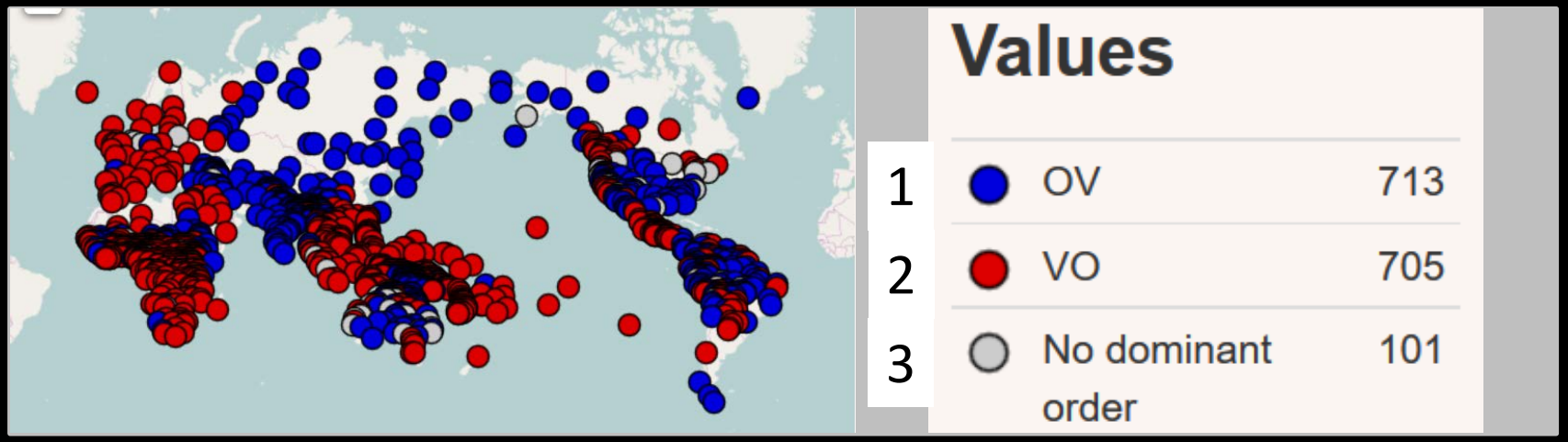
言語類型論的特徴からの潜在的2値パラメータの獲得

京都大学
村脇 有吾

言語処理学会 第23回年次大会@つくば (2017年3月16日)

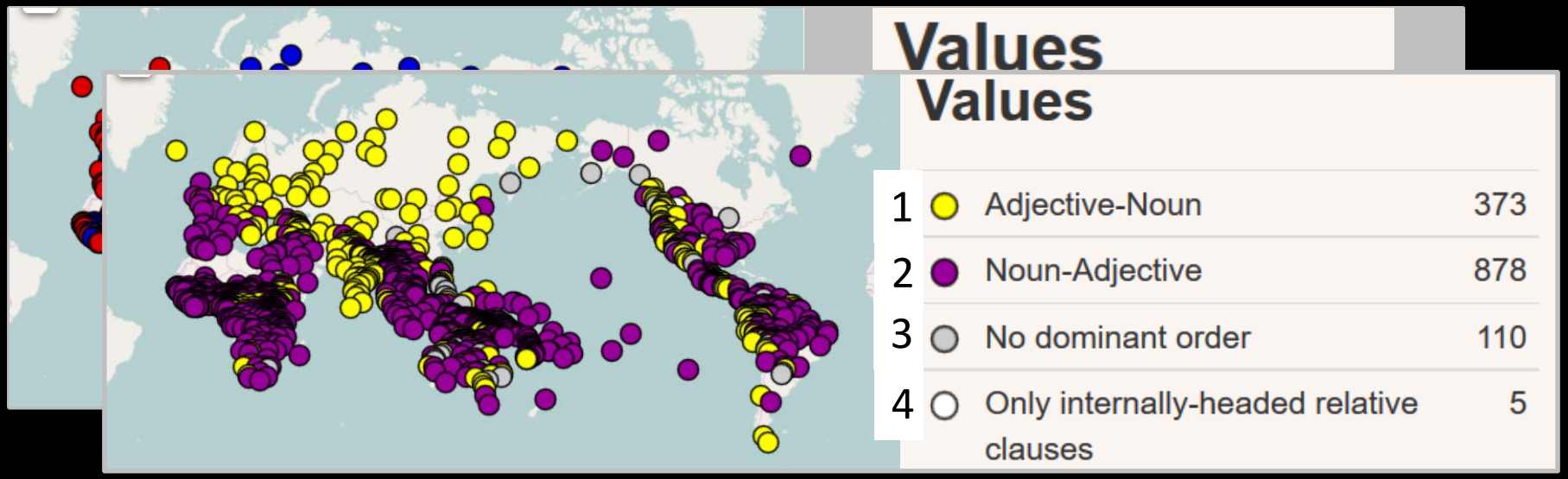
言語類型論的特徴



[Haspelmath+, 2005]

画像は http://wals.info/feature 以下の各ページから

言語類型論的特徴

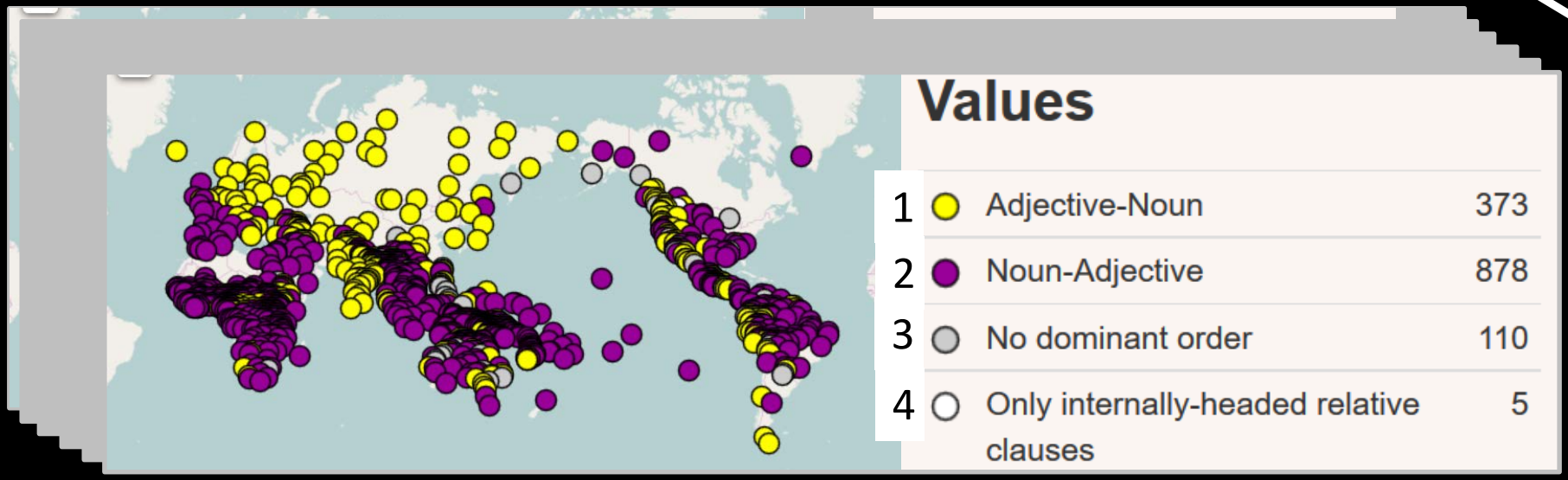


[Haspelmath+, 2005]

画像は http://wals.info/feature 以下の各ページから

言語類型論的特徴

(元データは192特徴)
$N = 104$ 個の特徴



[Haspelmath+, 2005]

画像は http://wals.info/feature 以下の各ページから

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

[Haspelmath+, 2005]

多値特徴数 N

言語数 L

1	2	…	2
2	1	…	3
1	4	…	3
⋮	⋮	⋮	⋮
3	1	…	1

X

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

[Haspelmath+, 2005]

多値特徴数 N

言語数 L

1	2	…	2
2	?	…	3
?	4	…	?
⋮	⋮	⋮	⋮
3	1	…	?

