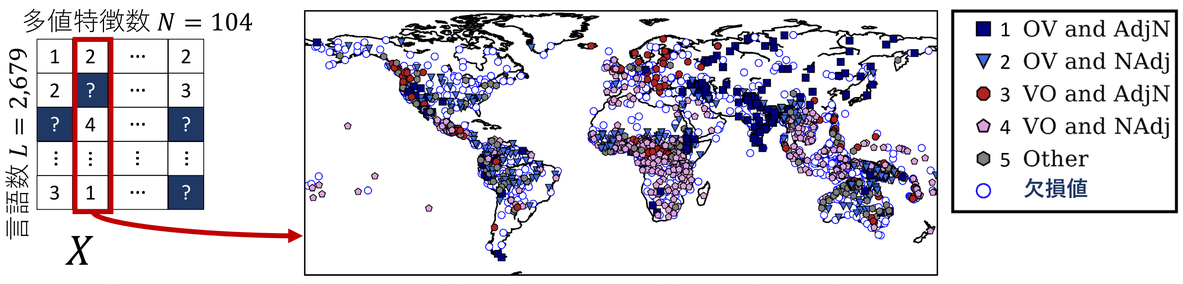
言語の構造的特徴はなぜ、どのように変化するのか

京都大学 KYOTO UNIVERSITY

村脇 有吾

言語の構造的特徴 (言語類型論)



今回扱うデータと特徴の地理的分布の例

- WALS: 世界中の言語の構造的特徴 (語順、声調の有無等) を記述したデータベース
- 特徴の歴史的変化、言語同士の系統関係を明らかにしたい
- データに**欠損値**が多すぎて扱いにくい (前処理後も**>70%**)

統計的手法の必要性

- 構造的特徴の歴史的変化は**不確実性**が高い
  - 語彙的特徴 (e.g. 英: waster, 独: Wasser) なら祖先から引き継いだか否かを人間が論証しやすい
  - 2つの言語が基本語順 (e.g. SVO) を共有していたとして、それが祖先から引き継がれたものかはよくわからない
- 統計的手法 (特に**ベイズ統計**) であれば、個々には不確実の高い手がかりを組み合わせ、より確からしい仮説を探せる
- 言語同士の系統関係は、従来は広い意味での語彙的手がかりを用いて立証されてきたが、構造的特徴の振る舞いが明らかになれば、系統推定への応用の可能性あり
  - 日本語の系統を明らかにしたい

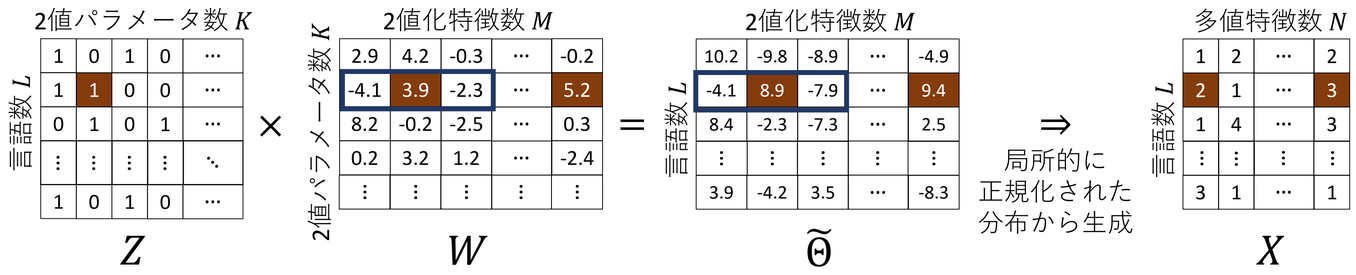
共時性と通時性を考慮した欠損値補完と潜在表現学習

(村脇, NLP2017) の拡張
To appear at IJCNLP 2017

確率的生成モデル: $P(A,Z,W,X) = \underbrace{P(A)P(Z|A)}_{\text{通時性}}\underbrace{P(W)P(X|Z,W)}_{\text{共時性}}$

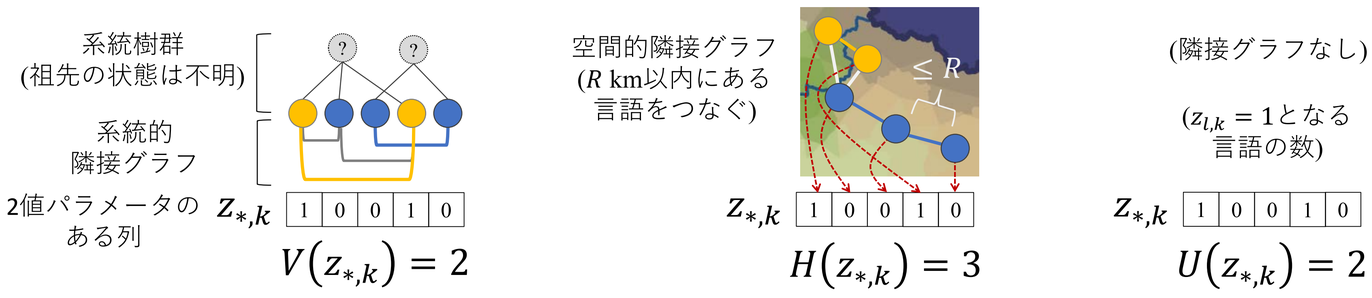
X: 表層的特徴行列, Z: 潜在的2値パラメータ行列
A: 通時性を表すパラメータ列, W: 共時性を表す重み行列

共時性のモデル化 (行の依存関係)



