Secrets of Matrix Factorization: Supplementary Document

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

Andrew Fitzgibbon
Microsoft, Cambridge, UK
awf@microsoft.com

Abstract

This document consists of details to which the original submission article has made references. These include a list of symbols, some matrix identities and calculus rules, a list of VarPro derivatives, symmetry analysis of the un-regularized VarPro Gauss-Newton matrix and MTSS computation details.

Symbol	Meaning
$\mathbb{R}^{p\times q}$	The space of real matrices of size $p \times q$
$\mathbb{S}^p$	The space of symmetric real matrices of size $p \times p$
$\mathtt{M}$	The measurement matrix $\in \mathbb{R}^{m\times n}$ (may contain noise and missing data)
$\mathtt{W}$	The weight matrix $\in \mathbb{R}^{m\times n}$
$\mathtt{U}$	The first decomposition matrix $\in \mathbb{R}^{m\times r}$
$\mathtt{V}$	The second decomposition matrix $\in \mathbb{R}^{n\times r}$
$\mathbf{u}$	$\operatorname{vec}(\mathtt{U}) \in \mathbb{R}^{mr}$
$\mathbf{v}$	$\operatorname{vec}(\mathtt{V}^\top) \in \mathbb{R}^{nr}$
m	The number of rows in $\mathtt{M}$
n	The number of columns in $\mathtt{M}$
p	The number of visible entries in $\mathtt{M}$
p_j	The number of visible entries in the j-th column of $\mathtt{M}$
f	The objective $\in \mathbb{R}$
$\boldsymbol{\varepsilon}$	The vector objective $\in \mathbb{R}^p$
$\mathtt{J}$	Jacobian
$\mathtt{H}$	Hessian
$\mathtt{V}^*$	$\arg\min_{\mathtt{V}} f(\mathtt{U},\mathtt{V}) \in \mathbb{R}^{n\times r}$
Π	Projection matrix $\in \mathbb{R}^{p\times mn}$ selecting nonzero entries of $\operatorname{vec}(\mathtt{W})$
$\tilde{\mathtt{W}}$	$\Pi \operatorname{diag}(\operatorname{vec}(\mathtt{W})) \in \mathbb{R}^{p\times mn}$
$\tilde{\mathbf{m}}$	$\tilde{\mathtt{W}} \operatorname{vec}(\mathtt{M}) \in \mathbb{R}^p$
$\tilde{\mathtt{U}}$	$\tilde{\mathtt{W}}(\mathtt{I}_n \otimes \mathtt{U}) \in \mathbb{R}^{p\times nr}$
$\tilde{\mathtt{V}}$	$\tilde{\mathtt{W}}(\mathtt{V} \otimes \mathtt{I}_m) \in \mathbb{R}^{p\times mr}$
$\mathtt{K}_{mr}$	The permutation matrix satisfying $\operatorname{vec}(\mathtt{U}^\top) = \mathtt{K}_{mr} \operatorname{vec}(\mathtt{U}) \in \mathbb{R}^{mr\times mr}$
$\mathtt{K}_{nr}$	The permutation matrix satisfying $\operatorname{vec}(\mathtt{V}^\top) = \mathtt{K}_{nr} \operatorname{vec}(\mathtt{V}) \in \mathbb{R}^{nr\times nr}$
$\mathtt{R}$	$\mathtt{W} \odot (\mathtt{U}\mathtt{V}^\top - \mathtt{M}) \in \mathbb{R}^{m\times n}$
$\mathtt{Z}$	$(\mathtt{W} \odot \mathtt{R}) \otimes \mathtt{I}_r \in \mathbb{R}^{mr\times nr}$

Table 1: A list of symbols used in [1].

1. Symbols

A list of symbols used in [1] can be found in Table 1.

2. Matrix identities and calculus rules

Some of the key matrix identities and calculus rules in [6, 5] are illustrated and extended below for convenience.

