Secrets of Matrix Factorization:
Further Derivations and Comparisons

Je Hyeong Hong
University of Cambridge
jhh37@cam.ac.uk

Andrew Fitzgibbon
Microsoft Research Cambridge
awf@microsoft.com

Abstract

This report illustrates the derivatives used by several state-of-the-art algorithms mentioned in [?] and compares each implementation in detail.

1. Derivatives for joint optimization

1.1. The Jacobian matrix $\mathtt{J}$

The Jacobian matrix for joint optimization is defined as $\mathtt{J} := \partial\boldsymbol{\varepsilon}/\partial\mathbf{x}$ where $\mathbf{x} := [\mathbf{u}; \mathbf{v}]$. From [?], we have

$$\mathtt{J} := \begin{bmatrix} \dfrac{\partial\boldsymbol{\varepsilon}_1(\mathbf{u},\mathbf{v})}{\partial\mathbf{u}} & \dfrac{\partial\boldsymbol{\varepsilon}_1(\mathbf{u},\mathbf{v})}{\partial\mathbf{v}} \\ \dfrac{\partial\boldsymbol{\varepsilon}_2(\mathbf{u})}{\partial\mathbf{u}} & \\ & \dfrac{\partial\boldsymbol{\varepsilon}_3(\mathbf{v})}{\partial\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \tilde{\mathtt{V}} & \tilde{\mathtt{U}} \\ \{\sqrt{\mu}\mathtt{I}\} & \\ & \{\sqrt{\mu}\mathtt{I}\} \end{bmatrix}. \tag{1}$$

1.2. The gradient vector $\mathbf{g}$

The gradient vector is defined as $\mathbf{g} := [\partial f/\partial\mathbf{x}]^\top$ where $\mathbf{x} := [\mathbf{u}; \mathbf{v}]$. Since $f(\mathbf{u},\mathbf{v}) := \|\boldsymbol{\varepsilon}(\mathbf{u},\mathbf{v})\|_2^2$, the gradient can be obtained by computing $2\mathtt{J}^\top\boldsymbol{\varepsilon}$ as follows:

$$\frac{1}{2}\mathbf{g} = \begin{bmatrix} \tilde{\mathtt{V}}^\top & \{\sqrt{\mu}\mathtt{I}\} & \\ \tilde{\mathtt{U}}^\top & & \{\sqrt{\mu}\mathtt{I}\} \end{bmatrix} \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \{\sqrt{\mu}\mathbf{u}\} \\ \{\sqrt{\mu}\mathbf{v}\} \end{bmatrix} = \begin{bmatrix} \tilde{\mathtt{V}}^\top\boldsymbol{\varepsilon}_1 + \{\mu\mathbf{u}\} \\ \tilde{\mathtt{U}}^\top\boldsymbol{\varepsilon}_1 + \{\mu\mathbf{v}\} \end{bmatrix} \tag{2}$$

1.3. The Hessian matrix $\mathtt{H}$

The Hessian matrix is defined as $\mathtt{H} := \partial\mathbf{g}/\partial\mathbf{x}$ with $\mathbf{x} = [\mathbf{u}; \mathbf{v}]$. This yields

$$\frac{1}{2}\mathtt{H} = \frac{1}{2}\begin{bmatrix} \dfrac{\partial\mathbf{g}}{\partial\mathbf{u}} & \dfrac{\partial\mathbf{g}}{\partial\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \tilde{\mathtt{V}}^\top\dfrac{\partial\boldsymbol{\varepsilon}_1}{\partial\mathbf{u}} + \{\mu\mathtt{I}\} & \tilde{\mathtt{V}}^\top\dfrac{\partial\boldsymbol{\varepsilon}_1}{\partial\mathbf{v}} + \dfrac{(\partial\tilde{\mathtt{V}})^\top\boldsymbol{\varepsilon}_1}{\partial\mathbf{v}} \\ \tilde{\mathtt{U}}^\top\dfrac{\partial\boldsymbol{\varepsilon}_1}{\partial\mathbf{u}} + \dfrac{(\partial\tilde{\mathtt{U}})^\top\boldsymbol{\varepsilon}_1}{\partial\mathbf{u}} & \tilde{\mathtt{U}}^\top\dfrac{\partial\boldsymbol{\varepsilon}_1}{\partial\mathbf{v}} + \{\mu\mathtt{I}\} \end{bmatrix}. \tag{3}$$

Applying tricks from [?] and [?] gives us

$$\begin{aligned}
\partial[\tilde{\mathtt{U}}]^\top \boldsymbol{\varepsilon}_1 &= (\mathtt{I}\otimes\partial\mathtt{U}^\top)\tilde{\mathtt{W}}^\top\boldsymbol{\varepsilon}_1 && \tag{4}\\
&= (\mathtt{I}\otimes\partial\mathtt{U}^\top)\operatorname{diag}(\operatorname{vec}\mathtt{W})\operatorname{vec}\mathtt{R} && /\!/\ \mathtt{R}(\mathtt{U},\mathtt{V}) := \mathtt{U}\mathtt{V}^\top - \mathtt{M} \tag{5}\\
&= (\mathtt{I}\otimes\partial\mathtt{U}^\top)\operatorname{vec}(\mathtt{W}\odot\mathtt{R}) && \tag{6}\\
&= \operatorname{vec}(\partial\mathtt{U}^\top(\mathtt{W}\odot\mathtt{R})) && /\!/\ \operatorname{vec}(\mathtt{AXB}) = (\mathtt{B}^\top\otimes\mathtt{A})\operatorname{vec}(\mathtt{X}) \tag{7}\\
&= ((\mathtt{W}\odot\mathtt{R})^\top\otimes\mathtt{I}_r)\operatorname{vec}(\partial\mathtt{U}^\top) && \tag{8}\\
&= \mathtt{Z}^\top\mathtt{K}_{mr}\partial\mathbf{u}, && /\!/\ \mathtt{Z} := (\mathtt{W}\odot\mathtt{R})\otimes\mathtt{I}_r,\ \ \operatorname{vec}(\partial\mathtt{U}^\top) = \mathtt{K}_{mr}\partial\mathbf{u} \tag{9}
\end{aligned}$$

and similarly

$$\begin{aligned}
\partial[\tilde{\mathtt{V}}]^\top \boldsymbol{\varepsilon}_1 &= (\partial\mathtt{V}^\top\otimes\mathtt{I})\tilde{\mathtt{W}}^\top\boldsymbol{\varepsilon}_1 && \tag{10}\\
&= (\partial\mathtt{V}^\top\otimes\mathtt{I})\operatorname{vec}(\mathtt{W}\odot\mathtt{R}) && \tag{11}\\
&= \operatorname{vec}((\mathtt{W}\odot\mathtt{R})\partial\mathtt{V}) && /\!/\ \operatorname{vec}(\mathtt{AXB}) = (\mathtt{B}^\top\otimes\mathtt{A})\operatorname{vec}(\mathtt{X}) \tag{12}\\
&= (\mathtt{I}_r\otimes(\mathtt{W}\odot\mathtt{R}))\operatorname{vec}(\partial\mathtt{V}) && \tag{13}\\
&= \left(\mathtt{K}_{mr}^\top((\mathtt{W}\odot\mathtt{R})\otimes\mathtt{I}_r)\mathtt{K}_{nr}\right)\mathtt{K}_{nr}^\top\operatorname{vec}(\partial\mathtt{V}^\top) && \tag{14}\\
&= \mathtt{K}_{mr}^\top\mathtt{Z}\partial\mathbf{v}. && /\!/\ \mathtt{Z} := (\mathtt{W}\odot\mathtt{R})\otimes\mathtt{I}_r \tag{15}
\end{aligned}$$

We observe that the above two terms in (9) and (15) arise when computing the exact Hessian as they do not appear in $\mathtt{J}^\top\mathtt{J}$. Combining these results with damping completes our unified derivation for joint optimization, yielding the colour-coded Hessian, or approximation, as

$$\frac{1}{2}\mathtt{H} = \begin{bmatrix} \tilde{\mathtt{V}}^\top\tilde{\mathtt{V}} + \{\mu\mathtt{I}\} + \langle\lambda\mathtt{I}\rangle & \tilde{\mathtt{V}}^\top\tilde{\mathtt{U}} + [\mathtt{K}_{mr}^\top\mathtt{Z}]_{FN} \\ \tilde{\mathtt{U}}^\top\tilde{\mathtt{V}} + [\mathtt{Z}^\top\mathtt{K}_{mr}]_{FN} & \tilde{\mathtt{U}}^\top\tilde{\mathtt{U}} + \{\mu\mathtt{I}\} + \langle\lambda\mathtt{I}\rangle \end{bmatrix}, \tag{16}$$

where $\mathtt{K}_{mr}$ and $\mathtt{Z}(\mathtt{U},\mathtt{V})$ are defined in [?]. This is equivalent to the Hessian matrix used by Buchanan and Fitzgibbon's Damped Newton algorithm [?].

