Estimation and Prediction Error in Supervised Setting 3.5

Zhuoyan Xu Jiaxin Hu

Aug 13 2019

The general supervised model is:

$$\begin{aligned} \text{logit}\left\{\mathbb{E}\left[\mathcal{Y}^{d_1 d_2 \ldots d_K}\right]\right\} &= \mathcal{G}^{r_1 r_2 \ldots r_K} \times_1 W^{d_1 r_1} \times_2 N_2^{d_2 r_2} \ldots \times_K N_K^{d_K r_K} \\ W^{d_1 r_1} &= X^{d_1 p} N_1^{p r_1} \end{aligned}$$

where $\mathcal{G}$ is the low rank core tensor of factorization. $W, N_2, \ldots, N_K$ are factor matrices. N_1 is the regression coefficient matrix for X on W.

We can write down the model in another view, which helps to compute:

$$\text{logit}\left\{\mathbb{E}\left[\mathcal{Y}^{d_1 d_2 \ldots d_K}\right]\right\} = \Theta \times_1 X^{d_1 p}$$

where Θ is coefficient tensor with tucker rank $(r_1, \ldots, r_K)$.

Definition (Restricted Isometry Property). The isometry constant of X is the smallest number δ_R such as the following holds for all Θ with Tucker rank at most R $= \max\{r_1, \ldots, r_K\}$.

$$(1-\delta_R)\|\Theta\|_F^2 \leq \|\Theta \times_1 X\|_F^2 \leq (1+\delta_R)\|\Theta\|_F^2$$

Using the notation of Sec2.2 in *Boundaries with Gaussian Width* in 8/8/2019. We have:

$$\begin{aligned} 0 &\leq \langle \mathcal{S}_Y^*(\Theta_{true} \times_1 X), (\Theta - \Theta_{true}) \times_1 X\rangle - \frac{\gamma_\alpha}{2}\|(\hat{\Theta} - \Theta_{true}) \times_1 X\|_F^2 \\ \|(\hat{\Theta} - \Theta_{true}) \times_1 X\|_F^2 &\leq \frac{2L_\alpha}{\gamma_\alpha}\langle L_\alpha^{-1}\mathcal{S}_Y^*(\Theta_{true} \times_1 X), (\Theta - \Theta_{true}) \times_1 X\rangle \end{aligned}$$

Define

$$\|X\|_{2\to\infty} = \max_{1\leq j\leq n}\sqrt{\sum_{i=1}^m |a_{ij}|^2}$$

which is the max column Euclidean norm of covariate matrix X, denoted as $2 \to \infty$ norm of X.

Define

$$\tilde{X} = \frac{X}{\|X\|_{2\to\infty}}$$

Use $\mathcal{S}$ denote $L_\alpha^{-1}\mathcal{S}_Y^*(\Theta_{true}\times_1 X)$. Then we have:

$$\begin{aligned}\langle L_\alpha^{-1}\mathcal{S}_Y^*(\Theta_{true}\times_1 X),&(\Theta-\Theta_{true})\times_1 X\rangle=\langle\mathcal{S},(\Theta-\Theta_{true})\times_1 X\rangle\\ &\leq\|X\|_{2\to\infty}\langle\mathcal{S}\times_1\frac{X^T}{\|X\|_{2\to\infty}},(\Theta-\Theta_{true})\rangle\\ &=\|X\|_{2\to\infty}\langle\mathcal{S}\times_1\tilde{X}^T,(\Theta-\Theta_{true})\rangle\\ &=\|X\|_{2\to\infty}\langle\mathcal{E},(\Theta-\Theta_{true})\rangle\end{aligned}$$

Since $\forall s\in\mathcal{S}$, where s denote any entry in $\mathcal{S}$, we have

$$\mathbb{E}(s)=0,\quad |s|\leq 1\Longrightarrow s\in\mathrm{sG}\,(1)$$

Consider:

$$\begin{aligned}\mathcal{E}&=\mathcal{S}\times_1\tilde{X}^T\\ \mathcal{E}_{i_1\cdots i_K}&=\sum_{j=1}^{d_1}\mathcal{S}_{ji_2\cdots i_K}\tilde{X}_{ji_1}\end{aligned}$$

Define

$$\mathcal{E}(u_1,u_2,\ldots,u_K)=\langle\mathcal{E},u_1\otimes u_2\otimes\ldots\otimes u_K\rangle$$

where $u_1\in B_2^p, u_k\in B_2^{d_k}$ for $k=2,\ldots,K$.

