Exploring phonological diversity through principal component analysis

Emily Clem and Lev Michael
University of California, Berkeley

eclen@berkeley.edu, levpmichael@berkeley.edu

LSA 2017
6 January 2017
Areal typology is undergoing a revolution as computational methods are being applied to new ‘big data’ regional datasets (e.g. O’Connor and Muysken, 2014; Reesink et al., 2009; Wichmann and Good, 2014)

Major goals:

- Identify typological structure in large regional datasets
- Identify areal vs. genetic patterns in such datasets (clarify role of language contact in regional typology)
A tool for areal typology: PCA (correlation and clustering)

- One promising tool to identify large-scale typological patterns is **principal component analysis** (PCA).
- PCA reorganizes a set of correlated variables into new subsets of linearly uncorrelated variables (the ‘principal components’).
- The principal components (PCs) ‘absorb’ the correlations in the original variables.
  - Examining the PCs informs one about correlations in the data.
A tool for areal typology: PCA (dimension reduction)

- PCs are also **ranked** in terms of what percent of the variance in the dataset each PC accounts for.
- In datasets where the original variables exhibit significant correlation:
  - the ‘early’ PCs (PC1, PC2, ...) account for much of the variance
  - while the later PCs account for little of the variance
- Later PCs can be discarded for many purposes, making PCA a tool for exploratory **dimension reduction** for high-dimension datasets.
- Since early PCs account for the major variance in the dataset, they identify major dimensions of **typological differentiation** in a dataset.
PCA and phonological areal typology in South America

- We demonstrate the utility of PCA by applying it to a dataset of South American phonological inventories (SAPhon), to answer the following questions:
  - How is the phonological diversity of South America structured?
    - What are the major typological parameters of differentiation?
  - What areal patterns are detectable?
  - What genetic patterns do we find?
Preview of results

- We present evidence for a strong areal signal in the Andean and Circum-Andean region and its subregions, separating it from Amazonia.
- We will also show evidence for a smaller linguistic area in Northwest Amazonia.
- We will argue that languages in South America differ as to whether the locus of phonological contrast is in their consonant system or vowel system.
- We will demonstrate that contrasts in nasality and length on vowels are both significant dimensions of differentiation for inventories in South America.
The data: SAPhon

- Our analysis of areality and typological patterning in South American phonologies is based on the SAPhon dataset
- SAPhon (South American Phonological Inventory Database) is an online database of phonological inventories of languages of South America (Michael et al., 2016)
- SAPhon houses phonological inventories for 363 languages
  - All languages for which published inventories are available (plus many for which they aren’t): ~90% of living South American languages
  - 104 more than the number of languages with ISO codes in South America (due to inclusion of extinct languages, and some finer-grained classification)
Modeling the SAPhon dataset using a vector space

- To apply PCA it is necessary to model the SAPhon inventories as points in a (301-dimensional) vector space:
  - The basis vectors (‘axes’) that define this space correspond to the 301 segments attested in the SAPhon dataset (\(\vec{p}, \vec{t}, \vec{k}, \vec{i}, \vec{e}, \vec{a} \ldots\))
  - For a given segment vector (e.g., \(\vec{p}\)), a language exhibits a magnitude of 1 if it has this segment in its inventory, and 0 if it doesn’t
  - The combination of these 1 and 0 values for the segment vectors positions each of the 363 SAPhon languages in the vector space
Understanding PCA

• PCA is a transformation (a rotation) of the original set of basis vectors into a new set of basis vectors (= the ‘principal components’)
  • This rotation eliminates correlations between the basis vectors in the dataset
  • These new vectors are oriented in the ‘directions’ of greatest variance in the dataset
• As with all rotations in a vector space, the new basis vectors are defined in terms of linear combinations of the old basis vectors
Understanding PCA

- Dataset with original set of variables (vectors)
Understanding PCA

- Dataset with original set of variables (vectors) and new PCs
- PC1 oriented in direction of greatest variance in the dataset
Understanding PCA

