Diachrony-aware Induction of Binary Latent Representations from Typological Features

Yugo Murawaki
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
murawaki@i.kyoto-u.ac.jp

Abstract

Although features of linguistic typology are a promising alternative to lexical evidence for tracing evolutionary history of languages, a large number of missing values in the dataset pose serious difficulties for statistical modeling. In this paper, we combine two existing approaches to the problem: (1) the synchronic approach that focuses on interdependencies between features and (2) the diachronic approach that exploits phylogenetically- and/or spatially-related languages. Specifically, we propose a Bayesian model that (1) represents each language as a sequence of binary latent *parameters* encoding inter-feature dependencies and (2) relates a language's parameters to those of its phylogenetic and spatial neighbors. Experiments show that the proposed model recovers missing values more accurately than others and that induced representations retain phylogenetic and spatial signals observed for surface features.

1 Introduction

Features of linguistic typology such as basic word order (examples are SVO and SOV) and the presence or absence of tone constitute a promising resource that can potentially be used to uncover the evolutionary history of languages. It has been argued that in exceptional cases, typological features can reflect a time span of 10,000 years or more (Nichols, 1994). Since typological features, by definition, allow us to compare an arbitrary pair of languages, they can be seen as the last hope for language isolates and tiny language families such as Ainu, Basque, and Japanese, for which lexicon-based historical-comparative linguistics[1] has failed to identify genetic relatives. Fortunately, the publication of a large typology database (Haspelmath et al., 2005) made it possible to take computational approaches to this area of study (Daumé III and Campbell, 2007).

Murawaki (2015) pursued a pipeline approach to utilizing typological features for phylogenetic inference. Exploiting interdependencies found among features, Murawaki (2015) first mapped each language, represented as a sequence of surface features, into a sequence of continuous latent components. It was in this continuous space that phylogenetic relations among languages were subsequently inferred. Murawaki (2015) argued that since the conversion and the resulting latent representations were designed to reflect typological naturalness, reconstructed ancestral languages were also likely to be typologically natural.

In this paper, however, we show that Murawaki (2015) rests on fragile underpinnings so that they need to be rebuilt. One of the most important problems underestimated by Murawaki (2015) is an alarmingly large number of missing values. The dataset is a matrix where languages are represented as rows and features as columns, but only less than 30% of the items are present after a modest preprocessing. What is worse, the situation is unlikely to change in the foreseeable future because of the thousands of languages in the world, there is ample documentation for only a handful. These missing values pose serious difficulties for statistical modeling. Ignoring uncertainty in data, however, Murawaki (2015) relied on point estimates of missing values provided by an existing method of imputation when inducing latent representations. In this paper, we take a Bayesian approach because it is known for its robustness in

[1] By lexicon-based historical-comparative linguistics, we mean broad topics including sound laws, cognates, and historical changes in inflectional paradigms.

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 451–461,
Taipei, Taiwan, November 27 – December 1, 2017

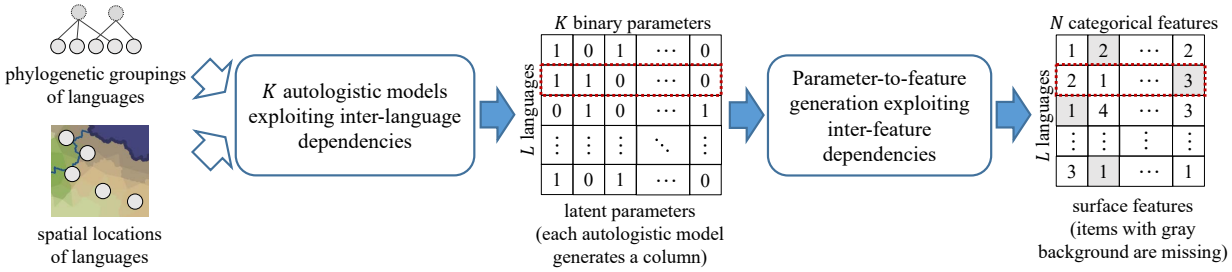

Figure 1: Overview of the proposed Bayesian generative model. Dotted boxes indicate the latent and surface representations of a language. Solid arrows show the direction of stochastic generation.

modeling uncertainties. We demonstrate that we can jointly infer missing values and latent representations.

Another question left unanswered is how good the induced representations are. In this paper, we present two quantitative analyses of the induced representations. The first one is rather indirect: we measure how well a model recovers missing values, with the assumption that good representations must capture regularity in surface features. We show that the proposed method outperformed the pipelined imputation method of Murawaki (2015) among others.

The second analysis involves geography. It is well known that the values of a surface feature do not distribute randomly in the world but reflect vertical (phylogenetic) transmissions from parents to children and horizontal (spatial or areal) transmissions between populations (Nichols, 1992). For example, languages of Mainland Southeast Asia are known for having similar tone systems even though they belong to different language families. To measure the degrees of the two modes of transmissions, we use an autologistic model that investigates dependencies among languages (Towner et al., 2012; Yamauchi and Murawaki, 2016). Since it requires the input to be discrete, we evaluate a new model that focuses on inter-feature dependencies in the same way as Murawaki (2015) but induces *binary* latent representations. We show that vertical and horizontal signals observed for surface features largely vanish from latent representations when only inter-feature dependencies are exploited. Although not directly applicable to the model of Murawaki (2015), our results suggest that the pipeline approach suffers from noise during phylogenetic inference. To address this problem, we extend the induction model to incorporate the autologistic model at the level of latent representations, rather than surface features. With this integrated model, we manage to let induced representations retain surface signals.

