HelmholtzZentrum münchen
German Research Center for Environmental Health

Approximate Bayesian Computation for Stochastic Single-Cell Time-Lapse Data Using Multivariate Test Statistics

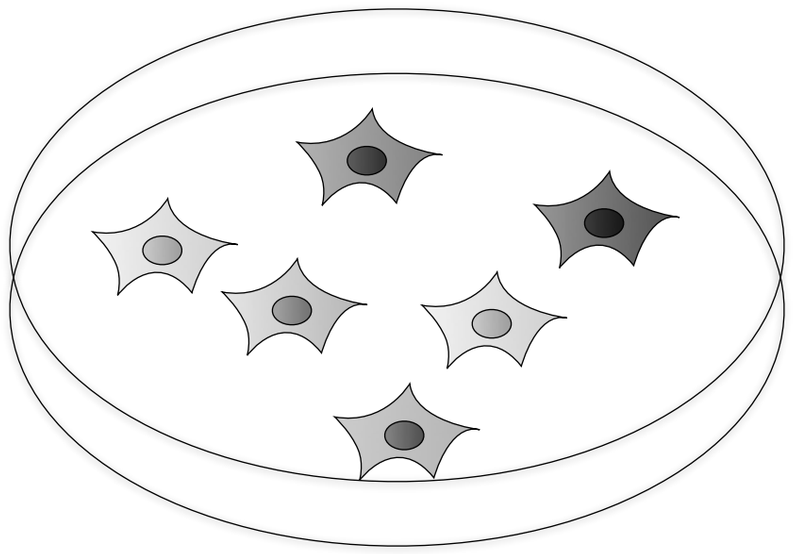
Carolin Loos, PhD student carolin.loos@helmholtz-muenchen.de
Institute of Computational Biology
Research Group Data-Driven Computational Modeling

Nantes, 16/09/15

Introduction

Introduction

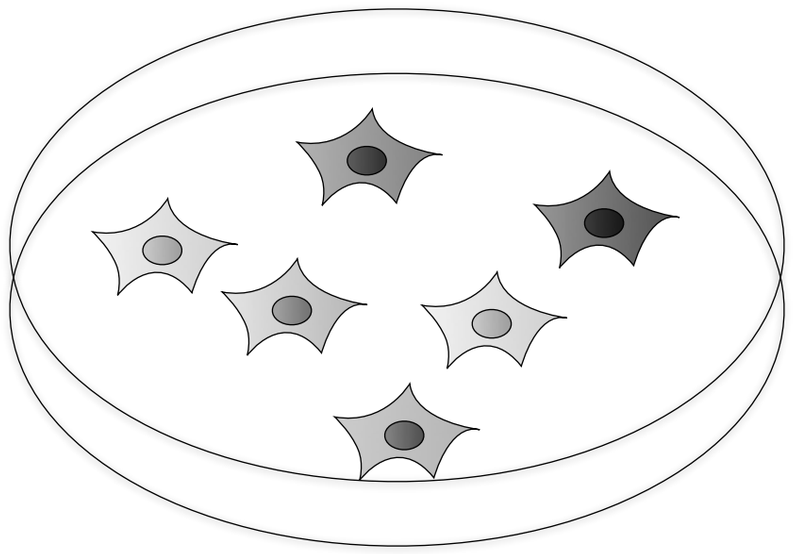
- Gene expression is known to be affected by different sources of variability[1]



1 Elowitz et al., Sience (2012)

Introduction

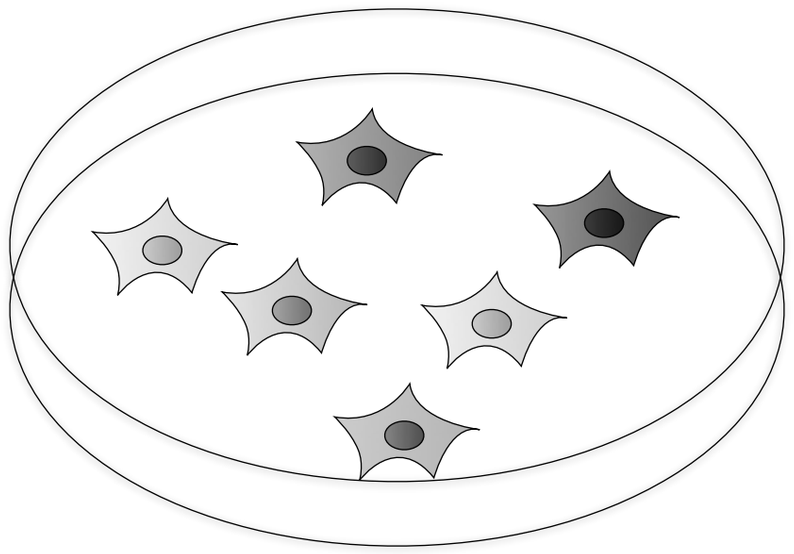
- Gene expression is known to be affected by different sources of variability[1]
- Heterogeneity and stochasticity can have major consequences for the understanding of the cellular mechanisms[2]



1 Elowitz et al., Sience (2012) **2** Wilkinson, Nature Reviews Genetics (2009)

Introduction

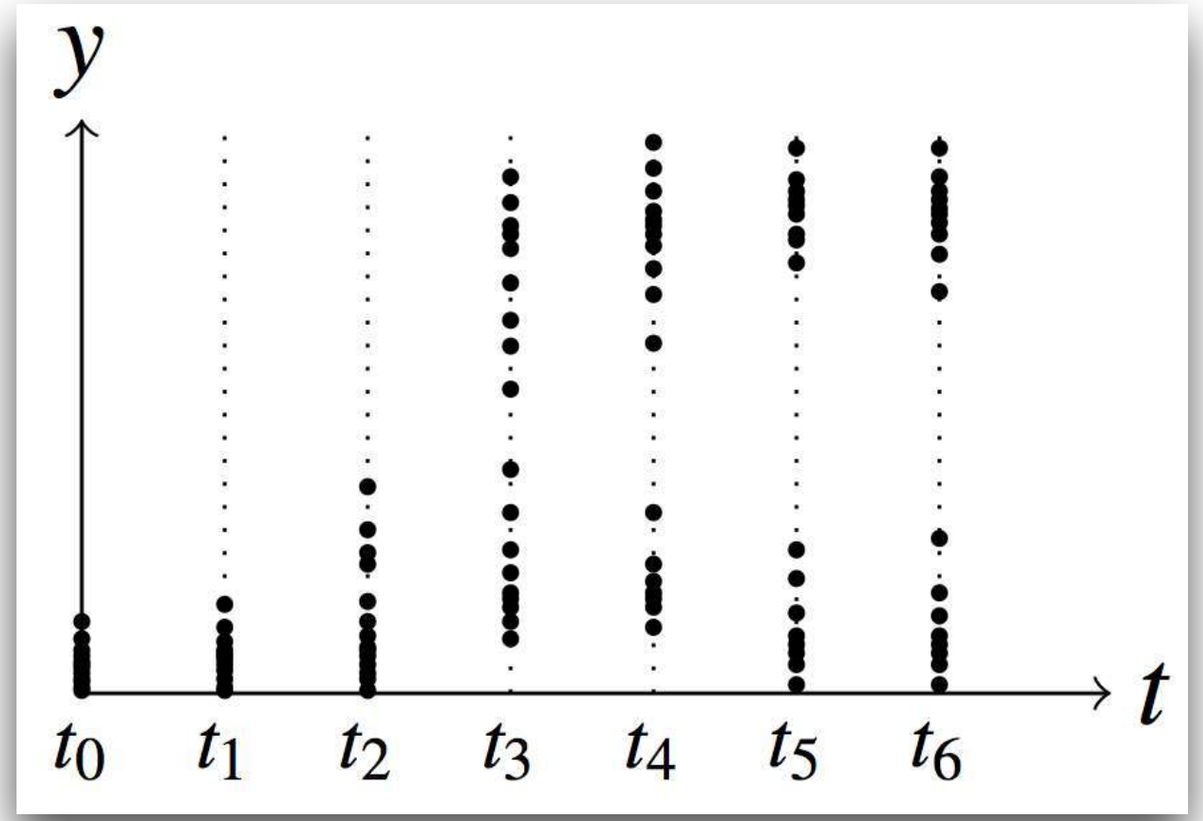
- Gene expression is known to be affected by different sources of variability[1]
- Heterogeneity and stochasticity can have major consequences for the understanding of the cellular mechanisms[2]
- Single-cell data for analysis of variability required



1 Elowitz et al., Sience (2012) **2** Wilkinson, Nature Reviews Genetics (2009)

Single-cell data

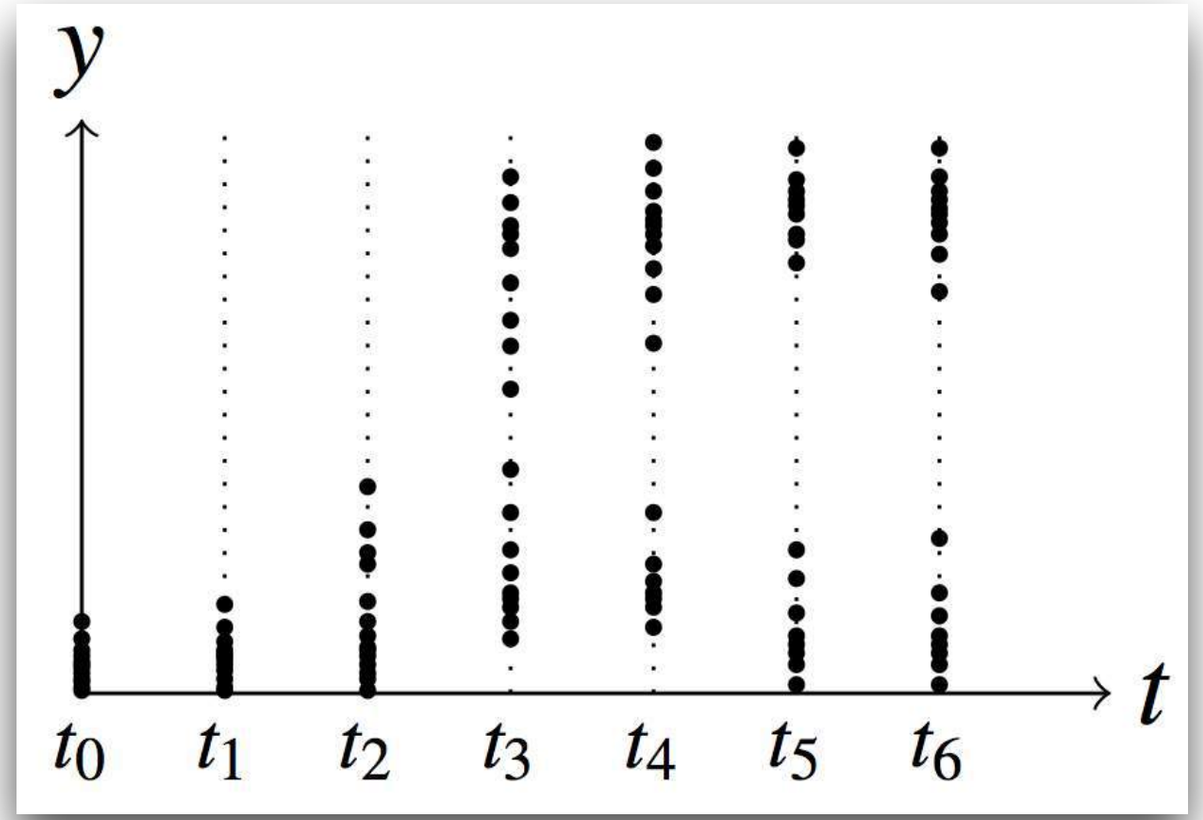
Single-cell data



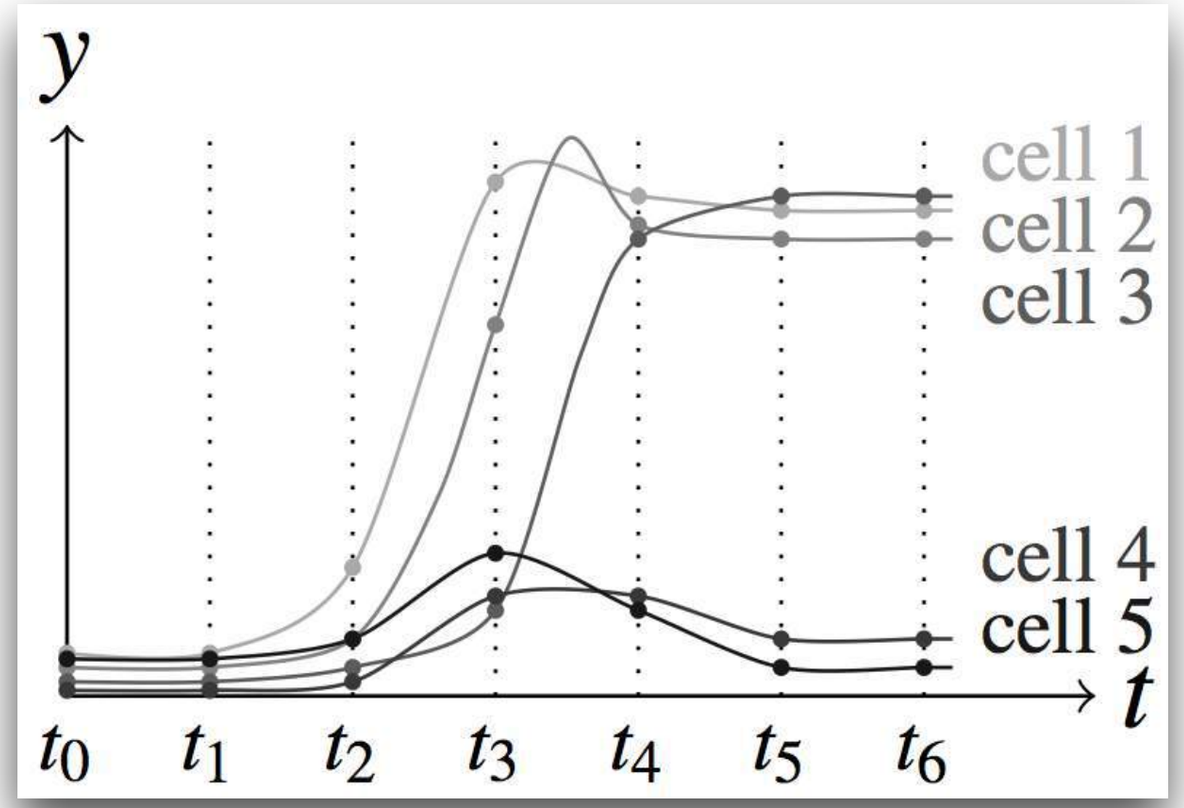
Single-cell snapshot data

Hasenauer, PhD thesis (2013)

Single-cell data



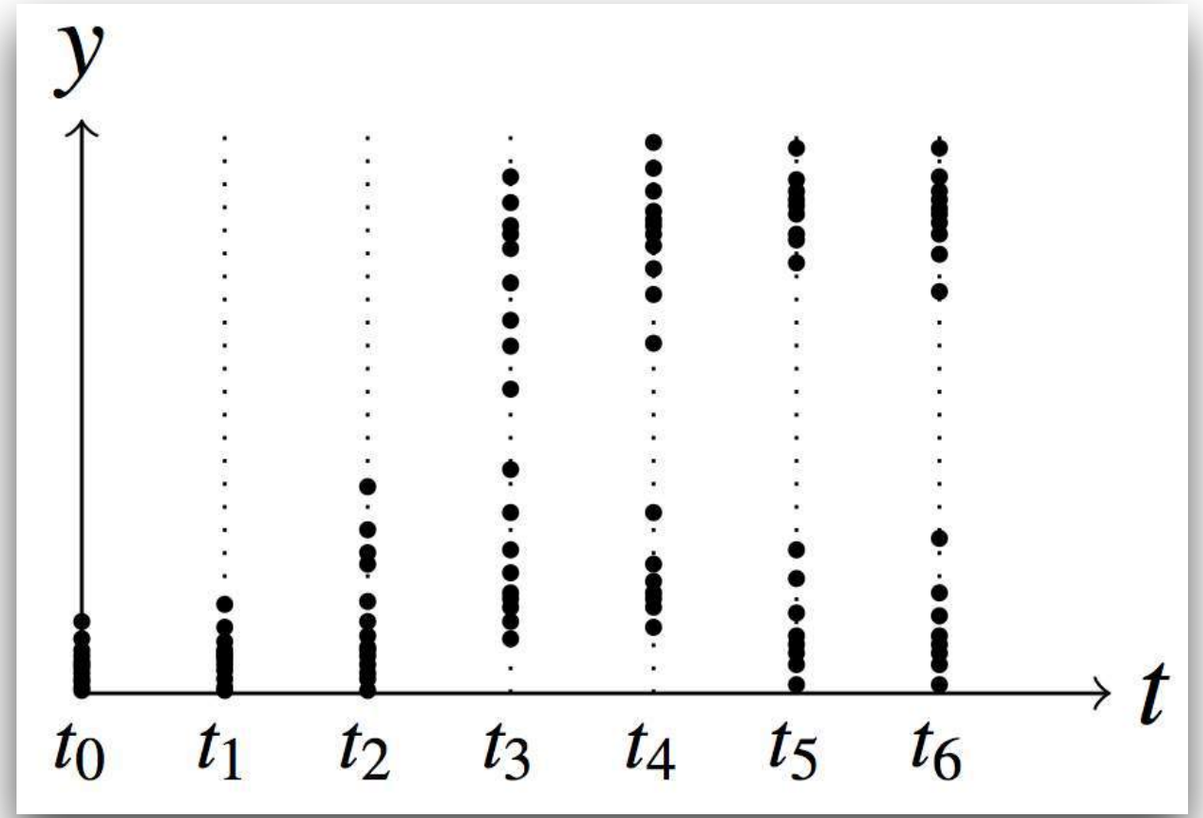
Single-cell snapshot data



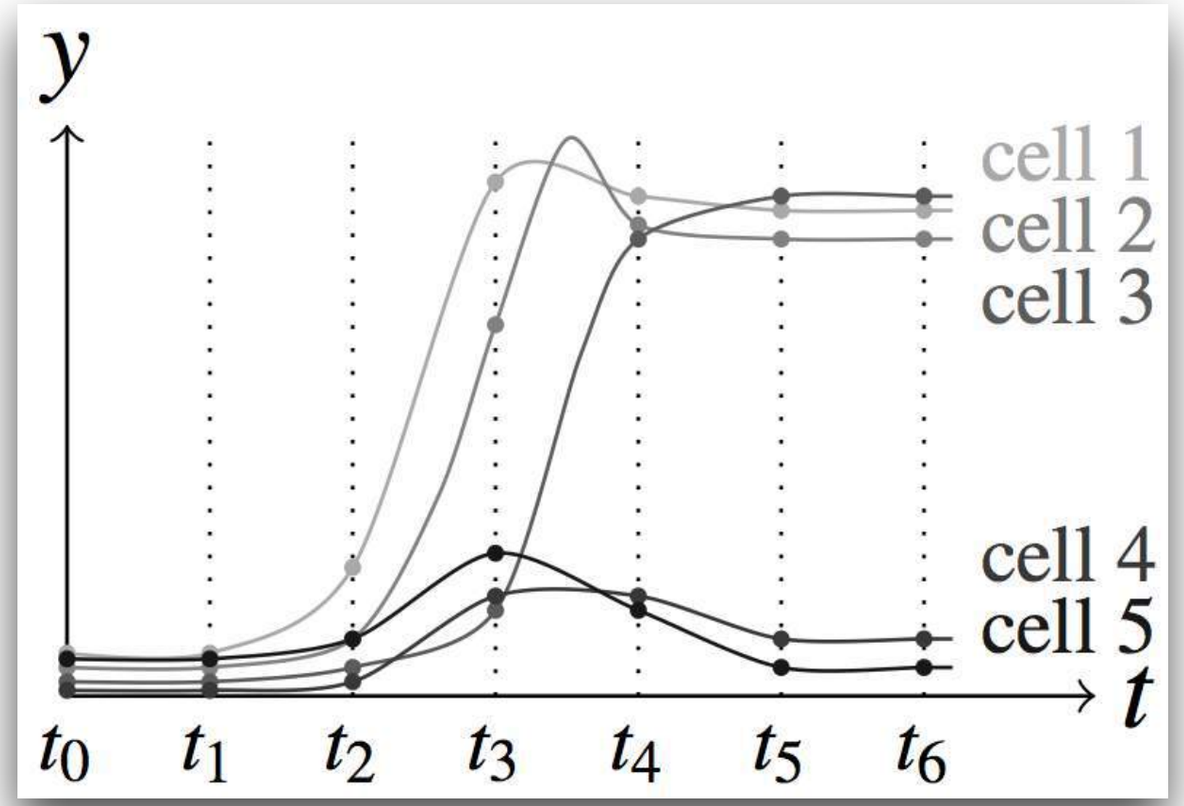
Single-cell time-lapse data

Hasenauer, PhD thesis (2013)

Single-cell data



Single-cell snapshot data



Single-cell time-lapse data

Single-cell time-lapse data provide information about temporal cross-correlation

Hasenauer, PhD thesis (2013)

Parameter inference for single-cell time-lapse data

Parameter inference for single-cell time-lapse data

General idea:

Derive model M with parameters θ and fit it to data D by e.g. maximizing the likelihood function or sampling from the posterior distribution p(θ|D)

Parameter inference for single-cell time-lapse data

General idea:

Derive model M with parameters θ and fit it to data D by e.g. maximizing the likelihood function or sampling from the posterior distribution p(θ|D)

- D: Single-cell time-series $\mathbf{x}_1, \ldots, \mathbf{x}_n$

Parameter inference for single-cell time-lapse data

General idea:

Derive model M with parameters θ and fit it to data D by e.g. maximizing the likelihood function or sampling from the posterior distribution p(θ|D)

- D: Single-cell time-series $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- M: Single-cell time-lapse data modeled with continuous time Markov chains (CTMCs)[1]

1 Gillespie, Annual Review of Physical Chemistry (2007)

Parameter inference for single-cell time-lapse data

General idea:

Derive model M with parameters θ and fit it to data D by e.g. maximizing the likelihood function or sampling from the posterior distribution p(θ|D)

- D: Single-cell time-series $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- M: Single-cell time-lapse data modeled with continuous time Markov chains (CTMCs)[1]
- θ: Parameters of process e.g. kinetic parameters, initial conditions, …

1 Gillespie, Annual Review of Physical Chemistry (2007)

Parameter inference for single-cell time-lapse data

Parameter inference for single-cell time-lapse data

- Evaluation of likelihood function for CTMCs often computationally too costly

Parameter inference for single-cell time-lapse data

- Evaluation of likelihood function for CTMCs often computationally too costly
- Approximate Bayesian Computation (ABC) methods required

Parameter inference for single-cell time-lapse data

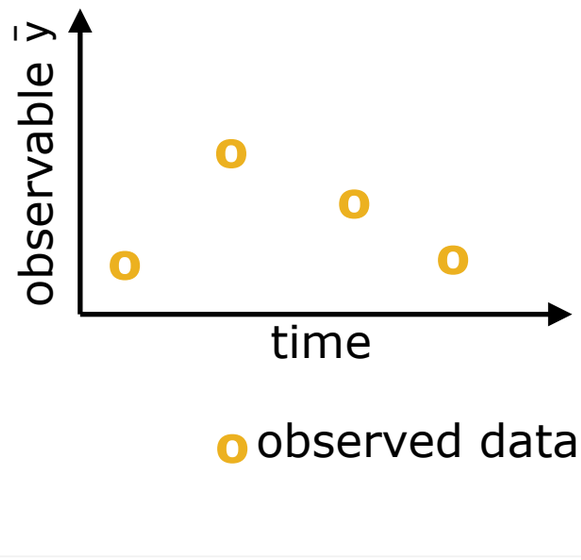
- Evaluation of likelihood function for CTMCs often computationally too costly

- Approximate Bayesian Computation (ABC) methods required

- ABC circumvents calculation of likelihood by comparing observed and simulated data set

Approximate Bayesian Computation rejection

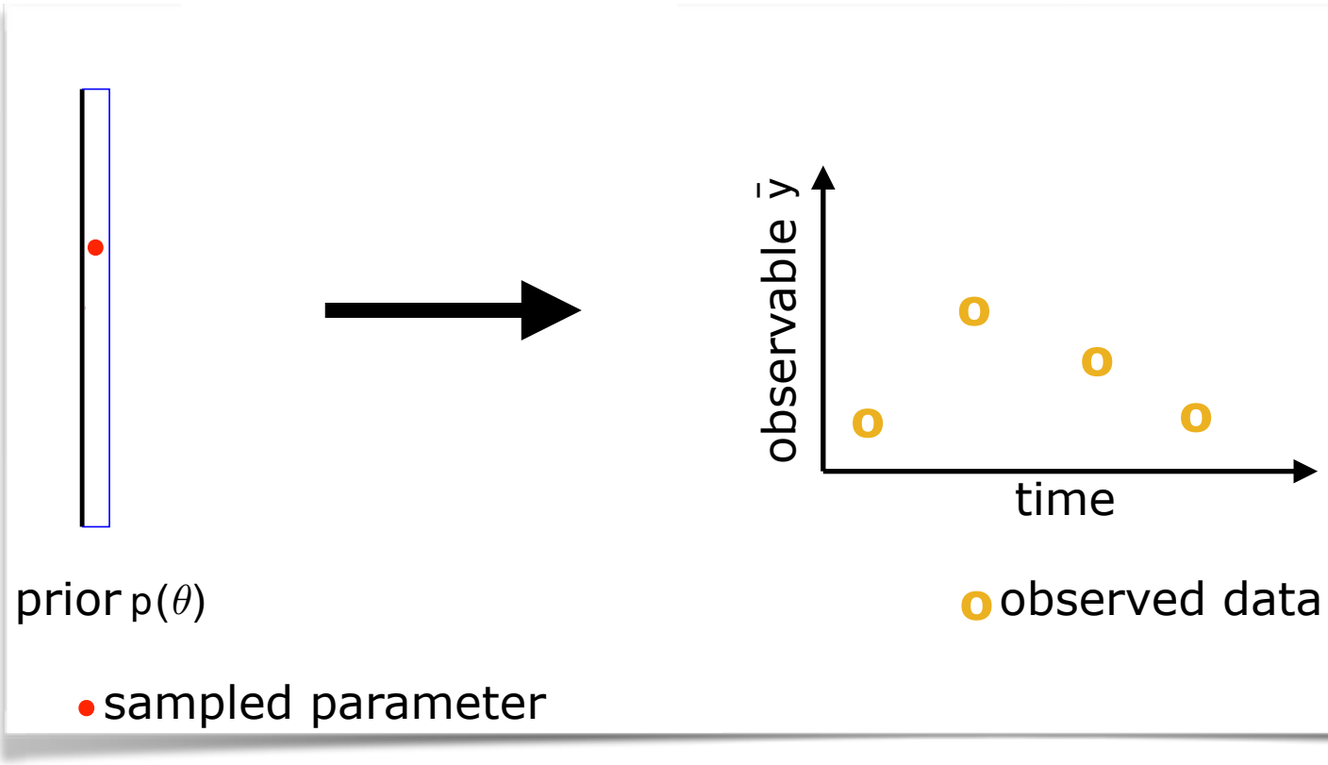
Approximate Bayesian Computation rejection



Goal: Approximate posterior distribution of θ given data D

Figure modified from Toni et al., Bioinformatics (2010)

Approximate Bayesian Computation rejection

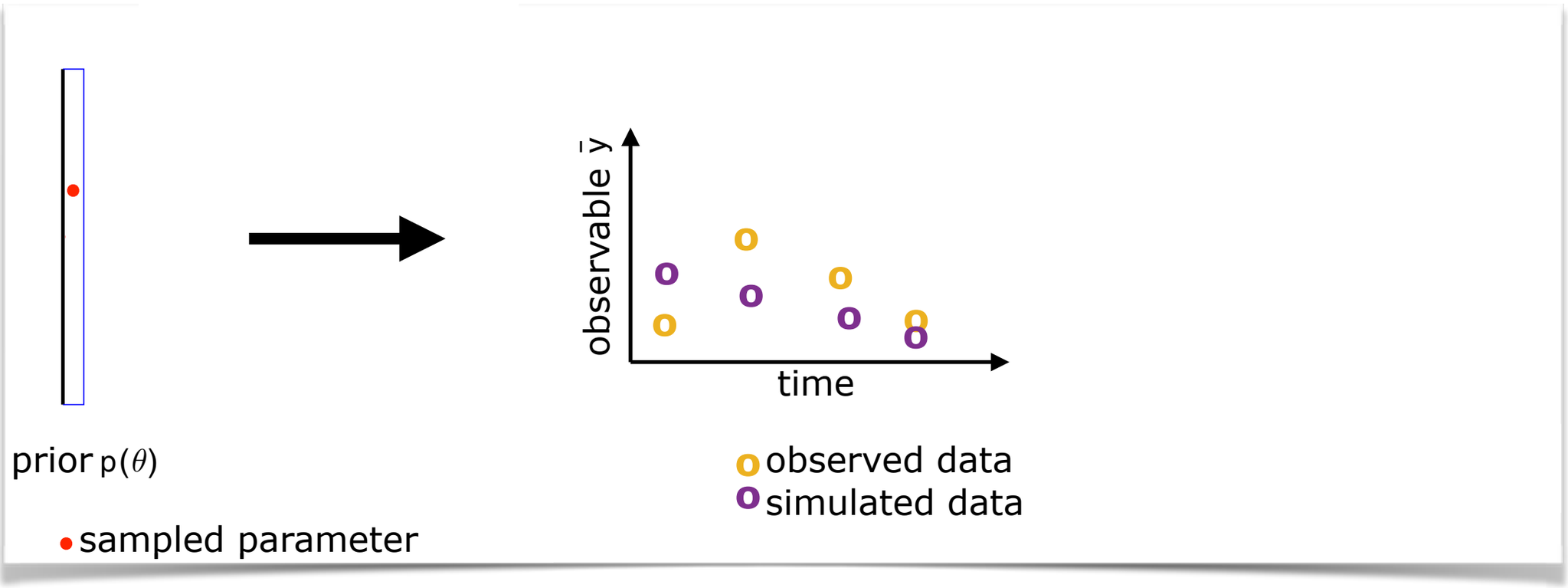


Goal: Approximate posterior distribution of θ given data D

1. Sample from prior distribution

Figure modified from Toni et al., Bioinformatics (2010)

Approximate Bayesian Computation rejection

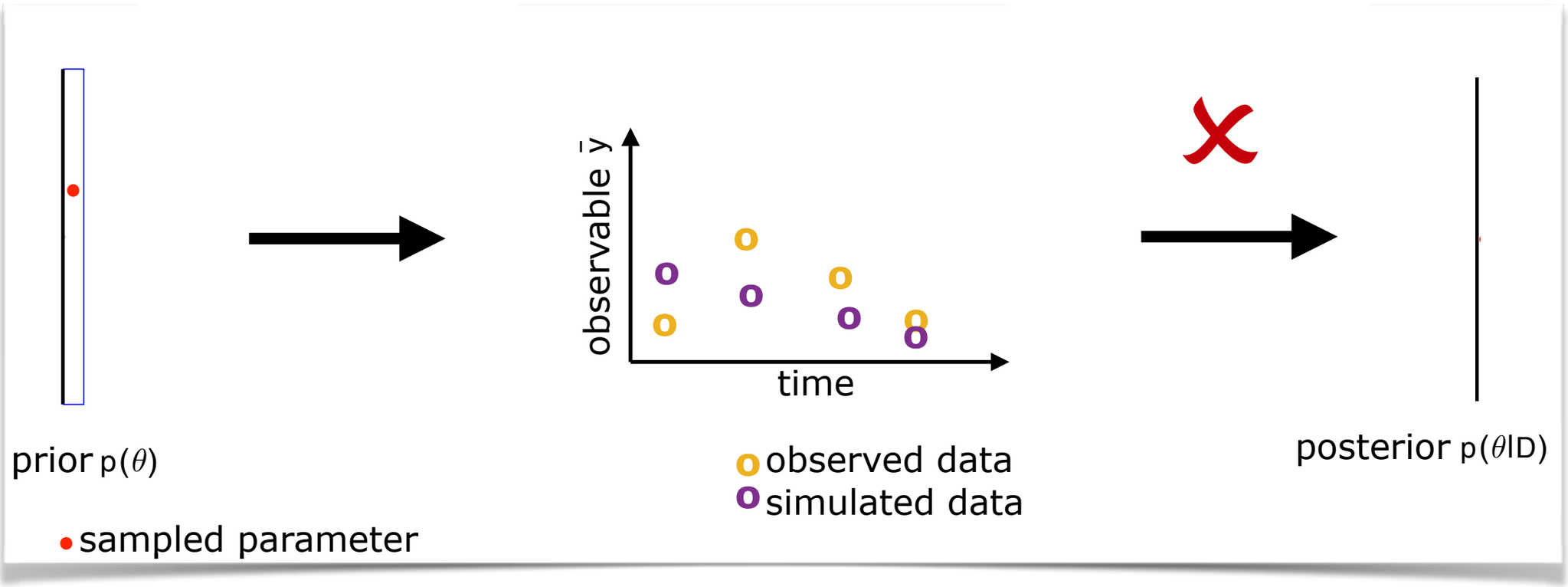


Goal: Approximate posterior distribution of θ given data D

1. Sample from prior distribution
2. Simulate data set

Figure modified from Toni et al., Bioinformatics (2010)

