**流形学习与auto encoder**

流形学习可用基于这样一个事实：许多应用中，数据集中在一个或多个低维的流形中。

流形（manifold）：是一个局部线性的拓扑空间。

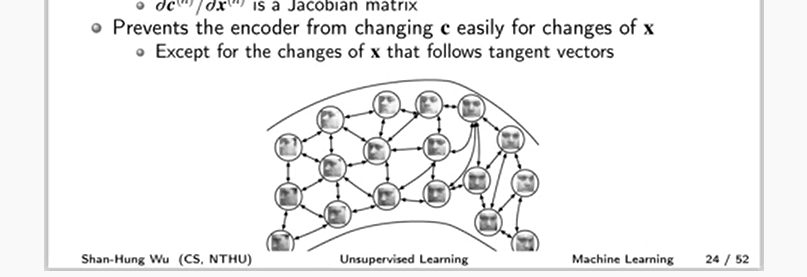
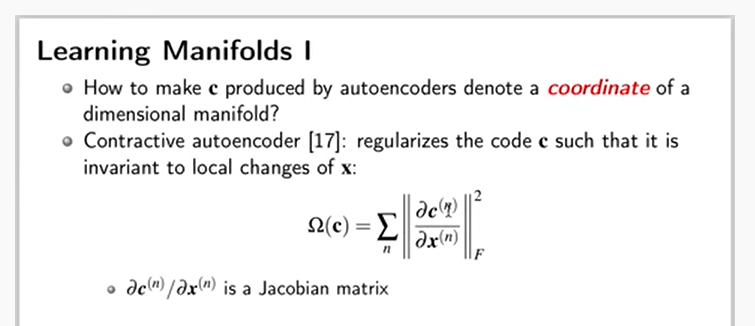
autoencoder why so good?

1.data's pattern is lower-dimensional, 本身是嵌入在高维空间中的低维流行，通过降维仍然可以保留原始数据的许多信息。

2. nn itself advantage：通过非线性层的叠加，nn用简单函数的层层复合，以及稳定长期的参数训练，nn最终可以学习到任意数据形式的表示。

how to make c produced by autoencoders denote a coordinate of a dimensional manifold?

Contractive autoencoder:

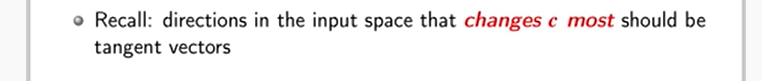


c轻易不会变，要变会随着tangent vector动，

In practice, it is easier to train a denoising autoencoder.

加random noise, c不轻易变,就含有 manifold coordinate.

how to get the tangent vector of a given point?



given a point x , let c be the code of x and J(x) be the Jacobian matrix of c at x,

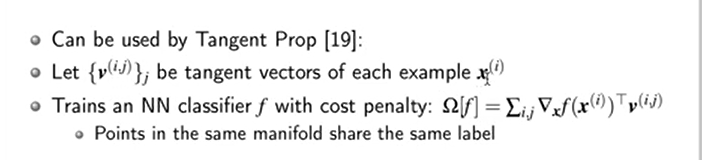
* J(x) summarizes how c changes in terms of x

1）分解J(x) using svd J(x)=UDV'

2) Let the tangent vector be the rows of V corresponding to the largest singular values in D



tangent vectors:



**对自编码器性质探究与光伏发电数据异常检测的总结**

从5月-6月初，查阅相关文献调研，自编码器是什么？自编码器的变种有哪些？自编码器的应用有哪些？然后将autoencoder用于光伏发电的数据的异常检测。

从6月-7月初，探究autoencoder的性质，得出的结果并不fancy.

**1. 光伏发电数据的异常检测**

由于自编码器的异常检测的步骤是用正常数据做训练，然后在测试阶段检测异常，然后通过loss的阈值去分离正常数据与异常数据。在这之前需要对原数据标定好是正常或异常，然后在训练完成后，判断结果与标签作对比得到准确率等判定模型好坏的指标。

模型设计很简单，采用一层非线性层的AE，由于正常数据与异常数据的差距十分明显，所以结果很好。

但是针对这个数据集，好多条记录的特征是缺失的，而且非常多，由于无法得知这些缺失的数据缺失的原因，所以我假定这些缺失的数据为异常数据，并且将缺失数据补全，为5850。但是如果不补全而是删去，数据量会很小。

