ABSTRACT
This paper presents a sparse approach of solving the one-sided Procrustes problem with special orthogonal constraint. By leveraging a planar decomposition common to all rotation matrices, a new constraint is introduced into this classical problem in the form of a sparsity-inducing norm. We call the resulting optimization problem the Rotational Lasso. Experimental results are presented from a synthetic dataset.

Index Terms— Orthogonal Procrustes problem, planar rotations, sparsity, estimation, gradient methods

1. INTRODUCTION
Consider the following Procrustes problem: given \( m \) examples of an \( n \)-dimensional vector and its image under a latent rotational map, we wish to estimate the unknown rotation matrix. We adhere to the classic approach of minimizing the sum of squared residuals. The favorable case is when \( m \gg n \), where we expect the well-known solution, given by the Singular Value Decomposition (SVD), to be well-defined. However, when \( m \ll n \) there are infinitely many optimal rotation matrices satisfying the minimization problem.

The focus of this paper is a sparse approach for solving the Procrustes problem:

\[
\arg\min_{Q \in SO(n)} \| B - QA \|_F^2, \tag{1}
\]

also known as “Wahba’s Problem” [1]. The notation \( \| \cdot \|_F \) denotes the Frobenius norm and the set \( SO(n) \) denotes the special orthogonal group of \( n \times n \) matrices, i.e., the set of square orthogonal matrices with determinant \(+1\). For full-rank \( A, B \in \mathbb{R}^{n \times m} \), the \( m \) corresponding columns of \( A \) and \( B \) are related by a latent \( n \)-dimensional rotation matrix, \( Q \). and we are interested in solving for an optimal matrix in the mean-square sense as an estimate of \( Q \). A solution to (1) is provided by the SVD of \( BA^T \). Using the full SVD, if \( BA^T = U \Sigma V^T \) with \( U^T U = V^T V = I \) and \( \Sigma \) is diagonal with \( \Sigma_{11} \geq \Sigma_{22} \geq \ldots \geq \Sigma_{nn} \geq 0 \), then an optimal rotation matrix is \( Q = U \text{diag}(1, \ldots, 1, \det(UV^T))V^T \). In general, given enough data, this solution is well defined because a full rank SVD will most likely possess unique singular vectors. In this paper we seek to solve this problem when there are few examples compared to the dimension of the data. Schönhemann addressed the issue of having an under-determined system [2], a necessary task since a full SVD is required. When the rank of \( BA^T \) is less than \( n \) the SVD approach will yield a valid minimizer. However, a consequence of having few examples is overfitting and this minimizer cannot be expected to generalize well. Therefore, we will refer to this solution as the naïve SVD (nSVD) solution.

The problem of being presented with few examples requires extra assumptions about the underlying structure of the latent quantity. For example, in [3] the authors assume that the unknown correlation matrix can be decomposed with relatively few Givens rotation matrices. In this paper we will assume that the unknown rotation matrix, \( Q_* \), is a product of applying a small number of 2D planar rotations. This approach is similar to [4], where the 2D planes are constructed from an orthonormal basis of \( \text{span}\{D_1, \ldots, D_m\} \) for \( D = B - A \) (\( D_i \) denotes the \( i \)-th column of \( D \)). The critical difference in our approach is the computation of the 2D planes.

Our main contribution is the formulation a direct approach for uncovering structure of a latent rotational map. Leveraging a canonical form exhibited by all elements of \( SO(n) \), an unbiased notion of sparsity emerges. Rather than optimize over \( SO(n) \), (1) is recast as an optimization problem over a Stiefel manifold with smaller dimensionality. With the addition of an \( \ell_1 \) penalty, the Rotational Lasso is formed. Promising results are featured in a simulated experiment.

The paper is organized as follows: §2 presents the theory behind planar rotations and how one can solve the associated optimization problem followed by §3, which introduces the Rotational Lasso. §4 exhibits results from a simulated example in dimension 200 and we conclude in §5.

