Causal Variable Selection Using Equilibrium Relations from Nonlinear Dynamics

Chris J. Oates[1,2,3]
[1] Centre for Complexity Science and
[2] Department of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Sach Mukherjee[3]
[3] Department of Biochemistry
Netherlands Cancer Institute
1066CX Amsterdam, The Netherlands

Abstract

Variable selection approaches are widely used in regression. In many applications, variable selection is used to prioritise variables that may have a causal influence on the response. However, many widely-used approaches are rooted in linear models that may limit ability to discern causal influences. Yet in a number of settings, nonlinear dynamical models of underlying processes are available that define classes of functional relationships between covariates and response. Here, we propose an approach for variable selection that is rooted in such nonlinear dynamical systems. Reversible jump Markov chain Monte Carlo is used to assess putative functional relationships and obtain posterior probabilities for the inclusion of individual covariates. We illustrate the approach in the context of protein biochemistry, using data simulated from a recent mechanistic model and in an application to protein data from cell lines. In the former, the true causal graph is known and is used for assessment, whilst in the latter results are compared against known biochemistry. We find that the proposed methodology is more effective at identifying causal variates than methods based on linear models. Finally we discuss opportunities and challenges involved in extending these ideas further.

1 INTRODUCTION

Variable selection approaches have been widely studied in the machine learning and statistical literature. In applications, variable selection typically serves two related yet distinct goals: (i) to improve prediction performance (e.g. in regression or classification); and (ii) to select variables that are influential in terms of the mechanisms underlying the response. Problem (i) has been extensively studied in machine learning and statistics over many years. However, in recent years problem (ii) has attracted increasing attention, and in many settings the results of variable selection are used to prioritise further work with explicitly causal goals. For instance, in economics, variables determined to be influential may be the focus of further work using natural experiments, while in biology, such variables may be selected for interventional experiments.

The distinction between variable selection problems (i) and (ii) mirrors the general distinction between regression and causal inference, as discussed extensively by Pearl (2009) and others. In terms of relevant variable subsets, the difference can be put simply: In variable selection for regression, the goal is to select variables that aid in prediction, while in causal variable selection the aim is to identify variables which actually drive the response. In general, the two sets of variables may be distinct. We are interested in problem (ii), which we refer to simply as *causal variable selection* or CVS. Accordingly, in both empirical examples we show below, we assess methodology with respect to ability to identify truly causal variables, rather than performance in the prediction sense of regression.

Many widely-used variable selection approaches employ regression formulations based on linear models, most often with additive Gaussian error. However, for CVS, linear formulations remain unsatisfactory for several reasons: (1) Variates may be highly correlated, often due to underlying dynamics. Flexibility inherent in the linear approach requires that modifications are made to the linear model in order to exclude non-causal but highly correlated variates (Cho and Fryzlewicz, 2012). (2) Symmetry of the linear equivalence in general limits identification of underlying causal relationships (Pearl, 2009; Peters *et al.*, 2011). (3) When the data generating model is nonlinear, the linear model may produce inefficient or inconsistent esti-

mation, attributing causal status to artifacts resulting from model misspecification (Heagerty and Kurland, 2001; Lv and Liu, 2010).

Yet, in many settings, nonlinear dynamical models of relevant data-generating processes are available. In economics, dynamics of many economic phenomena of interest have been formalized (e.g. dynamic stochastic general equilibrium models in macroeconomics). Equally, in biology, many processes have dynamical descriptions rooted in biochemical or biophysical models (e.g. enzyme kinetics). Where such models are available, it is natural to ask whether they may be exploited to facilitate CVS, since an appropriate nonlinear formulation may have enhanced power to exclude non-causal variates. Note however that due to the added complexity of nonlinear formulations, it is not *a priori* obvious that they must outperform simpler models, under practical conditions of sample size and measurement noise.

Recently, Peters *et al.* (2011) presented a theoretical study concerning certain classes of structural equations ("identifiable functional model classes" or IFMOCs) that permit identifiability of the underlying causal network structure. Our work shares motivation with theirs insofar as we are interested in using nonlinear functional forms to aid causal inference. However, we focus on the *inferential* challenge of identifying causal variables from noisy data, while Peters *et al.* do not consider inference from noisy data at all. Furthermore, both our examples involve settings that are not covered by their theory (due to cycles in the underlying causal graph). Thus, while our work shares the spirit of Peters *et al.*, and complements their work, the contribution we make is very much distinct.

In this article, we focus on the use of nonlinear dynamical formulations for the purpose of CVS. Our programme, in brief, is as follows. We consider a dynamical system $\dot{\boldsymbol{X}} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{U}; \boldsymbol{\theta})$ with state vector $\boldsymbol{X}$, external inputs or drivers $\boldsymbol{U}$ and unknown parameters $\boldsymbol{\theta}$. One component X_i of the state vector is of particular interest and is treated as the response. The corresponding component function f_i depends on a subset of the state variables; these are the parents of node i in an (unknown) causal graph for the dynamical system. The goal of CVS is to estimate the identity of these causal influences. Since the dynamical system is not usually known in detail *a priori*, we consider the practically applicable case in which f_i is known only to belong to a certain class $\mathcal{F}$, with all dynamical parameters remaining unknown. We perform model selection in a Bayesian framework, using reversible-jump Markov chain Monte Carlo (RJMCMC) to explore the functional model class $\mathcal{F}$. This yields posterior inclusion probabilities that are analogous to those obtained via conventional Bayesian variable selection. We then address two main empirical questions: (1) To what extent is CVS possible in practical settings, where data are noisy and dynamical parameters are unknown? (2) Does inference based on functional models offer advantages over the linear model?

The causal graph for a dynamical system depends on the time-scale on which the system is observed (Dash, 2003). Over infinitesimal time-scales, Iwasaki and Simon (1994) introduced *dynamic causal models* (DCM), associating causes with variables which appear as forcing terms in a system of differentials, as discussed in the previous paragraph. Causal graphs may be formally defined at equilibrium by applying the *equilibrate* operator to a DCM, which, informally speaking, has the effect of setting differential terms to zero; $\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{U}; \boldsymbol{\theta}) = \boldsymbol{0}$. In line with the mainstream variable selection literature, which has not emphasised time series data, we focus here on the equilibrium setting. We return in the Discussion to the important conceptual issues of identifiability and faithfulness at equilibrium.

To limit scope we develop these general ideas in the context of a specific application and associated class of dynamical systems. We focus on biochemistry, specifically regulation of protein state by phosphorylation. Protein phosphorylation is an advantageous test-bed in the context of CVS for dynamical systems. First, the kinetics of phosphorylation have been extensively studied, with dynamical formulations widely available in the literature (see e.g. Leskovac, 2003). Second, for some proteins and pathways, regulation has been studied in considerable causal and mechanistic detail. In many cases, it is known that certain proteins have the physical capacity to regulate a given target and corresponding mechanisms of interaction have been thoroughly explored *in vitro*. Third, there exist detailed computational models for canonical protein signalling pathways, which have been validated against experimental data (e.g. Schoeberl *et al.*, 2002; Xu *et al.*, 2010). Such models provide non-trivial simulation engines for assessment of methodology.

In order to allow gold-standard validation we focus on well-understood aspects of protein signalling. However it is important to note that while good models exist for certain pathways, a vast amount remains unknown concerning biological regulatory systems, including gene regulation and signalling. Indeed, a large part of ongoing efforts in mainstream biology seek to discover novel regulation in such systems. Traditionally, causation has been established on an edge-by-edge basis, based on low-dimensional interventional experiments. However, advances in data acquisition now permit studies with large numbers of variables (rather than a small number of pre-selected ones).

Such data offer for the first time an "unbiased" way to screen for novel influences. An automatic approach to causal variable selection, based only on observational data, could therefore greatly accelerate identification of novel regulatory influences. The associated inferential challenges are highly non-trivial and motivate the work here as well as many other recent statistical and machine learning efforts related to model selection.

Systems biology represents a key current application domain for machine learning and statistical approaches. However, to date, most methods applied to CVS-type problems have been rooted in discrete or linear formulations (Hill, 2012; Oates and Mukherjee, 2012; Sachs *et al.*, 2005). Recently Honkela *et al.* (2010) and Äijö and Lähdesmaki (2009) applied nonparametric (Gaussian process) regression to a related biological problem (gene regulation), but within a linear model.

The remainder of this paper is organised as follows. In Section 2 our approach is initially described in general terms, followed by a detailed exposition and application to protein phosphorylation. In Section 3 we present results on data simulated from a recently developed dynamical model of mitogen-activated protein kinase (MAPK) signalling, that has been validated against experimental data (Xu *et al.*, 2010). We then show results on real proteomic data from breast cancer cell lines. Finally, Section 4 closes with a discussion of practical implications and opportunities for CVS based on functional models, along with associated technical challenges.

2 METHODS

We begin in Section 2.1 by describing our approach in general terms. Section 2.2 then introduces relevant concepts in the application area of protein phosphorylation that we use to illustrate our approach. In particular we describe a class of dynamical systems describing phosphorylation based on ordinary differential equations (ODEs). Next, in Section 2.3 this model class is embedded into a statistical framework for observations obtained at dynamic equilibrium. Inference over model space is carried out in a Bayesian framework. This is facilitated by RJMCMC, with Section 2.4 dedicated to a presentation of our sampling scheme and a discussion of some key implementational details.

2.1 GENERAL FORMULATION

Let $\dot{\boldsymbol{X}} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{U}; \boldsymbol{\theta})$ describe the dynamics of a state vector $\boldsymbol{X} = (X_1, \ldots, X_p)$, with index set $\mathcal{X} = \{1, \ldots, p\}$, driven by external inputs $\boldsymbol{U} = (U_1, \ldots, U_q)$, with index set $\mathcal{U} = \{1, \ldots, q\}$; system parameters are given by $\boldsymbol{\theta}$.

Underlying the infinitesimal dynamics of $\boldsymbol{X}$ is an unknown causal graph G. One variable X_i is of particular interest; this represents the response in a regression sense. Let $\pi_i \subset \mathcal{X}$ denote the parents of variable i in the causal graph G. We seek to identify π_i using only steady-state observations of the state vector $\boldsymbol{X}$. For X_i, we have the reduced dynamical system $\dot{X}_i = f_i(\boldsymbol{X}_{\pi_i}, \boldsymbol{U}_{\gamma_i}; \boldsymbol{\theta}_i)$, where $\gamma_i \subseteq \mathcal{U}$ is the subset of drivers that influence variable X_i and for vector $\boldsymbol{Z}$ and set A, $\boldsymbol{Z}_A$ indicates selection of components of the vector whose indices are members of the set; parameters $\boldsymbol{\theta}_i$ are required to complete the specification of f_i. According to physical laws or modelling hypotheses, the class of parametric functional forms which are permissible for f_i is denoted by $\mathcal{F}$.

We consider situations where the characteristic timescale for the driving inputs $\boldsymbol{U}$ is much longer than the timescale of $\boldsymbol{X}$ dynamics; in this sense X_i reaches *equilibrium* with respect to U_i. We suppose that the dynamics converge to a unique stationary point, depending on the driver, given implicitly by the equilibrium relation $\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{U}; \boldsymbol{\theta}) = \boldsymbol{0}$. We then proceed as follows. Firstly, using knowledge of the functional class $\mathcal{F}$ of the dynamical system, we construct a family of possible equilibrium relations. Secondly, by comparing these relations to equilibrium data $\mathcal{D}$, an RJMCMC sampler (Green, 1995) is constructed over the space of relations. Finally, model selection proceeds by selecting variates $j \in \mathcal{X}$ which appear frequently under the sampler; these variates constitute an estimate of the true causes π_i.

2.2 PROTEIN PHOSPHORYLATION

We consider p proteins, each of which has an unphosphorylated form Y_i and a phosphorylated form X_i ($i \in \mathcal{X}$). In this work we do not allow for the possibility of latent variables; this assumption is valid for both the simulated and real data examples given below. The chemical reaction that gives product X_i from substrate Y_i is known as *phosphorylation* and is catalysed by *kinases* $E \in \mathcal{E}_i$. We consider the case in which the kinases themselves are phosphorylated proteins such that $\mathcal{E}_i \subset \mathcal{X}$ (if phosphorylation is not driven by a kinase in $\mathcal{X}$, we set $\mathcal{E}_i = \varnothing$). The ability of a kinase $E \in \mathcal{E}_i$ to catalyse phosphorylation of X_i may be tempered by *inhibitors* $I \in \mathcal{I}_{i,E} \subset \mathcal{X}$; the double subscript indicates that inhibition is specific to both substrate and kinase. Thus, the causal influences π_i on X_i comprise both the kinases and their inhibitors: $\pi_i = \mathcal{E}_i \cup \{\mathcal{I}_{i,E}\}_{E \in \mathcal{E}_i}$. Due to specificity of phosphorylation reactions, the underlying causal graph G is typically sparse, such that the number of causes π_i for variate X_i is usually low. An example is shown, using

a standard graphical representation, in Fig. 1a. In what follows we use Y_i, X_i to denote the concentrations of proteins Y_i, X_i respectively; $U_i = Y_i + X_i$ is then the total amount of protein i, which is approximately invariant over the timescale of phosphorylation dynamics.

For CVS, model selection takes place over variable subsets π_i. Accordingly, for our approach we require a dynamical system for any such subset (Fig. 1b). Following the biochemical literature (Kholodenko, 2006; Steijaert *et al.*, 2010), we use ODEs of the Michaelis-Menten type to provide a suitable class $\mathcal{F}$ of analytic approximations for phosphorylation dynamics; the dynamics f_i are given in general by

$$f_i(\boldsymbol{X}_{\pi_i}, U_i; \boldsymbol{\theta}_i) = \tag{1}$$
$$-V_0 X_i + \sum_{E \in \mathcal{E}_i} \frac{V_E X_E (U_i - X_i)}{(U_i - X_i) + K_E\big(1 + \sum_{I \in \mathcal{I}_{i,E}} \frac{X_I}{K_I}\big)}$$

where here the parameter vector $\boldsymbol{\theta}_i$ contains the maximum rates ($\boldsymbol{V}$) and Michaelis-Menten constants ($\boldsymbol{K}$) specific to phosphorylation of species i (dependence of $\boldsymbol{V}, \boldsymbol{K}$ on i is notationally suppressed for clarity). The set γ_i of driver variables is simply $\{i\}$. When $\mathcal{E}_i = \varnothing$ we instead define $f_i = \mu_i$, equal to the average response. Equilibrium relations for phosphorylation are given implicitly by solving $f_i(\boldsymbol{X}_{\pi_i}, U_i; \boldsymbol{\theta}_i) = 0$. Further details of the chemical kinetic formulation and underlying assumptions appear in Appendix A.

2.3 STATISTICAL FORMULATION

Inference proceeds based on a Bayesian formulation of the chemical kinetic model (Fig. 1c). Below we present the details of our formulation.

Consider independent observations $\boldsymbol{Y}, \boldsymbol{X}$ of protein expression obtained at equilibrium. To fix a characteristic scale (required for prior elicitation, see below), all data are normalised prior to inference such that each species attains unit mean. For a given protein i, a model M_i for phosphorylation describes putative kinases and associated inhibitors for protein i (note that M_i contains more information than the subset π_i, namely the specific mechanistic roles played by each variable in π_i). Given a model M_i and associated parameters $\boldsymbol{\theta}_i$, the roots of Eqn. 1 completely determine the equilibrium expression of X_i as a function of inputs U_i. This equilibrium relationship is formulated statistically using nonlinear regression, so that conditional on M_i and $\boldsymbol{\theta}_i$ we have

$$\log(X_i) = \tag{2}$$
$$\log\left(\sum_{E \in \mathcal{E}_i} \frac{(V_E/V_0) X_E Y_i}{Y_i + K_E\big(1 + \sum_{I \in \mathcal{I}_{i,E}} \frac{X_I}{K_I}\big)}\right) + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$ and the parameter vector $\boldsymbol{\theta}_i$ is augmented with σ. The logarithm of both predictor and response is taken in order to improve the normality assumption on the error ϵ.

In the Bayesian setting, prior probability distributions are required for parameters $\boldsymbol{\theta}_i$ and models M_i. For the parameters $\boldsymbol{\theta} = (\boldsymbol{V}, \boldsymbol{K}, \sigma)$, physical and statistical considerations require that $V_j, K_j, \sigma > 0$. Following Xu *et al.* (2010) we postulate that all biological processes must occur on an observable timescale, motivating, in the shape, scale parametrisation, the gamma priors $V_j \sim \Gamma(2, 1/2)$, $K_j \sim \Gamma(2, 1/2)$, each of unit mean and variance $1/2$. The noise parameter σ is inverse-gamma distributed *a priori* as $\sigma \sim \Gamma^{-1}(6, 1)$, with prior mean $1/5$ chosen to correspond to the magnitude of measurement noise in current proteomic technologies (Hennessey *et al.*, 2010), and variance $1/100$.

When expert opinion is available, rich subjective model priors may be elicited (see e.g., for graphical models, Mukherjee and Speed, 2008). In this work we employed a multiplicity correction prior, depending on a (possibly empty) prior model M_i^0 (Scott and Berger, 2010). For simplicity in the exposition, each variable X_j is constrained to appear in at most one of the sets $\mathcal{E}^M$, $\mathcal{I}_E^M$ (for clarity here and in the following section our notation emphasises dependence on the model M_i whilst suppressing dependence on the response protein i). Let π^M denote the variables included in model M and let π^0 denote the variables included in M^0. To avoid explicit computation of the normalising constant, which is not directly required by our methodology, the model prior is specified indirectly using prior odds $p(M')/p(M)$. These are uniquely determined by, and calculated from, iterative application of the following three criteria:

1. $p(M) = 0$ if M^0 is not nested in M. (M^0 is said to be nested in M if, for each $E \in \mathcal{E}^{M^0}$ we have $E \in \mathcal{E}^M$ and $\mathcal{I}_E^{M^0} \subseteq \mathcal{I}_E^M$.)

2. $p(M')/p(M) = (|\mathcal{X}| - |\pi^M|)^{-1}$ if M' differs to M by the addition of a single (uninhibited) kinase. ($\mathcal{E}^{M'} = \mathcal{E}^M \cup \{E\}$, $E \notin \mathcal{E}^M$, $\mathcal{I}_E^{M'} = \varnothing$ and $\mathcal{I}_{E'}^{M'} = \mathcal{I}_{E'}^M$ for all $E' \in \mathcal{E}^M$.) There are $|\mathcal{X}| - |\pi^M|$ such models M', so that the prior distribution is uniform over such models.

3. $p(M')/p(M) = ((|\mathcal{X}| - |\pi^M|) \times |\mathcal{E}^M|)^{-1}$ if M' differs to M in the addition of a single kinase inhibitor. ($\mathcal{E}^{M'} = \mathcal{E}^M$, $E \in \mathcal{E}^{M'}$, $\mathcal{I}_E^{M'} = \mathcal{I}_E^M \cup \{I\}$, $I \notin \mathcal{I}_E^M$ and $\mathcal{I}_{E'}^{M'} = \mathcal{I}_{E'}^M$ for all $E' \in \mathcal{E}^M$, $E' \neq E$.) Again, there are $(|\mathcal{X}| - |\pi^M|) \times |\mathcal{E}^M|$ such models M', with the prior distribution uniform over such models.

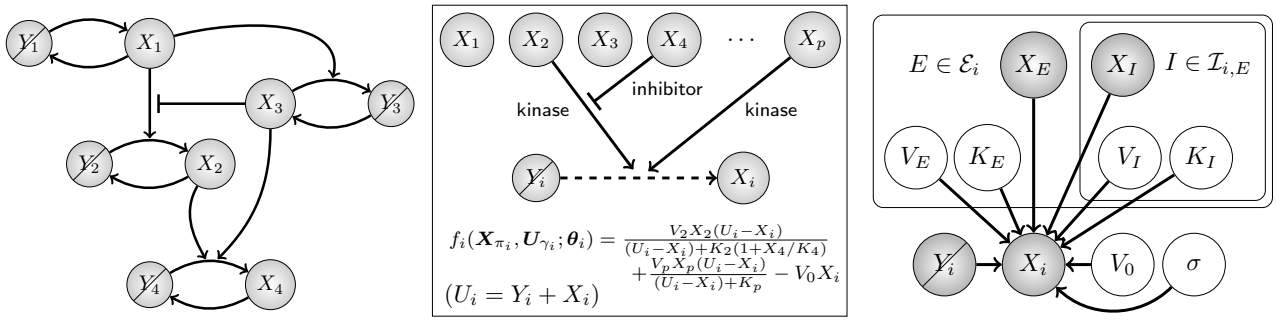


Figure 1: Overview of approach. (a) An example of a phosphorylation network. (b) Our approach couples automatic generation of dynamical models with variable selection to infer causes π_i of phosphorylation for species i. (c) A statistical formulation (graphical model) for equilibrium phosphorylation of species i is characterised by specifying kinases ($E \in \mathcal{E}_i$) and inhibitors ($I \in \mathcal{I}_{i,E}$) of kinases. Bounding boxes are used to indicate multiplicity of variables, shaded nodes are observed with noise.

2.4 REVERSIBLE JUMP MCMC

The dimensionality of the parameter vector $\boldsymbol{\theta}^M$ depends on the model M; $\dim(\boldsymbol{\theta}^M) = 1 + \dim(\boldsymbol{V}^M) + \dim(\boldsymbol{K}^M)$ where the latter quantities are functions of the numbers of kinases and inhibitors according to M. Since we seek models with high posterior probability, and noting that most models will provide insufficient explanatory power, we implement RJMCMC (Green, 1995) to reduce the effective size of model space. Following Green and Hastie (2009) we enumerate all possible models as $\{M^{(k)}\}_{k \in \mathcal{K}}$ and define the *across-model* state space

$$\mathcal{S} = \bigcup_{k \in \mathcal{K}} (\{k\} \times \Theta_k), \quad k = \bigtimes_{E \in \mathcal{E}^{M^{(k)}}} (\{E\} \times \mathcal{I}_E^{M^{(k)}}) \quad (3)$$

where parameters $\boldsymbol{\theta}^{M^{(k)}}$ for model $M^{(k)}$ belong to Θ_k and $\times$ denotes the Cartesian product. The reversible jump sampler constructs an ergodic Markov chain on $\mathcal{S}$ which has, as its stationary distribution, the posterior probability distribution $p(s|\mathcal{D}), s \in \mathcal{S}$. In particular the marginal $p(k|\mathcal{D})$ over the model index $k \in \mathcal{K}$ corresponds exactly to the posterior model probabilities $p(M^{(k)}|\mathcal{D})$. Construction of an efficient RJMCMC sampler requires an intuition for the across model state space. We adopt a deliberately transparent Metropolis-within-Gibbs approach (Roberts and Rosenthal, 2006), updating one coordinate of $\mathcal{S}$ at a time using a Metropolis-Hastings accept/reject probability of the form $\alpha(s, s') = \min(1, A(s, s')p(\mathcal{D}|s')/p(\mathcal{D}|s))$, where $A(s, s')$ is specified below. This calls for updates of two different kinds:

1. *Update the parameter values* $\boldsymbol{\theta}^M$. For convenience we used symmetric random walk proposal distributions which do not require density evaluation during a Metropolis-Hastings accept/reject step (i.e. $A(s, s') = 1$). All parameters were updated simultaneously, with rate constants, Michaelis-Menten parameters and noise parameter proposed according to $\log(V_i') \sim N(\log(V_i), \lambda_1^2)$, $\log(K_i') \sim N(\log(K_i), \lambda_2^2)$, $\log(\sigma') \sim N(\log(\sigma), \lambda_3^2)$ respectively. Specific values for $\boldsymbol{\lambda}$ are discussed below.

2. *Update the model index* $k \in \mathcal{K}$. This may be accomplished in one of the following four ways:

(a) *Add or remove a kinase* $E \notin \pi^0$. Move 2(a) begins by choosing to add or remove a kinase with equal probability, then selecting a candidate for the addition (or removal) uniformly from all candidates. If no candidate is available (e.g. there are no kinases left to remove) then no move is performed. To add an additional kinase ($\mathcal{E}^{M'} = \mathcal{E}^M \cup \{E\}$, $E \notin \mathcal{E}^M$, $\mathcal{I}_E^{M'} = \varnothing$) a new rate constant V_E and Michaelis-Menten parameter K_E are generated from the prior, whilst remaining parameters remain unchanged. Reversibility implies that removal of a kinase ($\mathcal{E}^{M'} = \mathcal{E}^M \backslash \{E\}$, $E \in \mathcal{E}^M$) corresponds to deletion of the removed kinase's parameters whilst the remaining parameters are left unchanged. We satisfy the dimension matching requirement by insisting new kinases initially have no inhibitors; conversely only kinases without inhibitors may be removed. (Informally, *dimension matching* requires that the number of parameters in the current model, plus the rank of the proposal distribution, must equal the number of parameters in the proposed model, plus the rank of the reverse proposal. This ensures all transformations are diffeomorphisms and thus *reversible*; see Green and Hastie (2009) for further details.) For addition of a kinase $A(s, s') = |\{E \in \mathcal{E}^{M'} : \mathcal{I}_E^{M'} = \varnothing,\ E \notin \mathcal{E}^{M^0}\}|^{-1}$ whereas for removal

$A(s,s') = |\{E \in \mathcal{E}^M : \mathcal{I}_E^M = \varnothing,\ E \notin \mathcal{E}^{M^0}\}|$.
(b) *Add or remove a kinase inhibitor* $I \notin \pi^0$. Addition of an inhibitor ($\mathcal{I}_E^{M'} = \mathcal{I}_E^M \cup \{I\}$, $I \notin \mathcal{I}_E^M$) proceeds analogously, selecting uniformly from available candidates and sampling a Michaelis-Menten parameter K_I from the prior, leaving the other parameters unchanged. Addition and removal ($\mathcal{I}_E^{M'} = \mathcal{I}_E^M \backslash \{I\}$, $I \in \mathcal{I}_E^M$) have respectively $A(s,s') = \left|\bigcup_{E \in \mathcal{E}^{M'}} \mathcal{I}_E^{M'} \backslash \pi^0\right|^{-1}$, $A(s,s') = \left|\bigcup_{E \in \mathcal{E}^M} \mathcal{I}_E^M \backslash \pi^0\right|$.
(c) *Swap one kinase for another* $E_1 \leftrightarrow E_2$, $E_1 \in \pi^M \backslash \pi^0, E_2 \notin \pi^M$. In this instance the new kinase E_2 is assigned the same inhibitors and parameters as the departing kinase (such a move is trivially reversible). The uniform proposal distribution was used, so that $A(s,s') = 1$.
(d) *Swap one inhibitor for another* $I_1 \leftrightarrow I_2$, $I_1 \in \pi^M \backslash \pi^0, I_2 \notin \pi^M$. The new inhibitor inherits all parameter values associated with the previous inhibitor. Again, the uniform proposal distribution was used so that $A(s,s') = 1$.

Moves 2(a) and 2(b) ensure irreducibility of the RJMCMC scheme over the space $\mathcal{K}$ of models. Irreducibility over the across-model space $\mathcal{S}$ follows almost immediately, so that ergodicity of the chain is assured. However, a theoretical guarantee of ergodicity does not guarantee practical convergence of the chain; in particular the mixing time will depend heavily on both the parameter proposal scales $\boldsymbol{\lambda}$ and the Gibbs schedule. Through experimentation we found that, for our datasets, the proposal scale parameters $\boldsymbol{\lambda} = (0.1, 0.1, 0.1)$ delivered acceptable mixing. At the Gibbs level, the deterministic schedule $1 \to 2(a) \to 1 \to 2(b) \to 1 \to 2(c) \to 1 \to 2(d) \to 1\ldots$ offered efficient mixing and was used for all computations in this paper. For applications, 30,000 iterations of the Gibbs sampler were performed, with 5,000 discarded as burn-in. Convergence was assessed using repeated runs from dispersed initial conditions.

3 RESULTS

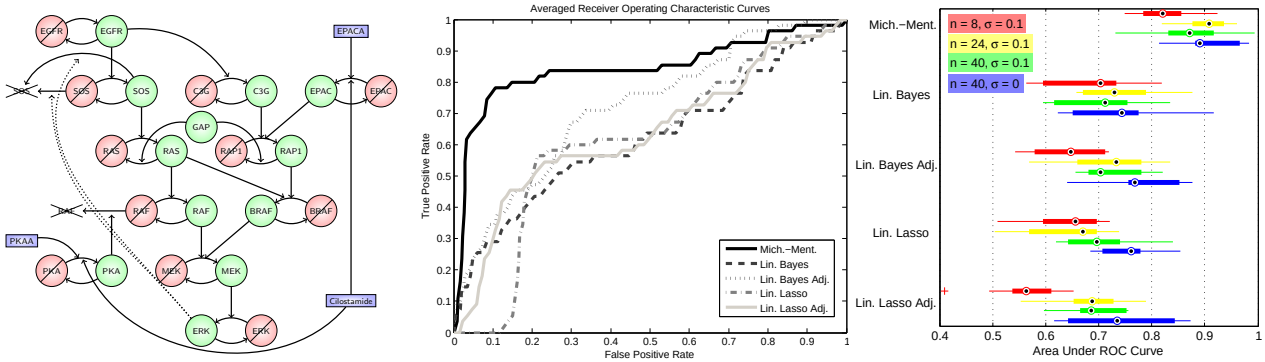
In this section we empirically assess our methodology and compare its performance with variable selection based on the linear model. In Section 3.1 we show results using a recently published dynamical model of a biological system due to Xu *et al.* (2010), where the underlying causal graph G is known exactly. In Section 3.2 we apply our approach to a real proteomic dataset, which has an unknown and likely more complex noise structure. Note that for both examples, in line with our goals, we assess methods exclusively in terms of ability to identify truly causal variables. Gold-standard assessment is possible due to the fact that, in the first example, causal relationships are known by design, whereas in the second example, they are known from extensive biochemical and biophysical experiments. The second example is particularly challenging since Michaelis-Menten functionals most likely represent a crude approximation to the true data-generating system. In both cases, for fair comparison between different methods, no informative model priors were used (i.e. we set $M^0 = \varnothing$).

3.1 SIMULATION STUDY

Data were generated from a computational model of the MAPK signaling pathway due to Xu *et al.* (2010), specified by a system of 25 nonlinear ODEs (Fig. 2a). The simulation gives covariates that are highly correlated, as would be expected in practice, whilst providing a known causal graph G for evaluation purposes. We note that, as is often the case in biological systems, the causal graph G is not acyclic (Fig. 2a, dashed edges). Some further details regarding the computational model are described in Appendix B. We introduced independent Gaussian measurement noise, additive on the log scale, of magnitude $\sigma = 0.1$, similar to current proteomic technologies (Hennessey *et al.*, 2010).

We benchmarked our approach against the linear-additive-Gaussian formulation $\log(\boldsymbol{X}_i) \sim N(\mathbf{1}\beta_0 + \boldsymbol{D}_M\boldsymbol{\beta}_M, \sigma^2\boldsymbol{I})$ with design matrix $\boldsymbol{D}_M = [\ldots \log(\boldsymbol{X}_j) \ldots]_{j \in \pi^M}$; the logarithm of a vector is taken component-wise. All variables were mean-variance standardised prior to inference. We consider two standard approaches to inference for the linear model, namely (1) the LASSO with penalty parameter set according to cross validation ("Lin. Lasso"), and (2) a conjugate Bayesian formulation ("Lin. Bayes"), based on the g-prior $\boldsymbol{\beta}_M \sim N(\mathbf{0}, n\sigma^2(\boldsymbol{D}_M'\boldsymbol{D}_M)^{-1})$, with a flat prior over the intercept $p(\beta_0) \propto 1$ and reference prior over the noise $p(\sigma) \propto 1/\sigma$ (Zellner, 1986). For the Bayesian approach we took a model prior $p(M)$ to be uniform over in-degree $d = \dim(\boldsymbol{\beta}_M)$ with the restriction $d \leqslant 3$. Model averaging was then used to obtain posterior inclusion probabilities. For each of the linear approaches (1) and (2) we also considered *adjusted* variants ("Lin. Lasso Adj." and "Lin. Bayes Adj.") where log-phospho-ratios $\log(X_i/Y_i)$ constitute the response; this can be motivated as a simple first order correction for variation in total protein levels.

For each phosphorylated or active species $i \in \mathcal{X}$ in the computational model, we sought to infer the causal variates π_i. For a fair comparison with the linear approaches that do not ascribe functional roles to variables, we did not distinguish between kinases and inhibitors during assessment. The resulting receiver operating characteristic (ROC) curves are shown in Fig.



(a) (b) (c)

Figure 2: Simulation study. (a) Computational model of the MAPK signalling pathway (Xu *et al.*, 2010). Circles represent proteins, rectangles represent interventions (drug treatments) used to perturb the system. For proteins, one strike-through represents inactivity, two strikes represent degradation. Dotted lines correspond to feedback mechanisms. (b) Average receiver operating characteristic (ROC) curves (sample size $n = 24$, noise $\sigma = 0.1$) using data generated from model (a). (c) Area under ROC curve (AUR) for each of the sample size (n) and noise (σ) regimes shown (boxplots over 5 datasets for each n, σ regime). [Legend: "Mich.-Ment": variable selection using equilibrium relations derived from Michaelis-Menten kinetics as described in text; "Lin. Bayes": Bayesian variable selection using linear model; "Lin. Lasso": variable selection using LASSO and linear model; "Lin. Bayes Adj." and "Lin. Lasso Adj.": as previous but corrected for total protein levels as described in text.]

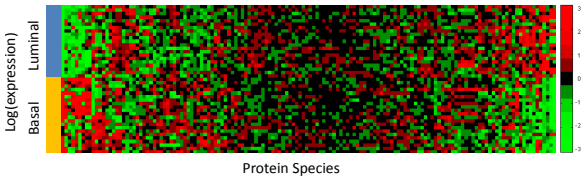
2b. For Bayesian analyses, ROC curves were generated by thresholding posterior edge probabilities at a level L which was then varied. Similarly LASSO coefficients were thresholded to produce a ROC curve. (Note that since the data are standardised, a larger coefficient represents a stronger indication that a covariate is implicated in the response.) Finally, ROC curves were combined by threshold averaging (Fawcett, 2006). Overall performance was quantified using area under the ROC curve (AUR), aggregated over all $i \in \mathcal{X}$. Results are shown over 5 datasets $\mathcal{D}$ for each of various combinations of sample size n and noise level σ (Fig. 2c). In all regimes our approach outperformed linear approaches; the latter did not perform well even in this low dimensional example. We note also that even in the least challenging regime ($n = 40$, $\sigma = 0$), none of the approaches were able to perfectly recover the entire causal graph G. The adjusted regressions, which model the log-phospho-ratio as the response, did not outperform the standard linear regressions.

3.2 CANCER PROTEOMIC DATA

Data were obtained using reverse-phase protein arrays (Hennessey *et al.*, 2010) applied to a panel of breast cancer cell lines (Neve *et al.*, 2006). Data $\mathcal{D}$ comprised equilibrium observations for $p = 38$ phosphorylated proteins, in addition to their unphosphoryated counterparts (Fig. 3a; data courtesy the lab of Gordon Mills, Department of Systems Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA). Total protein levels were used as a proxy for unphosphorylated protein concentrations Y_i. Cell lines were aggregated by subtype into *basal* ($n = 22$) and *luminal* ($n = 21$), with each member cell line comprising one sample. Fig. 3b displays the result of inference for the causes π_{S6} for phosphorylation of the protein S6 on amino acid residues Serine 235,236. We excluded measurements of S6 phosphorylation on residues Serine 240,244 since this correlates closely with phosphorylation on Serine 235,236, and therefore particularly challenges linear approaches. It is known in the biological literature that p70 S6 kinase (p70S6K) acts as the kinase for S6. However, only our approach ranks p70S6K highly (bold in Fig. 3b; ranking shown in Fig. 3c) and only for the basal subtype (Fig. 3c). In addition, it is noticeable that there is little agreement between the apparently similar linear formulations.

4 DISCUSSION

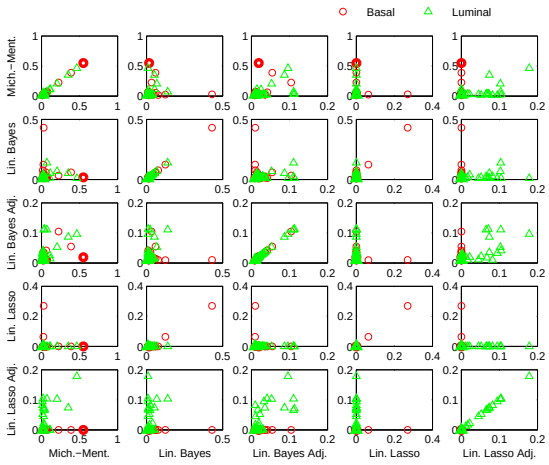
Motivated by the availability of classes of dynamical models in many settings and recent insights into causal inference using nonlinear models (Peters *et al.*, 2011), we sought to explore causal variable selection rooted in nonlinear dynamical systems. We focused on the setting of steady-state data, investigating whether nonlinear functional classes, derived from equilibrium analy-



(a)

Rank of the kinase p70S6K:	Basal	Luminal
Mich.-Ment.	1/36	24/36
Lin. Bayes	15/36	36/36
Lin. Bayes Adj.	8/36	36/36
Lin. Lasso	20/36	NA
Lin. Lasso Adj.	19/36	NA

(c)



(b)

Figure 3: Cancer protein data. (a) We investigated inference for the causal activators (kinases) of S6 using data obtained from cancer cell lines. (b) Comparison of methodologies: Each point in the scatter plots represents one phosphorylated protein, with points corresponding to the known causal influence p70S6K highlighted in bold. Here we display weights (posterior probabilities or absolute regression coefficients) attached to each protein. (c) Rank of p70S6K, the known activator (kinase), under different methodologies. For the luminal subtype, LASSO approaches selected the null model.

sis of dynamical models, could enhance the ability to infer causal influences compared to linear models.

We focused on an illustrative example from biology, namely protein phosphorylation. Phosphorylation is a key biochemical process where the availability of relatively sophisticated "gold-standard" models, extent of existing mechanistic insight and availability of relevant data combine to facilitate assessment of approaches for CVS. Our results, on simulated and real data, demonstrated that causal relationships in protein phosphorylation are estimated much more successfully under our approach than by conventional linear formulations. Note that the correct model hypothesis is violated in the real data example, since both our dynamical system (Michaelis-Menten kinetics) and noise model (Gaussian) are only crude approximations to the physical system. However, in this particular example the identity of the true causal regulator in indeed known from careful biophysical studies. It is therefore encouraging that we could correctly identify the causal regulator, despite model mis-specification.

It is important to note that, for the biochemical models we considered, all structural information regarding dynamics was identifiable from steady state observations. Of course, this does not hold in general and many systems do not enjoy this favourable property. However, a wide range of biological settings can be modelled using dynamical descriptions similar to the Michaelis-Menten functionals used here. Therefore, our work should generalise to a number of biological and chemical settings, including, among others, gene regulation. Thus, extensions of our work could offer the potential for substantive gains in a broad application domain (molecular and systems biology) where the machine learning approaches of variable selection and structural inference are currently widely used.

Attention was restricted to the equilibrium setting. In the general case, the equilibrium probability distribution can be *unfaithful* to the equilibrium graph so that the *do*-calculus (Pearl, 2009) may not apply. Dash (2003) formulated a criterion, known as *equilibration-manipulation commutability* (EMC), which characterises causal faithfulness at equilibrium. Put simply, for causal reasoning based on the equilibrium graph to be valid, the *equilibrate* and the *do* operators must commute. Our formulation of protein phosphorylation trivially satisfies the EMC criterion, since graph topology is invariant under the *equilibrate* operator. However, to fully generalise the ideas presented here would require investigation of these issues.

The linear approaches we used failed to perform well on simulated data and to identify known causal influences in real data. Further, we saw that apparently similar linear formulations can return very different recommendations for which predictors ought to be included in the model; one possible explanation for such disagreement may be model misspecification. In addition to superior performance in CVS, our formulation benefits from increased interpretability, ascribing mechanistic roles to variables and relating parameters to scientifically interpretable properties.

Peters *et al.* (2011) recently filled an important theoretical gap, demonstrating that within an *identifiable functional model class* (IFMOC) it is possible to consistently estimate causal relationships. This is an important step in thinking about causal inference using nonlinear models and emphasises the limitations that arise from symmetry inherent in the the linear-additive-Gaussian model. However, in order to formally show that a given functional class constitutes an IFMOC, the theory at present requires strong assumptions, including noise-free observation, that do not hold in the systems we considered here. We demonstrated that basing the likelihood on a relevant nonlinear dynamical system can lead to improved performance in CVS, under practical conditions of sample size and noise, even when observations are made at equilibrium only. In this sense, our contribution complements the theoretical results of Peters *et al.* (2011). We did not investigate whether our formulation also led to improved predictive performance (in the regression sense). It would be interesting to investigate whether improved CVS performance also confers improved predictive ability.

CVS is naturally facilitated by interventional experiments. Adequate modelling of the effects of intervention is important to ameliorate statistical confounding (Pearl, 2009). In testing, not presented here, we extended our methodology to incorporate *imperfect certain intervention*, where the interventional targets are assumed known, but the interventions may not completely block causal influences of their targets (see Eaton and Murphy, 2007, for a general discussion of interventions in graphical models). In the context of protein phosphorylation, kinases and their inhibitors can be intervened upon using agents (such as monoclonal antibodies or small molecule inhibitors). We modelled these effects by rescaling the effective concentration of interventional targets, in the presence of the treatment, as $X_j \mapsto \alpha_j X_j$ where $0 \leqslant \alpha_j \leqslant 1$ is an unknown parameter capturing interventional efficacy of the agent. Using this extended methodology we observed that interventional experiments were more informative than the global perturbation experiments considered here, leading to improved AUR scores.

Variable selection based on nonlinear models is computationally challenging. We considered low-to-moderate dimensional settings ($p = 12, 38$), for which the RJMCMC proved to be very effective. Many of the computations here are trivially parallelisable, and it may therefore be possible to extend our work to the high-dimensional setting, using tools of the kind discussed by Lee *et al.* (2010). In general, nonlinear approaches are clearly more burdensome than their linear counterparts, where highly efficient approaches, including those based on LASSO and related penalised likelihood schemes, allow rapid estimation even in high dimensions. We therefore view the methods presented here as complementary to variable selection based on linear models, allowing more refined exploration of causal influences in settings where some insight into underlying dynamics is available.

A PHOSPHORYLATION KINETICS

This short appendix describes assumptions needed to arrive at the chemical model for protein phosphorylation described above. For simplicity we did not explicitly distinguish between phosphorylation on different residues. The molecular mechanism of kinase inhibition entertained was competitive inhibition, where substrate (S) and inhibitor (I) compete for the same binding site on the enzyme (E), expressed chemically as $EI \rightleftharpoons E \rightleftharpoons ES$. Furthermore, when multiple inhibitors (I^A, I^B) are present they were assumed to act exclusively, competing for the same binding site on the enzyme $EI^A \rightleftharpoons E \rightleftharpoons EI^B$. Dephosphorylation was assumed to occur at a rate proportional to the amount of phosphorylated protein. The methodology which we presented can be generalised to other molecular mechanisms; in particular additional mechanisms such as noncompetitive, uncompetitive, hyperbolic and parabolic inhibition (Leskovac, 2003) could be readily integrated into our framework.

B COMPUTATIONAL MODEL

The computational model of Xu *et al.* (2010) contains 25 different species, of which $p = 12$ are *active* (shown in green in Fig. 2a). The model allows for treatments, exogenous to the statistical model, to be simulated, effectively resulting in global perturbations of the system (rather than interventions). Data were generated under combinatorial treatment with Cilostamide, PKAA and EPACA (blue in Fig. 2a; we direct the interested reader to the reference for full details of the model and treatments), with r samples taken at equilibrium under each treatment regime, giving a total of $n = 8r$ independent samples. For each sample, initial total protein levels were drawn independently from the uniform distribution $U_i \sim U[0, 1]$, mimicking, to some extent, natural variation due to transcriptional regulation.

Acknowledgments

We wish to thank the anonymous referees for suggestions which have improved the content and presentation of this article.

References

Äijö, T., Lähdesmaki, H. (2009) Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics, *Bioinformatics*, **25**(22), 2937-2944.

Cho, H., Fryzlewicz, P. (2012) High dimensional variable selection via tilting, *Journal of the Royal Statistical Society, Series B*, to appear.

Dash, D. (2003) Caveats for Causal Reasoning with Equilibrium Models, PhD thesis, Intelligent Systems Program, University of Pittsburgh.

Eaton D, Murphy K (2007) Exact Bayesian structure learning from uncertain interventions, *Proceedings of the 11th Conference on Artificial Intelligence and Statistics (AISTATS-07)*.

Fawcett, T. (2006) An introduction to ROC analysis, *Pattern Recognition Letters*, **27**(8), 861-874.

Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**(4), 711-732.

Green, P., Hastie, D. (2009) Reversible jump MCMC, technical report (http://www.maths.bris.ac.uk/mapjg/Papers.html).

Heagerty, P.J., Kurland, B.F. (2001) Misspecified Maximum Likelihood Estimates and Generalised Linear Mixed Models, *Biometrika*, **88**(4), 973-985.

Hecker, M. *et al.* (2009) Gene regulatory network inference: Data integration in dynamic models - A review, *Biosystems*, **96**(1), 86-103.

Hennessey, B.T. *et al.* (2010) A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Nonmicrodissected Human Breast Cancer, *Clinical Proteomics*, **6**, 129-151.

Hill, S. *et al.* (2012) Integrating biological knowledge into variable selection: an empirical Bayes approach with an application in cancer biology, *BMC Bioinformatics*, **13**(94), 1471-2105.

Honkela, A. *et al.* (2010) Model-based method for transcription factor target identification with limited data, *Proceedings of the National Academy of Sciences USA*, **107**(17), 77937798.

Iwasaki, Y., Simon, H.A. (1994) Causality and model abstraction, *Artificial Intelligence*, **67**(1), 143194.

Kholodenko, B.N. (2006) Cell-signalling dynamics in time and space, *Nature Reviews Molecular Cell Biology*, **7**(3), 165-176.

Lee, A. *et al.* (2010) On the Utility of Graphics Cards to Perform Massively Parallel Simulation of Advanced Monte Carlo Methods, *Journal of Computational and Graphical Statistics*, **19**(4), 769-789.

Leskovac, V. (2003) Comprehensive enzyme kinetics, Springer.

Lv, J., Liu, J.S. (2010) Model Selection Principles in Misspecified Models, Technical Report, arXiv:1005.5483v1.

Mukherjee, S., Speed, T.P. (2008) Network inference using informative priors, *Proceedings of the National Academy of Sciences USA*, **105**(38), 14313-14318.

Neve, R. *et al.* (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes, *Cancer Cell*, **10**(6), 515-527.

Oates, C.J., Mukherjee, S. (2012) Network inference and biological dynamics, *Annals of Applied Statistics*, to appear.

Pearl, J. (2009) Causal inference in statistics: An overview, *Statistics Surveys*, **3**, 96-146.

Peters, J. *et al.* (2011) Identifiability of Causal Graphs using Functional Models, *Proceedings of the 27th Annual Conference: Uncertainty in Artificial Intelligence (UAI-11)*, 589-598.

Roberts, G.O., Rosenthal, J.S. (2006) Harris Recurrence of Metropolis-within-Gibbs and Trans-Dimensional Markov Chains, *Annals of Applied Probability*, **16**(4), 2123-2139.

Sachs, K. *et al* (2005) Causal protein-signaling networks derived from multiparameter single-cell data, *Science*, **308**, 5239.

Schoeberl, B. *et al.* (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors, *Nature Biotechnology*, **20**(4), 370-375.

Scott, J.G., Berger, J.O. (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem, *Annals of Statistics*, **38**(5), 2587-2619.

Steijaert, M.N. *et al.* (2010) Computing the Stochastic Dynamics of Phosphorylation Networks, *Journal of Computational Biology*, **17**(2), 189-199.

Xu, T. *et al.* (2010) Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species, *Science Signaling*, **3**(113), ra20.

Zellner, A. (1986) On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions, *Bayesian Inference and Decision Techniques - Essays in Honor of Bruno de Finetti, eds. P. K. Goel and A. Zellner*, 233-24.