MULTI-RANK SPARSE AND FUNCTIONAL PCA
MANIFOLD OPTIMIZATION AND ITERATIVE DEFLATION TECHNIQUES

Michael Weylandt

Department of Statistics
Rice University
Houston, TX 77005
michael.weylandt@rice.edu

ABSTRACT

We consider the problem of estimating multiple principal components using the recently-proposed Sparse and Functional Principal Components Analysis (SFPCA) estimator. We first propose an extension of SFPCA which estimates several principal components simultaneously using manifold optimization techniques to enforce orthogonality constraints. While effective, this approach is computationally burdensome so we also consider iterative deflation approaches which take advantage of existing fast algorithms for rank-one SFPCA. We show that alternative deflation schemes can more efficiently extract signal from the data, in turn improving estimation of subsequent components. Finally, we compare the performance of our manifold optimization and deflation techniques in a scenario where orthogonality does not hold and find that they still lead to significantly improved performance. Index Terms—regularized PCA, orthogonality, deflation, sparsity, manifold optimization

1. INTRODUCTION

Principal Components Analysis (PCA, [1]) is a widely-used approach to finding low-dimensional patterns in complex data, enabling visualization, dimension reduction (compression), and predictive modeling. While PCA performs well in a wide range of low-dimensional settings, its performance degrades rapidly in high-dimensions, necessitating the use of regularized variants. Recently, Allen and Weylandt [2] proposed Sparse and Functional PCA (SFPCA), a unified regularization scheme that allows for simultaneous smooth (functional) and sparse estimation of both row and column principal components (PCs). The rank-one SFPCA estimator is given by

$$\arg\max_{u \in \mathbb{R}^n, v \in \mathbb{R}^n} u^T X v - \lambda_u P_u(u) - \lambda_v P_v(v)$$

where $P_u(\cdot)$ is a regularizer inducing sparsity in the row PCs, with strength controlled by $\lambda_u$; $\Omega_u$ is a positive semi-definite penalty matrix, typically a second- or fourth-order difference matrix; $S_u = I + \alpha_u \Omega_u$ is a smoothing matrix for the row PCs, with strength controlled by $\alpha_u$; and $\mathbb{E}_{S_u}$ is the unit ellipse of the $S_u$-norm, i.e., $\mathbb{E}_{S_u} = \{ u \in \mathbb{R}^n : u^T S_u u \leq 1 \}$. (Respectively, $P_v(\cdot)$, $\lambda_v$, $\Omega_v$, $\alpha_v$, $S_v$, and $\mathbb{E}_{S_v}$ for the column PCs.)

Allen and Weylandt [2] show that SFPCA unifies much of the existing regularized PCA literature [3–8] into a single framework, avoiding many pathologies of other approaches. Finally, they propose an efficient alternating maximization scheme with guaranteed global convergence to solve the bi-concave SFPCA problem (1). The SFPCA estimator only allows for a single pair of PCs to be estimated for a given data matrix $X$. Allen and Weylandt suggest applying SFPCA repeatedly to the deflated data matrix if multiple PCs are desired. While this approach performs acceptably in practice, it loses the interpretable orthogonality properties of standard PCA. In particular, the estimated PCs are no longer guaranteed to be orthogonal to each other, hindering the common interpretation of PCs as statistically independent sources of variance, or to the deflated data matrix, suggesting that additional signal remains uncaptured.

We extend the work of Allen and Weylandt [2] to address these shortcomings: first, in Section 2, we modify the SFPCA estimator to simultaneously estimate several sparse PCs subject to orthogonality constraints. The resulting estimator is constrained to a product of generalized Stiefel manifolds and we propose three efficient algorithms to solve the resulting manifold optimization problem. Next, in Section 3, we propose improved deflation schemes which provably remove all of the signal from the data matrix, allowing for more accurate iterative estimation of sparse PCs. Finally, we demonstrate the improved performance of our manifold estimators and deflation schemes in Section 4. Supplemental materials for this paper, including proofs, counter-examples, and additional algorithmic details, are available online at https://arxiv.org/abs/1907.12012.

