Computer Science and Application 计算机科学与应用, 2022, 12(9), 2094-2105
Published Online September 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.129212

Hans汉斯

基于非负矩阵分解的全球贸易缺失数据填补

宋丛威，张晓明

北京雁栖湖应用数学研究院，北京

收稿日期：2022年8月9日；录用日期：2022年9月7日；发布日期：2022年9月14日

摘　要

大数据时代，外贸企业对全球贸易数据高度依赖。但是数据缺失严重，给数据分析带来不便。本文提出用非负矩阵分解填补缺失数据；构造并实现了填补算法。实验通过和线性插值填补法进行对比，证明非负矩阵分解更适合应用于缺失数据填补，同时能够提取主题进出口矩阵，帮助人们理解贸易状况。

关键词

缺失数据填补，非负矩阵分解，Poisson分解，乘法更新规则，全球贸易

Imputation of Missing Data for Global Trade Based on Non-Negative Matrix Factorization

Congwei Song, Xiaoming Zhang

Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing

Received: Aug. 9th, 2022; accepted: Sep. 7th, 2022; published: Sep. 14th, 2022

Abstract

In the era of big data, foreign trade enterprises are highly dependent on global trade data. However, serious data loss brings inconvenience to data analysis. In this paper, non-negative matrix factorization is proposed to impute the missing data; an imputation algorithm is constructed and implemented. The experiment proves that non-negative matrix factorization is more suitable for imputation of missing data and can extract the topic import-export matrices to help people understand the trade situation by comparing with the imputation method by linear interpolation.

文章引用: 宋丛威, 张晓明. 基于非负矩阵分解的全球贸易缺失数据填补[J]. 计算机科学与应用, 2022, 12(9): 2094-2105. DOI: 10.12677/csa.2022.129212

Keywords

Imputation of Missing Data, Non-Negative Matrix Factorization, Poisson Factorization, Multiplicative Updating Rule, Global Trade

1. 引言

随着大数据时代的到来，海关数据也成为研究热点。海关数据目前主要掌控在政府部门手中，尚未被普遍使用。本文研究有较高的时效性。

全球贸易数据是海关数据的重要组成部分。外贸企业希望利用全球贸易数据指导其生产活动。然而某些国家的政府出于各种理由瞒报或误报贸易数据，导致全球贸易数据不完整。数据缺失可能对外贸企业做出合理决策造成阻碍。本文利用机器学习的方法填补缺失数据。填补结果可供企业利用。

数据抓取自国际数据网站 https://www.trademap.org/Index.aspx。原始数据的存储形式为 20 个 20×20 的矩阵，可表示为

$$A_t = \left\{ A_{a,b,t} \right\}_{a,b}$$

其中元素 $A_{a,b,t}$ 表示国家 a 国家 b，在 t (从 2001 年到 2020 年)年份的某货物(编号 72)出口总量。本文称 A_t 为(*进*)*出口矩阵*，等价于一个加权有向图，可称为“(进)出口图”。当绘制矩阵热力图时，进出口国家按照给定顺序排列，行为出口国，列为进口国。

本文的任务就是要填补数据 A_t 的缺失部分。A_t 中所有 0 值均被假定为缺失数据。

所有数据会事先处理成矩阵形式。首先将每个出口矩阵拉直成 400 维行向量。但其中 20 个国家有自己到自己的“进出口数据”即 0，这不是真实的。因此拉直时需要去除这 20 个 0，得到 380 维向量——对应一个非自反的有向图。最后组合成一个形状为 20×380 的矩阵 X。这个矩阵的每一行代表某年某货物出口状况。若出口矩阵的元素 $A_{a,b,t}$ 存储在行向量的第 j 位，则第 j 个属性名为“$a-b$”，其中 $a-b$ 分别为出、进口国名称。

注 特别强调，本文所有数据具有相同单位，但不指明该单位。此外，不透露编号为 72 的货物的真实名称。

正文组织如下。第二节介绍非负矩阵分解(NMF) [1] [2] [3] [4] [5]及基于其上的缺失数据填补的原理[6]。给出基于 Poisson 分解(PF) [7] [8]的统计学解释。第三节，根据已知的原理，设计缺失数据填补算法。算法的设计受 NMF 的乘法更新规则的启发。最后，实验结果展示在第四节，包括填补结果，误差计算和模型比较。

$\mathbb{R}_+$ 表示非负数。矩阵转置记为 A^{T}。$P(\lambda)$表示均值为 λ 的 Poisson 分布[9]，$B(p)$表示 Bernoulli 分布。矩阵的 Frobenius 范数定义为: $\|A\|_F^2 = \sum_{ij} |A_{ij}|^2$。只要没有分歧，所有累加符号都不指明求和范围，如 $\sum_{ij} A_{ij}$。

2. 数据填补原理

2.1. NMF 简介

非负矩阵定义为其所有元素非负的矩阵。所有大小为 $m \times n$ 的非负矩阵的集合记为 $\mathbb{R}_+^{m \times n}$。一个非负

矩阵 $X \in \mathbb{R}_+^{N\times p}$ 有如下近似分解：

$$X \sim WH \tag{1}$$

其中 W 和 H 分别属于 $\mathbb{R}_+^{N\times q}$ 和 $\mathbb{R}_+^{q\times p}$，一般 $q \le p$。这个近似分解称为 NMF。

