Pattern Recognition and Machine Learning

第12章　連続潜在変数

修士 1年
村下 昇平

BehaviorInNetworks
都市生活学・ネットワーク行動学研究室

* もくじ

- 主成分分析とは？
  - 一般的な主成分分析の目的と定式化
  - 主成分分析の応用
- 確率的主成分分析
  - 通常の最尤推定とEMアルゴリズムによる最尤推定
- ベイズ的主成分分析
- その他の話題
  - 因子分析やカーネル主成分分析
  - 非線形潜在変数モデル

0. 主成分分析とは？

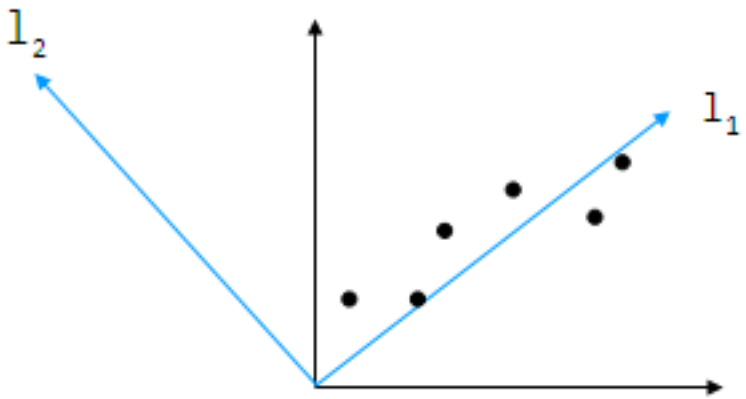
- この章では様々なアプローチによる主成分分析について扱います。
- っていうか、主成分分析ってよく聞くけど、そもそもなんなんですか？
- …というところについてはあまり説明されていないので、とりあえず勝手に調べてみました。

0. 主成分分析とは？:そもそもの目的。

ひとことでいえば「データ分布を扱う空間の基底を、より最適な別の基底に変換してから変量を解析する手法」。この「最適な基底」とは分散が最大になる方向で、それは分散共分散行列の固有ベクトルとなる(詳しくはこの後やります)。

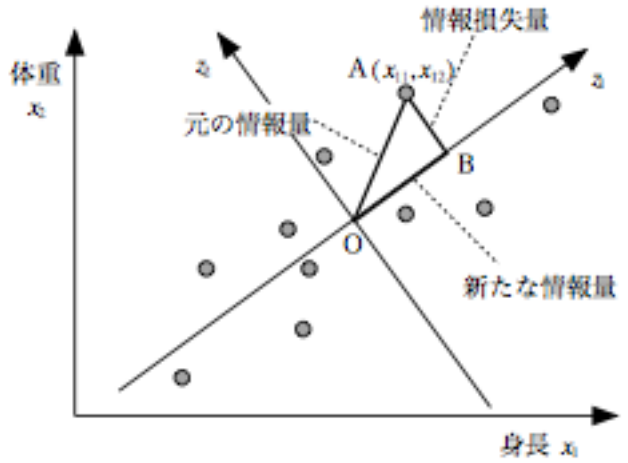
下図のばあい、基底変換後の l_2 軸周りにはほとんど情報量がなく、l_1 軸周りに情報が集中していることが分かる(というか、そうなるように基底を選ぶのが主成分分析だ！)。

l_2 は無視してしまうことで、2次元で扱っていたデータを 1次元だけで比較することが可能になる。…こうした(情報量の損失を最小化するという条件のもとでの)低次元化の手法が主成分分析！



0. 主成分分析とは？:分散が最大って？

じゃあなんで「分散が最大となる方向」で情報量の損失が最小化されるの？



いま、上図における点A(x1,x2)に注目すると、第1主成分z1のみでデータを代表させる場合の情報量はOBで与えられる(ここで、点Oはz1軸の原点であり、データの重心である)。このときOAが元の情報量であるが、このような各点における情報量の損失(OAとOBの情報量の差の総和、すなわち AB2=OA2 – OB2 の総和)を最小化するには、OBの総和、すなわち軸まわりのばらつきを抑えたい、ってことになる。
したらば、たしかに「分散が最大となる方向」にとった軸まわりでこの総和が最小となるよね…