In the end, the Bayesian generative model we propose induces binary latent representations by combining inter-feature dependencies and inter-language dependencies, with primacy given to the former (Figure 1). Whereas inter-feature dependencies are synchronic in nature, inter-language dependencies reflect diachrony. Thus we call the integrated model diachrony-aware induction.

Due to space limitation, we had to put technical details into the supplementary material. However, we would like to stress that the proposed model works only if it is armed with statistical techniques rarely found in the NLP literature. Together with missing values and binary representations, a large number of continuous variables that connect binary representations to surface features need to be inferred. Unfortunately, a naïve Metropolis-Hastings algorithm does not converge within realistic time scales. We solve this problem by adopting Hamiltonian Monte Carlo (Neal, 2011) since it enables us to efficiently sample a large number of continuous variables at once. Likewise, the autologistic model contains an intractable normalization term, which prevents the application of the standard Metropolis-Hastings sampler. We use an approximate sampler instead (Liang, 2010).

2 Related Work

2.1 Inter-feature Dependencies

Interdependencies among features have long been observed across the world's languages. For example, `OV` (object-verb) languages tend to be `AN` for the order of adjective and noun. Greenberg (1963) proposed dozens of such patterns known as *linguistic universals*. A statistical model for discovering Greenbergian universals was presented by Daumé III and Campbell (2007). Itoh and Ueda (2004) used the Ising model to model the interac-

tion between features. Although these studies entirely focused on surface patterns, they imply the presence of some latent structure behind these surface features.

Some generative linguists argue for the existence of binary latent *parameters* behind surface features although they are controversial even among generative linguists (Boeckx, 2014). We borrow the term *parameter* from generative linguistics because the name of *feature* is reserved for surface variables.

Parameters are part of the *principles and parameters* (P&P) framework (Chomsky and Lasnik, 1993), where, the structure of a language is explained by (1) a set of universal principles that are common to all languages and (2) a set of parameters whose values vary among languages. Here we skip the former since our focus is on structural variability. According to P&P, if we set specific values to all the parameters, then we obtain a specific language. Each parameter is binary and, in general, sets the values of multiple surface features in a deterministic manner. For example, the head directionality parameter is either `head-initial` or `head-final`. If `head-initial` is chosen, then surface features are set to `VO`, `NA` and `Prepositions`; otherwise the language in question becomes `OV`, `AN` and `Postpositions` (Baker, 2002). Baker (2002) discussed a number of parameters such as head directionality, polysynthesis, and topic prominent parameters.

Partly inspired by the P&P framework, we use a sequence of binary variables as the latent representation of a language. However, there are non-negligible differences between P&P and ours, which are discussed in Section S.2 of the supplementary material.

What the structure behind surface features looks like is almost exclusively discussed by generative linguists, but it should be noted that they are not the only group who attempts to explain surface patterns. Roughly speaking, generative linguists are part of the *synchronist* camp, as contrasted with *diachronists*, who consider that at least some patterns observed in surface features arise from common paths of diachronic development (Anderson, 2016). An important factor of diachronic development is grammaticalization, by which content words change into function words (Heine and Kuteva, 2007). For example, the correlation between the order of adposition and noun and the order of genitive and noun might be explained by the fact that adpositions often derive from nouns.

2.2 Inter-language Dependencies

The standard model for phylogenetic inference is the tree model, where a trait is passed on from a parent to a child with occasional modifications. In fact, the recent success in the applications of statistical models to historical linguistic problems is largely attributed to the tree model (Gray and Atkinson, 2003; Bouckaert et al., 2012). In linguistic typology, however, a non-tree-like mode of evolution has emerged as one of the central topics (Trubetzkoy, 1928; Campbell, 2006). Typological features, like loanwords, can be borrowed from one language to another, and as a result, vertical (phylogenetic) signals are obscured by horizontal (spatial) transmission.

The task of incorporating both vertical and horizontal transmissions within a statistical model of evolution is notoriously challenging because of the excessive flexibility of horizontal transmissions. This is the reason why previously proposed models are coupled with some very strong assumptions, for example, that a reference tree is given a priori (Nelson-Sathi et al., 2010), and that horizontal transmissions can be modeled through time-invariant areal clusters (Daumé III, 2009).

Consequently, we pursue a line of research in linguistic typology that draws on information on the current distribution of typological features without explicitly requiring the reconstruction of previous states (Nichols, 1992, 1995; Parkvall, 2008; Wichmann and Holman, 2009). The basic assumption is that if the feature in question is vertically stable, then a phylogenetically defined group of languages will tend to share the same value. Similarly, if the feature in question is horizontally diffusible, then spatially close languages would be expected to frequently share the same feature value. Since the current distribution of typological features is more or less affected by these factors, we need to disentangle the effects of each of these factors. To do this, Yamauchi and Murawaki (2016) adopted a variant of the autologistic model, which had been widely used to model the spatial distribution of a feature (Besag, 1974; Towner et al., 2012). The model was also used to impute missing values because the phylogenetic and spatial neighbors of a language had some predictive power over its feature values.

