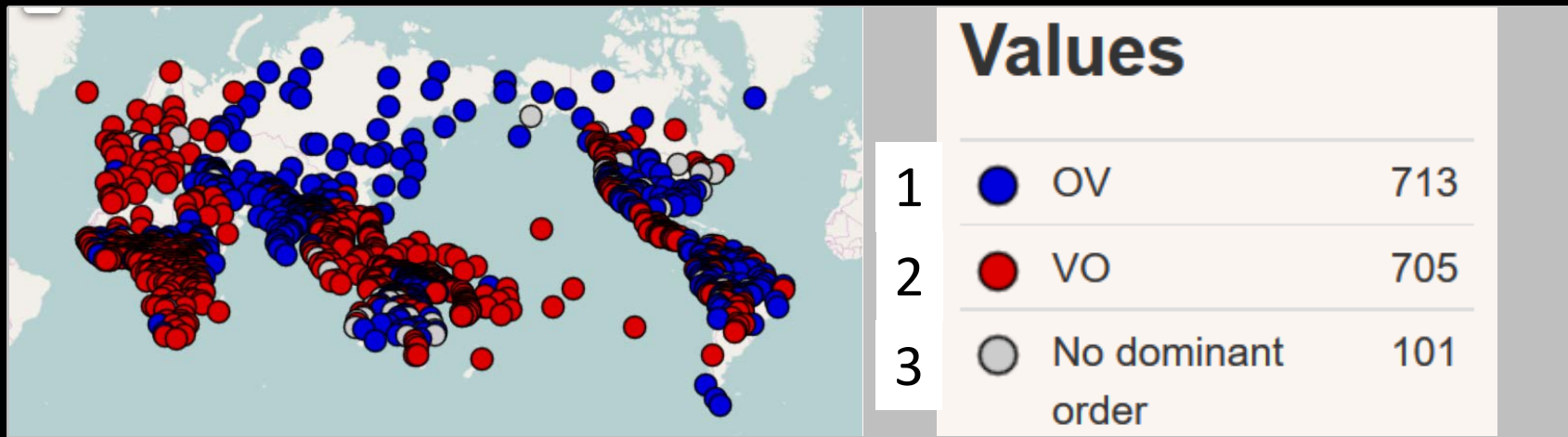


言語類型論的特徴からの 潜在的2値パラメータの獲得

京都大学
村脇 有吾

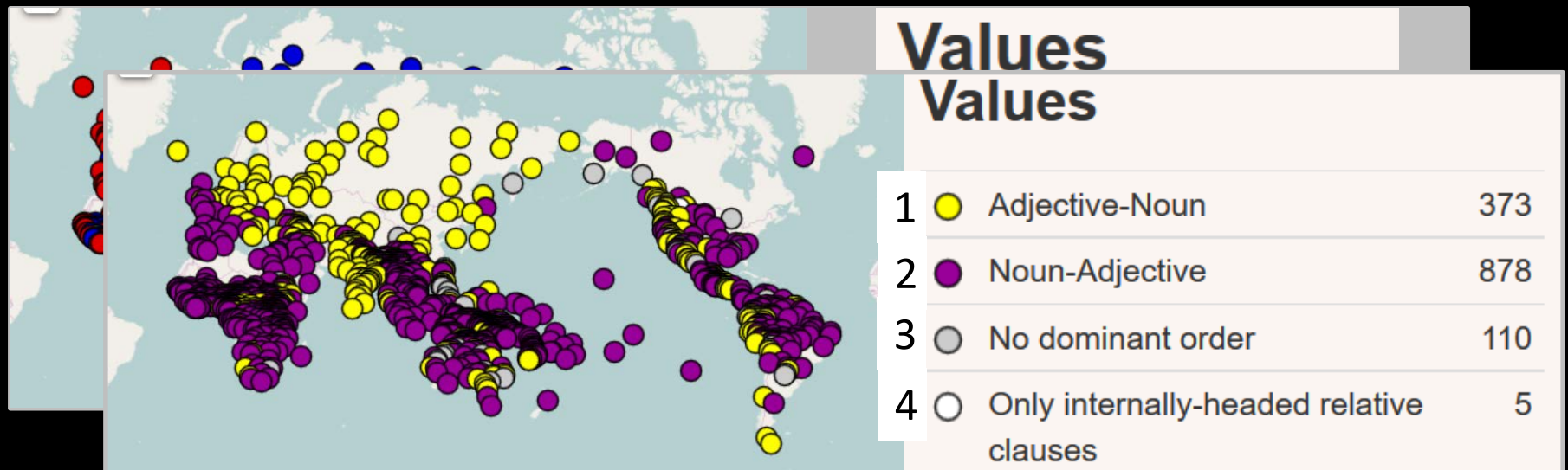


言語類型論的特徴



[Haspelmath+, 2005]

言語類型論的特徴

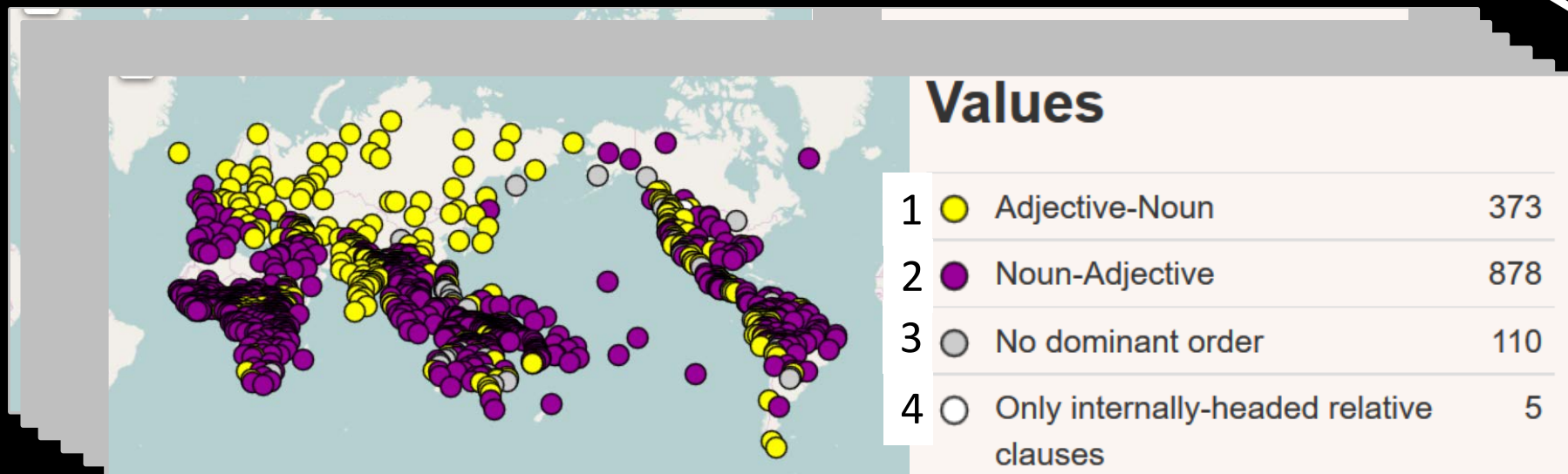


[Haspelmath+, 2005]

