

Stats 225: Bayesian Analysis

More on Gaussian Processes

Babak Shahbaba
UC Irvine

Overview

- ✿ In the previous lecture, we discussed Gaussian process models for regression and classification.
- ✿ In this lecture, we will discuss some advanced Gaussian process models.
- ✿ We specifically discuss the applications of GP in neuroscience, medical diagnostics, and population dynamics.

Longitudinal Data

Analysis

In Collaboration with Dan Gillen's Group

End-stage renal disease

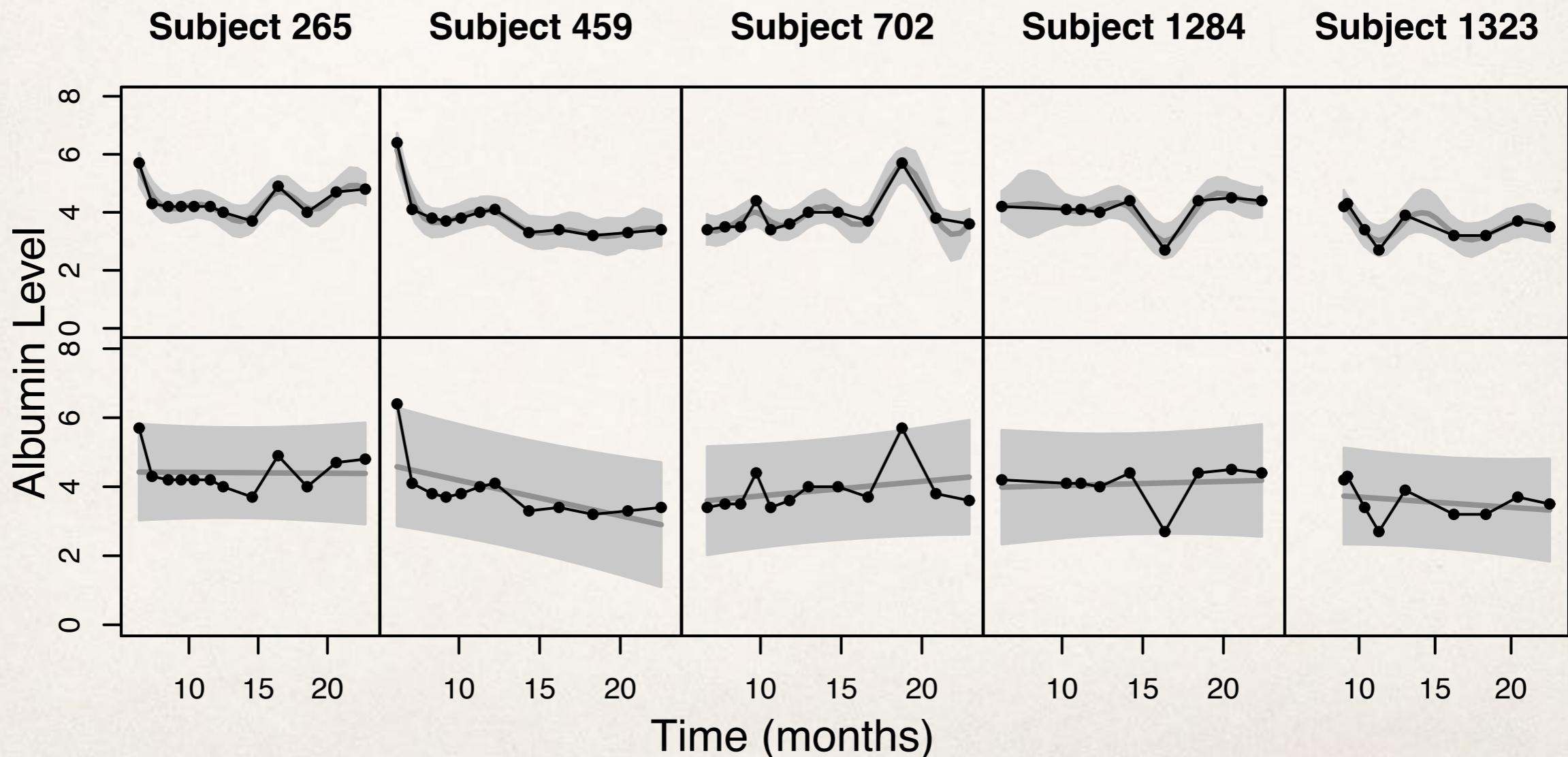
- ⌘ Over 850,000 persons in the United States are being treated for ESRD and many more suffer from early stage chronic kidney disease
- ⌘ The standard of care for adult ESRD patients that do not have access to a viable transplant is hemodialysis
- ⌘ Hemodialysis patients experience extremely high mortality rates.

Monitoring Albumin Level

- ❖ Multiple epidemiologic studies have shown that indices of protein-energy malnutrition (PEM), including serum albumin, are a strong predictor of total mortality among hemodialysis patients (cf. Fung, et al., 2002, Wong, et al. 2002).
- ❖ Fung et al found that among adult hemodialysis subjects, baseline albumin level and the slope of albumin over time were independent risk factors for mortality.
- ❖ Within-subject changes in serum albumin may also be associated with mortality.

Second Moment

- We hypothesized that high within-subject volatility in serum albumin measured over time may be indicative of increased mortality.



Model

- We could use a simple linear regression model

$$Y^L \sim N(X\beta, \sigma^2 I)$$

where $Y^L = (Y(t_1), \dots, Y(t_J))$ denotes albumin measurements at times t_1, \dots, t_J

- This model doesn't capture the volatility of the measurements
- We proposed the following model instead

$$Y^L = X\beta + \mathbf{W}(\mathbf{t}) + \epsilon$$

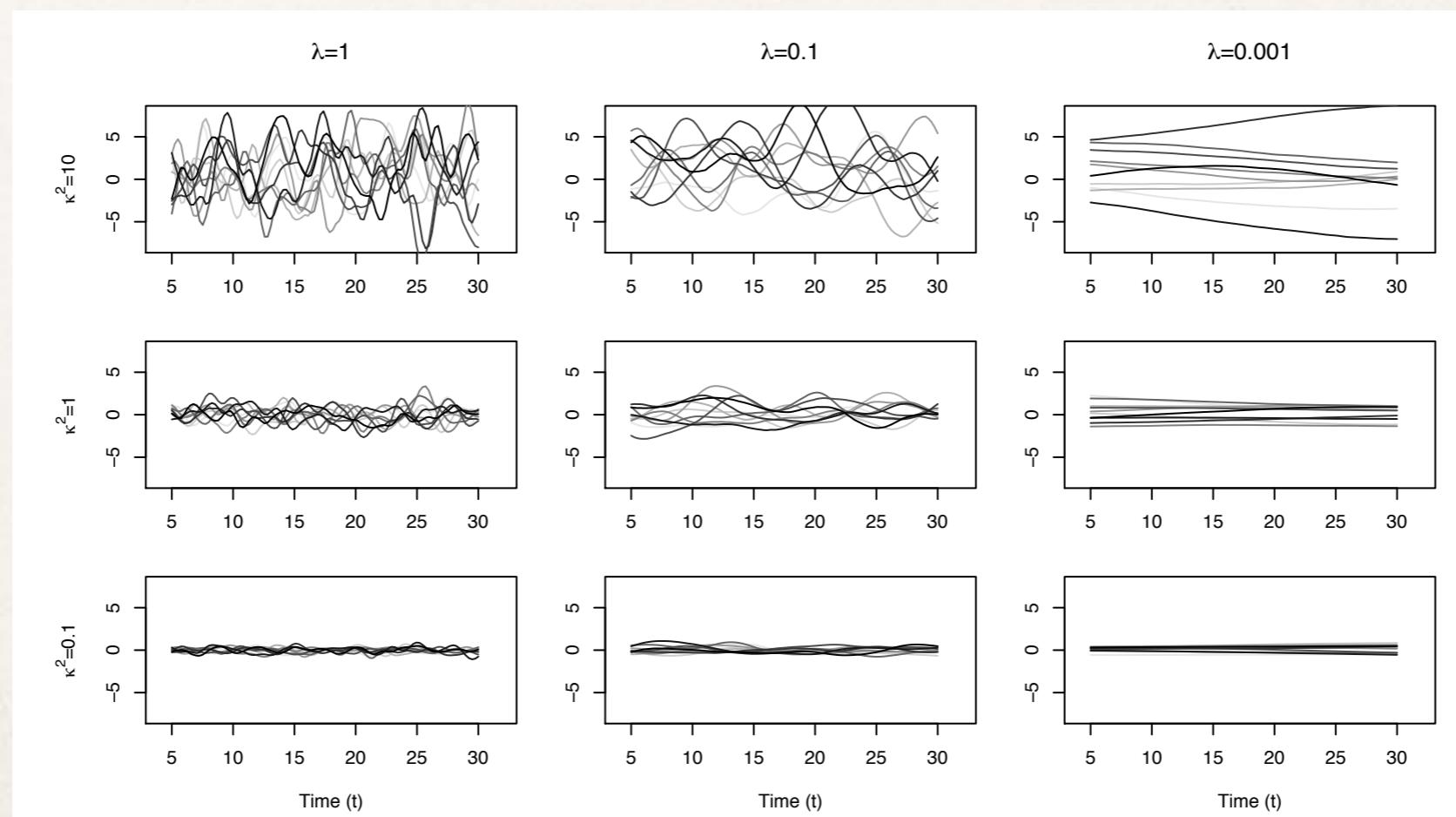
where $\mathbf{W}(\mathbf{t})$ are realizations from a Gaussian Process with mean zero and covariance function $\mathbf{C}(\mathbf{t}, \mathbf{t}')$

Model

- We used the following covariance function

$$C(t, t') = \kappa^2 e^{-\lambda|t-t'|^2}$$

where κ^2 controls the height of the oscillations and λ controls the correlation length between realizations.



Model

- ✿ Larger values of κ^2 produce higher volatility around the mean function
- ✿ Values of κ^2 near 0 produce nearly linear trajectories
- ✿ This makes the GP model a natural choice for the scientific problem being considered as we can focus on κ^2 as the functional of interest:

$$Y_i^L | \theta \sim N_{J_i}(X_i\beta_i, \kappa_i^2 K(\lambda) + \sigma^2 I)$$

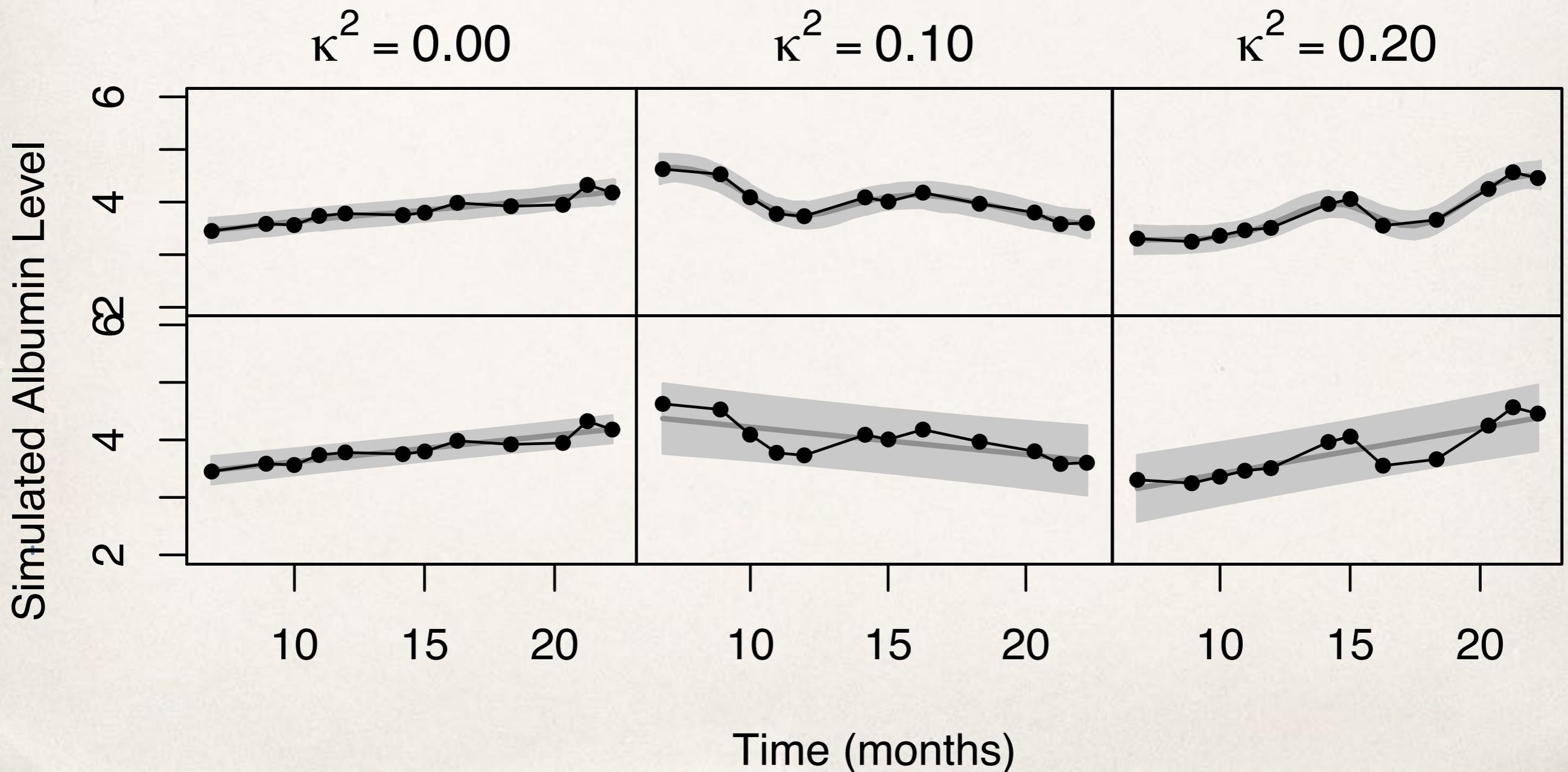
- ✿ Note that λ is shared by all subjects.

Model

- ❖ It is known that at least one of λ or κ^2 must have a (strongly) informative prior for identifiability (Berger et al., 2001)
- ❖ The parameter λ can be estimated by the correlation length of the data
- ❖ The practical range (defined as the distance where the correlation is 0.05) for the GP is given by $\sqrt{3/\lambda}$ (Diggle and Ribeiro, 2007)
- ❖ In our data, we take $\lambda=0.1$ so the practical range is 5.76 months

Results

- We found significant association between volatility of albumin measurements (κ^2) and mortality.