**总而言之，这个数据集不好。在接下来的异常检测任务中，或许需要挑选新的异常检测数据集。**

**2. 自编码器的探究实验**

另外！在做这个应用的过程中其实包含着对AE模型的理解和探索：

**（1）AE的隐空间Z有什么特殊的性质吗？是否可以通过某种数学语言描述一个分布或者有特殊的曲面？**

通过学习,查阅文献资料，发现对于普通的autoencoder（即对z不加约束或者说在损失函数中没有对z的惩罚项）的z向量的研究停留在它是一个原始数据的好的降维向量。它的各个维度也没有特定的物理意义。不同于VAE的是，AE的隐空间没有定义分布，所以不能像VAE一样在隐空间连续采样，生成有意义的结果。目前只能理解为将原始数据的簇映射为新的簇。原始数据之间距离近的在隐空间也近。

（**2）正常数据和异常数据的隐变量在隐空间有什么特点吗**？

通过实验发现，正常数据和异常数据的隐变量在隐空间聚成不同的簇。这些簇没有特定的形状。簇和簇之间是分开的。

（**3）阈值边界点数据的隐变量有什么特点，是否有分界面，将正常数据的z与异常数据的z分割开？这个分界面能用数学公式表示吗？**

通过实验发现，依据不同的阈值，阈值边界点的隐变量聚集在正常数据和异常数据的连线上，也是聚成簇。没有形成类似平面或曲面的分界面。不能用有意义的数学表示。

**（4）在loss = threshold这个方程中，我们可以解出阈值边界点集合X\_thres，那么这个阈值边界点是否构成某个曲面？有什么性质吗？**

依据曲面方程的定义，loss = threshold本身是一个曲面方程，所以通过牛顿法和遍历法求得的阈值边界点是在某个曲面上的。不同的激活函数，不同的隐变量个数形成不同的曲面和曲面方程，不具备特殊的性质。

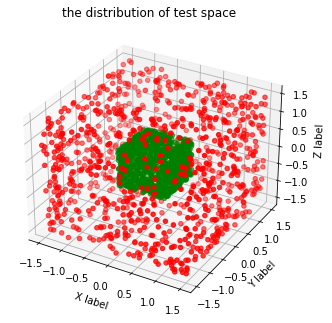
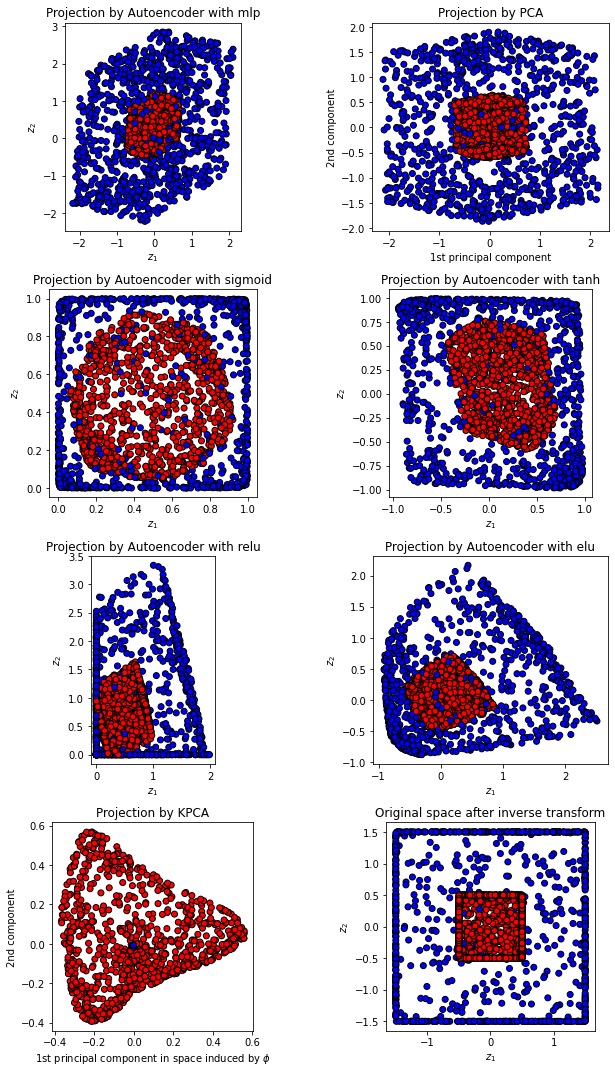
**（5）不同的数据形式是否会有不同的阈值边界点曲面？比如立方体？球？椭球？两球相交？**

