Modeling Multi-Dimensional Survival with Hazard Vector Fields

Daniel R. Corro

Abstract:

Traditional survival analysis deals with functions of one variable, "time." This paper explains the case of multiple and interacting aging metrics by introducing the notion of a hazard vector field. This approach is shown to provide a more general framework than traditional survival analysis, including the ability to model multi-dimensional censored data. A simple example illustrates how Green's Theorem in the plane applies to evaluate and even to theoretically optimize a course of action. One evident application is to the evaluation and promulgation of claim administration protocols.

Keywords: survival, vector field, hazard, gradient, line integral

Introduction

Although survival analysis has long recognized the need to account for different causes of death or failure, it recognizes only one way of measuring age. Consequently, survival functions—even "select" survival functions—are functions of one variable, typically denoted "t" and interpreted as "time". This paper explains the need to study observed lives from multiple perspectives. For example, a vehicle may burn several different types of fuel with varying and inter-related consumption patterns. The ability to determine whether a particular trip is possible and if so to find an efficient route may be best approached as a multi-dimensional problem. That is, it may not always be practical or revealing to reduce survival into functions of a single variable.

This work evolved from studying workers compensation insurance claims data and the motivation comes from that context. A quick claim resolution may not achieve a cost-effective result for either the insurer or the injured worker. A useful measure of "age" for the insurer may be the paid to date benefit cost of the claim while for the claimant the most important metric is likely his or her lost income. Traditional survival analysis can be helpful here, especially in dealing with open claims, i.e., "right-censored" data (c.f. [2], [4]). Simply taking "t" in the survival analysis models to be paid loss can yield useful reserve estimates (c.f. [4]). workers compensation claims typically involve both medical and wage replacement benefits. Each is expected to follow a distinctive payment pattern that need not be independent of the other. Indeed, that inter-relationship may prove to be a key cost driver. This paper illustrates how a multi-dimensional survival model can reveal those inter-relationships and their cost implications.

Consider, for instance, an issue from the ongoing debate over claim administration protocols. In the workers compensation context, is it better to pursue aggressive medical treatment quickly in an effort to minimize time lost from work, or is it more efficient to spend those resources another way, such as providing job retraining. Clearly the answer may vary tremendously based on the nature of the injury, the age of the worker, the applicable benefit provisions, and a myriad of other considerations.

The main conceptual result of this paper is that traditional survival analysis can be inherently limiting. This is established formally by showing that it is not always possible to define a survival function. The first section of the paper presents a generalization of the survival function to a function of several variables. Many of the basic formulas of survival analysis are readily generalized. The next section discusses censored data and shows how this can introduce new complications in the multi-dimensional context. The concept of a hazard vector field is defined and shown to provide a more general framework than traditional survival analysis. In particular, this framework is capable of dealing with multi-dimensional censored data. It is shown that the existence of a survival function conforms exactly to the "conservative force field" of classical physics. A simple example illustrates how Green's Theorem in the plane applies to comparing and even optimizing paths of action, e.g., as in evaluating claim administration protocols.

The concepts introduced in this paper may lead to the ability to help identify optimum claim handling practices. As noted in the section on further research, much additional work is required to test this approach. Some work that uses this approach to study the resolution pattern of workers compensation back strains shows some promise but is very preliminary. The examples presented here are only numeric illustrations; many have no practical application and some details are left to the reader. Those wishing additional details on the numerical examples or on the application to back strain cases may contact the author.

Section I: Basic Terminology and Notation

Let $\Re^+$ denote the set of nonnegative real numbers and $\Re^n$ denote n-dimensional space. For any $a = (a_1,...,a_n) \in \Re^n$, $\Im_a = \{(x_1,...,x_n) \,|\, x_i \geq a_i, 1 \leq i \leq n\}$; in particular, let $\Im = \Im_0$ denote the "positive quadrant." We regard $\Re^n$ as a model for "space-time" in which each coordinate represents an aging metric. The most natural case is when n=1 and the metric is time. For insurance applications, metrics to keep in mind would be cumulative payments or accumulation of some other quantity associated with claim resolution (e.g. x_1 = time from injury, x_2 = indemnity paid to date, x_3 = medical paid to date, x_4 = ALAE paid to date, etc.). We regard $\Im$ as all possible "failures" or "deaths", all of whose lives begin at the origin. More generally, $\Im_a$ represents the possible future (failure) values subsequent to attaining the point $a \in \Re^n$. Clearly $b \in \Im_a \Leftrightarrow \Im_b \subseteq \Im_a$.

Begin with a continuous probability density function (PDF) of "failures":

$$f : \Im \to \Re^+ \qquad \int_{\Im} f = 1.$$

It is natural to define a survival function as the probability of subsequent failure:

$$S : \Im \to \Re^+ \qquad S(a) = \int_{\Im_a} f = \frac{\int_{\Im_a} f}{\int_{\Im} f} \qquad a \in \Im$$

Observe that f and S uniquely determine one another; indeed, from the fundamental theorem of calculus:

$$f = (-1)^n \frac{\partial^n S}{\partial x_1 .. \partial x_n}.$$

For $b \in \Im_a$, define $f_a(b) = \frac{f(b)}{S(a)}$. This defines a PDF on $\Im_a$ in which the origin has been shifted to a and which has survival function $S_a(b) = \frac{S(b)}{S(a)}$, the conditional probability assuming survival to a.

