

作者	生姜 DrGinger
脚本	生姜 DrGinger
视频	崔崔 CuiCui
开源学习资源	https://github.com/Visualize-ML
平台	https://www.youtube.com/@DrGinger_Jiang https://space.bilibili.com/3546865719052873 https://space.bilibili.com/513194466

12.3 再聊主成分分析



本节你将掌握的核心技能：

- ▶ 判断是否需要标准化，依据是各特征标准差是否差异过大。
- ▶ 标准化后计算相关系数矩阵。
- ▶ 相关系数矩阵的特征值分解。
- ▶ 第一主成分解释最多方差信息。
- ▶ 区分载荷、因子得分。
- ▶ 每个主成分解释的方差比例及累计比例，用来评估降维效果。
- ▶ 前 p 个主成分近似还原原始数据的操作步骤与几何意义。

上一节从椭圆视角聊了聊主成分分析，相信大家已经理解把数据看作一个旋转椭圆的话，主成分分析就是找到合适的方向把旋转椭圆摆正。本节将采用真实数据和大家一起从应用角度再聊主成分分析。

数据

图 1 所示为不同年限 (比如, 0.5 年、1 年、2 年 ...) 利率每天涨跌数据。

⚠ 注意，图 1 中数据乘 100% 就是每天涨跌的百分比。



LA_12_03_01.ipynb 为本节配套代码，所有运算和可视化都在其中，请大家一边阅读本节内容，一边在 JupyterLab 实践。

查看 LA_12_03_01.ipynb，大家会发现数据矩阵 X 有 248 行、8 列，数据矩阵 X 是一个细高矩阵。

数据矩阵 X 的每一行（行向量）代表当天和上一天之间的涨跌， X 的每一列代表一个年限。

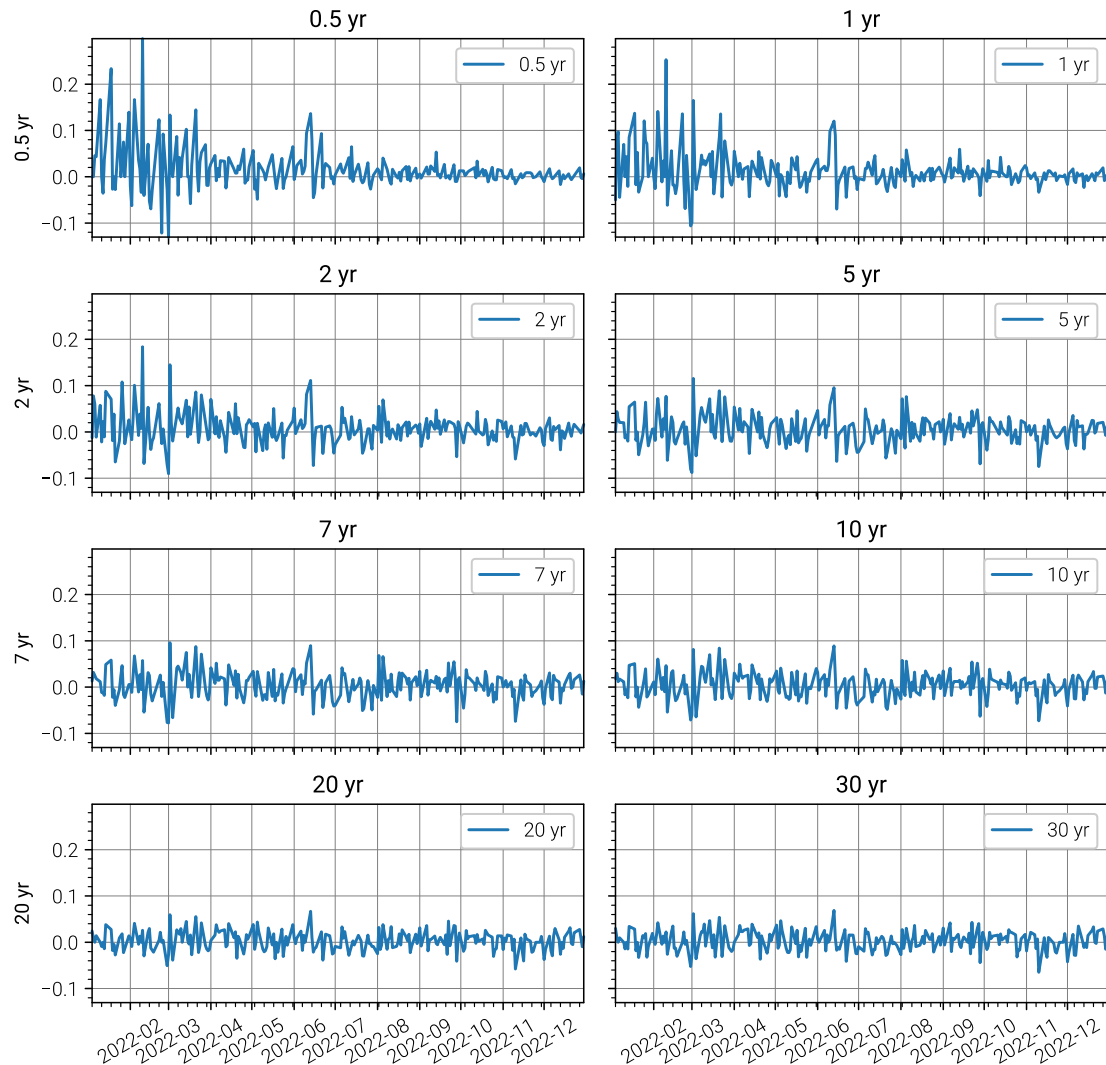


图 1. 不同年限利率每天涨跌线图

