A Sparse Bayesian Approach to the Identification of Nonlinear State-Space Systems

Wei Pan, Ye Yuan, Jorge Gonçalves, and Guy-Bart Stan

***Abstract*—This technical note considers the identification of nonlinear discrete-time systems with additive process noise but without measurement noise. In particular, we propose a method and its associated algorithm to identify the system nonlinear functional forms and their associated parameters from a limited number of time-series data points. For this, we cast this identification problem as a sparse linear regression problem and take a Bayesian viewpoint to solve it. As such, this approach typically leads to nonconvex optimizations. We propose a convexification procedure relying on an efficient iterative re-weighted ℓ_1-minimization algorithm that uses general sparsity inducing priors on the parameters of the system and marginal likelihood maximisation. Using this approach, we also show how convex constraints on the parameters can be easily added to the proposed iterative re-weighted ℓ_1-minimization algorithm. In the supplementary material available online (arXiv:1408.3549), we illustrate the effectiveness of the proposed identification method on two classical systems in biology and physics, namely, a genetic repressilator network and a large scale network of interconnected Kuramoto oscillators.**

***Index Terms*—Nonlinear system identification, re-weighted ℓ_1-minimization, sparse Bayesian learning.**

I. Introduction

Identification from time-series data of nonlinear discrete-time state-space systems with additive process noise is relevant to many different fields such as systems/synthetic biology, econometrics, finance, chemical engineering, social networks, etc. Yet, the development of general identification techniques remains challenging, especially due to the difficulty of adequately identifying nonlinear systems [2], [3]. Nonlinear dynamical system identification aims at recovering the set of nonlinear equations associated with the system from time-series observations. The importance of nonlinear dynamical system identification and its associated difficulties have been widely recognised [3], [4].

Since, typically, nonlinear functional forms can be expanded as sums of terms belonging to a family of parameterised functions (see [2, Sec. 5.4] and [3]), an usual approach to identify nonlinear state-space models is to search amongst a set of possible nonlinear terms (e.g., basis functions) for a parsimonious description coherent with the available data [5]. A few choices of basis functions are provided by classical functional decomposition methods such as Volterra expansion, Taylor polynomial expansion or Fourier series [2], [3], [6]. This is typically used to model systems such as those described by Wiener and Volterra series [6], [7], neural networks [8], nonlinear auto-regressive with exogenous inputs (NARX) models [9], and Hammerstein-Wiener [10] structures, to name just a few examples.

Recently, graphical models have been proposed to capture the structure of nonlinear dynamical networks. In the standard graphical models where each state variable represents a node in the graph and is treated as a random variable, the nonlinear relations among nodes can be characterised by factorising the joint probability distribution according to a certain directed graph [11]–[13]. However, standard graphical models are often not adequate for dealing with times series directly. This is mainly due to two aspects inherent to the construction of graphical models. The first aspect pertains to the efficiency of graphical models built using time series data. In this case, the building of graphical models requires the estimation of conditional distributions with a large number of random variables [14] (each time series is modelled as a finite sequence of random variables), which is typically not efficient. The second aspect pertains to the estimation of the moments of conditional distribution, which is very hard to do with a limited amount of data especially when the system to reconstruct is nonlinear. In the case of linear dynamical systems, the first two moments can sometimes be estimated from limited amount of data [15], [16]. However, higher moments typically need to be estimated if the system under consideration is nonlinear.

In this technical note, we propose a method to alleviate the problems mentioned above. This method relies on the assumption that there exits a finite set of candidate dictionary functions whose linear combination allows to describe the dynamics of the system of interest. In particular, we focus on discrete-time nonlinear systems with additive noise represented in a general state-space form. Based on this, we develop an identification framework that uses time series data and *a priori* knowledge of the type of system from which these time series data have been collected, e.g., biological, biochemical, mechanical or electrical systems. For example in Genetic Regulatory Network (GRN), only polynomial or rational nonlinear functional forms typically need to be considered in the identification process.

To identify the network efficiently given the available time series data, we cast this nonlinear system identification problem as a sparse linear regression problem [17]–[19]. Although such problems have been widely applied in the context of sparse coding, dictionary learning or image processing [20], [21], they have received little attention in nonlinear dynamical system identification. Besides the work presented here, one of the rare example of sparse estimation technique used for dynamical system identification is the multiple kernel-based regularisation method, which has been used to estimate finite impulse response models [22].

Furthermore, very few contributions are available in the literature that address the identification problem with *a priori* information or constraints on the parameters of the system [23], [24]. In contrast, our proposed framework allows us to incorporate convex constraints on the

Manuscript received July 7, 2014; revised November 7, 2014 and March 11, 2015; accepted April 14, 2015. Date of publication April 24, 2015; date of current version December 24, 2015. This work was supported by Microsoft Research, by the EPSRC (project EP/I03210X/1), by FNR and ERASysAPP, by the EPSRC-funded Centre for Synthetic Biology and Innovation at Imperial College London (project EP/G036004/1), and the EPSRC Fellowship for Growth of Dr Stan (project EP/M002187/1). Recommended by Associate Editor W. X. Zheng.

W. Pan and G-B. Stan are with the Centre for Synthetic Biology and Innovation and the Department of Bioengineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: w.pan11@imperial.ac.uk; g.stan@imperial.ac.uk).

Y. Yuan is with the Control Group, Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K. (e-mail: yy311@cam.ac.uk).

J. Gonçalves is with the Control Group, Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K. and the Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg City L-4362, Luxembourg (e-mail: jmg77@cam.ac.uk).

Digital Object Identifier 10.1109/TAC.2015.2426291

associated model parameters, e.g., equality or inequality constraints imposed among parameters, or *a priori* required stability conditions.

