1D3-1

多目的最適化法による適切なモデル群の探索

Exploring statistical model spaces using multiobjective genetic algorithms techniques

松香敏彦 *1
Toshihiko Matsuka

*1千葉大学
Chiba University

1. はじめに

統計モデルの多くは学説や理論を基に構築され、データとの適合度によって、モデルやその基となった理論の妥当性が検証されてきた。対象となる統計モデルが 1 つの場合は、統計的に有意な結果が得られた場合に、そのモデルの統計的妥当性（もしくは、統計的非妥当性の可能性は低いという結果）が示されたことになる。例えば、分散分析や回帰分析などでは、従属変数中の分散を独立変数で説明できる割合が有意に高いモデル、loglinear analysisn などでは、モデルの予測値と観測されたデータが有意に乖離しないモデルなどが妥当なモデルだとされる。複数のモデルを比較する際は、AIC などのように単なるデータとの適合度ではなく、モデルの複雑性を加味し、より汎化能力の高いモデルを採択することが一般的である。どちらの方法も、最終的には 1 つ、もしくはごく少数のモデルを「適切」なモデルとして採択し、その解析結果から変数間の関係や構造を理解・考察することが一般的である。これらの統計モデルによって様々な学説や理論が検証され科学は発展してきた。

一方で、逆のアプローチとしてデータを基に複数の適切なモデルを探索し、新たな仮説を生成するアプローチも考えられる。本研究では進化アルゴリズムを基礎とした多目的最適化法 [Deb, 01] を用いてモデルスペースを探索し、複数の適切なモデル群を探索・識別する例を紹介する。

1.1 多目的最適化法

新たな仮説の生成を目的としたデータ解析として、複数の適切なモデル群を識別するためには、複数の目的関数が必要となる。これらの複数の目的関数を総合的に評価することによって、Pareto-optimal なモデル集合、つまり、各モデルは他のどのモデルにも支配されない「適切なモデル群」を識別することが可能となる。図 1 は Pareto-optimal な解集合の例である。これは、20 の独立変数を持つ回帰モデルの Pareto-optimal な集合であり、目的関数は学習データへの適合度（X 軸: adjusted R^2）と新規テストデータにおける誤差（Y 軸: Sum of Squared Residual）である。つまり、図 1 はモデルの（学習データへの）適合度と汎化能力を目的とした場合の適切なモデル群を表している。

この例では、複雑度と誤差の 2 つを目的関数としたが、目的関数は 2 つ以上であれば、どのようなモデルの評価の指標であっても、またいくつ用いても（確率的）に Pareto-optimal な解集合を識別することが可能である。

連絡先: 松香敏彦，千葉大学文学部，263-8522 千葉県千葉市稲毛区弥生町 1.33，043.290.3578 (Voice)，043.290.2278 (Fax)，matsuka@cogsci.L.chiba-u.ac.jp

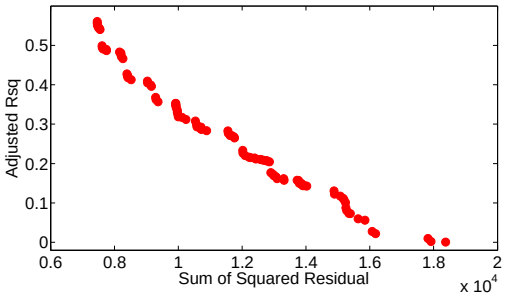

図 1: 実験 1A の結果、X 軸は学習データにおける適合度、Y 軸はテストデータにおける誤差を示している。

2. 実験

多目的最適化法によるモデル探索の有用性を検証するため、2 つの実験をおこなった。実験 1 では生成されたデータを基に回帰モデルを探索し、実験 2 では実データを基に、共分散構造分析モデルの探索をおこなった。

2.1 実験 1

実験 1 では、生成されたデータから、進化アルゴリズムを用いた多目的最適化法 (MOGA) を用いて、適切な回帰モデル群を探索する。データは 20 の一様分布 (a=0,b=1) に従う説明変数からなり、その内 10 変数を用いて目的変数を生成した。目的変数と従属の関係にある説明変数の回帰係数も一様分布 (a=10,b=15) に従うものを用いた。説明変数と関係のない目的変数の独自の要素は正規分布に従うものとした ($\mu = 0$,$\sigma = 10$)。実験 1A では、目的関数を学習データにおける適合度 (Adjusted R^2) とテストデータにおける誤差 (SSR) とし、実験 1B では、目的関数をベースモデル (目的変数と従属関係にある 10 の説明変数の内5 変数を含むモデル）との距離と SSR とした。実験 1A、1B 共に、学習データは 120、テストデータは 80 とした。

2.1.1 結果

図 1 は実験 1A の結果であり、学習データにおける適合度とテストデータにおける誤差を目的関数とした pareto-optimial な回帰モデル群であることを示している。図 2 は、これらのモデル群がどのように定義されているかを表している。図 3 は実験 1B の結果であり、ベースモデルとは異なる（距離のある）汎化能力のあるモデル群を模索するのに有益な情報となっていることが分かる。