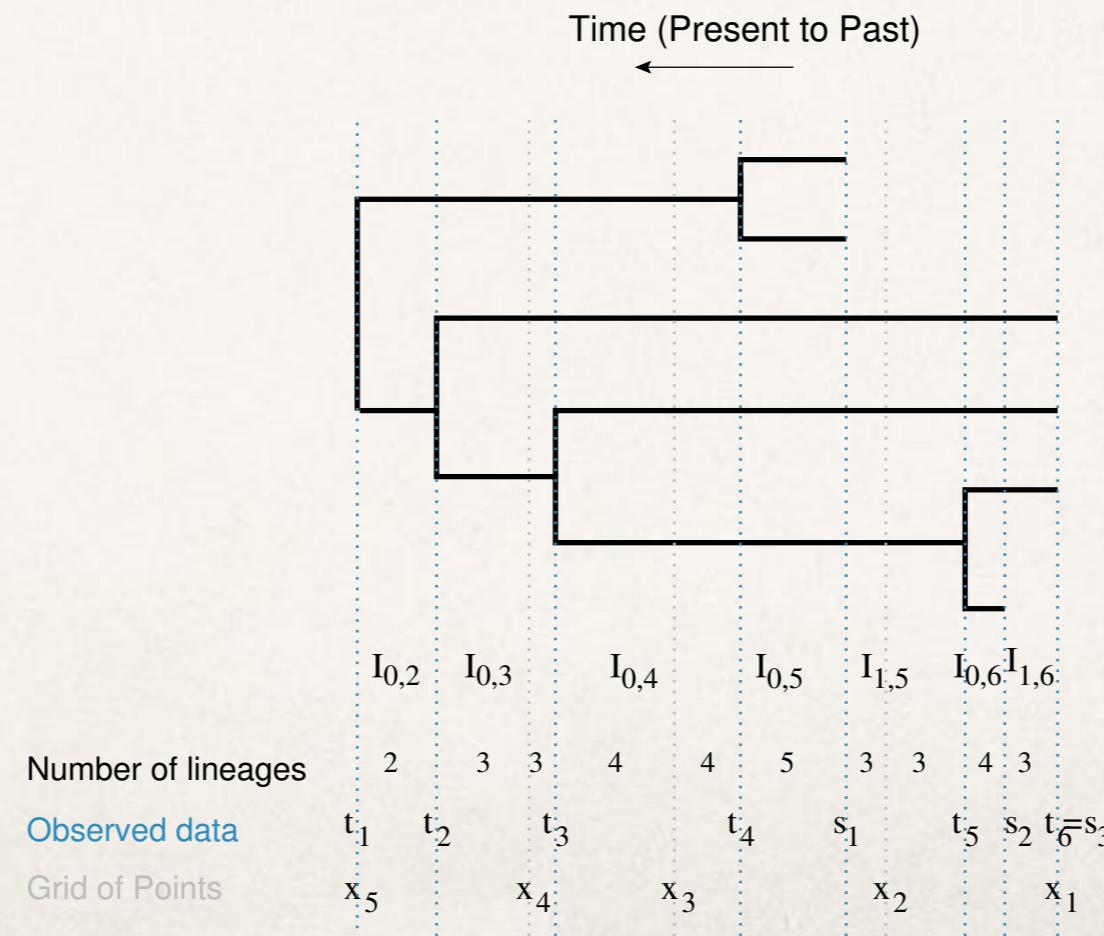
Population Dynamics

In Collaboration with Vladimir Minin's Group

Lan, S., Palacios, J., Karcher, M., Minin, V., Shahbaba, B. (2015) An Efficient Bayesian Inference Framework for Coalescent-Based Nonparametric Phylodynamics, *Bioinformatics*, 31(20), 3282-3289

Phyldynamics

- ❖ *Phyldynamics* focuses on reconstructing past population dynamics from current genetic samples taken from the population of interest.
- ❖ Phyldynamic inference uses a *coalescent* model that defines a probability density for the genealogy of randomly sampled individuals.



Likelihood

- The coalescent times, given *effective population size*, $N_e(t)$, have density

$$P[\mathbf{t} \mid \mathbf{s}, \mathbf{n}, N_e(t)] = \prod_{k=2}^n \frac{A_{0,k}}{N_e(t_{k-1})} \exp \left\{ - \int_{I_{0,k}} \frac{A_{0,k}}{N_e(t)} dt - \sum_{i=1}^m \int_{I_{i,k}} \frac{A_{i,k}}{N_e(t)} dt \right\}$$

where $A_{i,k} = \binom{\ell_{i,k}}{2}$ depends on the number of lineages $\ell_{i,k}$ in the interval $I_{i,k}$ defined by coalescent times and sampling times.

Prior

- ✿ We assumed log-Gaussian Process prior on effective population size, $N_e(t)$,

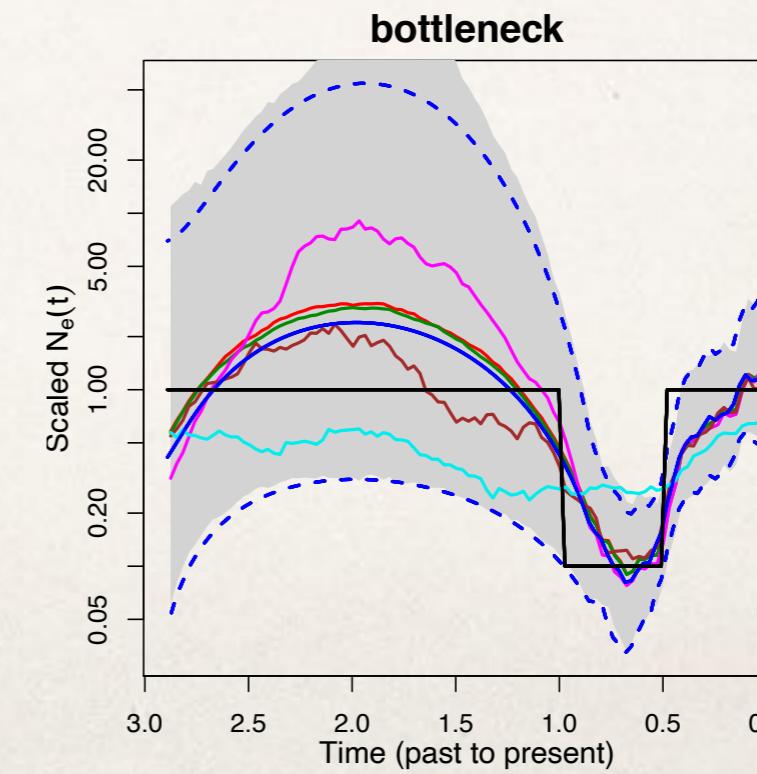
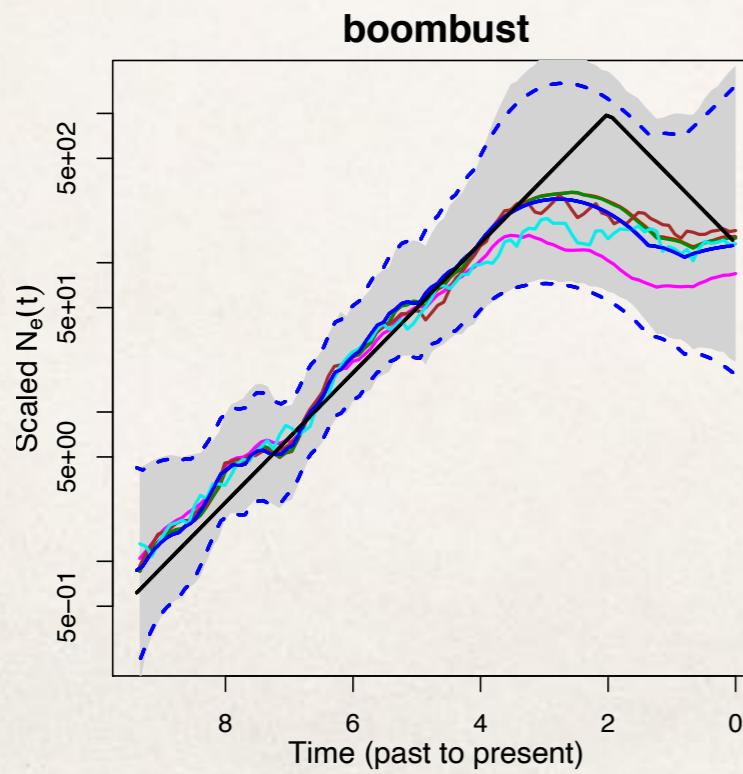
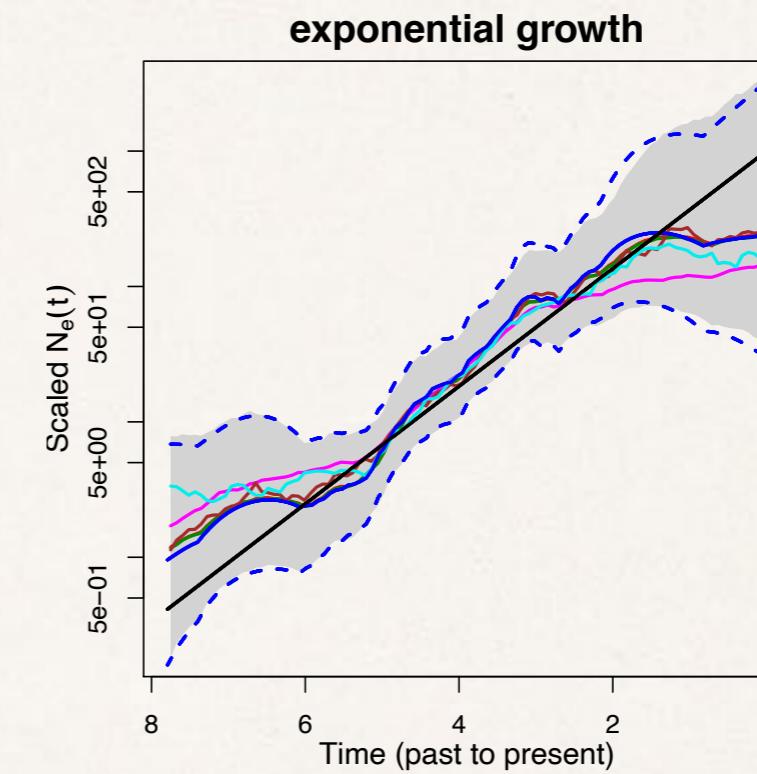
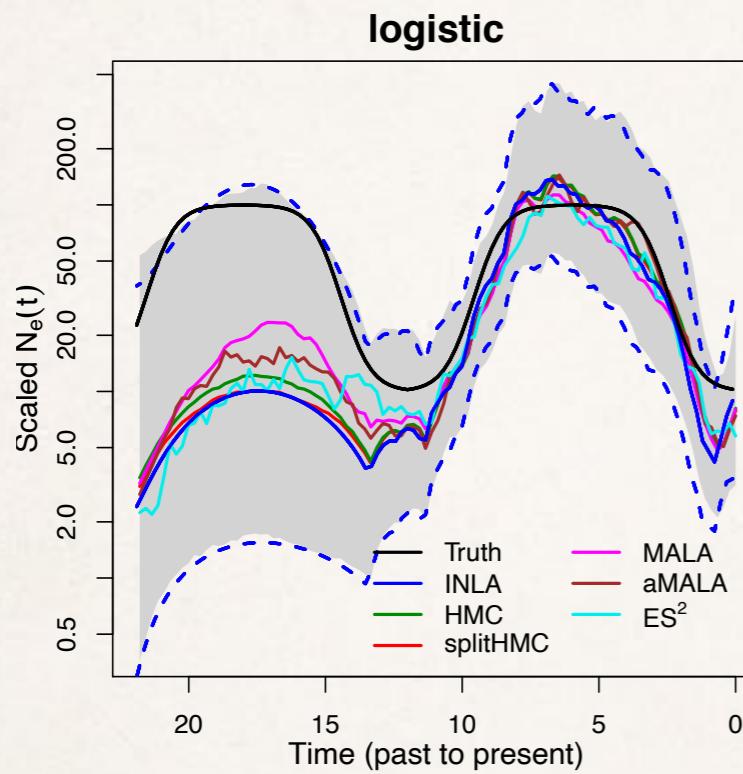
$$N_e(t) = \exp[f(t)], \quad f(t) \sim GP(\mathbf{0}, \mathbf{C})$$

Inference

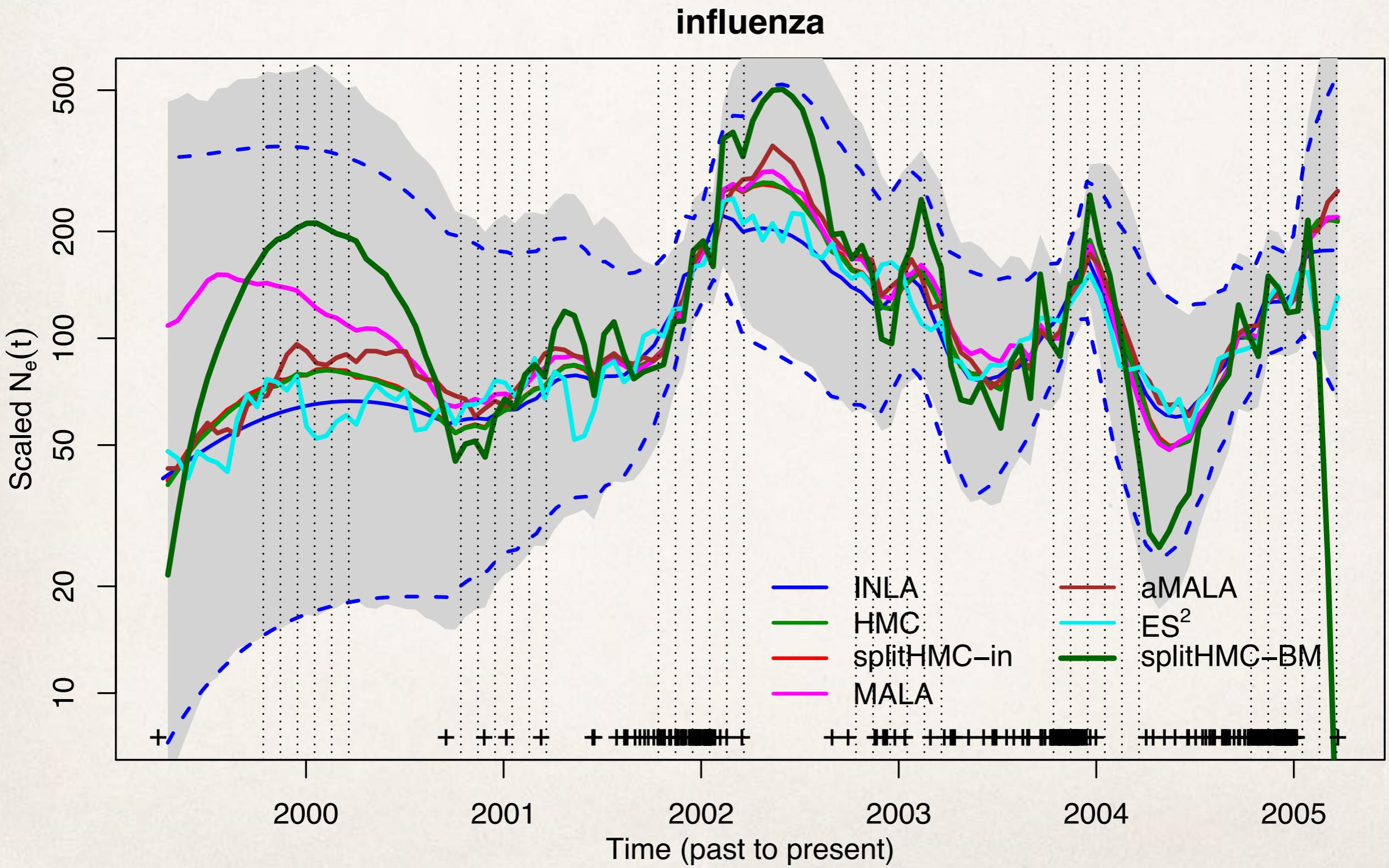
- ✿ We have developed a sampling algorithm, called split HMC, and compared to several existing methods:
 - ✿ Integrated Nested Laplace Approximation (INLA; Rue et al., 2009; Palacios and Minin, 2012)
 - ✿ Metropolis-Adjusted Langevin algorithm (MALA; Roberts and Tweedie, 1996)
 - ✿ Adaptive MALA (aMALA; Knorr-Held and Rue, 2002)
 - ✿ Hamiltonian Monte Carlo (HMC; Duane et al., 1987; Neal, 2010)
 - ✿ Elliptical Slice Sampler (ES²; Murray et al., 2010)

