Two problems in Astrostatistics:

Identification of Shell and Stream Substructures in Galaxy Debris[1]

&

Nonparametric Estimation in Contamination Models[2]

Bodhisattva Sen[3]
Department of Statistics
Columbia University, New York

Statistical Challenges in Modern Astronomy – VI
Carnegie Mellon University, Pittsburgh

9 June, 2016

[1] Joint work with R. Patra, R. Biswas, D. Hendel, K. Johnston
[2] Joint work with R. Patra
[3] Supported by NSF Grant DMS-1150435 (CAREER)

I. Identification of *shell* and *stream* substructures in galaxy debris

Tidal debris morphology: The *debris* from minor *mergers* can be broadly divided into *two* morphological categories – *streams* and *shells*

Figure : Galaxy NGC 5907 (left) and Galaxy NGC 474 (right)

- The goal of this collaborative project is to *identify* and *distinguish* different morphologies (i.e., shells and streams)

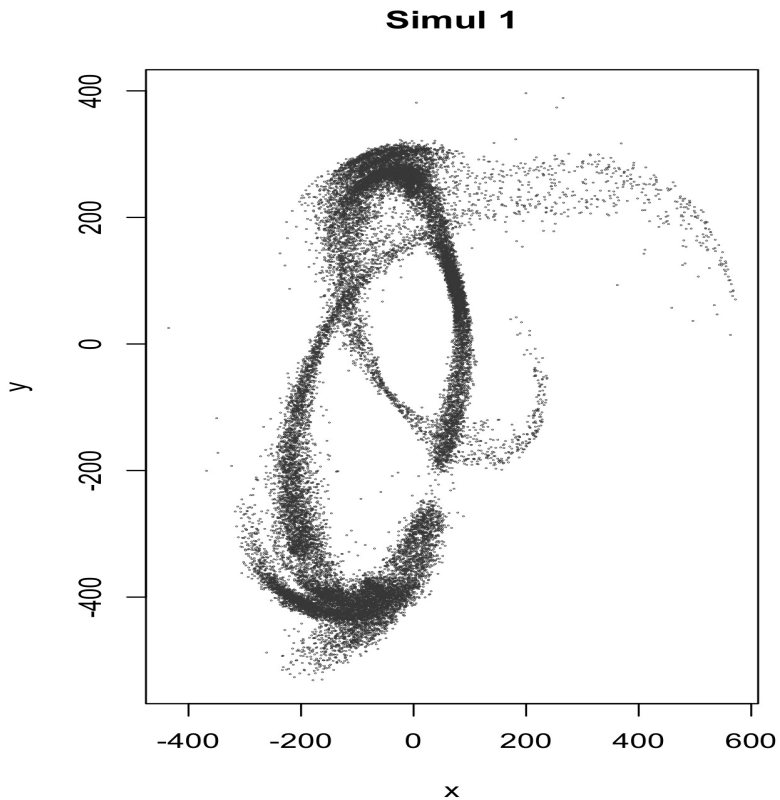
- We hope that this analysis will provide new insights into our understanding of *structure formation*

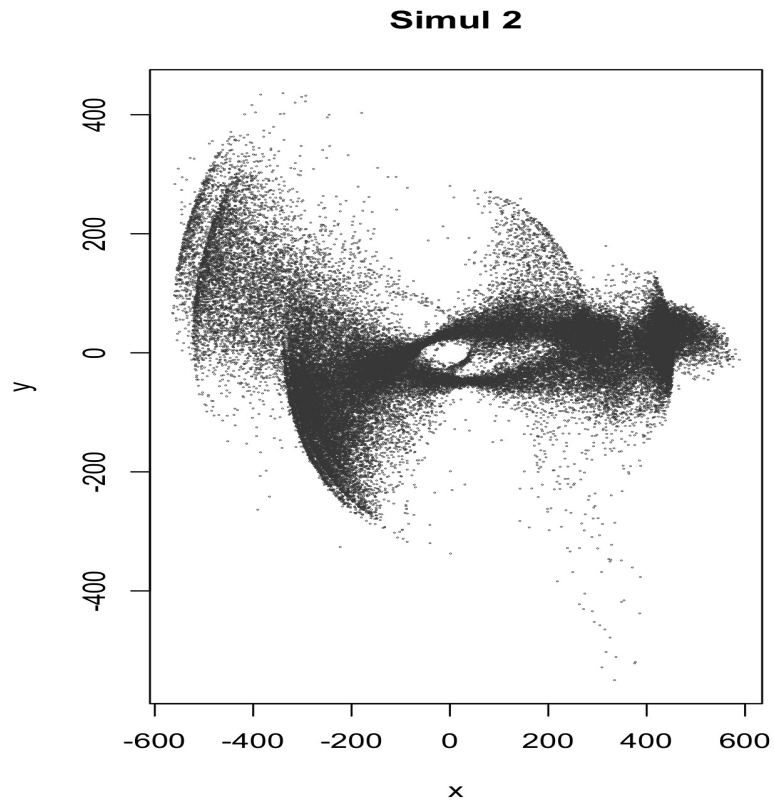
- *Stream*-like substructures stretch approximately along the progenitor's orbit, sometimes *wrapping* around the host multiple times

- *Shell*-like structures may extend both along and perpendicular to the path of their disrupted parent, forming an *umbrella*-shaped distribution of stars and/or *sharp* edges in the light distribution

- We study the outputs of *N-body* simulations (across a wide range of orbital and galactic parameters); Hendel and Johnston (2015)

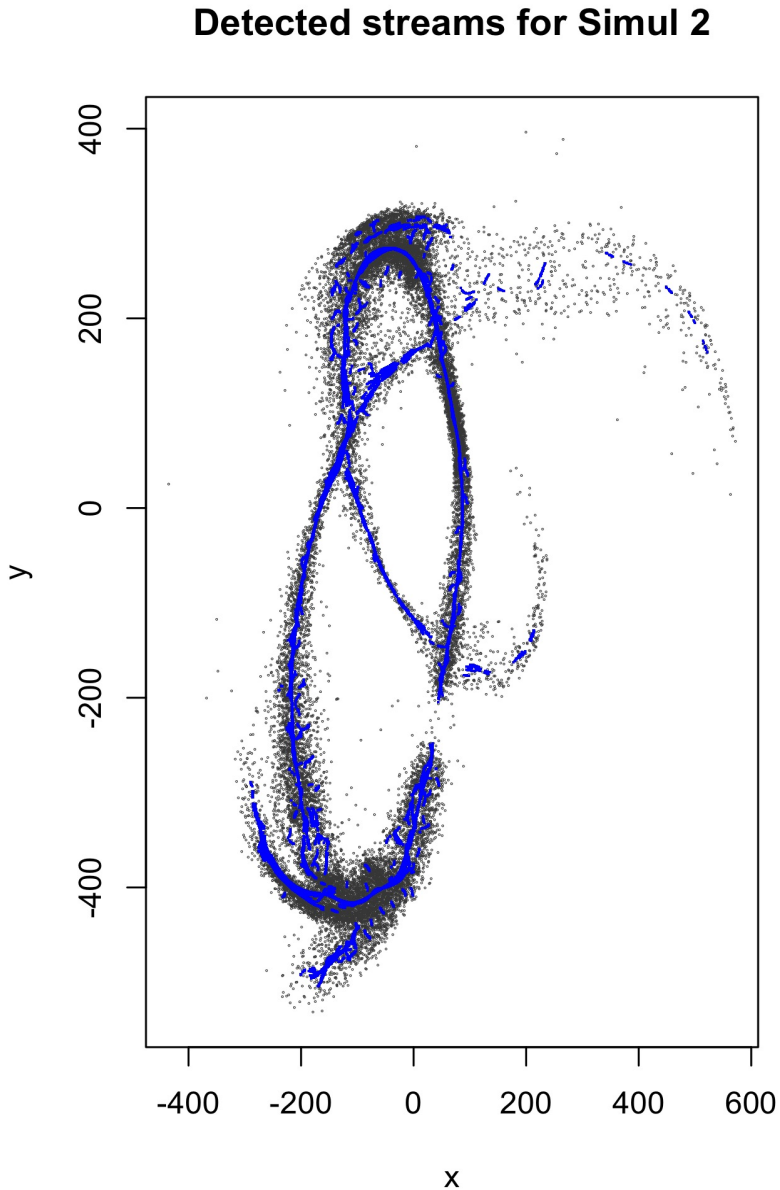
We will work with these *two* simulations

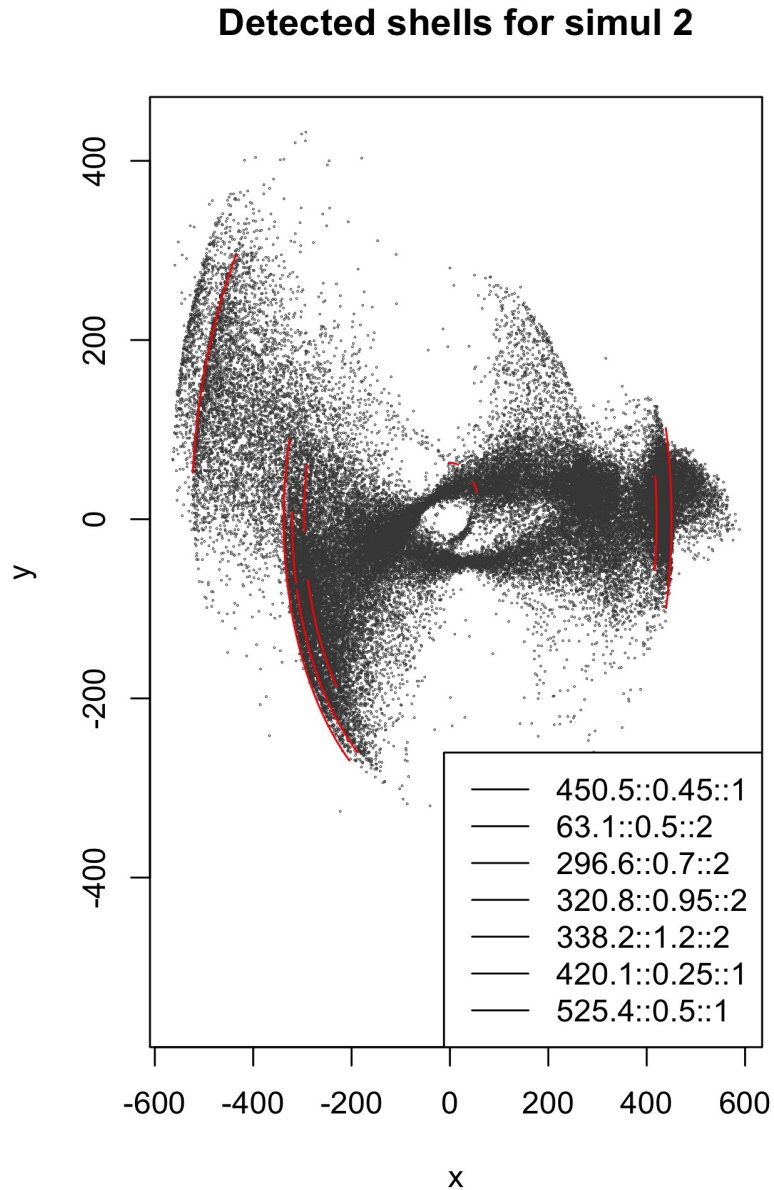




Shells versus Streams

- *Shells* have a *sharp edge* in surface brightness on one side and a *smooth decay* on the *other*
- *Streams* appear more *symmetric* around the *'ridge'*

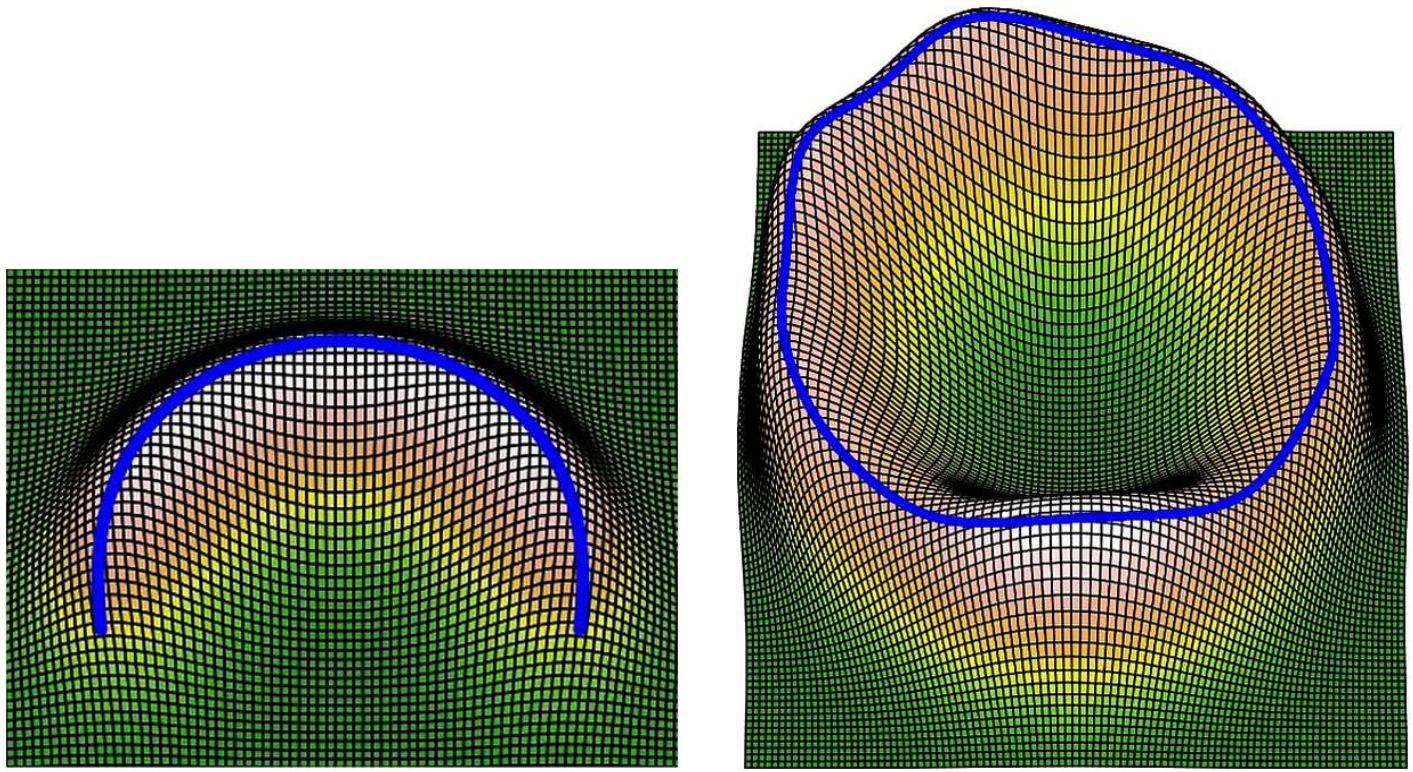

Detected streams for Simul 2
y
x
400
200
0
-200
-400
-400
-200
0
200
400
600


Detected shells for simul 2
y
x
400
200
0
-200
-400
-600
-400
-200
0
200
400
600
450.5::0.45::1
63.1::0.5::2
296.6::0.7::2
320.8::0.95::2
338.2::1.2::2
420.1::0.25::1
525.4::0.5::1

Shells versus Streams

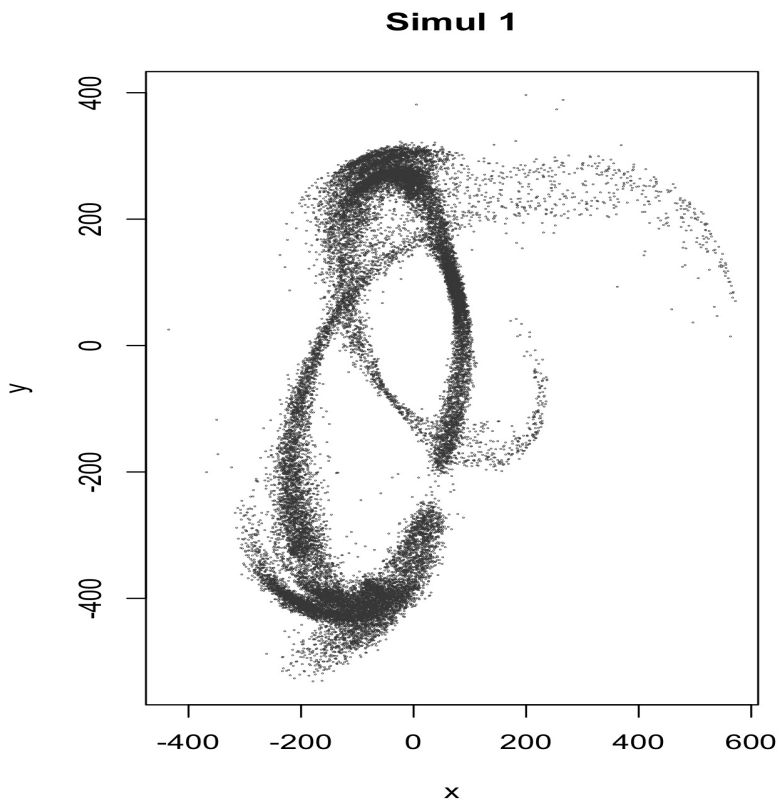
- *Shells* have a *sharp edge* in surface brightness on one side and *smooth decay* on the *other*
- *Streams* appear more *symmetric* around the *'ridge'*

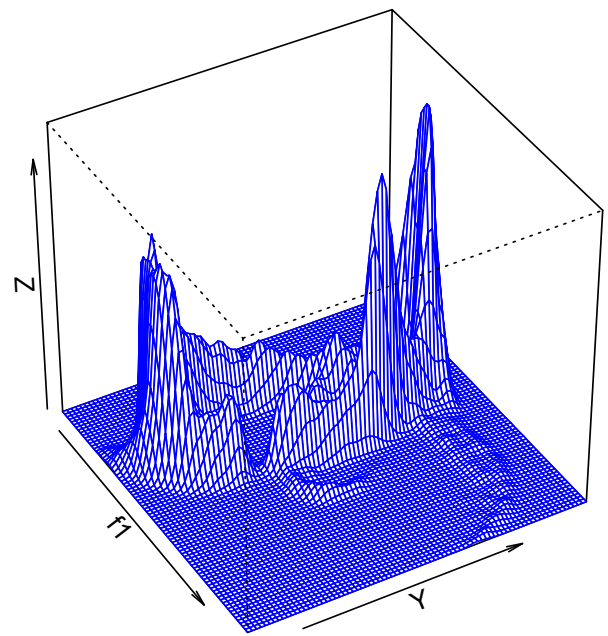
A ridge a *local maxima* in *one* direction (like a *mode*)



- **Data**: 2-D position of particles $(\{\vec{x}_i := (x_{i1}, x_{i2})\}_{i=1}^{n})$
- Construct *kernel density estimator* $p(\vec{x})$ to obtain the *surface density*, i.e.,

$$p(\vec{x}) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x_{i1} - x_1}{h}, \frac{x_{i2} - x_2}{h}\right)$$

