Online Appendices to "Data Aggregation and Demand Prediction"

Maxime C. Cohen, Renyu Zhang, Kevin Jiao

Appendix A: Standard Generalized Linear Models

We next discuss the standard generalized linear models for completeness. Interested readers are referred to McCullagh and Nelder (2019) for a comprehensive presentation of the GLM theory. In particular, we will use the decentralized model to illustrate the classical likelihood theory of generalized linear models. More specifically, for item i, we assume that the conditional distribution of $Y_{i,j}$ given the features $\boldsymbol{X}_{i,j}$ comes from an exponential family with the following density:

$$\mathbb{P}\big(Y_{i,j}|\boldsymbol{X}_{i,j}\big) = \exp\left\{\frac{Y_{i,j}\boldsymbol{X}'_{i,j}\boldsymbol{\beta}_i - H(\boldsymbol{X}'_{i,j}\boldsymbol{\beta}_i)}{H_2(\gamma)} + H_3(Y_{i,j},\gamma)\right\}, \qquad (18)$$

where $\gamma \in \mathbb{R}^+$ is a known scale parameter, and $H(\cdot)$, $H_2(\cdot)$, and $H_3(\cdot)$ are three real-valued normalization functions. The exponential family in Eq. (18) is very broad and includes Gaussian, binomial, Poisson, gamma, and inverse-Gaussian as special cases. It is straightforward to derive that, under the true parameter $\boldsymbol{\beta}_i$, the condition expectation of the outcome satisfies

$$\mathbb{E}\big[Y_{i,j}|\boldsymbol{X}_{i,j}\big] = H'(\boldsymbol{X}'_{i,j}\boldsymbol{\beta}_i) = G(\boldsymbol{X}'_{i,j}\boldsymbol{\beta}_i),$$

and the conditional variance of the outcome satisfies

$$\mathbb{V}\big(Y_{i,j}|\boldsymbol{X}_{i,j}\big) = H''(\boldsymbol{X}'_{i,j}\boldsymbol{\beta}_i) = G'(\boldsymbol{X}'_{i,j}\boldsymbol{\beta}_i)H_2(\gamma).$$

The log-likelihood function of parameter $\boldsymbol{b}_i$ for item i under model (18) is thus given by

$$\log\mathcal{L}_i(\boldsymbol{b}_i|\boldsymbol{Y}_i,\boldsymbol{X}_i) = \sum_{j=1}^{m}\left[\frac{Y_{i,j}\boldsymbol{X}'_{i,j}\boldsymbol{b}_i - H(\boldsymbol{X}'_{i,j}\boldsymbol{b}_i)}{H_2(\gamma)} + H_3(Y_{i,j},\gamma)\right] = \frac{1}{H_2(\gamma)}\cdot\sum_{j=1}^{m}\big[Y_{i,j}\boldsymbol{X}'_{i,j}\boldsymbol{b}_i - H(\boldsymbol{X}'_{i,j}\boldsymbol{b}_i)\big] + \text{constant},$$

where the constant is independent of the parameter $\boldsymbol{b}_i$. Therefore, the decentralized MLE $\hat{\boldsymbol{b}}_i$ is given by Eq. (3), which is equivalent to an iterative weighted least-squares procedure. The statistical theory of GLM and MLE establishes the asymptotic and finite sample properties of the decentralized estimator $\hat{\boldsymbol{b}}_i$. See Proposition 1 for more details.

Appendix B: Two Potential Methods

In this section, we introduce two potential methods to estimate the model in Eq. (1) and predict the demand, as well as discuss why these methods are not applicable to our setting.

B.1. Generalized ℓ_1-Regularized MLE

The first potential method we consider is the *generalized ℓ_1-/lasso-regularized MLE* (see, e.g., Tibshirani 1996, Tibshirani and Taylor 2011, Hastie et al. 2019) to estimate the coefficients. This approach revises the standard MLE by adding a generalized ℓ_1-regularizer. More specifically, the ℓ_1-regularized log-likelihood function of the aggregate model is given by:

$$\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\big[Y_{i,j}\boldsymbol{X}'_{i,j}\boldsymbol{\beta}_i - H(\boldsymbol{X}'_{i,j}\boldsymbol{\beta}_i)\big] - \lambda\left(\sum_{i\neq i'}\sum_{l=1}^{d}|\beta_{i,l} - \beta_{i',l}|\right),\ \lambda > 0, \qquad (19)$$

where $H(\cdot)$ is the normalization mapping that satisfies $H'(u) = G(u)$ (see Appendix A). A canonical result in the statistics literature shows that ℓ_1-regularization will shrink the regularized terms to 0 and, thus, generate sparse solutions (see, e.g., Tibshirani 1996, Tibshirani and Taylor 2011, Zou and Hastie 2005). As a result, adding a generalized ℓ_1-regularizer to MLE may potentially be helpful to capture the fact that a feature at the aggregate or cluster level shares the same coefficient for different items. We note that ℓ_1-regularized MLE is in a similar spirit to the *fused lasso regression* (see, e.g., Tibshirani and Taylor 2011, Tibshirani et al. 2005). In the retail demand forecasting literature, Huang et al. (2014) and Ma et al. (2016) develop Lasso-based methodological frameworks to overcome the problem of the ultra-high dimensionality of the feature space under multiple product categories.

As shown by Tibshirani and Taylor (2011) and Ramdas and Tibshirani (2016), generating the ℓ_1-regularized MLE typically involves solving the dual problem multiple times along the solution path. Given the high-dimensional nature of the convex optimization problem in Eq. (19) (i.e., the number of decision variables is nd, which is at the magnitude of thousand or more in practice), estimating the coefficients is computationally prohibitive even for a linear regression specification (i.e., $G(u) = u$) as it involves inverting $(nd) \times (nd)$-matrices in each step to construct the solution path (see, e.g., Tibshirani and Taylor 2011, Ramdas and Tibshirani 2016). Therefore, though theoretically plausible, using the ℓ_1-regularized MLE is not tractable for our problem in practical settings.

B.2. Direct Optimization

We next consider the direct optimization approach that directly formulates the problem as a nonlinear program to jointly estimate the data aggregation levels, cluster structures, and feature coefficients. Since the data aggregation levels and cluster structures are unknown apriori, we need to use one-hot encoding to represent the aggregation levels and cluster structures. More specifically, we use $\delta^s_{i,l}$ to denote the indicator variable for feature l of item i to be at the aggregate level, $\delta^n_{i,l}$ to denote the indicator variable for feature l of item i to be at the individual level, and $\delta^c_{i,l,\varsigma}$ to denote the indicator variable for feature l of item i to be at the cluster level and item i being in cluster $\mathcal{C}_{l,\varsigma}$. Thus, there is a total of $2nd + n\sum_{l=1}^{d} k_l$ binary decision variables. For expositional convenience, we consider the linear regression model (i.e., $G(u) = u$). Then, the mean squared loss minimization can be written as:

$$\min_{\boldsymbol{\beta},\boldsymbol{\delta}} \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(Y_{i,j} - \boldsymbol{X}'_{i,j}\boldsymbol{\beta}_i\right)^2 .$$

The constraints are not straightforward, so we next list them one by one. First, the δ variables are binary:

$$\delta^s_{i,l} \in \{0,1\},\ \delta^n_{i,l} \in \{0,1\},\ \text{and}\ \delta^c_{i,l,\varsigma} \in \{0,1\},\ \text{for all}\ 1 \le i \le n, 1 \le l \le d, 1 \le \varsigma \le k_l. \quad (20)$$

Second, a feature can be at one (and only one) data aggregation level, that is,

$$\delta^s_{i,l} + \delta^n_{i,l} + \sum_{\varsigma=1}^{k_l} \delta^c_{i,l,\varsigma} = 1,\ \text{for all}\ 1 \le i \le n, 1 \le l \le d. \quad (21)$$

Third, an aggregate-level feature should be at the aggregate level for all items:

$$\delta^s_{i,l} = \delta^s_{i',l},\ \text{for all}\ i \ne i', 1 \le l \le d. \quad (22)$$

Similarly, a SKU-level feature should be at the SKU level for all items:

$$\delta^n_{i,l} = \delta^n_{i',l}, \text{ for all } i \neq i', 1 \leq l \leq d. \tag{23}$$

By combining Eqs. (21), (22), and (23), it implies that a cluster-level feature should be at the cluster level for all items, namely,

$$\sum_{\varsigma=1}^{k_l} \delta^c_{i,l,\varsigma} = \sum_{\varsigma=1}^{k_l} \delta^c_{i',l,\varsigma}, \text{ for all } i \neq i', 1 \leq l \leq d.$$

The number of items in each cluster with respect to each cluster-level feature is at least two:

$$\sum_{i=1}^{n} \delta^c_{i,l,\varsigma} \geq \frac{2}{n} \sum_{\varsigma'=1}^{k_l} \sum_{i=1}^{n} \delta^c_{i,l,\varsigma'}, \text{ for all } 1 \leq l \leq d \text{ and } 1 \leq \varsigma \leq k_l. \tag{24}$$

We note that the left-hand side of Eq. (24) quantifies the total number of items in cluster ς with respect to feature l. Accordingly, one needs to consider two cases. First, if feature l is not at the cluster level, then the left-hand side of Eq. (24) should be equal to 0. By combining Eqs. (21), (22), and (23), this is equivalent to $\delta^c_{i,l,\varsigma} = 0$ for all item i and cluster ς, that is, the left-hand and right-hand sides of Eq. (24) are both equal to 0, and hence Eq. (24) holds in this case. Second, if feature l is at the cluster level, then the left-hand side of Eq. (24) should be at least equal to two. In fact, in this case, the RHS of Eq. (24) is

$$\frac{2}{n} \sum_{\varsigma'=1}^{k_l} \sum_{i=1}^{n} \delta^c_{i,l,\varsigma'} = \frac{2}{n} \sum_{i=1}^{n} \sum_{\varsigma'=1}^{k_l} \delta^c_{i,l,\varsigma'} = \frac{2}{n} \cdot n \cdot 1 = 2,$$

where the second equality follows from Eqs. (21), (22), and (23). Thus, the constraint in Eq. (24) is equivalent to requiring that the number of items in each cluster with respect to each cluster-level feature is at least equal to two.

For an aggregate-level feature, its coefficient should be the same for all items:

$$-2\bar{\beta}(1 - \delta^a_{i,l}) \leq \beta_{i,l} - \beta_{i',l} \leq 2\bar{\beta}(1 - \delta^a_{i,l}), \text{ for all } i \neq i' \text{ and } 1 \leq l \leq d, \tag{25}$$

where $\bar{\beta}$ is the maximum possible absolute value of the coefficients. Analogously, for a cluster-level feature and the items within the same cluster with respect to this feature, the coefficient should be identical:

$$-2\bar{\beta}(2 - \delta^c_{i,l,\varsigma} - \delta^c_{i',l,\varsigma}) \leq \beta_{i,l} - \beta_{i',l} \leq 2\bar{\beta}(2 - \delta^c_{i,l,\varsigma} - \delta^c_{i',l,\varsigma}), \text{ for all } i \neq i', \ 1 \leq l \leq d, \text{ and } 1 \leq \varsigma \leq k_l. \tag{26}$$

Based on (20), (21), (22), (23), (24), (25), and (26), we formulate the direct optimization approach as the following mixed-integer second-order conic program (SOCP):

$$
\begin{aligned}
\min_{\boldsymbol{\beta},\boldsymbol{\delta}} \quad & \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(Y_{i,j}-\boldsymbol{X}'_{i,j}\boldsymbol{\beta}_i\right)^2 \\
\text{s.t.} \quad & -\bar{\beta}\le \beta_{i,l}\le \bar{\beta}\ \ i=1,2,\ldots,n,\ \ l=1,2,\ldots,d, \\
& \delta^s_{i,l}\in\{0,1\},\ \delta^n_{i,l}\in\{0,1\},\ \text{and}\ \delta^c_{i,l,\varsigma}\in\{0,1\},\ \text{for all } 1\le i\le n, 1\le l\le d, 1\le \varsigma\le k_l, \\
& \delta^s_{i,l}+\delta^n_{i,l}+\sum_{\varsigma=1}^{k_l}\delta^c_{i,l,\varsigma}=1,\ \text{for all } 1\le i\le n, 1\le l\le d, \\
& \delta^s_{i,l}=\delta^s_{i',l},\ \text{for all } i\ne i', 1\le l\le d, \\
& \delta^n_{i,l}=\delta^n_{i',l},\ \text{for all } i\ne i', 1\le l\le d, \\
& \sum_{i=1}^{n}\delta^c_{i,l,\varsigma}\ge \frac{2}{n}\sum_{\varsigma'=1}^{k_l}\sum_{i=1}^{n}\delta^c_{i,l,\varsigma'},\ \text{for all } 1\le l\le d \text{ and } 1\le\varsigma\le k_l, \\
& -2\bar{\beta}(1-\delta^a_{i,l})\le \beta_{i,l}-\beta_{i',l}\le 2\bar{\beta}(1-\delta^a_{i,l}),\ \text{for all } i\ne i' \text{ and } 1\le l\le d, \\
& -2\bar{\beta}(2-\delta^c_{i,l,\varsigma}-\delta^c_{i',l,\varsigma})\le \beta_{i,l}-\beta_{i',l}\le 2\bar{\beta}(2-\delta^c_{i,l,\varsigma}-\delta^c_{i',l,\varsigma}),\ \text{for all } i\ne i',\ 1\le l\le d,\ \text{and } 1\le\varsigma\le k_l.
\end{aligned}
\tag{27}
$$

We note that the mixed-integer SOCP in Eq. (27) has nd continuous decision variables, $2nd+n\sum_{l=1}^{d}k_l$ binary decision variables, and $O\left(n^2d\left(\sum_{l=1}^{d}k_l\right)\right)$ linear constraints, which is intractable for a practical problem of a reasonable size.

