Matrix Sparse Coding

John Lafferty Min Xu

Carnegie Mellon University
University of Chicago

Outline

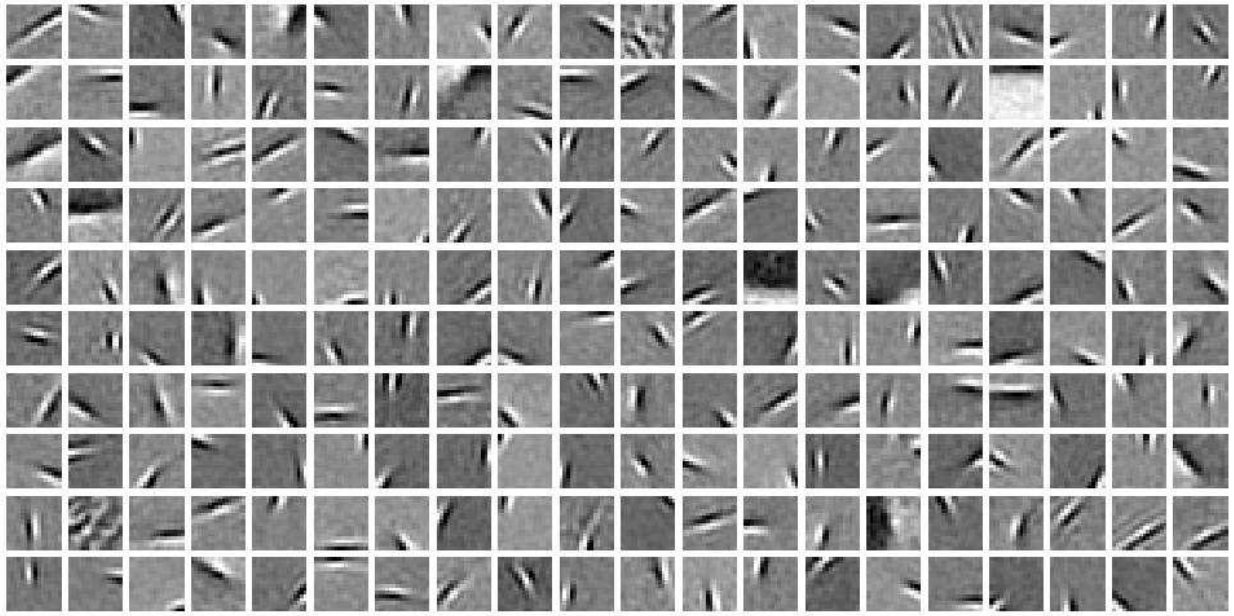
- Background on sparse coding
- Sparse coding for multivariate regression
- Algorithm & analysis
- Extensions

Sparse Coding

Motivation: understand neural coding (Olshausen and Field, 1996).

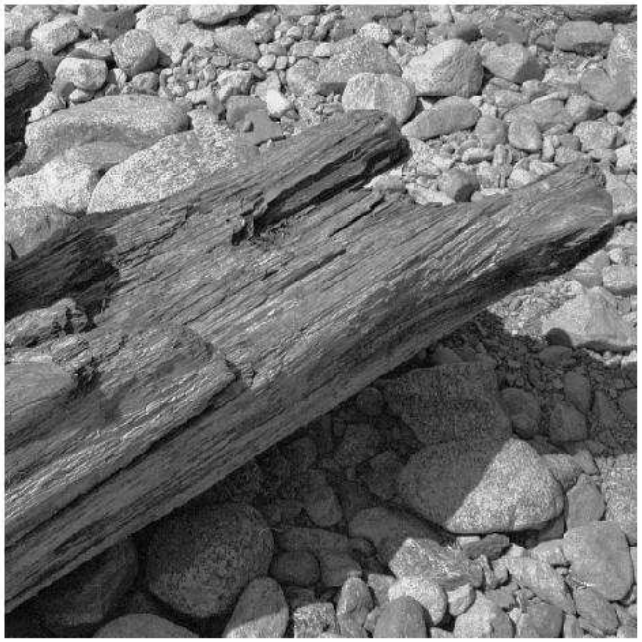
$$\min_{\alpha, X} \quad \sum_{g=1}^{G} \left\{ \frac{1}{2n} \left\| y^{(g)} - X\alpha^{(g)} \right\|_2^2 + \lambda \left\| \alpha^{(g)} \right\|_1 \right\}$$