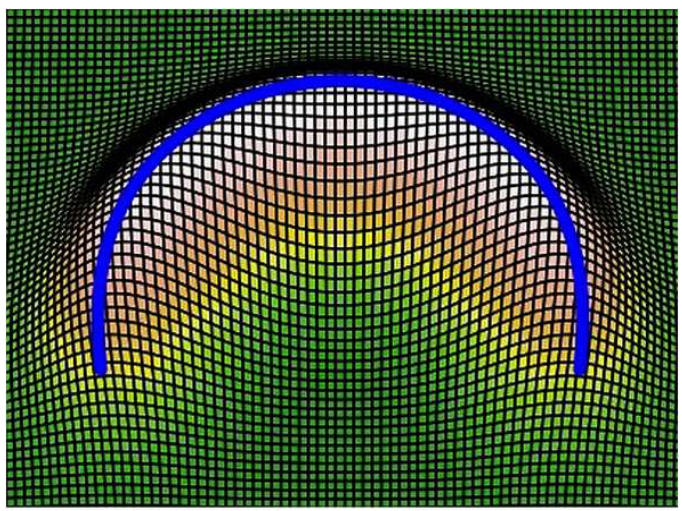


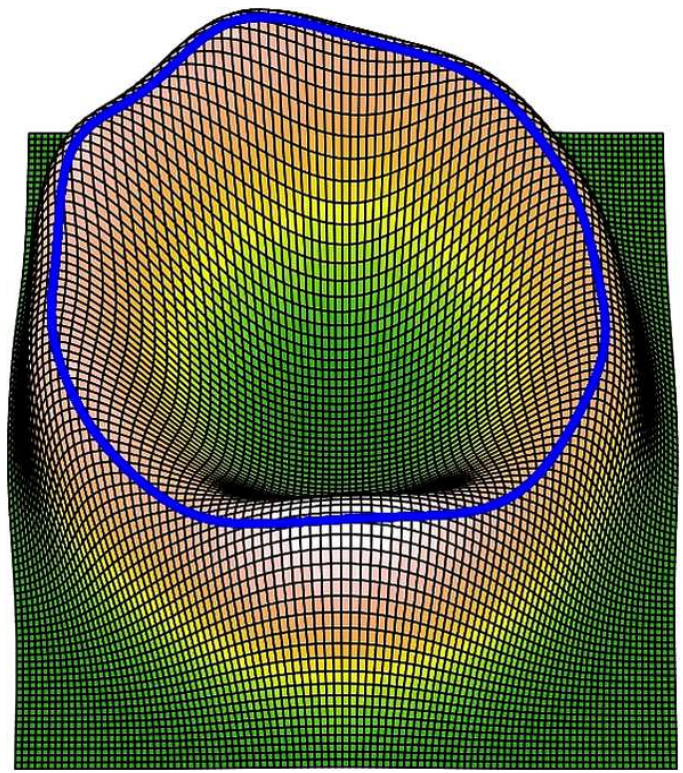


- *Hessian matrix*: $H(\vec{x}) = \frac{\partial^2}{\partial x_1 \partial x_2} p(\vec{x})$ at each point $\vec{x}$
- *Eigenvector* of the *smallest eigenvalue* of the $H(\vec{x})$: $\vec{v}(\vec{x})$
- The *ridge* is the set of points $\vec{x}_r$ such that

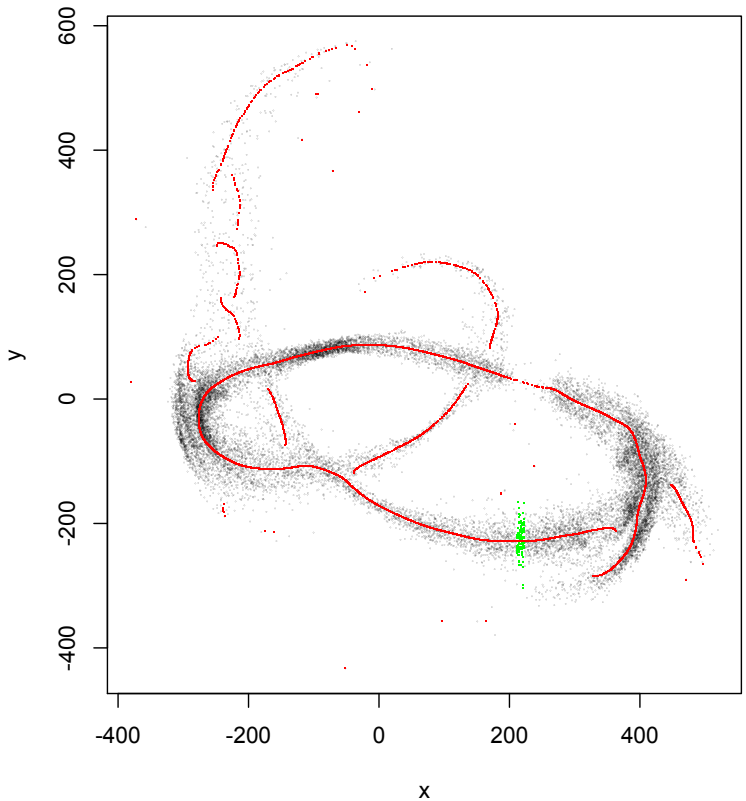
$$\vec{v}(\vec{x}_r)^\top \nabla p(\vec{x}_r) = 0 \quad \text{and} \quad \vec{v}(\vec{x}_r)^\top H(\vec{x}_r)\vec{v}(\vec{x}_r) < 0,$$

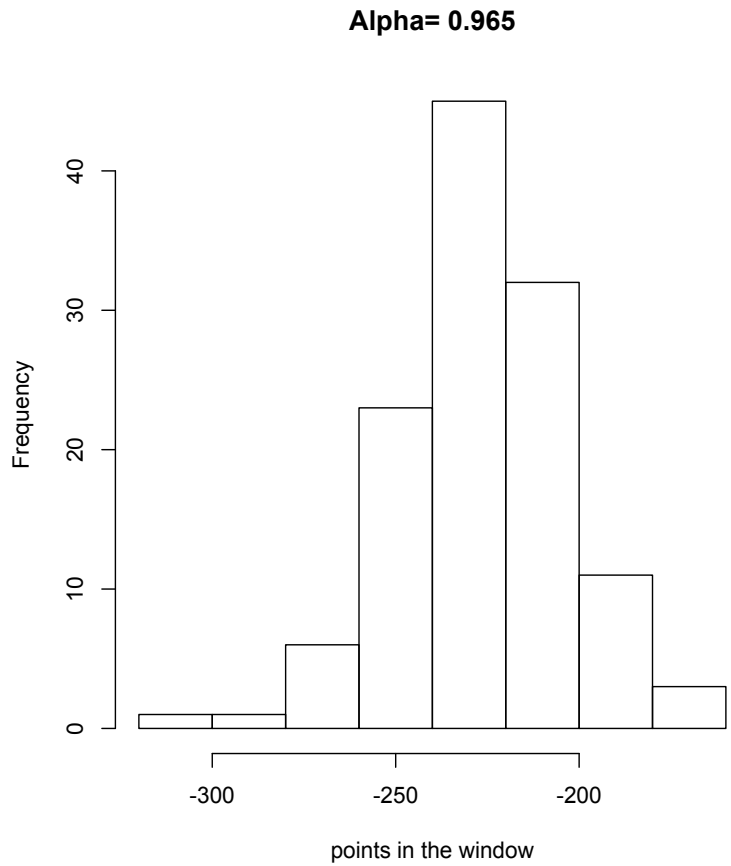
  local maxima in p along the *eigendirection* of the *smallest* eigenvalue
- *Subspace Constrained Mean Shift* [Ozertem, U. & Erdogmus, D. (2011)]

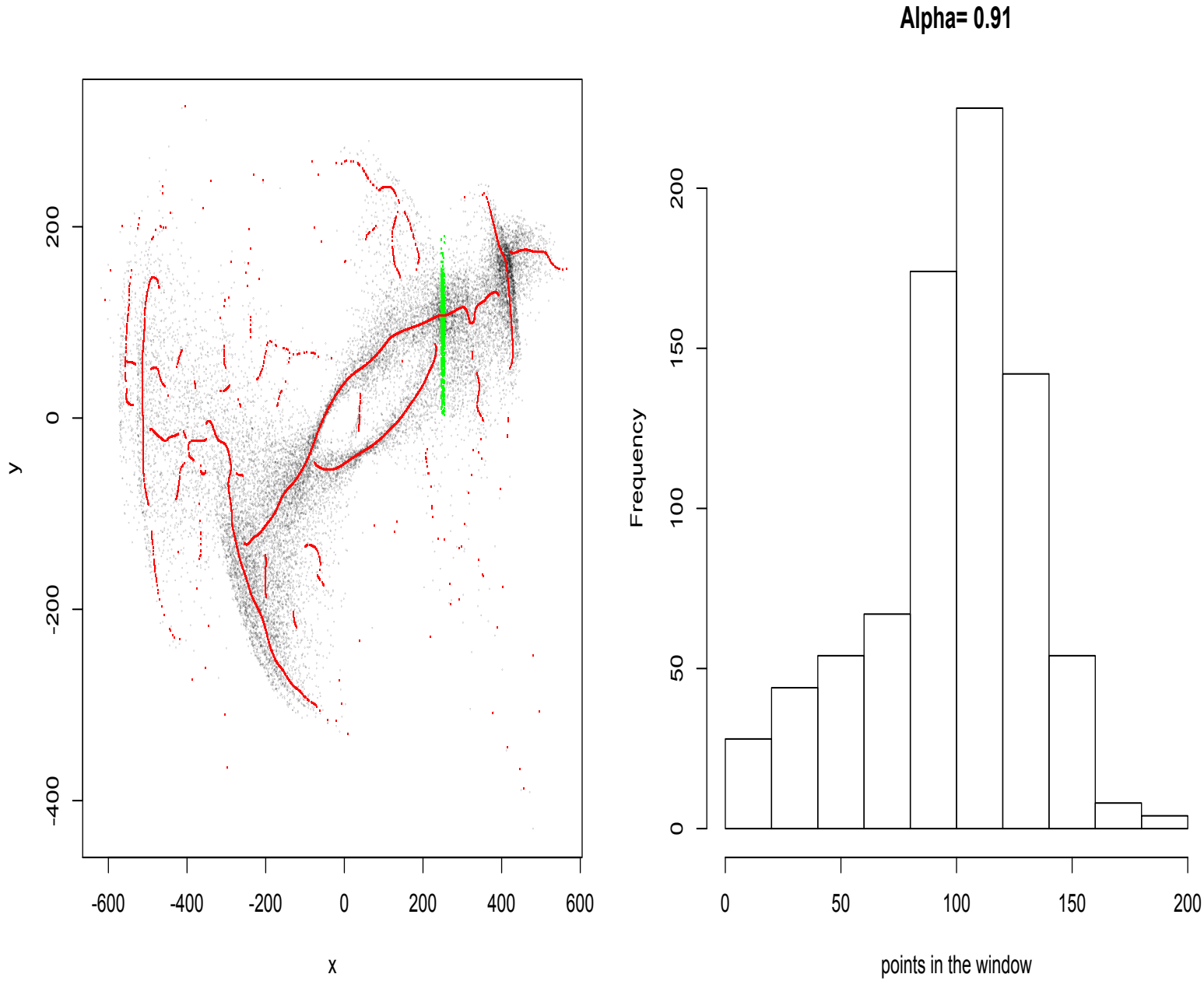


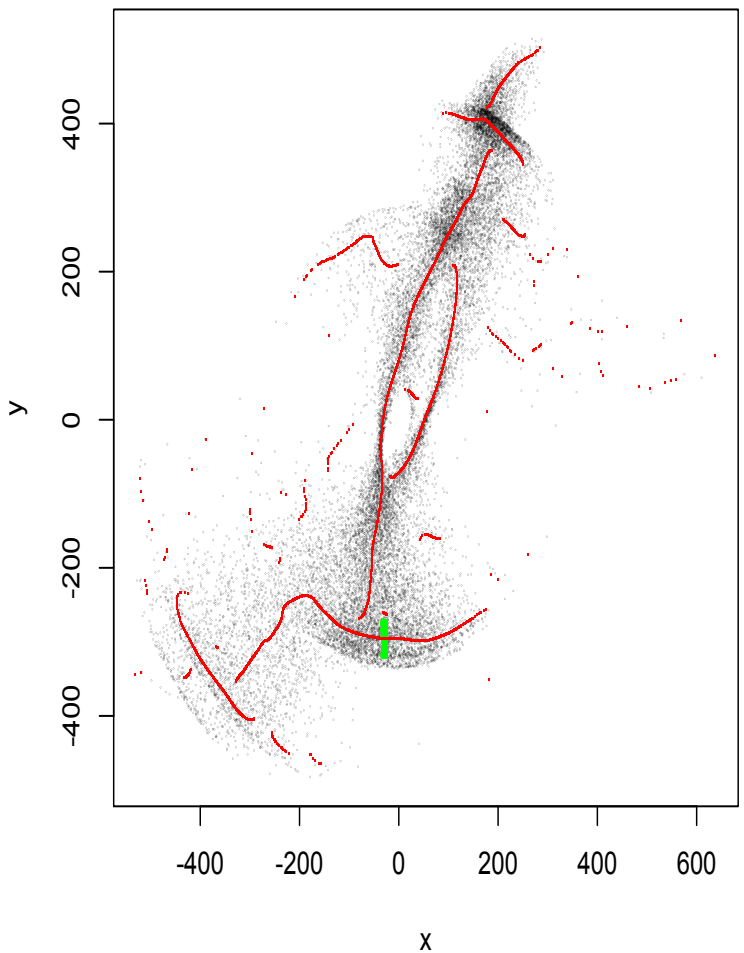


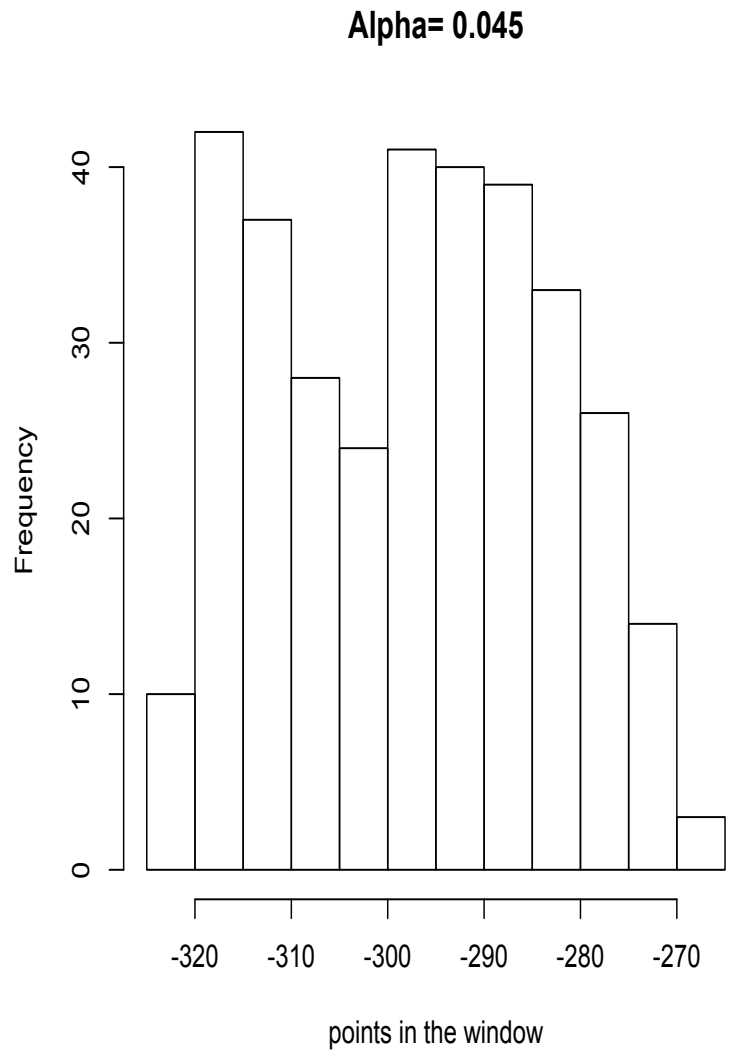
- Fix a *window* perpendicular to the ridge at a point (in the ridge)
- *Streams* are classified by looking for a symmetric *peak* in the window



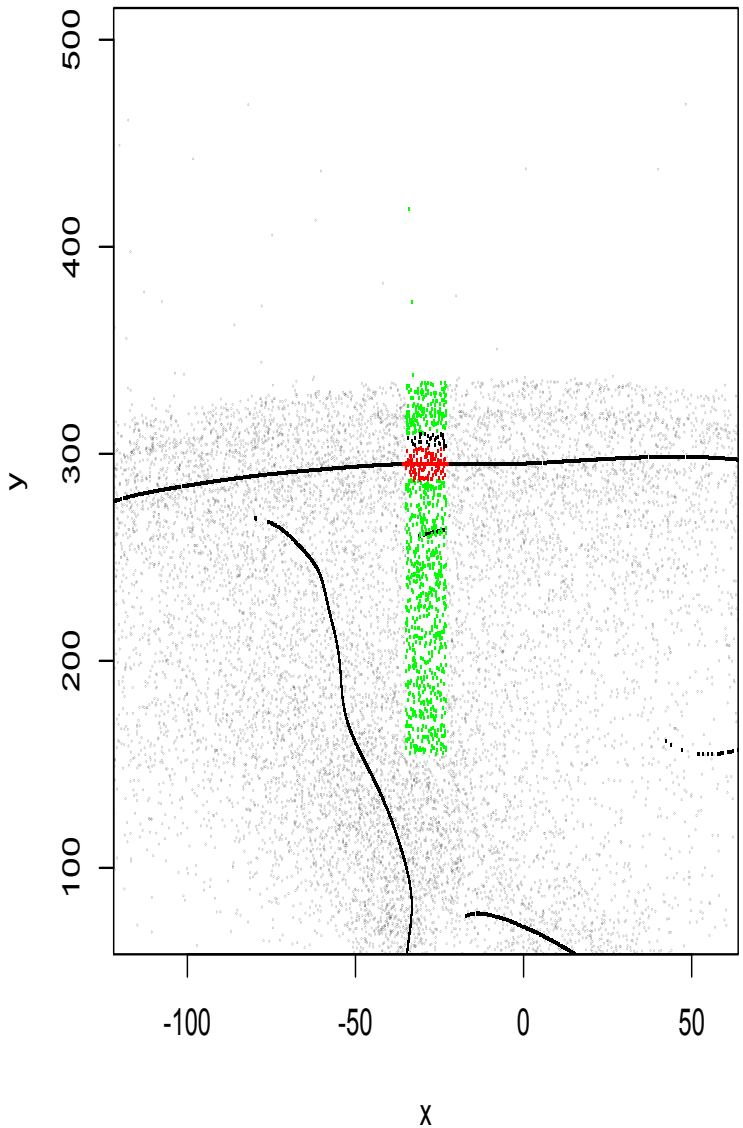


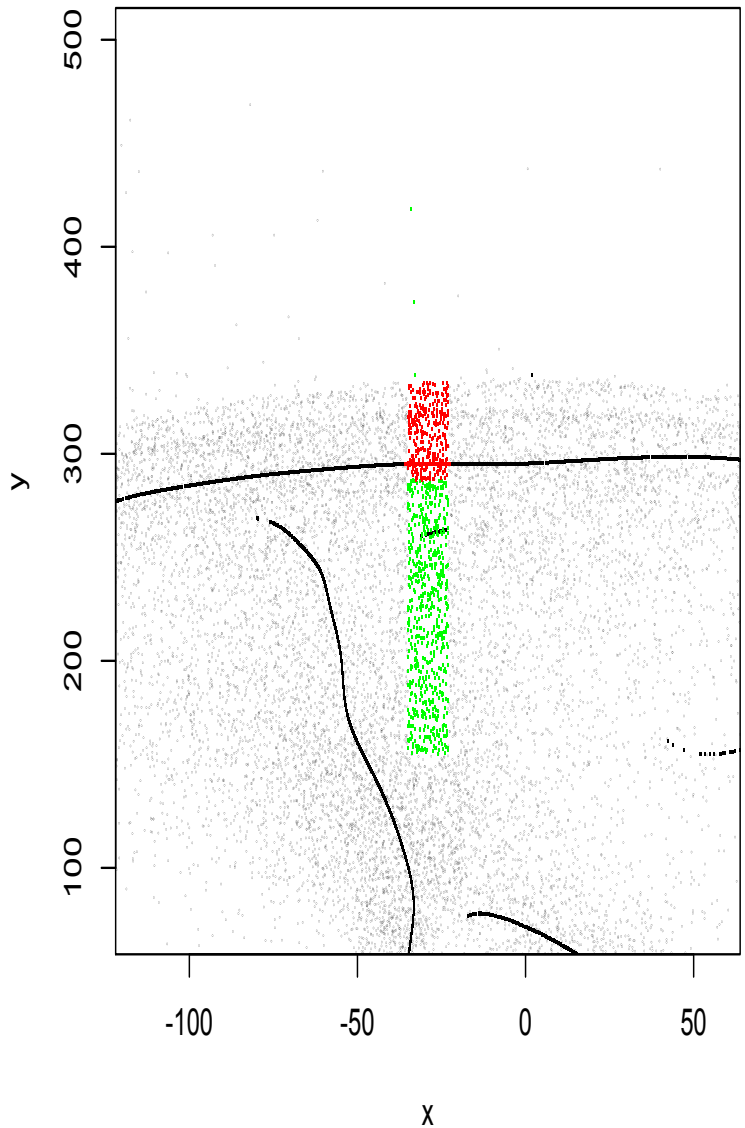

Alpha= 0.91
y
x
-600
-400
-200
0
200
400
600
-400
-200
0
200
Frequency
points in the window
0
50
100
150
200

