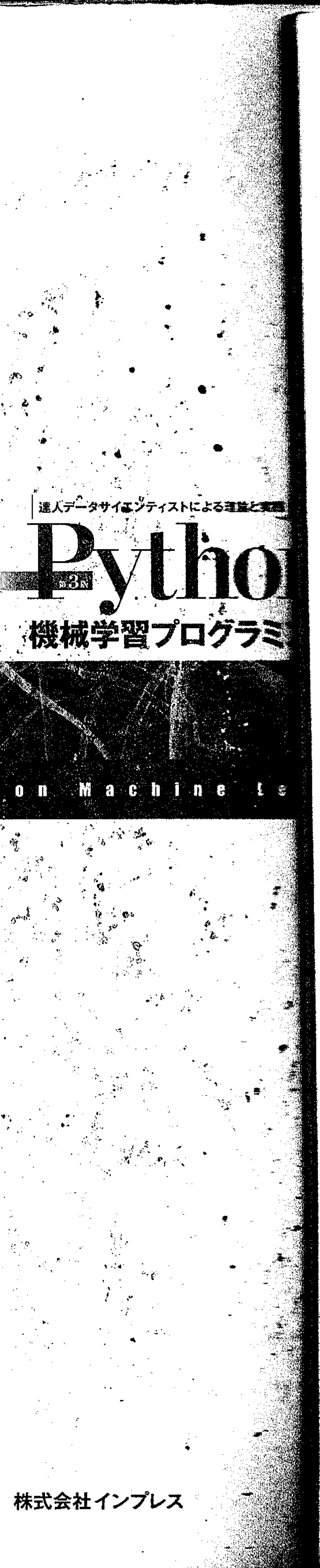
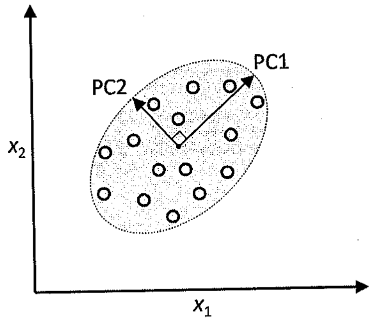


5.1.1 主成分分析の主要なステップ

ここでは、**主成分分析**(Principal Component Analysis : PCA)について説明する。
な分野にわたって広く使われている教師なし線形変換法であり、最もよく用いられる
出と次元削減である。それ以外にも、探索的データ解析や株取引での信号のノイズ除
マティクス分野でのゲノムデータや遺伝子発現量の解析にも応用されている。

PCAは、特徴量どうしの相関関係に基づいてデータからパターンを抽出するのに
うと、PCAの目的は、高次元データにおいて分散が最大となる方向を見つけ出し、
それよりも低い次元の新しい部分空間へ射影することである。次の図に示すように、
互いに直交するという制約があるとすれば、新しい部分空間の直交軸(主成分)を分
向と見なすことができる。ここで、x_1 と x_2 は元の特徴量軸であり、**PC1** と **PC2** は



PCAを次元削減に利用する場合は、$d \times k$ 次元(d 行 k 列)の変換行列 $\boldsymbol{W}$ を作成
ベクトル $\boldsymbol{x}$(訓練データの特徴量)を新しい k 次元の特徴量部分空間に写像できる。
量部分空間は、元の d 次元の特徴量空間よりも次元が低い。たとえば、特徴量ベク
に定義されているとすれば、

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_d],\ \boldsymbol{x} \in \mathbb{R}^d$$

変換行列 $\boldsymbol{W} \in \mathbb{R}^{d \times k}$ によって変換され、

$$\boldsymbol{xW} = \mathbf{z}$$

次の出力ベクトルが得られる。

$$\mathbf{z} = [z_1, z_2, \ldots, z_k],\ \mathbf{z} \in \mathbb{R}^k$$

元の d 次元のデータを新しい k 次元の部分空間に変換すると(通常は $k << d$ ※1)
散は最大となる。結果として生じるすべての主成分の分散が(それよりも前の主成
なるのは、他の主成分と相関がない(直交している)場合である。入力特徴量が相関
結果として生じる主成分は相互に直交した状態となる(他の主成分と相関がない)。
タのスケーリングに対して非常に敏感だ。特徴量が異なる尺度で計測されていて、

※1 [監注] $k << d$ は、d に比べて k がはるかに小さいことを表す。