such that $\quad \|X_j\|_2 \leq 1$



Sparse Coding for Natural Images

original image



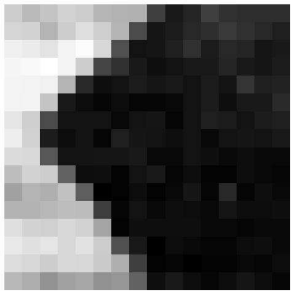
sparse representation

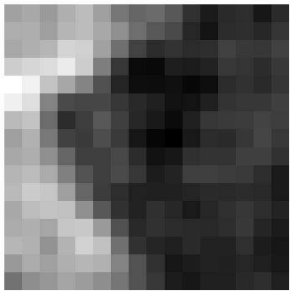
Codewords/patch 8.14, RSS 0.1894

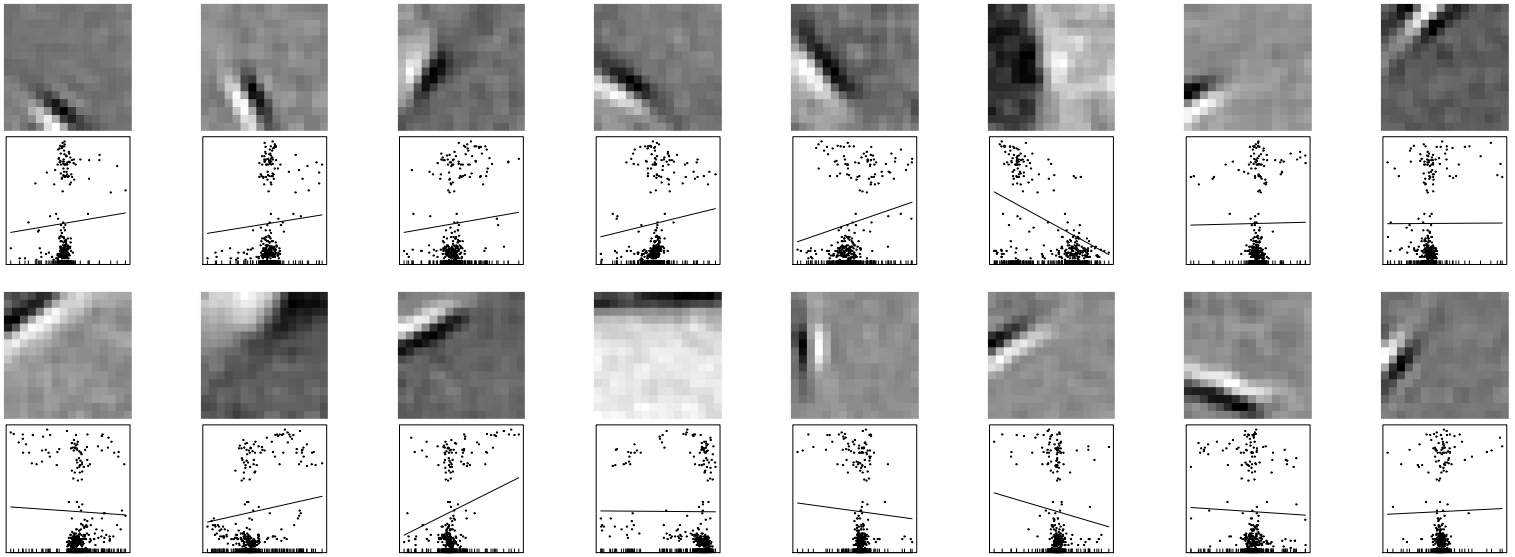
Sparse Coding for Natural Images

Original patch

Reconstruction
RSS $= 0.0906$





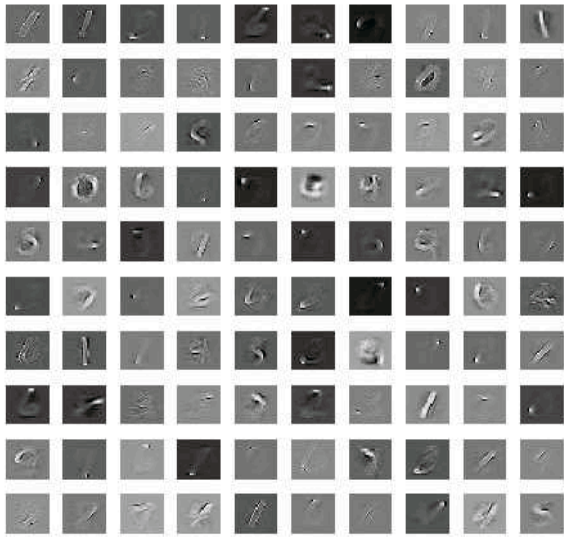


Properties

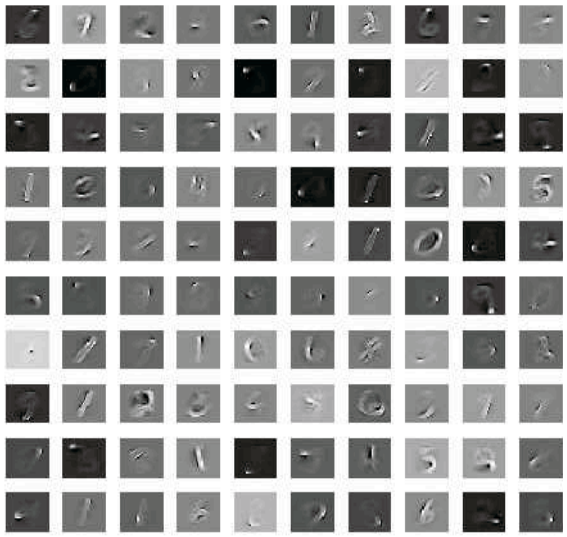
