線形判別式による文体の計量

石田 基広 *

概要

Stylometrics with Linear Discriminant

ISHIDA motohiro

This paper deals with subjects of authorship attribution through statistical analysis. Critics often assume that there are stylistic differences between authors, but it is obscure, how to discriminate the styles.

In this paper, 12 works by two German writers, Theodor Storm and Adalbert Stifter, are analysed through one of the multivariate analysis methods, Linear Discriminant Analysis.

The basis for analysis is a word list of 8 high frequent particles in the 12 works, which is made by Java and Perl programs. Based on the list, ten of the works are used to compute discriminants. The rest two works are set aside for "validation estimate". And the difference between the styles are computed with Mahalanobis distance.

Various equations can be composed according to combinations of selected variables, and they are compared based on statistical tests. The best equation is a non-linear one, but for simplicity, a second best equation, which is linear, is preferred. This equation, using two variables, manages to discriminate the two authors, and a validy of Discriminant Analysis for literary styles is confirmed.

* ishida-m@ias.tokushima-u.ac.jp, 2004/09

1 はじめに

筆者は, 線形および非線形の解析手法を応用した, 言語, 文体分析の実験を繰り返している. 本稿では, 19 世紀の 2 人のドイツ語作家の文体について, 2 群の線形判別分析を試みた結果についてまとめる. 文体の計量研究においては, そもそも何が統計量となりうるのか, またデータの量はどれほど必要なのか, などの問題の検討が続いているが, これらについては文献 [1, 5, 12, 14, 16] を参照頂きたい. *1

2 不変化詞の分布

2.1 対象テキストと不変化詞

本稿で取り上げた作家は, Adalbert Stifter (1805 - 1868) と Theodor Storm (1817 - 1888) である. この二人を取り上げたのは, コーパスデータが入手しやすいこと, また日本でも良く知られており, 邦語の解説文献が多くあるなどの理由による.*2 ちなみに文学史書の中には, この 2 人は, ともに「詩的リアリズム」派に属するかのような分類もなされているが [7] , 分類の客観的基準 (あるいはアルゴリズム) が明示されているわけではない (この文献に限った話ではないが).

さて本稿では, 2 人の作家の計 12 作品のコーパスから, 不変化詞 (Partikel, particles, あるいは function words) と言われる単語を抽出した. 作家判別, あるいは一般にテキストを分類するのに有効な変数, さらにはテキストの分量 (長さ) の選択にはさまざまな考え方があるが, 詳細は [6] を参照されたい. 本

*1 また "Literary and Linguistic Computing", "Computers and the Humanities" などでもしばしば取り上げられている.

*2 したがってランダムサンプリングの条件を満たすものではない.

番号	作家	作品名	発表年度	備考
1	Stifter	Bergkristal	1853	教師データ
2	Stifter	Bergmilch	1853	教師データ
3	Stifter	Granit	1853	教師データ
4	Stifter	Kalkstein	1853	教師データ
5	Stifter	Turmalin	1853	教師データ
6	Stifter	Katzensilber	1853	交差妥当化データ
1	Storm	Aquis submersus	1878	教師データ
2	Storm	Boetjer Basch	1887	教師データ
3	Storm	Die Regentrude	1866	教師データ
4	Storm	Die Soehne des Senators	1881	教師データ
5	Storm	Auf dem Staatshof	1859	教師データ
6	Storm	Auf der Universitaet	1862	交差妥当化データ

表1　対象としたテキストデータ

稿では表(1)にあるテキストを対象とした.[*3]

抽出にはJava言語とPerl言語による自作プログラムを利用した.[*4] 不変化詞の解析データとしては, 例えばn-gramを求めることも検討されるが, 今回は, 単純に出現回数を計測した. [11]

ところで不変化詞を抽出したのは, これらの頻度が, 一般にテキストの内容に影響されないと考えられているからである. [12, p. 13] 本稿では, 2人の作家の文体を, 不変化詞の頻度によって判別するのを目的とする. したがって頻度分布が作品, あるいはジャンルごとに異なるような語彙では, 「作家」判

*3 『石さまざま』は既出の短篇をまとめたものであるが, 単行本化にあたり加筆修正されており, 出版年度はすべて1853とした.

*4 http://www.ias.tokushima-u.ac.jp/linguistik
また Mason [13] を参照.

別にふさわしくない.

筆者の作成したプログラムが抽出する不変化詞は 23 種類ある. ただし本稿の実験では, ケース数 (作品数) が 10 作と少ないため, これらを一度に投入すると, 計算過程で必要となる逆行列が求められない. そこで先行研究などを踏まえ, 8 個の不変化詞を選びだした.[2, 3] この中から, ここで取り上げた 2 人の作家の文体判別に有効な不変化詞の最小の集合を発見するのが本稿の目的である. ただし, この結果が, 19 世紀のドイツ語作家, あるいはドイツ語の著作物一般にあてはまる訳ではない. また本来は, 8 種の不変化詞を選び出す手続きそれ自体を記述すべきであろうが, これらについては現在準備中の別稿に譲る.