言語類型論的特徴

(元データは192特徴)

$N = 104$ 個の特徴



[Haspelmath+, 2005]

言語類型論的特徴

(元データは192特徴)

$N = 104$ 個の特徴

$L = 2,679$ 個の言語

Values

- 1 ● Adjective-Noun 373
- 2 ● Noun-Adjective 878
- 3 ○ No dominant order 110
- 4 ○ Only internally-headed relative clauses 5

多値特徴数 N

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

言語数 L

X 2

(元データは192特徴)

$N = 104$ 個の特徴

[Haspelmath+, 2005]

言語類型論的特徴

(元データは192特徴)

$N = 104$ 個の特徴

$L = 2,679$ 個の言語

73.1%が欠損値

(元データは192特徴)

$N = 104$ 個の特徴

Values

- 1 ● Adjective-Noun 375
- 2 ● Noun-Adjective 878
- 3 ○ No dominant order 110
- 4 ○ Only internally-headed relative clauses 5

多値特徴数 N

1	2	...	2
2	?	...	3
?	4	...	?
⋮	⋮	⋮	⋮
3	1	...	?

言語数 L

X 2

[Haspelmath+, 2005]

言語類型論的特徴

(元データは192特徴)

$N = 104$ 個の特徴

$M = 723$ 個の2値化特徴

$L = 2,679$ 個の言語

73.1%が欠損値

(元データは192特徴)

$N = 104$ 個の特徴

Values

- 1 ● Adjective-Noun 375
- 2 ● Noun-Adjective 878
- 3 ○ No dominant order 110
- 4 ○ Only internally-headed relative clauses 5