2. PLANAR ROTATIONS
Given two orthonormal vectors, \( u, v \in \mathbb{R}^n \), a rotation of angle \( \theta \) in the 2-plane spanned by \( u \) and \( v \) is [5, 6]

\[
G([u|v], \theta) = I - (uv^T - vu^T) \sin \theta - (uu^T + vv^T)(1 - \cos \theta) \tag{2}
\]

\[
= \exp(-\theta(uv^T - vu^T)). \tag{3}
\]

The latter expression is derived in the Appendix. Fig. 1 exhibits a geometric interpretation of (2). Givens rotations restrict \( u \) and \( v \) to the standard basis vectors, and, hence, are a
subset of planar rotations. For $X \in \mathbb{R}^{n \times p}$ with orthonormal columns, let $\mathcal{R}(X)$ denote the set of all planar rotations derived from pairs of the $p$ columns. Thus, $\mathcal{R}(I_n)$ is the set of all Givens rotations in $n$ dimensions. Concatenate the orthonormal vectors $u$ and $v$ into $Y = [u|v]$. It follows that $Y \in \text{St}(n,2)$, where $\text{St}(n,p) = \{X : X \in \mathbb{R}^{n \times p}, X^T X = I_p\}$ with $p \leq n$, and that (2) can be rewritten as
\[
G(Y, \theta) = I - YW_\theta Y^T.
\]
where
\[
W_\theta = \begin{pmatrix}
1 - \cos \theta & \sin \theta \\
-\sin \theta & 1 - \cos \theta
\end{pmatrix}.
\]
We briefly address a sub-problem that will arise subsequently. Let $J = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. To find the optimal $\theta$ to minimize $\text{tr}(CYW_\theta Y^T)$ for given $Y \in \text{St}(n,2)$ and $C \in \mathbb{R}^{n \times n}$, we differentiate with respect to $\theta$:
\[
\text{tr}(CY \sin(\theta)I - \cos(\theta)J)Y^T = 0
\]
\[
\text{tr}(CYY^T) \sin \theta = \text{tr}(CYJY^T) \cos \theta
\]
\[
\tan \theta = \frac{\text{tr}(Y^TYJY)}{\text{tr}(Y^TCY)}
\]
The notion of sparsity with respect to a rotation needs to be explored. The connection of sparsity to $\ell_1$ regularization is an active field of research [7] and motivates the following two possible regularizations:
\[
\|B - QA\|_F^2 + \lambda \|Q\|_1 [Q \in \mathcal{SO}(n)]
\]
\[
\|B - e^{\Theta}A\|_F^2 + \lambda \|\Theta\|_1 [\Theta^T = -\Theta],
\]
where $\|X\|_F^2 = \sum_{i,j} |X_{ij}|^2$ for $X \in \mathbb{R}^{n \times m}$. For $\lambda$ large enough the optimal solution is $Q = I_n$ or $\Theta = 0_n$. As $\lambda$ decreases the first change we would expect to see is $Q$ to be a Givens rotation and, similarly, $\|\Theta\|_0 = 2$, implying $e^{\Theta}$ is also a Givens rotation. Thus, these forms of regularization may have a $\mathcal{R}(I)$ bias - a limitation in the effort of estimating the actual rotation matrix. We contend that sparsity in our setting should not lead to a “basis bias.” This leads to the following formulation.
For any skew-symmetric matrix, $\Theta \in \mathbb{R}^{n \times n}$ with $\Theta^T = -\Theta$, there exists a unitary matrix, $U \in \mathbb{R}^{n \times n}$, such that [8]
\[
\Theta = \begin{cases}
U \text{diag}(\theta_1 J, \ldots, \theta_n J) U^T & \text{n even} \\
U \text{diag}(\theta_1 J, \ldots, \theta_n J, 0) U^T & \text{n odd}
\end{cases}
\]
where $n = \lfloor n/2 \rfloor$ and $\theta_i \in \mathbb{R}$. The vector $\theta = [\theta_1, \ldots, \theta_n]^T$ can be thought of as an angle vector, where each angle $\theta_i$ is associated with the 2-plane spanned by column vectors $U_{2i-1}$ and $U_{2i}$. For $i = 1, \ldots, \bar{n}$, we construct the bivector $Y_i = [U_{2i-1} | U_{2i}]$ and arrive at the decomposition:
\[
\Theta = \sum_{i=1}^{\bar{n}} Y_i \times \theta_i J \times Y_i^T.
\]
We further note that $Y_i^TY_i = I_2$ and $Y_j^TY_j = 0_2$ for $i \neq j$, and thus for $k \geq 1$ we have $\Theta^k = \sum_{i=1}^{\bar{n}} Y_i(\theta_i J)^k Y_i^T$. After some algebra (see Appendix) we obtain the rotation matrix
\[
e^{\Theta} = I_n - \sum_{i=1}^{\bar{n}} Y_i W_{\theta_i} Y_i^T
\]
The sparsity of $\theta$ and the decomposition of (11) facilitates further simplification of the above results. Let $k$ denote the number of non-zero $\theta_i$, then for $\Theta \in \mathbb{R}^{n \times n}$ skew-symmetric, there exists a $U \in \text{St}(n,2k)$ such that $\Theta = U \text{diag}(\theta_1 J, \ldots, \theta_k J)^{\top}$. Subsequently, we obtain $e^{\Theta} = I - UVU^T$ where $V = \text{diag}(W_{\theta_1}, \ldots, W_{\theta_k})$. Thus, the value of $k$ is the number of 2D planar rotations we shall use to construct an estimate of $Q$. The estimate obtained will be of the form $I - UVU^T$ and the objective function we seek to minimize is $\|B - (I - UVU^T)A\|_F^2$. After expanding and simplifying, our optimization problem of interest is:
\[
\text{minimize} \quad \text{tr}(C(UVU^T))
\]
subject to $U \in \text{St}(n, 2k)$
\[
V = \text{diag}(W_{\theta_1}, \ldots, W_{\theta_k})
\]
where $C = AB^T$. The dimensionality of $\mathcal{SO}(n)$ is $n(n-1)/2$ and the dimensionality of $\mathcal{SO}(n, 2k)$ is $k(2n - 2k - 1)$. The dimensionality reduction factor is $(1 - \frac{2k}{n})(1 - \frac{2k}{n-2k})$, which, for $k \ll \bar{n}$, approximates to $1 - 4k/n$. In practice, the choice of $k$ is critical and leads to varying results. Given training data, one approach for selecting $k$ is cross validation.
We now address the method of solving (14). With relatively few examples the Hessian is ill-conditioned leading us to use [conjugate] gradient methods. The optimization techniques employed are detailed in [9, 10]. As there are two unknown quantities, $U$ and $\Theta$, we will proceed with alternating update equations. For $U$ fixed, the update for $\Theta$ has already been established in (8). The block-diagonal form of $V$ allows us to update each $\theta_i$ independently via $Y_i = [U_{2i-1} | U_{2i}]$. Updating $U$ involves projecting the unconstrained gradient to lie in the tangent plane at the point $U$ on the Stiefel manifold. The unconstrained gradient is $f_U = CUV + CTUV^T$ and projecting onto the tangent plane at $U$ yields [9]
\[
\nabla_U f = f_U - U f_U^{\top} U.
\]
With $\nabla_U f$ we can perform a line search. Let $\xi_\alpha = -\alpha \nabla_U f$ for some $\alpha \geq 0$. Projecting the point $U + \xi_\alpha$ onto $\text{St}(n, 2k)$ may yield a smaller objective value. This new point, given by the retraction $(U + \xi_\alpha)(I + \xi_\alpha^\top\xi_\alpha)^{-1/2}$, ensures that we will
Algorithm 1 Iterative method for solving (14)

