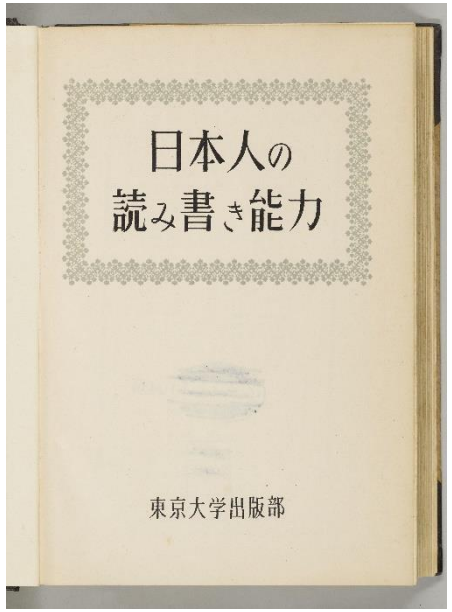
分布意味論の10分間解説例題を心理学者が作ってみた（失敗作？）

横山詔一（国立国語研究所言語変化研究領域教授）

NINJALサロン 2021年6月1日（火）15時30分～16時30分

司会：浅原正幸（コーパス開発センター教授）

浅原先生と横山をつなぐ
「点と線」
まず，点は･･･



文や文章を読んでいる時の眼球運動
測定装置「オフサルモグラフ」旧式

なぜ意味論？

- 国立国語研究所 言語学レクチャーシリーズ 動画教材 第7回「語の意味論」松本曜先生
- https://www.youtube.com/watch?v=v2f2lU7UFHY
- 動画教材から受けた刺激により古い記憶のふたがパカッとあいて，修士論文のテーマが「カテゴリー形成における情報統合」であったことを想い出した

きょうの進め方

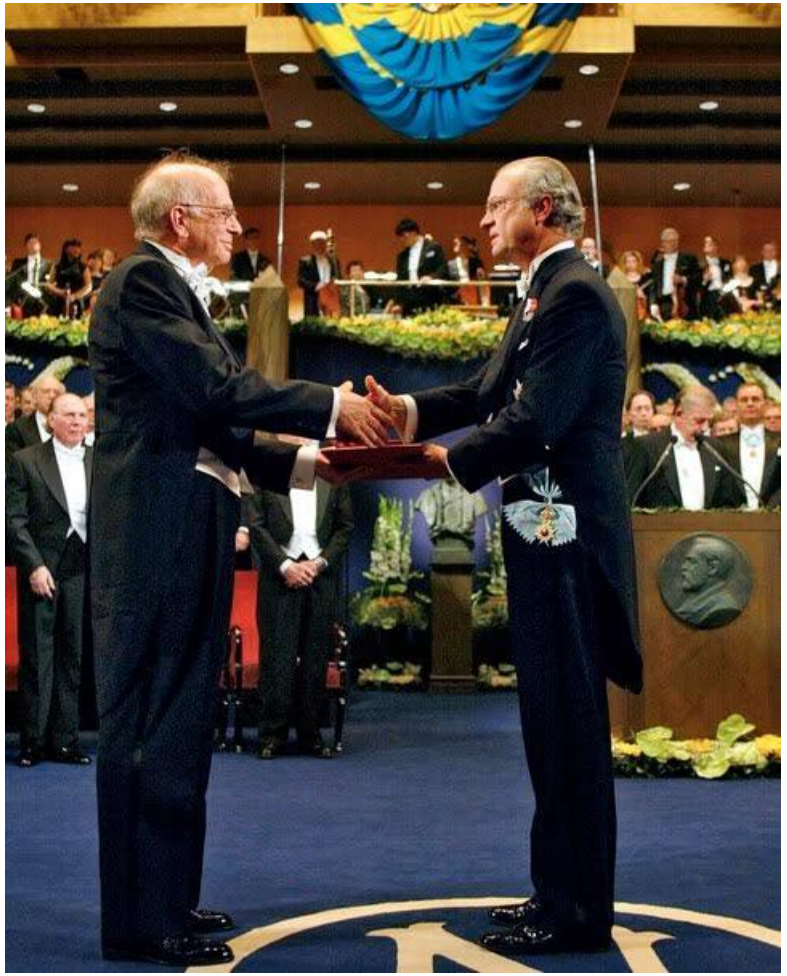
- 今回の発表は録画を流し，質疑応答からライブに切り替えるつもりでした
- 使用する録画は，青山学院大学の学部オンデマンド授業で使用した動画教材の一部を想定
- しかし，テストしてみたところ，音質が悪くなりそうなので，すべてライブにします

きょうの進め方

- 分布意味論の解説に入る前に，主成分分析PCAの解説をライブでおこないます
- すでに，解説YouTube動画（約16分間，主成分分析の説明は9分20秒あたりから最後まで約7分間）を公開中です

  => https://youtu.be/J_sHc4SyZ_o

類似性判断（similarity judgement）に関するモデルについて



- Kahneman & Tversky (1979) による計量経済学の研究が2002年にノーベル経済学賞を受賞
- Kahneman & Tverskyは両者とも認知心理学者
- 惜しいことにTverskyは受賞前に死去
- Tverskyは1992年に日本心理学会第56回大会で講演，通訳は繁桝算男先生（青学 繁桝江里先生のお父上）
- Tverskyは類似性判断に関するモデル（feature contrast model）でも有名 → 資料教材pdf