Let X be the random variable with PDF f and sample space $\Im$. Because $\Im$ is closed under vector addition (it is an additive semigroup), it is natural to consider the expression:

$$\mu = E(X) = \sum_{x \in \Im} f(x)x$$

as a candidate "expected failure vector". More generally, for $a \in \Im$ this suggests that the expected failure vector for survival beyond a be expressed as:

$$\rho(a) = \sum_{x \in \Im_a} f_a(x)(x-a)$$

This infinite weighted sum, properly interpreted as a limit, can be found (when finite) via integration. Let $\pi_i : \Re^n \to \Re$ denote the usual coordinate projection functions and $\{\varepsilon_i = (0,\ldots 0,1,0,\ldots 0) \mid 1 \le i \le n\}$ the usual set of coordinate unit vectors. Continuity and linearity imply:

$$\begin{aligned}\rho(a) &= \sum_{i=1}^{n} \left(\int_{\Im_a} \pi_i(f_a(x)(x-a)) \right) \varepsilon_i \\ &= \frac{1}{S(a)} \sum_{i=1}^{n} \left(\int_{a_1}^{\infty} \cdots \int_{a_n}^{\infty} f(x_1,\ldots, x_n)(x_i - a_i)\,dx_1 \ldots dx_n \right) \varepsilon_i\end{aligned}$$

The following integration result is a straightforward integration by parts:

Lemma: For any continuous function $g : \Re^+ \to \Re^+$, $b \in \Re^+$ with $\int_0^{\infty} g(t)dt < \infty$

$$\int_b^{\infty} (t-b)g(t)dt = \int_b^{\infty}\int_t^{\infty} g(w)dwdt$$

Proof: Let $b < c \in \Re^+$. Then we have:

$$\int_b^c\int_t^c g(w)dwdt = \int_b^c u\, dv$$

$$u(t) = \int_t^c g(w)dw \qquad v = t - b$$

$$du = -g(t)dt \qquad dv = dt$$

$$= [uv]_b^c - \int_b^c v\, du = \left[(t-b)\int_t^c g(w)dw\right]_b^c + \int_b^c (t-b)g(t)\, dt$$

$$= \int_b^c (t-b)g(t)\, dt$$

and the lemma follows by letting $c \to \infty$.

Define n functions:

$$g_1(t) = \int_{a_2}^{\infty} \dots \int_{a_n}^{\infty} f(t, x_2, \dots, x_n) dx_2 \dots dx_n$$

....

$$g_n(t) = \int_{a_1}^{\infty} \dots \int_{a_{n-1}}^{\infty} f(x_1, \dots, x_{n-1}, t) dx_1 \dots dx_{n-1}$$

Invoking the above lemma and rearranging the order of integration (Fubini's Theorem):

$$\rho(a) = \frac{1}{S(a)} \sum_{i=1}^{n} \left(\int_{a_i}^{\infty} (t - a_i) g_i(t) dt \right) \varepsilon_i$$

$$= \frac{1}{S(a)} \sum_{i=1}^{n} \left(\int_{a_i}^{\infty} \int_{t}^{\infty} g_i(x_i) dx_i dt \right) \varepsilon_i$$

$$= \frac{1}{S(a)} \sum_{i=1}^{n} \left(\int_0^{\infty} S(a + t\varepsilon_i) dt \right) \varepsilon_i = \sum_{i=1}^{n} \left(\int_0^{\infty} S_a(a + t\varepsilon_i) dt \right) \varepsilon_i,$$

which implies that this candidate for expected survival vector can be determined from conditional survival parallel to the coordinate axes. Note that $\rho : \Im \to \Im$ is a vector field and that:

$$\mu = \rho(0) = \sum_{i=1}^{n}\left(\int_0^{\infty} S(t\varepsilon_i)dt\right)\varepsilon_i$$

Recall that for n=1 the hazard function h can be defined as $h(t) = \dfrac{f(t)}{S(t)}$ or equivalently as $h = -\dfrac{d(\ln(S(t)))}{dt}$. While the first definition readily generalizes to define the *hazard function* $h = \dfrac{f}{S} : \Im \to \Re^{+}$ for any n, it is the second that is of greater interest. Given a survival function S on $\Im$ the corresponding *hazard vector field* is defined as:

$$\eta = \eta_S : \Im \to \Im \qquad \eta(x) = -\nabla(\ln(S(x))), \quad x \in \Im$$

where ∇ denotes the gradient operator.

For any $a \in \Im$, a *life path of a* is a continuous function $C:[0,1] \to \Im$ satisfying:

$$C(0) = 0$$
$$C(1) = a$$
$$0 \le t \le u \le 1 \Rightarrow C(u) \in \Im_{C(t)}$$

The latter condition simply means that the path progresses into the future. Note that for any $a \in \Im$ and life path C of a, we have:

$$\oint_C \eta = -\ln(S(C(1))) + \ln(S(C(0))) = -\ln(S(a)) \quad \Rightarrow \quad S(a) = e^{-\oint_C \eta}$$

We will, as is often done, occasionally confuse a path C with its image $\{C(t)\}$, implicitly exploiting independence of the line integral to path parameterization.

