Statistica Neerlandica (1994) Vol. 48, nr. 3, pp. 201–207

# Covariance formulas via marginal martingales

J. A. Wellner[1]

*Department of Statistics, GN-22*
*University of Washington*
*Seattle, WA 98195, USA*

Prenctice and Cai recently introduced and studied the function $C$ defined as the covariance function of the two marginal counting process martingales of a pair of dependent survival times $(T_1, T_2)$. They show that the function $C$ together with the marginal distributions determines the joint survival function $F$ of $(T_1, T_2)$. In this note we show how the key characterizing equation of Prentice and Cai yields a formula for the covariance of $T_1$ and $T_2$ in termsof the marginal mean residual life functions and $C$. The resulting formula generalizes a formula for the variance of a one-dimensional random variable $T$ due to Pyke (1965). We also explore several generalizations of the covariance formula, and obtain a valid $k$-dimensional version of the Prentice and Cai formula.

*Key Words & Phrases:* bivariate distribution, covariance, Hoeffding's formula, marginal martingales, mean residual life functions.

Let $T$ be a non-negative random variable with distribution function $F$: $F(t) = P(T \leq t)$. Then the survival function is $\bar{F}(t) = 1 - F(t) = P(T > t)$, and the cumulative hazard function is $\Lambda(t) = \int_{[0,t]} (1 - F_-)^{-1}\, \mathrm{d}F$. Let $N(t) = 1[T \leq t]$ be the corresponding elementary counting process, and let $M$ be the counting process martingale defined by

$$M(t) = N(t) - \int_0^t 1_{[T \geq s]}\, \mathrm{d}\Lambda(s), \qquad t \geq 0. \tag{1}$$

Define $V$ for $t \geq 0$ by

$$\begin{aligned} V(t) &\equiv \operatorname{Var}(M(t)) = E < M > (t) = \\ &= E \int_0^t 1_{[T \geq s]} (1 - \Delta\Lambda(s))\, \mathrm{d}\Lambda(s) = \int_0^t (1 - \Delta\Lambda(s))\, \mathrm{d}F(s) \end{aligned} \tag{2}$$

where $\Delta\Lambda(t) = \Lambda(t) - \Lambda(t-)$. Note that $V = F$ when $F$, and hence also $\Lambda$, is continuous. Assuming that $ET < \infty$, the *mean residual life function* $e$ of $T$ is defined by

$$e(t) = E(T - t \mid T > t) = \frac{\int_t^\infty \bar{F}(s)\, \mathrm{d}s}{\bar{F}(t)}, \qquad t \geq 0. \tag{3}$$

[1] Research supported in part by: National Science Foundation Grant DMS-9108409 and NWO Grant B 61-238.

When the distribution function $F$ of $T$ is continuous and $ET^2 < \infty$, PYKE (1965) discovered the following formula giving the variance of $T$ in terms of the mean residual life function $e$ and $V(t) \equiv \mathrm{Var}(M(t)) = F(t)$:

$$\mathrm{Var}(T) = \int e^2(t)\, \mathrm{d}V(t). \tag{4}$$

Note that $V \neq F$ when $F$ is discontinuous. The formula (4) was obtained for an arbitrary distribution function $F$ (in a somewhat different form) by HALL and WELLNER (1981). SHORACK and WELLNER (1986), page 283, generalized it to an arbitrary distribution function $F$ and an arbitrary square integrable function $a$ of $T$:

$$\mathrm{Var}[a(T)] = E[e_a^2(T)(1 - \Delta\Lambda(T))] = \int_0^\infty e_a^2(t)\, \mathrm{d}V(t) \tag{5}$$

where

$$e_a(t) = E(a(T) - a(t) \mid T > t) = -Ra(t). \tag{6}$$

Also see e.g. RITOV and WELLNER (1988), remark 2.2, page 201, or EFRON and JOHNSTONE (1990), for connections with the $R$ and $L$ operators arising in survival analysis ($e_a = -Ra$ and $L = R^T = R^{-1}$ is a martingale operator).

