スパース正準判別分析に基づく万葉短歌の作者の分類とその特徴付け

応用統計学　Vol. 48, No. 3 (2019), 1–13

研 究 論 文

スパース正準判別分析に基づく万葉短歌の作者の分類とその特徴付け

電気通信大学　川　野　秀　一

関西大学　村　田　右富実

要　旨　万葉歌の研究において，歌の音の使用傾向から歌人の特徴を捉える場合がある．それぞれの歌人の使用している音の癖を読み取ろうとするものである．しかし，これまでは歌内で多く使用されている，もしくはほとんど使用されていない単一の音のみに着目した単変量的な解析や主観的な判断がほとんどであった．本論文では，複数の音を考慮に入れた統計解析を実行し，歌人の分類ならびにその音に基づいた特徴付けについて考察する．具体的には，まず，柿本人麻呂，山上憶良，大伴旅人の 3 歌人の短歌に着目し，各短歌内で使用されている音節から特徴量を作成する．その後，得られたデータに対してスパース正準判別分析を適用することにより，歌人の分類と各歌人に特徴的な音節の選択を行う．

1. はじめに

万葉集とは現存する日本最古の歌集であり，8 世紀中頃から後半に完成したと言われている．万葉歌は，短歌，長歌，旋頭歌，仏足石歌の 4 種から成る．日本古典文学研究，特に上代文学（奈良時代以前の日本文学）研究においては，これらの歌の成り立ちや，文字を持たなかった日本語が書記されるまでの歴史，さらには言語に表出された感情・感性などについて研究を行っている．そうした研究の一環として，これまで，歌の音の使用傾向から歌作者の特徴を捉えてきた歴史がある．たとえば，五味 (1951) は，

ささなみの　志賀の唐崎　さきくあれど　大宮人の　船待ちかねつ　(1)

という，万葉集巻 1 の 30 番歌（柿本人麻呂作）について，

> 「ささなみの」から「さきくあれ」迄にサ行カ行の音が多く明快な感じがし，自然の姿が昔と変らぬといふ内容と相俟つて明るく言ひ下して来るのだが，「ど」の助詞一つを境に音も内容も急に暗転し，「大宮人の　船待ちかねつ」の沈んだ調子は，末句の緊縮によつて些かの弛みも見せず人の胸を衝いて来る．

と述べ，これを「動乱調」と名付けた．たしかに，この言辞は柿本人麻呂の歌の一つの特徴を直感的に捉えてはいるが，客観的な解析からはほど遠いと言わざるを得ない．客観的な解析を目指

した研究も確かにあるが，歌内で多く使用されている，もしくは，ほとんど使用されていない単一の音のみに着目した単変量的な解析 (村田, 2009) がある程度である．

本論文では，複数の音を考慮に入れた統計解析を実行し，歌人の分類ならびにその音に基づいた特徴付けについて考察する．特に，柿本人麻呂，山上憶良，大伴旅人の 3 歌人に着目し，短歌を対象とする．歌人を分類する統計解析手法には，正準判別分析 (Fisher, 1936; 小西, 2010) を考える．正準判別分析によって，3 歌人の分類の判断が視覚的に可能となる．しかし，正準判別分析では，どの音節が各歌人に特徴的であるか判断することは難しい．そこで，Witten and Tibshirani (2011) により提案されたスパース正準判別分析を用いることとする．スパース正準判別分析は，正準判別分析とスパース推定法の一つである lasso (Tibshirani, 1996) を組み合わせた方法であり，判別分析と特徴量選択を同時に実行する統計解析手法である．スパース正準判別分析によって，3 歌人の分類を行うと同時に，各歌人に特徴的な音節を選択することを試みる．

ここで，万葉歌に対して統計解析を行った研究をまとめておく．新井 (1998) は，万葉歌に含まれる 50 音分布を基にして，万葉集に残る人麻呂歌集歌を主成分分析およびクラスター分析により解析している．村田・川野 (2014) では，短歌に含まれる母音に着目し，混合効果モデル (Laird and Ware, 1982) に基づく解析を試みている．その結果，山上憶良，東歌，防人歌の固定効果項の挙動が，他の歌人とは異なることを指摘している．短歌内で使用されている 1,286 種類の文字に着目して解析を行った研究もある (村田・川野, 2016)．この研究では，1 クラスサポートベクターマシン (Schölkopf *et al.*, 2000) を用いた外れ値検出を実行し，吉田宜，防人歌，補修部が外れ値として検出されている．1 クラスサポートベクターマシンを用いた他の研究としては，村田・川野 (2017) がある．村田・川野 (2016) では短歌内の文字に着目したが，村田・川野 (2017) では短歌内で使用されている 31 種類の音に着目している．その結果，巻 14 と巻 16 に収録されている歌が，他の巻とは異なる様相を呈しているという知見を得ている．村田・川野 (2019) は，万葉歌の異伝注記の特徴をレーベンシュタイン距離で捉えた後，検定手法を用いることにより，異伝の分類を行っている．また，万葉歌とは異なるが，古事記の音読注に対する基本統計量の算出 (伊藤, 2008)，クラスター分析による日本書紀の分類 (松田, 2009)，基本統計量とテキストマイニングを用いた和泉式部日記の異本間の比較 (太刀岡, 2014)，階層型クラスタリングによる源氏物語の解析 (小野, 2015) といった研究もある．以上のように，これまで万葉歌に対していくつかの統計解析手法が適用されてきたが，複数の音を考慮に入れ，歌人の分類とともに特徴量選択までも行う研究は，本論文が初めてである．

