Longitudinal Spacekime Analytics:

Time Complexity & Inferential Uncertainty

Ivo D. Dinov

Statistics Online Computational Resource
Health Behavior & Biological Sciences
Computational Medicine & Bioinformatics
Neuroscience Graduate Program
Michigan Institute for Data Science
University of Michigan

http://SOCR.umich.edu

Joint work with Milen V. Velev (BTU)

Based on an upcoming book "*Data Science: Time Complexity and Inferential Uncertainty*"

SCHOOL OF NURSING
STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)
UNIVERSITY OF MICHIGAN

Slides Online:
"SOCR News"

Outline

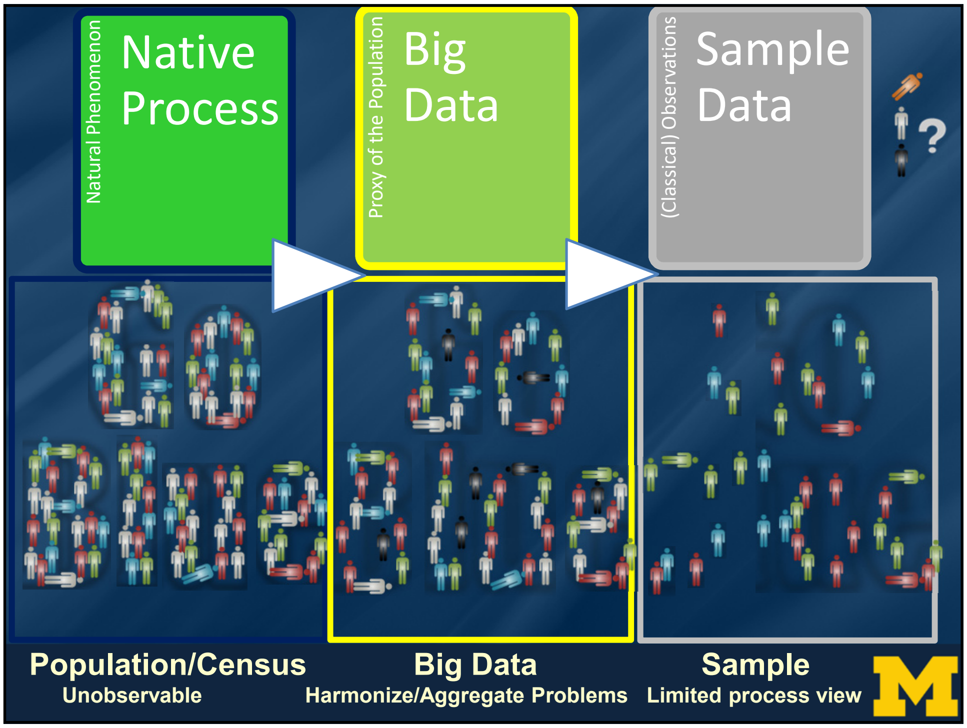
- ❑ Big Biomedical/Health Data Analytic Challenges
- ❑ Complex-Time (*kime*)
- ❑ Spacekime Mathematical Foundation
- ❑ Statistical Implications of Spacekime Analytics
  - ❑ Inferential Uncertainty
  - ❑ Bayesian Inference Representation
- ❑ Applications – Longitudinal Spacekime Data Analytics
  - ❑ Neuroimaging (UKBB, fMRI)
  - ❑ Air quality (UCI ML Air Quality Dataset)

Big Biomedical & Health Data Analytic Challenges

Data Analytics = Information Compression

- From 23 … to … 2^{23} (10M) $\left(\underbrace{23}_{2\ \#'s} \rightarrow \underbrace{2^{23}}_{8\ \#'s}\right)$
- Data Science: 1798 vs. 2019
- In the 18th century, Henry Cavendish used just 23 observations to answer a fundamental question – “What is the Mass of the Earth?” He estimated very accurately the mean density of the Earth/H_2O (5.483 ± 0.1904 g/cm^3)
- In the 21st century to achieve the same scientific impact, matching the reliability and the precision of the Cavendish’s 18th century prediction, requires a monumental community effort using massive and complex information perhaps on the order of 10M (2^{23}) bytes

Dinov (2016) JSMI



Characteristics of Big Biomed Data

IBM Big Data 4V's: Volume, Variety, Velocity & Veracity

Big Bio Data Dimensions	Tools
Size	Harvesting and management of vast amounts of data
Complexity	Wranglers for dealing with heterogeneous data
Incongruency	Tools for data harmonization and aggregation
Multi-source	Transfer and joint modeling of disparate elements
Multi-scale	Macro → meso → micro → nano scale observations
Time	Techniques accounting for longitudinal effects
Incomplete	Reliable management of missing data

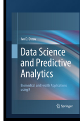
Example: analyzing observational data of 1,000's Parkinson's disease patients based on 10,000's signature biomarkers derived from multi-source imaging, genetics, clinical, physiologic, phenomics and demographic data elements

Software developments, student training, service platforms and methodological advances associated with the Big Data Discovery Science all present existing opportunities for learners, educators, researchers, practitioners and policy makers

Dinov, *GigaScience* (2016) PMID:26918190

Data Science & Predictive Analytics

- **Data Science**: an emerging extremely transdisciplinary field - bridging between the theoretical, computational, experimental, and applied areas. Deals with enormous amounts of complex, incongruent and dynamic data from multiple sources. Aims to develop algorithms, methods, tools, and services capable of ingesting such datasets and supplying semi-automated decision support systems
- **Predictive Analytics**: process utilizing advanced mathematical formulations, powerful statistical computing algorithms, efficient software tools, and distributed web-services to represent, interrogate, and interpret complex data. Aims to forecast trends, cluster patterns in the data, or prognosticate the process behavior either within the range or outside the range of the observed data (e.g., in the future, or at locations where data may not be available)

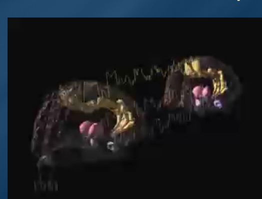


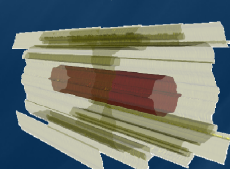
http://DSPA.predictive.space

Dinov, Springer (2018)

Longitudinal Data Analytics

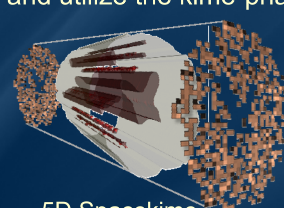
- **Neuroimaging**:
  - *4D fMRI:* time-series, represents measurements of hydrogen atom densities over a 3D lattice of spatial locations ($1 \leq x, y, z \leq 64$ pixels), about 3 × 3 millimeters^2 resolution. Data is recorded longitudinally over time ($1 \leq t \leq 180$) in increments of about 3 seconds & post-processing
  - *State-of-the-art Approaches*: Time-series modeling or Network analysis
  - *Spacekime Analytics*: 5D fMRI kime-series, representing the of hydrogen atom densities over the same 3D lattice of spatial locations, longitudinally over the 2D kime space, $\kappa = re^{i\varphi} \in \mathbb{C}$
  - *Differences*: Spacekime analytics estimate and utilize the kime-phases





4D Spacetime

4D/5D Reconstructions



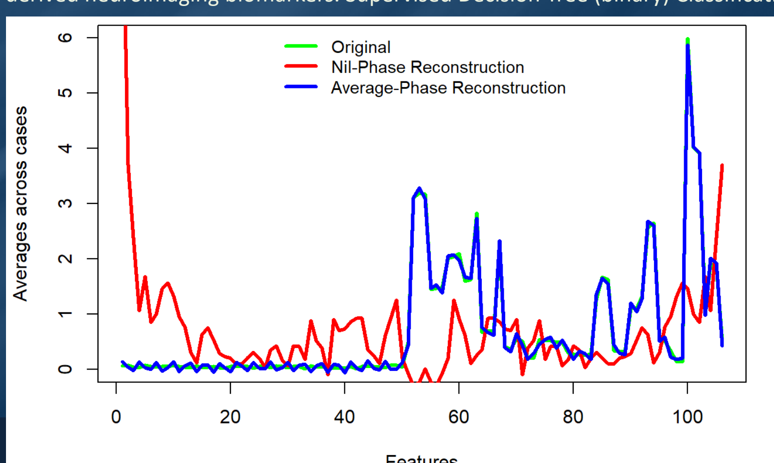
5D Spacekime

Dinov & Velev (2020)