2値化特徴数 M

多値特徴数 N

1	0	0	0	...	0
0	1	0	1	...	1
1	0	0	0	...	1
⋮	⋮	⋮	⋮	⋮	⋮
0	0	1	1	...	0

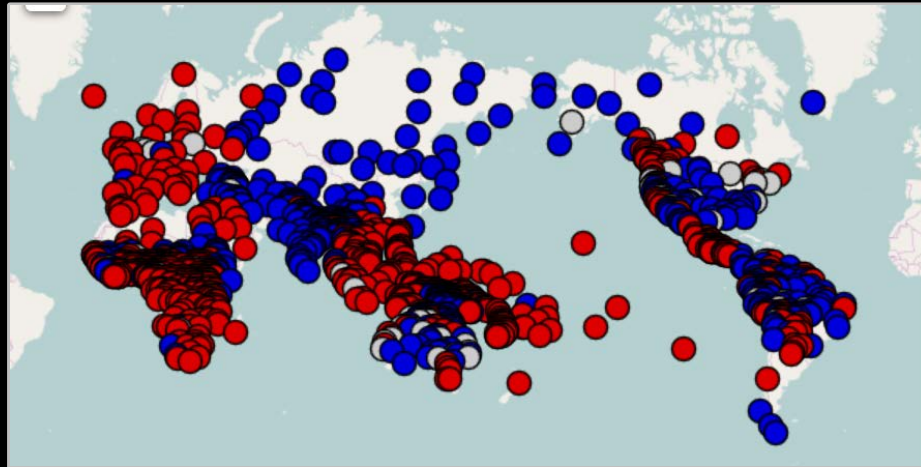
≡

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

$X \quad 2$

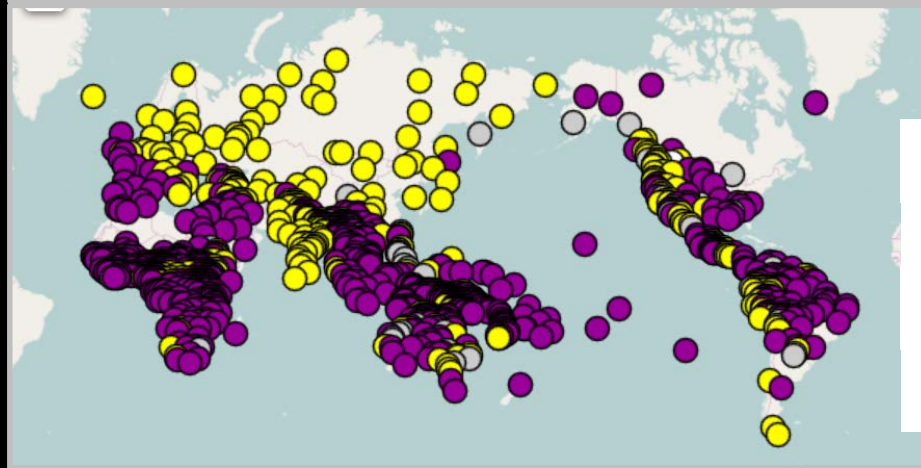
※これは
潜在的2値パラメータでは
ありません

類型論的特徴は構造を持つ



Values

1	● OV	713
2	● VO	705
3	○ No dominant order	101

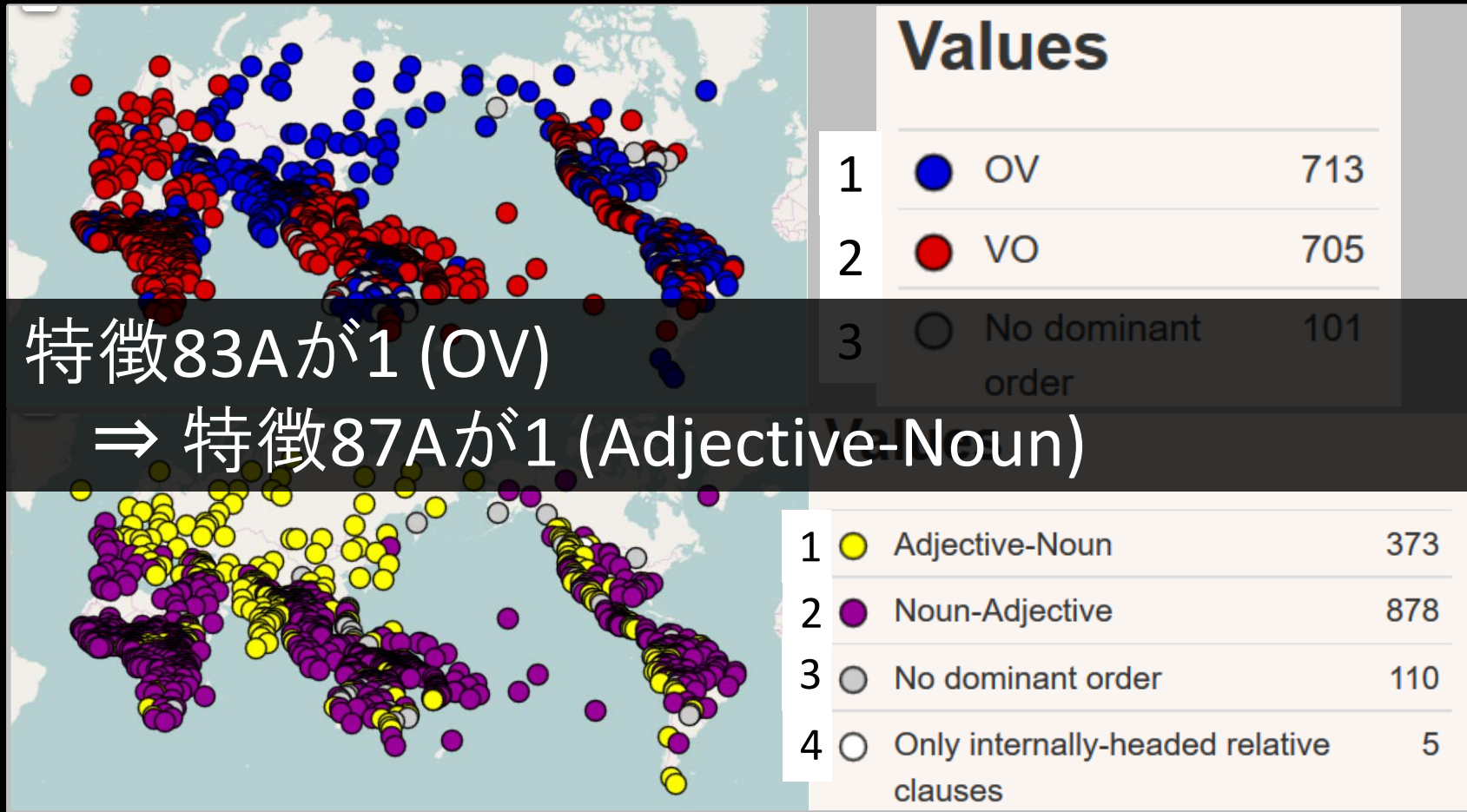


Values

1	● Adjective-Noun	373
2	● Noun-Adjective	878
3	○ No dominant order	110
4	○ Only internally-headed relative clauses	5

[Greenberg, 1963]

類型論的特徴は構造を持つ



[Greenberg, 1963]

構造を反映した潜在表現を仮定

Z

言語数 L

???
???
???
⋮
???



何らかの生成過程



何らかの潜在表現

多値特徴数 N

言語数 L

1	2	⋯	2
2	1	⋯	3
1	4	⋯	3
⋮	⋮	⋮	⋮
3	1	⋯	1

X 4

構造を反映した潜在表現を仮定

Z

言語数 L

???
???
???
⋮
???



何らかの生成過程



何らかの潜在表現

Q: どのような潜在表現?

多値特徴数 N

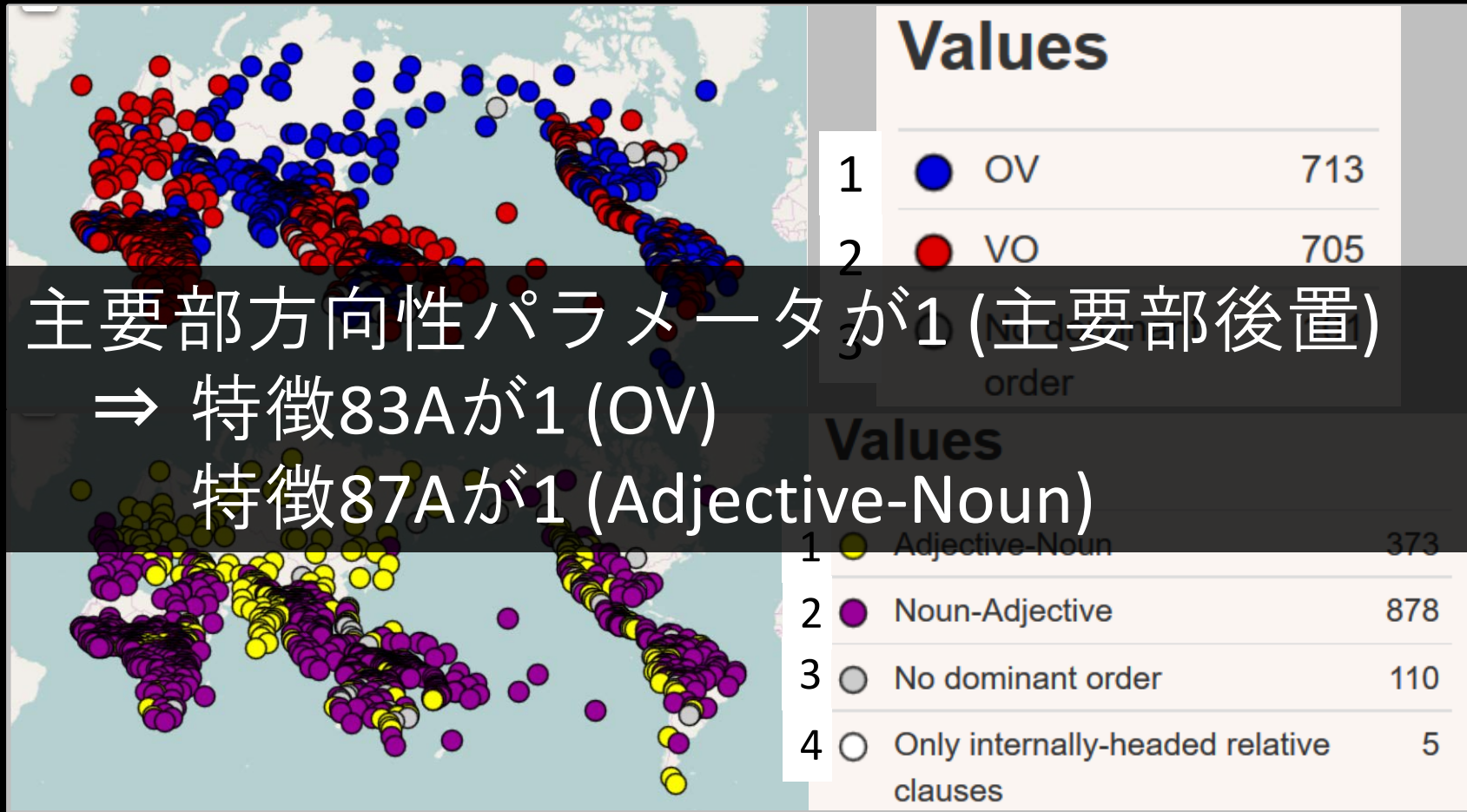
言語数 L

1	2	⋯	2
2	1	⋯	3
1	4	⋯	3
⋮	⋮	⋮	⋮
3	1	⋯	1

X

4

A: 潜在的2値パラメータ



[Baker, 2002]

