# 特征噪声下的非贪婪度量学习

## 引言

线性度量学习方法常旨在学习一个马氏度量矩阵刻画两个样本间的距离。由于矩阵是半正定的(positive semi-definite，PSD)，所以一些方法常将分解为，转而求解一个投影矩阵刻画样本距离。在非满秩的情况下，求解投影矩阵时的问题本质与降维(dimensionality reduction)相同 。因此很多诸如流形学习(manifold learning)一类的降维方法与度量学习有着较深的联系 。事实上，在高维空间中，学习一个刻画数据本征低维结构的度量，会使得算法的识别效果更好 。传统方法常采用L2距离损失作为样本损失，以刻画类内和类间散度，该类思想在样本观测分布准确无偏的假设前提下进行建模。但由于实际应用场景存在不可避免的噪声，L2损失加剧了噪声下的干扰程度，使得该类算法性能下降。

如何针对噪声环境构建一个鲁棒性的度量，是提升该情形下分类识别效果的关键 。纵观现有方法的出发点，传统方法本质上假设类内和类间样本服从两个不同高斯分布，由此该类方法的模型损失可等价为L2损失 。为了降低噪声的干扰尺度、避免L2损失加剧观测分布的有偏程度，一部分工作转而采用L1损失进行建模 。该类方法本质是假设类内和类间样本服从两个不同的拉普拉斯分布 。相对L2损失，L1损失的尺度较小，该类损失确实可以从一定程度上降低噪声的影响程度。因此，很多学者基于L1损失提出了一系列卓有成效的方法。例如，非监督情况下，PCA-L1 采用L1损失刻画总体样本方差，提升了算法的抗噪性。有监督情况下，基于L1损失的线性判别分析(linear discriminant analysis based on L1 loss, LDA-L1) 方法和Wang等人提出的度量学习方法 ，均分别利用L1损失统一刻画类内散度和类间散度，推导了一种逐向量的贪婪求解算法，取得了一定的效果；为了弥补贪婪算法的缺陷，Liu等人 在LDA-L1的基础上，推导了一种逐矩阵的非贪婪求解算法，提升了LDA-L1模型的性能。

然而，大部分上述方法存在两方面的问题。一方面，噪声使得观测距离偏大或偏小，对类内和类间样本优化方向具有不同的影响。统一采用L1损失只降低了距离尺度，并未分别考虑两类损失优化方向的差异性，不能从根本上提升度量的判别性。因此，考虑到这一点，扩大类内样本损失尺度、减小类间样本损失尺度，两者同时加大对两类样本的惩罚程度，可提升特征噪声下算法的判别性。与此同时，也有相关研究表明，类内样本和类间样本应考虑不同的损失 。另一方面，现有采用L1损失的方法大部分采用逐向量的求解方式，未能实现关于矩阵变量的整体优化，求解效果不够理想。为了改善LDA-L1模型的求解问题，Liu等人 推导了一种非贪婪的求解算法，但该方法是涉及到矩阵的梯度迭代，时间复杂度较高。

为此，基于边缘费歇尔分析方法 ，本章提出了一种采用L2/L1损失的非贪婪的鲁棒性度量学习方法。该方法模型在最小化类内样本L2损失的同时最大化类间样本L1损失，比基于L2损失的方法赋予类间样本更大的惩罚，比基于L1损失的方法赋予类内样本更大的惩罚，该模型可同时提升度量的判别性。由于模型的目标函数非凸使得求解较为困难，本章随后将目标函数转为迭代优化两个凸函数之差；受凸函数差算法 (difference of convex functions algorithm, DCA) 思想启发，本章设计了一个辅助函数，由此推导了一种非贪婪的迭代求解算法；随后所给的理论证明保证了算法的收敛性。在公开数据集上，仿真实验对同类方法进行了不同噪声程度和噪声类型下的对比，结果表明了所提的MFA-L2/L1方法的优越性与鲁棒性。该工作的主要贡献如下：

* 提出了一个基于L2/L1的边缘费歇尔分析模型，该模型利用L2损失刻画类内散度、L1损失刻画类间散度，解决噪声对类内和类间样本优化方向具有不同影响的问题，可同时提升度量的判别性；
* 设计并推导了一种非贪婪的迭代求解算法，在所构建的目标辅助函数下，解决了模型非凸的求解问题，该算法无需进行逐向量求解，具有较快的收敛性；
* 对迭代求解算法进行了理论分析，证明了其收敛性；
* 基于机器学习领域的公开数据集进行了特征噪声下的仿真实验，所提方法在对比实验中排名第一，结果验证了所提MFA-L2/L1方法的鲁棒性和有效性。

