Probabilistic generative models of language contact

Will Chang

Workshop on Quantitative Approaches to Areal Linguistic Typology
KNAW Amsterdam
13-14 December 2012

ABSTRACT

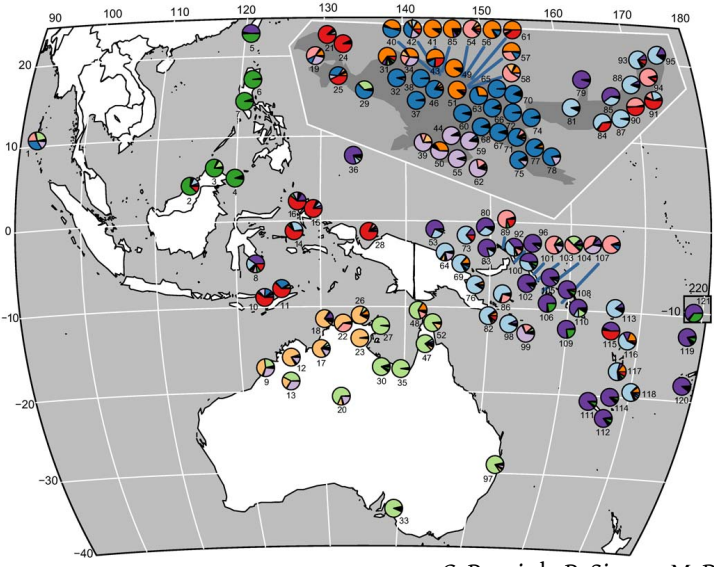
Ever since a model from population genetics was used on typological characters to illuminate the prehistory of Southeast Asia and the Pacific (Reesink et al., 2009), it became clear that such probabilistic models could be useful for a host of other linguistic problems as well. In this talk I will focus on the nuts and bolts of three such models, each an incremental improvment on, or a variant inspired by, the STRUCTURE model (Pritchard et al., 2000) that was used in the aforementioned study.

(1) I have applied the STRUCTURE model to data from the SAPhon database (Michael et al., 2012), consisting of the phonological inventories of 350+ South American languages. I will discuss three enhancements to the model that were necessary for good results: (i) a realistic enough prior for the feature frequencies in each ancestral population; (ii) a hierarchical Dirichlet prior (Teh et al., 2006) for language ancestries, so that the number of ancestral populations can be inferred automatically from the data, rather than set by the analyst; (iii) two methods for summarizing and visualizing the resulting posterior samples.

(2) The STRUCTURE model presupposes that all states of a character are equally amenable to borrowing. This assumption is awkward when it comes to modeling binary characters that represent the presence or absence of a feature. Lev Michael and I have constructed a rudimentary Relaxed Admixture Model to model features that one language can influence another to gain, but not to lose. When applied to the SAPhon database, this model proved useful for detecting the effects of relatively mild and recent contact.

(3) Rather than construct clusters of languages, as STRUCTURE essentially does, one can instead construct clusters of features — grouping together features that have similar distributions in the languages. This is especially effective when features are numerous and non-homoplastic. I have taken POLLEX, a comparative word list of Polynesian languages with over 4000 etyma (Biggs & Clark, 2006), and formed clusters with these etyma to induce isogloss bundles which reveal episodes of contact in the prehistory of these languages.

Explaining the Linguistic Diversity of Sahul using Population Models



G. Reesink, R. Singer, M. Dunn 2009

This talk is a show-and-tell of some of the things that I've been working on for the past two years. In some respects, the paper shown here (Reesink et al., 2009) is the point of departure for what I've been doing. In it, the authors used STRUCTURE (Pritchard et al., 2000), a model from population genetics, to explain the distribution of linguistic typological features in this part of the world. But for me things actually began when Lev Michael invited me do some analysis on a database that he was putting together, of the phonological inventories of South American languages. I hastily agreed because good databases are hard to come by in linguistics.

Talk overview

- Three models, two datasets.
- Datasets
  - SAPhon, a phonological inventory database for South American languages.
  - POLLEX, a etymological word list for Polynesian languages.
- Models
  - STRUCTURE, modified and used to analyze SAPhon.
  - Relaxed Admixture Model (RAM), also for analyzing SAPhon.
  - ETYMDIST, a clustering model for analyzing POLLEX.
- Focus
  - Model design.
  - Inference and implementation are not discussed.
  - Results are peripheral, for illustrating what the models do.
- Goals
  - Enable you to decide what these model are good for.
  - Hear how these models should evolve.

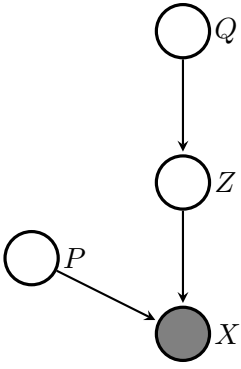
I will discuss two datasets and three models for analyzing them. I will focus on the design of the models and set aside matters of implementation and inference. I will also discuss some results, for the purpose of ullustrating the operation of the models. Some of the results are novel, but should be treated as provisional.

I hope that this talk will put you in a position to decide if any of these models would be useful to what you do. I also hope to garner feedback on how these models should evolve.

Terminology

What's a probabilistic generative model?

- *Generative*: The data (X) is generated (i.e. explained) by a set of underlying variables (P, Q, Z).
- *Probabilistic*: Variables are linked by probabilistic laws.



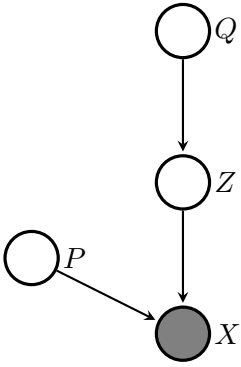
All of the models in this talk are probabilistic generative models. By *generative* I simply mean that the data is generated by a set of hidden or underlying variables, and by *probabilistic* I mean that all variables are related by probabilistic laws, as opposed to deterministically. The diagram shows such a model using a graph. By convention, an observed variable (i.e. data) is represented by a filled node. The diagram also shows that there is structure among the underlying variables. Q generates Z; and P and Z together generate X.

More terminology

We wish to *infer* values for P, Q, and Z. Bayesian inference gives a posterior distribution $P, Q, Z \mid X$.

- *Distribution*: A set of possibilities, each bearing a probability. The probabilities must add to one.
- *Posterior distribution*: Distributions for the underlying variables, in light of (i.e. posterior to seeing) the data.

Don't forget: before doing inference, we must state our *prior beliefs* ("priors") for the loose ends P and Q. What distributions do P and Q have before seeing the data?

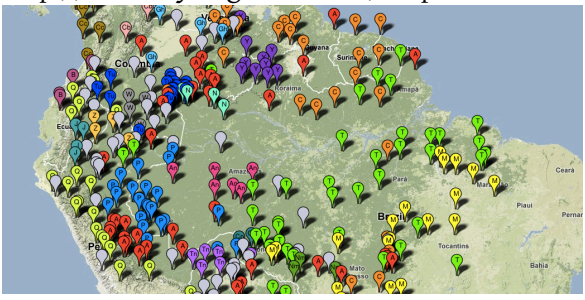


As is often the case with such models, the goal is to infer what the underlying variables (in this case, P, Q, and Z) might be, given what the data (X) is. In general, P, Q, and Z *could* be many things, but some of those possibilities are more likely than others. The possible values of P, Q, and Z jointly form a probability distribution, and since this is the distribution that is obtained *after* observing the data, it is called a posterior distribution. For each of the models in this talk, I use Bayesian inference to obtain an approximation to the posterior distribution for the underlying variables.

As a prerequisite of Bayesian inference, we must state our *prior beliefs* about the independent variables (in this case, P and Q) by specifying probability distributions over their possible values. These probability distributions encode what we know about P and Q *before* seeing the data X. Bayesian inference combines prior beliefs with the data to yield posterior beliefs about the underlying variables.

SAPhon: South American Phonological Inventory Database

- http://berkeley.linguistics.edu/~saphon



- Phonological inventories for South American languages (L. Michael, T. Stark, W. Chang, eds.).
- 355 languages, 297 features (most are phonemes).
- Reduce to a binary $N \times L$ matrix.
  - N languages, L features.
  - Each entry encodes whether language n has feature l.

SAPhon aims to be a high-quality, exhaustive database of the phonological inventories of the languages of South America. For my purposes I encode the data as simply a binary matrix, with N languages down the side and L features along the top. Each cell encodes the presence (1) or absence (0) of feature l in language n. The features are almost all phonemes, but a few encode for things such as the presence of tone or nasal harmony in the phonology of the language.

Some regularization has been done on the phonologies, to make them easier to compare. For example, /ɛ/ is replaced by /e/ whenever /e/ doesn't already exist, since the choice between /e/ and /ɛ/ may depend more on the linguist than on the language itself. Other information such as language family and geography are discarded during analysis, but are used in plotting results.

STRUCTURE backwards and forwards

X: Feature matrix

Barasana			b		t	d					k	g									s			h	w		j		ɾ	
Carapana		p	b		t	d					k	g									s		x		w		j		ɾ	
Guanano	pʰ	p	b	tʰ	t	d	tʃ			kʰ	k	g	ʔ								s			h	w		j		ɾ	
Karapana		p	b		t	d					k	g									s			h	w		j		ɾ	
Macuna			b		t	d					k	g									s			h	w		j		ɾ	
Piratapuyo		p	b		t	d					k	g	ʔ								s			h	w		j		ɾ	
Siriano		p	b		t	d					k	g									s			h	w		j		ɾ	
Tanimuca		p	b		t	d					k		ʔ								s			h	w		j		ɾ	
Tatuyo		p	b		t	d					k	g												h	w		j		ɾ	
Tucano	pʰ	p	b	tʰ	t	d				kʰ	k	g	ʔ								s			h	w		j		ɾ	
Tuyuca		p	b		t	d					k	g									s			h	w		j		ɾ	
Waimaha		p	b		t	d					k	g												h	w		j		ɾ	
Nukak		p	b		t	d		c	ɟ		k	g	ʔ											h	w				ɾ	
Daw		p	b		t	d		c	ɟ		k	g	ʔ	m	m'	n	n'	ɲ	ɲ'	ŋ		ʃ	x	h	w	w'	j	j'		l
Nadeb		p	b		t	d			ɟ		k	g	ʔ	m		n		ɲ		ŋ		ʃ		h	w		j		ɾ	

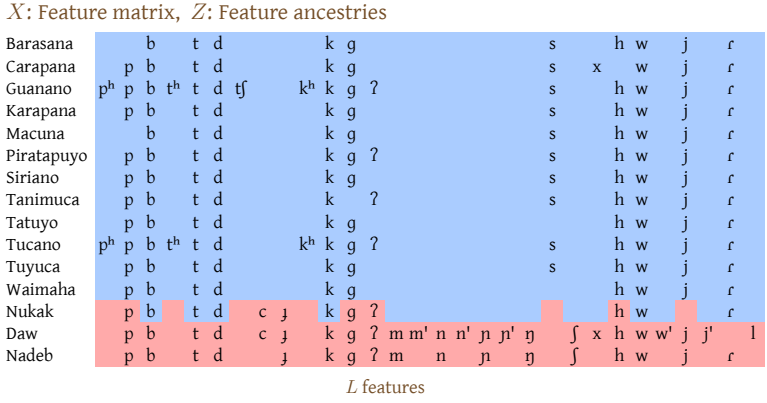
L features

N languages

Now I will discuss using the STRUCTURE model to analyze the SAPhon dataset. First I will review the core of the model, and afterwards explore the effect of different priors on feature frequencies and language ancestries.

Shown here is a binary matrix X containing a small subset of the data from SAPhon. For legibility I print the phoneme when it is present, rather than printing a bunch of zeroes and ones.

STRUCTURE backwards and forwards



For each cell STRUCTURE infers a *feature ancestry*, which indicates the source for that cell. In this example there are just two ancestries, blue and pink, but in general there may be more. This assignment of feature ancestries is denoted by Z. In this example, all languages except Nukak derive from just one ancestry. Nukak's inventory is analyzed as “mixed”, while the others are “pure”.

STRUCTURE backwards and forwards

P: Ancestral inventories (feature frequencies)

	pʰ	p	b	tʰ	t	d	tʃ	c	ɟ	kʰ	k	g	ʔ	m	m'	n	n'	ɲ	ɲ'	ŋ	s	ʃ	x	h	w	w'	j	j'	ɾ	l
Tucanoan	17	83	100	17	100	100	8	0	0	17	100	92	33	0	0	0	0	0	0	0	83	0	8	92	100	0	100	0	100	0
Nadahup	0	100	100	0	100	100	0	67	100	0	100	100	100	100	50	100	50	100	50	100	0	100	50	100	100	50	67	50	50	50

K clusters

X: Feature matrix, *Z*: Feature ancestries

	pʰ	p	b	tʰ	t	d	tʃ	c	ɟ	kʰ	k	g	ʔ	m	m'	n	n'	ɲ	ɲ'	ŋ	s	ʃ	x	h	w	w'	j	j'	ɾ	l
Barasana			b		t	d					k	g									s			h	w		j		ɾ	
Carapana		p	b		t	d					k	g									s		x		w		j		ɾ	
Guanano	pʰ	p	b	tʰ	t	d	tʃ			kʰ	k	g	ʔ								s			h	w		j		ɾ	
Karapana		p	b		t	d					k	g									s			h	w		j		ɾ	
Macuna			b		t	d					k	g									s			h	w		j		ɾ	
Piratapuyo		p	b		t	d					k	g	ʔ								s			h	w		j		ɾ	
Siriano		p	b		t	d					k	g									s			h	w		j		ɾ	
Tanimuca		p	b		t	d					k		ʔ								s			h	w		j		ɾ	
Tatuyo		p	b		t	d					k	g												h	w		j		ɾ	
Tucano	pʰ	p	b	tʰ	t	d				kʰ	k	g	ʔ								s			h	w		j		ɾ	
Tuyuca		p	b		t	d					k	g									s			h	w		j		ɾ	
Waimaha		p	b		t	d					k	g												h	w		j		ɾ	
Nukak		p	b		t	d		c	ɟ		k	g	ʔ											h	w				ɾ	
Daw		p	b		t	d		c	ɟ		k	g	ʔ	m	m'	n	n'	ɲ	ɲ'	ŋ		ʃ	x	h	w	w'	j	j'	ɾ	l
Nadeb		p	b		t	d			ɟ		k	g	ʔ	m		n		ɲ		ŋ		ʃ		h	w		j		ɾ	

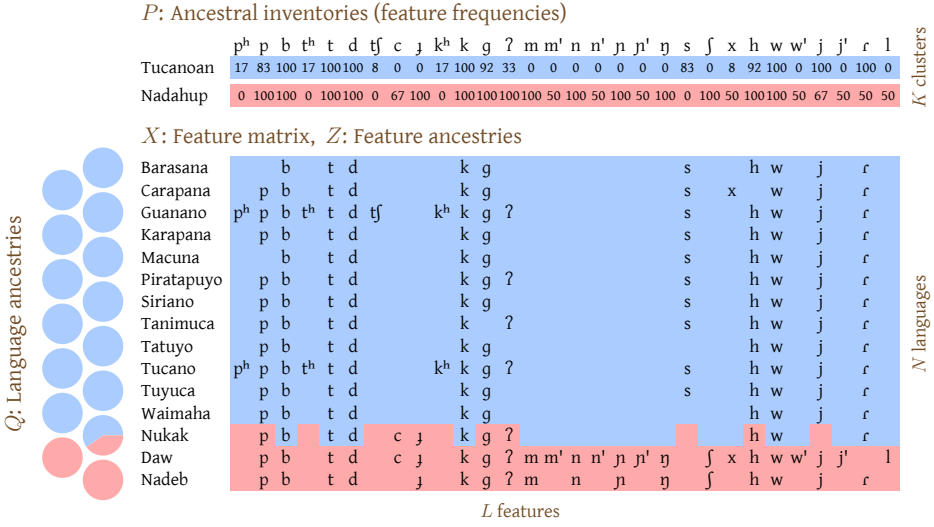
N languages

L features

STRUCTURE also infers *K ancestral population*, which are collectively denoted by *P*. (Each ancestral population is also referred to as an *ancestral inventory* or a *cluster*.) Each ancestral population is a bank of numbers, one for each feature. Each number indicates the frequency with which that feature appears in that ancestral population. This feature frequency corresponds to frequency with which the feature appears in the subset of modern languages that derive from the same ancestral population. For example, the feature frequency for /c/ in the pink ancestral population is 67%. This corresponds to the fact that, of the three languages whose value for /c/ derives from the pink ancestral population (Nukak, Daw, Nadeb), two of them have it.