Approximate Bayesian Computation rejection

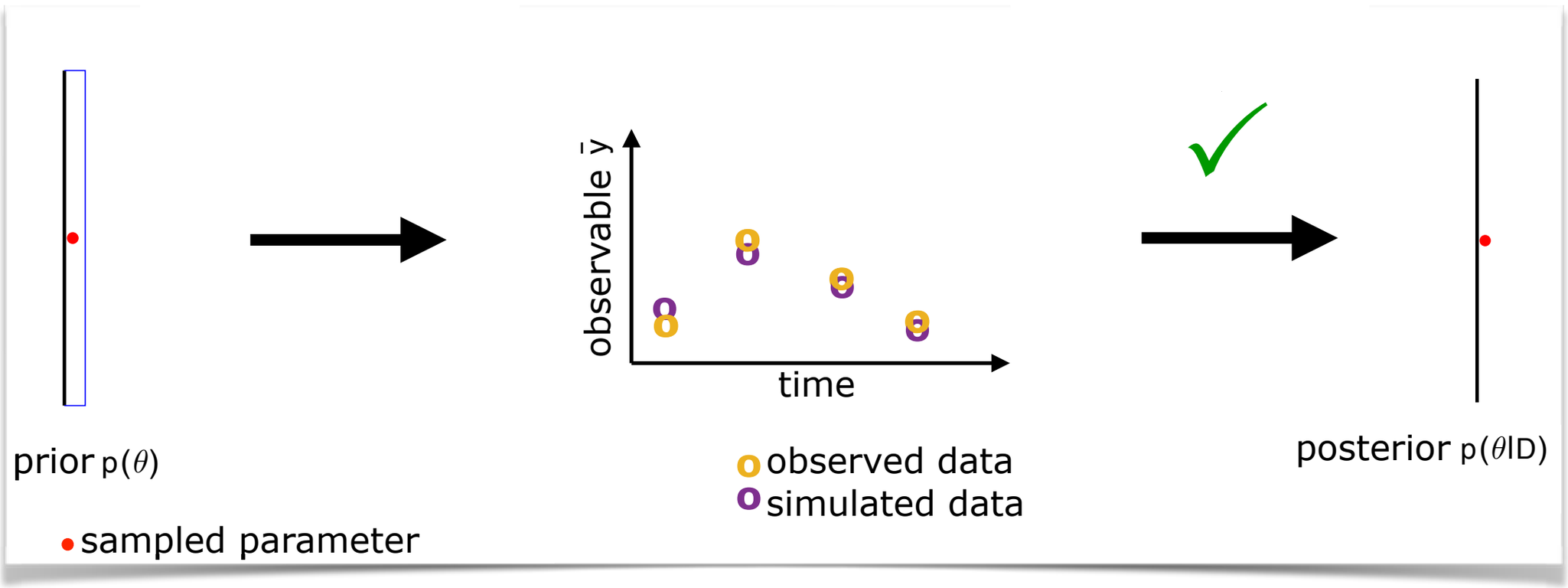


Goal: Approximate posterior distribution of θ given data D

1. Sample from prior distribution
2. Simulate data set
3. Accept or reject sampled parameter with respect to distance between simulated and observed data

Figure modified from Toni et al., Bioinformatics (2010)

Approximate Bayesian Computation rejection



Goal: Approximate posterior distribution of θ given data D

1. Sample from prior distribution
2. Simulate data set
3. Accept or reject sampled parameter with respect to distance between simulated and observed data

Figure modified from Toni et al., Bioinformatics (2010)

Approximate Bayesian Computation with Sequential Monte Carlo (ABC SMC)

Approximate Bayesian Computation with Sequential Monte Carlo (ABC SMC)

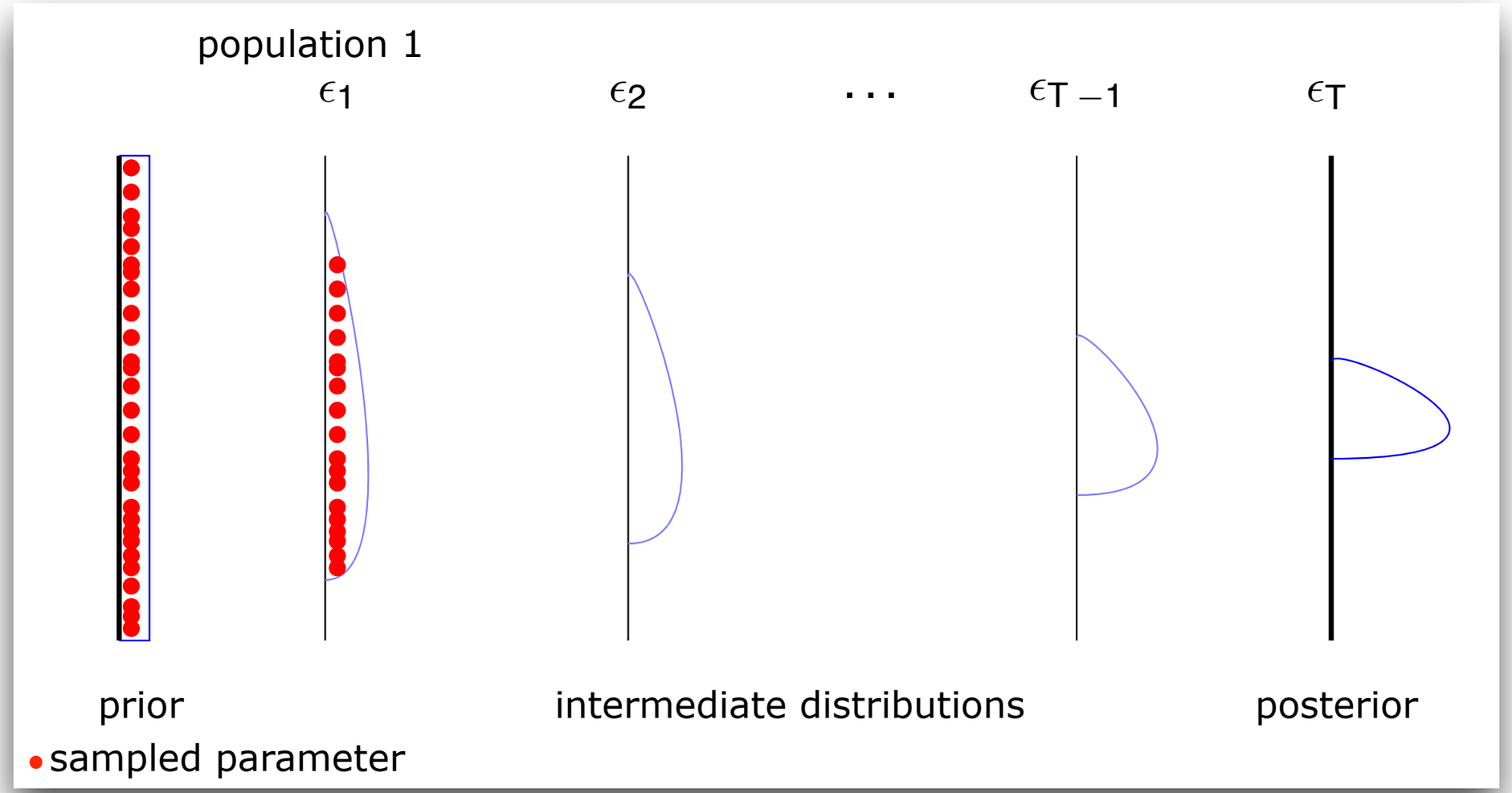


Figure modified from Toni et al., Bioinformatics (2010)

Approximate Bayesian Computation with Sequential Monte Carlo (ABC SMC)

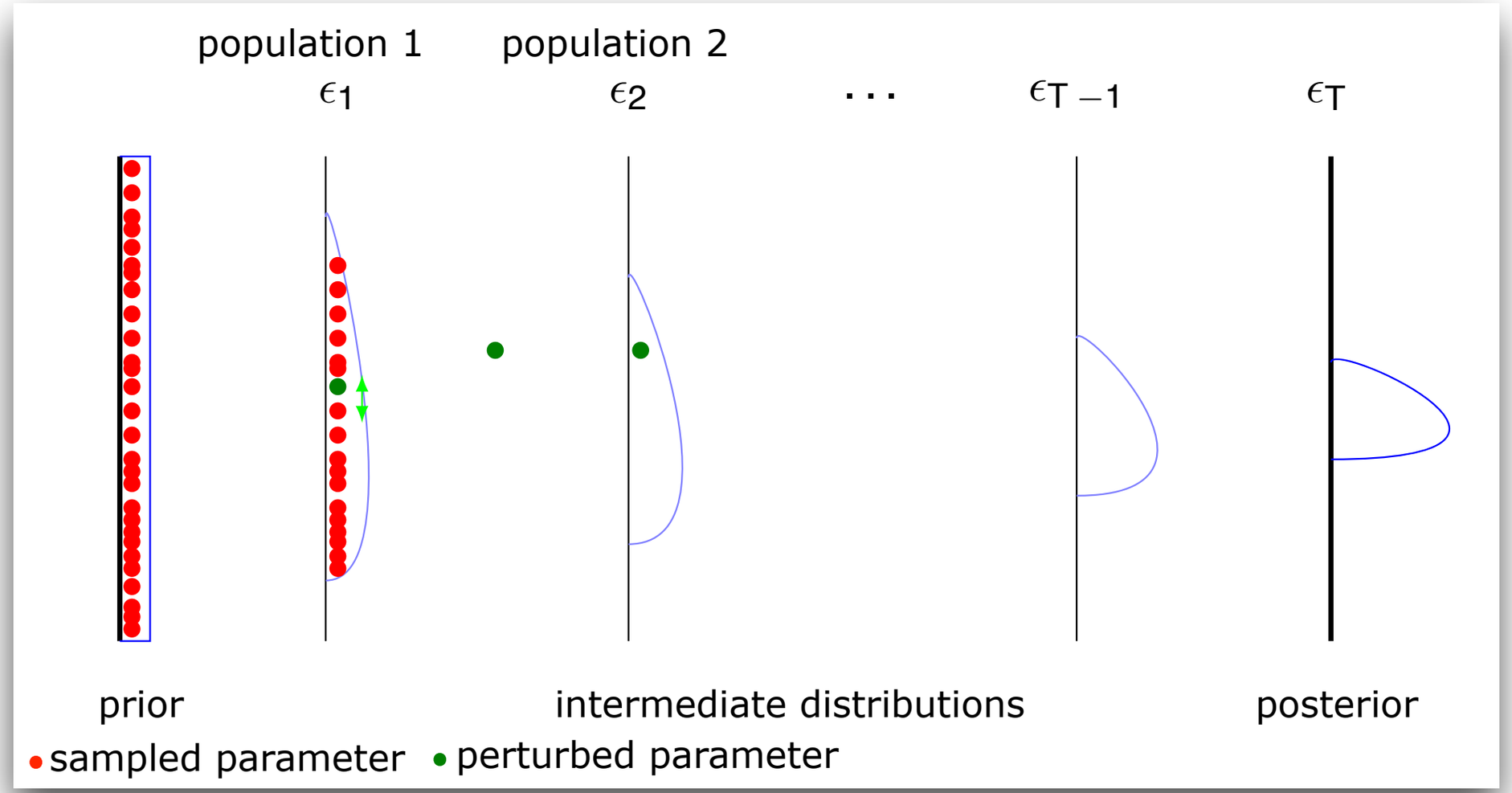


Figure modified from Toni et al., Bioinformatics (2010)

Approximate Bayesian Computation with Sequential Monte Carlo (ABC SMC)

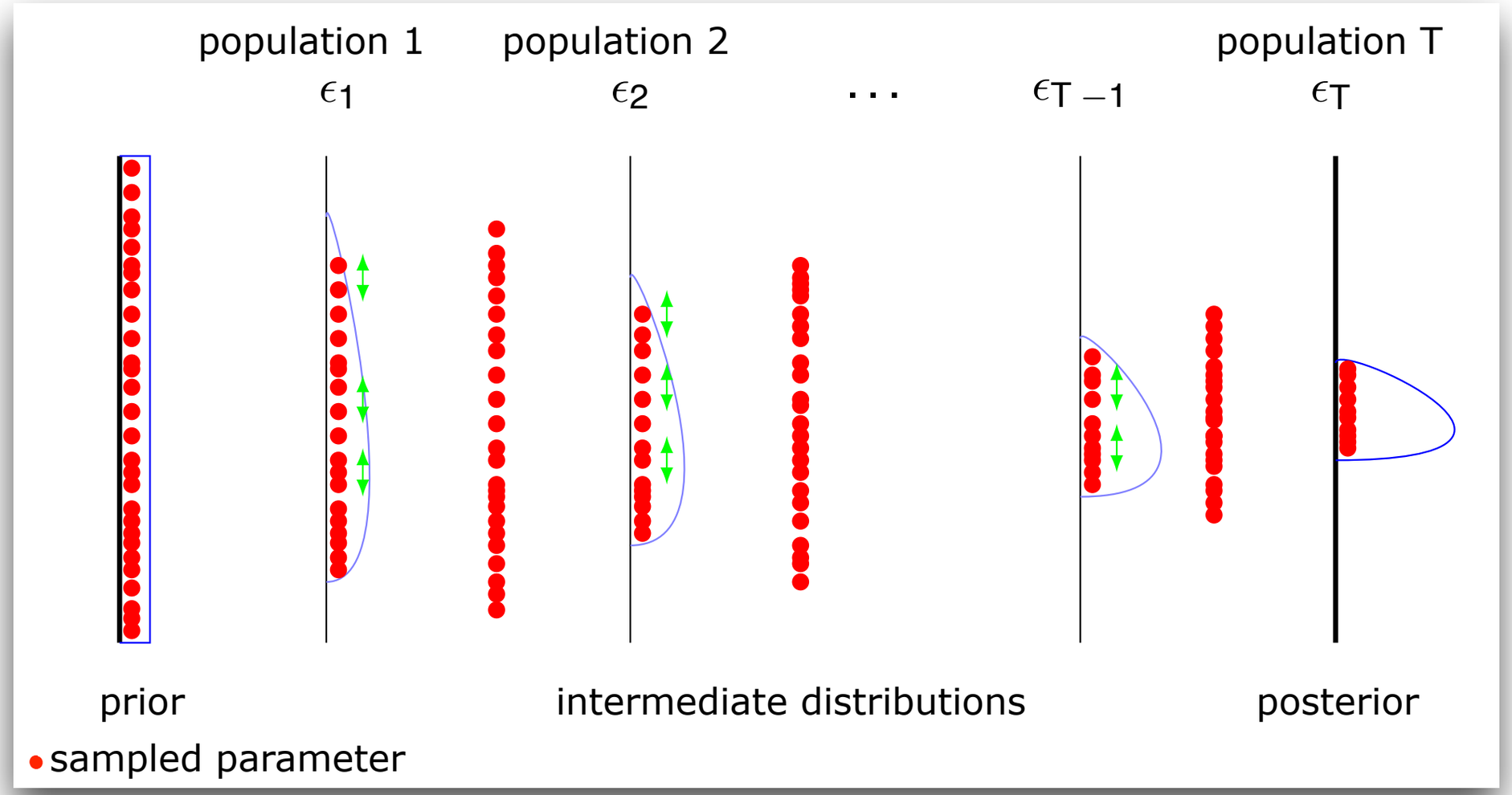
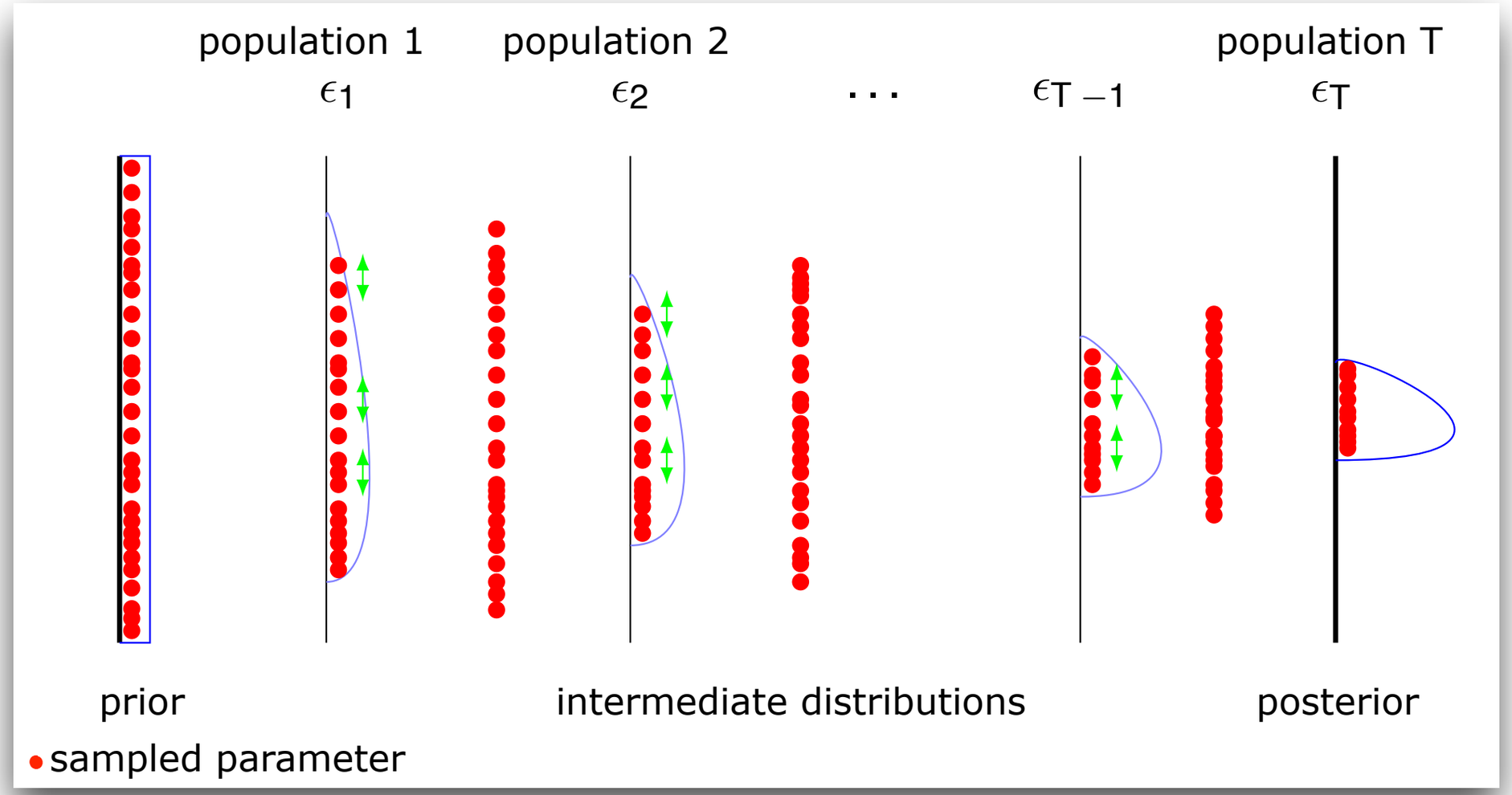


Figure modified from Toni et al., Bioinformatics (2010)

Approximate Bayesian Computation with Sequential Monte Carlo (ABC SMC)



Accept parameter samples with decreasing tolerance

Figure modified from Toni et al., Bioinformatics (2010)

ABC SMC for single-cell snapshot data

ABC SMC for single-cell snapshot data

- ABC SMC has been used to analyse flow cytometry (snapshot) data [1]

1 Lillacci et al., Bioinformatics (2013)

ABC SMC for single-cell snapshot data

- ABC SMC has been used to analyse flow cytometry (snapshot) data [1]
- Independent measurements for different time points

1 Lillacci et al., Bioinformatics (2013)

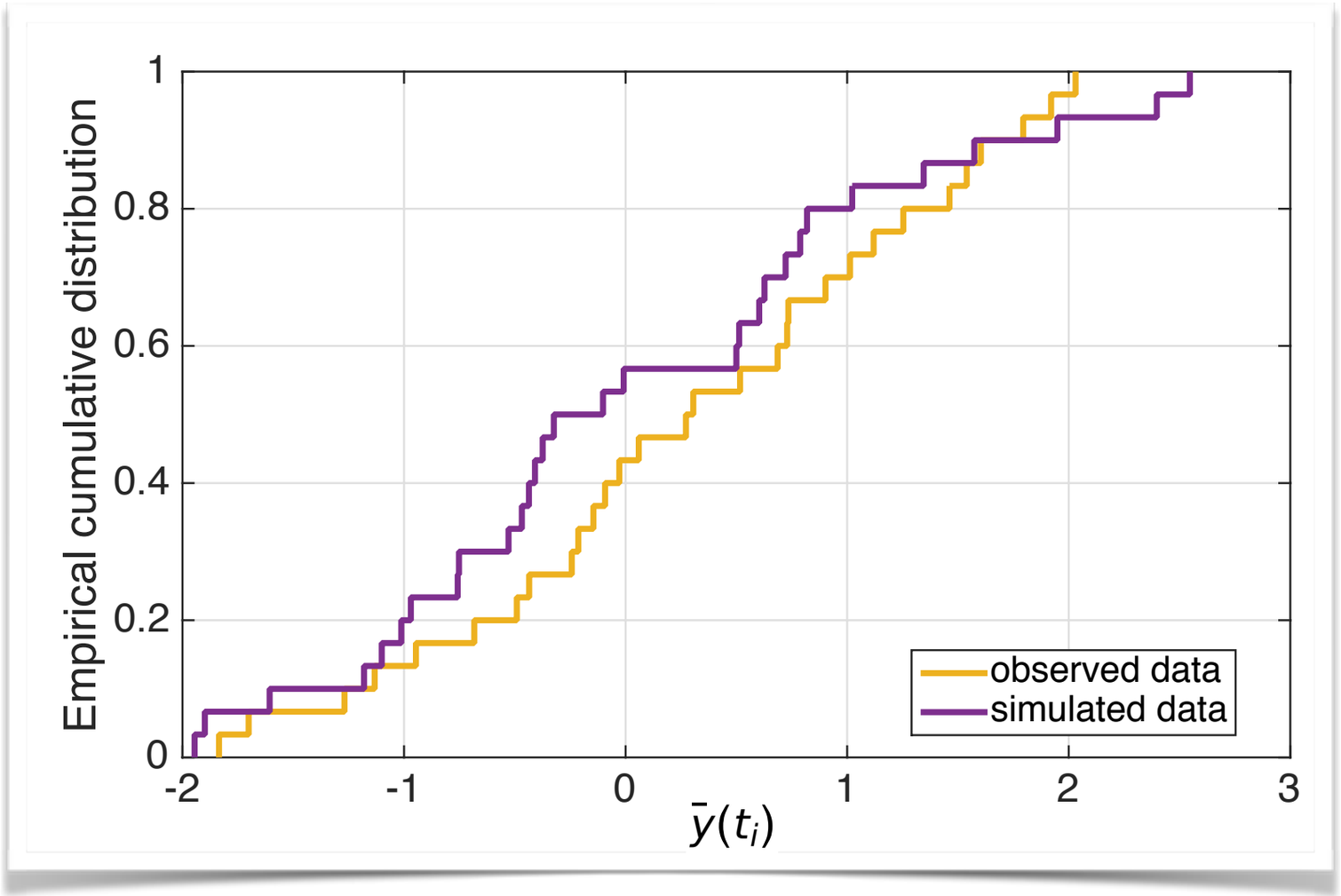
ABC SMC for single-cell snapshot data

- ABC SMC has been used to analyse flow cytometry (snapshot) data [1]
- Independent measurements for different time points
- Maximal Kolmogorov-Smirnov (KS) distance for every time point

1 Lillacci et al., Bioinformatics (2013)

ABC SMC for single-cell snapshot data

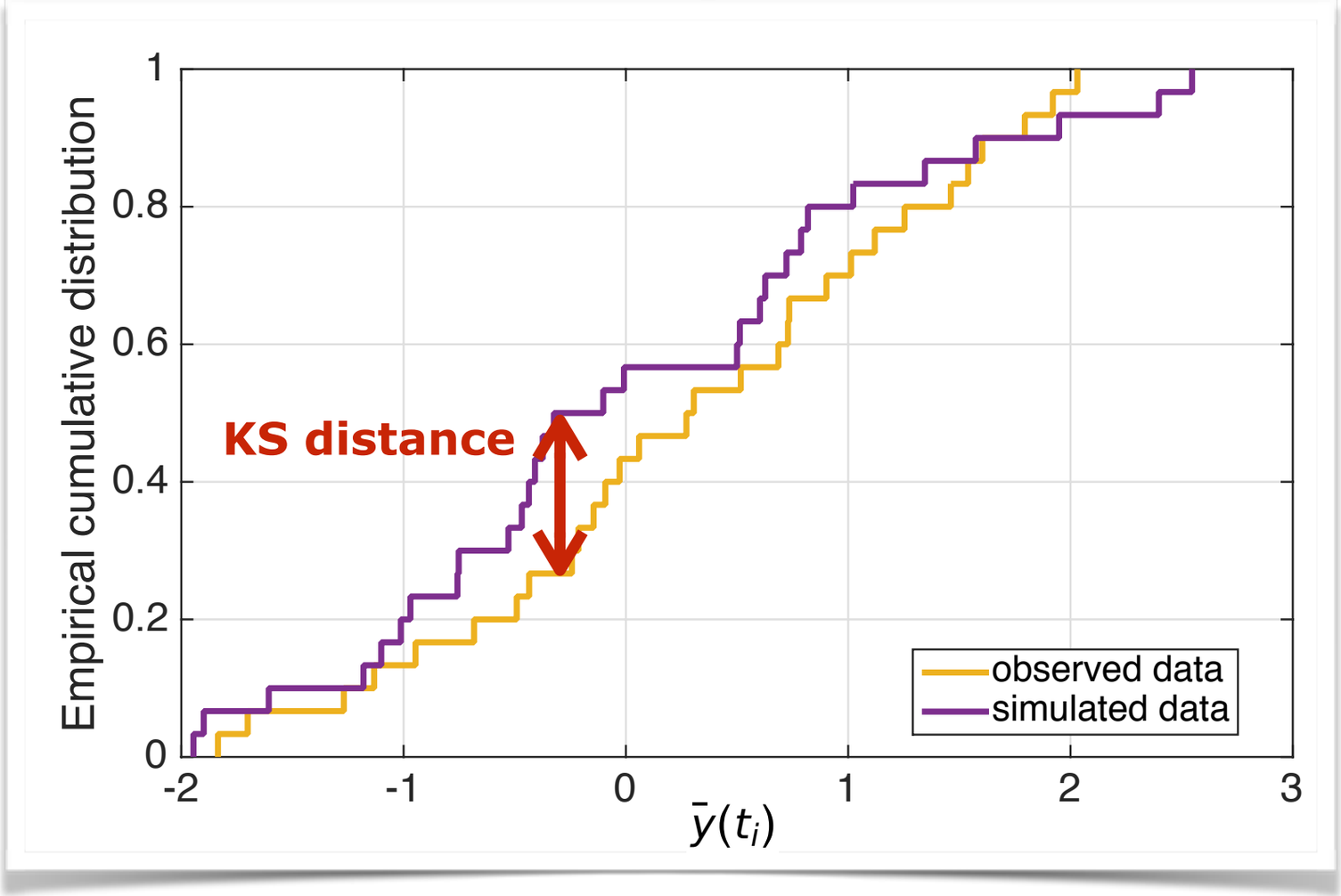
- ABC SMC has been used to analyse flow cytometry (snapshot) data [1]
- Independent measurements for different time points
- Maximal Kolmogorov-Smirnov (KS) distance for every time point



1 Lillacci et al., Bioinformatics (2013)

ABC SMC for single-cell snapshot data

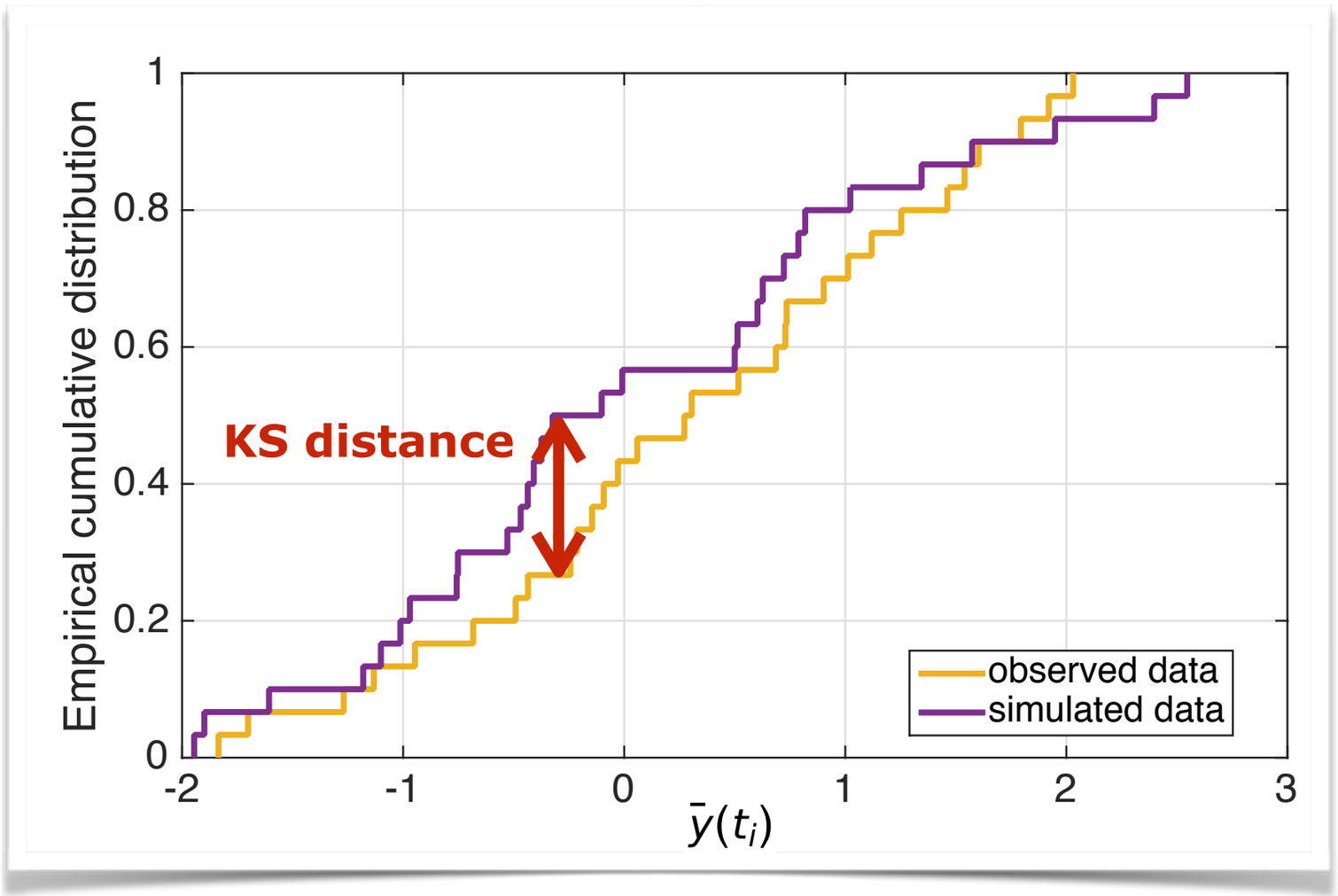
- ABC SMC has been used to analyse flow cytometry (snapshot) data [1]
- Independent measurements for different time points
- Maximal Kolmogorov-Smirnov (KS) distance for every time point