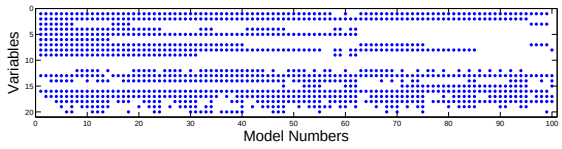

図 2: 実験 1A の結果:モデル定義。X 軸は SSR の低い順序で並べたモデルを表していて、Y 軸はモデルにおける各変数の有無を表している。SSR 値の低いモデルでは変数 1 10 の含有率が高く、SSR 値の高いモデルでは、変数 12 と 16 の含有率が高い。

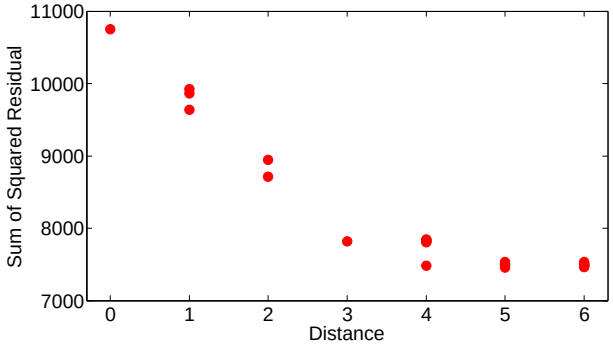

図 3: 実験 1B の結果。X 軸はベースモデルからの距離 (city-block distance)、Y 軸はテストデータにおける誤差を示している。ベースモデルに 4 6 変数を変更することによって、大幅にモデルの汎化能力を強化できることが分かる。

表 1: 実験 2 の結果: モデルの定義

Model	F	V1	V2	V3	V4	V5	V6	F2
Model1	F1	✓	✓	✓		✓	✓	
	F2	✓		✓		✓	✓	−
Model2	F1	✓	✓	✓		✓	✓	
	F2	✓	✓	✓				−
Model3	F1		✓	✓		✓		✓
	F2	✓			✓	✓		−
Model4	F1		✓	✓	✓			✓
	F2	✓			✓		✓	−
Model5	F1	✓	✓	✓	✓	✓	✓	
	F2	✓	✓			✓	✓	−
Model6	F1		✓	✓	✓	✓	✓	✓
	F2	✓				✓	✓	−
Model7	F1	✓	✓	✓	✓			✓
	F2	✓				✓	✓	−
Model8	F1	✓	✓	✓	✓			✓
	F2					✓	✓	−

表 2: 実験 2 の結果: モデルのフィット 指標

Model	Df	χ^2	P	P(RMSEA <0.5)
Model1	1	< .01	.984	.992
Model2	2	.07	.966	.991
Model3	3	.48	.922	.988
Model4	4	1.22	.875	.987
Model5	5	4.54	.474	.915
Model6	6	5.04	.531	.950
Model7	7	5.38	.614	.974
Model8	8	9.27	.320	.920

2.2 実験 2

実験 2 では、実データを基に共分散構造分析を探索的におこなった。共分散構造分析の入門書 [Kenny,98] にある、中学生 556 名の教育に関する 6 つ顕在変数からなるデータを用いた。変数の詳細は本研究の目的との関連性が低いため省略する。実験 1 と同様に、MOGA を用いて適切なモデル群を探索した。目的関数は複雑度(Df:自由度) とデータとの適合性(χ^2 値) を用いた。一世代の人口を 30 とし、1000 世代の進化アルゴリズムによる探索をおこなった。顕在変数の数が6 つということから、 パラメターの同定問題などを踏まえ、因子の数は最大 2 つまでとした。

2.2.1 結果

表 1 と 2 は、30 モデルの内、モデルで再現された共分散行列とデータから得られた共分散行列の乖離が統計的に有意でないもの、つまり、統計的に有意なモデルを示している。表 1 はモデルの定義を表しており(F は因子、V は顕在変数)、✓ の有無が因子と顕在変数、因子間の関係の有無を示している。表 2 はそれぞれのモデルのデータとの適合性を表す指標を示している。なお、このデータを解析したオリジナルの研究 [Kenny,98] では、表にある Model8 が採用されていた。自由度の値の近いモデル間では比較的似たモデルもあるが、全体として多様な統計的に有意なモデルを探索すること可能であることが示された。つまり、学説や理論からモデルを定義するのではなく、データからモデル探索を介し、仮説や学説を導く可能性が示唆された。

因子分析やパス解析など、共分散構造分析では同一のモデルを複数の異なった定義で表現することが可能である。例えば、Model 8 の場合、因子間の相関を排除しても、F1 と V5 および F1 と V 6 に従属関係があるとした場合も Model8 と全く同じ説明力をもつ同意義モデルとなる。実験 2 においても、ある複雑度において、複数の同意義モデルが得られた。このような一見異なってみえるが、実は同一のモデルであるなど、モデル定義によるモデル間の距離だけで多様性を扱うには注意が必要である。

参考文献

[Deb 01] Deb K. *Multi-objective optimisation using evolutionary algorithms* Wiley, Chichester. (201).

[Kenny 98] Kenny, R. B., *Principles and Practice of structural equation modelling.* Guilford, NY. (1991).