Subspace Clustering with a Twist

David Wipf and Yue Dong
Microsoft Research, Beijing
{davidwip,yuedong}@microsoft.com

Bo Xin
Peking University
boxin@pku.edu.cn

Abstract

Subspace segmentation or clustering can be defined as the process of assigning subspace labels to a set of data points assumed to lie on the union of multiple low-dimensional, linear subspaces. Given that each point can be efficiently expressed using a linear combination of other points from the same subspace, a variety of segmentation algorithms built upon ℓ_1, nuclear norm, and other convex penalties have recently shown state-of-the-art robustness on multiple benchmarks. However, what if instead of observing the original data points, we instead only have access to transformed, or 'twisted' so to speak, measurements? Here we consider underdetermined affine transformations that may arise in computer vision applications such as bidirectional reflectance distribution function (BRDF) estimation. Unfortunately most existing approaches, convex or otherwise, do not address this highly useful generalization. To fill this void, we proceed by deriving a probabilistic model that simultaneously estimates the latent data points and subspace memberships using simple EM update rules. Moreover, in certain restricted settings this approach is guaranteed to produce the correct clustering. Finally a wide range of corroborating empirical evidence, including a BRDF estimation task, speaks to the practical efficacy of this algorithm.

1 Introduction

As a data reduction or analysis tool, principal component analysis (PCA) is readily applicable whenever observable points lie on or near a low-dimensional linear subspace. Richer structures however may not conform to this model, and often we must consider ways of introducing additional complexity. For example, a natural extension of PCA is to consider that our data lie on a union of low-dimensional subspaces. In this expanded regime we may then consider the joint problem of estimating these subspaces and assigning each point to the closest one, a process commonly referred to as either subspace clustering or segmentation. Although unlike classical PCA a closed-form solution via the SVD is no longer possible, tractable approximations that succeed with high probability form a core component of numerous practical application domains. Examples include the analysis of social graphs (Jalali et al., 2011), network topology inference (Eriksson et al., 2012), user identification in movie rating systems Zhang et al. (2012), and a host of computer vision tasks such as image representation and compression, motion segmentation, and face clustering (Elhamifar & Vidal, 2013; Feng et al., 2014; Liu et al., 2013; Lu et al., 2012; Rao et al., 2010).

1.1 Problem Description

We define this problem more formally as follows. Let $\{\mathcal{S}_k\}_{k=1}^m$ denote a collection of m linear subspaces in $\mathbb{R}^d$, where $\dim[\mathcal{S}_k] = d_k < d \quad \forall k = 1, \ldots, m$. Moreover, suppose we have drawn n_k points from each subspace forming data matrices $\boldsymbol{X}_k \in \mathbb{R}^{d \times n_k}$. We then concatenate the points from each subspace, and the full arrangement of $n = \sum_{k=1}^m n_k$ points is reordered using an unknown permutation matrix $\boldsymbol{P} \in \mathbb{R}^{n \times n}$. Consequently, the entire data can be expressed as

$$\boldsymbol{X} \triangleq [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m]\boldsymbol{P} \in \mathbb{R}^{d \times n}. \quad (1)$$

Subspace clustering can then be described as the process of estimating a basis for each $\mathcal{S}_k$ as well as the subspace membership of each point $\boldsymbol{x}_j$.

One of the most robust approaches to obtaining such accurate data segmentations exploits the so-called *self-expressiveness property* of $\boldsymbol{X}$ (Elhamifar & Vidal, 2013), namely that any $\boldsymbol{x}_j$ can be represented as a linear combination of other data points in $\boldsymbol{X}$ within the same subspace. Moreover, if we can find such a representation using *only* points from the same subspace, then we have extracted vital information pertaining to the true latent segmentation.

One way to favor such cluster-aligned decompositions is by solving

$$\min_{\boldsymbol{Z}} \|\boldsymbol{Z}\|_0 \ \text{ s.t. } \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z},\ \text{diag}[\boldsymbol{Z}] = 0, \tag{2}$$

where $\|\boldsymbol{Z}\|_0$ is the matrix ℓ_0 norm, or a count of the number of nonzero elements in $\boldsymbol{Z}$, a penalty function that strongly favors zero-valued elements or a canonically sparse $\boldsymbol{Z}$. The diagonal constraint is required to prevent each point from using itself in the representation (e.g., the degenerate solution $\boldsymbol{Z}^* = \boldsymbol{I}$), enforcing that we must rely only on others in the same subspace. If we assume that each individual subspace satisfies $d_k < d$ for all k, and that sampled points are sufficiently dense in general position, then the solution to (2) will be block diagonal and aligned with the true clusters up to the permutation matrix $\boldsymbol{P}$, revealing subspace memberships. A final spectral clustering, post-processing step can further solidify the labels and is adopted by most recent methods (Elhamifar & Vidal, 2013). Of course solving (2) is non-convex, discontinuous, and NP-hard, so following the typical compressive sensing recipe it is desirable to replace the troublesome $\|\boldsymbol{Z}\|_0$ penalty with the convex relaxation $\|\boldsymbol{Z}\|_1$. This substitution is supported by rigorous theoretical arguments detailing conditions whereby subspace-aligned block-diagonal structure is guaranteed when we minimize $\|\boldsymbol{Z}\|_1$ over the constraint set (Soltanolkotabi & Candès, 2012).

Proceeding further, suppose that we are unable to observe $\boldsymbol{X}$ directly, but instead are only granted access to a measurement matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]$, where each column $\boldsymbol{y}_j$ is generated via the underdetermined system

$$\boldsymbol{y}_j = \boldsymbol{A}_j \boldsymbol{x}_j, \quad \forall j = 1, \ldots, n. \tag{3}$$

Here $\{\boldsymbol{A}_j\}$, with $\boldsymbol{A}_j \in \mathbb{R}^{p_j \times d}, p_j \leq d$ indicates a set of known, possibly overcomplete matrices with problem-dependent structure that can warp or *twist* each data point independently while mapping it onto the lower-dimensional observation space.[1] As described in depth later, such a situation commonly arises in computer vision applications such as bidirectional reflectance distribution function (BRDF) estimation, where each $\boldsymbol{A}_j$ is determined by lighting conditions and the surface geometry of an object with unknown BRDF we would like to obtain. Other possible scenarios include face clustering in subject-varying transform domains, or motion segmentation using approximations for perspective cameras. Additionally, if each $\boldsymbol{A}_j$ can be described as a matrix of zeroes with a single one in each row, then the resulting estimation problem is tantamount to subspace clustering with missing entries (Candès et al., 2014; Eriksson et al., 2012; Gruber & Weiss, 2004; Yang et al., 2015).

[1] We frequently use $\{\boldsymbol{M}_j\}$ to abbreviate a set of matrices $\{\boldsymbol{M}_j : j \in \mathcal{J}\}$, where the index set $\mathcal{J}$ should be clear from the context.

1.2 Naive Solutions

Clearly we can no longer directly rely on the original self-expressiveness property, because once we insert $\{\boldsymbol{A}_j\}$ into the pipeline, it no longer follows that each corresponding $\boldsymbol{y}_j$ can be compactly represented using only other points generated from the same subspace.[2] To compensate, several strategies immediately come to mind.

