Supplementary Material to "A likelihood-based approach for multivariate categorical response regression in high dimensions"

Aaron J. Molstad* and Adam J. Rothman†
Department of Statistics and Genetics Institute, University of Florida*
School of Statistics, University of Minnesota†

A Additional bivariate categorical response simulation studies and details

A.1 Alternative tuning parameter selection criterion

In this section, we present simulation study results under exactly the data generating models described in Section 6, but using a different tuning parameter selection criterion for each method. In these studies, we select tuning parameters by maximizing the log-likelihood evaluated on the validation set: for example, see equation (5) of Price et al. (2019). As in the main manuscript, we measure joint misclassification accuracy and average Kullback-Leibler divergence, the latter of which we define as

$$n_{\text{test}}^{-1}\sum_{i=1}^{n_{\text{test}}}\sum_{j=1}^{J}\sum_{k=1}^{K}\log\left(\frac{\hat{P}(Y_{i1}=j,Y_{i2}=k\mid x_i)}{P(Y_{i1}=j,Y_{i2}=k\mid x_i)}\right)\hat{P}(Y_{i1}=j,Y_{i2}=k\mid x_i)$$

where $\hat{P}(Y_{i1}=j,Y_{i2}=k\mid x_i)$ is an estimate of $P(Y_{i1}=j,Y_{i2}=k\mid x_i)$ based on some particular fitted model. We also record and report average test set Hellinger distance, which is defined as

$$\frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}\left(\frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K}\left[\{\hat{P}(Y_{i1}=j,Y_{i2}=k\mid x_i)\}^{1/2}-\{P(Y_{i1}=j,Y_{i2}=k\mid x_i)\}^{1/2}\right]^2\right)^{1/2}.$$

*Correspondence: amolstad@ufl.edu

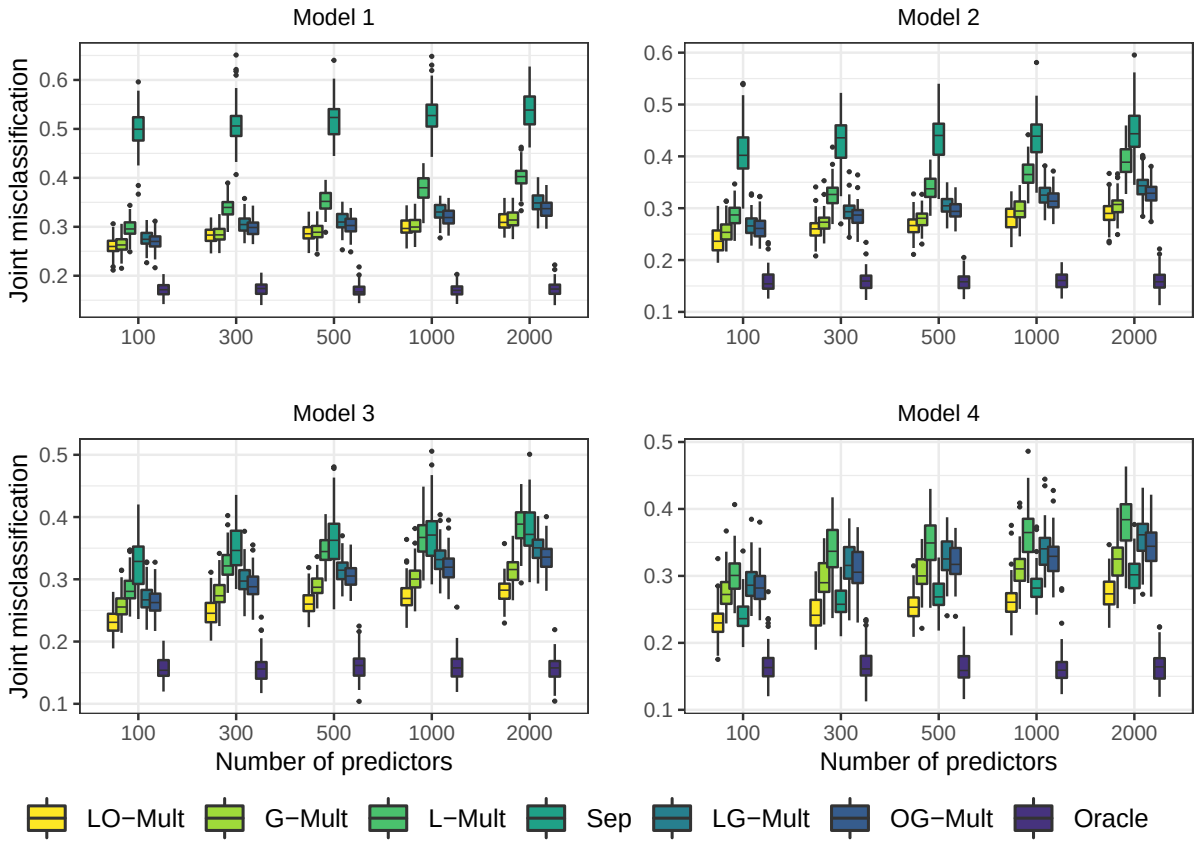

Figure 4: Joint misclassification rates under Models 1–4 with $p \in \{100, 300, 500, 1000, 2000\}$ and tuning parameters chosen to maximize the validation likelihood.

In Figures 4, 5, and 6 we display the joint misclassification rates, average KL divergence, and average Hellinger distance under exactly the data generating models in Section 6, but with tuning parameters chosen to maximize the validation likelihood. As can be seen comparing these results to those from Section 6, the metric used to select tuning parameters does have an effect on the results. While, the relative performances of each methods is essentially unchanged; and the classification accuracy decreases whereas the KL divergence and Hellinger distances are larger than when selecting tuning parameters by minimizing the validation misclassification rate.

A.2 Additional performance metrics and details

In Figure 7 and 8, we display the average test set Hellinger distances and marginal misclassification rates, respectively, under the same data generating models and tuning parameter selection criterion as in Section 6. In Figure 9, we provide a visualization of the groups being penalized by both the overlapping group lasso (`OG-Mult`) and latent group lasso (`LG-Mult`) estimators described in Section 6.

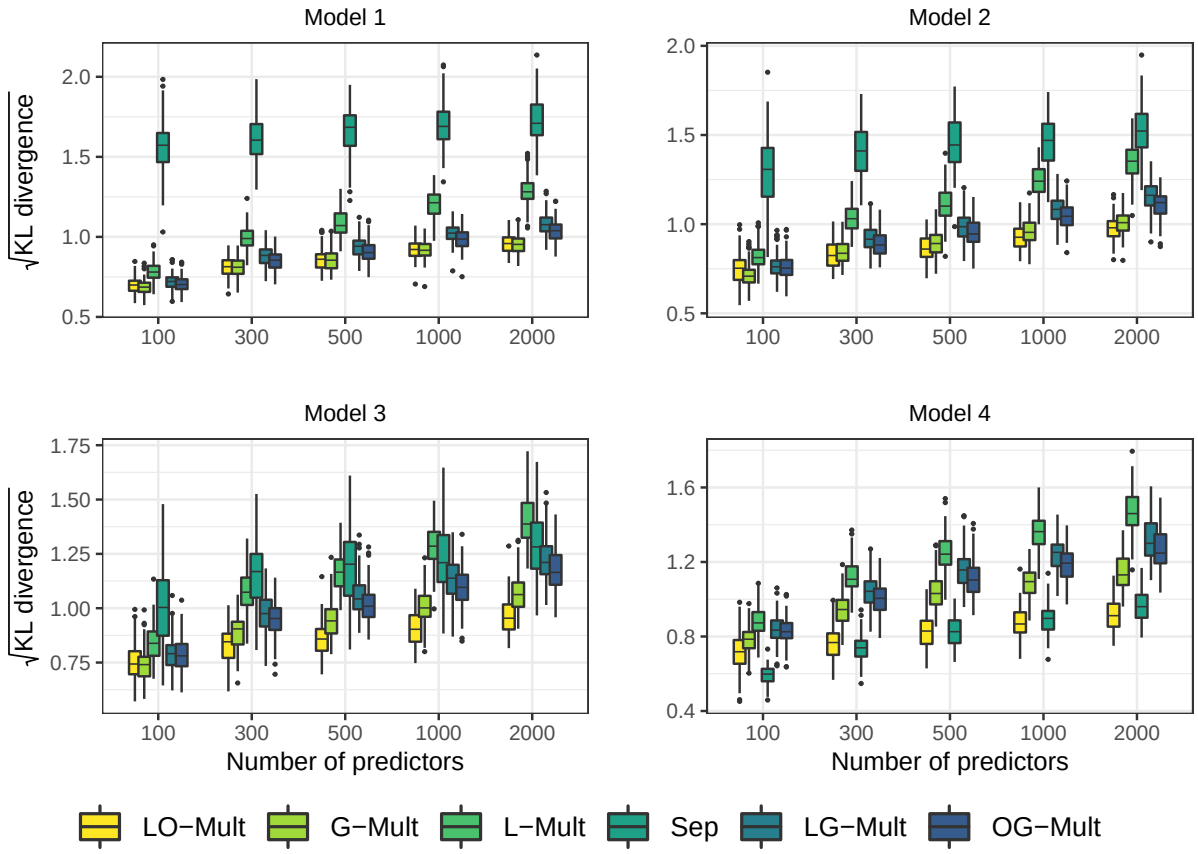

Figure 5: Square-root average Kullback-Leibler diverence under Models 1–4 with $p \in \{100, 300, 500, 1000, 2000\}$ and tuning parameters chosen to maximize the validation likelihood.

B Results with $J = 4$ and $K = 3$

In this section, we present simulation studies essentially identical to those from Section 6, but with $J = 4$ and $K = 3$. The data generating models differ only in how β^* is constructed under Models 2–4. In this setting, we simply find a V such that $V \in \text{Null}(D')$ and set the rows of β_* corresponding to predictors affecting only marginal distributions to be equal to $Vu \in \mathbb{R}^{12}$ where $u \in \mathbb{R}^6$ with each element drawn independently from Uniform$(-3, 3)$. This way, for each $\beta_{j,:} = Vu$, we have that $\beta_{j,:} \neq 0_{12}$, but $D'\beta_{j,:} = 0_{12}$.

Misclassification rates and average KL divergences are displayed in Figures 10 and 11. The performance of the methods relative to one another is quite similar to the settings where $J = 3$ and $K = 2$. In general, each method performs slightly worse, which can be easily explained by the fact that with more response categories, lower classification accuracy (even for the oracle) is expected.

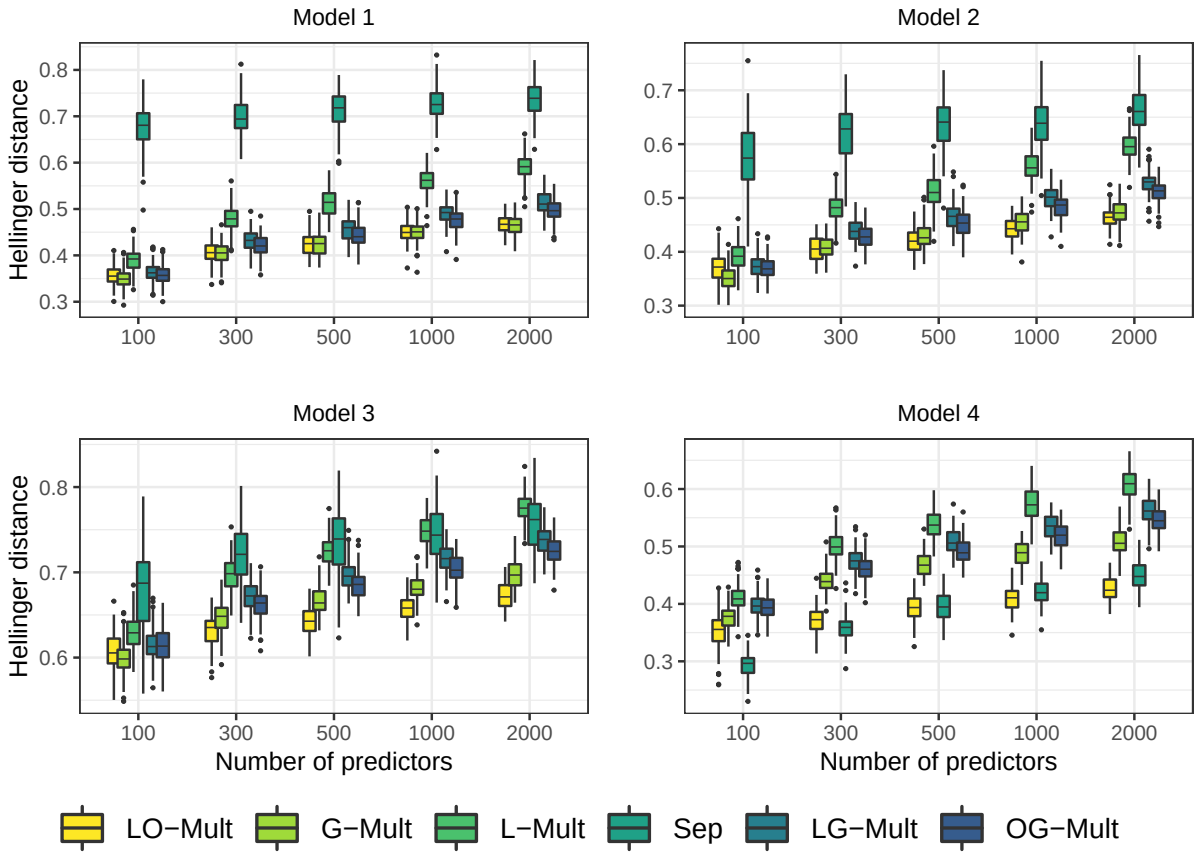

Figure 6: Average Hellinger distance under Models 1–4 with $p \in \{100, 300, 500, 1000, 2000\}$ and tuning parameters chosen to maximize the validation likelihood.

C Trivariate categorical response simulations

In this section, we present results from a simulation study in which we considered a trivariate response. That is, we have three response variables with $J = K = L = 2$ categories each and

$$P(Y_1 = j, Y_2 = k, Y_3 = l \mid x) = \frac{\exp(x'\boldsymbol{\beta}^*_{:,j,k,l})}{\sum_{s=1}^{J}\sum_{t=1}^{K}\sum_{u=1}^{L}\exp(x'\boldsymbol{\beta}^*_{:,s,t,u})}$$

for $(j, k, l) \in \{1, 2\} \times \{1, 2\} \times \{1, 2\}$. Define the matricized version of $\boldsymbol{\beta}^*$ as $\beta^* \in \mathbb{R}^{p \times JKL}$ where $\boldsymbol{\beta}^*_{:,j,k,l} = \beta^*_{:,h(j,k,l)}$ where $h(j, k, l) = (k-1)J + j + (l-1)JK$. We will compare four methods for estimating the mass function of $(Y_1, Y_2, Y_3 \mid x)$: `LO-Mult`, `G-Mult`, `L-Mult`, and `Sep`.

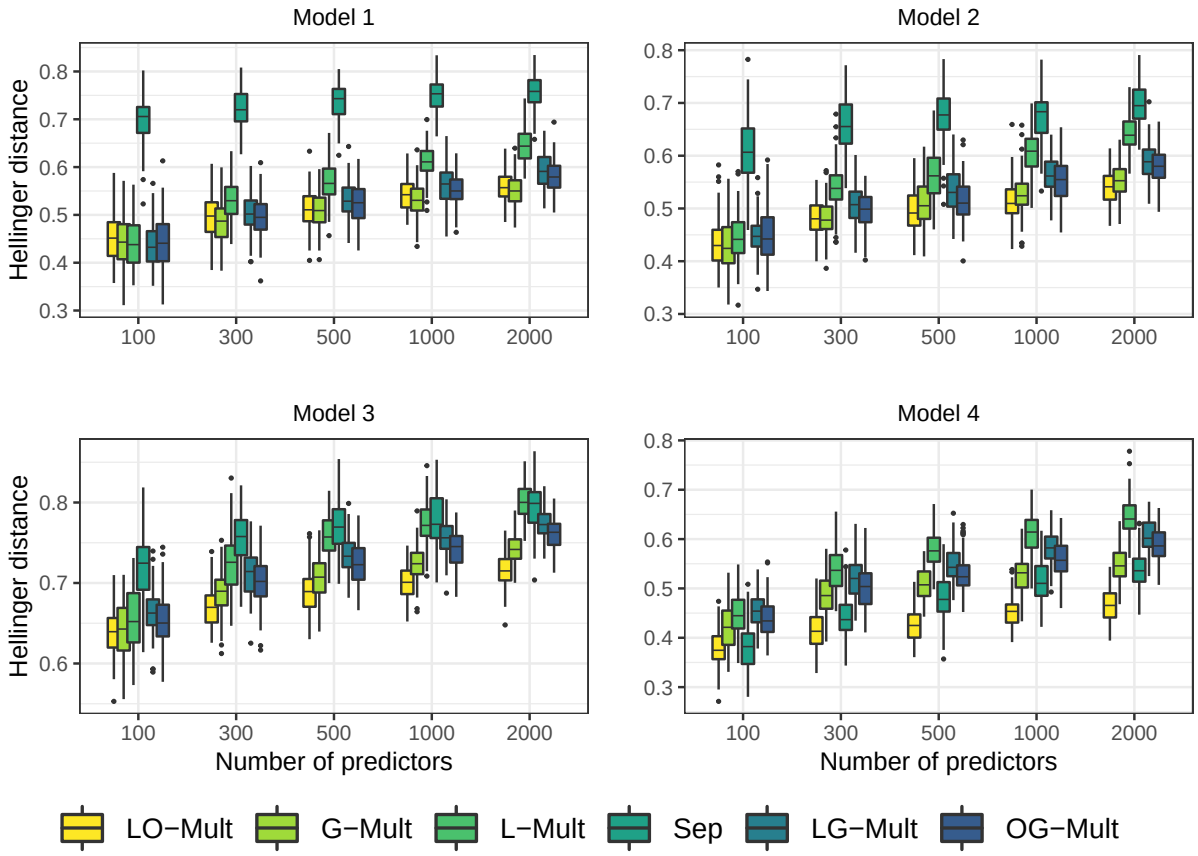

Figure 7: Average Hellinger distance under Models 1–4 with $p \in \{100, 300, 500, 1000, 2000\}$ and tuning parameters chosen as in Section 6.

