Convex Sparse Coding, Subspace Learning, and Semi-Supervised Extensions

Xinhua Zhang **Yaoliang Yu** **Martha White** **Ruitong Huang** **Dale Schuurmans**
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada
{xinhua2,yaoliang,whitem,ruitong,dale}@cs.ualberta.ca

Abstract

Automated feature discovery is a fundamental problem in machine learning. Although classical feature discovery methods do not guarantee optimal solutions in general, it has been recently noted that certain subspace learning and sparse coding problems can be solved efficiently, provided the number of features is not restricted *a priori*. We provide an extended characterization of this optimality result and describe the nature of the solutions under an expanded set of practical contexts. In particular, we apply the framework to a semi-supervised learning problem, and demonstrate that feature discovery can co-occur with input reconstruction and supervised training while still admitting globally optimal solutions. A comparison to existing semi-supervised feature discovery methods shows improved generalization and efficiency.

Introduction

Data representations, and transformations of data representations, are fundamental to machine learning. Expressing complex data objects, such as documents or images, as feature vectors—e.g. as bags of words, vectors of Fourier or wavelet coefficients, or indicators of nearest prototypes—can reveal important structure in a data collection, as well as in individual data items. Feature representations do not only facilitate understanding, they enable subsequent learning. (Kernel methods achieve similar goals by expressing data implicitly in an abstract feature space.) For any particular application, however, often one does not know which representation to use.

Automatically *discovering* useful features from data has been a long standing goal of machine learning research. Current feature discovery methods have already proved useful in many areas of data analysis, including text, image, and biological data processing. These methods differ primarily in the properties sought in any new data representation. Some approaches seek a low dimensional representation, such as principal components analysis (PCA) and modern variants (van der Maaten and Hinton 2008; Weinberger and Saul 2006). Others seek a representation where features behave independently, such as independent components analysis (ICA) (Comon 1994); or where the new feature vectors are sparse, such as sparse coding or vector quantization (Olshausen and Field 1997). Still others seek a representation that captures higher level, abstract features of the data that are invariant to low level transformations, such as in deep learning (Hinton 2007).

In each case, one key issue is whether an optimal feature representation can be recovered efficiently. The lack of an optimal feature discovery method can hamper the practical applicability of a principle—relegating its use to an artform. Globally solvable criteria such as PCA, by contrast, enjoy widespread use despite the numerous shortcomings, arguably because users need not understand the workings of any solver—it is sufficient to understand the principle being optimized. Recently, the development of sparsity inducing regularizers has made great inroads in achieving globally solvable forms of training. Indeed, convex reformulations have recently had a significant impact on many areas of machine learning, including multitask learning (Argyriou, Evgeniou, and Pontil 2008), collaborative filtering (Candes and Recht 2008; Srebro, Rennie, and Jaakkola 2004), and nonlinear dimensionality reduction (Weinberger and Saul 2006).

In this paper we contribute further progress to achieving tractable formulations of representation learning problems. First, we present an explicit formulation of convex feature discovery that encapsulates classical subspace learning and sparse coding as special cases. The main idea is to replace a bound on the number of features with a convex, sparsity inducing regularization scheme. We demonstrate how the standard regularizer used in sparse coding leads to a trivial form of vector quantization. However, other regularizers lead to elegant forms of subspace learning that encompass PCA and extensions within a simple unified framework. Very similar observations have already been made by (Bach, Mairal, and Ponce 2008). However, we provide further generalizations regarding the derived regularizers and recovered solutions, and present a more explicit framework that enables easier extension to other learning scenarios. Our second and more significant contribution is to develop a new convex formulation of semi-supervised representation learning within the general framework we develop. This formulation encompasses several previous proposals that deployed local training procedures (Mairal et al. 2008; Lee et al. 2009; Raina et al. 2007; Rish et al. 2007), but allows us to reexpress the problem in a joint, globally solvable form. Unlike (Goldberg et al. 2010), we recover an explicit feature

representation of the data that respects basis constraints, and can extend the formulation beyond transduction.

Preliminaries: Vector and Matrix Norms

In our technical development below we will need to make use of several vector and matrix norms, and their associated properties, so we centralize the definitions here.

For vectors, we use $\|\mathbf{x}\|$ to denote a norm on $\mathbf{x}$, and $\|\mathbf{y}\|^*$ to refer to its conjugate norm: $\|\mathbf{y}\|^* = \max_{\|\mathbf{x}\|\leq 1} \mathbf{x}'\mathbf{y}$. One can verify that $\|\mathbf{x}\|^{**}=\|\mathbf{x}\|$ (Rockafellar 1970, §15). We use $\|\mathbf{x}\|_p$ to denote a p-norm, $1\leq p\leq\infty$, whose conjugate norm is $\|\cdot\|_{p^*}$ such that $\frac{1}{p}+\frac{1}{p^*}=1$. Norms are always convex.

For matrices, we use $\|X\|$ to refer to a generic norm on X, and $\|Y\|^*$ to denote its conjugate norm. The conjugate satisfies $\|Y\|^* = \max_{\|X\|\leq 1} \text{tr}(X'Y)$ and $\|X\|^{**} = \|X\|$, where tr denotes trace. We will use $\|X\|_{(p,q)}$ to refer to the *induced norm* on X defined by $\|X\|_{(p,q)} = \max_{\|\mathbf{z}\|_p\leq 1} \|X\mathbf{z}\|_q$ (Horn and Johnson 1985, §5.6). The standard *spectral norm* will be denoted $\|X\|_{sp} = \|X\|_{(2,2)} = \sigma_{max}(X)$. The conjugate of the spectral norm is the *trace norm* $\|X\|_{tr} = \sum_i \sigma_i(X)$. The *Frobenius norm* is given by $\|X\|_F = \sqrt{\text{tr}(X'X)} = \sqrt{\sum_i \sigma_i^2(X)}$. We will also make use of the so called *block norm* $\|X\|_{r,s} = (\sum_i(\sum_j |X_{ij}|^r)^{\frac{s}{r}})^{\frac{1}{s}}$, whose conjugate is $\|X\|_{r,s}^* = \|X\|_{r^*s^*}$ such that $\frac{1}{r}+\frac{1}{r^*}=\frac{1}{s}+\frac{1}{s^*}=1$ (Bradley and Bagnell 2009). Finally, we require a preliminary fact.

Lemma 1 *For any bounded closed set $\mathcal{Z}\subset\mathbb{R}^n$ such that span$(\mathcal{Z})=\mathbb{R}^n$, and any $1\leq p\leq\infty$, the definition $\|X\|_{(\mathcal{Z},p)} = \max_{\mathbf{z}\in\mathcal{Z}} \|X\mathbf{z}\|_p$ establishes a norm on X.*

Proof: It is easy to verify $\|X\|_{(\mathcal{Z},p)}\geq 0$ and $\|\alpha X\|_{(\mathcal{Z},p)} = |\alpha|\|X\|_{(\mathcal{Z},p)}$. Next, note $\|X+Y\|_{(\mathcal{Z},p)} = \max_{\mathbf{z}\in\mathcal{Z}} \|(X+Y)\mathbf{z}\|_p \leq \max_{\mathbf{z}\in\mathcal{Z}} \|X\mathbf{z}\|_p + \|Y\mathbf{z}\|_p \leq \|X\|_{(\mathcal{Z},p)} + \|Y\|_{(\mathcal{Z},p)}$. Finally, if $\mathcal{Z}$ is not restricted to a subspace of $\mathbb{R}^n$ then $\max_{\mathbf{z}\in\mathcal{Z}} \|X\mathbf{z}\|_p = 0$ implies $X=0$. ■

Below we will need to use the three distinct types of matrix norms, $\|X\|_{p,q}$, $\|X\|_{(p,q)}$, and $\|X\|_{(\mathcal{P},q)}$, respectively.