图 1 这幅图很难看到不同年限涨跌是否存在某种趋势，我们还绘制了图 2。从图 2 这幅成对散点图中，我们可以看到不同年限利率涨跌存在很强的线性正相关。

下面，我们试着用主成分分析发现其中的规律。

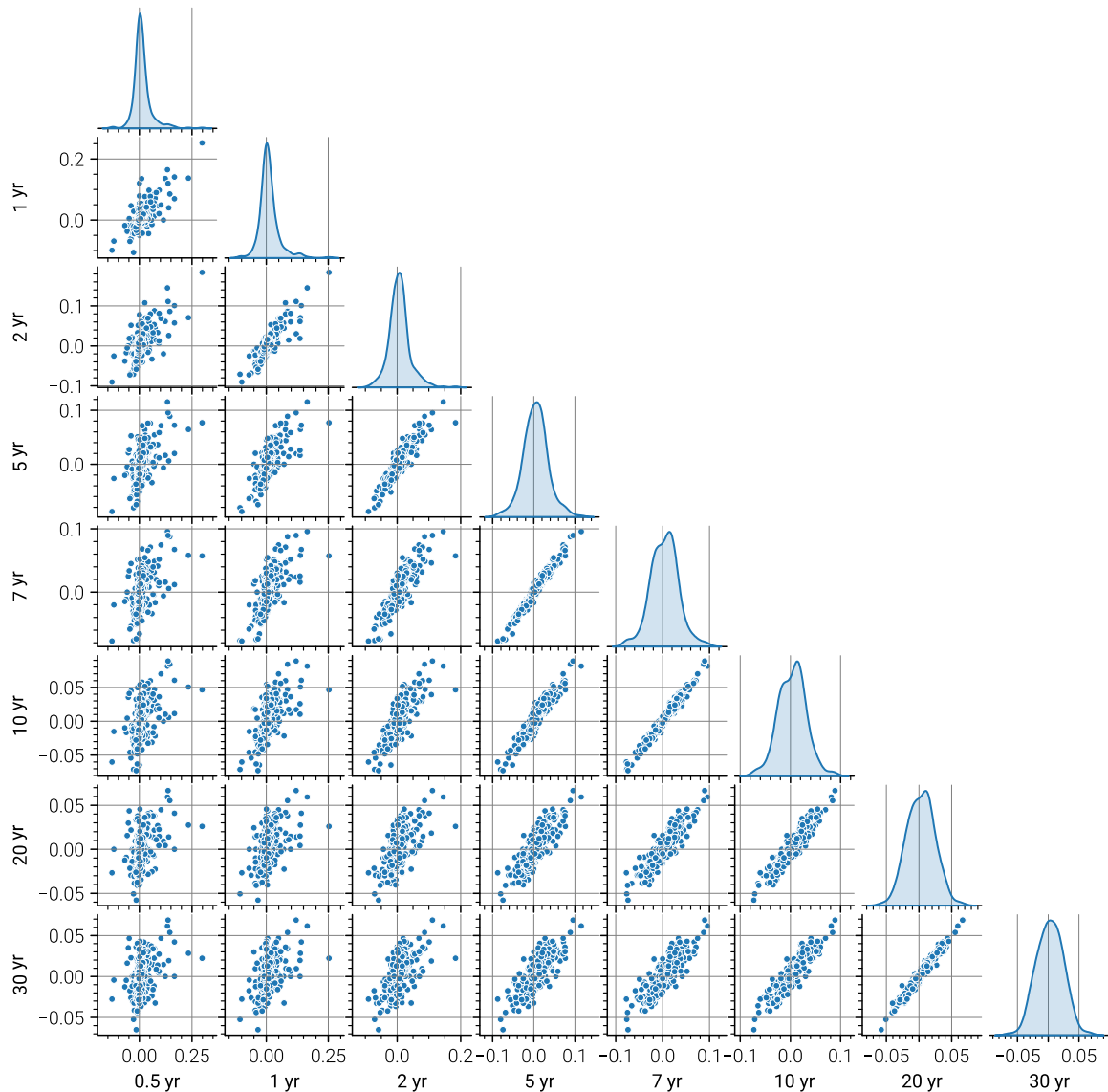


图 2. 不同年限利率每天涨跌成对散点图，不展示主对角线以上子图

协方差矩阵

数据矩阵 \mathbf{X} 中心化后得到 \mathbf{X}_c 。

$$\mathbf{X}_c = \mathbf{X} - \boldsymbol{\mu}^T \quad (1)$$

其中， $\boldsymbol{\mu}$ 为 \mathbf{X} 质心。

本章前文提过，协方差矩阵相当于 \mathbf{X}_c 的一种特殊格拉姆矩阵

$$\boldsymbol{\Sigma} = \frac{\mathbf{X}_c^T \mathbf{X}_c}{n-1} \quad (2)$$

图 3 所示为数据矩阵的协方差矩阵。

本章第 1 节介绍过协方差矩阵的主对角线元素为方差，平方根为标准差；协方差矩阵非主对角线元素为协方差。

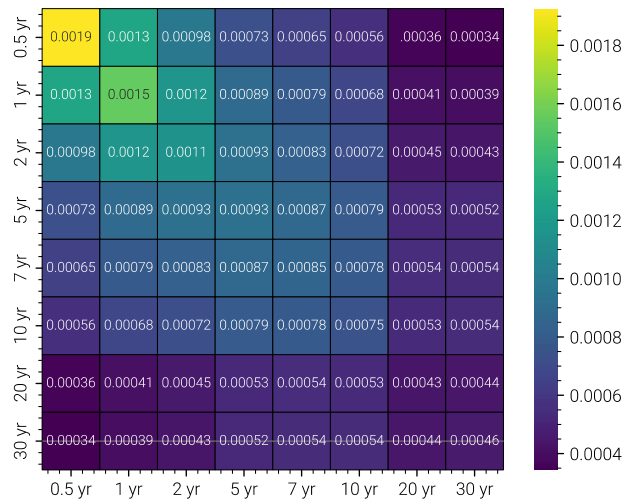


图 3. 利率涨跌数据的协方差矩阵热图

如图 4 所示，各个期限利率涨跌数据的标准差（波动率）相差很大。上一节提过，数据主成分分析时，如果不同特征的标准差差异过大，需要先对数据标准化。

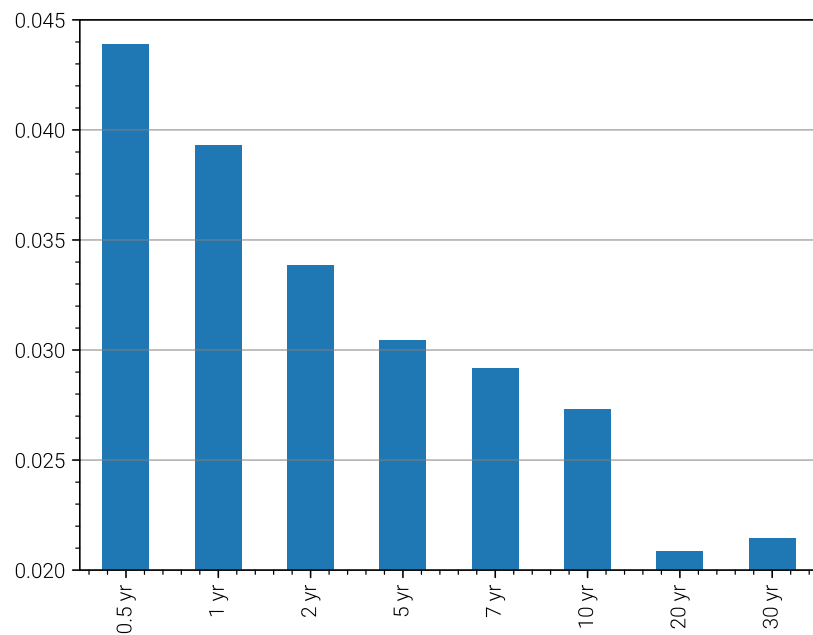


图 4. 标准差柱状图

标准化

标准化数据矩阵 \mathbf{Z} 为

$$\mathbf{Z} = \mathbf{X}_c \mathbf{D}^{-1} = (\mathbf{X} - \boldsymbol{\mu}^T) \mathbf{D}^{-1} \quad (3)$$

