Discussion of "CARS: Covariate assisted ranking and screening for large-scale two-sample inference" by Cai et al.

Guo Yu* Jacob Bien† Daniela Witten‡

December 18, 2018

We congratulate the authors on an elegant new proposal based on a clever insight: the usual two-sample t-test discards information that can be exploited to potentially increase power. Using side-information in multiple testing with FDR control has gained notable recent attention (see, e.g., Ramdas et al. 2017, Li & Barber 2017, Lei & Fithian 2018, Banerjee et al. 2018). Here, we wish to connect this work to another active area of research: *multi-view data*, in which multiple sets of variables (or "views") are measured for a single set of observations. In Section 4, Cai et al. (2019) mention that the CARS framework can be extended with a more general auxiliary statistic constructed using side-information. In this discussion, we demonstrate that one could form an auxiliary statistic based on a secondary view to improve power for multiple testing on the first view, while enjoying the easy computation of the CARS framework.

As in Cai et al. (2019), suppose that we are given i.i.d. observations of m random variables under two experimental conditions. In particular, in experimental 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 $\varepsilon_{ij}(\ell)$ are zero-mean random errors and for simplicity we suppress the common intercept. We assume that the random mean vectors $\boldsymbol{\mu}(1)$ and $\boldsymbol{\mu}(2)$ are sparse and wish to test the null hypothesis $H_{0j} : \mu_j(1) = \mu_j(2)$. 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, the random errors $\tilde{\varepsilon}_{ik}(\ell)$ have zero means, and again we suppress the common intercept. Suppose the second view provides information about the first view through a hierarchical sparsity constraint. In particular, for $j \in \{1, ..., m\}$ and $\ell \in \{1, 2\}$, with probability 1,

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

*Post-doctoral research associate, Department of Statistics, University of Washington, Seattle, Washington, 98195, gy63@uw.edu

†Assistant Professor of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, jbien@usc.edu

‡Professor of Statistics and Biostatistics, University of Washington, Seattle, Washington, 98195, dwitten@uw.edu

where $\sigma(j)$ is a mapping that points the jth entry of $\boldsymbol{\mu}(\ell)$ to its parent in $\tilde{\boldsymbol{\mu}}(\ell)$. A schematic is shown in Figure 1.

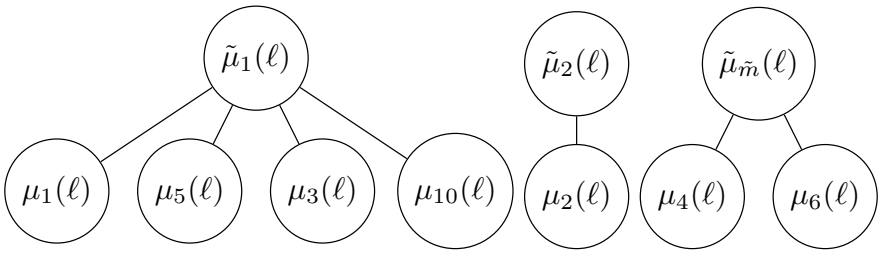

Figure 1: A schematic of the hierarchy. For example, here $\sigma(3) = 1$.

As a concrete example of this hierarchy, suppose that $X(\ell)$ is a set of protein expression measurements, and $Z(\ell)$ is a set of gene expression measurements. If transcripts for the gene that encodes 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 $(\mu_j(2), \tilde{\mu}_{\sigma(j)}(2))$ for all j. We further assume that for $j \in \{1, ..., m\}$, the random errors $(\varepsilon_{ij}(\ell), \tilde{\varepsilon}_{i\sigma(j)}(\ell))$ are bivariate normal and are independent across j, ℓ and i. Furthermore, we assume that all of the random errors are independent of $\boldsymbol{\mu}(\ell)$ and $\tilde{\boldsymbol{\mu}}(\ell)$.

Using the terminology of Cai et al. (2019), the "primary statistic" for testing $\mu_j(1) = \mu_j(2)$ 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. Note that R_j is the same as the auxiliary statistic T_{2j} in Cai et al. (2019), which only uses the internal information. By contrast, S_j is an auxiliary statistic constructed using external information. A small value of S_j provides evidence for $\tilde{\mu}_{\sigma(j)}(1) = \tilde{\mu}_{\sigma(j)}(2) = 0$, which by (1) facilitates the inference on whether $\mu_j(1) = \mu_j(2)$. In analogy to Proposition 1 in Cai et al. (2019), the oracle statistic can be simplified as

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

Moreover, $T_{OR}^{(j)}(t_j, r_j, s_j)$ enjoys the properties in Theorem 3 of Cai et al. (2019), and can be estimated using the strategies outlined in Sections 3.2 and 3.3 in Cai et al. (2019). We refer readers to the online supplementary material [1] for the proof of the results above. Finally, we note that if there is not a one-to-one mapping between $\sigma(j)$ and j, then special care will need to be paid to estimation of the oracle statistic, due to non-independence of the auxiliary test statistics.

[1] available at `https://hugogogo.github.io/paper/cars_discussion_supplement.pdf`

References

Banerjee, T., Mukherjee, G. & Sun, W. (2018), 'Adaptive sparse estimation with side information', *arXiv preprint arXiv:1811.11930* .

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

Lei, L. & Fithian, W. (2018), 'AdaPT: an interactive procedure for multiple testing with side information', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**(4), 649–679.

Li, A. & Barber, R. F. (2017), 'Accumulation tests for FDR control in ordered hypothesis testing', *Journal of the American Statistical Association* **112**(518), 837–849.

Ramdas, A., Barber, R. F., Wainwright, M. J. & Jordan, M. I. (2017), 'A unified treatment of multiple testing with prior knowledge using the p-filter', *arXiv preprint arXiv:1703.06222* .