Beyond Filters: Compact Feature Map for Portable Deep Model

ICML 2017

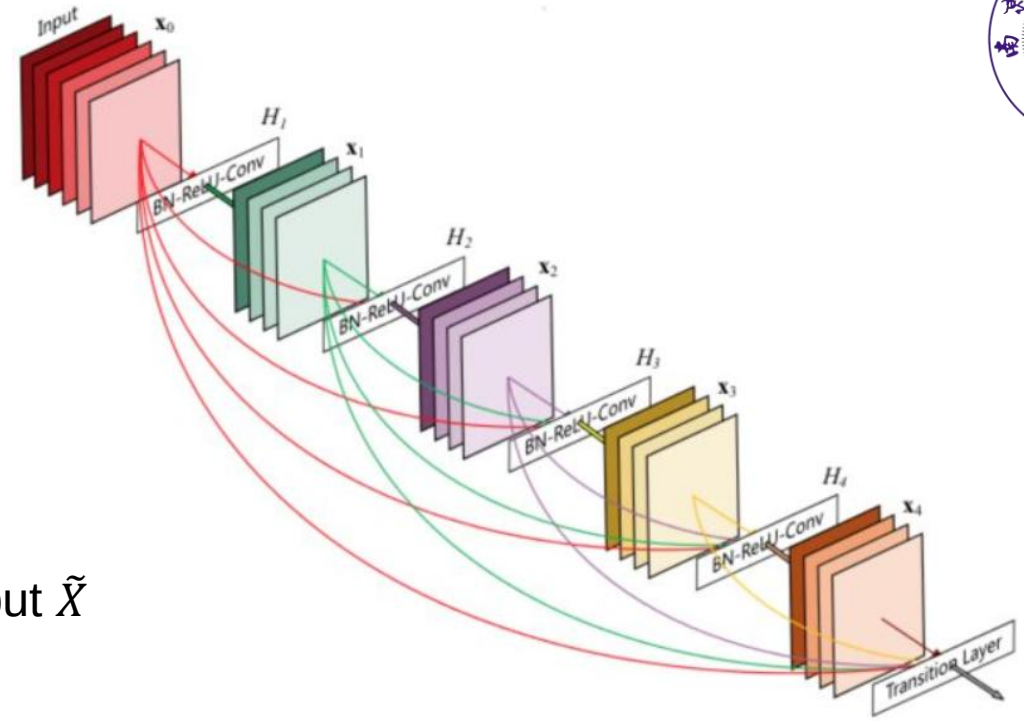
2018.6.20

Feature Map Y

Dimensionality Reduction

New Feature Map $\tilde{Y}$

Input $\tilde{X}$



CNN Layer Reconstruction

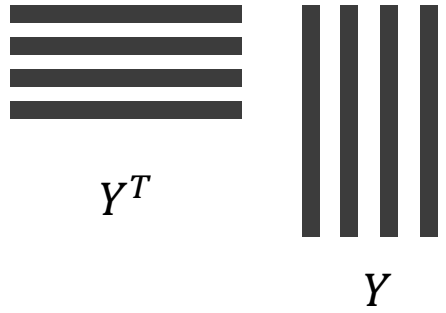
Save memory/computation resource

Feature Map Dimensionality Reduction

The most compact representation of a CNN should have no correlation between feature maps derived from different convolution filters.

$$\Theta(\mathbf{Y}) = ||\mathbf{Y}^T\mathbf{Y} \circ (\mathbf{1} - \mathbf{I})||_F^2 \qquad \min_{P,c} ||P\mathbf{Y}^T\mathbf{Y}P^T - C||_F^2, s.t. \quad C = \mathrm{diag}(c)$$

Preserve the distances between feature maps



$$\min_{P,c} ||P\mathbf{Y}^T\mathbf{Y}P^T - C||_F^2 + \lambda||\mathcal{D}(\tilde{\mathbf{Y}}) - \mathcal{D}(\mathbf{Y})||$$

$$s.t. \quad \tilde{\mathbf{Y}} = \mathbf{Y}P^T, \;\; C = \mathrm{diag}(c),$$

D_{ij} is the Euclidean distance between the i column and the j column of Y

Feature Map Dimensionality Reduction

$$\min_{P,c} ||P\mathbf{Y}^T\mathbf{Y}P^T - C||_F^2 + \lambda||\mathcal{D}(\tilde{\mathbf{Y}}) - \mathcal{D}(\mathbf{Y})||$$
$$s.t. \quad \tilde{\mathbf{Y}} = \mathbf{Y}P^T, \quad C = \mathrm{diag}(c),$$

$$\Longleftrightarrow$$

$$\min_{P} ||P\mathbf{Y}^T\mathbf{Y}P^T - C||_F^2,$$
$$s.t. \quad C = \mathrm{diag}(c), \quad PP^T = \mathbf{I}.$$