Unsupervised Representation Learning

First consider the problem of unsupervised feature discovery, where one is given an $n\times m$ matrix of data X such that each column is an n-dimensional observation and there are m observations. The goal is to learn an $n\times k$ dictionary B containing k basis vectors, and a $k\times m$ representation matrix Φ containing m new feature vectors of length k, so that X can be accurately reconstructed from $\hat{X}=B\Phi$. To measure approximation error we use a loss function $L(\hat{X};X)$ that is convex in its first argument. Conventional choices for L include sum of squared error $L(\hat{X};X)=\|\hat{X}-X\|_F^2$ or a sum of Bregman divergences $L(\hat{X};X)=\sum_j D(\hat{X}_{:j}\|X_{:j})$, but it is not necessary to restrict attention to any particular convex loss. Note that the factorization $\hat{X}=B\Phi$ is invariant to reciprocal rescalings of B and Φ, so to avoid degeneracy their individual magnitudes have to be controlled. We will assume that each column $B_{:j}$ of B is constrained to belong to a bounded closed convex set $\mathcal{B}$, hence $B\in\mathcal{B}^k$.

Subspace learning methods, such as PCA and variants seek a reconstruction matrix $\hat{X}=B\Phi$ that has reduced rank. *Sparse coding* methods, on the other hand, seek a reconstruction where each new feature vector $\Phi_{:j}$ is sparse; that is, $\hat{X_{:j}}$ is reconstructed from a small subset of basis vectors chosen from the dictionary B (Olshausen and Field 1997). For both, the generic training problem can be expressed

$$\min_{B\in\mathcal{B}^k} \min_{\Phi} L(B\Phi;X) + \alpha\|\Phi\|, \tag{1}$$

where L is a loss, $\alpha\geq 0$ is a parameter, $\|\cdot\|$ is a norm on the representation matrix Φ, and k is the number of features to be extracted. Specific choices of L, α, $\|\cdot\|$, and $\mathcal{B}$ yield standard forms of subspace learning and sparse coding. For example, $L(\hat{X};X)=\|\hat{X}-X\|_F^2$, $\alpha=0$, and $\mathcal{B}=\{\mathbf{b}:\|\mathbf{b}\|_2\leq 1\}$ yields PCA. Setting L to a Bregman divergence, $\alpha=0$, and $\mathcal{B}$ as above, yields exponential family PCA (Collins, Dasgupta, and Schapire 2001; Gordon 2002).[1] Unfortunately, (1) does not readily admit global training. Certainly, the problem is convex in B given Φ and vice versa, but it is not jointly convex. Beyond PCA, global training procedures are not generally known, and most research resorts to alternating minimization (Jenatton et al. 2010; Bradley and Bagnell 2008; Elad and Aharon 2006; Zou, Hastie, and Tibshirani 2006).

However, it has recently been observed that if the number of features is not bounded, and a sparse regularizer $\|\Phi\|$ is used to indirectly control their number, then B can be considered large but fixed and (1) becomes convex in Φ, making it amenable to a boosting approach that generates columns in B (Bradley and Bagnell 2009; Nowozin and Bakir 2008). More importantly, it has been realized that (1) can be solved directly in certain cases (Bach, Mairal, and Ponce 2008). We now elucidate this finding further and develop a general framework for solving the relaxed problem

$$\min_{B\in\mathcal{B}^\infty} \min_{\Phi} L(B\Phi;X) + \alpha\|\Phi\|. \tag{2}$$

Notation: We use $\min_{B\in\mathcal{B}^\infty}$ as a shorthand for $\min_{k\in\mathbb{N}} \min_{B\in\mathcal{B}^k}$.

Subspace Learning For subspace learning (i.e. dimensionality reduction), rather than bounding the number of columns in B, we allow it to grow as necessary and drop features implicitly by imposing a $\|\Phi\|_{2,1}$ regularizer. Such a regularizer will encourage entire rows $\Phi_{i:}$ (features) to become sparse (Argyriou, Evgeniou, and Pontil 2008) but otherwise only smooth the columns. To avoid degeneracy, we set $\mathcal{B}_2=\{\mathbf{b}:\|\mathbf{b}\|_2\leq 1\}$. Then, with these assumptions, the training problem (2) can be solved globally and efficiently.

Proposition 1 (Convex subspace learning)

$$\min_{B\in\mathcal{B}_2^\infty} \min_{\Phi} L(B\Phi;X) + \alpha\|\Phi\|_{2,1} \tag{3}$$

$$= \min_{\hat{X}} L(\hat{X};X) + \alpha\|\hat{X}\|_{tr}. \tag{4}$$

Given $\hat{X}$, a solution to (3) can be recovered by setting $B=U$ and $\Phi=\Sigma V'$, where $\hat{X}=U\Sigma V'$ is the SVD of $\hat{X}$.

[1] Note that (1) can easily accommodate *missing* entries in X by restricting the loss evaluation to observed entries (Srebro, Rennie, and Jaakkola 2004). This extension trivially available to every formulation discussed in this paper, so we do not emphasize it further.

Proof: (4) follows from Theorem 1 and Lemma 2 (22) below. Given $\hat{X}$, B and Φ must be optimal since these satisfy $B\Phi = \hat{X}$ and $\|\Phi\|_{2,1} = \text{tr}(\Sigma) = \|\hat{X}\|_{tr}$ respectively. ■

Therefore (3) can be solved globally by solving the convex problem (4), then recovering B and Φ from $\hat{X}$. The solution satisfies $\text{rank}(B) = \text{rank}(\Phi) = \text{rank}(\hat{X})$; thus, even though we allowed $B \in \mathcal{B}_2^\infty$, by reducing the rank of $\hat{X}$ via trace norm regularization one implicitly and efficiently controls the dimension of the result.[2] Interestingly, Proposition 1 can also be extended to remove sparse noise in X, generalizing the robust subspace learning formulations of (Candes et al. 2009; Xu, Caramanis, and Sanghavi 2010).

Corollary 1 (Convex robust subspace learning)

$$\min_{B\in\mathcal{B}_2^\infty} \min_{\Phi} \min_{S} L(B\Phi+S;X)+\alpha\|\Phi\|_{2,1}+\beta\|S\|_{1,1} \quad (5)$$

$$= \min_{\hat{X}} \min_{S} L(\hat{X}+S;X)+\alpha\|\hat{X}\|_{tr}+\beta\|S\|_{1,1}. \quad (6)$$

Sparse Coding For sparse coding, the goal is not to reduce dimension but instead to learn a sparse representation Φ. The standard regularizer used for this purpose has been $\|\Phi\|_{1,1}$, which encourages entry-wise sparsity in Φ (Mairal et al. 2008; Lee et al. 2009; Jenatton et al. 2010). To avoid degeneracy, one imposes the constraint $B_{:j} \in \mathcal{B}_q = \{\mathbf{b} : \|\mathbf{b}\|_q \leq 1\}$ for some $1 \leq q \leq \infty$. As above, the resulting training problem can be solved globally and efficiently.

