Horizon 2020 ITN project SUNDIAL

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 721463

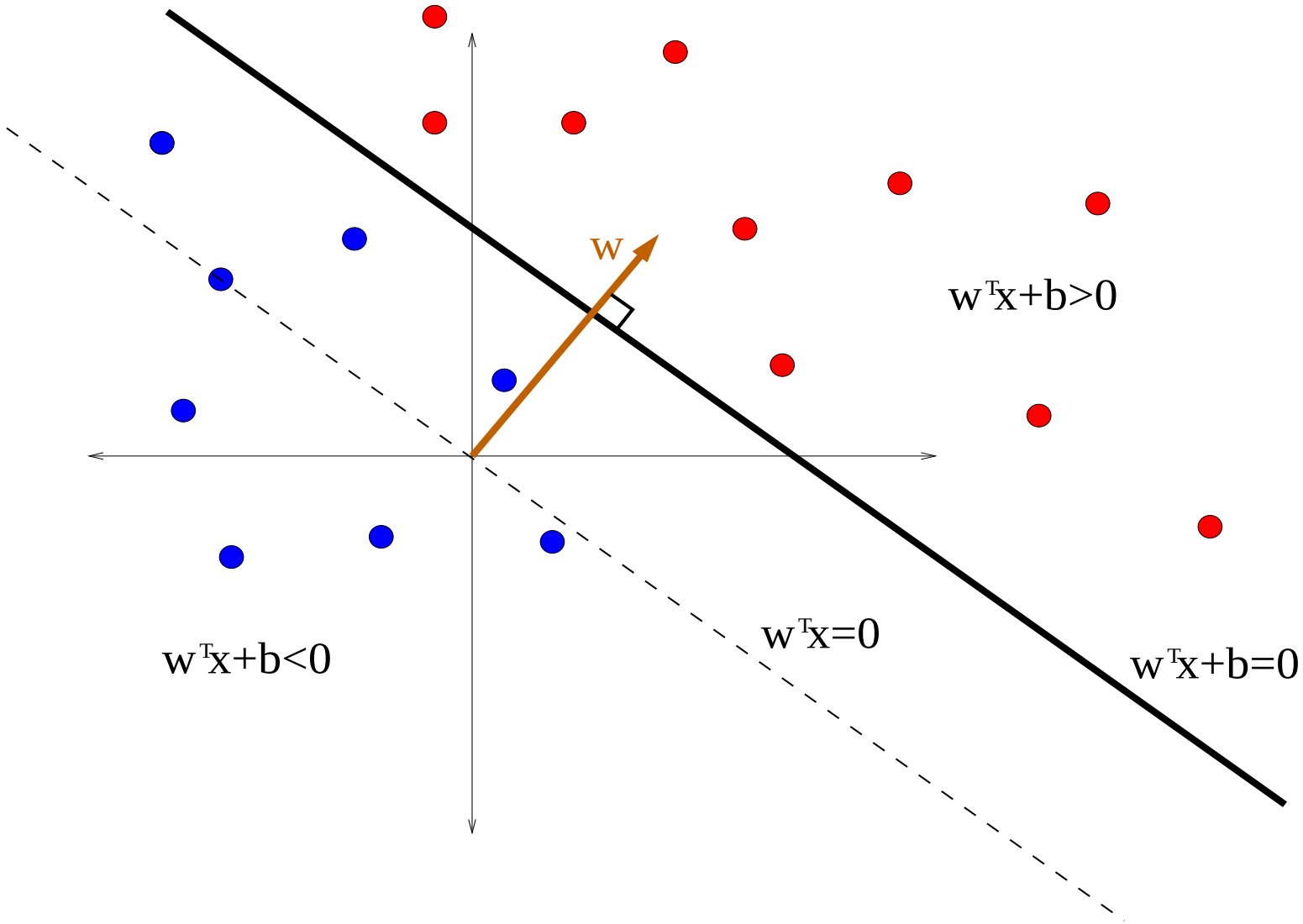
Dynamical Systems as Feature Representations for Learning from Temporal Data

Peter Tino and **Yuan Shen**
School of Computer Science
University of Birmingham
UK

Special thanks to ...

- Marco Canducci
- Huanhuan Chen
- Nick Gianniotis
- Zoe Kourtzi
- Krasimira Tsaneva-Atanasova
- Kerstin Bunte
- Jochen Steil
- Ata Kabán
- Xin Yao
- ...

Classification



Classification in feature space

$$y = sgn(\mathbf{w}^T \mathbf{x} + b)$$

often

$$\mathbf{w} = \sum_i \alpha_i \cdot \mathbf{x}_i$$

and so

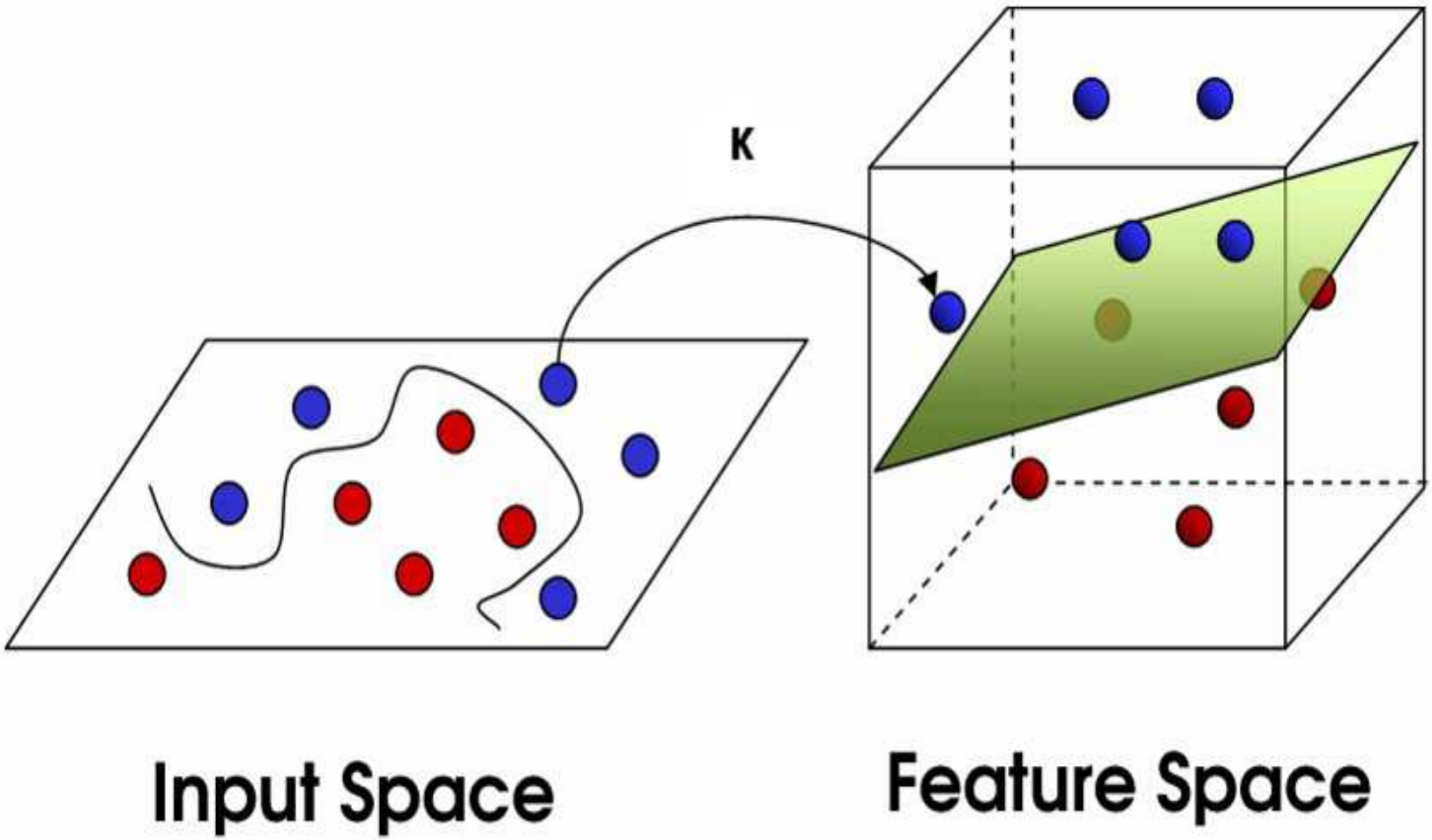
$$y = sgn\left(\sum_i \alpha_i \cdot \mathbf{x}_i^T \mathbf{x} + b\right)$$

or embed non-linearly to a high-dimensional feature space

$$y = sgn\left(\sum_i \alpha_i \cdot \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) + b\right) = sgn\left(\sum_i \alpha_i \cdot K(\mathbf{x}_i, \mathbf{x}) + b\right)$$

