Convex Feature Discovery

Dale Schuurmans

Xinhua Zhang, Yaoliang Yu, Martha White, Ruitong Huang

University of Alberta

Paper appearing in AAAI-2011

Discovering a feature representation from data

Examples

- dimensionality reduction (PCA, exponential family PCA)
- sparse coding
- independent component analysis
- deep learning

⋮

Usually involves learning both

a latent representation for data and a data reconstruction model

Context

could be: unsupervised, semi-supervised, or supervised

Challenge

Optimal feature discovery appears to be generally intractable

Have to jointly train

- latent representation
- data reconstruction model

Usually resort to alternating minimization

(sole exception: PCA)

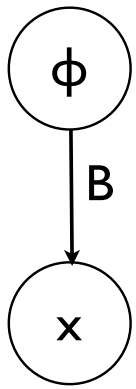
Main result

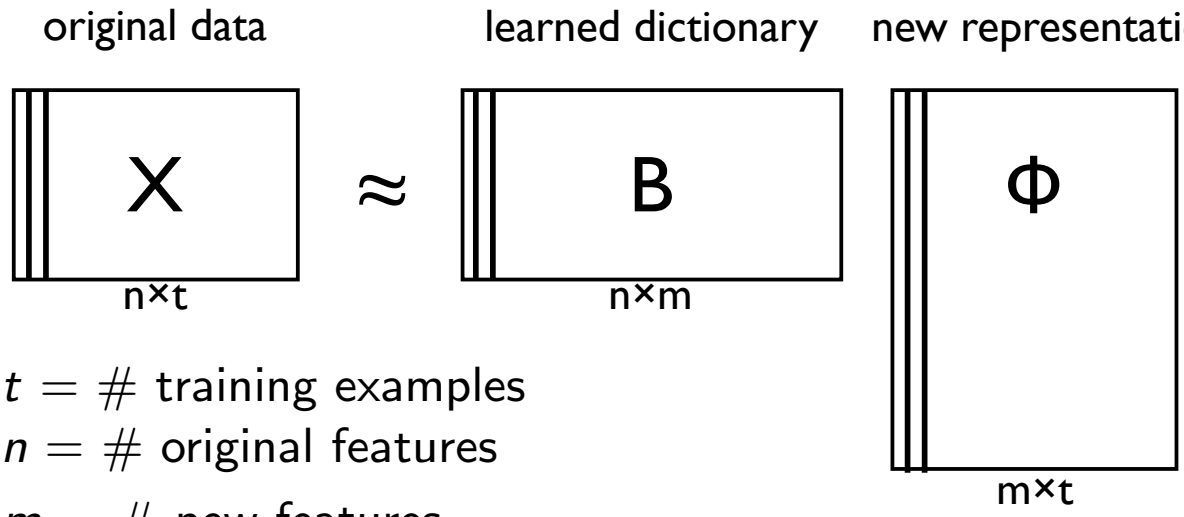
- Many feature discovery problems can be solved globally
- Requires a transformation of the joint training problem
- Can easily extend to semi-supervised and supervised scenarios

First consider unsupervised feature discovery

Unsupervised feature discovery

Single layer case = matrix factorization





t = # training examples
n = # original features
m = # new features

Choose B and Φ to minimize data reconstruction loss

$$L(B\Phi; X) = \sum_{i=1}^{t} L(B\Phi_{:i}; X_{:i})$$

Seek desired structure in latent feature representation

Φ low rank	:	dimensionality reduction
Φ sparse	:	sparse coding
Φ rows independent	:	independent component analysis

Generalized matrix factorization

Assume reconstruction loss $L(\hat{\mathbf{x}}; \mathbf{x})$ is convex in first argument

Bregman divergence

$$L(\hat{\mathbf{x}}; \mathbf{x}) = D_{F^*}(\hat{\mathbf{x}} \| \mathbf{x}) = D_F(f(\mathbf{x}) \| f(\hat{\mathbf{x}}))$$

(F strictly convex potential with transfer $f = \nabla F$)

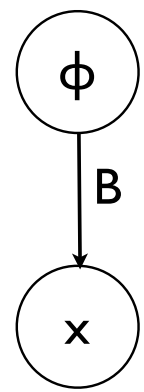
Tries to make $\hat{\mathbf{x}} \approx \mathbf{x}$

