Supplements for "Exponential tensor regression with covariates on multiple modes"

1 Proofs

Proof of Theorem **??**. Define $\ell(\mathcal{B})=\mathbb{E}(\mathcal{L}_{\mathcal{Y}}(\mathcal{B}))$, where the expectation is taken with respect to $\mathcal{Y}\sim\mathcal{B}_{\text{true}}$ under the model with true parameter $\mathcal{B}_{\text{true}}$. We first prove the following two conclusions:

C1. There exists two positive constants $C_1,\ C_2>0$, such that, with probability at least $1-\exp(-C_1\log K\sum_k p_k)$, the stochastic deviation, $\mathcal{L}_{\mathcal{Y}}(\mathcal{B})-\ell(\mathcal{B})$, satisfies

$$|\mathcal{L}_{\mathcal{Y}}(\mathcal{B})-\ell(\mathcal{B})|=|\langle\mathcal{E},\ \mathcal{B}\times_1\boldsymbol{X}_1\times_2\cdots\times_K\boldsymbol{X}_K\rangle|\leq C_2\|\mathcal{B}\|_F\log K\sqrt{\frac{\prod_k r_k}{\max_k r_k}\sum_k p_k}.$$

C2. The inequality $\ell(\hat{\mathcal{B}})-\ell(\mathcal{B}_{\text{true}})\leq-\frac{L}{2}\|\hat{\Theta}-\Theta^{\text{true}}\|_F^2$ holds, where $L>0$ is the lower bound for $\min_{|\theta|\leq\alpha}|b''(\theta)|$.

To prove C1, we note that the stochastic deviation can be written as:

$$\begin{aligned}\mathcal{L}_{\mathcal{Y}}(\mathcal{B})-\ell(\mathcal{B})&=\langle\mathcal{Y}-\mathbb{E}(\mathcal{Y}|\mathcal{X}),\ \Theta(\mathcal{B})\rangle\\&=\langle\mathcal{Y}-b'(\Theta^{\text{true}}),\ \Theta\rangle\\&=\langle\mathcal{E}\times_1\boldsymbol{X}_1^T\times_2\cdots\times_K\boldsymbol{X}_K^T,\ \mathcal{B}\rangle,\end{aligned}\tag{1}$$

where $\mathcal{E}=[\![\varepsilon_{i_1,\ldots,i_K}]\!]\stackrel{\text{def}}{=}\mathcal{Y}-b'(\Theta^{\text{true}})$. Based on Proposition 1, $\varepsilon_{i_1,\ldots,i_K}$ is sub-Gaussian-(ϕU). Let $\check{\mathcal{E}}\stackrel{\text{def}}{=}\mathcal{E}\times_1\boldsymbol{X}_1^T\times_2\cdots\times_K\boldsymbol{X}_K^T$. By the property of sub-Gaussian r.v's, $\check{\mathcal{E}}$ is a $(p_1,\ldots,p_K)$-dimensional sub-Gaussian tensor with parameter bounded by $C_2=\phi Uc_2^K$. Here $c_2>0$ is the upper bound of $\sigma_{\max}(\boldsymbol{X}_k)$. Applying Cauchy-Schwarz inequality to (1) yields

$$|\mathcal{L}_{\mathcal{Y}}(\mathcal{B})-\ell(\mathcal{B})|\leq\left\|\check{\mathcal{E}}\right\|_2\|\mathcal{B}\|_*,\tag{2}$$

where $\|\cdot\|_2$ denotes the tensor spectral norm and $\|\cdot\|_*$ denotes the tensor nuclear norm. The nuclear norm $\|\mathcal{B}\|_*$ is bounded by $\|\mathcal{B}\|_*\leq\sqrt{\frac{\prod_k r_k}{\max_k r_k}}\|\mathcal{B}\|_F$ (c.f. [**?**, 14]). The spectral norm $\left\|\check{\mathcal{E}}\right\|_2$ is bounded by $\left\|\check{\mathcal{E}}\right\|_2\leq C_1Uc^K\log K\sqrt{\sum_k p_k}$ with probability at least $1-\exp(-C_2\log K\sum_k p_k)$ (c.f. [**?**, **?**]). Combining these two bounds with (2), we have, with probability at least $1-\exp(-C_2\log K\sum_k p_k)$,

$$|\mathcal{L}_{\mathcal{Y}}(\mathcal{B})-\ell(\mathcal{B})|\leq C_1Uc_2^K\|\mathcal{B}\|_F\log K\sqrt{\frac{\prod_k r_k}{\max_k r_k}\sum_k p_k}.$$