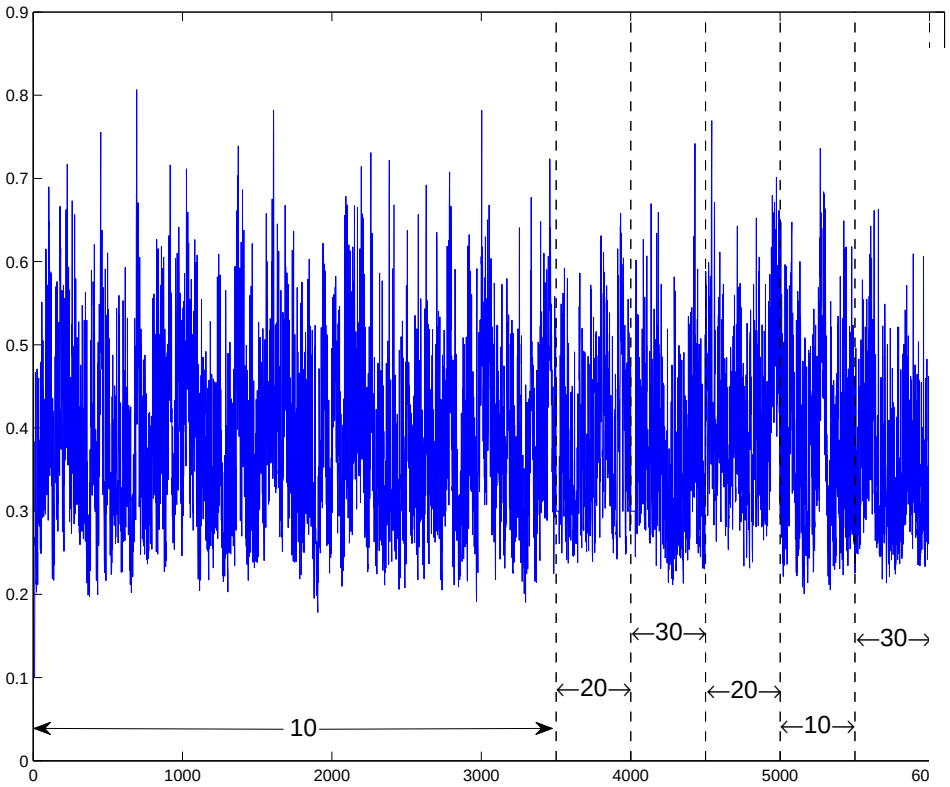
Appropriate input representations and their consistent treatment is the key!

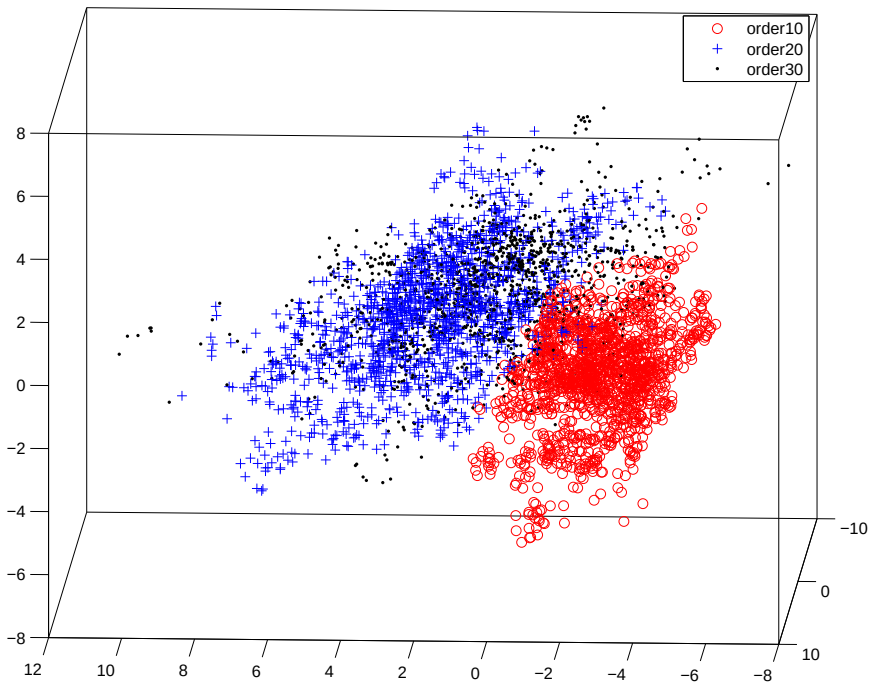
Classification in feature space



NARMA task - illustrative example

NARMA sequences - orders 10, 20 and 30 - represented as state space models





Learning in the Model Space Framework

- We do not assume all time series are collected on a fixed, regular time grid.
- Each data item is represented by a model that "explains" it
- Learning is formulated in the space of models (function space)
- Model class
  - flexible enough to represent variety of data items
  - sufficiently constrained to avoid overfitting

Parametric Dynamical Systems - ODE

Continuous-time deterministic dynamical system - mathematically represented as a multivariate Ordinary Differential Equation (ODE),

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{f}(\mathbf{x}_t; \boldsymbol{\psi}),$$

$\mathbf{x}_t \in X \subset \mathbb{R}^D$ state vector at time t
parameters $\boldsymbol{\psi}$ (include initial state $\mathbf{x}_0$).

Model parameters - $\boldsymbol{\theta} = \boldsymbol{\psi}$.

Parametric Dynamical Systems - SDE

Stochastic dynamical system - can be considered ODE driven by a multivariate random process parameterized by covariance matrix $\mathbf{\Sigma}$.

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t; \boldsymbol{\psi}) \ \mathrm{d}t + \mathbf{\Sigma} \ \mathrm{d}\mathbf{b}_t$$

vector $\mathbf{b}_t$ collects the D independent standard Brownian motions.

Each Measurement Sequence is Represented as a Model

Given a time series $\mathcal{Y} = \{(t_i, \mathbf{y}_i)\}_{i=1}^{L}$, a **Maximum Likelihood (ML) estimate** of $\boldsymbol{\theta}$ - maximize the likelihood function

$$p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{t}; \mathbf{R}) = \prod_{i=1}^{L} \mathcal{N}\left(\mathbf{y}_i \middle| \mathbf{x}_t(\boldsymbol{\theta}), t_i, \mathbf{R}\right)$$

for an ODE system and

$$p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{t}, \mathbf{R}) = \mathbb{E}_{\mathbf{x}_t|\boldsymbol{\theta}}\left[\prod_{i=1}^{L} \mathcal{N}\left(\mathbf{y}_i \middle| \mathbf{x}_t, t_i, \mathbf{R}\right)\right]$$

for an SDE system.