2. MANIFOLD OPTIMIZATION FOR SFPCA

One of the most attractive properties of PCA is the factors it extracts are orthogonal ($u_t^T u_t = v_t^T v_t = 1$ if $t = s$ and 0 otherwise). Because of this, PCs can be interpreted as separate sources of variance and, under an additional Gaussianity assumption, statistically independent. While this follows directly from the properties of eigendecompositions for standard PCA, it is much more difficult to obtain similar results for sparse PCA. Some authors have suggested that there exists a fundamental tension between orthogonality and sparsity, with Journée et al. [9] calling the goal of sparse and orthogonal estimation “questionable.” Indeed, ex post orthogonalization of sparse PCs, e.g., using a Gram-Schmidt step, destroys any sparsity in the estimated PCs. To avoid this, it is necessary to impose orthogonality directly in the estimation step, rather than trying to impose it afterwards.

We modify the SFPCA estimator to simultaneously estimate multiple PCs subject to an orthogonality constraint:

$$\arg\max_{U \in \mathbb{R}^n_{+ \times k}, V \in \mathbb{R}^n_{+ \times k}, \lambda_u, \lambda_v} \text{Tr}(U^T X V) - \lambda_u P_U(U) - \lambda_v P_V(\lambda_u, \lambda_v)$$

MW acknowledges support from the NSF Graduate Research Fellowship Program under grant number 1450681.
Method & Two-Way Orthogonality & One-Way Orthogonality & Subsequent Orthogonality & Robust to Scale of $u_t$, $v_t$
--- & $u_t^T X_t v_t = 0$ & $u_t^T X_{t,s} v_t = 0$ & $u_t^T X_{t,s} v_t = 0$ & $u_t^T X_{t,s} v_t = 0$
Hotelling’s Deflation (HD) & ✓ & ✗ & ✗ & ✗
Projection Deflation (PD) & ✓ & ✓ & ✗ & ✗
Schur Deflation (SD) & ✓ & ✓ & ✓ & ✓

Table 1: Properties of Hotelling’s Deflation (HD), Projection Deflation (PD), and Schur Complement Deflation (SD). Only SD captures all of the individual signal of each principal component without re-introducing signal at later iterations. Additionally, only SD allows for the non-unit-norm PCs estimated by SFPCA to be used without rescaling.

where $V^{S_u}_n$ is the generalized Stiefel manifold of order $k$ over $\mathbb{R}^n$, i.e.,

$$U \in V^{S_u}_n \iff U \in \mathbb{R}^{n \times k} \text{ and } U^T S_u U = I_k.$$  

The generalized Stiefel manifold constraint ensures orthogonality of the estimated PCs, while still allowing us to capture most of the variability in the data.\footnote{When estimating orthogonal factors, it is common to re-express the problem using a (generalized) Grassmannian manifold constraint to avoid identifiability issues. We cannot use the Grassmannian approach here as sparse estimation (implicitly) fixes a single coordinate system.} We note that, because we use a generalized Stiefel constraint, the estimated PCs will be orthogonal with respect to $S_u$, i.e., $u_t^T S_u u_s = 0$ for $t \neq s$, rather than orthogonal in the standard sense. This is commonly observed for functional PCA variants \cite{5, 6, 3}. and can be interpreted as orthogonality under the inner product generating the $S_u$-norm. If no roughness penalty is imposed ($\alpha_u = 0$ or $\alpha_v = 0$), then our method gives orthogonality in the standard sense.

To solve the Manifold SFPCA problem (2), we employ an alternating maximization scheme, first holding $V$ fixed while we update $U$ and vice versa, as described in Algorithm 1. Even with one parameter held fixed, the resulting sub-problems are still difficult manifold optimization problems, which require iterative approaches to obtain a solution \cite{10-17}.

Algorithm 1 Manifold SFPCA Algorithm

1. Initialize $\hat{U}, \hat{V}$ to the leading $k$ singular vectors of $X$
2. Repeat until convergence:
   (a) $U$-subproblem: Solve using Algorithm 2 or 3:
   $$\hat{U} = \arg\min_{U \in V^{S_u}_n} -\text{Tr}(U^T X V^T) + \lambda U P_U(U)$$
   (b) $V$-subproblem: Solve using Algorithm 2 or 3, with $U$ and $V$ reversed:
   $$\hat{V} = \arg\min_{V \in V^{S_v}_p} -\text{Tr}(U^T X V^T) + \lambda V P_V(V)$$
3. Return $\hat{U}$ and $\hat{V}$

