Complex-time (*Kime*) Representation of Longitudinal Data and Spacekime Analytics

Ivo D. Dinov

Statistics Online Computational Resource
Health Behavior & Biological Sciences
Computational Medicine & Bioinformatics
Michigan Institute for Data Science
University of Michigan
https://SOCR.umich.edu

Joint work with Milen V. Velev (BTU) & Yueyang Shen (UM)

Based on the book "*Data Science: Time Complexity, Inferential Uncertainty & Spacekime Analytics*"

STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)
UNIVERSITY OF MICHIGAN

Slides Online:
"SOCR News"

1

Big Data Characteristics & Challenges

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

Big Bio Data Dimensions	Specific Challenges
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, joint multivariate representation & modeling
Multi-scale	Interpreting macro → meso → micro → nano scale observations
Time	Techniques accounting for longitudinal effects (e.g., time corr)
Incomplete	Reliable management of missing data, imputation, obfuscation

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

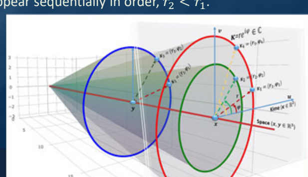
Software developments, student training, service platforms and methodological advances associated with the Big Data Discovery Science all present existing opportunities for learners, educators, researchers, practitioners and policy makers

Dinov, *GigaScience* (2016)   Gao *et al.*, *SciRep* (2018)

3

Complex-Time (*Kime*)

- At a given spatial location, 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 manifold is $\mathbb{R}^3 \times \mathbb{C}$:
  - (x, k_1) and (x, k_4) have the same spacetime representation, but different spacekime coordinates,
  - (x, k_1) and (y, k_1) share the same kime, but represent different spatial locations,
  - (x, k_2) and (x, k_3) have the same spatial-locations and kime-directions, but appear sequentially in order, $r_2 < r_1$.



4

Rationale for *Time* → *Kime* Extension

- **Math** – $Time$ is a special case of *kime*, $\kappa = |\kappa|e^{i\varphi}$ where $\varphi = 0$ (nil-phase)
  - algebraically a *multiplicative* (algebraic) group, (multiplicative) unity (identity) $= 1$
  - multiplicative inverses, multiplicative identity, associativity $t_1 * (t_2 * t_3) = (t_1 * t_2) * t_3$
  - The *time* domain ($\mathbb{R}^+$) is **not** a complete *algebraic field* $(+,*)$:
    - Additive unity (0), element additive inverse $(-t)$: $t + (-t) = 0$; is outside $\mathbb{R}^+$ (time-domain)
    - $x^2 + 1 = 0$ has no solutions in time (or in $\mathbb{R}$) ....

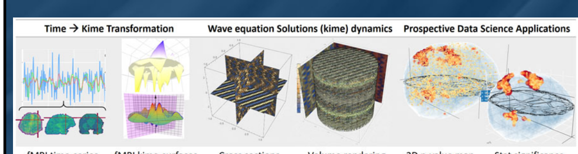
$$\text{Group}(*) \subseteq Ring\left(\overbrace{\underbrace{(+,*)}_{\text{associative \& distributive}}}^{\text{Compatible operations}}\right) \subseteq Field\ \overbrace{(+,*)}^{\text{Group}(+)}$$

  - Classical time ($\mathbb{R}^+$) is a *positive cone* over the field of the real numbers ($\mathbb{R}$)
  - Time forms a subgroup of the multiplicative group of the reals
  - Whereas kime ($\mathbb{C}$) is an algebraically *closed prime field* that naturally extends time
  - *Time* is ordered & *kime* is not – the kime magnitude preserves the intrinsic time order
  - Kime ($\mathbb{C}$) represents the smallest natural extension of time, complete filed that agrees with time
  - The *time* group is closed under addition, multiplication, and division (but not subtraction). It has the topology of $\mathbb{R}$ and the structure of a multiplicative topological group ≡ additive topological semigroup
- **Physics** –
  - Problem of time … (DOI 10.1007/978-3-319-58848-3)
  - $\mathbb{R}$ and $\mathbb{C}$ Hilbert-space quantum theories make different predictions (DOI: 10.1038/s41586-021-04160-4)
- **AI/Data Science** – Random IID sampling, Bayesian reps, tensor modeling of $\mathbb{C}$ kimesurfaces, novel analytics