\end{itemize}

Simulations



Human Influenza A in New York



Neuroscience

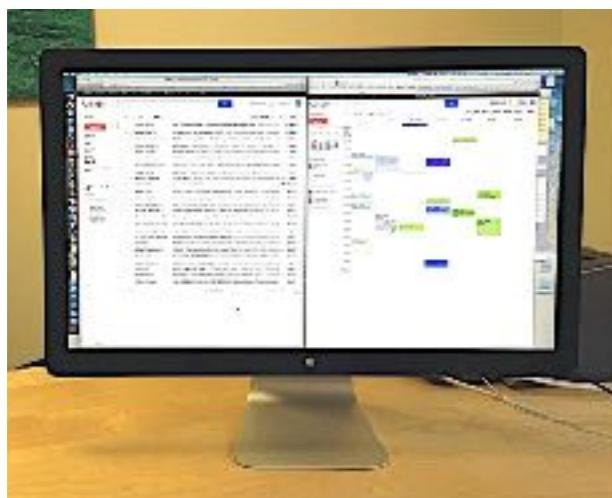
In Collaboration with Norbert Fortin's Lab

1. Zhou, B., Moorman, D. E., Behseta, S., Ombao, H., and Shahbaba, B. (2016), A Dynamic Bayesian Model for Characterizing Cross-Neuronal Interactions During Decision Making, *Journal of the American Statistical Association*, 111 (514), 459-471.
2. Lan, S., Holbrook, A., Elias, G.A., Fortin, N.J., Ombao, H., and Shahbaba, B. (2019+), Flexible Bayesian Dynamic Modeling of Correlation and Covariance Matrices, *Bayesian Analysis* (to appear).
3. Li, L., Pluta, D., Shahbaba, B., Fortin, N., Ombao, H., Baldi, P. (2019), Modeling Dynamic Functional Connectivity with Latent Factor Gaussian Processes, *NeurIPS 2019, Vancouver*.

Temporal Organization of Memories

- ❖ The memory for specific events is intrinsically tied to the **context** in which they occur, especially the spatial and temporal context

For example...



Event “A”

Catching up with emails



Event “B”

Getting some coffee

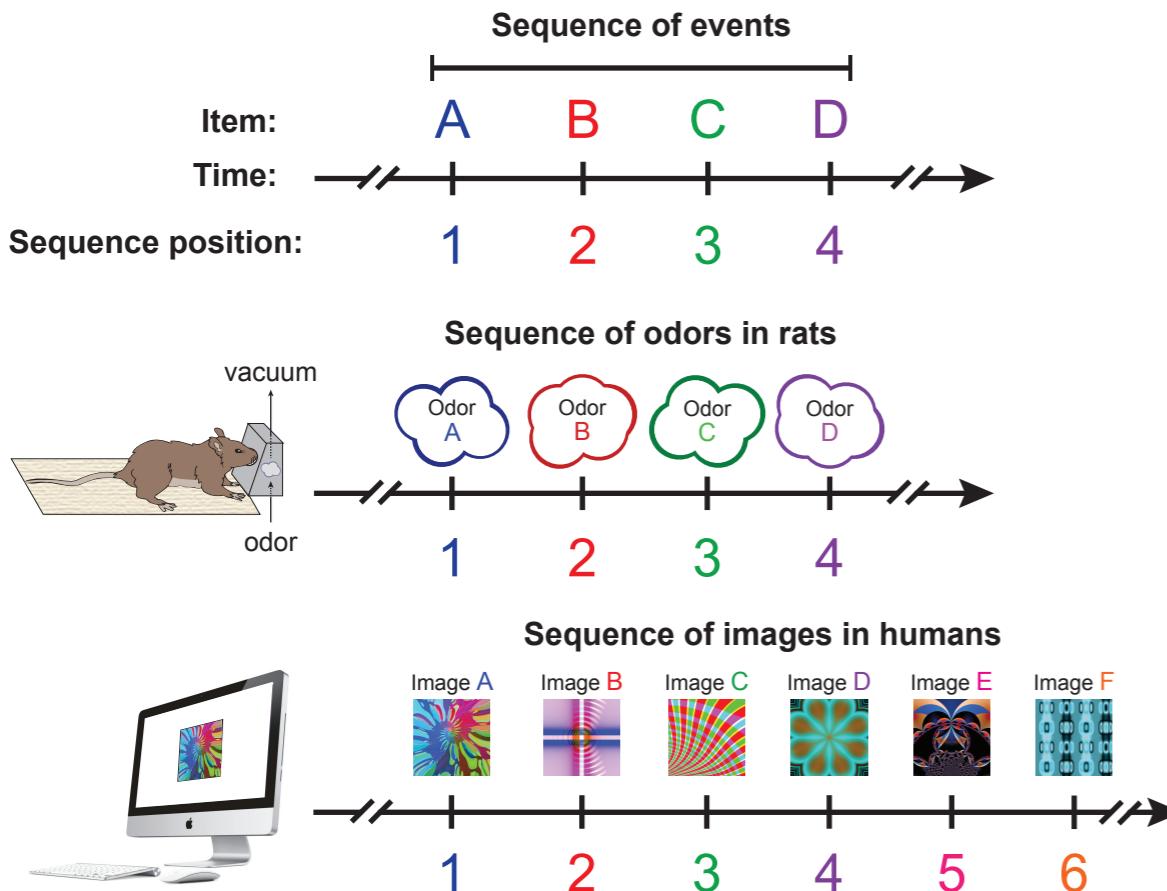


Event “C”

Going to a meeting

Temporal Organization of Memories

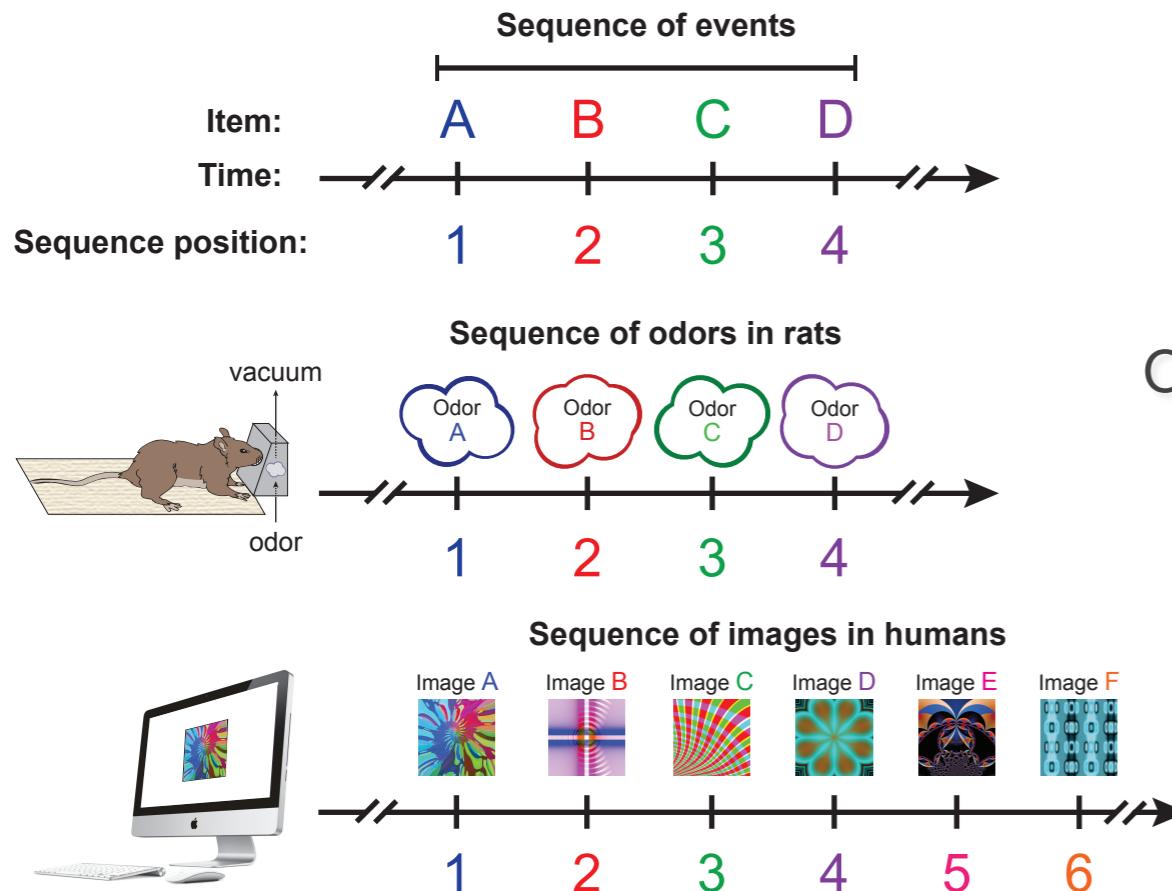
Task design



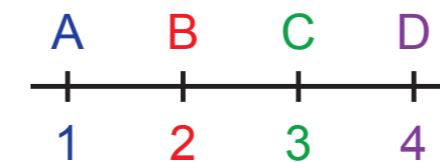
Each sequence is repeated many times during a session

Temporal Organization of Memories

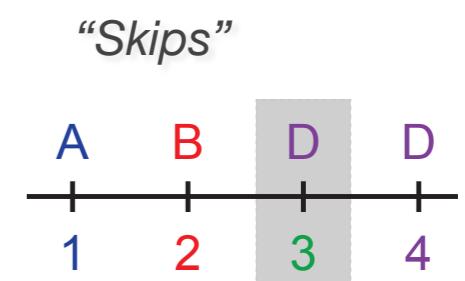
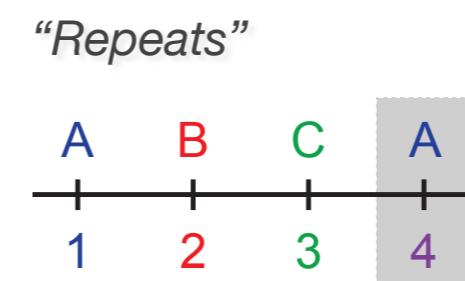
Task design



Most of the time, all items are “In Sequence” (**InSeq**)



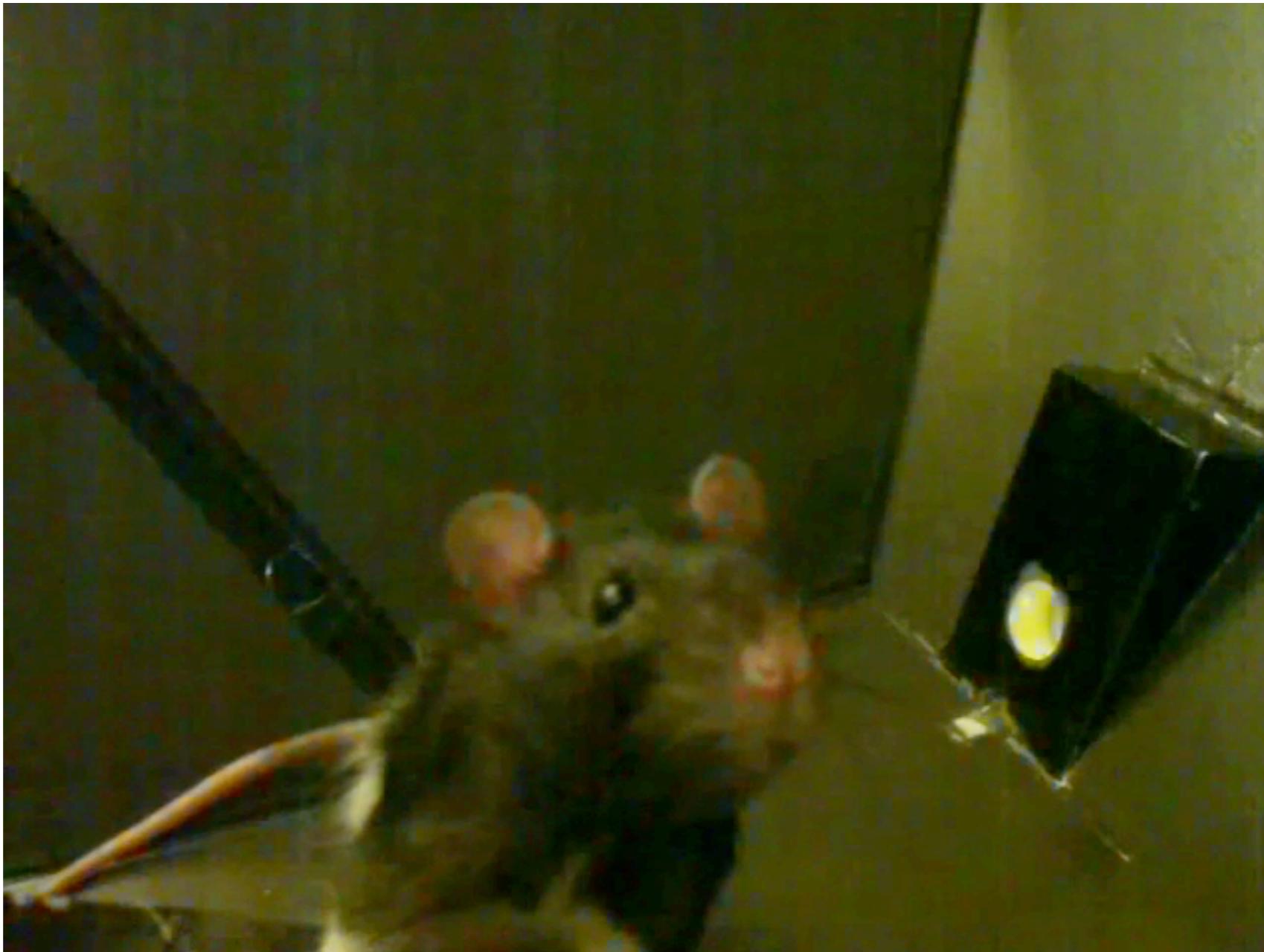
Occasionally, one item is “Out of Sequence” (**OutSeq**)