The traditional language of life contingencies refers to hazard as a "force of mortality". Of course, "force" is inherently a vector concept and the latter expression relates the force of failure η with the probability of survival S. This has a natural appeal as it relates survival to the amount of "work" done traversing a hazardous life path. It gives the term "life work" a new twist and suggests an almost Aristotelian concept of life-giving energy. The existence of a survival function, as defined here, corresponds to the case when the amount of work is independent of the path, analogous to a potential function measuring energy loss in classical physics.

We conclude this section with some simple examples for n=2, in which case we revert to the more conventional xy-plane notation.

Example 1: Let $(a,b) \in \Im$ be a vector in the positive quadrant. The *exponential survival function with parameter vector (a,b)* is defined as:

$$S(x,y)=e^{-ax-by} \qquad f(x,y)=abS(x,y) \quad h(x,y)=ab$$

Note that this models the case of constant expected survival, $\rho(x,y)=(\frac{1}{a},\frac{1}{b})$, and constant hazard field $\eta(x,y)=(a,b)$.

The generalization of Example 1 to n>2 is clear. It is not surprising that the expected survival vector is constant exactly when the hazard field is constant. The inverse relationship between the two in that event, $\eta \bullet \rho = n$, has an added geometric appeal since survival is "global" while hazard is "local" (See [3] for a more systematic discussion of the relationship between hazard and expected survival.)

Example 2: After a constant vector field (Example 1), the next simplest vector field is

$$\eta(x,y)=c(x,y) \text{ for some constant } \quad c\in\Re^{+}.$$

For this case,

$$S(x,y)=e^{-c\left(\frac{x^2+y^2}{2}\right)}, \quad f(x,y)=c^2xyS(x,y), \quad h(x,y)=c^2xy$$

We leave to the reader the verification that:

$$\rho(x,y)=\left(\sqrt{\frac{2\pi e^{cx^2}}{c}}(1-\Phi(\sqrt{c}x)),\sqrt{\frac{2\pi e^{cy^2}}{c}}(1-\Phi(\sqrt{c}y))\right),$$

where $\Phi(x)=\sqrt{\frac{1}{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$ is the standard normal cumulative density function.

Example 3: Suppose γ, T, g define another hazard field, survival, and PDF, respectively. Then $\eta+\gamma$ has survival function the product ST and PDF:

$$=S(x,y)g(x,y)+T(x,y)f(x,y)+\int_x^{\infty}f(t,y)dt\int_y^{\infty}g(x,t)dt+\int_y^{\infty}f(x,t)dt\int_x^{\infty}g(t,y)dt\geq 0.$$

Combining these examples, the "first degree equation" hazard field

$$\eta(x,y)=(a,b)+c(x,y)$$

has the "second degree" survival function $S(x,y)=e^{-\left(ax+by+c\left(\frac{x^2+y^2}{2}\right)\right)}$.

When n=1, a hazard function $h(t)$ is often viewed as belonging to a one-parameter family $\{ch(t)\mid c\in\Re^{+}\}$ of "proportional hazard" functions ([2] considers the mean survival over such a family). A proportional shift $h(t) \mapsto ch(t),\ c \in \Re^+$ in the hazard function corresponds to exponentiation of the survival function $S(t) \mapsto S(t)^c$. The next example shows that this concept becomes more complicated in higher dimensions.

Example 4: The function $S(x,y) = e^{-(x+1)(y+1)}$ is a survival function with PDF $f(x,y) = (xy + x+y)S(x,y)$ and hazard field $\eta(x,y) = (y+1, x+1)$. Letting $T = \sqrt{S}$, we let the reader verify that T is **not** a survival function, as defined here, since it would have PDF:

$$e^{-(x+1)(y+1)} \cdot \left(\frac{(x+1)(y+1)}{4} - \frac{1}{2}\right) < 0$$

for (x,y) sufficiently near the origin.

Section II: Censored data and Path Dependence

To make the discussion seem more concrete, let y measure wage replacement benefits and x medical benefits awarded on a workers compensation claim. For convenience, normalize costs so that the interval [0,2] covers the range of feasible amounts. Consider the following table of survival data:

Survival Data			
x	**y**	**Status**	**Count**
1	0	Open	378
1	1	Closed	393
2	2	Closed	229
Total			1,000

In this context "failure" means claim closure, as that corresponds to the end of the life of a claim. Cases open when the information is collected are regarded as censored. The reported values of x and y represent medical and indemnity paid to date figures at that evaluation. For closed cases, the final incurred costs are reported. Consistent with the assumed unit of payment, no case survives beyond (2,2).

Let P_a^b denote the probability of survival from point a to point b. The task is to determine the probability of survival from (0,0) to the point (1,1) = $P_{(0,0)}^{(1,1)}$. Note that there are no observed closures from (0,0) to (1,0) or to (0,1), so we must have $P_{(0,0)}^{(1,0)} = P_{(0,0)}^{(0,1)} = 1$. Since there are 393 failures at (1,1) among 1000 cases, none censored at (0,1), we find that $P_{(0,1)}^{(1,1)} = \frac{1000-393}{1000} = 0.607$. Taking into account the censoring at (1,0), however, implies that $P_{(1,0)}^{(1,1)} = \frac{1000-378-393}{1000-378} = 0.368$. Since $P_{(0,0)}^{(0,1)} P_{(0,1)}^{(1,1)} = 0.607 > 0.368 = P_{(0,0)}^{(1,0)} P_{(1,0)}^{(1,1)}$, this illustrates how censored data leads to a problem determining a probability of survival $S(1,1)$ from (0,0) to (1,1).

