Sebastian Springer

BAYESIAN INFERENCE BY INFORMATIVE GAUSSIAN FEATURES OF THE DATA

ACTA UNIVERSITATIS LAPPEENRANTAENSIS 950

Sebastian Springer

BAYESIAN INFERENCE BY INFORMATIVE GAUSSIAN FEATURES OF THE DATA

Dissertation for the degree of Doctor of Science (Technology) to be presented with due permission for public examination and criticism at Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland on the 5th of February, 2021, at noon.

Acta Universitatis
Lappeenrantaensis 950

Supervisor Professor Heikki Haario
LUT School of Engineering Science
Lappeenranta-Lahti University of Technology LUT
Finland

Reviewers Professor Timo Tiihonen
Faculty of Information Technology
University of Jyväskylä
Finland

Professor Nikolay Kuznetsov
Applied Cybernetics Department
Saint Petersburg State University
Russia

Opponent Professor Matti Lassas
Department of Mathematics and Statistics
University of Helsinki, Helsinki
Finland

ISBN 978-952-335-627-6
ISBN 978-952-335-628-3 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491

Lappeenranta-Lahti University of Technology LUT
LUT University Press 2021

Abstract

Sebastian Springer
Bayesian inference by informative Gaussian features of the data
Lappeenranta 2021
62 pages
Acta Universitatis Lappeenrantaensis 950
Diss. Lappeenranta-Lahti University of Technology LUT
ISBN 978-952-335-627-6, ISBN 978-952-335-628-3 (PDF), ISSN-L 1456-4491, ISSN 1456-4491

Given a set of measurements, a model can be calibrated to data by minimising a cost function that captures the model-data difference. The standard cost function for deterministic models is the residual sum of squares between the model output and the reference data. Sequential data assimilation methods can be used for stochastic models whenever it is possible to estimate the next state of the system from the current values. Chaotic dynamical systems are a specific type of models for which the classic approaches are often not available. Especially so, if the time gap between consecutive observations is higher than the predictable time interval. There is, therefore, a need for new methods for such problems, defined even as insolvable in the literature. In recent years, Haario et al. presented a solution to estimate the model parameters of chaotic systems by the so-called Correlation integral likelihood (CIL). The idea behind it is to use a generalisation of the correlation integral sum as a feature vector and build a Gaussian likelihood by estimating the variability from the repetitions of the experiment. In this work, we will further develop those ideas. We will first see how to generalise the CIL approach to use statistics that best fit the specific problem studied. Furthermore, we will see how it is possible to create different feature vectors from one or several data sets concerning the same phenomenon and combine them in a single likelihood to improve the estimation of the parameters. It will be shown how by using this method it is possible to estimate the parameters of chaotic systems when only the state vector is observed, or only a part of it, and when the measurements are arbitrarily sparse and scattered. We emphasise that the original distance-based CIL is by no means the only way to construct provably Gaussian feature vectors. Indeed, the selection of features to produce normally distributed empirical cumulative distribution function (eCDF) vectors should be carefully done case by case. Moreover, we will present tools to diagnose goodness-of-fit or a possible lack of fit after the posterior distribution has been obtained. To make our approach available also for cases with expensive forward models, we studied the combination with the Local Approximation MCMC (LA-MCMC) algorithms. The idea behind these types of sampling methods is to construct local polynomial approximations of the likelihood by regression over a set of neighbouring evaluations of the expensive likelihood, and make it more and more precise as the sampling proceeds. By this method, the number of expensive likelihood evaluations could be reduced by orders of magnitude without significantly impacting the precision of the resulting posterior distribution.

Keywords: Chaotic dynamical systems, Bayesian inference, Gaussian likelihood, feature vectors, LA-MCMC

Acknowledgements

This study was carried out in the department of Computational and Process Engineering at Lappeenranta LUT University, Finland, between 2016 and 2020.

First of all, I would like to express my gratitude to my supervisor, prof. Heikki Haario. I can state, with great confidence, that the decision of coming to Finland to learn from Heikki how to be a researcher has been one of the best choices I've made to date. I had the pleasure of appreciating his charisma and expertise that make him one of the leaders of this beautiful community who welcomed me warmly. Passion, motivation, respect, friendship and joy has been the leitmotiv that accompanied both our long research days and different leisure activities like hiking, biking, climbing and grilling.

A special thank goes to all my colleagues from LUT, especially Vladimir Shemyakin, Alexkey Kazarnikov, Alex Bibov, Peter Jones, Matylda Jablonska-Sabuka, prof Matti Heiliö, prof. Tuomo Kauranne, prof. Lassi Roininen, prof. Tapio Helin and prof. Bernardo Barbiellini. Their competence, loyalty and friendliness supported me to move forward in my studies.

I would like to extend my thanks to all the management and to all the secretaries of LUT who have always put research, teaching and the health of us researchers in the foreground, facilitating all the bureaucratic aspects.

Next I would like to thank all the people with whom I had the pleasure to collaborate during this years; from Finnish Meteorological Institute Janne Hakkarainen, Pirkka Ollinaho, Hannakaisa Lindqvist, Johanna Tammen and especially Jouni Susiluoto with whom I shared a great experience in Cambridge Massachusetts; all the members of the Finnish Inverse Problems Society; the MIT UQ lab and expecially prof. Youssef M. Marzouk and Andrew Davis; from the Montana University prof. Leonid Kalachev, prof. John Bardsley and Denis Shchepankin; from University of Helsinki prof. Eero Saksmann; from Università della Svizzera italiana prof. Antonietta Mira; Martin Simon from Deka Investment GmbH; prof. Andreas Hauptmann and the whole research group of the Oulu University.

I want to to express a special gratitude to my partner Ramona, our kids, both Alessandro and the little one we are ready to welcome soon; my siblings Christian and Serena, my parents Cristina and Daniel and all the other relatives; to my uncle Mario and all the other friends.

Sebastian Springer
Jannuary 2021
Helsinki, Finland

To all the people I love.

Sebastian

Contents

Nomenclature

Abbreviations

2D	Two dimensional
3D	Three dimensional
ABC	Approximative Bayesian Computation
AM	Adaptive Metropolis
BSL	Bayesian Synthetic Likelihood
CIL	Correlation Integral Likelihood
CLT	Central limit theorem
CPU	Central processing unit
DE	Differential Evolution
DRAM	Delayed Rejection Adaptive Metropolis
DR	Delayed Rejection
eCDF	Empirical cumulative distribution function
ECMWF	European Centre for Medium-Range Weather Forecasts
GLIF	Gaussian likelihood by informative feature vectors
IFS	Integrated Forecast System
KS	Kuramoto-Shivashinsky
LA-MCMC	Local approximation Markov chain Monte Carlo
MAP	Maximum a posteriori
MCMC	Markov chain Monte Carlo
MH	Metropolis Hasting
NLL	Negative log likelihood
ODE	Ordinary differential equation
PDE	Partial differential equation
QG	Quasi-geostrophic
RSS	Residual sum of squares
SDE	Stochastic differential equation
UQ	Uncertainty quantification

1 Introduction

There has always been the need to understand, study and predict the behaviour of the phenomena that surround us. Over the years, this has led to increasingly precise techniques for measuring and modelling events. Despite this, two fundamental problems remain: an error component is always introduced during the measurement collection process and the finite number of data collected always represent only one part of the phenomena studied. It follows that there is no single solution to the problem of estimating the parameters of a model. There is, therefore, a need to obtain the distribution of the possible models that characterise that specific set of measurements.
Given a set of measurements, the parameters of the reference model are typically, inferred by minimising a cost function that captures the model-data difference. The most classic of the cost functions is the residual sum of squares (RSS) between the model output and the reference data. In the Bayesian approach, a likelihood can be formulated by combining the cost function with the available prior knowledge to obtain the posterior distribution of the model parameters. In numerical practice, this is typically implemented through the use of Markov chain Monte Carlo (MCMC) methods, see Robert and Casella (2004). The posterior distribution provides a solution to the problem in a complete probabilistic sense. However, this approach is not always directly applicable.
Stochastic differential equations (SDE) are a particular type of equations whose dynamics can generally be divided into two components: drift or deterministic part and diffusion or stochastic part. Their general formula is given by

$$dx = f(x; t; \theta)dt + Lw(t), \tag{1.1}$$

where $x \in R^n$, $w(t)$ is a zero mean n-dimensional Gaussian process, θ denotes the parameters of interest and L is the diffusion term which can be constant or depend on x, t *or* θ. For basic settings - with a linear drift and constant diffusion - the standard RSS-based likelihood formulation can replaced by the so-called filter likelihood that is based on stepwise predictive integrations followed by a correction by data Särkkä (2013). For non-linear drifts, or cases where the diffusion term depends on the states of the system, several approaches of sequential data assimilation methods have been developed. They are available whenever it is possible to estimate the next state of the system from the current values. Typically this requires that measurements are available sufficiently dense in time. Examples of these methods are the Kalman filter, extended Kalman filter, ensemble Kalman filter and non-linear Gaussian filtering, see Särkkä (2013). All these methods are based on the idea of constraining the problem to short time intervals, so as to avoid any divergences from the data due to the stochastic nature of the system considered. However, the filtering algorithms introduce their own 'tuning' parameters that may have to be estimated together with the model parameters, such as the length of the assimilation window, the model error covariance matrix, covariance inflation, etc. This might have the impact of biasing the model parameter estimation, see Hakkarainen et al. (2013).

Chaotic phenomena provide a different class of models for which the standard methods may not be suitable. The reason again is that, even if deterministic, for long enough time

intervals, the system becomes unpredictable.

There are several specific situations where the parameters of a chaotic dynamical system can be estimated in more or less traditional ways, however. The estimation task trivially reduces to the classical deterministic setting, if enough data is given on a short time interval, where the system remains predictable. Another situation that enables a straightforward estimation is given by systems that can be written in the form $\dot{s} = G(s)\theta$, i.e. where the right-hand side depends on unknown parameters θ in a linear way. The classical 3D Lorenz system, for instance, is of this form. If all the components of the state s and the time derivatives $\dot{s}$ are known at enough measurement points, the system boils down to a matrix equation that can be solved for θ by linear regression. Moreover, classical statistical regression methods are available to determine the structure of $G(s)$ in case it is polynomial, see Brunton et al. (2016). However, the approach breaks down if only s without $\dot{s}$, or a subset of the components of $s, \dot{s}$ are observed.

The filtering methods developed for stochastic systems can also be applied to chaotic systems. In Hakkarainen et al. (2012), the parameters of a chaotic model were estimated using the filter likelihood provided by an extended Kalman filter. Even if the stochastic diffusion part of Eq. (1.1) is now missing, the filtering formalism is kept as such, interpreting the model error as the prediction error, or model bias. The tuning issues naturally appear again, as discussed in Hakkarainen et al. (2013). Conditional Gaussian statistics have been studied in Chen and Majda (2016) as a way of filtering nonlinear turbulent dynamical systems. In case the drift function has a suitable structure, it can be used together with filtering methods to the estimate model parameters of systems of the form of Eq. (1.1) even for chaotic drifts.

Operational ensemble prediction systems may be reformulated to include parameter variations and be employed for parameter estimation, even if no ensemble filtering is performed. See Jarvinen et al. (2011); Laine et al. (2011); Shemyakin and Haario (2018), where such approaches were developed and applied to the Integrated Forecast System (IFS) at the European Centre for Medium-Range Weather Forecasts (ECMWF). These methods, however, are heuristic and require short predictable time intervals again.

The above methods cease to be usable if only sparse data is available, in particular when the time between two consecutive measurements exceeds the predictable time interval. There is, therefore, a need for methods for parameter estimation in situations where the classical formulations cannot be employed.

An alternative approach was suggested by Wood (2010). The key idea was to use *synthetic likelihoods*: in order to test the compatibility of data against model simulations for a given model parameter value, a likelihood based on some summary statistics is recreated by repeated 'synthetic' model simulations. In a Gaussian case, this would amount to estimating the mean and covariance of a summary statistics of the simulations, and then testing whether the given data fits this likelihood. Summary statistics such as mean, autocorrelation or regression coefficients were used by Wood (2010). The approach has since been further studied and extended, see Price et al. (2018) and the references therein for discussions on Bayesian Synthetic Likelihood (BSL) as well as comparisons to the widely used Approximative Bayesian Computation (ABC) approach.

A related but 'opposite' approach was presented by Haario et al. (2015) to study chaotic

models. The difference is that, for Haario et al. (2015), a large set of data, or repetitions of a given experimental setting, it is assumed to be available. From the observations, it is then possible to estimate a likelihood for a summary statistics for a subset of the data. This is done offline, before starting the parameter estimation. When using the likelihood for inference, simulations are only needed for the respective subset amount of data. This can lead to substantial savings of computing time. Another key point of Haario et al. (2015) was the selection of the summary statistics. A variant of the concept of *Correlation Dimension*, a fractal dimension definition, e.g. Grassberger and Procaccia (1983b,a)), was developed. Rather than characterising an intrinsic dimension of a trajectory, the *Correlation Integral Likelihood* (CIL) , gives a statistical distance concept for two different trajectories. A remarkable property of the selected summary statistics is that it actually is the *empirical cumulative distribution function*, eCDF, of the distances of trajectory vectors. The theorems in the so-called U-statistics literature of Borovkova et al. (2001); Neumeyer (2004), are generalizations of classical versions of Central Limit Theorem, CLT, and guarantee that eCDF as summary statistics – or feature vector, FV, as it will be called – is theoretically Gaussian. In numerical calculations, the construction of the empirical likelihood thus reduces to estimating the mean and covariance of such feature vectors. The Gaussianity can be verified using standard normality tests.

The main advantage of both approaches is that by employing informative summary statistics of the data inference it can be performed even for complex data and models and in situations where standard likelihood constructions are not available. An advantage of BSL is that it works with a limited amount of data.The downside is that the procedure is computationally intensive, as for every new parameter candidate, the model needs to be integrated several times in order to compute the summary statistics. Naturally, this can be alleviated by parallelising the integration, if possible. On the other hand, the downside of the CIL approach is the requirement for a larger amount of data, while the (CPU) times are typically much smaller. As will be shown here, parallel calculations can be effectively used with the CIL approach as well.

In this work, we will further develop the ideas presented in Haario et al. (2015). We will first see how to generalise the CIL approach in order to use statistics that best fit the specific problem studied. Furthermore, we will see how it is possible to create different feature vectors from one or several data sets concerning the same phenomenon and combine them in a single likelihood to improve the estimation of the parameters. By using this method, it will be shown how it is possible to estimate the parameters of chaotic systems when only the state vector is observed, or only a part of it, and when the measurements are arbitrarily sparse and scattered. The dynamics of the system is identified if time-derivative information is available. Both low-dimensional and high-dimensional cases will be analysed to see the robustness of the approach, as the dimension of the state increases. We emphasise that the original distance-based CIL is by no means the only way to construct provably Gaussian feature vectors, indeed the selection of features to produce the normally distributed eCDF vectors should be carefully done case by case. Moreover, we will present tools to diagnostic goodness-of-fit or a possible lack of fit after the posterior distribution has been obtained.

A concrete problem regarding standard MCMC methods is that each likelihood evaluation needs a new model evaluation. Several thousand evaluations are often needed to have a reliable approximation of the posterior distribution. This makes these approaches impractical for models with a very high computational cost. To speed up the parameter estimation, we show how to combine the likelihoods studied here with the Local approximation MCMC (LA-MCMC) approach, see Conrad et al. (2016); Davis (2018); Davis et al. (2020). This algorithm is based on the idea of creating local polynomial approximations of the likelihood values at given points of the posterior. Regression surfaces are fitted to these expensively calculated'full' likelihood values, but the sampled chain values are obtained by the response surface values only, practically for free. The approximations are created using specific criteria to ensure that, at each step of the algorithm, the error introduced by the polynomial approximation is insignificant.
It will be shown how, by way of a couple hundred full likelihood evaluations, it is possible to obtain chains with two orders of magnitude more samples. This enables the MCMC sampling of parameters even for models with an excessively high computational cost for standard sampling methods.