For example, suppose we somehow knew the number of clusters m. Then let the set $\{\Omega_k\}$, with each $\Omega_k \subset \{1, \ldots, n\}$, denote a partitioning such that $\bigcup_{k=1}^m \Omega_k = \{1, \ldots, n\}$ and $\Omega_k \cap \Omega_{k'} = \emptyset$ for all pairs $\{k, k'\}$. Also let $\boldsymbol{X}_{\Omega_k}$ represent the columns of matrix $\boldsymbol{X}$ indexed by Ω_k. Now consider the joint optimization over all possible segmentations and latent points

$$\min_{\boldsymbol{X}, \{\Omega_k\}} \sum_{k=1}^m |\Omega_k| \text{rank}(\boldsymbol{X}_{\Omega_k}) \ \text{ s.t. } \boldsymbol{y}_j = \boldsymbol{A}_j \boldsymbol{x}_j,\ \forall j = 1, \ldots, n. \tag{4}$$

Then assuming the true latent $\boldsymbol{X}$ is composed of sufficient samples per subspace in general position, and that $\boldsymbol{A}_j$ contains a sufficient number of non-degenerate measurements, the solution to (4) will be such that $\{\Omega_k\}$ reflects the correct segmentation and $\boldsymbol{X}$ will be recovered. Unfortunately however, minimizing (4) requires an infeasible, combinatorial search over every possible clustering pattern.

Perhaps the most natural way to circumvent this problem is to invoke a two-stage procedure inspired by traditional matrix completion (Candès & Recht, 2008). The basic idea is to first obtain an estimate of the latent $\boldsymbol{X}$ by solving

$$\min_{\boldsymbol{X}} \text{rank}[\boldsymbol{X}], \quad \text{s.t. } \boldsymbol{y}_j = \boldsymbol{A}_j \boldsymbol{x}_j,\ \forall j = 1, \ldots, n, \tag{5}$$

which excludes any combinatorial search over labels. This represents an affine rank minimization problem that can be approximately solved by replacing the non-convex rank$[\boldsymbol{X}]$ penalty with the convex nuclear norm relaxation $\|\boldsymbol{X}\|_*$, or the sum of the singular values of $\boldsymbol{X}$. Once this solution is in hand, we may deploy any traditional subspace clustering algorithm on the resulting $\hat{\boldsymbol{X}}$.

The difficulty with this strategy is two-fold. First, unlike the data from individual subspaces $\boldsymbol{X}_k$, the matrix $\boldsymbol{X}$ may be full rank given that it is quite common to have $\sum_k d_k \geq d$. So in this situation we have no chance of obtaining a meaningful segmentation. However, even if the global solution to (5) does produce the correct $\boldsymbol{X}$, the nuclear norm relaxation required by a tractable implementation will be highly sensitive to both the correlation structure and relative column norm scaling of $\{\boldsymbol{A}_j\}$.

[2] An exception to this occurs when $\boldsymbol{A}_j$ is equal to some fixed $\boldsymbol{A}$ across all j, in which case the self-expressiveness property still holds and natural adaptations already exist (Patel et al., 2013; Wang et al., 2015a).

In fact existing theoretical guarantees for rank-nuclear norm equivalence place extremely strong conditions on the structure of the measurement process, which are unlikely to hold in practice here since in the problem instances we consider, each $\boldsymbol{A}_j$ is determined by physical properties of the experimental design. Moreover, unlike typical compressive sensing designs, we cannot even normalize columns of $\{\boldsymbol{A}_j\}$, because if we were to do so then a low-rank solution will no longer satisfy the constraint set in (5). Therefore we are left with a challenging NP-hard rank minimization problem as a required preprocessing step, a clearly undesirable starting point.

As an alternative to the above two-stage procedure, we could append an additional data fitting constraints to the canonical sparse subspace clustering objective from above and solve

$$\min_{\boldsymbol{X},\boldsymbol{Z}} \|\boldsymbol{Z}\|_1 \text{ s.t. } \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z},\ \text{diag}[\boldsymbol{Z}] = 0, \boldsymbol{y}_j = \boldsymbol{A}_j\boldsymbol{x}_j\ \forall j. \tag{6}$$

Although the penalty is convex in $\boldsymbol{Z}$ and the constraints are individually convex in $\boldsymbol{X}$ and $\boldsymbol{Z}$, the overall problem remains highly non-convex and difficult to optimize. Moreover it is unclear whether the global solution, even if somehow attainable, would guarantee that the correct clustering could be found. Additionally, in practical environments with noisy data, relaxing the additional equality constraints would require the inclusion of an additional trade-off parameter and application-specific tuning.

1.3 Overview of Contributions

To address the conceptual limitations of naive adaptations of existing subspace clustering approaches, in Section 2 we derive an alternative Bayesian approach specifically tailored to the proposed latent variable setting and loosely motivate its effectiveness. Next Section 3 derives expectation maximization (EM) udpdate rules that accommodate practical deployment. We then proceed to theoretical analysis of the underlying objective function in Section 4, followed by empirical validation in Section 5, practical deployment in Section 6, and final contextualization in Section 7. Overall, our contributions can be summarized as follows:

- We delineate an important generalization of subspace clustering to accommodate an underdetermined affine measurement process and derive a Bayesian algorithm that explicitly circumvents limitations of natural alternatives. Unlike existing state-of-the-art segmentation pipelines, our algorithm does not require a final spectral clustering step.
- We thoroughly unpack the proposed objective function and its customized mechanism for favoring the true subspace labels. This includes the exposition of specific conditions, albeit somewhat idealized, whereby a unique minimizing solution (global or local) will produce the correct segmentation.
- Although not our original intention, we demonstrate that our model can achieve state-of-the-art performance in more specialized domains when $\boldsymbol{A}_j$ displays certain additional structure. This includes traditional subspace clustering when each $\boldsymbol{A}_j = \boldsymbol{I}$ for all j, or subspace clustering with missing entries when each $\boldsymbol{A}_j$ is an all-zero matrix with a single one in each row.
- We provide strong empirical validation on a practical BRDF estimation problem that requires the full generality of the proposed affine observation model.

2 Model Description

We begin by decomposing the latent unobserved data as

$$\boldsymbol{X} = \sum_{i=1}^{m} \widetilde{\boldsymbol{X}}^{(i)}, \tag{7}$$

where each $\widetilde{\boldsymbol{X}}^{(i)} \in \mathbb{R}^{d\times n}$ can be interpreted as our estimate of the overall signal generated from the i-th subspace. Although we will eventually arrive at an objective function that is independent of this decomposition, it nonetheless serves as a useful tool for constructing hidden data for the EM algorithm described in Section 3. We next adopt the Gaussian likelihood function

$$p\left(\boldsymbol{Y}|\{\widetilde{\boldsymbol{X}}^{(i)}\};\lambda\right) \propto \exp\left[-\sum_j \tfrac{1}{2\lambda}\|\boldsymbol{y}_j - \boldsymbol{A}_j\sum_i \widetilde{\boldsymbol{x}}_j^{(i)}\|_2^2\right], \tag{8}$$

where λ is a noise parameter.[3] Additionally, in the limit as $\lambda \to 0$ this will enforce the same constraint set as in (3). Next we define an independent, zero-mean Gaussian prior distribution for each column of $\widetilde{\boldsymbol{X}}^{(i)}$, parameterized as $p\left(\{\widetilde{\boldsymbol{X}}^{(i)}\};\{\boldsymbol{\Gamma}_i\},\boldsymbol{W}\right) =$

$$\prod_{i,j} p(\widetilde{\boldsymbol{x}}_j^{(i)};\boldsymbol{\Gamma}_i, w_{ij}) = \prod_{i,j}\mathcal{N}\left(\widetilde{\boldsymbol{x}}_j^{(i)};\boldsymbol{0}, w_{ij}\boldsymbol{\Gamma}_i\right), \tag{9}$$

where each $\boldsymbol{\Gamma}_i$ represents a symmetric, positive semi-definite covariance basis matrix and the scalar coefficients w_{ij} constitute non-negative weighting factors that collectively form a parameter matrix $\boldsymbol{W}$.[4] Strictly speaking

[3] We could allow for a separate λ_j for each point which would ultimately allow us to learn outlier locations, but for space considerations we do not further pursue this direction here.