C.1 Implementation

In order to implement `LO-Mult`, we must first construct D as described in Section 5. Recalling that under the mapping h,

$$\beta = (\boldsymbol{\beta}_{:,1,1,1}, \boldsymbol{\beta}_{:,2,1,1}, \boldsymbol{\beta}_{:,1,2,1}, \boldsymbol{\beta}_{:,2,2,1}, \boldsymbol{\beta}_{:,1,1,2}, \boldsymbol{\beta}_{:,2,1,2}, \boldsymbol{\beta}_{:,1,2,2}, \boldsymbol{\beta}_{:,2,2,2}) \in \mathbb{R}^{p \times JKL},$$

so that we have

$$D' = \begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \\ 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 \end{pmatrix}. \tag{16}$$

Note that this is D matrix is constructed according to the discussion on Section 5.1. To apply Algorithm 1 to the trivariate setting, we need only consider how to solve (11) with D

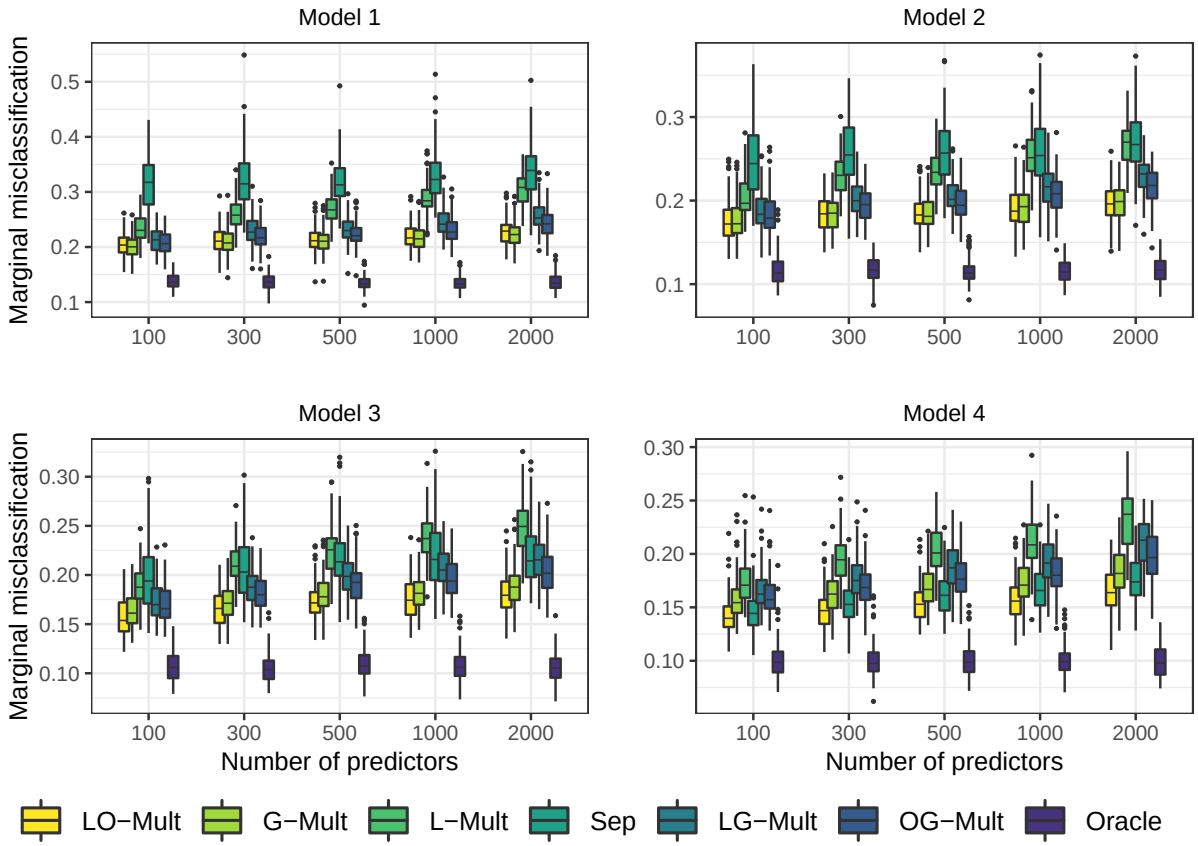

Figure 8: Marginal misclassification rates (for the J-category response variable) under Models 1–4 with $p \in \{100, 300, 500, 1000, 2000\}$.

$$\begin{pmatrix} \beta_{j,1,1} & \beta_{j,1,2} & \beta_{j,1,3} \\ \beta_{j,2,1} & \beta_{j,2,2} & \beta_{j,2,3} \end{pmatrix} \quad \begin{pmatrix} \beta_{j,1,1} & \beta_{j,1,2} & \beta_{j,1,3} \\ \beta_{j,2,1} & \beta_{j,2,2} & \beta_{j,2,3} \end{pmatrix} \quad \begin{pmatrix} \beta_{j,1,1} & \beta_{j,1,2} & \beta_{j,1,3} \\ \beta_{j,2,1} & \beta_{j,2,2} & \beta_{j,2,3} \end{pmatrix}$$

$$\begin{pmatrix} \beta_{j,1,1} & \beta_{j,1,2} & \beta_{j,1,3} \\ \beta_{j,2,1} & \beta_{j,2,2} & \beta_{j,2,3} \end{pmatrix} \quad \begin{pmatrix} \beta_{j,1,1} & \beta_{j,1,2} & \beta_{j,1,3} \\ \beta_{j,2,1} & \beta_{j,2,2} & \beta_{j,2,3} \end{pmatrix}$$

Figure 9: The groups of parameters which are penalized by both the overlapping and latent group-penalized multivariate multinomial estimators in (14) and (15) with $J = 2$ and $K = 3$ for $j = 2, \ldots, p$.

as defined above. For this purpose, we can straightforwardly apply Theorem 2; however, the closed form solution for (iii) in Proposition 1 no longer holds. In this setting, to obtain a τ which satisfies Theorem 2 (iii), we resort to a numeric root-solver to find τ. Note that the D in (16) has 4 non-zero singular values: their values are $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (\sqrt{12}, 2, 2, 2)$. Hence, by the same logic as in the proof of Proposition 1, letting $w_l = u_l'\nu$ (where u_l is the lth left

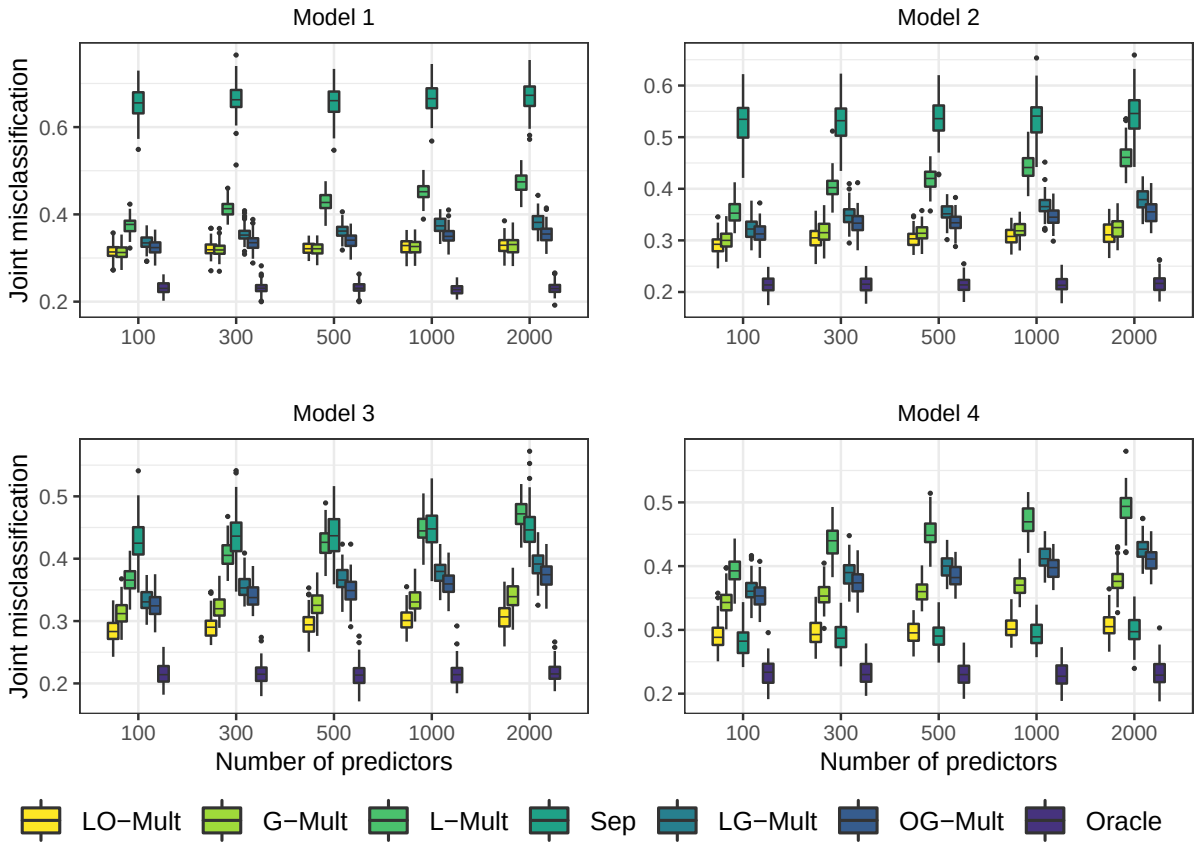

Figure 10: Joint misclassification rates under Models 1–4 with $p \in \{100, 300, 500, 1000, 2000\}$ with $J = 4$ and $K = 3$.

singular vector of D), we need τ such that

$$\sum_{l=1}^{\text{rank}(D)} \frac{w_l^2 \sigma_l^2}{(\sigma_l^2 + \tau)^2} = \lambda^2 \implies 12\frac{w_1^2}{(12+\tau)^2} + 4\sum_{l=2}^{4} \frac{w_l^2}{(4+\tau)^2} - \lambda^2 = 0.$$

Under the conditions of Theorem 2 (iii), such a $\tau > 0$ always exists and can be found using a numeric root-solver in R, e.g., `rootSolve`. For problems with moderately sized J, K, and L, this is can be done with reasonable efficiency.

C.2 Data generating models

To compare the various methods in the trivariate categorical response setting, we consider four data generating models similar to those from Section 6. Just as in Section B, we first obtain $V \in \text{Null}(D')$ for the D defined in (16). Then, we consider Models 5–8.

– **Model 5**: We randomly select 10 rows of $\beta^* \in \mathbb{R}^{p \times JKL}$ to be nonzero. Each of the

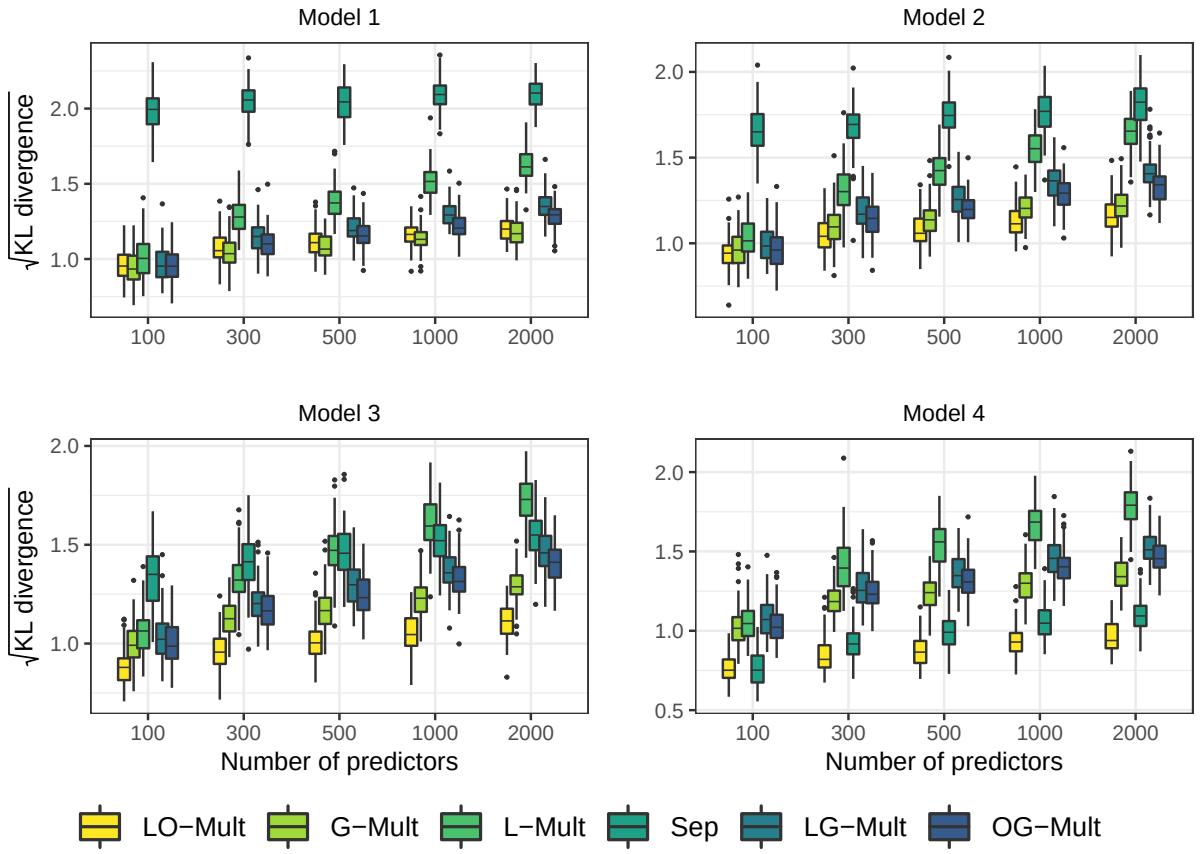

Figure 11: Square-root average Kullback-Leibler divergence under Models 1–4 with $p \in \{100, 300, 500, 1000, 2000\}$ with $J = 4$ and $K = 3$.

elements of these tens rows is set equal to independent realizations of a Uniform$(-3, 3)$ random variable.

– **Model 8**: We randomly select 10 rows of β^* to be nonzero. For each row independently, we generate four independent realizations of a Uniform$(-3, 3)$ random variable. Given these realizations, say (u_1, u_2, u_3, u_4), we set the row of β^* equal to Vu Under this construction, we can see $D'Vu = 0_6$.

Just as in Section 6, Models 6 and 7 are, in a sense, intermediate to Models 6 and 7.

– **Model 6**: We randomly select six rows of β^* to be nonzero and consist elements which are each independent realizations of a Uniform$(-3, 3)$ random variable. Then, we select an additional four rows of β^* to be generated in the same manner as Model 4.

– **Model 7**: We randomly select three rows of β^* to be nonzero and consist elements which are each independent realizations of a Uniform$(-3, 3)$ random variable. Then, we select an additional seven rows of β^* to be generated in the same manner as Model 4.