Matching loss

$$L(\hat{\mathbf{x}}; \mathbf{x}) = D_{F^*}(\hat{\mathbf{x}} \| f^{-1}(\mathbf{x})) = D_F(\mathbf{x} \| f(\hat{\mathbf{x}}))$$

Tries to make $f(\hat{\mathbf{x}}) \approx \mathbf{x}$

(A nonlinear predictor, but loss still convex in $\hat{\mathbf{x}}$)



Regular exponential family

$$L(\hat{\mathbf{x}}; \mathbf{x}) = -\log p_B(\mathbf{x}|\phi) = D_{F^*}(\hat{\mathbf{x}} \| f^{-1}(\mathbf{x})) - F(\mathbf{x}) - const$$

Training problem

$$\min_{B \in \mathbb{R}^{n \times m}} \quad \min_{\Phi \in \mathbb{R}^{m \times t}} \quad L(B\Phi; X)$$

How to impose desired structure on Φ?

Training problem

$$\min_{B \in \mathbb{R}^{n \times m}} \min_{\Phi \in \mathbb{R}^{m \times t}} L(B\Phi; X)$$

How to impose desired structure on Φ?

Dimensionality reduction

Fix # features $m < \min(n, t)$

- But only known to be tractable if $L(\hat{X}; X) = \|\hat{X} - X\|_F^2$ (PCA)
- No known efficient algorithm for other standard losses

Problem

$\mathrm{rank}(\Phi) = m$ constraint is too hard

Training problem

$$\min_{B \in \mathcal{B}_2^m} \min_{\Phi \in \mathbb{R}^{m \times t}} L(B\Phi; X) + \alpha \|\Phi\|_{2,1}$$

How to impose desired structure on Φ?

Relaxed dimensionality reduction (subspace learning)

Add rank reducing regularizer

$$\|\Phi\|_{2,1} = \sum_{j=1}^{m} \|\Phi_{j:}\|_2$$

Favors null rows in Φ

But need to add constraint to B

$$B_{:j} \in \mathcal{B}_2 = \{\mathbf{b} : \|\mathbf{b}\|_2 \leq 1\}$$

(Otherwise can make Φ small just by making B large)

Training problem

$$\min_{B \in \mathcal{B}_q^m} \min_{\Phi \in \mathbb{R}^{m \times t}} L(B\Phi; X) + \alpha \|\Phi\|_{1,1}$$

How to impose desired structure on Φ?

Sparse coding

Use sparsity inducing regularizer

$$\|\Phi\|_{1,1} = \sum_{j=1}^{m} \sum_{i=1}^{t} |\Phi_{ji}|$$

Favors sparse entries in Φ

Need to add constraint to B

$$B_{:j} \in \mathcal{B}_q = \{\mathbf{b} : \|\mathbf{b}\|_q \leq 1\}$$

(Otherwise can make Φ small just by making B large)

Training problem

$$\min_{B \in \mathbb{R}^{n \times m}} \min_{\Phi \in \mathbb{R}^{m \times t}} L(B\Phi; X) + \alpha D(\Phi)$$

How to impose desired structure on Φ?

Independent components analysis

Usually enforces $B\Phi = X$ as a constraint

- but interpolation is generally a bad idea
- Instead just minimize reconstruction loss
  plus a dependence measure $D(\Phi)$ as a regularizer

Difficulty

Formulating a reasonable convex dependence penalty

Training problem

Consider subspace learning and sparse coding

$$\min_{B \in \mathcal{B}^m} \min_{\Phi \in \mathbb{R}^{m \times t}} L(B\Phi; X) + \alpha \|\Phi\|$$

Choice of $\|\Phi\|$ and $\mathcal{B}$ determines type of representation recovered

Training problem

Consider subspace learning and sparse coding

$$\min_{B \in \mathcal{B}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|$$

Choice of $\|\Phi\|$ and $\mathcal{B}$ determines type of representation recovered

Problem

Still have rank constraint imposed by # new features m

Idea

Just relax $m \to \infty$

- Rely on sparsity inducing norm $\|\Phi\|$ to select features

Training problem

Consider subspace learning and sparse coding

$$\min_{B \in \mathcal{B}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|$$

Still have a problem

Optimization problem is not jointly convex in B and Φ

Training problem

Consider subspace learning and sparse coding

$$\min_{B \in \mathcal{B}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|$$

Still have a problem

Optimization problem is not jointly convex in B and Φ

Idea 1: Alternate!