例題１　主成分分析（因子分析）は役に立つのか

以下，フィクションです

- あなたは米国FBI心理分析官
- きょうの分析対象は国家機密データ，コード名は「性格検査結果」
- 「性格検査結果」をスパイが国外に持ち出そうとしているとの通報あり
- そこで，スパイ容疑者宅に踏み込み「例題１」のデータを押収
- 「例題１」は「性格検査結果」とまったく似ていない（暗号化されているのでは？）
- そのため，スパイ容疑者は国家機密なぞ知らぬ存ぜぬ，わたしは善良な米国市民だと言い張っている

Q1）「例題１」と「性格検査結果」の類似度（どのくらい似ているか）を示してください

コード名「性格検査結果」（実は公開データ，出典は後で明示します）

番号	外向性	社交性	積極性	知性	信頼性	素直さ
1	3	4	4	5	4	4
2	6	6	7	8	7	7
3	6	5	7	5	5	6
4	6	7	5	4	6	5
5	5	7	6	5	5	5
6	4	5	5	5	6	6
7	6	6	7	6	4	4
8	5	5	4	5	5	6
9	6	6	6	7	7	6
10	6	5	6	6	5	5
11	5	4	4	5	5	5
12	5	5	6	5	4	5
13	6	6	5	5	6	5
14	5	5	4	4	5	3
15	5	6	4	5	6	6
16	6	6	6	4	4	5
17	4	4	3	6	5	6
18	6	6	7	4	5	5
19	5	3	4	3	5	4
20	4	6	6	3	5	4

スパイ容疑者が持っていた「例題１」
作家20人の作品コーパスから得た使用率の6変数データだと主張

1	著者ID	格助詞は[,]	格助詞に	動詞思う	助動詞だ	ら抜き	名詞止め
2	1	0.150	0.250	0.100	0.100	0.040	0.040
3	2	0.300	0.400	0.175	0.150	0.070	0.070
4	3	0.300	0.250	0.125	0.125	0.070	0.060
5	4	0.300	0.200	0.150	0.175	0.050	0.050
6	5	0.250	0.250	0.125	0.175	0.060	0.050
7	6	0.200	0.250	0.150	0.125	0.050	0.060
8	7	0.300	0.300	0.100	0.150	0.070	0.040
9	8	0.250	0.250	0.125	0.125	0.040	0.060
10	9	0.300	0.350	0.175	0.150	0.060	0.060
11	10	0.300	0.300	0.125	0.125	0.060	0.050
12	11	0.250	0.250	0.125	0.100	0.040	0.050
13	12	0.250	0.250	0.100	0.125	0.060	0.050
14	13	0.300	0.250	0.150	0.150	0.050	0.050
15	14	0.250	0.200	0.125	0.125	0.040	0.030
16	15	0.250	0.250	0.150	0.150	0.040	0.060
17	16	0.300	0.200	0.100	0.150	0.060	0.050
18	17	0.200	0.300	0.125	0.100	0.030	0.060
19	18	0.250	0.150	0.125	0.075	0.040	0.040
20	19	0.200	0.150	0.125	0.150	0.060	0.040
21	Who?	0.300	0.200	0.125	0.150	0.070	0.050