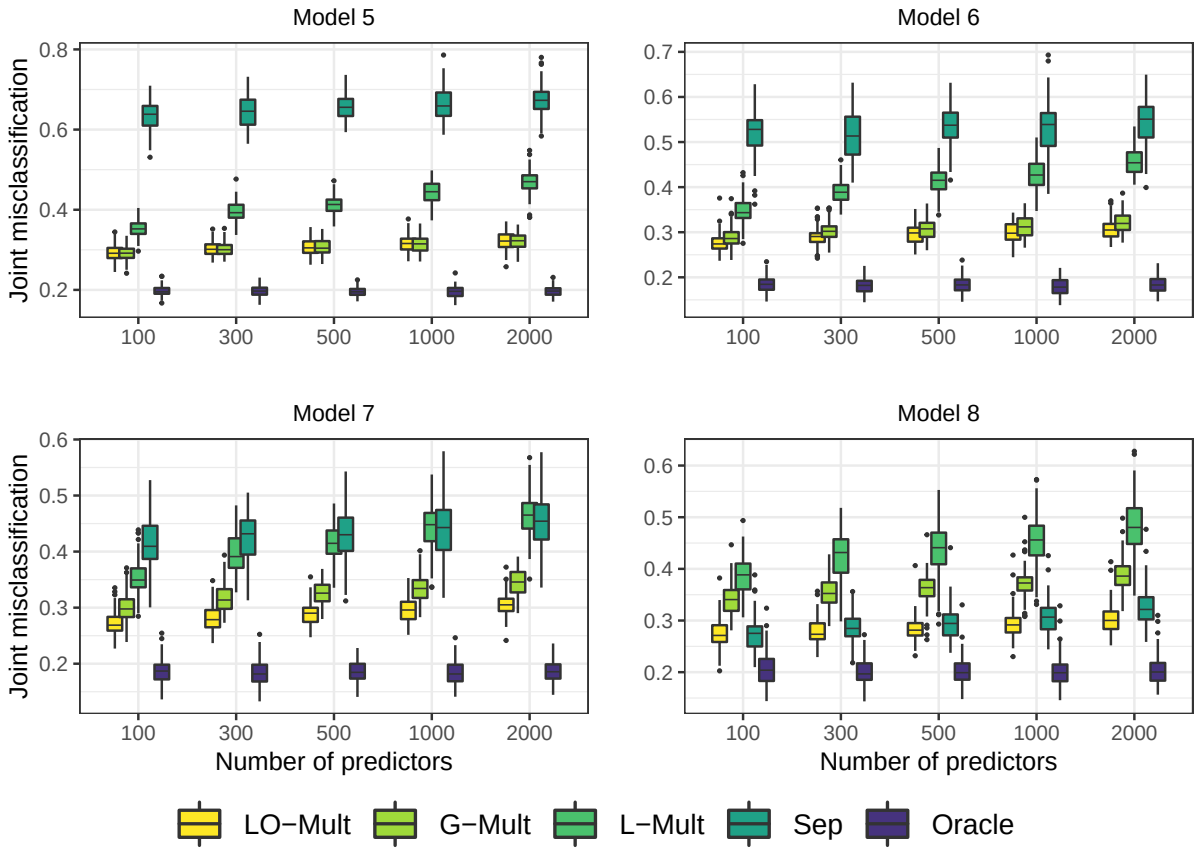

Figure 12: Joint misclassification rates under Models 5–8 with $p \in \{100, 300, 500, 1000, 2000\}$ with $J = K = L = 2$.

As mentioned, in these simulation studies, we only consider the estimators `LO-Mult`, `G-Mult`, `L-Mult`, `Sep`, and when appropriate, `Oracle`.

C.3 Results

In this section, we discuss results under Models 5–8. In Figure 12, we present the joint (i.e., trivariate) misclassification rates for each of the considered methods. Relative performances are essentially the same as in the various bivariate settings considered previously. Under Model 5, `LO-Mult` and `G-Mult` perform similarly – which is to be expected for the same reasons as described in Section 6. As we move from Model 5 to Models 6–8, we see that `LO-Mult` starts to outperform `G-Mult`. Meanwhile, `Sep` begins to perform better as we move from Model 5 towards Model 8: in Model 8, `Sep` – which correctly assumes the responses are independent – performs nearly as well as `LO-Mult`.

In Figure 13 and 14, we display both average Kullback-Leibler divergence and average Hellinger distances for the various methods. Just as with classification accuracy, performances largely agree with the bivariate setting. Of particular note is that as p grows, `LO-Mult` tends

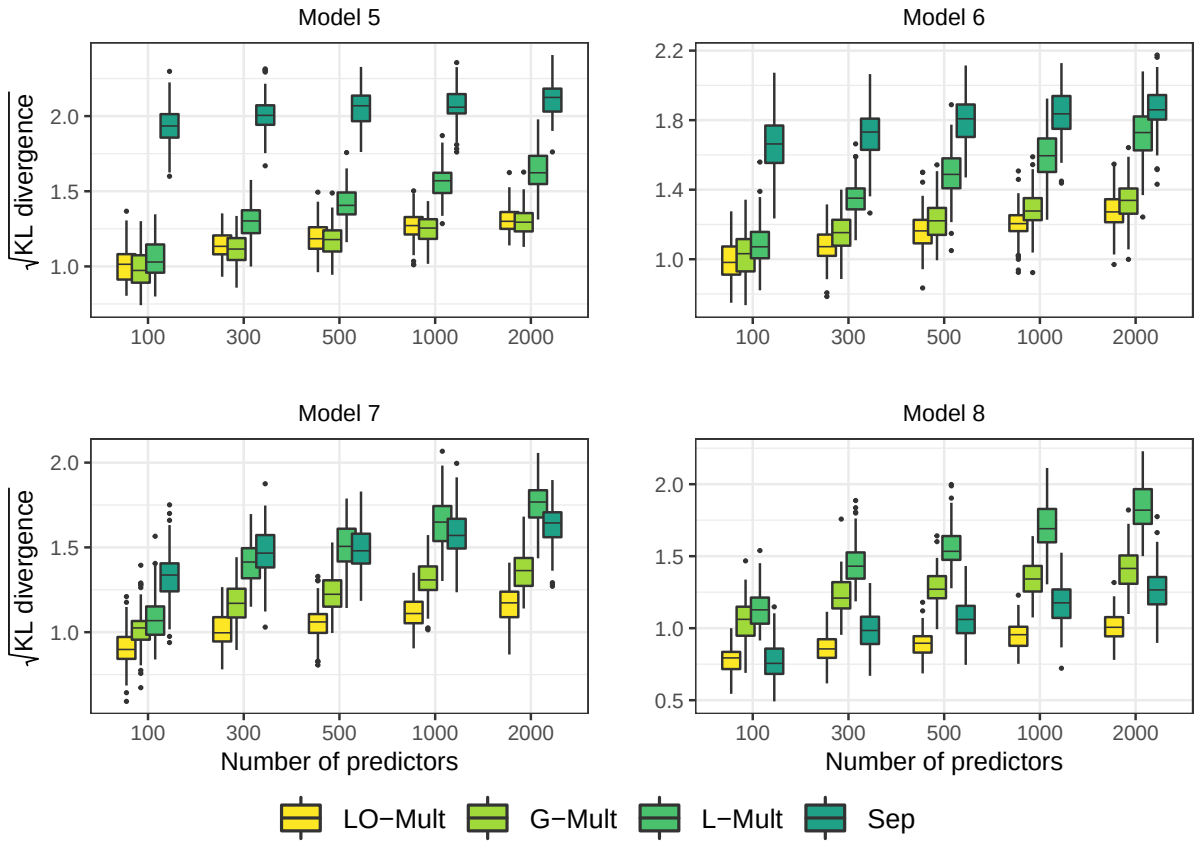

Figure 13: Square-root average Kullback-Leibler divergence under Models 5–8 with $p \in \{100, 300, 500, 1000, 2000\}$ with $J = K = L = 2$.

to outperform competitors more relative to when, say, $p = 100$.

D Proofs of results in Section 4

In this and the following sections, for ease of display, we omit the subscript on 0 when refering to a matrix or vector of zeros. The key to proving Theorem 2 is the following lemma[1], which reveals that we need only concern ourselves with computing $\hat{\eta}_{\bar{\lambda},0}$.

Lemma 1. *Let $\hat{\eta}_{\bar{\lambda},\bar{\gamma}}$ be a minimizer of (11) and let $\hat{\eta}_{\bar{\lambda},0}$ be the minimizer of (11) with $\bar{\gamma} = 0$. Then*

$$\hat{\eta}_{\bar{\lambda},\bar{\gamma}} = \begin{cases} \left(1 - \frac{\bar{\gamma}}{\|\hat{\eta}_{\bar{\lambda},0}\|_2}\right) \hat{\eta}_{\bar{\lambda},0} & : \|\hat{\eta}_{\bar{\lambda},0}\|_2 > \bar{\gamma} \\ 0 & : \|\hat{\eta}_{\bar{\lambda},0}\|_2 \leq \bar{\gamma} \end{cases} \tag{17}$$

Proof of Lemma 1. To prove Lemma 1, we show that first-order conditions for $\hat{\eta}_{\bar{\lambda},0}$ imply the first-order conditions for $\hat{\eta}_{\bar{\lambda},\bar{\gamma}}$ as defined in (17). First, recall that the zero subgradient

[1] A previous version of the Supplementary Materials contained a typo in the statement of this lemma.

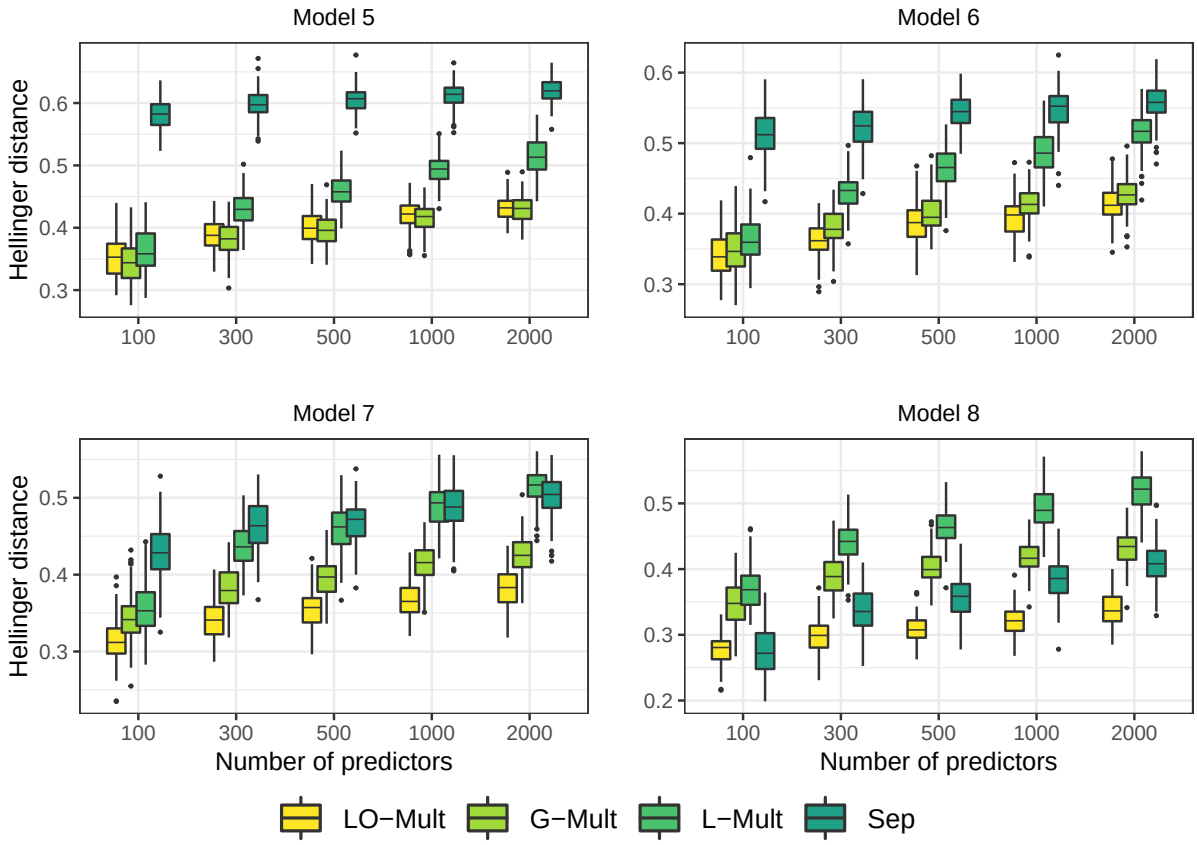

Figure 14: Average Hellinger distance under Models 5–8 with $p \in \{100, 300, 500, 1000, 2000\}$ with $J = K = L = 2$.

equation for $\hat{\eta}_{\bar{\lambda},0}$ is

$$0 = -\nu + \hat{\eta}_{\bar{\lambda},0} + \bar{\lambda} D \tilde{\phi} \tag{18}$$

for some $\tilde{\phi}$ such that $\tilde{\phi} = D'\hat{\eta}_{\bar{\lambda},0}/\|D'\hat{\eta}_{\bar{\lambda},0}\|_2$ if $[D'\hat{\eta}_{\bar{\lambda},0}] \neq 0$ and $\|\tilde{\phi}\|_2 \leq 1$ otherwise (i.e., $\tilde{\phi}$ is a subgradient of $\eta \mapsto \|D'\eta\|_2$ at $\hat{\eta}_{\bar{\lambda},0}$). Then, recall that the zero subgradient equation for $\hat{\eta}_{\bar{\lambda},\bar{\gamma}}$ is

$$0 = -\nu + \hat{\eta}_{\bar{\lambda},\bar{\gamma}} + \bar{\lambda} D \phi + \bar{\gamma} v, \tag{19}$$

for $(v, \phi) \in \mathbb{R}^{JK} \times \mathbb{R}^{\binom{J}{2}\binom{K}{2}}$ such that $v = \hat{\eta}_{\bar{\lambda},\bar{\gamma}}/\|\hat{\eta}_{\bar{\lambda},\bar{\gamma}}\|_2$ if $\hat{\eta}_{\bar{\lambda},\bar{\gamma}} \neq 0$ and $\|v\|_2 \leq 1$ otherwise; and $\phi = D'\hat{\eta}_{\bar{\lambda},\bar{\eta}}/\|D'\hat{\eta}_{\bar{\lambda},\bar{\eta}}\|_2$ if $D'\hat{\eta}_{\bar{\lambda},\bar{\eta}} \neq 0$ and $\|\phi\|_2 \leq 1$ otherwise.

We will consider three cases: (i) $\|\hat{\eta}_{\bar{\lambda},0}\|_2 > \bar{\gamma}$, (ii) $0 < \|\hat{\eta}_{\bar{\lambda},0}\|_2 \leq \bar{\gamma}$, and (iii) $\hat{\eta}_{\bar{\lambda},0} = 0$.

Case (i): We know from (18) that there exists a subgradient $\tilde{\phi}$ such that

$$0 = -\nu + \hat{\eta}_{\bar{\lambda},0} + \bar{\lambda} D \tilde{\phi}. \tag{20}$$

We assume that $\|\hat{\eta}_{\bar{\lambda},0}\|_2 > \bar{\gamma}$ so that $\hat{\eta}_{\bar{\lambda},\bar{\gamma}} = \hat{\eta}_{\bar{\lambda},0}(1 - \bar{\gamma}/\|\hat{\eta}_{\bar{\lambda},0}\|_2)$. We will show that this $\hat{\eta}_{\bar{\lambda},\bar{\gamma}}$

satisfies the first-order conditions (19). In particular, from (20), we have

$$\begin{aligned}
0 &= -\nu + \hat{\eta}_{\bar{\lambda},0} + \bar{\lambda} D\tilde{\phi} \\
\implies 0 &= -\nu + \hat{\eta}_{\bar{\lambda},0} + \bar{\lambda} D\tilde{\phi} + \bar{\gamma}\hat{\eta}_{\bar{\lambda},0}/\|\hat{\eta}_{\bar{\lambda},0}\|_2 - \bar{\gamma}\hat{\eta}_{\bar{\lambda},0}/\|\hat{\eta}_{\bar{\lambda},0}\|_2 \\
\implies 0 &= -\nu + \hat{\eta}_{\bar{\lambda},0}(1 - \bar{\gamma}/\|\hat{\eta}_{\bar{\lambda},0}\|_2) + \bar{\lambda} D\tilde{\phi} + \bar{\gamma}\hat{\eta}_{\bar{\lambda},0}/\|\hat{\eta}_{\bar{\lambda},0}\|_2 \\
\implies 0 &= -\nu + \hat{\eta}_{\bar{\lambda},0}(1 - \bar{\gamma}/\|\hat{\eta}_{\bar{\lambda},0}\|_2) + \bar{\lambda} D\tilde{\phi} + \bar{\gamma}\hat{\eta}_{\bar{\lambda},0}(1 - \bar{\gamma}/\|\hat{\eta}_{\bar{\lambda},0}\|_2)/\|\hat{\eta}_{\bar{\lambda},0}(1 - \bar{\gamma}/\|\hat{\eta}_{\bar{\lambda},0}\|_2)\|_2 \\
\implies 0 &= -\nu + \hat{\eta}_{\bar{\lambda},\bar{\gamma}} + \bar{\lambda} D\tilde{\phi} + \bar{\gamma}\hat{\eta}_{\bar{\lambda},\bar{\gamma}}/\|\hat{\eta}_{\bar{\lambda},\bar{\gamma}}\|_2 \qquad (21)
\end{aligned}$$

Since $\|\hat{\eta}_{\bar{\lambda},\bar{\gamma}}\|_2 > 0$ by assumption on $\hat{\eta}_{\bar{\lambda},0}$, we can take $v = \hat{\eta}_{\bar{\lambda},\bar{\gamma}}/\|\hat{\eta}_{\bar{\lambda},\bar{\gamma}}\|_2$. It only remains to check that $\tilde{\phi} = \phi$ where $\phi = D'\hat{\eta}_{\bar{\lambda},\bar{\gamma}}/\|D'\hat{\eta}_{\bar{\lambda},\bar{\gamma}}\|_2$ if $D'\hat{\eta}_{\bar{\lambda},\bar{\gamma}} \neq 0$ and $\|\phi\|_2 \leq 1$ otherwise. However, this is trivial since $\hat{\eta}_{\bar{\lambda},\bar{\gamma}}$ is a scalar multiple of $\hat{\eta}_{\bar{\lambda},0}$, so $D'\hat{\eta}_{\bar{\lambda},0}$ is a scalar multiple of $D'\hat{\eta}_{\bar{\lambda},\bar{\gamma}}$. Thus, if $D'\hat{\eta}_{\bar{\lambda},0} \neq 0$, then $D'\hat{\eta}_{\bar{\lambda},\bar{\gamma}} \neq 0$, whereas if $D'\hat{\eta}_{\bar{\lambda},0} = 0$, then $D'\hat{\eta}_{\bar{\lambda},\bar{\gamma}} = 0$. In either case, we can take $\phi = \tilde{\phi}$ so that finally, from (21),

$$0 = -\nu + \hat{\eta}_{\bar{\lambda},\bar{\gamma}} + \bar{\lambda} D\tilde{\phi} + \bar{\gamma}\hat{\eta}_{\bar{\lambda},\bar{\gamma}}/\|\hat{\eta}_{\bar{\lambda},\bar{\gamma}}\|_2 \implies 0 = -\nu + \hat{\eta}_{\bar{\lambda},\bar{\gamma}} + \bar{\lambda} D\phi + \bar{\gamma} v$$

which verifies that $\hat{\eta}_{\bar{\lambda},\bar{\gamma}}$ as defined in (17) satisfies the first-order optimality conditions for (11) when $\|\hat{\eta}_{\bar{\lambda},0}\|_2 > \bar{\gamma}$.