1 Lillacci et al., Bioinformatics (2013)

ABC SMC for single-cell snapshot data

- ABC SMC has been used to analyse flow cytometry (snapshot) data [1]
- Independent measurements for different time points
- Maximal Kolmogorov-Smirnov (KS) distance for every time point



Accept parameters that produce a low KS distance

1 Lillacci et al., Bioinformatics (2013)

ABC SMC with multivariate test statistics

ABC SMC with multivariate test statistics

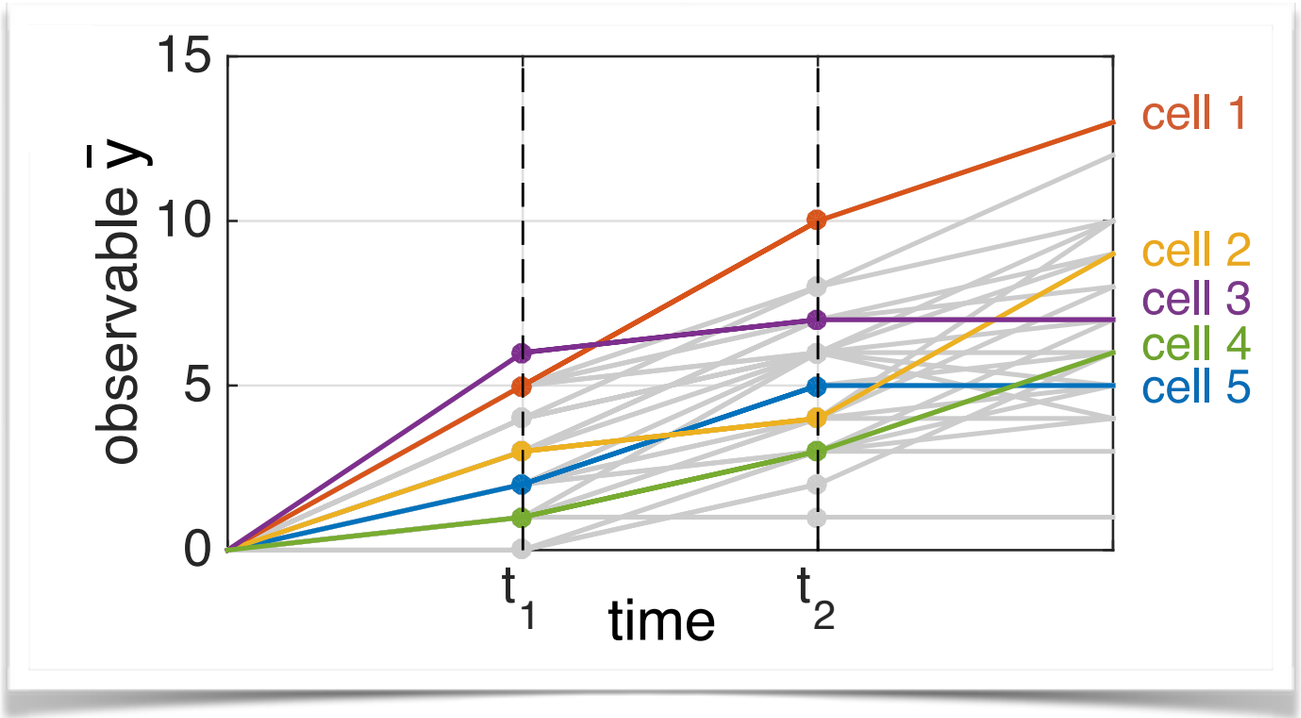
- Performance and convergence of ABC highly depends on the employed distance measure

ABC SMC with multivariate test statistics

- Performance and convergence of ABC highly depends on the employed distance measure
- Project single-cell time-series into high dimensional space

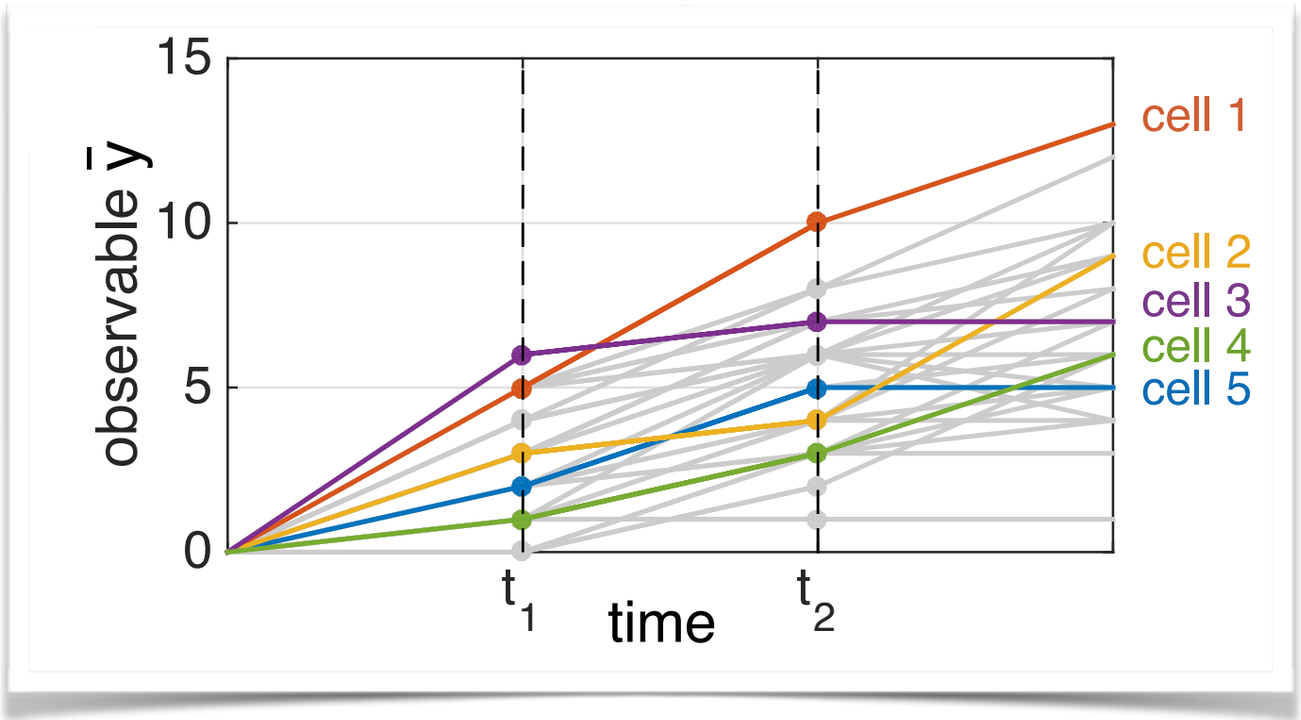
ABC SMC with multivariate test statistics

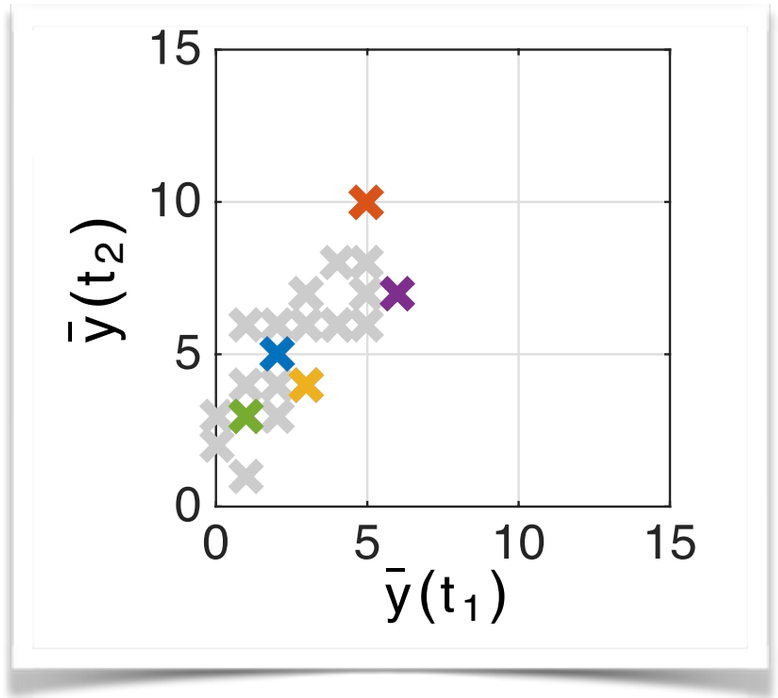
- Performance and convergence of ABC highly depends on the employed distance measure
- Project single-cell time-series into high dimensional space



ABC SMC with multivariate test statistics

- Performance and convergence of ABC highly depends on the employed distance measure
- Project single-cell time-series into high dimensional space

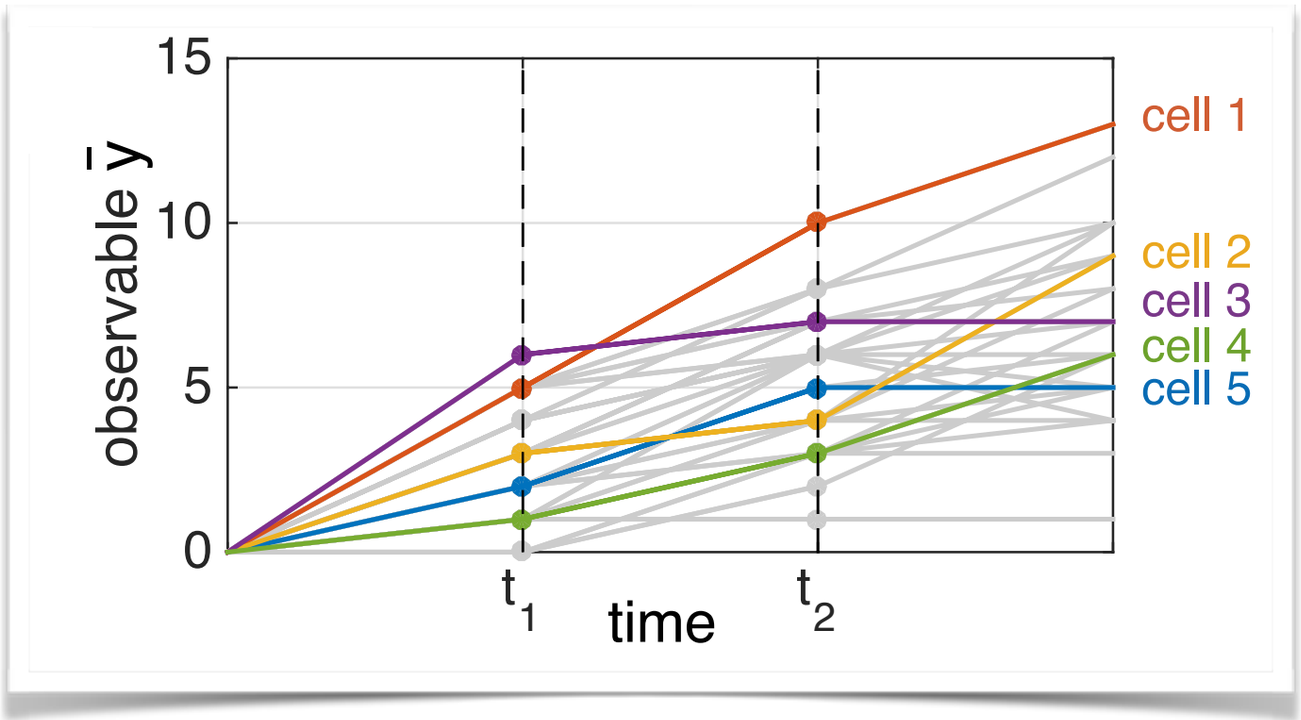


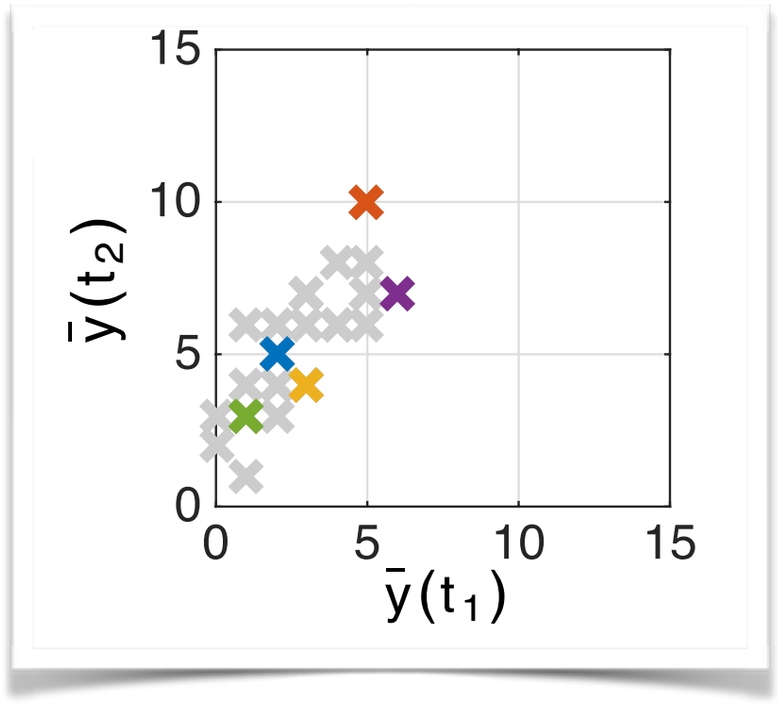


single-cell trajectory
measurement
two-dimensional representation of the trajectory

ABC SMC with multivariate test statistics

- Performance and convergence of ABC highly depends on the employed distance measure
- Project single-cell time-series into high dimensional space





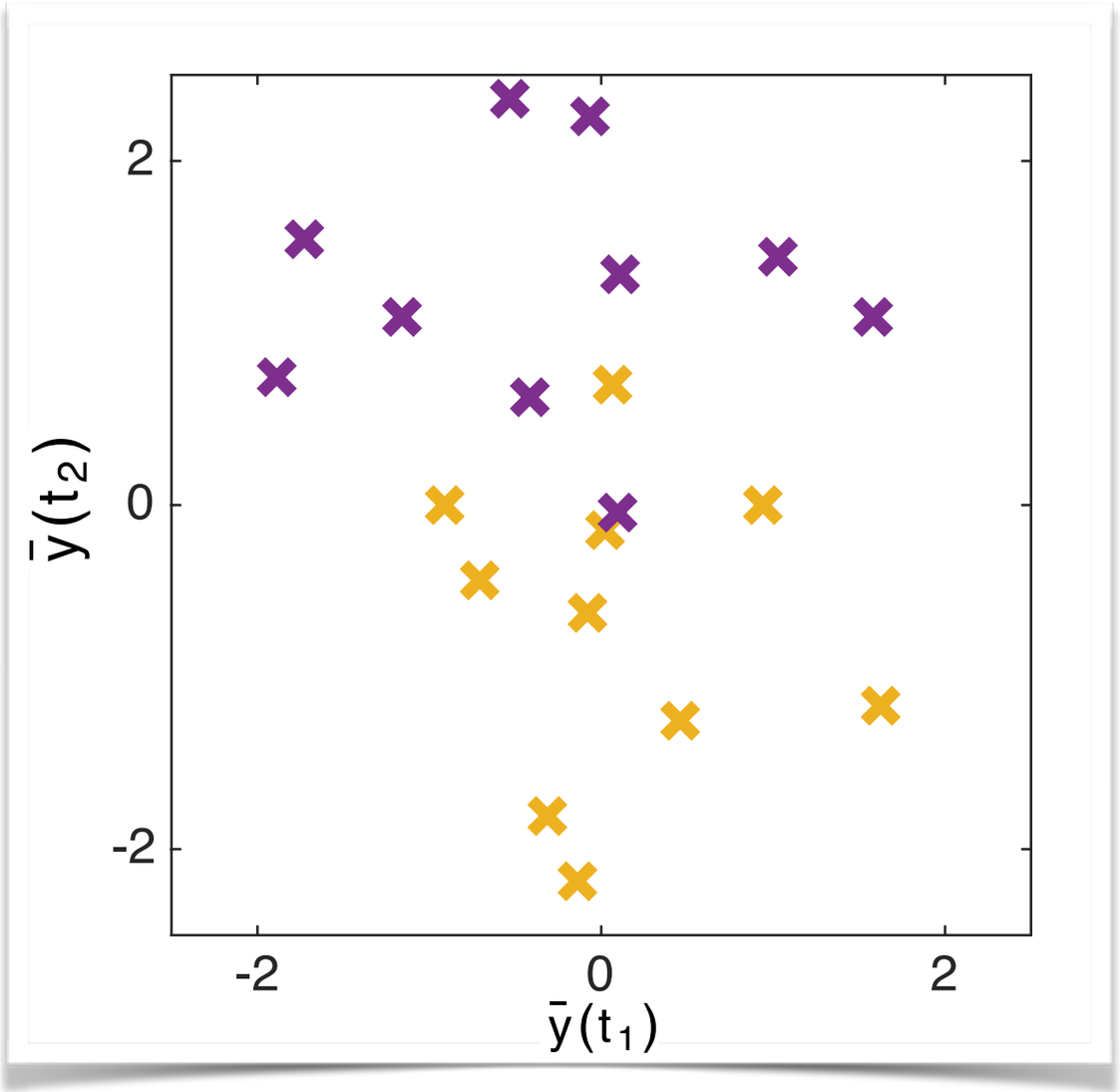
single-cell trajectory
measurement
two-dimensional representation of the trajectory

Use multivariate statistics on the distribution of single-cell trajectories in the rejection step of ABC SMC

Cross-match test

Cross-match test

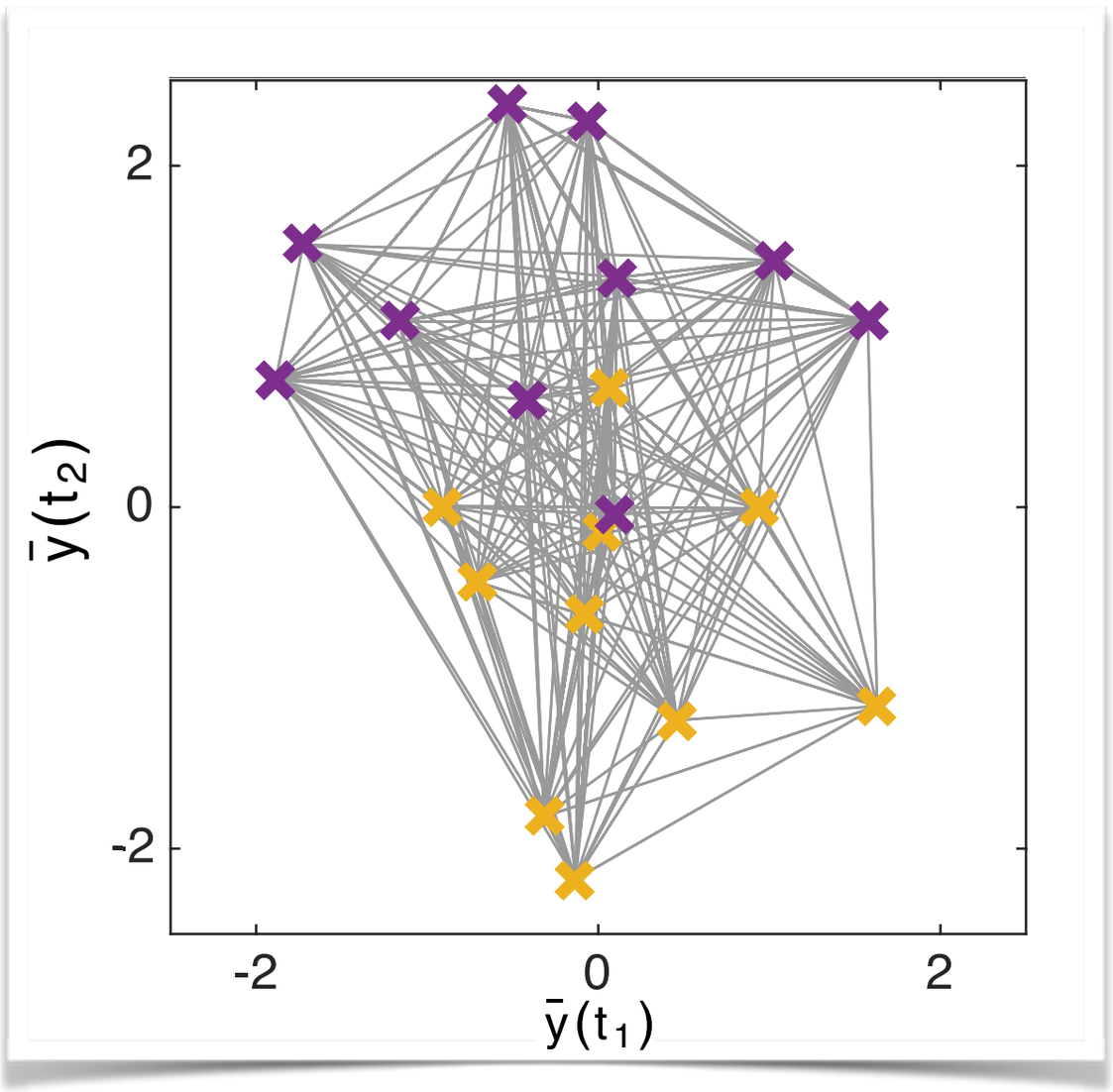
Cross-match test



observed samples
simulated samples

Rosenbaum, Journal of the Royal Statistical Society (2005)