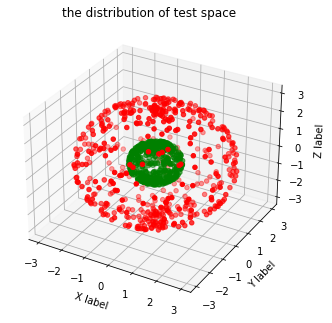
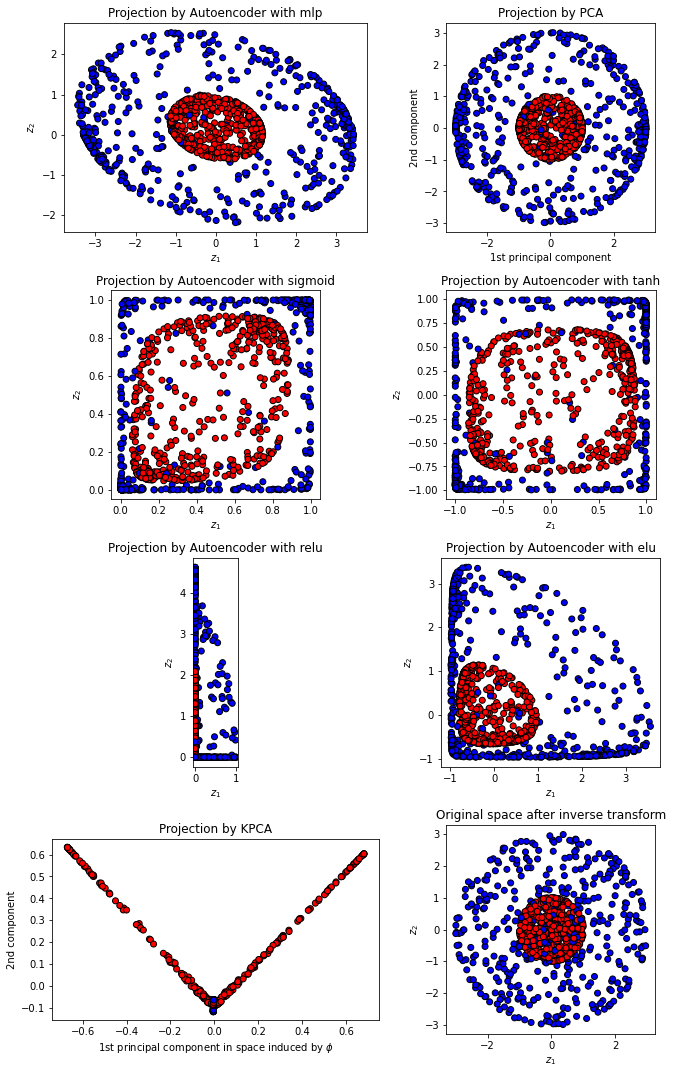
目前的答案是 否。因为曲面方程中激活函数的不同，隐变量个数的不同以及阈值的不同，所以有不同的曲面方程和曲面。不同形状的数据，如立方体面上的数据？球面上的数据？只是改变了在训练收敛后的不同参数值。这些参数值不会本质改变曲面方程，以及曲面。

至此，这些问题的解答都基于autoencoder的定义，性质与原理，工作方式决定，如果我们将autoencoder的定义，原理，工作方式全部理清，那么这些问题自然得到解答，而不是空想？

而问题（5）问了另外一个问题，autoencoder得到的Z和X\_thres是否会因为数据的形式的不同而不同？这种不同 指的不是序列数据（时序，文本），空间性数据（image），或者graph数据（分子结构，社交网络数据），而在解析集合的角度考虑，考虑数据是否满足某个曲面？球面？立方体？。。。

其实在深度学习模型，数据的形式是及其复杂的，维度是很高的，无法用一个曲面的表面去表示，更多是用某个分布去近似或者干脆不近似。所以数据满足不同的曲面，对于autoencoder是没有区别的。当然应该用实验加以验证，补。





一边在考虑上述问题，一边进行实验探究。首先生成toydata，然后训练一层的AE网络，然后可视化Z的分布，X\_thres有什么性质没有。

Step1:生成椭球内部的数据，变化各种激活函数，观察Z or X\_thres. 变换的激活函数有：MLP, Sigmoid, Tanh, ReLU, ELU.