Proposition 2 (Convex sparse coding)

$$\min_{B\in\mathcal{B}_q^\infty} \min_{\Phi} L(B\Phi;X) + \alpha\|\Phi\|_{1,1} \quad (7)$$

$$= \min_{\hat{X}} L(\hat{X};X) + \alpha\|\hat{X}'\|_{q,1}. \quad (8)$$

Given $\hat{X}$, setting $B = [\hat{X}_{:1}/\|\hat{X}_{:1}\|_q, ..., \hat{X}_{:m}/\|\hat{X}_{:m}\|_q]$ and $\Phi = diag([\|\hat{X}_{:1}\|_q, ..., \|\hat{X}_{:m}\|_q])$ provides a solution to (7).

Proof: (8) follows from Theorem 1 and Lemma 2 (23) below. Given $\hat{X}$, B and Φ must be optimal since these satisfy $B\Phi = \hat{X}$ and $\|\Phi\|_{1,1} = \|\hat{X}'\|_{q,1}$ respectively. ■

Therefore (7) can be solved efficiently by first solving (8), then recovering B and Φ as shown. Note that, contrary to common intuition, the solution is not over-complete. That is, we obtain a simple form of vector quantization that memorizes the (normalized) observations and codes the training data by a scaled indicator vector; an outcome also witnessed by (Bach, Mairal, and Ponce 2008). This property is an inherent weakness of $\|\cdot\|_{1,1}$ regularization that does not appear to be widely appreciated. Nevertheless, given a test point $\mathbf{x}$, one can recover $\boldsymbol{\phi} = \arg\min_{\boldsymbol{\phi}} \ell(B\boldsymbol{\phi}) + \alpha\|\boldsymbol{\phi}\|_1$, yielding a sparse representation in terms of the training observations.

General Formulation We now prove a general result that yields the previous propositions as special cases. This formulation is more general than (Bach, Mairal, and Ponce 2008), by allowing $B_{:j} \in \mathcal{B}$ for any bounded closed $\mathcal{B}$, more restrictive by considering $\|\Phi\|_{p,1}$ regularization, and delivers a more explicit characterization of the induced norm on $\hat{X}$ that we exploit in our semi-supervised extensions below.

Theorem 1 *For any $1 \leq p \leq \infty$, and any bounded closed set $\mathcal{B} \subset \mathbb{R}^n$ such that* $\text{span}(\mathcal{B}) = \mathbb{R}^n$

$$\min_{B\in\mathcal{B}^\infty} \min_{\Phi} L(B\Phi;X) + \alpha\|\Phi\|_{p,1} \quad (9)$$

$$= \min_{\hat{X}} L(\hat{X};X) + \alpha\|\hat{X}'\|^*_{(\mathcal{B},p^*)} \quad (10)$$

using the induced norm definition from Lemma 1.

Proof:

$$(9) = \min_{\hat{X}} \min_{B\in\mathcal{B}^\infty} \min_{\Phi:B\Phi=\hat{X}} L(\hat{X};X) + \alpha\|\Phi\|_{p,1} \quad (11)$$

$$= \min_{\hat{X}} L(\hat{X};X) + \alpha \min_{B\in\mathcal{B}^\infty} \min_{\Phi:B\Phi=\hat{X}} \|\Phi\|_{p,1}. \quad (12)$$

Now consider the inner minimization in (12). Fix any $\hat{X}$, $k \in \mathbb{N}$ and $B \in \mathcal{B}^k$, and observe

$$\min_{\Phi:B\Phi=\hat{X}} \|\Phi\|_{p,1} = \min_{\Phi} \max_{\Lambda} \|\Phi\|_{p,1} + \text{tr}(\Lambda'(\hat{X}-B\Phi)). \quad (13)$$

If B does not span the columns of $\hat{X}$ then the constraint $B\Phi = \hat{X}$ is infeasible, and (13) is unbounded above; hence such a B cannot participate in a minimum of (12). We conclude that any B selected in (12) must span the columns of $\hat{X}$. Given such a B, a feasible Φ exists, meaning Slater's condition is satisfied and strong Lagrange duality holds (Boyd and Vandenberghe 2004, §5.2.3). Thus for such a B

$$(13) = \max_{\Lambda} \min_{\Phi} \|\Phi\|_{p,1} + \text{tr}(\Lambda'(\hat{X}-B\Phi)). \quad (14)$$

Since the dual norm of $\|\cdot\|_{p,1}$ is $\|\cdot\|_{p^*,\infty}$, by norm duality:

$$(14) = \max_{\Lambda} \min_{\Phi} \max_{\|V\|_{p^*,\infty}\leq 1} \text{tr}(V'\Phi) + \text{tr}(\Lambda'(\hat{X}-B\Phi)) \quad (15)$$

$$= \max_{\Lambda} \max_{\|V\|_{p^*,\infty}\leq 1} \min_{\Phi} \text{tr}(\Lambda'\hat{X}) + \text{tr}(\Phi'(V-B'\Lambda)) \quad (16)$$

$$= \max_{\|V\|_{p^*,\infty}\leq 1} \max_{\Lambda:B'\Lambda=V} \text{tr}(\Lambda'\hat{X}) \quad (17)$$

$$= \max_{\Lambda:\|B'\Lambda\|_{p^*,\infty}\leq 1} \text{tr}(\Lambda'\hat{X}), \quad (18)$$

where (16) follows by (Rockafellar 1970, Cor. 37.3.2), (17) follows by eliminating Φ, and (18) follows by a straightforward substitution. Therefore, we have established

$$(12) = \min_{\hat{X}} L(\hat{X};X) + \alpha \min_{B\in\mathcal{B}^\infty} \max_{\Lambda:\|B'\Lambda\|_{p^*,\infty}\leq 1} \text{tr}(\Lambda'\hat{X}) \quad (19)$$

$$= \min_{\hat{X}} L(\hat{X};X) + \alpha \max_{\Lambda:\|\Lambda'\|_{(\mathcal{B},p^*)}\leq 1} \text{tr}(\Lambda'\hat{X}) \quad (20)$$

$$= \min_{\hat{X}} L(\hat{X};X) + \alpha\|\hat{X}'\|^*_{(\mathcal{B},p^*)}, \quad (21)$$

where (20) follows by the definition of $\|\cdot\|_{(\mathcal{B},p^*)}$ in Lemma 1, and (21) follows again by norm duality. ■

Lemma 2 *Let $\mathcal{B}_q = \{\mathbf{b} : \|\mathbf{b}\|_q \leq 1\}$. In the mapping established by Theorem 1, the induced regularizer $\|\hat{X}\|^*_{(\mathcal{B}_q,p^*)}$ has a simple closed form in the following special cases*

$$\|\Phi\|_{2,1}, \mathcal{B}_2 \mapsto \|\hat{X}'\|^*_{(\mathcal{B}_2,2)} = \|\hat{X}\|_{tr} \quad (22)$$

$$\|\Phi\|_{1,1}, \mathcal{B}_q \mapsto \|\hat{X}'\|^*_{(\mathcal{B}_q,\infty)} = \|\hat{X}'\|_{q,1} \quad (23)$$

$$\|\Phi\|_{p,1}, \mathcal{B}_1 \mapsto \|\hat{X}'\|^*_{(\mathcal{B}_1,p^*)} = \|\hat{X}\|_{p,1}. \quad (24)$$

(The first two cases correspond to propositions 1 and 2.)

[2] $\|X\|_{tr}$ is the convex envelope of $\text{rank}(X)$ over the set $\{X : \|X\|_{sp} \leq 1\}$, see e.g. (Recht, Fazel, and Parrilo 2007). It provides a convex relaxation of rank widely used in low rank matrix recovery (Candes and Recht 2008; Salakhutdinov and Srebro 2010).