2. Derivatives for variable projection

From [?], we have

$$\mathbf{v}^*(\mathbf{u}) := \arg\min_{\mathbf{v}} f(\mathbf{u},\mathbf{v}) = \tilde{\mathtt{U}}^{-\mu}\tilde{\mathbf{m}}. \tag{17}$$

2.1. The derivative of $\mathbf{v}^*(\mathbf{u})$

Substituting (17) to the original objective in [?] yields

$$\boldsymbol{\varepsilon}_1^*(\mathbf{u},\mathbf{v}^*(\mathbf{u})) = \tilde{\mathtt{U}}\mathbf{v}^* - \tilde{\mathbf{m}} = -(\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^{-\mu})\tilde{\mathbf{m}}. \tag{18}$$

Applying the μ-pseudo inverse rule in [?] yields

$$\begin{aligned}
\partial[\mathbf{v}^*] &= -\tilde{\mathtt{U}}^{-\mu}\partial[\tilde{\mathtt{U}}]\tilde{\mathtt{U}}^{-\mu}\tilde{\mathbf{m}} \\
&\quad + (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}} + \mu\mathtt{I})^{-1}\partial[\tilde{\mathtt{U}}]^\top(\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^{-\mu})\tilde{\mathbf{m}} && \tag{19}\\
&= -\tilde{\mathtt{U}}^{-\mu}\partial[\tilde{\mathtt{U}}]\mathbf{v}^* - (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}} + \mu\mathtt{I})^{-1}\partial[\tilde{\mathtt{U}}]^\top\boldsymbol{\varepsilon}_1^* && /\!/\ \textit{noting (17) and (18)} \tag{20}\\
&= -\tilde{\mathtt{U}}^{-\mu}\tilde{\mathtt{V}}^*\partial\mathbf{u} - (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}} + \mu\mathtt{I})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr}\partial\mathbf{u}, && /\!/\ \textit{noting bilinearity and using techniques in (9)} \tag{21}
\end{aligned}$$

and hence

$$\frac{d\mathbf{v}^*}{d\mathbf{u}} = -(\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}} + \mu\mathtt{I})^{-1}(\tilde{\mathtt{U}}^\top\tilde{\mathtt{V}}^* + \mathtt{Z}^{*\top}\mathtt{K}_{mr}). \tag{22}$$

2.2. The Jacobian matrix $\mathtt{J}^*$

The Jacobian matrix for variable projection is defined as

$$\mathtt{J}^* := \frac{d\boldsymbol{\varepsilon}^*(\mathbf{u},\mathbf{v}^*(\mathbf{u}))}{d\mathbf{u}} = \begin{bmatrix} \dfrac{d\boldsymbol{\varepsilon}_1^*(\mathbf{u},\mathbf{v}^*(\mathbf{u}))}{d\mathbf{u}} \\ \dfrac{d\boldsymbol{\varepsilon}_2^*(\mathbf{u})}{d\mathbf{u}} \\ \dfrac{d\boldsymbol{\varepsilon}_3^*(\mathbf{v}^*(\mathbf{u}))}{d\mathbf{u}} \end{bmatrix}. \tag{23}$$

Noting that

$$\frac{d\boldsymbol{\varepsilon}_1^*(\mathbf{u},\mathbf{v}^*(\mathbf{u}))}{d\mathbf{u}} = \left(\frac{\partial\boldsymbol{\varepsilon}_1^*}{\partial\mathbf{u}}\right) + \left(\frac{\partial\boldsymbol{\varepsilon}_1^*}{\partial\mathbf{v}^*}\right)\frac{d\mathbf{v}^*}{d\mathbf{u}}, \tag{24}$$

we obtain

$$\mathtt{J}^* = \begin{bmatrix} \tilde{\mathtt{V}}^* + \tilde{\mathtt{U}}\dfrac{d\mathbf{v}^*}{d\mathbf{u}} \\ \left\{\sqrt{\mu}\mathtt{I}\right\} \\ \left\{\sqrt{\mu}\dfrac{d\mathbf{v}^*}{d\mathbf{u}}\right\} \end{bmatrix}. \tag{25}$$

2.3. The gradient vector $\mathbf{g}^*$

The gradient for variable projection, $\mathbf{g}^*$ can be obtained by simplifying $2\mathtt{J}^{*\top}\boldsymbol{\varepsilon}^*$:

$$\frac{1}{2}\mathbf{g}^* = \left(\tilde{\mathtt{V}}^* + \tilde{\mathtt{U}}\frac{d\mathbf{v}^*}{d\mathbf{u}}\right)^\top \boldsymbol{\varepsilon}_1^* + \left\{\mu\mathbf{u} + \mu\left(\frac{d\mathbf{v}^*}{d\mathbf{u}}\right)^\top \mathbf{v}^*\right\} \qquad /\!/\ \boldsymbol{\varepsilon}_2^* = \mathbf{u},\ \ \boldsymbol{\varepsilon}_3^* = \mathbf{v}^*(\mathbf{u}) \tag{26}$$

2.4. The Gauss-Newton matrix $\mathtt{H}_{GN}^*$

Unlike for joint optimization, Hessian for regularized variable projection, $\mathtt{H}^* := d\mathbf{g}^*/d\mathbf{u}$, is difficult to compute analytically due to the presence of the term

$$\frac{d\left[\frac{d\mathbf{v}^*}{d\mathbf{u}}\right]^\top}{d\mathbf{u}}\mathbf{v}^*, \tag{27}$$

which exists when regularization is on. Boumal et al. [?] were able to bypass this issue by taking the directional derivative of the gradient, which is an efficient way to implement an iterative solver for the sub-problem. However, they used a different type of regularizer to ours. Nevertheless, we can obtain the damped Gauss-Newton matrix $\mathtt{H}_{GN}^*$ from $\mathtt{J}^{*\top}\mathtt{J}^*$:

$$\frac{1}{2}\mathtt{H}_{GN}^* = \left(\tilde{\mathtt{V}}^* + \tilde{\mathtt{U}}\frac{d\mathbf{v}^*}{d\mathbf{u}}\right)^\top \left(\tilde{\mathtt{V}}^* + \tilde{\mathtt{U}}\frac{d\mathbf{v}^*}{d\mathbf{u}}\right) + \left\{\mu\mathtt{I} + \mu\left(\frac{d\mathbf{v}^*}{d\mathbf{u}}\right)^\top\left(\frac{d\mathbf{v}^*}{d\mathbf{u}}\right)\right\} + \langle\lambda\mathtt{I}\rangle \tag{28}$$

A summary of this section can be found in Table 1. We will show in Section 3.5 that obtaining an analytic form for unregularized Hessian ($\mu = 0$) is less tricky.

3. Derivatives for un-regularized VarPro

This section illustrates the analytic derivatives used by un-regularized VarPro ($\mu = 0$) and their relations to Ruhe and Wedin's algorithms [?] applied to low-rank matrix completion. Having $\mu = 0$ simplifies $\mathbf{v}^*(\mathbf{u})$ in (17) to

$$\mathbf{v}^*(\mathbf{u}) = \tilde{\mathtt{U}}^\dagger\tilde{\mathbf{m}}. \tag{29}$$

Quantity	Definition	Gauss-Newton w/o {Regularization} w/o ⟨Damping⟩
$\mathbf{v}^*(\mathbf{u})$ $\mathbf{v}^* \in \mathbb{R}^{nr}$	$\mathbf{v}^* := \arg\min_{\mathbf{v}} f(\mathbf{u}, \mathbf{v})$	$\mathbf{v}^* = (\tilde{\mathtt{U}}^\top \tilde{\mathtt{U}} + \{\mu \mathtt{I}_{nr}\})^{-1} \tilde{\mathtt{U}}^\top \tilde{\mathbf{m}}$ ($\tilde{\mathbf{m}} \in \mathbb{R}^p$ consists of non-zero elements of $\tilde{\mathtt{W}}\mathbf{m}$.)
Derivative of $\mathbf{v}^*$ $\frac{d\mathbf{v}^*}{d\mathbf{u}} \in \mathbb{R}^{mr \times nr}$	$\frac{d\mathbf{v}^*}{d\mathbf{u}} := \frac{d\,\mathrm{vec}(\mathtt{V}^{*\top})}{d\,\mathrm{vec}(\mathtt{U})}$	$\frac{d\mathbf{v}^*}{d\mathbf{u}} = -(\tilde{\mathtt{U}}^\top \tilde{\mathtt{U}} + \{\mu \mathtt{I}_{nr}\})^{-1} (\tilde{\mathtt{U}}^\top \tilde{\mathtt{V}}^* + \mathtt{Z}^{*\top} \mathtt{K}_{mr})$ $(\mathtt{Z}^* := (\mathtt{W} \odot \mathtt{R}^*) \otimes \mathtt{I}_r)$
Cost vector $\boldsymbol{\varepsilon}^* \in \mathbb{R}^{p+\{N\}}$	$\boldsymbol{\varepsilon}^* := \begin{bmatrix} \boldsymbol{\varepsilon}_1^* \\ \left\{ \begin{matrix} \boldsymbol{\varepsilon}_2^* \\ \boldsymbol{\varepsilon}_3^* \end{matrix} \right\} \end{bmatrix}$	$\boldsymbol{\varepsilon}^* = \begin{bmatrix} \tilde{\mathtt{U}}\mathbf{v}^* - \mathbf{m} \\ \left\{ \begin{matrix} \mathbf{u} \\ \mathbf{v}^* \end{matrix} \right\} \end{bmatrix}$
Jacobian $\mathtt{J}^* \in \mathbb{R}^{p+\{N\} \times mr}$	$\mathtt{J}^* := \begin{bmatrix} \frac{d\boldsymbol{\varepsilon}_1^*}{d\mathbf{u}} \\ \left\{ \begin{matrix} \frac{d\boldsymbol{\varepsilon}_2^*}{d\mathbf{u}} \\ \frac{d\boldsymbol{\varepsilon}_3^*}{d\mathbf{u}} \end{matrix} \right\} \end{bmatrix}$	$\mathtt{J}^* = \begin{bmatrix} \tilde{\mathtt{V}}^* + \tilde{\mathtt{U}} \frac{d\mathbf{v}^*}{d\mathbf{u}} \\ \left\{ \begin{matrix} \sqrt{\mu} \mathtt{I}_{mr} \\ \sqrt{\mu} \frac{d\mathbf{v}^*}{d\mathbf{u}} \end{matrix} \right\} \end{bmatrix}$
Cost function $f^* \in \mathbb{R}$	$f^* := \lVert \boldsymbol{\varepsilon}^* \rVert_2^2$	$f^* = \lVert \mathtt{W} \odot (\mathtt{U}\mathtt{V}^{*\top} - \mathtt{M}) \rVert_F + \{\mu \lVert \mathtt{U} \rVert_F + \mu \lVert \mathtt{V}^* \rVert_F\}$
Gradient $\mathbf{g}^* \in \mathbb{R}^{mr}$	$\mathbf{g}^* := \frac{df^*}{d\mathbf{u}}$	$\frac{1}{2}\mathbf{g}^* = (\tilde{\mathtt{V}}^* + \tilde{\mathtt{U}} \frac{d\mathbf{v}^*}{d\mathbf{u}})^\top \boldsymbol{\varepsilon}_1^* + \left\{ \mu \left(\frac{d\mathbf{v}^*}{d\mathbf{u}} \right)^\top \mathbf{v}^* + \mu \mathbf{u} \right\}$
Gauss-Newton matrix $\mathtt{H}^* \in \mathbb{S}^{mr}$	$\mathtt{H}^* := \frac{d\mathbf{g}^*}{d\mathbf{u}} + \langle \lambda \mathtt{I} \rangle$	$\frac{1}{2}\mathtt{H}^* = (\tilde{\mathtt{V}}^* + \tilde{\mathtt{U}} \frac{d\mathbf{v}^*}{d\mathbf{u}})^\top (\tilde{\mathtt{V}}^* + \tilde{\mathtt{U}} \frac{d\mathbf{v}^*}{d\mathbf{u}})$ $+ \left\{ \mu \left(\frac{d\mathbf{v}^*}{d\mathbf{u}} \right)^\top \left(\frac{d\mathbf{v}^*}{d\mathbf{u}} \right) + \mu \mathtt{I}_{mr} \right\} + \langle \lambda \mathtt{I}_{mr} \rangle$

