Minimum Entropy Parameter Estimation of Bounded Nonlinear Dynamic Systems with Non-Gaussian State and Measurement Noise

George Papadopoulos * **Martin Brown** **

* *The Control Systems Centre, School of Electrical and Electronic Engineering, University of Manchester (email: g.papadopoulos@postgrad.manchester.ac.uk)*
** *The Control Systems Centre, School of Electrical and Electronic Engineering, University of Manchester (email: martin.brown@manchester.ac.uk)*

Abstract:
Parameter estimation plays an important role in Systems Biology in helping to understand the complex behavior of signal transduction networks. The problem becomes more complex as the inherent stochasticity of the signaling mechanism involves noise components of non-Gaussian nature. In this paper a novel stochastic parameter estimation method has been developed where the entropy of the joint residual PDF is used as a measure of the systems uncertainty. The optimal parameter values are selected as the ones corresponding to a minimal entropy value of the residual. The novelty of this approach lies in that the assumptions for the system involve both state and measurement noise of arbitrary distribution and the method is designed for general multivariable systems. The residual PDF is approximated using well known Kernel Density Estimation methods. The analysis of the method includes application to the RKIP regulated ERK signaling pathway and comparisons are drawn based on the Least Squares solution of the same problem.

Keywords: Minimum Entropy; Stochastic Parameter Estimation; ERK; RKIP

1. INTRODUCTION

The primary aim in systems biology is to understand the underlying complex structure of cellular signaling, transcription, translation or gene regulation (Wilkinson [2006]) and parameter estimation is a very important aspect of the modeling process involved. The additional difficulty of the modeling stage is the implicit stochasticity that is introduced since any level of abstraction of systems of such complexity will lead to models that are inherently stochastic instead of deterministic (Wilkinson [2006]). It could therefore be the case that a corresponding stochastic parameter estimation method could be more appropriate.

In the work presented in this paper, the modeling framework should be more accurately termed as stochastic kinetic modeling and the aim is to estimate the kinetic model parameters (rate constants) based on the assumption that the hypothetical system providing the target has been influenced by both state and measurement noise not necessarily of Gaussian nature. The deterministic model is a set of coupled ODEs that describe the time evolution of the biochemical process governing the cell, where the states are the species concentrations. The target response could be thought of as the direct, time course, stochastic output of a Gillespie stochastic simulator (Wilkinson [2006]).

Existing methods for stochastic parameter estimation like Gaussian Processes (GPs) and particle filters tackle the problem either by viewing it as regression task (GPs), where maximum likelihood is performed to obtain the optimal value of the hyperparameters that are linked with the original model parameters (Sanguinetti et al. [2006a,b]), or using state estimation first (particle filters) so as to obtain a posterior PDF for the update (Poyiadijis et al. [2003], Chan and Kouritzin [2001]) of the model observation and then proceed by finding the most probable value of the parameters. The difficulty associated with GPs is the assumed Gaussian prior in every step which generally is not considered to be a realistic assumption (Sanguinetti et al. [2006a]). Particle filters do overcome this limitation but substantial number of samples is needed for an accurate estimate, in addition to the Monte Carlo Integration step which can impose a considerable computational burden as the number of random variables increases (Doucet et al. [2001]).

The concept of minimum entropy has been used extensively in the past (Wang [2000]) in the area of control systems (Wang and Wang [2004], Yue et al. [2006]) with primary aim the uncertainty minimization of general bounded stochastic systems. It has also been shown (Wang [202], Yue and Wang [2003]) that when the noise is Gaussian then the the minimum entropy control corresponds to minimum variance. The application focused only on Single Input Single Output (SISO) systems only.

Minimum entropy parameter estimation is focused on obtaining high quality parameter estimates based on the target being influenced by both state and measurement noise and the examples included will show that it can outperform the corresponding solution obtained from a Least Squares (LS) estimator.

The minimum entropy parameter estimation framework presented here assumes state and measurement, i.i.d noise components. There are no other assumption regarding the structure of the noise PDFs. The modeling framework assumes a system of coupled Stochastic Differential Equations (SDEs) comprised of a deterministic and a stochastic part (the noise) describing the target generation or otherwise, the time evolution of the hypothetical target response of the biochemical reaction pathway. The underlying aim is to reach a state of minimum entropy by fine-tuning the parameters (stochastic parameter estimation process) of the deterministic model under the designers control. The uncertainty of such a state should also be minimal (Papoulis and Pillai [2002]). The advantage of using the entropy is that as a measure of uncertainty is much more general than the variance (Papoulis and Pillai [2002]) which is optimal only for cases of Gaussian noise assumptions.

This paper is organized as follows: in section 2 the general problem of parameter estimation in biochemical reaction pathways is described along with the assumptions of the modeling process. Section 3 describes the general Information Theory justification for the minimum entropy method. The formulation of the minimum entropy method is presented in section 4 and an analysis on the distribution of the parameter estimates is presented in section 5. Finally, the example on the RKIP regulated ERK pathway is included in section 6.

2. PROBLEM STATEMENT

The assumptions of the general modeling framework in biochemical reaction pathways is presented here, as well as justification and formal statement of the problem of finding an optimal parameter set based on state and measurement system noise.

2.1 The Stochastic Parameter Estimation Problem

The model of the system under study will be assumed to have a state space model as described in equation (1) with additive state and measurement i.i.d uncorrelated noise. There are no further restrictions on the nature of the random variables w and v.

$$\begin{aligned}\dot{\hat{\boldsymbol{x}}} &= f(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}) \\ \hat{\boldsymbol{y}} &= h(\hat{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}))\end{aligned} \quad \text{and the target} \quad \begin{aligned}\dot{\boldsymbol{x}} &= f(\boldsymbol{x}, \boldsymbol{\theta}^*) + w \\ \boldsymbol{y} &= h(\boldsymbol{x}(\boldsymbol{\theta}^*)) + v\end{aligned} \qquad (1)$$

The expression in (1) includes the assumption on the system on the left and the hypothetical system configuration that provided the target on the right. For simplicity, in the current scenario it is assumed that $h(\boldsymbol{x}(\boldsymbol{\theta})) = \boldsymbol{x}(\boldsymbol{\theta})$ and thus in general $\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\theta}) + v$. As it will be apparent from the analysis included section 3, the method is not restricted by this assumption.