A similar generalized clusterwise linear regression (CLR) model has been proposed by Park et al. (2017) to address a special case of our problem where all the features are at the cluster level and the cluster structure is the same across all features. Park et al. (2017) show that the generalized CLR is NP-hard and propose column generation and metaheuristic genetic algorithms to solve this problem. Since our problem in Eq. (27) is more general with unknown data aggregation levels, the estimation methods proposed by Park et al. (2017) are not applicable and tractable in our setting.

Alternatively, one may solve problem (27) via a procedure that iteratively estimates the continuous coefficients and the binary decision variables for aggregation levels and cluster structures (in a similar way as in Baardman et al. 2017). The iterative procedure will stop once the binary variables remain the same for two consecutive iterations. Baardman et al. (2017) address the demand prediction problem when there are only cluster-level features (i.e., no aggregate-level and no SKU-level features). In their setting, this iterative procedure was proved to converge to the true coefficients and cluster structure (i.e., the estimate is consistent). In our setting, however, we cannot guarantee the consistency of the iterative optimization approach due to the heterogeneous data aggregation levels and the unknown cluster structures in our model. As a result, the iterative procedure is not a viable approach to solve problem (27) and estimate our model.

In conclusion, both the generalized ℓ_1-regularized MLE and the direct optimization approaches cannot be used to solve our problem in practice.

Appendix C: Proofs of Statements

We next provide the proofs of all the technical results.

Proof of Proposition 1

Part (a). The proof of the consistency results follows from a standard result in statistics stating that the maximum-likelihood estimator (MLE) is consistent under some regularity conditions that are satisfied by a generalized linear model. See, for example, Fahrmeir et al. (1985) and McCullagh and Nelder (2019).

Part (b). We first show the asymptotic normality in Eq. (5). This is a standard result in the MLE literature, which follows directly from, e.g., Theorem 3 of Fahrmeir et al. (1985).

We next prove the finite-sample normality result in Eq. (4). Since the smallest eigenvalue of $\Sigma_i = \mathbb{E}\left[\boldsymbol{X_{i,j}X'_{i,j}}\right]$, $\lambda_{\min}(\Sigma_i)$, is strictly positive, by Proposition 1 of Li et al. (2017), there exists a (sufficiently large) threshold $\mathfrak{m}'_i$ such that, as long $m \ge \mathfrak{m}'_i$, the smallest eigenvalue of $\hat{\boldsymbol{V}}_i(m) := \sum_{j=1}^{m} \boldsymbol{X}_{i,j}\boldsymbol{X}'_{i,j}$, $\lambda_{\min}(\hat{\boldsymbol{V}}_i(m))$, can be arbitrarily large as m increases to infinity. Therefore, the condition of Theorem 1 in Li et al. (2017) (i.e., Equation (4) thereof) is satisfied.

We define $\boldsymbol{x} \in \mathbb{R}^d$ with $x_l = 1$ and all other $x_{l'} = 0$. Thus, $\boldsymbol{x}'(\hat{\boldsymbol{b}}(m)_i - \boldsymbol{\beta}_i) = \hat{b}_{i,l}(m) - \beta_{i,l}$ and the ℓ_2-norm of x associated with $\hat{V}_i(m)$ is

$$||\boldsymbol{x}||_{\hat{\boldsymbol{V}}_i(m)^{-1}} = \sqrt{\boldsymbol{x}'\hat{\boldsymbol{V}}_i(m)^{-1}\boldsymbol{x}} = \sqrt{(\hat{\boldsymbol{V}}_i(m)^{-1})_{l,l}} = \sqrt{\frac{1}{m}\left(\left(\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{X}_{i,j}\boldsymbol{X}'_{i,j}\right)^{-1}\right)_{l,l}} \le \sqrt{\frac{2(\Sigma_i^{-1})_{l,l}}{m}}, \tag{28}$$

when $m \ge \mathfrak{m}'_{i,l}$ for some threshold $\mathfrak{m}'_{i,l}$ by the strong law of large numbers. For any $\epsilon > 0$, we define

$$\delta := \exp\left(-\psi_{i,l}\cdot\epsilon^2\cdot m\right) \quad \text{where } \psi_{i,l} := \frac{\underline{g}_i^2}{18(\Sigma_i^{-1})_{l,l}\sigma^2}, \text{ and } \underline{g}_i := \inf\left\{G'(\boldsymbol{z}'\boldsymbol{b}_i) : \boldsymbol{z}\in\mathbb{R}^d, ||\boldsymbol{z}|| \le 1, ||\boldsymbol{b}_i - \boldsymbol{\beta}_i|| \le 1\right\} > 0.$$

Hence, δ satisfies

$$\frac{3\sigma}{\underline{g}_i}\cdot\sqrt{\log\left(\frac{1}{\delta}\right)}\cdot\sqrt{\frac{2(\Sigma_i^{-1})_{l,l}}{m}} = \epsilon.$$

Thus, if $m \ge \mathfrak{m}_{i,l} := \max\{\mathfrak{m}'_i, \mathfrak{m}'_{i,l}\}$, with probability at least $1 - 3\delta$, the following inequality holds:

$$|\hat{b}_{i,l}(m) - \beta_{i,l}| = |\boldsymbol{x}'(\hat{\boldsymbol{b}}(m)_i - \boldsymbol{\beta}_i)| \le \frac{3\sigma}{\underline{g}_i}\cdot\sqrt{\log\left(\frac{1}{\delta}\right)}\cdot||\boldsymbol{x}||_{\hat{\boldsymbol{V}}_i(m)^{-1}} \le \frac{3\sigma}{\underline{g}_i}\cdot\sqrt{\log\left(\frac{1}{\delta}\right)}\cdot\sqrt{\frac{2(\Sigma_i^{-1})_{l,l}}{m}} = \epsilon, \tag{29}$$

where the first inequality follows from Theorem 1 in Li et al. (2017) and the second from Eq. (28). Inequality (29) immediately implies that, for $m \ge \mathfrak{m}_{i,l}$, inequality (4) holds, hence proving Proposition 1. □

Proof of Proposition 2

The proof is similar to the proof of Proposition 1, so we only sketch it for brevity. We note that we can reformulate the aggregate model as a new GLM with $m \times n$ observations. We denote the outcome vector as $\tilde{\boldsymbol{Y}} \in \mathbb{R}^{mn}$. The feature design matrix $\tilde{\boldsymbol{X}}$ has $m \times n$ rows (representing observations) and d_x columns (representing the total number of features):

$$\tilde{Y}_j = G\left(\sum_{l=1}^{d_x}\tilde{X}_j^l\tilde{\beta}_l\right) + \epsilon_j, \tag{30}$$

where $\tilde{\beta}_l$ is the coefficient for feature l in the aggregate model and ϵ_j is the independent sub-Gaussian error term. With the new formulation in Eq. (30), the aggregate model can be viewed as a decentralized model with one item, d_x features, and $m \times n$ observations.

For **part (a)**, the proof follows directly from the standard MLE theory. For **part (b)**, both the finite sample and asymptotic normality follow from a similar argument as in the proof of Proposition 1. Finally, if $\mathcal{D}_c \cup \mathcal{D}_s \neq \emptyset$, then $\sqrt{m} b^a_{i,l}(m) = \sqrt{m} b^a_{i',l}(m)$ for $i \neq i'$ and $\mathcal{C}(i,l) = \mathcal{C}(i',l)$. In this case, the asymptotic distribution of $\sqrt{m}(\boldsymbol{b}^a(m) - \boldsymbol{\beta})$ is clearly degenerate, because it has some equal coordinates for all m. This concludes the proof of Proposition 2. □

Proof of Proposition 3

Part (a). We first show the inequality in Eq. (7).

- *Step 1.* If $\beta_{1,l} \neq \beta_{i,l}$ for some $i \neq 1$, then $\lim_{m \uparrow +\infty} \mathbb{P}\left[H^l_{1,i} \text{ is not rejected}\right] = 0$. This implies that the probability of Type-II error to falsely identify two different coefficients to be the same converges to 0.

We now assume that $\beta_{1,l} \neq \beta_{i,l}$ and use the notation $H^l_{1,i}(m)$ to make the dependence of $H^l_{1,i}$ on the sample size m explicit. The probability that $H^l_{1,i}(m)$ is not rejected (resp. is rejected) is denoted by $\mathfrak{p}(m)$ (resp. $\mathfrak{q}(m) := 1 - \mathfrak{p}(m)$). Since $\sqrt{m}\hat{b}_{1,l}(m)$ and $\sqrt{m}\hat{b}_{i,l}(m)$ are asymptotically normally distributed, there exists a constant $c_{i,l} > 0$ independent of m such that $H^l_{1,i}(m)$ is not rejected if and only if

$$|\sqrt{m}\hat{b}_{1,l}(m) - \sqrt{m}\hat{b}_{i,l}(m)| \leq c_{i,l}, \text{ for } m \text{ sufficiently large.}$$

We define $\varepsilon := \frac{1}{3} \cdot |\beta_{1,l} - \beta_{i,l}| > 0$. We assume that $m > \left(\frac{c_{i,l}}{\varepsilon}\right)^2$, i.e., $\frac{c_{i,l}}{\sqrt{m}} < \varepsilon$ and consider the case where $|\hat{b}_{1,l}(m) - \beta_{1,l}| \leq \varepsilon$ and $|\hat{b}_{i,l}(m) - \beta_{i,l}| \leq \varepsilon$. In this case,

$$|\hat{b}_{1,l}(m) - \hat{b}_{i,l}(m)| \geq \frac{1}{3}|\beta_{1,l} - \beta_{i,l}| = \varepsilon > \frac{c_{i,l}}{\sqrt{m}}.$$

Therefore, if $|\hat{b}_{1,l}(m) - \beta_{1,l}| \leq \varepsilon$ and $|\hat{b}_{i,l}(m) - \beta_{i,l}| \leq \varepsilon$, $H^l_{1,i}(m)$ will be rejected, which implies that if m is sufficiently large, we have

$$\begin{aligned}
\mathfrak{q}(m) \geq & \mathbb{P}\left[|\hat{b}_{1,l}(m) - \beta_{1,l}| \leq \varepsilon, |\hat{b}_{i,l}(m) - \beta_{i,l}| \leq \varepsilon\right] \\
= & 1 - \mathbb{P}\left[|\hat{b}_{1,l}(m) - \beta_{1,l}| > \varepsilon \text{ or } |\hat{b}_{i,l}(m) - \beta_{i,l}| > \varepsilon\right] \\
\geq & 1 - \mathbb{P}\left[|\hat{b}_{1,l}(m) - \beta_{1,l}| > \varepsilon\right] - \mathbb{P}\left[|\hat{b}_{i,l}(m) - \beta_{i,l}| > \varepsilon\right] \\
\geq & 1 - 3\exp(-\psi_{1,l}\varepsilon^2 m) - 3\exp(-\psi_{i,l}\varepsilon^2 m),
\end{aligned} \tag{31}$$

where the second inequality follows from the union bound and the last inequality from Eq. (4). Inequality (31) implies that $\lim_{m \uparrow +\infty} \mathfrak{q}(m) = 1$, or equivalently, $\lim_{m \uparrow +\infty} \mathfrak{p}(m) = 0$, which proves *Step 1.* This also implies that, as $m \uparrow +\infty$, the probability that the DAC_α algorithm mis-specifies an individual-level feature as a cluster- or aggregate- level one, or a cluster-level feature as an aggregate-level one will shrink to 0 exponentially fast.

- *Step 2.* If $l \in \mathcal{D}_c$, then Step 6 of Algorithm 1 will produce a consistent estimate of the cluster structure with respect to feature l, that is,

$$\lim_{m \uparrow +\infty} \mathbb{P}\left[(\hat{\mathcal{C}}_{1,l}, \hat{\mathcal{C}}_{2,l}, \ldots, \hat{\mathcal{C}}_{k_l,l}) \text{ is a permutation of } (\mathcal{C}_{1,l}, \mathcal{C}_{2,l}, \ldots, \mathcal{C}_{k_l,l})\right] = 0.$$

We now fix feature $l \in \mathcal{D}_c$. We note that, for any $i \in \mathcal{C}_\varsigma$ $(1 \leq \varsigma \leq k_l)$, the coefficient of feature l is $\beta^c_{\varsigma,l}$ and $\hat{b}_{i,l}$ converges to $\beta^c_{\varsigma,l}$ with a probability that exponentially decays in the sample size m. Thus, for the k-means algorithm $(k = k_l)$ applied to $\{\hat{b}_{1,l}, \hat{b}_{2,l}, ..., \hat{b}_{n,l}\}$, the centers of the k_l clusters $\{\hat{c}_{1,l}, \hat{c}_{2,l}, ..., \hat{c}_{k_l,l}\}$, where $\hat{c}_\varsigma$ is the center of cluster $\mathcal{C}_\varsigma$, will converge to the true coefficient vectors for cluster-level features $\beta^c_{\varsigma,l}$ up to a permutation on $\{1,2,...,k_l\}$. For notational convenience, we assume that $\hat{c}_{\varsigma,l}$ converges to $\beta^c_{\varsigma,l}$ for $1\le\varsigma\le k_l$.

