# 自由連想ネットワーク上の幾何学的な性質
# Geometric Properties on a Free Association Network

日高　昇平[†]
Shohei Hidaka

[†] Indiana University
shhidaka@indiana.edu

**Abstract**

A normative free association data is analyzed in order to describe the adjacency of semantic categories using a probabilistic model on a feature space. The structure estimated by the model suggests that a group of words form continuous and smooth clusters which are predicted by the past research. It supports the idea that semantic categories are optimally structured in terms of their discrimination and generalization.

**Keywords — Semantic Network, Free Association**

## 1. はじめに

カテゴリ化の最も基礎的な機能は、特徴を選択し情報を圧縮する事である．自然カテゴリの特定の種類のカテゴリ群は，特定の共通的な特徴を基にカテゴリ化される事が多い(e.g.，「動物」は４本足を持ち、毛皮をもつなど)．このようなカテゴリの構造を調べる事は，自然な概念形成を理解するための基礎となると考えられる．幼児は新奇な物体に新奇語が付与された場合に、既存の知識を活用して新奇語の参照する事例を推定する事が知られている[2]．このような幼児の新奇語汎用に関する先行研究から、著者は、自然カテゴリでは類似したカテゴリが類似した特徴空間における事例の分布パタンを持ち、類似していないカテゴリは異なる事例の分布パタンを持つのではないか、と仮説を立てた[1]．この自然カテゴリの大域構造を「滑らかなカテゴリ」と呼ぶ．滑らかなカテゴリの一つの利点は、新奇カテゴリに関する予測性の高さである．滑らかなカテゴリにおいて、類似したカテゴリが類似した事例の分布を持つので、新奇カテゴリの新奇な一事例からでも、既知の類似カテゴリから、新奇カテゴリの事例分布を予測することが可能となる[1]．

本研究では、自然カテゴリにおいて滑らかなカテゴリ構造が実際に存在するか、具体的にデータを分析することで検討した．本研究で検討したデータは成人の自由連想課題[4]である．この課題そのものには協力者の答えを制約するものは無いため、制約の小さい自由連想課題から、滑らかなカテゴリの構造が見つかれば、提案仮説の強い証拠となると考えられる．データの分析には、多次元正規分布を用いて連想パタンを近似する方法を用いた．このモデルでは、個々の単語の分布を多次元正規分布とみなし、二つの単語分布の確率的な重なりを連想確率とし、連想関係を多次元特徴空間上の分布として解釈する．

## 2. データ

本研究ではUniversity of Southern Florida Word Association Norm(USFWA)[3]を用いた．USFWA は英語の自由連想課題の結果を収録したもので、5000 以上の手がかり語、10000 以上の連想語を6000 人以上の協力者から得ている．公開されている中で最大級の連想データベースである．本研究では全ての語ではなく、最も連想される頻度の高い100 語(全ての連想回答のうち22%に当たる)を対象とし、手がかり語も同一の100 語を用いた．

## 3. 分析方法

与えられるデータにおいて、手がかりカテゴリ $i\left(i=1,2,\cdots,N\right)$ から連想カテゴリ $j\left(j=1,2,\cdots,N\right)$ が連想される確率を $Q\left(j\mid i\right)$，それに対応するモデルでの連想確率を $P\left(j\mid i\right)$ とする．モデルの連想確率 $P\left(j\mid i\right)$ はカテゴリ $i$の事前確率 $P\left(i\right)$, カテゴリ $i$, $j$ 間の確率的な重複を表す $F_{ij}$ を用いて以下のように表現される：$P(j\mid i)=P(j)\exp\left(F_{ij}\right)\Big/\left\{\sum_{j}^{M}P(j)\exp\left(F_{ij}\right)\right\}$

ただし、$F_{ij}$ はButtercherrya Bound[1]で、以下のよ

うに、多次元空間上でのカテゴリの平均ベクトル $\mu_i$ および共分散行列 $\sigma_i$ を用いて表される.

$$F_{ij} = -\frac{1}{8}(\mu_i - \mu_j)^T \sigma_{ij}^{-1}(\mu_i - \mu_j) - \log(|\sigma_{ij}|) + \frac{1}{2}\log(|\sigma_i||\sigma_j|)$$