## 相关工作

假设给定训练样本集，其中，，表示第个含有维特征的训练样本。对任意样本对，有或，其中，是正样本对集合（即相似样本对集合），是负样本对集合（即不相似样本对集合）。记待学习的投影矩阵为，表示其投影向量的个数。首先，这一节主要介绍与本章所提方法最相关的两个工作，基于L1损失的LDA-L1和基于L2损失的边缘费歇尔判别分析。

### 基于L1损失的线性判别分析方法

为了降低噪声对线性判别分析（linear discriminant analysis, LDA）的影响，一系列方法采用尺度较小的L1损失来改进LDA ，这类方法致力于在优化目标函数下学习一个投影矩阵：

其中，表示总体样本均值向量，表示第类样本均值向量；表示第类的样本数；表示总类别数，意味着样本属于第类；保证了投影矩阵的正交性。不同于传统LDA，该类目标函数同时涉及L1范数的最大化和最小化，无法采用特征向量分解进行求解，使得模型求解较为困难。针对此类模型的求解问题，LDA-L1采用逐向量优化的贪婪求解方法，该优化过程性能相对逐矩阵的优化算法略差 。针对这一问题，Liu等人近来提出了一种针对L1范数LDA的非贪婪算法(a non-greedy algorithm for L1-norm LDA, LDA-NgL1) ，针对同一模型 [[eq:L1LDA]](#eq:L1LDA)，该算法对矩阵进行整体求解，提升了模型的性能。但该方法涉及矩阵的梯度迭代，计算复杂度较高。

### 边缘费歇尔分析方法

由于真实应用较为复杂，数据分布不一定服从LDA关于高斯分布的假设。为了克服这一局限性，样本无需满足高斯分布，边缘费歇尔分析(marginal fisher analysis, MFA) 根据近邻样本的相似性重新定义了类内散度与类间散度，该方法取得了一定的成功和较为广泛的应用。基于样本近邻关系，该方法采用L2距离作为损失来定义类内散度和类间散度，在优化目标下学习投影矩阵：

此处，为便于理解不同模型之间的差异性，区别于MFA原始模型的矩阵表达方式，模型 [[eq:MFA]](#eq:MFA)采用样本对形式表示目标函数。其中，是正样本对或负样本对之差，行向量是关于列向量的转置。是图所连接的相似样本对集，而图是刻画类内散度的本质图(intrinsic graph)，即表示正样本对集；同理，是图所连接的不相似样本对集，而图是刻画类间散度的惩罚图(penalty graph)，即表示负样本对集。对于两个图的连接矩阵和，其定义分别如下：

$$\label{eq:MFA\_Ls}
\boldsymbol{A}\_{\mathcal{S}}(i,j)=\left\{
\begin{aligned}
&1,&&{\text{若}i \in N^{+}\_{k\_{\mathcal{S}}}(j)\text{或} j \in N^{+}\_{k\_{\mathcal{S}}}(i)};\\
&0,&&{\text{其它}}.
\end{aligned}
\right.\\$$

$$\label{eq:MFA\_Lb}
\boldsymbol{A}\_{\mathcal{D}}(i,j)=\left\{
\begin{aligned}
&1,&&{\text{若}i \in N^{-}\_{k\_{\mathcal{D}}}(j) \text{或}j \in N^{-}\_{k\_{\mathcal{D}}}(i)};\\
&0,&&{\text{其它}}.
\end{aligned}
\right.\\$$

其中，表示与第个样本具有相同标签的个近邻的样本集合，表示与其具有不同标签的个近邻的样本集合。该目标函数可直接通过特征向量分解实现求解。但由于MFA模型采用了L2距离作为损失，其性能易受噪声的影响。

## L2/L1损失的边缘费歇尔分析算法

基于MFA方法，针对噪声下观测分布的有偏性，本小节基于对噪声的影响分析提出特征噪声下的度量学习模型，对于模型求解的困难性提出了一种非贪婪的求解算法，并对算法收敛性进行了证明分析。