The state of the system will typically have the form $\{\hat{\boldsymbol{x}}(t) \in R^n, t = \{1, 2, .., N\}\}$ and it will be an n dimensional vector representing the participating species concentration while $\{\hat{\boldsymbol{y}}(t) \in R^n, t = \{1, 2, .., N\}\}$ is the n dimensional model measured output. The vectors $\{\boldsymbol{x}(t), \boldsymbol{y}(t) \in R^n, t = \{1, 2, .., N\}\}$ represent the hypothetical true state and observed vector. The function $f(.) : R^n \rightarrow R^n$ is the systems' deterministic transition equation and it represents the prior information that is available regarding the model. The initial conditions $\boldsymbol{x_0}$ for the system are assumed unknown and part of the parameter estimation process.

The residual vector can be formulated as :

$$\boldsymbol{\epsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}(\hat{\boldsymbol{x}}(\hat{\boldsymbol{\theta}})) \qquad (2)$$

and $[\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}] \in \boldsymbol{R}^{p+n}$, where p the number of parameters representing the systems' kinetic constants and n represents the additional parameters for the initial conditions of the system $\hat{\boldsymbol{x}}(0)$.

The stochastic modeling problem assumed here, is to estimate the optimal parameters $\hat{\boldsymbol{\theta}}$ based on minimizing the entropy of the joint residual PDF in (2), having inaccurate target observations $\boldsymbol{y}$ stemming from the inclusion of state and measurement noise. Figure 1 provides a schematic representation of the method.

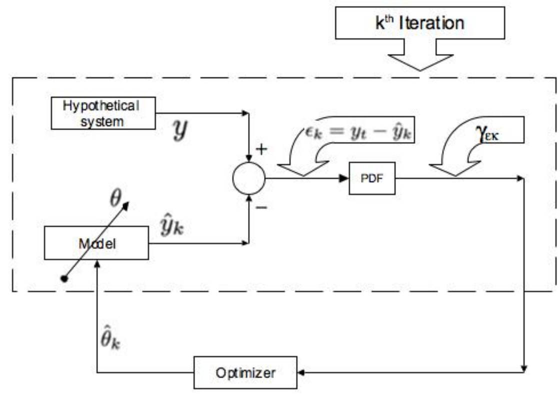

Fig. 1. The basic schematic representation of the entropy minimization process. This is a nonlinear optimization problem where k is the iteration identifier. On each run/simulation the optimizer feeds an optimal parameter vector $\hat{\theta_k}$ so as the resultant residual PDF γ_{ϵ_k} to have as minimum entropy as possible.

2.2 Modeling Assumptions

The assumptions for the current modeling framework are stated as follows:

(1) For the kinetic model to describe the biochemical network it is imperative that it behaves as a well-stirred reactor Szallasi et al. [2006] Wolkenhauer [April, 2005], meaning that there are no other compartments within the volume that the biochemical reaction network resides.
(2) The system is considered in a thermodynamic equilibrium resulting in an isochore and isothermal system Wolkenhauer [April, 2005] where there are no significant fluctuations in either the volume or the temperature of the system.
(3) The system is considered as infinitely expanding Wolkenhauer [April, 2005]. This assumption is necessary in order to avoid any surface effects Wolkenhauer [April, 2005] that might influence the modelling process.

Assumption 1 is necessary for the biochemical reaction system to be adequately described by a ODE system instead of a Partial Differential Equation model that would be the case if there were interactions between different compartments within the system. In addition, changes in temperature or volume (assumption 2) also affect the modeling process since it directly influences the species production. Surface effects (assumption 3) could also influence the representation of the system since they are essentially separate reaction systems in the surface of the volume where there might be differences in temperature and particle density.

3. INFORMATION THEORETIC JUSTIFICATION

This analysis aims at illustrating that the maximum likelihood solution for the general case of non Gaussian noise sources, is indeed the minimum entropy one. This is carried out for both state and measurement noise. In the case where the noise sources are of Gaussian nature then the minimum entropy solution coincides to the maximum likelihood solution (least squares) as also proven by Yue and Wang [2003].

Consider the general system formulation presented in (1), then the residual given in (2) can be further expressed as:

$$\boldsymbol{\epsilon} = \boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta})) + v = g(\boldsymbol{x}(\boldsymbol{\theta})) + v \tag{3}$$

where the nature of $g(.)$ depends on whether there is state noise in the system. The following analysis will show that the Minimum Entropy method has the same, uncertainty reducing, effect whether state noise is included or not. In the following sections, the two cases of a) measurement noise and b) measurement and state noise, will be examined from the information theory perspective so as to observe how the Minimum Entropy method corresponds to a system of minimized uncertainty and how it will effect the PDFs of the residual and the noise components involved.

Previous work (Papadopoulos and Brown [2008]) has been focused on the first case while the current formulation aims at proving the flexibility of the minimum entropy method by including state noise as well.

3.1 Measurement Noise

If no state noise is included then $g(\boldsymbol{x}(\boldsymbol{\theta}))$ is deterministic and the only random variable in the system is the measurement noise component v. From (3) it is possible to infer on the nature of the PDF of $\boldsymbol{\epsilon}$, $f_E(.)$ since it is linearly related with v and this is given by:

$$f_E(\epsilon) = f_V(\epsilon - g(\boldsymbol{x}(\boldsymbol{\theta}))) \tag{4}$$

where $f_V(.)$ is the PDF of the random variable v. Equation (4) is a simple linear transformation (Papoulis and Pillai [2002]) and it shows that the residual will have a similar PDF with that of the noise component but centered according to the constant component $g(.)$.