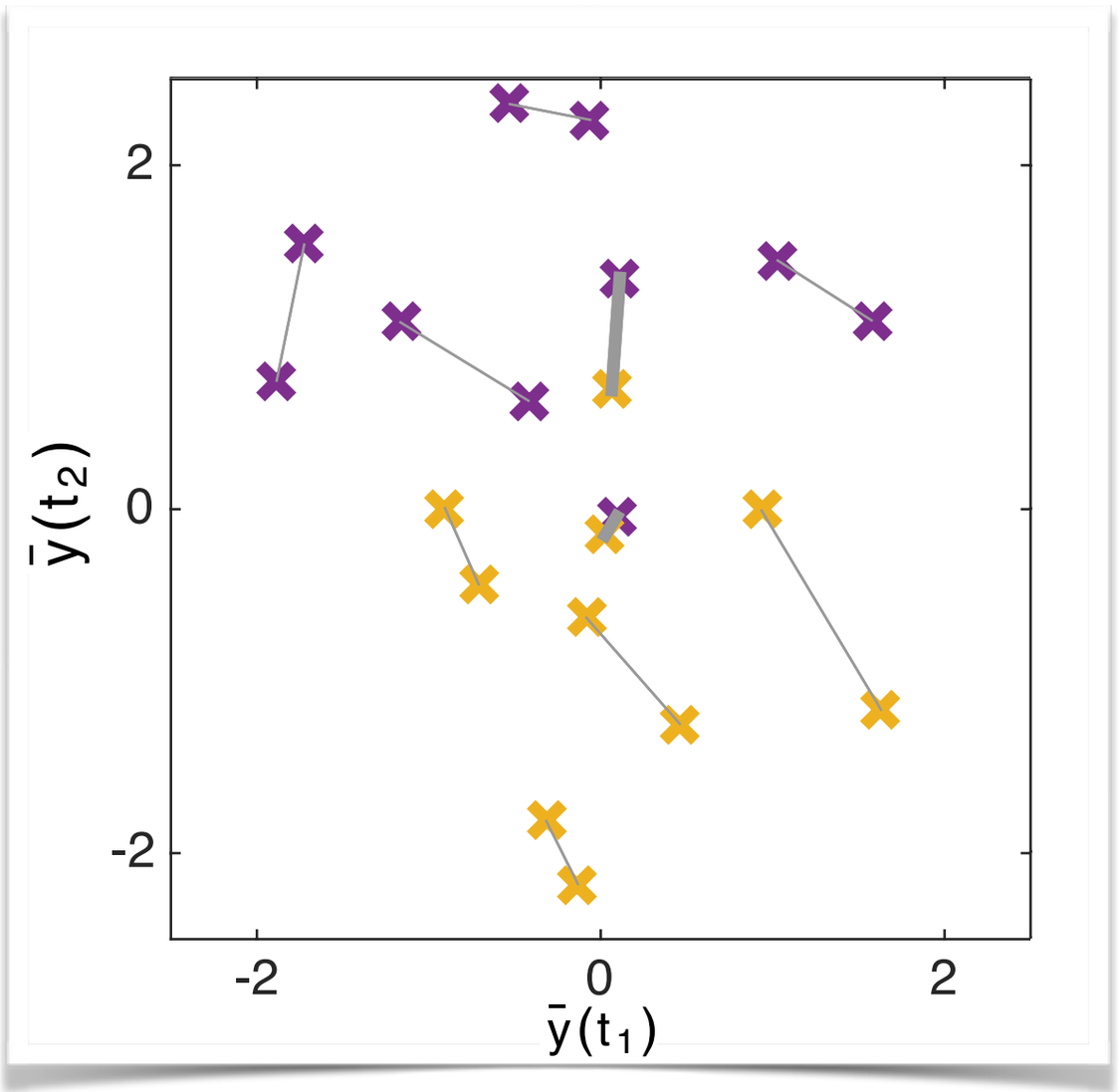
Cross-match test



observed samples
simulated samples

Perform a minimum weight non-bipartite matching on a complete graph

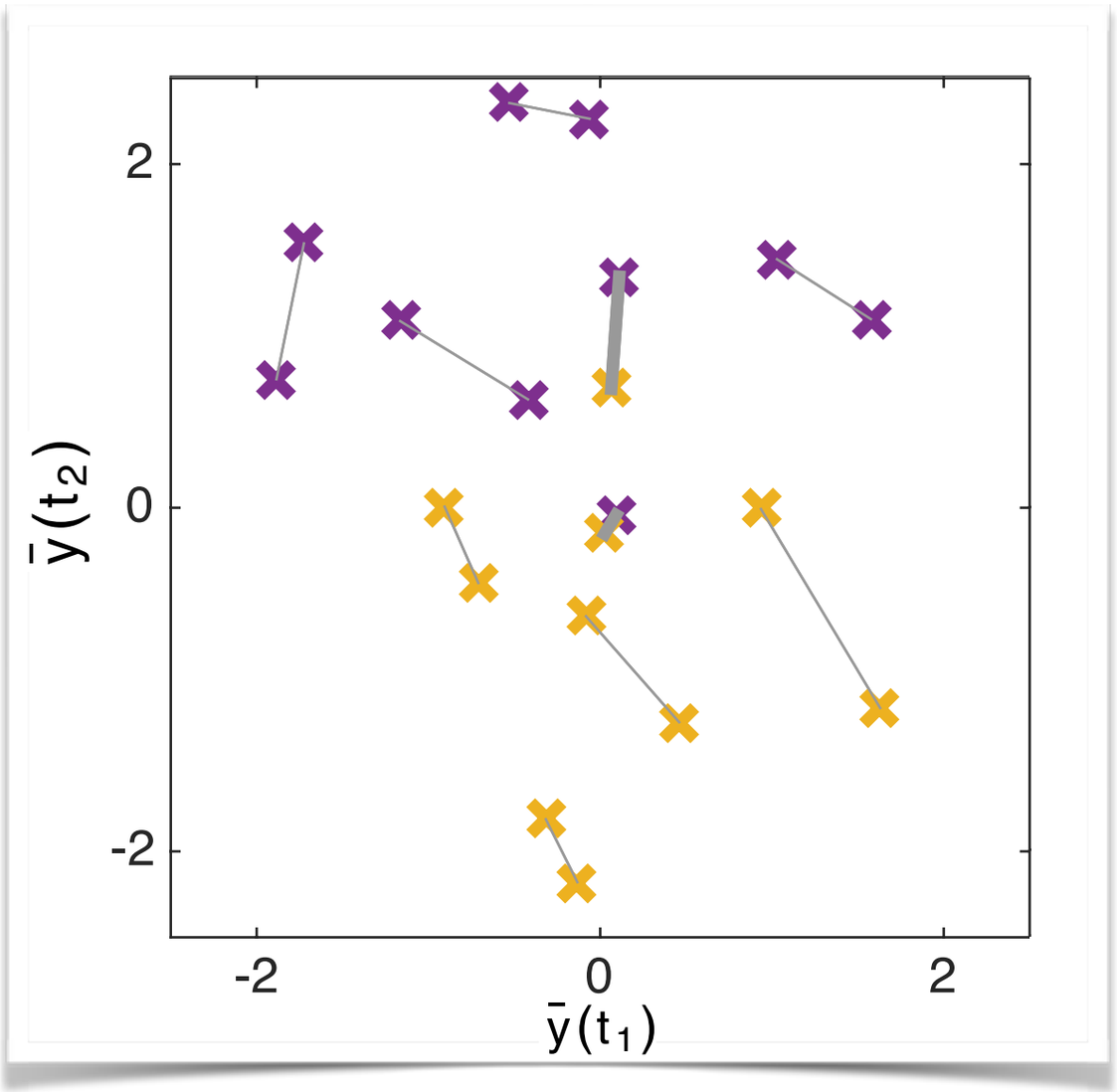
Cross-match test

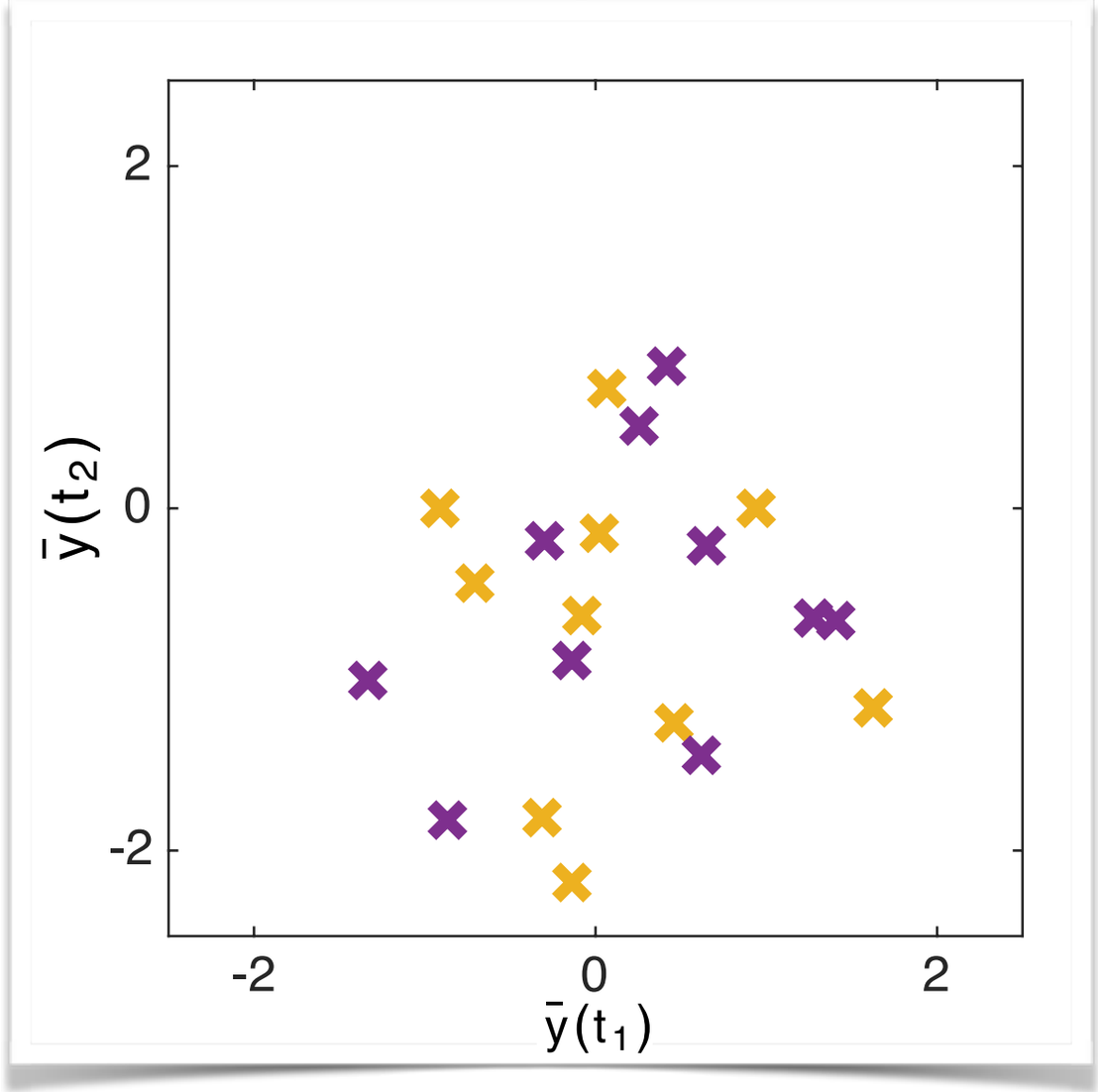


observed samples
simulated samples
no cross-match
cross-matches

Perform a minimum weight non-bipartite matching on a complete graph

Cross-match test



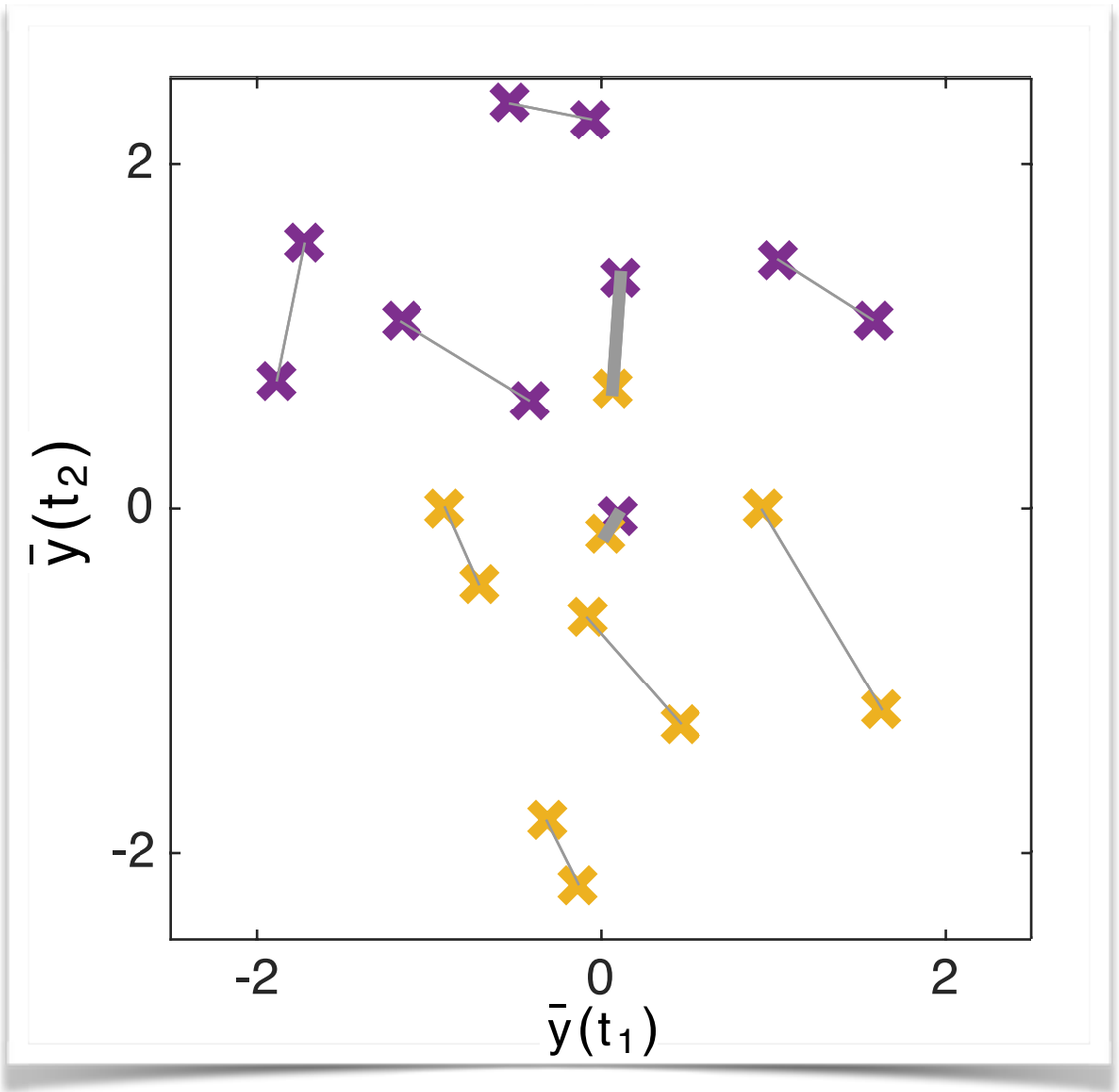


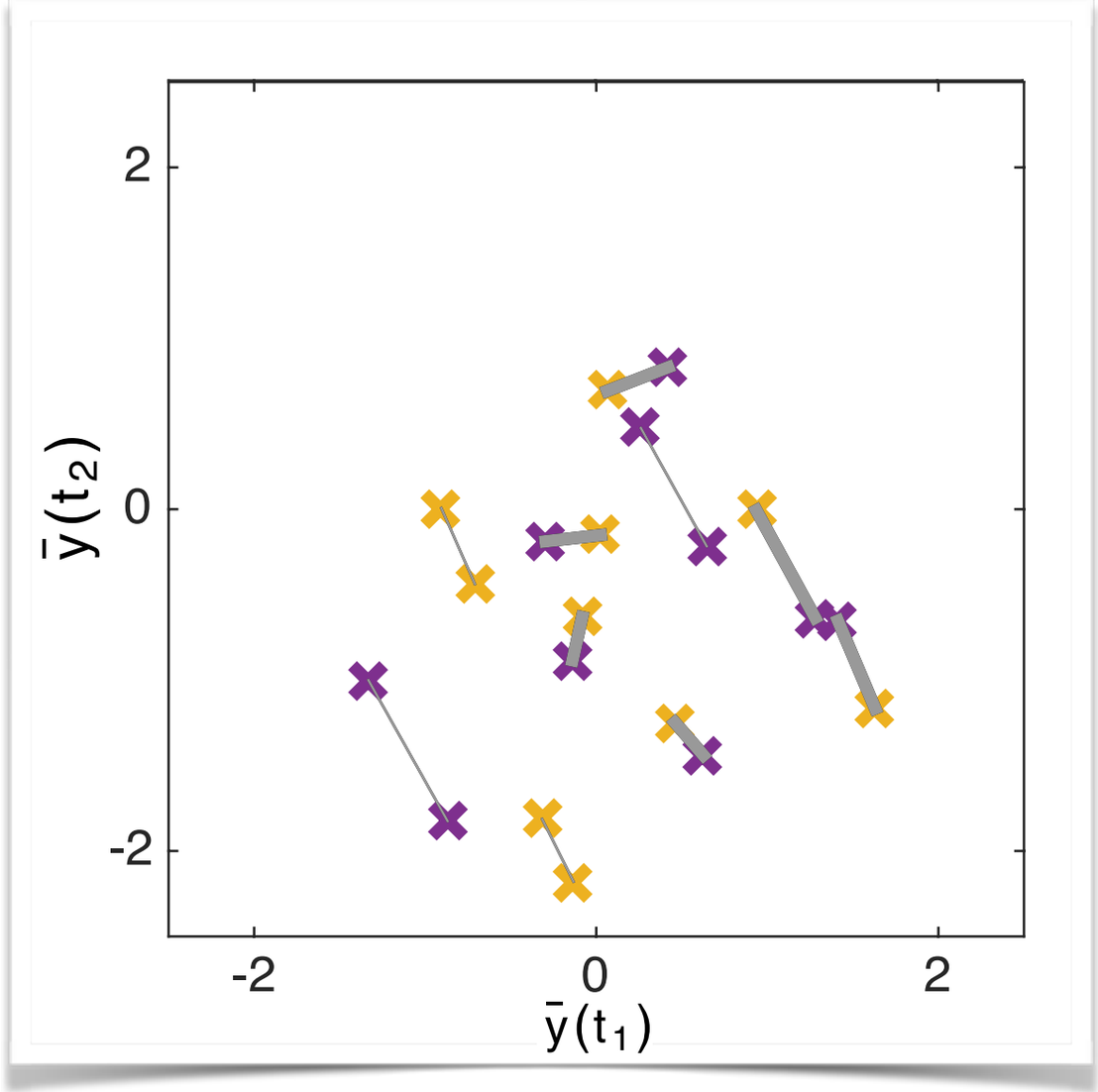
observed samples
simulated samples
no cross-match
cross-matches

Perform a minimum weight non-bipartite matching on a complete graph

Rosenbaum, Journal of the Royal Statistical Society (2005)

Cross-match test



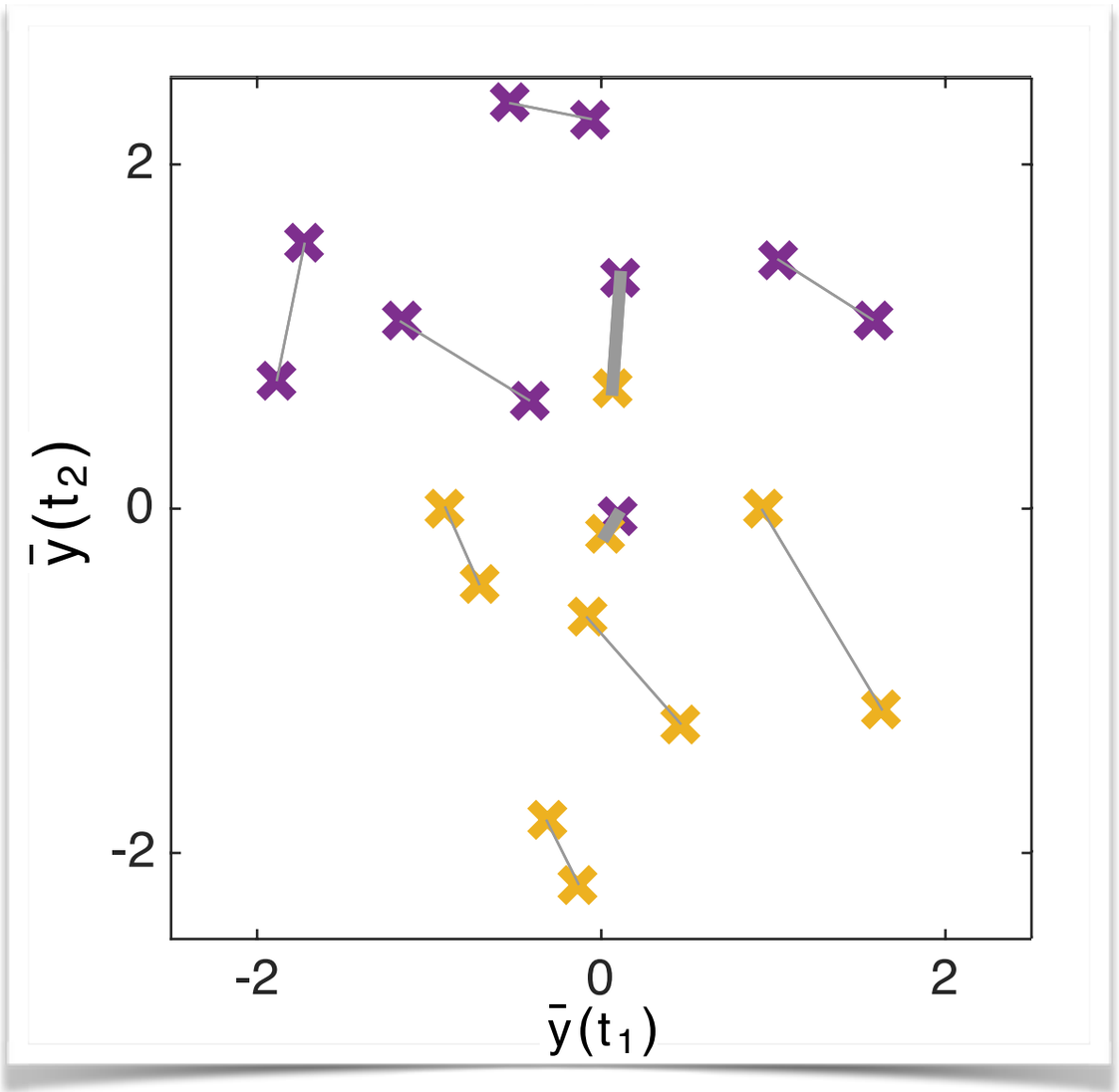


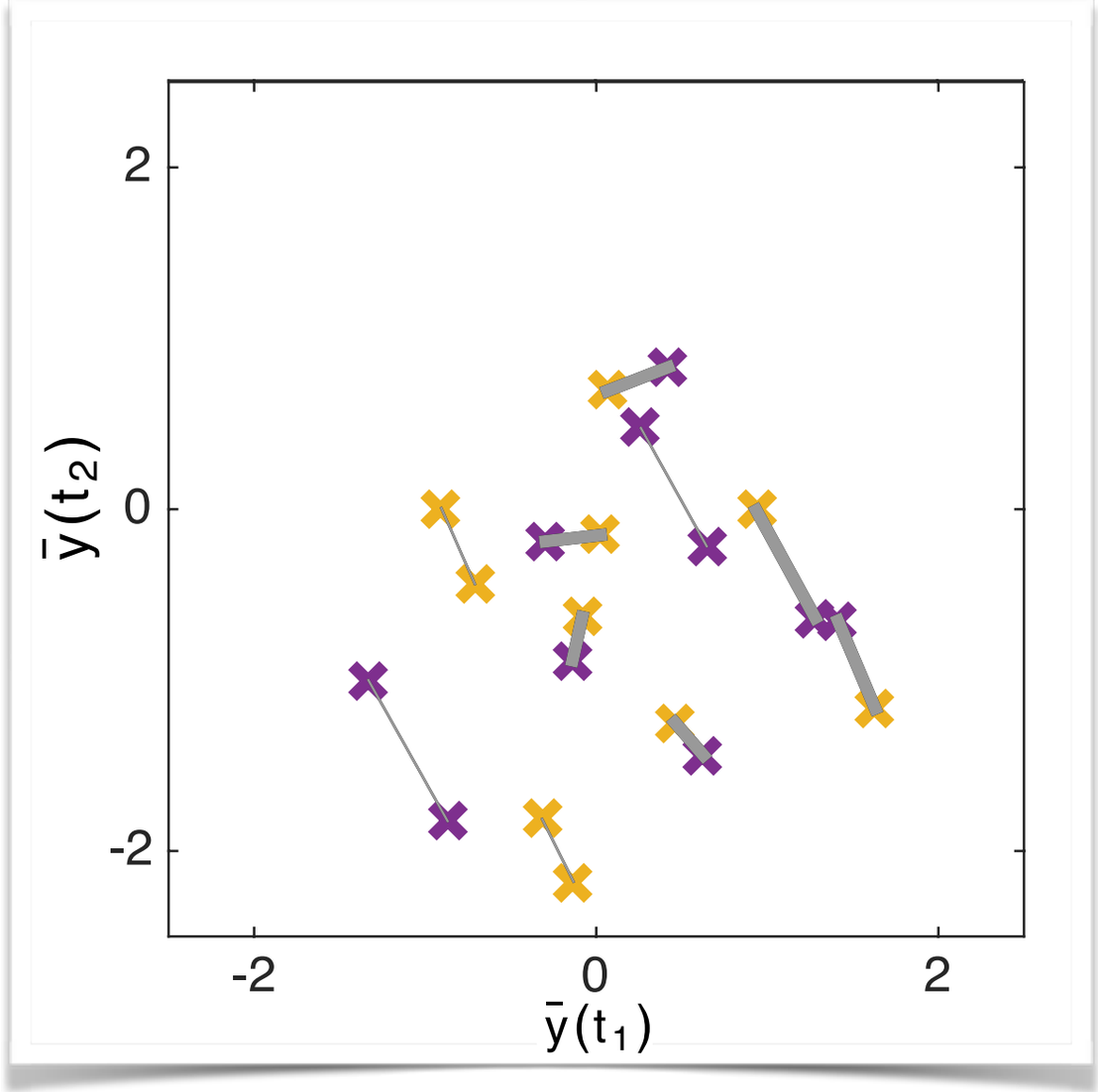
observed samples
simulated samples
no cross-match
cross-matches

Perform a minimum weight non-bipartite matching on a complete graph

Rosenbaum, Journal of the Royal Statistical Society (2005)

Cross-match test





observed samples
simulated samples
no cross-match
cross-matches

Perform a minimum weight non-bipartite matching on a complete graph

Accept parameters that produce more cross-matches

Rosenbaum, Journal of the Royal Statistical Society (2005)

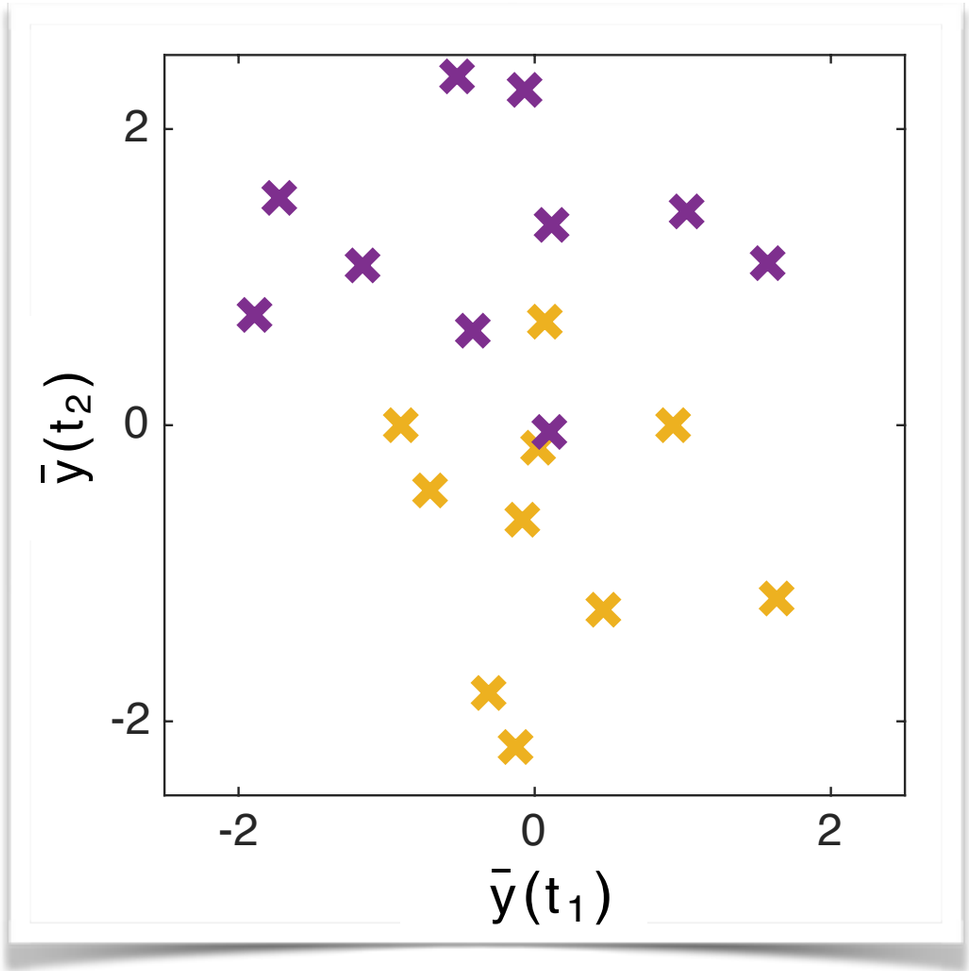
Maximum Mean Discrepancy (MMD)

Maximum Mean Discrepancy (MMD)

$$\text{MMD} = \left(\frac{1}{n^2} \sum_{i \neq j}^{n} k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{m^2} \sum_{i \neq j}^{m} k(\mathbf{y}_i, \mathbf{y}_j) - \frac{2}{nm} \sum_{i,j=1}^{n,m} k(\mathbf{x}_i, \mathbf{y}_j) \right)^{\frac{1}{2}}$$

Maximum Mean Discrepancy (MMD)

$$\mathrm{MMD} = \left(\frac{1}{n^2} \sum_{i \neq j}^{n} k(\mathrm{x}_i, \mathrm{x}_j) + \frac{1}{m^2} \sum_{i \neq j}^{m} k(\mathrm{y}_i, \mathrm{y}_j) - \frac{2}{nm} \sum_{i,j=1}^{n,m} k(\mathrm{x}_i, \mathrm{y}_j) \right)^{\frac{1}{2}}$$

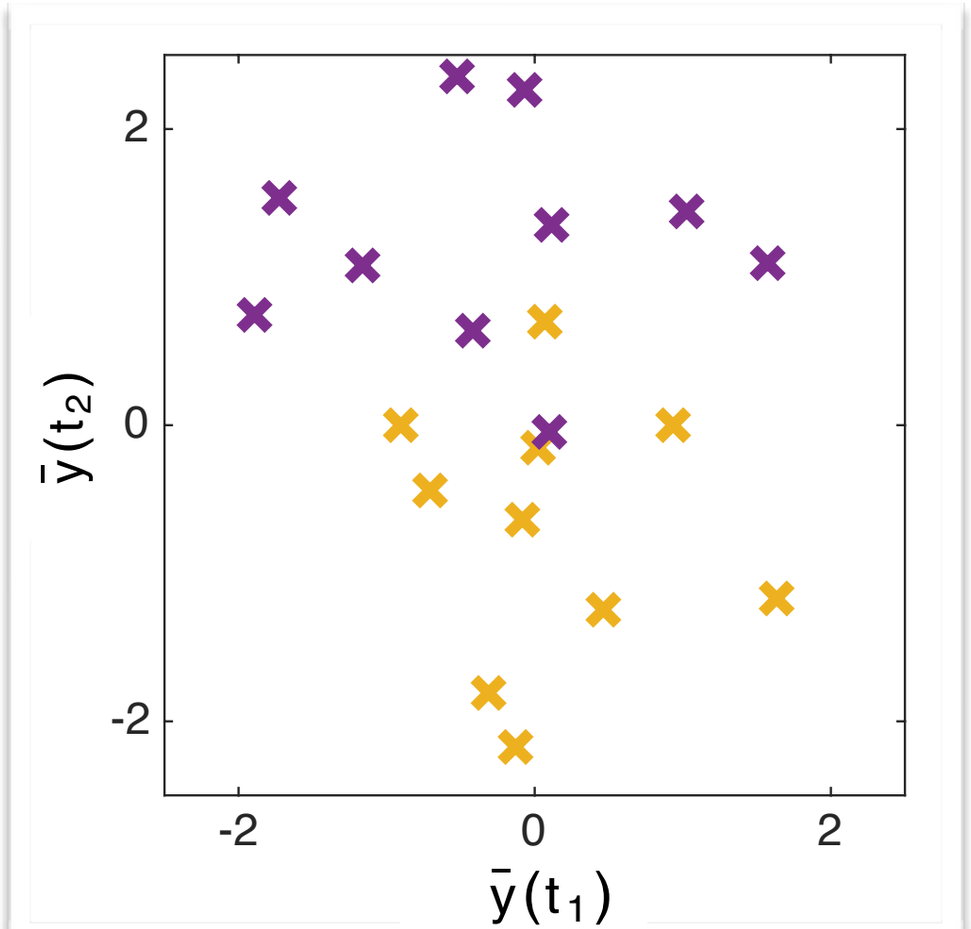


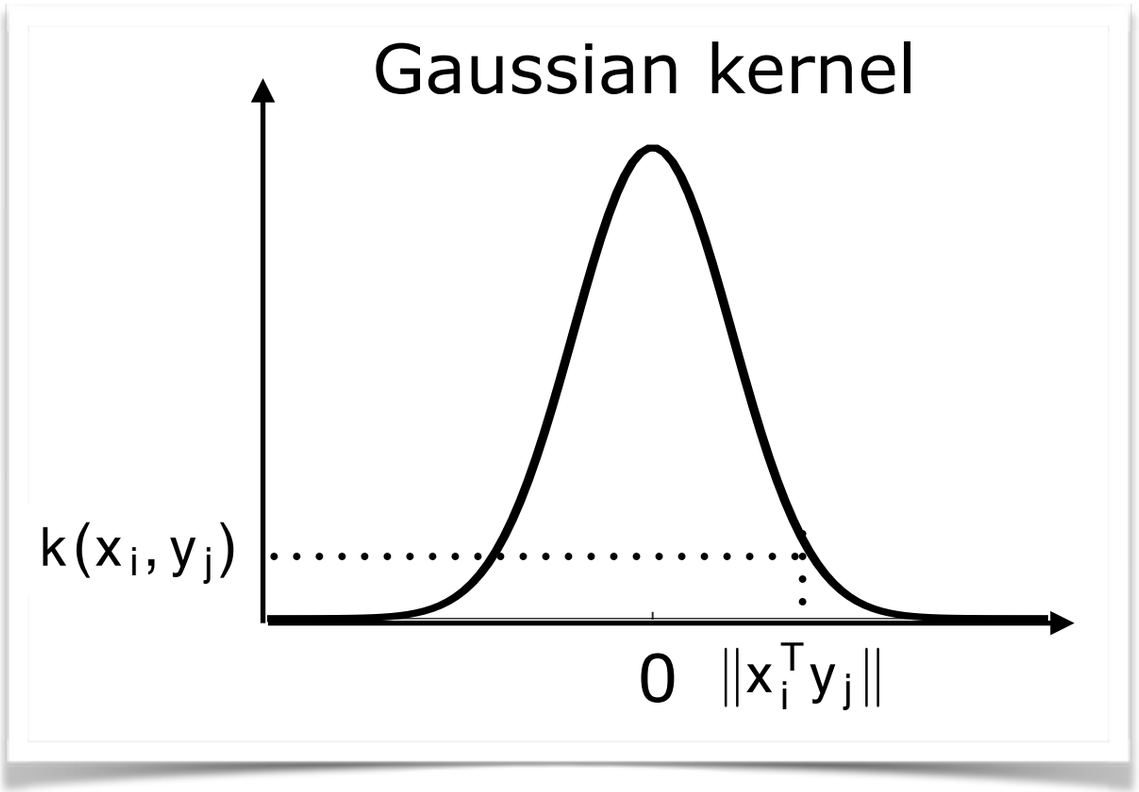
- observed samples $x_1, \ldots, x_n$
- simulated samples $y_1, \ldots, y_m$

Borgwardt et al., Bioinformatics (2006)

Maximum Mean Discrepancy (MMD)

$$\mathrm{MMD} = \left(\frac{1}{n^2} \sum_{i \neq j}^{n} k(\mathrm{x}_i, \mathrm{x}_j) + \frac{1}{m^2} \sum_{i \neq j}^{m} k(\mathrm{y}_i, \mathrm{y}_j) - \frac{2}{nm} \sum_{i,j=1}^{n,m} k(\mathrm{x}_i, \mathrm{y}_j) \right)^{\frac{1}{2}}$$



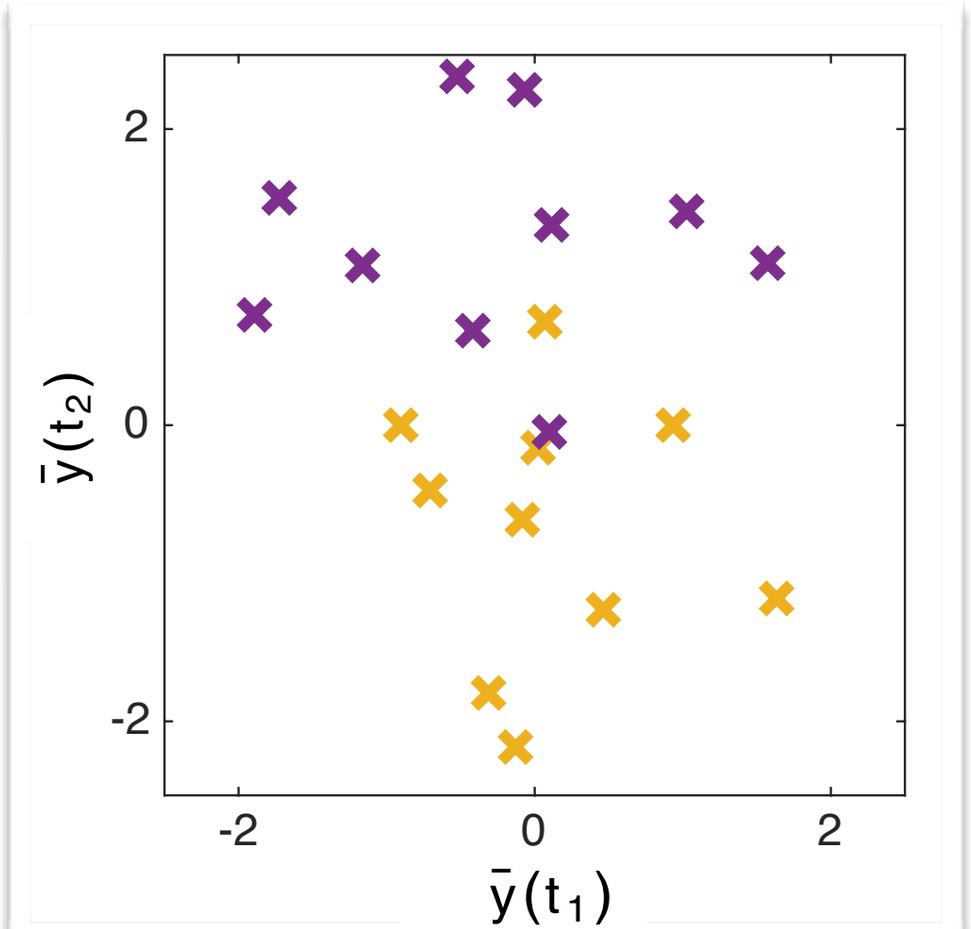


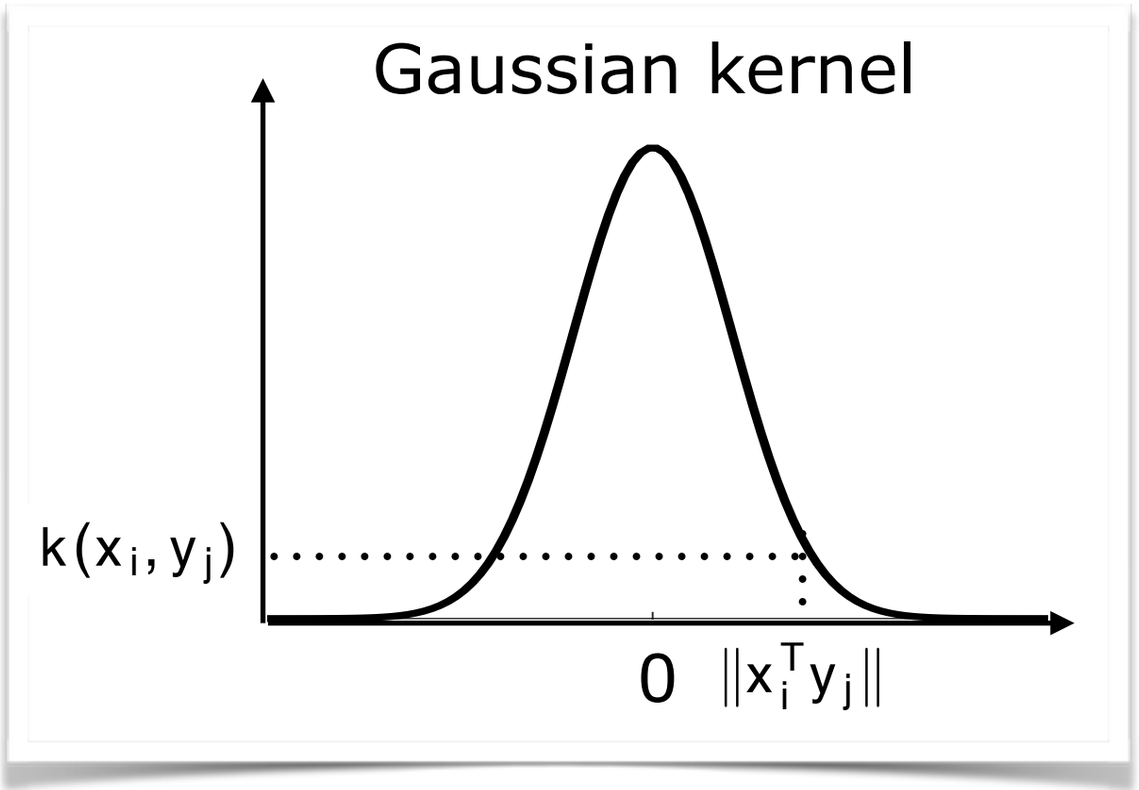
observed samples $x_1, \ldots, x_n$
simulated samples $y_1, \ldots, y_m$