If there is an item $i\in\mathcal{C}_{\varsigma,l}$ that is "mis-clustered" into $\hat{\mathcal{C}}_{\varsigma',l}$, we have, as $m\uparrow+\infty$, $\hat{b}_{i,l_1}$ converges to $\beta^c_{\varsigma,l}\neq\beta^c_{\varsigma',l}$, that is, for m sufficiently large,

$$|\hat{b}_{i,l}-\beta^c_{\varsigma,l}|<|\hat{b}_{i,l}-\beta^c_{\varsigma',l}|.$$

This implies that, for m sufficiently large,

$$|\hat{b}_{i,l}-\hat{c}_{\varsigma,l}|<|\hat{b}_{i,l}-\hat{c}_{\varsigma',l}|,$$

which contradicts the assumption that item $i\in\mathcal{C}_{\varsigma,l}$ is mis-clustered into $\hat{\mathcal{C}}_{\varsigma',l}$ and hence concludes the proof of *Step 2*. This also implies that, if $m\uparrow+\infty$, as long as a cluster-level feature is correctly specified, the cluster structure can also be correctly identified with probability 1.

- *Step 3.* Given any significance level $\alpha\in(0,1)$, the probability that the DAC_α algorithm mis-specifies any cluster-level feature as an individual one, or any aggregate-level feature as a cluster-level one or an individual-level one is upper bounded by $p(\alpha)>0$ as $m\uparrow+\infty$.

We first note that the probability that the DAC_α algorithm mis-specifies any cluster-level feature as an individual one, or any aggregate-level feature as a cluster-level one or an individual-level one is upper bounded by the probability of the event that all the features are at the aggregate level but the algorithm mis-specifies some feature to be at the cluster level or the individual level. We define the latter probability as $p(\alpha)$, which is the probability that $H^l_{1,i}$ is rejected for at least one (i,l) under $2\le i\le n$ and $1\le l\le d$ under the condition that all the features are at the aggregate level. For the rest of the proof of *Step 3*, we assume that all features are at the aggregate level

We next quantify $p(\alpha)$ using multiple hypothesis testing (MHT) in the asymptotic regime ($m\uparrow+\infty$). With a slight abuse of notation, we use $\hat{\boldsymbol{b}}\in\mathbb{R}^{nd}$ to denote a virtual estimator following the same distribution as the limiting distribution (i.e., $m\uparrow+\infty$) of $\sqrt{m}(\hat{\boldsymbol{b}}(m)-\boldsymbol{\beta})$. By Proposition 1(b), $\hat{\boldsymbol{b}}$ follows a zero-mean multivariate normal distribution with covariance matrix $\mathcal{V}:=\operatorname{diag}(\mathcal{I}_1(\boldsymbol{\beta}_1)^{-1},\mathcal{I}_2(\boldsymbol{\beta}_2)^{-1},...,\mathcal{I}_n(\boldsymbol{\beta}_n)^{-1})$, which is block diagonal. We define an $(n-1)d$-by-nd matrix $\mathcal{T}$ such that $\hat{\boldsymbol{t}}:=\mathcal{T}\cdot\hat{\boldsymbol{b}}$ is the joint estimator for Step 6 (Hypothesis Testing) of Algorithm 1, that is,

$$\hat{\boldsymbol{t}}:=\mathcal{T}\cdot\hat{\boldsymbol{b}}=\begin{pmatrix}\hat{b}_{1,1}-\hat{b}_{2,1}\\ \hat{b}_{1,1}-\hat{b}_{3,1}\\ \dots\\ \hat{b}_{1,1}-\hat{b}_{n,1}\\ \dots\\ \hat{b}_{1,d}-\hat{b}_{2,d}\\ \hat{b}_{1,d}-\hat{b}_{3,d}\\ \dots\\ \hat{b}_{1,d}-\hat{b}_{n,d}\end{pmatrix}\in\mathbb{R}^{(n-1)d}.$$

Therefore, $\hat{\boldsymbol{t}}$ is normally distributed with mean $\mathbf{0}\in\mathbb{R}^{(n-1)d}$ and covariance matrix $\tilde{\mathcal{V}}:=\mathcal{T}\cdot\mathcal{V}\cdot\mathcal{T}'$. Then, in Algorithm 1, $\mathcal{H}^l_{1,i}$ ($1\le l\le d$ and $2\le i\le n$) is rejected if and only if $\tilde{t}_{(n-1)(l-1)+i-1}=\hat{b}_{1,l}-\hat{b}_{i,l}$ is located outside the interval

$\mathbb{I}_j(\alpha):=\left[-\mathcal{V}_{j,j}\Phi^{-1}\left(1-\dfrac{\alpha}{2}\right),\mathcal{V}_{j,j}\Phi^{-1}\left(1-\dfrac{\alpha}{2}\right)\right]$, where $j:=(n-1)(l-1)+i-1$ and $\Phi^{-1}(\cdot)$ is the inverse $\Phi(\cdot)$.

We define $\mathbb{I}(\alpha) := \prod_{j=1}^{(n-1)d} \mathbb{I}_j(\alpha)$ as the Cartesian product of all $\mathbb{I}_j(\alpha)$. We then have

$$p(\alpha) = \mathbb{P}\left[\hat{\boldsymbol{t}} \notin \mathbb{I}(\alpha)\right], \text{ where } \hat{\boldsymbol{t}} \sim \mathcal{N}(\mathbf{0}, \tilde{V}).$$

We have now completed the proof of *Step 3*.

- *Step 4.* The probability $p(\alpha)$ is strictly decreasing in α with $\lim_{\alpha \downarrow 0} p(\alpha) = 0$.

It is clear by definition that $\mathbb{I}_j(\alpha_1) \subset \mathbb{I}_j(\alpha_2)$ for $\alpha_1 > \alpha_2$, so $\mathbb{I}(\alpha_1) = \prod_{j=1}^{(n-1)d} \mathbb{I}_j(\alpha_1) \subset \prod_{j=1}^{(n-1)d} \mathbb{I}_j(\alpha_2) = \mathbb{I}(\alpha_2)$ for $\alpha_1 > \alpha_2$. Therefore, for $\alpha_1 > \alpha_2$,

$$p(\alpha_1) = \mathbb{P}\left[\hat{\boldsymbol{t}} \notin \mathbb{I}(\alpha_1)\right] > \mathbb{P}\left[\hat{\boldsymbol{t}} \notin \mathbb{I}(\alpha_1)\right] = p(\alpha_2),$$

where the inequality follows from $\mathbb{I}(\alpha_1) \subset \mathbb{I}(\alpha_2)$. Finally, to prove that $\lim_{\alpha \downarrow 0} p(\alpha) = 0$, we note that $\lim_{\alpha \downarrow 0} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = +\infty$. Thus, $\lim_{\alpha \downarrow 0} \mathbb{I}(\alpha) = \mathbb{R}^{(n-1)d}$, which implies that

$$\lim_{\alpha \downarrow 0} p(\alpha) = \lim_{\alpha \downarrow 0} \mathbb{P}\left[\hat{\boldsymbol{t}} \notin \mathbb{I}(\alpha)\right] = 0,$$

where the last equality follows from the monotone convergence theorem. This completes the proof of *Step 4*. Furthermore, by combining *Step 1*, *Step 2*, *Step 3*, and *Step 4*, we conclude that inequality (7) holds.

- *Step 5.* The DAC_α estimator $\hat{\boldsymbol{\beta}}^\alpha$ is consistent.

If the data aggregation levels $(\mathcal{D}_s, \mathcal{D}_c, \mathcal{D}_n)$ and the cluster structure $\{\mathcal{C}_{1,l}, \mathcal{C}_{2,l}, \ldots, \mathcal{C}_{k_l,l}\}$ are correctly identified, Proposition 2 implies the consistency of $\hat{\boldsymbol{\beta}}^\alpha$. We next consider the following two cases: (i) A Type-I error occurs, under which Algorithm 1 falsely identifies two identical coefficients to be different; and (ii) A Type-II error occurs, under which Algorithm 1 falsely identifies two different coefficients to be the same. *Step 1* implies that the probability of Type-II error converges to 0 as the sample size m goes to infinity. For the case of Type-I error, the model is not mis-specified and, as a consequence, the same argument as in the proof of Proposition 1(a) implies that Step 7 of Algorithm 1 consistently estimates the true coefficient $\boldsymbol{\beta}$. This completes the proof of *Step 5* and of Proposition 3(a).

Part (b). Once again, we consider the following three cases:

- *Case 1.* The data aggregation levels $(\mathcal{D}_s, \mathcal{D}_c, \mathcal{D}_n)$ and the cluster structure $\{\mathcal{C}_{1,l}, \mathcal{C}_{2,l}, \ldots, \mathcal{C}_{k_l,l}\}$ are correctly identified. The event of this case is denoted by $\mathcal{E}_1(m)$, where we make the dependence on the sample size m explicit.

- *Case 2.* A Type-I error occurs but there is no Type-II error, under which Algorithm 1 falsely identifies two identical coefficients to be different. The event of this case is denoted by $\mathcal{E}_2(m)$, where we make the dependence on the sample size m explicit.

- *Case 3.* A Type-II error occurs, under which Algorithm 1 falsely identifies two different coefficients to be the same. The event of this case is denoted by $\mathcal{E}_3(m)$, where we make the dependence on the sample size m explicit.

We note that $\mathbb{P}\Big[\mathcal{E}_1(m)\cup\mathcal{E}_2(m)\cup\mathcal{E}_3(m)\Big]=1$ by definition. We first establish the following, for any $\epsilon>0$:

$$\begin{aligned}\mathbb{P}\Big[|\hat{\beta}^{\alpha}_{i,l}(m)-\beta_{i,l}|>\epsilon\Big] \le & \mathbb{P}\Big[|\hat{\beta}^{\alpha}_{i,l}(m)-\beta_{i,l}|>\epsilon,\mathcal{E}_1(m)\Big]+\mathbb{P}\Big[|\hat{\beta}^{\alpha}_{i,l}(m)-\beta_{i,l}|>\epsilon,\mathcal{E}_2(m)\Big]+\mathbb{P}\Big[|\hat{\beta}^{\alpha}_{i,l}(m)-\beta_{i,l}|>\epsilon,\mathcal{E}_3(m)\Big]\\ =&\sum_{j=1}^{3}\mathbb{P}\Big[|\hat{\beta}^{\alpha}_{i,l}(m)-\beta_{i,l}|>\epsilon\Big|\mathcal{E}_j(m)\Big]\mathbb{P}\Big[\mathcal{E}_j(m)\Big]\\ \le&\sum_{j=1}^{2}\mathbb{P}\Big[|\hat{\beta}^{\alpha}_{i,l}(m)-\beta_{i,l}|>\epsilon\Big|\mathcal{E}_j(m)\Big]+\mathbb{P}\Big[\mathcal{E}_3(m)\Big],\end{aligned} \tag{32}$$

where the first inequality follows from the union bound and from $\mathbb{P}\Big[\mathcal{E}_1(m)\cup\mathcal{E}_2(m)\cup\mathcal{E}_3(m)\Big]=1$, and the second inequality from $\mathbb{P}\Big[|\hat{\beta}^{\alpha}_{i,l}(m)-\beta_{i,l}|>\epsilon\Big|\mathcal{E}_2(m)\Big]\le 1$ and $P\Big[\mathcal{E}_j(m)\Big]\le 1$ $(j=1,2)$. The above equality follows from the definition of conditional probability. To prove Eq. (8), it suffices to show that there exist two constants $c_1>0$ and $c_2>0$, such that, for any $\epsilon>0$ and sufficiently large m, the following holds:

$$\mathbb{P}\Big[|\hat{\beta}^{\alpha}_{i,l}(m)-\beta_{i,l}|>\epsilon\Big|\mathcal{E}_j(m)\Big]\le c_1\exp(-c_2\epsilon^2 m),\ \ j=1,2, \tag{33}$$

$$\mathbb{P}\Big[\mathcal{E}_3(m)\Big]\le c_1\exp(-c_2\epsilon^2 m). \tag{34}$$

We next quantify the concentration bounds in Eqs. (33) and (34) for the three cases separately.

Case 1. In this case, $\mathcal{E}_1(m)$ holds true. Therefore, inequality (33) (for $j=1$) follows immediately from Proposition 2(b).

Case 2. In this case, $\mathcal{E}_2(m)$ holds true. There are $O(nd)$ sub-cases that differ on the estimation results of the data aggregation levels. For each sub-case, Proposition 2(b) also holds (though each with a different aggregate model to estimate). Thus, by applying the law of total probability, inequality (33) (for $j=2$) follows.

Case 3. In this case, inequality (31) implies that inequality (34) holds. Plugging inequalities (33) and (34) into (32) implies that there exist constants $\eta^{\alpha}_{i,l}>0$ and $\psi^{\alpha}_{i,l}>0$ such that inequality (8) holds for any $m>\mathfrak{m}^{\alpha}_{i,l}$ by setting the threshold $\mathfrak{m}^{\alpha}_{i,l}$ sufficiently large. It thus concludes the proof of Proposition 3(b). □

Proof of Proposition 4

Part (a). By Proposition 1(b) (Eq. (5) in particular), $\sqrt{m}(\hat{b}_{i,l}(m)-\beta_{i,l})$ converges in distribution to a single-dimensional normal distribution with mean 0 and variance $\kappa_{i,l}=(\mathcal{I}_i(\boldsymbol{\beta}_i)^{-1})_{l,l}$. Thus, we have

$$\lim_{m\uparrow+\infty} m\mathbb{E}(\hat{b}_{i,l}(m)-\beta_{i,l})^2=\lim_{m\uparrow+\infty}\mathbb{E}\Big[\sqrt{m}(\hat{b}_{i,l}(m)-\beta_{i,l})\Big]^2=\kappa_{i,l},$$

namely, Eq. (9) holds for all $1\le i\le n$ and $1\le l\le d$. This completes the proof of Proposition 4(a).