3 Data and Preprocessing

The dataset we used in the present study is the online edition[2] of the *World Atlas of Language Structures* (WALS) (Haspelmath et al., 2005). While Greenberg (1963) and generative linguists have manually induced patterns and parameters, WALS makes it possible to take computational approaches to modeling features (Daumé III and Campbell, 2007; Daumé III, 2009; Murawaki, 2015; Takamura et al., 2016; Murawaki, 2016).

WALS is essentially a matrix where languages are represented as rows and features as columns. As of 2017, it contained 2,679 languages and 192 surface features. It covered less than 15% of items in the matrix, however.

We removed sign languages, pidgins and creoles from the matrix. We imputed some missing values that could trivially be inferred from other features. We then removed features that covered less than 10% of the languages. After the preprocessing, the number of languages L was 2,607 while the number of features N was reduced to 104. The coverage went up to 26.9%, but the rate was still alarmingly low.

In WALS, languages are accompanied by additional information. We used the following fields to model inter-language dependencies. (1) genera, the lower of the two-level phylogenetic groupings, and (2) single-point geographical coordinates (longitude and latitude). By connecting every pair of languages within a genus, we constructed a phylogenetic neighbor graph. A spatial neighbor graph was constructed by linking all language pairs that were located within a distance of $R = 1000$ km. On average, each language had 30.8 and 89.1 neighbors, respectively.

The features in WALS are categorical. For example, Feature 81A, "Order of Subject, Object and Verb" has seven possible values: `SOV`, `SVO`, `VSO`, `VOS`, `OVS`, `OSV` and `No dominant order`, and each language incorporates one of these seven values. For each language, we arranged its features into a sequence. A sequence of categorical features can alternatively be represented as a binary sequence using the 1-of-F_i coding scheme: Feature i with F_i possible values was converted into F_i binary items among which only one item takes 1. The number of binarized features M was 723.

[2] http://wals.info/

L	# of languages
K	# of parameters
M	# of binarized features
N	# of categorical features
$Z \in \{0,1\}^{L \times K}$	Binary parameter matrix
$W \in \mathbb{R}^{K \times M}$	Weight matrix
$\tilde{\Theta} \in \mathbb{R}^{L \times M}$	Feature score matrix
$\Theta \in [0,1]^{L \times M}$	Feature probability matrix
$X \in \mathbb{N}^{L \times N}$	Categorical feature matrix

Table 1: Notations.

4 Proposed Method

Since the proposed model is rather complicated, we present two key components before going into the integrated model. Table 1 shows notations used in this paper. Surface features have two ways of indexing. First, feature values are serialized as $(1,1), \cdots, (1,F_1), (2,1), \cdots, (i,j), \cdots, (N,F_N)$, where (i,j) points to feature i's j-th value. Then they are given the flat index $1, \cdots, m, \cdots, M$ ($M = \sum_{i=1}^{N} F_i$). Two indices are mapped by the function $f(i,j) = m$. We need the flat representation because that is what latent parameters work on. A parameter is expected to capture the relation between one feature's particular value (e.g., `VO` for the order of object and verb) and another feature's particular value (`NA` for the order of adjective and noun).

4.1 Inter-feature Dependencies

Figure 2 illustrates how surface features are generated from binary latent parameters. We use matrix factorization (Srebro et al., 2005; Griffiths and Ghahramani, 2011) to capture inter-feature dependencies. Since categorical feature matrix X cannot directly be decomposed into two, we first construct (unnormalized) feature score matrix $\tilde{\Theta}$ and then stochastically generate X using $\tilde{\Theta}$.

$\tilde{\Theta}$ is a product of binary parameter matrix Z and weight matrix W. The generation of Z will be described in Section 4.2.[3] Each item of $\tilde{\Theta}$, $\tilde{\theta}_{l,m}$, is a score for language l's m-th binarized feature. It is affected only by parameters with $z_{l,k} = 1$ because

$$\tilde{\theta}_{l,m} = \sum_{k=1}^{K} z_{l,k} w_{k,m}. \quad (1)$$

[3] Although the natural choice for modeling binary latent matrices is an Indian buffet process (IBP) (Griffiths and Ghahramani, 2011), we do not take this approach for reasons we explain in Section S.1 of the supplementary material.

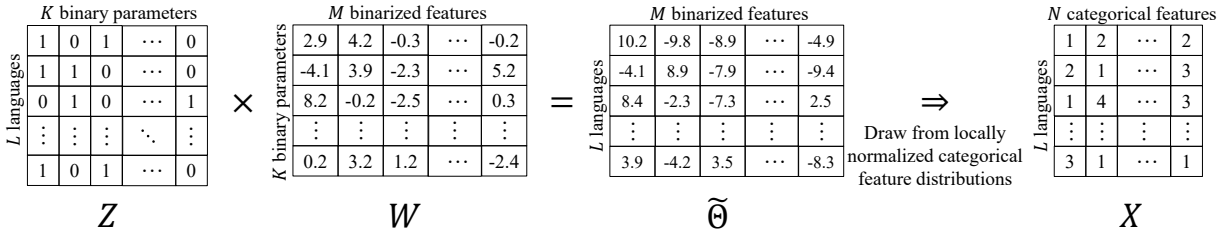

Figure 2: Stochastic parameter-to-feature generation. $\tilde{\Theta} = ZW$ encodes inter-feature dependencies.

