The Inner-Outer hypothesis of Indo-Aryan: a computational study

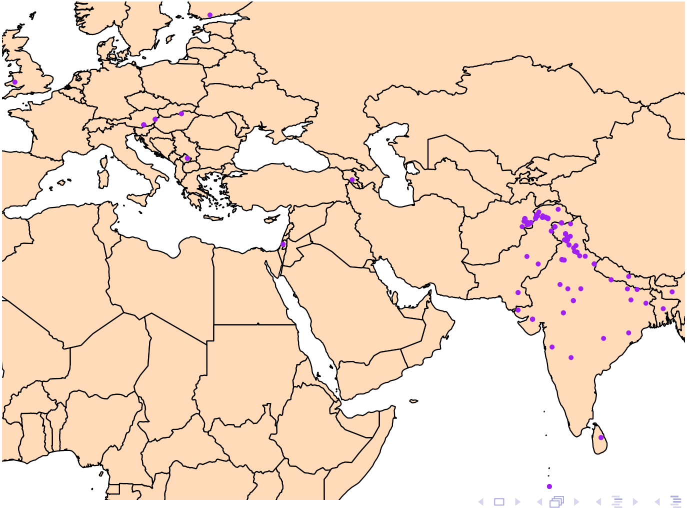
Chundra Aroor Cathcart
Department of Comparative Linguistics
University of Zurich

SALA 34 University of Konstanz 13 April 2018

Introduction

- The broad purpose of this research in progress: enhance the comparative method with probabilistic tools
- Comparative method: tried and tested in historical linguistics; however, it hits a wall when there is a great deal of noise (caused by contact, etc.)
- If exceptions to regular sound change are few in number, they can be dealt with on a piece-by-piece basis
- What do we do in the face of substantial uncertainty?
- Scope of this study: use a probabilistic methodology to address an unresolved hypothesis regarding the history of the Indo-Aryan languages

Indo-Aryan languages



Indo-Aryan languages I

- Generally agreed that virtually all Indo-Aryan speech varieties descend directly from attested Old Indo-Aryan, though it is possible to encounter one-off, highly sporadic instances of what appear to be archaisms
  - some MIA languages famously reflect a zero-grade form of the mediopassive participial suffix *-mīna-* < PIE **-mh_1no-*, against OIA *-māna-* < PIE **-meh_1no-* (von Hinüber 2001:301)
  - Additionally, while various voiced PIE clusters fell together as *kṣ* in Sanskrit (e.g., *$d^h g^{u̯h}$*er-* > OIA *kṣar-* versus Avestan *ɣžar-* 'flow'), some MIA languages show voiced reflexes of these clusters, e.g., Pali *paggharati*, Prakrit *jharaï*, though it is not clear whether this voicing is indeed archaic or secondary

Indo-Aryan languages II

  - "Dardic" languages of N. Pakistan/Afghanistan/NW India show curious behavior; minority view (Kogan 2005) argues that they are non-Indo-Aryan
  - Also notable is the Bangani language, which shows so-called *centum* treatment of the PIE palatovelars in a large portion of its vocabulary, as opposed to the *satem* treatment expected in Indo-Iranian. This behavior has been accounted for by some scholars by assuming that a population speaking a *centum* Indo-European dialect shifted to an Indo-Aryan speech variety and continued to use vocabulary items displaying centum treatment (Smith 2017)
- However, despite this general consensus, there is no agreement on subgrouping within Indo-Aryan

Indo-Aryan subgrouping hypotheses

- Hoernle 1880: proposes four groups nested within two higher-order groups
- Grierson (*LSI et seq*): continuation of H; argues for an Outer (peripheral) and Inner (core) group of languages on the basis of shared features
- Chatterji 1926: argues against all features proposed by G
- Zograf 1976: finds no evidence for Inner-Outer hypothesis
- Masica 1991: somewhat agnostic
- Southworth 2005: proponent of the Inner-Outer hypothesis

Problems for the comparative method in I-A I

Contact is arguably the biggest problem; dialectal variation is found at all chronological stages