For the sake of exposition I have labeled the two ancestral populations Tucanoan and Nadahup, though in reality STRUCTURE posits them with no regard to how linguists classify languages. Note that STRUCTURE posits a certain amount of *heterogeneity* in each ancestral population, as seen by the fact that not all of the feature frequencies are 0 or 100%. Nonetheless most of the feature frequencies are close to, if not actually, 0 or 100%, which means that these ancestral populations are actually quite homogeneous, and this accords with the notion that each ancestral population should have a distinct character.

STRUCTURE backwards and forwards

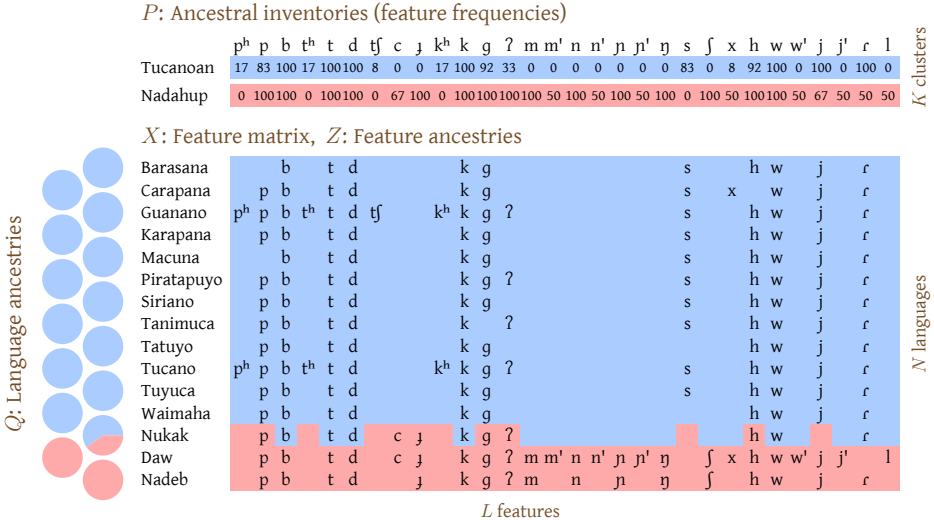


STRUCTURE also infers *N language ancestries*, which are collectively denoted by Q. In the abstract, each ancestry is simply a vector of K positive numbers that sum to one. I visualize them with pies. As noted earlier, all of the ancestries are pure except for that of Nukak. The ratio of pink to blue in its ancestry matches the ratio of pink to blue tiles in Nukak's feature ancestries.

The fact that most ancestries are pure accords with the principle that a model should be parsimonious. For Nukak, however, it is better to posit admixture, since none of the alternatives are appealing.

- If we put Nukak with Tucanoan, it would be the only language to have /c/ and /ɟ/, and to lack /s/.
- If we put Nukak with Nadahup, it would be the only language to lack phonemic nasals and /ʃ/.
- If we put Nukak in its own cluster, we would be neglecting its similarity to both Tucanoan and Nadahup.

STRUCTURE backwards and forwards



Sample to obtain joint posterior distribution $P, Q, Z \mid X$.

This is just one possible set of values for P, Q, and Z. Clearly others are possible, and perhaps quite probable as well. Bayesian inference will sample from among likely values for P, Q and Z to provide an estimate of the joint posterior distribution $P, Q, Z \mid X$.

STRUCTURE backwards and forwards

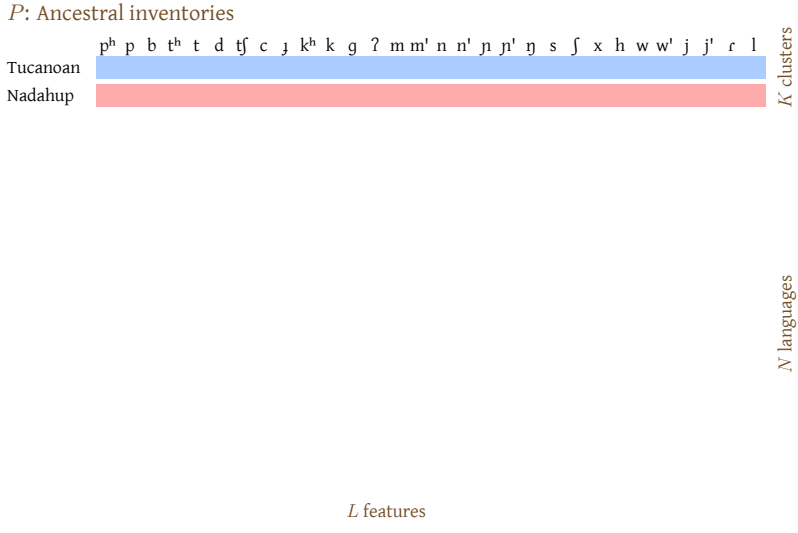
K clusters

N languages

L features

For a better understanding of STRUCTURE, we need to look at the data in a generative way. Let's generate some data according to the STRUCTURE model. We start with the givens: K clusters, N languages, and L features.

STRUCTURE backwards and forwards



First we generate P, the feature frequencies.

STRUCTURE backwards and forwards

P: Ancestral inventories (feature frequencies)

	pʰ	p	b	tʰ	t	d	tʃ	c	ɟ	kʰ	k	g	ʔ	m	m'	n	n'	ɲ	ɲ'	ŋ	s	ʃ	x	h	w	w'	j	j'	ɾ	l
Tucanoan	17	83	100	17	100	100	8	0	0	17	100	92	33	0	0	0	0	0	0	0	83	0	8	92	100	0	100	0	100	0
Nadahup	0	100	100	0	100	100	0	67	100	0	100	100	100	100	50	100	50	100	50	100	0	100	50	100	100	50	67	50	50	50

K clusters

N languages

L features

We draw $K \times L$ values for the feature frequencies from some prior distribution for P that is yet to be described. Most of the values should be near zero or one.

STRUCTURE backwards and forwards

P: Ancestral inventories (feature frequencies)

	pʰ	p	b	tʰ	t	d	tʃ	c	ɟ	kʰ	k	g	ʔ	m	m'	n	n'	ɲ	ɲ'	ŋ	s	ʃ	x	h	w	w'	j	j'	ɾ	l
Tucanoan	17	83	100	17	100	100	8	0	0	17	100	92	33	0	0	0	0	0	0	0	83	0	8	92	100	0	100	0	100	0
Nadahup	0	100	100	0	100	100	0	67	100	0	100	100	100	100	50	100	50	100	50	100	0	100	50	100	100	50	67	50	50	50

K clusters

Q: Language ancestries

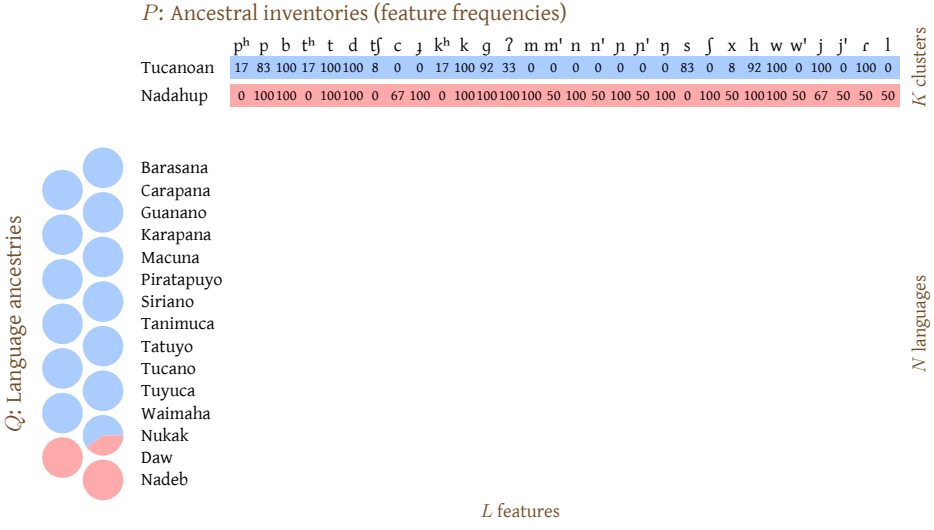
Barasana
Carapana
Guanano
Karapana
Macuna
Piratapuyo
Siriano
Tanimuca
Tatuyo
Tucano
Tuyuca
Waimaha
Nukak
Daw
Nadeb

N languages

L features

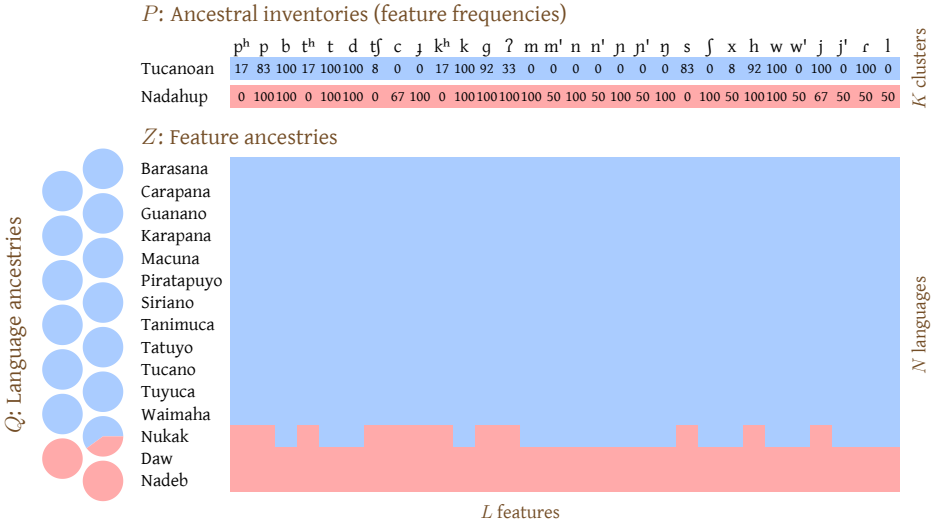
We also generate Q, the feature ancestries.

STRUCTURE backwards and forwards



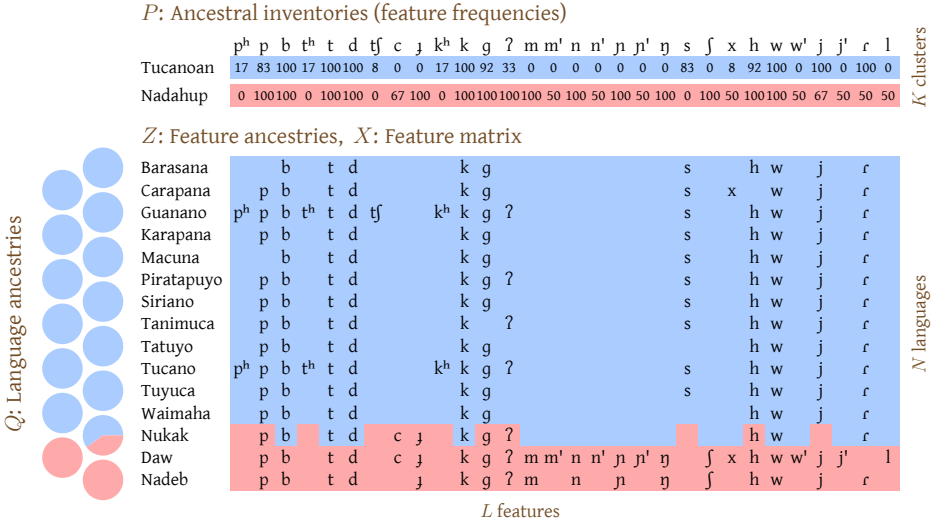
We draw N ancestries from some prior distribution for Q that is yet to be described. Most of the ancestries should be pure or nearly so.

STRUCTURE backwards and forwards



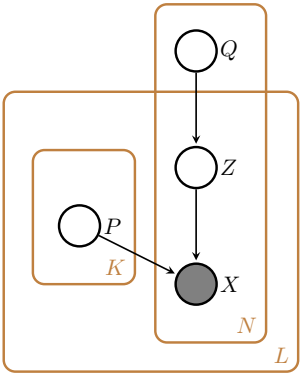
We consult Q to generate Z. For each language n and each feature l, we consult Q_n to generate Z_{nl}, the feature ancestry for each tile. Only for Nukak is there actually choice. Pink and blue are chosen randomly, with probability proportional to the size of pink and blue in Nukak's ancestry.

STRUCTURE backwards and forwards



Finally we consult Z and P to generate the data X. For each language n and each feature l, we first consult Z_{nl}, which is the feature ancestry for that tile. That tells us which ancestral population to consult. We then consult $P_{Z_{nl}l}$ (P indexed by Z_{nl} and l) to obtain the feature frequency for X_{nl}, which we generate with a weighted coin toss.