潜在的2値パラメータ

Z

2値パラメータ数 K

言語数 L

1	0	1	0	...
1	1	0	0	...
0	1	0	1	...
⋮	⋮	⋮	⋮	⋱
1	0	1	0	...



決定的な生成過程



例外があると破綻

多値特徴数 N

言語数 L

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

X 6

提案モデル: 確率的生成

$$\begin{matrix} & Z & & W & & \tilde{\Theta} \\ & \text{2値パラメータ数 } K & & \text{2値化特徴数 } M & & \text{2値化特徴数 } M \\ \text{言語数 } L & \begin{matrix} 1 & 0 & 1 & 0 & \dots \\ 1 & 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ 1 & 0 & 1 & 0 & \dots \end{matrix} & \times & \begin{matrix} \text{2値パラメータ数 } K \\ \begin{matrix} 2.9 & 4.2 & -0.3 & \dots & -0.2 \\ -4.1 & 3.9 & -2.3 & \dots & 5.2 \\ 8.2 & -0.2 & -2.5 & \dots & 0.3 \\ 0.2 & 3.2 & 1.2 & \dots & -2.4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{matrix} \end{matrix} & = & \begin{matrix} \text{言語数 } L \\ \begin{matrix} 10.2 & -9.8 & -8.9 & \dots & -4.9 \\ -4.1 & 8.9 & -7.9 & \dots & 9.4 \\ 8.4 & -2.3 & -7.3 & \dots & 2.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 3.9 & -4.2 & 3.5 & \dots & -8.3 \end{matrix} \end{matrix} \end{matrix}$$

局所的に正規化された
分布から生成



多値特徴数 N

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

X 7

提案モデル: 確率的生成

$$\begin{matrix} & Z & & W & & \tilde{\Theta} \\ & \text{2値パラメータ数 } K & & \text{2値化特徴数 } M & & \text{2値化特徴数 } M \\ \text{言語数 } L & \begin{matrix} 1 & 0 & 1 & 0 & \dots \\ 1 & 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ 1 & 0 & 1 & 0 & \dots \end{matrix} & \times & \begin{matrix} \text{2値パラメータ数 } K \\ \begin{matrix} 2.9 & 4.2 & -0.3 & \dots & -0.2 \\ -4.1 & 3.9 & -2.3 & \dots & 5.2 \\ 8.2 & -0.2 & -2.5 & \dots & 0.3 \\ 0.2 & 3.2 & 1.2 & \dots & -2.4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{matrix} \end{matrix} & = & \begin{matrix} \text{言語数 } L \\ \begin{matrix} 10.2 & -9.8 & -8.9 & \dots & -4.9 \\ -4.1 & 8.9 & -7.9 & \dots & 9.4 \\ 8.4 & -2.3 & -7.3 & \dots & 2.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 3.9 & -4.2 & 3.5 & \dots & -8.3 \end{matrix} \end{matrix} \end{matrix}$$

局所的に正規化された
分布から生成

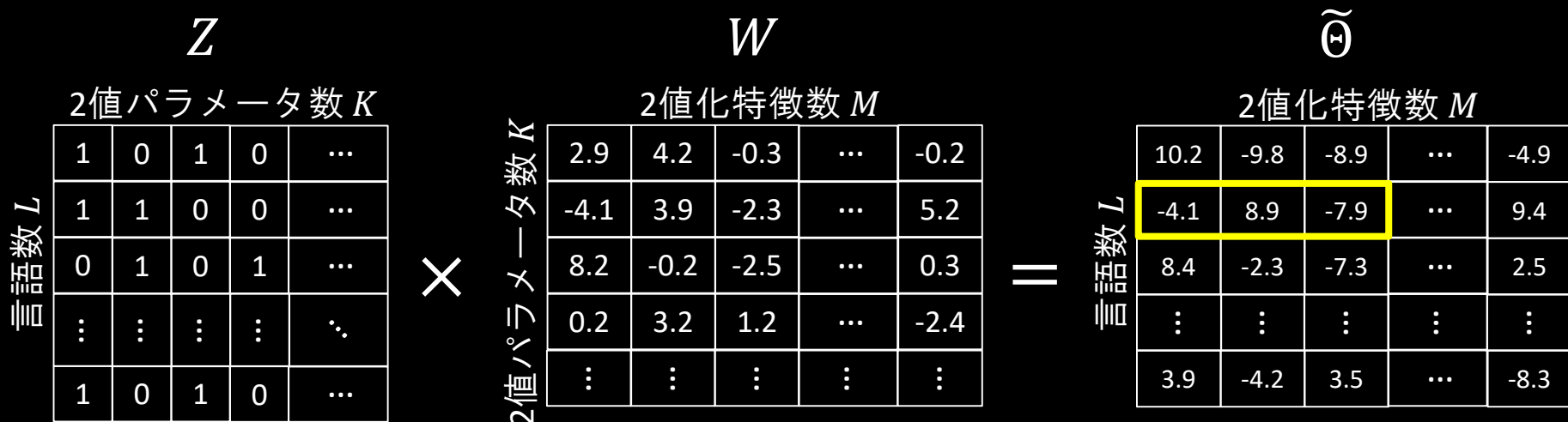


多値特徴数 N

1	2	...	2
2	1	...	3
1	4	...	3
...
3	1	...	1

X 7

提案モデル: 確率的生成



局所的に正規化された分布から生成



$$P(x_{2,1} = 2) = \frac{\exp(\tilde{\theta}_{2,2})}{\sum_{j=1}^3 \exp(\tilde{\theta}_{2,j})}$$

