Numerische Mathematik (2021) 147:611–650
https://doi.org/10.1007/s00211-021-01178-8

Numerische Mathematik

Minimum-correction second-moment matching: theory, algorithms and applications

Jing Lin[1] · **Pierre F. J. Lermusiaux**[1]

Received: 18 September 2019 / Revised: 7 November 2020 / Accepted: 24 December 2020 /
Published online: 6 February 2021

Abstract
We address the problem of finding the closest matrix $\tilde{\boldsymbol{U}}$ to a given $\boldsymbol{U}$ under the constraint that a prescribed second-moment matrix $\tilde{\boldsymbol{P}}$ must be matched, i.e. $\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{P}}$. We obtain a closed-form formula for the unique global optimizer $\tilde{\boldsymbol{U}}$ for the full-rank case, that is related to $\boldsymbol{U}$ by an SPD (symmetric positive definite) linear transform. This result is generalized to rank-deficient cases as well as to infinite dimensions. We highlight the geometric intuition behind the theory and study the problem's rich connections to minimum congruence transform, generalized polar decomposition, optimal transport, and rank-deficient data assimilation. In the special case of $\tilde{\boldsymbol{P}} = \boldsymbol{I}$, minimum-correction second-moment matching reduces to the well-studied optimal orthonormalization problem. We investigate the general strategies for numerically computing the optimizer and analyze existing polar decomposition and matrix square root algorithms. We modify and stabilize two Newton iterations previously deemed unstable for computing the matrix square root, such that they can now be used to efficiently compute both the orthogonal polar factor and the SPD square root. We then verify the higher performance of the various new algorithms using benchmark cases with randomly generated matrices. Lastly, we complete two applications for the stochastic Lorenz-96 dynamical system in a chaotic regime. In reduced subspace tracking using dynamically orthogonal equations, we maintain the numerical orthonormality and continuity of time-varying base vectors. In ensemble square root filtering for data assimilation, the prior samples are transformed into posterior ones by matching the covariance given by the Kalman update while also minimizing the corrections to the prior samples.

Mathematics Subject Classification 65F25 · 15A23 · 60H15

✉ Pierre F. J. Lermusiaux
pierrel@mit.edu

Jing Lin
linjing@mit.edu

[1] Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

1 Introduction

The second-moment matrix of $\boldsymbol{U} \in \mathcal{M}_{m\times n}(\mathbb{R})$[1] defined as $\boldsymbol{P} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Gamma}_m\boldsymbol{U} \in \mathcal{M}_{n\times n}$ contains all the pairwise inner products of $\boldsymbol{U}$'s columns weighted by $\boldsymbol{\Gamma}_m$. When we view $\boldsymbol{U}$'s columns as vectors in an inner product space, $\boldsymbol{P}$ is called the Gram matrix. When the $\boldsymbol{U}$'s rows are samples of a zero-mean random vector, $\boldsymbol{P}$ becomes the sample covariance matrix with $\boldsymbol{\Gamma}_m = (1/m)\boldsymbol{I}$. Here we unify them as second-moment matrices. In many computational problems, we encounter the task of correcting a given matrix $\boldsymbol{U}$ to some $\tilde{\boldsymbol{U}}$ that matches a prescribed second-moment matrix $\tilde{\boldsymbol{P}}$, i.e. $\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{\Gamma}_m\tilde{\boldsymbol{U}}$. However, such a $\tilde{\boldsymbol{U}}$ is not unique. When there is no physical information that favors one choice over another, a natural approach is to aim for a minimal correction $\tilde{\boldsymbol{U}} - \boldsymbol{U}$ to avoid introducing numerical artifacts.

Precisely, given a target symmetric positive definite (SPD) second-moment matrix $\tilde{\boldsymbol{P}}$, as well as an inner product on $\mathbb{R}^m$ and on $\mathbb{R}^n$ defined by the SPD weight matrix $\boldsymbol{\Gamma}_m \in \mathcal{M}_{m\times m}$ and $\boldsymbol{\Gamma}_n \in \mathcal{M}_{n\times n}$, respectively, we want to solve the optimization

$$\underset{\tilde{\boldsymbol{U}}\in\mathcal{M}_{m\times n}:\ \tilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{\Gamma}_m\tilde{\boldsymbol{U}}=\tilde{\boldsymbol{P}}}{\arg\min} \|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|^2_{\mathrm{F},\boldsymbol{\Gamma}_m,\boldsymbol{\Gamma}_n}, \tag{1}$$

where $\|\cdot\|_{\mathrm{F},\boldsymbol{\Gamma}_m,\boldsymbol{\Gamma}_n}$ is the Frobenius norm weighted by $\boldsymbol{\Gamma}_m$ and $\boldsymbol{\Gamma}_n$:

$$\begin{aligned}\|\boldsymbol{V}\|_{\mathrm{F},\boldsymbol{\Gamma}_m,\boldsymbol{\Gamma}_n} &\triangleq \mathrm{tr}(\sqrt{\boldsymbol{\Gamma}_n}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Gamma}_m\boldsymbol{V}\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}})\\ &= \mathrm{tr}(\sqrt{\boldsymbol{\Gamma}_m}\boldsymbol{V}\boldsymbol{\Gamma}_n\boldsymbol{V}^{\mathrm{T}}\sqrt{\boldsymbol{\Gamma}_m}^{\mathrm{T}}) = \|\boldsymbol{V}^{\mathrm{T}}\|_{\mathrm{F},\boldsymbol{\Gamma}_n,\boldsymbol{\Gamma}_m}.\end{aligned} \tag{2}$$

Here $\sqrt{\boldsymbol{A}}$ denotes a matrix square root of an SPD matrix $\boldsymbol{A}$, viewed as a self-adjoint positive operator, i.e. $\sqrt{\boldsymbol{A}}^{\mathrm{T}}\sqrt{\boldsymbol{A}} = \boldsymbol{A}$. Throughout this paper, "matrix square root" will refer to this definition (rather than $\sqrt{\boldsymbol{A}}^2 = \boldsymbol{A}$), unless otherwise mentioned. Such a square root is not unique and is subject to the unitary freedom, i.e. if $\boldsymbol{R}$ is a square root of $\boldsymbol{A}$, the set $\{\boldsymbol{Q}\boldsymbol{R} : \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{I}\}$ contains all square roots of $\boldsymbol{A}$. The choice of $\sqrt{\boldsymbol{A}}$ can be arbitrary but must be consistent. Moreover, we denote by $\boldsymbol{A}^{1/2}$ the unique SPD square root of $\boldsymbol{A}$.

For the optimization problem (1), we can always eliminate the inner product's weights $\boldsymbol{\Gamma}_m$ and $\boldsymbol{\Gamma}_n$ by variable substitution. If we take $\boldsymbol{W} = \sqrt{\boldsymbol{\Gamma}_m}\boldsymbol{U}\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}}$ and $\tilde{\boldsymbol{W}} = \sqrt{\boldsymbol{\Gamma}_m}\tilde{\boldsymbol{U}}\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}}$, then (1) reduces to

$$\underset{\tilde{\boldsymbol{W}}\in\mathcal{M}_{m\times n}:\ \tilde{\boldsymbol{W}}^{\mathrm{T}}\tilde{\boldsymbol{W}}=\sqrt{\boldsymbol{\Gamma}_n}\tilde{\boldsymbol{P}}\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}}}{\arg\min} \|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|^2_{\mathrm{F}}, \tag{3}$$

where $\|\boldsymbol{V}\|_{\mathrm{F}} = \mathrm{tr}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V})$ is the unweighted Frobenius norm and $\sqrt{\boldsymbol{\Gamma}_n}\tilde{\boldsymbol{P}}\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}}$ becomes the new target second-moment matrix. Therefore, without loss of generality, we will restrict ourselves to the simplified problem

[1] Since we will exclusively focus on real matrices in this paper, the explicit specification "$(\mathbb{R})$" of the field of matrix elements will be dropped hereafter.

$$\underset{\tilde{\boldsymbol{U}} \in \mathcal{M}_{m\times n}:\, \tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}}=\tilde{\boldsymbol{P}}}{\arg\min} \quad \|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F}}^2. \tag{4}$$

Note that we can also scale $\tilde{\boldsymbol{U}}$ to make $\tilde{\boldsymbol{P}} = \boldsymbol{I}$, but then $\boldsymbol{\Gamma}_n \neq \boldsymbol{I}$ in general. We choose to have $\boldsymbol{\Gamma}_n = \boldsymbol{I}$ and $\tilde{\boldsymbol{P}}$ arbitrary because it renders the problem's structure symmetric and leads to more insightful interpretations as will be seen in Sect. 2. How the two ways of scaling are connected will be mentioned in Sect. 2.3. Besides $\tilde{\boldsymbol{P}}$ being SPD, we further assume that $m \geq n$ and $\boldsymbol{U}$ has full column rank, so $\boldsymbol{P} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ is SPD. The rank-deficiency complications due to the violation of these conditions are addressed in Sect. 2.4.

Minimum-correction second-moment matching has many applications. In the special case of $\tilde{\boldsymbol{P}} = \boldsymbol{I}$, the task reduces to orthonormalization. Many algorithms involve tracking a varying orthogonal matrix [14,45], such as in optimization [1] or solving matrix differential equations with orthogonality constraints [22, sec. IV.9]. There does exist orthogonality-preserving algorithms for some applications. For example, [56] studies an orthogonality-preserving curvilinear search algorithm for optimization on a Stiefel manifold [14]. For time integration, [11,24] investigates orthogonality-preserving Gauss–Legendre Runge–Kutta schemes. However, such algorithms are typically constrained and the choices limited. Most matrix update algorithms prioritize other goals such as steepest descent optimization [50] or accurate time integration [10,30,39,46,53]. In such cases, after an orthogonal matrix is updated, a deviation from orthogonality is usually incurred by the numerical discretization of the matrix's continuous evolution. Therefore, we need to orthonormalize the matrix in such a way that this *entry-wise* continuous evolution is preserved [18,21]. A natural idea is thus to find the closest point to the updated matrix on the Stiefel manifold [2,11,25]. We will see a test case of this kind in Sect. 5.

In the general case with $\tilde{\boldsymbol{P}} \neq \boldsymbol{I}$, one application is minimum-correction covariance matching. For example, in an ensemble Kalman filter (EnKF) [16] for data assimilation, given the prior samples of a random vector in the rows of $\boldsymbol{U}$ and some observation data, the goal is to obtain posterior samples $\tilde{\boldsymbol{U}}$ such that their empirical mean and covariance match those of the posterior distribution obtained by the Kalman update. A variant proposed by [41] consists of updating the mean in the same way as EnKF but making the posterior sample variation $\tilde{\boldsymbol{U}}$ (with mean removed) as close as possible to the prior counterpart $\boldsymbol{U}$ under the constraint that the empirical covariance $(1/m)\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}}$ must match the one obtained by the Kalman update. We will use such a test case in Sect. 6.

Although the minimum-correction orthonormalization is relatively well understood with efficient algorithms developed, its general second-moment matching counterpart has not yet been studied and analyzed in a unified and systematic way. We aim to fill this gap in the present paper.

In Sect. 2, we solve the optimization (4) analytically with a constructive proof (Sect. 2.1 and "Appendix A"), provide geometric intuition, solve rank deficient cases (2.4), and discuss the connections to other topics including minimum congruence transform (Sect. 2.2) and generalized polar decomposition (Sect. 2.3). We also address the generalization to infinite dimensions in "Appendix B" and connect to optimal trans-

port in "Appendix C". In Sect. 3, we obtain numerical strategies for computing the optimal solution and review existing algorithms for polar decomposition and matrix square root that play an essential role in minimum-correction second-moment matching. We then modify and stabilize two Newton iterations deemed unstable previously, that can now be used to efficiently compute both the orthogonal polar factor and the SPD square root. We compare algorithms in terms of accuracy, cost and robustness. Results are benchmarked using randomly generated matrices in Sect. 4. In Sect. 5, we show that these algorithms can be used in a subspace tracking method for stochastic ODEs and PDEs, the dynamically orthogonal equations [17,46], to maintain both the orthonormality and the time continuity of the base vectors. Furthermore, we show in Sect. 6 how the algorithms for the general case can be applied to an ensemble-based filter for data assimilation, as introduced in [41]. Lastly, we conclude in Sect. 7.

2 Theory of minimum-correction second-moment matching

2.1 Analytic expression of the global optimizer

The constrained optimization (4) has a compact feasible region and a smooth objective function, so a global optimizer exists and is a critical point of the Lagrangian. Hence, we employ the approach of Lagrangian multipliers as in [41]. We complete a different proof, adopt more compact matrix notations, and provide new extensive discussions. The results are summarized in the following theorem with the proof in "Appendix A".

Theorem 2.1 (Minimum-correction second-moment matching) *The global optimum of* (4)*, assuming that* $m \geq n$ *and* $\boldsymbol{U}$ *has full column rank, is achieved by applying the unique* $n \times n$ *SPD linear transform to* $\boldsymbol{U}$ *that matches the second-moment matrix of the result* $\tilde{\boldsymbol{U}}$ *to the target* $\tilde{\boldsymbol{P}}$*. More precisely,*

$$\underset{\tilde{\boldsymbol{U}} \in \mathcal{M}_{m\times n}:\, \tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}}=\tilde{\boldsymbol{P}}}{\arg\min} \quad \|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F}}^2 = \boldsymbol{U}\boldsymbol{A}_*, \tag{5}$$

$$\|\boldsymbol{U}\boldsymbol{A}_* - \boldsymbol{U}\|_{\mathrm{F}}^2 = \mathrm{tr}\left[\boldsymbol{P} + \tilde{\boldsymbol{P}} - 2\left(\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}\right)^{1/2}\right] = \mathrm{tr}\left[\boldsymbol{P} + \tilde{\boldsymbol{P}} - 2\left(\boldsymbol{P}\tilde{\boldsymbol{P}}\right)^{1/2}\right] \tag{6}$$

[2] *with the SPD linear transform* $\boldsymbol{A}_*$ *being*

$$\begin{aligned} \boldsymbol{A}_* &= \sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}\left(\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}\right)^{-1/2}\sqrt{\tilde{\boldsymbol{P}}} \\ &= \sqrt{\boldsymbol{P}}^{-1}\left(\sqrt{\boldsymbol{P}}\tilde{\boldsymbol{P}}\sqrt{\boldsymbol{P}}^{\mathrm{T}}\right)^{1/2}\sqrt{\boldsymbol{P}}^{-\mathrm{T}}. \end{aligned} \tag{7}$$

[2] For square $\boldsymbol{A}$ diagonalizable with nonnegative eigenvalues and EVD $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}$, we define $\boldsymbol{A}^{1/2} \triangleq \boldsymbol{V}\boldsymbol{\Lambda}^{1/2}\boldsymbol{V}^{-1}$. Since $\boldsymbol{P}\tilde{\boldsymbol{P}} = \sqrt{\tilde{\boldsymbol{P}}}^{-1}\left(\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}\right)\sqrt{\tilde{\boldsymbol{P}}}$ is similar to $\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}$, $\boldsymbol{P}\tilde{\boldsymbol{P}}$, though not symmetric in general, is diagonalizable with positive eigenvalues and EVD $\boldsymbol{P}\tilde{\boldsymbol{P}} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}$. Therefore, $\mathrm{tr}((\boldsymbol{P}\tilde{\boldsymbol{P}})^{1/2}) = \sum_{i=1}^{n}\sqrt{\lambda_i} = \mathrm{tr}((\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}})^{1/2})$.

'Note that when $\tilde{\boldsymbol{P}} = \boldsymbol{P}$, we have $\boldsymbol{A}_* = \boldsymbol{I}$, which correctly gives the null correction in this trivial case. Moreover, $\tilde{\boldsymbol{U}}_* = \boldsymbol{U}\boldsymbol{A}_*$ implies $\mathrm{Col}(\tilde{\boldsymbol{U}}_*) = \mathrm{Col}(\boldsymbol{U})$, which is particularly important for covariance matching because a zero-mean $\boldsymbol{U}$ (i.e. $\boldsymbol{e}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{0}$, where $\boldsymbol{e}^{\mathrm{T}} = [1, \ldots, 1]$) will yield a zero-mean $\tilde{\boldsymbol{U}}_*$.

Since $\mathcal{M}_{m\times n}$ can be equipped with the inner product $\langle \boldsymbol{U}, \boldsymbol{V}\rangle_{\mathrm{F}} \triangleq \mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V})$ which induces the Frobenius norm $\|\cdot\|_{\mathrm{F}}$, Theorem 2.1 indeed characterizes the orthogonal projection $\mathcal{P}_{\tilde{\boldsymbol{P}}}$ of any $\boldsymbol{U} \in \mathcal{M}_{m\times n}$ onto the $\left(mn - \frac{1}{2}n(n+1)\right)$ dimensional sub-manifold $\mathcal{S}_{\tilde{\boldsymbol{P}}} = \{\tilde{\boldsymbol{U}} \in \mathcal{M}_{m\times n} : \tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{P}}\}$ as the unique $n \times n$ SPD linear transform that maps $\boldsymbol{U}$ into the sub-manifold.

Moreover, if we invert the first expression in (7), we obtain

$$\boldsymbol{A}_*^{-1} = \sqrt{\tilde{\boldsymbol{P}}}^{-1}(\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}})^{1/2}\sqrt{\tilde{\boldsymbol{P}}}^{-\mathrm{T}},$$

which is nothing but the second expression in (7) with $\tilde{\boldsymbol{P}}$ and $\boldsymbol{P}$ switched. This implies that $\tilde{\boldsymbol{U}}\boldsymbol{A}_*^{-1} = \boldsymbol{U}$ projects $\tilde{\boldsymbol{U}}$ orthogonally onto $\mathcal{S}_{\boldsymbol{P}}$, i.e.

$$\underset{\boldsymbol{U}\in\mathcal{M}_{m\times n}:\ \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{P}}{\arg\min} \|\boldsymbol{U} - \tilde{\boldsymbol{U}}\|_{\mathrm{F}}^2 = \tilde{\boldsymbol{U}}\boldsymbol{A}_*^{-1} \tag{8}$$

with $\boldsymbol{A}_*$ given by (7). Therefore, the orthogonal projection $\mathcal{P}_{\boldsymbol{P}}$ and $\mathcal{P}_{\tilde{\boldsymbol{P}}}$ specify a bijection between $\mathcal{S}_{\boldsymbol{P}}$ and $\mathcal{S}_{\tilde{\boldsymbol{P}}}$ such that

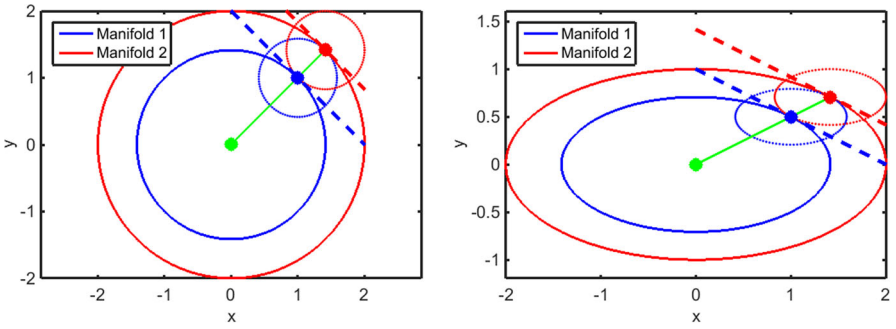
$$\mathcal{P}_{\boldsymbol{P}} \circ \mathcal{P}_{\tilde{\boldsymbol{P}}}\big|_{\mathcal{S}_{\boldsymbol{P}}} = \mathrm{Id}|_{\mathcal{S}_{\boldsymbol{P}}}, \qquad \mathcal{P}_{\tilde{\boldsymbol{P}}} \circ \mathcal{P}_{\boldsymbol{P}}|_{\mathcal{S}_{\tilde{\boldsymbol{P}}}} = \mathrm{Id}|_{\mathcal{S}_{\tilde{\boldsymbol{P}}}}. \tag{9}$$

In addition, since $\boldsymbol{A}_*$ depends on $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ only, both $\mathcal{P}_{\tilde{\boldsymbol{P}}}\big|_{\mathcal{S}_{\boldsymbol{P}}}$ and $\mathcal{P}_{\boldsymbol{P}}|_{\mathcal{S}_{\tilde{\boldsymbol{P}}}}$ are characterized by a constant SPD linear transform on $\mathbb{R}^n$.

Furthermore, the simple case of $n = 1$ helps us build a geometric intuition of the above results, as illustrated in Fig. 1. When $n = 1$, $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ reduce to a positive real number, while $\mathcal{S}_{\boldsymbol{P}}$ and $\mathcal{S}_{\tilde{\boldsymbol{P}}}$ reduce to two concentric $(m-1)$-spheres (or two $(m-1)$-ellipsoids with the same eccentricity if $\boldsymbol{\Gamma}_m \neq \lambda\boldsymbol{I}$). Hence, it is straightforward to see that the rays from the origin provide a bijection between the two spheres. This mapping is equivalent to scaling one sphere by a constant ratio. As Fig. 1 shows, the points on the two manifolds can be paired up such that in each point pair, the blue point is the closet point on the blue manifold to the red one and vice versa.

However, the case of $n = 1$ also over-simplifies some aspects of the problem and could be misleading. For example, Fig. 1 could create the illusion that the projection operator is transitive, i.e. projecting from $\mathcal{S}_{\boldsymbol{P}_1}$ to $\mathcal{S}_{\boldsymbol{P}_3}$ is equivalent to first projecting from $\mathcal{S}_{\boldsymbol{P}_1}$ to some intermediate $\mathcal{S}_{\boldsymbol{P}_2}$ and then from $\mathcal{S}_{\boldsymbol{P}_2}$ to $\mathcal{S}_{\boldsymbol{P}_3}$. We thus emphasize that in general,

$$\mathcal{P}_{\boldsymbol{P}_3} \circ \mathcal{P}_{\boldsymbol{P}_2}\big|_{\mathcal{S}_{\boldsymbol{P}_1}} \neq \mathcal{P}_{\boldsymbol{P}_3}\big|_{\mathcal{S}_{\boldsymbol{P}_1}}, \qquad \boldsymbol{A}_*^{1\to 2}\boldsymbol{A}_*^{2\to 3} \neq \boldsymbol{A}_*^{1\to 3}, \tag{10}$$



(a) $\boldsymbol{\Gamma}_m = \mathrm{diag}(1, 1)$ **(b)** $\boldsymbol{\Gamma}_m = \mathrm{diag}(1, 4)$

Fig. 1 A geometric picture of Theorem 2.1 for $m = 2$ and $n = 1$

since $\boldsymbol{A}_*^{1\to 3}$ is symmetric but $\boldsymbol{A}_*^{1\to 2}\boldsymbol{A}_*^{2\to 3}$ is not in general. This is especially clear when $\boldsymbol{P}_2 = \boldsymbol{I}$, which implies $\boldsymbol{A}_*^{1\to 2} = \boldsymbol{P}_1^{-1/2}$ and $\boldsymbol{A}_*^{2\to 3} = \boldsymbol{P}_3^{1/2}$ according to (7). Hence, whenever $\boldsymbol{P}_1$ and $\boldsymbol{P}_3$ are not commutative, $\boldsymbol{A}_*^{1\to 2}\boldsymbol{A}_*^{2\to 3}$ is not symmetric.

Finally, with (3), we obtain the global optimizer for the general setting (1).

