UNIVERSITY OF CAMBRIDGE

Latent Variables: The Power of Assuming Missing Information

Thomas Marge

Trinity BA Seminars

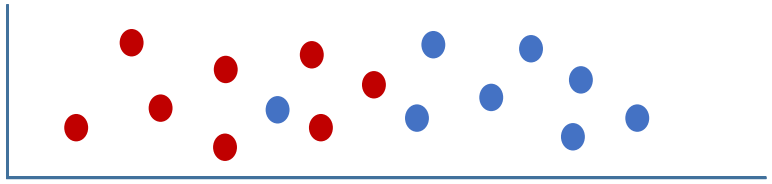
April 29, 2019

Advisor: John Aston

Audio recordings of ancestral languages do not exist. Are these sounds hidden in how we speak today?

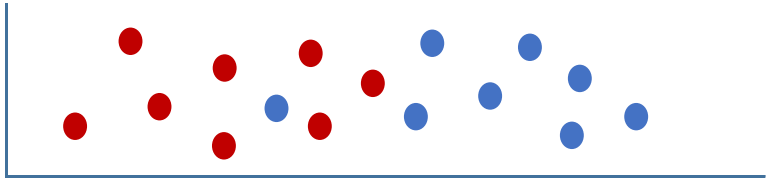
DISCLAIMER: NOT A LINGUIST

General Multiple Distribution Modeling

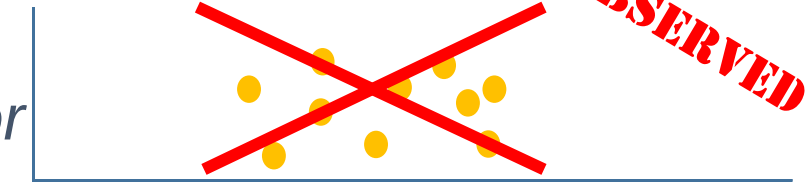


Latent Tree Model

Now



Predecessor

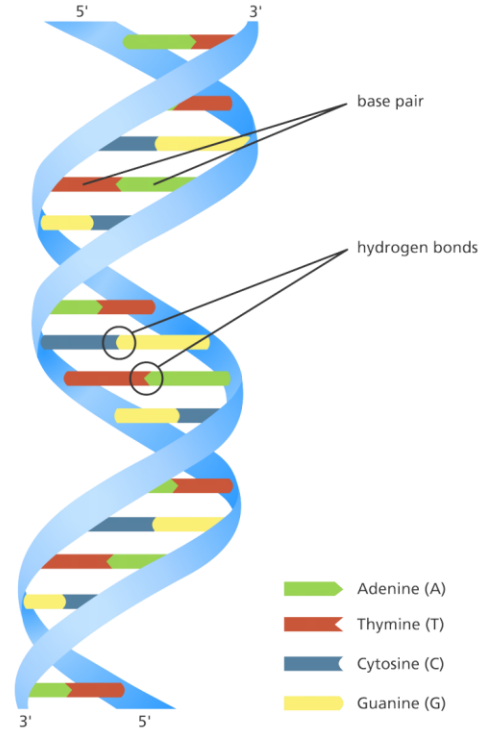


Questions
-how can we model the distribution of multiple data sets?
-how can we distinguish between multiple data sets?
-does it appear that there are multiple data sets?

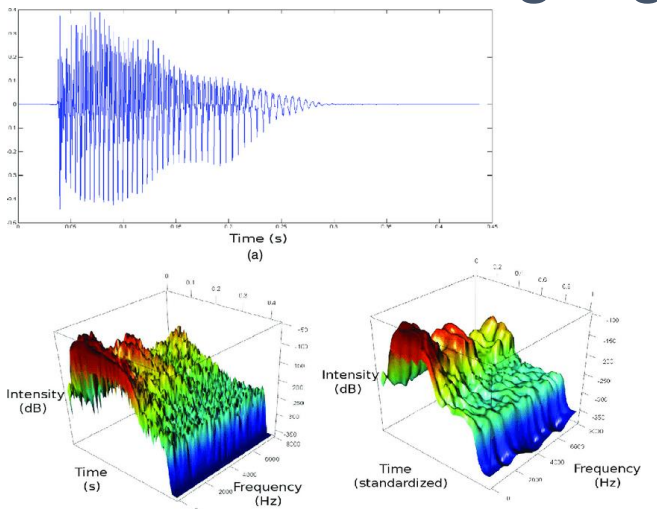
Additional Questions
-is the data consistent with a shared ancestor?
-can we describe a distribution for the common ancestor based on the observed nodes?