We locally apply normalization to $\tilde{\Theta}$ to obtain Θ, in which $\theta_{l,i,j}$ is the probability of language l taking value j for categorical feature i

$$\theta_{l,i,j} = \frac{\exp(\tilde{\theta}_{l,f(i,j)})}{\sum_{j'} \exp(\tilde{\theta}_{l,f(i,j')})}. \tag{2}$$

Finally, language l's i-th categorical feature, $x_{l,i}$, is generated from this distribution.

$$P(x_{l,i} \mid \mathrm{z}_{l,*}, W) = \theta_{l,i,x_{l,i}}, \tag{3}$$

where $\mathrm{z}_{l,*} = (z_{l,1}, \cdots, z_{l,K})$.

Combining Eqs. (1) and (2), we obtain

$$\begin{aligned} \theta_{l,i,j} &\propto \exp(\sum_{k=1}^{K} z_{l,k} w_{k,f(i,j)}) \\ &= \prod_{k=1}^{K} \exp(z_{l,k} w_{k,f(i,j)}). \end{aligned} \tag{4}$$

We can see from Eq. (4) that this is a product-of-experts model (Hinton, 2002). If $z_{l,k} = 0$, parameter k has no effect on $\theta_{l,i,j}$ because $\exp(z_{l,k} w_{k,f(i,j)}) = 1$. Otherwise, if $w_{k,f(i,j)} > 0$, it makes $\theta_{l,i,j}$ larger, and if $w_{k,f(i,j)} < 0$, it lowers $\theta_{l,i,j}$.

Suppose that for parameter k, a certain group of languages takes $z_{l,k} = 1$. If two categorical feature values (i_1, j_1) and (i_2, j_2) have positive weights (i.e., $w_{k,f(i_1,j_1)} > 0$ and $w_{k,f(i_2,j_2)} > 0$), the pair must often co-occur in these languages. Likewise, the fact that two feature values do not co-occur can be encoded as a positive weight for one value and a negative weight for the other.

4.2 Inter-language Dependencies

The autologistic model is used to generate each column of Z, $\mathrm{z}_{*,k} = (z_{1,k}, \cdots, z_{L,k})$. To construct the model, we use two neighbor graphs and the corresponding three counting functions, as illustrated in Figure 3. $V(\mathrm{z}_{*,k})$ returns the number of pairs sharing the same value in the phylogenetic neighbor graph, and $H(\mathrm{z}_{*,k})$ is the spatial equivalent of $V(\mathrm{z}_{*,k})$. $U(\mathrm{z}_{*,k})$ gives the number of languages that take the value 1.

We now introduce the following variables: vertical stability $v_k > 0$, horizontal diffusibility $h_k > 0$, and universality $-\infty < u_k < \infty$ for each feature k. Then the probability of $\mathrm{z}_{*,k}$ conditioned on v_k, h_k and u_k is given as

$$P(\mathrm{z}_{*,k} \mid v_k, h_k, u_k) = \frac{\exp\Big(v_k V(\mathrm{z}_{*,k}) + h_k H(\mathrm{z}_{*,k}) + u_k U(\mathrm{z}_{*,k})\Big)}{\sum_{\mathrm{z}'_{*,k}} \exp\Big(v_k V(\mathrm{z}'_{*,k}) + h_k H(\mathrm{z}'_{*,k}) + u_k U(\mathrm{z}'_{*,k})\Big)}.$$

The denominator is a normalization term, ensuring that the sum of the distribution equals one.

The autologistic model can be interpreted in terms of the competition associated with possible assignments of $\mathrm{z}_{*,k}$ for the probability mass 1. If a given value, $\mathrm{z}_{*,k}$, has a relatively large $V(\mathrm{z}_{*,k})$, then setting a large value for v_k enables it to appropriate fractions of the mass from its weaker rivals. However, if too large a value is set for v_k, then it will be overwhelmed by its stronger rivals.

To acquire further insights into the model, let us consider the probability of language l taking value $b \in \{0, 1\}$, conditioned on the rest of the languages, $\mathrm{z}_{-l,k}$:

$$\begin{aligned} &P(z_{l,k} = b \mid \mathrm{z}_{-l,k}, v_k, h_k, u_k) \propto \\ &\exp\left(v_k V_{l,k,b} + h_k H_{l,k,b} + u_k b\right), \end{aligned} \tag{5}$$

where $V_{l,k,b}$ is the number of language l's phylogenetic neighbors that assume value b, and $H_{l,k,b}$ is its spatial counterpart. $P(z_{l,k} = b \mid \mathrm{z}_{-l,k}, v_k, h_k, u_k)$ is expressed by the weighted linear combination of the three factors in the log-space. It will increase with a rise in the number of phylogenetic neighbors that assume value b. However, this probability depends not only on the phy-