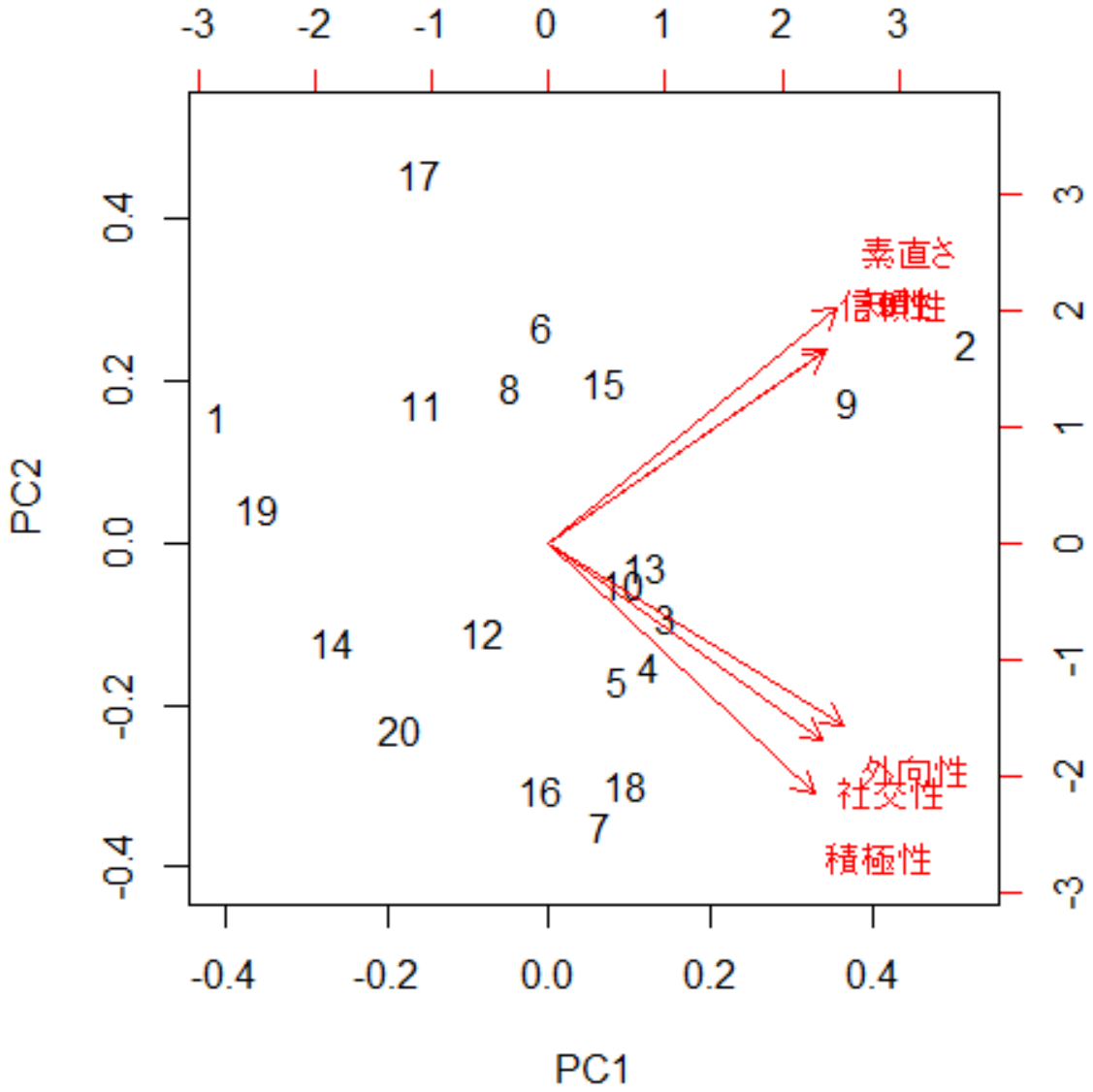
「性格検査結果」の一部と「例題１」の一部を並べてみると

1. まったく似ていない感じもするが，行や列を入れ替えているのかも
2. 「性格検査結果」と「例題１」の眼に見えない構造を視覚化したい

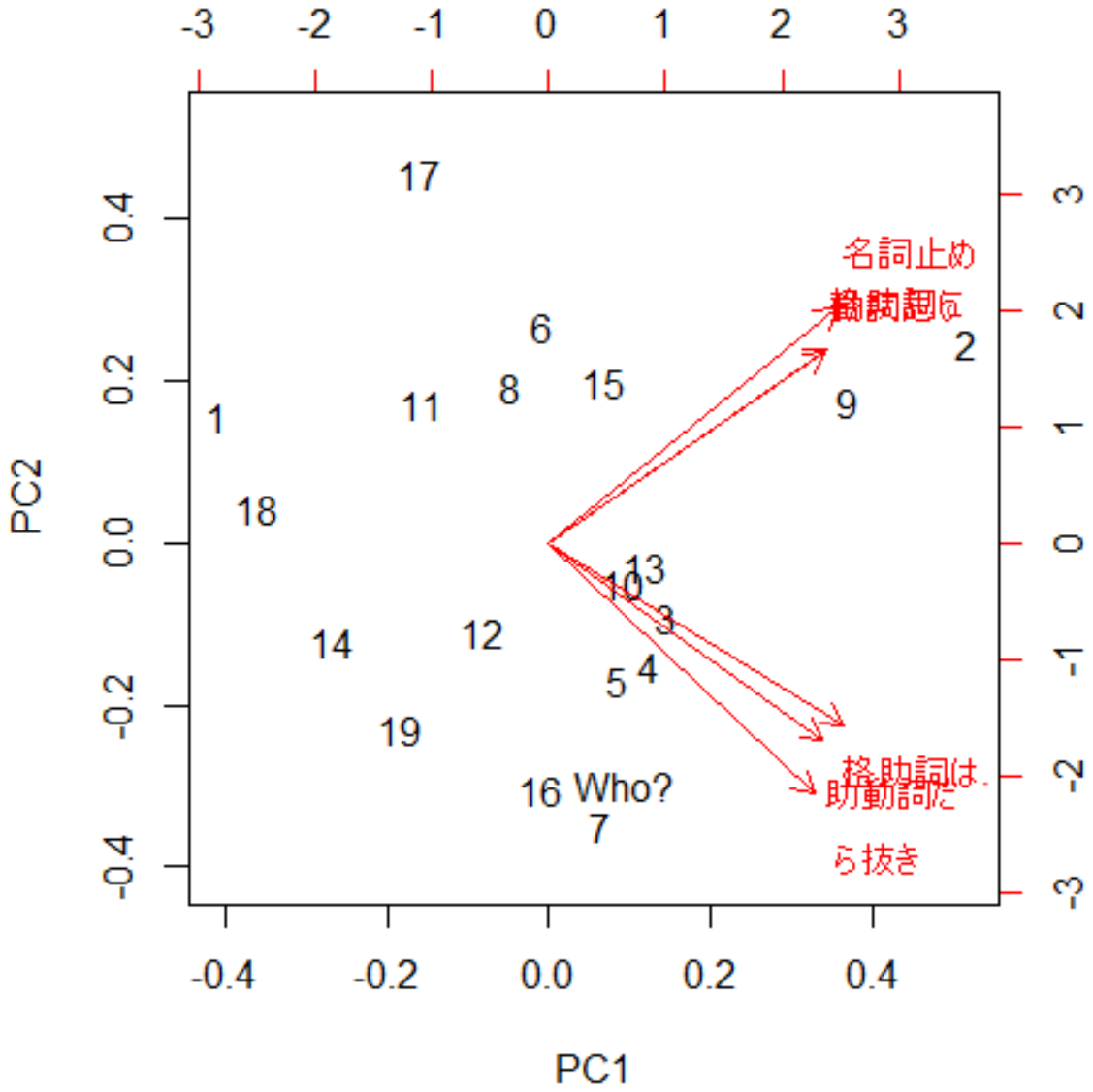
番号	外向性	社交性	積極性	知性	信頼性	素直さ
1	3	4	4	5	4	4
2	6	6	7	8	7	7
3	6	5	7	5	5	6
4	6	7	5	4	6	5
5	5	7	6	5	5	5
6	4	5	5	5	6	6
7	6	6	7	6	4	4

