# 20150906

想法：

分离热风炉换风扰动，分别进行PCA；

从波形中提取信息，要求：假设原始信号x服从正态分布，提取出的波形信息f(x)也服从正态分布。

# 20150908

**RBM基础：**

二值分布的能量函数



连续值高斯分布



将v通过参数a/b/w压缩成h

符合统计意义的v，压缩成h后再重构会比较精确；而不符合统计意义的v，压缩成h后，其重构很可能不精确。

# 20150909

想法：估计变量概率分布密度，对输入进行非线性变换，使之符合正态分布，

可能方向：压缩感知。

想法：通过矩阵分解得到反映系统状态的特征，包括线性变换（矩阵相乘）和非线性变换（非线性回归）

# 20150914

想法：通过分析主元的时序变化曲线来预测和诊断故障。

Deep的奇淫技巧：dropout、autocode、最后一次参数改为相乘

# 20150916

想法：考虑样本在得分区间T上的动态变化过程，即考虑考察在多维空间T上的曲线变化趋势。必须要考虑时间相似性和空间相似性以及变化趋势的相似性。

想法：由于变量值会随时间产生漂移，每次对当前输入u(t)减去前m时刻的输入均值（更多的考虑其变化，而不是状态）。

想法：既然分布只是发生了漂移，那把之前的估计的分布加上漂移的位移不就行啦？

# 20150918

如何分清一个变量随时间的变化是因为工作点发生了变化还是因为该变量本身就不稳定。

# 20150920

目标：跟踪高炉的动态变化过程

设：

训练集D1，测试集D2

两个数据集的均值和标准差分别为：m1、m2、s1、s2

D1经过PCA得到P1，T1

D2经过PCA得到P2，T2

即：

假设（待验证）：在无故障状态下，尽管高炉的工作的发生了迁移（m1/s1与m2/s2不等），但T1和T2同分布。

用训练集的P1计算：

则有：

的绝对值如下图所示



目标：跟踪统计量T的独立性

# 20150921

想法：

多PCA模型：

进行迭代PCA时，对于每个样本点x，若对于每个PCA模型的T方统计量或者SPE都超过了阈值，则新建一个PCA 模型；否则，选择SPE最小的PCA模型加入该样本点。

维护的模型样本点数为50~20000个

想法：

对数据进行类PCA的autoencoder，当一个节点的输出大于阈值时，则对其进行分裂。

模仿CNN里面的卷积，对时间序列进行卷积，最后池化。

# 20150922

通过小实验证明了改写的rbm的准确性，但是参数的调整对结果的影响较大，特别是隐层神经元的个数。

# 20150923

终于把sparse autoencoder改对了，程序在求梯度时w2忘了更新；

问题：

高炉的状态漂移比较严重（收到矿物种类、设定产量等影响）

富氧量等指标受到人工操作，与其他传感器变量不同。

# 20150925

经过一系列测试检验了模型性能，下一步用于正常状态的高炉

# 20151013

想法：将高炉的输入看做图像，用于cnn；为了避免工作点随时间的漂移，考虑先进行sfa，再作为cnn的输入。

想法：用dnn来做高炉的stage分段。

# 20151016

解决工作点漂移的必然可行性：

将工作点的漂移可以类比为图像的明暗变化，从而必然可以用cnn得到较好的压缩和诊断效果。

# 20151017

设计cnn结构：

输入大小：720\*26

卷积层：26\*26

池化层：10\*1

# 20151018

受到手写字体识别文献的启发，认识到故障的检测和分离也有同样的区别：

故障检测更倾向于估计输入的概率分布，根据概率分布来判断是否发生了故障；而故障分离则不需要知道输入的分布，而是需要给出不同输入下输出的概率分布。

印象里有篇论文讲到：如果数据集足够大，有监督的深度学习是不需要autoencoder作为pretraining也同样可以自动学习特征的，并避免陷入局部最优解。但是，在故障诊断领域，要提供足够多的故障样本是显然不可能的，因此使用autoencoder的pretraining就很重要。此外，由于样本不足，也容易导致过拟合，因此dropout、denoise等抑制过拟合的方法也是必须的。

# 20151024

三种思路：

1. 用正常数据建模，检验故障数据在该模型上的输出和重构误差。该模型对
2. 用正常数据和故障数据分别建模，考察不同类别数据在不同模型上的表现。
3. 用所有数据统一建模，考察不同类别数据在该模型上的输出和重构误差。

所以，首先得建立两种数据集：

1. 全部数据，可能不能一次性读取，只能分成文件储存在一个文件夹中
2. 按故障类别分的数据，储存在不同的文件夹中

只取输出量：2:5,11:15,22:26

考虑提取的特征：

最近一段时间内的均值、标准差

最近0~L、0~2L、……时间内的导数的均值、标准差

# 20151026

考虑最近的0~L、0~2L、0~4L、0~8L内的直线拟合和曲线拟合，用拟合参数和拟合误差作为特征。

想法：

以最近一段时间内的变量均值为输入，主成分分析的P为输出，训练dnn

或者输入对输出拟合。

# 20151029

想法：

有监督学习时，对于故障发生前的样本，距离发生故障越近的标签越接近于1

开题答辩：

故障样本稀少，无监督学习（自编码）的重要性

dnn：

auto-encoder：denoise、contractive

结构LSTM、RNN、CNN

信号的前期处理和特征工程：

PCA：Multy-model or multy-staged PCA

SFA：

# 20151106

LSTM的用法：直接给数据打标签，然后BP；2、将高炉的下一步输出作为标签，训练网络参数（pre-training），再将标签改为故障类别，并加上softmax层。

# 20151111

LSTM在语音识别领域的成功应用证明了其对时序数据的建模能力，其使用方法和建模过程都对LSTM在过程工业故障诊断领域的应用很有参考价值，但在进行具体的高炉异常炉况的诊断时，与语音识别还是有一些不同，主要有以下几点：

* 语音识别的一段时间序列中往往有很多不同的标签（比如一段话中有很多不同的字词），而高炉的一段时间序列中炉况的分类则比较统一（往往只有从正常炉况到某个异常炉况的标签序列）。
* 语音信号的样本远远多于异常炉况的样本，异常炉况的训练时需要更加注意避免过拟合，这就限制了对异常炉况建模的模型复杂度。
* 语音信号是一维信号；而高炉是多维信号，而且信号间的相互关系比较复杂，往往存在较大的误差和时延。
* 语音数据的均值和波形都比较稳定，数值准确，噪声较小；而高炉炉况数据中各个运行变量的准确度、信噪比、相对于炉况变化的时延等等都各不相同，而且没有明确固定的波形变化，再加上工作点的漂移，这些都给end-to-end的训练模型方式带来了完全不同的挑战。

综上所述，在研究基于深度学习的大型高炉异常炉况诊断时，既可以参考语音识别领域的诸多技术和优化方法，又需要针对过程工业和高炉的特点进行相应的调整和技术创新。

# 2015.11.15

熟练使用caffe、Theano、torch等编程框架。