(1)的等价写法是，$x_i \sim \sum_{k=1}^{q} w_{ik}h_k$，其中 x_i 是 X 的第 i 行，h_k 是 H 的第 k 行。这意味着任何 $\mathbb{R}^p$ 上的样本 x_i 近似为基 $\{h_k\}$ 的线性组合[5]。

注 补充说明一下，H 或其行向量组不一定是基，因为没有要求它具有独立性。这里还是习惯称它为基。

注 有的作者把 NMF 写成 $X\sim WH^{\mathrm{T}}$。有的把 W 看成基。这些都不是本质性的。

注 人们习惯用 X 的行向量表示单个样本；这些行向量构成的矩阵表示整个数据集。在本文中，单个样本对应一个出口矩阵。

给定损失函数 l。求解下述优化问题可得 W，H 的估计。

$$\min_{W,H} \sum_{ij} l\left(X_{ij},(WH)_{ij}\right) \tag{2}$$

常用损失函数选为平方误差 $l(x,y)=|x-y|^2$ 或者散度 $l(x,y)=d(x\|y)=x\log(x/y)-x+y$。本文选择后者。

散度下的优化问题和所谓的 Poisson 分解模型等价：

$$X \sim P(WH) \tag{3}$$

即在已知 W，H 的条件下，x_{ij} 独立地服从 Poisson 分布 $P((WH)_{ij})$。(2)的解正好是(3)的极大似然估计。此外，本文的 X 是非负整数值，其经济学意义就是货物金额的计量，而且有固定的单位。假定它服从 Poisson 分布是非常合理的。本文把 PF 作为 NMF 模型的统计学解释。

H 的行向量相当于 SVD 中的特征向量，在本文中，称为“主题向量”，反映进出口的内在规则。某年某货物的进出口情况由这些规则非负线性求和得到。H 的行向量也会重组为出口矩阵——*主题出口矩阵*。重构结果 $\hat{X}=WH$ 称为*重构出口矩阵*。

2.2. 非负矩阵三因素分解

本文把 NMF 写成非负矩阵三因素分解(NMTF) [10]的形式。NMTF 是 NMF 的一个变种。其分解形式如下

$$X \sim WDH \tag{4}$$

其中所有矩阵非负。本文规定，W 的所有列向量求和为 1，H 的所有行向量求和为 1，而 D 是对角矩阵，其对角线元素降序排列。获得 NMTF 形式的方式，可以是为其中的三个矩阵构造乘法更新规则，也可以先通过 NMF 得到 W，H，再进行归一化操作，得到(4)中的 W，H，而归一化系数构成对角矩阵 D。(4)中的 H 依然称为主题向量，WD 是分解系数，而 D 中元素反映对应主题向量的重要性。类似于主成分分析(PCA)，可严格定义第 k 个主题的重要性：

$$s_k = \frac{d_k}{\sum_{ij} x_{ij}} \tag{5}$$

利用重要性的概念，可发现最值得关注的主题出口矩阵，并确定分解的秩 q。

2.3. NMF 缺失数据填补原理

如果 X 包含缺失值，那么不能直接进行 NMF 分解。需要用合理的办法补全缺失值。有趣的是，NMF

本身就有填补缺失值的功能。

注 本文 X 不存在整行或整列的缺失情况。否则算法无法正常执行。而且这种情况应该需要进行预测，而不是补全。

设 X'为 X 的填补结果，即缺失位置补上合理数字，非缺失位置与 X 值相等。X 缺失情况用*缺失矩阵* R 表示，其中

$$R_{ij}=\begin{cases}1, & \text{如果}X_{ij}\text{已知}\\ 0, & \text{如果}X_{ij}\text{缺失}\end{cases} \tag{6}$$

每一种填补，都有对应的 NMF 分解 $X'\sim WH$，并产生一个误差 $L(X')=\min_{W,H}\sum_{ij}l\left(X'_{ij},(WH)_{ij}\right)$。合理的填补应该使这个误差最小。于是，缺失数据填补转化为下述优化问题。

$$\min_{X',W,H}\sum_{ij}R_{ij}l\left(X'_{ij},(WH)_{ij}\right)=\min_{W,H}\sum_{R_{ij}=1}l\left(X'_{ij},(WH)_{ij}\right)=\min_{W,H}\sum_{ij}R_{ij}l\left(X_{ij},(WH)_{ij}\right) \tag{7}$$

其中 X'是对 X 的某个填补结果，而 W，H 是非负矩阵，且仅当 $R_{ij}=0$ 时，

$$X'_{ij}\sim(WH)_{ij} \tag{8}$$

为了提高拟合性能，在(7)的基础上加入正则项：

$$\min_{W,H}\sum_{ij}R_{ij}l\left(X_{ij},(WH)_{ij}\right)+\alpha_W\|W\|_F^2+\alpha_H\|H\|_F^2 \tag{9}$$