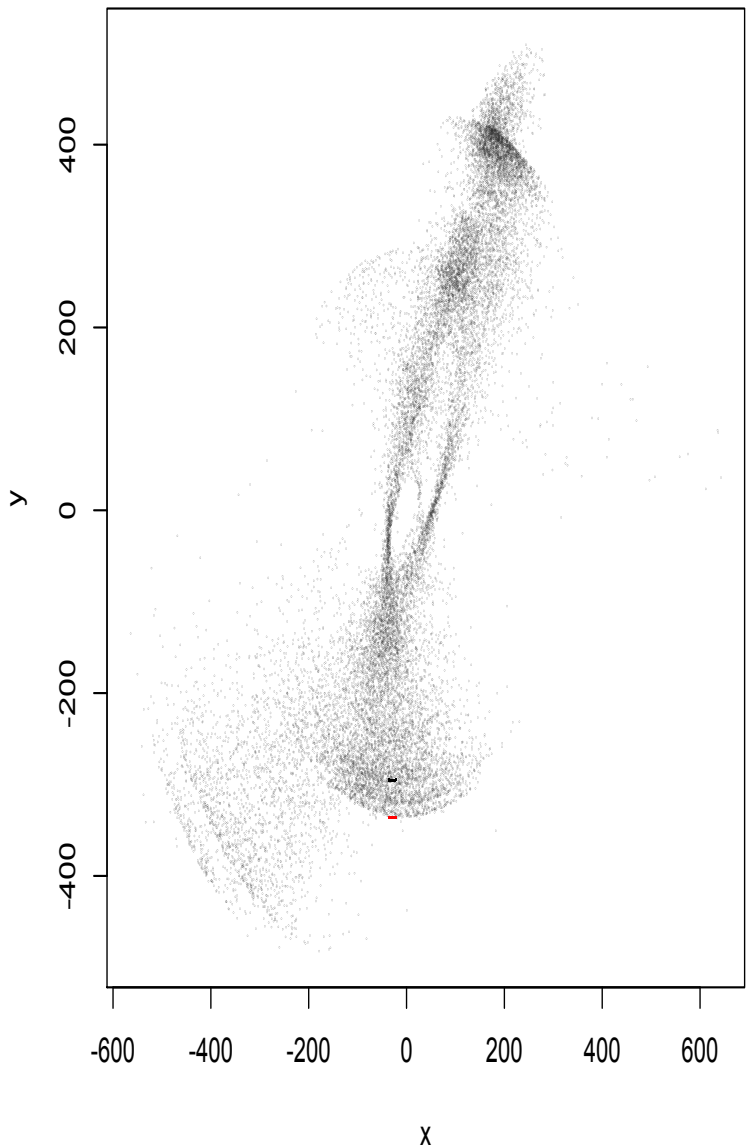


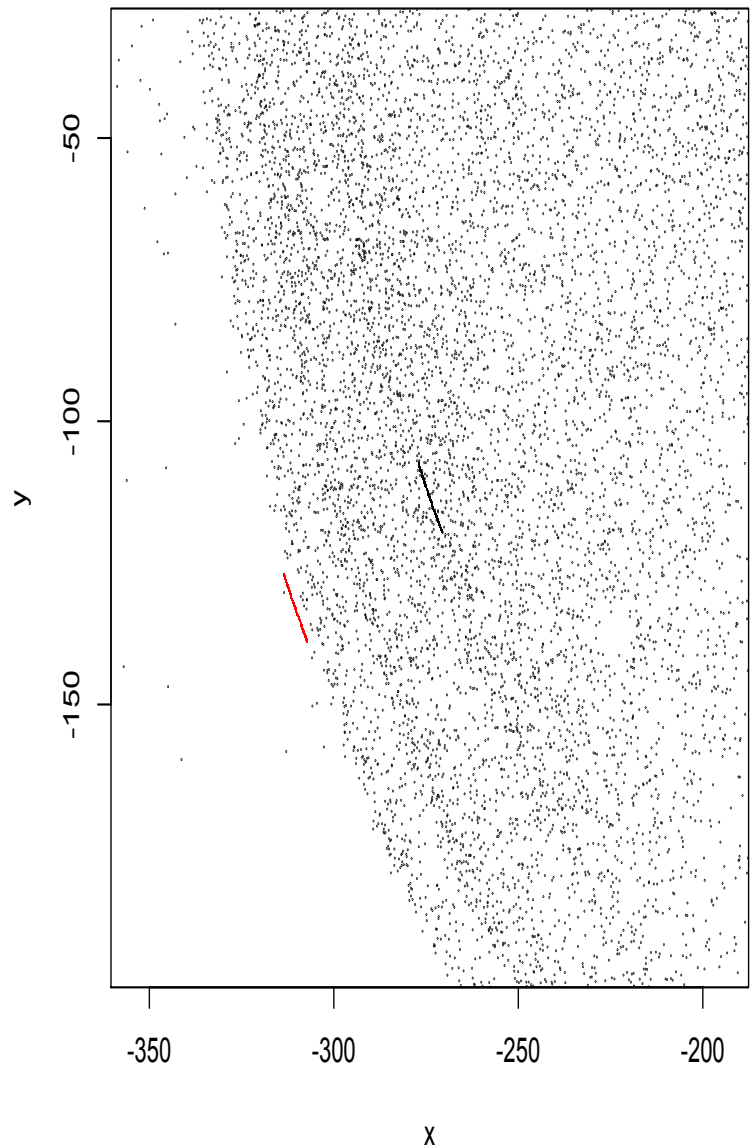


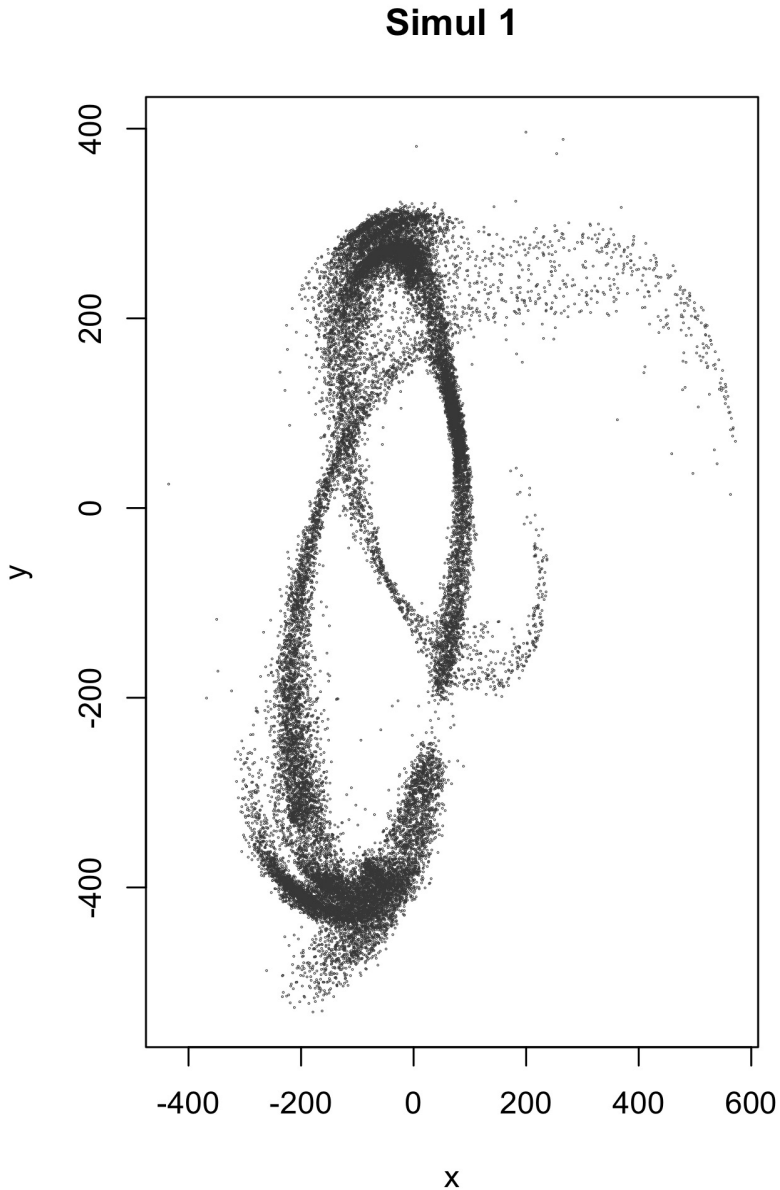
If a *peak* is not detected, then we try to look for a *shell* nearby