Dinov & Velev (2021)

5

Kime transforms → PDEs → AI



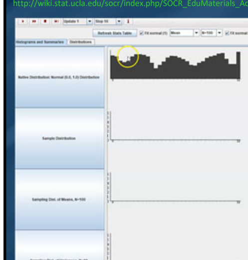
Wang *et al.*, 2022 | Dinov & Velev (2021)

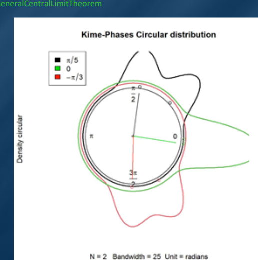
15

Hidden Variable Theory & Random Sampling

- **Phase sampling**: Kime phase distributions are mostly symmetric, random observations

http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_GeneralCentralLimitTheorem



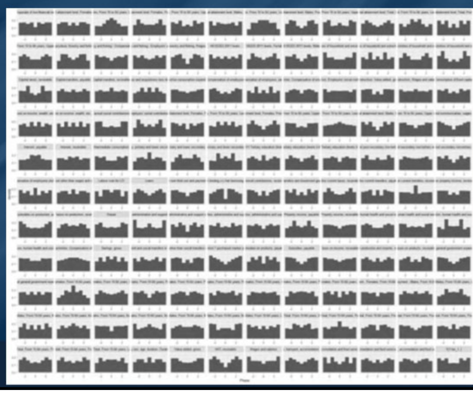


https://www.socr.umich.edu/TCIU/HTMLs/Chapter6_Kime_Phases_Circular.html

Dinov, Christou & Sanchez (2008)   Dinov & Velev (2021)

16

Phase Distributions



132 features - EU Econ Data; Dinov & Velev (2021)

17

(Many) Spacekime Open Math Problems

- **Ergodicity**

Let's look at particle velocities in the 4D Minkowski spacetime (X), a measure space where gas particles move spatially and evolve longitudinally in time. Let $\mu = \mu_x$ be a measure on X, $f(\boldsymbol{x}, t) \in L^1(X, \mu)$ be an integrable function (e.g., velocity of a particle), and $T: X \to X$ be a measure-preserving transformation at position $\boldsymbol{x} \in \mathbb{R}^3$ and time $t \in \mathbb{R}^+$.

A pointwise ergodic theorem argues that in a measure theoretic sense, the average of f (e.g., velocity) over all particles in the gas system at a fixed time, $\bar{f} = E_t(f) = \int_{\mathbb{R}^3} f(\boldsymbol{x}, t) d\mu_x$, will be equal to the average f of just one particle ($\boldsymbol{x}$) over the entire time span,

$$\tilde{f} = \lim_{n \to \infty} \left(\frac{1}{n} \sum_{m=0}^{n} f(T^m \boldsymbol{x})\right), \text{ i.e., (show) } \bar{f} \equiv \tilde{f}.$$

The spatial probability measure is denoted by μ_x and the transformation $T^m \boldsymbol{x}$ represents the dynamics (time evolution) of the particle starting with an initial spatial location $T^o \boldsymbol{x} = \boldsymbol{x}$.

