Lifetime Data Anal (2014) 20:459–480
DOI 10.1007/s10985-013-9278-0

Parametric survival densities from phase-type models

Eric V. Slud · Jiraphan Suntornchost

Received: 13 June 2012 / Accepted: 25 July 2013 / Published online: 22 August 2013

Abstract After a brief historical survey of parametric survival models, from actuarial, biomedical, demographical and engineering sources, this paper discusses the persistent reasons why parametric models still play an important role in exploratory statistical research. The phase-type models are advanced as a flexible family of latent-class models with interpretable components. These models are now supported by computational statistical methods that make numerical calculation of likelihoods and statistical estimation of parameters feasible in theory for quite complicated settings. However, consideration of Fisher Information and likelihood-ratio type tests to discriminate between model families indicates that only the simplest phase-type model topologies can be stably estimated in practice, even on rather large datasets. An example of a parametric model with features of mixtures, multiple stages or 'hits', and a trapping-state is given to illustrate simple computational tools in R, both on simulated data and on a large SEER 1992–2002 breast-cancer dataset.

Electronic supplementary material The online version of this article (doi:10.1007/s10985-013-9278-0) contains supplementary material, which is available to authorized users.

Disclaimer This paper describes research of its authors, and is released to inform interested parties and encourage discussion. Results and conclusions are the authors' and have not been endorsed by the Census Bureau.

E. V. Slud
Census Bureau CSRM, Washington, DC, USA
e-mail: Eric.V.Slud@census.gov

E. V. Slud
University of Maryland, College park, MD, USA

J. Suntornchost (✉)
Chulalongkorn University, Bangkok, Thailand
e-mail: Jiraphan.S@chula.ac.th

Keywords EM algorithm · Flowgraph model · Latent class model · Markov chain · Maximum likelihood · Right-censored survival data · Transition intensity

1 Historical introduction

Parametric survival analysis models originated from several directions of applied statistical work: actuarial and demographic summaries of human mortality patterns, biological and epidemiologic summaries of mortality and disease incidence, and engineering failure testing and reliability studies. Although latter-day durational analyses can also be found in economics, sociology, and other fields, our intuitions about mechanisms and hazard rates have generally followed these older sources. Beginning with parametric models designed to reflect specific qualitative features of observed hazard rates, each discipline has added levels of detail to reflect distinct objectives. In recent years, there has been a resurgence of interest in models based on threshold crossings, cumulative damage, and other hitting-times expressed in terms of underlying stochastic processes.

In this article we provide background and an assessment of the state of parametric survival modeling, from the particular vantage point of models which can be fitted to biomedical datasets of moderate to large size. We consider simple conceptual tools, primarily Fisher information and likelihood ratio type tests (Kullback-Leibler distances), for describing relationships between different models and for resolving which parameters can be stably identified from data. We focus on *phase-type models*, the class of models of absorption times by continuous-time discrete state Markov chains, and some variants.

1.1 Sources from actuarial science and demography

Human mortality description in actuarial science begins with life tables (Halley 1693). Patterns of mortality by age were formulated initially through *age-specific death-rates* denoted nowadays by $q_x = P(x < T < x+1 \,|\, T \geq x)$ for one-year time intervals associated with continuous lifetime random variables T. Models of mortality were later formulated more or less interchangeably either through death rates or through the hazard rate function $h(x)$ (termed *force of mortality* by actuaries), so as to facilitate the numerical calculation of expected present values under constant rates of compound interest. These models include the influential *exponential law* $h(x) = bA^x$ of Gompertz (1825), where $b > 0$ and A is slightly greater than 1, to which Makeham later (1860) added a constant term, as well as the Weibull power law $h(x) = b\,k\,x^{k-1}$. (For these and other historical references, see Bowers et al. 1997 and Lin and Liu 2007.) These models are naturally unified (Brillinger 1961) as being among the small class of *extreme-value* distributions characterized by the fundamental Fisher-Tippett-Gnedenko (1927–1948) Theorem (Feller 1972, vol.2) as possible distributional limits of maxima $\max_{1\leq i\leq n} X_i$ of independent identically distributed sequences $\{X_i\}$.

In practice, actuaries and demographers generally rely on nonparametric techniques to do justice to their large volumes of mortality data, using *'Graduation'*—the actuarial term for Whittaker-Henderson smoothing splines and related methods (Bowers et al.

1997)—to fit sequences q_x of age-specific death-rates at integer ages. Yet, especially to characterize the changes in mortality between different population cohorts, actuaries and demographers have continued to propose parametric models, like the Heiligman and Pollard (1980) eight-parameter model. While that model fits important qualitative features of modern life tables, it is known to be messy to fit to real data and its parameters are not readily interpretable, so it is not considered really practical. Another, more tractable parametric approach is given by Lin and Liu (2007), which will be discussed below under the heading of phase-type models.

Demographers, like actuaries, generally prefer large-population mortality datasets to be reflected in highly parameterized models, verging on the nonparametric, which can be used to summarize and forecast secular variations of death-rates over time and birth-cohort. The most famous model in this area is that of Lee and Carter (1992), which expresses the logarithm of the age-x specific death rate $q_{x,t}$ at calendar time t as the sum of an underlying age-effect α_x and a bilinear form $\beta_x\,\gamma_t$. The model thus includes a period-effect and a period-specific rate of change in the main age-effect. It is widely used as a benchmark for analysis, generalizing the older linear t-projections of x-mortality, although often the time t rate of change γ_t in age-x slopes are found to be roughly linear. Later developments of this model have primarily focused on specialized methods of fitting it from data. Extensions and variants of the model itself can be found in Bongaarts (2005); Koissi et al. (2006), and Booth and Tickle (2008). Suntornchost et al. (2011) have recently found that the model works better on several decades of US cause-specific mortality data when the time t rates of change are parameterized separately for several distinct age-intervals.

1.2 Biomedical and epidemiologic sources

The very fruitful *multihit model* of cancer incidence was formulated by Armitage and Doll (1954) based on their observation that cancer incidence for many different sites and populations approximately follows a power law as a function of age. The multihit model essentially says that before a malignant tumor becomes clinically observable, its precursor cells must have passed successively through a series of independent stages, conceptualized as mutations or newly initiated developmental events. The key contribution of this model was a mechanism 'explaining' the observed power law: when the k successive transition rates λ are identical, the power dependence on age is the term x^{k-1} in the Gamma (k, λ) density for the sum of k Expon (λ) waiting times.

This model already displays the key features that later characterize phase-type models for mortality: independent, latent stages with exponentially distributed durations. For example, Knudson (1971) advanced a Markovian model (with 6 states and 7 transition-rate parameters) for retinoblastoma development which was later substantially validated. See Moolgavkar (2004) for references and background on the 50 years of further development of the original multi-hit idea, which showed satisfying agreement between conceptualized latent stages and the mutations discovered through molecular genetics. Moolgavkar (2004) claims that multistage cancer models of causation have now become explanatory, supported by genetic and other biological

evidence, but that more accurate descriptive transition models must still be developed. Other Markov-chain models of cancer incidence times or death times following diagnosis and initial treatment have been introduced by many different authors for several different cancers, such as the Manton and Stallard (1980) model of breast-cancer mortality. (See additional references of Moolgavkar (2004) and Manton and Stallard (1980) for other examples.)

1.3 Sources from reliability

There is a long tradition in engineering reliability to view failure as the result of physical but nondeterministic cumulative-damage processes such as crack-spreading or corrosion. Accordingly, many reliability models have been developed (Singpurwalla 1996) in which an underlying unobservable stochastic process ('degradation' or 'damage') $X(t)$ determines the failure time random variable T for a device as the first time when $X(t)$ crosses threshold a, which may itself be a random characteristic of the device. Specific classes of stochastic processes X crossing a constant threshold determine well known failure time distributions. The best known example, when X is the Wiener process with drift, is the 2-parameter *Inverse Gaussian* distribution. Lee and Whitmore (2006) discuss several other such models, in some of which a process $Y(t)$ correlated with the underlying process $X(t)$ can be observed: their approach to survival data analysis is to choose a tractable process X and model survival times through regression models for the threshold a or initial point $x_0 = X(0)$ in terms of observable covariates. More ambitiously, Aalen and Gjessing (2001) study threshold-crossing times for a much wider class of continuous-time continuous-state Markov processes with the objective of deriving qualitative properties of the hazard functions for crossing times from the underlying process properties. They produce some interesting results, but there are few examples where this program is analytically tractable.

The idea of failure as a process of arriving at a death-state by way of time-homogeneous Markovian transitions between discrete states is a simpler version of the threshold-crossing idea, and is the definition of a phase-type model. In a wide array of applied-probability and statistical investigations, these models have proven useful, and they are the ones we focus on for the rest of the paper.

1.4 Phase-type models

A *phase-type* random variable T is defined as the absorption time into a termination- or death-state in a continuous-time homogeneous finite-state Markov chain. If the chain with transition intensity matrix $Q = \{Q_{ab} : a, b = O, 1, \ldots, K, D\}$ starts in the initial state denoted O, with the terminal death-state (only one is needed) denoted D, and the other states $\{1, \ldots, K\}$, then

$$P(T \le t) = (\exp(tQ))_{OD} = P_{OD}(t)$$

Such models are natural for cascades of discrete states reflecting accumulation of mutation or damage, which accordingly experience different rates of immediate failure. However, we must distinguish conceptually the counting-process models with observable states and nonparametrically estimated general transition intensities from the latent-state Markovian models for which only the survival duration (often right-censored) can be observed.

Models of this type were proposed separately in Queueing Theory (Neuts 1975) and in the classic Illness-Death model (see Andersen et al. 1993 for references). In both of these settings, the state transitions may be observable. However, the terminology of *phase-type models* also refers to waiting times for failure or other events in which intermediate states are latent or unobservable. Thus, in pharmacokinetic compartmental modeling, the overall time for a drug or chemical to remain in the human body may be idealized as a succession of sojourns in organs or other subsystems which can be observed indirectly if at all, and the parameters of the duration distribution are of primary interest. (References can be found in Macheras and Iliadis (2006)).

The phase-type distributions introduced by Neuts in 1975 as a generalization of the Erlang distribution have been widely used in stochastic models in queueing and telecommunication (Sengupta 1989; Asmussen 1992; Ausin et al. 2004), traffic flow (Thümmler et al. 2006), actuarial science (Lin and Liu 2007), health care (Faddy and McClean 1999), and survival analysis (Aalen 1995; Olsson 1996).

Among the other phase-type applications cited in the previous paragraph, the actuarial phase-type model of Lin and Liu (2007) deserves special mention because of its connection to actuarial and demography lifetime models discussed in Sect. 1.1. These authors provide a model representing human mortality as an ordered sequence of many intermediate states $\{1, 2, \ldots, n\}$ in which each state also has a possible direct transition to the death-state D. To maintain parametric parsimony, Lin and Liu define a simple common parameterization of transitions $k \mapsto k+1$ as well as a power-law form of death-transitions $Q_{kD} = b + ak^c$ reminiscent of the Weibull death-rate function. They show that this parameterization, with as few as 6–9 total parameters, is remarkably successful in reproducing subtle features of the age-specific death-rate curves of three historical Swedish population cohorts.

The phase-type distributions are known to be dense (in the sense of pointwise convergence of distribution functions) among all continuous distributions on the positive half-line. They are appealing because they include several of the most important constructions used by applied probabilists to describe realistically complex waiting-time phenomena: as shown in the following Proposition, the phase-type class is closed under finite mixtures, as well as under minima, maxima, and sums of independent waiting-time random variables.

Proposition 1 *(Neuts 1981) Suppose that* $T_1, T_2, \ldots, T_m$ *are phase-type random variables, with respective densities* $f_{T_j}(\cdot)$.

(a) If $(p_1, \ldots, p_m)$ *is a probability vector, then the mixture random variable* T_* *with density* $\sum_{j=1}^{m} p_j f_{T_j}(x)$ *is also a phase-type variable.*

(b) The sum $T_1 + \cdots + T_m$ *is a phase-type random variable.*

(c) Both $\min\{T_j : j = 1, \ldots, m\}$ *and* $\max\{T_j : j = 1, \ldots, m\}$ *are phase-type random variables.*

Proof Let the states, initial distribution, and transition intensities of the phase-type Markov chains M_j whose absorption times are T_j be denoted respectively, for $1 \le j \le m$, by $s \in S_j$, by $\pi_j(s)$, and by $Q_j(s_1, s_2)$ for $s_1, s_2 \in S_j$. Denote the terminal (death) state in the j'th chain by D_j. In the first two parts of the proof, we define a Markov chain M with states $\cup_{j=1}^m S_j$, after identifying certain states and defining a suitable initial distribution, for which the absorption time into a designated death-state D is the desired random variable.

(a) Now the initial distribution is defined for all $j = 1, \ldots, m$ and $s \in S$ by $\pi(s) = \sum_{j=1}^m p_j \pi_j(s) I_{[s \in S_j]}$. Define the state $D \equiv \cup_{j=1}^m \{D_j\}$ by lumping the death-states of all the chains M_j into a single death-state. The chain M (with intensity matrix Q) allows only the transitions $s \mapsto s'$ allowed (for $s, s' \in S_j$ for some j) by the separate chains M_j, with the intensity

$$Q(s, s') = \sum_{j=1}^{m} I_{[s,s' \in S_j]}\, Q_j(s, s') \qquad \text{for} \qquad s,\ s' \in S$$

All other transitions are impossible, i.e., are given transition intensity 0. In this chain, the waiting time to absorption is exactly T_j if the initial state lies in S_j, which is an event of probability p_j. Therefore the unconditional absorption time is distributed according to the mixture with probabilities p_j of the distributions of the respective times T_j, as desired.

(b) In this case, the initial distribution is defined to be $\pi_1(\cdot)$ on S, and the overall death state for the new chain is defined as D_m. Moreover, in the newly defined chain, each transition $s \mapsto D_j$ for $j = 1, \ldots, m-1$ and $s \in S_j$ is disallowed (given intensity 0), and new transitions $(j, s) \mapsto (j+1, s')$ for all $s' \in S_{j+1}$ are included, with intensities

$$Q(s, s') = \sum_{j=1}^{m-1} I_{[s \in S_j,\, s' \in S_{j+1}]}\, Q_j(s, D_j) \cdot \pi_{j+1}(s')$$

That is, in this new chain the transitions to intermediate death-states D_j at the expiration of the successive waiting-times T_j are replaced by transitions to the starting states for the T_{j+1} chain, with probabilities according to the initial distribution for the $j+1$ chain.

(c) For each of the desired constructions in this part, the state space consists of the cartesian-product space $S' = S_1 \times S_2 \times \ldots \times S_m$, the initial distribution defined by

$$\pi(s_1, s_2, \ldots, s_m) = \prod_{j=1}^{m} \pi_j(s_j)$$

and the allowed transitions given, for $s_k \in S_k$, $k = 1, \ldots, m$, by

$$(s_1, s_2, \ldots, s_m) \mapsto (s_1, \ldots, s_{j-1}, s', s_{j+1}, \ldots, s_m) \qquad \text{for} \qquad s' \in S_j$$

with Q-matrix intensity equal to $Q_j(s_j, s')$. For this Markov Chain definition, the absorbing terminal state-set is defined to be

$$D \equiv \{(s_1, \ldots, s_m) : \ s_j = D_j \quad \text{for some} \quad j = 1, \ldots, m\}$$

in order to achieve $\min(T_1, \ldots, T_m)$ as overall absorption time; and the terminal state-set is defined as

$$D \equiv \{(s_1, \ldots, s_m) : \ s_j = D_j \quad \text{for all} \quad j = 1, \ldots, m\}$$

in order to achieve $\max(T_1, \ldots, T_m)$ as overall absorption time. □

Thus, mixtures of exponentials are of phase type, which is the main avenue for the introduction of decreasing hazard rate failure times into the class. Note also that phase-type models with general initial distributions can always be viewed as mixtures of waiting times for models with fixed initial state. (The part of the chain beginning from each state k is newly defined as a separate branch or path.) Because of (b), the time until crossing of a fixed threshold by a random walk whose times between jumps are phase-type is also a phase-type random variable. Since Wiener processes with drift are distributional limits of random walks with exponentially distributed time-steps, also inverse-Gaussian and other cumulative-damage models are naturally viewed as examples or distributional limits of phase-type models. The multi-hit models considered by Armitage and Doll were essentially pure-birth Markov-chain models of this type, leading to Gamma and Erlang random variables, and more general pure-birth models have been considered in the carcinogenesis modeling literature (Moolgavkar 2004).