[4] Note that (Tipping & Bishop, 1999; Wang et al., 2015b) describe alternative probabilistic mixture models that can be applied to clustering; however, the parameterizations and underlying inference algorithms are completely different from ours, do not apply to the latent affine model we consider, and do not lead to any of the desired properties discussed herein.

we should require that each $w_{ij}\boldsymbol{\Gamma}_i$ factor be positive definite such that the implied matrix inverse included with this distribution is well-defined. However, we can adjust for the semi-definite case with a more refined definition of the prior. First, if some $w_{ij} = 0$, then we simply define that $\widetilde{\boldsymbol{x}}_j^{(i)} = \boldsymbol{0}$ with probability one. For the $w_{ij} > 0$ case, without loss of generality assume that $\boldsymbol{\Gamma}_i = \boldsymbol{R}_i\boldsymbol{R}_i^\top$ for some matrix $\boldsymbol{R}_j$. We then stipulate that $p\left(\widetilde{\boldsymbol{x}}_j^{(i)}; \boldsymbol{\Gamma}_i, w_{ij}\right) = 0$ if $\widetilde{\boldsymbol{x}}_j^{(i)} \notin \text{span}[\boldsymbol{R}_j]$, and $p\left(\widetilde{\boldsymbol{x}}_j^{(i)}; \boldsymbol{\Gamma}_i, w_{ij}\right) \propto \exp\left[-\frac{1}{2}(\widetilde{\boldsymbol{x}}_j^{(i)})^\top(\boldsymbol{R}_j^\top)^\dagger\boldsymbol{R}_j^\dagger\widetilde{\boldsymbol{x}}_j^{(i)}\right]$ otherwise. These refinements are tacitly assumed in many related Bayesian models, and can be viewed as a natural limiting case whereby a degenerate covariance enforces that all probability mass reside in a low-dimensional subspace of the full ambient space.

Given that both the likelihood function and prior distribution are Gaussians, the posterior distribution is also a Gaussian with closed-form moments. While expressing these moments in full is slightly cumbersome from a notational standpoint, the marginalized posterior of each $\widetilde{\boldsymbol{x}}_j^{(i)}$ is given by

$$p\left(\widetilde{\boldsymbol{x}}_j^{(i)}|\boldsymbol{y}_j; \boldsymbol{\Gamma}_i, w_{ij}, \lambda\right) = \mathcal{N}\left(\widetilde{\boldsymbol{x}}_j^{(i)}; \boldsymbol{\mu}_j^{(i)}, \boldsymbol{\Sigma}_j^{(i)}\right) \quad (10)$$

with means and covariances defined by

$$\begin{aligned}\boldsymbol{\mu}_j^{(i)} &= w_{ij}\boldsymbol{\Gamma}_i\boldsymbol{A}_j^\top\left(\lambda\boldsymbol{I} + \boldsymbol{A}_j\boldsymbol{\Psi}_j\boldsymbol{A}_j^\top\right)^{-1}\boldsymbol{y}_j, \quad (11)\\ \boldsymbol{\Sigma}_j^{(i)} &= w_{ij}\boldsymbol{\Gamma}_i - w_{ij}^2\boldsymbol{\Gamma}_i\boldsymbol{A}_j^\top\left(\lambda\boldsymbol{I} + \boldsymbol{A}_j\boldsymbol{\Psi}_j\boldsymbol{A}_j^\top\right)^{-1}\boldsymbol{A}_j\boldsymbol{\Gamma}_i,\end{aligned}$$

where

$$\boldsymbol{\Psi}_j = \sum_i w_{ij}\boldsymbol{\Gamma}_i, \quad \forall j = 1, \ldots, n, \quad (12)$$

Although it is easily shown that each $\widetilde{\boldsymbol{x}}_j^{(i)}$ is independent across data points j, they may be highly correlated across the basis index i. However, for purposes of the EM algorithm derived in Section 3, only the moments from (11) will be required.

The rationale for the chosen parameterization of the prior $p\left(\{\widetilde{\boldsymbol{X}}^{(i)}\}; \{\boldsymbol{\Gamma}_i\}, \boldsymbol{W}\right)$ becomes partially evident upon inspection of the posterior mean expression from (11). Suppose each $\boldsymbol{\Gamma}_i$ spans the k-th unknown subspace we would like to recover. And moreover, suppose each $\boldsymbol{w}_j$ (the j-th column of $\boldsymbol{W}$) is a vector of zeros with a single nonzero in the position corresponding with the true subspace membership of $\boldsymbol{x}_j$. Then by virtue of the left multiplication in (11), $\boldsymbol{x}_j$ will have a posterior mean constrained to the correct subspace, with zero covariance (or posterior mass) leaking into other, errant subspaces. Hence under the stated conditions a posterior mean estimator will produce minimal reconstruction error.

Of course all of this is predicated on our ability to actually obtain a basis set $\{\boldsymbol{\Gamma}_i\}$ and weight matrix $\boldsymbol{W}$ fulfilling the stringent subspace-aware criterion described above. Hence we have merely shifted our original goal of estimating $\boldsymbol{X}$ and clustering its columns, to the task of learning subspace-aware covariances $\{\boldsymbol{\Gamma}_i\}$ and a column-sparse weight matrix $\boldsymbol{W}$ with support aligned with the true subspaces. While certainly not immediately obvious, the remainder of this paper will demonstrate that a standard marginalization strategy is quite effective for this purpose.