- The PCs are a linear combination of the original basis vectors:

\[ P\hat{C}_1 = \sin(\alpha)\hat{v}_1 + \cos(\alpha)\hat{v}_2 = \frac{1}{\sqrt{2}}\hat{v}_1 + \frac{1}{\sqrt{2}}\hat{v}_2 \]

\[ P\hat{C}_2 = \sin(\beta)\hat{v}_1 + \cos(\beta)\hat{v}_2 = \frac{1}{\sqrt{2}}\hat{v}_1 - \frac{1}{\sqrt{2}}\hat{v}_2 \]
The PCs obtained by carrying out PCA on the original SAPhon basis vectors are thus linear combinations of the segment vectors.

The linear sum specifies the positive and negative weights accorded each segment in calculating the given PC:

- e.g. \( PC_1 = 0.140l + 0.137ts + 0.128f + 0.114a + 0.122a: \)
  \[\ldots - 0.359\ddot{i} - 0.348\ddot{a} - 0.315\ddot{e} - 0.312\ddot{u} - 0.306\ddot{o} \]

Each PC resembles a pair of weighted phonological inventories:

- A **positive inventory** that characterizes the positive extremum of that PC
- A **negative inventory** that characterizes the negative extremum of that PC
Variance explained by PCs

- A comparatively small number of the 301 PCs are responsible for explaining the majority of the variance in the dataset
- PC1 = 14.2%, PC2 = 7.6%, PC3 ≈ 5.4%, …
Variance explained by PCs

- A comparatively small number of the 301 PCs are responsible for explaining the majority of the variance in the dataset
- PC1 = 14.2%, PC2 = 7.6%, PC3 ≈ 5.4%, ... 
- Due to the rapid decrease in variance explained by the successive PCs, we can focus on the largest PCs to identify the significant ‘typological structure’ of the SAPhon phonological inventories
- For each of the first 5 PCs, we can examine both areal and genetic patterns that are revealed
PC1: positive coefficients

- PC1 explains 14.2% of the variance in the data
- The segments with the largest positive coefficients include affricates, palatals, and laterals
PC1: negative coefficients

- Negative coefficients are slightly larger than positive coefficients
- The segments with the largest negative coefficients include nasal vowels, ñ, and mid vowels
PC1: areal signal

- PC1 yields a strong positive signal in the Andean and Circum-Andean region
PC1: genetic signal

- PC1 shows a negative genetic signal including the Tucanoan ($p<1.0\text{E}-13$), Tupí ($p<1.0\text{E}-11$), and Macro-Ge ($p<1.0\text{E}-4$) families (using Kolmogorov-Smirnov Test)
PC1: genetic signal

- PC1 shows a negative genetic signal including the Tucanoan ($p<1.0E-13$), Tupí ($p<1.0E-11$), and Macro-Ge ($p<1.0E-4$) families (using Kolmogorov-Smirnov Test)
PC1: summary

- Positive segments: alveolar and palatal laterals, affricates, fricatives, and nasals
- Negative segments: nasal vowels, ɨ
- Positive component yields a strong areal signal in the Andes and Circum-Andean area, including Patagonia
  - Independently identifies this phonological area, first found using a Naive Bayesian Classifier method (Michael et al., 2014)
- Negative component shows a genetic signal from Tucanoan, Macro-Ge, and Tupí families
PC2: positive coefficients

- PC2 explains 7.6% of the variance in the data
- The segments with the largest positive coefficients include the voiced stops
PC2: negative coefficients

- Negative coefficients are larger than positive coefficients
- All of the segments with the largest negative coefficients are long vowels
**PC2: areal signal**

- PC2 does not display a strong areal signal distinct from genetic relationships.
PC2: genetic signal

- PC2 shows negative genetic signal from Arawak \((p<1.0\text{E}-5)\) and Yanomam \((p<1.0\text{E}-4)\)
PC2: genetic signal

- PC2 shows negative genetic signal from Arawak (p < 1.0E-5) and Yanomam (p < 1.0E-4)
PC2: summary