Step2：生成高斯分布数据，

Step3: 生成两球相交的数据。

Step4: 生成球面与立方体表面的数据（限3维）

除了观察Z和求解X\_thres，还有其他可探究的实验吗？暂时还没想到。

目前通过可视化的方式对Z和X的边界点进行探究

是否可以在manifold learning的领域做一些解释。

通过上述对autoencoder的探究，基本上可以回答上述几个问题。但是问题的答案不是吸引人的，现实是残酷的，并不能得出promising的结果。

但是比较最初对AE的认识，现在有了更清晰的认识：他就是通过训练一个identity function（这个在众多学者叙述中都是如此）即。然后得到的Z 亦称为code,即是X的低维表示。Z难以有可以解释的含义。

在1989的文章中，学者理论上证明了无bias的MLP等价于PCA。MLP的局部最优解是X'X的z\_dim个特征向量的组合，MLP的全局最优是X'X的前z\_dim个特征向量的组合。

在Manifold Learning的观点中，AE可以算是一种manifold learning,

**3. 接着回到异常检测问题中。---**如何借助AE实现一个好的异常检测任务。

如果有一个异常检测问题，我们能判定它是否适合用AE解决吗？

1. 用数据来试，结果好就用，结果不好就不用。中间尝试各种网络结构选择。

应该不止用数据去试，可以分析不同的问题，是有标签无标签，或者是什么类型的数据。而且是否可以分析数据是什么流形，能否应用manifold learning。

1. 是否能预先判定AE可用/不可用？

首先要分析AE对数据有无要求，有没有天然缺陷？然后再判断能否对某些异常检测有用，有些不可用？也就是说异常检任务测在这个方面怎么分类

在我的观点中，AE对于异常检测的训练方式就如上述。即找到正常数据去训练就可以。

3）AE适合应用于那类异常检测问题？从数学的角度？（立方体？球？椭球？）

当然首先异常检测问题的分类是怎样的分类的？它都有哪些分类方式？（[有一篇知乎文章做了科普](https://zhuanlan.zhihu.com/p/266513299)）

*异常检测（Deep Anomaly Detection）*

从定义而言就是一种**识别不正常情况与挖掘非逻辑数据**的技术，也叫**outliers**.

在异常检测问题中选择怎样的网络结构很大部分取决于原始数据（raw/input data）的类型，原始数据在广义上我们分为连续性（Sequential）与非连续型（Non-sequential）.如何选择相应的模型，简单如下表所示：

|  |  |  |
| --- | --- | --- |
| 原始数据类型 | 举例 | 异常检测模型选择 |
| 连续型 | 视频，DNA序列，自然语言文本 | CNN，RNN类（包括LSTM） |
| 非连续型 | 图片，传感器 | CNN，AE及其变种 |

依照数据标签的可用性（availability of labels）

数据的标签是非常重要的事情，标签代表着正常（normal）数据或者异常数据。异常检测因此也根据标签的内容广义分为监督，半监督，和无监督。因为有时候数据的标签难以获取，所以无监督的方法占据主流。自编码器是无监督方法的核心，是基于重构的方法检测；还有一种方法是基于预测的方法。

依照数据异常的类型

1.点集：就是指正常数据中，某一个中间数据突然显示异常

2.连续集：就是值上下文相关的连续数据出现异常。

3.团队集：举个例子，如果某天你的信用卡突然短时间内不停地消费50元，那机器可能会发现，这些团队数据集的消费出现了异常，这种情况我们也在其他场合经常遇到。

所以说，AE对某个异常检测能否使用，就看能否使用重构的方法检测异常。

目前的光伏电场数据适合做的有**园区光伏原始数据7-10月**和**青海光伏数据**。但是首先得弄明白当时青海的光伏数据是用来做什么用的？有没有异常数据？

2021.7.22

基于图像或结构化的数据，异常检测方法基本还是autoencoder框架。正样本去训练，正样本和负样本（或者说异常样本）要在预测阶段准确的检测出来。因此要求模型（1）对异常数据敏感（2）对正常数据不敏感。对于模型好坏的评价指标：用**confusion matrix**(混淆矩阵)。模型的超参选择采用：**grid search**.