### 特征噪声下的度量学习模型

由于特征噪声下样本的观测分布偏离真实分布，因此在该情形下样本的观测距离相对真实情况可能会偏大或偏小，使得噪声对类内和类间样本优化方向具有不同的影响。结合图 [1.1](#fig:NsyPoint)进行具体分析，对类内散度而言，偏小的观测距离可能使得模型对正样本对的惩罚程度不足。因此，相对L1损失，采用尺度较大的L2损失刻画类内散度，加大对正样本对的惩罚程度更为合适。相反，对类间散度而言，观测距离偏大可能使得模型对负样本对的惩罚程度不足，采用尺度较小的L1损失刻画类间散度，加大对负样本对的惩罚程度更为合适。

特征噪声下的观测分布与真实分布对比图

特征噪声下的观测分布与真实分布对比图

考虑MFA方法无需数据满足高斯分布的前提优势，本章提出一种基于L2/L1损失的边缘费歇尔分析(marginal fisher analysis based on L2/L1 loss, MFA-L2/L1)模型：

与MFA的目标函数 [[eq:MFA]](#eq:MFA)类似，表示基于近邻关系的正样本对集或负样本对集的样本差向量；表示对矩阵取迹的算子；矩阵的行向量为所有来自正样本对集的行向量；矩阵的行向量为所有来自负样本对集的行向量；需要说明和区分的是，为了表示方便，本章参考文献 的表示方式，对任意大小为矩阵,定义其范数$\footnote{本章中该形式矩阵范数均只采用这一种定义，其形式虽与矩阵的列和范数形式相同，但计算方式并不相同，请注意区分。}$ 。

与LDA-L1、MFA两个方法对比，MFA-L2/L1方法的目标函数不仅非凸，且其同时涉及L2范数最小化和L1范数最大化。因此，LDA-L1、LDA-NgL1和MFA的求解算法都不将再适合本章所提的MFA-L2/L1模型。为了解决该模型的求解问题，下一小节将作进一步分析。

### 非贪婪的求解算法

分析MFA-L2/L1模型，其目标函数是关于两个凸函数的商形式，且这两个刻画散度的凸函数值始终非零。因此，受文献 中的定理1和定理2启发，为使得目标函数 [[eq:MFA21]](#eq:MFA21)下降，本章设计了一个关于和的求解形式，转为迭代两个凸函数之差：

在公式 [[eq:RMFA]](#eq:RMFA)中，，，是第次迭代时关于投影矩阵的目标函数值：

为了更为简洁的表达，用函数表示，则公式 [[eq:RMFA]](#eq:RMFA)的目标函数可记为：

分析公式 [[eq:RMFA\_itr2]](#eq:RMFA_itr2)，目标函数是两个凸函数和之差，在公式中是常数，不影响第次迭代时的推导分析。由于公式 [[eq:RMFA\_itr2]](#eq:RMFA_itr2)的目标函数具有与DCA思想 相同的差形式，受该思想启发，公式 [[eq:RMFA\_itr2]](#eq:RMFA_itr2)在第次迭代求解时，可生成两个变量序列，即中间变量序列和求解变量序列（为了与外部的交替迭代次序区分开来，上标表示在固定第次迭代时采用DCA思想求解的第次内部迭代）：

其中，是在处的次梯度；则可看作是在处的一阶展开式。记，由此，对于序列 [[eq:RMFA\_SEQ]](#eq:RMFA_SEQ)有如下定理：

[theory:theory\_1] 序列 [[eq:RMFA\_SEQ]](#eq:RMFA_SEQ)中的使得成立。

*Proof.* 由于是在处的次梯度，所以在邻域内对任意有: 成立。则：

$$\label{ineq::INEQ12}
\begin{aligned}
F(\boldsymbol{W}^{k+1})&=G(\boldsymbol{W}^{k+1})-\lambda\_{t-1} L(\boldsymbol{W}^{k+1})\notag\\
&\leq G(\boldsymbol{W}^{k+1})-\lambda\_{t-1} [ L(\boldsymbol{W}^k)+H(\boldsymbol{W^{k+1}},\boldsymbol{W}^k)]\notag\\
&\leq G(\boldsymbol{W}^{k})-\lambda\_{t-1} [ L(\boldsymbol{W}^k)+H(\boldsymbol{W^k},\boldsymbol{W}^k)]\notag\\
&=G(\boldsymbol{W}^k)-\lambda\_{t-1} L(\boldsymbol{W}^k)=F(\boldsymbol{W}^k).
\end{aligned}$$

 ◻

根据定理 [[theory:theory\_1]](#theory:theory_1)可知，由序列 [[eq:RMFA\_SEQ]](#eq:RMFA_SEQ)可得到关于公式 [[eq:RMFA\_itr2]](#eq:RMFA_itr2)中目标函数的下降序列，因而由序列 [[eq:RMFA\_SEQ]](#eq:RMFA_SEQ)可推导出具体迭代为：