Now let $(T_1, T_2)$ be a pair of survival times with joint survival function $\bar{F}$: $\bar{F}(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$. Let $N_1(t_1) = 1[T_1 \le t_1]$ and $N_2(t_2) = 1[T_2 \le t_2]$ be the two marginal counting processes, and let $M_1$, $M_2$ be the marginal counting process martingales defined by

$$M_i(t_i) = N_i(t_i) - \int_0^{t} 1_{[T_i \ge s]}\, \mathrm{d}\Lambda_i(s), \qquad t_i \ge 0, \qquad i = 1, 2. \tag{7}$$

PRENTICE and CAI have recently introduced and studied the function $C$ defined by

$$C(t_1, t_2) \equiv E\{M_1(t_1) M_2(t_2)\}, \qquad t_i \ge 0, \qquad i = 1, 2. \tag{8}$$

They show that $C$, together with the marginal survival functions $\bar{F}_i(t_i) = P(T_i > t_i)$, $i = 1, 2$, determine the joint survival function $\bar{F}$ of $T_1$, $T_2$ by the following formula:

$$\bar{F}(t_1, t_2) = \bar{F}_1(t_1)\bar{F}_2(t_2)\left[1 + \int_0^{t_1}\int_0^{t_2} \frac{1}{\bar{F}_1(s_1)\bar{F}_2(s_2)}\, C(\mathrm{d}s_1, \mathrm{d}s_2)\right]; \tag{9}$$

see equation (8) of PRENTICE and CAI (1992a) or (1992b). This interesting representation of $\bar{F}$ in terms of $C$ and the marginal distributions $F_1$ and $F_2$ has applications to a wide range of bivariate estimation problems, including the bivariate random right censoring model. By use of (8) one can obtain estimators of $\bar{F}$; see e.g. PRENTICE and CAI (1992a), (1992b), and GILL, VAN DER LAAN and WELLNER (1994).

PRENTICE and CAI (1992a) also derive a $k$-dimensional extension of (9); see their formula (23), page 508. However, as pointed out by GILL (1992), their $k$-dimensional formula is incorrect: see PRENTICE and CAI (1993).

The object of this note is to show how (9) yields a formula analogous to (4) for

$\operatorname{Cov}(T_1, T_2)$ in terms of $C$ and the marginal mean residual life functions $e_i$, $i = 1, 2$, defined for $i = 1, 2$ by

$$e_i(t_i) = E(T_i - t_i \mid T_i > t_i) = \frac{\int_{t_i}^{\infty} \bar{F}_i(s_i)\,\mathrm{d}s_i}{\bar{F}_i(t_i)}, \qquad t_i \geq 0. \tag{10}$$

We will also give a bivariate analogue of (5), and a multivariate generalization thereof. The latter yields, in turn, a correct multivariate generalization of (9).

Proposition 1. *Suppose that* $(T_1, T_2)$ *have joint survival function* $\bar{F}$, *marginal survival and mean residual life functions* $\bar{F}_i$, $e_i$, $i = 1, 2$, *and that* $C$ *is defined by (8). Then, if* $\operatorname{Var}(T_i) < \infty$, $i = 1, 2$,

$$\operatorname{Cov}(T_1, T_2) = \int_0^{\infty}\int_0^{\infty} e_1(t_1) e_2(t_2) C(\mathrm{d}t_1, \mathrm{d}t_2). \tag{11}$$

*Furthermore, the bivariate analogue of (5) holds: for any measurable functions* $a_i(T_i)$, $i = 1, 2$ *with* $\operatorname{Var}(a_i(T_i)) < \infty$, $i = 1, 2$,

$$\operatorname{Cov}(a_1(T_1), a_2(T_2)) = \int_0^{\infty}\int_0^{\infty} e_{a_1}(t_1) e_{a_2}(t_2) C(\mathrm{d}t_1, \mathrm{d}t_2), \tag{12}$$

*where* $e_{a_i}$, $i = 1, 2$ *are given by (6) (with* $a, T, t$ *replaced by* $a_i, T_i, t$, $i = 1, 2$).

Proof. Our first proof of (11) and (12) will proceed directly from (9); the second proof will yield (12) directly via the martingale representations of Ritov and Wellner (1988). By a well-known formula due to Hoeffding (see e.g. Lehmann, 1966, or Shorack and Wellner, 1986, formula (A.8.2), page 862),

$$\operatorname{Cov}(T_1, T_2) = \int_0^{\infty}\int_0^{\infty} (\bar{F}(t_1, t_2) - \bar{F}_1(t_1)\bar{F}_2(t_2))\,\mathrm{d}t_1\,\mathrm{d}t_2. \tag{a}$$

Hoeffding's original version of this formula applies to distribution functions, but the current version in terms of survival functions is an elementary consequence.