Proof: Note $\|\hat{X}'\|^*_{(\mathcal{B}_2,2)} = \|\hat{X}'\|^*_{(2,2)} = \|\hat{X}'\|^*_{sp} = \|\hat{X}'\|_{tr}$, proving (22). Then by (Steinberg 2005, §1.3.1) and Hölder's inequality one can show $\|Y\|^*_{(1,r)} = \max_{\|B\|_{(1,r)}\leq 1} \mathrm{tr}(B'Y) = \max_{B:\|B_{:j}\|_r\leq 1 \forall j} \sum_j B'_{:j}Y_{:j} = \sum_j \|Y_{:j}\|_{r^*} = \|Y'\|_{r^*,1}$; so for (24) one obtains $\|\hat{X}'\|^*_{(\mathcal{B}_1,p^*)} = \|\hat{X}'\|^*_{(1,p^*)} = \|\hat{X}\|_{p,1}$. Finally, $\|Y'\|_{(q^*,p^*)} = \|Y\|_{(p,q)}$ (Steinberg 2005, §1.2.2), so for (23): $\|\hat{X}'\|^*_{(\mathcal{B}_q,\infty)} = \|\hat{X}'\|^*_{(q,\infty)} = \|\hat{X}\|^*_{(1,q^*)} = \|\hat{X}'\|_{q,1}$. ■

Theorem 1 captures a wide range of formulations, including for example standard sparse coding and sparse PCA (Bradley and Bagnell 2009; Bach, Mairal, and Ponce 2008; Zou, Hastie, and Tibshirani 2006). However, for (10) to admit an efficient global optimization procedure, the derived norm $\|\hat{X}'\|^*_{(\mathcal{B},p^*)}$ must be efficiently computable for given $\hat{X}$. We have already seen that this is achievable in Propositions 1 and 2, where the derived norm reduced to standard, efficiently computable norms on $\hat{X}$. Unfortunately the induced norm $\|\hat{X}'\|^*_{(\mathcal{B},p^*)}$, although convex, is not always efficiently computable (Hendrickx and Olshevsky 2010; Steinberg 2005). In particular, this norm is not known to be efficiently computable for the mixed regularizers considered in (Bach, Mairal, and Ponce 2008; Bradley and Bagnell 2009; Zou, Hastie, and Tibshirani 2006); hence these previous works had to introduce relaxations, heuristic basis generators, or alternating minimization, respectively. Nevertheless, we will see that there remain many important and useful cases where $\|\hat{X}'\|^*_{(\mathcal{B},p^*)}$ can be computed efficiently.

Semi-supervised Representation Learning

Our main contribution in this paper is to demonstrate a global form of semi-supervised representation learning that can be achieved by applying the general framework.

Consider a setting where we are given an $n \times m_u$ matrix of unlabeled data X_u, an $n \times m_l$ matrix of labeled data X_l, and a $c \times m_l$ matrix of target values Y_l. We would like to learn a $k \times (m_l + m_u)$ representation matrix $\Phi = [\Phi_l, \Phi_u]$ and an $n \times k$ basis dictionary B such that $X = [X_l, X_u]$ can be reconstructed from $\hat{X} = B\Phi$, while simultaneously learning a $c \times k$ prediction model W such that Y_l can be reconstructed from $\hat{Y}_l = W\Phi_l$. Let $L_u(B\Phi; X)$ and $L_s(W\Phi_l; Y_l)$ denote unsupervised and supervised losses respectively, which we assume are convex in their first argument. To avoid degeneracy we impose the constraints $B_{:j} \in \mathcal{B}$ and $W_{:j} \in \mathcal{W}$ for bounded closed convex sets $\mathcal{B}$ and $\mathcal{W}$. The joint training problem can then be expressed as a convex program.

Proposition 3

$$\min_{B\in\mathcal{B}^\infty} \min_{W\in\mathcal{W}^\infty} \min_{\Phi_l,\Phi_u} L_u(B[\Phi_l, \Phi_u]; X) + \beta L_s(W\Phi_l; Y_l) + \alpha\|[\Phi_l, \Phi_u]\|_{p,1} \quad (25)$$

$$= \min_{U\in\mathcal{U}^\infty} \min_{\Phi} L_c(U\Phi; Z) + \alpha\|\Phi\|_{p,1} \quad (26)$$

$$= \min_{\hat{Z}} L_c(\hat{Z}; Z) + \alpha\|\hat{Z}'\|^*_{(\mathcal{U},p^*)}, \quad (27)$$

where $Z = \begin{bmatrix} X_l & X_u \\ Y_l & 0 \end{bmatrix}$, $U = \begin{bmatrix} B \\ W \end{bmatrix}$, $\hat{Z} = U[\Phi_l, \Phi_u]$, $\mathcal{U} = \mathcal{B}\times\mathcal{W}$, *and* $L_c(\hat{Z}; Z) = L_u(B\Phi; X) + \beta L_s(W\Phi_l; Y_l)$.

Proof: (27) follows immediately from Theorem 1. ■

Unlike staged training procedures that separate the unsupervised from the supervised phase (Lee et al. 2009), and previous work on semi-supervised dimensionality reduction that relies on alternating minimization (Rish et al. 2007; Pereira and Gordon 2006), Proposition 3 provides a jointly convex formulation that allows all components to be trained simultaneously. Whether (27) can be solved efficiently and B, W, Φ recovered from $\hat{Z}$ depends on the structure of the derived norm $\|\hat{Z}'\|^*_{(\mathcal{U},p^*)}$. We show two significant cases where this can be achieved.

Sparse Coding Formulation If in (25) we choose $p = 1$ and constrain the basis dictionary and prediction model to $\mathcal{B}_{q_1} = \{\mathbf{b} : \|\mathbf{b}\|_{q_1} \leq 1\}$ and $\mathcal{W}_{q_2} = \{\mathbf{w} : \|\mathbf{w}\|_{q_2} \leq \gamma\}$ respectively, then an efficient characterization of the induced norm in (27) can be obtained. Thus, here we are considering $\|\Phi\|_{1,1}$ regularization (hence $p^* = \infty$). Let Λ denote a dual matrix the same size as $\hat{Z}$, where Λ^X denotes the upper and Λ^Y the lower parts respectively, and let $\mathcal{U}^{q_1}_{q_2} = \mathcal{B}_{q_1} \times \mathcal{W}_{q_2}$. The dual of the induced norm can be easily derived first.

Lemma 3 $\|\Lambda'\|_{(\mathcal{U}^{q_1}_{q_2},\infty)} = \max_j \|\Lambda^X_{:j}\|_{q_1^*} + \gamma\|\Lambda^Y_{:j}\|_{q_2^*}$.

Proof:

$$\|\Lambda'\|_{(\mathcal{U}^{q_1}_{q_2},\infty)} = \max_{\mathbf{u}\in\mathcal{U}^{q_1}_{q_2}} \|\Lambda'\mathbf{u}\|_\infty = \max_{\mathbf{u}\in\mathcal{U}^{q_1}_{q_2}} \max_j \Lambda'_{:j}\mathbf{u} \quad (28)$$

$$= \max_j \max_{\|\mathbf{b}\|_{q_1}\leq 1} \max_{\|\mathbf{w}\|_{q_2}\leq\gamma} \mathbf{b}'\Lambda^X_{:j} + \mathbf{w}'\Lambda^Y_{:j}. \quad (29)$$

The lemma then follows by norm duality. ■

The induced norm on $\hat{Z}$ is then easy to determine.

Corollary 2 $\|\hat{Z}'\|^*_{(\mathcal{U}^{q_1}_{q_2},\infty)} = \sum_j \max(\|\hat{Z}^X_{:j}\|_{q_1}, \frac{1}{\gamma}\|\hat{Z}^Y_{:j}\|_{q_2})$.