- convex in B given Φ
- convex in Φ given B

Could use any other form of local training

Training problem

Consider subspace learning and sparse coding

$$\min_{B \in \mathcal{B}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|$$

Still have a problem

Optimization problem is not jointly convex in B and Φ

Idea 2: Boost!

- Implicitly fix B to universal dictionary
- Keep row-wise sparse Φ
- Incrementally select column in B ("weak learning problem")
- Update sparse Φ

Can prove convergence under broad conditions

Training problem

Consider subspace learning and sparse coding

$$\min_{B \in \mathcal{B}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|$$

Still have a problem?

Optimization problem is not jointly convex in B and Φ

Idea 3: Solve!

- Can easily solve for globally optimal joint B and Φ
- **But** requires a significant reformulation

Key observation

Equivalent reformulation

Theorem

$$\min_{B \in \mathcal{B}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|_{p,1}$$
$$= \min_{\hat{X} \in \mathbb{R}^{n \times t}} L(\hat{X}; X) + \alpha |||\hat{X}|||$$

- $|||\cdot|||$ is an induced matrix norm on $\hat{X}$ determined by $\mathcal{B}$ and $\|\cdot\|_{p,1}$

Important fact

Norms are always convex

Computational strategy

1. Solve for optimal response matrix $\hat{X}$ first (convex minimization)
2. Then recover optimal B and Φ from $\hat{X}$

Independently observed by Bach et al. (arXive 2008)

Example: subspace learning

$$\min_{B \in \mathcal{B}_2^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|_{2,1}$$
$$= \min_{\hat{X} \in \mathbb{R}^{n \times t}} L(\hat{X}; X) + \alpha \|\hat{X}\|_{tr}$$

Recovery

- Let $U\Sigma V' = \mathrm{svd}(\hat{X})$
- Set $B = U$ and $\Phi = \Sigma V'$

Preserves optimality

- $\|B_{:j}\|_2 = 1$ hence $B \in \mathcal{B}_2^n$
- $\|\Phi\|_{2,1} = \|\Sigma V'\|_{2,1} = \sum_j \sigma_j \|V_{:j}\|_2 = \sum_j \sigma_j = \|\hat{X}\|_{tr}$

Thus

$$L(\hat{X}; X) + \alpha \|\hat{X}\|_{tr} \quad = \quad L(B\Phi; X) + \alpha \|\Phi\|_{2,1}$$

Example: sparse coding

$$\min_{B \in \mathcal{B}_q^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|_{1,1}$$

$$= \min_{\hat{X} \in \mathbb{R}^{n \times t}} L(\hat{X}; X) + \alpha \|\hat{X}'\|_{q,1}$$

Recovery

$$B = \left[\frac{1}{\|\hat{X}_{:1}\|_q} \hat{X}_{:1}, \ldots, \frac{1}{\|\hat{X}_{:t}\|_q} \hat{X}_{:t} \right] \quad \text{(rescaled columns)}$$

$$\Phi = \begin{bmatrix} \|\hat{X}_{:1}\|_q & & 0 \\ & \ddots & \\ 0 & & \|\hat{X}_{:t}\|_q \end{bmatrix} \quad \text{(diagonal matrix)}$$

Preserves optimality

- $\|B_{:j}\|_q = 1$ hence $B \in \mathcal{B}_q^t$
- $\|\Phi\|_{1,1} = \sum_j \|\hat{X}_{:j}\|_q = \|\hat{X}'\|_{q,1}$

Thus

$$L(\hat{X}; X) + \alpha \|\hat{X}'\|_{q,1} = L(B\Phi; X) + \alpha \|\Phi\|_{1,1}$$

Example: sparse coding

Outcome

Sparse coding with $\|\cdot\|_{1,1}$ regularization $=$ vector quantization

- drops some examples
- memorizes remaining examples

Optimal solution is not overcomplete

Could not make these observations using local solvers

Simple extensions

- Missing observations in X
- Robustness to outliers in X

$$\min_{S \in \mathbb{R}^{n \times t}} \min_{\hat{X} \in \mathbb{R}^{n \times t}} L(\ (\hat{X} + S)_\Omega\ ;\ X_\Omega\ ) + \alpha \|\hat{X}\|_* + \beta \|S\|_{1,1}$$

$\Omega =$ observed indices in X

$S =$ speckled outlier noise

(jointly convex in $\hat{X}$ and S)

