Web-based Supplementary Materials for "Smooth" Semiparametric Regression Analysis for Arbitrarily Censored Time-to-Event Data

Min Zhang* **and Marie Davidian**

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, U.S.A.

**email:* mzhang4@stat.ncsu.edu

Web Appendix A: Properties of the SNP Density Estimator

In this appendix, we give more detail on the SNP density estimator, review work establishing its properties, and describe what is known about its performance when it has been embedded in various complex statistical models. We refer the reader to the references cited, especially Gallant and Nychka (1987) and Fenton and Gallant (1996, 1996b), for technical details and further developments.

The SNP density estimator is a truncation (or sieve) estimator based on a Hermite series expansion and was originally introduced by Gallant and Nychka (1987) in the context of representing the nonparametric part of nonlinear structural models popular in econometric analysis. These models can be rather complicated and would ordinarily also include a finite-dimensional parametric component, as in the semiparametric time-to-event regression models we consider. Since its introduction, the SNP has been used in numerous applications with great success, where it has been embedded in various complex statistical models involving possibly numerous additional parameters of interest. These include in econometric models for stock volatility (Gallant, Hansen, and Tauchen, 1990), as a model for a bivariate distribution in binary choice models for labor-force participation (Gabler, Laisney, and Lechner, 1993), as the underpinning of methods for nonlinear time series analysis (Gallant and Tauchen, 1990; Gallant, Rossi, and Tauchen, 1993), and as a representation of the density of a vector of random effects in various mixed effects (e.g., Davidian and Gallant, 1992, 1993; Zhang and Davidian, 2001; Chen, Zhang, and Davidian, 2002) and joint longitudinal-survival data models (Song, Davidian, and Tsiatis, 2002). In all of these settings, empirical studies suggest that, via a maximum likelihood approach analogous to that proposed in the main paper for the semiparametric time-to-event regression, fitting is computationally stable and feasible and valid inferences may be obtained, as discussed further below.

Gallant and Nychka (1987) considered the general case of a k-variate density in statistical models where both the density and a finite-dimensional vector of parameters are to be estimated. They described the class $\mathcal{H}$ in which the true density f_0 is assumed to lie in terms of a weighted Sobolev norm, depending on the number of derivatives f_0 is assumed to possess, and they provided a rigorous statement of the conditions under which the SNP estimators for f_0 and other parameters should be consistent in some sense for the true values, assuming the parametric part of the model is correctly specified. In particular, they showed that, as long as the truncation rule (choice of K) is such that $K = K_n$, say, converges to infinity with n, the SNP density estimator is consistent with respect to Sobolev norm and that this implies that functionals of the true density, such as the distribution function, as well as the finite-dimensional parameters in the model, are also estimated consistently. See Gallant and Nychka (1987) for technical details and discussion and Davidian and Gallant (1993, Section 3) for a summary. From a practical point of view, a main consideration in the use of SNP as a representation for the density of a model component is the degree of smoothness the true density is thought to enjoy as reflected by the degree of differentiability it is thought to possess, as for other density estimation methods.

Estimation of f_0 in the case $k = 1$, of interest in the main paper, has been studied in some detail. Fenton and Gallant (1996) specialized the consistency results of Gallant and Nychka

(1987) to the univariate case when estimation of f_0 is to be based on an iid sample from f_0, and they carried out an extensive battery of empirical studies demonstrating the ability of the SNP density estimator to approximate a wide range of true densities, including some exhibiting rather extreme behavior. They and other authors mentioned below focused on the estimator based on the normal base density, as it has been used extensively in econometric applications. They noted that, for $k = 1$, the class $\mathcal{H}$ of densities defined by Gallant and Nychka is spanned by

$$\mathcal{H}_n = \left\{ f_n : f_n(z, \boldsymbol{a}) = \left(\sum_{j=0}^{K_n} a_j z^j \right)^2 e^{-z^2/2} + \epsilon_0 \varphi(z) \right\}, \tag{A.1}$$