Thus:

$$\begin{aligned}\mathbb{E}\left[\exp\left\{t\mathcal{E}(u_1,u_2,\ldots,u_K)\right\}\right]&=\mathbb{E}\left[\exp\left\{t\sum_{i_1=1}^{p}\sum_{i_2=1}^{d_2}\cdots\sum_{i_K=1}^{d_K}\mathcal{E}_{i_1i_2\cdots i_K}u_{1i_1}u_{2i_2}\cdots u_{Ki_K}\right\}\right]\\ &=\mathbb{E}\left[\exp\left\{t\sum_{i_1=1}^{p}\sum_{i_2=1}^{d_2}\cdots\sum_{i_K=1}^{d_K}\sum_{j=1}^{d_1}\mathcal{S}_{ji_2\cdots i_K}\tilde{X}_{ji_1}u_{1i_1}u_{2i_2}\cdots u_{Ki_K}\right\}\right]\\ &=\prod_{i_1=1}^{p}\prod_{i_2=1}^{d_2}\cdots\prod_{i_K=1}^{d_K}\prod_{j=1}^{d_1}\mathbb{E}\left[\exp\left\{t\mathcal{S}_{ji_2\cdots i_K}\tilde{X}_{ji_1}u_{1i_1}u_{2i_2}\cdots u_{Ki_K}\right\}\right]\end{aligned}$$

Since

$$\mathbb{E}\left[e^{t\mathcal{S}_{j_1\cdots i_K}}\right]\leq e^{t^2/2}$$

Then for a given X and fixed $u_k(k=1,\ldots,K)$, we have:

$$\begin{aligned}\mathbb{E}\left[\exp\left\{t\mathcal{E}(u_1,u_2,\ldots,u_K)\right\}\right]&\leq\prod_{i_1=1}^{p}\prod_{i_2=1}^{d_2}\cdots\prod_{i_K=1}^{d_K}\prod_{j=1}^{d_1}\left[\exp\left\{\frac{1}{2}t^2\tilde{X}_{ji_1}^2u_{1i_1}^2u_{2i_2}^2\cdots u_{Ki_K}^2\right\}\right]\\ &=e^{t^2/2}\end{aligned}$$

According to theorem 1 in [1], we have with probability at least $1-\exp\left(-C_1\log K(p+\sum_{k=2}^K d_k)\right)$:

$$\|\mathcal{E}\|_\sigma\leq C_2\log K\sqrt{p+\sum_{k=2}^K d_k}\tag{1}$$

According to our bounds on Gaussian width, we have:

$$\langle \mathcal{E}, (\Theta - \Theta_{true}) \rangle \leq C_2 \sqrt{\sum_{k=2}^{K} r_k (\sum_{k=2}^{K} d_k + p)} \|\hat{\Theta} - \Theta_{\text{true}}\|_F$$

Thus, we have:

$$\begin{aligned} \left\| \left(\hat{\Theta} - \Theta_{\text{true}} \right) \times_1 X \right\|_F^2 \leq & \frac{2L_\alpha}{\gamma_\alpha} \|X\|_{2\to\infty} \langle \mathcal{E}, (\Theta - \Theta_{\text{true}}) \rangle && (2) \\ \leq & \frac{2L_\alpha C_2}{\gamma_\alpha} \|X\|_{2\to\infty} \sqrt{\sum_{k=2}^{K} r_k (\sum_{k=2}^{K} d_k + p)} \|\hat{\Theta} - \Theta_{\text{true}}\|_F && (3) \end{aligned}$$

