Received: 5 October 2017 | Revised: 15 January 2018 | Accepted: 1 February 2018

DOI: 10.1111/stan.12140

SPECIAL ISSUE ARTICLE

WILEY

Statistical modelling of cell movement

Diana Giurghita | Dirk Husmeier

School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK

Correspondence
Diana Giurghita, School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK.
Email: d.giurghita.1@research.gla.ac.uk

Present Address
School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK.

Funding information
Engineering and Physical Sciences Research Council (EPSRC), Grant/Award Number: EP/N014642/1

Collective cell movement affects vital biological processes in the human organism such as wound healing, immune response, and cancer metastasis. A better understanding of the mechanisms driving cell movement is then essential for the advancement of medical treatments. In this paper, we demonstrate how the unscented Kalman filter, a technique used extensively in engineering in the context of filtering, can be applied to estimate random or directed cell movement. Our proposed model, formulated using stochastic differential equations, is fitted on data describing the movement of *Dictyostelium* cells, an amoeba routinely used as a proxy for eukaryotic cell movement.

KEYWORDS
cell movement, chemotaxis, stochastic differential equations, unscented Kalman filter

1 | INTRODUCTION

Many important biological processes, such as wound healing, tissue development, and cancer cell invasion, are based on the collective movement of cells. One of the main mechanisms for directed cell movement is chemotaxis, where cells follow chemical gradients (chemoattractants) present in their environment. These gradients might arise from the presence of a local source of chemoattractant or due to local depletion of the chemical in the environment (Tweedy, Knecht, Mackay, & Insall, 2016). An example of the former scenario is the migration of breast tumour cells that respond to the epidermal growth factor released by macrophages (Wyckoff et al., 2004). In an attempt to acquire a deeper understanding of the mechanisms behind cell movement, many population-based models have been formulated using partial differential equations, for example, in the form of advection–diffusion equations, with very few of them attempting to fit these models to actual experimental data (Ferguson, Matthiopoulos, Insall, & Husmeier, 2017).

In this paper, we propose a model that describes the movement of any individual cell being driven by an external resource gradient using stochastic differential equations (SDEs) of the following form:

$$dX_t = \sigma dB_t^X \tag{1}$$

$$dY_t = \frac{\alpha\beta \exp\left[-\beta(Y_t - \gamma t)\right]}{\{1 + \exp\left[-\beta(Y_t - \gamma t)\right]\}^2} dt + \sigma dB_t^Y. \tag{2}$$

Equations 1 and 2) describe the evolution in time of the x and y coordinates of a cell in 2D space. σdB_t^X and σdB_t^X are Brownian motion terms, which represent in this model the intrinsic randomness in a cell's movement. The coordinate in the y direction has a drift term that is described by three parameters: α, the amplitude of the resource gradient; β, the steepness of the gradient; and γ, which indicates how fast the gradient changes over time. The strength of the random component in the cell movement equations is indicated by the diffusion coefficient σ.

Inference in nonlinear dynamical systems, such as the one we propose in Equations 1 and 2, poses numerous challenges due to the stochastic nature of the data, intractable likelihoods, and unidentifiable parameters. Recent developments have tackled this problem using likelihood-free methods such as sequential Monte Carlo approximate Bayesian computation (SMC ABC), or computational methods (particle Markov Chain Monte Carlo [MCMC]; Golightly & Wilkinson, 2011); however, these can become too computationally expensive as the number of time points or parameters increases. More specifically, in high-dimensional spaces (either in terms of data or parameters), approximate Bayesian computation (ABC) methods become inefficient due to high rejection rates, which means that a large number of model simulations with proposed parameter values will be required for posterior inference (Beaumont, Zhang, & Balding, 2002). Furthermore, whilst summary statistics can potentially reduce the dimensionality of the problem, selecting appropriate measures is a problem-specific task, meaning that the algorithm needs to be adapted to each application (Aeschbacher, Beaumont, & Futschik, 2012). Lastly, particle MCMC, although an effective and easy-to-generalise approach, is very computationally intensive because a particle filter is run at each iteration of the algorithm, and furthermore, thousands of iterations are normally required (Andrieu, Doucet, & Holenstein, 2010). A great deal of research is currently focused on easing the computational burden by parallelising particle MCMC algorithms (Mingas, Bottolo, & Bouganis, 2017).

State space models (SSMs) are widely used representations of physical processes described as first-order differential equations in terms of inputs, outputs, and parameters. Applications of SSMs typically involve noisy observations in the form of time series representing a latent process, governed by unknown parameters. The problem of interest (estimation in SSMs), also known as "filtering", refers to inferring the latent process and the unknown parameters by making use of the observations in a recursive or "online" form, as observations become available. In simple cases, where the problem can be formulated as a linear Gaussian SSM (LG-SSM), the Kalman filter has been shown to provide an optimal solution to the filtering problem, in a sense that the estimates produced minimise the root mean squared error (MSE) of the parameters. One of the earliest applications of the Kalman filter was in navigation systems and object tracking, such as airplanes, missiles, and spaceships, using noisy radar data. Perhaps the most prominent contribution to the field of aerospace was the use of the Kalman filter for trajectory estimation in the Apollo mission in the 1960s (Grewal & Andrews, 2010). Further applications of the Kalman filter are found in time series forecasting, image and signal processing, sensor data fusion, robotics (Choset & Nagatani, 2001), economics (Bahmani-Oskooee & Brown, 2004), etc.

However, most often, the applications will not meet the linearity or Gaussian assumptions, meaning that the classical Kalman filter cannot be applied, and this gave rise to a different class

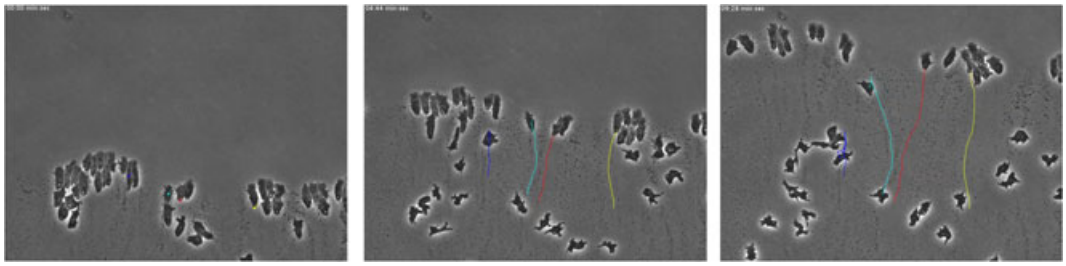

FIGURE 1 Three frames from the high-resolution microscopy video recording *Dictyostelium* cells movement, corresponding to times (min:sec): 00:00, 04:44, 09:28. The data used in this study consist of two cell paths in the form of 200 spatial locations (x and y coordinates) extracted at 4-s intervals. The two chosen cell paths correspond to the red (strong directed movement) and dark blue (weak directed movement) lines in the three time frames

of nonlinear Kalman filters that can be applied to nonlinear system representations. One of these methods is the unscented Kalman filter (UKF), an online Bayesian filtering method for nonlinear systems introduced by Julier and Uhlmann (1997) as an alternative to the widely used extended Kalman filter. The former provides improved performance in terms of stability and accuracy of estimates without any linearisation. Furthermore, it can easily be scaled up to higher dimensions. Intuitively, the classical Kalman filter starts from the initial distribution of the state vector, drawn from a multivariate normal distribution, which is then iterated through a prediction and updating step for each measurement available using the transition and observation models. Recent applications of the UKF include modelling soft tissue mechanics of the heart (Xi et al., 2011), chemical kinetics (Baker, Poskar, & Junker, 2011), modelling beetle flying techniques (Mohamad, 2015), neural network training (van der Merwe & Wan, 2001), and nonlinear dynamical system identification (Sitz, Schwarz, Kurths, & Voss, 2002; Voss, Timmer, & Kurths, 2004) .