なお今回の解析では, 抽出した 12 作品のうち, データ解析には 10 作品のみを計算対象とした. これを「教師データ」とする. 残りの 2 作は, 10 作品による解析結果が, どの程度実用的かを調べるためのデータ である. これを「交差妥当化用データ」とする. [22] 先の表 (1) の備考にその区別を記しておいたが, それぞれの作家について, 番号 6 の作品は解析用データには含めない. これらは, 解析結果が出た段階で, 解析結果を応用するためのデータとして利用する.

表 (2), 表 (3) は, 選択した不変化詞が 2 作家のテキストで使われている頻度である.*5 テキストそれぞれの総語数は異なるので, ここでは 1000 語単位に換算してある. なお抽出にあたっては, テキスト全体ではなく, いわゆる「地の文」のみを対象としている. すなわち主人公の発話などの「会話文」は省いている. *6

*5 紙幅の都合で小数点 2 位までを表示する.

*6 ここでは単純に, 引用符に囲まれた部分はすべて会話文として処理している. この方法では, テキスト解釈としては不適当な分類を避けられないだろうが, 主観的な判断よりは, 誤差が少なくなると考えた.

stifter	aber	auch	man	noch	nun	sich	so	und
英訳	but	too	man	still	now	oneself	so	and
1	5.47	5.74	4.94	3.74	1.80	9.42	8.82	42.16
2	5.62	4.31	9.28	4.83	1.43	11.63	8.36	42.49
3	3.93	3.70	1.62	2.54	2.77	6.94	4.63	45.86
4	2.51	6.76	3.06	3.45	1.80	8.33	6.21	38.05
5	5.61	6.16	5.61	3.87	1.50	9.32	5.21	36.60

表 2　Stifter における不変化詞の頻度 (1000 語単位)

storm	aber	auch	man	noch	nun	sich	so	und
1	10.55	5.91	1.02	5.10	2.29	4.53	8.16	29.78
2	10.40	4.54	0.23	5.70	1.61	8.93	4.16	40.45
3	8.18	3.89	0.19	5.64	1.55	12.46	5.64	36.22
4	8.79	4.04	0.59	5.58	1.66	9.03	2.73	30.18
5	6.31	2.25	1.57	4.73	1.35	6.99	4.96	32.37

表 3　Storm における不変化詞の頻度 (1000 語単位)

3　判別分析

3.1　手法の選択

本稿の目的は, 2 人の作家の文体を, 不変化詞の頻度によって分類しようというものである. したがって, 目的変量はカテゴリカルなデータであり, 説明変量は連続値である. 本稿では判別の方法として, 多変量解析法の一つである「判別分析」を用いる.[19, 21] なお判別分析と重回帰分析の間には密接な

関係のあることが知られている.[8, 17] そこで, 2 作家の判別に有効な判別式が導出された後, 分析手法をあらため, 重回帰分析を適用し, 変数選択の合理性について, 異なった観点からの検証を試みる.

さて解析そのものは, ソフトウェアパッケージに任せれば済むのだが, 以下, R 言語によるデータ解析の紹介の意味も含め, 少し細かく解析過程を記述する.[10, 18]

ところで 2 群のデータを判別分析するには, 大きく三つの方法がある. 第一に全体の分散に対する群間分散の比 (相関比) を最大にする方法がある. この方法で求めた相関比は, 重回帰分析における決定係数に等しく, また判別関数の係数と相関比の積から, 重回帰分析における偏回帰係数が求められる. [8] 第二に, 各群の重心を求め, その重心に対する各ケースの距離を求める方法 (マハラノビスの汎距離) がある. この方法では, 母集団の性質によって, 判別関数が一次の直線で求められる場合と, 二次曲線になる場合とがある. 第三に, 対数尤度比と閾値との差によって判別する方法である (ベイズ判別ルール).[15]

本稿では, 二つ目のマハラノビスの汎距離により, 8 種の変量から, 文体判別に最良の組合せを検討する. はじめに, 各変量を単独で説明変数に採用した場合の結果と, 係数の F 値を検討する. この結果をもとに, 複数の変数の組合せを順次検討していく. ただし 8 組すべてを使った組合せを検討するのではなく, 1 変量による解析結果を参考に, 幾つかの候補に絞りこみ, それらの組合せを検討する.

4 変量の選択

4.1 変量間の相関

はじめに 8 種の変数について, 相互の相関係数を求めておく. ただし紙幅の都合から, ここでは絶対値が 0.60 以上の相関係数を示す 4 種, すなわち

相関行列	Writer	aber	man	noch	und
Writer	1.00	0.83	−0.74	0.81	−0.69
aber	0.83	1.00	−0.49	0.83	−0.53
man	−0.74	−0.49	1.00	−0.31	0.42
noch	0.81	0.83	−0.31	1.00	−0.56
und	−0.69	−0.53	0.42	−0.56	1.00

表 4　aber, man, noch, und の相関係数

aber, man, noch, und (と作家 Writer) についてのみ, その相関行列を表 (4) に示す. (なおカテゴリカルな変量である作家名 Writer には, ダミー変数 (0, 1) を与えている.)

それぞれ作家 Writer との間に高い相関を示すが, さらに, この 4 種の変数相互の相関も高く, 「多重共線性」を示す可能性がある. ただし man と noch の相関は比較的小さく, それぞれの Writer に対する相関は大きいことが見て取れる.