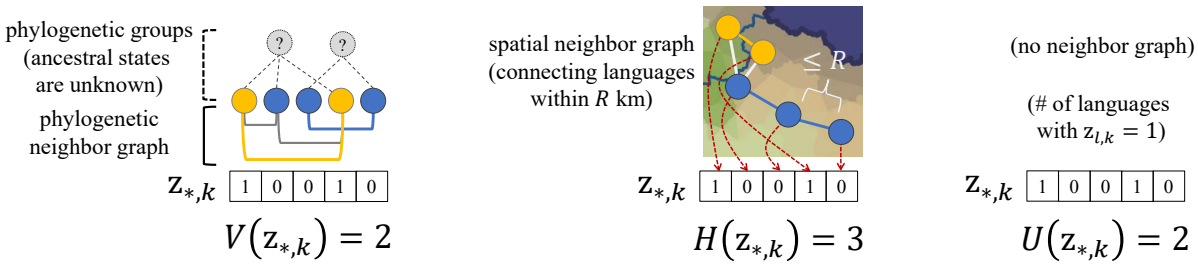

Figure 3: Neighbor graphs and counting functions used to encode inter-language dependencies.

logenetic neighbors of language l, but it also depends on its spatial neighbors and on universality. How strongly these factors affect the stochastic selection is controlled by v_k, h_k, and u_k.

4.3 Integrated Model

Now we complete the generative model by integrating the two types of dependencies. The joint distribution is defined as

$$P(A, Z, W, X) = P(A)P(Z|A)P(W)P(X|Z, W),$$

where hyperparameters are omitted for brevity and A is a set of latent variables that control the generation of Z:

$$P(A) = \prod_{k=1}^{K} P(v_k)P(h_k)P(u_k).$$

Their prior distributions are: $v_k \sim \mathrm{Gamma}(\kappa, \theta)$, $h_k \sim \mathrm{Gamma}(\kappa, \theta)$, and $u_k \sim \mathcal{N}(0, \sigma^2)$.[4]

Next, $z_{*,k}$'s are generated as described in Section 4.2:

$$P(Z \mid A) = \prod_{k=1}^{K} P(z_{*,k} \mid v_k, h_k, u_k).$$

The generation of Z is followed by that of the corresponding weight matrix $W \in \mathbb{R}^{K \times M}$, and then we obtain the feature score matrix $\tilde{\Theta} = ZW$. Each item of W, $w_{k,m}$, is generated from Student's t-distribution with 1 degree of freedom. We choose this distribution for two reasons. First, it has heavier tails than the Gaussian distribution and allows some weights to fall far from 0. Second, our inference algorithm demands that the negative logarithm of the probability density function be differentiable (see Section S.4 for details).

[4] In the experiments, we set shape $\kappa = 1$, scale $\theta = 1$, and standard deviation $\sigma = 10$. These priors were not non-informative, but they were sufficiently gentle in the regions where these parameters typically resided.

The t-distribution satisfies the condition while the Laplace distribution does not.

Finally, X is generated using $\tilde{\Theta} = ZW$, as described in Section 4.1:

$$P(X \mid Z, W) = \prod_{l=1}^{L} \prod_{i=1}^{N} P(x_{l,i} \mid z_{l,*}, W).$$

4.4 Inference

As usual, we use Gibbs sampling to perform posterior inference. Given observed values $x_{l,i}$, we iteratively update $z_{l,k}$, v_k, h_k, u_k, and $w_{k,*}$ as well as missing values $x_{l,i}$.

Update $x_{l,i}$. $x_{l,i}$ is sampled from Eq. (3).

Update $z_{l,k}$. The posterior probability $P(z_{l,k} \mid -)$ is proportional to Eq. (5) times the product of Eq. (3) for all feature i's of language l.

Update v_k, h_k and u_k. We want to sample v_k (and h_k and u_k) from $P(v_k \mid -) \propto P(v_k)P(z_{*,k} \mid v_k, h_k, u_k)$. This belongs to a class of problems known as sampling from doubly-intractable distributions (Møller et al., 2006; Murray et al., 2006). While it remains a challenging problem in statistics, it is not difficult to approximately sample the variables if we give up theoretical rigorousness (Liang, 2010). The details of the algorithm we use can be found in Section S.3 of the supplementary material.

Update $w_{k,*}$. The remaining problem is how to update $w_{k,m}$. Since the number of weights is very large ($K \times M$), the simple Metropolis-Hastings algorithm (Görür et al., 2006; Doyle et al., 2014) is not a workable option. To address this problem, we block-sample $w_{k,*} = (w_{k,1}, \cdots, w_{k,M})$ using Hamiltonian Monte Carlo (HMC) (Neal, 2011). A sketch of the algorithm can be found in Section S.4 of the supplementary material.

5 Experiments

5.1 Missing Value Imputation

We indirectly evaluated the proposed model, called SYNDIA, by means of missing value imputation. If it predicts missing feature values better than reasonable baselines, we can say that the induced parameters are justified. Although no ground truth exists for the missing portion of the dataset, missing value imputation can be evaluated by hiding some observed values and verifying the effectiveness of their recovery. We conducted a 10-fold cross-validation.

We ran SYNDIA with two different settings: $K = 50$ and 100. We performed posterior inference for 500 iterations. After that, we collected 100 samples of $x_{l,i}$ for each language, one per iteration. For each missing value $x_{l,i}$, we output the most frequent value among the 100 samples. The HMC parameters ϵ and S were set to 0.05 and 10, respectively.

We applied simulated annealing to the sampling of $z_{l,k}$. For the first 100 iterations, the inverse temperature was increased from 0.1 to 1.0.

We compared SYNDIA with several baselines.

MFV For each categorical feature i, always output the most frequent value among observed $x_{l,i}$.