Table 1: Derivatives for variable projection applied to low-rank matrix factorization with missing data

3.1. The derivative of $\mathbf{v}^*(\mathbf{u})$

Deleting μ-related terms in (22) gives

$$\frac{d\mathbf{v}^*}{d\mathbf{u}} = -(\tilde{\mathtt{U}}^\top \tilde{\mathtt{U}})^{-1} (\tilde{\mathtt{U}}^\top \tilde{\mathtt{V}}^* + \mathtt{Z}^{*\top} \mathtt{K}_{mr}) \tag{30}$$

$$= -\tilde{\mathtt{U}}^\dagger \tilde{\mathtt{V}}^* - (\tilde{\mathtt{U}}^\top \tilde{\mathtt{U}})^{-1} \mathtt{Z}^{*\top} \mathtt{K}_{mr}. \tag{31}$$

3.2. The Jacobian matrix $\mathtt{J}_1^*$

Removing μ terms in (25) and noting (31) yields

$$\mathtt{J}_1^* = \tilde{\mathtt{V}}^* + \tilde{\mathtt{U}} \frac{d\mathbf{v}^*}{d\mathbf{u}} \tag{32}$$

$$= (\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger) \tilde{\mathtt{V}}^* - \tilde{\mathtt{U}}^{\dagger\top} \mathtt{Z}^{*\top} \mathtt{K}_{mr}. \tag{33}$$

3.3. The gradient vector $\mathbf{g}_1^*$

Deleting μs and noting (26), (33) and (29) gives the gradient vector:

$$\frac{1}{2}\mathbf{g}_1^* = \tilde{\mathtt{V}}^{*\top} (\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger) \boldsymbol{\varepsilon}_1^* + \mathtt{K}_{mr}^\top \mathtt{Z}^* \tilde{\mathtt{U}}^\dagger \boldsymbol{\varepsilon}_1^* \tag{34}$$

$$= \tilde{\mathtt{V}}^{*\top} \boldsymbol{\varepsilon}_1^* \qquad /\!/\ \tilde{\mathtt{U}}^\dagger \boldsymbol{\varepsilon}_1^* = 0 \ \ \textit{since} \ \ \boldsymbol{\varepsilon}_1^* = \tilde{\mathtt{U}}\mathbf{v}^* - \tilde{\mathbf{m}} \tag{35}$$

Above can also be rearranged into matrix to produce $\nabla_{\mathtt{U}} f_1^*(\mathtt{U})$:

$$\frac{1}{2}\nabla_{\mathtt{U}} f_1^*(\mathtt{U}) = \mathrm{unvec}(\tilde{\mathtt{V}}^{*\top} \boldsymbol{\varepsilon}_1^*) = (\mathtt{W} \odot \mathtt{R}^*)\mathtt{V}^* \qquad /\!/\ \textit{similar to (13)} \tag{36}$$

3.4. The Gauss-Newton matrix $\mathtt{H}^*_{GN1}$

The Gauss-Newton matrix can be obtained by computing $\mathtt{J}_1^{*\top}\mathtt{J}_1^*$:

$$\begin{aligned}\frac{1}{2}\mathtt{H}^*_{GN1} &= \tilde{\mathtt{V}}^{*\top}(\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^* + \mathtt{K}_{mr}^\top\mathtt{Z}^*\tilde{\mathtt{U}}^\dagger\tilde{\mathtt{U}}^{\dagger\top}\mathtt{Z}^{*\top}\mathtt{K}_{mr} \\ &\quad + \tilde{\mathtt{V}}^{*\top}(\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{U}}^{\dagger\top}\mathtt{Z}^{*\top}\mathtt{K}_{mr} + \mathtt{K}_{mr}^\top\mathtt{Z}^*\tilde{\mathtt{U}}^\dagger(\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^* \quad /\!/\ \mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger = (\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)^\top = (\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)^2 \qquad (37) \\ &= \tilde{\mathtt{V}}^{*\top}(\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^* + \mathtt{K}_{mr}^\top\mathtt{Z}^*(\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr} \quad /\!/\ \tilde{\mathtt{U}}^\dagger\tilde{\mathtt{U}}^{\dagger\top} = (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\ \textit{and}\ \tilde{\mathtt{U}}^\dagger\tilde{\mathtt{U}} = \mathtt{I} \qquad (38)\end{aligned}$$

3.5. The Hessian matrix $\mathtt{H}_1^*$

Taking the partial derivative of $\mathbf{g}_1^*$ yields

$$\begin{aligned}\frac{1}{2}\partial[\mathbf{g}_1^*] &= \partial[\tilde{\mathtt{V}}^*]^\top\boldsymbol{\varepsilon}_1^* + \tilde{\mathtt{V}}^{*\top}\partial[\boldsymbol{\varepsilon}_1^*] \qquad (39) \\ &= \mathtt{K}_{mr}^\top\mathtt{Z}^*\partial\mathbf{v}^* + \tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{U}}\partial\mathbf{v}^* + \tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{V}}^*\partial\mathbf{u}, \quad /\!/\ \textit{noting bilinearity and using techniques in (15)} \qquad (40)\end{aligned}$$

and hence

$$\begin{aligned}\frac{1}{2}\mathtt{H}_1^* &= \tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{V}}^* + \tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{U}}\frac{d\mathbf{v}^*}{d\mathbf{u}} + \mathtt{K}_{mr}^\top\mathtt{Z}^*\frac{d\mathbf{v}^*}{d\mathbf{u}} \qquad (41) \\ &= \tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{V}}^* - \tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger\tilde{\mathtt{V}}^* - \tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{U}}^{\dagger\top}\mathtt{Z}^{*\top}\mathtt{K}_{mr} - \mathtt{K}_{mr}^\top\mathtt{Z}^*\tilde{\mathtt{U}}^\dagger\tilde{\mathtt{V}}^* - \mathtt{K}_{mr}^\top\mathtt{Z}^*(\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr}. \qquad (42)\end{aligned}$$

Colourizing the above expression and adding damping factor $\langle\lambda\mathtt{I}\rangle$ gives

$$\frac{1}{2}\mathtt{H}_1^* = \tilde{\mathtt{V}}^{*\top}(\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^* + [-1]_{FN} \times \mathtt{K}_{mr}^\top\mathtt{Z}^*(\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr} - [\tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{U}}^{\dagger\top}\mathtt{Z}^{*\top}\mathtt{K}_{mr} + \mathtt{K}_{mr}^\top\mathtt{Z}^*\tilde{\mathtt{U}}^\dagger\tilde{\mathtt{V}}^*]_{FN} + \langle\lambda\mathtt{I}\rangle. \qquad (43)$$

3.6. Column-wise derivatives

Some algorithms [?, ?] use what we refer to as column-wise derivatives. We may view the unregularized objective, $f_1^*(\mathbf{u},\mathbf{v}^*(\mathbf{u}))$, as the the sum of squared norm of individual columns in $\mathtt{R}^*(\mathbf{u},\mathbf{v}^*(\mathbf{u}))$. In other words,

$$f_1^*(\mathbf{u},\mathbf{v}^*(\mathbf{u})) := \|\mathtt{R}^*\|_F^2 = \sum_{j=1}^n \|\mathbf{r}_j^*\|_2^2 = \sum_{j=1}^n \|\mathbf{w}_j \odot (\mathtt{U}\mathbf{v}_j^* - \mathbf{m}_j)\|_2^2, \qquad (44)$$

where $\mathbf{r}_j^* \in \mathbb{R}^m$ is the j-th column of $\mathtt{R}^*$, $\mathbf{w}_j \in \mathbb{R}^m$ is the j-th column of $\mathtt{W}$, $\mathbf{v}_j^* \in \mathbb{R}^r$ is the j-th row of $\mathtt{V}^*(\mathtt{U})$ and $\mathbf{m}_j \in \mathbb{R}^m$ is the j-th column of $\mathtt{M}$. Using identities in [?] gives

$$f_1^*(\mathbf{u},\mathbf{v}^*(\mathbf{u})) = \sum_{j=1}^n \|\Pi_j\,\mathrm{diag}(\mathbf{w}_j)(\mathtt{U}\mathbf{v}_j^* - \mathbf{m}_j)\|_2^2 := \sum_{j=1}^n \|\tilde{\mathtt{W}}_j(\mathtt{U}\mathbf{v}_j^* - \mathbf{m}_j)\|_2^2, \qquad (45)$$

where, for each j, Π_j is a fixed $p_j \times m$ projector matrix, where p_j is the number of visible elements in the j-th column, which eliminates known-zero entries from $\boldsymbol{\varepsilon}_1^*$. $\tilde{\mathtt{W}}_j$ is defined as $\Pi_j\,\mathrm{diag}(\mathbf{w}_j)$. Now define

$$\begin{aligned}\tilde{\mathtt{U}}_j &:= \tilde{\mathtt{W}}_j\mathtt{U} \qquad (46) \\ \mathbf{u}_j &:= \mathrm{vec}(\tilde{\mathtt{U}}_j) = \mathrm{vec}(\tilde{\mathtt{W}}_j\mathtt{U}) = (\mathtt{I}\otimes\tilde{\mathtt{W}}_j)\mathbf{u} \qquad (47) \\ \tilde{\mathbf{m}}_j &:= \tilde{\mathtt{W}}_j\mathbf{m}_j, \text{ and} \qquad (48) \\ \boldsymbol{\varepsilon}_{1j}^* &:= \tilde{\mathtt{U}}_j\mathbf{v}_j^* - \tilde{\mathbf{m}}_j. \qquad (49)\end{aligned}$$