- Provides high dimensional, nonlinear representation
- Sparsity enables codewords to specialize, isolate particular “features”
- Overcomplete basis, adapted to data automatically
- Frequentist form of topic modeling, soft VQ

Sparse Coding for Computer Vision

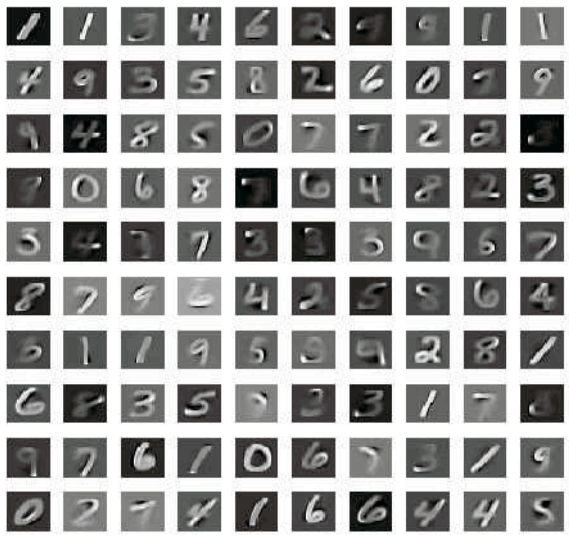
(source: Kai Yu)



Error: 4.54%



Error: 3.75%



Error: **2.64%**

- Best accuracy when learned codewords are like digits
- Advanced versions are state-of-art for object classification