Therefore both the induced norm and its dual are efficiently computable in this case, resulting in an efficient "sparse coding" formulation of semi-supervised representation learning. The training problem can be solved globally by first solving the convex optimization (27), then given $\hat{Z}$, recovering B, W and Φ by setting $U = \hat{Z}D^{-1}$ and $\Phi = D$, where D is a diagonal matrix such that $D_{jj} = \max(\|\hat{Z}^X_{:j}\|_{q_1}, \frac{1}{\gamma}\|\hat{Z}^Y_{:j}\|_{q_2})$. This solution must be an optimum since $U\Phi = \hat{Z}$ and $\|\Phi\|_{1,1} = \|\hat{Z}'\|^*_{(\mathcal{U}^{q_1}_{q_2},\infty)}$. Unfortunately, as in the unsupervised case, we reach the conclusion that $\|\Phi\|_{1,1}$ regularization leads to a trivial form of vector quantization, unless the number of features is explicitly restricted (but imposing this restriction leads to intractability).

Subspace Learning Formulation Fortunately, the situation for subspace learning is more interesting. If instead in (25) we choose $p = 2$ and constrain the basis dictionary and prediction model to $\mathcal{B}_2 = \{\mathbf{b} : \|\mathbf{b}\|_2 \leq 1\}$ and $\mathcal{W}_2 = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq \gamma\}$ respectively, then we achieve an efficient characterization of the derived norm, and thus recover a novel and effective convex formulation of semi-supervised dimensionality reduction.

To derive a concrete characterization of the induced norm we need to introduce the following definitions. Let $\mathcal{U}^2_2 = \mathcal{B}_2 \times \mathcal{W}_2$, and let I_n denote an $n \times n$ identity matrix. Define two diagonal indicator matrices $I^X = \mathrm{diag}([\mathbf{1}_n; \mathbf{0}_c])$

and $I^Y = \text{diag}([\mathbf{0}_n, \mathbf{1}_c])$, such that $I^X + I^Y = I_{n+c}$. Also define the parameterized diagonal matrix $D_\rho = \text{diag}([\sqrt{1+\gamma^2\rho}\,\mathbf{1}_n; \sqrt{\gamma^2+1/\rho}\,\mathbf{1}_c])$ for $\rho \geq 0$. As above, it will be easier to first derive the dual norm, before using duality again to recover the target norm on $\hat{Z}$.

Lemma 4 $\|\Lambda'\|_{(\mathcal{U}_2^2,2)} = \min_{\rho\geq 0} \|D_\rho \Lambda\|_{sp}$.

Proof:
$$\|\Lambda'\|^2_{(\mathcal{U}_2^2,2)} = \max_{\mathbf{h}:\|\mathbf{h}^X\|_2=1,\,\|\mathbf{h}^Y\|_2=\gamma} \mathbf{h}'\Lambda\Lambda'\mathbf{h} \quad (30)$$
$$= \max_{H:H\succeq 0,\,\text{tr}(HI^X)=1,\,\text{tr}(HI^Y)=\gamma^2} \text{tr}(H\Lambda\Lambda') \quad (31)$$
$$= \min_{\lambda\geq 0,\nu\geq 0:\Lambda\Lambda'\preceq \lambda I^X+\nu I^Y} \lambda + \gamma^2\nu \quad (32)$$
$$= \min_{\lambda\geq 0,\nu\geq 0:\|D_{\nu/\lambda}\Lambda\|^2_{sp}\leq \lambda+\gamma^2\nu} \lambda + \gamma^2\nu \quad (33)$$
$$= \min_{\lambda\geq 0,\nu\geq 0} \|D_{\nu/\lambda}\Lambda\|^2_{sp} = \min_{\rho\geq 0} \|D_\rho\Lambda\|^2_{sp}, \quad (34)$$

where (31) follows by the substitution $H = \mathbf{h}\mathbf{h}'$ and (32) is its Lagrange dual.[3] To explain (33) note that for $\lambda \geq 0$ and $\nu \geq 0$, the relation $\Lambda\Lambda' \preceq \lambda I^X + \nu I^Y$ holds if and only if $D_{\frac{\nu}{\lambda}}\Lambda\Lambda' D_{\frac{\nu}{\lambda}} \preceq D_{\frac{\nu}{\lambda}}(\lambda I^X + \nu I^Y) D_{\frac{\nu}{\lambda}} = (\lambda+\gamma^2\nu) I_{n+c}$. ■

Not only does the dual norm have a simple characterization, it is efficiently computable: it can be evaluated by a simple power method iteration that renormalizes each part $\mathbf{h}^X$ and $\mathbf{h}^Y$ of $\mathbf{h}$ independently.[4] Given this dual formulation, the target norm can then be easily characterized.

Lemma 5 $\|\hat{Z}'\|^*_{(\mathcal{U}_2^2,2)} = \max_{\rho\geq 0} \|D_\rho^{-1}\hat{Z}\|_{tr}$.

Proof:
$$\|\hat{Z}'\|^*_{(\mathcal{U}_2^2,2)} = \max_{\|\Lambda'\|_{(\mathcal{U}_2^2,2)}\leq 1} \text{tr}(\Lambda'\hat{Z}) \quad (35)$$
$$= \max_{\rho\geq 0} \max_{\Lambda:\|D_\rho\Lambda\|_{sp}\leq 1} \text{tr}(\Lambda'\hat{Z}) \quad (36)$$
$$= \max_{\rho\geq 0} \max_{\tilde{\Lambda}:\|\tilde{\Lambda}\|_{sp}\leq 1} \text{tr}(\tilde{\Lambda}' D_\rho^{-1}\hat{Z}) = \max_{\rho\geq 0} \|D_\rho^{-1}\hat{Z}\|_{tr}, \quad (37)$$

using the definitions of the dual norms (norm duality). ■

So using this induced norm, the objective (27) can be optimized to recover $\hat{Z}$. Although $\|\hat{Z}\|^*_{(\mathcal{U}_2^2,2)}$ can be computed by a line search over $\rho \geq 0$,[5] it is far more efficient to work with the dual norm given in Lemma 4.

Computational Method In our experiments below we solve the learning problem (27) for the subspace case by working with a more efficient dual formulation (Rockafellar 1970, Theorem 31.1 and 31.3). In particular, we first recover the matrix $\hat{\Lambda}$ by solving the dual problem

$$\min_{\Lambda} L_c^\star(\Lambda; Z) + \alpha^\star \|\Lambda'\|_{(\mathcal{U}_2^2,2)} \quad (38)$$

where $L_c^\star(\Lambda; Z)$ is the Fenchel conjugate of $L_c(\hat{Z}; Z)$, and $\alpha^\star$ is a dual regularization parameter.[6] Then, with $\hat{\Lambda}$ available, we recover $\hat{Z}^X$ and $\hat{Z}_l^Y$ by solving

$$\min_{\hat{Z}^X,\hat{Z}_l^Y} L_u(\hat{Z}^X; X) + \beta L_s(\hat{Z}_l^Y; Y_l) - \text{tr}(\hat{Z}^{X\prime}\hat{\Lambda}^X) - \text{tr}(\hat{Z}_l^{Y\prime}\hat{\Lambda}_l^Y).$$

Since $\hat{Z}_u^Y$ does not affect $L_c(\hat{Z}; Z)$ in (27), it can be recovered by minimizing $\|\hat{Z}'\|^*_{(\mathcal{U}_2^2,2)}$ keeping $\hat{Z}^X$ and $\hat{Z}_l^Y$ fixed.