Corollary 2.1 (General setting with inner product weighted by $\boldsymbol{\Gamma}_m$ and $\boldsymbol{\Gamma}_n$)

$$\underset{\tilde{\boldsymbol{U}}\in\mathcal{M}_{m\times n}:\, \tilde{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{\Gamma}_m\tilde{\boldsymbol{U}}=\tilde{\boldsymbol{P}}}{\arg\min} \quad \|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F},\boldsymbol{\Gamma}_m,\boldsymbol{\Gamma}_n}^2 = \boldsymbol{U}(\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}}\boldsymbol{A}_*\sqrt{\boldsymbol{\Gamma}_n}^{-\mathrm{T}}) \tag{11}$$

$$\|\tilde{\boldsymbol{U}}_* - \boldsymbol{U}\|_{\mathrm{F}}^2 = \mathrm{tr}[\boldsymbol{P}_\Gamma + \tilde{\boldsymbol{P}}_\Gamma - 2(\boldsymbol{P}_\Gamma\tilde{\boldsymbol{P}}_\Gamma)^{1/2}] \tag{12}$$

$$\begin{aligned}\boldsymbol{A}_* &= \sqrt{\tilde{\boldsymbol{P}}_\Gamma}^{\mathrm{T}}\,(\sqrt{\tilde{\boldsymbol{P}}_\Gamma}\boldsymbol{P}_\Gamma\sqrt{\tilde{\boldsymbol{P}}_\Gamma}^{\mathrm{T}})^{-1/2}\sqrt{\tilde{\boldsymbol{P}}_\Gamma} \\ &= \sqrt{\boldsymbol{P}_\Gamma}^{-1}(\sqrt{\boldsymbol{P}_\Gamma}\tilde{\boldsymbol{P}}_\Gamma\sqrt{\boldsymbol{P}_\Gamma}^{\mathrm{T}})^{1/2}\sqrt{\boldsymbol{P}_\Gamma}^{-\mathrm{T}}.\end{aligned} \tag{13}$$

$$\tilde{\boldsymbol{P}}_\Gamma = \sqrt{\boldsymbol{\Gamma}_n}\tilde{\boldsymbol{P}}\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}}, \qquad \boldsymbol{P}_\Gamma = \sqrt{\boldsymbol{\Gamma}_n}\boldsymbol{P}\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}}. \tag{14}$$

We have shown previously that the choice of square root $\sqrt{\tilde{\boldsymbol{P}}_\Gamma}$ and $\sqrt{\boldsymbol{P}_\Gamma}$ in (13) is irrelevant. Here we can show that the choice of square root $\sqrt{\boldsymbol{\Gamma}_n}$ is also irrelevant. Under the unitary freedom $\sqrt{\boldsymbol{\Gamma}_n} \to \boldsymbol{Q}\sqrt{\boldsymbol{\Gamma}_n}$ and thus $\boldsymbol{P}_\Gamma \to \boldsymbol{Q}\boldsymbol{P}_\Gamma\boldsymbol{Q}^{\mathrm{T}}$, we can choose $\sqrt{\boldsymbol{Q}\boldsymbol{P}_\Gamma\boldsymbol{Q}^{\mathrm{T}}}$ to be $\boldsymbol{Q}\sqrt{\boldsymbol{P}_\Gamma}\boldsymbol{Q}^{\mathrm{T}}$. Therefore, the $\boldsymbol{Q}$'s will pair up with the $\boldsymbol{Q}^{\mathrm{T}}$'s and cancel each other, which renders the linear transform $\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}}\boldsymbol{A}_*\sqrt{\boldsymbol{\Gamma}_n}^{-\mathrm{T}}$ invariant. However, the choice of the inner product weights $\boldsymbol{\Gamma}_m$ and $\boldsymbol{\Gamma}_n$ themselves does affect the optimum, although they play quite different roles.

$\boldsymbol{\Gamma}_m$ defines the inner product on $\mathbb{R}^m$ for computing the second-moment matrices. This inner product usually has physical meaning and is commonly determined by the underlying problem. For example, if each column of $\boldsymbol{U}$ is a discrete representation of a function $f(x)$ over some domain Ω, then $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Gamma}_m\boldsymbol{v}$ could approximate the inner product $\int_\Omega u(x)v(x)\mathrm{d}x$. If each row of $\boldsymbol{U}$ is a sample of a zero-mean random vector, then $\boldsymbol{\Gamma}_m = \frac{1}{m}\mathrm{diag}(1, \ldots, 1)$ will make $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Gamma}_m\boldsymbol{v}$ the empirical covariance matrix. Note

that $\boldsymbol{A}_*$ does not depend on $\boldsymbol{\Gamma}_m$ explicitly, but only through how $\tilde{\boldsymbol{P}}$ and $\boldsymbol{P}$ are computed based on $\boldsymbol{\Gamma}_m$.

In contrast, $\boldsymbol{\Gamma}_n$ is more of a tunable numerical parameter. It controls how we want to weight the columns of $\boldsymbol{U}$ differently so that when we evaluate the closeness between $\tilde{\boldsymbol{U}}$ and $\boldsymbol{U}$, we give priority to some columns over others. Such differentiation could be advantageous if, for example, some of the columns of $\boldsymbol{U}$ contain more reliable information (e.g. due to smaller numerical errors) than others, or the columns are different by several orders of magnitude and the small columns barely contribute to the closeness metric. If the magnitude is of concern, a natural choice of a scaling matrix is either $\boldsymbol{\Gamma}_n = \boldsymbol{P}^{-1}$ or $\boldsymbol{\Gamma}_n = \tilde{\boldsymbol{P}}^{-1}$:

$$\boldsymbol{\Gamma}_n = \boldsymbol{P}^{-1} \Longrightarrow \sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}} \boldsymbol{A}_* \sqrt{\boldsymbol{\Gamma}_n}^{-\mathrm{T}} = \sqrt{\boldsymbol{P}}^{-1} (\sqrt{\boldsymbol{P}}^{-\mathrm{T}} \tilde{\boldsymbol{P}} \sqrt{\boldsymbol{P}}^{-1})^{1/2} \sqrt{\boldsymbol{P}}$$
$$\boldsymbol{\Gamma}_n = \tilde{\boldsymbol{P}}^{-1} \Longrightarrow \sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}} \boldsymbol{A}_* \sqrt{\boldsymbol{\Gamma}_n}^{-\mathrm{T}} = \sqrt{\tilde{\boldsymbol{P}}}^{-1} (\sqrt{\tilde{\boldsymbol{P}}}^{-\mathrm{T}} \boldsymbol{P} \sqrt{\tilde{\boldsymbol{P}}}^{-1})^{-1/2} \sqrt{\tilde{\boldsymbol{P}}}.$$

Surprisingly, as [41] points out, these two choices of $\boldsymbol{\Gamma}_n$ lead to the same optimal mapping. By the properties (8) and (9) of the projection operators, we can quickly show the equivalence between these two expressions by noticing that either one can be obtained by inverting the other and switching $\boldsymbol{P}$ with $\tilde{\boldsymbol{P}}$. Last but not least, note that the optimal mapping $\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}} \boldsymbol{A}_* \sqrt{\boldsymbol{\Gamma}_n}^{-\mathrm{T}}$ explicitly depends on $\boldsymbol{\Gamma}_n$ and when $\boldsymbol{\Gamma}_n$ is not a scalar matrix, the optimal mapping is in general not symmetric.

2.2 Minimum congruence transform

In the optimization problem (4), the candidate $\tilde{\boldsymbol{U}}$ can come from all of $\mathcal{M}_{m\times n}$, but the global optimizer $\boldsymbol{U}\boldsymbol{A}_*$ turns out to share the column space of $\boldsymbol{U}$. This implies that if we restrict the candidate set to the n^2-dimensional subspace $\{\tilde{\boldsymbol{U}} \in \mathcal{M}_{m\times n} : \mathrm{Im}(\tilde{\boldsymbol{U}}) \subset \mathrm{Im}(\boldsymbol{U})\} = \{\tilde{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{A} : \boldsymbol{A} \in \mathcal{M}_{n\times n}\}$ of the mn-dimensional $\mathcal{M}_{m\times n}$, the global optimizer is still given by (7). This is equivalent to requiring the second-moment matching to be achieved by an n-by-n linear transform. With this restriction, the objective function in (4) becomes

$$\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F}}^2 = \|\boldsymbol{U}\boldsymbol{A} - \boldsymbol{U}\|_{\mathrm{F}}^2 = \mathrm{tr}\left((\boldsymbol{A}-\boldsymbol{I})^{\mathrm{T}} \boldsymbol{P} (\boldsymbol{A}-\boldsymbol{I})\right) = \|\boldsymbol{A}-\boldsymbol{I}\|_{\mathrm{F},\boldsymbol{P}}^2, \quad (15)$$

where the minimum correction of $\boldsymbol{U}$ translates to the minimum operation of $\boldsymbol{A}$, quantified by being the closest linear transform to the identity mapping. Besides, the second-moment matching constraint is now on $\boldsymbol{A}$

$$\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{U}}^{\mathrm{T}} \tilde{\boldsymbol{U}} = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{U}^{\mathrm{T}} \boldsymbol{U} \boldsymbol{A} = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{P} \boldsymbol{A}.$$

Therefore, Theorem 2.1 implies the following corollary.

Corollary 2.2 (Minimum congruence transform)

$$\underset{\boldsymbol{A}\in\mathcal{M}_{n\times n}:\ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{A}=\tilde{\boldsymbol{P}}}{\arg\min} \|\boldsymbol{A}-\boldsymbol{I}\|_{\mathrm{F},\boldsymbol{P}}^2 = \boldsymbol{A}_* \quad (16)$$

with $\boldsymbol{A}_*$ given by (7) for any SPD $\tilde{\boldsymbol{P}}$ and $\boldsymbol{P}$ in $\mathcal{M}_{n\times n}$.

Note that this optimization problem is related to $\boldsymbol{U}$ only through its second-moment matrix $\boldsymbol{P}$. We can indeed interpret it in a way that does not involve $\boldsymbol{U}$. Since $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{A} = \tilde{\boldsymbol{P}}$ indicates that $\tilde{\boldsymbol{P}}$ is congruent with $\boldsymbol{P}$ through the congruence transform $\boldsymbol{A}$, (16) can be viewed as finding the minimum congruence transform between two SPD matrices. Here "minimum" again refers to the minimum action of an operator.

2.3 Generalized polar decomposition

An important and familiar special case of the problem (4) is when $\tilde{\boldsymbol{P}} = \boldsymbol{I}$ and the task reduces to minimum-correction orthonormalization, i.e. finding the closest orthogonal matrix $\tilde{\boldsymbol{U}}$ to an arbitrary $\boldsymbol{U} \in \mathcal{M}_{m\times n}$ with full column rank. In this case, the candidate solutions form the Stiefel manifold $\mathcal{S}_{\boldsymbol{I}} = \{\boldsymbol{V} \in \mathcal{M}_{m\times n} : \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}\}$. Therefore, Theorem 2.1 reduces to the well-known fact that a matrix can be orthogonally projected onto the Stiefel manifold through its polar decomposition [2], summarized in the following corollary.

Corollary 2.3 (Orthogonal projection onto the Stiefel manifold by polar decomposition) *Given $\boldsymbol{U} \in \mathcal{M}_{m\times n}$ with full column rank and $\boldsymbol{P} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$, we have*

$$\tilde{\boldsymbol{U}}_* = \underset{\tilde{\boldsymbol{U}}\in\mathcal{S}_{\boldsymbol{I}}}{\arg\min} \|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F}}^2 = \boldsymbol{U}\boldsymbol{P}^{-1/2}, \tag{17}$$

where $\boldsymbol{U} = \tilde{\boldsymbol{U}}_*\boldsymbol{P}^{1/2}$ gives the unique polar decomposition of $\boldsymbol{U}$.

Note that Corollary 2.1 implies that the above result readily generalizes to the case with a Frobenius norm weighted by $\boldsymbol{\Gamma}_m$ and $\boldsymbol{\Gamma}_n$; the polar decomposition then becomes the weighted polar decomposition investigated in [57]. However, our results with $\tilde{\boldsymbol{P}} \neq \boldsymbol{I}$ in Theorem 2.1 indeed generalize the polar decomposition even further.

Corollary 2.4 *Given $\boldsymbol{U} \in \mathcal{M}_{m\times n}$ with full column rank and $\tilde{\boldsymbol{P}} \in \mathcal{M}_{n\times n}$ that is SPD, we have a unique polar-like decomposition*

$$\boldsymbol{U} = \tilde{\boldsymbol{U}}\boldsymbol{\Lambda} \tag{18}$$

such that $\tilde{\boldsymbol{U}} \in \mathcal{M}_{m\times n}$, $\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{P}}$ and $\boldsymbol{\Lambda} \in \mathcal{M}_{n\times n}$ is SPD. Moreover, this decomposition orthogonally projects $\boldsymbol{U}$ onto $\tilde{\boldsymbol{U}}$ in the sub-manifold $\mathcal{S}_{\tilde{\boldsymbol{P}}}$.

This corollary enriches the polar decomposition with the symmetry between $\boldsymbol{U}$ and $\tilde{\boldsymbol{U}}$. Note that this symmetry is preserved by our choice of scaling that simplifies (1)–(4). Should we scale $\tilde{\boldsymbol{P}}$ rather than $\boldsymbol{\Gamma}_n$ to $\boldsymbol{I}$, we would instead obtain the weighted polar decomposition in [57].

2.4 Complications due to degeneracy

In Sect. 2.1, we have restricted ourselves to the nice full-rank case with $m \geq n = \mathrm{rank}(\tilde{\boldsymbol{P}}) = \mathrm{rank}(\boldsymbol{U})$. In practice, sometimes this is not satisfied, such as in some

scenarios of data assimilation where the number m of samples of a random vector is smaller than the dimension n of the random vector itself (see Sect. 6). Hence, here we will reveal the implications of degeneracy due to any of these full-rank assumptions being violated. Before moving on, we introduce the notations $r = \mathrm{rank}(\boldsymbol{U})$ and $\tilde{r} = \mathrm{rank}(\tilde{\boldsymbol{P}})$.

Step 1 First of all, since $\tilde{r} \leq n$, when $m < n$, it is possible that $m < \tilde{r}$, in which case $\mathrm{rank}(\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}}) \leq m < \tilde{r}$, so there exists no $\tilde{\boldsymbol{U}} \in \mathcal{M}_{m\times n}$ whose second-moment matrix can match $\tilde{\boldsymbol{P}}$ due to rank deficiency. If $m < \tilde{r}$, an easiest fix is to find a best rank-m proxy to $\tilde{\boldsymbol{P}}$. For example, the leading rank-m truncation of the SVD of $\tilde{\boldsymbol{P}}$ (which is also the EVD here since $\tilde{\boldsymbol{P}}$ is semi-SPD) serves as the closest rank-m approximation under any unitarily invariant norm [28]. See [28, sec. 6] for a review on other variants of a nearest lower-rank approximation. With $\tilde{\boldsymbol{P}}$ replaced by such a proxy when $m < \tilde{r}$, we can always ensure $m \geq \tilde{r}$ for the modified problem.

Step 2 Next, given $m \geq \tilde{r}$, if $\tilde{r} < n$, a variable substitution by a projection from $\mathbb{R}^n$ onto $\mathrm{Row}(\tilde{\boldsymbol{P}})$ will eliminate the rank deficiency in $\tilde{\boldsymbol{P}}$. If we construct $\tilde{\boldsymbol{Z}}$ with its columns forming an orthonormal basis of $\mathrm{Row}(\tilde{\boldsymbol{P}})$, we have $\tilde{\boldsymbol{P}} = \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{P}}_{\tilde{r}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}$ such that $\tilde{\boldsymbol{P}}_{\tilde{r}} = \tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}} \in \mathcal{M}_{\tilde{r}\times\tilde{r}}$ has full rank. To have $\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{P}}$, we must require $\mathrm{Row}(\tilde{\boldsymbol{U}}) = \mathrm{Row}(\tilde{\boldsymbol{P}})$, so $\tilde{\boldsymbol{U}}$ can be uniquely represented by $\tilde{\boldsymbol{S}} \in \mathcal{M}_{m\times\tilde{r}}$ such that $\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{S}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}$. If we complete $\tilde{\boldsymbol{Z}}$ into an orthogonal $[\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{Z}}_\perp] \in \mathcal{M}_{n\times n}$, since

$$\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F}}^2 = \|(\tilde{\boldsymbol{U}} - \boldsymbol{U})[\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{Z}}_\perp]\|_{\mathrm{F}}^2 = \|\tilde{\boldsymbol{S}} - \boldsymbol{U}\tilde{\boldsymbol{Z}}\|_{\mathrm{F}}^2 + \|\boldsymbol{U}\tilde{\boldsymbol{Z}}_\perp\|_{\mathrm{F}}^2,$$

the original optimization problem (4) can be equivalently formulated as

$$\underset{\tilde{\boldsymbol{S}}\in\mathcal{M}_{m\times\tilde{r}}:\ \tilde{\boldsymbol{S}}^{\mathrm{T}}\tilde{\boldsymbol{S}}=\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}}{\arg\min} \|\tilde{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}}^2 \tag{19}$$

with $\boldsymbol{S} = \boldsymbol{U}\tilde{\boldsymbol{Z}}$. Since we already have $m \geq \tilde{r}$, with this further manipulation, we can always reduce a degenerate problem into one with $m \geq n = \mathrm{rank}(\tilde{\boldsymbol{P}})$.

Step 3 Finally, given $m \geq n = \mathrm{rank}(\tilde{\boldsymbol{P}})$, let's consider the case of $\mathrm{rank}(\boldsymbol{U}) = r < n$. If we construct $\boldsymbol{Z}$ with its columns forming an orthonormal basis of $\mathrm{Row}(\boldsymbol{U})$, $\boldsymbol{U}$ can be uniquely represented by a full-column-rank $\boldsymbol{W} \in \mathcal{M}_{m\times r}$ such that $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{Z}^{\mathrm{T}}$. Moreover, since $\boldsymbol{Z}$ can be augmented to an orthogonal $[\boldsymbol{Z}, \boldsymbol{Z}_\perp] \in \mathcal{M}_{n\times n}$, $\tilde{\boldsymbol{U}}$ can be uniquely represented by $\tilde{\boldsymbol{W}} = \tilde{\boldsymbol{U}}\boldsymbol{Z}$ and $\tilde{\boldsymbol{W}}_\perp = \tilde{\boldsymbol{U}}\boldsymbol{Z}_\perp$, which yield $\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{W}}\boldsymbol{Z}^{\mathrm{T}} + \tilde{\boldsymbol{W}}_\perp\boldsymbol{Z}_\perp^{\mathrm{T}}$ and

$$\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F}}^2 = \|(\tilde{\boldsymbol{U}} - \boldsymbol{U})[\boldsymbol{Z}, \boldsymbol{Z}_\perp]\|_{\mathrm{F}}^2 = \|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_{\mathrm{F}}^2 + \|\tilde{\boldsymbol{W}}_\perp\|_{\mathrm{F}}^2.$$

Note that $\|\tilde{\boldsymbol{W}}_\perp\|_{\mathrm{F}}^2 = \mathrm{tr}(\boldsymbol{Z}_\perp^{\mathrm{T}}\tilde{\boldsymbol{P}}\boldsymbol{Z}_\perp^{\mathrm{T}})$ is a constant thanks to the second-moment constraint. Therefore, the original problem (4) can be reduced to

$$\underset{\tilde{\boldsymbol{W}},\tilde{\boldsymbol{W}}_\perp:\ \tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}}=\tilde{\boldsymbol{P}}}{\arg\min} \|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_{\mathrm{F}}^2, \tag{20}$$

where $\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{P}}$ if and only if

$$\tilde{\boldsymbol{W}}^{\mathrm{T}}\tilde{\boldsymbol{W}} = \boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{P}}\boldsymbol{Z}, \qquad \tilde{\boldsymbol{W}}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\perp} = \boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{P}}\boldsymbol{Z}_{\perp}, \qquad \tilde{\boldsymbol{W}}_{\perp}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\perp} = \boldsymbol{Z}_{\perp}^{\mathrm{T}}\tilde{\boldsymbol{P}}\boldsymbol{Z}_{\perp}. \tag{21}$$

Note the subtlety that although the objective function depends on $\tilde{\boldsymbol{W}}$ only, it is not obvious whether we can decouple the optimization over $\tilde{\boldsymbol{W}}$ and that over $\tilde{\boldsymbol{W}}_{\perp}$, because it is possible that for the optimizer $\tilde{\boldsymbol{W}}_*$ to

$$\underset{\tilde{\boldsymbol{W}}\in\mathcal{M}_{m\times r}:\ \tilde{\boldsymbol{W}}^{\mathrm{T}}\tilde{\boldsymbol{W}}=\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{P}}\boldsymbol{Z}}{\arg\min} \|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_{\mathrm{F}}^2, \tag{22}$$

there exists no $\tilde{\boldsymbol{W}}_{\perp}$ such that the second and third constraint in (21) are satisfied. These two matrix constraints consist of $r(n-r)$ linear scalar equations and $(n-r)(n-r+1)/2$ quadratic ones, which amount to $(n-r)(n+r+1)/2$ independent scalar constraints. On the other hand, we have $(n-r)m$ degrees of freedom in $\tilde{\boldsymbol{W}}_{\perp}$. Therefore, under our assumption of $m \geq n \geq (r+1)$, we have $(n-r)(n+r+1)/2 \leq (n-r)n \leq (n-r)m$ so there should be enough degrees of freedom to ensure the existence of a valid $\tilde{\boldsymbol{W}}_{\perp}$ if all the constraints are compatible with each other. In the following, we will show that indeed for any $\tilde{\boldsymbol{W}}$ satisfying the first constraint in (21), there exists a valid $\tilde{\boldsymbol{W}}_{\perp}$ satisfying the other two. Moreover, we will characterize the set of all feasible $\tilde{\boldsymbol{W}}_{\perp}$'s. This justifies reducing (20)–(22), which is of exactly the same form as (4) and does satisfy all the full-rank assumptions we have made for (4).

To identify all the solutions satisfying the last two constraints in (21), first we introduce the notation

$$\begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{12}^{\mathrm{T}} & \boldsymbol{B}_{22} \end{bmatrix} \triangleq \begin{bmatrix} \tilde{\boldsymbol{W}}^{\mathrm{T}} \\ \tilde{\boldsymbol{W}}_{\perp}^{\mathrm{T}} \end{bmatrix} \left[\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{W}}_{\perp}\right] = \begin{bmatrix} \boldsymbol{Z}^{\mathrm{T}} \\ \boldsymbol{Z}_{\perp}^{\mathrm{T}} \end{bmatrix} \tilde{\boldsymbol{P}} \left[\boldsymbol{Z}, \boldsymbol{Z}_{\perp}\right].$$

The general solution to the linear constraint $\tilde{\boldsymbol{W}}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\perp} = \boldsymbol{B}_{12}$ can be written as $\tilde{\boldsymbol{W}}_{\perp} = (\tilde{\boldsymbol{W}}^{\mathrm{T}})^{+}\boldsymbol{B}_{12} + \tilde{\boldsymbol{W}}_{\mathrm{c}}\boldsymbol{V}$ with "$+$" denoting the Moore-Penrose pseudo-inverse (a.k.a. the generalized inverse), the columns of $\tilde{\boldsymbol{W}}_{\mathrm{c}} \in \mathcal{M}_{m\times(m-r)}$ forming an orthonormal basis of $\mathrm{Col}(\tilde{\boldsymbol{W}})^{\perp}$, and $\boldsymbol{V} \in \mathcal{M}_{(m-r)\times(n-r)}$ arbitrary. Since $\boldsymbol{B}_{11}$ has full rank r, $\tilde{\boldsymbol{W}}$ has full column rank r and thus $\tilde{\boldsymbol{W}}^{+} = (\tilde{\boldsymbol{W}}^{\mathrm{T}}\tilde{\boldsymbol{W}})^{-1}\tilde{\boldsymbol{W}}^{\mathrm{T}} = \boldsymbol{B}_{11}^{-1}\tilde{\boldsymbol{W}}^{\mathrm{T}}$ and $(\tilde{\boldsymbol{W}}^{\mathrm{T}})^{+} = (\tilde{\boldsymbol{W}}^{+})^{\mathrm{T}} = \tilde{\boldsymbol{W}}\boldsymbol{B}_{11}^{-1}$. Hence, $\tilde{\boldsymbol{W}}_{\perp} = \tilde{\boldsymbol{W}}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12} + \tilde{\boldsymbol{W}}_{\mathrm{c}}\boldsymbol{V}$. Now the quadratic constraint $\tilde{\boldsymbol{W}}_{\perp}^{\mathrm{T}}\tilde{\boldsymbol{W}}_{\perp} = \boldsymbol{B}_{22}$ is equivalent to $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^{\mathrm{T}}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}$, where the right hand side is an SPD Schur complement. A valid $\boldsymbol{V}$ exists if and only if $(m-r) \geq (n-r) \geq 1$. Therefore, the above analysis confirms our previous speculation based on comparing the number of degrees of freedom with the number of constraints. Besides, when a valid $\boldsymbol{V}$ exists, it is not unique.

We summarize the above three steps in the following theorem.