其中， \mathbf{D} 为对角方阵，主对角线元素为标准差

$$\mathbf{D} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix} \quad (4)$$

图 5 所示为标准化后的线图。

本书前文提过，标准化数据的最大特点是消除了原始特征的单位，使不同量纲、不同尺度的数据可以在同一个尺度上进行比较。

如图 5 所示，经过标准化处理后，每个子图的数据的均值变为 0，标准差变为 1。

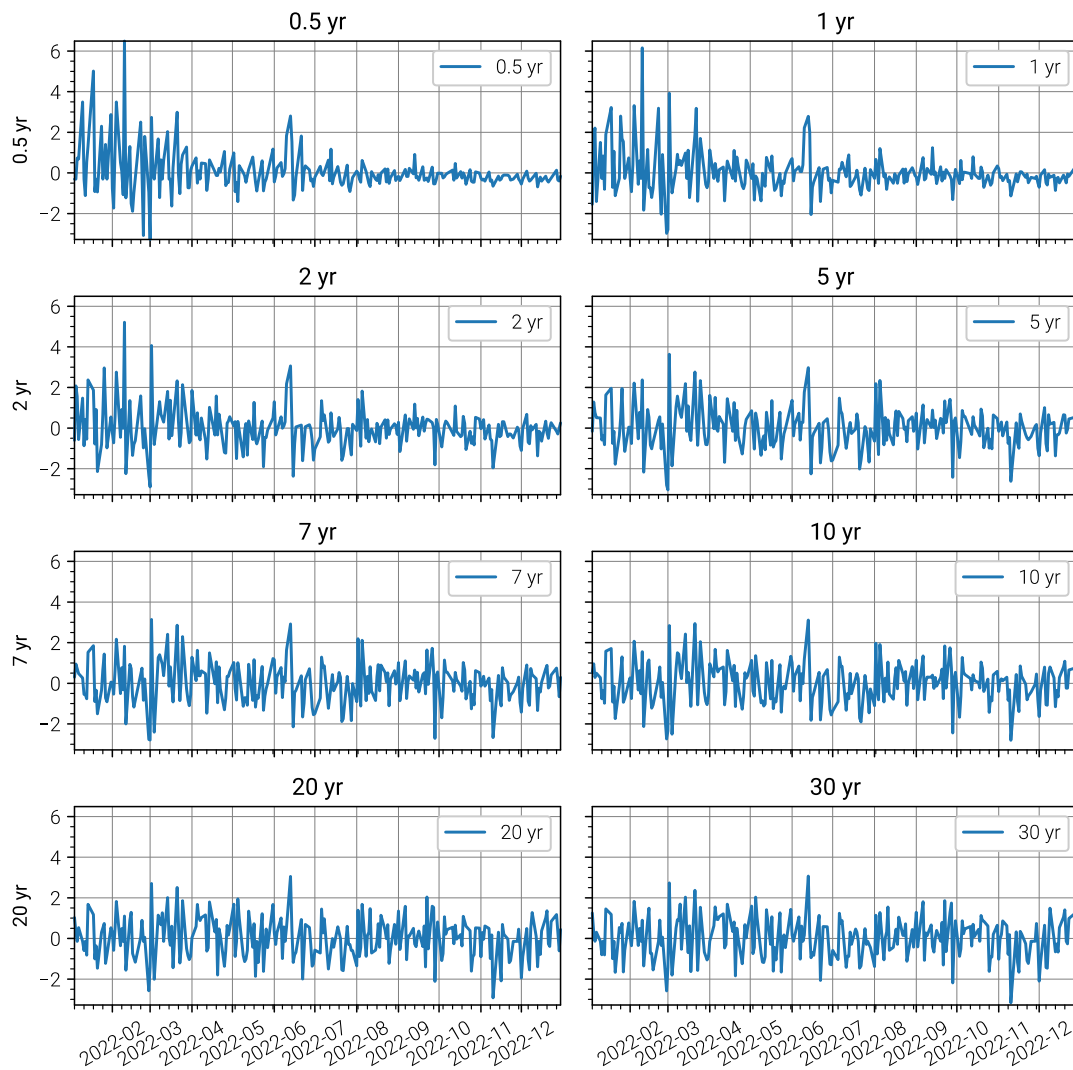


图 5. 标准化数据 (均值为 0, 标准差为 1) 线图

图 6 所示为标准化后数据的成对散点图。

比较图 2、图 6，我们会发现标准化不会改变数据的分布形状，但会对中心位置和离散程度进行统一，保留了样本之间的相对位置和结构。

标准化后的数据不再依赖原始的度量单位。

⚠ 需要注意的是，标准化对异常值较敏感，因为均值和标准差本身容易受到极端值的影响。

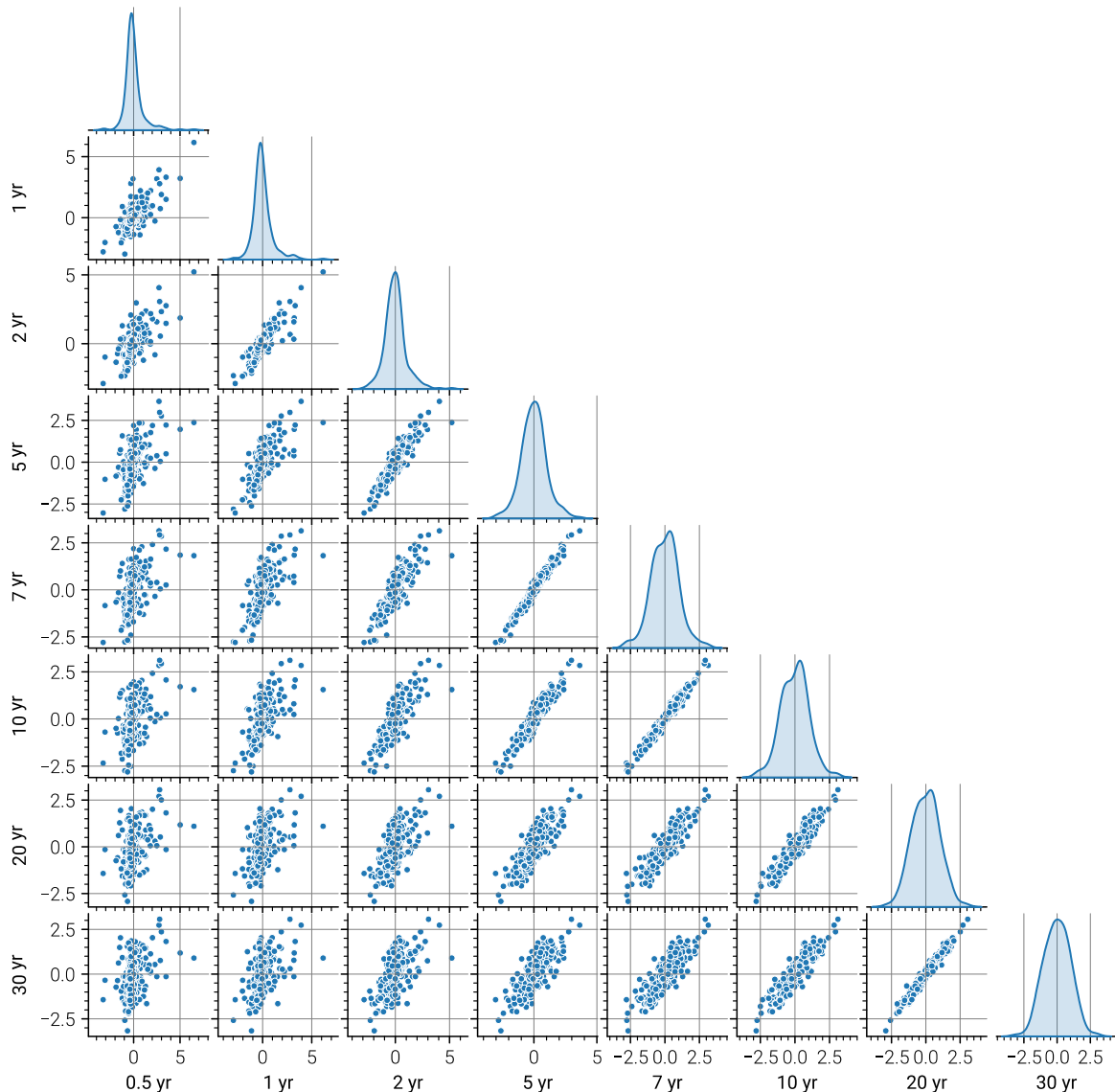


图 6. 标准化数据成对散点图，不展示主对角线以上子图

线性相关性系数矩阵

前文提过，线性相关性系数是用来衡量两个变量之间线性关系强弱和方向的一个数值，取值范围在 -1 到 1 之间。 1 表示完全正相关， -1 表示完全负相关， 0 表示无线性关系。