The component functions of a hazard vector field η determined from a survival function S are readily expressed in terms of S and the PDF f. For example when n=2 we have:

$$\eta(x,y) = (P(x,y), Q(x,y)) = -\nabla(\ln(S(x,y))$$
$$= -\left(\frac{\partial \ln(S(x,y))}{\partial x}, \frac{\partial \ln(S(x,y))}{\partial y}\right) = \left(\frac{\int_y^\infty f(x,v)dv}{S(x,y)}, \frac{\int_x^\infty f(u,y)du}{S(x,y)}\right)$$

in which the common denominator $S(x,y)$ measures the probability of survival to (x,y) and the numerators the observed "marginal failures" subsequent to (x,y).

In the case of censored data, consider a decomposition:

$$f(x,y) = f_0(x,y) + f_1(x,y)$$

into censored and uncensored observations. Consistent with how (right) censored data is handled in survival analysis when n=1, it is natural to consider

$$\eta_1(x,y) = (P_1(x,y), Q_1(x,y)) = \left(\frac{\int_y^\infty f_1(x,v)dv}{S(x,y)}, \frac{\int_x^\infty f_1(u,y)du}{S(x,y)}\right)$$

in which the numerators measure only observed failures.

Example 5: Begin with:

$$S(x,y) = \frac{1}{(x+1)(y+1)}$$
$$f(x,y) = \left[\frac{1}{(x+1)(y+1)}\right]^2 = S(x,y)^2$$
$$\eta(x,y) = (P(x,y), Q(x,y)) = \frac{1}{2}\left(\frac{1}{(x+1)^2(y+1)}, \frac{1}{(x+1)(y+1)^2}\right)$$

and decompose $f(x,y)$ as:

$$f_0(x,y)=\frac{1+x+y}{(x+1)^3(y+1)^3}$$

$$f_1(x,y)=\frac{xy}{(x+1)^3(y+1)^3}$$

$$\eta_1(x,y)=(P_1(x,y),Q_1(x,y))=\left(\frac{\int_y^\infty f_1(x,v)dv}{S(x,y)},\frac{\int_x^\infty f_1(u,y)du}{S(x,y)}\right)$$

$$=\left(\frac{x(2y+1)}{2(x+1)^2(y+1)},\frac{(2x+1)y}{2(x+1)(y+1)^2}\right)$$

$$\Rightarrow \frac{\partial P_1}{\partial y}-\frac{\partial Q_1}{\partial x}=\left(\frac{x-y}{2}\right)f(x,y)$$

It follows that the vector field η_1 does not have a potential function and in particular does not have the form $-\nabla \ln S_1$ for any survival function S_1. This points out the need to generalize our definitions, as is done in the next section.

Section III: Definition of the Generalized Survival Model

Let $\Gamma=\{C_a \mid a\in\Im; C_a \text{ a life path for a}\}$, $\eta:\Im\to\Im$ a continuous vector field. The corresponding *generalized survival function* $S:\Gamma\to\Re^+$ is determined from

$$S(C_a)=e^{-\oint_{C_a}\eta}$$

The pair η, S is referred to as a *generalized survival model* on $\Im$.

Observe that if η, S and γ, T are generalized survival models, then so is $a\eta+b\gamma, S^aT^b$, for any $a,b\in\Re^+$. In particular, this generalized survival formulation captures situations that cannot be modeled with PDF's, from both this formal arithmetic perspective and as regards the ability to relate survival with choice of life path.

Of course, even for n=1, any continuous function $h:\Re^+\to\Re^+$ can formally define a survival function as $S(t)=e^{-\int_0^t h(w)dw}$ but setting $f(t)=h(t)S(t)$ need not yield a continuous PDF, as considered here. Indeed, $\int_0^\infty f(t)dt=1-p$ where $p=\lim_{t\to 0}S(t)$ can be greater than 0. In that case it is easy to augment $f(t)$ by a point mass of probability p

to achieve a mixed PDF based model. For $n>1$, the relationship between path dependence and the existence of a PDF based model lies somewhat deeper.

Also, for $n=1$, the hazard is interpreted as the instantaneous rate of failure. Consider now the case $n=2$. Following standard convention, express the hazard field as $\eta(x,y)=(P(x,y),Q(x,y))$ and assume also that P and Q are continuously differentiable. Note that for any $t>0$, any life path of $(a+t,b)$ passing through (a,b) has the form $C+D_t$ where C is a life path of (a,b) and $D_t(s)=(a+s,b),\ 0\le s\le t$. It follows that the conditional probability of survival from (a,b) to $(a+t,b)$ is uniquely determined as:

$$e^{-\oint_{D_t}\eta}=\frac{S(C+D)}{S(C)}=p(t)$$

The mean rate of failure $\alpha(t)$ per horizontal unit along D_t is also independent of the choice of C:

$$\alpha(t)=\frac{\left(\dfrac{S(C)-p(t)S(C)}{S(C)}\right)}{t}=\frac{1-p(t)}{t}$$

using the fact that the curve D_t is parameterized by arc length.