2.1. Vec and Kronecker product identities

Given that each small letter in bold is an arbitrary column vector of appropriate size, we have

$$\begin{aligned}
\operatorname{vec}(\mathtt{A} \odot \mathtt{B}) &= \operatorname{diag}(\operatorname{vec}\mathtt{A})\operatorname{vec}(\mathtt{B}), && (1)\\
\operatorname{vec}(\mathtt{AXB}) &= (\mathtt{B}^\top \otimes \mathtt{A})\operatorname{vec}(\mathtt{X}), && (2)\\
\operatorname{vec}(\mathtt{AB}) &= (\mathtt{B}^\top \otimes \mathtt{I})\operatorname{vec}(\mathtt{A}) = (\mathtt{I}^\top \otimes \mathtt{A})\operatorname{vec}(\mathtt{B}), && (3)\\
(\mathtt{A} \otimes \mathtt{B})(\mathtt{C} \otimes \mathtt{D}) &= \mathtt{AC} \otimes \mathtt{BD}, && (4)\\
(\mathtt{A} \otimes \mathbf{b})\mathtt{C} &= \mathtt{AC} \otimes \mathbf{b}, && (5)\\
\operatorname{vec}(\mathtt{A}^\top) &= \mathtt{K}_{\mathtt{A}}\operatorname{vec}(\mathtt{A}), && (6)\\
(\mathtt{A} \otimes \mathtt{B}) &= \mathtt{K}_1^\top(\mathtt{B} \otimes \mathtt{A})\mathtt{K}_2, \text{ and} && (7)\\
(\mathtt{A} \otimes \mathbf{b}) &= \mathtt{K}_1^\top(\mathbf{b} \otimes \mathtt{A}). && (8)
\end{aligned}$$

2.2. Differentiation of μ-pseudo inverse

Let us first define μ-pseudo inverse $\mathtt{X}^{-\mu} := (\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\mathtt{X}^\top$. The derivative of μ-pseudo inverse is

$$\begin{aligned}
\partial[\mathtt{X}^{-\mu}] &= \partial[(\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}]\mathtt{X}^\top + (\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\partial[\mathtt{X}]^\top && (9)\\
&= -(\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\partial[\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I}](\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\mathtt{X}^\top + (\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\partial[\mathtt{X}]^\top && (10)\\
&= -(\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}(\partial[\mathtt{X}^\top]\mathtt{X} + \mathtt{X}^\top\partial[\mathtt{X}])(\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\mathtt{X}^\top + (\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\partial[\mathtt{X}]^\top && (11)\\
&= -(\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\partial[\mathtt{X}^\top]\mathtt{X}\mathtt{X}^{-\mu} - (\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\mathtt{X}^\top\partial[\mathtt{X}]\mathtt{X}^{-\mu} + (\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\partial[\mathtt{X}]^\top && (12)\\
&= -\mathtt{X}^{-\mu}\partial[\mathtt{X}]\mathtt{X}^{-\mu} + (\mathtt{X}^\top\mathtt{X} + \mu\mathtt{I})^{-1}\partial[\mathtt{X}]^\top(\mathtt{I} - \mathtt{X}\mathtt{X}^{-\mu}). && (13)
\end{aligned}$$

The derivative of pseudo-inverse $\mathtt{X}^\dagger := (\mathtt{X}^\top\mathtt{X})^{-1}\mathtt{X}^\top$ is then simply the above for $\mu = 0$.

3. Derivatives for variable projection (VarPro)

A full list of derivatives for un-regularized VarPro can be found in Table 2. Derivations are illustrated in [2]. The regularized setting is also discussed in [2].

3.1. Differentiating $\mathbf{v}^*(\mathbf{u})$

From [1], we have

$$\mathbf{v}^*(\mathbf{u}) := \arg\min_{\mathbf{v}} f(\mathbf{u}, \mathbf{v}) = \tilde{\mathtt{U}}^{-\mu}\tilde{\mathbf{m}}. \tag{14}$$

Substituting (14) to the original objective yields

$$\boldsymbol{\varepsilon}_1^*(\mathbf{u}, \mathbf{v}^*(\mathbf{u})) = \tilde{\mathtt{U}}\mathbf{v}^* - \tilde{\mathbf{m}} = -(\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^{-\mu})\tilde{\mathbf{m}}. \tag{15}$$