X

言語類型論的特徴

(元データは192特徴)

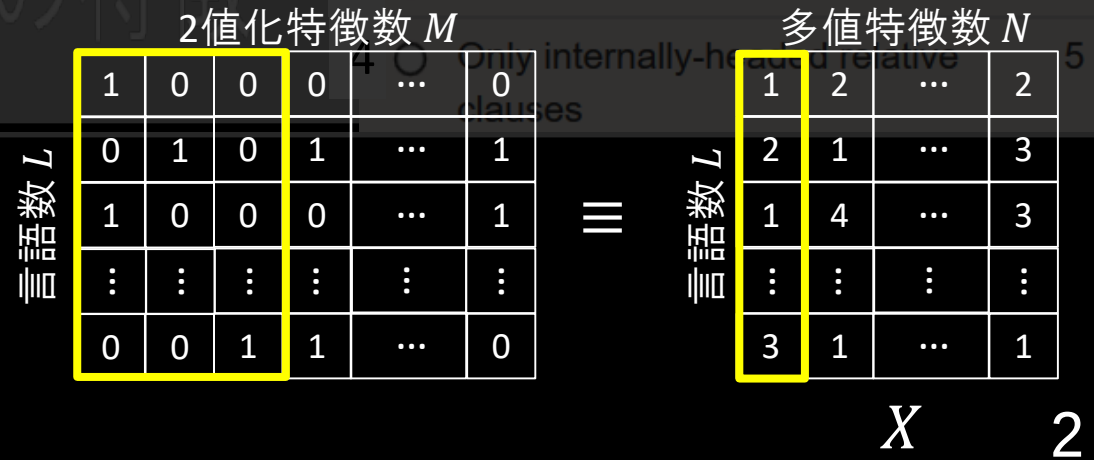
$N = 104$ 個の特徴

$M = 723$ 個の2値化特徴

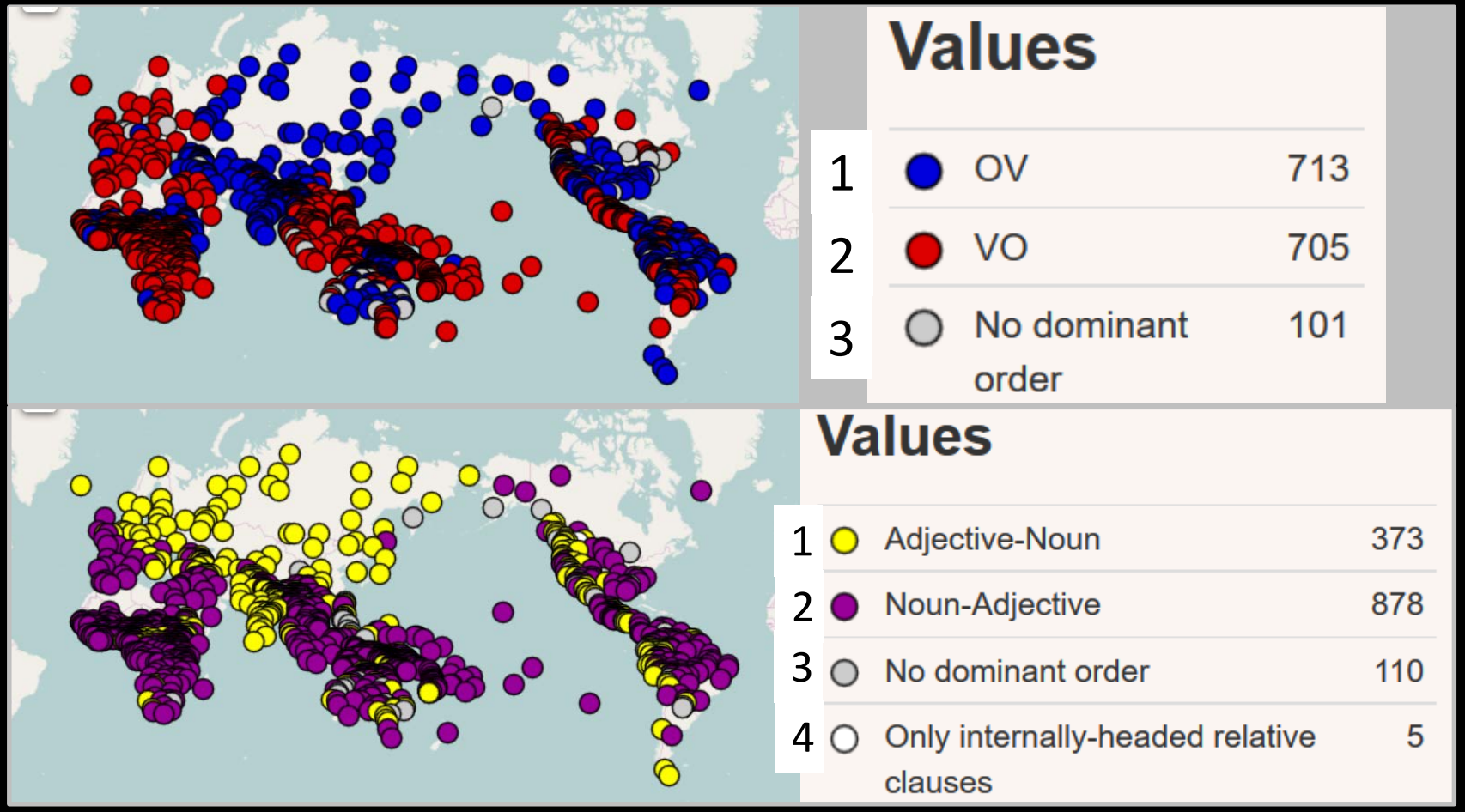
$L = 2{,}679$ 個の言語

73.1%が欠損値

※これは
潜在的2値パラメータでは
ありません

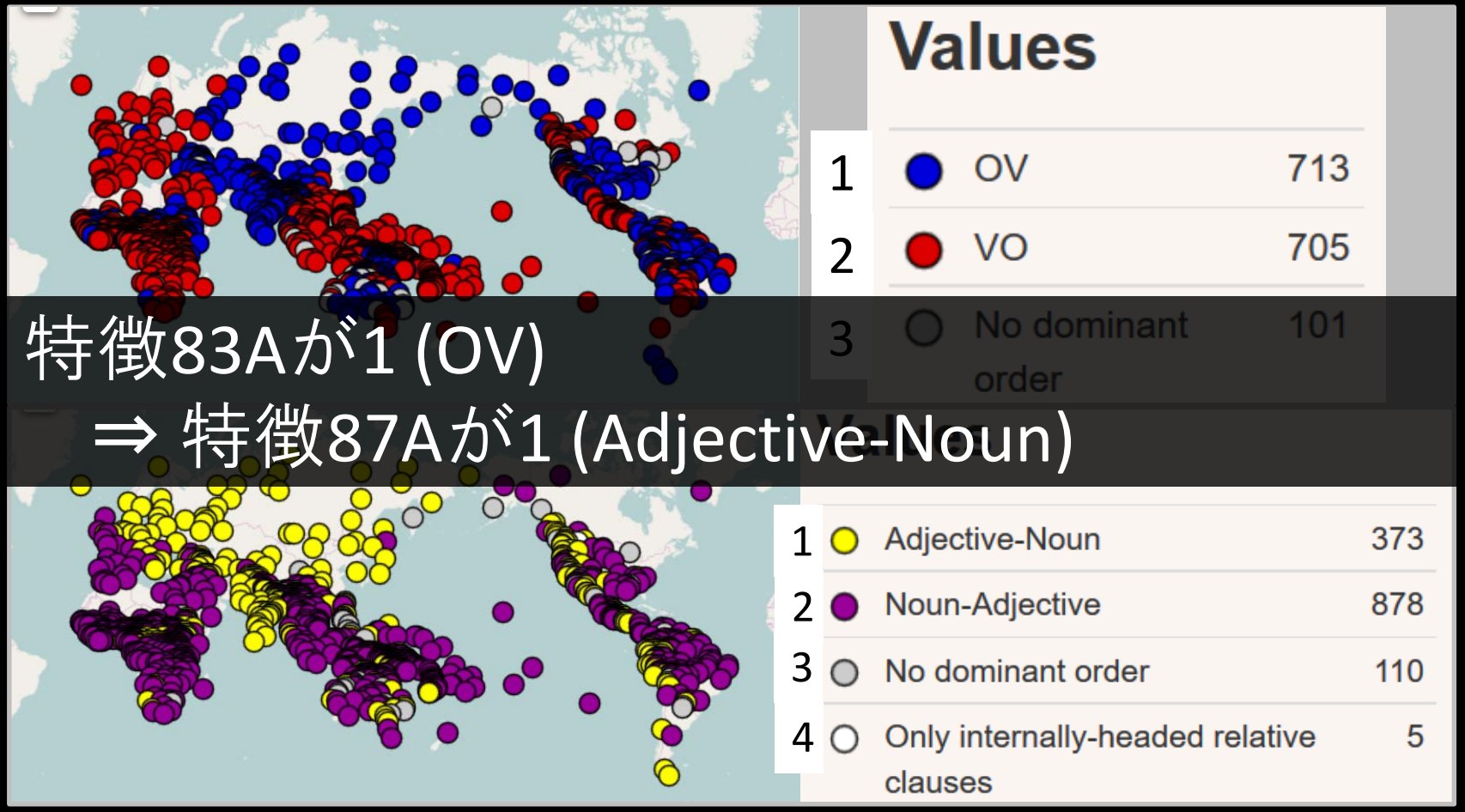


類型論的特徴は構造を持つ



[Greenberg, 1963]