Allen and Weylandt \cite{2} developed a custom projected proximal gradient algorithm to solve the $u$- and $v$-subproblems of the rank-one SFPCA estimator. Assuming $P_U$ and $P_V$ are positive homogeneous, (e.g., $P(\lambda) = \|A\|_p$ for arbitrary $p \geq 1$ and $A$), they establish convergence to a stationary point. In order to extend this idea to the multi-rank (manifold) case, we use the recently-proposed Manifold Proximal Gradient (ManPG) scheme of Chen et al. \cite{14}, detailed in Algorithm 2. ManPG proceeds in two-steps: first, we solve for a descent direction $D$ of the objective along the tangent space of the generalized Stiefel manifold, subject to the tangency constraint of $D^T S_u U^{(k)}$ being skew-symmetric; secondly, back-tracking line search is used to determine a step-size $\alpha$, after which the estimate is projected back onto the generalized Stiefel manifold using a retraction anchored at the previous $U^{(k)}$. The retraction, which plays the same role as the projection step in the original SFPCA algorithm \cite[Algorithm 1]{2}, can be computed using a Cholesky factorization \cite[Algorithm 3.1]{17}. Chen et al. \cite{15} showed that a single step of ManPG is sufficient to ensure convergence despite the bi-concave objective: we refer to their approach as Alternating ManPG (A-ManPG).

Algorithm 2 Manifold Prox. Gradient (ManPG) for $U$-Subproblem

1. Initialize $U^{(k)} = \hat{U}$
2. Repeat until convergence:
   - Solve, subject to $D^T S_u U^{(k)} + (U^{(k)})^T S_u D = 0$:
     $$D = \arg\min_{D \in \mathbb{R}^{n \times k}} -(X \hat{V}, D)_F + \lambda U P_U(U^{(k)} + D)$$
   - Select $\alpha$ by Armijo-back-tracking
   - $U^{(k+1)} = \text{Retr}_{U^{(k)}}(\alpha D)$
3. Return $\hat{U}$

Note that in general manifold proximal gradient schemes \cite{14, 15} impose a maximum step-size to ensure that linearization of the smooth portion of the objective actually leads to descent: because the smooth portion of our objective function in linear in $U$ and $V$, we can omit this term from Algorithm 2.

While efficient when tuned properly, we have found the performance of ManPG on the $U$- and $V$-subproblems quite sensitive to infeasibility in the descent direction. A more robust scheme can be derived by using the Manifold ADMM (MADMM) scheme of Kovalsky et al. \cite{13} to solve the subproblems. Like standard ADMM schemes, MADMM allows us to split a problem into two parts, each of which can be solved more easily than the global problem. When applied to the $U$- and $V$-subproblems, the MADMM allows us to separate the manifold constraint from the sparsity inducing regularizer, thereby side-stepping the orthogonality / sparsity tension at the heart of this paper. After this splitting, the smooth update can be shown to equivalent to the unbalanced Procrustes problem \cite{18, 19} with a closed-form update: $U^{(k+1)} = S_u^{1/2} A B^T$ where $A \Delta B^T$ is the SVD of $S_u^{1/2} X \hat{V} + p S_u^{1/2} (W^{(k)} - Z^{(k)})$. The sparse update is simply the proximal operator of $P_U(\cdot)$, typically a thresholding step \cite[Chapter 6]{20}. To the best of our knowledge, the convergence of MADMM has not yet been established, but we have not observed significant non-convergence problems in our experiments.
Algorithm 3 Manifold ADMM (MADMM) for $\bar{U}$-Subproblem

1. Initialize $U^{(k)} = W^{(k)} = \bar{U}$, $Z^{(k)} = 0$ and $k = 1$
2. Repeat until convergence:
   
   $U^{(k+1)} = \arg\min_{U \in \mathbb{R}^{n \times k}} -\text{Tr}(U^T X V) + \frac{\rho}{2} ||U - W^{(k)} + Z^{(k)}||_F^2$

   $W^{(k+1)} = \arg\min_{W \in \mathbb{R}^{n \times k}} \lambda_U \text{Tr}(W) + \frac{\rho}{2} ||U^{(k+1)} - W + Z^{(k)}||_F^2$