STRUCTURE core



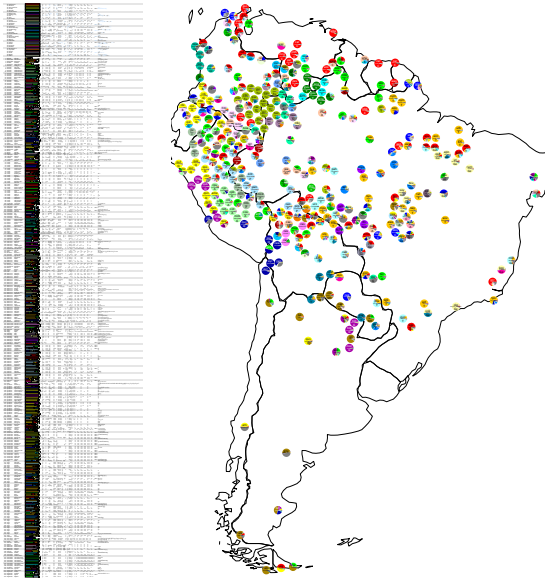
$Q_n \in \Delta^{K-1}$	Ancestry of language n.
$Z_{nl} \sim \text{Cat}(Q_n) \in \{1, \ldots, K\}$	Ancestry of feature l in language n.
$P_{kl} \in [0, 1]$	Frequency of feature l in population k.
$X_{nl} \sim \text{Bernoulli}(P_{Z_{nl}l}) \in \{0, 1\}$	Feature l present in language n?

This *plate diagram* is a more formal way to represent the model (see Bishop, 2006:363 for more on plate diagrams). The plates, which are drawn with brown rectangles, indicate two things. (1) They indicate the dimensions of the variables that they contain. (2) They indicate that the elements of the variables that they contain are independently generated. For example, the fact that P is inside the rectangles labeled K and L means that P has $K \times L$ elements. These elements are generated independently from some (as yet unspecified) prior distribution.

The annotations to the right of each variable indicate the range of each element of the variable. Each element of P, denoted by P_{kl}, falls in the interval $[0, 1]$. Each element of Q, denoted by Q_n, is a member of Δ^{K-1}, which denotes the set of all ancestries with K elements. Incidentally, each Q_n is actually a vector of K elements, but since these K elements must sum to one, they are not independently generated. Hence, I do not enclose Q in the plate labeled with K.

This diagram also describes how the variables are related. The expression $Z_{nl} \sim \text{Cat}(Q_n)$ means that each Z_{nl} is drawn from a categorical distribution parameterized by Q_n, where the probability of each outcome k is proportional to the kth element of Q_n. The expression $X_{nl} \sim \text{Bernoulli}(P_{Z_{nl}l})$ means that each X_{nl} is one with probability $P_{Z_{nl}l}$, or else zero. The generation of P and Q is unspecified, but will be discussed soon.

STRUCTURE results preview



These are the results when STRUCTURE is fed SAPhon in its entirety. (Please zoom in to see details.) On the right are pies showing the language ancestries for each language. (This is the information contained in Q.) I use 36 colors to show the 36 largest ancestral populations. The rest are lumped together and drawn with a dark gray.

STRUCTURE seems to be doing many things right. For example Nukak ('mbr', near the south end of Colombia) is predominantly olive-colored. The olive cluster consists of most of the Tucanoan languages. Most of the rest of Nukak's colors are associated with the Nadahup languages further to the southeast. This comports with the idea that Nukak has been influenced by both groups in historical times (CITE).

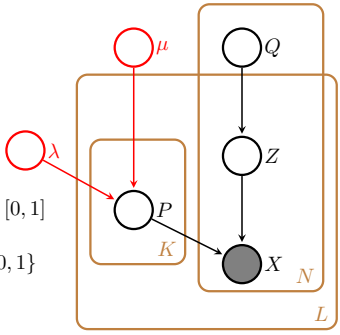
Another example of a small success is the pie for Yánesha ('ame', in Central Peru) which is colored yellow and pastel green. Yellow is the color of most Quechuan languages and pastel green is the color of the Kampan Arawak languages in its vicinity. This comports with what we know about Yánesha as a genetically Arawak language in the foothills of the Andes that has been heavily influenced by Quechuan (CITE).

On the left are two bits of information. At the top (where there are bars of one color) I show the 36 largest ancestral populations. To the left of each bar, I print the language that best exemplifies the ancestral inventory. To the right I show the feature frequencies for each feature with the color of the letters. (This is the information contained in P.) Features with feature frequencies in the second, third, and fourth quarters of the interval [0,1] are colored, respectively, light blue, blue, and black. Note that these ancestral inventories look more or less like real inventories.

Further below on the left I display language ancestries using bars rather than pies. The languages have been sorted by family, and it is possible to tell using this plot that Arawak phonologies are quite diverse, while Carib or Quechuan phonologies are much less so.

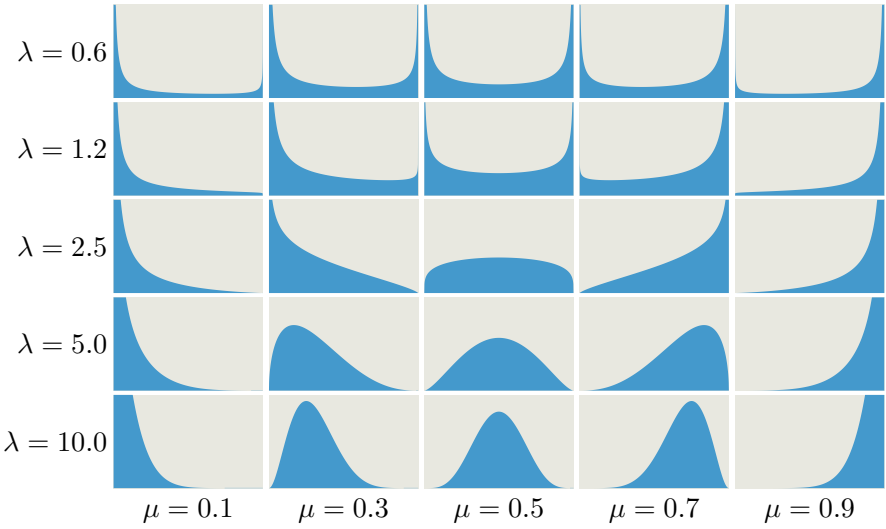
Naive feature frequency prior

Basic frequency of any feature.	$\mu \in [0, 1]$
Heterogeneity parameter.	$\lambda \in (0, \infty)$
Frequency of feature l in population k.	$P_{kl} \sim \text{Beta}(\lambda\mu, \lambda(1-\mu)) \in [0, 1]$
feature l present in language n?	$X_{nl} \sim \text{Bernoulli}(P_{Z_{nl}l}) \in \{0, 1\}$



Among the simplest possible priors for P is to draw each element independently from the same beta distribution. The next slide explains the peculiar way in which the beta distribution as been parameterized.

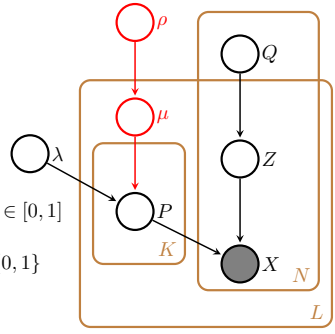
Density function of Beta$(\lambda\mu, \lambda(1-\mu))$ on $[0, 1]$



The mean of each distribution is given by μ, while λ determines how concentrated the mass is. A high λ yields feature frequencies that are all close to μ, while a low λ yields feature frequencies that are close to zero or one. Since most features are rare, we would expect μ to be low in real life; and since ancestral populations should be relatively homogeneous, we would expect λ to be low as well. When we allow the model to estimate μ and λ, that's exactly what we get.

Improved feature frequency prior

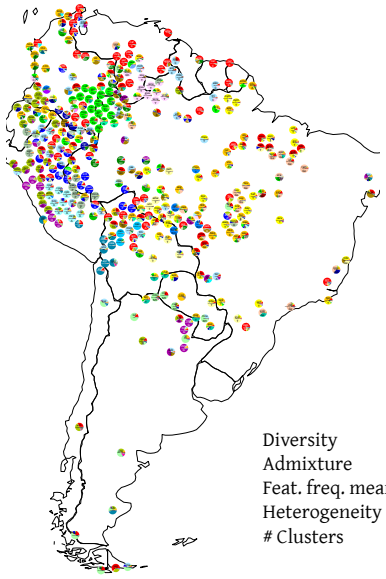
Shape parameter for universal feature frequencies.	$\rho \in (0, \infty)$
Universal frequency of feature l.	$\mu_l \sim \text{Beta}(\rho, 1) \in [0, 1]$
Heterogeneity parameter.	$\lambda \in (0, \infty)$
Frequency of feature l in population k.	$P_{kl} \sim \text{Beta}(\lambda\mu_l, \lambda(1-\mu_l)) \in [0, 1]$
Feature l present in language n?	$X_{nl} \sim \text{Bernoulli}(P_{Z_{nl}l}) \in \{0, 1\}$



We can improve the prior by respecting the fact that when we consider each feature individually, its frequency in each ancestral population will tend to be correlated. That is to say, $P_{1l}, P_{2l}, \ldots, P_{kl}$ will tend to reflect some universal feature frequency for feature l. To model this, we can posit a universal feature frequency μ_l for each feature, and generate $\mu_1, \ldots, \mu_L$ from a beta distribution parameterized by μ_l.

This prior is a special case of the prior in the *model with correlated allele frequencies* discussed in the appendix of Pritchard et al. (2000).

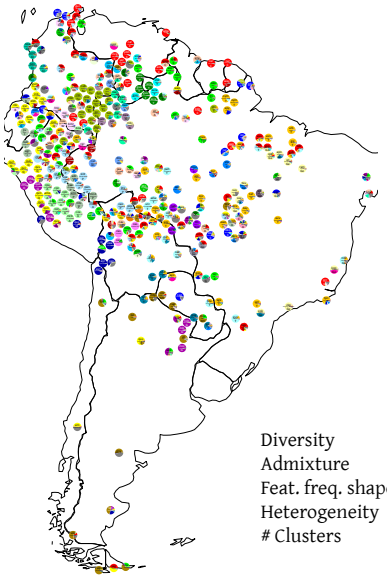
STRUCTURE with naive feature frequency prior



		Mean	95% HPD
Diversity	γ	21.4	[16.4, 26.3]
Admixture	α	0.835	[0.731, 0.957]
Feat. freq. mean	μ	0.102	[0.095, 0.109]
Heterogeneity	λ	0.069	[0.054, 0.084]
# Clusters		51.6	[45, 58]

This result is from using the model with the simple prior for P. Starting from the 25th inventory (exemplified by Karajá) the ancestral inventories are impossible in real life, as they have low feature frequencies for common sounds, and large collections of rare sounds that are unlikely to co-occur in a single population (see top-left of figure).

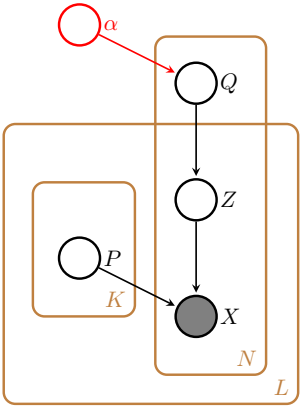
STRUCTURE results with full model



		Mean	95% HPD
Diversity	γ	34.0	[24.5, 46.2]
Admixture	α	0.490	[0.407, 0.565]
Feat. freq. shape	ρ	0.287	[0.251, 0.319]
Heterogeneity	λ	0.234	[0.190, 0.278]
# Clusters		69.6	[61, 76]

This result is from using the more sophisticated prior for P. There are no longer implausible ancetral inventories as in the previous slide, because having estimated universal feature frequencies $\mu_1, \ldots, \mu_L$ makes it hard to have low P_{kl} common features and high P_{kl} for rare features.

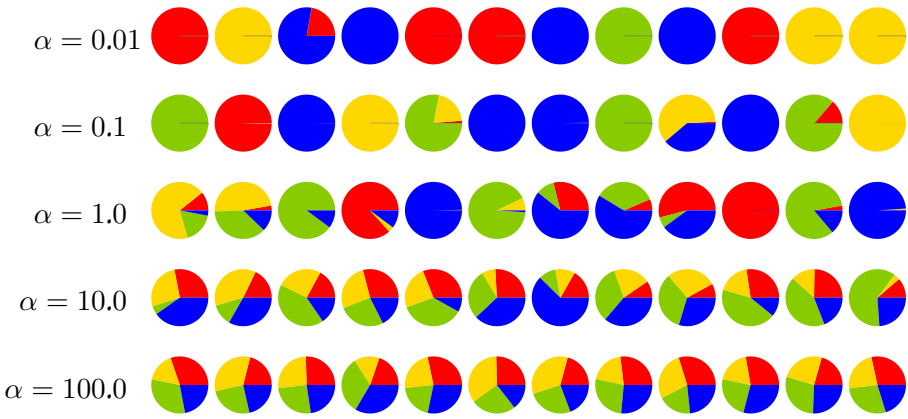
Dirichlet ancestry prior



$\alpha \in (0, \infty)$	Admixture parameter.
$Q_n \sim \text{Dirichlet}(\frac{\alpha}{K}) \in \Delta^{K-1}$	Ancestry of language n.
$Z_{nl} \sim \text{Cat}(Q_n) \in \{1, \ldots, K\}$	Ancestry of feature l in language n.
$P_{kl} \in [0, 1]$	Frequency of feature l in population k.
$X_{nl} \sim \text{Bernoulli}(P_{Z_{nl}l}) \in \{0, 1\}$	Feature l present in language n?

The simplest kind of prior for Q is a Dirichlet distribution, from which "pies" can be drawn independently, one for each language. The next slide illustrates the significance of the parameter α. The factor of $\frac{1}{K}$ need not concern us right now, as K is a constant.

Draws from Dirichlet($\frac{\alpha}{K}$) with $K = 4$



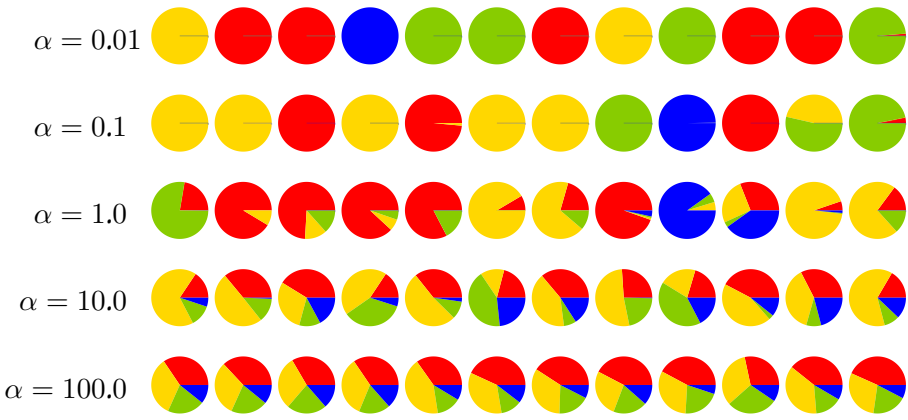
Each row shows multiple draws from a different parameterization of a Dirichlet distribution. The first row shows that when α is small, one gets mostly pure ancestries, and even when there is a second element, it is unlikely that there is a third. The next two rows show that as α increases, so does the amount of admixture in the pies that get drawn. When α is very high as in the last row, there is a nearly equal K-way split in each pie.