Case (ii): Assume $0 < \|\hat{\eta}_{\bar{\lambda},0}\|_2 \leq \bar{\gamma}$. We will show that $\hat{\eta}_{\bar{\lambda},\bar{\gamma}} = 0$ satisfies the first-order conditions for (11) given in (19). Recall that by definition, there exists a subgradient $\tilde{\phi}$ such that

$$0 = -\nu + \hat{\eta}_{\bar{\lambda},0} + \bar{\lambda} D\tilde{\phi}. \qquad (22)$$

Since $\|\hat{\eta}_{\bar{\lambda},0}\|_2 \leq \bar{\gamma}$, $1 \leq \bar{\gamma}/\|\hat{\eta}_{\bar{\lambda},0}\|_2$, so we can write $1 = \bar{\gamma}/\|\hat{\eta}_{\bar{\lambda},0}\|_2 - z_1$ for some $z_1 \geq 0$ and thus, (22) implies

$$0 = -\nu + \hat{\eta}_{\bar{\lambda},0}\left(\frac{\bar{\gamma}}{\|\hat{\eta}_{\bar{\lambda},0}\|_2} - z_1\right) + \bar{\lambda} D\tilde{\phi}$$

which in turn implies

$$0 = -\nu + \hat{\eta}_{\bar{\lambda},\bar{\gamma}} + \bar{\lambda} D\tilde{\phi} + \bar{\gamma}\left(\frac{\hat{\eta}_{\bar{\lambda},0}}{\|\hat{\eta}_{\bar{\lambda},0}\|_2} - \frac{z_1\hat{\eta}_{\bar{\lambda},0}}{\bar{\gamma}}\right) \qquad (23)$$

since $\hat{\eta}_{\bar{\lambda},\bar{\gamma}} = 0$ by assumption. Then, because we must have $\|\phi\|_2 \leq 1$, we can simply take $\phi = \tilde{\phi}$ since $\|\tilde{\phi}\|_2 \leq 1$ regardless of whether $D'\hat{\eta}_{\bar{\lambda},0} = 0$ or $D'\hat{\eta}_{\bar{\lambda},0} \neq 0$. Thus, (23) suggets that $\hat{\eta}_{\bar{\lambda},\bar{\gamma}} = 0$ satisfies the first-order conditions for (11) as long as

$$\left\| \frac{\hat{\eta}_{\bar{\lambda},0}}{\|\hat{\eta}_{\bar{\lambda},0}\|_2} - \frac{z_1\hat{\eta}_{\bar{\lambda},0}}{\bar{\gamma}} \right\|_2 \leq 1.$$

Letting $z_2 = \hat{\eta}_{\bar{\lambda},0}/\|\hat{\eta}_{\bar{\lambda},0}\|_2$ so that $\|z_2\|_2 = 1$, we have

$$\left\| \frac{\hat{\eta}_{\bar{\lambda},0}}{\|\hat{\eta}_{\bar{\lambda},0}\|_2} - \frac{z_1 \hat{\eta}_{\bar{\lambda},0}}{\bar{\gamma}} \right\|_2 = \|z_2(1 - \bar{\gamma}^{-1} z_1 \|\hat{\eta}_{\bar{\lambda},0}\|_2)\|_2 = \|z_2\|_2 \left(1 - \frac{z_1 \|\hat{\eta}_{\bar{\lambda},0}\|_2}{\bar{\gamma}}\right) = \left(1 - \frac{z_1}{1+z_1}\right) \leq 1.$$

Therefore, with $v = \frac{\hat{\eta}_{\bar{\lambda},0}}{\|\hat{\eta}_{\bar{\lambda},0}\|_2} - \frac{z_1 \hat{\eta}_{\bar{\lambda},0}}{\bar{\gamma}}$, from (23) we can conclude,

$$0 = -\nu + \hat{\eta}_{\bar{\lambda},\bar{\gamma}} + \bar{\lambda} D\tilde{\phi} + \bar{\gamma} \left(\frac{\hat{\eta}_{\bar{\lambda},0}}{\|\hat{\eta}_{\bar{\lambda},0}\|_2} - \frac{z_1 \hat{\eta}_{\bar{\lambda},0}}{\bar{\gamma}} \right) \implies 0 = -\nu + \hat{\eta}_{\bar{\lambda},\bar{\gamma}} + \bar{\lambda} D\phi + \bar{\gamma} v$$

for a $(v, \phi) \in \mathbb{R}^{JK} \times \mathbb{R}^{\binom{J}{2}\binom{K}{2}}$ such that $\|v\|_2 \leq 1$ and $\|\phi\|_2 \leq 1$, which is exactly the zero subgradient equation when $\hat{\eta}_{\bar{\lambda},\bar{\gamma}} = 0$.

Case (iii): This case is trivial: to see that zero subgradient equation for $\hat{\eta}_{\bar{\lambda},0} = 0$ implies the zero subgradient equation for $\hat{\eta}_{\bar{\lambda},\bar{\gamma}} = 0$, simply take $\phi = \tilde{\phi}$ and $v = 0$. ■

With Lemma 1 in place, we are ready to prove Theorem 2.

Proof of Theorem 2. Recall that the zero subgradient equation for $\hat{\eta}_{\bar{\lambda},\bar{\gamma}}$ is

$$0 = -\nu + \hat{\eta}_{\bar{\lambda},\bar{\gamma}} + \bar{\lambda} D\phi + \bar{\gamma} v, \tag{24}$$

where

$$v \in \{v \in \mathbb{R}^{JK} : v = \hat{\eta}_{\bar{\lambda},\bar{\gamma}}/\|\hat{\eta}_{\bar{\lambda},\bar{\gamma}}\|_2 \text{ if } \hat{\eta}_{\bar{\lambda},\bar{\gamma}} \neq 0 \text{ and } \|v\|_2 \leq 1 \text{ otherwise}\},$$

and

$$\phi \in \{\phi \in \mathbb{R}^{\binom{J}{2}\binom{K}{2}} : \phi = D'\hat{\eta}_{\bar{\lambda},\bar{\eta}}/\|D'\hat{\eta}_{\bar{\lambda},\bar{\eta}}\|_2 \text{ if } D'\hat{\eta}_{\bar{\lambda},\bar{\eta}} \neq 0 \text{ and } \|\phi\|_2 \leq 1 \text{ otherwise}\}.$$

We consider each of the three cases set out in the statement of Theorem 2. To deal with cases (ii) and (iii), we focus on the solution for $\hat{\eta}_{\bar{\lambda},0}$ and then apply Lemma 1.

Case (i): If $\|\nu\|_2 \leq \bar{\gamma}$, we can set $\hat{\eta}_{\bar{\lambda},\bar{\gamma}} = 0$, $\phi = 0$, and $v = \nu/\bar{\gamma}$, so that $\|v\|_2 \leq 1$, and thus, $\hat{\eta}_{\bar{\lambda},\bar{\gamma}} = 0$ would satisfy the first-order conditions (24).

Case (ii): We consider the dual problem of (11) with $\bar{\gamma} = 0$ (e.g., see the derivation of a related dual problem in Section 4 of Tibshirani et al. (2011)):

$$\hat{u} \in \arg\min_u \|\nu - Du\|_2^2, \quad \|u\|_2 \leq \bar{\lambda},$$

where $\hat{\eta}_{\bar{\lambda},0} = \nu - D\hat{u}$. Hence, if $\|(D'D)^- D'\nu\|_2 \leq \bar{\lambda}$, $\hat{u} = (D'D)^- D'\nu$, so it would follow that

$\hat{\eta}_{\bar{\lambda},0} = \nu - D(D'D)^{-}D'\nu = \mathcal{P}^{\perp}_{D,0}\nu$. An application of Lemma 1 yields the second result.

Case (iii): We again consider the dual problem of (11) with $\bar{\gamma} = 0$. If $\|(D'D)^{-}D'\nu\|_2 > \bar{\lambda}$, it must be that the minimizer $\hat{u}$ is only the boundary of the constraint set $\{u : \|u\|_2 \leq \bar{\lambda}\}$, or equivalently, $\|\hat{u}\|_2^2 = \bar{\lambda}^2$. Then, because there is a one-to-one correspondence between the constrained version of ridge regression and its Lagrangian form when the constraint is active, we know there exists a $\tau > 0$ such that for every $\bar{\lambda}$ satisfying the condition of (iii),

$$\hat{u} = \underset{u:\|u\|_2^2 \leq \bar{\lambda}^2}{\arg\min} \|\nu - Du\|_2^2 = \underset{u}{\arg\min} \|\nu - Du\|_2^2 + \tau\|u\|_2^2,$$

and thus, since $(D'D+\tau I)^{-1}D'\nu$ minimizes the rightmost objective function above, if $\|(D'D+\tau I)^{-1}D'\nu\|_2^2 = \bar{\lambda}^2$, we know $\hat{u} = (D'D+\tau I)^{-1}D'\nu$. The result then follows from $\nu - D(D'D+\tau I)^{-1}D'\nu = \mathcal{P}^{\perp}_{D,\tau}\nu$ and Lemma 1. ■

Next, we provide a sketch of the proof of Proposition 1.

Proof of Proposition 1. Let $U\text{Diag}\left(\{\sigma_l\}_{l=1}^k\right)V'$ be the singular value decomposition of D where $k = \min(JK, \binom{J}{2}\binom{K}{2})$, $U'U = I_k$, $V'V = I_k$, and $\sigma_l \geq 0$ for $l \in [k]$. Note that by construction, only the first $r = (J-1)(K-1)$ singular values of D are nonzero (e.g., see discussion of D versus $\mathcal{D}$ in Section 2). Then, letting $\Sigma = \text{Diag}\left(\{\sigma_l\}_{l=1}^k\right)$, we can write

$$(D'D+\tau I)^{-1}D'\nu = V(\Sigma^2+\tau I)^{-1}\Sigma U'\nu$$

so that

$$\|(D'D+\tau I)^{-1}D'\nu\|_2 = \bar{\lambda} \iff \nu'U\Sigma(\Sigma^2+\tau I)^{-2}\Sigma U'\nu = \bar{\lambda}^2.$$

Letting u_l denote the lth column of U, we can define $w = (w_1, \ldots, w_k)' \in \mathbb{R}^k$ where $w_l = u_l'\nu \in \mathbb{R}$ so that we may write

$$\nu'U\Sigma(\Sigma^2+\tau I)^{-2}\Sigma U'\nu = w'Aw,$$

where A is diagonal with (l,l)th entry $(\sigma_l^2+\tau)^{-2}\sigma_l^2$. Thus, it follows that

$$w'Aw = \sum_{l=1}^{r} \frac{w_l^2\sigma_l^2}{(\sigma_l^2+\tau)^2},$$

which yields the first result. Then because for each $l \in [r]$, $\sigma_l = \sqrt{JK}$, it further follows that

$$\sum_{l=1}^{r} \frac{w_l^2\sigma_l^2}{(\sigma_l^2+\tau)^2} = \lambda^2 \implies JK\sum_{l=1}^{r} \frac{w_l^2}{(JK+\tau)^2} = \lambda^2.$$

And thus, the previous equality implies

$$\tau = \frac{\sqrt{JK \sum_{l=1}^{r} w_l^2}}{\bar{\lambda}} - JK.$$

It is easy to check that under the conditions of (iii), this τ must be positive. ■

Proof of Theorem 3. We again appeal to Lemma 1, which will give us the result for $\hat{\eta}_{\bar{\lambda},\bar{\gamma}}$ once we have obtained the expression for $\hat{\eta}_{\bar{\lambda},0}$. We thus focus on the solution for $\hat{\eta}_{\bar{\lambda},0}$. Recall that when $J = K = 2$, $D'\hat{\eta}_{\bar{\lambda},0} \in \mathbb{R}$ and $\tilde{\phi} \in \mathbb{R}$, where $\tilde{\phi} = \text{sign}(D'\hat{\eta}_{\bar{\lambda},0})$ if $D'\hat{\eta}_{\bar{\lambda},0} \neq 0$ and $\tilde{\phi} \in [-1, 1]$ otherwise. We consider all three cases enumerated in the statement of Theorem 3. Let $\ddot{\nu} = \nu_1 - \nu_2 - \nu_3 + \nu_4$ and recall in this setting, $D = (1, -1, -1, 1)'$.

Case (iii): Suppose $\ddot{\nu} < -4\bar{\lambda}$. If we let $\hat{\eta}_{\bar{\lambda},0} = (\nu_1 + \bar{\lambda}, \nu_2 - \bar{\lambda}, \nu_3 - \bar{\lambda}, \nu_4 + \bar{\lambda})'$, then the gradient of the objective is

$$-\nu + \hat{\eta}_{\bar{\lambda},0} + \bar{\lambda} D \text{sign}(D'\hat{\eta}_{\bar{\lambda},0}) = -\begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \end{pmatrix} + \begin{pmatrix} \nu_1 + \bar{\lambda} \\ \nu_2 - \bar{\lambda} \\ \nu_3 - \bar{\lambda} \\ \nu_4 + \bar{\lambda} \end{pmatrix} - \bar{\lambda} D$$

since

$$\text{sign}(D'\hat{\eta}_{\bar{\lambda},0}) = \text{sign}(\nu_1 + \bar{\lambda} - (\nu_2 - \bar{\lambda}) - (\nu_3 - \bar{\lambda}) + \nu_4 + \bar{\lambda}) = \text{sign}(\ddot{\nu} + 4\bar{\lambda}) = -1$$

by our assumption $\ddot{\nu} < -4\bar{\lambda}$. Hence, because

$$-\begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \end{pmatrix} + \begin{pmatrix} \nu_1 + \bar{\lambda} \\ \nu_2 - \bar{\lambda} \\ \nu_3 - \bar{\lambda} \\ \nu_4 + \bar{\lambda} \end{pmatrix} - \bar{\lambda} \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} = 0,$$

when $\ddot{\nu} < -4\bar{\lambda}$, the first-order conditions

$$-\nu + \hat{\eta}_{\bar{\lambda},0} + \bar{\lambda} D \text{sign}(D'\hat{\eta}_{\bar{\lambda},0}) = 0$$

are satisfied with $\hat{\eta}_{\bar{\lambda},0} = (\nu_1 + \bar{\lambda}, \nu_2 - \bar{\lambda}, \nu_3 - \bar{\lambda}, \nu_4 + \bar{\lambda})'$.

Case (ii): When $\ddot{\nu} > 4\bar{\lambda}$, the result follows from a nearly identical proof as in case (iii).