本論文の構成は次の通りである．第 2 節では，解析に用いる万葉短歌とその特徴量の作成方法について述べる．第 3 節では，解析手法であるスパース正準判別分析を紹介する．第 4 節では，上代特殊仮名遣いに触れた後，解析結果を示す．第 5 節でまとめと今後の課題を述べる．

2. 万葉短歌データセット

2.1. 万葉短歌

万葉集は全 20 巻から成り，収録されている歌にはそれぞれ歌番号が付されている．歌体の種類は先にも述べたように，短歌，長歌，旋頭歌，仏足石歌の 4 種である．短歌はよく知られているように，五七五七七の 5 句 31 音で表現されており，長歌は五七の 2 句が数回続き最後に七で終

わる形を基本とする．旋頭歌と仏足石歌は，ともに 6 句 38 音の歌であるが，旋頭歌は五七七五七七，仏足石歌は五七五七七七という句形になっている．また，収録されている歌数は，短歌は約 4,200 首，長歌は 265 首，旋頭歌は 62 首，仏足石歌は 1 首となっている．本論文では，一番歌数の多い短歌に着目して論を進める．

万葉短歌の原文は，たとえば前掲の (1) の歌であれば，

樂浪之　思賀乃辛碕　雖幸有　大宮人之　船麻知兼津

のように，すべて漢字である．この漢字の文字列を読み下すと，

ささなみの　しがのからさき　さきくあれど　おほみやひとの　ふねまちかねつ　　(2)

となる．このように一首ずつ読み下したテキストは，現在多数あるが，本論文では鶴・森山 (1972) を用いる．

2.2. 特徴量の作成と対象作者

万葉短歌に統計解析手法を適用するためには，各歌から特徴量を作成して数値データ化する必要がある．ここでは，各歌の音節パターンにより特徴量を作成する．具体的には，「あ」から「を」（濁音を含む）までの 68 種類の音節（なお，奈良時代には「っ」（促音），「ん」（撥音），「ゃ，ゅ，ょ」（拗音）はない）に着目し，これらの音節が一首中に何回用いられているかカウントすることによって特徴量を作成する．先程の (2) の歌を例に取ると，「あ」は 1 回，「か」は 2 回，「さ」は 4 回使用されており，「い」や「う」は 0 回使用されている．この操作によって，各歌を 68 次元の特徴ベクトルで特徴付けることができる．

解析の対象となる作者は，柿本人麻呂，山上憶良，大伴旅人とする．柿本人麻呂は，後世歌聖と呼ばれる万葉集を代表する歌人であり，山上憶良は一般的な万葉歌人とは違って，家族愛や貧困を歌う異色の歌人である．また，大伴旅人は，先の二人とは違って政府の中枢の一人であり，かつ，典型的な万葉歌人の一人である．これらの歌人を対象とした理由は，歌数がほぼ同数（柿本人麻呂が 70 首，山上憶良が 62 首，大伴旅人が 55 首）であり，かつ，前述したように万葉集の代表的歌人であるところによる．

3. データ解析手法

本節では，解析に用いたスパース正準判別分析を紹介する．スパース正準判別分析を紹介する前に，正準判別分析について触れておく．

3.1. 正準判別分析

いま，p 次元特徴ベクトル $\boldsymbol{x}$ に対して，群 G_j $(j = 1, \ldots, g)$ から n_j 個のデータ $\boldsymbol{x}_1^{(j)}, \ldots, \boldsymbol{x}_{n_j}^{(j)}$ が得られたとする．$n = n_1 + \cdots + n_g$ とするとき，群内分散は

$$\hat{\Sigma}_w = \frac{1}{n} \sum_{j=1}^{g} \sum_{i=1}^{n_j} \left(\boldsymbol{x}_i^{(j)} - \bar{\boldsymbol{x}}_j \right) \left(\boldsymbol{x}_i^{(j)} - \bar{\boldsymbol{x}}_j \right)^T \tag{3}$$

で与えられる．ここで，$\bar{\boldsymbol{x}}_j = \sum_{i=1}^{n_j} \boldsymbol{x}_i^{(j)} / n_j$ は群 G_j における標本平均である．また，群間分散は

$$\hat{\Sigma}_b = \frac{1}{n}\sum_{j=1}^{g} n_j \left(\bar{\boldsymbol{x}}_j - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}_j - \bar{\boldsymbol{x}}\right)^T \tag{4}$$

で与えられる．ここで，$\bar{\boldsymbol{x}} = \sum_{j=1}^{g}\sum_{i=1}^{n_j} \boldsymbol{x}_i^{(j)}/n$ である．

正準判別分析とは，各群のデータを超平面 $y = w_1 x_1 + \cdots + w_p x_p$ に射影することを考え，射影されたデータが群毎によく分離されるように超平面を逐次的に決定する方法である．まず，第 1 判別ベクトル $\boldsymbol{w}_1 = (w_{11}, \ldots, w_{1p})^T$ は，(3) 式と (4) 式を用いることによって定式化される最大化問題

$$\max_{\boldsymbol{w}_1} \boldsymbol{w}_1^T \hat{\Sigma}_b \boldsymbol{w}_1 \ \text{ subject to } \ \boldsymbol{w}_1^T \hat{\Sigma}_w \boldsymbol{w}_1 = 1 \tag{5}$$