$$||y_1P^T||_2 = ||y_1||_2 \qquad ||y_1P^T - y_2P^T||_2^2 = ||y_1 - y_2||_2^2$$

$\Longrightarrow$ Dimensionality has not been reduced since P is a square matrix

$$\min_{\tilde{\mathbf{Y}}} \frac{1}{2}||\tilde{\mathbf{Y}} - \mathbf{Y}P^T||_F^2 + \lambda||\tilde{\mathbf{Y}}||_{2,1} \longrightarrow \min_{P,c} ||P\mathbf{Y}^T\mathbf{Y}P^T - C||_F^2 + \beta||c||_1$$
$$s.t. \quad C = \mathrm{diag}(c), \quad PP^T = \mathbf{I},$$

$$\min_{P,c} ||P\mathbf{Y}^T\mathbf{Y}P^T - C||_F^2 + \beta||c||_1$$
$$s.t. \quad C = \text{diag}(c), \quad PP^T = \mathbf{I},$$

$$\min_{p,c} ||P\mathbf{Y}^T\mathbf{Y}P^T - C||_F^2 + \alpha||PP^T - \mathbf{I}||_F^2 + \beta||c||_1$$
$$s.t. \quad P = \text{circ}(p), \quad C = \text{diag}(c),$$

$$\text{circ}(p) := \begin{bmatrix} p_1 & p_d & \cdots & p_3 & p_2 \\ p_2 & p_1 & p_d & & p_3 \\ \vdots & p_2 & p_1 & \ddots & \vdots \\ & & \ddots & \ddots & \\ p_{d-1} & & \ddots & \ddots & p_d \\ p_d & p_{d-1} & \cdots & p_2 & p_1 \end{bmatrix}$$

Early Active Learning via Robust Representation and Structured Sparsity

IJCAI 2013

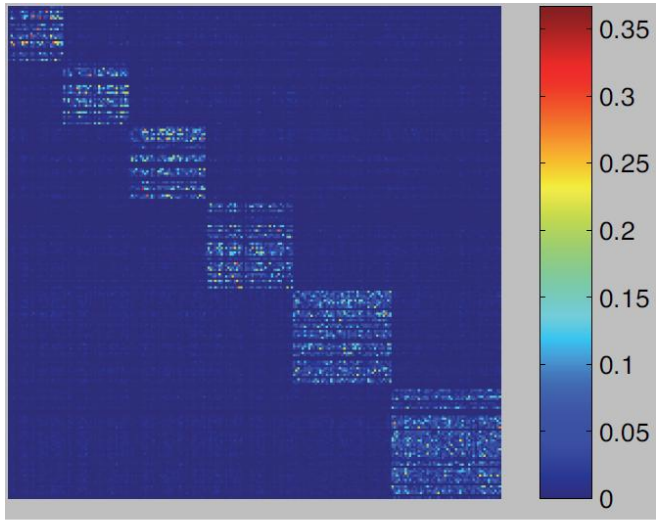
$$\min_{A,B} \sum_{i=1}^{n} (\|x_i - Ba_i\|_2^2 + \gamma \|a_i\|_2^2) \qquad s.t. \quad A = [a_1, \dots, a_n] \in R^{m \times n}, B \subset X, |B| = m$$

$$\min_{A} \sum_{i=1}^{n} (\|x_i - Xa_i\|_2^2 + \gamma \|a_i\|_2^2) \qquad s.t. \quad A = [a_1, \dots, a_n] \in R^{n \times n}, \qquad \|A\|_{2,0} = m$$

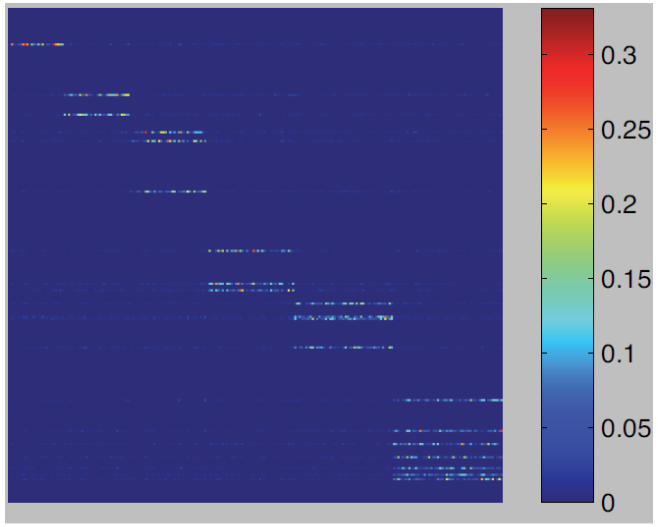
$$\min_{A} \sum_{i=1}^{n} (\|x_i - Xa_i\|_2^2) + \gamma \|A\|_{2,1} \qquad \text{sensitive to data outliers}$$

