Method Comparison: Matrix response regression

Jiaxin Hu

May 9, 2021

1 Model comparison (STD)

1.1 Our model

Let $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote the order-K observation tensor, and $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}, k \in [K]$ denote the feature matrices. Assume the entries in $\mathcal{Y}$ are from exponential family. Our model is stated as

$$\mathbb{E}[\mathcal{Y}|\boldsymbol{X}_1, ..., \boldsymbol{X}_K] = f(\mathcal{B} \times \{\boldsymbol{X}_1, ..., \boldsymbol{X}_K\}),$$

where the coefficient tensor $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ satisfies the low-rank structure

$$\mathcal{B} = \mathcal{C} \times \{\boldsymbol{M}_1, ..., \boldsymbol{M}_K\},$$

with core tensor $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$, $\boldsymbol{M}_k \in \mathbb{R}^{p_k \times r_k}$, and $\boldsymbol{M}_k^T \boldsymbol{M}_k = \boldsymbol{I}_{r_k}$.

The model under the special case with $K = 3, d_1 = d_2 = n, d_3 = N, \boldsymbol{X}_1 = \boldsymbol{X}_2 = \boldsymbol{I}_n$ is

$$\mathbb{E}[\mathcal{Y}|\boldsymbol{X}_3] = f(\mathcal{B} \times_3 \boldsymbol{X}_3), \tag{1}$$

where $\mathcal{B}$ has low-rank structure.

1.2 Matrix response regression (Mreg)

Let $\boldsymbol{Y}_i \in \mathbb{R}^{n \times n}, i \in [N]$ denote the observed matrices, and $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ denote the feature matrix for the N subjects. Let $\boldsymbol{X}_i$ denote the i-th column of $\boldsymbol{X}$. Assume the entries in $\boldsymbol{Y}_i$ are from exponential family. The matrix-wise regression model is

$$\mathbb{E}[\boldsymbol{Y}_i|\boldsymbol{X}_i] = f(\Theta + \mathcal{B} \times_3 \boldsymbol{X}_i),$$

where $\mathcal{B} \in \mathbb{R}^{n \times n \times p}$ is the sparse coefficient tensor with sparsity $\|\mathcal{B}\|_0 \leq s$ and Θ is the intercept matrix with low-rank structure satisfying the SVD $\Theta = U\Sigma V^T, \Sigma \in \mathbb{R}^{r \times r}$.

Stuck the matrix observations together as a tensor $\mathcal{Y} \in \mathbb{R}^{n \times n \times N}$ with $\mathcal{Y}_{..i} = \boldsymbol{Y}_i$. We rewrite the matrix regression model into the tensor form as following.

$$\mathbb{E}[\mathcal{Y}|\boldsymbol{X}] = f(\tilde{\Theta} + \mathcal{B} \times_3 \boldsymbol{X}),$$

where $\tilde{\Theta} \in \mathbb{R}^{n \times n \times N}$ is "intercept" tensor with $\tilde{\Theta}_{..i} = \Theta$. Further, we can move the intercept into the coefficient tensor $\mathcal{B}$ by replacing the covariate matrix $\boldsymbol{X}$ to $\tilde{\boldsymbol{X}} = [1_N, \boldsymbol{X}] \in \mathbb{R}^{N \times (p+1)}$. Then, we have

$$\mathbb{E}[\mathcal{Y}|\boldsymbol{X}] = f(\tilde{\mathcal{B}} \times_3 \tilde{\boldsymbol{X}}), \tag{2}$$

where $\tilde{\mathcal{B}} \in \mathbb{R}^{n \times n \times (p+1)}$ with $\tilde{\mathcal{B}}_{..1} = \Theta$.

2 Simulation Set up

2.1 Basic points

1. Our model (1) and the matrix regression model (2) are quite similar. The main differences between these two include (1) the coefficient tensor $\tilde{\mathcal{B}}$ may not have tensor low-rank structure (2) the existence of intercept matrix Θ. Note that the Mreg is also capable to assume $\mathcal{B}$ has low-rank structure. But the paper mainly focuses on the sparsity of $\mathcal{B}$.