Each sequence is repeated many times during a session

Temporal Organization of Memories

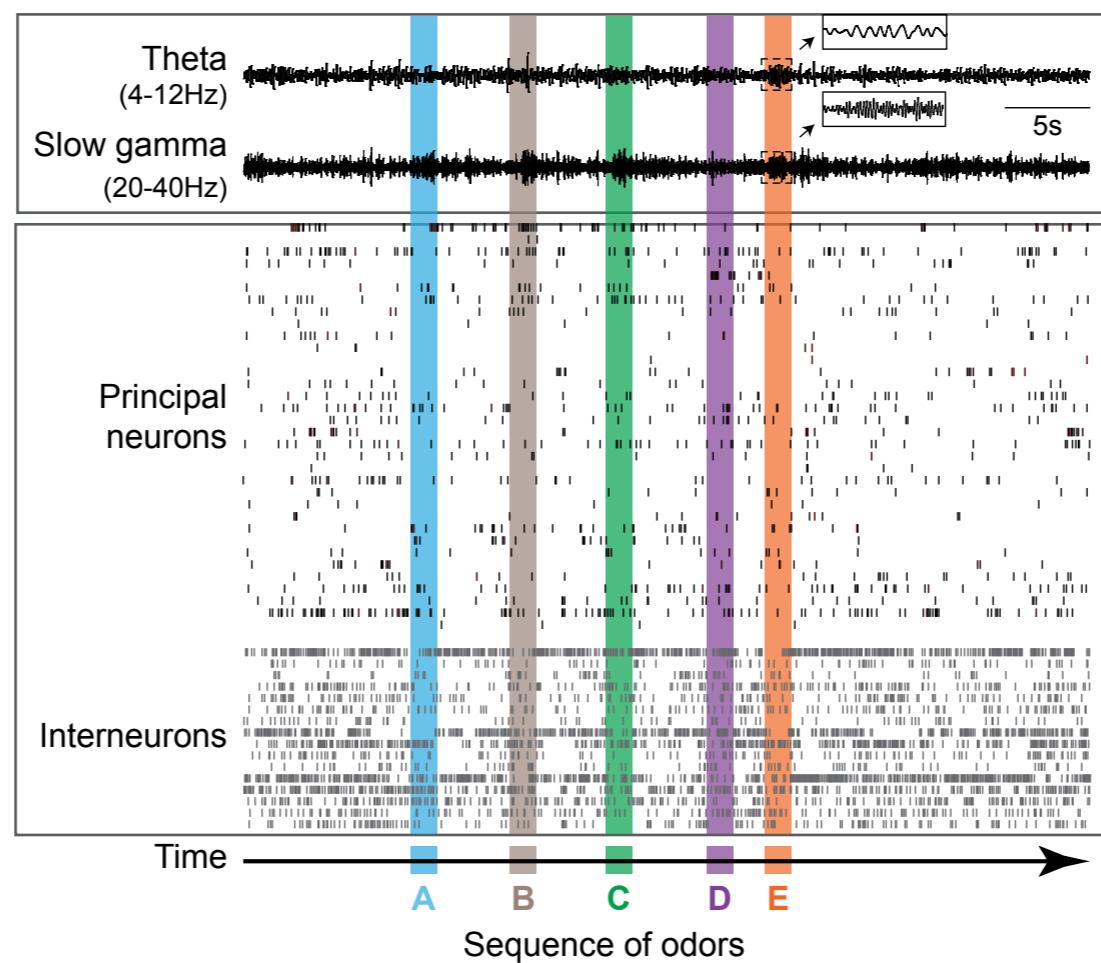
Sequence 1: A → B → C → D



Data

- From such experiments, we collect some continuous time series data, called Local Field Potential (LFP), and some discrete time series data, called Spike Trains.

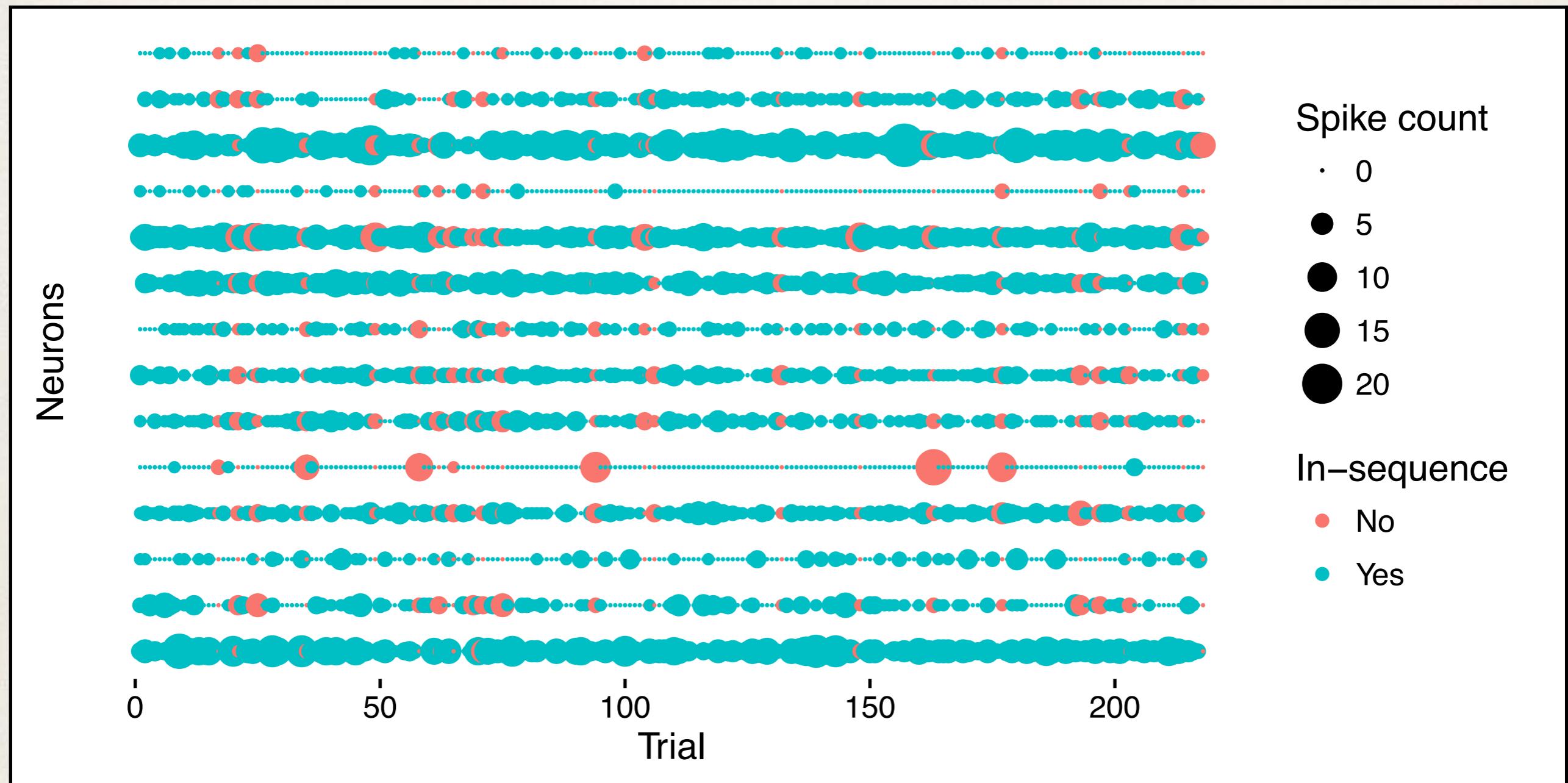
Representative recording during
one sequence presentation



Spike Trains

Discrete Time Series

Data



Modeling Time-varying Firing Rates

- For each neuron, we assume that the firing rate depends on an underlying latent variable, $u(t)$, which has a Gaussian process prior:

$$u(t) \sim \text{GP}(0, C)$$

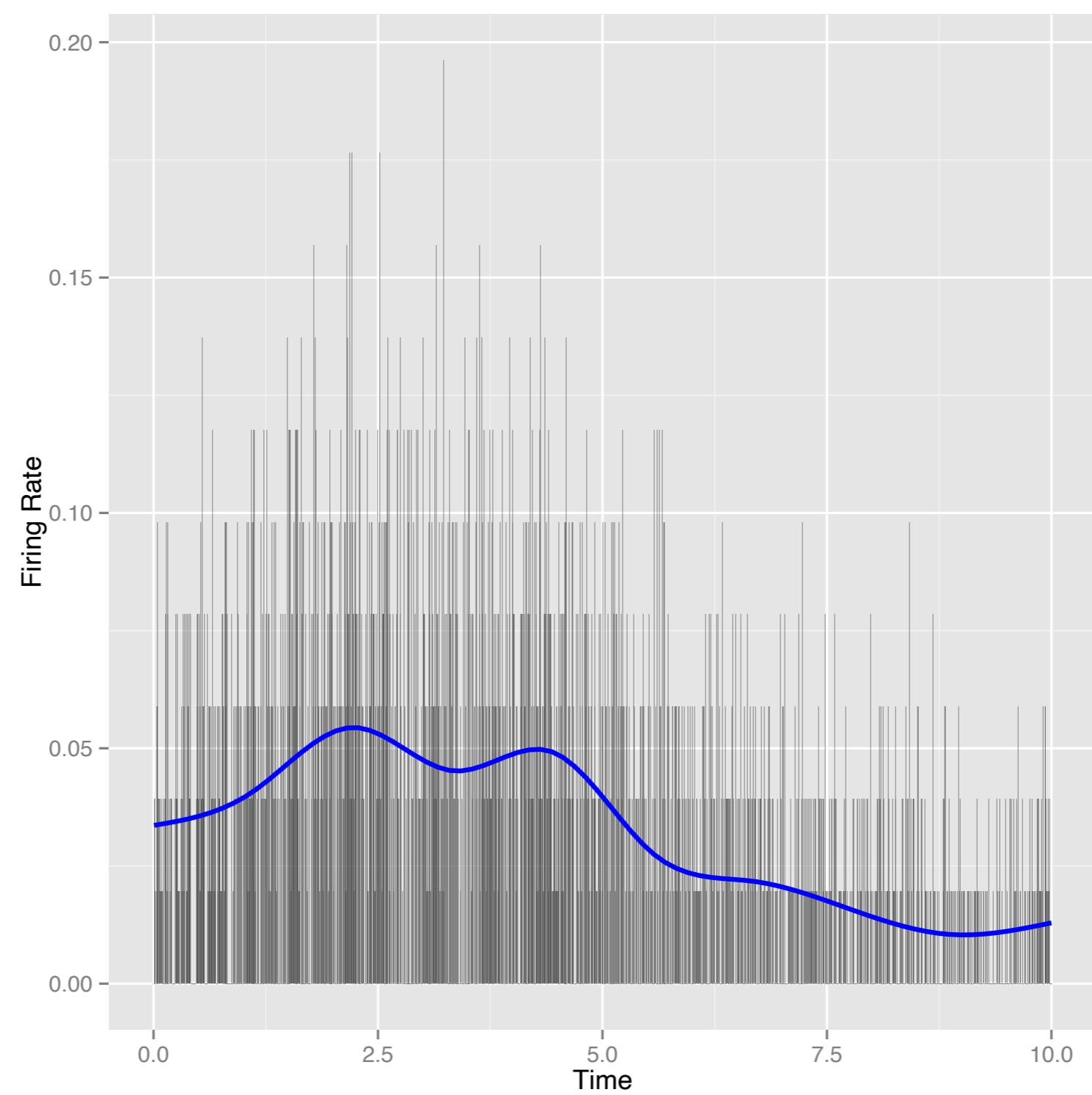
$$C_{ij} = \text{Cov}[u(t_i), u(t_j)]$$

$$C_{ij} = \kappa^2 \exp[-\lambda(t_i - t_j)^2] + \delta_{ij}\xi^2$$

- We specify the spike probability, p_t , within time interval t in terms of $u(t)$ through the following transformation:

$$p_t = \frac{1}{1 + \exp[-u(t)]}$$

Modeling Time-varying Firing Rates



Model for Detecting Synchrony

- For a given time bin, t , we model the joint firing probability of multiple spike trains,

$$\Pr_r(Y_{1t} = y_{1t}, \dots, Y_{nt} = y_{nt})$$

as a function of the marginal probabilities.