多値特徴数 N

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

X 7

提案モデル: 確率的生成

$$\begin{matrix} & Z & & W & & \tilde{\Theta} \\ & \text{2値パラメータ数 } K & & \text{2値化特徴数 } M & & \text{2値化特徴数 } M \\ \text{言語数 } L & \begin{matrix} 1 & 0 & 1 & 0 & \dots \\ 1 & 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ 1 & 0 & 1 & 0 & \dots \end{matrix} & \times & \begin{matrix} \text{2値パラメータ数 } K \\ \begin{matrix} 2.9 & 4.2 & -0.3 & \dots & -0.2 \\ -4.1 & 3.9 & -2.3 & \dots & 5.2 \\ 8.2 & -0.2 & -2.5 & \dots & 0.3 \\ 0.2 & 3.2 & 1.2 & \dots & -2.4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{matrix} \end{matrix} & = & \begin{matrix} \text{言語数 } L \\ \begin{matrix} 10.2 & -9.8 & -8.9 & \dots & -4.9 \\ -4.1 & 8.9 & -7.9 & \dots & 9.4 \\ 8.4 & -2.3 & -7.3 & \dots & 2.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 3.9 & -4.2 & 3.5 & \dots & -8.3 \end{matrix} \end{matrix} \end{matrix}$$

局所的に正規化された
分布から生成

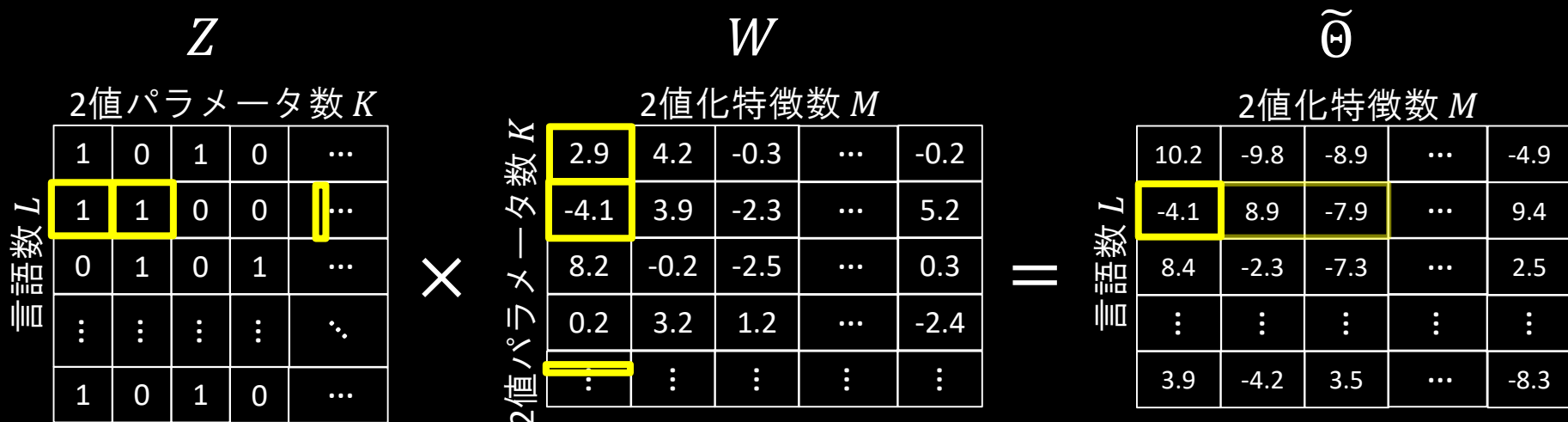


多値特徴数 N

1	2	...	2
2	1	...	3
1	4	...	3
...
3	1	...	1

X 8

提案モデル: 確率的生成



局所的に正規化された分布から生成



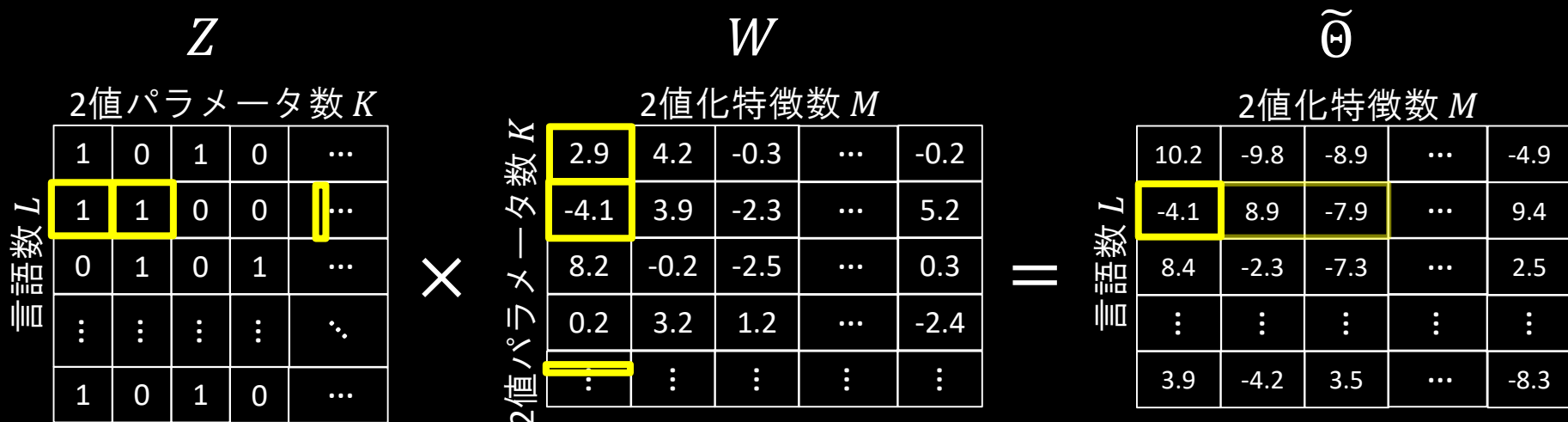
$$\tilde{\theta}_{2,1} = \sum_k z_{2,k} \times w_{k,1}$$