线性相关性系数反映的是两个变量在去除单位、标准化之后的变化趋势是否一致，因此不受变量原始单位和尺度的影响。

如图 7 所示，线性相关性系数矩阵相当于汇总了多个特征之间两两线性相关性系数

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,D} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D} & \rho_{2,D} & \cdots & 1 \end{bmatrix} \quad (5)$$

线性相关性系数矩阵的每一个元素代表了相应两个特征之间的线性相关程度。对角线上的值恒为 1，因为每个特征和自身完全正相关。

线性相关性系数矩阵相当于标准化数据的协方差矩阵，即

$$\mathbf{P} = \Sigma_z = \frac{\text{Gram matrix } \mathbf{Z}^T \mathbf{Z}}{n-1} \quad (6)$$

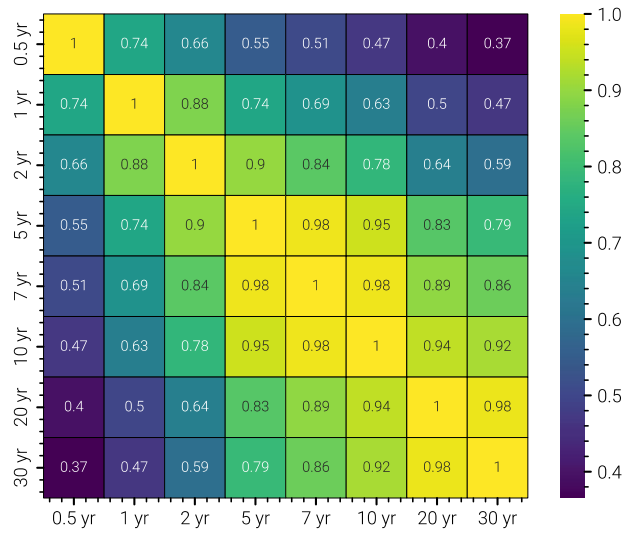


图 7. 线性相关性系数矩阵

从正交投影角度来看， \mathbf{Z} 在 \mathbf{e}_1 投影的得到数据 z_1 为

$$\mathbf{z}_1 = \mathbf{Z}\mathbf{e}_1 = \begin{bmatrix} z_1 & z_2 & \cdots & z_D \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (7)$$

计算 z_1 方差

$$\text{var}(z_1) = \frac{\mathbf{z}_1^T \mathbf{z}_1}{n-1} = \frac{(\mathbf{Z}\mathbf{e}_1)^T \mathbf{Z}\mathbf{e}_1}{n-1} = \mathbf{e}_1^T \frac{\mathbf{Z}^T \mathbf{Z}}{n-1} \mathbf{e}_1 = \mathbf{e}_1^T \mathbf{P} \mathbf{e}_1 = 1 \quad (8)$$

上式相当于取出 \mathbf{P} 的第 1 行、第 1 列元素，这意味着 z_1 的方差为 1。

⚠ 注意， \mathbf{Z} 的每一列已经中心化（去均值），因此 (8) 中计算方差时，不必重复中心化。

类似地， \mathbf{Z} 在 \mathbf{e}_1 投影的得到数据 z_2

$$\mathbf{z}_2 = \mathbf{Z}\mathbf{e}_2 \quad (9)$$

z_1 、 z_2 的协方差为

$$\text{cov}(z_1, z_2) = \text{cov}(z_2, z_1) = \frac{z_2^T z_1}{n-1} = \frac{(Ze_2)^T Ze_1}{n-1} = e_2^T \frac{Z^T Z}{n-1} e_1 = e_2^T P e_1 = \rho_{1,2} \quad (10)$$

特征值分解

图 8 所示为线性相关性系数矩阵 P 的特征值分解，即

$$P = VAV^T \quad (11)$$

线性相关性系数矩阵为对称矩阵，所以对线性相关性系数矩阵的特征值分解是谱分解。图 8 中 V 为正交矩阵，即规范正交基。

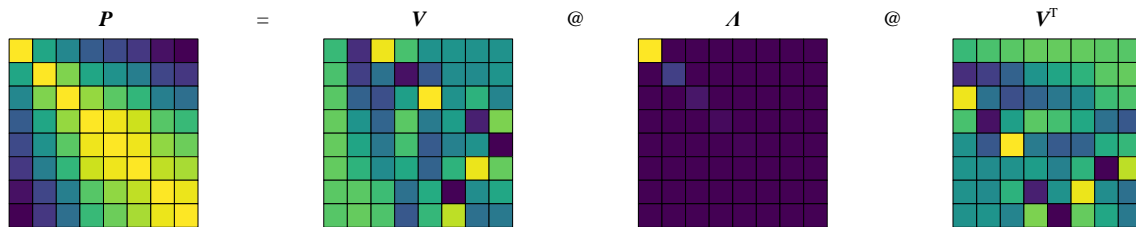


图 8. 线性相关性系数矩阵的特征值分解

线性相关性系数矩阵 P 的对角化可以写成

$$A = V^T P V \quad (12)$$

将 V 写成行向量展开 (12) 得到

$$\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix} = \begin{bmatrix} v_1^T P v_1 & v_1^T P v_2 & \cdots & v_1^T P v_d \\ v_2^T P v_1 & v_2^T P v_2 & \cdots & v_2^T P v_d \\ \vdots & \vdots & \ddots & \vdots \\ v_d^T P v_1 & v_d^T P v_2 & \cdots & v_d^T P v_d \end{bmatrix} \quad (13)$$

从正交投影角度来看， Z 在 v_1 投影的得到数据 y_1 为

$$y_1 = Z v_1 \quad (14)$$

而 y_1 方差为

$$\text{var}(y_1) = \frac{y_1^T y_1}{n-1} = \frac{(Z v_1)^T Z v_1}{n-1} = v_1^T \frac{Z^T Z}{n-1} v_1 = v_1^T P v_1 = \lambda_1 \quad (15)$$