- Positive segments: voiced stops
- Negative segments: long vowels
- Negative component shows a strong genetic signal associated with Yanomam, and other families also cluster together
- We see a large negative dispersion with the most strongly negative languages displaying vowel length contrasts for many vowels
PC3: positive coefficients

- PC3 explains 5.4% of the variance in the data
- The segments with the largest positive coefficients include the voiced stops, tone, and aspirated stops
- Positive coefficients are much larger than negative coefficients
PC3: negative coefficients

- The segments with the largest negative coefficients include the nasal stops and approximants
PC3: areal signal

- PC3 shows a strong positive signal in Northwest Amazonia
PC3: genetic signal

- PC3 shows a strong negative genetic signal associated with Tupí ($p<1.0E-5$)
PC3: genetic signal

- PC3 shows a strong negative genetic signal associated with Tupí ($p<1.0E-5$)
PC3: summary

- Positive segments: voiced stops, tone
- Negative segments: nasal stops, approximants
- Positive component yields a strong areal signal in Northwest Amazonia
  - Identifies this well known linguistic area (see, e.g. Aikhenvald, 2002) on the basis of phonological inventories alone
- This positive signal reflects languages that have processes of nasal harmony rather than underlying nasal stops
- Negative component shows a genetic signal from Tupí
PC4: positive coefficients

- PC4 explains 4.6% of the variance in the data
- The segments with the largest positive coefficients include β, i, palatals, and mid vowels
PC4: negative coefficients

- Negative coefficients are larger than positive coefficients
- The segments with the largest negative coefficients are ejectives, aspirated stops, laterals, and uvulars
PC4: areal signal

- PC4 shows a strong negative signal in the Southern Andean and Circum-Andean region, including Patagonia
PC4: genetic signal

- PC4 shows a positive genetic signal associated with Tupí ($p<1.0\text{E}-5$) and Panoan ($p<1.0\text{E}-4$)
PC4: genetic signal

• PC4 shows a positive genetic signal associated with Tupí (p<1.0E-5) and Panoan (p<1.0E-4)
PC4: summary

- Positive segments: β, i, palatals
- Negative segments: ejectives, aspirated stops, laterals, uvulars
- Negative component shows further support for a strong areal signal in Southern Andean and Circum-Andean region
  - Identifies this important sub-area of the Andean and Circum-Andean area also found by Michael et al. (2014)
- Positive component shows a genetic signal from Tupí and Panoan
PC5: positive coefficients

- PC5 explains 3.9% of the variance in the data
- The segments with the largest positive coefficients are mid vowels, w, and glottals
PC5: negative coefficients

- Negative coefficients are slightly larger than positive coefficients
- The segments with the largest negative coefficients are palatals, affricates, and ː
PC5: areal signal

- PC5 shows a negative signal in the Central-Northern Andean region and a positive signal in the east
PC5: genetic signal

- PC5 shows a west/east divide illustrated by the Quechua ($p<1.0E-9$) and Carib ($p<1.0E-4$) families
PC5: genetic signal

- PC5 shows a west/east divide illustrated by the Quechua ($p<1.0E-9$) and Carib ($p<1.0E-4$) families
**PC5: summary**

- Positive segments: mid vowels
- Negative segments: palatals, affricates, ɨ
- There is a general divide, both genetic and areal, between a negative signal in the west and a positive signal in the east
  - The western negative region corresponds to the Central-Northern Andean and Circum-Andean sub-region identified by Michael et al. (2014)
- Overall, the signal is becoming weak by this point
Interim Summary: Areal results

- Examination of geographical distribution of extremal values of PCs 1–5 have identified known phonological areas in South America:
  1. Andean and Circum-Andean phonological area
     1.1 Southern Andean and Circum-Andean phonological sub-area
     1.2 Central-Northern Andean and Circum-Andean phonological sub-area
  2. Northwest Amazonian phonological area