Our Work

- Extend intuition of sparse coding to multivariate regression with grouped data
- Show how matrices (or vectors) sharing common structure are easier to approximate, when dictionary has similar structure
- Analyze rates of convergence under different formulations of “common structure”

Problem Formulation

- Data fall into G groups, indexed by $g = 1, \ldots, G$
- Covariate $X_i^{(g)} \in \mathbb{R}^p$ and response $Y_i^{(g)} \in \mathbb{R}^q$, model

$$Y_i^{(g)} = B^{*(g)} X_i^{(g)} + \epsilon_i^{(g)}$$

$$\epsilon^{(g)} \sim \mathcal{N}(0, \sigma^2 I_q)$$

- Goal: estimate $B^{*(g)} \in \mathbb{R}^{q \times p}$ with

$$\widehat{B}^{(g)} = \sum_{k=1}^{K} \widehat{\alpha}_k^{(g)} D_k$$

where each D_k is low rank, $\widehat{\alpha}^{(g)} = (\widehat{\alpha}_1^{(g)}, \ldots, \widehat{\alpha}_K^{(g)})$ is sparse

Conditional Sparse Coding

- Objective function:

$$f(\alpha, D) = \frac{1}{G} \sum_{g=1}^{G} \left\{ \frac{1}{n} \left\| Y^{(g)} - \left(\sum_{k=1}^{K} \alpha_k^{(g)} D_k \right) X^{(g)} \right\|_F^2 + \lambda \|\alpha^{(g)}\|_1 \right\}$$

  minimized over $D_k \in \mathcal{C}(\tau)$,

$$\mathcal{C}(\tau) = \left\{ D \in \mathbb{R}^{q \times p} : \|D\|_* \leq \tau \text{ and } \|D\|_2 \leq 1 \right\}$$

- Dictionary entries D_k are shared across groups; nuclear norm constraint forces them to be low rank

Conditional Sparse Coding

Input: Data $\{(Y^{(g)}, X^{(g)}\}_{g=1,...,G}$, parameters λ and τ

1. Initialize dictionary $\{D_1, ..., D_K\}$ as random rank one matrices
2. Alternate following steps until convergence of $f(\alpha, D)$:
   a. Encoding step: $\{\alpha^{(g)}\} \leftarrow \arg\min_{\alpha^{(g)}} f(\alpha, D)$
   b. Learning step: $\{D_k\} \leftarrow \arg\min_{D_k \in \mathcal{C}(\tau)} f(\alpha, D)$

$$f(\alpha, D) = \frac{1}{G} \sum_{g=1}^{G} \left\{ \frac{1}{n} \left\| Y^{(g)} - \left(\sum_{k=1}^{K} \alpha_k^{(g)} D_k \right) X^{(g)} \right\|_F^2 + \lambda \|\alpha^{(g)}\|_1 \right\}$$

Summary of Results

With $B^{0(g)}$ as optimal s-sparse dictionary approximation of $B^{*(g)}$,

$$\|B^{*(g)} - \widehat{B}^{(g)}\|_F \leq \underbrace{\|B^{*(g)} - B^{0(g)}\|_F}_{\text{approximation error}} + \underbrace{\|B^{0(g)} - \widehat{B}^{(g)}\|_F}_{\text{estimation error}}$$

Assumptions on $\{B^{*(g)}\}_{g=1}^{G}$	Approximation error $\|B^{*(g)} - B^{0(g)}\|_F$	Total error $\|B^{*(g)} - \widehat{B}^{(g)}\|_F$
No structure	$\left(\frac{1}{K}\log(GKs)\right)^{s/pq}$	$\sqrt{\frac{pq\log(npqG)}{n}}$
Low rank	$r\left(\frac{1}{K}\log(GKs)\right)^{s/r(p+q)}$	$\sqrt{\frac{r(p+q)\log(nr(p+q)G)}{n}}$
Union of Subspaces	$\gamma\left(\frac{\Gamma^{\gamma_D}}{K}\log(GKs)\right)^{s/\gamma}$	$\sqrt{\frac{\gamma\gamma_D\log(n\Gamma G\gamma)}{n}}$