也就是说， v_1 (第一主成分方向) 方向上， Z 的投影结果方差最大；也就是说， v_1 解释了数据中最多的方差。

在主成分分析中， V 的列向量也叫载荷 (loading)。

注意，很多文献中，载荷的另一种是将 V 的列向量乘以对应主成分的特征值的平方根，比如第一主成分的载荷为

$$v_1 \sqrt{\lambda_1} \quad (16)$$

因子得分 (score) 表示每个样本在主成分坐标系下的投影坐标, 即 y_i 。

因子得分构成的矩阵对应如下乘法

$$Y = ZV \quad (17)$$

也就是说, Z 朝 V 正交投影得到 Y 。

? 请画出 (17) 对应的矩阵运算热图。

将 V 写成列向量, 展开 (17) 得到

$$[y_1 \ y_2 \ \cdots \ y_D] = [Zv_1 \ Zv_2 \ \cdots \ Zv_D] \quad (18)$$

而 Y 的协方差矩阵就是 (11) 谱分解得到的对角方阵, 即

$$\Sigma_Y = \frac{Y^T Y}{n-1} = \frac{(ZV)^T ZV}{n-1} = V^T \frac{Z^T Z}{n-1} V = V^T P V = A \quad (19)$$

? 请画出 (19) 对应的矩阵运算热图。

特征值

线性相关性系数矩阵 P 的主对角线之和为 P 的迹, 即 $\text{trace}(P) = 8$ 。这相当于标准化数据 Z 的方差总和。

前文提过, A 的迹也是 8, 即

$$\text{trace}(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_8 = 8 \quad (20)$$

图 9 所示为特征值从大到小排列。

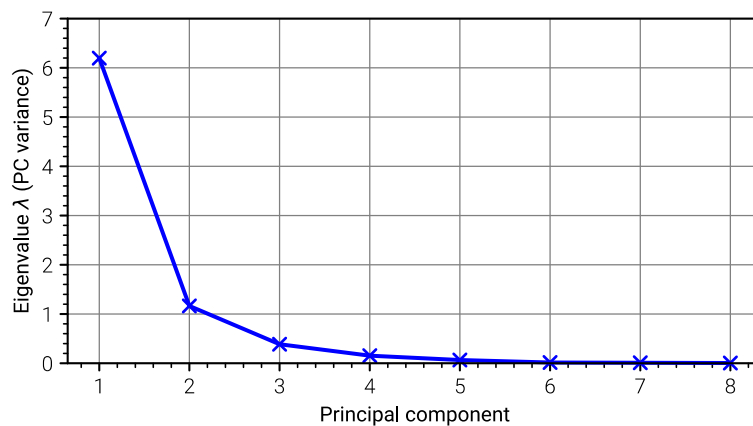


图 9. 特征值从大到小排列

第 j 个特征值 λ_j 对方差总和的贡献百分比为:

$$\frac{\lambda_j}{\sum_{j=1}^D \lambda_j} 100\% \quad (21)$$

上式分母是数据总方差，即 $\text{trace}(\mathbf{P}) = \text{trace}(\mathbf{A})$ 。

(21) 这个比值可以用来衡量第 j 个主成分对数据的解释能力。

前 p 个特征值累积解释总方差百分比 (cumulative percentage of explained variance) 为：

$$\frac{\sum_{j=1}^p \lambda_j}{\sum_{j=1}^D \lambda_j} 100\% \quad (22)$$

如图 10 所示，这个百分比代表前 p 个主成分所能解释的已释方差之和占有所有主成分已释方差之和的比例。累计已释方差和百分比能够用来评估 PCA 的降维效果，它衡量了前 p 个主成分能够解释数据方差的比例。

通常来说，我们希望通过选择适当的主成分数 p ，使累计已释方差和百分比达到预设的阈值（比如 80% 或 90%），以保留尽可能多的原始数据信息。

我们发现第一主成分已经解释接近 80% 的总方差，

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \dots + \lambda_8} \times 100\% = 77.477\% \quad (23)$$

而前两个主成分解释的总方差百分比超过 90%

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \dots + \lambda_8} \times 100\% = 92.038\% \quad (24)$$

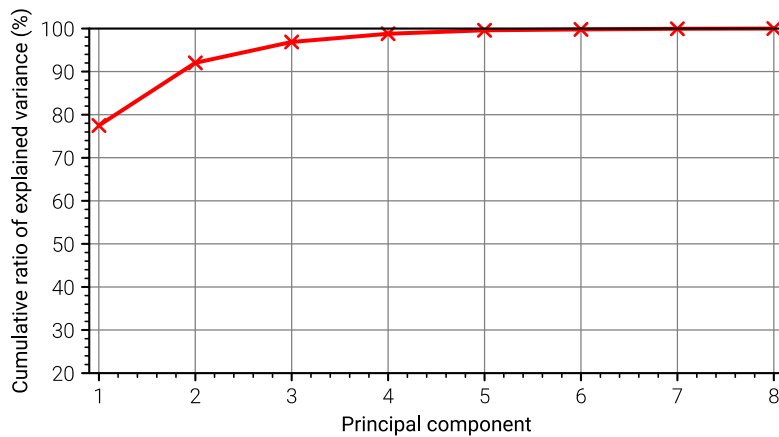


图 10. 前 p 个特征值累积解释总方差百分比

近似还原

如果仅用第一主元信息近似还原 \mathbf{Z} (还不是 \mathbf{X})，对应矩阵乘法

$$\hat{\mathbf{Z}} = \mathbf{Z}_1 = \mathbf{y}_1 @ \mathbf{v}_1^T = \mathbf{Z} @ (\mathbf{v}_1 @ \mathbf{v}_1^T) \quad (25)$$

上式相当于 Z 仅在 v_1 方向的正交投影。

Z 和 Z_1 之间的误差为

$$Z - Z_1 = Z - Z @ (v_1 @ v_1^T) = Z(I - v_1 @ v_1^T) \quad (26)$$

这也是一个正交投影。

为了用 Z_1 近似还原 X ，我们还需要“缩放 → 平移”

$$\hat{X} = Z_1 D + \mu^T = (y_1 @ v_1^T) D + \mu^T = (Z @ (v_1 @ v_1^T)) D + \mu^T \quad (27)$$

也就是说，上式代表：

- ▶ Z_1 近似还原 Z (标准化数据);
- ▶ 然后通过对角方阵 D (对角线元素为标准差) 缩放，近似还原中心化数据 X_c ;
- ▶ 最后再用质心 μ^T 平移 (广播原则)，近似原始数据矩阵 X 。

把 (3) 带入 (27) 得到

$$\hat{X} = ((X - \mu^T) D^{-1} @ (v_1 @ v_1^T)) D + \mu^T \quad (28)$$

上式对应的几何操作为“平移 → 缩放 → 投影 → 缩放 → 平移”。具体来说，

- ▶ 平移：数据中心化。
- ▶ 缩放：数据标准化。
- ▶ 投影：标准化数据朝第一主元 v_1 投影。
- ▶ 缩放：第二步逆操作。
- ▶ 平移：第一步逆操作。