where $\varphi(z)$ is the standard normal density as in the main paper, and $\boldsymbol{a}$ are such that $\int f_n(z, \boldsymbol{a})\, dz = 1$; choices other than $e^{-z^2/2}$ and $\varphi(z)$ are also permitted, as would be the case in the main paper. In (A.1), ϵ_0 is a small positive number, and K_n depends on n; it is possible to rewrite (A.1) in terms of Hermite polynomials. As discussed by Gallant and Nychka (1987) and Davidian and Gallant (1993), the second term in (A.1) acts as a lower bound that governs tail behavior, ensuring that $\int \log f_n(z, \boldsymbol{a}) f_0(z)\, dz$ exists for all $f_n \in \mathcal{H}_n$, required in order to establish the results in Gallant and Nychka (1987); see this paper for further discussion. The lower bound is usually ignored in practice, as in the main paper, and vast empirical evidence has shown that this practice leads to reasonable results..

Fenton and Gallant (1996b) established rates of convergence in L_1 where $K_n = O(n^\alpha)$ for $\alpha > 0$. Coppejans and Gallant (2002) derived the convergence rate under the Hellinger metric and investigated the use of cross-validation as an alternative to information-criterion based selection of the truncation point. As noted by Kim (2007), an SNP estimator may not achieve the optimal convergence rate established by Stone (1990) for log-spline density estimators; however, it has several advantages, including computational ease and convenience; a straightforward means of simultaneous estimation of finite-dimensional parameters when the density is part of an overall semiparametric model; and the ability to evaluate whether or not the parametric model corresponding to the base density is sufficient to represent the data, as described in our particular context at the end of Section 3 of the main paper. Fenton and Gallant (1996b, Erratum) note that, while it is not possible to demonstrate that SNP density estimators have the same convergence rate as kernel density estimators, the extensive available empirical evidence suggests that they are qualitatively and asymptotically similar to kernel estimators.

Regarding asymptotic normality of estimators for finite-dimensional parameters and functionals in SNP-based semiparametric models, formal, theoretical results for general semiparametric models are not available. As noted by Kim (2007), this is probably because of the fact that the SNP density estimator is "parametric" for any fixed degree of truncation. There is extensive empirical evidence in different statistical models (e.g., Gallant and Tauchen, 1990; Zhang and Davidian, 2001; Song et al., 2002), as well as theoretical evidence in specific settings (e.g., Eastwood and Gallant, 1991; Fan, Zhang, and Zhang, 2001) that, if one treats the degree of truncation as fixed, so that the model involves a finite-dimensional "parameter," as proposed in the main paper, standard errors and confidence intervals may be constructed using standard parametric asymptotic theory. As shown by Eastwood and Gallant (1991) in a simpler setting, this requires that the degree of truncation be chosen adaptively; these authors show that the use of information-criterion-based (so adaptive) truncation rules, as proposed in the main paper, will result in such inferences being asymptotically correct. As noted by Coppejans and Gallant (2002) and Kim (2007), the practice of basing inferences on standard parametric large sample theory following adaptive choice of the truncation point is widely accepted to yield reasonable inferences in general problems and is standard in applications in analyses based on SNP.

In summary, two decades of experience suggest that use of SNP to represent ordinarily unspecified or latent components of general complex statistical models, as proposed for the specific case of semiparametric time-to-event regression models in the main paper, leads to reliable inferences under conditions similar to those assumed for competing approaches.

As noted in the Discussion of the main paper, a rigorous proof of the theoretical properties of the SNP approach proposed in the main paper is an open problem. We conjecture that it should be possible to prove that the SNP-based estimator for $\boldsymbol{\beta}$ is root-n consistent. For the PH and PO models, which are members of the linear transformation model class, this is true when one is completely nonparametric with respect to the unknown baseline distribution and uses nonparametric maximum likelihood to estimate it in these models. Thus, we expect that, under appropriate conditions, it is true for the SNP approach as well. Our simulation results do not contradict this supposition. We conjecture that this is also true for the AFT model, as it is possible to show such results for, e.g., rank-based methods. This model is a bit more problematic than the other two in that a fully efficient approach where one is completely nonparametric about the unknown survival distribution would require the support points of the distribution to depend on $\boldsymbol{\beta}$. We suspect that the undercoverage of Wald confidence intervals for $\boldsymbol{\beta}$ we report on in this case for smaller samples may be related to this structural phenomenon somehow.

