Proceedings of the 9th International Symposium on
Dynamics and Control of Process Systems (DYCOPS 2010),
Leuven, Belgium, July 5-7, 2010
Mayuresh Kothare, Moses Tade, Alain Vande Wouwer, Ilse Smets (Eds.)

Nonlinear System Identification from Small Data Sets

R. Bhushan Gopaluni [*] **Devin Marshman** [*]

[*] *Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, Canada V6T 1Z3 (Tel: 604 827 5668; e-mail: bhushan.gopaluni@ubc.ca, dmarshman@chbe.ubc.ca).*

Keywords: Nonlinear Systems, Maximum Likelihood Parameter Estimation, Compressed Sensing.

Abstract: We propose a novel algorithm for identification of structured nonlinear systems using a compressive sampling approach. Compressive sampling is an approach to reconstruct randomly sampled signals from small data sets. The proposed algorithm provides empirical evidence to suggest that parameters can potentially be estimated from small data sets using compressive sampling. This approach is illustrated through a simulated example.

1. INTRODUCTION

Nonlinear system identification has been an active area of research for many years (Ljung [1999], Haber and Keviczky [1999], Bates and Watts [1998], Gopaluni [2008], Varziri et al. [2008]). A majority of these identification techniques rely on some form of maximum likelihood estimation (MLE). It is well known that for consistency and minimum variance of parameter estimates, maximum likelihood approaches require large amounts of data. For small data sets, MLE often leads to biassed parameter estimates (Casella and Berger [1990]). A recently developed sampling and signal reconstruction paradigm called compressive sampling (Candes and Wakin [2005]) can potentially be used to estimate reliable parameters from very small data sets.

Let us consider a commonly used representation of nonlinear state space models with linear measurement equation

$$\begin{aligned} x_{t+1} &= f(x_t, u_t, \theta) + w_t \\ y_t &= x_t + v_t \end{aligned} \tag{1}$$

where $x_t \in \mathbf{R}^{n\times 1}$ is the n-dimensional state vector, $u_t \in \mathbf{R}^{s\times 1}$ is the s-dimensional input vector, $y_t \in \mathbf{R}^{m\times 1}$ is the m-dimensional output or measurement vector, and w_t, v_t are independent and identically distributed Gaussian noise sequences of appropriate dimension and variances Q and R respectively, $\theta \in \mathbf{R}^{p\times 1}$ is a p-dimensional parameter vector and $f(.)$ is some nonlinear function that describes the dynamics of the process. The subscript t in the above variables indicates the time. The nonlinear function $f(.)$ is typically obtained using physical laws such as energy and mass balance expressions for the process. In some processes, due to their complexity, it is difficult to develop accurate and reliable nonlinear functions. In such processes, either unstructured approximations of $f(.)$ or an input-output model of the following form is used,

$$y_t = h(y_{t-1}, y_{t-2}, \cdots, y_{t-d_y}, u_{t-1}, \cdots, u_{t-d_u}) + v_t \tag{2}$$

where d_y and d_u represent the number of past outputs and inputs used in the model. $h(.)$ is some nonlinear function of the inputs and the outputs. Whether a process is given by a structured or an unstructured state-space model or an input-output model, often the functions $f(.)$, and $h(.)$ are approximated using a set of basis functions. This article provides an algorithm for parameter estimation in such models from smaller data sizes than those required in traditional estimation procedures that use some form of least squares.

Identification of input-output models can often be transformed to a problem of curve fitting using a set of basis functions (for instance wavelets, laguerre-volterra models etc.). Typically, input-output models are obtained by choosing a large number of basis functions and then dropping those corresponding to smaller coefficients. The coefficients of the basis functions are often obtained by minimizing the 2-norm of prediction errors. Since the coefficients of significant basis functions are not known a priori, this approach often involves estimating a large number of coefficients while a small portion of them are ultimately used in a model. As is well known, the larger the number of estimated coefficients, the larger should the data set be for good variance properties on the estimated coefficients (or parameters). In other words, a large number of coefficients are estimated despite knowing the fact that only a small portion of them will be used. Using the technique of compressive sampling it is possible to estimate only the significant coefficients and thus avoid the requirement of large number of samples.

