Discussion of “CARS: Covariate assisted ranking and screening for large-scale two-sample inference”

Guo Yu[*]　Jacob Bien[†]　Daniela Witten[‡]

December 19, 2018

In this discussion, we connect the authors’ elegant proposal to *multi-view data*, in which multiple sets of variables (or “views”) are measured on the same observations. Using ideas from Section 4 of Cai et al. (2019), we show that we can exploit a secondary view to improve power for testing on the first view.

Consider i.i.d. observations of m random variables under two conditions. In condition $\ell \in \{1, 2\}$, observation $i \in \{1, ..., n_\ell\}$ of variable $j \in \{1, ..., m\}$ is given by

$$\text{(View 1)} \quad X_{ij}(\ell) = \mu_j(\ell) + \varepsilon_{ij}(\ell),$$

where $\varepsilon_{ij}(\ell)$ is zero-mean, and we suppress the common intercept. The random mean vectors $\boldsymbol{\mu}(1)$ and $\boldsymbol{\mu}(2)$ are sparse. Furthermore, for the same individuals, we also observe a second view of $\tilde{m}$ variables,

$$\text{(View 2)} \quad Z_{ik}(\ell) = \tilde{\mu}_k(\ell) + \tilde{\varepsilon}_{ik}(\ell) \quad \text{for} \quad k \in \{1, ..., \tilde{m}\}.$$

The mean vectors $\tilde{\boldsymbol{\mu}}(\ell)$ are sparse, $\tilde{\varepsilon}_{ik}(\ell)$ is zero-mean, and again we suppress the intercept. Suppose the two views satisfy a hierarchical sparsity constraint: for $j \in \{1, ..., m\}$ and $\ell \in \{1, 2\}$,

$$\tilde{\mu}_{\sigma(j)}(\ell) = 0 \implies \mu_j(\ell) = 0, \tag{1}$$

where $\sigma(j)$ maps the jth entry of $\boldsymbol{\mu}(\ell)$ to its parent in $\tilde{\boldsymbol{\mu}}(\ell)$; see Figure 1.

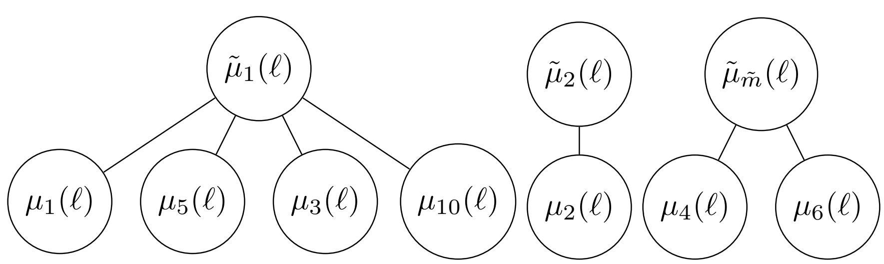

Figure 1: Schematic of (1), with $\sigma(3) = 1$.

[*]Statistics Department, University of Washington, Seattle, gy63@uw.edu
[†]Data Sciences and Operations Department, University of Southern California, Los Angeles, jbien@usc.edu
[‡]Statistics and Biostatistics Departments, University of Washington, Seattle, dwitten@uw.edu

Concretely, suppose $X(\ell)$ and $Z(\ell)$ contain protein and gene expression measurements, respectively. If transcripts that encode the jth protein are absent (i.e. $\tilde{\mu}_{\sigma(j)}(\ell) = 0$), then the jth protein cannot be present (i.e. $\mu_j(\ell) = 0$).

Suppose that $(\mu_j(1), \tilde{\mu}_{\sigma(j)}(1))$ is independent of $(\mu_j(2), \tilde{\mu}_{\sigma(j)}(2))$. Further assume that the random errors $(\varepsilon_{ij}(\ell), \tilde{\varepsilon}_{i\sigma(j)}(\ell))$ are bivariate normal and independent across j, ℓ and i, and independent of $\boldsymbol{\mu}(\ell)$ and $\tilde{\boldsymbol{\mu}}(\ell)$.

Using the terminology of Cai et al. (2019), the "primary statistic" for testing $H_{0j} : \mu_j(1) = \mu_j(2)$ is

$$T_j = C_j \left(\bar{X}_j(1) - \bar{X}_j(2) \right)$$

for some constant C_j. We consider a pair of "auxiliary statistics,"

$$R_j = D_j \left(\bar{X}_j(1) + \frac{n_2 \mathrm{Var}(\varepsilon_{ij}(1))}{n_1 \mathrm{Var}(\varepsilon_{ij}(2))} \bar{X}_j(2) \right), \qquad S_j = E_j \left(\bar{Z}_{\sigma(j)}(1) + \frac{n_2 \mathrm{Cov}(\varepsilon_{ij}(1), \tilde{\varepsilon}_{i\sigma(j)}(1))}{n_1 \mathrm{Cov}(\varepsilon_{ij}(1), \tilde{\varepsilon}_{i\sigma(j)}(2))} \bar{Z}_{\sigma(j)}(2) \right),$$

for some constants D_j and E_j. R_j is the same as T_{2j} in Cai et al. (2019), whereas S_j is constructed using the second data view. A small value of $|S_j|$ provides evidence for $\tilde{\mu}_{\sigma(j)}(1) = \tilde{\mu}_{\sigma(j)}(2) = 0$, which by (1) suggests that $\mu_j(1) = \mu_j(2)$. In analogy to Proposition 1 in Cai et al. (2019), the oracle statistic is

$$\begin{aligned} T_{OR}^{(j)}(t_j, r_j, s_j) \equiv \Pr(\theta_{1j} = 0 | T_j = t_j, R_j = r_j, S_j = s_j) &= \frac{f(t_j, r_j, s_j | \theta_{1j} = 0) \Pr(\theta_{1j} = 0)}{f(t_j, r_j, s_j)} \\ &= \frac{f(t_j | \theta_{1j} = 0) f(r_j, s_j | \theta_{1j} = 0) \Pr(\theta_{1j} = 0)}{f(t_j, r_j, s_j)}. \end{aligned}$$

Moreover, $T_{OR}^{(j)}(t_j, r_j, s_j)$ enjoys the properties in Theorem 3 of Cai et al. (2019). Detailed proofs are available at `https://hugogogo.github.io/paper/cars_discussion_supplement.pdf`. If there is not a one-to-one mapping between $\sigma(j)$ and j, then $T_{OR}^{(j)}(t_j, r_j, s_j)$ must be estimated carefully.

References

Cai, T. T., Sun, W. & Wang, W. (2019), 'CARS: Covariate assisted ranking and screening for large-scale two-sample inference'.