CDA 统计课程

特征筛选、主成分回归

==========

---

1. 特征筛选

1.1 特征筛选流程

从业务导向、简单回归、相关分析、动态回归、主成分五个方面阐述。如图所示，数据库特征选择过程将从两个方面入手，一方面数据库特征选择流程的特点？另一方面软件具体的实施步骤及注意事项。

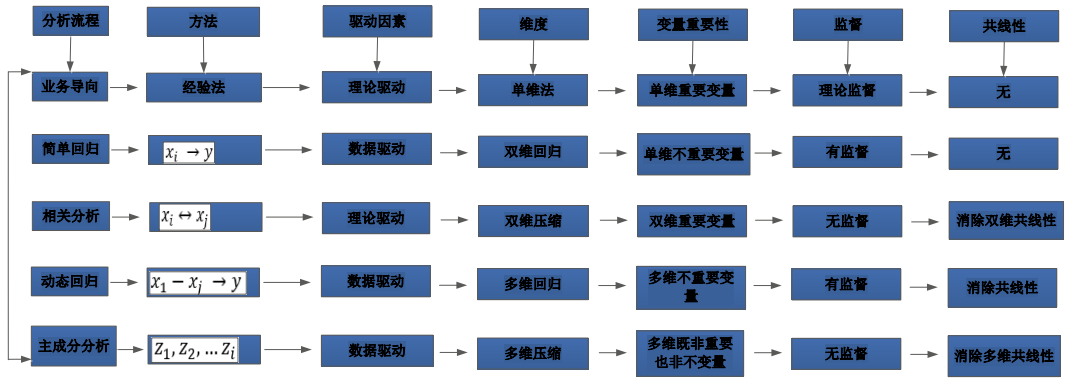

图 数据库特征选择流程图

1.2 DB 特征筛选合理性

（1）方法

通过经验法从业务上判断变量重要性，这样有助于与业务环境相契合，不至于模型脱节于业务，造成与同事、老板的经验不一致。

（2）驱动因素

理论驱动与数据驱动相间出现，并永远要把理论放在第一位，数据分析或探索放在第二位。第一位的目的是要找到重要变量，第二位是为了节省时间，快速的将大型数据中的弱维变量删除。

（3）维度

维度分析的整体规律是从低维到多维。描述性统计侧重于单维分析，并通过图形过渡到对双变量的处理，将双维问题推广到多维，这是统计分析的一般性流程。

（4）变量重要性

业务导向和（第三步）相关分析都是通过业务准则判断变量重要性，此处筛选出的变量比较少，需要执行的时间很久；而第二步的简单回归和第四步的动态回归是，依回归系数检验大幅删除变量的方式进行的，可以有效地节省时间。

第五步的主成分分析是一种压缩变量的技术，在压缩过程中会损失变量信息，因此尽量不要对重要变量压缩，又因为压缩过程需要借助变量间的相关性，所以不重要的变量间又很难产生这种相关，通常也不会有理想的结果。

（5）监督

监督表示有共同的指向，在特征筛选中监督是常态，因为每个自变量共享潜在目标。业务导向、简单回归和动态回归都属于监督方法，分别从业务和数据两个角度执行监督。

相关分析和主成分分析是没有监督的，这与后文的共线性处理有关。

（6）共线性

特征选择本身具有处理内生性问题，选择合适的自变量，避免自变量和残差间的相关，也正因为涉及到多个自变量，共线性问题也是绕不开的话题，因此加入对共线性的处理。

1.3 DB 特征筛选方法步骤

---

2. 主成分回归

2.1 主成分原理

线性回归分析中，我们发现最小二乘技术实际是使点的垂直距离，即误差平方和（ss_e）最小，也就是下图散点距离回归线的垂直距离最小。

如果是向因变量均值做垂直线，那么分解的是每一个点在因变量上的方差，因此这种做法其实是假设因变量比较重要。当然相反，如果我们向水平均值即自变量均值作垂直线，其实是假设自变量很重要，如右图所示，因此现在我们要询问一个问题，主成分分析能否假设变量重要性？回答是非监督降维当然不行，因为多变量技术中不区分因变量和自变量的角色，而是将所有变量视为同等重要的自变量。