類型論的特徴は構造を持つ



特徴83Aが1 (OV)

⇒ 特徴87Aが1 (Adjective-Noun)

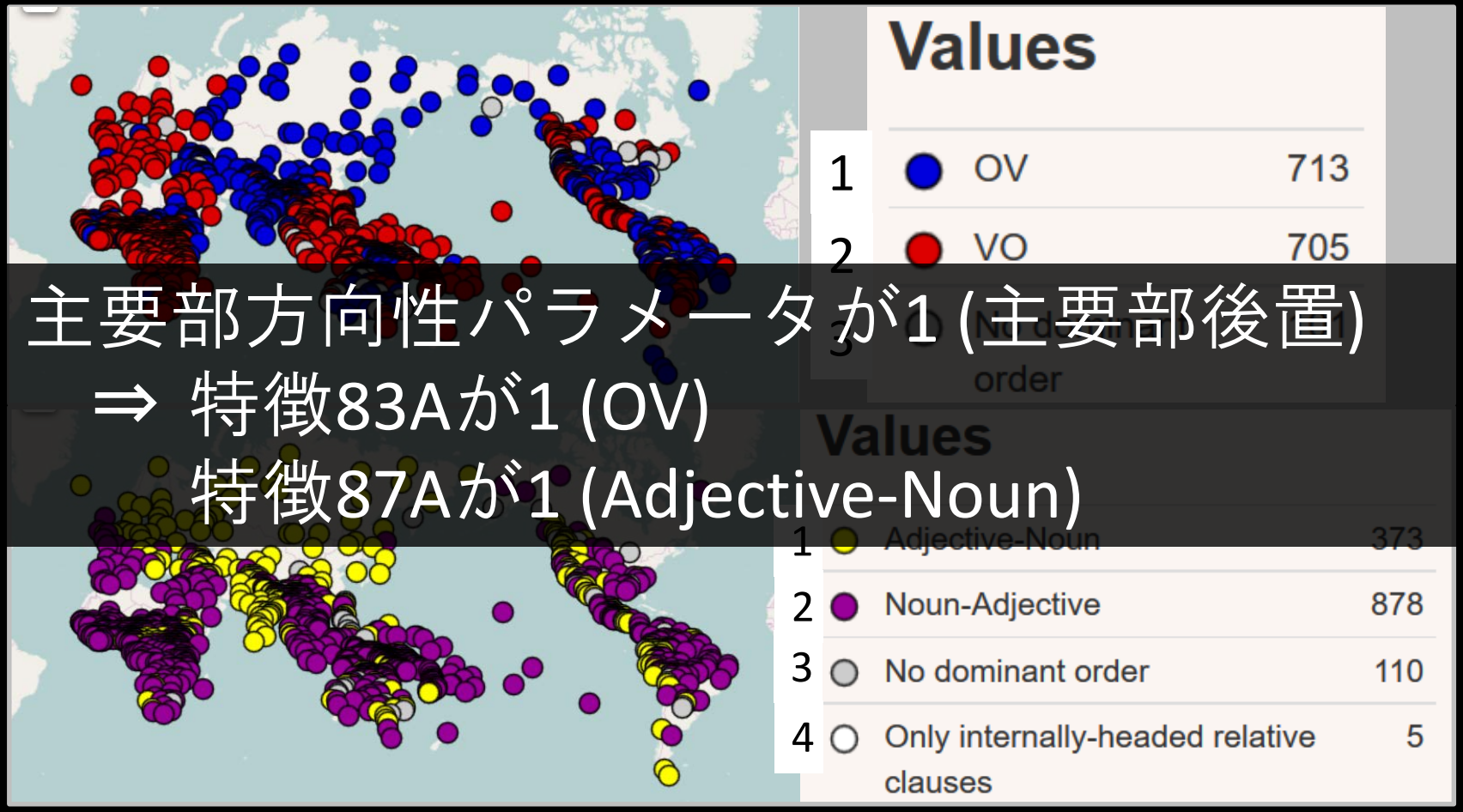
[Greenberg, 1963]

構造を反映した潜在表現を仮定

構造を反映した潜在表現を仮定

Q: どのような潜在表現?