We are interested in the instantaneous horizontal rate of failure at (a,b), which is just the limit:

$$\alpha=\lim_{t\to 0}\alpha(t)=-\lim_{t\to 0}\frac{p(t)-p(0)}{t-0}=-\frac{dp}{dt}_{|t=0}$$

On the other hand:

$$p(t)=e^{-\oint_{D_t}\eta}\quad\Rightarrow$$

$$-\ln p(t)=\oint_{D_t}\eta=\int_{(a,b)}^{(a+t,b)}P(a+x,b)dx+Q(a+x,b)dy$$

$$=\int_0^t P(a+s,b)ds$$

since $dy=0$ along D_t. First differentiating by t and then letting $t\to 0$:

$$\frac{1}{p(t)}\left(-\frac{dp}{dt}\right)=P(a+t,b)\quad\Rightarrow\alpha=P(a,b)$$

We conclude that:

$P(a,b)$=instantaneous rate of failure per horizontal age unit at (a,b)

$Q(a,b)$=instantaneous rate of failure per vertical age unit at (a,b).

This is readily generalized to higher dimensions and provides a means to calculate the component functions of the hazard vector field. Clearly the hazard vector field determines a generalized survival function. This discussion shows the converse: a generalized survival function determines conditional probability, whence failure rates parallel to the coordinate axes, which in turn determine a hazard vector field.

For any n and $a \in \Im$, define the curve $D_{a,i,t}(s) = a + s\varepsilon_i,\ 0 \le s \le t,\ 1 \le i \le n$. The above discussion on failure rate noted that conditional survival parallel to a coordinate axis is independent of choice of path and the discussion in Section I then suggests the following definition for the expected survival vector

$$\rho(a) = \sum_{i=1}^{n} \left(\int_0^{\infty} e^{-\oint_{D_{a,i,t}} \eta} \, dt \right) \varepsilon_i \quad for \quad a \in \Im$$

When n=2 and $\eta(x,y) = (P(x,y), Q(x,y))$ the reader can readily verify that

$$\rho(a,b) = \left(\int_0^{\infty} e^{-\int_0^t P(a+s,b)ds} \, dt,\ \int_0^{\infty} e^{-\int_0^t Q(a,b+s)ds} \, dt \right)$$

$$= \left(\int_a^{\infty} e^{-\int_a^t P(s,b)ds} \, dt,\ \int_b^{\infty} e^{-\int_b^t Q(a,s)ds} \, dt \right)$$

Note that for $c>a$:

$$\int_a^{\infty} e^{-\int_a^t P(s,b)ds} \, dt = \int_a^{c} e^{-\int_a^t P(s,b)ds} \, dt + \int_c^{\infty} e^{-\int_a^t P(s,b)ds} \, dt$$

$$\le c - a + e^{-\int_a^c P(s,b)ds} \int_c^{\infty} e^{-\int_c^t P(s,b)ds} \, dt$$

$$\le c - a + \int_c^{\infty} e^{-\int_c^t P(s,b)ds} \, dt$$

$$\Rightarrow a + \int_a^{\infty} e^{-\int_a^t P(s,b)ds} \, dt \le c + \int_c^{\infty} e^{-\int_c^t P(s,b)ds} \, dt$$

$$\Rightarrow (c,b) + \rho(c,b) \in \Im_{(a,b)+\rho(a,b)}$$

and by symmetry, for $d>b$:

$$(c,d) + \rho(c,d) \in \Im_{(c,b)+\rho(c,b)} \subseteq \Im_{(a,b)+\rho(a,b)}$$

The corresponding result for $n=1$ is a special case of $n=2$ and the case $n>2$ is a straightforward induction using the case $n=2$. In general, we have:

$$b \in \Im_a \quad \Rightarrow \quad b+\rho(b) \in \Im_{a+\rho(a)}$$

This is intuitively what one would expect and has implications to the task of determining a hazard vector field approximating empirical data (c.f. [3]).

Again, the section concludes with an example:

Example 6: Consider the vector field

$$\eta(x,y)=\left(\frac{y^2}{2},x^2\right).$$

and consider the following line segment paths:

$$\begin{aligned}
&C_1 \textit{ from } (0,0) \textit{ to } (1,0)\\
&C_2 \textit{ from } (0,0) \textit{ to } (0,1)\\
&C_3 \textit{ from } (1,0) \textit{ to } (1,1)\\
&C_4 \textit{ from } (0,1) \textit{ to } (1,1)\\
&C_5 \textit{ from } (0,0) \textit{ to } (1,1)
\end{aligned}$$