多値特徴数 N

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

X 8

提案モデル: 確率的生成



局所的に正規化された分布から生成



$$\tilde{\theta}_{2,1} = \sum_k z_{2,k} \times w_{k,1}$$

$z_{l,k} = 1$ に対応する w のみ影響

多値特徴数 N

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

提案モデル: 確率的生成

$$\begin{matrix} & Z & & & W & & & \tilde{\Theta} \\ & \text{2値パラメータ数 } K & & & \text{2値化特徴数 } M & & & \text{2値化特徴数 } M \\ \text{言語数 } L & \begin{matrix} 1 & 0 & 1 & 0 & \dots \\ 1 & 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ 1 & 0 & 1 & 0 & \dots \end{matrix} & \times & & \begin{matrix} \text{2値パラメータ数 } K \\ \begin{matrix} 2.9 & 4.2 & -0.3 & \dots & -0.2 \\ -4.1 & 3.9 & -2.3 & \dots & 5.2 \\ 8.2 & -0.2 & -2.5 & \dots & 0.3 \\ 0.2 & 3.2 & 1.2 & \dots & -2.4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{matrix} \end{matrix} & = & & \begin{matrix} \text{言語数 } L \\ \begin{matrix} 10.2 & -9.8 & -8.9 & \dots & -4.9 \\ -4.1 & 8.9 & -7.9 & \dots & 9.4 \\ 8.4 & -2.3 & -7.3 & \dots & 2.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 3.9 & -4.2 & 3.5 & \dots & -8.3 \end{matrix} \end{matrix} \end{matrix}$$

局所的に正規化された
分布から生成

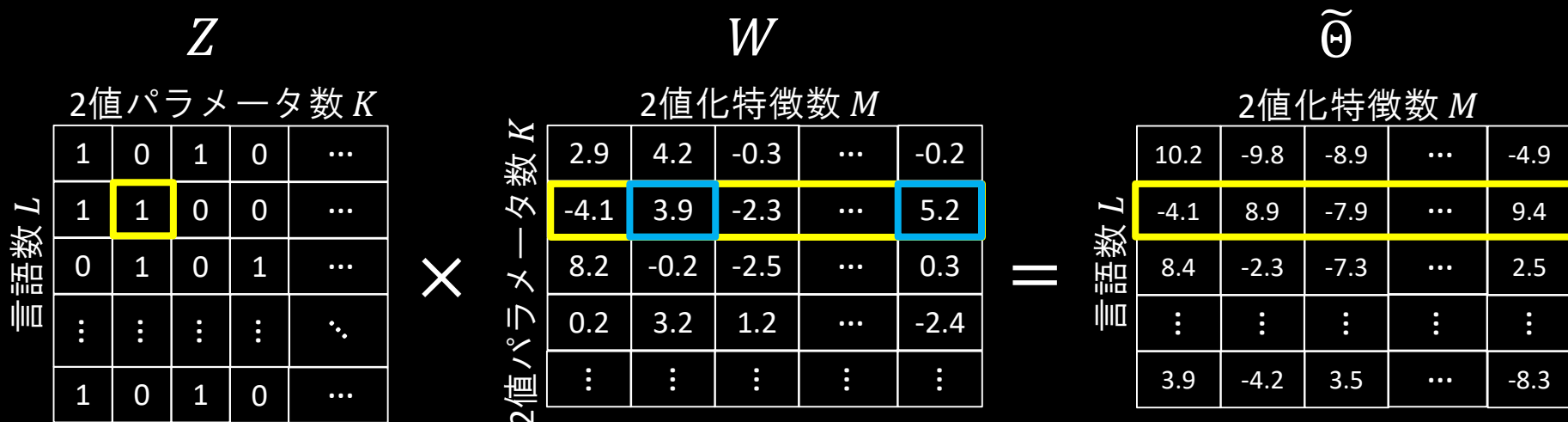


多値特徴数 N

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

X 9

提案モデル: 確率的生成



局所的に正規化された分布から生成



このパラメータで1をとる言語群ではある特徴のある値と別の特徴のある値を同時にとりやすい

多値特徴数 N

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

提案モデル: 確率的生成

$$\begin{array}{c} \text{言語数 } L \\ \times \\ \begin{array}{c} Z \\ \text{2値パラメータ数 } K \\ \begin{array}{|c|c|c|c|c|} \hline 1 & 0 & 1 & 0 & \dots \\ \hline 1 & 1 & 0 & 0 & \dots \\ \hline 0 & 1 & 0 & 1 & \dots \\ \hline \vdots & \vdots & \vdots & \vdots & \ddots \\ \hline 1 & 0 & 1 & 0 & \dots \\ \hline \end{array} \end{array} \end{array} \times \begin{array}{c} \begin{array}{c} W \\ \text{2値化特徴数 } M \\ \begin{array}{|c|c|c|c|c|} \hline 2.9 & 4.2 & -0.3 & \dots & -0.2 \\ \hline -4.1 & 3.9 & -2.3 & \dots & 5.2 \\ \hline 8.2 & -0.2 & -2.5 & \dots & 0.3 \\ \hline 0.2 & 3.2 & 1.2 & \dots & -2.4 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline \end{array} \end{array} \end{array} = \begin{array}{c} \begin{array}{c} \tilde{\Theta} \\ \text{2値化特徴数 } M \\ \begin{array}{|c|c|c|c|c|} \hline 10.2 & -9.8 & -8.9 & \dots & -4.9 \\ \hline -4.1 & 8.9 & -7.9 & \dots & 9.4 \\ \hline 8.4 & -2.3 & -7.3 & \dots & 2.5 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline 3.9 & -4.2 & 3.5 & \dots & -8.3 \\ \hline \end{array} \end{array} \\ \text{言語数 } L \end{array}$$

局所的に正規化された
分布から生成



行列 Z はインディアンビューツフェ過程
[Griffiths+, 2011] から生成

多値特徴数 N

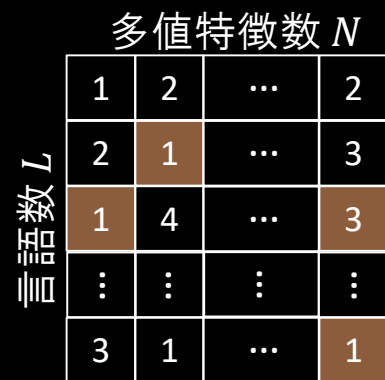
1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