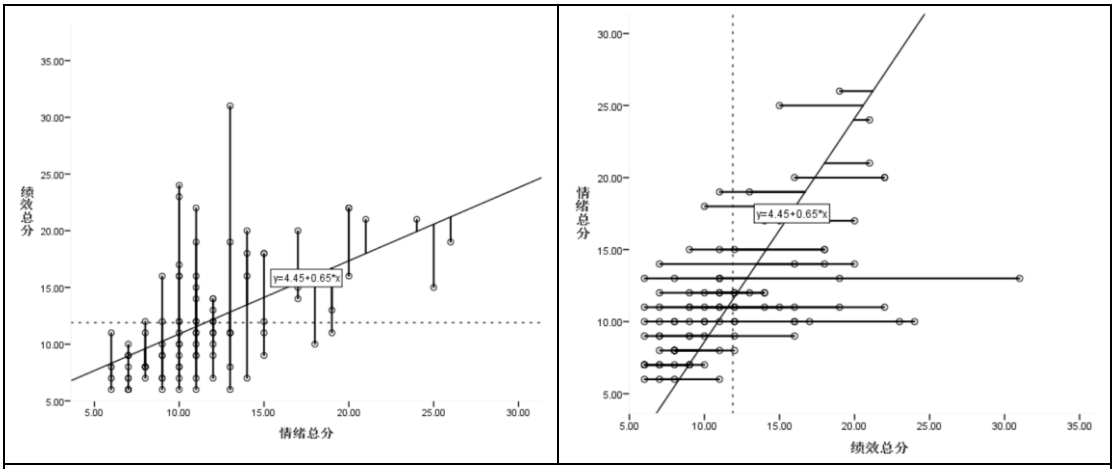

图 回归类算法的变量假设

这点启示帮助我们理解主成分算法的特征，那就是若想分解误差，我们既不能水平也不能纵向垂直作误差线，而是需要向中间的某个方向作误差线，如下图，暂时将图中的长实线视为第一主成分，短的视为第二主成分。某个散点X_{ij}向第一主成分做垂线，该垂线既不水平也不纵向垂直，能够确保行列的重要性假设相同[1]。

如果我们假设所有散点，同时向第一主成分作垂线，并约束其误差平方和最小，此时我们估计的回归方程就是第一主成分模型。由于第一主成分信息对应于第一束光投影后的方差，所以投影的方差拥有 80%的维度信息，就是第一主成分身上携带的数据信息，同理可得第二主成分携带了剩余的 20%的数据方差（n个变量对应n个主成分）。同时如果我们有理由认为 20%的信息并不重要，我们只需使用第一主成分，这就完成了降维——通过损失信息获得更少维度的过程。

主成分其实也是广义上的回归，既然是回归就有自己能够解释的部分，所以可以将主成分解释的方差看成回归中的判定系数或信度系数R^2，于是我们就可以根据$R^2/(1-R^2)$近似构造主成分的特征值，由于特征值取值是严格递减的，因此如果R^2以 0.5 为界的话，特征值以 1 为界，进而判断选择多少个主成分。

[1] 《实用多元统计分析》的样本主成分的近似几何意义。

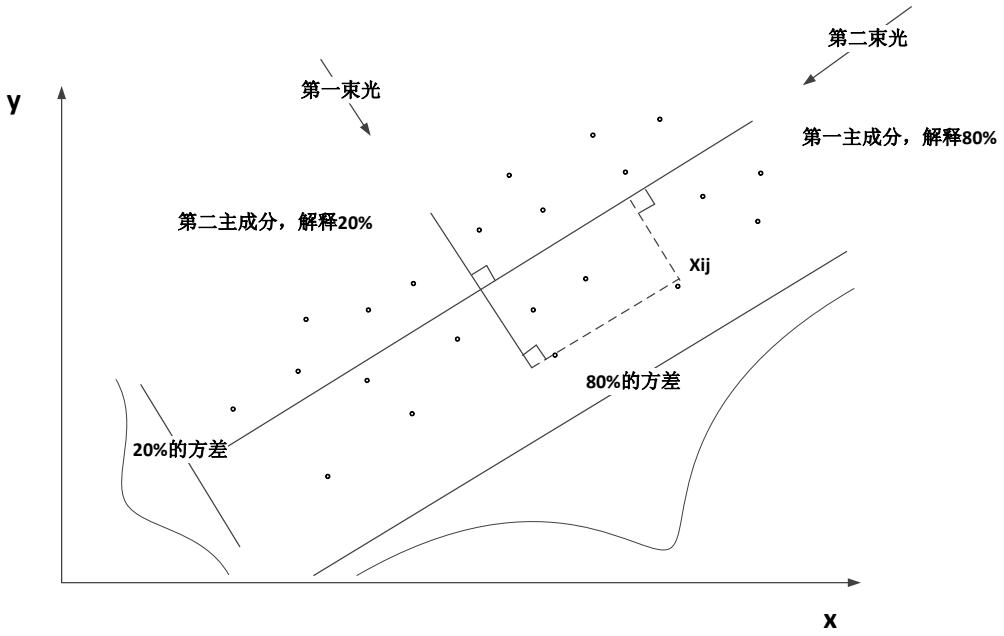

图 主成分算法示意图

2.2 主成分判断标准

主成分分析优劣的判定标准:

第一，尽量将更多变量压缩在第一主成分和第二主成分之内;

第二，第一主成分解释的信息能够超过 50%;

第三，第一主成分和第二主成分解释的信息总和超过 70%;

第四，第一主成分除以第二主成分的比值大于 3;

第五，用更少的主成分代表更多的变量。

2.3 主成分的应用场景

场景 1：假设经过 5 个步骤的特征选择后，模型上限依然是 15 个自变量，此时已经选出 6 个自变量，并需要对 30 个自变量进行主成分分析。如果最终压缩的变量是 3 个（如 z_1, z_2, z_3），并且解释 70%的信息，问此结果如何？回答是可以接受；如果解释了 50%的信息呢？回答依然是可以接受；如果是 40%的信息？回答是不可接受。这是因为主成分很少接受累积解释比例低于 50%的场景，但 50%又远远没有达到 70%的要求，并且 3 个主成分也超出第一条准则，为什么可以接受呢？主要原因是此处的 30 个变量首先并不重要，否则也可能直接删除，其次是 3 代表 30 可谓以少示多，满足更少和更多的要求。

场景 2：如果原始变量只有 5 个，进行压缩后产生 3 个主成分，并且保证其解释的总信息高达 85%。若问结果如何，回答是不可接受，因为这不满足以更少的主成分代表更多的变量，除非降维成一个主成分，解释高达 70%以上，才能免强满足所谓更少和更多的含义。

场景 3：如果主成分目的是分析 20 个变量间的行为偏好特征，例如观察组间异方差、正态分布、数据稀疏等问题，此时需要将数据尽量压缩在第一维和第二维之内，因为超过二维后的数据很难直观判断数据特征，所以若压缩成 3 个主成分，并且累积解释 70%，通常也是不可接受的；若将其压缩成 2 个主成分，并且累积解释了 60%，则视其为可以接受。

场景 4：如果研究目的是获得综合评分，此时将数据压缩成几个主成分并不重要，因为可以对多个主成分进行加权求和，而重要的是最终累积解释比例，要求越高越好，例如如果对 20 个变量进行压缩，最终产生 2 个主成分，并且保证累计解释比例为 70%，但此结果不可接受，若产生 6 个主成分，并累计解释比例 90%，问结果如何？回答是可以接受。

场景 5：如果主成分目的是为了判断数据是否存在 4 个维度的结构效度，而此时主成分的结果是 3 个维度，并累积解释 90%以上的信息，问结果如何？回答是不可接受；如果主成分的结构和数据原始结构一致，并且只累积解释 60%的信息，问此结果如何？回答是可以接受。

2.4 主成分与因子

测量学认为，如果一个指标无测量误差，如收入，通常视为显变量；如果变量存在测量误差，很难用一个指标直接测量，如幸福感，则将其视为潜变量。因子即潜变量之意，意为不可直接测量、不可见但实际存在的变量。主成分或因子就是实现这种潜在评分的算法。由于主成分对本身的关注程度并不高，故无需命名，但因子即潜变量，潜变量需要命名，因为也这是结构意义所在。

（1）主成分：无名

从主成分的应用标准来看，如果累积解释 70%就已经很好了，但我们要问剩下的 30%哪去了？其实是通过直接丢弃的方式，来获得数据维度的精简。当然如果在变量不重要的情况下，问题并不大，但如果变量很重要，丢失了如此之多的信息，显然是不可取的。因此可见，主成分分析的变量既非重要，又非不重要，这是告诉我们变量其实并不重要。

如果变量不重要，那么它对业务的解释固然也不重要，因此我们为什么要费时费力去解释其业务意义呢？主成分的存在，只是为了 “榨取”数据的剩余信息，尽可能的提取信息，至于它与业务产生了什么关系，其实不是重点，所以主成分具有工具变量的性质。