Part (b). The proof relies on analyzing the log-likelihood functions of the decentralized and aggregate models carefully. Hence, we first introduce some notation. We define the (empirical average) log-likelihood of item i with the data sample $\{(Y_{i,j},\boldsymbol{X}_{i,j}):j=1,2,\dots,m\}$ as

$$\mathfrak{L}_i(\boldsymbol{b}_i;m):=\frac{1}{m}\sum_{j=1}^{m}\log\mathcal{L}_i(\boldsymbol{\beta}_i|Y_{i,j},\boldsymbol{X}_{i,j})=\frac{1}{m}\sum_{j=1}^{m}[Y_{i,j}\boldsymbol{X}'_{ij}\boldsymbol{b}_i-H(\boldsymbol{X}'_{i,j}\boldsymbol{b}_i)],$$

where we ignore a constant independent of data for the last equality and $H'(u)=G(u)$. Thus, the decentralized estimator for item i is $\hat{\boldsymbol{b}}_i(m)=\arg\max_{\boldsymbol{b}_i}\mathfrak{L}_i(\boldsymbol{b}_i;m)$.

We also denote the gradient and Hessian of the log-likelihood function associated with item i by $\nabla\mathfrak{L}_i(\boldsymbol{b}_i;m)$ and $\nabla_2\mathfrak{L}_i(\boldsymbol{b}_i;m)$, respectively. The Fisher information matrix with respect to the decentralized model of item i is thus given by $\mathcal{I}_i(\boldsymbol{b}_i) = -\mathbb{E}[\nabla_2\mathfrak{L}_i(\boldsymbol{b}_i;1)]$, where the expectation is taken with respect to the true value of $\boldsymbol{b}_i = \boldsymbol{\beta}_i$ and the true distribution of $(Y_{i,1},\boldsymbol{X}_{i,1})$. By using the law of large numbers, we have $\lim_{m\uparrow+\infty}\nabla_2\mathfrak{L}_i(\boldsymbol{b}_i;m) = -\mathcal{I}_i(\boldsymbol{b}_i)$. Likewise, we define the log-likelihood of all the items as follows:

$$\mathfrak{L}(\boldsymbol{b};m) = \sum_{i=1}^{n}\mathfrak{L}_i(\boldsymbol{b}_i;m) := \sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\log\mathcal{L}_i(\boldsymbol{b}_i|Y_{i,j},\boldsymbol{X}_{i,j}) = \sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}[Y_{i,j}\boldsymbol{X}'_{ij}\boldsymbol{b}_i - H(\boldsymbol{X}'_{i,j}\boldsymbol{b}_i)].$$

Hence, the aggregate estimator is defined by $\hat{\boldsymbol{b}}^a(m) = \arg\max_{\boldsymbol{b}\in\Xi}\mathfrak{L}(\beta;m)$, where the feasible parameter set Ξ is defined as in Eq. (6). We denote the gradient and Hessian of $\mathfrak{L}(\boldsymbol{b};m)$ as $\nabla\mathfrak{L}(\boldsymbol{b};m) = \sum_{i=1}^{n}\nabla\mathfrak{L}_i(\boldsymbol{b};m)$ and $\nabla_2\mathfrak{L}(\boldsymbol{b};m) = \sum_{i=1}^{n}\nabla_2\mathfrak{L}_i(\boldsymbol{b};m)$, respectively. We are now ready to prove Proposition 4(b) in different steps.

- *Step 1.* The aggregate estimator $\hat{\boldsymbol{b}}^a(m)$ satisfies the following expected squared error:

$$\lim_{m\uparrow+\infty} m\cdot\mathbb{E}(b^a_{i,l}(m)-\beta_{i,l})^2 = \left(\frac{1}{n_{i,l}}\right)^2\cdot\left(\sum_{i'\in\mathcal{C}(i,l)}\kappa_{i',l}\right),\ \text{for all } 1\le i\le n \text{ and } 1\le l\le d, \tag{35}$$

where $\kappa_{i,l}$'s are defined in Proposition 4.

Based on the decentralized estimator, $\hat{\boldsymbol{b}}(m)$, we first construct the following new estimator $\hat{\theta}_{i,l}(m)$ for each item i and feature l:

$$\hat{\theta}_{i,l}(m) = \frac{1}{n_{i,l}}\sum_{i'\in\mathcal{C}(i,l)}\hat{b}_{i',l}(m).$$

By the consistency and asymptotic normality of $\hat{\boldsymbol{b}}(m)$ (Proposition 1), we have $\hat{\theta}_{i,l}(m)\xrightarrow{p}\beta_{i,l}$, for each i and l; and

$$\begin{aligned}\lim_{m\to+\infty} m\cdot\mathbb{E}(\hat{\theta}_{i,l}(m)-\beta_{i,l})^2 &= \lim_{m\to+\infty} m\cdot\mathbb{E}\Big(\frac{1}{n_{i,l}}\sum_{i'\in\mathcal{C}(i,l)}\hat{b}_{i',l}(m)-\beta_{i,l}\Big)^2\\ &= \left(\frac{1}{n_{i,l}}\right)^2\cdot\lim_{m\to+\infty} m\cdot\mathbb{E}\Big(\sum_{i'\in\mathcal{C}(i,l)}(\hat{b}_{i',l}(m)-\beta_{i,l})\Big)^2\\ &=\left(\frac{1}{n_{i,l}}\right)^2\cdot\sum_{i'\in\mathcal{C}(i,l)}\Big[\lim_{m\to+\infty} m\cdot\mathbb{E}\Big(\hat{b}_{i',l}(m)-\beta_{i,l}\Big)^2\Big]\\ &=\left(\frac{1}{n_{i,l}}\right)^2\cdot\Big(\sum_{i'\in\mathcal{C}(i,l)}\kappa_{i',l}\Big),\ \text{for } i=1,2,\ldots,n,\ \text{and } l=1,2,\ldots,d,\end{aligned} \tag{36}$$

where the third equality follows from the fact that the demand of different items are independent and the last equality follows from Proposition 4(a).

We next apply the Taylor expansion of $\nabla\mathfrak{L}_i(\cdot;m)$ around the true parameter value $\boldsymbol{\beta}_i$ for each i:

$$\nabla\mathfrak{L}_i(\hat{\boldsymbol{b}}_i(m);m) = \nabla\mathfrak{L}_i(\boldsymbol{\beta}_i;m) + \nabla_2\mathfrak{L}_i(\boldsymbol{\beta}_i;m)\cdot(\hat{\boldsymbol{b}}_i(m)-\boldsymbol{\beta}_i) + o(||\hat{\boldsymbol{b}}_i(m)-\boldsymbol{\beta}_i||),$$

where $o(\cdot)$ refers to the standard “Little-o Notation” applied to each component of the vector.

Since $\hat{\boldsymbol{b}}_i(m)$ is the maximizer of $\mathfrak{L}_i(\cdot;m)$, the first-order condition applies, that is, $\nabla\mathfrak{L}_i(\hat{\boldsymbol{b}}_i(m);m) = 0$. Thus, by plugging this into the Taylor expansion of $\nabla\mathfrak{L}_i(\cdot;m)$ we obtain

$$\nabla\mathfrak{L}_i(\boldsymbol{\beta}_i;m) + \nabla_2\mathfrak{L}_i(\boldsymbol{\beta}_i;m)\cdot(\hat{\boldsymbol{b}}_i(m)-\boldsymbol{\beta}_i) + o(||\hat{\boldsymbol{b}}_i(m)-\boldsymbol{\beta}_i||) = 0,\ \text{for each } i. \tag{37}$$

Analogously, we apply the Taylor expansion of $\nabla\mathfrak{L}(\cdot;m)$ around the true parameter value $\boldsymbol{\beta}$:

$$\nabla\mathfrak{L}(\hat{\boldsymbol{b}}^a;m) = \nabla\mathfrak{L}(\boldsymbol{\beta};m) + \nabla_2\mathfrak{L}(\boldsymbol{\beta};m)\cdot(\hat{\boldsymbol{b}}^a(m)-\boldsymbol{\beta}) + o(||\hat{\boldsymbol{b}}^a(m)-\boldsymbol{\beta}||),$$

that is,

$$\sum_{i=1}^{n}\nabla\mathfrak{L}_i(\hat{\boldsymbol{b}}_i^a(m);m) = \sum_{i=1}^{n}\nabla\mathfrak{L}_i(\boldsymbol{\beta};m) + \sum_{i=1}^{n}\nabla_2\mathfrak{L}_i(\boldsymbol{\beta}_i;m)\cdot(\hat{\boldsymbol{b}}_i^a(m)-\boldsymbol{\beta}_i) + o(||\hat{\boldsymbol{b}}^a(m)-\boldsymbol{\beta}||).$$

Since $\hat{\boldsymbol{b}}^a(m)$ is the maximizer of $\mathfrak{L}(\cdot;m)$ under the constraint that $\beta_{i,l} = \beta_{i',l}$ for all $i' \in \mathcal{C}(i,l)$, we have $\sum_{i'\in\mathcal{C}(i,l)}\nabla^l\mathfrak{L}_{i'}(\hat{\beta}_{i'}(m);m) = 0$, where the operator ∇^l refers to the partial derivative with respect to feature l. Thus, for each item i, it follows that

$$\sum_{i'\in\mathcal{C}(i,l)}\nabla^l\mathfrak{L}_{i'}(\hat{\boldsymbol{b}}_{i'}^a(m);m) = \sum_{i'\in\mathcal{C}(i,l)}\nabla^l\mathfrak{L}_{i'}(\boldsymbol{\beta}_{i'};m) + \sum_{i'\in\mathcal{C}(i,l)}\nabla_2^l\mathfrak{L}_{i'}(\boldsymbol{\beta}_{i'};m)\cdot(\hat{\boldsymbol{b}}_{i'}^a(m)-\boldsymbol{\beta}_{i'}) + o(||\hat{\boldsymbol{b}}(m)-\boldsymbol{\beta}||) = 0, \tag{38}$$

where ∇_2^l is the l-th row of the Hessian.

From Eq. (37), we have for each i and each l,

$$\sum_{i'\in\mathcal{C}(i,l)}\nabla^l\mathfrak{L}_{i'}(\hat{\boldsymbol{b}}_{i'}(m);m) = \sum_{i'\in\mathcal{C}(i,l)}\nabla^l\mathfrak{L}_{i'}(\boldsymbol{\beta}_{i'};m) + \sum_{i'\in\mathcal{C}(i,l)}\nabla_2^l\mathfrak{L}_{i'}(\boldsymbol{\beta}_{i'};m)\cdot(\hat{\boldsymbol{b}}_{i'}(m)-\boldsymbol{\beta}_{i'}) + o(||\hat{\boldsymbol{b}}(m)-\boldsymbol{\beta}||) = 0. \tag{39}$$

We note that $\hat{b}_{i',l}^a(m) = \hat{b}_{i,l}^a(m)$ for all $i' \in \mathcal{C}(i,l)$, so that in total there are $n_{i,l} = |\mathcal{C}(i,l)|$ coefficients identical to $\hat{b}_{i,l}^a(m)$. By plugging this identity into Eq. (38) and subtracting Eq. (39), we obtain the following, for each i and each l:

$$\left|n_{i,l}\hat{b}_{i,l}^a(m) - \sum_{i'\in\mathcal{C}(i,l)}\hat{b}_{i,l}(m)\right| = o(m^{-\frac{1}{2}}),\ \text{i.e.,}\ |\hat{b}_{i,l}^a(m) - \hat{\theta}_{i,l}(m)| = o(m^{-\frac{1}{2}}), \tag{40}$$

where we used the facts that $||\hat{\boldsymbol{b}}(m)-\boldsymbol{\beta}|| = O(m^{-\frac{1}{2}})$ and $||\hat{\boldsymbol{b}}^a(m)-\boldsymbol{\beta}|| = O(m^{-\frac{1}{2}})$ (by applying the strong law of large numbers, namely, $\lim_{m\uparrow+\infty}\nabla_2^l\mathfrak{L}_i(\boldsymbol{\beta}_i;m) = -\mathcal{I}_i^l(\boldsymbol{\beta}_i)$ for any i and l). Thus, by Eq. (40),

$$\mathbb{E}\Big(\hat{b}_{i,l}^a(m) - \hat{\theta}_{i,l}(m)\Big)^2 = o(m^{-1})\ \text{for each } i \text{ and each } l. \tag{41}$$

For each i and each l, we have

$$\begin{aligned} & m\cdot\mathbb{E}\Big(\hat{b}_{i,l}^a(m) - \beta_{i,l}\Big)^2 \\ =& m\cdot\mathbb{E}\Big(\hat{b}_{i,l}^a(m) - \hat{\theta}_{i,l}(m) + \hat{\theta}_{i,l}(m) - \beta_{i,l}\Big)^2 \\ =& m\cdot\Big[\mathbb{E}\Big(\hat{b}_{i,l}^a(m) - \hat{\theta}_{i,l}(m)\Big)^2 + \mathbb{E}\Big(\hat{\theta}_{i,l}(m) - \beta_{i,l}\Big)^2 + 2\mathbb{E}\Big(\hat{b}_{i,l}^a(m) - \hat{\theta}_{i,l}(m)\Big)\Big(\hat{\theta}_{i,l}(m) - \beta_{i,l}\Big)\Big] \end{aligned} \tag{42}$$

