Supplement Proof for the Hubness Problem

Chao Zhao

January 26, 2018

We denote by $||\cdot||_F$ and $||\cdot||$ the Frobenius norm and the spectral norm of matrices. As already seen in Section 2.3, we want to get the $\mathbf{W}$ by ridge regression:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}} ||\mathbf{XW}-\mathbf{Y}||_F + \lambda||\mathbf{W}||_F \tag{1}$$

which has an analytical solution $\hat{\mathbf{W}} = (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$.

Based on the statement that the spectral norm of a centered matrix can be interpreted as an indicator of the variance of data along its principal axis, Shigeto et al. (2015) suggests that to show $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{W}}$ has lower variance to the original $\mathbf{Y}$, we just need to show $||\hat{\mathbf{Y}}|| < ||\mathbf{Y}||$. They also give a sketchy proof. We write up a more detailed one here. One more difference is that we use the Frobenius norm, rather than the spectral norm to depict the variance of the whole matrix, since the spectral norm can only depict the variance along the principal axis.

Proposition 1. *For two matrices $\mathbf{X}\in\mathbb{R}^{n\times d_1}$ and $\mathbf{Y}\in\mathbb{R}^{n\times d_2}$, we learn a mapping matrix $\mathbf{W}$ using the ridge regression: $\hat{\mathbf{W}} = \operatorname{argmin}_{\mathbf{W}\in\mathbb{R}^{d_1\times d_2}} ||\mathbf{XW}-\mathbf{Y}||_F+\lambda||\mathbf{W}||_F$, where $\lambda>0$, then we have $||\mathbf{X}\hat{\mathbf{W}}||_F < ||\mathbf{Y}||_F$.*

Proof. We first plug the solution of $\hat{\mathbf{W}}$ in $\mathbf{X}\hat{\mathbf{W}}$:

$$\begin{aligned} ||\mathbf{X}\hat{\mathbf{W}}||_F &= ||\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}||_F \\ &\leq ||\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T||\cdot||\mathbf{Y}||_F. \end{aligned} \tag{2}$$

To obtain the last step, we see that for two matrices $\mathbf{A}$ and $\mathbf{B}$, $||\mathbf{AB}||_F^2 = \sum_j ||\mathbf{AB}_j||^2 \leq \sum_j ||\mathbf{A}||^2\cdot||\mathbf{B}_j||^2 = ||\mathbf{A}||^2\sum_j||\mathbf{B}_j||^2 = ||\mathbf{A}||_2^2||\mathbf{B}||_F^2$.

We write the singular value decomposition of matrix $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{USV}^T. \tag{3}$$

Note that both $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices. In other words, $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$. Then

$$\mathbf{X^T X} = \mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{V}\mathbf{S}^2\mathbf{V}^T, \tag{4}$$

And therefore

$$\begin{aligned}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}) &= \mathbf{V}\mathbf{S}^2\mathbf{V}^T + \lambda\mathbf{V}\mathbf{I}\mathbf{V}^T \\ &= \mathbf{V}(\mathbf{S}^2 + \lambda\mathbf{I})\mathbf{V}^T.\end{aligned} \tag{5}$$

Now we plug in:

$$\begin{aligned}\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T &= \mathbf{U}\mathbf{S}\mathbf{V}^T[\mathbf{V}(\mathbf{S}^2 + \lambda\mathbf{I})\mathbf{V}^T]^{-1}\mathbf{V}\mathbf{S}^T\mathbf{U}^T \\ &= \mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}^{-T}(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^{-1}\mathbf{V}\mathbf{S}^T\mathbf{U}^T \\ &= \mathbf{U}\underbrace{\mathbf{S}(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\mathbf{S}^T}_{\mathbf{\Lambda}}\mathbf{U}^T.\end{aligned} \tag{6}$$

Note that $\mathbf{S}$ is an $n \times d_1$ matrix with diagonal entries $(\alpha_1, \cdots, \alpha_{d_1})$, it is easy to show that $\mathbf{\Lambda} = \mathbf{S}(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\mathbf{S}^T$ is also a $n \times d_1$ diagonal matrix with entries $(\frac{\alpha_1^2}{\alpha_1^2 + \lambda}, \cdots, \frac{\alpha_n^2}{\alpha_n^2 + \lambda})$. Then,

$$||\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T|| = ||\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T|| = \frac{\alpha_1^2}{\alpha_1^2 + \lambda} < 1 \tag{7}$$

And therefore $||\hat{\mathbf{Y}}||_F < ||\mathbf{Y}||_F$ holds. □

If $\lambda = 0$, then $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, and $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is exactly the matrix of the orthogonal projection onto the column space of $\mathbf{X}$, as the authors point out. And therefore the conclusion that $\hat{\mathbf{Y}}$ has lower variance than $\mathbf{Y}$ can also be explained in this way.

References

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 135–151. Springer, 2015.