0. 主成分分析とは？:簡単な定式化

既に説明した通り、主成分分析とはD次元データを M<D であるようなM個の変数を用いて近似することである。すなわち xn を以下の近似式によって近似するということになる。

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni}\mathbf{u}_i + \sum_{i=M+1}^{D} b_i\mathbf{u}_i$$

ここでu_iはD次元の正規直交基底(互いに直角な単位ベクトル：第1～第D主成分)である。また、$\{z_{ni}\}$はその特定のデータ点に依存している（失われていない情報）が、一方$\{b_i\}$はすべてのデータ点に共通な定数と考える。

この近似式を用いると、情報損失は次の歪み尺度Jによって表される。

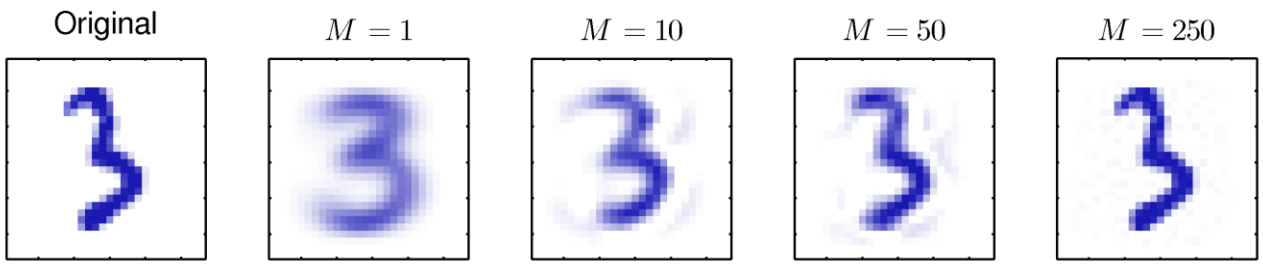
$$J = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|^2$$

すなわち主成分分析の目的はこのJの最小化であると言える。
…で、肝心の最小化(第1, 第2節)に関しては省略しちゃいますが、結局共分散行列の（相対的に大きな）固有ベクトルとなるわけです。

1.3. 主成分分析の応用

- 次元削減をするということは…
- 特徴抽出、すなわち文字認識の際に位置や大きさのばらつきを正規化するのに用いられたり…(主成分を潜在変数とみなし、得られたデータがこれに従うと考えることで、「情報損失」を「ノイズ」として解釈することになる)

