Gaussian Process Models of Sound Change in Indo-Aryan Dialectology

LChange 2019, Florence 2 August 2019

Chundra A. Cathcart
Department of Comparative Linguistics
University of Zurich

Introduction

Digitized etymological resources and quantitative models can aid us in better understanding language relationships.

Indo-Aryan languages are a group with a long history of scholarship.

It is generally agreed that all contemporary Indo-Aryan languages descend from some attested variant of Sanskrit or Old Indo-Aryan (OIA, Emeneau 1966), though some minority views exist (e.g., Kogan, 2005).

No truly conclusive understanding of Indo-Aryan subgrouping: a number of hypotheses have been advanced, but no single proposal has emerged as the winner (for summary of challenges, see Southworth 1964; Jeffers 1976; Toulmin 2009).

Sound change is assigned a great deal of explanatory power in Indo-Aryan dialectology (Masica, 1991); a number of sound changes thought to be probative with respect to Indo-Aryan dialectology have been put forth (Hock, 2016)

This is in part due to the fact that IA languages have developed in close contact with each other, and intimate lexical borrowing between closely related languages has been widespread (Turner, 1967).

Broad goal of current work: use sound changes extracted from a large etymological database to shed light on IA dialectology

Narrow goal: use a Gaussian Process prior over sound changes to capture shared patterns across different historical phonological developments

Hypotheses

Hoernle (1880): four groups nested within two higher-order groups

Grierson, *Linguistic Survey of India*: Inner-Outer hypothesis, on the basis of several morphosyntactic and phonological innovations

Chatterji (1926): argues that innovations proposed by Grierson are too chronologically shallow to be meaningful for subgrouping purposes

Zograf (1976): no evidence for I-O hypothesis

Masica (1991): skeptical of I-O hypothesis, inconclusive on I-A subgrouping ("overlapping genetic zones")

Southworth (2005): attempts to revive I-O hypothesis on basis of additional evidence

Zoller (2016, forthcoming): also argues for the I-O hypothesis (but with different formulation and evidence)

Still no clear consensus (Hock, 2016), perhaps because traditional comparative method cannot deal with the noise IA exhibits.

Quantitative approaches to Indo-Aryan subgrouping are underused (with some exceptions, e.g., Borin et al. 2014; Peterson 2017)

This is striking, given the wealth of digitized resources for South Asian languages (e.g., Turner, 1966a)

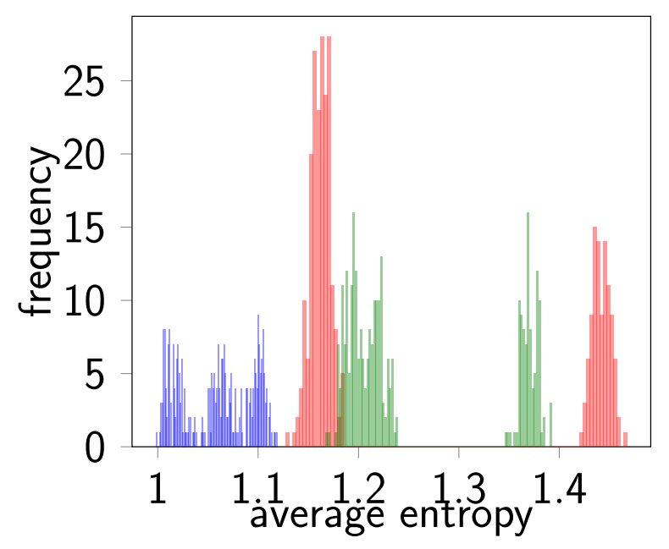

Figure: Average word-level component assignment entropies from posterior samples for each model (Diagonal = blue, BGP = red, GGP = green).