2. The algorithm for Mreg outputs the estimation of Θ and $\mathcal{B}$. Therefore, we may compare the estimation error and the prediction error.

3. The Mreg model suffers the lack of invariance to covariate encoding due to the sparsity constraint.

   Consider the toy example with following set-to-0 covariate matrix and true parameters.

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \Theta = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \quad \mathcal{B}_{..1} = \begin{bmatrix} 0 & 0 \\ 0 & -2 \end{bmatrix}, \quad \text{and} \quad \mathcal{B}_{..2} = \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix},$$

   where we have the true sparsity $\|\mathcal{B}\|_0 = 2$.

   Consider the sum-to-0 encoding of the covariate $\tilde{X}$. To make two models equivalent, the true parameters for sum-to-0 constraint should be

$$\tilde{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}, \quad \tilde{\Theta} = \Theta - \tilde{\mathcal{B}}_{..1} - \tilde{\mathcal{B}}_{..2},$$

$$\tilde{\mathcal{B}}_{..1} = \frac{2\mathcal{B}_{..1} - \mathcal{B}_{..2}}{3} = \frac{1}{3}\begin{bmatrix} 0 & 0 \\ 1 & -4 \end{bmatrix}, \quad \tilde{\mathcal{B}}_{..2} = \frac{-\mathcal{B}_{..1} + 2\mathcal{B}_{..2}}{3} = \frac{1}{3}\begin{bmatrix} 0 & 0 \\ -2 & 2 \end{bmatrix}.$$

   Note that the sparsity of the coefficient tensor is $\left\|\tilde{\mathcal{B}}\right\|_0 = 4$. If we apply previous sparsity $\|\mathcal{B}\|_0 = 2$ to the algorithm, we always have biased estimations with second encoding. Therefore, Mreg suffers the invariance to covariate encoding.

2.2 Simulation setup

Basic setting: In this simulation, we mimic the low-rank brain network setting. Let $d_1 = d_2 = 20$ and $d_3 = 50$. We consider the Gaussian model first, i.e., the link function $f(x) = x$.

Experiment 1: Invariance of covariate encoding
In this experiment, we want to show that Mreg is not invariant to the covariate encoding while our STD is invariant. The design idea follows the toy example in last subsection. Target figure: Network heat-maps for the fitted value $\hat{Y}_i \in \mathbb{R}^{20\times 20}$. I believe the heat-maps can show the difference of fitted value directly.

agree.

To check whether the method is invariant to the feature encodings or not, we compare the fitted values with the same given observation tensor and different encoded feature matrix $\boldsymbol{X}$. For better comparison, we consider the Gaussian observation for Mreg and our STD model, under which the numerical solutions are more stable.

For fair comparison:
1. simper X: 2-category feature (gender).
Encoding A: 1 for female 0 for male; Encoding B: 1 for male 0 for female
2. generate data as you did in Mreg
3. fit both Mreg and our method using the same set {X (encoding A), Y}
4. fit using {X (encoding B), Y}
5. visualize as you did now.

For Mreg model, we consider the setting $r = 2, n = 10, N = 15, p = 2$, and $s = 0.1$. Let $\boldsymbol{X} \in \mathbb{R}^{15\times 2}$ denote the discrete feature matrix. For example, $\boldsymbol{X}$ be the membership matrix representing the assignment of the individuals to 3 groups. Suppose that first 5 individuals belong to group 1, second 5 individuals belong to group 2, and the rest belongs to group 3. There are two possible way to encode the group assignment

$$\boldsymbol{X}_1 = \begin{bmatrix} 1_5 & 0 \\ 0 & 1_5 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{X}_2 = \begin{bmatrix} 1_5 & 0 \\ 0 & 1_5 \\ -1_5 & -1_5 \end{bmatrix}.$$