ただし、$\sigma_{ij} = (\sigma_i + \sigma_j)/2$ で、$|X|$ は $X$ の行列式である. パラメタ $\{\mu_i, \sigma_i, P(i)\}(i = 1, 2, \cdots, N)$ の推定は以下の対数尤度 $L$ を最大化することで行った.

$$L = \sum_j^M \sum_i^N Q(j \mid i) \log(P(j \mid i))$$

**滑らかさ指標**

もし，カテゴリが滑らか(2つのカテゴリが類似であるほど、類似の分布パタンを持つ)であれば，カテゴリの中心間の距離(i.e.，平均ベクトルノルム) とカテゴリ尤度の類似性(i.e.，共分散行列ノルム) には正の相関があるはずである．滑らかさ指標滑らかさの指標として，2 カテゴリの平均ベクトル間ノルムと2 カテゴリの共分散行列間のノルムの相関を分析した.

## 4. 結果・考察

最も連想される頻度の高い100語の連想パタンを、個々のカテゴリを多次元正規分布として分析した. 10次元特徴空間における平均・共分散によって単語を表現するモデルは、与えられた連想確率パタンの全分散のうち28%を説明可能であった. 推定された10次元のうち、2次元におけるカテゴリの分布パタンを図 1 に示す. 図1では、中心にあるカテゴリほど分散が小さく、また周辺にあるカテゴリは中心方向に大きな分散を持つ傾向があった．この傾向は、10次元全ての特徴空間で一貫していた．このパタンは、ある特定の連想関係を持つものが方向付けられた(ある特徴軸に特化した)分散を持ち、より一般的に連想されるものが中心にある事を意味している．次に、カテゴリ滑らかさの指標として、カテゴリ平均間の距離と共分散行列間の距離の相関を分析した．その結果、有意な高い相関が得られた(R=0. 67)．これは、連想ネットワークにおいて、カテゴリの滑らかな構造、すなわち類似のカテゴリが類似の分布パタンを持つ構造を示唆している．カテゴリの滑らかな構造は、理論的には、カテゴリの識別性、一般性を最適化した結果発生する大域的構造であると考えられる[1]．これを踏まえて考察すると、自由連想は、課題としては何ら制約を持たないが、そのパタンは意味知識の幾何学的な最適化を反映しているのではないかと考えられる.

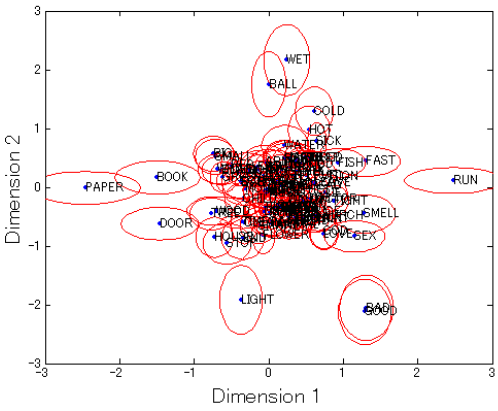


図 1: 推定された特徴空間(次元 1，2)．カテゴリ平均を点(および単語)で、共分散行列(0. 3 標準偏差領域)を楕円で示す.

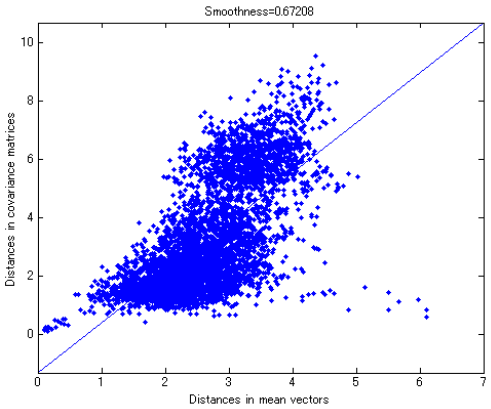


図 2: カテゴリ平均間のユークリッド距離(x 軸)と共分散行列間のユークリッド距離(y 軸).

## 参考文献


[1] Hidaka, S. & Smith, L. B.(2008) How Features Create Knowledge of Kinds. In *Proceedings of The Thirtyth Annual Meeting of Cognitive Science Society*, 1029–1035.

[2] Markman, E. (1989). Categorization and Naming in Children: Problems of Induction. Cambridge, MA: MIT Press.

[3] Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms.