Structured Proportional Jump Processes

Tal El-Hay **Omer Weissbrod** **Elad Eban**[*]
Machine Learning for Healthcare
and Life Sciences
IBM Research – Haifa

Maurizio Zazzi
Department of Medical Biotechnologies
University of Siena, Italy

Francesca Incardona
InformaPRO S.r.l.
EuResist Network GEIE, Italy

Abstract

Learning the association between observed variables and future trajectories of continuous-time stochastic processes is a fundamental task in dynamic modeling. Often the dynamics are non-homogeneous and involve a large number of interacting components. We introduce a conditional probabilistic model that captures such dynamics, while maintaining scalability and providing an explicit way to express the interrelation between the system components. The principal idea is a factorization of the model into two distinct elements: one depends only on time and the other depends on the system configuration. We developed a learning procedure, given either full or point observations, and tested it on simulated data. We applied the proposed modeling scheme to study large cohorts of diabetes and HIV patients, and demonstrate that the factorization helps shed light on the dynamics of these diseases.

1 INTRODUCTION

Studies of dynamic systems often attempt to investigate the dependency of these dynamics on a set of static explanatory variables. In many cases, the studied process is composed of interrelated components that evolve continuously in time; hence, inter-component interactions are of interest as well. Examples appear in diverse fields, ranging from medicine to computational biology and economics.

Inferring such conditional dynamics of a real life system involves several challenges. We illustrate these challenges by our motivating example of studying disease progression in patients infected with *Human Immunodeficiency Virus* (HIV). The two common measures of the severity of HIV infection, viral load and the immune system CD4 protein count, are interrelated. A higher viral load weakens the immune system, while a weakened immune system potentially affects viral dynamics. A weakened immune system also increases the risk of death, either directly or by increasing the risk of contracting other diseases. These dynamics are depicted in Figure 1.

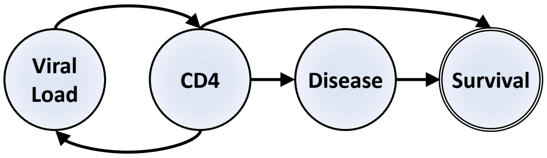


Figure 1: A graphical representation of a SCUP model for HIV. Directed arcs indicate directions of influence.

The typical properties of dynamic processes are a non-constant states transition rate (*non-homogeneity*), and the ability to observe the process states at only a finite set of time-points (*point observations*). Additionally, the processes may include highly diverse explanatory variables whose distribution is often difficult to learn. For example, this might include the type of drugs taken by each individual and their viral genome at that time. Due to these properties, a modeling framework for such processes should account for non-homogeneity, deal with point observations, be scalable in the number of components, and provide a robust way to account for observed explanatory variables without modeling their distribution.

The seminal work of (Cox, 1972) laid the foundations for rigorous analysis of the dynamics of non-homogeneous irreversible processes by introducing the *proportional hazard model*. A key point of this model is its focus on modeling the dynamics of a single binary-valued variable *conditioned* on some set of background variables. The proportional hazard framework proposed by Cox turned out to be extremely useful in modeling processes such

[*] Currently at: The Selim and Rachel Benin School of Computer Science and Engineering. The Hebrew University of Jerusalem.

as survival after medical procedures, how specific drugs affect a disease, the failure of manufactured components, and many more.

In recent years, several extensions of this model have been proposed (e.g., (Du et al., 2013)). One notable extension of the Cox model is Multi-State models (MSTMs), which model single component processes that can occupy one of a finite number of states at each time point (Putter et al., 2007). MSTMs support non-homogeneity and learning from point observations, and allow us to condition the dynamics on explanatory variables, resolving the difficulties in explicitly modeling covariates.

MSTMs are increasingly being used in medical and epidemiological studies (e.g., (Looker et al., 2013; Walker et al., 2013; Taghipour et al., 2013)). Nevertheless, MSTMs are not naturally suited for analyzing multi-component processes, because they require defining a state space corresponding to the Cartesian product of the individual components, resulting in a representation that is exponential in the number of components.

In this paper we consider modeling the conditional distribution of a non-homogeneous multi-dimensional continuous-time Markov process $\mathbf{Y}(t) = Y_1(t), \ldots, Y_n(t)$ given a set of covariates $\mathbf{x} \in R^p$, which we refer to as background covariates. Our goal is to construct a modeling language that is compact, interpretable, and scalable, meaning that it allows learning dependencies of specific components on specific covariates as well as on other components, while allowing efficient inference and learning.

A Continuous-Time Bayesian Network (CTBN) (Nodelman et al., 2002), the continuous-time extension of a dynamic Bayesian network, is a framework that enables the modeling of high-dimensional processes with complex dependencies; these dependencies are expressed via an interpretable network topology. CTBNs naturally deal with missing data, using exact inference for small topologies, and a variety of approximate methods for large topologies. Therefore, the principles that underlie CTBNs can serve as a basis to scale up MSTMs, by introducing structured representation and accompanying mathematical machinery from CTBNs.

In this work, we define StruCtured proportional jUmp Processes (SCUP), a new model combining ideas from the fields of proportional hazard models, MSTMs, and CTBNs. Our key modeling assumption decomposes the dynamics of the process into two elements. The first element is the *effect of time* on the dynamics of each individual component, independently of the others. The second element represent the dependence of the *evolution of each component* on the background covariates, as well as on the state of the other components. This decomposition allows a compact representation of the combined effect of non-homogeneity, background variables, and interactions among components. We show how this model can be learned from point observations and demonstrate the properties of our approach on synthetic data, as well as on large cohorts of data from diabetes and HIV patients. Our analysis helps identify reliable markers that may predispose diabetes and HIV patients to medical complications. Namely, we find that routine blood tests can serve as a biomarker for an increase in glycated hemoglobin, which is a highly reliable marker for complications among diabetes patients.

2 BACKGROUND

A multi-component continuous-time stochastic process over a discrete state space is a collection of vector-valued random variables $\{\mathbf{Y}(t) = Y_1(t), \ldots, Y_n(t) | t \geq 0\}$ where for each component i, $Y_i(t)$ takes on values from a finite set S_i. We say that such a process is *Markovian* if, for all sequences $t_1 \leq t_2 \leq \ldots \leq t_k$, it satisfies

$$\begin{aligned}\Pr(\mathbf{Y}(t_k) = \mathbf{y}_k | \mathbf{Y}(t_{k-1}) = \mathbf{y}_{k-1}, \ldots, \mathbf{Y}(t_1) = \mathbf{y}_1) \\ = \Pr(\mathbf{Y}(t_k) = \mathbf{y}_k | \mathbf{Y}(t_{k-1}) = \mathbf{y}_{k-1})\end{aligned}$$

Continuous time Markov processes are completely characterized by the rate function $q(t; \mathbf{y}, \mathbf{y}')$, which determines the instantaneous transition probability between states:

$$\Pr(\mathbf{Y}(t+\Delta t) = \mathbf{y}' | \mathbf{Y}(t) = \mathbf{y})) = 1_{\mathbf{y}=\mathbf{y}'} + q(t; \mathbf{y}, \mathbf{y}')\Delta t + o(\Delta t) \quad (1)$$

where 1 is the indicator function and $o(\Delta t)$ is a function that converges to zero faster than its argument, i.e., $\lim_{\Delta t \downarrow 0} \frac{o(\Delta t)}{\Delta t} = 0$. The rate functions are non-negative for every $\mathbf{y} \neq \mathbf{y}'$. The diagonal elements $q(t; \mathbf{y}, \mathbf{y})$ are the exit rates from state $\mathbf{y}$ at time t that satisfy $q(t; \mathbf{y}, \mathbf{y}) = -\sum_{\mathbf{y}' \neq \mathbf{y}} q(t; \mathbf{y}, \mathbf{y}')$. A Markov jump process is *homogeneous* if the rates do not depend on time, i.e., $q(t; \mathbf{y}, \mathbf{y}') = q_{\mathbf{y},\mathbf{y}'}$, otherwise it is *non-homogeneous*

Continuous time Bayesian networks (CTBNs) provide a compact representation for homogeneous Markov processes where only one component can change at a time, and where the instantaneous dynamics of each component i are influenced by a small set of parent components denoted by $pa(i)$. We refer to $pa(i)$ as the *context* of the component i. These assumptions are encoded by setting $q(t; \mathbf{y}, \mathbf{y}') = 0$ when $\mathbf{y}$ and $\mathbf{y}'$ differ by more than one component, and $q(t; \mathbf{y}, \mathbf{y}') = q_{y_i, y_i' | \mathbf{y}_{pa(i)}}$ when they differ in component i, where y_i and $\mathbf{y}_{pa(i)}$ are the states of component i and of the subset $pa(i)$, respectively. This dependency structure is represented by a directed graph $\mathbf{G}$ over the nodes labeled $1 \ldots n$, where the parents of node i are $pa(i)$. We note that the graph G need not be a DAG. In recent years, several approximate methods that exploit this structured representation have been developed (Saria et al., 2007; Cohn et al., 2010; El-Hay et al., 2010; Celikkaya et al., 2011; Rao and Teh, 2011b; Opper and Sanguinetti, 2007).

3 STRUCTURED PROPORTIONAL JUMP PROCESSES

Consider a system of interacting components with two additional characteristics: (1) The dynamics of each component depends on a set of background variables $\mathbf{x} \in R^P$; and (2) Transition rates are non-homogeneous. This work deals with modeling and learning the interactions between the components as well as the relation between the background variable $\mathbf{x}$ and the dynamics of the system represented by $\mathbf{Y}(t)$. As in regression and conditional models, the distribution of the background covariates $\mathbf{x}$ will not be modeled .

Assuming Markovian dynamics, such systems are characterized by a conditional rate function $q(t; \mathbf{y}, \mathbf{y}'|\mathbf{x})$. To model this rate in a compact manner we first assume that, as in CTBNs, $\mathbf{Y}$ has local dynamics, namely is governed by conditional rate functions for all $\mathbf{y} \neq \mathbf{y}'$:

$$q(t; \mathbf{y}, \mathbf{y}'|\mathbf{x}) \equiv q^i(t; y_i, y_i'|\mathbf{y}_{pa(i)}, \mathbf{x}) \cdot 1_{\{j: y_j \neq y_j'\} = \{i\}} .$$

Next, we need to capture the dependency of these dynamics on both time and covariates. To do this, we decompose the rate into two elements: the dependence on time and the joint dependence on context and background variables.

The effect of the time on the transition is captured by the notion of the *baseline rate*. For each component i and pair of states y_i to y_i', we denote by $r^i_{y_i, y_i'}(t)$ the non-negative time dependent functions. The effect of the joint state of the covariates $\mathbf{x}$ and the context $\mathbf{y}_{pa(i)}$ is mediated through a set of weight vectors $\mathbf{w}^i_{y_i, y_i'} \in R^N$. Combining these elements, we define the conditional transition of the SCUP model:

$$q^i(t; y_i, y_i'|\mathbf{y}_{pa(i)}, \mathbf{x}) \equiv r^i_{y_i, y_i'}(t) \cdot \exp\{\mathbf{w}^i_{y_i, y_i'} \cdot \phi^i(\mathbf{x}, \mathbf{y}_{pa(i)})\}, \tag{2}$$

where $\phi^i(\mathbf{x}, \mathbf{y}_{pa(i)})$ is a mapping of the covariates and parent states into an N-dimensional feature vector (where in general N could depend on i). This representation does not explicitly specify the dependency structure of the components on $\mathbf{x}$ because it does not have a significant effect on the inference computational complexity, as shown in Section 4. Note that the time-dependent effect is common to the entire population, meaning that it does not depend on the covariates $\mathbf{x}$ and $\mathbf{y}_{pa(i)}$. On the other hand, the covariates, as well as the parent components, modulate the transition rate between states y_i and y_i' through the second element, independently of time.

To gain some insight into the assumptions encoded in this model, we consider three examples. First, we note that by setting the baseline rates to a constant value, removing the background variables, and setting ϕ^i to be a vector of indicators of the parents' joint state, we obtain a CTBN.

The second example is the *Cox proportional hazard model* (Cox, 1972). This model has a single binary outcome Y where $Y = 1$ represents a base state and $Y = 0$ represents a terminal *failure* state such as death. The rates in such a system are:

$$q(t; 1, 0|\mathbf{x}) \equiv r_0(t) e^{\mathbf{w} \cdot \mathbf{x}} \quad \text{and} \quad q(t; 0, 1|\mathbf{x}) \equiv 0 \ , \tag{3}$$

where $r_0(t)$ is the baseline rate. In this model $q(t; 1, 0|\mathbf{x})$ is called the *hazard function* and the probability of surviving for a time greater than t is

$$\Pr(Y(t) = 1|\mathbf{x}, Y(0) = 1) = e^{-\int_0^t q(s;1,0|\mathbf{x}) ds} .$$

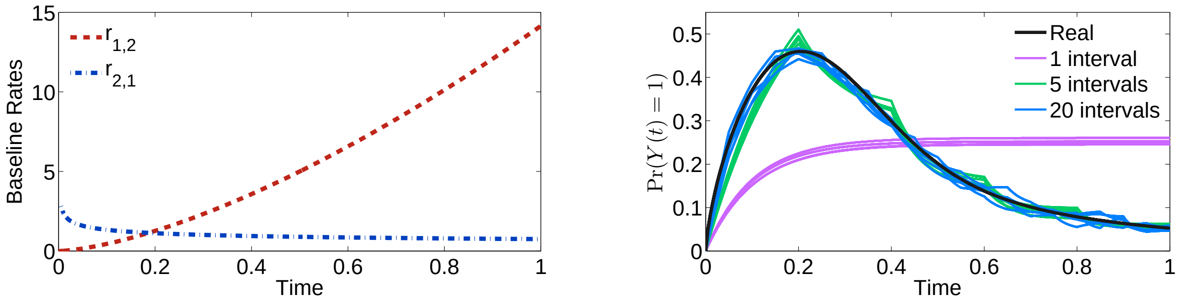
In case the baseline is constant, the survival time distribution is exponential. A monomial baseline, $r_0(t) = \lambda k (\lambda t)^{k-1}$, gives a Weibull distribution. A common approach is to model the baseline in a non-parametric manner (see the seminal work of (Kaplan and Meier, 1958)).

The Cox model encapsulates an assumption that the failure rate proportion for two individuals with attributes $\mathbf{x}_1$ and $\mathbf{x}_2$ is time invariant as $q(t; 1, 0|\mathbf{x}_1)/q(t; 1, 0|\mathbf{x}_2) = e^{\mathbf{w} \cdot (\mathbf{x}_1 - \mathbf{x}_2)}$. This approach is generalized in *multi-state-models* (Putter et al., 2007), which involve a single component and define $q(y, y'|\mathbf{x}; t) = r_{y,y'}(t) e^{\mathbf{w}_{y,y'} \cdot \mathbf{x}}$, resulting in a proportion of $e^{\mathbf{w}_{y,y'} \cdot (\mathbf{x}_1 - \mathbf{x}_2)}$.

The proportionality assumption in SCUP is conditional, that is, if we fix $\mathbf{Y}_{pa(i)}(t) = \mathbf{y}_{pa(i)}$ the proportion between conditional rates is $\exp\{\mathbf{w}^i_{y_i, y_i'} \cdot (\phi^i(\mathbf{x}_1, \mathbf{y}_{pa(i)}) - \phi^i(\mathbf{x}_2, \mathbf{y}_{pa(i)}))\}$ for all t. However, the proportion of the actual marginal rate of moving from y_i to y_i' is time dependent because it is marginalized with time dependent weights, $\Pr(\mathbf{Y}_{pa(i)}(t)|\mathbf{x})$. A time invariance property also holds for proportions between transition rates that are conditioned on two different parent assignments for a fixed $\mathbf{x}$.

The third example deals with an HIV patient model, as shown in Figure 1. The proposed model contains components corresponding to the viral load (VL), CD4 concentration, the status of a certain disease of interest, and an absorbing survival component. The model topology encodes the assumption that the VL and CD4 components affect each other directly, whereas the effect of VL on survival is mediated through CD4 and the disease.

As a concrete example, the disease state space can be {"none', "mild", "severe"}, with the possible transitions "none"↔"mild", and "mild"↔"severe", and the CD4 state space can be {"high", "low"}. The CD4→disease arc encodes a parameter for each combination of the CD4 level and one of the four disease transitions. Notably, the



(a) True baseline rates

(b) Approximation of predictive probability

Figure 2: A Piecewise linear approximation for a non-homogeneous process.

ratio between the transition rates given CD4="high" and given CD4="low" is time independent, and determined solely according to the mapping $\phi^{disease}$(CD4) and the parameters $w^{disease}$ for each transition.

3.1 REPRESENTATION OF BASELINE RATES

Time dependent baseline rates can either be represented non-parametrically as in the classical Cox model, or assume a parametric representation. Examples include Weibull hazard function $r(t) = \lambda k(\lambda t)^{k-1}$, log-logistic hazard, $r(t) = \frac{\lambda k t^{k-1}}{1+\lambda t^k}$ and more. In this work we will adopt a piecewise constant representation, which can approximate well behaved processes with a high degree of accuracy.

To characterize such processes, we consider single-component models with a time-dependent state $Y(t)$ (every model can be represented as a single-component model whose state space is the Cartesian product of the components state spaces). Denote by $\mathbf{P}^{\mathbf{Q}}(s,t)$ the transition matrix whose y, y' entry is $\Pr(Y(t) = y \,|\, Y(s) = y')$, and by $\mu_y^{\mathbf{Q}}(t) \equiv \Pr^{\mathbf{Q}}(Y(t) = y)$ the time-dependent marginal probability. We say that a matrix $\mathbf{P}$ is embeddable if there exists a matrix $\mathbf{A}$ such that $\mathbf{P} = e^{\mathbf{A}}$. Let $\tau_0 < \tau_1 < \ldots < \tau_K$ be an ordered set of time points, and suppose that $\mathbf{P}_{\mathbf{Q}}(\tau_{k-1}, \tau_k)$ is embeddable for every $k = 1, \ldots, K$. From the Markov property, it follows that there exists a piecewise constant rate matrix function $\hat{\mathbf{Q}}(t) = \mathbf{Q}_k, \forall \tau_{k-1} \le t < \tau_k$ that satisfies $\mu_y^{\mathbf{Q}}(\tau_k) = \mu_y^{\hat{\mathbf{Q}}}(\tau_k)$. Moreover, the following lemma bounds the error for rate matrices with bounded transition rates:

Lemma 3.1 :*Let $Y(t)$ be a non-homogeneous process with bounded transition rates $\mathbf{Q}(t)$ and an embeddable rate matrix. Denote $\rho_k = max_{y,\tau_{k-1} \le t < \tau_k} |q_{y,y}(t)|$, $\hat{\rho}_k = max_y |\hat{q}_{y,y}|$. Then, for all y and $\tau_{k-1} \le t \le \tau_k$, $|\mu_y^{\mathbf{Q}}(t) - \mu_y^{\hat{\mathbf{Q}}}(t)| < (\rho + \hat{\rho}) \cdot (\tau_k - \tau_{k-1})$.*

This lemma, proven in the appendix, suggests that the number of intervals required to bound the bias by ϵ scales inversely linear with $1/\epsilon$. We note that tight approximations exist in the case of non-embeddable processes (Davies, 2010).

As an example, consider a two state model with time dependent baseline rates depicted in Figure 2a. This process induces a non-monotone marginal probability $\mu_1^{\mathbf{Q}}(t)$ given an initial condition $Y(0) = 2$, as shown by the smooth black line in Figure 2b. The initial rise follows from the relation $r_{2,1}(t) > r_{1,2}(t)$, and the subsequent decline from the opposite relation. The colored lines show estimated probabilities given by piecewise constant models with 1, 5 and 20 intervals of constant rates that were trained on 1000 simulated trajectories.

4 LEARNING

Generally, training data may include a mixture of point observations on some components and full (complete) trajectories of others. For example CD4 and viral load are point observations measured periodically, whereas time of death is usually exactly recorded resulting in a continuous observation on survival. To learn from such data, we will adapt the approach taken for CTBNs, which handles unseen state trajectories between observations as missing data and employs Expectation Maximization (EM). The first step is to derive the likelihood of the model given complete trajectories.

4.1 LIKELIHOOD FUNCTION

A fully observed trajectory is represented using the sequence $t_0, \ldots, t_M$ and states $y_0, \ldots, y_{M-1}$ such that $Y(t) = y_k$ for $t \in [t_k, t_{k+1})$. We denote such a trajectory by $y_{[0,t_M]}$. The likelihood of a non-homogeneous process with a set of rates $\mathcal{M} = \{q(t; y, y')\}_{y,y'}$ is

$$l(\mathcal{M}|y_{[0,t_M]}) = \exp\left\{\int_{t_{M-1}}^{t_M} q(t; y_{M-1}, y_{M-1})dt\right\} \quad (4)$$
$$\cdot \prod_{k=0}^{M-2}\left[\exp\left\{\int_{t_k}^{t_{k+1}} q(t; y_k, y_k)dt\right\} q(t_{k+1}; y_k, y_{k+1})\right]$$

where $q(t; \mathbf{y}, \mathbf{y}) = -\sum_{y' \neq y} q(t; y, y')$.

Let $\mathcal{D}$ be a data set that includes pairs of trajectories $y^{[c]}$ and covariates $\mathbf{x}_c$, where $c = 1, \ldots, N_{\text{sequences}}$ and denote by $ll(\mathcal{M}|\mathcal{D})$ be the log-likelihood.

The log-likelihood is unbounded if there are no constraints on the baseline rate functions. Consider for example the survival model described in Equation 3, and suppose that no background variable is involved. In this case,

$$ll(r_0(t)|\mathcal{D}) = \sum_c \left[-\int_0^{t_c} r_0(t)dt + \log r_0(t_c)\right] .$$

One can construct a series of baseline rates such that this term approaches infinity as $r_0(t) \quad \sum_c a_c\delta(t - t_c)$, implying that a naive maximum likelihood procedure tends to overfit $r_0(t)$ to a function that imposes transitions at the observed times if no constraints are put in place. An alternative approach for non-parametric estimation of a possibly arbitrary baseline is to use partial-likelihood (Cox, 1972, 1975). However, this direction does not generalize naturally to partially observed data. Two possible approaches for placing constraints use either a restricted parametric form or regularized baseline.

4.2 PARTIALLY OBSERVED DATA

To deal with partially observed data, we perform an EM procedure. On each iteration we compute the expected log-likelihood of a new model with respect to the posterior distribution of the current model $\mathcal{M}_0$. The posterior distribution of a Markov process $\mathcal{M}_0$ given a sequence σ_c is characterized by a set of time-dependent functions (Cohn et al., 2010)

$$\begin{aligned}\mu_y(t|c) &= \Pr(Y(t) = y|c, \mathcal{M}_0)\\ \gamma_{y,y'}(t|c) &= \lim_{\Delta t \to 0} \frac{\Pr(Y(t) = y, Y(t + \Delta t) = y'|c, \mathcal{M}_0)}{\Delta t} .\end{aligned}$$

$\mu_y(t|c)$ is the singleton probability that the process is in state y at time t. $\gamma_{y,y'}(t|c)$ is the *intensity* of the pairwise probability of being in state y and then moving to y' at time t.

Using this characterization, taking the expectation on the log of the likelihood function in Equation 4 and plugging in the decomposition of the conditional intensities depicted in Equation 2, gives the expected log-likelihood of a multi-component model:

$$E_{\mathcal{M}_0}[ll(\mathcal{M}, \mathcal{D})] = \sum_c \sum_{y_i, \mathbf{y}_{pa(i)}} \sum_{y'} \quad (5)$$
$$\Bigg[-\exp\{\mathbf{w}^i_{y_i,y'_i} \cdot \phi^i(\mathbf{x}_c, \mathbf{y}_{pa(i)})\} \int_t \mu_{y_i,\mathbf{y}_{pa(i)}} r^i_{y_i,y'_i} dt$$
$$+ \int_t \gamma_{y_i,y'_i|\mathbf{y}_{pa(i)}} \left(\log r^i_{y_i,y'_i} + \mathbf{w}^i_{y_i,y'_i} \cdot \phi^i(\mathbf{x}_c, \mathbf{y}_{pa(i)})\right) dt\Bigg]$$

where we omit t and c from μ, γ and r, $\mu_{y_i,\mathbf{y}_{pa(i)}}$ is the marginalization of the posterior distribution to the subset of components $i, pa(i)$, and similarly $\gamma_{y_i,y'_i|\mathbf{y}_{pa(i)}} = \sum_{\{\hat{\mathbf{y}}|\hat{y}_i = y_i \hat{\mathbf{y}}_{pa(i)} = \mathbf{y}_{pa(i)}\}} \gamma_{\hat{\mathbf{y}},[\hat{\mathbf{y}}_{\setminus i}, y'_i]}$ is a marginalization of pairwise probability intensities. Additional details are given in the Appendix. An exact computation of these functions and their integrals is feasible for systems with a small number of components. Otherwise, a variety of approximate methods are available (Saria et al., 2007; Cohn et al., 2010; El-Hay et al., 2010; Celikkaya et al., 2011; Rao and Teh, 2011b; Opper and Sanguinetti, 2007).

4.3 OPTIMIZATION

The gradient of the log-likelihood with respect to $\mathbf{w}$ is:

$$\frac{\partial E_{\mathcal{M}_0}[ll(\mathcal{M}, \mathcal{D})]}{\partial \mathbf{w}^i_{y_i,y'_i}} =$$
$$\sum_c \sum_{\mathbf{y}_{pa(i)}} \phi^i(\mathbf{x}_c, \mathbf{y}_{pa(i)})(M^c_{\mathbf{y}_{pa(i)}} - MP^c_{\mathbf{y}_{pa(i)}})$$

where the first term $M^c_{\mathbf{y}_{pa(i)}} = \int_t \gamma_{y_i,y'_i|\mathbf{y}_{pa(i)}} dt$ is the expected number of transitions from y_i to y'_i given the state of the parents $\mathbf{y}_{pa(i)}$, M_0 and the evidence in sequence c (see (Cohn et al., 2010)). The second term

$$MP^c_{\mathbf{y}_{pa(i)}} = \exp\{\mathbf{w}^i_{y_i,y'_i} \cdot \phi^i(\mathbf{x}_c, \mathbf{y}_{pa(i)})\} \int_t \mu^i_{y_i,\mathbf{y}_{pa(i)}} r^i_{y_i,y'_i} dt$$

is the integral of the probability of being in state $[y_i, \mathbf{y}_{pa(i)}]$, multiplied by the transition rate. Hence, this term can be interpreted as the expected number of *potential transitions*. The gradient weighs the feature vectors $\phi^i(\mathbf{x}_c, \mathbf{y}_{pa(i)})$ using the difference between the expected number of actual and potential transitions.

Optimization of the baseline that assumes a parametric form $r^i_{y_i,y'_i}(t) = r^i_{y_i,y'_i}(t; \theta)$ involves computation of its gradient with respect to the parameters θ

$$\frac{\partial E_{\mathcal{M}_0}[ll(\mathcal{M}, \mathcal{D})]}{\partial \theta} =$$
$$\sum_c \sum_{\mathbf{y}_{pa(i)}} \int_t \Bigg[-\exp\{\mathbf{w}^i_{y_i,y'_i} \cdot \phi^i(\mathbf{x}, \mathbf{y}_{pa(i)})\}\mu_{y_i,\mathbf{y}_{pa(i)}}$$
$$+ \frac{\gamma_{y_i,y'_i|\mathbf{y}_{pa(i)}}}{r^i_{y_i,y'_i}}\Bigg] \frac{\partial r^i_{y_i,y'_i}}{\partial \theta} dt \quad .$$

In the simplest case, if the baseline is constant or piecewise constant, the stationary point solution has a closed from.

A maximum likelihood estimator can be found using an EM procedure iterating between expectation and maximization steps. Expectation steps compute the functions that represent the posterior distribution, μ and γ. Maximization steps involve optimizing the covariate and cross component influence weights $\mathbf{w}^{i}_{y_i,y'_i}$ using standard optimization methods and the gradient derived in Equation 6, as well as optimizing the baseline rates using the gradient in Equation 6. While the overall target function is not convex, the optimization of $\mathbf{w}^{i}_{y_i,y'_i}$ is a convex problem given fixed baselines and posterior distributions, and so is the case for many choices of the baseline rates .

5 EXPERIMENTAL RESULTS

5.1 LEARNING EVALUATION

Our initial experiments test the validity of SCUP. To this end, we created synthetic SCUP data sets. We then trained SCUP using these data sets, and compared the similarity of the learned models with the actual ones.

The topology for all data generating models was similar to the HIV disease topology (Figure 1) with the exclusion of the survival state. All models included a single randomly drawn binary covariate. The baseline rates followed a Weibull rate, with a shape parameter $\kappa = 2$ and a scale parameter drawn from an inverse Gamma distribution (the Weibull distribution conjugate prior), with shape and scale parameters both equal to 2. For each component with parents y_1, y_2, and a covariate x, we used the feature mapping $\phi(x, y_1, y_2) = (x, \mathbf{1_{y_1=2}}, \mathbf{1_{y_2=2}})$, with feature coefficients drawn from $\mathcal{N}(0,1)$.

We evaluated learning performance as a function of dataset size and sampling rate. During the training, we divided the time interval [0,1] into 5 equally sized intervals, and learned piecewise-constant baseline rates for each one. We considered both fully observed data and point observations, with observation times for each trajectory drawn uniformly from [0,1]. All trajectories were observed at times $t = 0$ and $t = 1$. Our evaluation compared the similarity of the learned models to the true generating models through the root mean square error (RMSE) of the learned coefficients. We also compared the integral of each baseline rate across the time interval [0,1], to its true value. The baseline integral was used because it does not depend on parametric form, and because it is used in inference and learning tasks, rather than the baseline itself. Figure 3 shows that learning accuracy increases with sample size and sampling rate, as expected. As a further measure of validity, we verified that $\log(\text{RMSE})$ decreases linearly with log dataset size, with slopes close to -0.5, indicating consistency (data not shown).

5.2 THE EFFECT OF NON-HOMOGENEITY

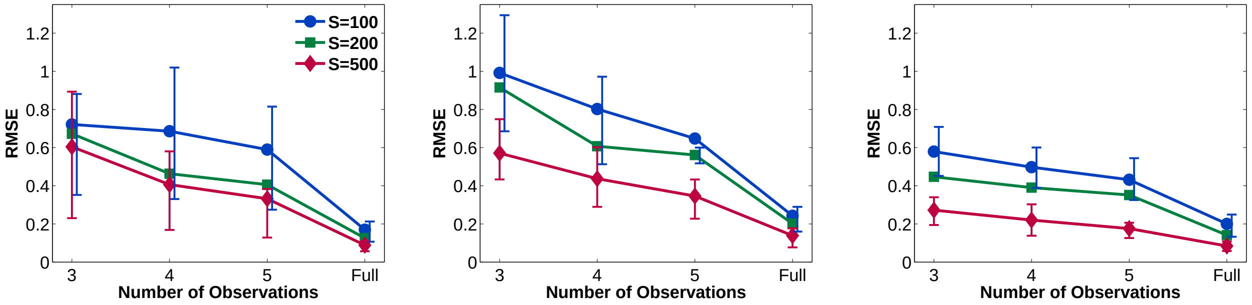
To test the effect of non-homogeneity, we generated data from homogeneous and non-homogeneous SCUP models. We then trained the models with different levels of non-homogeneity on the generated datasets, and evaluated learning performance. The datasets were generated from two SCUP architectures similar to those described in the previous section, with the exception that the first architecture used a homogeneous constant baseline rate for all transitions, whereas the second one used baseline rates as previously described. The baseline rates for the first architecture were generated from a Gamma distribution, with scale and shape parameters equal to 1.0. We generated five models from each architecture, and generated a dataset of 500 trajectories using each model. Every trajectory was observed at times $t = 0$ and $t = 1$, and at three other uniformly drawn time-points. We trained SCUP models with increasing numbers of piecewise-constant baseline rate. Notably, models with one baseline rate are equivalent to CTBNs. We evaluated the learning performance via a five-fold cross validation of out of sample (OOS) likelihood.

The results, shown in Figure 4, demonstrate that homogeneous models cannot capture complex dynamics that change over time. Increasing the number of baseline rates leads to greater flexibility on the one hand, but to the risk of overfitting on the other.

5.3 COMPARISON WITH OTHER METHODS

To assess the relative performance of SCUP, we compared it to two competing methods, which can both be derived as special instances of SCUP: A factored model and a multi-state model. The factored model (FM) is a SCUP model with several independent components. There are no arcs between components, and thus transition probabilities are affected only by covariates and baseline rates. The multi-state model (MSTM) follows the implementation of a package called MSM (Jackson, 2011). It can be viewed as a SCUP model with a single component, whose state space is the Cartesian product of the SCUP components state spaces. We verified empirically that our implementation yields the same results as MSM on a wide variety of scenarios. SCUP can be seen as an intermediate method between these two extremes, balancing between compactness and expressiveness. Notably, all three methods fully support non-homogeneous dynamics.

We generated five models for each of the three architectures, each having a single binary covariate, with Weibull baseline rates and randomly drawn coefficients, as described in the previous section. The SCUP models were generated and used as described in the previous sections. The FM models contained three binary components, and the MSM models contained a single eight-state component,



(a) Baseline rates

(b) Parent coefficients

(c) Covariate coefficients

Figure 3: Root mean square error of estimated parameters for various sampling rates, and the 75% confidence intervals (confidence intervals for S=200 are omitted for clarity).

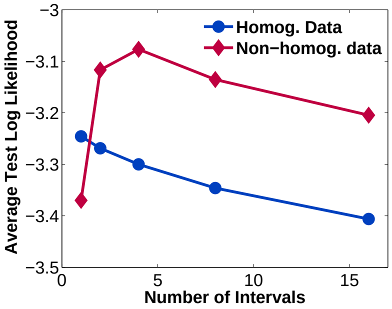


Figure 4: Out of sample likelihood for models trained with increasing number of piecewise-constant baseline rates.

Table 1: The number of parameters learned by FM, SCUP, and MSM. The number of piecewise-constant baseline rates is denoted by t.

	FM	SCUP	MSM
Cov. coefficients	6	6	56
Parents coefficients	0	12	0
Baseline rates	$6t$	$6t$	$56t$
Total number	$6{+}6t$	$18{+}6t$	$56{+}56t$

with one state corresponding to each assignment of the components' states in the SCUP model. Both the MSM and FM models used the feature mapping $\phi(x) = x$.

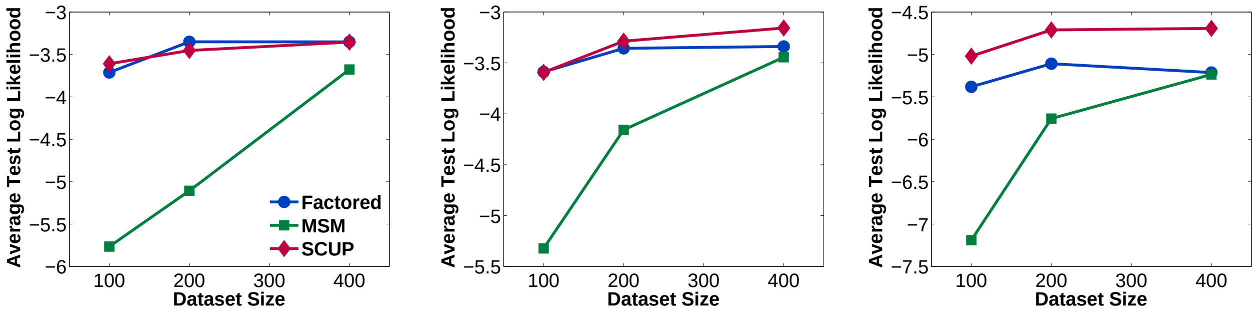
We generated datasets of 1,000 trajectories using each of the 15 models. We then examined how well a model from each architecture can be trained on each dataset, via a three-fold cross validation of OOS likelihood. The trajectories were observed as described in the previous section. All trained models used five piecewise constant baseline rates. The number of parameters for the three models is shown in Table 1, demonstrating that SCUP bridges between the two extremes.

Figure 5 demonstrates that SCUP is more flexible than the other two methods, allowing it to represent data generated by different architectures, while retaining compactness. MSM exhibits poor learning capabilities for smaller datasets; this holds true even for data created by a model with the same architecture, demonstrating overfitting due to model complexity. The factored model does not suffer from overfitting, but has limited expressiveness, and thus cannot capture mutual influences between components.

5.4 ANALYSIS HIV DATA

We evaluated the performance of SCUP by analyzing real data from a data set containing lab measurements of HIV patients who took medication on a regular basis, previously described in (Rosen-Zvi et al., 2008). We defined models with two components corresponding to the two main measures of HIV severity, viral load (VL) and CD4 lymphocytes concentration, as well as a continuously observed binary absorbing component, representing survival. The resulting model is similar to the one described in Figure 1, with the omission of the disease component, and the addition of a VL→survival arc, which was added to obtain a fully connected topology. Following previous works, the CD4 level was dichotomized to have 2 states, using a threshold of 200 (D'Amico et al., 2011). The VL level was also dichotomized to have 2 states, using a threshold of 500, as previously done in analyses of this data (Rosen-Zvi et al., 2008).

For the analysis, we randomly selected 2000 patients whose VL and CD4 levels were both observed at each observation point. The resulting dataset contained 5.14 observations per patient on average (standard deviation 3.37). For every patient, we included covariates corresponding to age, sex, and whether the patient had undertaken a different therapy in the past. Feature mappings consisted of a concatenation of the covariates, with a binary 0/1 feature for each parent component. The initial time t=0 was set as the therapy start time. For patients who underwent several successive therapies, only observations taken during the period of the first one were included in the analysis. All patients had



(a) FM generated data

(b) SCUP generated data

(c) MSM generated data

Figure 5: Test likelihoods of data learned by different models.

an observation at time t=0, using the closest measurement within a month from the therapy start date.

We computed the average OOS log likelihood obtained via a five-fold cross validation, with increasing numbers of piecewise constant baseline rates. The results, shown in Figure 6, clearly demonstrate the powerful effect of non-homogeneity, and the importance of modeling it correctly. MSM has an advantage when using a small number of baseline rates, owing to its richer model, which can capture richer interaction patterns between the system components. However, SCUP steadily improves as the number of baseline rates increases, until it eventually surpasses MSM. This increase indicates the presence of strong non-homogeneous dynamics. MSM can also capture non-homogeneous dynamics, but is hindered by its large number of parameters. The FM model exhibits weaker performance than the other methods for every number of baseline rates tested. This is due to the fact that it cannot capture the dynamics stemming from mutual influences between the system components. The decrease in OOS likelihood for FM when using 16 baseline rates may stem from overfitting, which occurs because it is trying to incorrectly capture mutual influences between the system components via baseline rates.

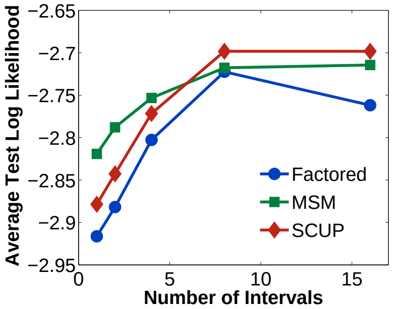


Figure 6: Performance of SCUP, FM, and MSM on the HIV dataset.

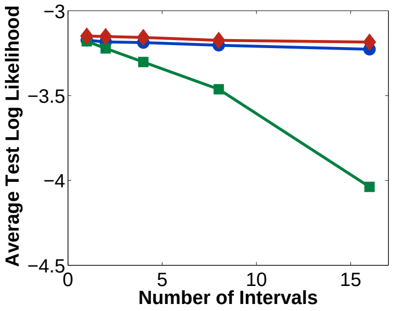


Figure 7: Performance of SCUP, FM, and MSM on the diabetes dataset.

5.5 ANALYSIS OF DATA FROM DIABETES PATIENTS

We evaluated SCUP on a large cohort of diabetes patients, previously described by (Neuvirth et al., 2011). Following (Neuvirth et al., 2011), we define the main outcome of interest as the glycated hemoglobin (HbA1c) blood test, which is a reliable indicator of diabetes severity status. A higher HbA1c indicates increased risk of developing complications.

Our goal was to learn the interaction patterns between the HbA1c level and other potential diabetes biomarkers commonly measured in routine blood tests. The ability to predict HbA1c levels from routine blood tests can improve early detection of the disease progression. To this end, we defined a SCUP model with binary components for HbA1c, low-density lipoprotein (LDL), and triglycerides levels. The two states of each component correspond to normal and abnormal clinical status, with the thresholds for HbA1c, LDL, and triglycerides set to 7, 130 and 200, respectively. We used a fully connected topology, and included the age and sex of each patient as covariates.

For the analysis, we randomly chose 1,000 patients with non-missing values for the components of interest at every observation point. Every patient had 3.25 observations on average (standard deviation 1.52). Feature mappings consisted of a concatenation of the covariates, with a binary 0/1 feature for each parent component. The time $t = 0$ for each patient was determined according to the first observation time.

We computed the average OOS likelihood obtained via a five-fold cross validation, with increasing numbers of intervals. The results, shown in Figure 7, demonstrate that SCUP can scale to rich models without overfitting. The factored model, although scalable, does not capture the interactions between components, leading to weaker prediction power. The MSM model tends to suffer from overfitting due to its complexity. The lack of increase in OOS likelihood for increased numbers of intervals indicates that the components tend to follow homogeneous dynamics in this dataset. Nevertheless, SCUP does not overfit when trained with a large number of intervals, indicating its robustness to the type of underlying dynamics in the data.

To further investigate the different methods, we examined the coefficients describing mutual influence between the system components; these were learned across the different folds. We examined the models that assumed one piecewise-constant interval, as they had the best fit for this data. For every pair of components, we computed the coefficient describing the influence of one on a transition of the other. For MSM, we averaged the two corresponding coefficients over the two possible states of the third component. The results, shown in Table 2, demonstrate that SCUP models learned across the different folds are more consistent with each other, leading to substantially smaller variance.

The results demonstrate rich interaction patterns across the components. For example, increased triglycerides levels are associated with an increase in HbA1C, whereas increased HbA1C is associated with stabilization of the triglycerides levels via a reduction of their transition rate. Such observations cannot be performed directly in FM nor MSM, due to their lack of modular structure.

6 DISCUSSION

We proposed a proportional modeling scheme for non-homogeneous multi-component processes, by combing factorizations of CTBNs with a decomposition dating back to proportional hazard models. The key modeling assumption is a decomposition of the process into a time-dependent non-homogeneous component that does not depend on the model topology, and a time-independent component that depends on the model topology and additional features. This is a natural extension of classic hazard models, which can be considered as special SCUP instances with no underlying topology. This decomposition leads to compact models that can capture complex dynamics, as well as an efficient learning scheme, and easily interpretable results.

Table 2: The average coefficients of parent influence on increase (↑) and decrease (↓) learned in the diabetes dataset, and the minimum and maximum values obtained across the five folds.

	SCUP	MSM
LDL$\rightarrow$A1C$\uparrow$	.17 (.08, .25)	-.24 (-.73, .27)
Trig.$\rightarrow$A1C$\uparrow$	.31 (.09, .45)	.12 (-.65, .62)
LDL$\rightarrow$A1C$\downarrow$	.09 (.04, .16)	-.20 (-.50, .06)
Trig.$\rightarrow$A1C$\downarrow$	-.17 (-.23, .02)	-.22 (-.42, .22)
A1C$\rightarrow$LDL$\uparrow$	.34 (.20, .45)	1.15 (.86, 1.31)
Trig.$\rightarrow$LDL$\uparrow$	.57 (.33, .79)	.68 (.31, 1.14)
A1C$\rightarrow$LDL$\downarrow$	-.04 (-.23, .14)	-.18 (-.98, .33)
Trig.$\rightarrow$LDL$\downarrow$	-.38 (-.49, -.25)	.67 (.16, 1.46)
A1C$\rightarrow$Trig.$\uparrow$	-.28 (-.49, -.09)	-.51 (-.90, -.37)
LDL$\rightarrow$Trig.$\uparrow$	.82 (.67, .92)	-.54 (-1.13, .25)
A1C$\rightarrow$Trig.$\downarrow$	-.59 (-.71, -.50)	-1.24 (-1.82, -.89)
LDL$\rightarrow$Trig.$\downarrow$	.63 (.40, .82)	-.72 (-1.79, .09)

Our theoretical and empirical results demonstrate that non-homogeneous dynamics can be captured accurately using a piecewise homogeneous approximation. It would be interesting to compare this baseline rates representation to parametric forms. Learning such models is straightforward and can be performed by plugging in the partial derivative of a specific parametric form to the gradient in Equation 6.

Baseline rates can be regularized via spline approximations (Commenges, 2002; Joly et al., 2009; Farewell and Tom, 2012) or Gaussian process priors (Rao and Teh, 2011a). Splines can also be naturally adapted to regularize piecewise constant rates. This can be done by bounding the difference between rates in adjacent time intervals, or the rate of change of this difference, which is analogous to bounding the first and second derivative, respectively. Regularization of other model parameters, such as the covariate or parents coefficients, can potentially be handled using standard regularization methods such as elastic nets, as recently proposed for Cox regression (Simon et al., 2011).

In this work we studied moderately sized systems. Adapting approximate inference methods developed for CTBNs that support non-homogeneity, such as (Rao and Teh, 2011b) or (El-Hay et al., 2010), could scale up this framework to arbitrarily large systems.

Acknowledgments

We would like to thank our anonymous reviewers for useful comments, Hani Neuvirth-Telem for helping with processing of the diabetes data, the EuResist Network study group for providing the HIV data, Gal Elidan for insightful suggestions on earlier versions of the manuscript, and Amir Globerson for helpful discussions.

References

E. B. Celikkaya, C. R. Shelton, and W. Lam. Factored filtering of continuous-time systems. In *UAI*, 2011.

I. Cohn, T. El-Hay, N. Friedman, and R. Kupferman. Mean field variational approximation for continuous-time Bayesian networks. *Journal of Machine Learning Research*, 11:2745 2783, 2010.

D. Commenges. Inference for multi-state models from interval-censored data. *Stat Methods Med Res*, 11(2): 167–182, Apr 2002.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

G. D'Amico, G. Di Biase, J. Janssen, and R. Manca. HIV evolution: a quantification of the effects due to age and to medical progress. *Informatica*, 22(1):27–42, 2011.

E. B. Davies. Embeddable Markov matrices. *Electronic Journal of Probability*, 15:1474–1486, 2010.

N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha. Scalable influence estimation in continuous-time diffusion networks. In *NIPS*, pages 3147–3155, 2013.

T. El-Hay, I. Cohn, N. Friedman, and R. Kupferman. Continuous-time belief propagation. In *ICML*, pages 343–350, 2010.

V. T. Farewell and B. D. Tom. The versatility of multi-state models for the analysis of longitudinal data with unobservable features. *Lifetime Data Anal*, Dec 2012.

C. H. Jackson. Multi-state models for panel data: the MSM package for R. *Journal of Statistical Software*, 38(8):1–29, 2011.

P. Joly, C. Durand, C. Helmer, and D. Commenges. Estimating life expectancy of demented and institutionalized subjects from interval-censored observations of a multi-state model. *Stat Model*, 9(4): 345–360, 2009.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

H. C. Looker, S. O. Nyangoma, D. T. Cromie, J. A. Olson, et al. Predicted impact of extending the screening interval for diabetic retinopathy: the Scottish Diabetic Retinopathy Screening programme. *Diabetologia*, May 2013.

H. Neuvirth, M. Ozery-Flato, J. Hu, J. Laserson, et al. Toward personalized care management of patients at risk: the diabetes case study. In *KDD*, pages 395–403. ACM, 2011.

U. Nodelman, C.R. Shelton, and D. Koller. Continuous time Bayesian networks. pages 378–387, 2002.

M. Opper and G. Sanguinetti. Variational inference for Markov jump processes. In *NIPS*, 2007.

H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*, 26(11):2389–2430, May 2007.

V. Rao and Y. W. Teh. Gaussian process modulated renewal processes. In *NIPS*, pages 2474–2482, 2011a.

V. Rao and Y. W. Teh. Fast MCMC sampling for Markov jump processes and continuous time Bayesian networks. In *UAI*, pages 619–626, 2011b.

M. Rosen-Zvi, A. Altmann, M. Prosperi, et al. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. *Bioinformatics*, 24(13):399–406, Jul 2008.

S. Saria, U. Nodelman, and D. Koller. Reasoning at the right time granularity. In *UAI*, pages 326–334, 2007.

N. Simon, J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 3 2011. ISSN 1548-7660.

S. Taghipour, D. Banjevic, A. B. Miller, N. Montgomery, A. K. Jardine, and B. J. Harvey. Parameter estimates for invasive breast cancer progression in the Canadian National Breast Screening Study. *Br. J. Cancer*, 108(3): 542–548, Feb 2013.

A. Walker, S. Doyle, J. Posnett, and M. Hunjan. Cost-effectiveness of single-dose tamsulosin and dutasteride combination therapy compared with tamsulosin monotherapy in patients with benign prostatic hyperplasia in the UK. *BJU Int.*, Jan 2013.