- Note that the joint probability has the following simplex constraint:

$$\sum_{(y_{1t}, \dots, y_{nt}) \in (0,1)^n} \Pr_r(Y_{1t} = y_{1t}, \dots, Y_{nt} = y_{nt}) = 1, \text{ for each } t = 1, 2, \dots, T$$

- To preserve the constraint, we use the following model:

$$Y_{it} = 1_{(-\infty, \tau_{it}]}(u_{it}) = \begin{cases} 1, & \text{if } u_{it} \leq \tau_{it} \\ 0, & \text{otherwise.} \end{cases} \quad (u_{1t}, \dots, u_{nt})^T \sim N(0, \Sigma)$$

Model

- ✿ Our model guarantees the simplex constraint by mapping the probabilities in the 2^n quadrants to the 2^n joint probabilities of n neurons.
- ✿ We can also show that the spike trains preserve the dependence structure of the latent variables.
- ✿ Note that the joint probabilities of spike trains depend on both the latent variables and thresholds.
- ✿ While the latent variables specify the dependence structure as discussed above, the thresholds determine the marginal probabilities of firing for each neuron.

$$u_{\{i\}t} \sim N(0, \Sigma)$$

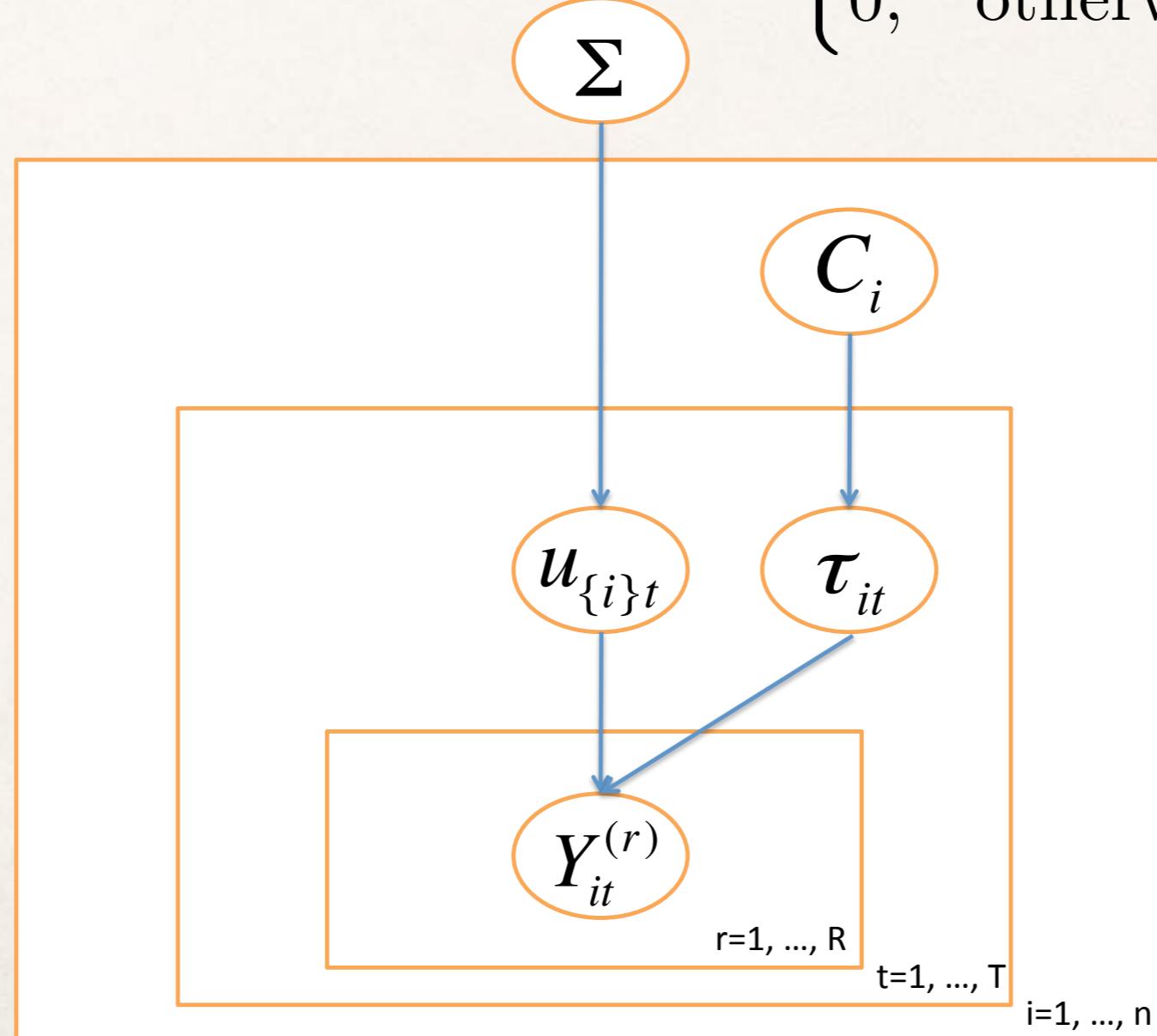
$$\tau_i(t) \sim GP(0, C_i)$$

Model

$$u_{\{i\}t} \sim N(0, \Sigma)$$

$$\tau_i(t) \sim GP(0, C_i)$$

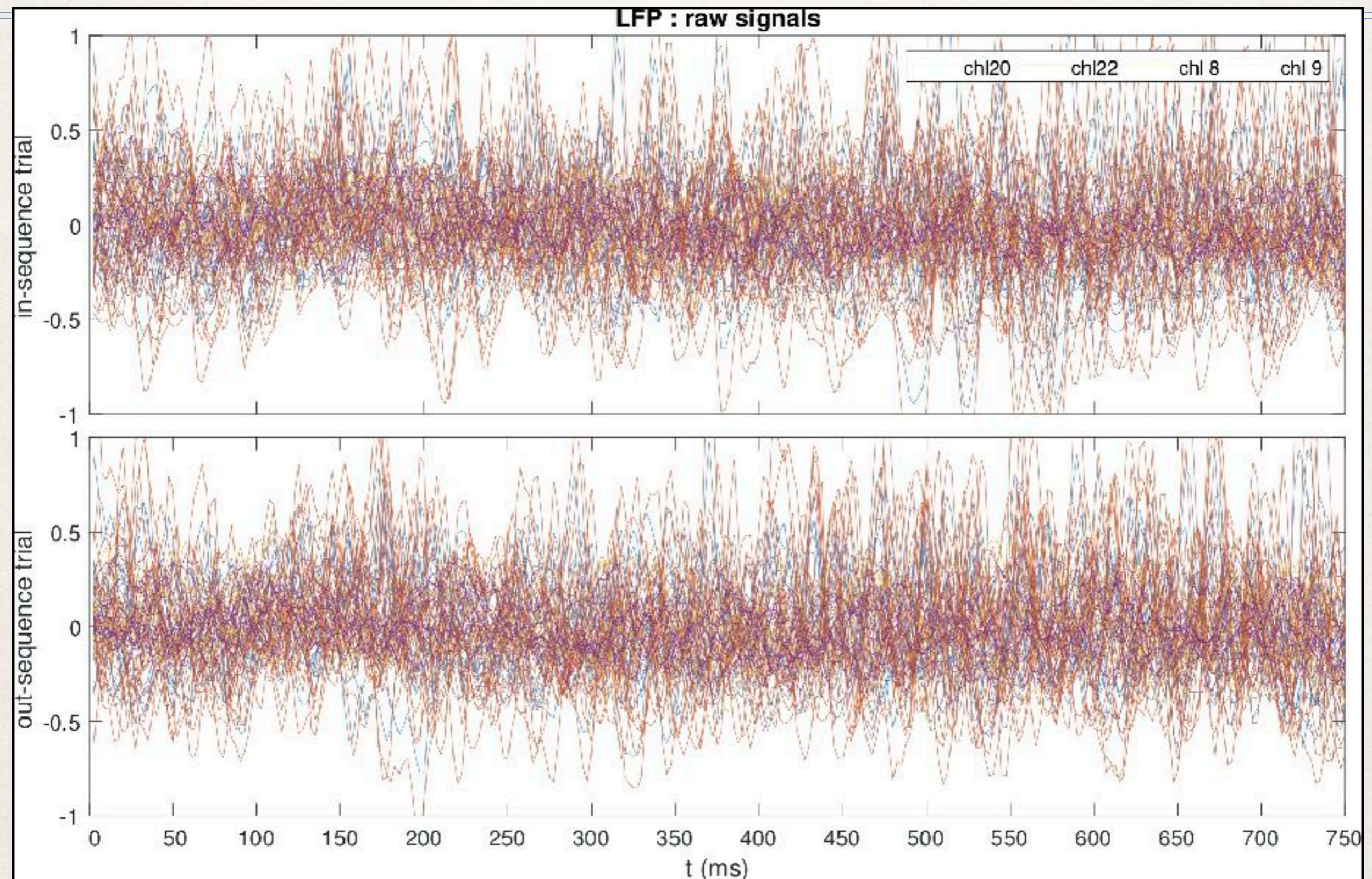
$$Y_{it} = 1_{(-\infty, \tau_{it}]}(u_{it}) = \begin{cases} 1, & \text{if } u_{it} \leq \tau_{it} \\ 0, & \text{otherwise.} \end{cases}$$



Local Field Potentials

Continuous Time Series

Data



Dynamic Functional Connectivity

- ✿ Dynamic functional connectivity is an active area of research in the neuroscience community
- ✿ Let $X_n(t)$ be a p -variate LFP time series for observations $n=1, \dots, N$.
- ✿ Let $K_n(t)$ be the $p \times p$ covariance matrix of $X_n(t)$ for observation n at time t .
- ✿ The questions of interest are
 - ✿ Does $K_n(t)$ varies significantly across t ?
 - ✿ Are $K_m(t)$ and $K_n(t)$ differ significantly when m and n are associated with different experimental conditions?

Alternative Models

- ✿ Previous modeling approaches include:
 - ✿ Sliding window (SW) methods, often with PCA (Lindquist et al, 2014, Leonardi et al., 2015)
 - ✿ Variants of the Hidden Markov Model (HMM, Cabral et al., 2017)
 - ✿ Latent factor models, such as Latent Factor Stochastic Volatility (LFSV, Lopes et al., 2004, Kastner et al, 2017)

Model

- ✿ For multiple time series, we have

$$X_n(t) \sim \mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t) \text{ where } \Sigma_t = [\sigma_{ij}(t)]_{D \times D} > 0$$

$$\mu_i(t) \sim \mathcal{GP}(0, \kappa_\mu(t, \theta)), \quad i = 1, \dots, D$$

$$\log \sigma_{ii}(t) \sim \mathcal{GP}(0, \kappa_\sigma(t, \theta))$$

- ✿ Spatial dependence is coded in Σ
- ✿ Temporal evolution is modeled by various Gaussian process (GP) models
- ✿ We need to ensure the positive-definiteness of the covariance matrix over time

Decomposition

- ❖ We can use Cholesky decomposition of the covariance matrix:

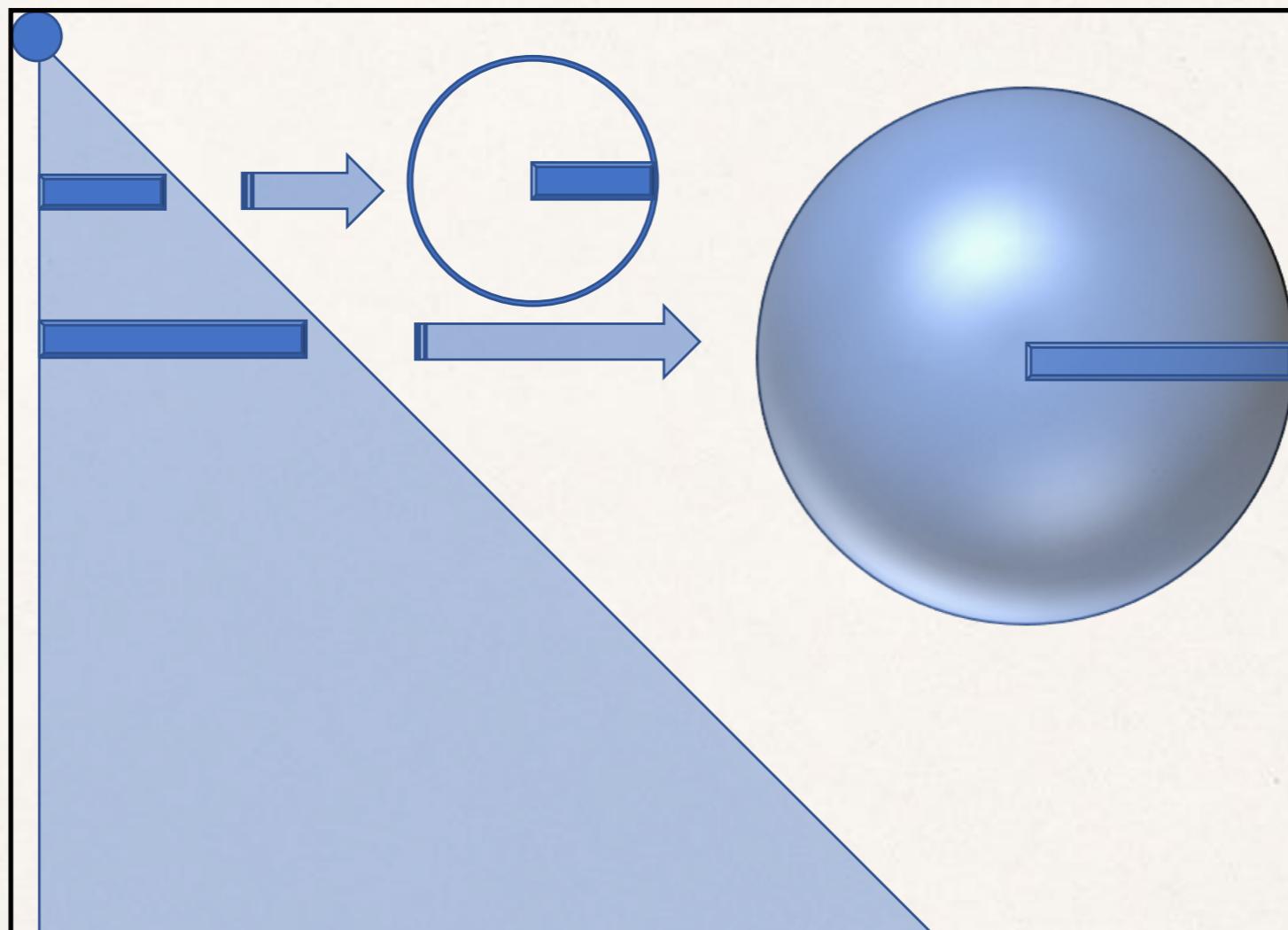
$$\Sigma = \mathbf{L}\mathbf{L}^T, \quad \sigma_{ij} = \sum_{k=1}^{\min\{i,j\}} l_{ik}l_{jk}, \quad \mathbf{L} = \begin{bmatrix} * \\ ** \\ * * * \end{bmatrix}$$

$$\sigma_i^2 := \sigma_{ii} = \sum_{k=1}^i l_{ik}^2 = \|\mathbf{l}_i\|^2, \quad \mathbf{L} = \begin{bmatrix} \mathbf{l}_1 \\ \vdots \\ \mathbf{l}_D \end{bmatrix}$$

- ❖ Or work on the correlation matrix instead:

$$\mathbf{P} := \text{diag}(\Sigma)^{-\frac{1}{2}} \Sigma \text{diag}(\Sigma)^{-\frac{1}{2}} = \mathbf{L}^*(\mathbf{L}^*)^T, \quad \rho_{ij} = \sum_{k=1}^{\min\{i,j\}} l_{ik}^*l_{jk}^*$$

Geometric View



$$(\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_D) \in \mathcal{S}_0^0(\sigma_1) \times \mathcal{S}_0^1(\sigma_2) \cdots \times \mathcal{S}_0^{D-1}(\sigma_D)$$

$$(\mathbf{l}_1^*, \mathbf{l}_2^*, \dots, \mathbf{l}_D^*) \in \mathcal{S}_0^0 \times \mathcal{S}_0^1 \cdots \times \mathcal{S}_0^{D-1}$$