Complex-Time (*kime*)
&
Spacekime Foundations

Preview: Get to UKBB Big Data Analytics Study, but first we need some background …

9,914 UKBB participants; 7,614 clinical measurements, phenotypic features, and derived neuroimaging biomarkers. Supervised Decision Tree (binary) Classification



The Fourier Transform

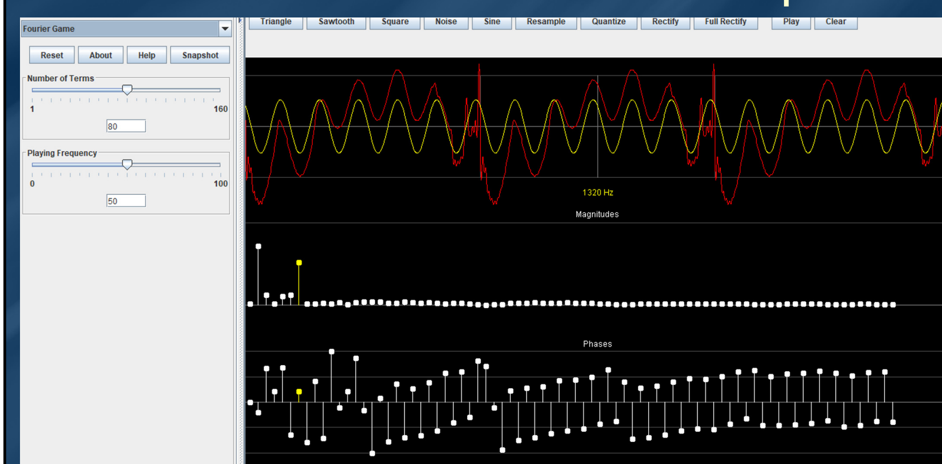
By separability, the classical **spacetime Fourier transform** is just four Fourier transforms, one for each of the four spacetime dimensions, $(\boldsymbol{x}, t) = (x, y, z, t)$. The FT is a function of the angular frequency ω that propagates in the wave number direction $\boldsymbol{k}$ (space frequency). Symbolically, the forward and inverse Fourier transforms of a 4D ($n = 4$) spacetime function f, are defined by:

$$FT(f) = \hat{f}(\boldsymbol{k}, \omega) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int f(\boldsymbol{x}, t) e^{i(\omega t - \boldsymbol{kx})} dt d^3\boldsymbol{x},$$

$$IFT(\hat{f}) = \hat{\hat{f}}(\boldsymbol{x}, t) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int \hat{f}(\boldsymbol{k}, \omega) e^{-i(\omega t - \boldsymbol{kx})} d\omega d^3\boldsymbol{k}.$$

$$\left[\hat{\hat{f}}(\boldsymbol{x}, t) = IFT(\hat{f}) = IFT(FT(f)) = f(\boldsymbol{x}, t) \right]$$

1D Fourier Transform Example



SOCR 1D Fourier / Wavelet signal decomposition into *magnitudes* and *phases (Java applet)*

Top-panel: original signal (image), white-color curve drawn manually by the user and the reconstructed synthesized (IFT) signal, red-color curve, computed using the user modified magnitudes and phases

Bottom-panels: the Fourier analyzed signal (FT) with its magnitudes and phases

http://www.socr.ucla.edu/htmls/game/Fourier_Game.html (Java Applet)

2D Fourier Transform – The Importance of Magnitudes & Phases

Fourier Analysis

(real part of the Forward Fourier Transform)

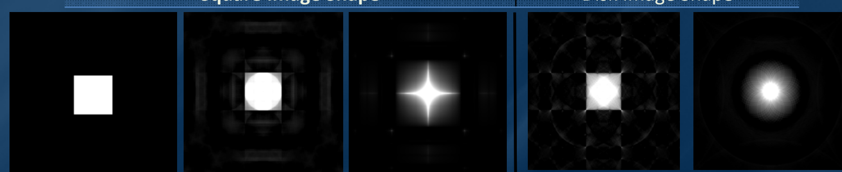
Square Image Shape	Disk Image Shape

2D image 1 (square) | Re(FT(square)) | Magnitude FT(square) | Phase FT(square) | 2D image 2 (disc) | Re(FT(disc)) | Magnitude FT(disc) | Phase FT(disc)

Fourier Synthesis

(real part of the Inverse Fourier Transform)

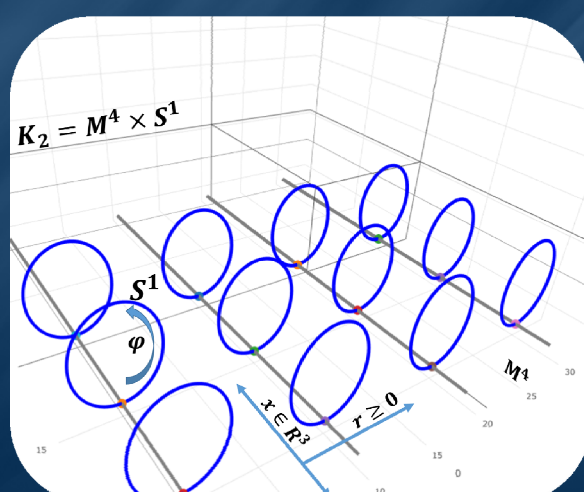
Square Image Shape	Disk Image Shape



IFT(FT(square)) ≡ square	**IFT using square-magnitude & disc-phase**	**IFT using square-magnitude & nil-phase**	**IFT using disc-magnitude & square-phase**	**IFT using disc-magnitude & nil-phase**

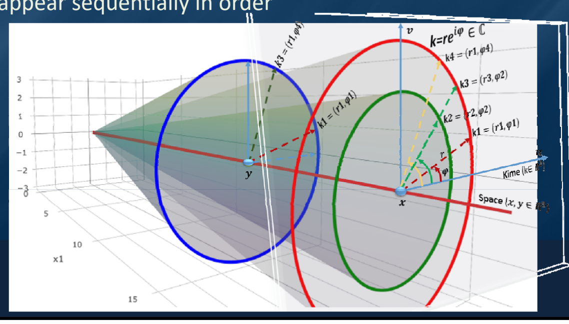
Kaluza-Klein Theory

- Theodor Kaluza developed (1921) an extension of the classical general relativity theory to 5D. This included the metric, the field equations, the equations of motion, the stress-energy tensor, and the cylinder condition. Oskar Klein (1926) interpreted Kaluza's 3D+2D theory in quantum mechanical space and proposed that the fifth dimension was curled up and microscopic.
- The topology of the 5D Kaluza-Klein spacetime is $K_2 \cong M^4 \times S^1$, where M^4 is a 4D Minkowski spacetime and S^1 is a circle (non-traversable).



Complex-Time (*Kime*)

- At a given spatial location, $\boldsymbol{x}$, complex time (*kime*) is defined by $\kappa = re^{i\varphi} \in \mathbb{C}$, where:
  - the magnitude represents the longitudinal events order ($r > 0$) and characterizes the longitudinal displacement in time, and
  - event phase ($-\pi \le \varphi < \pi$) is an angular displacement, or event direction
- There are multiple alternative parametrizations of kime in the complex plane
- Space-kime universe ($R^3 \times \mathbb{C}$):
  - $(\boldsymbol{x}, k1)$ and $(\boldsymbol{x}, k4)$ have the same spacetime representation, but different spacekime coordinates,
  - $(\boldsymbol{x}, k1)$ and $(\boldsymbol{y}, k1)$ share the same kime, but represent different spatial locations,
  - $(\boldsymbol{x}, k2)$ and $(\boldsymbol{x}, k3)$ have the same spatial-locations and kime-directions, but appear sequentially in order