One of the modeling ideas often used in phase-type models is to allow transition steps from many intermediate states directly to the absorbing death-state. Models with such arcs are often called *Coxian* (following Cox 1955), and such direct transitions are often modeled to have the same or functionally related transition-rate parameters. Including Coxian arcs in phase-type models has been shown to prevent numerical difficulties in calculating phase-type distributional quantities, and such models have found extensive applications in many fields of study, in particular, in studies of duration of hospital stays, for which a good review can be found in Marshall and Zenga (2009). Further extensions of Coxian models to incorporate regression terms involving covariates into transition rates have also been studied, for example in Faddy and McClean (1999).

Phase-type models have also been adapted to analyze censored data. Olsson (1996) extended the EM algorithm of Asmussen et al. (1996) to find parameter estimates from either right censored or interval censored data. Aslett and Wilson (2011) proposed a Bayesian method to fit phase-type distributions to right censored observations.

In a phase-type model, each waiting time to leave a state is exponentially distributed. A natural further generalization is to define 'flow-graph' models which allow the time spent in each state j to be a parametric non-exponentially distributed random variable as in Huzurbazar (1999), which may depend on j and may also depend on the next state to be visited (the hallmark of a semi-Markov model, Huzurbazar (2005), chap. 7).

In this way, whole sections of a phase-type diagram with a single entry state and a single exit state can be replaced by random variables with simple parametric forms.

1.5 Methods of fitting phase-type models

Although the Phase-type distributions are flexible, it is known that the representations of distributions they provide are not unique (O'Cinneide 1989). In other words, Phase-type representations may be badly over-parameterized. Considerations of model parsimony have led many authors to constrain many of the phase-type transition rates to be the same or functionally related. Examples include Lin and Liu (2007) in connecting the direct ('Coxian') failure rates from a succession of internal states to follow a power-law plus additive constant, as well as the restriction by Bobbio et al. (2003) to a model subclass called acyclic phase type (APH) distributions, and by Thümmler et al. (2006) to a small subset of mixtures of Erlang distributions.

Many fitting methods for general Phase-Type distributions or subclasses have been proposed. Four main methods are Moment Matching (Bobbio et al. 2005), Numerical nonlinear minimization (Johnson 1993), Expectation-Maximization (EM) algorithms (Asmussen et al. 1996; Olsson 1996; Lee and Lin 2010), and Bayesian methods (Bladt et al. 2003; Ausin et al. 2004; McGrory et al. 2009).

Several software packages are available for likelihood inference in phase-type models. For example, `EMpht` (Olsson 1998) is a C-language program that implements the EM algorithm studied in Asmussen et al. (1996) and Olsson (1996) both for uncensored and right-censored waiting-time data. The `R` package `PhaseType` (Aslett 2011) embodies the Bayesian-based methods discussed in Bladt et al. (2003) and Aslett and Wilson (2011). The `KPC`-Toolbox (Casale et al. 2010) is a `MATLAB` library that fits Markovian arrival processes and can be adapted to fit Phase-Type distributions. The `PhFit` general Phase-Type fitting tool (Horváth and Telek 2002) approximates data distributions by continuous and discrete Phase-Type models. The `R` package `actuar` (Dutang et al. 2008) provides many actuarial functions including basic characteristics of general Phase-Type distributions.

2 Motivations for parametric densities

Some motivation must still be given for low-dimensional parametric models, in the current era when biostatistical and reliability techniques (Andersen et al. 1993) allow right-censored large-sample survival data analyses to accommodate nonparametric baseline hazards for various semiparametric regression models. Parametric models still have a very useful role to play, for at least four different reasons:

(I) Subjects in large epidemiologic studies or databases (like the SEER cancer database) are often extensively cross-classified into many cells of moderate size (100 or less) with widely varying prognosis; in such settings, low-dimensional parametric models may be the best available in individual cells, and the parameter vectors themselves can usefully describe the cellwise survival differences.

(II) It is desirable to describe qualitative features of whole-population survival curves which can arise from simple mechanisms (like the pure-birth successive-mutation hypothesis of Armitage and Doll), and to distinguish them from mixtures of distinct phenomena which strongly indicate two or more different diseases within a single named disease entity. This aspect of parametric, and also of phase-type, modeling is introduced explicitly in a paper of Manton and Stallard (1980) and entails a search for mixture components, which is meaningful within parametric but not nonparametric distributional classes. In the case of breast cancer, biomedical research has shown clearly (Anderson et al. 2006) that subpopulations defined by estrogen receptor (ER) status, positive versus negative, account for two different peaks in the survival density.

(III) Models with interpretable transitions between latent states along disease pathways can be scientifically interesting in suggesting new observations to make, transition steps which are influenced by particular risk-factors, or separate disease entities represented by specific states.

(IV) More specifically, models (whether fully parametric or not) including regression parameters for the direct influence of covariates on single transition rates can be used to investigate the relative importance of various exposure variables on single transition steps toward disease incidence and mortality.

In the remainder of this paper, we define a phase-type survival-density class with 3 to 8 parameters for which densities and likelihoods are explicitly computed; describe tools based on likelihood ratios and Fisher Information to clarify which parameters can be stably identified within parametric densities; and illustrate the fitting and interpretation of the model on breast-cancer data from the SEER 1992–2002 database. Within the model class we study, we indicate how survival regression models would be naturally introduced. Finally, we bring together our own computational experience with that found in the literature on phase-type modeling to draw some conclusions about the feasibility and advisability of fitting complicated phase-type models to right-censored survival data.

3 Parameters and likelihood for a simple phase-type model

We study statistical inference within a moderately parameterized phase-type model family. The particular topology we consider, displayed in Fig. 1 below and cited as `Model F`, seems to us particularly appropriate in a survival setting for which the time origin and initial state O correspond to diagnosis and first treatment for a serious disease like a cancer. Immediately after treatment, direct transitions to death (state D) or a cure/quiescent state C are possible, but there may also begin a slower process of migration or mutation of existing diseased or precursor cells, along one or more pathways determined either by new internal biological events (e.g., mutations related to environmental or radiologic exposures) or by genetics (alleles related to disease susceptibility). Because of our motivating breast cancer data illustration in the following Section, we are interested in allowing the data to impose a model structure involving two separate disease paths, paths which are known (Anderson et al. 2006) to correspond to positive and negative Estrogen Receptor status in breast cancer. The

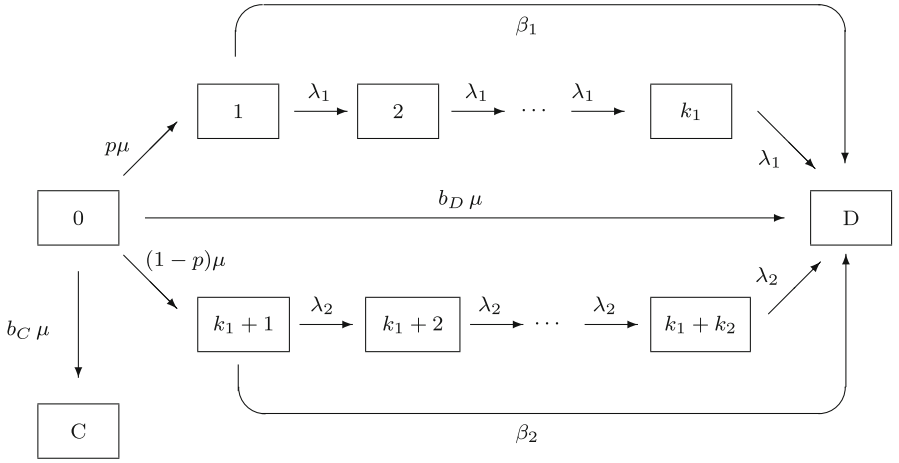

Fig. 1 Markov transition diagram for `Model F` with immediate cures and failures, additional direct failures from states 1, 2, and two failure pathways

Markov chain transition intensities are given in Fig. 1, and can be understood to say that the chain begins by waiting in state O for a time $T_1 \sim \text{Expon}((1+b_C+b_D)\,\mu)$, and then jumps to one of the states C, D, 1, or k_1+1, with respective probabilities defined by

$$(p_C,\ p_D,\ p_1,\ p_2) = \frac{1}{1+b_C+b_D}\,(b_C,\ b_D,\ p,\ 1-p) \tag{3.1}$$