2 Bayesian inference

2.1 Markov chain Monte Carlo

Let us denote by

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = f(\mathbf{u}, \theta), \;\; \mathbf{u}(t=0) = \mathbf{u}_0, \tag{2.1}$$

a dynamical system with state $\mathbf{u}(t) \in \mathbb{R}^n$, initial state $\mathbf{u}_0 \in \mathbb{R}^n$, and parameters $\theta \in \mathbb{R}^d$. The time-discretised system, with time steps $t_i \in \{t_1, \ldots, t_\tau\}$ giving the observation points, can be written as

$$\mathbf{u}_i = F(t_i; \mathbf{u}_0, \theta) + \epsilon_i. \tag{2.2}$$

The model–observation mismatch at a collection of times t_i, can be written as

$$\epsilon_i = \mathbf{u}_i - F(t_i; \mathbf{u}_0, \theta). \tag{2.3}$$

Given the uncertainty in the measurement error, a point estimate of determining θ by u cannot be considered as a full answer to this problem.
In the Bayesian parameter estimation approach, θ is interpreted as a random variable and the goal is to find the *posterior distribution* $\pi(\theta|\mathbf{u})$ of the parameters which gives the probability density for values of θ, given measurements $\mathbf{u}$. Using the Bayes' formula, the posterior density is

$$\pi(\theta|\mathbf{u}) = \frac{l(\mathbf{u}|\theta)p(\theta)}{\int l(\mathbf{u}|\theta)p(\theta)d\theta},$$

where $l(\mathbf{u}|\theta)$ is the *likelihood* and $p(\theta)$ is the *prior* distribution.
The likelihood gives the probability density of observing $\mathbf{u}$ given the parameter value θ, the prior contains all the experts information about the parameters while the integral in the denominator is the normalisation constant. In this work, we will consider uniform priors, $p(\theta) \propto 1$ within the case specific bound constraints.
The MCMC methods aim at generating a sequence of random samples $(\theta_1, \theta_2, ..., \theta_N)$, whose distribution asymptotically approaches the posterior distribution as N increases. The samples form a *Markov Chain* as every new point θ_{i+1} depends only on the previous point θ_i. The averages computed from the MCMC samples converge to the correct expected value as the number of samples increases. Let π be the target density. An MCMC algorithm is said to be ergodic if, for a function satisfying sufficient conditions $f : \mathbb{R}^d \to \mathbb{R}$ and initial parameter value θ_0, it holds that

$$\lim_{n\to\infty} \frac{1}{n+1}(f(\theta_0) + f(\theta_1) + \ldots + f(\theta_n)) = \int_{R^d} f(\theta)\pi(\theta)d\theta,$$

where $(\theta_0, ..., \theta_n)$ are the samples produced by the MCMC algorithm. See Brooks et al. (2011) for further details.
The Metropolis algorithm with a Gaussian proposal, with covariance matrix $\mathbf{C}$ and initial point $\theta_{old} = \theta_0$, can be written as follows:

1. Generate a candidate value $\theta_{new} \sim N(\theta_{old}, \mathbf{C})$ and compute the residual sum of

squares $SS(\theta_{new})$.

2. Accept the candidate if $u < \min\left(1, \frac{\pi(\hat{\theta}_{new})}{\pi(\theta_{old})}\right)$ where $u \sim U(0,1)$.

   (a) If accepted, add θ_{new} to the chain and set $\theta_{old} := \theta_{new}$ and $\pi(\theta_{old}) := \pi(\theta_{new})$.

   (b) If rejected, repeat θ_{old} in the chain.

3. Go to step 1 until a desired chain length is achieved.

Note that the normalisation constant $\int l(\mathbf{u}|\theta)p(\theta)d\theta$ is not needed, which drastically reduces the complexity of the problem as the number of model parameter increases. Let us consider the model for the mismatch Eq. (2.3) with $\epsilon \sim N(0, \sigma^2 I)$, the likelihood can be written as:

$$l(\mathbf{u}|\theta) \propto \prod_{i=1}^{n} l(u_i|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[\mathbf{u}_i - F(t_i; \mathbf{u}_0, \theta)]^2\right).$$

where $SS(\theta) = \sum_{i=1}^{n}[\mathbf{u}_i - F(t_i; \mathbf{u}_0, \theta)]^2$ is the sum of squares, and the measurement error variance σ^2 can be estimated using repeated measurements.
The posterior density can be written as

$$\pi(\theta|\mathbf{u}) \propto l(\mathbf{u}|\theta)p(\theta) \propto \exp\left(-\frac{1}{2\sigma^2}SS(\theta)\right).$$

and, therefore, the acceptance ratio reduces to

$$\min\left(1, \frac{\pi(\theta_{new})}{\pi(\theta_{old})}\right) = \min\left(1, \exp\left(-\frac{1}{2\sigma^2}\left(SS(\theta_{new}) - SS(\theta_{old})\right)\right)\right).$$

Adaptive metropolis One of the challenges of the MCMC approach is to find a proper proposal distribution for the given target posterior. The Adaptive metropolis (AM) algorithm aims at adapting the proposal as the chain proceeds to better fit the target density. In this algorithm, the empirical covariance matrix is used after an initial burn in period $n_0 > 0$ in which the covariance is fixed as in the basic version of the MCMC:

$$\mathbf{C}_n = \begin{cases} \mathbf{C}_0, & n \leq n_0 \\ s_d \mathrm{Cov}(\theta_0, \ldots, \theta_{n-1}) + s_d \xi \mathbf{I}_d, & n > n_0. \end{cases}$$

The intuitive idea behind the adaptation is that as the accepted chain values sample the underlying unknown posterior distribution, the covariance of the chain should increasingly well approximate the size and shape of the posterior during the run, and thus provide a proposal with improving mixing. The longer the chain gets, the less the new members of it impact the covariance. Thus, the adaptation diminishes during the run, and asymptotically the AM sampling boils down to usual (MH) sampling. See Haario et al. (2001) for a proof of the ergodicity of the AM algorithm.

Naturally, one should use a good initial parameter value to start the chain as well as a good proposal distribution, if available. The standard way to achieve this is to first perform an optimisation to find a maximum a posteriori (MAP) or maximum likelihood point for the parameter vector, and use it as the starting point. The Jacobian matrix, a derivative-based linearisation of the negative log likelihood function, provides an approximation of the Hessian matrix and yields the usual choice for an initial proposal. The likelihoods discussed in this work turn out to be stochastic, so the standard way is not directly applicable. However, the same overall idea can be implemented using population-based optimisation algorithms, that are able to give both good initial chain values and proposal distributions.

Delayed Rejection Adaptive Metropolis In cases where the target distribution is strongly correlated, a wide initial proposal distribution might lead to a slow adaptation before reaching an optimal mixing of the chain. As a consequence, a long chain is needed to obtain a good estimate of the posterior distribution. A better choice usually is to start from a relatively narrow proposal distribution and let the adaptation find a well-mixing proposal. However, for posteriors with strongly nonlinear correlations, the adaptation may even decrease the acceptance rate: while the overall size of the posterior is correctly found, a single Gaussian covariance is again too wide, resulting in frequent proposals outside the 'thin' posterior. In Haario et al. (2006), the Delayed Rejection Adaptive Metropolis (AM) algorithm is introduced as a solution to this situation, based on the use of several proposal distributions. This algorithm is a combination of the Delayed Rejection (DR) algorithm proposed by Mira (2001) and Adaptive Metropolis. In the delayed rejection, the algorithm is divided into stages. For every chain value, a new candidate is proposed from a proposal distribution. If the candidate is accepted, the chain continues, while if the candidate is rejected, a new candidate is proposed from a different proposal distribution that is usually a scaled version of the previous one. The proposal depends on the rejected candidate, so the acceptance rate has to be rewritten accordingly. If this second candidate is accepted, it is added to the chain and the algorithm continues by proposing a new candidate from the first stage; otherwise, if it is again rejected, this is repeated for a prescribed number of stages, see Haario et al. (2006) for further details.
The proposal distribution at the first stage of the DRAM algorithm is adapted as in AM, the covariance matrix C_n^1 at the sampling step n for the proposal distribution is updated by the sampled chain, no matter from which stage the chain members has been accepted. We only use two proposals, the covariance of the second stage is given by $C_n^2 = C_n^1/\gamma$, with fixed scaling factor $\gamma > 1$. The default value used is $\gamma = 10$.

2.2 Local Approximation MCMC

Despite the precautions introduced with AM and DRAM, in addition to the many other variants present in the literature, the biggest problem for many MCMC methods is that a large number of samples is necessary in order to obtain a good approximation of the posterior distribution.
This fact makes their use complicated or even impractical in cases where the computational cost of a single likelihood is very large. To cope with this situation, we employ

a *surrogate model method* which replaces many of the likelihood evaluations with less expensive approximations without affecting the quality of the approximation of the posterior distribution. This method, first introduced by Conrad et al. (2016), is called *Local approximation MCMC* (LA-MCMC).
The method consists of building local surrogate models of likelihood or negative log likelihood (NLL) during the creation of the chain. These models become more and more precise as sampling continues, especially in areas where the posterior probability has higher values. To make this approach efficient, the main characteristics that these surrogate models must have are that they are computationally much cheaper than the model that they approximate and that they maintain at least locally a behaviour similar to that of the latter (in our specific case the NLL). Given these assumptions, a natural choice is to consider an approximation given by *local polynomials* (in our case quadratic) calculated by regression starting from a set of *full NLL* evaluations present in a neighbourhood of the value of interest.
It was shown in the works of Davis (2018); Davis et al. (2020) that it is possible to obtain a linear decay of the error with respect to the length of the chain by carefully choosing the points where to calculate the full NLL. As the sampling proceeds, more and more precise approximations are obtained, converging asymptotically to the target distribution, at a decidedly more advantageous cost than the classic MCMC methods.
On a practical level, a reduction of the number of evaluations of the full NLL could be found, quantifiable in different orders of magnitude, see Conrad et al. (2016); Davis (2018); Davis et al. (2020). This will also be confirmed in this work where, although the NLL target is not deterministic, a few hundred points were sufficient to obtain chains of two orders of magnitude longer. Below is a brief summary of the LA-MCMC algorithm, leaving Davis et al. (2020) as a reference in case the reader is interested in further details. For each element of the chain:

1. Sample a new candidate using the previous parameter of the chain and the proposal distribution;

2. Calculate the local regression. If the reliability of the regression model is insufficient, re-evaluate the expensive NLL in the neighbourhood of the new candidate to reduce the regression error;

3. Calculate the acceptance probability using the approximation of the NLL obtained from the local polynomial regression of the current candidate and the previous one in the chain;

4. Accept or reject the candidate.

We remind the reader that all the evaluations of the expensive NLL are saved as they are calculated in the so-called *support set* $\{(\theta_i, NLL(\theta_i))\}_{i=1}^{N_{supp}}$ where θ_i are the parameters for which the NLL was calculated. This set is separate from the sampled chain, only the parameters sampled from the surrogate surface are included in the chain. At each step of the algorithm, the parameters of the new candidate's neighbourhood will be selected from

those present within this set. An optimal refinement strategy is essential to find the right balance between the error in the acceptance probability of the point $3)$ and the number of expensive NLL evaluations saved in the support set during the point $2)$. In other words, the strategy must ensure that, at each step of the algorithm, the bias introduced by the surrogate model is not the dominant source of error. The strategy consists of various refinement criteria that are triggered in the presence of an insufficient surrogate model with respect to the thresholds given by indicators that become more and more severe as the sampling proceeds. See Conrad et al. (2016); Davis (2018); Davis et al. (2020) for further details on the specific criteria used.

3 Bayesian inference by informative Gaussian features of the data

As already discussed in Sections 1 and 2, the classical approach to parameter estimation is based on minimising, as θ varies, the sum of square difference between the integrated model realisations and the data in possession at all times t_i.
However, there are scenarios in which this approach is not valid. As an example, let us consider the case of an oscillator formed by two coupled chaotic systems with a specific coupling strength which lead to the intermittent bursts, for more details see Eq. (3.4). In such scenario, the residual signal between the two oscillators may have a low magnitude for most of the time and then has sudden peaks, unpredictable given the chaotic nature of the system. It can be noted that, in this case, the sum of the residuals is not very informative: the same RSS can be obtained with quite different residuals.

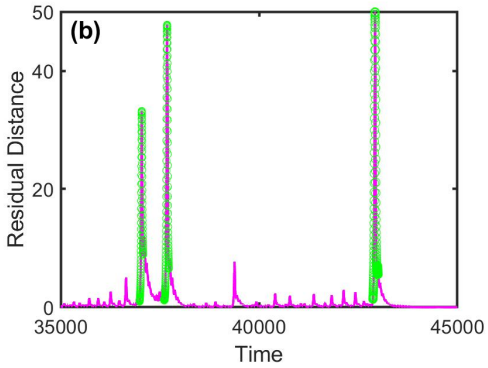


Figure 3.1: Difference between two coupled oscillators. The signal is having low magnitude most of the time with sudden outbursts.

Nevertheless, the residuals still maintain a certain coherence in the distribution of the peaks, as can be found by integrating this model multiple times with fixed parameters and varying the initial state. It, therefore, becomes attractive to consider the *distribution* of the residues to characterise the variability of the model. It will be shown in this work how the parameters of chaotic systems can be correctly estimated by using the empirical cumulative distribution function (eCDF) of the residuals or other scalar valued functions of data, supposing that the eCDF can be empirically estimated from the data. This is the intuition behind the Bayesian inference based on the informative statistics of data.

3.1 Gaussian likelihoods by informative feature vectors

In the emerging era of big data, measurements may be so abundant that subsampling of data is needed before (UQ) analysis is possible, due to memory issues. Another motive for creating likelihoods for subsamples of data only stems from excessive computing times: numerically heavy models may be simulated so that the predictions cover subsets

of the collected data, while covering all the samples of data would be exhaustively expensive. We discuss now a general approach on how likelihoods may be constructed by data subsampling.

Let us suppose that a data set $U = (u_1, ..., u_\tau)$ of vectors u is available. We want to construct a summary statistics for subsets of size $N < \tau$. For this purpose, the set U is divided in n_{epo} subsets each of size N, that is $U = \{u^k\}$ where u^k denotes the sets $u^k = (u_1^k, ..., u_N^k), k = 1, ..., n_{epo}$. The division into the subsets can be, in principle, arbitrary. However, our examples deal with dynamical models where the subsets consist of time-wise consequtive measurements on time intervals called epochs, and n_{epo} denotes the number of them.

Next, suppose we have a scalar-valued mapping $y : U \to R^1$. The mapping may be computed in various ways; some examples will be presented in the following sections. It can be computed by single vectors of each epoch that provides N scalar values for each epoch, or by pairs of vectors between two different epochs, that yields N^2 scalar values. The statistical distribution of these scalars will be used to construct the likelihoods throughout this work. The distribution is expressed as the cumulative distribution function (eCDF) of the scalars given by y.

The distribution of the eCDF vector is estimated by repeating the calculations by y over all the epochs used. The number of available repetitions depends on the mapping y. For example, a construction based on single epochs would produce n_{epo} samples of eCDF feature vectors, while a construction based on pairs of different epochs gives $\binom{n_{\mathrm{epo}}}{2}$ vector samples. It turns out that the distribution of the eCDF vector is in fact asymptotically Gaussian in all the examples we deal with. This is a consequence of generalisations of the central limit theorem (CLT). Thus, the estimation of the distribution of the eCDF vector amounts to calculating the mean and covariance (μ, Σ) of the sample vectors created. More details are given in the following sections.

Opposite to the BSL approach, here the likelihood is constructed offline only once, as a summary statistics of the available data. For the purposes of high-CPU inference, this approach has obvious advantages. When simulating the model for new model parameter values, the computational cost is limited to that needed for integration over one time epoch only. Moreover, we will later show how a surrogate posterior approximation method can be used to drastically reduce the CPU times. However, such an approximation requires a fixed likelihood construction, while in the BSL approach, the likelihood is recreated for each new parameter evaluation.

Another benefit comes from the way how the likelihood is constructed by eCDF vectors. The normality is guaranteed theoretically, and it can – and should – always be numerically verified by normality tests before using the likelihood for parameter estimation.

The approach naturally has limitations as well. It is intended for cases where the data set is large enough to allow the subsampling to be done in a reliable manner. The data should be representative, that is, cover the possible rare events well enough. Later on, we will discuss ways to diagnose the possible pitfalls, whether they are due to insufficient data or a lack-of-fit between the model and data. However, we note that for limited data (and CPU times allowing) you can also resort to the BSL construction, using the feature vectors

suggested here. Using the above notations, we would then have data over one epoch only. To test a new model parameter value youwould simulate n_{epo} times the model over the time interval of data, create the likelihood as discussed here and then accept or reject the parameter value depending on how well the data agrees with the likelihood. While the approach takes n_{epo} times more computing time, it allows inference by limited data sets as well.

Obviously, the central limit theorem directly provides Gaussianity (supposing bounded data), if instead of a scalar mapping y we use a component-wise mean as a summary statistics. However, it turns out that, in many cases, such a 'naive' summary will average out too much information. The challenge then is to find mappings that provide a transformation from the – often statistically intractable – original data to feature vectors that at the same time preserve the information of the original data, and give Gaussian distributions. We will thus call, in general, the approach by the acronym (GLIF), Gaussian likelihood by informative feature vectors. In the next section, a first version of such a likelihood, the Correlation integral likelihood Haario et al. (2015); Springer et al. (2019); Maraia et al. (2020); Kazarnikov and Haario (2020), will be discussed.

3.2 Correlation integral likelihood

Estimating the parameters of chaotic dynamical systems has been a challenging task since their introduction to numerical simulations by Lorenz (1963). Most of the approaches presented in the literature are based on having enough dense measurements within a predictable time interval, or are applicable only in limited special situations, see the discussion in the Introduction. The Correlation integral likelihood (CIL) Haario et al. (2015); Springer et al. (2019); Maraia et al. (2020) is a GLIF type approach that is free of such limitations. It is based on a generalisation of the correlation integral, introduced in the mathematical physics literature to characterise the fractal dimension of an attractor. However, rather than estimating the intrinsic dimension of trajectories, we are interested in characterising the similarities or distances between different trajectories. The scalar mapping y is provided by the Euclidean distance between vectors of trajectories. So, constructing the CIL likelihood requires three steps: (i) computing the distances between observation vectors from pairs of different trajectories in the training data; (ii) evaluating the empirical cumulative distribution function (eCDF) in the log-scale of these distances and (iii) estimating the mean and covariance of the ensuing log-eCDF vectors by repeating steps (i) and (ii) over all different data epochs pairs.