While the literature on nonlinear system identification is replete with algorithms for identification of input-output models, there are very few results on identification of nonlinear state-space models (Gopaluni [2008]). Identification of nonlinear state-space models is more challenging due to the hidden states. If the state and measurement dynamics are linear and the noise is Gaussian, then parameters can be estimated using subspace identification methods (Van Overschee and Moor [1996]) or the expectation maximization algorithm (Shumway and Stoffer [2000]). On the other hand, if the state and measurement dynamics are nonlinear and if the noise is non-Gaussian, then approximations of expectation maximization algorithm have to be employed for parameter estimation (Gopaluni [2008], Schön et al. [2006], Goodwin and Agüero [2005]). All these approaches to identification of nonlinear state-space models require large sets of data for consistent estimation of parameters.

In many chemical engineering processes, especially biotechnology processes, only a small number of measurements are available. In this article, we propose a novel algorithm that makes use of small number of samples for parameter estimation and yet is expected to perform as well as standard identification techniques.

This article is divided into following sections: In section 2, a short introduction to the recently developed sampling paradigm called compressive sampling is presented. In section 3, the main algorithm is presented and applied to states-space models. In section 4, illustrative examples are presented and the article is concluded with a summary in section 5.

2. COMPRESSIVE SAMPLING

Compressive sampling or compressed sensing is a recently developed sampling paradigm that allows one to reconstruct a time series signal from a small fraction of samples. Consider a time series signal $\{y_t\}$ for $t = \{1, \cdots, T\}$ where y_t represents the signal value at time t and T is the total number of sample times at which an estimate of the signal, y_t is desired. Let us assume that a set of basis functions, $\{\phi_1(t), \cdots, \phi_N(t)\}$ where N denotes the total number of basis functions, can be used to represent this time signal. Then one can write the time series signal, y_t as a linear combination of basis functions as follows:

$$y_t = \sum_{i=1}^{N} c_i \phi_i(t) \tag{3}$$

where c_i are some constant coefficients and they are denoted by a vector, $c = [c_1 \ \cdots \ c_N]$. In chemical engineering, wavelet basis functions and radial basis functions have been widely used for data compression, signal reconstruction, and black-box modeling. Most time series signals, while dense in time domain, can be represented compactly using some basis functions. In general, these time series signals can be represented by an extremely sparse vector c. In other words, only a small fraction of the constants c_i are nonzero (and significant) and the rest are either zero or close to zero (or insignificant). Typically, we do not know these constant coefficients nor do we know to which basis functions the non-zero constant coefficients correspond to. Hence, the commonly used approach to estimate the coefficients, c_i, is linear least squares.

Given the time series sequence of y_t, it is possible to formulate a linear least squares problem that has a unique solution if $T \geq N$. However, if $T < N$ then it is well known that the set of linear equations (in terms of c_i in (3)) do not have a unique solution. In problems where only a limited amount of data (or measurements) are available, such an under determined system is a common occurrence. The linear least squares objective function to estimate the constant, c, is the 2-norm of the difference between the predicted, y_t, and the actual measurements. This is a quadratic objective function of the following form

$$J_{ls}(c) = \sum_{t=1}^{T} (y_t - \hat{y}_t(c))'((y_t - \hat{y}_t(c)) \tag{4}$$

where $\hat{y}_t(c)$ is the predicted value of the measurement y_t and is a function of the coefficient vector. Now using vector notation, and denoting the vector of measurements with Y and the vector of predictions by $\hat{Y}(c)$, the quadratic objective function in (4) simply becomes the following 2-norm

