第一个要讲的概念是生成模型和判别模型，概率学统计学在构建模型的过程中，一般有两种思路，一种是生成模型，一种是判别模型，假设观测值 o 和模型 q，如果对 P(o|q)建模，就是 Generative 模型，首先建立样本的概率密度模型，再利用模型推理预测，就如在 rbm 中，先建立了 Energy-Based model，（p27），然后训练参数使之接近输入数据真实分布，代码里面直接将 input 代入，作为 training distribution 真实分布，探讨的是样本如何产生，估计的是联合概率分布，反映同类数据本身的相似度。如果对 P(q|o)建模，就是 Discrminative 模型，建立的是预测模型，是条件概率密度，找到不同数据的最优分类面（斯坦福课上讲的），能够清晰的分辨多类之间区别，但是不能够把数据本身的特性找出来，像之前的 mlp 就是这样的。就如当给出一张手写体图片，生成模型关注的是各个像素点之间的相对关系，给出最后是哪个数字，而判别模型就是通过训练的参数，当给出 test 输入，它会根据分类成哪个数字。

第二个概念是 pca，主成分分析，主要的目的是将输入的高维数据映射到低维上面来，但是包含同样多的信息，针对一个手写体图片而言，图片大小为 28*28，那么它的维度就是 784，但是这 784 个维度不是每一个都有用，我们输出手写体的 input 数据时会发现有很多个 0，或者说很多空白地方都是无用的维度，例如四个角上的像素值，因此希望能够把数据变换到一个新的坐标系统中，这个坐标系统就是由主成分构成的。（画出图 1，2）首先有两种方式来解释 pca 的意义，我们采用第一种，最大化它们投影的方差，也就是找到一个方向使点在这个方向上投影最分散，那么就是说明数据在这个方向上变化最大。而这些方向，这些主成分我们可以看成特征值，就是一系列的这些特征值的线性组合构成了一张图片（例如人脸），因此得到了公式 1.5，I(x,y)都只是一个 patch，一个 patch 里面的像素点有 784 个，si 相当于 I（x，y）在 Wi 上的投影，作为输出。xy 是具体像素点的位置。特征数是我们直接假定就等于 pixel 个数。

$$I(x,y)=\sum_{i=1}^{n} A_i(x,y)s_i \quad 1.4$$

$$s_i=\sum_{(x,y)} W_i(x,y)I(x,y) \quad 1.5$$

1.4 则是这些特征值 W 的线性组合构成了 I（x，y），一般来说，主成分特征值的个数应该与输入的维数一致，因为相当于只是进行了坐标变换，但是我们只取前四分之一的主成分，因为这些成分就大概包括了所有需要改变的地方，所以只取了前 n 个，因此，我们求出了前 n 个 W，以及能够求出对应的 zi，而不能够求出所有的 si。故求 pca 的方式如下。

$$var(s)=E\{s^2\}-(E\{s\})^2 \quad 5.3$$

$$\|W\|=\sqrt{\sum_{x,y} W(x,y)^2}=1 \quad 5.4$$

$$\frac{1}{T}\sum_{i=1}^{T}\left(\sum_{x,y} W(x,y)I_t(x,y)\right)^2 \quad 5.5$$

公式 5.3 求的是主成分特征值，然后再计算得到主成分 si，求最大的方差，也就是公式 5.5，从公式 1.5 可以得出，假如不加限制，那么最大方差会在 W 趋近于无穷大时才达到，因为我们确定的只有 w 的方向，大小未定，因此加上限定条件为公式 5.4 等于 1，求第二个主成分，公式未变，限定条件会加上与之前的所有 w 都正交，这就是求主成分的方法。接下来就是白化过程（白化的理由暂时没弄懂，过程如此），让得到的主成分 si 方差为 1，因此用公式 5.16。以上就是关于 pca 的部分内容，由于我们在 ica 中只关注到了以上几点，即在 ica 中需要利用 pca 进行规范预处理。

$$z_i=\frac{y_i}{\sqrt{var(y_i)}} \quad 5.16$$

关于 ica，它也是一种生成模型。主要目的是将独立分布的量分离出来，较为典型的例子是假设有两个声音源波形是不一样的，ica 的功能是将两个声音源波形分离出来（自己画一下），在我们的例子中，手写体识别或者说图像识别，变量 si 就是分布独立，因此我们要分离的便是这些 si。回到模型，模型将映射得到的 si 当成了隐藏的独立变量，其中的 I(x,y)只是一个 patch，因此 A 这个特征值对于每一个 patch 而言都是相同的，但是 si 却在 patch 之间不相同，因为每个 patch 最后的表现形式是不同的。同时 ica 基于三个假设：1.si 是统计独立的随机变量 2.si 的分布是非高斯的，因为数据是稀疏分布的，而稀疏分布不是高斯分布的形式 3.Ai 决定的线性系统是可逆的，我们需要计算得到的部分是 Ai，等价于需要求得的部分是 Wi，Wi 是 Ai 的逆矩阵。可以看到，si 的个数是 m，wi 个数也为 m，当 m 与输入 pixels 个数也就是 784 相等时，系统是可逆的。