其中 α_W，α_H 是合理的正则项系数。

注 多数文献把正则项写成 $1/2\alpha_W\|W\|_F^2$ 的形式。这不是本质性的。

实际上，PF 更容易给出数据缺失模型的统计学解释：

$$X_{ij}\sim P\left((WH)_{ij}\right),R_{ij}=1 \tag{10}$$

而当 $R_{ij}=0$ 时，没有观测到 X_{ij} 的样本。算得 PF 模型的对数似然为(已省略次要常数)

$$\sum_{R_{ij}=1}\left(x_{ij}\ln\left((WH)_{ij}\right)-(WH)_{ij}\right) \tag{11}$$

其中参数 W,H 的极大似然估计就是(7)的解。对 PF 来说，填补过程可以理解成生成过程：$X'_{ij}\sim P\left((WH)_{ij}\right)$，其中 $R_{ij}=0$。

最后，填补值 X'_{ij} 应该都是非负整数。这里两种处理方法：用 Poisson 分布生成 X'，其中 X'一定是整数值的；若直接用 NMF 的重构矩阵 WH，则因为其不一定是整数值的，所以需要进行取整处理。

注 X 不能出现整行整列缺失的情况，否则(7)无法正常填补所有缺失值。

3. 算法设计与实现

3.1. NMF 缺失数据填补算法

(7)相当于一个加权损失函数，其中缺失数据的权重为 0：$\sum_{ij}R_{ij}l\left(X_{ij},Y_{ij}\right)$。而标准的 NMF 算法很容易改造成加权形式[5] [11] [12]。

和标准的 NMF 一样，W，H 需要用迭代方法近似求解。迭代算法就是循环执行 W，H 的乘法更新规则[3] [4] [5]。在散度损失下，乘法更新规则为：

$$\begin{aligned}W_{ik}&\leftarrow W_{ik}\frac{\sum_j H_{kj}X_{ij}^R/(WH)_{ij}}{\sum_l H_{kl}R_{il}}\\ H_{kj}&\leftarrow H_{kj}\frac{\sum_i W_{ik}X_{ij}^R/(WH)_{ij}}{\sum_l W_{lk}R_{lj}}\end{aligned} \tag{12}$$

其中 $X_{ij}^{R}=X_{ij}R_{ij}$。

(12)是加权 NMF 乘法更新规则的加权形式。和标准 NMF 的乘法更新规则相比，(12)增加了一次和权重矩阵的 Hadamard 乘法(点态乘法)运算，故其计算复杂度和标准 NMF 相当。在填补的情形中，R 就是填补矩阵。当 R 的所有元素为 1 时，退化为标准 NMF。

考虑正则项后，本文对(12)进行如下改进：

$$\begin{aligned} W_{ik} &\leftarrow W_{ik}\frac{\sum_j H_{kj}X_{ij}^{R}\big/(WH)_{ij}+\alpha}{\sum_l H_{kl}R_{il}+\alpha_W W_{ik}+\alpha} \\ H_{kj} &\leftarrow H_{kj}\frac{\sum_i W_{ik}X_{ij}^{R}\big/(WH)_{ij}+\alpha}{\sum_l W_{lk}R_{ij}+\alpha_H H_{kj}+\alpha} \end{aligned} \tag{13}$$

其中 α_W，α_H，α 是三个合理的非负超参数。α 可以调节迭代速度，还可以避免除数为 0 的错误；α_W，α_H 是正则项系数。

NMF 填补算法

输入有缺失数据的 X，输出 W，H 和填补结果 X'

1. 初始化：用对每一列数据选择一个合理常数填补得到 X'，对 X'执行 NMF 得到 W，H 初始值
2. 用(12)或(13)计算更新 W，H
3. 重复执行 2 直到收敛
4. 计算 WH，X 的缺失部分用 WH 填补得到 X' (必要时，对填补结果进行整数化处理)

3.2. 程序实现

本文的算法用 Python3.8 实现，运行于 MacOS10.15 上。程序设计遵从 scikit-learn API 设计原则[13]。事实上，加权 NMF (缺失数据填补 NMF)直接继承自 scikit-learn 的 NMF 类，并重写 fit 等主要方法。

源代码、数据和实验结果均托管在 https://gitee.com/williamzjc/nmf-missing-data 上。

4. 实验

我们用一个数组存储 2001~2020 年编号 72 的货物的贸易数据。根据前文所述，数组可被表示为大小 20×380 的正整数值矩阵 X。缺失数据约占 4.4%。主题个数选为 $q=3$。

实验分两个大部分。第一部分是直接对缺失数据填补。第二部分通过进行人为的随机缺失把已知数据分成训练数据和测试数据。测试数据可用来估计预测误差，并和插值型填补法进行比较。