By combining the results from (45) and (49), it is evident that

$$f_1^*(\mathbf{u},\mathbf{v}^*(\mathbf{u})) = \sum_{j=1}^n \|\boldsymbol{\varepsilon}_{1j}^*(\mathbf{u}_j,\mathbf{v}_j^*(\mathbf{u}))\|_2^2. \qquad (50)$$

Hence,

$$\mathbf{v}_j^* := \arg\min_{\mathbf{v}_j} \sum_{j=1}^{n} \|\tilde{\mathtt{U}}_j \mathbf{v}_j^* - \tilde{\mathbf{m}}_j\|_2^2 \tag{51}$$

$$= \arg\min_{\mathbf{v}_j} \|\tilde{\mathtt{U}}_j \mathbf{v}_j^* - \tilde{\mathbf{m}}_j\|_2^2 = \tilde{\mathtt{U}}_j^{\dagger} \tilde{\mathbf{m}}_j, \tag{52}$$

which shows that each row of $\mathtt{V}^*$ is independent of other rows. Also, this leads to

$$\boldsymbol{\varepsilon}_{1j}^* = -(\mathtt{I} - \tilde{\mathtt{U}}_j \tilde{\mathtt{U}}_j^{\dagger})\tilde{\mathbf{m}}_j. \tag{53}$$

3.6.1 The derivative of $\mathbf{v}_j^*(\mathbf{u})$

We define

$$\tilde{\mathtt{V}}_j^* := \tilde{\mathtt{W}}_j(\mathbf{v}_j^{*\top} \otimes \mathtt{I}_m) \in \mathbb{R}^{p_j \times mr} \text{ and} \tag{54}$$

$$\mathtt{Z}_j^* := (\mathbf{w}_j \odot \mathbf{r}_j^*) \otimes \mathtt{I}_r \in \mathbb{R}^{mr \times r}. \tag{55}$$

Computing the the derivative of $\mathbf{v}_j^*$ in a similar way to (20) yields

$$\partial[\mathbf{v}_j^*] = -\tilde{\mathtt{U}}_j^{\dagger} \partial[\tilde{\mathtt{U}}_j]\mathbf{v}_j^* - (\tilde{\mathtt{U}}_j^{\top} \tilde{\mathtt{U}}_j)^{-1} \partial[\tilde{\mathtt{U}}_j]^{\top} \boldsymbol{\varepsilon}_{1j}^* \tag{56}$$

$$= -\tilde{\mathtt{U}}_j^{\dagger} \tilde{\mathtt{V}}_j^* \partial\mathbf{u} - (\tilde{\mathtt{U}}_j^{\top} \tilde{\mathtt{U}}_j)^{-1} \mathtt{Z}_j^{*\top} \mathtt{K}_{mr} \partial\mathbf{u}, \tag{57}$$

and hence

$$\frac{d\mathbf{v}_j^*}{d\mathbf{u}} = -\tilde{\mathtt{U}}_j^{\dagger} \tilde{\mathtt{V}}_j^* - (\tilde{\mathtt{U}}_j^{\top} \tilde{\mathtt{U}}_j)^{-1} \mathtt{Z}_j^{*\top} \mathtt{K}_{mr} \tag{58}$$

$$= -\tilde{\mathtt{U}}_j^{\dagger} \tilde{\mathtt{V}}_j^* - (\tilde{\mathtt{U}}_j^{\top} \tilde{\mathtt{U}}_j)^{-1} (\mathtt{I} \otimes (\mathbf{w}_j \odot \mathbf{r}_j^*)^{\top}) \tag{59}$$

$$= -\tilde{\mathtt{U}}_j^{\dagger} \tilde{\mathtt{V}}_j^* - (\tilde{\mathtt{U}}_j^{\top} \tilde{\mathtt{U}}_j)^{-1} \otimes (\mathbf{w}_j \odot \mathbf{r}_j^*)^{\top}. \tag{60}$$

Above is the column-wise derivative of $\mathbf{v}_j^*$ with respect to entire $\mathbf{u}$. However, we note from (52) that not all entries of $\mathbf{u}$ are used in $\mathbf{v}_j^*$ for each j. Hence, (60) can be truncated.

Taking the derivative with respect to $\mathbf{u}_j$ yields

$$(\mathtt{I} \otimes \tilde{\mathtt{W}}_j) \frac{d\mathbf{u}}{d\mathbf{u}_j} = \mathtt{I}. \tag{61}$$

Since $\tilde{\mathtt{W}}_j \tilde{\mathtt{W}}_j^{\top} = \mathtt{I}$,

$$\frac{d\mathbf{u}}{d\mathbf{u}_j} = \mathtt{I} \otimes \tilde{\mathtt{W}}_j^{\top}. \tag{62}$$

This gives the truncated derivative of $\mathbf{v}_j^*$

$$\frac{d\mathbf{v}_j^*}{d\mathbf{u}_j} = \frac{d\mathbf{v}_j^*}{d\mathbf{u}} \frac{d\mathbf{u}}{d\mathbf{u}_j} = \frac{d\mathbf{v}_j^*}{d\mathbf{u}} (\mathtt{I} \otimes \tilde{\mathtt{W}}_j^{\top}). \tag{63}$$

3.6.2 The column-wise Jacobian matrix $\mathtt{J}_{1j}^*$

Since the Jacobian matrix is independent for each column, we can obtain an expression for this, $\mathtt{J}_{1j}^* \in \mathbb{R}^{p_j \times mr}$, as follows:

$$\mathtt{J}_{1j}^* = \tilde{\mathtt{V}}_j^* + \tilde{\mathtt{U}}_j \frac{d\mathbf{v}_j^*}{d\mathbf{u}} \tag{64}$$

$$= (\mathtt{I} - \tilde{\mathtt{U}}_j \tilde{\mathtt{U}}_j^{\dagger})\tilde{\mathtt{V}}_j^* - \tilde{\mathtt{U}}_j^{\dagger\top} \mathtt{Z}_j^{*\top} \mathtt{K}_{mr} \tag{65}$$

$$= (\mathtt{I} - \tilde{\mathtt{U}}_j \tilde{\mathtt{U}}_j^{\dagger})\tilde{\mathtt{V}}_j^* - \tilde{\mathtt{U}}_j^{\top} \left((\tilde{\mathtt{U}}_j^{\top} \tilde{\mathtt{U}}_j)^{-1} \otimes (\mathbf{w}_j \odot \mathbf{r}_j^*)^{\top} \right) \tag{66}$$

Also, the truncated form of the above is

$$\frac{d\boldsymbol{\varepsilon}_{1j}^*}{d\mathbf{u}_j} = \mathtt{J}_{1j}^* \frac{d\mathbf{u}}{d\mathbf{u}_j} = \mathtt{J}_{1j}^* (\mathtt{I} \otimes \tilde{\mathtt{W}}_j^{\top}). \tag{67}$$

3.6.3 The column-wise gradient vector $\mathbf{g}_1^*$

Noting $\mathbf{g}_1^* := df_1^*/d\mathbf{u}$ yields

$$\frac{1}{2}\mathbf{g}_1^* = \frac{1}{2}\frac{d}{d\mathbf{u}}\sum_{j=1}^{n}\|\boldsymbol{\varepsilon}_{1j}^*\|_2^2 \tag{68}$$

$$= \sum_{j=1}^{n}\left(\frac{d\boldsymbol{\varepsilon}_{1j}^*}{d\mathbf{u}}\right)^\top \boldsymbol{\varepsilon}_{1j}^* = \sum_{j=1}^{n}\mathtt{J}_{1j}^{*\top}\boldsymbol{\varepsilon}_{1j}^* = \sum_{j=1}^{n}\tilde{\mathtt{V}}_j^{*\top}\boldsymbol{\varepsilon}_{1j}^* \qquad \textit{// using (53)} \tag{69}$$