We can think of α as a parameter that determines how much admixture there is to be in the language ancestries. I mentioned before that in a parsimonious model, language ancestries should be relatively pure. It's good, then, that when the model is used to estimate α, it is quite low.

How can we improve on this prior for Q? One unsatisfactory thing about the prior is that it assumes that there are K equally-sized clusters. (A wedge of any color has a mean size of $\frac{1}{K}$.) On the other hand, our data probably does not consist in K equally-sized clusters, no matter how large we set K to be. If we lined up the clusters from the largest to the smallest, we would find that they are of unequal sizes, and there would be a large number of very small clusters at the end. Given that this is a clustering model with admixture, the smallest clusters may consist in no more than a few sounds in one or two languages.

The first step to modeling clusters of unequal sizes is to use an asymmetrical Dirichlet distribution.

Draws from Dirichlet$(0.4\alpha, 0.3\alpha, 0.2\alpha, 0.1\alpha)$

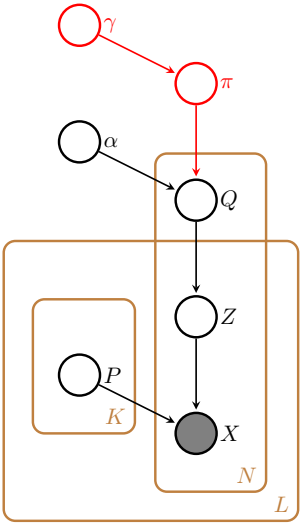


An asymmetrical Dirichlet distribution is just like a symmetrical Dirichlet distribution, except that there are now K parameters. Each parameter corresponds to a color, and the mean size of a wedge of that color is proportional to it.

These distributions have been parameterized so that the parameters always have the ratio $4:3:2:1$. I vary a multiplier α and draw repeatedly from each distribution. As expected, in each case, there is about four times as much red as there is blue, three times as much yellow, and twice as much green. As before, a small α leads to one color preponderating, and a large α leads to pies that are divided almost exactly according to the $4:3:2:1$ ratio.

We can use an asymmetrical Dirichlet distribution to generate language ancestries, but how do we generate the parameter ratio? We use another Dirichlet distribution, of course.

Hierarchical Dirichlet ancestry prior

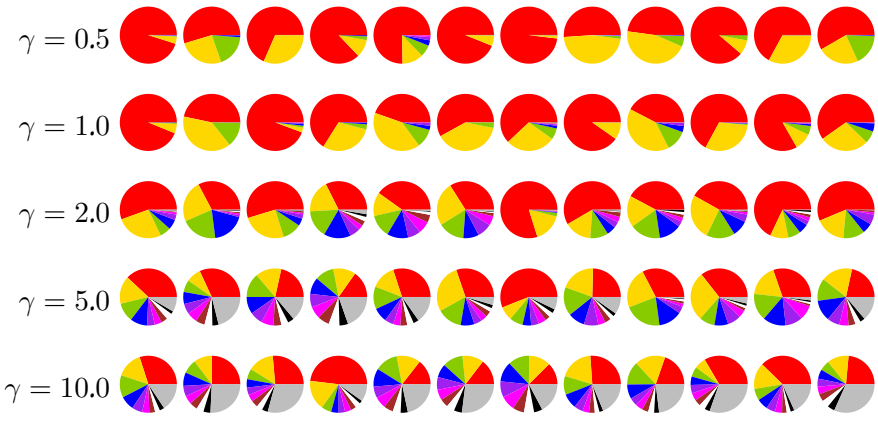


$\gamma \in (0, \infty)$	Diversity parameter.
$\pi \sim \text{Dirichlet}(\frac{\gamma}{K}) \in \Delta^{K-1}$	Relative cluster sizes.
$\alpha \in (0, \infty)$	Admixture parameter.
$Q_n \sim \text{Dirichlet}(\alpha\pi_1, \ldots, \alpha\pi_K) \in \Delta^{K-1}$	Ancestry of language n.
$Z_{nl} \sim \text{Cat}(Q_n) \in \{1, \ldots, K\}$	Ancestry of feature l in language n.
$P_{kl} \in [0, 1]$	Frequency of feature l in population k.
$X_{nl} \sim \text{Bernoulli}(P_{Z_{nl}l}) \in \{0, 1\}$	Feature l present in language n?

In the *hierarchical Dirichlet* ancestry prior, I use an asymmetrical Dirichlet distribution to generate language ancestries, and I use another Dirichlet distribution to generate the asymmetry. The ratios between cluster sizes is denoted by $\pi = (\pi_1, \ldots, \pi_K)$, whose elements sum to one. Language ancestries are independently drawn from $\text{Dirichlet}(\alpha\pi_1, \ldots, \alpha\pi_K)$, and π is drawn from a symmetrical Dirichlet distribution parameterized by $\frac{\gamma}{K}$.

The next two slides will show the effects of varying γ and K. It will be seen that as $K \to \infty$, the distribution $\text{Dirichlet}(\frac{\gamma}{K})$ converges in an important respect.

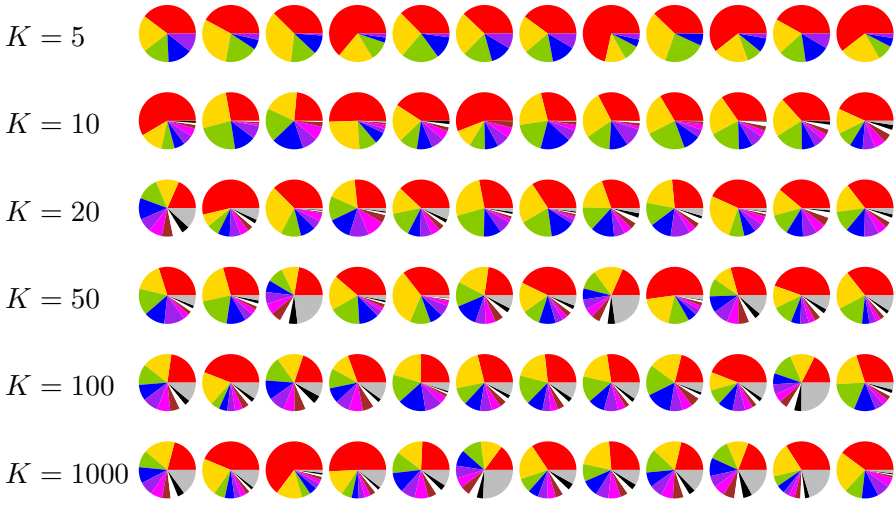
Sorted draws from Dirichlet($\frac{\gamma}{K}$) with $K = 100$



Each row shows multiple *sorted* draws from a different parameterization of a Dirichlet distribution. The ancestry elements have been sorted so that red is always used for the largest element, yellow for the second largest, etc. The largest nine elements each have their own color, and the rest are lumped together as gray.

As expected, when γ is small, red preponderates, but the pies become more finely divided as γ increases. We can regard γ as a *diversity parameter* that determines how fragmented the languages are.

Sorted draws from Dirichlet($\frac{5}{K}$)

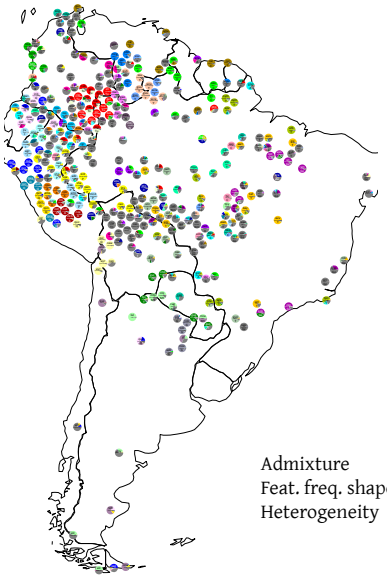


The size of ith largest slice converges in law as $K \to \infty$.

Now γ is fixed at 5, and K is varied. Note that there is no observable difference between the distributions for $K = 100$ and $K = 1000$. The upshot is that for any value of γ, it does not matter what K is set to as long as it is set large enough. This is suited to modeling a dataset such as SAPhon, where it does not make sense to think of diversity in terms of how many clusters there are (since they tail off into lots of small ones), but rather in terms of how quickly the cluster sizes tail off, as governed by γ.

Of theoretical interest is that as $K \to \infty$, the hierarchical Dirichlet prior converges to well-studied objects in probability (Teh et al., 2006).

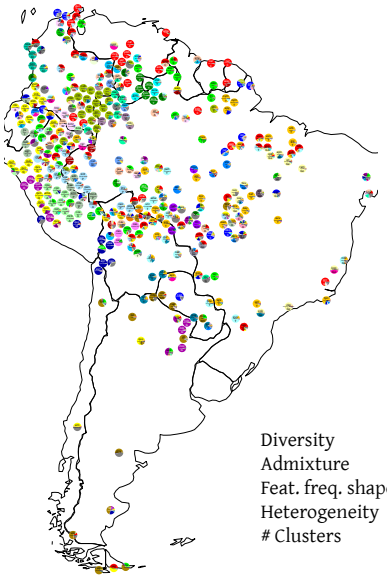
STRUCTURE with simple Dirichlet ancestry prior ($K = 90$)



		Mean	95% HPD
Admixture	α	0.423	[0.354, 0.487]
Feat. freq. shape	ρ	0.260	[0.229, 0.290]
Heterogeneity	λ	0.165	[0.127, 0.204]

This is the outcome of using a simple Dirichlet prior for Q. $K = 90$ is the setting that approximately maximized the marginal likelihood, using the inference method for K given in the appendix of Pritchard et al. (2000). ($K = 90$ was better than $K = 80$ or $K = 100$.)

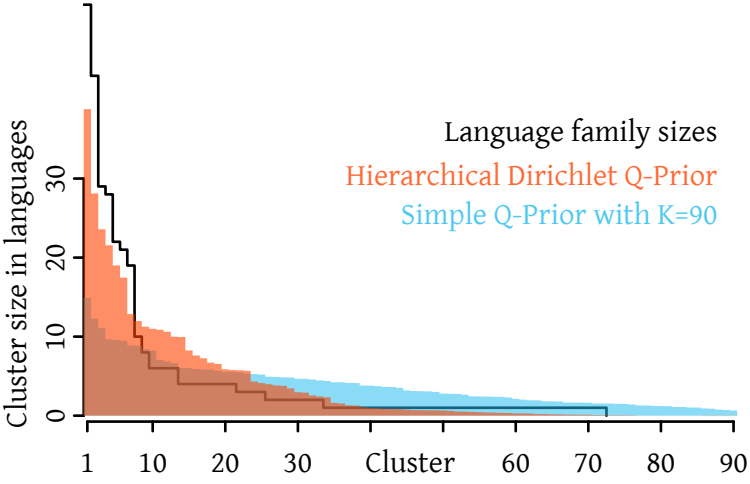
STRUCTURE results with full model



		Mean	95% HPD
Diversity	γ	34.0	[24.5, 46.2]
Admixture	α	0.490	[0.407, 0.565]
Feat. freq. shape	ρ	0.287	[0.251, 0.319]
Heterogeneity	λ	0.234	[0.190, 0.278]
# Clusters		69.6	[61, 76]

This is the outcome of using the hierarchical Dirichlet prior. (We've seen this twice before.)

Cluster size histograms



- The simple prior splits too aggressively with large clusters.
- The hierarchical prior may lump too aggressively with small clusters.

On this plot are superimposed three things:

- The cluster sizes that result from using the simple Q-prior, in blue.
- The cluster sizes that result from using the hierarchical Dirichlet Q-prior, in red.
- And for reference, a histogram of language family sizes, in black.

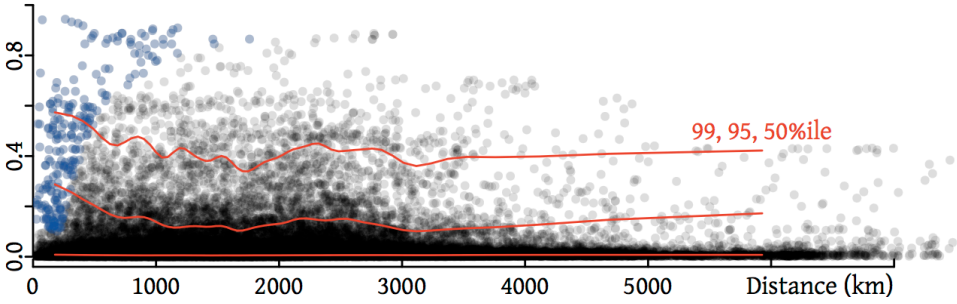
The simple Q-prior prefers clusters of similar sizes, as can be seen from the relative flatness of the blue histogram. If the language family classifications are anything to go by, it's probably the case that the simple Q-prior splits large clusters too aggressively. It also seems to have too wide of a tail, with too many clusters consisting of 2 to 5 languages. The hierarchical Dirichlet Q-prior follows the language family classifications more closely, though perhaps its tail tapers too quickly.

I lack accurate marginal likelihood estimates for the priors, but browsing the results, I see places where one or the other is better. For example, the simple prior breaks the Quechuan languages up into too many clusters; but the hierarchical Dirichlet prior proposes some implausible connections that the simple prior does not, such as between Cabiyarí and Pémono, about which more will be said later.

Finally, I should mention that the simple prior has one more disadvantage, which is that with it, the analyst has to do multiple runs to find the optimal value for K. This is a considerable weakness, as each run takes about a month on a fast machine.

Inter-family feature-sharing frequency, by distance

- Let's plot a subset of these language pairs on a map.