Web Appendix B: Parametrization of the SNP Representation

In this appendix, we give a more detailed description of how the "standard" SNP density representation in Equation (1) of the main paper may be parameterized in terms of $\boldsymbol{\phi}$. See Zhang and Davidian (2001) for the general case. For fixed K and base density $\psi(z)$, the representation is

$$h_K(z) = P_K^2(z)\,\psi(z) = (a_0 + a_1 z + a_2 z^2 + \cdots + a_K z^K)^2\,\psi(z), \tag{B.1}$$

subject to constraint

$$\int (a_0 + a_1 z + a_2 z^2 + \cdots + a_K z^K)^2 \, \psi(z) \, dz = 1. \tag{B.2}$$

Let $\boldsymbol{a} = (a_0, a_1, \cdots, a_K)^T$ of length $K+1$ as in the main paper, define $\boldsymbol{w} = (1, z, z^2, \ldots, z^K)$, and define the random vector $\boldsymbol{W} = (1, Z, Z^2, \ldots, Z^K)^T$, where Z is a random variable with density $\psi(z)$. Then note that we can write the polynomial squared in (B.1) as

$$(a_0 + a_1 z + a_2 z^2 + \cdots + a_K z^K)^2 = \boldsymbol{a}^T \boldsymbol{w} \boldsymbol{w}^T \boldsymbol{a}.$$

Therefore, the constraint (B.2) is equivalent to requiring that

$$\boldsymbol{a}^T \boldsymbol{A} \boldsymbol{a} = 1, \quad \boldsymbol{A} = E(\boldsymbol{W}\boldsymbol{W}^T).$$

For $\psi(z)$ either the standard normal or exponential densities, the matrix $\boldsymbol{A}$ is known and positive definite, so that we can write $\boldsymbol{A} = \boldsymbol{B}^T \boldsymbol{B}$ for some positive definite matrix $\boldsymbol{B}$. Thus, write $\boldsymbol{a}^T \boldsymbol{A} \boldsymbol{a} = \boldsymbol{a} \boldsymbol{B}^T \boldsymbol{B} \boldsymbol{a}$, so that with $\boldsymbol{c} = \boldsymbol{B}\boldsymbol{a}$, $\boldsymbol{a}^T \boldsymbol{A} \boldsymbol{a} = \boldsymbol{c}^T \boldsymbol{c} = 1$. Thus, $\boldsymbol{c}$ lies on the unit sphere, which suggests the spherical transformation

$$\begin{aligned}
c_1 &= \sin(\phi_1), \\
c_2 &= \cos(\phi_1)\sin(\phi_2), \\
&\vdots \\
c_K &= \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{K-1})\cos(\phi_K), \\
c_{K+1} &= \cos(\phi_1)\cos(\phi_2)\cdot\cos(\phi_{K-1})\cos(\phi_K),
\end{aligned}$$

given in the Section 2 of the main paper, where $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_K)^T$, $-\pi/2 < \phi_j \leq \pi/2$, $j = 1, \ldots, K-1$, $0 \leq \phi_k \leq 2\pi$.