$$J = \min_{A} \ \|(X - XA)^T\|_{2,1} + \gamma \|A\|_{2,1}$$

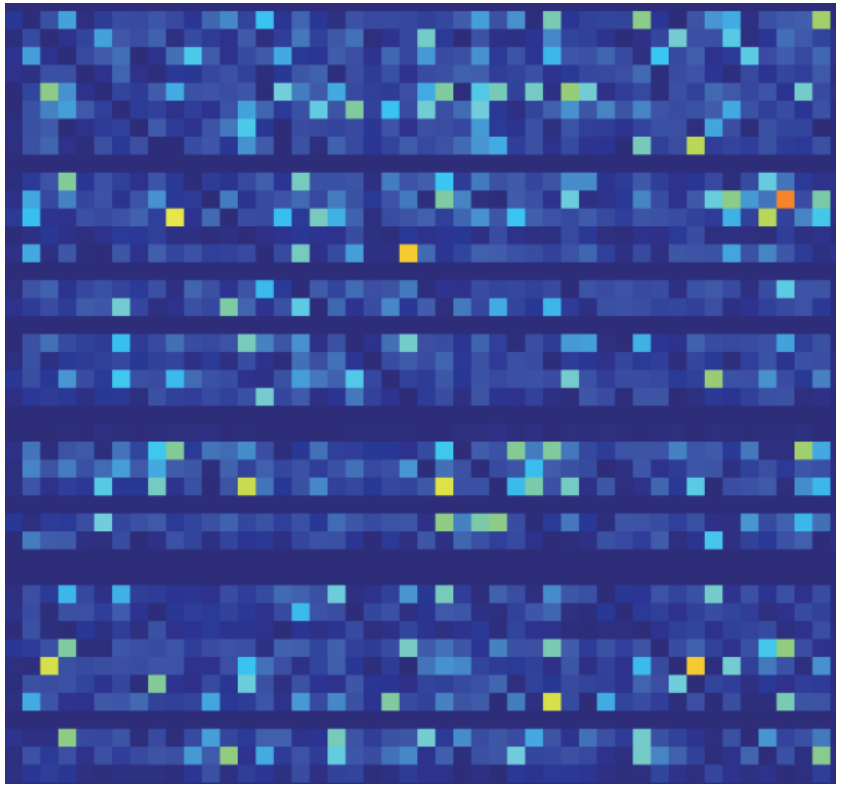
Sort the rows of A by **the row-sum values of the absolute** A in the **decreasing order**. Therefore, the active learning task can be performed **by selecting the m samples** corresponding to the **top m rows** of A.



