BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 8 2013, pages 1052–1059
doi:10.1093/bioinformatics/btt097

Systems biology

Advance Access publication February 22, 2013

An improved constraint filtering technique for inferring hidden states and parameters of a biological model

Syed Murtuza Baker[1,*], C. Hart Poskar[1], Falk Schreiber[2,3] and Björn H. Junker[1]

[1]Systems Biology Group, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Gatersleben, Germany, [2]Institute of Computer Science, Martin Luther University, 06120 Halle-Wittenberg, Germany and [3]Plant Bioinformatics Group, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Gatersleben, Germany

Associate Editor: Martin Bishop

ABSTRACT

Motivation: In systems biology, kinetic models represent the biological system using a set of ordinary differential equations (ODEs). The correct values of the parameters within these ODEs are critical for a reliable study of the dynamic behaviour of such systems. Typically, it is only possible to experimentally measure a fraction of these parameter values. The rest must be indirectly determined from measurements of other quantities. In this article, we propose a novel statistical inference technique to computationally estimate these unknown parameter values. By characterizing the ODEs with non-linear state-space equations, this inference technique models the unknown parameters as hidden states, which can then be estimated from noisy measurement data.

Results: Here we extended the square-root unscented Kalman filter SR-UKF proposed by Merwe and Wan to include constraints with the state estimation process. We developed the constrained square-root unscented Kalman filter (CSUKF) to estimate parameters of non-linear state-space models. This probabilistic inference technique was successfully used to estimate parameters of a glycolysis model in yeast and a gene regulatory network. We showed that our method is numerically stable and can reliably estimate parameters within a biologically meaningful parameter space from noisy observations. When compared with the two common non-linear extensions of Kalman filter in addition to four widely used global optimization algorithms, CSUKF is shown to be both accurate and computationally efficient. With CSUKF, statistical analysis is straightforward, as it directly provides the uncertainty on the estimation result.

Availability and implementation: Matlab code available upon request from the author.

Contact: baker@ipk-gatersleben.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 5, 2012; revised on February 7, 2013; accepted on February 18, 2013

1 INTRODUCTION

Living organisms develop a highly interconnected web of dynamic mechanisms between different levels of molecular components, which are organized into diverse biochemical reaction networks. A better understanding of the general principles of these biological processes requires a system-level study of the underlying mechanisms (Sun *et al.*, 2008). The integration of computational and mathematical methods with experimental efforts is necessary to gain such systems level understanding. In systems biology, the specific type of modelling that studies the detailed quantitative analysis of the dynamics of cellular processes is mechanistic or kinetic modelling (Stelling, 2004). Such models are often characterized by sets of ordinary differential equations (ODEs) (Sitz *et al.*, 2002) that describe the underlying interactive mechanisms of the system in a quantitative manner (Liu and Niranjan, 2012). They account for the change of concentration of certain chemical species such as the metabolites, messenger RNA (mRNA) or proteins over time. These changes are specified through the rate laws, which are determined from *in vitro* biochemical experiments. The rate laws may have multiple parameters that need to be determined correctly to faithfully mimic observations (Lillacci and Khammash, 2010; Liu and Niranjan, 2012). While it might be possible to experimentally measure some of these parameters, for most of them, conducting biological experiments to determine their values are difficult, expensive, time-consuming or not currently possible (Lillacci and Khammash, 2010). As a result, different computational techniques have been developed for the indirect determination of these unknown parameter values from some other measured quantities (Moles *et al.*, 2003).

In general, the parameter estimation problem is formulated as a non-linear optimization problem that aims at minimizing the difference between simulated and measured values from experimental time-course data (Arisi *et al.*, 2006). Starting with an initial guess, these methods exhaustively search the parameter space to find the set of values that produce the best fit between the simulated and the measured data. Extensive research exists on the application of different optimization techniques, such as non-linear least square fitting and methods following steepest descent gradient techniques (Mendes and Kell, 1998), methods suitable for global optimization including simulated annealing (SA; Kirkpatrick *et al.*, 1983) and evolutionary computation (Moles *et al.*, 2003). But due to the non-linear nature of the dynamics, the optimization problems of biochemical networks are most often multimodal (Moles *et al.*, 2003). Exploring such parameter space is computationally expensive for these optimization techniques. Moreover, they often get stuck in local minima (Moles *et al.*, 2003). Their performance also suffers in the presence of large measurement errors (Lillacci and Khammash, 2010). This is compounded as few optimization techniques are

*To whom correspondence should be addressed.

able to take into consideration the system noise (Wilkinson, 2007).

An alternative to traditional optimization methods is to use the Bayesian inference method. This method takes both the system and measurement noise into consideration when extracting information from noisy or uncertain data (Wilkinson, 2007). The major advantage is its ability to infer the full probability distribution of the states and the parameters instead of making a simple point estimate. This method tries to estimate the posterior distribution of the parameter θ conditioned on the observed data y, $p(\theta|y)$. However, the calculation of this posterior density involves high-dimensional integration for which analytical expressions are generally not available (Barnes *et al*., 2011). Their numerical solution might also be computationally challenging as the problem is translated to a combinatorial summation problem (Lillacci and Khammash, 2010). The Markov chain Monte Carlo (MCMC) (Brooks, 1998) method, which works with batch data, has the capability to represent the posterior distribution numerically (Barnes *et al*., 2011). Recently, MCMC techniques have been widely used in biological systems for predicting different targets (Barenco *et al*., 2006; Jayawardhana *et al*., 2008; Vyshemirsky and Girolami, 2008). Sequential approaches such as the sequential Monte Carlo (SMC), also known as particle filters (Doucet *et al*., 2001; Lang *et al*., 2007), and Kalman filtering type approaches (Kalman, 1960; Julier and Uhlmann, 1997) were primarily developed for data produced in a sequential fashion. These methods yield better approximations compared with the batch processing algorithms when applied to the whole data (Liu and Niranjan, 2012). There has been extensive research in recent years to apply the sequential approximation methods to estimate states and parameters in biological systems. The work of Liu and Niranjan (2012); Mahsuni and Haris (2009); Nakamura *et al*. (2009) demonstrate the applicability of the SMC in the state estimation of biological models. Although SMC or particle filters are more sophisticated than the Kalman filtering methods, they involve calculation of several hyper parameters and have to solve numerous sets of ODEs, which leads to a high computational cost (Quach *et al*., 2007). Nakamura *et al*. (2009) reported that parameter estimation using particle filtering for a regular-size biological model requires the cost of petascale computing. Such large-scale computing can only be performed with high-end clusters and supercomputers. In that respect, Kalman filtering methods are much more suitable for biological systems, as they are less resource intensive in terms of computational cost. To date, efforts have been made to apply Kalman filtering–based methods to estimate parameters of biological systems (Lillacci and Khammash, 2010; Lillaci and Valigi, 2007; Sisson *et al*., 2007; Sun *et al*., 2008; Quach *et al*., 2007; Zeng *et al*., 2011 to name a few).

In this article, we propose a new form of the unscented Kalman filter for the estimation of unknown parameter values of biological systems within a constrained boundary. Unlike other non-linear extensions to the Kalman filter (KF), this algorithm remains numerically stable. We illustrate the efficiency of this method by estimating the unknown parameters and hidden states of two different systems. The first system is the upper part of Glycolysis in yeast (Klipp *et al*., 2005), modified from Hynne *et al*. (2001). With this, we study the parameter estimation of a medium-size kinetic model. The second model is from the Dream6 contest on the estimation of model parameter challenge (Dream6, 2012). The result from the parameter estimation of these models clearly demonstrates the applicability of our newly proposed algorithm to much larger and real-life biological models.

2 MODELLING BIOLOGICAL NETWORKS AS STATE-SPACE MODELS

To apply Kalman filtering for parameter estimation in biological models, the system must first be formulated as a non-linear state-space model (Quach *et al*., 2007). The state-space description uses a set of first-order differential equations to provide a convenient and powerful representation of the non-linear dynamics of a system. This representation consists of state variables and observed variables as well as their different components and interactions. State variables represent the total state of the system at any given time. Generally some state variables are not directly observable (hidden states). The observed variables are those that can be directly measured. The general form of the non-linear state-space model is

$$\begin{aligned} \dot{x} &= F(x, \theta, u, t) + w, x(t_0) = x_0 \\ y &= H(x, \theta, t) + v \end{aligned} \tag{1}$$

The evolution of these state variables, $x = [x_1, x_2, \cdots, x_n]$, over time is defined by the state equation, F, which also depends on the model parameters $\theta = [\theta_1, \theta_2, \cdots, \theta_m]$, starting from x_0 at time t_0 (Quach *et al*., 2007). For a biological system, this state vector x might represent the change of concentration of biochemical species like metabolite or mRNA. The state equation may also be influenced by external input, $u = [u_1, u_2, \cdots, u_p]$. The observation equation, H, relates the state variables to the output variables, $y = [y_1, y_2, \cdots, y_k]$, describing how the state variables are observed through their measurements. The process noise w is uncorrelated Gaussian white noise with probability distribution, $p(w) \sim N(0, Q)$. This is an extrinsic noise describing the amount of confidence we have in our model. The measurement noise, v, explains the reliability of the measurement data and is also uncorrelated Gaussian white noise with probability distribution, $p(v) \sim N(0, R)$.

2.1 Parameter estimation in non-linear state space

Parameter estimation of kinetic models represented with state-space models can be addressed with state observers by extending the state-space definition to use an augmented state vector $x^{aug} = [x \quad \theta]$ (Sitz *et al*., 2002). This augmented state vector is constructed by considering the parameters, θ as additional state variables with a zero rate of change, i.e. constants. In this manner, the parameters are considered as constant functions of time instead of as mere constant numbers (Lillacci and Khammash, 2010). The augmented state-space equations for parameter estimation are

$$\begin{aligned} \dot{x} &= F(x, \theta, t) + w, x(t_0) = x_0 \\ \dot{\theta} &= 0 \qquad\qquad\quad , \theta(t_0) = \theta_0 \end{aligned} \tag{2}$$

For notational simplicity, we will use x to denote x^{aug}.

2.2 Deriving non-linear state space from ODEs

Dynamics of the biological systems are characterized by a set of ODEs, where the state evolution is described as

$$\dot{x} = F(x, u, t) + w, x(t_0) = x_0 \tag{3}$$

There are a number of ways to link this ODE with the non-linear state-space modelling. One issue to consider is that, although the system is a continuous time process, it is observed only at discrete times. Sitz *et al*. (2002) solved this problem by numerically integrating the state dynamics between the temporal time points at which the state is being observed, using the parameter values estimated at the previous iteration:

$$\begin{aligned} f(x(k), u) &= x(k) + \int_{t_k}^{t_{k+1}} F(x(\tau), u)d\tau \\ x(k+1) &= f(x(k), u) + w(k) \end{aligned} \tag{4}$$

2.3 Parameter estimation with filtering techniques

In this article, we apply the Bayesian framework to infer the hidden states and parameters of a biological system. We seek to calculate the joint probability distribution of the parameters and states, given the measurements, denoted as $p(\theta, x(k)|y(k))$. The algorithm we propose is derived from the sequential estimation procedure based on filtering.

The filtering technique sequentially calculates the posterior probability of the state, $p(x(k)|y(k))$, given the measurement data upto the k-th time step. The complete calculation consists of a sequential and alternative calculation of the state probability at time k having the measurement data upto time step $k-1$, called the *prediction step*

$$p(x(k)|y(k-1)) = \int p(x(k)|x(k-1))p(x(k-1)|y(k-1))dx(k-1) \tag{5}$$

and the calculation of the state probability when the measurement data is available at the k-th time step is called the *correction step*.

$$p(x(k)|y(k)) = \frac{p(y(k)|x(k))p(x(k)|y(k-1))}{p(y(k)|y(k-1))} \tag{6}$$

At the start, the prediction step $p(x(0)|y(0:-1))$ is defined as the density of the initial state vector, $(x(0))$.

The Kalman filter optimally calculates the posterior mean and covariance when f and h are linear Markov processes. This algorithm is invalid if f and h are non-linear. Several extensions of Kalman filter has been developed to work under non-linearity such as the extended Kalman filtering (EKF) (van der Merwe, 2004) and unscented Kalman filtering (UKF) (Julier and Uhlmann, 1996; Merwe and Wan, 2001). Among the different non-linear extensions, UKF has demonstrated a much higher accuracy in the state estimation of dynamic systems (Julier and Uhlmann, 2004).

The key idea in UKF is the 'unscented transformation' where a set of points called sigma-points are selected deterministically to capture the probabilistic spread of the state variables. These sigma–points are then propagated through the process function f. An approximation of the transformed mean $x(k|\hat{k}-1)$ and covariance $P(k|k-1)$ are then calculated from these sigma–points. At the correction step, the final state estimation $x(k)$ uses the classical Kalman filtering approach. The complete algorithm for UKF is given in Supplement 6.

A major drawback when applying the UKF to biological systems is that it becomes numerically unstable when the estimation covariance matrix is not positive definite (Merwe and Wan, 2001). The square-root implementation of UKF solves the problem of numerical instability by propagating the square root of the covariance matrix at each iteration (Merwe and Wan, 2001).

3 CONSTRAINED SQUARE-ROOT UNSCENTED KALMAN FILTER

Constraints play a major role in biological systems, as they include previously known information about the parameter ranges. As biological parameters vary within a biophysically plausible range, the inclusion of constraints makes the state estimation possible within a biologically meaningful space. One method for introducing constraints into the KF is the moving horizon estimator. However, this approach requires the solution of a non-recursive quadratic system at each step (Teixeira *et al*., 2008). Alternatively, the interval constrained unscented transformation (ICUT) proposed by Teixeira *et al*. (2008) introduces the constraints at each of the sigma–points individually. This constrains the sigma–points and the mean value to stay within the given boundaries. That is, if a sigma point falls outside of the boundary, it is projected onto the boundary. The latter approach is a more efficient method for introducing the state inequality constraints into the state estimation process (Teixeira *et al*., 2008).

Applying ICUT directly to UKF would ensure a constrained state estimation, and the instability of this technique would still pose a problem during the estimation. The filtering technique proposed in this article includes state inequality constraints by modifying the square-root unscented Kalman filter. This new filtering technique, the constrained square-root unscented Kalman filter (CSUKF), is both numerically stable and can estimate parameters within a biologically plausible parameter space.

In CSUKF, we define the state constraints for an n-dimensional state vector at iteration k, $x(k) \in \mathbb{R}^n$, with an n-dimensional hypercube according to

$$L(k) \leq x(k) \leq U(k) \tag{7}$$

Where for iteration k, $L(k) \in \mathbb{R}^n$ is the vector of lower bounds and $U(k) \in \mathbb{R}^n$ is the vector of upper bounds. (Although typically not the case for biological systems, the boundaries of each sigma-point may vary with each iteration.) For an unbounded region, the limits are set to infinity.

3.1 Selection of sigma-points

The direction of the sigma-points are defined through the matrix $S = [+\sqrt{P} \quad -\sqrt{P}]$, where P is the current estimate of the covariance matrix. This allows the positive and negative square roots to be handled separately. The step size vector, ζ, is then defined as

$$\zeta_j \triangleq min(col_j(\Theta)) \tag{8}$$

$$\Theta_{i,j} \triangleq \begin{cases} \sqrt{n+\lambda} & \text{if } S_{i,j}=0 \\ min(\sqrt{n+\lambda}, \frac{U_i(k)-\hat{x}_i(k-1)}{S_{i,j}}) & \text{if } S_{i,j}>0 \\ min(\sqrt{n+\lambda}, \frac{L_i(k)-\hat{x}_i(k-1)}{S_{i,j}}) & \text{if } S_{i,j}<0 \end{cases}$$

where $\lambda = \alpha^2(n+\kappa) - n$ is a scaling parameter. The other scaling parameters α and κ are described in Supplement 1. When none of the sigma-points are outside of the boundary, the step size is $\sqrt{n+\lambda}$, the regular step size for the scaled UT (Julier and Industries, 2002). However, if any of the sigma-points violate the boundary, then the step size is set to the scaled distance to the nearest boundary. These result in the constrained sigma-point matrix, lying either on or inside the boundary, and defined by

$$\mathcal{X} = \begin{cases} \hat{x}(k-1) & ,j=0 \\ \hat{x}(k-1) + \zeta_j col_j(S) & ,1 \le j \le n \end{cases} \quad (9)$$

The constrained sigma-points may not be symmetric. In this case, we need to adjust the weights associated with these non-symmetric sigma-points. The weights, W^m for the mean and W^c for the covariance, are calculated such that when sigma-points do not violate the constraints, the regular weights are selected. However, if the sigma-points are propagated onto the boundary, the weights are varied linearly with the step size. Supplement 7 illustrates the sigma-point selection, and the detailed calculation of the weights is given with the description of the algorithm.

3.2 Formulation of CSUKF

For the implementation of CSUKF, we applied the same techniques as SR-UKF, namely QR decomposition, Cholesky factor updating and pivot-based least squares. The Cholesky factor, $V(k)$, of the compound matrix formed by the positively weighted sigma-points and the matrix square root of the process noise are calculated using the QR decomposition. A similar approach is used to update the Cholesky factor of the measurement update covariance matrix. The Cholesky factor update is used to downdate the Cholesky factor for the negatively weighted sigma-points. Finally, the pivot-based least squares are used to calculate the Kalman gain $K(k)$ (see Supplement 1 for a detailed description). By adapting the method from the square-root UKF, $V(k)$ is propagated at each iteration instead of the full covariance matrix P, eliminating the need to explicitly calculate this square root at each iteration.

We start the CSUKF by initializing the state vector $\hat{x}(0) = E[x(0)]$ and the state covariance matrix $P(0) = E[(x(0) - \hat{x}(0))(x(0) - \hat{x}(0))]$, and then calculating its square root, $V(0)$, via Cholesky factorization. As the weights in the constrained sigma-point calculation are asymmetric, they may vary in magnitude and sign, i.e. they can be positive or negative. Thus the calculation of the square-root factor of the covariance matrix needs to be decomposed into two parts, one for the positive weights of W^C denoted as $V_{pos}(k)$ and one for the negative weights of W^C denoted as $V_{neg}(k)$. This propagated Cholesky factor is used to calculate the sigma-points, $\mathcal{X}$. At each iteration, they are calculated as

$$V_{pos}(k) = \left[\sqrt{W_j^c}(\mathcal{X}_j(k|k-1) - \hat{x}(k)) \quad Q\right]_{j \in I^+} \quad (10)$$

$$V_{neg}(k) = \left[\sqrt{|W_j^c|}(\mathcal{X}_j(k|k-1) - \hat{x}(k))\right]_{j \in I^-} \quad (11)$$

where, $I^+ = \{j | W_j^C \ge 0\}$ and $I^- = \{j | W_j^C < 0\}$ are used as a helper index to identify the positive and negative weights, respectively. The covariance matrix P can be recovered from V as

$$P(k) = V_{pos}(k)(V_{pos}(k))^T - V_{neg}(k)(V_{neg}(k))^T$$

QR decomposition was used to express V_{pos} in terms of an orthogonal matrix O and an upper triangular matrix G

$$V_{pos}(k) = O(k)(G(k))^T \quad (12)$$

The state covariance matrix then becomes

$$\begin{aligned} P(k) &= G(k)(O(k))^T O(k)(G(k))^T - V_{neg}(k)(V_{neg}(k))^T \\ &= G(k)(G(k))^T - V_{neg}(k)(V_{neg}(k))^T \end{aligned} \quad (13)$$

where, $(O(k))^T(O(k)) = I$. Thus, the state covariance matrix can be formed from the upper triangular matrix generated during QR decomposition, i.e. $G(k) = qr(V_{pos}(k))$. Supplement 4 explains the recovery of the covariance matrix in detail.

A rank 1 downgrade to Cholesky factorization is performed to include the effect of $V_{neg}(k)$ in the calculation of the square root. This approach is applied to update the square-root factor in both the time and measurement update steps. The final step for each iteration is the calculation of the state estimation and the square-root factor of the estimation covariance. We next discuss the complete algorithm of CSUKF.

3.3 Initialization

To start the filter, we need to first initialize the states and the error covariance matrix. As we do not have any measurement data to estimate $x(0)$, we set the initial state estimate to its expected value.

$$\hat{x}(0) = \mathbb{E}[x(0)] \quad (14)$$

It follows the calculation of the error covariance matrix as $P(0) = \mathbb{E}[(x(0) - \hat{x}(0))(x(0) - \hat{x}(0))^T]$. In CSUKF, we only propagate the triangular form of the square root of the error covariance matrix. This square root is calculated at the initialization step using the Cholesky decomposition.

$$V(0) = chol\{P(0)\} \quad (15)$$

We iterate the next steps for the whole time-series data from $k=1$ to $k=T$.

3.4 Time update (prediction stage)

The time update equations are first applied to obtain the current a priori state estimate, which we denote as $x^-(k)$. This calculation is based on the process function f, considering all the data available upto step $k-1$, $y(1:k-1)$. To calculate this $x^-(k)$, we need to first generate $2n+1$ sigma-points. These sigma-points, $\mathcal{X}(k|k-1)$, are calculated based on the state estimate made at $(k-1)^{th}$ time step, $\hat{x}(k-1)$ and the Cholesky factor of the error

covariance matrix $V(k-1)$. The sigma-points satisfying $L \le \mathcal{X}(k|k-1) \le U$ are then selected as

$$\mathcal{X}(k|k-1) = \begin{cases} \hat{x}(k-1) & ,j=0 \\ \hat{x}(k-1) + \zeta_j col_j(S) & ,1 \le j \le 2n \end{cases} \quad (16)$$

where $\mathcal{X}$ is based on the direction, $S = [V(k-1) \quad -V(k-1)]$. ζ is the step size, which is calculated as

$$\zeta_j \triangleq min(col_j(\Theta)) \quad (17)$$

$$\Theta_{i,j} \triangleq \begin{cases} \sqrt{n+\lambda} & \text{if } S_{i,j} = 0 \\ min(\sqrt{n+\lambda}, \frac{U_i(k) - \hat{x}_i(k-1)}{S_{i,j}}) & \text{if } S_{i,j} > 0 \\ min(\sqrt{n+\lambda}, \frac{L_i(k) - \hat{x}_i(k-1)}{S_{i,j}}) & \text{if } S_{i,j} < 0 \end{cases}$$

With the adjustment of the sigma-points, the weights are to be adjusted as well. These weights vary linearly with the step size of the sigma-points. They are calculated as

$$W_0^m = b \quad (18)$$

$$W_0^c = b + (1 - \alpha^2 + \beta) \quad (19)$$

$$W_j^m = W_j^c = a\zeta_j + b, \quad 1 \le j \le 2n \quad (20)$$

such that $\sum W = 1$, as defined in the general weighting scheme of UT (Julier and Uhlmann, 1996). Solving for a and b then yields

$$a = \frac{n}{(n+\lambda)} \frac{1-2\lambda}{\sum_{j=1}^{2n} \zeta_j}$$

$$b = \frac{\lambda}{n+\lambda}$$

The detailed calculation of a and b is given in Supplement 5. These sigma-points are transformed through the non-linear state function.

$$\mathcal{X}^x(k|k-1) = f[\mathcal{X}(k|k-1)] \quad (21)$$

These transformed sigma-points are then used to calculate the a priori mean and covariance of the estimation.

$$\hat{x}^-(k) = \mathcal{X}^x(k|k-1)(W^m)^T \quad (22)$$

$$G^x(k) = qr\{\left[\sqrt{W_j^c}(\mathcal{X}_j^x(k|k-1) - \hat{x}^-(k)) \quad \sqrt{Q}\right]_{j \in I^+}\} \quad (23)$$

$$V_{neg}^x(k) = \left[\sqrt{|W_j^c|}(\mathcal{X}_j^x(k|k-1) - \hat{x}^-(k))\right]_{j \in I^-} \quad (24)$$

The prior cholesky factor for the time update, $V^{x-}(k)$, is a downdate of the positive and negative square roots

$$V^{x-}(k) = cholupdate(G^x(k), V_{neg}^x(k), '-')$$

3.5 Measurement update (correction stage)

The correction stage updates the a priori state estimate $\hat{x}^-(k)$ by taking the current measurement data, $y(k)$, into consideration. At this stage, sigma-points are redrawn to incorporate the measurement noise, R

$$\mathcal{X}(k|k-1) = [\hat{x}^-(k) \quad \hat{x}^-(k) \pm \sqrt{(n+\lambda)}V^{x-}(k)]$$

These sigma-points are now propagated through the observation function h. The same approach is applied to calculate the cholesky factor for the measurement covariance, $V^{y-}(k)$.

$$\mathcal{X}^y(k|k-1) = h[\mathcal{X}(k|k-1)] \quad (25)$$

$$\hat{y}^-(k) = \mathcal{X}^y(k|k-1)(W^m)^T \quad (26)$$

$$G^y(k) = qr\left\{\left[\sqrt{W_j^c}(\mathcal{X}_j^y(k|k-1) - \hat{y}^-(k)) \quad \sqrt{R}\right]_{j \in I^+}\right\} \quad (27)$$

$$V_{neg}^y(k) = \left[\sqrt{|W_j^c|}(\mathcal{X}_j^y(k|k-1) - \hat{y}^-(k))\right]_{j \in I^-} \quad (28)$$

$$V^{y-}(k) = cholupdate(G^y(k), V_{neg}^y(k), '-') \quad (29)$$

3.6 State and covariance estimation

The same general formulation used in the UKF for the final state and covariance estimation is applied. For that we need to first calculate the cross-correlation covariance matrix,

$$P^{xy}(k) = \sum_{j=0}^{2n} W_j^c[\mathcal{X}_j^x(k|k-1) - \hat{x}^-(k)][\mathcal{X}_j^y(k|k-1) - \hat{y}^-(k)]^T \quad (30)$$

The Kalman gain, which describes the relative certainty of the current state estimate and the measurement, is calculated as

$$K(k) = P^{xy}(k)/(V^{y-}(k))^T V^{y-}(k) \quad (31)$$

For measurement data $y_{meas}(k)$ at iteration k, the final posterior state estimation is

$$\hat{x}(k) = \hat{x}^-(k) + K(k)(y_{meas}(k) - \hat{y}^-(k)) \quad (32)$$

The square-root factor of the estimation covariance matrix $V(k)$ is then updated

$$V(k) = cholupdate\{V^{x-}(k), K(k)V^{y-}(k), '-'\} \quad (33)$$

If at the end of the measurement update stage, any state value violates the constraints, they are projected back onto the constraint boundary before the start of the next iteration.

The choice of probability density function in CSUKF ultimately depends on the variance; as the mean of the non-negative state approaches zero, the distribution either remains Gaussian or if sufficiently large, may be better described by a Weibull distribution (which is one sided, i.e. 0 for $x<0$). The Weibull distribution with parameters $\lambda \approx mean$ and $k>2$ has been shown to approximate the Gaussian distribution (at $k=2$, it is equated with the Rayleigh distribution and moves towards a Dirac delta centered at $x=\lambda$ as k approaches infinity). As such, we feel the trade off of this simplified approach using the Gaussian pdf is appropriate both theoretically and, moreover, has held up in the actual application of the algorithm when applied to parameter estimation problems.

Table 1. Comparison of Kalman filter extensions; EKF, UKF and CSUKF calculated from 50 repetitions of each algorithm with the parameters initialized to random values between zero and one. The standard deviation (SD) is shown in the parentheses next to the mean value

Parameter name	Actual value	EKF	UKF	CSUKF
		Mean (SD)	Mean (SD)	Mean (SD)
k_2	2.26	3.58 (0.03)	2.26 (0.03)	2.27 (0.08)
$V^f_{max,3}$	140.28	92.65 (0.21)	140.31 (0.45)	140.24 (0.68)
$V_{max,4}$	44.73	36.13 (0.14)	44.71 (0.16)	44.73 (0.28)
$k_{8,r}$	133.33	120.05 (0.34)	133.33 (0.32)	133.33 (0.38)

In the UKF, the sigma-points are symmetric. When calculating the Taylor series accuracy, these symmetric sigma-points result in all odd order moments being zero Julier and Uhlmann (2004). In CSUKF, when sigma-points violate the constraints, they are projected back onto the feasible space and may result in asymmetric sigma-points. To preserve the Gaussian distribution of the state variables, the weights (W^m and W^c) are adjusted to compensate for the projected sigma-points (Supplement 4). Thus, the weighted symmetric sigma-points from this modified distribution maintain the same order of accuracy for the CSUKF. This preservation of accuracy is illustrated by the comparison of UKF and CSUKF results, as contrasted with those of EKF (Table 1).

4 RESULTS AND DISCUSSION

4.1 Glycolysis model from yeast

The impetus of this work is to develop an algorithm that can estimate unknown parameter values for kinetic models of biological systems. To test the algorithm, the kinetic model from the upper part of glycolysis in yeast (Klipp *et al.*, 2005), modified from Hynne *et al.* (2001), was chosen. This model, shown in Figure 1, consists of the first four reactions from the upper part of glycolysis. This model explains the degradation of glucose in the process of yielding energy. The model has 15 parameters of which four parameters, k_2, $V^f_{max,3}$, $V_{max,4}$ and k_{8r}, were selected to be estimated. These parameters were chosen to cover a large range of values, from 2.26 (k_2) to 140.28 ($V^f_{max,3}$). The other two parameters are $V_{max,4} = 44.73$ and $k_{8,r} = 133.33$. Parameters can not have negative value so we set the lower bound for all parameters as $L(x) \geq 0$. As there is no upper bound for the parameters, this was set to infinity $U(x) \leq \infty$. Synthetic measurement data y_{meas} with uncorrelated white noise was generated to validate the model.

Results: Figure 2 shows the estimation trajectory of the four parameters from a single run of the CSUKF. The dashed line represents the actual parameter value and the solid line represents the estimation trajectory of the parameters. The CSUKF estimations quickly approach the actual parameter values, regardless of their magnitude. Despite the noisy data, the parameter fitting continues to improve over time as does the accuracy of the estimate. This is again consistent with the observed behaviour of the original UKF. The same parameters were then

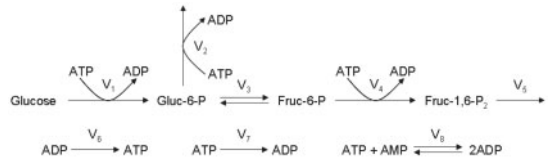

Fig. 1. Simplified glycolysis model adapted from Klipp *et al.* (2005). The fluxes are denoted v_i. Each flux is represented with specific rate laws depending on their enzymatic reactions. Details of the rate laws can be found in the Supplement 2

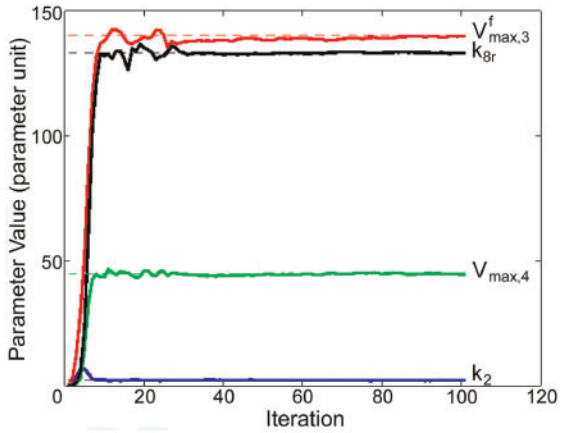

Fig. 2. Parameter estimation trajectory from CSUKF. The actual values of the four unknown parameters are shown with dashed lines, and the recursive estimations are shown with solid lines. The sample time is $\Delta t = 0.25$ corresponding to 100 data points

estimated with the EKF and UKF, providing an additional set of comparative values. The summary statistics from 50 runs (with random initial state values between zero and one) are listed in Table 1. The CSUKF again performs similarly to the UKF, both of which yield accurate and precise estimations. In this result, the UKF had marginally lower standard deviations compared with the CSUKF, indicating a slightly better estimation accuracy. However, it suffers from numerical instability when initialized to higher values as shown in (Supplementary Table S4). To determine the effect that the choice of initial parameter values has on each non-linear KF variant, 50 sets of randomly selected starting values were generated for each parameter in the range of [0 1] and a second 50 in the range of the actual parameter values, [0 140] (Supplementary Table S2). Supplementary Table S1 presents the estimations from running each algorithm with each set of initial parameter values. One major difference was that for this model, the UKF became numerically unstable, when initialized with higher values (Supplementary Table S4), failing 68% of the time. The CSUKF was found to be independent of the starting values yielding 100% solutions, with only one outlier (Supplementary Table S3). Furthermore, without constraints, the UKF periodically considered negative parameter values to be valid. Both UKF and CSUKF provide more accurate estimates than EKF while requiring <3% of the processing time, which is consistent with previous comparisons (Kandepu *et al.*, 2007; Merwe and Wan, 2001). Using the MATLAB profiler, the bottleneck in EKF was found to be the calculation of the Jacobian, required for linearization. EKF was also found to

be most dependent on the initial values, with the estimation result showing significant variation in all parameters when starting in the extended range. In the limited starting range of [0 1], the EKF produced poor but consistent results (Supplementary Table S5). Analysing these results, CSUKF was found to have the most consistent and accurate results of these three KF extensions. Additional comparisons were made with four probabilistic global optimization algorithm, evolutionary programming (EP), genetic algorithm (GA), simulated annealing (SA) and particle swarm optimization (PSO). Supplementary Table S1 summarizes these results. For this comparison, the model was implemented in Copasi 4.8 with the default settings used for each of these algorithms. To summarize, the overall estimation performance of CSUKF exceeded that of the other algorithms, both in terms of runtime and accuracy (Supplementary Table S1).