用第一主元信息近似还原 X 的误差为

$$X - \hat{X} = X - ((X - \mu^T) D^{-1} @ (v_1 @ v_1^T)) D + \mu^T \quad (29)$$

图 11 中，蓝色曲线为原始数据、橘色曲线为第一主元信息还原得到的近似原始数据；而黑色代表误差。

图 12 所示为第一主元信息近似还原 X 数据的成对散点图；每幅子图的散点位于同一条直线上，这是因为我们仅用一个主元信息还原原始数据。

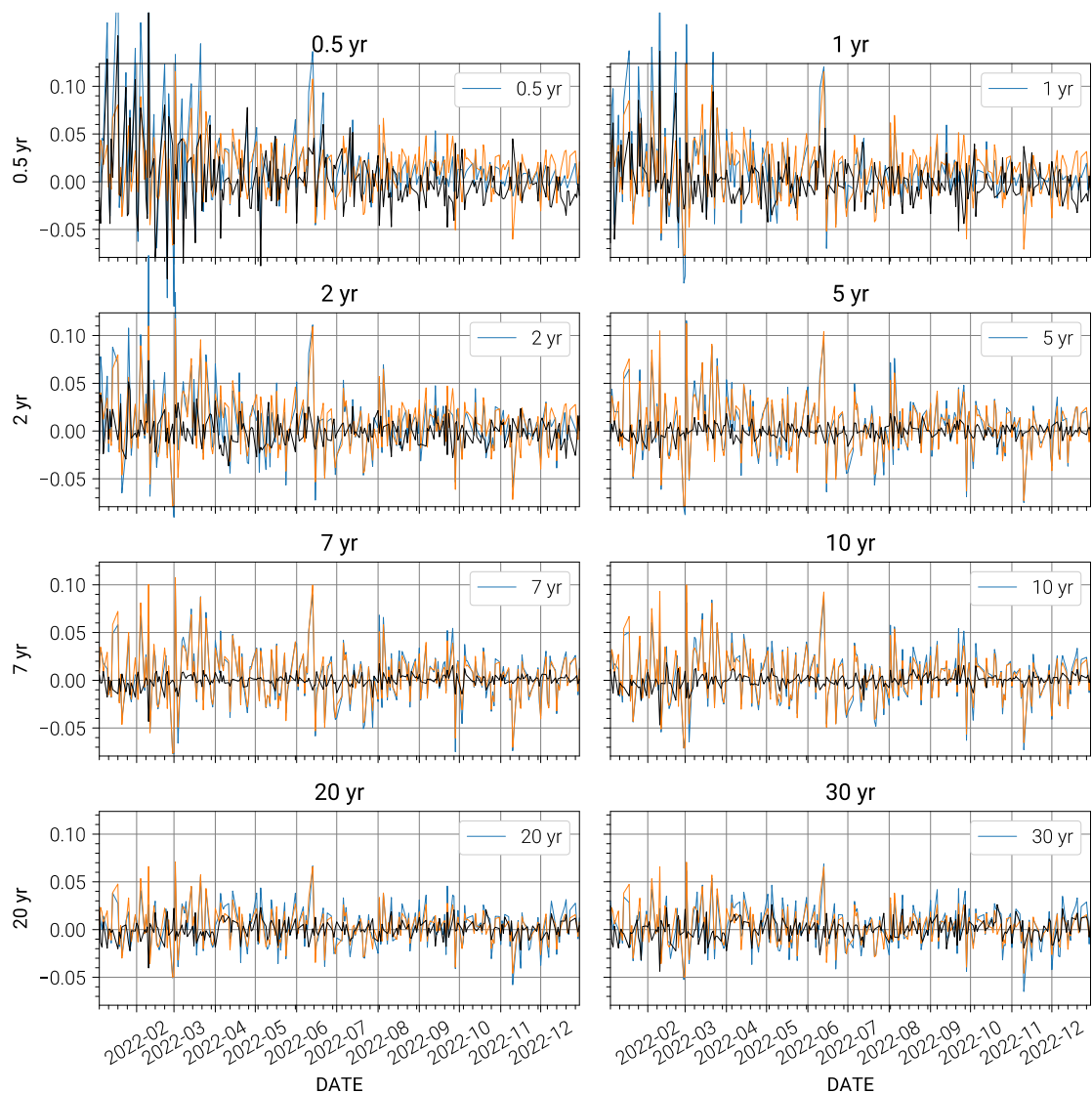


图 11. 用线图比较原始数据、近似还原数据，第一主元

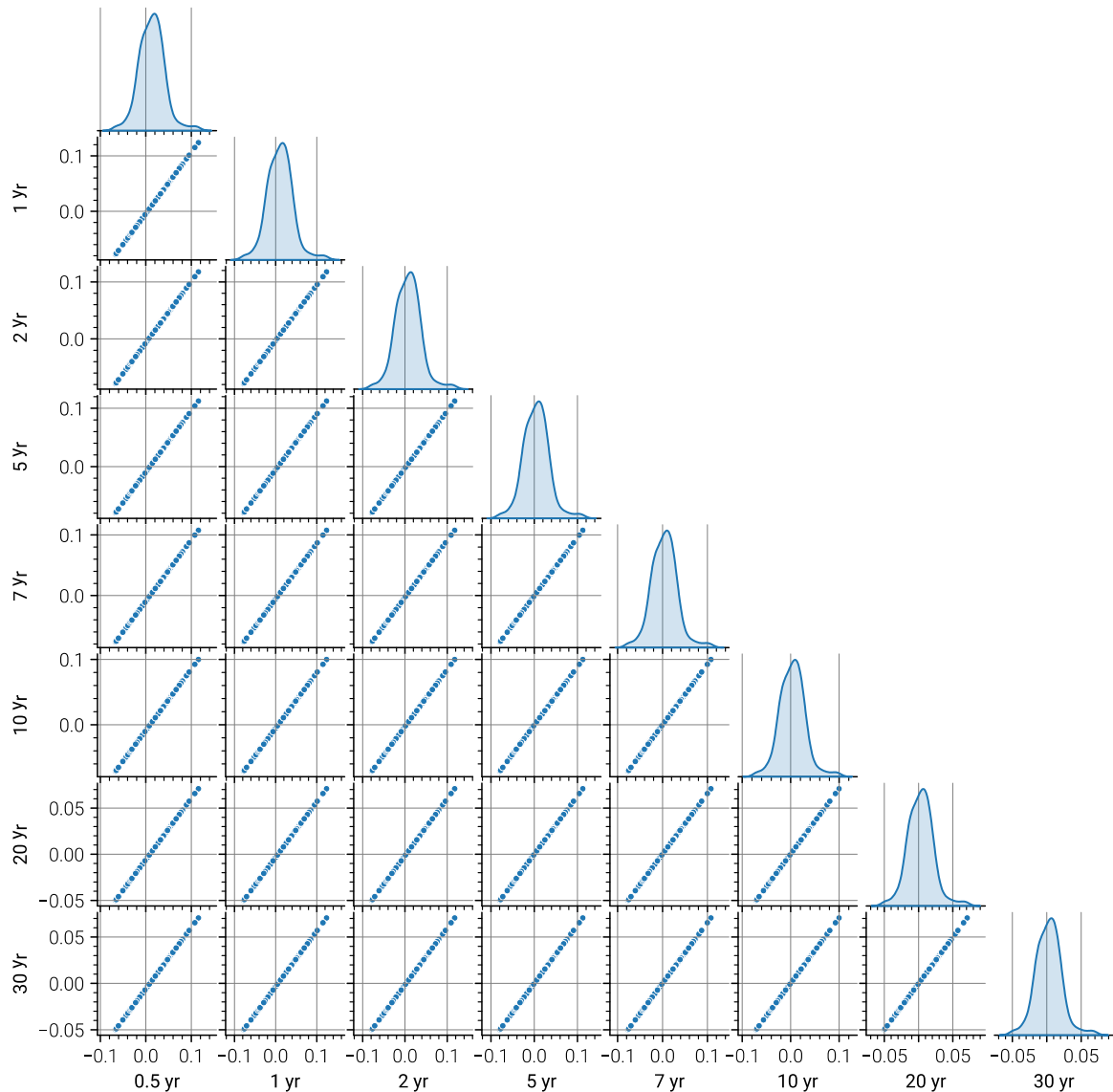


图 12. 第一主元近似还原数据成对散点图，不展示主对角线以上子图

为了比较原数据与近似还原数据之间的相似程度，我们绘制如图 13 所示的散点图。

图 13 中，近似还原的数据作为横轴，原始数据作为纵轴。如果还原效果越好，说明每个点的横纵坐标越接近，也就是说，它们在图上会越接近于一条对角线 $x_2 = x_1$ 。因此，我们在图中添加 $x_2 = x_1$ 这条参考线（用红色标出），以此作为“完美还原”的标准线。若大多数散点聚集在这条红线附近，表示近似还原与原始数据高度相似。这个图形化手段可以直观判断还原质量。

