題目　混合正規分布モデル――回帰分析への応用――

氏名　道塚　駿

指導教員　結城雅樹

クラスター分析は，観測値を複数のクラスターに分類する手法群で，分類基準や階層的手法・非階層的手法の別により様々な分類手法が存在し，それらはスポーツなどの行動分類や通販サイトのレコメンデーションをはじめとして幅広い目的に使われている．非階層的なクラスター分析として利用される手法の一つに，混合分布モデルがある．混合分布モデルは本来，多峰分布を記述するものであるが，複数の確率分布(要素分布)から生じた観測値が混在し，どの観測値がどの要素分布に所属するかが不明であるときにも用いられるモデルで，ある観測値が複数のクラスターに確率的に分類されると考えるファジィクラスタリングの一つとして位置づけられる．本研究では 1 変量混合正規分布モデルを扱い，このモデルを回帰分析に応用する．回帰直線が複数混在し，観測値がどの回帰直線のまわりに分布するか不明な状況下で，観測値の分布する回帰直線とそのパラメータの推定のための手段として，混合分布モデルを「混合回帰分析モデル」として応用する．本研究では，混合回帰分析を定式化するとともに，大別して 3 つのシミュレーションを行った．　混合分布や回帰直線のパラメータの推定には EM アルゴリズムを利用した．まず，混合回帰分析モデルの基礎となる混合正規分布モデルの性質を調べるためのシミュレーションである．混合正規分布モデルによる，2 つの要素分布のパラメータ推定と誤判別の頻度が，分布間の距離にどの程度依存するかを調べた．次に，混合回帰分析モデルの推定精度を調べるため，2 本の回帰直線のまわりに観測値が分布する人工データを生成し，各回帰直線のパラメータと誤判別の頻度が 2 直線間の距離に依存するかを調べた．最後に，混合回帰分析モデルの活用できる状況とモデルの限界を調べるため，フィッシャーのアイリスデータにこの手法を適用した．以上の結果，混合正規分布モデルでは，推定の精度は要素分布間の距離に依存し，混合回帰分析モデルでは，回帰直線間の最遠距離が大きいほど推定精度が良い傾向にあることがわかった．また，混合回帰分析モデルは，独立変数の定義域にわたって誤差分散が一定であることを前提としているため，混在する回帰直線ごとに独立変数の観測値が分布する範囲が大きく異なる場合には回帰係数を正しく推定できないことがある．