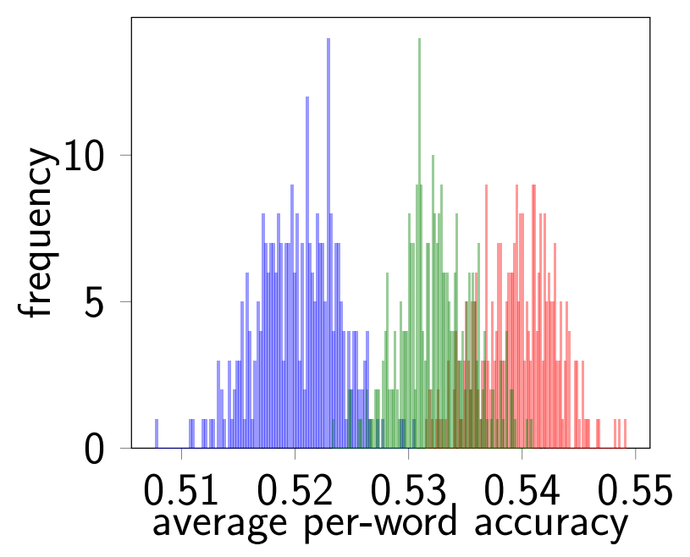

Figure: Average per-word accuracies from posterior samples for each model (Diagonal = blue, BGP = red, GGP = green).

Work in progress

Cathcart fthc uses a mixed-membership model closely related to the Structure model of Pritchard et al. (2000) or LDA (Blei et al., 2003) to assess the evidence for the I-O hypothesis, finding at least partial support; mixed-membership models can be helpful in addressing questions in contact linguistics (cf. Reesink et al., 2009; Syrjänen et al., 2016) provided that the features analyzed display chronological stability and relatively low homoplasy.

This work contrasts two priors over sound change:

A Dirichlet prior, which has a history of use in quantitative approaches to sound change (Bouchard-Côté et al., 2007)

A Partitioned Logistic Normal prior, capable of expressing covariance between outcomes within and across distributions.

However, no major difference in behavior were found between these models in terms of posterior predictive checks.

This procedure relied on a fixed covariance matrix for the Logistic Normal distribution; this covariance matrix may be inappropriate.

The covariance matrix can be modeled via a Gaussian Process.

Data

We extracted all modern Indo-Aryan (NIA) forms from Turner (1966b) along with the OIA headwords from which these reflexes descend (Middle Indo-Aryan languages such as Prakrit and Pali were excluded).

The data were pre-processed to facilitate the extraction of sound changes.

We restrict our analysis to changes affecting OIA ʃ, ʋ, ɲ, ɳ, ʂ, ɽ, h, i, iː, j, kʂ, l, n, r, s, u, uː, which are thought to play a meaningful role in Indo-Aryan dialectology (Southworth, 2005; Hock, 2016).

Model and results

We employ mixed-membership models in order to tease apart admixture between IA languages on the basis of sound changes; model assumes that EACH WORD in EACH LANGUAGE is generated by one of K latent dialect components, according to the relevant sound changes whose operation the word displays

Key parameters are $\boldsymbol{\theta}$ (language-level distributions over dialect components) and $\boldsymbol{\phi}$ (component-level collections of distributions over sound changes). The stochastic generative process we assume to underlie the data looks as follows:

```
For $\boldsymbol{w}_i : i \in \{1, ..., W\}$, the vector of relevant inputs in each OIA etymon
  For each language $l \in \{1, ..., L\}$ continuing $\boldsymbol{w}_i$
    $z_{i,l} \sim \text{Categorical}(\boldsymbol{\theta}_l)$ [Draw a dialect component label]
    For each OIA input $w_{i,t}$ in etymon $\boldsymbol{w}_i$ at index $t : \{1, ..., |\boldsymbol{w}_i|\}$
      $y_{i,l,t} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{w,l}, w_{i,t}, \cdot})$ [Generate each output]
```

Across models, $\boldsymbol{\phi}$ consists of a logistic normal prior over sound change with three different types of covariance: (1) diagonal (cf. Srivastava and Sutton, 2017); "binary" GP prior with ARD kernel, dependent on whether sounds involved are the same or different; and (3) "granular" GP prior, dependent on featural dissimilarity of sounds involved.

