resume（スライド補足資料）:「J-POP」のトレンド 1960-2019

—構造的トピックモデルによる推定—

小牧和哉（大阪大学人間科学研究科 M1）

1　分析モデル

トピックモデルは, クラスタリングにおいて, 特定の文書が 1 トピック「のみ (single)」に分類されるのではなく, 「複数の (mixed)」トピックに潜在的・確率的に分布していると考えるモデルである. トピックモデルは細分類として, LDA(Latent Dirichlet Allocation), CTM(Correlated Topic Model) 等が存在するが, ここで用いたのはトピック同士に相関を仮定し, ディリクレ分布などの事前分布の代わりに, 共変量 (covariates) を投入して回帰モデルに近似させる STM(Structural Topic Model) を用いた.

STM の数式的な表現は以下の通りである.

トピックの割合 (Topic Prevalence) における事前分布仮定:

$$\mu_{d,k} = X_{d\gamma k}$$

$$\gamma_k \sim Normal(0, \sigma_k^2), for \;\; k = 1...K-1$$
$$\sigma_k^2 \sim Gamma(s^\gamma, r^\gamma)$$

確率論的言語モデルにおける分布仮定:

$$\vec{\theta_d} \sim LogisticNormal(\Gamma' X_d', \sum)$$

$$z_{d,n} \sim Multinomial(\theta_d), for\ n = 1...N_d$$

$$w_{d,n} \sim Multinomial(\beta_{z_{d,n}}), for\ n = 1...N_d$$

ここで言う θ_d は, 文書ごとのトピックの出現確率分布のことを指しており, $\Gamma^T = [\gamma_1...\gamma_k]$ の document(文書)×K (トピック数)-1 の係数行列で表現される。Σ はトピックの分散共分散行列を指している. これらのパラメーターは, ロジスティック正規分布に従うことを仮定しており, Softmax 関数により, 合計は「1」となる. そして, この, document(d) 中の単語 (n) 単位における潜在トピック $z_{d,n}$ は θ_d をパラメーターとした多項分布に従う.

最後に, $z_{d,n}$ によって生成された観察者の手元にある観察される単語 $w_{d,n}$ は, 潜在トピックが割り当てられた単語分布 $\beta_{z_{d,n}}$ をパラメーターとした多項分布に従う.

トピックの中身における分布仮定:

$$\beta_{d,v}^{k} \propto exp(m_{v} + k_{v}^{y} + k_{v}^{k} + k_{v}^{y,k})$$

$$k_{v}^{y,k} \sim Laplace(0, \tau_{v}^{y,k})$$

$$\tau_{v}^{y,k} \sim Gamma(s^{k}, r^{k})$$

トピック K が与えられた状態でのドキュメントに出現する単語の確率分布 $\beta_{d,v}^{k}$ は, 単語の出現率の対数分布を表すパラメーター m_{v}, トピック独自の分散 k_{v}^{y}, 共変量の分散 k_{v}^{k} そしてそれらの交互作用による分散 $k_{v}^{(}y, k)$ によって表現される. このモデルは通称 SAGE（Sparse Additive Generative Model）と呼ばれる推定手法である. β の事前分布を新たに仮定することなく, トピックの固有性を対数加法モデルで表現することによって計算量を削減できることが特徴である. 分散はラプラス分布によってスムージングされており, スパースなデータにおいても推定値のロバスト性を維持できる. ちなみに, 共変量が用意されていない場合はトピックの確率を単純推定するのみとなる.

これらの階層的な分布仮定を置いたうえで, STM は, 上記に挙げた文書ごとのトピック比率 θ_{d} と単語の出現確率分布 $\beta_{d,v}^{k}$ の両パラメーターに加え, 外生変数である共変量（文書が生成された年, 著者, 文書の分類, 評価等）で構造化し（structuralize）, 数学的なアルゴリズム手法 (変分ベイズ, マルコフ連鎖モンテカルロ, EM 等) を用いて, モデル収束 convergence させることによって, 推定値を算出する.

