# Evidence theory on prediction error

## Zhuoyan Xu

## Aug 10 2019

Consider we have an extra covariate matrix $X^{d1 \times p}$(accounting for features), which contains the information of countries. We want to connect the membership matrix(or factor matrix) A and B with the information in tensor X.

The general form is:

$$\text{logit}\left\{\mathbb{E}\left[\mathcal{Y}^{d_1 d_2 \ldots d_K}\right]\right\} = \Theta = \mathcal{G}^{r_1 r_2 \ldots r_K} \times_1 W_1^{d_1 r_1} \times_2 W_2^{d_2 r_2} \times_3 N_3^{d_2 r_2} \ldots \times_K N_K^{d_K r_K}$$

$$W_1^{d_1 r_1} = X_1^{d_1 p} N_1^{p r_1}$$

$$W_2^{d_2 r_2} = X_1^{d_2 p} N_2^{p r_2}$$

where $\mathcal{G}$ is the low rank core tensor of factorization. $W_1, W_2, \ldots, N_K$ are factor matrices. Without out loss of generality, $N_i$ is the regression coefficient matrix for $X_i$ on $W_i$.

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

$$\Theta = \mathcal{C} \times_1 X_1 \times_2 X_2$$

$$\mathcal{C} = \mathcal{G}^{r_1 r_2 \ldots r_K} \times_1 N_1^{d_1 r_1} \times_2 N_2^{d_2 r_2} \ldots \times_K N_K^{d_K r_K}$$

where $\mathcal{C}$, a tensor with tucker rank $(r_1, \ldots, r_K)$, is our coefficient tensor. $X_i$ is our predictor.

# 1 Frobenius Norm Loss

Without loss of generality, we choose logit as link function, use $\pi = f(\Theta) = \text{logit}^{-1}(\Theta)$ to denote true probability of our tensor.

The log-likelihood function is:

$$\mathcal{L}_\mathcal{Y}(\pi) = \sum_{i_1, \ldots, i_K} \left[ \mathbf{1}_{\left\{y_{i_1, \ldots, i_K} = 1\right\}} \log\left(\pi_{i_1, \ldots, i_K}\right) + \mathbf{1}_{\left\{y_{i_1, \ldots, i_K} = 0\right\}} \log\left\{\pi_{i_1, \ldots, i_K}\right\} \right]$$

Thus we have:

$$\frac{\partial \mathcal{L}_\mathcal{Y}}{\partial \pi_{i_1, \ldots, i_K}} = \frac{1}{\pi_{i_1, \ldots, i_K}} \mathbf{1}_{\left\{y_{i_1, \ldots, i_K} = 1\right\}} - \frac{1}{1 - \pi_{i_1, \ldots, i_K}} \mathbf{1}_{\left\{y_{i_1, \ldots, i_K} = 0\right\}}$$

$$\frac{\partial \mathcal{L}_\mathcal{Y}^2}{\partial \pi_{i_1, \ldots, i_K}^2} = -\frac{\mathbf{1}_{\left\{y_{i_1, \ldots, i_K} = 1\right\}}}{\pi_{i_1, \ldots, i_K}^2} - \frac{\mathbf{1}_{\left\{y_{i_1, \ldots, i_K} = 0\right\}}}{(1 - \pi_{i_1, \ldots, i_K})^2}$$

$$\frac{\partial \mathcal{L}_\mathcal{Y}^2}{\partial \pi_{i_1, \ldots, i_K} \pi_{i_1', \ldots, i_K'}'} = 0 \quad \text{if} \quad (i_1, \ldots, i_K) \neq (i_1', \ldots, i_K')$$

Without loss of generality, for any observation y in binary tensor and its corresponding probability $\pi$, we have:

$$\frac{\partial \mathcal{L}_y}{\partial \pi} = \frac{y}{\pi} - \frac{1-y}{1-\pi} \leq \frac{1}{\pi(1-\pi)}$$

Since the constrain $\theta = \text{logit}(\pi) \leq \alpha$(constrain on the max norm of ground truth tensor) and the symmetric property of link function f, we have:

$$\pi(1-\pi) = f(\theta)(1-f(\theta)) \geq f(\alpha)(1-f(\alpha)) = c$$

where c is a constant. Thus:

$$\frac{\partial \mathcal{L}_y}{\partial \pi} \leq \frac{1}{c}$$

similarly, consider

$$\frac{\partial \mathcal{L}_y^2}{\partial \pi^2} = -\frac{y}{\pi^2} - \frac{1-y}{(1-\pi)^2} \leq -1$$