を解くことによって得ることができる．次に，第 2 判別ベクトル $\boldsymbol{w}_2 = (w_{21}, \ldots, w_{2p})^T$ は，最大化問題

$$\max_{\boldsymbol{w}_2} \boldsymbol{w}_2^T \hat{\Sigma}_b \boldsymbol{w}_2 \ \text{ subject to } \ \boldsymbol{w}_2^T \hat{\Sigma}_w \boldsymbol{w}_2 = 1, \boldsymbol{w}_2^T \hat{\Sigma}_w \boldsymbol{w}_1 = 0 \tag{6}$$

を解くことによって得ることができる．この操作を繰り返すことによって，$k = 2, \ldots, K$ に対して，第 k 判別ベクトル $\boldsymbol{w}_k = (w_{k1}, \ldots, w_{kp})^T$ は最大化問題

$$\max_{\boldsymbol{w}_k} \boldsymbol{w}_k^T \hat{\Sigma}_b \boldsymbol{w}_k \ \text{ subject to } \ \boldsymbol{w}_k^T \hat{\Sigma}_w \boldsymbol{w}_k = 1, \boldsymbol{w}_k^T \hat{\Sigma}_w \boldsymbol{w}_i = 0 \qquad \forall i < k \tag{7}$$

を解くことによって得ることができる．ここで，$K = \min(g-1, p)$ である．

Witten and Tibshirani (2011) は，最大化問題 (7) とその等式制約を不等式制約に緩めた以下の最大化問題

$$\max_{\boldsymbol{w}_k} \boldsymbol{w}_k^T \hat{\Sigma}_b \boldsymbol{w}_k \ \text{ subject to } \ \boldsymbol{w}_k^T \hat{\Sigma}_w \boldsymbol{w}_k \le 1, \boldsymbol{w}_k^T \hat{\Sigma}_w \boldsymbol{w}_i = 0 \qquad \forall i < k \tag{8}$$

が同値であることを示している (Witten and Tibshirani, 2011, Appendix A.1)．さらに，彼女らは同様に最大化問題 (8) と最大化問題

$$\max_{\boldsymbol{w}_k} \boldsymbol{w}_k^T \hat{\Sigma}_b^k \boldsymbol{w}_k \ \text{ subject to } \ \boldsymbol{w}_k^T \hat{\Sigma}_w \boldsymbol{w}_k \le 1 \tag{9}$$

が同値であることも示している (Witten and Tibshirani, 2011, Appendix A.2)．ここで，

$$\hat{\Sigma}_b^k = \frac{1}{n} X^T Y (Y^T Y)^{-1/2} P_k^{\perp} (Y^T Y)^{-1/2} Y^T X \tag{10}$$

であり，X は $n \times p$ の計画行列，Y は i 番目のデータが j 群に属するならば $y_{ij} = 1$ となる $n \times g$ の指示行列，$P_k^{\perp}$ は $k = 1$ ならば $P_1^{\perp} = I_g$，$k > 1$ ならば $(Y^T Y)^{-1/2} Y^T X \hat{\boldsymbol{w}}_i$ $(i < k)$ と直交する空間への射影行列である．なお，I_n は $n \times n$ 単位行列，$\hat{\boldsymbol{w}}_i$ $(i < k)$ は (9) 式の解を表す．この定式化によって等式制約が外れ，パラメータ $\boldsymbol{w}_k$ の最適化が容易になるという利点を有している．

3.2. スパース正準判別分析

特徴ベクトル $\boldsymbol{x}$ の次元が高くなるにつれて，群内分散 $\hat{\Sigma}_w$ の推定が不安定，さらには求めることができなくなる．また，それぞれの判別ベクトルに寄与している特徴量を判断することも困難になる．

このような問題点を克服するために，Witten and Tibshirani (2011) は第 k 判別ベクトルを得る

方法として

$$\max_{\boldsymbol{w}_k}\left\{\boldsymbol{w}_k^T\hat{\Sigma}_b^k\boldsymbol{w}_k-\lambda_k\sum_{j=1}^{p}\left|\hat{\sigma}_j w_{kj}\right|\right\}\ \text{subject to}\ \boldsymbol{w}_k^T\tilde{\Sigma}_w\boldsymbol{w}_k\le 1 \tag{11}$$

の最大化問題を提案し，この方法をスパース正準判別分析と呼んだ．ここで，$\tilde{\Sigma}_w=\mathrm{diag}\left(\hat{\sigma}_1^2,\ldots,\hat{\sigma}_p^2\right)$，$\hat{\sigma}_j^2$ は群内分散 $\tilde{\Sigma}_w$ の第 (j,j) 成分，λ_k は正の値を取る正則化パラメータである．(11) 式から，判別ベクトルのいくつかの要素をぴったりと 0 と推定する，つまり，スパース推定することができ，判別ベクトルに寄与している/ 寄与していない特徴量を選択することができる．また，(11) 式では，正則化パラメータを $\lambda_1,\ldots,\lambda_K$ と K 個用意する必要があるが，Witten and Tibshirani (2011) は $\lambda_k=\lambda\|\tilde{\Sigma}_w^{-1/2}\hat{\Sigma}_b^k\tilde{\Sigma}_w^{-1/2}\|$ とすることで 1 個の正則化パラメータ λ を用意する方法を提案しており，本論文でもこの方法を採用する．ここで，ノルム $\|\cdot\|$ は行列の最大固有値を表すことに注意しておく．

Witten and Tibshirani (2011) のスパース正準判別分析以外にも，正準判別分析の判別ベクトルをスパース推定する方法はいくつか知られている．Clemmensen *et al.* (2011) は，最適スコア法 (Hastie *et al.*, 1995) の考えにしたがい，推定するパラメータの最適化問題にスパース制約を課すことによって，正準判別分析のスパース推定を提案している．Qiao *et al.* (2009) では，Zou *et al.* (2006) のスパース主成分分析の損失関数と同様に，回帰分析の枠組みの下で正準判別分析の損失関数を提案し，その損失関数とスパース正則化項の同時最小化によりスパース正準判別分析を実現している．また，Clemmensen *et al.* (2011) による方法は，統計解析ソフトウェア R に含まれるパッケージ **sparseLDA** により実行可能である．