Taking $\partial[\mathbf{v}^*]$ and applying (13) gives us

$$\begin{aligned}
\partial[\mathbf{v}^*] &= -\tilde{\mathtt{U}}^{-\mu}\partial[\tilde{\mathtt{U}}]\tilde{\mathtt{U}}^{-\mu}\tilde{\mathbf{m}} + (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}} + \mu\mathtt{I})^{-1}\partial[\tilde{\mathtt{U}}]^\top(\mathtt{I} - \tilde{\mathtt{U}}\tilde{\mathtt{U}}^{-\mu})\tilde{\mathbf{m}} && (16)\\
&= -\tilde{\mathtt{U}}^{-\mu}\partial[\tilde{\mathtt{U}}]\mathbf{v}^* - (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}} + \mu\mathtt{I})^{-1}\partial[\tilde{\mathtt{U}}]^\top\boldsymbol{\varepsilon}_1^* && /\!/ \textit{ noting } \mathbf{v}^*(\mathbf{u}) = \tilde{\mathtt{U}}^{-\mu}\tilde{\mathbf{m}} \textit{ and (15)} && (17)\\
&= -\tilde{\mathtt{U}}^{-\mu}\tilde{\mathtt{V}}^*\partial\mathbf{u} - (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}} + \mu\mathtt{I})^{-1}\mathtt{Z}^{*\top}\mathtt{K}_{mr}\partial\mathbf{u}, && /\!/ \textit{ noting bilinearity and the results in [2]} && (18)
\end{aligned}$$

and hence

$$\frac{d\mathbf{v}^*}{d\mathbf{u}} = -(\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}} + \mu\mathtt{I})^{-1}(\tilde{\mathtt{U}}^\top\tilde{\mathtt{V}}^* + \mathtt{Z}^{*\top}\mathtt{K}_{mr}). \tag{19}$$