Borgwardt et al., Bioinformatics (2006)

Maximum Mean Discrepancy (MMD)

$$\text{MMD} = \left(\frac{1}{n^2} \sum_{i \neq j}^{n} k(\mathrm{x}_i, \mathrm{x}_j) + \frac{1}{m^2} \sum_{i \neq j}^{m} k(\mathrm{y}_i, \mathrm{y}_j) - \frac{2}{nm} \sum_{i,j=1}^{n,m} k(\mathrm{x}_i, \mathrm{y}_j) \right)^{\frac{1}{2}}$$





- observed samples $x_1, \ldots, x_n$
- simulated samples $y_1, \ldots, y_m$

Accept parameters that produce a small MMD

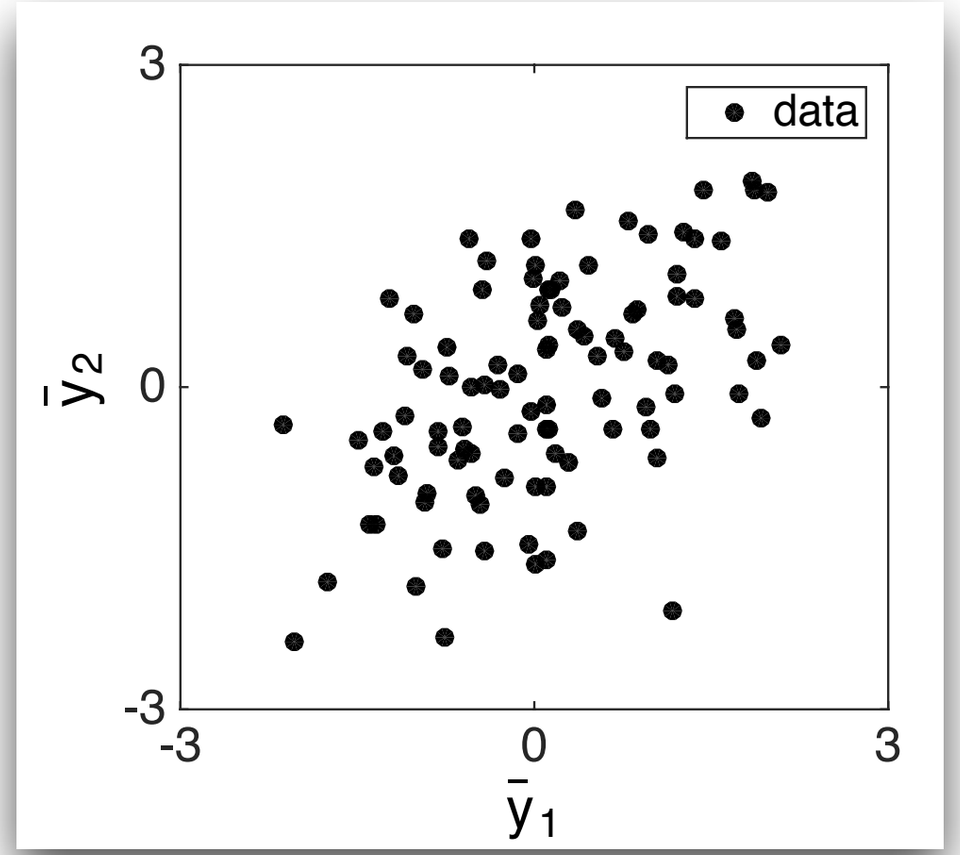
Borgwardt et al., Bioinformatics (2006)

Example: Bivariate Normal Distribution

Example: Bivariate Normal Distribution

Data: 100 samples of a bivariate normal distribution $\bar{y} \sim N(\mu, \Sigma)$

with $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_2 & \theta_1 \end{pmatrix}$

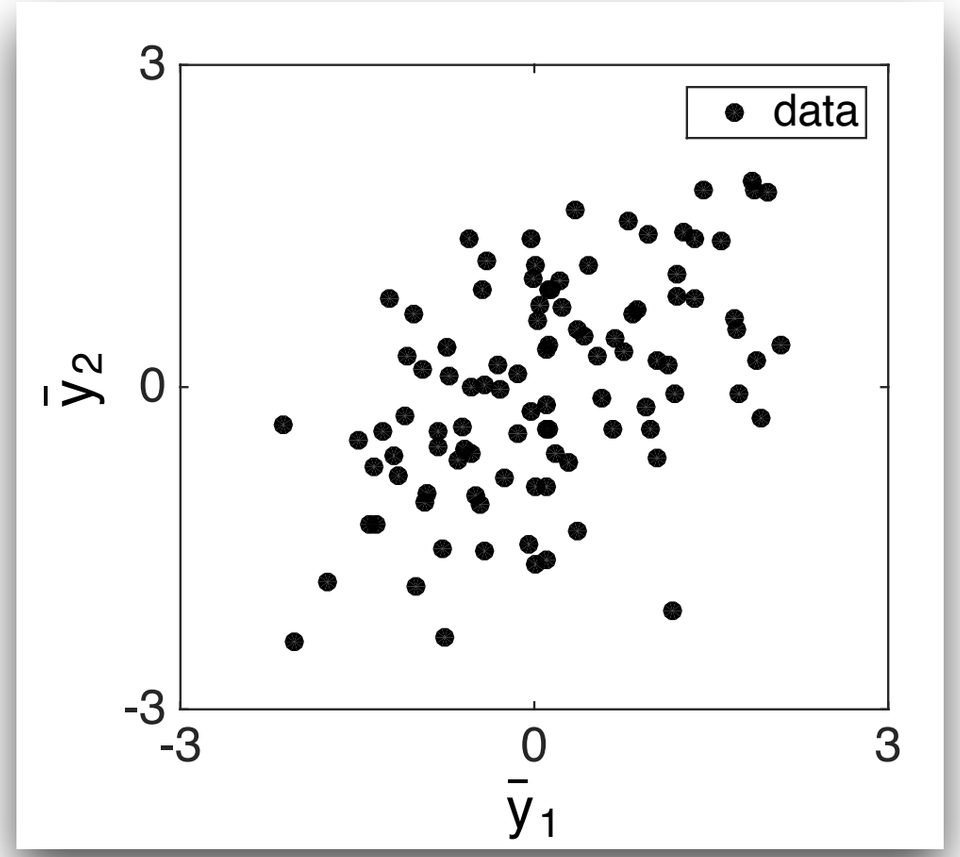


Example: Bivariate Normal Distribution

Data: 100 samples of a bivariate normal distribution $\bar{y} \sim N(\mu, \Sigma)$

with $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_2 & \theta_1 \end{pmatrix}$

Goal: Approximate posterior distributions of θ_1 and θ_2 using ABC SMC

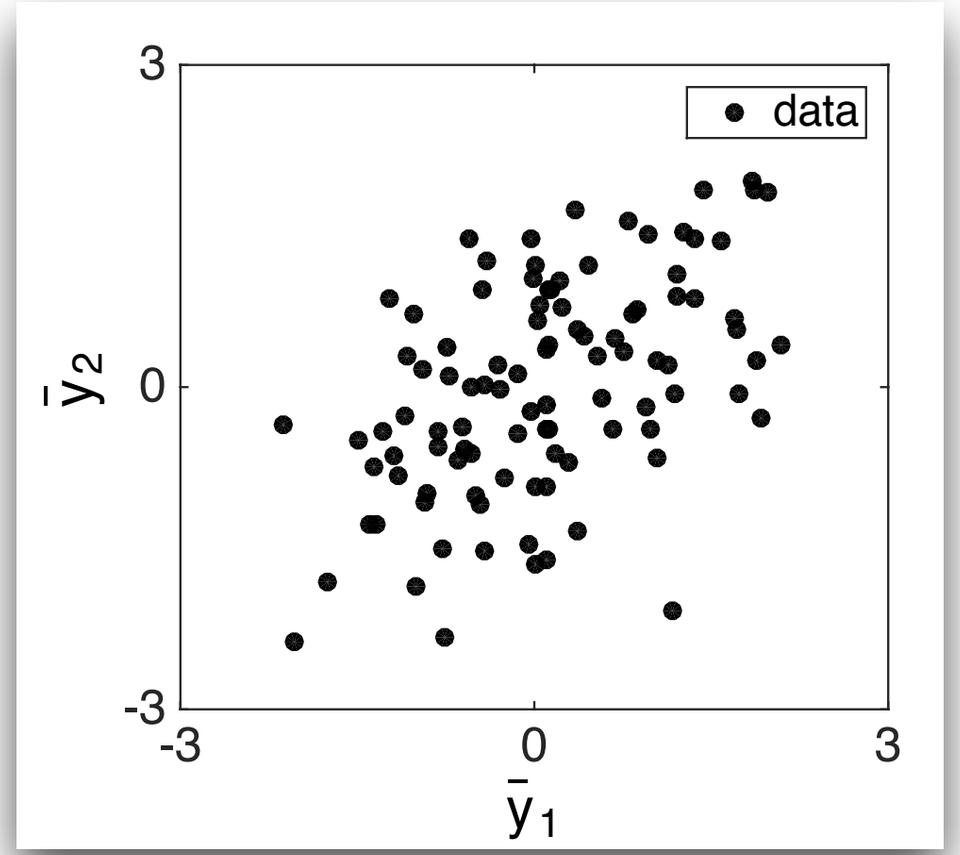


Example: Bivariate Normal Distribution

Data: 100 samples of a bivariate normal distribution $\bar{y} \sim N(\mu, \Sigma)$

with $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_2 & \theta_1 \end{pmatrix}$

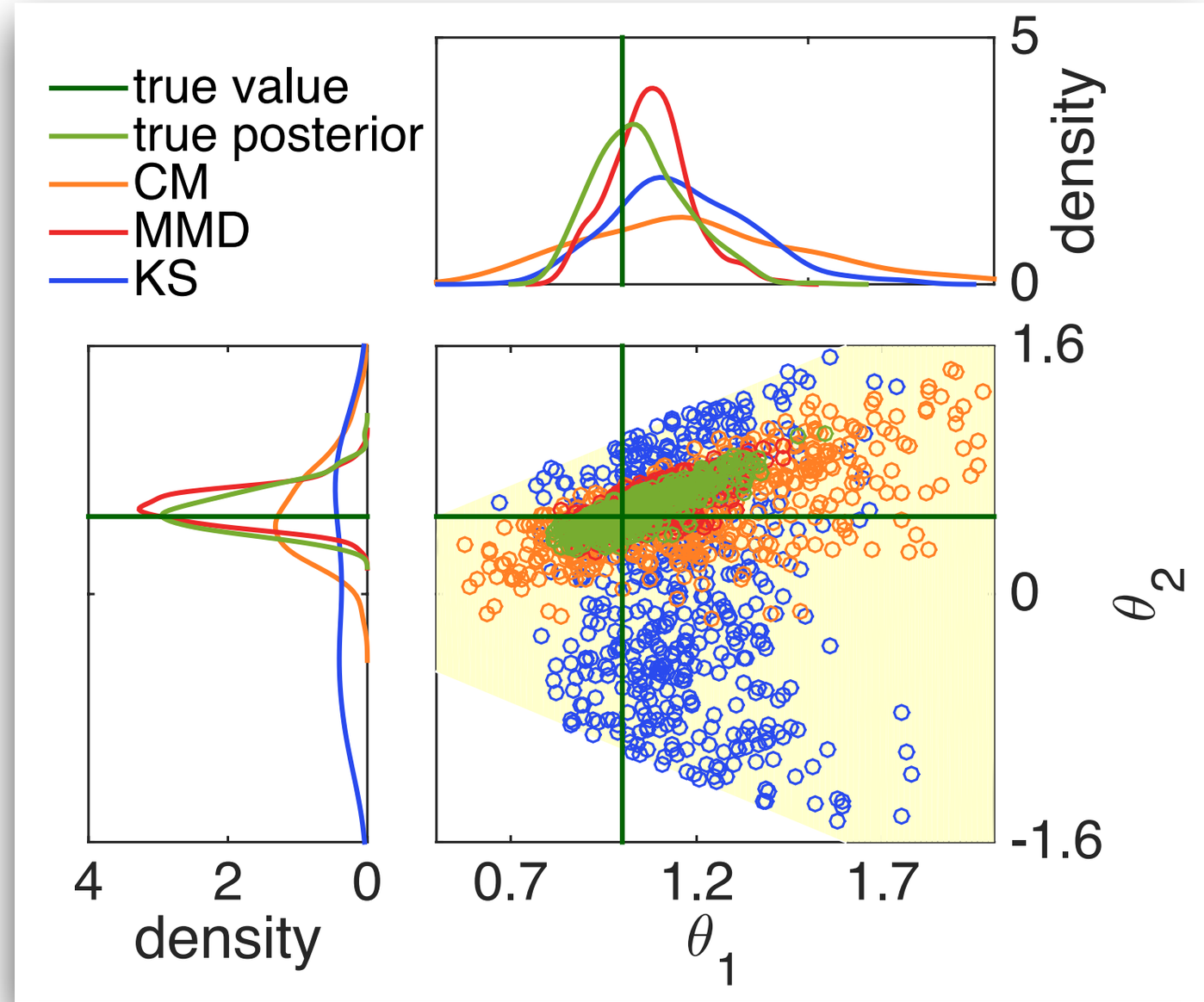
Goal: Approximate posterior distributions of θ_1 and θ_2 using ABC SMC



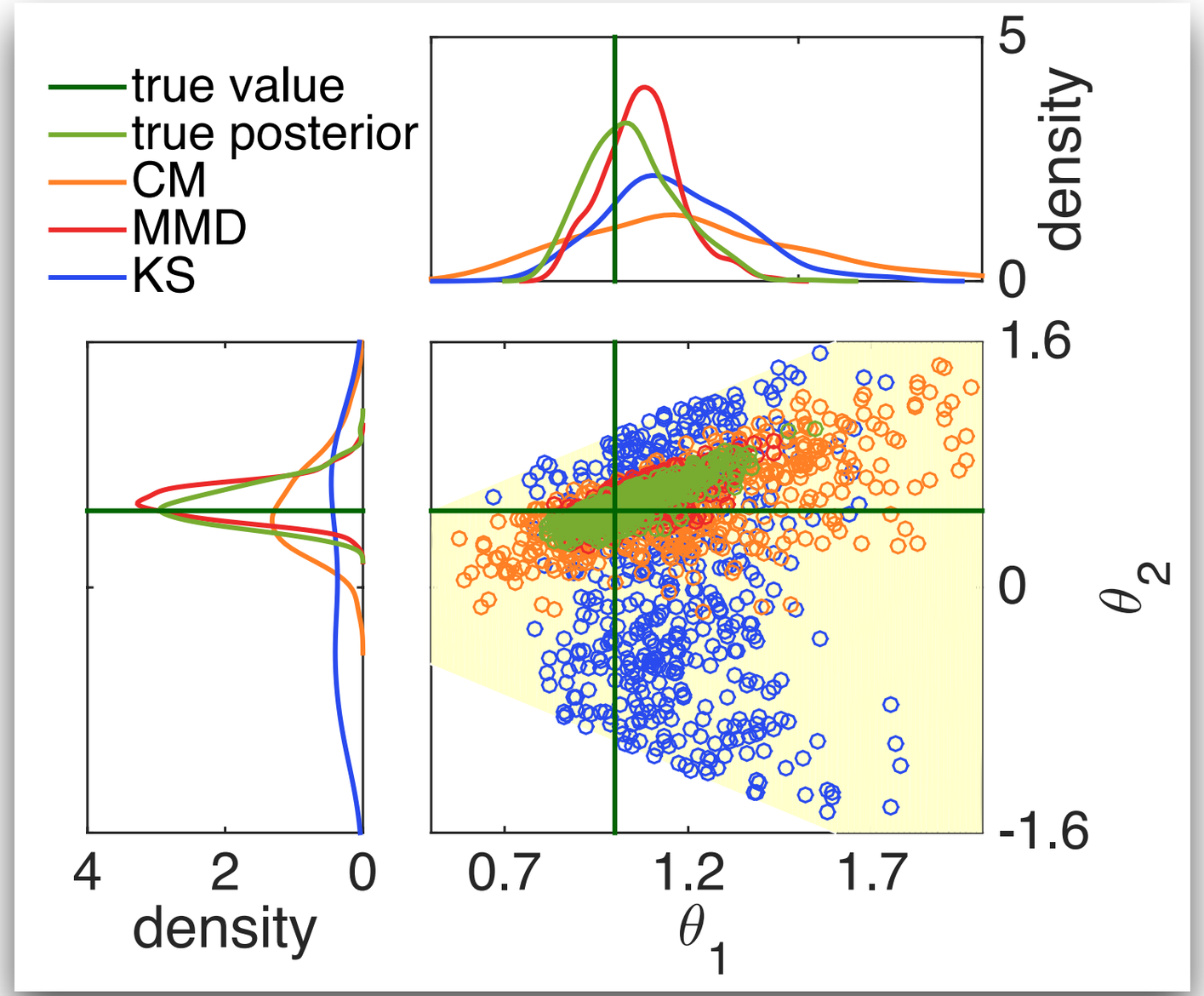
Compare ABC SMC with cross-match test (CM), maximum mean discrepancy (MMD), and maximal Kolmogorov-Smirnov distance (KS)

Posterior approximations

Posterior approximations

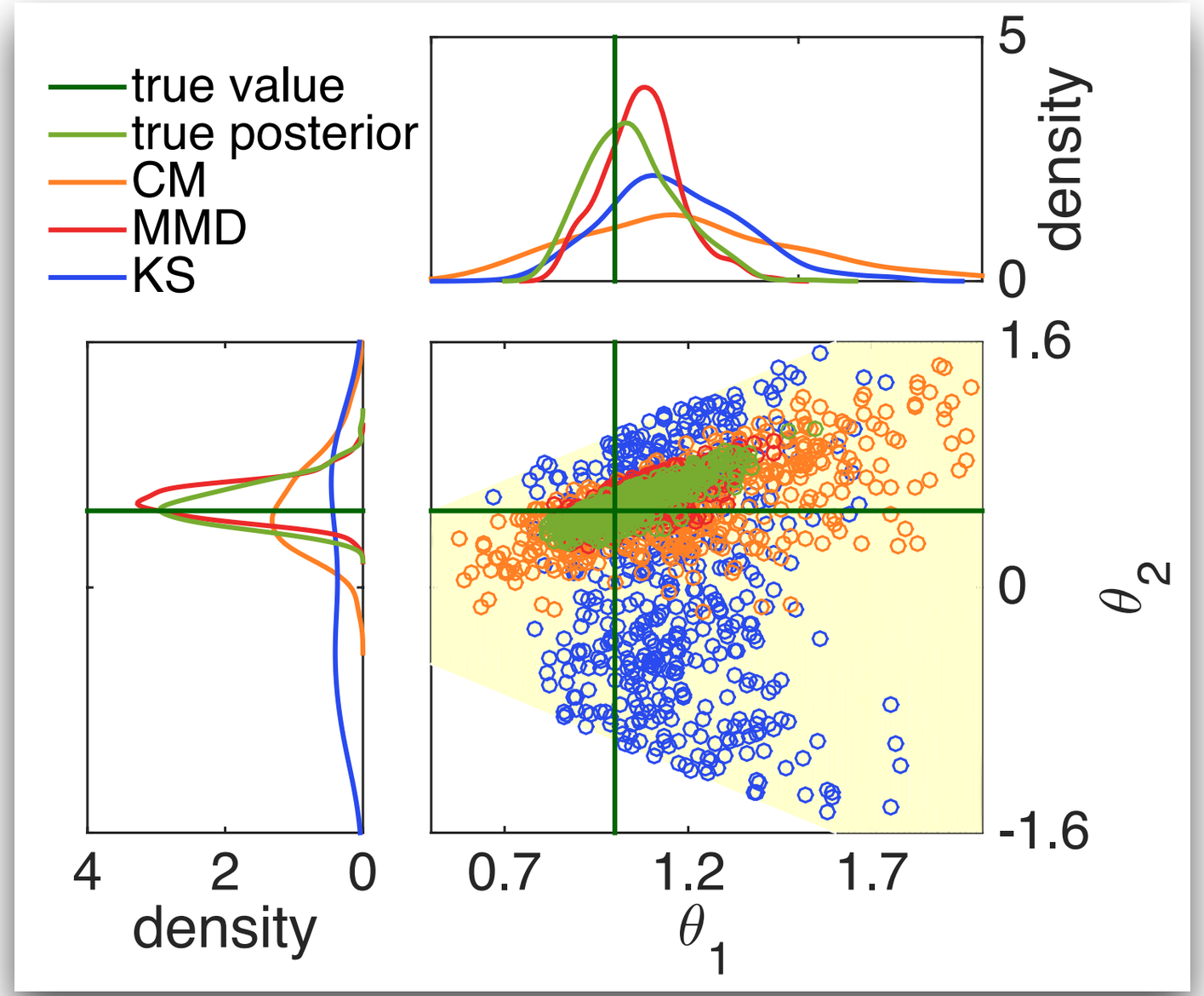


Posterior approximations



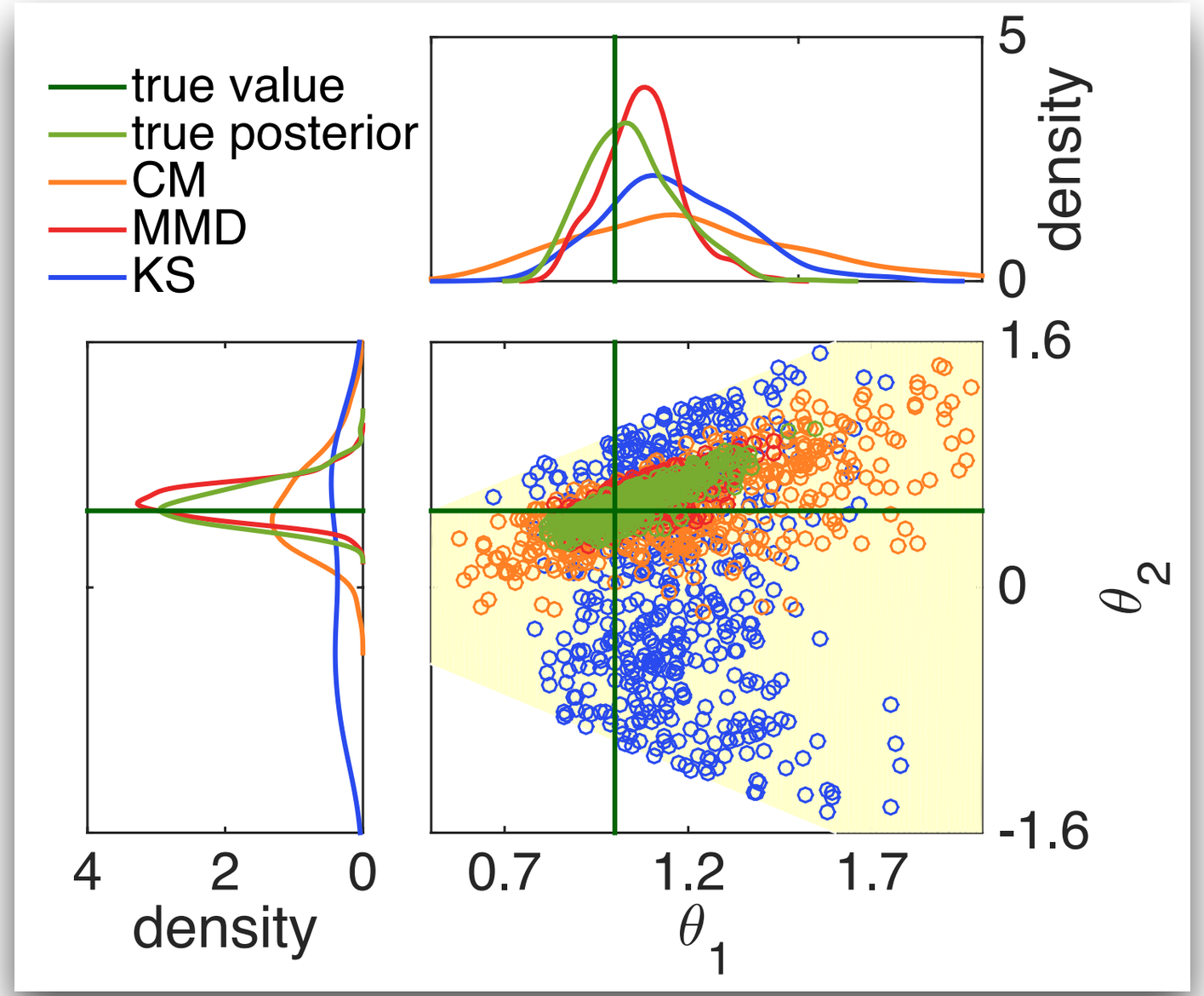
- For KS only θ_1 is identifiable

Posterior approximations



- For KS only θ_1 is identifiable
- CM gives a wider posterior approximation

Posterior approximations



- For KS only θ_1 is identifiable
- CM gives a wider posterior approximation
- Only MMD gives a reasonable posterior approximation

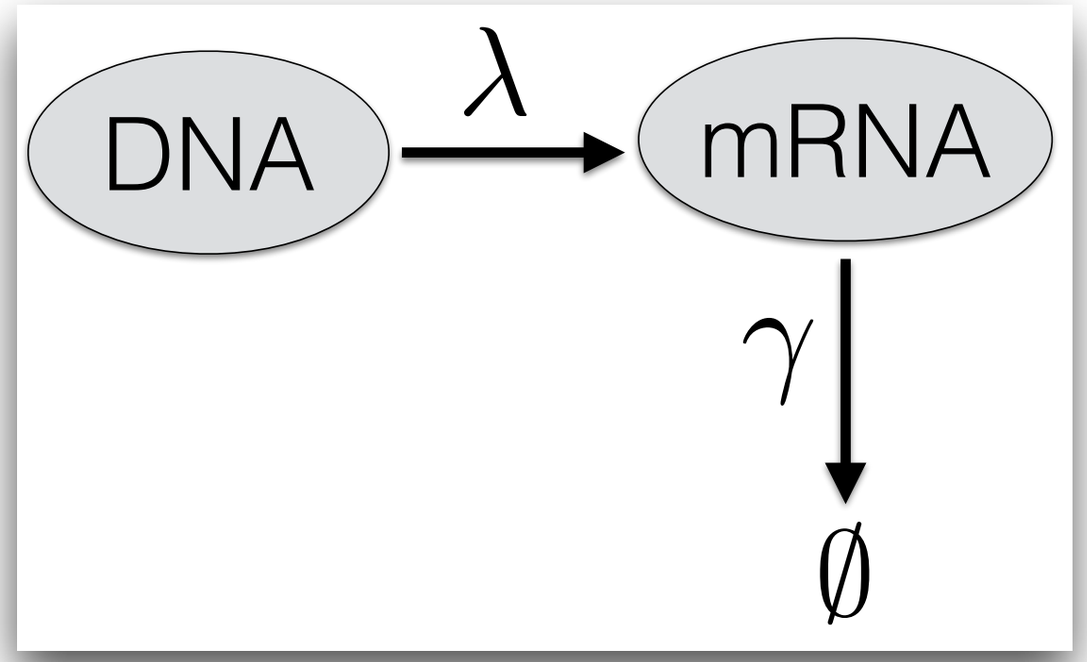
Artificial single-cell time-lapse data

Artificial single-cell time-lapse data

- One stage model of gene expression with mRNA synthesis rate λ and degradation rate γ

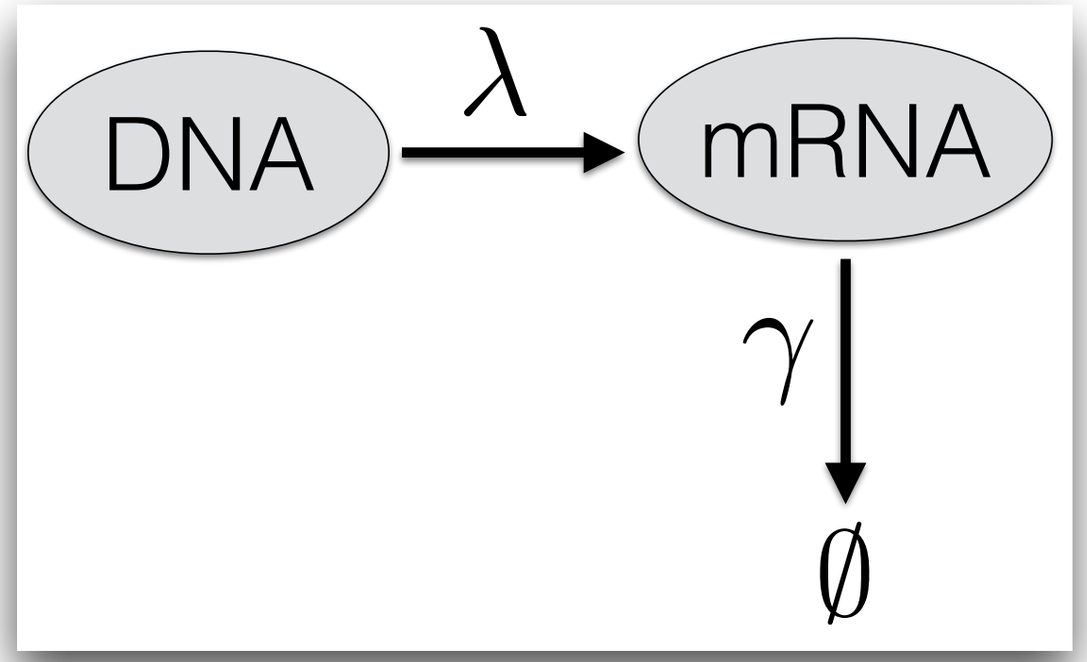
Artificial single-cell time-lapse data

- One stage model of gene expression with mRNA synthesis rate λ and degradation rate γ