In sparse linear regression problems, finding the sparsest solution is desirable but typically NP-hard. The classic "Lasso" or ℓ_1-minimization algorithm are typically used as a relaxation to alleviate this numerical difficulty [25]. However, these algorithms usually only work well or have performance guarantees when the considered dictionary matrix has certain properties such as the *restricted isometry property* (RIP) [18], [26] or the *incoherence* property [27]. Loosely speaking, these properties require that the columns of the dictionary matrix are orthogonal, or nearly so. Unfortunately, such properties are hardly guaranteed for nonlinear identification problems and, as a consequence, ℓ_1-relaxation based algorithms typically do not work well when these conditions are not satisfied.

In this technical note, we shall explain, from a probabilistic viewpoint, how a Bayesian approach can attenuate problems arising in the case of high correlations between columns of the dictionary matrix. In particular, the main contributions of this technical note are:

- To formulate the problem of reconstructing discrete-time nonlinear systems with additive noise into a sparse linear regression problem. The model class in this technical note covers a large range of systems, e.g., systems with multiple inputs and multiple outputs, systems with memory in their states and inputs, and autoregressive models.
- To derive a sparse Bayesian formulation of the nonlinear system identification problem, which is casted into a nonconvex optimization problem.
- To develop an iterative re-weighted ℓ_1-minimization algorithm to convexify the nonconvex optimization problem and solve it efficiently. This formulation can also take into account additional convex constraints on the parameters of the model.

The generality of our framework allows it to be applied on a broad class of nonlinear system identification problems. In particular, to illustrate our results, we applied our approach to two examples: (1) the Genetic Repressilator Network, where we identify nonlinear regulation relationships between genes, transcriptional and translational strengths and degradation rates, and (2) a network of Kuramoto Oscillators, where we identify the network topology and nonlinear coupling functions. Details about these examples can be found in the supplementary material [1].

This technical note is organized as follows. Section II-A introduces the class of nonlinear models considered. Section II-B formulates the nonlinear identification problem into a sparse linear regression problem. Section III re-interprets the sparse problem from a Bayesian point of view, while Section IV shows how the resulting nonconvex optimization problem can be convexified and solved efficiently using an iterative re-weighted ℓ_1-minimization algorithm. Finally, we conclude and discuss several future open questions.

II. Formulation of the Nonlinear Identification Problem

A. Considered Nonlinear Dynamical Model Class

We consider dynamical systems described by discrete-time nonlinear state-space equations driven by additive Gaussian noise. The discrete-time dynamics of the i-th state variable x_i, $i = 1, \ldots, n_\mathbf{x}$ is assumed to be described by

$$\begin{aligned} x_i(t_{k+1}) &= \mathbf{F}_i\left(\mathbf{x}(t_k), \mathbf{u}(t_k)\right) + \xi_i(t_k) \\ &= \sum_{s=1}^{N_i} v_{is} f_{is}\left(\mathbf{x}(t_k), \mathbf{u}(t_k)\right) + \xi_i(t_k) \\ &= \mathbf{f}_i^\top\left(\mathbf{x}(t_k), \mathbf{u}(t_k)\right)\mathbf{v}_i + \xi_i(t_k) \end{aligned} \qquad (1)$$

where $\mathbf{x} = [x_1, \ldots, x_{n_\mathbf{x}}]^\top \in \mathbb{R}^{n_\mathbf{x}}$ denotes the state vector, $\mathbf{u} = [u_1, \ldots, u_{n_\mathbf{u}}]^\top \in \mathbb{R}^{n_\mathbf{u}}$ denotes the input vector, and $\mathbf{F}_i(\cdot) : \mathbb{R}^{n_\mathbf{x}+n_\mathbf{u}} \to \mathbb{R}$ is a smooth nonlinear function which is assumed to be represented as a linear combination of several dictionary functions $f_{is}(\mathbf{x}(t_k), \mathbf{u}(t_k)) : \mathbb{R}^{n_\mathbf{x}+n_\mathbf{u}} \to \mathbb{R}$ (see [2, Sec. 5.4]). These constituent dictionary functions can be monomial, polynomial, constant or any other functional form such as rational, exponential, trigonometric etc. $\mathbf{f}_i(\mathbf{x}(t_k), \mathbf{u}(t_k))$ is the vector of considered dictionary functions (which does not contain unknown parameters) while $\mathbf{v}_i \in \mathbb{R}^{N_i}$ appearing in (1) is the weight vector associated with the dictionary functions vector. The additive noise $\xi_i(t_k)$ is assumed to be i.i.d. Gaussian distributed with zero mean: $\xi_i(t_k) \sim \mathcal{N}(0, \lambda_i)$, with $\mathbb{E}(\xi_i(t_p)) = 0$, $\mathbb{E}(\xi_i(t_p)\xi_i(t_q)) = \lambda_i \delta_{pq}$, where $\delta_{pq} = \begin{cases} 1, & p = q, \\ 0, & p \neq q \end{cases}$. $\xi_i(\cdot)$ and $\xi_j(\cdot)$ are assumed independent $\forall i \neq j$.

Remark 1: The class of systems considered in (1) can be extended to the more general dynamics class $\mathbf{x}_i(t_{k+1}) = \mathbf{F}_i(\mathbf{x}(t_k), \ldots, \mathbf{x}(t_{k-m_\mathbf{x}}), \mathbf{u}(t_k), \ldots, \mathbf{u}(t_{k-m_\mathbf{u}})) + \mathbf{x}_i(t_k)$, where the "orders" $m_\mathbf{x}$ and $m_\mathbf{u}$ are assumed to be known *a priori*, and $\mathbf{F}_i(\cdot) : \mathbb{R}^{(m_\mathbf{x}+1)n_\mathbf{x}+(m_\mathbf{u}+1)n_\mathbf{u}} \to \mathbb{R}$. An example of such system can be found in the supplementary material [1] (see Example 1). In particular, MIMO nonlinear autoregressive models belong to such descriptions.