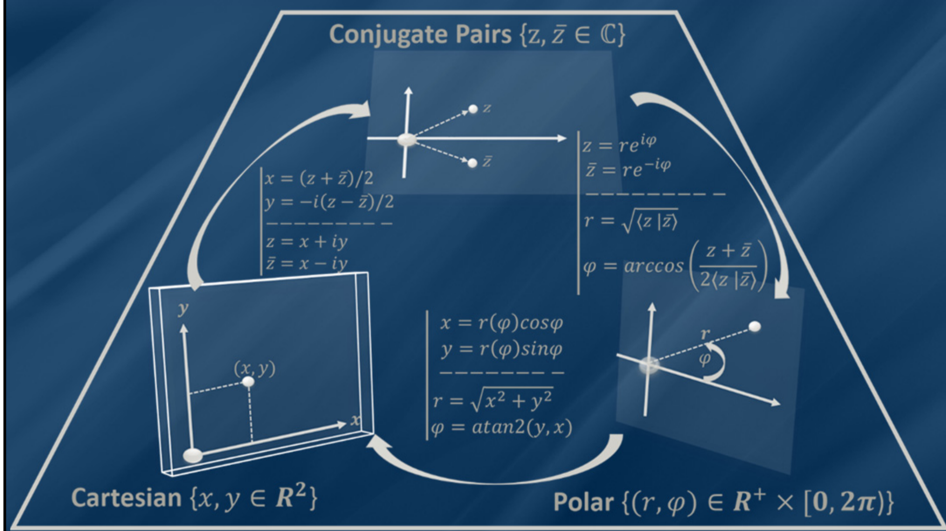
The Spacekime Manifold

- **Spacekime:** $(\boldsymbol{x}, \boldsymbol{k}) = \left(\underbrace{x^1, x^2, x^3}_{\text{space}}, \underbrace{c\kappa_1 = x^4, c\kappa_2 = x^5}_{\text{kime}} \right) \in X, \quad c \sim 3 \times 10^8 \; m/s$
- **Kevents** (*complex events*): points (or states) in the spacekime manifold X. Each kevent is defined by where ($\boldsymbol{x} = (x, y, z)$) it occurs in space, what is its *causal longitudinal order* $\left(r = \sqrt{(x^4)^2 + (x^5)^2}\right)$, and in what *kime-direction* ($\varphi = \text{atan2}(x^5, x^4)$) it takes place.
- **Spacekime interval** (ds) is defined using the general Minkowski 5×5 metric tensor $\left(\lambda_{ij}\right)_{i=1,j=1}^{5,5}$, which characterizes the geometry of the *curved spacekime* manifold:

$$ds^2 = \sum_{i=1}^{5} \sum_{j=1}^{5} \lambda_{ij} dx^i dx^j = \lambda_{ij} dx^i dx^j$$

- *Euclidean (flat) spacekime* metric corresponds to the tensor: $\left(\lambda_{ij}\right) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}$
  - Spacelike intervals correspond to $ds^2 > 0$, where an inertial frame can be found such that two kevents $a, b \in X$ are simultaneous. An object can't be present at two kevents which are separated by a spacelike interval.
  - Lightlike intervals correspond to $ds^2 = 0$. If two kevents are on the line of a photon, then they are separated by a lightlike interval and a ray of light could travel between the two kevents.
  - Kimelike intervals correspond to $ds^2 < 0$. An object can be present at two different kevents, which are separated by a kimelike interval.

Kime Parameterizations



Spacekime Math Generalizations

- Spacekime generalization of Lorentz transform between two reference frames, K & K': (ds is Lorentz transform invariant)

$$\underbrace{\begin{pmatrix} x' \\ y' \\ z' \\ k_1' \\ k_2' \end{pmatrix}}_{\in K'} = \begin{pmatrix} \zeta & 0 & 0 & -\frac{c^2}{v_1}\beta^2\zeta & -\frac{c^2}{v_2}\beta^2\zeta \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -\frac{1}{v_1}\beta^2\zeta & 0 & 0 & 1+(\zeta-1)\frac{c^2}{(v_1)^2}\beta^2 & (\zeta-1)\frac{c^2}{v_1 v_2}\beta^2 \\ -\frac{1}{v_2}\beta^2\zeta & 0 & 0 & (\zeta-1)\frac{c^2}{v_1 v_2}\beta^2 & 1+(\zeta-1)\frac{c^2}{(v_2)^2}\beta^2 \end{pmatrix} \underbrace{\begin{pmatrix} x \\ y \\ z \\ k_1 \\ k_2 \end{pmatrix}}_{\in K}$$

- Kime (Wirtinger) derivative & acceleration (second order kime-derivative at $\boldsymbol{k} = (r, \varphi)$):

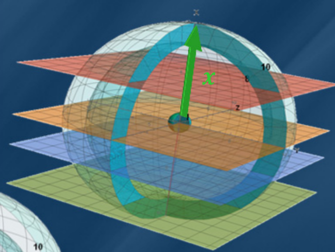
$$f''(\boldsymbol{k}) = \frac{1}{2}\left(\underbrace{cos(2\varphi)\frac{\partial^2 f}{\partial r^2} - \frac{2}{r}sin2\varphi\frac{\partial^2 f}{\partial r\partial\varphi} - \frac{1}{r^2}cos(2\varphi)\frac{\partial^2 f}{\partial \varphi^2}}_{real} \underbrace{-\,i\left(sin(2\varphi)\frac{\partial^2 f}{\partial r^2} + \frac{2}{r}cos(2\varphi)\frac{\partial^2 f}{\partial r\partial\varphi} - \frac{1}{r^2}sin(2\varphi)\frac{\partial^2 f}{\partial \varphi^2}\right)}_{imaginary} \right)$$

- Derived other spacekime concepts: calculus of differentiation & integration, law of addition of velocities, energy-momentum conservation law, stability conditions for particles moving in spacekime, conditions for nonzero rest particle mass, and causal structure of spacekime …

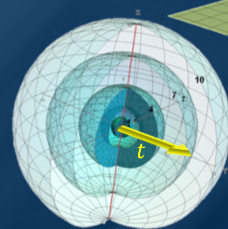
Dinov & Velev (2020)

Spacekime Foliations

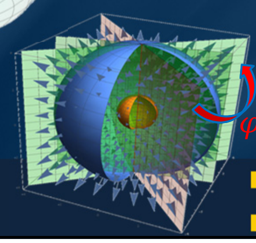
<u>Manifold foliation</u> (*spacekime slicing*) is a covering decomposition into hypersurfaces of lower dimension *(e.g., n-1)* paired with a smooth scalar field (regular with non-trivial gradient), so that each hypersurface (leaf) is a level surface of the scalar field.



- ❑ Space (x) Foliation of Spacekime:



- ❑ (Radial, t) Time-Foliation of Spacekime:



- ❑ (Angular, φ) Phase-Foliation of *Kime*:

Spacekime Connection to Data Analytics?

Mathematical-Physics ⟹ Data Science

Mathematical-Physics	Data Science
A **particle** is a small localized object that permits observations and characterization of its physical or chemical properties	An **object** is something that exists by itself, actually or potentially, concretely or abstractly, physically or incorporeal (e.g., person, subject, etc.)
An **observable** a dynamic variable about particles that can be measured	A **feature** is a dynamic variable or an attribute about an object that can be measured
Particle **state** is an observable particle characteristic (e.g., position, momentum)	**Datum** is an observed quantitative or qualitative value, an instantiation, of a feature
Particle **system** is a collection of independent particles and observable characteristics, in a closed system	**Problem**, aka Data System, is a collection of independent objects and features, without necessarily being associated with apriori hypotheses
Wave-function	**Inference-function**
Reference-Frame **transforms** (e.g., Lorentz)	Data **transformations** (e.g., wrangling, log-transform)
State of a system is an observed measurement of all particles ~ wavefunction	**Dataset (data)** is an observed instance of a set of datum elements about the problem system, $O = \{X, Y\}$
A **particle system is computable** if (1) the entire system is logical, consistent, complete and (2) the unknown internal states of the system don't influence the computation (wavefunction, intervals, probabilities, etc.)	**Computable data object** is a very special representation of a dataset which allows direct application of computational processing, modeling, analytics, or inference based on the observed dataset
…	…

Mathematical-Physics ⟹ Data Science

Math-Physics

Wavefunction

Wave equ problem:

$$\left(\frac{\partial^2}{\partial x^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t}\right)\psi(x,t) = 0$$

Complex Solution: $\psi(x,t) = Ae^{i(kx-wt)}$

where $\left|\frac{w}{k}\right| = v$,

represents a traveling wave

Data Science

Inference function - describing a solution to a specific data analytic system (a problem). For example,