Theorem 2.2 (Degenerate counterpart of Theorem 2.1 *Given* $\boldsymbol{U} \in \mathcal{M}_{m\times n}$ *with* $\mathrm{rank}(\boldsymbol{U}) = r$ *and* $\tilde{\boldsymbol{P}} \in \mathcal{M}_{n\times n}$ *with* $\mathrm{rank}(\tilde{\boldsymbol{P}}) = \tilde{r} \leq m$, *we have*

$$\underset{\tilde{\boldsymbol{U}}\in\mathcal{M}_{m\times n}:\ \tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}}=\tilde{\boldsymbol{P}}}{\arg\min} \|\tilde{\boldsymbol{U}}-\boldsymbol{U}\|_{\mathrm{F}}^2=\tilde{\boldsymbol{U}}_*=\tilde{\boldsymbol{S}}_*\tilde{\boldsymbol{Z}}^{\mathrm{T}}=(\tilde{\boldsymbol{W}}_*\boldsymbol{Z}^{\mathrm{T}}+\tilde{\boldsymbol{W}}_{\perp *}\boldsymbol{Z}_{\perp}^{\mathrm{T}})\tilde{\boldsymbol{Z}}^{\mathrm{T}}, \quad (23)$$

where $\tilde{\boldsymbol{S}}_*=\tilde{\boldsymbol{W}}_*\boldsymbol{Z}^{\mathrm{T}}+\tilde{\boldsymbol{W}}_{\perp *}\boldsymbol{Z}_{\perp}^{\mathrm{T}}\in\mathcal{M}_{m\times\tilde{r}}$. *Moreover, the columns of* $\tilde{\boldsymbol{Z}}\in\mathcal{M}_{n\times\tilde{r}}$ *form an orthonormal basis of* $\mathrm{Row}(\tilde{\boldsymbol{P}})$ *and the columns of* $\boldsymbol{Z}\in\mathcal{M}_{\tilde{r}\times r'}$ *form an orthonormal basis of* $\mathrm{Row}(\boldsymbol{U}\tilde{\boldsymbol{Z}})$ *with*

$$r'=\tilde{r}-\dim(\mathrm{Row}(\tilde{\boldsymbol{P}})\cap\mathrm{Row}(\boldsymbol{U})^{\perp})=r-\dim(\mathrm{Row}(\boldsymbol{U})\cap\mathrm{Row}(\tilde{\boldsymbol{P}})^{\perp}). \quad (24)$$

Finally, $\tilde{\boldsymbol{W}}_*$ *and* $\tilde{\boldsymbol{W}}_{\perp *}$ *are given by*

$$\tilde{\boldsymbol{W}}_*=\underset{\tilde{\boldsymbol{W}}\in\mathcal{M}_{m\times r'}:\ \tilde{\boldsymbol{W}}^{\mathrm{T}}\tilde{\boldsymbol{W}}=\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}\boldsymbol{Z}}{\arg\min} \|\tilde{\boldsymbol{W}}-\boldsymbol{U}\tilde{\boldsymbol{Z}}\boldsymbol{Z}\|_{\mathrm{F}}^2 \quad (25)$$

and $\tilde{\boldsymbol{W}}_{\perp *}=\tilde{\boldsymbol{W}}_*\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}+\tilde{\boldsymbol{W}}_{\mathrm{c}*}\boldsymbol{V}_*\in\mathcal{M}_{m\times(\tilde{r}-r')}$ *with the columns of* $\tilde{\boldsymbol{W}}_{\mathrm{c}*}\in\mathcal{M}_{m\times(m-r')}$ *forming an orthonormal basis of* $\mathrm{Col}(\tilde{\boldsymbol{W}}_*)^{\perp}$ *and* $\boldsymbol{V}_*\in\mathcal{M}_{(m-r')\times(\tilde{r}-r')}$ *arbitrary as long as* $\boldsymbol{V}_*^{\mathrm{T}}\boldsymbol{V}_*=\boldsymbol{B}_{22}-\boldsymbol{B}_{12}^{\mathrm{T}}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}$. *The* $\boldsymbol{B}_{ij}$ *blocks are defined as:*

$$\begin{bmatrix}\boldsymbol{B}_{11} & \boldsymbol{B}_{12}\\ \boldsymbol{B}_{12}^{\mathrm{T}} & \boldsymbol{B}_{22}\end{bmatrix}\triangleq\begin{bmatrix}\tilde{\boldsymbol{W}}^{\mathrm{T}}\\ \tilde{\boldsymbol{W}}_{\perp}^{\mathrm{T}}\end{bmatrix}\left[\tilde{\boldsymbol{W}},\tilde{\boldsymbol{W}}_{\perp}\right]=\begin{bmatrix}\boldsymbol{Z}^{\mathrm{T}}\\ \boldsymbol{Z}_{\perp}^{\mathrm{T}}\end{bmatrix}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}\left[\boldsymbol{Z},\boldsymbol{Z}_{\perp}\right].$$

Besides, the global minimum can be expressed in exactly the same way as in (6)*:*

$$\|\tilde{\boldsymbol{U}}_*-\boldsymbol{U}\|_{\mathrm{F}}^2=\mathrm{tr}[\boldsymbol{P}+\tilde{\boldsymbol{P}}-2(\boldsymbol{P}\tilde{\boldsymbol{P}})^{1/2}]. \quad (26)$$

Proof Everything other than (26) ensues readily from knitting into one piece step 2 and step 3 shown right before we state this theorem. To prove (26), notice that

$$\|\tilde{\boldsymbol{U}}_*-\boldsymbol{U}\|_{\mathrm{F}}^2-\mathrm{tr}(\boldsymbol{P})-\mathrm{tr}(\tilde{\boldsymbol{P}})=-2\mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{U}}_*)=-2\mathrm{tr}((\boldsymbol{U}\tilde{\boldsymbol{Z}})^{\mathrm{T}}\tilde{\boldsymbol{S}}_*)=-2\mathrm{tr}((\boldsymbol{U}\tilde{\boldsymbol{Z}}\boldsymbol{Z})^{\mathrm{T}}\tilde{\boldsymbol{W}}_*).$$

Since Theorem 2.1 applies to (25), the last expression in the above can be computed by (6) to yield $\|\tilde{\boldsymbol{U}}_*-\boldsymbol{U}\|_{\mathrm{F}}^2=\mathrm{tr}(\boldsymbol{P})+\mathrm{tr}(\tilde{\boldsymbol{P}})-2\mathrm{tr}[(\boldsymbol{P}_{r'}\tilde{\boldsymbol{P}}_{r'})^{1/2}]$, where $\boldsymbol{P}_{r'}=\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{P}\tilde{\boldsymbol{Z}}\boldsymbol{Z}$ and $\tilde{\boldsymbol{P}}_{r'}=\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}\boldsymbol{Z}$. Hence, what remains to be shown is $\mathrm{tr}[(\boldsymbol{P}_{r'}\tilde{\boldsymbol{P}}_{r'})^{1/2}]=\mathrm{tr}[(\boldsymbol{P}\tilde{\boldsymbol{P}})^{1/2}]$ and it suffices to show that $\boldsymbol{P}_{r'}\tilde{\boldsymbol{P}}_{r'}$ and $\boldsymbol{P}\tilde{\boldsymbol{P}}$ share the same nonzero eigenvalues with the same algebraic multiplicity. To demonstrate this, first compute

$$\boldsymbol{P}_{r'}\tilde{\boldsymbol{P}}_{r'}=\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{P}\tilde{\boldsymbol{Z}}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{P}})\tilde{\boldsymbol{Z}}\boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{P}(\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{P}})\tilde{\boldsymbol{Z}}\boldsymbol{Z}=\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{P}\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}\boldsymbol{Z}$$

using the property of the orthogonal projections $\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$ and $\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}$. Next, notice that $(\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{P})(\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}\boldsymbol{Z})$ and $(\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}\boldsymbol{Z})(\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{P})$ have identical nonzero eigenvalues with the same algebraic multiplicity. Since $(\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}\boldsymbol{Z})(\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{P})=\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{P}=\tilde{\boldsymbol{P}}\boldsymbol{P}$, we conclude that the nonzero spectrum of $\boldsymbol{P}_{r'}\tilde{\boldsymbol{P}}_{r'}$ and of $\boldsymbol{P}\tilde{\boldsymbol{P}}$ coincide in terms of eigenvalues and their algebraic multiplicity, which completes the proof. □

Remark 1 $\tilde{\boldsymbol{U}}_*$ is unique if and only if $r' = \tilde{r}$, i.e. $\mathrm{Row}(\tilde{\boldsymbol{P}}) \cap \mathrm{Row}(\boldsymbol{U})^\perp = 0$. In this case, step 3 is not needed. Given $r' = \tilde{r}$, the reverse problem seeking the closest point to $\tilde{\boldsymbol{U}}_*$ with second-moment matrix $\boldsymbol{P}$ will also have a unique solution if and only if $r' = r$, i.e. $\mathrm{Row}(\boldsymbol{U}) \cap \mathrm{Row}(\tilde{\boldsymbol{P}})^\perp = 0$. Therefore, the elements in $\mathcal{S}_{\boldsymbol{P}} = \{\boldsymbol{U} \in \mathcal{M}_{m\times n} : \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{P}\}$ and those in $\mathcal{S}_{\tilde{\boldsymbol{P}}}$ can be paired up bijectively and unambiguously by this closest-point correspondence if and only if $\mathrm{rank}(\boldsymbol{P}) = \mathrm{rank}(\tilde{\boldsymbol{P}}) = \mathrm{rank}(\boldsymbol{P}\,\tilde{\boldsymbol{P}})$. In practice, one common special case of such is when $\mathrm{Row}(\boldsymbol{U}) = \mathrm{Row}(\tilde{\boldsymbol{P}})$, as will be seen in Sect. 6. If this is true, besides fixing rank deficiency, step 2 also preserves the value of the objective function, i.e. $\|\tilde{\boldsymbol{S}} - \boldsymbol{S}\|_{\mathrm{F}}^2 = \|\tilde{\boldsymbol{U}}\tilde{\boldsymbol{Z}} - \boldsymbol{U}\tilde{\boldsymbol{Z}}\|_{\mathrm{F}}^2 \equiv \|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F}}^2$.

Remark 2 One extreme case of rank deficiency is when $\mathrm{Row}(\tilde{\boldsymbol{P}}) \perp \mathrm{Row}(\boldsymbol{U})$ and thus $r' = 0$. This renders the optimization (23) trivial because the objective function is now constant: $\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F}}^2 \equiv \mathrm{tr}(\boldsymbol{P} + \tilde{\boldsymbol{P}})$, so all feasible $\tilde{\boldsymbol{U}}$'s are equally good.

Remark 3 Although the degeneracy significantly complicates the characterization of the optimizer $\tilde{\boldsymbol{U}}_*$, the minimum of the objective function (26) surprisingly shares the same simple formula as (6) in the non-degenerate case.

3 Algorithms for minimum-correction second-moment matching

3.1 General computation strategies

To compute $\tilde{\boldsymbol{U}}_*$ in Theorem 2.1, there are two routes. We can either compute $\tilde{\boldsymbol{U}}_* = \boldsymbol{U}\boldsymbol{A}_*$ using

$$\boldsymbol{A}_* = \sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}(\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}})^{-1/2}\sqrt{\tilde{\boldsymbol{P}}} = \sqrt{\boldsymbol{P}}^{-1}(\sqrt{\boldsymbol{P}}\tilde{\boldsymbol{P}}\sqrt{\boldsymbol{P}}^{\mathrm{T}})^{1/2}\sqrt{\boldsymbol{P}}^{-\mathrm{T}} \tag{27}$$

or directly obtain $\tilde{\boldsymbol{U}}_*$ without explicitly forming $\boldsymbol{P}$ or computing $\boldsymbol{A}_*$.

Computing $\tilde{\boldsymbol{U}}_*$ through $\boldsymbol{A}_*$ To take the $\boldsymbol{A}_*$ route, we need a pre-processing step $\boldsymbol{P} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$, a post-processing step $\tilde{\boldsymbol{U}}_* = \boldsymbol{U}\boldsymbol{A}_*$ and the key step (27). The drawback is that forming $\boldsymbol{P} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ indeed squares the condition number, i.e. $\kappa(\boldsymbol{P}) = \kappa(\boldsymbol{U})^2$. Therefore, if $\boldsymbol{U}$ is ill-conditioned, the non-$\boldsymbol{A}_*$ route is preferred.

Computing $\tilde{\boldsymbol{U}}_*$ without $\boldsymbol{A}_*$ The non-$\boldsymbol{A}_*$ route is based on directly computing the $\mathcal{S}_{\tilde{\boldsymbol{P}}}$ polar factor $\tilde{\boldsymbol{U}}_*$ of the generalized PD (polar decomposition) $\boldsymbol{U} = \tilde{\boldsymbol{U}}_*\boldsymbol{C}$ introduced in Corollary 2.4. First, we need to reduce the rectangular PD to a square one by identifying an orthonormal basis of $\boldsymbol{U}$'s columns, i.e. $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{R}$ with $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{R} \in \mathcal{M}_{n\times n}$. This can be achieved by a Householder QR factorization [51, ch. 10]. Next we reduce computing the $\mathcal{S}_{\tilde{\boldsymbol{P}}}$ factor $\tilde{\boldsymbol{R}}_*$ of the generalized PD $\boldsymbol{R} = \tilde{\boldsymbol{R}}_*\boldsymbol{C}$ to computing the orthogonal polar factor $\boldsymbol{Q}$ of the PD $\boldsymbol{R}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}} = \boldsymbol{Q}\boldsymbol{S} = (\tilde{\boldsymbol{R}}_*\sqrt{\tilde{\boldsymbol{P}}}^{-1})(\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{C}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}})$. Finally, we can assemble the desired $\tilde{\boldsymbol{U}}_*$ as $\tilde{\boldsymbol{U}}_* = \boldsymbol{V}\tilde{\boldsymbol{R}}_* = \boldsymbol{V}\boldsymbol{Q}\sqrt{\tilde{\boldsymbol{P}}}$.

To summarize, the $\boldsymbol{A}_*$ route is eventually reduced to computing an SPD square root, while the non-$\boldsymbol{A}_*$ route is reduced to performing a polar decomposition. These two tasks are indeed intimately related and an algorithm for one task usually corresponds to one for the other, as will be seen in the next subsection.

In the above, we have assumed the problem is non-degenerate, as required by Theorem 2.1. If that is not the case, the constructive proof of Theorem 2.2 already provides an algorithm to reduce a degenerate case to a non-degenerate one.

3.2 Algorithms for the SPD square root and polar decomposition

Given an invertible $\boldsymbol{R} \in \mathcal{M}_{n\times n}$ with polar decomposition $\boldsymbol{R} = \boldsymbol{Q}\boldsymbol{C}$, where $\boldsymbol{Q}$ and $\boldsymbol{C}$ are the orthogonal and SPD polar factor, respectively, then $\boldsymbol{C}$ is the SPD square root of $\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}$, i.e. $\boldsymbol{C} = (\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R})^{1/2}$. On the other hand, given an SPD $\boldsymbol{P} \in \mathcal{M}_{n\times n}$ with an arbitrary square root $\sqrt{\boldsymbol{P}}$, $\boldsymbol{P}^{1/2}$ will also be the SPD polar factor of $\sqrt{\boldsymbol{P}}$. Therefore, computing an SPD square root and performing a polar decomposition can be reduced to each other in exactly the same vein as how the Cholesky decomposition is related to the QR factorization.

3.2.1 Algorithms based on SVD/EVD

For the polar decomposition of $\boldsymbol{R} \in \mathcal{M}_{n\times n}$, a straightforward SVD-based algorithm stems from the fact that the SVD of $\boldsymbol{R}$ can be readily manipulated to yield the polar decomposition:

$$\boldsymbol{R} = \boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} = (\boldsymbol{Q}\boldsymbol{V}^{\mathrm{T}})(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}).$$

This SVD-based algorithm corresponds to using the EVD (eigenvalue decomposition) to obtain the SPD square root of an SPD $\boldsymbol{P} \in \mathcal{M}_{n\times n}$ by $\boldsymbol{P} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} = (\boldsymbol{V}\boldsymbol{\Lambda}^{1/2}\boldsymbol{V}^{\mathrm{T}})^2$.

Since we have efficient and numerically stable algorithms for SVD and EVD [20, 51], the above SVD/EVD-based algorithms are easy to implement and robust to use. Despite this, there are cases where a non-SVD/EVD-based algorithm is desirable. It might be because, for example, we have no access to an efficient SVD/EVD procedure, we want to have more control over the accuracy of the result, or we want to utilize some special features particular to a problem at hand to further boost the efficiency. A non-SVD/EVD-based algorithm typically involves explicitly carrying out a matrix fixed-point iteration, as will be discussed next.

3.2.2 Algorithms based on fixed-point iterations

For computing the SPD square root and polar decomposition, there are also non-SVD/EVD-based algorithms that take the form of a matrix fixed-point iteration $\boldsymbol{X}_{k+1} = \boldsymbol{F}(\boldsymbol{X}_k)$ with $\boldsymbol{F} : \mathcal{M}_{n\times n} \to \mathcal{M}_{n\times n}$ involving only matrix addition, multiplication, and inversion.

If we view an iteration as a discrete dynamical system[3], the minimal requirements for a matrix iteration $\boldsymbol{X}_{k+1} = \boldsymbol{F}(\boldsymbol{X}_k)$ to be usable in practice for seeking a target matrix $\boldsymbol{X}_*$ include:

1. $\boldsymbol{X}_*$ is a neutrally stable fixed point of $\boldsymbol{F}$.

[3] See [48, sec. 5.1] for the definition of terms related to dynamical systems.

2. A subset $\mathcal{A}_{\text{ini}}$ of the attraction basin $\mathcal{A}_{\boldsymbol{F},\boldsymbol{X}_*} \subset \mathcal{M}_{n\times n}$ an be identified.

Here by definition, the attraction basin contains all the points that will converge to $\boldsymbol{X}_*$ under $\boldsymbol{F}$, i.e. $\boldsymbol{X} \in \mathcal{A}_{\boldsymbol{F},\boldsymbol{X}_*}$ if and only if $\boldsymbol{X}_k \to \boldsymbol{X}_*$ with $\boldsymbol{X}_0 = \boldsymbol{X}$ and $\boldsymbol{X}_{k+1} = \boldsymbol{F}(\boldsymbol{X}_k)$. If these two requirements are met, for any initial guess $\boldsymbol{X}_0 \in \mathcal{A}_{\text{ini}}$, we have $\boldsymbol{X}_k \to \boldsymbol{X}_*$ analytically and the neutral stability of $\boldsymbol{X}_*$ guarantees that any numerical errors incurred in carrying out $\boldsymbol{X}_{k+1} = \boldsymbol{F}(\boldsymbol{X}_k)$ will stay bounded once $\boldsymbol{X}_k$ approaches $\boldsymbol{X}_*$, i.e. entering some ε-neighborhood of $\boldsymbol{X}_*$.

As we shall see, usually we can find $\mathcal{A}_{\text{ini}}$ as a stable manifold containing $\boldsymbol{X}_*$ and invariant under $\boldsymbol{F}$. Sometimes $\mathcal{A}_{\text{ini}}$ can even extend to infinity. However, this should not be confused with global convergence and it does not even guarantee the numerical stability of the algorithm because $\mathcal{A}_{\text{ini}}$ as a submanifold of $\mathcal{M}_{n\times n}$ is a boundary set, i.e. it has no interior point. Hence, the numerical errors $(\hat{\boldsymbol{X}}_k - \boldsymbol{X}_k)$, although typically small in magnitude but in general unconstrained in structure, will drive the actual iterates $\hat{\boldsymbol{X}}_k$'s away from $\mathcal{A}_{\text{ini}}$ and might make them fail to converge to anything close to $\boldsymbol{X}_*$. This is why we need the neutral stability of $\boldsymbol{X}_*$ to ensure that if the iterates manage to approach $\boldsymbol{X}_*$ before the numerical errors accumulate significantly, these errors will remain bounded hereafter.

Since the dynamical system $\boldsymbol{X}_{k+1} = \boldsymbol{F}(\boldsymbol{X}_k)$ is n^2-dimensional, we can reshape each $\boldsymbol{X}_k$ into an $\mathbb{R}^{n\times n}$ vector and view $\boldsymbol{F}$ as an operator from $\mathbb{R}^{n\times n}$ to itself. Denote the Jacobian matrix of $\boldsymbol{F}$ at $\boldsymbol{X}_*$ by $\mathrm{J}\boldsymbol{F} \in \mathcal{M}_{n^2\times n^2}$[4]. Then a necessary condition for $\boldsymbol{X}_*$ to be neutrally stable is $\rho(\mathrm{J}\boldsymbol{F}) \le 1$, where $\rho(\cdot)$ denotes the spectral radius, while the strict inequality $\rho(\mathrm{J}\boldsymbol{F}) < 1$ is a sufficient one. However, viewing $\boldsymbol{X}_k$ as a vector typically makes the expression of $\boldsymbol{F}$ awkwardly complicated and renders cumbersome the analysis of $\mathrm{J}\boldsymbol{F}$'s spectrum. [43] bypasses this difficulty elegantly by keeping $\boldsymbol{X}_k$ as a matrix and taking a functional viewpoint of $\boldsymbol{X}_{k+1} = \boldsymbol{F}(\boldsymbol{X}_k)$. From this perspective, the counterpart of $\mathrm{J}\boldsymbol{F}$ is the Fréchet derivative $\mathrm{D}\boldsymbol{F}$ of the operator $\boldsymbol{F}$ at $\boldsymbol{X}_*$. Here $\mathrm{D}\boldsymbol{F} : \mathcal{M}_{n\times n} \to \mathcal{M}_{n\times n}$ is a linear operator. Since $\mathrm{J}\boldsymbol{F}$ and $\mathrm{D}\boldsymbol{F}$ are essentially the same linear operator, they share the same spectrum. It turns out that identifying all the eigenvalues and eigenvectors of $\mathrm{D}\boldsymbol{F}$ by direct observation can be made possible by suitable linear transforms.

Next, we review the most common iterations for computing the i) SPD square root of an SPD matrix and ii) orthogonal polar factor of a square matrix. To do so, we analyze and compare iterations in a novel unified framework. We also obtain new insights into their properties and propose a slight modification that surprisingly remedies some iterations deemed numerically unstable previously.

(i) A unified framework for four SPD square root Iterations

Given an SPD $\boldsymbol{P}$, the following four iterations can be used to compute the SPD square root $\boldsymbol{P}^{1/2}$ and $\boldsymbol{P}^{-1/2}$ (see [27,43] and [20, sec. 9.4.2]):

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \Delta\boldsymbol{X}_k, \quad \boldsymbol{X}_k\Delta\boldsymbol{X}_k + \Delta\boldsymbol{X}_k\boldsymbol{X}_k = \boldsymbol{P} - \boldsymbol{X}_k^2; \tag{28}$$

$$\boldsymbol{X}_{k+1} = (1/2)(\boldsymbol{X}_k + \boldsymbol{X}_k^{-1}\boldsymbol{P}); \tag{29}$$

$$\boldsymbol{X}_{k+1} = (1/2)(\boldsymbol{X}_k + \boldsymbol{Y}_k^{-1}), \quad \boldsymbol{Y}_{k+1} = (1/2)(\boldsymbol{Y}_k + \boldsymbol{X}_k^{-1}); \tag{30}$$

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + (1/2)\boldsymbol{X}_k(\boldsymbol{I} - \boldsymbol{X}_k\boldsymbol{P}\boldsymbol{X}_k). \tag{31}$$

[4] The explicit dependence on $\boldsymbol{X}_*$ is omitted in the notation since it should be clear from context.

Among these iterations, (28) stems from a direct application of the Newton's method to the matrix equation $\boldsymbol{G}(\boldsymbol{X}) = \boldsymbol{X}^2 - \boldsymbol{P} = \boldsymbol{0}$ with the Fréchet derivative of $\boldsymbol{G}$ at $\boldsymbol{X}_k$ being the linear operator $\mathrm{D}\boldsymbol{G}_{\boldsymbol{X}_k}\boldsymbol{X} = \boldsymbol{X}_k\boldsymbol{X} + \boldsymbol{X}\boldsymbol{X}_k$. (29) is a simplification of (28) by making the assumption $\boldsymbol{X}_k\Delta\boldsymbol{X}_k = \Delta\boldsymbol{X}_k\boldsymbol{X}_k$, which is true for certain choices of $\boldsymbol{X}_0$ as we shall see. (30) is an extension of (29) by coupling (29) with a mirror iteration to compute $\boldsymbol{P}^{1/2}$ and $\boldsymbol{P}^{-1/2}$ simultaneously. Finally, (31) is quite different from the others, as it comes from mimicking the Newton iteration for the scalar equation $g(x) = 1/x^2 - \lambda = 0$ [43].