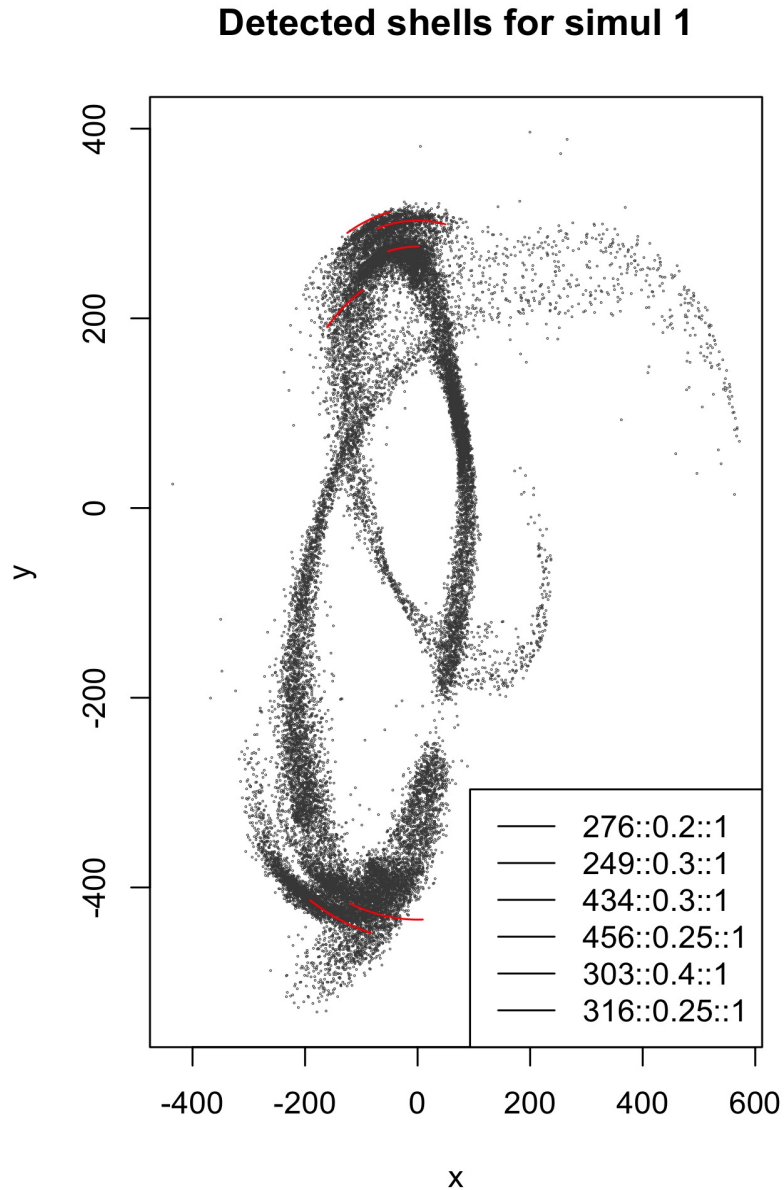

500
400
300
200
100
y
-100
-50
0
50
x

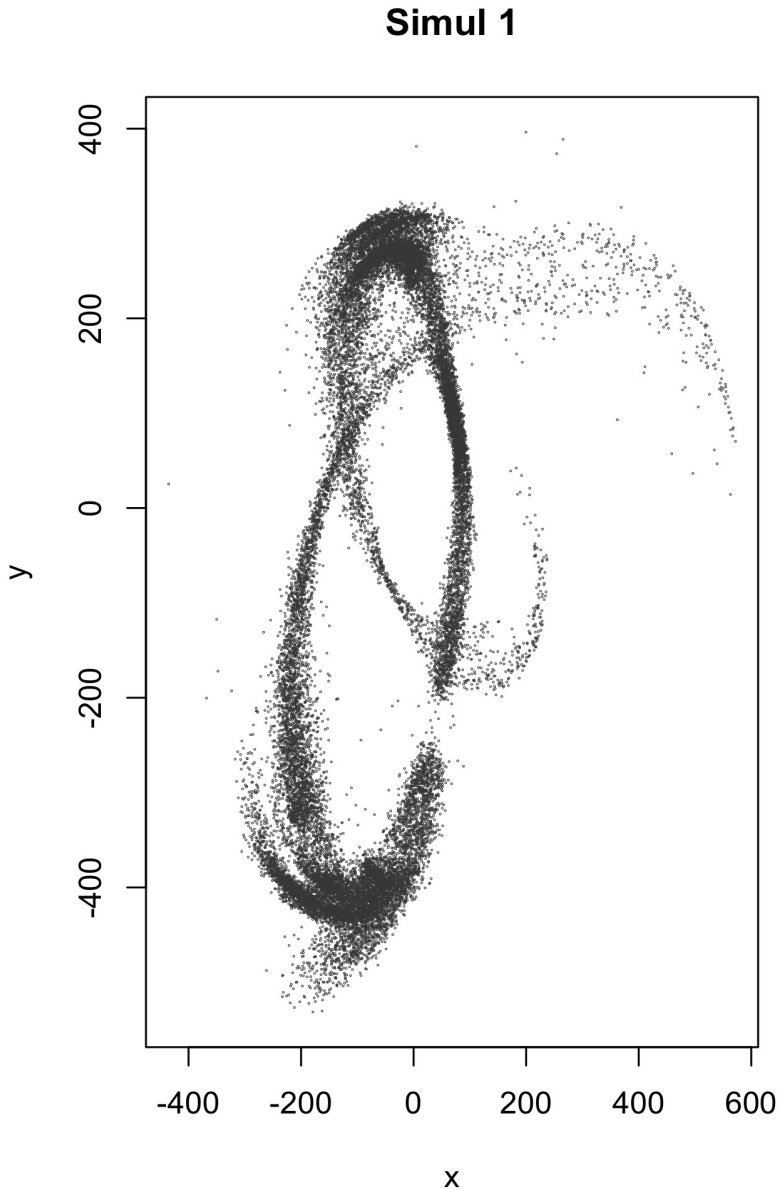

500
400
300
200
100
y
-100
-50
0
50
x

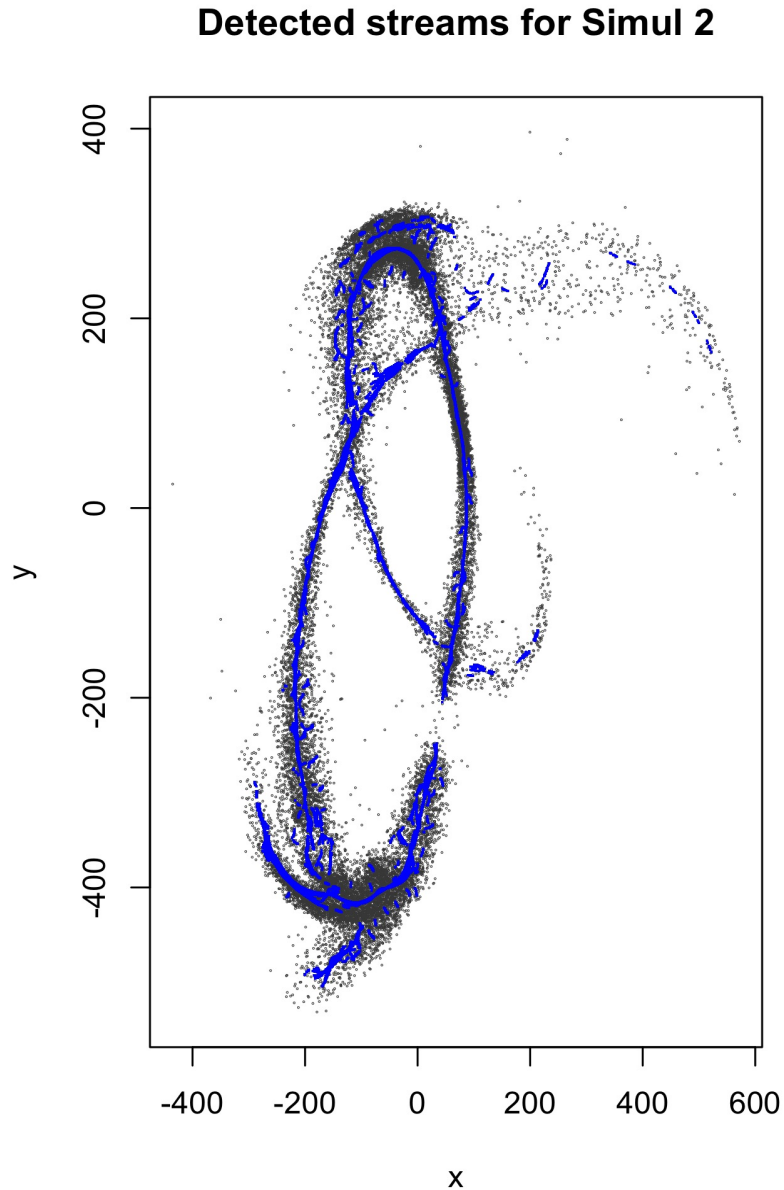

y
400
200
0
-200
-400
-600
-400
-200
0
200
400
600
x

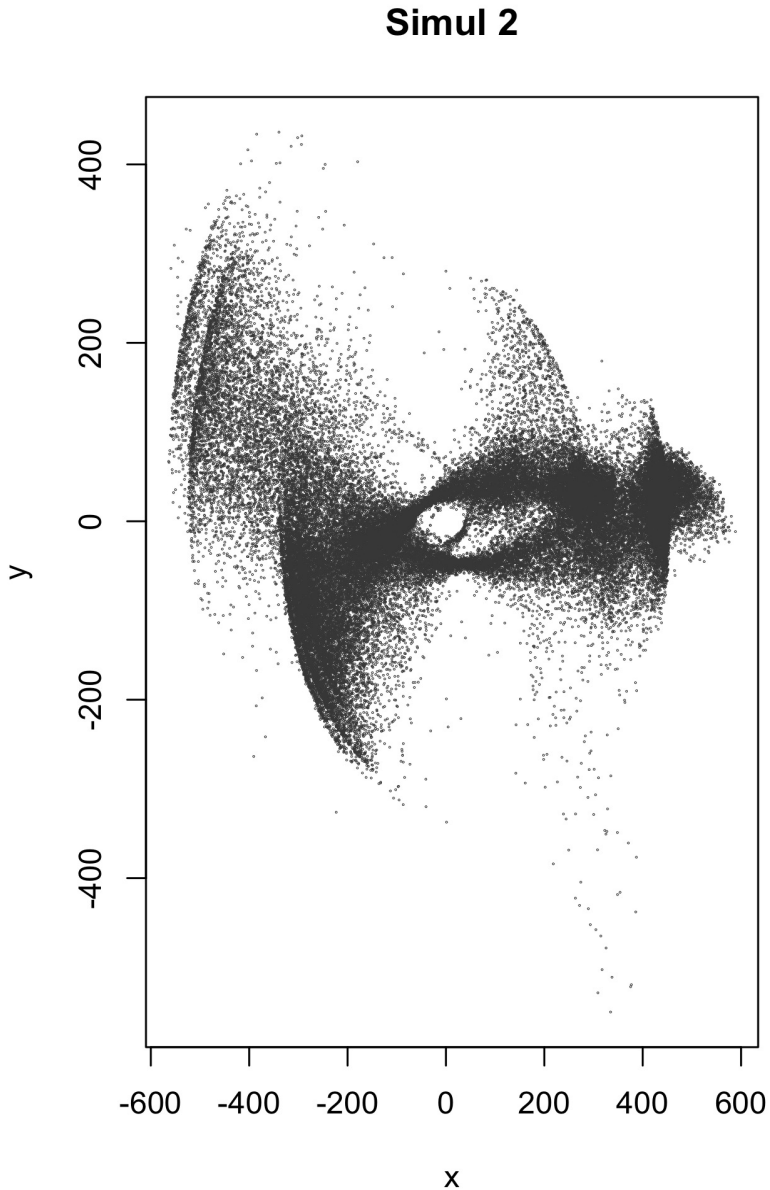

y
-50
-100
-150
-350
-300
-250
-200
x

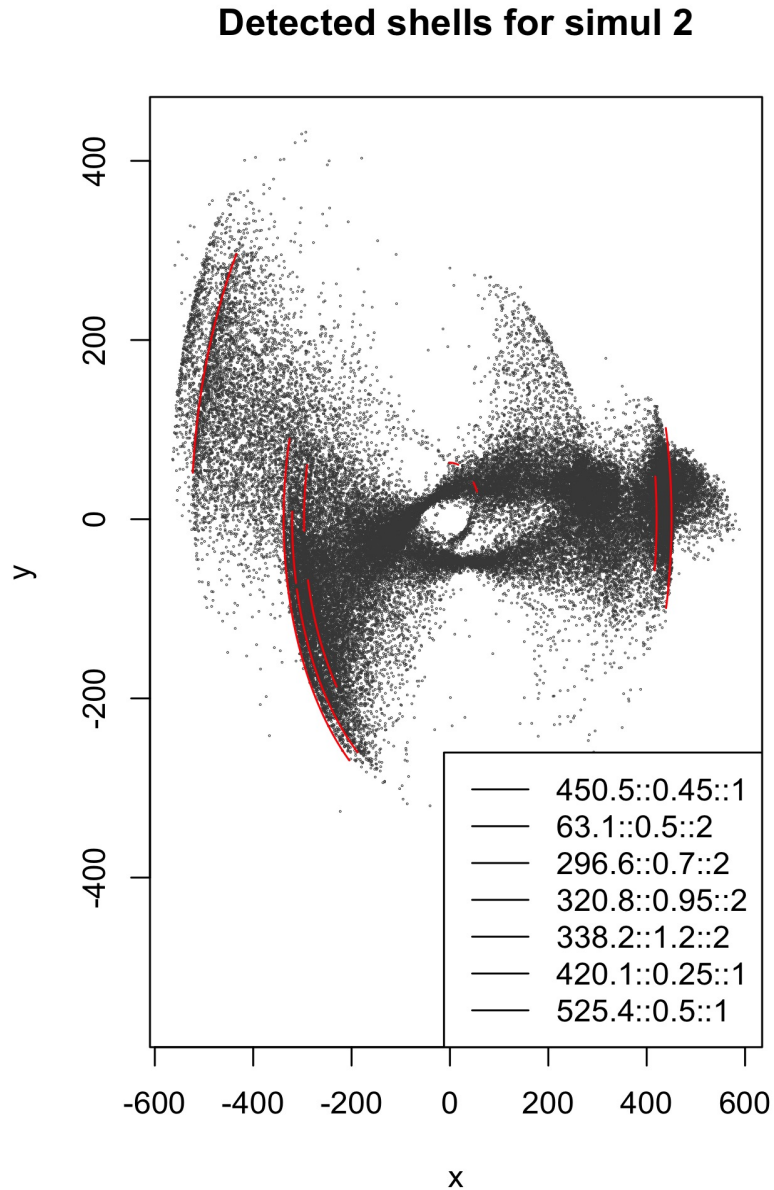

Simul 1
400
200
0
-200
-400
y
-400
-200
0
200
400
600
x

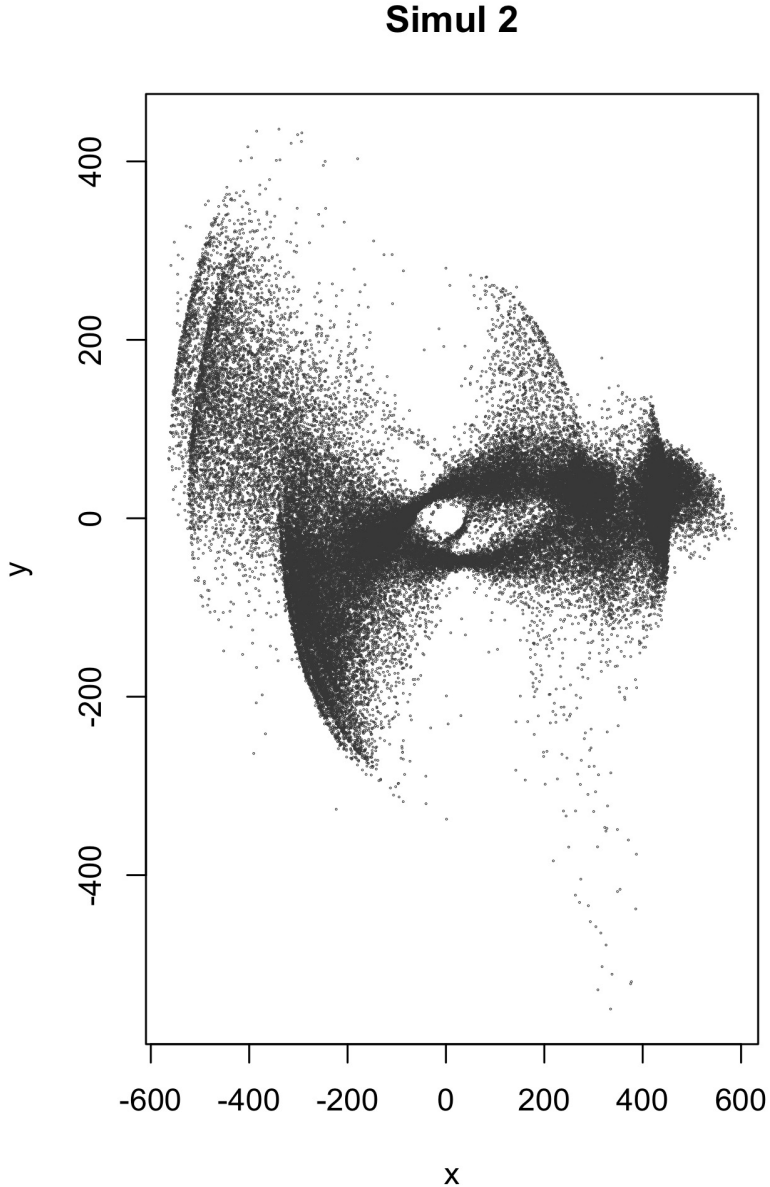

Detected shells for simul 1
400
200
0
-200
-400
y
-400
-200
0
200
400
600
x
276::0.2::1
249::0.3::1
434::0.3::1
456::0.25::1
303::0.4::1
316::0.25::1

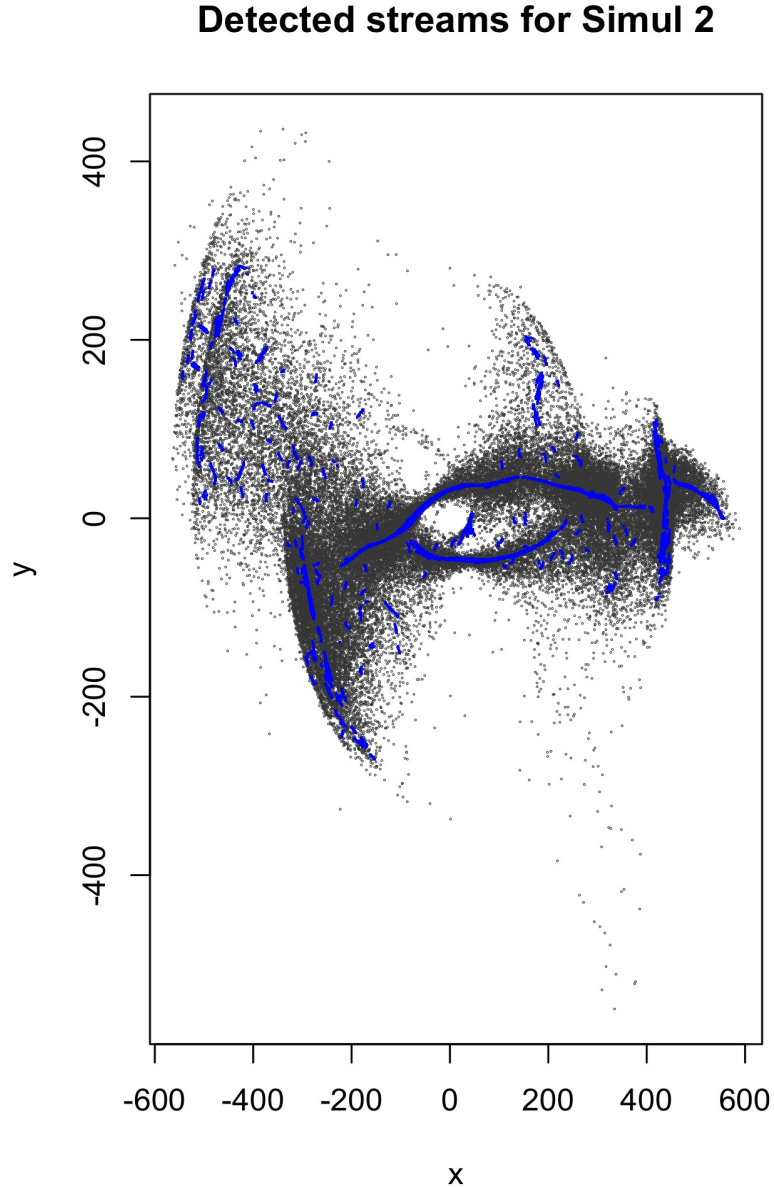

Simul 1
400
200
0
-200
-400
y
-400
-200
0
200
400
600
x


Detected streams for Simul 2
400
200
0
-200
-400
y
-400
-200
0
200
400
600
x

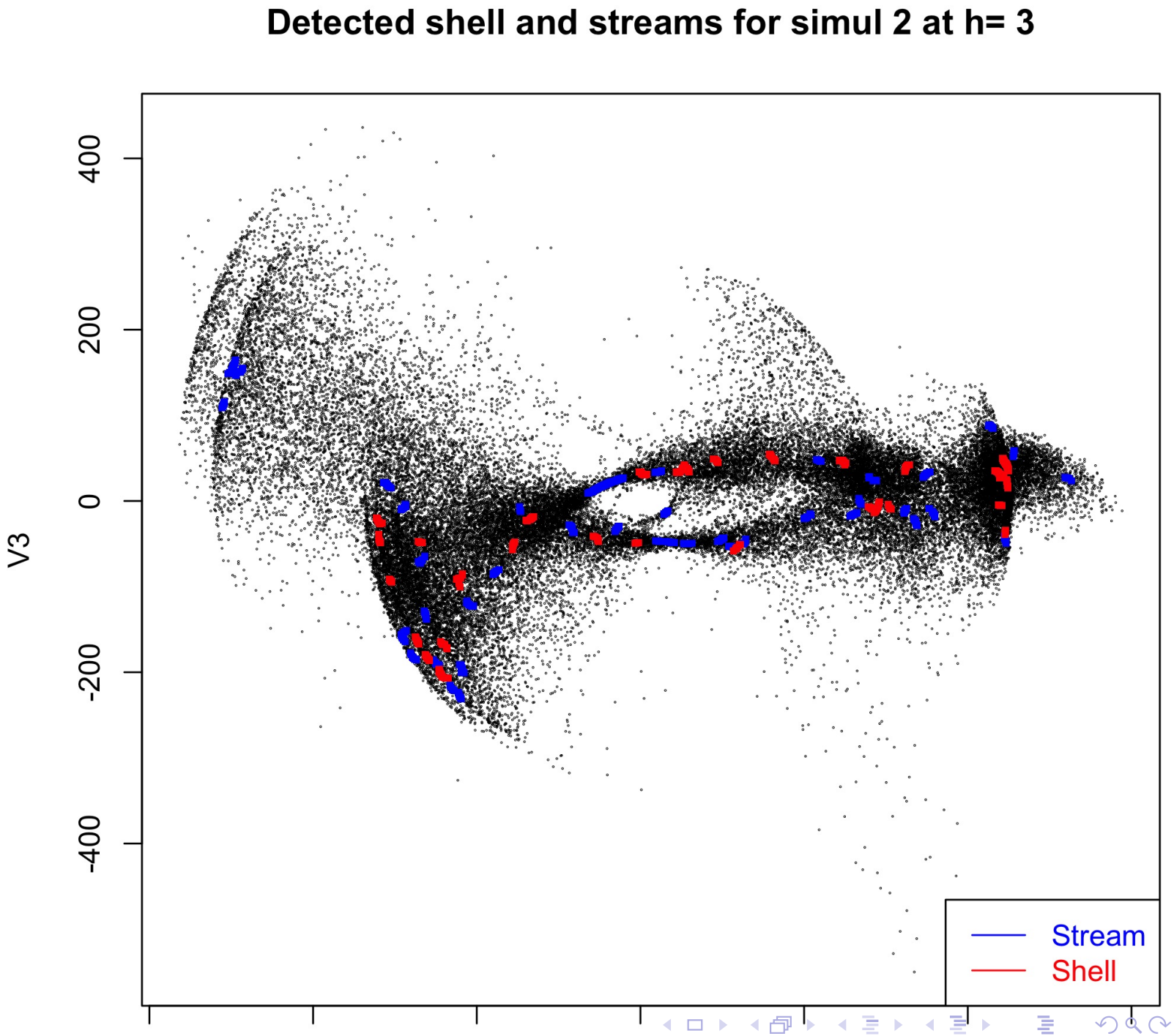

Simul 2
y
x
400
200
0
-200
-400
-600
-400
-200
0
200
400
600

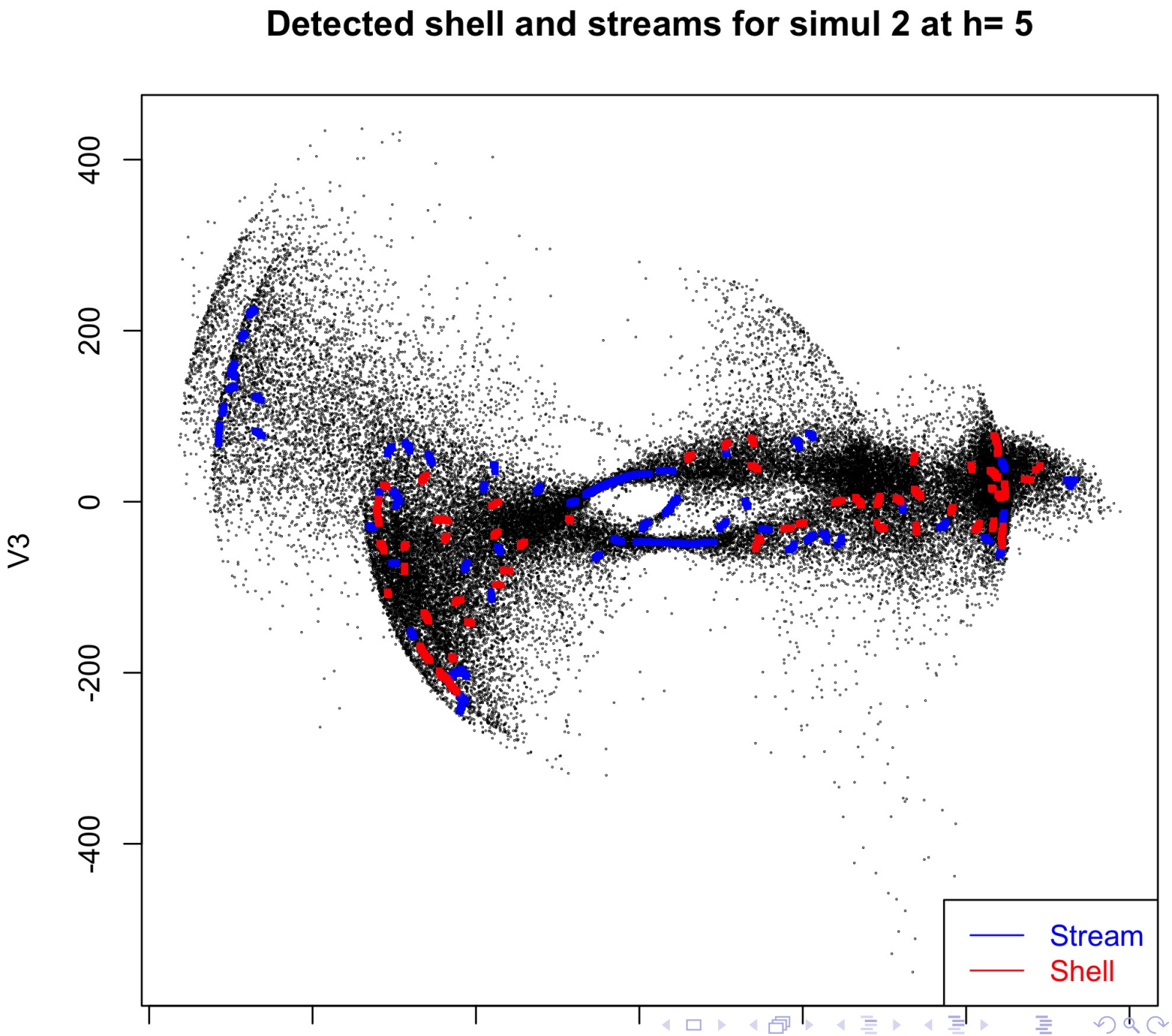

Detected shells for simul 2
y
x
400
200
0
-200
-400
-600
-400
-200
0
200
400
600
450.5::0.45::1
63.1::0.5::2
296.6::0.7::2
320.8::0.95::2
338.2::1.2::2
420.1::0.25::1
525.4::0.5::1