- A linear (GLM) model represents a solution of a prediction inference problem, $Y = X\beta$, where the inference function quantifies the effects of all independent features (X) on the dependent outcome (Y), data: $O = \{X, Y\}$:

$$\psi(O) = \psi(X,Y) = \hat{\beta} = \hat{\beta}^{OLS} = \langle X|X\rangle^{-1}\langle X|Y\rangle = \left(X^T X\right)^{-1} X^T Y.$$

- A non-parametric, non-linear, alternative inference is SVM classification. If $\psi_x \in H$, is the lifting function $\psi: R^\eta \to R^d$ ($\psi: x \in R^\eta \to \tilde{x} = \psi_x \in H$), where $\eta \ll d$, the kernel $\psi_x(y) = \langle x|y\rangle: O \times O \to R$ transformed non-linear to linear separation, the observed data $O_i = \{x_i, y_i\} \in R^\eta$ are lifted to $\psi_{O_i} \in H$. Then, the SVM prediction operator is the weighted sum of the kernel functions at ψ_{O_i}, where β^* is a solution to the SVM regularized optimization:

$$\langle \psi_O | \beta^* \rangle_H = \sum_{i=1}^{n} p_i^* \langle \psi_O | \psi_{O_i} \rangle_H .$$

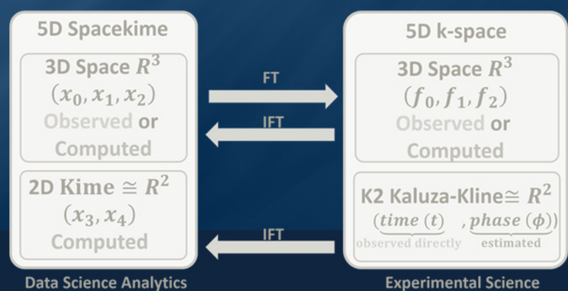
The linear coefficients, p_i^*, are the dual weights that are multiplied by the label corresponding to each training instance, $\{y_i\}$.

Inference always depends on the (input) data; however, it does not have 1-1 and onto bijective correspondence with the data, since the inference function quantifies predictions in a probabilistic sense.

GLM/SVM: http://DSPA.predictive.space | Dinov, Springer (2018)

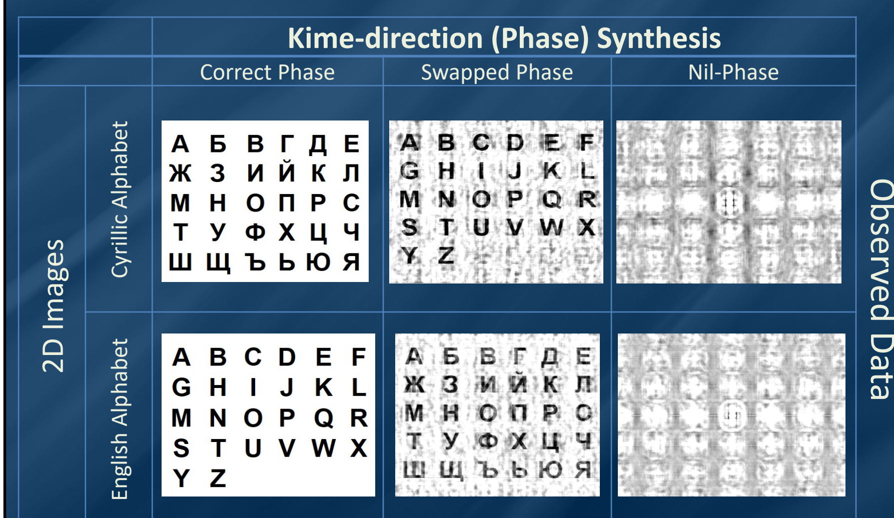
Spacekime Analytics

- Let's assume that we have:
  (1) Kime extension of Time, and
  (2) Parallels between wavefunctions ↔ inference functions
- Often, we can't directly observe (record) data natively in 5D spacekime.
- Yet, we can measure quite accurately the kime-magnitudes (r) as event orders, "times".
- To reconstruct the 2D spatial structure of kime, borrow tricks used by crystallographers [1] to resolve the structure of atomic particles by only observing the magnitudes of the diffraction pattern in k-space. This approach heavily relies on (1) prior information about the kime directional orientation (that may be obtained from using similar datasets and phase-aggregator analytical strategies), or (2) experimental reproducibility by repeated confirmations of the data analytic results using longitudinal datasets.



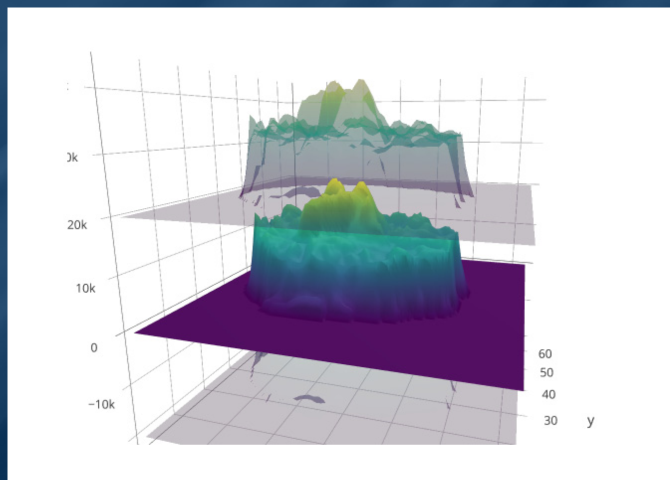
[1] Rodriguez, Ivanova, Nature 2015

2D Image Analysis / Character Recognition



Dinov & Velev (2020)

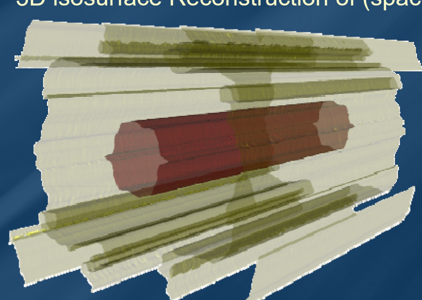
Back to fMRI (4D spacetime data)

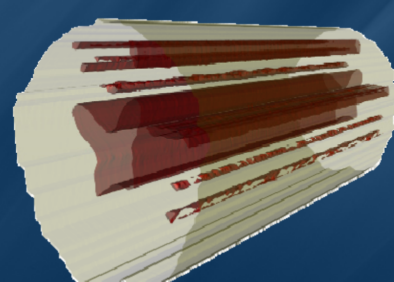


3D rendering of 3 time cross-sections of
the fMRI series over a 2D spatial domain

Spacekime Analytics: fMRI Example

- 3D isosurface Reconstruction of (space=2, time=1) fMRI signal





<u>4D spacetime</u>: Reconstruction using trivial phase-angle; kime=time=(magnitude, 0)

<u>5D Spacekime</u>: Reconstruction using correct kime=(magnitude, phase)

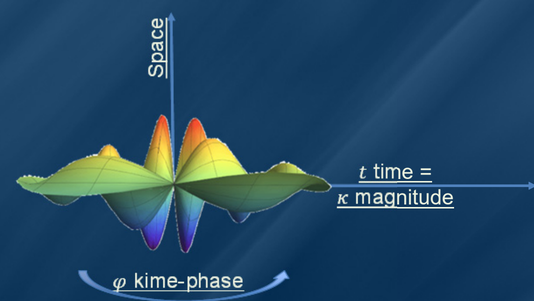
3D pseudo-spacetime reconstruction:

$$f = \hat{h}\,(\underbrace{x_1, x_2}_{space},\ \underbrace{t}_{time})$$

Spacekime Analytics: Kime-series = Surfaces (not curves)