4. 解析結果

本節では，音節の情報に基づき作成された万葉短歌データに対して，スパース正準判別分析を適用した結果について述べる．スパース正準判別分析の計算には，R に含まれるパッケージ **penalizedLDA** を用いた．解析に用いたソースコードは `https://sites.google.com/site/shuichikawanoja/publications_japanese/JSAS_KawanoMurata.R` よりダウンロードすることができる．

4.1. 上代特殊仮名遣いについて

解析結果を述べる前に，上代特殊仮名遣いに触れておく必要がある．先程，(2) の歌の読みを

ささなみの　しがのからさき　さきくあれど　おほみやひとの　ふねまちかねつ

と記した．これは，古典文学一般に用いられている，いわゆる旧仮名遣いと呼ばれる仮名遣いで記したものであるが，奈良時代までは，「き（ぎ），け（げ），こ（ご），そ（ぞ），と（ど），の，ひ（び），へ（べ），み，め，よ，ろ，も」に，2 種類の書き分けが存在していたことが判明している（「も」は古事記においてのみ書き分けられる）．たとえば，「秋」の「き」には「岐，伎，吉，棄，枳」といった文字が用いられる一方，「月」の「き」には「紀，記，忌，幾，奇」などが用いられ，これらが混用されることはない．すなわち，旧仮名遣いよりも多くの音節を区別したことが知られている．この現象は上代特殊仮名遣いと呼ばれ，前者を「きの甲類」，後者を「きの乙類」と称

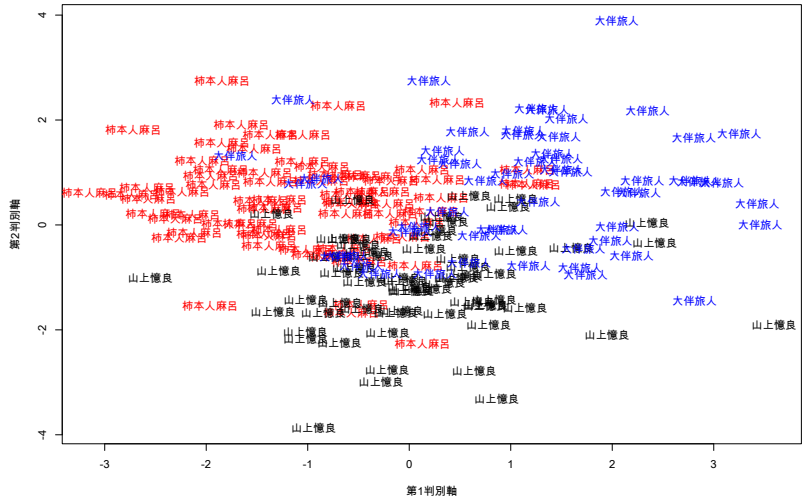

図 1. スパース正準判別分析によって 2 次元平面上に射影したデータ．赤色が柿本人麻呂，黒色が山上憶良，青色が大伴旅人を表す．

している．

いま，該当する音節について，甲類をひらがな，乙類をカタカナにして，先程の歌を記すと，

ささなみノ　しがノからさき　さきくあれド　おほみやひトノ　ふねまちかねつ

となる．必然的に，上代特殊仮名遣いを区別しない場合に 68 種類だった音節の種類は，87 種類に増えることになる．つまり，われわれが普段目にしている旧仮名遣いで解析した場合，それは我々が読んだときの差異を映し出すことになる．一方，我々は上代特殊仮名遣いの違いを認識できない．したがって，上代特殊仮名遣いを区別した場合の解析が区別しない場合の解析結果と同じであれば，奈良時代の人々も現代人と同じように 3 歌人の特徴を捉えていたことになる．このような観点から，以下，上代特殊仮名遣いを区別しない場合と，区別した場合のそれぞれについて論を進める．

4.2. 上代特殊仮名遣いを区別しない場合

はじめに，上代特殊仮名遣いを区別しないデータに対する結果を述べる．データは `https://sites.google.com/site/shuichikawanoja/publications_japanese/dataset.txt` よりダウンロードすることができる．音節「ぼ」は対象としている 3 歌人すべてにおいて不使用であったため，解析から除外した．スパース正準判別分析に含まれる正則化パラメータの値は 10 分割交差検証法で決定しようと試みたが，分割のパターンを変更すると最適値が $\lambda = 0.02$ から $\lambda = 0.13$ と大きく変動した．そこで，交差検証法ではなく，いくつかの候補を試した後に $\lambda = 0.09$ と主観的に決定した．

図 1 は，スパース正準判別分析を用いて柿本人麻呂，山上憶良，大伴旅人の 3 歌人を 2 次元平面上に射影したものである．ここで，横軸が第 1 判別ベクトル $\boldsymbol{w}_1$ により構成される第 1 判別軸 $\boldsymbol{w}_1^T\boldsymbol{x}$，縦軸が第 2 判別ベクトル $\boldsymbol{w}_2$ により構成される第 2 判別軸 $\boldsymbol{w}_2^T\boldsymbol{x}$ を表している．この図よ

表 1.　3 歌人に対する混同行列.