Case (i): Suppose $|\ddot{\nu}| \leq 4\bar{\lambda}$. Let $\hat{\eta}_{\bar{\lambda},0} = (\nu_1 - \ddot{\nu}/4, \nu_2 + \ddot{\nu}/4, \nu_3 + \ddot{\nu}/4, \nu_4 - \ddot{\nu}/4)'$. We want to

show that

$$-\nu + \hat{\eta}_{\bar{\lambda},0} + \bar{\lambda} D u = 0 \tag{25}$$

for some $u \in [-1, 1]$. Notice,

$$-\nu + \hat{\eta}_{\bar{\lambda},0} + \bar{\lambda} D u = -\begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \end{pmatrix} + \begin{pmatrix} \nu_1 + \ddot{\nu}/4 \\ \nu_2 - \ddot{\nu}/4 \\ \nu_3 - \ddot{\nu}/4 \\ \nu_4 + \ddot{\nu}/4 \end{pmatrix} + \bar{\lambda} \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} u = \begin{pmatrix} \ddot{\nu}/4 \\ -\ddot{\nu}/4 \\ -\ddot{\nu}/4 \\ \ddot{\nu}/4 \end{pmatrix} + \bar{\lambda} \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} u.$$

Therefore, if we set $u = -\ddot{\nu}/(4\bar{\lambda})$, we know $u \in [-1, 1]$ by assumption and thus,

$$-\nu + \hat{\eta}_{\bar{\lambda},0} + \bar{\lambda} D u = \begin{pmatrix} \ddot{\nu}/4 \\ -\ddot{\nu}/4 \\ -\ddot{\nu}/4 \\ \ddot{\nu}/4 \end{pmatrix} - \bar{\lambda} \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \ddot{\nu}/(4\bar{\lambda}) = 0$$

so that the first-order conditions (25) are satisfied. ■

E Proofs of results in Section 5

Proof of Lemma 1. It is straightforward to show, e.g., see Agresti (2002), that (12) implies a). To show that the latter two log odds constraints imply b), notice with a) holding,

$$P(Y_1 = j, Y_2 = 1 \mid x, Y_3 = l) = P(Y_1 = j \mid x, Y_3 = l) P(Y_2 = 1 \mid x, Y_3 = l), \quad (j, l) \in [J] \times [L],$$

so that we can write, for all $(j, l) \in [J-1] \times [L-1]$,

$$\begin{aligned} &\frac{P(Y_1 = j \mid x, Y_3 = l) P(Y_1 = j+1 \mid x, Y_3 = l+1)}{P(Y_1 = j+1 \mid x, Y_3 = l) P(Y_1 = j \mid x, Y_3 = l+1)} \\ &\quad = \frac{P(Y_1 = j \mid x, Y_3 = l) P(Y_1 = j+1 \mid x, Y_3 = l+1)}{P(Y_1 = j+1 \mid x, Y_3 = l) P(Y_1 = j \mid x, Y_3 = l+1)} \frac{P(Y_2 = 1 \mid x, Y_3 = l) P(Y_2 = 1 \mid x, Y_3 = l+1)}{P(Y_2 = 1 \mid x, Y_3 = l) P(Y_2 = 1 \mid x, Y_3 = l+1)} \\ &\qquad = \frac{P(Y_1 = j, Y_2 = 1 \mid x, Y_3 = l) P(Y_1 = j+1, Y_2 = 1 \mid x, Y_3 = l+1)}{P(Y_1 = j+1, Y_2 = 1 \mid x, Y_3 = l) P(Y_1 = j, Y_2 = 1 \mid x, Y_3 = l+1)} \end{aligned}$$

and thus,

$$\log \left(\frac{\pi^*_{j,1,l}(x) \pi^*_{j+1,1,l+1}(x)}{\pi^*_{j+1,1,l}(x) \pi^*_{j,1,l+1}(x)} \right) = 0, \quad (j, l) \in [J-1] \times [L-1]$$

implies

$$\log\left(\frac{P(Y_1 = j \mid x, Y_3 = l)P(Y_1 = j+1 \mid x, Y_3 = l+1)}{P(Y_1 = j+1 \mid x, Y_3 = l)P(Y_1 = j \mid x, Y_3 = l+1)}\right) = 0, \quad (j,l) \in [J-1] \times [L-1]$$

which implies the left expression in b). The right expression in b) follows from the same set of arguments, reversing the roles of Y_1 and Y_2. It is immediate that a) and b) together imply (12). ■

F Proof of Theorem 1

F.1 Main proof

We first provide a number of key lemmas which we use to establish the result in Theorem 1. We provide proofs of these lemmas in the subsequent subsection.

In order to obtain our error bound, we use a property of the multinomial negative log-likelihood closely related to *self-concordance*. We begin with a lemma from Bach (2010), which defines the notion of ν-self-concordance and establishes an upper bound on the Taylor expansion of any function satisfying the conditions of 2-self-concordance.

Lemma 2. *(Proposition 1, Bach (2010)) Let $F : \mathbb{R}^q \to \mathbb{R}$ be a convex, three times differentiable function such that for all $w, v \in \mathbb{R}^q$, the function $g(t) = F(w + tv)$ satisfies for all $t \in \mathbb{R}$, $|\nabla^3 g(t)| \leq R\|v\|_2 \cdot [\nabla^2 g(t)]^{\nu/2}$ for some fixed constants $\nu > 0$ and $R \geq 0$. Then, if such a $R \geq 0$ exists for a given ν, F is said to be ν-self-concordant. Moreover, if F is 2-self-concordant, then for all $w \in \mathbb{R}^q$ and $v \in \mathbb{R}^q$*

$$F(w+v) \geq F(w) + \operatorname{tr}\{v'\nabla F(w)\} + \frac{v'\nabla^2 F(w)v}{R^2\|v\|_2^2}(e^{-R\|v\|_2} + R\|v\|_2 - 1),$$

for the corresponding $R \geq 0$.

Following the proof of Lemma 4 from Tran-Dinh et al. (2015), we establish that $\mathcal{G}$, the (scaled) multinomial negative log-likelihood, is a 2-self concordant function. For completeness, we include a proof in the next subsection.

Lemma 3. *The function $\tilde{\mathcal{G}} : \mathbb{R}^{p \times JK} \to \mathbb{R}$ satisfies the definition of 2-self-concordance with $R = \sqrt{6}\max_{i \in [n]} \|X_{i,:}\|_2$.*

Combining Lemma 2 and Lemma 3, we have that for any $\beta^\dagger$ and Δ,

$$\begin{aligned}\tilde{\mathcal{G}}(\beta^\dagger + \Delta) - \tilde{\mathcal{G}}(\beta^\dagger) \geq \operatorname{tr}\{\Delta'\nabla\tilde{\mathcal{G}}(\beta^\dagger)\} \\ + \frac{\operatorname{vec}(\Delta)'\nabla^2\tilde{\mathcal{G}}(\beta^\dagger)\operatorname{vec}(\Delta)}{d_n^2\|\Delta\|_F^2}\left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right),\end{aligned} \tag{26}$$

where $d_n = \sqrt{6}\max_{i\in[n]} \|X_{i,:}\|_2$. With (26) in hand, we then apply the proof technique outlined in Negahban et al. (2012). First, we need another lemma, Lemma 4, which states that when the tuning parameters are chosen appropriately, the error $\hat{\beta} - \beta^\dagger$ belongs to the set $\mathbb{C}(\mathcal{S}, \phi)$. The proof of Lemma 4 is given in the next subsection.

Lemma 4. *If $\lambda = \phi_2\gamma$ and $\gamma > \phi_1\|\nabla\tilde{\mathcal{G}}(\beta^\dagger)\|_{\infty,2}$ where $\|A\|_{\infty,2} = \max_j \|A_{j,:}\|_2$, then $\hat{\Delta} = \hat{\beta} - \beta^\dagger$ belongs to the set $\mathbb{C}(\mathcal{S}, \phi)$.*

Lemma 5. *Let*

$$\gamma = \frac{\phi_1\epsilon\kappa(\mathcal{S}, \phi)}{c\{(\phi_1 + 1)\sqrt{|S_L| + |S_M|} + \phi_1\phi_2\Psi_{J,K}(S_L)\}},$$

for some fixed constants $c > 2$, $\phi_1 > 1$, and $\phi_2 > 0$. If $\gamma > \phi_1\|\nabla\mathcal{G}(\beta^\dagger)\|_{\infty,2}$ and $\epsilon > 0$ is sufficiently close to zero such that $e^{-d_n\epsilon} + d_n\epsilon - d_n^2\epsilon^2/c - 1 > 0$, then $\|\hat{\beta} - \beta^\dagger\|_F \leq \epsilon$.

Finally, we need to assign a probability to the event $\gamma > \phi_1\|\nabla\tilde{\mathcal{G}}(\beta^\dagger)\|_{\infty,2}$ for a particular choice of γ. Along these lines, we have the following lemma.

Lemma 6. *Under assumption A1 and A2,*

$$P\left\{\|\nabla\tilde{\mathcal{G}}(\beta^\dagger)\|_{\infty,2} \leq \sqrt{\frac{JK}{4n}} + \sqrt{\frac{\log(p/\alpha)}{n}}\right\} \geq 1 - \alpha.$$

With all the pieces in place, we are now ready to prove Theorem 1.

Proof of Theorem 1. To prove Theorem 1, we combine Lemma 5 and Lemma 6. Specifically, let $\gamma = \phi_1\{JK/(4n)\}^{1/2} + \phi_1\{\log(p/\alpha)/n\}^{1/2}$, $\lambda = \phi_2\gamma$, and (following the first equality in the statement of Lemma 5) take

$$\begin{aligned}\epsilon &= \frac{\gamma c\{(\phi_1 + 1)\sqrt{|S_L| + |S_M|} + \phi_1\phi_2\Psi_{J,K}(S_L)\}}{\phi_1\kappa(\mathcal{S}, \phi)} \\ &= \frac{c\{(\phi_1 + 1)\sqrt{|S_L| + |S_M|} + \phi_1\phi_2\Psi_{J,K}(S_L)\}}{\kappa(\mathcal{S}, \phi)}\left\{\sqrt{\frac{JK}{4n}} + \sqrt{\frac{\log(p/\alpha)}{n}}\right\}\end{aligned}$$

where $c > 2$ is a fixed constant. Then, under Condition 1, $e^{-d_n\epsilon} + d_n\epsilon - d_n^2\epsilon^2/c - 1 > 0$ so that it follows from applications of Lemma 5 and 6 that

$$P(\|\hat{\beta} - \beta^*\|_F \leq \epsilon) \geq P\left\{\|\nabla\tilde{\mathcal{G}}(\beta^\dagger)\|_{\infty,2} \leq \sqrt{\frac{JK}{4n}} + \sqrt{\frac{\log(p/\alpha)}{n}}\right\} \geq 1 - \alpha. \quad \blacksquare$$

F.2 Proofs of results in Section F.1

Proof of Lemma 3. Our proof uses the same steps as the proof of Lemma 4 from Tran-Dinh et al. (2015), although our result is different (by a factor of n). Let $\tilde{g}(t) = \tilde{\mathcal{G}}(A + tB)$ for

matrices $A \in \mathbb{R}^{p\times JK}$ and $B \in \mathbb{R}^{p\times JK}$. Then, we write $\tilde{g}$ as

$$\tilde{g}(t) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{i,j,k}[x_i' A_{\cdot,j,k} + t(x_i' B_{\cdot,j,k})] + \frac{1}{n}\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{J}\sum_{k=1}^{K}\exp\left[x_i' A_{\cdot,j,k} + t(x_i' B_{\cdot,j,k})\right]\right\}.$$

Our objective is to show that $\tilde{g}$ satisfies the conditions from Lemma 2. However, note that the second and third derivatives of $\tilde{g}$ depend only on the second term, so we show the conditions hold instead for

$$g(t) = \frac{1}{n}\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{J}\sum_{k=1}^{K}\exp\left[x_i' A_{\cdot,j,k} + t(x_i' B_{\cdot,j,k})\right]\right\},$$

which would be sufficient for the desired result. Letting $\mu_{i,j,k}(t) = \exp\{x_i' A_{\cdot,j,k} + t(x_i' B_{\cdot,j,k})\}$ and $b_{i,j,k} = x_i' B_{\cdot,j,k}$, we have

$$\nabla^2 g(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{J}\sum_{k=1}^{K} b_{i,j,k}^2 \mu_{i,j,k}(t)}{\sum_{j=1}^{J}\sum_{k=1}^{K}\mu_{i,j,k}(t)} - \left[\frac{\sum_{j=1}^{J}\sum_{k=1}^{K} b_{i,j,k}\mu_{i,j,k}(t)}{\sum_{j=1}^{J}\sum_{k=1}^{K}\mu_{i,j,k}(t)}\right]^2\right\}$$

and also

$$\nabla^3 g(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{J}\sum_{k=1}^{K} b_{i,j,k}^3 \mu_{i,j,k}(t)}{\sum_{j=1}^{J}\sum_{k=1}^{K}\mu_{i,j,k}(t)} + 2\left[\frac{\sum_{j=1}^{J}\sum_{k=1}^{K} b_{i,j,k}\mu_{i,j,k}(t)}{\sum_{j=1}^{J}\sum_{k=1}^{K}\mu_{i,j,k}(t)}\right]^3 - \frac{3\left[\sum_{j=1}^{J}\sum_{k=1}^{K} b_{i,j,k}^2\mu_{i,j,k}(t)\right]\left[\sum_{j=1}^{J}\sum_{k=1}^{K} b_{i,j,k}\mu_{i,j,k}(t)\right]}{\left[\sum_{j=1}^{J}\sum_{k=1}^{K}\mu_{i,j,k}(t)\right]^2}\right\}$$

Next, we simplify $\nabla^2 g(t)$. Letting $\mu_i(t) = \sum_{j=1}^{J}\sum_{k=1}^{K}\mu_{i,j,k}(t)$, and letting $\sum_{j,k}(\text{resp.}\sum_{s,t})$ denote $\sum_{j=1}^{J}\sum_{k=1}^{K}$ (resp. $\sum_{s=1}^{J}\sum_{t=1}^{K}$) for ease of display,

$$\begin{aligned}\nabla^2 g(t) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mu_i(t)\left[\sum_{j,k} b_{i,j,k}^2\mu_{i,j,k}(t)\right] - \left[\sum_{j,k} b_{i,j,k}\mu_{i,j,k}(t)\right]^2}{\mu_i(t)^2}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j,k}\sum_{s,t}(b_{i,j,k} - b_{i,s,t})^2\mu_{i,j,k}(t)\mu_{i,s,t}(t)}{2\mu_i(t)^2}\right\} = \frac{1}{n}\sum_{i=1}^{n}\nabla^2 g_i(t). \qquad (27)\end{aligned}$$

Based on (27), we can see that the second derivative is positive since the $\mu_{i,j,k}(t)$ are all positive. It can also be verified that

$$\nabla^3 g(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j,k}\sum_{s,t}(b_{i,j,k}-b_{i,s,t})^2\mu_{i,j,k}(t)\mu_{i,s,t}(t)\left[\sum_{l,m}(b_{i,j,k}+b_{i,s,t}-2b_{i,l,m})\mu_{i,l,m}(t)\right]}{2\mu_i(t)^3}\right\},$$

so that using the same approach from Tran-Dinh et al. (2015), we see

$$\begin{aligned}
|\nabla^3 g(t)| &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\left\{\frac{\sum_{j,k}\sum_{s,t}(b_{i,j,k}-b_{i,s,t})^2\mu_{i,j,k}(t)\mu_{i,s,t}(t)\left[\sum_{l,m}(b_{i,j,k}+b_{i,s,t}-2b_{i,l,m})\mu_{i,l,m}(t)\right]}{2\mu_i(t)^3}\right\}\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j,k}\sum_{s,t}(b_{i,j,k}-b_{i,s,t})^2\mu_{i,j,k}(t)\mu_{i,s,t}(t)\left[\sum_{l,m}\mu_{i,l,m}(t)\sqrt{6(b_{i,j,k}^2+b_{i,s,t}^2+b_{i,l,m}^2)}\right]}{2\mu_i(t)^3}\right\}
\end{aligned}$$

so that taking $b_i = (b_{i,1,1},\ldots,b_{i,J,K})' \in \mathbb{R}^{JK}$, the previous inequality implies

$$\begin{aligned}
|\nabla^3 g(t)| &\leq \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j,k}\sum_{s,t}(b_{i,j,k}-b_{i,s,t})^2\mu_{i,j,k}(t)\mu_{i,s,t}(t)\left[\sqrt{6}\|b_i\|_2\sum_{l,m}\mu_{i,l,m}(t)\right]}{2\mu_i(t)^3}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\sqrt{6}\|b_i\|_2\left\{\frac{\sum_{j,k}\sum_{s,t}(b_{i,j,k}-b_{i,s,t})^2\mu_{i,j,k}(t)\mu_{i,s,t}(t)}{2\mu_i(t)^2}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\sqrt{6}\|b_i\|_2\nabla^2 g_i(t) = \frac{\sqrt{6}}{n}\sum_{i=1}^{n}\|x_i'B\|_2\nabla^2 g_i(t) \leq \frac{\sqrt{6}}{n}\sum_{i=1}^{n}\|X_{i,:}\|_2\|B\|_F\nabla^2 g_i(t) \\
&\leq \sqrt{6}\max_{i\in[n]}\|X_{i,:}\|_2\|B\|_F\left(\frac{1}{n}\sum_{i=1}^{n}\nabla^2 g_i(t)\right) = \sqrt{6}\max_{i\in[n]}\|X_{i,:}\|_2\|B\|_F\nabla^2 g(t)
\end{aligned}$$

and thus, with $R = \sqrt{6}\max_{i\in[n]}\|X_{i,:}\|_2$, we have the desired result

$$|\nabla^3 g(t)| \leq R\|B\|_F\nabla^2 g(t). \quad \blacksquare$$

We prove Lemma 4 after Lemma 5 since it relies on arguments outlined in Lemma 5.