States C and D are absorbing; from state 1 the time to absorption in D is equal with probability $q_1 = \beta_1/(\beta_1+\lambda_1)$ to a r.v. $T_{1D} \sim \text{Expon}(\lambda_1+\beta_1)$ and with probability $1-q_1$ to the sum of T_{1D} and an independent variable $G_1 \sim \text{Gamma}(k_1-1,\ \lambda_1)$; and from state 2 the time to absorption in D is equal with probability $q_2 = \beta_2/(\beta_2+\lambda_2)$ to a r.v. $T_{2D} \sim \text{Expon}(\lambda_2+\beta_2)$ and with probability $1-q_2$ to the sum of T_{2D} and an independent variable $G_2 \sim \text{Gamma}(k_2-1, \lambda_2)$. Note that if $\beta_1 = 0$, then the overall waiting time from state 1 to reach D is distributed as Gamma (k_1, λ_1). The decomposition into waiting times T_{1D} and G_1 accounts separately for the time to leave state 1 and to progress from 2 to D after the transition $1 \mapsto 2$. In this description, the properties of Markov chains and exponential waiting times ensure that at all branches, the branching events result from trials which are independent of all waiting times. If either of the Gamma shape parameters k_j is equal to 1, then the corresponding intensity pair (β_j, λ_j) is unidentifiable and the two transition arcs with these intensities can be replaced by a single arc with transition intensity $\beta_j+\lambda_j$. Thus, if $k_j = 1$, without loss of generality $\beta_j = 0$.

It is apparent from the foregoing paragraph and the Definitions (3.1) that the absorption time density into state D for the pictured `Model F` Markov chain is a mixture with weights p_D, p_1q_1, p_2q_2, $p_1(1-q_1)$, and $p_2(1-q_2)$ of the Expon$((1+b_C+b_D)\mu)$, Expon$((1+b_C+b_D)\mu) * \text{Expon}(\beta_1+\lambda_1)$, Expon$((1+b_C+b_D)\mu) * \text{Expon}(\beta_2+\lambda_2)$, Expon$((1+b_C+b_D)\mu) * \text{Expon}(\beta_1+\lambda_1) * \text{Gamma}(k_1-1, \lambda_1)$,

and $\text{Expon}((1+b_C+b_D)\mu) * \text{Expon}(\beta_2+\lambda_2) * \text{Gamma}(k_2-1, \lambda_2)$ densities, where $*$ denotes convolution. The weights in this mixture add up to $1 - p_C < 1$ because of the positive probability p_C with which the chain is absorbed at C and never hits D. The convolutions in these densities are in fact easy to write down in closed form, for positive integers k_1, k_2, which makes the densities and survival functions fully explicit and easy to compute in vectorial form in the likelihood for `Model F` based on right-censored survival data. Computing formulas that allow these calculations to be implemented rapidly in `R` code can be found in the Supplement to this paper (Slud and Suntornchost 2013).

The `Model F` Markov chains include a variety of cure models along with the Erlang-type multi-hit model considered by Armitage and Doll (1954), including special cases of that model with up to 3 distinct rates for successive mutation 'hits'. Models of these types can all be accommodated within cases of `Model F` for which $p = 0$ or $p = 1$, and we refer to the resulting phase-type absorption times as 'single-path Model F' densities. As a matter of notation, we refer to the single-path `Model F` absorption density with $p = 1$ in Fig. 1 as the $(b_C, b_D, \mu, \beta_1, \lambda_1)$ single-path density, with shape-parameter k_1 generally fixed. The `Model F` class was designed to include such single-path densities as well as a large class of two-component mixtures of them, which we will find to be particularly useful in the data illustration of Sect. 4. The formal result justifying this idea is the following Lemma.

Lemma 1 *The mixture with weights p and $1-p$ of the single-path model F densities which have respective parameters $(b_C, b_D, \mu, \beta_1, \lambda_1)$ with shape k_1 and $(\tilde{b}_C, \tilde{b}_D, \tilde{\mu}, \beta_2, \lambda_2)$ with shape k_2, is again a* `model F` *phase-type density if and only if $(1+\tilde{b}_C+\tilde{b}_D)\,\tilde{\mu} = \bar{\mu} \equiv (1+b_C+b_D)\,\mu$.*

Proof The stated condition is necessary because the two single-path models respectively have $\text{Expon}((1+b_C+b_D)\,\mu)$ and $\text{Expon}((1+\tilde{b}_C+\tilde{b}_D)\,\tilde{\mu})$ distributed waiting times until exit from the initial state. See the discussion immediately following Fig. 1 to see that each of the phase-type single-path densities is itself a mixture of an exponential density with other convolved-density components; the mixture of two such mixtures cannot be of the same type unless the single exponential density term in both mixture components is the same.

Now suppose that the condition of the Lemma holds, and that $p \neq 1$. Then the expression of the `Model F` absorption time density with parameters

$$(p^*,\ b_C^*,\ b_D^*,\ \bar{\mu}/(1+b_C^*+b_D^*),\ \beta_1,\ \beta_2,\ \lambda_1,\ \lambda_2)$$

as a mixture of an exponential density and convolutions is the same as the expression for the mixture with weights $p,\ 1-p$ of the two single-path model F densities as long as all three of the following equalities hold

$$\frac{p^*}{1+b_C^*+b_D^*} = \frac{p}{1+b_C+b_D}, \qquad \frac{1-p^*}{1+b_C^*+b_D^*} = \frac{1-p}{1+\tilde{b}_C+\tilde{b}_D}$$
$$\frac{b_C^*}{1+b_C^*+b_D^*} = \frac{p}{1+b_C+b_D}\,b_C + \frac{1-p}{1+\tilde{b}_C+\tilde{b}_D}\,\tilde{b}_C.$$

We solve these equations explicitly for parameters $p^* \in [0,1]$, b_C^*, b_D^*. First, taking ratios of the first two of these equations leads to the equality

$$\frac{p^*}{1-p^*} = \frac{1+\tilde{b}_C+\tilde{b}_D}{1+b_C+b_D} \cdot \frac{p}{1-p}$$

which uniquely determines $p^* \neq 1$. Next, substituting the first two equalities in the third shows that $b_C^* = p^* b_C + (1-p^*)\tilde{b}_C$. Also, subtracting the sums of the three equalities from 1 on each side shows that the third equality holds with C's and D's reversed, from which it follows that $b_D^* = p^* b_D + (1-p^*)\tilde{b}_D$. The proof of the Lemma is complete. □

3.1 Simulations illustrating model comparisons

We consider parameter estimation based on several simulations with phase-type models of the type of Fig. 1, with the objective of illustrating the extent to which different phase-type models can be distinguished based on reasonable sized datasets. In our simulations, we restrict to cases with $b_D = b_C = 0$, so that immediate transitions $O \mapsto D$ and $O \mapsto C$ are rendered impossible. In the first series of simulations, we also fix parameters $\beta_1 = \beta_2 = 0$ and $(p, \mu, \lambda_1, \lambda_2) = (0.30,\ 2.00,\ 0.20,\ 0.30)$, with $k_1 = 4$, $k_2 = 3$, and generate one uncensored data-sample for each of the sample-sizes $N = 100,\ 1000,\ 10000,\ 20000, 100000$. The purpose here is not to illustrate the mechanics of convergence, but rather to use simulations to document the convergence of the per-observation Fisher Information matrix, and thereby to indicate that certain parameters—i.e., linear combinations of parameters which are close to lying in the span of the larger-eigenvalue eigenvectors—are estimated reasonably well in moderate samples, while others are estimated badly even in large samples. In these examples, the survival time is a mixture of convolutions of exponential and Gamma random variables, and the likelihood is explicit, so that ML computations by quasi-Newton-Raphson optimization are very fast. ML estimators of the transformed parameters $\vartheta = (\mathrm{logit}(p),\ \log(\mu),\ \log(\lambda_1),\ \log(\lambda_2))$ were calculated for each of the simulated datasets, with the results shown in Table 1 along with estimated standard errors (SE's, in parentheses). The table shows how remarkably large the sample size must be for the MLE's to be close to the true values. This is a feature of latent-class models generally, and can be studied more precisely by analyzing the Fisher Information for ϑ.