By Eq. (36), we have

$$\lim_{\uparrow+\infty} m\cdot\mathbb{E}\Big(\hat{b}_{i,l}^a(m) - \hat{\theta}_{i,l}(m)\Big)^2 = \Big(\frac{1}{n_{i,l}}\Big)^2\cdot\Big(\sum_{i'\in\mathcal{C}(i,l)}\kappa_{i',l}\Big). \tag{43}$$

By Eq. (41), we have

$$\lim_{m\uparrow+\infty} m\cdot\mathbb{E}\Big(\hat{\theta}_{i,l}(m) - \beta_{i,l}\Big)^2 = 0. \tag{44}$$

By Eq. (36) and Eq. (41), we have

$$\begin{aligned}&\lim_{m\uparrow+\infty} m\cdot\left|\mathbb{E}\left(\hat{b}^a_{i,l}(m)-\hat{\theta}_{i,l}(m)\right)\left(\hat{\theta}_{i,l}(m)-\beta_{i,l}\right)\right|\\ \leq&\lim_{m\uparrow+\infty} 2m\cdot\sqrt{\mathbb{E}\left(\hat{b}^a_{i,l}(m)-\hat{\theta}_{i,l}(m)\right)^2\cdot\mathbb{E}\left(\hat{\theta}_{i,l}(m)-\beta_{i,l}\right)^2}\\ =&2\sqrt{\left(\lim_{m\uparrow+\infty} m\cdot\mathbb{E}\left(\hat{b}^a_{i,l}(m)-\hat{\theta}_{i,l}(m)\right)^2\right)\cdot\left(\lim_{m\uparrow+\infty} m\cdot\mathbb{E}\left(\hat{\theta}_{i,l}(m)-\beta_{i,l}\right)^2\right)}\\ =&0,\end{aligned}\tag{45}$$

where the inequality follows from the Cauchy-Schwartz inequality, the first equality from the fact that both limits exist, and the last from Eqs. (43) and (44). Finally, we plug Eqs. (43), (44), and (45) into Eq. (42) to obtain Eq. (35), and this concludes the proof of *Step 1.*

- *Step 2.* For the DAC_α estimator $\hat{\boldsymbol{\beta}}^\alpha$, inequality (10) holds.

We consider the three cases defined in the proof of Proposition 3: (i) $\mathcal{E}_1(m)$ (i.e., data aggregation levels and cluster structures corrected identified by DAC_α), (ii) $\mathcal{E}_2(m)$ (Type-I error made but no Type-II error made by DAC_α), and (iii) $\mathcal{E}_3(m)$ (Type-II error made by DAC_α).

Since $\mathbb{P}\Big[\mathcal{E}_1(m)\cup\mathcal{E}_2(m)\cup\mathcal{E}_3(m)\Big]=1$, we have

$$\begin{aligned}\mathbb{E}\Big[\hat{\beta}^\alpha_{i,l}(m)-\beta_{i,l}\Big]^2\leq&\mathbb{E}\Big[\left(\hat{\beta}^\alpha_{i,l}(m)-\beta_{i,l}\right)\mathbf{1}_{\mathcal{E}_1(m)}\Big]^2+\mathbb{E}\Big[\left(\hat{\beta}^\alpha_{i,l}(m)-\beta_{i,l}\right)\mathbf{1}_{\mathcal{E}_2(m)}\Big]^2+\mathbb{E}\Big[\left(\hat{\beta}^\alpha_{i,l}(m)-\beta_{i,l}\right)\mathbf{1}_{\mathcal{E}_3(m)}\Big]^2\\ =&\sum_{j=1}^{3}\mathbb{E}\Big[\left(\hat{\beta}^\alpha_{i,l}(m)-\beta_{i,l}\right)^2\Big|\mathcal{E}_j(m)\Big]\mathbb{P}\Big[\mathcal{E}_j(m)\Big]\\ \leq&\mathbb{E}\Big[\left(\hat{\beta}^\alpha_{i,l}(m)-\beta_{i,l}\right)^2\Big|\mathcal{E}_1(m)\Big]\mathbb{P}\Big[\mathcal{E}_1(m)\Big]+\mathbb{E}\Big[\left(\hat{\beta}^\alpha_{i,l}(m)-\beta_{i,l}\right)^2\Big|\mathcal{E}_2(m)\Big]\mathbb{P}\Big[\mathcal{E}_2(m)\Big]+4\bar{\beta}^2\mathbb{P}\Big[\mathcal{E}_3(m)\Big],\end{aligned}\tag{46}$$

where the first inequality follows from the union bound, the second from $\left(\hat{\beta}^\alpha_{i,l}(m)-\beta_{i,l}\right)^2\leq 4\bar{\beta}^2$, and the equality from the definition of conditional expectation.

We next bound each of the three terms in Eq. (46). By Proposition 3 (inequality (7) in particular), Eq. (9), Eq. (35) and, $\mathbb{P}\Big[\mathcal{E}_1(m)\Big]+\mathbb{P}\Big[\mathcal{E}_2(m)\Big]\leq 1$, we obtain

$$\begin{aligned}&\lim_{m\uparrow+\infty} m\mathbb{E}\Big[\left(\hat{\beta}^\alpha_{i,l}(m)-\beta_{i,l}\right)^2\Big|\mathcal{E}_1(m)\Big]\mathbb{P}\Big[\mathcal{E}_1(m)\Big]+\lim_{m\uparrow+\infty} m\mathbb{E}\Big[\left(\hat{\beta}^\alpha_{i,l}(m)-\beta_{i,l}\right)^2\Big|\mathcal{E}_2(m)\Big]\mathbb{P}\Big[\mathcal{E}_2(m)\Big]\\ \leq&(1-p(\alpha))\left(\frac{1}{n_{i,l}}\right)^2\cdot\left(\sum_{i'\in\mathcal{C}(i,l)}\kappa_{i',l}\right)+p(\alpha)\kappa_{i,l}.\end{aligned}\tag{47}$$

By invoking inequality (34) in the proof of Proposition 3, we have

$$\lim_{m\uparrow+\infty} 4\bar{\beta}^2 m\mathbb{P}\Big[\mathcal{E}_3(m)\Big]\leq 4\bar{\beta}^2\cdot\lim_{m\uparrow+\infty}\Big[mc_1\exp(-c_2 m)\Big]=0.\tag{48}$$

Plugging inequalities (47) and (48) into inequality (46) implies that inequality (10) holds. This completes the proof of *Step 2*, and, thus, the proof of Proposition 4(b).

Part (c). Since $\left(\frac{1}{n_{i,l}}\right)^2\cdot\Big(\sum_{i'\in\mathcal{C}(i,l)}\kappa_{i',l}\Big)<\kappa_{i,l}$ and $0<p(\alpha)<1$ (by Proposition 3), we have

$$(1-p(\alpha))\left(\frac{1}{n_{i,l}}\right)^2\cdot\Big(\sum_{i'\in\mathcal{C}(i,l)}\kappa_{i',l}\Big)+p(\alpha)\kappa_{i,l}<\kappa_{i,l}.$$

Hence, by Proposition 4(a) and (b), there exists a threshold $\bar{\mathfrak{m}}_{i,l}$ such that

$$m \cdot \mathbb{E}\left[\hat{\beta}^{\alpha}_{i,l}(m) - \beta_{i,l}\right]^2 < m \cdot \mathbb{E}\left[\hat{b}_{i,l}(m) - \beta_{i,l}\right]^2 \text{for all } m \geq \bar{\mathfrak{m}}_{i,l},$$

which implies Eq. (11). This completes the proof of Proposition 4(c). □

Proof of Proposition 5

Part (a). Since $\lambda_{\min}(\Sigma_i) > 0$ for any item i, the matrix $\boldsymbol{X}_i'\boldsymbol{X}_i$ is of full rank (i.e., rank$= d$) when the sample size m is sufficiently large. We denote the MSE of item i with respect to an estimator $\hat{\boldsymbol{\beta}}_i$ as

$$\widehat{\mathcal{MSE}}_i(\hat{\boldsymbol{\beta}}_i) = \frac{\sum_{j=1}^{m}(Y_{i,j} - \boldsymbol{X}_{i,j}\hat{\boldsymbol{\beta}}_i)^2}{m}.$$

Applying Theorem 2.2 and its proof from Rigollet (2015) to the decentralized estimator of item i $\hat{\boldsymbol{b}}_i$ implies that

$$\mathbb{E}\left[\widehat{\mathcal{MSE}}_i(\hat{\boldsymbol{\beta}}_i)\right] \leq \frac{d\sigma^2}{m}. \tag{49}$$

One should also observe the identity that

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathcal{MSE}}_i(\hat{\boldsymbol{b}}_i) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{i,j} - \boldsymbol{X}_{i,j}\hat{\boldsymbol{\beta}}_i)^2}{nm} = \widehat{\mathcal{MSE}}(\hat{\boldsymbol{b}}). \tag{50}$$

Plugging Eq. (49) into the expectation of Eq. (50) yields inequality (12) and, thus, proves Proposition 5(a).

Part (b). We prove this result in different steps.

- *Step 1.* The MSE of the aggregate estimator $\hat{\boldsymbol{b}}^a$ has the following bound:

$$\mathbb{E}\left[\widehat{\mathcal{MSE}}(\hat{\boldsymbol{b}}^a)\right] \leq \frac{4d_x\sigma^2}{nm}. \tag{51}$$

Following the proof of Proposition 2, the aggregate model can be formulated as Eq. (30) with $G(u) = u$. Furthermore, the dimension of the design matrix $\tilde{\boldsymbol{X}}$ is of dimension $n \times m$ by d_x. One can easily check that $\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}$ has rank d_x. Thus, by Theorem 2.2 from Rigollet (2015), Eq. (51) holds. This proves *Step 1*.

- *Step 2.* Under the DAC algorithm, we denote the event that only a Type-I error occurs but a Type-II error does not occur (i.e., the algorithm may only falsely identify two identical coefficients to be different) as $\mathcal{E}_1$ and the total number of coefficients to estimate in Step 7 of Algorithm 1 is denoted as the random variable $\tilde{\mathfrak{d}}_x(\alpha)$. We then have

$$\mathbb{E}\left[\widehat{\mathcal{MSE}}(\hat{\boldsymbol{\beta}}^{\alpha})\middle|\mathcal{E}_1, \tilde{\mathfrak{d}}_x(\alpha) = \mathfrak{d}_x\right] \leq \frac{4\mathfrak{d}_x\sigma^2}{nm}, \text{ for all } \mathfrak{d}_x = d_x, d_x+1, ..., nd. \tag{52}$$

Conditioned on $\mathcal{E}_1$ and $\tilde{\mathfrak{d}}_x(\alpha) = \mathfrak{d}_x$, the DAC_{α} can be viewed as the aggregate estimator in a revised aggregate model. Therefore, by applying Eq. (52) to this model implies that Eq. (52) holds, and this proves *Step 2*.

- *Step 3.* Inequality (13) holds for the DAC_α estimator.

We denote $\mathcal{E}_2$ as the event where a Type-II error occurs (i.e., the DAC algorithm falsely identifies different coefficients to be identical). Hence, $\mathbb{P}[\mathcal{E}_1 \cup \mathcal{E}_2] = 1$. We thus have

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathcal{MSE}}(\hat{\boldsymbol{\beta}}^\alpha)\right] \leq & \mathbb{E}\left[\widehat{\mathcal{MSE}}(\hat{\boldsymbol{\beta}}^\alpha)\cdot \mathbf{1}_{\mathcal{E}_1}\right] + \mathbb{E}\left[\widehat{\mathcal{MSE}}(\hat{\boldsymbol{\beta}}^\alpha)\cdot \mathbf{1}_{\mathcal{E}_2}\right] \\
= & \mathbb{E}\left[\widehat{\mathcal{MSE}}(\hat{\boldsymbol{\beta}}^\alpha)\middle|\mathcal{E}_1\right]\mathbb{P}\left[\mathcal{E}_1\right] + \mathbb{E}\left[\widehat{\mathcal{MSE}}(\hat{\boldsymbol{\beta}}^\alpha)\middle|\mathcal{E}_2\right]\mathbb{P}\left[\mathcal{E}_2\right] \\
= & \sum_{\mathfrak{d}_x = d}^{d_x} \mathbb{E}\left[\widehat{\mathcal{MSE}}(\hat{\boldsymbol{\beta}}^\alpha)\middle|\mathcal{E}_1, \tilde{\mathfrak{d}}_x(\alpha) = \mathfrak{d}_x\right]\mathbb{P}\left[\mathcal{E}_1, \tilde{\mathfrak{d}}_x(\alpha) = \mathfrak{d}_x\right] + \mathbb{E}\left[\widehat{\mathcal{MSE}}(\hat{\boldsymbol{\beta}}^\alpha)\middle|\mathcal{E}_2\right]\mathbb{P}\left[\mathcal{E}_2\right] \\
\leq & \sum_{\mathfrak{d}_x = d}^{d_x} \frac{4\mathfrak{d}_x\sigma^2}{nm}\mathbb{P}\left[\tilde{\mathfrak{d}}_x(\alpha) = \mathfrak{d}_x\right] + 4\bar{\beta}^2 c_1 \exp(-c_2 m) \\
= & \frac{4d_x(\alpha)\sigma^2}{nm} + o(m^{-1}),
\end{aligned} \tag{53}$$

where the first inequality follows from the union bound, the second from inequality Eq. (52) and inequality (34), and the last equality from the definition of $d_x(\alpha)$ and the identity $\lim_{m\uparrow+\infty} m\exp(-c_2 m) = 0$. By Eq. (53), to prove Eq. (13), it suffices to show that $d_x(\alpha) < nd$, which holds by the support of $\tilde{\mathfrak{d}}_x(\alpha)$. This completes the proof of *Step 3*.