其中，$\emph{sign}(\cdot)$是关于矩阵逐元素的函数：

$$\begin{split}
\emph{sign}(x)=\left\{
\begin{aligned}
1, &\text{若} \quad x \geq 0,\\
-1, &\text{其它}.
\end{aligned}
\right.\\
\end{split}$$

值得注意的是，求解时，在序列 [[eq:SqObj2]](#eq:SqObj2)中关于的常数项部分已忽略。将序列 [[eq:SqObj2]](#eq:SqObj2)的目标函数关于矩阵求导，并令导数为零矩阵，得到：

解之得：

按关于的序列 [[eq:SqObj1]](#eq:SqObj1)和关于的迭代公式 [[eq:SoluW]](#eq:SoluW)进行计算，直至序列 [[eq:SqObj2]](#eq:SqObj2)迭代停止。该迭代过程所得到的就是第次迭代所求的，进一步可由公式 [[eq:RMFA\_itr1]](#eq:RMFA_itr1)的原理计算。

为保证矩阵的正交性约束，该问题可通过正交强迫一致问题(orthogonal procrustes problem) 来解决，即：。该问题的解为，其中，和分别是来自奇异值分解(Singular Value Decomposition, SVD)中大小为和的正交矩阵：，其对应的非零奇异值个数为。总结以上推导过程，MFA-L2/L1模型的具体求解步骤如算法 [[alg:MFA-L2/L1]](#alg:MFA-L2/L1)所示。

**输入:** , ,总内部迭代次数,外部总迭代次数,投影矩阵向量数; **初始化**. 更新 ; ; 计算$\boldsymbol{Y}^k = \boldsymbol{B}^\mathrm{T}\emph{sign}(\boldsymbol{B}\boldsymbol{W}^{k})$; 计算; ; 计算SVD分解： ; 更新; . **输出:**投影矩阵.

分析迭代流程，序列的作用是构造一个关于公式 [[eq:RMFA\_itr2]](#eq:RMFA_itr2)中目标函数的辅助函数，即序列 [[eq:SqObj2]](#eq:SqObj2)中关于的目标函数，以确保迭代过程中公式 [[eq:SqObj2]](#eq:SqObj2)的目标函数下降。整个算法始终对进行整体求解，属于非贪婪算法。由于以上分析仅限于在固定、求解时目标函数 [[eq:RMFA\_itr2]](#eq:RMFA_itr2)的单调性情况，下一小节将联合变量和分析算法的整体收敛性。

### 算法收敛性及计算开销分析

[theory:theory\_2] 算法 [[alg:MFA-L2/L1]](#alg:MFA-L2/L1)使得目标函数 [[eq:MFA21]](#eq:MFA21)下降。

*Proof.* 当固定第次外部迭代时，关于的梯度为零时有公式 [[eq:GradW]](#eq:GradW)成立；此外，由定理 [[theory:theory\_1]](#theory:theory_1)可知，使得目标函数下降，将拆分成关于投影向量形式的表达，用表示矩阵的第个行向量，则成立：

$$\begin{aligned}
\label{eq:covg1}
&\displaystyle \sum\_{l=1}^m \left(\boldsymbol{w}\_l\right)\_{t}^\mathrm{T}\boldsymbol{A}^\mathrm{T}\boldsymbol{A}\left(\boldsymbol{w}\_l\right)\_{t}-\lambda\_{t-1}\displaystyle \sum\_{l=1}^m \sum\_{i} \frac{\left(\boldsymbol{w}\_l\right)\_t^\mathrm{T}{\left( \boldsymbol{b}^i\right) }^\mathrm{T}\boldsymbol{b^i}\left(\boldsymbol{w}\_l\right)\_{t-1}}{|\boldsymbol{b}^i\left(\boldsymbol{w}\_l\right)\_{t-1}|}\leq \notag\\
&\displaystyle \sum\_{l=1}^m \left(\boldsymbol{w}\_l\right)\_{t-1}^\mathrm{T}\boldsymbol{A}^\mathrm{T}\boldsymbol{A} \left(\boldsymbol{w}\_l \right)\_{t-1}-\lambda\_{t-1}\displaystyle \sum\_{l=1}^m\sum\_{i} \frac{\left(\boldsymbol{w}\_l\right)^\mathrm{T}\_{t-1}{\left( \boldsymbol{b}^i\right) }^\mathrm{T}\boldsymbol{b}^i\left(\boldsymbol{w}\_l\right)\_{t-1}}{|\boldsymbol{b}^i\left(\boldsymbol{w}\_l\right)\_{t-1}|}.
\end{aligned}$$