In this paper, we present our approach to fitting the SDE model in Equations 1 and 2 to cell movement data using the UKF, a nonlinear Bayesian filter. We provide some insights into the particularities of this model using simulated data and discuss the results of this analysis from a real data set, describing the movement of *Dictyostelium* cells (see Figure 1).

2 | METHODS

2.1 | The Kalman filter in general

We introduce the UKF by referring to a general SSM, composed of a *state equation* as follows:

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}, \boldsymbol{\epsilon}_t) \tag{3}$$

and a *measurement equation* as follows:

$$\mathbf{y}_t = \mathbf{g}(\mathbf{x}_t, \boldsymbol{\nu}_t), \tag{4}$$

where $\mathbf{x}_t$ represents the vector of the hidden states, $\mathbf{y}_t$ are the measurements, $\boldsymbol{\epsilon}_t$ is the process noise at time t, with covariance matrix $\boldsymbol{Q}_t$, $\boldsymbol{\nu}_t$ is the observation noise at time t, with covariance matrix $\boldsymbol{R}_t$, and the functions $\mathbf{f}$ and $\mathbf{g}$ represent the transition and the observation models, respectively. Whilst non-Gaussian or correlated noise sources can be accommodated using nonlinear Kalman filters, in simpler applications, $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\nu}_t$ usually represent independent and identically distributed additive Gaussian noise. Furthermore, if the observation and transition models are

linear, Equations 3 and 4 can be simplified to give rise to an LG-SSM or a linear dynamic model, as follows:

$$\mathbf{x}_t = \mathbf{A}_t\mathbf{x}_{t-1} + \boldsymbol{\epsilon}_t,\ \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_t) \tag{5}$$

$$\mathbf{y}_t = \mathbf{B}_t\mathbf{x}_t + \boldsymbol{\nu}_t,\ \boldsymbol{\nu}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R}_t). \tag{6}$$

In Equations 5 and 6, $\mathbf{A}_t$ and $\mathbf{B}_t$ represent the transition and measurement matrices. The importance of the LG-SSM model comes from the fact that it supports exact inference via the Kalman filter: The Gaussian prior representing the initial state of the system, $p(\mathbf{x}_1) = \mathcal{N}(\boldsymbol{\mu}_{1|0}, \boldsymbol{\Sigma}_{1|0})$ is propagated to the following states that will also be Gaussian, $p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \mathcal{N}(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ (for notational simplicity).

In general, the filtering problem just refers to estimating the latent states $\mathbf{x}_t$ given the noisy observations $\mathbf{y}_t$, that is, calculating the density $p(\mathbf{x}_t|\mathbf{y}_{1:t})$. This can be performed online or, as the data stream in, by sequentially applying Bayes rule in the form of an *update* step, as follows:

$$p\left(\mathbf{x}_t|\mathbf{y}_t, \mathbf{y}_{1:t-1}\right) \propto p(\mathbf{y}_t|\mathbf{x}_t) \cdot p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \tag{7}$$

and a *prediction* step, as follows:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1}. \tag{8}$$

The advantage of the Kalman filter comes from the fact that the probability distribution of the predictor step $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ and the probability distribution of the updating step $p(\mathbf{x}_t|\mathbf{y}_t, \mathbf{y}_{1:t-1})$ can be obtained in closed form using the properties of Gaussian distribution and linearity assumptions as follows:

$$p\left(\mathbf{x}_t|\mathbf{y}_{1:t-1}\right) = \int p\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right) p\left(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}\right) d\mathbf{x}_{t-1} \tag{9}$$

$$= \int \mathcal{N}(\mathbf{x}_t|\mathbf{A}_t\mathbf{x}_{t-1}, \mathbf{Q}_t)\mathcal{N}\left(\mathbf{x}_{t-1}|\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}\right) d\mathbf{x}_{t-1} \tag{10}$$

$$= \mathcal{N}\left(\mathbf{x}_t|\mathbf{A}_t\boldsymbol{\mu}_{t-1}, \mathbf{A}_t\boldsymbol{\Sigma}_{t-1}\mathbf{A}_t^T + \mathbf{Q}_t\right), \tag{11}$$

and so,

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \mathcal{N}(\mathbf{x}_t|\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\Sigma}}_t), \tag{12}$$

where $\bar{\boldsymbol{\mu}}_t$ and $\bar{\boldsymbol{\Sigma}}_t$ are the prediction mean and covariance matrix at time t.

For the update step, we make use of the known result for Bayes rule applied to linear Gaussian systems (Murphy, 2012, p. 119), which means that $p(\mathbf{x}_t|\mathbf{y}_t, \mathbf{y}_{1:t-1})$ from (7) has a closed-form solution and is also a Gaussian distribution, as follows:

$$p\left(\mathbf{x}_t|\mathbf{y}_t, \mathbf{y}_{1:t-1}\right) = \mathcal{N}\left(\mathbf{x}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\right), \tag{13}$$

where $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ are the update mean and covariance matrix at time t:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t\mathbf{r}_t \tag{14}$$

$$\boldsymbol{\Sigma}_t = \left(\mathbf{I} - \mathbf{K}_t\mathbf{B}_t\right)\boldsymbol{\Sigma}_{t|t-1}. \tag{15}$$

In Equations 14 and 15, $\mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{B}_t^T(\mathbf{R}_t + \mathbf{B}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{B}_t^T)^{-1}$ represents the Kalman gain matrix and $\mathbf{r_t} = \mathbf{y}_t - \hat{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{B}_t\boldsymbol{\mu}_{t|t-1}$ represents the residual or the difference between the predicted and the actual observations. For step-by-step derivations of the Kalman filter equations, see p. 643 in Murphy (2012).

Therefore, the algorithm essentially updates the mean and covariance of the Gaussian distribution of the state vector at each iteration, meaning that the Kalman filter equations are entirely dependent on the predicted means and covariances of $\mathbf{x}_t$ and $\mathbf{y}_t$, given in (12).

As such, Julier and Uhlmann (1997) pointed out that in a nonlinear scenario (i.e., when **f** or **g** is nonlinear), the Kalman filter challenge essentially amounts to calculating the first two moments of a random variable that has to undergo a nonlinear transformation.

2.2 | Unscented transformation

In the UKF algorithm, the approximation of the Gaussian distribution is made using the unscented transform, which consists of a set of deterministically chosen sigma points that are passed through the nonlinear function and weighted to obtain the mean and covariance of the Gaussian (Sitz et al., 2002). The unscented transform was developed based on the intuition that it is easier to approximate a Gaussian distribution with a fixed number of parameters than to calculate an approximation to a nonlinear transformation (Julier & Uhlmann, 2004). However, because no random sampling is involved, this is not to be confused with a Monte Carlo method, which also means that a smaller number of sampled points are required (see Figure 2 for an illustration of these methods for a 2D system) . More specifically, in the context of UKF, the location and weights for the sigma points are chosen to match the first two moments (mean and covariance) of the prior distribution. The points are then transformed using the nonlinear function, and the statistics of interest (mean and covariance) are calculated (Julier & Uhlmann, 2004).

In general, if $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{y} = f(\mathbf{x})$ is a nonlinear function, and we want to estimate $p(\mathbf{y})$, the unscented transformation achieves this by first calculating a set of $2d + 1$ sigma points, as follows:

$$\mathcal{X} = \left(\boldsymbol{\mu}, \left\{ \boldsymbol{\mu} + \left(\sqrt{(d+\lambda)\boldsymbol{\Sigma}} \right)_i \right\}_{i=1}^{d}, \left\{ \boldsymbol{\mu} - \left(\sqrt{(d+\lambda)\boldsymbol{\Sigma}} \right)_i \right\}_{i=1}^{d} \right), \tag{16}$$

where $\boldsymbol{\Sigma}_i$ denotes the ith column of the matrix $\boldsymbol{\Sigma}$, λ is a scaling parameter, and d is the dimension of the vector $\mathbf{x}$.

To obtain the transformed sigma points, we propagate them through the nonlinear function $\mathcal{Y}_i = f(\mathcal{X}_i)$, and then, we obtain the mean and covariance of $p(\mathbf{y})$ by weighting the transformed sigma points, as follows:

$$\boldsymbol{\mu}_{\mathbf{y}} = \sum_{i=0}^{2d} w_i \mathcal{Y}_i \tag{17}$$

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \sum_{i=0}^{2d} w_i \left(\mathcal{Y}_i - \boldsymbol{\mu}_{\mathbf{y}} \right) \left(\mathcal{Y}_i - \boldsymbol{\mu}_{\mathbf{y}} \right)^T, \tag{18}$$

where the weights are defined as follows:

$$w_0 = \frac{\lambda}{d+\lambda} \tag{19}$$

$$w_i = \frac{1}{2(d+\lambda)}. \tag{20}$$

2.3 | Nonaugmented UKF

The unscented transformation is used twice for each iteration of the UKF algorithm. First, the algorithm approximates the predictive density $p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ by using the previous

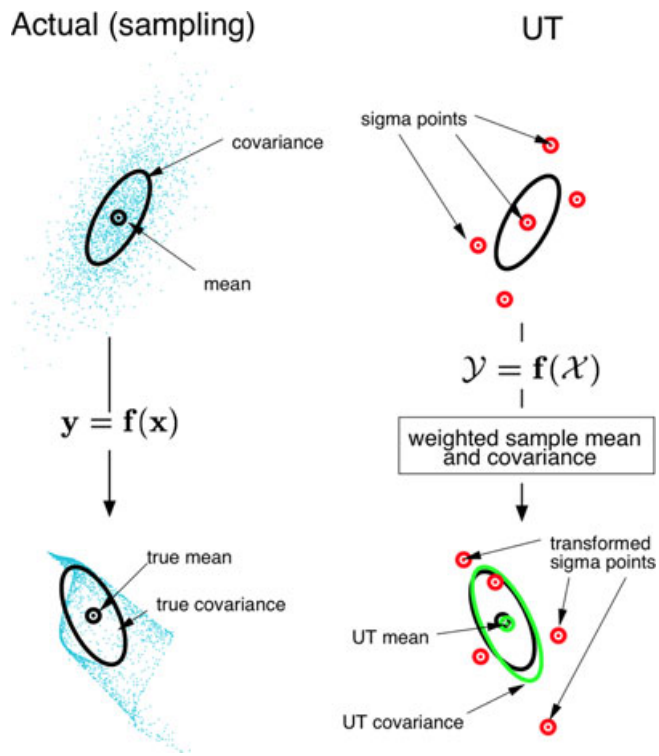

FIGURE 2 A comparison of Monte Carlo sampling (left panel) and the unscented transformation (denoted as UT on the right panel) for the propagation of the mean and covariance of a vector $\mathbf{x}$ through a nonlinear function $\mathbf{f}$. The left panel shows a sample from a Monte Carlo simulation (displayed as blue points), which is transformed using the nonlinear function $\mathbf{f}$ and then used to calculate the true mean and covariance (displayed as a black circle and a black ellipse). Note that with only five sigma points, shown in the right panel (which is orders of magnitude less than a typical Monte Carlo sampler, shown by the blue points in the left panel), the UT provides good estimates for the mean and covariance of the transformed variable $\mathbf{y}$. The unscented transformation steps shown on the right-hand side are as follows: A set of five sigma points (red circles at the top) is chosen to represent the 2D system, which is then passed individually through the nonlinear function to obtain the set of transformed sigma points $\mathcal{Y} = \mathbf{f}(\mathcal{X})$ (red circles at the bottom). Calculating the weighted sample mean and covariance of the transformed sigma points (plotted as a green circle and a green ellipse) will provide an estimate for the true mean and covariance of the variable of interest $\mathbf{y}$ (obtained using Monte Carlo simulation and displayed as a black circle and a black ellipse; Wan and van der Merwe, 2000)

time point belief state $\mathcal{N}(\mathbf{x}_{t-1}|\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ and by passing it through the transition model $\mathbf{f}$, as follows:

$$\boldsymbol{\mathcal{X}}_t^* = \left(\boldsymbol{\mu}_{t-1}, \left\{\boldsymbol{\mu}_{t-1} + \left(\sqrt{(d+\lambda)\boldsymbol{\Sigma}_{t-1}}\right)_i\right\}_{i=1}^{d}, \left\{\boldsymbol{\mu}_{t-1} - \left(\sqrt{(d+\lambda)\boldsymbol{\Sigma}_{t-1}}\right)_i\right\}_{i=1}^{d}\right) \tag{21}$$

$$\boldsymbol{\mathcal{X}}_t = \mathbf{f}\left(\boldsymbol{\mathcal{X}}_t^*\right) \tag{22}$$

$$\bar{\boldsymbol{\mu}}_t = \sum_{i=0}^{2d} w_i \boldsymbol{\mathcal{X}}_i \tag{23}$$

$$\bar{\boldsymbol{\Sigma}}_t = \sum_{i=0}^{2d} w_i \left(\boldsymbol{\mathcal{X}}_i - \bar{\boldsymbol{\mu}}_t\right)\left(\boldsymbol{\mathcal{X}}_i - \bar{\boldsymbol{\mu}}_t\right)^T + \boldsymbol{Q}. \tag{24}$$