Suppose $\mathbf{S}$ is a data set comprising observations of the dynamical system of interest in a time interval $[0, \tau]$. As above, $\mathbf{S}$ is split into n_{epo} different epochs. The epochs can, in principle, be any subsets from the reference data set $\mathbf{S}$. We restrict the epochs to be time-consecutive intervals of N evenly spaced observations. Denote $\mathbf{s}^k = \{\mathbf{s}_i^k\}_{i=1}^N$ and $\mathbf{s}^l = \{\mathbf{s}_j^l\}_{j=1}^N$, with $1 \leq k, l \leq n_{\text{epo}}$ and $k \neq l$, be two such disjoint epochs. The individual observable vectors $\mathbf{s}_i^k \in \mathbb{R}^d$ and $\mathbf{s}_j^l \in \mathbb{R}^d$ comprising each epoch come from the time intervals $[t_{kN+1}, t_{(k+1)N}]$ and $[t_{lN+1}, t_{(l+1)N}]$, respectively. In other words, superscripts refer to different epochs and subscripts refer to the time points within those epochs. Haario et al. (2015) define the *modified correlation integral sum* $C(R, N, \mathbf{s}^k, \mathbf{s}^l)$ by counting all pairs

of observations that are less than a distance $R > 0$ from each other:

$$C(R, N, \mathbf{s}^k, \mathbf{s}^l) = \frac{1}{N^2} \sum_{i,j \leq N} \mathbb{1}_{[0,R]} \left(\left\| \mathbf{s}_i^k - \mathbf{s}_j^l \right\| \right), \tag{3.1}$$

where $\mathbb{1}$ denotes the indicator function and $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$. In the physics literature, evaluating Eq. (3.1) in the limit $R \to 0$, with $k = l$ and $i \neq j$, numerically approximates the fractal dimension of the attractor that produced $\mathbf{s}^k = \mathbf{s}^l$ (Grassberger and Procaccia, 1983b,a). Here, we instead use Eq. (3.1) to characterise the distribution of distances between $\mathbf{s}^k$ and $\mathbf{s}^l$ at all the relevant scales. We assume that the state space is bounded; therefore, a radius R_0 covering all pairwise distances in Eq. (3.1) exists.

Note that for a prescribed set of radii $R = R_m = R_0 b^{-m}$, with $b > 1$ and $m = 0, \ldots, M$, Eq. (3.1) defines indeed a discretisation of the empirical CDF of the distances $\|\mathbf{s}_i^k - \mathbf{s}_j^l\|$, with histogram bins given by the numbers R_m. So Eq. (3.1) gives an example of the GLIF approach, with the scalar mapping y given by the distances.

Now we define $y_m^{k,l} = C(R_m, N, \mathbf{s}^k, \mathbf{s}^l)$ as components of a *statistic* T, which for any $\mathbf{s}^k, \mathbf{s}^l$ is given by $T(\mathbf{s}^k, \mathbf{s}^l) = \mathbf{y}^{k,l} \coloneqq (y_0^{k,l}, \ldots, y_M^{k,l})$. This statistic is the feature vector used for the CIL approach. It turns out that, under mild conditions on the independence of the samples, the feature vector is in fact Gaussian (see the discussion in the next section). We characterise this normal distribution by subsampling the data set $\mathbf{S}$. Specifically, we approximate the mean μ and covariance Σ of T
by the sample mean and sample covariance of the set $\{\mathbf{y}^{k,l} : 1 \leq k, l \leq n_{\text{epo}}, k \neq l\}$, evaluated for all $\binom{n_{\text{epo}}}{2}$ pairs of epochs $(\mathbf{s}^k, \mathbf{s}^l)$ using fixed values of R_0, b, M, and N.

The Gaussian distribution of T effectively characterises the geometry of the attractor represented in the data set $\mathbf{S}$. Now, we wish to use this distribution to infer the parameters θ. Given a parameter value $\tilde{\theta}$, we use the model to generate states $\mathbf{s}^*(\tilde{\theta}) = \{\mathbf{s}_i^*(\tilde{\theta})\}_{i=1}^N$ for the length of a single epoch. We then evaluate the statistics $y_m^{k,*} = C(R_m, N, \mathbf{s}^k, \mathbf{s}^*(\tilde{\theta}))$ as in Eq. (3.1), by computing the distances between elements of $\mathbf{s}^*(\tilde{\theta})$ and the states of an epoch $\mathbf{s}^k$ selected from the data $\mathbf{S}$. Combining these statistics into a feature vector $\mathbf{y}^{k,*}(\tilde{\theta}) = (y_m^{k,*})_{m=0}^M$, we can write a noisy estimate of the log-likelihood function:

$$\log p(\tilde{\theta} | \mathbf{s}^k) = -\frac{1}{2} \left(\mathbf{y}^{k,*}(\tilde{\theta}) - \mu \right)^\top \Sigma^{-1} \left(\mathbf{y}^{k,*}(\tilde{\theta}) - \mu \right) \tag{3.2}$$

The expression is noisy for two reasons: for a fixed model parameters, we randomise the initial values (and numerical solver options, possibly) and thus obtain different realisations for each model simulation. Moreover, the selection of K introduces additional randomness. Therefore, the feature vectors $\mathbf{y}^{k,*}$ are random for any finite N. A partial reduction of the randomness can be obtained by having more data, i.e. having more epochs in the data set with larger N. This, however, comes with increasing computational cost. Other options to reduce the noise of the likelihood evaluations are discussed in the next section.

The CIL approach interprets the time series of observations vectors of chaotic trajectories as samples from a distribution in the state space, i.e. from the underlying attractor. A natural requirement is that the attractor is assumed to be fixed. In addition, the observations

should give a representative sample from the attractor. In Section 3.5, we discuss cases where this is not satisfied. We note that, since only the state vector values are used, no information of the time instants of observations is needed, and the observations may be arbitrarily far away from each other. On the other hand, the information of the transient behaviour is also lost. In Section 4.1, we discuss ways to get the dynamical information back in the estimation.

3.3 Gaussianity of the feature vectors, one and two sample U-statistics

Here, we discuss in more detail the theoretical background of Gaussianity of the summary statistics used, in the specific non-i.i.d situation in question. Basic results in U-statistics (see e.g. Borovkova et al. (2001); Neumeyer (2004) and references therein) yield Gaussianity of the quantity

$$\sqrt{N}\Big(\frac{2}{N(N-1)}\sum_{1\le i<j\le N} f(X_i, X_j) - A\Big)$$

in the limit $N \to \infty$ (assuming non-degeneracy). Here, $X_1, \ldots, X_N$ are i.i.d. or i.d. weakly dependent random variables (or vectors) and f, say, a bounded function, and $A := \mathbf{E}\, f(X_1, X_2)$. These results generalise the classical result of Donsker (1951), which applies to i.i.d. samples from a scalar-valued distribution. Note that the expression only has N independent samples, so the convergence rate $N^{-1/2}$ is as expected.
This result transfers in a standard way to the Gaussian statistics of the cumulative distribution function. Another variant of this is where you considers instead two-sample U-statistics, with a similar statement for

$$\sqrt{N}\Big(\frac{1}{N^2}\sum_{1\le i,j\le N} f(X_i, Y_j) - A\Big)$$

where $X_1, \ldots, X_N$ is an i.i.d set of variables, and $Y_1, \ldots, Y_N$ is another i.i..d set of variables, possibly with different distribution and independent from the X_j's, see Neumeyer (2004). In Borovkova et al. (2001), the results for one-sample U-statistics were proved assuming only asymptotic independence for the variables X_k. Similar results may naturally be proved for two-sample U-statistics.
Our situation corresponds to the case of two-sample U-statistics. In our case, $X_i = s_i^k, Y_j = s_j^l$ for any fixed $k \neq l$, with $1 \le k, l \le n_{epo}$. The function f is vector valued, with components

$$f(X, Y, m) = 1_{[0,R_m]}(\|X - Y\|_2), \quad m = 0, ..., M.$$

Recall that the mean μ and the covariance Σ of f are calculated by the $\binom{n_{\mathrm{epo}}}{2}$ vectors of the set $\left\{\mathbf{Y}^{k,l}\right\}_{1\le k\neq l\le n_{epo}}$. Again, the convergence is covered by U-statistics, yielding the convergence rate of order $n_{epo}^{-1/2}$.
Given a fixed data set of n_{epo} epochs, various options are available based on selecting more

than two epochs for the creation of the feature vectors. Even if the theoretical convergence rate is limited by available epochs, using several of them can be beneficial, as a larger subset of epochs used to create the feature vectors decreases the numerical variability in the likelihood. Instead of pairs, we can consider the distances between one epoch and several other epochs. Maximally, we can use $(n_{epo}-2)$ others while creating the summary statistic. Leaving one epoch out is necessary to have enough , $\binom{n_{\mathrm{epo}}}{n_{\mathrm{epo}}-2} = \binom{n_{\mathrm{epo}}}{2}$ possible combinations from which to obtain a good estimate of the feature vector variability. This 'one to many' modification of correlation integral sum can be written as

$$C(R, N, \mathbf{s}^k, S \setminus \{\mathbf{s}^q\}) = \frac{1}{(n_{epo}-2)N^2} \sum_{l \leq N, l \neq k, q} \sum_{i,j \leq N} \mathbb{1}_{[0,R]} \left(\left\| \mathbf{s}_i^k - \mathbf{s}_j^l \right\| \right), \tag{3.3}$$

The increase in the computational cost due to distance computations is negligible in cases where the model integration dominates the time needed in the likelihood evaluation. In the sampling phase, each new trajectory integrated using a candidate parameter is compared now to random $n_{epo}-2$ epochs from the training set to obtain the summary statistics as in Eq. (3.3). Note that while the noise is reduced by taking the average over several epochs of the training set, the gain is limited: the noise due to the evaluation of the new sample (corresponding to $\mathbf{s}^k$ in Eq. (3.3)) remains. Naturally, this source of noise can be limited by creating n_{epo} independent simulation samples for each proposed new parameter value by, e.g. parallel calculations. Anyway, this option would increase the computational cost, and is not considered in this work.

3.4 Selection of informative features for specific tasks

Now that the Correlation integral likelihood has been extensively discussed, let us return to other possibilities of using the GLIF approach. The distance-based CIL is by no means the only alternative to employ the key underlying idea, the normality of the eCDF vectors. Indeed, you have to carefully select the feature – the mapping y – that best fits to the specific problem. In addition to various options for the feature vector, we discuss ways to combine different statistics into a Gaussian likelihood. Ways to avoid introducing bias in the likelihood creation are discussed in the next section, together with ways to diagnose the possible pitfalls.

Selecting an appropriate statistic that is informative with respect to the data and model is critical to make this approach work best. It is, therefore, important to have a good level of prior knowledge of the system being treated, as by selecting a certain statistic a strong assumption is made that will affect the final result. We use here several examples to demonstrate this argument, but leave the more detailed numerical examples to be presented in section 4.

Time dependent information The system (4.3) in Section 4 gives an example where the basic form of the CIL partially fails. Here, the classical Lorenz 3D system is equipped with an additional parameter K that scales the right-hand side of the equation. While the standard parameters σ, ρ, β determine the shape of the chaotic attractor, the portion of the state space occupied by the computed trajectories, the fourth parameter only affects the

average speed with which the system evolves. If we produce the posterior distribution of all four parameters (for the details see section 4), what happens is that the first three parameters are correctly identified, while the fourth remains unidentified. What is the reason for this partial defeat? The answer can be found by analysing the steps by which the CIL statistic is created. As previously described, the procedure involves the calculation of the distances between two data epochs, and the subsequent mapping of these distances in a log-eCDF. During this last step, the temporal information of the measurements is completely ignored, resulting in the loss of information regarding the speed of evolution of the system. While the remaining information is enough in many situations, and even beneficial in cases where the measurement time instants are poorly known, it here leads to the failure to identify the parameter K. A remedy is to extend the state s by the derivative of it, ds/dt, and to create a feature vector for it by (3.1), with ds/dt instead of s. To do this by measured data, we have to assume that the measurements contain pairs of measurements, where the temporal difference is small to allow an approximation of the derivative values. Otherwise, the time lag separating such pairs can again be arbitrarily high. This second feature vector, however, does not contain sufficient information on the geometry of the chaotic attractor, resulting in a lack of identification of the first three parameters of the model. The solution is to consider both features together, in a way that will be discussed below.

Back to basics: feature vectors based on residuals, means and standard deviations Computing the distances between observations at different time points is not always needed. Obviously, no mapping y is needed in cases where the system already is scalar-valued. The classical Ornstein-Uhlenbeck system provides an example of this, same as any one-dimensional stochastic differential equation SDE systems. For such systems, we can directly form the eCDF vectors of the state values by data, as well as the stochastic derivative values, in case pairwise measurements close enough in time are available. The GLIF lilkelihood can then be estimated by repeated evaluations of the eCDF vectors, and numerically verified to be Gaussian by the χ^2 test again.

An example where the calculation of distances would be downright counterproductive is given by swarming. The Kuramoto equation, Kuramoto and Araki (1975); Kuramoto (1984) describes a well-known system of oscillators rotating on the unit circle, each with its own frequency. With increasing the value of a coupling constant, however, the rotation speeds converge to a common value, and the oscillators pack together in a close swarm. The right features that characterises the closeness of the swarm are then simply the mean and std, the standard deviation, of the x, y coordinates of the swarm members (the related concepts of order and potential are used in the Kuramoto literature). Again, we can form eCDF vectors of both the mean and std values, and arrive at a Gaussian likelihood. Obviously, calculating distances between the swarm members at different time instants does not produce any meaningful information on the relative closeness of the members of the pack.

Synchronisation of two chaotic systems provides another example where the distances between the systems at the same time instants, i.e. the usual residual norms, characterise the degree of synchronizations in a natural way. However, in cases of weak or emerging synchronization the geometry of the joint attractor changes with the coupling strength,

so the CIL likelihood can be expected to improve the estimation accuracy. Again, the combination of two feature vectors should be employed.

How to combine different feature vectors As indicated by the above examples, selecting a single feature vector for a particular data set may be too limiting, and may lead to only partial and incomplete results. Using more than one statistics often leads to more precise results. If the selected statistics are independent, you can compute two or more likelihoods separately and multiply them to create the joint likelihood. Given that different feature vectors most often are not independent, we propose the following. *Concatenate the samples of two or more feature vectors calculated for the same epochs, and verify numerically that the combined feature vector sample is Gaussian.*
The concatenated statistics may consist of different types of measurements of the system under consideration. The important thing is to always analyse the problem studied and the data available to select informative summary statistics. In the section for numerical experiments, we will present several scenarios in order to demonstrate how to use this method in the best possible way.

3.5 Diagnostics: the likelihood and goodness of fit

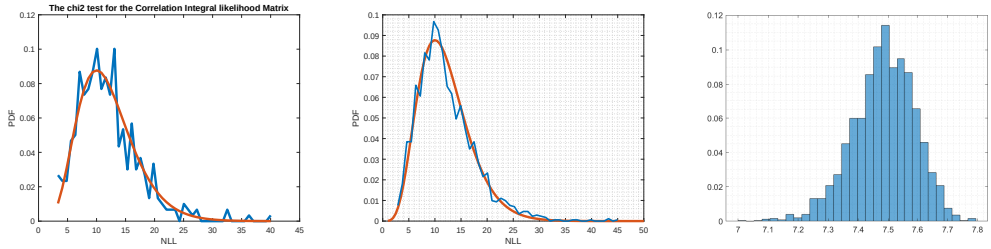
As discussed in Section 3.1, the construction of the eCDF feature vector likelihood is based on n_{epo} data sets, each consisting of N vectors. Two numerical approximations are performed based on these numbers, the computation of individual eCDF vectors that depends on N, and estimation μ and Σ, the mean and covariance, by repetitions of the eCDF vectors, that depends on n_{epo}. Obviously, both approximations need large enough sample sizes to be numerically stable. If the feature vector mapping y is computed from combinations of k epochs, we get $\binom{n_{\text{epo}}}{k}$ different vectors for the estimation of μ and Σ (even if they mey not be independent, see Sec. 3.3). In the case of CIL, we take either $k = 2$ or $k = n_{epo} - 2$ which gives N^2 or $(n_{epo} - 2)N^2$ scalar values for the single eCDF vector estimations. Typically, selections of the order $n_{epo} = 40, ..., 70$ and $N = 1000, ..., 2000$ lead to stable results. In this section, we demonstrate well-working cases as well as the pitfalls that appear if either of the numbers is too small. Moreover, we show how the χ^2 test can be used to reveal a lack-of-fit between the training data and model simulations.
As the demonstration example, we use the Lorenz system, but extended to represent two identical chaotic oscillators unidirectionally coupled. The equations governing the temporal evolution of this system are given as

$$\begin{cases} \dot{x_1} = \sigma(y_1 - x_1); \\ \dot{y_1} = \rho x_1 - y_1 - x_1 z_1; \\ \dot{z_1} = x_1 y_1 - \beta z_1; \\ \\ \dot{x_2} = \sigma(y_2 - x_2) + K_c(x_1 - x_2); \\ \dot{y_2} = \rho x_2 - y_2 - x_2 z_2; \\ \dot{z_2} = x_2 y_2 - \beta z_2. \end{cases} \tag{3.4}$$

The reference parameters are the standard ones, $(\sigma = 10, \rho = 28, \beta = 8/3)$. A large enough coupling parameter, roughly $K_c > 8$, leads to complete synchronisation, i.e. starting from different initial values the two subsystems coincide after a long enough time interval. But for a somewhat weaker coupling parameter $K_c = 7.5$ synchronisation remains incomplete, the trajectories of the 3D subsystems mostly stay close, but experience intermittent outbursts at irregular times. Figure 3.1 shows examples of norms of the residuals between the two subsystems. Indeed, as the feature of interest here is how close the two 3D trajectories are at given time points, a natural feature mapping y in this case is the internal distance between the subsystems given by (3.4), rather than distances between different 6D trajectories of (3.4) as used in the CIL case. But this means that $k = 1$ in the subsampling of the $\binom{n_{\text{epo}}}{k}$ epochs, so that only n_{epo} repetitions of the eCDF vectors are available, each giving N values to calculate each eCDF.
First, we present an ideal case, with sufficiently high values for N and n_{epo} and no bias between the data and model. We set $n_{epo} = 400$ and $N = 2000$. It is important that the measurements within an epoch are temporally consecutive points, not randomly selected from a larger set of measurements as the risk of having epochs without the irregular spikes would then increase. In this example, we take the time points as $t_{i+1} - t_i = 1$ for all i. By carrying out steps by the guidelines given in Sec. 3.1, a set of 400 eCDF vectors is obtained, each computed by 2000 distance values between the 3D subsystems. The mean μ and the covariance Σ of the eCDFs' are substituted in the formula Eq. (3.2) so as to obtain the cost function. The normality of the obtained feature vectors is tested numerically, and the results are given in Fig. 3.2a. The picture presents the histogram of the values obtained by data, and the χ^2 whose degrees of freedom is given by the dimension (number of bins) of the eCDF vectors. We see that the test values perfectly follow the theoretical χ^2 density function. For simplicity, we only sample here the coupling parameter K_c keeping the rest of the model parameters fixed in their reference values. The posterior distribution estimate of K_c shown in Fig.3.2c has a peak around the reference parameter value 7.5 with relative error of approximately 10%.