对于矩阵的第个行向量，根据柯西-施瓦兹不等式，在第次迭代与第次迭代之间，对投影矩阵的第个投影列向量成立：

将矩阵的所有行向量关于以上形式的不等式变形，并进行叠加，则成立：

$$\begin{aligned}
\label{eq:covg3}
&\displaystyle \sum\_{i} \frac{\left(\boldsymbol{b}^i\left(\boldsymbol{w}\_l\right)\_t\right)^\mathrm{T}\boldsymbol{b}^i\left(\boldsymbol{w}\_l\right)\_{t-1}}{|\boldsymbol{b}^i\left(\boldsymbol{w}\_l \right)\_{t-1}|}-\displaystyle \sum\_{i} |\boldsymbol{b}^i\left(\boldsymbol{w}\_l\right)\_{t}|\leq \notag\\
&\displaystyle \sum\_{i} \frac{\left(\boldsymbol{b}^i\left(\boldsymbol{w}\_l\right)\_{t-1}\right)^\mathrm{T}\boldsymbol{b}^i\left(\boldsymbol{w}\_l\right)\_{t-1}}{|\boldsymbol{b}^i\left(\boldsymbol{w}\_l\right)\_{t-1}|}-\displaystyle \sum\_{i} |\boldsymbol{b}^i\left(\boldsymbol{w}\_l\right)\_{t-1}|=0.
\end{aligned}$$

将不等式 [[eq:covg3]](#eq:covg3)带入不等式 [[eq:covg1]](#eq:covg1)，得到:

$$\begin{aligned}
\label{eq:covg4}
&\displaystyle \sum\_{l=1}^m \left(\boldsymbol{w}\_l\right)\_{t}^\mathrm{T}\boldsymbol{A}^\mathrm{T}\boldsymbol{A}\left(\boldsymbol{w}\_l\right)\_{t}-\lambda\_{t-1}\displaystyle \sum\_{l=1}^m \| B\left(\boldsymbol{w}\_l\right)\_{t}\|\_1\leq \notag\\
&\displaystyle \sum\_{l=1}^m \left(\boldsymbol{w}\_l\right)\_{t-1}^\mathrm{T}\boldsymbol{A}^\mathrm{T}\boldsymbol{A}\left(\boldsymbol{w}\_l\right)\_{t-1}-\lambda\_{t-1}\displaystyle \sum\_{l=1}^m \| B\left(\boldsymbol{w}\_l\right)\_{t-1}\|\_1=0.
\end{aligned}$$

最后成立：

即目标 [[eq:MFA21]](#eq:MFA21)下降。 ◻

通过算法1可知，该算法的时间复杂度主要来自于外部循环步骤4、11、12与内部循环步骤7、8这五个步骤。若矩阵的类内样本对差向量个数为，矩阵的类内样本对差向量个数为，SVD分解中非零奇异值个数为，则五个步骤的计算开销与外部迭代次数、内部迭代次数、投影向量个数和空间维数有关。分析可知步骤4、7、8、11、12的时间复杂度分别为、、、、。因此，为该算法的时间复杂度，是内部循环次数时所对应的时间复杂度。根据实验观察，表明内部迭代仅需几次即可终止。因此，算法的时间复杂度可参考内部循环次数时所对应的时间复杂度。

## 实验分析

为了验证MFA-L2/L1方法的鲁棒性和解法的有效性，实验在5个UCI数据集和7个人脸数据集上添加了不同程度的特征噪声，与相关方法进行了对比。为保证实验的公正性，所有实验均从训练样本集中学习一个度量矩阵或投影矩阵，基于所学度量采用最近邻算法对测试样本进行分类。

### UCI数据集实验

为分析所提算法在不同噪声程度下的性能，实验分别在UCI数据集$\footnote{http://archive.ics.uci.edu/ml/datasets.php}$中选用Statlog (Heart)、Parkinsons 、Seeds、Sonar、Wine加入了比例为、、、左右的特征噪声，进行同类方法对比。以上UCI数据集信息见表 [[tab:chap2\_UCIdata]](#tab:chap2_UCIdata)。为避免特征尺度与噪声尺度不一致而影响评价的公正性，实验将数据集每个特征向量的模标准化为1，参考椒盐噪声的方式，实验按噪声比例的平方根进行样本数和特征数取整，在数据矩阵中随机选取对应样本数下对应特征数的矩阵元素，将其置为0或1作特征噪声。每个数据集按7：3的比例随机划分为训练集和测试集，该随机过程重复100次，报告平均准确率及方差作为实验结果。