Lemma 3.1 *The converging point and a corresponding $\mathcal{A}_{\text{ini}}$ of feasible initial values for iterations Eqs. (28)–(31) are summarized as*

$$(28), (29) : \boldsymbol{X}_k \to \boldsymbol{P}^{1/2} \quad \text{for } \boldsymbol{X}_0 \in \mathcal{C}_{\boldsymbol{P}} \triangleq \{\boldsymbol{X} \text{ SPD} : \boldsymbol{X}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{X}\}; \tag{32}$$

$$(30) : \boldsymbol{X}_k \to \boldsymbol{P}^{1/2}, \quad \boldsymbol{Y}_k \to \boldsymbol{P}^{-1/2} \quad \text{for } \boldsymbol{X}_0 \in \mathcal{C}_{\boldsymbol{P}}, \quad \boldsymbol{Y}_0 = \boldsymbol{P}^{-1}\boldsymbol{X}_0; \tag{33}$$

$$(31) : \boldsymbol{X}_k \to \boldsymbol{P}^{-1/2} \quad \text{for } \boldsymbol{X}_0 \in \{\boldsymbol{X} \in \mathcal{C}_{\boldsymbol{P}} : \|\boldsymbol{X}\|_2 < \sqrt{3/\|\boldsymbol{P}\|_2}\}. \tag{34}$$

Proof Observe that $\boldsymbol{X}_k \in \mathcal{C}_{\boldsymbol{P}}$ (and $\boldsymbol{Y}_k \in \mathcal{C}_{\boldsymbol{P}}$ for (30)) implies $\boldsymbol{X}_{k+1} \in \mathcal{C}_{\boldsymbol{P}}$ (and $\boldsymbol{Y}_{k+1} \in \mathcal{C}_{\boldsymbol{P}}$), so $\mathcal{C}_{\boldsymbol{P}}$ is an invariant submanifold under these iterations. Therefore, there exists an EVD $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}$ that also diagonalizes the iterates $\boldsymbol{X}_k = \boldsymbol{Q}\boldsymbol{\Lambda}_k\boldsymbol{Q}^{\mathrm{T}}$ (and $\boldsymbol{Y}_k = \boldsymbol{Q}\boldsymbol{\Sigma}_k\boldsymbol{Q}^{\mathrm{T}}$) and the matrix iteration can be decoupled into n scalar ones of the same form:

$$(28), (29) : \lambda_{k+1} = (1/2)(\lambda_k + \lambda/\lambda_k) : \ \lambda_k \to \sqrt{\lambda} \quad \text{for } \lambda_0 > 0 \tag{35}$$

$$(30) : \lambda_{k+1} = (1/2)(\lambda_k + 1/\sigma_k), \quad \sigma_{k+1} = (1/2)(\sigma_k + 1/\lambda_k) : \tag{36}$$

$$\lambda_k \to \sqrt{\lambda_0/\sigma_0}, \quad \sigma_k \to \sqrt{\sigma_0/\lambda_0} \quad \text{for } \lambda_0, \sigma_0 > 0$$

$$(31) : \lambda_{k+1} = (1/2)\lambda_k(3 - \lambda_k^2\lambda) : \ \lambda_k \to 1/\sqrt{\lambda} \quad \text{for } 0 < \lambda_0 < \sqrt{3/\lambda} \tag{37}$$

Here $\lambda > 0$ is any diagonal entry of $\boldsymbol{\Lambda}$, which is also an eigenvalue of $\boldsymbol{P}$. □

Note that (36) can be reduced to two decoupled iterations of the same form as (35) due to the observation that $\lambda_k/\sigma_k \equiv \lambda_0/\sigma_0$. Since (35) is nothing but the well-known Newton square-root iteration that solves $g(x) = x^2 - \lambda = 0$ while (37) is the Newton iteration for solving $g(x) = 1/x^2 - \lambda = 0$ [43], we know both scalar iterations converge quadratically. Furthermore, this implies that the matrix iterations Eqs. (28)–(31) also converge quadratically provided their initial values satisfy Eqs. (32)–(34).

Lemma 3.1 only guarantees the analytic convergence when roundoff errors are absent. To ensure the stability of an iteration $\boldsymbol{X}_{k+1} = \boldsymbol{F}(\boldsymbol{X}_k)$, we require the Fréchet derivative $\mathrm{D}\boldsymbol{F}$ of the operator $\boldsymbol{F}$ at the limit $\boldsymbol{X}_*$ to have spectral radius $\rho(\mathrm{D}\boldsymbol{F}) \leq 1$.

Lemma 3.2 *The Fréchet derivatives for Eqs. (28)–(31) are summarized as*

$$(28) : \mathrm{D}\boldsymbol{F}\boldsymbol{X} = \boldsymbol{0}; \tag{38}$$

$$(29), (31) : \mathrm{D}\boldsymbol{F}\boldsymbol{X} = (1/2)(\boldsymbol{X} - \boldsymbol{\Lambda}^{-1/2}\boldsymbol{X}\boldsymbol{\Lambda}^{1/2}); \tag{39}$$

$$(30) : \mathrm{D}\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{Y}) = (1/2)(\boldsymbol{X} - \boldsymbol{\Lambda}^{1/2}\boldsymbol{Y}\boldsymbol{\Lambda}^{1/2}, \boldsymbol{Y} - \boldsymbol{\Lambda}^{-1/2}\boldsymbol{X}\boldsymbol{\Lambda}^{-1/2}). \tag{40}$$

These linear operators have eigenvectors $\mathbf{Z}_{ij}$'s and eigenvalues μ_{ij}'s as below:

$$(39): \mathbf{Z}_{ij} = \boldsymbol{E}_{ij}, \quad \mu_{ij} = (1/2)(1 - \sqrt{\lambda_j/\lambda_i}); \tag{41}$$

$$(40): \mathbf{Z}_{ij} = (\boldsymbol{E}_{ij}, 1/\sqrt{\lambda_i\lambda_j}\boldsymbol{E}_{ij}), \quad \mu_{ij} = 0,$$

$$\mathbf{Z}'_{ij} = (\boldsymbol{E}_{ij}, -1/\sqrt{\lambda_i\lambda_j}\boldsymbol{E}_{ij}), \quad \mu'_{ij} = 1. \tag{42}$$

Here $\boldsymbol{E}_{ij} \in \mathcal{M}_{n\times n}$ is the matrix having its (i, j)-th entry equal to 1 as the only nonzero entry.

Proof (38) is a standard result for an exact Newton iteration, which must have $\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{X} - (\mathrm{D}\boldsymbol{G}_{\boldsymbol{X}})^{-1}\boldsymbol{G}(\boldsymbol{X})$. Hence, $(\mathrm{D}\boldsymbol{F}_{\boldsymbol{X}_*})\boldsymbol{X} = \boldsymbol{X} - (\mathrm{D}\boldsymbol{G}_{\boldsymbol{X}_*})^{-1}(\mathrm{D}\boldsymbol{G}_{\boldsymbol{X}_*})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} \equiv \boldsymbol{0}$. (39) and (40) are obtained by diagonalizing $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}$ with a linear variable substitution $\boldsymbol{X}' = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{Q}$. For example, for (39), originally we have $\boldsymbol{Y} = \mathrm{D}\boldsymbol{F}\boldsymbol{X} = (1/2)(\boldsymbol{X} - \boldsymbol{P}^{-1/2}\boldsymbol{X}\boldsymbol{P}^{1/2})$. After the variable substitution, $\boldsymbol{Y} = \mathrm{D}\boldsymbol{F}\boldsymbol{X}$ becomes $\boldsymbol{Y}' = \boldsymbol{Q}^{\mathrm{T}}(\mathrm{D}\boldsymbol{F}(\boldsymbol{Q}\boldsymbol{X}'\boldsymbol{Q}^{\mathrm{T}}))\boldsymbol{Q} = (1/2)(\boldsymbol{X}' - \boldsymbol{\Lambda}^{-1/2}\boldsymbol{X}'\boldsymbol{\Lambda}^{1/2})$, which yields (39). This simplification can be justified by the fact that the spectrum of $\mathrm{D}\boldsymbol{F}$ is invariant under a linear variable substitution. Note that $\boldsymbol{X}'$ is not diagonal in general because in stability analysis, we can no longer assume $\boldsymbol{X} \in \mathcal{C}_{\boldsymbol{P}}$.

Since each term in (39) and (40) has only one non-diagonal matrix factor, we can readily identify all the eigenvectors $\mathbf{Z}_{ij}$'s and eigenvalues μ_{ij}'s by observation. □

Therefore, if we want $\rho(\mathrm{D}\boldsymbol{F}) = (1/2)(\sqrt{\kappa(\boldsymbol{P})} - 1) < 1$ for (41), we need the 2-norm condition number $\kappa(\boldsymbol{P})$ to satisfy $\kappa(\boldsymbol{P}) < 9$. This is very unsatisfactory because it puts a stringent constraint on the $\boldsymbol{P}$'s to which (29) and (31) are applicable. [43] shows that for (31), a manual symmetrification step in the end of each iteration can alleviate this stability constraint to $\kappa(\boldsymbol{P}) < 17+6\sqrt{8}$, which is only a nonessential and minor improvement. This makes sense because $\mathcal{C}_{\boldsymbol{P}}$ requires being commutative with $\boldsymbol{P}$ besides being SPD, so symmetrification itself does not retract a perturbed iterate back to $\mathcal{C}_{\boldsymbol{P}}$ in general, though may bring it closer.

On the other hand, (42) always has $\rho(\mathrm{D}\boldsymbol{F}) = 1$. Note that (30) has the special feature that $\boldsymbol{P}$ does not show up in the iteration at all and the converging point $(\boldsymbol{X}_*, \boldsymbol{Y}_*)$ can vary with the initial value $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ continuously. Indeed, the fixed points of (30) are not isolated but form an n^2-dimensional submanifold corresponding to the n^2 degrees of freedom in $\boldsymbol{X}_0\boldsymbol{Y}_0^{-1}$ for specifying the initial values. Moreover, the n^2 eigenvectors associated with $\mu'_{ij} = 1$ in (42) span the tangent space of this fixed-point submanifold at $\boldsymbol{X}_*$. Any perturbation along these directions will neither grow nor diminish, but will simply remain constant and eventually reflect itself on a shift in the converging point, while perturbations along the eigenvectors for $\mu_{ij} = 0$ will decay quadratically.

Before moving on, we make several remarks on the practical use of Eqs. (28)–(28). First, solving the Sylvester equation for $\Delta\boldsymbol{X}_k$ in (28) is nontrivial and expensive [31, sec. 13.3], which renders (28) of little practical value. Second, the extra constraint $\|\boldsymbol{X}_0\|_2 < \sqrt{3/\|\boldsymbol{P}\|_2}$ for the convergence of (31) indeed poses no difficulty because such an $\boldsymbol{X}_0$ is easy to find. Note that $\|\boldsymbol{X}_0\|_2 < \sqrt{3/\|\boldsymbol{P}\|_2}$ can be replaced by a more stringent inequality $\|\boldsymbol{X}_0\| < \sqrt{3/\|\boldsymbol{P}\|}$ since $\|\boldsymbol{X}_0\|_2 = \rho(\boldsymbol{X}_0) \le \|\boldsymbol{X}_0\|$ for any operator norm $\|\cdot\|$ and symmetric $\boldsymbol{X}_0$. Hence, we can replace the 2-norm by another

operator norm (e.g. 1-norm or ∞-norm) that can be more easily calculated. A simple choice of $\boldsymbol{X}_0$ can be, for example, $\boldsymbol{X}_0 = (1/\sqrt{\|\boldsymbol{P}\|_1})\boldsymbol{I}$.
(ii) Stablizing two unstable SPD square root iterations by combining with orthogonal polar factor iterations

Next, we review two iterations for computing the orthogonal polar factor of an invertible $\boldsymbol{R} \in \mathcal{M}_{n\times n}$ (see [26] and [20, sec. 9.4.3]):

$$\boldsymbol{X}_{k+1} = (1/2)(\boldsymbol{X}_k + \boldsymbol{X}_k^{-\mathrm{T}}); \tag{43}$$

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + (1/2)\boldsymbol{X}_k(\boldsymbol{I} - \boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k), \tag{44}$$

where the initial value is set to $\boldsymbol{X}_0 = \boldsymbol{R}$ for both iterations. With the polar decomposition $\boldsymbol{X}_0 = \boldsymbol{Q}_0\boldsymbol{S}_0$, Eqs. (43) and (44) is analytically equivalent to using the same iterations with initial value $\boldsymbol{X}_0 = \boldsymbol{S}_0$ to compute the square root of $\boldsymbol{I}$. Note that Eqs. (43) and (44) are almost the same as Eqs. (29) and (31) with $\boldsymbol{P} = \boldsymbol{I}$ except for the presence of an extra transpose. Analytically this makes no difference when the initial value $\boldsymbol{X}_0$ is SPD since all the iterates will be symmetric. However, surprisingly it turns out that this extra transpose makes a huge difference in numerical stability because it indeed lifts the stringent $\kappa(\boldsymbol{P}) < 9$ stability constraint for Eqs. (29) and (31).

We will expound this significant improvement by generalizing Eqs. (43) and (44) to the following multi-purpose (SPD square root or orthogonal polar factor) iterations:

$$\boldsymbol{X}_{k+1} = (1/2)(\boldsymbol{X}_k + \boldsymbol{X}_k^{-\mathrm{T}}\boldsymbol{P}); \tag{45}$$

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + (1/2)\boldsymbol{X}_k(\boldsymbol{I} - \boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}_k), \tag{46}$$

where $\boldsymbol{P}$ is some SPD matrix. Equations (45) and (46) can be viewed as the counterparts of Eqs. (29) and (31) for computing the square root of $\boldsymbol{P}$ in the sense of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{P}$ rather than $\boldsymbol{X}^2 = \boldsymbol{P}$.

Lemma 3.3 *Suppose the initial value $\boldsymbol{X}_0$ is invertible and has polar decomposition $\boldsymbol{X}_0 = \boldsymbol{Q}_0\boldsymbol{S}_0$ (or $\boldsymbol{X}_0^{\mathrm{T}} = \boldsymbol{Q}_0^{\mathrm{T}}\boldsymbol{S}_0$). Then we have the following convergence properties:*

$$(45):\ \boldsymbol{X}_k \to \boldsymbol{Q}_0\boldsymbol{P}^{1/2} \quad \text{for } \boldsymbol{X}_0 = \boldsymbol{Q}_0\boldsymbol{S}_0,\ \boldsymbol{S}_0 \in \mathcal{C}_{\boldsymbol{P}}; \tag{47}$$

$$(46):\ \boldsymbol{X}_k \to \boldsymbol{P}^{-1/2}\boldsymbol{Q}_0 \quad \text{for } \boldsymbol{X}_0 = \boldsymbol{S}_0\boldsymbol{Q}_0,\ \boldsymbol{S}_0 \in \mathcal{C}_{\boldsymbol{P}},\ \|\boldsymbol{X}_0\|_2 < \sqrt{3/\|\boldsymbol{P}\|_2}. \tag{48}$$

In other words, $\boldsymbol{X}_k$ converges to a particular (inverse) square root of $\boldsymbol{P}$ that shares the same orthogonal polar factor with $\boldsymbol{X}_0$. In particular, if we set $\boldsymbol{P} = \boldsymbol{I}$, $\boldsymbol{X}_k$ will converge to the orthogonal polar factor $\boldsymbol{Q}_0$ of $\boldsymbol{X}_0$, while if we instead set $\boldsymbol{Q}_0 = \boldsymbol{I}$ (i.e. $\boldsymbol{X}_0 \in \mathcal{C}_{\boldsymbol{P}}$), $\boldsymbol{X}_k$ will converge to the (inverse) SPD square root of $\boldsymbol{P}$. Hence, these two multi-purpose iterations reflect the inherent connections between matrix square root and polar decomposition and unify the two on a algorithmic level.

Using the same methodology for stability analysis as before, we obtain:

Lemma 3.4 *The Fréchet derivative at $\boldsymbol{X}_*$ for both Eqs. (45) and (46):*

$$\mathrm{D}\boldsymbol{F}\boldsymbol{X} = (1/2)(\boldsymbol{X} - \boldsymbol{\Lambda}^{-1/2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}), \tag{49}$$

Table 1 Recommended algorithms for computing the minimum-correction second-moment matching solution $\tilde{\boldsymbol{U}}_*$ in Theorem 2.1. $\tilde{\boldsymbol{U}}_*$-based algorithms avoid forming $\boldsymbol{P} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ and directly operate on $\boldsymbol{U}$

Algorithm	Description
$\boldsymbol{A}_*$-EVD	Use (27) for $\boldsymbol{A}_*$ with $\sqrt{\cdot}$ computed by Cholesky factorization and $(\cdot)^{1/2}$ computed by EVD (see Sect. 3.2.1)
$\boldsymbol{A}_*$-NtSqr	Use (27) for $\boldsymbol{A}_*$ with $\sqrt{\cdot}$ computed by Cholesky factorization and $(\cdot)^{1/2}$ computed by the stablized Newton iteration (45) or (46)
$\tilde{\boldsymbol{U}}_*$-NtPD	Perform Householder QR factorization $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{R}$ Compute the orthogonal polar factor $\boldsymbol{Q}$ of $\boldsymbol{R}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}$ using the stablized Newton iteration (45) or (46). Form $\tilde{\boldsymbol{U}}_* = \boldsymbol{V}\boldsymbol{Q}\sqrt{\tilde{\boldsymbol{P}}}$

where the diagonal $\boldsymbol{\Lambda}$ contains the eigenvalues of $\boldsymbol{P}$. Again, by observation, we obtain all the eigenvalues and eigenvectors for the linear operator $\mathrm{D}\boldsymbol{F}$ *as:*

$$\begin{aligned} \boldsymbol{Z}_{ij} &= \boldsymbol{E}_{ij} + \sqrt{\lambda_i/\lambda_j}\boldsymbol{E}_{ji}, \quad \mu_{ij} = 0, \quad i \le j, \\ \boldsymbol{Z}_{ij} &= \boldsymbol{E}_{ij} - \sqrt{\lambda_i/\lambda_j}\boldsymbol{E}_{ji}, \quad \mu_{ij} = 1, \quad i > j. \end{aligned} \tag{50}$$

Therefore, instead of the stringent constraint $\kappa(\boldsymbol{P}) < 9$ for Eqs. (29) and (31) due to (41), now we have $\rho(\mathrm{D}\boldsymbol{F}) \equiv 1$ regardless of $\kappa(\boldsymbol{P})$. Moreover, the fixed points are no longer isolated and they form a $n(n-1)/2$-dimensional submanifold corresponding to the unitary freedom encoded by $\boldsymbol{Q}_0$. The $n(n-1)/2$ eigenvectors associated with eigenvalue 1 span the tangent space of this fixed-point submanifold at $\boldsymbol{X}_*$ and any perturbation orthogonal to this tangent space decays quadratically to 0.

Again, as discussed in the end of part *i)*, the extra initial value constraint $\|\boldsymbol{X}_0\|_2 < \sqrt{3/\|\boldsymbol{P}\|_2}$ for (46) poses no practical difficulty. The trade-off between (45) and (46) lies in the relative cost between matrix inversion $\boldsymbol{A}^{-1}\boldsymbol{B}$ and multiplication $\boldsymbol{A}\boldsymbol{B}$ for square $\boldsymbol{A}$ and $\boldsymbol{B}$. (45) costs one inversion while (46) costs three multiplications. In practice, computing $\boldsymbol{A}^{-1}\boldsymbol{B}$ is about 1.5 times the cost of computing $\boldsymbol{A}\boldsymbol{B}$ (based on testing in MATLAB and [29]), so overall Eqs. (45) and (46) are equally good alternatives to the SVD/EVD approach for computing the SPD square root or the orthogonal polar factor.

Finally, we emphasize that unlike Eqs. (28)–(31) which aim to solve $\boldsymbol{X}^2 = \boldsymbol{P}$, Eqs. (45) and (46) target at solving $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{P}$. Consequently, Eqs. (45) and (46) will not yield the major square root of a general $\boldsymbol{P}$ with eigenvalues not on the negative real axis.

The various new algorithms recommended for solving the minimum-correction second-moment matching problem in Theorem 2.1 are summarized in Table 1.

3.3 The special case of orthonormalization

Now we turn to the special case of orthonormalization, i.e. $\tilde{\boldsymbol{P}} = \boldsymbol{I}$, where (27) reduces to $\boldsymbol{A}_* = \boldsymbol{P}^{-1/2}$ and $\boldsymbol{U} = \tilde{\boldsymbol{U}}_*\boldsymbol{A}_*^{-1}$ becomes the polar decomposition. In this case, [28] provides an a priori bound for $\|\tilde{\boldsymbol{U}}_* - \boldsymbol{U}\|_{\mathrm{F}}$:

$$\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}-\boldsymbol{I}\|_{\mathrm{F}}/(\|\boldsymbol{U}\|_2+1)\le\|\tilde{\boldsymbol{U}}_*-\boldsymbol{U}\|_{\mathrm{F}}\le\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}-\boldsymbol{I}\|_{\mathrm{F}}. \tag{51}$$

When $\|\boldsymbol{U}\|_2$ is of order 1 or smaller, the above yields a tight estimate of the order of magnitude of $\|\tilde{\boldsymbol{U}}_*-\boldsymbol{U}\|_{\mathrm{F}}$. One common scenario of such is with $\boldsymbol{U}$ being nearly orthogonal, i.e. $\boldsymbol{P}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\approx\boldsymbol{I}$, when, for example, a path on a Stiefel manifold is tracked numerically for subspace tracking or optimization with orthogonality constraints.

For orthonormalization, there are many widely-used algorithms that yield an orthogonal matrix not being the closest to the original $\boldsymbol{U}$. We clarify next how they are related to the minimum-correction solution.

3.3.1 Comparing with QR-based orthonormalization

If we only seek an orthogonal $\tilde{\boldsymbol{U}}$ close to $\boldsymbol{U}$, but not necessarily the closest one, we can stop after the QR factorization step $\boldsymbol{U}=\boldsymbol{V}\boldsymbol{R}$ in Sect. 3.1 and take $\tilde{\boldsymbol{U}}=\boldsymbol{V}$. This corresponds to the Cholesky decomposition $\boldsymbol{P}=\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}$. The continuity of the Cholesky factor $\boldsymbol{R}$ with respect to $\boldsymbol{P}$ [47, sec. 12.1.3] guarantees that $\boldsymbol{P}\approx\boldsymbol{I}$ implies $\boldsymbol{R}\approx\boldsymbol{I}$ and thus $\boldsymbol{V}\approx\boldsymbol{U}$. More precisely,

$$\|\tilde{\boldsymbol{U}}_*-\boldsymbol{U}\|_{\mathrm{F}}\le\|\boldsymbol{V}-\boldsymbol{U}\|_{\mathrm{F}}\le\frac{1+\|\boldsymbol{U}\|_2}{\sqrt{2}(1-\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}-\boldsymbol{I}\|_2)}\|\tilde{\boldsymbol{U}}_*-\boldsymbol{U}\|_{\mathrm{F}}, \tag{52}$$

which is credited to Jiguang Sun by [25], bounds the closeness of the orthogonal QR factor $\boldsymbol{V}$ to $\boldsymbol{U}$ by a multiple of the minimum $\|\tilde{\boldsymbol{U}}_*-\boldsymbol{U}\|_{\mathrm{F}}$ (see also [9, sec. 4]). In particular, we have approximately $\|\boldsymbol{V}-\boldsymbol{U}\|_{\mathrm{F}}\le\sqrt{2}\|\tilde{\boldsymbol{U}}_*-\boldsymbol{U}\|_{\mathrm{F}}$ when $\boldsymbol{P}\approx\boldsymbol{I}$.

This closeness property of the QR factorization stems from the nested subspace correspondence between $\boldsymbol{V}$ and $\boldsymbol{U}$, meaning that the first $k=1,\dots,n$ columns of $\boldsymbol{V}$ and $\boldsymbol{U}$ span the same subspace in $\mathbb{R}^m$, which constrains $\boldsymbol{V}$ to be close to $\boldsymbol{U}$. This also implies that when $\boldsymbol{U}$'s columns are permuted, the corresponding orthogonal QR factor $\boldsymbol{V}$ cannot be obtained by applying the same permutation to the original one, unlike the orthogonal polar factor $\tilde{\boldsymbol{U}}_*$.