Next we prove C2. Applying Taylor expansion to $\ell(\mathcal{B})$ around $\mathcal{B}_{\text{true}}$,

$$\ell(\mathcal{B})=\ell(\mathcal{B}_{\text{true}})-\frac{1}{2}\text{vec}(\mathcal{B}-\mathcal{B}_{\text{true}})^T\mathcal{H}_{\mathcal{Y}}(\check{\mathcal{B}})\text{vec}(\mathcal{B}-\mathcal{B}_{\text{true}}),\tag{3}$$

where $\mathcal{H}_{\mathcal{Y}}(\check{\mathcal{B}})$ is the (non-random) Hession of $\frac{\partial\ell^2(\mathcal{B})}{\partial^2\mathcal{B}}$ evaluated at $\check{\mathcal{B}}=\alpha\text{vec}(\alpha\mathcal{B}+(1-\alpha)\mathcal{B}_{\text{true}})$ for some $\alpha\in[0,1]$. Recall that $b''(\theta)=\text{Var}(y|\theta)$, because $y\in\mathbb{R}$ follows the exponential family distribution with function $b(\cdot)$. By chain rule and the fact that $\Theta = \Theta(\mathcal{B}) = \mathcal{B} \times_1 \boldsymbol{X}_1 \cdots \times_K \boldsymbol{X}_K$, the equation (4) implies that

$$\ell(\mathcal{B}) - \ell(\mathcal{B}_{\text{true}}) = -\frac{1}{2}\sum_{i_1,\ldots,i_K} b''(\check{\theta}_{i_1,\ldots,i_K})(\theta_{i_1,\ldots,i_K} - \theta_{\text{true},i_1,\ldots,i_K})^2 \le -\frac{L}{2}\|\Theta - \Theta^{\text{true}}\|_F^2, \tag{4}$$

holds for all $\mathcal{B} \in \mathcal{P}$, provided that $\min_{|\theta|\le\alpha} |b''(\theta)| \ge L > 0$. In particular, the inequality (4) also applies to the constrained MLE $\hat{\mathcal{B}}$. So we have

$$\ell(\hat{\mathcal{B}}) - \ell(\mathcal{B}_{\text{true}}) \le -\frac{L}{2}\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2. \tag{5}$$

Now we have proved both C1 and C2. Note that $\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) \ge 0$ by the definition of $\hat{\mathcal{B}}$, This implies that

$$\begin{aligned}
0 &\le \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) \\
&\le \left(\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \ell(\hat{\mathcal{B}})\right) - (\mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) - \ell(\mathcal{B}_{\text{true}})) + \left(\ell(\hat{\mathcal{B}}) - \ell(\mathcal{B}_{\text{true}})\right) \\
&\le \langle \mathcal{E},\ \Theta - \Theta^{\text{true}}\rangle - \frac{L}{2}\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2,
\end{aligned}$$

where the second line follows from (5). Therefore,

$$\begin{aligned}
\|\hat{\Theta} - \Theta^{\text{true}}\|_F &\le \frac{2}{L}\Big\langle \mathcal{E},\ \frac{\hat{\Theta} - \Theta^{\text{true}}}{\|\hat{\Theta} - \Theta^{\text{true}}\|_F}\Big\rangle \\
&\le \frac{2}{L} \sup_{\Theta: \|\Theta\|_F = 1, \Theta = \mathcal{B}\times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K} \langle \mathcal{E},\ \Theta\rangle \\
&\le \frac{2}{L} \sup_{\mathcal{B}\in\mathcal{P}: \|\mathcal{B}\|_F \le \prod_k \sigma_{\min}^{-1}(\boldsymbol{X}_k)} \langle \mathcal{E},\ \ \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K\rangle.
\end{aligned} \tag{6}$$

Combining (6) with C1 yields the desired conclusion.

