Supplementary Material: Representational Similarity Learning

1. Extending methods to matrices that are not Positive Semi-Definite

For any symmetric $\boldsymbol{S}$ we have an eigendecomposition $\boldsymbol{UDU}^T$, where the eigenvalues may be negative. We can think of $\boldsymbol{W}$ as having the form $\boldsymbol{BDB}^T$ for some $\boldsymbol{B}$ and using the $\boldsymbol{D}$ from $\boldsymbol{S}$.

So we consider the following optimizations instead.

$$\min_{\boldsymbol{B}} \|\boldsymbol{S} - \boldsymbol{XBDBX}^T\|_F^2 \ ,$$

and

$$\min_{\boldsymbol{B}} \|\boldsymbol{U} - \boldsymbol{XB}\|_F^2 \ .$$

Then given a $\boldsymbol{B}$ construct $\boldsymbol{W} = \boldsymbol{BDB}^T$. The rest remains the same.

2. Clustering properties of GrOWL with absolute error loss function

Proof of Theorem 3.1

Proof. The proof is divided into two steps. First, we show $\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| = \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$ and then we further show that the rows are equal. We proceed by contradiction. Assume $\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| \neq \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$ and, without loss of generality, suppose $\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| > \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$. We see that there exists a modification of the solution with a smaller GrOWL norm and same data-fitting term, and thus smaller overall objective value which contradicts our assumption that $\widehat{\boldsymbol{B}}$ is the minimizer of $L(\boldsymbol{B}) + G(\boldsymbol{B})$.

Consider the modification, $\boldsymbol{V} = \widehat{\boldsymbol{B}}$ except $\widehat{\boldsymbol{v}}_{j\cdot} = \widehat{\boldsymbol{\beta}}_{j\cdot} - \boldsymbol{\varepsilon}$ and $\widehat{\boldsymbol{v}}_{k\cdot} = \widehat{\boldsymbol{\beta}}_{k\cdot} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} = \delta\widehat{\boldsymbol{\beta}}_{j\cdot}$ and δ is chosen such that $\|\boldsymbol{\varepsilon}\| \in \big(0, \frac{\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| - \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|}{2}\big]$
Let $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{XB}\|_1 = \|\boldsymbol{Y}' - \boldsymbol{x}_{\cdot j}\widehat{\boldsymbol{\beta}}_{j\cdot} - \boldsymbol{x}_{\cdot k}\widehat{\boldsymbol{\beta}}_{k\cdot}\|_1$ where $\boldsymbol{Y}'$ is the residual term given by $\boldsymbol{Y}' = \boldsymbol{Y} - \sum_{i\neq j,k} \boldsymbol{x}_{\cdot i}\widehat{\boldsymbol{\beta}}_{i\cdot}$. Since $\boldsymbol{x}_{\cdot j} = \boldsymbol{x}_{\cdot k}$, L is invariant under this transformation, *i.e.,* $L(\boldsymbol{V}) = L(\widehat{\boldsymbol{B}})$. Same is true for $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{XB}\|_F^2$.

Observe that the GrOWL norm of $\boldsymbol{B}$ is equal to the OWL norm of the vector of euclidean norms of rows of $\boldsymbol{B}$. Since $\|\boldsymbol{v}_{k\cdot}\| = \|\boldsymbol{\beta}_{k\cdot} + \boldsymbol{\varepsilon}\| \leq \|\boldsymbol{\beta}_{k\cdot}\| + \|\boldsymbol{\varepsilon}\|$, this transformation is equivalent to that defined in Lemma 3.1 and we have

$$G(\widehat{\boldsymbol{B}}) - G(\boldsymbol{V}) \quad \geq \quad \Delta\|\boldsymbol{\varepsilon}\|$$

This leads to a contradiction to our assumption that $\widehat{\boldsymbol{B}}$ is the minimizer of $L(\boldsymbol{B}) + G(\boldsymbol{B})$ and completes the proof that $\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| = \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$. Now, let $\widehat{\boldsymbol{\beta}}_{j\cdot} + \widehat{\boldsymbol{\beta}}_{k\cdot} = \boldsymbol{z}$, then the minimizer satisfies

$$\min_{\widehat{\boldsymbol{\beta}}_{j\cdot}, \widehat{\boldsymbol{\beta}}_{k\cdot}} w_j\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| + w_k\|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$$

$$\text{such that } \widehat{\boldsymbol{\beta}}_{j\cdot} + \widehat{\boldsymbol{\beta}}_{k\cdot} = \boldsymbol{z} \text{ and } \|\widehat{\boldsymbol{\beta}}_{j\cdot}\| = \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$$

It is easy to see that the solution to this optimization is $\widehat{\boldsymbol{\beta}}_{j\cdot} = \widehat{\boldsymbol{\beta}}_{k\cdot} = \boldsymbol{z}/2$ □

Proof of Theorem 3.2

Proof. The proof is similar to the identical columns theorem. By contradiction and without loss of generality, suppose $\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| > \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$. We show that there exists a transformation of $\widehat{\boldsymbol{B}}$ such that the increase in the data fitting term is smaller than the decrease in the GrOWL norm.