B. Identification Problem Statement

If M data samples satisfying (1) can be obtained from the system of interest, the system in (1) can be written as $\mathbf{y}_i = \mathbf{\Psi}_i \mathbf{v}_i + \boldsymbol{\xi}_i$, $i = 1, \ldots, n_\mathbf{x}$, where $\mathbf{y}_i \triangleq [x_i(t_1), \ldots, x_i(t_M)]^\top \in \mathbb{R}^{M \times 1}$, $\mathbf{v}_i \triangleq [v_{i1}, \ldots, v_{iN_i}]^\top \in \mathbb{R}^{N_i \times 1}$, $\boldsymbol{\xi}_i \triangleq [\xi_i(t_0), \ldots, \xi_i(t_{M-1})]^\top \in \mathbb{R}^{M \times 1}$, and $\mathbf{\Psi}_i \in \mathbb{R}^{M \times N_i}$ represents the dictionary matrix with its j-th column being $[f_{ij}(\mathbf{x}(t_0), \mathbf{u}(t_0)), \ldots, f_{ij}(\mathbf{x}(t_{M-1}), \mathbf{u}(t_{M-1}))]^\top$.

In this framework, the identification problem amounts to finding $\mathbf{v}_i \in \mathbb{R}^{N_i \times 1}$ given the measured data stored in $\mathbf{y}_i$. This, in turn, amounts to solving a linear regression problem, which can be done using standard least square approaches, provided that the structure of the nonlinearities in the model are known, i.e., provided that $\mathbf{\Psi}_i$ is known. In what follows, we make the following assumption on the measurements contained in $\mathbf{y}_i$.

Assumption 1: The system (1) is fully measurable, i.e., time series data of all the state variables x_i can be obtained.

Depending on the field for which the dynamical model needs to be built, only a few typical nonlinearities specific to this field need to be considered. In what follows, we gather in a matrix $\mathbf{\Phi}_i$ similar to $\mathbf{\Psi}_i$ the set of *all* candidate/possible dictionary functions that we want to consider for identification

$$\mathbf{y}_i = \mathbf{\Phi}_i \mathbf{w}_i + \boldsymbol{\xi}_i, \qquad i = 1, \ldots, n_\mathbf{x}. \qquad (2)$$

The solution $\mathbf{w}_i$ to (2) is typically going to be sparse, which is mainly due to the potential introduction of non-relevant and/or non-independent dictionary functions in $\mathbf{\Phi}_i$.

Since the $n_\mathbf{x}$ linear regression problems in (2) are independent, for simplicity of notation, we omit the subscript i used to index the state variable and simply write:

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \boldsymbol{\xi}. \qquad (3)$$

It should be noted that N, the number of dictionary functions or number of columns of the dictionary matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$, can be very large, at least larger than the number of observations M. Moreover, since $\mathbf{y}$ is constructed from time series data, typically two or more of the columns of the $\mathbf{\Phi}$ matrix are highly correlated. In this case standard methods, which involve some form of ℓ_1-regularised minimization, often yield poor performance on system identification [28].

III. Bayesian Viewpoint on the Reconstruction Problem

A. Sparsity Inducing Priors

Bayesian modelling treats all unknowns as stochastic variables with certain probability distributions [29]. For $\mathbf{y} = \mathbf{\Phi w} + \boldsymbol{\xi}$, it is assumed that the stochastic variables in the vector $\boldsymbol{\xi}$ are Gaussian i.i.d. with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \lambda\mathbf{I})$. In such case, the likelihood of the data given $\mathbf{w}$ is $\mathcal{P}(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{\Phi w}, \lambda\mathbf{I}) \propto \exp[-(1/2\lambda)\|\mathbf{y} - \mathbf{\Phi w}\|_2^2]$. We define a prior distribution $\mathcal{P}(\mathbf{w})$ as $\mathcal{P}(\mathbf{w}) \propto \exp[-(1/2)\sum_j g(w_j)] = \prod_j \exp[-(1/2)g(w_j)] = \prod_j \mathcal{P}(w_j)$, where $g(w_j)$ is a given function of w_j. To enforce sparsity on $\mathbf{w}$, the function $g(\cdot)$ is usually chosen as a concave, non-decreasing function of $|w_j|$. Examples of such functions $g(\cdot)$ include Generalised Gaussian priors and Student's t priors (see [30] for details).

Computing the posterior mean $\mathbb{E}(\mathbf{w}|\mathbf{y})$ is typically intractable because the posterior $\mathcal{P}(\mathbf{w}|\mathbf{y})$ is highly coupled and non-Gaussian. To alleviate this problem, ideally one would like to approximate $\mathcal{P}(\mathbf{w}|\mathbf{y})$ as a Gaussian distribution for which efficient algorithms to compute the posterior exist [29]. Another approach consists in considering *super-Gaussian* priors, which yield a lower bound for the priors $\mathcal{P}(w_j)$ [30]. The sparsity inducing priors mentioned above are *super-Gaussian*. More specifically, if we define $\boldsymbol{\gamma} \triangleq [\gamma_1, \ldots, \gamma_N]^\top \in \mathbb{R}_+^N$, we can represent the priors in the following relaxed (variational) form: $\mathcal{P}(\mathbf{w}) = \prod_{j=1}^n \mathcal{P}(w_j)$, $\mathcal{P}(w_j) = \max_{\gamma_j > 0} \mathcal{N}(w_j|0, \gamma_j)\varphi(\gamma_j)$, where $\varphi(\gamma_j)$ is a nonnegative function which is treated as a hyperprior with γ_j being its associated hyperparameters. Throughout, we call $\varphi(\gamma_j)$ the "*potential function*." This Gaussian relaxation is possible if and only if $\log \mathcal{P}(\sqrt{w_j})$ is concave on $(0, \infty)$. The following proposition provides a justification for the above:

Proposition 1 [30]: A probability density $\mathcal{P}(w_j) \equiv \exp(-g(w_j^2))$ can be represented in the convex variational form: $\mathcal{P}(w_j) = \max_{\gamma_j > 0} \mathcal{N}(w_j|0, \gamma_j)\varphi(\gamma_j)$ if and only if $-\log \mathcal{P}(\sqrt{w_j}) = g(w_j)$ is concave on $(0, \infty)$. In this case the potential function takes the following expression: $\varphi(\gamma_j) = \sqrt{2\pi/\gamma_j}\exp(g^*(\gamma_j/2))$ where $g^*(\cdot)$ is the concave conjugate of $g(\cdot)$. A symmetric probability density $\mathcal{P}(w_j)$ is said to be super-Gaussian if $\mathcal{P}(\sqrt{w_j})$ is log-convex on $(0, \infty)$.

B. Marginal Likelihood Maximisation

For a fixed $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]$, we define a relaxed prior, which is a joint probability distribution over $\mathbf{w}$ and $\boldsymbol{\gamma}$, as $\mathcal{P}(\mathbf{w}; \boldsymbol{\gamma}) = \prod_j \mathcal{N}(w_j|0, \gamma_j)\varphi(\gamma_j) = \mathcal{P}(\mathbf{w}|\boldsymbol{\gamma})\mathcal{P}(\boldsymbol{\gamma}) \leq \mathcal{P}(\mathbf{w})$, where $\mathcal{P}(\mathbf{w}|\boldsymbol{\gamma}) \triangleq \prod_j \mathcal{N}(w_j|0, \gamma_j), \mathcal{P}(\boldsymbol{\gamma}) \triangleq \prod_j \varphi(\gamma_j)$. Since the likelihood is $\mathcal{P}(\mathbf{y}|\mathbf{w})$ is Gaussian, we can get a relaxed posterior which is also Gaussian $\mathcal{P}(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}) = ((\mathcal{P}(\mathbf{y}|\mathbf{w})\mathcal{P}(\mathbf{w}; \boldsymbol{\gamma}))/(\int \mathcal{P}(\mathbf{y}|\mathbf{w})\mathcal{P}(\mathbf{w}; \boldsymbol{\gamma})d\mathbf{w})) = \mathcal{N}(\mathbf{m}_\mathbf{w}, \mathbf{\Sigma}_\mathbf{w})$. Defining $\mathbf{\Gamma} \triangleq \mathrm{diag}[\boldsymbol{\gamma}]$, the posterior mean and covariance are given by

$$\mathbf{m}_\mathbf{w} = \mathbf{\Gamma\Phi}^\top(\lambda\mathbf{I} + \mathbf{\Phi\Gamma\Phi}^\top)^{-1}\mathbf{y} \tag{4}$$

$$\mathbf{\Sigma}_\mathbf{w} = \mathbf{\Gamma} - \mathbf{\Gamma\Phi}^\top(\lambda\mathbf{I} + \mathbf{\Phi\Gamma\Phi}^\top)^{-1}\mathbf{\Phi}. \tag{5}$$

Now the key question is how to choose the most appropriate $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}} = [\hat{\gamma}_1, \ldots, \hat{\gamma}_N]$ to maximise $\prod_j \mathcal{N}(w_j|0, \gamma_j)\varphi(\gamma_j)$ such that $\mathcal{P}(\mathbf{w}|\mathbf{y}, \hat{\boldsymbol{\gamma}})$ can be a "good" relaxation to $\mathcal{P}(\mathbf{w}|\mathbf{y})$. Using the product rule for probabilities, we can write the full posterior as: $\mathcal{P}(\mathbf{w}, \boldsymbol{\gamma}|\mathbf{y}) \propto \mathcal{P}(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma})\mathcal{P}(\boldsymbol{\gamma}|\mathbf{y}) = \mathcal{N}(\mathbf{m}_\mathbf{w}, \mathbf{\Sigma}_\mathbf{w}) \times \mathcal{P}(\mathbf{y}|\boldsymbol{\gamma})\mathcal{P}(\boldsymbol{\gamma})/\mathcal{P}(\mathbf{y})$. Since $\mathcal{P}(\mathbf{y})$ is independent of $\boldsymbol{\gamma}$, the quantity $\mathcal{P}(\mathbf{y}|\boldsymbol{\gamma})\mathcal{P}(\boldsymbol{\gamma}) = \int \mathcal{P}(\mathbf{y}|\mathbf{w})\mathcal{P}(\mathbf{w}|\boldsymbol{\gamma})$ $\mathcal{P}(\boldsymbol{\gamma})d\mathbf{w}$ is the prime target for variational methods [31]. This quantity is known as evidence or marginal likelihood. A good way of selecting $\hat{\boldsymbol{\gamma}}$ is to choose it as the minimizer of the sum of the misaligned probability mass, e.g.,

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &= \arg\min_{\boldsymbol{\gamma} \geq \mathbf{0}} \int \mathcal{P}(\mathbf{y}|\mathbf{w})\,|\mathcal{P}(\mathbf{w}) - \mathcal{P}(\mathbf{w}; \boldsymbol{\gamma})|\,d\mathbf{w} \\ &= \arg\max_{\boldsymbol{\gamma} \geq \mathbf{0}} \int \mathcal{P}(\mathbf{y}|\mathbf{w}) \prod_{j=1}^n \mathcal{N}(w_j|0, \gamma_j)\varphi(\gamma_j)d\mathbf{w}.\end{aligned} \tag{6}$$