Goodness of fit As a way to verify the goodness of fit, i.e. how well the variability of the data set is represented by the model variability with the estimated parameter, it is advisable to conduct the following study. Given μ and Σ from the training set and the parameter posterior distribution, we evaluate the cost function Eq. 3.2 for feature vectors obtained by integrating the model with the MAP of the distribution a couple of thousand times, and compare its distribution with that of the data – or the reference χ^2 density function , as the latter two should be known to agree. Here, the NLL has been re-evaluated 2000 times using the MAP as a reference value. The NLL values obtained form a distribution which practically matches the reference χ^2 distribution, as can be seen in Fig. 3.2b. This indicates, first of all, that the number of epochs as well as the number of measurements per epoch present in the training set cover the variability of the model well, and that this latter is the correct model to be considered for the dynamics that created the data (because we simulate a situation in which data is given to us and it is up to us to find the best model that represents it). Note that the MAP of the parameter distribution in this case practically coincides with the parameter $K_c = 7.5$ with which the synthetic data

were created, so the result is as expected. We also point out that simulating the model with K_c values at the tails or outside of the distribution yields cost function values at the tails or outside of the χ^2 distribution, respectively.



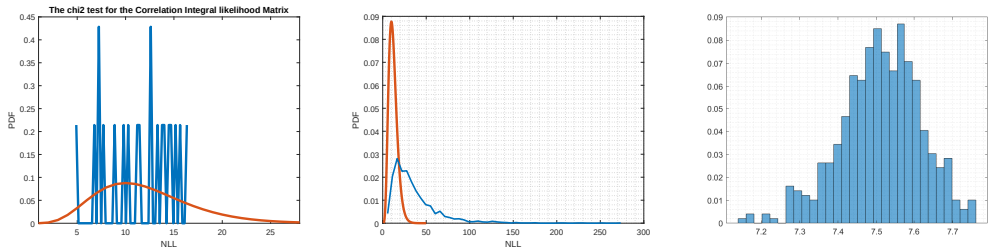
(a) (b) (c)

Figure 3.2: (a) Comparison between the NLL of the feature vectors obtained from the training set and the respective χ^2 distribution. (b) Comparison between the NLL of the feature vectors obtained by reintegrating the system using the MAP as reference parameter and the respective χ^2 distribution. (c) Posterior distribution.

Next, let us discuss the impacts of having either N or n_{epo} too small as well as how the above goodness-of-fit test can reveal different discrepancies between the data and model simulations.

Insufficient number of epochs We repeat a situation similar to the previous ideal case, with the difference that we only create $n_{epo} = 20$ epochs of $N = 2000$ measurements each. The likelihood can be constructed as such. However, the normality test of the feature vectors obtained from the training set does not produce a proper result due to the limited number of repetitions, as reported in Fig. 3.3a. However, we can test how the MCMC sampling performs. The posterior distribution is shown in Fig. 3.3c. Somewhat surprisingly, the reference parameter is again around the MAP of the posterior distribution and the relative error is approximately 10%. The main difference with respect to the ideal case is a drop in the acceptance ratio in sampling the parameter chain, from 26% to around 5%. This may be attributed to the crude estimate of the mean μ and covariance Σ of the feature vectors. Intuitively, the training set does not contain a representative sample of the dynamics, so even parameters close to the correct value may produce trajectories missing in the training set, thereby leading to rejection. This can be numerically verified by using the goodness-of-fit test again, by creating new trajectories at the MAP value, whose feature vectors are inserted in the cost function to obtain the variability of the NLL. Figure 3.3b shows the discrepancy between the variability of the thus resulting histogram and the ideal one given by the χ^2 distribution. In particular, the distribution of histogram is shifted towards and outside the tail of the χ^2 distribution, while some values are found within the variability of the data. To summarise, the model dynamics contains more variability than that present in the data. In particular, it may happen that new integrated trajectories have more or less outbursts of different magnitude than those

few epochs in the training set. In this case the goodness-of-fit test actually shows a lack of data, rather than any bias between data and model.



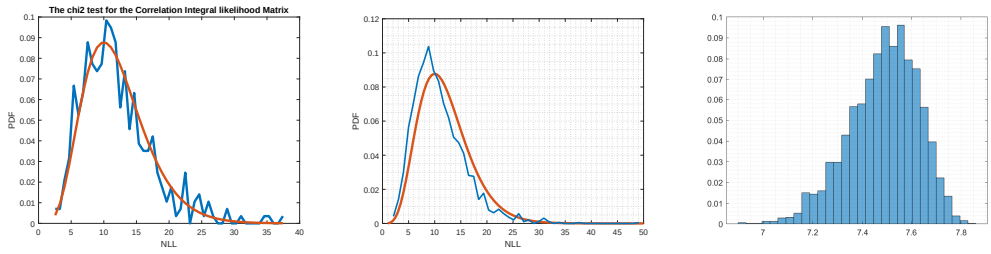
(a) (b) (c)

Figure 3.3: (a) Comparison between the NLL of the feature vectors obtained from the training set and the respective χ^2 distribution. (b) Comparison between the NLL of the feature vectors obtained by reintegrating the system using the MAP as reference parameter and the respective χ^2 distribution. (c) Posterior distribution.

Next, we will analyse situations with different types of bias between the model and data, and how they can be diagnosed.

Complex data, simple model, case 1 We first discuss the situation where data contains more rich dynamics than the model is able to produce. Consider a setup otherwise the same as in the ideal one, but with the difference that the training data consists of epochs created with varying 'hyperparameter' coupling values, $K_c \sim N(7.5, 0.1)$. However, the data is sampled with a model containing a single K_c. Figure 3.4a shows again how the feature vectors obtained from the training set follow a Gaussian distribution. The posterior distribution is given in Fig. 3.4c, with MAP close to the reference parameter and relative error of 13%. A slight increase in the relative error is expected given the higher variability in the training set.
The goodness of fit test, the distribution of the NLL associated to feature vectors obtained from trajectories integrated using the MAP as a reference parameter is shown in 3.4b. As we can see, the values are shifted to the left side of the χ^2 distribution. This indicates a higher variability in the training set feature vectors than what could be produced by reintegrating the model repeatedly with a fixed parameter K_c. This could be even seen by plotting the feature vectors of the training data against those obtained with the model: the first one contains more variability than the one obtained for the verification, with all the trajectories coming with a fixed model parameter.

Complex data, simple model, case 2 We create a biased situation by using different values for the coupling constant to create the training data, and to fit a model to it. Here, the training data is obtained with $K_c = 7.5$, so it consists of coupled 3D trajectories that are close to synchronised most of the time, but 'escape' from each other for short periods at irregularly distributed time points. But we model the data with a model that is incapable

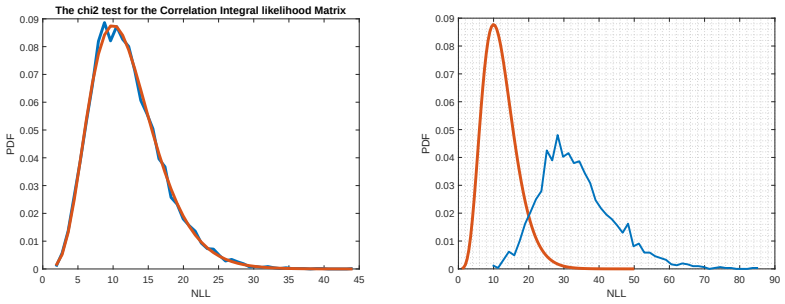


(a) (b) (c)

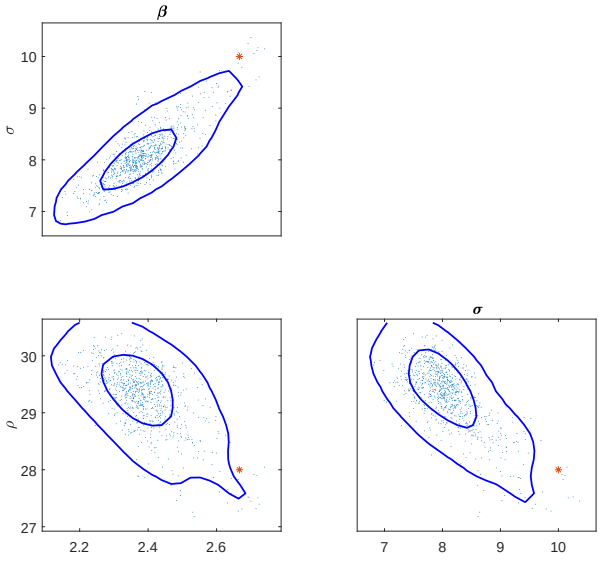
Figure 3.4: (a) Comparison between the NLL of the feature vectors obtained from the training set and the respective χ^2 distribution. (b) Comparison between the NLL of the feature vectors obtained by reintegrating the system using the MAP as reference parameter and the respective χ^2 distribution. (c) Posterior distribution. Red stars give the correct parameter values

of producing such intermittent outburst, using the value $K_c = 9$. Now, we try to estimate the parameters (β, σ, ρ) as the K_c parameter is kept fixed. As the likelihood, we now use the CIL construction discussed in Section 3.2. Indeed, while K_c directly impacts the residuals between the 3D subsystems, the parameters (σ, ρ, β) impact the geometry of the underlying attractor, and the CIL option gives much smaller parameter posteriors.
The results are given in the figures 3.5. While the χ^2 test for data is perfect and the parameter posterior distributions seem fine, we can see that the true parameter values lie outside (or at the extreme tail) of the posterior distribution. The goodness-of-fit test clearly verifies the situation. It would reveal the bias, even if the true parameter value would not be known.

Simple data, complex model For completeness, we also report the posterior that is obtained in the opposite situation.
The data here consists of a 6-dimensional trajectory and the study has been conducted using the CIL cost function. In the training set, we keep the K_c parameter fixed to 9. In this settings, no 'rare events' are created. In the sampling phase, we use $K_c = 7.5$, so the model creates more complex dynamics, including those intermittent outbursts. As can be seen in Fig. 3.6, the model parameters get wrongly identified, and the true reference parameter lies clearly outside the sampled posterior distribution. However the χ^2 test for data and the parameter posterior distributions seem fine again. Based on them alone, you could easily assume that the correct posterior has been obtained. The goodness-of-fit test reveals the discrepancy, however. The distributions based on data (or the overlapping χ^2 distribution) and simulations intersect only at their tails. Furthemore, in this type of scenarios, the acceptance rate of the MCMC sampling gets very low.

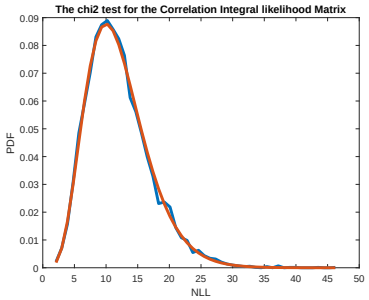


(a) (b)

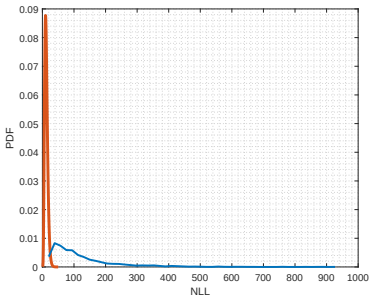


(c)

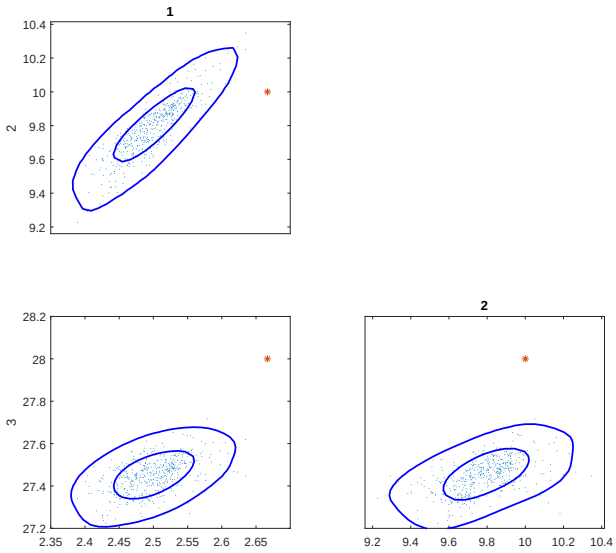
Figure 3.5: (a) Comparison between the NLL of the feature vectors obtained from the training set and the respective χ^2 distribution. (b) Comparison between the NLL of the feature vectors obtained by reintegrating the system using the MAP as reference parameter and the respective χ^2 distribution. (c) Posterior distribution. Red stars give the correct parameter values



(a)



(b)



(c)

Figure 3.6: (a) Comparison between the NLL of the feature vectors obtained from the training set and the respective χ^2 distribution. (b) Comparison between the NLL of the feature vectors obtained by reintegrating the system using the MAP as reference parameter and the respective χ^2 distribution. (c) Posterior distribution.

4 Experimental studies

Parameter estimation So far, we have discussed the properties of several GLIF type likelihood constructions assuming that the 'right' parameter value is already known. Naturally this is not the situation in most modelling tasks. Rather, the model parameters have to be estimated by data, starting from some initial guess that might be far away from the final estimate. The usual procedure with deterministic models with known measurement noise statistics is to find the maximum likelihood estimate by standard optimisers. A linearisation–based approximation of the parameter covariance matrix can then be used get an initial proposal distribution to start the MCMC sampling in order to get the whole parameter posterior distribution determined. In our situation the same approach can be used, but using optimisers that fit the special type of stochastic likelihoods at hand. The method used in this work is the Differential Evolution (DE). It constructs a population of parameter vectors, evaluates the cost function at each of them, and then created the next generation parameter population by a 'survival of fittest' heuristics. For more details, see the version introduced in Shemyakin and Haario (2018), that especially takes into account the stochasticity of the cost function. A special characteristics in our situation is that the cost function is normalised to be the χ^2 distribution with a given degree of freedom: the stopping criteria for the DE-optimiser can be set to be the step when all the members of the population have cost function values roughly inside the numerical range of the χ^2 distribution. An important side product of using th DE optimiser is that the parameter samples of the last few population generations are close to the final estimates and thus can be used calculate an excellent initial proposal covariance for the ensuing MCMC run. See Figure 4.3b for an example of this issue.

Construction of the eCDF vectors The GLIF approach is based on calculating repeated examples of the eCDF vectors of the scalar values provided by a mapping y. Even if the construction of empirical CDF curves by given scalar data is a most basic task in statistics, a few remarks of good practices should be kept in mind. Most of the examples of this section use the CIL approach, so we focus the discussion on it. Similar comments are valid for other GLIF versions, however.

We suppose that the considered trajectories remain bounded, so all the distances between the vectors are inside a ball of finite radius. The range of values of the bins $R_m,\ m = 0, \ldots, M$ is selected by the distances of the training set, keeping in mind that a positive variance is needed for every bin to avoid a singular covariance matrix. Therefore, the largest radius R_0 can be obtained by

$$R_0 = \min_{k \neq l} \left\{ \max_{i,j} \left\| s_i^k - s_j^l \right\| \right\} \tag{4.1}$$

over the disjoint subsets of the samples $\mathbf{s}^k$ and $\mathbf{s}^l$ of length N, $k, l = 1, ..., n_{epo}$. As always with histograms, the number of bins M has to be selected first. A too small value gives a crude histogram, but a too large M a 'peaky' one. In our case, too large M increases the stochasticity of the likelihood resulting in MCMC sampling with poor acceptance.