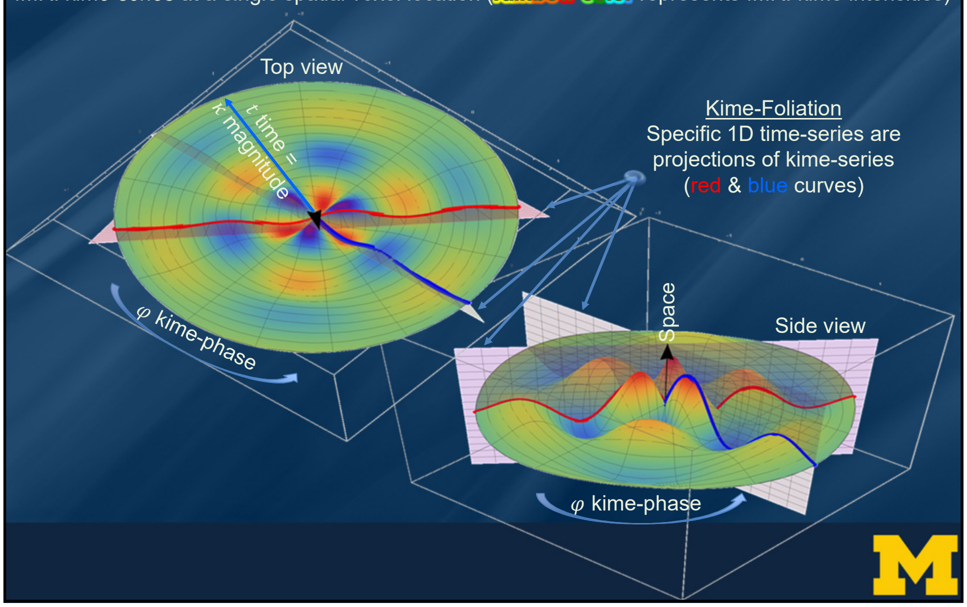
In the 5D spacekime manifold, time-series curves extend to kime-series, i.e., surfaces parameterized by kime-magnitude (t) and the kime-phase (φ).

Kime-phase aggregating operators that can be used to transform standard time-series curves to spacekime kime-surfaces, which can be modeled, interpreted, and predicted using advanced spacekime analytics.



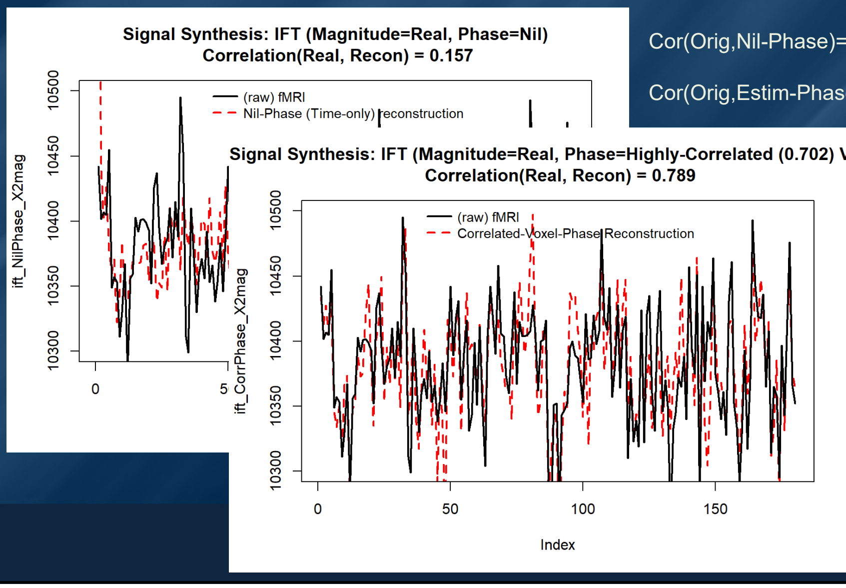
Spacekime Analytics: fMRI kime-series

fMRI *kime-series* at a single spatial voxel location (rainbow color represents fMRI kime intensities)



Spacekime Analytics: fMRI Example

Reconstruction of the fMRI timeseries at a single spatial voxel location



Cor(Orig,Nil-Phase)=0.16

Cor(Orig,Estim-Phase)=0.79

Statistical Implications of Spacekime Analytics

Uncertainty

- Quantum Mechanics: $\|D_x u\|\,\|xu\| = \langle \frac{\hbar}{i}\partial_x u \mid ixu\rangle = \frac{\hbar}{2}\|u\|^2 > 0$, i.e., non-commutation of the unbounded *operators* $D_x = \frac{\hbar}{i}\partial_x$ and x, (multiplication by x).
- Signal processing: Functions can't be time-limited **and** band-limited. Alternatively, a function and its Fourier transform cannot both have bounded domains $\sigma_t \times \sigma_\omega \geq 1/(4\pi)$, where σ_t, σ_ω are the time and frequency SDs.
- Entropic uncertainty: Entropy can be used just like the SD to quantify distribution structure. For instance, for angular, bimodal, or divergent-variance distributions, Entropy may be a better measure of dispersion than SD. For $FT(f)(\omega) = \hat{f}(\omega)$ and $IFT(\hat{f})(x) = \hat{\hat{f}}(x)$, the Shannon information entropies:

$$H_x = \int \hat{\hat{f}}(x)\log\left(\hat{\hat{f}}(x)\right)dx \text{ and } H_\omega = \int \hat{f}(\omega)\log\left(\hat{f}(\omega)\right)d\omega .$$

  satisfy: $H_x + H_\omega \geq \log(e/2)$.
- $L^2(\mathbb{R})$ uncertainty: it is impossible for $f \in L^2$ and $\hat{f}$ to both decrease extremely rapidly. If both have rapidly decreasing tails: $|f(x)| \leq C(1+|x|)^n e^{-a\pi x^2}$ and $\left|\hat{f}(\omega)\right| \leq C(1+|\omega|)^n e^{-b\pi\omega^2}$, for some constant C, polynomial power n, and $a, b \in \mathbb{R}$, then $f = 0$ (when $ab > 1$); $f(x) = P_k(x)e^{-a\pi x^2}$ and $\hat{f}(\omega) = \widehat{P_k}(\omega) * e^{-\omega^2/4\pi a}$, where $\deg(P_k) \leq n$ (when $ab = 1$); or (when $ab < 1$).

Heisenberg's Uncertainty in Spacekime?

- Heisenberg's uncertainty is resolved in 5D spacekime
- We can derive the classical 4D spacetime Heisenberg uncertainty as a reduction of Einstein-like 5D deterministic dynamics:
  - The math is terse – it involves deriving the equations of motion by maximizing the distance (integral along the geodesic) between two points in 5D spacekime
  - The inner product $du^\mu\, dx_\mu = \frac{dx^\mu dx_\mu}{L} = \frac{ds^2}{L}$. Since $\frac{ds}{L} \to 1$ near the leaf membrane, $du^\mu\, dx_\mu = L = \frac{h}{mc}$. Replacing the change in velocity (du^μ) by the change in momentum (dp^μ) yields: $dp^\mu\, dx_\mu = h$.
  - This relation is similar to the quantum mechanics uncertainty principle in 4D Minkowski spacetime; however, it is obtained from 5D Einstein deterministic dynamics. In other words, in spacetime, Heisenberg's uncertainty principal manifests simply because of the one degree of freedom (kime-phase), i.e., lack of sufficient information about the second kime dimension.
  - In 5D spacekime, the conventional geodesic motion is perturbed by an extra force f^μ that can be split into two parts $f^\mu = f_\perp^\mu + f_\parallel^\mu$. The normal component $f_\perp^\mu$ is similar to other conventional forces and obeys the usual orthogonality condition $f_\perp^\mu\, u^\mu = 0$. However, the parallel component $f_\parallel^\mu$ has no analog in 4D spacetime. In general, it has a non-trivial inner product with the 4-velocity u^μ, $f_\parallel^\mu u^\mu \neq 0$.
- In Minkowski 4D spacetime, the lack of kime-phase data naturally leaves one degree of freedom in the system causing Heisenberg's uncertainty. However, the latter can be explicated by information knowledge of the fifth component (kime-phase).

Bayesian Inference Representation

- Suppose we have a single spacetime observation $X = \{x_{i_o}\} \sim p(x \mid \gamma)$ and $\gamma \sim p(\gamma \mid \varphi = \text{phase})$ is a process parameter (or vector) that we are trying to estimate.
- Spacekime analytics aims to make appropriate inference about the process X.
- The sampling distribution, $p(x \mid \gamma)$, is the distribution of the observed data X conditional on the parameter γ and the prior distribution, $p(\gamma \mid \varphi)$, of the parameter γ before the data X is observed, $\varphi = \text{phase aggregator}$.
- Assume that the hyperparameter (vector) φ, which represents the kime-phase estimates for the process, can be estimated by $\hat{\varphi} = \varphi'$.
- Such estimates may be obtained from an oracle, approximated using similar datasets, acquired as phases from samples of analogous processes, or derived via some phase-aggregation strategy.
- Let the posterior distribution of the parameter γ given the observed data $X = \{x_{i_o}\}$ be $p(\gamma|X, \varphi')$ and the process parameter distribution of the kime-phase hyperparameter vector φ be $\gamma \sim p(\gamma \mid \varphi)$.