The second unscented transform is performed to approximate the likelihood $p(\mathbf{y}_t|\mathbf{x}_t)$ by using the prior $\mathcal{N}(\mathbf{x}_t|\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\Sigma}}_t)$ and the observation model $\mathbf{g}$, as follows:

$$\boldsymbol{\mathcal{Y}}_t^* = \left(\bar{\boldsymbol{\mu}}_t, \left\{ \bar{\boldsymbol{\mu}}_t + \left(\sqrt{(d+\lambda)\bar{\boldsymbol{\Sigma}}_t} \right)_i \right\}_{i=1}^{d}, \left\{ \bar{\boldsymbol{\mu}}_t - \left(\sqrt{(d+\lambda)\bar{\boldsymbol{\Sigma}}_t} \right)_i \right\}_{i=1}^{d} \right) \tag{25}$$

$$\boldsymbol{\mathcal{Y}}_t = \mathbf{g}\left(\boldsymbol{\mathcal{Y}}_t^*\right) \tag{26}$$

$$\hat{\mathbf{y}}_t = \sum_{i=0}^{2d} w_i \boldsymbol{\mathcal{Y}}_i \tag{27}$$

$$\mathbf{S}_t = \sum_{i=0}^{2d} w_i \left(\boldsymbol{\mathcal{Y}}_i - \hat{\mathbf{y}}_t\right)\left(\boldsymbol{\mathcal{Y}}_i - \hat{\mathbf{y}}_t\right)^T + \boldsymbol{R}. \tag{28}$$

We then obtain the posterior density $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ using Bayes rule (see Murphy, 2012, chapter 18.5, for derivations) as follows:

$$\boldsymbol{\Sigma}_t^{xy} = \sum_{i=0}^{2d} w_i \left(\boldsymbol{\mathcal{X}}_i - \bar{\boldsymbol{\mu}}_t\right)\left(\boldsymbol{\mathcal{Y}}_i - \hat{\mathbf{y}}_t\right)^T \tag{29}$$

$$\mathbf{K}_t = \boldsymbol{\Sigma}_t^{xy}\mathbf{S}_t^{-1} \tag{30}$$

$$\boldsymbol{\mu}_t = \bar{\boldsymbol{\mu}}_t + \mathbf{K}_t\left(\mathbf{y}_t - \hat{\mathbf{y}}_t\right) \tag{31}$$

$$\boldsymbol{\Sigma}_t = \bar{\boldsymbol{\Sigma}}_t - \mathbf{K}_t\mathbf{S}_t\mathbf{K}_t^T. \tag{32}$$

An extension to the classical UKF proposed by Sitz et al. (2002) indicates that model parameters $\boldsymbol{\theta}$ can be included as dynamical variables in the hidden states vector $\mathbf{x}_t$, which means that these can be estimated at every time point alongside other unobserved system states. Thus, we can use the UKF in an augmented form as a unified methodology for system state tracking and parameter estimation.

2.4 | Augmented UKF

Wu, Hu, Wu, and Hu (2005) presented an augmented version of the UKF and demonstrated the advantages of process and measurement noise augmentation to the state vector in the case of nonadditive or non-Gaussian noise. The reason behind this is that the measurement and process noise added to the state vector capture better odd-moment information through the computation of additional sigma points. Applications and implementations of this method can be found in Hartikainen, Solin, & Särkkä (2011) and Sitz et al. (2002).

The augmented UKF can be reformulated from the standard UKF SSM by merging the signal states $\mathbf{x}_t$, parameters $\boldsymbol{\theta}_t$, process noise $\boldsymbol{\epsilon}_t$, and measurement noise $\boldsymbol{\nu}_t$ and by constructing an *augmented state equation* as follows:

$$\mathbf{x}_t^a = \begin{bmatrix} \mathbf{x}_t \\ \boldsymbol{\theta}_t \\ \boldsymbol{\epsilon}_t \\ \boldsymbol{\nu}_t \end{bmatrix} = \begin{bmatrix} \mathbf{f}\left(\mathbf{x}_{t-1}, \boldsymbol{\epsilon}_{t-1}\right) \\ \boldsymbol{\theta}_{t-1} \\ \boldsymbol{\epsilon}_{t-1} \\ \boldsymbol{\nu}_{t-1} \end{bmatrix} = \mathbf{f}^a\left(\mathbf{x}_{t-1}^a\right)$$

and an *augmented measurement equation*, as follows:

$$\mathbf{y}_t^a = \mathbf{g}^a\left(\mathbf{x}_t^a\right) = \mathbf{g}^a\left(\mathbf{x}_t, \boldsymbol{\nu}_t\right).$$

Note that the augmented UKF algorithm equations can be derived in a similar manner to Equations 21–28, except that in the prediction step, the sigma points are calculated using the augmented state vector $\mathbf{x}_t^a$ and the corresponding augmented covariance matrix, as follows:

$$\mathbf{\Sigma}_t^a = \begin{bmatrix} \mathbf{\Sigma_t} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{\theta_t} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Q_t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{R_t} \end{bmatrix}, \tag{33}$$

where $\mathbf{\Sigma}_{\theta_t}$ is the covariance matrix for the parameters. Also, the number of sigma points is now $2d_a + 1$, where d_a is the size of the augmented state vector $\mathbf{x}_t^a$.

3 | SIMULATION RESULTS

In this section, we include results from various simulations we carried out using the augmented and nonaugmented UKF implementations from the MATLAB toolbox developed by Hartikainen et al. (2011).

First, we apply the Euler–Maruyama discretisation to bring the system described in Equations 1 and 2 into the standard SSM described in Section 2, as follows:

$$X_t = X_{t-1} + \sigma \Delta B_t^X \tag{34}$$

$$Y_t = Y_{t-1} + \frac{\alpha\beta \exp[-\beta(Y_{t-1} - \gamma t)]}{\{1 + \exp[-\beta(Y_{t-1} - \gamma t)]\}^2} \Delta t + \sigma \Delta B_t^Y. \tag{35}$$

Here, ΔB_t^X and ΔB_t^Y are just sums of random normal increments between time $t-1$ and t. As such, in a more general notation (dropping the superscript for ease),

$$\Delta B_t = \sum_{k=1}^{\frac{\Delta t}{\delta t}} dB_k, \text{ where } dB_k \sim \sqrt{\delta t}\mathcal{N}(0,1).$$

In the notation above, Δt is the sampling time step and δt is the integration time step required by the discretisation of the Brownian motion path (for more details on the discretisation of SDEs, see Higham, 2001). We fix these time steps at values $\Delta t = 0.1, \delta t = 0.001$ for all the simulations presented in this section.

Equations 34 and 35 thus define the transition function $\mathbf{f}$ from (3). In this scenario, we assume that the process is observed with a small amount of additive Gaussian noise $\nu_t \sim \mathcal{N}(0, 0.1^2)$; hence, $\mathbf{g}$ from (4) is $\mathbf{y}_t = \mathbf{g}(\mathbf{x}_t, \boldsymbol{\nu}_t) = \mathbf{x}_t + \boldsymbol{\nu}_t$.

We then fit the nonaugmented UKF to a synthetic data set using the following parameters: $\alpha = 5, \beta = 2, \gamma = 0.5, \sigma = 0.1$. The results summarised in Figure 3 indicate good agreement between the estimated UKF path and the true cell path.

The UKF also provides good estimates for the parameters $\hat{\beta} = 1.88, \hat{\gamma} = 0.51, \hat{\sigma} = 0.04$ with relatively small standard errors $0.17, 0.83, 0.06$, except for $\hat{\alpha}$ where the estimates indicate a more substantial deviation from the true parameter (bias: 0.44, and standard error is 0.35).

3.1 | Profile likelihood

A potential source of bias as the one observed in Figure 3 can be investigated by looking at the likelihood, that is, marginal likelihood with respect to the hidden states. In order to do that, we

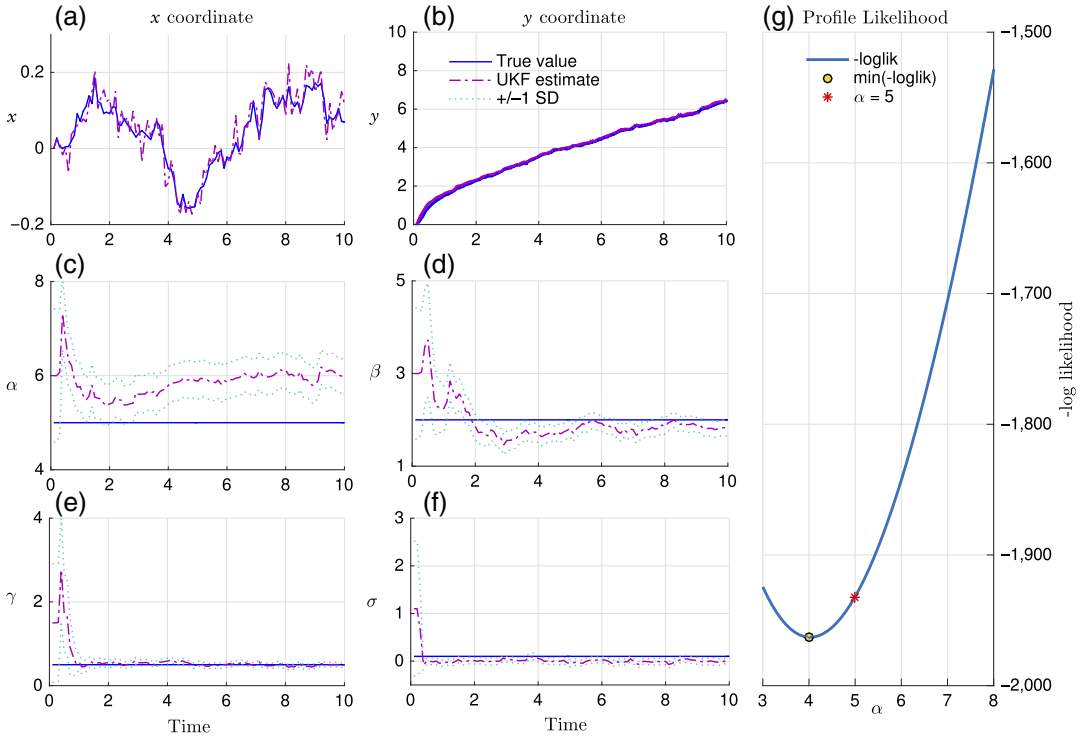

FIGURE 3 Simulation results: Nonaugmented unscented Kalman filter (UKF) tracking of cell coordinates x (subplot a) and y (subplot b), and parameters α (subplot c), β (subplot d), γ (subplot e), σ (subplot f) for time interval $[0, 10]$, displayed on the y axes. Parameter estimates are shown as dash-dotted purple line, ± 1 standard error bounds displayed as a blue dotted line, and true parameter values are shown as blue continuous lines. On the right-hand side, negative log profile likelihood plot for α, obtained by fixing the other three parameters at their true values (subplot g)

first derive the probability of the observed system at time t conditional on the state of the system at time $t-1$ by integrating out the latent variable $\mathbf{x}_t$, as follows:

$$p(\mathbf{y}_t|\mathbf{y}_{t-1}) = \int p(\mathbf{y}_t|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{y}_{t-1})\, d\mathbf{x}_t \tag{36}$$

$$= \int \mathcal{N}(\mathbf{y}_t|\mathbf{x}_t, \mathbf{R}_t)\, \mathcal{N}(\mathbf{x}_t|\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\Sigma}}_t) d\mathbf{x}_t. \tag{37}$$

Using the Gaussian convolution integral results (Bishop, 2006, chapter 2.3), we simplify (37) to the following:

$$p(\mathbf{y}_t|\mathbf{y}_{t-1}) = \mathcal{N}(\mathbf{y}_t|\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\Sigma}}_t + \mathbf{R}_t), \tag{38}$$

where $\bar{\boldsymbol{\mu}}_t$ and $\bar{\boldsymbol{\Sigma}}_t$ are the predicted mean and covariance at time t, and $\mathbf{R}_t$ is the measurement noise covariance matrix at time t.

The log likelihood is then as follows:

$$\mathcal{L} = \log \prod_t p(\mathbf{y}_t|\mathbf{y}_{t-1}) = \sum_t \log \mathcal{N}\left(\mathbf{y}_t|\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\Sigma}}_t + \mathbf{R}_t\right) \tag{39}$$

$$\propto \sum_t \left(\log\det(2\pi\boldsymbol{\Sigma}_t) + 0.5\left(\mathbf{y}_t - \bar{\boldsymbol{\mu}}_t\right)^T \boldsymbol{\Sigma}_t^{-1}\left(\mathbf{y}_t - \bar{\boldsymbol{\mu}}_t\right)\right), \tag{40}$$

where $\mathbf{\Sigma}_t$ and $\bar{\mathbf{\Sigma}}_t$ are defined in Equations 32 and 24, respectively. We evaluate the marginal log likelihood in (40) by considering a grid of values for each parameter in the model and by fitting the UKF with each parameter combination. The results summarising the profile likelihood for the α parameter in Figure 3 can be used to calculate the Cramér–Rao lower bound, which provides an indication of the intrinsic uncertainty specific to the problem. In this case, the minimum standard deviation attainable by an estimator of α is 0.14. The standard error obtained from the UKF estimation for α is 0.35; this then indicates that the estimated value of the parameter is reasonably close to the true value.