□

Proposition 1 (sub-Gaussian residual)**.** *Define the residual tensor $\mathcal{E} = [\![\varepsilon_{i_1,\ldots,i_K}]\!] = \mathcal{Y} - b'(\Theta) \in \mathbb{R}^{d_1\times\cdots\times d_K}$. Under the Assumption A2, $\varepsilon_{i_1,\ldots,i_K}$ is a sub-Gaussian random variable with sub-Gaussian parameter bounded by ϕU, for all $(i_1,\ldots,i_K) \in [d_1]\times\cdots\times[d_K]$.*

Proof. The proof is similar to Lemma 3 in [**?**]. For ease of presentation, we drop the subscript $(i_1,\ldots,i_K)$ and simply write $\varepsilon\ (= y - b'(\theta))$. For any given $t \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}(\exp(t\varepsilon|\theta) &= \int c(x)\exp\left(\frac{\theta x - b(\theta)}{\phi}\right)\exp\left(t(x - b'(\theta))\right)dx \\
&= \int c(x)\exp\left(\frac{(\theta+\phi t)x - b(\theta+\phi t) + b(\theta+\phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right)dx \\
&= \exp\left(\frac{b(\theta+\phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right) \\
&\le \exp\left(\frac{\phi U t^2}{2}\right),
\end{aligned}$$

where $c(\cdot)$ and $b(\cdot)$ are known functions in the exponential family corresponding to y. Therefore, ε is sub-Gaussian-(ϕU). □

Proof of Theorem **??**. The proof is similar to [3]. We sketch the main steps here for completeness. Recall that $\ell(\mathcal{B}) = \mathbb{E}(\mathcal{L}_{\mathcal{Y}}(\mathcal{B}))$. By the definition of KL divergence, we have that,

$$\begin{aligned}
\ell(\hat{\mathcal{B}}) &= \ell(\mathcal{B}_{\text{true}}) - \sum_{(i_1,\dots,i_K)} KL(\theta_{\text{true},i_1,\dots,i_K}, \hat{\theta}_{i_1,\dots,i_K}) \\
&= \ell(\mathcal{B}_{\text{true}}) - \mathrm{KL}(\mathbb{P}_{\mathcal{Y}_{\text{true}}},\ \mathbb{P}_{\hat{\mathcal{Y}}}),
\end{aligned}$$

where $\mathbb{P}_{\mathcal{Y}_{\text{true}}}$ denotes the distribution of $\mathcal{Y}|\mathcal{X}$ with true parameter $\mathcal{B}_{\text{true}}$, and $\mathbb{P}_{\hat{\mathcal{Y}}}$ denotes the distribution with estimated parameter $\hat{\mathcal{B}}$. Therefore

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\mathcal{Y}_{\text{true}}},\ \mathbb{P}_{\hat{\mathcal{Y}}}) &= \ell(\mathcal{B}_{\text{true}}) - \ell(\hat{\mathcal{B}}) \\
&= \frac{1}{2}\sum_{i_1\dots,i_K} b''(\check{\theta}_{i_1,\dots,i_K})(\theta_{i_1,\dots,i_K} - \theta_{\text{true},i_1,\dots,i_K})^2 \\
&\le \frac{U}{2}\|\Theta - \Theta^{\text{true}}\|_F^2 \\
&\le \frac{U}{2}c_2^{2K}\|\mathcal{B} - \mathcal{B}_{\text{true}}\|_F^2,
\end{aligned}$$

where the second line comes from (4), and $c_2 > 0$ is the upper bound for the $\sigma_{\max}(\boldsymbol{X}_k)$. The result then follows from Theorem **??**. □

Proof of Global Convergence. How to solve the problem that stationary points are not isolated?? □

Proof of Linear Local Convergence. Define the differential mapping $S : S(\mathcal{A}^{(t)}) = \mathcal{A}^{(t+1)}$. Let $\boldsymbol{H}$ be the Hessian matrix of $\mathcal{L}(\mathcal{A})$ at the local maximum $\mathcal{A}^*$. We partition the $\boldsymbol{H}$:

$$d^2\mathcal{L}(\mathcal{A}^*) = d^2\mathcal{L}(\mathcal{C}^*, M_1^*, \cdots, M_K^*) = \begin{pmatrix} d^2_{CC}\mathcal{L} & d^2_{CM_1}\mathcal{L} & \cdots & d^2_{CM_K}\mathcal{L} \\ d^2_{M_1C}\mathcal{L} & d^2_{M_1M_1}\mathcal{L} & \cdots & d^2_{M_1M_K}\mathcal{L} \\ \vdots & \vdots & \ddots & \vdots \\ d^2_{M_KC}\mathcal{L} & d^2_{M_KM_1}\mathcal{L} & \cdots & d^2_{M_KM_K}\mathcal{L} \end{pmatrix} = L + D + L^\top,$$