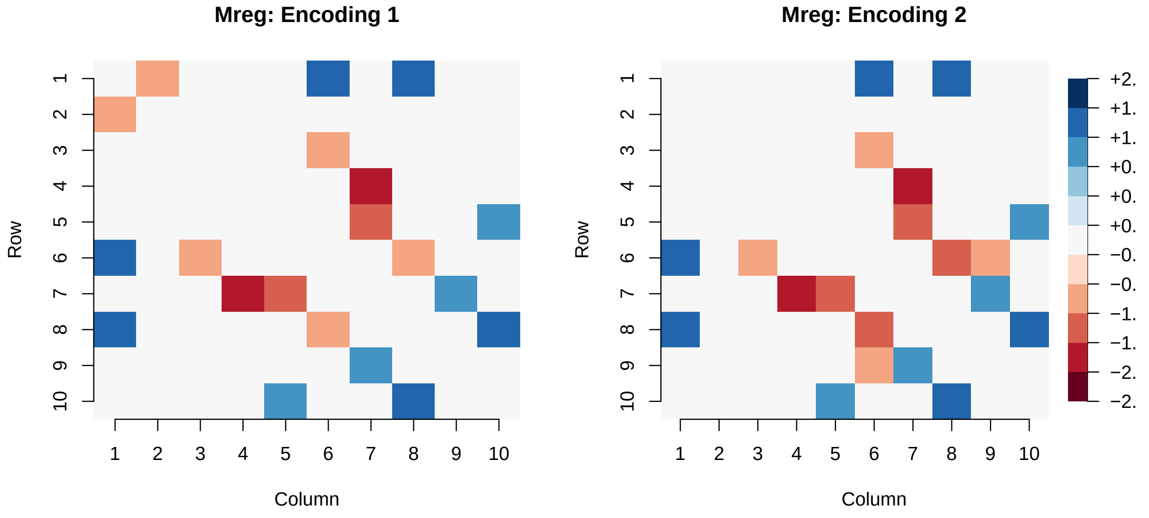
We say that the first encoding $\boldsymbol{X}_1$ has set-to-0 constraint, and the second encoding $\boldsymbol{X}_2$ has sum-to-0 constraint. We generate the data using the first encoding

$$\mathcal{Y} = \tilde{\Theta} + \mathcal{B} \times_3 \boldsymbol{X}_1 + \epsilon,$$

where $\epsilon_{i,j,k} \sim_{i.i.d.} N(0,1)$ and s proportion of $\mathcal{B}$'s entries are assigned with value 2 randomly while other entries keep 0.

If the model is invariant to feature encodings, then the fitted values should be the same and thus the estimation of the difference between group 1 and group 2 should be the same. Figure 1 visualizes the difference $\hat{\mathcal{B}}_{..1} - \hat{\mathcal{B}}_{..2}$ of Mreg model under different encodings respectively.

Is the plotted B from your own calculation or the software output?


Mreg: Encoding 1

Mreg: Encoding 2

Figure 1: Fitted value difference $\hat{\mathcal{B}}_{..2} - \hat{\mathcal{B}}_{..1}$ of Mreg model with different feature encodings. The left panel uses set-to-0 constraint $\boldsymbol{X}_1$, and the right panel uses sum-to-0 constraint to $\boldsymbol{X}_2$.

Similarly, for our STD model, we consider the setting $\boldsymbol{r} = (3, 3, 3), d_1 = d_2 = 10, d_3 = 15, p = 3$ and difference encodings

$$\boldsymbol{X}_1' = \begin{bmatrix} 1_5 & 1_5 & 0 \\ 1_5 & 0 & 1_5 \\ 1_5 & 0 & 0 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{X}_2' = \begin{bmatrix} 1_5 & 1_5 & 0 \\ 1_5 & 0 & 1_5 \\ 1_5 & -1_5 & -1_5 \end{bmatrix}.$$

Figure 2 visualizes the difference between group 1 and group 2 $\mathcal{B}_{..3} - \mathcal{B}_{..2}$ of STD model under different encodings respectively.