Observe that, with the usual notational conventions, C_1+C_3, C_2+C_4 and C_5 can be taken as life paths for the point $(1,1)$. For example,

$$\oint_{C_1} \eta = \int_{(0,0)}^{(1,0)} \frac{y^2}{2}dx + x^2 dy \qquad y=0, dy=0$$

$$=\int_0^1 0dx + x^2(0) = 0 \quad \Rightarrow S(C_1)=1$$

The reader can readily verify the following observations:

$$S(C_2)=1$$

$$S(C_1+C_3)=\frac{1}{e}\approx 0.368$$

$$S(C_2+C_4)=S(C_5)=\sqrt{\frac{1}{e}}\approx 0.607,$$

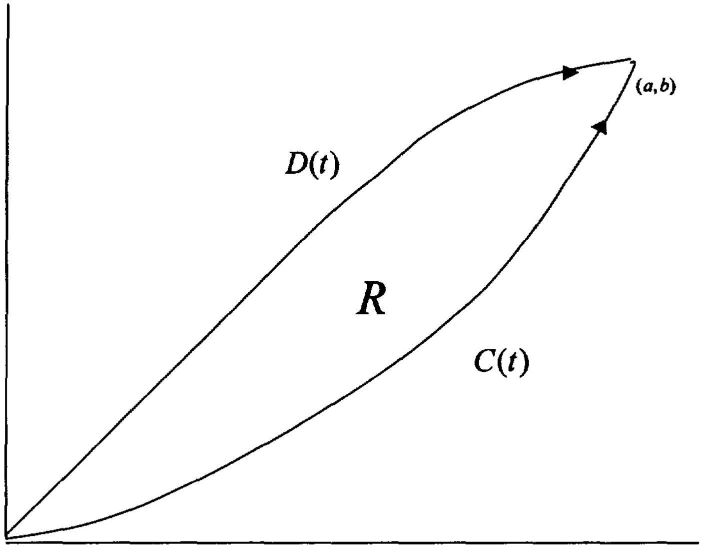
which may explain the rather odd choices for the survival data in the previous section. Note also that

$$\rho(x,y)=\left(\frac{2}{y^2},\frac{1}{x^2}\right)=\left(\frac{1}{P(x,y)},\frac{1}{Q(x,y)}\right)\Rightarrow \rho(x,y)\bullet\eta(x,y)=2.$$

Clearly, a hazard vector field $\eta(x,y)=(P(x,y),Q(x,y))$ and the corresponding expected survival vector field $\eta(x,y)$ should be "inversely related" in some sense. In this example, as in Example 1, their component functions are found to be multiplicative inverses of each other. The interested reader can verify that this is characteristic of the case when $\frac{\partial P}{\partial x}=\frac{\partial Q}{\partial y}=0$ (c.f. [3]).

Section IV: An Application of Green's Theorem in the Plane

Again consider the case n=2 and let $(a,b)\in\Im$ be a point in the positive quadrant with life paths C and D. We are interested in comparing $S(C)$ with $S(D)$. The case of most interest is when (a,b) is the "first" point beyond the origin at which the life paths meet and so assume further that C lies beneath D in the rectangle $[0,a]\times[0,b]$. The picture is:



We are interested in comparing the probabilities of failure/survival over the two paths. As in the previous section, express the hazard field as $\eta(x,y)=(P(x,y),Q(x,y))$ and assume P and Q are continuously differentiable. Under these conditions, C-D is a closed curve enclosing a simply connected region R. Green's theorem, a topic covered in most advanced calculus courses, relates the line integral over the boundary with an integral over the enclosed region. In this case, it states that:

$$\oint_C \eta - \oint_D \eta = \oint_{C-D} \eta = \iint_R \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\, dx\, dy$$

Letting $r(x, y) = \dfrac{\partial Q}{\partial x} - \dfrac{\partial P}{\partial y}$—sometimes called the "rotation" of η at (x,y)—it follows that:

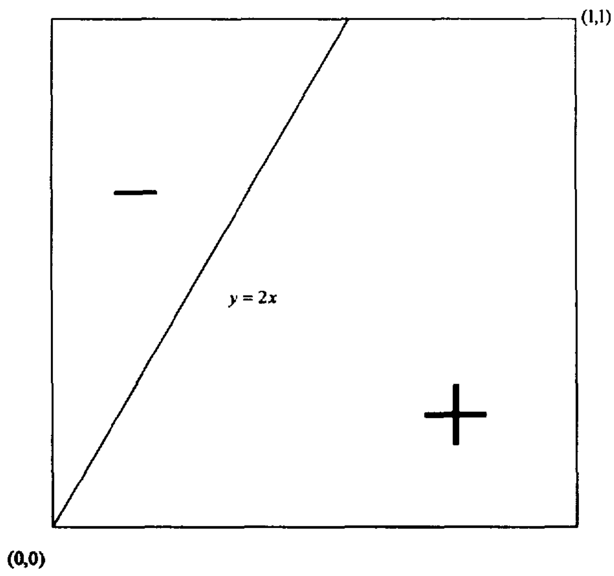
$$S(D) = e^{\alpha} S(C) \quad where \quad \alpha = \iint_R r$$

In particular,

$$r(x,y) \geq 0 \; on \; R \Rightarrow S(D) \geq S(C)$$
$$r(x,y) \leq 0 \; on \; R \Rightarrow S(D) \leq S(C)$$

with strict inequality holding when r does not vanish on R. Clearly, the function $r(x,y)$ is key to the task of identifying paths of least or greatest resistance, i.e., optimum paths for failure or survival.