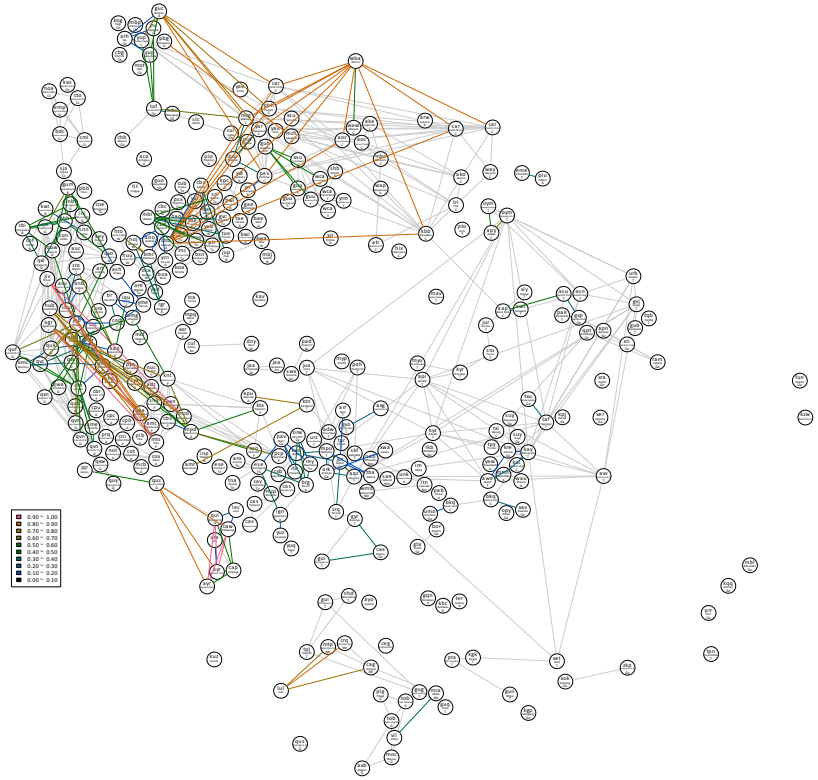


This is another way to understand the results of the STRUCTURE analysis. Each dot represents a pair of languages that are not in the same family. The x-axis is the distance between them, in kilometers, and the y-axis is the mean fraction of features that they share the same source for. Also shown are quantile lines, conditioned on inter-language distance. There are three reassuring things in this plot.

- The median is very small. Essentially, two languages from different families, chosen at random, do not share the same source for any sounds.
- The median does not get higher as the distance gets smaller. This means that STRUCTURE results does not simply recapitulate geography.
- The higher quantile lines do get higher as the distance gets smaller. This means that proximity makes it likelier for two languages to have been in contact, and STRUCTURE was able to infer this, despite having been given no geographical information.

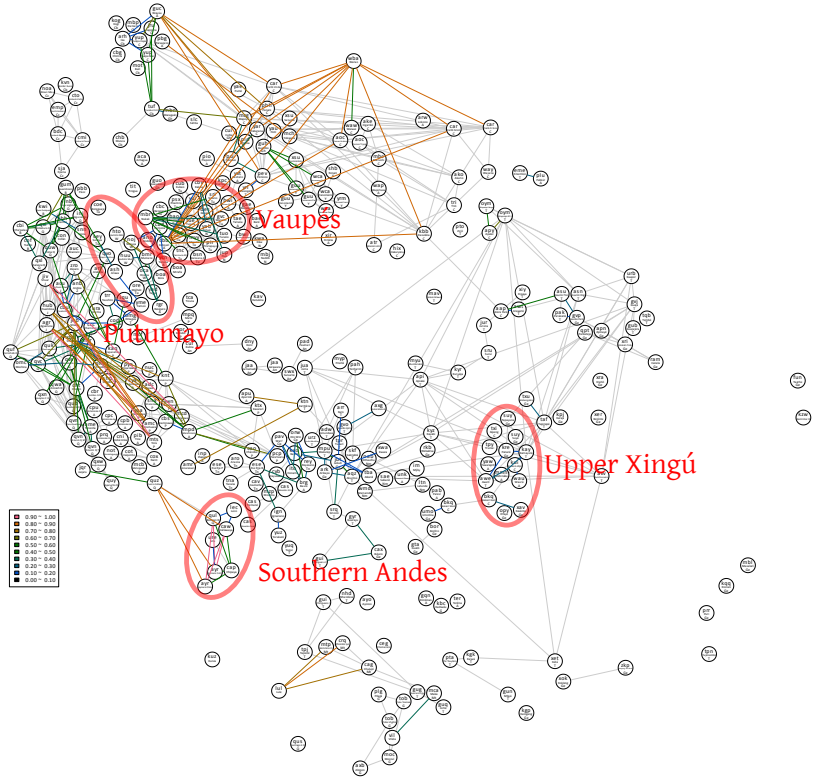
Naturally we would like to know which pairs of languages are high in the plot. What I will do, is plot on a map a line between two languages if they share the same source for many of their features. I cannot do this for too many pairs of languages without creating an unreadable plot, so I will do it for the arbitrary subset of the dots shown in blue. For languages that are near each other, I will plot them if they share the same source for 10% of their features; and I increase this threshold as the distance gets larger.

STRUCTURE lines plot



This ‘line plot’ is an alternative to the pie plot that throws inter-family connections into relief. Inter-family connections are drawn in color; intra-family connections are in light gray. A line between two languages means that one has borrowed from the other; or that they both obtained features from a third source.

STRUCTURE lines plot



This plot provides backing for some proposed South American linguistic areas. Four appear in this slide. However, these linguistic areas are often obscured by lines that run longer distances. For example, the Vaupes is obscured by amber-colored lines, which denote pairs that share the same source for 80 to 90% of their features. It turns out that these connections are specious.

Cabiyarí and Pémono

- Cabiyarí's inventory is not too unusual for an Arawak language.
- Pémono's inventory is very typical for a Carib language.

Cabiyarí

Consonants	Bilabial	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Stop/affricate	p	t̪	t	tʃ		k	ʔ
Fricative							h
Nasal	m		n				
Approximant	w				j		
Tap, flap			ɾ				

Vowels	Front	Central	Back
High	i		u
Mid	e		o
Low		a	

Pémono

Consonants	Bilabial	Alveolar	Palatal	Velar	Glottal
Stop	p	t		k	ʔ
Fricative		s			h
Nasal	m	n	ɲ		
Approximant	w		j		
Tap, flap		ɾ			

Vowels	Front	Central	Back
High	i	ɨ	u
Mid	e	ə	o
Low		a	

- STRUCTURE puts them in the same cluster. 86% of their features are deemed to come from the same ancestral population.

Let's have a look at a particularly egregious false positive: the supposed connection between Cabiyarí and Pémono. STRUCTURE reckons that these two languages share the same source for 86% of their features. Their inventories do not seem particularly close, but STRUCTURE lumps them together because it does not realize that Arawak phonological inventories are very diverse, and Cabiyarí's ancestry suffices to explain its inventory without the need to posit borrowing. (Recall that STRUCTURE is not given information on a language's classification.)

Relaxed Admixture Model (RAM)

From STRUCTURE to RAM:

- Make use of linguistic classifications, to weed out spurious connections such as between Cabiyarí and Pémono.
- Posit *relaxed admixture* in order to detect areas such as Upper Xingú.

Relaxed admixture: a model of contact where features can be gained, but not lost, due to contact.

- Clearly not realistic.
  - Nukak lacked phonemic nasal stops, presumably due to influence from Tucanoan languages.
  - Ashéninka (Apurucayali) lost mid vowels, presumably due to contact with Quechuan languages.
- But, useful for detecting mild and recent contact.

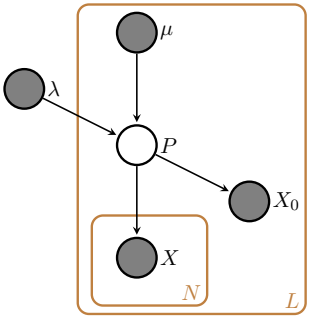
The Relaxed Admixture Model (RAM) is an attempt to take a step back, by using a model that is more modest than STRUCTURE in scope, but targets some of its shortcomings. Rather than seek to model the entire dataset coherently, as STRUCTURE does, RAM looks at pairs of languages only, and asks, if this were the only contact scenario in all of South America, how much explanatory power do we gain by positing it? Unlike STRUCTURE, RAM is given the linguistic family for each language, and it employs a simple model for how a language gets its features from its family.

RAM was designed to be sensitive to recent contact, and so it holds that the presence of a sound can be borrowed, but that the absence of a sound cannot — a property that we (Lev and I) termed *relaxed admixture*. We were banking on the notion that gaining a sound can easily happen in instances of superficial contact, but that losing a sound often meant a structural change in the phonology of the language, which necessitates more intense contact.

Relaxed admixture is clearly naive — as noted here concerning Nukak and Ashéninka — but it does prove useful.

Model $\mathcal{M}_0$: Inheritance only

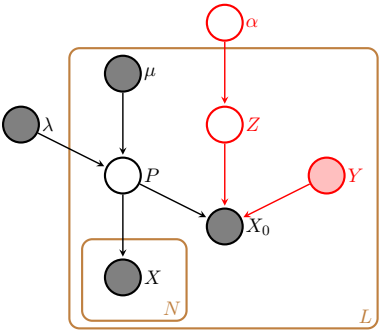
Universal frequency of feature l.	$\mu_l \in (0, 1)$
Generality of universal feature frequencies.	$\lambda \in (0, \infty)$
Frequency of feature l in this family.	$P_l \sim \text{Beta}(\lambda\mu_l, \lambda(1-\mu_l)) \in (0, 1)$
Feature l in target language	$X_{0l} \sim \text{Bernoulli}(P_l) \in \{0, 1\}$
Feature l in language n, which is in the target's family.	$X_{nl} \sim \text{Bernoulli}(P_l) \in \{0, 1\}$



- Model $\mathcal{M}_0$ explains X_0 as the result of inheritance alone.
- Universal feature frequencies $\mu_1, \ldots, \mu_L$ and generality parameter λ are fixed based on previous runs of STRUCTURE.
- μ and λ are significant when the family is small, or when the target is an isolate.

A RAM analysis is set up as a model evaluation problem. For each pair of languages, one acting as potential donor and the other as potential target, we propose two models: one with contact, and one without. Here is the null hypothesis — the model without contact. X_0 is the target inventory. Each feature is generated via a Bernoulli distribution, parameterized by a feature frequency that is obtained by analyzing all the members of the language family. There is not enough data to infer the hyperparameters μ and λ, so these are inferred from a run of STRUCTURE and fixed in this model. They become quite important when the target is a linguistic isolate, or is from a small family.

Model $\mathcal{M}_1$: Relaxed admixture model (RAM)

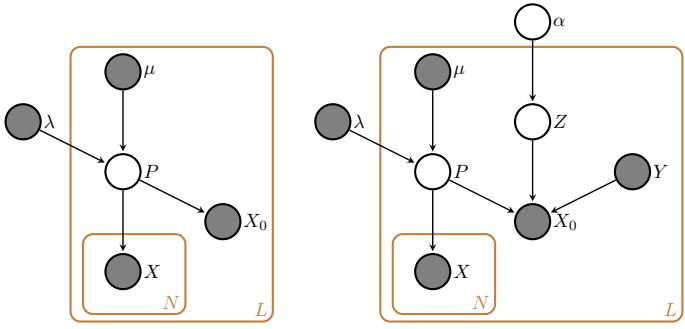


$\alpha \sim \text{Beta}(\frac{1}{2}, \frac{1}{2}) \in (0, 1)$	Admixture parameter.
$Z_l \sim \text{Bernoulli}(\alpha) \in \{0, 1\}$	Can feature l be borrowed?
$Y_l \in \{0, 1\}$	Feature l in donor language.
$\begin{cases} X_{0l} = 1 \text{ if } Z_l = 1 \text{ and } Y_l = 1, \\ X_{0l} \sim \text{Bernoulli}(P_l) \text{ otherwise.} \end{cases}$	Feature l in target language.
$X_{nl} \sim \text{Bernoulli}(P_l) \in \{0, 1\}$	Feature l in language n, which is in the target's family.

- Model $\mathcal{M}_1$ explains X_0 as the result of inheritance and borrowing from Y.
- Only the presence of a feature can be borrowed.
- Admixture parameter α denotes the fraction of features *present* in Y that are borrowed into X_0.

The alternative hypothesis is that the target inventory X_0 is a product of both inheritance and borrowing. The variable Z is binary vector, with $Z_l = 1$ if X_0 *may* borrow feature l from the donor inventory Y. The feature is actually borrowed only if the feature is present in Y, i.e. if $Z_l = 1$ and $Y_l = 1$.

Borrowing score



$\mathcal{M}_0$: Just inheritance. $\mathcal{M}_1$: Borrowing too.

- Which model explains X_0 better? Compute the Bayes factor.

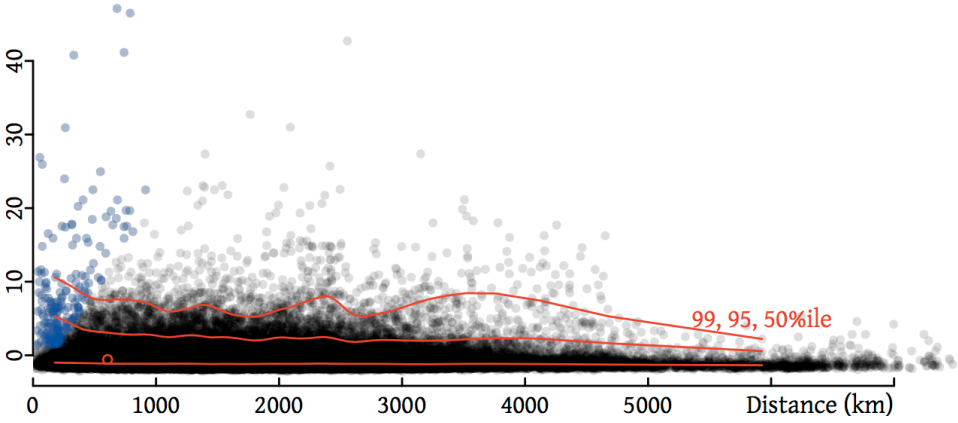
$$\mathcal{K} = \frac{\mathbb{P}(X_0 \mid \mathcal{M}_1)}{\mathbb{P}(X_0 \mid \mathcal{M}_0)} = \frac{\sum_\alpha \sum_P \mathbb{P}(P, \alpha, X, X_0 | \lambda, \mu, \mathcal{M}_1)}{\sum_P \mathbb{P}(P, X, X_0 | \lambda, \mu, \mathcal{M}_0)}$$

- Use $\log \mathcal{K}$ as a *borrowing score* for each donor-target pair.
- Note: $\mathcal{M}_1$ for one donor-target pair may be incompatible with $\mathcal{M}_1$ for another.

Both models are simple enough that their marginal likelihoods $\mathbb{P}(X_0 \mid \mathcal{M}_1)$ and $\mathbb{P}(X_0 \mid \mathcal{M}_0)$ can be computed exactly. We take the log of the Bayes factor as a *borrowing score*, which indicates how advantageous it is to posit borrowing. When the borrowing score is greater than zero, the alternative model is favored. It is important to keep in mind that unlike STRUCTURE, RAM is a local model: the $\mathcal{M}_1$ for one donor-target pair is not necessarily compatible with the $\mathcal{M}_1$ for another.

Inter-family borrowing scores by distance

- For any pair of languages not in the same family, plot the larger borrowing score.