1	著者ID	格助詞は[,]	格助詞に	動詞思う	助動詞だ	ら抜き	名詞止め
2	1	0.150	0.250	0.100	0.100	0.040	0.040
3	2	0.300	0.400	0.175	0.150	0.070	0.070
4	3	0.300	0.250	0.125	0.125	0.070	0.060
5	4	0.300	0.200	0.150	0.175	0.050	0.050
6	5	0.250	0.250	0.125	0.175	0.060	0.050
7	6	0.200	0.250	0.150	0.125	0.050	0.060

「性格検査結果」を主成分分析し，結果を2次元空間に布置→次のスライドと比較



「例題１」を主成分分析→前のスライドと比較（18，19，20，Who?，ら抜き等に注目）



例題１　主成分分析（因子分析）の有用性について

- 「例題１」は「性格検査結果」とまったく似ていない
- 主成分分析（Principal Component Analysis：PCA）で解析し，結果を2次元空間に布置すると，両者はきれいに一致することが判明

Q1）「例題１」と「性格検査結果」の類似度（どのくらい似ているか）を示してください

A1）　両者の潜在構造（眼には見えない構造）は同じ，類似度は100%

✓ 人工知能が語の意味を学習し，理解するときの基礎にPCAがある

✓ 「性格検査」データは小塩真司（おしお・あつし）先生がネットで公開しているものです
http://www.f.waseda.jp/oshio.at/edu/data_b/top.html　心理データ解析 第8回(2)分析例1（主因子解・バリマックス回転）

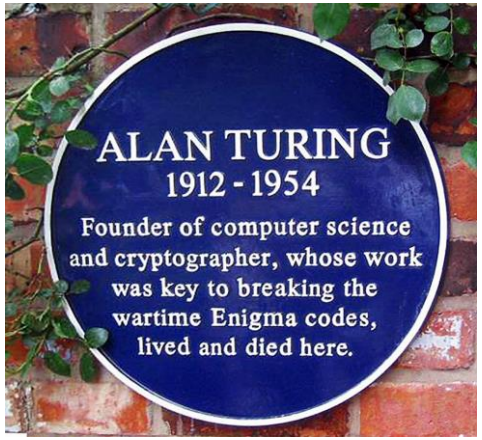
きょうの進め方

- では，分布意味論の解説をライブでおこないます
- すでに，解説YouTube動画（約30分間，分布意味論の解説は9分10秒あたりから最後までの約20分間）を公開中

  ⇒ https://youtu.be/lDL9Ru8VGk8

- 10分間解説に失敗した!?

例題2　意味空間（Semantic Space ）を体験



以下，フィクションです

- 時は1942年，あなたは英国海軍情報部の言語心理分析官（コード名：くまのPoohさん）
- 分析対象はドイツ軍潜水艦U-Boatが発信する大量の暗号文
- 沈没したU-Boatから暗号解読のコードブックを入手することに成功したが，大部分の文字が消えていた
- 現時点で判明しているのは，（1）「V1，V2，・・・」は他動詞，（2）「N1，N2,・・・」は目的語，（3）「N1 ＝ Katze = Cat」

例題2　意味空間（Semantic Space ）を体験

以下，フィクションです

- 現時点で判明しているのは，（1）「V1，V2，・・・」は他動詞，（2）「N1，N2,・・・」は目的語，（3）「N1 ＝ Katze = Cat」

Q）今回の任務は「N19」が何を意味する名詞か推理すること

1. 他動詞と目的語の共起頻度をカウントし，その共起頻度表を「例題２」と命名
2. 「例題２」を主成分分析にかける => 意味空間を描く
3. 意味空間における単語の布置を注意深く観察する