where L is strictly block lower triangle matrix and D is the diagonal part. By condition *A2*, every diagonal block of H is negative definite and thus $L + D$ is invertible. According to Bezdek(2003), we have $dS(\mathcal{A}^*) = -(L+D)^{-1}L$. Therefore, the spectral radius of $dS(\mathcal{A}^*)$ is $\rho = max_i|\lambda_i(-(L+D)^{-1}L)| \in (0,1)$. According to Jensen's inequality:

$$\begin{aligned}
\left\|S(\mathcal{A}^{(t)}) - S(\mathcal{A}^*)\right\|_F &= \left\|\int_0^1 dS(\mathcal{A}^* - u(\mathcal{A}^* - \mathcal{A}^{(t)}))(\mathcal{A}^* - \mathcal{A}^{(t)})du\right\|_F \\
&\le \int_0^1 \left\|dS(\mathcal{A}^* - u(\mathcal{A}^* - \mathcal{A}^{(t)}))(\mathcal{A}^* - \mathcal{A}^{(t)})\right\|_F du \\
&\le \rho\left\|\mathcal{A}^{(t)} - \mathcal{A}^*\right\|_F.
\end{aligned}$$

That implies S is a contraction mapping on the space $(\mathcal{A}, \|\cdot,\cdot\|_F)$. By contraction principle, the iterates $S(\mathcal{A}^{(t)}) = \mathcal{A}^{(t+1)}$ linearly converges to the point $\mathcal{A}^*$ that $S(\mathcal{A}^*) = \mathcal{A}^*$,

$$\left\|\mathcal{A}^{(t)} - \mathcal{A}^*\right\|_F \le \rho^t\left\|\mathcal{A}^{(0)} - \mathcal{A}^*\right\|_F,$$

for $\mathcal{A}^{(0)}$ sufficiently close to $\mathcal{A}*$.

□

2 Time complexity

The computational complexity of our tensor regression model is $O(d^3+d)$ for each loop of iterations, where $d = \prod_k d_k$ is the total size of the response tensor. More precisely, the update of core tensor costs $O(r^3 d^3)$, where $r = \sum_k r_k$ is the total size of the core tensor. The update of factor matrix $\boldsymbol{M}_k$ involves solving p_k separate GLMs. Solving those GLMs requires $O(r_k^3 p_k + p_k r_k^3 d d_k^{-1})$, and therefore the cost for updating K factors in total is $O(\sum_k r_k^3 p_k d_k + d \sum_k r_k^3 p_k d_k^{-1}) \approx O(\sum p_k d_k + d) \approx O(d)$.

3 Additional results for real data analysis

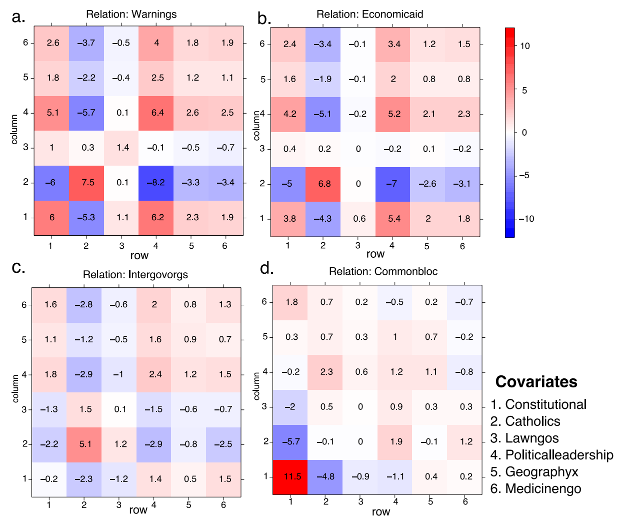
Here we provide additional results for the real data analysis.

3.1 HCP data analysis

Supplement Figure S1 compares the estimated coefficients from our method (tensor regression) with those from classical GLM approach. A classical GLM is to regress the brain edges, one at a time, on the individual-level covariates, and this logistic model is repeatedly fitted for every edge $\in [68] \times [68]$. As we can see in the figure, our tensor regression shrinkages the coefficients towards center, thereby enforcing the sharing between coefficient entries.