If we treat $\{\boldsymbol{\Gamma}_i\}$ and $\boldsymbol{W}$ as the key parameters of interest and $\{\widetilde{\boldsymbol{X}}^{(i)}\}$ as nuisance latent variables, then a common Bayesian inference strategy is to marginalize over $\{\widetilde{\boldsymbol{X}}^{(i)}\}$ and then maximize the resulting likelihood function with respect to remaining unknown parameters (Tipping, 2001; Wipf et al., 2011; Xin & Wipf, 2015). This involves solving

$$\max_{\boldsymbol{\Gamma}_i \in H^+ \forall i, \boldsymbol{W} \geq 0} \int p(\boldsymbol{Y}|\boldsymbol{X}; \lambda)p(\boldsymbol{X}; \{\boldsymbol{\Psi}_j\})d\boldsymbol{X}, \quad (13)$$

where H^+ denotes the set of positive semi-definite and symmetric $d \times d$ matrices. After a $-2\log$ transformation and application of a standard convolution-of-Gaussians integration (Tipping, 2001), solving (13) is equivalent to minimizing the cost function

$$\mathcal{L}(\{\boldsymbol{\Gamma}_i\}, \boldsymbol{W}) = \sum_j \boldsymbol{y}_j^\top\boldsymbol{\Sigma}_{y_j}^{-1}\boldsymbol{y}_j + \log\left|\boldsymbol{\Sigma}_{y_j}\right|, \quad (14)$$

where

$$\boldsymbol{\Sigma}_{y_j} = \sum_i w_{ij}\boldsymbol{A}_j\boldsymbol{\Gamma}_i\boldsymbol{A}_j^\top + \lambda\boldsymbol{I}. \quad (15)$$

The latter represents the covariance of $\boldsymbol{y}_j$ conditioned on $\{\boldsymbol{\Gamma}_i\}$ and $\boldsymbol{w}_j$.

3 Algorithm Derivation

To optimize $\mathcal{L}(\{\boldsymbol{\Gamma}_i\}, \boldsymbol{W})$ we may treat $\{\widetilde{\boldsymbol{X}}^{(i)}\}$ as hidden data and execute a straightforward EM procedure (Dempster et al., 1977) similar to that proposed in (Tipping, 2001). For the E-step we need only compute the posterior moments given by (11). For the M-step we must solve

$$\min_{\{\boldsymbol{\Gamma}_i\}, \boldsymbol{W}} \text{E}\left[-\log p\left(\{\widetilde{\boldsymbol{X}}^{(i)}\}, \boldsymbol{Y}; \{\boldsymbol{\Gamma}_i\}, \boldsymbol{W}, \lambda\right)\right], \quad (16)$$

where the expectation is with respect to $p\left(\{\widetilde{\boldsymbol{X}}^{(i)}\}|\boldsymbol{Y}; \{\boldsymbol{\Gamma}_i'\}, \boldsymbol{W}', \lambda\right)$, which represents the posterior distribution obtained using moments parameterized with fixed values $\{\boldsymbol{\Psi}_j'\}$ and $\boldsymbol{W}'$ computed from the previous iteration. After a few algebraic manipulations, this is equivalent to solving

$$\min_{\{\boldsymbol{\Gamma}_i\}, \boldsymbol{W}} \sum_{i,j} \text{tr}\left[\text{E}\left[\widetilde{\boldsymbol{x}}_j^{(i)}(\widetilde{\boldsymbol{x}}_j^{(i)})^\top\right](w_{ij}\boldsymbol{\Gamma}_i)^{-1}\right] + \log|w_{ij}\boldsymbol{\Gamma}_i|, \quad (17)$$

where $\mathrm{E}\left[\widetilde{\boldsymbol{x}}_j^{(i)}(\widetilde{\boldsymbol{x}}_j^{(i)})^\top\right] = \boldsymbol{\mu}_j^{(i)}\left(\boldsymbol{\mu}_j^{(i)}\right)^\top + \boldsymbol{\Sigma}_j^{(i)}$. Unfortunately (17) has no closed-form solution. However, we can first optimize over $\{\boldsymbol{\Gamma}_i\}$ with $\boldsymbol{W}$ fixed, and then optimize over $\boldsymbol{W}$ with $\{\boldsymbol{\Gamma}_i\}$ fixed, both of which have closed-form solutions. Although these updates could be iterated until convergence, the EM algorithm does not actually require full completion of both E and M steps. In fact partial minimization, or incremental variants, are adequate to ensure cost function descent (Neal & Hinton, 1999).[5]

For the $\{\boldsymbol{\Gamma}_i\}$ update, we can solve for each $\boldsymbol{\Gamma}_i$ independently via

$$\boldsymbol{\Gamma}_i^* = \arg\min_{\boldsymbol{\Gamma}_i} \mathrm{tr}\left[\Theta_i \boldsymbol{\Gamma}_i^{-1}\right] + n\log|\boldsymbol{\Gamma}_i| \;=\; \tfrac{1}{n}\Theta_i \qquad (18)$$

where

$$\Theta_i = \sum_j \tfrac{1}{w_{ij}} \mathrm{tr}\left[\left(\boldsymbol{\mu}_j^{(i)}\left(\boldsymbol{\mu}_j^{(i)}\right)^\top + \boldsymbol{\Sigma}_j^{(i)}\right)\right]. \qquad (19)$$

Likewise for $\boldsymbol{W}$ we can solve independently for each element using

$$w_{ij}^* = \arg\min_{\boldsymbol{\Gamma}_i} \beta_{ij} w_{ij}^{-1} + d\log w_{ij} \;=\; \tfrac{1}{d}\beta_{ij}, \qquad (20)$$

where

$$\beta_{ij} = \mathrm{tr}\left[\left(\boldsymbol{\mu}_j^{(i)}\left(\boldsymbol{\mu}_j^{(i)}\right)^\top + \boldsymbol{\Sigma}_j^{(i)}\right)\boldsymbol{\Gamma}_i^{-1}\right]. \qquad (21)$$

To summarize then, we need only iterate (11), (18), and (20) to descend the objective function (14). With some attention to details, this can be accomplished with updates that are linear in n and m, the number of points and clusters respectively), and cubic in d (ambient space dimension).

A final point worth addressing is initialization. Assuming complete agnosticism regarding subspaces and labels, the selection $\boldsymbol{\Gamma}_i = \boldsymbol{I}$ and $w_{ij} = 1$ for all i and j seems like the most natural choice. However, we require some small degree of symmetry breaking randomness to initiate a non-degenerate descent. We simply use $w_{ij} \sim 1 + \mathrm{U}\left[0, 10^{-3}\right]$ for all initializations, although results are not sensitive to this choice.

4 Cost Function Analysis

While perhaps counterintuitive, the proposed objective function (14) has a number of desirable attributes that justify its usage for latent subspace clustering. As motivation for this claim, it is helpful to map the arguments of (14) to a criterion of subspace optimality. More formally, we say that $\{\{\boldsymbol{\Gamma}_i^*\}, \boldsymbol{W}^*\}$ is a *subspace optimal solution* iff

1. For all $i = 1, \ldots, m$, span$[\boldsymbol{\Gamma}_i^*]$ equals some true $\mathcal{S}_k$, and no two $\boldsymbol{\Gamma}_i^*$ span the same subspace.
2. For all $j = 1, \ldots, n$, $\|\boldsymbol{w}_j^*\|_0 = 1$, with nonzero element aligned with the correct subspace.

Such a solution guarantees that an accurate estimate of $\boldsymbol{X}$ can be obtained via (11), and that the correct subspace labels will be recovered. The remainder of this section will quantify how such solutions relate to minima of (14).

To begin, using duality arguments from (Wipf et al., 2011), there is a close association between global minima of (4) and (14) in terms of the recovered subspaces and labels. However, none of this is suggestive of why we might prefer dealing with the latter over say, brute force combinatorial optimization of the former. For this purpose we need to actually describe conditions whereby (14) is more likely to produce subspace optimal solutions without getting stuck at local optimal. While it is quite challenging to address this situation in sweeping terms for such a coupled, non-convex probabilistic model, we will nonetheless describe at least one scenario where bad local optimal can be fully eradicated, followed by more general conditions whereby optimal non-increasing solution paths exist.

For convenience, let $\{\Omega_k^*\}$ denote the true partitioning of $\boldsymbol{X}$, aligned with the presumed generative subspace labels. We then have the following:

Theorem 1. *Suppose that we have a data matrix $\boldsymbol{X}$ which follows the model from (1), we observe the affine measurements $\boldsymbol{y}_j = \boldsymbol{A}_j\boldsymbol{x}_j$ for all j, and that the true latent $\boldsymbol{X}$ is such that $d_k = 1$ for all subspaces. Furthermore assume that $\{\boldsymbol{A}_j\}$ satisfies $p_j > 1$ for all j and $\cap_{j\in\Omega_k^*} null[\boldsymbol{A}_j] = \emptyset$ for all k. Then any local or global minimizer $\{\{\boldsymbol{\Gamma}_i^*\}, \boldsymbol{W}^*\}$ of (14) in the limit $\lambda \to 0$ is subspace optimal.*

At least in low-noise/stylized conditions, this result specifies a relatively broad regime whereby no suboptimal minima exist, meaning any minimizer (local or global) will always return the correct clustering as well as a unique basis spanning each true subspace.[6] And this result is emblematic of a wider range of operating circumstances whereby subspace optimal solutions are closely aligned with minima of (14). Certainly our empirical evidence provided in Sections 5 and 6 suggests this to be the case.

Interestingly though, neither of the naive approaches discussed in Section 1 can satisfy something similar. In fact,

[5] While technically these updates are guaranteed to reduce or leave-unchanged the objective function until a fixed point is reached, to formally guarantee convergence of the EM algorithm to a local minima requires additional effort, such as the demonstration that the conditions of Zangwill's Global Convergence Theorem have been satisfied (Zangwill, 1969). We do not pursue a detailed theoretical investigation to this effect here, although it is possible to do so.

[6] In the context of affine rank minimization and a single subspace, i.e., $m = 1$, it has been shown that under similar conditions no bad local minima will exist with a probabilistic PCA-like model (Xin & Wipf, 2015); however, this is a much simpler problem and the same analysis/proof techniques do not apply.

under the stated conditions of Theorem 1, (4) can have numerous suboptimal local minima, while (6) can have both suboptimal local and global minima, both of which can return incorrect labels and cluster bases.

Likewise, if we replace the rank function with the nuclear norm in (5), then even with all other theorem specifications in place, it is still possible that we recover the wrong estimate for $\boldsymbol{X}$ such that no correct clustering is possible via any secondary step. As an example, it is a simple matter to design adversarial conditions on $\boldsymbol{A}_j$ via simple transformations such as $\boldsymbol{A}_j \rightarrow \boldsymbol{A}_j\boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal scaling matrix to which the nuclear norm solution will be highly sensitive. And as stated previously, we cannot negate the impact of $\boldsymbol{D}$ via normalization without destroying the low-rank assumption with which estimating $\boldsymbol{X}$ is predicated on to begin with. Moreover, it is also possible to have a full rank $\boldsymbol{X}$ consistent with the setting of Theorem 1 such that it is formally unidentifiable even with (5) unaltered.

Moving forward, it is considerably more difficult to guarantee that no bad local minima exist under broader conditions, e.g., when $d_k > 1$. However, we can still analyze non-increasing paths between a family of initializations (or intermediate points in some optimization trajectory) and subspace optimal solutions. This simplified analysis criteria yields the following:

Theorem 2. *Suppose* $\sum_i w_{ij}\boldsymbol{\Gamma}_i = \alpha_j\boldsymbol{U}\boldsymbol{U}^\top$ *for all* j*, where* $\boldsymbol{U}$ *represents any orthonormal basis spanning* $\bigoplus_{k=1}^m \mathcal{S}_k$ *and each* $\alpha_j > 0$ *is a scalar weighting factor. Then in the limit* $\lambda \rightarrow 0$*, if each* α_j *is suitably large there exists a non-increasing path from this point to some subspace optimal solution* $\{\{\boldsymbol{\Gamma}_i^*\}, \boldsymbol{W}^*\}$.

Corollary 1. *In the simplified scenario when* $\boldsymbol{A}_j = \boldsymbol{I}$ *for all* j *(i.e., canonical subspace clustering where the latent* $\boldsymbol{X} = \boldsymbol{Y}$ *are now fully observable), Theorem 2 holds without any size restrictions on each* $\alpha_j > 0$.

Because we can always choose to initialize with $\boldsymbol{\Gamma}_i = \boldsymbol{I}$ for all i, or more generally $\boldsymbol{\Gamma}_i$ equal to some suitable $\boldsymbol{U}\boldsymbol{U}^\top$, then a byproduct of Theorem 2 is the insurance that a path exists from a computable point to the correct clustering that is devoid of local minima even when $\boldsymbol{W}$ is initialized arbitrarily. And this result can be generalized with additional effort to quantify a broader class of locations such that such paths to optimal solutions exist. Of course obviously a result of this type is still quite limited in that it does not guarantee that such a path can be found, or rule out the existence of saddle points along the way. But it is nonetheless another indicator of the appropriateness of (14) in addressing even basic subspace clustering problems for which it was not initially designed. And similar to previous arguments, neither of the naive approaches, i.e., solving either (4) or (6), can satisfy something similar.

5 Simulation Experiments

We now present illustrative synthetic experiments tailored to showcase generic abilities, with designs and dimensions inspired by (Soltanolkotabi & Candès, 2012; Yang et al., 2015).

5.1 Fully Observable Model

We begin by investigating the original subspace clustering problem where $\boldsymbol{Y} = \boldsymbol{X}$. In particular, we examine challenging conditions where there exists a significant degree of subspace overlap similar to an experimental design from (Soltanolkotabi & Candès, 2012). Data are generated as follows. Three subspaces of dimension $d_1 = d_2 = d_3 = 20$ are embedded in $\mathbb{R}^{25}$, each containing 50 data points. This is accomplished for each subspace by generating $\boldsymbol{X}_k = \boldsymbol{U}_k\boldsymbol{V}_k^\top$, where $\boldsymbol{U}_k \in \mathbb{R}^{25\times 20}$ and $\boldsymbol{V} \in \mathbb{R}^{50\times 20}$ have iid $\mathcal{N}(0,1)$ elements. With probability one the resulting $\boldsymbol{X}$ will be full rank with significantly overlapping subspace magisteria. We then normalize each column to have unit ℓ_2 norm, and apply the state-of-the-art ℓ_1-norm based subspace clustering mentioned in Section 1, denoted ℓ_1-SSC, to sort out subspace labels. This algorithm involves solving (2) with $\|\boldsymbol{Z}\|_1$ replacing $\|\boldsymbol{Z}\|_0$, forming the symmetric affinity matrix $|\hat{\boldsymbol{Z}}| + |\hat{\boldsymbol{Z}}^\top|$ using the estimated $\hat{\boldsymbol{Z}}$, followed by a separate spectral clustering step with knowledge of the true number of clusters m (Elhamifar & Vidal, 2013). For our algorithm we assign cluster labels based on the index of the largest value of each estimated $\boldsymbol{w}_j$ (typically though there is only a single entry significantly larger than zero when the clustering is successful); no spectral clustering heuristic is required.