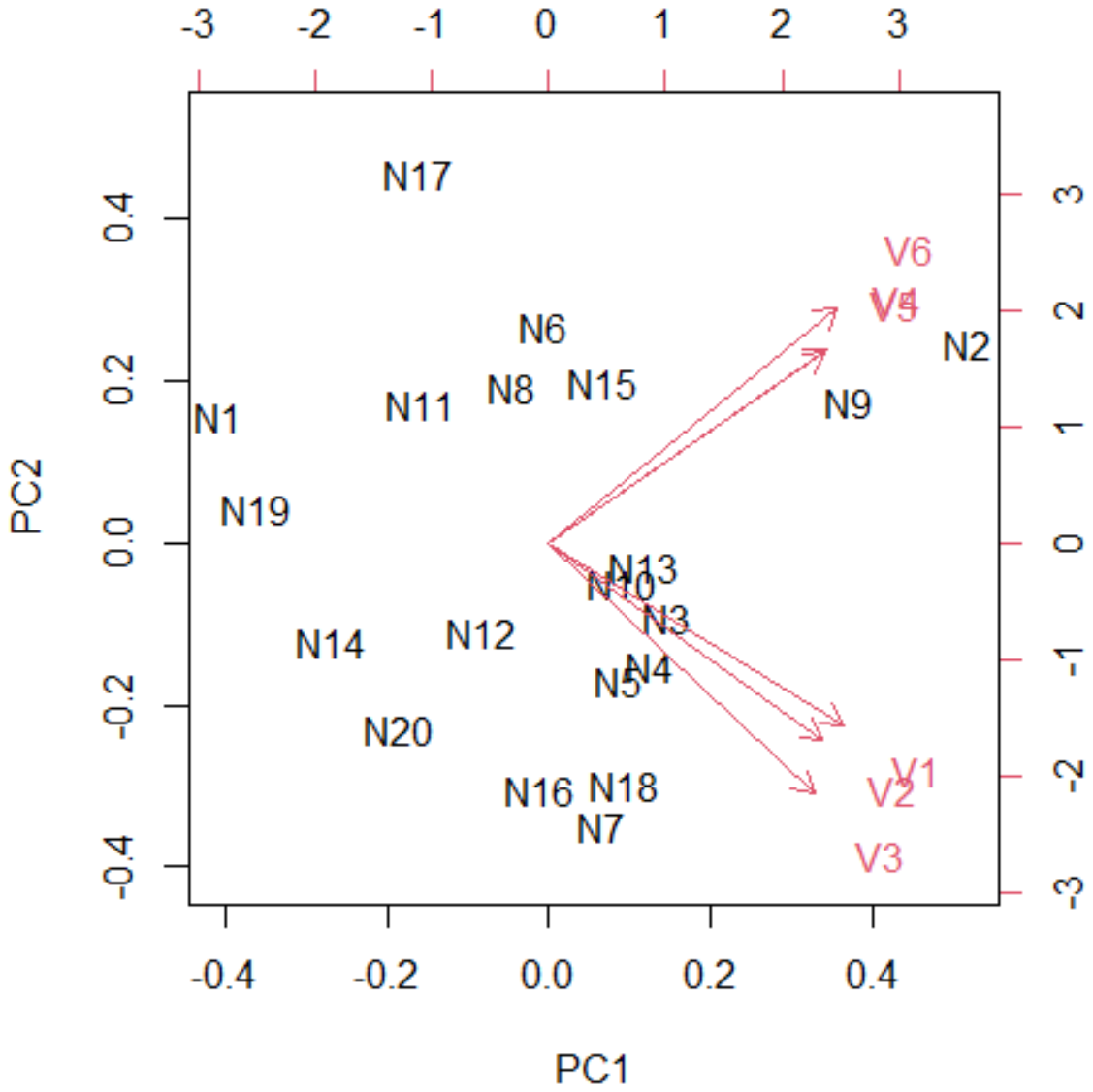
例題２

横山詔一2021作成

	V1	V2	V3	V4	V5	V6
N1	0	10	10	20	0	10
N2	30	30	40	50	30	40
N3	30	20	40	20	10	30
N4	30	40	20	10	20	20
N5	20	40	30	20	10	20
N6	10	20	20	20	20	30
N7	30	30	40	30	0	10
N8	20	20	10	20	10	30
N9	30	30	30	40	30	30
N10	30	20	30	30	10	20
N11	20	10	10	20	10	20
N12	20	20	30	20	0	20
N13	30	30	20	20	20	20
N14	20	20	10	10	10	0
N15	20	30	10	20	20	30
N16	30	30	30	10	0	20
N17	10	10	0	30	10	30
N18	30	30	40	10	10	20
N19	20	0	10	0	10	10
N20	10	30	30	0	10	10

N19とN1は似ていないように見える

「例題２」を主成分分析し，結果を2次元空間に布置



例題2　分布意味論とは

- 人工知能による単語の意味処理は Distributional Hypothesis（分布仮説）に立脚している。これは「同じ文脈に出現する単語は，似たような意味を持つ傾向がある」という仮説である（Harris, 1954）
- 上記の仮説に基づく意味論を「分布意味論（Distributional Semantics）」という
- 今回は他動詞と目的語（名詞）の共起頻度をカウントしたデータを主成分分析（Principal Component Analysis：PCA）で解析