As shown by Chen [**?**], above can also be represented as the sum of truncated gradient-layers as follows:

$$\frac{1}{2}\mathbf{g}_1^* = \sum_{j=1}^{n}\left(\frac{d\boldsymbol{\varepsilon}_{1j}^*}{d\mathbf{u}_j}\frac{d\mathbf{u}_j}{d\mathbf{u}}\right)^\top \boldsymbol{\varepsilon}_{1j}^* \tag{70}$$

$$= \sum_{j=1}^{n}\left(\frac{d\mathbf{u}_j}{d\mathbf{u}}\right)^\top (\mathtt{I}\otimes\tilde{\mathtt{W}}_j)\mathtt{J}_{1j}^{*\top}\boldsymbol{\varepsilon}_{1j}^* \qquad \textit{// noting (67)} \tag{71}$$

$$= \sum_{j=1}^{n}(\mathtt{I}\otimes\tilde{\mathtt{W}}_j)^\top(\mathtt{I}\otimes\tilde{\mathtt{W}}_j)\mathtt{J}_{1j}^{*\top}\boldsymbol{\varepsilon}_{1j}^* \qquad \textit{// noting } \mathbf{u}_j = (\mathtt{I}\otimes\tilde{\mathtt{W}}_j)\mathbf{u} \tag{72}$$

$$:= \frac{1}{2}\sum_{j=1}^{n}(\mathtt{I}\otimes\tilde{\mathtt{W}}_j^\top)\mathbf{g}_{1j}^* \tag{73}$$

where we define

$$\frac{1}{2}\mathbf{g}_{1j}^* : = \left(\frac{d\boldsymbol{\varepsilon}_{1j}^*}{d\mathbf{u}_j}\right)^\top \boldsymbol{\varepsilon}_{1j}^*. \tag{74}$$

3.6.4 The column-wise Hessian matrix $\mathtt{H}_1^*$

From the definition $\mathtt{H}_1^* := d\mathbf{g}_1^*/d\mathbf{u}$, the summation in $\mathbf{g}_1^*$ applies to $\mathtt{H}_1^*$ also. Noting the analytical structure of $\mathtt{J}_{1j}^{*\top}\mathtt{J}_{1j}^{*\top}$, (43) can be converted to

$$\frac{1}{2}\mathtt{H}_1^* = \sum_{j=1}^{n}\Big[\tilde{\mathtt{V}}_j^{*\top}(\mathtt{I}-\tilde{\mathtt{U}}_j\tilde{\mathtt{U}}_j^\dagger)\tilde{\mathtt{V}}_j^* + [-1]_{FN}\times\mathtt{K}_{mr}^\top\mathtt{Z}_j^*(\tilde{\mathtt{U}}_j^\top\tilde{\mathtt{U}}_j)^{-1}\mathtt{Z}_j^{*\top}\mathtt{K}_{mr}$$

$$-[\tilde{\mathtt{V}}_j^{*\top}\tilde{\mathtt{U}}_j^{\dagger\top}\mathtt{Z}_j^{*\top}\mathtt{K}_{mr}+\mathtt{K}_{mr}^\top\mathtt{Z}_j^*\tilde{\mathtt{U}}_j^\dagger\tilde{\mathtt{V}}_j^*]_{FN}\Big] + \langle\lambda\mathtt{I}\rangle. \tag{75}$$

All the $\mathtt{Z}_j^{*\top}\mathtt{K}_{mr}$ terms can again be replaced with $\mathtt{I}\otimes(\mathbf{w}_j\odot\mathbf{r}_j^*)$ using the matrix identities shown in [**?**].

As shown by Chen [**?**], above can be represented as the sum of truncated Hessian-layers as follows:

$$\frac{1}{2}\mathtt{H}_1^* := \frac{1}{2}\frac{d\mathbf{g}_1^*}{d\mathbf{u}} + \langle\lambda\mathtt{I}\rangle \tag{76}$$

$$= \frac{1}{2}\sum_{j=1}^{n}\left[(\mathtt{I}\otimes\tilde{\mathtt{W}}_j^\top)\frac{d\mathbf{g}_{1j}^*}{d\mathbf{u}_j}\frac{d\mathbf{u}_j}{d\mathbf{u}}\right] + \langle\lambda\mathtt{I}\rangle \tag{77}$$

$$= \frac{1}{2}\sum_{j=1}^{n}\left[(\mathtt{I}\otimes\tilde{\mathtt{W}}_j^\top)\mathtt{H}_{1j}^*(\mathtt{I}\otimes\tilde{\mathtt{W}}_j)\right] + \langle\lambda\mathtt{I}\rangle \tag{78}$$

where we define

$$\begin{aligned}
\mathtt{H}^*_{1j} &:= \frac{d\mathbf{g}^*_{1j}}{d\mathbf{u}_j} && (79)\\
&= (\mathtt{I} \otimes \tilde{\mathtt{W}}_j)\frac{d(\tilde{\mathtt{V}}_j^{*\top}\boldsymbol{\varepsilon}^*_{1j})}{d\mathbf{u}_j} \qquad \textit{// from (73)} && (80)\\
&= (\mathtt{I} \otimes \tilde{\mathtt{W}}_j)\frac{d\mathbf{g}^*_1}{d\mathbf{u}}\frac{d\mathbf{u}}{d\mathbf{u}_j} && (81)\\
&= (\mathtt{I} \otimes \tilde{\mathtt{W}}_j)\frac{d\mathbf{g}^*_1}{d\mathbf{u}}(\mathtt{I} \otimes \tilde{\mathtt{W}}_j^\top). && (82)
\end{aligned}$$

Here, $\mathtt{H}^*_{1j}$ is the compact j-th layer of the Hessian matrix.

4. Manifold optimization on un-regularized VarPro

The un-regularized variable projection objective has the property $f^*_1(\mathtt{U}) = f^*_1(\mathtt{UA})$ for any invertible matrix $\mathtt{A}$. This means that $\mathtt{U}$ lies on a Grassmann manifold embedded in Euclidean space which is a type of Riemannian manifold with its own metrics.

The book by Absil et al. [?] extensively covers optimization strategies on Riemannian manifolds. Boumal et al. [?] concisely illustrates how to incorporate Grassmann or Stiefel manifold projection into the optimization framework:

1. If the solution is outside the underlying manifold, perform retraction.
2. Project the gradient to the tangent space of current $\mathtt{U}$ by multiplying the manifold projection matrix to the original gradient $\mathbf{g}^*_1$. This gives $\mathbf{g}^*_p$.
3. Project Hessian to the tangent space of current $\mathtt{U}$ by multiplying the manifold projection matrix to the derivative of the projected gradient $\mathbf{g}^*_p$. One may add some damping when using Levenberg-Marquardt, and this yields $\mathtt{H}^*_p$.
4. Obtain solution $\Delta\mathbf{u}$ for the projected sub-problem using a second-order solver:

   $$\Delta\mathbf{u} = \arg\min_{\boldsymbol{\delta}\perp\mathbf{u}} \mathbf{g}_p^{*\top}\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^\top\mathtt{H}^*_p\boldsymbol{\delta} \tag{83}$$

   where $\boldsymbol{\delta}\perp\mathbf{u}$ is the linear constraint $\mathtt{U}^\top\,\mathrm{unvec}(\boldsymbol{\delta}) = 0$. $\Delta\mathbf{u}$ can be unvectorized to give $\Delta\mathtt{U}_p$.
5. Since $\mathtt{U} + \Delta\mathtt{U}_p$ is outside the manifold, perform retraction using geodastic distance or implicit retraction such as q-factor.

Absil et al. [?] states that taking the q-factor of the updated variable is an efficient way of retracting back to either the Stiefel or the Grassmann manifold. i.e.

$$\mathtt{U}^* = qf(\mathtt{U} + \Delta\mathtt{U}_p) \tag{84}$$

For our case of the Grassmann manifold , the original tangent space-projection matrix is $\mathtt{I} - \mathtt{U}(\mathtt{U}^\top\mathtt{U})^{-1}\mathtt{U}^\top$. When using the q factor-based retraction, $\mathtt{U}^\top\mathtt{U} = \mathtt{I}$ and thus the projection matrix is reduced to $\mathtt{I} - \mathtt{UU}^\top$. As we are working with vectorized quantities, we also need to obtain an equivalent form for the projection matrix. Since the following relationship must hold

$$\begin{aligned}
\mathbf{g}^*_p &= \mathrm{vec}\left((\mathtt{I} - \mathtt{UU}^\top)\frac{df^*_1(\mathbf{u}))}{d\mathtt{U}}\right) && (85)\\
&= (\mathtt{I} \otimes (\mathtt{I} - \mathtt{UU}^\top))\,\mathrm{vec}\left(\frac{df^*_1(\mathbf{u})}{d\mathtt{U}}\right) && (86)\\
&= (\mathtt{I} \otimes (\mathtt{I} - \mathtt{UU}^\top))\mathbf{g}^*_1, && (87)
\end{aligned}$$

we can deduce that $\mathtt{P}_p = \mathtt{I} \otimes (\mathtt{I} - \mathtt{UU}^\top) = \mathtt{I} \otimes \mathtt{U}_\perp\mathtt{U}_\perp^\top$.