3.3.2 Comparing with plain SVD orthonormalization

[52, app. B] proposes a plain SVD algorithm where the left singular vectors $\boldsymbol{V}$ of an SVD $\boldsymbol{U}=\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{Z}^{\mathrm{T}}$ is directly taken as $\tilde{\boldsymbol{U}}$. The prime issue with this approach is the non-uniqueness of an SVD. The singular vectors associated with the same (or numerically close) singular value are only determined up to an orthogonal transform, which can flip signs, permute the order and even rotate the vectors. This is especially problematic when $\boldsymbol{U}$ is nearly orthogonal because all its singular values will cluster around 1. In this case, $\|\boldsymbol{V}-\boldsymbol{U}\|_{\mathrm{F}}$ can be of order 1 even if $\boldsymbol{U}$ is nearly orthogonal. To remedy this issue, [18] suggest a gradient descent algorithm for orthonormalization. Here we analyze it in "Appendix D" and conclude that due to its linear convergence and stringent stability constraint, it is unattractive compared to the algorithms in Sect. 3.2.

Table 2 Description of the various orthonormalization algorithms benchmarked. Here $\boldsymbol{U}$ is to be orthonormalized and $\boldsymbol{P} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ is its Gram matrix

Algorithm	Description
SVD–EVD	Plain SVD ortho. (see Sect. 3.3) by EVD of $\boldsymbol{P}$
PD-EVD	PD ortho. by EVD of $\boldsymbol{P}$ (see Sect. 3.2.1)
PD-SVD*	PD ortho. by SVD of $\boldsymbol{U}$ (see Sect. 3.2.1)
PD-NtSqr	PD ortho. by computing $\boldsymbol{P}^{1/2}$ using Newton iteration (45)
PD-NtISqr	PD ortho. by computing $\boldsymbol{P}^{-1/2}$ using Newton iteration (46)
PD-GD	PD ortho. by computing $\boldsymbol{P}^{-1/2}$ using gradient descent (87)
PD-NtPD*	PD ortho. by Householder QR $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{R}$ and computing the orthogonal polar factor of $\boldsymbol{R}$ using Newton iteration (45)
QR-CGS*	QR ortho. by classical Gram-Schmidt
QR-MGS*	QR ortho. by modified Gram-Schmid
QR-Chol	QR ortho. by Cholesky decomposition of $\boldsymbol{P}$
QR-HH*	QR ortho. by Householder QR $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{R}$

The algorithms that do not form $\boldsymbol{P}$ and directly operate on $\boldsymbol{U}$ are marked by a "*" in their names

4 Benchmarks of the algorithms' performance

In this section, we benchmark the performance of the various new minimum-correction second-moment matching algorithms introduced in Sect. 3 for the special case of orthonormalization (i.e. $\tilde{\boldsymbol{P}} = \boldsymbol{I}$) and compare them to existing algorithms. All the algorithms to be tested are described in Table 2. Note that PD-NtSqr, PD-NtISqr and PD-NtPD are based on our new multi-purpose Newton iteration (45) and (46), while others are existing algorithms from literature. Comparing Table 2 to Table 1, here PD-EVD corresponds to $\boldsymbol{A}_*$-EVD in Table 1, PD-NtSqr and PD-NtISqr to $\boldsymbol{A}_*$-NtSqr, while PD-NtPD corresponds to $\tilde{\boldsymbol{U}}_*$-NtPD.

For all algorithms, the fixed-point iteration terminates once $\|\boldsymbol{X}_k - \boldsymbol{X}_{k-1}\| < 10^{-14}\|\boldsymbol{X}_{k-1}\|$ is satisfied or the maximum number of iterations is reached. The numerical tests are performed under the MATLAB® environment on a machine with an Intel® Core™ i7-4702MQ processor and 16GB of memory.

The algorithms are tested in three cases. The first is with a $10^6 \times 10^2$ matrix $\boldsymbol{U}$ (m-by-n, $m \gg n$) of a small condition number $\kappa = 1.5$, so the tall thin $\boldsymbol{U}$ is nearly orthogonal. We form $\boldsymbol{U} = \boldsymbol{U}_0\boldsymbol{\Lambda}\boldsymbol{Q}_0$ by random generation of an orthogonal m-by-n $\boldsymbol{U}_0$ and orthogonal n-by-n $\boldsymbol{Q}_0$. A random orthogonal matrix is obtained by applying QR Householder to a random matrix whose entries are independent and uniformly distributed over $[-1, 1]$. $\boldsymbol{\Lambda}$ is a diagonal matrix of condition number κ and of diagonal entries decaying geometrically to 1. Each algorithm is tested using the same 10 samples of $\boldsymbol{U}$ and the average performance metrics are listed in Table 3. For the second and third cases, the same tests are performed, but for a square well-conditioned $\boldsymbol{U}$ and a rectangular ill-conditioned $\boldsymbol{U}$, respectively. Results are in Table 4.

Overall, Tables 3 and 4 confirm several key messages:

Table 3 Performance of the various algorithms for orthonormalization in the case of well-conditioned (nearly orthogonal) rectangular matrices (m-by-n, $m \gg n$)

	$m = 10^6,\ n = 10^2,\ \kappa(\boldsymbol{U}) = 1.5$			
	$\|\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}} - \boldsymbol{I}\|$	$\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|$	$\|\boldsymbol{A} - \boldsymbol{A}^{\mathrm{T}}\|$	(#Ite) time
SVD–EVD	2e−14	16	12	0.6s
PD-EVD	3e−14	2.75	1e−15	0.6s
PD-SVD*	3e−14	2.75	1e−15	5s
PD-NtSqr	8e−15	2.75	8e−16	(6) 2.3s
PD-NtISqr	9e−15	2.75	1e−15	(8) 0.6s
PD-GD	8e−13	2.75	3e−15	(131) 0.6s
PD-NtPD*	1e−14	2.75	1e−15	(7) 5s
QR-CGS*	6e−15	3.04	1.9	38s
QR-MGS*	6e−15	3.04	1.9	38s
QR-Chol	7e−15	3.04	1.9	2s
QR-HH*	1e−14	3.04	1.9	4s

Here $\boldsymbol{A}$ is the linear transform such that $\tilde{\boldsymbol{U}} = \boldsymbol{U}\boldsymbol{A}$. The Frobenius norm is exclusively used. For the algorithms based on a fixed-point iteration, the average number of iterations is recorded in the parenthesis in front of the run time. In this case, we have $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{I}\| = 6.5$ for all 10 repetitions, which can be compared to $\|\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}} - \boldsymbol{I}\|$

Table 4 Performance of the algorithms on well-conditioned square matrices ($\kappa(\boldsymbol{U}) = 1.5,\ \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{I}\| = 29$) and on ill-conditioned rectangular matrices ($\kappa(\boldsymbol{U}) = 10^6,\ \|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{I}\| = 1.5 \times 10^{12}$), respectively

	$m = n = 2 \times 10^3,\ \kappa(\boldsymbol{U}) = 1.5$			$m = 10^6,\ n = 10^2,\ \kappa = 10^6$	
	$\|\tilde{\boldsymbol{P}} - \boldsymbol{I}\|$	$\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|$	(#Ite) time	$\|\tilde{\boldsymbol{P}} - \boldsymbol{I}\|$	(#Ite) time
PD-EVD	3e−13	12.3	2s	4e−5	0.6s
PD-SVD*	4e−13	12.3	4s	3e−14	5s
PD-NtSqr	4e−14	12.3	(7) 3s	4e−5	(100) 2s
PD-NtInv	6e−14	12.3	(9) 6s	4e−5	(100) 0.6s
PD-GD	3e−11	12.3	(422) 280s	9	(1e4) 3s
PD-NtPD*	7e−14	12.3	(9) 4s	2e−14	(26) 5s
QR-CGS*	8e−14	13.6	13s	6e−5	38s
QR-MGS*	6e−14	13.6	13s	2e−10	38s
QR-Chol	3e−14	13.6	0.3s	3e−5	2s
QR-HH*	7e−14	13.6	0.4s	1e−14	4s

Notations are as those in Table 3

i The plain SVD algorithm does not promote closeness between $\tilde{\boldsymbol{U}}$ and $\boldsymbol{U}$, as is verified by $\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|$ of SVD–EVD being much larger than that of the QR-based and of the new minimum-correction (PD-based) algorithms, see Table 3.

ii The new PD-based algorithms achieve the minimum correction as is verified by the agreement between their $\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|$ values and the analytical one computed using (6) as $\|\tilde{\boldsymbol{U}}_* - \boldsymbol{U}\| = (\mathrm{tr}\boldsymbol{I} + \mathrm{tr}\boldsymbol{P} - 2\mathrm{tr}\boldsymbol{P}^{1/2})^{1/2} = (n + \|\boldsymbol{U}\|_{\mathrm{F}}^2 - 2\|\boldsymbol{U}\|_*)^{1/2}$, where $\|\cdot\|_*$

is the nuclear norm, being the sum of singular values. The symmetry of $\boldsymbol{A}_*$ is verified by the machine epsilon values of $\|\boldsymbol{A} - \boldsymbol{A}^{\mathrm{T}}\|$. The $\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|$ values of the QR-based algorithms are larger than $\|\tilde{\boldsymbol{U}}_* - \boldsymbol{U}\|$, but within the bound (52).

iii For ill-conditioned matrices, the algorithms based on forming $\boldsymbol{P} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ suffer from poor accuracy due to squaring the condition number. This is revealed by the large $\|\tilde{\boldsymbol{P}} - \boldsymbol{I}\|$ values of the algorithms without a "*" in their names, see Table 4. Among them, those based on a fixed-point iteration have their solutions stagnate with large orthogonality errors after only a few iterations. The iterates neither converge nor blow up and just bounce around, even after as many as 10^5 iterations. CGS and MGS are known to suffer from the same orthogonality issue in this case. Hence, the only algorithms that can still reliably orthonormalize $\boldsymbol{U}$ are those based on performing Householder QR or SVD on $\boldsymbol{U}$.

iv For tall thin matrices, the algorithms based on forming $\boldsymbol{P} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ tend to be more efficient when $\boldsymbol{U}$ is well-conditioned, because other than computing the small $\boldsymbol{P}$ in the beginning and applying $\boldsymbol{A}$ to $\boldsymbol{U}$ in the end, they only deal with small n-by-n matrices. This strategy avoids performing QR factorization or SVD to a large m-by-n matrix, which could be relatively expensive in this case.

There is yet another algorithm for computing the SPD matrix square root that takes a completely different route from fixed-point iterations. [23] utilizes the matrix version of the Cauchy integral formula for the square root function and combines conformal maps and the trapezoidal rule in computing the contour integral. However, its exponential/geometric convergence rate is only equivalent to a linearly converging fixed-point iteration, which is significantly slower than the quadratic convergence of Eqs. (45) and (46) proposed previously. Indeed, this can be verified by comparing to the results based on the Pascal matrices shown in Fig. 7 of [23]. It takes 9 and 20 Newton iterations to obtain the SPD square root of the 3-by-3 and 8-by-8 Pascal matrix, respectively, while the contour-integral approach requires 13 and 34 quadrature points to reach the same accuracy. Here every extra quadrature point has the equivalent cost of one iteration step because both involve solving a linear system of the same size. Besides efficiency, using fixed-point iteration has the extra advantage of conceptual and algorithmic simplicity.

5 Application to matrix differential equations that preserve orthogonality

A matrix differential equation

$$\mathrm{d}_t \boldsymbol{U} = \boldsymbol{F}(\boldsymbol{U}) \tag{53}$$

for $\boldsymbol{U}(t) \in \mathcal{M}_{m\times n}$ is said to preserve orthogonality if $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$ at $t = 0$ implies $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$ at all $t > 0$. We encounter such equations in applications such as subspace tracking [5,8,17,30,32,34], where we track the time evolution of an n-dimensional subspace of $\mathbb{R}^m$, which may come from a discretization of an infinite-dimensional dynamical system, by evolving an orthonormal basis with its base vectors forming $\boldsymbol{U}$'s columns. In this section, we showcase our new minimum-correction second-moment

matching algorithms in the special case of $\tilde{\boldsymbol{P}} = \boldsymbol{I}$ for re-orthonormalization required in solving such equations. Before the examples, we analyze the error due to the loss of numerical orthogonality and clarify the role played by re-orthonormalization.

Although (53) preserves orthogonality, its time-discretized counterpart $\hat{\boldsymbol{U}}_{n+1} = \hat{\boldsymbol{F}}(\hat{\boldsymbol{U}}_n)$ in general does not [11,24,25,53]. Here, $\hat{\boldsymbol{U}}_n$ is a numerical approximation of $\boldsymbol{U}_n = \boldsymbol{U}(t_n)$ and the map $\hat{\boldsymbol{F}}$ stems from a particular time-marching scheme, which is said to preserve orthogonality if $\hat{\boldsymbol{U}}_n^{\mathrm{T}}\hat{\boldsymbol{U}}_n = \boldsymbol{I} \implies \hat{\boldsymbol{U}}_{n+1}^{\mathrm{T}}\hat{\boldsymbol{U}}_{n+1} = \boldsymbol{I}$. According to [11,24], the only known family of orthogonality-preserving time-marching schemes is the one of Gauss–Legendre Runge–Kutta schemes. Unfortunately, such schemes are difficult to implement, expensive to use and not widely available in numerical software [25]. Moreover, they only preserve orthogonality for certain $\boldsymbol{F}$, which renders them even less attractive. Therefore, a more practical strategy is to integrate (53) by an arbitrary scheme and re-orthonormalize the $\hat{\boldsymbol{U}}_n$'s at the end of each time step.

We denote by $\hat{\boldsymbol{U}}$ the numerical solution at a particular time obtained by some time-marching scheme without re-orthonormalization, and by $\Delta\boldsymbol{U} = \hat{\boldsymbol{U}} - \boldsymbol{U}$ the time-marching error. If we choose $\|\cdot\|$ as the matrix 2-norm, we have

$$\|\hat{\boldsymbol{U}}^{\mathrm{T}}\hat{\boldsymbol{U}} - \boldsymbol{I}\| = \|(\boldsymbol{U} + \Delta\boldsymbol{U})^{\mathrm{T}}(\boldsymbol{U} + \Delta\boldsymbol{U}) - \boldsymbol{I}\| \leq 2\,\|\Delta\boldsymbol{U}\|\,\|\boldsymbol{U}\| + \|\Delta\boldsymbol{U}\|^2\,,$$

since $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$. If the scheme is p-th order accurate, i.e. $\|\Delta\boldsymbol{U}\| = \mathrm{O}((\Delta t)^p)$, since $\|\boldsymbol{U}\| = \mathrm{O}(1)$, we have $\|\hat{\boldsymbol{U}}^{\mathrm{T}}\hat{\boldsymbol{U}} - \boldsymbol{I}\| = \mathrm{O}((\Delta t)^p)$, which is of the same order as the time-marching error. This seems to imply that as long as the time-marching scheme is accurate enough, there is no need to worry about the orthogonality error. However, the fact that (53) preserves orthogonality only implies that the Stiefel manifold $\mathcal{S}_{\boldsymbol{I}} = \{\boldsymbol{U} : \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}\}$ is invariant under (53), but $\mathcal{S}_{\boldsymbol{I}}$ may not be a stable manifold. Therefore, any deviation of $\boldsymbol{U}$ from $\mathcal{S}_{\boldsymbol{I}}$ may not stay bounded under (53). It might even hit an unstable direction and diverge exponentially. In addition, in many cases, the derivation and thus the validity of (53) itself relies on $\boldsymbol{U}$ remaining orthogonal, so a violation of this condition may unexpectedly ruin the numerical solution.

The above analysis justifies re-orthonormalizing $\hat{\boldsymbol{U}}$ after each time step. This procedure should ideally not undermine the convergence order of the time-marching scheme or introduce significant numerical artifacts. Hence, we need to analyze how re-orthonormalization affects the numerical errors.

Suppose $\tilde{\boldsymbol{U}} \in \mathcal{S}_{\boldsymbol{I}}$ is the orthonormalized solution obtained from $\hat{\boldsymbol{U}}$. To maintain the order of convergence, it suffices to have $\|\tilde{\boldsymbol{U}} - \hat{\boldsymbol{U}}\| \leq C\|\hat{\boldsymbol{U}} - \boldsymbol{U}\|$ for some $C > 0$ because then by triangle inequality, we have $\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\| \leq (C+1)\|\hat{\boldsymbol{U}} - \boldsymbol{U}\|$. This can be satisfied if we choose $\tilde{\boldsymbol{U}}$ to be $\tilde{\boldsymbol{U}}_*$, the orthogonal projection of $\hat{\boldsymbol{U}}$ onto $\mathcal{S}_{\boldsymbol{I}}$, because, as $\boldsymbol{U} \in \mathcal{S}_{\boldsymbol{I}}$, Corollary 2.3 implies $\|\tilde{\boldsymbol{U}}_* - \hat{\boldsymbol{U}}\| \leq \|\hat{\boldsymbol{U}} - \boldsymbol{U}\|$ and thus $\|\tilde{\boldsymbol{U}}_* - \boldsymbol{U}\| \leq 2\|\hat{\boldsymbol{U}} - \boldsymbol{U}\|$. Indeed, since $\|\hat{\boldsymbol{U}} - \boldsymbol{U}\|$ is small, $\tilde{\boldsymbol{U}}_*$, $\hat{\boldsymbol{U}}$, and $\boldsymbol{U}$ are almost the vertices of a right triangle ("almost" because $\mathcal{S}_{\boldsymbol{I}}$ is not flat), so in practice, $\|\tilde{\boldsymbol{U}}_* - \boldsymbol{U}\| \leq \|\hat{\boldsymbol{U}} - \boldsymbol{U}\|$. Numerical examples in [11] show that in some cases $\|\tilde{\boldsymbol{U}}_* - \boldsymbol{U}\|$ can even be smaller than $\|\hat{\boldsymbol{U}} - \boldsymbol{U}\|$ by an order of magnitude. Another viable option for $\tilde{\boldsymbol{U}}$ is the orthogonal QR factor $\boldsymbol{V}$ of $\hat{\boldsymbol{U}}$ because (52) guarantees $\|\boldsymbol{V} - \hat{\boldsymbol{U}}\| \leq \sqrt{2}\|\tilde{\boldsymbol{U}}_* - \hat{\boldsymbol{U}}\| \leq \sqrt{2}\|\hat{\boldsymbol{U}} - \boldsymbol{U}\|$.

To showcase the new algorithms and demonstrate the above results, we consider the stochastic Lorenz-96 system [38] (the index i is circular, i.e. $x_i = x_{i+m}$),

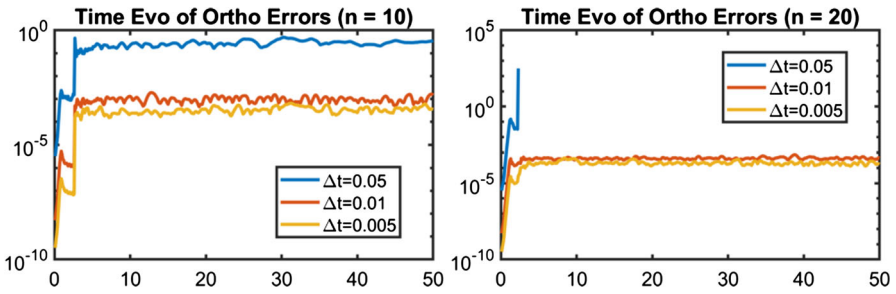


Fig. 2 Orthogonality errors in the absence of re-orthonormalization. The error saturates at a higher level with larger time step Δt. When $\Delta t = 0.05$, the orthogonality is completely lost for $n = 10$ (left), while the solution even blows up for $n = 20$ (right)

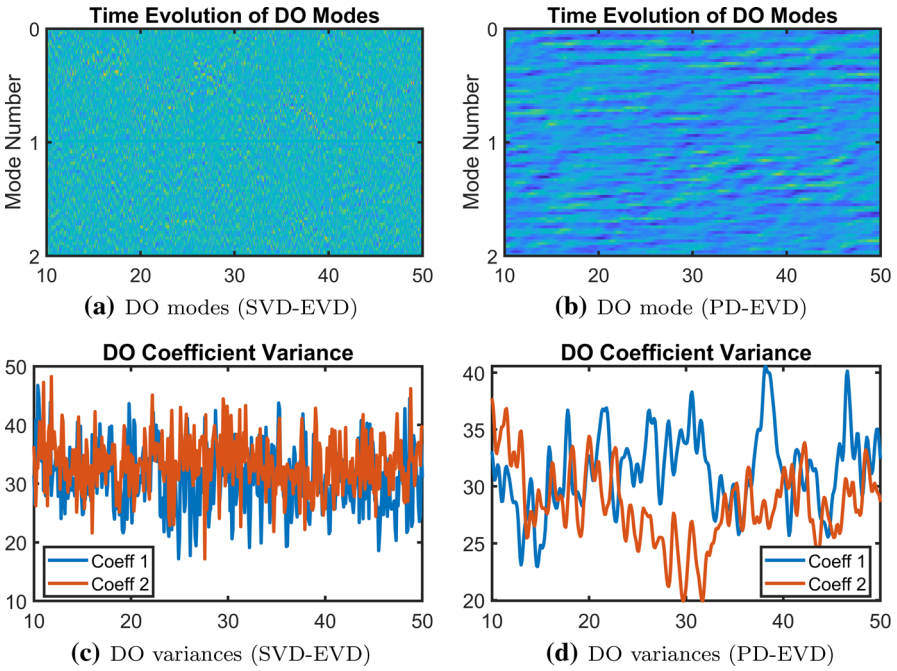
$$\mathrm{d}_t x_i = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F, \qquad i = 1, \ldots, m. \tag{54}$$

This toy model mimics atmospheric convection along a mid-latitude circle. We set $m = 40$ and $F = 8$ which lead to chaotic dynamics. In all simulations, the initial condition is $x_i = 0$ for $i \neq 1$ while x_1 is uniformly distributed over $[-0.01, 0.01]$. The classical explicit RK-4 scheme is used for time integration.

To solve (54), we employ the dynamically orthogonal (DO) equations [46]. They approximate the random vector $\boldsymbol{u}(t;\omega) = [x_1, \ldots, x_m]^{\mathrm{T}} \in \mathbb{R}^m$ by a low rank expansion $\boldsymbol{u}(t;\omega) = \bar{\boldsymbol{u}}(t) + \sum_{i=1}^{n} \phi_i(t;\omega)\boldsymbol{u}_i(t) = \bar{\boldsymbol{u}} + \boldsymbol{U}\boldsymbol{\phi}$ where $\bar{\boldsymbol{u}}$ is the mean, $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n] \in \mathcal{M}_{m\times n}$ contains the orthonormal base vectors and $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_n]^{\mathrm{T}}$ contains the random scalar coefficients. The evolution of all these components is governed by a dynamical system, the DO equations, that preserves and whose validity is predicated on $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} \equiv \boldsymbol{I}$. Such techniques are useful for uncertainty quantification in high-dimensional systems governed by time-dependent stochastic PDEs [18,49]. Related works include the dynamic low-rank approximation of a time-varying matrix [30], geometric analyses of the DO equations based on matrix manifold theory [17,19] and dynamically bi-orthogonal equations [10].

In Fig. 2, we first show the time evolution of the orthogonality error $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{I}\|_{\max}$ ($\|\boldsymbol{A}\|_{\max} \triangleq \max\{A_{ij}\}$) for the numerical integration of the DO equations without any re-orthonormalization. Results are shown for $n = 10,\ 20$ and three time step sizes. As we can see, the orthogonality error grows with Δt since a larger Δt implies a larger discrepancy between the time-discrete system and the originally orthogonality-preserving continuous one. Moreover, when $\Delta t = 0.005$ and 0.01, $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{I}\|_{\max}$ tends to saturate at a low level ($< 10^{-3}$) that corresponds to only a negligible deviation from orthogonality. However, when Δt increases to 0.05, $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{I}\|_{\max}$ can either level off at a magnitude of 1 for $n = 10$ that signals a complete loss of orthogonality or even drive the simulation unstable and cause the solution to blow up for $n = 20$. This exemplifies the necessity of re-orthonormalization.