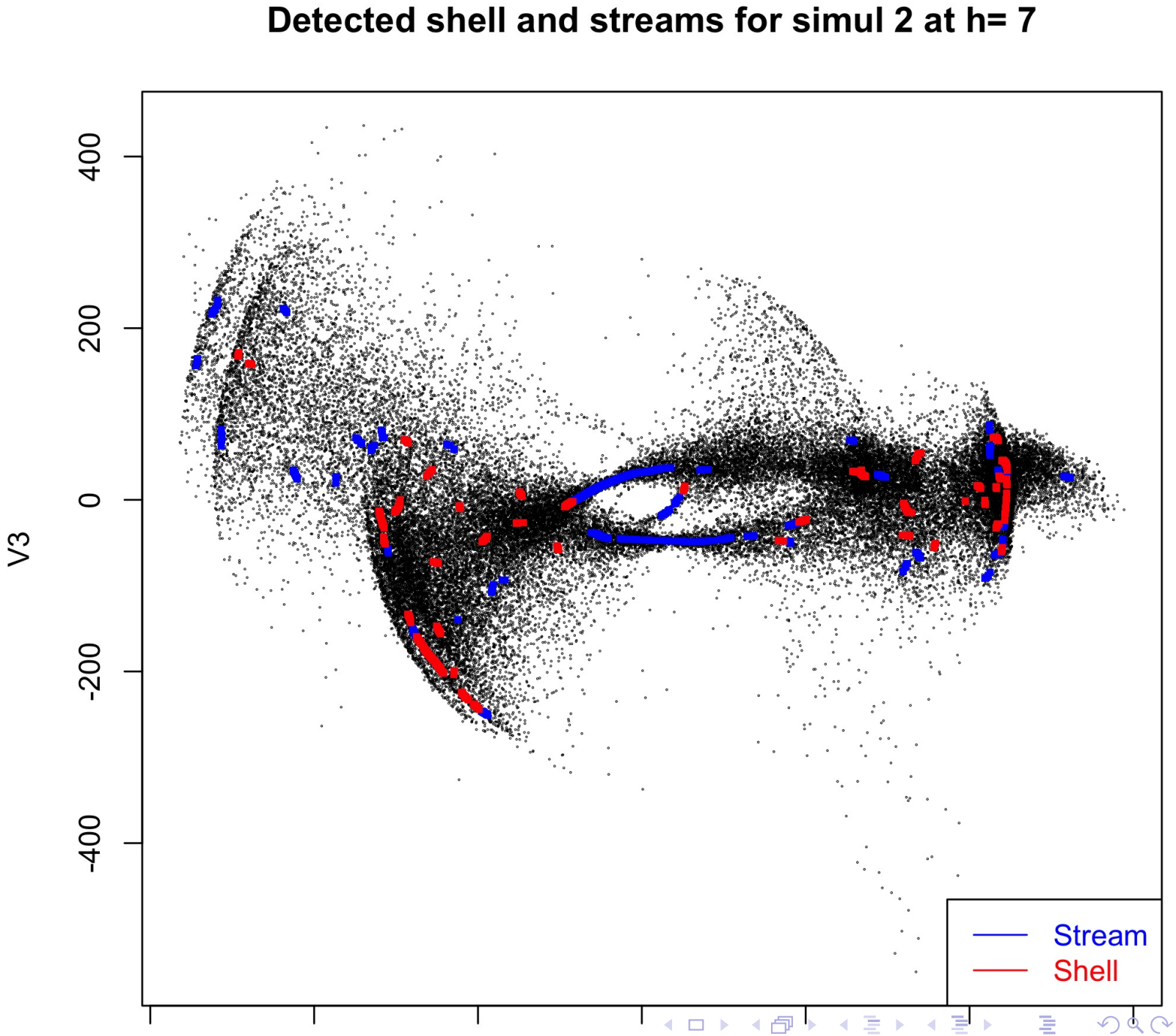

Simul 2
y
x
400
200
0
-200
-400
-600
-400
-200
0
200
400
600

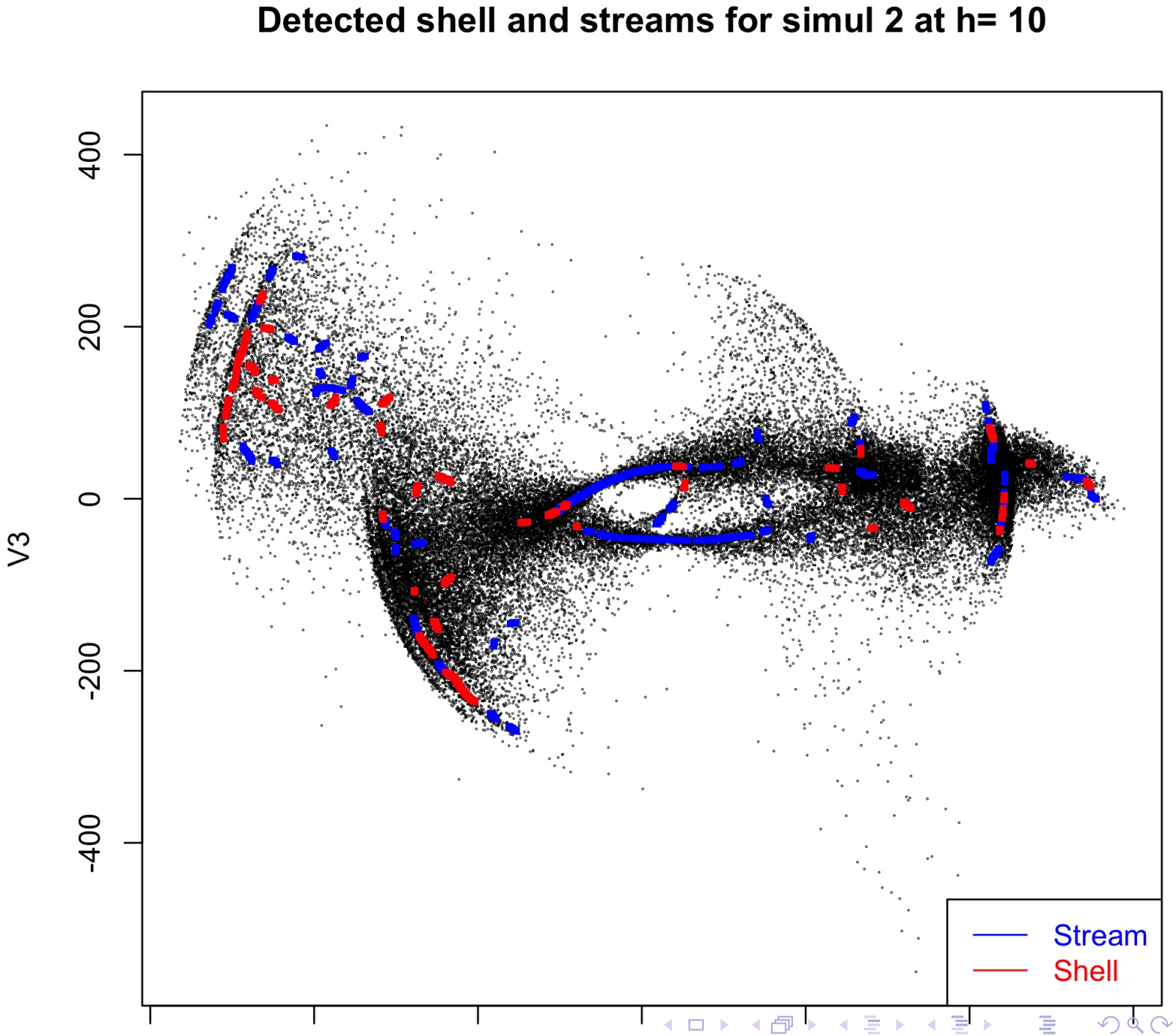

Detected streams for Simul 2
y
x
400
200
0
-200
-400
-600
-400
-200
0
200
400
600

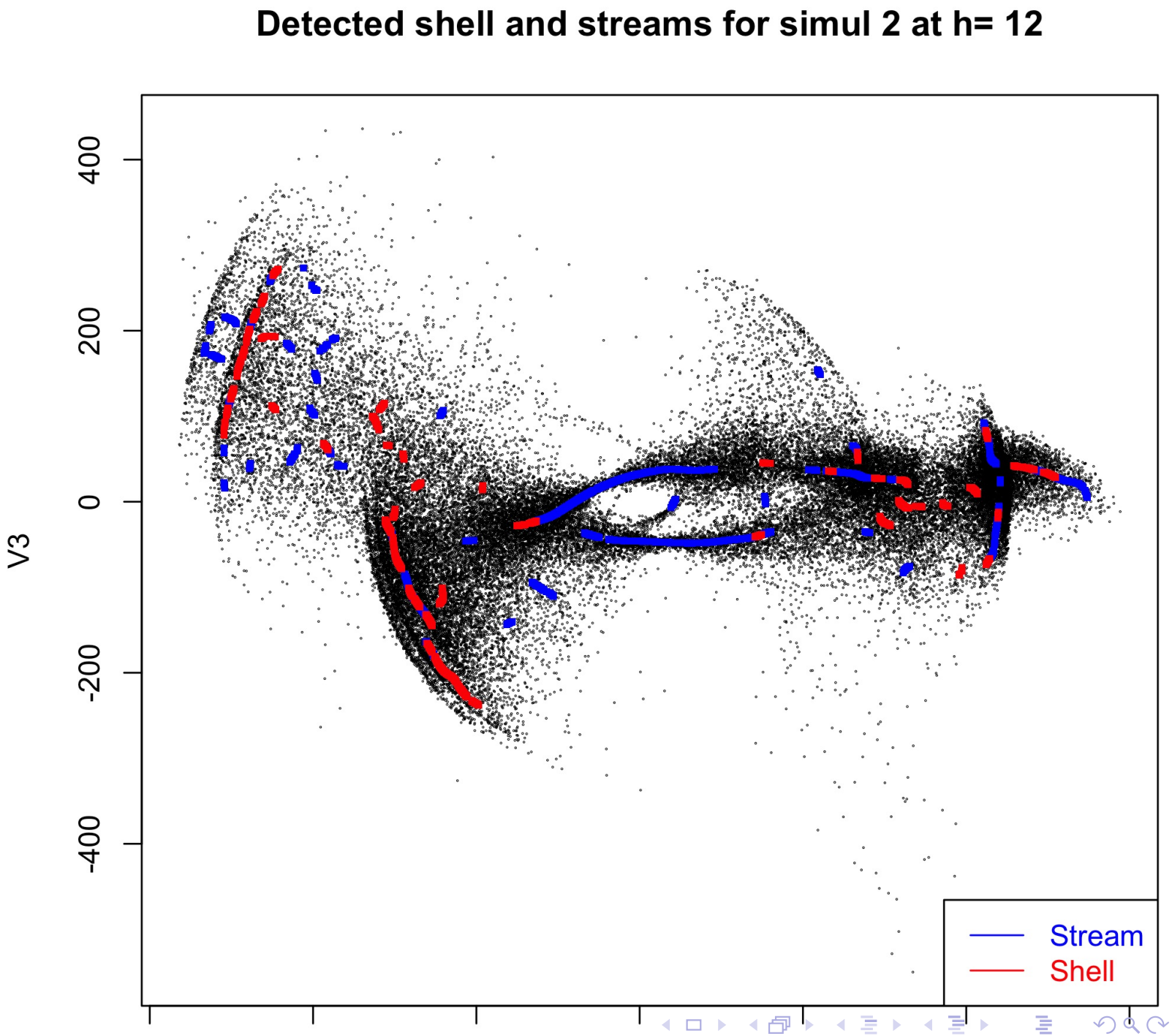
Multi-scale idea

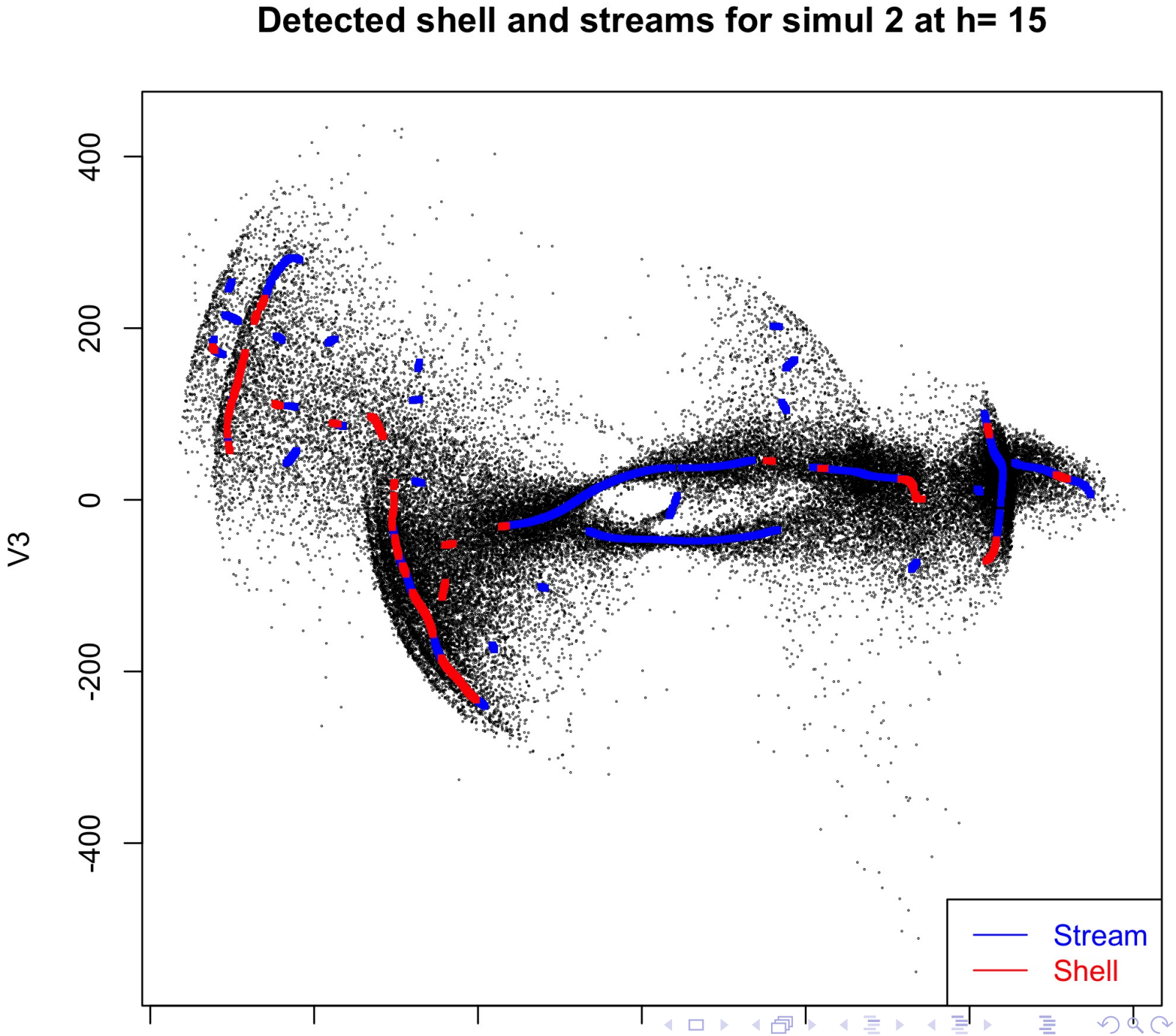

Detected shell and streams for simul 2 at h= 3
V3
400
200
0
-200
-400
Stream
Shell

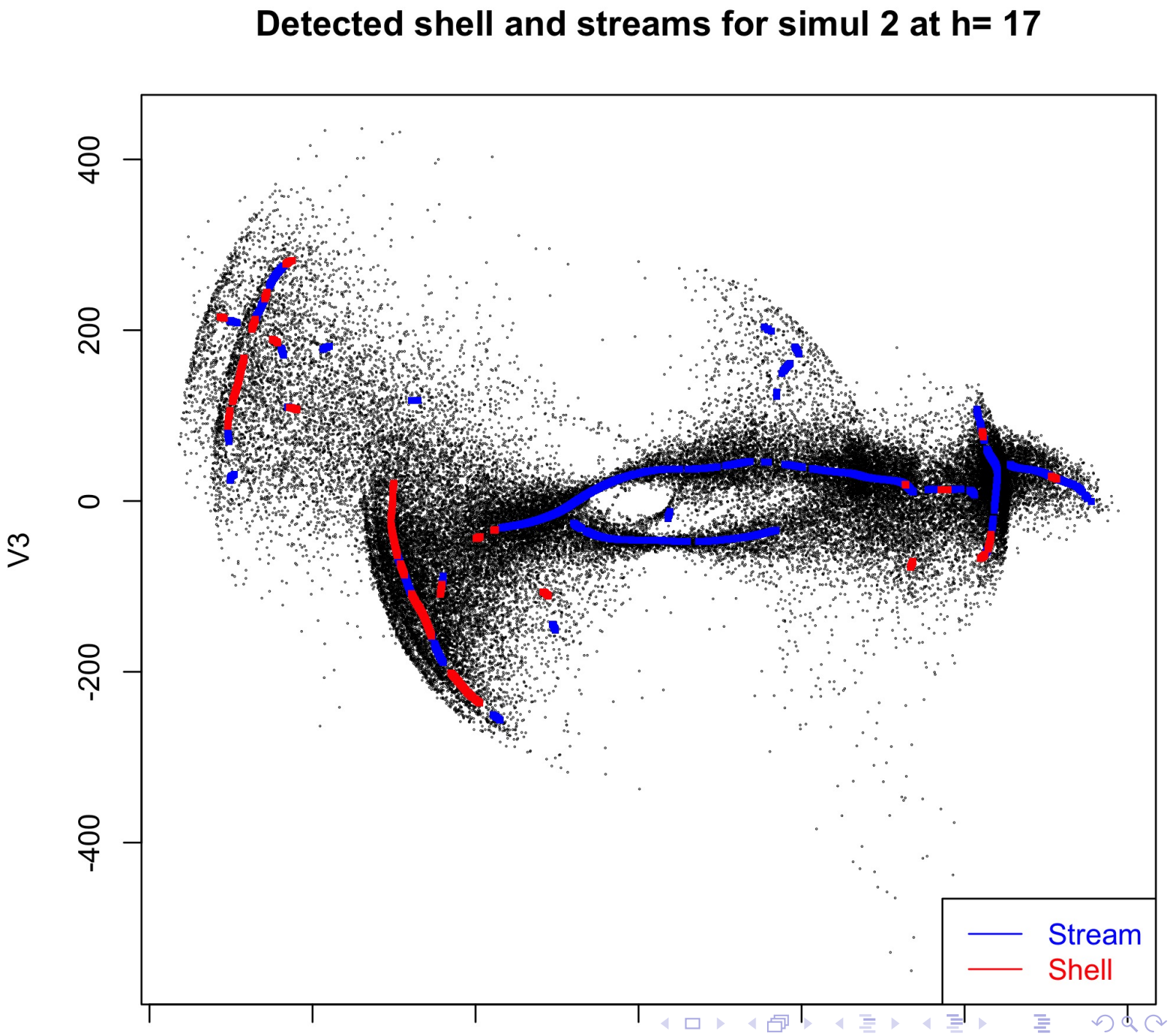

Detected shell and streams for simul 2 at h= 5
V3
400
200
0
-200
-400
Stream
Shell

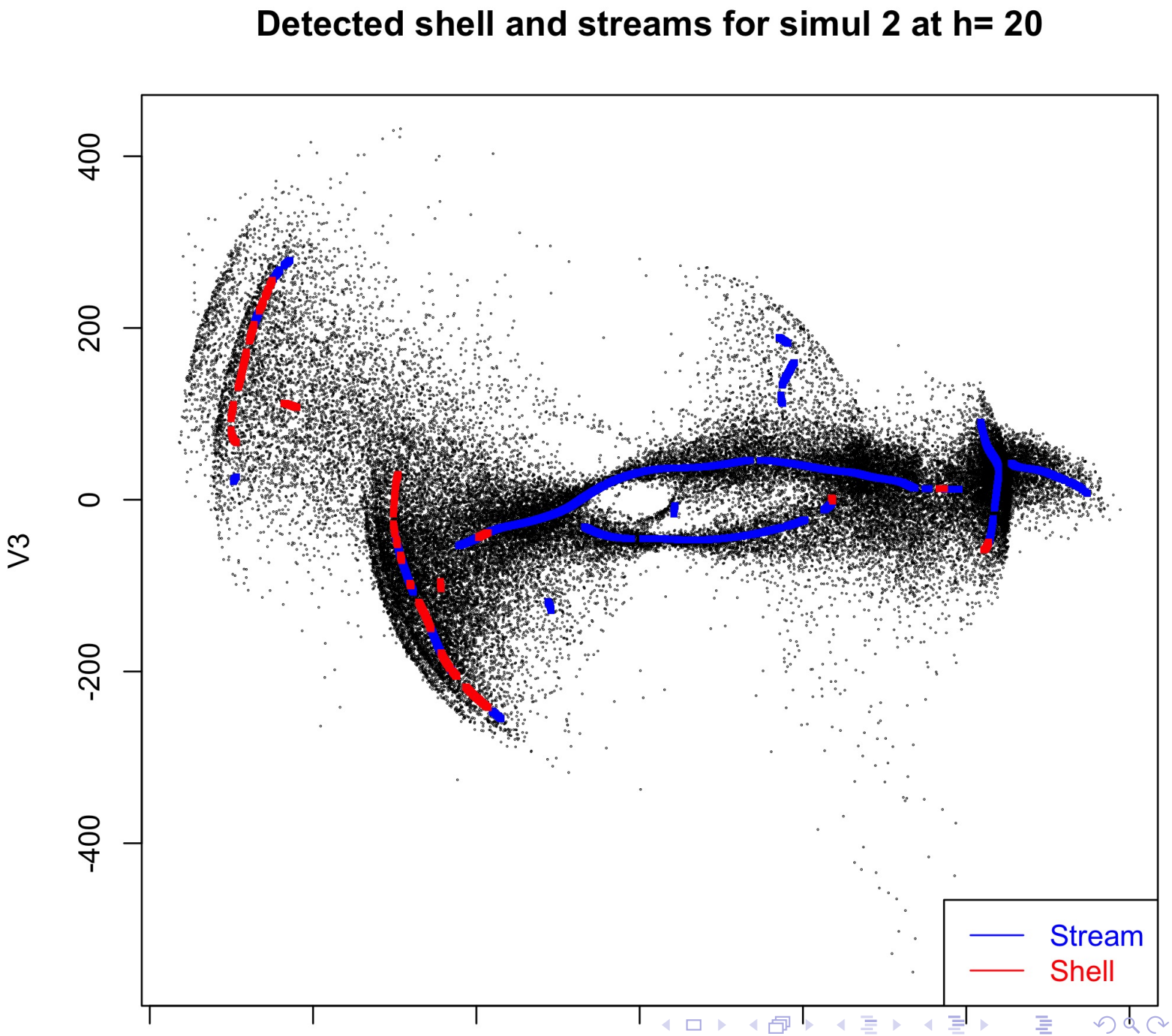

Detected shell and streams for simul 2 at h= 7
V3
400
200
0
-200
-400
Stream
Shell