実際の解析にはいる前に, 複数の群の判別式を求める前提として, そもそも群間に差があると仮定できるのか, 各群の母平均は異なるのか, さらには母分散共分散は同じであるのか, 検定しておかねばならない. 本稿では, 手順が逆になるが, 先に群分けに有効な変数の組を予め推定し, その変数を説明変量とした 2 群のデータを対象に, 母平均の差, 群間の差の検定を行なうこととする.

4.2　1 変量による判別

まず 8 種の変量それぞれについて, 単独で説明変量として選び出した場合のマハラノビスの距離を求める. また選択された変数について求めた係数の F 値を比較する. マハラノビスの距離による判別分析では, データそれぞれの

変数	マハラノビスの距離	F 値
aber	7.17903	17.94758
auch	0.8631681	2.157920
man	3.975088	9.93772
noch	6.473473	16.18368
nub	0.1371687	0.3429218
sich	0.09545074	0.2386269
so	0.6104945	1.526236
und	3.057304	7.64326

表 5　各変数のマハラノビスの距離と, 係数の F 値

等分散性が問題となる. もしも等分散性が仮定できるのであれば, 判別式は一次式で表せるが, 棄却された場合, マハラノビスの距離をそのまま用いることになる.

ここで説明変量として不変化詞を一つだけ選択した場合のマハラノビスの距離と, 選択された変数の係数の F 値を表 (5) に示す. マハラノビスの距離は, 2 群をプールした不偏分散で, それぞれの平均値の差を割り, これに 2 群の平均の差を乗じて求める. 自由度 $r, n_1 + n_2 - p - 1$ の F 分布の 5% 有意点は 5.32 である. ただしここで $r = 1, p = 1$ である. 表 (5) からは, aber, noch, man, und の順に有力な係数となることが見て取れる. 具体的に aber の解析手順と結果を以下に示す.

```
    [Aber を単独で説明変量とした場合]
> stormAber
[1] 10.557454 10.403822  8.180756  8.793821  6.317690
> stifterAber
[1] 5.479818 5.622385 3.937920 2.515723 5.613092
```

```
[それぞれの分散]
> stifterAberVar <- var(stifterAber )
> stormAberVar <- var(stormAber )
[プールされた分散]
> stifstomAberVar <-
(4* stifterAberVar + 4* stormAberVar) / 8
[係数]
> stifstomAberKeisu <-
(mean(stifterAber) - mean(stormAber))
/ stifstomAberVar
[２群の平均値をまとめたデータ行列]
> stifstomAberMean <-
rbind(mean(stifterAber), mean(stormAber))
> stifstomAberMean
          [,1]
[1,] 4.633788
[2,] 8.850708
[切片]
> stifstomAberSeppen <-
-mean(stifstomAberMean2 %*% stifstomAberKeisu)
[マハラノビスの距離]
> stifstomAberMaha <-
(mean(stifterAber) - mean(stormAber))
* stifstomAberKeisu
> stifstomAberMaha
[1] 7.17903
[ F 値]
> (5*5*(5+5-1-1))/((5+5)*(5+5-2)*1) *  7.17903
```

```
[1] 17.94758
```

しかしながら説明変量として aber だけを用いた判別分析では, 教師データも完全には分類できない. 求めた一次式に, もともとのデータを代入すれば, Stifter の場合は正の実数が, Storm であれば負の実数が求まる. ところが以下に示すように, ここで得られた一次式では, Storm の 5 番目の作品が, 誤って Stifter の作品と分類されている.

```
  [教師データの判別]
> (as.matrix(stifterAber)
%*% stifstomAberKeisu)+ stifstomAberSeppen
          [,1]
[1,] 2.149204
[2,] 1.906493
[3,] 4.774184
[4,] 7.195380
[5,] 1.922314
> (as.matrix(stormAber)
%*% stifstomAberKeisu)+ stifstomAberSeppen
           [,1]
[1,] -6.495137
[2,] -6.233589
[3,] -2.448965
[4,] -3.492667
[5,]  0.722783 *
```

F 値のもっとも高い aber を説明変数とした場合でも, 単独では, 教師データも完全に判別することができなかった. そこで F 値の高かった変数を組み合わせてみる.

4.3 aber と noch の組合せ

まずは 2 変量 {aber, noch} による判別分析を試みる. ちなみにこの二つの変量の単相関係数は 0.8 強であった.

判別式を求める前に, {aber, noch} 2 変量を説明変量とした 2 群の母分散共分散行列が等しいか検定を行なう. 等分散性が確認されれば, 判別式は一次式で表される. ここでは Box M 検定を行なう.

4.3.1 Box M 検定

Box M 検定では, それぞれの分散共分散行列と, プールされた分散共分散行列を求め, それぞれについて行列式を求める. これらを以下の式に当てはめる.