However, this ignores uncertainty around the model estimate. In cases **where only noisy and/or sparse data are available**, any point estimate of the model parameter is not a sufficient representation of the partially observed dynamical system.

Each Measurement Sequence is Represented as Posterior Over Models

Each measured sequence $\mathcal{Y}$ will be represented by

$$p(\boldsymbol{\theta}|\mathcal{Y}, \mathbf{R}) = p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{t}, \mathbf{R}) \propto p(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{t}, \mathbf{R}) \cdot p(\boldsymbol{\theta}),$$

where $p(\boldsymbol{\theta})$ is the prior over $\boldsymbol{\theta}$.

Supervised Learning - e.g. LiMS Classifier

Assume no additional relevant information for the classification could be extracted from observation noise or observation times, i.e. the observation noise and observation times processes are not conditional on the class label.

$$
\begin{aligned}
p(c|\mathcal{Y}) &= \iint \mathrm{d}\mathbf{x}_t \, \mathrm{d}\boldsymbol{\theta} \; p(c, \mathbf{x}_t, \boldsymbol{\theta}|\mathcal{Y}) \\
&= \iint \mathrm{d}\mathbf{x}_t \, \mathrm{d}\boldsymbol{\theta} \; p(c|\mathbf{x}_t, \boldsymbol{\theta}, \mathcal{Y}) \; p(\mathbf{x}_t, \boldsymbol{\theta}|\mathcal{Y}) \\
&= \int \mathrm{d}\boldsymbol{\theta} \; p(c|\boldsymbol{\theta}) \int \mathrm{d}\mathbf{x}_t \; p(\mathbf{x}_t, \boldsymbol{\theta}|\mathcal{Y})
\end{aligned}
$$

Key point of LiMS - all the relevant information in $(\mathbf{x}_t, \boldsymbol{\theta}, \mathcal{Y})$ for the class label prediction can be collapsed into the model $\boldsymbol{\theta}$.

LiMS Classifier

$$
\begin{aligned}
p(c|\mathcal{Y}) &= \int \mathrm{d}\boldsymbol{\theta}\ p(c|\boldsymbol{\theta}) \int \mathrm{d}\mathbf{x}_t\ p(\mathbf{x}_t, \boldsymbol{\theta}|\mathcal{Y}) \\
&= \int \mathrm{d}\boldsymbol{\theta}\ p(c|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\mathcal{Y}) \\
&= \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{Y})}\Big[p(c|\boldsymbol{\theta})\Big]
\end{aligned}
$$

Note that **the classifier $p(c|\mathcal{Y})$ operates on posterior distributions over models, $p(\theta|\mathcal{Y}_i)$, but is formulated based on classifier $p(c|\theta)$ operating in the model space**.

Training the LiMS Classifier

- Data set:

$$\mathcal{D} = \{(\mathcal{Y}_1, c_1), (\mathcal{Y}_2, c_2), ..., (\mathcal{Y}_N, c_N)\}$$

- Transformed data set:

$$\widetilde{\mathcal{D}} = \{(p(\boldsymbol{\theta}|\mathcal{Y}_1), c_1), (p(\boldsymbol{\theta}|\mathcal{Y}_2), c_2), ..., (p(\boldsymbol{\theta}|\mathcal{Y}_N), c_N)\}$$

- Pick your favourite (probabilistic) classifier.

- Pick a loss function $\mathcal{L}$.

- Derive learning equations for $\mathbf{w}$ by plugging in the LiMS classifier

$$p(c_i|\mathcal{Y}_i) = \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{Y}_i)}\Big[p(c_i|\boldsymbol{\theta}; \mathbf{w})\Big]$$

Training the LiMS Classifier

For example,

$$\mathbf{w}_{ML} = \underset{\mathbf{w}}{\text{argmax}} \prod_{i=1}^{N} \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{Y}_i)} \Big[p(c_i|\boldsymbol{\theta}; \mathbf{w}) \Big]$$

Stochastic Double-Well (SDW) system

$$\mathrm{d}x_t = \underbrace{4(x_t - a)(d^2 - x_t^2)}_{f(x_t)} + \kappa^2 \cdot \mathrm{d}b_t,$$

b_t - univariate standard Brownian motion

$\boldsymbol{\theta} = (d, \kappa, a)$ model parameters, well location parameter d, well asymmetry parameter a and standard deviation κ of the dynamic noise.

Drift term $f(x_t)$ is not explicitly time-dependent $\implies$ dynamics governed by potential $u(x)$ with $f(x) = -\nabla_x u(x)$.

Equilibrium probability distribution of x is $p^{\mathbf{eq}}(x) \propto \exp\left(-\frac{u(x)}{\kappa^2}\right)$.

Stochastic dynamics

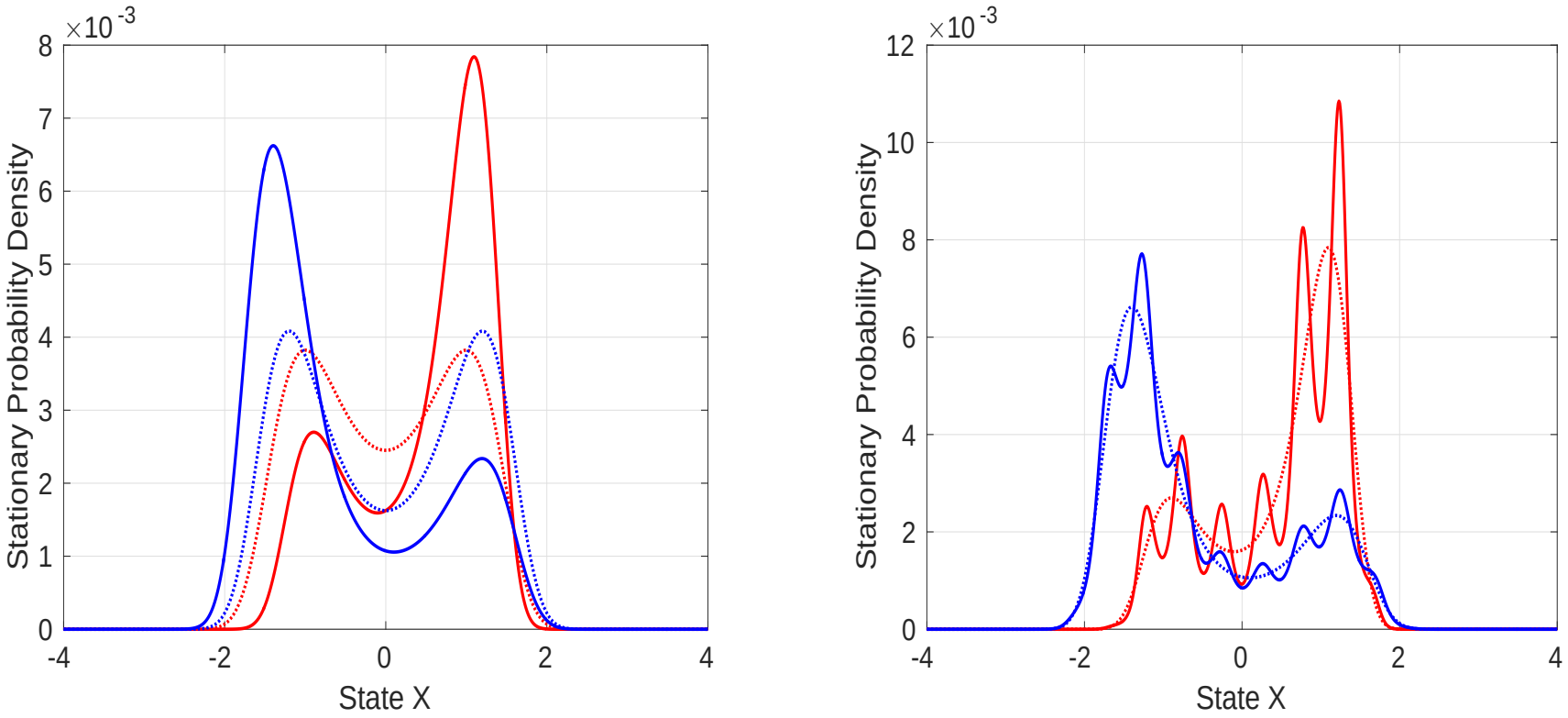


Figure: Left Panel: Equilibrium probability distribution of states x of four example Stochastic Double-Well Systems. Right Panel: The same as in in Left Panel but for Stochastic Multi-Well Systems.