例題2　分布意味論に基づく意味空間

- 多くの単語をなるべく少ない次元の空間に布置 => これを「分布意味論に基づく意味空間」という

✓ PCAにより空間の次元数を低減でき，データ操作が楽になる

✓ PCAにより単語の潜在的意味（人間には理解不能！）を数値で表現できるようになる

Q）今回の任務は「N19」が何を意味する名詞か推理すること

1. 「N19」に意味が近いのは「N1 ＝ Katze = Cat」
2. よって「N19」は「Hund ＝ Dog」かもネ

- 人工知能が語の意味を学習し，理解するときの基礎にPCAがある

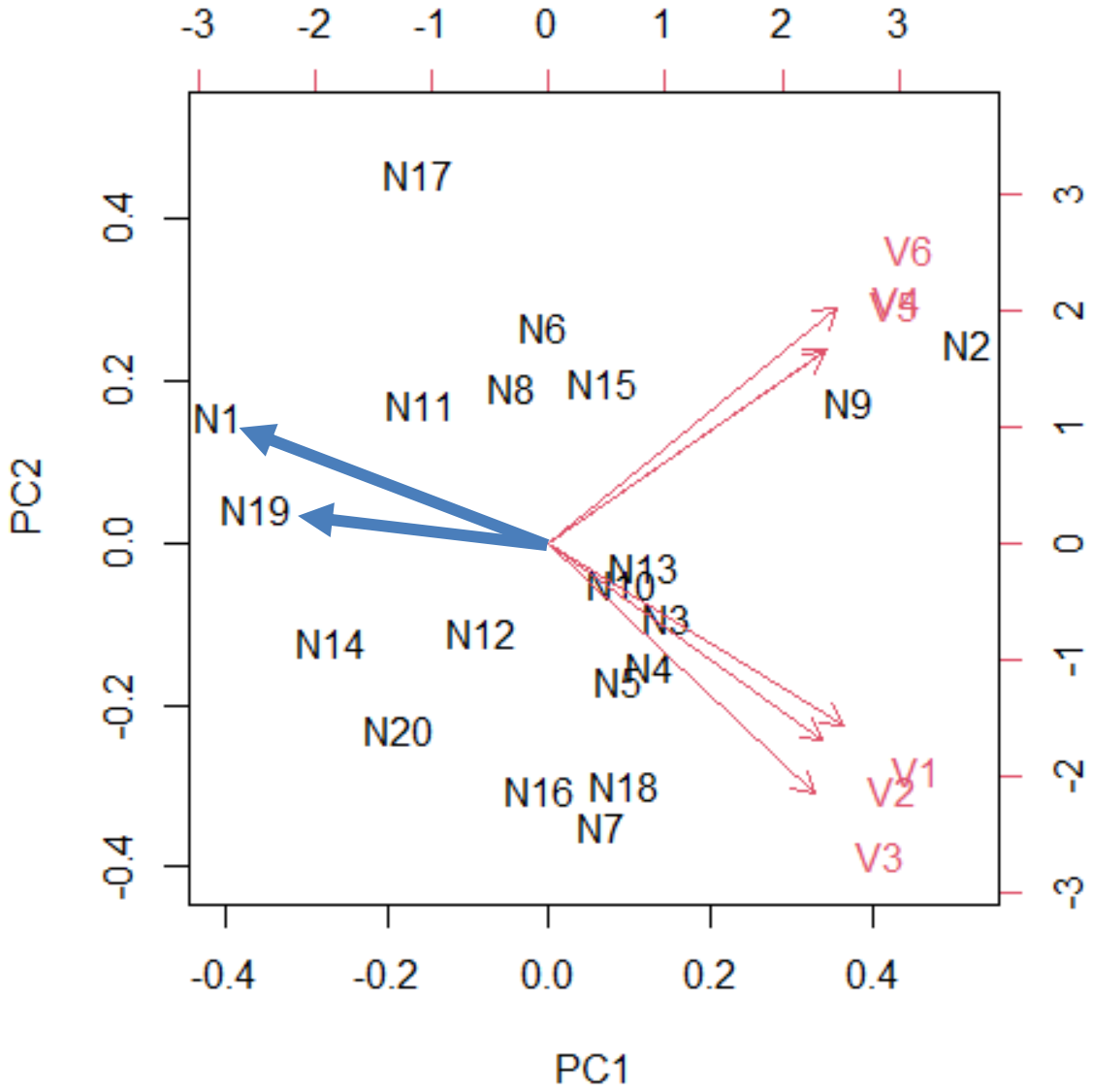
例題2

分布意味論に基づく意味空間と単語のベクトル表現

- 人工知能が語の意味を学習し，理解するときの基礎にPCAがある
- 単語の意味をどう表現・記述するか

1. 国語辞典には語釈，用例，品詞などが記述されている
2. 人工知能は意味空間の中心（原点）から単語の座標に矢印を引く => 単語をベクトルで表現・記述する方法（数値で表現できる），次のスライド参照