4.2 Gene regulatory network

To verify the applicability and accuracy of CSUKF, it was applied to estimate the parameters of a gene regulatory network. This model was part of the Dream6 estimation of model parameters challenge. The network structure of the model and the corresponding rate laws are given in Supplement 3. This model follows linear kinetics for mRNA degradation and protein synthesis and degradation. Hill-type kinetics were assumed for the activation and repression. Protein production was modelled in two steps, the transcription step and the translation step. Considering the two Hill kinetic parameters, θ_h, θ_{kd}, and the mRNA synthesis rate $\theta_{mRNA,\,syn}$, the transcription can be described as

$$v_{mRNA3} = \theta_{mRNA3,\,syn} \cdot \frac{(\frac{p1}{\theta_{3kd}})^{\theta_{3h}}}{(1+\frac{p1}{\theta_{3kd}})^{\theta_{3h}}} \cdot \frac{1}{1+(\frac{p2}{\theta_{4kd}})^{\theta_{4h}}} - \theta_{mRNA,\,deg} \cdot mRNA3 \qquad (34)$$

The network topology and reaction rates were given by the organizers. Initially, the contest provides a limited amount of data representing the time-course of mRNA concentration for the wild-type variety. To reflect an actual scientific scenario, additional datasets can be purchased by spending a virtual budget allocated to each team at the time of registration. The model has 30 parameters among which 29 are unknown and, therefore, have to be estimated. For this experiment, we additionally bought the protein concentration of the wild type and also the mRNA and protein concentration for the mutant with increase of RBS4 activity by 100%. All time-series data are over the interval 0 to 20 s with a 1 s step size. We set the constraints as $10^{-6} \leq \theta \leq 10^{6}$. The experiment was divided into two phases. In the first phase, we used the mutant data to estimate the mean and covariance of the parameters. During this step, the states were initialized with small random numbers. Similarly, the process noise covariance matrix, Q, was also initialized with small random numbers. The measurement noise covariance matrix R was initialized with the noise model supplied by the organizers, which is $y_{noisy} = max[0, y + 0.1 + 0.2 \cdot y]$. The estimated values from phase 1 are used to formulate the prior distribution of the parameters for the second phase of the experiment where the data from the wild type was used. Figure 3 compares the final estimation result of the parameters with the actual values (which were made available at the end of the contest). Supplementary Table S6 describes the numerical value of the estimation along with the standard deviation. As can be seen from Figure 3 and Supplementary Table S6, CSUKF can estimate most of the parameters with a reasonably high accuracy for this model. A comparative result of the parameter estimation for this network with GA, EP, SA and EKF along with CSUKF are given in Supplementary Table S7. One point to note from Supplementary Table S6 is that parameters having high standard deviation deviates more from the actual value compared with the other parameters, indicating that some of the parameters might be non-identifiable. However, non-identifiability is not considered under the scope of this article. For a more in-depth view of the estimation, state trajectory of the metabolites with the estimated parameters and actual parameters, a full set of plots are provided in Supplement 8.

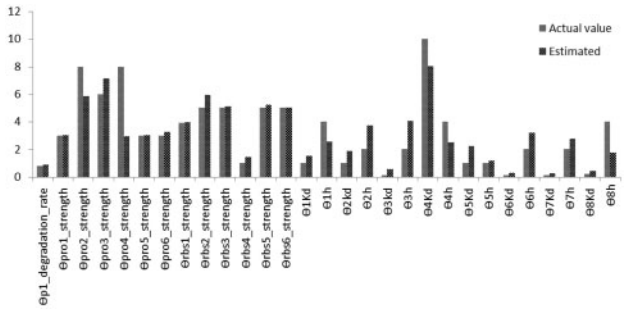

Fig. 3. Parameter estimation trajectory from CSUKF. The actual values of the four unknown parameters are shown with dashed lines, and the recursive estimations are shown with solid lines. The sample time is $\Delta t = 0.25$ corresponding to 100 data points

5 CONCLUSION

In systems biology, the modelling of biochemical networks require the estimation of mechanistic parameters that cannot be determined directly from measurements. These parameters are inferred from other measured quantities. Compounding the problem is the fact that these measurements are inherently noisy, the networks are non-linear and traditional optimization space is multimodal.

In this article, we describe a novel filtering method, the CSUKF, which provides a powerful yet computationally cost-effective method to tackle complex parameter inference problems for biological models. While filtering approaches are generally of use in applications suffering from noisy measurement data, the constrained parameter estimation and insensitivity to initial conditions combine to make CSUKF particularly well suited to biological applications. To illustrate this point, the CSUKF was shown to successfully estimate the parameters of a metabolic network (the Glycolysis model) and a gene regulatory network (the Dream6 challenge). CSUKF systematically propagates uncertainty while exploring the parameter space, giving the estimation covariance matrix along with the parameter approximations. Thus this approach makes it easier to conduct direct statistical analysis over the estimation. Furthermore, the CSUKF provides this enhanced parameter estimation while ensuring numerical stability with the same runtime complexity of the less-stable unscented Kalman filter.

The CSUKF has been specifically designed for the needs of systems biology; however, these needs are not unique to biological modelling. Thus the CSUKF can be generally applied to any filtering or state-space problem.

ACKNOWLEDGEMENT

We would like to thank Dr Christian Krach and Dr Jan Huege for their fruitful discussion on the modelling approaches.