Let the PDF of a general instantiation of the output y corresponding to any given parameter value $\boldsymbol{\theta}$, be denoted as $f(\boldsymbol{y}|\boldsymbol{\theta})$, then, as also pointed out by Akaike [1973] and also included in Ta and DeBrunner [2004], obtaining the optimal parameters $\boldsymbol{\theta}^*$ corresponds to maximization of the log likelihood function:

$$\hat{\boldsymbol{\theta}}_{ML} = argmax_{\boldsymbol{\theta}}\left(\sum_{i=1}^{N} \log f(y_i|\boldsymbol{\theta})\right) \tag{5}$$

and is linked with minimization of the K-L distance from $f(\boldsymbol{y}|\boldsymbol{\theta})$ to $f(\boldsymbol{y}|\boldsymbol{\theta}^*)$. The K-L distance can be formed as:

$$D = \int f(\boldsymbol{y}|\boldsymbol{\theta}^*) \log \frac{f(\boldsymbol{y}|\boldsymbol{\theta}^*)}{f(\boldsymbol{y}|\boldsymbol{\theta})} d\boldsymbol{y} \tag{6}$$

since the only random variable in the systems is v, $f(\boldsymbol{y}|\boldsymbol{\theta})$ can be given as:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = f_V(\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta}))) \tag{7}$$

substituting (7) into (6) the expression for D can be rewritten as:

$$D = \int f_V(\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta}^*))) \cdot \log \frac{f_V(\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta}^*)))}{f_V(v + g(\boldsymbol{x}(\boldsymbol{\theta})))} d\boldsymbol{y}$$

Noting that:

$$H(v) = \int f_V(v) \log f_V(v) dv$$

the expression for D simplifies to:

$$D = -H(v) - \int f_V(v) \log f_V(v + g(\boldsymbol{x}(\boldsymbol{\theta}))) dv \tag{8}$$

changing variables $v \leftarrow v + g(\boldsymbol{x}(\boldsymbol{\theta}))$, (8), the expression for D simplifies to:

$$D = -H(v) - \int f_V(v - g(\boldsymbol{x}(\boldsymbol{\theta}))) \log f_V(v) dv \tag{9}$$

At this point it is necessary to take into account the linear transformation that exists between the two variables ϵ and v. Since in general $y_0 = T(x_0) \implies x_0 = T^{-1}(y_0)$ and since $T(\cdot)$ in the present case is a linear transformation the following expression holds (Peebles [1993]):

$$f_E(e) = f_V(e - g(\boldsymbol{x}(\boldsymbol{\theta}))) \implies f_V(e) = f_E(\epsilon + g(\boldsymbol{x}(\boldsymbol{\theta}))) \tag{10}$$

combining (10) and (9) yields:

$$\begin{aligned} D &= -H(v) - \int f_E(v) \log f_V(v) dv \implies \\ &= -H(v) - \int f_E(v) \log \frac{f_E(v)}{f_V(v)} dv + H(\epsilon) \end{aligned} \tag{11}$$

since in general the entropy is defined as $H(\epsilon) = \int f_E(p) \log f_E(p) dp$ and the integral is independent of the variable. Equation (11) simplifies to :

$$D = D(f_E \| f_V) - H(\epsilon) + H(v) \tag{12}$$

where $\|$ denotes the K-L distance. From (12) and based on the definition of the residual given in (3) one can express the mutual information, $I(\epsilon, g(\cdot))$, encapsulated in the residual as:

$$I(\epsilon, g(\cdot)) = H(\epsilon) - H(\epsilon|g(\cdot)) = H(\epsilon) - H(v) \tag{13}$$

and by combining (13) and (12) results in:

$$D = D(f_E \| f_V) + I(\epsilon, g(\cdot)) \tag{14}$$

Equation (14) shows that by minimizing the distance D is to get $f_E(\cdot)$ as close as possible to $f_V(\cdot)$ while making the two random variables ϵ and v as uncorrelated as possible. Since both quantities $D(f_E\|f_V)$ and $I(\epsilon, g(\cdot))$ are positive, minimizing one will lead to minimization of the other.

In addition, since $f_V(\cdot)$ is not available it is not possible to minimize $D(f_E\|f_V)$ directly. The quantity $I(\epsilon, g(\cdot))$ however is controlled only by $H(\epsilon)$ since $H(v)$ is constant throughout the process. Therefore, minimization of $H(\epsilon)$ corresponds to the maximum likelihood solution for the general case of noise components and uncertainty reduction.

3.2 Measurement and State Noise

When state noise is included in the system then the residual as formed in (3) is jointly dependent on both the state and measurement noise components. Due to the presence of the sate noise component w, $\boldsymbol{x}(\boldsymbol{\theta})$ is actually a random variable with PDF $f_X(\cdot)$ and in general the residual $\boldsymbol{\epsilon}$ is also a function of both $\boldsymbol{x}$ and v with PDF $f_V(\cdot)$ respectively. Since the noise components are uncorrelated the residual PDF can have the following form:

$$f_E(\epsilon) = \int f_X(\boldsymbol{x}(\theta)) \cdot f_V(\epsilon - g(\boldsymbol{x}(\boldsymbol{\theta})))d\boldsymbol{x} \quad (15)$$

In (15), if both $f_X(\cdot)$ and $f_V(\cdot)$ were known then the usual Maximum Likelihood (ML) approach could also be used here, however, since in most cases these components will not be available, one can again use the same framework as in (6) only using $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ instead of just $f(\boldsymbol{y}|\boldsymbol{\theta})$. The corresponding K-L distance now has the form:

$$D = \int f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}^*) \log \frac{f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}^*)}{f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})} d\boldsymbol{x}d\boldsymbol{y} \quad (16)$$

due to the independence of w and v, $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ can be given as:

$$f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) = f_X(\boldsymbol{x}(\boldsymbol{\theta})) \cdot f_V(\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta}))) \quad (17)$$

and following a similar course as previously, by substituting (7) into (16) the expression for D can be rewritten as:

$$D = \int f_X(\boldsymbol{x}(\boldsymbol{\theta})) \cdot f_V(\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta}^*))) \cdot \\ \cdot \log \frac{f_V(\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta}^*))}{f_V(\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta}))} d\boldsymbol{x}d\boldsymbol{y}$$