Define

$$\mathcal{S}_y\left(\pi_{\text{true}}\right) = \left[\left[\frac{\partial \mathcal{L}_y}{\pi_{i_1,\dots,i_K}}\right]\right]\Bigg|_{\pi=\pi_{\text{true}}} \quad \text{and} \quad \mathcal{H}_y\left(\pi_{\text{true}}\right) = \left\|\frac{\partial \mathcal{L}_y^2}{\partial \pi_{i_1,\dots,i_K}\pi_{i'_1,\dots,i'_K}}\right\|\Bigg|_{\pi=\pi_{\text{true}}}$$

Use Taylor expansion and we have:

$$\mathcal{L}_y(\pi) = \mathcal{L}_y\left(\pi_{\text{true}}\right) + \langle \mathcal{S}_y\left(\pi_{\text{true}}\right), \pi - \pi_{\text{true}}\rangle + \frac{1}{2}\text{vec}\left(\pi - \pi_{\text{true}}\right)^T \mathcal{H}_y(\check{\pi})\,\text{vec}\left(\pi - \pi_{\text{true}}\right)$$

Use $\hat{\pi}$ to denote MLE prediction, we have:

$$0 \leq \mathcal{L}_y(\pi) - \mathcal{L}_y\left(\pi_{\text{true}}\right) = \langle \mathcal{S}_y\left(\pi_{\text{true}}\right), \pi - \pi_{\text{true}}\rangle + \frac{1}{2}\text{vec}\left(\pi - \pi_{\text{true}}\right)^T \mathcal{H}_y(\check{\pi})\,\text{vec}\left(\pi - \pi_{\text{true}}\right)$$

$$\leq \langle \mathcal{S}_y\left(\pi_{\text{true}}\right), \pi - \pi_{\text{true}}\rangle - \frac{1}{2}\left\|\pi - \pi_{\text{true}}\right\|_F^2$$

Apply the theorem on Gaussian width, we have our prediction error:

$$\|\hat{\pi} - \pi\|_F \leq \frac{2C_2}{f(\alpha)(1-f(\alpha))}\sqrt{\prod_{k=1}^{K-1} r_k \sum_{k=1}^{K} d_k}$$

## 2   KL-Divergence

Consider $\pi = f(\Theta)$, where f is the link function. We use Without loss of generality, for any observation $y_i$ in binary tensor and its corresponding ground truth $\theta_i$ and probability $\pi_i$, we have:

$$\mathcal{L}_y(\Theta) = \sum_i \left\{y_i \log[f(\theta)] + (1-y_i)\log[1-f(\theta)])\right\}$$

Use $\Theta^*$ to denote the true parameter, then we have:

$$\mathcal{L}_\mathcal{Y}(\Theta) - \mathcal{L}_\mathcal{Y}(\Theta^*) = \sum_i \left\{ y_i \log\left[\frac{f(\theta)}{f(\theta^*)}\right] + (1-y_i)\log\left[\frac{1-f(\theta)}{1-f(\theta^*)}\right]\right) \right\}$$

Thus, we have:

$$\mathbb{E}_{\Theta^*}\{\mathcal{L}_\mathcal{Y}(\Theta) - \mathcal{L}_\mathcal{Y}(\Theta^*)\} = \sum_{y_i=1}\left\{ f(\theta^*)\log\left[\frac{f(\theta)}{f(\theta^*)}\right]\right\} + \sum_{y_i=0}\left\{ [1-f(\theta^*)]\log\left[\frac{1-f(\theta)}{1-f(\theta^*)}\right]\right\}$$

$$= \sum_i \left\{ \pi^*\log\left[\frac{\pi}{\pi^*}\right]\right\}$$

$$= -D_{\mathrm{KL}}(\pi^*\|\pi)$$

Thus we have:

$$\mathbb{E}_{\pi^*}\{\mathcal{L}_\mathcal{Y}(\pi) - \mathcal{L}_\mathcal{Y}(\pi^*)\} = -D_{\mathrm{KL}}(\pi^*\|\pi)$$