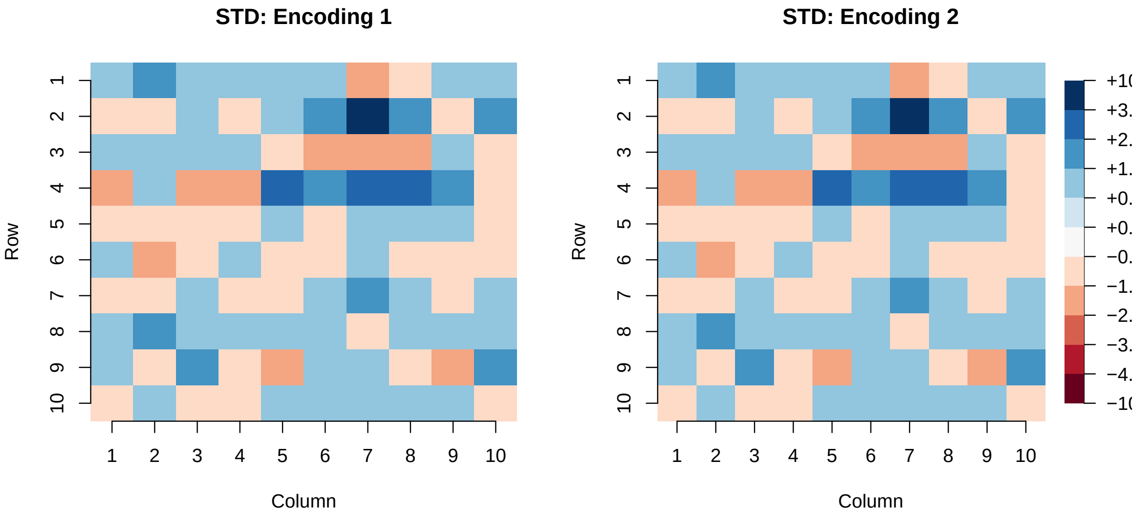


Figure 2: Fitted value difference $\hat{\mathcal{B}}_{..3} - \hat{\mathcal{B}}_{..2}$ of STD model with different feature encodings. The left panel uses set-to-0 constraint $\boldsymbol{X}_1'$, and the right panel uses sum-to-0 constraint to $\boldsymbol{X}_2'$.

For clearly comparison, let matrix in left panels subtracts the matrix in right panels. Figure 3 shows the subtractions. According to Figure 3, our STD model is invariant to the feature encodings while Mreg is not invariant. The possible reasons for lack of invariance of Mreg is the sparsity assumption to the coefficient tensor $\mathcal{B}$.