   $Z^{(k+1)} = Z^{(k)} + U^{(k+1)} - W^{(k+1)}$
3. Return $\bar{U}$ and $\bar{V}$

3. ITERATIVE DEFLATION FOR SFPCA

We next consider the use of iterative deflation schemes for multi-rank SFPCA. As discussed by Mackey [21], the attractive orthogonality properties of standard (Hotelling’s) deflation depend critically on the estimated PCs being exact eigenvectors of the covariance matrix. Because the PCs estimated by sparse PCA schemes are almost surely not eigenvectors, Mackey [21] proposes several alternate deflation schemes which retain some of the attractive properties of Hotelling’s deflation even when non-eigenvectors are used. We extend these to the low-rank model and allow for deflation by several PCs, possibly with non-unit norm, simultaneously, e.g., as produced by ManSFPCA. To ease exposition, we first work in the vector setting and consider the general case at the end of this section. The properties of our proposed deflation schemes are summarized in Table 1 above.

The simplest deflation scheme is essentially that used by Hotelling [1], extended to the low-rank model:

$$X_t := X_{t-1} - d_t u_t v_t^T$$

(HD)

For two-way sparse PCA variants [7, 8], this deflation gives a deflated matrix which is “two-way” orthogonal to the estimated PCs, i.e., $u_t^T X_t v_t = 0$. We may interpret this as Hotelling’s deflation (HD) capturing all the signal jointly associated with the pair $(u_t, v_t)$. We may also ask if HD captures all of the signal associated with $u_t$ or only the signal which is also associated with $v_t$. If HD captures all of the signal associated with $u_t$, then we would expect $u_t^T X_t \tilde{v} = 0$ for all $\tilde{v} \in \mathbb{R}^p$, or equivalently, $u_t^T X_t = 0$. Interestingly, HD does not have this left-orthogonality property, suggesting that it leaves additional $u_t$-signal in the deflated matrix $X_t$.

HD fails to yield left- and right-orthogonality because it is not based on a projection operator. To address this in the covariance model, Mackey [21] proposed a deflation scheme which projects the covariance matrix onto the orthogonal complement of the estimated principal component. We extend this idea to the low-rank model by projecting the column- and row-space of the data matrix into the orthogonal complement of the left- and right-PCs respectively, giving two-way projection deflation (PD):

$$X_t := (I_n - u_t u_t^T) X_{t-1} (I_p - v_t v_t^T).$$

(PD)

Unlike HD, PD captures all of the linear signal associated with $u_t$ and $v_t$ individually ($u_t X_t \tilde{v} = u_t X_t v_t = 0, \forall \tilde{u} \in \mathbb{R}^n, \tilde{v} \in \mathbb{R}^p$).

If we use PD repeatedly, however, the multiply-deflated matrix will not continue to be orthogonal to the PCs: that is, $u_t^T X_{t+s} \neq 0$ for $s \geq 1$. This suggests that repeated application of PD can reintroduce signal in the direction of the PCs that we previously removed. This occurs because PD works by sequentially projecting the data matrix, but in general the composition of two orthogonal projections is not another orthogonal projection without additional assumptions. To address this, Mackey [21] proposed a Schur complement deflation (SD) technique, which we now extend to the low-rank (two-way) model:

$$X_t := X_{t-1} - X_{t-1} v_t u_t^T X_{t-1} / u_t^T X_{t-1} v_t .$$

(SD)

While Mackey motivates this approach using conditional distributions and a Gaussianity assumption on $X$, it can also be understood as an alternate projection construct which is more robust to scaling and non-orthogonality.

So far, we have only considered the behavior of the proposed deflation schemes for two-way sparse PCA. If we consider SFPCA in general, however, $u_t$ and $v_t$ are unit vectors under the $S_u$- and $S_v$-norms, not under the Euclidean norm. Consequently, the projections used by PD may not be actual projections and a PD deflated matrix may fail to be two- or one-way orthogonal. Normalizing the estimated PCs before deflation addresses this problem and is recommended in practice; conversely, because its deflation term is invariant under rescalings of $u_t$ and $v_t$, SD works without renormalization.

