Appendix

主成分分析の理論

A.1. 主成分の導出

ここでは、主成分分析の理論について直観的な説明とそれに伴ういくつかの概念の定義を与えます。今、N 個のサンプル数を持つ P 次元の標本を X とします。すなわち、X は以下の算式で表されるとします。

$$\mathrm{X} = \{x_{i,j} | i = 1,2,\cdots,N, j = 1,2,\cdots,P\}$$

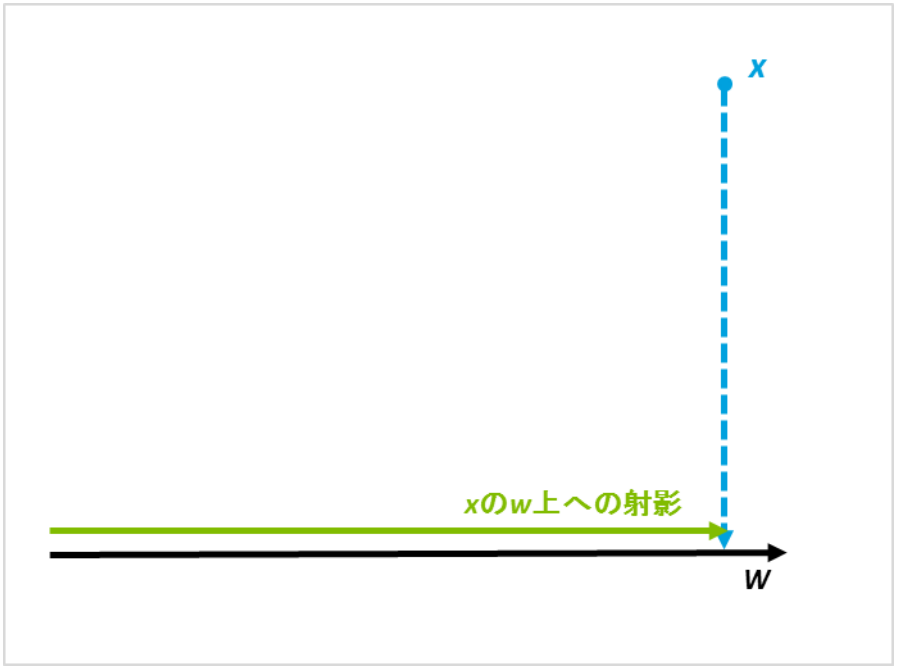
また、

$$\boldsymbol{x_i} = \{x_{i,j} | j = 1,2,\cdots,P\}$$

とおきます。

X は各成分のサンプル平均$\overline{x_p}(=\frac{1}{N}\sum_{i=1}^{N} x_{i,p})$が 0 であることを仮定します。（この仮定は、データのばらつきを考える場合は中心からの各サンプルの距離が重要になること、また、それは平行移動によって変わらないことを考えて議論の論理性を損なうものではありません。）X に含まれるサンプルは無作為に抽出されたものであり特別な規則はないとします。ある一つの「軸」$\boldsymbol{w} = (w_1,\cdots w_P)$を仮定したときに X に含まれる要素がどれだけ $\boldsymbol{w}$ によって特徴づけられるかは、X に含まれる各要素の $\boldsymbol{w}$ に対する射影、すなわち内積によって表すことができます。射影を考える場合は、各要素の $\boldsymbol{w}$ 上での長さが保存されている必要があるので、${w_1}^2 + \cdots + {w_P}^2 = 1$であるとします。

（図 A－１）



X の各要素の $\boldsymbol{w}$ 上への射影は、$t_n = \sum_{j=1}^{P} X_{n,j} \cdot w_j \quad (\mathrm{n} = 1,2,\cdots,\mathrm{N})$で与えられるので $\boldsymbol{w}$ 上での X の標本不偏分散は、

$$\frac{1}{N-1}\sum_{n=1}^{N} {t_n}^2 = \boldsymbol{w} \cdot \left(\frac{1}{N-1} \cdot {}^{\tau}XX\right) \cdot {}^{\tau}\boldsymbol{w}$$

と表現することができます。X 全体のばらつきを最も説明することができる軸は、直感的にいうと、X 全体のばらつきを最も多く切り取ることができるものであるため、上述の算式を最大にするような $\boldsymbol{w}$ だと考えられます。

ここで、$\mathrm{V} = \left(\frac{1}{N-1} \cdot {}^{\tau}XX\right)$とおくと、V は非負値対称行列となり右辺は V に対する二次形式となっているため、これを最大にする $\boldsymbol{w}$ は V の最大固有値に対応する固有ベクトルであることが一般的に知られています。このような $\boldsymbol{w}$ を X に対する**第一主成分**と呼んでいます。同様にして、二番目に大きい固有値に対応する固有ベクトルを**第二主成分**、三番目に大きい固有値に対応する固有ベクトルを**第三主成分**とそれぞれ呼びます。さて、これらは V が非負値対称行列であることから、正規直交基底となることも知られています。すなわち、主成分分析とは標本のばらつきを大きく切り取ることができる正規直交系への座標変換を求めることであり、それらの正規直交基底は標本から得られる分散共分散行列の固有ベクトルであると言い替えることができます。

A.2. 寄与率

では、X のばらつきのうち、どの程度の割合が各主成分によって切り取られているのでしょうか。行列の演算を考えると上述の算式より、

$$\frac{1}{N-1}\sum_{n=1}^{N} |\boldsymbol{x_n}|^2 = Tr(V)$$

であることが分かります。左辺は X の標本不偏分散を表しており、右辺は行列の固有値の総和に等しくなります。このことは、X のばらつき全体は各主成分上への射影によって完全に分解することができるということを意味しています。このことを踏まえて、各主成分が標本において切り取ることが出来る分散の割合は、各主成分に対応する固有値を$\{\lambda_l | l = 1,2,\cdots,P\}$とおくと、

$$\mathrm{C}_m = \frac{\lambda_m}{\sum_{l=1}^{P} \lambda_l} = \frac{\lambda_m}{Tr(V)}$$

であると考えられます。このC_mを**寄与率**と呼びます。また、第一主成分から第m主成分までの累計により、どの程度のばらつきを表現できるかについての指標は

$$\sum_{l=1}^{m} C_l$$

と表すことができ、これを**累積寄与率**と呼びます。