Note that in the expansion $\boldsymbol{u} = \bar{\boldsymbol{u}} + \boldsymbol{U}\boldsymbol{\phi}$, there is also the unitary freedom $\boldsymbol{U}\boldsymbol{\phi} = (\boldsymbol{U}\boldsymbol{Q})(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\phi})$ for any orthogonal $\boldsymbol{Q}$, under which all realizations $\boldsymbol{u}(t;\omega)$ are unchanged. Therefore, if we apply such a $\boldsymbol{Q}$ at some time t and use $\boldsymbol{U}\boldsymbol{Q}$ and $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\phi}$ as the initial values of the modes and coefficients for the subsequent time integra-



(a) DO modes (SVD-EVD) **(b)** DO mode (PD-EVD)

(c) DO variances (SVD-EVD) **(d)** DO variances (PD-EVD)

Fig. 3 Comparison between orthonormalization using SVD–EVD (left) with that using polar decomposition PD-EVD (right). The evolution (from $t = 10$ to 50, $\Delta t = 0.05$) of the first two DO modes $[\boldsymbol{u}_1^{\mathrm{T}}, \boldsymbol{u}_2^{\mathrm{T}}]^{\mathrm{T}}$ stacked together are shown on the top and the evolution of their coefficient variances $\mathrm{Var}(\phi_1)$ and $\mathrm{Var}(\phi_2)$ on the bottom

tion, the future evolution of $\boldsymbol{u}(t;\omega)$ will remain the same. However, this breaks the time continuity of $\boldsymbol{U}$ and $\boldsymbol{\phi}$ at time t, as illustrated in Fig. 3, where the SVD–EVD and PD-EVD (see Table 2) algorithm are compared. Due to the non-uniqueness of an SVD, SVD–EVD not only removes the orthogonality error, but also applies to $\boldsymbol{U}$ some $\boldsymbol{Q}$ that may not at all be close to $\boldsymbol{I}$. As a result, we lose the time continuity of the modes and coefficients as indicated by the "random" pattern in the left half of Fig. 3. In contrast, PD-EVD minimizes such numerical artifacts. Maintaining time continuity is especially important when we use multi-step time-marching schemes[5] because their validity relies on the smooth evolution of $\boldsymbol{U}$ and $\boldsymbol{\phi}$ across different time steps. A "random" $\boldsymbol{Q}$ applied at each time step destroys this smoothness and can cause the time integration to diverge.

Finally we show in Fig. 4 how the orthogonality error can be controlled by PD-EVD, PD-NtSqr, PD-NtISqr, QR-Chol and SVD–EVD for re-orthonormalization. These algorithms were described in Table 2. They are the best choices in this case based on the discussion in Sect. 4 because the $\hat{\boldsymbol{U}}$ to be re-orthonormalized after each time step is a tall thin matrix and is already nearly orthogonal. Here we exclusively use $\Delta t = 0.05$ and $n = 10$ for all simulations. The left of Fig. 4 indicates that the orthogonality errors are kept at almost machine epsilon. The two EVD-based algorithms have slightly larger

[5] For example, [49] uses the leapfrog scheme and [13] uses the implicit-explicit backward difference schemes for the DO equations.

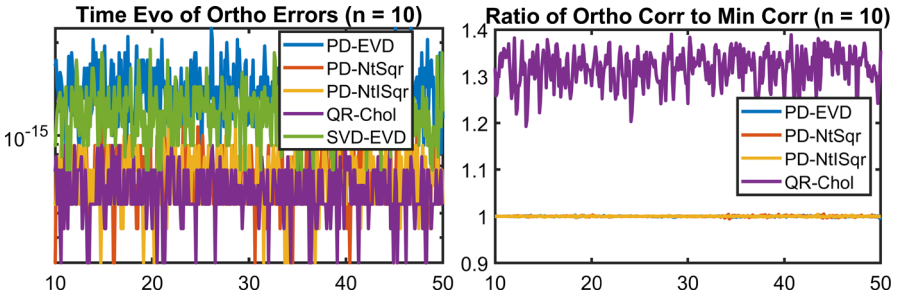


Fig. 4 The orthogonality errors after re-orthonormalization by various algorithms (left) and the ratio of the correction $\|\tilde{\boldsymbol{U}} - \hat{\boldsymbol{U}}\|$ to the minimum $\|\tilde{\boldsymbol{U}}_* - \hat{\boldsymbol{U}}\|$ computed by (6) (right). The ratios for SVD–EVD are of order 10^6 and thus not plotted. Here $\Delta t = 0.05$ and $n = 10$ are used for all the simulations

errors, most likely due to the convergence criterion for the built-in EVD subroutine, but this is not essential in practice. The right plot shows the ratio of the actual correction $\|\tilde{\boldsymbol{U}} - \hat{\boldsymbol{U}}\|$ to the theoretical minimum $\|\tilde{\boldsymbol{U}}_* - \hat{\boldsymbol{U}}\| = (n + \mathrm{tr}\,\boldsymbol{P} - 2\mathrm{tr}\,\boldsymbol{P}^{1/2})^{1/2}$ computed by (6). As we can see, all three PD-based algorithms achieve a ratio of 1, while the QR-based algorithm has a ratio between 1.2 and 1.4, agreeing with the bound (52). Not surprisingly, the plain SVD orthonormalization yields a huge ratio of order 10^6 (not plotted) because it produces corrections of order 1 even when $\|\hat{\boldsymbol{U}}^{\mathrm{T}}\hat{\boldsymbol{U}} - \boldsymbol{I}\|$ is only of order 10^{-6}.

6 Application to ensemble square root filters for data assimilation

Another application of the minimum-correction second-moment matching is in the ensemble square root filters for data assimilation [3,6,15,16,35,41,44]. Given an n-dimensional dynamical system $\mathrm{d}_t\boldsymbol{u} = \boldsymbol{f}(\boldsymbol{u})$ as well as a linear observation model $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{u} + \boldsymbol{e} \in \mathbb{R}^d$ with Gaussian noise $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$, the task of data assimilation is to combine the information content of observations $\hat{\boldsymbol{y}}$ with that of model predictions to refine the (probabilistic) estimate of the state $\boldsymbol{u}$. At each observation time t, we denote a forecast ensemble with m realizations for $\boldsymbol{u}$ by $\boldsymbol{U}_- \in \mathcal{M}_{n\times m}$[6] and its sample covariance by $\boldsymbol{P}_- = (1/m)\tilde{\boldsymbol{U}}_-\tilde{\boldsymbol{U}}_-^{\mathrm{T}}$. Here, $\boldsymbol{U}_-$ is decomposed into the mean and the fluctuation as $\boldsymbol{U}_- = \bar{\boldsymbol{u}}\boldsymbol{1}^{\mathrm{T}} + \tilde{\boldsymbol{U}}_-$, where $\boldsymbol{1}$ is a vector with all entries being 1.

The task of filtering, the most common type of data assimilation, is to obtain an analysis ensemble $\boldsymbol{U}_+ \in \mathcal{M}_{n\times m}$, which integrates information in $\boldsymbol{U}_-$ with the observations collected at t denoted by $\hat{\boldsymbol{y}}$, so that $\boldsymbol{U}_+$ better captures the underlying true state than $\boldsymbol{U}_-$ can. A popular family of filters is that of ensemble square root filters, which update the mean in the same ways as a Kalman filter:

$$\bar{\boldsymbol{u}}_+ = \bar{\boldsymbol{u}}_- + \boldsymbol{K}(\hat{\boldsymbol{y}} - \boldsymbol{H}\bar{\boldsymbol{u}}_-), \qquad \boldsymbol{K} = \boldsymbol{P}_-\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{P}_-\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R})^{-1}, \tag{55}$$

[6] Here the role of rows and columns flips compared to previous sections, since we want to follow the notation convention in the field of data assimilation.

and update the ensemble spread in some way such that the sample covariance of $\boldsymbol{U}_+$ matches the one given by a Kalman filter, i.e.

$$\boldsymbol{P}_+ = (1/m)\tilde{\boldsymbol{U}}_+\tilde{\boldsymbol{U}}_+^{\mathrm{T}} = (\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H})\boldsymbol{P}_-. \tag{56}$$

Many variants for obtaining such a $\tilde{\boldsymbol{U}}_+$ have been proposed, including the ensemble Kalman filter [15,16], error subspace statistical estimation [33,35], ensemble adjustment filter [3] and ensemble transform filter [6], to name a few [54]. Here, we demonstrate a particular variant proposed by [41], which choose $\tilde{\boldsymbol{U}}_+$ to be the one closest to $\tilde{\boldsymbol{U}}_-$, reducing the task to minimum-correction second-moment matching. This choice may be seen as retaining as much physical information from the prior ensemble $\boldsymbol{U}_-$ as possible [41].

More precisely, the update formula for $\tilde{\boldsymbol{U}}_+$ proposed by [41] is

$$\tilde{\boldsymbol{U}}_+ = \underset{\boldsymbol{V}:\ \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}=m\boldsymbol{P}_+}{\arg\min} \|\boldsymbol{V} - \tilde{\boldsymbol{U}}_-\|_{\mathrm{F},\boldsymbol{P}_-^{-1}}\ . \tag{57}$$

Corollary 2.1 implies that $\tilde{\boldsymbol{U}}_+ = \boldsymbol{A}\tilde{\boldsymbol{U}}_-$ with (here $\boldsymbol{\Gamma}_n = \boldsymbol{P}_-^{-1}$)

$$\boldsymbol{A} = \sqrt{\boldsymbol{\Gamma}_n}^{-1}\boldsymbol{A}_*\sqrt{\boldsymbol{\Gamma}_n} = \sqrt{\boldsymbol{P}_-}^{\mathrm{T}}(\sqrt{\boldsymbol{P}_-}^{-\mathrm{T}}\boldsymbol{P}_+\sqrt{\boldsymbol{P}_-}^{-1})^{1/2}\sqrt{\boldsymbol{P}_-}^{-\mathrm{T}}\ . \tag{58}$$

This is almost the same as $\boldsymbol{A} = \boldsymbol{P}_-^{1/2}(\boldsymbol{P}_-^{-1/2}\boldsymbol{P}_+\boldsymbol{P}_-^{-1/2})^{1/2}\boldsymbol{P}_-^{-1/2}$, which is eq. A14 in [41], except that all the square root $\sqrt{\cdot}$ were unnecessarily restricted to the unique SPD one $(\cdot)^{1/2}$ in the expression used by [41].

We will again use the Lorenz-96 system (54) as an example with exactly the same numerical setups as those in Sect. 5, except that $F = 4$ is used now. Hence, we have $n = 40$, $m = 1000$. Note that n now denotes the system's dimension while m the number of realizations, differing from Sect. 5. The observation data include every other variable, i.e. $x_1, x_3, \ldots, x_{39}$, at time $t = 10, 15, 20, \ldots, 50$. The observation noise covariance $\boldsymbol{R} = 0.01\boldsymbol{I}$ is assumed known.

Here four filters are tested and compared: "EnSQR" is the one proposed by [41] and based on (57); "EnKF" is the classical ensemble Kalman filter [16]; "KF" is the same as EnKF except that the filtering update is replaced by fitting a Gaussian to the prior ensemble, applying the classical Kalman update and re-sampling from the posterior Gaussian. These three filters are based on brute force Monte-Carlo simulations of the Lorenz-96 system without any dimension reduction. "EnSQR-DO" is EnSQR applied to a DO simulation with a 10-dimensional reduced subspace. The performance of these four filters are compared in Fig. 5.

The RMSE (root mean squared errors) on the left shows that all filters perform similarly except that EnSQR-DO has a larger error due to dimension reduction (the attractor is not well contained in a 10-dimensional linear subspace globally). The ratios of the ensemble correction (fluctuation part) $\|\tilde{\boldsymbol{U}}_+ - \tilde{\boldsymbol{U}}_-\|$ at each filtering step to the theoretical minimum $\|\tilde{\boldsymbol{U}}_{+*} - \tilde{\boldsymbol{U}}_-\|$ for posterior covariance matching are plotted on the right. The minimum is computed by (12) being $\|\tilde{\boldsymbol{U}}_{+*} - \tilde{\boldsymbol{U}}_-\|^2 = \mathrm{tr}(\boldsymbol{I} + \tilde{\boldsymbol{P}}_\Gamma - 2\tilde{\boldsymbol{P}}_\Gamma{}^{1/2})$, where $\tilde{\boldsymbol{P}}_\Gamma = \sqrt{\boldsymbol{P}_-}^{-\mathrm{T}}\boldsymbol{P}_+\sqrt{\boldsymbol{P}_-}^{-1}$. As we can see, the two minimum-correction-based

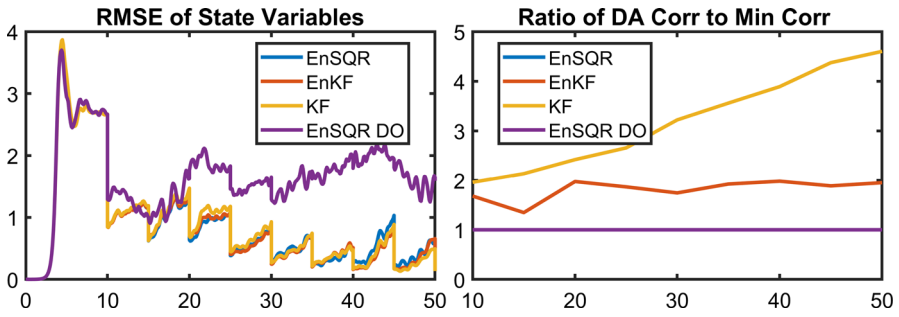


Fig. 5 Time evolution of RMSE of the various filters (left) and the ratio of the filtering correction $\|\tilde{\boldsymbol{U}}_+ - \tilde{\boldsymbol{U}}_-\|$ to the minimum $\|\tilde{\boldsymbol{U}}_{+*} - \tilde{\boldsymbol{U}}_-\|$ computed by (12) (right). The first three cases are without dimension reduction while the last one is reduced to a 10-dimensional subspace using DO. Here $\Delta t = 0.05$ is used for all the simulations

filters (EnSQR, EnSQR-DO) achieve the unit ratio while the other two do not. KF has the largest ratio because its re-sampling step completely breaks any realization correspondence between the prior and posterior ensemble. In contrast, EnKF adopts a realization-wise update strategy, which yields a ratio between those of KF and of the minimum-correction filters. For further details on the performance of the minimum-correction-based ensemble square root filter (EnSQR), please refer to section 5.3.3 and 5.7 of [36].

7 Conclusions

In this paper, we develop the theory for minimum-correction second-moment matching. We solve the optimization analytically for both the full-rank and rank-deficient case. The finite-dimensional result is generalized to the infinite-dimensional setting and a connection to optimal transport is drawn. We also show how the general problem can reduce to the familiar one of optimal orthonormalization and we accordingly generalize the polar decomposition, which is known to solve this special case.

We obtain numerical schemes for computing the optimizer. We show the instrumental role played by the algorithms for polar decomposition and SPD matrix square root and we analyze existing ones in the literature. We modify two Newton iterations deemed unstable before and significantly improve their stability. The resulting two new multi-purpose schemes (45) and (46) can be used to efficiently compute both the orthogonal polar factor and the SPD square root. These iterations play a key role in minimum-correction second-moment matching using algorithm $\boldsymbol{A}_*$-NtSqr and $\tilde{\boldsymbol{U}}_*$-NtPD in Table 1 and in their counterparts PD-NtSqr, PD-NtISqr, and PD-NtPD in Table 2 for orthonormalization. We verify the higher performance of the new algorithms using benchmarks with random matrices.

Finally, we showcase results in two applications. In reduced subspace tracking for uncertainty quantification, we maintain the numerical orthogonality and continuity of a time-varying orthonormal basis. In an ensemble square root filter for data assimilation,

we obtain the unique posterior ensemble that matches a given covariance matrix while also minimizing its distance to the prior one.

Acknowledgements We thank the members of our MSEAS group at MIT. We especially thank Florian Feppon for several suggestions and for pointing us to the connection to the polar decomposition. We are grateful to the Office of Naval Research for support under grants N00014-14-1-0476 (Science of Autonomy – LEARNS) and N00014-14-1-0725 (Bays-DA) to the Massachusetts Institute of Technology. We also want to thank the anonymous referees for reading our manuscript carefully and providing helpful feedback to help us improve the presentation.

A: Proof of theorem 2.1

Proof First, since the objective function of (4) can be rewritten as

$$\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F}}^2 = \mathrm{tr}[(\tilde{\boldsymbol{U}} - \boldsymbol{U})^{\mathrm{T}}(\tilde{\boldsymbol{U}} - \boldsymbol{U})] = \mathrm{tr}\,\tilde{\boldsymbol{P}} + \mathrm{tr}\,\boldsymbol{P} - 2\mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{U}}),$$

we have

$$\underset{\tilde{\boldsymbol{U}}\in\mathcal{M}_{m\times n}:\ \tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}}=\tilde{\boldsymbol{P}}}{\arg\min} \|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathrm{F}}^2 = \underset{\tilde{\boldsymbol{U}}\in\mathcal{M}_{m\times n}:\ \tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}}=\tilde{\boldsymbol{P}}}{\arg\min} -2\mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{U}}). \tag{59}$$

The Lagrangian of this optimization is thus

$$L(\tilde{\boldsymbol{U}}, \boldsymbol{\Lambda}) = -2\mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{U}}) + \mathrm{tr}(\boldsymbol{\Lambda}^{\mathrm{T}}(\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}} - \tilde{\boldsymbol{P}})), \tag{60}$$

where $\boldsymbol{\Lambda} \in \mathcal{M}_{n\times n}$ is the Lagrangian multiplier for the constraint $\tilde{\boldsymbol{U}}^{\mathrm{T}}\tilde{\boldsymbol{U}} = \tilde{\boldsymbol{P}}$. The global optimizer should be one of the critical points, so [7]

$$\nabla_{\tilde{\boldsymbol{U}}} L\big|_{\tilde{\boldsymbol{U}}_*,\boldsymbol{\Lambda}_*} = -2\boldsymbol{U} + 2\tilde{\boldsymbol{U}}_*\boldsymbol{\Lambda}_* = \boldsymbol{0}, \qquad \nabla_{\boldsymbol{\Lambda}} L\big|_{\tilde{\boldsymbol{U}}_*,\boldsymbol{\Lambda}_*} = \tilde{\boldsymbol{U}}_*^{\mathrm{T}}\tilde{\boldsymbol{U}}_* - \tilde{\boldsymbol{P}} = \boldsymbol{0}. \tag{61}$$

The first condition gives $\boldsymbol{U} = \tilde{\boldsymbol{U}}_*\boldsymbol{\Lambda}_*$ and the second is simply the second-moment constraint on $\tilde{\boldsymbol{U}}_*$. The symmetry of this constraint implies the symmetry of the Lagrangian multiplier $\boldsymbol{\Lambda}_*$. Inserting these results into the objective function (59), we obtain

$$-2\mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{U}}_*) = -2\mathrm{tr}(\boldsymbol{\Lambda}\tilde{\boldsymbol{U}}_*^{\mathrm{T}}\tilde{\boldsymbol{U}}_*) = -2\mathrm{tr}(\boldsymbol{\Lambda}_*\tilde{\boldsymbol{P}}) = -2\mathrm{tr}(\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{\Lambda}_*\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}). \tag{62}$$

The reason for symmetrifying the matrix in the last step will soon become clear. Since

$$\boldsymbol{P} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{\Lambda}_*\tilde{\boldsymbol{U}}_*^{\mathrm{T}}\tilde{\boldsymbol{U}}_*\boldsymbol{\Lambda}_* = \boldsymbol{\Lambda}_*\tilde{\boldsymbol{P}}\boldsymbol{\Lambda}_*,$$

the second-moment constraint reduces to $\boldsymbol{P} = \boldsymbol{\Lambda}_*\tilde{\boldsymbol{P}}\boldsymbol{\Lambda}_*$. To relate the objective function (62) to this constraint, we use a small trick:

[7] See [42] for how these matrix gradients are computed or refer to "Appendix A" of [37].

$$
\begin{aligned}
(\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{\Lambda}_*\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}})^2 &= (\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{\Lambda}_*\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}})(\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{\Lambda}_*\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}) \\
&= \sqrt{\tilde{\boldsymbol{P}}}(\boldsymbol{\Lambda}_*\tilde{\boldsymbol{P}}\boldsymbol{\Lambda}_*)\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}} \\
&= \sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}.
\end{aligned}
$$

Therefore, $\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{\Lambda}_*\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}$ must be a symmetric square root of $\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}$, which is not unique. This characterizes all the critical points of (4), among which is the desired global optimizer $\tilde{\boldsymbol{U}}_* = \boldsymbol{U}\boldsymbol{\Lambda}_*^{-1}$.

Next, we identify the global optimizer by comparing the values of the objective function evaluated at these critical points. Since $\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{\Lambda}_*\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}$ is symmetric, it has an eigendecomposition

$$
\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{\Lambda}_*\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}} = \boldsymbol{V}\,\mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)\,\boldsymbol{V}^{\mathrm{T}}
$$

and its square $\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}$ will have the corresponding eigendecomposition

$$
\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}} = \boldsymbol{V}\,\mathrm{diag}\left(\lambda_1^2, \ldots, \lambda_n^2\right)\boldsymbol{V}^{\mathrm{T}}. \tag{63}
$$

If we denote the n positive eigenvalues of $\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}$ by $\sigma_1, \ldots, \sigma_n$, then without loss of generality, we must have $\sigma_i = \lambda_i^2$ and hence $\lambda_i = \pm\sqrt{\sigma_i}$. Therefore, the objective function (62) reduces to

$$
\begin{aligned}
-2\mathrm{tr}\left(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{U}}_*\right) &= -2\mathrm{tr}\left(\boldsymbol{V}\,\mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)\,\boldsymbol{V}^{\mathrm{T}}\right) \\
&= -2\sum_{i=1}^{n} \pm\sqrt{\sigma_i} \geq -2\sum_{i=1}^{n}\sqrt{\sigma_i}.
\end{aligned} \tag{64}
$$

We can see that the lower bound is attained if and only if $\lambda_i = \sqrt{\sigma_i}$, $i = 1, \ldots, n$, i.e. when $\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{\Lambda}_*\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}$ takes the unique SPD square root:

$$
\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{\Lambda}_*\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}} = (\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}})^{1/2}. \tag{65}
$$

Therefore, the global minimizer to (4) is $\tilde{\boldsymbol{U}}_* = \boldsymbol{U}\boldsymbol{\Lambda}_*^{-1}$ with

$$
\boldsymbol{\Lambda}_*^{-1} = \sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}(\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}})^{-1/2}\sqrt{\tilde{\boldsymbol{P}}}. \tag{66}
$$

Note that the particular choice of square root $\sqrt{\tilde{\boldsymbol{P}}}$ is irrelevant because both the spectrum of $\sqrt{\tilde{\boldsymbol{P}}}\boldsymbol{P}\sqrt{\tilde{\boldsymbol{P}}}^{\mathrm{T}}$ and the global optimizer (66) are invariant under the unitary freedom $\sqrt{\tilde{\boldsymbol{P}}} \to \boldsymbol{Q}\sqrt{\tilde{\boldsymbol{P}}}$, due to the fact that $\boldsymbol{Q}\boldsymbol{A}^{1/2}\boldsymbol{Q}^{\mathrm{T}} = (\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^{\mathrm{T}})^{1/2}$ for any SPD $\boldsymbol{A}$.

We remark that an alternative to using (62) is to use

$$-2\mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\tilde{\boldsymbol{U}}_*) = -2\mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Lambda}_*^{-1}) = -2\mathrm{tr}\left(\boldsymbol{P}\boldsymbol{\Lambda}_*^{-1}\right) = -2\mathrm{tr}(\sqrt{\boldsymbol{P}}\boldsymbol{\Lambda}_*^{-1}\sqrt{\boldsymbol{P}}^{\mathrm{T}})$$

and correspondingly $\tilde{\boldsymbol{P}} = \boldsymbol{\Lambda}_*^{-1}\boldsymbol{P}\boldsymbol{\Lambda}_*^{-1}$. This simply switches the role of $\boldsymbol{P}$ and $\tilde{\boldsymbol{P}}$ and replaces $\boldsymbol{\Lambda}_*$ with $\boldsymbol{\Lambda}_*^{-1}$. Therefore, this gives an alternative expression of $\boldsymbol{\Lambda}_*^{-1}$ obtained in (66), i.e.

$$\boldsymbol{\Lambda}_*^{-1} = \sqrt{\boldsymbol{P}}^{-1}(\sqrt{\boldsymbol{P}}\tilde{\boldsymbol{P}}\sqrt{\boldsymbol{P}}^{\mathrm{T}})^{1/2}\sqrt{\boldsymbol{P}}^{-\mathrm{T}}. \tag{67}$$

□

B: Generalization from $\mathbb{R}^m$ to a Hilbert space

In Theorem 2.1, each column of $\boldsymbol{U}$ or $\tilde{\boldsymbol{U}}$ is an element in $\mathbb{R}^m$ and the second-moment matrices are computed based on an inner product on $\mathbb{R}^m$. A natural question is whether Theorem 2.1 still holds when $\mathbb{R}^m$ is replaced by an infinite-dimensional Hilbert space $\mathcal{H}$ equipped with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$.