The second equality is a consequence of $\mathcal{P}(\mathbf{w}; \boldsymbol{\gamma}) \leq \mathcal{P}(\mathbf{w})$. The procedure in (6) is referred to as evidence maximization or type-II maximum likelihood [32]. It means that the marginal likelihood can be maximized by selecting the most probable hyperparameters able to explain the observed data. Once $\hat{\boldsymbol{\gamma}}$ is computed, an estimate of the unknown weights can be obtained by setting $\hat{\mathbf{w}}$ to the posterior mean (4) as $\hat{\mathbf{w}} = \mathbb{E}(\mathbf{w}|\mathbf{y}; \hat{\boldsymbol{\gamma}}) = \hat{\mathbf{\Gamma}}\mathbf{\Phi}^\top(\lambda\mathbf{I} + \mathbf{\Phi}\hat{\mathbf{\Gamma}}\mathbf{\Phi}^\top)^{-1}\mathbf{y}$, with $\hat{\mathbf{\Gamma}} \triangleq \mathrm{diag}[\hat{\boldsymbol{\gamma}}]$. If an algorithm can be proposed to compute $\hat{\boldsymbol{\gamma}}$ in (6), we can, based on it, obtain an estimation of the posterior mean $\hat{\mathbf{w}}$.

C. Enforcing Additional Constraints on $\mathbf{w}$

It is often important to be able to impose constraints on $\hat{\mathbf{w}}$ when formulating the optimization problem (6) used to compute $\hat{\mathbf{w}}$ from $\hat{\boldsymbol{\gamma}}$. In physical and biological systems, positivity of the parameters $\mathbf{w}$ of the system is an example of such constraints. Another example of constrained optimization comes from stability considerations, which emerge naturally when the underlying system is known *a priori* to be stable.[1] Yet, only a few contributions in the literature address the problem of how to take into account *a priori* information on system stability in the context of system identification [23], [24]. To be able to integrate constraints on $\mathbf{w}$ into the problem formulation, we consider the following assumption on $\mathbf{w}$.

Assumption 2: Constraints on the weights $\mathbf{w}$ can be described by a set of convex functions:

$$\begin{aligned} H_i^{[I]}(\mathbf{w}) &\leq 0, \quad i = 1, \ldots, m_I \\ H_j^{[E]}(\mathbf{w}) &= 0, \quad j = 1, \ldots, m_E \end{aligned} \tag{7}$$

where the convex functions $H_i^{[I]} : \mathbb{R}^N \to \mathbb{R}$ are used to define inequality constraints, whereas the convex functions $H_j^{[E]} : \mathbb{R}^N \to \mathbb{R}$ are used to define equality constraints.

IV. Nonconvex Optimization for Identification Problems

In this section, we derive a sparse Bayesian formulation of the problem of system identification with convex constraints, which is casted into a nonconvex optimization problem. The nonconvex optimization problem can be dealt by an iterative re-weighted ℓ_1-minimization algorithm.

A. Nonconvex Objective Function in Hyperparameter

Theorem 1: The optimal hyperparameters $\hat{\boldsymbol{\gamma}}$ in (6) can be obtained by minimizing the following objective function:

$$\mathcal{L}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = \log|\lambda\mathbf{I} + \mathbf{\Phi\Gamma\Phi}^\top| + \mathbf{y}^\top(\lambda\mathbf{I} + \mathbf{\Phi\Gamma\Phi}^\top)^{-1}\mathbf{y} + \sum_{j=1}^N p(\gamma_j) \tag{8}$$

where $p(\gamma_j) = -2\log\varphi(\gamma_j)$ and $|\mathbf{A}|$ denotes the determinant of matrix $\mathbf{A}$. The posterior mean is then given by $\hat{\mathbf{w}} = \hat{\mathbf{\Gamma}}\mathbf{\Phi}^\top(\lambda\mathbf{I} + \mathbf{\Phi}\hat{\mathbf{\Gamma}}\mathbf{\Phi}^\top)^{-1}\mathbf{y}$, where $\hat{\mathbf{\Gamma}} = \mathrm{diag}[\hat{\boldsymbol{\gamma}}]$.

[1] Many stability conditions can be formulated as convex optimization problems (see for example [33], [34]).

Proof 1: See Section A in the Appendix of [1].

Lemma 1: The objective function in the hyperparameter $\boldsymbol{\gamma}$-space, $\mathcal{L}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})$ in (8), is nonconvex.

Proof 2: See Section B in the Appendix of [1].

B. Nonconvex Objective Function in $\mathbf{w}$ *With Convex Constraints*

Based on the analysis in Section IV-A, we first derive a dual objective function in the $\mathbf{w}$-space with convex constraints by considering the equivalent objective function of (8) in the $\boldsymbol{\gamma}$-space. We then show that this equivalent objective function is also nonconvex.

Theorem 2: The estimate for $\mathbf{w}$ with constraints can be obtained by solving the optimization problem

$$\min_{\mathbf{w}} \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 + \lambda g_{\mathrm{SB}}(\mathbf{w}), \quad \text{subject to (7)} \tag{9}$$

where $g_{\mathrm{SB}}(\mathbf{w}) = \min_{\boldsymbol{\gamma}\geq\mathbf{0}}\{\mathbf{w}^\top\boldsymbol{\Gamma}^{-1}\mathbf{w} + \log|\lambda\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^\top| + \sum_{j=1}^{N} p(\gamma_j)\}$ and the estimate of the stochastic variable $\mathbf{w}$ is given by the poseterior mean $\mathbf{m}_{\mathbf{w}}$ defined in (4).

Proof 3: See Section C in the Appendix of [1].

Although all the constraint functions are convex in Theorem 2, we show in the following Lemma that the objective function in (9) is nonconvex since it is the sum of convex and concave functions.