- Old Indo-Aryan (OIA) period:
  - Oral transmission of oldest texts; vernacularisms imposed, e.g., _śithirá-_ 'loose' < _*śr̥thirá-_ (Tedesco 1960, Elizarenkova 1989)
  - Regional variation (Witzel 1989)
- Middle Indo-Aryan (MIA) period:
  - Attested MIA languages show striking similarity in terms of the changes they have undergone
  - Radical simplification of consonant clusters; operation of the "2-mora rule"

Problems for the comparative method in I-A II

  - However, some variation: Pali shows a conservative layer where consonant clusters are repaired via epenthesis rather than assimilation (e.g., Pa. *palavati* ∼ *pavati*, Pkt. *pavaï* < OIA *plávate* 'swims'; cf. Oberlies 2001:112–3). It also displays different instantiations of the 2-mora rule, e.g., CV̄C ∼ CVC (Hock 2016:33)
- New Indo-Aryan (NIA) period:
  - Generally assumed that non-peripheral languages of South Asia such as Hindi, Panjabi, Bengali, etc., all descend from speech varieties akin to those attested during the MIA period
  - At this point, many languages have begun to take on drastically individual characteristics, such as the Assamese sound change to *x* from the sibilant of its MIA predecessor *š

Problems for the comparative method in I-A III

- In particular, the singleton and geminate consonants brought about by the two-mora rule have been subjected to additional changes. Hindi, on one hand, typically undergoes the merger VCC, $\bar{\text{V}}$C > $\bar{\text{V}}$C, while Panjabi generally goes in the opposite direction, going as far as to assimilate non-IA words into this template (e.g., the allegro pronunciation [pənɟəbːi] ∼ [pənɟabi], originally a Persian loan)

- These problems have been acknowledged by contemporary scholars (e.g., Masica 1991:460): “Perhaps a wiser course would be to recognize a number of overlapping genetic zones,each defined by specific criteria.”
- At the same time, many have had success with the comparative method for smaller problems in I-A (cf. Toulmin 2009)

Southworth 2005

- Southworth rejects what he sees as Masica's fatalism, and attempts to update and revive Grierson's Inner-Outer hypothesis
- S adduces evidence that Grierson did not or was unable to consider
- Makes the languages associated with each group more or less explicit

S's evidence I

- Morphosyntactic isoglosses uniting Outer group
  - Future marker developed from OIA gerundive suffix *-tavya-*
  - Past tense in *-l-*
- Phonological isoglosses
  - OIA vocalic *r̥* is realized primarily as *a* in the Outer group and *i* in the Inner group, though reflexes are not entirely regular
  - Builds on proposal of Turner (1916) that inner languages placed stress on the rightmost non-final heavy syllable, whereas outer languages had fixed stress on the initial syllable

S's evidence II

- $l > n$ in Outer languages; S admits that $l > n$ changes in the Outer group are not a unified phenomenon, in terms of conditioning environments, but concludes nevertheless that the developments are unlikely to be independent of each other
- Loss of non-initial post-consonantal *h* in Outer group
- Phonemic distinction between $u/\bar{u}$ and $i/\bar{\imath}$ neutralized in Outer group
- Stress on rightmost non-final heavy syllable in Inner languages versus fixed stress on initial syllable in Outer languages (after Turner 1916)

S's evidence III

- A potential isogloss unnoticed by S: if Emeneau is correct that “it is not ruled out that [Marathi use of *jaṇa*] owes its inception to some stimulus ultimately deriving from the full-fledged [numeral classifier] system of Magadhan”, then the presence a classifier-like construction derived from *jana* may unite Outer languages (cf. Sinhala *denā*, Assamese *zan*, etc.)

- S has selected evidence that he considers probative with respect to this hypothesis, namely evidence that he believes to be sufficiently archaic as to reflect the division between the two ancient groups.
- Other evidence is arguably not probative: it doesn't matter that Hindi and Panjabi, both members of the Inner group, differ in their treatment of VCC and $\bar{\text{V}}$C sequences, as these developments are conceivably of a post-MIA date.
- Alternations between *r* and *l*, on the other hand, can be found in the earliest chronological period of I-A; however, this does not necessarily mean that all such changes are old
- It is not clear if every sound change considered deserves the diagnostic power ascribed to it by S.