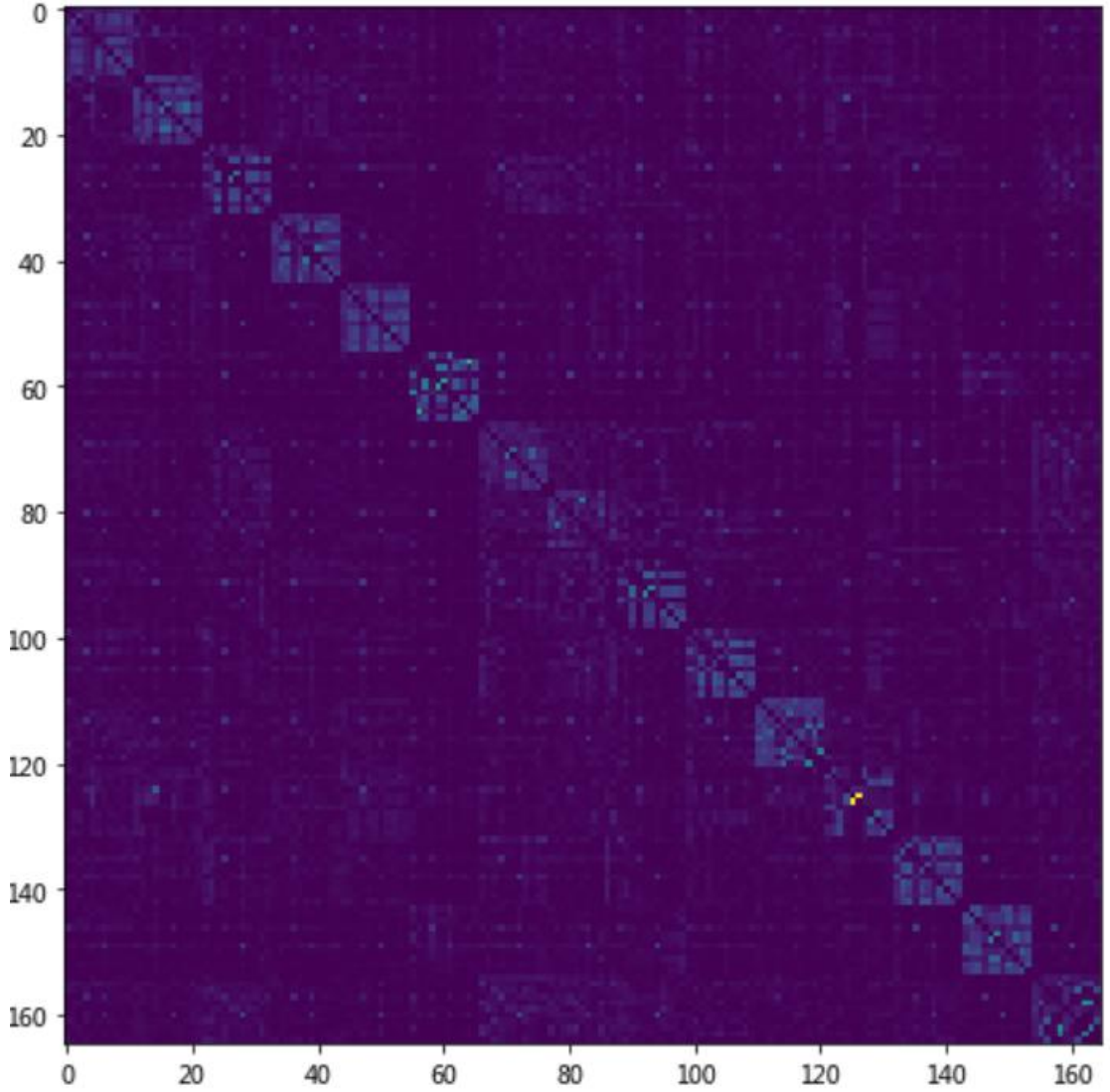
(b) Learned matrix A

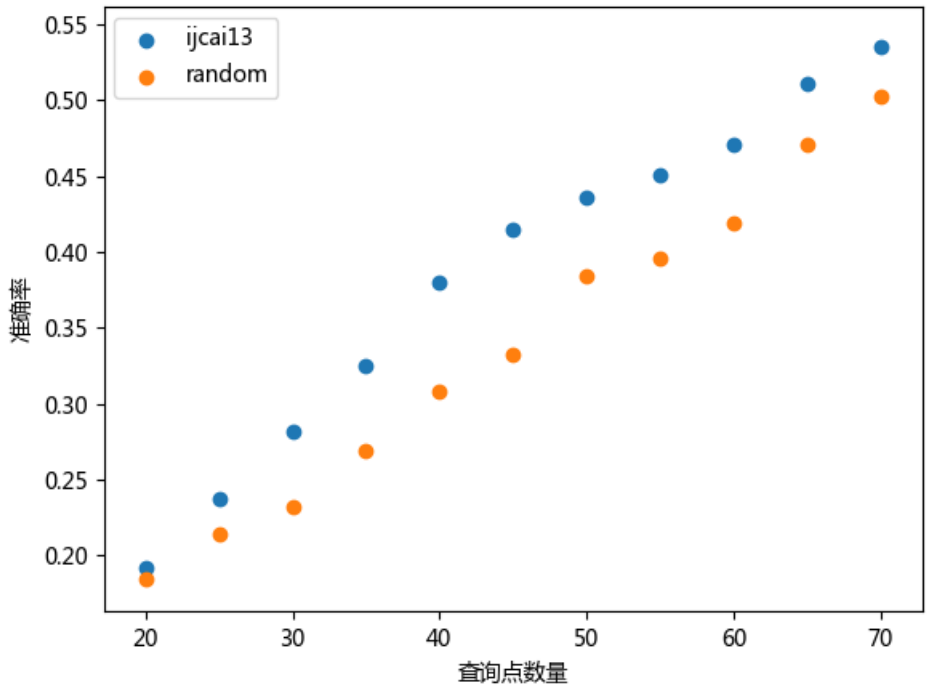


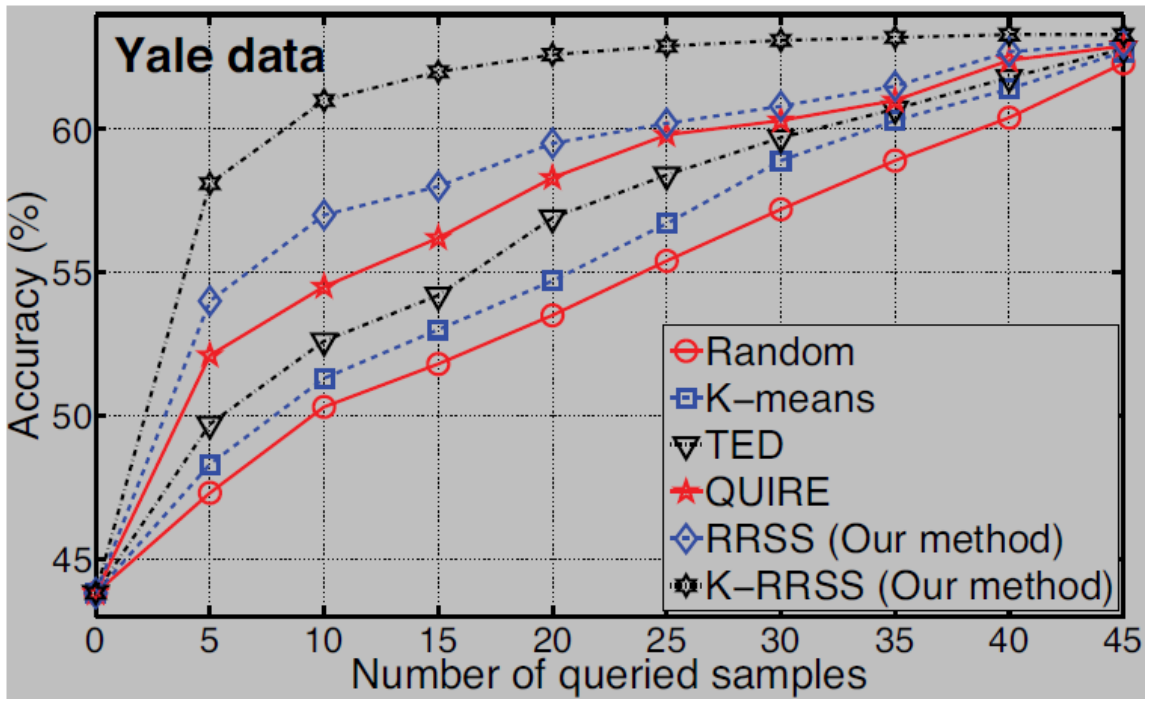
(c) The top 20 rows of A



yalefaces



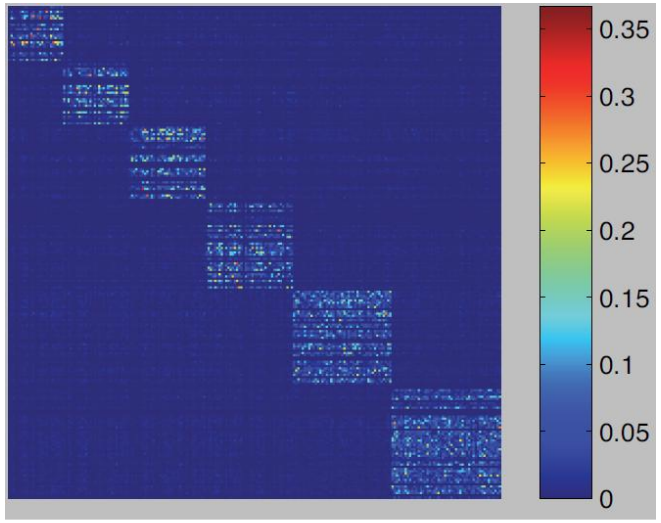

ijcai13
random
准确率
查询点数量
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
20
30
40
50
60
70


Yale data
Accuracy (%)
Number of queried samples
Random
K–means
TED
QUIRE
RRSS (Our method)
K–RRSS (Our method)
60
55
50
45
0
5
10
15
20
25
30
35
40
45