		予測した歌人		
		柿本人麻呂	山上憶良	大伴旅人
実際の歌人	柿本人麻呂	52	10	8
	山上憶良	6	47	9
	大伴旅人	4	10	41

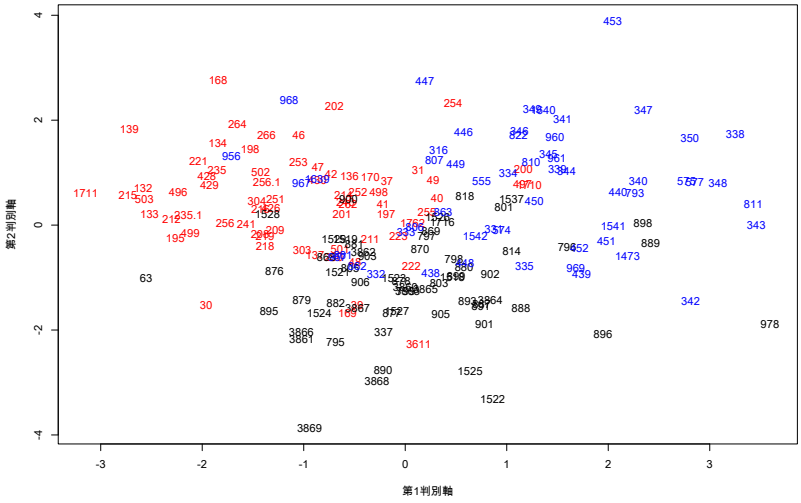

図 2.　スパース正準判別分析によって 2 次元平面上に射影した歌番号のデータ．赤色が柿本人麻呂，黒色が山上憶良，青色が大伴旅人の歌を表す．

り，大方，第 1 判別軸で柿本人麻呂と大伴旅人が分かれ，第 2 判別軸で山上憶良とその他が分かれていると考えられる．また，左上に柿本人麻呂，中央下に山上憶良，右上に大伴旅人が固まっていることがわかり，視覚的にもこれら 3 歌人はうまく分かれていることが見て取れる．表 1 は学習データに対する混同行列であり，行が実際の歌人，列が予測した歌人を表している．柿本人麻呂と誤判別することは小さく，山上憶良と誤判別することはやや大きいことがわかるがそこまで大きな値ではない．また，見かけ上の誤判別率は 25.1%であった．

図 2 は，図 1 の歌人の名前を歌番号に変えたものであり，各歌人の集団から外れている歌に対して次のような考察ができる．3611 番歌，1710 番歌（200 番歌のすぐ下にある歌）は柿本人麻呂の集団から大きく外れているが，これらの歌は柿本人麻呂らしくない歌ということで知られている（武田, 1957; 小島他, 1996; 伊藤, 1998; 稲岡, 2002）．山上憶良の歌で大きく外れている 1528 番歌は，山上憶良らしくないことで知られている（土屋, 1951）．また，63 番歌は山上憶良が一番若いときに詠んだ歌（土屋, 1949），978 番歌は山上憶良が亡くなる直前に詠んだ歌（斎藤, 1938）であることが知られており，これらの歌が x 座標の意味で正反対の位置にあるのは興味深い．大伴旅人で大きく外れている 956 番歌は，大伴旅人の中でも柿本人麻呂らしい歌と言われている（佐竹, 2000）．967 番歌，968 番歌は選別餞別歌への返歌で，類型性が高いため大伴旅人らしくないことが知られている．以上より，誤判別されている歌には何かしら理由があるものが多く，このことは解析結果が妥当であることを裏付けている．

表 2. 降順した第 1 判別ベクトルの推定値．はじめの 6 要素を記載．

音節	と	な	る	し	べ	む
推定値	0.358	0.289	0.268	0.242	0.181	0.172

表 3. 昇順した第 1 判別ベクトルの推定値．はじめの 6 要素を記載．

音節	み	ま	ど	お	い	や
推定値	–0.344	–0.305	–0.224	–0.210	–0.178	–0.161

表 4. 昇順した第 2 判別ベクトルの推定値．はじめの 6 要素を記載．

音節	ぶ	ね	ら	ち	べ	で
推定値	–0.313	–0.294	–0.269	–0.203	–0.176	–0.162

表 5. 第 1 判別ベクトルと第 2 判別ベクトルにおいて推定値が 0 となった音節．

第 1 判別ベクトル	き，く，こ，す，せ，そ，た，つ，ほ，よ，が，げ，ざ，ず，ぜ，づ，ば，び
第 2 判別ベクトル	や，へ，あ，ぞ，だ，こ，ほ，よ，が，ざ，ず，づ，ひ，め，を，ぐ，ろ，り，え，む

表 2 は，第 1 判別ベクトルを降順に並べ替え，そのはじめの 6 要素を示したものである．図 1 より，第 1 判別軸の値が大きくなるにつれて大伴旅人と判別されることがわかるため，表 2 に載せている音節は大伴旅人を特徴付けるものと考えられる．実際に，音節「し」は大伴旅人を特徴付ける音節と言われている（大久間他, 1982; 井村, 1984）．表 3 は，第 1 判別ベクトルを昇順に並べ替え，そのはじめの 6 要素を示したものである．第 1 判別軸の値が小さくなるにつれて柿本人麻呂と判別されるため，表 3 の音節は柿本人麻呂を特徴付けるものと考えられる．音節「ど」や「や」が柿本人麻呂で特徴的に使用されていることが，稲岡 (1985) によって示唆されている．表 4 は，第 2 判別ベクトルを昇順に並べ替え，そのはじめの 6 要素を示したものである．第 2 判別軸の値が小さくなるにつれて山上憶良と判別されるため，表 3 の音節は山上憶良を特徴付けるものと考えられる．音節「ら」は山上憶良が特徴的に使用している音節として知られている (高木, 1956).