Discrete Model: Phylogenetics

-approximate probabilities of peptide mismatch between generations known for highly conserved regions
-number of generation may also be accessible
-finitely many base pairs



Continuous Model: Language



-continuous distribution with timeframe mismatch
-cross language influences makes finding conserved pronunciation difficult
-“thousand year rule”: rapid change leads to more noise

DNA IMAGE SOURCE: https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwjrlem0w-bgAhVoh-AKHWsmCWwQjRx6BAgBEAU&url=https%3A%2F%2Fwww.yourgenome.org%2Ffacts%2Fwhat-is-dna&psig=AOvVaw0KyV6a_tUdZMrh1-sAUCPw&ust=1551721546803534
SPECTROGRAM SOURCE: https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwjTosvuw-bgAhUvWN8KHdIjCAkQjRx6BAgBEAU&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2Fa-Raw-record-b-raw-log-spectrogram-and-c-smoothed-and-aligned-log-spectrogram-for_fig1_328234781&psig=AOvVaw3g7qCfricOqEcYy-Se24Jf&ust=1551721655816795

Language Processing: Time Synchronization

-Speakers communicate at different speeds

-More than just start time and end time for spoken words, interior of words may be warped as well

Muller and Tang Time Warp

-Aligning different words from different speakers across languages using measurement of alignment of peaks

-Results show that looking at a subset of comparisons is enough

-Searching for a universally comparable time

real time *universal time*

$$y_{ij} = Y_i(t_j) + \epsilon_{ij} = X_i\{h_i^{-1}(t_j)\} + \epsilon_{ij},\ t_j \in \mathcal{T},$$

underlying distribution

$$X_i(t) = \mu(t) + \delta Z_i(t),\ \text{for } t \in \mathcal{T},$$

deviation in distribution (ie speaker, language)

linear spline warping function

$$h(t) = \begin{cases} \tau_1 t/a_1, & \text{for } 0 \le t < a_1, \\ (\tau_j - \tau_{j+1})(t - a_j)/(a_j - a_{j+1}) + \tau_j, & \text{for } a_j \le t < a_{j+1},\ j = 1, ..., p-1, \\ (\tau_p - T)(t - T)/(a_p - T) + T, & \text{for } a_p \le t \le T. \end{cases}$$

TIME WARP SOURCE: Pairwise Curve Synchronization for Functional Data By RONG TANG Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, Rockville, Maryland 20850, U.S.A. rong.tang@fda.hhs.govand HANS-GEORG MU¨LLERDepartment of Statistics, University of California, Davis, California 95616, U.S.A. mueller@wald.ucdavis.edu

Muller and Tang Time Warp (continued)

-differentiability constraints
-well ordering of time constraints
-independence constraints

-proof of computational time
-proof of local minima

$$h(t) = \Theta^{\mathrm{T}} A(t).$$ ← *a way to store spline coefficients*

$$g_{ik}(t) = h_i\{h_k^{-1}(t)\}$$ ← *transforming between languages*

$$\widetilde{\Theta}_{g_{ik}} = \arg\min_{\Theta \in \Omega} C_\lambda(Y_i, Y_k, \Theta),$$

objective function →

$$C_\lambda(Y_i, Y_k, \Theta) = E\Big(\int_{\mathcal{T}} \Big[\{Y_i(\Theta^{\mathrm{T}} A(t)) - Y_k(t)\}^2 + \lambda\{\Theta^{\mathrm{T}} A(t) - t\}^2\Big] dt \Big| Y_i, Y_k\Big)$$

measure difference in peaks

limit deviation from true time coordinates to stop "single peaks"

TIME WARP SOURCE: Pairwise Curve Synchronization for Functional Data By RONG TANG Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, Rockville, Maryland 20850, U.S.A. rong.tang@fda.hhs.govand HANS-GEORG MU¨LLERDepartment of Statistics, University of California, Davis, California 95616, U.S.A. mueller@wald.ucdavis.edu

Muller and Tang Time Warp (continued)

-differentiability constraints
-well ordering of time constraints
-independence constraints

-proof of computational time
-proof of local minima

$$h(t) = \Theta^{\mathrm{T}} A(t). \quad \longleftarrow \quad \textit{a way to store spline coefficients}$$

$$g_{ik}(t) = h_i\{h_k^{-1}(t)\} \quad \longleftarrow \quad \textit{transforming between languages}$$