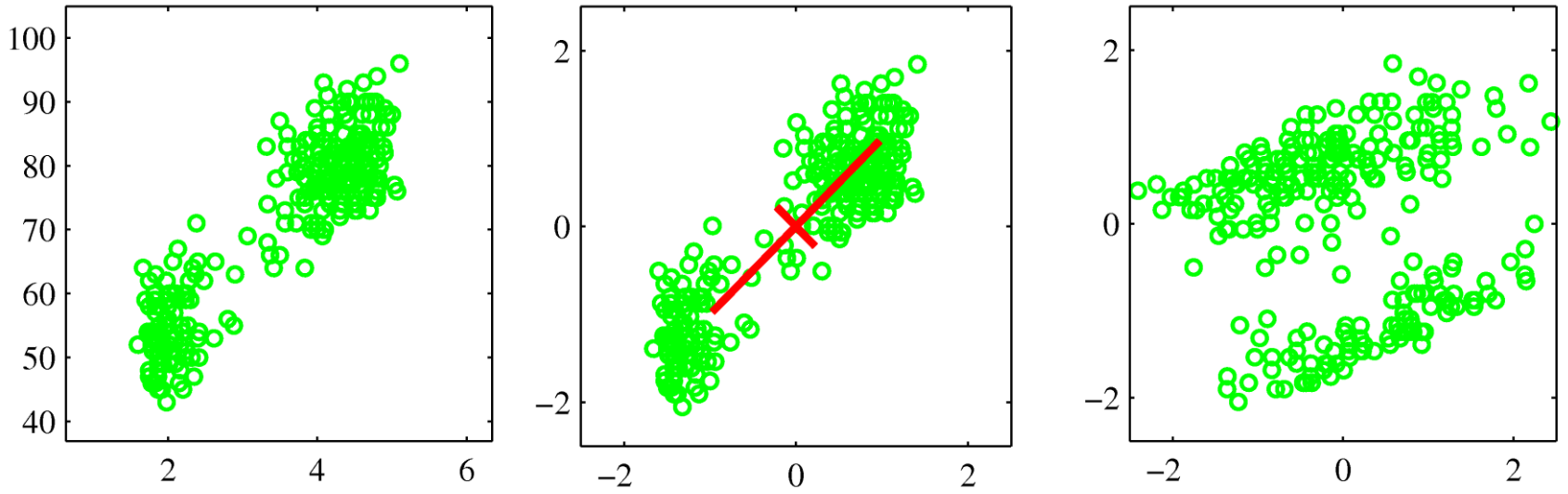
- 非可逆データ圧縮に利用したりできる。



データ圧縮の例。原画像は28pixcel×28pixcel、すなわちD=784次元のデータであるが、M=250で十分に表現されていることがわかる。

1.3. 主成分分析の応用

- また、必ずしも次元削減だけでなく…
- 平均を0、共分散行列（各々の分散だけじゃない！）を単位行列にするような白色化(球状化)を行うことで異なる変数を無相関化できたりする。



様々な前処理の例。左が元データ。中央は個々の変数について平均を0、分散を1に標準化したもの（赤線は規格化されたデータ集合に対する主軸となっている）。そして右は主成分分析による白色化（平均が0、共分散行列が単位行列）。

2.確率的主成分分析

- というわけで、本節では主成分分析が確率的潜在変数モデルの最尤解としても表現されることを示す。
- このような形で定式化された主成分分析を確率的主成分分析と呼ぶ。
- 確率的主成分分析の利点は次の通りである。
  - モデルがデータ集合の主要な相関の構造を捉えることができることに加え、（制約付きのガウス分布に基づいているため）自由パラメータの数を制限できる。
  - 主成分分析を行うためのEMアルゴリズムを導くことができる。これは上位の固有ベクトルのみが必要な（Mが小さい）状況では計算効率が良く、途中でデータ共分散行列を計算する必要もない。
  - 確率モデルとEM法の組み合わせにより、データ集合内の欠損値を扱える。
  - 確率的主成分分析の混合モデルをより見通しのよい方法で定式化でき、EMアルゴリズムを用いて訓練できる。
  - 主成分分析のベイズ的取り扱いの基礎を与える。ベイズ的取り扱いでは、主成分空間の次元を自動的にデータから見いだすことができる。
  - 尤度関数が得られるので、他の確率密度モデルとの直接の比較が出来る。これは、通常の主成分分析で計算できる「再現コスト」という量がしばしば誤解を招く結果を与えることと対照的である。
  - クラスで条件づけられた確率密度のモデル化に利用できる。
  - データサンプルを分布から得るための生成モデルとして利用できる。

2. 確率的主成分分析

- 確率的主成分分析は、すべての周辺分布と条件付き分布がガウス分布になっている線形ガウスモデルの枠組みの単純な例である。
- 確率的主成分分析を定式化するには…
  - まず主部分空間に対応する潜在変数**z**を明示的に導入する。
  - 次にガウス分布を仮定した潜在変数**z**についての事前分布 $p(\mathbf{z})=\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$ および…
  - 潜在変数の値で条件付けられた観測変数xについてのガウス分布である条件付き分布 $p(\mathbf{x}|\mathbf{z})= \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z}+\boldsymbol{\mu},\sigma^2\mathbf{I})$ を定義する。（パラメータの詳細については後述する）
- この枠組みは伝統的な主成分分析の見方と対照的である。すなわち、潜在変数空間からデータ空間への写像に基づいているのである。

