Geometric fluid approximation for general continuous-time Markov chains

Michalis Michaelides
mic.michaelides@ed.ac.uk

Guido Sanguinetti Jane Hillston

School of Informatics, University of Edinburgh

April 2019

Continuous-time Markov chains

- $X(t)$ time-dependent random variable.
- At times t_n, observe jumps $X([t_n, t_{n+1})) = \xi_n$.
- $\xi_n \in I$, countable state-space.
- For transition rate matrix Q, elements q_{ij} and $q_i = -\sum_j q_{ij}$,

$$p(\xi_n \mid \xi_{n-1}, \cdots \xi_0) = p(\xi_n \mid \xi_{n-1});$$
$$p(\xi_n = j \mid \xi_{n-1} = i) = q_{ij}/-q_i;$$
$$t_n - t_{n-1} \sim \exp(-q_i).$$

Chapman-Kolmogorov

Would like $P_{ij}(s;t) = P(X(t) = j \mid X(s) = i)$ where $t > s$ for all states (matrix $P(s;t)$).

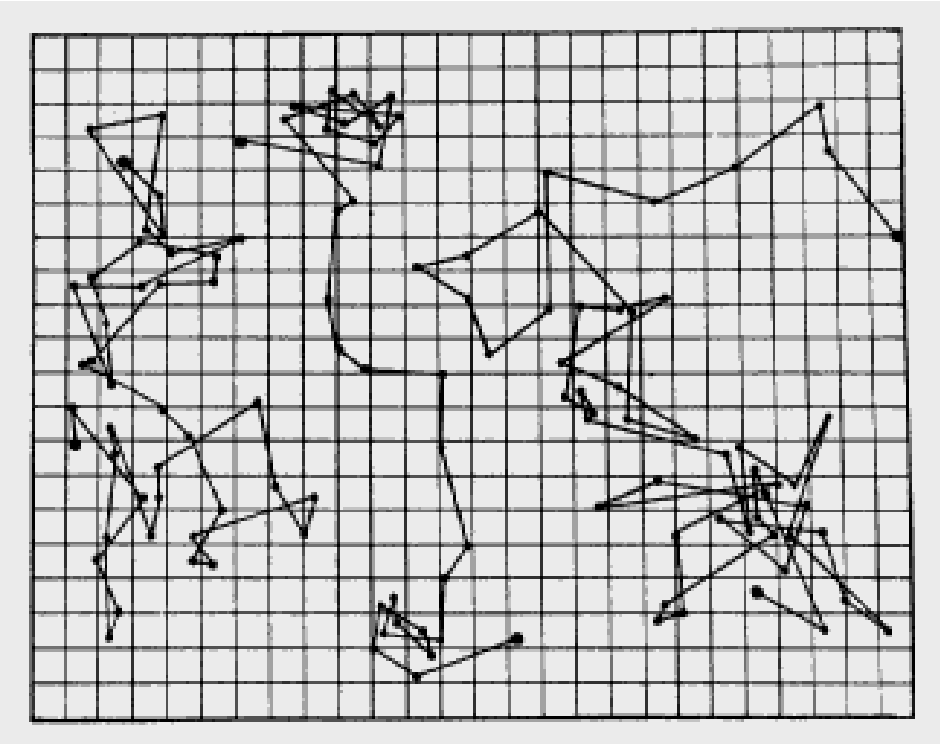
- Easy! Just solve Chapman-Kolmogorov equations:

$$\frac{\partial P_{ij}}{\partial t}(s;t) = \sum_k P_{ik}(s;t) Q_{kj}$$

- Not so easy when $|I|$ is large.

Brownian motion

[Einstein 1905, Langevin 1908] – The birth of stochastic calculus



- Isotropic jumps;
- unbounded continuous domain;
- vanishing transition probabilities.

Fluid limit

- Solving CK is hard.
- Fluid limit:

$$\frac{dX}{dt} = \beta(X) + \textit{noise}$$

- deterministic + stochastic.
- If stochastic $<<$ deterministic, solve classical ODE.
- ODE solution $\rightarrow$ mean behaviour, as $N \rightarrow \infty$.

[Fokker 1914, Planck 1917; Kolmogorov 1931]
Probability distribution evolution.

Fluid limit

States naturally ordered in $\mathbb{R}^d$; and
transitions expressible in terms of states:

Differential CK equation = Master equation $\approx$ Fokker-Planck equation.

- $\mathbf{x} : I \to \mathbb{R}^d$.
- $q(\xi, \xi') \to q(\mathbf{x}, \mathbf{x} + \Delta\mathbf{x})$
- Birth/death process, chemical reaction systems, etc.
- Under some scaling, conditions for fluid limit fulfilled.
- Approximation becomes exact in some limit of infinite state system size (i.e. infinite state density, transition distance $\to 0$).

[Van Kampen, Kurtz]

Examples of pCTMCs

Figure 1: Predator-prey systems in ecology: the Lotka-Volterra model.

Differential equation approximations for Markov chains

[Darling & Norris, 2008]

- Map $\mathbf{x} : I \to \mathbb{R}^d$, I discrete state-space.
- Define *drift vector*

$$\beta(\xi) \approx \sum_{\xi' \neq \xi} (\mathbf{x}(\xi') - \mathbf{x}(\xi))\, q(\xi, \xi')$$

- Then mapped process

$$\mathbf{x}(\xi(t)) = \mathbf{x}(\xi(0)) + M(t) + \int_0^t \beta(\xi(s))ds$$

Differential equation approximations for Markov chains

- Construct *drift vector field* $b(\mathbf{x})$ over continuous $\mathbb{R}^d$.
- Solution to $\dot{\mathbf{x}}_t = b(\mathbf{x}_t)$:

$$\mathbf{x}_t = \mathbf{x}_0 + \int_0^t b(\mathbf{x}_s) ds,$$

converges to mapped CTMC solution

$$\langle \mathbf{x}(\xi(t)) \rangle$$

(under scaling, etc.).

Differential equation approximations for Markov chains

- Construct *drift vector field* $b(\mathbf{x})$ over continuous $\mathbb{R}^d$.
- Solution to $\dot{\mathbf{x}}_t = b(\mathbf{x}_t)$:

$$\mathbf{x}_t = \mathbf{x}_0 + \int_0^t b(\mathbf{x}_s) ds,$$

converges to mapped CTMC solution

$$\langle \mathbf{x}(\xi(t)) \rangle$$

(under scaling, etc.).

No general way to construct $\mathbf{x}$, $b(\mathbf{x})$; manual construction. Address this:
(1) algorithm to embed state-space,
(2) infer *drift vector field*.