实验选用LDA-L1 、LDA-NgL1 、MFA-NgL1、MFA、discriminative null space(DNS) 作为对比方法，与MFA-L2/L1进行比较。需要指出说明的是，MFA-NgL1是在MFA的样本关系下，采用LDA-NgL1相同的L1损失与求解算法。这些方法中，LDA-L1、LDA-NgL1、MFA-NgL1和MFA-L2/L1均属于迭代方法；MFA和DNS属于非迭代方法。LDA-L1的梯度步长设为0.01，LDA-NgL1和MFA-NgL1设置均遵循LDA-NgL1 原代码。经实验测试，固定矩阵的第次迭代，MFA-L2/L1序列内部迭代仅需一次即可收敛，因此最大次数设置为；LDA-NgL1和MFA-NgL1采用的解法属于矩阵梯度迭代法，需要较高的迭代次数，因此第次迭代时将内部迭代次数设置为100；当这三个迭代方法的迭代次数满足或迭代误差满足以下条件时，算法终止:

对于需要矩阵初始化的方法，均在同一随机种子下生成同一初始矩阵，该过程基于每次划分训练集和测试集时的不同随机种子生成。实验事先采用PCA计算贡献率的成分数，设置投影矩阵输出维度为。最终结果见图 [[fig:chap2\_UCI]](#fig:chap2_UCI)，算法排名结果见表 [[tab:chap2\_UCIRk]](#tab:chap2_UCIRk)。

根据图 [[fig:chap2\_UCI]](#fig:chap2_UCI)和表 [[tab:chap2\_UCIRk]](#tab:chap2_UCIRk)的的报告结果可知，在不同的噪声程度下，MFA-L2/L1相对其它同类方法更优，具有较好的准确性和鲁棒性。在Parkinson、Seeds和Wine三个数据集上，MFA-L2/L1在四种噪声程度下的性能均排名第一。观察Wine数据集上的结果，所提方法与次优方法LDA-NgL1的差异尤为明显，在至三个噪声程度下性能两者性能差异超过了。在图 [1.2](#fig:chap2_UCI1)中，所提方法虽然在Heart和Sonar数据集上性能略低于LDA-NgL1、MFA-NgL1和LDA-L1，但观察图 [1.3](#fig:chap2_UCI2)至图 [1.5](#fig:chap2_UCI4)可发现：随噪声程度增大，MFA-L2/L1在这两个数据集上逐渐胜过其它方法，其对于次优方法的优势逐渐增大。因此，基于本次实验中不同数据集、不同噪声程度的对比结果说明，随噪声程度增加，MFA-L2/L1表现了较好的抗噪性和识别性。

### AR数据集实验

鉴于MFA-L2/L1和MFA两者之间的联系，为了更全面地比较这两个方法的性能，该小节在AR数据集上进行实验对比。AR数据集 是人脸识别领域较为经典的数据集之一，选取该数据集裁剪后的2600张人脸图像进行实验，共有100人（男性、女性各50人）。实验中，每个人的图像属于一类，每类共含26张图像，具有光照、表情、是否佩戴眼镜或围巾等变化。

实验事先将所有图像转为50×40像素大小的灰度图像，随机从每个类别中选取6张图像，分别加入比例的高斯块和椒盐噪声两个类型噪声。随后从每个类别中随机选取13张图像作为训练样本，剩余13张图像作为测试样本。为避免实验结果的随机性，该过程重复10次后取平均结果进行报告。所有样本采用PCA保留贡献率作为前处理，其余实验设置与上一小节关于UCI数据集的实验设置相同。实验对比了投影矩阵维度在80维至225维之间的性能变化，将其结果报告如图 [[fig:AR]](#fig:AR)。

观察图 [[fig:AR]](#fig:AR)可知，在高斯块噪声情况下，MFA-L2/L1性能整体优于MFA。在椒盐噪声情况下，MFA的准确率随维度增加发生了微弱下降；MFA-L2/L1的准确率逐渐上升，超过MFA之后保持稳定。由相关工作中和方法介绍可知，两种方法均采用L2损失定义类内散度，两者的区别在于对类间散度定义不同。与MFA相比，MFA-L2/L1采用L1损失定义类间散度，提升了该方法的判别性。该实验结果表明，由于采用了L1损失对类间样本赋予更大的惩罚，MFA-L2/L1较MFA具有更好的性能。