Observed trajectories

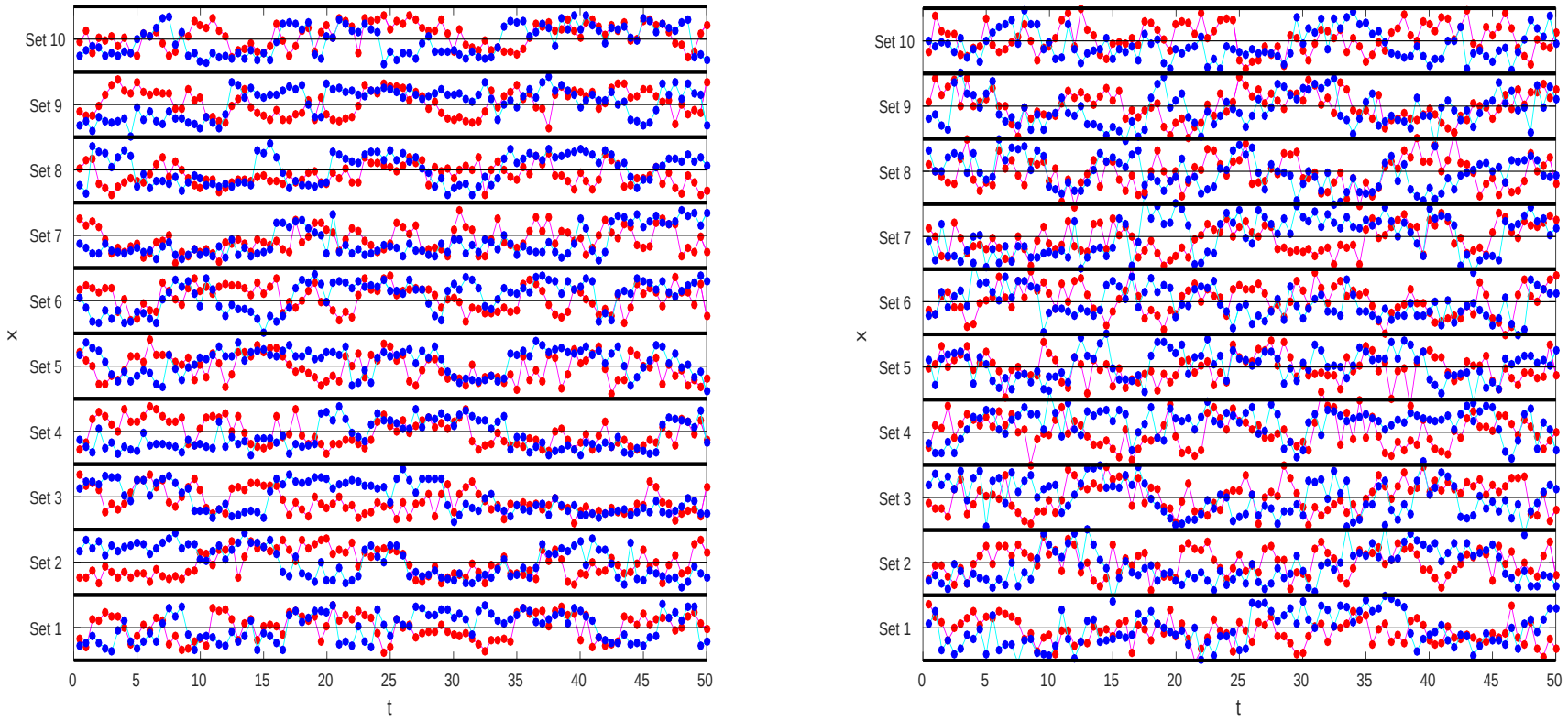


Figure: Variance σ^2 of Gaussian distributed observation noise is 0.04 (left) and 0.36 (right).