- Even if the objections given above are unreasonable, the evidence supplied is still arguably selective. Although S envisions two sociolinguistic groups that experienced greater intra-group than inter-group communication, he also concedes that this integrity broke down at a later date, prior to the changes that began to distinguish NIA speech varieties from each other.
- Nitpicking regarding minor details aside, the innovations presented are convincing; S's approach is thorough and far from cavalier. However, arguments worded with probabilistic language (p. 148: "the number of detailed similarities makes independent innovation unlikely") call for explicit probabilistic methodologies. We require a means of allocating credibility to the Inner-Outer hypothesis using a large body of data that has not been selected by hand.

- I take to heart (Masica's 1991:457) observation, however qualified, that one "non-arbitrary way of [establishing dialect groups] might appear to lie in giving priority to phonology"; by using a large number of Modern IA forms extracted from Turner's (1962–66) Comparative Dictionary of the Indo-Aryan Languages and a probabilistic methodology, we hope to alleviate some of the woes that the author lists on the same page (such as the problem of widespread dialect admixture).

Rationale I

- Extract information regarding sound change from a large data set
- Use a Bayesian mixed-membership model to infer associations between particular sound changes and latent dialectal “components” representing the Inner and Outer group
    - Assumption: language contact happens at the word level (i.e., via lexical borrowing)
- Assess correlation between geography and a language’s association with the Inner/Outer group
- Assess amount of inter-group communication; are individual languages associated uniformly with each component, or are there skews?

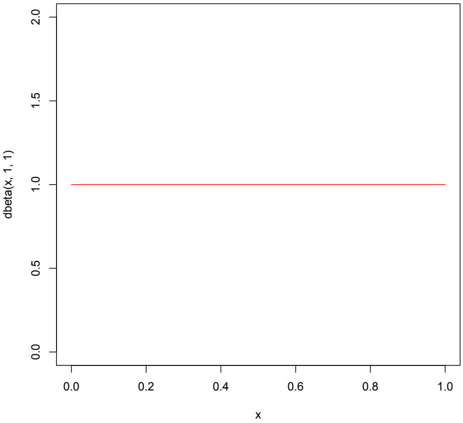
Bayesian statistics in 30 seconds

- Fundamental goals of Bayesian inference:
  - Allocate credibility to hypotheses, usually in the form of a distribution (with some degree of uncertainty)
  - Incorporate prior knowledge (when needed)
- Key concepts
  - **Prior** distribution over hypotheses
  - **Likelihood**: probability that a given hypothesis
  - **Posterior** distribution over hypotheses: posterior probability of a hypothesis: proportional to prior times likelihood

I have a coin, and want to know the distribution over probabilities that the coin comes up heads

Prior

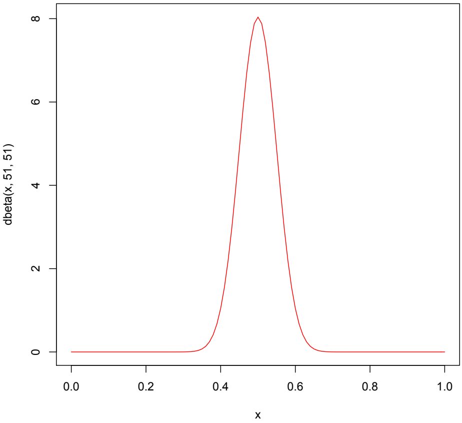
I start with an uninformative prior: all probabilities (e.g., hypotheses/parameter values) are equally probable



Likelihood

I flip the coin 100 times and see 50 heads

Posterior



- Posterior distribution is peaked at the Maximum a posterior (MAP) value.