Finally, given an optimal solution $\hat{Z}$, we recover U and Φ by a simple cut algorithm that greedily generates columns for U: Recall from the proof of Theorem 1 that if U and Φ are optimal they must satisfy

$$\|\hat{Z}'\|^*_{(\mathcal{U}_2^2,p^*)} = \min_{U\in(\mathcal{U}_2^2)^\infty} \min_{\Phi:U\Phi=\hat{Z}} \|\Phi\|_{2,1} \quad (39)$$
$$= \min_{U\in(\mathcal{U}_2^2)^\infty} \min_{\Phi} \max_{\Lambda} \|\Phi\|_{2,1} + \text{tr}(\Lambda'(\hat{Z} - U\Phi)) \quad (40)$$
$$= \max_{\Lambda:\|\mathbf{u}'\Lambda\|_2\leq 1\ \forall \mathbf{u}\in\mathcal{U}_2^2} \text{tr}(\Lambda'\hat{Z}). \quad (41)$$

Thus, given a current U, a minimum cost Φ and corresponding Lagrange multiplier Λ can be recovered by solving the inner problem in (39). If Λ were optimal it would have to satisfy $\|\mathbf{u}'\Lambda\|_2 \leq 1\ \forall \mathbf{u} \in \mathcal{U}_2^2$ in (41); hence a maximally violated constraint can be efficiently computed by solving $\mathbf{u}^* \in \arg\max_{\mathbf{u}\in\mathcal{U}_2^2} \mathbf{u}'\Lambda\Lambda'\mathbf{u}$ (using the same power method as before). If $\|\mathbf{u}^{*\prime}\Lambda\|_2 \leq 1+\epsilon$ the procedure halts. Otherwise $\mathbf{u}^*$ is added as a new column to U, and the procedure repeats. Each iteration makes maximum greedy progress in (39) and convergence to the optimum is not hard to establish.

Therefore, by applying these computational methods we obtain an efficient global procedure to solve a semi-supervised representation learning problem in the spirit of (Rish et al. 2007; Raina et al. 2007; Lee et al. 2009; Mairal et al. 2008). Recently, (Goldberg et al. 2010) has proposed a transductive dimensionality reduction formulation that is also convex. However, that formulation does not provide extraction of the representation Φ nor the prediction model W—instead it only recovers the analog of $\hat{Z}$ containing transductive predictions on the unlabeled data—and it cannot enforce individual constraints on the supervised (W) and unsupervised (B) parts of the model respectively.

Experimental Results

Algorithms To evaluate the proposed convex sparse semi-supervised learning method (CS^3), we compared its performance to two local and one convex approach respectively: alternation (ALT), staged-alternation (STAGE) and a transductive matrix completion method (Goldberg et al. 2010). In the alternating approach, two of the three variables, B, Φ, W, are fixed and the other optimized, repeating optimization over each variable until convergence. In the staged-alternator, the optimizer alternates between B and Φ until convergence and then optimizes the prediction model, W. Note that the staged-alternator is the approach taken by (Lee et al. 2009); however, we included their implementation in the results for completeness.[7] The implementation of the transductive matrix completion method follows the settings outlined by (Goldberg et al. 2010).

[3] When maximizing a convex function of H one of the extreme points in $\{H : H\succeq 0, \text{tr}(HI_n)=1, \text{tr}(HI_c)=\gamma^2\}$ must be optimal. It is known that these extreme points have rank at most 1 (Pataki 1998), hence the rank constraint can be dropped in (31).

[4] The objective $\|D_\rho\Lambda\|_{sp}$ is also quasi-convex in $\rho \in (0,\infty)$.

[5] The objective $\|D_\rho^{-1}\hat{Z}\|_{tr}$ is quasi-concave in $\rho \in (0,\infty)$.

[6] In our experiments we therefore fix $\alpha^\star$, recover $\hat{\Lambda}$ and $\hat{Z}$ by the dual formulation given above, then recover the corresponding α in the primal problem by $\alpha = -(L_c^\star(\hat{\Lambda}; Z) + L_c(\hat{Z}; Z))/\|\hat{Z}'\|^*_{(\mathcal{U}_2^2,2)}$.

[7] http://www.eecs.umich.edu/∼honglak/softwares/fast_sc.tgz

Datasets We investigated six classification datasets: (i) A synthetic dataset with features and labels generated analogously to (Goldberg et al. 2010), which contains 20 features and 400 samples. The rank of the generated feature matrix is 4, and zero mean independent Gaussian noise with variance $\sigma^2 = 0.1$ was added to the features. (ii) A UCI dataset, Wisconsin Breast Cancer (WBC), which contains 10 features and 683 samples.[8] (iii) A UCI dataset, Ionosphere, which contains 34 features and 351 samples.[9] (iv) Three semi-supervised learning benchmark datasets, BCI, COIL and g241n, which all contain 1500 samples, with 117, 241 and 241 features respectively.[10] For experiments on transductive learning, we randomly selected from each dataset L examples as the labeled data and U examples as the unlabeled data. Both L and U are reported in Table 2. Then we measured the transductive error on the U examples, and this error was further averaged over runs on five different random choices of the L and U examples.

Parameter selection Cross validation is ineffective here due to the small number of training points. Instead, we iterated over several different parameter settings for the two regularization parameters, β and α, and the infinity norm bound, γ, and chose the one that produced lowest test label error, individually for each algorithm. For example, in Table 2, we chose the best $\beta \in \{10^{-5},\ 10^{-3},\ 10^{-2},\ 0.1,\ 1\}$. The number of bases for the staged and alternating algorithms was set to 100 for a balance between runtime and accuracy.

Comparison 1: Optimization We first investigated CS^3's ability to obtain lower objective values, by setting the loss functions and parameters to common choices across the algorithms. In particular, we set the supervised loss to the square Frobenius norm, the unsupervised loss to the logistic loss, and the regularizer to $||\Phi||_{2,1}$. We minimized (38) by L-BFGS. For the local optimization methods ALT and STAGE a projected gradient method was used to enforce norm constraints on B and a constrained optimization was used for the infinity norm on W. To evaluate optimization quality, we fixed $\gamma = 1$ and tested on $\beta \in \{0.1, 10\}$ and $\alpha^\star \in \{0.1, 10\}$ (using the recovered α to train ALT and STAGE). Two thirds of the examples were used as labeled data while the rest used as unlabeled data. Table 1 shows that CS^3 outperforms the non-convex approaches in terms of both the objective value attained and the training cost on all the data sets. Interestingly, ALT and STAGE occasionally find a solution that is very close to the global optimum.

Comparison 2: Transductive error We then evaluated the transductive generalization error attained by the different methods. In this case, we considered different choices for the loss functions, and report the results for the best, using either a soft-margin support vector machine (hinge loss) or smooth logistic loss for the prediction model; and either projecting or not projecting B and W in the local optimization methods. Limited memory BFGS was used in all cases with a smooth loss function, excluding (Lee et al. 2009) who used a conjugate gradient method with a smooth ϵ-L_1 regularizer. In Table 2, the best results for the alternators were obtained with a hinge loss with an unprojected optimization.

Table 1: Minimum objective values in Equation (25) obtained by the training methods on six data sets. The numbers in the parenthesis indicate runtime in seconds.