Surface-DIA An autologistic model applied to surface features (Yamauchi and Murawaki, 2016). The details of the model are presented in Section S.5 of the supplementary material.

DPMPM A Dirichlet process mixture of multinomial distributions with a truncated stick-breaking construction (Si and Reiter, 2013) used by Blasi et al. (2017). It assigns a single categorical latent variable to each language. As an implementation, we used the `R` package *NPBayesImpute*.

MCA A variant of multiple correspondence analysis (Josse et al., 2012) used by Murawaki (2015). We used the `imputeMCA` function of the `R` package *missMDA*.

SYN A simplified version of SYNDIA, with v_k and h_k removed from the model. See Section S.6 of the supplementary material for details.

MFV and Surface-DIA can be seen as the models of inter-language dependencies while DPMPM, MCA and SYN are these of inter-feature dependencies.

Table 2 shows the result. We can see that SYNDIA with $K = 50$ performed the best.

Type	Model	Accuracy
Lang.	MFV	60.95%
	Surface-DIA	66.22%
Feat.	DPMPM ($K^* = 50$)	69.08%
	MCA	69.88%
	SYN ($K = 50$)	73.83%
	SYN ($K = 100$)	72.87%
Both	SYNDIA ($K = 50$)	**74.46%**
	SYNDIA ($K = 100$)	74.00%

Table 2: Accuracy of missing value imputation. The first column indicates the types of dependencies the models exploit: inter-language dependencies, inter-feature dependencies and both.

Model	Accuracy
Full model (SYNDIA)	74.46%
-vertical	73.89%
-horizontal	**74.47%**
-vertical -horizontal (SYN)	73.83%

Table 3: Ablation experiments for missing value imputation. $K = 50$.

Smaller K yielded higher accuracy although the likelihood $P(X \mid Z, W)$ went up as K increased. Due to the high ratio of missing values, the model might have overfitted the data with larger K.

The fact that SYN outperformed Surface-DIA suggests that inter-feature dependencies have more predictive power than inter-language dependencies in the dataset. However, they are complimentary in nature as SYNDIA outperformed SYN.

We can confirm the limited expressive power of single categorical latent variables because DPMPM performed poorly even if a small value was set to the truncation level K^* to avoid overfitting. MCA employs more expressive representations of a sequence of continuous variables for each language. It slightly outperformed DPMPM but was beaten by SYN by a large margin. We conjecture that MCA was more sensitive to initialization than the Bayesian model armed with MCMC sampling. In any case, this result indicates that the latent representations Murawaki (2015) obtained were of poorer quality than those of SYN, not to mention those of SYNDIA.

We also conducted ablation experiments by removing either v_k or h_k from the model. The result is shown in Table 3. It turned out that the horizontal factor had stronger predictive power than the vertical factor, which has a negative implication on typology-based phylogenetic inference.

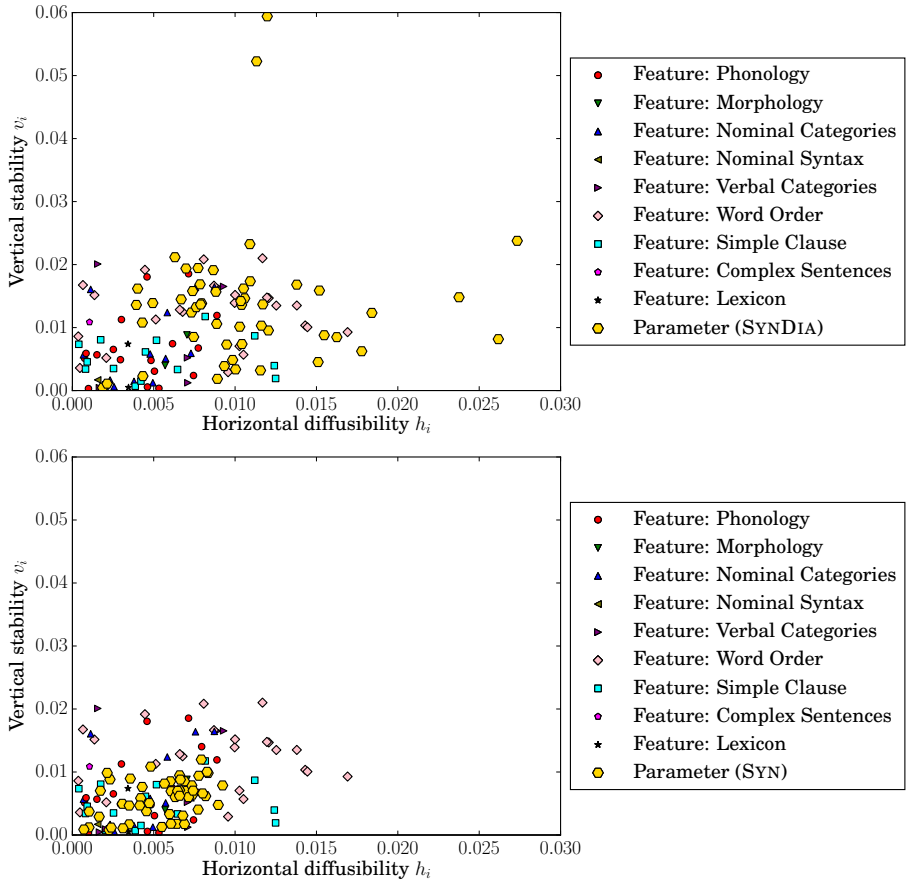

Figure 4: Scatter plots of surface features and induced parameters, with vertical stability v_i (v_k) as the y-axis and horizontal diffusibility h_i (h_k) as the x-axis. Larger v_i (h_i) indicates that feature i is more stable (diffusible). Comparing the absolute values of a v_i and an h_i makes no sense because they are tied with different neighbor graphs. Features are classified into 9 broad categories (called *Area* in WALS). v_k (and h_k) is the geometric mean of the 100 samples. The induction models are SYNDIA (Top) and SYN (Bottom). For both models, $K = 50$.