Results - DWS - dominant well

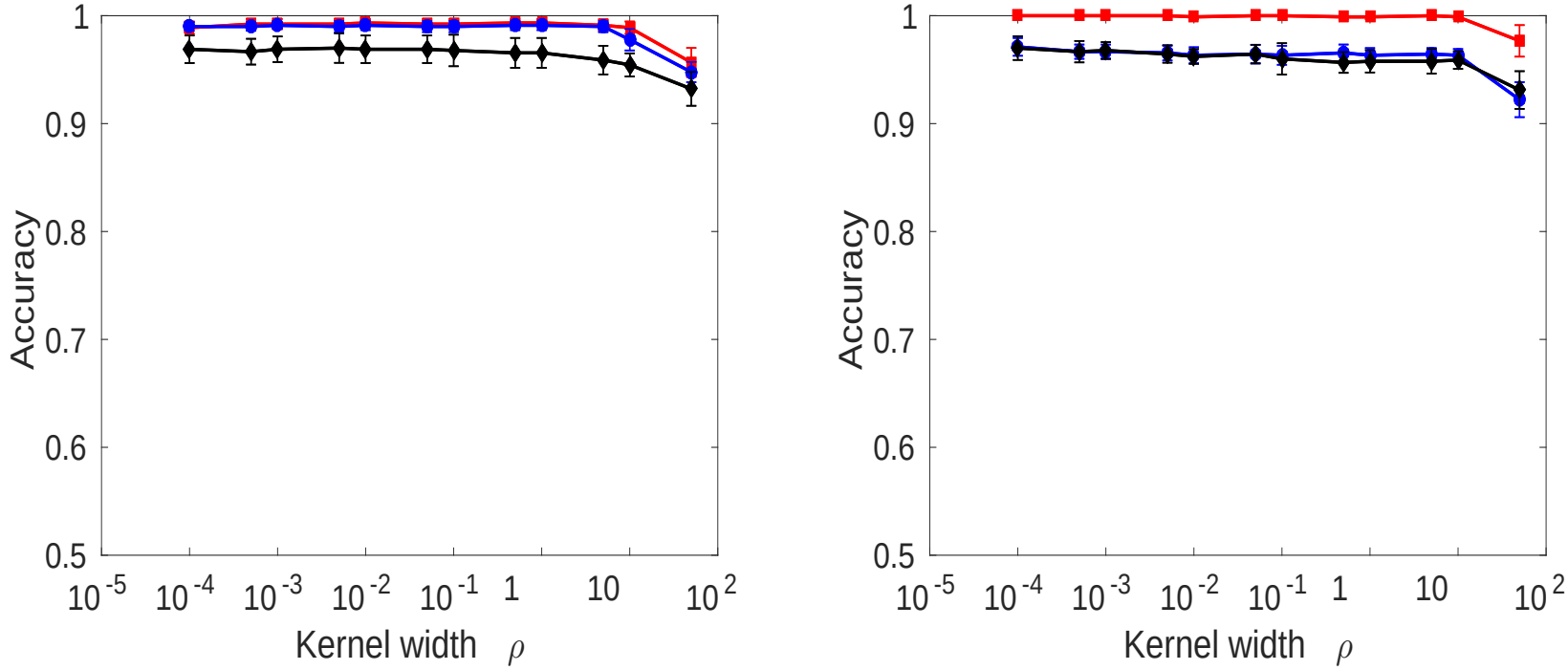
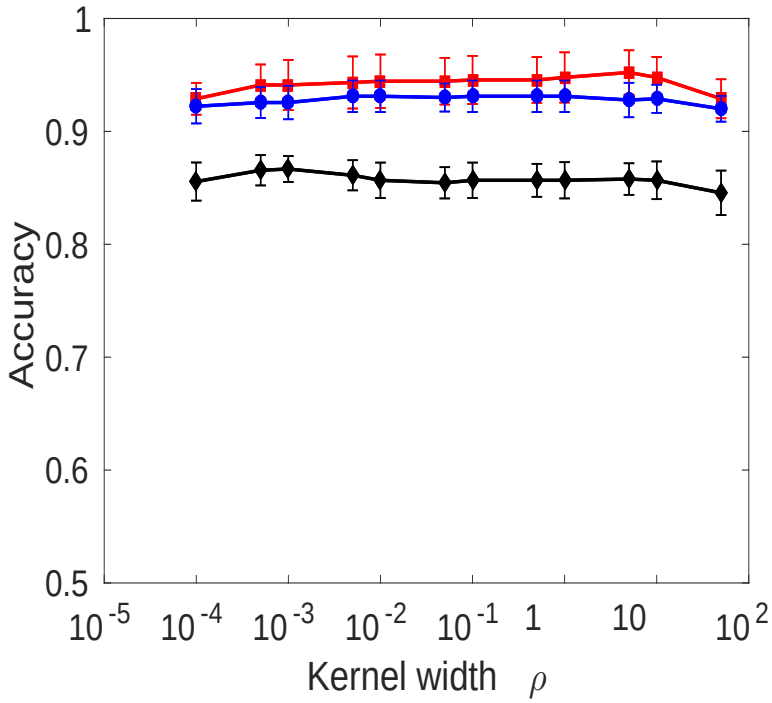
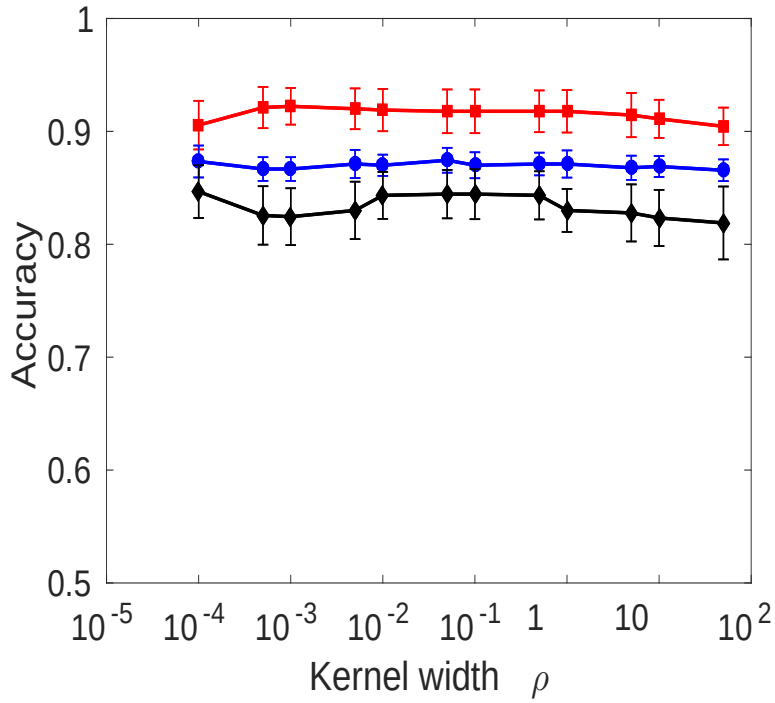


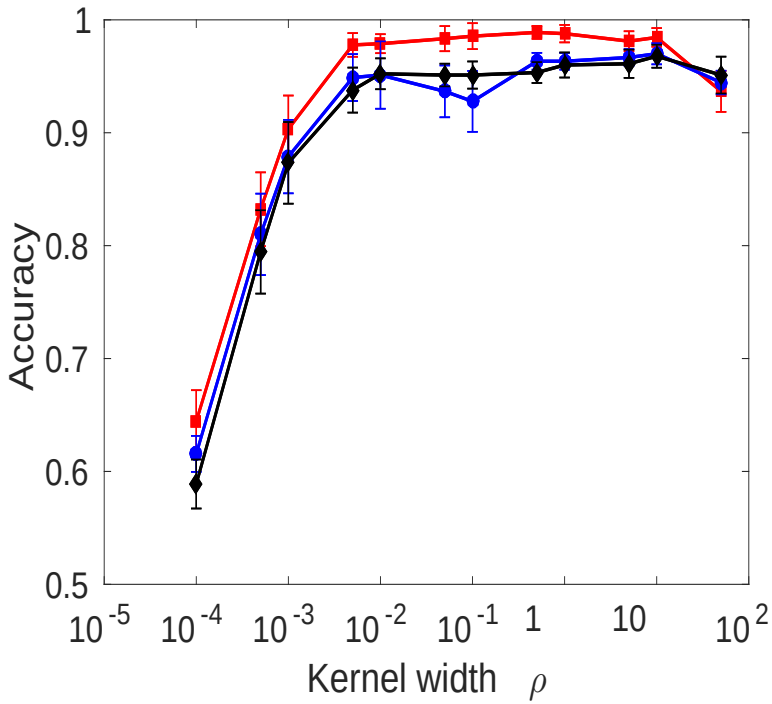
Figure: Gaussian distributed observation noise, $\sigma^2 = 0.04$ (left) and $\sigma^2 = 0.36$ (right). Random observation time subsampling, observation time frequency high (red) $\rightarrow$ low (black, blue).

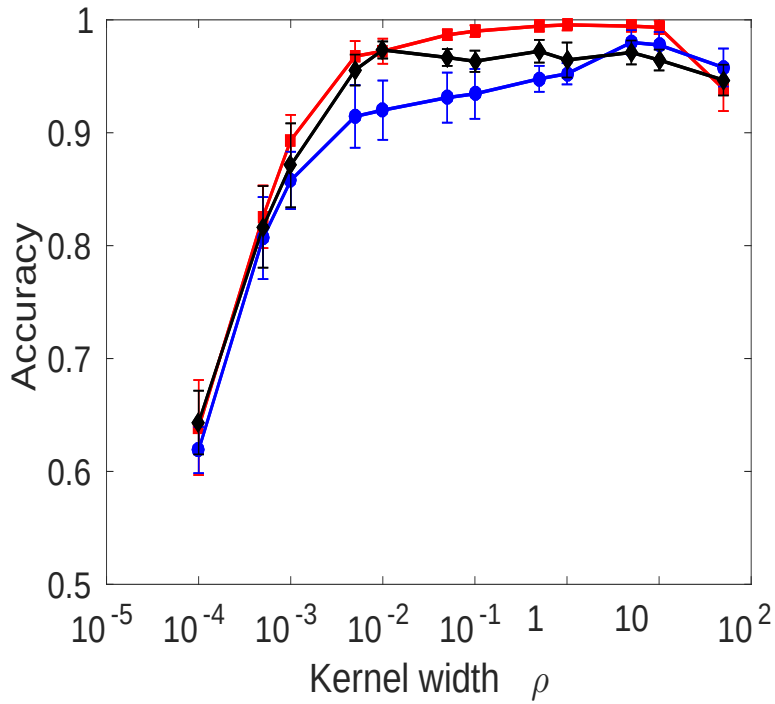
Results - DWS - only weakly dominant well