This is a trivial and uninteresting example: often we want to find the MAP values of many interacting parameters

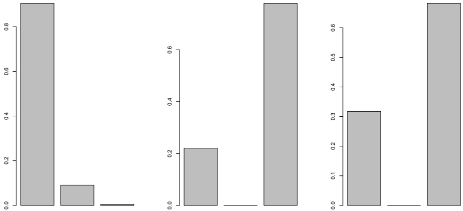
Why priors matter: the Dirichlet distribution I

- Multinomial distribution: probabilities of N events, all summing to one
- The probability of a sound change (or non-change) in a speech community can be seen as a multinomial probability: P(s > s) vs. P(s > h) vs. ... in some conditioning environment (pace Neogrammarians)
- But sound change cannot be characterized by **any** multinomial distribution, but rather **sparse** distributions, where one outcome gets most of the probability mass: e.g., $P(s > s)=.98$, $P(s > h)=.02$
- This should be contrasted with **smooth** distributions, where probabilities are almost uniform across events
- The Dirichlet is a popular prior distribution for the multinomial distribution

Why priors matter: the Dirichlet distribution II

- N-length vector of **concentration** parameters
- In a symmetric Dirichlet (assumed henceforth), all concentration parameters are equal
- If concentration parameter is greater than one, smooth multinomials are generated
- If less than one, sparse multinomials are generated

Random draws from Dirichlet(.1,.1,.1)



- Does not matter which outcome gets the most probability mass, only that one outcome gets the most probability mass!

Bayesian generative model

We tell a story about how our data were generated:

Bayesian generative model

We tell a story about how our data were generated:

- For each dialect group $k \in \{1, ..., K\}$ (K=2)

Bayesian generative model

We tell a story about how our data were generated:

- For each dialect group $k \in \{1, ..., K\}$ (K=2)
    - $\phi^k \sim$ Dirichlet($\alpha < 1$) [draw a group-level collection of sparse Multinomial distributions over sound changes]

Bayesian generative model

We tell a story about how our data were generated:

- For each dialect group $k \in \{1, ..., K\}$ (K=2)
  - $\phi^k \sim$ Dirichlet($\alpha < 1$) [draw a group-level collection of sparse Multinomial distributions over sound changes]
- For each language $l \in \{1, ..., L\}$

Bayesian generative model

We tell a story about how our data were generated:

- For each dialect group $k \in \{1, ..., K\}$ (K=2)
  - $\boldsymbol{\phi}^k \sim$ Dirichlet($\alpha < 1$) [draw a group-level collection of sparse Multinomial distributions over sound changes]
- For each language $l \in \{1, ..., L\}$
  - $\boldsymbol{\theta}^l \sim$ Dirichlet(β) [draw a language-level distribution over dialect group makeup]

Bayesian generative model

We tell a story about how our data were generated:

- For each dialect group $k \in \{1, ..., K\}$ (K=2)
    - $\phi^k \sim$ Dirichlet($\alpha < 1$) [draw a group-level collection of sparse Multinomial distributions over sound changes]
- For each language $l \in \{1, ..., L\}$
    - $\boldsymbol{\theta}^l \sim$ Dirichlet(β) [draw a language-level distribution over dialect group makeup]
    - For word w_i^l in language l

Bayesian generative model

We tell a story about how our data were generated:

- For each dialect group $k \in \{1, ..., K\}$ (K=2)
  - $\phi^k \sim \text{Dirichlet}(\alpha < 1)$ [draw a group-level collection of sparse Multinomial distributions over sound changes]
- For each language $l \in \{1, ..., L\}$
  - $\boldsymbol{\theta}^l \sim \text{Dirichlet}(\beta)$ [draw a language-level distribution over dialect group makeup]
  - For word w_i^l in language l
    - $z_i^l \sim \text{Cat}(\boldsymbol{\theta}^l)$ [choose a dialect group from which the word comes]

Bayesian generative model

We tell a story about how our data were generated:

- For each dialect group $k \in \{1, \ldots, K\}$ (K=2)
    - $\phi^k \sim \text{Dirichlet}(\alpha < 1)$ [draw a group-level collection of sparse Multinomial distributions over sound changes]
- For each language $l \in \{1, \ldots, L\}$
    - $\boldsymbol{\theta}^l \sim \text{Dirichlet}(\beta)$ [draw a language-level distribution over dialect group makeup]
    - For word w_i^l in language l
        - $z_i^l \sim \text{Cat}(\boldsymbol{\theta}^l)$ [choose a dialect group from which the word comes]
        - For each sound change $\sigma_{i,j}^l$

Bayesian generative model

We tell a story about how our data were generated:

- For each dialect group $k \in \{1, ..., K\}$ (K=2)
  - $\phi^k \sim$ Dirichlet($\alpha < 1$) [draw a group-level collection of sparse Multinomial distributions over sound changes]
- For each language $l \in \{1, ..., L\}$
  - $\boldsymbol{\theta}^l \sim$ Dirichlet(β) [draw a language-level distribution over dialect group makeup]
  - For word w_i^l in language l
    - $z_i^l \sim$ Cat($\boldsymbol{\theta}^l$) [choose a dialect group from which the word comes]
    - For each sound change $\sigma_{i,j}^l$

      $\sigma_{i,j}^l \sim$ Cat($\phi^{z_i^l}$) [generate the sound change according to z_i^l]

Bayesian generative model

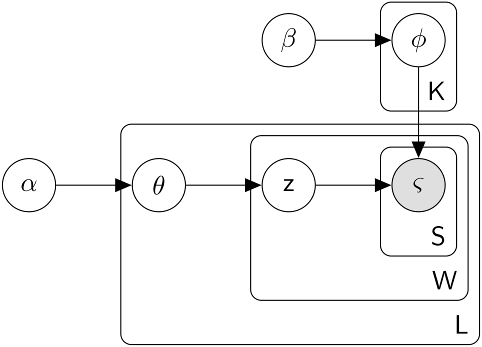
We tell a story about how our data were generated:

- For each dialect group $k \in \{1, ..., K\}$ (K=2)
  - $\phi^k \sim$ Dirichlet($\alpha < 1$) [draw a group-level collection of sparse Multinomial distributions over sound changes]
- For each language $l \in \{1, ..., L\}$
  - $\boldsymbol{\theta}^l \sim$ Dirichlet(β) [draw a language-level distribution over dialect group makeup]
  - For word w_i^l in language l
    - $z_i^l \sim$ Cat($\boldsymbol{\theta}^l$) [choose a dialect group from which the word comes]
    - For each sound change $\sigma_{i,j}^l$

      $\sigma_{i,j}^l \sim$ Cat($\phi^{z_i^l}$) [generate the sound change according to z_i^l]

(Very similar to Latent Dirichlet Allocation)

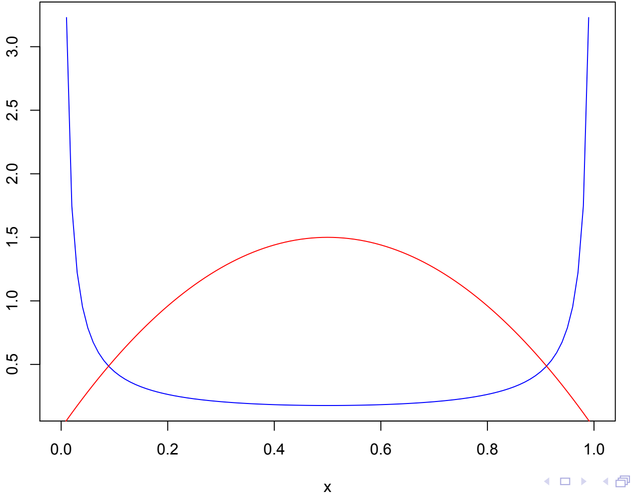
Plate diagram



- Prior distribution over a language's component membership is Dirichlet distributed (technically Beta distributed, since K=2)
- Recall that a symmetric Dirichlet distribution has one parameter, the concentration parameter
- We call the concentration parameter of the prior over languages' component membership β and treat it as an unknown quantity whose posterior distribution we wish to infer
- If $\beta > 1$, the distribution is "smooth"; if $\beta < 1$, the distribution is "sparse"

Visualization in 2-D

blue=.1; red=2



Operationalization of I/O hypothesis

If the Inner/Outer hypothesis as formulated by S is valid, then $\beta < 1$ is expected

Parameter inference

- Specifying a generative model defines parameters of interest
- We wish to infer posterior distributions for these parameters (crucially, we want to infer the posterior distribution of β, and it would be nice to have other information as well)
- Many MCMC algorithms are easy to implement, but inefficient
- Probabilistic programming languages such as Stan, PyMC3, and Edward allow specialists to specify the generative model, then use efficient algorithms to fit the parameters of interest
- PyMC3 used for this study

