第十三章 线性因子模型

典型的线性因子模型

- 因子分析
- 概率PCA
- 主成分分析（PCA）
- 独立成分分析（ICA）
- 慢特征分析
- 稀疏编码

PCA的流形解释

第十三章 线性因子模型

线性因子模型通过随机线性解码器函数来定义，该函数通过对h的线性变换以及添加噪声来生成x

数据生成过程：

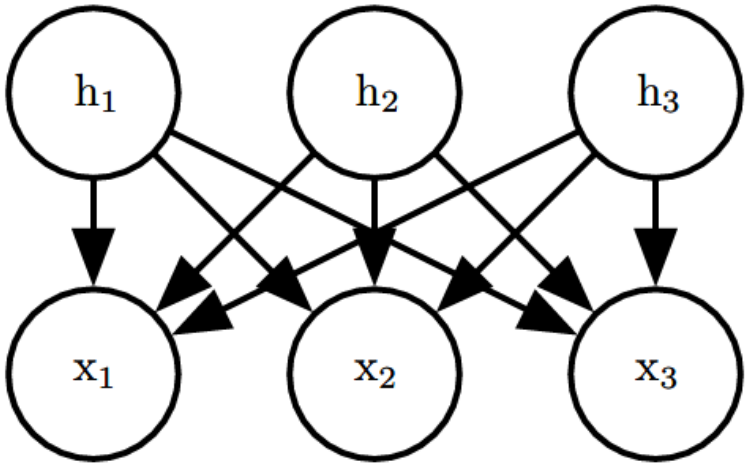
- 从一个分布中抽取解释性因子h

$$\mathbf{h} \sim p(\boldsymbol{h})$$

- 然后在给定因子的情况下，对实值的可观察变量进行采样

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} + noise$$

- *noise*通常是对角化的，且服从高斯分布



$$\mathbf{x} = \boldsymbol{W}\mathbf{h} + \boldsymbol{b} + \mathbf{noise}$$

因子分析

潜变量的先验分布$p(\boldsymbol{h})$是一个方差为单位矩阵的高斯分布

$$\mathbf{h} \sim \mathcal{N}(\boldsymbol{h}; \mathbf{0}; \boldsymbol{I})$$

- 设噪声是从对角协方差矩阵的高斯分布中抽出的，协方差矩阵为

$$\psi = \text{diag}(\boldsymbol{\sigma}^2)$$

- 潜变量的作用是捕获不同观测变量x_i之间的依赖关系

- x 服从多维正态分布

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{b}, \boldsymbol{W}\boldsymbol{W}^T + \psi)$$

概率PCA

潜变量的先验分布$p(\boldsymbol{h})$是一个方差为单位矩阵的高斯分布

$$\mathbf{h} \sim \mathcal{N}(\boldsymbol{h}; \mathbf{0}, \boldsymbol{I})$$

- 条件方差σ_i^2为同一个值，这样$\boldsymbol{x}$的协方差就变为
$$\boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I}$$
- x 的条件分布
$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{b}, \boldsymbol{W}\boldsymbol{W}^T + \sigma^2 \boldsymbol{I})$$
- 等价地
$$\mathbf{x} = \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} + \sigma \boldsymbol{z}$$
- 其中
$$\mathbf{z} \sim \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \boldsymbol{I})$$

- $\sigma \to 0$时，概率PCA 退化为PCA

ICA

将观察到的信号分离成许多潜在信号，这些潜在信号通过缩放和叠加可以恢复成观察数据

- 动机：
  - 通过选择一个独立的$p(\boldsymbol{h})$，我们可以尽可能恢复接近独立的潜在因子

- ICA 的所有变种均要求$p(\boldsymbol{h})$ 是非高斯的
  - 这样的选择在 0 附近具有比正态分布更高的峰值
  - ICA 常用于学习稀疏特征

- 变种
  - 非线性独立成分估计
    - 能高效地计算每个变换的Jacobian 行列式
  - 独立子空间分析
    - 鼓励组内统计依赖关系、抑制组间依赖关系来学习特征组

慢特征分析

使用来自时间信号的信息学习不变特征的线性因子模型

基本思想

- 场景的重要特性通常变化得非常缓慢
- 将模型正则化，从而能够学习到那些随时间变化较为缓慢的特征

引入慢性原则

- 向代价函数添加以下项：

$$\lambda \sum_{t} L(f(\boldsymbol{x}^{(t+1)}), f(\boldsymbol{x}^{(t)}))$$

- λ是确定慢度正则化强度的超参数项
- t是样本时间序列的索引
- f是需要正则化的特征提取器

慢特征分析十分高效

- 应用于线性特征提取器
- 可以通过闭式解训练

慢特征分析

SFA 算法

- 先将$f(x;\theta)$定义为线性变换
- 然后求解如下优化问题

$$\min_{\theta} \mathbb{E}_t \left(f\left(\boldsymbol{x}^{(t+1)}\right)_i - f\left(\boldsymbol{x}^{(t)}\right)_i \right)^2$$