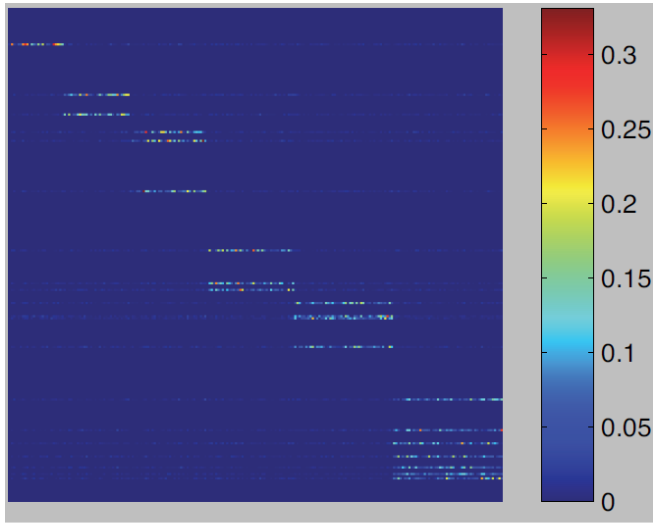
$$\theta(Y) = \|Y^T Y \circ (\mathbf{1} - I)\|_F^2$$

$$\theta(A) = \|AA^T \circ (\mathbf{1} - I)\|_F^2$$

$$= \|AA^T - C\|_F^2$$



(b) Learned matrix A



(c) The top 20 rows of A

$$J = \min_A \ \|(X - XA)^T\|_{2,1} + \gamma\|A\|_{2,1}$$

$$\frac{\partial J}{\partial A} = X^T XAU - X^T XU + \gamma VA = 0$$

$$J = \min_{A,c} \ \|(X - XA)^T\|_{2,1} + \gamma\|AA^T - C\|_F^2 + \lambda\|c\|_1$$

$$J = \min_{A,C} \ \|(X - XA)^T\|_{2,1} + \gamma\|AA^T - C\|_F^2 + \lambda\|c\|_1$$