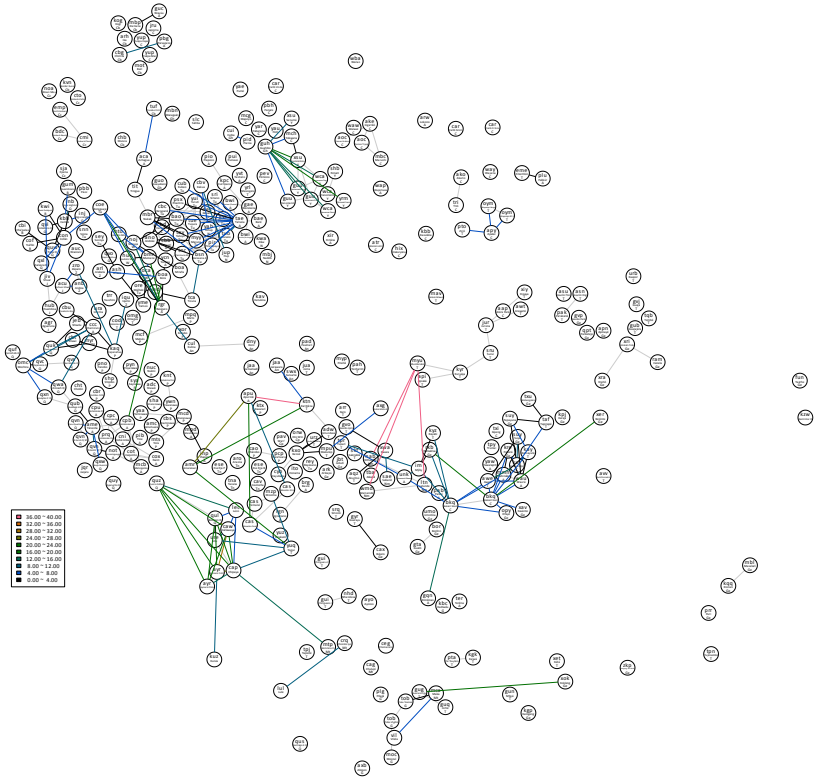
This plot shows, for each pair of languages not in the same family, the higher of the two borrowing scores involving the pair, plotted against the distance between the pair in kilometers. Also shown are quantile lines, conditioned on distance. Again there are three things about this plot that seem right.

- The median line is well below zero. Two languages chosen at random will be unlikely to have a positive borrowing score.
- The median line is relatively flat, indicating that borrowing score is not merely a function of proximity.
- The higher-quantile lines get higher as the distance decreases. RAM finds that the closer two languages are, the profitable it is to posit borrowing.

Also, RAM does not yield false positives like Cabiyarí and Pémono. That particular pair has been plotted as a red circle. It can be seen that its borrowing score is negative. RAM accounts for the fact that Awarak is a diverse language family, so that inheritance suffices to explain Cabiyarí's features.

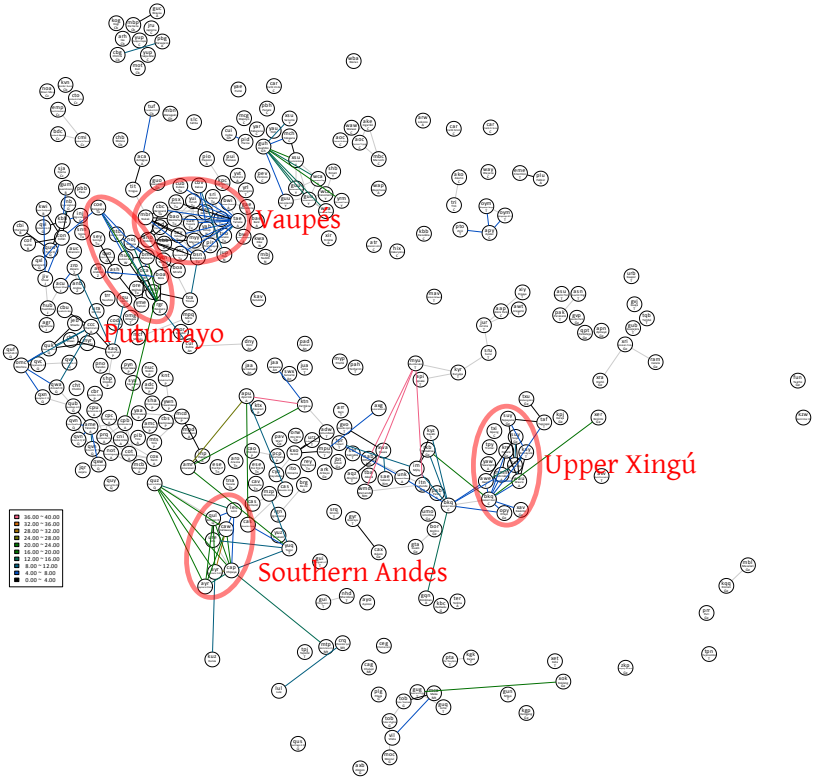
As with STRUCTURE, it is possible to plot higher borrowing scores as lines on a map. I do this for the arbitrary subset of dots colored blue.

RAM line plot



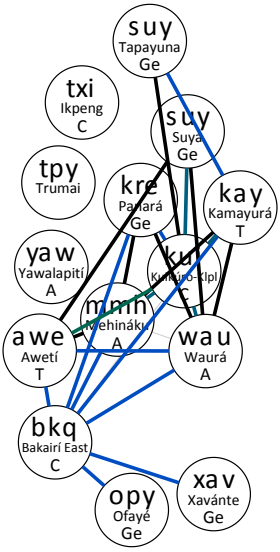
Gone are many of the distracting false positives seen in the line plot for STRUCTURE.

RAM line plot



The proposed linguistic areas appear more clearly on this plot.

Borrowing in Upper Xingú: Nasal vowels



Segments identified by RAM as likely borrowed, contingent on contact:

Donor	Target	
Kamayurá (T)	Waura (A)	ɨ̃ ẽ ĩ ũ ã ɨ
Panara (Ge)	Waura (A)	ɨ̃ ẽ ũ ĩ ã ɨ
Awetí (T)	Waura (A)	ɣ ɨ̃ ẽ ũ ĩ ã ɨ
Suya (Ge)	Waura (A)	ɣ ɨ̃ ẽ ũ ĩ ã ɨ
Kamayurá (T)	Kuikuro (C)	ts ŋ õ ɨ̃ ẽ ũ ĩ ã
Awetí (T)	Kuikuro (C)	ts ɣ ŋ õ ɨ̃ ẽ ũ ĩ ã l
Panara (Ge)	Kuikuro (C)	õ ɨ̃ ẽ ũ ĩ ã
Suya (Ge)	Kuikuro (C)	ɣ ŋ õ ɨ̃ ẽ ũ ĩ ã
Tapayuna (Ge)	Kuikuro (C)	ŋ õ ɨ̃ ẽ ũ ĩ ã

Feature frequencies for nasal vowels in S. America:

Macro-Ge	~55%	Carib	~16%
Tupian	~85%	Arawak	~16%

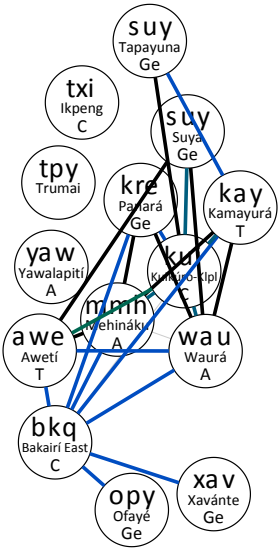
This is a blowup of the Upper Xingú. For each pair of languages in the table, RAM was made to print out the features most likely to have been borrowed, contigent on there having been borrowing between the pair of languages.

One useful characteristic of probabilistic generative models is that one can examine the marginal distributions for any underlying variable. Here, we seek features l such that the posterior expecation of borrowing $\mathbb{E}(Z_l|X_0, X, Y, \mu, \lambda, \mathcal{M}_1)$ is high.

In the results we see that RAM deems nasal vowels to have been borrowed into the Arawak and Carib languages of the Upper Xingú. This makes sense, given that they occur infrequently in Arawak and Carib languages in the continent as a whole.

Note that RAM suggests several candidate donors for each target. While it is impossible to decide which languages were the actual donors, it is entirely plausible that this kind of exchange has been taking place.

Borrowing in Upper Xingú: /ɨ/



Segments identified by RAM as likely borrowed, contingent on contact:

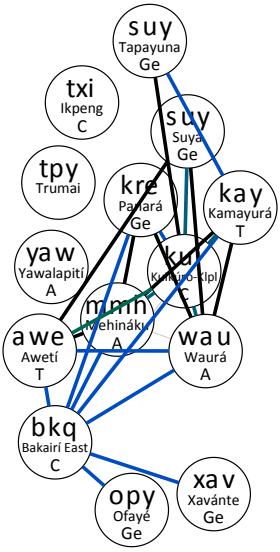
Donor	Target	
Kamayurá (T)	Waura (A)	ɨ̃ ẽ ĩ ũ ã ɨ
Panara (Ge)	Waura (A)	ɨ̃ ẽ ũ ĩ ã ɨ
Awetí (T)	Waura (A)	ɣ ɨ̃ ẽ ũ ĩ ã ɨ
Suya (Ge)	Waura (A)	ɣ ɨ̃ ẽ ũ ĩ ã ɨ
Kamayurá (T)	Kuikuro (C)	ts ŋ õ ɨ̃ ẽ ũ ĩ ã
Awetí (T)	Kuikuro (C)	ts ɣ ŋ õ ɨ̃ ẽ ũ ĩ ã l
Panara (Ge)	Kuikuro (C)	õ ɨ̃ ẽ ũ ĩ ã
Suya (Ge)	Kuikuro (C)	ɣ ŋ õ ɨ̃ ẽ ũ ĩ ã
Tapayuna (Ge)	Kuikuro (C)	ŋ õ ɨ̃ ẽ ũ ĩ ã

Feature frequencies for /ts/ in S. America:

Carib	~90%	Arawak	~30%
Macro-Ge	~70%		
Tupian	~92%		

Similarly, RAM infers that /ɨ/ was borrowed into Waura, an Arawak language. Its frequency in Arawak languages elsewhere is relatively low.

Borrowing in Upper Xingú: /ts/



Segments identified by RAM as likely borrowed, contingent on contact:

Donor	Target	
Waura (A)	Kuikuro (C)	ɣ ts ɨ̃ ẽ ũ ã ĩ l
Kamayurá (T)	Kuikuro (C)	ts ŋ ɨ̃ õ ẽ ũ ã ĩ
Awetí (T)	Kuikuro (C)	ɣ ts ŋ ɨ̃ ũ ã ĩ l
Mehinaku (A)	Kuikuro (C)	ts ɨ̃ ẽ ũ ĩ ã l
Waura (A)	Kamayurá (T)	ts
Waura (A)	Awetí (T)	ɣ l ts

Feature frequencies for /ɨ/ in S. America:

Arawak	~63%	Carib	~3%
		Macro-Ge	~7%
		Tupian	~15%

And finally, RAM infers /ts/ to have been borrowed into various Carib and Tupían languages in the region.

In general, RAM is most useful for investigating linguistic areas where languages have just begun to influence one another. The Upper Xingú is not the kind of linguistic area that is defined by distinctive features that are not found anywhere else — at least in its phonological inventories. If we simply plot the occurrence of various features on a map, the Upper Xingú will fail to appear. Nor is the Upper Xingú an area of obvious homogeneity. The best way to detect it is to look for convergence between its languages, against the backdrop of their respective genetic profiles.

POLLEX: Polynesian Lexicon Project

- http://pollex.org.nz/
- An etymological word list for Polynesian languages and a small number of Oceanic neighbors (B. Biggs, R. Clark)
- 40+ languages, 4000+ etyma, 40,000+ forms.
- Reduce to a binary $N \times L$ matrix.
  - N etyma, L languages.
  - Each entry encodes whether etymon n is attested in language l.
  - Treat cognates and loans the same.

Polynesian Lexicon Project Online

Protoform: QAROFA.A [MP] Love, pity, compassion

Description:	Love, pity, compassion
Reconstruction:	Reconstructs to MP: Malayo-Polynesian
Notes:	*4 POC *qaro-qopa. *5 PMP *harep "gernhaben" (Dpf. 1938). *5 PMP *herap "to like" (Dpf). *7 Cf. PPN *qofa "greeting".

Pollex entries:

Language	Reflex	Description	Source
Anuta	Aropa	Pity, compassion	(Yen)
East Futuna	ʔAlofa-ʔia	Love, pity, compassion	(Bgs)
East Uvea	ʔOfa	Amitie, affection, amour, faveur	(Rch)
East Uvea	ʔAlofa	Salut	(Rch)
Easter Island	ʔAroha	Commiseration, compassion, condolence	(Pts)
Emae	Faka/arofa	Poor	(Clk)
Fijian	Garo	Desire **Problematic**	(Cpl)
Hawaiian	Aloha	Love, pity, compassion	(Pkl)

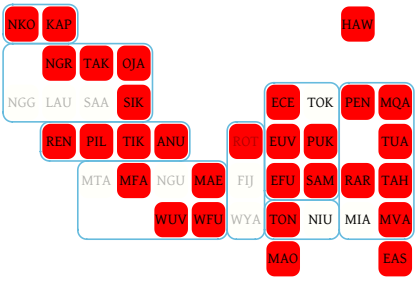
And now, I will transition abruptly to talking about ETYMDIST, starting with the dataset to be analyzed.

POLLEX consists of data of a much different sort than SAPhon, but we will still reduce it to a binary matrix, with each entry denoting the presence or absence of a particular etymon in a particular language. However, the matrix is transposed, in the sense that the languages run along the top rather than down the side. Accordingly, I write L languages (rather than N languages as before) and N etyma, as a way to intimate that ETYMDIST clusters by etyma rather than by languages.

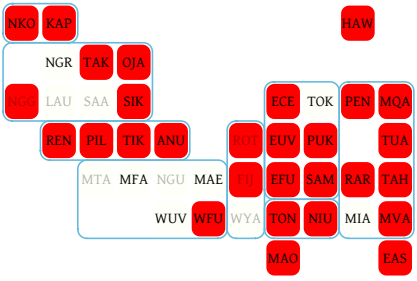
ETYMDIST overview

- Take observed distributions of etyma in POLLEX, e.g.:

*qariki ‘chief’



*qarofa ‘love, pity, compassion’



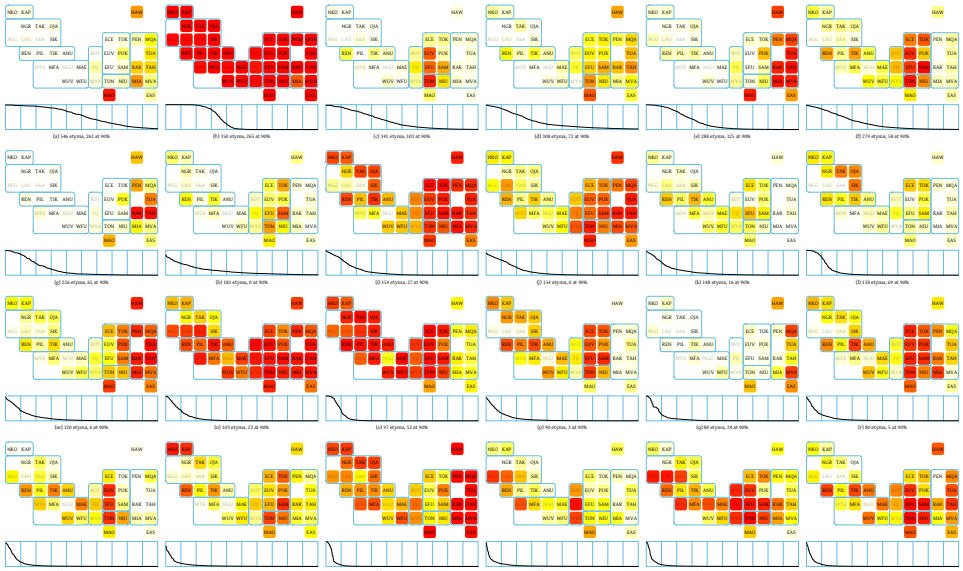
- Infer actual distributions, using some model of lexicographic coverage.
- Cluster etyma by posited actual distributions.