需要满足三个约束

- 一是学习到的特征要具有零均值，这样优化问题才会具有唯一解

$$\mathbb{E}_t f\left(\boldsymbol{x}^{(t)}\right)_i = 0$$

- 二是学习到的特征要具有单位方差，以防止所有的特征趋近于 0 的病态解

$$\mathbb{E}_t \left[f\left(\boldsymbol{x}^{(t)}\right)_i^2 \right] = 1$$

- 最后一个是要求学习到的特征之间必须彼此线性去相关

$$\forall i < j, \mathbb{E}_t \left[f\left(\boldsymbol{x}^{(t)}\right)_i f\left(\boldsymbol{x}^{(t)}\right)_j \right] = 0$$

稀疏编码

将一个信号表示为一组基的线性组合

- 稀疏编码：
  - 在模型中推断$\boldsymbol{h}$值的过程
- 稀疏建模：
  - 设计和学习模型的过程
- 线性的解码器加上噪声的方式得到$\boldsymbol{x}$的重构，通常假设线性因子有一个各向同性精度为的高斯噪声

$$p(\boldsymbol{x}|\boldsymbol{h}) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}, \frac{1}{\beta}\boldsymbol{I}\right)$$

- 分布$p(\boldsymbol{h})$通常选取为一个峰值很尖锐且接近0的分布
  - 可分解的Laplace分布
  - Cauchy分布
  - 可分解的Student-t 分布

- 稀疏编码中的编码器不是参数化的编码器

稀疏编码

稀疏编码中的编码器不是参数化的编码器

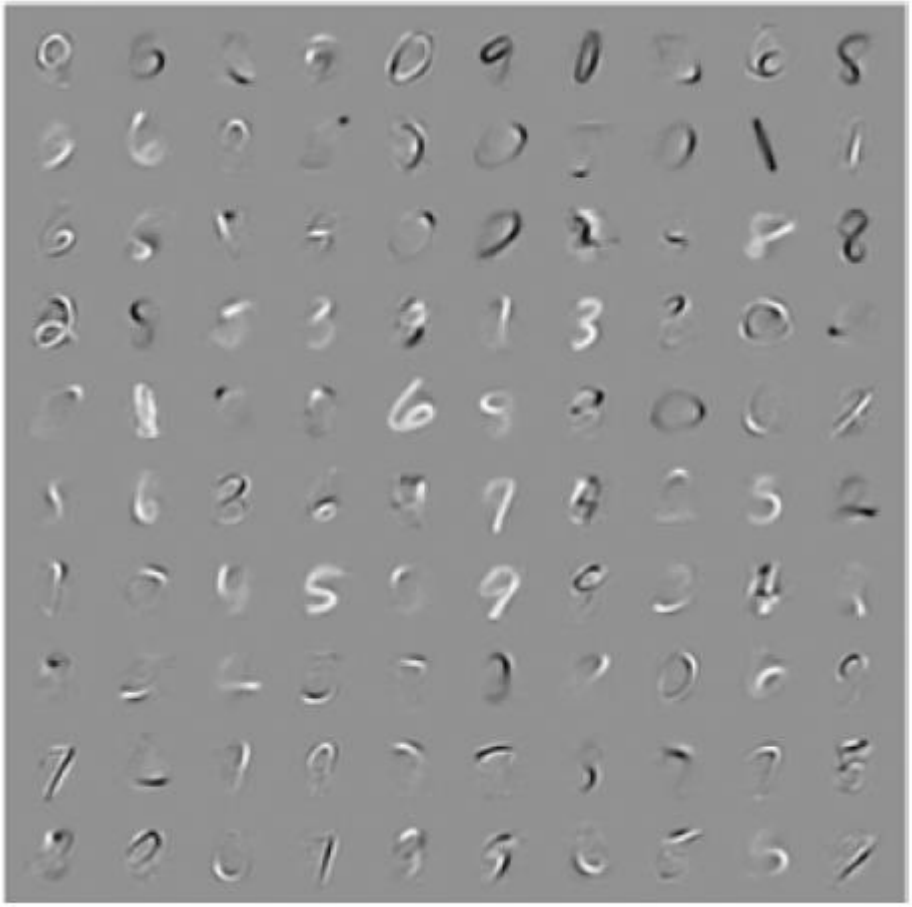
非参数编码器的优点

- 比任何特定的参数化编码器更好地最小化重构误差和对数先验的组合
- 编码器没有泛化误差

非参数编码器的主要缺点

- 在给定$\boldsymbol{x}$的情况下需要大量的时间来计算$\boldsymbol{h}$
- 稀疏编码经常产生糟糕的样本
  - 每个单独的特征可以很好地被学习到，但是隐藏编码值的因子先验会导致模型包括每个生成样本中所有特征的随机子集