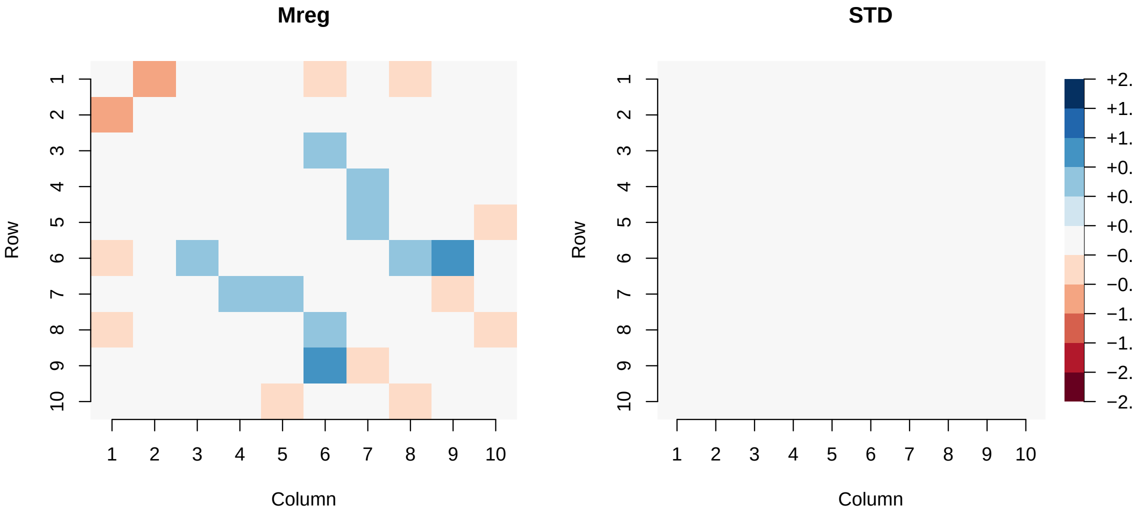
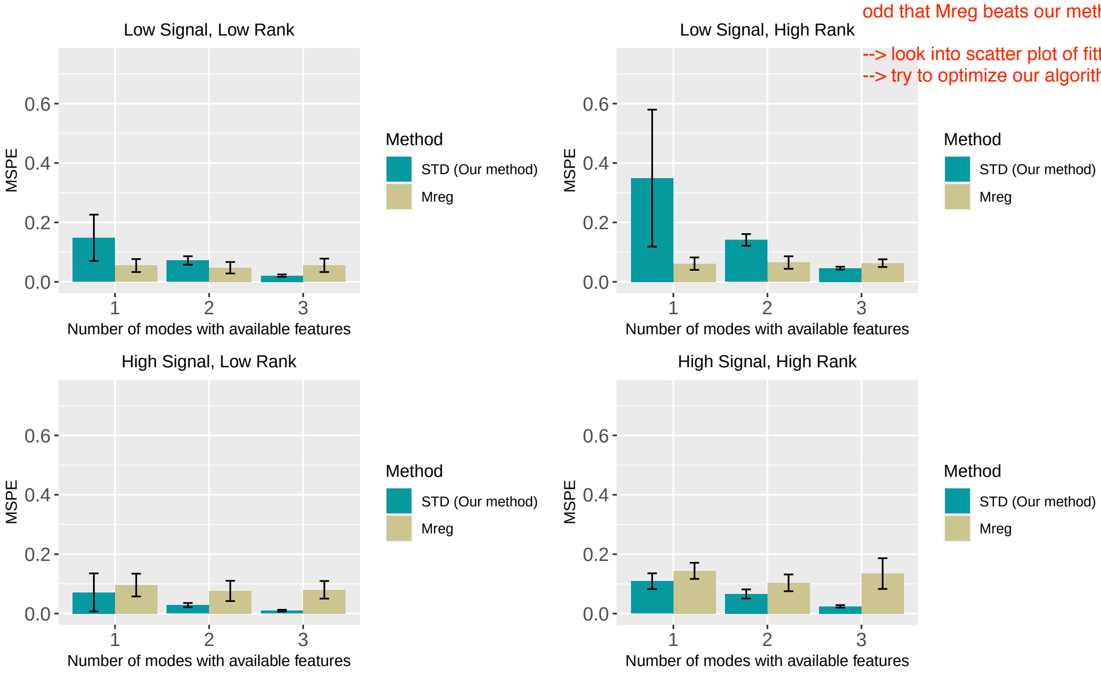


Figure 3: Difference between two encodings for Mreg model and STD model. The matrices are the subtraction of the left panels between the right panels in Figure 1 and Figure 2.

Experiment 2: Prediction error

In this experiment, we check the prediction error for these two method via MPSE. The comparison is assessed from three aspects: (a) benefit of incorporating features from multiple modes; (b) prediction error with respect to sample size.

Since the code for Mreg can only deal with binary data, we generate observation tensor with logistic link through `sim_data`.
Target figure: Similar to Figure 5. Histograms that compare the MPSE under different cases.



odd that Mreg beats our method when # feature =1.

--> look into scatter plot of fitted B
--> try to optimize our algorithm (e.g. by taking benefit from Mr

Figure 4: Comparison between our STD method and Mreg method versus the number of available informative modes. We consider rank $\boldsymbol{r} = (3,3,3)$(low), $\boldsymbol{r} = (4,5,6)$(high), and signal $\alpha = 3$(low), $\alpha = 6$(high).

For high signal case only, see figure 7.

Why is our model worse than Mreg even under correctly specified model?
--> investigate the reason & try to optimize our algorithm.