We note however that by drawing the data using such iid Gaussian isotropic sources (as is typically done for experimental purposes), the data within each subspace will lack any significant structure or correlation, to which the ℓ_1 norm solution can be highly sensitive. Hence to deviate from the relatively easier, isotropic situation, we gradually experiment with increasing the degree of intra-subspace correlation by adding a rank-one component $\alpha\|\boldsymbol{U}_k\boldsymbol{V}_k^\top\|_2\boldsymbol{a}_k\boldsymbol{b}_k^\top$ to each subspace, where vectors $\boldsymbol{a}_k \in \mathbb{R}^{25\times 1}$ and $\boldsymbol{b}_k \in \mathbb{R}^{50\times 1}$ are also iid $N(0,1)$ and α is a non-negative scalar that weights the contribution.

Figure 1 displays the clustering errors (percentage of mislabeled points) for both ℓ_1-SSC and our method averaged over 10 trials. We observe that when the correlation parameter $\alpha = 0$, both methods perform well, but as soon as α begins increasing, the quality of ℓ_1-SSC solutions degrades significantly, unlike our algorithm which is stable across all values. Hence even when no latent affine structure or the twist is present (the fully observable case with $\boldsymbol{A}_j = \boldsymbol{I}$), minimizing (14) represents a principled objective function for subspace clustering.

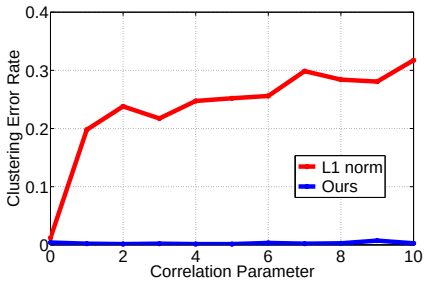

Figure 1: Comparisons using a fully-observable model as the within-subspace correlation is increased.

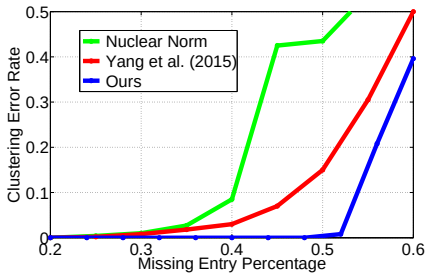

Figure 2: Comparisons using a partially-observable model as the number of missing entries is increased.

5.2 Missing Entries

Next we address the case were we have observed some $\boldsymbol{X}$ with a certain proportion of missing entries. As discussed in Section 1, this is equivalent to assuming that each $\boldsymbol{A}_j$ is a matrix of zeros with a single one per row. For this particular special case, (Yang et al., 2015) has proposed a modification of ℓ_1-SSC whereby missing entries are set to zero but partially compensated for using a special projection step.[7] Although this method cannot be extended to general $\{\boldsymbol{A}_j\}$, we can nonetheless evaluate our approach against this missing entry specialization. For this purpose, we select the most difficult clustering test from (Yang et al., 2015), whereby the latent $\boldsymbol{X}$ is full rank and the number of missing entries grows large.

Following (Yang et al., 2015), we generate $m = 5$ subspaces, each of dimension 5, embedded in $d = 25$ dimensional space. Next 50 points are drawn from each subspace using the same Gaussian factorization from above (and $\alpha = 0$). The fraction of missing entries is then gradually increased to test performance. Figure 2 displays the results, including a common baseline nuclear norm estimate of $\boldsymbol{X}$ followed by ℓ_1-SSC subspace clustering. Again we observe that, even without any spectral clustering step as used by others, our algorithm outperforms state-of-the-art existing approaches, including all variety of algorithms from (Yang et al., 2015) that were specifically designed for this problem.

5.3 General Affine Model

Finally, we consider our original motivating scenario where $\{\boldsymbol{A}_j\}$ can be arbitrary. To this end, we repeat the experiment from above, but with the binary sampling matrices replaced with elements of each $\boldsymbol{A}_j$ drawn iid from $\mathcal{N}(0,1)$. We also fix the number of measurements per point to $p_j = 15$ for each j; other dimensions remain unchanged.

[7] Actually (Yang et al., 2015) presents multiple approaches for handling missing entries, including another method from (Candès et al., 2014); however, we compare against the best performing variant among all of these.

Figure 3 explores the ability to recover the true latent $\boldsymbol{X}$ compared to the minimum nuclear norm solution, which represents the most viable existing alternative. We also tried minimizing (6); however, the results where quite poor (much worse than the nuclear norm solution) because of unavoidable convergence to local minima.

While the true number of clusters is $m = 5$, we vary the number of clusters assumed by our algorithm from $\hat{m} = 1, \ldots, 20$ and record the normalized MSE given by $\langle \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_{\mathcal{F}}^2 / \|\boldsymbol{X}\|_{\mathcal{F}}^2 \rangle$ averaged over 10 trials. Moreover, if we successfully recover $\boldsymbol{X}$ with $\hat{m} \neq m$, it is trivial to either fuse redundant clusters or split merged clusters using simple existing subspace clustering approaches to obtain labels if required.

Note that it is possible to exactly recover $\boldsymbol{X}$ with $\hat{m} \neq m$, provided $\hat{m}$ is sufficiently large such that $\boldsymbol{X}$ is identifiable. More specifically, to even have a chance of recovery for any possible algorithm, it must be the case that for all clusters k, the number of degrees-of-freedom in each associated low rank $\boldsymbol{X}_k$ (the points within cluster k) is less than the number of measurements of $\boldsymbol{X}_k$. For example, in the present case each $\boldsymbol{X}_k \in \mathbb{R}^{25 \times 50}$ has $5 \times (25 + 50) - 5^2 = 350$ degrees-of-freedom, and we have $\sum_{j \in \Omega_k} p_j = 15 \times 50 = 750$ measurements per subspace to work with (more than double the d.o.f.), which should be sufficient if $\hat{m} = 5$. However, when $\hat{m} < 5$, then two or more subspaces must be merged for estimation purposes leading to at least one $5 + 5 = 10$ dimensional subspace with $50 + 50 = 100$ points, and $10 \times (25 + 100) - 10^2 = 1150$ degrees-of-freedom, but only $15 \times 100 = 1500$ measurements. Even if we knew the true subspace labels, recovering $\boldsymbol{X}$ would still then be extremely challenging given how close the number of measurements are to the degrees-of-freedom.

But of course we still need to learn the labels as well, compounding the difficulty dramatically such that success by any possible algorithm is suspect. Therefore we should expect failure with $\hat{m} < 5$ on theoretical grounds, and indeed, from Figure 3 the error increases monotonically as $\hat{m}$ is decreased below 5. In contrast, for $\hat{m} > 5$, we observe that over-segmentation has minimal effect in disrupting the es-

timation of $\boldsymbol{X}$, and our algorithm has dramatically lower MSE than the nuclear norm solution; it only actually begins to rise appreciably for $\hat{m} > 17$. At this point presumably the large degree of superfluous over-segmentation may increase the risk of local minima as the parameter space becomes unnecessarily large.