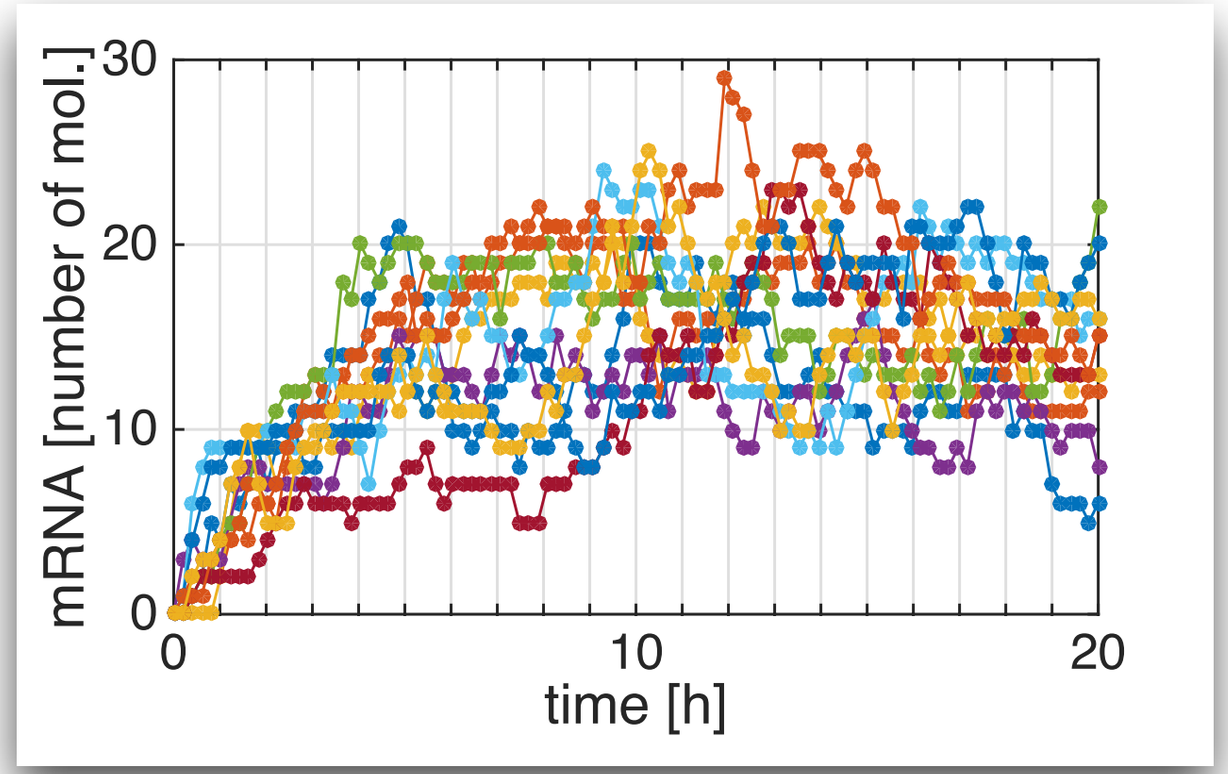
Artificial single-cell time-lapse data

- One stage model of gene expression with mRNA synthesis rate λ and degradation rate γ
- Comparison of ABC SMC with multivariate statistics (MMD) and univariate statistics (KS)



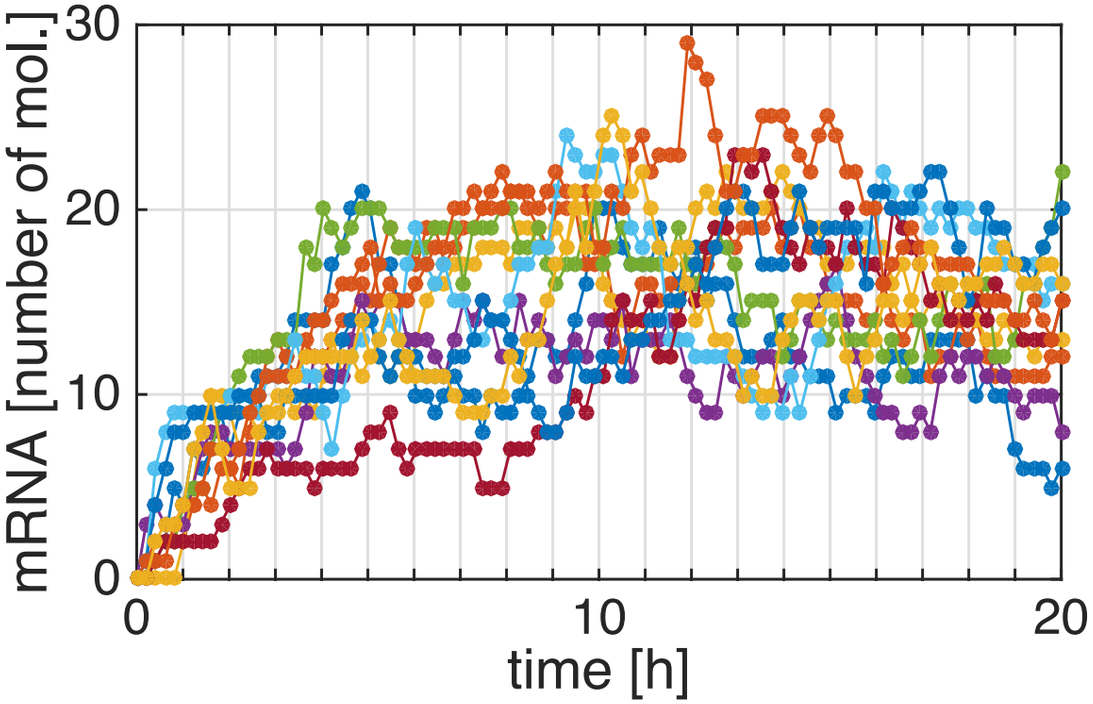
Artificial single-cell time-lapse data

Artificial single-cell time-lapse data

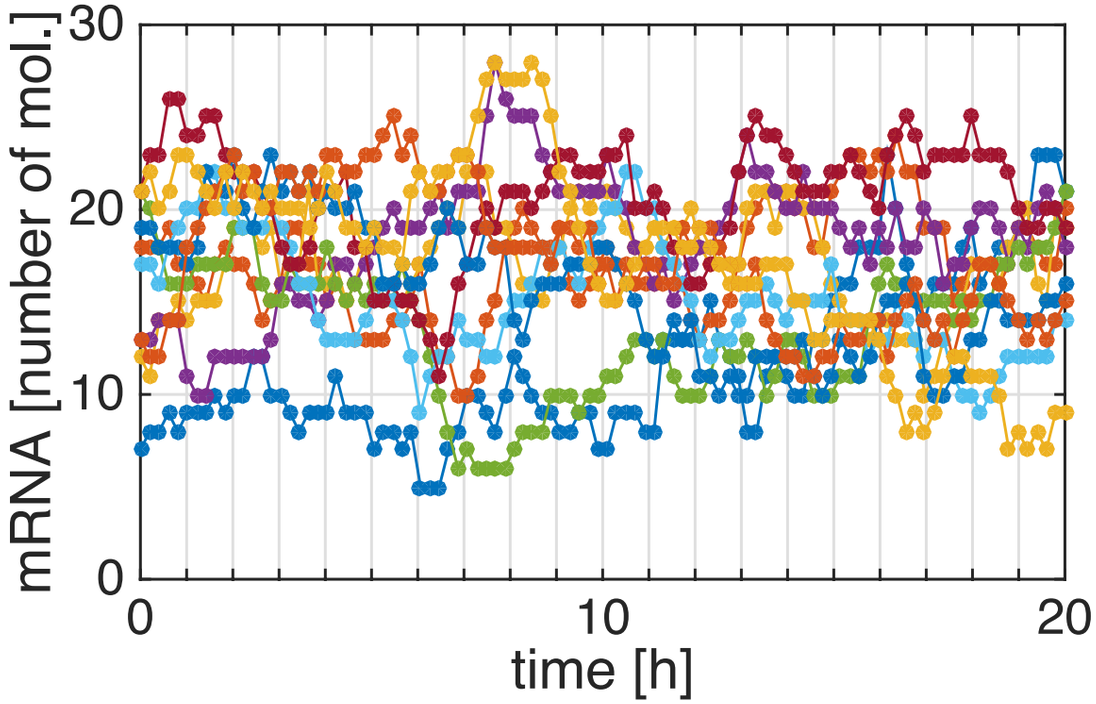


Non-equilibrium time-series

Artificial single-cell time-lapse data

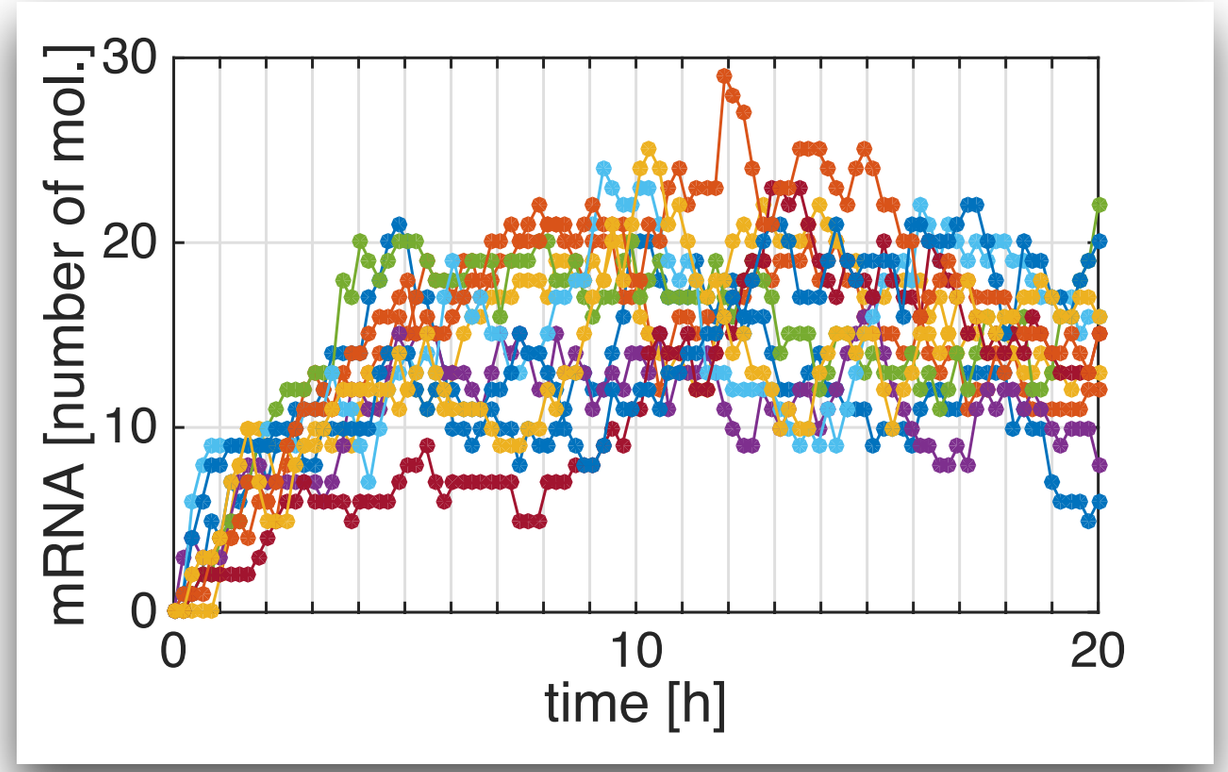


Non-equilibrium time-series

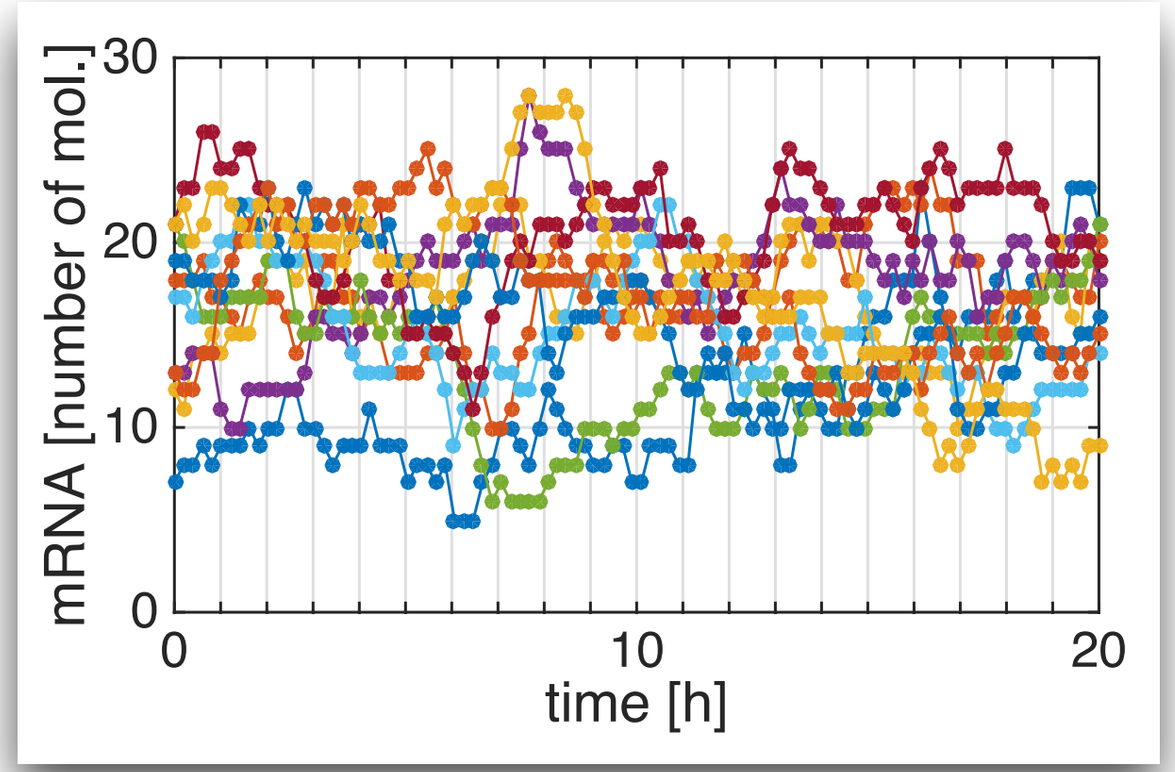


Equilibrium time-series

Artificial single-cell time-lapse data



Non-equilibrium time-series

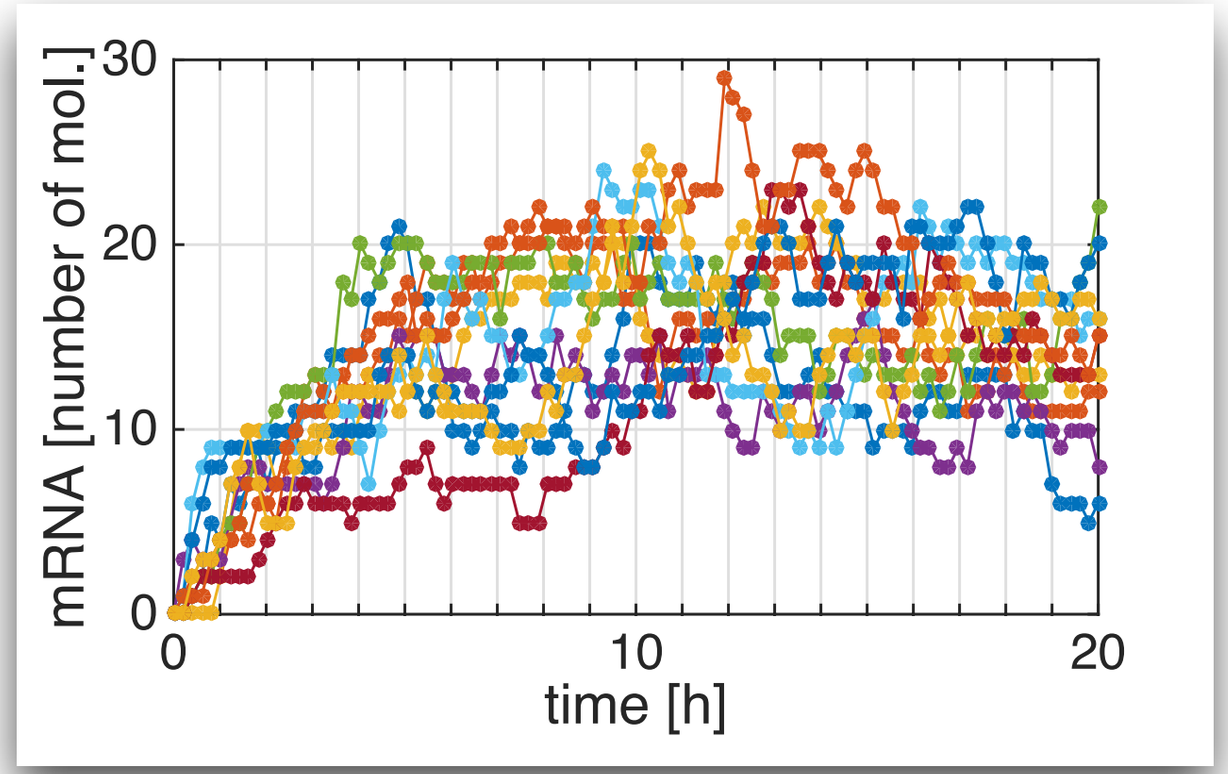


Equilibrium time-series

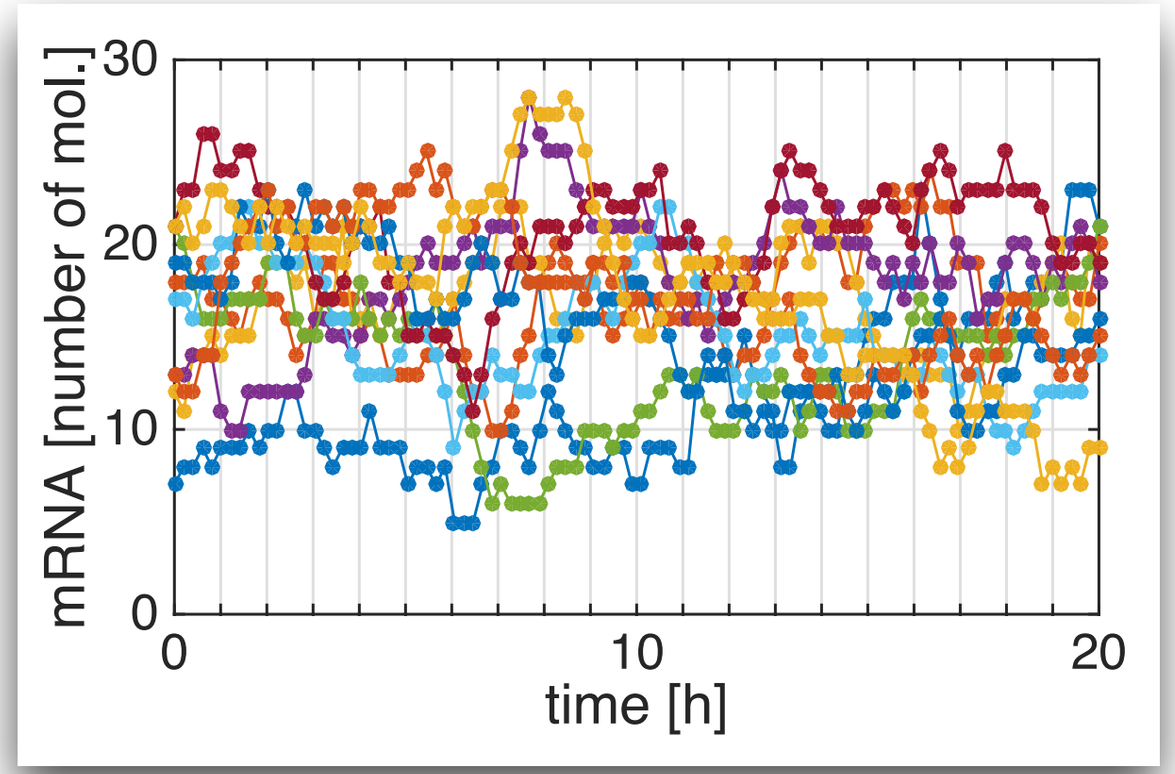
- Generate time-series using the Stochastic Simulation Algorithm [1]

1 Gillespie, Annual Review of Physical Chemistry (2007)

Artificial single-cell time-lapse data



Non-equilibrium time-series



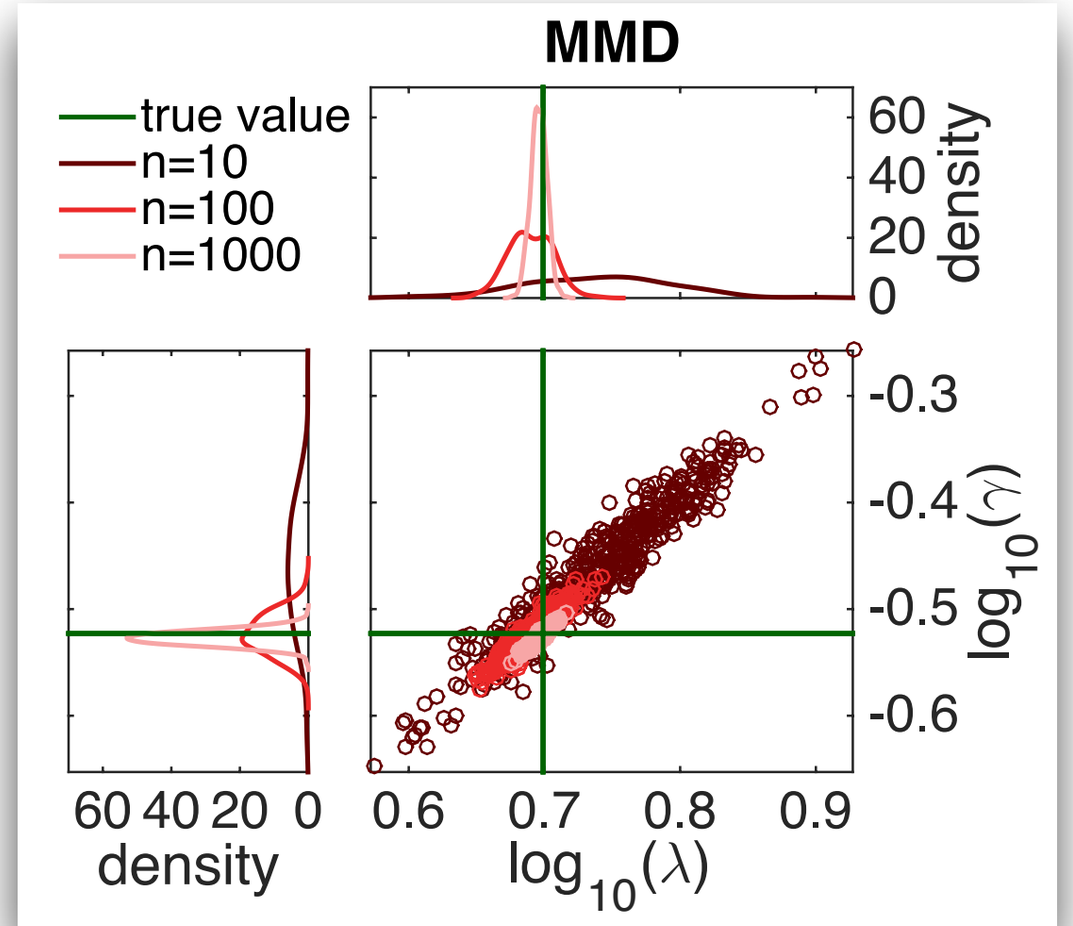
Equilibrium time-series

- Generate time-series using the Stochastic Simulation Algorithm [1]
- Consider 10, 100, 1000 cells and equilibrium and non equilibrium time-series

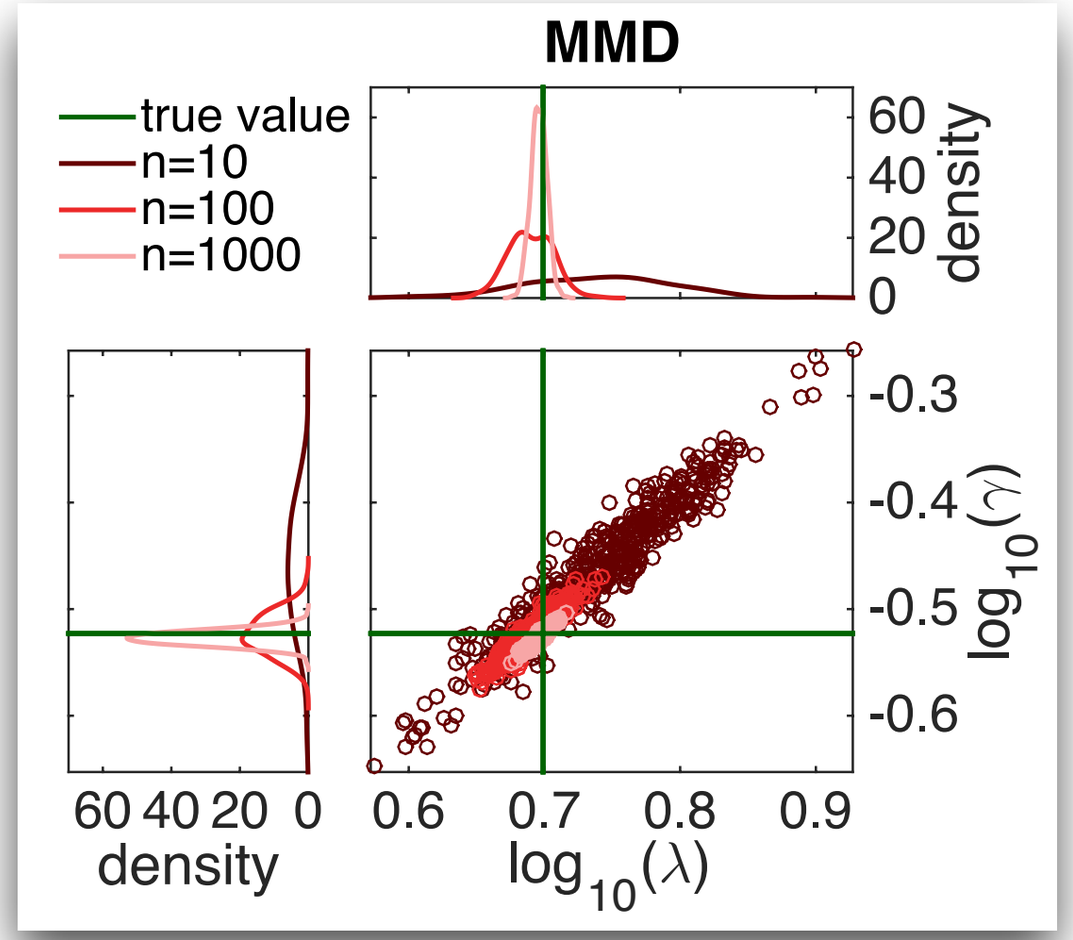
1 Gillespie, Annual Review of Physical Chemistry (2007)

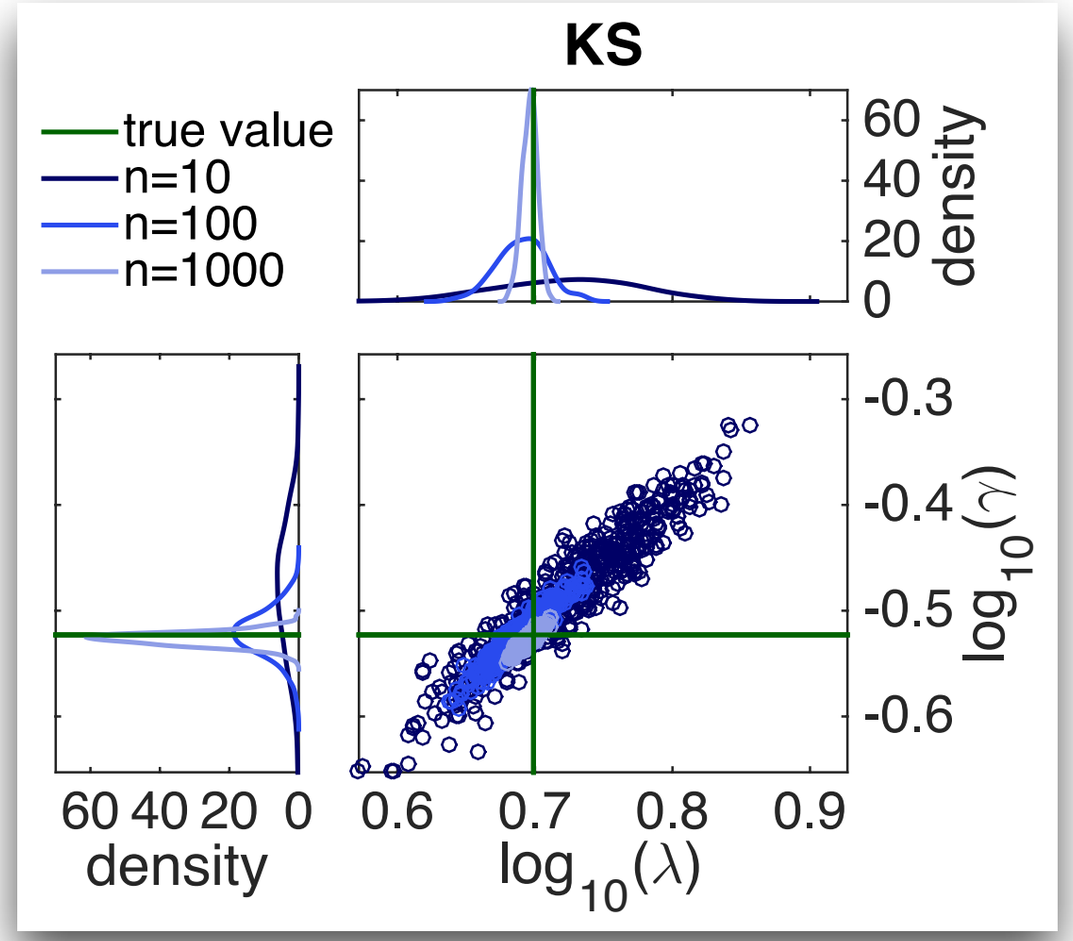
Posterior approximations for non-equilibrium time-series