The extension of these techniques to the multi-rank case is straightforward. We give the normalized variants here:

$$X_{1D} := X_{t-1} - U_t (U_t^T U_t)^{-1} U_t^T X_{t-1} V_t (V_t^T V_t)^{-1} V_t^T X_t$$

$$X_{1D} := (I_n - U_t (U_t^T U_t)^{-1} U_t^T) X_{t-1} (I_p - V_t (V_t^T V_t)^{-1} V_t^T)$$

$$X_{1D} := X_{t-1} - X_{t-1} V_t (U_t^T X_{t-1} V_t)^{-1} U_t^T X_{t-1}$$

As in the covariance model, if $u_t$ and $v_t$ are true singular vectors, all three deflation schemes are equivalent.

4. SIMULATION STUDIES

In this section, we compare the performance of Manifold SFPCA and the iterative rank-one deflation schemes proposed above in illustrative simulation studies. Manifold SFPCA using Manifold ADMM (Algorithm 3) to solve the subproblems achieves better solutions than Manifold Proximal Gradient or A-ManPG (Algorithm 2) in less time. Furthermore, despite the additional flexibility of the iterative rank-one variants, Manifold SFPCA achieves both better signal recovery and a higher proportion of variance explained, even when the orthogonality assumptions are violated.

We first consider the relative performance of the three algorithms proposed for solving the Manifold SFPCA problem (2). We generate data in $X^* = U^* D^* V^*$ in $\mathbb{R}^{250 \times 100}$ with three distinct PCs: the left PCs ($U^*$) are localized sinuoids of varying frequency; the right PCs ($V^*$) are non-overlapping sawtooth waves. (See Figure 1.) We add independent standard Gaussian noise (E) to give a signal-to-noise ratio (SNR) of $||X^*||/||E|| \approx 1.2$. We fix $\lambda_k = \lambda_n = 1$ and $\alpha_u = \alpha_v = 3$ which is near optimal for all three schemes. This is a favorable setting for Manifold SFPCA as the underlying signals are orthogonal, sparse, smooth, and of comparable magnitude.

Table 2 shows the performance of our three Manifold SFPCA algorithms on several metrics, averaged over 100 replicates. Overall, the Manifold ADMM (MADMM [13]) and Alternating Manifold Proximal Gradient (A-ManPG [15]) variants perform best, handily beating the Manifold Proximal Gradient scheme [14] on all measures. MADMM achieved the best objective value on every replicate.
In terms of signal recovery, MADMM achieves slightly better performance than A-ManPG on the right singular vectors, while A-ManPG is slightly better on the left singular vectors. Computationally, MADMM dominates both proximal gradient variants even though it requires many more matrix decompositions. The descent direction subproblems of ManPG and A-ManPG are rather expensive to solve repeatedly and their performance is very sensitive to the solver used. Overall, the MADMM variant of Manifold SFPCA achieves the best optimization and statistical performance in far less time than the proximal gradient-based variants.

Next, we compare Manifold SFPCA with the iterative deflation schemes proposed in Section 3 in two different scenarios: the favorable scenario used above and a less-favorable scenario where the true PCs are shifted and no longer orthogonal. \( (n = p = 100 \text{ and } ||(U^*)^T(V^*) - I_3||, ||(V^*)^TU^* - I_3|| \approx 0.37) \) We compare the proportion of variance explained using Manifold SFPCA with iterative rank-one SFPCA using the normalized Hotelling, Projection, and Schur Complement deflation strategies. As can be seen in Table 3, PD and SD consistently dominate HD. Because PD and SD fully remove the signal associated with estimated PCs, the subsequent PCs are able to capture different signals and explain a larger fraction of variance. By ensuring that the signal is never re-introduced, SD does even better than PD as we consider higher ranks. Interestingly, while PD and SD perform about as well when the underlying signals are orthogonal, SD performs much better in the non-orthogonal scenario. PD and SD perform about as well when the underlying signals are orthogonal, SD performs much better in the non-orthogonal scenario. Interestingly, while PD and SD perform as well when the underlying signals are orthogonal, SD performs much better in the non-orthogonal scenario.

By estimating all three PCs simultaneously, Manifold SFPCA is able to find a better set of PCs than any of the greedy deflation methods.