To demonstrate how this transformation works, we give two explicit examples. In the first example, suppose $K = 2$ and let $\phi(z)$ be the standard normal density. In this case,

$\boldsymbol{c} = (c_1, c_2, c_3)^T$, and $c_1 = \sin(\phi_1)$, $c_2 = \cos(\phi_1)\sin(\phi_2)$, $c_3 = \cos(\phi_1)\cos(\phi_2)$, so that $\boldsymbol{\phi} = (\phi_1, \phi_2)^T$. It is straightforward to show that

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{pmatrix},$$

in which case

$$\boldsymbol{B} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}^{-1} = \begin{pmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1/\sqrt{2} \end{pmatrix}.$$

Now

$$\boldsymbol{a} = \boldsymbol{B}^{-1}\boldsymbol{c} = \begin{pmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} \sin(\phi_1) \\ \cos(\phi_1)\sin(\phi_2) \\ \cos(\phi_1)\cos(\phi_2) \end{pmatrix}. \tag{B.3}$$

Thus note that we can express the polynomial in (B.1) in terms of $\boldsymbol{\phi}$ as $a_0 + a_1 z + a_2 z^2 = \boldsymbol{a}^T(1, z, z^2)^T$, where $\boldsymbol{a}$ is given in (B.3). This may be substituted in (B.1) to give the representation $h_2(z; \boldsymbol{\phi})$ in terms of $\boldsymbol{\phi}$.

As a second example, take again $K = 2$ but with $\psi(z)$ the standard exponential density. Again we have $\boldsymbol{a} = (a_0, a_1, a_2) = \boldsymbol{B}^{-1}\boldsymbol{c}$, where $\boldsymbol{c}$ is as before. It is straightforward to show that now

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 6 \\ 2 & 6 & 24 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 4 \\ 0 & 0 & 2 \end{pmatrix}, \quad \boldsymbol{B}^{-1} = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1/2 \end{pmatrix}.$$

Web Appendix C: Achieving the Global Maximum/Starting Values

In this appendix, we describe the approaches we have used successfully to obtain starting values for parameters for maximizing the SNP loglikelihood for each model (AFT, PH, or

PO) for fixed K and base density. For a given K and base density, the corresponding SNP likelihood $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$ involves the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\mu, \sigma, \boldsymbol{\phi}^T)^T$, and maximization requires starting values for all of these parameters. The SNP loglikelihood typically is quite complex and is replete with local maxima. Thus, we require a procedure that offers assurance that the global maximum has been identified. This suggests using "waves" of starting values, as has been proposed with SNP in other contexts (e.g., Gallant and Tauchen, 1990). We thus obtain different sets of starting values that hopefully traverse a likely region of the parameter space where the global maximum lies by fixing $\boldsymbol{\phi}$ at each value over a grid of possible values and then deriving corresponding starting values for the remaining parameters $(\mu, \sigma, \boldsymbol{\beta}^T)^T$ (μ, σ, β) depending on the model (PH, AFT, PO) being fitted, as we describe shortly. For each set of starting values so obtained, $\ell_K(\beta, \boldsymbol{\theta})$ is maximized. The maximizing values of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ leading to the largest value of $\ell_K(\beta, \boldsymbol{\theta})$ are assumed to yield to the global maximum and are taken to be the final estimates. Often, many of the sets of starting values will lead to the same maximized value of $\ell_K(\beta, \boldsymbol{\theta})$ and the same estimates, engendering confidence that the global maximum has indeed been identified. We have found that, although elements of $\boldsymbol{\phi}$ are restricted to certain ranges, as long as the grid of starting values is chosen as recommended, one may use unconstrained optimization of $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$ with assurance that the resulting estimates are such that $h_K(z; \boldsymbol{\phi})$ evaluated at the estimates is a valid density.

Our recommended grid points become less dense as K increases owing to the increasing computational cost of repeated maximizations. For $K = 0$, there is no $\boldsymbol{\phi}$, and starting values for $(\mu, \sigma, \boldsymbol{\beta}^T)^T$ may be found as described below, where $E(Z)$ and $\text{var}(Z)$ are known constants. Because for $K > 0$ each element of $\boldsymbol{\phi}$ must satisfy $-\pi/2 < \phi_j \leq \pi/2$, $j = 1, \ldots, K$, for $K = 1$ we choose the grid to be the 16 values in $(-1.5, -1.3, -1.1, \cdots, 1.3, 1.5)$. For $K = 2$, we fix $\boldsymbol{\phi} = (\phi_1, \phi_2)$ over 16 values of $(-1.5, -0.5, 0.5, 1.5) \times (-1.5, -0.5, 0.5, 1.5)$. We have demonstrated in our simulations and applications that choosing the grid points in this way yields reliable results (i.e., plausible estimates that appear to represent the global maximum) with feasible computation times.

