Estimation and Prediction Error in Supervised Setting 2.0

Zhuoyan Xu

Aug 11 2019

The general supervised model is:

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

where $\mathcal{G}$ is the low rank core tensor of factorization. $W, N_2, \dots, N_K$ 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:

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

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

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

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

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

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

Then we define:

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

which is the maximum column norm of the matrix.

Use $\mathcal{S}$ denote $L_\alpha^{-1}\mathcal{S}_Y^*(\Theta_{true}\times_1 X)$, use $\tilde{X}$ denote $\frac{X}{\|X\|_\infty}$. Then we have:

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

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

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

Consider:

$$\begin{aligned}
\mathcal{E} =&\mathcal{S}\times_1 \tilde{X}^T\\
\mathcal{E}_{i_1\cdots i_K} =&\sum_{j_1=1}^{d_1}\mathcal{S}_{j_1 i_2\cdots i_K}\tilde{X}_{j_1 i_1}
\end{aligned}$$

Since

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

then for a given X, we have:

$$\mathbb{E}\left[exp\{t\mathcal{S}_{j_1 i_2\cdots i_K}\tilde{X}_{j_1 i_1}\}\right]\le exp\{(\tilde{X}_{j_1 i_1})^2 t^2/2\}$$

Thus, we have:

$$\begin{aligned}
\mathbb{E}\left[exp\{t\mathcal{E}_{i_1\cdots i_K}\}\right] \le& exp\{\frac{t^2}{2}\sum_{j_1=1}^{d_1}(\tilde{X}_{j_1 i_1})^2\}\\
=&e^{t^2/2}
\end{aligned}$$

Therefore, $\mathcal{E}$ is also a random sub-Gaussian tensor that $\forall$ entries $\varepsilon\in\mathcal{E}$, $\varepsilon\in\mathrm{sG}\,(1)$. According to our bounds on Gaussian width, we have:

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

Thus, we have:

$$\begin{aligned}
\left\|\left(\hat{\Theta}-\Theta_{\mathrm{true}}\right)\times_1 X\right\|_F^2 \le& \frac{2L_\alpha}{\gamma_\alpha}\|X\|_\infty\left\langle \mathcal{E},(\Theta-\Theta_{\mathrm{true}})\right\rangle \qquad (1)\\
\le& \frac{2L_\alpha C_2}{\gamma_\alpha}\|X\|_\infty\sqrt{\sum_{k=2}^{K} r_k(\sum_{k=2}^{K} d_k+p)}\|\hat{\Theta}-\Theta_{\mathrm{true}}\|_F \qquad (2)
\end{aligned}$$

1 Coefficient Estimation Error

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

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

2 Prediction Error

According to RIP, we have:

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

According to (2),

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

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

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

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.