2010 年度前期　応用統計学　第 1 回

多変量解析と応用 ー イントロダクション

関係を説明するー多変量解析の考え方

統計学は，ランダム現象によって生じた分布するデータ，すなわち「ばらばら」なデータの集まりを扱う学問です．人は，観察される現象が「どんな仕組みで」起きているのかを理解したいという欲求を，常にもってきました．この「仕組みの理解」こそが「科学」であり，仕組みを理解することによって未知の現象を予測することができます．

分布するデータに対しても，そのばらばらなデータが「どのように」ばらばらか，を理解しようと努めてきました．そして，さまざまな現象に対して「ばらばらのなりかたの型」，すなわち確率分布モデルを考え，それを使って「集団の一部のデータを調べて，集団全体を推測する」という統計的推測の技術を開発してきました．例えば，試験の点数の分布が正規分布モデルで表せると考えると，一部の人の点数から全体の点数の分布を推測することができます．

では，「試験の点数」というひとつの数値からなるデータではなく，「各受験生の，英語の点数と数学の点数の組」といった，複数の項目の組からなるデータの場合を考えてみましょう．この場合，複数の項目間の関係を説明することで，「どうばらばらか」をより詳細に説明できます．例えば，英語の点だけ，数学の点だけみれば，それぞれ正規分布にしたがって分布している場合でも，英語と数学の関係を見れば「英語の得意な人は数学も得意」といった両項目の関係を説明することができます．

上で述べたデータの「項目」を統計学では**変量**といい，複数の変量の組からなるデータを，変量間の関係をふまえて分析する統計的手法を**多変量解析**といいます．この講義では，さまざまな多変量解析手法と，その画像処理・パターン認識・感性情報科学への応用を説明します．

回帰分析ー変量間の関係を推定する

> 各都市の緯度と平均気温の組の観測データをみると，緯度が高い（低い）ときは気温は低い（高い）ようである．緯度と平均気温の関係は，どのように式であらわされるだろうか？

このような場合に，「平均気温」(y) を「緯度」(x) の式で表そうというのが，**回帰分析**の手法です．いちばん簡単なのは，図 1 のように $y = a + bx$ という 1 次式の関係があるという「モデル」を考えて，観測された x, y の組にこの式がもっともよくあてはまるように a と b を決めるというものです（この図 1 を散布図といいます)．

もちろん，各都市での気温の違い，すなわち気温の分布は緯度だけで説明されるものではありません．そこで，より精密に気温の分布を説明するために，例えば「気温が緯度と標高の 1 次式で表される」というモデルを考えます．このように，複数の変量の組でひとつの変量が定まると考える分析法を**重回帰分析**といいます．

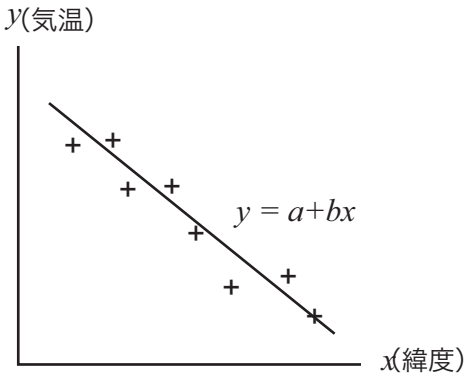

図 1: 散布図と回帰

主成分分析ー多数の変量を代表する変量を求める

> クラスで数学と国語（各々１００点満点）の試験を行った．数学の成績は０点から１００点の範囲に散らばっているが，国語の成績は全員が０点から２０点の範囲に入っていた．どのように総合得点を求めて各学生を評価すればよいだろうか？

この例では，数学の成績の散らばりに比べて国語の成績の散らばりが小さくなっています．このとき，両科目を単純に合計して総合得点を求めるよりも，散らばりの大きい数学の成績により大きな重みをつけて総合得点を求めるほうが，総合得点の散らばりが大きくなり，各受験生の評価をつけやすくなります．

このように，複数の変量をもっともうまく代表するようなひとつの（あるいは，少数の）新しい変量を求める方法が**主成分分析**です．「もっともうまく代表する」というのは，別の見方でいうと「もとのデータをなるべく損なわずに，少ない変量で表す」ということで，主成分分析は，画像データ等をその品質をなるべく保ったままデータ量を圧縮する技術の基礎にもなっています．

判別分析とパターン認識ー多変量データを分類する

> ある病気の診断のため，血圧，血糖値などのいくつかの検査を行なった．これまでの症例から，健康な人と病気の人それぞれの，各検査項目の値の分布はわかっている．このとき，新たに検査した患者の検査結果から，この患者が健康か病気かを判定するにはどうすればよいだろうか？

まず簡単のために，検査項目がひとつの場合を考えてみましょう．ごく単純に考えれば，健康な人の検査値の平均と病気の人の検査値の平均を求め，新たな患者の検査値がそのどちらの値に近いかによって，健康か病気かを判断すればよいことになります．しかし，これでよいのでしょうか？