(1)

$$\chi_0^2 = \left\{1 - \left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n_1 + n_2 - 2}\right)\frac{2p^2 + 3p - 1}{6(p+1)}\right\} log_e \frac{|S|^{n_A + n_B - 2}}{|S|^{n_A - 1} + |S|^{n_B - 1}}$$

求めた値を, 自由度 $p(p+1)$, すなわち 6 の χ^2 分布と照合する.

```
  [ aber と noch の分散共分散行列]
> stormAberNochCov <- cov (stormAberNoch)
> stifterAberNochCov <- cov (stifterAberNoch)
[プールされた分散共分散行列とその逆行列]
> stifstomAberNochCov <-
(4*stifterAberNochCov + 4* stormAberNochCov)/8
> stifstomAberNochCovSolve <-
solve(stifstomAberNochCov)
[それぞれの行列式]
> stifterAberNochCovDet <- det(stifterAberNochCov)
```

```
> stormAberNochCovDet <- det(stormAberNochCov)
> stifstomAberNochCovDet <- det(stifstomAberNochCov)
```

以上の計算結果を式 (1) に当てはめ, Box M 検定を行なう.

```
  [Box M 検定]
> (stifstomAberNochCovDet^8)
/(stormAberNochCovDet^4 * stifterAberNochCovDet^4)
[1] 11.13076
> (1 - (1/4 + 1/4 - 1/(5+5-2))
* ((2*2*2 + 3*2 -1) / (6*(2+1))))
 * log(11.13076)
[1] 1.757082
```

計算結果の 1.757 は自由度 6 の χ^2 分布の 5 % 有意点 12.60 よりも小さな数値であり, 仮説は棄却されない. すなわち 2 群の分散共分散は等しいと仮定できる.

そこで判別関数として一次の判別式を求め, この判別直線によって 2 作家の判別を試みることとする.

4.3.2 aber, noch による一次式

以下, 2 群の判別関数を求める計算手順を簡単に示す. まず 2 群それぞれの分散共分散行列, またプールされた分散共分散行列を求める. [20, p. 74] 次に後者の逆行列を求め, これに 2 群の平均値の差を乗じ, 変量それぞれの係数を求める.

```
  [係数]
> stifstomAberNochKeisu <- stifstomAberNochCovSolve
%*% (stifterAberNochMean  - stormAberNochMean )
> stifstomAberNochKeisu
```

```
            [,1]
aber -1.193122
noch -2.461299
```

[切片]

```
> stifstomAberNochSeppen <-
-mean(stifstomNochAberMean
 %*%  stifstomAberNochKeisu )
> stifstomAberNochSeppen
[1] 19.17835
```

ところで先の相関行列から, aber, noch の間には高い相関関係が認められた. この相関の度合が, 判別関数に矛盾をもたらしていないか検討する. すなわち「多重共線性」が生じていないか確認する. この確認方法は簡単で, 係数の符号を, 両者の平均の差と比較すれば良い. 以下に示すように符号は一致しているので, ここでは多重共線性は生じていないと判断する.

[多重共線性のチェック]

```
> (stifterAberNochMean  - stormAberNochMean )
          [,1]
aber -4.216921
noch -1.662803
```

求めた係数と切片で, 教師データを判別してみる. 実測値を当てはめて計算した値が, 正の実数であれば Stifter, 逆に負の実数であれば Storm である.

[教師データの判別]

```
> (as.matrix(stifterAberNoch)
%*%  stifstomAberNochKeisu) +  stifstomAberNochSeppen
          [,1]
  1 3.4293056
```

```
     2 0.5627276
     3 8.2083820
     4 7.6628607
     5 2.9466106
> (as.matrix(stormAberNoch)
%*% stifstomAberNochKeisu) + stifstomAberNochSeppen
            [,1]
  1 -5.97114109
  2 -7.27105954
  3 -4.48522091
  4 -5.06075218
  5 -0.02171283
```

判別に成功している. 続けて交差妥当化用データを検証してみる.

```
     [交差妥当化用データの判別]
> (as.matrix(stifterstormTestAberNoch )
%*% stifstomAberNochKeisu)+ stifstomAberNochSeppen
          [,1]
  1  5.893237
  2 -3.927882
```

こちらも正しく判別している.

4.3.3 判別結果の検定

{aber, noch} による一次式は, 教師データと交差妥当化データの両方を正しく判別している. しかしこの結果に満足せず, aber だけの場合に比べて, 判別の精度が有意に上がっているか検定してみる. まずはマハラノビスの距離を

求める. 計算には次の行列式を利用する.

(2) $D^2 = t(\mathbf{x} - \mu)\Sigma^{-1}(\mathbf{x} - \mu)$

```
  [マハラノビスの距離]
t(stifterAberNochMean - stormAberNochMean)
%*% stifstomAberNochCovSolve
%*% (stifterAberNochMean - stormAberNochMean)
  [1,] 9.123955
```

ちなみに, この計算は以下のようにしても同じである.

```
> t(stifstomAberNochKeisu)
 %*% (stifterAberNochMean - stormAberNochMean)
  [1,] 9.123955
```

先の表 (5) に記載した, aber だけの場合のマハラノビスの距離 7.18 と, 2 変量に増やした場合のマハラノビスの距離を比較し, その F 値を求める. F 値は式 (3) で求められる. ただしここで $r = 1$ である.

(3) $$F = \frac{n_1 + n_2 - p - 1}{r} \frac{n_1 n_2 (D^2(p) - D^2(p - r))}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D^2(p - r)}$$

式 (3) に求めた値を代入し, 計算結果を自由度 $r, n_1 + n_2 - p - 1$, すなわち 1, 5+5-2-1 の F 分布の 5% 有意点 5.59 と比較する.

```
  [追加変数の F 値]
>  ((5+5-2-1)*5*5*(9.123955 - 7.17903 ))
/ ((5+5)*(5+5-2) + 5*5*7.17903 )
[1] 1.311729
```

新たに加えた変数 noch の係数が 0 との仮説が棄却されないので, 判別に効果の無い可能性が否定できない.