「例題２」を主成分分析し，結果を2次元空間に布置→単語のベクトル表現の例



N1の意味はN1ベクトルで，
N19の意味はN19ベクトル
で表現・記述されている

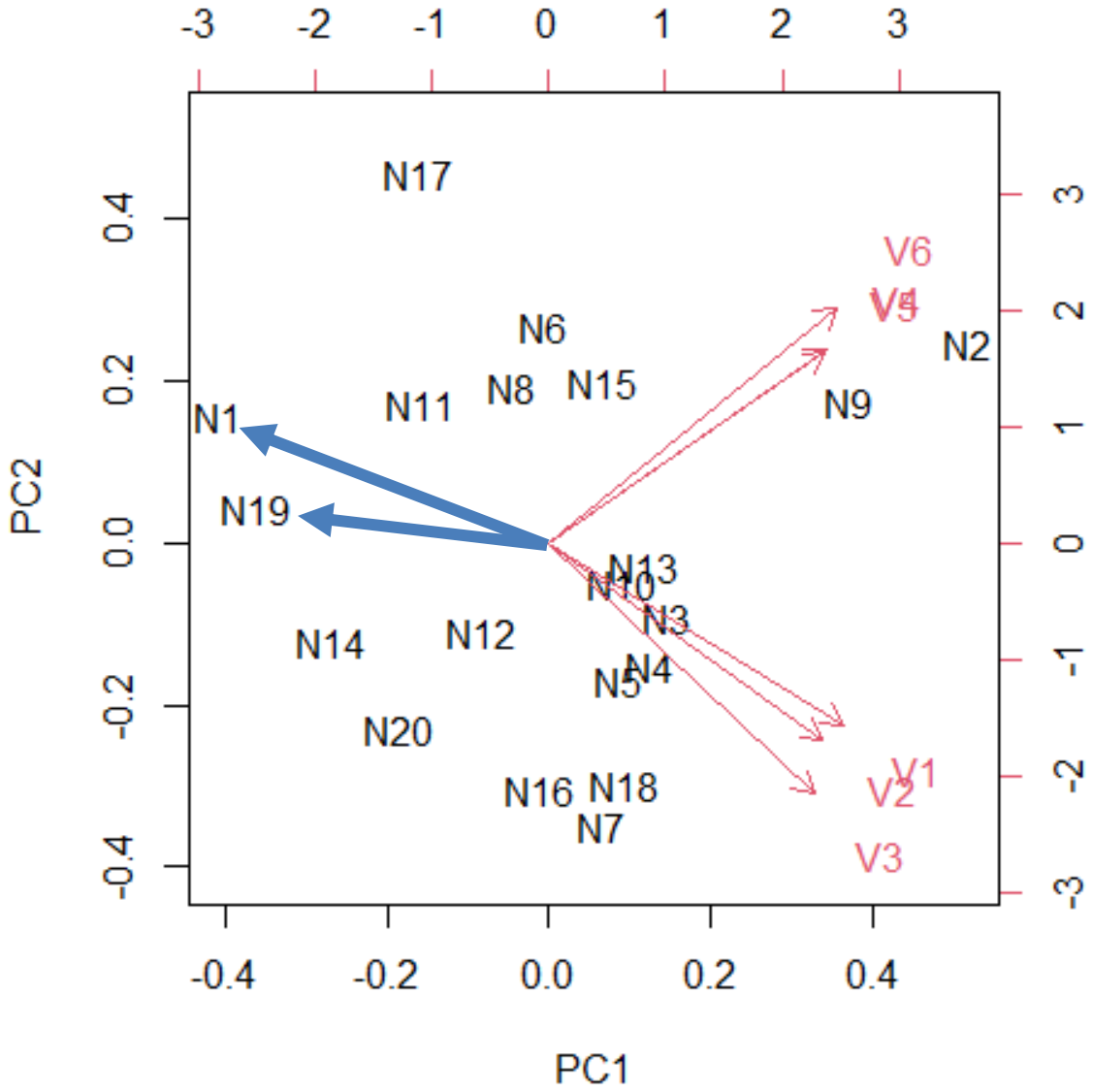
例題2

分布意味論に基づく意味空間と単語埋め込み（Word Embedding）

1. 単語のベクトル表現を「分散表現」とか「単語埋め込み」ということがある
   - 単語分散表現のイメージ => 単語の意味はその単語自体にあるのではなく，周囲の単語との共時性，布置，文脈のなかに散らばって存在する
   - 単語埋め込みのイメージ => 単語のベクトル表現には周囲の単語との共時性，布置，文脈の情報のエッセンスが圧縮されて埋め込まれている
2. ベクトル表現だとベクトルの足し算や引き算が簡単にできる
3. その結果，ご利益が！　驚くような性能を発揮