Noting that:

$$H(v) = \int f_V(v) \log f_V(v) dv \text{ and also } \int f_X(x)dx = 1$$

the expression for D simplifies to:

$$D = -H(v) - \int f_X(\boldsymbol{x}(\boldsymbol{\theta})) \cdot f_V(v) \log f_V(\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta}))d\boldsymbol{x}dv \quad (18)$$

At this point it is important to note that $f_V(\boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta}))$ depends on v hence one can change variables for $v \leftarrow \boldsymbol{y} - h(\boldsymbol{x}(\boldsymbol{\theta}))$. After substituting, equation (18) becomes:

$$\begin{aligned} D &= -H(v) - \int f_X(\boldsymbol{x}(\boldsymbol{\theta})) \cdot f_V(v) \log f_V(v) d\boldsymbol{x}dv \\ &= -H(v) + \int f_E(v) \log f_V(v) dv \\ &= -H(v) + \int f_E(v) \log \frac{f_E(v)}{f_V(v)} dv + H(\epsilon) \\ &= -H(v) + D(f_E \| f_V) + H(\epsilon) \end{aligned} \quad (19)$$

where $D(f_E \| f_V)$ in (19) denotes the K-L distance between $f_E(\cdot)$ and $f_V(\cdot)$. Setting $k = g(\boldsymbol{x}(\boldsymbol{\theta}))$ and using basic entropy theory, the mutual information between $\boldsymbol{\epsilon}$ and k can be expressed as:

$$I(\epsilon, k) = H(\epsilon) - H(\epsilon|k) \quad (20)$$

where $H(\epsilon|k)$ denotes the conditional entropy of ϵ given k, or more accurately, the entropy of the residual having observed the value of the state. However, knowing that $\epsilon = k + v$, the conditional entropy can be expressed as:

$$\begin{aligned} H(\epsilon|k) &= \int f(k + v) \log f(k + v) d\epsilon \\ &= \int f(v) \log f(v) dv \\ &= H(v) \end{aligned} \quad (21)$$

since the entropy is translation invariant. Utilizing (20) and (21), one can express (19) as:

$$D = D(f_E \| f_V) + I(\epsilon, k) \quad (22)$$

It can be seen by observing equations (14) and (22) that indeed the Minimum Entropy method can be equally applied in both cases of state and measurement noise. The same conclusions as stated in the previous section hold here as well with the only difference being the nature of the components k and $g(\cdot)$ contained in the mutual information for each case. It can be concluded that, since the conditional entropy $H(\epsilon|k)$ or $H(\epsilon|g(\cdot))$ will always assume either k or $g(\cdot)$ as given, the mutual information component is heavily dependent on the entropy measure $H(\epsilon)$ only, and indeed in both cases minimizing the entropy of the residual results in minimization of the systems uncertainty.

4. MINIMUM ENTROPY PARAMETER ESTIMATION

4.1 Introduction

In cases where the uncertainty in a stochastic system is generally non-Gaussian then the variance can no longer be used and a more general measure of uncertainty has to be employed (Papoulis and Pillai [2002]). The proposed novel stochastic parameter estimation approach is designed for a multivariate environment, focusing on minimizing the joint residual PDF $\hat{\gamma}_\epsilon(\boldsymbol{\epsilon}|\boldsymbol{\theta})$ between the target response y and the model output $\hat{y}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\theta}})$ as presented in (1). The joint residual PDF is formed by all the outputs of the system taking into account state and measurement noise.

4.2 Minimum Entropy Parameter Estimation: Top Level Algorithm

The schematic representation of the method is depicted in figure 1. At the top level the algorithm is structured as follows:

1. Initialization stage: Obtain the target vector $\boldsymbol{y}$ and set the initial value of the parameter vector $\boldsymbol{\theta_0}$ for the optimization procedure.
2. Feed $\boldsymbol{\theta_0}$ into the optimizer which outputs the next iteration $\hat{\boldsymbol{\theta}}$.
3. The parameter estimate is passed to the model and after solving the ODE system the new $\hat{\boldsymbol{y}}(\hat{\boldsymbol{\theta}})$ is created.
4. $\hat{\boldsymbol{y}}(\hat{\boldsymbol{\theta}})$ is passed into the system and the residual is formulated and its PDF $\hat{\gamma}_\epsilon(\ldots)$ approximated using KDE.
5. The residual PDF is now passed to the optimizer where the nonlinear optimization problem in (23) is solved and a new optimal $\hat{\boldsymbol{\theta}}$ is estimated.
6. If $\hat{\boldsymbol{\theta}}$ is acceptable, terminate, otherwise go to 2

Steps 4 and 5 are the most important in the process. In step 4 the residual PDF is approximated using KDE and gaussian kernels centered on each data point (more on that in section 4.5). Step 5 is the optimization stage where the simplex method is used to derive the optimal $\boldsymbol{\theta}$ for the current iteration (section 4.4).

4.3 Minimum Entropy Parameter Estimation: Assumptions

The assumptions are formulated as follows:

1. The system is assumed to be bounded in a given interval $[\alpha, \beta]$.
2. The *residual PDF* γ_e is assumed to be measurable and differentiable in all its arguments within the interval $[\alpha, \beta]$.
3. The system is assumed to be Observable and Identifiable in $[\alpha, \beta]$.
4. Both state and measurement noise components, w and v respectively, are assumed to be identically and independently distributed with unknown PDF structure.

(5) The state variable $\boldsymbol{x}$ is sufficiently varied for the optimal parameters $\boldsymbol{\theta}$ to be estimated.

Assumption 2 is there to assure continuity of the function since this is necessary for the optimization process. Observability (assumption 3) is necessary for the simulation of the SDE system. Usually having presented with a set of initial conditions the system is assumed observable, however in the current case the initial conditions are assumed unknown and are part of the optimization process, for this reason this assumptions is really there to ensure that a set of initial conditions does exist. In linear state space models this condition should be sufficient for identifiability however the same cannot be said for nonlinear systems.