Therefore, some hand-tuning might be required for M, while the selection of the other parameters then follow in a rather straightforward way. A typical choice of M in our examples has been in the range from 10 to 30. The smallest radius is selected by requiring that for all the possible pairs $(\mathbf{s}^k, \mathbf{s}^l)$, it holds that $\mathcal{B}_{R_M}(\mathbf{s}_i^k) \cap \mathbf{s}^l \neq \emptyset$, where $\mathcal{B}_{R_M}(\mathbf{s}_i^k)$ is the ball of radius R_M centred at $\mathbf{s}_i^k$. That is,

$$R_M = \max_{k \neq l} \left\{ \min_{i,j} \left\| s_i^k - s_j^l \right\| \right\} \tag{4.2}$$

For the log–log feature vectors, the base value b is finally obtained by $R_M = R_0 b^{-M}$ and via that all the other radii R_m ar determined as well. Here, we assume that the measurements are representative, that is each epoch should cover all the corners of the chaotic attractor. See the discussion in section 3.5 for possible pitfalls and for diagnostics in this respect.
As discussed earlier, a precaution to take into account in computing the distances is the order of magnitude of the different components of the state. In case they are considerably different, we suggest scaling the states before computing the distances. Otherwise, there is a risk of losing valuable information regarding the variability of the components with lower magnitude. Recall also that the measurements of the dynamic system must be semi-independent, i.e. the autocorrelation between consecutive measurements must drop quickly enough. Failing this condition leads to the loss of the Gaussianity of the feature vectors, which can be verified by the χ^2 test of the training set again, before on embarking further calculations.

4.1 Low dimensional chaotic dynamical systems

In this section, dedicated entirely to numerical experiments, we will analyse different scenarios that will help to explore the potential of this approach.

Lorenz 63 The most classic of the example of a chaotic system is certainly the one introduced by the E.Lorenz in the early 1960s as an extreme simplification of a meteorological model. The system of equations governing its time evolution is given by

$$\begin{cases} \dot{X} = K(\beta \cdot (Y - X)) \\ \dot{Y} = K(X \cdot (\gamma - Z) - Y) \\ \dot{Z} = K(X \cdot Y - \alpha \cdot Z). \end{cases} \tag{4.3}$$

Its most standard version is the one with the parameters $(\alpha = 8/3, \beta = 10, \gamma = 28)$ which form the famous chaotic attractor in the shape of a 'butterfly'.
As seen in the Figure 4.1, the trajectories of the same model integrated by arbitrarily close initial states separate after a predictable time interval. This effect is commonly described as the 'butterfly effect'. However, in the state space, the thus obtained trajectories remain within the region defining the chaotic attractor of the system.
For demonstration purposes, we have modified the system by multiplying each time derivative in (4.3) by a positive factor K. This simply scales the time variable, so it

changes the speed by which the trajectory evolves along the attractor, but not the attractor itself. In the following examples, we will consider the standard parameters together $K = 1$ and $K = 10$.
Let us suppose a set of measurements of the system at time points$t_0, t_1, ..., t_\tau$, sufficient to obtain a training set formed by $n_{epo} = 64$ epochs of $N = 2000$ measurements each. For demonstration purposes we will consider that the difference between observation points is greater than the predictable time interval of the system, remembering, however, that this condition is not necessary for the method to work. We estimate the parameters of the model in different cases using the CIL method.

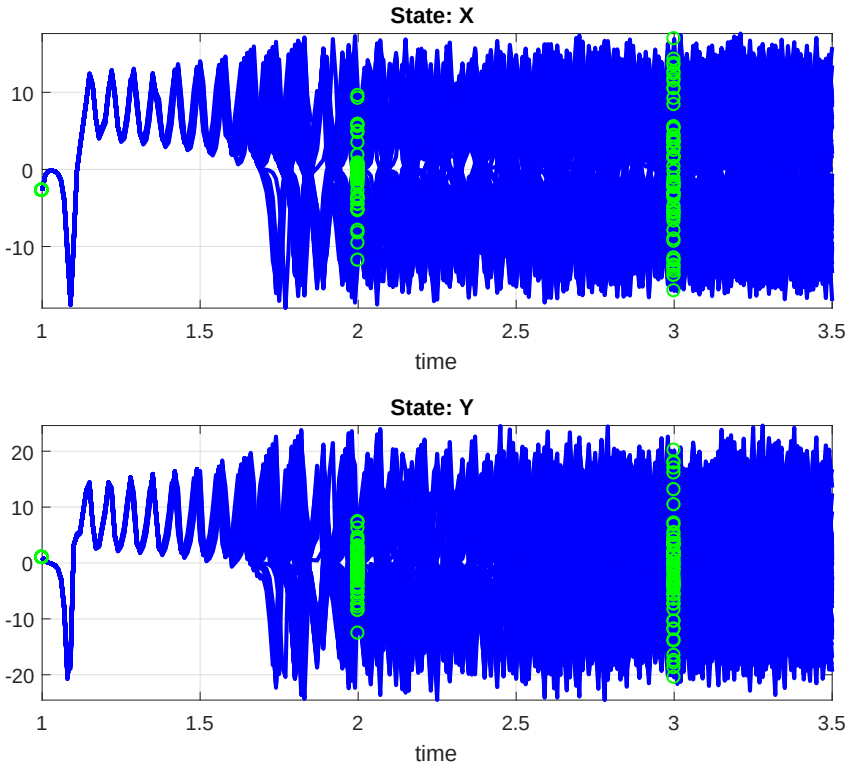
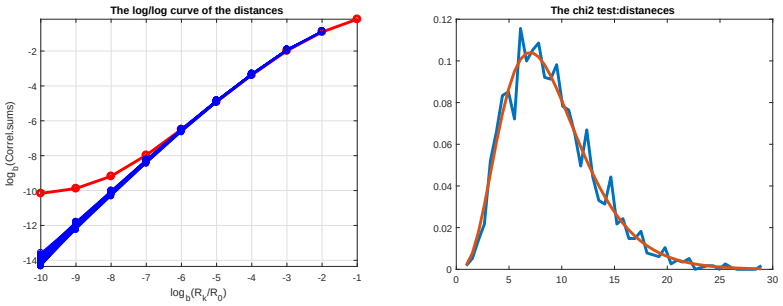


Figure 4.1: Example of the possible measurements of the states (X, Y) done beyond the deterministic intervals for the Lorenz 63 chaotic attractor for 64 simulations with slightly randomised initial conditions.

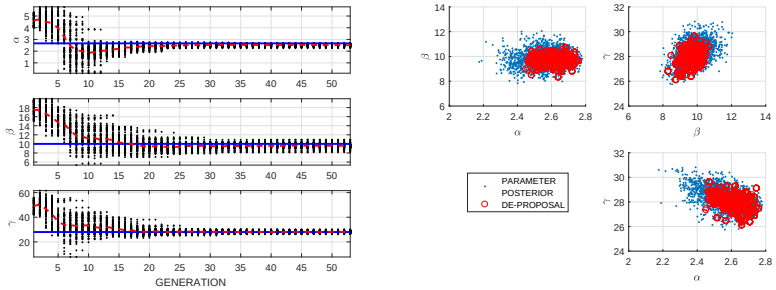
Construction of the likelihood Once the number of radii $M = 10$ has been fixed, the other constants of the method can be obtained as described above, giving $R_0 = 2.51$, $R_M = 0.002$ and $b = 2.04$. The set of feature vectors $\left\{y^{k,l} | k, l \leq n_{epo}\right\}$ is shown in Figure 4.2a, while in Figure 4.2b we find the verification that the assumption of normality

of feature vectors is valid. Given the chaotic nature of the system, a different value follows each new likelihood evaluation.



(a) (b)

Figure 4.2: Lorenz 63: (a) Feature vectors in log-log scale of the measurements. (b) χ^2-test of the log eCDFs for the vectors shown in Figure 4.2a.



(a) (b)

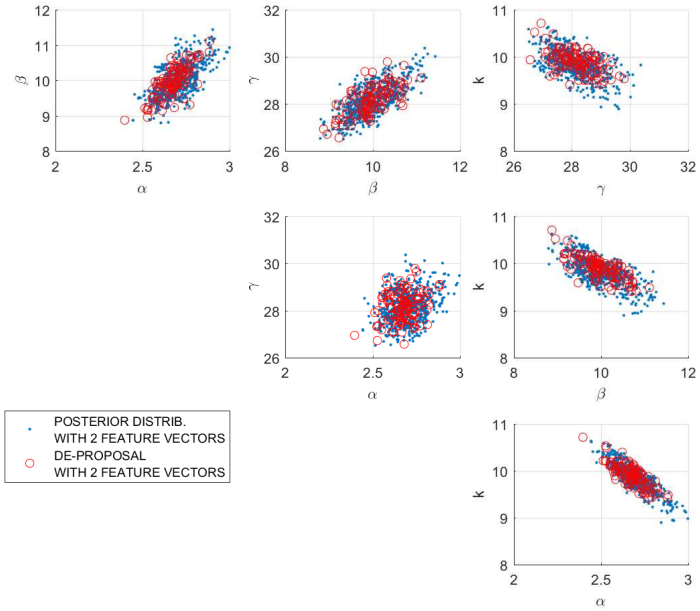
Figure 4.3: Lorenz 63: (a) DE population convergence from a poor initial guess to a narrow region around the right parameter values. (b) Marginal posterior distributions and their respective DE proposals.

Parameter estimation, $K = 1$, one feature vector We demonstrate here the estimation of the model parameters, using the constructed negative likelihood function Eq. (3.2) as the cost function, and starting with initial guesses far away from the true parameter values. Since the cost function is stochastic, it is necessary to use a derivative free optimiser, such as the slightly modified DE algorithm. Figure 4.3a shows the convergence of the algorithm, with the y-axis giving the population of values of each parameter together with the means, and the x-axis the steps of the algorithm, the consequtive 'generations'. The initial parameter values are randomly generated between the given lower and upper bounds, so that the true values are outside those intervals. The size of populations is 45 members. As can be seen, the populations converge close to the true values but then keep

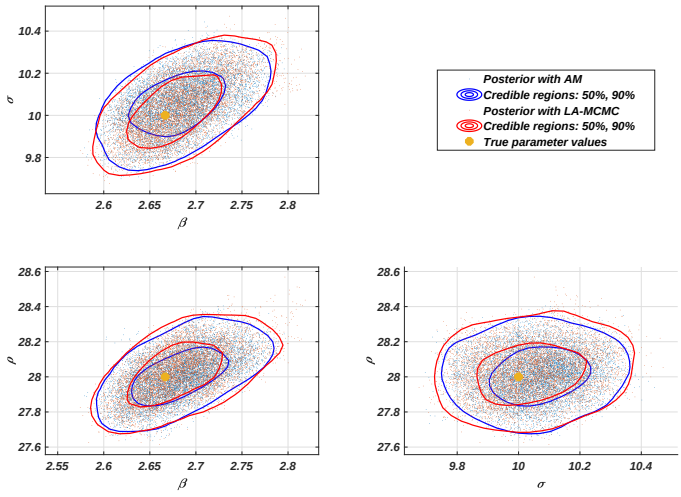
varying in a stationary manner. This is due to the stochastic character of the cost function, and the modification, a recalculation step, introduced in Shemyakin and Haario (2018). The outcome of the optimisation is next used to initialise the MCMC posterior sampling. The mean of the last - or a few last - generations can be used to start the sampling, and the Gaussian initial proposal distribution is given by the mean and the covariance matrix of a few last generation values. The posterior distribution is obtained using the AM algorithm. The two dimensional (2D) marginals are shown in Figure 4.3b. In addition, the Figure shows the parameter values generated by the DE optimiser for the initial covariance. In this case, those values indeed provide a remarkably good approximation for the final posterior distribution.

Two feature vectors, $K = 10$ While the standard three parameters are well identified by the likelihood presented, it is clear that the parameter K would remain undetermined, as the information regarding the time instants of the measurements is not used at all. In order to also correctly identify this parameter, it is necessary to introduce a different feature vector.
Suppose that, in addition to the measurements described above, we add measurements instantly after them, i.e. take double measurements $t_0, t_0 + \delta t, t_1, t_1 + \delta t, ..., t_\tau, t_{\tau+\delta t}$ with $\delta t << 1$. Given an appropriate choice of δt that takes into account the possible measurement error, the new data enables you to approximate the derivative values at the points t_i. Therefore, the dynamics of the system can also be estimated. For each pair of epochs, the CIL feature vectors are computed separately for the states and for the approximations of the derivatives. Once obtained, these two statistics are concatenated to obtain the final feature vector.
As in the first example, the DE-algorithm is used to obtain the MAP and a good approximation of the initial proposal distribution. From the posterior distribution obtained by the AM algorithm, it is clear that all four parameters are correctly identified by the method, see Fig. 4.4a. Again, the last generations of the DE samples already give a good approximation of the final posterior. A comparison of the posteriors in Figures 4.3 and 4.4 shows that the estimation accuracy of the standard three parameters is also improved by considering two feature vectors, although the gain is rather moderate. This is case-dependent, however. In other examples, the improvement is quite dramatic.

Sampling with the LA-MCMC method A known difficulty of using MCMC methods to provide parameters posteriors of a model is that the model must be integrated for each likelihood evaluation, and that chains of several thousand elements are needed to obtain a good estimate of the posterior distribution. For models whose computational cost is very, high this becomes an almost insurmountable problem. The LA-MCMC methods described in Sec. 2 can give a decisive drop of the CPU demand. Let us repeat the parameter estimate in the case of Lorenz using the LA-MCMC method, considering only the first three parameters and a single feature vector. To get an idea of the computational savings achieved with LA-MCMC, several MCMC chains of length 100000 were generated. The cumulative number of forward model evaluations required by the two algorithms is shown in Fig. 4.5. For the selected chain length of 100000, the total number of full likelihood



(a)



(b)

Figure 4.4: Lorenz 63: (a) In blue the 2D marginals of the chain obtained by both feature vectors. In red, the last 25 generations of the parameter population are coming from the DE-optimiser. (b) The 2D marginal of the posterior distributions of the Lorenz 63 model obtained with AM (ewd) and LA-MCMC (blue).

evaluation for the LA-MCMC method were between 955 and 1016. A drop of two orders of magnitude has been achieved with no significant difference between sampled posterior distributions, see the 2D marginals obtained by AM and those obtained with LA-MCMC in Figure 4.4.

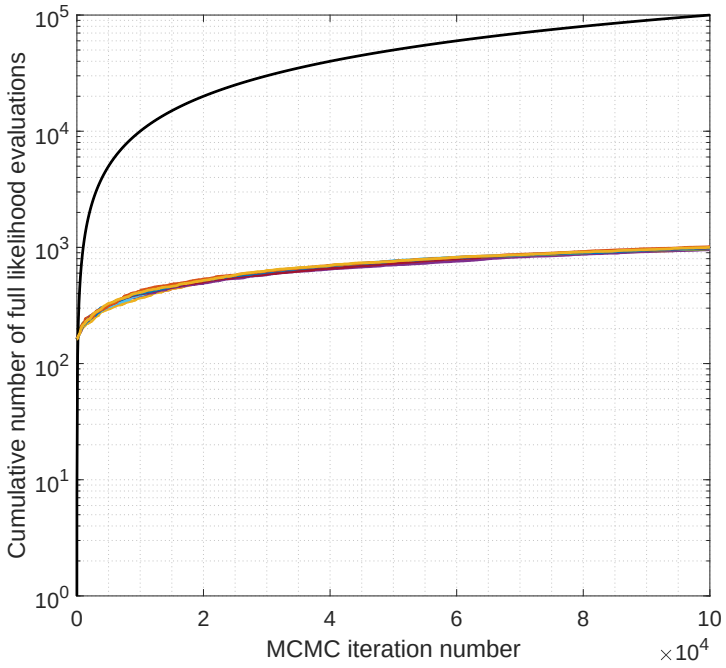


Figure 4.5: Comparison of the cumulative number of full likelihood evaluations while using AM (black line) and LA-MCMC (coloured lines). Every coloured line corresponds to a different chain obtained with LA-MCMC by using the same likelihood.

Before considering cases with much higher dimensional states, we considered it appropriate to verify the robustness of the method on different three-dimensional chaotic attractors. It has emerged that in all the cases considered (that have a stable attractor, see the 'pathological' example of Bouali below), the method has been able to correctly identify the parameters of the model. In addition, the use of the second feature vector for the derivative always decreased the uncertainty of the parameters of the models. The amount of improvement depends on to what extent some parameters affect the speed with which the trajectory travels through the underlying attractor. Here, we report two examples that we considered particularly interesting, namely the attractor introduced in Wang et al. (2009) and the Chua7 attractor, see Zhong et al. (2002). For many more examples, see Springer et al. (2019).

Wang The equations governing the evolution over time of Wang's system are given by

$$\begin{cases} \dot{X} = \alpha \cdot X + \gamma \cdot Y \cdot Z \\ \dot{Y} = \beta \cdot X + \delta \cdot Y - X \cdot Z \\ \dot{Z} = \epsilon \cdot Z + \zeta \cdot X \cdot Y, \end{cases} \tag{4.4}$$

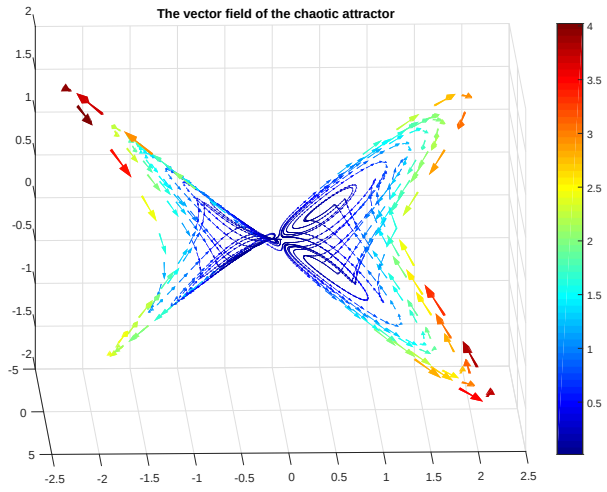
with reference parameters $(\alpha, \beta, \gamma, \delta, \epsilon, \zeta) = (0.2, -0.01, 1, -0.4, -1, -1)$. The system creates a double butterfly-shaped attractor. Suppose that there are again a sufficient number of measurements so that we can obtain $n_{epo} = 64$ epochs of $N = 8000$ data each. In this case, more measurements were required given the more complex nature of the chaotic attractor. Using fewer measurements is possible, but the risk of not covering all the parts of the chaotic attractor with each trajectory increases. An example is shown in Figure 4.6a. Once the number of radii M = 10 is fixed both for the statistic concerning the states and the one concerning their derivatives, it is possible to again obtain the constants of the method, respectively $(R_0 = 2.17, b = 2.16)$ and $(R0 = 1.23, b = 2.61)$.