Fluid approximations for Q matrices

General idea

- $\exists$ fluid approximations for structured CTMCs (e.g. populations, queues, etc.).
- Automate fluid approximation:
  - *embedding of states* in $\mathbb{R}^d$.
  - *construction of drift vector* for dynamics in $\mathbb{R}^d$.
- Procedure should be:
  (1) close to manual constructions in simple cases;
  (2) *good enough* for other cases (where no manual approximations exist).

The trivial embedding

- One can trivially embed into $\mathbb{R}^{|I|}$.
- Simple (trivial) calculation shows that the embedded mean satisfies

$$\partial_t \langle x_t \rangle = Q^\top \langle x_t \rangle \ ;$$

  i.e. the C-K Equation!
- This is an exact representation (obviously).
- Slightly less trivial calculation: if $Q = Q^\top$ (which generally does not hold), any rotation would work.

The trivial embedding

- One can trivially embed into $\mathbb{R}^{|I|}$.
- Simple (trivial) calculation shows that the embedded mean satisfies

$$\partial_t \langle x_t \rangle = Q^\top \langle x_t \rangle \ ;$$

  i.e. the C-K Equation!
- This is an exact representation (obviously).
- Slightly less trivial calculation: if $Q = Q^\top$ (which generally does not hold), any rotation would work.
- Spectral analysis of Q might help.
- Link to manifold learning.

Laplacian eigenmaps

Construct map $\mathbf{x} : I \to \mathbb{R}^d$, I discrete state-space.

Think of Q as a network (nodes are states, edges allowed transitions)

Allows Laplacian eigenmaps [Mikhail Belkin, 2003]

Properties:

- Preserve locality $\implies$ limit transition size.
- Generally, as $|I|$ increases, states get closer.

Laplacian eigenmaps

Procedure:

- For network Q, construct unweighted Laplacian matrix L where

$$L_{ij} = 1 - \delta_{q_{ij},0} \qquad \forall i \neq j$$
$$\text{and } L_{ii} = -\sum_{j} L_{ij}.$$

- Take d eigenvectors with d smallest eigenvalues (except 0).
- Give state coordinates of d-dimensional embedding.

Gaussian process regression

Construct *drift vector field* $b(\mathbf{x})$ over continuous $\mathbb{R}^d$.

- Have
$$b(\mathbf{x}(\xi)) = \beta(\xi)$$
only defined at $\mathbf{x}(\xi)$ points.
- What about the rest of the $\mathbb{R}^d$ domain?

Gaussian process regression

Construct *drift vector field* $b(\mathbf{x})$ over continuous $\mathbb{R}^d$.

- Have
$$b(\mathbf{x}(\xi)) = \beta(\xi)$$
only defined at $\mathbf{x}(\xi)$ points.
- What about the rest of the $\mathbb{R}^d$ domain?
- Regress! Gaussian processes can estimate values in between.
- Gaussian processes are Lipschitz continuous!

Some challenges:

- Kernel choice.
- Hyperparameter choice.
- Unknown Lipschitz constant.

Laplacian eigenmaps + GP: sanity check

Does it produce standard embeddings (e.g. for pCTMCs)?

Theorem

Let $\mathcal{C}$ be a pCTMC, whose underlying transition graph is a multi-dimensional grid graph. The unweighted Laplacian fluid approximation of $\mathcal{C}$ coincides with the canonical fluid approximation in the hydrodynamic scaling limit.

A foray into *diffusion maps*

Generalisation[1] of Laplacian eigenmaps.
Deals with:

- non-uniform sampling on the manifold $p = e^{-U(\mathbf{x})}$;
- (extension[2]) asymmetric graphs $Q \neq Q^\top$.

Backward diffusion operators:

$$-\partial_t = \mathcal{H}^{(\alpha)}_{aa} = \Delta + (\mathbf{r} - 2(1-\alpha)\nabla U)\cdot\nabla, \qquad \text{and}$$
$$-\partial_t = \mathcal{H}^{(\alpha)}_{ss} = \Delta - 2(1-\alpha)\nabla U\cdot\nabla.$$

[1] Ronald R. Coifman and Stéphane Lafon. "Diffusion maps". In: *Applied and Computational Harmonic Analysis* 21.1 (July 2006), pp. 5–30.

[2] Dominique C. Perrault-joncas and Marina Meila. "Directed Graph Embedding: an Algorithm based on Continuous Limits of Laplacian-type Operators". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor et al. Curran Associates, Inc., 2011, pp. 990–998.

Geometric fluid approximation

- Start with general CTMC with generator matrix Q.
- Embed network in $\mathbb{R}^d$ using diffusion maps.
- Define drift vector on embedded nodes by pushing forward transitions.
- Use these as observations in a GP-regression model.
- Completely general, does not require a special population structure.

Geometric fluid approximation

- Start with general CTMC with generator matrix Q.
- Embed network in $\mathbb{R}^d$ using diffusion maps.
- Define drift vector on embedded nodes by pushing forward transitions.
- Use these as observations in a GP-regression model.
- Completely general, does not require a special population structure.
- We call this the *Geometric Fluid Approximation*.

Example: birth and death

2 species birth-death process, embedded in $\mathbb{R}^2$ space.