稀疏编码



PCA的流形解释

线性因子模型，包括PCA 和因子分析，可以理解为学习一个流形：

- 编码器表示为 $\boldsymbol{h} = f(\boldsymbol{x}) = \boldsymbol{W}^T(\boldsymbol{x} - \boldsymbol{\mu})$
- 解码器负责计算重构 $\hat{\boldsymbol{x}} = g(\boldsymbol{h}) = \boldsymbol{b} + \boldsymbol{V}\boldsymbol{h}$
- 最小化重构误差 $\mathbb{E}[||\boldsymbol{x} - \hat{\boldsymbol{x}}||^2]$

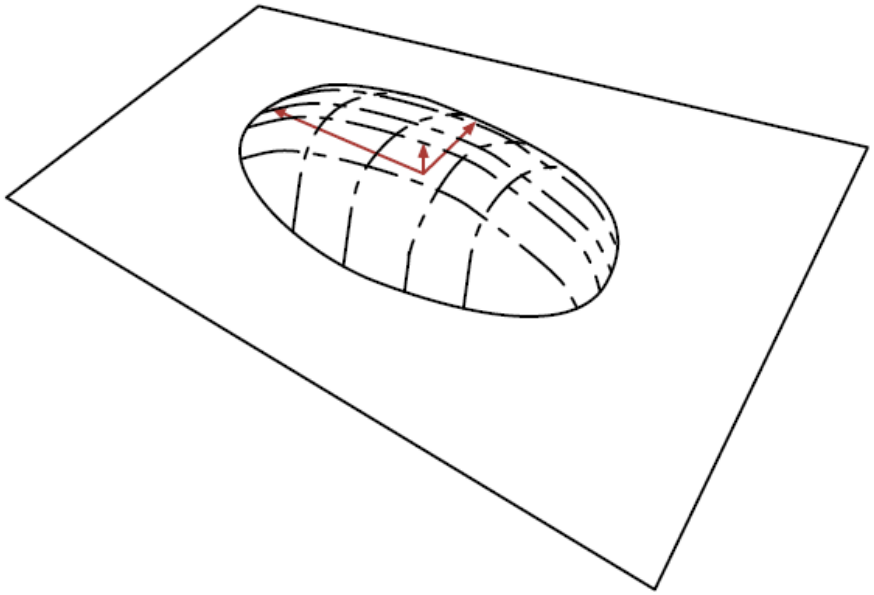


图 13.3: 平坦的高斯能够描述一个低维流形附近的概率密度。此图表示了“流形平面”上“馅饼”的上半部分，并且这个平面穿过了馅饼的中心。正交于流形方向（指向平面外的箭头方向）的方差非常小，可以被视作是“噪声”，其他方向（平面内的箭头）的方差则很大，对应了“信号”以及降维数据的坐标系统。

PCA的流形解释

线性因子模型，包括PCA 和因子分析，可以理解为学习一个流形：

- 最优线性编码器和解码器的选择对应着$\boldsymbol{V} = \boldsymbol{W}, \boldsymbol{\mu} = \boldsymbol{b} = \mathbb{E}[\boldsymbol{x}]$
- $\boldsymbol{W}$的列形成一组标准正交基，这组基的生成子空间与协方差矩阵$\boldsymbol{C}$的主特征向量的生成子空间相同

$$\boldsymbol{C} = \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T]$$

- $\boldsymbol{C}$的特征值λ_i对应了$\boldsymbol{x}$在特征向量$v^{(i)}$方向上的方差
- 如果$x \in \mathbb{R}^D, h \in \mathbb{R}^d$，并且满足$d < D$，则最佳误差为

$$\min \mathbb{E}\left[\left|\left|\boldsymbol{x} - \hat{\boldsymbol{x}}\right|\right|^2\right] = \sum_{i=d+1}^{D} \lambda_i$$

- 如果协方差矩阵的秩为d，则特征值λ_{d+1}到λ_D都为0，则重构误差为0