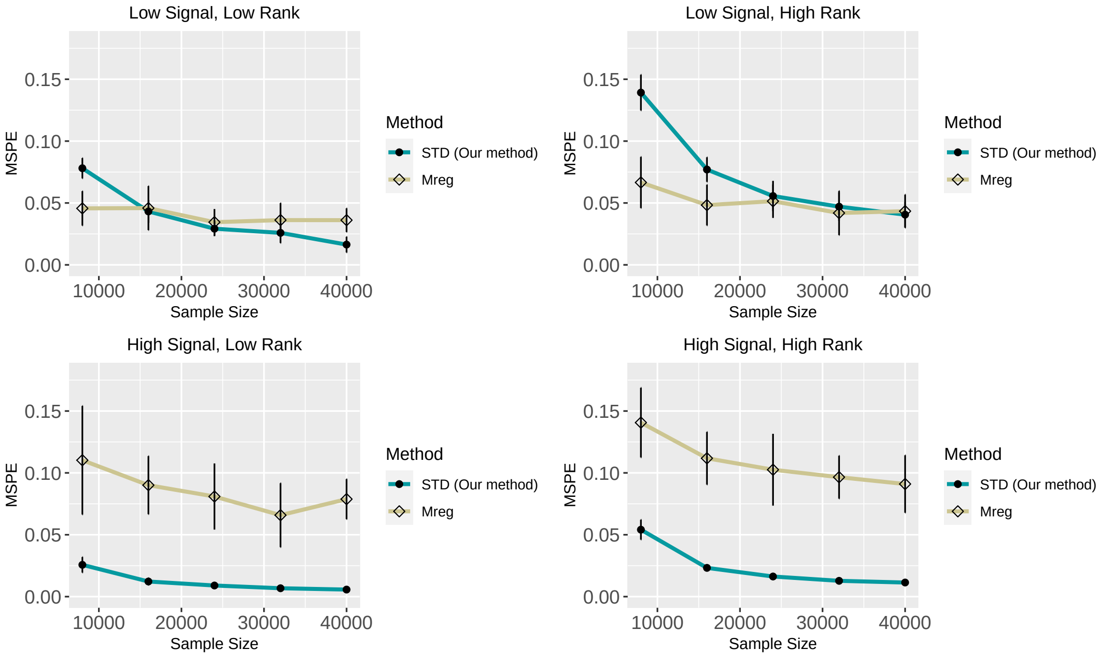
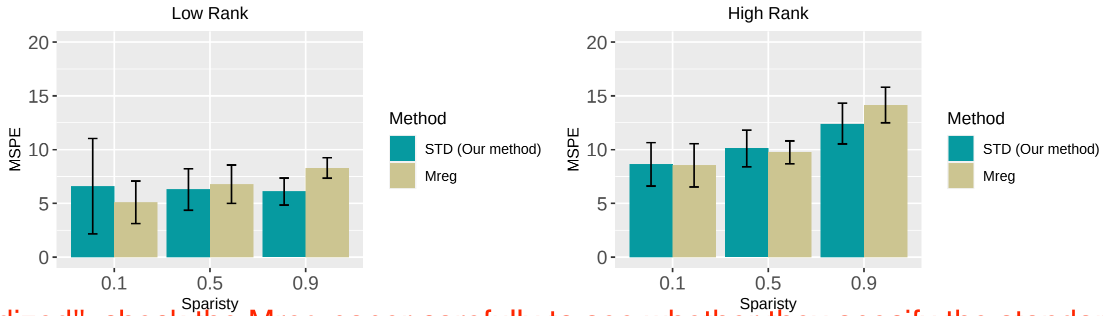


Figure 5: Comparison between our STD method and Mreg method versus the sample size. We consider rank $\boldsymbol{r} = (3,3,3)$(low), $\boldsymbol{r} = (4,5,6)$(high), and signal $\alpha = 3$(low), $\alpha = 6$(high).



``standardized'': check the Mreg paper carefully to see whether they specify the standardization. If so, our earlier claim on lack of invariance may not hold.