Explaining the main result

Theorem

$$\min_{B \in \mathcal{B}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|_{p,1}$$
$$= \min_{\hat{X} \in \mathbb{R}^{n \times t}} L(\hat{X}; X) + \alpha |||\hat{X}|||$$

for an induced matrix norm $|||\hat{X}||| = \|\hat{X}'\|^*_{(\mathcal{B}, p^*)}$

Explaining the main result

Theorem

$$\min_{B \in \mathcal{B}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|_{p,1}$$
$$= \min_{\hat{X} \in \mathbb{R}^{n \times t}} L(\hat{X}; X) + \alpha |||\hat{X}|||$$

for an induced matrix norm $|||\hat{X}||| = \|\hat{X}'\|^*_{(\mathcal{B},p^*)}$

A dual norm

$$\|\hat{X}'\|^*_{(\mathcal{B},p^*)} = \max_{\|\Lambda'\|_{(\mathcal{B},p^*)} \leq 1} \mathrm{tr}(\Lambda' \hat{X})$$

(standard definition of a dual norm)

Explaining the main result

Theorem

$$\min_{B \in \mathcal{B}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|_{p,1}$$
$$= \min_{\hat{X} \in \mathbb{R}^{n \times t}} L(\hat{X}; X) + \alpha |||\hat{X}|||$$

for an induced matrix norm $|||\hat{X}||| = \|\hat{X}'\|^*_{(\mathcal{B},p^*)}$

A dual norm

$$\|\hat{X}'\|^*_{(\mathcal{B},p^*)} = \max_{\|\Lambda'\|_{(\mathcal{B},p^*)} \leq 1} \operatorname{tr}(\Lambda' \hat{X})$$

(standard definition of a dual norm)

of a vector-norm induced matrix norm

$$\|\Lambda'\|_{(\mathcal{B},p^*)} = \max_{\mathbf{b} \in \mathcal{B}} \|\Lambda' \mathbf{b}\|_{p^*}$$

(easy to prove this yields a norm on matrices)

Proof outline

$$\min_{B\in\mathcal{B}^\infty} \min_{\Phi\in\mathbb{R}^{\infty\times t}} L(B\Phi; X) + \alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} \min_{B\in\mathcal{B}^\infty} \min_{\Phi: B\Phi=\hat{X}} L(\hat{X}; X) + \alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X}; X) + \alpha \min_{B\in\mathcal{B}^\infty} \min_{\Phi: B\Phi=\hat{X}} \|\Phi\|_{p,1}$$

Proof outline

$$\min_{B\in\mathcal{B}^\infty} \min_{\Phi\in\mathbb{R}^{\infty\times t}} L(B\Phi; X) + \alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} \min_{B\in\mathcal{B}^\infty} \min_{\Phi: B\Phi=\hat{X}} L(\hat{X}; X) + \alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X}; X) + \alpha \min_{B\in\mathcal{B}^\infty} \min_{\Phi: B\Phi=\hat{X}} \|\Phi\|_{p,1}$$

For any $B \in \mathcal{B}^\infty$ that spans the columns of $\hat{X}$

$$\min_{\Phi: B\Phi=\hat{X}} \|\Phi\|_{p,1} = \min_{\Phi} \max_{\Lambda} \max_{\|V\|_{p^*,\infty}\le 1} \operatorname{tr}(V'\Phi) + \operatorname{tr}(\Lambda'(\hat{X} - B\Phi))$$

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

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

$$= \max_{\Lambda: \|B'\Lambda\|_{p^*,\infty}\le 1} \operatorname{tr}(\Lambda'\hat{X})$$

Proof outline

$$\min_{B\in\mathcal{B}^\infty} \min_{\Phi\in\mathbb{R}^{\infty\times t}} L(B\Phi; X) + \alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} \min_{B\in\mathcal{B}^\infty} \min_{\Phi: B\Phi=\hat{X}} L(\hat{X}; X) + \alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X}; X) + \alpha \min_{B\in\mathcal{B}^\infty} \max_{\Lambda: \|B'\Lambda\|_{p^*,\infty}\leq 1} \operatorname{tr}(\Lambda'\hat{X})$$