Identifiability means that for a given set of parameters the result is unique. When the system in question is linear in the parameters then this assumption means that in the whole of the time span used for simulation say $[t_1, t_2]$ the composite matrix of each $f_i(\mathbf{x})$ where i = *time instants* is not singular or it is relatively well conditioned.

4.4 Entropy Minimization

The task of finding an optimal $\boldsymbol{\theta}$ is posed as an optimization problem, that of finding an optimal parameter vector corresponding to the residual component $\boldsymbol{\epsilon}$ having minimum entropy:

$$\begin{aligned} J(\hat{\boldsymbol{\theta}}) &= \int_{\boldsymbol{\epsilon}\in[\alpha,\beta]} \hat{\gamma_e}(\boldsymbol{\epsilon}(\hat{\boldsymbol{\theta}})) \log(\hat{\gamma_e}(\boldsymbol{\epsilon}(\hat{\boldsymbol{\theta}})))d\boldsymbol{\epsilon} \\ &\text{subject to} \int_{\boldsymbol{\epsilon}\in[\alpha,\beta]} \hat{\gamma_e}(\boldsymbol{\epsilon}(\hat{\boldsymbol{\theta}}))d\boldsymbol{\epsilon} = 1 \end{aligned} \tag{23}$$

where $\int_{\boldsymbol{\epsilon}\in[\alpha,\beta]} \hat{\gamma_\epsilon}(\boldsymbol{\epsilon}(\hat{\boldsymbol{\theta}}))d\boldsymbol{\epsilon} = 1$ is there simply to provide the necessary condition for the estimate $\hat{\gamma_e}(\ldots)$ to be a PDF and $\boldsymbol{\epsilon}$ is there to denote all the residual components from 1 up to n. The schematic representation of the method formulation is given in figure 1. Each iteration k, the PDF of the residual (2) is estimated using KDE (section 4.5) and constitutes a run/simulation, based on a given time span, of the ODE dynamic model of the biochemical reaction system.

This is a nonlinear optimization problem and the function $\hat{\gamma_\epsilon}(\boldsymbol{\epsilon}(\hat{\boldsymbol{\theta}}))$ is in general a nonlinear function of $\boldsymbol{\theta}$. For that reason, as well as for the nature of the general ODE model structure in (1), the optimization problem presented in (23) has no closed form solution for the optimal parameter vector $\hat{\boldsymbol{\theta}}$. One could however use the formulation,

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} = \hat{\boldsymbol{\theta}}_{\boldsymbol{k-1}} - \lambda \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\boldsymbol{k-1}}} \tag{24}$$

where $\lambda > 0$ is the learning rate. This is a standard gradient approach which guarantees local convergence assuming that λ is selected sufficiently small. However slow convergence can be a problem when the Hessian is not well conditioned and the difference between the maximum and the minimum eigenvalue is large. From (23) having dealt with the constraint in the PDF estimation stage detailed in section (4.5), it follows that,

$$\frac{\partial J(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \int_{\boldsymbol{\epsilon}\in[\alpha,\beta]} \frac{\partial \hat{\gamma_\epsilon}(\boldsymbol{\epsilon}(\hat{\boldsymbol{\theta}}))}{\partial \boldsymbol{\theta}} \left[1 + \ln \hat{\gamma_\epsilon}(\boldsymbol{\epsilon}(\hat{\boldsymbol{\theta}}))\right] \tag{25}$$

Combining (24) and (25) one can estimate recursively the optimal $\hat{\boldsymbol{\theta}}$ update, in reality however the optimization problem posed in (23) can be solved directly using any nonlinear optimization routine. In this case a variant of the Simplex method was used since it does not require evaluation of the Hessian of (23).

In addition, any indirect search methods like BFGS require approximation of the inverse Hessian to estimate the next parameter update, something that might not always be available. In the case however that the Hessian can be provided, indirect search methods can converge faster. The BFGS in particular has been shown Rao [2006] to have superlinear convergence near the optimal parameter estimate.

4.5 Density Estimation Using KDE

For the general case of multivariable systems, KDE methods offer a very attractive alternative to the usual parametric methods where the assumption of a general PDF model is necessary. KDE does not need any prior knowledge of the system apart form the samples from which the estimate of the PDF is evaluated. The general formulation for the multivariate case used here involves product kernels of the form,

$$\hat{\gamma}(\boldsymbol{x}) = \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{d} \boldsymbol{K}\left(\frac{x_i - x_{ij}}{h_j}\right) \right\} \tag{26}$$

where $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ is the random variable, h_j is the smoothing parameter for each dimension d and the data is assumed to have been collected in an $n \times d$ matrix. Geometrically, the estimate here places a probability mass if $1/n$ centered on each sample point.

The kernel function $K(.)$ has been chosen as an exponential of the form $\exp(-\frac{1}{2}|\boldsymbol{x} - \boldsymbol{x}'|^2)$, in view of the fact that the choice of the smoothing parameter h is far more important Scott [1992] and a smooth, symmetric about the origin kernel would be the most sensible choice since it would produce a continuous PDF estimate. Furthermore, the entropy of the PDF estimate in (26) can be evaluated directly in terms of the kernels used to approximate the PDF without the need to estimate the multidimensional integral presented in (23).

The selection of KDE for the density estimation method was also motivated by the fact that virtually all non parametric methods are asymptotically kernel methods Scott [1992].

5. ASYMPTOTIC DISTRIBUTION OF PARAMETER ESTIMATES.

After the parameter estimation process converges is usually required to infer on the quality of the parameter estimates. This is most commonly done by examining the nature of the distribution of the parameters at optimality, to that end, one must analyze the bias and the variance of the parameters.

Let the general system formulation have a structure as presented in (1), a residual formulation as in (2) and the objective function used for the parameter estimation process as in (23).