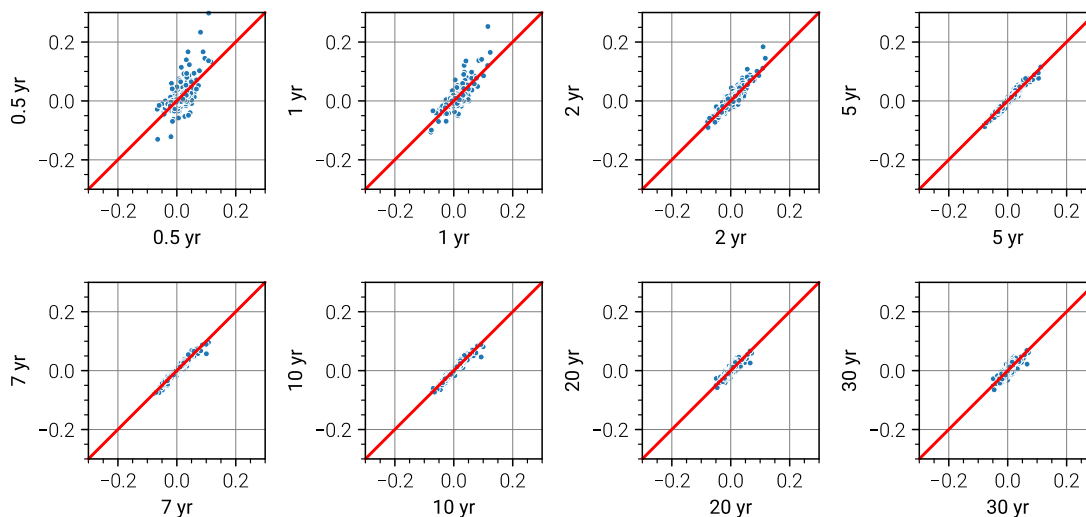


图 13. 用散点图比较原始数据、近似还原数据，第一主元

用前两个主元信息近似还原 \mathbf{X} ，对应的矩阵运算为

$$\hat{\mathbf{X}} = (\mathbf{Z}_1 + \mathbf{Z}_2)\mathbf{D} + \boldsymbol{\mu}^T = (\mathbf{y}_1 @ \mathbf{v}_1^T + \mathbf{y}_2 @ \mathbf{v}_2^T)\mathbf{D} + \boldsymbol{\mu}^T = \mathbf{Z} @ (\mathbf{v}_1 @ \mathbf{v}_1^T + \mathbf{v}_2 @ \mathbf{v}_2^T)\mathbf{D} + \boldsymbol{\mu}^T \quad (30)$$