$Tr(A^T A) = \|A\|_F^2$

$$\|AA^T - C\|_F^2 = Tr\left((AA^T - C)^T(AA^T - C)\right) = Tr((AA^T - C)(AA^T - C))$$

$$= Tr(AA^T AA^T - AA^T C - CAA^T - C^2) = Tr(AA^T AA^T) - Tr(AA^T C + CAA^T)$$

$Tr(ABC) = Tr(BCA) = Tr(CAB)$

$$= Tr(AA^T AA^T) - 2Tr(CAA^T)$$

$$\frac{\partial}{\partial X} Tr(BXX^T) = BX + B^T X$$

$$\frac{\partial}{\partial A} Tr(CAA^T) = 2CA$$

1. 矩阵函数$U=F(X),V=G(X),W=H(X)$乘积的微分矩阵为:

$$d(UVW)=(dU)VW+U(dV)+UV(dW)$$

2. 矩阵转置的微分矩阵等于原矩阵的微分矩阵的转置　　$d(X^T)=d(X)^T$

3. 矩阵的迹的矩阵微分等于矩阵微分的迹　　$d(Tr(X))=Tr(dX)$

$$d\left(Tr(AA^TAA^T)\right)=Tr\left(d(AA^TAA^T)\right)$$

$$Tr(ABC)=Tr(BCA)=Tr(CAB)$$

$$=Tr(d(A)A^TAA^T)+Tr(A(dA)^TAA^T)+Tr(AA^T(dA)A^T)+Tr(AA^TA(dA)^T)$$

$$=2Tr\left(A^TAA^Td(A)\right)+2Tr(AA^TA(dA)^T)$$

$$Tr(A^TB) = Tr(B^TA)$$

$$Tr(AA^TA(dA)^T) = Tr\big((dA)(AA^TA)^T\big) = Tr\big((dA)(A^TAA^T)\big) = Tr\big((A^TAA^T)(dA)\big)$$

$$2Tr\big(A^TAA^Td(A)\big) + 2Tr(AA^TA(dA)^T) = 4Tr\big(A^TAA^Td(A)\big)$$

$dTr(X^TX) = Tr(2X^TdX)$　　由命题3.2.1直接得$Tr(X^TX)$关于X的梯度矩阵为

$$\frac{\partial Tr(X^TX)}{\partial X} = (2X^T)^T = 2X$$

$$\frac{\partial Tr\big(A^TAA^Td(A)\big)}{\partial A} = AA^TA$$

$$J = \min_{A} \ \|(X - XA)^T\|_{2,1} + \gamma\|A\|_{2,1} \qquad X \in R^{d\times n}$$

$$\frac{\partial J}{\partial A} = X^T XAU - X^T XU + \gamma VA = 0$$

$$U: u_{ii} = \frac{1}{2\|x_i - Xa_i\|_2}, \qquad V: v_{ii} = \frac{1}{2\|a_i\|_2}$$

$$u_{ii}X^T Xa_i - u_{ii}X^T x_i + \gamma V a_i = 0$$

$$a_i = u_{ii}(u_{ii}X^T X + \gamma V)^{-1}X^T x_i$$

Input: The data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$.
Initialize $\mathbf{A} \in \mathbb{R}^{n\times n}$;
while *not converge* **do**
  1. Calculate the diagonal matrix $\mathbf{U}$, where the i-th diagonal element of $\mathbf{U}$ is $u_{ii} = \frac{1}{2\|\mathbf{x}_i - \mathbf{X}\mathbf{a}_i\|_2}$.
  Calculate the diagonal matrix $\mathbf{V}$, where the i-th diagonal element of $\mathbf{V}$ is $v_{ii} = \frac{1}{2\|\mathbf{a}^i\|_2}$. ;
  2. For each $i(1 \le i \le n)$, update $\mathbf{a}_i$ by $\mathbf{a}_i = u_{ii}(u_{ii}\mathbf{X}^T\mathbf{X} + \gamma\mathbf{V})^{-1}\mathbf{X}^T\mathbf{x}_i$;
end
Output: The matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.