At this point it should be mentioned that in the case that the parameter estimates are not normally distributed, the mean and the variance is not sufficient to describe their distribution and since the distributions of he noise sources w, v is not known beforehand it is difficult to infer on their properties. For this reason, in this analysis, it is assumed that the noise components are normally distributed and have zero mean. At optimality the following condition should hold (also included in L.Ljung [1999]):

$$J'(\hat{\boldsymbol{\theta}}_N) \equiv \frac{\partial J(\hat{\boldsymbol{\theta}}_{\boldsymbol{N}})}{\partial \boldsymbol{\theta}_N} = 0 \tag{27}$$

where $\hat{\boldsymbol{\theta}}_N$ denotes the estimate of $\boldsymbol{\theta}$ at sample point N. Using Taylor series to expand (27) around the optimal $\boldsymbol{\theta}^*$ results in:

$$\begin{aligned} &J'(\boldsymbol{\theta}^*) + J''(\boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = 0 \\ &(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = -[J''(\boldsymbol{\theta})]^{-1} J'(\boldsymbol{\theta}^*) \end{aligned} \tag{28}$$

that constitutes an expression for $Bias(\hat{\boldsymbol{\theta}}_N)$. Analyzing the derivative $J'(\boldsymbol{\theta}^*)$ as presented in (25) results in:

$$J'(\boldsymbol{\theta}^*) = \int \frac{\partial \gamma_\epsilon(\boldsymbol{\epsilon}(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\epsilon}} \frac{\partial \boldsymbol{\epsilon}}{\partial \boldsymbol{\theta}^*} \Big[1 + \ln \gamma_\epsilon(\boldsymbol{\epsilon}(\boldsymbol{\theta}^*)) \Big] \tag{29}$$

the expression in equation (29) represents the continuous sum of random variables as $N \to \infty$; since the assumption on their nature is that they are independent it is a direct consequence of the Central Limit Theorem (CLT) that :

$$\int \frac{\partial \gamma_\epsilon(\boldsymbol{\epsilon}(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\epsilon}} \frac{\partial \boldsymbol{\epsilon}}{\partial \boldsymbol{\theta}^*} \Big[1 + \ln \gamma_\epsilon(\boldsymbol{\epsilon}(\boldsymbol{\theta}^*)) \Big] \in AsN(0, Q) \tag{30}$$

and the covariance Q is given by :

$$Q = \lim_{N \to \infty} N \cdot E\{[J'(\boldsymbol{\theta}^*)][J'(\boldsymbol{\theta}^*)]^T\} \tag{31}$$

The result presented in (30) states that, provided independence, the random variable converges to a normal distribution and if the mean of the associated noise components is zero then the normal distribution will have zero mean as well and covariance Q. In addition, it can be proven that (as also included in L.Ljung [1999]) if (30), (31) hold, then the following is true for the bias and the variance of the parameter estimates:

$$\begin{aligned} &\sqrt{N}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \in AsN(0, P_\theta) \\ &P_\theta = [J''(\boldsymbol{\theta}^*)]^{-1} Q [J''(\boldsymbol{\theta}^*)]^{-1} \end{aligned} \tag{32}$$

The second order derivative $J''(\boldsymbol{\theta}^*)$ for the entropy case is given as:

$$\begin{aligned} J''(\boldsymbol{\theta}^*) = &\int \Big\{ \frac{\partial^2 \gamma_\epsilon(\boldsymbol{\epsilon}(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\epsilon}^2} \frac{\partial \boldsymbol{\epsilon}}{\partial \boldsymbol{\theta}^*} + \frac{\partial \gamma_\epsilon(\boldsymbol{\epsilon}(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\epsilon}} \frac{\partial^2 \boldsymbol{\epsilon}}{\partial (\boldsymbol{\theta}^*)^2} \Big\} \cdot \\ &\Big[1 + \ln \gamma_\epsilon(\boldsymbol{\epsilon}(\boldsymbol{\theta}^*)) \Big] d\boldsymbol{\epsilon} + \int \frac{\partial \gamma_\epsilon(\boldsymbol{\epsilon}(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\epsilon}} \frac{\partial \boldsymbol{\epsilon}}{\partial \boldsymbol{\theta}^*} \cdot \\ &\Big[\frac{\partial \gamma_\epsilon(\boldsymbol{\epsilon}(\boldsymbol{\theta}^*))}{\partial \boldsymbol{\epsilon}} \frac{1}{\gamma_\epsilon(\boldsymbol{\epsilon}(\boldsymbol{\theta}^*))} \Big] d\boldsymbol{\epsilon} \end{aligned} \tag{33}$$

This concludes the asymptotic description for the variance and bias of the parameter estimates.

6. THE RKIP REGULATED ERK SIGNALING PATHWAY

6.1 Basic Definitions

The Raf Kinase Inhibitor Protein (RKIP) regulated Extracellular signal Regulated Kinase (ERK) model shown in figure 2 and its mathematical formulation presented in the Appendix, is an 11 state and 11 parameter model and in this version is somewhat restricted in that it only represents the ERK pathway regulated by RKIP. In figure 2 , circles represent the concentration of the states and rectangular blocks the kinetic constants that are used as parameters. If a block contains two parameters they correspond to the forward and reverse direction of the reaction. A more detailed description is presented in Cho et al. [2003].

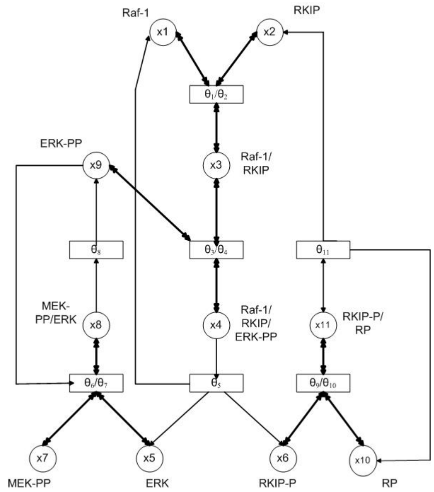

Fig. 2. The RKIP regulated ERK signal transduction pathway having 11 states and 11 parameters.

The mathematical ODE model of the ERK pathway (the function $f(\cdot)$ is the RHS of the model) is provided in the Appendix section and it constitutes the formal representation of the 11 states and 11 parameters state space continuous model where $x_i, i = \{1, 2, ..., 11\}$ represent each species concentration and $\theta_i, i = \{1, 2, ..., 11\}$ the parameters. The function $h(\cdot)$ representing the combination of the states that are observed is simply $\hat{\boldsymbol{y}} = \hat{\boldsymbol{x}}(\hat{\boldsymbol{\theta}})$ as also stated in section (2.1).