X 10

Gibbsサンプリングによる推論



局所的に正規化された
分布から生成



Gibbsサンプリングによる推論

$$\begin{array}{c} Z \\ \text{2値パラメータ数 } K \\ \text{言語数 } L \\ \begin{array}{|c|c|c|c|c|} \hline 1 & 0 & 1 & 0 & \dots \\ \hline 1 & 1 & 0 & 0 & \dots \\ \hline 0 & 1 & 0 & 1 & \dots \\ \hline \vdots & \vdots & \vdots & \vdots & \ddots \\ \hline 1 & 0 & 1 & 0 & \dots \\ \hline \end{array} \\ \times \\ \begin{array}{c} W \\ \text{2値化特徴数 } M \\ \text{2値パラメータ数 } K \\ \begin{array}{|c|c|c|c|c|} \hline 2.9 & 4.2 & -0.3 & \dots & -0.2 \\ \hline -4.1 & 3.9 & -2.3 & \dots & 5.2 \\ \hline 8.2 & -0.2 & -2.5 & \dots & 0.3 \\ \hline 0.2 & 3.2 & 1.2 & \dots & -2.4 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline \end{array} \\ \end{array} \\ = \\ \begin{array}{c} \tilde{\Theta} \\ \text{2値化特徴数 } M \\ \text{言語数 } L \\ \begin{array}{|c|c|c|c|c|} \hline 10.2 & -9.8 & -8.9 & \dots & -4.9 \\ \hline -4.1 & 8.9 & -7.9 & \dots & 9.4 \\ \hline 8.4 & -2.3 & -7.3 & \dots & 2.5 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline 3.9 & -4.2 & 3.5 & \dots & -8.3 \\ \hline \end{array} \\ \end{array}
 \end{array}$$

局所的に正規化された
分布から生成



- ナイーブな手法を用いると W が現実的な時間では収束しない
- Hamiltonian Monte Carlo [Neal, 2011] により効率的にブロック・サンプリング可能

多値特徴数 N

1	2	...	2
2	1	...	3
1	4	...	3
...
3	1	...	1

X 11

Q: 獲得されたパラメータの妥当性?

Z

2値パラメータ数 K

言語数 L

1	0	1	0	...
1	1	0	0	...
0	1	0	1	...
⋮	⋮	⋮	⋮	⋱
1	0	1	0	...

×

2値パラメータ数 K

W

2値化特徴数 M

2.9	4.2	-0.3	...	-0.2
-4.1	3.9	-2.3	...	5.2
8.2	-0.2	-2.5	...	0.3
0.2	3.2	1.2	...	-2.4
⋮	⋮	⋮	⋮	⋮

=

$\tilde{\Theta}$

2値化特徴数 M

言語数 L

10.2	-9.8	-8.9	...	-4.9
-4.1	8.9	-7.9	...	9.4
8.4	-2.3	-7.3	...	2.5
⋮	⋮	⋮	⋮	⋮
3.9	-4.2	3.5	...	-8.3

局所的に正規化された
分布から生成



多値特徴数 N

言語数 L

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

X 12

Q: 獲得されたパラメータの妥当性?

Z

2値パラメータ数 K

言語数 L

1	0	1	0	...
1	1	0	0	...
0	1	0	1	...
⋮	⋮	⋮	⋮	⋮
1	0	1	0	...

×

W

2値化特徴数 M

2値パラメータ数 K

2.9	4.2	-0.3	...	-0.2
-4.1	3.9	-2.3	...	5.2
8.2	-0.2	-2.5	...	0.3
0.2	3.2	1.2	...	-2.4
⋮	⋮	⋮	⋮	⋮

=

$\tilde{\Theta}$

2値化特徴数 M

言語数 L

10.2	-9.8	-8.9	...	-4.9
-4.1	8.9	-7.9	...	9.4
8.4	-2.3	-7.3	...	2.5
⋮	⋮	⋮	⋮	⋮
3.9	-4.2	3.5	...	-8.3