Proof outline

$$\min_{B\in\mathcal{B}^\infty} \min_{\Phi\in\mathbb{R}^{\infty\times t}} L(B\Phi;X) + \alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} \min_{B\in\mathcal{B}^\infty} \min_{\Phi:B\Phi=\hat{X}} L(\hat{X};X) + \alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X};X) + \alpha \underbrace{\min_{B\in\mathcal{B}^\infty} \max_{\Lambda:\|B'\Lambda\|_{p^*,\infty}\le 1}}_{} \operatorname{tr}(\Lambda'\hat{X})$$

Proof outline

$$\min_{B\in\mathcal{B}^\infty} \min_{\Phi\in\mathbb{R}^{\infty\times t}} L(B\Phi; X) + \alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} \min_{B\in\mathcal{B}^\infty} \min_{\Phi: B\Phi=\hat{X}} L(\hat{X}; X) + \alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X}; X) + \alpha \underbrace{\min_{B\in\mathcal{B}^\infty} \max_{\Lambda: \|B'\Lambda\|_{p^*,\infty}\leq 1}} \operatorname{tr}(\Lambda'\hat{X})$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X}; X) + \alpha \max_{\Lambda: \|\Lambda'\|_{(\mathcal{B},p^*)}\leq 1} \operatorname{tr}(\Lambda'\hat{X})$$

Proof outline

$$\min_{B\in\mathcal{B}^\infty} \min_{\Phi\in\mathbb{R}^{\infty\times t}} L(B\Phi;X)+\alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} \min_{B\in\mathcal{B}^\infty} \min_{\Phi:B\Phi=\hat{X}} L(\hat{X};X)+\alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X};X)+\alpha \underbrace{\min_{B\in\mathcal{B}^\infty} \max_{\Lambda:\|B'\Lambda\|_{p^*,\infty}\le 1} \operatorname{tr}(\Lambda'\hat{X})}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X};X)+\alpha \underbrace{\max_{\Lambda:\|\Lambda'\|_{(\mathcal{B},p^*)}\le 1} \operatorname{tr}(\Lambda'\hat{X})}$$

Proof outline

$$\min_{B\in\mathcal{B}^\infty} \min_{\Phi\in\mathbb{R}^{\infty\times t}} L(B\Phi;X)+\alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} \min_{B\in\mathcal{B}^\infty} \min_{\Phi:B\Phi=\hat{X}} L(\hat{X};X)+\alpha\|\Phi\|_{p,1}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X};X)+\alpha \underbrace{\min_{B\in\mathcal{B}^\infty} \max_{\Lambda:\|B'\Lambda\|_{p^*,\infty}\le 1} \operatorname{tr}(\Lambda'\hat{X})}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X};X)+\alpha \underbrace{\max_{\Lambda:\|\Lambda'\|_{(\mathcal{B},p^*)}\le 1} \operatorname{tr}(\Lambda'\hat{X})}$$

$$= \min_{\hat{X}\in\mathbb{R}^{n\times t}} L(\hat{X};X)+\alpha\|\hat{X}'\|^*_{(\mathcal{B},p^*)}$$

done

Closed form induced norms

Theorem

$$\min_{B \in \mathcal{B}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L(B\Phi; X) + \alpha \|\Phi\|_{p,1}$$
$$= \min_{\hat{X} \in \mathbb{R}^{n \times t}} L(\hat{X}; X) + \alpha \|\hat{X}'\|^*_{(\mathcal{B}, p^*)}$$

Special cases

$$\mathcal{B}_2, \|\Phi\|_{2,1} \mapsto \|\hat{X}'\|^*_{(\mathcal{B}_2,2)} = \|\hat{X}\|_{tr} \quad \text{(subspace learning)}$$
$$\mathcal{B}_q, \|\Phi\|_{1,1} \mapsto \|\hat{X}'\|^*_{(\mathcal{B}_q,\infty)} = \|\hat{X}'\|_{q,1} \quad \text{(sparse coding)}$$
$$\mathcal{B}_1, \|\Phi\|_{p,1} \mapsto \|\hat{X}'\|^*_{(\mathcal{B}_1,p^*)} = \|\hat{X}\|_{p,1}$$

Simple experiments

Experimental results

Alternate	:	repeatedly optimize over B, Φ successively
Global	:	recover global joint minimizer over B, Φ

Experimental results: Sparse coding

Objective value achieved

	data set				
	COIL	WBC	BCI	Ionos	G241N
Alternate	1.314	4.918	0.898	1.612	1.312
Global	0.207	0.659	0.306	0.330	0.207

$\times 10^{-2}$

(squared loss, $q = 2$, $\alpha = 10^{-5}$)