Lemma 2: The penalty function $g_{\mathrm{SB}}(\mathbf{w})$ in Theorem 2 is a non-decreasing, concave function of $|\mathbf{w}|$, which promotes sparsity on the weights $\mathbf{w}$.[2]

Proof 4: The proof uses the duality lemma (see [35, Sec. 4.2]). See Section D in the Appendix of [1].

C. Lasso Type Algorithm

We define the terms excluding $h^*(\boldsymbol{\gamma}^*)$ as

$$\mathcal{L}_{\boldsymbol{\gamma}^*}(\boldsymbol{\gamma}, \mathbf{w}) \triangleq \frac{1}{\lambda}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 + \sum_j \left(\frac{w_j^2}{\gamma_j} + \gamma_j^*\gamma_j\right). \tag{10}$$

For a fixed $\boldsymbol{\gamma}^*$, we notice that $\mathcal{L}_{\boldsymbol{\gamma}^*}(\boldsymbol{\gamma}, \mathbf{w})$ is jointly convex in $\mathbf{w}$ and $\boldsymbol{\gamma}$ and can be globally minimized by solving over $\boldsymbol{\gamma}$ and then $\mathbf{w}$. Since $w_j^2/\gamma_j + \gamma_j^*\gamma_j \geq 2w_j\sqrt{\gamma_j^*}$, for any $\mathbf{w}$, $\gamma_j = |w_j|/\sqrt{\gamma_j^*}$ minimizes $\mathcal{L}_{\boldsymbol{\gamma}^*}(\boldsymbol{\gamma}, \mathbf{w})$. When $\gamma_j = |w_j|/\sqrt{\gamma_j^*}$ is substituted into $\mathcal{L}_{\boldsymbol{\gamma}^*}(\boldsymbol{\gamma}, \mathbf{w})$, $\hat{\mathbf{w}}$ can be obtained by solving the following weighted convex ℓ_1-minimization procedure:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 + 2\lambda\sum_{j=1}^{N}\sqrt{\gamma_j^*}|w_j|\right\}. \tag{11}$$

We can then set $\gamma_j = |\hat{w}_j|/\sqrt{\gamma_j^*}$, $\forall j$. As a consequence, $\mathcal{L}_{\boldsymbol{\gamma}^*}(\boldsymbol{\gamma}, \mathbf{w})$ will be minimized for any fixed $\boldsymbol{\gamma}^*$. Due to the concavity of $g_{\mathrm{SB}}(\mathbf{w})$, the objective function in (9) can be optimised using a re-weighted ℓ_1-minimization in a similar way as was considered in (11). The updated weight at the kth iteration is then given by $u_j^{(k)} \triangleq ((\partial g_{\mathrm{SB}}(\mathbf{w}))/(2\partial|w_j|))|_{\mathbf{w}=\mathbf{w}^{(k)}} = \sqrt{\boldsymbol{\gamma}_j^{*(k)}}$.

We can now explain how the update of the parameters can be performed based on the above. We start by setting the iteration count k to zero and $u_j^{(0)} = 1$, $\forall j$. At this stage, the solution is a typical ℓ_1-minimization solution. Then at the kth iteration, we initialise $u_j^{(k)} = \sqrt{\boldsymbol{\gamma}_j^{*(k)}}$, $\forall j$ and then minimize over $\boldsymbol{\gamma}$ using $\gamma_j = |w_j|/\sqrt{\gamma_j^*}$, $\forall j$. Consider again $\mathcal{L}_{\boldsymbol{\gamma},\mathbf{w}}(\boldsymbol{\gamma}, \mathbf{w})$. For any fixed $\boldsymbol{\gamma}$ and $\mathbf{w}$, the tightest bound can be obtained by minimizing over $\boldsymbol{\gamma}^*$. The tightest value of $\boldsymbol{\gamma}^* = \hat{\boldsymbol{\gamma}}^*$ equals the gradient of the function $h(\boldsymbol{\gamma}) \triangleq \log|\lambda\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^\top| + \sum_{j=1}^{N} p(\gamma_j)$ defined in Lemma 1 at the current $\boldsymbol{\gamma}$. $\boldsymbol{\gamma}^*$ has the following analytical expression: $\hat{\boldsymbol{\gamma}}^* = \nabla_{\boldsymbol{\gamma}}(\log|\lambda\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^\top| + \sum_{j=1}^{N} p(\gamma_j)) = \mathrm{diag}[\boldsymbol{\Phi}^\top(\lambda\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^\top)^{-1}\Phi] + p'(\boldsymbol{\gamma})$, where $p'(\boldsymbol{\gamma}) = [p'(\gamma_1), \ldots, p'(\gamma_N)']^\top$. The optimal $\boldsymbol{\gamma}^{*(k+1)}$ can then be obtained as $\boldsymbol{\gamma}^{*(k+1)} = \mathrm{diag}[\boldsymbol{\Phi}^\top(\lambda\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}^{(k)}\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{\Phi}] + p'(\boldsymbol{\gamma}^{(k)})$, where $\boldsymbol{\Gamma}^{(k)} \triangleq \mathrm{diag}[\boldsymbol{\gamma}^{(k)}]$. After computing the estimation of $\gamma_j^{(k)} = |w_j^{(k)}|/\sqrt{\gamma_j^{*(k)}}$, we can compute $\boldsymbol{\gamma}^{*(k+1)}$, which gives $\gamma_j^{*(k+1)} = \boldsymbol{\Phi}_j^\top (\lambda\mathbf{I} + \boldsymbol{\Phi}\mathbf{U}^{(k)}\mathbf{W}^{(k)}\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{\Phi}_j + p'(\gamma_j^{(k)})$, $\mathbf{U}^{(k)} \triangleq \mathrm{diag}[\mathbf{u}^{(k)}]^{-1} = \mathrm{diag}[\sqrt{\boldsymbol{\gamma}^{*(k)}}]^{-1}$, $\mathbf{W}^{(k)} \triangleq \mathrm{diag}[|\mathbf{w}^{(k)}|]$. We can then define $u_j^{(k+1)} \triangleq \sqrt{\gamma_j^{*(k+1)}}$ for the next iteration of the weighted ℓ_1-minimization. The above described procedure is summarized in Algorithm 1.