The per-observation Fisher Information matrices are estimated as the negative Hessian matrices evaluated at the ML (transformed) parameter values, divided by sample size. This *observed Fisher Information* $\hat{I}_1(\hat{\vartheta})$ stabilizes nicely for the successively larger sample sizes, and for $N = 10^5$ has eigenvalues $1.2601,\ 0.770,\ 0.0054,\ 0.0012$. As a consequence, for sample size N, large-sample theory predicts standard errors (obtained by taking reciprocal square roots) $0.891/\sqrt{N}$, $1.139/\sqrt{N}$, $13.550/\sqrt{N}$, and $28.911/\sqrt{N}$, respectively for the linear combinations of the parameter estimates $v_1\,\mathrm{logit}(\hat{p}) + v_2 \log(\hat{\mu}) + v_3 \log(\hat{\lambda}_1) + v_4 \log(\hat{\lambda}_2))$ for each of the four unit eigenvectors of the Information matrix. We use the estimated information matrix and eigenvectors at the largest simulated sample size (10^5) in place of the theoretical values. Thus, at

Table 1 Parameter $(p,\ \mu,\ \lambda_1,\ \lambda_2)$ MLE's and SE's (in parentheses) on transformed scale (logit for p, log for others) by sample size N, for single simulated datasets with $k_1 = 4, k_2 = 3$

	True	$N = 100$	$N = 1,000$	$N = 10,000$	$N = 20,000$	$N = 100,000$
logit(p)	−0.8473	−0.4273	0.1693	−0.8167	−0.9354	−0.7541
(SD)		(0.5218)	(0.6391)	(0.1859)	(0.1469)	(0.0571)
log(μ)	0.6931	−1.1358	−0.9828	0.4574	0.7304	0.5782
(SD)		(0.9058)	(1.0260)	(0.2325)	(0.1878)	(0.0815)
log(λ_1)	−1.6094	−1.4574	−1.3981	−1.6000	−1.6227	−1.5952
(SD)		(0.0941)	(0.1538)	(0.0332)	(0.0256)	(0.0101)
log(λ_2)	−1.2040	−0.6957	−0.6187	−1.1835	−1.2213	−1.1769
(SD)		(0.2788)	(0.3469)	(0.0465)	(0.0313)	(0.0149)

the moderately large sample size of $N = 1000$, the first eigenvector parameter combination $0.216\,\text{logit}(\hat{p}) - 0.075\ \log(\hat{\mu}) - 0.425\ \log(\hat{\lambda}_1) - 0.877\ \log(\hat{\lambda}_2))$ is fairly well estimated at 1.246 with predicted standard error of 0.028, while the fourth eigenvector combination $0.482\,\text{logit}(\hat{p}) - 0.859\,\log(\hat{\mu}) + 0.082\,\log(\hat{\lambda}_1) + 0.152\ \log(\hat{\lambda}_2)$ is very badly estimated at 0.717 with predicted standard error of 0.914. Ill-conditioning in the Information matrix results in dramatic differences between the quality of identification of parameters. In particular, the parameter $\text{logit}(p)$ which has large eigenvector coefficients only for the third and fourth eigenvectors, is very badly resolved, with standard error $(\hat{I}(\hat{\vartheta})^{-1})_{11}/\sqrt{N} = 0.571$, while the parameter $\log(\lambda_1/\lambda_2)$ is fairly well identified, with estimated value of -0.779 and standard error 0.083.

4 Model fitting on breast cancer data

In this Section, we fit the `Model F` parametric class of densities to the White Female SEER dataset on mortality in 13-registry US databases of breast cancer cases diagnosed between 1992 and 2001 and followed through 2002. Details concerning the data, a spline-based fitting methodology, and discussion can be found in Anderson et al. (2006). Of the complete dataset of 243,808 cases, we analyzed only the 198,785-case subset of White females with age at diagnosis from 30 to 89, for breast-cancer mortality. Although the primary focus of the Anderson et al. study was to understand the shape of post-diagnosis hazard as a mixture of the disaggregated disease types indicated by Estrogen Receptor (ER) status, we omitted that covariate from our analysis, since our objective is to learn what a purely parametric statistical analysis using the model of Sect. 3 could have told about the likely mixture components of breast-cancer mortality in the combined population.

While Anderson et al. (2006) directly created spline-fitted hazard functions for their combined and ER-disaggregated study populations, we performed a slightly more complicated preliminary analysis designed to correct for year-of-diagnosis mortality differences, since Kaplan-Meier curves for the data stratified by diagnosis year (`DiagYr`) showed a small but clear trend in decrease of hazards with `DiagYr`. The cumulative hazards were nearly linear for the datasets with `DiagYr` after 1996, with

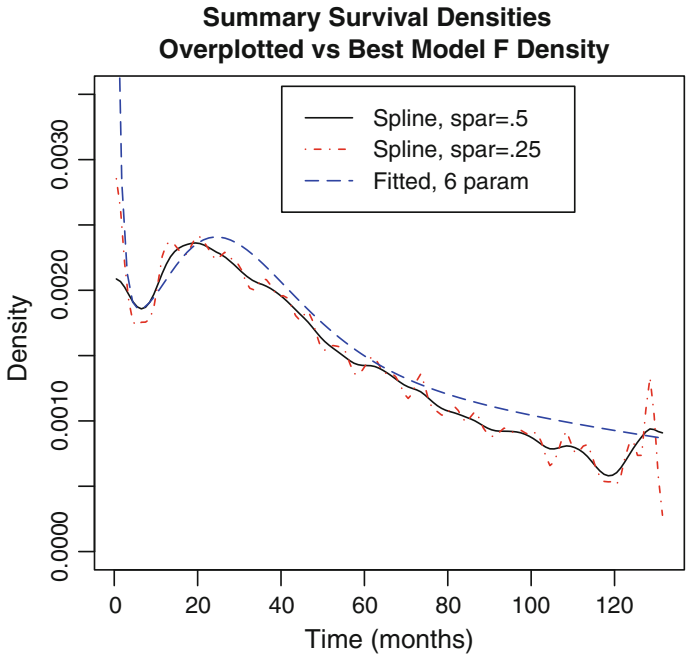

Fig. 2 Spline and fitted density functions to the SEER 1992–2002 data on US white 30–89 female breast-cancer mortality following diagnosis

a slight concavity over times 6–11 years for earlier `DiagYr`'s. Since the nonparametrically fitted hazards were therefore approximately proportional across `DiagYr`, we fitted a Cox proportional hazards model with a dummy variable for `DiagYr` as the only covariate, finding contrast effects for `DiagYr` versus 1992 to be 0.007, −0.024, −0.065, −0.093, −0.138, −0.161, −0.236, −0.285, −0.292.

We present as our basic nonparametric mortality curve the summary survival curve for that Cox model, plotted at the raw survival times of 0:131 months plus 0.5, to which we fitted a smoothing spline using the `R` function `smooth.spline`, with smoothing parameter `spar=0.5`. Figure 2 shows the corresponding survival density, along with one done the same way but with less smoothing (`spar=0.25`), along with the best fit (plotted in blue) that we were able to find to the data, a 6-parameter model within Model F for which the Coxian rates β_1, β_2 are fixed equal to 0. In this fitted model, as in all those treated below, $k_1 = 4$ and $k_2 = 1$. (A 5-parameter variant model which looks visually identical to the 6-parameter density in Fig. 2, but fits slightly worse, is obtained by letting the μ rate-parameter in Fig. 1 go to ∞, which has the effect of making the limiting `model F` density a mixture of two exponential components and two convolutions.)