It then follows immediately, by using Prentice and Cai's formula (9) to evaluate the integrand on the right side in (a), and then using Fubini's theorem, that

$$\operatorname{Cov}(T_1, T_2) = \iint \bar{F}_1)t_1)\bar{F}_2(t_2) \int_0^{t_1}\int_0^{t_2} \frac{1}{\bar{F}_1(s_1)\bar{F}_2(s_2)}\, C(\mathrm{d}s_1, \mathrm{d}s_2)\,\mathrm{d}t_1\,\mathrm{d}t_2 \tag{b}$$

$$= \iiiint \frac{1_{[s_1 \leq t_1]} 1_{[s_2 \leq t_2]} \bar{F}_1(t_1)\bar{F}_2(t_2)}{\bar{F}_1(s_1)\bar{F}_2(s_2)}\, C(\mathrm{d}s_1, \mathrm{d}s_2)\,\mathrm{d}t_1\,\mathrm{d}t_2$$

$$= \iint \frac{\int_{[s_1 < t_1]} \bar{F}_1(t_1)\,\mathrm{d}t_1}{\bar{F}_1(s_1)} \frac{\int_{[s_2 < t_2]} \bar{F}_2(t_2)\,\mathrm{d}t_2}{\bar{F}_2(s_2)}\, C(\mathrm{d}s_1, \mathrm{d}s_2)$$

$$= \iint e_1(s_1) e_2(s_2) C(\mathrm{d}s_1, \mathrm{d}s_2). \tag{c}$$

The second formula (12) follows by first observing that (9) implies that it holds for the indicator functions $a_i(T_i) = 1_{[T_i > t_i]}$, $i = 1, 2$ for any $(t_1, t_2)$, since $e_{a_t}(s) = -R_F a_t(s) = -1_{[0,t]}(s)\bar{F}(t)/\bar{F}(s)$ if $a_t(s) = 1_{[s>t]}$. By bilinearity of both sides of the equation, this yields (12) by the standard argument of obtaining measurable functions as limits of simple functions.

Here is a second proof of (12) via the martingale representations of RITOV and WELLNER (1988). By (2.8) and (2.7) of RITOV and WELLNER (1988),

$$a_1(T_i) - Ea_i(T_i) = L_i \circ R_i a_i(T_i) = \int_0^\infty R_i a_i \, dM_i, \qquad i = 1, 2 \tag{d}$$

where $L_i$ and $R_i$ are the $L$ and $R$ operators corresponding to the marginal distribution functions $F_i$, $i = 1, 2$ respectively; here $M_i$, $i = 1, 2$ are the marginal counting process martingales given by (7). Also, from formula (2.4) in RITOV and WELLNER (1988), $R_i a_i = -e_{a_i}$.

The representations given in (d) yield

$$\begin{aligned} \operatorname{Cov}(a_1(T_1), a_2(T_2)) &= E\left(\int_0^\infty R_1 a_1 \, dM_1 \int_0^\infty R_2 a_2 \, dM_2\right) \qquad \text{(e)} \\ &= \int_0^\infty \int_0^\infty R_1 a_1(t_1) R_2 a_2(t_2) C(dt_1, dt_2) \\ &= \int_0^\infty \int_0^\infty e_{a_1}(t_1) e_{a_2}(t_2) C(dt_1, dt_2) \end{aligned}$$

by application of Fubini's theorem. □

The second proof of (12) given above immediately yields the following multivariate generalization of proposition 1:

PROPOSITION 2. *Suppose that $T = (T_1, \ldots, T_k)$ has distribution function $F$ on $R^{+k}$ and that $a_i(T_i)$ are measurable functions with $E(a_i^k(T_i)) < \infty$, $i = 1, \ldots, k$. Then*

$$\begin{aligned} &E\left\{\prod_{i=1}^k (a_i(T_i) - Ea_i(T_i))\right\} \\ &= (-1)^k \int_0^\infty \ldots \int_0^\infty \left\{\prod_{i=1}^k e_{a_i}(t_i)\right\} C(dt_1, \ldots, dt_k) \end{aligned} \tag{13}$$

*where*

$$C(t_1, \ldots, t_k) \equiv E \prod_{i=1}^k M_i(t_i). \tag{14}$$

PROOF. Use the representation in (d), now for $i = 1, \ldots, k$, and proceed exactly as in (e). □

When specialized to the indicator functions $a_i(T_i) = 1_{[T_i > t_i]}$, $i = 1, \ldots, k$, (13) yields one possible $k$-variate generalization of (9) as an immediate corollary:

COROLLARY.

$$\bar{F}(t_1, \ldots, t_k) = \psi_k(t_1, \ldots, t_k) + \bar{F}_1(t_1) \ldots \bar{F}_k(t_k) \int_0^{t_1} \ldots \int_0^{t_k} \frac{1}{\bar{F}_1(s_1) \ldots \bar{F}_k(s_k)} C(\mathrm{d}s_1, \ldots, \mathrm{d}s_k) \tag{15}$$