问题的关键是针对我们的需求，和数据本身，怎样设计一个好的模型？

我们的需求：

（1）对异常数据敏感（2）对正常数据不敏感。

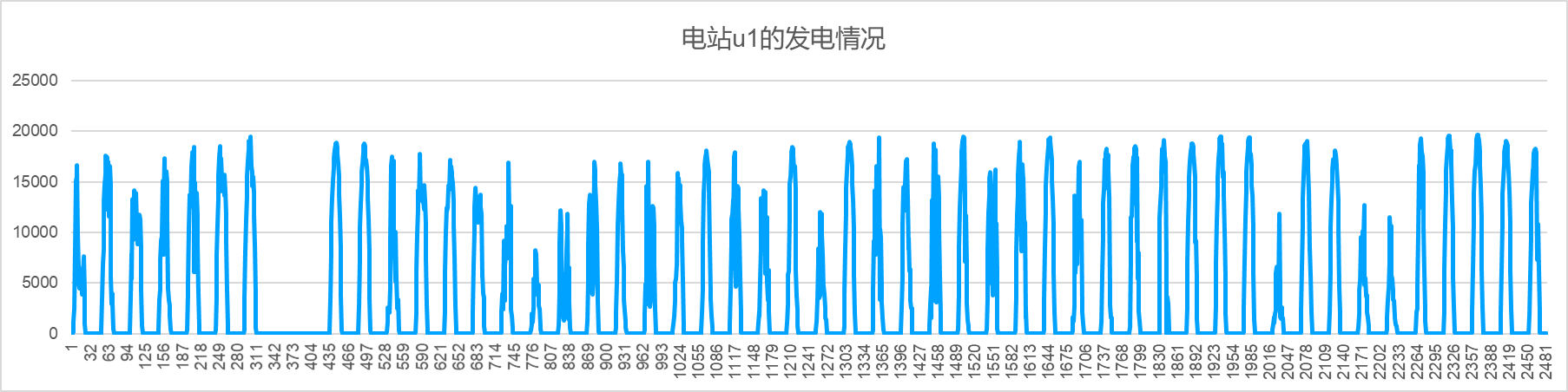
数据描述：

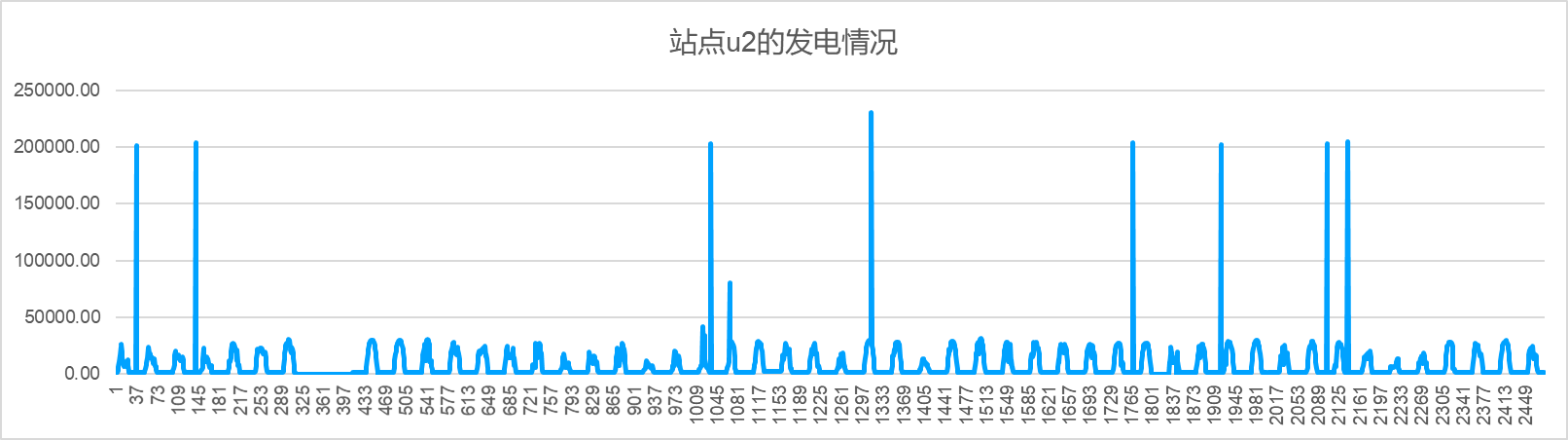
光伏发电的数据，不同其他形式的数据，它的数据记录一般是在白天，也就是说在白天的时间段是有数据，在晚上为0。这样数据就具有了周期性。