3.2 | Comparison of the augmented and nonaugmented UKF

It was also of interest to investigate the performance of the augmented UKF (including the error terms in the state vector) compared with the nonaugmented UKF, as the former has been reported in the literature to provide better estimation (Hartikainen et al., 2011; Wu et al., 2005).

A simulation study was carried out to explore the differences in estimation of the augmented versus nonaugmented UKF on the cell movement system parameters in Equations (34) and (35). Both methods of interest were applied to 20 data instantiations, where all the system parameters $\alpha = 5, \beta = 2, \gamma = 0.5, \sigma = 0.1$, as well as the integration and discretisation parameters $\delta t = 0.001, \Delta t = 0.1$, were kept fixed. The difference between the 20 data sets would then be due to the intrinsic stochasticity of the sampled Brownian motion paths $\Delta B_t^X, \Delta B_t^Y$.

The comparison of the two methods is performed by comparing the relative bias (R.Bias) and relative MSE (R.MSE), calculated as follows:

$$\text{Relative Bias:} \quad \text{R.Bias} = \frac{(\hat{\theta} - \theta)}{\theta} \tag{41}$$

$$\text{Relative MSE:} \quad \text{R.MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right)^2, \tag{42}$$

where $\hat{\theta}$ denotes the UKF estimate for each parameter of interest, θ is the true parameter value, and N is the number of observations in the simulation.

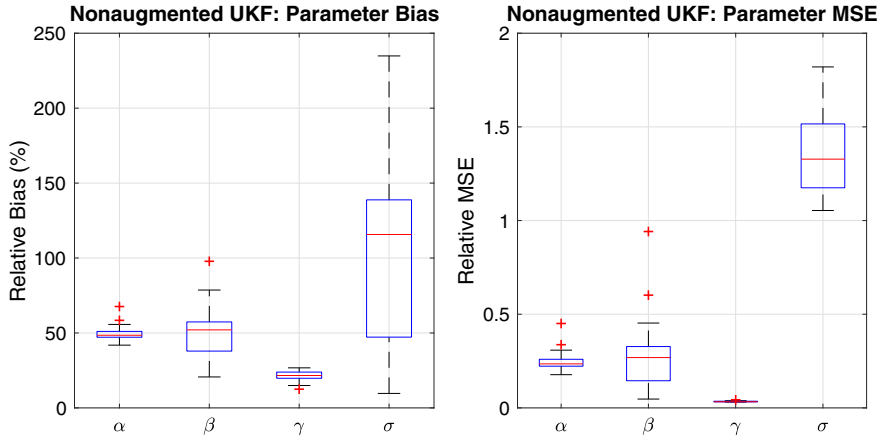

FIGURE 4 Simulation results of *nonaugmented unscented Kalman filter (UKF)* for the cell movement system parameters $\alpha, \beta, \gamma, \sigma$. Left-hand side: relative bias (%). Right-hand side: relative mean squared error (MSE). Results were obtained from 20 data set instantiations

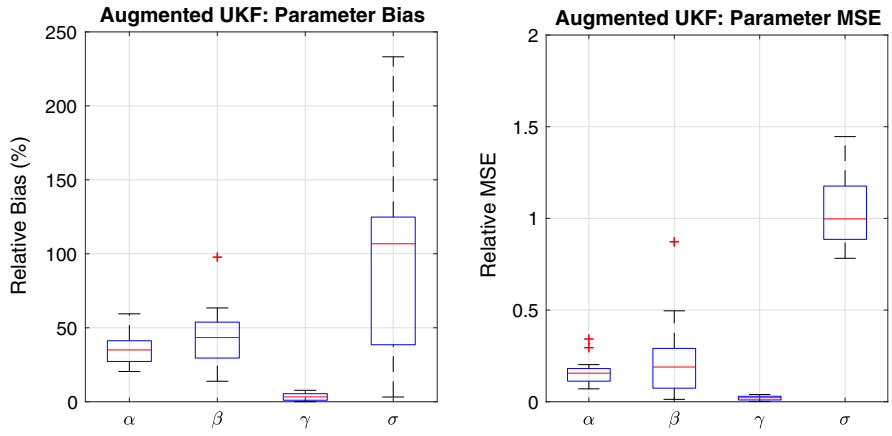

FIGURE 5 Simulation results of *augmented unscented Kalman filter (UKF)* for the cell movement system parameters $\alpha, \beta, \gamma, \sigma$. Left-hand side: relative bias (%). Right-hand side: relative mean squared error (MSE). Results were obtained from 20 data set instantiations

TABLE 1 Comparison of average *relative mean squared error* (R.MSE) obtained from 20 data set instantiations between the augmented and nonaugmented unscented Kalman filter (UKF) for the four parameters in the model: $\alpha, \beta, \gamma, \sigma$

Parameter	Augmented UKF R.MSE	Nonaugmented UKF R.MSE
α	0.15	0.23
β	0.19	0.27
γ	0.02	0.03
σ	1.03	1.32

Figure 4 displays boxplots of the relative bias (left-hand side) and the relative MSE (right-hand side) for the nonaugmented Kalman filter constructed with the estimates from the 20 data set instantiations. Similarly, Figure 5 displays the same measures for the augmented Kalman filter. In terms of relative bias, the augmented version provides an average improvement of 13% for α, 9% for β, 18% for γ, and 10% for σ. In terms of R.MSE, an improvement is also observed from the augmented UKF for each of the parameters in the cell movement system (see Table 1).

4 | APPLICATION ON *DICTYOSTELIUM* CELLS MOVEMENT

Dictyostelium cells are widely used in experiments as proxies for understanding the mechanisms of human disease because of their similarities to important human cells (leukocytes and cancer cells) in terms of biology and response to chemotaxis (Tweedy et al., 2016).

Data on cell movement is available in the form of a high-resolution microscopy video of a group of *Dictyostelium* amoebae. These data were produced in a dish of agar containing spatially homogeneous levels of the chemoattractant folate, which can be consumed and depleted by the cells to create a resource gradient. *Dictyostelium* cells were added to a trough cut in the centre of the dish, and the movement of these cells once they had moved under the agar and out of this trough was filmed under a microscope. The data used in this section consist of two cell paths

corresponding to 200 spatial locations as x and y coordinates extracted at 4-s intervals using the ImageJ software (Rasband, 1997).

The choice of cells for this study was motivated by their movement behaviour; One displays a more random movement, due to weaker drift, and one displays a more directed movement, due to stronger drift. These cells can be identified visually from Figure 1 as belonging to either the group of front cells or the group of cells that linger behind. These specific movement patterns arise as a consequence of the fact that the front cells are driven upwards by the availability of the chemoattractant, which they deplete locally, whereas the cells behind show a more random movement because the chemoattractant they experience has been partly depleted by the front group.