3.2 Nations data analysis

To investigate the effects of dyadic attributes towards connections, we depicted the estimated coefficients $\hat{\mathcal{B}} = [\![\hat{b}_{ijk}]\!]$ for several relation types (Supplement Figure S2). Note that entries $\hat{b}_{ijk}$ can be interpreted as the contribution, at the logit scale, of covariate pair (i, j) (ith covariate for the "sender" country and jth covariate for the "receiver" country) towards the connection of relation k. Several interesting findings emerge from the observation. We found that relations belonging to a same cluster tend to have similar covariate effects. For example, the relations *warnings* and *ecnomicaid* are classified into Cluster II, and both exhibit similar covariate pattern (Supplement Figure S2a-b). Moreover, the majority of the diagonal entries $\hat{\mathcal{B}}(i, i, k)$ positively contribute to the connection. This suggests that countries with coherent attributes tend to interact more often than others. We also found that the *constitutional* attribute is an important predictor for the *commonbloc* relation, whereas the effect is weaker for other relations (Supplement Figure S2d). This is not surprising, as the block partition during Cold War is associated with the *constitutional* attribute.



Supplementary Figure S2: Effect estimation in the *Nations* data. Panels (a)-(d) represent the estimated effects of country-level attributes towards the connection probability, for relations *warnning*, *economicaid*,*intergovorg*, and *commonblock*, respectively.

Supplement table S1 summarizes the K-means clustering of the 56 relations based on the 3$^{\text{rd}}$ mode factor $\boldsymbol{M}_3 \in \mathbb{R}^{56\times 4}$ in the tensor regression model.

Cluster I	officialvisits, intergovorgs, militaryactions, violentactions, duration, negativebehavior, boycottembargo, aidenemy, negativecomm, accusation, protestsunoffialacts, nonviolentbehavior, emigrants, relexports, timesincewar, commonbloc2, rintergovorgs3, relintergovorgs
Cluster II	economicaid, booktranslations, tourism, relbooktranslations, releconomicaid, conferences, severdiplomatic, expeldiplomats, attackembassy, unweightedunvote, reltourism, tourism3, relemigrants, emigrants3, students, relstudents, exports, exports3, lostterritory, dependent, militaryalliance, warning
Cluster III	treaties, reltreaties, exportbooks, relexportbooks, weightedunvote, ngo, relngo, ngoorgs3, embassy, reldiplomacy, timesinceally, independence, commonbloc1
Cluster IV	commonbloc0, blockpositionindex

Supplementary Table S1: K-means clustering of relations based on factor matrix in the coefficient tensor.

References

[1] Will Wei Sun and Lexin Li. STORE: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.

[2] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

[3] Nicolai Baldin and Quentin Berthet. Optimal link prediction with matrix logistic regression. *arXiv preprint arXiv:1803.07054*, 2018.

[4] Peter D Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295, 2005.

[5] Guillaume Rabusseau and Hachem Kadri. Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems*, pages 1867–1875, 2016.

[6] Rose Yu and Yan Liu. Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning*, pages 373–381, 2016.

[7] Qibin Zhao, Cesar F Caiafa, Danilo P Mandic, Zenas C Chao, Yasuo Nagasaka, Naotaka Fujii, Liqing Zhang, and Andrzej Cichocki. Higher order partial least squares (HOPLS): a generalized multilinear regression method. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1660–1673, 2012.

[8] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[9] David Hong, Tamara G Kolda, and Jed A Duersch. Generalized canonical polyadic tensor decomposition. *SIAM Review, in press. arXiv:1808.07452*, 2019.

[10] Anru Zhang and Dong Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 2018.

[11] Han Chen, Garvesh Raskutti, and Ming Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–208, 2019.

[12] Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146, 2017.

[13] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[14] Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear Algebra and Its Applications*, 520:44–66, 2017.

[15] K Ruben Gabriel. Generalised bilinear regression. *Biometrika*, 85(3):689–700, 1998.

[16] Richard F Potthoff and SN Roy. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3-4):313–326, 1964.

[17] Muni S Srivastava, Tatjana von Rosen, and Dietrich Von Rosen. Models with a kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics*, 17(4):357–370, 2008.

[18] Jingfei Zhang, Will Wei Sun, and Lexin Li. Network response regression for modeling population of networks with covariates. *arXiv preprint arXiv:1810.03192*, 2018.

[19] Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

[20] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[21] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

[22] Lucy F Robinson, Lauren Y Atlas, and Tor D Wager. Dynamic functional connectivity using state-based dynamic community structure: Method and application to opioid analgesia. *NeuroImage*, 108:274–291, 2015.