$$J = \min_{A} \ \|(X - XA)^T\|_{2,1} + \gamma \|A\|_{2,1} \qquad X \in R^{d \times n}$$

$$\frac{\partial J}{\partial A} = X^T XAU - X^T XU + \gamma VA = 0 \qquad U: u_{ii} = \frac{1}{2\|x_i - Xa_i\|_2}, \qquad V: v_{ii} = \frac{1}{2\|a_i\|_2}$$

4.2 An Efficient Algorithm to Solve the Constrained Problem

The Lagrangian function of the problem in Eq. (14) is

$$\mathcal{L}(\mathbf{U}) = \|\mathbf{U}\|_{2,1} - Tr(\mathbf{\Lambda}^T(\mathbf{AU} - \mathbf{Y})).$$

NIPS 2010

Taking the derivative of $\mathcal{L}(\mathbf{U})$ w.r.t $\mathbf{U}$, and setting the derivative to zero, we have:

$$\frac{\partial \mathcal{L}(\mathbf{U})}{\partial \mathbf{U}} = 2\mathbf{DU} - \mathbf{A}^T\mathbf{\Lambda} = \mathbf{0},$$

where $\mathbf{D}$ is a diagonal matrix with the i-th diagonal element as[1]

$$d_{ii} = \frac{1}{2\,\|\mathbf{u}^i\|_2}.$$

$$J = \min_{A,c} \ \|(X - XA)^T\|_{2,1} + \gamma\|AA^T - C\|_F^2 + \lambda\|c\|_1$$

$$\frac{\partial J}{\partial A} = X^T XAU - X^T XU + 4(AA^T A - \gamma CA) = 0 \qquad U: u_{ii} = \frac{1}{2\|x_i - Xa_i\|_2}$$

For each i compute:

u_{ii}

For each i compute:

$a_i = A[:, i]$

(update M)

update M

compute :

$\min_{A,c} \gamma\|AA^T - C\|_F^2 + \lambda\|c\|_1$

$$u_{ii}X^T Xa_i - u_{ii}X^T x_i + 4\gamma(M - C)a_i = 0$$

$M = AA^T$

$$a_i = u_{ii}[u_{ii}X^T X + 4\gamma(M - C)]^{-1}X^T x_i$$

$M = AA^T A$ $\quad u_{ii}X^T Xa_i - u_{ii}X^T x_i + 4\gamma m_i - 4\gamma Ca_i = 0$

$$a_i = [u_{ii}X^T X - 4\gamma C]^{-1}[u_{ii}X^T x_i - 4\gamma m_i]$$

$$\min_{A,c} \gamma \|AA^T - C\|_F^2 + \lambda \|c\|_1$$

A

A^T

对于式(11.12), 可先计算 $\boldsymbol{z} = \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k)$, 然后求解

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 . \tag{11.13}$$

令 x^i 表示 $\boldsymbol{x}$ 的第 i 个分量, 将式(11.13)按分量展开可看出, 其中不存在 $x^i x^j$ $(i \neq j)$ 这样的项, 即 $\boldsymbol{x}$ 的各分量互不影响, 于是式(11.13)有闭式解

$$x_{k+1}^i = \begin{cases} z^i - \lambda/L, & \lambda/L < z^i ; \\ 0, & |z^i| \leqslant \lambda/L ; \\ z^i + \lambda/L, & z^i < -\lambda/L , \end{cases} \tag{11.14}$$

其中 x_{k+1}^i 与 z^i 分别是 $\boldsymbol{x}_{k+1}$ 与 $\boldsymbol{z}$ 的第 i 个分量. 因此, 通过 PGD 能使 LASSO 和其他基于 L_1 范数最小化的方法得以快速求解.