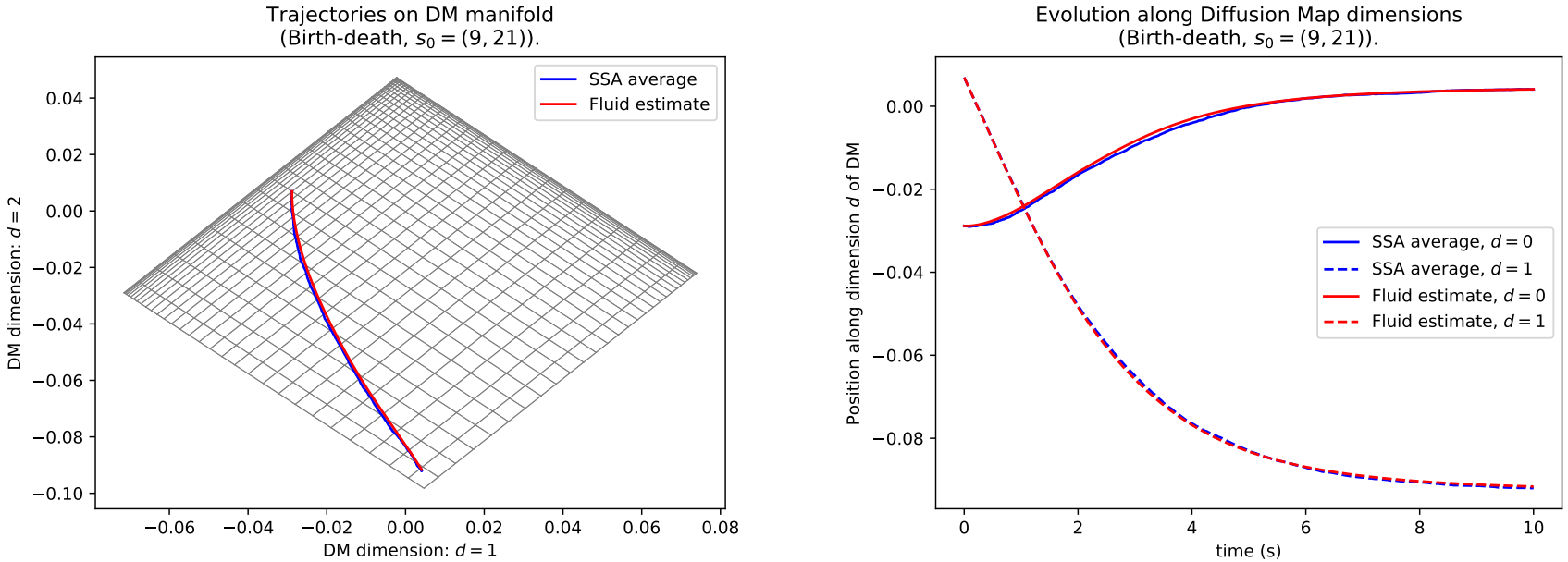


Figure 2: Sanity check – expect good agreement.

Example: birth and death

2 species birth-death process, embedded in $\mathbb{R}^{|I|}$ space.

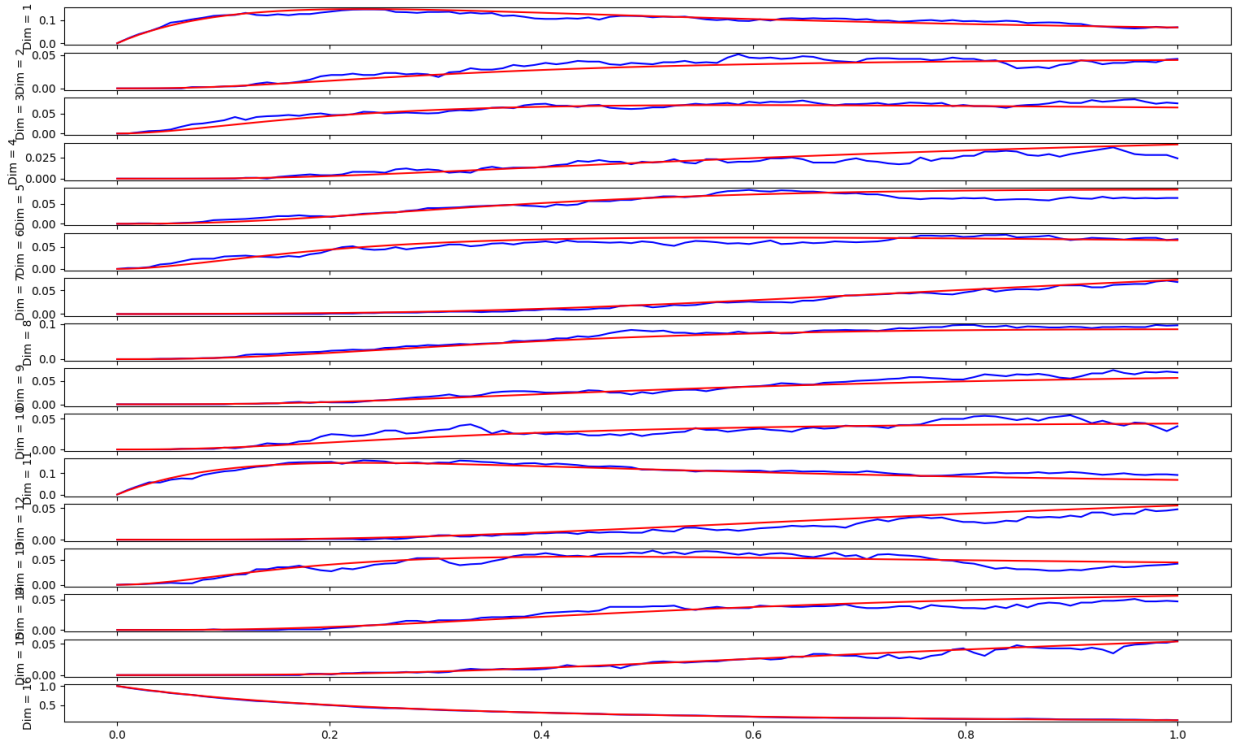


Figure 3: Every state a dimension – expect perfect agreement.

Example: Lotka-Volterra model

Foxes consume rabbits and decay.

$$R \xrightarrow{b=0.1} 2R;$$
$$R + F \xrightarrow{c=0.01} 2F;$$
$$F \xrightarrow{d=0.2} \emptyset.$$

Example: Lotka-Volterra model

2 species Lotka-Volterra process, embedded in $\mathbb{R}^2$ space.