More precisely, assume $u_1, \ldots, u_n \in \mathcal{H}$ are linearly independent and form an n-vector of $\mathcal{H}$ denoted by $\boldsymbol{u} = [u_1, \ldots, u_n]^{\mathrm{T}} \in \mathcal{H}^n$. Then we can compute the second-moment matrix (a.k.a. the Gram matrix) of $\boldsymbol{u}$ as $\boldsymbol{P} = \mathrm{Gr}(\boldsymbol{u}, \boldsymbol{u})$, where

$$\mathrm{Gr}(\boldsymbol{u}, \boldsymbol{v}) \triangleq \left\langle \boldsymbol{u}, \boldsymbol{v}^{\mathrm{T}} \right\rangle_{\mathcal{H}} \in \mathcal{M}_{n\times n} \tag{68}$$

i.e. with the (i, j)-th element of $\mathrm{Gr}(\boldsymbol{u}, \boldsymbol{v})$ being $\langle u_i, v_j\rangle_{\mathcal{H}}$. If $\mathcal{H}^n$ is equipped with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}^n}$ defined as

$$\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\mathcal{H}^n} \triangleq \mathrm{tr}\left(\sqrt{\boldsymbol{\Gamma}_n}\mathrm{Gr}(\boldsymbol{u}, \boldsymbol{v})\sqrt{\boldsymbol{\Gamma}_n}^{\mathrm{T}}\right) \tag{69}$$

with again an SPD weight matrix $\boldsymbol{\Gamma}_n$, we want to solve the following minimum-correction second-moment matching problem

$$\underset{\tilde{\boldsymbol{u}}\in\mathcal{H}^n:\mathrm{Gr}(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{u}})=\tilde{\boldsymbol{P}}}{\arg\min} \|\tilde{\boldsymbol{u}} - \boldsymbol{u}\|_{\mathcal{H}^n}^2 \tag{70}$$

for a given SPD $\tilde{\boldsymbol{P}} \in \mathcal{M}_{n\times n}$. Here the norm is induced by the inner product (69). As a common scenario, if $\mathcal{H} = L_2(\Omega)$, which is the space of square-integrable real functions on some measure space $(\Omega, \mathcal{A}, m)$ equipped with inner product $\langle u(\cdot), v(\cdot)\rangle_{\mathcal{H}} = \int_\Omega u(\omega)v(\omega)\mathrm{d}m(\omega)$, then the inner product (69) on $\mathcal{H}^n$ can be rewritten as

$$\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\mathcal{H}^n} = \int_\Omega \boldsymbol{u}(\omega)^{\mathrm{T}}\boldsymbol{\Gamma}_n\boldsymbol{v}(\omega)\mathrm{d}m(\omega). \tag{71}$$

The definitions (69) and (71) are simply the Hilbert space counterparts of the two expressions in (2). Since we can get rid of $\boldsymbol{\Gamma}_n$ by scaling the elements in $\mathcal{H}^n$ by $\sqrt{\boldsymbol{\Gamma}_n}$ as before, we will assume $\boldsymbol{\Gamma}_n = \boldsymbol{I}$ hereafter.

Theorem 2.1 does not apply to (70), but as in Sect. 2.2, if we restrict the candidate set to $\{\tilde{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{u} : \boldsymbol{A} \in \mathcal{M}_{n\times n}\}$, the Hilbert space complication will be confined to computing the Gram matrices only. Since

$$\mathrm{Gr}(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{B}\boldsymbol{v}) = \left\langle \boldsymbol{A}\boldsymbol{u}, (\boldsymbol{B}\boldsymbol{v})^{\mathrm{T}} \right\rangle_{\mathcal{H}} = \boldsymbol{A} \left\langle \boldsymbol{u}, \boldsymbol{v}^{\mathrm{T}} \right\rangle_{\mathcal{H}} \boldsymbol{B}^{\mathrm{T}} = \boldsymbol{A}\mathrm{Gr}(\boldsymbol{u}, \boldsymbol{v})\boldsymbol{B}^{\mathrm{T}}$$

due to the bilinearity of an inner product, we have

$$\|\tilde{\boldsymbol{u}} - \boldsymbol{u}\|_{\mathcal{H}^n}^2 = \|(\boldsymbol{A} - \boldsymbol{I})\boldsymbol{u}\|_{\mathcal{H}^n}^2 = \mathrm{tr}\left((\boldsymbol{A} - \boldsymbol{I})\mathrm{Gr}(\boldsymbol{u}, \boldsymbol{u})(\boldsymbol{A} - \boldsymbol{I})^{\mathrm{T}}\right) = \|\boldsymbol{A} - \boldsymbol{I}\|_{\mathrm{F},\boldsymbol{P}}^2 .$$

Hence (70) reduces to (16) and Corollary 2.2 applies.

To establish the analog of Theorem 2.1 for (70), we need some techniques from the calculus of variation to generalize optimization in $\mathbb{R}^m$ to that in a Hilbert space.

Theorem B.1 (Minimum-correction 2nd-moment matching on a Hilbert space) *Given $\boldsymbol{u} \in \mathcal{H}^n$ whose entries are linearly independent and $\tilde{\boldsymbol{P}}$ which is SPD, we have*

$$\underset{\tilde{\boldsymbol{u}}\in\mathcal{H}^n:\mathrm{Gr}(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{u}})=\tilde{\boldsymbol{P}}}{\arg\min} \|\tilde{\boldsymbol{u}} - \boldsymbol{u}\|_{\mathcal{H}^n}^2 = \boldsymbol{A}_*\boldsymbol{u} \tag{72}$$

with $\boldsymbol{A}_*$ given by (7).

Proof Compared to the case in $\mathbb{R}^m$, although $\mathcal{S}_{\tilde{\boldsymbol{P}}} = \{\tilde{\boldsymbol{u}} \in \mathcal{H}^n : \mathrm{Gr}(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}) = \tilde{\boldsymbol{P}}\}$ is still closed and bounded ($\|\tilde{\boldsymbol{u}}\|_{\mathcal{H}^n}^2 \equiv \mathrm{tr}(\tilde{\boldsymbol{P}})$ on $\mathcal{S}_{\tilde{\boldsymbol{P}}}$), $\mathcal{S}_{\tilde{\boldsymbol{P}}}$ is no longer compact due to the infinite dimensionality. Therefore, although the objective function is continuous and bounded from below, its infimum may not be attainable. In the following, we first assume that a global minimizer $\tilde{\boldsymbol{u}}_*$ exists and derive an analytic expression for it. After that, we will show that this expression, which is well-defined whether or not a global minimum exists, is indeed the unique global minimizer to (70).

Here the objective function is

$$F(\tilde{\boldsymbol{u}}) = \|\tilde{\boldsymbol{u}} - \boldsymbol{u}\|_{\mathcal{H}^n}^2 = \mathrm{tr}\left(\mathrm{Gr}(\tilde{\boldsymbol{u}} - \boldsymbol{u}, \tilde{\boldsymbol{u}} - \boldsymbol{u})\right) = \mathrm{tr}(\tilde{\boldsymbol{P}}) + \mathrm{tr}(\boldsymbol{P}) - 2\mathrm{tr}\left(\mathrm{Gr}(\tilde{\boldsymbol{u}}, \boldsymbol{u})\right),$$

so its first-order variation is

$$\delta F(\tilde{\boldsymbol{u}}, \delta\tilde{\boldsymbol{u}}) = -2\mathrm{tr}\left(\mathrm{Gr}(\delta\tilde{\boldsymbol{u}}, \boldsymbol{u})\right).$$

The constraint on $\tilde{\boldsymbol{u}}$ is $\mathrm{Gr}(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}) = \tilde{\boldsymbol{P}}$, so the constraint on $\delta\tilde{\boldsymbol{u}}$ is

$$\mathrm{Gr}(\delta\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}) + \mathrm{Gr}(\tilde{\boldsymbol{u}}, \delta\tilde{\boldsymbol{u}}) = 0, \tag{73}$$

which is simply saying that $\mathrm{Gr}(\delta\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}})$ is anti-symmetric. If $\tilde{\boldsymbol{u}}$ is a local minimizer to (72), we must have $\delta F(\tilde{\boldsymbol{u}}, \delta\tilde{\boldsymbol{u}}) \equiv 0$ for any $\delta\tilde{\boldsymbol{u}} \in \mathcal{H}^n$ that satisfies (73).

Now we prove by contradiction that each entry of $\boldsymbol{u}$ must be a linear combination of the entries of a local minimizer $\tilde{\boldsymbol{u}}$, which means $\boldsymbol{u} = \boldsymbol{A}\tilde{\boldsymbol{u}}$ for some $\boldsymbol{A} \in \mathcal{M}_{n\times n}$. If this was not the case, without loss of generality, we can assume $u_1 \notin \text{span}\{\tilde{u}_1, \ldots, \tilde{u}_n\} \subset \mathcal{H}$. Therefore, the projection of u_1 onto $\text{span}\{\tilde{u}_1, \ldots, \tilde{u}_n\}^{\perp}$ must be nonzero and we denote it by v. Now we construct $\delta\tilde{\boldsymbol{u}} = [v, 0, \ldots, 0]^{\mathrm{T}}$. We can see that $\text{Gr}(\delta\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}) = 0$ because $\langle v, \tilde{u}_i\rangle_{\mathcal{H}} = 0,\ i = 1, \ldots, n$. Hence, (73) is satisfied. On the other hand, since $\langle v, u_1\rangle_{\mathcal{H}} = \|v\|_{\mathcal{H}}^2 > 0$, we have

$$\delta F(\tilde{\boldsymbol{u}}, \delta\tilde{\boldsymbol{u}}) = -2\text{tr}\left(\text{Gr}(\delta\tilde{\boldsymbol{u}}, \boldsymbol{u})\right) = -2\sum_{i=1}^{n}\langle\delta\tilde{u}_i, u_i\rangle_{\mathcal{H}} = -2\langle v, u_1\rangle_{\mathcal{H}} < 0,$$

which contradicts the assumption that $\tilde{\boldsymbol{u}}$ is a local minimizer. Therefore, any local minimizer $\tilde{\boldsymbol{u}}$ must satisfy $\boldsymbol{u} = \boldsymbol{A}\tilde{\boldsymbol{u}}$ for some $\boldsymbol{A} \in \mathcal{M}_{n\times n}$. Since the entries of $\boldsymbol{u}$ are linearly independent, $\boldsymbol{A}$ must be invertible and thus $\tilde{\boldsymbol{u}} = \boldsymbol{A}^{-1}\boldsymbol{u}$.

Since we have already shown that with the candidate set restricted to $\{\tilde{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{u}\}$, Corollary 2.2 applies. This completes the proof of (72) when a global minimum exists.

Note that no matter whether a global minimum exists or not, we can always construct $\tilde{\boldsymbol{u}}_* = \boldsymbol{A}_*\boldsymbol{u} \in \mathcal{S}_{\tilde{\boldsymbol{P}}}$ with $\boldsymbol{A}_*$ given by (7). Therefore, if we can show $\|\tilde{\boldsymbol{u}}_* - \boldsymbol{u}\|_{\mathcal{H}^n}$ is indeed a lower bound of $\|\tilde{\boldsymbol{u}} - \boldsymbol{u}\|_{\mathcal{H}^n}$ on $\mathcal{S}_{\tilde{\boldsymbol{P}}}$, we can prove the existence and uniqueness of a global minimizer.

Since $F(\tilde{\boldsymbol{u}}) - \text{tr}(\tilde{\boldsymbol{P}}) - \text{tr}(\boldsymbol{P}) = -2\text{tr}\left(\text{Gr}(\tilde{\boldsymbol{u}}, \boldsymbol{u})\right) = -2\langle\tilde{\boldsymbol{u}}, \boldsymbol{u}\rangle_{\mathcal{H}^n}$, we want to show $\langle\tilde{\boldsymbol{u}}, \boldsymbol{u}\rangle_{\mathcal{H}^n} \le \langle\tilde{\boldsymbol{u}}_*, \boldsymbol{u}\rangle_{\mathcal{H}^n}$ for any $\tilde{\boldsymbol{u}} \in \mathcal{S}_{\tilde{\boldsymbol{P}}}$ by the Cauchy-Schwarz inequality. Since $\tilde{\boldsymbol{u}}_*$ is not parallel to $\boldsymbol{u}$ in general, we need two linear transforms $\tilde{\boldsymbol{u}} = \boldsymbol{A}_*^{1/2}\tilde{\boldsymbol{v}}$ and $\boldsymbol{u} = \boldsymbol{A}_*^{-1/2}\boldsymbol{v}$ to make sure that the Cauchy-Schwarz inequality attains its equal sign when $\tilde{\boldsymbol{u}} = \boldsymbol{A}_*\boldsymbol{u}$. Therefore, we have

$$\langle\tilde{\boldsymbol{u}}, \boldsymbol{u}\rangle_{\mathcal{H}^n} = \text{tr}\left(\boldsymbol{A}_*^{1/2}\left\langle\tilde{\boldsymbol{v}}, \boldsymbol{v}^{\mathrm{T}}\right\rangle_{\mathcal{H}}\boldsymbol{A}_*^{-1/2}\right) = \text{tr}\left(\left\langle\tilde{\boldsymbol{v}}, \boldsymbol{v}^{\mathrm{T}}\right\rangle_{\mathcal{H}}\right) = \langle\tilde{\boldsymbol{v}}, \boldsymbol{v}\rangle_{\mathcal{H}^n} \le \|\tilde{\boldsymbol{v}}\|_{\mathcal{H}^n}\|\boldsymbol{v}\|_{\mathcal{H}^n},$$

where the last step is the Cauchy-Schwarz inequality. On the other hand, we have

$$\begin{aligned}
\|\boldsymbol{v}\|_{\mathcal{H}^n}^2 &= \left\langle\boldsymbol{A}_*^{1/2}\boldsymbol{u}, \boldsymbol{A}_*^{1/2}\boldsymbol{u}\right\rangle_{\mathcal{H}^n} = \text{tr}\left(\boldsymbol{A}_*^{1/2}\left\langle\tilde{\boldsymbol{u}}, \boldsymbol{u}^{\mathrm{T}}\right\rangle_{\mathcal{H}}\boldsymbol{A}_*^{1/2}\right) = \text{tr}\left(\boldsymbol{A}_*\boldsymbol{P}\right),\\
\|\tilde{\boldsymbol{v}}\|_{\mathcal{H}^n}^2 &= \left\langle\boldsymbol{A}_*^{-1/2}\tilde{\boldsymbol{u}}, \boldsymbol{A}_*^{-1/2}\tilde{\boldsymbol{u}}\right\rangle_{\mathcal{H}^n} = \text{tr}\left(\tilde{\boldsymbol{P}}\boldsymbol{A}_*^{-1}\right) = \text{tr}\left(\boldsymbol{A}_*\boldsymbol{P}\right),\\
\langle\tilde{\boldsymbol{u}}_*, \boldsymbol{u}\rangle_{\mathcal{H}^n} &= \langle\boldsymbol{A}_*\boldsymbol{u}, \boldsymbol{u}\rangle_{\mathcal{H}^n} = \text{tr}\left(\boldsymbol{A}_*\boldsymbol{P}\right),
\end{aligned}$$

where the last equal sign in the second line is due to the second-moment matching constraint $\boldsymbol{A}_*\boldsymbol{P}\boldsymbol{A}_* = \tilde{\boldsymbol{P}}$. Combining the above results, we have shown

$$\langle\tilde{\boldsymbol{u}}, \boldsymbol{u}\rangle_{\mathcal{H}^n} \le \text{tr}\left(\boldsymbol{A}_*\boldsymbol{P}\right) = \langle\tilde{\boldsymbol{u}}_*, \boldsymbol{u}\rangle_{\mathcal{H}^n}$$

for any $\tilde{\boldsymbol{u}} \in \mathcal{S}_{\tilde{\boldsymbol{P}}}$, which completes the proof. □

Theorem 2.2 can also be generalized to the case with $\mathbb{R}^m$ replaced by a Hilbert space $\mathcal{H}$. This continuous setup can be intuitively viewed as the discrete one with

$m = \infty$, which implies that we always have $m > n \geq \tilde{r}$ so the modification in step 1 is never needed. We state this generalization in the following theorem, whose proof is omitted since it is totally in parallel with that of Theorem 2.2.

Theorem B.2 (Degenerate counterpart of Theorem B.1) *Given* $\boldsymbol{u} \in \mathcal{H}^n$ *with* $\boldsymbol{P} = \mathrm{Gr}(\boldsymbol{u}, \boldsymbol{u}) \in \mathcal{M}_{n\times n}$ *and* $\mathrm{rank}(\boldsymbol{P}) = \dim(\mathrm{span}\{u_1, \ldots, u_n\}) = r$, *and semi-SPD* $\tilde{\boldsymbol{P}} \in \mathcal{M}_{n\times n}$ *with* $\mathrm{rank}(\tilde{\boldsymbol{P}}) = \tilde{r}$, *we have*

$$\underset{\tilde{\boldsymbol{u}}\in\mathcal{H}^n:\mathrm{Gr}(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{u}})=\tilde{\boldsymbol{P}}}{\arg\min} \|\tilde{\boldsymbol{u}} - \boldsymbol{u}\|^2_{\mathcal{H}^n} = \tilde{\boldsymbol{u}}_* = \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{s}}_* = \tilde{\boldsymbol{Z}}(\boldsymbol{Z}\tilde{\boldsymbol{w}}_* + \boldsymbol{Z}_\perp\tilde{\boldsymbol{w}}_{\perp *}) \tag{74}$$

where $\tilde{\boldsymbol{s}}_* = \boldsymbol{Z}\tilde{\boldsymbol{w}}_* + \boldsymbol{Z}_\perp\tilde{\boldsymbol{w}}_{\perp *} \in \mathcal{H}^{\tilde{r}}$. *Moreover, the columns of* $\tilde{\boldsymbol{Z}} \in \mathcal{M}_{n\times\tilde{r}}$ *form an orthonormal basis of* $\mathrm{Row}(\tilde{\boldsymbol{P}})$ *and the columns of* $\boldsymbol{Z} \in \mathcal{M}_{\tilde{r}\times r'}$ *form an orthonormal basis of* $\mathrm{Row}(\tilde{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{P}\tilde{\boldsymbol{Z}})$ *with*

$$r' = \tilde{r} - \dim(\mathrm{Row}(\tilde{\boldsymbol{P}}) \cap \mathrm{Row}(\boldsymbol{P})^\perp) = r - \dim(\mathrm{Row}(\boldsymbol{P}) \cap \mathrm{Row}(\tilde{\boldsymbol{P}})^\perp). \tag{75}$$

Finally, $\tilde{\boldsymbol{w}}_*$ *and* $\tilde{\boldsymbol{w}}_{\perp *}$ *are given by*

$$\tilde{\boldsymbol{w}}_* = \underset{\tilde{\boldsymbol{w}}\in\mathcal{H}^{r'}:\ \mathrm{Gr}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{w}})=\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}\boldsymbol{Z}}{\arg\min} \|\tilde{\boldsymbol{w}} - \boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}\boldsymbol{u}\|^2_{\mathcal{H}^{r'}} \tag{76}$$

and $\tilde{\boldsymbol{w}}_{\perp *} = \boldsymbol{B}_{12}^{\mathrm{T}}\boldsymbol{B}_{11}^{-1}\tilde{\boldsymbol{w}}_* + \boldsymbol{v}_* \in \mathcal{H}^{\tilde{r}-r'}$ *with arbitrary* $\boldsymbol{v}_*$ *such that* $\mathrm{Gr}(\boldsymbol{v}_*, \tilde{\boldsymbol{w}}_*) = \boldsymbol{0}$, *i.e.* $v_{*,i} \in \mathrm{span}\{\tilde{w}_{*,1}, \ldots, \tilde{w}_{*,r'}\}^\perp \subset \mathcal{H}$ *for* $i = 1, \ldots, (\tilde{r} - r')$, *and* $\mathrm{Gr}(\boldsymbol{v}_*, \boldsymbol{v}_*) = \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^{\mathrm{T}}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}$. *The* $\boldsymbol{B}_{ij}$ *blocks are defined as:*

$$\begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{12}^{\mathrm{T}} & \boldsymbol{B}_{22} \end{bmatrix} \triangleq \mathrm{Gr}\left(\begin{bmatrix} \tilde{\boldsymbol{w}} \\ \tilde{\boldsymbol{w}}_\perp \end{bmatrix}, \begin{bmatrix} \tilde{\boldsymbol{w}} \\ \tilde{\boldsymbol{w}}_\perp \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{Z}^{\mathrm{T}} \\ \boldsymbol{Z}_\perp^{\mathrm{T}} \end{bmatrix} \tilde{\boldsymbol{Z}}^{\mathrm{T}}\tilde{\boldsymbol{P}}\tilde{\boldsymbol{Z}}\left[\boldsymbol{Z}, \boldsymbol{Z}_\perp\right].$$

Besides, the global minimum of the objective function is

$$\|\tilde{\boldsymbol{u}}_* - \boldsymbol{u}\|^2_{\mathcal{H}^n} = \mathrm{tr}[\boldsymbol{P} + \tilde{\boldsymbol{P}} - 2(\boldsymbol{P}\tilde{\boldsymbol{P}})^{1/2}]. \tag{77}$$

All previous remarks regarding Theorem 2.2 apply here as well. Moreover, this theorem also extends our optimal transport interpretation in "Appendix C" to the most general case with arbitrary semi-SPD covariance matrix $\boldsymbol{\Sigma}_\mu$ and $\boldsymbol{\Sigma}_\nu$, unlike most previous literature (e.g. [12,40]) on this topic, which essentially restrict themselves in cases with $\mathrm{Row}(\boldsymbol{\Sigma}_\nu) \subset \mathrm{Row}(\boldsymbol{\Sigma}_\mu)$. In particular, Theorem B.2 implies that the optimal coupling π_* between μ and ν (which is a joint distribution with μ and ν being its marginals) for matching only the first two moments is deterministic (i.e. corresponds to a bijection between $\boldsymbol{u} \sim \mu$ and $\tilde{\boldsymbol{u}} \sim \nu$) if and only if $\mathrm{rank}(\boldsymbol{\Sigma}_\mu) = \mathrm{rank}(\boldsymbol{\Sigma}_\nu) = \mathrm{rank}(\boldsymbol{\Sigma}_\mu\boldsymbol{\Sigma}_\nu)$. The optimal coupling is deterministic in the direction from μ to ν (i.e. π_* induces a map from $\boldsymbol{u}$ to $\tilde{\boldsymbol{u}}$ or the conditional $\pi_{\nu|\mu}$ is always a Dirac delta) if and only if $\mathrm{Row}(\boldsymbol{\Sigma}_\nu) \cap \mathrm{Row}(\boldsymbol{\Sigma}_\mu)^\perp = 0$, of which the assumption $\mathrm{Row}(\boldsymbol{\Sigma}_\nu) \subset \mathrm{Row}(\boldsymbol{\Sigma}_\mu)$ usually made in previous literature is a special case.

C: Connections to optimal transport

Our main results Theorems 2.1 and B.1 are also closely connected to the problem of optimal transport (a.k.a. the Monge-Kantorovich minimization problem [55, p. 10]), which concerns the coupling (in the form of a joint distribution) between two given probability distributions that minimizes an associated transport cost. More precisely, in a particular setting that is relevant to our context, given two PDFs $\mu(\cdot)$ and $\nu(\cdot)$ with finite second moments over $\mathbb{R}^n$, the problem concerns the joint distribution $\pi(\cdot,\cdot)$ over $\mathbb{R}^n \times \mathbb{R}^n$, which can be degenerate, such that π has μ and ν as marginals and the L_2 distance between the two marginal random variables are minimized, i.e.

$$\pi_* = \underset{\pi \in \Pi_{\mu,\nu}}{\arg\min} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}, \tag{78}$$

where $\|\cdot\|$ is the Euclidean distance in $\mathbb{R}^n$ and

$$\Pi_{\mu,\nu} = \left\{ \pi(\cdot,\cdot) \in \text{PDFs} : \int_{\mathbb{R}^n} \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} = \mu(\boldsymbol{x}), \int_{\mathbb{R}^n} \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} = \nu(\boldsymbol{y}) \right\}. \tag{79}$$

Under the above conditions, a unique optimizer π_* exists [55, p. 11] and it turns out to be deterministic, i.e. $\pi_*(\boldsymbol{x}, \boldsymbol{y}) = \delta_{\boldsymbol{y}=\boldsymbol{T}_*(\boldsymbol{x})}\mu(\boldsymbol{x})$ for some mapping $\boldsymbol{T}_* : \mathbb{R}^n \to \mathbb{R}^n$ such that $\int \pi_*(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} = \nu(\boldsymbol{y})$. Moreover, the quantity

$$W_2(\mu, \nu) = \min_{\pi \in \Pi_{\mu,\nu}} \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \right)^{1/2} \tag{80}$$

is called the L_2 Wasserstein distance between μ and ν [55, ch. 6].