$$J_{ls}(c) = \|Y - \hat{Y}(c)\|_2^2 \tag{5}$$

J_{ls} is convex and has a unique solution when $T \geq N$. However, if $T < N$, then there is more than one vector, c, that satisfies (3). Minimizing $J_{ls}(c)$ will only provide one such c that is not necessarily the true c. In other words, there is a high chance that the least squares estimate of c will be biased. On the other hand, over the last few years it has been shown (Candes and Wakin [2005]) that if the vector c is sparse (as is usually the case with most physical time series signals), then solving the following 1-norm optimization will result in a unique unbiased sparse solution of c for some $T > S$ where S is of the order of $K \log(N)$ with K being the number of non-zero coefficients in c,

$$\min \ \|c\|_1 \tag{6}$$

$$\text{subject to } Y = \hat{Y}(c) \tag{7}$$

If K is much smaller than N, then S will be much smaller than N and hence it is possible to obtain a unique sparse solution of the vector c using 1-norm instead of the commonly used 2-norm in minimizing the error between actual measurements and the predictions. Moreover, it is easy to see that the above optimization with respect c is a convex problem and hence there are many efficient algorithms for its minimization.

However, not all $T > S$ samples of the signal will provide an unbiased estimate of c. It was shown in Candes and Wakin [2005] that any T randomly sampled measurements will allow us to find an unbiased estimate of c by minimizing (7), if certain conditions on the basis functions are satisfied.

In summary, the idea behind compressive sampling is to sample the time signal completely randomly and then reconstruct it using a 1-norm objective function. If the original time signal can be expressed as a linear combination of basis functions using a sparse vector c, then it is possible to reconstruct time signal from a small randomly sampled data set with a high degree of accuracy. The number of random samples (T) required to reconstruct a signal depend both on the number of non-zero coefficients in c (K) and the total number of basis functions being used (N). Therefore, if a signal from a process can be represented using a sparse c vector and some known basis functions, then a small number of samples would often suffice to reconstruct the signal.

In fact, this result can be extended to treat noisy signals (please see Candes and Wakin [2005] for details). Let us consider a time signal, y_t, that can be expressed as follows:

$$y_t = \sum_{i=1}^{N} c_i \phi_i(t) + v_t \tag{8}$$

Then the vector c can be identified highly accurately by solving the following convex optimization problem,

$$\begin{aligned} &\min \ \|c\|_1 \\ &\text{subject to } \|Y - \hat{Y}(c)\|_2^2 \leq \sigma^2 \end{aligned} \tag{9}$$

where σ^2 is a constant proportional to the covariance of v_t. The optimization problem in (9) is often called *basis pursuit with denoising*. In the next section, we present an algorithm that makes use of this idea behind compressed sensing to identify parameters in nonlinear models of the form described in (1),(2). It must be pointed out that this presentation of compressive sampling is over-simplified. There are other technical constraints that the process of sampling and the basis functions need to satisfy for an accurate estimation of c. In particular, the combination of sampling and basis function "matrices" need to satisfy the so-called *Restricted Isometry Property* for reliable estimation of c. The presentation in this article does not provide theoretical guarantees for accurate reconstruction of c. However, the approach appears promising and we provide empirical evidence to support it.

3. PROPOSED ALGORITHM

The state equation in (1) is nonlinear, while the measurement equation is linear. Since the states are corrupted with noise in the state dynamic equation, even though the measurement equation is linear, it is not possible to estimate the parameters using nonlinear least squares. However, maximum likelihood approaches such as those developed in Gopaluni [2008] can be used for parameter estimation if the structure of the function $f(.)$ is known and if large sets of data are available. These maximum likelihood methods fail if the available data is small compared to the missing data. In such a scenario, a compressed sensing based algorithm, developed below, is expected to provide better estimates than maximum likelihood approaches.