事後分布は以下になる (詳細は Robers et al. 2016b, 2019).

$$p(\theta, z, k, \gamma, \Sigma | w, X, Y) =$$

$$\left(\prod_{d=1}^{D} LogisticNormal(\theta_{d} | \mu = X_{d}\gamma, \Sigma) \left(\prod_{n=1}^{N} Multinormal(z_{n,d} | \theta_{d}) \times Multinormal(w_{n} | \beta_{d,k=z_{d,n}} \right) \right)$$

$$\prod p(k) \times \prod Normal(\gamma_{k} | \sigma_{k}^{2}) \times \prod Gamma(\sigma_{k}^{2} | r^{\gamma})$$

where, d: 文書,n: 単語,k: トピック,w: 単語集合,θ_{d}: 文書中のトピック分布,$\mu = X_{d}\gamma, \Sigma$: 文書-トピック行列と分散・共分散行列,$z_{n,d}$: 各文書中の各単語における潜在変数 (トピック),$\beta_{d,k=z_{d,n}}$: トピックに割り当てられる単語分布,γ: 過剰適合を防ぐ事前分布

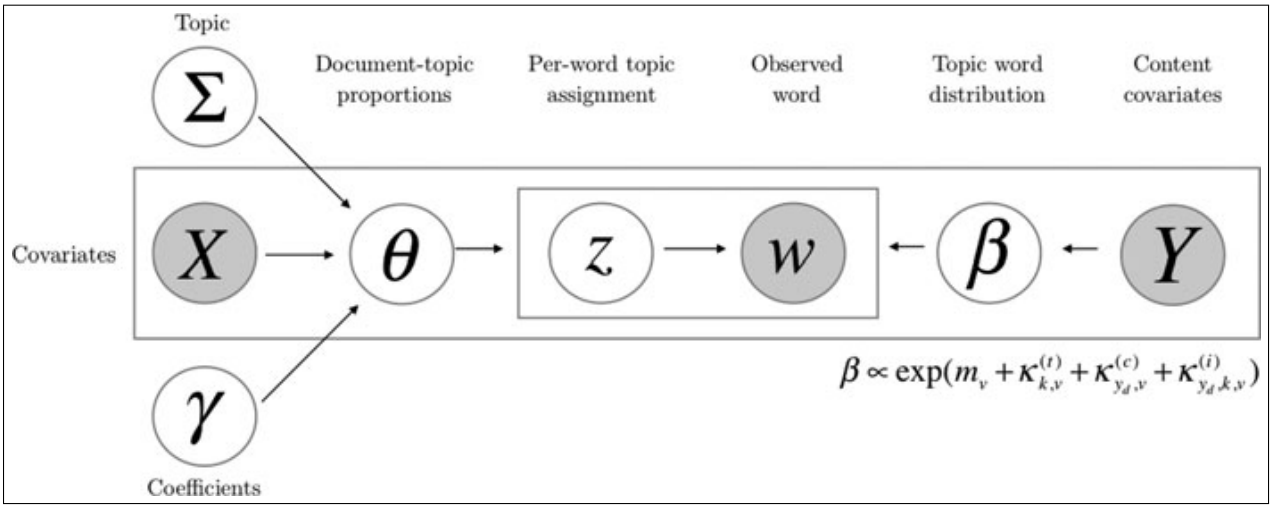

Figure 1: STM のグラフィカル・モデル (Roberts et al. 2016a)

1.1 記述統計量

Table 1: 記述統計表

	mean	sd	min	max	range
年	1989.16	17.20	1960	2019	59
発売後経過年数	30.21	17.38	0	90	90
歌手性別（女性）	0.40	0.49	0	1	1
歌手性別（男性）	0.56	0.50	0	1	1
作詞者性別（女性）	0.23	0.42	0	1	1
作詞者性別（男性）	0.75	0.43	0	1	1
総単語数	201.21	85.26	14	830	816