6.2 Parameter Estimation Using Minimum Entropy

The method was applied to the RKIP regulated ERK signaling pathway which is an 11 state and 11 parameter system. The schematic representation of the systems is given in figure (2) and the deterministic model is included in the Appendix in (35). The nominal values for the initial species concentrations are presented in table (2) and the initial values for the parameters in table (3).

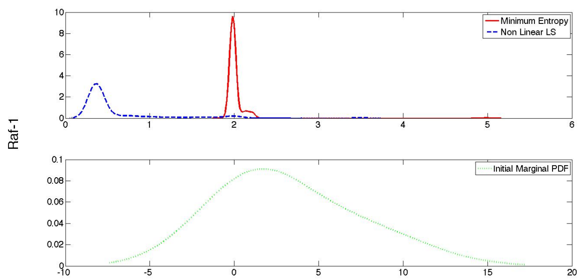

Fig. 3. The marginal PDF of the Raf-1 species.

To illustrate the validity of the method a comparison was drawn with the broadly known non linear least squares (LS) that is optimal when Gaussian noise is included. In the present scenario the noise components w, v were selected as being of Poissonian nature with a PDF as in (34) with λ residing in the interval $(0, 1]$ and α representing the number of occurrences of an event. An example of the marginal PDF included in the target response in the species Raf-1 is depicted in figure 5. The estimated parameter values are presented in table (1).

Figures 3 and 4 depict the differences in the marginal residual PDF for the species Raf-1 and Raf-1/RKIP respectively for

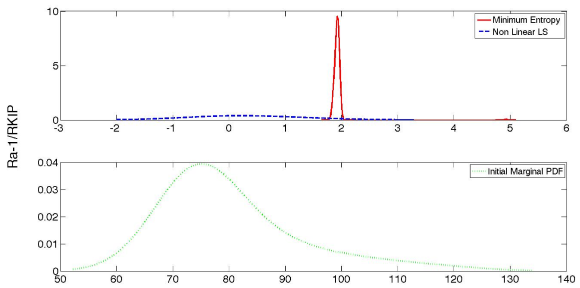

Fig. 4. The marginal PDF of the species Raf-1/RKIP.

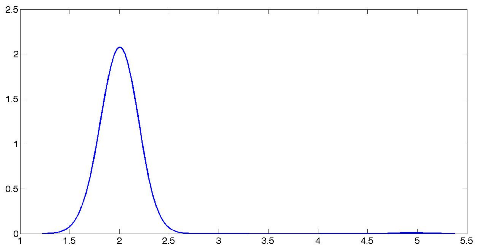

Fig. 5. This is an example of a marginal Poissonian PDF with $\lambda = 0.1$.

Table 1. Estimated parameter values

Parameter	Estimated Value	Parameter	Estimated Value
$\theta 1$	0.53	$\theta 7$	0.00769
$\theta 2$	0.732	$\theta 8$	0.0736
$\theta 3$	0.625	$\theta 9$	0.9189
$\theta 4$	0.00248	$\theta 10$	0.001213
$\theta 5$	0.326	$\theta 11$	0.878
$\theta 6$	0.799		

both LS (blue dotted line) and the Minimum Entropy method (solid red line), both plots include a representation of the initial marginal PDF of the residual at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. It can be seen that the the residual marginals obtained with the LS method are more widely distributed indicating a higher degree of uncertainty while the Minimum Entropy counterpart is more narrow and closely resembling the marginal Poissonian originally included.

In addition, the figure (6) illustrate how the Entropy is minimized as the optimization process evolves. It can be seen that after sufficient number of iterations is allowed the Entropy converges to the optimal one.

$$\gamma_p(w) = 3\left[e^{-\lambda}\sum_{\alpha=0}^{\infty}\frac{\lambda^{-\alpha}}{\alpha!}\delta(w-\alpha)\right] + 2 \tag{34}$$

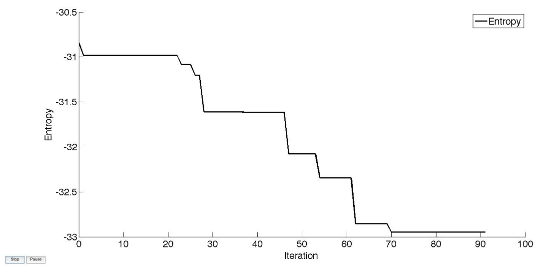

Fig. 6. A schematic representation of how the entropy value changes as a function of the optimizer iterations.

7. CONCLUSIONS

The inherent stochasticity of the cellular mechanism suggests that both state and measurement noise in biochemical reaction pathways is not always of Gaussian nature. For this reason the parameter estimation method has to robust enough to be used for a wider variety of noise distributions.

In this paper a novel stochastic parameter estimation process has been developed for general, dynamic, multivariable stochastic systems representing the biochemical process governing the cell where the noise is not necessarily of Gaussian nature. The current formulation of the method assumes both state and measurement noise and focuses on minimizing the joint PDF of the residual between the model and the given target response. After the residual is formed its PDF is approximated using well known KDE methods and the optimization process converges to those parameters that correspond to the residual having minimum entropy. The general assumption for the development of the method do not include any structure on either the state or the measurement noise components.

Future work will focus on how the minimum entropy method can be extended so as to include additional regularization parameters to influence the models complexity. The study will aim in determining how the addition of a 1-norm (Brown and Costen [2005]) imposed on the parameters can influence the sparseness in the parameter space, leading to a sparse stochastic parameter estimation scenario where a parameter will be included based on its influence on the entropy measurement.