組合せ	マハラノビスの距離	F 値
aber, man	16.90259	6.557927
aber, noch	9.123955	1.311729
aber, und	11.63335	3.004159
man, noch	43.62023	26.88045
man, und	8.73087	4.63973
noch, und	9.609885	2.269597

表 6　各組のマハラノビスの距離と, 追加変数の係数の F 値

4.4　aber, man の組み合わせ

そこで, 別の組合せを検討することになる. 今, 変量の数は 4 種であるから, すべてのペアを検討するのであれば, 上記の解析を $\binom{4}{2}$ 通り行なうことになる. ここで結果を表 (6) に示すと, 単独で F 値の高かった {aber, man} の組合せについては, 変量 man を増やした効果を表す F 値は 6.56 であり, また {man, noch} の場合, man を追加変量とした場合の F 値は最高の 26.88 であった.

しかしながら下に示すように, {man, noch} の組合せでは, BoxM 検定で仮説が棄却される. つまり母分散共分散行列が等しいとはみなせなくなり, 判別式は一次式では表せない.

[それぞれの分散共分散行列式と, プールされた分散共分散行列式]

```
> stifterManNochCov <- cov(stifterManNoch )
> stormManNochCov <- cov(stormManNoch )
> stifstomManNochCov <-
(4*stifterManNochCov + 4*stormManNochCov) / 8
```

```
[行列式]
> stifterManNochCovDet <- det( stifterManNochCov)
> stormManNochCovDet <- det( stormManNochCov)
> stifstomManNochCovDet <- det( stifstomManNochCov)
[比]
>  stifstomManNochCovDet^8 /
( stifterManNochCovDet^4 * stormManNochCovDet^4)
[1] 632242749230
```

この結果を式 (1) に当てはめる.

```
  [Box M 検定]
>  (1 - (1/4 + 1/4 - 1/(5+5-2))
* ((2*2*2 + 3*2 -1) / (6*(2+1))))
* log ( 632242749230 )
[1] 19.81331
```

結果の値 19.81 は, 自由度 p(p+1) の χ^2 分布の 5% 有意点 12.60 よりも大きく, 帰無仮説は棄却される. すなわち {man,noch} の組み合わせの場合, 判別式は一次式では表せない. これは本質的な問題ではないのだが, 本稿では簡単のため, 一次式で表せる判別式を求めたい.

そこで, {aber,man} の組合せを検討する. はじめに Box M 検定を示し, 判別式が一次式で求められるか確認する.

```
  [Box M 検定]
[1] 7.800828
> qchisq(0.95,6)
[1] 12.59159
```

Box M 検定の結果より, 判別式は一次式としてよい. そこで以下, 係数, 切片, マハラノビスの距離を求める.

```
  [aber man]
[係数]
> stifstomAberManKeisu <- stifstomAberManCovSolve
%*% (stifterAberManMean  - stormAberManMean )
> stifstomAberManKeisu
            [,1]
  aber -2.436124
  man   1.585307
[切片]
> stifstomAberManSeppen <-
-mean(stifstomAberManMean %*%  stifstomAberManKeisu )
> stifstomAberManSeppen
[1] 11.96248
[マハラノビスの距離]
> t(stifterAberManMean - stormAberManMean)
%*% stifstomAberManCovSolve
 %*% (stifterAberManMean - stormAberManMean)
            [,1]
  [1,] 16.90259
```

新たに加えた説明変量 man の係数が有意であるかを検定するため, 式 (3) によってF値を求める.

```
  [係数の F 値]
>  ((5+5-2-1)*5*5*(16.90259  - 7.17903 ))
/ ((5+5)*(5+5-2)+5*5* 7.17903)
[1] 6.557927
```

計算結果を自由度 $r, n_1 + n_2 - p - 1$, すなわち 1, 5+5-2-1 の F 分布の 5% 有

意点5.59と比較すると有意である.

さらに多重共線性をチェックしておく.

```
[多重共線性]
> (stifterAberManMean  - stormAberManMean )
            [,1]
aber -4.216921
man   4.181936
```

係数の符号と一致する.

4.4.1 母平均および母群間の差の検定, 誤判別率

ここで2群の間に有意な差が認められるかを検定する. さらに求めた判別式によって標本を判別する際, 誤った判別を行なう確率を計算する. はじめに2群の母平均の差の検定を行なう. まず先に求めたように, それぞれの群の重心からのマハラノビスの距離は16.90であった. この値を式(4)に代入する. ここで D^2 は, 2群の重心間のマハラノビスの距離である.

$$F = \frac{n_1 n_2 (n_1 + n_2 - p - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)p} D^2 \tag{4}$$

```
[F 値]
> (5*5*(5+5-2-1))/((5+5)*(5+5-2)*2) * 16.90259
[1] 18.48721
```

このF値は, 自由度 p, $n_1 + n_2 - p - 1$ のF分布に従うから, 5% 有意点4.74と比較するが, 十分に有意である. よって2群の母平均には差がある.