4.1. 填补结果

用本文算法对 X 进行填补。因为数据量较大，所以表 1 只列出 2001 年部分缺失数据的填补值，其余见本文提供的 Gitee 链接。

Table 1. The filling values of the missing data in 2020
表 1. 2020 年缺数数据填补值

出口国家	进口国家	填补值
墨西哥	新加坡	1676
墨西哥	沙特阿拉伯	6244
阿联酋	中国	6617

Continued

阿联酋	美国	58,935
…	…	…
沙特阿拉伯	俄罗斯	851
沙特阿拉伯	瑞士	2222

图 1 是一个直观展示填补效果的热力图，其中只显示缺失数据的填补值。热力图可以使我们直观了解出口矩阵的基本性质。

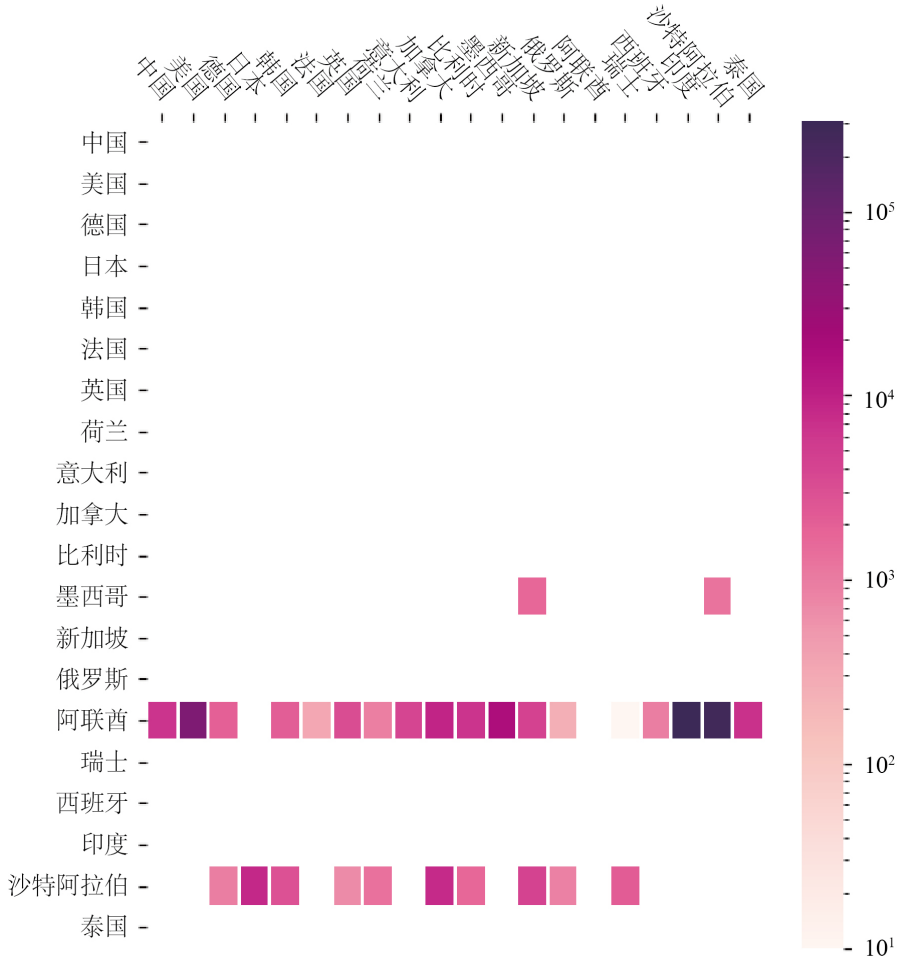


Figure 1. Heat map of the completion values in the import-export matrix (in 2020)
图 1. 进出口矩阵(2020 年)中填补值热力图

正如前文所述，重要性系数可以帮助我们发现值得关注的主题出口矩阵。根据图 2，选择 $q=4$ 是非常合理的。这也就是说，所有出口矩阵都近似地是三个主题出口矩阵的非负线性组合。

NMF 填补算法的结果中包含了主题出口矩阵。根据设置，一共有 4 个主题出口矩阵(见图 3)。

最后对主题出口矩阵做一个直观的解释：第一个主题出口矩阵主要反映中国向韩国出口情况；第二个主要反映亚洲各国之间、欧洲各国之间和北美洲各国之间的贸易，且出口热力图展现出明显的对称性；