4.1. The projected gradient vector

We can obtain the projected gradient $\mathbf{g}_p^*$ by multiplying the tangent space-projection matrix $\mathtt{P}_p$ to the original gradient $\mathbf{g}_1^*$. Doing so initially gives us

$$\begin{aligned}
\mathbf{g}_p^* &:= (\mathtt{I} \otimes (\mathtt{I} - \mathtt{UU}^\top))\mathbf{g}_1^* && (88)\\
&= \mathbf{g}_1^* - 2(\mathtt{I} \otimes \mathtt{UU}^\top)\tilde{\mathtt{V}}^{*\top}\boldsymbol{\varepsilon}_1^* && (89)\\
&= \mathbf{g}_1^* - 2(\mathtt{I} \otimes \mathtt{UU}^\top)\operatorname{vec}((\mathtt{W} \odot \mathtt{R}^*)\mathtt{V}^*) && \textit{// noting (36)} \quad (90)\\
&= \mathbf{g}_1^* - 2\operatorname{vec}(\mathtt{UU}^\top(\mathtt{W} \odot \mathtt{W} \odot (\mathtt{UV}^{*\top} - \mathtt{M}))\mathtt{V}^*) && \textit{// } (\mathtt{B}^\top \otimes \mathtt{A})\operatorname{vec}(\mathtt{X}) = \operatorname{vec}(\mathtt{AXB}) \textit{ [?]} \quad (91)\\
&= \mathbf{g}_1^* - 2(\mathtt{V}^{*\top} \otimes \mathtt{U})\operatorname{vec}(\mathtt{U}^\top(\mathtt{W} \odot \mathtt{R}^*)) && (92)
\end{aligned}$$

Vectorizing $\mathtt{U}^\top(\mathtt{W} \odot \mathtt{R}^*)$ yields

$$\begin{aligned}
\operatorname{vec}(\mathtt{U}^\top(\mathtt{W} \odot \mathtt{R}^*)) &= (\mathtt{I} \otimes \mathtt{U}^\top)\operatorname{vec}(\mathtt{W} \odot \mathtt{R}^*) && (93)\\
&= (\mathtt{I} \otimes \mathtt{U}^\top)\operatorname{diag}(\operatorname{vec}\mathtt{W})\operatorname{vec}\mathtt{R}^* && (94)\\
&= (\mathtt{I} \otimes \mathtt{U}^\top)\tilde{\mathtt{W}}^\top\boldsymbol{\varepsilon}_1^* && (95)\\
&= \tilde{\mathtt{U}}^\top(\tilde{\mathtt{U}}\mathbf{v}^* - \tilde{\mathbf{m}}) && (96)\\
&= \tilde{\mathtt{U}}^\top(\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger\tilde{\mathbf{m}} - \tilde{\mathbf{m}}) = 0 && (97)
\end{aligned}$$

giving

$$\mathbf{g}_p^* = \mathbf{g}_1^*. \qquad (98)$$

This means that the gradient is already on the tangent space of $\mathtt{U}$ and thus no additional computation is required.

4.2. The projected Hessian matrix

Projected Hessian $\mathtt{H}_p^*$ can be obtained by multiplying the tangent-space projection matrix $\mathtt{P}_p$ to the derivative of the projected gradient $\mathbf{g}_p^*$. Since we know that $\mathbf{g}_p^* = \mathbf{g}_1^*$, we simply need to project original undamped Hessian to the tangent space of $\mathtt{U}$:

$$\mathtt{H}_p^* := \mathtt{P}_p\frac{d\mathbf{g}_1^*}{d\mathbf{u}} = (\mathtt{I} - \mathtt{I} \otimes \mathtt{UU}^\top)\frac{d\mathbf{g}_1^*}{d\mathbf{u}} \qquad (99)$$

There are 4 terms present in $\mathtt{H}_1^*$ as shown in (43). First, note that

$$\begin{aligned}
(\mathtt{I} \otimes \mathtt{U}^\top)(\tilde{\mathtt{V}}^{*\top}(\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^*) &= (\mathtt{V}^{*\top} \otimes \mathtt{U}^\top)\tilde{\mathtt{W}}^\top(\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^* && \textit{// noting } \tilde{\mathtt{V}}^* = \tilde{\mathtt{W}}(\mathtt{V}^* \otimes \mathtt{I}) \quad (100)\\
&= (\mathtt{V}^{*\top} \otimes \mathtt{I})(\mathtt{I} \otimes \mathtt{U}^\top)\tilde{\mathtt{W}}^\top(\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^* && (101)\\
&= (\mathtt{V}^{*\top} \otimes \mathtt{I})\tilde{\mathtt{U}}^\top(\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^* = 0. && (102)
\end{aligned}$$

Hence, we observe that the first term $(\tilde{\mathtt{V}}^{*\top}(\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^*)$ remains the same after projection. Furthermore, note that

$$\begin{aligned}
(\mathtt{I} \otimes \mathtt{U}^\top)(\mathtt{K}_{mr}^\top\mathtt{Z}^*) &= \mathtt{K}_{mr}^\top(\mathtt{U}^\top \otimes \mathtt{I})\mathtt{K}_{mr}\mathtt{K}_{mr}^\top((\mathtt{W} \odot \mathtt{R}^*) \otimes \mathtt{I}) && (103)\\
&= \mathtt{K}_{mr}^\top(\mathtt{U}^\top(\mathtt{W} \odot \mathtt{R}^*) \otimes \mathtt{I}) = 0. && \textit{// noting (94) – (97)} \quad (104)
\end{aligned}$$

Combining the results of (102) and (104) yields the projected Hessian matrix

$$\begin{aligned}
\frac{1}{2}\operatorname{Hess}f^*(\mathbf{u}) = \tilde{\mathtt{V}}^{*\top}(\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^* + [-1]_{FN} \times \mathtt{K}_{mr}^\top\mathtt{Z}^*(\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr} \qquad (105)\\
- \mathtt{P}_p[\tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{U}}^{\dagger\top}\mathtt{Z}^{*\top}\mathtt{K}_{mr}]_{FN} - [\mathtt{K}_{mr}^\top\mathtt{Z}^*\tilde{\mathtt{U}}^\dagger\tilde{\mathtt{V}}^*]_{FN}.
\end{aligned}$$

Once again, this shows that the projected Gauss-newton matrix, $\mathtt{H}_{pGN}^*$, is analytically identical to the original Gauss-Newton matrix, $\mathtt{H}_{1GN}^*$, as the first two terms in (105) are the same.

Above expression is essentially equal to the directional derivative form of Hessian derived by Boumal et al. [?] with no regularization (see §5.5). However, $\operatorname{Hess}f^*(\mathbf{u})$ is asymmetric. One way to tackle this is to consider the manifold of the subproblem as will be seen in the next section.

4.3. The projected subproblem

The undamped Newton solution $\Delta\mathbf{u}$ to the subproblem mentioned in (83) satisfies

$$\frac{1}{2}\mathrm{Hess}f^*(\mathbf{u})\Delta\mathbf{u} = -\frac{1}{2}\mathbf{g}_p^*, \tag{106}$$

where $\Delta\mathbf{u}$ lies on the tangent space of $\mathtt{U}$. This means that

$$\mathtt{U}^\top\Delta\mathtt{U} = (\mathtt{I}\otimes\mathtt{U}^\top)\Delta\mathbf{u} = 0, \tag{107}$$

leading to

$$(\mathtt{I}\otimes(\mathtt{I}-\mathtt{U}\mathtt{U}^\top))\Delta\mathbf{u} = \mathtt{P}_p\Delta\mathbf{u} = \Delta\mathbf{u}. \tag{108}$$

Substituting the result from (108) to (106) yields

$$\frac{1}{2}\mathrm{Hess}f^*(\mathbf{u})\mathtt{P}_p\Delta\mathbf{u} = -\frac{1}{2}\mathbf{g}_p^*. \tag{109}$$

The matrix product to the left of $\Delta\mathbf{u}$ in (109) is

$$\begin{aligned}\frac{1}{2}\mathrm{Hess}f^*(\mathbf{u})\mathtt{P}_p = {} & \tilde{\mathtt{V}}^{*\top}(\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^* + [-1]_{FN}\times\mathtt{K}_{mr}^\top\mathtt{Z}^*(\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr} \\ & - \mathtt{P}_p[\tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{U}}^{\dagger\top}\mathtt{Z}^{*\top}\mathtt{K}_{mr}]_{FN} - [\mathtt{K}_{mr}^\top\mathtt{Z}^*\tilde{\mathtt{U}}^\dagger\tilde{\mathtt{V}}^*]_{FN}\mathtt{P}_p. \qquad /\!/ \textit{ noting (102) and (104)}\end{aligned} \tag{110}$$

Note that the Gauss-Newton matrix still remains unchanged. We now define the symmetric projected Hessian with damping $\mathtt{H}_p^*$:

$$\frac{1}{2}\mathtt{H}_p^* := \frac{1}{2}\mathrm{Hess}f^*(\mathbf{u})\mathtt{P}_p + \langle\lambda\mathtt{I}\rangle = \frac{1}{2}\mathtt{P}_p\frac{d\mathbf{g}_p^*}{d\mathbf{u}}\mathtt{P}_p + \langle\lambda\mathtt{I}\rangle \tag{111}$$

4.4. Constraining the subproblem

The subproblem mentioned in (83) is rank-deficient since the update is orthogonal to $\mathtt{U}$ and thus the update dimension is r^2 less than that of $\mathbf{u}\in\mathbb{R}^{mr}$. Two approaches are known to handle this issue.

4.4.1 Reducing the subproblem dimension

Originally proposed by Manton et al. [?], the first approach is to reduce the entire subproblem dimension by projecting it to the orthogonal space of $\mathtt{U}$. Noting that $\mathtt{I}-\mathtt{U}\mathtt{U}^\top = \mathtt{U}_\perp\mathtt{U}_\perp^\top$, we define the reduced projection matrix

$$\mathtt{P}_r = \mathtt{I}\otimes\mathtt{U}_\perp^\top \in \mathbb{R}^{(m-r)r\times mr}, \tag{112}$$

which makes $\mathtt{P}_r\mathtt{P}_p = \mathtt{P}_r$. Defining the reduced update $\Delta\mathbf{u}_\perp \in \mathbb{R}^{mr-r^2}$ yields ${\mathtt{P}_r}^\top\Delta\mathbf{u}_\perp = \Delta\mathbf{u}_p$. Hence, projecting the subproblem to the reduced dimension yields

$$\begin{aligned} & \frac{1}{2}\mathtt{P}_r\mathtt{H}_p^*\Delta\mathbf{u}_p = -\frac{1}{2}\mathtt{P}_r\mathbf{g}_p^* && (113)\\ \Rightarrow\quad & -\frac{1}{2}\mathtt{P}_r\mathbf{g}_p^* = \mathtt{P}_r\left(\frac{1}{2}\mathtt{P}_p\frac{d\mathbf{g}_p^*}{d\mathbf{u}}\mathtt{P}_p + \langle\lambda\mathtt{I}\rangle\right)\Delta\mathbf{u}_p && (114)\\ & = \mathtt{P}_r\left(\frac{1}{2}\frac{d\mathbf{g}_p^*}{d\mathbf{u}}\mathtt{P}_p + \langle\lambda\mathtt{I}\rangle\right)({\mathtt{P}_r}^\top\Delta\mathbf{u}_\perp) && (115)\\ & = \mathtt{P}_r\left(\frac{1}{2}\frac{d\mathbf{g}_1^*}{d\mathbf{u}} + \langle\lambda\mathtt{I}\rangle\right){\mathtt{P}_r}^\top\Delta\mathbf{u}_\perp && (116)\\ & = (\frac{1}{2}\mathtt{P}_r\mathtt{H}_1^*{\mathtt{P}_r}^\top + \langle\lambda\mathtt{I}\rangle)\Delta\mathbf{u}_\perp \qquad /\!/\ \mathtt{P}_r{\mathtt{P}_r}^\top = (\mathtt{I}\otimes\mathtt{U}_\perp^\top)(\mathtt{I}\otimes\mathtt{U}_\perp) = \mathtt{I} && (117)\end{aligned}$$

As the update is now forced to be orthogonal to current $\mathtt{U}$, the reduced problem is now better-conditioned than the original at the expense of increased computational complexity. This approach is incorporated to Chen's LM_M series of algorithms [?].