Proof of Lemma 5. First, we define the set $\mathcal{B}_{\epsilon,\phi} = \left\{\Delta \in \mathbb{R}^{p\times JK} : \|\Delta\|_F = \epsilon, \Delta \in \mathbb{C}(\mathcal{S},\phi)\right\}$ and the function $H(\Delta) = \mathcal{F}_{\lambda,\gamma}(\beta^\dagger+\Delta) - \mathcal{F}_{\lambda,\gamma}(\beta^\dagger)$. Following the same argument as in Molstad and Rothman (2018), since the objective function in (6), $\mathcal{F}_{\lambda,\gamma}$, is convex and because $\hat{\beta}$ is its global minimizer, as long as $\gamma > \phi_1\|\nabla\tilde{\mathcal{G}}(\beta^\dagger)\|_{\infty,2}$, we know that $\inf\{H(\Delta) : \Delta \in \mathcal{B}_{\epsilon,\phi}\} > 0$ implies $\|\hat{\beta}-\beta^\dagger\|_F \leq \epsilon$. See Lemma 4 of Negahban et al. (2012) for a proof of this fact. Hence,

our goal is to show $H(\Delta) > 0$ for all $\Delta \in \mathcal{B}_{\epsilon,\phi}$ under the conditions of the lemma statement. First, we have

$$H(\Delta) = \underbrace{\mathcal{G}(\beta^\dagger + \Delta) - \mathcal{G}(\beta^\dagger)}_{T_1} + \underbrace{\gamma(\|\beta^\dagger + \Delta\|_{1,2} - \|\beta^\dagger\|_{1,2})}_{T_2} + \underbrace{\gamma\phi_2(\|\beta^\dagger D + \Delta D\|_{1,2} - \|\beta^\dagger D\|_{1,2})}_{T_3} \tag{28}$$

We begin by bounding T_1. Applying Lemma 3, using (26) and assumption A2, it follows that

$$\begin{aligned} T_1 &\geq \operatorname{tr}\left\{\Delta' \nabla \mathcal{G}(\beta^\dagger)\right\} + \frac{\operatorname{vec}(\Delta)' \nabla^2 \tilde{\mathcal{G}}(\beta^\dagger) \operatorname{vec}(\Delta)}{d_n^2 \|\Delta\|_F^2} \left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right) \\ &\geq -\|\Delta\|_{1,2} \|\nabla \mathcal{G}(\beta^\dagger)\|_{\infty,2} + \frac{\operatorname{vec}(\Delta)' \nabla^2 \tilde{\mathcal{G}}(\beta^\dagger) \operatorname{vec}(\Delta)}{d_n^2 \|\Delta\|_F^2} \left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right) \end{aligned} \tag{29}$$

where (29) follows from Hölder's inequality. Then, since $\Delta \in \mathcal{B}_{\epsilon,\phi}$ implies $\Delta \in \mathbb{C}(\mathcal{S}, \phi)$, by definition of $\kappa(\mathcal{S}, \phi)$, the inequality (29) implies

$$\begin{aligned} T_1 &\geq -\|\Delta\|_{1,2} \|\nabla \tilde{\mathcal{G}}(\beta^\dagger)\|_{\infty,2} + \frac{\kappa(\mathcal{S}, \phi)}{d_n^2} \left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right). \\ &\geq -\frac{\gamma}{\phi_1}\|\Delta\|_{1,2} + \frac{\kappa(\mathcal{S}, \phi)}{d_n^2} \left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right). \end{aligned} \tag{30}$$

where (30) holds because $\gamma > \phi_1 \|\nabla \mathcal{G}(\beta^\dagger)\|_{\infty,2}$ by assumption. Next, we bound T_2 and T_3. Recall that S_L, S_M, and S_I are sets of predictors where $\beta^\dagger_{S_L,:} \neq 0$, $\beta^\dagger_{S_M,:} \neq 0$, and $\beta^\dagger_{S_I,:} = 0$; $\beta^\dagger_{S_L,:} D \neq 0$, $\beta^\dagger_{S_M,:} D = 0$, and $\beta^\dagger_{S_I,:} D = 0$. By the triangle inequality, we have

$$\begin{aligned} T_2 &= \gamma(\|\beta^\dagger + \Delta\|_{1,2} - \|\beta^\dagger\|_{1,2}) \\ &= \gamma(\|\beta^\dagger_{S_L \cup S_M,:} + \Delta_{S_L \cup S_M,:}\|_{1,2} + \|\Delta_{S_I,:}\|_{1,2} - \|\beta^\dagger_{S_L \cup S_M,:}\|_{1,2}) \\ &\geq \gamma(\|\Delta_{S_I,:}\|_{1,2} - \|\Delta_{S_L \cup S_M,:}\|_{1,2}) \end{aligned}$$

Similarly, for T_3,

$$T_3 = \gamma\phi_2(\|\beta^\dagger D + \Delta D\|_{1,2} - \|\beta^\dagger D\|_{1,2}) \geq \gamma\phi_2(\|\Delta_{S_I \cup S_M,:} D\|_{1,2} - \|\Delta_{S_L,:} D\|_{1,2}).$$

Then, putting (30) together with the bounds for T_2 and T_3,

$$\begin{aligned} H(\Delta) &\geq -\frac{\gamma}{\phi_1}\|\Delta\|_{1,2} + \frac{\kappa(\mathcal{S},\phi)}{d_n^2}\left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right) + T_2 + T_3 \qquad (31)\\ &\geq -\frac{\gamma}{\phi_1}(\|\Delta_{S_I,:}\|_{1,2} + \|\Delta_{S_L\cup S_M,:}\|_{1,2}) + \frac{\kappa(\mathcal{S},\phi)}{d_n^2}\left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right)\\ &\qquad + \gamma(\|\Delta_{S_I,:}\|_{1,2} - \|\Delta_{S_L\cup S_M,:}\|_{1,2}) + T_3\\ &\geq \frac{\kappa(\mathcal{S},\phi)}{d_n^2}\left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right) - \frac{\gamma(\phi_1+1)}{\phi_1}\left(\|\Delta_{S_L\cup S_M,:}\|_{1,2}\right) + T_3. \end{aligned}$$

By plugging in the bound for T_3, this implies

$$\begin{aligned} H(\Delta) &\geq \frac{\kappa(\mathcal{S},\phi)}{d_n^2}\left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right) - \frac{\gamma(\phi_1+1)}{\phi_1}\left(\|\Delta_{S_L\cup S_M,:}\|_{1,2}\right)\\ &\qquad + \gamma\phi_2(\|\Delta_{S_I\cup S_M,:}D\|_{1,2} - \|\Delta_{S_L,:}D\|_{1,2})\\ &\geq \frac{\kappa(\mathcal{S},\phi)}{d_n^2}\left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right) - \frac{\gamma(\phi_1+1)}{\phi_1}\left(\|\Delta_{S_L\cup S_M,:}\|_{1,2}\right) - \gamma\phi_2\|\Delta_{S_L,:}D\|_{1,2}. \end{aligned}$$

Then, since $\Psi_{J,K}(S_L) = \sup_{M\neq 0, M\in\mathbb{R}^{p\times JK}} \|M_{S_L,:}D\|_{1,2}/\|M\|_F$, and using the fact that $\|\Delta_{S_L\cup S_M,:}\|_{1,2} \leq \sqrt{|S_L|+|S_M|}\|\Delta\|_F$, the previous inequality implies

$$H(\Delta) \geq \frac{\kappa(\mathcal{S},\phi)}{d_n^2}\left(e^{-d_n\|\Delta\|_F} + d_n\|\Delta\|_F - 1\right) - \gamma\|\Delta\|_F\left\{\frac{(\phi_1+1)}{\phi_1}\sqrt{|S_L|+|S_M|} + \phi_2\Psi_{J,K}(S_L)\right\}$$

so that for $\Delta \in \mathcal{B}_{\epsilon,\phi}$, i.e., $\|\Delta\|_F = \epsilon$ and $\Delta \in \mathbb{C}(\mathcal{S},\phi)$,

$$= \frac{\kappa(\mathcal{S},\phi)}{d_n^2}\left(e^{-d_n\epsilon} + d_n\epsilon - 1\right) - \gamma\epsilon\left\{\frac{(\phi_1+1)}{\phi_1}\sqrt{|S_L|+|S_M|} + \phi_2\Psi_{J,K}(S_L)\right\}.$$

Thus, for constant $c > 2$, with

$$\gamma = \frac{\phi_1\epsilon\kappa(\mathcal{S},\phi)}{c\{(\phi_1+1)\sqrt{|S_L|+|S_M|} + \phi_1\phi_2\Psi_{J,K}(S_L)\}},$$

it follows that

$$H(\Delta) \geq \frac{\kappa(\mathcal{S},\phi)}{d_n^2}\left(e^{-d_n\epsilon} + d_n\epsilon - 1\right) - \frac{\kappa(\mathcal{S},\phi)d_n^2}{d_n^2 c}\epsilon^2 = \frac{\kappa(\mathcal{S},\phi)}{d_n^2}\left(e^{-d_n\epsilon} + d_n\epsilon - \frac{d_n^2\epsilon^2}{c} - 1\right)$$

so that for ϵ sufficiently close to zero,

$$\left(e^{-d_n\epsilon} + d_n\epsilon - \frac{d_n^2\epsilon^2}{c} - 1\right) > 0,$$

which yields the desired result. ■

Proof of Lemma 4. Note that letting $\hat{\Delta} = \hat{\beta} - \beta^\dagger$, we know that $H(\hat{\Delta})$ as defined in (28) is non-positive. Hence, because $e^{-x} + x - 1 > 0$ for all $x > 0$, by the arguments used to obtain (31),

$$\begin{aligned} 0 \geq H(\hat{\Delta}) \geq -\frac{\gamma}{\phi_1}(\|\hat{\Delta}_{S_I,:}\|_{1,2} + \|\hat{\Delta}_{S_L \cup S_M,:}\|_{1,2}) + \gamma(\|\hat{\Delta}_{S_I,:}\|_{1,2} - \|\hat{\Delta}_{S_L \cup S_M,:}\|_{1,2}) \\ + \gamma\phi_2(\|\hat{\Delta}_{S_I \cup S_M,:}D\|_{1,2} - \|\hat{\Delta}_{S_L,:}D\|_{1,2}) \end{aligned}$$

which implies

$$0 \geq \frac{(\phi_1 - 1)}{\phi_1}\|\hat{\Delta}_{S_I,:}\|_2 - \frac{(\phi_1 + 1)}{\phi_1}\|\hat{\Delta}_{S_L \cup S_M,:}\|_{1,2} + \phi_2(\|\hat{\Delta}_{S_I \cup S_M,:}D\|_{1,2} - \|\hat{\Delta}_{S_L,:}D\|_{1,2})$$

so that

$$\frac{(\phi_1 + 1)}{\phi_1}\|\hat{\Delta}_{S_L \cup S_M,:}\|_{1,2} + \phi_2\|\hat{\Delta}_{S_L,:}D\|_{1,2} \geq \frac{(\phi_1 - 1)}{\phi_1}\|\hat{\Delta}_{S_I,:}\|_{1,2} + \phi_2\|\hat{\Delta}_{S_I \cup S_M,:}D\|_{1,2},$$

the desired result. ■

We prove Lemma 6 below. First, we state an important inequality which is key to our proof.

McDiarmid's Inequality. *Let $X_1, \ldots, X_n$ be independent random variables each taking values in the set $\mathcal{X}$. Let $f : \mathcal{X} \times \cdots \times \mathcal{X} \to \mathbb{R}$. If for each $i \in [n]$, the function f satisfies*

$$|f(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n) - f(X_1, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n)| \leq c_i$$

for all $(X_1, \ldots, X_n)$ and any $\tilde{X}_i \in \mathcal{X}$, then, for every $\epsilon > 0$,

$$P\left\{f(X_1, \ldots, X_n) \geq \mathbb{E}f(X_1, \ldots, X_n) + \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

We are now ready to prove Lemma 6.

Proof of Lemma 6. First, notice that $\nabla\tilde{\mathcal{G}}(\beta^\dagger) = n^{-1}X'W$ where $X = (x_1, \ldots, x_n)' \in \mathbb{R}^{n\times p}$ and the ith row of W, $W_{i,:} \in \mathbb{R}^{JK}$, can be expressed $W_{i,:} = \text{vec}\{\pi^*(x_i)\} - \text{vec}(\mathcal{Y}_i)$ for $i \in [n]$. To simplify notation, we will let $v_i = \text{vec}(\mathcal{Y}_i) \in \mathbb{R}^{JK}$ and $\pi_i^* = \text{vec}\{\pi^*(x_i)\} = (\pi^*_{1,1}(x_i), \ldots, \pi^*_{J,K}(x_i))' \in \mathbb{R}^{JK}$. We will use $v_{i,j}$ denote the jth element of v_i and similarly for π_i^* so that $W_{i,j} = v_{i,j} - \pi^*_{i,j}$ for each $j \in [JK]$. Note that under A1, each $W_{i,:}$ is independent but not identically distributed.

Our objective is to find a γ such that with high probability

$$P\left(\frac{1}{n}\left\|X'W\right\|_{\infty,2} \leq \gamma\right).$$

Starting with the union bound, we have

$$P\left(\frac{1}{n}\left\|X'W\right\|_{\infty,2} \leq \gamma\right) = 1 - P\left(\frac{1}{n}\max_{j\in[p]}\left\|W'X_{:,j}\right\|_2 > \gamma\right) \geq 1 - \sum_{j=1}^{p} P\left(\frac{1}{n}\left\|W'X_{:,j}\right\|_2 > \gamma\right). \tag{32}$$

To bound the probability in the final term, we apply McDiarmid's inequality. We first establish the component-wise deviation bound c_i. Notice, taking $f(W_1, \ldots, W_n) = \|W'X_{:,j}\|_2/n$, we have that for any pair $(W_{i,:}, \tilde{W}_{i,:})$ letting $\tilde{W}$ denote W with ith row replaced with $\tilde{W}_{i,:}$,

$$\left|\left\|W'X_{:,j}\right\|_2 - \left\|\tilde{W}'X_{:,j}\right\|_2\right| \leq \left\|(W-\tilde{W})'X_{:,j}\right\|_2$$

by the reverse triangle inequality. Then, because $W_{k,:} = \tilde{W}_{k,:}$ for all $k \neq i$,

$$\left\|(W-\tilde{W})'X_{:,j}\right\|_2 = \sqrt{x_{i,j}^2\sum_{l=1}^{JK}\left(\pi^*_{i,l} - v_{i,l} - \pi^*_{i,l} + \tilde{v}_{i,l}\right)^2} = \sqrt{x_{i,j}^2\sum_{l=1}^{JK}\left(\tilde{v}_{i,l} - v_{i,l}\right)^2} \leq \sqrt{2}|x_{i,j}|$$

since v_i and $\tilde{v}_i$ differ by one in at most two coordinates by definition (since each $\mathcal{Y}_i$ can have only one component equal to one and all others equal to zero). Hence, for each $i \in [n]$, we have

$$\left|f(W_{1,:}, \ldots, W_{i,:}, \ldots, W_{n,:}) - f(W_{1,:}, \ldots, \tilde{W}_{i,:}, \ldots, W_{n,:})\right| \leq \frac{\sqrt{2}|x_{i,j}|}{n}$$