[23] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

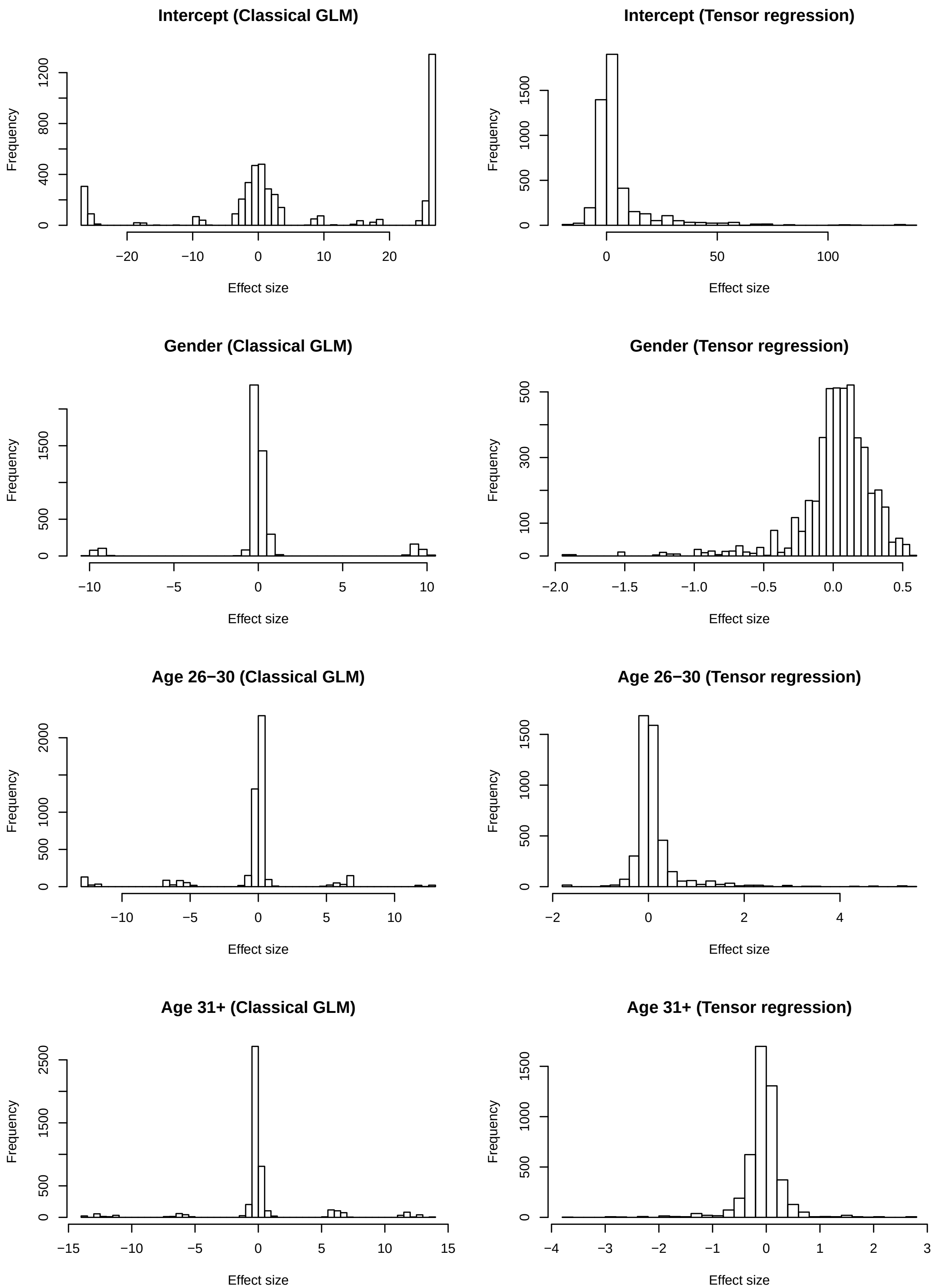
[24] Miaoyan Wang and Yuchen Zeng. Multiway clustering via tensor block models. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019). arXiv:1906.03807*, 2019.

[25] Linda Geddes. Human brain mapped in unprecedented detail. *Nature*, 2016.

[26] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

[27] Madhura Ingalhalikar, Alex Smith, Drew Parker, Theodore D Satterthwaite, Mark A Elliott, Kosha Ruparel, Hakon Hakonarson, Raquel E Gur, Ruben C Gur, and Ragini Verma. Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences*, 111(2):823–828, 2014.

[28] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning*, volume 11, pages 809–816, 2011.



Supplementary Figure S1: Comparison of coefficient estimation in the HCP data.