The solid spline-fitted curve in Fig. 2 closely resembles the summary all-patients survival hazard pictures in Anderson et al. (2006). The spline fit to the same Cox-model summary survival, but with lesser smoothing (dotted curve in Fig. 2), shows more clearly the overall features of the density which a parametric model should seek to reproduce. These features include a high initial spike in hazard, a density peak near

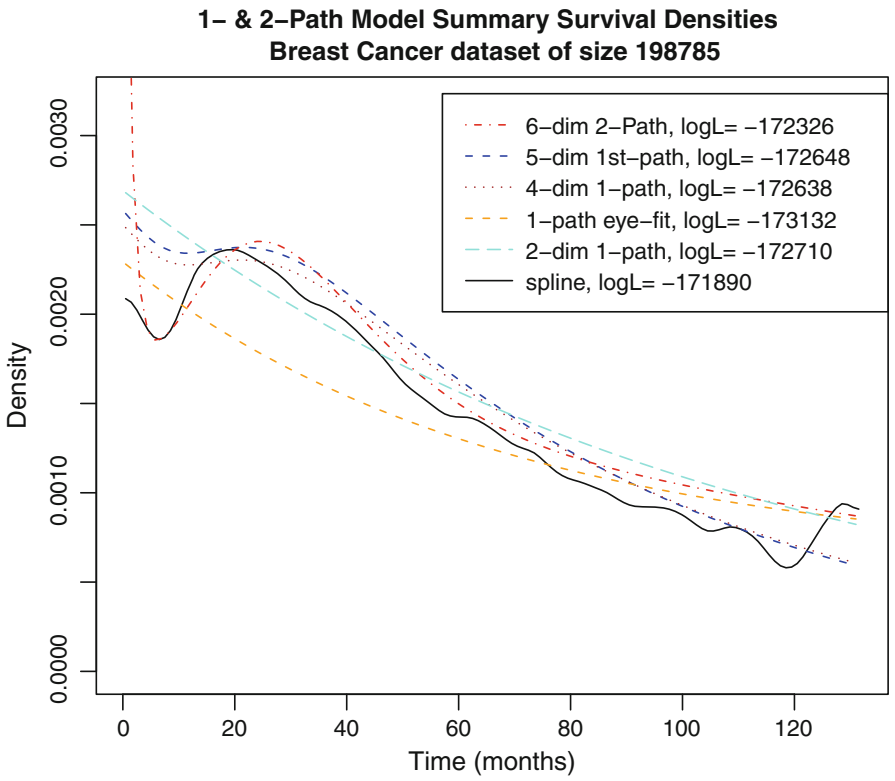

Fig. 3 Spline and 5 parametric Model F density functions ML fitted to the SEER 1992–2002 data as in Fig. 2 on US female breast-cancer mortality following diagnosis

20 months, an approximately linear decrease of density between 20 and 120 months, and a final turn up in density between 120 and 130 months. Presumably the initial hazard spike is due to immediate adverse outcomes from surgery and untreatable advanced-stage cancers, and the peak and density decrease from 20 to 120 months are due to the recent successes in treating a large fraction of cancers detected at early stages. But we cannot account for the final upturn in hazard, which our models do not address at all.

We describe next the computational steps used to fit successively more complicated models within the `model F` class to the Breast Cancer dataset. First of all, we added to all of the 198,785 survival times, which are whole numbers of weeks ranging from 0 to 131, random variates supported in [0, 1) which make the death-hazards constant on each interval $[k, k+1)$ for the resulting continuous variates, without changing the empirical all-cause survival function at the discrete integer endpoints. Thus, if $\hat{S}(k)$ denotes the proportion of discrete survival times greater than or equal to k, and $\hat{\mu}_k \equiv \log(\hat{S}(k)/\hat{S}(k+1))$ for $k = 0, 1, \ldots, 131$, then we added to all survival times $T_i = k$ independent random variates τ_i with density $\hat{\mu}_k\, e^{-s\hat{\mu}_k}/(1-e^{-\hat{\mu}_k})$, $0 \le s < 1$. The reason for doing this is to avoid the artificiality of 0 survival times or, if a constant like 0.5 were to be added to all survival times, to avoid distorting the `model F` parameter estimation because of the absence of deaths on [0, 0.5).

Most of the computational work in finding good starting parameter values for fitting the models displayed in Fig. 3 was done on a single set of 20,000 patient records randomly selected from the full dataset of 198,785 records. (Although purely a numerical optimization strategy in the present context, this use of parametric log-likelihood on a sampled dataset is well known within survey-sampling theory as a *pseudolikelihood* estimation method, Binder (1983)). Within each model class of fixed parameter dimension, the right-censored survival-data log-likelihood coded in `R` (R Development Core Team 2011) was maximized by a quasi-Newton-Raphson algorithm using the `R` function `nlm`, convergence of which was rapid (of the order of 1 minute for the dataset of size 20,000, on a 2.10 `gHz` laptop) but also sensitive to the choice of starting parameter values.

In order to avoid the selection of an incorrect local peak by the ML optimization, it was necessary to fit a nested succession of models within `model F`, increasing the parameter dimension by at most 1 or 2 at each step. The precise sequence used in fitting the models displayed in Fig. 3 was as follows:

(i) The initial, 2-dimensional model, was based on a single path ($p = 1$, which makes β_2, λ_2 irrelevant), with rates 0 along Coxian arcs ($b_D = \beta_1 = 0$); thus, only the parameters $b_C,\ \mu,\ \lambda_1$ remained to be estimated, and initially we restricted $\lambda_1 = e^{30}$, so the model provides probabilities $b_C/(1+b_C)$ of "cure" (or infinite time to hit D) and $1/(1+b_C)$ of an $Expon(\mu(1+b_C))$ waiting time. The fitted values were: $\hat{b}_C = 2.3551,\ \hat{\mu} = 0.0028$.

(ii) The next (4-dimensional) model allowed free parameters $(b_C,\ b_D,\ \mu,\ \lambda_1)$, with other parameters fixed as in (i), and was initialized at the values $(2 \cdot 2.3551,\ 1,\ 0.0028,\ \exp(-2))$. Note that with $b_D = 1$, the choice $b_C = 2 \cdot 2.3551$ yields the same cure-probability $b_C/(1 + b_C + b_D) = 0.702$ as in the fitted model of (i). The optimized ML parameters in this model were: $(11.5502,\ 2.7578,\ 0.0009,\ 0.1857)$.

(iii) A 5-dimensional one-path model augmented the parameters of (ii) by allowing β_1 to range freely, initialized close to 0 (say, 0.0001). The optimized model had approximately the same log-likelihood as the model in (iv), but gave a visually different density. (See Fig. 3, 5-dim fit below).

(iv) Guided by Lemma 1, we next sought a single-path `model F` density with $p = 0$ and with parameters $\tilde{b}_C,\ \tilde{b}_D,\ \tilde{\mu}$ satisfying $(1 + \tilde{b}_C + \tilde{b}_D)\,\tilde{\mu} = 0.01433$, which is the corresponding parameter from the fitted model (iii). We varied parameters $(\tilde{b}_C,\ \tilde{b}_D,\ \tilde{\mu},\ \lambda_2)$ based visually on density plots, subject to this restriction, until arriving at the curve labeled "1-path eye-fit" in Fig. 3 as the lower-path `model F` to mix with the single upper-path model (iii) via Lemma 1. Note that because $k_2 = 1$, there was no loss of generality in restricting $\beta_2 = 0$ in this and the next model.

(v) Finally, we used Lemma 1 to define a 7-dimensional mixture parameter to initialize the search for a general `model F` parameter (for $\beta_2 = 0$), with $p = 0.9$. The model found by quasi-Newton-Raphson maximization of log-likelihood is the 7-dimensional model which turned out to be the best fit to the Breast Cancer data within `model F`. This model turned out to have an extremely small (and

Table 2 Parameters and log-likelihoods for models in Fig. 3, with $k_1 = 4, k_2 = 1, \beta_2 = 0$

# par.	p	μ	λ_1	λ_2	β_1	b_C	b_D	logLik
2	1	0.00269	100	100	0	2.360	0	−172,710
4	1	0.00086	0.1862	100	0	12.064	2.911	−172,638
5	1	0.00094	0.1856	100	0.00023	11.549	2.758	−172,648
6	0.090	0.3015	0.1180	0.0056	0	1.811	0.022	−172,326

non-significant) value of β_1, so re-fitting it with the restriction $\beta_1 = 0$ gave essentially the same fit as in (v), with a 6-dimensional parameter.

At stages (iii) and (v) of the model-fitting, we varied the choice of k_1, and at the final stage (both the 6- and 7-dimensional versions) also k_2, re-fitting the other parameters for each (k_1, k_2) choice. We found that the log-likelihood was larger by 0.1 to 0.2 as k_1 moved from 4 to 2 (based on sample size 20,000), while $k_2 = 1$ could not be improved. Since the optimization of the `model F` parameters became slightly less stable (due to the near-nonidentifiability of β_1 as k_2 became smaller), we judged the change in k_2 to be not worth making.