Therefore, by McDiarmid's inequality,

$$P\left(\frac{1}{n}\left\|W'X_{:,j}\right\|_2 \geq \frac{1}{n}\mathbb{E}\left\|W'X_{:,j}\right\|_2 + \epsilon\right) \leq \exp\left(\frac{-2n^2\epsilon^2}{2\sum_{i=1}^n x_{i,j}^2}\right) \leq \exp\left(-n\epsilon^2\right),$$

where the second inequality follows from $\sum_{i=1}^n x_{i,j}^2 \leq n$, i.e., assumption A2. It remains only

to bound the expectation. Notice,

$$\mathbb{E}\|W'X_{:,j}\|_2 = \mathbb{E}\sqrt{\sum_{l=1}^{JK}\left\{\sum_{i=1}^{n} x_{i,j}\left(\pi^*_{i,l} - v_{i,l}\right)\right\}^2} \leq \sqrt{\sum_{l=1}^{JK}\mathbb{E}\left[\left\{\sum_{i=1}^{n} x_{i,j}\left(\pi^*_{i,l} - v_{i,l}\right)\right\}^2\right]}$$

by Jensen's inequality. Furthermore, letting $\mathbb{V}$ denote the variance, each term under the rightmost square-root can be bounded since

$$\begin{aligned}\mathbb{E}\left[\left\{\sum_{i=1}^{n} x_{i,j}\left(\pi^*_{i,l} - v_{i,l}\right)\right\}^2\right] &= \mathbb{V}\left\{\sum_{i=1}^{n} x_{i,j}\left(\pi^*_{i,l} - v_{i,l}\right)\right\} + \left[\mathbb{E}\left\{\sum_{i=1}^{n} x_{i,j}\left(\pi^*_{il} - v_{i,l}\right)\right\}\right]^2 \\ &= \mathbb{V}\left\{\sum_{i=1}^{n} x_{i,j}\left(\pi^*_{i,l} - v_{i,l}\right)\right\} = \sum_{i=1}^{n} x_{i,j}^2\mathbb{V}(v_{i,l}) \leq \frac{1}{4}\sum_{i=1}^{n} x_{i,j}^2 \leq \frac{n}{4}\end{aligned}$$

since $n^{-1}\mathbb{E}(v_{i,l}) = \pi^*_{i,l}$, $\mathbb{V}(v_{i,l}) = \pi^*_{il}(1-\pi^*_{il}) \leq 1/4$ and $\sum_{i=1}^{n} x_{i,j}^2 \leq n$ by assumption A2. Therefore, we have that $n^{-1}\mathbb{E}\left\|W'X_{:,j}\right\|_2 \leq \{JK/(4n)\}^{1/2}$ and thus

$$P\left(\frac{1}{n}\left\|W'X_{:,j}\right\|_2 \geq \sqrt{\frac{JK}{4n}} + \epsilon\right) \leq \exp\left(-n\epsilon^2\right),$$

so that taking $\epsilon = \{\log(p/\alpha)/n\}^{1/2}$, it follows from (32) that

$$P\left(\|\nabla\tilde{\mathcal{G}}(\beta^\dagger)\|_{\infty,2} \leq \sqrt{\frac{JK}{4n}} + \sqrt{\frac{\log(p/\alpha)}{n}}\right) \geq 1 - p\exp\left(-\frac{n\log(p/\alpha)}{n}\right) = 1-\alpha. \qquad \blacksquare$$

F.3 Proofs of Corollaries and Remarks

Proof of Remark 1 By definition, $\Psi_{J,K}(S) = \sup_{M\in\mathbb{R}^{p\times JK}, M\neq 0}\frac{\|M_{S,:}D\|_{1,2}}{\|M\|_F}$. Recall that $M_{S,:}$ is the submatrix of M containing only rows whose indices belong to the set S. By the Cauchy-Schwarz inequality,

$$\|M_{S,:}D\|_{1,2} = \sum_{j=1}^{p}\mathbf{1}(j\in S)\|M_{j,:}D\|_2 \leq \sqrt{\sum_{j=1}^{p}\mathbf{1}(j\in S)^2}\sqrt{\sum_{j=1}^{p}\|M_{j,:}D\|_2^2} = \sqrt{|S|}\|MD\|_F.$$

Thus,

$$\sup_{M\in\mathbb{R}^{p\times JK},M\neq 0}\frac{\|M_{S,:}D\|_{1,2}}{\|M\|_F}\leq \sup_{M\in\mathbb{R}^{p\times JK},M\neq 0}\frac{\sqrt{|S|}\|MD\|_F}{\|M\|_F}$$
$$=\sup_{U\in\mathbb{R}^{p\times JK},\|U\|_F=1}\sqrt{|S|}\|UD\|_F=\sup_{U\in\mathbb{R}^{p\times JK},\|U\|_F=1}\sqrt{|S|\mathrm{tr}(UDD'U')}.$$

Letting $U_{j,:}\in\mathbb{R}^{JK}$ be the jth row of U; and letting $\varphi_1(DD')$ be the largest eigenvalue of DD', we have

$$\Psi_{J,K}(S)\leq\sup_{\|U\|_F=1}\sqrt{|S|\sum_{j=1}^{p}U'_{j,:}(DD')U_{j,:}}\leq\sup_{\|U\|_F=1}\sqrt{|S|\varphi_1(DD')\sum_{j=1}^{p}U'_{j,:}U_{j,:}}=\sqrt{|S|\varphi_1(DD')}.$$

The result follows from the fact that $\varphi_1(DD')=JK$ for all J and K. ■

Proof of Corollary 1. As before, let $\hat{\Delta}=\hat{\beta}-\beta^\dagger$. We know that by definition of the disjoint sets S_I, S_L, and S_M,

$$\|\hat{\beta}-\beta^\dagger\|_{1,2}=\|\hat{\Delta}\|_{1,2}=\|\hat{\Delta}_{S_I,:}\|_{1,2}+\|\hat{\Delta}_{S_L\cup S_M,:}\|_{1,2}. \tag{33}$$

Lemma 4 ensures that on the event $\gamma>\phi_1\|\nabla\tilde{\mathcal{G}}(\beta^\dagger)\|_{\infty,2}$, $\hat{\Delta}\in\mathbb{C}(\mathcal{S},\phi)$, so $\gamma>\phi_1\|\nabla\tilde{\mathcal{G}}(\beta^\dagger)\|_{\infty,2}$ equivalently implies (after some algebra)

$$\|\hat{\Delta}_{S_I,:}\|_{1,2}\leq\frac{(\phi_1+1)\|\hat{\Delta}_{S_L\cup S_M,:}\|_{1,2}+\phi_1\phi_2(\|\hat{\Delta}_{S_L,:}D\|_{1,2}-\|\hat{\Delta}_{S_I\cup S_M,:}D\|_{1,2})}{\phi_1-1}$$
$$\leq\frac{(\phi_1+1)\|\hat{\Delta}_{S_L\cup S_M,:}\|_{1,2}+\phi_1\phi_2\|\hat{\Delta}_{S_L,:}D\|_{1,2}}{\phi_1-1}. \tag{34}$$

Thus, by (33) and (34), we have

$$\|\hat{\beta}-\beta^\dagger\|_{1,2}=\|\hat{\Delta}_{S_I,:}\|_{1,2}+\|\hat{\Delta}_{S_L\cup S_M,:}\|_{1,2}$$
$$\leq\frac{(\phi_1+1)\|\hat{\Delta}_{S_L\cup S_M,:}\|_{1,2}+\phi_1\phi_2\|\hat{\Delta}_{S_L,:}D\|_{1,2}}{\phi_1-1}+\frac{\phi_1-1}{\phi_1-1}\|\hat{\Delta}_{S_L\cup S_M,:}\|_{1,2}$$
$$\leq\frac{2\phi_1\|\hat{\Delta}_{S_L\cup S_M,:}\|_{1,2}+\phi_1\phi_2\|\hat{\Delta}_{S_L,:}D\|_{1,2}}{\phi_1-1}$$
$$\leq\frac{2\phi_1\sqrt{|S_L|+|S_M|}\|\hat{\Delta}\|_F+\phi_1\phi_2\Psi_{J,K}(S_L)\|\hat{\Delta}\|_F}{\phi_1-1}$$

so that the previous inequality finally implies

$$\|\hat{\beta} - \beta^{\dagger}\|_{1,2} \leq \left\{ \frac{2\phi_1 \sqrt{|S_L| + |S_M|} + \phi_1 \phi_2 \Psi_{J,K}(S_L)}{\phi_1 - 1} \right\} \|\hat{\Delta}\|_F. \tag{35}$$

Since $\gamma > \phi_1 \|\nabla \tilde{\mathcal{G}}(\beta^{\dagger})\|_{\infty,2}$ implies both (35) and $\|\hat{\Delta}\|_F \leq \Phi_n$, the probability of

$$\|\hat{\beta} - \beta^{\dagger}\|_{1,2} \leq \left\{ \frac{2\phi_1 \sqrt{|S_L| + |S_M|} + \phi_1 \phi_2 \Psi_{J,K}(S_L)}{\phi_1 - 1} \right\} \Phi_n$$

is greater than or equal to the probability of $\gamma > \phi_1 \|\nabla \tilde{\mathcal{G}}(\beta^{\dagger})\|_{\infty,2}$, which under the specification in Theorem 1, occurs with probability at least $1 - \alpha$. ■

Proof of Corollary 2. The proof of Corollary 2 follows an identical series of arguments as the Proof of Theorem 1. We simply redefine $\beta \in \mathbb{R}^{p \times \check{K}}$ and $\Psi_{\{K_j\}_{j=1}^G}$ according to the appropriate D matrix. This modifies Condition 1, which depends on the $\Psi_{\{K_j\}_{j=1}^G}$, n, p, ϕ_1, ϕ_2, S_L and S_M; modifies $\mathbb{C}(\mathcal{S}, \phi)$; and modifies the restricted eigenvalue, which is based on the $p\check{K} \times p\check{K}$ Hessian of $\tilde{\mathcal{G}}$ with respect to the vectorization of its matrix-valued argument. Thus, all that is required is to determine the value of γ such that $\gamma > \phi_1 \|\nabla \mathcal{G}(\beta^{\dagger})\|_{\infty,2}$ for (scaled) negative log-likelihood $\tilde{\mathcal{G}} : \mathbb{R}^{p \times \check{K}} \to \mathbb{R}$. It is easy to see that modifying Lemma 6 would require only replacing $\sum_{l=1}^{JK}$ with $\sum_{l=1}^{\check{K}}$. Thus, by an identical set of arguments as those in the proof of Lemma 6, with $W \in \mathbb{R}^{n \times \check{K}}$, we would have that $\mathbb{E}\|W' X_{:,j}\|_2 / n \leq \{\check{K}/(4n)\}^{1/2}$, which implies

$$P\left(\|\nabla \tilde{\mathcal{G}}(\beta^{\dagger})\|_{\infty,2} \leq \sqrt{\frac{\check{K}}{4n}} + \sqrt{\frac{\log(p/\alpha)}{n}} \right) \geq 1 - \alpha.$$

Hence, applying Lemma 4 and 5 would lead to the stated conclusion. ■

G Additional details

G.1 Need for constraint matrix D

If instead of penalizing $\|D'\beta_{m,:}\|_2$, one penalized $\|\mathcal{D}_1'\beta_{m,:}\|_2$ or $\|\mathcal{D}_2'\beta_{m,:}\|_2$ (where $\mathcal{D}_1$ and $\mathcal{D}_2$ correspond to different minimal sets of odds-ratios), the solution path (i.e., set of candidate models) would depend on which sets of odds ratios are encoded in the constraint matrices $\mathcal{D}_1$ and $\mathcal{D}_2$. This may be problematic because at many points along the solution path $\mathcal{D}_1'\beta_{m,:} \neq 0$, but the penalty will encourage $\mathcal{D}_1'\beta_{m,:}$ to be small in Euclidean norm. This may or may not correspond to $\mathcal{D}_2'\beta_{m,:}$ being small. For this reason, selecting one particular minimal set to construct $\mathcal{D}_1$ may favor estimates with certain log odds ratios being small (but non-zero), but

does not enforce (directly, at least) shrinkage of others. The use of D avoids this problem entirely: all log odds-ratios are shrunken to an equal degree.

Regarding the theory, the results would be effectively unchanged if we used some $\mathcal{D}$ instead of D. The sets S_I, S_L, S_M (and their cardinalities) would be no different: only $\mathbb{C}(\mathcal{S}, \phi)$ would have D replaced with $\mathcal{D}$. In addition, we would redefine $\Psi_{J,K}$ with $\mathcal{D}$ replacing D in the numerator. However, the bound in Remark 1 would not be improved by replacing D with $\mathcal{D}$. Examining the proof of Remark 1, it can be seen that the bound depends on the largest eigenvalue of DD' (or $\mathcal{D}\mathcal{D}'$). It can be verified that in both cases, this is equal to JK.[2]

G.2 Explicit form of $\beta^\dagger$

Consider that for any $\beta \in \mathcal{F}_\pi$, the matrix $\beta_a = \beta - a1'_{JK}$ also belongs to $\mathcal{F}_\pi$ for any $a \in \mathbb{R}^p$. Hence, given any $\beta \in \mathcal{F}_\pi$ (i.e., any β which leads to the "true" probabilities), our definition of $\beta^\dagger$ can be expressed

$$\beta^\dagger = \beta - \tilde{a}1'_{JK}, \qquad \text{where} \qquad \tilde{a} = \arg\min_{a \in \mathbb{R}^p} \|\beta - a1'_{JK}\|_{1,2}.$$

Fortunately, we can find an explicit form for $\tilde{a}$. Notice

$$\tilde{a} = \arg\min_{a \in \mathbb{R}^p} \|\beta - a1'_{JK}\|_{1,2} = \arg\min_{a \in \mathbb{R}^p} \sum_{j=1}^{p} \|\beta_{j,:} - a_j 1_{JK}\|_2$$

so that the jth element of $\tilde{a}$ is given by

$$\tilde{a}_j = \arg\min_{a_j \in \mathbb{R}} \|\beta_{j,:} - a_j 1_{JK}\|_2 = \arg\min_{a_j \in \mathbb{R}} \|\beta_{j,:} - a_j 1_{JK}\|_2^2$$

from which we can easily see that $\tilde{a}_j = (JK)^{-1} \sum_{m=1}^{JK} \beta_{j,m}$. This reveals that given any $\beta \in \mathcal{F}_\pi$, $\beta^\dagger = \beta - (\beta 1_{JK}/JK)1'_{JK}$, i.e., $\beta^\dagger$ is simply the version of β with row-wise average zero, which is uniquely defined for a particular $\mathcal{F}_\pi$ (and easily computed given any $\beta \in \mathcal{F}_\pi$).

G.3 More than one replicate per subject

At the suggestion of a referee, we explored the effects of additional replicates on the theoretical results from Section 3. Here, we prove that additional replicates (with the number of unique

[2]The largest eigenvalues of DD' and $\mathcal{D}\mathcal{D}'$ match, but the second through $(J-1)(K-1)$th largest eigenvalues do not. For DD' in the bivariate response case, these eigenvalues are equal to the largest: this is not true of $\mathcal{D}\mathcal{D}'$.

subjects in the dataset fixed) can improve the error bound. Specifically, we show the restricted eigenvalue condition is always more plausible (in a sense to be described momentarily) with additional replicates than it is for a dataset with the same number of distinct subjects[3], but each having a single replicate.

Lemma 7. *Let $\kappa(\mathcal{S}, \phi)$ be the restricted eigenvalue for a dataset with $n_i = 1$ for all $i \in [n]$. Let $\ddot{\kappa}(\mathcal{S}, \phi)$ be the restricted eigenvalue for the same dataset with the same n subjects and at least one subject having more than one replicate, i.e., $n_i \geq 2$ for at least one $i \in [n]$. Then $\ddot{\kappa}(\mathcal{S}, \phi) \geq \kappa(\mathcal{S}, \phi)$ almost surely.*

Proof of Lemma 7. Recall that the restricted eigenvalue is defined as

$$\kappa(\mathcal{S}, \phi) = \inf_{\Delta \in \mathbb{C}(\mathcal{S}, \phi)} \frac{\text{vec}(\Delta)' \nabla^2 \tilde{\mathcal{G}}(\beta^\dagger) \text{vec}(\Delta)}{\|\Delta\|_F^2},$$

where

$$\mathbb{C}(\mathcal{S}, \phi) = \big\{ \Delta \in \mathbb{R}^{p \times JK} : \Delta \neq 0, (\phi_1 + 1)\|\Delta_{S_L \cup S_M, :}\|_{1,2} + \phi_1 \phi_2 \|\Delta_{S_L, :} D\|_{1,2} \geq (\phi_1 - 1)\|\Delta_{S_I, :}\|_{1,2} + \phi_1 \phi_2 \|\Delta_{S_I \cup S_M, :} D\|_{1,2} \big\}.$$

Note first that for a dataset with $n_i = 1$ for all $i \in [n]$

$$\nabla^2 \tilde{\mathcal{G}}(\beta^\dagger) = n^{-1} \sum_{i=1}^{n} \{ P^*_{\beta^\dagger}(x_i) \otimes x_i x_i' \}$$

where letting $\tilde{\pi}^*_{i, f(j,k)} = \pi^*_{j,k}(x_i)$,

$$P^*_{\beta^\dagger}(x_i) = \begin{pmatrix} \tilde{\pi}^*_{i,f(1,1)}(1 - \tilde{\pi}^*_{i,f(1,1)}) & -\tilde{\pi}^*_{i,f(1,1)} \tilde{\pi}^*_{i,f(2,1)} & \cdots & \cdots & -\tilde{\pi}^*_{i,f(1,1)} \tilde{\pi}^*_{i,f(J,K)} \\ -\tilde{\pi}^*_{i,f(2,1)} \tilde{\pi}^*_{i,f(1,1)} & \tilde{\pi}^*_{i,f(2,1)}(1 - \tilde{\pi}^*_{i,f(2,1)}) & -\tilde{\pi}^*_{i,f(2,1)} \tilde{\pi}^*_{i,f(3,1)} & \cdots & -\tilde{\pi}^*_{i,f(2,1)} \tilde{\pi}^*_{i,f(J,K)} \\ \vdots & \cdots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ -\tilde{\pi}^*_{i,f(J,K)} \tilde{\pi}^*_{i,f(1,1)} & -\tilde{\pi}^*_{i,f(J,K)} \tilde{\pi}^*_{i,f(2,1)} & \cdots & \cdots & \tilde{\pi}^*_{i,f(J,K)}(1 - \tilde{\pi}^*_{i,f(J,K)}) \end{pmatrix} \in \mathbb{R}^{JK \times JK}.$$