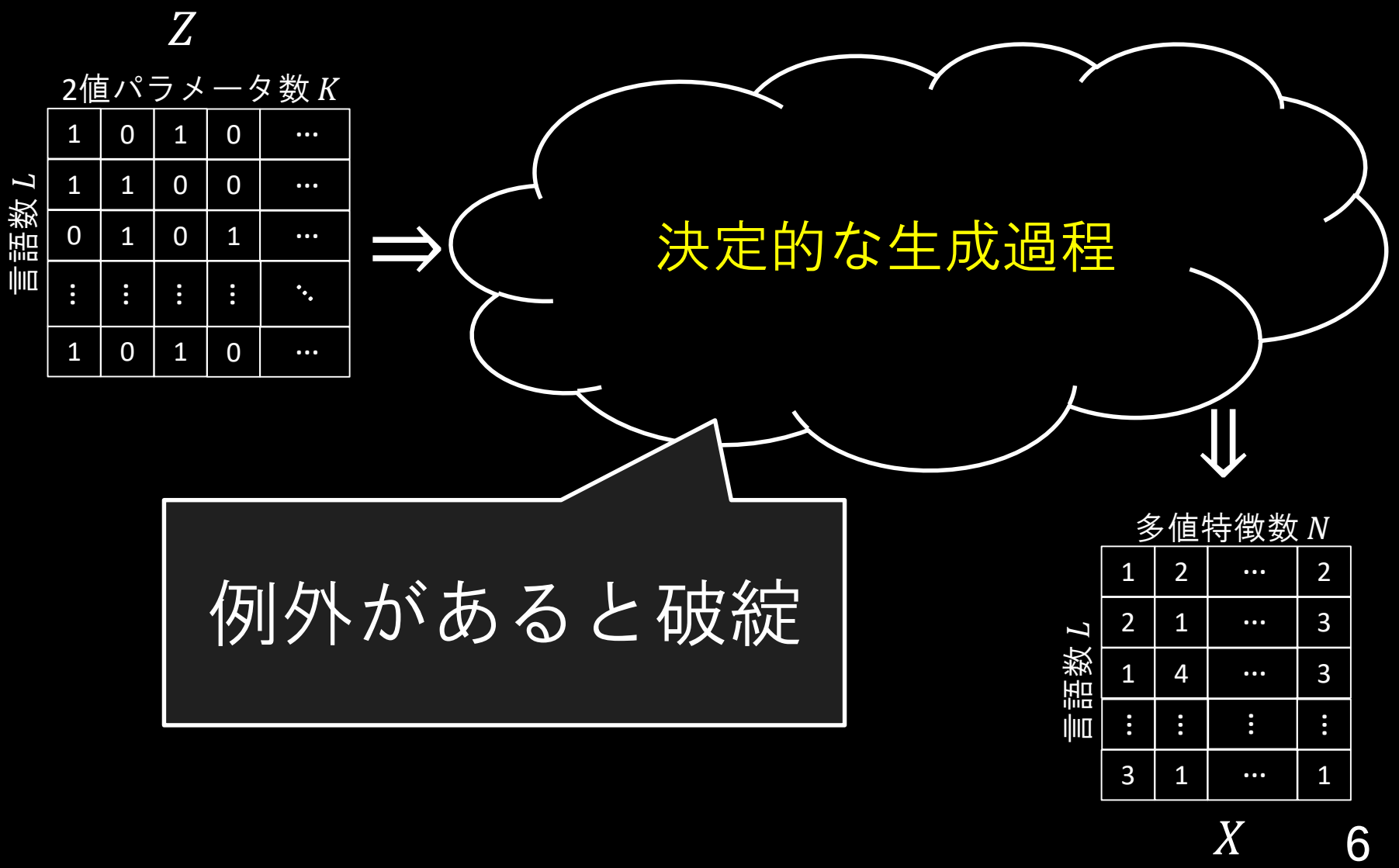
A: 潜在的2値パラメータ



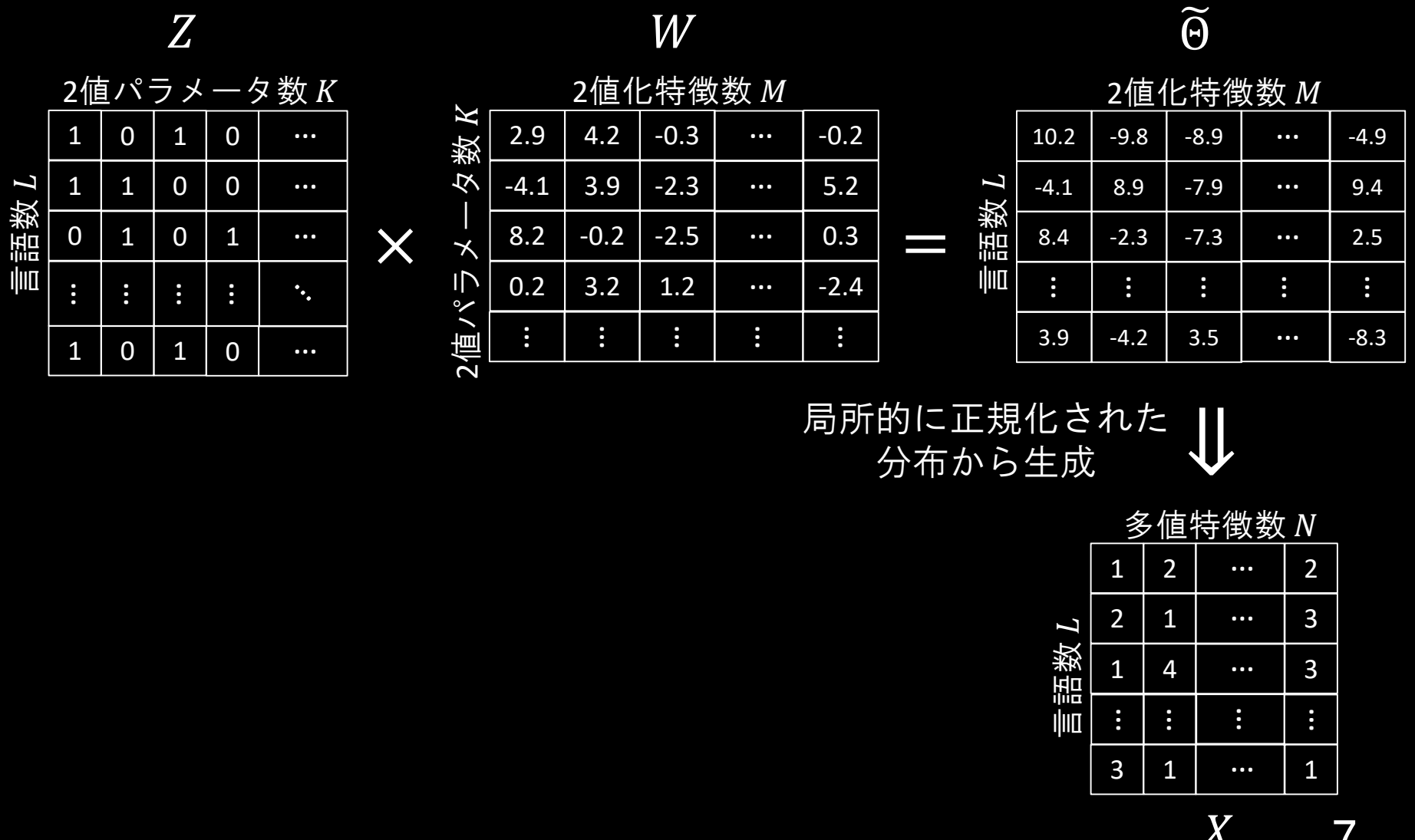
主要部方向性パラメータが1 (主要部後置)
⇒ 特徴83Aが1 (OV)
特徴87Aが1 (Adjective-Noun)

[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

提案モデル: 確率的生成


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
⇓
多値特徴数 N
言語数 L
1 2 … 2
2 1 … 3
1 4 … 3
⋮ ⋮ ⋮ ⋮
3 1 … 1
X