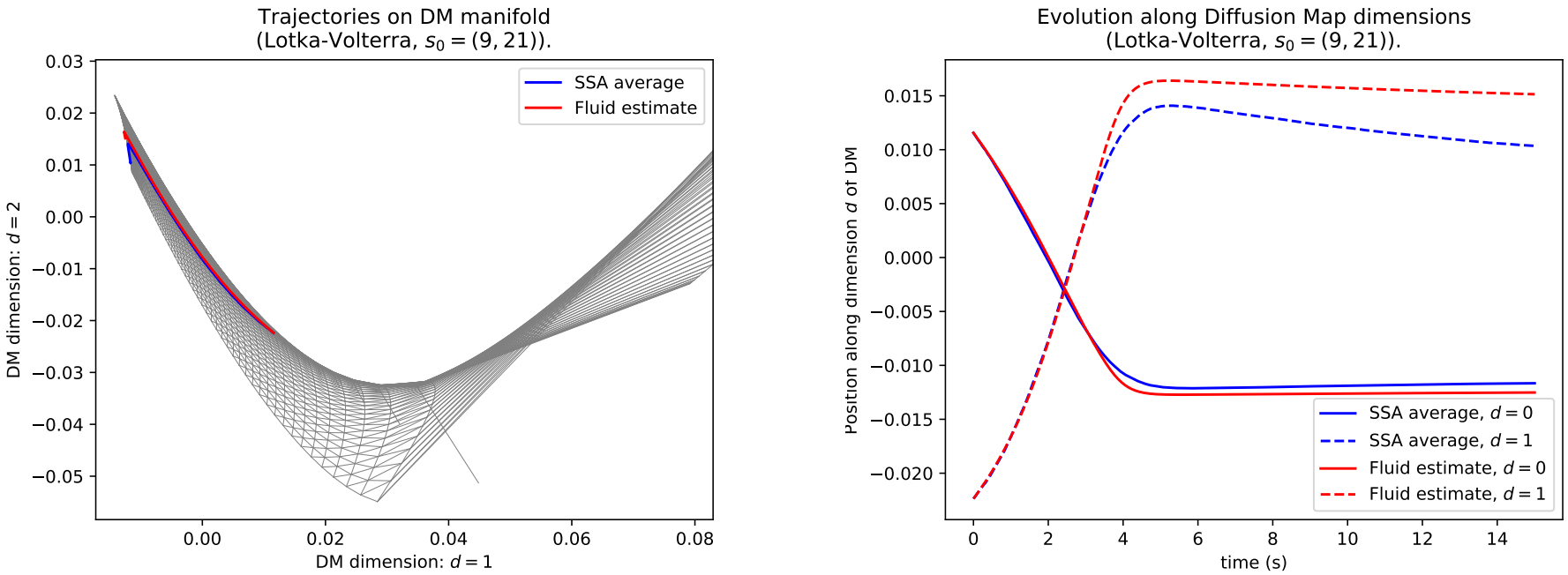


Figure 4: LV with oscillations.

Example: Lotka-Volterra perturbed

Perturbed 2 species Lotka-Volterra process, embedded in $\mathbb{R}^2$ space.

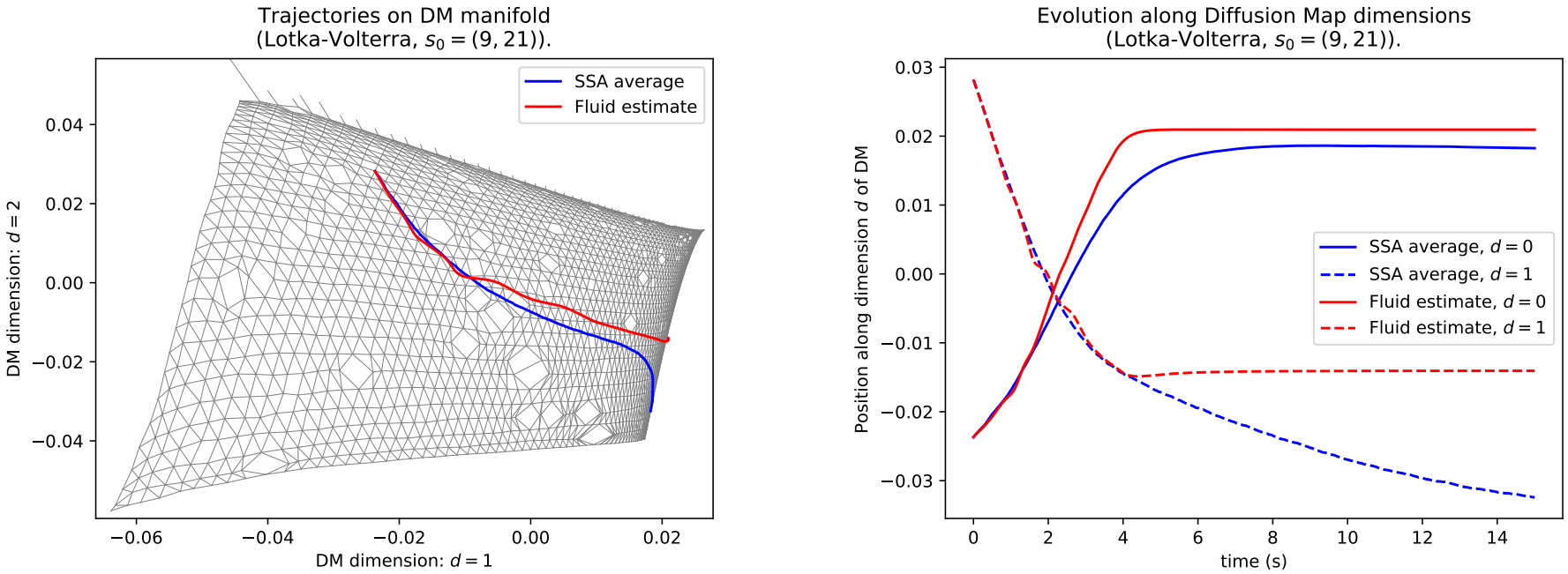


Figure 5: LV with oscillations, rates corrupted by $|0.5\eta|, \eta \sim \mathcal{N}(0, 1)$ noise.