*where*

$$\psi_k(t_1, \ldots, t_k) = E\left\{\sum_{j=1}^{k} \left(\prod_{i=1}^{j-1} 1_{[T_i > t_i]}\right) \bar{F}_j(t_j) \prod_{i=j+1}^{k} (1_{[T_i > t_i]} - \bar{F}_i(t_i))\right\} \tag{16}$$

*is a function of the $k-1$ and lower dimensional marginal survival functions of $F$.*

PROOF. First note that for $a_i(T_i) = 1_{[T_i > t_i]}$,

$$e_{a_i}(s_i) = -R_{F_i} a_i(s) = -1_{[0, t_i]}(s_i) \frac{\bar{F}_i(t_i)}{\bar{F}_i(s_i)}, \qquad i = 1, \ldots, k.$$

Hence (13) yields, with this choice of the functions $a_i$,

$$E \prod_{i=1}^{k} (1_{[T_i > t_i]} - \bar{F}_i(t_i))$$

$$= \bar{F}_1(t_1) \ldots \bar{F}_k(t_k) \int_0^{t_1} \ldots \int_0^{t_k} \frac{1}{\bar{F}_1(s_1) \ldots \bar{F}_k(s_k)} C(\mathrm{d}s_1, \mathrm{d}s_k). \tag{a}$$

Rewrite the left side of (a) as

$$\bar{F}(t_1, \ldots, t_k) + E \prod_{i=1}^{k} (1_{[T_i > t_i]} - \bar{F}_i(t_i)) - \bar{F}(t_1, \ldots, t_k).$$

Now use the formula

$$\prod_{j=1}^{k} a_j - \prod_{j=1}^{k} b_j = \sum_{j=1}^{k} \left(\prod_{i=1}^{j-1} a_i\right)(a_j - b_j)\left(\prod_{i=j+1}^{k} b_i\right)$$

to compute

$$E \prod_{i=1}^{k} (1_{[T_i > t_i]} - \bar{F}_i(t_i)) - \bar{F}(t_1, \ldots, t_k)$$

$$= -E\left\{\prod_{i=1}^{k} 1_{[T_i > t_i]} - \prod_{i=1}^{k} (1_{[T_i > t_i]} - \bar{F}_i(t_i))\right\}$$

$$= -E\left\{\sum_{j=1}^{k} \left(\prod_{i=1}^{j-1} 1_{[T_i > t_i]}\right) \bar{F}_j(t_j) \prod_{i=j+1}^{k} (1_{[T_i > t_i]} - \bar{F}_i(t_i))\right\}.$$

This yields (15). □

Here are three simple examples of (11).

EXAMPLE 1. Suppose that $F$ is Gumbel's bivariate exponential distribution, as considered by PRENTICE and CAI (1992). Thus

$$\bar{F}(t_1, t_2) = e^{-(t_1+t_2)}(1 + \theta(1 - e^{-t_1})(1 - e^{-t_2}))$$

and, as noted by PRENTICE and CAI,

$$C(t_1, t_2) = \frac{\theta}{4}(1 - e^{-2t_1})(1 - e^{-2t_2})$$

$$C(\mathrm{d}t_1, \mathrm{d}t_2) = \theta e^{-2(t_1+t_2)}\,\mathrm{d}t_1\,\mathrm{d}t_2.$$

Since the marginal distributions are both exponential, $e_1(t_1) = e_2(t_2) = 1$, and we find that

$$\int\int e_1(t_1)e_2(t_2)C(\mathrm{d}t_1, \mathrm{d}t_2) = \int_0^\infty\int_0^\infty 1\,C(\mathrm{d}t_1, \mathrm{d}t_2) = \frac{\theta}{4}.$$

On the other hand, by direct calculation the density is

$$f(t_1, t_2) = e^{-(t_1+t_2)}(1 + \theta(1 - 2e^{-t_1})(1 - 2e^{-t_2}))1_{[0,\infty)}(t_1)1_{[0,\infty)}(t_2),$$

and this yields

$$E(T_1T_2) = 1 + \frac{\theta}{4}$$

and hence $\mathrm{Cov}(T_1, T_2) = \theta/4$.