Quantity	Definition	ALS (RW3) + $[\text{Approx. Gauss-Newton (RW2)}]_{RW2}$ + $[\text{Full Gauss-Newton (RW1)}]_{RW1}$ + $[\text{Full Newton}]_{FN}$ w/o $\langle\text{Damping}\rangle$ w/o $\langle\text{Projection constraint}\rangle_P$
$\mathbf{v}^*(\mathbf{u}) \in \mathbb{R}^{nr}$	$\mathbf{v}^* := \arg\min_{\mathbf{v}} f(\mathbf{u}, \mathbf{v})$	$\mathbf{v}^* = \tilde{\mathtt{U}}^\dagger \tilde{\mathbf{m}}$ ($\tilde{\mathbf{m}} \in \mathbb{R}^p$ consists of non-zero elements of $\mathbf{m}$.)
$\frac{d\mathbf{v}^*}{d\mathbf{u}} \in \mathbb{R}^{nr \times mr}$	$\frac{d\mathbf{v}^*}{d\mathbf{u}} := \frac{d\,\mathrm{vec}(\mathtt{V}^{*\top})}{d\,\mathrm{vec}(\mathtt{U})}$	$\frac{d\mathbf{v}^*}{d\mathbf{u}} = -[\tilde{\mathtt{U}}^\dagger \tilde{\mathtt{V}}^*]_{RW2} - [(\tilde{\mathtt{U}}^\top \tilde{\mathtt{U}})^{-1} \mathtt{Z}^{*\top} \mathtt{K}_{mr}]_{RW1}$ ($\mathtt{Z}^* := (\mathtt{W} \odot \mathtt{R}^*) \otimes \mathtt{I}_r$)
Cost vector $\boldsymbol{\varepsilon}_1^* \in \mathbb{R}^p$	$\boldsymbol{\varepsilon}_1^*$	$\boldsymbol{\varepsilon}_1^* := \tilde{\mathtt{U}}\mathbf{v}^* - \mathbf{m}$
Jacobian $\mathtt{J}_1^* \in \mathbb{R}^{p \times mr}$	$\mathtt{J}_1^* := \frac{d\boldsymbol{\varepsilon}_1^*}{d\mathbf{u}}$	$\mathtt{J}_1^* = (\mathtt{I}_p - [\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger]_{RW2})\tilde{\mathtt{V}}^* - [\tilde{\mathtt{U}}^{\dagger\top} \mathtt{Z}^{*\top} \mathtt{K}_{mr}]_{RW1}$
Cost function $f_1^* \in \mathbb{R}$	$f_1^* := \|\boldsymbol{\varepsilon}_1^*\|_2^2$	$f_1^* = \|\mathtt{W} \odot (\mathtt{U}\mathtt{V}^{*\top} - \mathtt{M})\|_F$
Ordinary gradient $\mathbf{g}_1^* \in \mathbb{R}^{mr}$	$\mathbf{g}_1^* := \frac{df_1^*}{d\mathbf{u}}$	$\frac{1}{2}\mathbf{g}_1^* = \tilde{\mathtt{V}}^{*\top} \boldsymbol{\varepsilon}_1^* = \mathrm{vec}((\mathtt{W} \odot \mathtt{R}^*)\mathtt{V}^*)$
Projected gradient $\mathbf{g}_p^* \in \mathbb{R}^{mr}$	$\mathbf{g}_p^* := \mathtt{P}_p \mathbf{g}_1^*$ ($\mathtt{P}_p := \mathtt{I}_{mr} - \mathtt{I}_r \otimes \mathtt{U}\mathtt{U}^\top$)	$\frac{1}{2}\mathbf{g}_p^* = \frac{1}{2}\mathbf{g}_1^*$
Reduced gradient $\mathbf{g}_r^* \in \mathbb{R}^{mr-r^2}$	$\mathbf{g}_r^* := \mathtt{P}_r \mathbf{g}_1^*$ ($\mathtt{P}_r := \mathtt{I}_r \otimes \mathtt{U}_\perp^\top$)	$\frac{1}{2}\mathbf{g}_r^* = \frac{1}{2}\mathtt{P}_r \mathbf{g}_1^*$
Ordinary Hessian $\mathtt{H}_{1GN}^*, \mathtt{H}_1^* \in \mathbb{S}^{mr}$	$\frac{1}{2}\mathtt{H}_{1GN}^* := \mathtt{J}_1^{*\top}\mathtt{J}_1^* + \langle\lambda\mathtt{I}\rangle$ $\frac{1}{2}\mathtt{H}_1^* := \frac{1}{2}\frac{d\mathbf{g}_1}{d\mathbf{u}} + \langle\lambda\mathtt{I}\rangle$	$\frac{1}{2}\mathtt{H}_{1GN}^* = \tilde{\mathtt{V}}^{*\top}(\mathtt{I}_p - [\tilde{\mathtt{U}}\tilde{\mathtt{U}}^\dagger]_{RW2})\tilde{\mathtt{V}}^* + [\mathtt{K}_{mr}^\top \mathtt{Z}^* (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1} \mathtt{Z}^{*\top} \mathtt{K}_{mr}]_{RW1} + \langle\lambda\mathtt{I}_{mr}\rangle$ $\frac{1}{2}\mathtt{H}_1^* = \frac{1}{2}\mathtt{H}_{1GN}^* - [2\mathtt{K}_{mr}^\top \mathtt{Z}^* (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1} \mathtt{Z}^{*\top} \mathtt{K}_{mr}]_{FN} - [\mathtt{K}_{mr}^\top \mathtt{Z}^* \tilde{\mathtt{U}}^\dagger \tilde{\mathtt{V}}^* + \tilde{\mathtt{V}}^{*\top} \tilde{\mathtt{U}}^{\dagger\top} \mathtt{Z}^{*\top} \mathtt{K}_{mr}]_{FN} + \langle\lambda\mathtt{I}_{mr}\rangle$
Projected Hessian $\mathtt{H}_{pGN}^*, \mathtt{H}_p^* \in \mathbb{S}^{mr}$	$\frac{1}{2}\mathtt{H}_{pGN}^* := \mathtt{P}_p \mathtt{J}^{*\top} \mathtt{J}^* \mathtt{P}_p + \langle\lambda\mathtt{I}\rangle$ $\frac{1}{2}\mathtt{H}_p^* := \frac{1}{2}\mathtt{P}_p \frac{d\mathbf{g}_p}{d\mathbf{u}} \mathtt{P}_p + \langle\lambda\mathtt{I}\rangle$ ($\mathtt{P}_p := \mathtt{I}_r \otimes (\mathtt{I}_m - \mathtt{U}\mathtt{U}^\top)$)	$\frac{1}{2}\mathtt{H}_{pGN}^* = \frac{1}{2}\mathtt{H}_{1GN}^*$ $\frac{1}{2}\mathtt{H}_p^* = \frac{1}{2}\mathtt{H}_{pGN}^* - [2\mathtt{K}_{mr}^\top \mathtt{Z}^* (\tilde{\mathtt{U}}^\top\tilde{\mathtt{U}})^{-1} \mathtt{Z}^{*\top} \mathtt{K}_{mr}]_{FN} - [\mathtt{K}_{mr}^\top \mathtt{Z}^* \tilde{\mathtt{U}}^\dagger \tilde{\mathtt{V}}^* \mathtt{P}_p + \mathtt{P}_p \tilde{\mathtt{V}}^{*\top} \tilde{\mathtt{U}}^{\dagger\top} \mathtt{Z}^{*\top} \mathtt{K}_{mr}]_{FN} + \langle\alpha\mathtt{I}_r \otimes \mathtt{U}\mathtt{U}^\top\rangle_P + \langle\lambda\mathtt{I}_{mr}\rangle$
Reduced Hessian $\mathtt{H}_r^* \in \mathbb{S}^{(mr-r^2)}$	$\frac{1}{2}\mathtt{H}_r^* := \frac{1}{2}\mathtt{P}_r \mathtt{H}_p^* \mathtt{P}_r{}^\top$ ($\mathtt{P}_r := \mathtt{I}_r \otimes \mathtt{U}_\perp^\top$)	$\frac{1}{2}\mathtt{H}_r^* = \frac{1}{2}\mathtt{P}_r \mathtt{H}_p^* \mathtt{P}_r{}^\top = \frac{1}{2}\mathtt{P}_r \mathtt{H}_1^* \mathtt{P}_r{}^\top = \frac{1}{2}\mathtt{P}_r \frac{d\mathbf{g}_1^*}{d\mathbf{u}} \mathtt{P}_r{}^\top + \langle\lambda\mathtt{I}_{mr}\rangle$ $\begin{cases} \mathtt{K}_{mr}\,\mathrm{vec}(\mathtt{U}) = \mathrm{vec}(\mathtt{U}^\top) \\ \mathtt{R}^* := \mathtt{W} \odot (\mathtt{U}\mathtt{V}^{*\top} - \mathtt{M}) \\ \mathtt{Z}^* := (\mathtt{W} \odot \mathtt{R}^*) \otimes \mathtt{I}_r \end{cases}$
Ordinary update	Ordinary update for $\mathtt{U}$	$\mathtt{U} \leftarrow \mathtt{U} - \mathrm{unvec}(\mathtt{H}_1^{*-1}\mathbf{g}_1^*)$
Projected update	Grassmann manifold retraction	$\mathtt{U} \leftarrow qf(\mathtt{U} - \mathrm{unvec}(\mathtt{H}_p^{*-1}\mathbf{g}_p^*))$
Reduced update	Grassmann manifold retraction	$\mathtt{U} \leftarrow qf(\mathtt{U} - \mathtt{P}_r{}^\top \mathrm{unvec}(\mathtt{H}_r^{*-1}\mathbf{g}_r^*))$