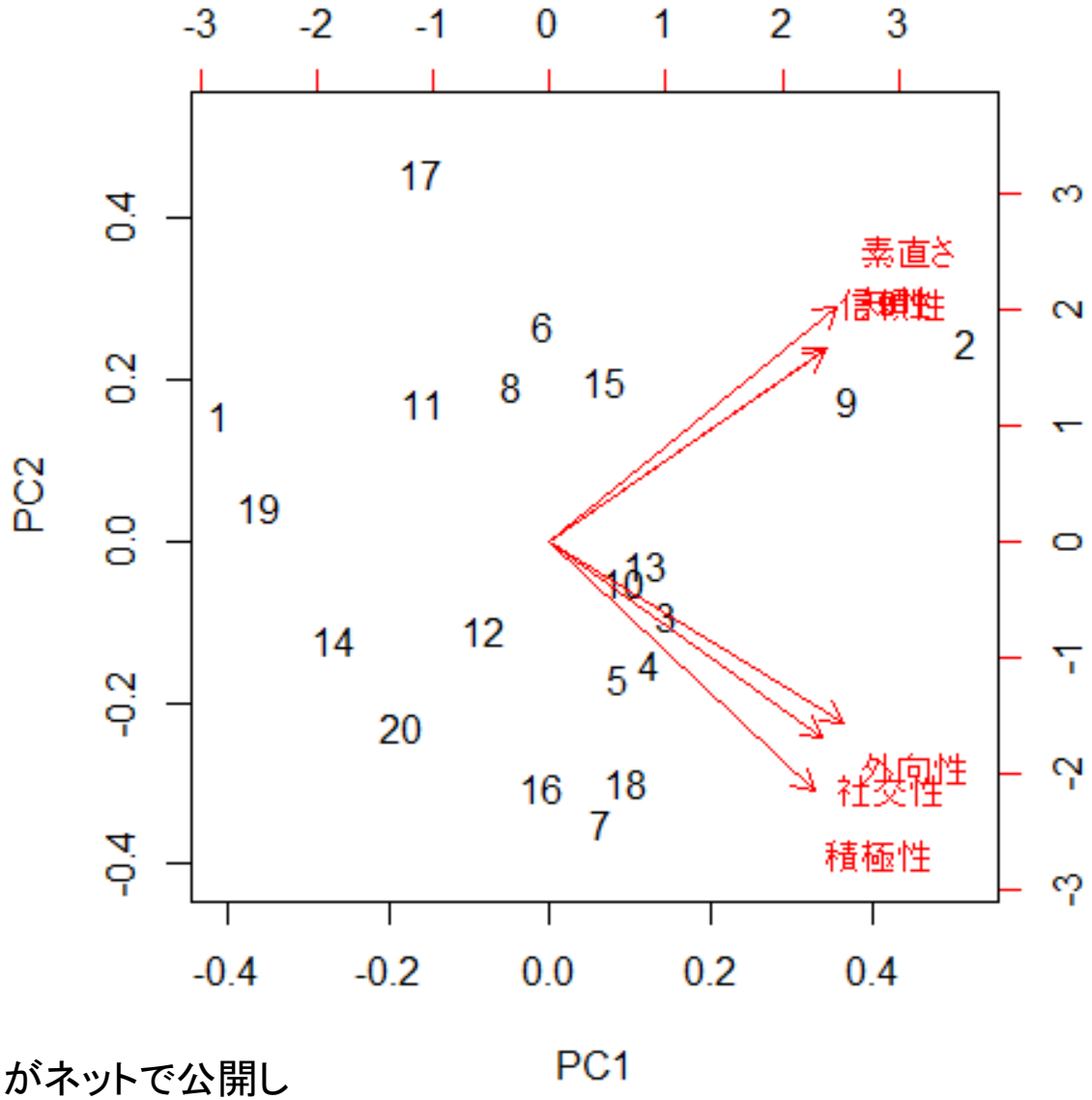
   ✓ たとえば「王様ー男性＋女性＝女王」などが瞬時にできる

「例題２」を主成分分析し，結果を2次元空間に布置→次のスライドと比較



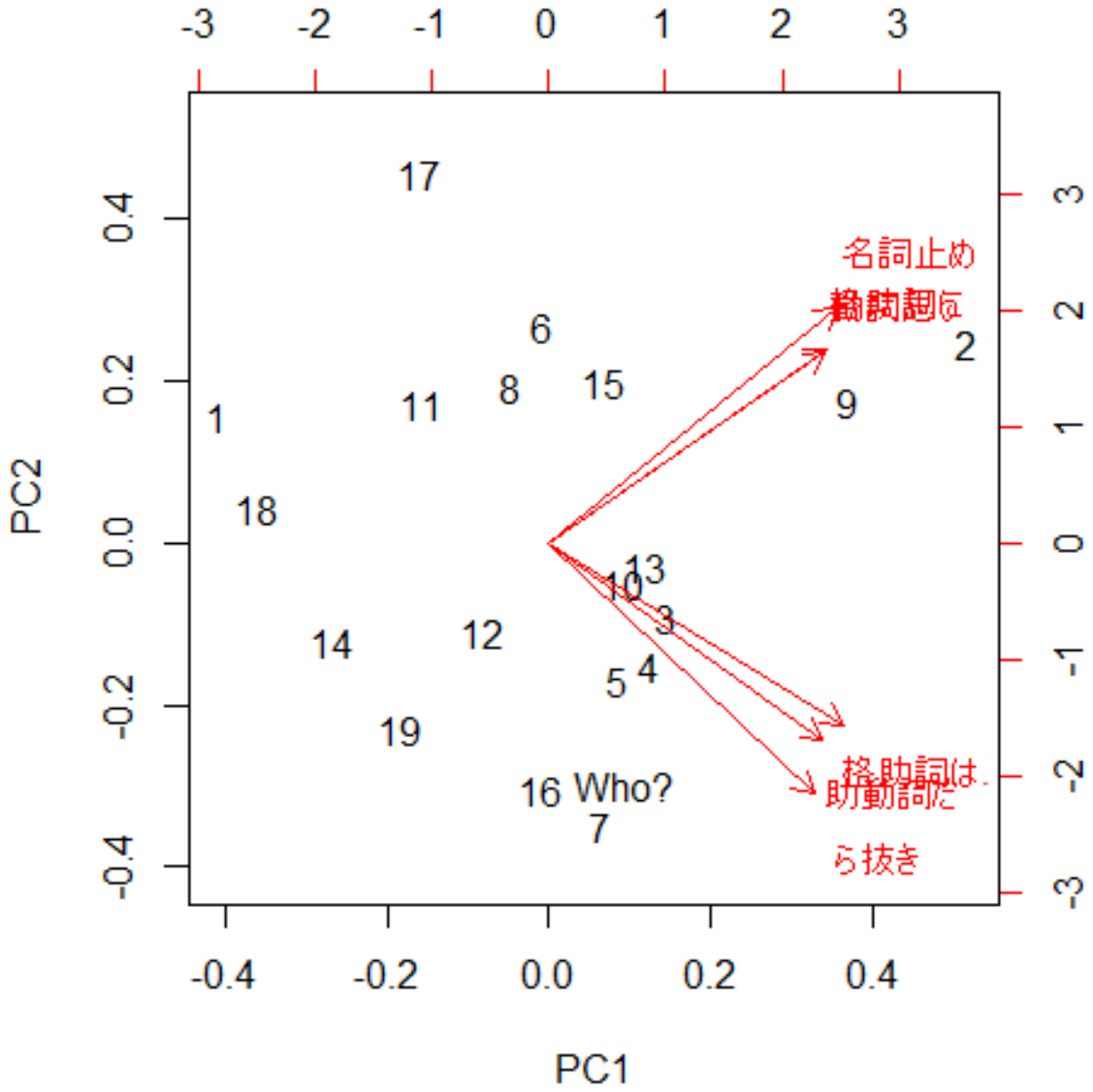
N1の意味はN1ベクトルで，
N19の意味はN19ベクトル
で表現・記述されている

「性格検査結果」を主成分分析し，結果を2次元空間に布置→次のスライドと比較



「性格検査」データは小塩真司先生がネットで公開しているものです
http://www.f.waseda.jp/oshio.at/edu/data_b/top.html

「例題１」を主成分分析→前のスライドと比較（18，19，20，Who?，ら抜き等に注目）



きょうの進め方

- これで発表は終わりです
- それでは，これから質疑応答をお願いいたします
- その前に，参考図書と謝辞について

参考図書（古典）

- 主成分分析や因子分析をベクトルで解説

  柳井晴夫・岩坪秀一（1976）『複雑さに挑む科学―多変量解析入門』（ブルーバックス），講談社

- 分散分析（ANOVA）の原理をもっともシンプルに解説

  佐藤　信（1983）『推計学のすすめ：決定と計画の科学』（ブルーバックス），講談社

浅原先生と横山をつなぐ
「点と線」
線は・・・

修了証

横山詔一殿

あなたはコーパス開発センター
主催の python 勉強会 2018 を修
了しましたのでこれを証します。

2018 年 1 月 26 日

国立国語研究所
コーパス開発センター
准教授
浅原 正幸

謝辞

- 浅原勉強会（道場？）でご指導くださった岡 照晃先生と大村 舞先生，そして勉強会参加者のみなさまに御礼を申し上げます
- JPS科研費21K00551（研究代表者：横山詔一）の成果の一部である