第三个主要反映日本向亚洲各国出口的贸易；第四个主要反映亚洲各国之间的贸易，以日本为主。

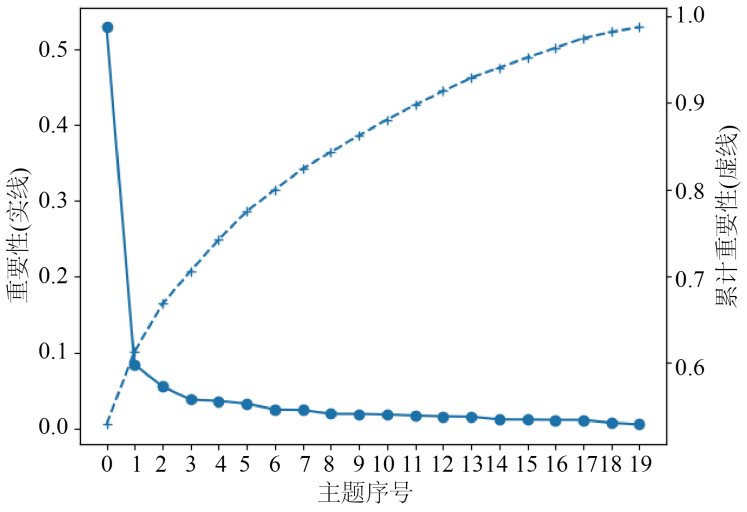
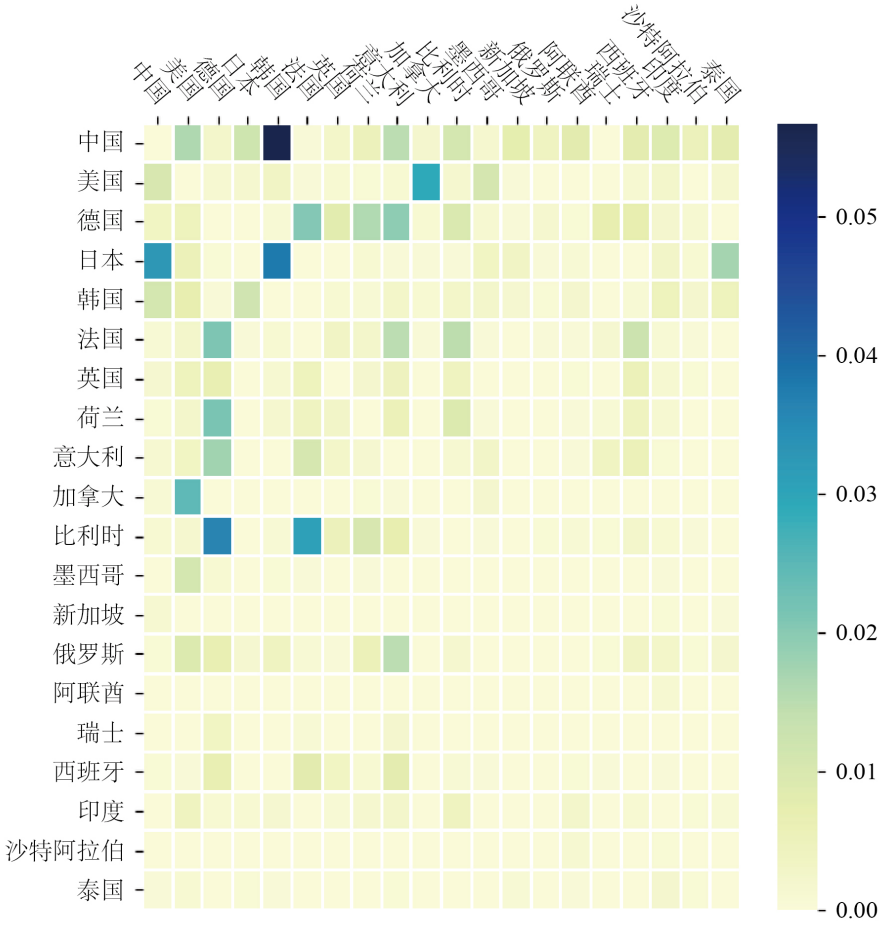
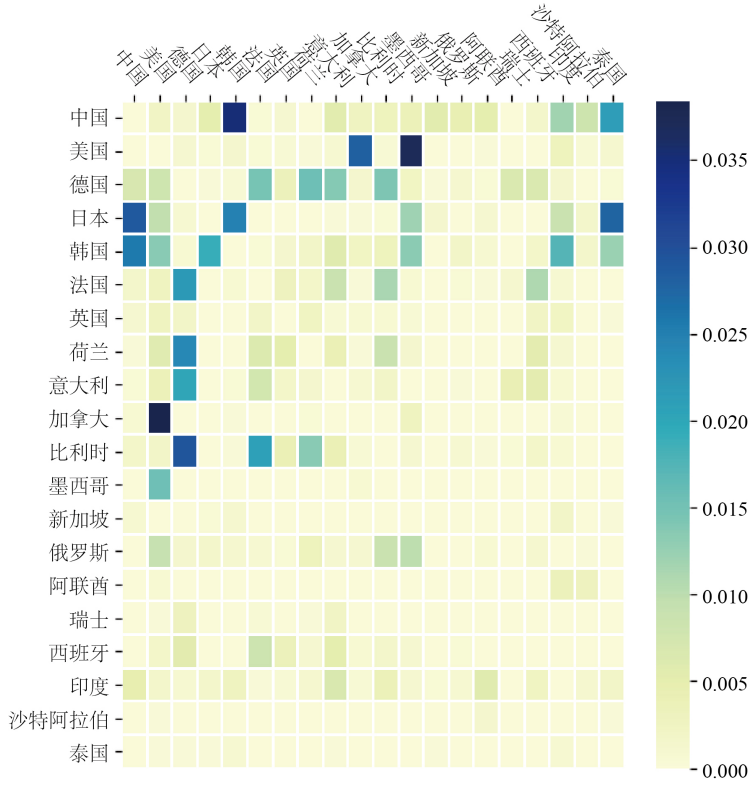