- *Step 4.* The function $d_x(\alpha)$ is decreasing in α with $\lim_{\alpha\downarrow 0} d_x(\alpha) = d_x$.

By *Steps 3 and 4* from the proof of Proposition 3(a), Step 2 of Algorithm 1 is less likely to reject $H^l_{1,i}$ for each i and each l with a smaller α, for any sample path of $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$. Thus, for any sample path of $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$, $\tilde{\mathfrak{d}}_x(\alpha)$ is decreasing in α. Hence, $d_x(\alpha) = \mathbb{E}\left[\tilde{\mathfrak{d}}_x(\alpha)\right]$ is decreasing in α as well. Finally, as $\alpha \downarrow 0$, the probability that Step 2 of Algorithm 1 will reject $H^l_{1,i}$ converges to 0, which implies that $\tilde{\mathfrak{d}}_x(\alpha)$ converges to d_x with probability 1. Therefore, $\lim_{\alpha\downarrow 0} d_x(\alpha) = \lim_{\alpha\downarrow 0}\mathbb{E}\left[\tilde{\mathfrak{d}}_x(\alpha)\right] = d_x$, where the second equality follows from the monotone convergence theorem. This completes the proof of *Step 4* and of Proposition 5(b). □

Proof of Proposition 6

As a first step, we adopt the bias-variance decomposition (e.g., Eq. (7.9) in Hastie et al. 2019) to evaluate the generalization error of each item i. For any estimation algorithm π, we have

$$\begin{aligned}
\mathcal{GE}_i(\pi) = & \mathbb{E}\Big[Y_{i,m_i+1} - \sum_{l\in D}\hat{\beta}_{i,l}(\pi, \mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]^2 \\
= & \sigma^2 + \mathbb{E}\Big(\sum_{l\in D}\beta_{i,l}X^l_{i,m_i+1} - \mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\pi, \mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2 \\
& + \mathbb{E}\Big(\sum_{l\in D}\hat{\beta}_{i,l}(\pi, \mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1} - \mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\pi, \mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2,
\end{aligned} \tag{54}$$

where the first term is referred to as the irreducible error, the second term as the bias (which we denote as $\mathbb{B}_i(\pi)$), and the third term as the variance (which we denote as $\mathbb{V}_i(\pi)$). In the following, we will evaluate the bias and variance terms for the Dec and DAC estimators. We also use Agg to denote the aggregate estimator.

Before completing the proof of Proposition 6, we show that Agg is also an OLS estimator. The true data-generating process (of the training set) can be specified as

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}(\mathsf{tr})\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\epsilon}}, \tag{55}$$

where $\tilde{\boldsymbol{Y}}$ is a (τm)-dimensional vector that concatenates $\boldsymbol{Y}_i$ for all items $i = 1, 2, ..., n$, $\tilde{\boldsymbol{\epsilon}}$ is a (τm)-dimensional error vector that concatenates $\boldsymbol{\epsilon}_i$ for all items, and $\tilde{\boldsymbol{X}}(\mathsf{tr})$ is the $(\tau m) \times d_x$-dimensional feature matrix defined as follows:

$$\tilde{\boldsymbol{X}}(\mathsf{tr}) := \begin{pmatrix} \boldsymbol{X}_1^s(\mathsf{tr}), & \boldsymbol{X}_1^n(\mathsf{tr}), & \boldsymbol{0}, & ..., & \boldsymbol{0}, & \tilde{\boldsymbol{X}}_1^c \\ \boldsymbol{X}_2^s(\mathsf{tr}), & \boldsymbol{0}, & \boldsymbol{X}_2^n(\mathsf{tr}), & ..., & \boldsymbol{0}, & \tilde{\boldsymbol{X}}_2^c \\ ..., & ..., & ..., & ..., & ..., & ... \\ \boldsymbol{X}_n^s(\mathsf{tr}), & \boldsymbol{0}, & \boldsymbol{0}, & ..., & \boldsymbol{X}_n^n(\mathsf{tr}), & \tilde{\boldsymbol{X}}_n^c \end{pmatrix},$$

where $\tilde{\boldsymbol{X}}_i^c$ ($m_i \times (\sum_{l \in \mathcal{D}_c} k_l)$-dimensional) is the design matrix block of the aggregate model with respect to item i ($i = 1, 2, ..., n$), which is constructed in a similar fashion as the aggregate-level and individual-level features. For conciseness, we do not write out $\tilde{\boldsymbol{X}}_i^c$ in full detail. Thus, $\tilde{\boldsymbol{\beta}}$ is a d_x-dimensional vector where the first d_s entries are the coefficients of the features at the aggregate level, the next nd_n entries are the coefficients of the features at the individual level for each of the n items, and the last $\sum_{l \in \mathcal{D}_c} d_l$ entries are the coefficients of the features at the cluster level. By the model specification in Eq. (55), the coefficient estimates for the aggregate model are given by:

$$\hat{\tilde{\boldsymbol{\beta}}}(\mathsf{Agg}, \mathfrak{D}(\mathsf{tr})) = (\tilde{\boldsymbol{X}}(\mathsf{tr})^T \tilde{\boldsymbol{X}}(\mathsf{tr}))^{-1} \tilde{\boldsymbol{X}}(\mathsf{tr})^T \tilde{\boldsymbol{Y}}(\mathsf{tr}) = \tilde{\boldsymbol{\beta}} + (\tilde{\boldsymbol{X}}(\mathsf{tr})^T \tilde{\boldsymbol{X}}(\mathsf{tr}))^{-1} \boldsymbol{X}(\mathsf{tr})^T \tilde{\boldsymbol{\epsilon}} \in \mathbb{R}^{d_x}, \tag{56}$$

which is essentially an OLS estimator (see also McCullagh and Nelder 2019 for more details).

Part (a). The Dec and Agg estimators are ordinary least squares (OLS). It is a standard result in the statistics and econometrics literature that OLS is an unbiased estimator, that is, $\mathbb{B}_i(\pi) = 0$ for $\pi \in \{\mathsf{Agg}, \mathsf{Dec}\}$ or, equivalently,

$$\sum_{l \in D} \beta_{i,l} X_{i,m_i+1}^l - \mathbb{E}\Big[\sum_{l \in D} \hat{\beta}_{i,l}(\pi) X_{i,m_i+1}^l\Big] = 0, \text{ for } \pi \in \{\mathsf{Agg}, \mathsf{Dec}\}. \tag{57}$$

We next quantify the variance term $\mathbb{V}_i(\pi)$ for item i and estimation algorithm π. For the training data of item i, we use the $m_i \times d$ matrix $\boldsymbol{X}_i(\mathsf{tr}) := (X_{i,j}^l(\mathsf{tr}) : 1 \le l \le d, 1 \le j \le m)$ as the feature matrix, and the m dimensional vector $\boldsymbol{Y}_i(\mathsf{tr}) := (Y_{i,j}(\mathsf{tr}) : 1 \le j \le m)$ as the label. For each item i, the decentralized estimator is given by:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_i(\mathsf{Dec}, \mathfrak{D}_i(\mathsf{tr})) =& (\boldsymbol{X}_i(\mathsf{tr})^T X_i(\mathsf{tr}))^{-1} \boldsymbol{X}_i(\mathsf{tr})^T \boldsymbol{Y}_i(\mathsf{tr}) \\ =& (\boldsymbol{X}_i(\mathsf{tr})^T \boldsymbol{X}_i(\mathsf{tr}))^{-1} \boldsymbol{X}_i(\mathsf{tr})^T (\boldsymbol{X}_i(\mathsf{tr}) \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i) \\ =& \boldsymbol{\beta}_i + (\boldsymbol{X}_i(\mathsf{tr})^T \boldsymbol{X}_i(\mathsf{tr}))^{-1} \boldsymbol{X}_i(\mathsf{tr})^T \boldsymbol{\epsilon}_i, \end{aligned} \tag{58}$$

where the first equality follows from $\boldsymbol{Y}_i(\mathsf{tr}) = \boldsymbol{X}_i(\mathsf{tr})\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$. We are now ready to evaluate the variance of the Dec estimator:

$$\begin{aligned}\mathbb{V}_i(\mathsf{Dec}) =&\mathbb{E}\Big(\sum_{l\in D}\hat{\beta}_i(\mathsf{Dec},\mathfrak{D}_i(\mathsf{tr}))X^l_{i,m_i+1} - \mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\mathsf{Dec},\mathfrak{D}_i(\mathsf{tr}))X^l_{i,m_i+1}\Big]\Big)^2\\ =&\mathbb{E}\Big[(\boldsymbol{X}^T_{i,m_i+1}(\boldsymbol{X}_i(\mathsf{tr})^T\boldsymbol{X}_i(\mathsf{tr}))^{-1}\boldsymbol{X}_i(\mathsf{tr})^T)\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T(\boldsymbol{X}^T_{i,m_i+1}(\boldsymbol{X}_i(\mathsf{tr})^T\boldsymbol{X}_i(\mathsf{tr}))^{-1}\boldsymbol{X}_i(\mathsf{tr})^T)^T\Big]\\ =&\mathbb{E}\Big[\mathbb{E}_{\boldsymbol{\epsilon}_i}\Big[(\boldsymbol{X}^T_{i,m_i+1}(\boldsymbol{X}_i(\mathsf{tr})^T\boldsymbol{X}_i(\mathsf{tr}))^{-1}\boldsymbol{X}_i(\mathsf{tr})^T)\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T(\boldsymbol{X}^T_{i,m_i+1}(\boldsymbol{X}_i(\mathsf{tr})^T\boldsymbol{X}_i(\mathsf{tr}))^{-1}\boldsymbol{X}_i(\mathsf{tr})^T)^T\Big|\boldsymbol{X}_{i,m_i+1},\boldsymbol{X}_i(\mathsf{tr})\Big]\Big]\\ =&\sigma^2\mathbb{E}\Big[(\boldsymbol{X}^T_{i,m_i+1}(\boldsymbol{X}_i(\mathsf{tr})^T\boldsymbol{X}_i(\mathsf{tr}))^{-1}\boldsymbol{X}_i(\mathsf{tr})^T)(\boldsymbol{X}^T_{i,m_i+1}(\boldsymbol{X}_i(\mathsf{tr})^T\boldsymbol{X}_i(\mathsf{tr}))^{-1}\boldsymbol{X}_i(\mathsf{tr})^T)^T\Big]\\ =&\sigma^2\mathbb{E}\Big[\boldsymbol{X}^T_{i,m_i+1}(\boldsymbol{X}_i(\mathsf{tr})^T\boldsymbol{X}_i(\mathsf{tr}))^{-1}\boldsymbol{X}_{i,m_i+1}\Big]\\ =&\frac{\sigma^2}{m_i}\mathbb{E}\Big[\boldsymbol{X}^T_{i,m_i+1}\Big(\frac{1}{m_i}\boldsymbol{X}_i(\mathsf{tr})^T\boldsymbol{X}_i(\mathsf{tr})\Big)^{-1}\boldsymbol{X}_{i,m_i+1}\Big],\end{aligned}$$

where the second equality follows from Eq. (58), the third from the law of iterated expectations, and the fourth from the fact that $\epsilon_{i,j}$ are *i.i.d.* with mean 0 and variance σ^2. By the strong law of large numbers and the dominated convergence theorem, we have $\lim\limits_{m_i\uparrow+\infty}\frac{1}{m_i}\boldsymbol{X}_i(\mathsf{tr})^T\boldsymbol{X}_i(\mathsf{tr}) = \boldsymbol{Id}_d$ and we can interchange the limit and expectation operators, where $\boldsymbol{Id}_d$ is the identity matrix with dimension d, observing that the features are *i.i.d.* with mean 0 and variance 1. Thus, we have

$$\lim_{m_i\uparrow+\infty} m_i\cdot\mathbb{V}_i(\mathsf{Dec}) = \sigma^2\mathbb{E}\Big[\boldsymbol{X}_i^T\boldsymbol{X}_i\Big] = \sigma^2\cdot d(=\sigma^2\cdot(d_s+d_n+d_c)),$$

where the second equality follows from the fact that $\boldsymbol{X}_{i,m_i+1}$ has d features which are *i.i.d.* with mean 0 and variance 1. For item i, $m_i = m\tau_i$, so we obtain

$$\lim_{m\uparrow+\infty} m\cdot\mathbb{V}_i(\mathsf{Dec}) = \frac{\sigma^2 d}{\tau_i}. \tag{59}$$

Hence, by plugging Eqs. (57) and (59) into Eq. (54), we conclude that Eq. (14) holds. This proves Proposition 6(a).

Part (b). We decompose the proof of this part into several steps.