Recalling Taylor expansion in previous section:

$$\mathcal{L}_\mathcal{Y}(\pi) = \mathcal{L}_\mathcal{Y}(\pi^*) + \langle S_\mathcal{Y}(\pi^*), \pi - \pi^* \rangle + \frac{1}{2}\mathrm{vec}\,(\pi - \pi^*)^T \mathcal{H}_\mathcal{Y}(\breve{\pi})\,\mathrm{vec}\,(\pi - \pi^*)$$

Recalling the definition of $S_\mathcal{Y}(\pi^*)$ and $\mathcal{H}_\mathcal{Y}(\breve{\pi})$, we have

$$\mathbb{E}_{\pi^*} S_\mathcal{Y}(\pi^*) = 0$$

And for $\forall \pi_i$, we have:

$$-\mathbb{E}_{\pi^*}\left[\frac{\partial\mathcal{L}_y^2}{\partial\pi_i^2}\right] = \mathbb{E}_{\pi^*}\left[\frac{y}{\pi_i^2} + \frac{1-y}{(1-\pi_i)^2}\right] = \left[\frac{\pi^*}{\pi_i^2} + \frac{1-\pi^*}{(1-\pi_i)^2}\right]$$

Similar to max norm on ground truth tensor $\Theta$, we have

$$-\mathbb{E}_{\pi^*}\left[\frac{\partial\mathcal{L}_y^2}{\partial\pi_i^2}\right] = \left[\frac{\pi^*}{\pi_i^2} + \frac{1-\pi^*}{(1-\pi_i)^2}\right] \leq \frac{2}{\pi_i^2(1-\pi_i)^2} \leq \frac{2}{f(\alpha)^2(1-f(\alpha))^2}$$

Thus:

$$D_{\mathrm{KL}}(\pi^*\|\pi) = -\mathbb{E}_{\pi^*}\{\mathcal{L}_\mathcal{Y}(\pi) - \mathcal{L}_\mathcal{Y}(\pi^*)\}$$

$$= -\frac{1}{2}\mathbb{E}_{\pi^*}\{\mathrm{vec}\,(\pi - \pi^*)^T \mathcal{H}_\mathcal{Y}(\breve{\pi})\,\mathrm{vec}\,(\pi - \pi^*)\}$$

$$\leq \frac{1}{f(\alpha)^2(1-f(\alpha))^2}\|\hat{\pi} - \pi\|_F^2$$

According to our result in previous section, we have:

$$D_{\mathrm{KL}}(\pi^*\|\pi) \leq \frac{1}{f(\alpha)^2(1-f(\alpha))^2}\|\hat{\pi} - \pi\|_F^2 \leq \frac{4C_2^2}{f(\alpha)^4(1-f(\alpha))^4}\prod_{k=1}^{K-1} r_k \sum_{k=1}^{K} d_k$$

3

# 3    Hellinger Loss

According to the definition of Hellinger loss, we have:

$$
\begin{aligned}
d_H^2(\pi, \pi^*) &= \sum_{i=1} \left\{ (\sqrt{\pi_i} - \sqrt{\pi_i^*})^2 + (\sqrt{1 - \pi_i} - \sqrt{1 - \pi_i^*})^2 \right\} \\
&\leq \sum_{i=1} \left\{ |\sqrt{\pi_i} - \sqrt{\pi_i^*}|(\sqrt{\pi_i} + \sqrt{\pi_i^*}) + |\sqrt{1 - \pi_i} - \sqrt{1 - \pi_i^*}|(\sqrt{1 - \pi_i} + \sqrt{1 - \pi_i^*}) \right\} \\
&= 2 \sum_{i=1} |\pi_i - \pi_i^*|
\end{aligned}
$$

According to Pinsker's inequality, we have:

$$
\sum_{i=1} |\pi_i - \pi_i^*| \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P\|Q)}
$$

Thus, we have:

$$
d_H^2(\pi, \pi^*) \leq \sqrt{2 D_{\mathrm{KL}}(P\|Q)} \leq \frac{2C_2}{f(\alpha)^2 (1 - f(\alpha))^2} \sqrt{2 \prod_{k=1}^{K-1} r_k \sum_{k=1}^{K} d_k}
$$