Consider the modification, $\boldsymbol{V}$, as defined in the proof of Theorem 3.1. By triangle inequality, the difference in loss function L that results from this modification satisfies

$$L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) \quad \leq \quad \big\|\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k}\big\|_1 \|\boldsymbol{\varepsilon}\|_1$$

Invoking Lemma 3.1 as in the previous theorem and $\|\boldsymbol{\varepsilon}\|_1 \leq \sqrt{r}\|\boldsymbol{\varepsilon}\|$, we get

$$\begin{aligned} L(\boldsymbol{V}) + G(\boldsymbol{V}) &- (L(\widehat{\boldsymbol{B}}) + G(\widehat{\boldsymbol{B}})) \\ &\leq \sqrt{r}(\big\|\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k}\big\|_1 - \frac{\Delta}{\sqrt{r}})\|\boldsymbol{\varepsilon}\| < 0 \end{aligned}$$

This contradicts our assumption that $\widehat{\boldsymbol{B}}$ is the minimizer of $L(\boldsymbol{B}) + G(\boldsymbol{B})$ and completes the proof for absolute loss. The proof with squared Frobenius loss can easily be extended using the inequality derived in Appendix B. □

Proof of Theorem 3.3

Proof. The proof is similar to the identical columns theorem. By contradiction, suppose $\|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\| \geq \frac{8\phi\|\widehat{\boldsymbol{\beta}}_{k\cdot}\|}{4\phi^2+1} \geq \frac{2\|\widehat{\boldsymbol{\beta}}_{k\cdot}\|}{\phi}$. We show that there exists a transformation of $\widehat{\boldsymbol{B}}$ such that the increase in the data fitting term is smaller than the decrease in the GrOWL norm.

Consider the modification, $\boldsymbol{V}$, as defined in the proof of Theorem 3.1 with $\boldsymbol{\varepsilon} = \frac{\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}}{2}$. By triangle inequality, the difference in loss function L that results from this modification satisfies

$$L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) \quad \leq \quad \left\|\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k}\right\|_1 \|\boldsymbol{\varepsilon}\|_1$$

We now bound the decrease in the GrOWL norm. Note by parallelogram law,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}}_{j\cdot} + \widehat{\boldsymbol{\beta}}_{k\cdot}\|^2 \\
&= 2\|\widehat{\boldsymbol{\beta}}_{j\cdot}\|^2 + 2\|\widehat{\boldsymbol{\beta}}_{k\cdot}\|^2 - \|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\|^2 \\
&\leq 2\|\widehat{\boldsymbol{\beta}}_{j\cdot}\|^2 + 2\|\widehat{\boldsymbol{\beta}}_{k\cdot}\|^2 + \left(\frac{1}{4\phi^2} - \frac{1}{4\phi^2} - 1\right)\|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\|^2 \\
&\leq 4\|\widehat{\boldsymbol{\beta}}_{j\cdot}\|^2 + \left(\frac{\|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\|}{2\phi}\right)^2 - \frac{1+4\phi^2}{4\phi^2}\|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\|^2 \\
&\leq 4\|\widehat{\boldsymbol{\beta}}_{j\cdot}\|^2 + \left(\frac{\|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\|}{2\phi}\right)^2 - 2\frac{\|\widehat{\boldsymbol{\beta}}_{j\cdot}\|\|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\|}{\phi} \\
&\leq \left(\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| + \|\widehat{\boldsymbol{\beta}}_{k\cdot}\| - \frac{\|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\|}{2\phi}\right)^2
\end{aligned}$$

Thus, we have

$$\begin{aligned}
G(\widehat{\boldsymbol{B}}) - G(\boldsymbol{V}) &\geq \Delta\left(\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| + \|\widehat{\boldsymbol{\beta}}_{k\cdot}\| - \|\widehat{\boldsymbol{\beta}}_{j\cdot} + \widehat{\boldsymbol{\beta}}_{k\cdot}\|\right) \\
&\geq \frac{\Delta\|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\|}{2\phi} = \frac{\Delta\|\boldsymbol{\varepsilon}\|}{\phi}
\end{aligned}$$