The proposed algorithm for parameter estimation in nonlinear structured state-space models is presented below:

- **Step 1**: Estimate the state sequence x_t assuming that it can be written compactly as follows
$$x_t = \sum_{i=1}^{N} c_i \phi_i(t) \tag{10}$$
by solving the optimization problem
$$\begin{aligned} &\min \ \|c\|_1 \\ &\text{subject to } \|Y - \hat{X}(c)\|_2^2 \leq \sigma^2 \end{aligned} \tag{11}$$
where $\hat{X}(c)$ is a vector of estimated states, $\hat{x}_t$ [1]
- **Step 2**: Once an estimated state sequence is obtained, the parameter vector is obtained by solving the following nonlinear optimization problem
$$\min_{\theta} \sum_{t=1}^{N-1} (\hat{x}_{t+1} - f(\hat{x}_t, u_t, \theta))'(\hat{x}_{t+1} - f(\hat{x}_t, u_t, \theta)) \tag{12}$$

In the first step of the algorithm the idea behind compressed sensing is used to reconstruct the unknown state signal. If this state signal is sparse in a known basis set, then only a small number of random measurements of y_t will suffice to reconstruct the state signal. Once the hidden state is reconstructed, the parameter vector is estimated in the second step using standard nonlinear least squares. If the state-space model is unstructured, an obvious extension of this method can be used. If an input-output model (in (2)) is desired, then the algorithm can be modified as follows,

- **Step 1**: Assume that the function $h(.)$ can be approximated using some basis functions as follows
$$\begin{aligned} y_t = &\sum_{i=1}^{N} c_i \phi_i(y_{t-1}, y_{t-2}, \cdots, y_{t-d_y}, u_{t-1}, \cdots, u_{t-d_u}) \\ &+ v_t \end{aligned}$$
- **Step 2**: The parameter vector, c is obtained by solving the following nonlinear optimization problem
$$\begin{aligned} &\min \ \|c\|_1 \\ &\text{subject to } \|Y - \hat{Y}(c)\|_2^2 \leq \sigma_v^2 \end{aligned} \tag{13}$$
where σ_v^2 is a constant proportional to the output noise variance.

One obvious disadvantage with the above approach is that one has to choose appropriate values for σ^2 and σ_v^2 while in maximum likelihood approaches, the noise and state covariance matrices can automatically be estimated.

[1] Please note the difference between $\hat{x}_t$ and x_t. $\hat{x}_t$ is the estimated state sequence after performing the optimization, while x_t is the state as a function of the vector c.

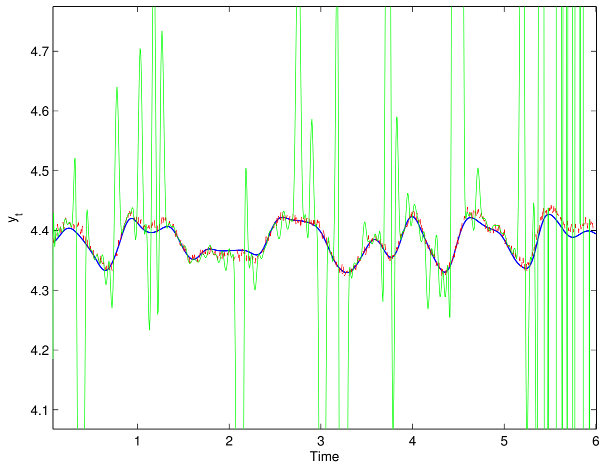

Fig. 1. Example 1: Reconstructed signal. The solid stable line is the reconstructed signal using compressed sensing algorithm, the dotted line is the signal of interest, and the solid unstable line is the least squares reconstructed signal.

4. ILLUSTRATIVE EXAMPLES

4.1 Example 1

The first example is designed to show the power of compressed sensing in reconstructing a signal from a very small data set. A simulated time signal is randomly sampled and reconstructed by solving the optimization problem in (7). A large number of radial basis functions (1000) are used in reconstructing this signal. The original signal shown in figure 1 consists of 1000 samples. However, only about 202 random samples are measured and used to reconstruct the original 1000 sample signal. The signal reconstructed using (7), as seen in figure 1, approximates the original signal very well. On the other hand, the signal reconstructed using (4) is very poor and the predictions of the reconstructed signal are unstable. The instability of the predictions from the model obtained using standard least squares algorithm is due to the ill-conditioned regression matrix (which in turn is due to too few samples and too many basis functions).