Indeed, computation times are entirely manageable. For example, the typical time to fit all three models (PH, AFT, PO) to one data set with $n = 200$ and 25% right censoring using our SAS implementation, where maximizations are carried out using the SAS IML optimizer `nlpqn`, including maximization at each set of starting values for each K-base density combination for each model followed by selection of the preferred model-K-base density combination using HQ, is 100 seconds on a 1.73 GHz PC.

AFT Model

As in Equation (7) of the main paper, the AFT model is

$$\log(T_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} + e_i, \quad e_i \text{ iid.} \tag{C.1}$$

The SNP approach represents the AFT model (C.1) as

$$\log(T_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} + e_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \mu + \sigma Z_i, \tag{C.2}$$

where e_i and Z_i are iid, and the density of Z_i may be well-approximated by the two SNP formulations described in main paper. To get a rough estimate of $(\mu, \sigma, \boldsymbol{\beta})$ for each fixed $\boldsymbol{\phi}$, we pretend that the e_i follows a normal distribution and fit (C.2) using SAS `proc lifereg` to obtain estimates of $\boldsymbol{\beta}$ and the mean and variance of e_i, which we denote by $\boldsymbol{\beta}_e$, μ_e, and σ_e^2, respectively. We use $\boldsymbol{\beta}_e$ as the starting value for $\boldsymbol{\beta}$ and obtain starting values of μ and σ by solving the equations

$$\begin{aligned} \mu_e &= \mu + \sigma E(Z) \\ \sigma_e^2 &= \sigma^2 \text{var}(Z), \end{aligned}$$

for μ and σ. Here, $E(Z)$ and $\text{var}(Z)$ are functions of $\boldsymbol{\phi}$ for each K-base density combination ($K > 0$) and hence for a given $\boldsymbol{\phi}$ grid point are fixed constants. E.g., for the standard normal base density and $K = 2$, $E(Z) = 2a_0a_1 + 6a_1a_2$ and $\mathrm{var}(z) = a_0^2 + 3(2a_0a_2 + a_1^2) + 15a_2^2 - \{E(Z)\}^2$, where $\boldsymbol{a}$ is a function of $\boldsymbol{\phi}$ as in Web Appendix B and hence is fixed once $\boldsymbol{\phi}$ is fixed.

When $K = 0$, there is no $\boldsymbol{\phi}$. To obtain multiple starting values, we solve for μ and σ as above, where $E(Z)$ and $\mathrm{var}(Z)$ are known constants for both base densities. We use three sets of starting values: the solution (μ, σ) so determined, $(\mu - \sigma/2, \sigma)$, and $(\mu + \sigma/2, \sigma)$.

PH Model

To obtain a starting value for $\boldsymbol{\beta}$, we use Cox's partial likelihood method implemented in SAS `proc phreg`. The procedure `proc phreg` also gives an estimate of the baseline survival function $S_0(t)$. To obtain starting values for μ and σ for a fixed $\boldsymbol{\phi}$, we pretend that $\log(T_0)$ in Equation (2) of the main paper is normally distributed, so that T_0 is lognormal. Now $E(T_0) = \int_0^\infty S_0(t)dt$ and $E(T_0^2) = \int_0^\infty 2tS_0(t)dt$, and by substituting the estimated baseline survival function into these expressions, we obtain estimates of $E(T_0)$ and $E(T_0^2)$. This calculation is simple, as the estimated baseline survival function is a step function and thus the two integrals reduce to summations. If we denote the mean and variance of $\log(T_0)$ as μ_e and σ_e^2, using the relationships $E(T_0^m) = \exp(m\mu_e + m^2\sigma_e^2/2)$, $m = 1, 2$, we may obtain rough estimates of μ_e and σ_e^2 by solving two equations. Once these are obtained, we may proceed as described before for the AFT model to find starting values for μ and σ for each $K \geq 0$.