ETYMDIST clusters etyma by each etymon’s *distribution*, i.e. its pattern of presence or absence among the languages. We distinguish between *observed* distributions and *actual* distributions.

The observed distribution of an etymon is the set of languages for which the etymon is attested in POLLEX. In this slide I show the observed distributions of two etyma. The languages have been laid out roughly geographically.

The actual distribution of an etymon is a matter of inference, based on a simple model of the amount of lexicographic coverage received by each language. ETYMDIST will cluster etyma by what it infers their actual distributions to be.

ETYMDIST results



The results are shown here. Each plot is a cluster of etyma. The size of the cluster (i.e. the number of etyma that participate in the cluster) is given in the caption. The geographical extent of the cluster is shown by the map. A language that is colored red means that all of the etyma in the cluster are posited to exist in the language. A language that is colored orange-yellow means that half of the etyma in the cluster are posited to exist in the language.

The membership of each etymon is probabilistic: an etymon may be in a cluster with less than probability one. The histogram below each map shows the probability of membership for the thousand etyma with the highest probabilities of membership. A slow drop in the histogram, as in the first cluster, indicates that the category has “fuzzy boundaries”; whereas a sharp drop, as in the second cluster, indicates that it is relatively clear-cut which etyma belong to the cluster.

Before examining some of these clusters in detail, I will first describe how ETYMDIST works.

ETYMDIST rationales

Why not just use STRUCTURE?

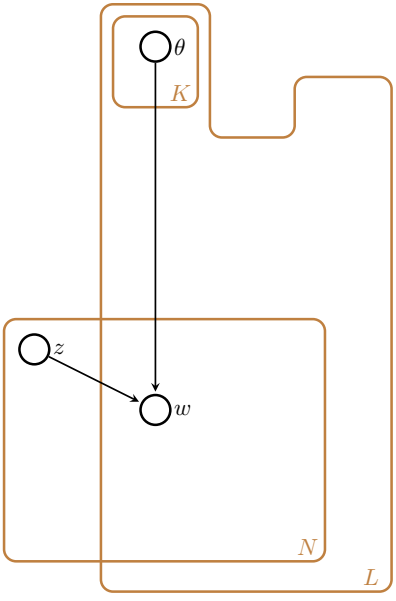
- Q: Why not just throw the binary matrix into STRUCTURE, treating etyma like features in SAPhon?
- A: Etyma presence is not homoplastic. Can the same etymon really arise from more than one ancestral population?
- Q: Why not use STRUCTURE, but treat each etymon as a SAPhon language, and each language as a SAPhon feature?
- A: Clustering with admixture is strictly more complex than ETYMDIST, which is pure clustering. Still, it could be worth trying.

In any case, we'd have to build into STRUCTURE the concept of lexicographic coverage.

I should start by explaining why I didn't just use STRUCTURE to analyze POLLEX. The answer is that I did not want to treat etyma the same ways as I treated phonemes in SAPhon, because I did not want an analysis that permitted etymon state to be homoplastic. That is, I did not want the model to allow an etymon to come from more than one ancestral population. This is quite different than how it is for phonemes — a good model has to posit /t/ in all or nearly all ancestral populations. But in the case of etyma, I wanted each etymon to be accounted for by a single ancestral population. This is why ETYMDIST clusters by etymon rather than by language.

One way to cluster by etymon would be to feed STRUCTURE a transposed data matrix, but I wanted to do simple clustering rather than clustering with admixture, in part because I could not think of a reasonable interpretation for clustering with admixture when the data matrix was transposed.

ETYMDIST



$\theta_{kl} \in (0, 1)$
Intensity of class k in language l.

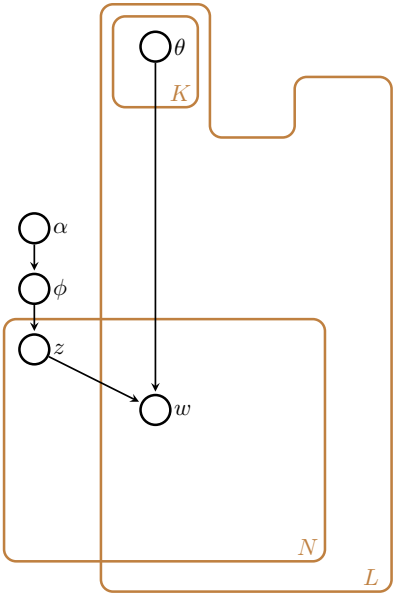
$z_n \in \{1, \ldots, K\}$
Class of eytmon n.

$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

This is the heart of the model. The variable z_n denotes the cluster to which etymon n belongs. Each cluster k is defined by a bank of *intensities* $\theta_{k1}, \ldots, \theta_{kL}$, one for each language l. The intensity θ_{kl} denotes the probability that an etymon of class k will exist in language l.

The variable w is an $N \times L$ binary matrix, with each entry w_{nl} denoting whether etymon n exists in language l. To generate w_{nl}, first look up the cluster of the etymon z_n, then look up the intensity of that cluster in that language ($\theta_{z_n l}$, i.e. θ indexed by z_n and l), and then perform a weighted coin toss.

ETYMDIST

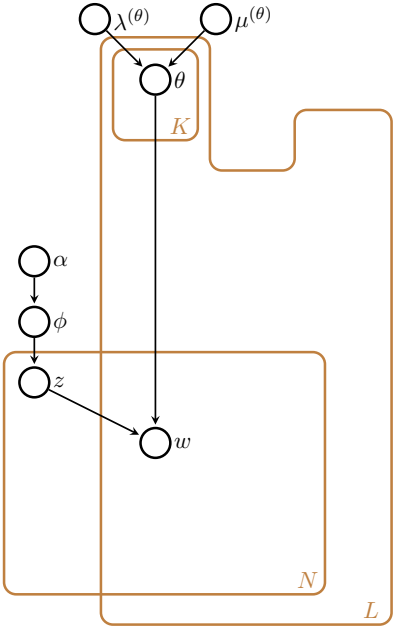


$\theta_{kl} \in (0, 1)$
Intensity of class k in language l.

$\alpha \in (0, \infty)$
Diversity parameter.
$\phi \in \Delta^{K-1}, \phi \sim \text{Dirichlet}(\alpha/K)$
Relative cluster size.
$z_n \in \{1, \ldots, K\},\ z_n \sim \text{Cat}(\phi)$
Class of eytmon n.
$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

The cluster variables z_n are generated via a categorical distribution parameterized by the relative cluster sizes ϕ. This in turn is drawn from a Dirichlet distribution parameterized by α/K. So that the number of clusters can be inferred from the data, K is set to infinity. (In the implementation of the model, z is generated by a Chinese restaurant process parameterized by α.)

ETYMDIST



$\lambda^{(\theta)} \in (0, \infty),\ \mu^{(\theta)} \in (0, 1)$
Hyperparameters for intensity.

$\theta_{kl} \in (0, 1),\ \theta_{kl} \sim \text{Beta}(\mu^{(\theta)}\lambda^{(\theta)}, (1 - \mu^{(\theta)})\lambda^{(\theta)})$
Intensity of class k in language l.

$\alpha \in (0, \infty)$
Diversity parameter.

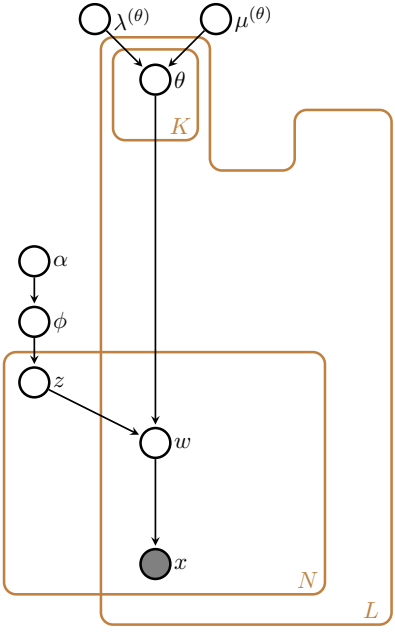
$\phi \in \Delta^{K-1}, \phi \sim \text{Dirichlet}(\alpha/K)$
Relative cluster size.

$z_n \in \{1, \ldots, K\},\ z_n \sim \text{Cat}(\phi)$
Class of eytmon n.

$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

The intensities are all generated from the same beta distribution, parameterized by $\lambda^{(\theta)}$ and $\mu^{(\theta)}$. I strongly suspect that there are easy ways to improve this prior, but I have yet to think of any.

ETYMDIST



$\lambda^{(\theta)} \in (0, \infty),\ \mu^{(\theta)} \in (0, 1)$
Hyperparameters for intensity.

$\theta_{kl} \in (0, 1),\ \theta_{kl} \sim \text{Beta}(\mu^{(\theta)}\lambda^{(\theta)}, (1 - \mu^{(\theta)})\lambda^{(\theta)})$
Intensity of class k in language l.

$\alpha \in (0, \infty)$
Diversity parameter.

$\phi \in \Delta^{K-1}, \phi \sim \text{Dirichlet}(\alpha/K)$
Relative cluster size.

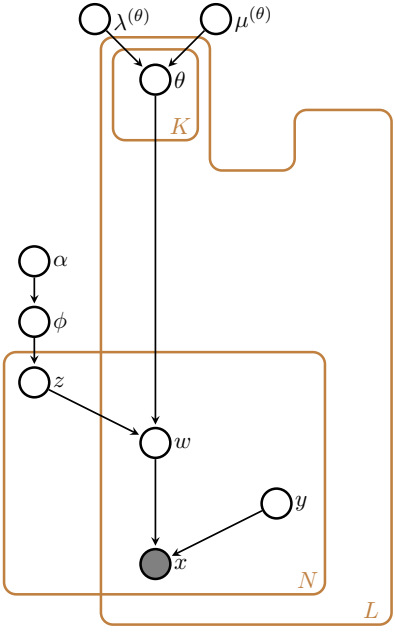
$z_n \in \{1, \ldots, K\},\ z_n \sim \text{Cat}(\phi)$
Class of eytmon n.

$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

$x_{nl} \in \{0, 1\}$
Is etymon n observed in language l?

The difference between actual and observed distributions is the difference between w and x. The variable x_{nl} indicates whether etymon n is attested in language l. It is a function of w_{nl}, but also of ...

ETYMDIST



$\lambda^{(\theta)} \in (0, \infty),\ \mu^{(\theta)} \in (0, 1)$
Hyperparameters for intensity.

$\theta_{kl} \in (0, 1),\ \theta_{kl} \sim \text{Beta}(\mu^{(\theta)}\lambda^{(\theta)}, (1 - \mu^{(\theta)})\lambda^{(\theta)})$
Intensity of class k in language l.

$\alpha \in (0, \infty)$
Diversity parameter.

$\phi \in \Delta^{K-1}, \phi \sim \text{Dirichlet}(\alpha/K)$
Relative cluster size.

$z_n \in \{1, \ldots, K\},\ z_n \sim \text{Cat}(\phi)$
Class of eytmon n.

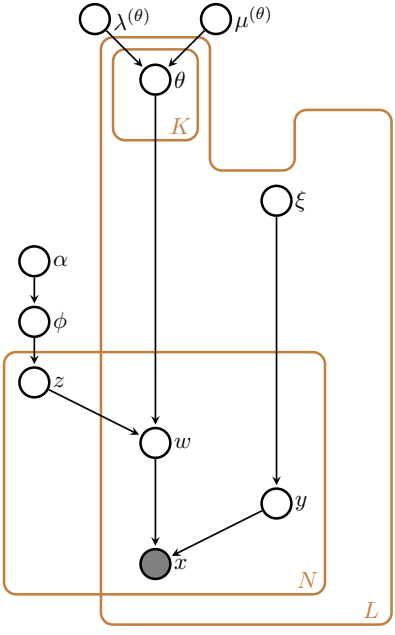
$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

$y_{nl} \in \{0, 1\}$
Observability of etymon n in language l.

$x_{nl} \in \{0, 1\},\ x_{nl} = w_{nl} \cdot y_{nl}$
Is etymon n observed in language l?

...y_{nl}, which indicates whether an etymon n in language l would be observable, if it were already to exist in language l.

ETYMDIST



$\lambda^{(\theta)} \in (0, \infty),\ \mu^{(\theta)} \in (0, 1)$
Hyperparameters for intensity.

$\theta_{kl} \in (0, 1),\ \theta_{kl} \sim \text{Beta}(\mu^{(\theta)}\lambda^{(\theta)}, (1 - \mu^{(\theta)})\lambda^{(\theta)})$
Intensity of class k in language l.

$\xi_l \in (0, 1)$
Coverage for language l.

$\alpha \in (0, \infty)$
Diversity parameter.

$\phi \in \Delta^{K-1}, \phi \sim \text{Dirichlet}(\alpha/K)$
Relative cluster size.

$z_n \in \{1, \ldots, K\},\ z_n \sim \text{Cat}(\phi)$
Class of eytmon n.

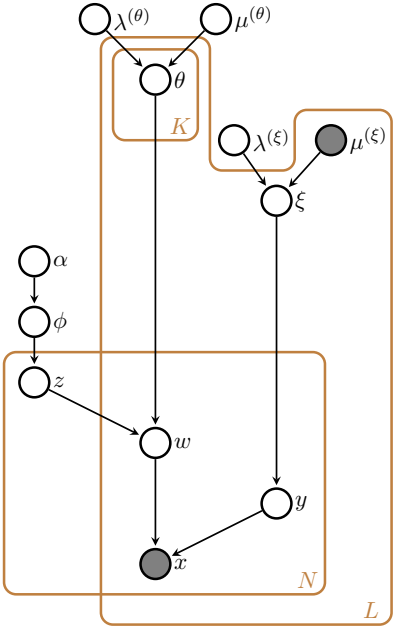
$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

$y_{nl} \in \{0, 1\},\ y_{nl} \sim \text{Bernoulli}(\xi_l)$
Observability of etymon n in language l.