The posterior distribution was computed three times. The first using only the information concerning the states, the second also using the information concerning the approximations of the derivatives from noisy data, and the last using the noise-free derivatives obtained directly from the equations of the system instead of the approximations assuming zero error in the measurements. The marginal 2D of the posterior distribution presented in Figure 4.6 show how using two statistics in the creation of the feature vector greatly reduces the uncertainty in the posterior distribution. Noise-free derivative values further increase the accuracy, but not as dramatically for most parameters.

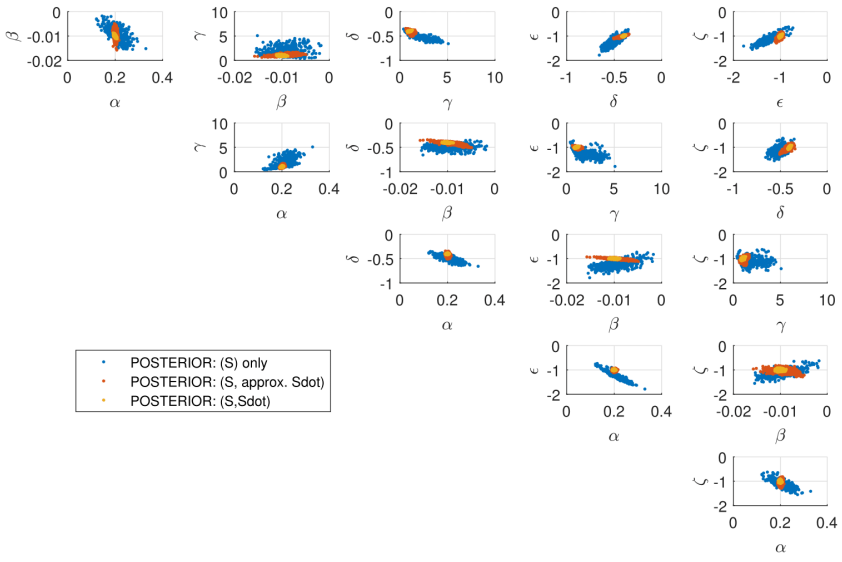
Chua7 The equations governing the evolution over time of the Chua7 system are given by

$$\begin{cases} \dot{X} &= \alpha \cdot (Y - h); \\ \dot{Y} &= X - Y + Z; \\ \dot{Z} &= -\beta \cdot Y; \\ h &= m_7 \cdot X + 0.5 \\ &\times \ ((m_0 - m_1) \cdot (|X + c_1| - |X - c_1|) \\ &+ \ (m_1 - m_2) \cdot (|X + c_2| - |X - c_2|) \\ &+ \ (m_2 - m_3) \cdot (|X + c_3| - |X - c_3|) \\ &+ \ (m_3 - m_4) \cdot (|X + c_4| - |X - c_4|) \\ &+ \ (m_4 - m_5) \cdot (|X + c_5| - |X - c_5|) \\ &+ \ (m_5 - m_6) \cdot (|X + c_6| - |X - c_6|) \\ &+ \ (m_6 - m_7) \cdot (|X + c_7| - |X - c_7|)) \, . \end{cases} \tag{4.5}$$

The system generates seven distinct 'zones' in the three dimensional (3D) state space. There are 17 parameters of interest that we are going to estimate, with the reference



(a)



(b)

Figure 4.6: Wang: (a) Wang model attractor with magnitudes of the state vector derivatives indicated by colours. (b) Comparison between the posteriors obtained from noisy data with a (5%) noise level. The posterior in blue is produced by measured state values only. The posterior in red is produced by using state and derivative values. For comparison, the posterior in yellow is given by noiseless state and derivative values.

values

$$(\alpha, \beta, m_0, m_1, m_2, m_3, m_4, m_5, m_6, m_7, c_1, c_2, c_3, c_4, c_5, c_6, c_7) = \\ (14, 20, 0.9/7, -3/7, 0.5, -0.3429, 0.36, -0.24, 0.36, -0.24).$$

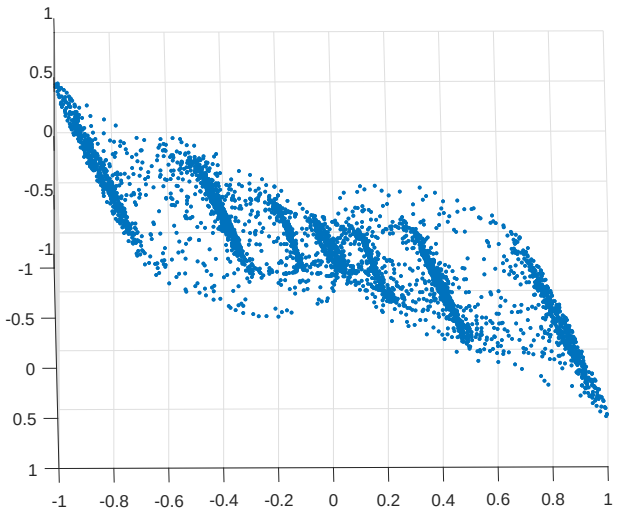
Suppose again a sufficient number of measurements available, so that we can obtain $n_{epo} = 64$ epochs of $N = 4000$ measurements each. In this case both the states and their derivatives are also used in the parametric inference. Two studies were conducted in this case. In the first case, it was assumed that the measurement error was absent, while in the second case, we added a Gaussian relative error of 1%. As shown in Figure 4.7 in both cases all 17 parameters of the model have been identified correctly. The main difference is that in the case without error the posterior ones have a maximum of 2% of relative error, while in the case in which the error has been added this relative error in the posterior increases to become at most 6%. The accuracy of the result is somewhat surprising. We believe this is due to the fact that the model is very sensitive to parametric variations. In particular, it is enough to slightly vary the parameters outside the posterior distributions sampled, with the impact that the number of 'zones' that determine the attractor goes from seven to five or three.

A pathological case Finally, we discuss an example where efforts to identify the model by state space information fail. The CIL approach is based on the idea that time series data is interpreted as samples from an underlying attractor. While the dynamics of the system is unpredictable in time, the attractor in the state space is assumed to be fixed. Moreover, a successful parameter estimation requires that the geometry of the attractor depends on the model parameters in a continuous way. In the present 'pathological' example introduced in Bouali (2014), this is not true. The system is given by the equations

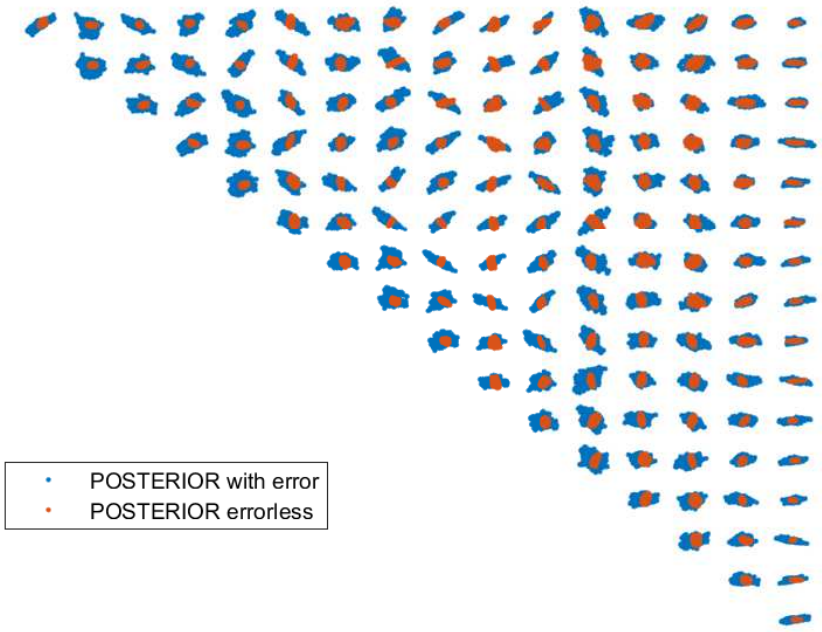
$$\begin{cases} \dot{X} = aX \cdot (1 - Y) - bZ; \\ \dot{Y} = -cY \cdot (1 - X^2); \\ \dot{Z} = dX. \end{cases} \tag{4.6}$$

We use the reference parameters values used in Bouali (2014), $(a = 3, b = 2.2, c = 2, d = 0.001)$. Integrating the model with a fixed parameter but slightly different initial values naturally leads to diverging trajectories. However, unlike for all the other chaotic systems described in this work or in Springer et al. (2019), now the *attractor itself* radically changes as well. Figure 4.8 shows three examples obtained in this way. Even if the integration time interval is chosen to be as long as practically possible, the system seems to be converging to various different stable regions, depending on small changes in the initial conditions. This might naturally happen if the initial values happen to be just on the boundary of domains of attraction. However, the same behaviour can be verified with quite different initial state values. On the other hand, quite different parameter values a, b, c, d may lead to similarly behaving trajectories.

In this situation the creation of the CIL likelihood fails. Already the first step, the verification that the feature vectors of the training set would follow the χ^2 test, demonstrate the



(a)



(b)

Figure 4.7: Chua 7: (a) an example of the states of the system. (b) 2D posterior marginal distributions.

opposite, the empirical distribution of data has a longer tail than the respective χ^2 density. We note that, here, the same as in several other 3D test cases employed in this work, equation (4.6) can be written in the linear form $\dot{S} = A(S)\theta$, where $S = (X, Y, Z)$, $\theta = (a, b, c, d)$ and $A(S)$ is the respective matrix depending on S. As discussed in the introduction of the present paper, in the special case that all the components of S and the time derivatives $\dot{S}$ are known at sufficiently many time instants, the values of θ can be solved trivially by a linear regression. However, indeed, the solution is limited to this special case, while the CIL approach enables the parameter estimation with partial observations and without the derivative information, provided the system has a stable attractor.

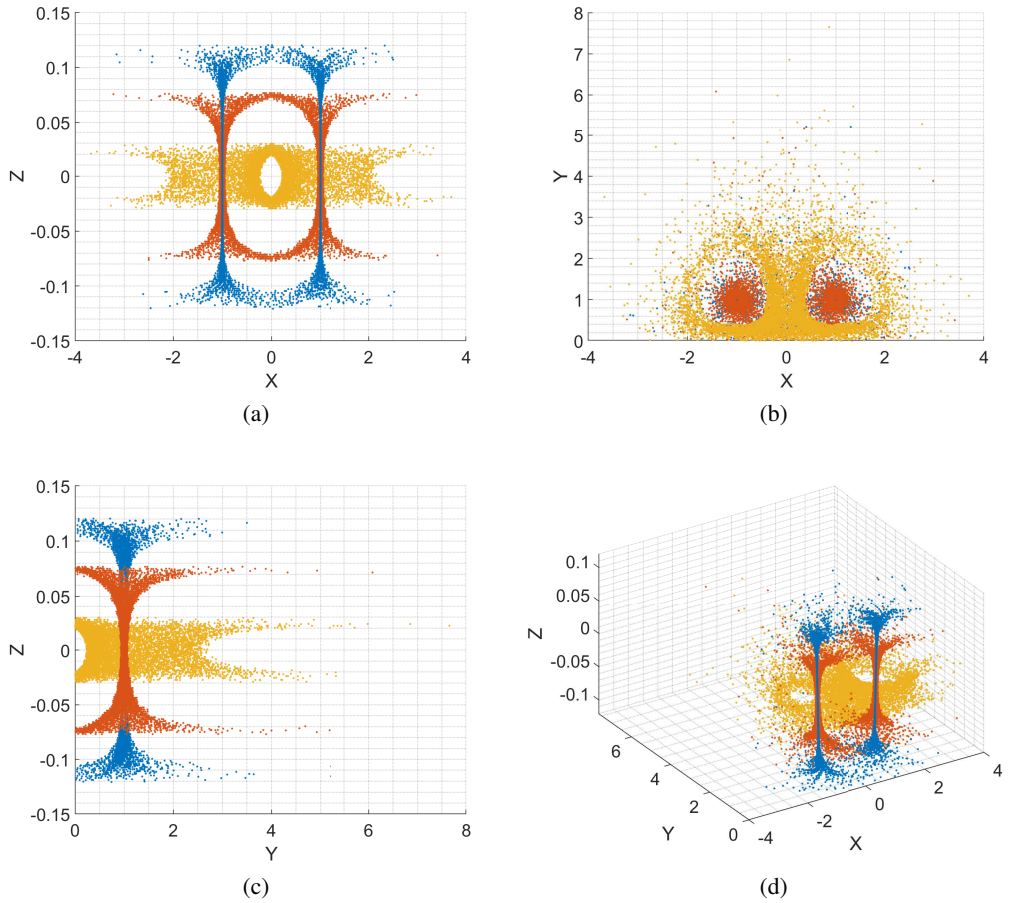


Figure 4.8: The trajectories were obtained with fixed model parameter $(a = 3, b = 2.2, c = 2, d = 0.001)$ and by randomising the initial condition (1;1;-0.02) with 1% relative Gaussian multiplicative noise

4.2 High dimensional chaotic dynamical systems

The Kuramoto-Shivashinsky model In the following paragraph, we will analyse the multidimensional chaotic system introduced in (Kuramoto and Yamada, 1976; Kuramoto, 1978; Sivashinsky, 1977) as a model for phase turbulence in reaction-diffusion systems and instabilities of laminar flames. The equation governing the temporal evolution of this chaotic PDE, denoted by (KS) in what follows, is given by

$$s_t = -ss_x - \frac{1}{\eta}s_{xx} - \gamma s_{xxxx}, \tag{4.7}$$

where $s = s(x,t)$, $x \in \mathbb{R}$ and $t \in \mathbb{R}_+$. It is also assumed that $s(x+L,t) = s(x,t)$, i.e., that the solution s is spatially periodic with period L.

In our experiments, we use the parameterisation introduced in (Yiorgos Smyrlis, 1996) that maps the spatial domain $[-\frac{L}{2}, \frac{L}{2}]$ into $[-\pi, \pi]$ by the change of variable $\tilde{x} = \frac{2\pi}{L}x$ and $\tilde{t} = \left(\frac{2\pi}{L}\right)^2 t$. Fixed L = 100, reference parameters $\gamma = (\pi/50)^2$ and $\eta = \frac{1}{2}$ are obtained. Furthermore, let us assume that the solution of this problem can be represented by truncating the Fourier series

$$s(x,t) = \sum_{j=0}^{\infty}\left[A_j(t)\sin\left(\frac{2\pi}{L}jx\right) + B_j(t)\cos\left(\frac{2\pi}{L}jx\right)\right]. \tag{4.8}$$

It follows that the initial problem boils down to solving the system of ODEs

$$\dot{A}_j(t) = \alpha_1 j^2 A_j(t) + \alpha_2 j^4 A_j(t) + F_1(A(t)) \tag{4.9}$$

$$\dot{B}_j(t) = \beta_1 j^2 B_j(t) + \beta_2 j^4 B_j(t) + F_2(B(t)), \tag{4.10}$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are polynomials of the vectors A and B, whose coefficients $A_j(t)$ and $B_j(t)$, are unknown. For further details, see (Huttunen et al., 2018). An immediate advantage of writing the problem in this form is that it is possible to calculate the solution in parallel efficiently using graphics cards currently available. In our case, using an Nvidia GTX 1070 graphical card, it is possible to simulate several thousand simulations in parallel when the x-dimension discretization contains about 500 points. In what follows, we will use a 256 dimensional discretisation.

Our experiment begins by creating the reference sample which consists of 64 epochs of the 256-dimensional system of KS of 1024 time points each. For each of the epochs, the system is integrated for the time interval $[0, 150000]$. After discarding a predictable part [0 500], 1024 equidistant values are selected as the observation instants, i.e. such that $\Delta_t \approx 146$ is the interval between the observations. The CIL approach is used, with one feature vector that consists of distances between all or different subsets of the 256 components of the system. After scaling, the states of the system in the multidimensional cube $[-1,1]^{256}$ we get the constants to construct the eCDF vectors with $R_0 = 1801.7$, $M = 32$, and $b = 1.025$. Unlike the cases discussed so far, where integrating the model once took a fraction of a second, for the KS model it takes 103 seconds for obtaining a

solution for the time interval $[0, 150000]$. It follows that, to obtain a chain of 10^5 model parameter values, it would take about 4 months of calculation, provided that the laptop does not give up first.
A partial solution could be to use several parallel chain, but this option will not be elaborated in this work. A valid alternative is to break the integration of the system into subintervals. This is possible as the GLIF methods simply require that the samples used to calculate the feature vectors are a good representatives of the underlying attractor. In particular, we will divide the time integration into 128 parts that are computed in parallel using randomised initial values. This amounts to integrating over time intervals $[0, 4500]$. From each of these, we remove the predictable time frame $[0, 500]$ and take 8 equidistant measurements that yields a total of 1024 points to use to create the feature vectors. In this case, we can afford to take a step $\Delta_t = 500$ between measurements, i.e. to have more independent state samples than in the original setting. In this way, the time required to obtain a feature vector drops from 103 seconds to 2.5 seconds, reducing the cost of obtaining the chain of 10^5 elements to about 20 hours. This allows us to compute the parameter posterior in a standard way using the AM algorithm for sampling.