### FEI数据集实验

由于MFA-L2/L1与LDA-NgL1同属非贪婪算法，且基于UCI数据集上都具有较好的排名与稳定性，该小节实验基于FEI 数据集对两者进行对比。该数据集同样为公开的人脸识别领域数据集，共包含200人（男性、女性各100人），每人采集了14张RGB图像，包含了不同角度、光照等14个场景变化，实验选取6个固定场景的1200张图像（均为正面人脸）进行算法比较。

由于前两小节实验已考察了不同噪声程度和噪声类型下的算法性能，本小节只考虑一种噪声类型。实验事先将1200张图像统一转为32×32像素大小的灰度图像，每个类别随机选取2张图像进行加噪。参考文献的方式，随机选取图像的像素点，将其从原像素值翻转至作像素噪声。每个类别随机选取3张干净图像用作测试样本，其余样本全用作训练，该过程重复10次。实验设置与前两小节保持一致，实验对比了投影矩阵维度在100维至250维之间的性能变化，其结果见图 [1.8](#fig:FEI)。

MFA-L2/L1与LDA-NgL1两个方法基于FEI数据随维度变化的准确率对比

MFA-L2/L1与LDA-NgL1两个方法基于FEI数据随维度变化的准确率对比

观察图 [1.8](#fig:FEI)，随维度增加，MFA-L2/L1和LDA-NgL1两个方法的性能差距逐渐变大。LDA-NgL1的性能随维度增加发生了一定的波动和下降；MFA-L2/L1在不同维度下的性能始终高于LDA-NgL1，且相对稳定。由相关工作和方法介绍可知，两种方法均采用L1损失刻画类间散度，两者区别主要在于对类内散度定义不同。与LDA-NgL1相比，MFA-L2/L1采用L2损失刻画类内散度以加大惩罚力度，所以具有更好的判别性。

### 人脸数据集综合实验

为了更全面比较MFA-L2/L1与相关方法的性能，该小节基于5个人脸数据集作了进一步的实验对比。除了LDA-NgL1、MFA-NgL1、LDA-L1、MFA和DNS以外，对比方法加入了经典的度量学习方法KISSME 。在这些方法中，DNS主要学习类内零空间作为度量矩阵；依赖于正样本对和负样本对，KISSME主要计算出两个协方差的逆矩阵之差作为度量矩阵。这两个方法的度量矩阵维数固定、无需设定。鉴于此，本小节均在固定维度下进行实验对比。实验设置仍与之前一致，对训练和测试过程重复10次，取实验平均结果作为报告。考虑上一小节实验未进行PCA处理，算法间并无较大性能差异，该节保留PCA中贡献率，对所有数据进行预处理。5个数据集的具体信息和加噪方式具体如下：

1) Senthil数据集$\footnote{http://www.geocities.ws/senthilirtt/Senthil\%20Face\%20Database\%20Version1}$： 该数据集由5个人的80张人脸图像构成，每个类别各有16张图像，所有图像均转为32×32像素大小的灰度图像。实验从每个类别中随机选取4张图像作为噪声样本，与上一节加噪方式类似，加入的像素噪声。为保证训练集和测试集各占一半，每个类别中随机选取6个干净样本、2个噪声样本用于训练，其余样本全用于测试。该数据集的投影矩阵输出维度设置为25。

2) Yale数据集$\footnote{http://vision.ucsd.edu/content/yale-face-database}$ ： 该数据集由15个人的165张人脸图像构成，每个类别各有11张图像，所有图像均转为32×32像素大小的灰度图像。实验从每个类别中随机选取3张图像作为噪声样本，加入面积比的高斯块噪声。为保证训练集和测试集大约各占一半，实验从每个类别中随机选取5张干净样本用于测试，其余样本全用于训练。该数据集的投影矩阵输出维度设置为40。

3) ORL数据集$\footnote{http://cam-orl.co.uk/facedatabase.html}$ ： 该数据集由40个人的400张人脸图像构成，每个类别各有10张图像，所有图像均转为33×28像素大小的灰度图像。实验从每个类别中随机选取3张图像作为噪声样本，加入面积比的高斯块噪声。为保证训练集和测试集各占一半，每个类别随机选取5张干净图像用于测试，其余全用于训练。该数据集的投影矩阵输出维度设置为35。