$x_{nl} \in \{0, 1\},\ x_{nl} = w_{nl} \cdot y_{nl}$
Is etymon n observed in language l?

Each y_{nl} is derived via a Bernoulli distribution parameterized by the lexicographic coverage for language l, ξ_l.

ETYMDIST



$\lambda^{(\theta)} \in (0,\infty),\ \mu^{(\theta)} \in (0,1)$
Hyperparameters for intensity.

$\theta_{kl} \in (0,1),\ \theta_{kl} \sim \text{Beta}(\mu^{(\theta)}\lambda^{(\theta)}, (1-\mu^{(\theta)})\lambda^{(\theta)})$
Intensity of class k in language l.

$\lambda^{(\xi)} \in (0,\infty),\ \mu_l^{(\xi)} = 0.9N_l/\max\{N_1, N_2, \ldots, N_L\}$
N_l = #entries for language l.

$\xi_l \in (0,1),\ \xi_l \sim \text{Beta}(\mu_l^{(\xi)}\lambda^{(\xi)}, (1-\mu_l^{(\xi)})\lambda^{(\xi)})$
Coverage for language l.

$\alpha \in (0,\infty)$
Diversity parameter.

$\phi \in \Delta^{K-1}, \phi \sim \text{Dirichlet}(\alpha/K)$
Relative cluster size.

$z_n \in \{1,\ldots,K\},\ z_n \sim \text{Cat}(\phi)$
Class of eytmon n.

$w_{nl} \in \{0,1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

$y_{nl} \in \{0,1\},\ y_{nl} \sim \text{Bernoulli}(\xi_l)$
Observability of etymon n in language l.

$x_{nl} \in \{0,1\},\ x_{nl} = w_{nl} \cdot y_{nl}$
Is etymon n observed in language l?

It stands to reason that ξ_l correlates positively with N_l, the number of forms in POLLEX for language l. This correlation is encoded via a beta distribution. The degree of correlation is encoded in the hyperparameter $\lambda^{(\xi)}$.

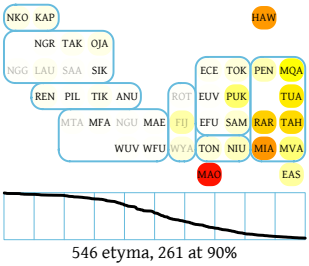
Note 1: One reason that the correlation is imperfect, is that for non-Polynesian languages in POLLEX, the lexicographic coverage is much higher than N_l would suggest. The reason is that POLLEX selectively contains etyma that appear in Polynesian languages, which artificially limits the N_l for non-Polynesian languages.

Note 2: Supplying the model with some knowledge of N_l seemed critical for getting the model to learn reasonable values for ξ_l; or perhaps I did not try hard enough to get it to work without N_l.

Hawaiian

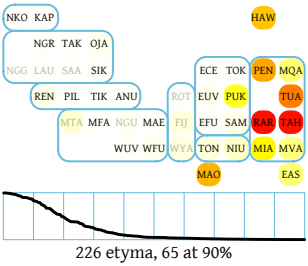
- Traditionally Hawaiian is classified with MQA and MVA as a Marquesic language, but lexically it is very mixed.
- Hawaiian may have participated in up to three Eastern Polynesian dialect chains.

Southern chain



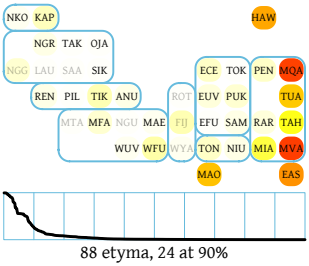
546 etyma, 261 at 90%

Northern chain



226 etyma, 65 at 90%

Eastern chain

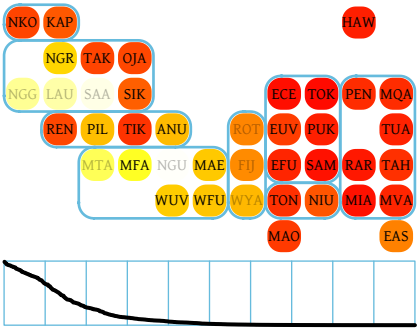


88 etyma, 24 at 90%

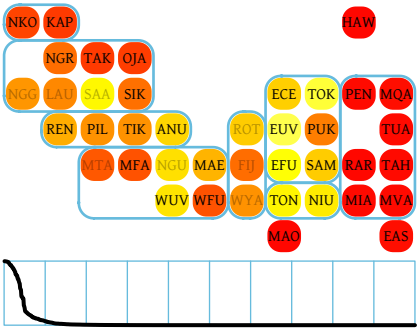
And now for a rundown of some interesting clusters. Hawaiian is conventially classified as subgrouping with Marquesan (MQA) and Mangarevan (MVA), but these clusters suggest that Hawaiian participated in three Eastern Polynesian dialect chains. Moreover, Hawaiian was not the only language to do so: Tuamotoan, and perhaps Tahitian, seem to be part of all three as well.

Northern Outliers

- Polynesian Outliers are geographically Micronesian or Melanesian.
- The Northern Outliers may form a genetic subgroup with Eastern Polynesian languages.



159 etyma, 27 at 90%



47 etyma, 20 at 90%

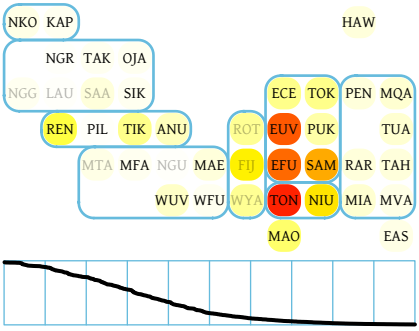
The cluster on the right shows etyma that were retained in the Northern Polynesian Outliers (NKO, KAP, NGR, TAK, OJA, SIK) and Eastern Polynesia, but were mostly absent in Western Polynesia. This pattern may have come about due to genetic descent: there is comparative evidence that the Northern Outliers and the Eastern Polynesian languages share a common ancestor to the exclusion of other languages (Wilson, 1985).

The cluster on the left looks similar to the one of the right, but now the etyma are robustly present in Western Polynesia. My guess is that it consists of etyma that were widespread in Polynesia, but were partially lost in the Southern Polynesian Outliers of Vanuatu and New Caledonia (MFA, MAE, WUV, WFU).

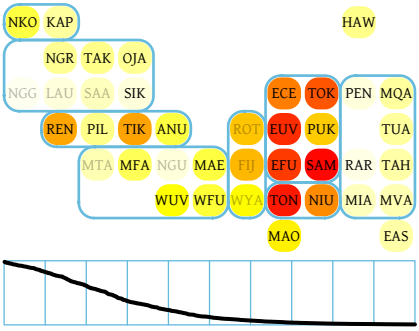
It is a bit unsatisfactory that two clusters that look so similar should receive such different interpretations, but I lack a better explanation for these results.

Lexical borrowing in Western Polynesia

- Polynesian splits into Tongic (TON & NIU) and Nuclear Polynesian (the rest).
- These patterns reflect borrowing between the two subgroups in geographical Western Polynesia.



341 etyma, 103 at 90%

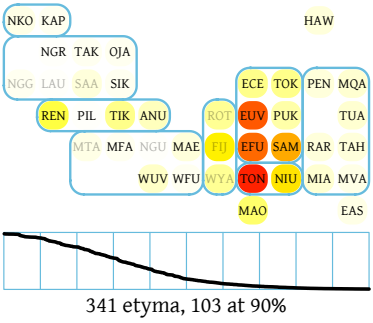


279 etyma, 58 at 90%

Western Polynesia (consisting of ECE, TOK, EUV, PUK, EFU, SAM, TON, NIU) has been a locus of long-standing cultural exchange. These two clusters show etyma that have diffused throughout the region as a consequence of that exchange. One could alternatively argue that these clusters consist of Proto Polynesian etyma that were conserved in the Western Polynesian languages, but lost elsewhere; but this is not plausible when the etyma number in the hundreds.

Most stable cluster members

- Definition: an etymon's degree of membership is its posterior probability of being in the cluster.
- Show the most stable members, which all have degree of membership > 0.97.



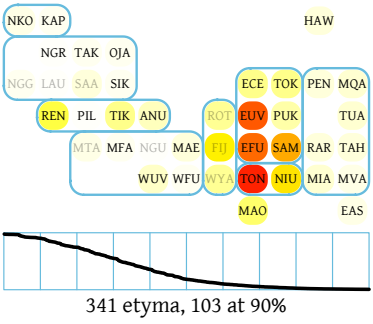
341 etyma, 103 at 90%

*kou$_2$	'Uncastrated pig'
*efe	'Refuse of grated coconut or kava'
*liku$_2$	'Short, of a woman's skirt'
*mapa	'A kind of tree (Diospyros or Maba sp.)'
*kea$_3$	'Breadfruit'
*fauqigo	'Hibiscus sp.'
*kita$_{2b}$	'Relapse following an illness'
*kaka$_2$	'Deceive, cheat'

With ETYMDIST, it is possible to list the best (most probable) members of each cluster. I have done so for the smaller Western Polynesian cluster. The glosses of these etyma do not seem to be particularly significant, though with more investigation, a pattern may emerge.

Most stable cluster members

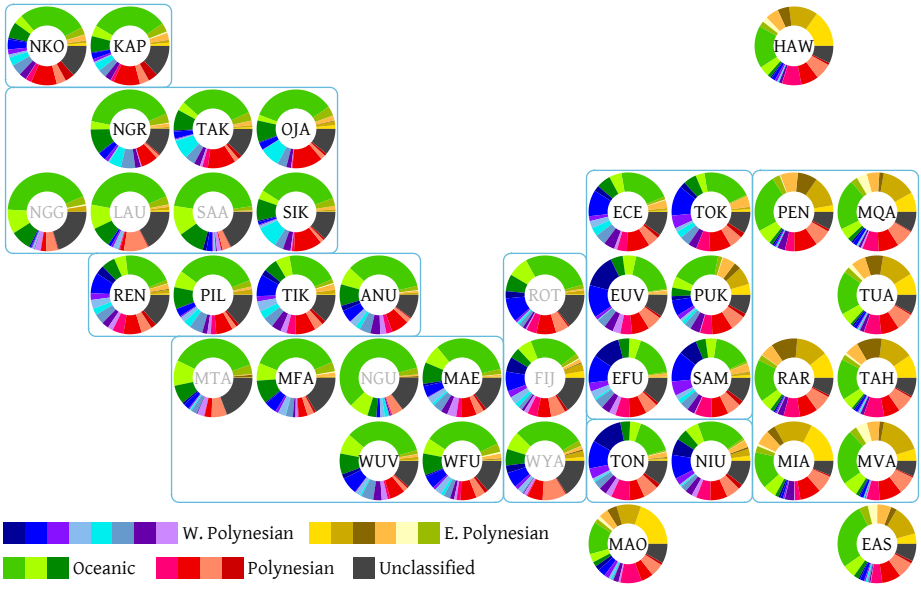
- There are ~143 members attested in TON and SAM.
- Show the most stable members, which all have a degree of membership > 0.91.



341 etyma, 103 at 90%

*fakavaka	'A handle, provide with a handle'
*fesikiqaki	'(Ex)change places'
*munua	'A fish'
*fakameomeo	'Displeased'
*fakakaukau	'Consider carefully'
*kalia	'Double canoe'
*sela	'Asthma; gasp for breath'
*takafalu	'A small tree (Micromelum' minutum)'

Here are etyma from the same cluster, but I have restricted the list to etyma attested in both Tongan (TON) and Samoan (SAM). Tongan-Samoan loanwords are hard to detect because the sound correspondences are such that it is always possible to reconstruct a protoform in a common ancestor. A handful of loans have been identified by cognate set distribution, parallel semantic shifts, or unexpected phonological correspondences (Marck, 2000, p. 69; Pawley, 2009), but this cluster provides evidence that there was a massive amount of word-borrowing between the two societies.



This plot is a visual experiment: it shows what happens when the clusters are collapsed together into one map. The 21 largest clusters are given distinct colors. I use shades of blue for clusters centered in Western Polynesia; shades of yellow for those in Eastern Polynesia; shades of green for clusters that have high intensities in non-Polynesian languages; and shades of red for clusters that are widespread in Polynesian languages, but not in other Oceanic languages. The remaining clusters are lumped together as dark gray.

I do not know how informative this plot really is, but it does make the point that Pukapukan (PUK) has a very admixed vocabulary, with as many elements from the East as are from the West.

Bibliography I

Biggs, Bruce & Ross Clark. 2006. POLLEX: The comparative Polynesian lexicon project. Computer file, University of Auckland.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. Springer, New York.

Marck, Jeff. 2000. *Topics in Polynesian Language and Culture History*. Pacific Linguistics.

Michael, Lev, Tammy Stark & Will Chang (compilers). 2012. South American Phonological Inventory Database v1.1.1. Survey of California and Other Indian Languages Digital Resource. Berkeley: University of California.

Pawley, Andrew. 2009. Polynesian paradoxes: Subgroups, wave models and the dialect geography of proto-Polynesian. Paper presented at 11th International Conference on Austronesian Linguistics, Aussois, France, June 2009.

Bibliography II

Pritchard, Jonathan K., Matthew Stephens & Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2). 945--959.

Reesink, Ger, Ruth Singer & Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biology* 7(11).

Teh, Y.W., M.I. Jordan, M.J. Beal & D.M. Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association* 101(476). 1566--1581.

Wilson, William H. 1985. Evidence for an Outlier source for the Proto Eastern Polynesian pronominal system. *Oceanic Linguistics* 85--133.

STRUCTURE as applied to phonological inventories

STRUCTURE: Given a sample of N specimens, find K clusters, with admixture.

Domain:	Population genetics N specimens L genetic loci M-fold ploidy J alleles per locus	Phonological inventories N languages L features Features are absent/present.
Given:	$X \in \{1, \ldots, J\}^{N \times L \times M}$	$X \in \{1, 0\}^{N \times L}$
Infers:	$P \in (\Delta^{J-1})^{K \times L}$ For each of K ancestral populations, what is the frequency of each allele at each locus? $Q \in (\Delta^{K-1})^{N}$ For each of N specimens, what fraction of its alleles derive from each ancestral population?	$P \in [0, 1]^{K \times L}$ For each of K ancestral inventories, What is the frequency of each feature? $Q \in (\Delta^{K-1})^{N}$ For each of N languages, what fraction of its features derive from each ancestral inventory?

Modeling ideals

- Responsiveness to data: avoid hard *a priori* assumptions.
  - E.g., let the number of clusters be inferrable from the data.
  - E.g., account for incompleteness in the data.
- Interpretability: real-world interpretations for model elements.
  - Have units whenever possible.
  - Make bold predictions.
- Transparency: avoid treating models as black boxes.
  - E.g., what did this ancestral population look like?
  - E.g., what words did language X borrow from language Y?