APPENDIX

The mathematic formulation of the RKIP regulated ERK model is given in (35) along with the set of initial conditions for the ODE solver in Table 2 and the initial values for the parameters in the optimization process in Table 3

$$\begin{aligned}
\frac{dx_1}{dt} &= -\theta_1 x_1(t)x_2(t) + \theta_2 x_3(t) + \theta_5 x_4(t)\\
\frac{dx_2}{dt} &= -\theta_1 x_1(t)x_2(t) + \theta_2 x_3(t) + \theta_1 1 x_1 1(t)\\
\frac{dx_3}{dt} &= \theta_1 x_1(t)x_2(t) - \theta_2 x_3(t) + \theta_3 x_3(t)x_9(t) + \theta_4 x_4(t)\\
\frac{dx_4}{dt} &= \theta_3 x_3(t)x_9(t) - \theta_4 x_4(t) - \theta_5 x_4(t)\\
\frac{dx_5}{dt} &= \theta_5 x_4(t) - \theta_6 x_5(t)x_7(t) - \theta_7 x_8(t)\\
\frac{dx_6}{dt} &= \theta_5 x_4(t) - \theta_9 x_6(t)x_{10}(t) - \theta_{10} x_{11}(t)\\
\frac{dx_7}{dt} &= -\theta_6 x_5(t)x_7(t) + \theta_2 x_8(t) + \theta_8 x_8(t)\\
\frac{dx_8}{dt} &= \theta_6 x_5(t)x_7(t) - \theta_7 x_8(t) - \theta_8 x_8(t)\\
\frac{dx_9}{dt} &= -\theta_3 x_3(t)x_9(t) + \theta_4 x_4(t) + \theta_8 x_8(t)\\
\frac{dx_{10}}{dt} &= -\theta_9 x_6(t)x_{10}(t) + \theta_{10} x_{11}(t) + \theta_{11} x_{11}(t)\\
\frac{dx_{11}}{dt} &= \theta_9 x_5(t)x_{10}(t) - \theta_{10} x_{11}(t) - \theta_{11} x_{11}(t)
\end{aligned} \tag{35}$$

Table 2. Initial values for the species concentrations for the ERK ODE system.

Species	Initial Value	Species	Initial Value
$x1$	121.2	$x7$	88.67
$x2$	48.48	$x8$	3.737
$x3$	1.515	$x9$	242.602
$x4$	4.848	$x10$	162.105
$x5$	12.524	$x11$	2.727
$x6$	4.848		

Table 3. Initial parameter values for the optimization process

Parameter	Initial Value	Parameter	Initial Value
$\theta 1$	0.5353	$\theta 7$	0.0076
$\theta 2$	0.0073	$\theta 8$	0.0717
$\theta 3$	0.6312	$\theta 9$	0.9292
$\theta 4$	0.0025	$\theta 10$	0.0012
$\theta 5$	00318	$\theta 11$	0.8787
$\theta 6$	0.8080		

REFERENCES

H Akaike. Information theory and an extention of the maximum likelihood principle. *Proc, 2nd Symposium on Information Theory*, pages 267–281, 1973.

M. Brown and N. Costen. Exploratory basis pursuit regularization. *Pattern Recognition Letters*, 26(12):1907–1915, 2005.

H. Y. Chan and M. A. Kouritzin. Particle filters for combined state and parameter estimation. *Proc. of SPIE*, 4380:244–252, 2001.

K H Cho, S Y Shin, W Kolch, O Wolkenhauer, and B McFerran. *Mathematical Modeling of the influence of RKIP on the ERK signaling pathway*. Proceedings of the 1st international workshop on computational methods in Systems Biology, 2003.

A Doucet, N de freitas, and N Gordon. *Sequential Monte Carlo Methods in Practice*. Springer Verlag, 2001.

L.Ljung. *Systems Identification: Theory for the User*. Prentice Hall, 1999.

G. Papadopoulos and M. Brown. Minimum entropy parameter estimation: Application to the rkip regulated erk signaling pathway. *IEEE World Congress on Computational Intelligence*, Accepted for publication, 2008.

A Papoulis and U Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw Hill, fourth edition, 2002.

P.Z Jr Peebles. *Probability, Random Variables and Random Signal Principles*. McGraw -Hill, 1993.

G Poyiadijis, A Doucet, and S S Singh. Maximum likelihood parameter estimation in general state space models using particle filters. *Ann. of the Inst. of Stat. Math.*, 55(2):409–422, 2003.

S. Rao. *Engineering Optimization: Theory and Practice*. New Age International Publishers, revised third edition, 2006.

G Sanguinetti, N D Lawrence, and M Rattray. Probabilistic inference of trancription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22(22):2775–2781, 2006a.

G Sanguinetti, M Rattray, and N D Lawrence. A probabilistic dynamical model for quantitative inference of the regulatory mechanism of trascription. *Bioinformatics*, 22(14):1753–1759, 2006b.

D W Scott. *Multivariate Density Estimation*. Wiley and sons, 1992.

Z Szallasi, J Stelling, and V Periwal. *System Modelling in Cell Biology*. MIT Press, 2006.

M Ta and V DeBrunner. Minimum entropy estimation as a near maximum-likelihood method and its application in systems identification with non-gaussian noise. *ICASSP*, 2:545–8, May 2004.

A Wang and H Wang. Minimizing entropy and mean tracking control for affine nonlinear and non-gaussian dynamic systems. *IEE proc. Control Theory Appl.*, 151(4), July 2004.

H Wang. Minimum entropy control of non-gaussian dynamic stochastic systems. *Automatic Control, IEEE transactions on,*, 47(2):398–403, 202.

H Wang. *Bounded Dynamic Stochastic Systems: Modelling and Control*. Springer Verlag, 2000.

D.J. Wilkinson. *Stochastic Modeling for Systems Biology*. Chapman and Hall/CRC, 2006.

O Wolkenhauer. *Systems Biology: Dynamic Pathway Modelling*. Draft Manuscript, April, 2005.

H Yue and H Wang. Minimum entropy control of closed loop tracking errors for dynamic stochastic systems. *Automatic Control, IEEE transactions on,*, 48(1):118–122, 2003.

H. Yue, J. Zhou, and H. Wang. Minimum entropy of b-spline pdf systems with mean constraint. *Automatica*, 42:989–994, Feb 2006.