对输入数据进行规范预处理。1，移除 dc component，类似于直流分量，在图片中意味着各个像素点的灰度均值，灰度值在手写体里面的表现就是 0～255，导致了 si 是零均值的，公式 5.1，原本数据是零均值，映射到另一个坐标系中也大概是零均值的。2，计算 principal components。3，只取前 n 个 principal components，n 一般等于原始数据维度的四分之一。4，主成分通过公式来分类，得到 whittened 的 si。结果得到了一个 n 维的向量用 z 表示。规范预处理后的 zi 是 si 的线性变换，为公式 7.5。公式 7.5 两边乘以 b 矩阵的逆矩阵，这个地方需要注意的是虽然 bij 不是方阵，但是它仍然存在逆矩阵成为广义逆矩阵。那么就变成了公式 7.6，vi 就是每一个行向量。也就是说 zi 与 si 是两个东西，si 是独立组件，但是 zi 不是，不然我们就可

以直接用 zi 了，只是 zi 可以用来表示 si。

$$z_i=\sum_{j=1}^{m} b_{ij}s_j \quad 7.5$$

$$s_i=\sum_{j=1}^{n} v_{ij}z_j \quad 7.6$$

关于 ica 的具体内容，首先假设我们已知 si 的概率密度函数，来求 zi 的概率密度函数，求出 V 的行列式的绝对值，给出 z 向量的概率密度函数，公式 7.13，其中参数变量是 vij，同时各个 patch 选取是随机的，可以视为相互独立，每个 zt 都是第 t 个 patch 经过预处理获得的，T 是观测到的 patch 数量，因此得到似然函数 7.14，然后取对数得到似然函数的对数，然后我们需要求出最大化的 7.15，

7.5—1

$$p(\boldsymbol{z})=|det(\boldsymbol{V})|\prod_{i=1}^{n} p_i(\boldsymbol{v_i^T z})=|det(\boldsymbol{V})|\prod_{i=1}^{n} p_i(\sum_{j=1}^{n} v_{ij}z_j) \quad 7.13$$

$$L(\boldsymbol{v_1},\cdots,\boldsymbol{v_n})=\prod_{t=1}^{T} p(\boldsymbol{z_t})=\prod_{t=1}^{T}\left[|det(\boldsymbol{V})|\prod_{i=1}^{n} p_i(\boldsymbol{v_i^T z})\right] \quad 7.14$$

$$\log L(\boldsymbol{v_1},\cdots,\boldsymbol{v_n})=T\log|det(\boldsymbol{V})|+\sum_{i=1}^{n}\sum_{t=1}^{T}\log p_i(\boldsymbol{v_i^T z}) \quad 7.15$$

第一项可以视为常数然后舍去，要求 vij，那么就对每一个求偏导令式子等于 0，就像常规求解的过程一样。pi 在 ica 的论文中选取的是 cosh（ ）。

关于我们需要实现的论文，它提到了 ica 的两个缺陷，一是很难学习到超完备（overcomplete）的特征值，即特征值的维度不能大于输入数据的维度（784）这点我们可以从定义式 1.4 中看到，这种方式假设了特征 Ai 与输入 pixels 个数一致，第二是预处理过程降低了输入数据的相关性，这都是因为在 pca 处理过程中需要特征值 w 之间正交。因为正交所以不完备。ica 对白化很敏感体现在白化高维数据不太可行，白化公式是 5.16，假如 784，分解协方差矩阵次数会达到 784*784。

因此论文中提出了用 reconstruction cost（重建损失）来替换掉正交限制。传统的标准 ica 如下 1，论文中提出的修正算法为 2，m 是 pixels 个数.

$$\underset{W}{minimize}\sum_{i=1}^{m}\sum_{j=1}^{k} g(W_j x^{(i)}) \quad \text{subject to WW^T = I} \qquad 1$$

$$\underset{W}{minimize}\frac{\lambda}{m}\sum_{i=1}^{m}\|W^T W x^{(i)}-x^{(i)}\|_2^2+\sum_{i=1}^{m}\sum_{j=1}^{k} g(W_j x^{(i)}) \quad 2$$