局所的に正規化された
分布から生成



- パラメータそのものを定量評価するのは難しい
- 既知の特徴の一部を隠して欠損値推定
- 正しく復元できたなら、データに潜む規則性を捉えられていると言える

多値特徴数 N

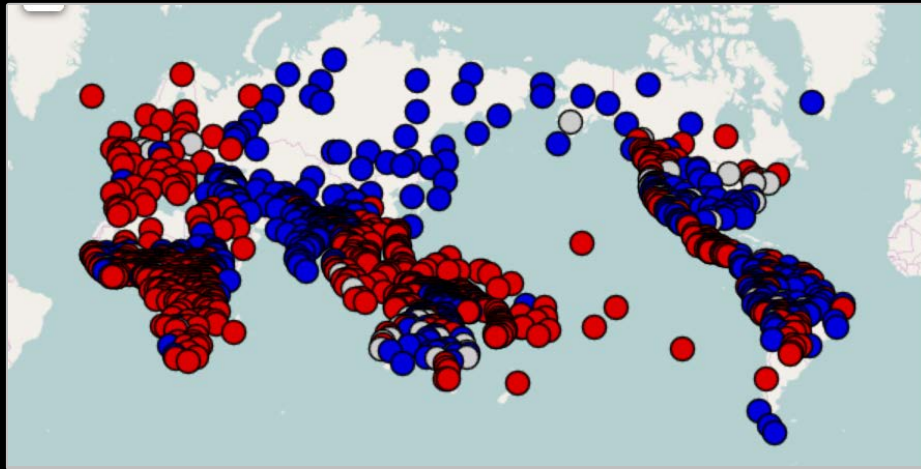
言語数 L

1	2	...	2
2	1	...	3
1	4	...	3
⋮	⋮	⋮	⋮
3	1	...	1

X 12

欠損値推定結果

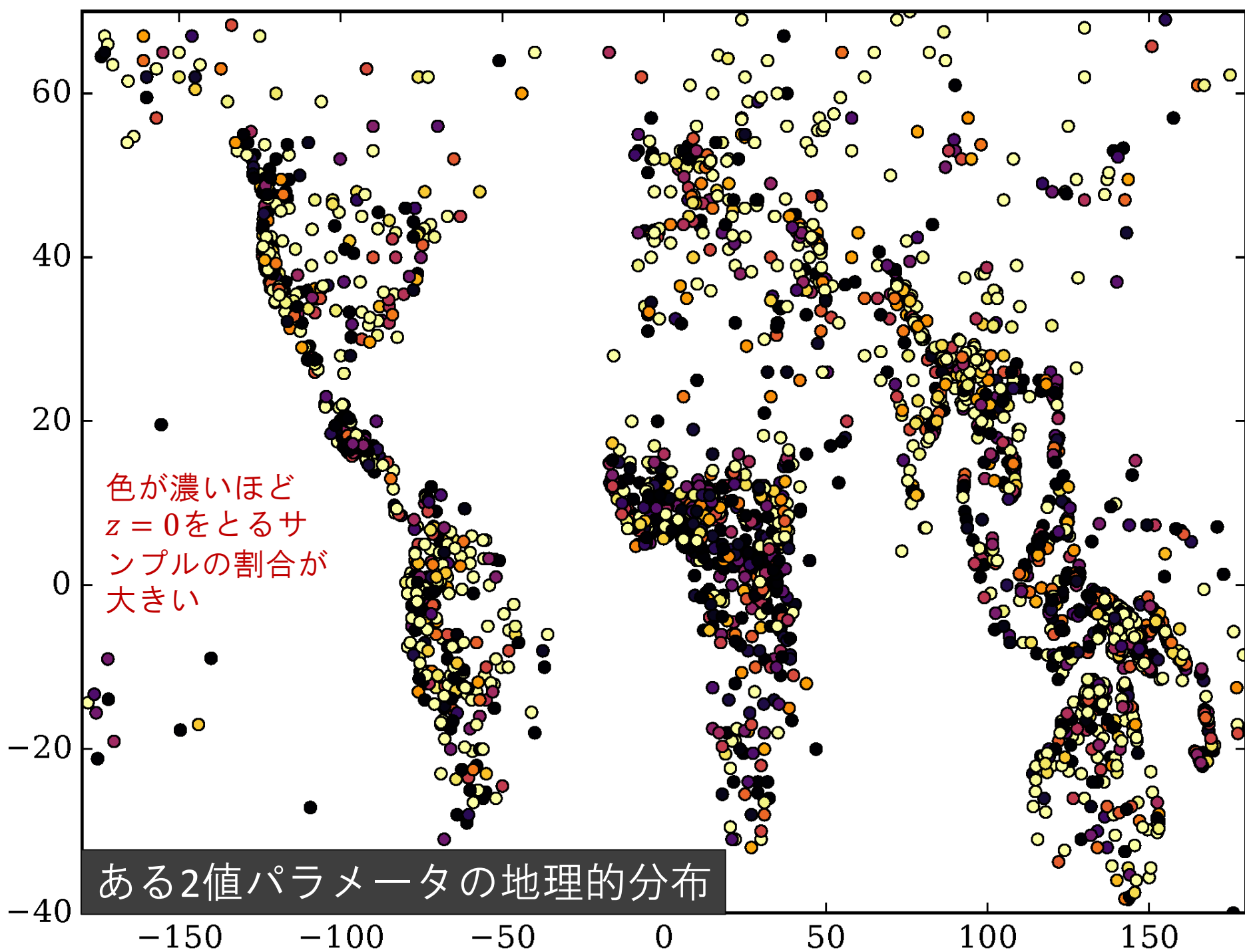
モデル	精度
最頻値	60.9%
多重対応分析の一種 [Josse+, 2012]	69.9%
提案手法 ($K_0 = 50$)	72.6%
提案手法 ($K_0 = 100$)	71.0%



Values

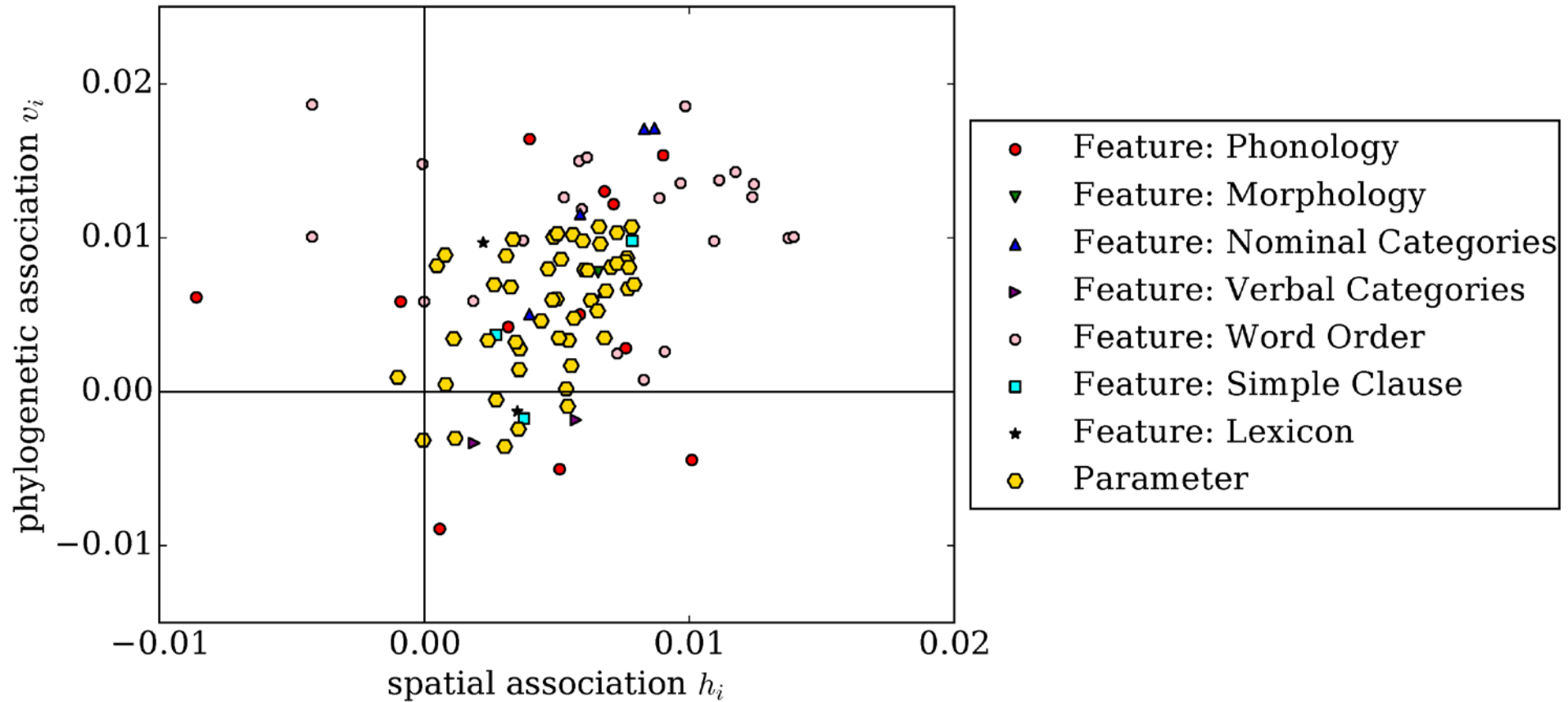
1	● OV	713
2	● VO	705
3	○ No dominant order	101

Q: 一部の表層的特徴には
系統的・地理的シグナルが見られるが、
潜在的2値パラメータはどうか?



Autologistic model [Yamauchi+, 2016]

による定量的分析



まとめ

- 多値の表層的な類型論的特徴列から潜在的2値パラメータ列を獲得するモデルを提案
- Hamiltonian Monte Carloを用いたブロックサンプリングにより推論を実現
- 得られた2値パラメータ列はデータの規則性を捉えている
- 一部の表層的特徴に見られる系統的・地理的シグナルが、なぜか2値パラメータからは消える傾向