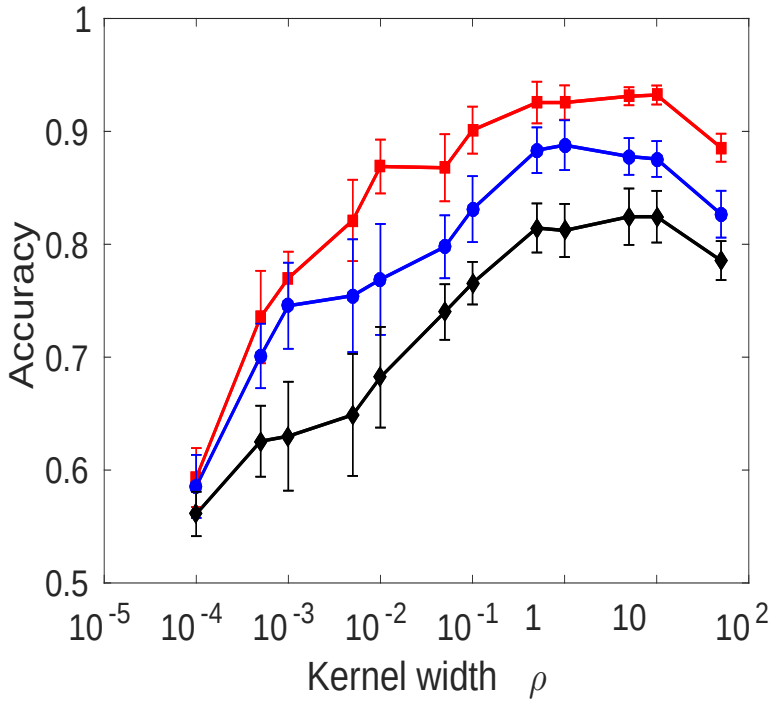


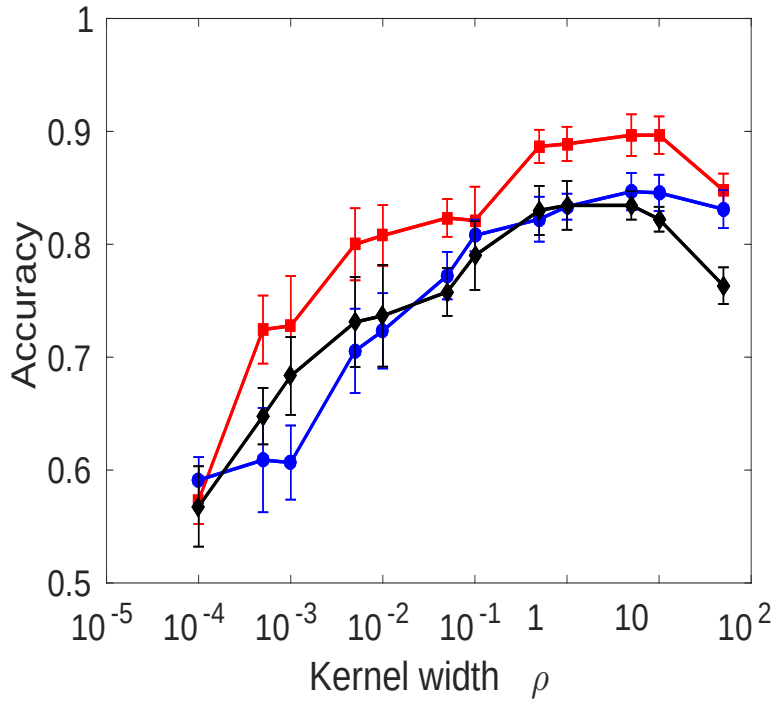
Results - Multi-Well - dominant well





Results - Multi-Well - only weakly dominant well





GnRH model hierarchy

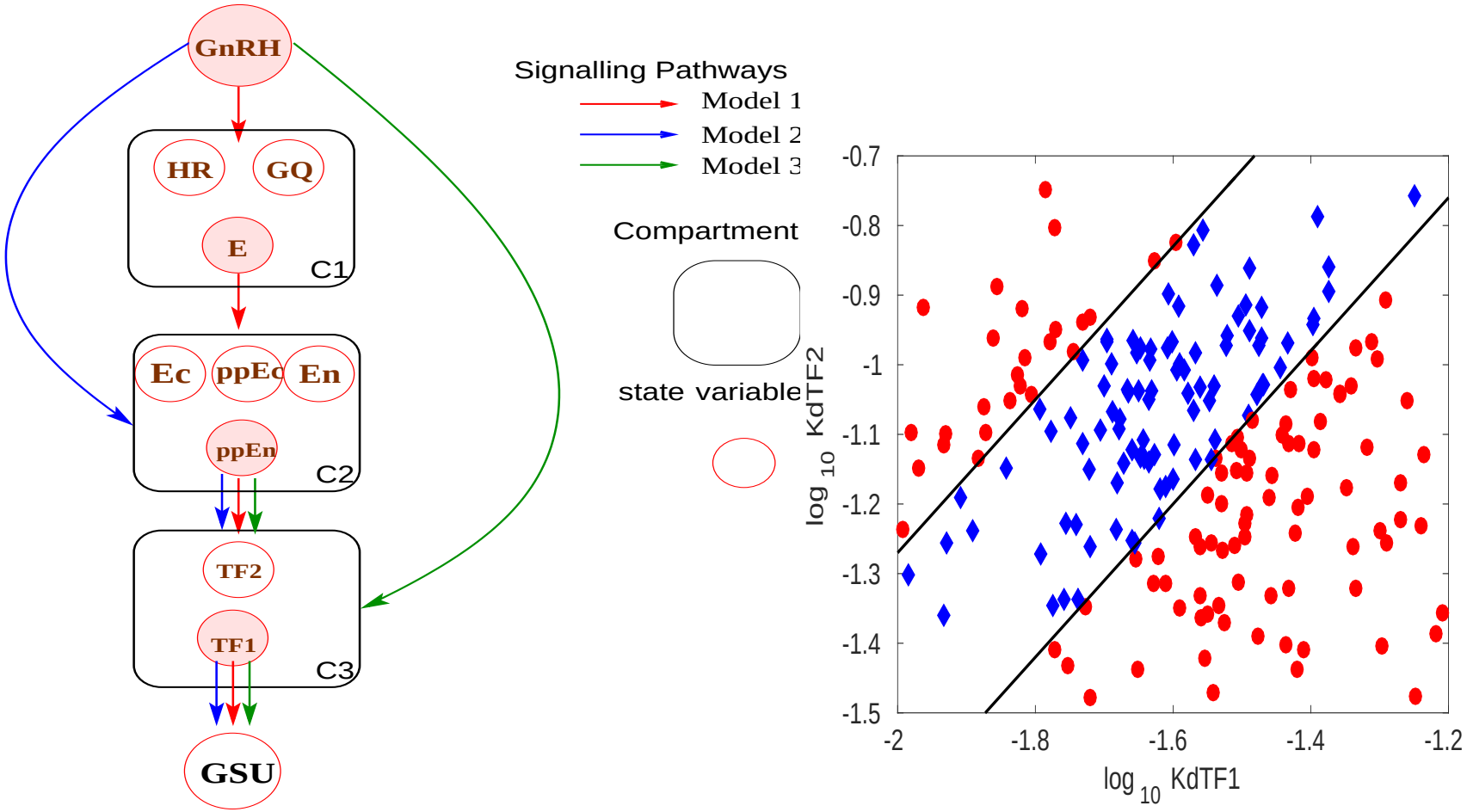


Figure: Left - 3 nested GnRH signalling models. Right - classes of GnRH signalling models: normal (Blue Diamonds) and ubnormal (Red Disks).