EXAMPLE 2. Suppose that $F$ is the Morgenstern copula function (distribution function on $[0, 1]^2$ with uniform marginals); thus the distribution function is

$$F(u, v) = uv(1 + \theta(1 - u)(1 - v))$$

where $|\theta| < 1$; the density function is

$$f(u, v) = 1 + \theta(1 - 2u)(1 - 2v).$$

It is easy to compute $\mathrm{Cov}(U, V) = \frac{\theta}{36}$. The marginal distributions are uniform $(0, 1)$, and

$$e_1(u) = e_2(u) = \frac{1}{2}(1 - u).$$

Straightforward calculation yields $C(\mathrm{d}u, \mathrm{d}v) = \theta(1 - u)(1 - v)\,\mathrm{d}u\,\mathrm{d}v$, and hence

$$\begin{aligned}
&\int_0^1\int_0^1 e_1(u)e_2(v)C(\mathrm{d}u, \mathrm{d}v)\\
&= \frac{1}{4}\int_0^1\int_0^1 (1 - u)(1 - v)\theta(1 - u)(1 - v)\,\mathrm{d}u\,\mathrm{d}v\\
&= \frac{\theta}{4}\left(\int_0^1 (1 - u)^2\,\mathrm{d}u\right)^2 = \frac{\theta}{36}.
\end{aligned}$$

Example 3. Suppose that $F$ puts mass $1/4$ at each of the four points $(1,2)$, $(2,1)$, $(2,2)$, and $(3,3)$. Then $ET_1 = 2 = ET_2$, $\mathrm{Var}(T_1) = 1/2 = \mathrm{Var}(T_2)$, $E(T_1 T_2) = 17/4$, and $\mathrm{Cov}(T_1, T_2) = 1/4$.

In this case the two mean residual life functions are given by

$$e_i(t) = (2-t)1_{[0,1)}(t) + (7/3 - t)1_{[1,2)}(t) + (3-t)1_{[2,3]}(t), \qquad i = 1,2.$$

so that $e_i(1) = 1$, $e_i(2) = 4/3$, $i = 1,2$. The function $C$ has jumps at the nine points on the lattice determined by the marginal distributions, and we calculate $C(\Delta 1, \Delta 1) = -1/16$, $C(\Delta 1, \Delta 2) = 1/12 = C(\Delta 1, \Delta 2)$, $C(\Delta 2, \Delta 1) = 5/36$.

Note that the other 5 points (with at least one coordinate equal to 3) do not affect the outcome of our calculation since $e_1(3) = e_2(3) = 0$. Thus we calculate

$$\begin{aligned} &\int_0^\infty \int_0^\infty e_1(t_1) e_2(t_2) C(\mathrm{d}t_1, \mathrm{d}t_2) \\ &= \frac{4}{3}\,\frac{4}{3}\left(-\frac{1}{16}\right) + \frac{4}{3}\,1\,\frac{1}{12} + \frac{4}{3}\,1\,\frac{1}{12} + 1 \cdot 1\,\frac{5}{36} \\ &= \frac{1}{4} = \mathrm{Cov}(T_1, T_2). \end{aligned}$$

**Acknowledgements**

I owe thanks to Bruce Turnbull for suggesting the relevance of Hoeffding's formula, Richard Gill for help with example 3, and a referee for help in finding the correct version of formula (16).

**References**

Efron, B. and I. Johnstone (1990), Fisher's information in terms of the hazard rate, *Annals of Statistics 18,* 38–62.

Gill, R. D., *Personal communication,* 1992.

Gill, R. D., M. J. van der Laan and J. A. Wellner (1994), Inefficient estimators of the bivariate survival function for three models to appear in: *Annales de l'Institut Henri Poincaré.*

Hall, W. J. and J. A. Wellner (1981), Mean residual life, in: M. Csorgo, D. A. Lawson, J. N. K. Rao and A. K. Md. E. Saleh (eds.) *Statistics and related topics,* 169–184, North Holland, Amsterdam.

Lehmann, E. L. (1966), Some concepts of dependence, *Annals of Mathematical Statistics 37,* 1137–1153.

Prentice, R. L. and J. Cai (1992a), Covariance and survivor function estimation using censored multivariate failure time data, *Biometrika 79,* 495–512.

Prentice, R. L. and J. Cai (1992b), Marginal and conditional models for the analysis of multivariate failure time data, in: J. P. Klein and P. K. Goel (eds.) *Survival Analysis: State of the Art,* 393–406, Kluwer, Dordrecht.

Prentice, R. L. and J. Cai (1993), Amendments and Corrections to: Covariance and survivor function estimation using censored multivariate failure time data, *Biometrika 80,* 711–712.

Pyke, R. (1965), Spacings, *Journal of the Royal Statistical Society, Series B,* 395–449.

Ritov, Y. and J. A. Wellner (1988), Censoring, martingales, and the Cox model, *Contemporary Mathematics 80,* 191–219, American Mathematical Society, Providence.

Shorack, G. R. and J. A. Wellner (1986), *Empirical processes with applications to statistics,* Wiley, New York.

Received: March 1992. Revised: September 1993.