- *Step 1.* For the aggregate estimator $\hat{\boldsymbol{b}}^a$, the generalized error satisfies

$$\lim_{m\uparrow+\infty} m\cdot\big(\mathcal{GE}_i(\mathsf{Agg})-\sigma^2\big) = \Big(\sum_{l=1}^{d}\frac{1}{\tau(i,l)}\Big)\cdot\sigma^2,\ \text{for } i=1,2,...,n. \tag{60}$$

Without loss of generality, we assume that the first d_s features $\{1,2,...,d_s\}$ are at the aggregate level, the next d_n features $\{d_s+1,d_s+2,...,d_s+d_n\}$ are at the individual level, and the remaining d_c features $\{d_n+d_s+1,d_n+d_s+2,...,d\}$ are at the cluster level. For item i, we denote by $\boldsymbol{X}_i^s(\mathsf{tr}) = (X^l_{1,j}(\mathsf{tr}) : 1\le l\le d_s, 1\le j\le m_i)$ the feature matrix at the aggregate level, $\boldsymbol{X}_i^n(\mathsf{tr}) = (X^l_{i,j}(\mathsf{tr}) : d_s+1\le l\le d_s+d_n, 1\le j\le m_i)$ the feature matrix at the individual level, and $\boldsymbol{X}_i^c(\mathsf{tr}) = (X^l_{i,j}(\mathsf{tr}) : d_s+d_c+1\le l\le d, 1\le j\le m_i)$ the feature matrix at the cluster level.

We next analyze the aggregate model (55). To evaluate $\mathbb{V}_i(\mathsf{Agg})$, we define an auxiliary d_x-dimensional random vector for each item i, $\tilde{\boldsymbol{X}}_{i,m_i+1}$ follows the same distribution as the $\sum_{j=1}^{i-1} m_j + 1$ row of $\tilde{\boldsymbol{X}}(\mathsf{tr})$. We also denote the non-zero entries of $\tilde{\boldsymbol{X}}_{i,m_i+1}$ as $\tilde{\boldsymbol{\mathfrak{X}}}_i = ((\tilde{\boldsymbol{X}}^s_{i,m_i+1})^T, (\tilde{\boldsymbol{X}}^n_{i,m_i+1})^T, (\tilde{\boldsymbol{X}}^c_{i,m_i+1})^T)^T \in \mathbb{R}^d$. We are now ready to compute $\mathbb{V}_i(\mathsf{Agg})$:

$$\begin{aligned}
&\mathbb{V}_i(\mathsf{Agg}) = \mathbb{E}\Big(\sum_{l=1}^{d_x} \hat{\tilde{\beta}}_i(\mathsf{Agg}, \mathfrak{D}_i(\mathsf{tr}))\tilde{X}^l_{i,m_i+1} - \mathbb{E}\Big[\sum_{l=1}^{d_x} \hat{\tilde{\beta}}_{i,l}(\mathsf{Agg}, \mathfrak{D}_i(\mathsf{tr}))\tilde{X}^l_{i,m_i+1}\Big]\Big)^2 \\
=&\mathbb{E}\Big[(\tilde{\boldsymbol{X}}^T_{i,m_i+1}(\tilde{\boldsymbol{X}}_i(\mathsf{tr})^T\tilde{\boldsymbol{X}}_i(\mathsf{tr}))^{-1}\tilde{\boldsymbol{X}}_i(\mathsf{tr})^T)\tilde{\boldsymbol{\epsilon}}_i\tilde{\boldsymbol{\epsilon}}_i^T(\tilde{\boldsymbol{X}}^T_{i,m_i+1}(\tilde{\boldsymbol{X}}_i(\mathsf{tr})^T\tilde{\boldsymbol{X}}_i(\mathsf{tr}))^{-1}\tilde{\boldsymbol{X}}_i(\mathsf{tr})^T)^T\Big] \\
=&\mathbb{E}\Big[\mathbb{E}_{\tilde{\boldsymbol{\epsilon}}_i}\Big[(\tilde{\boldsymbol{X}}^T_{i,m_i+1}(\tilde{\boldsymbol{X}}_i(\mathsf{tr})^T\tilde{\boldsymbol{X}}_i(\mathsf{tr}))^{-1}\tilde{\boldsymbol{X}}_i(\mathsf{tr})^T)\tilde{\boldsymbol{\epsilon}}_i\tilde{\boldsymbol{\epsilon}}_i^T(\tilde{\boldsymbol{X}}^T_{i,m_i+1}(\tilde{\boldsymbol{X}}_i(\mathsf{tr})^T\tilde{\boldsymbol{X}}_i(\mathsf{tr}))^{-1}\tilde{\boldsymbol{X}}_i(\mathsf{tr})^T)^T\Big|\tilde{\boldsymbol{X}}_{i,m_i+1}, \tilde{\boldsymbol{X}}_i(\mathsf{tr})\Big]\Big] \\
=&\sigma^2\mathbb{E}\Big[\tilde{\boldsymbol{X}}^T_{i,m_i+1}(\tilde{\boldsymbol{X}}_i(\mathsf{tr})^T\tilde{\boldsymbol{X}}_i(\mathsf{tr}))^{-1}\tilde{\boldsymbol{X}}_{i,m_i+1}\Big] \\
=&\frac{\sigma^2}{\tau m}\mathbb{E}\Big[\tilde{\boldsymbol{\mathfrak{X}}}_i^T\Big(\frac{1}{\tau m}\widehat{\boldsymbol{\mathfrak{M}}}_i(\mathsf{tr})\Big)^{-1}\tilde{\boldsymbol{\mathfrak{X}}}_i\Big],
\end{aligned} \tag{61}$$

where the second equality follows from Eq. (56), the third from the law of iterated expectations, and the fourth from the fact that $\epsilon_{i,j}$ are *i.i.d.* with mean 0 and variance σ^2. We now compute the matrix $\widehat{\boldsymbol{\mathfrak{M}}}_i(\mathsf{tr})$. By the strong law of large numbers, $\widehat{\boldsymbol{\mathfrak{Z}}}_i := \lim_{m\uparrow+\infty} \frac{1}{\tau m}\widehat{\boldsymbol{\mathfrak{M}}}_i(\mathsf{tr})$ is a diagonal matrix with the following properties: (i) if $l \in \mathcal{D}_s$, then

$$(\widehat{\boldsymbol{\mathfrak{Z}}}_i)_{l,l} = \lim_{m\uparrow+\infty} \frac{1}{\tau m}\sum_{i'=1}^{n}(\boldsymbol{X}^l_{i'}(\mathsf{tr}))^T\boldsymbol{X}^l_{i'}(\mathsf{tr}) = 1; \tag{62}$$

(ii) if $l \in \mathcal{D}_n$, then

$$(\widehat{\boldsymbol{\mathfrak{Z}}}_i)_{l,l} = \frac{\tau_i}{\tau}\cdot \lim_{m\uparrow+\infty}\frac{1}{\tau_i m}(\boldsymbol{X}^l_i(\mathsf{tr}))^T\boldsymbol{X}^l_i(\mathsf{tr}) = \frac{\tau_i}{\tau}; \tag{63}$$

(iii) if $l \in \mathcal{D}_c$, then

$$(\widehat{\boldsymbol{\mathfrak{Z}}}_i)_{l,l} = \frac{\tau(i,l)}{\tau}\cdot \lim_{m\uparrow+\infty}\frac{1}{\tau(i,l)m}\sum_{i'\in\mathcal{C}(i,l)}(\boldsymbol{X}^l_{i'}(\mathsf{tr}))^T\boldsymbol{X}^l_{i'}(\mathsf{tr}) = \frac{\tau(i,l)}{\tau}. \tag{64}$$

Hence, $(\widehat{\boldsymbol{\mathfrak{Z}}}_i)^{-1}$ is also diagonal with $((\widehat{\boldsymbol{\mathfrak{Z}}}_i)^{-1})_{l,l} = ((\widehat{\boldsymbol{\mathfrak{Z}}}_i)_{l,l})^{-1}$. Therefore, by the dominated convergence theorem, we plug Eqs. (62), (63), and (64) into Eq. (61) and interchange the limit and expectation operators to obtain

$$\lim_{m\uparrow+\infty} m\cdot\mathbb{V}_i(\mathsf{Agg}) = \sigma^2\cdot\left(\frac{d_s}{\tau}+\frac{d_n}{\tau_i}+\sum_{l\in\mathcal{D}_c}\frac{1}{\tau(i,l)}\right) = \sigma^2\cdot\left(\sum_{l=1}^{d}\frac{1}{\tau(i,l)}\right). \tag{65}$$

Thus, by plugging Eqs. (57) and (65) into Eq. (54), we conclude that Eq. (60) holds, and this proves *Step 1.*

- *Step 2.* The generalization error of the DAC estimator satisfies Eq. (15).

We consider the three cases defined in the proof of Propositions 3 and 4: (i) $\mathcal{E}_1$ (i.e., data aggregation levels and cluster structures are correctly identified by DAC_α), (ii) $\mathcal{E}_2$ (Type-I error made but no Type-II error made by DAC_α), and (iii) $\mathcal{E}_3$ (Type-II error made by DAC_α). We define $\bar{\beta}$ as the maximum possible value of the coefficients for all items and all features.

We quantify $\mathbb{B}_i(\mathsf{DAC}_\alpha)$ and $\mathbb{V}_i(\mathsf{DAC}_\alpha)$ separately. Since $\mathbb{P}\Big[\mathcal{E}_1\cup\mathcal{E}_2\cup\mathcal{E}_3\Big]=1$, we have

$$
\begin{aligned}
\mathbb{B}_i(\mathsf{DAC}_\alpha) \le & \mathbb{E}\Big[\Big(\sum_{l\in D}\beta_{i,l}X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\mathsf{DAC}_\alpha,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\mathbf{1}_{\mathcal{E}_1}\Big]\\
&+\mathbb{E}\Big[\Big(\sum_{l\in D}\beta_{i,l}X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\mathsf{DAC}_\alpha,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\mathbf{1}_{\mathcal{E}_2}\Big]\\
&+\mathbb{E}\Big[\Big(\sum_{l\in D}\beta_{i,l}X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\mathsf{DAC}_\alpha,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\mathbf{1}_{\mathcal{E}_3}\Big]\\
&=\sum_{j=1}^{3}\mathbb{E}\Big[\Big(\sum_{l\in D}\beta_{i,l}X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\mathsf{DAC}_\alpha,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\Big|\mathcal{E}_j\Big]\cdot\mathbb{P}\Big[\mathcal{E}_j\Big]\\
&\le 4d\bar{\beta}^2\mathbb{P}\Big[\mathcal{E}_3\Big],
\end{aligned}
\tag{66}
$$

where the first inequality follows from the union bound, the second from the fact that OLS is unbiased for a correctly specified model and the definition of $\bar{\beta}$, and the equality from the definition of conditional expectation. Therefore, by inequality (34) in the proof of Proposition 3 and inequality (66) , we have

$$
\lim_{m\uparrow+\infty}\mathbb{B}_i(\mathsf{DAC}_\alpha)\le\lim_{m\uparrow+\infty}4d\bar{\beta}^2m\mathbb{P}\left[\mathcal{E}_3\right]\le 4d\bar{\beta}^2\cdot\lim_{m\uparrow+\infty}\Big[mc_1\exp(-c_2m)\Big]=0. \tag{67}
$$

We next evaluate $\mathbb{V}_i(\mathsf{DAC}_\alpha)$ using a similar strategy:

$$
\begin{aligned}
\mathbb{V}_i(\mathsf{DAC}_\alpha) \le & \mathbb{E}\Big[\Big(\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\mathbf{1}_{\mathcal{E}_1}\Big]\\
&+\mathbb{E}\Big[\Big(\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\mathbf{1}_{\mathcal{E}_2}\Big]\\
&+\mathbb{E}\Big[\Big(\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\mathbf{1}_{\mathcal{E}_3}\Big]\\
&=\sum_{j=1}^{3}\mathbb{E}\Big[\Big(\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\Big|\mathcal{E}_j\Big]\cdot\mathbb{P}\Big[\mathcal{E}_j\Big]\\
&\le\mathbb{E}\Big[\Big(\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\Big|\mathcal{E}_1\Big]\cdot\mathbb{P}\Big[\mathcal{E}_1\Big]\\
&+\mathbb{E}\Big[\Big(\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\Big|\mathcal{E}_2\Big]\cdot\mathbb{P}\Big[\mathcal{E}_2\Big]\\
&+4d\bar{\beta}^2\mathbb{P}\Big[\mathcal{E}_3\Big],
\end{aligned}
\tag{68}
$$

where the first inequality follows from the union bound, the second from the definition of $\bar{\beta}$, and the equality from the definition of conditional expectation.

We now bound each of the three terms in Eq. (68). By Proposition 3 (inequality (7) in particular), Eqs. (59), (65) and, $\mathbb{P}\left[\mathcal{E}_1\right]+\mathbb{P}\left[\mathcal{E}_2\right]\le 1$, we have

$$
\begin{aligned}
&\lim_{m\uparrow+\infty}m\cdot\mathbb{E}\Big[\Big(\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\Big|\mathcal{E}_1\Big]\cdot\mathbb{P}\Big[\mathcal{E}_1\Big]\\
&+\lim_{m\uparrow+\infty}m\cdot\mathbb{E}\Big[\Big(\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}-\mathbb{E}\Big[\sum_{l\in D}\hat{\beta}_{i,l}(\pi,\mathfrak{D}_{\mathsf{tr}})X^l_{i,m_i+1}\Big]\Big)^2\Big|\mathcal{E}_2\Big]\cdot\mathbb{P}\Big[\mathcal{E}_2\Big]\\
&\le p(\alpha)\cdot\frac{d\cdot\sigma^2}{\tau_i}+(1-p(\alpha))\cdot\Big(\sum_{l=1}^{d}\frac{1}{\tau(i,l)}\Big)\cdot\sigma^2.
\end{aligned}
\tag{69}
$$

Plugging inequalities (69) and (67) into inequality (68) implies that

$$\lim_{m\uparrow+\infty} m\cdot\mathbb{V}_i(\mathsf{DAC}_\alpha)\le p(\alpha)\cdot\frac{d\cdot\sigma^2}{\tau_i}+(1-p(\alpha))\cdot\left(\sum_{l=1}^{d}\frac{1}{\tau(i,l)}\right)\cdot\sigma^2. \tag{70}$$

By combining inequalities (67) and (70) with the bias-variance decomposition from Eq. (54), we conclude that inequality (15) holds. This completes the proof of *Step 2*, and, thus, the proof of Proposition 6(b).

