## 如何在非负矩阵分解中选择最佳潜在因子数量

有两个良好的标准：1）显着相关系数； 2）比较残差平方和与一组等级的随机数据的比较（也许有一个名字，但我不记得了）

Cophenetic相关系数： 您对每个等级重复NMF几次，并计算结果的相似程度。换句话说，考虑到初始种子是随机的，识别出的簇的稳定性如何。在共模系数下降之前，选择最高的K。

**针对随机数据的RSS** 对于任何降维方法，与原始数据相比（RSS估计）总是会丢失信息。现在执行NMF以增加K，并使用原始数据集和随机数据集计算RSS。当比较RSS与K的函数时，RSS在原始数据集中随着K的增加而减小，但是对于随机数据集来说情况则更少。通过比较两个斜率，交叉处应该有一个K。换句话说，在噪声之内，您能损失多少信息（=最高K）。

参考文献：

1，吉恩 Brunet，Pablo Tamayo，Todd R. Golub和Jill P. Mesirov。使用矩阵分解进行元基因和分子模式发现。美国国家科学院学报，101（12）：4164-4169，2004。

2，阿蒂拉·弗里吉西（Attila Frigyesi）和马蒂亚斯·霍格隆（Mattias Hoglund）。用于分析复杂基因表达数据的非负矩阵分解：临床相关肿瘤亚型的鉴定。Cancer Informatics，6：275-292，2008。

**要在非负矩阵分解中选择最佳数量的潜在因子，请使用交叉验证。**

如你写，NMF的目的是要找到低维WW和HH以最小化重构误差的所有非负元素∥V−WH∥2‖V−WH‖2。想象一下，我们遗漏了VV一个元素，例如VabVab，并对缺少一个单元的结果矩阵执行NMF运算。这意味着找到WW和HH可使所有非缺失像元的重构误差最小：

∑ij≠ab(Vij−[WH]ij)2.∑ij≠ab(Vij−[WH]ij)2.

一旦做到这一点，我们可以预测左外元件VabVab，通过计算[WH]ab[WH]ab，并计算预测误差

eab=(Vab−[WH]ab)2.eab=(Vab−[WH]ab)2.

可以重复此过程，一次只删除所有元素VabVab，并对所有aa和bb的预测误差求和。这将导致整体PRESS值（预测的残差平方和）E(k)=∑abeabE(k)=∑abeab，这将取决于kk。希望函数E(k)E(k)可以用作“最佳”kk的最小值。

请注意，这可能会在计算上造成很高的成本，因为必须为每个遗漏的值重复NMF，并且编程也可能很棘手（取决于使用缺失值执行NMF的难易程度）。在PCA中，可以通过省略完整的VV行来解决此问题（这会大大加快计算速度），请参阅[如何对PCA执行交叉验证以确定主分量的数量中的](https://qastack.cn/stats/93845)答复[。](https://qastack.cn/stats/93845)，但这在这里是不可能的。

当然，所有交叉验证的通常原理都适用于此，因此一个人一次可以省去很多单元（而不是一个），并且/或者只对一些随机单元重复该过程，而不是遍历所有单元。两种方法都可以帮助加快流程。

**编辑（2019年3月）：**参见[@AlexWilliams的](https://qastack.cn/stats/users/35917/alex-williams)这张非常精美的插图[文章](https://qastack.cn/stats/users/35917/alex-williams)：[http](http://alexhwilliams.info/itsneuronalblog/2018/02/26/crossval) : [//alexhwilliams.info/itsneuronalblog/2018/02/26/crossval](https://qastack.cn/stats/users/35917/alex-williams)。Alex 对于缺少值的NMF 使用<https://github.com/kimjingu/nonnegfac-python>。