, \quad \text{etc.}$$

$$\hat\omega_{\mathrm{plug}} = \frac{\hat\gamma^\mathsf{T}\Sigma\hat\gamma + d\hat\sigma_B^2/N}{\hat\gamma^\mathsf{T}\Sigma\hat\gamma + d\hat\sigma_B^2/N + d\hat\sigma_S^2/n}$$

$$\hat\omega_{\mathrm{plug},h} = \frac{\hat\gamma^\mathsf{T}\Sigma\hat\gamma + h(\hat\sigma_B^2)}{\hat\gamma^\mathsf{T}\Sigma\hat\gamma + h(\hat\sigma_B^2) + d\hat\sigma_S^2/n}$$

EG $h = 0$ or any $h \geqslant 0$ with $\mathbb{E}(h) < \infty$

Theorem

$X_S \in \mathbb{R}^{n\times d}, \qquad X_B \in \mathbb{R}^{N\times d}$ fixed full rank $\qquad X_S^\mathsf{T} X_S = n\Sigma$

$Y_S \sim \mathcal{N}(X_S\beta, \sigma_S^2 I_n) \qquad Y_B \sim \mathcal{N}(X_B(\beta+\gamma), \sigma_B^2 I_N)$ indep

If $d \geqslant 5$ and $n - d \geqslant 10$ then

$$\mathbb{E}\bigl(\|X_T(\hat\beta(\hat\omega) - \beta)\|^2\bigr) \; < \; \mathbb{E}\bigl(\|X_T(\hat\beta_S - \beta)\|^2\bigr)$$

For any $X_T^\mathsf{T} X_T = \Sigma$ and any $\hat\omega = \hat\omega_{\mathrm{plug},h}$

Chen, O, Shi (2012) on arXiv and

`http://research.google.com/pubs/pub41010.html`

Conclusions

There is something to gain by using data from a closely related sample

For $d \geqslant 5$ and $n - d \geqslant 10$ (and Gaussian data) ignoring the related sample is inadmissible

A key step is Stein's lemma which requires Gaussian data

We suspect the benefits extend beyond the Gaussian case

The algorithms but maybe not the theory extend to binary responses

Thanks

1) Co-authors Aiyou Chen and Minghui Shi of Google Inc.

2) Google for supporting this work

3) Helpful comments: Penny Chu, Corinna Cortes, Tony Fagan, Yijia Feng, Jerome Friedman, Jim Koehler, Diane Lambert, Elissa Lee & Nicolas Remy

4) Session chair: Steve Marron

5) Organizer: Yichao Wu