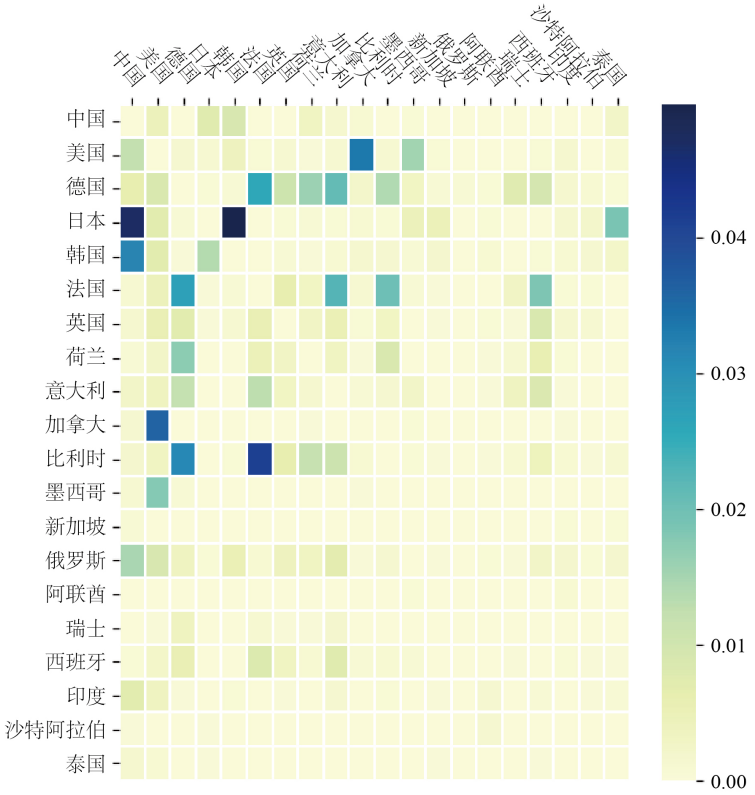
Figure 2. The 20 significance coefficients
图 2. 20 个重要性系数



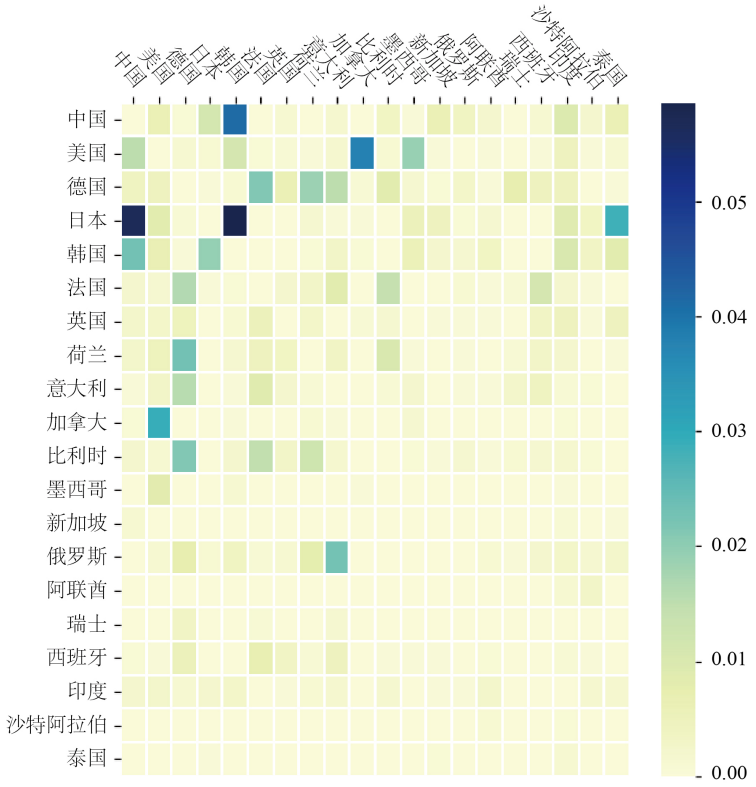
(a)


中国
美国
德国
日本
韩国
法国
英国
荷兰
意大利
加拿大
比利时
墨西哥
新加坡
俄罗斯
阿联酋
瑞士
西班牙
印度
沙特阿拉伯
泰国
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000

(b)


中国
美国
德国
日本
韩国
法国
英国
荷兰
意大利
加拿大
比利时
墨西哥
新加坡
俄罗斯
阿联酋
瑞士
西班牙
印度
沙特阿拉伯
泰国
0.04
0.03
0.02
0.01
0.00

(c)



(d)

Figure 3. The most four significant topic export matrices (normalized)
图 3. 最重要的 4 个主题出口矩阵(已归一化)

4.2. 填补误差

4.2.1. 每年平均预测误差比较

我们把 X 中的所有元素分成三部分：训练数据、测试数据、缺失数据。

如前所述，缺失数据约占 4.4%。测试数据是从已知数据(约占 95.6%)中分离出来的，用来估计误差，评判模型性能；也就是说已知数据包含训练数据和测试数据。从算法角度讲，测试数据也是缺失数据，因为训练阶段没有直接用在参数估计上。对缺失数据进行填补时，我们会用到所有已知数据，而且只能给出填补值，不知道真实误差。我们用下述公式直观解释数据的划分：

$$\text{数据} = \text{已知数据} + \text{缺失数据} = \text{训练数据} + \text{测试数据} + \text{缺失数据}$$

注 这里的数据指 X 中的元素，而不是 X 的行向量或整个矩阵。

缺失矩阵 R 可区别已知数据和缺失数据。测试和训练都是对 $R_{ij}=1$ 的数据进行的。我们随机删失 10%的数据，作为测试数据。此操作用随机缺失矩阵 M 表示，满足 $M_{ij}\sim B(0.9)$。若原数据的缺失矩阵用 R 表示，则训练时的缺失矩阵为 MR。因此，训练数据占已知数据的 90%，约占全数据的 90% × 95.6%。