The models compared visually in this Section can further be understood through their log-likelihood values on the SEER breast-cancer data. (We made all calculations of these log-likelihoods using the discrete survival time in weeks plus the random variates τ_i between 0 and 1 week, mentioned above.) We first clarify the relationship between visual fidelity of fitted survival densities and purely statistical model comparisons via likelihood ratio tests. Table 2 displays ML estimated model-parameters and log-likelihoods for the SEER data used in producing Fig. 3, i.e., the SEER data on breast cancer mortality following diagnosis for white females aged 30–89. The log-likelihood differences between the models are large, because of the large sample size. For purposes of comparison, the log-likelihoods on the same data for the models whose densities are plotted in Fig. 2 are $-171,453$ for the spline-fitted survival density with `spar`$= 0.25$, $-171,899$ for the spline-fitted survival density with `spar`$= 0.5$, and $-172,326$ for the best-fitting (6-parameter, 2-path) model.

The Figures and log-likelihoods shown, and the results of other analyses not shown, demonstrate clearly that the essential features of the density curves up to 120 months can be captured only by 2-path models, in other words mixture models, within the phase-type `model F` class. Figure 3 also indicates that each increase in parameter dimension allows an additional visual feature of the empirical smoothed density—which the spline fit displays—to be captured by the parametric model: the 2-parameter one-path model captures the early and late density levels; the 4-parameter one-path model additionally captures the approximate curvilinear pattern of decrease of the density; the 5-parameter model begins to capture the initial hook (decrease and then increase to local peak); and the 2-path 6-parameter model follows (and even exaggerates) the initial hook, although the less-smoothed spline picture in Fig. 2 does show a sharp initial density decrease) while closely following the local peak near 20 months.

It is well known that latent-class and mixture models often have poorly identified parameters, sometimes even for strikingly large sample sizes. We have seen the same

phenomenon in the Information matrices for the simulated data discussed in Sect. 3.1 above. So we focus next on the Fisher Information matrices and parameter standard errors for the fitted models, expressed for the transformed parameters (which are subvectors of) $\vartheta = (\text{logit}(p), \log(\mu), \log(\lambda_1), \log(\lambda_2), \log(\beta_1), \log(b_C), \log(b_D))$. For models with respectively 2, 4, and 6 parameters, the ranges of eigenvalues of the respective observed information matrices $\hat{I}(\hat{\vartheta})$ were found to be $(1313, 15686)$, $(46, 28882)$, and $(58, 22387)$. Thus, in all of the models the most accurate linear parameter-combinations with unit-vector coefficients have SE's of order 0.007, while the least accurate for the 4- and 6-dimensional models have SE's of 0.13 or larger. For example, the three models give estimated SE's for $\log(\mu)$ respectively 0.010, 0.094, and 0.072; and SE's for $\log(b_C)$ are respectively 0.027, 0.057, and 0.077.

While the phase-type models fitted to the large SEER dataset have strikingly ill-conditioned Fisher information matrices—and therefore at least some parameters which are very badly identified—one can with some assurance achieve the qualitatively important finding, that at least two mixture components are needed for a high-quality parametric fit. The fact that in these data the ER-status now represents a medically *observable* identifier of two distinct mixture components (which is essentially the point of the Anderson et al. (2006) article) corroborates this conclusion, and suggests the potential usefulness in new applications of a similar parametric statistical approach in detecting the presence of two separate diseases within a single diagnostic category.

5 Computational experience with the EM approach

In this study, our main estimation method has been direct quasi-Newton-Raphson optimization of an accurately calculated log-likelihood. This method was applicable because of the relative simplicity of `Model F`, where paths do not connect except at the $O, D, 1,$ and $k_1 + 1$ states. Table 1 illustrated the need for large sample size to estimate all parameters accurately. More numerical studies of performance of ML estimates in Monte Carlo simulations can be found in the Supplement to this paper (Slud and Suntornchost 2013). The Supplement also displays histograms allowing the reader to assess the rather slow rate of convergence of distributions of ML estimators to normality as sample sizes get large.

The most common method of estimation of parameters applicable in principle to general phase-type models, follows an EM approach introduced by Asmussen et al. (1996). The idea of the EM method is first to write down the log-likelihood function for the complete observations, i.e., the absorption-time dataset augmented as though all of the intermediate transition times had also been observed. This log-likelihood, as a function of the free parameter ϑ, is then replaced (the *E-step*) by its conditional expectation given the actually observed data, taken with respect to a hypothetical fixed parameter vector ϑ_k. Then the conditional expected log-likelihood given observed data is maximized over ϑ (the *M-step*), yielding the next iteration ϑ_{k+1} in the estimated-parameter sequence. The E and M steps are repeated until the sequence ϑ_k converges. The calculation of conditional expectations in the E-step is performed in the phase-type model by setting up a system of differential equations related to the intensity matrix,

for the unknown transition-intensity parameters, and these equations are solved by a Runge-Kutta method. Details of this EM implementation can be found in Asmussen et al. (1996) and the documentation of an associated C-language program, `EMpht`, in Olsson (1998). The `EMpht` program did not specify convergence criteria, so following it we re-implemented the algorithm in the `R` platform using the `R`-function `rk` for the Runge-Kutta equation solver. Our convergence criteria involve smallness of changes in log-likelihood of the order of accuracy 10^{-10}. As mentioned in Asmussen et al. (1996), some drawbacks of the EM algorithm are its slow convergence rate (up to 10,000 iterations often being required for reasonable convergence), and its occasional convergence to a local maximum or saddle point. Another drawback is that the E-step calculation must be performed for each observation, which is computationally burdensome in large samples.

The Supplement (Slud and Suntornchost 2013) studies a specific application of the Asmussen et al. (1996) EM algorithm, to the phase-type model which is the mixture of $\text{Exp}(\alpha_1) * \text{Gamma}(4, \lambda_1)$ and $\text{Exp}(\alpha_2) * \text{Gamma}(2, \lambda_2)$, with sample sizes of 100 and 1,000. We found that very long CPU times are required to achieve EM convergence in the case of sample sizes as large as 1,000, even in low-dimensional parametric examples. However, the EM algorithm gave reasonable fit in the case of small sample size. The parameter estimation results and standard errors are given in the Supplement, along with an implementation of the method of Oakes (1999) to obtain the Fisher Information matrix from the EM algorithm computations.

Our overall conclusions are that the EM algorithm method of Asmussen et al. (1996) and Olsson (1996) for fitting phase-type survival densities to right-censored survival data is primarily of theoretical interest, because the method places no restriction on the complexity of the underlying Markov chain. But in practice, even when the models are very simple, simpler than `model F` of Fig. 1, the computation times are prohibitively large even for moderately large datasets, and they scale roughly proportionately to sample size.

Other EM approaches exist, for example one embodied in the `PhFit` tool based on the work of Horváth and Telek (2000). This tool enables a search for the number n of states (phases) and generally increases n when model-fitting is not satisfactory, but the search—like ours in the Breast Cancer data analysis—is not fully automatic and relies on some visual inspections of fit. As these authors explain in their paper, larger values n result in greater computation time and does not necessarily increase the quality of fit. All of these findings are consistent with ours, but in view of the ill-posed information matrices arising even in Phase-type models of moderate dimension, we would discourage the use of any automatic method which often generates highly parameterized models.

6 Summary and discussion

We have surveyed the broad field of parametric models for survival densities, from the vantage point of the special class of latent-state stochastic transition models known as Phase-type models. Our numerical illustrations and data analysis of a real breast cancer dataset show that even for relatively low-dimensional models of this type,

the Fisher Information matrices can be strikingly ill conditioned, and yet that certain parameters reflecting qualitative features of the fitted models—especially the presence or absence of extra 'paths' or mixture components—can be estimated adequately and have important interpretations. Our overall point is that visual features of survival densities may reflect important structure about underlying mechanism of transition among minimally parameterized latent states, structure with biomedical importance for the suggestion of future research directions, such as the search for multiple diseases underlying a single diagnostic category.