Related Work

- Low-rank regression: Yuan et al. (2007), Negahban and Wainwright (2011)
- Task-driven sparse coding: Mairal, Bach and Ponce (2010)
- Multi-task learning: Evgeniou and Pontil (2004), Maurer and Pontil (2010)
- Sparse representation: Jeong and Kim (2009)

Analysis Overview: Low Rank Case

- Assumptions: Each $B^{*(g)}$ is of rank at most r. Dictionary elements $D_1, \ldots, D_K$ are random rank one matrices.

- Suppose that $s < \min\{r(p+q), rK/2\}$. Then with probability at least $1 - 1/K$,

$$\max_g \|B^{*(g)} - B^{0(g)}\|_F \leq c_1 r \left(\frac{\log(GKs)}{c_2 K} \right)^{s/r(p+q)}$$

Analysis Overview: Union of Subspaces

- Assumptions: Each $B^{*(g)}$ is a γ-sparse combination of rank one generator matrices $A_1, \ldots, A_\Gamma$. Dictionary elements $D_1, \ldots, D_K$ are random γ_D-sparse combinations of $\{A_k\}$.

- Suppose that $K > 4\Gamma^{\gamma_D}(\log\Gamma + \log G\gamma)$ and $s < \gamma$. Then with probability at least $1 - 1/K - 1/\Gamma$,

$$\max_g \|B^{*(g)} - B^{0(g)}\|_F \leq c_1\gamma\left(\frac{\Gamma^{\gamma_D}\log(GKs)}{c_2K}\right)^{s/\gamma}$$

Estimation Error

- Under restricted eigenvalue assumptions, if the regularization level is chosen to be $\lambda_n = c_3\sqrt{\sigma^2 \frac{\log(GK)}{n}}$, then the lasso solution satisfies, with probability at least $1 - 1/K$,

$$\max_g \|\widehat{B}^{(g)} - B^{0(g)}\|_F \leq c_1\sqrt{\frac{\sigma^2 s \log(GK)}{n}}$$

- Due to sparsity, estimation error grows only with $\log K$.

Total Error

- Consider as parameters the sparsity level s and dictionary size K; adjusted to get best possible rates