5.2 Vertical and Horizontal Signals

Hereafter we use all observed features to perform posterior inference. We examined how vertically stable and horizontally diffusible the induced parameters were. For SYNDIA, we simply extracted v_k and h_k from posterior samples. For comparison, we used Surface-DIA to estimate vertical stability and horizontal diffusibility of surface features. The same autologistic model was used to estimate v_k and h_k of SYN *after* the posterior inference. For details, see Sections S.5 and S.6.2 of the supplementary material.

Figure 4 summarizes the results. We can see that the most vertically stable latent parameters of SYNDIA are comparable to the most vertically stable surface features. The same holds for the most horizontally diffusible ones. Thus we can conclude that the induced representations retain vertical and horizontal signals observed for surface features.

On the other hand, SYN halved vertical stability and horizontal diffusibility when transforming surface features into latent parameters. A plausible explanation of this failure is that for many scarcely documented languages, we simply did not have enough observed surface features to determine their latent representations only from inter-feature dependencies. Due to the inherent uncertainty, $z_{l,k}$ swung between 0 and 1 during posterior inference, regardless of the states of their neighbors. As a result, these languages seem to have blocked vertical and horizontal signals. By contrast, SYNDIA appears to have flipped $z_{l,k}$ without disrupting inter-language dependencies when there were.

In summary, the experimental results have neg-

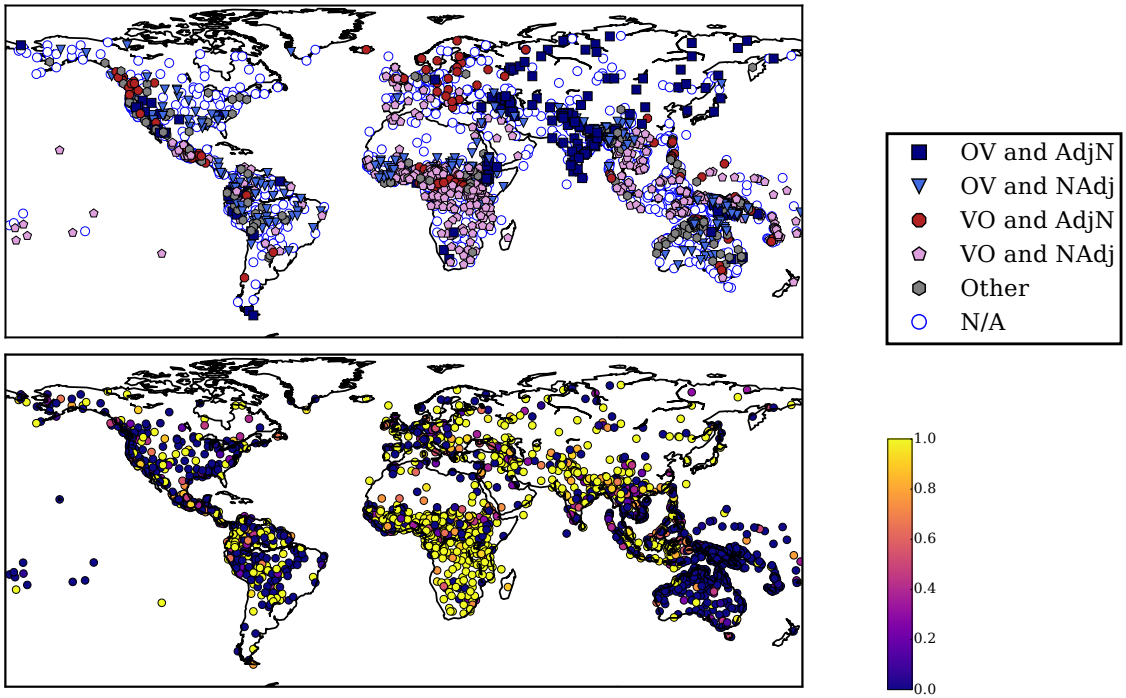

Figure 5: A comparison of a surface feature and a latent parameter in terms of geographical distribution. Each point denotes a language. (Top) Feature 97A, "Relationship between the Order of Object and Verb and the Order of Adjective and Noun." Missing values are denoted as N/A. (Bottom) A parameter of SYNDIA with $K_0 = 50$. Lighter nodes indicate higher frequencies of $z_{l,k} = 1$ among 100 samples.

ative implications for the pipeline approach pursued by Murawaki (2015), where the inter-feature dependency-based induction of latent representations is followed by phylogenetic inference. Fortunately, evidence presented up to this point suggests that it can be readily replaced with the proposed model.