In general, the optimal transport mapping $\boldsymbol{T}_*(\cdot)$ is nonlinear. However, if we relax the constraint of marginal matching to only first- and second-moment matching, Theorem B.1 implies that $\boldsymbol{T}_*(\cdot)$ is indeed affine with the linear part being SPD, even when μ and ν do not have a well-defined PDF (i.e. not absolutely continuous with respect to the Lebesgue measure). To show this, first notice that the optimization (78) over joint distributions on $\mathbb{R}^n \times \mathbb{R}^n$ can be translated into an equivalent problem over random vectors in $\mathbb{R}^n$:

$$\tilde{\boldsymbol{u}}_* = \underset{\tilde{\boldsymbol{u}} \sim \nu}{\arg\min}\, \mathbb{E}\left[\|\tilde{\boldsymbol{u}} - \boldsymbol{u}\|^2 \right] \tag{81}$$

with some $\boldsymbol{u} \sim \mu$. Next, denote the mean of μ and ν by $\boldsymbol{m}_\mu$ and $\boldsymbol{m}_\nu$, respectively, and also the covariance matrices by $\boldsymbol{\Sigma}_\mu$ and $\boldsymbol{\Sigma}_\nu$. Then the objective function in (81) becomes

$$\mathbb{E}\left[\|\tilde{\boldsymbol{u}} - \boldsymbol{u}\|^2 \right] = \|\boldsymbol{m}_\mu - \boldsymbol{m}_\nu\|^2 + \mathbb{E}\left[\|\tilde{\boldsymbol{u}}' - \boldsymbol{u}'\|^2 \right],$$

where the primes indicate the centered random vectors. Now if we relax the distribution matching constraint $\tilde{\boldsymbol{u}} \sim \nu$ to only mean and covariance matching, i.e. $\mathbb{E}[\tilde{\boldsymbol{u}}] = \boldsymbol{m}_\nu$

and $\mathbb{E}[\tilde{\boldsymbol{u}}'\tilde{\boldsymbol{u}}'^{\mathrm{T}}] = \boldsymbol{\Sigma}_\nu$, the optimal transport problem (81) is simplified to the task of minimum-correction second-moment matching:

$$\tilde{\boldsymbol{u}}'_* = \underset{\mathbb{E}[\tilde{\boldsymbol{u}}'\tilde{\boldsymbol{u}}'^{\mathrm{T}}]=\boldsymbol{\Sigma}_\nu}{\arg\min} \ \mathbb{E}\left[\|\tilde{\boldsymbol{u}} - \boldsymbol{u}\|^2\right]. \tag{82}$$

If we set $\mathcal{H} = L_2(\Omega)$ as the Hilbert space of all square-integrable random variables defined over the probability space $(\Omega, \mathcal{A}, \mathcal{P})$ on which $\boldsymbol{u}$ is defined, then Theorem B.1 applies to (82) and yields $\tilde{\boldsymbol{u}}_* = \boldsymbol{T}_*(\boldsymbol{u}) = \boldsymbol{A}_*(\boldsymbol{u} - \boldsymbol{m}_\mu) + \boldsymbol{m}_\nu$ with

$$\boldsymbol{A}_* = \sqrt{\boldsymbol{\Sigma}_\mu}^{-1}(\sqrt{\boldsymbol{\Sigma}_\mu}\boldsymbol{\Sigma}_\nu\sqrt{\boldsymbol{\Sigma}_\mu}^{\mathrm{T}})^{1/2}\sqrt{\boldsymbol{\Sigma}_\mu}^{-\mathrm{T}}. \tag{83}$$

Moreover, since

$$\mathbb{E}\left[\|\tilde{\boldsymbol{u}}' - \boldsymbol{u}'\|^2\right] = \mathrm{tr}(\boldsymbol{\Sigma}_\mu) + \mathrm{tr}(\boldsymbol{\Sigma}_\nu) - 2\mathrm{tr}[\mathrm{Cov}(\tilde{\boldsymbol{u}}, \boldsymbol{u})] \tag{84}$$

and

$$\mathrm{tr}[\mathrm{Cov}(\tilde{\boldsymbol{u}}_*, \boldsymbol{u})] = \mathrm{tr}[\mathrm{Cov}(\boldsymbol{A}_*\boldsymbol{u}', \boldsymbol{u}')] = \mathrm{tr}[(\sqrt{\boldsymbol{\Sigma}_\mu}\boldsymbol{\Sigma}_\nu\sqrt{\boldsymbol{\Sigma}_\mu}^{\mathrm{T}})^{1/2}] = \mathrm{tr}[(\boldsymbol{\Sigma}_\mu\boldsymbol{\Sigma}_\nu)^{1/2}],$$

we have

$$W_2^2(\mu, \boldsymbol{T}_*\mu) = \|\boldsymbol{m}_\mu - \boldsymbol{m}_\nu\|^2 + \mathrm{tr}[\boldsymbol{\Sigma}_\mu + \boldsymbol{\Sigma}_\nu - 2(\boldsymbol{\Sigma}_\mu\boldsymbol{\Sigma}_\nu)^{1/2}], \tag{85}$$

where the distribution $\boldsymbol{T}_*\mu$ is the push-forward of μ by $\boldsymbol{T}_*$. Note that although $\boldsymbol{T}_*\mu$ share the same mean and covariance with ν, in general $\boldsymbol{T}_*\mu \neq \nu$ and $W_2(\mu, \boldsymbol{T}_*\mu) < W_2(\mu, \nu)$, so (85) indeed provides a lower bound for $W_2(\mu, \nu)$ based on the first two moments of μ and ν.

If we partition the set of all distributions (possibly degenerate) over $\mathbb{R}^n$ that have finite second moments into equivalent classes by their means and covariances then (85) induces a metric on this quotient space. Denote by $\mathcal{S}_{\boldsymbol{m},\boldsymbol{\Sigma}}$ the equivalent class of all distributions with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$. Given $\mathcal{S}_{\boldsymbol{m}_2,\boldsymbol{\Sigma}_2}$ and $\mu \in \mathcal{S}_{\boldsymbol{m}_1,\boldsymbol{\Sigma}_1}$, there exists a unique $\tilde{\mu} = \boldsymbol{T}_*\mu \in \mathcal{S}_{\boldsymbol{m}_2,\boldsymbol{\Sigma}_2}$ that is the closest to μ under the distance $W_2(\cdot,\cdot)$. As a generalization to the one-to-one correspondence of (9) and the intuition illustrated in Fig. 1, the members in any two equivalent classes $\mathcal{S}_{\boldsymbol{m}_1,\boldsymbol{\Sigma}_1}$ and $\mathcal{S}_{\boldsymbol{m}_2,\boldsymbol{\Sigma}_2}$ can be paired up by such minimum-W_2-distance matching. Moreover, since this distance between μ and $\boldsymbol{T}_*\mu$ depends only on $\mathcal{S}_{\boldsymbol{m}_1,\boldsymbol{\Sigma}_1}$ and $\mathcal{S}_{\boldsymbol{m}_2,\boldsymbol{\Sigma}_2}$ but not on the particular member μ, $W_2(\mu, \boldsymbol{T}_*\mu)$ can be used to define the distance between $\mathcal{S}_{\boldsymbol{m}_1,\boldsymbol{\Sigma}_1}$ and $\mathcal{S}_{\boldsymbol{m}_2,\boldsymbol{\Sigma}_2}$. This is sometimes called the Fréchet distance [12]. Note that the covariance part of (85) can also be isolated and used as a metric among SPD matrices.

Although in general (85) is only a lower bound for $W_2(\mu, \nu)$, if both μ and ν belong to a family of distributions closed under affine transforms and each member of this family is uniquely determined by its mean and covariance, then we have $\boldsymbol{T}_*\mu = \nu$ and thus $W_2(\mu, \nu)$ coincides with (85). [12] mentions one typical example of such where the distribution family is characterized by a radial kernel $\phi : [0, \infty) \to [0, \infty)$

that satisfies $0 < \int_0^\infty r^{2n+1}\phi(r^2)\mathrm{d}r < \infty$, which guarantees the finiteness of second moments. In this family, a member μ's PDF takes the form $\mu(\boldsymbol{x}) \propto \phi((\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{m}))$ with some SPD $\boldsymbol{A}$, where $\boldsymbol{m}$ is the mean and $\boldsymbol{A}^{-1}$ is a multiple of the covariance. In other words, this family is generated by pushing forward a non-degenerate elliptically contoured distribution that has finite second moments by invertible affine transforms. It can be checked that within such a family, a distribution can be uniquely identified by its mean and covariance. Therefore, given μ and ν, we always have $\boldsymbol{T}_*\mu = \nu$ and thus $W_2(\mu, \nu)$ can be computed by (85). The most familiar example of such is probably the family of Gaussian distributions corresponding to the exponential kernel $\phi(r) = e^{-r}$, for which the above results are well known.

To the best of our knowledge, the earliest work that derives the right hand side expression in (85) is in [12,40]. However, they did not connect this result to the problem of optimal transport. Moreover, they translate the optimization into one over the cross-covariance matrix $\boldsymbol{B} = \mathrm{Cov}(\boldsymbol{u}, \tilde{\boldsymbol{u}})$, which does not necessarily uniquely determine $\tilde{\boldsymbol{u}}$ given $\boldsymbol{u}$ if the optimizer $\boldsymbol{B}_*$ does not correspond to a deterministic affine coupling between the two. In Theorem B.1, although the objective function is also simplified to $\mathrm{tr}(\mathrm{Cov}(\boldsymbol{u}, \tilde{\boldsymbol{u}}))$ (or its empirical counterpart (59) in the discrete case of Theorem 2.1), we have provided an alternative proof based on optimizing on $\tilde{\boldsymbol{u}}$ directly rather than on $\boldsymbol{B} = \mathrm{Cov}(\boldsymbol{u}, \tilde{\boldsymbol{u}})$.

Finally, note also that the more general result Theorem B.1 on a Hilbert space indeed contains the special case on $\mathbb{R}^m$ described by Theorem 2.1. Moreover, if we view Theorem 2.1 through the lens of Theorem B.1, we can identify a discrete counterpart of the optimal transport interpretation for Theorem 2.1. The key observation is that Theorem 2.1 is equivalent to Theorem B.1 with $\mathcal{H} = \mathbb{R}^m$, which on the other hand, corresponds to (82) with μ being a discrete distribution over $\mathbb{R}^n$ with m particles, i.e. with its (generalized) PDF being the sum of m Dirac deltas. Therefore, the optimal transport interpretation of Theorem B.1 implies that $\tilde{\boldsymbol{u}}_*$ also follows a discrete distribution with its m particles given by the rows of $\tilde{\boldsymbol{U}}_*$ in Theorem 2.1. This is actually a stronger result than Theorem 2.1 because in the latter, we have restricted the feasible $\tilde{\mu}$'s within the m-particle discrete distributions a priori.

For an exposition of other connections between the optimal transport and the polar decomposition, see [4,7].

D: Gradient descent for matrix square root

[18] proposes to solve the equation $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X} = \boldsymbol{I}$ for an $\boldsymbol{X}$ close to $\boldsymbol{I}$ by solving the unconstrained optimization $\arg\min_{\boldsymbol{X}} \|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X} - \boldsymbol{I}\|_{\mathrm{F}}^2$ iteratively with the initial guess $\boldsymbol{X}_0 = \boldsymbol{I}$.

The objective function attains its global minimum 0 at every $\boldsymbol{X}$ that solves $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X} = \boldsymbol{I}$. The hope is that if we use an iterative scheme and start from $\boldsymbol{I}$, the iterates will converge to a solution $\boldsymbol{X}$ close to $\boldsymbol{I}$. [18] proposes to use a simple gradient descent iteration. Since $f(\boldsymbol{X}) = \|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X} - \boldsymbol{I}\|^2 = \mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}) - 2\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}) + \mathrm{tr}(\boldsymbol{I})$,

we can compute the gradient analytically as

$$\nabla_{\boldsymbol{X}} f = 4\boldsymbol{P}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X} - \boldsymbol{I}). \tag{86}$$

Therefore, the gradient descent iteration takes the form

$$\boldsymbol{X}_{n+1} = \boldsymbol{X}_n - \gamma\,\boldsymbol{P}\boldsymbol{X}_n(\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}_n - \boldsymbol{I}) \tag{87}$$

with γ being a tunable step size. This is Algorithm 3 in [18, sec. 3.5]. However, [18] does not mention how to pick γ, when the iteration converges, and how fast.

We can analyze this iteration by exactly the same techniques used in Sect. 3.2. First, we have $\boldsymbol{X}_k \to \boldsymbol{P}^{-1/2}\boldsymbol{Q}_0$ for

$$\boldsymbol{X}_0 = \boldsymbol{S}_0\boldsymbol{Q}_0, \quad \boldsymbol{S}_0 \in \mathcal{C}_{\boldsymbol{P}}, \quad \|\boldsymbol{X}_0\|_2 \le 1/\|\boldsymbol{P}\|_2^{1/2}, \quad \gamma \le 1/(2\|\boldsymbol{P}\|_2), \tag{88}$$

since the matrix iteration can be reduced to the scalar one $\sigma_{k+1} = \sigma_k - \gamma\lambda\sigma_k(\sigma_k^2 - 1)$[8], where λ is an eigenvalue of $\boldsymbol{P}$ and $\sigma_k/\sqrt{\lambda}$ is the corresponding singular value of $\boldsymbol{X}_k$. Next, the stability analysis shows that the Fréchet derivative at the limit $\boldsymbol{X}_*$ is

$$\mathrm{D}\boldsymbol{F}(\boldsymbol{X}) = \boldsymbol{X} - \gamma\boldsymbol{\Lambda}\boldsymbol{X} - \gamma\boldsymbol{\Lambda}^{1/2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}, \tag{89}$$

which has eigenvectors and eigenvalues as

$$\begin{aligned} \boldsymbol{Z}_{ij} &= \boldsymbol{E}_{ij} + \sqrt{\lambda_j/\lambda_i}\boldsymbol{E}_{ji}, \quad \mu_{ij} = 1 - \gamma(\lambda_i + \lambda_j), \quad i \le j, \\ \boldsymbol{Z}_{ij} &= \boldsymbol{E}_{ij} - \sqrt{\lambda_i/\lambda_j}\boldsymbol{E}_{ji}, \quad \mu_{ij} = 1, \quad i > j. \end{aligned} \tag{90}$$

Here the λ_i's are the diagonal entries of $\boldsymbol{\Lambda}$ and are also the eigenvalues of $\boldsymbol{P}$. Therefore, for $\rho(\mathrm{D}\boldsymbol{F}) < 1$, we need $\gamma < 1/\|\boldsymbol{P}\|_2$, which is less stringent than (88). Moreover, the iteration converges linearly with the asymptotic error diminishing coefficient not smaller than $(\lambda_{\max} - \lambda_{\min})/(\lambda_{\max} + \lambda_{\min})$. Therefore, the gradient descent is considered unattractive here compared to the quadratically converging iterations (45) and (46).

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2009)
2. Absil, P.A., Malick, J.: Projection-like retractions on matrix manifolds. SIAM J. Optim. **22**(1), 135–158 (2012)
3. Anderson, J.L.: An ensemble adjustment Kalman filter for data assimilation. Mon. Weather Rev. **129**(12), 2884–2903 (2001)
4. Benamou, J.D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numerische Mathematik **84**(3), 375–393 (2000)
5. Bergemann, K., Gottwald, G., Reich, S.: Ensemble propagation and continuous matrix factorization algorithms. Q. J. R. Meteorol. Soc. **135**(643), 1560–1572 (2009)

[8] Since $(1 - \sigma_{k+1}) = (1 - \sigma_k)(1 - \gamma\lambda\sigma_k(\sigma_k + 1))$, if $0 < \sigma_0 \le 1$ and $0 < \gamma \le 1/(2\lambda)$, σ_k monotonically increases and converges to 1.

6. Bishop, C.H., Etherton, B.J., Majumdar, S.J.: Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. Mon. Weather Rev. **129**(3), 420–436 (2001)
7. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Commun. Pure Appl. Math. **44**(4), 375–417 (1991)
8. Brockett, R.W.: Dynamical systems that learn subspaces. In: Mathematical System Theory, pp. 579–592. Springer (1991)
9. Chandrasekaran, S., Ipsen, I.C.: Backward errors for eigenvalue and singular value decompositions. Numerische Mathematik **68**(2), 215–223 (1994)
10. Cheng, M., Hou, T.Y., Zhang, Z.: A dynamically bi-orthogonal method for time-dependent stochastic partial differential equations I: derivation and algorithms. J. Comput. Phys. **242**, 843–868 (2013)
11. Dieci, L., Russell, R.D., Van Vleck, E.S.: Unitary integrators and applications to continuous orthonormalization techniques. SIAM J. Numer. Anal. **31**(1), 261–281 (1994)
12. Dowson, D., Landau, B.: The Fréchet distance between multivariate normal distributions. J. Multivar. Anal. **12**(3), 450–455 (1982)
13. Dutt, A.: High order stochastic transport and Lagrangian data assimilation. Master's thesis, Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, Massachusetts (2018)
14. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**(2), 303–353 (1998)
15. Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. Oceans **99**(C5), 10143–10162 (1994)
16. Evensen, G.: The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn. **53**(4), 343–367 (2003)
17. Feppon, F., Lermusiaux, P.F.J.: A geometric approach to dynamical model-order reduction. SIAM J. Matrix Anal. Appl. **39**(1), 510–538 (2018). https://doi.org/10.1137/16M1095202
18. Feppon, F., Lermusiaux, P.F.J.: Dynamically orthogonal numerical schemes for efficient stochastic advection and Lagrangian transport. SIAM Rev. **60**(3), 595–625 (2018). https://doi.org/10.1137/16M1109394
19. Feppon, F., Lermusiaux, P.F.J.: The extrinsic geometry of dynamical systems tracking nonlinear matrix projections. SIAM J. Matrix Anal. Appl. **40**(2), 814–844 (2019). https://doi.org/10.1137/18M1192780
20. Golub, G.H., Van Loan, C.F.: Matrix Comput, 4th edn. JHU Press, Baltimore (2013)
21. Govaerts, W., Werner, B.: Continuous bordering of matrices and continuous matrix decompositions. Numerische Mathematik **70**(3), 303–310 (1995)
22. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, vol. 31. Springer, Berlin (2006)
23. Hale, N., Higham, N.J., Trefethen, L.N.: Computing A^{α}, $\log(A)$, and related matrix functions by contour integrals. SIAM J. Numer. Anal. **46**(5), 2505–2523 (2008)
24. Higham, D.J.: Runge-Kutta type methods for orthogonal integration. Appl. Numer. Math. **22**(1–3), 217–223 (1996)
25. Higham, D.J.: Time-stepping and preserving orthonormality. BIT Numer. Math. **37**(1), 24–36 (1997)
26. Higham, N.J.: Computing the polar decomposition-with applications. SIAM J. Sci. Stat. Comput. **7**(4), 1160–1174 (1986)
27. Higham, N.J.: Newton's method for the matrix square root. Math. Comput. **46**(174), 537–549 (1986)
28. Higham, N.J.: Matrix Nearness Problems and Applications. Department of Mathematics, University of Manchester, Tech. rep (1988)
29. Higham, N.J., Schreiber, R.S.: Fast polar decomposition of an arbitrary matrix. SIAM J. Sci. Stat. Comput. **11**(4), 648–655 (1990)
30. Koch, O., Lubich, C.: Dynamical low-rank approximation. SIAM J. Matrix Anal. Appl. **29**(2), 434–454 (2007)
31. Laub, A.J.: Matrix Analysis for Scientists and Engineers. SIAM, New Delhi (2005)
32. Lermusiaux, P.F.J.: Error subspace data assimilation methods for ocean field estimation: theory, validation and applications. Ph.D. thesis, Harvard University, Cambridge, Massachusetts (1997)
33. Lermusiaux, P.F.J.: Data assimilation via error subspace statistical estimation, part II: Mid-Atlantic Bight shelfbreak front simulations, and ESSE validation. Mon. Weather Rev. **127**(7), 1408–1432 (1999). https://doi.org/10.1175/1520-0493
34. Lermusiaux, P.F.J.: Evolving the subspace of the three-dimensional multiscale ocean variability: Massachusetts Bay. J. Mar. Syst. **29**(1), 385–422 (2001)

35. Lermusiaux, P.F.J., Robinson, A.R.: Data assimilation via error subspace statistical estimation, part I: theory and schemes. Mon. Weather Rev. **127**(7), 1385–1407 (1999). https://doi.org/10.1175/1520-0493
36. Lin, J.: Bayesian learning for high-dimensional nonlinear dynamical systems: methodologies, numerics and applications to fluid flows. Ph.D. thesis, Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, Massachusetts (2020)
37. Lin, J.: Minimum-correction second-moment matching: theory, algorithms and applications. Master's thesis, Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, Massachusetts (2020)
38. Lorenz, E.N.: Predictability: a problem partly solved. In: Proc. Seminar on Predictability, vol. 1 (1996)
39. Musharbash, E., Nobile, F.: Dual dynamically orthogonal approximation of incompressible Navier Stokes equations with random boundary conditions. J. Comput. Phys. **354**, 135–162 (2018)
40. Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrices. Linear Algebra Appl. **48**, 257–263 (1982)
41. Ott, E., Hunt, B.R., Szunyogh, I., Zimin, A.V., Kostelich, E.J., Corazza, M., Kalnay, E., Patil, D., Yorke, J.A.: A local ensemble Kalman filter for atmospheric data assimilation. Tellus A **56**(5), 415–428 (2004)
42. Petersen, K.B., Pedersen, M.S.: The matrix cookbook. Technical University of Denmark (2012)
43. Philippe, B.: An algorithm to improve nearly orthonormal sets of vectors on a vector processor. SIAM J. Algebr. Discrete Methods **8**(3), 396–403 (1987)
44. Reich, S., Cotter, C.: Probabilistic Forecasting and Bayesian Data Assimilation. Cambridge University Press, Cambridge (2015)
45. Rheinboldt, W.C.: On the computation of multi-dimensional solution manifolds of parametrized equations. Numerische Mathematik **53**(1–2), 165–181 (1988)
46. Sapsis, T.P., Lermusiaux, P.F.J.: Dynamically orthogonal field equations for continuous stochastic dynamical systems. Physica D: Nonlinear Phenomena **238**(23–24), 2347–2360 (2009). https://doi.org/10.1016/j.physd.2009.09.017
47. Schatzman, M.: Numerical Analysis: A Mathematical Introduction. Oxford University Press, Oxford (2002)
48. Strogatz, S.H.: Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. Westview Press, Boulder (1994)
49. Subramani, D.N.: Probabilistic regional ocean predictions: stochastic fields and optimal planning. Ph.D. thesis, Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, Massachusetts (2018)
50. Tagare, H.D.: Notes on Optimization on Stiefel Manifolds. Yale University, New Haven (2011)
51. Trefethen, L.N., Bau, D.I.I.I.: Numerical Linear Algebra, vol. 50. SIAM, New Delhi (1997)
52. Ueckermann, M., Lermusiaux, P., Sapsis, T.: Numerical schemes and computational studies for dynamically orthogonal equations. MSEAS Report 11, Department of Mechanical Engineering, Massachusetts Institute of Technology (2011). http://mseas.mit.edu/?p=1928
53. Ueckermann, M.P., Lermusiaux, P.F.J., Sapsis, T.P.: Numerical schemes for dynamically orthogonal equations of stochastic fluid and ocean flows. J. Comput. Phys. **233**, 272–294 (2013). https://doi.org/10.1016/j.jcp.2012.08.041
54. Vetra-Carvalho, S., van Leeuwen, P.J., Nerger, L., Barth, A., Altaf, M.U., Brasseur, P., Kirchgessner, P., Beckers, J.M.: State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems. Tellus A: Dyn. Meteorol. Oceanogr. **70**(1), 1–43 (2018). https://doi.org/10.1080/16000870.2018.1445364
55. Villani, C.: Optimal Transport: Old and New, vol. 338. Springer, Berlin (2008)
56. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Math. Program. **142**(1–2), 397–434 (2013)
57. Yang, H., Li, H.: Weighted polar decomposition and WGL partial ordering of rectangular complex matrices. SIAM J. Matrix Anal. Appl. **30**(2), 898–924 (2008)