Funding: This work was supported by the German Federal Ministry of Education and Research (BMBF) (grant 0315295).

Conflict of Interest: none declared.

REFERENCES

Arisi,I. *et al.* (2006) Parameter estimate of signal transduction pathways. *BMC Neurosci.*, **7** (**Suppl. 1**), S6.

Barenco,M. *et al.* (2006) Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biol.*, **7**, R25.

Barnes,C.P. *et al.* (2011) Bayesian design strategies for synthetic biology. *Interface Focus*, **1**, 895–908.

Brooks,D.J. (1998) Bayesian methods in bioinformatics and computational systems biology. *The Statistician*, **47**, 69–100.

Doucet,A. *et al.* (2001) *Sequential Monte Carlo Methods in Practice*. Springer, USA.

Dream6 (2012) *Estimation of Model Parameters Challenge*. Dialogue for Reverse Engineering Assessments and Methods. http://www.the-dream-project.org/ (6 March 2013, date last accessed).

Hynne,F. *et al.* (2001) Full-scale model of glycolysis in Saccharomyces cerevisiae. *Biophys. Chem.*, **94**, 121–163.

Jayawardhana,B. *et al.* (2008) Bayesian inference of the sites of perturbations in metabolic pathways via Markov chain Monte Carlo. *Bioinformatics*, **24**, 1191–1197.

Julier,S.J. and Industries,I. (2002) The scaled unscented transformation. In *Proceedings of IEEE American Control Conference*, pp. 4555–4559.

Julier,S.J. and Uhlmann,J.K. (1996) A general method for approximating nonlinear transformations of probability distributions. In *Technical report*. Robotics Research Group, Department of Engineering Science, University of Oxford.

Julier,S.J. and Uhlmann,J.K. (1997) A new extension of the Kalman filter to nonlinear systems. In *International Symposium on Aerospace/Defense Sensing. Simulation and Controls*, *Orlando*, *FL*, pp. 182–193.

Julier,S.J. and Uhlmann,J.K. (2004) Unscented filtering and nonlinear estimation. *Proc. IEEE*, **92**, 401–422.

Kalman,R.E. (1960) A new approach to linear filtering and prediction problems. *Trans. ASME, J. Basic Eng.*, **82** (Series D), 35–45.

Kandepu,R. *et al.* (2007) Applying the unscented Kalman filter for nonlinear state estimation. *J. Process Control*, **18**, 753–768.

Kirkpatrick,S. *et al.* (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Klipp,E. *et al.* (2005) *Systems Biology in Practice: Concepts, Implementation and Application*. Wiley-VCH, Weinheim, Germany.

Lang,L. *et al.* (2007) Bayesian estimation via sequential Monte Carlo sampling - constrained dynamic systems. *Automatica*, **43**, 1615–1622.

Lillaci,G. and Valigi,P. (2007) State observers for the estimation of mRNA and protein dynamics. In *Life Science Systems and Applications Workshop, 2007. LISA 2007. IEEE/NIH*, pp. 108–111.

Lillacci,G. and Khammash,M. (2010) Parameter estimation and model selection in computational biology. *PLoS Comput. Biol.*, **6**, e1000696.

Liu,X. and Niranjan,M. (2012) State and parameter estimation of the heat shock response system using Kalman and particle filters. *Bioinformatics*, **28**, 1501–1507.

Mahsuni,G. and Haris,V. (2009) A particle filtering algorithm for parameter estimation in real-time biosensor arrays. In *IEEE International Workshop on Genomic Signal Processing and Statistics*. Minneapolis, USA.

Mendes,P. and Kell,D. (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, **14**, 869–883.

Merwe,R.V.D. and Wan,E.A. (2001) The square-root unscented Kalman filter for state and parameter-estimation. In *International Conference on Acoustics, Speech, and Signal Processing*. Salt Lake city, USA, pp. 3461–3464.

Moles,C.G. *et al.* (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.*, **13**, 2467–2474.

Nakamura,K. *et al.* (2009) Parameter estimation of in silico biological pathways with particle filtering towards a petascale computing. In *Pacific Symposium on Biocomputing*. Hawaii, pp. 227–238.

Quach,M. *et al.* (2007) Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. *Bioinformatics*, **23**, 3209–3216.

Sisson,S.A. *et al.* (2007) Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, **104**, 1760–1765.

Sitz,A. *et al.* (2002) Estimation of parameters and unobserved components for nonlinear systems from noisy time series. *Phys. Rev. E*, **66**, 16210.

Stelling,J. (2004) Mathematical models in microbial systems biology. *Curr. Opin. Microbiol.*, **7**, 513–518.

Sun,X. *et al.* (2008) Extended Kalman filter for estimation of parameters in nonlinear state-space models of biochemical networks. *PLoS One*, **3**, e3758.

Teixeira,B.O.S. *et al.* (2008) Unscented filtering for interval-constrained nonlinear systems. In *47th IEEE Conference on Decision and Control*, Cancun, Mexico, pp. 5116–5121.

van der Merwe,R. (2004) Sigma-point Kalman filters for probabilistic inference in dynamic state-space models. In Ph.D. thesis, OGI School of Science and Engineering, Oregon Health and Science University, Oregon, USA.

Vyshemirsky,V. and Girolami,M. (2008) Biobayes: a software package for Bayesian inference in systems biology. *Bioinformatics*, **24**, 1933–1934.

Wilkinson,S.P. (2007) Markov chain Monte Carlo methods and its application. *Brief. Bioinform.*, **8**, 109–116.

Zeng,N. *et al.* (2011) Inference of nonlinear state-space models for sandwich-type lateral flow immunoassay using extended Kalman filtering. *IEEE Trans. Biomed. Eng.*, **58**, 1959–1966.