5.3 Discussion

Figure 5 compares a latent parameter of SYNDIA with a surface feature on the world map. Some surface features show several geographic clusters of large size, telling something about the evolutionary history of languages. Even with a large number of missing values, SYNDIA yielded comparable geographic clusters for some parameters. Some geographic clusters were also produced by SYN, especially when the estimation of $z_{l,k}$ was stable. In our subjective evaluation, SYNDIA appeared to show clearer patterns than SYN. Needless to say, not all surface features were associated with clear geographic patterns, and not all latent parameters were. Overall, the results shed a positive light on the applicability of the induced representations to phylogenetic inference.

We also checked the weight matrix W (Figure S.2). It is not easy to analyze qualitatively but it deserves future investigation.

6 Conclusion

In this paper, we presented a Bayesian model that induces binary latent parameters from surface features of linguistic typology. We combined inter-language dependencies with inter-feature dependencies to obtain the latent representations of better quality. Gathering various statistical techniques, we managed to create the complex but workable model. The source code is publicly available at https://github.com/murawaki/latent-typology.

We pointed out that typology-based phylogenetic inference proposed by Murawaki (2015) had weak foundations, and we rebuilt them from scratch. The whole long paper was needed to do so, but our ultimate goal is the same as the one stated by Murawaki (2015). In the future, we would like to utilize the new latent representations to uncover the evolutionary history of languages.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 26730122.

References

Stephen R. Anderson. 2016. Synchronic versus diachronic explanation and the nature of the language faculty. *Annual Review of Linguistics*, 2:1–425.

Mark C. Baker. 2002. *The Atoms of Language: The Mind's Hidden Rules of Grammar*. Basic Books.

Julian Besag. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Damián E. Blasi, Susanne Maria Michaelis, and Martin Haspelmath. 2017. Grammars are robustly transmitted even during the emergence of creole languages. *Nature Human Behaviour*.

Cedric Boeckx. 2014. What principles and parameters got wrong. In M. Carme Picallo, editor, *Treebanks: Building and Using Parsed Corpora*, pages 155–178. Oxford University Press.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

Lyle Campbell. 2006. Areal linguistics. In *Encyclopedia of Language and Linguistics, Second Edition*, pages 454–460. Elsevier.

Noam Chomsky and Howard Lasnik. 1993. The theory of principles and parameters. In Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld, and Theo Vennemann, editors, *Syntax: An International Handbook of Contemporary Research*, volume 1, pages 506–569. De Gruyter.

Hal Daumé III. 2009. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 593–601.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72.

Gabriel Doyle, Klinton Bicknell, and Roger Levy. 2014. Nonparametric learning of phonological constraints in optimality theory. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1103.

Dilan Görür, Frank Jäkel, and Carl Edward Rasmussen. 2006. A choice model with infinitely many latent features. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 361–368.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Joseph H. Greenberg, editor. 1963. *Universals of language*. MIT Press.

Thomas L. Griffiths and Zoubin Ghahramani. 2011. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224.

Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press.

Bernd Heine and Tania Kuteva. 2007. *The Genesis of Grammar: A Reconstruction*. Oxford University Press.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

Yoshiaki Itoh and Sumie Ueda. 2004. The ising model for changes in word ordering rules in natural languages. *Physica D: Nonlinear Phenomena*, 198(3):333–339.

Julie Josse, Marie Chavent, Benot Liquet, and François Husson. 2012. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, 29(1):91–116.

Faming Liang. 2010. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022.

Jesper Møller, Anthony N. Pettitt, R. Reeves, and Kasper K. Berthelsen. 2006. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Yugo Murawaki. 2015. Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 324–334.

Yugo Murawaki. 2016. Statistical modeling of creole genesis. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. 2006. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366.

Radford M. Neal. 2011. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. CRC Press.

Shijulal Nelson-Sathi, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. 2010. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B: Biological Sciences*.

Johanna Nichols. 1992. *Linguistic Diversity in Space and Time*. University of Chicago Press.

Johanna Nichols. 1994. The spread of language around the Pacific rim. *Evolutionary Anthropology: Issues, News, and Reviews*, 3(6):206–215.

Johanna Nichols. 1995. Diachronically stable structural features. In Henning Andersen, editor, *Historical Linguistics 1993. Selected Papers from the 11th International Conference on Historical Linguistics, Los Angeles 16–20 August 1993*. John Benjamins Publishing Company.

Mikael Parkvall. 2008. Which parts of language are the most stable? *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, 61(3):234–250.

Yajuan Si and Jerome P. Reiter. 2013. Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38(5):499–521.

Nathan Srebro, Jason D. M. Rennie, and Tommi S. Jaakkola. 2005. Maximum-margin matrix factorization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pages 1329–1336.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 69–76.

Mary C. Towner, Mark N. Grote, Jay Venti, and Monique Borgerhoff Mulder. 2012. Cultural macroevolution on neighbor graphs: Vertical and horizontal transmission among western north American Indian societies. *Human Nature*, 23(3):283–305.

Nikolai Sergeevich Trubetzkoy. 1928. Proposition 16. In *Acts of the First International Congress of Linguists*, pages 17–18.

Søren Wichmann and Eric W. Holman. 2009. *Temporal Stability of Linguistic Typological Features*. Lincom Europa.

Kenji Yamauchi and Yugo Murawaki. 2016. Contrasting vertical and horizontal transmission of typological features. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 836–846.