同じく, ウィルクスの Λ 統計量によって2群間の差が有意と見なせるか検討する. Λ は, 全体の平方和・積和行列式で群内平方和・積和行列式を割った値である. ここで全体の平方和・積和行列は以下に示す通りである.

```
[全体の平方和・積和行列]
```

```
> stifstomAberManT <- (9 * cov(stifstomAberMan))
> stifstomAberManT
          aber       man
aber  64.27201 -34.91629
man  -34.91629  78.91786
[群内平方和・積和行列]
> stifstomAberManW <- 4*stifterAberManCov
+ 4 * stormAberManCov
> stifstomAberManW
          aber       man
aber 19.815959  9.170948
man   9.170948 35.196382
```

Λ は, それぞれの行列式の比である.

```
  [Wilks のラムダ]
> det(stifstomAberManW) / det (stifstomAberManT)
[1] 0.1591834
```

この Λ を以下の式 (5) に代入して F 値を求める.

$$(5) \qquad F = \frac{n-p-1}{p}\,\frac{1-\Lambda}{\Lambda}$$

```
  [F 値]
> ((10-2-1) / 2) * ((1 - 0.1591834 ) / 0.1591834)
[1] 18.48722
```

ちなみに, この値は先に求めた母平均の差の F 値と一致する.

誤判別率は, マハラノビスの距離 16.90 を用い, 以下のように求められる.

```
  [誤判別率]
```

```
> 1 - pnorm(sqrt(16.90259 ) / 2)
[1] 0.01990871
```

判別に失敗する確率は 2 % で, 高い成功率が得られた.

5 結果

5.1 求めた式によるデータの判別

今回は, Stifter と Storm の文体判別にもっとも有効な不変化詞の最小の組を求めるのが目的であるから, {aber, man} の組合せが適当と判断される.

念のため, aber と man で教師データを判別すると, 以下に示すように, 100% の判別率である. また交差妥当化データの判別結果をあげると, こちらも判別に成功している.

```
  [教師データの判別]
> (stifterAberManMean  - stormAberManMean )
          [,1]
aber -4.216921
man   4.181936
[切片]
> stifstomAberManSeppen <-
-mean(stifstomAberManMean %*%  stifstomAberManKeisu )
> stifstomAberManSeppen
[1] 11.96248
[教師データの判別]
> (as.matrix(stifterAberMan)
%*%  stifstomAberManKeisu) +  stifstomAberManSeppen
         [,1]
```

```
     1  6.452630
     2 12.982809
     3  4.939793
     4 10.694481
     5  7.186769
> (as.matrix(stormAberMan)
%*% stifstomAberManKeisu) + stifstomAberManSeppen
          [,1]
     1 -12.139699
     2 -13.015999
     3  -7.658064
     4  -8.518399
     5  -0.924321
[交差妥当化データの判別]
> (as.matrix (stifterstormTestAberMan)
%*% stifstomAberManKeisu) + stifstomAberManSeppen
         [,1]
     1 14.40589
     2 -7.86371
```

判別式によれば, 負の値が Storm に分類されるが, 交差妥当化データは正しく判別されている.

結局, Stifter, Storm の 2 作家を判別するのに最良の一次式は以下のようになる. ここで x_1 は aber を, x_2 は man を意味する.

(6) $z = -2.44\,x_1 + 1.59\,x_2 + 11.96$

5.2 散布図の作成

2 変数 {aber ,man} によって Stifter, Storm の二人を分けることに成功した. これを散布図 (1) に描き, 2 作家を分ける判別直線 (6) を引いてみる. なお散布図の R は Stifter を, M は Storm を表す.