Investigate the ergodic properties of various transformations in the 5D **spacekime**:

$$\underbrace{\bar{f} \equiv E_\kappa(f) = \frac{1}{\mu_x(X)} \int f\left(\boldsymbol{x}, \underbrace{t, \phi}_{\kappa}\right) d\mu_x}_{\text{space averaging}} \stackrel{?}{=} \underbrace{\lim_{t \to \infty} \left(\frac{1}{t} \sum_{m=0}^{t} \left(\int_{-\pi}^{+\pi} f(T^m \boldsymbol{x}, t, \phi) d\Phi\right)\right) \equiv \tilde{f}}_{\text{kime averaging}}$$

Dinov & Velev (2021)

18

Spacekime ←→ Data Science

20

Mathematical-Physics ⟹ Data Science & AI

Physics	Data Science/AI
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 a priori 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, $\boldsymbol{O} = \{\boldsymbol{X}, \boldsymbol{Y}\}$
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
…	…

Dinov & Velev (2021)

21

Mathematical-Physics ⟹ Data Science & AI

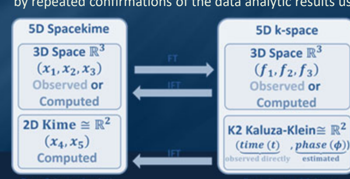
Physics	Data Science
Wavefunction Wave equ problem: $\left(\frac{\partial^2}{\partial x^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t}\right)\psi(x,t) = 0$ Complex Solution: $\psi(x,t) = Ae^{i(kx-wt)}$ where $\left\lvert\frac{w}{k}\right\rvert = v$, represents a traveling wave	*Inference function* - 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, $\boldsymbol{Y} = \boldsymbol{X}\beta$, where the inference function quantifies the effects of all independent features ($\boldsymbol{X}$) on the dependent outcome ($\boldsymbol{Y}$), data: $\boldsymbol{O} = \{\boldsymbol{X}, \boldsymbol{Y}\}$: $\psi(\boldsymbol{O}) = \psi(X,Y) \Rightarrow \hat{\beta} = \hat{\beta}^{OLS} = \langle X \vert X\rangle^{-1}\langle X \vert Y\rangle = (X^TX)^{-1}X^TY.$ • A non-parametric, non-linear, alternative inference is SVM classification. If $\psi_x \in \boldsymbol{H}$, is the lifting function $\psi: \boldsymbol{R}^\eta \to \boldsymbol{R}^d$ ($\psi: x \in \boldsymbol{R}^\eta \to \tilde{x} = \psi_x \in \boldsymbol{H}$), where $\eta \ll d$, the kernel $\psi_x(y) = \langle x \vert y\rangle: \boldsymbol{O} \times \boldsymbol{O} \to \boldsymbol{R}$ transformes non-linear to linear separation, the observed data $\boldsymbol{O}_i = \{x_i, y_i\} \in \boldsymbol{R}^\eta$ are lifted to $\psi_{O_i} \in \boldsymbol{H}$. 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: $\langle \psi_O \vert \beta^* \rangle_H = \sum_{i=1}^{n} p_i^* \langle \psi_O \vert \psi_{O_i} \rangle_H$ 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: https://DSPA2.predictive.space | Dinov, Springer (2018)

22

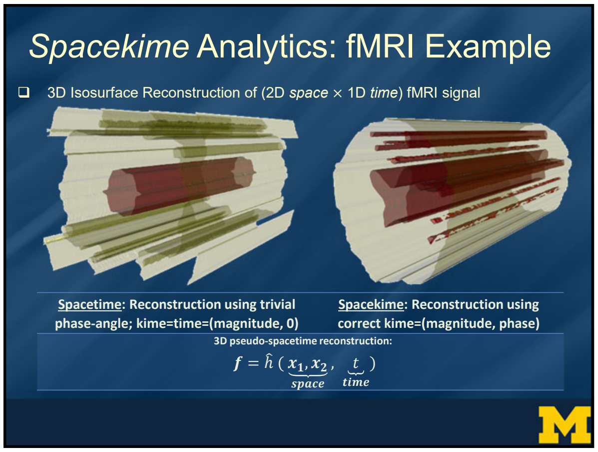
Spacekime Analytics

- 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 models, using similar datasets, or phase-aggregator analytical strategies), or (2) experimental reproducibility by repeated confirmations of the data analytic results using longitudinal datasets