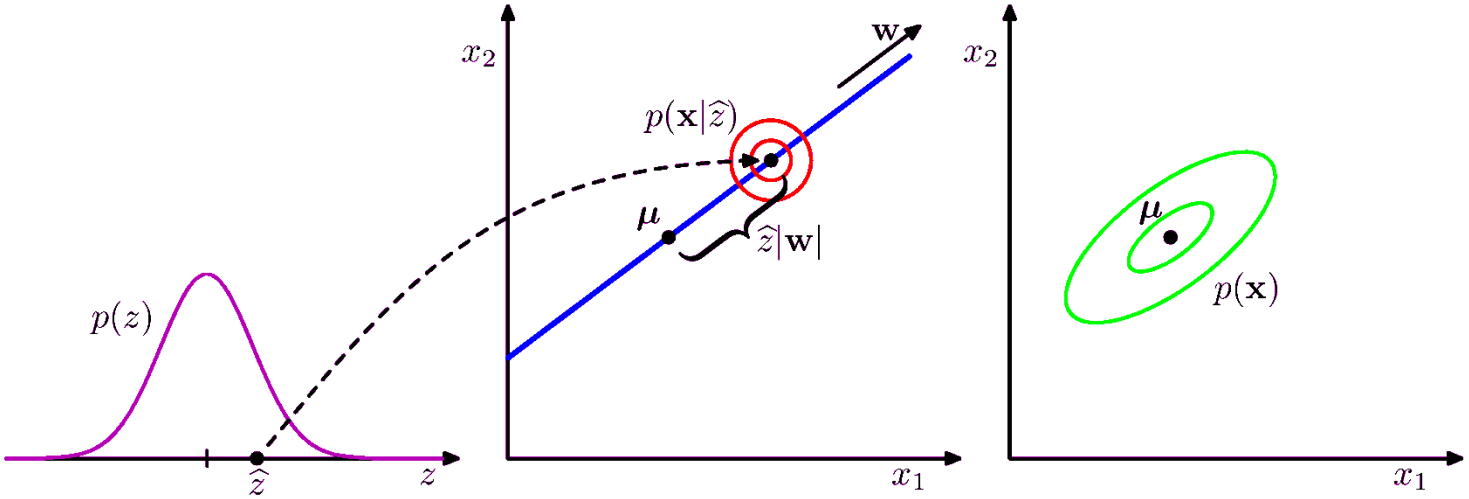
2. 確率的主成分分析

- 生成モデルの観点から確率的主成分分析を眺めることができる。
- つまり、まず潜在変数の値をひとつ選び、その値で条件付けつつ観測変数をサンプリングすることで、観測変数のサンプル値が得られる。D次元の観測変数xは、M次元の潜在変数zの線形変換にガウス分布による「ノイズ」が加えられたもので定義される。
- これは次式で表される。ここで z はM次元の潜在変数であり、ガウス分布に従う。また、εはD次元の、平均0で共分散が $\sigma^2\mathbf{I}$ のガウス分布に従うノイズの変数である。

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

2. 確率的主成分分析

図でかくと…



最初に潜在変数 z の値を事前分布p(z)からひとつ抽出し、次に **x** の値を平均 **wz+μ** 、共分散$\sigma^2\mathbf{I}$ の当方的なガウス分布（赤の円）から抽出することにより、観測データ点 **x** を生成する。

2. 確率的主成分分析

- 「お前は何を言っているんだ」と感じられたと思いますので、ここで自分なりの解釈を書いておきます。

- つまり、従来の主成分分析では…
  - データがまずあって、そこから特徴(=情報損失が最小になるような基底)を見つけ出す。
  - M+1～Dの主成分に頼る部分は「損失」
- しかし確率的主成分分析の考え方では…
  - まず特徴(=主成分:潜在変数)があって、与えられたデータはそれにノイズが乗ったものと考える。
  - M+1～Dの主成分に頼る部分は「ノイズ」
  - たぶん、ノイズが乗ったデータに対する、ベイズ的フィッティングと似たようなもんなんじゃないかな。