n=2,934

2　結果

2.1　補足：トピック数選定（１）

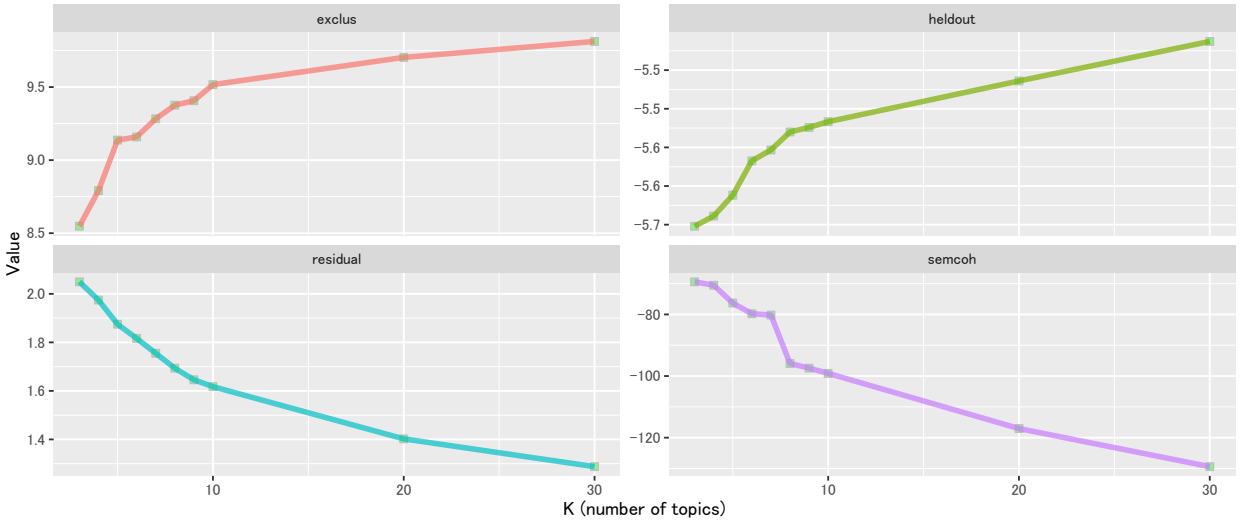

Fig.2　トピック数（K）別の各指標の推定値

- semantic coherence…意味論的なまとまり
- exclusivity…トピックの排他性
- residual…残差
- heldout…クロスバリデーションに基づく尤度