最後に表 5 は，第 1 判別ベクトルと第 2 判別ベクトルにおいて，推定値が 0 となった音節をまとめたものである．第 1 判別ベクトルに含まれる「げ，ざ，ぜ，づ，ば，び」，第 2 判別ベクトルに含まれる「だ，ざ，づ，ぐ，ろ，り」といった濁音およびラ行の音は，日本語においては自立語の語頭に来ることがなく，必然的に使用例は少ないと考えられる．また，第 1 判別ベクトルの「が」，第 2 判別ベクトルの「や，へ，ぞ，が，を」は助詞であり，日本語においてその頻度は極めて高い．さらに両方に登場している「ず」は，いわゆる「打ち消しのず」であり，助詞同様，登場頻度が高い．このように，歌を詠む際にごくありふれた音節，ならびにほとんど使用されていない音節に対して，推定値が 0 となっていることがわかる．

4.3. 上代特殊仮名遣いを区別した場合

上代特殊仮名遣いを区別したデータに対する結果を述べる．このデータの解析の目的は，前述したように，4.2 節で議論した内容の奈良時代以前への一般化にある．データは `https://sites.`

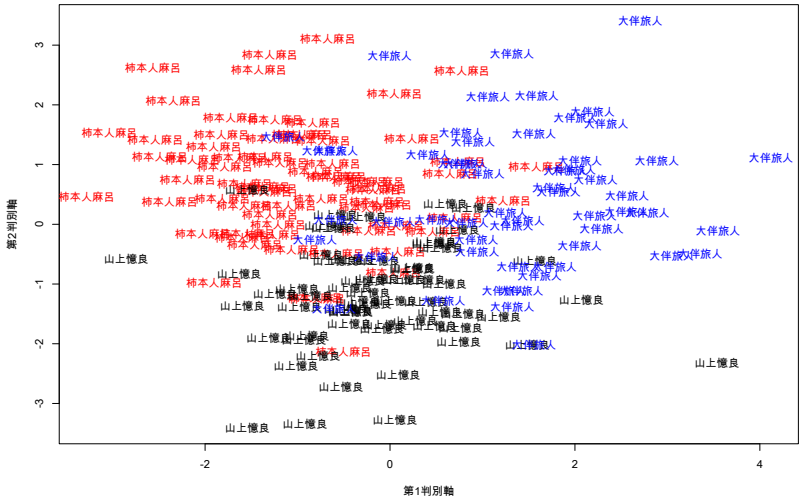

図 **3.** スパース正準判別分析によって **2** 次元平面上に射影したデータ．データは上代特殊仮名遣いを区別した場合．赤色が柿本人麻呂，黒色が山上憶良，青色が大伴旅人を表す．

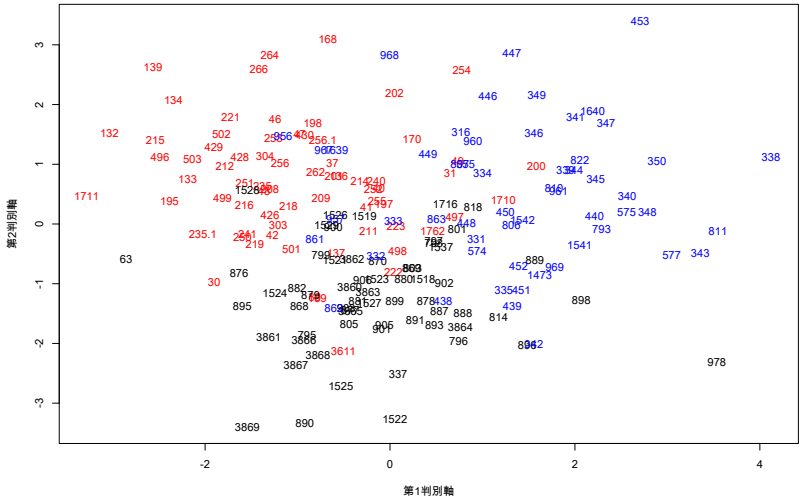

図 **4.** スパース正準判別分析によって **2** 次元平面上に射影した歌番号のデータ．データは上代特殊仮名遣いを区別した場合．赤色が柿本人麻呂，黒色が山上憶良，青色が大伴旅人の歌を表す．

google.com/site/shuichikawanoja/publications_japanese/dataset_tokushu.txt よりダウンロードすることができる．4.2 節と同様に，音節「ぼ」は解析から除外し，正則化パラメータの値は $\lambda = 0.09$ とした．なお，解析を実行すると，第 2 判別ベクトルの符号が 4.2 節で得られた第 2 判別ベクトルの符号と反対であったため，以降，本節で得られた第 2 判別ベクトルに -1 倍を掛けて論を進める．