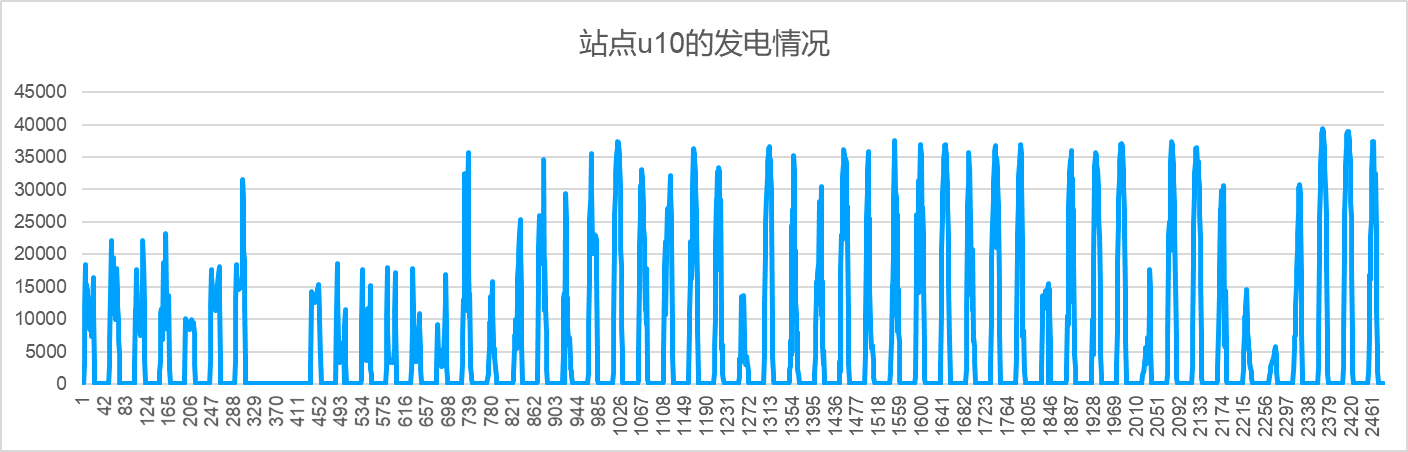
1. 青海数据

青海数据总共有10个发电站，总体数据良好。

如下图所示为青海光伏发电数据集：







可以发现站点U1和 U10的数据属于正常的波动，正常的数据有两种数据模型，一个表示在白天正常发电状态，一个表示晚上的静默状态。

然而，站点u2的发电情况显示，发现u2在某些时间会出现异常高的数值。这些数据是否可以看作是异常数据。而且是单点异常数据。其中第36个数据为200000，而前后时刻均为静默数据，在这个时段，其他站点也是静默状态；而第1313个数据也出现异常，前后数据均是白天的正常记录数据，在这个时段，其他站点的记录也是正常的。而且通过画图，我们发现好像就只有这一类异常数据。因此，异常数据会在任意时段（不论白天或晚上），任意站点都会发生异常。

（通过了解实验室同学，发现光伏发电的数据收集得到之后，没有用作过异常检测，除了王shujie用kmeans做过一点，方法也待商榷；也没用于其他实验。一直是待用状态）。

通过对青海数据的初步探究，发现数据规整，异常数据很少，所以不适宜用于异常检测。

1. 园区数据7月-10月

**该数据集共36列，即36个站点。而且每个站点的数据范围不同，甚至差距很大。异常情况有两种：突然变大的单点异常数据和连续长时间不变的静默数据。**

对该数据集有以下疑问：

1. 这些站点是独立收集数据的吗？他们为什么有点数值均值很大，有的很小？是什么原因，这些站点之间有什么联系没？这就间接的说明是不是可以用graph去做呢？如果是独立的就没必要用了。