Gonadotropin-Releasing Hormone Signalling model

an example pathway model

GnRH signal is the chemical signal which stimulates the reproductive endocrine system.

ODE system - **11 state variables**:
- concentrations of gonadotropin releasing hormones [GnRH] (driving input)
- gonadotropin hormones [GSU] (measurable output)

Remaining state variables - grouped into 3 compartments along the signalling pathway:
- C1 for GnRH binding process
- C2 for extracellular signal regulated kinase (ERK) activation
- C3 for transcription factor (TF) activation.

Issues to Study

We'd like to study two important issues for classifying partially observed dynamical systems (PODS):

- **The influence of model uncertainty on classification in the model space.**

  It is natural to expect that the posterior over possible models, given the observations, is a better (model space) representation of the observed time series than a single model, e.g. MAP point estimate.
  It is also natural to expect that the classification performance will increase with reducing model uncertainty.
  We use the level of observation noise, or the number of observations as surrogate uncertainty measures.

Issues to Study

- **Performance degradation when the inferential model class used to represent the observed time series through posteriors is a reduced sub-model class of the true model class generating the training and test data.**

  For example, in real-world applications, it is inevitable that there is a gap between the real-world and the mathematical model developed to account for it. *A reduced model could be used to represent time series, as long as it captures characteristics relevant for the given classification task.*

Experimental Setup

Generated two independent sets of GnRH models for training and testing (200 labelled models each).

Randomly sampled 400 parameter vectors $\boldsymbol{\theta}_{\mathbf{GnRH}} = \left(\log K_{d_{\mathrm{TF}_1}}, \log K_{d_{\mathrm{TF}_2}}, t_p\right)$ of the GnRH model. Each of the three model parameters are sampled from the corresponding Gaussian distribution truncated to the permissible range.

Generated a variety of observation time series with different observation settings (number of observations, observation times and observational noise level).

Experimental Setup

Simulated GnRH (8-hour window) and recorded [GSU] trajectory at six different pulse frequencies.

Initial values of state variables in GnRH model were fixed but the trajectory over the first half an hour is discarded. The transient behaviour has been ignored and only the attractor part of individual trajectories is used for sampling observations. Hence, initialisation of the GnRH model has little influence on inferring the underlying model from observations.

15 observation sets using different pairs of observation noise level σ and the inter-sample interval (ISI). The observation sets are organised in three groups (5 sets in each group):

Observation Sets - Group 1

Observations sampled regularly every $ISI = 75$ minutes over 7.5 hours, yielding 6 observation times.

The level σ of observation noise in the 5 observation sets was set to 0.1, 0.03, 0.01, 0.005 and 0.001.

The observation sets in this group correspond to the partially observed GnRH model with **five different levels of model uncertainty controlled by** σ.

Observation Sets - Group 2

Observation noise fixed to $\sigma = 0.03$.

We varied the number of observation times within the 7.5 hour window. In particular, the 5 observation sets contained 5, 6, 10, 15 and 30 observation times with $ISI = 90$, 75, 45, 30 and 15, respectively. The observation times were placed randomly with uniform distribution over the 7.5 hour window

Five different levels of model uncertainty are controlled by the sparsity of observations.

Results - LiMS, Groups 1 and 2

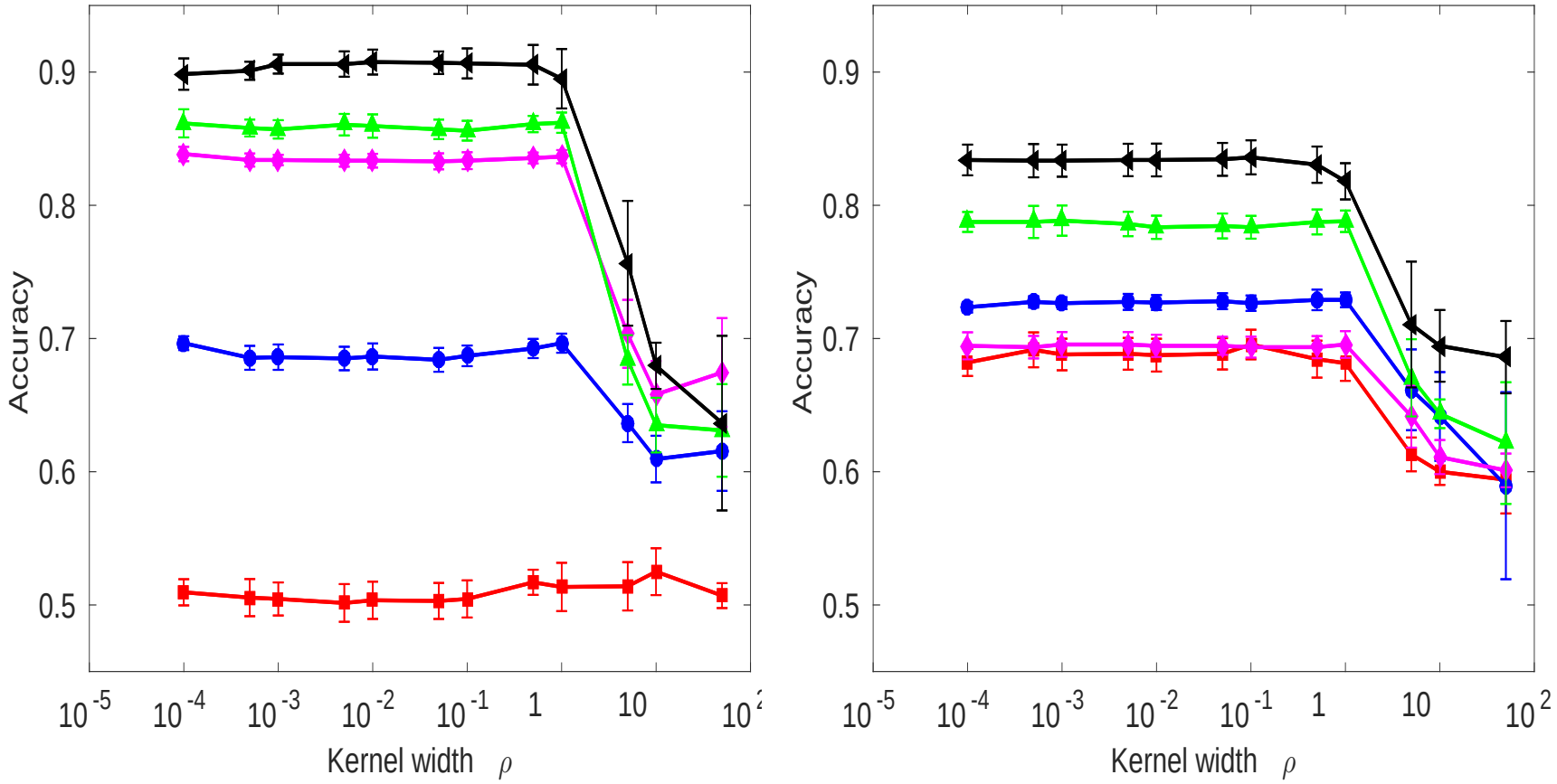


Figure: Left: $ISI = 75$ (regular sampling), noise st dev varies across $\{0.1, 0.03, 0.01, 0.005, 0.001\}$ (red, blue, magenta, green, and black, respectively. Right: St dev of observation noise fixed to 0.3, ISI varies across $\{15, 30, 45, 75, 90\}$ (black, green, magenta, blue, and red, respectively). The observation times are random and the ISI-values given are the expected value.