? 请思考 (30) 对应的几何操作。并根据 (29) 计算用前两个主元信息近似还原 \mathbf{X} 误差。

图 14、图 15、图 16 三幅图展示的是“第一 + 第二”主元近似还原原始数据的可视化，请大家自行分析。

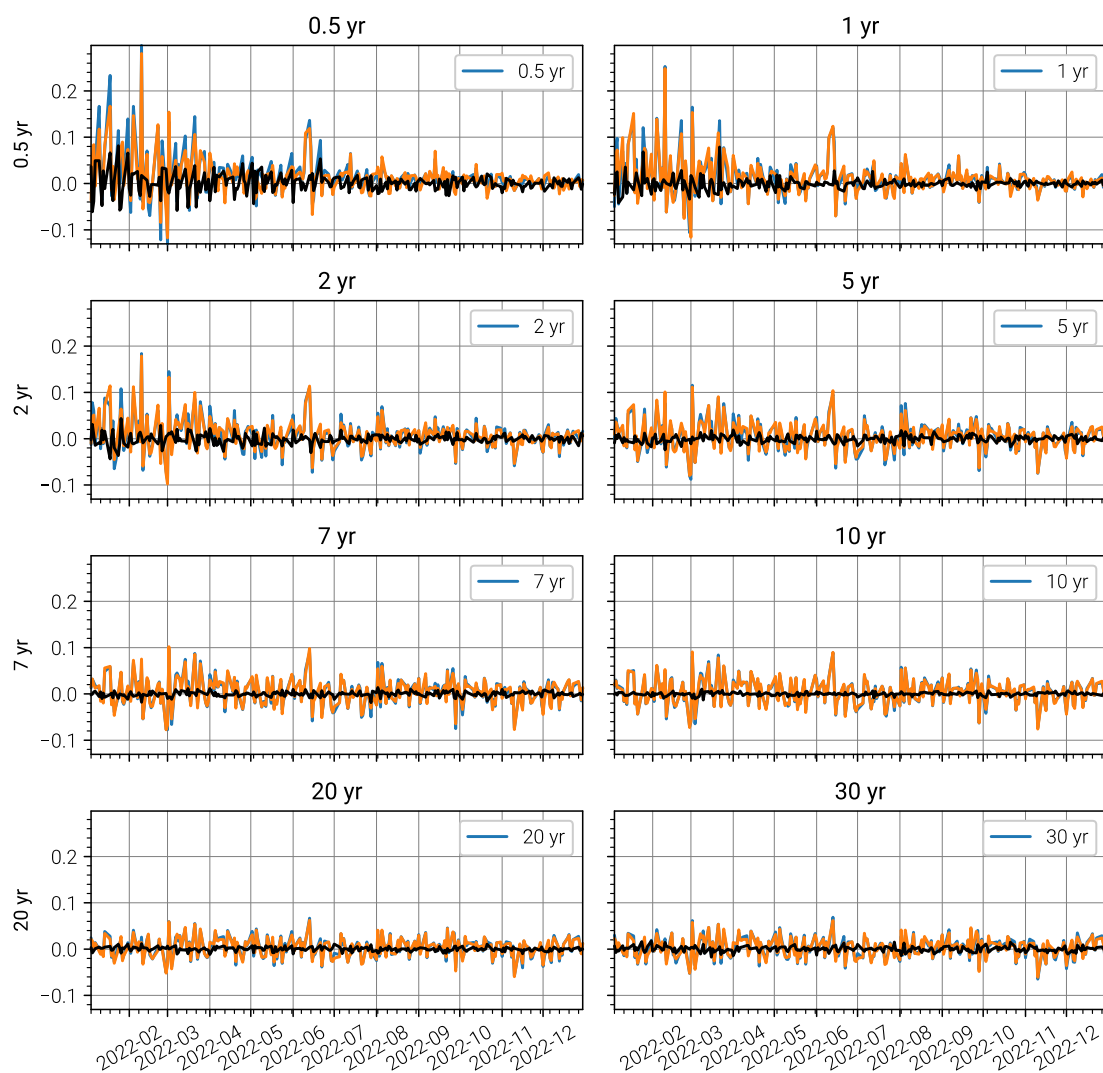


图 14. 用线图比较原始数据、近似还原数据, “第一 + 第二”主元

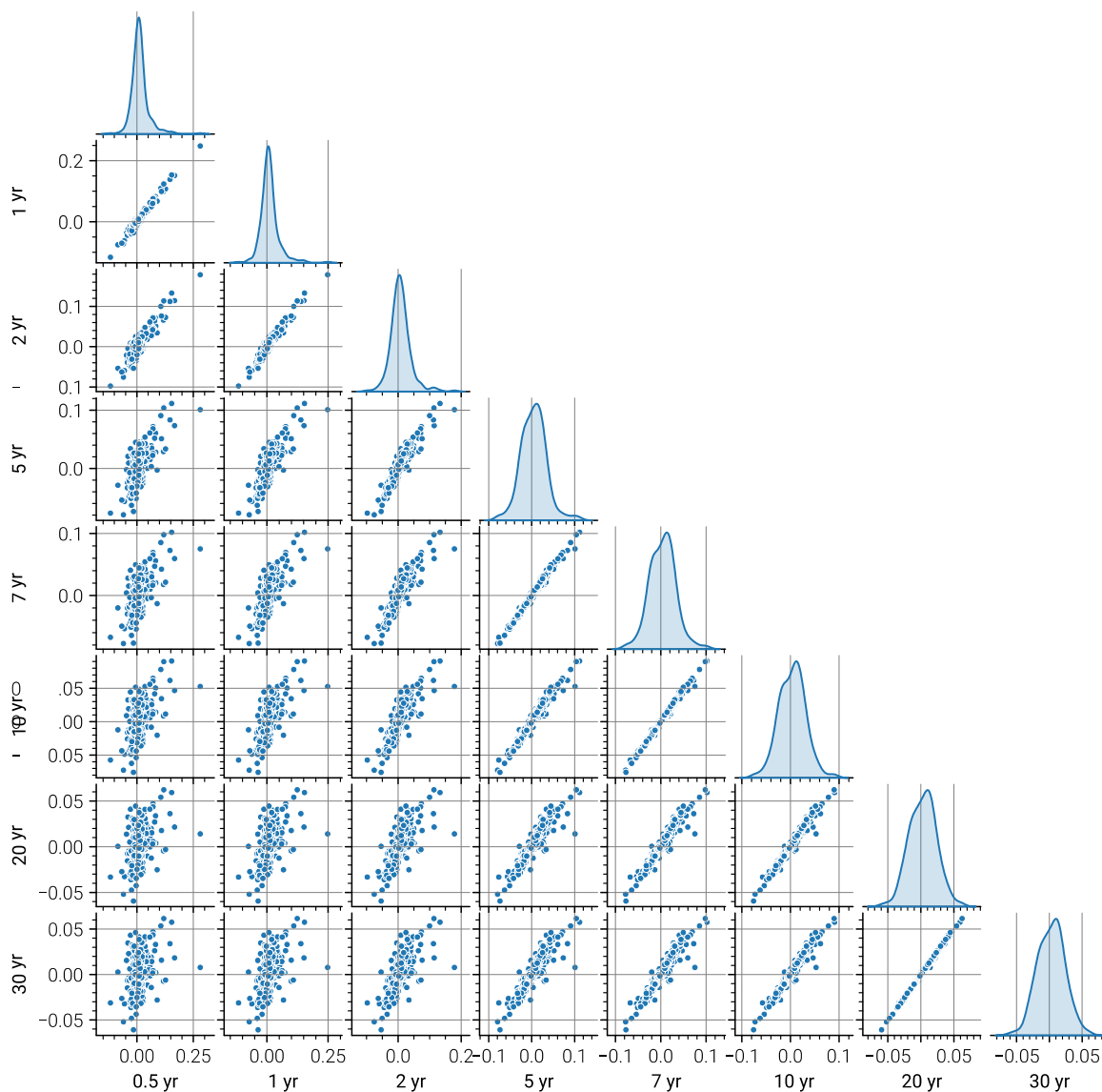


图 15. 用散点图比较原始数据、近似还原数据, “第一 + 第二”主元

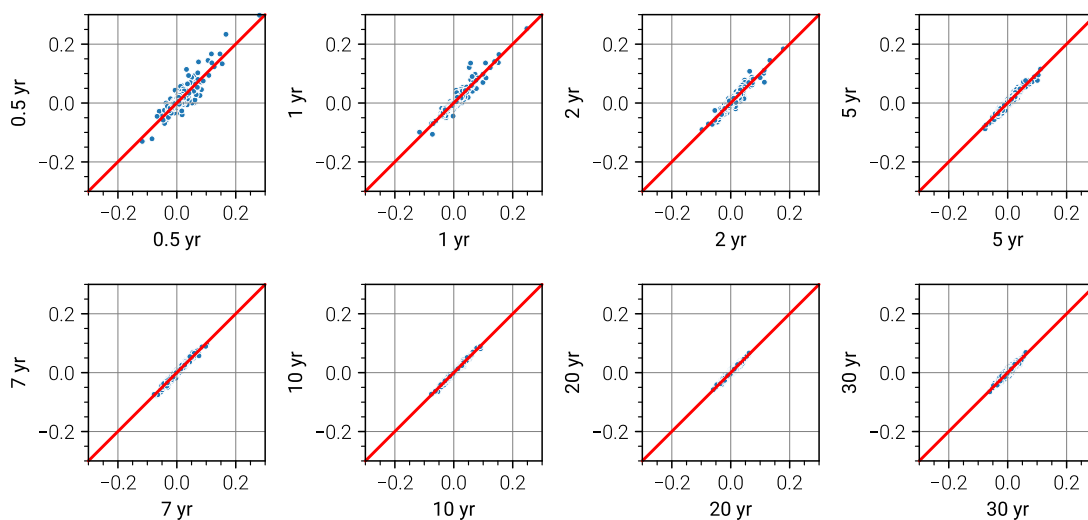


图 16. 用散点图比较原始数据、近似还原数据, “第一 + 第二”主元

本 PDF 文件为作者草稿, 发布目的为方便读者在移动终端学习, 终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有, 请勿商用, 引用请注明出处。

代码及 PDF 文件下载: <https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教, 本书专属邮箱: jiang.visualize.ml@gmail.com



LA_12_03_02.ipynb 用 `statsmodels.multivariate.pca()` 完成主成分分析。

LA_12_03_01.ipynb 相当于复刻了 LA_12_03_02.ipynb 的结果。请大家自学 LA_12_03_02.ipynb。请注意，特征向量符号差异。



请大家用 DeepSeek/ChatGPT 等工具完成本节如下习题。

Q1. 请修改 LA_12_03_01.ipynb，用前 3、4 个主元信息近似还原原始数据。

Q2. 请把 LA_12_03_01.ipynb 所有可视化方案用在 LA_12_03_02.ipynb。

Q3. 请学习使用 Scikit-learn 中 PCA 函数完成本节示例：

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>