Posterior approximations for non-equilibrium time-series

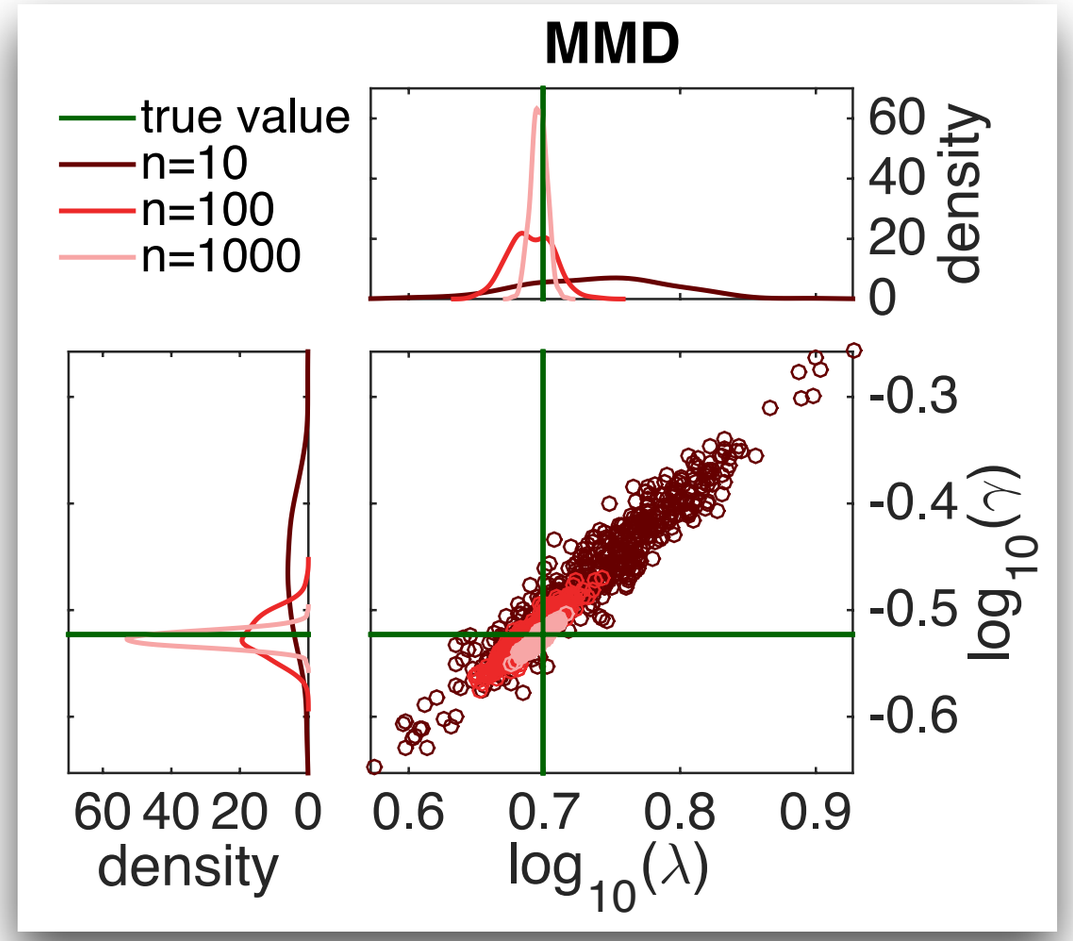


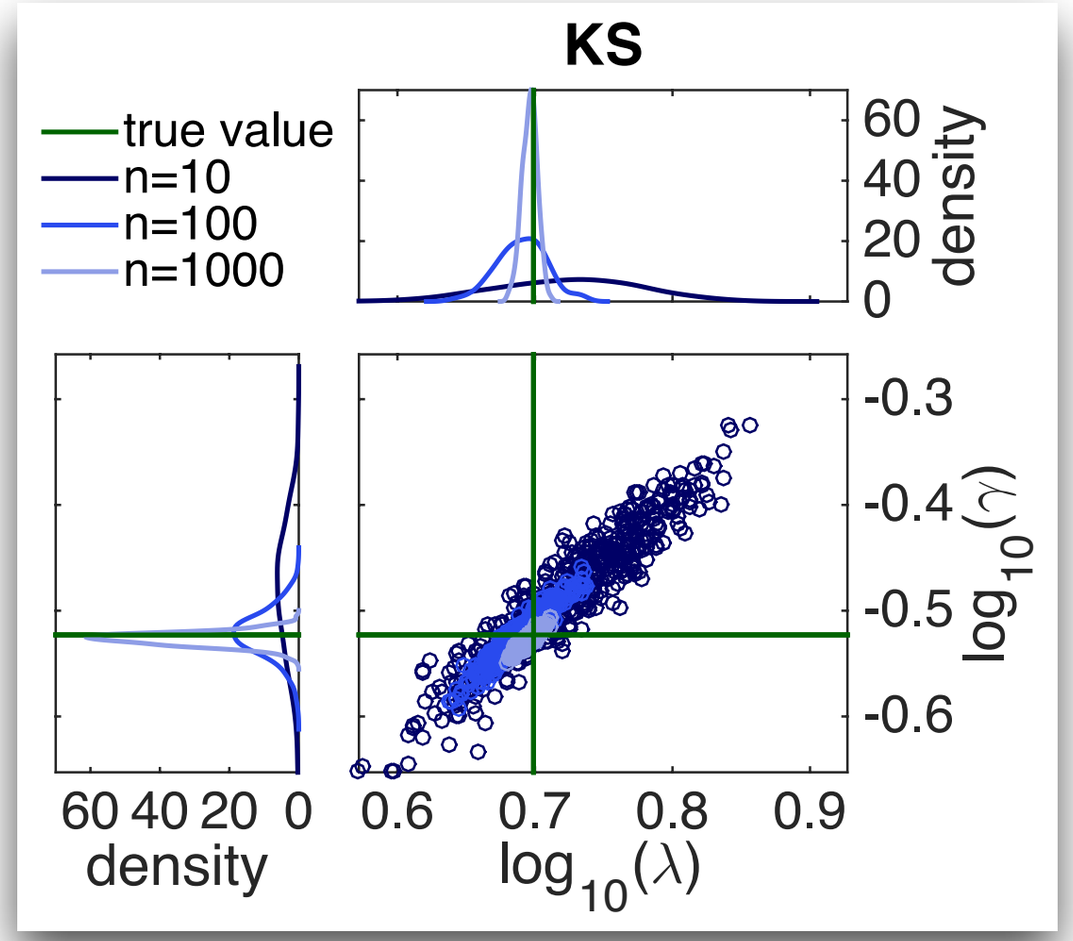
Posterior approximations for non-equilibrium time-series





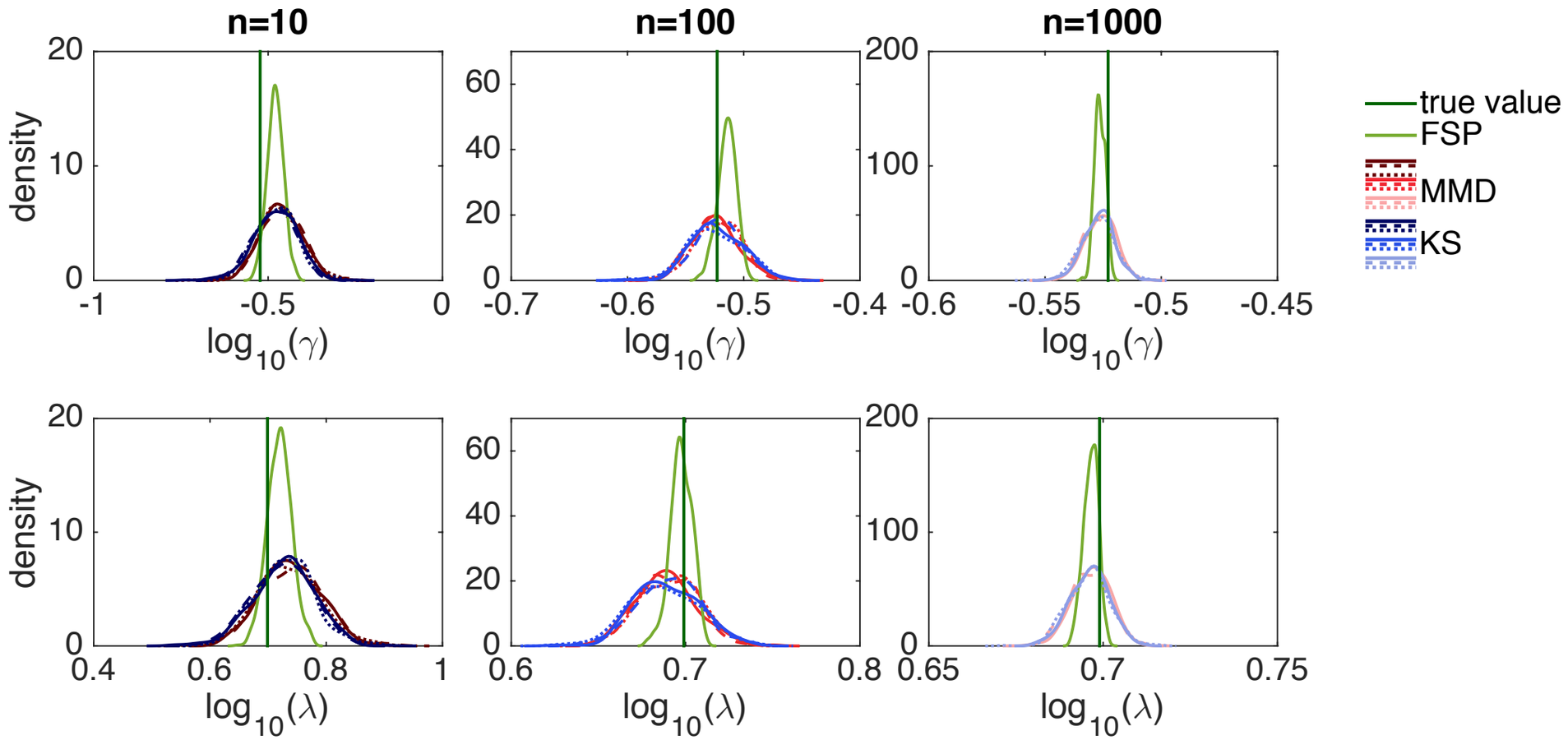
Posterior approximations for non-equilibrium time-series



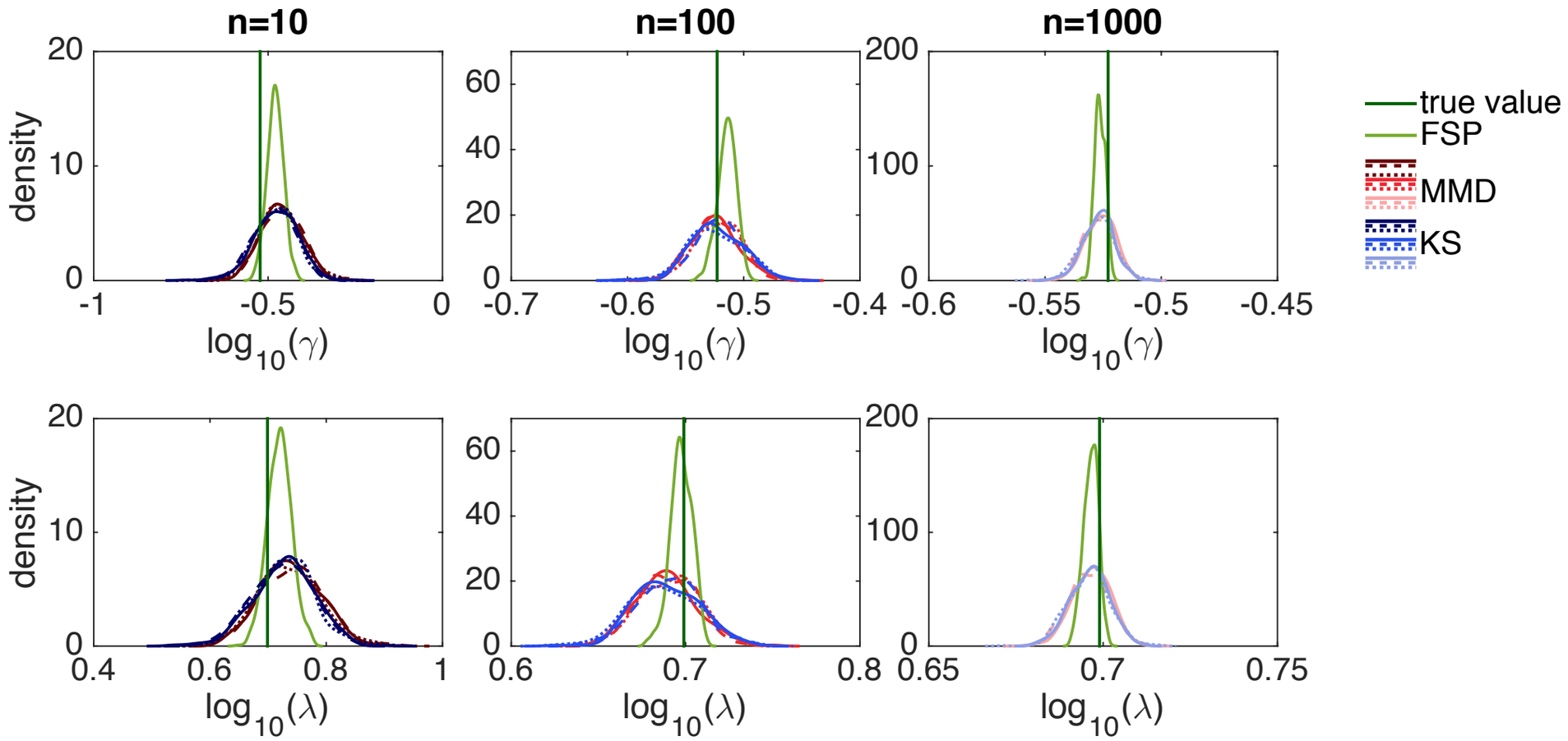


MMD and KS perform similar

Posterior approximations for non-equilibrium time-series



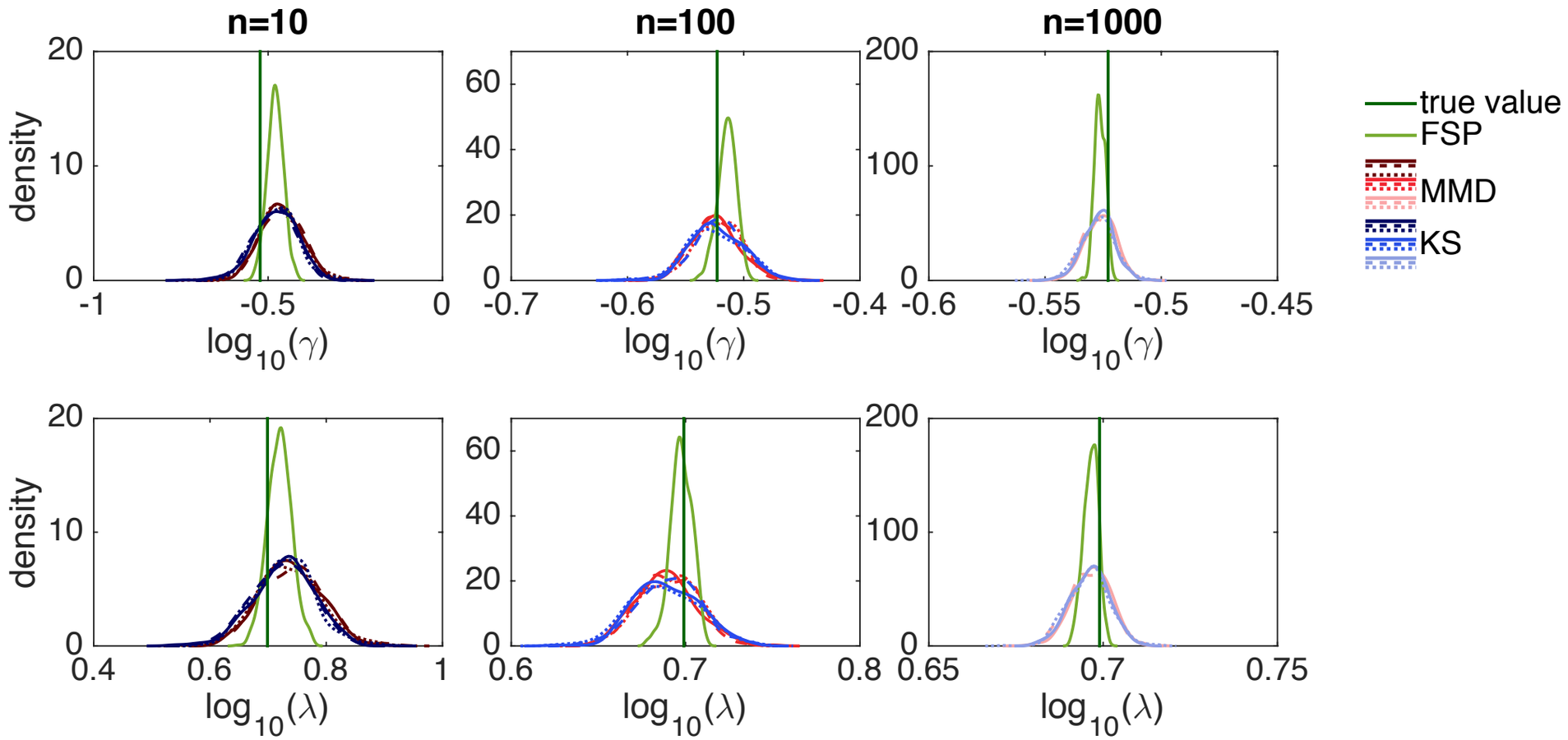
Posterior approximations for non-equilibrium time-series



- Comparison of posterior approximation with Finite State Projection (FSP)[1]

1 Munsky et al., J. Chem. Phys. (2006)

Posterior approximations for non-equilibrium time-series

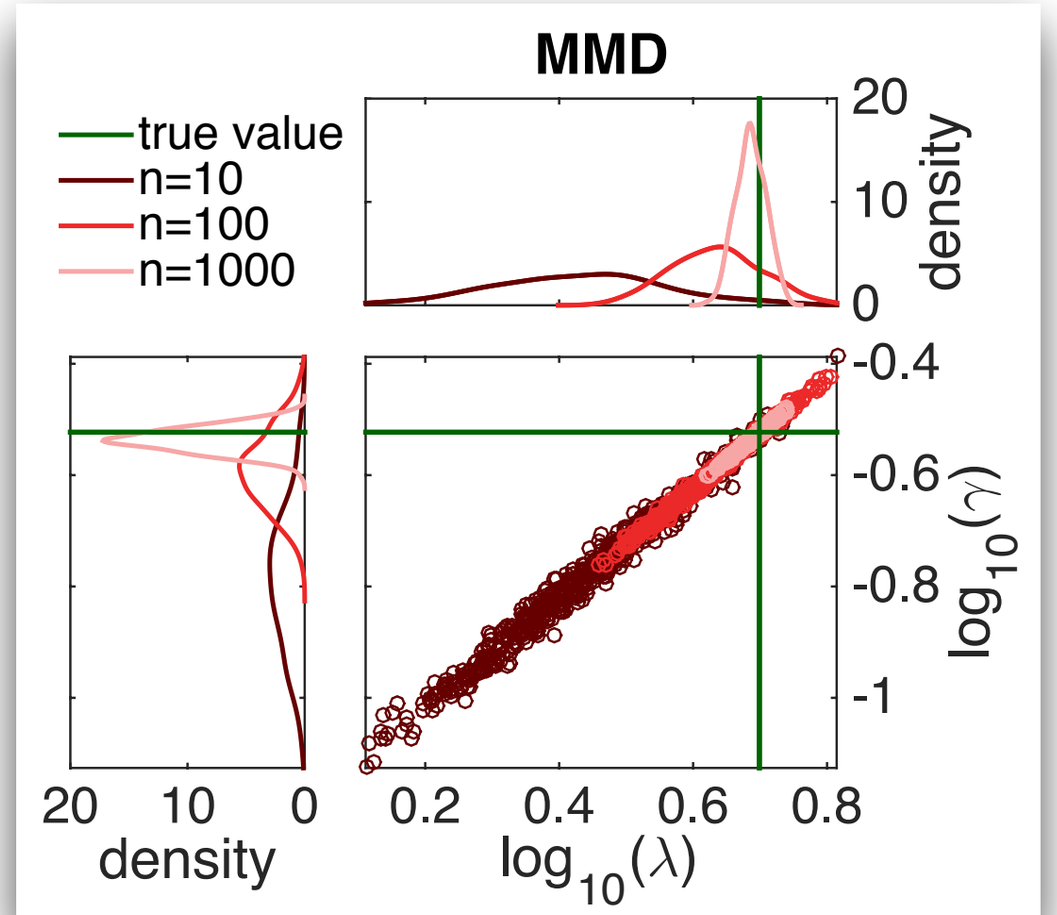


- Comparison of posterior approximation with Finite State Projection (FSP)[1]
- FSP gives narrower posterior approximations compared to MMD and KS

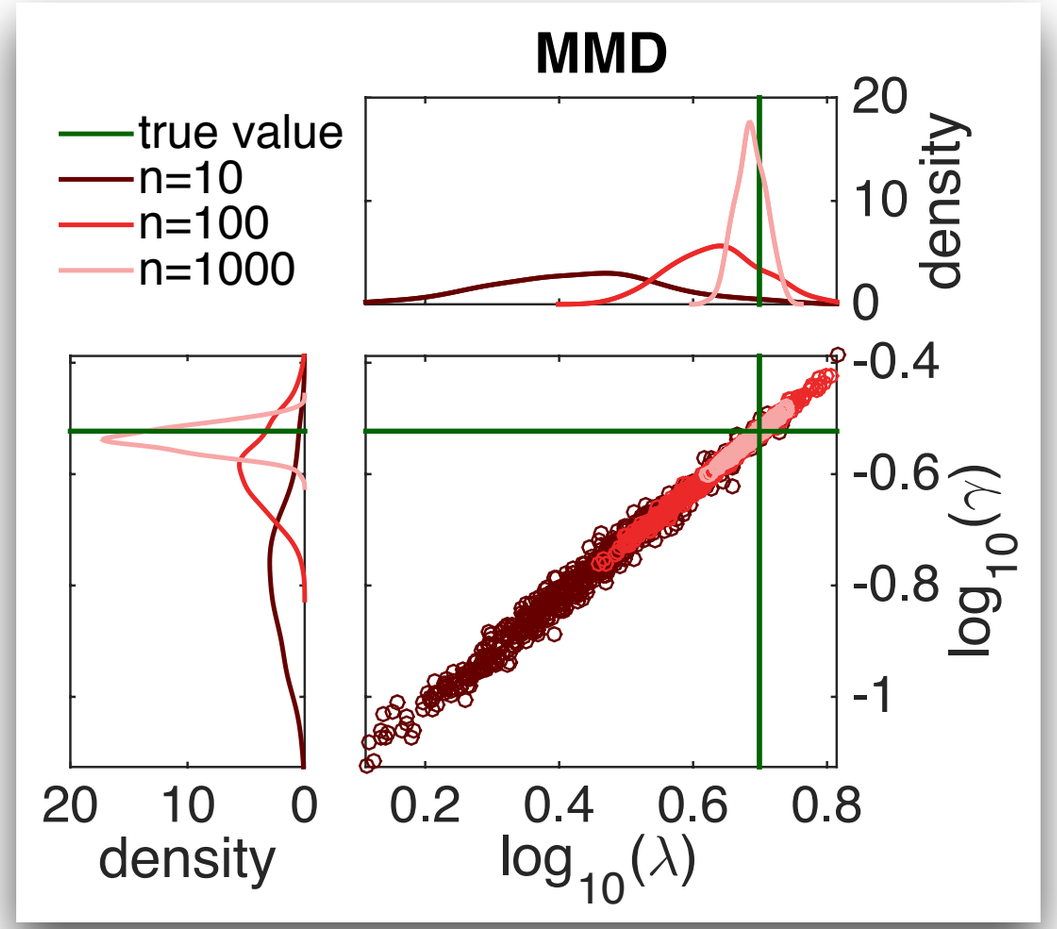
1 Munsky et al., J. Chem. Phys. (2006)

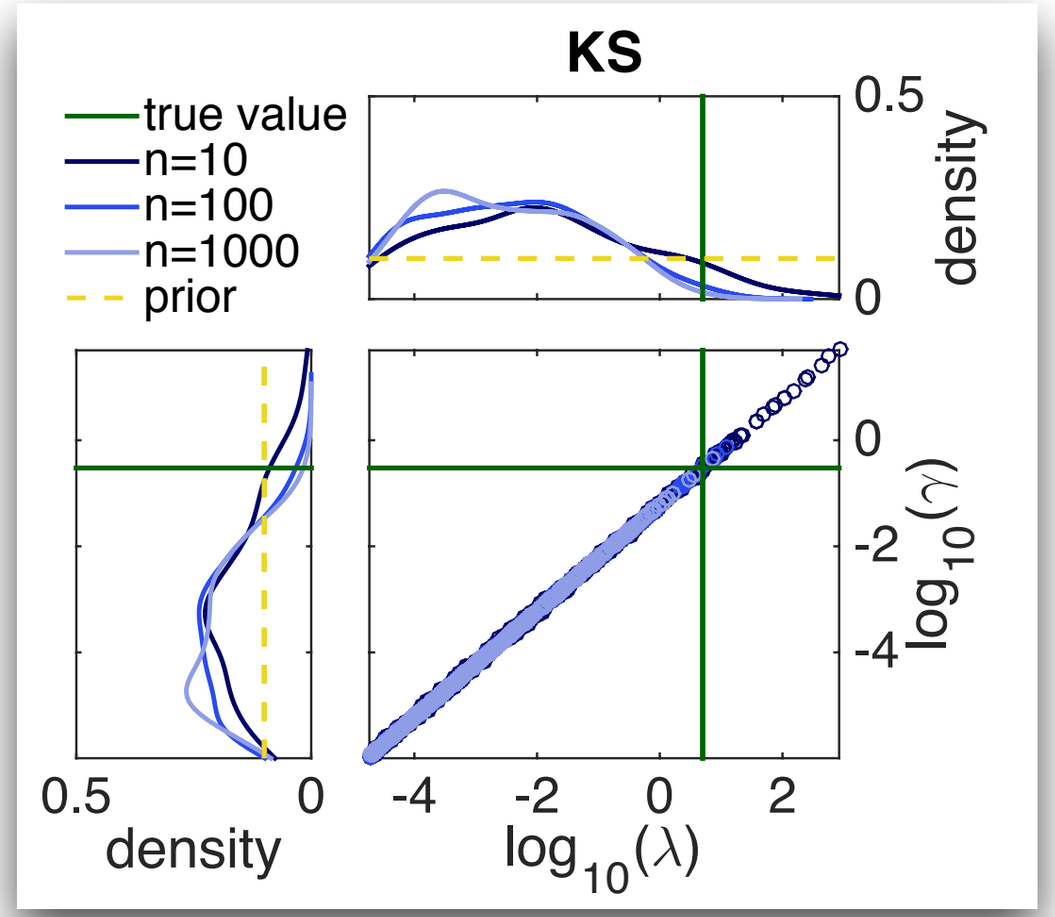
Posterior approximations for equilibrium time-series

Posterior approximations for equilibrium time-series

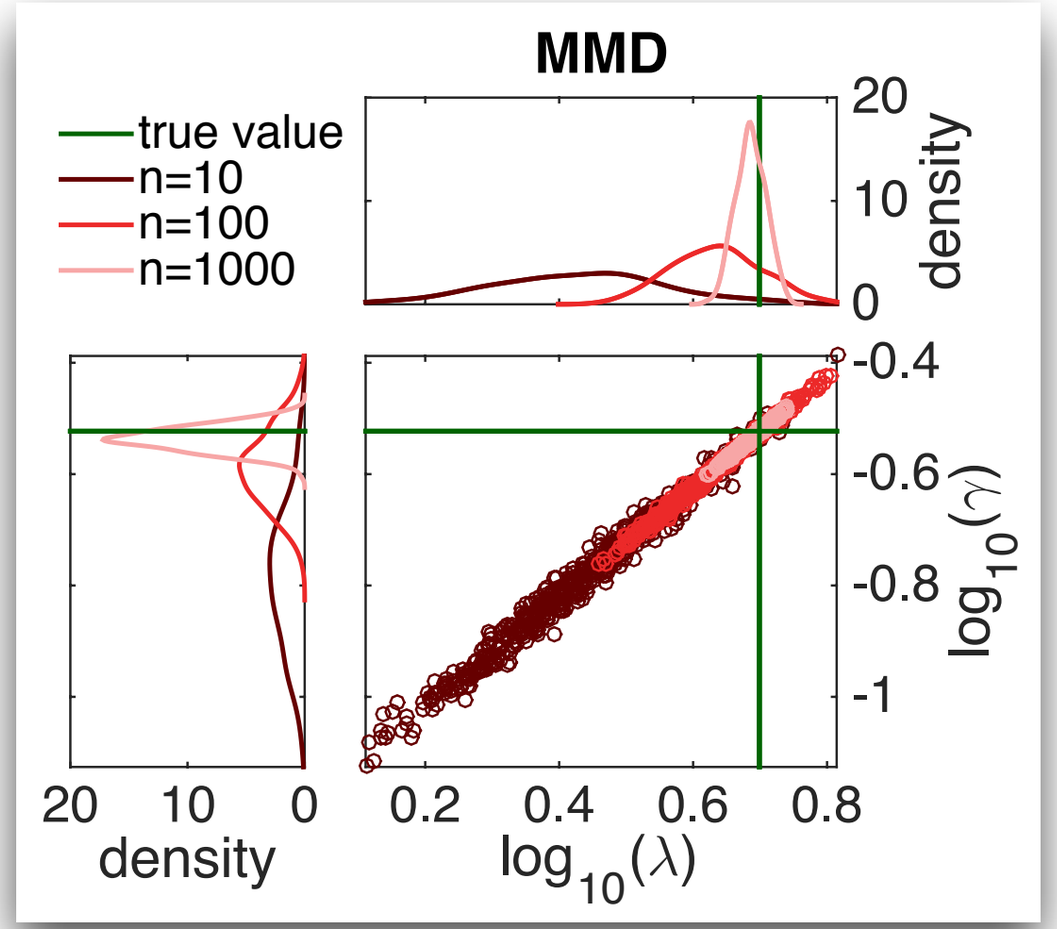


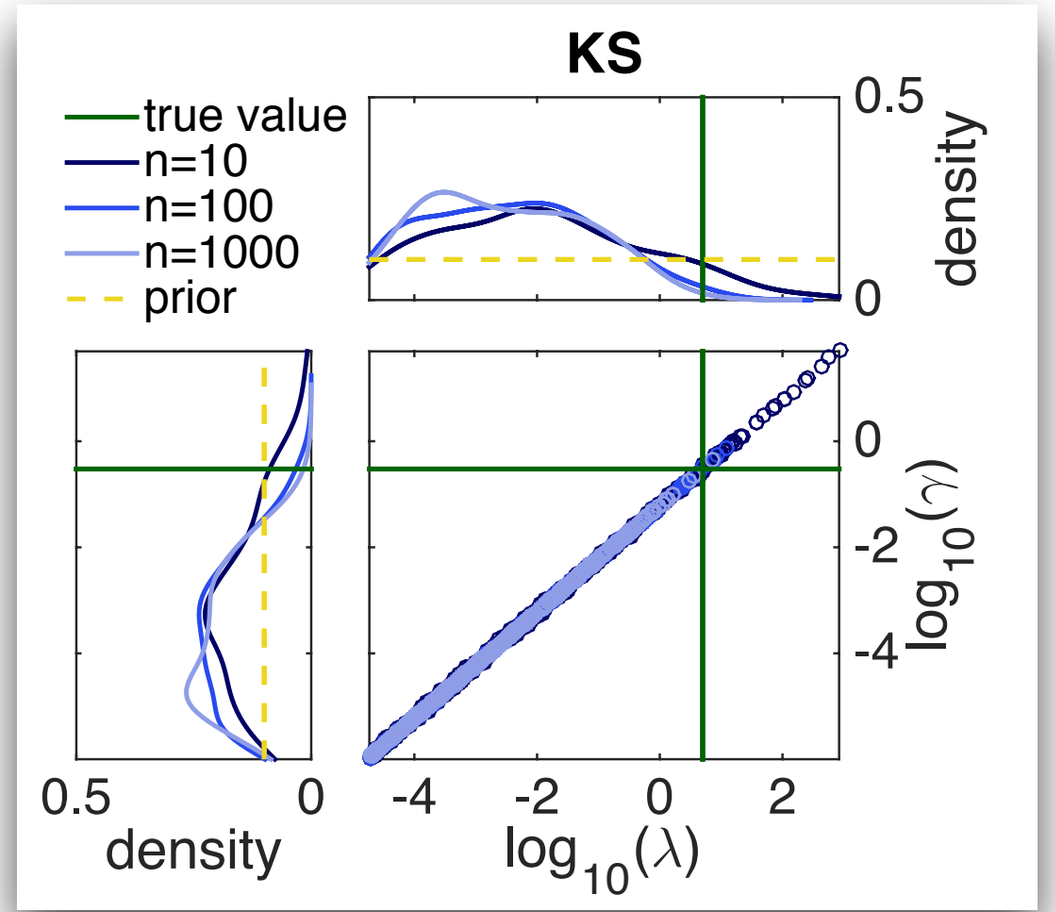
Posterior approximations for equilibrium time-series