Inverse-Wishart

- We can show that the conjugate Inverse-Wishart prior can be presented in this framework
- However, our proposed approach allows for specifying more flexible priors
- If two variables y_i and y_j known to be uncorrelated a priori, then we can choose a prior that encourages l_i and l_j to be perpendicular
- For example, we can specify priors $p(l_i)$ that concentrate on the poles of S_0^{i-1} ,

$$p(l_i) \propto |l_{ii}|, \quad i = 2, \dots, D$$

- This leads to fewer non-zero off-diagonal elements, which in turn leads to a larger number of perpendicular variables

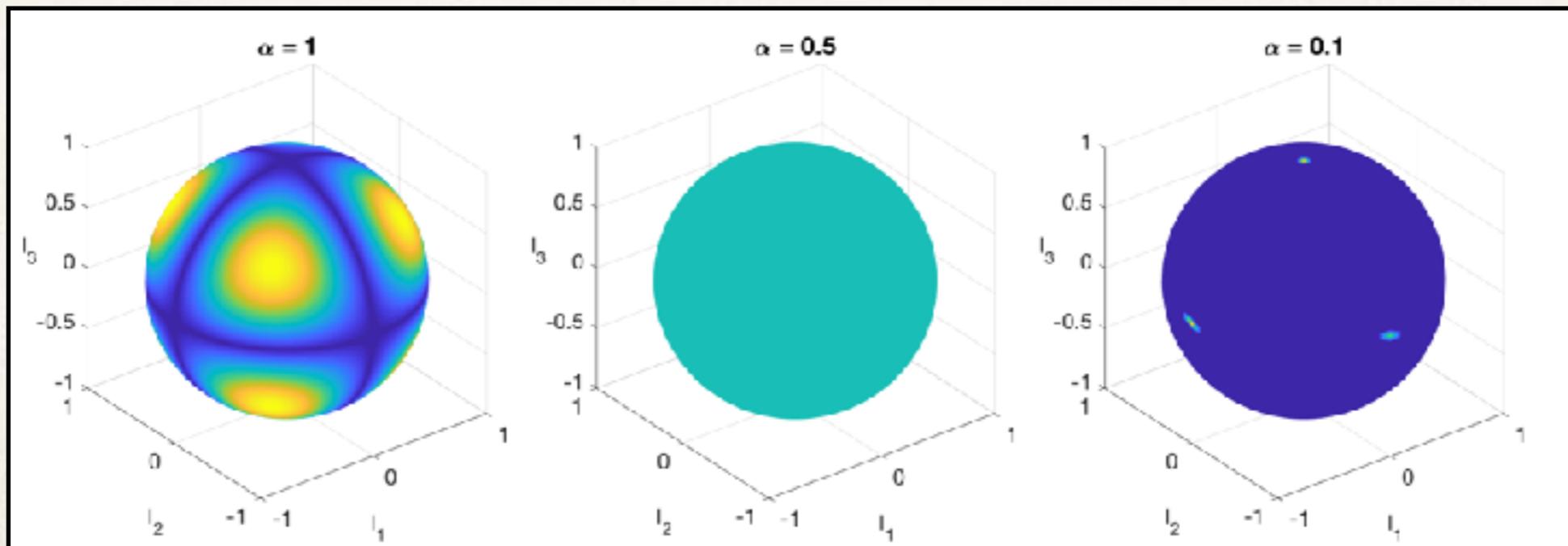
Squared-Dirichlet Distribution

- In general, we can map a probability distribution defined on the simplex onto the sphere
- Squared-Dirichlet distribution

$$\mathbf{l}_i^2 := (l_{i1}^2, l_{i2}^2, \dots, l_{ii}^2) \sim \text{Dir}(\boldsymbol{\alpha}_i)$$

$$\mathbf{l}_i \sim \text{Dir}^2(\boldsymbol{\alpha}_i)$$

$$p(\mathbf{l}_i) = p(\mathbf{l}_i^2) |2\mathbf{l}_i| \propto (\mathbf{l}_i^2)^{\boldsymbol{\alpha}_i - 1} |\mathbf{l}_i| = |\mathbf{l}_i|^{2\boldsymbol{\alpha}_i - 1} := \prod_{k=1}^i |l_{ik}|^{2\alpha_{ik} - 1}$$



Uniform Prior

- Setting $\alpha_i = (\frac{1}{2}, \dots, \frac{1}{2}, 1)$, we obtain the special case,

$$p(\mathbf{l}_i) \propto |l_{ii}|, \quad i = 2, \dots, D$$

- Setting $\alpha_i = (\frac{1}{2}\mathbf{1}_{i-1}^\top, \alpha_{ii})$, $\alpha_{ii} = \frac{(i-2)D - 1}{2}$ leads to a marginally uniform prior

$$\rho_{ij} \sim \text{Unif}(-1, 1), \quad i \neq j$$

$$\alpha_i = (\frac{1}{2}\mathbf{1}_{i-1}^\top, \alpha_{ii}), \quad \alpha_{ii} = \frac{D-i}{2} + 1$$

- Setting $p(\mathbf{P}) \propto 1$ leads to a joint uniform prior

Unit-vector Gaussian Distribution

- ❖ Another natural spherical prior can be obtained by constraining a multivariate Gaussian random vector to have unit norm.
- ❖ Unit-vector Gaussian distribution:

$$p(\mathbf{l}_i \mid \|\mathbf{l}_i\|_2 = 1) = \frac{1}{(2\pi)^{\frac{i}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{l}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{l}_i - \boldsymbol{\mu}) \right\}, \quad \|\mathbf{l}_i\|_2 = 1$$

- ❖ Setting $\Sigma = I$, we obtain the von Mises-Fisher distribution as a special case.

Dynamic Model

- To model the covariance (or correlation) matrix dynamically, we use a time-varying Cholesky matrix, L_t
- Since each row of L_t has to be on a sphere of certain dimension, we consider a multivariate process, called unit-vector process (uvP), satisfying the unit-norm requirement

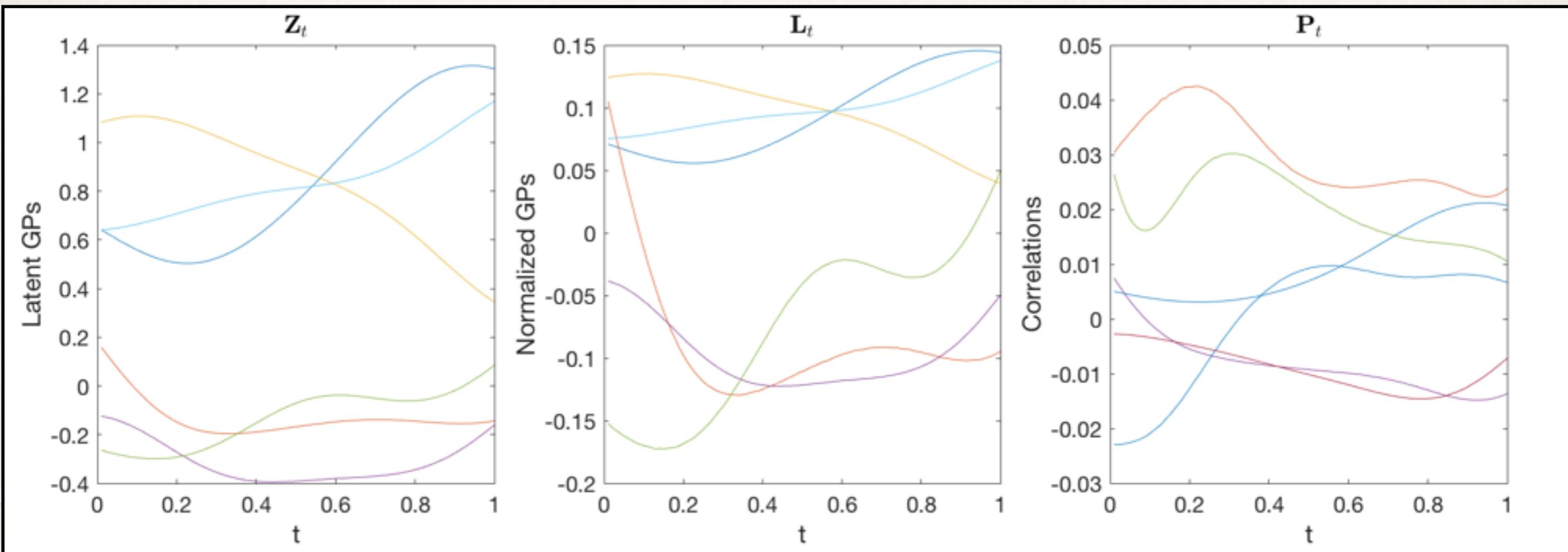
$$\|\mathbf{l}_i(t)\| \equiv 1, \quad \forall t \in \mathcal{T}$$

- More specifically, we are using A D-dimensional vector Gaussian process

$$\mathbf{Z}(x) \sim \mathcal{GP}_D(\boldsymbol{\mu}, \mathcal{C}, \mathbf{V}_{D \times D})$$

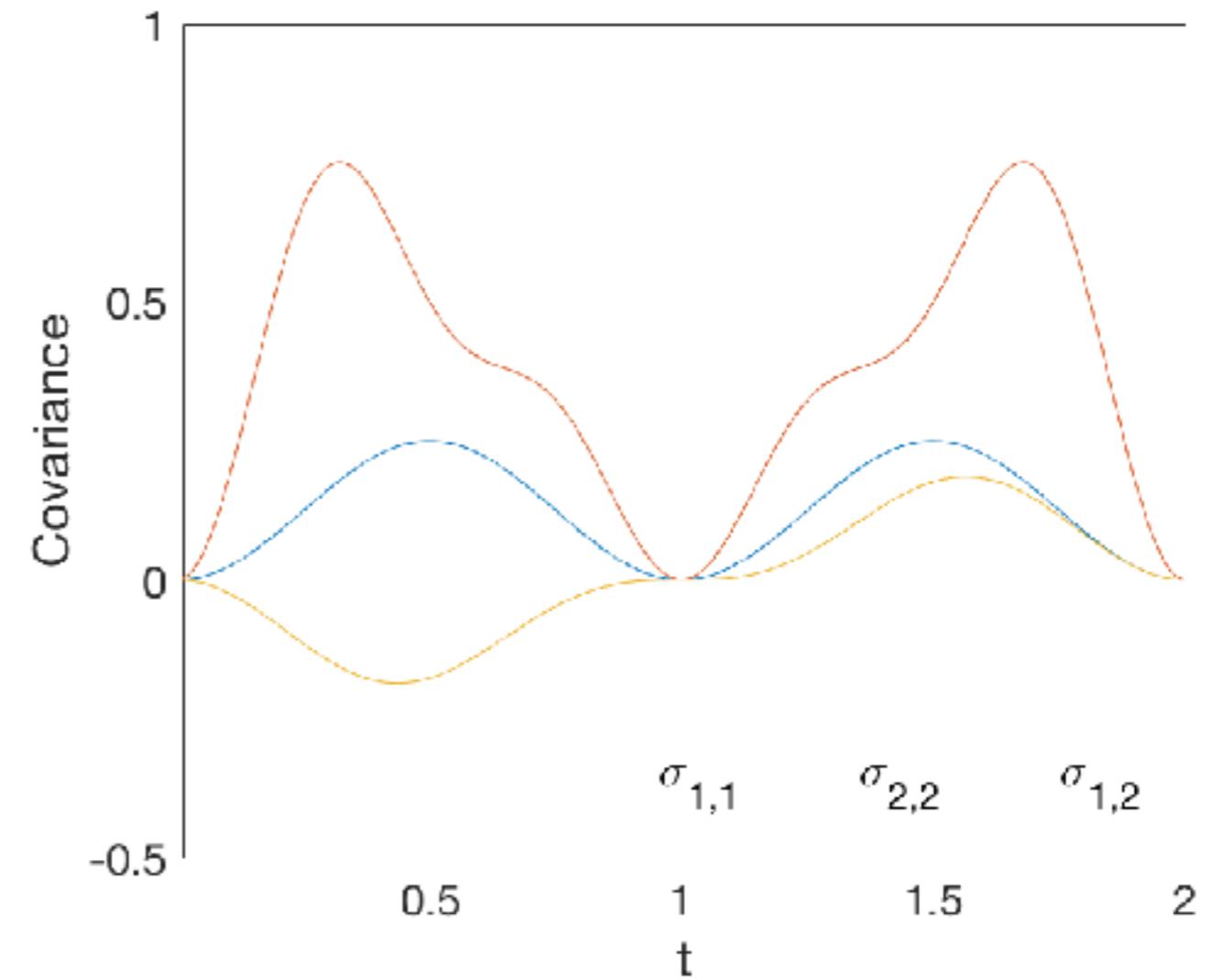
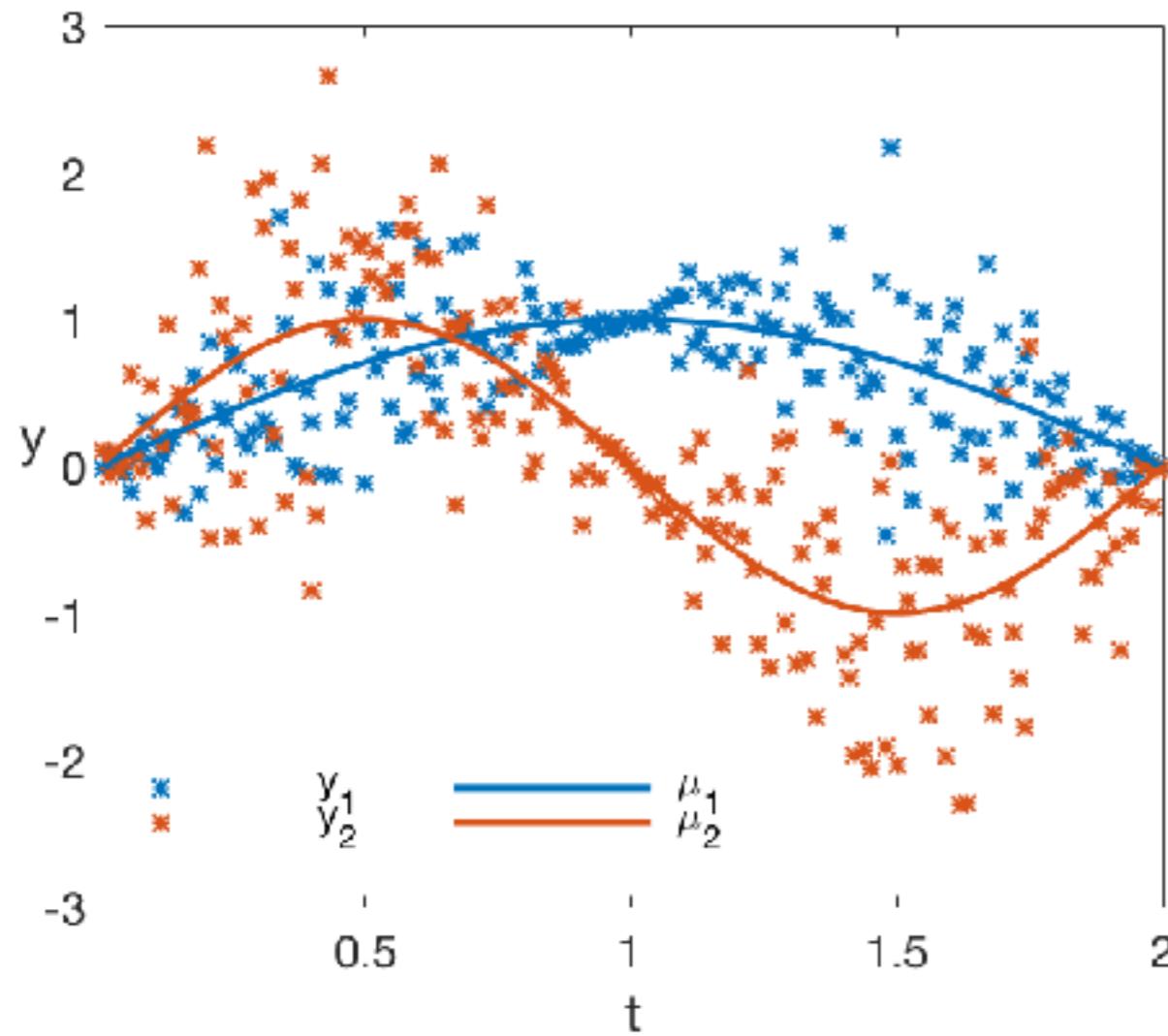
which is constrained to unit-sphere to obtain a unit-vector Gaussian process (uvGP)

Example



A realization of vector GP (left), its normalization (forming rows of) L_t (middle) and the induced correlation process.