4) Caltech数据集$\footnote{http://www.vision.caltech.edu/html-files/archive.html}$ 该数据集由31个人的450张人脸图像构成，其中有5个类别均只含一张图像，其余每个类别各包含5到29张图像不等。实验先将其裁剪为193×162像素大小的灰度图像，后转为32×27像素大小。考虑该数据集类别的不平衡性，实验将样本数不超过5的类别全用作噪声样本，从其余每个类别中随机选取5张图像作噪声样本，与上一小节加噪方式类似，加入的像素噪声。实验将仅含一个样本的类别用于训练；对于其余每个类别，实验从中随机选取一半的样本用于训练，剩下一半样本用于测试。该数据集的投影矩阵输出维度设置为60。

5) UMIST数据集$\footnote{https://www.visioneng.org.uk/datasets/}$ ： 该数据集由20个人的575张人脸图像构成，每个类别包含19到48张图像不等。每幅图像转为32×32像素大小的灰度图像。对每个类别随机选取10张图像作为噪声样本，与上一小节加噪方式类似，加入的像素噪声。对每个类别随机选取大约一半样本数的干净样本作测试，其余样本均用作训练。该数据集的投影矩阵输出维度设置为25。

基于5个人脸数据集的实验结果见表 [[tab:chap2\_facedata]](#tab:chap2_facedata)，整体对比得知，在不同数据集的不同噪声情形下，MFA-L2/L1方法均取得了最优性能。在MFA-L2/L1、LDA-NgL1、MFA-NgL1和LDA-L1四个迭代算法中，LDA-L1的性能较差。这一现象的原因在于，该方法采用逐向量方式求解的贪婪解法，所以算法精确度不佳，该现象从另一层面上也反映了非贪婪算法的优势。

比较LDA-NgL1、MFA-NgL1和MFA三个方法，发现前两者并没有太大的优势。这一现象说明，在该实验噪声环境下，采用L1损失刻画类内散度和类间散度带来的提升并不明显。DNS方法在对比中性能较差，其原因在于每个数据集的线性零空间是有限的，导致其在该情况下的表达能力有限。由于KISSME方法计算的度量矩阵与原空间空大小一致，未能有效刻画数据潜在的低维结构，限制了其相似性的刻画能力，该缺点在文献中有所验证。综上分析表明，MFA-L2/L1通过利用不同的距离作为损失以刻画类内散度和类间散度，使得其在该本实验中优于其它算法，具有较好的鲁棒性和判别性。

### 算法运行时间分析

为了对算法收敛性情况进行对比，实验记录了Caltech人脸数据集上各算法的运行时间，具体报告见表格 [[tab:time\_rmfa]](#tab:time_rmfa)。从计算方式上分析，DNS和KISSME属于闭式求解算法，无需迭代，两者运行的时间显然最短。而其余四种方法均属于迭代方法，最大外部迭代次数均为100次，所运行时间相对非迭代方法更长。其中，采用同一种模型和求解方式的LDA-NgL1和MFA-NgL1所运行时间最长，虽然LDA-NgL1属于非贪婪式方法，但其求解方式涉及矩阵的梯度迭代计算，收敛过程缓慢。由于MFA-NgL1基于相似性生成的样本对，其训练样本数量实质远大于LDA-L1和LDA-NgL1中的样本数，所以其运行时间最长。而更为有趣的是，在同样的样本对数量下，所提方法MFA-L2/L1的运行时间低于非贪婪方法LDA-L1。该现象的原因在于，MFA-L2/L1虽属于非贪婪方法，但其内部迭代只需一次即可收敛，每步迭代实质是闭式求解，无需进行LDA-NgL1的梯度计算，该方法收敛性较快。综合分析，由于其简洁的求解过程，MFA-L2/L1所运行的时间较短，这从一定程度上也反映了算法具有较快的收敛性。

## 本章小结

针对特征噪声对度量学习的影响，本章提出了一种非贪婪度量学习方法MFA-L2/L1，随后推导了一种高效的求解算法。MFA-L2/L1方法采用L2损失和L1损失分别对应定义类内散度和类间散度，同时赋予两类样本更大的惩罚，以提升投影矩阵的判别性。针对目标函数非凸带来的求解困难，本章推导了一种非贪婪的迭代求解算法，并证明了算法的收敛性。在5个UCI数据集的不同噪声程度、AR和FEI数据集的不同维度、Senthil等5个人脸数据集上进行了大量仿真实验，结果分析表明MFA-L2/L1方法具有一定抗噪性和判别性。