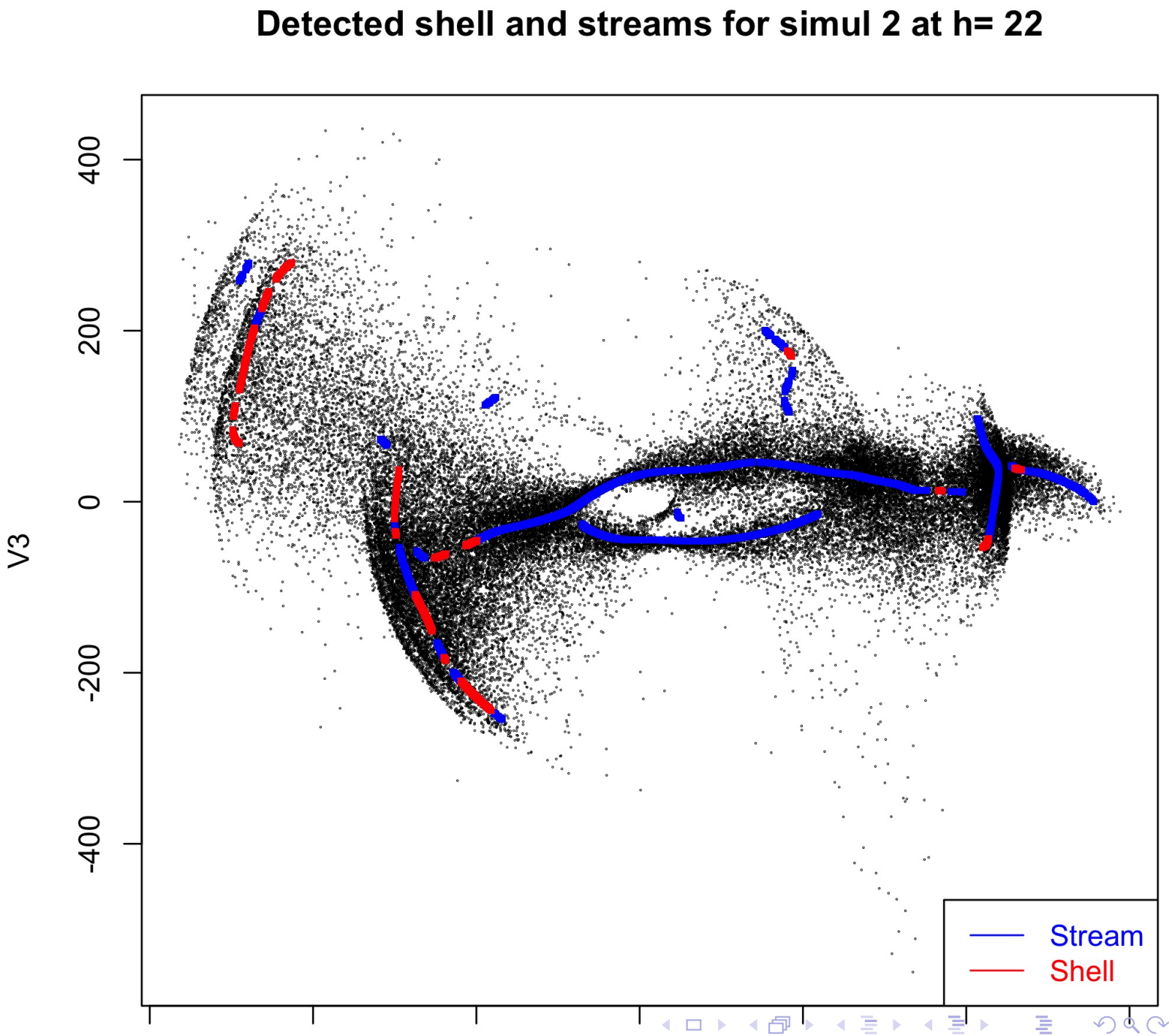

Detected shell and streams for simul 2 at h= 10
V3
400
200
0
-200
-400
Stream
Shell

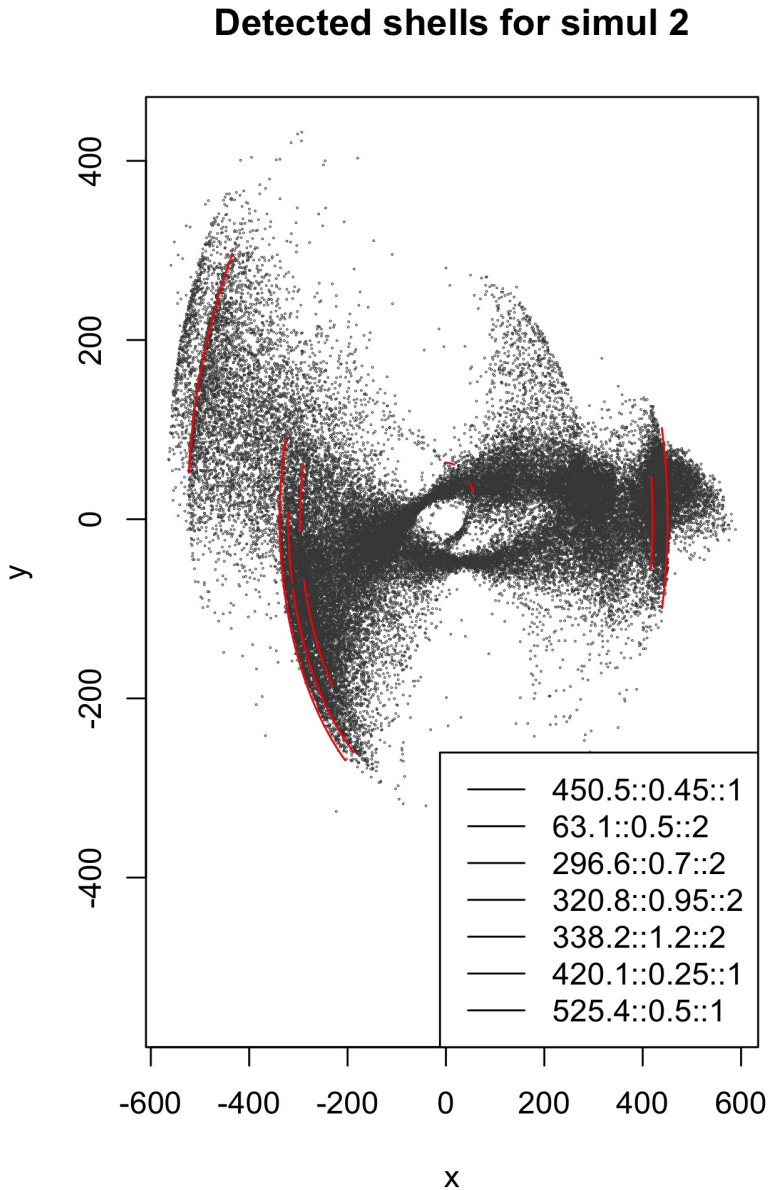

Detected shell and streams for simul 2 at h= 12
V3
400
200
0
-200
-400
Stream
Shell

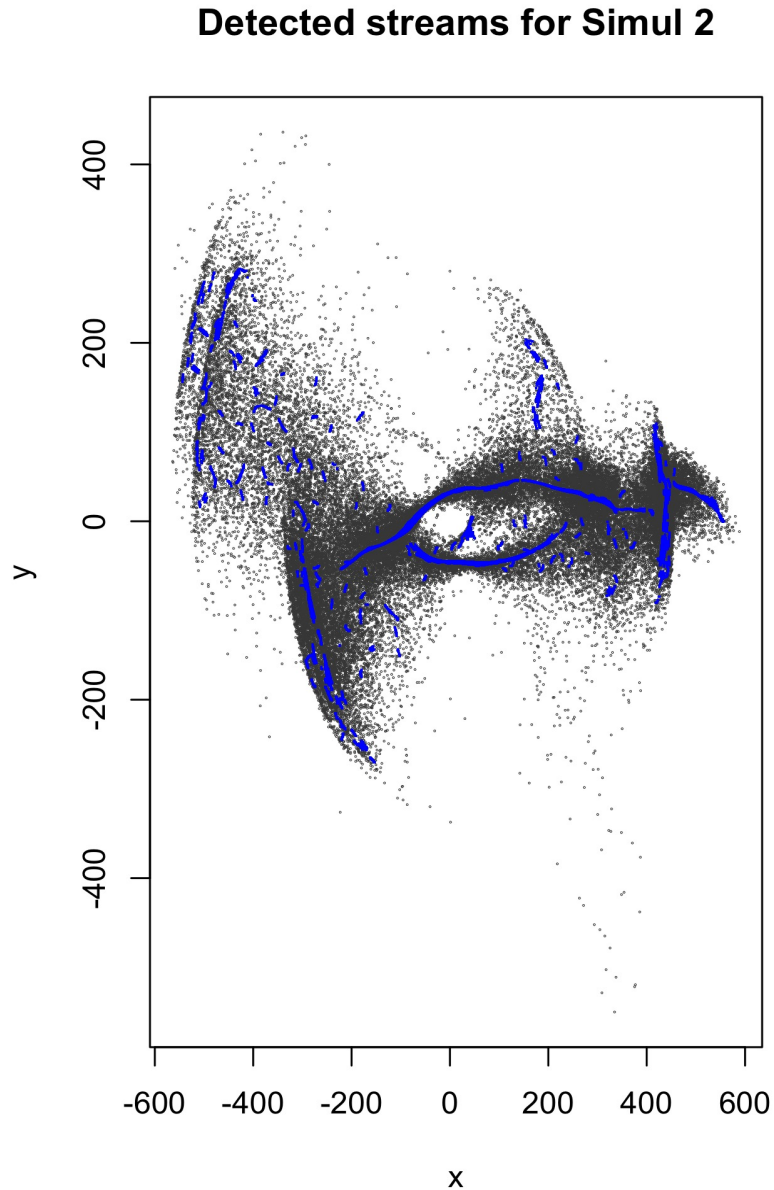

Detected shell and streams for simul 2 at h= 15
V3
400
200
0
-200
-400
Stream
Shell

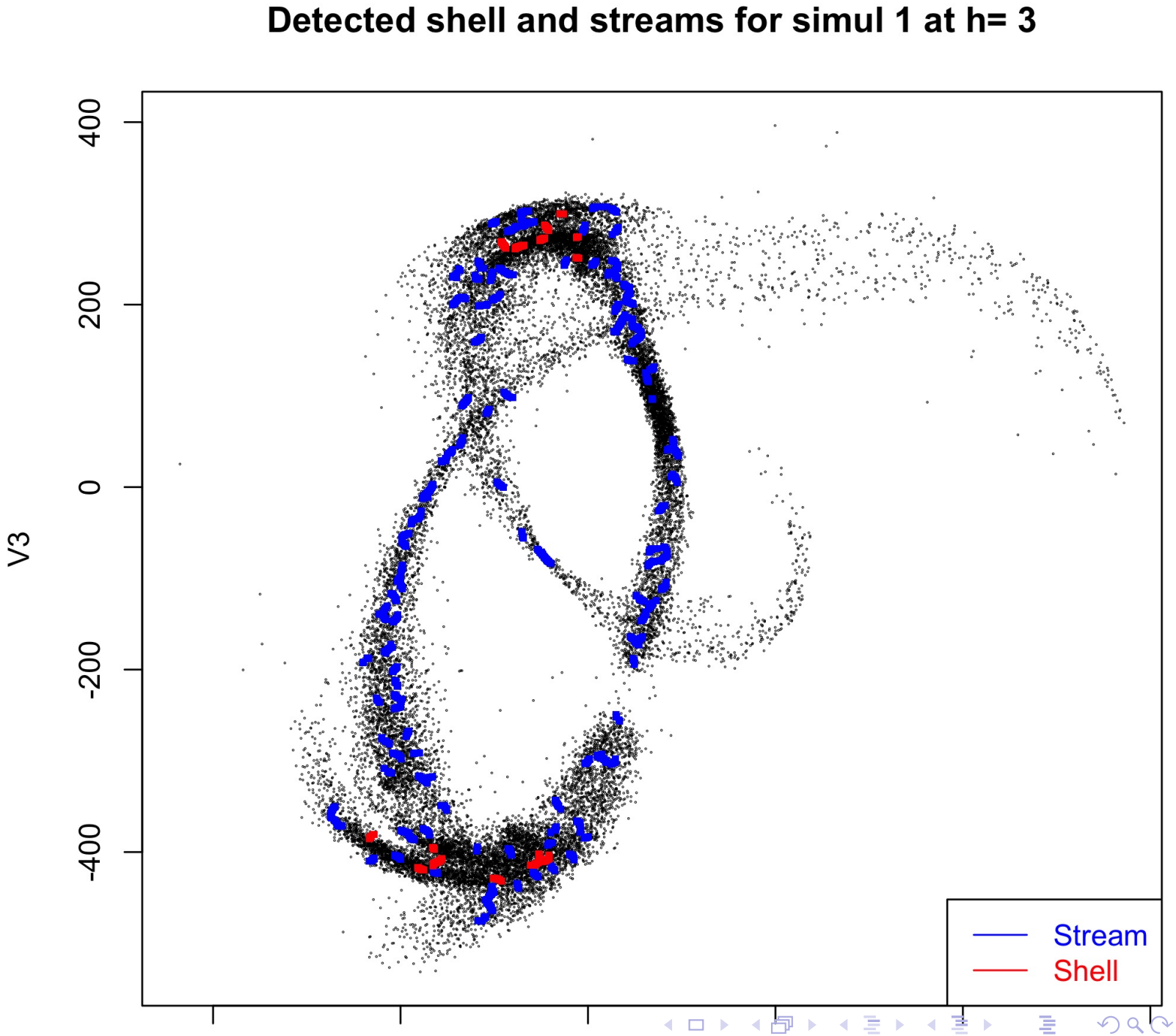

Detected shell and streams for simul 2 at h= 17
V3
400
200
0
-200
-400
Stream
Shell

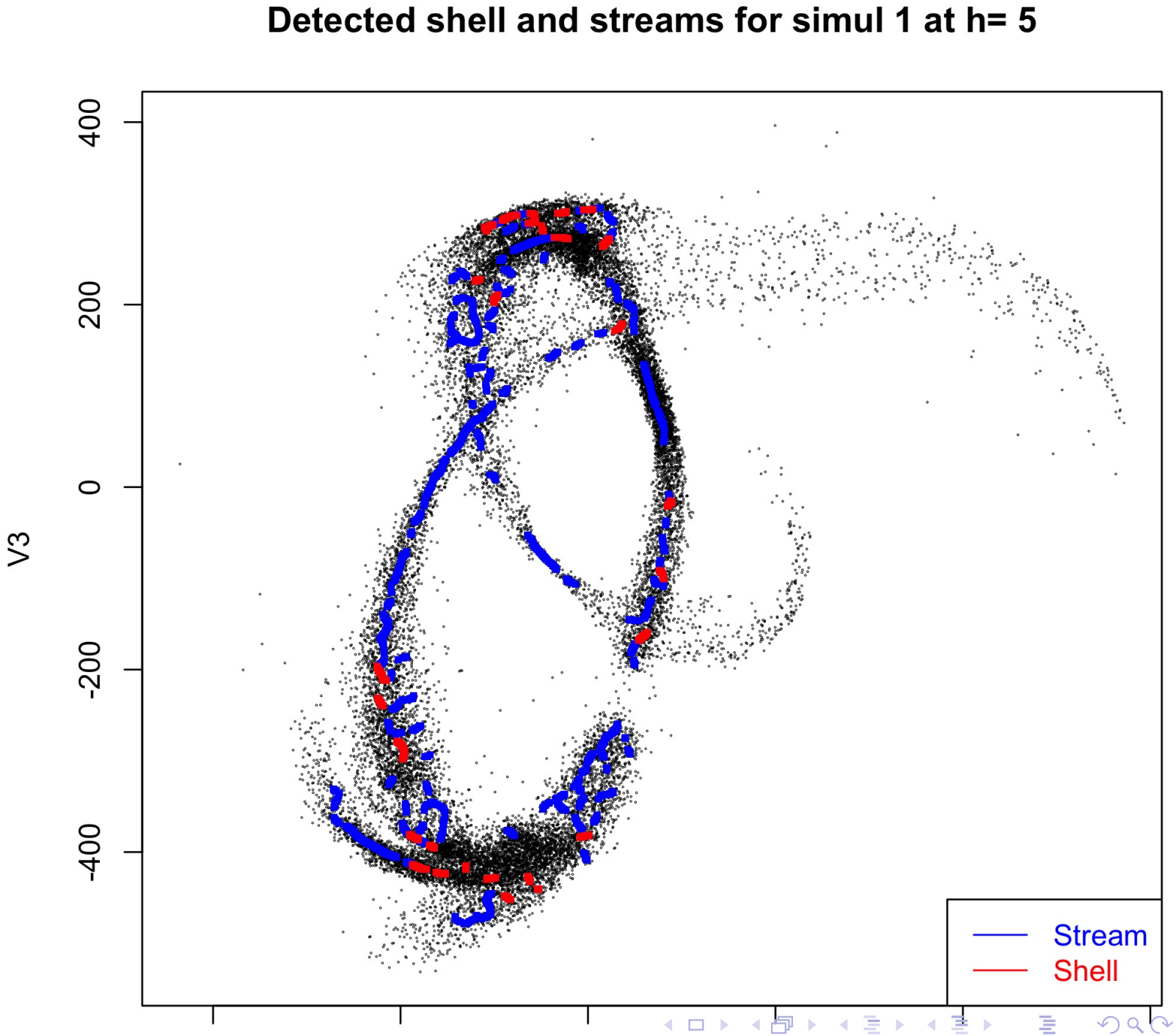

Detected shell and streams for simul 2 at h= 20
V3
400
200
0
-200
-400
Stream
Shell

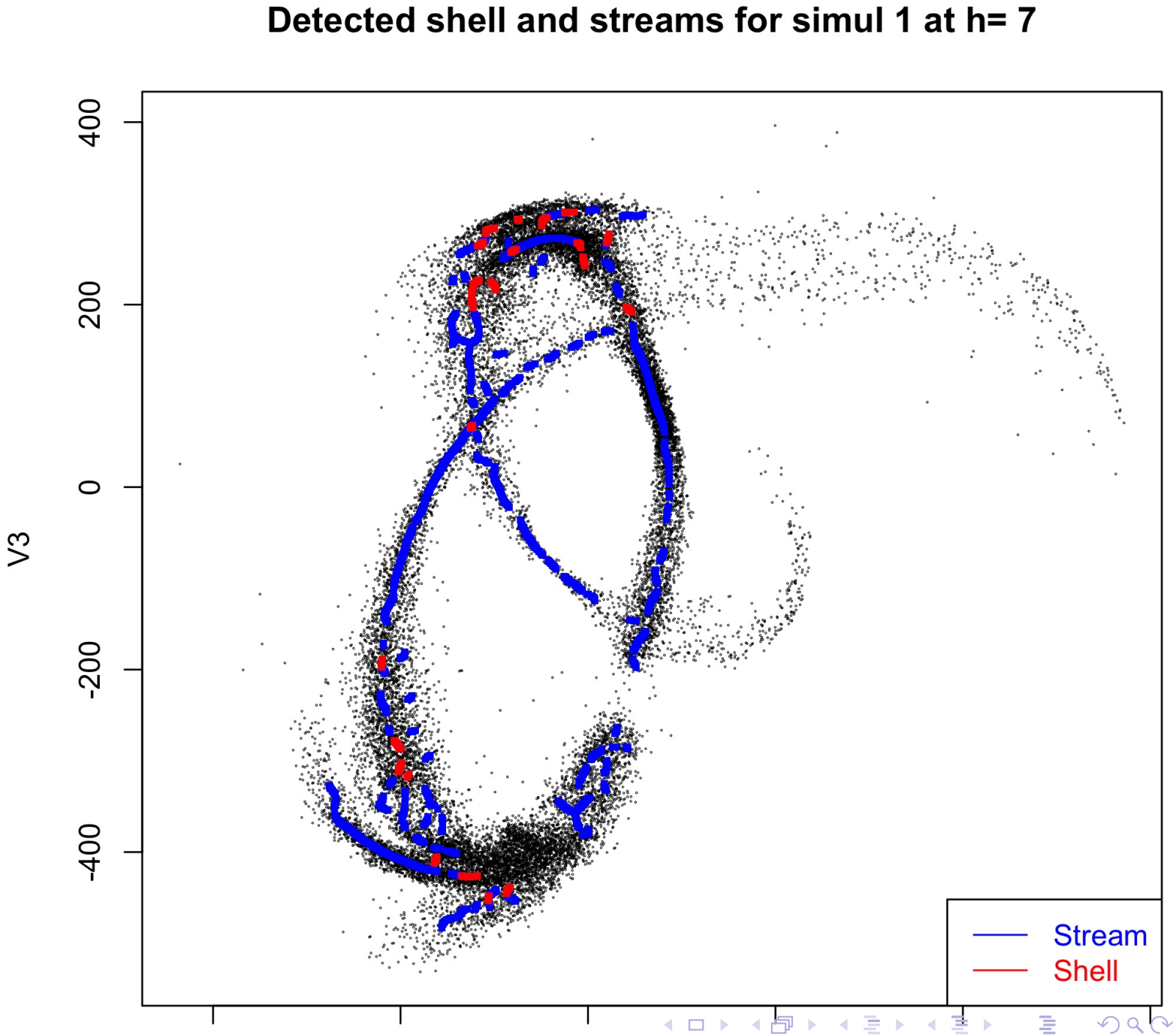

Detected shell and streams for simul 2 at h= 22
V3
400
200
0
-200
-400
Stream
Shell