2.2　補足：トピック数選定（２）

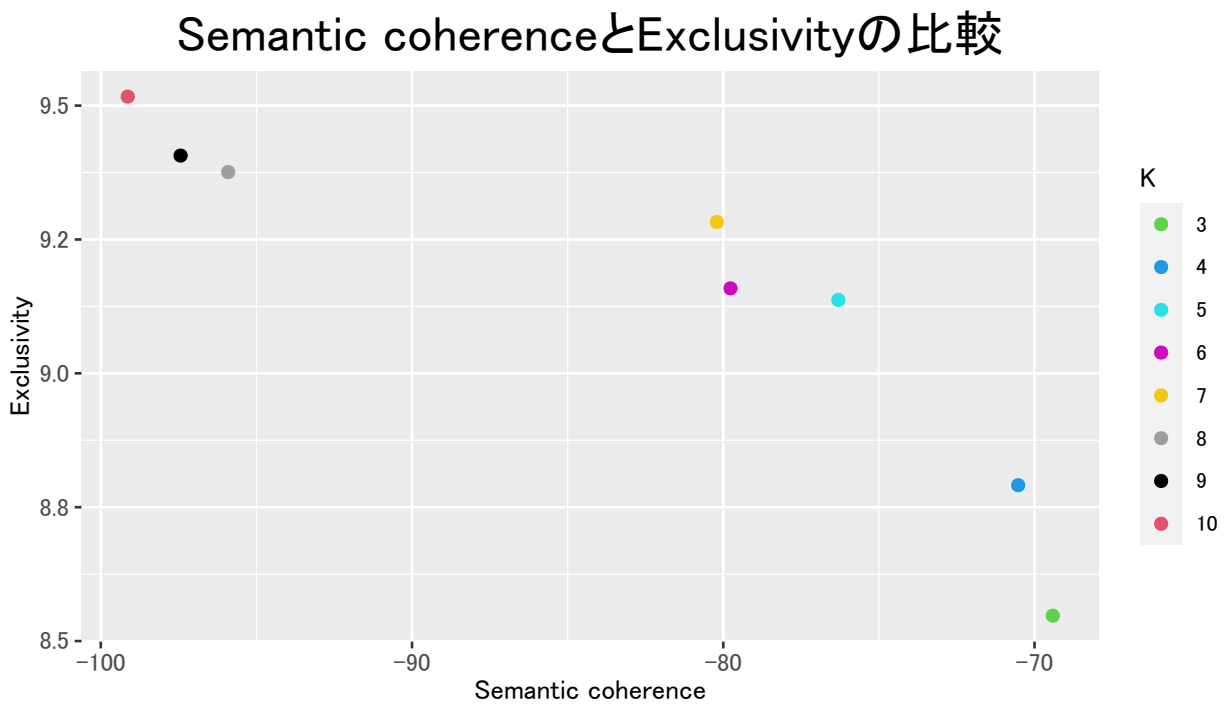

Table 2: トピック数における各指標の推移

K（トピック数）	Exclusivity	Semantic Coherence	heldout	residual
3	8.6	-69	-5.7	2.0
4	8.8	-71	-5.7	2.0
5	9.1	-76	-5.7	1.9
6	9.2	-80	-5.6	1.8
7	9.3	-80	-5.6	1.8
8	9.4	-96	-5.6	1.7
9	9.4	-97	-5.6	1.6
10	9.5	-99	-5.6	1.6
20	9.7	-117	-5.5	1.4
30	9.8	-129	-5.5	1.3

2.3 補足: トピック数選定（３）

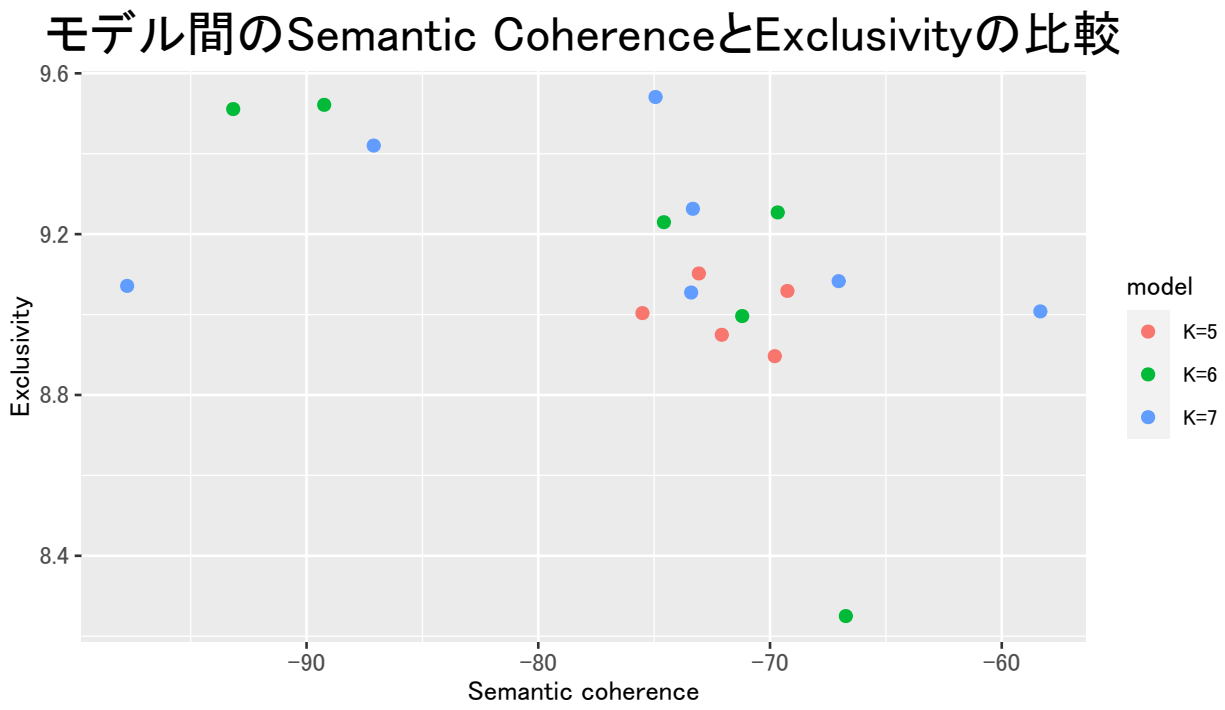

Table 3: モデル間のトピック数における各指標の推移

K（トピック数）	Exclusivity	Semantic Coherence	model
1	9.1	-69	1
2	9.0	-76	1
3	8.9	-70	1
4	8.9	-72	1
5	9.1	-73	1
1	9.2	-70	2
2	9.0	-71	2
3	9.5	-93	2
4	9.2	-75	2

5	8.2	-67	2
6	9.5	-89	2
1	9.3	-73	3
2	9.1	-67	3
3	9.1	-73	3
4	9.1	-98	3
5	9.4	-87	3
6	9.0	-58	3
7	9.5	-75	3

2.4 Semantic Coherence

数式的には以下のようになる (Mimno, D., et. al. 2011).

$$C(K; V^{(k)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(k)}, v_l^{(k)}) + 1)}{D(v_l^{(k)})}$$

$D(v)$ は v という単語が出現する文書の頻度, $D(v_m^{(k)}, v_l^{(k)}) + 1)$ は単語 v と v' が共起する文書の頻度を示す. なお, $V^k = (v_1^k, ..., v_M^k)$ はトピック K において, M という単語が最も出現する確率である単語のリストである。数式中の「1」は 0 の場合の対数値が計算できるように加えてある.