2. 两种不同的异常类型，如何在模型中检测出来呢？

-----------------------------------------------------------------------------

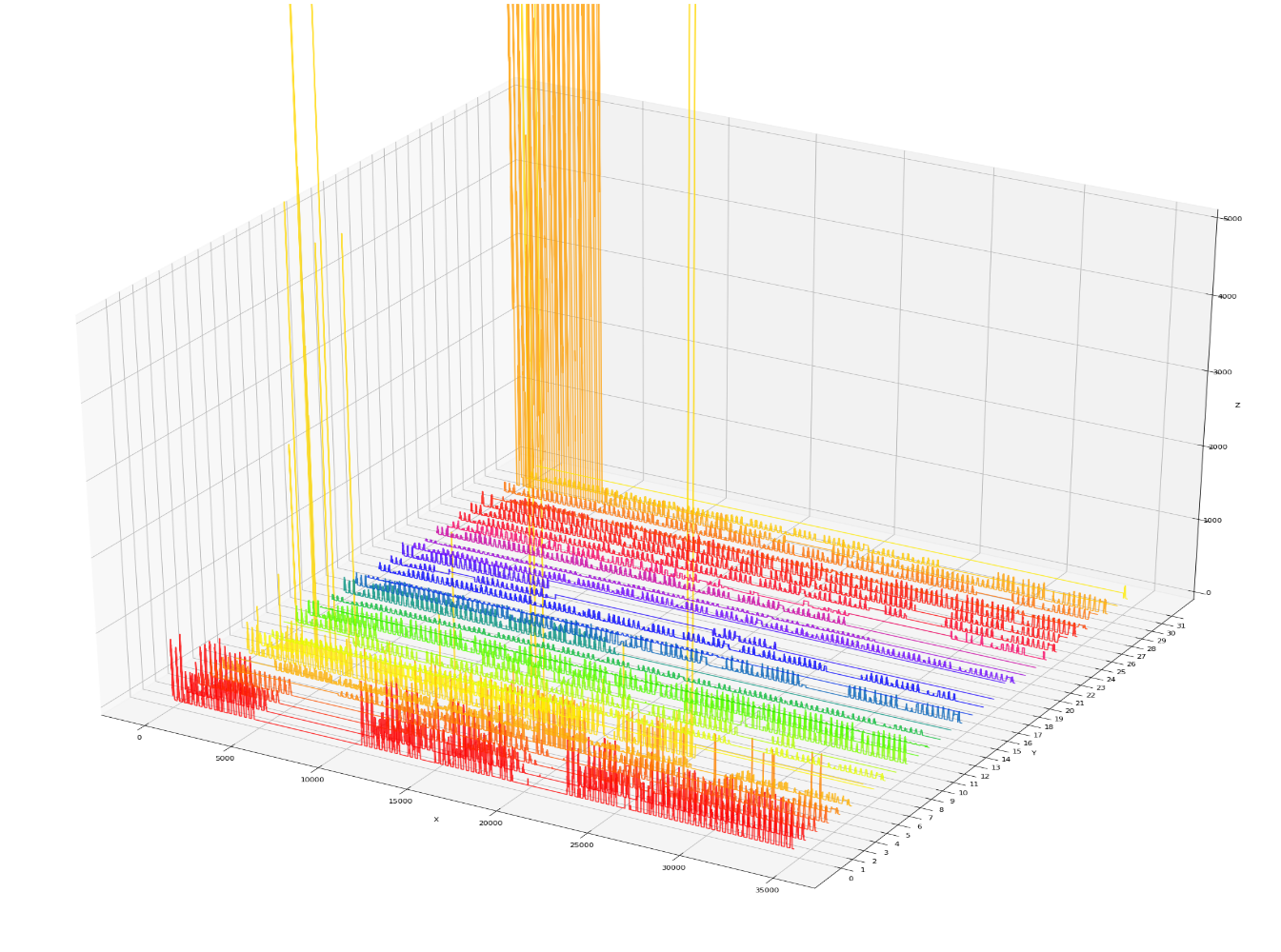
首先把autoencoder的球上的数据补全碓好✔

然后写anomaly detection的代码

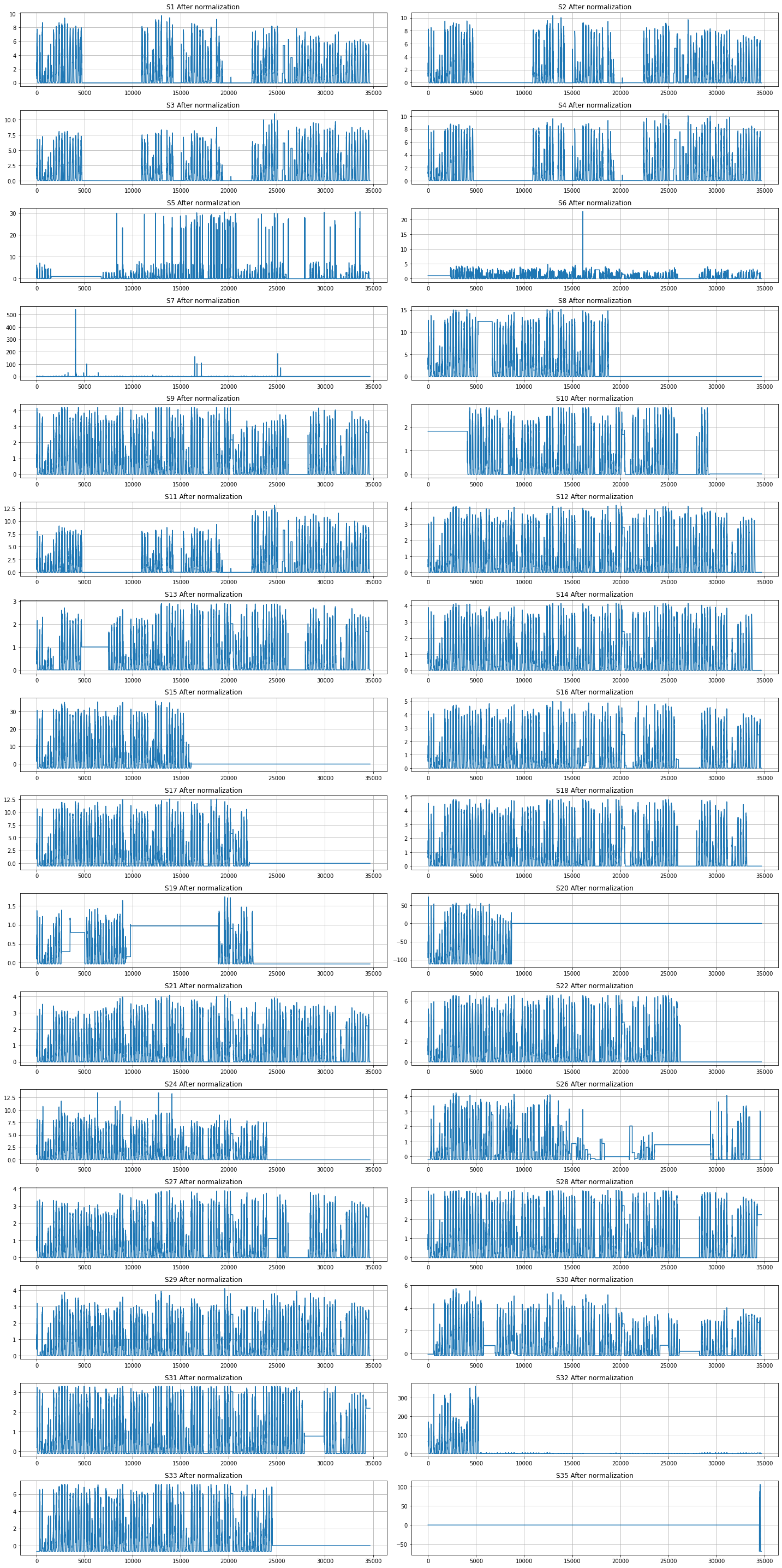
基于预测，基于重构，ARIMA模型（周期性数据）

还要思考：manifold learning how to be used in anomaly detection

通过数据描述发现一些列或者说一些站点数据无效，因此剔除掉S23, S25, S34, S36.这样剩下32个站点数据。



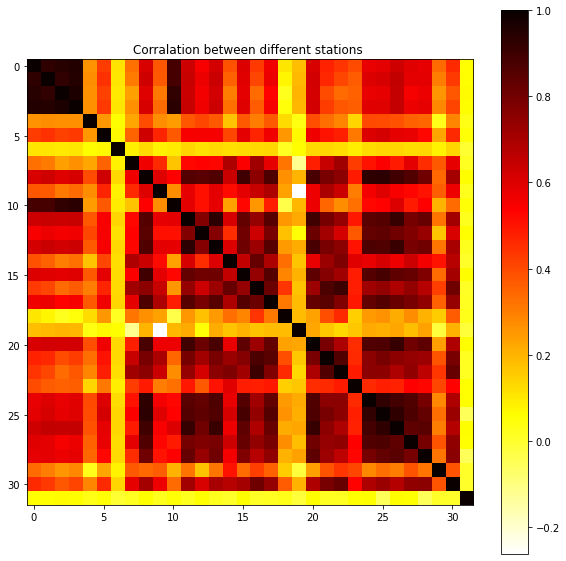




通过上述的数据可视化发现，每个站点可能出现的异常时间不同，比如有的在开始阶段异常后面正常波动，而有的在中间异常，有的则在最后异常。而且每个站点的取值范围也是不同的，**所以说基于重构的方法很难实行（因为要挑选正常数据进行训练）**；再者，一个站点的坏数据不能影响这个时刻其他站点记录的好数据。**因此需要对逐站点进行异常分析？！**

**因此，尝试使用基于预测的方法。**

因此，我们分析了各个站点相关性，观察各个站点之间是否有某种相关性，如下图heatmap：



发现部分站点之间具有较强的相关性。其他地方的相关性较弱的可能受异常数据的影响。

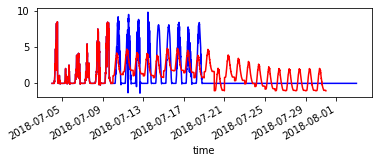
分析：

如果各站点有相关性是否可以用graph的方法识别异常？？？

因此先假设各站点无相关性，彼此独立。

然后设计模型：

1.周期性时间预测模型（ARIMA）：



蓝色为原始处理后的数据，红色为预测数据。

2.LSTM模型：