Ultimately, the binary GP model performs the best in terms of posterior predictive checks (see figures).

If we figure out the best way to probabilistically represent sound change, digitized etymological resources can tell us a great deal about the evolution of linguistic groups such as Indo-Aryan!

References

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning research 3*, 993–1022.

Borin, L., A. Saxena, T. Rama, and B. Comrie (2014). Linguistic landscaping of south asia using digital language resources: Genetic vs. areal linguistics. In *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3137–3144.

Bouchard-Côté, A., P. Liang, T. Griffiths, and D. Klein (2007). A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, pp. 887–896. Association for Computational Linguistics.

Cathcart, C. A. (fthc). A probabilistic assessment of the Indo-Aryan Inner-Outer Hypothesis. *Journal of Historical Linguistics*.

Chatterji, S. K. (1926). *The Origin and Development of the Bengali Language*. Calcutta: Calcutta University Press.

Emeneau, M. B. (1966). The dialects of Old-Indo-Aryan. In J. Puhvel (Ed.), *Ancient Indo-European dialects*, pp. 123–138. Berkeley: University of California Press.

Hock, H. H. (2016). The languages, their histories, and their genetic classification. In H. H. Hock and E. Bashir (Eds.), *The Languages and Linguistics of South Asia: A Comprehensive Guide*, pp. 9–240. Berlin, Boston: De Gruyter.

Hoernle, A. F. R. (1880). *A comparative grammar of the Gaudian languages*. London: Trübner and Co.

Jeffers, R. J. (1976). The position of the Bihārī dialects in Indo-Aryan. *Indo-Iranian Journal 18*(3-4), 215–225.

Kogan, A. I. (2005). *Dardskie jazyki. Genetičeskaja xarakteristika*. Moscow: Vostočnaja Literatura.

Masica, C. P. (1991). *The Indo-Aryan languages*. Cambridge: Cambridge University Press.

Peterson, J. (2017). Fitting the pieces together: Towards a linguistic prehistory of eastern-central South Asia (and beyond). *Journal of South Asian Languages and Linguistics 4*, 211–257.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000). Inference of population structure using multilocus genotype data. *Genetics 155*(2), 945–959.

Reesink, G., R. Singer, and M. Dunn (2009). Explaining the linguistic diversity of Sahul using population models. *PLoS Biology 7*, e1000241.

Southworth, F. (1964). Family-tree diagrams. *Language 40*(4), 557–565.

Southworth, F. C. (2005). *Linguistic Archaeology of South Asia*. London: Routledge.

Srivastava, A. and C. Sutton (2017). Autoencoding variational inference for topic models. In *International Conference on Learning Representations (ICLR)*.

Syrjänen, K., T. Honkola, J. Lehtinen, A. Leino, and O. Vesakoski (2016). Applying population genetic approaches within languages: Finnish dialects as linguistic populations. *Language Dynamics and Change 6*, 235–283.

Toulmin, M. (2009). *From linguistic to sociolinguistic reconstruction: the Kamta historical subgroup of Indo-Aryan*. Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Turner, R. L. (1962–1966a). *A comparative dictionary of Indo-Aryan languages*. London: Oxford University Press.

Turner, R. L. (1962–1966b). *A comparative dictionary of Indo-Aryan languages*. London: Oxford University Press.

Turner, R. L. (1975 [1967]). Geminates after long vowel in Indo-aryan. In *R.L. Turner: Collected Papers 1912–1973*, pp. 405–415. London: Oxford University Press.

Zograf, G. A. (1976). *Morfologičeskij stroj novyx indoarijskix jazykov*. Moscow: Nauka.

Zoller, C.-P. (2016). Outer and Inner Indo-Aryan, and northern India as an ancient linguistic area. *Acta Orientalia 77*, 71–132.