PO Model

Similar to the procedure for the AFT model, we first assume a parametric model for the baseline event time to estimate $\boldsymbol{\beta}$. In order to use a standard SAS procedure to fit a PO model, we exploit the fact that when the "errors" in an AFT model, e_i, $i = 1, \ldots, n$, are iid with a logistic distribution, this model is also a PO model. That is, by letting e_i in (C.2) be iid with a logistic distribution, we are equivalently specifying a PO model with baseline event time T_0 from a log-logistic distribution. Thus, we may use SAS `proc lifereg` to obtain estimates we denote as $\boldsymbol{\beta}_{aft}$, μ_l and σ_l, where the subscript "aft" indicates that the fitted coefficient is with respect to the AFT model, and subscript l indicates μ_l and σ_l are parameters characterizing a logistic distribution. Obtaining estimates of the mean and variance of e_i, denoted by μ_e and σ_e as before, is straightforward by using the relationships $\mu_e = \mu_l$ and $\sigma_e^2 = \pi^2\sigma_l^2/3$. Starting values for for μ and σ may be obtained in the same way as described previously for $K \geq 0$. As for the starting value for $\boldsymbol{\beta}$, the coefficient the coefficient corresponding to the PO model, one can easily derive that $\boldsymbol{\beta}$ is equal to $-\boldsymbol{\beta}_{aft}/\sigma_l$, and thus the obvious approach is to substitute the fitted values from `proc lifereg` into this expression.

Web Appendix D: Extension of the AFT Model to "Heteroscedastic Errors"

For transformed event-time models such as (C.1), a standard assumption is that the deviations e_i are iid, made in virtually all studies of these models (a recent exception is Huang, Ma, and Xie, 2005). Stare, Heinzl, and Harrell (2000) discuss the potential for biased inference on $\boldsymbol{\beta}$ if this is violated. The SNP approach readily handles so-called "heteroscedastic errors" and provides a mechanism for testing departures from the iid assumption, which may be difficult to detect graphically (Stare et al., 2000).

In (C.2), the SNP representation implies that $E\{\log(T_i)|\boldsymbol{X}_i\} = \{\mu + \sigma E(Z_i)\} + \boldsymbol{X}^T\boldsymbol{\beta}$ and $\text{var}\{\log(T_i)|\boldsymbol{X}_i\} = \sigma^2\text{var}(Z_i)$, where $E(Z_i)$ and $\text{var}(Z_i)$ are calculated assuming either Z or $Z^* = e^Z$ has density $h \in \mathcal{H}$, so that under a fixed K-base density combination are known functions of the corresponding $\boldsymbol{\phi}$. This suggests an equivalent formulation with "centered errors;" i.e., writing (2) in the main paper instead as $\log(T_0) = \mu + \sigma\{Z - E(Z)\}$ and again taking $e_i = \log(T_{0i})$ in (C.2) yields $\log(T_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} + \mu + \sigma\{Z_i - E(Z_i)\}$, so the mean is reparameterized as $E\{\log(T_i)|\boldsymbol{X}_i\} = \mu + \boldsymbol{X}_i^T\boldsymbol{\beta}$ while still $\text{var}\{\log(T_i)|\boldsymbol{X}_i\} = \sigma^2\text{var}(Z_i)$. Viewing $\{Z_i - E(Z_i)\}$ as a mean-zero deviation, then, permits the immediate extension of (C.2) given by