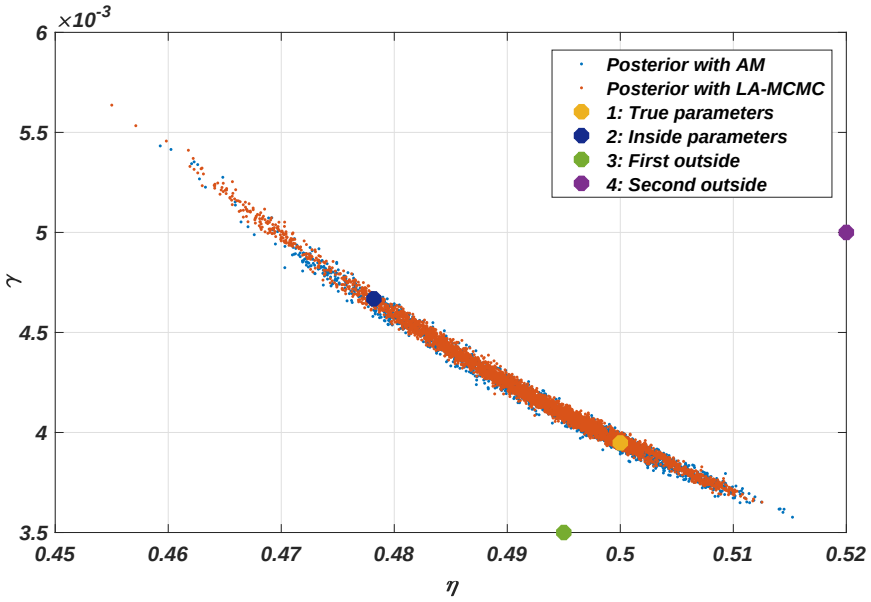


Figure 4.9: Posterior distribution of the parameters of the KS system. The parameter values are shown in Table 4.1, while examples of the respective integrated trajectories are given in Fig. 4.10.

Despite this important reduction, 20 hours is still high. Therefore, we further reduce the time necessary to obtain the posterior distribution using the LA-MCMC method. The experiment was repeated several times to see the robustness of the solution. In all the simulated cases, there was no significant difference with respect to the posterior distribution obtained by AM. However, there was a drastic reduction in the number of likelihood

Table 4.1: Parameter values of the four parameter vectors used in the forward KS model simulation examples in Fig. 4.10. The parameter vector in the first column labelled 1 are the true parameters, and the second one resides inside the posterior. The last two are outside the posterior, Fig. 4.9.

	Case 1	Case 2	Case 3	Case 4
η	0.5	0.4782	0.495	0.52
γ	0.00395	0.00467	0.0035	0.005

evaluations necessary to obtain the a posterior distribution. The same as in the 3D Lorenz case, the reduction was about two decades, more specifically, the full likelihood evaluations varied between 1131 and 1221 in the experiments conducted.
By combining LA-MCMC and the reduction obtained by integrating the solution in parallel sub-intervals, it was possible to reduce the time required to obtain a chain of 10^5 elements from 4 months to about 45 minutes. Figure 4.9 shows the posterior distribution obtained by the two methods, standard AM and adaptive LA-MCMC.

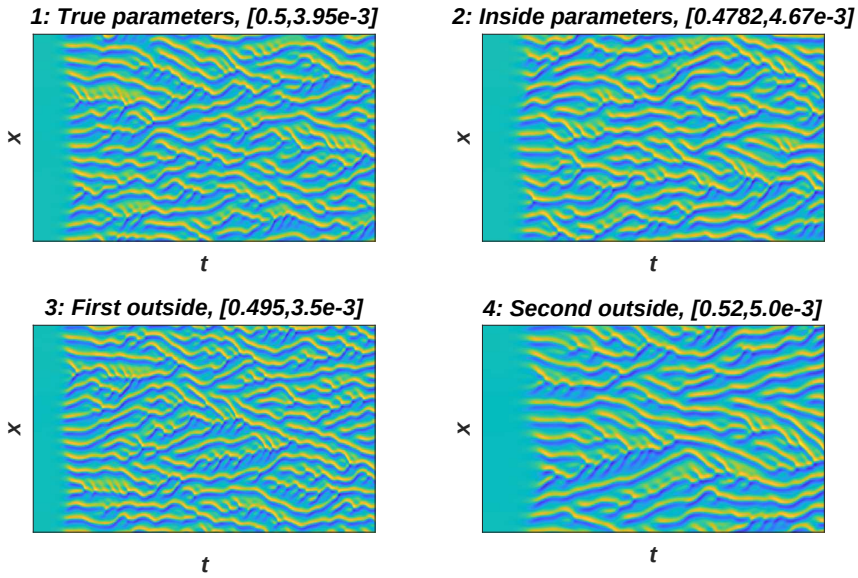


Figure 4.10: Example model trajectories from the KS system. Figure 1) shows simulation using the true parameters, the parameters used for Figure 4) are inside the posterior distribution, and Figures 2) and 3) are generated from simulations with parameters outside the posterior distribution, shown in Fig. 4.9. The values of the parameter vectors 1, 2, 3 and 4 are given in Table 4.1. The y-axis shows the 256-dimensional state vector, and the x-axis the time evolution of the system.

To study possible visual differences between the trajectories, three parameter cases were

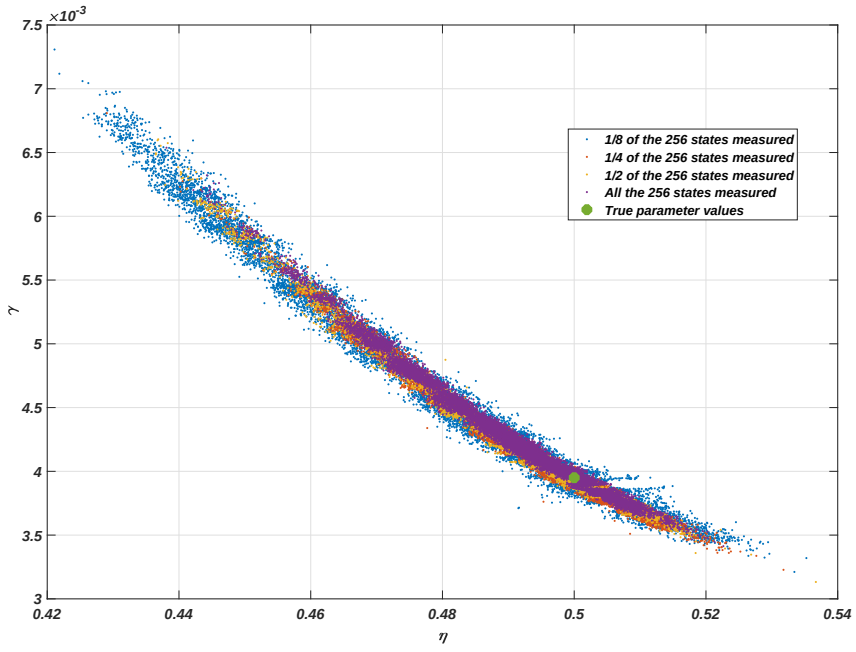


Figure 4.11: Comparison between the KS system's posterior distribution in case that all or one part of the states are observed

selected on basis of the sampled posterior, in addition to the reference one. Table 4.1 gives the values, which are also plotted in Figure 4.9.
In two cases, the parameters lie within the 99% confidence region while the other two lie outside it. From a visual analysis, it can be said that only the fourth case, most far away from the posterior, has visibly different trajectory characteristics as compared to the others. This indicates that likelihood is able to distinguish differences that are difficult to perceive with the naked eye.
A second set of experiments was conducted to see the stability of the solution when the amount of observed states of the system decreases. By keeping the rest of the experiment the same as in the first case, the observed states were reduced from 256 to 128,64 and 32. The results shown in Figure 4.11 show how the posterior increases with decreasing observed states.

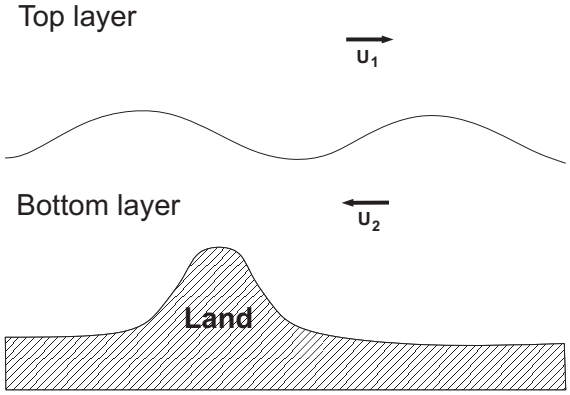


Figure 4.12: An example of the layer structure of the two-layer quasi-geostrophic model. The terms U_1 and U_2 denote mean zonal flows respectively in the top and the bottom layer.

Quasi-geostrophic model The second multidimensional example studied is the quasi-geostrophic (QG) model by Fandry and Leslie (1984); Pedlosky (1987). The quasi-geostrophic two-layer model simulates the behaviour of the wind on a cylindrical surface divided into two layers, where the underlying layer is influenced by a tomography, chosen by the user to simulate the surface landscape. In addition, the model parameters include the thickness of the two layers of the atmosphere, denoted by H_1 and H_2. The Coriolis forces acting on the system are denoted by f_0. In Figure 4.13, we can see an example of a snapshot of the QG system, while in Figure 4.12 an example of the two-layer geometry is given.

In a non-dimensional form the QG system can be written as

$$q_1 = \Delta\psi_1 - F_1(\psi_1 - \psi_2) + \beta y, \tag{4.11}$$
$$q_2 = \Delta\psi_2 - F_2(\psi_2 - \psi_1) + \beta y + R_s, \tag{4.12}$$

where q_i are the potential vorticities, and ψ_i the stream functions with indexes $i = 1, 2$ for the upper and the lower layers, respectively. Both the q_i and ψ_i are functions of time t and spatial coordinates x, and y. The coefficients $F_i = \frac{f_0^2 L^2}{\acute{g} H_i}$ control how much the model layers interact, β is the northward gradient of the Coriolis force that gives rise to faster cyclonic flows closer to the poles, L is a length scale constant and $\acute{g}$ is a gravity constant. Finally, $R_s(x, y) = \frac{S(x,y)}{\eta H_2}$ defines the topography for the lower layer where $\eta = \frac{U}{f_0 L}$ is the Rossby number of the system with L and U designating the length and speed scales, respectively. For further details, see (Fandry and Leslie, 1984) and (Pedlosky, 1987).

It is assumed that the motion determined by the model is geostrophic, essentially meaning that potential vorticity of the flow is preserved in both layers:

$$\frac{\partial q_i}{\partial t} + u_i \frac{\partial q_i}{\partial x} + v_i \frac{\partial q_i}{\partial y} = 0. \tag{4.13}$$

Here, u_i and v_i are velocity fields, which are functions of both space and time. They are

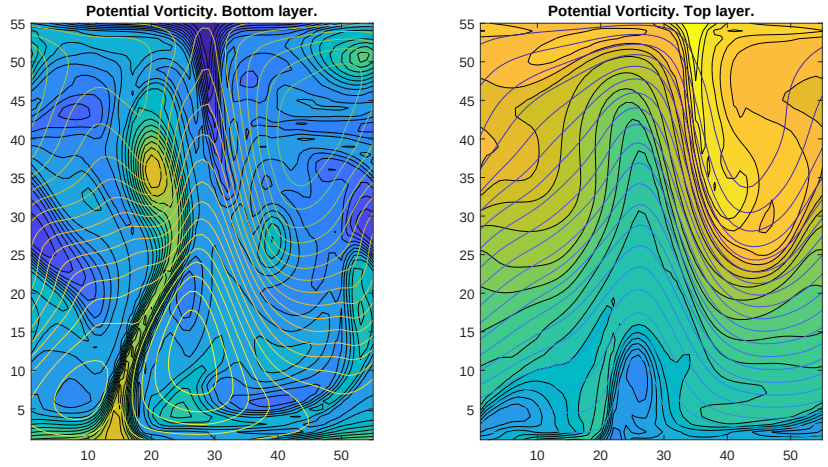


Figure 4.13: An example of the 6050-dimensional state of the quasi-geostrophic model. Note the cylindrical boundary conditions.

obtained from the stream functions ψ_i via

$$u_i = -\frac{\partial \psi_i}{\partial y}, \qquad v_i = \frac{\partial \psi_i}{\partial x}. \tag{4.14}$$

Equations (4.11)–(4.14) define the spatio-temporal evolution of the quantities q_i, ψ_i, $i = 1, 2$.

The numerical integration of this system is carried out using a Semi-Lagrangian scheme, where the potential vorticities q_i are computed according to Eq. (4.13) for given velocities u_i and v_i. With these q_i the stream functions can then be obtained from Eq. (4.11) and (4.12) with a two-stage finite difference scheme. Finally, the velocity field is updated by Eq. (4.14) for the next iteration round. For more details, see (Fandry and Leslie, 1984).

To obtain more detailed and consistent chaotic fields the spatial discretisation of the grid has been increased from the often used 20×40 to a more dense 55×55. In Figure 4.13 a snapshot of the 6050-dimensional trajectory thus obtained is shown.

The training data was generated by integrating the system 64 times on the time interval $[0, 8192]$, and extracting 1000 equidistant observations by skipping the initial predictable interval $[0, 192]$. For this model, $\Delta t = 1$ corresponds to 6h of 'real' time, which means that the time interval $[192, 8192]$ amounts to a long-range integration of roughly 5-6 years of a climate model. For this setting, the integration time of a trajectory is about 10 minutes, thus making about two years of calculation time to obtain a chain of 10^5 elements. The only way to obtain the posterior distribution in a reasonable time is, therefore, to use the methods of LA-MCMC shown in the previous examples. The model states observed for both layers consists of two distinct fields that depend on each other, namely the vorticity and stream function. We will create one feature vector for the potential vorticity on both layers and the other for the stream function by following the suggestions discussed

in (Haario et al., 2015), where it revealed to be useful to construct separate feature vectors to characterise two different parts of the Lorenz95 system. Our feature vector will be again the correlation integral. The CIL type of likelihood is obtained by stacking these two feature vectors one after another. The Gaussianity of the thus obtained $2(M+1)$ dimensional FV is again verified numerically. The number of bins selected was $M = 32$ leading to the constants $R_0 = 55$ and $b = 1,075$ for potential vorticity, and $R_0 = 31$, and $b = 1.046$ for the stream function. Compared to the other cases studied so far, there is a higher variability in the FV attributable due to the higher dimensionality of the model, and a more complex dynamics of the underlying system.

In our experiments, the parameters of interest were the layer heights, with the reference values used to construct the training data $H_1 = 5500$ and $H_2 = 4500$. Using the LA-MCMC algorithm, we generated a few chains of 10^5 elements each. The number of full likelihood evaluations varied between 682 and 762, which would translate to around five days of computing time on the laptops used. As with Kuramoto-Sivashinsky, the forward model integration can be split to segments computed in parallel, which reduces the required time for computing the likelihood to around 20 seconds, corresponding to 4 h for generating the MCMC chain. The time required to compute the pairwise distances for generating the feature vectors was negligible if compared to the model integration time. In Figure 4.14 we present the posterior distribution thus obtained.

To summarise, a calculation time of roughly 2 years by brute force can be reduced to 4 h by the LA-MCMC method combined with parallel integrations. We note that a further reduction is possible by running parallel MCMC chains. That option was, however, not used in this work.

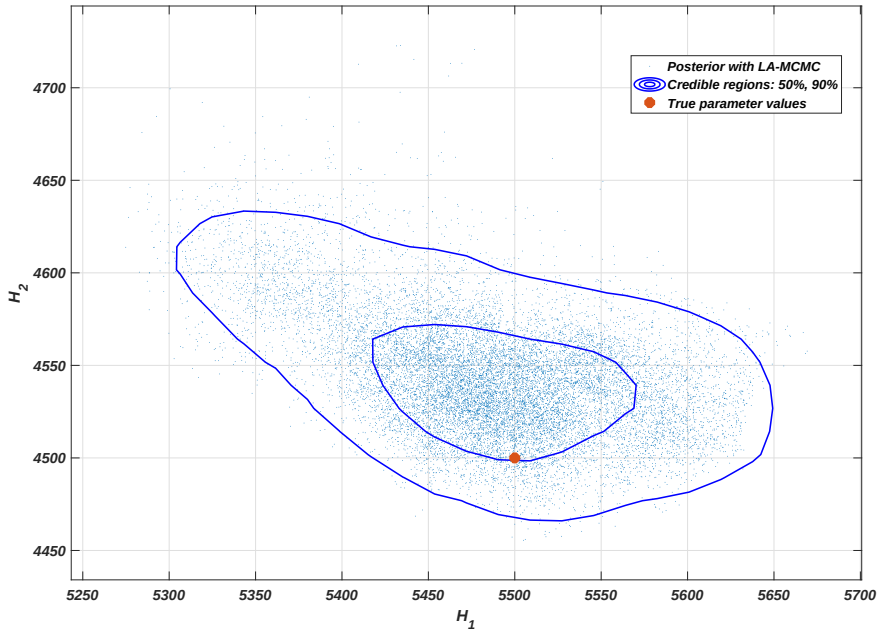


Figure 4.14: The posterior distribution of the H_1 and H_2-parameters of the quasi-geostrophic system shows how these parameters slightly anticorrelate with each other.