Results - PPK, Groups 1 and 2

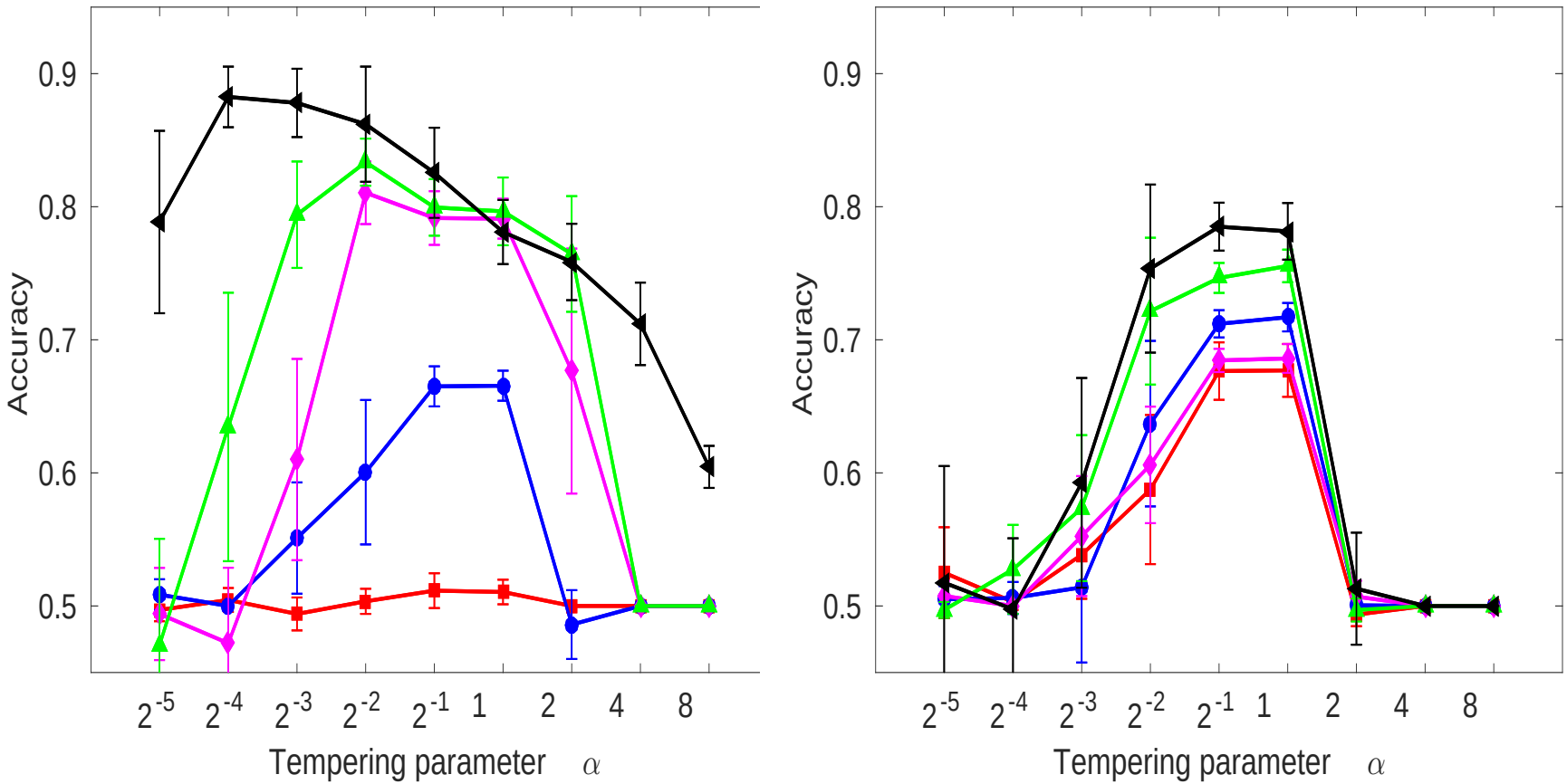


Figure: Classification performance as function of tempering parameter.

Comparing LiMS and PPK

Entropy	(σ, ISI)	p-value
1.7	(0.001, 75)	**0.01**
4.3	(0.005, 75)	**0.00**
5.2	(0.01, 75)	**0.02**
5.6	(0.03, 15)	**0.00**
6.0	(0.03, 30)	**0.00**
6.2	(0.03, 45)	**0.01**
6.4	(0.03, 75)	**0.00**
6.5	(0.03, 90)	0.07
7.3	(0.1, 75)	0.15

Table: Sign-rank tests (p-values) at different levels of model uncertainty. One-sided hypothesis: LiMS outperforms PPK.

Missmatch Between the Generative and Inferential Models

Data Sets	(σ, ISI)	*M1*	*M2*	*M3*
Group 1	(0.001, 75)	0.91 ± 0.02	0.88 ± 0.03	0.90 ± 0.01
	(0.01, 75)	0.84 ± 0.01	0.84 ± 0.01	0.83 ± 0.01
	(0.1, 75)	0.52 ± 0.01	0.54 ± 0.01	0.54 ± 0.01
Group 2	(0.03, 90)	0.83 ± 0.02	0.82 ± 0.02	0.82 ± 0.01
	(0.03, 30)	0.69 ± 0.01	0.71 ± 0.01	0.71 ± 0.01
	(0.03, 30)	0.68 ± 0.02	0.69 ± 0.01	0.68 ± 0.02

Table: LiMS using inferential GnRH models *M1*,*M2*, and *M3*.

.

Missmatch Between the Generative and Inferential Models

Data Sets	(σ, ISI)	*M1*	*M2*	*M3*
Group 1	(0.001, 75)	0.91 ± 0.02	0.88 ± 0.03	0.90 ± 0.01
	(0.01, 75)	0.84 ± 0.01	0.84 ± 0.01	0.83 ± 0.01
	(0.1, 75)	0.52 ± 0.01	0.54 ± 0.01	0.54 ± 0.01
Group 2	(0.03, 90)	0.83 ± 0.02	0.82 ± 0.02	0.82 ± 0.01
	(0.03, 30)	0.69 ± 0.01	0.71 ± 0.01	0.71 ± 0.01
	(0.03, 30)	0.68 ± 0.02	0.69 ± 0.01	0.68 ± 0.02

Table: LiMS using inferential GnRH models *M1*,*M2*, and *M3*.

.

Missmatch Between the Generative and Inferential Models

For classification of PODS via Learning in the Model Space framework, it is **not necessary for the inferential model structure to be a perfect model** of the underlying dynamical system generating the data, **as long as the reduced complexity inferential model structure captures the essential characteristics needed for the given classification task**.

Interested?

- K. Bunte, D.J. Smith, M.J. Chappell, Z.K. Hassan-Smith, J.W. Tomlinson, W. Arlt, P. Tino: **Learning Pharmacokinetic Models for in vivo Glucocorticoid Activation**. Journal of Theoretical Biology, 455, pp. 222-231, 2018.
- P. Tino: **Asymptotic Fisher Memory of Randomized Linear Symmetric Echo State Networks**. Neurocomputing, 298, pp. 48, , 2018.
- Y. Shen, P. Tino, K. Tsaneva-Atanasova: **Classification framework for partially observed dynamical systems**. Physical Review E, 95, 043303,2017.
- H. Chen, F. Tang, P. Tino, A. G. Cohn, X. Yao: **Model Metric Co-learning for Time Series Classification**. IJCAI 2015, pp. 3387-3394, AAAI Press, 2015.
- H. Chen, P. Tino, X. Yao, A. Rodan: **Learning in the Model Space for Fault Diagnosis**. IEEE TNNLS, 25(1), pp. 124-136, 2014.