4.4.2 Relaxing the orthogonality constraint

The second approach relaxes the constraint $\mathtt{U}^\top \Delta\mathtt{U} = 0$ by adding a term that promotes the orthogonality of the update to the subproblem as shown:

$$\Delta\mathbf{u}_p = \arg\min_{\boldsymbol{\delta}\perp\mathbf{u}} \mathbf{g}_p^{*\top}\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^\top \mathtt{H}_p^*\boldsymbol{\delta} + \langle \alpha \|\mathtt{U}^\top\Delta\|_F^2\rangle_P \tag{118}$$

where $\Delta \in \mathbb{R}^{m\times r}$ is the unvectorized form of δ. Differentiating $\|\mathtt{U}^\top\Delta\|_F^2$ with respect to δ gives

$$\begin{aligned}\frac{d\|\mathtt{U}^\top\Delta\|_F^2}{d\delta} &= \frac{d\|\operatorname{vec}(\mathtt{U}^\top\Delta)\|_2^2}{d\delta} &(119)\\ &= \frac{d\|(\mathtt{I}\otimes\mathtt{U}^\top)\delta\|_2}{d\delta} &(120)\\ &= \frac{d(\delta^\top(\mathtt{I}\otimes\mathtt{U}\mathtt{U}^\top)\delta)}{d\delta} &(121)\\ &= (\mathtt{I}\otimes\mathtt{U}\mathtt{U}^\top)\delta. &(122)\end{aligned}$$

Hence the Hessian matrix has an added term

$$\mathtt{H}_p^* \leftarrow \mathtt{H}_p^* + \langle \alpha\mathtt{I}\otimes\mathtt{U}\mathtt{U}^\top\rangle_P. \tag{123}$$

When $\alpha = 1$, the term in $\langle$angle-bracketed red$\rangle_P$ is equivalent to the rank-filling term introduced by Okatani et al. in their Damped Wiberg algorithm [**?**]. Hence, above illustrates a connection between their implementation and the Grassmann manifold constraint.

4.5. Overall remark

In viewpoint of manifold optimization, the aforementioned analytic results imply that performing a Gauss Newton-variant un-regularized VarPro on low-rank matrix factorization problems is automatically incorporating the Grassmann manifold optimization framework. This was possible as both the gradient and the Gauss-Newton matrix (or its approximation discussed in §5.1) are already in the tangent space of $\mathtt{U}$. The algorithms that perform orthorgonalization using (orth) or q-factor can be interpreted as retracting back to the solution manifold.

Damped Wiberg [**?**] and LM_M_GN [**?**] have enjoyed greater success by unambiguously incorporating the manifold constraint $\mathtt{U}^\top\Delta\mathtt{U} = 0$.

5. Connections to existing algorithms

5.1. Ruhe & Wedin's algorithms

Ruhe and Wedin [**?**] proposed a set of general VarPro algorithms based on Gauss-Newton solver some of which make certain assumptions on the Gauss-Newton matrix.

Applying their algorithms to our un-regularized problem yields the following results for the set of matrices defined in [**?**]:

$$\begin{aligned}\mathtt{B} &= \frac{\partial[\tilde{\mathtt{U}}]^\top\mathbf{v}^*}{\partial\mathbf{u}} = \tilde{\mathtt{V}}^* && \textit{// noting bilinearity} &(124)\\ \mathtt{C} &= \tilde{\mathtt{U}}^{\dagger\top}\frac{\partial[\tilde{\mathtt{U}}]^\top\boldsymbol{\varepsilon}_1^*}{\partial\mathbf{u}} = \tilde{\mathtt{U}}^{\dagger\top}\mathtt{Z}^{*\top}\mathtt{K}_{mr}. && \textit{// noting the techniques from (9)} &(125)\\ \mathtt{P}_F &:= \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger &&&(126)\\ \mathtt{Q}_F &:= \mathtt{I} - \mathtt{P}_F = \mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger &&&(127)\\ \mathbf{g} &= \tilde{\mathbf{m}} &&&(128)\\ \mathtt{Q}_F\mathbf{g} &= (\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathbf{m}} && \textit{// from (18) and noting } \mu = 0 &(129)\end{aligned}$$

Transforming the $\mathbf{u}$-update of Algorithm 1 (RW1) to our notation gives

$$\begin{aligned}\delta\mathbf{u} &= (\mathtt{Q}_F\mathtt{B} + \mathtt{P}_F\mathtt{C})^\dagger\mathtt{Q}_F\mathbf{g} &&(130)\\ &= -\left((\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^* + \tilde{\mathtt{U}}^{\dagger\top}\mathtt{Z}^{*\top}\mathtt{K}_{mr}\right)^\dagger\boldsymbol{\varepsilon}_1^* &&(131)\\ &= -(\mathtt{J}_1^{*\top}\mathtt{J}_1^*)^{-1}\mathtt{J}_1^{*\top}\boldsymbol{\varepsilon}_1^* \ = -(\mathtt{J}_1^{*\top}\mathtt{J}_1^*)^{-1}\tilde{\mathtt{V}}^{*\top}\boldsymbol{\varepsilon}_1^*. \qquad \textit{// from (35)} &&(132)\end{aligned}$$

Above is identical to the Gauss-Newton solver on VarPro, and manifold projection is automatically introduced (see §4.2).

Algorithm 2 (RW2) approximates $\mathtt{J}_1^*$ by neglecting the term $\mathtt{P}_F\mathtt{C}$, which is ultimately equivalent to setting

$$d\mathbf{v}^*/d\mathbf{u} \approx -\tilde{\mathtt{U}}^\dagger\tilde{\mathtt{V}}^*. \tag{133}$$

This yields

$$\begin{align}
\delta\mathbf{u} &= (\mathtt{Q}_F\mathtt{B})^\dagger\mathtt{Q}_F\mathbf{g} \tag{134}\\
&= -\left((\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^*\right)^\dagger \boldsymbol{\varepsilon}_1^* \tag{135}\\
&= -\left(\tilde{\mathtt{V}}^{*\top}(\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^*\right)^{-1}\tilde{\mathtt{V}}^{*\top}(\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\boldsymbol{\varepsilon}_1^* \qquad \textit{// noting } (\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger) = (\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)^\top = (\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)^2 \tag{136}\\
&= -\left(\tilde{\mathtt{V}}^{*\top}(\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathtt{V}}^*\right)^{-1}\tilde{\mathtt{V}}^{*\top}\boldsymbol{\varepsilon}_1^* \qquad \textit{// noting } \boldsymbol{\varepsilon}_1^* = -(\mathtt{I}-\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger)\tilde{\mathbf{m}}. \tag{137}
\end{align}$$

This approximation is identical to that used by CSF [**?**].

Lastly, Algorithm 3 (RW3) ignores the variable change and solve for one variable whilst fixing the other. This is done by approximating $\mathtt{J}_1^* \approx \tilde{\mathtt{V}}^*$ leading to

$$\delta\mathbf{u} = -\left(\tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{V}}^*\right)^{-1}\tilde{\mathtt{V}}^{*\top}\boldsymbol{\varepsilon}_1^*.$$

The $\mathbf{u}$-update is then

$$\begin{align}
\mathbf{u} &\leftarrow \mathbf{u} + \delta\mathbf{u} \tag{138}\\
&= \mathbf{u} - \left(\tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{V}}^*\right)^{-1}\tilde{\mathtt{V}}^{*\top}\boldsymbol{\varepsilon}_1^* \tag{139}\\
&= \mathbf{u} - \left(\tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{V}}^*\right)^{-1}\tilde{\mathtt{V}}^{*\top}(\tilde{\mathtt{V}}^*\mathbf{u}-\tilde{\mathbf{m}}) \tag{140}\\
&= \tilde{\mathtt{V}}^\dagger\tilde{\mathbf{m}}. \tag{141}
\end{align}$$

Above is identical to the $\mathbf{u}$-update in un-regularized alternation (ALS). Since the expression for $\mathbf{v}^*(\mathbf{u})$ is also the same as that in ALS, this makes Algorithm 3 (RW3) the un-regularized ALS as mentioned in [**?**].

5.2. Chen's algorithms

Chen's algorithms in [**?**, **?**] apply truncated column-wise derivatives (see §3.6). These algorithms contain several redundant computations which can be removed by applying some of the matrix identities listed in [**?**].

The LM_S series apply no manifold constraint whilst LM_M series apply the hard constraint mentioned in §4.4.1. Inverses are computed using backslash in `MATLAB`.

5.3. CSF

CSF [**?**] applies full (non-truncated) column-wise derivatives (see §3.6). No manifold constraint is used, and inverses are computed using backslash in `MATLAB`.