提案モデル: 確率的生成

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 ⇓

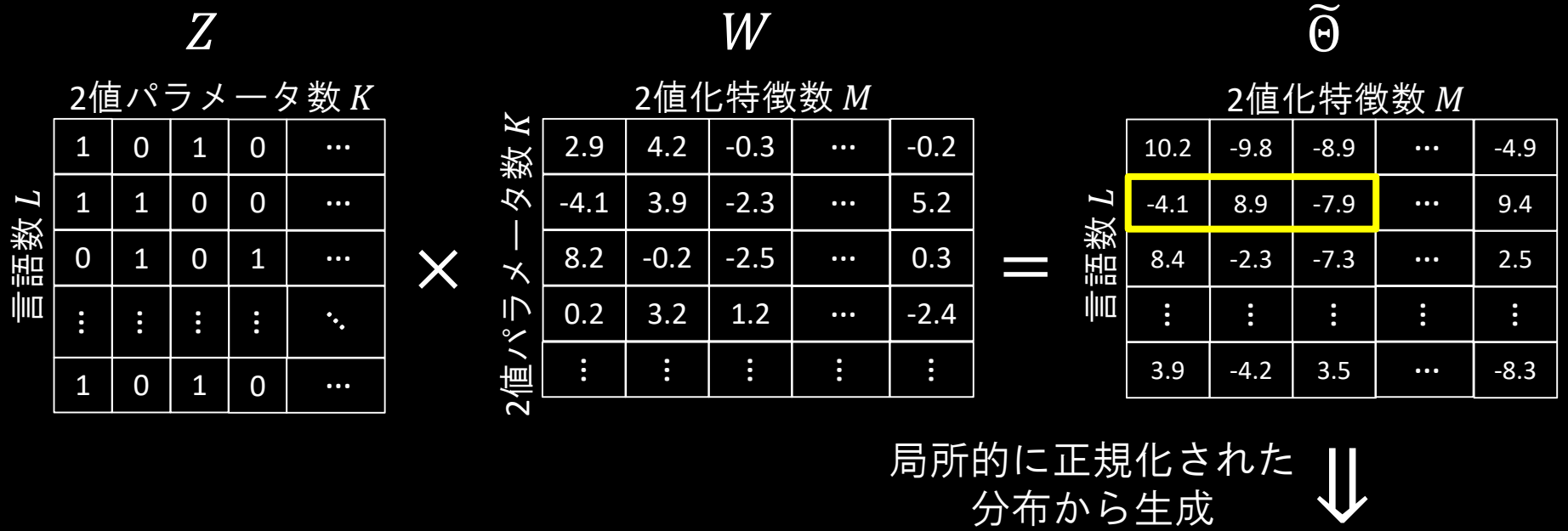
多値特徴数 N

言語数 L

1	2	…	2
2	1	…	3
1	4	…	3
⋮	⋮	⋮	⋮
3	1	…	1

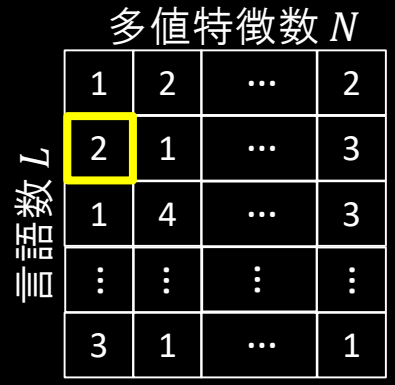
X

提案モデル: 確率的生成



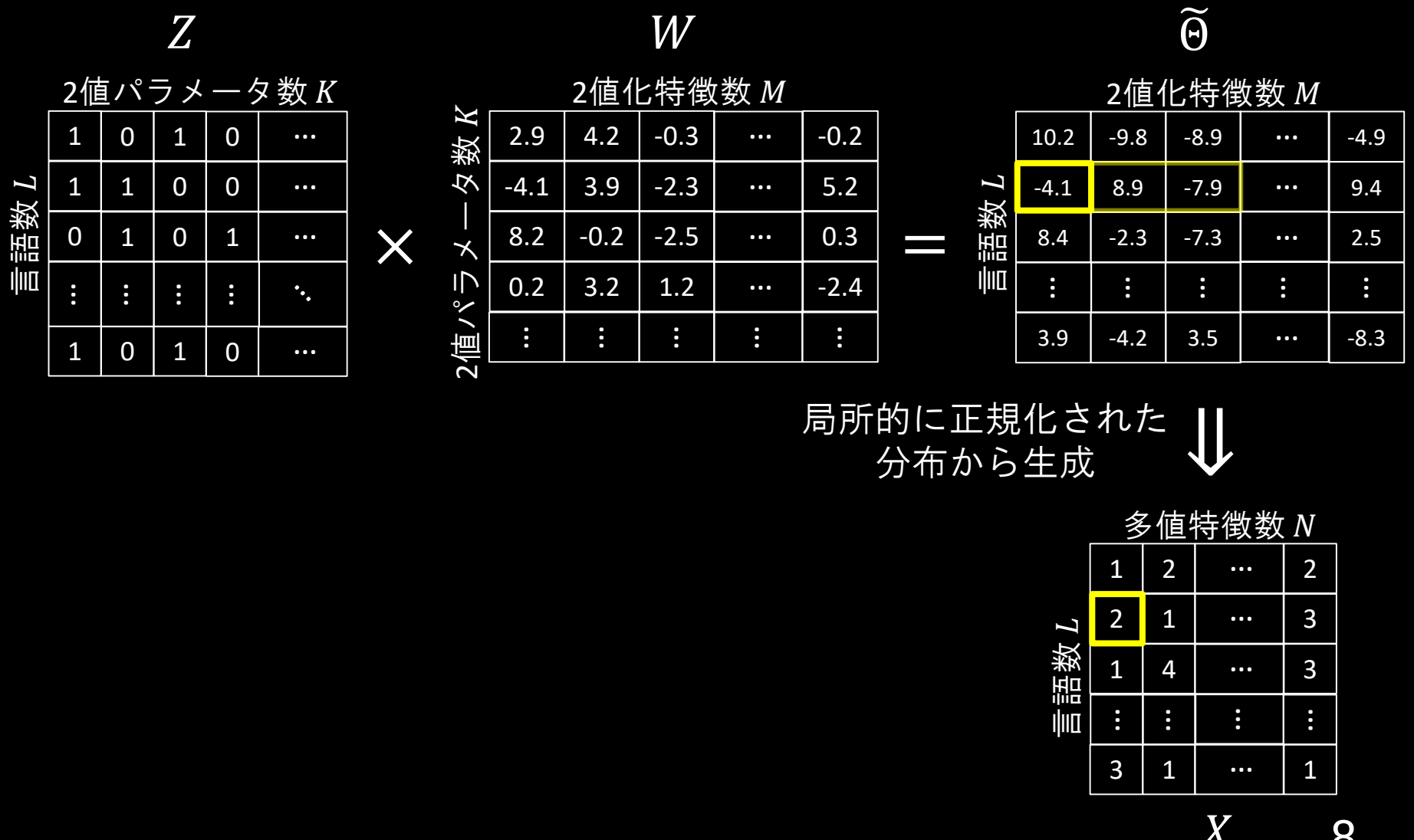
局所的に正規化された
分布から生成 ⇓

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

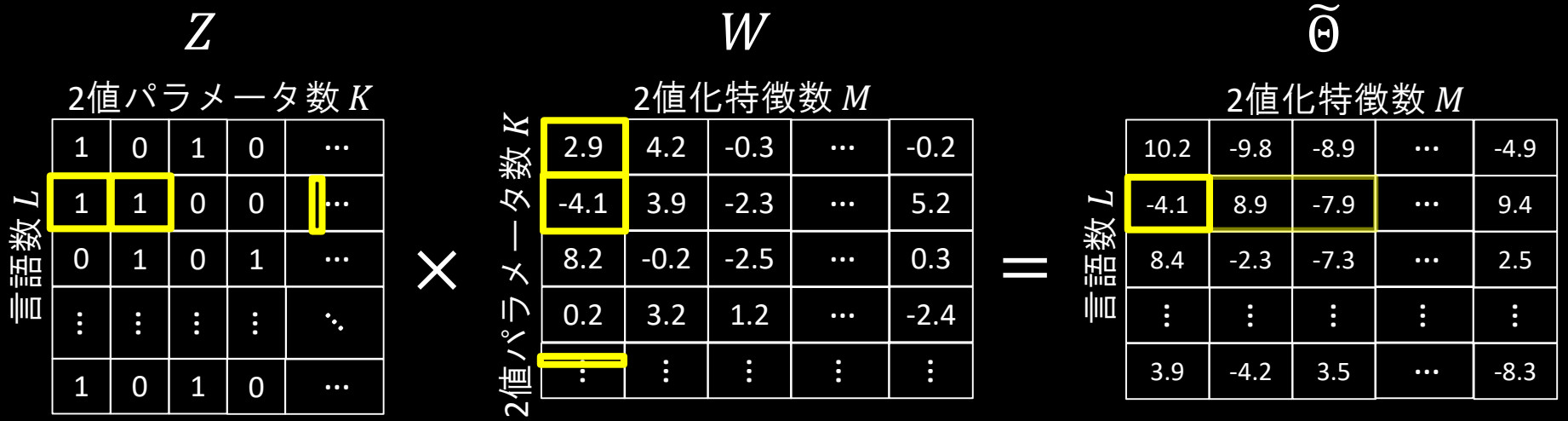


X

提案モデル: 確率的生成

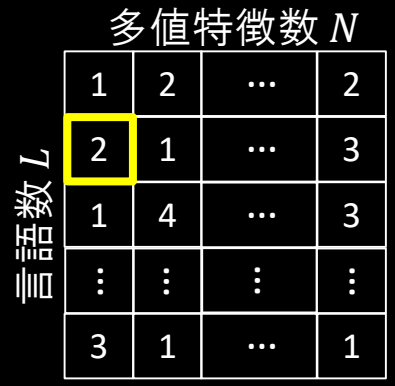

Z
2値パラメータ数 K
言語数 L
1 0 1 0 …
1 1 0 0 …
0 1 0 1 …
⋮ ⋮ ⋮ ⋮ ⋱
1 0 1 0 …
×
W
2値化特徴数 M
2値パラメータ数 K
2.9 4.2 -0.3 … -0.2
-4.1 3.9 -2.3 … 5.2
8.2 -0.2 -2.5 … 0.3
0.2 3.2 1.2 … -2.4
⋮ ⋮ ⋮ ⋮ ⋮
=
$\tilde{\Theta}$
2値化特徴数 M
言語数 L
10.2 -9.8 -8.9 … -4.9
-4.1 8.9 -7.9 … 9.4
8.4 -2.3 -7.3 … 2.5
⋮ ⋮ ⋮ ⋮ ⋮
3.9 -4.2 3.5 … -8.3
局所的に正規化された
分布から生成
⇓
多値特徴数 N
言語数 L
1 2 … 2
2 1 … 3
1 4 … 3
⋮ ⋮ ⋮ ⋮
3 1 … 1
X