仮に，病気にかかっている人の検査値の分散が非常に大きく，健康な人の検査値の分散は非常に小さいとしてみましょう．このときは，新しい患者が健康ならば，その検査値は健康な人の平均に非常に近い値になるはずで，そこから少しでも離れている場合は病気であると判断したほうがよいはずです（図2）．このように「健康」か「病気」かどちらに「近い」かを表現するには，それぞれの分布を考慮した距離を定義する必要があります．このような方法であるデータがどの群に属するかを判定する方法を**判別分析**といいます．

判別分析の考え方は，コンピュータで文字を識別するといった**パターン認識**技術の基礎となります．

図 2: 判別分析ー健康か，病気か

それは，画像データや音声データは，いずれもそのさまざまな特徴が多数の変量で表現された多変量データと考えることができるからです．この講義では，パターン認識の分野で使われている判別の手法についても触れます．

クラスタリングー似たデータをグループ化する

たくさんのデータが集まったとき，似たデータをグループにして分類してゆくと，その集まりの特徴が見えてきます．このように「似ているデータをひとつのグループにまとめる」という操作が**クラスタリング**です．「互いに似ているデータ」とは，「散布図上で距離が近い」と考えることができますが，多変量データの場合は散布図は高次元になるので，「似ているデータをまとめる」のはそう簡単ではありません．そのため，クラスタリングは現在でも新しい手法が研究され続けている分野です．

因子分析ー多変量データに潜む構造を見いだす

> 数学，国語，英語の試験を行なった．この３科目の成績を，もっと少ない項目，例えば「学力」という１つの項目，あるいは「数理能力」と「言語能力」という２つの項目で表現することはできないのだろうか？

もしも上のようなことができるなら，それは３科目の試験結果の中に潜んでいるより単純な構造を探り出していることになります．極端な場合，どの受験者も数学と英語の成績が同じなら，各受験者の成績は国語と英語の２項目で表されることになります．こんなに極端でなくても，各科目間の成績に相関があることはよくありますから，１項目あるいは２項目で成績を表せる可能性は十分にあります．上の「総合能力」や「数理能力」「言語能力」のような項目を因子といい，少数の因子を使ってデータを表現する方法を**因子分析**といいます．

例えば，３科目の成績を，１つの「総合能力因子」であらわせると考えたとしましょう．すると，

（各人の数学の成績）　＝　（「総合能力因子」の各人の点数）×（数学に対応する定数）＋（誤差）
（各人の国語の成績）　＝　（「総合能力因子」の各人の点数）×（国語に対応する定数）＋（誤差）
（各人の英語の成績）　＝　（「総合能力因子」の各人の点数）×（英語に対応する定数）＋（誤差）

とあらわせることになります．ここで（「総合能力因子」の各人の点数）は科目間で共通であることに注意してください．各人の「総合能力」が，「総合能力因子」の点数で表されています．この（「総合能力因子」の各人の点数）を因子得点，（（科目）に対応する定数）を因子負荷量といいます．因子分析では，線形代数学の手法を使って，最適な因子数を決め，因子負荷量と因子得点を求めます．

画像科学・感性情報科学への応用

デジタル画像は，各画素を変量とする多変量データと考えられます．そこで，これまでに述べたような画像データ圧縮やパターン認識に多変量解析の手法が用いられます．また，人間の感性を定量的に把握し，工業的に利用しようとする感性情報科学では，製品の性能・特徴を表す多変量データと，人の感性による評価との間を結びつけるために，因子分析などの多変量解析手法が用いられます．これらについても，応用例を紹介します．

今日の演習

ずいぶん前のものですが，次の記事について，統計学的におかしな点があれば述べてください．

> 「サメ肌水着」五輪で圧勝　競泳金メダルの 6 割獲得　－　ミズノと東レ、英スポーツ用品メーカーのスピード社が開発し、「サメ肌水着」として注目された「スピード　ファーストスキン」の着用者が、シドニー五輪の競泳競技で、金メダル総数の 6 割にあたる 31 個を獲得した。メーカー各社の激しい開発競争を制したミズノの上治丈太郎取締役（現地責任者）は、「技術力の高さが証明でき満足している」と話し、早くも 4 年後のアテネ五輪へ向けた開発に意欲をみせている。
>
> ファーストスキンの着用者は、出場選手 1677 人の 6 割にあたる 1006 人（138 カ国）に上った。メダルは、151 個のうちの 100 個 (66%) を獲得。「金」は、男子が 1500 メートル自由形のグラン・ハケット（オーストラリア）ら 26 個中の 14 個 (54%)、女子が 50 メートル、100 メートル自由形のインヘ・デブルーイン（オランダ）ら 25 個中の 17 個 (68%) を占めた。世界記録は、着用者が２、３人いたリレーも含め、12 種目で樹立に貢献したという。（朝日新聞 asahi.com 2000 年 9 月 25 日）