### Table 2: Comparison of Manifold-ADMM, Manifold Proximal Gradient, and Alternating Manifold Proximal Gradient approaches for Manifold SFPCA (2).

|                         | MADMM | ManPG | A-ManPG |
|-------------------------|-------|-------|---------|
| Time (s)                | 19.46 | 217.40| 175.99  |
| Suboptimality           | 0     | 4.58  | 12.23   |
| rSS-Error \( U \)       | 68.66%| 74.03%| 64.54%  |
|                         | 36.85%| 50.69%| 43.17%  |
| TPR \( U \)             | 87.75%| 69.72%| 89.99%  |
|                         | 94.75%| 70.30%| 89.87%  |
| FPR \( U \)             | 12.25%| 30.28%| 10.01%  |
|                         | 5.25% | 29.70%| 10.13%  |

### Table 3: Cumulative Proportion of Variance Explained (CPVE) of Rank-One SFPCA with (normalized) Hotelling, Projection, and Schur Complement Deflation and of (order 3) Manifold SFPCA. SD gives the best CVPE of the iterative approaches and appears to be more robust to violations of orthogonality (Scenario 2). Manifold SFPCA outperforms the iterative methods in both scenarios.

### 5. DISCUSSION

We have introduced two practical extensions to Sparse and Functional PCA: first, we presented a multi-rank scheme which estimates multiple PCs simultaneously and proposed algorithms to solve the resulting manifold optimization problems. The resulting estimator inherits many of the attractive properties of rank-one SFPCA and is, to the best of our knowledge, the first multi-rank PCA scheme for the low-rank model. Manifold SFPCA combines both the flexibility and superior statistical performance of rank-one SFPCA with the superior interpretability of orthogonal (non-regularized) PCA.

Secondly, we re-considered the use of Hotelling’s deflation, and developed two additional deflation schemes which have attractive theoretical properties and empirical performance. Our schemes extend the results of Mackey [21] in several ways: they allow for deflation by multiple PCs in a single step, they are applicable to the low-rank model and the covariance model, and they are robust to non-orthogonality and non-unit-scaling common to Functional PCA variants. While developed for SFPCA, these deflation schemes are useful for any regularized PCA model.

We note here that our results can be extended naturally to other multivariate analysis techniques which can be expressed in a regularized SVD framework (e.g., PLS, CCA, etc.). Our deflation approaches can be also extended to the higher-order / multi-way context and may be particularly useful in the context of regularized CP decompositions [22, 23]. In the tensor setting, ManSFPCA is a sparse and smooth version of a Tucker decomposition, which suggests several interesting extensions we leave for future work [24, 25].
6. REFERENCES

[1] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” Journal of Educational Psychology, vol. 24, no. 6, pp. 417–441, 1933. DOI: 10.1037/h0071325.

[2] G. I. Allen and M. Weylandt, “Sparse and functional principal components analysis,” in DSW 2019: Proceedings of the 2nd IEEE Data Science Workshop, G. Karypis, G. Michailidis, and R. Willett, Eds., Minneapolis, Minnesota: IEEE, 2019, pp. 11–16. DOI: 10.1109/DSW.2019.8755778.

[3] B. W. Silverman, “Smoothened functional principal components analysis by choice of norm,” Annals of Statistics, vol. 24, no. 1, pp. 1–24, 1996. DOI: 10.1214/aos/1033066196.

[4] H. Shen and J. Z. Huang, “Sparse principal component analysis via regularized low rank matrix approximation,” Journal of Multivariate Analysis, vol. 99, no. 6, pp. 1015–1034, 2008. DOI: 10.1016/j.jmva.2007.06.007.

[5] J. Z. Huang, H. Shen, and A. Buja, “Functional principal components analysis via penalized rank one approximation,” Electronic Journal of Statistics, vol. 2, pp. 678–695, 2008. DOI: 10.1214/08-EJJS218.

[6] ——, “The analysis of two-way functional data using two-way regularized singular value decompositions,” Journal of the American Statistical Association, vol. 104, no. 488, pp. 1609–1620, 2009. DOI: 10.1198/jasa.2009.tm08024.

[7] D. M. Witten, R. Tibshirani, and T. Hastie, “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,” Biostatistics, vol. 10, no. 3, pp. 515–534, 2009. DOI: 10.1093/biostatistics/kxp008.

[8] G. I. Allen, L. Grosenick, and J. Taylor, “A generalized least-square matrix decomposition,” Journal of the American Statistical Association, vol. 109, no. 505, pp. 145–159, 2014. DOI: 10.1080/01621459.2013.852978.