Data: Automated sound change extraction I

- I attempt to extract sound changes operating between Sanskrit and modern I-A languages (Dardic and Romani excluded, as in S 2005)
- Sanskrit words and their modern descendants are extracted from a digitized version of Turner (1962–66)
- Orthography normalized (OIA kṣ treated as one segment); morphological mismatches accounted for (e.g., verbal endings stripped)
- Segments in word pairs must be properly aligned; modified version Needleman-Wunsch based on List (2012), Jäger (2013) used; e.g., Sanskrit /aːntra/ 'entrails' > Nepali /aːn-ro/
- Aligned sequences used to extract sound changes as rewrite rules, e.g., Sanskrit t > Nepali ∅ / n _ r

Data: Automated sound change extraction II

- Many phenomena that this approach **cannot** capture (e.g., metathesis, multiple intermediate layers)
- Insensitive to telescoped changes:
  - Assamese /x/, the reflex of OIA *s, ś, ṣ*, is thought to develop from intermediate **ś* (Kakati 1941:224)
  - Marathi change *ch* > *s* affects certain words containing MIA **ch* < OIA *kṣ* as well as OIA *ch*
  - Some instances of Hindi *bh* < **mh* (cf. Oberlies 2005:48)
- At present most economic approach

- Potential problems with Dirichlet prior over sound changes: no explicit way to model correlation
  - Insensitive to similarity between outcomes: we may wish to treat ʃ and x as more similar to each other than to s
  - Cannot capture trends across non-identical sounds sharing many phonological features (e.g., would struggle to model Grimm's law)
- Many of these issues discussed in work of A. Bouchard-Côté and colleagues

Inference

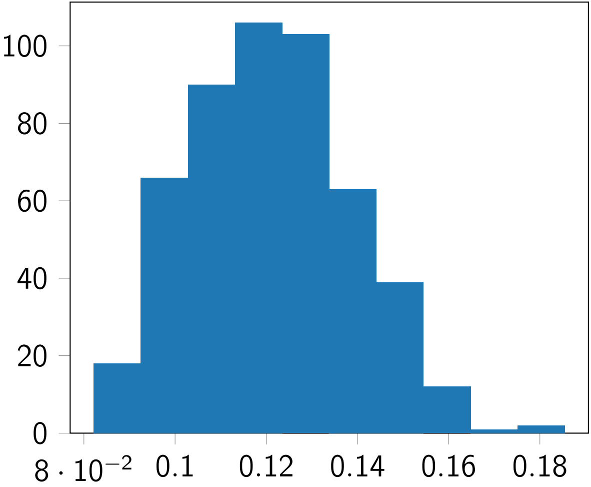
- Automatic differentiation variational inference (ADVI) as implemented in PyMC3 used to infer posterior distributions of parameters; mean-field ADVI approximates the posterior distribution of each parameter of interest by fitting a spherical Gaussian distribution to the true posterior distribution
- I initially wished to take into account all sound changes in data set; this was computationally infeasible
- Linguistically informed approach: all sound changes affecting OIA i, iː, u, uː, v, j, kʂ, l, n, ɳ, ʂ, s, ɕ that occur more than 5 times across data set (currently do not have stress marked in OIA)
- Still a large data set: 25573 words, 1149 sound changes

Implementation

- For relatively simple parameters, PyMC3 can interpret a model specification that looks exactly like the generative model we defined above
- This model however has a large number of free parameters, some of them discrete; because ADVI is gradient-based (i.e., derivatives are calculated to steer inference in the direction of high posterior probability), it cannot sample discrete variables; hence discrete variables must be marginalized out (summed over); this is more efficient

Results: β

Are posterior values for β greater or less than one? Are language-level distributions over dialect components smooth or sparse?