Table 2: List of derivatives and extensions for un-regularized variable projection on low-rank matrix factorization with missing data. This includes notions to Ruhe and Wedin's algorithms.

4. Additional symmetry of the Gauss-Newton matrix for un-regularized VarPro

To observe the additional symmetry, we first need to know how the un-regularized VarPro problem can be reformulated in terms of column-wise derivatives. Below is a summary of these derivatives found in [2].

4.1. Summary of column-wise derivatives for VarPro

The unregularized objective $f_1^*(\mathbf{u}, \mathbf{v}^*(\mathbf{u}))$ can be defined as the sum of squared norm of each column in $\mathtt{R}^*(\mathbf{u}, \mathbf{v}^*(\mathbf{u}))$. i.e.

$$f_1^*(\mathbf{u}, \mathbf{v}^*(\mathbf{u})) := \sum_{j=1}^{n} \|\tilde{\mathtt{W}}_j(\mathtt{U}\mathbf{v}_j^* - \mathbf{m}_j)\|_2^2, \tag{20}$$

where $\mathbf{v}_j^*$ is the j-th row of $\mathtt{V}^*(\mathtt{U})$, $\mathbf{m}_j$ is the j-th column of $\mathtt{M}$ and $\tilde{\mathtt{W}}_j$ is the non-zero rows of $\operatorname{diag}\operatorname{vec}(\mathbf{w}_j)$, where $\mathbf{w}_j \in \mathbb{R}^m$ is the j-th column of $\mathtt{W}$. Hence, $\tilde{\mathtt{W}}_j$ is a $p_j \times m$ weight-projection matrix, where p_j is the number of non-missing entries in column j.

