Hierarchical heterogeneous analysis of high-dimensional data based on interactions

November 25, 2021

1 Model

Consider n independent subjects. For subject i, denote y_i be the response variable, and $X_i = (X_{i1}, \ldots, X_{ip})^\top$ and $Z_i = (Z_{i1}, \ldots, Z_{iq})^\top$ be the p- and q-dimensional vectors of environment (E) and image (I) measurements. Consider the heterogeneity model:

$$y_i = X_i^\top \beta_i + Z_i^\top \alpha_i + \sum_{s=1}^{q} W_i^{(s)} \eta_{is} + \varepsilon_i, \tag{1}$$

where $W_i^{(s)} = (Z_{is}X_{i1}, \ldots, Z_{is}X_{ip})$, $\beta_i = (\beta_{i1}, \ldots, \beta_{ip})^\top$, $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ip})^\top$, and $\eta_{is} = (\eta_{is1}, \ldots, \eta_{isp})^\top$ are the regression coefficients of subject i for the main E, main G, and their interactions respectively, and ε_i's are the random errors. Here the intercept term is omitted to simplify notation. To respect the "main effects, interactions" hierarchical constraint, we conduct the decomposition of η_{isj} as $\eta_{isj} = \beta_{ij}\gamma_{isj}$. Denote $\gamma_{is} = (\gamma_{is1}, \ldots, \gamma_{isp})^\top$. Then model (1) can be rewritten as

$$y_i = X_i^\top \beta_i + Z_i^\top \alpha_i + \sum_{s=1}^{q} W_i^{(s)} (\beta_i \odot \gamma_{is}) + \varepsilon_i, \tag{2}$$

where $\odot$ is the component-wise product. Let $B_i = (\beta_i^\top, \alpha_i^\top)^\top$ and $\gamma_i = (\gamma_{i1}^\top, \ldots, \gamma_{iq}^\top)^\top$.

We consider a hierarchical subgroup structure. Specifically, B_i's define a first-level heterogeneity structure with K_1 subgroups, and γ_i's define a second-level heterogeneity structure with K_2 subgroups. Each first-level subgroup is a union of one or several subgroups in the second level. Denote $\{\mathcal{G}_1, \ldots, \mathcal{G}_{K_1}\}$ as the collection of subject index sets of the K_1 first-level subgroups, and $\{\mathcal{T}_1, \ldots, \mathcal{T}_{K_2}\}$ as the collection of subject index sets of the K_2 second-level subgroups. Then there exists a partition of $\{1, \ldots, K_2\}$: $\{\mathcal{H}_1, \ldots, \mathcal{H}_{K_1}\}$ satisfying $\mathcal{G}_{k_1} = \bigcup_{k_2 \in \mathcal{H}_{k_1}} \mathcal{T}_{k_2}$ for $1 \leq k_1 \leq K_1$ and $1 \leq k_2 \leq K_2$. The subjects within k_1th first-level subgroup have the identical coefficients of main E and G, denoted as B_{k_1}, and those within k_2th second-level subgroup have the identical coefficients of interactions, denoted as γ_{k_2}. Overall, the response variable y_i's satisfy the following distribution:

$$f(y_i|X_i) = \sum_{k_1=1}^{K_1} \pi_{k_1} \sum_{k_2 \in \mathcal{H}_{k_1}} \frac{\pi_{k_2}}{\pi_{k_1}} f(y_i|X_i, B_{k_1}, \gamma_{k_2}) \tag{3}$$

$$= \sum_{k_2=1}^{K_2} \pi_{k_2} f(y_i|X_i, B_{\mathcal{F}(k_2)}, \gamma_{k_2}), \tag{4}$$

where π_{k_1} and π_{k_2} are unknown mixture probability of first- and second-level subgroups, respectively, and $\mathcal{F}(\cdot)$ is a mapping from second-level subgroup index sets to first-level subgroup index sets, i.e., $\mathcal{F}(k_2) = k_1$ for $k_2 \in \mathcal{H}_{k_1}$.

2 Penalized estimation

We propose the penalized objective functions:

$$\mathcal{L}(\boldsymbol{B}, \boldsymbol{\gamma}, \pi|\boldsymbol{X}, y) = -\sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k(y_i|X_i, B_k) + \mathcal{P}(\boldsymbol{B}, \boldsymbol{\gamma}), \tag{5}$$

where $\boldsymbol{X}$ and y denote the collection of observed data. $\boldsymbol{B} = (B_1, \dots, B_K)$, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)$, and $\pi = (\pi_1, \dots, \pi_k)^\top$,

$$\mathcal{P}(\boldsymbol{B}, \boldsymbol{\gamma}) = \sum_{k=1}^{K} \sum_{j=1}^{p} pen(|\beta_{kj}|, \lambda_1) + \sum_{k=1}^{K} \sum_{s=1}^{q} \sum_{j=1}^{p} pen(|\gamma_{ksj}|, \lambda_1) + \sum_{k<k'} pen(\sqrt{\|B_k - B_{k'}\|_2^2 + \|\gamma_k - \gamma_{k'}\|_2^2}, \lambda_2) \tag{6}$$

$$+ \sum_{k<k'} pen(\sqrt{\|B_k - B_{k'}\|_2^2}, \lambda_3), \tag{7}$$

$pen(\cdot, \lambda)$ is concave penalty function with tuning parameter $\lambda > 0$. In our numerical study, we adopt minimax concave penalty (MCP). K is a known constant that satisfies $K > K_2$. The proposed estimate is defined as the minimizer of objection function (5), denoted by $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\pi})$.