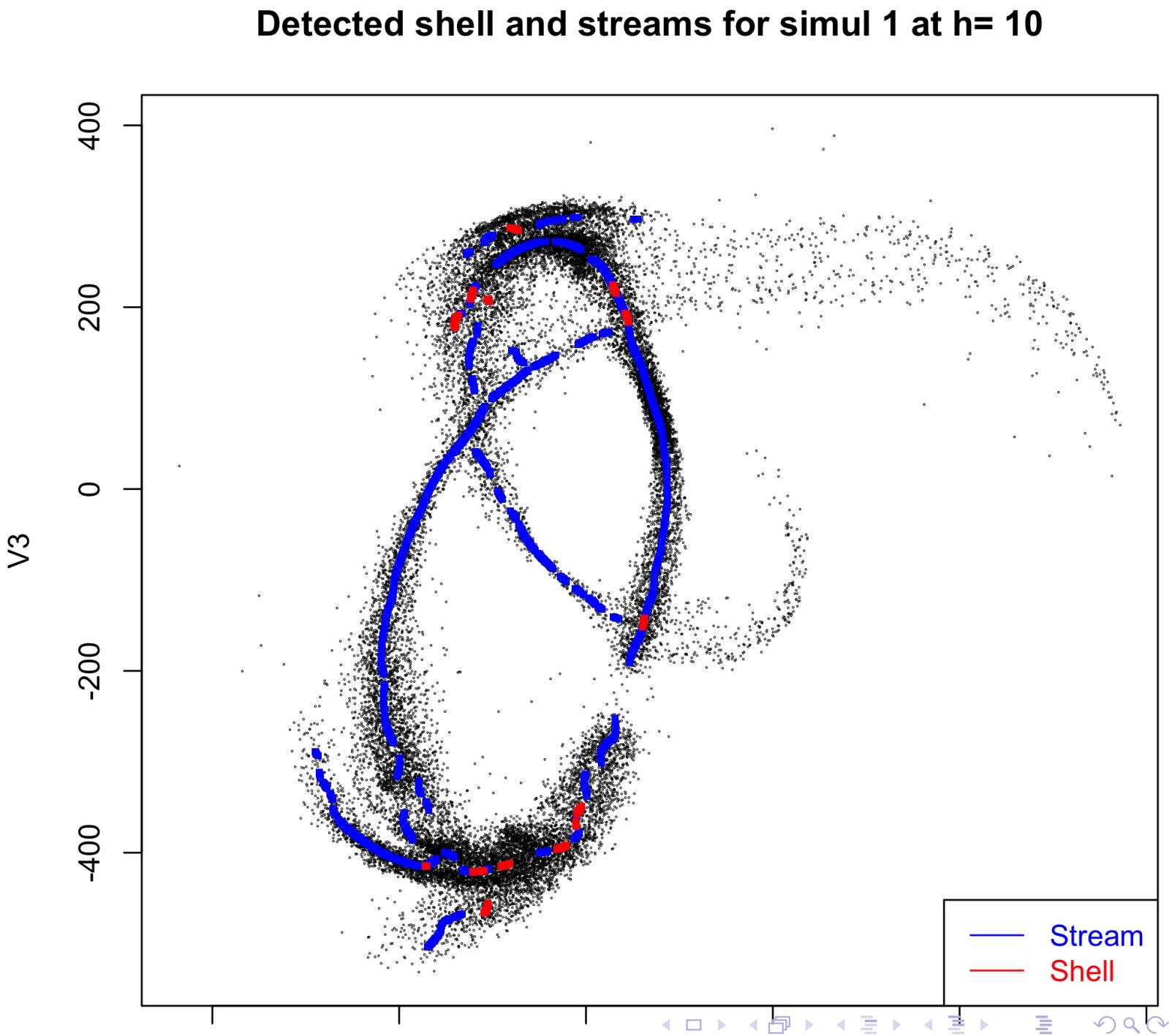

Detected shells for simul 2
y
x
400
200
0
-200
-400
-600
-400
-200
0
200
400
600
450.5::0.45::1
63.1::0.5::2
296.6::0.7::2
320.8::0.95::2
338.2::1.2::2
420.1::0.25::1
525.4::0.5::1

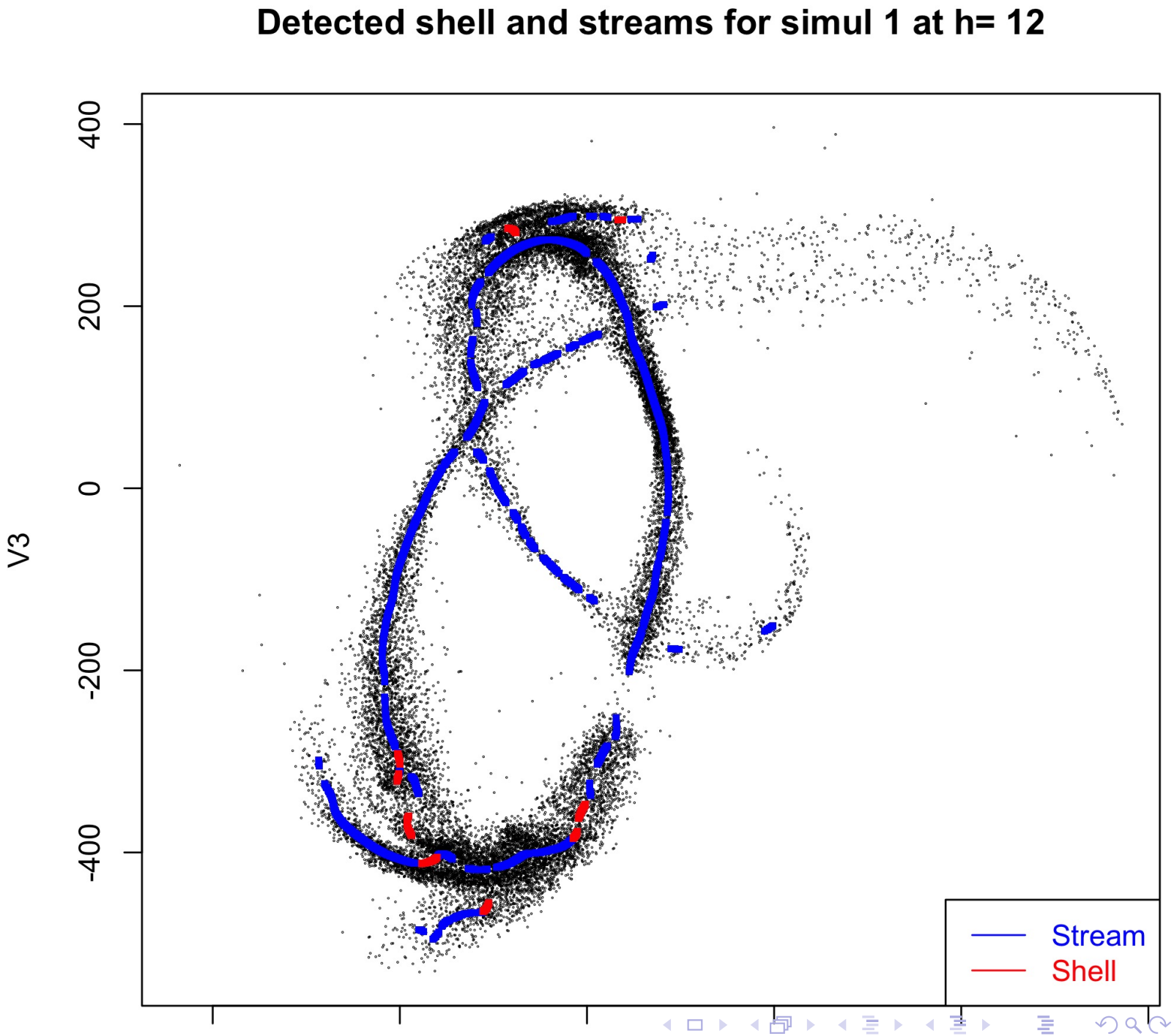

Detected streams for Simul 2
y
x
400
200
0
-200
-400
-600
-400
-200
0
200
400
600

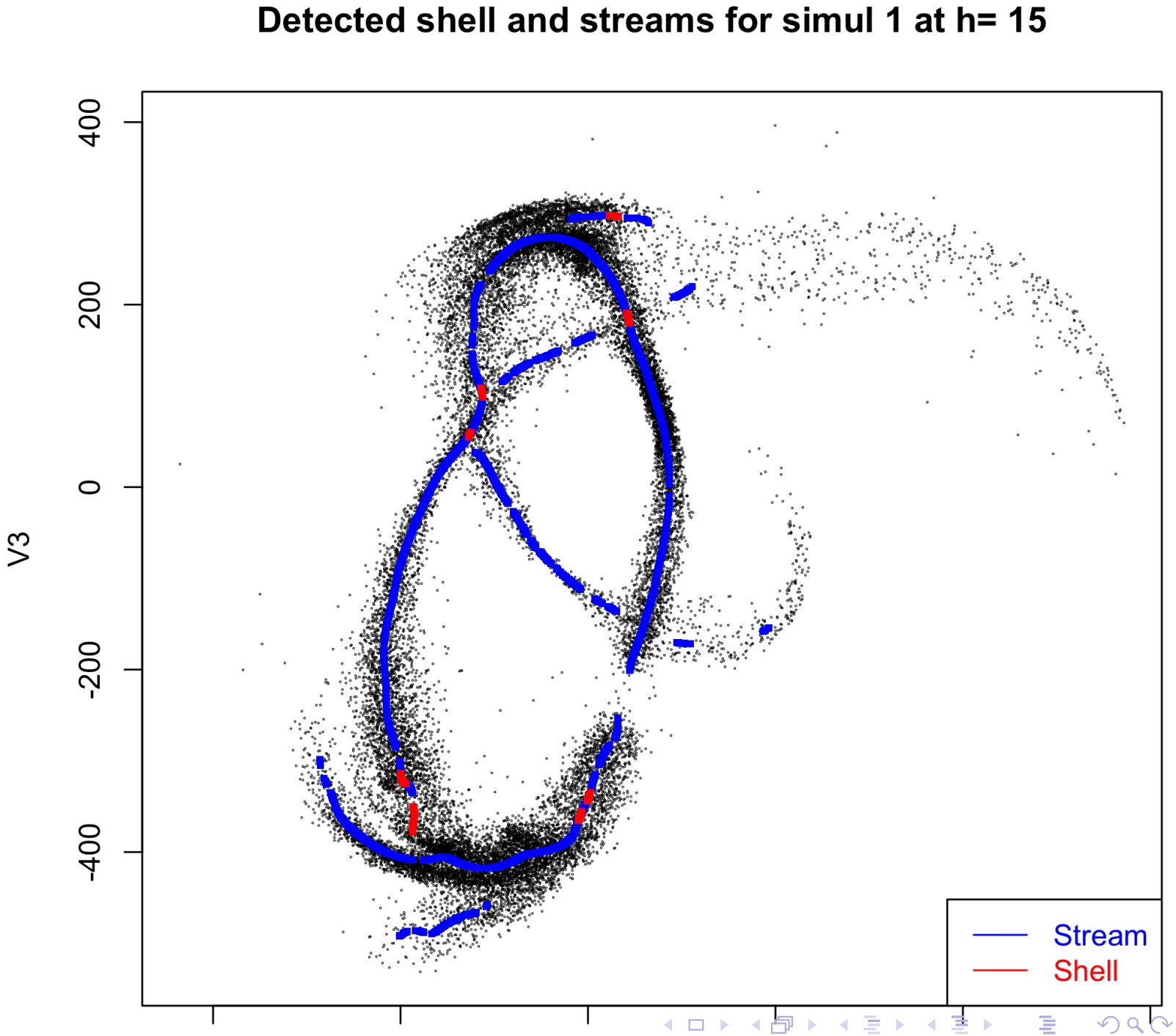

Detected shell and streams for simul 1 at h= 3
V3
400
200
0
-200
-400
Stream
Shell

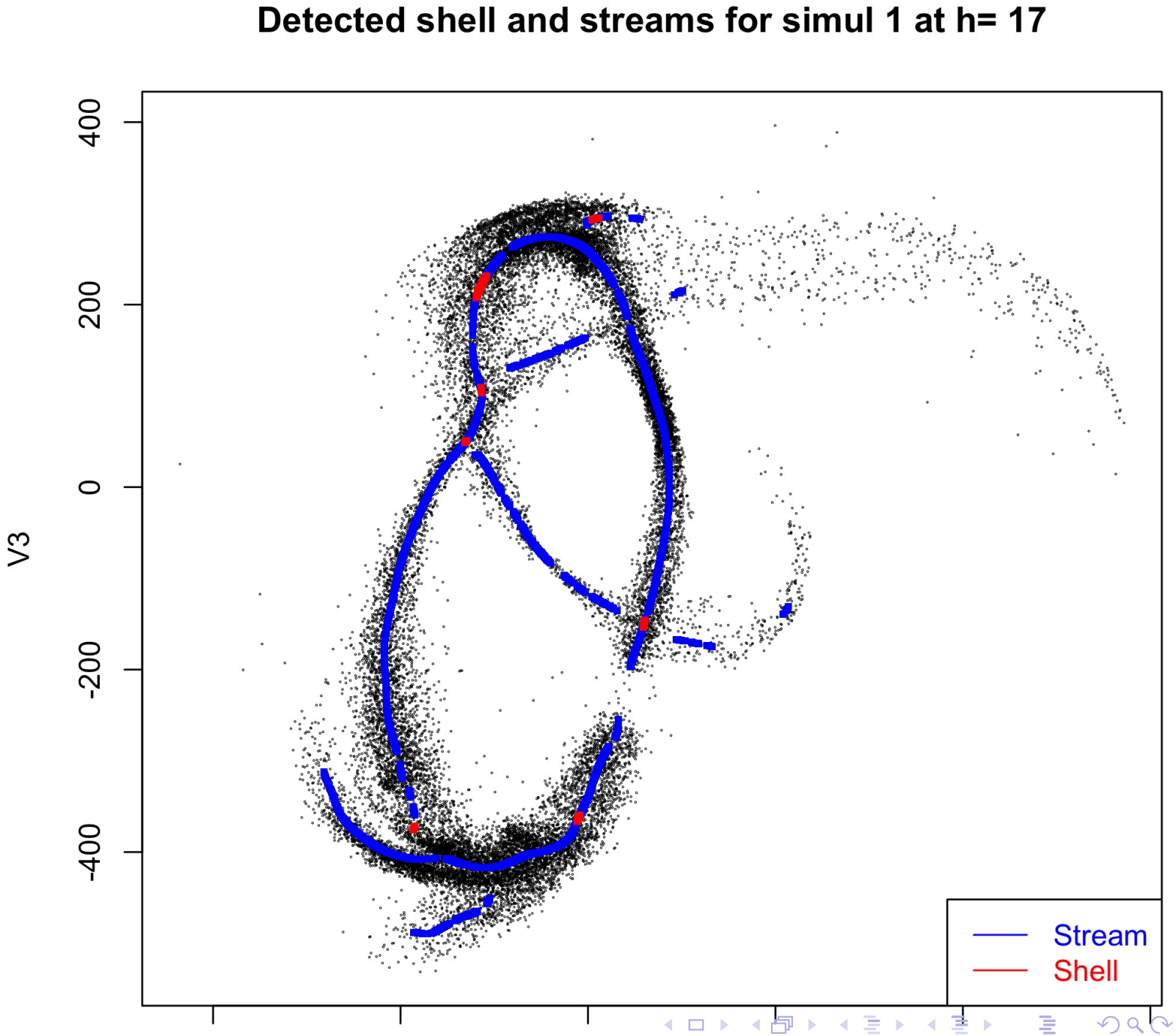

Detected shell and streams for simul 1 at h= 5
V3
400
200
0
-200
-400
Stream
Shell

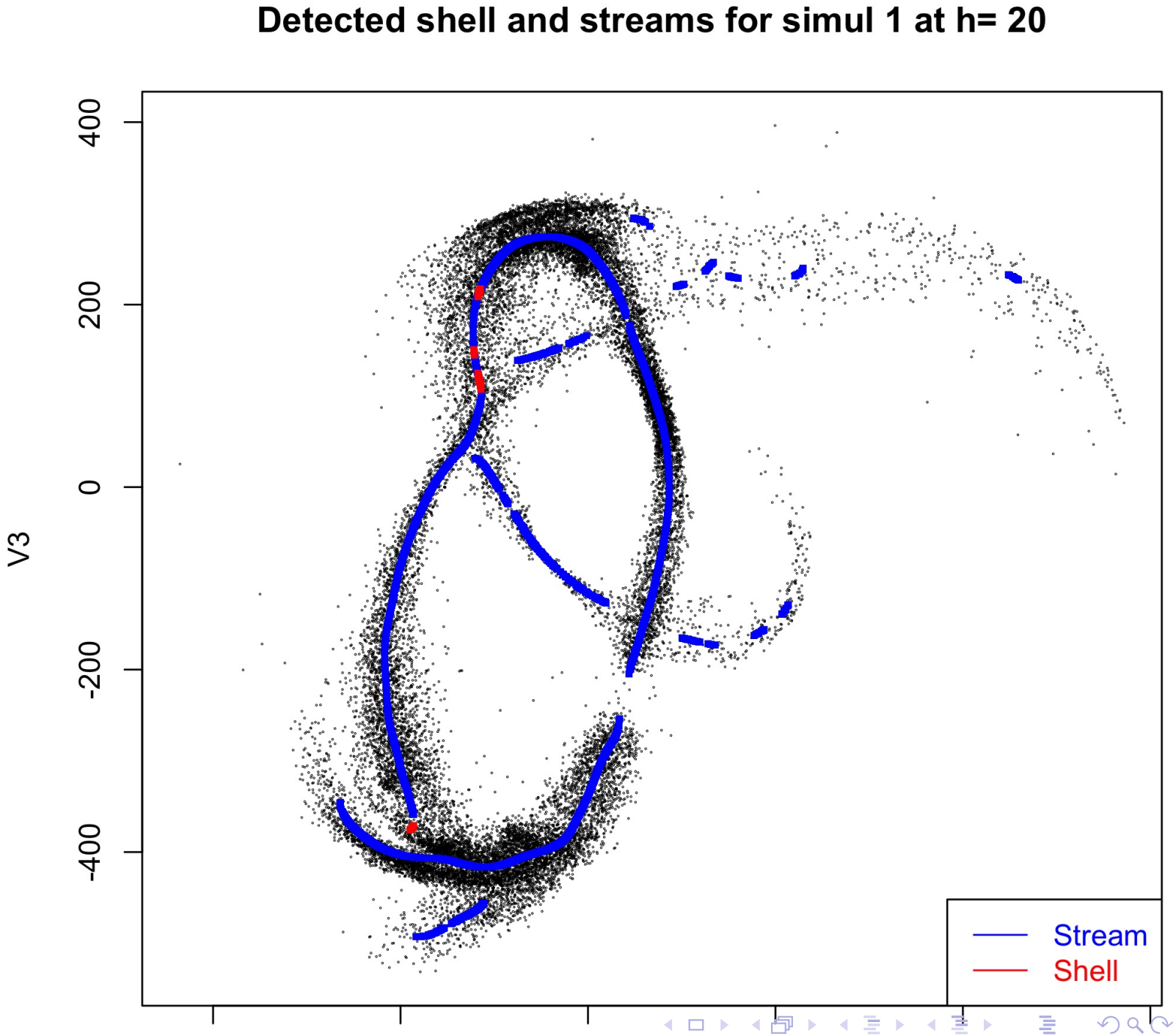

Detected shell and streams for simul 1 at h= 7
V3
400
200
0
-200
-400
Stream
Shell