Then we show $\|X\|_{2\to\infty}$ can be bound by $(1 + \delta_R)$ according to RIP. Without loss of generality, assume j-th column of X (denoted as X_j) has max Euclidean norm:

$$\|X_j\|_2 = \|X\|_{2\to\infty}$$

Since RIP holds for all Θ with Tucker rank at most R = $\max\{r_1, \ldots, r_K\}$. For a fixed Θ, pick any mode-1 fiber of Θ (denoted as $\theta_{\mathbf{i}}$), let j-th entry in $\theta_{\mathbf{i}}$ to be 1 and other entry to be zero. We also set any other fiber expect $\theta_{\mathbf{i}}$ to be 0. Then we have:

$$\|X\|_{2\to\infty}^2 = \|X_j\|_2^2 = \|\Theta \times_1 X\|_F^2 \leq (1 + \delta_R) \|\Theta\|_F^2 = (1 + \delta_R)$$

1 Coefficient Estimation Error

According to (2) and RIP property, we can conclude the boundary of estimation error is:

$$\begin{aligned} \|(\hat{\Theta} - \Theta_{true})\|_F^2 &\leq \frac{1}{1 - \delta_{2R}(X)} \|(\hat{\Theta} - \Theta_{true}) \times_1 X\|_F^2 \\ &\leq \frac{2L_\alpha C_2 \|X\|_{2\to\infty}}{\gamma_\alpha (1 - \delta_{2R}(X))} \sqrt{\sum_{k=2}^{K} r_k (\sum_{k=2}^{K} d_k + p)} \|\hat{\Theta} - \Theta_{\text{true}}\|_F \\ \|(\hat{\Theta} - \Theta_{true})\|_F &\leq \frac{2L_\alpha C_2 \sqrt{1 + \delta_R}}{\gamma_\alpha (1 - \delta_{2R}(X))} \sqrt{\sum_{k=2}^{K} r_k (\sum_{k=2}^{K} d_k + p)} \end{aligned}$$

2 Prediction Error

According to RIP, we have:

$$\|(\hat{\Theta} - \Theta_{true})\|_F \leq \frac{1}{\sqrt{1 - \delta_{2R}(X)}} \|(\hat{\Theta} - \Theta_{true}) \times_1 X\|_F$$

According to (2),

$$
\begin{aligned}
\left\|\left(\hat{\Theta}-\Theta_{\text{true}}\right)\times_1 X\right\|_F^2 \leq & \frac{2L_\alpha C_2}{\gamma_\alpha}\|X\|_{2\to\infty}\sqrt{\sum_{k=2}^K r_k(\sum_{k=2}^K d_k+p)}\|\hat{\Theta}-\Theta_{\text{true}}\|_F \\
\leq & \frac{2L_\alpha C_2}{\gamma_\alpha}\sqrt{1+\delta_R}\sqrt{\sum_{k=2}^K r_k(\sum_{k=2}^K d_k+p)}\frac{1}{\sqrt{1-\delta_{2R}(X)}}\|(\hat{\Theta}-\Theta_{true})\times_1 X\|_F \\
\left\|\left(\hat{\Theta}-\Theta_{\text{true}}\right)\times_1 X\right\|_F \leq & \frac{2L_\alpha C_2\sqrt{1+\delta_R}}{\gamma_\alpha\sqrt{1-\delta_{2R}(X)}}\sqrt{\sum_{k=2}^K r_k(\sum_{k=2}^K d_k+p)}
\end{aligned}
$$

According to the Taylor Expansion, we can conclude the prediction error in Frobenius term is:

$$
\begin{aligned}
\|\mathbb{E}[\hat{Y}]-\mathbb{E}[Y]\|_F &= \|f(\Theta_{true}\times_1 X)-f(\hat{\Theta}\times_1 X)\|_F \\
&\leq \frac{2L_\alpha C_2 M\sqrt{1+\delta_R}}{\gamma_\alpha\sqrt{1-\delta_{2R}(X)}}\sqrt{\sum_{k=2}^K r_k(\sum_{k=2}^K d_k+p)}
\end{aligned}
$$

where $M=\text{Sup}_x(f(x))$ and d is link function.

Similarly, we can get the prediction loss in K-L loss and Hellinger distance through Frobenius norm.

References

[1] Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.