1: \( U \leftarrow U_{\text{init}} \in \text{St}(n, 2k) \), \( V \leftarrow V_{\text{init}} \)
2: \( \text{while } \) Stopping Criterion \( = \) false \( \text{do} \)
3: \( f_U \leftarrow C U V + C^T U V^T \)
4: \( \nabla_U f \leftarrow f_U - U f_U^T U \)
5: \( \text{solve (16) and update } U \)
6: \( \text{for } i = 1 \text{ to } k \text{ do} \)
7: \( \theta_i \leftarrow \tan^{-1} \left( \frac{\text{tr}(Y^T C Y_\perp J)}{\text{tr}(Y^T C Y_\perp)} \right) \)
8: \( \text{end for} \)
9: \( V \leftarrow \text{diag}(W_{\theta_1}, \ldots, W_{\theta_k}) \)
10: \( \text{end while} \)

stay on \( \text{St}(n, 2k) \) due to \( \xi_\alpha \) lying in the tangent plane [10]. Thus, a line search attempts to solve
\[
\arg \min_{\alpha \geq 0} f \left( (U + \xi_\alpha)(I + \xi^T \xi_{\alpha})^{-1/2} \right) . \tag{16}
\]

We experimented with several line search techniques, including ones that advance geodesically. Ultimately, we found that (16) yielded the best results. Algorithm 1 details the procedure used to solve (14).