Figure 6 displays the augmented versus nonaugmented UKF estimation results for a *Dictyostelium* cell displaying more random (weak drift) movement. We emphasise that the

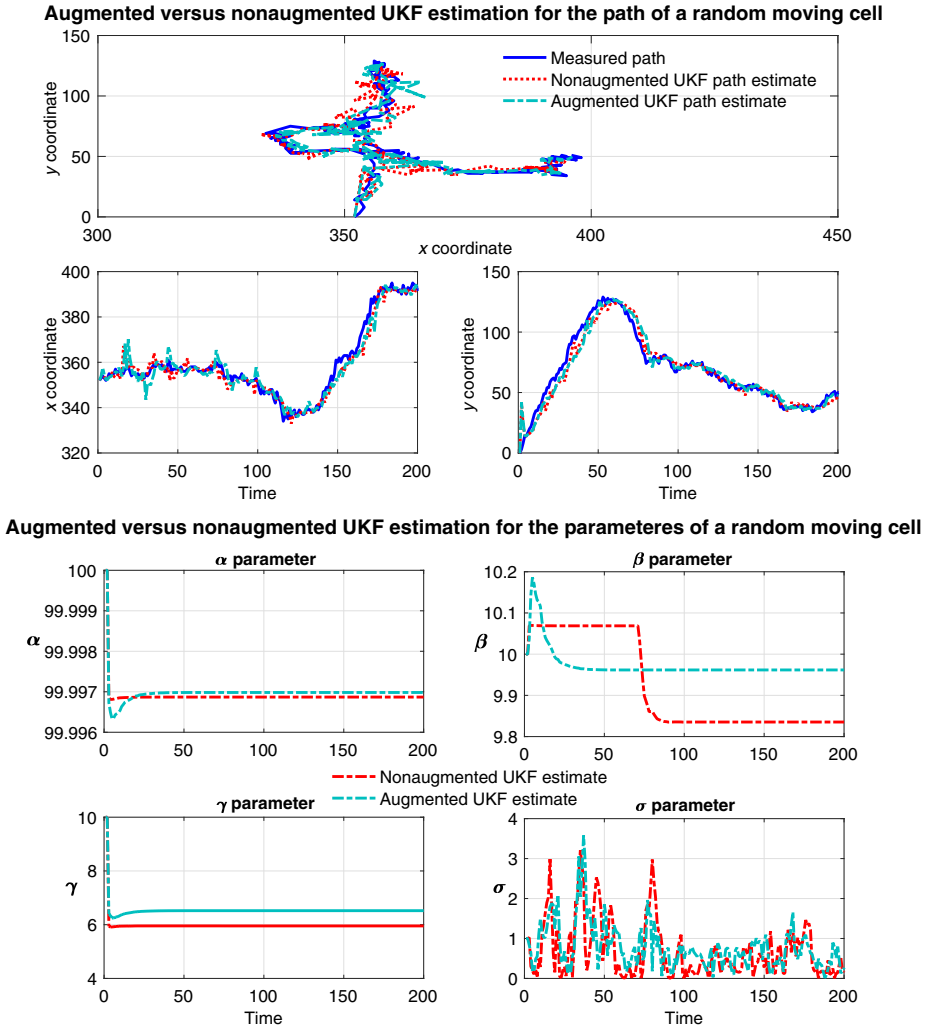

FIGURE 6 Comparison of augmented and nonaugmented unscented Kalman filter (UKF) for the path estimation (top panel) and parameter estimation (bottom panel) for a *Dictyostelium* cell displaying random (weak drift) movement

main interest for biological applications is the inference of the parameters, which are included in Figure 6 (bottom). However, because the true parameters for the real data are unknown, we use the tracking of the cell trajectories as a proxy for assessing the accuracy of inference

TABLE 2 Comparison of MSE between the augmented and nonaugmented unscented Kalman filter (UKF), for the weak drift and strong drift movement of *Dictyostelium* cells represented by their x and y coordinates

Cell type	System state	Augmented UKF MSE	Nonaugmented UKF MSE
Weak drift movement	x coordinate	9.52	12.22
Weak drift movement	y coordinate	51.33	53.01
Strong drift movement	x coordinate	14.59	26.46
Strong drift movement	y coordinate	14.40	15.49

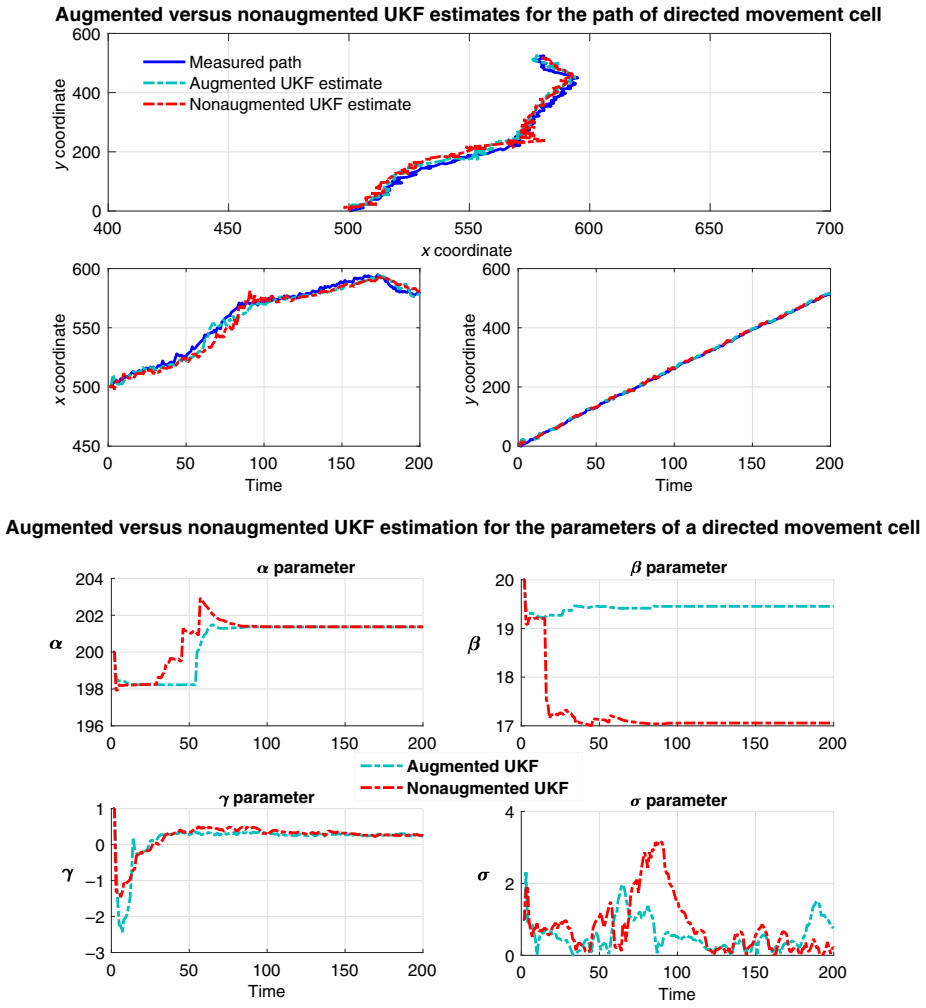

FIGURE 7 Comparison of augmented and nonaugmented unscented Kalman filter (UKF) for the path estimation (top panel) and parameter estimation (bottom panel) for a *Dictyostelium* cell displaying directed (strong drift) movement

(see Figure 6, top). In terms of MSE, the augmented UKF provides a marginal improvement over the nonaugmented UKF (see Table 2), which is in agreement with the findings from visually inspecting the cell path estimates.