Algorithm 1 Nonlinear Identification Algorithm

```
Collect time series data from the system of interest (assuming the system can be described by (1));
Select the candidate dictionary functions that will be used to construct the dictionary matrix described in Section II-B;
Initialise $u_j^0 = 1, \forall j$
for $k = 0, \ldots, k_{\max}$ do
    Solve the weighted $\ell_1$-minimization problem with convex constraints on $\mathbf{w}$
        $$\min_{\mathbf{w}} \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|_2^2 + 2\lambda\sum_j u_j^{(k)}|w_j|, \quad \text{subject to (7)}$$
    Set $\mathbf{U}^{(k)} \triangleq \mathrm{diag}[\mathbf{u}^{(k)}]^{-1}$, $\mathbf{W}^{(k)} \triangleq \mathrm{diag}[|\mathbf{w}^{(k)}|]$;
    Update weights $u_j^{(k+1)}$ for the next iteration $u_j^{(k+1)} = [\boldsymbol{\Phi}_j^\top(\lambda\mathbf{I} + \boldsymbol{\Phi}\mathbf{U}^{(k)}\mathbf{W}^{(k)}\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{\Phi}_j + p'(\gamma_j^{(k)})]^{1/2}$;
    if a stopping criterion is satisfied then
        Break;
    end if
end for
```

Remark 2: There are two important aspects of the re-weighted ℓ_1-minimization algorithm presented in Algorithm 1. First, for convex optimization, there will be no exact zeros during the iterations and strictly speaking, we will always get a solution without any zero entry even when the RIP condition holds. However, some of the estimated weights will have very small magnitudes compared to those of other weights, e.g., $\pm 10^{-5}$ compared to 1, or the "energy" some of the estimated weights will be several orders of magnitude lower than the average "energy," e.g., $\|w_j\|_2^2 \ll \|\mathbf{w}\|_2^2$. Thus a threshold needs to be defined *a priori* to prune "small" weights at each iteration. The second aspect concerns the computational complexity of this approach. The repeated execution of Algorithm 1 is very cheap computationally since it scales as $\mathcal{O}(MN\|\mathbf{w}^{(k)}\|_0)$ (see [36], [37]). Since at each iteration certain weights are estimated to be zero, certain dictionary functions spanning the corresponding columns of $\boldsymbol{\Phi}$ can be pruned out for the next iteration.

[2] $|\mathbf{w}|$ denotes the vector whose elements are $|w_j|, \forall j$.

D. Convergence

It is natural to investigate the convergence properties of this iterative re-weighted ℓ_1-minimization procedure. Let $\mathcal{A}(\cdot)$ denote a mapping that assigns to every point in $\mathbb{R}_+^N$ the subset of $\mathbb{R}_+^N$ which satisfies Steps 5 and 6 in Algorithm 1. Then the convergence property can be established as follows.

Theorem 3: Given the initial point $\boldsymbol{\gamma}^{(0)} \in \mathbb{R}_+^n$ a sequence $\{\boldsymbol{\gamma}^{(k)}\}_{k=0}^{\infty}$ is generated such that $\boldsymbol{\gamma}^{(k+1)} \in \mathcal{A}(\boldsymbol{\gamma}^{(k)}),\ \forall k$. This sequence is guaranteed to converge to a local minimum (or saddle point) of $\mathcal{L}_{\boldsymbol{\gamma}}$ in (8).

Proof 5: The proof is in one-to-one correspondence with that of the Global Convergence Theorem [38]. See Section E in the Appendix of [1].

V. Illustrative Numerical Examples

To implement Algorithm 1, we use CVX, a popular package for specifying and solving convex programs [39]. To illustrate our results, the approach is applied to two classic examples: 1) the Genetic Repressilator Network, where we identify nonlinear regulation relationships between genes, transcriptional and translational strengths and degradation rates and 2) a network of Kuramoto Oscillators, where we identify the network topology and nonlinear coupling functions. More details about these two examples and algorithmic comparisons with other algorithms described in [40] in terms of the Root of the Normalised Mean Square Error (RNMSE) and computational running time for different Signal-to-Noise Ratios (SNR) can be found in the supplementary material [1]. Importantly, this comparison shows that Algorithm 1 outperforms other classical algorithms [40] in terms of RNMSE, when used to identify the nonlinear systems associated with these illustrative examples. The corresponding code is available at https://github.com/panweihit/BSID.

VI. Conclusion and Discussion

This technical note proposed a new method for the identification of nonlinear discrete-time state-space systems with additive process noise. This method only required time-series data and some prior knowledge about the type of system from which these data have been acquired (e.g., biochemical, mechanical or electrical). Based on this prior knowledge, candidate nonlinear functions (dictionary functions) can be selected for the particular type of system to be identified.

Due to the typical sparsity in terms of number of dictionary functions used to describe the dynamics of nonlinear systems and the fact that the number of measurements is typically small (at least smaller than the number of candidate nonlinear functions), the corresponding identification problem falls into the class of sparse linear regression problems. We considered this problem in a Bayesian framework and solved it efficiently using an iterative re-weighted ℓ_1-minimization algorithm. This approach also allowed us to easily add convex constraints from prior knowledge of some properties of the system (e.g., positivity of certain variables, stability of the system, etc.). Finally, we illustrated how this approach can be efficiently used to accurately reconstruct discrete-time nonlinear models of the genetic repressilator and of Kuramoto networks.