提案モデル: 確率的生成

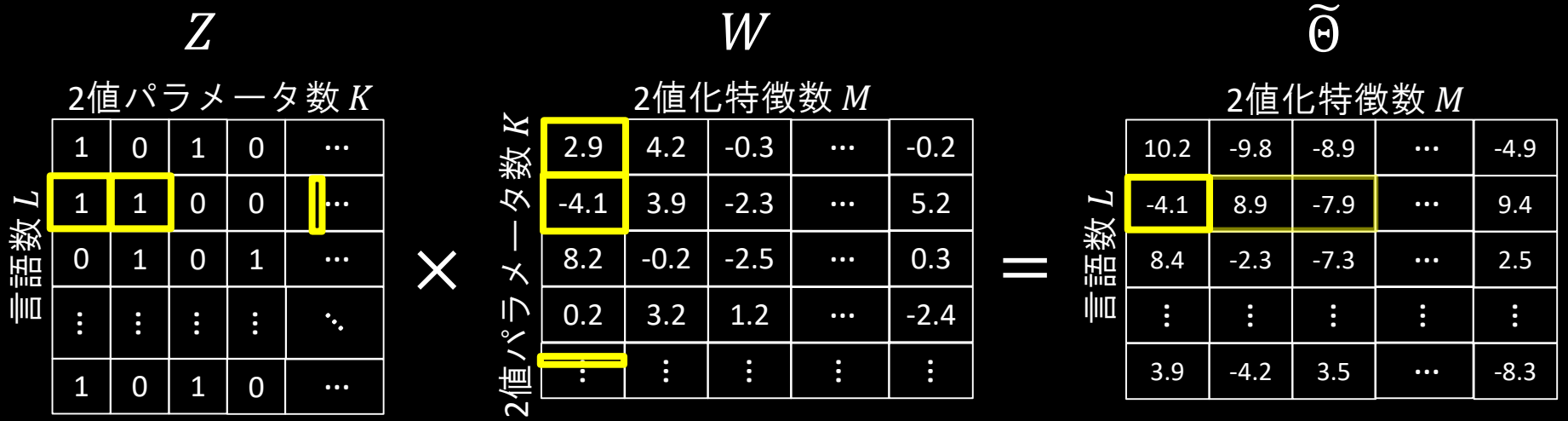


$$\tilde{\theta}_{2,1} = \sum_{k} z_{2,k} \times w_{k,1}$$

局所的に正規化された
分布から生成



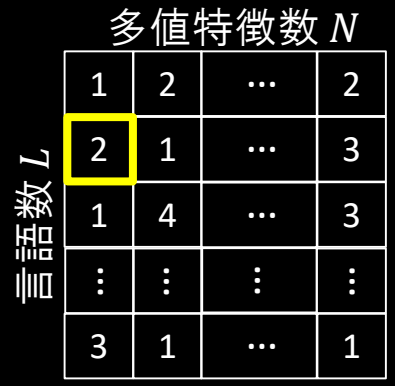
提案モデル: 確率的生成



局所的に正規化された
分布から生成 ⇓

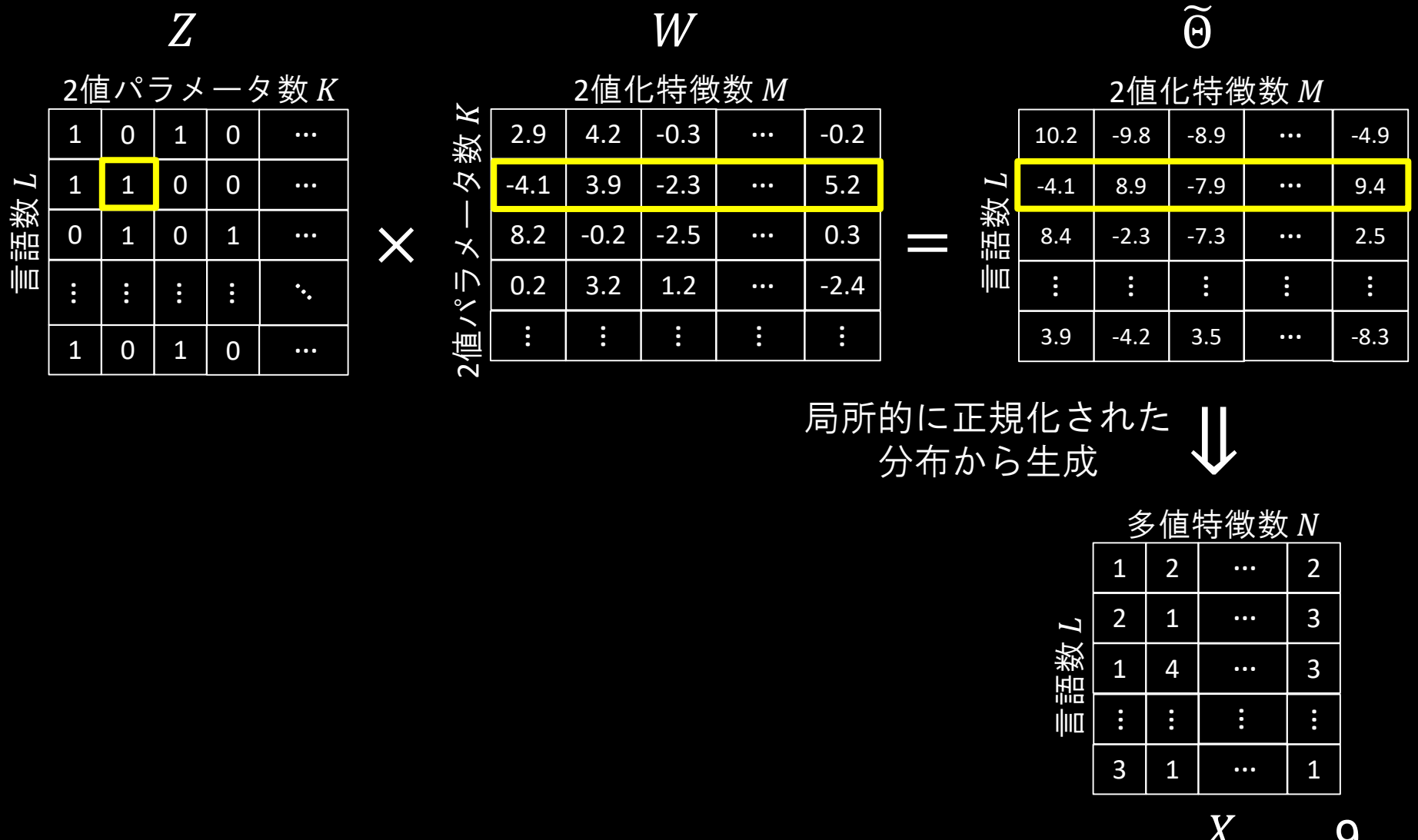
$$\tilde{\theta}_{2,1} = \sum_{k} z_{2,k} \times w_{k,1}$$

$z_{l,k} = 1$に対応するwのみ影響

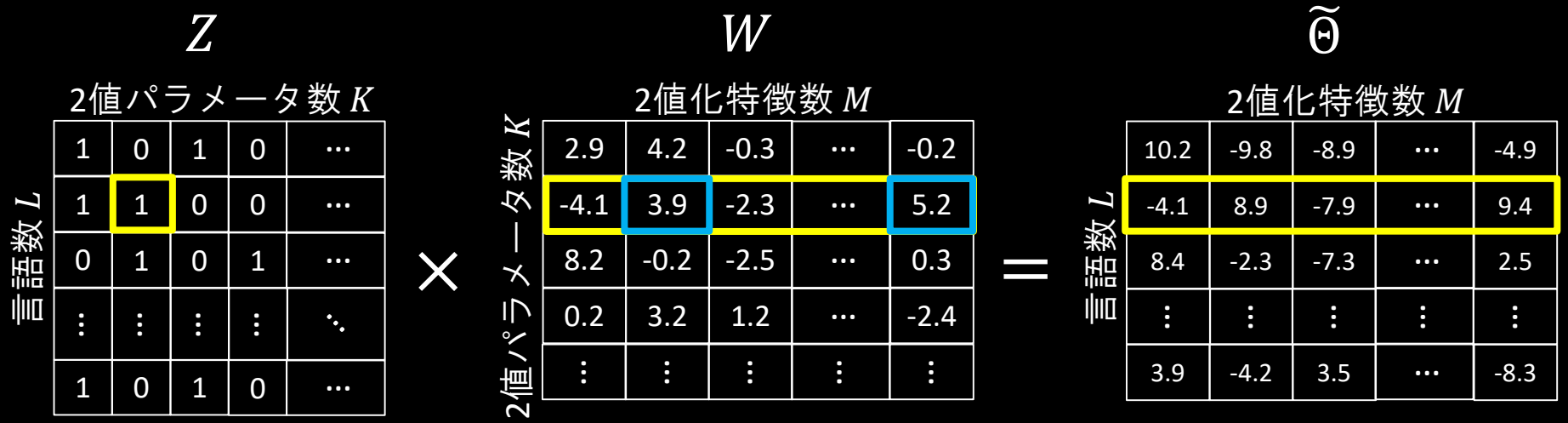


X

提案モデル: 確率的生成

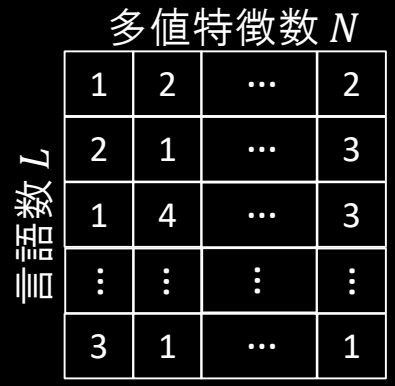

Z
2値パラメータ数 K
言語数 L
1 0 1 0 …
1 1 0 0 …
0 1 0 1 …
⋮ ⋮ ⋮ ⋮ ⋱
1 0 1 0 …
×
W
2値化特徴数 M
2値パラメータ数 K
2.9 4.2 -0.3 … -0.2
-4.1 3.9 -2.3 … 5.2
8.2 -0.2 -2.5 … 0.3
0.2 3.2 1.2 … -2.4
⋮ ⋮ ⋮ ⋮ ⋮
=
$\tilde{\Theta}$
2値化特徴数 M
言語数 L
10.2 -9.8 -8.9 … -4.9
-4.1 8.9 -7.9 … 9.4
8.4 -2.3 -7.3 … 2.5
⋮ ⋮ ⋮ ⋮ ⋮
3.9 -4.2 3.5 … -8.3
局所的に正規化された
分布から生成
⇓
多値特徴数 N
言語数 L
1 2 … 2
2 1 … 3
1 4 … 3
⋮ ⋮ ⋮ ⋮
3 1 … 1
X