Method	$\beta = 0.1$		$\beta = 10$	
	$\alpha^\star = 0.1$	$\alpha^\star = 10$	$\alpha^\star = 0.1$	$\alpha^\star = 10$
COIL				
CS^3	0.071 (97)	0.070 (156)	6.934 (82)	6.809 (93)
ALT	0.075 (2044)	0.072 (2084)	7.226 (5188)	6.951 (3439)
STAGE	3.611 (95)	3.052 (94)	6.934 (630)	6.934 (597)
WBC				
CS^3	0.119 (2)	0.113 (27)	6.981 (4)	6.711 (4)
ALT	0.122 (360)	0.114 (292)	7.245 (1520)	6.796 (872)
STAGE	0.119 (167)	0.119 (166)	6.981 (163)	6.981 (163)
BCI				
CS^3	0.074 (30)	0.069 (43)	6.936 (10)	6.483 (12)
ALT	0.077 (243)	0.069 (288)	7.200 (992)	6.534 (528)
STAGE	0.074 (83)	0.074 (105)	6.936 (150)	6.936 (158)
IONOSPHERE				
CS^3	0.084 (5)	0.078 (75)	6.946 (3)	6.434 (5)
ALT	0.087 (194)	0.078 (167)	7.242 (722)	6.477 (320)
STAGE	0.084 (58)	0.084 (71)	6.946 (119)	6.946 (103)
G241N				
CS^3	0.071 (73)	0.070 (111)	6.934 (66)	6.809 (94)
ALT	0.075 (1731)	0.217 (1886)	7.242 (4951)	6.966 (3782)
STAGE	3.601 (92)	3.054 (81)	6.934 (621)	6.934 (568)
SYNTHETIC				
CS^3	0.094 (3)	0.090 (15)	6.956 (2)	6.503 (4)
ALT	0.097 (193)	0.090 (155)	7.216 (819)	6.556 (384)
STAGE	0.094 (51)	0.094 (83)	6.956 (137)	6.956 (128)

One can see that in every case CS^3 is either comparable to or outperforms the other competitors. Surprisingly, though the approach in (Goldberg et al. 2010) is the most similar to CS^3, it performs noticeably worse than CS^3. ALT performs surprisingly poorly for WBC and Ionosphere; one possible reason is that the mixed supervised classification and unsupervised regression losses create poor local minima. This suggests that for an alternating minimization approach, separating the problem into a factorization step and a classification learning step is more appropriate. We can also see that LEE ET AL. often performs better than STAGE, despite having the same objective; this result is likely due to optimizations in their code, such as the smoothed sparse regularizer.

Conclusion

We have developed a general framework for expressing convex representation learning problems. For subspace learning, we showed that trace norm regularization is the natural consequence of using $\|\Phi\|_{2,1}$. For sparse coding, we found the sparse regularizer $\|\Phi\|_{1,1}$ leads to vector quantization if the number of features is not restricted. Our general framework admits many other formulations, and we demonstrated a new convex formulation of semi-supervised subspace learning that shows the benefits of globally training multiple components. For future work, we are investigating

[8] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

[9] http://archive.ics.uci.edu/ml/datasets/Ionosphere

[10] http://www.kyb.tuebingen.mpg.de/ssl-book/benchmarks.html

Table 2: Average test (transductive) error of sparse coding techniques on a variety of datasets ($\pm$ standard deviation).

	COIL (n=241, L=10, U=100)	WBC (n=10, L=10, U=50)	BCI (n=117, L=10, U=200)	IONOSPHERE (n=34, L=10, U=300)	G241N (n=241, L=10, U=100)	SYNTHETIC (n=20, L=40, U=360)
ALT	0.464 ± 0.036	0.388 ± 0.156	0.440 ± 0.028	0.457 ± 0.075	0.478 ± 0.053	0.464 ± 0.019
STAGED	0.476 ± 0.037	0.200 ± 0.043	0.452 ± 0.041	0.335 ± 0.050	0.484 ± 0.050	0.417 ± 0.035
LEE ET AL.	0.414 ± 0.029	0.168 ± 0.100	0.436 ± 0.093	0.350 ± 0.042	0.452 ± 0.073	0.411 ± 0.027
GOLDBERG	0.484 ± 0.068	0.288 ± 0.105	0.540 ± 0.025	0.338 ± 0.053	0.524 ± 0.022	0.496 ± 0.018
CS^3	0.388 ± 0.043	0.134 ± 0.072	0.380 ± 0.069	0.243 ± 0.042	0.380 ± 0.036	0.341 ± 0.013

extensions to structured and hierarchical sparsity (Jenatton et al. 2010), factored sparsity (Jia, Salzmann, and Darrell 2010), and robust formulations (Candes et al. 2009).

Acknowledgments

Thanks to Özlem Aslan for her assistance with this research. Work supported by NSERC, AICML, the University of Alberta, MITACS, and the Canada Research Chairs program.

References

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Mach. Learn.* 73:243–272.

Bach, F.; Mairal, J.; and Ponce, J. 2008. Convex sparse matrix factorizations. arXiv:0812.1869v1.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge U. Press.

Bradley, D., and Bagnell, J. 2008. Differentiable sparse coding. In *NIPS 22*.

Bradley, D., and Bagnell, J. 2009. Convex coding. In *UAI*.

Candes, E., and Recht, B. 2008. Exact matrix completion via convex optimization. *Found. Comput. Math.* 9:717–772.

Candes, E.; Li, X.; Ma, Y.; and Wright, J. 2009. Robust principal component analysis? arXiv:0912.3599.

Collins, M.; Dasgupta, S.; and Schapire, R. 2001. A generalization of principal component analysis to the exponential family. In *NIPS*.

Comon, P. 1994. Independent component analysis, a new concept? *Signal Processing* 36(3):287–314.

Elad, M., and Aharon, M. 2006. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. on Image Processing* 15:3736–3745.

Goldberg, A.; Zhu, X.; Recht, B.; Xu, J.; and Nowak, R. 2010. Transduction with matrix completion: Three birds with one stone. In *NIPS 23*.

Gordon, G. 2002. Generalized^2 linear^2 models. In *NIPS 15*.

Hendrickx, J., and Olshevsky, A. 2010. Matrix p-norms are NP-hard to approximate if $p \neq 1, 2, \infty$. *SIAM J. Matrix Anal. Appl.* 31(5):2802–2812.

Hinton, G. 2007. Learning multiple layers of representations. *Trends in Cognitive Sciences* 11:428–434.

Horn, R., and Johnson, C. 1985. *Matrix Analysis*. Cambridge.

Jenatton, R.; Mairal, J.; Obozinski, G.; and Bach, F. 2010. Proximal methods for sparse hierarchical dictionary learning. In *ICML*.

Jia, Y.; Salzmann, M.; and Darrell, T. 2010. Factorized latent spaces with structured sparsity. In *NIPS 23*.

Lee, H.; Raina, R.; Teichman, A.; and Ng, A. 2009. Exponential family sparse coding with application to self-taught learning. In *IJCAI*.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; and Zisserman, A. 2008. Supervised dictionary learning. In *NIPS 21*.

Nowozin, S., and Bakir, G. 2008. A decoupled approach to exemplar-based unsupervised learning. In *ICML*.

Olshausen, B., and Field, D. 1997. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* 37:3311–3325.

Pataki, G. 1998. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Math. Oper. Res.* 23(2):339–358.

Pereira, F., and Gordon, G. 2006. The support vector decomposition machine. In *ICML*.

Petz, D. 2004. A survey of trace inequalities. In *Functional Analysis and Operator Theory*, 287–298. Banach Center.

Raina, R.; Battle, A.; Lee, H.; Packer, B.; and Ng, A. 2007. Self-taught learning: Transfer learning from unlabeled data. In *ICML*.

Recht, B.; Fazel, M.; and Parrilo, P. 2007. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Review* 52(3):471–501.