Several important questions remain currently open for further research. Possibly, the most important is the assumption that the system is fully measurable. Typically, only part of the state is measured [41], [42], and, in particular, the number of hidden/unobservable nodes and their position in the network are usually unknown. We are currently investigating partial-measurement extensions of the method presented in this technical note. Meanwhile, our algorithm is relatively more computationally expensive than other algorithms such as those in [40] but outperforms them all in terms of the accuracy of the identification as measured by the RNMSE. In future work, we plan to improve our proposed algorithm by exploiting further the structure of the optimization problem at hand and reducing the associated algorithmic complexity. Another issue is that we assume that only process noise is present, and thus do not directly take into account measurement noise. We are currently working on an extension of the method allowing the incorporation of measurement noise into the presented framework.

Acknowledgment

The authors would like to thank Dr. W. Dai, Prof. M. Barahona, and Dr. A. Sootla (Imperial College London) for helpful discussions.

References

[1] Appendix and Supplementary Material, arXiv:1408.3549.

[2] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.

[3] S. A. Billings, *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, Spatio-Temporal Domains*. New York, NY, USA: Wiley, 2013.

[4] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson, and A. Juditsky, "Nonlinear black-box modeling in system identification: A unified overview," *Automatica*, vol. 31, no. 12, pp. 1691–1724, 1995.

[5] R. Haber and H. Unbehauen, "Structure identification of nonlinear dynamic systems: Survey on input/output approaches," *Automatica*, vol. 26, no. 4, pp. 651–677, 1990.

[6] M. Barahona and C. Poon, "Detection of nonlinear dynamics in short, noisy time series," *Nature*, vol. 381, no. 6579, pp. 215–217, 1996.

[7] N. Wiener, *Nonlinear Problems in Random Theory*, vol. 1. Cambridge, MA, USA: MIT Press, Aug. 1996, p. 142.

[8] K. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Trans. Neural Netw.*, vol. 1, no. 1, pp. 4–27, Jan. 1990.

[9] I. Leontaritis and S. Billings, "Input-output parametric models for nonlinear systems part i: Deterministic nonlinear systems," *Int. J. Control*, vol. 41, no. 2, pp. 303–328, 1985.

[10] E. Bai, "An optimal two-stage identification algorithm for hammerstein-wiener nonlinear systems," *Automatica*, vol. 34, no. 3, pp. 333–338, 1998.

[11] D. Kollar and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[12] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, USA: Morgan Kaufmann, 1988.

[13] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, Search*, vol. 81. Cambridge, MA, USA: MIT Press, 2000.

[14] D. Barber and A. Cemgil, "Graphical models for time-series," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 18–28, 2010.

[15] F. R. Bach and M. I. Jordan, "Learning graphical models for stationary time series," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2189–2199, Aug. 2004.

[16] D. Materassi and M. V. Salapaka, "On the problem of reconstructing an unknown topology via locality properties of the wiener filter," *IEEE Trans. Autom. Control*, vol. 57, no. 7, pp. 1765–1777, Jul. 2012.

[17] W. Pan, Y. Yuan, J. Gonçalves, and G.-B. Stan, "Reconstruction of arbitrary biochemical reaction networks: A compressive sensing approach," in *Proc. IEEE 51st Annu. Conf. Decision and Control (CDC)*, 2012, pp. 2334–2339.

[18] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[19] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[20] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. CVPR*, 2008, pp. 1–8.

[21] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.

[22] T. Chen, M. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto, "System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques," *IEEE Trans. Autom. Control*, vol. 59, no. 11, pp. 2933–2945, Nov. 2014.

[23] V. Cerone, D. Piga, and D. Regruto, "Enforcing stability constraints in set-membership identification of linear dynamic systems," *Automatica*, vol. 47, no. 11, pp. 2488–2494, 2011.

[24] M. Zavlanos, A. Julius, S. Boyd, and G. Pappas, "Inferring stable genetic networks from steady-state data," *Automatica*, vol. 47, no. 6, pp. 1113–1122, 2011.

[25] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc., Ser. B (Methodological)*, pp. 267–288, 1996.

[26] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.

[27] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, 2003.

[28] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[29] C. Bishop, *Pattern Recognition and Machine Learning*, vol. 4. New York, NY, USA: Springer-Verlag, 2006.

[30] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, "Variational EM algorithms for non-Gaussian latent variable models," *Advanc. Neur. Inform. Process. Syst.*, vol. 18, p. 1059, 2006.

[31] M. Wainwright and M. Jordan, "Graphical models, exponential families, variational inference," *Found. Trends in Mach. Learning.*, vol. 1, no. 1/2, pp. 1–305, 2008.

[32] M. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[33] S. Boyd, L. El Ghaoul, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, vol. 15. Philadelphia, PA, USA: Society for Industrial Mathematics, 1987.

[34] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[35] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.

[36] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimisation," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[37] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 317–329, 2010.

[38] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1969.

[39] M. Grant, S. Boyd, and Y. Ye, CVX: MATLAB Software for Disciplined Convex Programming, 2008. [Online]. Available: http://cvxr.com

[40] D. L. Donoho, V. C. Stodden, and Y. Tsaig, About SparseLab 2007. [Online]. Available: http://sparselab.stanford.edu

[41] Y. Yuan, G. Stan, S. Warnick, and J. Goncalves, "Robust dynamical network structure reconstruction," *Automatica (Special Issue on System Biology)*, vol. 47, pp. 1230–1235, 2011.

[42] Y. Yuan, "Decentralised network prediction and reconstruction algorithms," Ph.D. dissertation, 2012.