Bayesian Inference Representation

- We can formulate spacekime inference as a Bayesian parameter estimation problem:

$$\underbrace{p(\gamma|X,\varphi')}_{\text{posterior distribution}} = \frac{p(\gamma,X,\varphi')}{p(X,\varphi')} = \frac{p(X|\gamma,\varphi') \times p(\gamma,\varphi')}{p(X,\varphi')} = \frac{p(X|\gamma,\varphi') \times p(\gamma,\varphi')}{p(X|\varphi') \times p(\varphi')} =$$

$$= \frac{p(X|\gamma,\varphi')}{p(X|\varphi')} \times \frac{p(\gamma,\varphi')}{p(\varphi')} = \frac{p(X|\gamma,\varphi') \times p(\gamma|\varphi')}{\underbrace{p(X|\varphi')}_{\text{observed evidence}}} \propto \underbrace{p(X|\gamma,\varphi')}_{\text{likelihood}} \times \underbrace{p(\gamma|\varphi')}_{\text{prior}}.$$

- In Bayesian terms, the posterior probability distribution of the unknown parameter γ is proportional to the product of the likelihood and the prior.
- In probability terms, the posterior = likelihood times prior, divided by the observed evidence, in this case, a single spacetime data point, x_{i_o}.

Bayesian Inference Representation

- Spacekime analytics based on a single spacetime observation x_{i_o} can be thought of as a type of Bayesian prior-predictive *or* posterior-predictive distribution estimation problem.
- Prior predictive distribution of a new data point x_{j_o}, marginalized over the *prior* – i.e., the sampling distribution $p(x_{j_o}|\gamma)$, weight-averaged by the pure *prior* distribution):

$$p(x_{j_o}|\varphi') = \int p(x_{j_o}|\gamma) \times \underbrace{p(\gamma|\,\varphi')}_{\text{prior distribution}} \, d\,\gamma \; .$$

- Posterior predictive distribution of a new data point x_{j_o}, marginalized over the *posterior* ; i.e., the sampling distribution $p(x_{j_o}|\gamma)$, weight-averaged by the *posterior* distribution:

$$p(x_{j_o}|x_{i_o},\varphi') = \int p(x_{j_o}|\gamma) \times \underbrace{p(\gamma|x_{i_o},\varphi')}_{\text{posterior distribution}} \, d\,\gamma \; .$$

- The difference between these two predictive distributions is that
  - the posterior predictive distribution is updated by the observation $X = \{x_{i_o}\}$ and the hyperparameter, φ (phase aggregator),
  - whereas the prior predictive distribution only relies on the values of the hyperparameters that appear in the prior distribution.

Bayesian Inference Representation

- The posterior predictive distribution may be used to sample or forecast the distribution of a prospective, yet unobserved, data point x_{j_o}.
- The posterior predictive distribution spans the entire parameter state-space ($\mathrm{Domain}(\gamma)$), just like the wavefunction represents the distribution of particle positions over the complete particle state-space.
- Using maximum likelihood or maximum a posteriori estimation, we can also estimate an individual parameter point-estimate, γ_o. In this frequentist approach, the point estimate may be plugged into the formula for the distribution of a data point, $p(x \mid \gamma_o)$, which enables drawing IID samples or individual outcome values.

Bayesian Inference Simulation

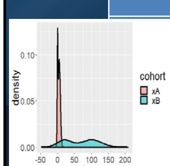
- Simulation example using 2 random samples drawn from mixture distributions each of $n_A = n_B = 10\text{K}$ observations:
  - $\{X_{A,i}\}_{i=1}^{n_A}$, where $X_{A,i} = 0.3U_i + 0.7V_i$, $U_i \sim N(0,1)$ and $V_i \sim N(5,3)$, and
  - $\{X_{B,i}\}_{i=1}^{n_B}$, where $X_{B,i} = 0.4P_i + 0.6Q_i$, $P_i \sim N(20,20)$ and $Q_i \sim N(100,30)$.

- The intensities of cohorts A and B are independent and follow different mixture distributions. We'll split the first cohort (A) into training (C) and testing (D) subgroups, and then:
  - Transform all four cohorts into Fourier k-space,
  - Iteratively randomly sample single observations from cohort C,
  - Reconstruct the data into spacetime using a single kime-magnitude value and alternative kime-phase estimates derived from cohorts B, C, and D, and
  - Compute the classical spacetime-derived population characteristics of cohort A and compare them to their spacekime counterparts obtained using a single C kime-magnitude paired with B, C, or D kime-phases.

Bayesian Inference Simulation

Summary statistics for the original process (cohort A) and the corresponding values of their counterparts computed using the spacekime reconstructed signals based on kime-phases of cohorts B, C, and D. The estimates for the latter three cohorts correspond to reconstructions using a single spacetime observation (i.e., single kime-magnitude) and alternative kime-phases (in this case, kime-phases derived from cohorts B, C, and D).

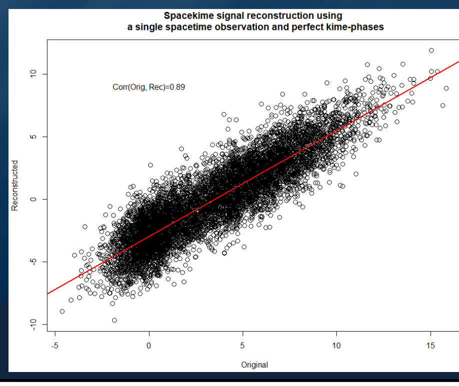
	Spacetime	Spacekime Reconstructions (single kime-magnitude)		
Summaries	(A) Original	(C) Phase=True	(B) Phase=Diff. Process	(D) Phase=Independent
Min	-2.38798	-2.98116	-3.798440	-2.69808
1st Quartile	-0.89359	-0.76765	-0.636799	-0.76453
Median	0.03311	-0.05982	0.009279	-0.08329
Mean	0.00000	0.00000	0.000000	0.00000
3rd Quartile	0.75772	0.72795	0.645119	0.69889
Max	3.61346	3.64800	3.986702	3.22987
Skewness	0.348269	0.2372526	0.001021943	0.31398
Kurtosis	-0.68176	-0.4452207	0.2149918	-0.3270084



Bayesian Inference Simulation

The correlation between the original data (A) and its <u>reconstruction using a single kime magnitude</u> and the correct kime-phases (C) is $\rho(A, C) = 0.89$.

This strong correlation suggests that a substantial part of the A process <u>energy can be recovered using only a single observation</u>. In this case, to reconstruct the signal back into spacetime and compute the corresponding correlation, we used a single kime-magnitude (sample-size=1) and process C kime-phases.



Bayesian Inference Simulation

Let's demonstrate the Bayesian inference corresponding to this spacekime data analytic problem using a simulated bimodal experiment:

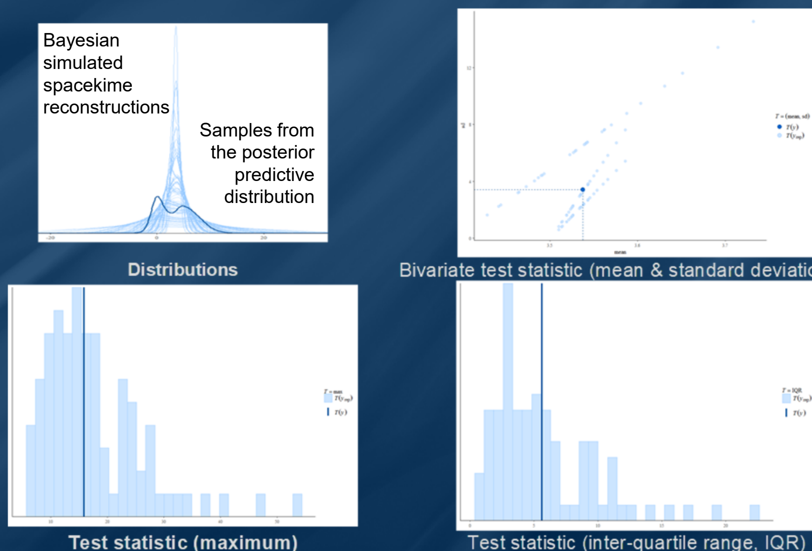
$X_A = 0.3U + 0.7\mathrm{V}$, where $U \sim N(0,1)$ and $V \sim N(5,3)$