If we observe n_j replicates for the jth subject, we could express the Hessian for the (scaled)

[3]By "distinct subjects", we mean subjects who have distinct measured predictors.

negative log-likelihood, denoted $\ddot{\tilde{\mathcal{G}}}$, as

$$\begin{aligned}\nabla^2\ddot{\tilde{\mathcal{G}}}(\beta^\dagger) &= n^{-1}\sum_{i=1}^{n}\left[\sum_{j=1}^{n_i}\{P^*_{\beta^\dagger}(x_i)\otimes x_i x_i'\}\right] \\ &= \underbrace{n^{-1}\sum_{i=1}^{n}\{P^*_{\beta^\dagger}(x_i)\otimes x_i x_i'\}}_{\nabla^2\tilde{\mathcal{G}}(\beta^\dagger)} + \underbrace{n^{-1}\sum_{i=1}^{n}(n_i-1)\{P^*_{\beta^\dagger}(x_i)\otimes x_i x_i'\}}_{Q}\end{aligned}$$

where $\tilde{\mathcal{G}}$ is the (scaled) negative log-likelihood for the dataset with $n_i = 1$ for all $i \in [n]$. Of course, Q is symmetric and non-negative definite so that that

$$\begin{aligned}\ddot{\kappa}(\mathcal{S},\phi) &= \inf_{u\in\mathbb{C}(\mathcal{S},\phi)}\frac{\mathrm{vec}(u)'\nabla^2\ddot{\tilde{\mathcal{G}}}(\beta^\dagger)\mathrm{vec}(u)}{\|u\|_F^2} \\ &= \inf_{u\in\mathbb{C}(\mathcal{S},\phi)}\frac{\mathrm{vec}(u)'\{\nabla^2\tilde{\mathcal{G}}(\beta^\dagger)+Q\}\mathrm{vec}(u)}{\|u\|_F^2} \\ &= \inf_{u\in\mathbb{C}(\mathcal{S},\phi)}\left\{\frac{\mathrm{vec}(u)'\nabla^2\tilde{\mathcal{G}}(\beta^\dagger)\mathrm{vec}(u)}{\|u\|_F^2}+\frac{\mathrm{vec}(u)'Q\mathrm{vec}(u)}{\|u\|_F^2}\right\} \\ &\geq \inf_{u\in\mathbb{C}(\mathcal{S},\phi)}\frac{\mathrm{vec}(u)'\nabla^2\tilde{\mathcal{G}}(\beta^\dagger)\mathrm{vec}(u)}{\|u\|_F^2}+\inf_{w\in\mathbb{C}(\mathcal{S},\phi)}\frac{\mathrm{vec}(w)'Q\mathrm{vec}(w)}{\|w\|_F^2}\end{aligned}$$

and since $\nu'Q\nu \geq 0$ for all unit vectors ν, the previous inequality implies

$$\ddot{\kappa}(\mathcal{S},\phi) \geq \inf_{u\in\mathbb{C}(\mathcal{S},\phi)}\frac{\mathrm{vec}(u)'\nabla^2\tilde{\mathcal{G}}(\beta^\dagger)\mathrm{vec}(u)}{\|u\|_F^2} = \kappa(\mathcal{S},\phi)$$

from which the conclusion follows. ■

However, we caution against this result being interpreted as "having few subjects with many replicates is better than more subjects with fewer replicates". In the $n_i > 1$ case, X would consist of duplicated rows. In general, duplicated rows lead to a lower rank $\nabla^2\tilde{\mathcal{G}}(\beta^\dagger)$ (relative to a version of X of the same dimension with entirely distinct rows), which in turn leads to a smaller restricted eigenvalue and thus, worse error bound.

Hence, if one dataset has X with n rows based on n_1 distinct subjects and another dataset has X of the same dimension based on n_2 ($n_2 > n_1$) distinct subjects, we would expect that the restricted eigenvalue condition would be more plausible for the latter dataset, in general. That is to say, there is a tradeoff between the benefit of replicates and the number of distinct subjects in a dataset. More replicates are beneficial (as Lemma 7 reveals), but not at the expense of more distinct subjects in the dataset.

G.4 Additional computational details for competitors

Here, we very briefly discuss how we compute `OG-Mult` and `LG-Mult`. As discussed in the main manuscript, for both we use an accelerated proximal gradient descent algorithm. In each step of both algorithms, we must solve the respective proximal operators for the two penalties. For the overlapping group penalty, we use the algorithm proposed by Yuan et al. (2013). In brief, this is an iterative procedure which solves the dual of the proximal operator via accelerated gradient descent. For the latent-group lasso penalty, we use a blockwise coordinate descent algorithm to solve the proximal operator (e.g., Algorithm 2 of Yan and Bien (2017)).

G.5 Candidate tuning parameters

In this section, we discuss the construction of the set of candidate tuning parameters for `LO-Mult`. For the remainder of this discussion, let $\hat{\beta}_{\lambda,\gamma}$ denote the minimizer of (6) with tuning parameters (λ, γ) and recall that $\|A\|_{\infty,2} = \max_j \|A_{j,:}\|_2$ for a matrix A.

First, we pre-specify a set of candidate λ: we found that $\lambda \in [10^{-4}, 10^{-1}]$ covered all interesting models (i.e., those with smallest cross-validation error) across all the settings we considered. As a default, we suggest $\lambda \in \{10^x : x \in \{-4, -3.75, -3.50, -3.25, \ldots, -1\}\}$. Then, to determine a set of candidate γ, we use the fact that if $\hat{\beta}_{0,\gamma} = (\tilde{\beta}_0, 0_{JK\times p-1})'$ (where $\tilde{\beta}_0 \in \mathbb{R}^{JK}$ is the unpenalized maximum likelihood estimator from the intercept only model) for a particular γ, then $\hat{\beta}_{\lambda,\gamma} = (\tilde{\beta}_0, 0_{JK\times p-1})'$ for that same γ for any $\lambda > 0$. To simplify notation, let $\hat{\beta}_{0,\infty} = (\tilde{\beta}_0, 0_{JK\times p-1})'$. Based on the first-order optimality conditions for $\hat{\beta}_{\lambda,\gamma}$, it can be checked that if $\gamma \geq \|\nabla\tilde{\mathcal{G}}(\hat{\beta}_{0,\infty})\|_{\infty,2}$ then $\hat{\beta}_{\lambda,\gamma} = (\tilde{\beta}_0, 0_{JK\times p-1})'$ for all λ. Thus, we first compute $\gamma_{\max} = \|\nabla\tilde{\mathcal{G}}(\hat{\beta}_{0,\infty})\|_{\infty,2}$, and then consider candidate set $\gamma \in [\delta\gamma_{\max}, \gamma_{\max}]$ (equally spaced on the log-base-2 scale) where $\delta < 1$. In our simulation studies, we found $\delta = 0.05$ worked well. In practice, we suggest a user try a larger value of δ with fewer candidate γ, then based on the cross-validation errors, refine δ and rerun with more candidate γ values.

H Semi-supervised categorical response regression

In practice, when there are multiple categorical responses variables, it is often the case that one or more are costly or difficult to observe. To address these situations, we extend our method to settings where some response variables are missing or unobserved. As before, we focus on the bivariate categorical response regression model, but our developments can be generalized to three or more categorical response variables as will be discussed in a subsequent section.

Throughout this section, let $y_{(1)i} \in \mathbb{R}^J$ and $y_{(2)i} \in \mathbb{R}^K$ denote the observed response category counts for ith subject's first and second response variables, respectively (treating all responses as completely observed). As before, we assume that $n_i = 1$ for each $i \in [n]$ for simplicity. Let $(\mathcal{L}_1, \mathcal{U}_1)$ and $(\mathcal{L}_2, \mathcal{U}_2)$ be pairs of partitions of $[n]$ where $i \in \mathcal{L}_k$ if $y_{(k)i}$ is observed and $i \in \mathcal{U}_k$ if $y_{(k)i}$ is unobserved for $(i, k) \in [n] \times \{1, 2\}$. Then, the observed data

negative log-likelihood (divided by n) is given by

$$\mathcal{G}_{\mathcal{U},\mathcal{L}}(\boldsymbol{\beta}) = -\frac{1}{n}\left[\sum_{i\in\mathcal{L}_1\cap\mathcal{L}_2}\log\left\{\sum_{j,k}\frac{\exp\left(x_i'\boldsymbol{\beta}_{:,j,k}\right)y_{(1)i,j}y_{(2)i,k}}{\sum_{s,t}\exp\left(x_i'\boldsymbol{\beta}_{:,s,t}\right)}\right\} + \sum_{i\in\mathcal{L}_1\cap\mathcal{U}_2}\log\left\{\sum_{j,k}\frac{\exp\left(x_i'\boldsymbol{\beta}_{:,j,k}\right)y_{(1)i,j}}{\sum_{s,t}\exp\left(x_i'\boldsymbol{\beta}_{:,s,t}\right)}\right\} + \sum_{i\in\mathcal{U}_1\cap\mathcal{L}_2}\log\left\{\sum_{j,k}\frac{\exp\left(x'\boldsymbol{\beta}_{:,j,k}\right)y_{(2)i,k}}{\sum_{s,t}\exp\left(x'\boldsymbol{\beta}_{:,s,t}\right)}\right\}\right].$$

The observed data likelihood consists of the joint probability mass function for subjects with both responses observed, and the marginal probability mass function for those with only one of the two responses observed.

To fit the multivariate multinomial logistic regression model with partially unobserved responses, we propose to minimize a penalized version of $\mathcal{G}_{\mathcal{U},\mathcal{L}}$ using the penalties motivated in Section 2

$$\underset{\beta\in\mathbb{R}^{p\times JK}}{\arg\min}\left\{\mathcal{G}_{\mathcal{U},\mathcal{L}}(\boldsymbol{\beta}) + \lambda\sum_{m=2}^{p}\|D'\beta_{m,:}\|_2 + \gamma\sum_{m=2}^{p}\|\beta_{m,:}\|_2\right\}. \tag{36}$$

Fortunately, we need not resort to an expectation-maximization algorithm to compute (36). In fact, we can solve this (possibly non-convex) optimization problem directly using a modified version of the monotone accelerated proximal gradient descent proposed in Li and Lin (2015). Specifically, we will need to compute the gradient of $\tilde{\mathcal{G}}_{\mathcal{U},\mathcal{L}}$, the version of $\mathcal{G}_{\mathcal{U},\mathcal{L}}$ taking a matrix-valued input. The gradient of $\tilde{\mathcal{G}}_{\mathcal{U},\mathcal{L}}$ can be expressed $\nabla\tilde{\mathcal{G}}_{\mathcal{U},\mathcal{L}}(\beta^{(t)}) = n^{-1}X'W_{\mathcal{L},\mathcal{U}}(\beta^{(t)})$ where $W_{\mathcal{L},\mathcal{U}}(\beta^{(t)})$ has entries

$$[W_{\mathcal{L},\mathcal{U}}(\beta^{(t)})]_{i,f(j,k)} = \begin{cases} \pi^{(t)}_{i,j,k} - y_{(1)i,j}y_{(2)i,k} & : i\in\mathcal{L}_1\cap\mathcal{L}_2 \\ \pi^{(t)}_{i,j,k}(1-y_{(1)i,j}) + (\pi^{(t)}_{i,j,k} - \pi^{(t)}_{(2)i,k|j})y_{(1)i,j} & : i\in\mathcal{L}_1\cap\mathcal{U}_2 \\ \pi^{(t)}_{i,j,k}(1-y_{(2)i,k}) + (\pi^{(t)}_{i,j,k} - \pi^{(t)}_{(1)i,j|k})y_{(2)i,k} & : i\in\mathcal{L}_2\cap\mathcal{U}_1, \end{cases}$$

where

$$\pi^{(t)}_{i,j,k} = \frac{\exp(x_i'\boldsymbol{\beta}^{(t)}_{:,j,k})}{\sum_{s=1}^{J}\sum_{t=1}^{K}\exp(x_i'\boldsymbol{\beta}^{(t)}_{:,s,t})}, \quad \pi^{(t)}_{(1)i,j|k} = \frac{\exp(x_i'\boldsymbol{\beta}^{(t)}_{:,j,k})}{\sum_{s=1}^{J}\exp(x_i'\boldsymbol{\beta}^{(t)}_{:,s,k})}, \quad \pi^{(t)}_{(2)i,k|j} = \frac{\exp(x_i'\boldsymbol{\beta}^{(t)}_{:,j,k})}{\sum_{t=1}^{K}\exp(x_i'\boldsymbol{\beta}^{(t)}_{:,j,t})},$$

for $(i,j,k)\in[n]\times[J]\times[K]$. Computing the gradient of $\tilde{\mathcal{G}}_{\mathcal{U},\mathcal{L}}$ is only slightly more computationally intensive than computing the gradient of $\tilde{\mathcal{G}}$. In addition to computing joint probabilities, we see that computing the gradient involves computing both marginal and conditional probabilities. For example, $\pi^{(t)}_{(1)i,j|k}$ denotes the estimated conditional probability $P(Y_1 = j \mid x, Y_2 = k)$ at $\boldsymbol{\beta}^{(t)}$. To apply Algorithm 1 of Li and Lin (2015), we need only use that their updating equations (11) and (12) are instances of our (11), for which we can apply Theorem 2.

I Additional figures and tables referenced in Section 7

In this section, we provide a figure and table referenced in the main document, but omitted for the sake of space. In Table 2, we provide counts for both cancer types and 5-year survival status of the 420 subjects included in our data analysis in Section 8. In Figure 15, we present Kaplan-Meier survival curves for the three cancer types, and for all three combined (in purple).

5-year status	KICH	KIRC	KIRP	Total
Alive	37	152	40	229
Deceased	8	148	35	191
Total	45	300	75	420

Table 2: Counts for the two multinomial response variables in the pan-kidney cancer data we analyze in Section 7.

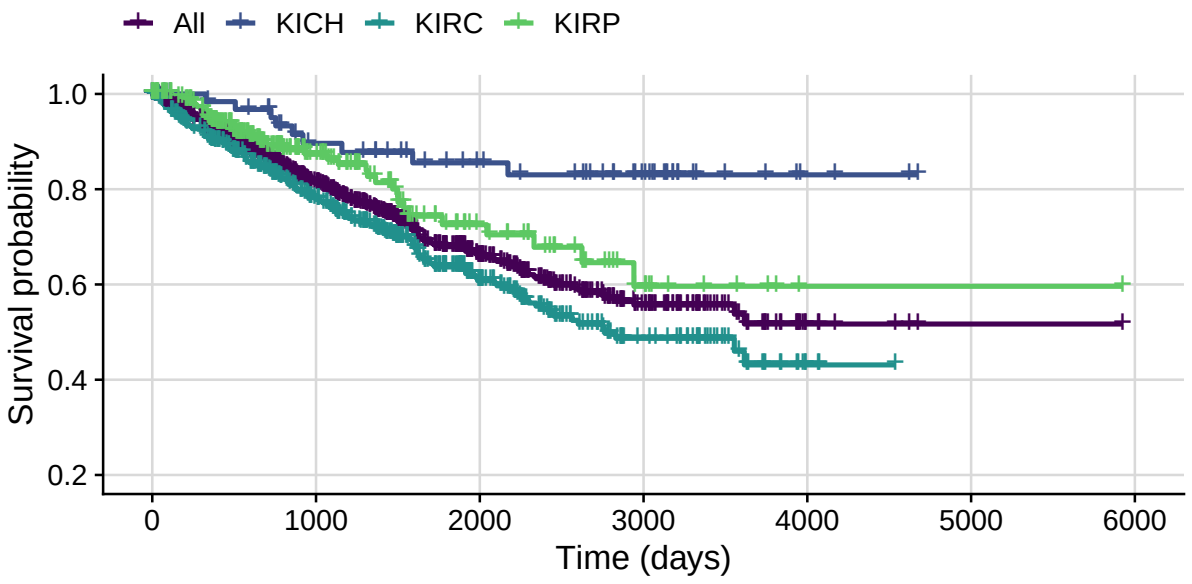

Figure 15: Kaplan-Meier survival curves for the TCGA pan-kidney cancer cohort with all three types combined (purple) and the three distinct cancer subtypes.

References

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons, Inc., 2nd edition.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414.

Li, H. and Lin, Z. (2015). Accelerated proximal gradient methods for nonconvex programming. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Molstad, A. J. and Rothman, A. J. (2018). Shrinking characteristics of precision matrix estimators. *Biometrika*, 105(3):563–574.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Price, B. S., Geyer, C. J., and Rothman, A. J. (2019). Automatic response category combination in multinomial logistic regression. *Journal of Computational and Graphical Statistics*, 28(3):758–766.

Tibshirani, R. J., Taylor, J., et al. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371.

Tran-Dinh, Q., Li, Y.-H., and Cevher, V. (2015). Composite convex minimization involving self-concordant-like cost functions. In *Modelling, Computation and Optimization in Information Systems and Management Sciences*, pages 155–168. Springer.

Yan, X. and Bien, J. (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, 32(4):531–560.

Yuan, L., Liu, J., and Ye, J. (2013). Efficient methods for overlapping group lasso. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2104–2116.