Now define the followings:

$$\tilde{\mathtt{U}}_j := \tilde{\mathtt{W}}_j\mathtt{U} \tag{21}$$

$$\tilde{\mathbf{m}}_j := \tilde{\mathtt{W}}_j\mathbf{m}_j, \text{ and} \tag{22}$$

$$\boldsymbol{\varepsilon}_{1j}^* = \tilde{\mathtt{U}}_j\mathbf{v}_j^* - \tilde{\mathbf{m}}_j. \tag{23}$$

Hence,

$$\mathbf{v}_j^* := \arg\min_{\mathbf{v}_j} \sum_{j=1}^{n} \|\tilde{\mathtt{U}}_j\mathbf{v}_j^* - \tilde{\mathbf{m}}_j\|_2^2 \tag{24}$$

$$= \arg\min_{\mathbf{v}_j} \|\tilde{\mathtt{U}}_j\mathbf{v}_j^* - \tilde{\mathbf{m}}_j\|_2^2 = \tilde{\mathtt{U}}_j^\dagger\tilde{\mathbf{m}}_j, \tag{25}$$

which shows that each row of $\mathtt{V}^*$ is independent of other rows.

4.1.1 Some definitions

We define

$$\tilde{\mathtt{V}}_j^* := \tilde{\mathtt{W}}_j(\mathbf{v}_j^{*\top} \otimes \mathtt{I}_m) \in \mathbb{R}^{p_j \times mr}, \text{ and} \tag{26}$$

$$\mathtt{Z}_j^* := (\mathbf{w}_j \odot \mathbf{r}_j^*) \otimes \mathtt{I}_r \in \mathbb{R}^{mr \times r}. \tag{27}$$

4.1.2 Column-wise Jacobian $\mathtt{J}_{1j}^*$

The Jacobian matrix is also independent for each column. From [2], the Jacobian matrix for column j, $\mathtt{J}_{1j}^*$, is

$$\mathtt{J}_{1j}^* = (\mathtt{I} - \tilde{\mathtt{U}}_j\tilde{\mathtt{U}}_j^\dagger)\tilde{\mathtt{V}}_j^* - \tilde{\mathtt{U}}_j^{\dagger\top}\mathtt{Z}_j^{*\top}\mathtt{K}_{mr} \tag{28}$$

$$= (\mathtt{I} - \tilde{\mathtt{U}}_j\tilde{\mathtt{U}}_j^\dagger)\tilde{\mathtt{V}}_j^* - \tilde{\mathtt{U}}_j^\top\left((\tilde{\mathtt{U}}_j^\top\tilde{\mathtt{U}}_j)^{-1} \otimes (\mathbf{w}_j \odot \mathbf{r}_j^*)^\top\right). \tag{29}$$

4.1.3 The column-wise form of Hessian $\mathtt{H}_1^*$

From [2], the column-wise form of $\mathtt{H}_1^*$ is

$$\frac{1}{2}\mathtt{H}_1^* = \sum_{j=1}^{n}\Big[\tilde{\mathtt{V}}_j^{*\top}(\mathtt{I} - [\tilde{\mathtt{U}}_j\tilde{\mathtt{U}}_j^\dagger]_{RW2})\tilde{\mathtt{V}}_j^* + [-1]_{FN} \times [\mathtt{K}_{mr}^\top\mathtt{Z}_j^*(\tilde{\mathtt{U}}_j^\top\tilde{\mathtt{U}}_j)^{-1}\mathtt{Z}_j^{*\top}\mathtt{K}_{mr}]_{RW1}$$

$$- [\tilde{\mathtt{V}}_j^{*\top}\tilde{\mathtt{U}}_j^{\dagger\top}\mathtt{Z}_j^{*\top}\mathtt{K}_{mr} + \mathtt{K}_{mr}^\top\mathtt{Z}_j^*\tilde{\mathtt{U}}_j^\dagger\tilde{\mathtt{V}}_j^*]_{FN}\Big] + \langle\lambda\mathtt{I}\rangle. \tag{30}$$

All $\mathtt{Z}_j^{*\top}\mathtt{K}_{mr}$ can be replaced with $\mathtt{I} \otimes (\mathbf{w}_j \odot \mathbf{r}_j^*)^\top$ using (8).