提案モデル: 確率的生成



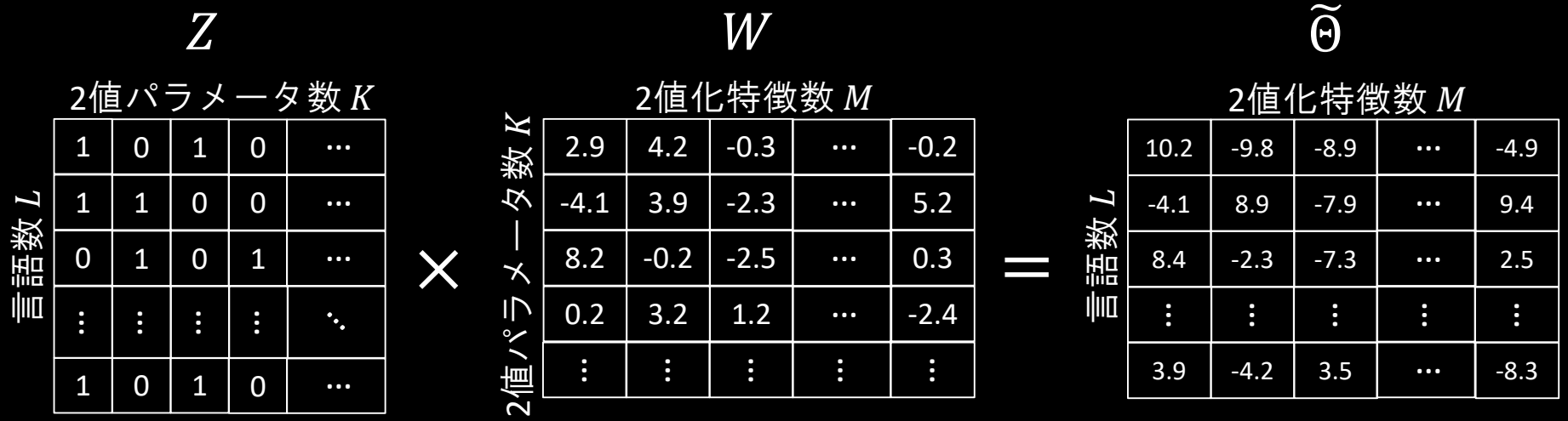
このパラメータで1をとる言語群では
ある特徴のある値と別の特徴のある
値を同時にとりやすい

局所的に正規化された
分布から生成 ⇓



X

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 ⇓

行列Zはインディアンビュッフェ過程 [Griffiths+, 2011] 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

Gibbsサンプリングによる推論

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 ⇓

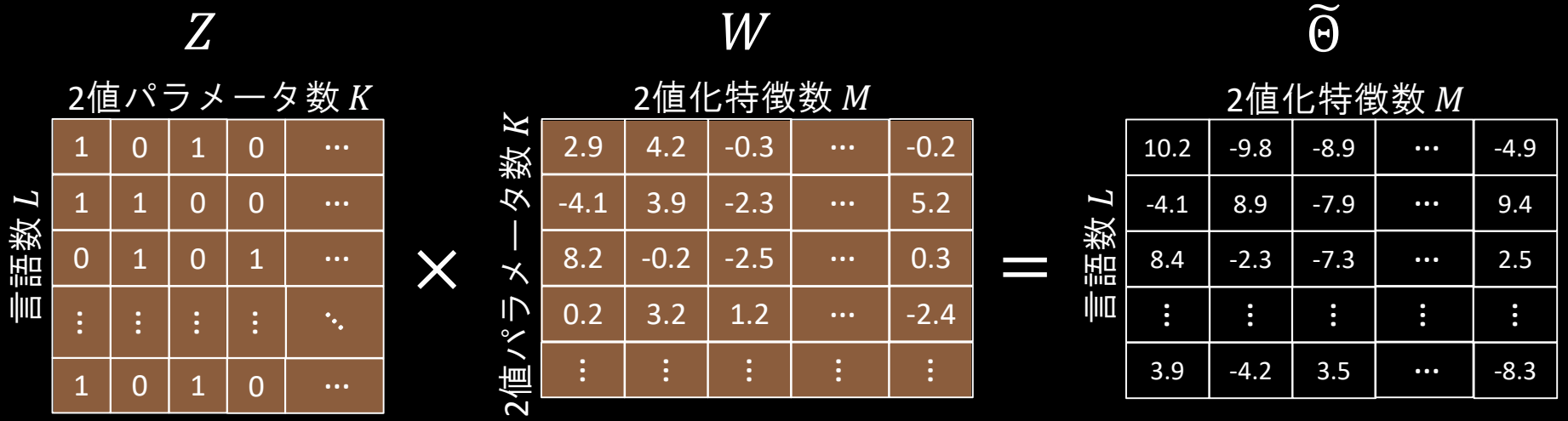
多値特徴数 N

言語数 L

1	2	…	2
2	1	…	3
1	4	…	3
⋮	⋮	⋮	⋮
3	1	…	1

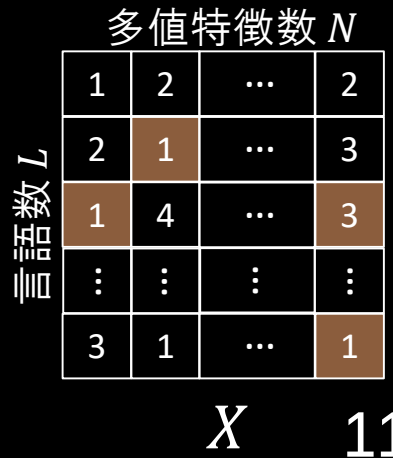
X

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 ⇓

- 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



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	

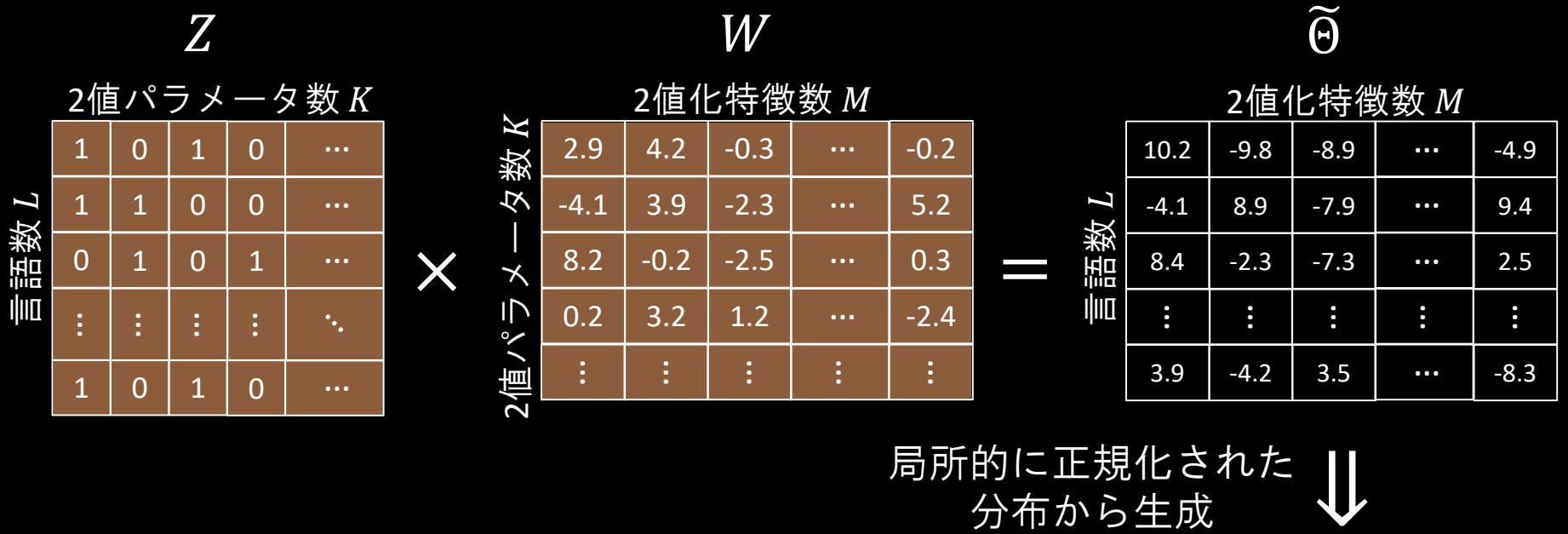
局所的に正規化された
分布から生成 ⇓

多値特徴数 N

言語数 L				
1	2	⋯	2	
2	1	⋯	3	
1	4	⋯	3	
⋮	⋮	⋮	⋮	
3	1	⋯	1	

X

Q: 獲得されたパラメータの妥当性?