Experimental results: Sparse coding

Run time (seconds)

	data set				
	COIL	WBC	BCI	Ionos	G241N
Alternate	1.95	10.54	0.88	1.71	2.37
Global	0.06	0.01	0.01	0.01	0.09

(squared loss, $q = 2$, $\alpha = 10^{-5}$)

Experimental results: Subspace learning

Objective value achieved

	data set				
	COIL	WBC	BCI	Ionos	G241N
Alternate	1.314	4.957	0.903	1.632	1.313
Global	0.072	0.072	0.092	0.079	0.205

$\times 10^{-2}$

(squared loss, $\alpha = 10^{-5}$)

Experimental results: Subspace learning

Run time (seconds)

	data set				
	COIL	WBC	BCI	Ionos	G241N
Alternate	2.40	9.31	1.12	0.47	2.43
Global	2.18	0.06	0.19	0.06	2.11

(squared loss, $\alpha = 10^{-5}$)

Catch

Every norm is convex

But not every induced matrix norm is tractable

$$\|X\|_2 = \sigma_{max}(X)$$

$$\|X\|_1 = \max_j \sum_i |X_{ij}|$$

$$\|X\|_\infty = \max_i \sum_j |X_{ij}|$$

$\|X\|_p$ NP-hard to approximate for $p \neq 1, 2, \infty$

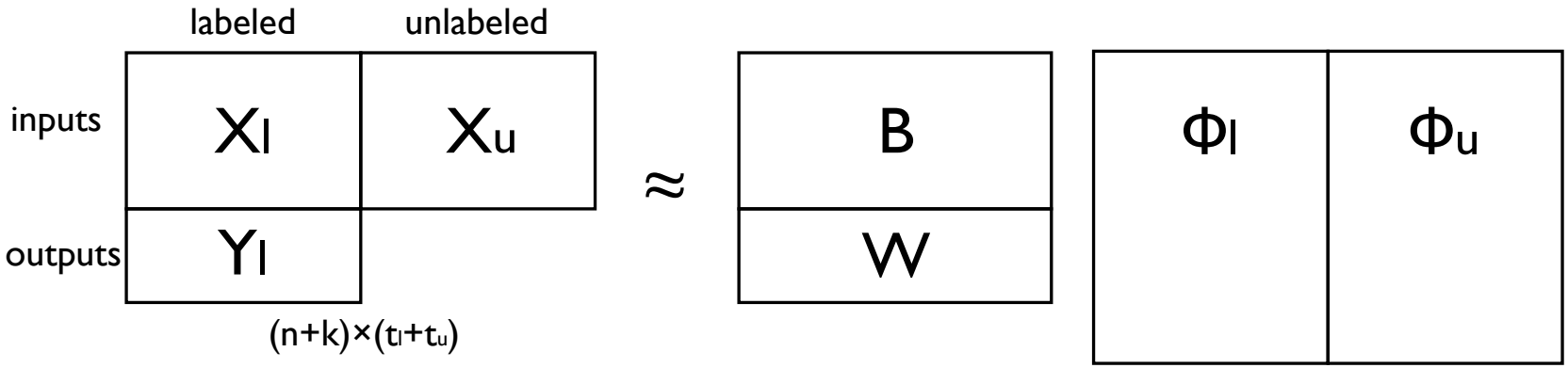
Question

Any other useful induced matrix norms that are tractable?

Yes!

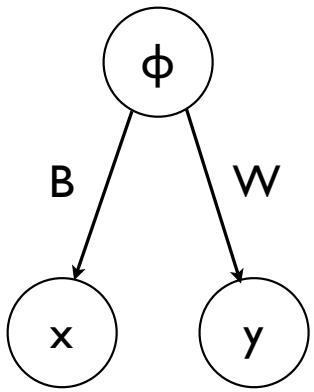
Semi-supervised feature discovery

Semi-supervised feature discovery



$t_l = \#$ labeled
$t_u = \#$ unlabeled
$t = t_l + t_u$

$n = \#$ original features
$k = \#$ output dimensions



Learn

$\Phi = [\Phi_l, \Phi_u]$ data representation
$B =$ input reconstruction model $f(B\Phi) \approx X$
$W =$ output reconstruction model $h(W\Phi_l) \approx Y_l$