Example: gene ON-OFF

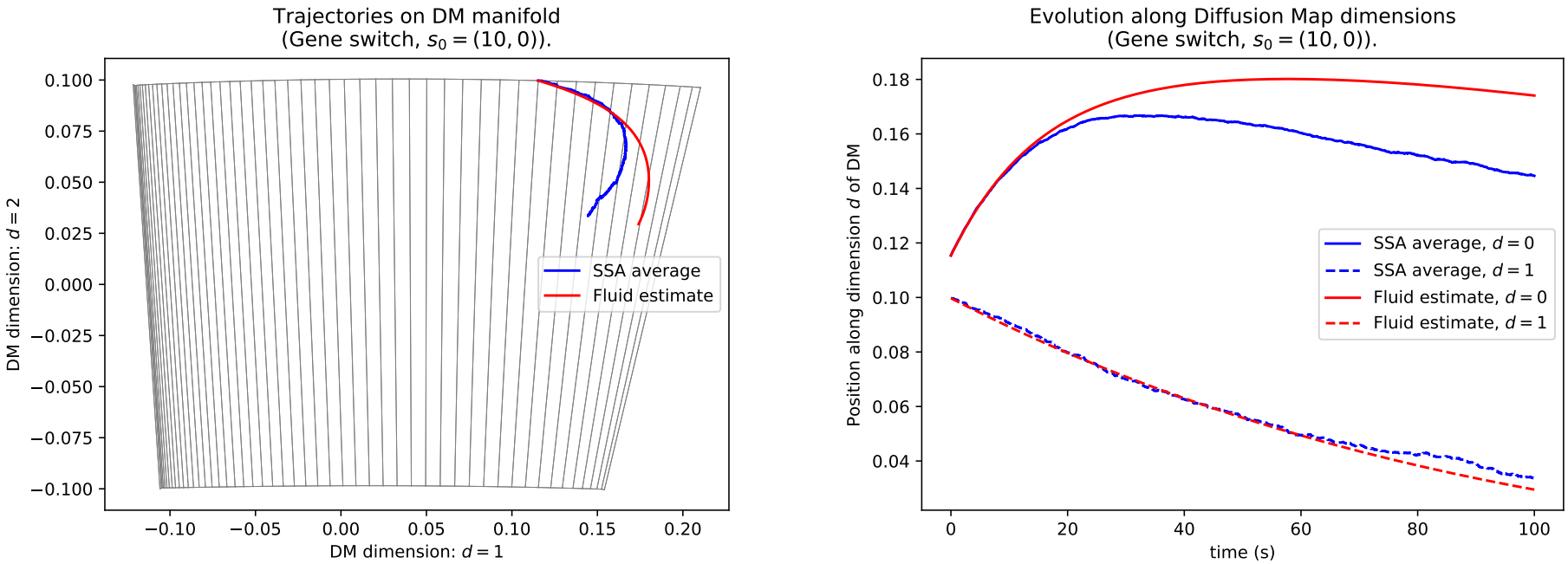


Figure 6: The genetic switch model with a faster switching rate ($5 \cdot 10^{-3}\text{s}^{-1}$), showing how the fluid solution (red) diverges from the projected mean evolution (blue) after $t \approx 20$s; the qualitative aspects of the trajectory remain similar.

Example: SIRS model

SIRS model Embedded in $\mathbb{R}^3$ space.

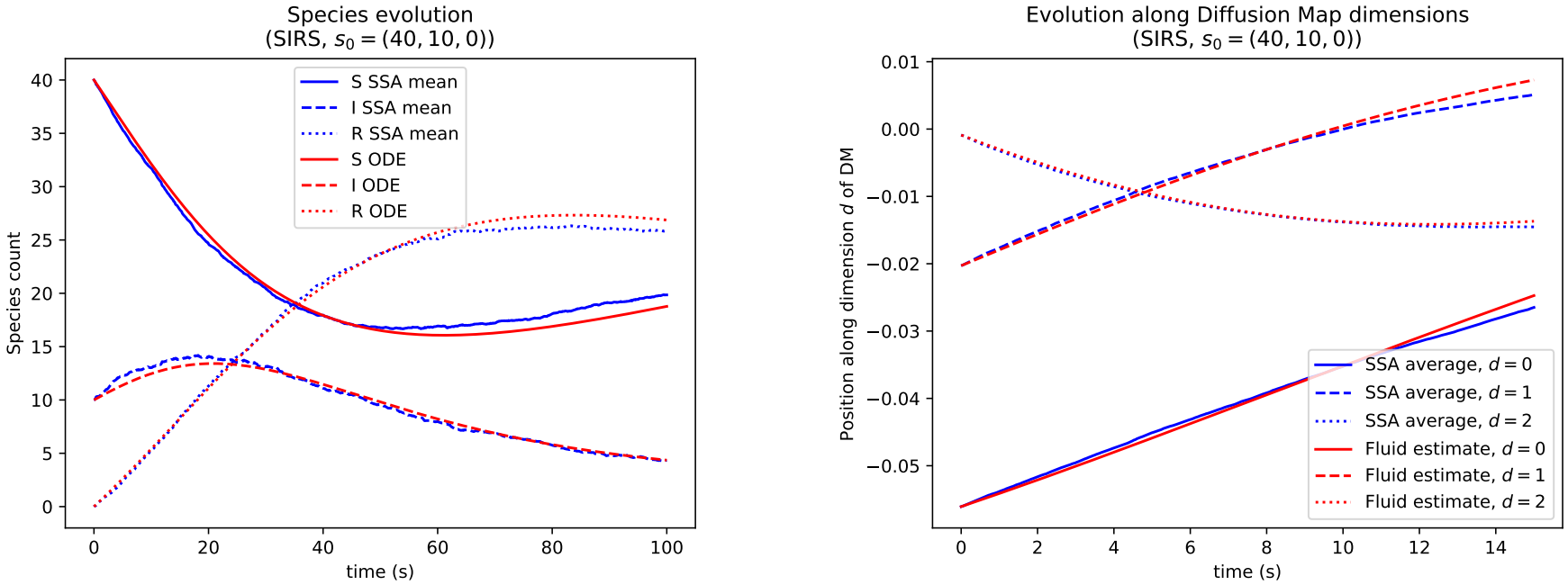


Figure 7: Left: SIRS classical fluid embedding in $\mathbb{R}^3$. Right: SIRS DM embedding in $\mathbb{R}^3$.

First passage time (FPT) distribution

SIRS model Embedded in $\mathbb{R}^3$ space.