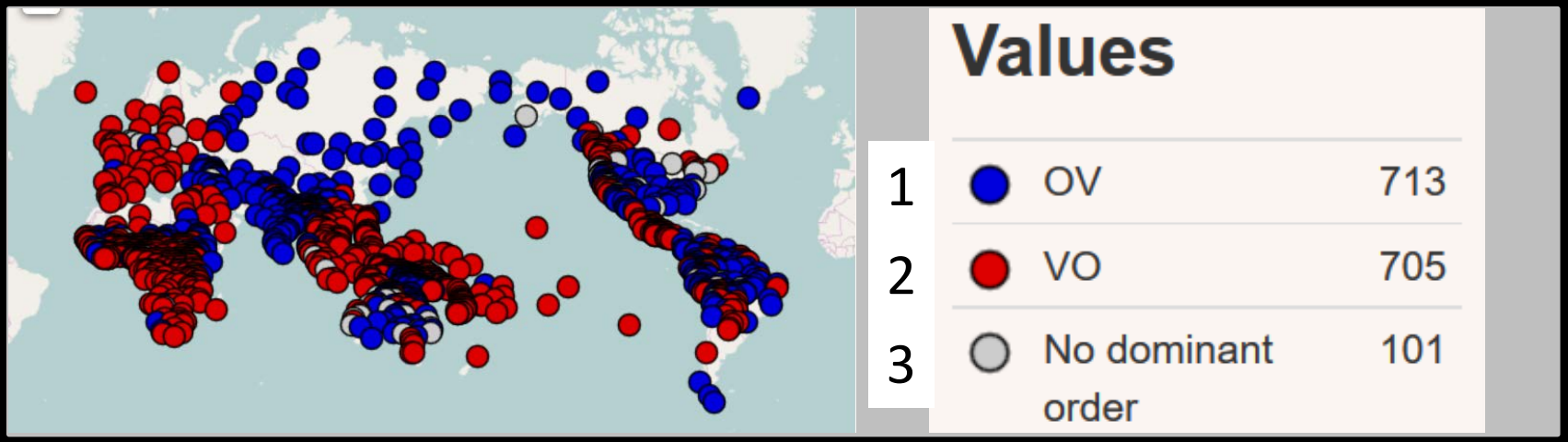


- パラメータそのものを定量評価するのは難しい
- 既知の特徴の一部を隠して欠損値推定
- 正しく復元できたなら、データに潜む規則性を捉えられていると言える

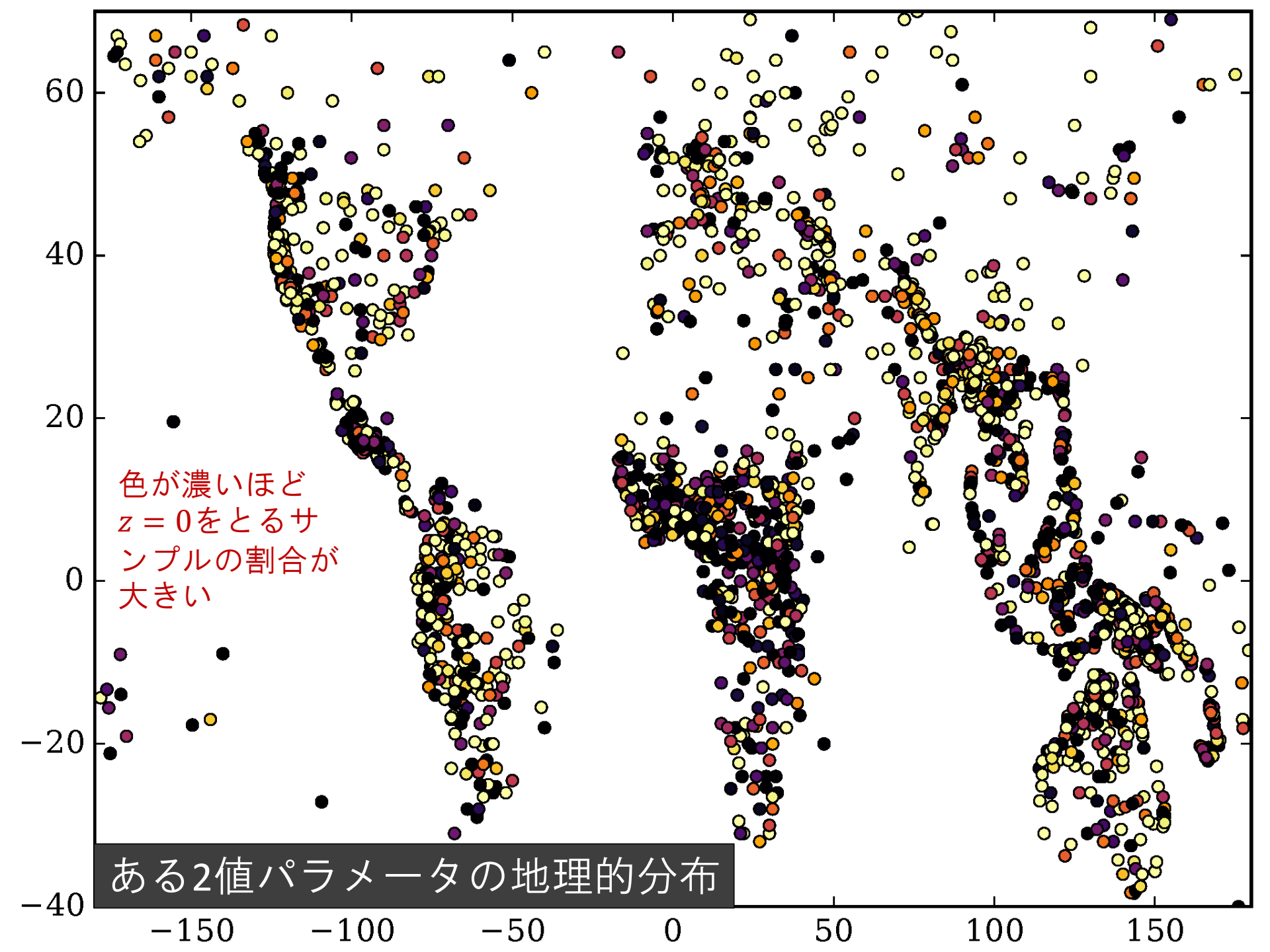
欠損値推定結果

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%

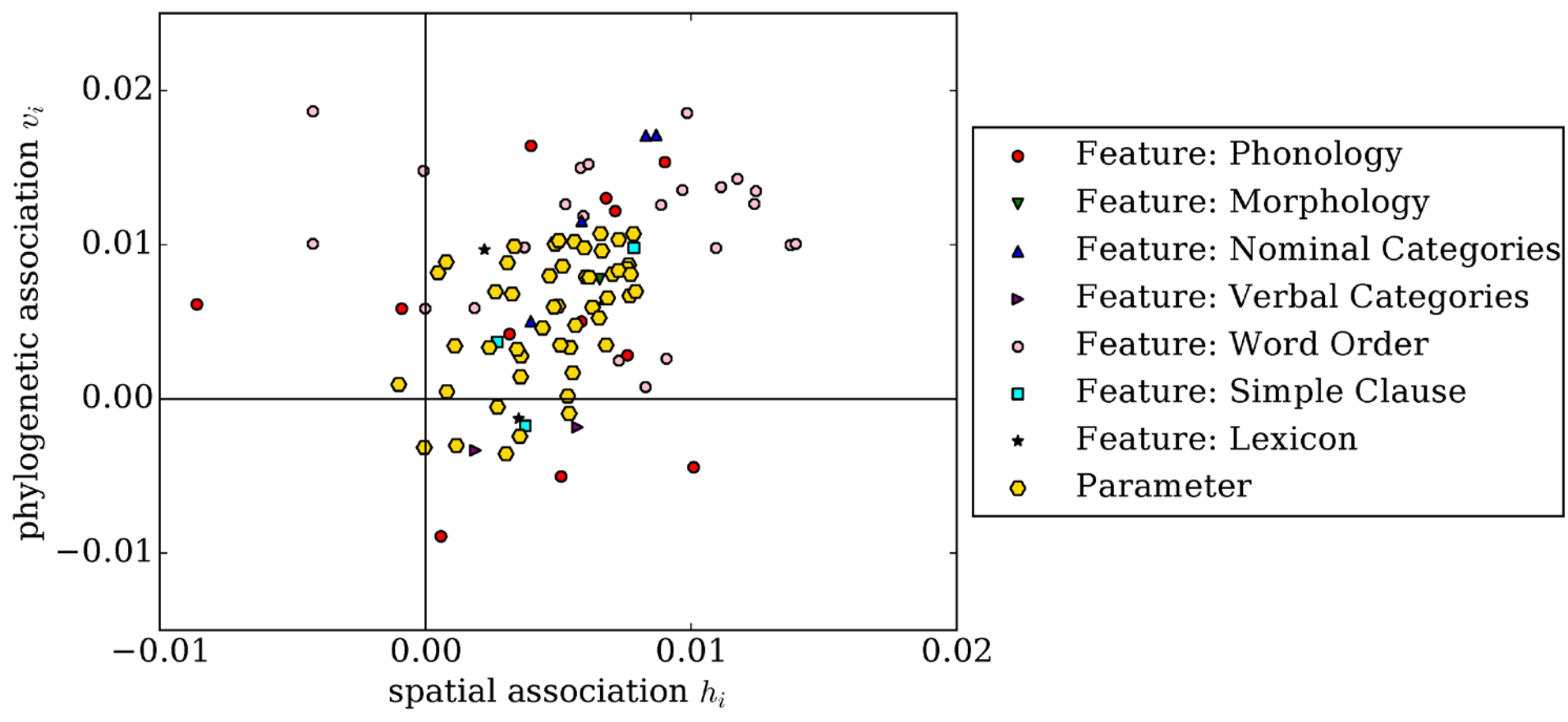
※予稿集から変更あり



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?


60
40
20
0
−20
−40
−150
−100
−50
0
50
100
150
色が濃いほど
$z = 0$をとるサンプルの割合が大きい
ある2値パラメータの地理的分布

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