2.1. 確率的主成分分析:パラメータの最尤推定

というわけで、パラメータ W, μ, σ^2 の値を最尤推定を使って決定する。

まず対数尤度関数は以下で表される。（先ほども述べたとおり、z(のパラメータ) から）**X** が生成される、と考えている）

$$\ln p(\mathbf{X}\,|\,\mu,\mathbf{W},\sigma^2) = \sum_{n=1}^{N} \ln p(x_n\,|\,\mathbf{W},\mu,\sigma^2)$$

$$= \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{x}\,|\,\mu,\mathbf{C})$$

p(x)の平均が**μ**であるのは先ほど見た通り。
分散 **C** はxの共分散cov[**x**] = cov[**Wz**+**μ**+**ε**] = **E**[(**Wz**+**ε**)(**Wz**+**ε**)$^{\mathrm{T}}$]。**z**と**ε**は独立であるから単純に和をとればよく、結局 **C**=**WW**$^{\mathrm{T}}$+σ^2**I** となる。

最終的に対数尤度関数は以下のように書き下せる。（対数をとったガウス分布の和を書き出しただけ！）

$$= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\mu)^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}_n-\mu)$$

2.1. 確率的主成分分析:パラメータの最尤推定

平均については単純なガウス分布に対する尤度関数なので、これを最大化するような μ はデータ平均と一致する。これを代入すると、対数尤度関数は次のように書ける。

$$\ln p(\mathbf{X}|\mathbf{W},\mu,\sigma^2) = -\frac{N}{2}\left\{D\ln(2\pi) + \ln|\mathbf{C}| + \mathrm{Tr}\left(\mathbf{C}^{-1}\mathbf{S}\right)\right\}.$$

ここで S はデータに対する共分散行列であり、次式で与えられる。

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}}.$$

また、Wとσ^2 についてはもっと複雑であるが閉形式の厳密解が存在し、その求め方などについて最近では研究がすすんできているらしい。

2.2. EMアルゴリズムによる主成分分析

- こうして厳密な閉形式の形で最尤パラメータの値を得られるんならそれを使えばいいじゃん、話はそれでおわりじゃん、って話もあるけれど、大規模な問題（高次元空間）においては、サンプルの共分散行列を直に扱うよりもEMアルゴリズムを用いた方が計算量的に有利になったりする。

- というわけで、EMアルゴリズムによる主成分分析であるが、これは以下のステップを経て尤度関数を最大化する。

1. パラメータの初期化
2. Eステップ　：潜在変数空間の事後分布の十分統計量を計算
3. Mステップ　：パラメータ値の更新

…以下、1と2を繰り返していく。

どのように定式化するかについては省略。

2.3. ベイズ的主成分分析

- これまでは主部分空間の次元Mが与えられたものとして考えてきたが、そもそもこのMをどのように選ぶか、という問題がある。
  - 比較的大きな固有値と比較的小さな固有値の間にはっきりした境目があるなら、その境目までの固有値を用いるのが自然であるが…実際にはそんなに明らかに変わってきたりしない。
  - 公差確認法によって確認用データ集合の尤度関数が最大になるように次元の値を選ぶこともできるが、計算量的に高くついてしまう。

- ベイズ的な手法でモデル選択を行えば、これを解決することができる！（らしい）

- 詳細は略。エビデンス近似うんぬん。

2.4. その他の話題

- 因子分析
  - 確率的主成分分析と深い関係がある
- カーネル主成分分析
  - カーネル置換を主成分分析に適用
- 非線形潜在変数モデル
  - これまでの主成分分析は線形ガウス分布に基づいていたが、ここでは非線形・非ガウス的なモデルを考える。
  - 独立成分分析
    - 潜在変数と観測変数の関係が線形だが、潜在変数の分布が非ガウス分布であるモデル
  - 自己連想ニューラルネットワーク
    - NNの教師なし学習への応用で、次元削減などに用いられる。ここでは入力数と出力数を同じにしたネットワークを使い、この誤差に関する指標を比較する。
  - 非線形多様体のモデル化
    - 自然に得られるデータ源は（ある程度のノイズは別にして）高次元の観測データの空間のなかに埋め込まれた低次元非線形多様体に対応する場合が多い。ここでは明示的にこの性質を把握する手法についていくつか紹介している。