Simulation

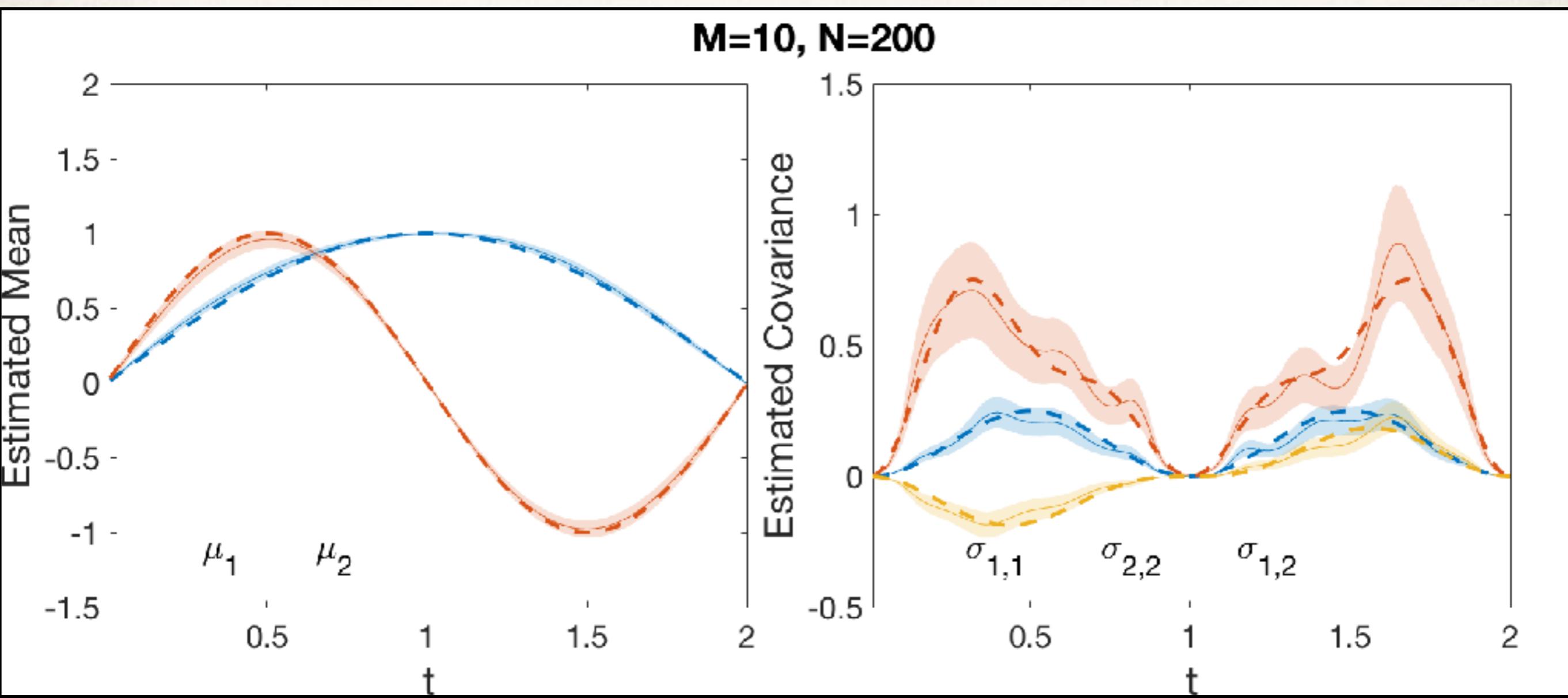


$$y(t) \sim \mathcal{N}_D(\mu(t), \Sigma(t)), \quad \Sigma(t) = L(t)L(t)^\top \circ S, \quad t \in [0, 2]$$

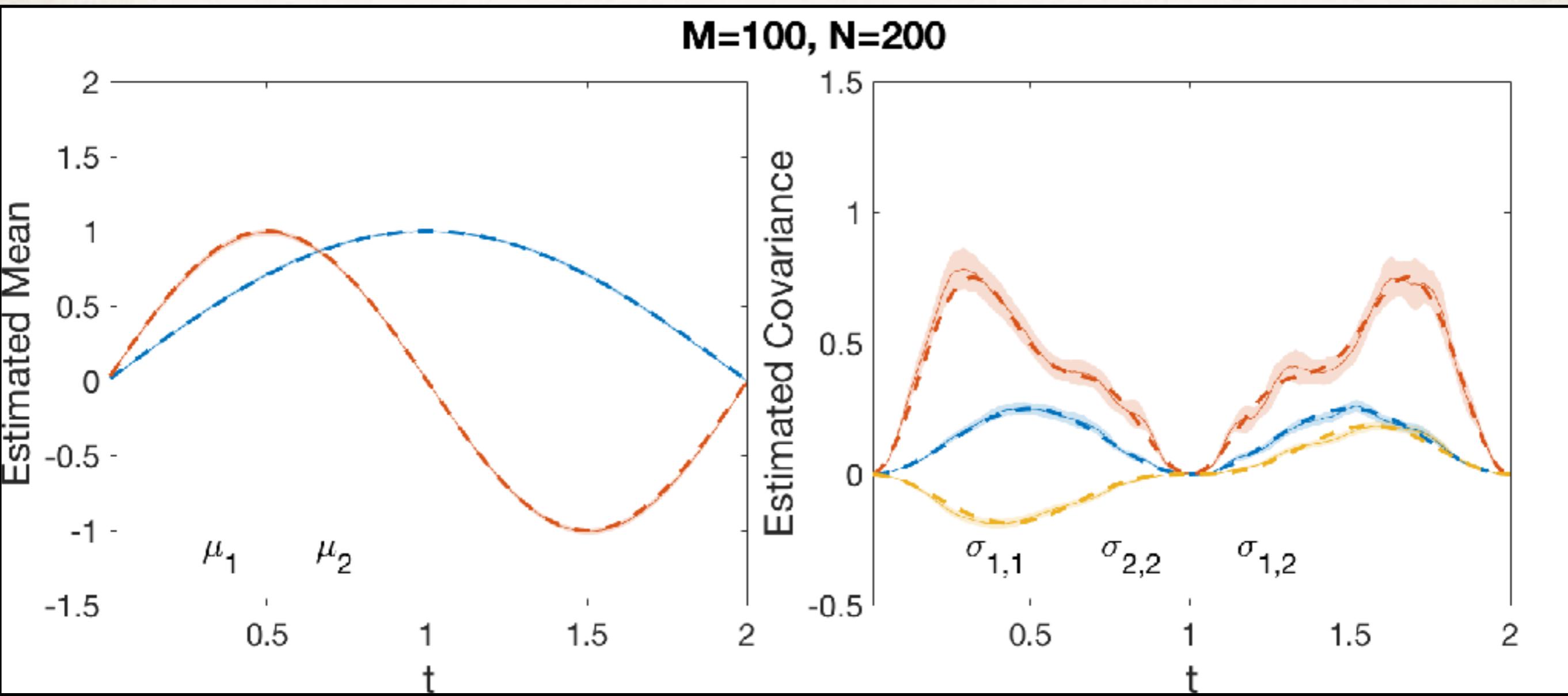
$$\mu_i(t) = \sin(it\pi/D), \quad L_{ij}(t) = (-1)^i \sin(it\pi/D)(-1)^j \cos(jt\pi/D), \quad j \leq i = 1, \dots$$