Similarly, the corresponding results for the directed movement (strong drift) of a *Dictyostelium* cell and the visual estimates in Figure 7 (top) are in close agreement with the MSE calculation in Table 2; the augmented version provides marginally better estimates for the cell's x and y coordinates. As before, we include the comparison of parameter estimates in Figure 7 (bottom), with the mention that a gold standard (i.e., true parameter value) is not available, and we rely on simulations from the system equations in (34) and (35) to assess goodness of fit.

5 | CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrate the application of the UKF, a Bayesian filtering technique that adequately trades off accuracy versus computational efficiency, to a real-world problem potentially relevant to cancer research: the movement of *Dictyostelium* cells, which has not been tackled at individual cell level before.

First, we have investigated the performance of the noise augmented UKF in comparison with the nonaugmented UKF through a simulation study. Our results indicated a small reduction of the estimation bias for cell movement parameters between, on average, 9% and 18%. Second, we applied both of these methods to real data consisting of the movement of two *Dictyostelium* cells, one displaying strong drift (more directed) movement and the other one displaying weak drift (more random) movement. Our results indicate that both methods estimate the cell paths well, with the augmented version providing a marginal improvement.

In conclusion, our results indicate that the UKF can be successfully used for parameter inference and tracking cells displaying various movement patterns. Future research will extend this work by applying the UKF to a population of cells. Additionally, we plan to use this framework of parameter and state estimation to fit models describing alternative movement mechanisms, such as the self-induced gradient model described by Tweedy et al. (2016), and employ model selection criteria to choose the best model.

ACKNOWLEDGEMENTS

The authors thank Robert Insall and his research group at the Beatson Institute for sharing the *Dictyostelium* cell microscopy images.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

FINANCIAL DISCLOSURE

This work is part of the research programme of the Centre for Multiscale Soft Tissue Mechanics with application to heart & cancer (SofTMech), funded by the Engineering and Physical Sciences Research Council (EPSRC) of the UK, Grant EP/N014642/1.

REFERENCES

Aeschbacher, S., Beaumont, M. A., & Futschik, A. (2012). A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics*, *192*(3), 1027–1047.

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(3), 269–342.

Bahmani-Oskooee, M., & Brown, F. (2004). Kalman filter approach to estimate the demand for international reserves. *Applied Economics*, *36*(15), 1655–1668.

Baker, S. M., Poskar, C. H., & Junker, B. H. (2011). Unscented Kalman filter with parameter identifiability analysis for the estimation of multiple parameters in kinetic models. *EURASIP Journal on Bioinformatics and Systems Biology*, *2011*(1), 7. https://doi.org/10.1186/1687-4153-2011-7

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Choset, H., & Nagatani, K. (2001). Topological simultaneous localization and mapping (SLAM): Toward exact localization without explicit localization. *IEEE Transactions on Robotics and Automation*, *17*(2), 125–137.

Ferguson, E. A., Matthiopoulos, J., Insall, R. H., & Husmeier, D. (2017). Statistical inference of the mechanisms driving collective cell movement. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(4), 869–890.

Golightly, A., & Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, *1*(6), 807–820.

Grewal, M. S., & Andrews, A. P. (2010). Applications of Kalman filtering in aerospace 1960 to the present [historical perspectives]. *IEEE Control Systems*, *30*(3), 69–78.

Hartikainen, J., Solin, A., & Särkkä, S. (2011). Optimal filtering with Kalman filters and smoothers: A manual for the Matlab toolbox EKF/UKF. Espoo, Finland: Department of Biomedica Engineering and Computational Sciences, Aalto University School of Science.

Higham, D. J. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, *43*(3), 525–546.

Julier, S. J., & Uhlmann, J. K. (1997). *A new extension of the Kalman filter to nonlinear systems*. Paper presented at the AeroSense '97, Orlando, FL, pp. 182–193.

Julier, S. J., & Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, *92*(3), 401–422.

Mingas, G., Bottolo, L., & Bouganis, C. S. (2017). Particle MCMC algorithms and architectures for accelerating inference in state-space models. *International Journal of Approximate Reasoning*, *83*, 413–433.

Mohamad, Z. (2015). *Modeling of beetle mimicking ornithopter using mixed unscented Kalman filter (UKF) and differential evolution (DE) method*. Kuala Lumpur, Malaysia: Kulliyyah of Engineering, International Islamic University Malaysia. Retrieved from https://books.google.co.uk/books?id=bHPKnQAACAAJ

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

Rasband, W. (1997–2012). ImageJ. Bethesda, MD: National Institutes of Health.

Sitz, A., Schwarz, U., Kurths, J., & Voss, H. U. (2002). Estimation of parameters and unobserved components for nonlinear systems from noisy time series. *Physical Review E*, *66*(1), 016210.

Tweedy, L., Knecht, D. A., Mackay, G. M., & Insall, R. H. (2016). Self-generated chemoattractant gradients: attractant depletion extends the range and robustness of chemotaxis. *PLOS Biology*, *14*(3): e1002404.

van der Merwe, R., & Wan, E. A. (2001). *Efficient derivative-free Kalman filters for online learning*. Paper presented at ESANN 2001 Proceedings, Bruges, Belgium, pp. 205–210.

Voss, H. U., Timmer, J., & Kurths, J. (2004). Nonlinear dynamical system identification from uncertain and indirect measurements. *International Journal of Bifurcation and Chaos*, *14*(6), 1905–1933.

Wan, E. A., & van der Merwe, R. (2000). *The unscented Kalman filter for nonlinear estimation*. Paper presented at the Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium, Alberta, Canada, pp. 153–158.

Wu, Y., Hu, D., Wu, M., & Hu, X. (2005). Unscented Kalman filtering for additive noise case: Augmented vs. non-augmented. *IEEE Signal Processing Letters*, *12*(5), 4051–4055.

Wyckoff, J., Wang, W., Lin, E. Y., Wang, Y., Pixley, F., Stanley, E. R., … Condeelis, J. (2004). A paracrine loop between tumor cells and macrophages is required for tumor cell migration in mammary tumors. *Cancer Research*, *64*(19), 7022–7029.

Xi, J., Lamata, P., Lee, J., Moireau, P., Chapelle, D., & Smith, N. (2011). Myocardial transversely isotropic material parameter estimation from in-silico measurements based on a reduced-order unscented Kalman filter. *Journal of the Mechanical Behavior of Biomedical Materials*, *4*(7), 1090–1102.

How to cite this article: Giurghita D, Husmeier D. Statistical modelling of cell movement. *Statistica Neerlandica*. 2018;72:265–280. https://doi.org/10.1111/stan.12140