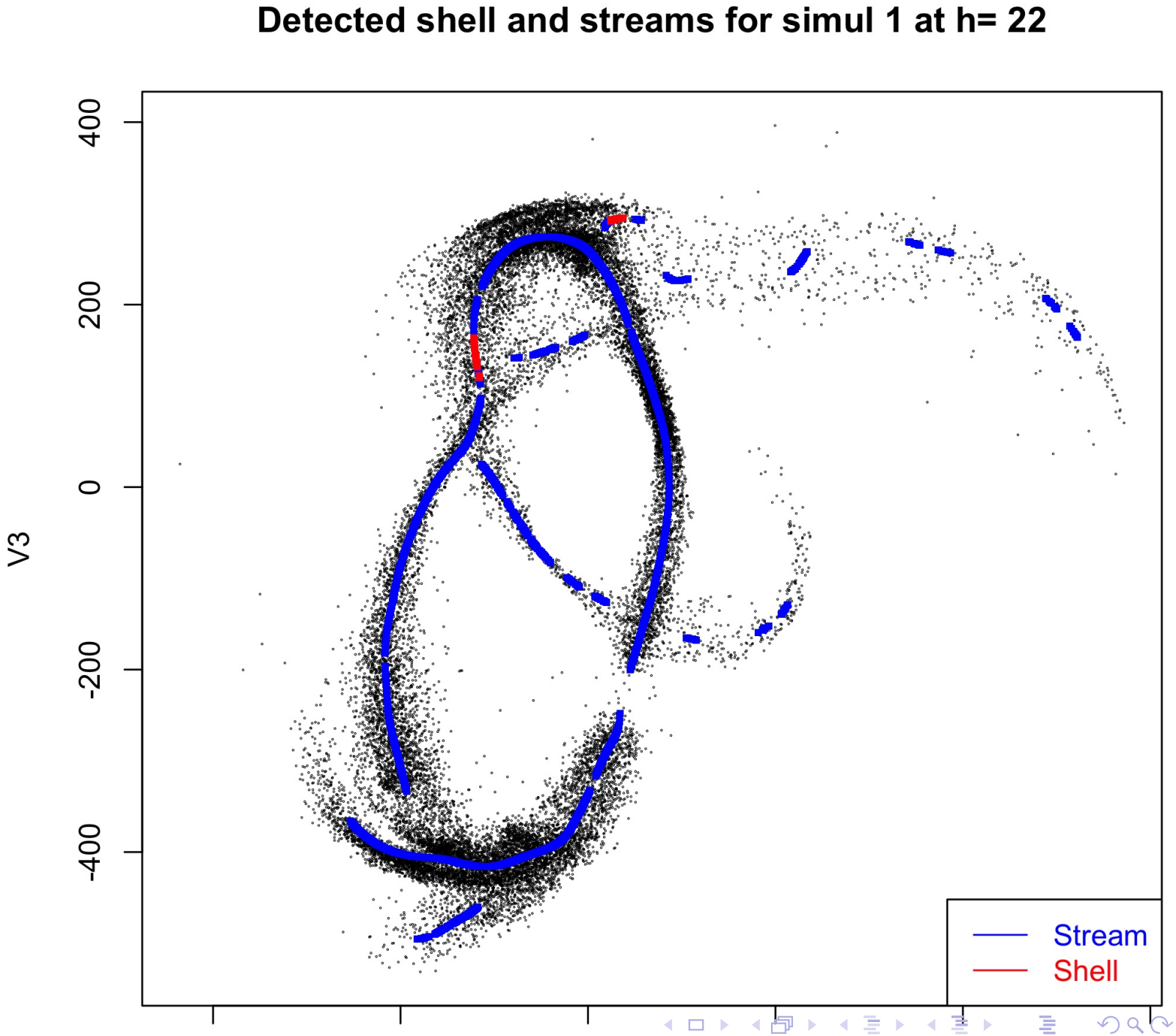

Detected shell and streams for simul 1 at h= 10
V3
400
200
0
-200
-400
Stream
Shell

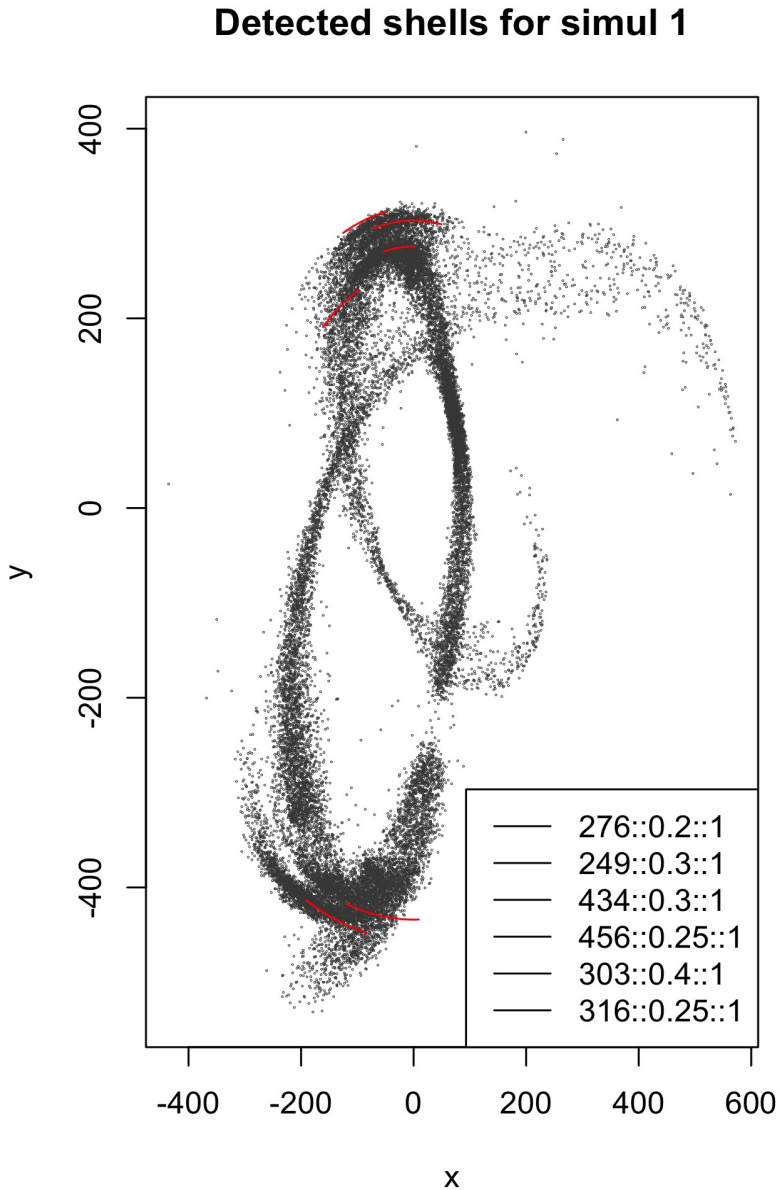

Detected shell and streams for simul 1 at h= 12
V3
400
200
0
-200
-400
Stream
Shell

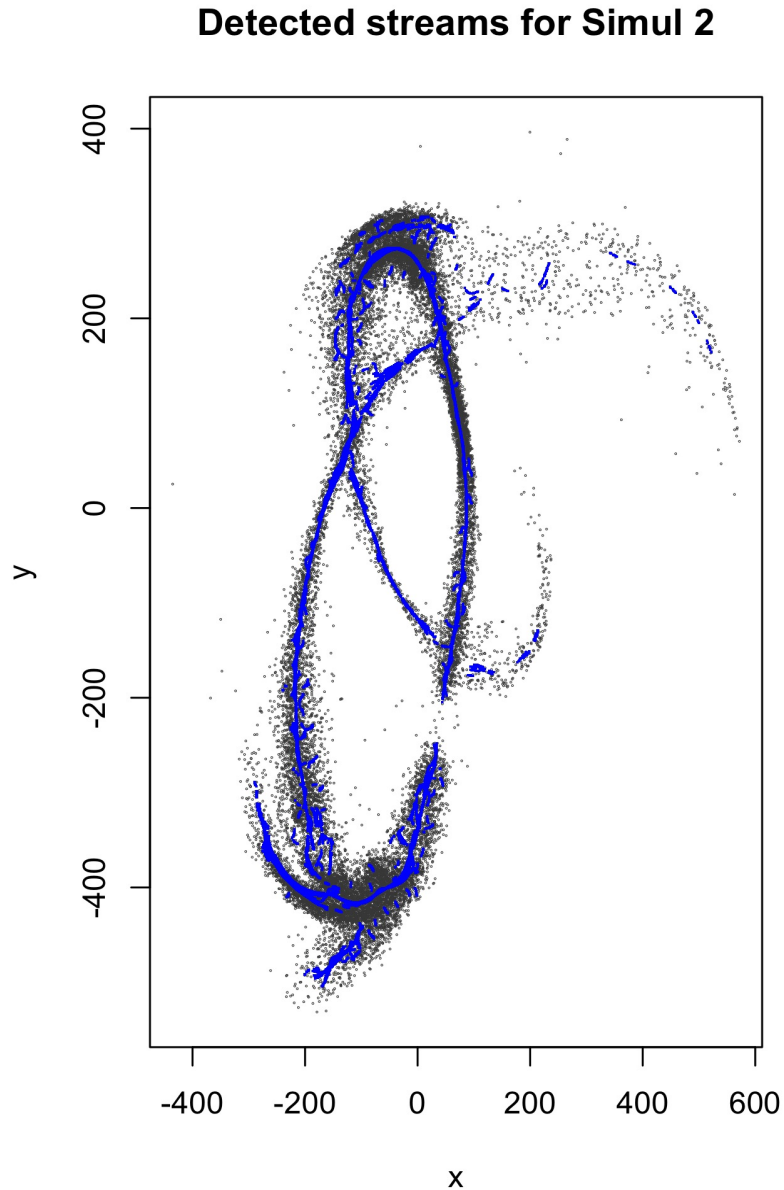

Detected shell and streams for simul 1 at h= 15
V3
400
200
0
-200
-400
Stream
Shell


Detected shell and streams for simul 1 at h= 17
V3
400
200
0
-200
-400
Stream
Shell


Detected shell and streams for simul 1 at h= 20
V3
400
200
0
-200
-400
Stream
Shell


Detected shell and streams for simul 1 at h= 22
V3
400
200
0
-200
-400
Stream
Shell


Detected shells for simul 1
y
x
400
200
0
-200
-400
-400
-200
0
200
400
600
276::0.2::1
249::0.3::1
434::0.3::1
456::0.25::1
303::0.4::1
316::0.25::1


Detected streams for Simul 2
y
x
400
200
0
-200
-400
-400
-200
0
200
400
600

Issues

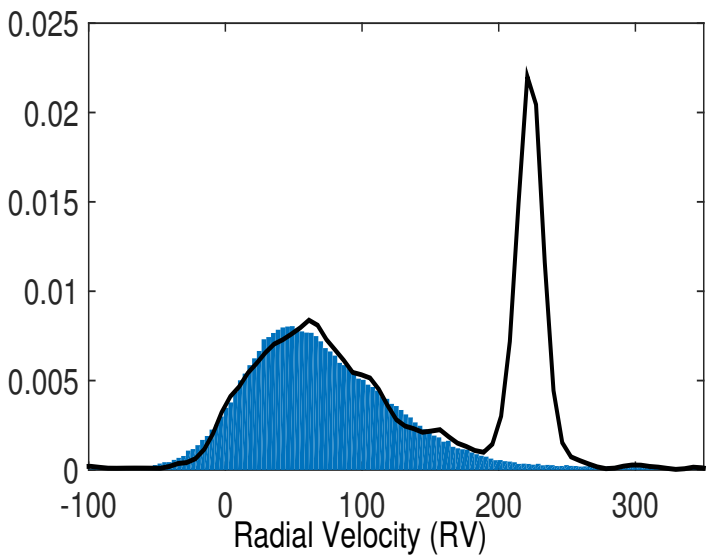
- How to choose the *bandwidth* (crucial)? Use a *multi-scale* idea
- Peaks are not always well-defined
- How to detect the *boundary* of the shell? Look at the *drop* in density
- How should we choose the width/length of the window?
- *Dip test* (for unimodality) used to choose the *length* of the window

II. Contamination Models

- Data: *Radial velocity* (RV) of stars ($n = 1215$) from Carina (*dSph*), *contaminated* with Milky Way stars in the field of view.

- **Question**: Find the distribution of RV of stars in *Carina* — F_s and the *proportion* of stars from Carina α

- F_b: The distribution of RV of the contaminating stars; known from the Besancon Milky Way model (Robin et. al, 2003)



Plot of histogram of F_b (blue) overlaid with the (scaled) *KDE* of data

Mixture model with two components

$$F(x) = \alpha F_s(x) + (1 - \alpha) F_b(x)$$

- F_b is a *known* distribution function (DF)
- *Unknowns*: Mixing proportion $\alpha \in (0, 1)$ and DF F_s $(\neq F_b)$
- Problem: Given a random *sample* from $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} F$ (and F_b), we wish to (*nonparametrically*) estimate F_s and the parameter α

Applications

- In *contamination* problems — application in astronomy
- In *multiple* testing problems — the p-values are *uniformly* distributed on [0,1], under H_0, while their distribution associated with H_1 is *unknown*; see e.g., Efron (2010)

Identifiability

- When α is *unknown*, the problem is *ill-posed*

- If $F = \alpha F_s + (1-\alpha)F_b$ for some F_b (known) and α (unknown), then the mixture model can be re-written as

$$F = (\alpha+\eta)\left(\frac{\alpha}{\alpha+\eta}F_s + \frac{\eta}{\alpha+\eta}F_b\right) + (1-\alpha-\eta)F_b,$$

for $0 \leq \eta \leq 1-\alpha$, and the term $(\alpha F_s + \eta F_b)/(\alpha+\eta)$ can be thought of as the *unknown* DF. Thus, if the model holds for α then it holds for $\alpha+\eta$, for every $\eta \leq 1-\alpha$.

Identifiable parameter

- We redefine the mixing proportion as

$$\alpha_0 := \inf\left\{\gamma \in (0,1] : \frac{F-(1-\gamma)F_b}{\gamma} \text{ is a valid DF}\right\}$$

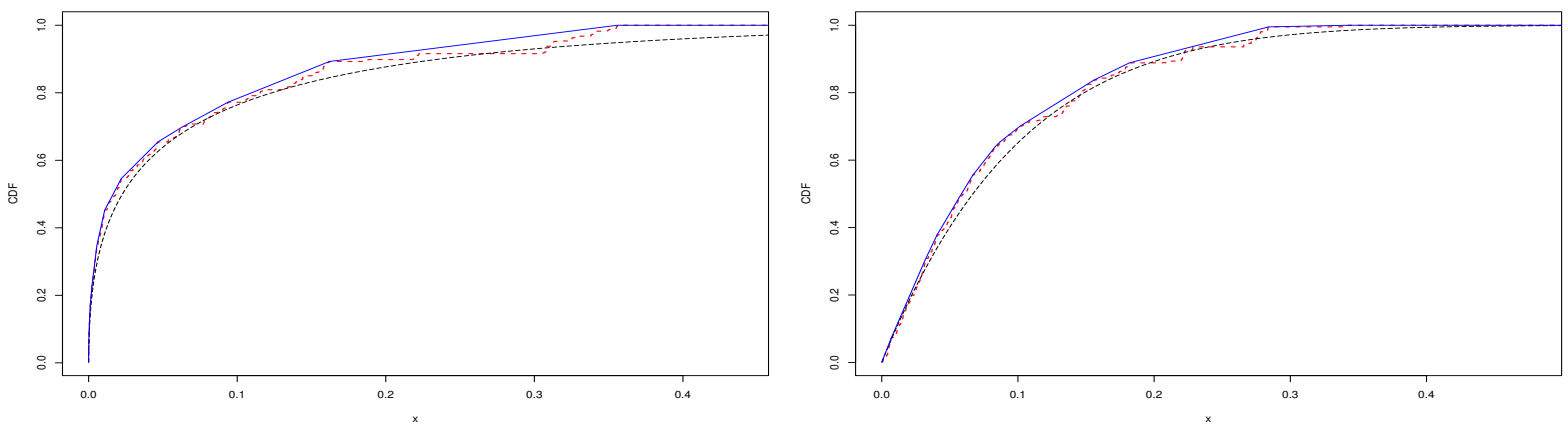
- Intuitively, this definition makes sure that the "signal" distribution F_s does *not* include any contribution from the known "background" F_b

Estimation of α_0 and F_s

- Can develop a *consistent* estimator $\hat{\alpha}_n$ of α_0
- Can construct *tuning parameter free* nonparametric estimators of F_s and f_s — the density of F_s (under some conditions)

Result

- If the model is *identifiable*, then $\hat{F}_{s,n}$ and $\hat{f}_{s,n}$ are consistent



Plot of $\hat{F}_{s,n}$ (in dotted red) and F_s (in dashed black) when $n = 300$

Lower confidence bound for α_0

- We can construct a *finite sample* (honest) *lower confidence bound* $\hat{\alpha}_L$ with the property

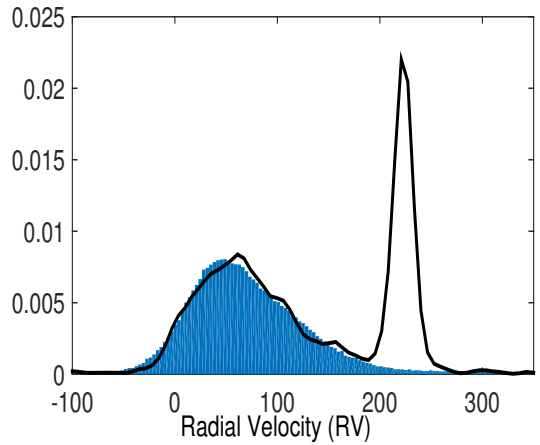
$$\mathbb{P}(\alpha_0 \geq \hat{\alpha}_L) \geq 1 - \beta, \qquad \text{for all } n,$$

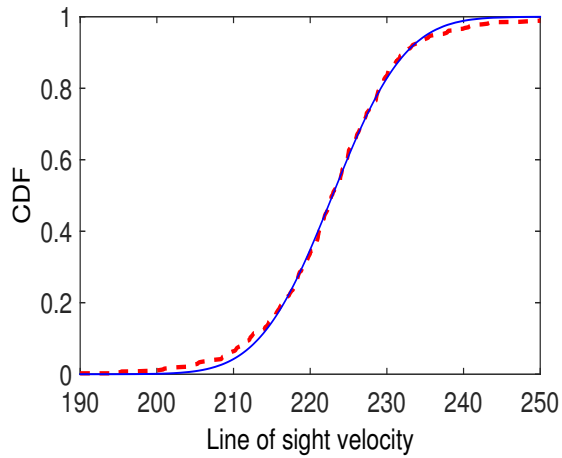
for a specified confidence level $(1 - \beta)$, $0 < \beta < 1$

- Would allow one to assert, with a specified level of confidence, that the proportion of "signal" is *at least* $\hat{\alpha}_L$
- $\hat{\alpha}_L$ is defined in terms of the $(1 - \beta)$-*quantile* of a distribution H_n
- H_n is *distribution-free*, when F is continuous, and can be simulated
- Requires *no tuning* parameters; lower bound holds *for all n*

Astronomy application

- Data: *Radial velocity* (RV) of stars ($n = 1215$) from Carina (*dSph*), *contaminated* with Milky Way stars in the field of view.





- Astronomers usually assume the distribution of the radial velocities for these dSph galaxies to be *Gaussian* in nature.
- The right panel shows $\hat{F}_{s,n}$ (in dashed red) overlaid with the *closest* Gaussian distribution (in blue).
- Our estimate $\hat{\alpha}_n$ of α_0 turns out to be 0.356, while the lower bound $\hat{\alpha}_L$ (at level 0.05) is found to be 0.322.

References

[1] Biswas, R., Hendel, D., Johnston, K. V., Patra, R. K., and **Sen, B.** (2016).
Statistical tools to categorize debris morphologies in a galaxy.
in preparation.

[2] Hendel, David and Johnston, Kathryn V. (2015).
Tidal debris morphology and the orbits of satellite galaxies, *Monthly Notices of the Royal Astronomical Society*, **454**,2472–2485.

[3] Ozertem, U. and Erdogmus, D. (2011).
Locally defined principal curves and surfaces. *The Journal of Machine Learning Research*, **12**, 1249–1286.

[4] Patra, R. K. and **Sen, B.** (2015).
Estimation of a two-component mixture model with ap- plications to multiple testing.
J. Roy. Statist. Soc. Ser. B (to appear).

Thank You! Questions?