5.4. Damped Wiberg

The framework of the Damped Wiberg algorithm is included in [**?**]. Derivations illustrated by Okatani et al. [**?**, **?**] are essentially the same as in our submission [**?**, **?**]. The small differences are as follows:

1. The definitions of $\mathbf{u}$ and $\mathbf{v}$ are swapped. i.e. Damped Wiberg assumes thin measurement matrix.
2. Instead of $\mathbf{u} := \mathrm{vec}(\mathtt{U})$, Damped Wiberg uses $\mathbf{u} := \mathrm{vec}(\mathtt{U}^\top)$ introducing additional permutation in the update of $\mathbf{u}$. i.e.

$$\frac{1}{2}\mathtt{H}_{\mathrm{DW}}^* = \mathtt{K}_{mr}\tilde{\mathtt{V}}^{*\top}(\mathtt{I}-[\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger]_{RW2})\tilde{\mathtt{V}}^*\mathtt{K}_{mr}^\top + \langle\mathtt{K}_{mr}(\mathtt{I}\otimes\mathtt{U}\mathtt{U}^\top)\mathtt{K}_{mr}^\top\rangle_P + \langle\lambda\mathtt{I}\rangle. \tag{142}$$

3. The Gauss-Newton matrix used by Damped Wiberg is an approximation of the original Gauss-Newton matrix, and this approximation is equivalent to that used by Ruhe and Wedin's Algorithm 2.

The Damped Wiberg algorithm uses QR-decomposition for several inversion processes, which seems to improve both success rates and overall algorithm speed as the decomposed set of Q-blocks and R-blocks can be used multiple times.

5.5. RTRMC

RTRMC [**?**] uses regularized VarPro with full Hessian and Grassmann manifold projection. Boumal et al. has overcome the difficulty of analytically deriving full Hessian (see §2.4) by incorporating an indirect subproblem solver, which only requires directional derivatives. Instead of applying the regularization $\mu(\|\mathtt{U}\|_F^2 + \|\mathtt{V}\|_F^2)$, RTRMC uses $\mu\|\mathtt{U}\mathtt{V}^\top\|_{\not\Omega}^2$ [**?**] where $\not\Omega$ is a set of missing entries.

We will show that their Hessian operator for the un-regularized case matches our expression for Hessian in §3.5. From [**?**], the unregularized projected gradient is

$$\frac{1}{2}\mathbf{g}_p^* := \mathrm{vec}((\mathtt{W} \odot \mathtt{R}^*)\mathtt{V}^*). \tag{143}$$

Hence, $\mathrm{grad} f_1^*(\mathtt{U}) := \nabla_{\mathtt{U}} f_1^*(\mathtt{U})$ is simply the unvectorized quantity $2(\mathtt{W} \odot \mathtt{R}^*)\mathtt{V}^*$.

Noting $\mathtt{R}^*(\mathtt{U}) = \mathtt{U}\mathtt{V}^{*\top}(\mathtt{U}) - \mathtt{M}$, the Hessian operator, Hess $f_1^*(\mathtt{U})[\eta]$, can be obtained by taking the directional derivative of $\mathrm{grad} f_1^*(\mathtt{U})$ towards direction η:

$$\begin{aligned}
\frac{1}{2}\text{Hess } f_1^*(\mathtt{U})[\eta] &:= \frac{1}{2}D\text{grad } f_1^*(\mathtt{U})[\eta] && (144)\\
&= (\mathtt{W} \odot \mathtt{W} \odot \eta\mathtt{V}^{*\top})\mathtt{V}^* + (\mathtt{W} \odot \mathtt{W} \odot \mathtt{U}\; D\mathtt{V}^{*\top}(\mathtt{U})[\eta])\mathtt{V}^* + (\mathtt{W} \odot \mathtt{R}^*)\; D\mathtt{V}^*(\mathtt{U})[\eta] && (145)
\end{aligned}$$

Vectorizing the first term produces

$$\begin{aligned}
\mathrm{vec}((\mathtt{W} \odot \mathtt{W} \odot \eta\mathtt{V}^{*\top})\mathtt{V}^*) &= (\mathtt{V}^{*\top} \otimes \mathtt{I})\,\mathrm{vec}(\mathtt{W} \odot \mathtt{W} \odot \eta\mathtt{V}^{*\top}) && (146)\\
&= (\mathtt{V}^{*\top} \otimes \mathtt{I})\tilde{\mathtt{W}}^\top\tilde{\mathtt{W}}\,\mathrm{vec}(\eta\mathtt{V}^{*\top}) && (147)\\
&= (\mathtt{V}^{*\top} \otimes \mathtt{I})\tilde{\mathtt{W}}^\top\tilde{\mathtt{W}}(\mathtt{V}^{*\top} \otimes \mathtt{I})\,\mathrm{vec}(\eta) && (148)\\
&= \tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{V}}^*\,\mathrm{vec}(\eta), && (149)
\end{aligned}$$

which is the RW3 (ALS) term. Vectorizing the second term generates

$$\begin{aligned}
\mathrm{vec}(\mathtt{W} \odot \mathtt{W} \odot \mathtt{U}\; D\mathtt{V}^{*\top}(\mathtt{U})[\eta])\mathtt{V}^* &= (\mathtt{V}^{*\top} \otimes \mathtt{I})\,\mathrm{vec}(\mathtt{W} \odot \mathtt{W} \odot \mathtt{U}\; D\mathtt{V}^{*\top}(\mathtt{U})[\eta]) && (150)\\
&= (\mathtt{V}^{*\top} \otimes \mathtt{I})\tilde{\mathtt{W}}^\top\tilde{\mathtt{W}}\,\mathrm{vec}(\mathtt{U}\; D\mathtt{V}^{*\top}(\mathtt{U})[\eta])\\
&= (\mathtt{V}^{*\top} \otimes \mathtt{I})\tilde{\mathtt{W}}^\top\tilde{\mathtt{W}}(\mathtt{I} \otimes \mathtt{U})\,\mathrm{vec}(\; D\mathtt{V}^{*\top}(\mathtt{U})[\eta]) && (151)\\
&= \tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{U}}\,\mathrm{vec}(\; D\mathtt{V}^{*\top}(\mathtt{U})[\eta]), && (152)
\end{aligned}$$

and vectorizing the last term yields

$$\begin{aligned}
\mathrm{vec}((\mathtt{W} \odot \mathtt{R}^*)\; D\mathtt{V}^*(\mathtt{U})[\eta]) &= (\mathtt{I} \otimes (\mathtt{W} \odot \mathtt{R}^*))\,\mathrm{vec}(D\mathtt{V}^*(\mathtt{U})[\eta]) && (153)\\
&= (\mathtt{I} \otimes (\mathtt{W} \odot \mathtt{R}^*))\mathtt{K}_{nr}^\top\,\mathrm{vec}(D\mathtt{V}^{*\top}(\mathtt{U})[\eta]) && \textit{// from [?]} \quad (154)\\
&= \mathtt{K}_{mr}^\top((\mathtt{W} \odot \mathtt{R}^*) \otimes \mathtt{I})\mathtt{K}_{nr}\mathtt{K}_{nr}^\top\,\mathrm{vec}(D\mathtt{V}^{*\top}(\mathtt{U})[\eta]) && \textit{// from [?]} \quad (155)\\
&= \mathtt{K}_{mr}^\top\mathtt{Z}^*\,\mathrm{vec}(D\mathtt{V}^{*\top}(\mathtt{U})[\eta]). && (156)
\end{aligned}$$

We can obtain $\mathrm{vec}(D\mathtt{V}^{*\top}(\mathtt{U})[\eta])$ by reformulating the expression for $D\mathtt{V}^{*\top}(\mathtt{U})[\eta]$ given by (3.62) in [**?**]. Bearing in mind that the regularization parameter is set to zero here and that the second decomposition matrix $\mathtt{V}$ is transposed in [**?**], we yield

$$\begin{aligned}
\mathrm{vec}(D\mathtt{V}^{*\top}(\mathtt{U})[\eta]) &= -(\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\,\mathrm{vec}(\eta^\top(\mathtt{W} \odot \mathtt{R}^*) + \mathtt{U}^\top(\mathtt{W} \odot \mathtt{W} \odot (\eta\mathtt{V}^{*\top})) && (157)\\
&= -(\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}((\mathtt{W} \odot \mathtt{R}^*)^\top \otimes \mathtt{I})\,\mathrm{vec}(\eta^\top)\\
&\quad - (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}(\mathtt{I} \otimes \mathtt{U}^\top)(\mathtt{W} \odot \mathtt{W} \odot \eta\mathtt{V}^{*\top})\\
&= -(\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr}\,\mathrm{vec}(\eta) - (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\tilde{\mathtt{U}}^\top\tilde{\mathtt{W}}(\mathtt{V}^* \otimes \mathtt{I})\,\mathrm{vec}(\eta)\\
&= -((\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr} + (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\tilde{\mathtt{U}}^\top\tilde{\mathtt{V}}^*)\,\mathrm{vec}(\eta). && (158)
\end{aligned}$$

Hence, (152) is reduced to

$$-\tilde{\mathtt{V}}^{*\top}\tilde{\mathtt{U}}((\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr} + (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\tilde{\mathtt{U}}^\top\tilde{\mathtt{V}}^*)\,\mathrm{vec}(\eta), \tag{159}$$

and (156) is reduced to

$$-\mathtt{K}_{mr}^\top\mathtt{Z}^*((\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr} + (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1}\tilde{\mathtt{U}}^\top\tilde{\mathtt{V}}^*)\,\mathrm{vec}(\eta). \tag{160}$$

It is then easy to check that adding all three terms (149), (159) and (160) yield $\frac{1}{2}\mathtt{H}_1^*\,\mathrm{vec}(\eta)$, where $\mathtt{H}_1^*$ is the full Hessian for unregularized VarPro.

References