簡単に言えば, 共起語の条件つき確率の総和を求めて, それを意味論的なつながりの指標として捉えていることが分かる.

2.5 Exclusivity

数式としては以下のようになる (Bischof & Airoldi 2012).

$$\varphi_{f,k} = \frac{\beta_{f,k}}{(\sum_{j \in S} \beta_{f,j})}$$

$\beta_{f,k}$ はトピック K における単語の頻度を指し, $\sum_{j \in S} \beta_{f,j}$ はトピック集合 S における j トピックにおける単語の頻度の総和を意味している.

2.6 各トピックの頻出単語

```
## Topic 1 Top Words:
##       Highest Prob: 恋, 女, 花, 男, 泣い, 娘, ひとり
##       FREX: 娘, 女, 酒, 子, 男, 東京, 唄
##       Lift: 三, 酒, 娘, あんた, 東京, 難, 有
##       Score: 難, 女, 娘, 男, 有, 酒, 東京
## Topic 2 Top Words:
##       Highest Prob: あなた, 私, 好き, 愛, 恋, 愛し, わたし
##       FREX: あなた, 私, 好き, わたし, あたし, ほしい, 彼
##       Lift: あなた, 神様, 私, あたし, 彼, お願い, わたし
##       Score: あなた, 私, 神様, わたし, あたし, 好き, 愛し
## Topic 3 Top Words:
##       Highest Prob: 今, 明日, 自分, 一, 未来, ずっと, 時
##       FREX: 自分, 未来, 世界, 全て, 僕ら, 先, 強く
##       Lift: 走れ, 歩, 進む, 自分, 未来, 全て, 現実
##       Score: 走れ, 僕ら, 自分, 未来, ずっと, 全て, 場所
## Topic 4 Top Words:
```

```
##       Highest Prob: 愛，俺，心，胸，恋，抱い，夜
##       FREX: 俺，お前，おまえ，あいつ，気分，抱い，今夜
##       Lift: お前，そいつ，日本，気分，あいつ，踊り，俺
##       Score: そいつ，俺，お前，おまえ，貴方，あいつ，日本
## Topic 5 Top Words:
##       Highest Prob: 夢，風，空，夏，星，胸，消え
##       FREX: 空，夏，風，海，雲，春，星
##       Lift: 足，雲，桜，虹，青空，彼方，空
##       Score: 足，空，夏，風，春，太陽，雲
## Topic 6 Top Words:
##       Highest Prob: 君，僕，好き，愛，今，愛し，心
##       FREX: 君，僕，会い，気持ち，守り，触れ，側
##       Lift: 超，君，僕，守り，側，触れ，会い
##       Score: 君，僕，超，ずっと，好き，気持ち，愛し
## Topic 7 Top Words:
##       Highest Prob: 人，二，日，涙，忘れ，時，夜
##       FREX: 二，人，雪，想い出，逢い，頃，忘れ
##       Lift: せつな，雪，二，逢い，降る，人，想い出
##       Score: せつな，人，二，雪，想い出，逢い，さよなら
```

なお, それぞれの指標の算出式は以下の通りである.

$$Lift = \frac{\beta_{k,v}}{w_v/\Sigma_v w_v}$$

v は単語, k はトピックを示す. トピックにおける単語分布を, 単語出現率でわったもの.

$$FREX_{k,v} = (\frac{\omega}{FCDF(\beta_{k,v}/\Sigma_{j=1}^{K}\beta_{j,v})} + \frac{1-\omega}{ECDF(\beta_{k,v})})^{-1}$$

トピックにおける頻繁かつ排他的な単語を示す指標. 単語のランクの調和平均.

$$SCORE = \beta_{v,k}(log\beta_{w,k} - 1/K\Sigma_k{}'log\beta_{v,k'})$$

スコア関数は, あるトピックにおける単語頻度の対数をとったものを, 他のトピックにおけるあるトピックと同じ単語頻度の対数で除したものを示す.