Posterior approximations for equilibrium time-series

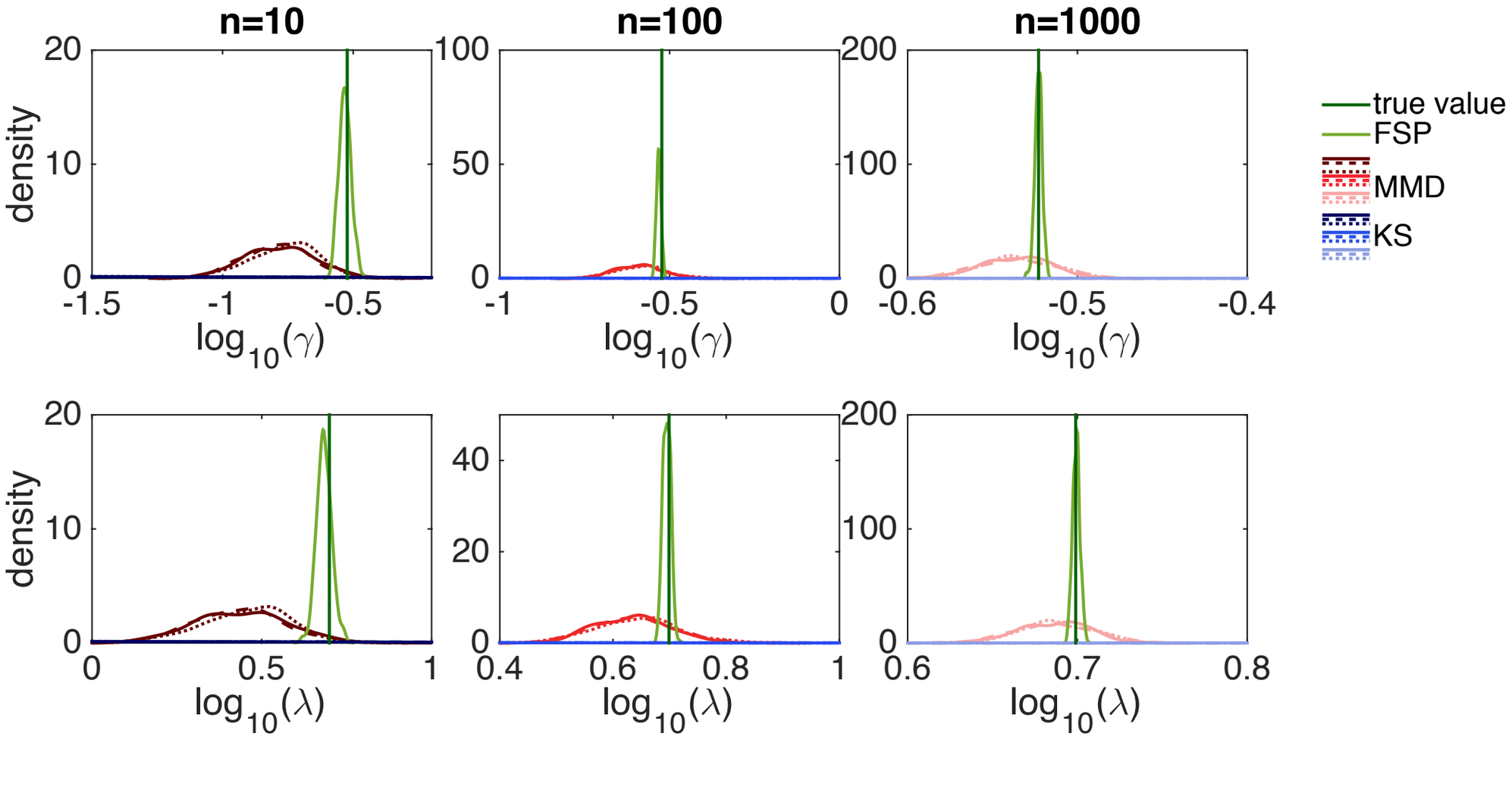




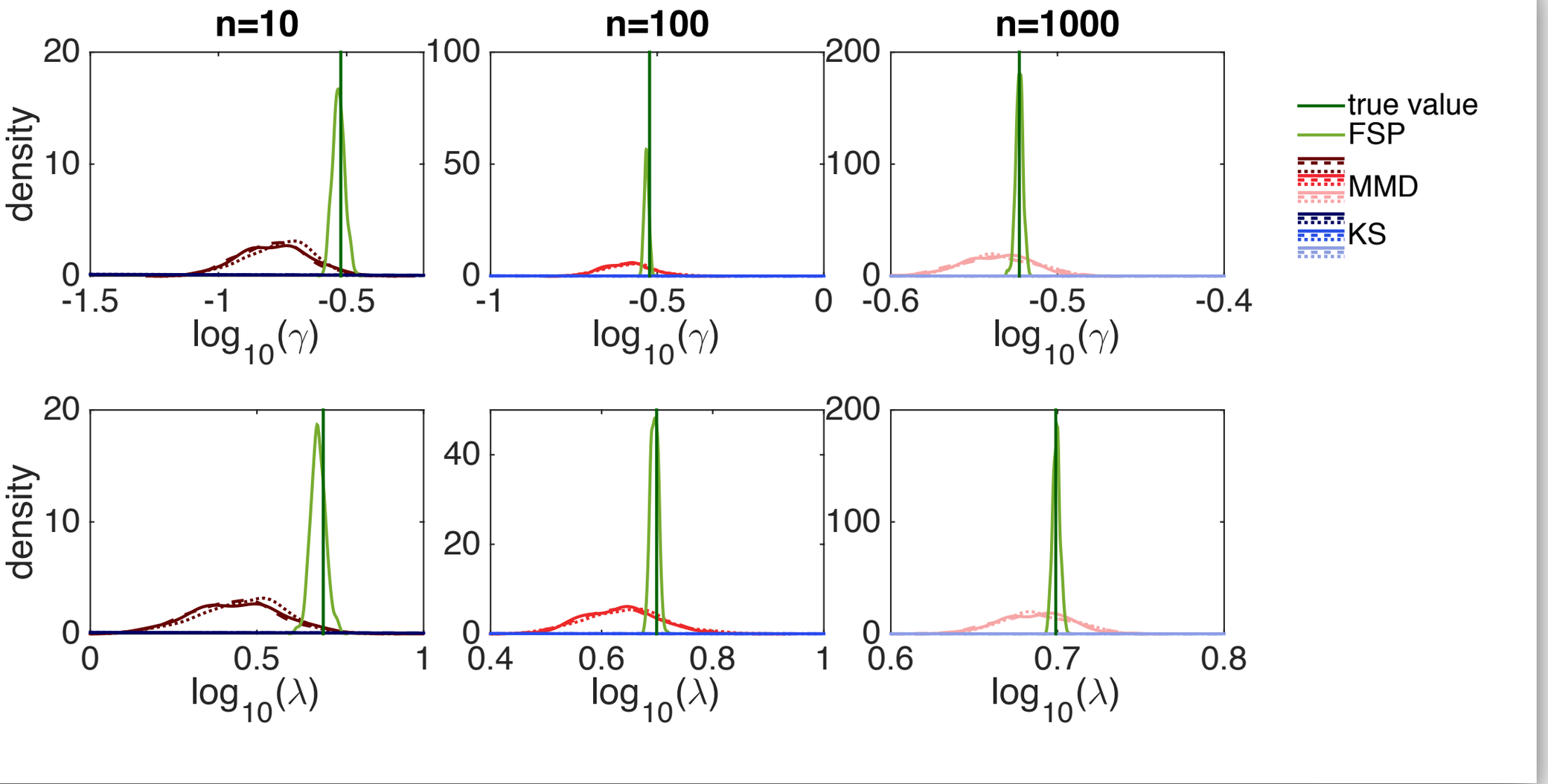
KS can only estimate ratio of parameters

Posterior approximations for equilibrium time-series

Posterior approximations for equilibrium time-series



Posterior approximations for equilibrium time-series

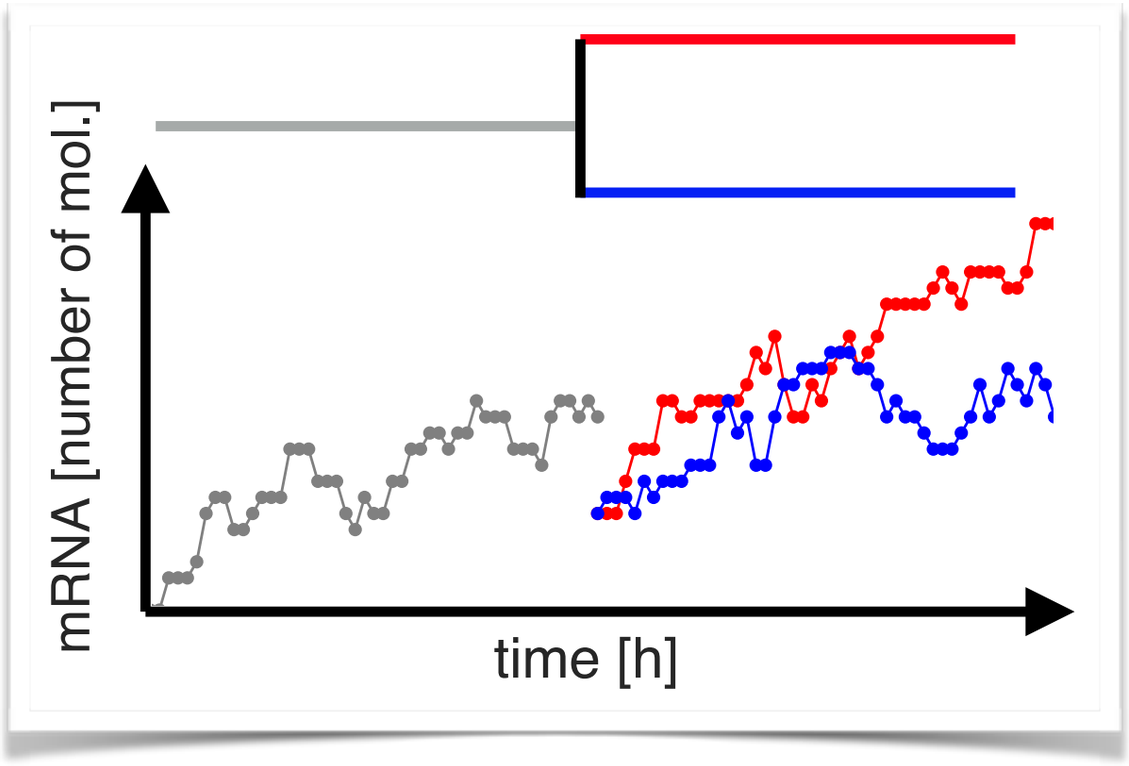


Bigger difference between FSP and MMD for equilibrium data

Tree-structured data

Tree-structured data

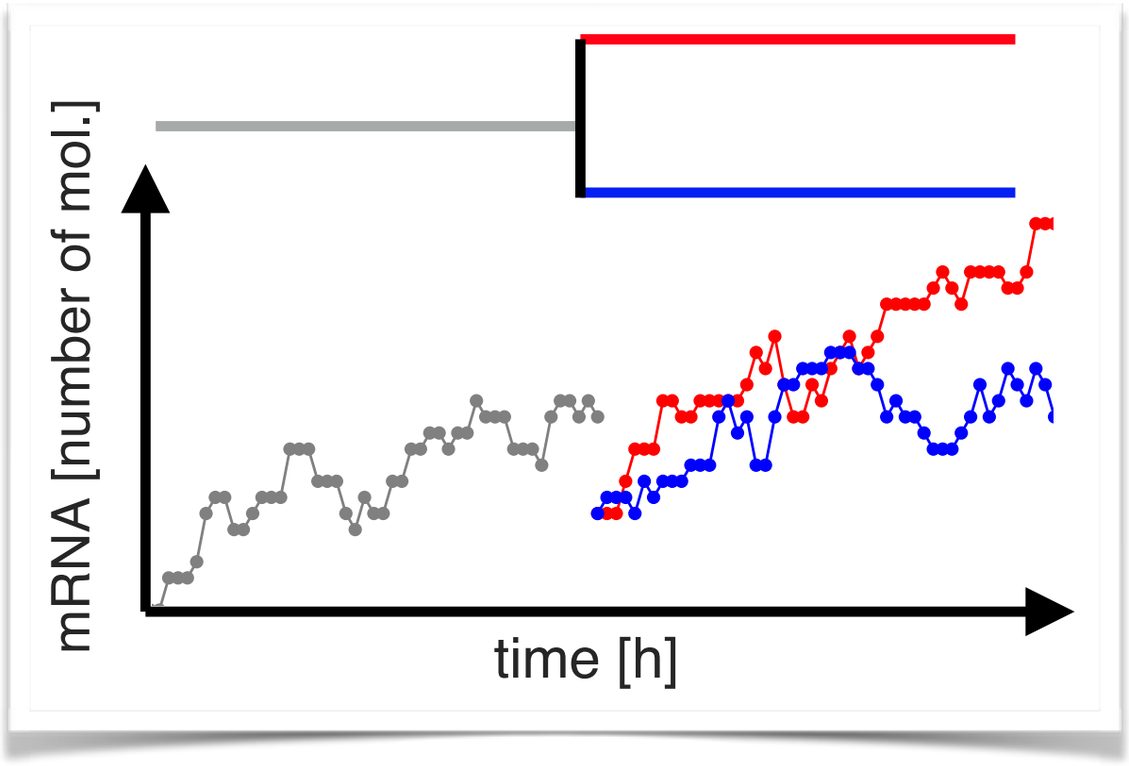
- Data often contain information about the ancestors of a cell[1]



1 Etzrodt et al., Cell Stem Cell (2014)

Tree-structured data

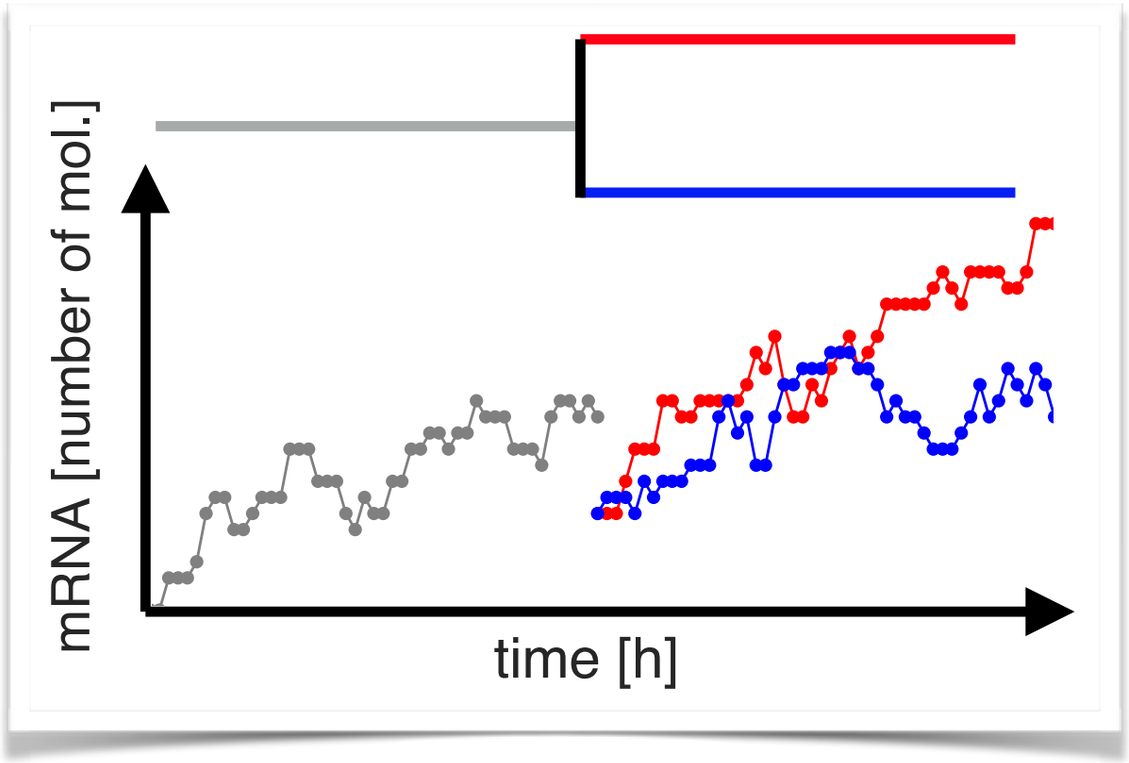
- Data often contain information about the ancestors of a cell[1]
- Include lineage information of the cells



1 Etzrodt et al., Cell Stem Cell (2014)

Tree-structured data

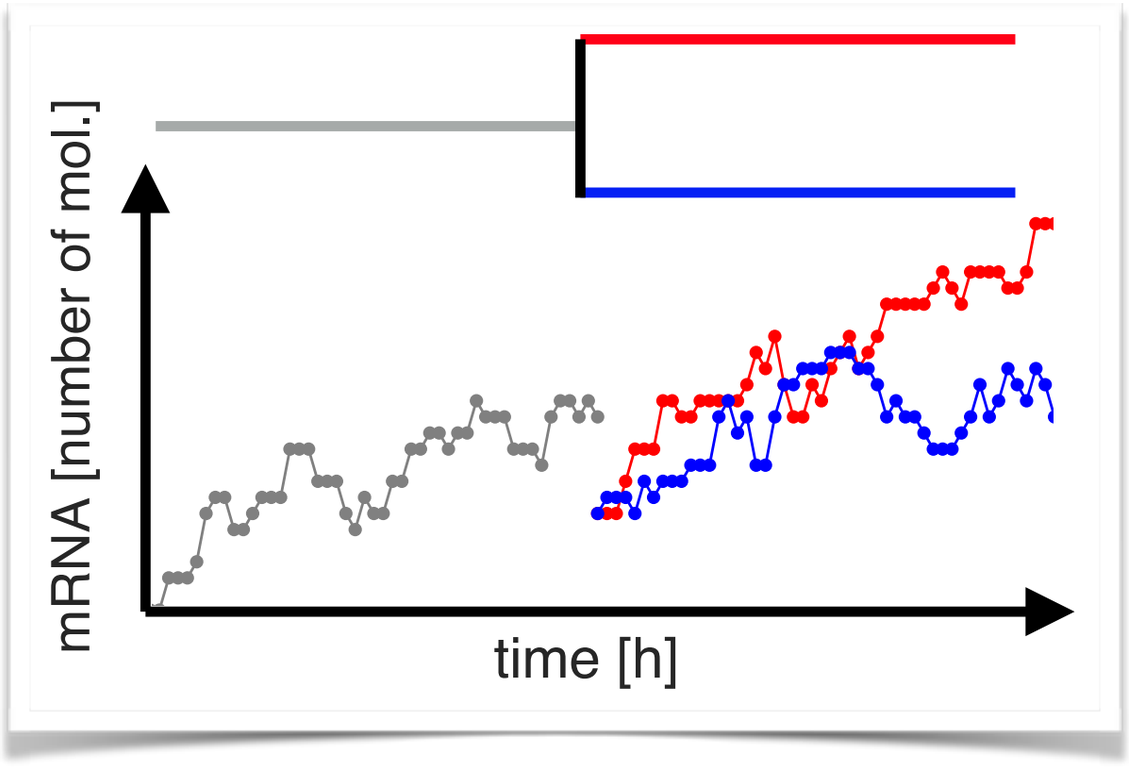
- Data often contain information about the ancestors of a cell[1]
- Include lineage information of the cells
- One sample has the form $x_i = (x_{i,mother}, x_{i,daughter_1}, x_{i,daughter_2})$

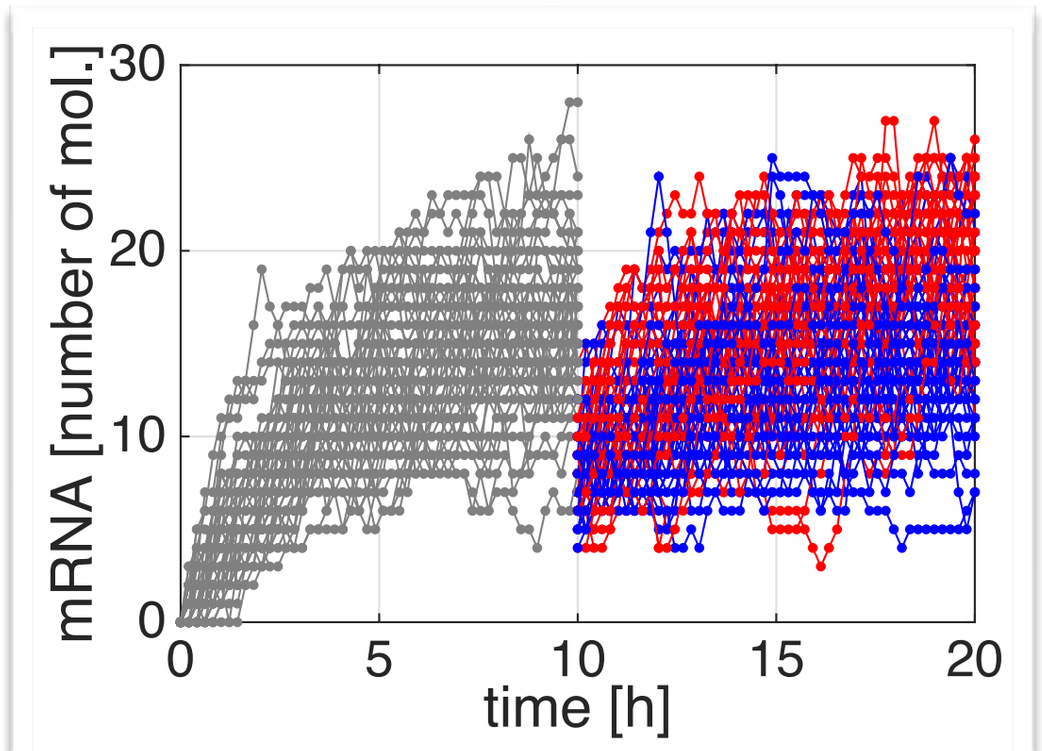


1 Etzrodt et al., Cell Stem Cell (2014)

Tree-structured data

- Data often contain information about the ancestors of a cell[1]
- Include lineage information of the cells
- One sample has the form $x_i = (x_{i,mother}, x_{i,daughter_1}, x_{i,daughter_2})$
- Artificial data of 50 simple trees

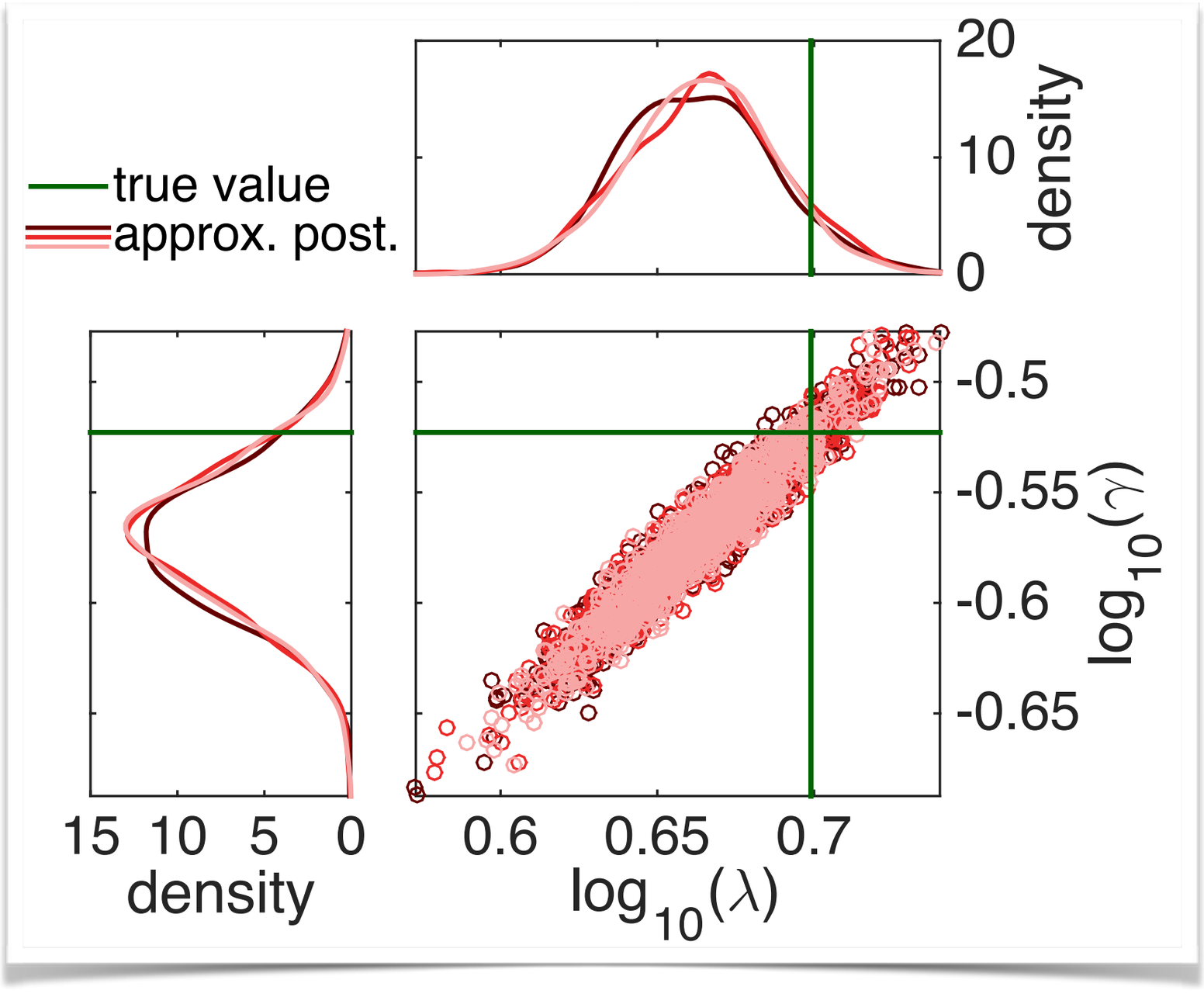




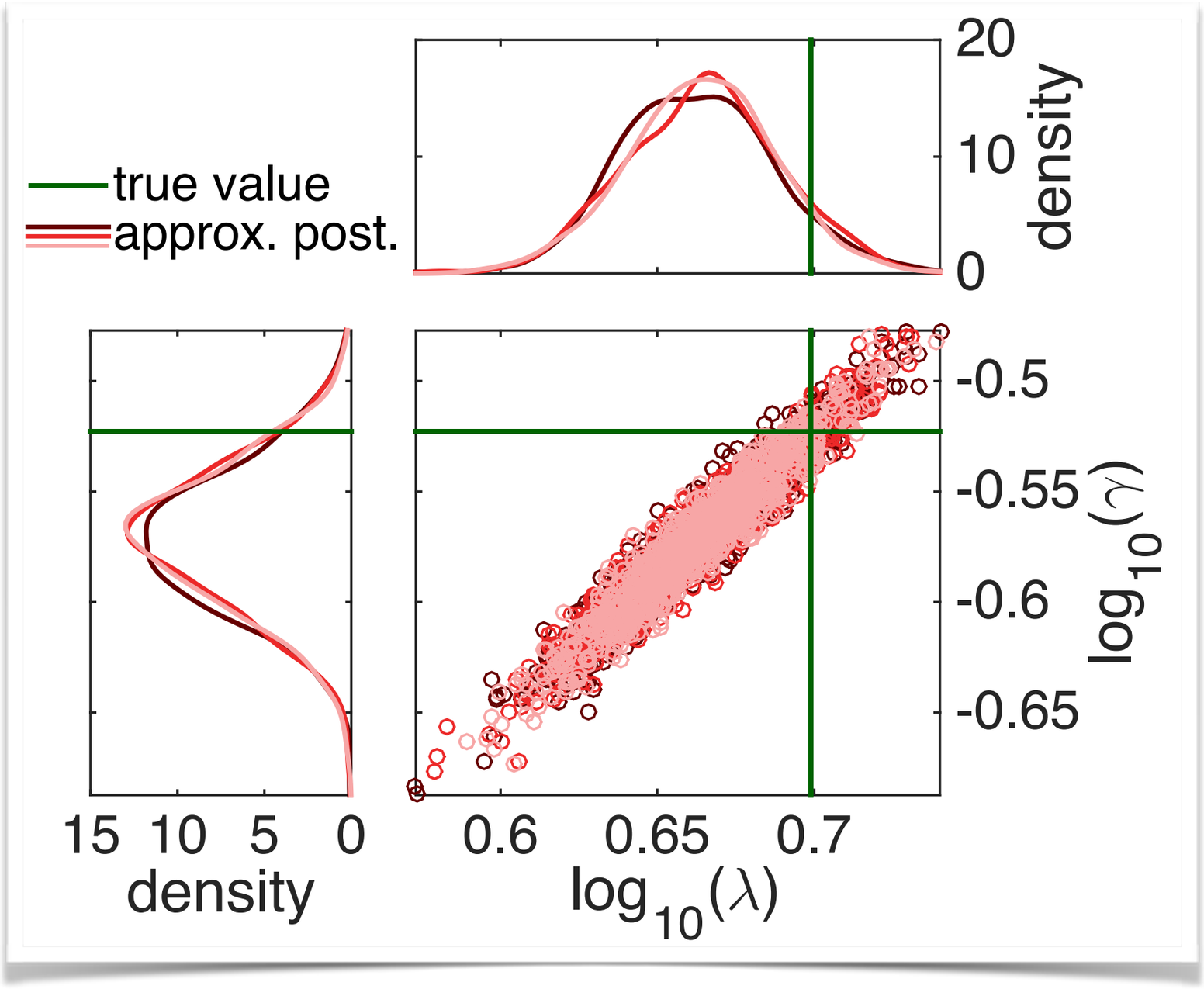
1 Etzrodt et al., Cell Stem Cell (2014)

Tree-structured data

Tree-structured data



Tree-structured data



Use the method to exploit lineage information

Outlook

Outlook

- Evaluate information content of lineage information

Outlook

- Evaluate information content of lineage information
- Incorporate efficient simulations

Outlook

- Evaluate information content of lineage information
- Incorporate efficient simulations
  - Tau-leaping[1]

1 Gillespie, Annual Review of Physical Chemistry (2007)

Outlook

- Evaluate information content of lineage information
- Incorporate efficient simulations
  - Tau-leaping[1]
  - Method of conditional moments [2]

1 Gillespie, Annual Review of Physical Chemistry (2007)
2 Hasenauer et al., Journal of Mathematical Biology (2014)

Outlook

- Evaluate information content of lineage information
- Incorporate efficient simulations
  - Tau-leaping[1]
  - Method of conditional moments [2]
- Further improvements by tuning parameters of ABC

1 Gillespie, Annual Review of Physical Chemistry (2007)
2 Hasenauer et al., Journal of Mathematical Biology (2014)

Outlook

- Evaluate information content of lineage information
- Incorporate efficient simulations
  - Tau-leaping[1]
  - Method of conditional moments [2]
- Further improvements by tuning parameters of ABC
  - Appropriate threshold sequence[3]

1 Gillespie, Annual Review of Physical Chemistry (2007)
2 Hasenauer et al., Journal of Mathematical Biology (2014)
3 Silk et al., Stat. Appl. Genet. Mol. (2013)

Outlook

- Evaluate information content of lineage information
- Incorporate efficient simulations
  - Tau-leaping[1]
  - Method of conditional moments [2]
- Further improvements by tuning parameters of ABC
  - Appropriate threshold sequence[3]
  - Study perturbation kernels

1 Gillespie, Annual Review of Physical Chemistry (2007)
2 Hasenauer et al., Journal of Mathematical Biology (2014)
3 Silk et al., Stat. Appl. Genet. Mol. (2013)

Outlook

- Evaluate information content of lineage information
- Incorporate efficient simulations
  - Tau-leaping[1]
  - Method of conditional moments [2]
- Further improvements by tuning parameters of ABC
  - Appropriate threshold sequence[3]
  - Study perturbation kernels
- Study adaptive number of simulations

1 Gillespie, Annual Review of Physical Chemistry (2007)
2 Hasenauer et al., Journal of Mathematical Biology (2014)
3 Silk et al., Stat. Appl. Genet. Mol. (2013)

Summary

Summary

- Method to analyse single-cell time-lapse data

Summary

- Method to analyse single-cell time-lapse data
- Account for temporal information of individual cells

Summary

- Method to analyse single-cell time-lapse data
- Account for temporal information of individual cells
- Identifiability in several cases only with multivariate statistics

Summary

- Method to analyse single-cell time-lapse data
- Account for temporal information of individual cells
- Identifiability in several cases only with multivariate statistics
- MMD is a suitable test statistic

Summary

- Method to analyse single-cell time-lapse data
- Account for temporal information of individual cells
- Identifiability in several cases only with multivariate statistics
- MMD is a suitable test statistic
- Enables incorporation of lineage information

Summary

- Method to analyse single-cell time-lapse data
- Account for temporal information of individual cells
- Identifiability in several cases only with multivariate statistics
- MMD is a suitable test statistic
- Enables incorporation of lineage information

Flexible framework for the analysis of single-cell time-lapse data, which might help getting deeper insight into the cellular mechanisms and therefore advance e.g. stem cell research.

Acknowledgement

Jan Hasenauer
Carsten Marr
Fabian Theis
Dennis Rickert

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Institute of Computational Biology
Data-Driven Computational Modelling

Acknowledgement

Jan Hasenauer
Carsten Marr
Fabian Theis
Dennis Rickert

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Institute of Computational Biology
Data-Driven Computational Modelling

carolin.loos@helmholtz-muenchen.de