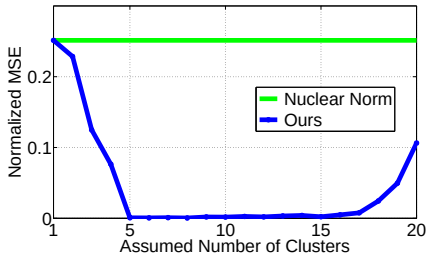

Figure 3: Comparisons using the general affine model as the number of assumed clusters is varied. Minimizing (6) led to a normalized MSE above 1.0 for all cases and initializations we tried (not shown).

6 Application Example: BRDF Estimation

One interesting application of the proposed method is surface reflectance reconstruction. Here surface reflectance simply refers to how a given surface reflects light. Moreover, if we have access to an accurate estimate, then we can compute exactly how a given object and material will appear under any lighting condition and viewing direction, which is extremely useful in many computer vision and graphics domains. From a technical standpoint, surface reflectance properties can be quantified by a spatially varying bi-directional reflectance distribution function (BRDF), which encodes the ratio between the *incoming* radiance from lighting direction θ_{in} and the *outgoing* radiance to the viewing direction θ_{out}, at each surface point j on some object or scene of interest. Although the BRDF represents an inherent property of the underlying materials, it is quite difficult to acquire since what we actually perceive from an object is jointly dependent on lighting conditions, viewing direction, and the BRDF itself.

More concretely, the observed outgoing radiance $y_j(\theta_{out})$ at direction θ_{out} can be expressed as the product of the surface BRDF $\rho_j(\theta_{out}, \theta_{in})$ at point j and the incoming radiance $r(\theta_{in})$ from direction θ_{in} integrated over all lighting directions $\mathcal{D}$, giving

$$y_j(\theta_{out}) = \int_{\mathcal{D}} \rho_j(\theta_{out}, \theta_{in}) r(\theta_{in}) d\theta_{in}. \qquad (22)$$

Moreover, the surface reflectance of each surface pixel can be expressed, to close approximation, as a linear combination of basis functions via

$$\rho_j(\theta_{out}, \theta_{in}) = \sum_{i=1}^{21} x_{ij} \rho_i(\theta_{out}, \theta_{in}), \qquad (23)$$

where $\boldsymbol{x}_j = [x_{1j}, \ldots, x_{21j}]^\top$ are weights and each $\rho_i(\theta_{out}, \theta_{in})$ represents a Cook-Torrance BRDF basis function for $i \in \{1, \ldots, 20\}$ and a Lambertian reflectance function for $i = 21$ (Lawrence et al., 2006; Dong et al., 2010; Chen et al., 2014). Combining with (22) this yields

$$y_j(\theta_{out}) = \sum_{i=1}^{21} x_{ij} \int_{\mathcal{D}} \rho_i(\theta_{out}, \theta_{in}) r(\theta_{in}) d\theta_{in}. \qquad (24)$$

With known lighting conditions and incoming radiance $r(\theta_{in})$, and the fixed known basis $\rho_i(\theta_{out}, \theta_{in})$, the integral components of (24) can be pre-computed. Additionally, measurements from multiple viewing directions can be packed into the vector $\boldsymbol{y}_j = [y_j(\theta_{out_1}), \ldots, y_j(\theta_{out_p})]^\top$ for each point j, and the corresponding integrals of the basis function can also be packed similarly in to a matrix $\boldsymbol{A}_j$ with $(A_j)_{ti} = \int_{\mathcal{D}} \rho_i(\theta_{out_t}, \theta_{in}) r(\theta_{in}) d\theta_{in}$, producing the affine model $\boldsymbol{y}_j = \boldsymbol{A}_j \boldsymbol{x}_j$, which is of course equivalent to (3). Note that both the viewing and lighting directions are defined in local coordinates of the surface point j, and therefore the transformation $\boldsymbol{A}_j$ will necessarily change with pixel position.

The estimation goal is to recover each latent weight vector $\boldsymbol{x}_j$ for all j, from which we can compute the BRDF using (23). Here we make the reasonable assumption that at any given location, the number of unknown materials is limited to a small number, consistent with many real-world objects (in fact, it is quite common that only a single material may be present in many object regions). Moreover, given that the BRDF of each unknown base material can be closely approximated using (23) with a fixed weight vector for each material, it follows that the corresponding unknown weights $\boldsymbol{x}_j$ will each lie in a union of low-dimensional subspaces, conforming with the proposed subspace clustering model (Lawrence et al., 2006; Dong et al., 2010; Chen et al., 2014).

We test our algorithm as follows. Data acquisition is accomplished using physically-based path-tracing (Wenzel, 2010; Pharr & Humphreys, 2010), which accurately reproduces the physical capturing process. Importantly, this gains us access to the ground-truth such that quantitative comparisons are possible. We prepare two datasets to evaluate the performance of the proposed algorithm, one *checker* dataset, which consists of four different materials positioned in a checker-board pattern, and one *blend* dataset, that has four representative materials and each surface point represents a blending between two of the four possible materials. In both cases we mapped the materials onto a sphere with known geometry comprised of a total of $n = 104074$ points. The lighting was produced using the *Grace Cathedral* environment map (Debevec & Malik, 1997). Finally, we capture images of the object under 5 different view directions, resulting in 5 observations per visible surface point for a total of $5 \times n = 520370$ measurements. We compare our algorithm against a similar frame-