Semi-supervised feature discovery

Let

$$Z = \begin{bmatrix} X_l & X_u \\ Y_l & \emptyset \end{bmatrix} \quad U = \begin{bmatrix} B \\ W \end{bmatrix} \quad \mathcal{U} = \begin{bmatrix} \mathcal{B} \\ \mathcal{W} \end{bmatrix}$$

Formulation

$$\min_{B \in \mathcal{B}^\infty} \min_{W \in \mathcal{W}^\infty} \min_{\Phi \in \mathbb{R}^{\infty \times t}} L_u(B\Phi; X) + \beta L_s(W\Phi_l; Y_l) + \alpha \|\Phi\|_{p,1}$$

$$= \min_{\hat{Z} \in \mathbb{R}^{(n+k) \times t}} \tilde{L}(\hat{Z}; Z) + \alpha \|\hat{Z}'\|^*_{(\mathcal{U}, p^*)}$$

Note

Imposing separate constraints on B and W

Questions

- Is the induced norm $\|\hat{Z}'\|^*_{(\mathcal{U}, p^*)}$ efficiently computable?
- Can optimal B, W, Φ be recovered from optimal $\hat{Z}$?

Example: sparse coding formulation

Regularizer: $\|\Phi\|_{1,1}$

Constraints:

$$\begin{aligned} \mathcal{B}_{q_1} &= \{\mathbf{b} : \|\mathbf{b}\|_{q_1} \leq 1\} \\ \mathcal{W}_{q_2} &= \{\mathbf{w} : \|\mathbf{w}\|_{q_2} \leq \gamma\} \\ \mathcal{U}_{q_2}^{q_1} &= \mathcal{B} \times \mathcal{W} \end{aligned}$$

Theorem

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

efficiently computable

Recovery

$$\begin{aligned} \Phi_{jj} &= \max\left(\|\hat{Z}^X_{:j}\|_{q_1},\ \tfrac{1}{\gamma}\|\hat{Z}^Y_{:j}\|_{q_2}\right) \quad \text{(diagonal matrix)} \\ U &= \hat{Z}\Phi^{-1} \end{aligned}$$

Preserves optimality

But still reduces to a form of vector quantization

Example: subspace learning formulation

Regularizer: $\|\Phi\|_{2,1}$

Constraints:

$$\begin{aligned} \mathcal{B}_2 &= \{\mathbf{b} : \|\mathbf{b}\|_2 \leq 1\} \\ \mathcal{W}_2 &= \{\mathbf{w} : \|\mathbf{w}\|_2 \leq \gamma\} \\ \mathcal{U}_2^2 &= \mathcal{B} \times \mathcal{W} \end{aligned}$$

Theorem

$$\|\hat{Z}'\|^*_{(\mathcal{U}_2^2,\infty)} = \max_{\rho \geq 0} \|D_\rho^{-1} \hat{Z}\|_{tr} \quad \text{where } D_\rho = \begin{bmatrix} \sqrt{1+\gamma\rho}\, I & 0 \\ 0 & \sqrt{\frac{1+\gamma\rho}{\rho}}\, I \end{bmatrix}$$

efficiently computable: quasi-concave in ρ

Example: subspace learning formulation

Lemma: dual norm

$$
\begin{aligned}
\|\Lambda'\|^2_{(\mathcal{U}_2^2,2)} &= \max_{\mathbf{h}:\|\mathbf{h}^X\|_2=1,\ \|\mathbf{h}^Y\|_2=\gamma} \mathbf{h}'\Lambda\Lambda'\mathbf{h} \\
&= \max_{H:H\succeq 0,\ tr(HI^X)=1,\ tr(HI^Y)=\gamma} \mathrm{tr}(H\Lambda\Lambda') \\
&= \min_{\lambda\geq 0,\ \nu\geq 0} \ \min_{\Lambda:\Lambda\Lambda'\preceq \lambda I^X+\nu I^Y} \lambda+\gamma\nu \\
&= \min_{\lambda\geq 0,\ \nu\geq 0} \ \min_{\Lambda:\|D_{\nu/\lambda}\Lambda\|^2_{sp}\leq \lambda+\gamma\nu} \lambda+\gamma\nu \\
&= \min_{\lambda\geq 0,\ \nu\geq 0} \ \|D_{\nu/\lambda}\Lambda\|^2_{sp} \\
&= \min_{\rho\geq 0} \ \|D_{\rho}\Lambda\|^2_{sp}
\end{aligned}
$$