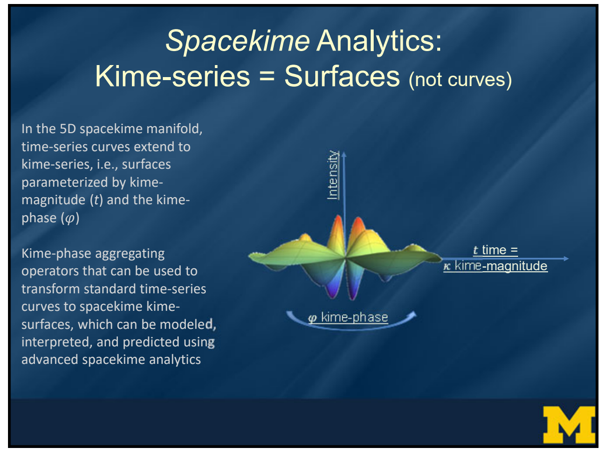


[1] Rodriguez, Ivanova, Nature (2015)

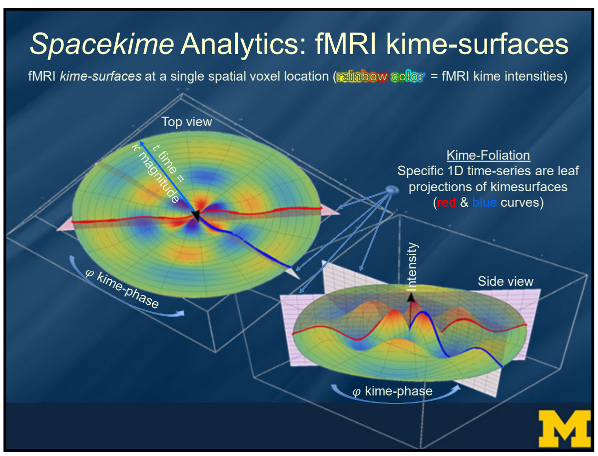
23



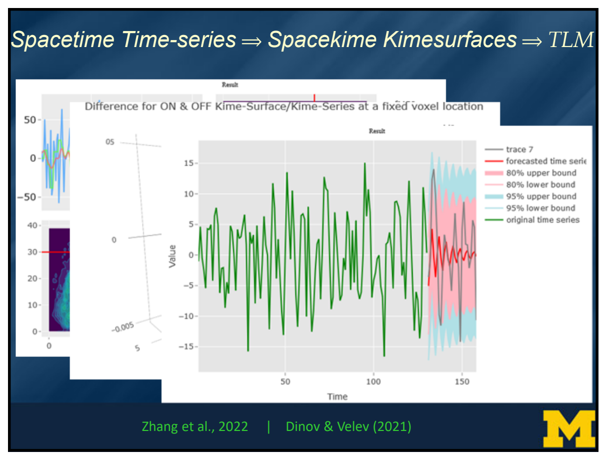
24



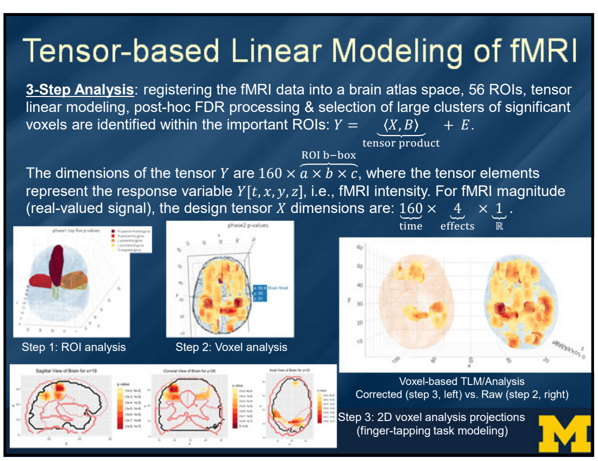
25



26



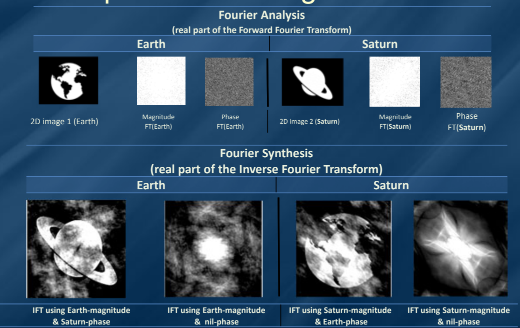
27



28

29

2D Fourier Transform – The Importance of Magnitudes & Phases



32

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 (model distribution), approximated using similar datasets, acquired as phases from samples of analogous processes, derived via some phase-aggregation strategy, or computed via Laplace transform.
- 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 $\gamma \sim p(\gamma \mid \varphi)$.

34

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}.

35

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.

36

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 ($\text{Domain}(\gamma)$), 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.

37

Bayesian Inference Simulation

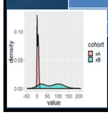
- Simulation example using 2 random samples drawn from mixture distributions each of $n_A = n_B = 10\text{K}$ observations:
  - $\{X_{A,i}\}_{i=1}^{n_A}$, where $X_{A,i} = 0.3U_i + 0.7V_i$, $U_i \sim N(0,1)$ and $V_i \sim N(5,3)$, and
  - $\{X_{B,i}\}_{i=1}^{n_B}$, where $X_{B,i} = 0.4P_i + 0.6Q_i$,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

38

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

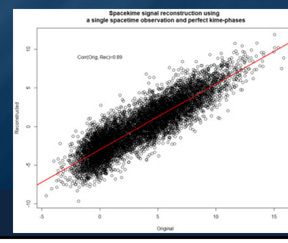


39

Bayesian Inference Simulation

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

This strong correlation suggests that a substantial part of the A process energy can be recovered using only a single observation. 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.



40

Bayesian Inference Simulation