4.2 Biological Example

In the second example, a signal transcription pathway that is defined by four nonlinear differential equations and has four parameters is used. The Janus family of kinases (JAK) - signal transducer and activator of transcription (STAT) pathway describes a series of reactions taking place across cytoplasm and nucleus of a cell to trigger the transcription of key genes. The signaling pathway occurs through multiple cell surface receptors, including the erythropoietin receptor (EpoR). EpoR plays an important role in the proliferation and differentiation of erythroid progenitor cells (Swameye et al. [2003]), which refer to cells that are able to grow into a specific type of cell - in this case, red blood cell - through cell-division (Lackie [2007]). Figure 2 shows the diagram of the JAK-STAT signal transduction pathway. Through a series of reactions, EpoR creates docking sites for STAT5, a latent transcription factor. The mathematical model of JAK-STAT signaling pathway was originally developed in Swameye et al. [2003]. There are four state variables which represent the concentrations of unphosphorylated STAT5 (x_1), tyrosine phosphorylated monomeric STAT5 (x_2), tyrosine phosphorylated dimeric STAT5 (x_3) and STAT5 within the nucleus (x_4). The exogenous input variable of the model, $u(t)$, is the concentration of EpoR. The key actions taken by STAT5 are phosphorylation (x_1 to x_2, in figure 2), formation of dimers (x_2 to x_3, in figure 2), and migration from cytoplasm into nucleus (x_3 to x_4, in figure 2). Once present in the nucleus, STAT5 is able to trigger the transcription of target genes. There are several hypotheses for the termination mechanism of JAK-STAT pathway, including degradation of STAT5 within the nucleus and migration of STAT5 from nucleus back to cytoplasm.

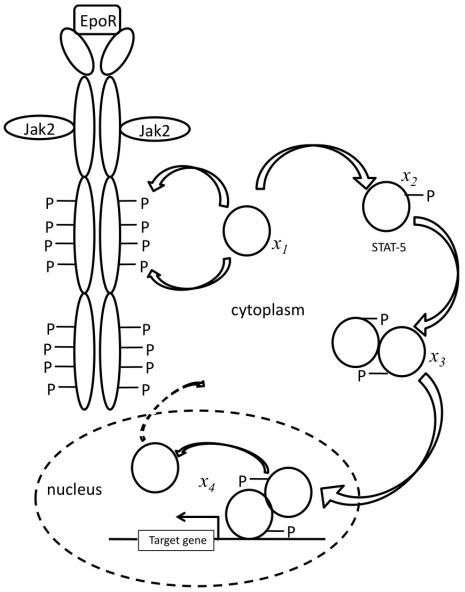

Fig. 2. JAK-STAT signal transduction pathway diagram (Swameye et al. [2003]).

Initially, EpoR creates docking sites for STAT5. This triggers a series of STAT5 reactions where unphosphorylated monomeric STAT5 (x_1) becomes phosphorylated monomeric STAT5 (x_2), which in turn forms phosphorylated dimeric STAT5 (x_3) that migrates into the nucleus. Once inside the nucleus, phosphorylated dimeric STAT5 (x_4) triggers the expression of target gene. The signal transduction pathway terminates, by migration of STAT5 from nucleus back to cytoplasm.

A model for the JAK-STAT signal transduction pathway was adopted following the suggestion in Zi and Klipp [2006], Quach et al. [2007] and expressed as a set of four coupled ordinary differential equations as follows.