三种模型每一年的误差如图 1 所示。(线性)插值填补对每列数据进行线性插值，会在两个已知数据中间形成一个等差数列，显然是一种过于简陋的填补方法。类似也可以构造二次函数插值填补，不过高次插值会产生负数值，需要额外处理。本文不会在这方面花费篇幅，只选用线性插值法。

对于测试数据(即满足 $M_{ij}=0$，$R_{ij}=1$ 的 X_{ij})，我们知道缺失部分的实际值。因此，定义每个年份缺失数据的填补值和实际值相对误差：

$$e_i := \frac{\sum_{j:M_{ij}=0,R_{ij}=1}\left|X_{ij}-\hat{X}_{ij}\right|}{\sum_{j:M_{ij}=0,R_{ij}=1}\left|X_{ij}\right|} \tag{14}$$

其中 $\hat{X}_{ij}$ 为 X_{ij} 的填补值。根据这个定义，测试误差也称为*填补误差*。它比训练误差，即(7)或(11)的函数值，更能反映算法性能。

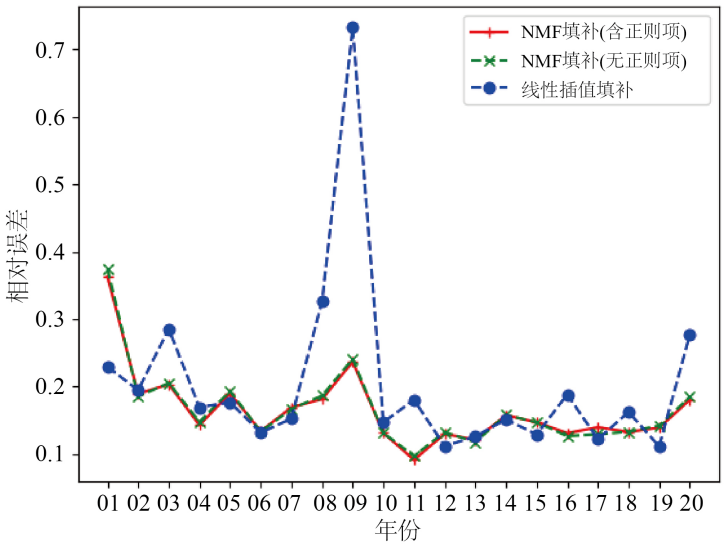


Figure 4. Comparison of the filling errors
图 4. 填补误差比较

上述程序将被运行 150 次。每个年份都算得 150 个误差，然后我们取其平均值作为最终的误差值。每次运行，随机缺失矩阵 M 都会被重新生成，避免出现样本偏差。图 4 表明，在多数年份，NMF 填补法的测试误差明显小于插值填补法的测试误差；只有在个别年份，NMF 填补法稍微劣于插值填补法；正则化和无正则化的 NMF 填补误差比较接近，但是通过计算均值，发现正则化确实起到了减少过拟合现象的作用。正则项系数设为 $\alpha_W = \alpha_H = 0.25$。

4.2.2. 总误差比较

下面用公式 $\sum_i e_i / N$ 计算总误差，并比较算法的总体性能。结果见表 2。正如前文所述，正则化减少了过拟合现象，降低了总误差。

Table 2. Comparison of the performance of the models (Note that the linear interpolation filling method has no learning process and no training error)
表 2. 模型整体性能对比(注意线性插值填补法没有学习过程，不存在训练误差)

模型/算法	训练总误差	测试总误差
NMF 填补(含正则项)	0.1152	0.1652
NMF 填补(无正则项)	0.1141	0.1663
线性插值填补	-	0.2049

5. 结束语

本文用 NMF 对国际贸易缺失数据进行填补。实验证明 NMF 填补法的性能显著优于插值填补法。本

文还增加了正则项，进一步降低了填补误差。

除了非负约束，NMF 的基 H 本身没有结构限制(至多只有正则项的限制)。H 的结构应该包含某种先验知识，即主题出口矩阵结构上的限制。

为了限制 H 的结构，可以考虑使用张量分解：

$$x_i = \sum_k w_{ik} \mathrm{vec}\left(h_k^{(1)} \circ h_k^{(2)}\right)$$

其中 $h_k^{(1)}$，$h_k^{(2)}$ 分别是 20，20 维向量，向量张量积运算 $a \circ b = \left\{a_i b_j\right\}$。也就是说基向量本质上是两个向量的外积。如果不对进出口数据做向量化处理，那么有矩阵 $x_i = \sum_k w_{ik} \mathrm{vec}\left(h_k^{(1)} \circ h_k^{(2)}\right)$，其张量表示为

$$X \sim \left[\left[W, H^{(1)}, H^{(2)}\right]\right] = \sum_k w_k \otimes h_k^{(1)} \otimes h_k^{(2)}$$

其中 W，$H^{(1)}$，$H^{(2)}$，分别是向量族$\{w_k\}$，$\left\{h_k^{(1)}\right\}$，$\left\{h_k^{(2)}\right\}$ 构成的矩阵。这就是著名的 CP 分解[14] [15] [16]。因为原始数据显然可以看作一个 3 阶张量，其中张量的三个维度(mode)分别为年份、出口国和入口国，所以我们预期可构造出基于张量分解的更有效的填补算法。这种分解方法不仅在数据的表示上更为自然，而且减少了参数数量，避免过拟合。