[9] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, “Generalized power method for sparse principal component analysis,” Journal of Machine Learning Research, vol. 11, pp. 517–553, 2010. Available at: http://www.jmlr.org/papers/v11/journee10a.html.

[10] P.-A. Absil, R. Mahony, and R. Sepulchre, Optimization Algorithms on Matrix Manifolds. Princeton University Press, 2007, ISBN: 978-0-691-13298-3.

[11] Z. Wen and W. Yin, “A feasible method for optimization with orthogonality constraints,” Mathematical Programming, vol. 142, no. 1-2, pp. 397–434, 2013. DOI: 10.1007/s10107-012-0584-1.

[12] R. Lai and S. Osher, “A splitting method for orthogonality constrained problems,” Journal of Scientific Computing, vol. 58, no. 2, pp. 431–449, 2014. DOI: 10.1007/s10915-013-9740-x.

[13] A. Kovalsky, K. Glashoff, and M. M. Bronstein, “MADMM: A generic algorithm for non-smooth optimization on manifolds,” in ECCV 2016: Proceedings of the 14th European Conference on Computer Vision, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., ser. Lecture Notes in Computer Science, vol. 9909, Springer, 2016, pp. 680–696. DOI: 10.1007/978-3-319-46454-1_41.

[14] S. Chen, S. Ma, A. M.-C. So, and T. Zhang, “Proximal gradient method for nonsmooth optimization over the Stiefel manifold,” ArXiv Pre-Print 1811.00980, Available at: http://arxiv.org/abs/1811.00980.

[15] S. Chen, S. Ma, L. Xue, and H. Zou, “An alternating manifold proximal gradient method for sparse PCA and sparse CCA,” ArXiv Pre-Print 1903.11576, 2019. Available at: http://arxiv.org/abs/1903.11576.

[16] K. Benidis, Y. Sun, P. Babu, and D. P. Palomar, “Orthogonal sparse PCA and covariance estimation via Procrustes reformulation,” IEEE Transactions on Signal Processing, vol. 64, no. 23, pp. 6211–6226, 2016. DOI: 10.1109/TSP.2016.2605073.

[17] H. Sato and K. Aihara, “Cholesky QR-based retraction on the generalized Stiefel manifold,” Computational Optimization and Applications, vol. 72, no. 2, pp. 293–308, 2019. DOI: 10.1007/s10589-018-0046-7.

[18] P. H. Schönherrmann, “A generalized solution of the orthogonal Procrustes problem,” Psychometrika, vol. 31, no. 1, pp. 1–10, 1966. DOI: 10.1007/BF02289451.

[19] L. Eldén and H. Park, “A Procrustes problem on the Stiefel manifold,” Numerische Mathematik, vol. 82, no. 4, pp. 599–619, 1999. DOI: 10.1007/s002110050432.

[20] A. Beck, First-Order Methods in Optimization, ser. MOS-SIAM Series on Optimization. 2017. DOI: 10.1137/1.9781611974997.

[21] L. Mackey, “Deflation methods for sparse PCA,” in NIPS 2008: Advances in Neural Information Processing Systems 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2008, pp. 1017–1024. Available at: https://papers.nips.cc/paper/3575-deflation-methods-for-sparse-pca.

[22] G. I. Allen, “Sparse higher-order principal components analysis,” in AISTATS 2012: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics, N. D. Lawrence and M. Girolami, Eds., vol. 22, Canary Islands, Spain: PMLR, 2012, pp. 27–36. Available at: http://proceedings.mlr.press/v22/allen12.html.

[23] ——, “Multi-way functional principal components analysis,” in CAMSAP 2013: Proceedings of the 8th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, V. Cevher and S. Gazor, Eds., St. Martin, France: IEEE, 2013, pp. 220–223. DOI: 10.1109/CAMSAP.2013.6714047.

[24] L. R. Tucker, “Some mathematical notes on three-mode factor analysis,” Psychometrika, vol. 31, no. 3, pp. 279–311, 1966. DOI: 10.1007/BF02289464.

[25] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” SIAM Review, vol. 51, no. 3, pp. 455–500, 2009. DOI: 10.1137/07070111X.