3. THE ROTATIONAL LASSO

The problem of selecting \( k \) presents a serious limitation when \( n \) is large. Without any prior knowledge of \( k \), any sort of validation technique can be costly. Therefore, we introduce an \( \ell_1 \) approach [11]. Set \( k = k_{\text{max}} \leq \lfloor n/2 \rfloor \) and solve the \( \ell_1 \) minimization problem:
\[
\text{minimize} \quad \text{tr}(C U V U^T) + \lambda \|\theta\|_1
\]
\[
\text{subject to} \quad U \in \text{St}(n, 2k_{\text{max}}), \quad V = \text{diag}(W_{\theta_1}, \ldots, W_{\theta_{k_{\text{max}}}}) \tag{17}
\]

for \( \lambda \geq 0 \). When the regularization parameter, \( \lambda \), is properly selected, the expectation is that the resulting \( \theta \in \mathbb{R}^{k_{\text{max}}} \) is sparse. In turn, the sparsity of \( \theta \) may reflect the true number of 2D planar rotations and does not impose a basis bias as seen in (9) and (10). The integer \( k_{\text{max}} \) is an upper bound placed on the true sparsity. In a situation where no assumptions of the actual \( k \) can be made, one can set \( k_{\text{max}} = \lfloor n/2 \rfloor \). The tradeoff, however, is that now \( U \in \text{St}(n, 2k_{\text{max}}) \) has many degrees of freedom.

We use the same template for solving (17) as we do (14), except that now we have an additional \( \lambda \|\theta\|_1 \) term. Referring back to (6), the \( \theta_i \) update involves solving
\[
\text{tr}(Y^T C Y_\perp) \sin \theta_i - \text{tr}(Y^T C Y_\perp J) \cos \theta_i + \lambda \text{sgn} \theta_i = 0, \tag{18}
\]

where \( \text{sgn} \theta \) is a subgradient of \(|\theta|\). Let \( \alpha_i = \text{tr}(Y^T C Y_\perp) \), \( \beta_i = \text{tr}(Y^T C Y_\perp J) \), and \( \gamma_i = \sqrt{\alpha_i^2 + \beta_i^2} \). After some manipulations, we obtain the update equation:
\[
\theta_i \leftarrow \cos^{-1} \left( \frac{\text{clip}(\beta_i, \lambda)}{\gamma_i} \right) - \cos^{-1} \left( \frac{\beta_i}{\gamma_i} \right) \tag{19}
\]
\[+ 2||\alpha_i, -\beta_i, \lambda|| \sin^{-1} \left( \frac{\alpha_i}{\gamma_i} \right) - 2||\alpha_i, \beta_i, \lambda|| \sin^{-1} \left( \frac{\beta_i}{\gamma_i} \right) \]

where \( \text{clip}(\beta, \lambda) = \text{sgn}(\beta) \times \min(|\beta|, \lambda) \) (see Fig. 2). The two trailing terms provide necessary adjustments to account for negative \( \alpha_i \). The “\([\alpha, \beta, \lambda]\)” term denotes the indicator \( 1 \{||\beta| > |\lambda|, \alpha < 0, \beta > 0\} \). Essentially, the continuous shrinkage results from the “clipping” of \( \beta_i \). For \( \alpha_i \geq 0 \) and \( \gamma_i \geq \lambda \), it can be shown that the \( \ell_1 \) regularization leads to a soft-threshold of \( \sin^{-1} \left( \frac{\lambda}{\gamma_i} \right) \) to the unregularized \( \theta_i \) (when \( \gamma_i < \lambda \), the unregularized \( \theta_i \) is set to 0).

4. EXPERIMENTS

In this section we compare our technique alongside two other methods for solving (1). All computations were done in MATLAB\textsuperscript{\textregistered}. The first method, nSVD, uses Matlab’s \texttt{svd()} function to compute a full SVD of \( BA^T \). The second method, abbreviated “RFC” to credit the authors of [4], computes an ON basis of \( \text{span}\{B_1 - A_1, \ldots, B_m - A_m\} \), successively takes 2 vectors at time (without replacement) to form 2-planes, and subsequently computes the optimal angle associated with each 2-plane. This method is similar to our approach but differs in how the 2-planes are computed.