Example: subspace learning formulation

Can easily derive target norm from dual norm

$$
\begin{aligned}
\|\hat{Z}'\|^*_{(\mathcal{U}_2^2,2)} &= \max_{\|\Lambda'\|_{(\mathcal{U}_2^2,2)}\leq 1} \operatorname{tr}(\Lambda'\hat{Z}) \\
&= \max_{\rho\geq 0} \max_{\Lambda:\|D_\rho\Lambda\|_{sp}\leq 1} \operatorname{tr}(\Lambda'\hat{Z}) \\
&= \max_{\rho\geq 0} \max_{\tilde{\Lambda}:\|\tilde{\Lambda}\|_{sp}\leq 1} \operatorname{tr}(\tilde{\Lambda}' D_\rho^{-1}\hat{Z}) \\
&= \max_{\rho\geq 0} \|D_\rho^{-1}\hat{Z}\|_{tr}
\end{aligned}
$$

(proves theorem)

Example: subspace learning formulation

Computational strategy

Solve in dual, since $\|\Lambda'\|_{(\mathcal{U}_2^2,\infty)}$ can be computed efficiently via partitioned power method iteration

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

Given $\hat{\Lambda}$

- Recover $\hat{Z}^X$ and $\hat{Z}_l^Y$ by solving

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

- Recover $\hat{Z}_u^Y$ by minimizing $\|\hat{Z}'\|_{(\mathcal{U}_2^2,2)}$ (keeping $\hat{Z}^X$, $\hat{Z}_l^Y$ fixed)

Example: subspace learning formulation

Recovery

Given optimal $\hat{Z}$, recover U and Φ iteratively by repeating:

- $(\Phi^{(\ell)}, \Lambda^{(\ell)}) \in \arg\min_{\Phi} \max_{\Lambda} \|\Phi\|_{2,1} + \mathrm{tr}(\Lambda'(\hat{Z} - U^{(\ell)}\Phi))$
- $\mathbf{u}^{(\ell+1)} \in \arg\max_{\mathbf{u} \in \mathcal{U}_2^2} \|\mathbf{u}'\Lambda^{(\ell)}\|_2$
- $U^{(\ell+1)} = [U^{(\ell)}, \mathbf{u}^{(\ell+1)}]$

Converges to optimal U and Φ

- $U^{(\ell)}\Phi^{(\ell)} = \hat{Z}$ for all ℓ
- $\|\Phi^{(\ell)}\|_{2,1} \to \|\hat{Z}'\|^*_{(\mathcal{U}_2^2,2)}$

Experiments

Experimental results: Subspace learning

Staged	:	first locally optimize B, Φ, then optimize W
Alternate	:	repeatedly optimize over B, W, Φ successively
Global	:	recover joint global minimizer over B, Φ, W

Experimental results: Subspace learning

Objective value achieved

	data set				
	COIL	WBC	BCI	Ionos	G241N
Staged	1.384	1.321	0.799	0.769	1.381
Alternate	0.076	0.122	0.609	0.081	0.076
Global	0.070	0.113	0.069	0.078	0.070

(1/3 labeled, 2/3 unlabeled, squared loss, $\alpha^\star = 10$, $\beta = 0.1$)

Experimental results: Subspace learning

Run time (seconds)

	data set				
	COIL	WBC	BCI	Ionos	G241N
Staged	272	73	45	28	290
Alternate	2352	324	227	112	2648
Global	106	8	25	61	94

(1/3 labeled, 2/3 unlabeled, squared loss, $\alpha^\star = 10$, $\beta = 0.1$)

Experimental results: Subspace learning

Transductive generalization error

	data set				
	COIL	WBC	BCI	Ionos	G241N
Staged	0.476	0.200	0.452	0.335	0.484
Alternate	0.464	0.388	0.440	0.457	0.478
Global	0.388	0.134	0.380	0.243	0.380
(Lee et al. 2009)	0.414	0.168	0.436	0.350	0.452
(Goldberg et al. 2010)	0.484	0.288	0.540	0.338	0.524

(1/3 labeled, 2/3 unlabeled, squared loss, $\alpha^\star = 10$, $\beta = 0.1$)

Conclusion

Conclusion

Global training can be more efficient than local training

Alternation is inherently slow to converge

Global training simplifies practical application

- no under-training
- only need to guard against over-fitting
- can use standard regularization techniques

Future work

- Learning structured or factored latent representations
- ICA
- Deep learning