4.2. Symmetry analysis

As derived in [2], the Gauss-Newton matrix for un-regularized VarPro is identical to its projection to the tangent space of $\mathtt{U}$. The column-wise representation of this is

$$\frac{1}{2}\mathtt{H}^*_{GN1} = \sum_{j=1}^{n}\left[\tilde{\mathtt{V}}_j^{*\top}(\mathtt{I} - [\tilde{\mathtt{U}}_j\tilde{\mathtt{U}}_j^{\dagger}]_{RW2})\tilde{\mathtt{V}}_j^* + [\mathtt{K}_{mr}^{\top}\mathtt{Z}_j^*(\tilde{\mathtt{U}}_j^{\top}\tilde{\mathtt{U}}_j)^{-1}\mathtt{Z}_j^{*\top}\mathtt{K}_{mr}]_{RW1}\right] + \langle\lambda\mathtt{I}\rangle. \tag{31}$$

Simplifying the first term yields

$$\begin{aligned}\tilde{\mathtt{V}}_j^{*\top}\tilde{\mathtt{V}}_j^* &= (\mathbf{v}_j^* \otimes \mathtt{I})\tilde{\mathtt{W}}_j^{\top}\tilde{\mathtt{W}}_j(\mathbf{v}_j^{*\top} \otimes \mathtt{I}) && (32)\\ &= (\mathbf{v}_j^* \otimes \tilde{\mathtt{W}}_j^{\top})(\mathbf{v}_j^{*\top} \otimes \tilde{\mathtt{W}}_j) = \mathbf{v}_j^*\mathbf{v}_j^{*\top} \otimes \tilde{\mathtt{W}}_j^{\top}\tilde{\mathtt{W}}_j, && /\!/ \textit{ noting (5)} \quad (33)\end{aligned}$$

which is the Kronecker product of two symmetric matrices.

Now defining $\tilde{\mathtt{U}}_{j\mathtt{Q}} := qf(\tilde{\mathtt{U}}_j)$ gives $\tilde{\mathtt{U}}_j\tilde{\mathtt{U}}_j^{\dagger} = \tilde{\mathtt{U}}_{j\mathtt{Q}}\tilde{\mathtt{U}}_{j\mathtt{Q}}^{\top}$. Hence, simplifying the second term yields

$$\begin{aligned}-\tilde{\mathtt{V}}_j^{*\top}(\tilde{\mathtt{U}}_j\tilde{\mathtt{U}}_j^{\dagger\top})\tilde{\mathtt{V}}_j^* &= -(\mathbf{v}_j^* \otimes \tilde{\mathtt{W}}_j^{\top})(\tilde{\mathtt{U}}_{j\mathtt{Q}}\tilde{\mathtt{U}}_{j\mathtt{Q}}^{\top})(\mathbf{v}_j^{*\top} \otimes \tilde{\mathtt{W}}_j) && (34)\\ &= -\mathbf{v}_j^*\mathbf{v}_j^{*\top} \otimes \tilde{\mathtt{W}}_j^{\top}\tilde{\mathtt{U}}_{j\mathtt{Q}}\tilde{\mathtt{U}}_{j\mathtt{Q}}^{\top}\tilde{\mathtt{W}}_j. && /\!/ \textit{ noting (5)} \quad (35)\end{aligned}$$

Given that $\tilde{\mathtt{U}}_j = \tilde{\mathtt{U}}_{j\mathtt{Q}}\tilde{\mathtt{U}}_{j\mathtt{R}}$, we can transform $(\tilde{\mathtt{U}}_j^{\top}\tilde{\mathtt{U}}_j)^{-1}$ to $\tilde{\mathtt{U}}_{j\mathtt{R}}^{-1}\tilde{\mathtt{U}}_{j\mathtt{R}}^{-\top}$. The last term then becomes

$$\begin{aligned}\mathtt{K}_{mr}^{\top}\mathtt{Z}_j^*(\tilde{\mathtt{U}}_j^{\top}\tilde{\mathtt{U}}_j)^{-1}\mathtt{Z}_j^{*\top}\mathtt{K}_{mr} &= (\mathtt{I} \otimes (\mathbf{w}_j \odot \mathbf{r}_j^*))\tilde{\mathtt{U}}_{j\mathtt{R}}^{-1}\tilde{\mathtt{U}}_{j\mathtt{R}}^{-\top}(\mathtt{I} \otimes (\mathbf{w}_j \odot \mathbf{r}_j^*)^{\top}) && /\!/ \textit{ noting (8)} \quad (36)\\ &= \tilde{\mathtt{U}}_{j\mathtt{R}}^{-1}\tilde{\mathtt{U}}_{j\mathtt{R}}^{-\top} \otimes (\mathbf{w}_j \odot \mathbf{r}_j^*)(\mathbf{w}_j \odot \mathbf{r}_j^*)^{\top}. && /\!/ \textit{ noting (5)} \quad (37)\end{aligned}$$

Hence, all the terms inside the square brackets in (31) can be represented as the Kronecker product of two symmetric matrices, which has an additional internal symmetry that can be exploited. This means that computing the Gauss-Newton matrix for un-regularized VarPro can be sped up by factor of 4 at most.