目前，NMF 填补法仅应用于 72 号货物的进出口数据。未来，我们会将其应用于所有进出口数据。利用张量分解方法，可以同时对所有数据进行处理。还可以考虑迁移学习的方法。将目前得到的结果迁移到其他数据。

此外，我们会重新考虑 X 的分布。本文假设 X 的元素服从 Poisson 分布。但也可改用 0-膨胀 Poisson 分布，非负二次分布等[7] [8]。还可以为 W，H 设置先验分布[17]，而先验分布起到正则项的作用。

基金项目

浙江省自然科学基金(LQ19F050004)。

参考文献

[1] Cichocki, A., Zdunek, R., Phan, A.H. and Amari, S. (2009) Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multiway Data Analysis and Blind Source Separation. John Wiley & Sons, Ltd., Hoboken. https://doi.org/10.1002/9780470747278

[2] Lee, D.D. and Seung, H.S. (2000) Algorithms for Non-Negative Matrix Factorization. *Neural Information Processing Systems* 2000, Vol. 13, 556-562.

[3] Lin, C.J. (2007) On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization. *IEEE Transactions on Neural Networks*, **18**, 1589-1596. https://doi.org/10.1109/TNN.2007.895831

[4] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning. 2nd Edition, Springer, Berlin, 44-49. https://doi.org/10.1007/978-0-387-84858-7

[5] Wang, Y.X. and Zhang, Y.J. (2013) Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge and Data Engineering*, **25**, 1336-1353. https://doi.org/10.1109/TKDE.2012.51

[6] Brouwer, T. (2017) Bayesian Matrix Factorisation: Inference, Priors, and Data Integration. University of Cambridge, Cambridge.

[7] Gouvert, O., Oberlin, T. and Fevotte, C. (2020) Negative Binomial Matrix Factorization. *IEEE Signal Processing Letters*, **27**, 815-819. https://doi.org/10.1109/LSP.2020.2991613

[8] Simchowitz, M. (2013) Zero-Inflated Poisson Factorization for Recommendation Systems. Princeton Department of Mathematics, Princeton. https://msimchowitz.github.io/JuniorPaper.pdf

[9] Poisson Distribution. https://en.wikipedia.org/wiki/Poisson_distribution

[10] Yoo, J. and Choi, S. (2009) Probabilistic Matrix Tri-Factorization. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, 19-24 April 2009, 1553-1556. https://doi.org/10.1109/ICASSP.2009.4959893

[11] Kim, Y. and Choi, S. (2009) Weighted Nonnegative Matrix Factorization. 2009 *IEEE International Conference on*

Acoustics, *Speech and Signal Processing*, Taipei, 19-24 April 2009, 1541-1544. https://doi.org/10.1109/ICASSP.2009.4959890

[12] Zhang, S., Wang, W., James, F.J. and Makedon, F. (2006) Learning from Incomplete Ratings Using Non-Negative Matrix Factorization. *Proceedings of the* 2006 *SIAM International Conference on Data Mining* (*SDM*), Bethesda, 20-22 April 2006, 549-553. https://doi.org/10.1137/1.9781611972764.58

[13] Buitinck, L., Louppe, G. and Blondel, M. (2013) API Design for Machine Learning Software: Experiences from the Scikit-Learn Project. *ECML PKDD Workshop*: *Languages for Data Mining and Machine Learning*, Prague, September 2013, 108-122.

[14] Harshman, R.A. (1970) Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multimodal Factor Analysis.

[15] Kim, J., He, Y. and Park, H. (2014) Algorithms for Nonnegative Matrix and Tensor Factorizations: A Unified View Based on Block Coordinate Descent Framework. *Journal of Global Optimization*, **58**, 285-319. https://doi.org/10.1007/s10898-013-0035-4

[16] Kolda, T.G. and Bader, B.W. (2009) Tensor Decompositions and Applications. *SIAM Review*, **51**, 455-500. https://doi.org/10.1137/07070111X

[17] Hoffman, M.D. (2012) Poisson-Uniform Nonnegative Matrix Factorization. *IEEE International Conference on Acoustics*, *Speech and Signal Processing*, Kyoto, 25-30 March 2012, 5361-5364. https://doi.org/10.1109/ICASSP.2012.6289132

附　录

乘法更新规则(12)的矩阵形式为，

$$H \leftarrow H \circ \left(W^{\mathrm{T}}\left(X \circ R \oslash WH\right)+\alpha\right) \oslash \left(W^{\mathrm{T}}R+\alpha\right)$$

$$W \leftarrow W \circ \left(\left(X \circ R \oslash WH\right)H^{\mathrm{T}}+\alpha\right) \oslash \left(RH^{\mathrm{T}}+\alpha\right)$$

其中$\circ$，$\oslash$分别是 Hadamard 乘法、除法运算，α为非负超参数。矩阵形式可以用 Python 数值计算库 numpy 轻松实现。在矩阵形式上加入正则项也是非常容易的。本文源代码就是基于这个形式。读者可以根据自己的风格和掌握的计算机语言实现或改进乘法更新规则。