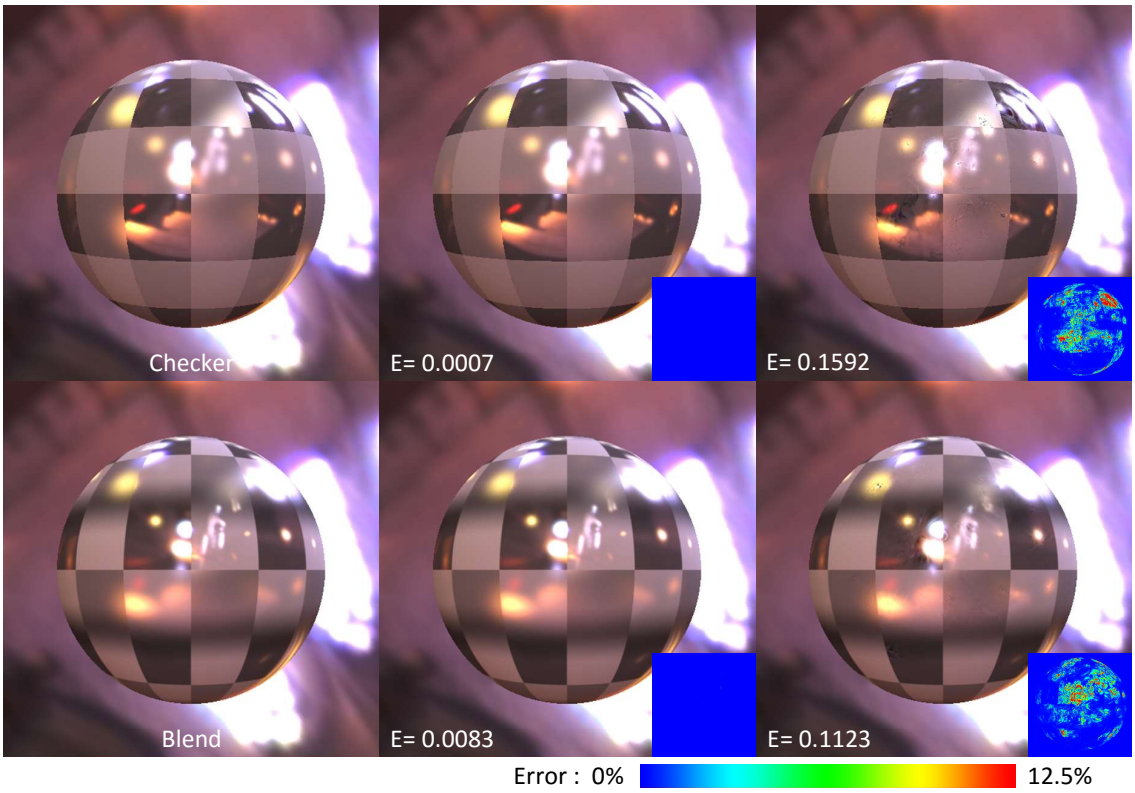

Figure 4: Spatially-varying BRDF reconstruction results. *Left column:* Ground truth reference under novel environmental lighting. *Middle column:* Rendering using our model. *Right column:* Rendering using the nuclear norm. Normalized MSE and difference maps are also included as an insert for both algorithms. Rendering errors best viewed by zooming.

work built upon the nuclear norm (Chen et al., 2014).

Figure 4 compares the renderings based on the reconstructed BRDFs under novel environmental lighting conditions (not those used to actually learn the BRDFs). We observe that with only 5 measurements per surface point, we can accurately reconstruct the BRDF without producing any visual artifacts. On the contrary, when using the nuclear norm Chen et al. (2014), the limited measurements cannot produce an accurate reconstruction and visual artifacts are clearly evident (zoom in for better viewing). The problem is compounded by the fact that the measurement matrices $\{\boldsymbol{A}_j\}$ are highly ill-conditioned as indicated by Figure 5, which displays the singular values of each $\boldsymbol{A}_j$ averaged across all j as compared to those from ideal matrices sampled iid from $\mathcal{N}(0,1)$. The nuclear norm is quite sensitive to this distinction which likely accounts, at least in part, for its poor performance. Note that accurate reconstruction from few measurements is a crucial ingredient of practical, inexpensive systems because it implies that fewer cameras are needed and/or a shorter acquisition time.

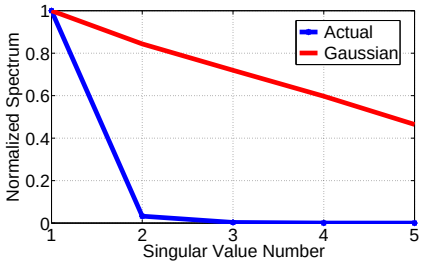

Figure 5: Singular values averaged across all $\boldsymbol{A}_j$ used in the BRDF estimation experiments. Ideal Gaussian data of equivalent dimensions is included for comparison. A fast singular value decay can be highly disruptive to nuclear-norm-based recovery algorithms.

7 Conclusions

In this paper we have introduced a practically-relevant, affine twist into the standard subspace clustering pipeline. We then derived a new, Bayesian-inspired algorithm that accounts for this added confound when necessary, while still defaulting to a principled state-of-the-art approach when deployed on existing segmentation problems with fully observable data, or when missing entries are present. Our framework, which does not require the typical spectral clustering post-processing step, is supported both by theoretical arguments and a large-scale, real-world application involving BRDF estimation and subsequent rendering.

References

Candès, E. and Recht, B. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, 9, 2008.

Candès, E., Mackey, L., and Soltanolkotabi, M. From robust subspace clustering to full-rank matrix completion. *Extended Abstract*, 2014.

Chen, G., Dong, Y., Peers, P., Zhang, J., and Tong, X. Reflectance scanning: Estimating shading frame and BRDF with generalized linear light sources. *ACM Trans. Graphics*, 33(4), 2014.

Debevec, P.E. and Malik, J. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH*, 1997.

Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

Dong, Y., Wang, J., Tong, X., Snyder, J., Lan, Y., Ben-Ezra, M., and Guo, B. Manifold bootstrapping for SVBRDF capture. *ACM Trans. Graphics*, 29(4), 2010.

Elhamifar, E. and Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(11), 2013.

Eriksson, B., Balzano, L., and Nowak, R. High-rank matrix completion. *International Conference on Artificial Intelligence and Statistics*, 2012.

Feng, J., Lin, Z., Xu, H., and Yan, S. Robust subspace segmentation with block-diagonal prior. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

Gruber, A. and Weiss, Y. Multibody factorization with uncertainty and missing data using the em algorithm. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.

Jalali, A., Chen, Y., Sanghavi, S., and Xu, H. Clustering partially observed graphs via convex optimization. *International Conference on Machine Learning*, 2011.

Lawrence, J., Ben-Artzi, A., DeCoro, C., Matusik, W., Pfister, H., Ramamoorthi, R., and Rusinkiewicz, S. Inverse shade trees for non-parametric material representation and editing. *ACM Trans. Graphics*, 25(3), 2006.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(1), 2013.

Lu, C.Y., Min, H., Zhao, Z.Q., Zhu, L., Huang, D.S., and Yan, S. Robust and efficient subspace segmentation via least squares regression. *European Conference on Computer Vision*, 2012.

Neal, R.M. and Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, 1999.

Patel, V.M., Nguyen, H.V., and Vidal, R. Latent space sparse subspace clustering. *International Conference on Computer Vision*, 2013.

Pharr, M. and Humphreys, G. *Physically Based Rendering: From Theory To Implementation*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2010.

Rao, S., Tron, R., Vidal, R., and Ma, Y. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(10), 2010.

Soltanolkotabi, M. and Candès, E. A geometric analysis of subspace clustering with outliers. *Annals of Statistics*, 40(4), 2012.

Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, 2001.

Tipping, M.E. and Bishop, C.M. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11 (2), 1999.

Wang, Y., Wang, Y.X., and Singh, A. A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data. *International Conference on Machine Learning*, 2015a.

Wang, Y., Wipf, D., Ling, Q., Chen, W., and Wassell, I. Multi-task learning for subspace segmentation. *International Conference on Machine Learning*, 2015b.

Wenzel, J. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org.

Wipf, D.P., Rao, B.D., and Nagarajan, S. Latent variable Bayesian models for promoting sparsity. *IEEE Trans. Information Theory*, 57(9), 2011.

Xin, B. and Wipf, D.P. Pushing the limits of affine rank minimization by adapting probabilistic PCA. *International Conference on Machine Learning*, 2015.

Yang, C., Robinson, D., and Vidal, R. Sparse subspace clustering with missing entries. *International Conference on Machine Learning*, 2015.

Zangwill, W.I. *Nonlinear Programming: A Unified Approach*. Prentice Hall, New Jersey, 1969.

Zhang, A., Fawaz, N., Ioannidis, S., and Montanari, A. Guess who rated this movie: Identifying users through subspace clustering. *International Conference on Uncertainty in Artificial Intelligence*, 2012.