Theory works very nicely in continuous world…
But we never have truly continuous observations

$$\widetilde{\Theta}_{g_{ik}} = \arg\min_{\Theta \in \Omega} C_\lambda(Y_i, Y_k, \Theta),$$

objective function →

$$C_\lambda(Y_i, Y_k, \Theta) = E\Big(\int_{\mathcal{T}} \Big[\{Y_i(\Theta^{\mathrm{T}} A(t)) - Y_k(t)\}^2 + \lambda \{\Theta^{\mathrm{T}} A(t) - t\}^2 \Big] dt \Big| Y_i, Y_k \Big)$$

measure difference in peaks

limit deviation from true time coordinates to stop "single peaks"

TIME WARP SOURCE: Pairwise Curve Synchronization for Functional Data By RONG TANG Division of Biostatistics, Center for Devices and Radiological Health, Food and Drug Administration, Rockville, Maryland 20850, U.S.A. rong.tang@fda.hhs.govand HANS-GEORG MU¨LLERDepartment of Statistics, University of California, Davis, California 95616, U.S.A. mueller@wald.ucdavis.edu

Motivation: Why Gaussian?

-DEFINITION: we say that our observed random variables have a tree structure iff they are conditionally independent from each other given "predecessor" unobserved random variables.

$$\Omega := \Sigma^{-1}$$ ← *Inverse covariance*

$\Omega_{ij} = 0$ if and only if X_i and X_j are conditionally independent given all other coordinates of X.

($\Sigma_{ij} = 0$ if and only if X_i and X_j are independent.)

WHY? … PDF of multivariate normal

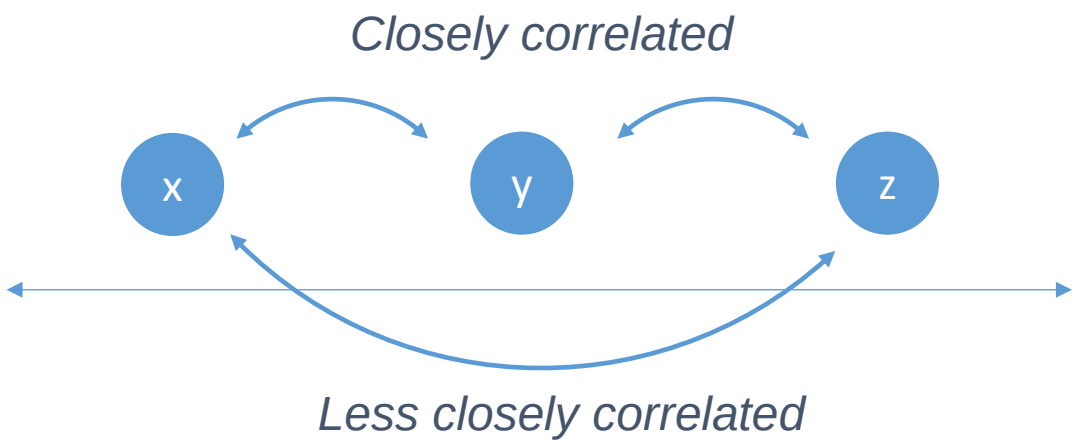
$$(2\pi)^{-\frac{k}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

EQUATION SOURCE: https://stephens999.github.io/fiveMinuteStats/normal_markov_chain.html, Wikipedia

Interpreting the Positivity Constraint

$$\forall_{i<j<k} \quad \sigma_{ij}\sigma_{ik}\sigma_{jk} \geq 0.$$

Roughly equates to an "ordered" covariance structure



ORIGINAL PAPER: Gaussian tree constraints applied to acoustic linguistic functional data Nathaniel Shiers a, John A.D. Aston b,∗, Jim Q. Smith a, John S. Coleman c a Department of Statistics, University of Warwick, Coventry, UK b Statistical Laboratory, University of Cambridge, Cambridge, UK c Phonetics Laboratory, University of Oxford, Oxford, UK

Language Model Overview

-we have observations from speaker i, from word j, from language k

$$X_{ijk}$$

-each observation has amplitude observations for each frequency y at time z

$$X_{ijk}^{yz}$$

-each of these combinations defines a Gaussian random variable

-for our purposes we can think of the model as viewing words and speakers to be observations of language level variability

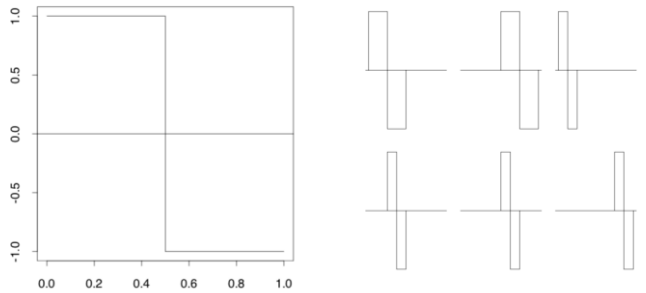
What is a wavelet?