Rish, I.; Grabarnik, G.; Cecchi, G.; Pereira, F.; and Gordon, G. 2007. Closed-form supervised dimensionality reduction with generalized linear models. In *ICML*.

Rockafellar, R. 1970. *Convex Analysis*. Princeton U. Press.

Salakhutdinov, R., and Srebro, N. 2010. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *NIPS 23*.

Srebro, N.; Rennie, J.; and Jaakkola, T. 2004. Maximum-margin matrix factorization. In *NIPS 17*.

Steinberg, D. 2005. Computation of matrix norms with applications to robust optimization. Master's thesis, Technion.

van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR* 9:2579–2605.

Weinberger, K., and Saul, L. 2006. Unsupervised learning of image manifolds by semidefinite programming. *IJCV* 70(1):77–90.

Xu, H.; Caramanis, C.; and Sanghavi, S. 2010. Robust PCA via outlier pursuit. In *NIPS 23*.

Zou, H.; Hastie, T.; and Tibshirani, R. 2006. Sparse principal component analysis. *JCGS* 15(2):262–286.

Appendix

Correctness of solution recovery

Recall: given $\hat{Z}$, the recovery procedure computes an optimal U and Φ (and Λ) such that $U\Phi = \hat{Z}$ and $\|\Phi\|_{2,1} = \|\hat{Z}\|^*_{(\mathcal{U}_2^2,2)}$. The procedure is iterative: given a current $U_k \in (\mathcal{U}_2^2)^k$ the following updates are computed

$$(\Phi_k, \Lambda_k) \in \arg\min_{\Phi}\max_{\Lambda} \|\Phi\|_{2,1} + \text{tr}(\Lambda'(\hat{Z} - U_k\Phi)) \quad (42)$$

$$\mathbf{u}_{k+1} \in \arg\max_{\mathbf{u}\in\mathcal{U}_2^2} \|\mathbf{u}'\Lambda_k\|_2 \quad (43)$$

$$U_{k+1} = [U_k, \mathbf{u}_{k+1}]. \quad (44)$$

The procedure halts when $\|\mathbf{u}'_{k+1}\Lambda_k\|_2 \leq 1 + \epsilon$.

Proposition 4 *If $\|\mathbf{u}'_{k+1}\Lambda_k\|_2 \leq 1$ then the procedure must halt, and (U_k, Φ_k, Λ_k) must be an optimal solution.*

Proof: Halting is obvious. To establish optimality, note

$$\begin{aligned}
&\|\hat{Z}\|^*_{(\mathcal{U}_2^2,2)} \\
&= \min_{U\in(\mathcal{U}_2^2)^\infty} \min_{\Phi:U\Phi=\hat{Z}} \|\Phi\|_{2,1} && (45)\\
&\leq \min_{\Phi:U_k\Phi=\hat{Z}} \|\Phi\|_{2,1} && (46)\\
&= \min_{\Phi}\max_{\Lambda} \|\Phi\|_{2,1} + \text{tr}(\Lambda'(\hat{Z} - U_k\Phi)) && (47)\\
&= \max_{\Lambda}\min_{\Phi} \|\Phi\|_{2,1} + \text{tr}(\Lambda'(\hat{Z} - U_k\Phi)) && (48)\\
&= \max_{\Lambda}\min_{\Phi} \max_{\Psi:\|\Psi\|_{2,\infty}\leq 1} \text{tr}(\Psi'\Phi) + \text{tr}(\Lambda'(\hat{Z} - U_k\Phi)) && (49)\\
&= \max_{\Lambda} \max_{\Psi:\|\Psi\|_{2,\infty}\leq 1} \min_{\Phi} \text{tr}(\Psi'\Phi) + \text{tr}(\Lambda'(\hat{Z} - U_k\Phi)) && (50)\\
&= \max_{\Psi:\|\Psi\|_{2,\infty}\leq 1} \max_{\Lambda:U_k'\Lambda=\Psi} \text{tr}(\Lambda'\hat{Z}) && (51)\\
&= \max_{\Lambda:\|U_k'\Lambda\|_{2,\infty}\leq 1} \text{tr}(\Lambda'\hat{Z}) && (52)\\
&= \text{tr}(\Lambda_k'\hat{Z}), && (53)
\end{aligned}$$

where (53) follows from the optimality of Λ_k. Now using the fact that $\|\mathbf{u}'_{k+1}\Lambda_k\|_{sp} \leq 1$ by assumption, and hence $\max_{\mathbf{u}\in\mathcal{U}_2^2} \|\mathbf{u}'\Lambda_k\|_{sp} \leq 1$, we get $\|\mathbf{u}'\Lambda_k\|_{sp} \leq 1\ \forall \mathbf{u} \in \mathcal{U}_2^2$. That is, Λ_k must be a feasible point in the constraint set $\{\Lambda : \|\mathbf{u}'\Lambda\|_{sp} \leq 1\ \forall \mathbf{u} \in \mathcal{U}_2^2\}$. Therefore

$$\begin{aligned}
\text{tr}(\Lambda_k'\hat{Z}) &\leq \max_{\Lambda:\|\mathbf{u}'\Lambda\|_{sp}\leq 1\ \forall \mathbf{u}\in\mathcal{U}_2^2} \text{tr}(\Lambda'\hat{Z}) && (54)\\
&= \|\hat{Z}\|^*_{(\mathcal{U}_2^2,2)}, && (55)
\end{aligned}$$

where (55) follows from (39)-(41). Thus we conclude that the expressions (45)-(55) are all equal, which implies that $\|\Phi_k\|_{2,1} = \|\hat{Z}\|^*_{(\mathcal{U}_2^2,2)}$. ■

Proposition 5 *If $\epsilon = 0$ the recovery procedure must converge to an optimal solution.*

Proof: Given Proposition 4, the only way the iteration (42)-(44) could fail to converge to an optimal solution would be if there existed a $\delta > 0$ and a natural number K such that $\|\mathbf{u}'_{k+1}\Lambda_k\|_{sp} \geq 1 + \delta$ for all $k \geq K$. Assume this to be the case. But for any such k we would then have $\|\mathbf{u}'_{k+1}\Lambda_\ell\|_{sp} \leq 1$ for all $\ell > k$ by the construction of Λ_ℓ; see e.g. (52). Then by the triangle inequality

$$\begin{aligned}
\delta &\leq \|\mathbf{u}'_{k+1}\Lambda_k\|_{sp} - \|\mathbf{u}'_{k+1}\Lambda_\ell\|_{sp} && (56)\\
&\leq \|\mathbf{u}'_{k+1}(\Lambda_k - \Lambda_\ell)\|_{sp} && (57)\\
&\leq \|\mathbf{u}_{k+1}\|_2\ \|\Lambda_k - \Lambda_\ell\|_{sp} && (58)\\
&= \sqrt{2}\|\Lambda_k - \Lambda_\ell\|_{sp}, && (59)
\end{aligned}$$

since $\mathbf{u}_{k+1} \in \mathcal{U}_2^2$. Therefore for all $k \geq K$, Λ_k maintains a ball of radius $\delta/2$ (in spectral norm) that contains no Λ_ℓ for $\ell > k$. But every Λ_k is also a feasible point in the set $\{\Lambda : \|U_0'\Lambda\|_{2,\infty} \leq 1\}$, which is bounded (if U_0 spans the space). This leads to a contradiction. ■

In the practical implementation, we stop the iteration when $\|\mathbf{u}'_{k+1}\Lambda_k\|_2 \leq 1 + \epsilon$. The proof of Proposition 5 immediately implies that the algorithm must halt within a finite number of iterations.