$$S_{ij} = (|i - j| + 1)^{-1}, \quad i, j = 1, \dots, D$$

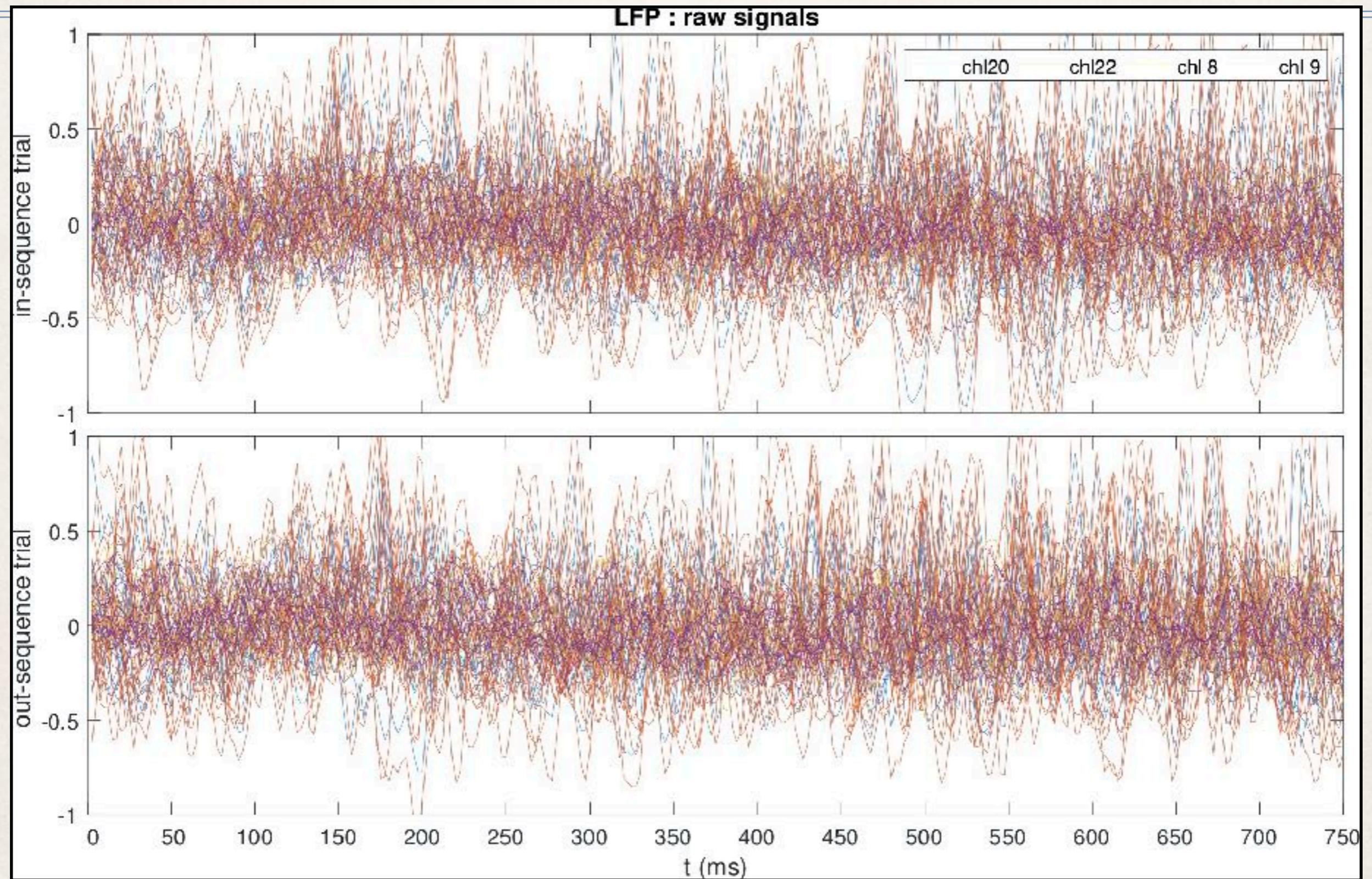
Simulation



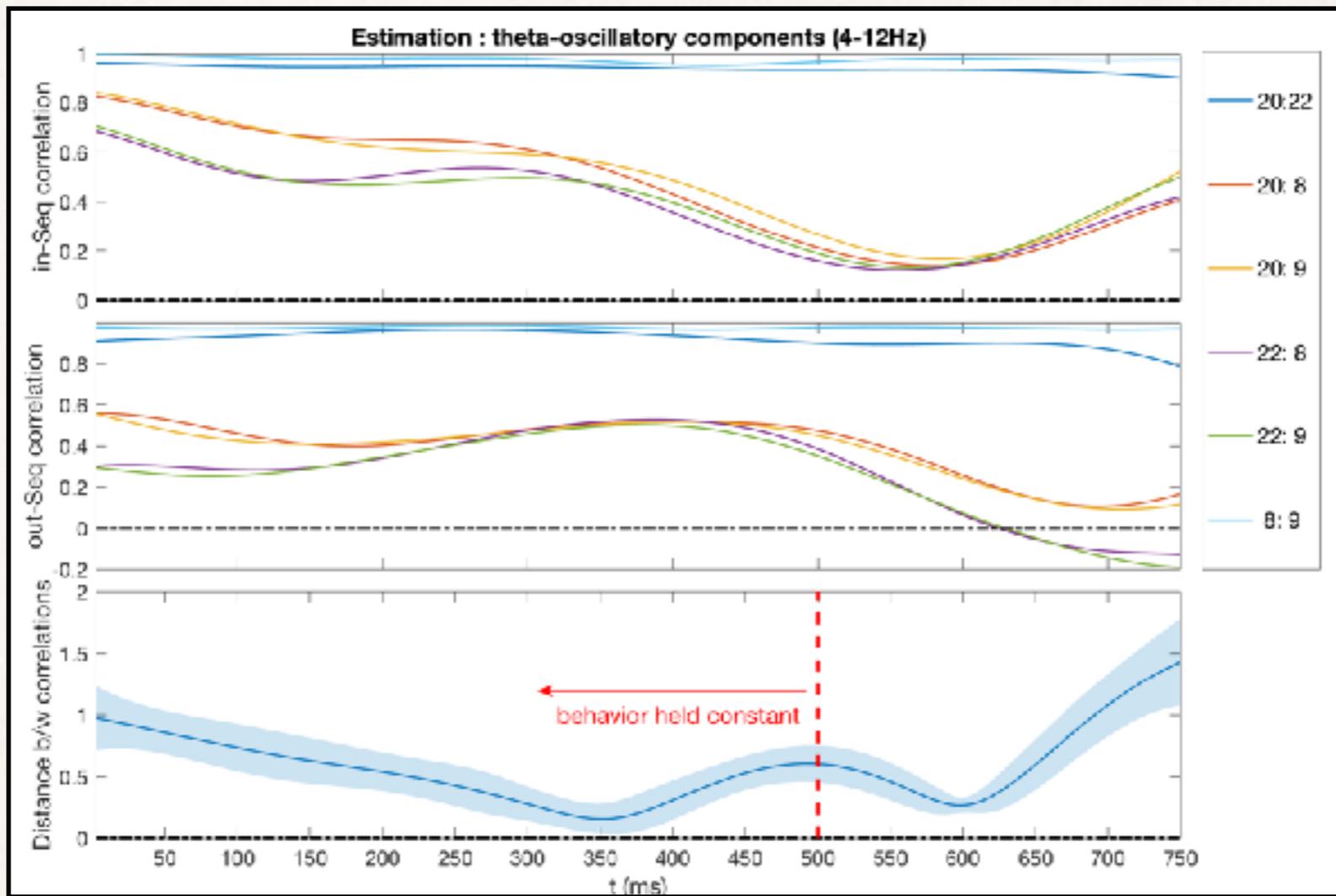
Simulation



Results for the Rat Experiment

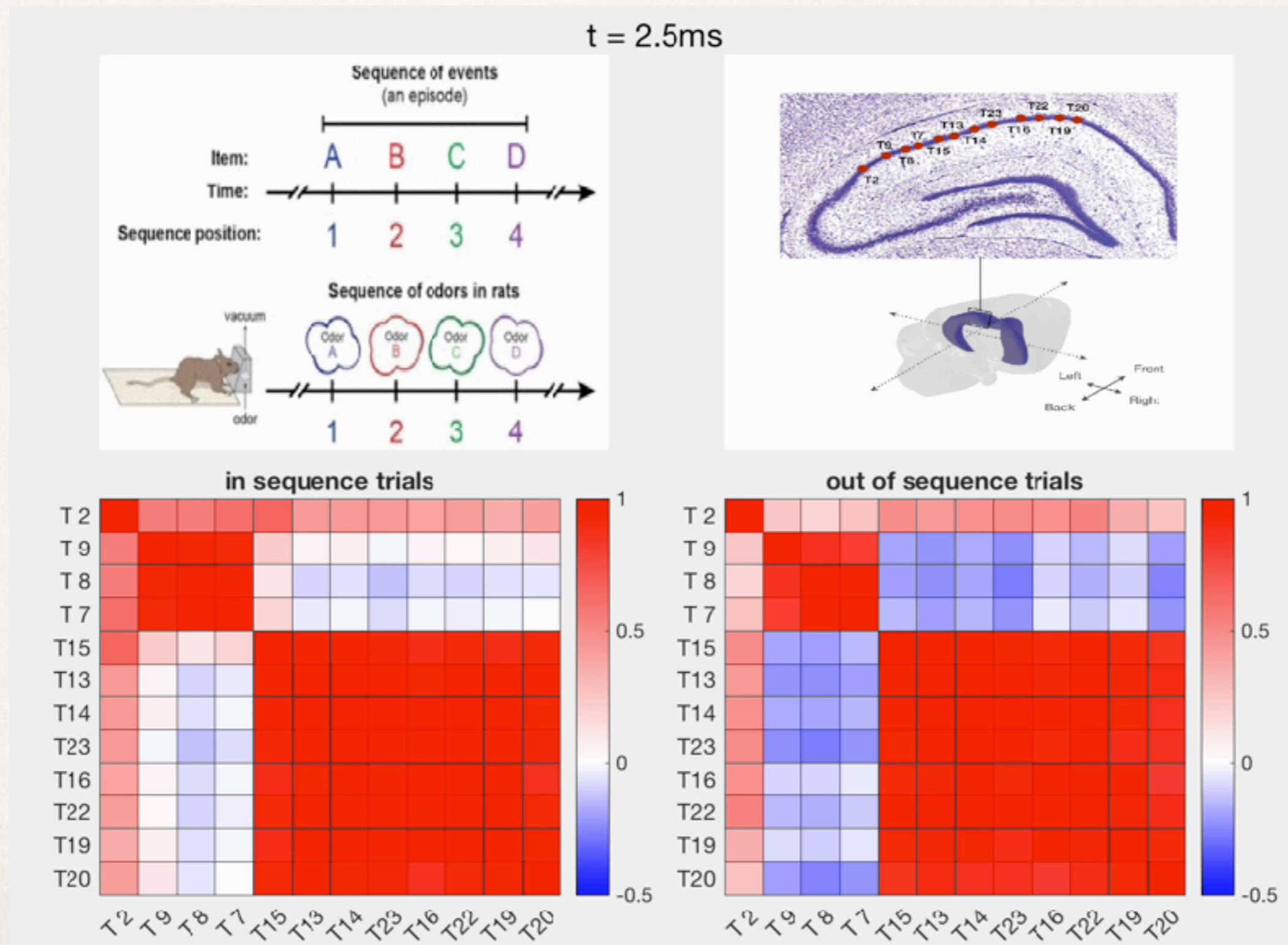


4 Tetrodes



- Nearby electrodes (20:22 and 8:9) displayed remarkably higher correlations in LFP compared to distant pairs (20:8, 20:9, 22:8, and 22:9).
- InSeq and OutSeq activity was very similar at the beginning (e.g., before 350ms) but maximally different at the end.

All Tetrodes

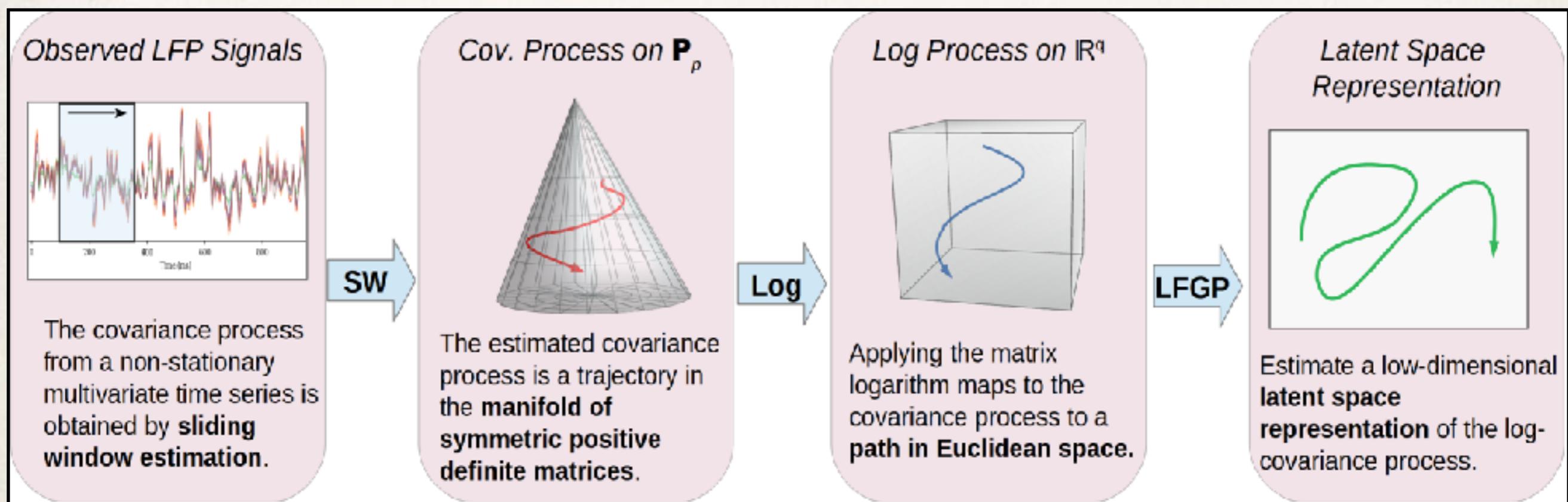


Latent Factor Gaussian Process Model

- ❖ The method discussed above does not scale well for high dimensional data
- ❖ Alternatively, we could use a latent factor model, but instead of placing a factor structure on the observed time series (as in the LFSV model), we can represent the covariance process itself as a linear combination of factors:
 1. Estimate the covariance process via sliding windows, which gives a time series of symmetric positive definite matrices.
 2. Apply the matrix logarithm to the estimated covariance process, giving a time series of real symmetric matrices.
 3. Fit a multivariate time series factor model to the vectorized upper triangle of the log-covariance process.

Latent Factor Gaussian Process Model

Latent Factor Gaussian Process Model



Li et. al. (2019), *NeurIPS*

Latent Factor Gaussian Process Model

$$X_n(t) \sim \mathcal{D}(0, K_n(t)) \text{ where } K_n(t) = \exp(\overrightarrow{\mathbf{u}}^{-1}(Z_n(t)))$$

$$Z_n(t) = B \cdot F_n(t) + \epsilon_n \text{ where } \epsilon_n \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2 I)$$

$$F_n(t) \sim \mathcal{GP}(0, \kappa(t; \theta))$$

$$B \sim p_1, \sigma^2 \sim p_2, \theta \sim p_3,$$

Here, $Z_i(t)$ is the vectorized Log-covariance, $F_i(t)$ is the r -length factor vector, B is the $p(p+1)/2 \times r$ loading matrix, and p_1, p_2, p_3 are priors.

Latent Factor Gaussian Process Model

Sketch of Sampling Algorithm

To generate posterior draws from the model, we can apply a combination of Gibbs sampling and Metropolis or slice sampling.

Input: Vectorized log-covariance $Y_i(t), i = 1, \dots, n$.

Input: GP kernel κ and priors p_1, p_2, p_3 .

Output: Posterior samples of factors F and factor loadings B .

Initialize values for $B_0, \sigma_0^2, \theta_0$.

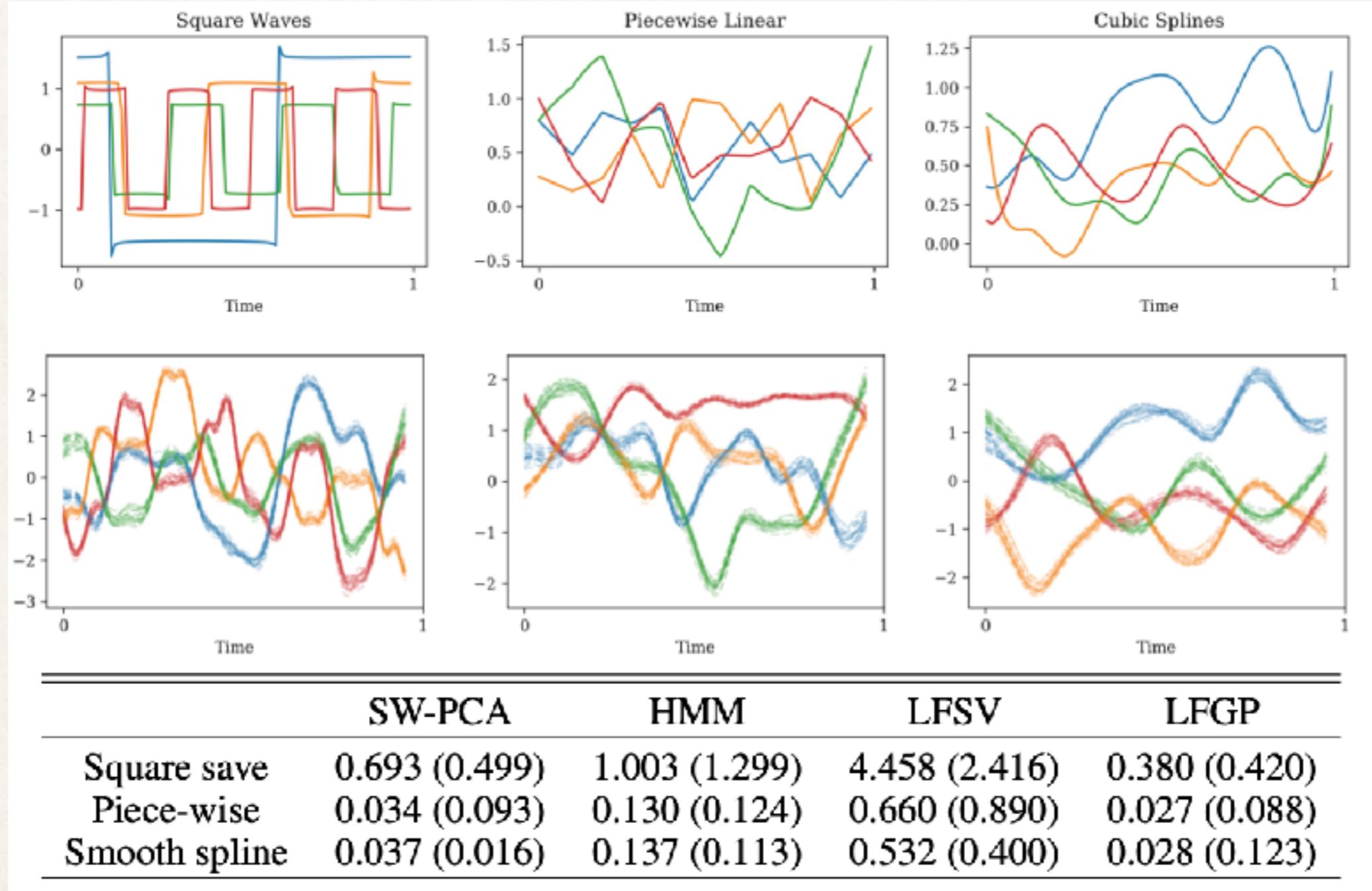
for $s = 1, \dots, S$ **do**

 | Sample $F_s | B_{s-1}, \sigma_{s-1}^2, \theta_{s-1}$ from $p(F | Y)$.

 | Sample $B_s, \sigma_s^2, \theta_s$ conditional on F_s

end

Simulation Results



The evaluation metric is reconstruction loss of the covariance as measured by the Log-Euclidean distance.

Results for the Rat Experiment