大数据应用场景中，主成分主要用于主成分回归这种模式。如果从主成分回归的应用角度来看，主要是来解决老样本预测，即更关注预测值及市场细分能力。而获得市场细分评估，与我们是否去解释主成分没有任何影响，所以没有必要解释。此外若单结构问题而言，可以通过重要的自变量与因变量间的β来解释，那么解释β与我们控制的主成分的实际意义自然也没有关系。

综上所述，我们发现主成分也是一种潜变量，只是这种潜变量在使用过程中，并没有用到名称及业务意义，看起来有些形式大于内容，所以无需命名。

（2）因子分析：有名

潜变量与指标间的关系有反映型和形成型两种，如图 6-3 的左图为反映型，即形式为“面→点”，使用多个指标反应潜变量，比较符合社会科学对概念的定义；右图为形成型测量，即“点→面”的形式，不过对于社会科学概念的定义，多少有点“以偏概全”的意味，因此通常建议使用反映型测量。

反映型测量的模型为

$$\xi_1 = \lambda_{x11} * x_1 + \delta_1 \ ,$$
$$\xi_1 = \lambda_{x21} * x_2 + \delta_2 \ ,$$
$$\xi_1 = \lambda_{x31} * x_3 + \delta_3 \ ,$$

而形成型测量的模型为

$$\xi_1 = \lambda_{x11} * x_1 + \lambda_{x21} * x_2 + \lambda_{x31} * x_3 + \zeta_1 \quad 。$$

形成型测量其实类似于传统统计模型中的回归，而我们为什么不使用回归来测量潜变量呢？回归中的预测值也是潜变量，只使用一个误差，误差源来至于偏残差，使用潜变量即预测得分来判断影响因素与因变量间的关系，表示对未来行为的潜在评估，但反映型测量更侧重于对概念本身范畴的定义，并假设每个显变量携带不同的误差源（每个显变量与潜变量的相关之余），因此如果强调对未来行为的预测或潜在行为的倾向性，可以使用形成型测量，而如果强调概念本身的测量可以使用反映型测量。

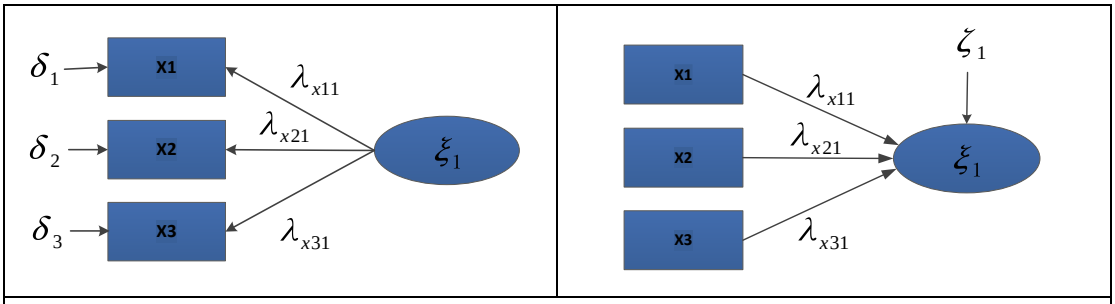

图 潜变量测量

2.5 主成分回归：“回归+回归”模式

主成分回归出现的背景是，统计学逐渐从点和线的问题转移到面的优化问题，因此本质上属于多模型联合，但当时这种联立并不一定具有合理控制误差的特点，诸如联立方程组和结构方程能够合理控制误差，但路径分析和主成分回归等技术则没能合理控制，这直接导致合理控制误差的模型适应了小数据的生态环境，如精确性、归因、结构等问题，而不能合理控制误差的模型也恰好适应于大数据的生态，如速度研究、工具性质、整合性等问题。

主成分回归和路径分析本质上也是没有区别的。就路径分析而言，模型本身也是多个回归的组合。

$$y_1 = \gamma_{11} * x_1 + \gamma_{12} * x_2 + \zeta_1 \quad , \qquad (1\text{-}1)$$

$$y_2 = \gamma_{21} * x_1 + \gamma_{22} * x_2 + \beta_{21} * y_1 + \zeta_2 \quad , \qquad (1\text{-}2)$$

进一步来看，如果我们将公式 1-2 的意义作拓展，x_1和x_2都是自然字段，但 y_1是根据公式 1-1 的回归计算而来，而主成分也是回归，所以若将y_1视为主成分z_1，公式 1-2 岂不就是主成分回归吗？