5. Computing Mean time to second success (MTSS)

As mentioned in [1], computing MTSS is based on the assumption that there exists a known best optimum on a dataset to which thousands or millions of past runs have converged.

Given that we have such dataset, we first draw a set of random starting points, usually between 20 to 100, from an isotropic Normal distribution (i.e. `U = randn(m, r)` with some predefined seed). For each sample of starting points, we run each algorithm and observe two things,

1. whether the algorithm has successfully converged to the best-known optimum, and
2. the elapsed time.

Since the samples of starting points have been drawn in a pseudo-random manner with a monotonically-increasing seed number, we can now estimate how long it would have taken each algorithm to observe two successes starting from each sample number. The average of these yields MTSS.

An example is included in the next section to demonstrate this procedures more clearly.

5.1. An example illustration

Table 3 shows some experimental data produced by an algorithm on a dataset of which the best-known optimum is 1.225. For each sample of starting points, the final cost and the elapsed time have been recorded.

We first determine which samples have yielded *success*, which is defined as convergence to the best-known optimum (1.225). Then, for each sample number, we can calculate up to which sample number we should have run to obtain two successes. e.g. when starting from sample number 5, we would have needed to run the algorithm up to sample number 7. This allows us to estimate the times required for two successes starting from each sample number (see Table 3). MTSS is simply the average of these times, $432.4/7 = 61.8$.

Sample no.	1	2	3	4	5	6	7	8	9	10
Final cost	1.523	1.225	1.647	1.225	1.52	1.225	1.225	1.647	1.225	1.774
Elapsed time (s)	23.2	15.1	24.7	19.5	25.4	16.3	15.5	21.2	17.8	21.0
Success (S)		S		S		S	S		S	
Sample numbers of next 2 successes	2 & 4	2 & 4	4 & 6	4 & 6	6 & 7	6 & 7	7 & 9	N/A	N/A	N/A
Time required for 2 successes (s)	82.5	59.3	85.9	61.2	57.2	31.8	54.5	N/A	N/A	N/A

Table 3: Above example data is used to compute MTSS of an algorithm. We assume that each sample starts from randomly-drawn initial points and that the best-known optimum for this dataset is 1.225 after millions of trials. In this case, MTSS is the average of the bottom row values which is 61.8.

6. Results

The supplementary package includes a set of csv files that were used to generate the results figures in [1]. They can be found in the `<main>/Results` directory.

6.1. Limitations

Some results are incomplete for the following reasons:

1. RTRMC [3] fails to run on FAC and Scu datasets. The algorithm has also failed on several occasions when running on other datasets. The suspected cause of such failure is the algorithm's use of Cholesky decomposition, which is unable to decompose non-positive definite matrices.

2. The original Damped Wiberg code (TO_DW) [7] fails to run on FAC and Scu datasets. The way in which the code implements QR-decomposition requires datasets to be *trimmed* in advance so that each column in the measurement matrix has at least the number of visible entries equal to the estimated rank of the factorization model.

3. CSF [4] runs extremely slowly on UGb dataset.

4. Our Ceres-implemented algorithms are not yet able to take in sparse measurement matrices, which is required to handle large datasets such as NET.

References

[1] Authors. Secrets of matrix factorization: Approximations, numerics and manifold optimization. 2015. 1, 2, 5, 6

[2] Authors. Secrets of matrix factorization: Further derivations and comparisons. Technical report, further.pdf, 2015. 2, 4, 5

[3] N. Boumal and P.-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pages 406–414. 2011. 6

[4] P. F. Gotardo and A. M. Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(10):2051–2065, Oct 2011. 6

[5] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. 3rd edition, 2007. 2

[6] T. P. Minka. Old and new matrix algebra useful for statistics. Technical report, Microsoft Research, 2000. 2

[7] T. Okatani, T. Yoshida, and K. Deguchi. Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms. In *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*, pages 842–849, 2011. 6