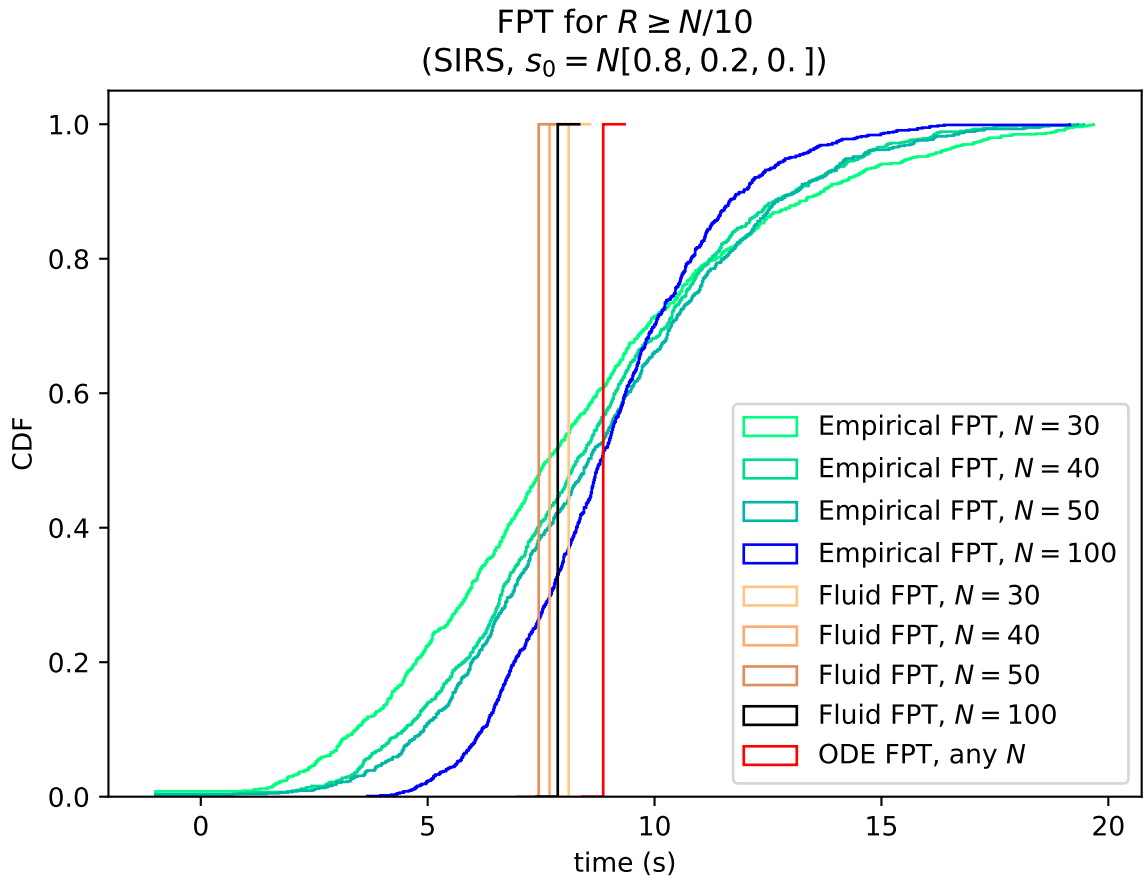


Figure 8: FPT CDF $\rightarrow$ classical ODE estimate. Our construction is close.

Summary

- No general way from $\partial_t P = Q^\top P \ \rightarrow \ \partial_t p = \mathcal{H} p$.
- Ingredients:
  - $\mathbf{x} : I \rightarrow \mathbb{R}^d$ — diffusion maps;
  - $\beta(\xi) \rightarrow b(\mathbf{x})$ — Gaussian process regression.
- State/transition agnostic bridge from discrete CTMC to continuous diffusion process.

Thank you

Feedback very welcome!

Continuity conditions

For a master equation to converge to FPE:

1. Jump sizes must vanish

$$\lim_{\Delta t \to 0} \left\{ \frac{1}{\Delta t} p(x, t + \Delta t \mid z, t) \right\} = W(x \mid z, t);$$

$$\lim_{N \to \infty} W(x \mid z, t) = 0;$$

   uniformly in $x,\ z,\ t$ for $|x - z| \geq \epsilon$.

2. Drift vector field is

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|x-z|<\epsilon} dx\ (x_i - z_i)\ p(x,\ t + \Delta t \mid z, t) = A_i(z,\ t) + O(\epsilon)\ ;$$

   and

3. Diffusion matrix field is

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|x-z|<\epsilon} dx\ (x_i - z_i)(x_j - z_j)\ p(x,\ t + \Delta t \mid z, t) = B_{ij}(z,\ t) + O(\epsilon)$$

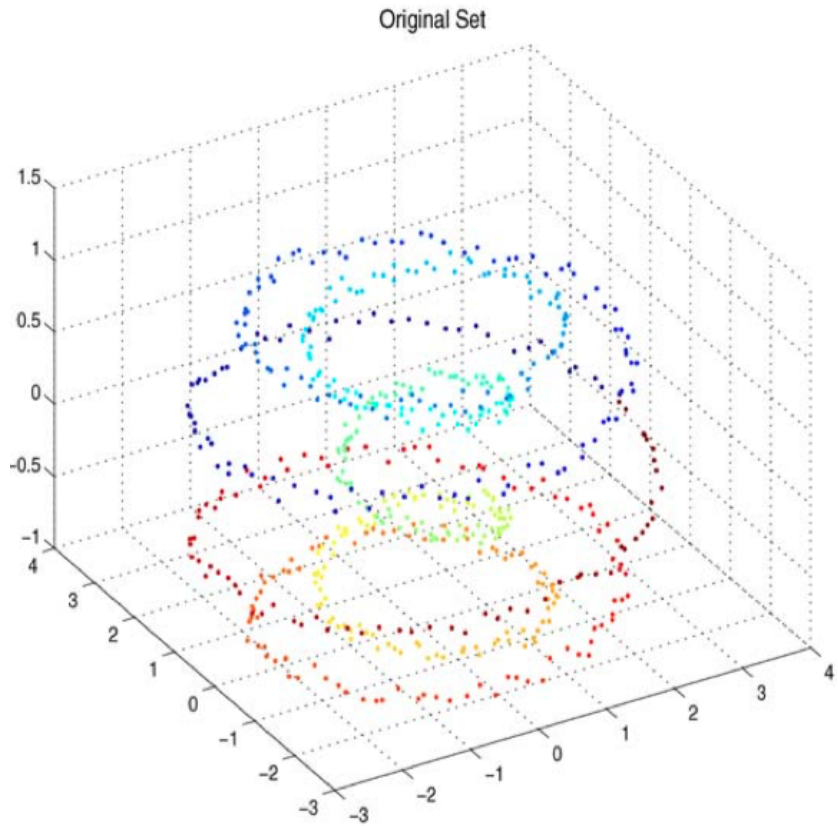
   uniformly in $z,\ \epsilon,\ t$.

Unsupervised learning problem

High-dimensional data, $\mathbf{x} \in \mathbb{R}^m$.

Want to find

- low-dimensional projection,
- clusters.



3

[3] Coifman and Lafon, see n. 1.

Usual assumptions

- Data $\mathcal{D}$ lie on lower dimensional manifold $\mathcal{M} \subset \mathbb{R}^m$.
- $\mathcal{M}$ is continuous.
- Data not necessarily uniformly sampled on manifold: density $\mu(\mathbf{x}) = q(\mathbf{x}) = e^{-U(\mathbf{x})}$.
- What to examine...
- ... in order to recover map $\Psi : \mathcal{D} \to \mathcal{M}$, and potential U?

Kernel, kernel, on the wall
how similar am I to the other data points?

Capture geometry by constructing similarity graph W,

$$W_{ij} = k(\mathbf{x}_i, \mathbf{x}_j),$$

with $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_{>0}$, and

- symmetry, $k(x, y) = k(y, x)$,
- p.s.d., $k(x, y) \geq 0$.

Usually pick $k_\epsilon(x, y) = \exp\left(-\|x - y\|^2/\epsilon\right)$.

Graph Laplacians and Random Walks

Normalised kernel (graph Laplacian)
as discrete-time Markov chain where:

$$p_\epsilon(x,y) = \frac{k_\epsilon(x,y)}{d_\epsilon(x) = \int_{\mathcal{M}} k_\epsilon(x,y)dy} \ ,$$

Transition probability after t steps:

$$p_t(x,y) = P^t,$$

where

$$P\,|f\rangle = \int_{\mathcal{M}} p(x,y)f(y)d\mu(y).$$

Random Walks

Leverage spectral theory for MCs that are:

- ergodic,
- aperiodic, (stationary distribution is $\pi(x)$)
- reversible.