2.7 J-POP のトレンド

トピック割合を従属変数とした回帰分析の結果を以下に示す.

2.7.1 経過年をダミー変数として投入した場合

Table 4: 各トピック割合を従属変数とした重回帰分析（係数は非標準化回帰係数

topic	term	coef.	S.E.	p value
1	定数項	−0.10	0.03	0.001
1	2005-19	0.06	0.02	0.004
1	1990-2014	−0.00	0.01	0.781
1	1960-74	0.05	0.02	0.001
1	歌手男性ダミー	0.01	0.01	0.098
1	歌手男女ダミー	−0.02	0.02	0.318

1	総単語数（歌詞一つあたり）	0.00	0.00	0.421
1	歌詞古さ	0.01	0.00	0.000
1	作詞者男性ダミー	0.02	0.01	0.004
1	作詞者男女ダミー	0.02	0.02	0.420
2	定数項	0.27	0.03	0.000
2	2005-19	−0.07	0.02	0.003
2	1990-2014	−0.06	0.02	0.000
2	1960-74	−0.02	0.02	0.226
2	歌手男性ダミー	−0.17	0.01	0.000
2	歌手男女ダミー	−0.07	0.02	0.004
2	総単語数（歌詞一つあたり）	0.00	0.00	0.439
2	歌詞古さ	0.00	0.00	0.456
2	作詞者男性ダミー	−0.03	0.01	0.005
2	作詞者男女ダミー	−0.01	0.03	0.704
3	定数項	0.22	0.03	0.000
3	2005-19	0.10	0.02	0.000
3	1990-2014	0.09	0.01	0.000
3	1960-74	0.02	0.01	0.201
3	歌手男性ダミー	−0.00	0.01	0.565
3	歌手男女ダミー	−0.02	0.02	0.189
3	総単語数（歌詞一つあたり）	0.00	0.00	0.000
3	歌詞古さ	−0.00	0.00	0.000
3	作詞者男性ダミー	−0.01	0.01	0.236
3	作詞者男女ダミー	−0.02	0.02	0.329
4	定数項	0.20	0.03	0.000
4	2005-19	−0.12	0.02	0.000
4	1990-2014	−0.07	0.01	0.000
4	1960-74	−0.09	0.01	0.000
4	歌手男性ダミー	0.05	0.01	0.000
4	歌手男女ダミー	0.04	0.02	0.022
4	総単語数（歌詞一つあたり）	−0.00	0.00	0.036
4	歌詞古さ	−0.00	0.00	0.175
4	作詞者男性ダミー	0.01	0.01	0.464
4	作詞者男女ダミー	0.00	0.02	0.972
5	定数項	0.13	0.03	0.000
5	2005-19	0.05	0.02	0.021
5	1990-2014	0.04	0.01	0.004
5	1960-74	−0.02	0.01	0.191
5	歌手男性ダミー	0.02	0.01	0.002
5	歌手男女ダミー	0.03	0.02	0.067
5	総単語数（歌詞一つあたり）	−0.00	0.00	0.000
5	歌詞古さ	0.00	0.00	0.137
5	作詞者男性ダミー	0.01	0.01	0.175
5	作詞者男女ダミー	0.01	0.02	0.818
6	定数項	0.15	0.03	0.000
6	2005-19	0.00	0.03	0.960
6	1990-2014	0.02	0.02	0.158
6	1960-74	0.04	0.02	0.019
6	歌手男性ダミー	0.09	0.01	0.000
6	歌手男女ダミー	0.04	0.02	0.060
6	総単語数（歌詞一つあたり）	0.00	0.00	0.249

6	歌詞古さ	−0.00	0.00	0.000
6	作詞者男性ダミー	−0.00	0.01	0.612
6	作詞者男女ダミー	0.02	0.03	0.516
7	定数項	0.13	0.02	0.000
7	2005-19	−0.03	0.02	0.108
7	1990-2014	−0.01	0.01	0.379
7	1960-74	0.02	0.01	0.107
7	歌手男性ダミー	−0.01	0.01	0.359
7	歌手男女ダミー	0.00	0.02	0.995
7	総単語数（歌詞一つあたり）	−0.00	0.00	0.134
7	歌詞古さ	0.00	0.00	0.032
7	作詞者男性ダミー	0.00	0.01	0.900
7	作詞者男女ダミー	−0.01	0.02	0.682