1. Fix the others and update J by setting $J = \arg\min_J \frac{1}{\mu}||J||_* + \frac{1}{2}||J - (Z + Y_2/\mu)||_F^2$.

2. Fix the others and update S by setting $S = \arg\min_S \frac{1}{\mu}||S||_* + \frac{1}{2}||S - (L + Y_3/\mu)||_F^2$.

3. Fix the others and update Z by setting $Z = (\mathtt{I} + X^TX)^{-1}(X^T(X - LX - E) + J + (X^TY_1 - Y_2)/\mu)$.

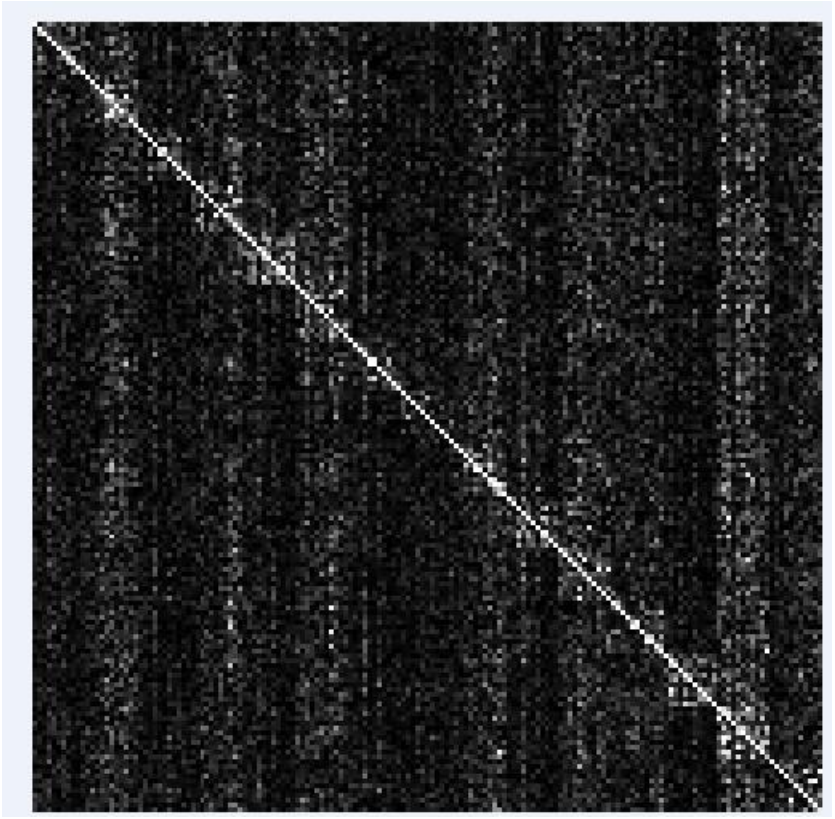
4. Fix the others and update L by setting $L = ((X - XZ - E)X^T + S + (Y_1X^T - Y_3)/\mu)(\mathtt{I} + XX^T)^{-1}$.

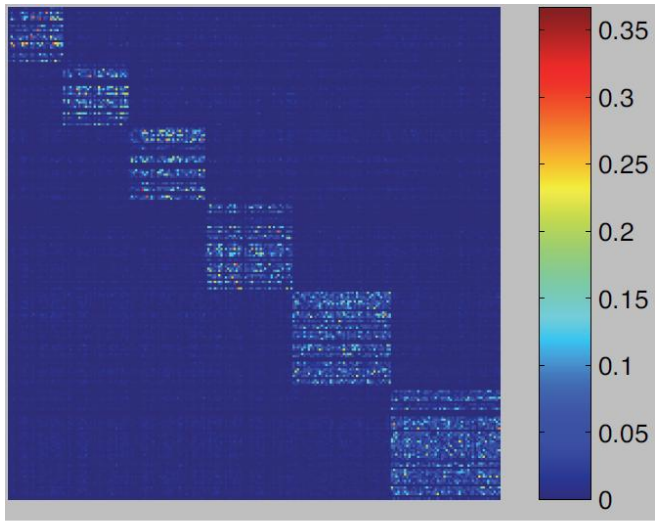
5. Fix the others and update E by setting $E = \arg\min_E \lambda/\mu||E||_1 + 0.5||E - (X - XZ - LX + Y_1)/\mu||_F^2$.

6. Update the multipliers by $Y_1 = Y_1 + \mu(X - XZ - LX - E)$, $Y_2 = Y_2 + \mu(Z - J)$, $Y_3 = Y_3 + \mu(L - S)$.

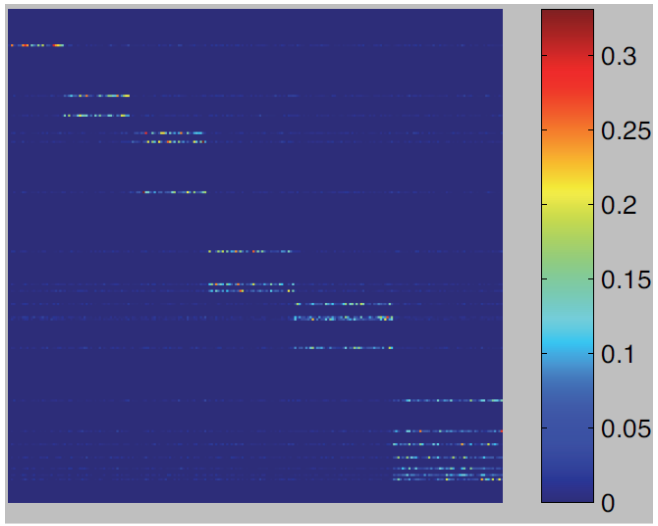
实验

不收敛





(b) Learned matrix A



(c) The top 20 rows of A

$$\underset{i}{\operatorname{argmax}} \left\{ \|a_i\|_1 - \alpha * \max_j \langle a_i, a_i^s \rangle \right\}$$