Parametric models built from mixtures are notoriously difficult to identify from moderate sample-size data. The consequence of this observation for Phase-type survival models is that only models with relatively simple path-structure and state descriptions can have a realistic chance of being fitted stably. For this reason, it may be misguided in biomedical applications to fit the complicated multistate phase-type models for which the EM methods of parameter estimation were devised. As a consequence, if only models at most of the order of complexity of our `Model F` are to be fitted, then direct likelihood computation methods based on simple properties of exponential variates and mixtures of their convolutions will be applicable.

The phase-type `Model F` can readily be extended to incorporate biomedical covariates into a regression for log transition rates such as $\log(\mu)$ or $\log(\lambda_1)$. Such survival regression models increase flexibility for joint models of nonhomogeneous populations, in the spirit of the threshold regression models of Lee and Whitmore (2006). Analogous regressions for Coxian parameters were found to increase the model likelihood in Faddy and McClean (1999). However, the introduction of unknown coefficients for covariates might also result in ML parameter estimates with large variances. The identification of some non-intercept regression coefficients might be strong, as we have seen for log ratios of transition-rates. The empirical and numerical study of such parametric regression models is a subject of our further research.

Acknowledgements We are grateful to Drs. Philip Rosenberg and William Anderson for an introduction to the breast cancer dataset and research questions of Anderson et al. (2006), and for their encouragement on this project.

References

Aalen OO (1995) Phase type distributions in survival analysis. Scand J Stat 22:447–463

Aalen OO, Gjessing H (2001) Understanding the shape of the hazard rate: a process point of view. Stat Sci 16:1–22

Andersen P, Borgan Ø, Gill R, Keiding N (1993) Statistical models for counting processes. Springer, New York

Anderson W, Chen B, Jatoi I, Rosenberg P (2006) Effects of estrogen receptor expression and histopathology on annual hazard rates of death from breast cancer. Breast Cancer Res Treat 100:121–126

Armitage P, Doll R (1954) The age distribution of cancer and a multistage theory of carcinogeneisis. Br J Cancer 8:1–12

Aslett LJM (2011) Inference for phase-type distributions. http://cran.r-project.org/web/packages/PhaseType/PhaseType.pdf

Aslett LJM, Wilson SP (2011) Markov Chain Monte Carlo for inference on phase-type models. Proceedings of ISI

Asmussen S (1992) Phase-type representations in random walk and queueing problems. Ann Prob 20:772–789

Asmussen S, Nerman O, Olsson M (1996) Fitting phase-type distributions via the EM algorithm. Scand J Stat 23:419–441
Ausin M, Wiper M, Lillo R (2004) Bayesian estimation for the M/G/1 queue using a phase-type approximation. J Stat Plan Inference 118:83–101
Binder D (1983) On the variances of asymptotically normal estimators from complex surveys. Int Stat Rev 51:279–292
Bladt M, Gonzalez A, Lauritzen SL (2003) The estimation of phase-type related functionals using Markov Chain Monte Carlo methods. Scand Actuar J 4:280–300
Bobbio A, Horváth A, Scarpa M, Telek M (2003) Acyclic discrete phase type distributions: properties and a parameter estimation algorithm. Perform Evaluation 54:1–32
Bobbio A, Horváth A, Telek M (2005) Matching three moments with minimal acyclic phase-type distributions. Stoch Models 21:303–326
Bongaarts J (2005) Long-range trends in adult mortality: models and projection methods. Demography 42(1):23–49
Booth H, Tickle L (2008) Mortality modelling and forecasting: a review of methods. Ann Actuar Sci 3:3–43
Bowers N, Gerber H, Hickman J, Jones D, Nesbitt C (1997) Actuarial mathematics, 2nd edn. Society of Actuaries, Schaumburg
Brillinger D (1961) A justification of some common laws of mortality. Trans Soc Actuar XIII(116–119):1961
Casale G, Zhang EZ, Smirni E (2010) Kpc-toolbox: best recipes for automatic trace fitting using markovian arrival processes. Perform Evaluation 67:873–896
Cox DR (1955) A use of complex probabilities in the theory of stochastic processes. Proc Camb Philos Soc 51:313–319
Dutang C, Goulet V, Pigeon M (2008) Actuar: an R package for actuarial science. J Stat Softw 25(7):38
Faddy M, McClean S (1999) Analysing data on lengths of stay of hospital patients using phase-type distributions. Appl Stoch Models in Bus Ind 15:311–317
Feller W (1972) Introduction to probability theory and its applications, 2nd edn. Wiley, New York
Gompertz B (1825) On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philos Trans R Soc Lond 115:513–585
Halley E (1693) An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw; with an attempt to ascertain the price of annuities upon lives. Philos Trans R Soc Lond 17:596–610
Heiligman L, Pollard J (1980) The age pattern of mortality. J Inst Actuar 107:49–175
Horváth A, Telek M (2000) Approximating heavy trailed behavior with phase-type distributions. In 3rd international conference on matrix-analytic methods in stochastic models, Leuven, Belgium
Horváth A, Telek M (2002) Phfit: a general phase-type fitting tool. Lecture notes in computer science 2324:82–91
Huzurbazar AV (1999) Flowgraph models for generalized phase type distributions with non-exponential waiting times. Scand J Stat 26:145–457
Huzurbazar AV (2005) Flowgraph models for multistate time-to-event data. Wiley, New York
Johnson MA (1993) Selecting parameters of phase distributions: combining nonlinear programming, heuristics, and erlang distributions. ORSA J Comput 5:69–83
Knudson A (1971) Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad USA 68:820–823
Koissi MC, Shapiro AF, Högnäs G (2006) Evaluating and extending the Lee-Carter model for mortality forecasting: bootstrap confidence interval. Insur Math Econ 38:1–20
Lee M-L, Whitmore G (2006) Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. Stat Sci 21:501–513
Lee R, Carter L (1992) Modeling and forecasting us mortality. J Am Stat Assoc 87:659–675
Lee S, Lin XS (2010) Modeling and evaluating insurance losses via mixtures of Erlang distributions. North Am Actuar J 14:107–130
Lin XS, Liu X (2007) Markov aging process and phase-type law of mortality. North Am Actuar J 11:92–109
Macheras P, Iliadis A (2006) Modeling in biopharmaceutics, pharmacokinetics, and pharmacodynamics. Springer, New York
Makeham WM (1860) On the law of mortality and the construction of annuity tables. J Inst Actuar Assur 8:301–310
Manton K, Stallard E (1980) A two disease model of female breast cancer: mortality in 1969 among white females in the United States. J Natl Cancer Inst 64:9–16

Marshall AH, Zenga M (2009) Simulating coxian phase-type distributions for patient survival. Int Trans Oper Res 16:213–226
McGrory C, Pettitt A, Faddy M (2009) A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. J Comput Stat Data Anal 53:4311–4321
Moolgavkar SH (2004) Fifty years of the multistage model: remarks on a landmark paper. Int J Epidemiol 33:1182–1183
Neuts MF (1975) Probability of phase type. Liber Amicorum Prof. Emeritus H. Florin, Department of Mathematics. Belgium: University of Louvain, pp. 173–206
Neuts MF (1981) Matrix geometric solutions in stochastic models: an algorithmic approach. The Johns Hopkins University Press, Baltimore
Oakes D (1999) Direct calculation of the information matrix via the EM algorithm. J R Stat Soc B 61:479–482
O'Cinneide CA (1989) On non-uniqueness of representations of phase-type distributions. Stoch Models 5:247–259
Olsson M (1996) Estimation of phase-type distributions from censored data. Scand J Stat 23:443–460
Olsson M (1998) The EMpht programme. http://home.imf.au.dk/smus/dl/EMusersguide.ps
R Development Core Team (2011) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org
Sengupta B (1989) Markov processes whose steady-state distribution is matrix-exponential with an application to the GI/G/1 queue. Adv Appl Probab 21:159–180
Singpurwalla N (1996) On competing risk and degradation processes. IMS Lecture notes: monograph series 49:229–240
Slud EV, Suntornchost J (2013) Monte Carlo studies for evaluating phase-type parameter estimation, and a Fisher information calculation, (Preprint)
Suntornchost J, Slud EV, Wei R (2011) Modeling US cause-specific mortality using an age-segmented Lee-Carter model, (Preprint)
Thümmler A, Buchholz P, Telek M (2006) A novel approach for phase-type fitting with the EM algorithm. IEEE Trans Depend Secur Comput 3:245–258