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

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

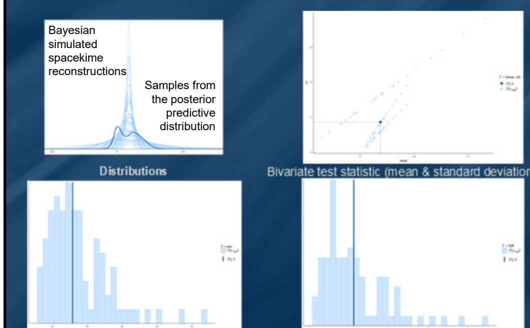
Specifically, we will illustrate the Bayesian inference using repeated single spacetime observations from cohort A, $X = \{x_{i_o}\}$, and varying kime-phase priors ($\theta = \text{phase aggregator}$) obtained from cohorts B, C, or D, using different posterior predictive distributions.

Relations between the empirical data distribution (dark blue) and samples from the posterior predictive distribution, representing Bayesian simulated spacekime reconstructions (light-blue). 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 signal compression can be exploited by subsequent model-based or model-free data analytic strategies for retrospective prediction, prospective forecasting, ML classification, AI derived clustering, and other spacekime inference methods.

41

Bayesian Inference Simulation



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

42

Spacekime Analytics: Demos

- Pubs & Tutorials
  - https://socr.umich.edu/people/dinov/publications.html
  - https://TCIU.predictive.space
  - https://SpaceKime.org
- R Package
  - https://cran.rstudio.com/web/packages/TCIU
- GitHub
  - https://github.com/SOCR/TCIU

43

Acknowledgments

Slides Online: "SOCR News"

Funding

NIH: UL1TR002240, R01CA233487, R01MH121079, R01MH126137, T32GM141746

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

Collaborators

- SOCR: **Milen Velev, Yueyang Shen, Daxuan Deng, Zijing Li, Yongkai Qiu, Zhe Yin, Yufei Yang, Yuxin Wang, Rongqian Zhang, Yuyao Liu, Yupeng Zhang, Yunjie Guo, Simeone Marino**
- **UMich MIDAS/MCAIM Centers**: Lydia Bieri, Kayvan Najarian, Chris Monk, Issam El Naqa, HV Jagadish, Brian Athey

https://SOCR.umich.edu

44