Combining this with $\|\boldsymbol{\varepsilon}\|_1 \leq \sqrt{r}\|\boldsymbol{\varepsilon}\|$, we get

$$\begin{aligned}
L(\boldsymbol{V}) + G(\boldsymbol{V}) - &(L(\widehat{\boldsymbol{B}}) + G(\widehat{\boldsymbol{B}})) \\
&\leq \left(\sqrt{r}\left\|\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k}\right\|_1 - \frac{\Delta}{\phi}\right)\|\boldsymbol{\varepsilon}\| < 0
\end{aligned}$$

This contradicts our assumption that $\widehat{\boldsymbol{B}}$ is the minimizer of $L(\boldsymbol{B}) + G(\boldsymbol{B})$ and completes the proof for absolute loss. The proof with squared Frobenius loss can easily be extended using the inequality derived in Appendix B. $\square$

3. Clustering properties of GrOWL with squared Frobenius loss function

In this section, we consider the optimization

$$\min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 \ + \ G(\boldsymbol{B}) \tag{1}$$

Here we derive an upper bound on the increase in the squared loss term after applying the transformation, $\boldsymbol{V}$. We assume that the columns of the matrix, $\boldsymbol{X}$, are normalized to a common norm, *i.e.,* ($\|\boldsymbol{x}_{\cdot i}\| = c$ for $i = 1, \cdots, p$). Define $L(\boldsymbol{X}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 = \|\boldsymbol{Y}' - \boldsymbol{x}_{\cdot j}\boldsymbol{\beta}_{j\cdot} - \boldsymbol{x}_{\cdot k}\boldsymbol{\beta}_{k\cdot}\|_F^2$ where $\boldsymbol{Y}'$ is again the residual term.

Lemma 1. *Let $\widehat{\boldsymbol{B}} \in \mathbb{R}^{p\times r}$ and if $\boldsymbol{V}$ is as defined in the respective theorems, then we have*

$$L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) \leq \|\boldsymbol{\varepsilon}\|\|\boldsymbol{Y}'\|_F\|\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k}\|$$

Proof.

$$\begin{aligned}
L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) = &\frac{1}{2}\|\boldsymbol{Y}' - \boldsymbol{x}_{\cdot j}(\widehat{\boldsymbol{\beta}}_{j\cdot} - \boldsymbol{\varepsilon}) - \boldsymbol{x}_{\cdot k}(\widehat{\boldsymbol{\beta}}_{k\cdot} + \boldsymbol{\varepsilon})\|_F^2 \\
&- \frac{1}{2}\|\boldsymbol{Y}' - \boldsymbol{x}_{\cdot j}\widehat{\boldsymbol{\beta}}_{j\cdot} - \boldsymbol{x}_{\cdot k}\widehat{\boldsymbol{\beta}}_{k\cdot}\|_F^2
\end{aligned}$$

Expanding the Frobenius norm terms, canceling the common $\frac{1}{2}\|\boldsymbol{Y}'\|_F^2$ terms and using the common norm of columns ($\|\boldsymbol{x}_{\cdot i}\| = c$ for $i = 1, \cdots, p$) we get

$$\begin{aligned}
&L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) \\
&= \frac{c^2}{2}\mathrm{tr}((\widehat{\boldsymbol{\beta}}_{j\cdot} - \boldsymbol{\varepsilon})(\widehat{\boldsymbol{\beta}}_{j\cdot} - \boldsymbol{\varepsilon})^T + (\widehat{\boldsymbol{\beta}}_{k\cdot} + \boldsymbol{\varepsilon})(\widehat{\boldsymbol{\beta}}_{k\cdot} + \boldsymbol{\varepsilon})^T \\
&- \widehat{\boldsymbol{\beta}}_{j\cdot}\widehat{\boldsymbol{\beta}}_{j\cdot}^T - \widehat{\boldsymbol{\beta}}_{k\cdot}\widehat{\boldsymbol{\beta}}_{k\cdot}^T) + \mathrm{tr}(\boldsymbol{Y}'^T(\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k})\boldsymbol{\varepsilon}) \\
&+ \mathrm{tr}((\widehat{\boldsymbol{\beta}}_{j\cdot} - \boldsymbol{\varepsilon})\boldsymbol{x}_{\cdot j}^T\boldsymbol{x}_{\cdot k}(\widehat{\boldsymbol{\beta}}_{k\cdot} + \boldsymbol{\varepsilon})^T - \widehat{\boldsymbol{\beta}}_{j\cdot}\boldsymbol{x}_{\cdot j}^T\boldsymbol{x}_{\cdot k}\widehat{\boldsymbol{\beta}}_{k\cdot}^T)
\end{aligned}$$