図 3 は図 1 の 3 歌人の散布図，図 4 は図 2 の歌番号の散布図，表 6 は表 1 の混同行列にそれぞ

表 6.　3 歌人に対する混同行列．データは上代特殊仮名遣いを区別した場合．

		予測した歌人		
		柿本人麻呂	山上憶良	大伴旅人
実際の歌人	柿本人麻呂	55	8	7
	山上憶良	6	50	6
	大伴旅人	7	7	41

表 7.　降順した第 1 判別ベクトルの推定値．はじめの 6 要素を記載．データは上代特殊仮名遣いを区別した場合．

音節	ト	る	な	し	に	べ
推定値	0.319	0.312	0.265	0.246	0.196	0.160

表 8.　昇順した第 1 判別ベクトルの推定値．はじめの 6 要素を記載．データは上代特殊仮名遣いを区別した場合．

音節	ま	み	お	ド	の	ね
推定値	–0.308	–0.250	–0.218	–0.205	–0.178	–0.167

表 9.　昇順した第 2 判別ベクトルの推定値．はじめの 6 要素を記載．データは上代特殊仮名遣いを区別した場合．

音節	ら	ぶ	ね	べ	ち	ミ
推定値	–0.270	–0.268	–0.236	–0.183	–0.173	–0.157

表 10.　第 1 判別ベクトルと第 2 判別ベクトルにおいて推定値が 0 となった音節．データは上代特殊仮名遣いを区別した場合．

第 1 判別ベクトル	ら，べ，き，げ，た，ぜ，せ，ゲ，く，け， ば，こ，ほ，が，ざ，ず，づ，コ，へ，ゾ， キ，す，う，ノ
第 2 判別ベクトル	し，さ，メ，え，り，ヨ，ひ，じ，ぐ，を， エ，こ，ほ，が，ざ，ず，づ，コ，へ，だ， ヒ，よ，あ，そ，へ，や，ド，お，ま

れ対応している．これらの図表から，図 1，図 2，表 1 とほとんど大差ないことがわかり，4.2 節で議論した内容が本節でも成り立つと考えられる．

第 1 判別ベクトルならびに第 2 判別ベクトルの各要素について考える．表 7，表 8，表 9 は，順に 4.2 節で議論した，表 2，表 3，表 4 に対応している．表 2 と表 7 を見比べると，音節「な，る，し，べ」の 4 種類が共通している．さらに，表 2 の「と」が，表 7 の「ト」の乙類に置き換わっていることは興味深い．次に，表 3 と表 8 を見比べると，音節「ま，み，お」の 3 種類が共通している．ここでは，表 3 の「ど」が，表 8 の「ド」の乙類に置き換わっている．「ト・ド」はどちらも助詞の「と・ど」に該当する音である．「ト」についてははっきりしないが，「ド」はいわゆる逆接をあらわす（「〜あれど」など）．これは冒頭に記したように，五味 (1951) のいう「動乱調」の中心をなしているのは，この「ド」による逆接であるといってよい．最後に，表 4 と表 9 を見比べると，「ら，ぶ，ね，べ，ち」の 5 種類もの音節が共通している．このように，順番が前後しているものの，基本的に上代特殊仮名遣いを区別しないときの結果と同じ音節が挙がってきている．したがって，ここでも 4.2 節で議論した内容が成り立つと考えられる．

表 10 は表 5 の推定値が 0 になった音節に対応しており，上代特殊仮名遣いを区別した方が特徴量が多いため，より多くの音節が必要ないと判断されている．また，表 5 と表 10 から，第 1 判別ベクトルでは「き，く，こ，す，せ，た，ほ，が，げ，ざ，ず，ぜ，づ，ば」の 14 音節が共通

しており，第 2 判別ベクトルでは「や，へ，あ，だ，こ，ほ，よ，が，ざ，ず，づ，ひ，を，ぐ，り，え」の 16 音節が共通しており，多くの音節が共通していることがわかる．

以上より，上代特殊仮名遣いを区別したときの解析結果は，区別しないときの解析結果と同じような傾向を持っているため，4.2 節での解析結果は奈良時代以前にも一般化できるのではないかと考えられる．

5. まとめと今後の課題

本論文では，万葉短歌の作者の分類とその特徴を捉えるために，スパース正準判別分析による統計解析を試みた．データの特徴量は歌の音節のパターンに基づき作成し，作者として柿本人麻呂，山上憶良，大伴旅人の 3 歌人を考えた．解析の結果，スパース正準判別分析によって 3 歌人は分類され，また，各歌人に特徴的な音節が選択された．選択された音節の中には過去の研究で示唆されている音節もいくつか含まれており，今回の解析結果は，これまでの万葉歌研究の観点から見ても妥当であることがわかった．

スパース正準判別分析を適用する際，本論文では，10 分割交差検証法の結果を参考にして正則化パラメータの値を主観的に決定した．分割のパターンに依存することなく，かつ，客観的に正則化パラメータの値を決定するためには，10 分割交差検証法を 1 個抜き交差検証法に変更することが考えられる．しかし，将来新たな万葉短歌が大量に得られ，それらを識別・判別することはほぼ考えられないため，予測を目的とする交差検証法を採用することには疑問が残る．予測を目的としないときの正則化パラメータの決定方法については，これまで研究がほとんど進んでおらず，そのような決定方法に関する研究を今後推し進める必要がある．また，解析により選択されたいくつかの音節については，その妥当性が過去の研究により示唆されているが，多くの音節については選択された根拠がはっきりしていない．加えて，非常に柿本人麻呂らしいと言われている 200 番歌が，柿本人麻呂の集団から大きく外れており，その原因はいまのところ不明である．以上の問題点については，今後の課題としたい．

本論文の解析では，柿本人麻呂，山上憶良，大伴旅人以外のデータは用いていない．解析に用いていない大量のデータを活用するために，これらのデータを教師なしデータと見なし，半教師あり学習 (Chapelle *et al.*, 2006) を行うことが考えられる．しかし，3 歌人以外のデータにも既に作者情報が付与されているため，単に半教師あり学習を適用するだけでは有用な情報を抽出することは難しい．そこで，柿本人麻呂，山上憶良，大伴旅人と特徴量の分布が似ている歌を事前に抽出し，その歌を教師なしデータとして半教師あり学習に組み込むといった新たなモデリング手法の開発が望まれる．さらに，解析を行う際，本論文では音の並び，すなわち歌の時系列情報は全く考慮に入れていない．これらの情報を考慮に入れた時系列統計モデルや，近年発展が目覚ましい深層学習，特に，再帰型ニューラルネットワーク (Rumelhart *et al.*, 1986) への適用を考えることにより，興味深い知見を得ることができると予想している．これらの更なる解析方法についても，今後の課題としたい．