- Whereas previous identification of these areas have relied on methods that require human insight and intuition (Naive Bayes Classifier approaches; Michael et al. (2014)), in the PCA approach can these results emerge more directly from the data.
Identifying phonological ‘types’

- Since the PCs identify the major dimensions of phonological variation in South America, we can develop a ‘continental typology’ of phonological inventories.
- Major phonological types can be identified by examining how inventories cluster with respect to PC1 and PC2.
  - Hierarchical clustering using a Euclidean distance measure and Ward’s clustering criterion.
- Sampling languages in each cluster allows us to identify their major features.
Phonological types in South America

- What major types can we extrapolate from PCA?
- What are the clusters and what are their general profiles?
Quadrant 1: +PC1, +PC2

- Features: large consonant inventories (laterals, affricates, voiced stops) and small vowel inventories
- Example language: Salasaca Quechua (Quechua)
Quadrant 2: +PC1, -PC2

- Features: large consonant inventories (laterals, affricates), moderate vowel inventories (long vowels)
- Example language: Chamicuro (Arawak)

<table>
<thead>
<tr>
<th>Consonants</th>
<th>Bilabial</th>
<th>Alveolar</th>
<th>Post-alveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
<td>p t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affricate</td>
<td></td>
<td>ts tf ts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td></td>
<td>s j s h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>m n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td></td>
<td>j w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tap, flap</td>
<td></td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral</td>
<td></td>
<td>l η</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vowels</th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>i i:</td>
<td>u u:</td>
<td></td>
</tr>
<tr>
<td>Mid</td>
<td>e e:</td>
<td>o o:</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>a a:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Quadrant 3: -PC1, -PC2

• Features: small consonant inventories, large vowel inventories (nasal and long vowels)
• Example language: Karitiâna (Tupí)

<table>
<thead>
<tr>
<th>Consonants</th>
<th>Bilabial</th>
<th>Alveolar</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
<td>p</td>
<td>t</td>
<td>k</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>s</td>
<td></td>
<td></td>
<td>h</td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>n</td>
<td>ñ</td>
<td>ñ</td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td></td>
<td></td>
<td></td>
<td>w</td>
<td></td>
</tr>
<tr>
<td>Tap, flap</td>
<td>r</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vowels</th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>i ī i: ī:</td>
<td>i ī i: ī:</td>
<td></td>
</tr>
<tr>
<td>Mid</td>
<td>e ē e: ē:</td>
<td></td>
<td>o ō o: ō:</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td>a ā a: ā:</td>
<td></td>
</tr>
</tbody>
</table>
**Quadrant 4: -PC1, +PC2**

- **Features:** moderate consonant inventories (voiced stops), moderate vowel inventories (nasal vowels)
- **Example language:** Siona (Tucanoan)

<table>
<thead>
<tr>
<th>Consonants</th>
<th>Bilabial</th>
<th>Alveolar</th>
<th>Post-alveolar</th>
<th>Palatal</th>
<th>Velar</th>
<th>Labio-velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop/affricate</td>
<td>p</td>
<td>b</td>
<td>tʃ</td>
<td>k</td>
<td>g</td>
<td>kʷ gʰ</td>
<td>?</td>
</tr>
<tr>
<td>Fricative</td>
<td></td>
<td>s</td>
<td>z</td>
<td></td>
<td></td>
<td>h</td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td>h</td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td>j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>w</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vowels</th>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>ññ</td>
<td>ññ</td>
<td>u ū</td>
</tr>
<tr>
<td>Mid</td>
<td>ññ</td>
<td></td>
<td>o ō</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td>a ſ</td>
</tr>
</tbody>
</table>
Conclusion

- In South America, whether languages make a large number of contrasts in their vowels is very significant in ‘typing’ languages
  - Nasal vs. oral is one of the most significant dimensions of variation
  - Length contrasts are also an important parameter of differentiation
- PCA is successful in producing this continental typology as well as in identifying important linguistic areas such as the Andean and Circum-Andean region and Northwest Amazonia
- This work provides a starting point for more quantitatively rigorous analyses of areality