The following experiment was done via simulated data. The dimension of the data was \( n = 200 \) and the unrotated vectors were generated from sampling \( N(0, I) \) followed by normalizing to the unit sphere. The actual number of 2D planar rotations is 8. After the rotated vectors are computed, all vectors are corrupted by Gaussian noise and randomly perturbed on the spherical surface. The noise powers were selected to simulate a 0-mean angular error with standard deviation \( \sigma_a \) [\( \sigma_a \)] degrees for the unrotated [rotated] vectors. The complete data set consists of 1040 vector pairs (unrotated and rotated) of which we trained on 40 and tested on the other 1000. With only 40 examples to train on we set \( k_{\text{max}} = 40/2 = 20 \) for the Rotational Lasso and selected \( \lambda \) via cross validation on each training set. After 100 train/test random partition assignments we computed the results as presented in Table 1. For all cases, except when a 1° angular
noise power was present, a sparsity level of 8 was observed at least 92% of the time. While the testing error remains comparably small, the increasing angular noise causes the sparsity distribution to spread.

5. CONCLUSION
We have developed a sparse approach for solving the special orthogonal Procrustes problem. The critical assumption is that the rotation matrix is a product of a relatively small number of planar rotations. The resulting optimization problem, the Rotational Lasso, is also suitable when a dataset possesses few examples. A potential improvement would be to add an $\ell_2$ penalty on $\theta$ and form a “Rotational Elastic Net” [12]. The resulting $\theta$, update equation would rely on numerical techniques (consider solving $x = \cos(x)$). In a situation with few examples where the sparsity cannot be assumed to be highly sparse, the $\ell_2$ regularization may prove advantageous.

6. APPENDIX
A planar rotation requires three inputs: two orthonormal vectors ($u$ and $v$) and an angle ($\theta$). Let $Y = [u\ v] \in \mathbb{R}^{n \times 2}$, implying $Y \in \text{St}(n, 2)$. It is a well-known result that $e^Z \in \text{SO}(n)$ when $Z \in \mathbb{R}^{n \times n}$ is skew-symmetric. Let $Z = Y \Theta Y^T$ where $\Theta = \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}$. Thus, $\Theta^T = -\Theta$ and $Z^T = Y \Theta^T Y^T = -Y \Theta Y^T = -Z$, i.e., $Z$ is skew-symmetric. To evaluate $e^Z = I + Z/1! + Z^2/2! + Z^3/3! + \ldots$, we note that $Z^k = Y \Theta^k Y^T$ ($k = 1, 2, \ldots$) and

\[
Z = Y \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix} Y^T
\]

\[
\frac{1}{2} Z^2 = Y \begin{pmatrix} -\theta^2/2 & 0 \\ 0 & -\theta^2/2 \end{pmatrix} Y^T
\]

\[
\frac{1}{6} Z^3 = Y \begin{pmatrix} 0 & +\theta^3/6 \\ -\theta^3/6 & 0 \end{pmatrix} Y^T
\]

\[
\frac{1}{24} Z^4 = Y \begin{pmatrix} 0 & +\theta^4/24 \\ -\theta^4/24 & 0 \end{pmatrix} Y^T
\]

\[
\vdots
\]

Summing the above yields

\[
Z + Z^2/2 + Z^3/6 + \ldots = Y \begin{pmatrix} \cos \theta - 1 & -\sin \theta \\ \sin \theta & \cos \theta - 1 \end{pmatrix} Y^T,
\]

from which we obtain $e^Z = I - Y W_{\theta} Y^T$. For $r \in \text{span}\{u\ v\}$, we have $e^Z r = r$ because $Y^T r = (0 \ 0)^T$. Additionally, the cosine of the angle between $u$ and $e^Z u$ is $u^T e^Z u = u^T u - (1 - \cos \theta) = \cos \theta$ because $Y^T u = (0 \ 0)^T$. Similarly, the cosine of the angle between $v$ and $e^Z u$ is $v^T e^Z u = v^T u - (1 - \cos \theta) = \cos \theta$ because $Y^T v = (0 \ 0)^T$. These properties are represented in Fig. 1.

7. REFERENCES

Table 1. Simulation Results: For each ($\sigma_a$, $\sigma_b$) pair we list the mean training error (Frobenius norm of residual) and the corresponding standard deviation in brackets. Directly underneath the training results are the testing results.