Eigen-decompose R.W. on $\mathcal{D}$:

$$P\psi_k = \lambda_k \psi_k,$$

s.t. $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N-1} \geq 0$.

Diffusion distance

Define *diffusion distance*:

$$D_t(x, y) \triangleq \|p_t(x, \cdot) - p_t(y, \cdot)\|^2_{L^2(\mathcal{M}, d\mu/\pi)} \ ;$$

demand it matches Euclidean distance in mapped space:

$$\|\Psi_t(x) - \Psi_t(y)\| = D_t(x, y) = \left[\sum_k \lambda_k^{2t} \left(\psi_k(x) - \psi_k(y)\right)^2\right]^{1/2} .$$

Satisfy up to precision δ with

$$\boxed{\Psi_t(\mathbf{x}) \triangleq \left(\lambda_1^t \psi_1(\mathbf{x}),\ \lambda_2^t \psi_2(\mathbf{x}),\ \ldots,\ \lambda_d^t \psi_d(x)\right)},$$

where $d = \max\left\{k \in \mathbb{N} \mid \lambda_k^t > \delta \lambda_1^t\right\}$.

Anisotropic diffusion

What if $\mu(\mathbf{x}) = q(\mathbf{x}) = e^{-U(\mathbf{x})}$ is non-uniform?

- Sampling $\mathcal{M}$ is biased,
- need to consider effects of $q(\mathbf{x}) = e^{-U(\mathbf{x})}$.

Anisotropic kernel

$$k_\epsilon^{(\alpha)}(x, y) = \frac{k_\epsilon(x, y)}{q_\epsilon^\alpha(x) q_\epsilon^\alpha(y)},$$

$\alpha \in \mathbb{R}_{\geq 0}$ can separate geometry from density.

- $\alpha = 0$: normalised graph Laplacian (Laplacian eigenmaps);
- $\alpha = 1/2$: Fokker-Planck as limit operator (more later);
- $\alpha = 1$: Laplace-Beltrami operator Δ.

Limit operators[4]

Think of $\mathcal{D}$ as samples of

$$\dot{\mathbf{x}} = -\nabla U(\mathbf{x}) + \sqrt{2}\dot{\mathbf{w}} \ .$$

In limit $\epsilon \to 0,\ N \to \infty$
evolution operators of $|f\rangle$:

$$\frac{\partial}{\partial t}|f\rangle = \mathcal{H}_f^{(\alpha)}|f\rangle = \Big[\Delta - 2\alpha\nabla U\cdot\nabla + (2\alpha-1)(\|\nabla U\|^2 - \Delta U)\Big]|f\rangle\,,$$
$$-\frac{\partial}{\partial t}|f\rangle = \mathcal{H}_b^{(\alpha)}|f\rangle = [\Delta - 2(1-\alpha)\nabla U\cdot\nabla]|f\rangle \ .$$

If $\mu(\mathbf{x}) \neq c$, α matters.

[4] Boaz Nadler et al. "Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators". In: *Advances in Neural Information Processing Systems 18*. Ed. by Y. Weiss, B. Schölkopf, and J. C. Platt. MIT Press, 2006, pp. 955–962.

Anisotropic projections[5]