$$
\begin{aligned}
\frac{dx_t^{(1)}}{dt} &= -a_1 x_t^{(1)} u_t + 2a_4 x_t^{(4)} I_{\{t\geq\tau\}} + w_t^{(1)},\\
\frac{dx_t^{(2)}}{dt} &= a_1 x_t^{(1)} u_t - 2(x_t^{(2)})^2 + w_t^{(2)},\\
\frac{dx_t^{(3)}}{dt} &= -a_3 x_t^{(3)} + (x_t^{(2)})^2 + w_t^{(3)},\\
\frac{dx_t^{(4)}}{dt} &= a_3 x_t^{(3)} - a_4 x_t^{(4)} I_{\{t\geq\tau\}} + w_t^{(4)}, \qquad (14)
\end{aligned}
$$

where $I_{\{t\geq\tau\}}$ is an indicator function that is equal to zero when $t < \tau$ and is equal to one when $t \geq \tau$ (assumed to 200 units of time in this example). $w^{(i)}$ denotes noise in the ith state equation. a_1, a_3, a_4 are constants whose values from literature are taken to be $0.0515; 3.39; 0.35$ respectively. There is a time delay between the initial addition of EpoR into the system that triggers the activation of STAT5 signal transduction pathway, and the migration of STAT5 into nucleus. τ, in the model, accounts for this time delay. The following output variables are assumed to be measured [2] ,

$$
\begin{aligned}
y_t^{(1)} &= x_t^{(1)} + v_t^{(1)},\\
y_t^{(2)} &= x_t^{(2)} + v_t^{(2)},\\
y_t^{(3)} &= x_t^{(3)} + v_t^{(3)},\\
y_t^{(4)} &= x_t^{(4)} + v_t^{(4)}.
\end{aligned}
\qquad (15)
$$

where $y_t^{(i)}$ and $v_t^{(i)}$ denote the corresponding measurements and the associated noise for state i. 1000 samples are generated from a discretized model of the JAT-STAT differential equation model. Using only 400 randomly chosen samples, the parameters are estimated from the proposed method are found to be $0.0517; 3.3895; 0.3291$. The estimated parameters are close to the true parameters, and hence the reconstructed states and the measurements show a good fit. In figures 3-6 plots of x_t and its reconstructed infinite horizon estimates are shown. The initial state is not estimated and is assumed to be zero while reconstructing the infinite horizon estimates, and hence the infinite horizon estimates take a few samples before the prediction error becomes small.

5. CONCLUSIONS

A new approach to parameter estimation in nonlinear stochastic systems is presented. This method makes use of the idea of compressed sensing to reconstruct the noise corrupted state signal. The reconstructed state signal is then used in a nonlinear optimization problem to estimate the parameter vector. This approach is illustrated through a simulation example.

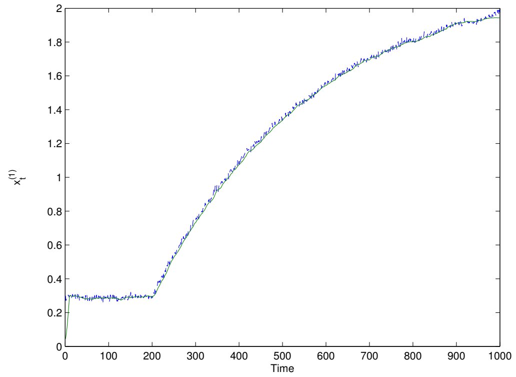

Fig. 3. The smooth line is the reconstructed state, $x_t^{(1)}$, and the noisy signal is the measurement $y_t^{(1)}$.

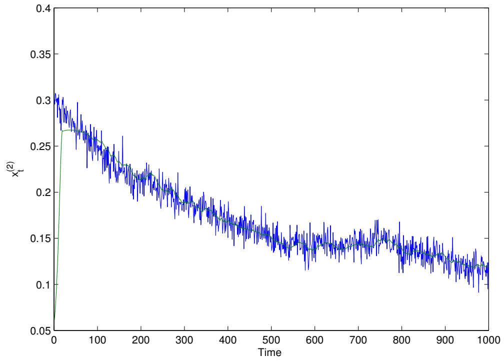

Fig. 4. The smooth line is the reconstructed state, $x_t^{(2)}$, and the noisy signal is the measurement $y_t^{(2)}$.