2.7.2 補足: 多重共線性について

本来, 説明変数間の多重共線性が通常の回帰モデルでは発生するが, 機械学習においては, L1 正則化ペナルティ項を加えることで (回帰モデルのパラメーター推定でいえば, Lasso 回帰にこれがあたる), 次元圧縮を行い, 多重共線性を回避できる. トピックモデルにおける正則化式は以下のようになる (Roberts, M. E., et. al 2016: 18).

$$argminS = \sum_{v} |\beta_{k,v}^{ref} - \beta_{k,v}^{cand}|$$

$argminS$ は, 誤差を最小化する損失関数, $\beta_{k,v}^{ref}$ はターゲットモデル, $\beta_{k,v}^{cand}$ は推定モデルを指す.

3 引用文献

Benoit K., Watanabe K., Wang H., Nulty P., Obeng A., Muller S., & Matsuo A., 2018, "quanteda: An R package for the quantitative analysis of textual data," *Journal of Open Source Software*, 3(30), 774.

Bischof & Airoldi., 2012, "Summarizing topical content with word frequency and exclusivity," *In Proceedings of the International Conference on Machine Learning*.

Blei, D. M., Ng, A. Y., & Jordan, M. I., 2003, "Latent dirichlet allocation," *Journal of machine Learning research*, 3, 993-1022.

Blei, D. M., 2012, "Probabilistic topic models," *Communications of the ACM*, 55(4): 77-84.

DiMaggio P., 2015, "Adapting computational text analysis to social science", *Big Data & Society*, 2(2): 1-5.

Giddens, A., 1991, *Modernity and self-identity: self and society in the late modern age*, Stanford University Press. (秋吉美都・安藤太郎・筒井淳也訳, 2005, 『モダニティと自己アイデンティティ後期近代における自己と社会』ハーベスト社.)

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D., 2015, "The extent and consequences of p-hacking in science," *PLoS Biol*, 13(3): 1-15.

北田暁大, 2004, 『〈意味〉への抗いメディエーションの文化政治学』せりか書房.

久保正敏, 1995, 「ニューミュージックに見る恋愛風景」『情報処理学会研究報告人文科学とコンピュータ（CH）』, 25: 49-57.

増田聡, 2006, 『聴衆をつくる音楽批評の解体文法』青土社.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M. & McCallum, A., 2011, “Optimizing Semantic Coherence in Topic Models,” *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP ’11, 262–272.

見田宗介, 1978, 『近代日本の心情の歴史—流行歌の社会心理史』講談社.

小川博司, 2004, 「ポピュラー音楽へのアプローチ」
井上俊編『新版　現代文化を学ぶ人のために』世界思想社, 161-183.

大出彩・松本文子・金子貴昭, 2013, 「流行歌から見る歌詞の年代別変化」『情報処理学会』, 4: 103 110.

Mohr, John W., and Petko, B., 2013, “Introduction—Topic models: What they are and why they matter”, *Poetics*, 41(6): 545-569.

Motohiro Ishida, 2020, *RMeCab: interface to MeCab*, R package version 1.05.

R Core Team, 2020, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing: Vienna, Austria.

Roberts, E. M., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. & Rand, D. G., 2014, “Structual Topic Models for Open-Ended Survey Responses,” *American Journal of Political Science*, 58(4), 1064-1082.

Roberts, E. M., Stewart, B. M. & Airoldi, E. M., 2016a, “A Model of Text for Experimentation in the Social Sciences,” *Journal of the American Statistical Association*, 111(515), 988-1003.

Roberts, M., Stewart, B., & Tingley, D., 2016b, “Navigating the Local Modes of Big Data: The Case of Topic Models”, R. Alvarez eds., *Computational Social Science: Discovery and Prediction*, Cambridge: Cambridge University Press, 51-97.

Roberts, M. E., Stewart, B. M. & Tingley, S. D., 2019, “stm: An R Package for Structural Topic Models,” *Journal of Statistical Software*, 91(2), 1-40.

佐藤一誠, 2018,『トピックモデルによる統計的潜在意味解析』コロナ社.

瀧川裕貴, 2019, 「戦後日本社会学のトピックダイナミクス」『理論と方法』, 34(2): 238-261.

烏賀陽弘道, 2005, 『J ポップとは何か』岩波書店.