- Low rank case: For sufficiently large K,

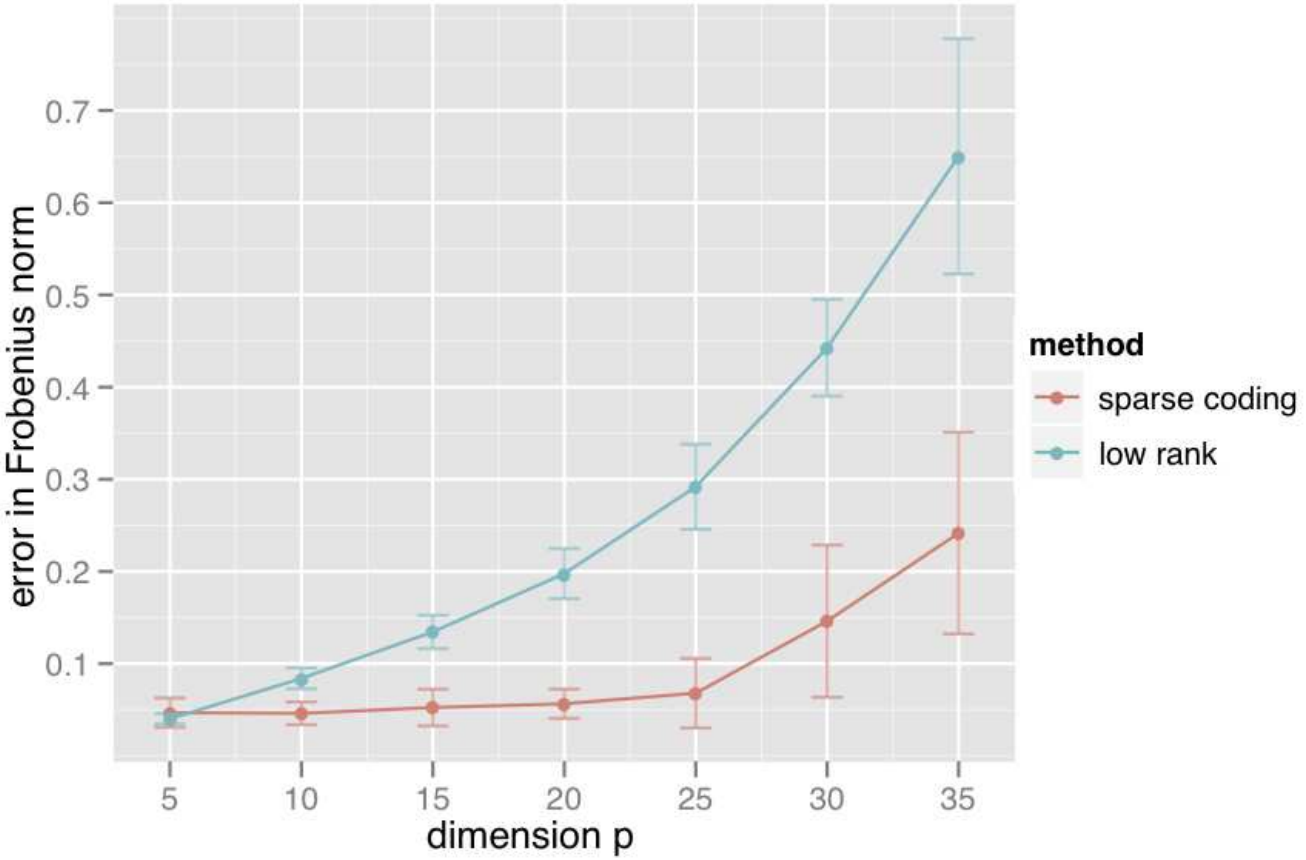
$$\max_g \|B^{*(g)} - \widehat{B}^{(g)}\|_F = O_P\left(\sigma\sqrt{\frac{r(p+q)\log(\sigma^2 nr(p+q)G)}{n}}\right)$$

- Union of subspaces case:

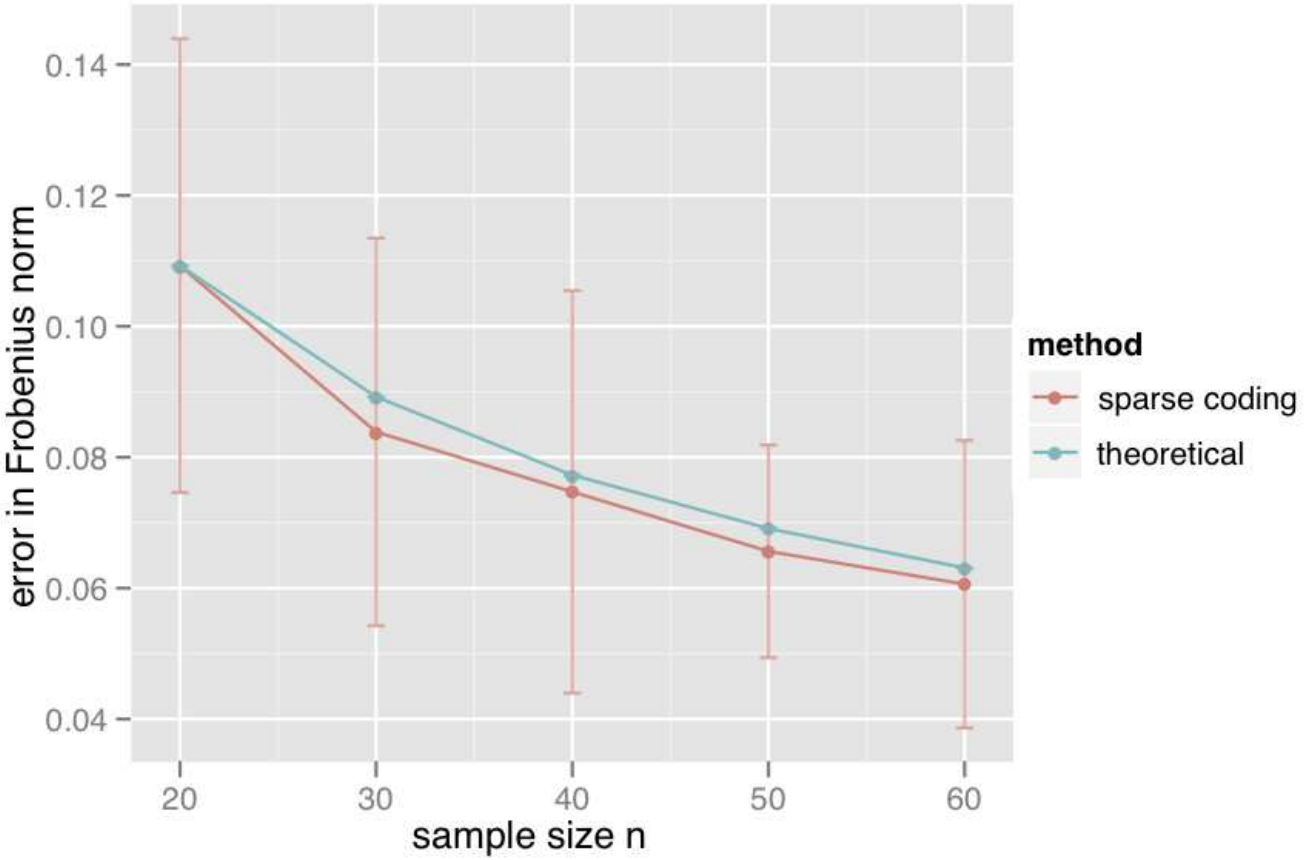
$$\max_g \|B^{*(g)} - \widehat{B}^{(g)}\|_F = O_P\left(\sigma\sqrt{\frac{\gamma\gamma_D\log(n\Gamma G\gamma)}{n}}\right)$$

(ignoring extra $\log$ factors)

Experiments: Union of Subspaces



Experiments: Union of Subspaces



Experiments with Equities Data

- 29 companies in single industry sector, from 2002 to 2007
- One day log returns, $Y_t = \log S_t/S_{t-1}$, X_t lagged values
- Grouped in 35 day periods

	30 days back	50 days back	90 days back	Sparse Coding
Correlation	-0.000433	0.0527	0.0513	**0.0795**
Predictive R^2	-0.0231	-0.0011	0.00218	**0.0042**

Covariance Coding

- Sparse code the group sample covariance matrices

$$\widehat{S}_n^{(g)} = \frac{1}{n}\sum_{i=1}^{n} Y_i^{(g)} Y_i^{(g)T}$$

- Objective function:

$$f(\alpha, \beta, D) = \frac{1}{G}\sum_{g=1}^{G}\left\{\frac{1}{n}\left\|\widehat{S}_n^{(g)} - \mathrm{diag}(\beta) - \sum_{k=1}^{K}\alpha_k^{(g)} D_k\right\|_F^2 + \lambda\|\alpha^{(g)}\|_1\right\}$$

minimized over $D_k \in \mathcal{C}(\tau)$,

$$\mathcal{C}(\tau) = \{D \succeq 0, \|D\|_* \leq \tau \text{ and } \|D\|_2 \leq 1\}$$

- Optimization over $\alpha^{(g)}$ by solving semidefinite program or nonnegative lasso

Summary

- Sparse coding for multivariate regression
- Analysis of idealized procedure under different assumptions on shared structure
- Open: Analysis of alternating minimization algorithm
- Sparse coding techniques are promising for a wide range of statistical problems