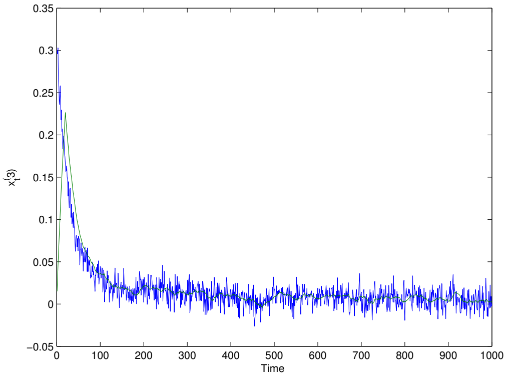

Fig. 5. The smooth line is the reconstructed state, $x_t^{(3)}$, and the noisy signal is the measurement $y_t^{(3)}$.

6. ACKNOWLEDGMENTS

The authors would like to thank Shelly Jang for bringing JAK-STAT model to their attention.

REFERENCES

D.M. Bates and D.G. Watts. *Nonlinear Regression Analysis and its Applications*. John Wiley and Sons, Inc., 1998.

[2] Please note that in practice only combinations of these outputs are measured. However, to illustrate the usefulness of the proposed algorithm all the states are assumed to be measured and arbitrary random binary signal is used as the input.

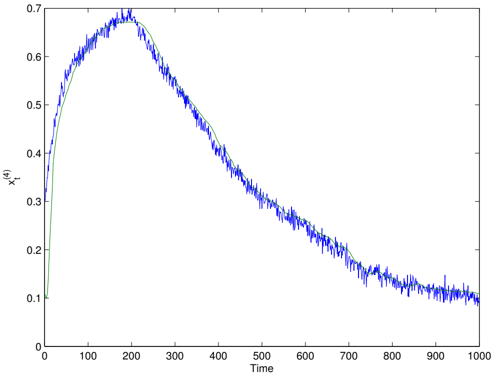

Fig. 6. The smooth line is the reconstructed state, $x_t^{(4)}$, and the noisy signal is the measurement $y_t^{(4)}$.

E.J. Candes and M.B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2005.

G. Casella and R.L. Berger. *Statistical Inference.* Brooks/Cole Publishing Company, Pacific Grove, California, 1990.

G.C. Goodwin and J.C. Agüero. Approximate EM algorithms for paramter and state estimation in nonlinear stochastic models. *Proceedings of IEEE Conference on Decision and Control*, pages 368–373, 2005.

R.B. Gopaluni. A particle filter approach to identification of nonlinear processes under missing observations. *Canadian Journal of Chemical Engineering*, 86(6):1081–1092, 2008.

R. Haber and L. Keviczky. *Nonlinear System Identification: Input-Ouput Modelling Approach.* Kluwer Academic Publishers, 1999.

J. Lackie. *The Dicitonary of Cell and Molecular Biology.* Academic Press, 2007.

L. Ljung. *System Identification: Theory for the user.* Prentice Hall, 1999.

M. Quach, N. Brunel, and F. d'Alche Buc. Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference. *Bioinformatics*, 23(23):3209–3216, 2007.

T.B. Schön, A. Wills, and B. Ninness. Maximum likelihood nonlinear system estimation. In *Proceedings of IFAC Symposium on System Identification*, pages 1003–1008, 2006.

R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications.* Springer, 2000.

I. Swameye, T. G. Muller, J. TImmer, O. Sandra, and U. Klingmuller. Identification of Nucleocytoplasmic Cycling as a Remote Sensor in Cellular Signaling by Databased Modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):1028–1033, 2003.

P. Van Overschee and B. De Moor. *Subspace identification for linear systems.* Kluwer, 1996.

M.S. Varziri, K.B. McAuley, and J.P. McLellan. Parameter estimation in continuous time dynamic models in the presence of unmeasured states and non-stationary disturbances. *Industrial Engineering and Chemistry Research*, 47:380–393, 2008.

Z. Zi and E. Klipp. SBML-PET: a Systems Biology Markup Language-based parameter estimation tool. *Bioinformatics*, 22(21):2704–2705, 2006.