Results: sound change distributions

Cannot show results for all 1149 changes; some ones that distinguish strongly across dialect groups (posterior means given):

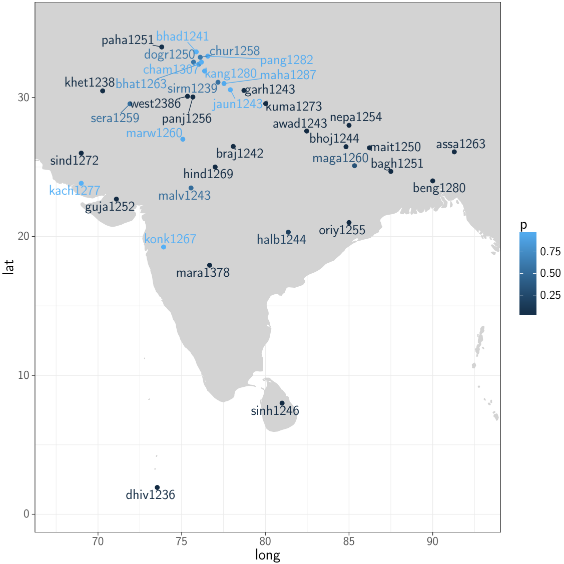
change	component 1 prob	component 2 prob
ʃ > s / # _ i	0.01814	0.9958
ʃ > ʃ / # _ i	0.9434	0.0012
ʃ > x / # _ i	0.0123	0.0010
ʃ > c^{h} / # _ i	0.0109	0.0008
ʃ > h / # _ i	0.0150	0.0009
ʃ > s / # _ ʊ	0.0532	0.9938
ʃ > ʃ / # _ ʊ	0.9467	0.0061
n > ɴ / a _ d^{ɦ}	0.0198	0.9938
n > ∅ / a _ d^{ɦ}	0.0164	0.0017
n > n / a _ d^{ɦ}	0.9530	0.0024
n > n^{ɦ} / a _ d^{ɦ}	0.0106	0.0020

Other changes do not cut as cleanly across the two groups as hoped:

change	component 1 prob	component 2 prob
kʂ > k^h / # _ u	0.4690	0.9933
kʂ > c^h / # _ u	0.5309	0.0066
n > n / a _ aː	0.3372	0.9676
n > ɳ / a _ aː	0.6627	0.0323
uː > u / p _ t	0.3566	0.6141
uː > uː / p _ t	0.3130	0.2266
uː > o / p _ t	0.3303	0.1592

- Somewhat surprisingly, the Dirichlet distribution is better at capturing correlated developments of specific sounds within a given dialect group (e.g., ʃ changes overwhelmingly to s in component 2, regardless of conditioning environment)
- Strong associations picked up within and across words

Results: geographic distribution I



Results: geographic distribution II

- No real core-periphery pattern produced
- Some disturbing results: Sindhi patterning with “outer” group; Marathi and Konkani in different groups
- Methodology still needs some debugging before it can be used to assess the Inner/Outer hypothesis

Discussion I

- No support was found for the Inner/Outer hypothesis, but that does not mean that the hypothesis is wrong
- Outstanding problems:
    - Hardware problems necessitated underuse of data
    - No way of teasing apart “old” and “recent” sound changes
    - Potential problems with the Dirichlet distribution: no explicit way of uncovering connections between types of sound change
- Potential future directions:
    - Use different prior over sound change distributions
    - PyMC3 implementation in theory makes it easy to toggle between different priors

Discussion II

- Analysis may benefit from modeling correlation between reflexes of sound change; e.g., if s $>$ x in a particular dialect group, a change s $>$ h in the same group should not surprise us
- The Dirichlet distribution is incapable of modeling such correlation
- However, the Logistic Normal distribution (Atchison and Shen 1980) could model such correlations; a multivariate normal distribution (for which covariance can be expressed) is transformed into a multinomial distribution, so correlation can be modeled between reflexes of Sanskrit sounds in a particular environment

Discussion III

- Partitioned/Shared Logistic Normal distribution (Cohen and Smith 2009) takes this correlation even further, and can model similar behavior between similar Sanskrit sounds
- Harder to model sparsity, but attempts to constrain the Dirichlet prior to yield sparse sound change distributions was not particularly successful

Conclusion

- While most assumption of traditional historical linguistics (in my opinion) are well-founded, their use — particularly for highly noisy and complex phenomena — can be arbitrary
- Probabilistic methods may be able to help guide the method in such situations
- We are still very very far from making probabilistic models match traditional historical-comparative questions to a satisfactory degree
- Hopefully incremental progress can be made!