Specifically, we will illustrate the Bayesian inference using repeated single spacetime observations from cohort A, $X = \{x_{i_o}\}$, and varying kime-phase priors ($\theta = \text{phase aggregator}$) obtained from cohorts B, C, or D, using different posterior predictive distributions

Relations between the empirical data distribution (<u>**dark blue**</u>) and samples from the posterior predictive distribution, representing Bayesian simulated spacekime reconstructions (<u>**light-blue**</u>). The derived Bayesian estimates do not perfectly match the empirical distribution of the simulated data, yet there is clearly information encoding that is captured by the spacekime data reconstructions

This <u>signal compression</u> can be exploited by subsequent model-based or model-free data analytic strategies for retrospective prediction, prospective forecasting, ML classification, derived clustering, and other spacekime inference methods

Bayesian Inference Simulation

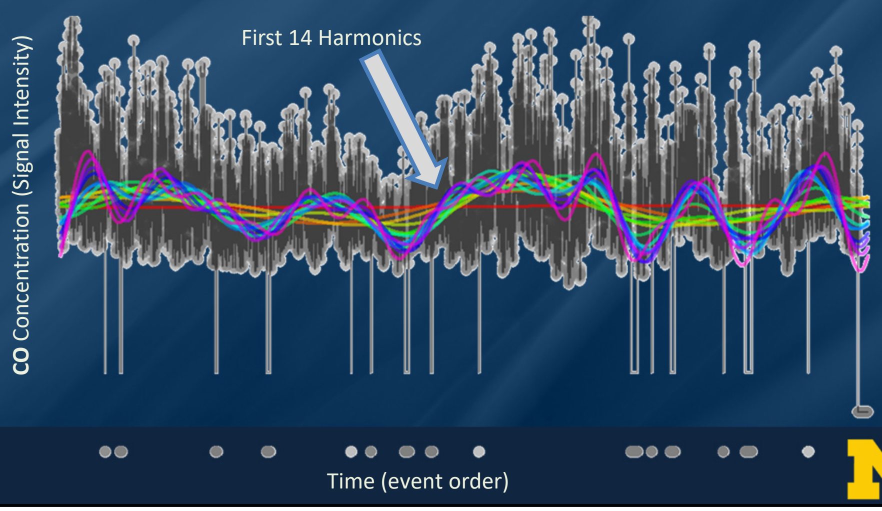


Relations between the empirical data distribution (**dark blue**) and samples from the posterior predictive distribution, Bayesian simulated spacekime reconstructions (**light-blue**).

Applications – Longitudinal Spacekime Data Analytics

Exogenous Feature Time-series Analysis

ARIMAX modeling of UCI ML Air Quality Dataset (9,358 hourly-averaged CO responses from an array of sensors). Demonstrate the effect of kime-direction on the analysis of the longitudinal data.



Exogenous Feature Time-series Analysis

CO ARIMAX models derived on 3 different signal reconstructions based on alternative kime-direction estimates

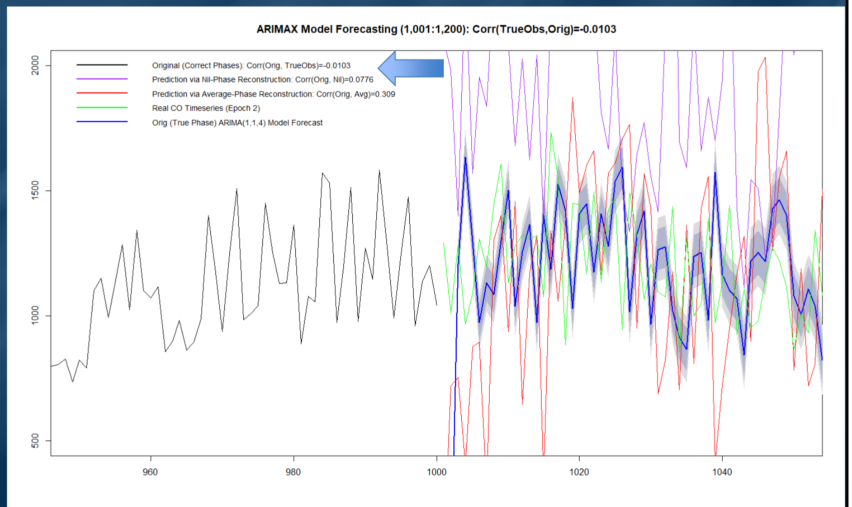
Phase	Nil	Average	True=original
Model Estimate	ARIMA(2,0,1)	ARIMA(2,0,3)	ARIMA(1,1,4)
AIC	13179	14183	10581
ar1	1.11406562	0.329482302	0.2765312
ar2	-0.14565048	0.238363531	.
ma1	-0.78919188	0.267291585	-0.88913497
ma2	.	-0.006079386	0.12679494
ma3	.	0.15726556	0.03043726
ma4	.	.	-0.17655728
intercept	503.3455144	742.800113	.
xreg1	-0.40283891	0.58379483	0.08035744
xreg2	0.13656613	0.280936931	6.14947902
xreg3	-0.51457636	-0.649722755	0.09859223
xreg4	1.09611981	1.239910298	0.01634736
xreg5	1.21946209	-0.026110332	-0.04816591
xreg6	1.30628469	1.081777956	-0.01104142
xreg7	1.20868397	0.254018471	0.1832854
xreg8	1.14905809	0.306524131	0.17648482
xreg9	-0.48233756	-0.405204908	6.53739782
xreg10	0.03145281	0.351063312	1.79388326
xreg11	-0.46395772	-0.457689796	-12.06965578

ARIMAX (*p,d,q*)

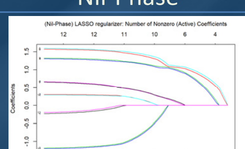
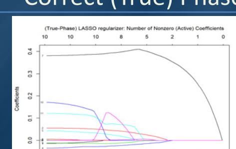
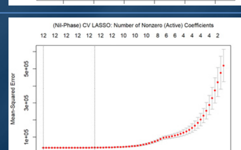
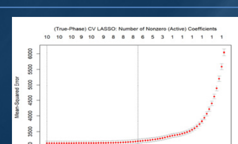
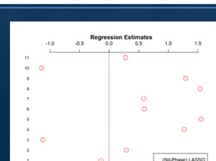
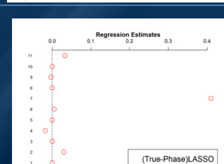
p = order (# of time lags) of the AR part
d = differencing (# of past values subtractions)
q = order of MA part

Exogenous Feature Timeseries Forecasting

CO ARIMAX models derived on 3 different signal reconstructions based on alternative kime-direction estimates



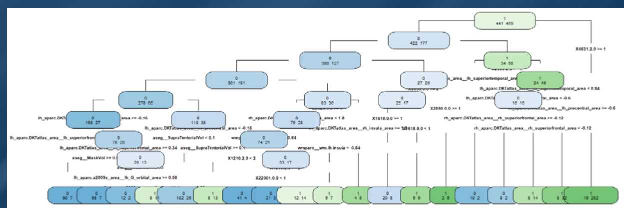
Exogenous Feature Time-series Analysis

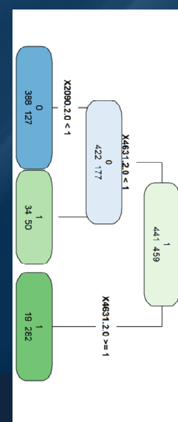
	Synthesis Approach	
	Nil-Phase	Correct (True) Phase
Number of Nonzero (Active) LASSO Coefficients		
LASSO Mean Square Error CV Error of Model Coefficients		
LASSO Regression Model Coefficients		

Results of **regularized linear modeling** of CO-concentration using LASSO penalty

Big Data Analytics Study – UKBB

- 9,914 UKBB participants; 7,614 features:
  Features: clinical+phenotypic variables (5K) and derived neuroimaging biomarkers (2.5K)
- Supervised Decision Tree (binary Dx) Classification – **Correct Kime-Phase Estimates**





Zhou, et al. SREP (2019)