Example 6 (Continued): Here $r(x, y) = 2x - y$ and as before the focus stays on survival to the point $(a, b) = (1,1)$. All life paths are contained within the unit square where the sign of r is depicted below:



The picture suggests considering the life path defined as:

$$C_6(t) = \begin{cases} (t,2t) & 0 \le t \le \frac{1}{2} \\ (t,1) & \frac{1}{2} \le t \le 1 \end{cases}$$

The reader can readily verify—directly or using Green—that:

$$\oint_{C_6} \eta = \frac{5}{12} \Rightarrow S(C_6) = e^{-\frac{5}{12}} \approx .659$$

Consider a deformation of $C_6 \mapsto \breve{C}_6$ downward that would invade the region for which r>0. Taking $\breve{C}_6 = C, C_6 = D$ in the above, we find that $S(C_6) > S(\breve{C}_6)$. On the other hand, any deformation of $C_6 \mapsto \hat{C}_6$ upward would invade the region for which r<0. Taking $\hat{C}_6 = D, C_6 = C$ in the above, we find that $S(C_6) > S(\hat{C}_6)$. It follows that the life path C_6 provides the maximum probability of survival to $(1,1)$. A similar argument shows that the life path $C_1 + C_3$ provides the minimum probability of survival to $(1,1)$. Finally, consider, as in Section II, the interpretation when values of x and y represent medical and indemnity benefits paid to date. Subject to this hazard function, the path $C_1 + C_3$ (which corresponds to the "sports medicine" approach of first focusing all resources to medical care) maximizes the probability of claim resolution at $(1,1)$.

It is apparent from the example that optimal paths can be expected to trace along solutions to r(x,y)=0 and the boundary of the rectangle. Observe that in the interpretation of Example 6, time was not included among the coordinates. Instead, time was relegated to the role of parameter of life paths. That is appropriate provided the focus is more on costs than on their specific timing. If, for example, it is desired to estimate expected time to failure, it would make sense to include time among the coordinates and look particularly at the expected survival vector component in that direction. Similarly, if the timing of payments is at issue, such as with claim administration protocols, it is natural to explicitly include time as a coordinate in the model. Given the way data is collected, time stamped payment information is the most natural source for capturing a life path and time is the most natural parameter.

Green's theorem comes neatly into play when considering alternative paths for getting to the same place, i.e., when resources are already allocated and it is a question of optimizing their effect on claim resolution. Logically prior to this, of course, is the issue of allocating resources, as illustrated in yet another revisit to the example:

Example 6 (Continued): Suppose we have fixed resources $\beta > 0$ and we consider the portion $\alpha \in [0,1]$ to be spent on medical. Clearly this involves considering life paths to the line $x + y = \beta$. So let $C_{\alpha,\beta}$ denote the linear life path from (0,0) to $(\alpha\beta,(1-\alpha)\beta)$. We leave to the reader the verification that:

$$\gamma(\alpha,\beta) = \oint_{C_{\alpha,\beta}} \eta = \frac{(\alpha - \alpha^3)\beta^3}{6}$$

It follows that for any $\beta > 0$, $\gamma(\alpha,\beta)$ has a relative maximum at $\alpha = \sqrt{\frac{1}{3}}$ and so allocating that portion of every dollar to medical would follow along the straight path

$$\left\{\left(\sqrt{\frac{1}{3}}t, \left(1-\sqrt{\frac{1}{3}}\right)t\right) \quad | \quad t \geq 0\right\}$$

that maximizes the probability of resolving the claim.

There is also the converse issue, suppose you are confronted with a claim that requires a certain amount of work to close, how can you minimize the cost outlay? This related allocation problem is illustrated in our final revisit to the example:

Example 6 (Concluded): Suppose we have a fixed amount of work $\beta > 0$ needed to close a claim and we wish to find a life path that requires the least possible total payment $x + y$. We simplify the problem and only consider straight-line solutions and let α denote the slope. Let C_a denote the linear life path from (0,0) to $(a, \alpha a)$. The reader can verify that our constraint translates into:

$$\beta = \oint_{C_a} \eta = \int_0^a \frac{\alpha^2 u^2}{2} + \alpha u^2 du = \left(\frac{\alpha^2}{2} + \alpha\right)\left(\frac{a^3}{3}\right)$$

and that the outlay $a + \alpha a$ is minimized when $\alpha = \sqrt{3} - 1$. We find that in the most cost-effective solution, the (constant) portion spent on medical $= \frac{\sqrt{3}-1}{\sqrt{3}}$ is independent of the fixed amount of work β required to close the claim.