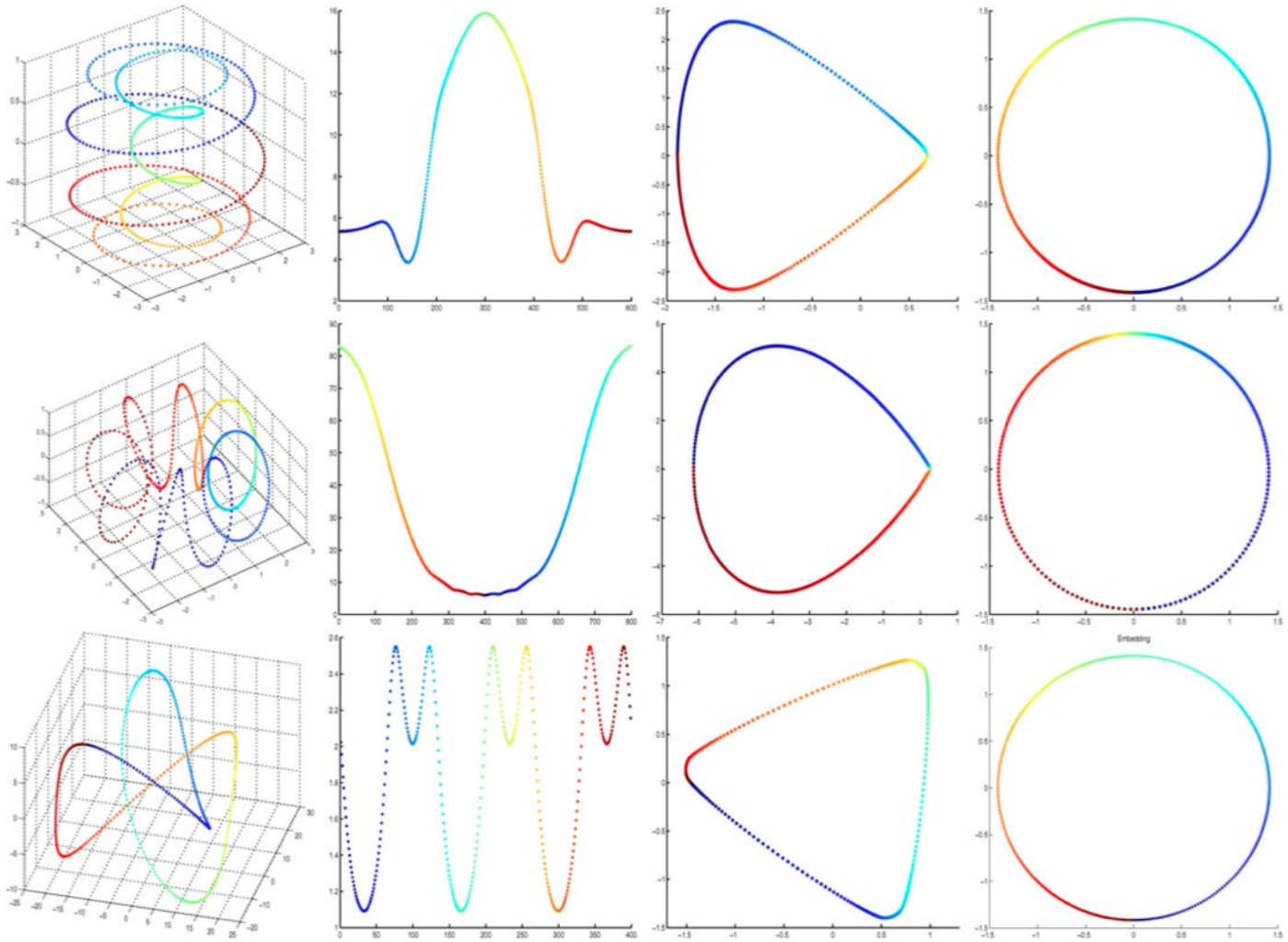


Fig. 4. From left to right: original curves, the densities of points, the embeddings via the graph Laplacian ($\alpha = 0$) and the embeddings via the Laplace–Beltrami approximation ($\alpha = 1$). In the latter case, the curve is embedded as a perfect circle and the arclength parametrization is recovered.

[5] Coifman and Lafon, see n. 1.

Extension to directed graphs

Directed Graph Embedding: an Algorithm based on Continuous Limits of Laplacian-type Operators

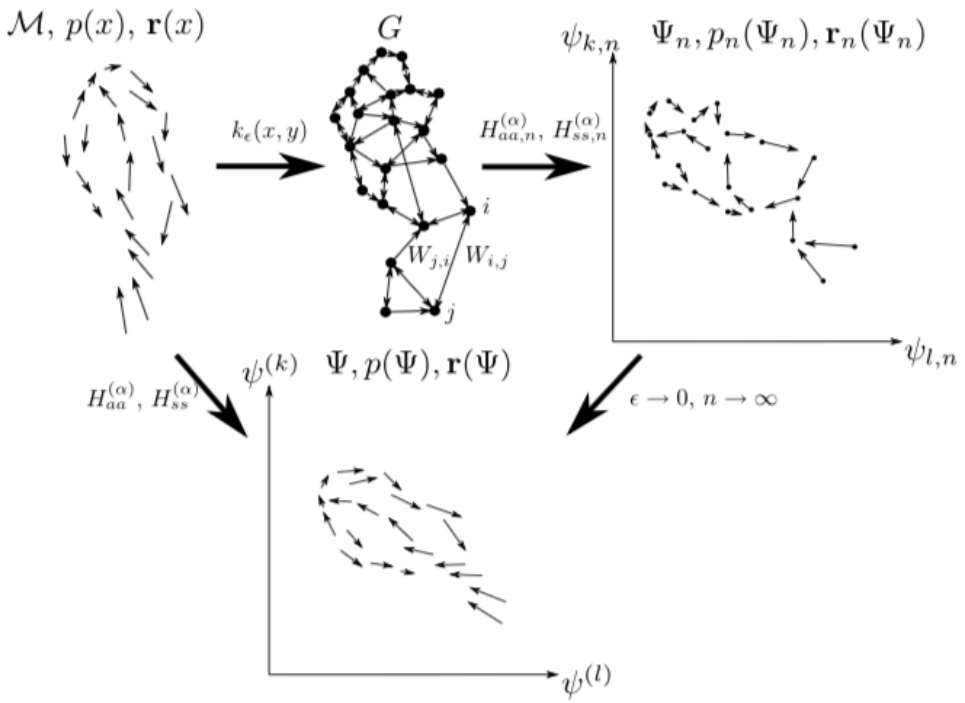
Dominique C. Perrault-Joncas
Department of Statistics
University of Washington
Seattle, WA 98195
dcpj@stat.washington.edu

Marina Meilă
Department of Statistics
University of Washington
Seattle, WA 98195
mmp@stat.washington.edu

6

[6] Perrault-joncas and Meila, see n. 2.

Extension to directed graphs



From left to right: the graph generative process mapping the sample on $\mathcal{M}$ to geometric random graph G via the kernel $k_\epsilon(x, y)$, then the subsequent embedding Ψ_n of G by operators $H^{(\alpha)}_{aa,n}$, $H^{(\alpha)}_{ss,n}$ (defined in section 3.1). As these operators converge to their respective limits, $H^{(\alpha)}_{aa}$ and $H^{(\alpha)}_{ss}$, so will $\Psi_n \to \Psi$, $p_n \to p$, and $\mathbf{r}_n \to \mathbf{r}$.
We design an algorithm that, given G, produces the top right embedding (Ψ_n, p_n, and $\mathbf{r}_n$).

Figure 1: Schematic of our framework.

The question is then as follows: can the generative process' geometry $\mathcal{M}$, distribution $p = e^{-U}$, and directionality $\mathbf{r}$, be recovered from G? In other words, is there an embedding of G in $\mathbb{R}^m$, $m \geq d$ that approximates all three components of the process and that is also consistent as sample size increases and the bandwidth vanishes? In the case of undirected graphs, the theory of Laplacian

Asymmetric kernel

$$k_\epsilon^{(\alpha)}(x,y) = h_\epsilon(x,y) + a_\epsilon(x,y);$$
$$h_\epsilon(x,y) = \frac{k_\epsilon(x,y)}{q_\epsilon^\alpha(x) q_\epsilon^\alpha(y)},$$
$$a_\epsilon(x,y) = -a_\epsilon(y,x) = \frac{\mathbf{r}(x,y)}{2}(y-x)h_\epsilon(x,y).$$

Backward diffusion operators:

$$-\partial_t = \mathcal{H}_{aa}^{(\alpha)} = \Delta + (\mathbf{r} - 2(1-\alpha)\nabla U)\cdot\nabla, \qquad \text{and}$$
$$-\partial_t = \mathcal{H}_{ss}^{(\alpha)} = \Delta - 2(1-\alpha)\nabla U\cdot\nabla.$$

Stochastic processes and associated generators

Related as limiting cases[7].

Case	**Operator**	**Stochastic Process**
$\epsilon > 0$ $N < \infty$	finite $N \times N$ matrix P	R.W. in discrete space discrete in time (DTMC)
$\epsilon > 0$ $N \to \infty$	operators T_f, T_b	R.W. in continuous space discrete in time
$\epsilon \to 0$ $N < \infty$	infinitesimal generator matrix $Q \in \mathbb{R}^{N \times N}$	Markov jump process; discreet in space, continuous in time
$\epsilon \to 0$ $N \to \infty$	infinitesimal generator $\mathcal{H}_f$	diffusion process continuous in space & time

[7] Boaz Nadler et al. "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems". In: *Applied and Computational Harmonic Analysis* 21.1 (July 2006), pp. 113–127.