Raw Decision Tree

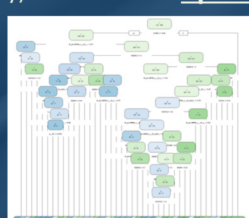
```
## Prediction   0   1
##        0     362  60
##        1      79 399
##              Accuracy : 0.8456
##           95% CI : (0.82, 0.87)
##    No Information Rate : 0.51
##    P-Value [Acc > NIR] : <2e-16
##                Kappa : 0.6907
## Mcnemar's Test P-Value : 0.1268
##           Sensitivity : 0.8209
##           Specificity : 0.8693
##         Detection Rate : 0.4022
##    Detection Prevalence : 0.4689
##       Balanced Accuracy : 0.8451
```

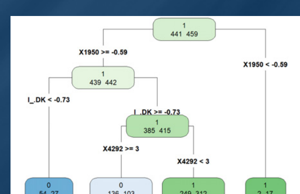
Pruned Decision Tree

```
## Prediction   0   1
##        0     388 127
##        1      53 332
##              Accuracy : 0.8
##           95% CI : (0.77, 0.83)
##    No Information Rate : 0.51
##    P-Value [Acc > NIR] : < 2.2e-16
##                Kappa : 0.6012
## Mcnemar's Test P-Value : 5.295e-08
##           Sensitivity : 0.8798
##           Specificity : 0.7233
##   Detection Prevalence : 0.5722
##      Balanced Accuracy : 0.8016
```

Big Data Analytics Study – UKBB

- 9,914 UKBB participants (11 epochs of 900 cases); 7,614 clinical measurements, phenotypic features, and derived neuroimaging biomarkers Supervised Decision Tree (binary) Classification – **Epoch-average Kime-Phases**





Raw Decision Tree

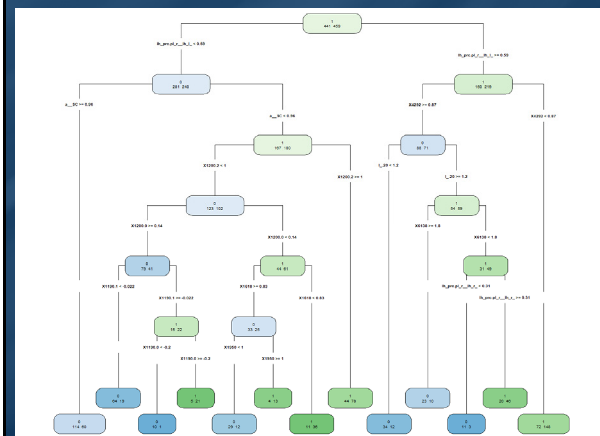
```
##        Reference
## Prediction   0   1
##        0 354  85
##        1  87 374
##              Accuracy : 0.8089
##           95% CI : (0.78, 0.83)
##    No Information Rate : 0.51
##    P-Value [Acc > NIR] : <2e-16
##                Kappa : 0.6176
## Mcnemar's Test P-Value : 0.9392
##           Sensitivity : 0.8027
##           Specificity : 0.8148
##         Detection Rate : 0.3933
##    Detection Prevalence : 0.4878
##       Balanced Accuracy : 0.8088
```

Pruned Decision Tree

```
##        Reference
## Prediction   0   1
##        0 190 130
##        1 251 329
##              Accuracy : 0.5767
##           95% CI : (0.54, 0.61)
## No Information Rate : 0.51
## P-Value [Acc > NIR] : 3.501e-05
##                Kappa : 0.1484
## Mcnemar's Test P-Value : 7.857e-10
##           Sensitivity : 0.4308
##           Specificity : 0.7168
##         Detection Rate : 0.2111
## Detection Prevalence : 0.3556
##   Balanced Accuracy : 0.5738
```

Big Data Analytics Study – UKBB

- 9,914 UKBB participants; 7,614 clinical measurements, phenotypic features, and derived neuroimaging biomarkers
- Supervised Decision Tree (binary) Classification – **Nil-average Kime-Phases**



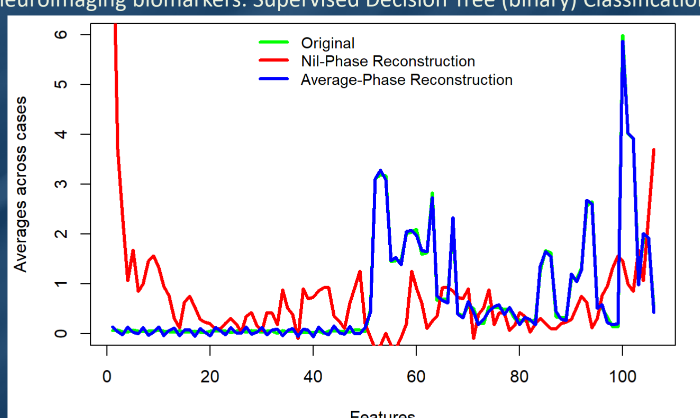
Raw Decision Tree

```
##          Reference
## Prediction   0   1
##         0 341  86
##         1 100 373
##         Accuracy : 0.7933
##           95% CI : (0.77, 0.82)
##    No Information Rate : 0.51
##    P-Value [Acc > NIR] : <2e-16
##              Kappa : 0.5862
##  Mcnemar's Test P-Value : 0.3405
##           Sensitivity : 0.7732
##           Specificity : 0.8126
##         Detection Rate : 0.3789
##    Detection Prevalence : 0.4744
##      Balanced Accuracy : 0.7929
```

Pruned Decision Tree (not shown) was degenerate

Big Data Analytics Study – UKBB

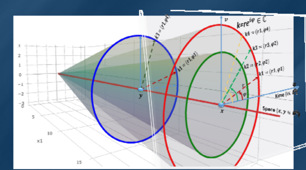
- 9,914 UKBB participants; 7,614 clinical measurements, phenotypic features, and derived neuroimaging biomarkers. Supervised Decision Tree (binary) Classification



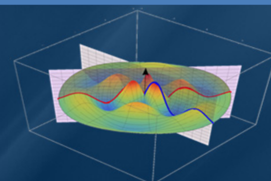
Overall feature averages across cases for the 3 complementary kime-reconstruction analytic strategies

Summary

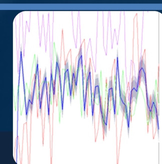
- ❑ Need new methods to tackle substantial Big Biomed/Health Data Challenges
- ❑ *Spacekime* representation makes a difference in predictive analytics



- ❑ Math models useful for representation & analysis of complex-temporal data
- ❑ *Spacekime transform* enables small sample inference



- ❑ Optimal *kime-phase aggregators*?
- ❑ Spacekime analytics representation has lots of *"Open problems" (math, stats, DS)*



Interested in Spacekime Analytics?

- ❑ Check www.SpaceKime.org
- ❑ Contact me
- ❑ We have lots of "Open Problems"

Acknowledgments

Slides Online: "SOCR News"

Funding

NIH: P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, UL1TR002240, R01CA233487

NSF: 1916425, 1734853, 1636840, 1416953, 0716055, 1023115

Collaborators

- SOCR: **Milen Velev, Yongkai Qiu, Zhe Yin, Yufei Yang**, Alexandr Kalinin, Selvam Palanimalai, Syed Husain, Matt Leventhal, Ashwini Khare, Rami Elkest, Abhishek Chowdhury, Patrick Tan, Pratyush Pati, Brian Zhang, Juana Sanchez, Dennis Pearl, Kyle Siegrist, Rob Gould, Nicolas Christou, Hanbo Sun, Tuo Wang, Yi Wang, Lu Wei, Lu Wang, Simeone Marino
- LONI/INI: Arthur Toga, Roger Woods, Jack Van Horn, Zhuowen Tu, Yonggang Shi, David Shattuck, Elizabeth Sowell, Katherine Narr, Anand Joshi, Shantanu Joshi, Paul Thompson, Luminita Vese, Stan Osher, Stefano Soatto, Seok Moon, Junning Li, Young Sung, Carl Kesselman, Fabio Macciardi, Federica Torri
- UMich MIDAS/MNORC/AD/PD Centers: Cathie Spino, Chuck Burant, Ben Hampstead, Kayvan Najarian, Stephen Goutman, Stephen Strobbe, Hiroko Dodge, Hank Paulson, Bill Dauer, Roger Albin, Chris Monk, Issam El Naqa, HV Jagadish, Brian Athey

http://SOCR.umich.edu