Figure 6: Comparison between our STD method and Mreg method versus the sparsity of $\mathcal{B}$ in Mreg model. The observation tensors are generated by model $\mathcal{Y} = Ber(\Theta + \mathcal{B} \times_3 \boldsymbol{X})$, where $\boldsymbol{X} \sim N(0,1)$ and $\boldsymbol{X}$ is standardized. We consider the rank of Θ: $r = 3$(low) and $r = 6$(high). We fit STD model with $\boldsymbol{r} = (3,3,3)$, $\boldsymbol{r} = (6,6,6)$, $\boldsymbol{r} = (9,9,9)$ respectively and choose the best results. We duplicate each setting for 5 times.

why larger than 2?

Bhat = [Theta, B]\ in R^{d * d * 2}

ground truth rank of Bhat <= {r+s, r+s, 2}, where s = number of non-zero entries in B, and r = rank of Theta.

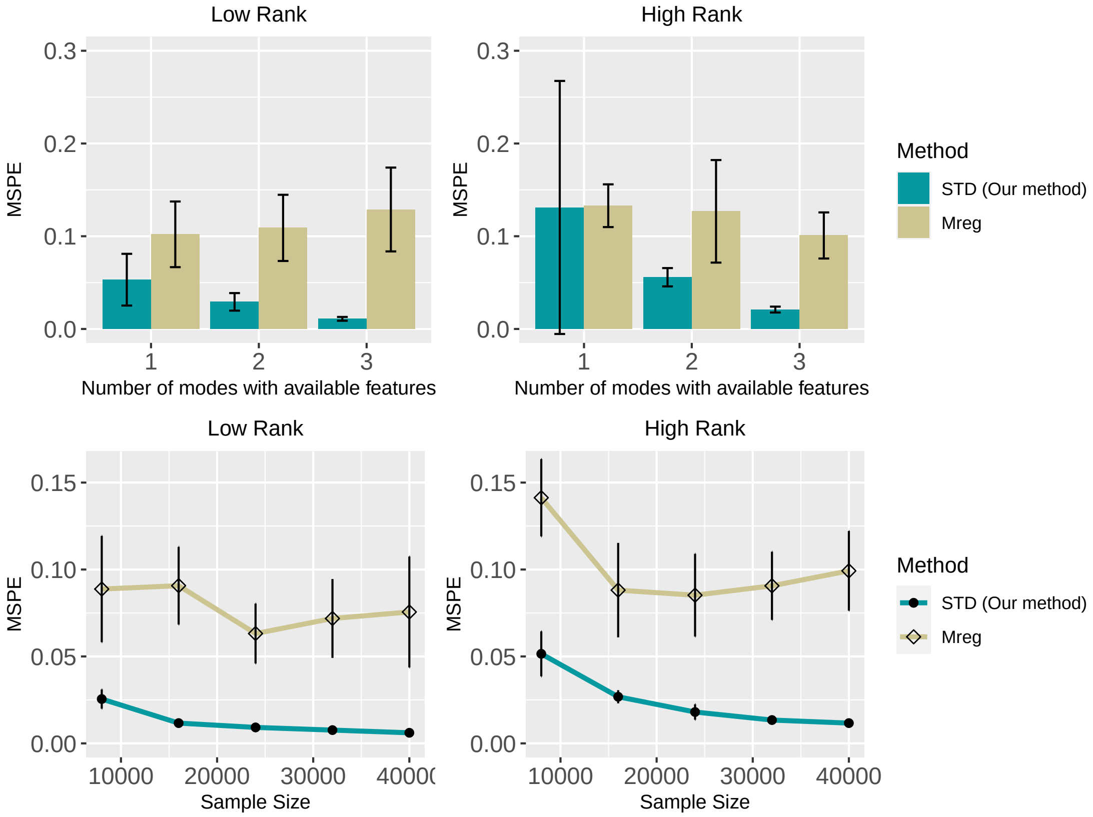


Figure 7: Comparison between our STD method and Mreg method. Higher panels plot MSPE versus the number of modes with available features. Lower panels plot MSPE versus the effective sample size d^2. We consider rank $\boldsymbol{r} = (3,3,3)$(low) and $\boldsymbol{r} = (4,5,6)$(high).