- 表層的特徴間に依存関係がある (Greenberg, 1963) (Baker, 2002) ほか
- 依存関係を行列分解によりモデル化

通時性のモデル化 (列の依存関係)



- 類型論では系統的変化以外に**接触**による変化も重要 (Trubetzkoy, 1928) ほか
- 2種類の隣接グラフを用いた自己ロジスティックモデルにより、現在の2値パラメータの分布をモデル化 (Yamauchi+, COLING2016)

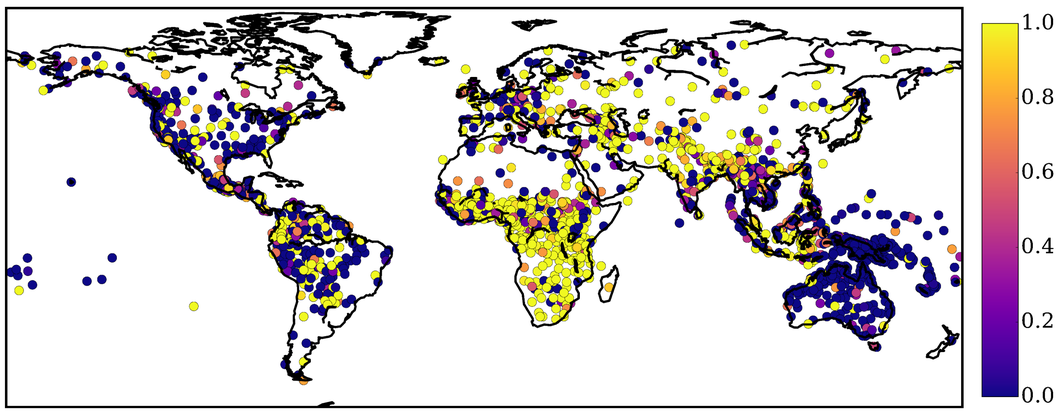
$$P(z_{l,k} = b | \mathbf{z}_{-l,k}, v_k, h_k, u_k) \propto \exp(v_k V_{l,k,b} + h_k H_{l,k,b} + u_k b)$$

- ある言語のとる値は、その値をとる系統的・空間的隣接ノードの数に対数空間で比例
- 系統的安定性、空間的拡散性の強さはそれぞれ$v_k, h_k > 0$が制御

MCMC法による事後推論の結果

既知の特徴を隠した際の欠損値補完精度
(10分割交差確認)

種類	モデル	精度 (%)
通時	最頻値	60.95
	表層DIA (Yamauchi+, 2016)	66.22
共時	MCA (Josse+, 2012)	69.88
	SYN ($K = 50$)	73.83
	SYN ($K = 100$)	72.87
共時・通時	SYNDIA ($K = 50$)	**74.46**
	SYNDIA ($K = 100$)	74.00



推定された潜在的2値パラメータの地理的分布の例
(色が薄いほどサンプル中の$z_{l,k} = 1$の割合が高い)

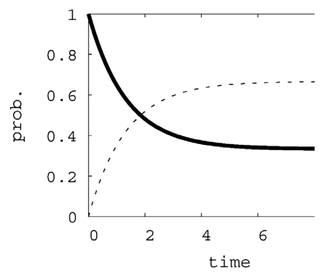
- 欠損値が高精度に補完できた
- 表層的特徴に見られる**地理的シグナル**が潜在的2値パラメータでも維持されている
  - 通時性のみをモデル化した場合 (SYN) は、潜在空間への写像時に地理的シグナルが破壊された

歴史的変化の解明: 課題の整理

- 言語は「進歩」しているのか? − No?
  - 「未開」社会の言語が単純であるという証拠は確認できない
- なぜ世界の言語は多様なのか?
  - 生物多様性は環境の多様性で説明できそう
  - 言語の構造的特徴に対する言語外の影響は、いくつか可能性が指摘されているものがあるが、全体としては限定的
    - 放出子音の有無と標高との相関 (Everett, 2013)
    - 高周波音が伝わりにくい環境 (高温等) の言語では共鳴音の割合が高い (Maddieson+, 2015)
- どんな風にでも変化するのか? − No (ではなぜ?)
  - 特徴間に依存関係があり、とり得る組み合わせは限られる
- 変化に方向性はあるか? − 少なくとも部分的にYes?
  - ミクロな変化要因の一つは文法化 (逆変化を考えにくい)
    - e.g. 指示詞から定冠詞が、1を表す数詞から不定冠詞が発達
  - 循環的変化の可能性が指摘されている
    - 膠着語 → 屈折語 → 孤立語 → 膠着語 (19世紀言語学の議論)

今後の方向性

$P(x = 0 | \pi(x) = 1, t) = \exp(tQ)_{1,0}$
1から始まり、時間 t 後に0となる確率
遷移率行列 $Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$



- 系統樹と葉ノードの状態が与えられれば、連続時間マルコフ連鎖によるモデル化で、遷移率行列と祖先の状態が推定できる
  - 系統樹: Glottolog の階層的分類 (WALSの簡易系統樹より詳細)
  - 葉ノードの状態: 今回事後推論により得られたサンプル
- 個別の表層的特徴を調べた研究はある
  - 基本語順の遷移率行列と祖語の語順の推定 (Maurits+, 2014)
  - 各特徴の系統的安定性の推定 (Dediu, 2010) (Greenhill+, 2010)
  - 特徴**対**の共進化 (Dunn+, 2011)
- 表層的特徴列ではなく、潜在的2値パラメータを使って遷移率行列と祖先の状態を推定し、次に表層的特徴列を生成してみる
  - 共時的に自然な状態の祖語を推定できるはず