We conclude this section with a formulation of Green's theorem suitable for comparing survival along any two life paths C and D of $(a,b) \in \mathfrak{I}$. For any $x \in [0,a]$ let $L_x = \{(x,t) \mid t \in [0,b]\}$ be the vertical line segment above x. Our assumptions imply that:

$$L_x \cap D = \{(x,t) \mid t \in [d_1(x), d_2(x)]\}$$
$$L_x \cap C = \{(x,t) \mid t \in [c_1(x), c_2(x)]\}$$

And we may define:

$$\delta(x,y) = \begin{cases} -1 & d_2(x) < c_1(x) \\ +1 & c_2(x) < d_1(x) \\ 0 & otherwise \end{cases}$$

Pictorially, δ is 1 when C lies below D and -1 when D lies below C, in effect flagging the two possible orientations the life paths can traverse around the region R they enclose. All life paths to (a,b) lie in the closed rectangle $[0,a]\times[0,b]$ and the path:

$$\hat{C}(t)=\begin{cases}(0,2bt) & 0\le t\le \frac{1}{2}\\ (2a(t-\frac{1}{2}),b) & \frac{1}{2}\le t\le 1\end{cases}$$

is the "top" top life path. Let

$$R_1 = \text{simply connected region enclosed by } C\text{-}\hat{C}$$
$$R_2 = \text{simply connected region enclosed by } D\text{-}\hat{C}$$
$$R=(R_1\cup R_2)-(R_1\cap R_2)$$

By Green's theorem:

$$\oint_{C-D}\eta = \oint_{C-\hat{C}}\eta - \oint_{D-\hat{C}}\eta = \iint_{R_1} r - \iint_{R_2} r = \iint_{R}\delta r$$

$$\Rightarrow S(D)=e^{\alpha}S(C) \quad where \quad \alpha=\iint_R \delta r$$

This provides a general comparison formula that is amenable to numeric evaluation. In practice, though, a simple chart of the sign of $r(x,y)$ over the applicable rectangle is the best starting point. The key, therefore, to identifying optimal paths is a representation of η that yields a sufficiently accurate picture of r.

Section V: Further Research

The question remains how to determine a hazard field from empirical data. One simple approach is to restrict the domain of the function to regions over which the hazard vector is assumed constant and then approximate it by estimating the coordinate failure rates. For this, traditional survival analysis methods suffice. SAS, for example, is well suited since its survival analysis procedures can be performed over cells of data and its time variable can be set to measure changes parallel to the coordinate variables (see [1]). General curve fitting techniques can then be used to paste the pieces together. Clearly a more systematic approach, especially a computer algorithm, would be useful. An alternative is to first estimate the expected survival vector field ρ—which is more straightforward in concept—and then "invert" that field in some fashion to derive the hazard vector field η (this is considered in [3]).

A generalized survival model can be used to assign a case reserve "vector". Unlike traditional reserve formulas, the vector would account for the interaction of component cost liabilities. Properly formulated, it would provide integrated benchmarks for both the prospective duration and various dollar costs of a claim. Note that the definition of expected survival vector field presented here is strictly prospective. It would be interesting to see whether the theory can yield a "tangent reserve vector" (or higher derivative vectors) defined on life paths and sensitive to the prior history of the claim.

It would also be interesting and potentially very valuable to determine whether an insurer has any tendency to follow paths of "greatest or least resistance" in resolving cases. The ability to identify optimum paths might eventually yield valuable information on protocols for case management. Example 6 is indicative of how to exploit Green's Theorem in such an investigation, not to mention first semester calculus.

Example 4 illustrates that the concept of a proportional hazard relationship becomes more complicated in higher dimensions. Indeed, the concept itself can be blown up n^2-fold from scalar to matrix multiplication. Further research is needed to determine what concepts work best. The Cox proportional hazard model (see [1]) is the standard tool for relating explanatory variables ("covariates") with the hazard function. Because each component along a life path implies essentially the same failure **sequence**, the Cox model will typically associate the same covariate proportional shift irrespective of which coordinate x_i is chosen as the time t variable. Alternatively, a parameter for the life path could be used as time t. As a result, the Cox model can be used in this context but only with the understanding that the proportional effect is assumed to be uniform over all values of all components. By the same token, so-called "time dependent" interventions can also be analyzed using the Cox model provided the intervention is consistently defined among the n components. This should not be a problem with time-stamped data where time is used to parameterize the life paths.

Of particular value would be a generalization of the Cox model approach that avoids such strong "inter-dimensional" assumptions on constant proportionality. The ideal solution would be the ability to model covariate impact on the hazard vector field via pre or post multiplication by a constant matrix. Presumably, determining the "best" such matrix would involve constructing appropriate maximum likelihood functions. The discussion in Section III, however, suggests that this may not be straightforward.

Sometimes all of $\Im$ may exceed the "natural" sample space for a particular problem. A subset (e.g. manifold as in [5]) might be more appropriate and the "Stokes type" theorems may prove useful in that context, analogous to the use of Green's theorem in the simple example discussed here. Applications of "advanced calculus" have traditionally been the purview of physicists and engineers, not actuaries. Use of multivariate survival models may help level that playing field.

References

[1] Allison, Paul D., *Survival Analysis Using the SAS® System*: A Practical Guide, The SAS Institute, Inc., 1995.

[2] Corro, Dan, *Calculating the Change in Mean Duration of a Shift in the Hazard Rate Function*, to appear in CAS Forum, Winter 2001.

[3] Corro, Dan, *A Note on the Inverse Relationship of Hazard with Life Expectancy*, in preparation.

[4] Corro, Dan, *Modeling Loss Development with Micro Data*, CAS Forum, Fall 2000.

[5] Spivak, Michael, *Calculus on Manifolds: A Modern Approach to Classical Theorems of Advanced Calculus*, Perseus Books, 1965.