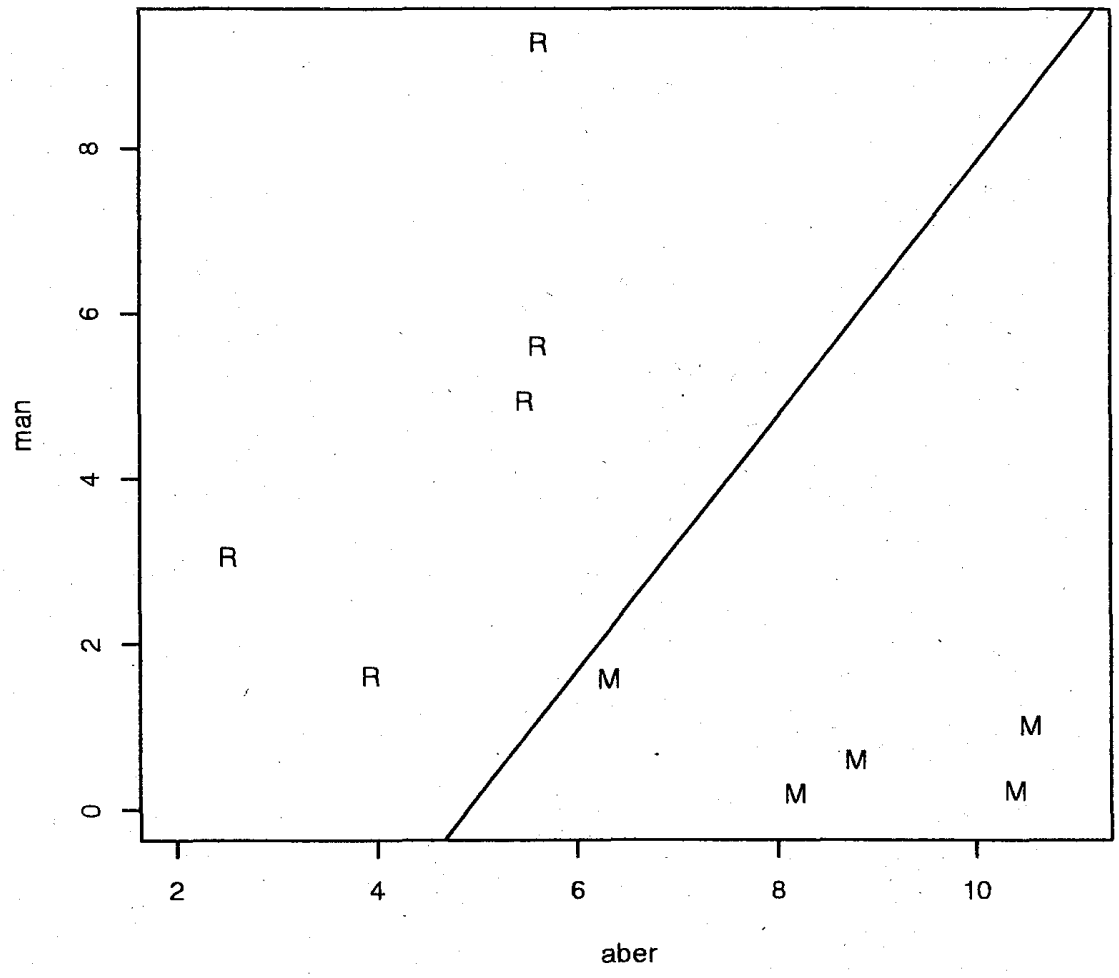

図 1 aber, man の度数による Stifter, Storm の分布

5.3 判別式の有効性について

以上, 線形判別式によって二人のドイツ語作家の「文体」を判別することに成功した. ただし, この方法の一般化については留保が必要である. そもそもここで解析された 12 作品は, ランダムに選ばれたのではなく, たまたまコーパスが入手可能であったからにすぎない. その意味で, 今回の線形式が二人の別の作品には有効ではない可能性も否定できない. *7

また仮にランダムサンプリングの原則に従うことができ, 二人の作家の判別に有効な式が算出されたとしても, aber, man の 2 変数に基づく判別式が, そのまま同時代の別のドイツ語作家, いわんやドイツ語作家一般に適用可能とは思えない. また判別対象が 2 作家以上となれば, 必要とされる説明変数の数も当然増えるであろう. もちろん, この集合の濃度が 2 で足りるとは到底思えない.*8

今回の結果である aber, man の有効性は, ある程度は一般的な傾向と考えられ, 他の作家の文体判別においても有力な説明変数の候補であろう. しかし他方で, そもそも文体が個性だとすれば, 判別に有効な変数の組合せが常に同じとは思えない. いずれによせ, 実際の解析作業では, 対象となる作家ごとに, その文体から基礎統計量を算出し, その分布について十分な調査が必要であろう.

なお文体の判別手法として, 最近では, ニューラルネットワークなどの手法も応用されている. [23] 実は, 今回のデータに関しては, 別にニューラルネットワーク解析を試みており, 十分満足いく解析結果を得ている. ニューラル

---

*7 また筆者は独語テキストデータを http://gutenberg.spiegel.de/ からダウンロードしたが, html 形式のデータは, 例えばウムラウトや引用符の扱いが統一されておらず, 解析プログラムの作成に苦労した. チェックをくり返したつもりであるが, 想定外の書式が含まれている可能性もあり, 解析には誤差が想定される.

*8 Hoover[4, 9] らの研究にあるように, 説明変数の数が増えれば, その分, 判別成績は改善する. 数百語の頻度を用いれば, 解析結果はほぼ完全となる.

ネットワークは強力な解析手法であるが, しかしながら最適な解が得られる過程が, 分析者にはブラックボックスであり, また解析結果の意味付けもほとんど不可能である.

その点で, 多少解析結果は劣っても, 線形の判別分析は, 「文学」解釈という次元からすれば, より望ましい手法でないかと思われる.

5.4 補遺 1 - Fisher の線形判別式

ちなみに統計解析言語 R では, 古典的な 「Fisher の線形判別式」による分析が可能である. 以下, その結果だけを示す.

```
  [Fisher の線形判別式]
> stifterstormAberMan2.lda
Call:
lda(as.factor(Writer)
~ aber + man, data = stifterstormAberMan2)
Prior probabilities of groups:
stifterter   storm
    0.5      0.5
Group means:
              aber       man
stifterter 4.633788 4.9058601
storm    8.850708 0.7239239
Coefficients of linear discriminants:
              LD1
aber   0.5925469
man   -0.3855996
```

切片は以下のように求める.

```
  [切片]
> -mean(stifterstormAberMan2.lda$means
%*% stifterstormAberMan2.lda$scaling)
[1] -2.909677
```

この分析によって教師データを分類すると, 以下のように完全に判別されている..

```
  [教師データの分類]
> predict(stifterstormAberMan2.lda  )
$class
 [1] stifter stifter stifter stifter stifter
 [6] storm   storm   storm   storm
Levels: stifter storm
```

この場合, aber の係数は 0.59, man は -0.39 と求められた. 切片の方は以下のようにして計算する.

```
  [切片]
> -mean(stifterstormAberMan2.lda$means
%*% stifterstormAberMan2.lda$scaling)
[1] -2.909677
```

よって一次式は以下のように求まる.