謝　辞　本稿の改訂に当たって，査読者の方から大変貴重なご指摘と適切なご意見をいただきました．ここに記して御礼申し上げます．本研究は日本学術振興会科学研究費補助金 (19K11854) の助成を受けたものです．

参 考 文 献

新井皓士 (1998).『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，**119**(3)，307–325．

Chapelle, O., Scholkopf, B. and Zien, A. (2006). *Semi-Supervised Learning*. MIT Press.

Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, **53**(4), 406–413.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), 179–188.

五味智英 (1951). 古代和歌. 至文堂.

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics*, **23**(1), 73–102.

井村哲夫 (1984). 万葉集全注（巻第五）．有斐閣．

稲岡耕二 (1985). 万葉集の作品と方法．岩波書店．

稲岡耕二 (2002). 和歌文学大系２．明治書院．

伊藤博 (1998). 萬葉集釋注八＜巻第十五＞＜巻第十六＞．集英社．

伊藤雅光 (2008).『古事記』の音読注に見られる「言語の経済性」について．計量国語学会，**26**(6)，177–195．

小島憲之，東野治之，木下正俊校注 (1996). 新編日本古典文学全集（9）萬葉集（4）．小学館．

小西貞則 (2010). 多変量解析入門―線形から非線形へ―．岩波書店．

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**(4), 963–974.

松田信彦 (2009). 日本書紀「区分論」の新たな展開―多変量解析の考え方を参考にして―．古代文芸論叢：青木周平先生追悼，239–254．

村田右富実 (2009). カイ二乗検定を用いた万葉短歌の声調の分析．萬葉，**202**，23–41．

村田右富実，川野秀一 (2014). 多変量解析を用いた万葉短歌の声調外在化について．美夫君志，**88**，29–37．

村田右富実，川野秀一 (2016). 多変量解析を用いた万葉歌の筆録者同定の可能性試論．上代文学，**117**，91–103．

村田右富実，川野秀一 (2017). 多変量解析から見る万葉短歌の一般性と特殊性―巻を単位として―．文学・語学，**220**，14–24．

村田右富実，川野秀一 (2019). 万葉歌における異伝注記の特徴―依拠情報不明異伝と文字情報依拠異伝の差異を基点に―．萬葉，**227**，73–92．

小野洋平 (2015). 源氏物語成立論の統計科学的再考察―村上・今西 (1999) を中心に―．計量国語学会，**29**(8)，296–312．

大久間喜一郎，森淳司，針原孝之編 (1982). 万葉集歌人辞典．雄山閣．

Qiao, Z., Zhou, L. and Huang, J. Z. (2009). Sparse linear discriminant analysis with applications to high dimensional low sample size data. *IAENG International Journal of Applied Mathematics*, **39**(1), 48–60.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

斎藤茂吉 (1938). 万葉秀歌．岩波書店．

佐竹昭広，工藤力男，山崎福之，山田英雄，大谷雅夫 (2000). 萬葉集〈2〉（新日本古典文学大系）．岩波書店．

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J. and Platt, J. C. (2000). Support vector method for novelty detection. *Advances in Neural Information Processing Systems*, **12**, 582–588.

太刀岡勇気 (2014). 中古日記文学の計量国語学的分析と異本間の関係性の客観分析―『和泉式部日記』と『更級日記』を題材に―．計量国語学会，**29**(6)，187–210．

高木市之助 (1956). 孤語．文学・語学，**2**，10–16．

武田祐吉 (1957). 増訂万葉集全註釈＜第十一巻＞＜本文篇九＞（巻の十五・十六・十七）．角川書店．

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58**(1), 267–288.

土屋文明 (1949). 万葉集私注＜第三巻＞．筑摩書房．

土屋文明 (1951). 万葉集私注＜第八巻＞．筑摩書房．

鶴久，森山隆 (1972). 万葉集．おうふう．

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society Series B*, **73**(5), 753–772.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.

著者連絡先：〒 182–8585 東京都調布市調布ヶ丘 1–5–1
川野秀一（Tel. 042–443–5620）
E-mail: skawano@ai.lab.uec.ac.jp

Japanese J. Appl. Statist. 48 (3) (2019), 1–13

Classifying and characterizing poets for *Manyo tanka* via sparse canonical discriminant analysis

Shuichi Kawano[1,*] **and Migifumi Murata**[2]

[1] The University of Electro-Communications
[2] Kansai University

Abstract

Manyo tanka is a short Japanese poem included in *Manyoshu* which is the oldest collection of Japanese poems. Since the short poems are composed by several poets, each poem has characteristics for each poet. Until now, the characteristics have been subjectively studied or have been investigated based on a single sound. In this paper, we use a statistical method to study the characteristics based on multiple sounds in *Manyo tanka*. In particular, we analyze the *Manyo tanka* dataset using sparse canonical discriminant analysis. This analysis uncovers inherent properties of poets for *Manyo tanka*.

Key words: Classical literature, Feature selection, Sparsity.

[*]Corresponding author
E-mail address: skawano@ai.lab.uec.ac.jp