Expanding terms and making further cancellations gives

$$\begin{aligned}
&L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) \\
&= \mathrm{tr}(\boldsymbol{Y}'^T(\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k})\boldsymbol{\varepsilon}) - (c^2 - \boldsymbol{x}_{\cdot j}^T\boldsymbol{x}_{\cdot k})\,\mathrm{tr}((\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot} - \boldsymbol{\varepsilon})\boldsymbol{\varepsilon}^T) \\
&\leq \mathrm{tr}(\boldsymbol{Y}'^T(\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k})\boldsymbol{\varepsilon}) \\
&- (c^2 - \boldsymbol{x}_{\cdot j}^T\boldsymbol{x}_{\cdot k})\|\boldsymbol{\varepsilon}\|(\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| - \|\widehat{\boldsymbol{\beta}}_{k\cdot}\| - \|\boldsymbol{\varepsilon}\|) \\
&\leq \mathrm{tr}(\boldsymbol{Y}'^T(\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k})\boldsymbol{\varepsilon}^T) \\
&\leq \|\boldsymbol{Y}'\|_F\|(\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k})\boldsymbol{\varepsilon}\|_F \\
&= \|\boldsymbol{\varepsilon}\|\|\boldsymbol{Y}'\|_F\|\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k}\|
\end{aligned}$$

where the first inequality follows from simplification and Cauchy-Schwarz inequality. The second inequality follows from $c^2 > \boldsymbol{x}_{\cdot j}^T\boldsymbol{x}_{\cdot k}$ and $\|\widehat{\boldsymbol{\beta}}_{j\cdot}\|_2 - \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|_2 - \|\boldsymbol{\varepsilon}\| > 0$ (by assumption). The third inequality follows, again, by Cauchy-Schwarz inequality.

$\square$

Using this Lemma one can easily extend the clustering properties of GrOWL to the optimization in (1).

4. Proximal algorithms for GrOWL

Proof. Outline: the proof proceeds by finding a lower bound for the objective function in (5) and then we show that the proposed solution achieves this lower bound.

First, note that the following is true for any $\boldsymbol{B}$ and $\boldsymbol{V}$,

$$\begin{aligned}\|\boldsymbol{B}-\boldsymbol{V}\|_F^2 &= \sum_{i=1}^{p} \|\boldsymbol{\beta}_{i\cdot} - \boldsymbol{v}_{i\cdot}\|^2 \\ &\geq \sum_{i=1}^{p} (\|\boldsymbol{\beta}_{i\cdot}\| - \|\boldsymbol{v}_{i\cdot}\|)^2 = \|\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{v}}\|^2\end{aligned}$$

where the inequality follows from reverse triangle inequality.

Combining this with $G(\boldsymbol{B}) = \Omega_{\boldsymbol{w}}(\tilde{\boldsymbol{\beta}})$, we have a lower bound on the objective function in (5). For all $\boldsymbol{B} \in \mathbb{R}^{p\times r}$

$$\frac{1}{2}\|\boldsymbol{B}-\boldsymbol{V}\|_F^2 + G(\boldsymbol{B}) \geq \frac{1}{2}\|\mathrm{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}) - \tilde{\boldsymbol{v}}\|^2 + \Omega_{\boldsymbol{w}}(\mathrm{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}))$$

Finally, we show that $\boldsymbol{B} = \widehat{\boldsymbol{V}}$ achieves this lower bound,

$$\begin{aligned}&\frac{1}{2}\|\widehat{\boldsymbol{V}}-\boldsymbol{V}\|_F^2 + G(\widehat{\boldsymbol{V}}) \\ &= \frac{1}{2}\sum_{i=1}^{p} \|(\mathrm{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}))_i \frac{\boldsymbol{v}_{i\cdot}}{\|\boldsymbol{v}_{i\cdot}\|} - \boldsymbol{v}_{i\cdot}\|_2^2 + \Omega_{\boldsymbol{w}}(\mathrm{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}})) \\ &= \frac{1}{2}\|\mathrm{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}) - \tilde{\boldsymbol{v}}\|_2^2 + \Omega_{\boldsymbol{w}}(\mathrm{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}))\end{aligned}$$

$\square$