5 Summary and future work

The verification of scientific models against real-world observations is a central part of science. However, as only a finite number of measurements are available, which are always more or less noisy and often biased, it is impossible to identify model parameters uniquely. In the standard deterministic setting, where the measurement error is assumed to be known and Gaussian, an informative cost function that captures the model-data difference can be written using the residual sum of squares (RSS). This approach is not directly applicable for chaotic or stochastic systems. In case of stochastic differential equations (SDE), an approach that became standard is to infer model parameters using one of the sequential data assimilation approaches widely presented in the literature, the Kalman filter as the basic example. For chaotic systems, the sequential data assimilation approach is applicable, too. This is well documented by everyday weather forecasts: a forecast is created by current knowledge, the prediction is corrected when new data becomes available, and the corrected state is again the starting point for a next forecast. However, such a process is possible only if measurements of the system are available at a frequency dense enough, so that meaningful predictions can be made. For long-time predictions, such as is needed for climate modelling, for example, this is not the case. There is, therefore, a need for new methodologies that enable parameter estimation also for situations in which it is not possible to predict the dynamics for a time range of interest by the initial values currently known. Chaotic systems with time-wise sparse measurements provide a generic example of such situations.

A set of time series measurements can also be considered as a set of samples from an underlying distribution (e.g., the underlying chaotic attractor). One possibility is, therefore, to characterise the difference in distribution between two set of measurements in state space instead of doing it point by point in time. We propose here an approach to infer model parameters based on such features of the data.

Different features extracted from the data can be designed so as to capture specific characteristics of the underlying system studied. We have shown how to combine different feature vectors into one single likelihood so as to increase the accuracy of the posterior distribution. The theorems in the U-statistics literature state that, under general assumptions, the studied feature vectors are asymptotically Gaussian. It is, therefore, possible to construct a likelihood by estimating the mean and covariance of the feature vectors created by data, and use it to sample the posterior distribution of model parameters. In this way problems previously declared as unsolvable in the literature find a solution. To test the robustness of our approach, it has been tested on a set of different problems both in low and high dimensional. In particular, we performed parameter identification for chaotic (ODE) and chaotic (PDE) systems. In cases where the forward model is expensive, the MCMC approach become often computationally infeasible. To overcome these difficulties, we combined our type of likelihoods with the Local Approximation MCMC. The intuition behind this recently introduced type of MCMC algorithm is to substitute many of the expensive likelihood evaluations with inexpensive local polynomial approximations of a target likelihood based on few neighbouring evaluations of this last. By using the LA-MCMC approach, the computational cost revealed to be reduced by orders

of magnitude. Given that the likelihood is implicit, based on features of data rather than a direct comparison to model simulations, a lack-of-fit might remain unnoticed in cases where the estimation 'technically' seems fine. To detect the discrepancy between data or model simulations in such situations, a diagnostic test for goodness-of-fit fit is proposed and demonstrated.

There are many potential directions for extension of this work. A different type of problems might require a different type of data reduction to feature vectors to be informative with respect to model parameters. A further improvement could be obtained by having algorithms that extract the most informative feature vectors from the data so as to improve the precision of the posterior distribution. Moreover, other approaches based on feature vectors, such as Bayesian Synthetic Likelihood (BSL), could potentially benefit from having a more rich list of possible type of feature vectors to be used for parameter estimation. Our approach requires a certain number of repeated experiments to be implementable while the BSL approach could be often used even with less data at a higher computational cost. Further resources might be allocated in inspecting the limits within which it is better to use one or the other of the approaches. In addition to the examples presented in this work, our approach has been shown to be useful in quite different applications, such as in identifying model parameters of reaction-diffusion systems by using the steady-state solutions of the Turing patterns only, see Kazarnikov and Haario (2020) as well as for chaotic SDE systems, see Maraia et al. (2020). Examples of further possible types of problems include financial time series, image classification, synchronization or climate models. Although answering these questions requires further work, the research presented in this thesis provides a promising step towards quantifying uncertainties for problems that, until now, had no solution.

References

Borovkova, S., Burton, R., and Dehling, H. (2001). Limit theorems for functionals of mixing processes with applications to U-statistics and dimension estimation. *Transactions of the American Mathematical Society*, 353(11), pp. 4261–4318. doi:10.1090/S0002-9947-01-02819-7.

Bouali, S. (2014). A 3D Strange Attractor with a Distinctive Silhouette. The Butterfly Effect Revisited. *arXiv*, pp. 1–10. url: https://arxiv.org/abs/1311.6128.

Brooks, S., Gelman, A., Jones, G., and Meng, X.L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15), pp. 3932–3937. doi:10.1073/pnas.1517384113, url: https://www.pnas.org/content/113/15/3932.

Chen, N. and Majda, A.J. (2016). Filtering Nonlinear Turbulent Dynamical Systems through Conditional Gaussian Statistics. *Monthly Weather Review*, 144(12), pp. 4885–4917. doi:10.1175/MWR-D-15-0437.1.

Conrad, P.R., Marzouk, Y.M., Pillai, N.S., and Smith, A. (2016). Accelerating Asymptotically Exact MCMC for Computationally Intensive Models via Local Approximations. *Journal of the American Statistical Association*, 111(516), pp. 1591–1607. doi:10.1080/01621459.2015.1096787, url: https://doi.org/10.1080/01621459.2015.1096787.

Davis, A., Marzouk, Y., Smith, A., and Pillai, N. (2020). Rate-optimal refinement strategies for local approximation MCMC. *arXiv e-prints*, arXiv:2006.00032.

Davis, A.D. (2018). *Prediction under uncertainty: from models for marine-terminating glaciers to Bayesian computation*. MIT Doctoral thesis.

Donsker, M. (1951). An invariance principle for certain probability limit theorems. *Memoirs of the American Mathematical Society*, 6.

Fandry, C.B. and Leslie, L.M. (1984). A Two-Layer Quasi-Geostrophic Model of Summer Trough Formation in the Australian Subtropical Easterlies. *Journal of the Atmospheric Sciences*, 41(5), pp. 807–818. doi:10.1175/1520-0469(1984)041¡0807:ATLQGM¿2.0.CO;2, url: https://doi.org/10.1175/1520-0469(1984)041<0807:ATLQGM>2.0.CO;2.

Grassberger, P. and Procaccia, I. (1983a). Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1), pp. 189 – 208. ISSN 0167-2789, doi:https://doi.org/10.1016/0167-2789(83)90298-1, url: http://www.sciencedirect.com/science/article/pii/0167278983902981.

Grassberger, P. and Procaccia, I. (1983b). Procaccia, I.: Estimation of the Kolmogorov entropy from a chaotic signal. Phys. Rev. A 28, 2591-2593. *Physical Review A - PHYS REV A*, 28, pp. 2591–2593. doi:10.1103/PhysRevA.28.2591.

Haario, H., Kalachev, L., and Hakkarainen, J. (2015). Generalized correlation integral vectors: A distance concept for chaotic dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(6), p. 063102. doi:10.1063/1.4921939.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4), pp. 339–354. ISSN 1573-1375, doi:10.1007/s11222-006-9438-0.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2), pp. 223–242.

Hakkarainen, J., et al. (2012). On closure parameter estimation in chaotic systems. *Nonlinear Processes in Geophysics*, 19(1), pp. 127–143. doi:10.5194/npg-19-127-2012.

Hakkarainen, J., et al. (2013). A dilemma of the uniqueness of weather and climate model closure parameters. *Tellus A: Dynamic Meteorology and Oceanography*, 65(1), p. 20147. doi:10.3402/tellusa.v65i0.20147.

Huttunen, J., Kaipio, J., and Haario, H. (2018). Approximation error approach in spatiotemporally chaotic models with application to Kuramoto-Sivashinsky equation. *Computational Statistics & Data Analysis*, 123, pp. 13 – 31. ISSN 0167-9473, doi:https://doi.org/10.1016/j.csda.2018.01.015, url: http://www.sciencedirect.com/science/article/pii/S0167947318300240.

Jarvinen, H., Laine, M., Solonen, A., and Haario, H. (2011). Ensemble prediction and parameter estimation system: the concept. *Quarterly Journal of the Royal Meteorological Society*, 138(663), pp. 281–288. doi:10.1002/qj.923, url: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.923.

Kazarnikov, A. and Haario, H. (2020). Statistical approach for parameter identification by Turing patterns. *Journal of Theoretical Biology*, 501, p. 110319. ISSN 0022-5193, doi:https://doi.org/10.1016/j.jtbi.2020.110319, url: http://www.sciencedirect.com/science/article/pii/S0022519320301740.

Kuramoto, Y. and Araki, H.e. (1975). *Lecture Notes in Physics, International Symposium on Mathematical Problems in Theoretical Physics*. Springer-Verlag, New York.

Kuramoto, Y. (1978). Diffusion-Induced Chaos in Reaction Systems. *Progress of Theoretical Physics Supplement*, 64, pp. 346–367. doi:10.1143/PTPS.64.346, url: http://dx.doi.org/10.1143/PTPS.64.346.

Kuramoto, Y. (1984). *Chemical Oscillations, Waves, and Turbulence*. Springer-Verlag, New York.

Kuramoto, Y. and Yamada, T. (1976). Turbulent State in Chemical Reactions. *Progress of Theoretical Physics*, 56(2), pp. 679–681. doi:10.1143/PTP.56.679, url: http://dx.doi.org/10.1143/PTP.56.679.

Laine, M., Solonen, A., Haario, H., and Jarvinen, H. (2011). Ensemble prediction and parameter estimation system: the method. *Quarterly Journal of the Royal Meteorological Society*, 138(663), pp. 289–297. doi:10.1002/qj.922, url: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.922.

Lorenz, E.N. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2), pp. 130–141. doi:10.1175/1520-0469(1963)020¡0130:DNF¿2.0.CO;2, url: https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Maraia, R., et al. (2020). Parameter estimation of stochastic chaotic systems. *International Journal for Uncertainty Quantification*. doi: https://doi.org/10.1615/Int.J.UncertaintyQuantification.2020032807.

Mira, A. (2001). On Metropolis-Hastings algorithms with delayed rejection. *Metron*, LIX(34), pp. 231–241.

Neumeyer, N. (2004). A central limit theorem for two-sample U-processes. *Statistics & Probability Letters*, 67(1), pp. 73 – 85. ISSN 0167-7152, doi: https://doi.org/10.1016/j.spl.2002.12.001.

Pedlosky, J. (1987). *Geophysical Fluid Dynamics, 22-57*. Springer-Verlag, New York.

Price, L.F., Drovandi, C.C., Lee, A., and Nott, D.J. (2018). Bayesian Synthetic Likelihood. *Journal of Computational and Graphical Statistics*, 27(1), pp. 1–11. doi:10.1080/10618600.2017.1302882, url: https://doi.org/10.1080/10618600.2017.1302882.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.

Shemyakin, V. and Haario, H. (2018). Online identification of large scale chaotic system. *Nonlinear Dynamics*. doi:DOI: 10.1007/s11071-018-4239-5.

Sivashinsky, G. (1977). Nonlinear analysis of hydrodynamic instability in laminar flames-I. Derivation of basic equations. *Acta Astronautica*, 4(11), pp. 1177 – 1206. ISSN 0094-5765, doi:https://doi.org/10.1016/0094-5765(77)90096-0, url: http://www.sciencedirect.com/science/article/pii/0094576577900960.

Springer, S., et al. (2019). *Robust parameter estimation of chaotic systems*. doi: $10.3934/ipi.2019053$, ISSN $1930 - 8337$.

Wang, Z., et al. (2009). A 3-D four-wing attractor and its analysis. *Brazilian Journal of Physics*, 39, pp. 547 – 553. ISSN 0103-9733.

Wood, S.N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310), p. 1102.

Yiorgos Smyrlis, D.P. (1996). Computational study of chaotic and ordered solutions of the Kuramoto-Shivashinsky equation. *NASA Contractor Report 198283*, 96-12.

Zhong, G.Q., Man, K.F., and Chen, G. (2002). Asystematic approach to generating n-scroll attractors. *International Journal of Bifurcation and Chaos*, 12(12), pp. 2907–2915. doi:10.1142/S0218127402006230, url: https://www.worldscientific.com/doi/abs/10.1142/S0218127402006230.

ACTA UNIVERSITATIS LAPPEENRANTAENSIS

912. KINNUNEN, SINI-KAISU. Modelling the value of fleet data in the ecosystems of asset management. 2020. Diss.

913. MUSIKKA, TATU. Usability and limitations of behavioural component models in IGBT short-circuit modelling. 2020. Diss.

914. SHNAI, IULIIA. The technology of flipped classroom: assessments, resources and systematic design. 2020. Diss.

915. SAFAEI, ZAHRA. Application of differential ion mobility spectrometry for detection of water pollutants. 2020. Diss.

916. FILIMONOV, ROMAN. Computational fluid dynamics as a tool for process engineering. 2020. Diss.

917. VIRTANEN, TIINA. Real-time monitoring of membrane fouling caused by phenolic compounds. 2020. Diss.

918. AZZUNI, ABDELRAHMAN. Energy security evaluation for the present and the future on a global level. 2020. Diss.

919. NOKELAINEN, JOHANNES. Interplay of local moments and itinerant electrons. 2020. Diss.

920. HONKANEN, JARI. Control design issues in grid-connected single-phase converters, with the focus on power factor correction. 2020. Diss.

921. KEMPPINEN, JUHA. The development and implementation of the clinical decision support system for integrated mental and addiction care. 2020. Diss.

922. KORHONEN, SATU. The journeys of becoming ang being an international entrepreneur: A narrative inquiry of the "I" in international entrepreneurship. 2020. Diss.

923. SIRKIÄ, JUKKA. Leveraging digitalization opportunities to improve the business model. 2020. Diss.

924. SHEMYAKIN, VLADIMIR. Parameter estimation of large-scale chaotic systems. 2020. Diss.

925. AALTONEN, PÄIVI. Exploring novelty in the internationalization process - understanding disruptive events. 2020. Diss.

926. VADANA, IUSTIN. Internationalization of born-digital companies. 2020. Diss.

927. FARFAN OROZCO, FRANCISCO JAVIER. In-depth analysis of the global power infrastructure - Opportunities for sustainable evolution of the power sector. 2020. Diss.

928. KRAINOV, IGOR. Properties of exchange interactions in magnetic semiconductors. 2020. Diss.

929. KARPPANEN, JANNE. Assessing the applicability of low voltage direct current in electricity distribution - Key factors and design aspects. 2020. Diss.

930. NIEMINEN, HARRI. Power-to-methanol via membrane contactor-based CO2 capture and low-temperature chemical synthesis. 2020. Diss.

931. CALDERA, UPEKSHA. The role of renewable energy based seawater reverse osmosis (SWRO) in meeting the global water challenges in the decades to come. 2020. Diss.

932. KIVISTÖ, TIMO. Processes and tools to promote community benefits in public procurement. 2020. Diss.

933. NAQVI, BILAL. Towards aligning security and usability during the system development lifecycle. 2020. Diss.

934. XIN, YAN. Knowledge sharing and reuse in product-service systems with a product lifecycle perspective. 2020. Diss.

935. PALACIN SILVA, VICTORIA. Participation in digital citizen science. 2020. Diss.

936. PUOLAKKA, TIINA. Managing operations in professional organisations – interplay between professionals and managers in court workflow control. 2020. Diss.

937. AHOLA, ANTTI. Stress components and local effects in the fatigue strength assessment of fillet weld joints made of ultra-high-strength steels. 2020. Diss.

938. METSOLA, JAAKKO. Good for wealth or bad for health? Socioemotional wealth in the internationalisation process of family SMEs from a network perspective. 2020. Diss.

939. VELT, HANNES. Entrepreneurial ecosystems and born global start-ups. 2020. Diss.

940. JI, HAIBIAO. Study of key techniques in the vacuum vessel assembly for the future fusion reactor. 2020. Diss.

941. KAZARNIKOV, ALEXEY. Statistical parameter identification of reaction-diffusion systems by Turing patterns. 2020. Diss.

942. SORMUNEN, PETRI. Ecodesign of construction and demolition waste-derived thermoplastic composites. 2020. Diss.

943. MANKONEN, ALEKSI. Fluidized bed combustion and humidified gas turbines as thermal energy conversion processes of the future. 2020. Diss.

944. KIANI OSHTORJANI, MEHRAN. Real-time efficient computational approaches for hydraulic components and particulate energy systems. 2020. Diss.

945. PEKKANEN, TIIA-LOTTA. What constrains the sustainability of our day-to-day consumption? A multi-epistemological inquiry into culture and institutions. 2021. Diss.

946. NASIRI, MINA. Performance management in digital transformation: a sustainability performance approach. 2021. Diss.

947. BRESOLIN, BIANCA MARIA. Synthesis and performance of metal halide perovskites as new visible light photocatalysts. 2021. Diss.

948. PÖYHÖNEN, SANTERI. Variable-speed-drive-based monitoring and diagnostic methods for pump, compressor, and fan systems. 2021. Diss.

949. ZENG, HUABIN. Continuous electrochemical activation of peroxydisulfate mediated by single-electron shuttle. 2021. Diss.

LUT University

ISBN 978-952-335-627-6
ISBN 978-952-335-628-3 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491
Lappeenranta 2021