Part (c). By subtracting Eq. (14) from Eq. (15), we have

$$\lim_{m\uparrow+\infty} m\cdot\Big(\mathcal{GE}_i(\mathsf{Dec})-\mathcal{GE}_i(\mathsf{DAC}_\alpha)\Big)>(1-p(\alpha))\cdot\left(d_s\left(\frac{1}{\tau_i}-\frac{1}{\tau}\right)+\sum_{l\in\mathcal{D}_c}\left(\frac{1}{\tau_i}-\frac{1}{\tau(i,l)}\right)\right)\cdot\sigma^2,$$

where the inequality follows from the fact that, if a Type-I error occurs, it may still identify some (if not all) of the identical coefficients. Since $0<p(\alpha)<1$ (by Proposition 3), $\tau>\tau_i$, and $\tau(i,l)\ge\tau_i$ for each i and l, for a sufficiently large m, we have

$$m\cdot\Big(\mathcal{GE}_i(\mathsf{Dec})-\mathcal{GE}_i(\mathsf{DAC}_\alpha)\Big)>(1-p(\alpha))\cdot\left(d_s\left(\frac{1}{\tau_i}-\frac{1}{\tau}\right)+\sum_{l\in\mathcal{D}_c}\left(\frac{1}{\tau_i}-\frac{1}{\tau(i,l)}\right)\right)\cdot\sigma^2>0,$$

namely, Eq. (16) holds. Finally, by taking the first- and second-order partial derivatives of $g_i(\boldsymbol{\tau})$ we obtain, for $i'\ne i$,

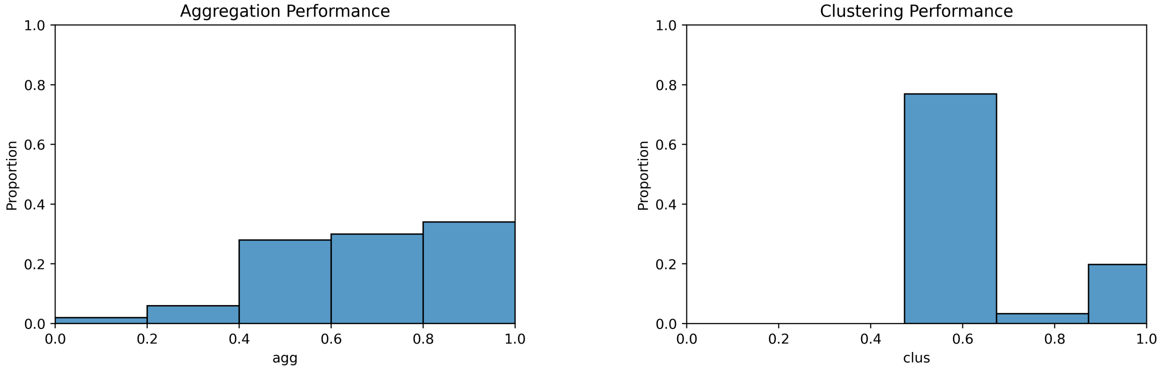
$$\frac{\partial g_i(\boldsymbol{\tau})}{\partial\tau_i}<0,\ \ \frac{\partial g_i(\boldsymbol{\tau})}{\partial\tau_{i'}}>0,\ \ \frac{\partial^2 g_i(\boldsymbol{\tau})}{\partial_2\tau_i}>0,\ \text{and}\ \ \frac{\partial^2 g_i(\boldsymbol{\tau})}{\partial_2\tau_{i'}}<0.$$

This suggests that $g_i(\boldsymbol{\tau})$ is convexly decreasing in τ_i and concavely increasing in $\tau_{i'}$. We have thus completed the proof of Proposition 6(c). □

Appendix D: Additional Plots for Section 5

In this section, we report the simulation results to evaluate the performance of the DAC algorithm to identify the data aggregation levels of the features and recover the cluster structure of the items with respect to cluster-level features. The simulation details are presented in Section 5.1.

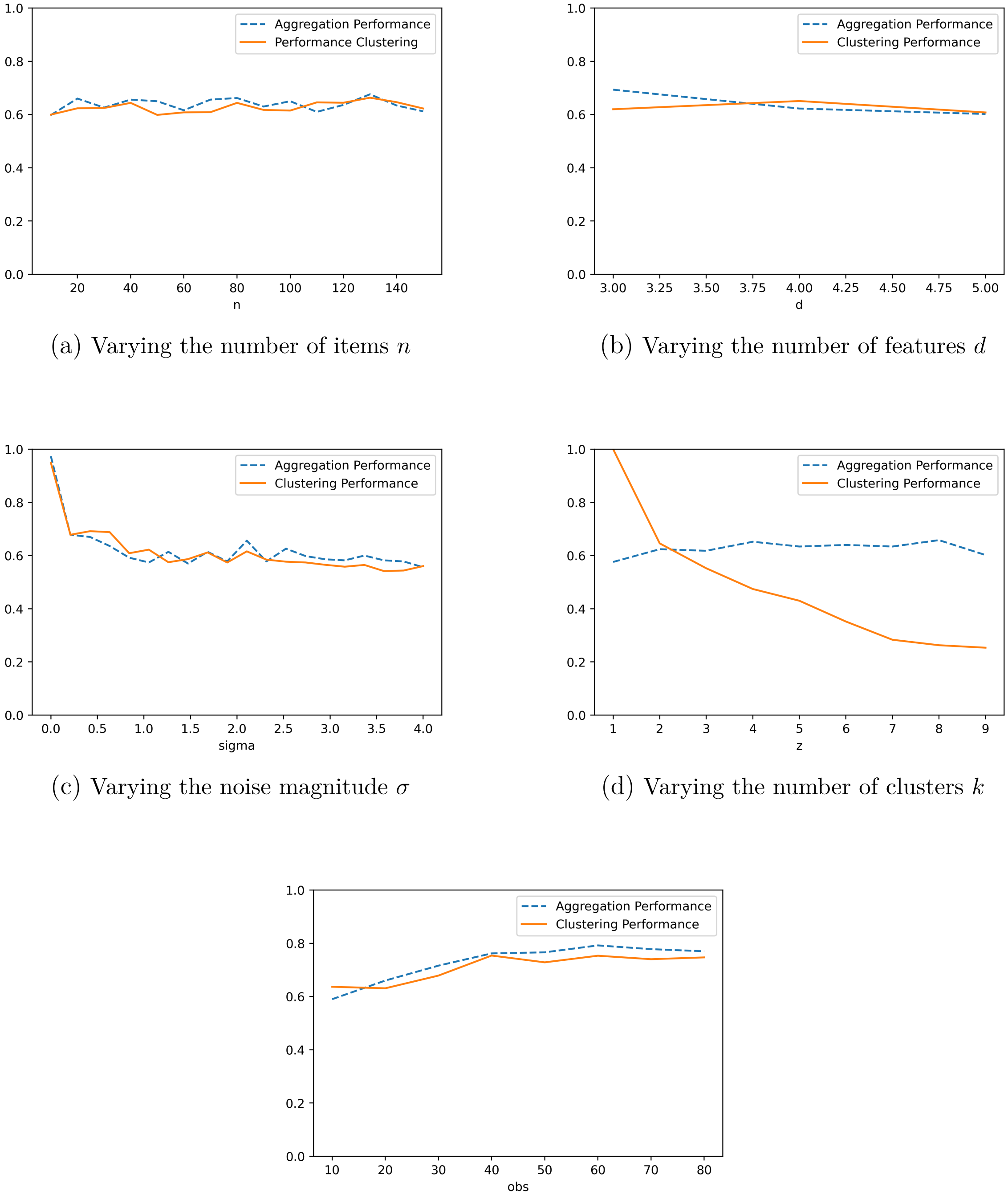
In Figure 7(a), we compute the aggregation performance (i.e., the capability to correctly identify the data aggregation levels of the features), whereas in Figure 7(b), we compute the clustering performance (i.e., the capability to accurately recover the cluster structure of the items with respect to cluster-level features). As discussed in Section 5.1, the aggregation performance is captured by the *Accuracy* metric and the clustering performance by the *Rand index* (more details about these two metrics including the formulas can be found in Section 5.1). Figure 7 reports the values of these two metrics (in the form of a histogram) for the instances presented in Table 1. Specifically, the y-axis represents the proportion and the x-axis reports the value obtained for each instance. Both the Accuracy and the Rand index metrics are evaluated at the problem instance level. More specifically, for each problem instance specified in Table 1, we evaluate the *Accuracy* as the proportion of the d features that our DAC algorithm correctly identifies their aggregation levels (i.e., among the d features, how many of them are correctly classified in terms of their aggregation level). Likewise, for each problem instance specified in Table 1, we evaluate the *Rand index* as the proportion of item-pairs that our DAC algorithm correctly identifies (i.e., whether they are in the same cluster or not). In our simulation, we set the cluster structure invariant with respect to different features.



(a) Identifying aggregation levels

(b) Recovering the cluster Structure

Figure 7 Performance in identifying aggregation levels and recovering the cluster structure.



(a) Varying the number of items n

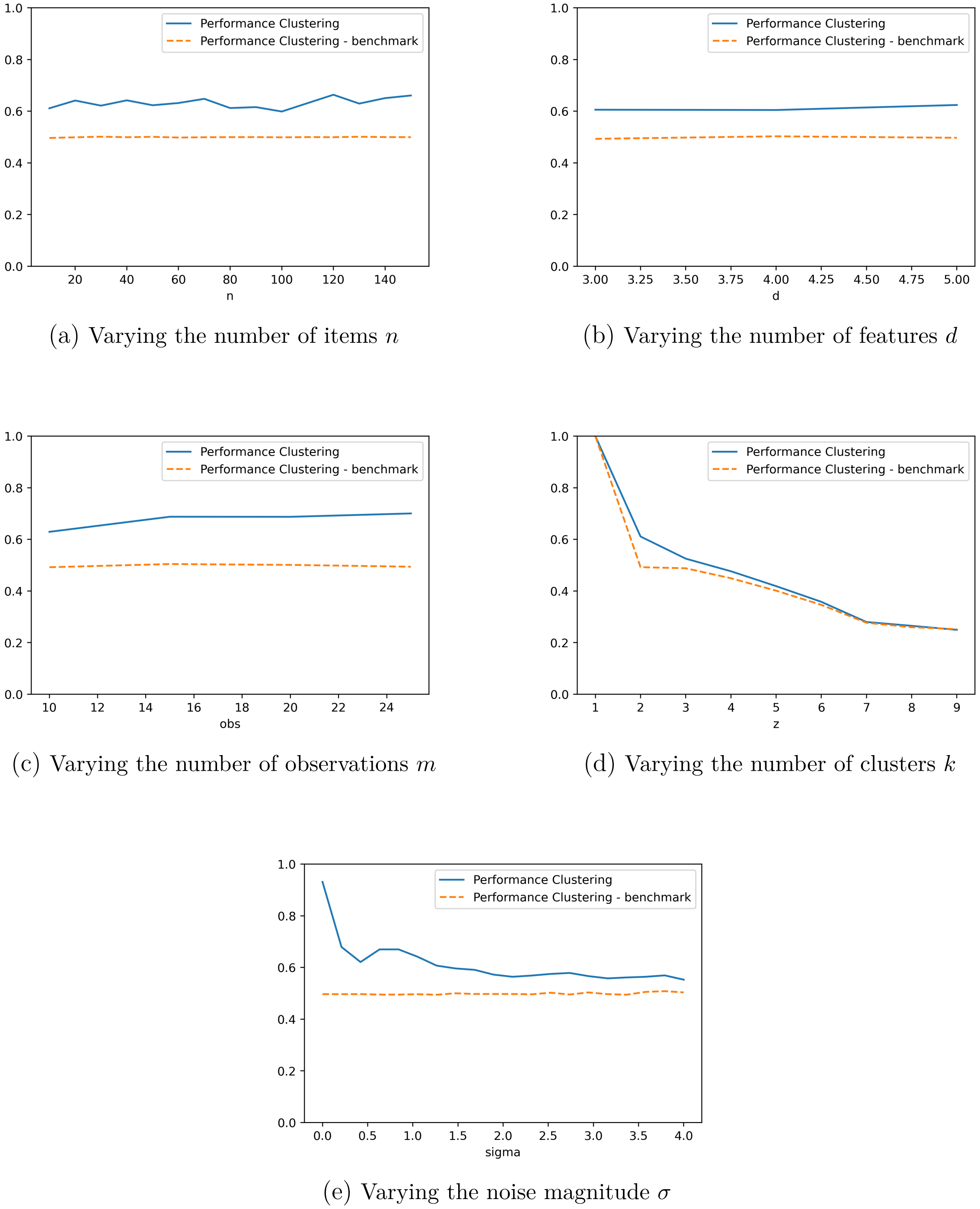
(b) Varying the number of features d

(c) Varying the noise magnitude σ

(d) Varying the number of clusters k

(e) Varying the number of observations m

Figure 8 **Sensitivity analysis on the DAC performance.**



(a) Varying the number of items n

(b) Varying the number of features d

(c) Varying the number of observations m

(d) Varying the number of clusters k

(e) Varying the noise magnitude σ

Figure 9 **Comparing** DAC **with the** k**-means benchmark.**