$z = 0.59x_1 - 0.39x_2 - 2.91$

5.5 補遺 2 - 重回帰分析による検定

さて 2 群の判別分析は, 重回帰分析と密接な関係があるのが知られている [17, p. 291]. ただし, ここでは選択した 2 変量を説明変量とする重回帰分析によって, 妥当な分析結果が得られるかを簡単に示すにとどめる.

今, カテゴリカルな変量である作家名 Writer に実数を与える. ここでは Stifter に 0.5 ($= n_2/(n_1 + n_2)$) を, また Storm には -0.5 ($= -n_1/(n_1 + n_2)$)) を与え, 重回帰分析を試みた.

[重回帰分析]

```
> summary(lm(writer
~ aber + man, data = stifterstormAberMan))
Call:
lm(formula = writer
~ aber + man, data = stifterstormAberMan)
Residuals:
     Min       1Q   Median       3Q      Max
-0.45402 -0.10835  0.03595  0.14623  0.25427
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.59507    0.29272   2.033  0.08156 .
aber        -0.12118    0.03412  -3.551  0.00933 **
man          0.07886    0.03079   2.561  0.03751 *
---
Residual standard error: 0.2384 on 7 degrees of freedom
Multiple R-Squared: 0.8408,Adjusted R-squared: 0.7953
F-statistic: 18.49 on 2 and 7 DF,  p-value: 0.001609
```

以上のように, 係数はともに 5 % 以下の水準で有意である. 決定係数は 0.84 とさほど高くは無いが, しかしながら判別式における変数選択 {aber, man} は, Stifter, Storm を分類するための説明変量として適当であったとみなして良いであろう.

参考文献

[1] アンソニー・ケニィ. 文章の計量—文学研究のための計量文体学入門. 南雲堂, 1995.

[2] 新井皓士. 『人麿歌集』と『ヘンリー六世』の帰属について - 多変量解析の計量言語学的応用の試み -. 一橋論叢, Vol. 119, No. 3, pp. 307–325, 2001.

[3] J. F. Burrows. *Computation into Criticism.* Oxford, 1987.

[4] J. F. Burrows. Delta: a Measure of Stylistic Difference and a Guide to likely Authorship. *Literary and Linguistic Computing*, Vol. 17, No. 3, pp. 267–287, 2003.

[5] Ross Clement and David Sharp. Ngram and Bayesian Classification of Documents for Topic and Authorship. *Literary and Linguistic Computing*, Vol. 18, No. 4, pp. 423–447, 2003.

[6] Richards Forsyth and David I. Holmes. Feature-Finding for Text Classification. *Literary and Linguistic Computing*, Vol. 11, No. 4, pp. 163–174, 1996.

[7] フリッツ・マルティーニ. ドイツ文学史. 三修社, 1979.

[8] 長谷川勝也. 判別分析および数量化理論 2 類の重回帰分析との関連性. 産業能率大学紀要, Vol. 15, No. 2, pp. 33–57, 1994.

[9] David L. Hoover. Multivariate Analysis and the Study of Style Variation. *Literary and Linguistic Computing*, Vol. 18, No. 4, pp. 341–360, 2003.

[10] 石田基広. 統計解析で探るシュティフター『石さまざま』: R 言語による分散分析と多変量解析. ドイツ語情報処理研究, Vol. 15, pp. 1–17, 2004.

[11] 金明哲. 助詞の n-gram モデルに基づいた書き手の判別. 計量国語学, Vol. 23, No. 5, pp. 225–240, 2002.

[12] 金明哲, 村上征勝, 永田昌明, 大津起夫, 山西建司. 言語と心理の統計. 岩

波書店, 2003.

[13] Oliver Mason. *Programming for Corpus Linguistics : How to Do Text Analysis with Java.* Edinburgh, 2000.

[14] 村上征勝. 真贋の科学. 朝倉書店, 1994.

[15] 長畑秀和. 多変量解析へのステップ. 共立出版, 2001.

[16] Michael P. Oakes. *Statistics for Corpus Linguistics.* Edinburgh, 1998.

[17] 奥野忠一, 久米均, 芳賀敏郎, 吉澤正. 多変量解析法:改訂版. 日科技連出版社, 1981.

[18] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3.

[19] 田中豊, 脇本和昌. 多変量統計解析法. 現代数学社, 1983.

[20] 田中豊, 垂水共之, 脇本和昌. パソコン統計解析ハンドブック 2 多変量解析編. 共立出版, 1984.

[21] 豊田秀樹. 非線形多変量解析—ニューラルネットによるアプローチ. 朝倉書店, 1996.

[22] 豊田秀樹. 金鉱を掘り当てる統計学—データマイニング入門. 講談社, 2001.

[23] Fiona J. Tweedie, David I. Holmes, and Simon Singh. Neural Network Application in Stylometry: the Federalist Papers. *Computers and the Humanities*, Vol. 30, pp. 1–10, 1996.