-An attempt to extract “features” in data

-hope that some wavelet coefficients are noise and others contain essential information

Desired Filter Structure

$$1 = (\text{const})^2 \int \psi^2(2^j x - k)dx$$



Raw

Wavelet Coefficients (in 2 dimensions)

Inverted Image after 25% of Pixels Thresholded

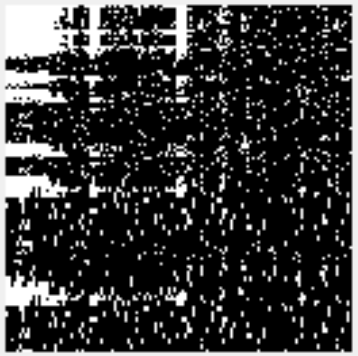
SOURCE: Wavelets For Kids A Tutorial Introduction. Brani Vidakovic and Peter Mueller. Duke University
IMAGE SOURCE: http://eeweb.poly.edu/~yao/EL5123/image/lena_gray.bmp

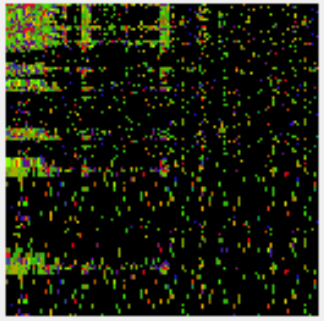
Understanding the Analysis

-visualize which wavelets are tree consistent, as well as which tree is best fit by different coefficients

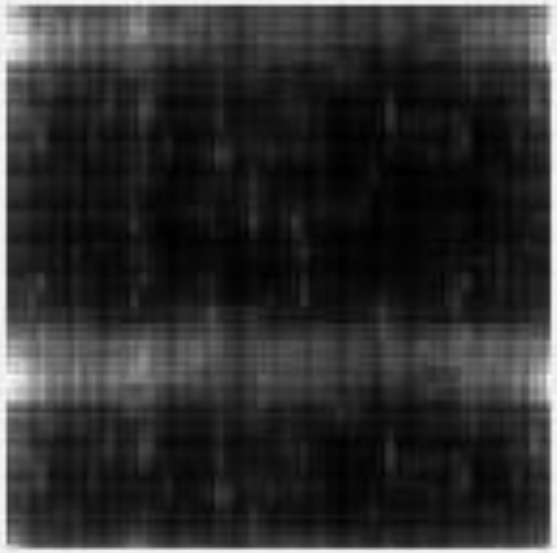
-invert wavelets weighted by their correlation to get a sense for where the tree amenable information maps to

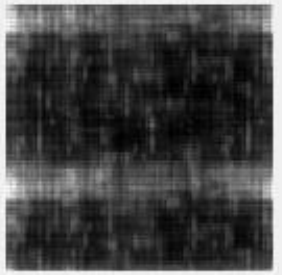
In Wavelet Coefficient Coordinates, which tree?

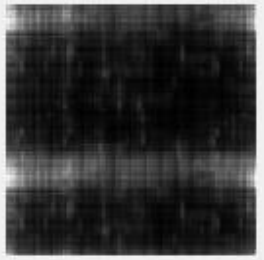


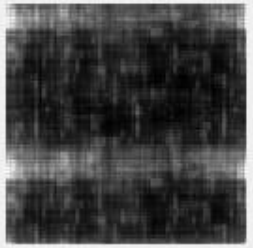


In Spectrogram Coordinates, which frequencies and times









Advantages

-clearly outlines certain sections of wavelet coefficients and sections of original spectrogram

-presents a couple of trees somewhat supported by data and consistent with linguistics research

Setbacks

-tree amenable coefficients do not allow for a strong reconstruction, suggesting either a lack of overall tree amenability or poor feature extraction of wavelet basis

Future Directions

-create better understood hypothesis testing both for the Gaussian Latent Tree Model as well as other tree models

UNIVERSITY OF CAMBRIDGE

Funding and Support

UNIVERSITY OF CAMBRIDGE