$$\log(T_i) = \boldsymbol{X}^T\boldsymbol{\beta} + e_i, \quad e_i = \mu + \sigma v(\boldsymbol{X}_i, \boldsymbol{\alpha})\{Z_i - E(Z_i)\}, \tag{C.1}$$

where $v(\boldsymbol{x}, \boldsymbol{\alpha}) > 0$ for all $\boldsymbol{x}$ is a parametric variance function such that $v(\boldsymbol{x}, \boldsymbol{\alpha}) \equiv 1$ if $\boldsymbol{x} = \boldsymbol{0}$ or $\boldsymbol{\alpha} = \boldsymbol{0}$, so that $\text{var}\{\log(T_i)|\boldsymbol{X}_i\} = \sigma^2\text{var}(Z_i)v^2(\boldsymbol{X}_i, \boldsymbol{\alpha})$. Although it may not be possible to postulate a "correct" model $v(\boldsymbol{x}, \boldsymbol{\alpha})$, a parsimonious, flexible variance function may be a useful way to capture at least the predominant features of potential heterogeneity (Carroll and Ruppert, 1988, Ch. 3). E.g., a model popular in ordinary regression for this purpose is $v(\boldsymbol{x}, \boldsymbol{\alpha}) = \exp(\boldsymbol{x}^T\boldsymbol{\alpha})$ (or similar form depending on a subset of $\boldsymbol{x}$). Again assuming Z or $Z^* = e^Z$ has density $h \in \mathcal{H}$, it is straightforward to derive SNP approximations to the conditional survival and density functions of $T|\boldsymbol{X}$ based on (C.1), as we now show.

In what follows, we present the conditional survival and density functions of T given $\boldsymbol{X}$, suppressing the subscript i. Considering the case where Z_i in (C.1) is taken to have the standard normal base density SNP representation, letting

$$r = \frac{\log(t) - \boldsymbol{X}^T\boldsymbol{\beta} - \mu}{\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})} + E(Z),$$

the conditional density and survival distribution are given by

$$\begin{aligned} f_K(t\,|\,\boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \{t\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})\}^{-1}P_K^2(r)\,\varphi(r), \\ S_K(t\,|\,\boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \int_r^\infty P_K^2(z)\,\varphi(z)\,dz. \end{aligned} \tag{C.2}$$

The integral in (C.2) may be calculated straightforwardly using the recursive formulæ given after Equation (3) in the main paper. The term $E(Z)$ may written as a function of $\boldsymbol{\phi}$ as before.

For the SNP representation using the standard exponential base density, we assume that the density of $Z^* = e^Z$ may be approximated by this representation. That is, the density of Z^* is represented as $h_K(z^*) = P_K^2(z^*)\mathcal{E}(z^*) = (a_0 + a_1 z^* + \cdots + a_K z^{*K})^2 e^{-z^*}$. Let

$$r = \exp\left\{ \frac{\log(t) - \boldsymbol{X}^T\boldsymbol{\beta} - \mu}{\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})} + E(Z) \right\}.$$

It may then be shown that the conditional density and survival function of $T|\boldsymbol{X}$ are

$$\begin{aligned} f_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \{t\sigma v(\boldsymbol{X}, \boldsymbol{\alpha})\}^{-1} r P_K^2(r)\mathcal{E}(r), \\ S_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) &= \int_r^\infty P_K^2(z)\,\mathcal{E}(z)\,dz. \end{aligned} \tag{C.3}$$

Again, the integral in (C.3) is calculable by the recursion described in the main paper. As r involves $E(Z) = E\{\log(Z^*)\}$, we present explicitly this calculation for $K = 0, 1, 2$, where as before a_0, a_1, a_2 are the coefficient in the polynomial $P_K(z)$, which are in turn expressed in terms of $\boldsymbol{\phi}$. With Euler's constant $\gamma = 0.57721566490153286060$, defining $H_1 = -\gamma$, $H_2 = 1 - \gamma$,$H_3 = 3 - 2\gamma$, $H_4 = 11 - 6\gamma$, $H_5 = 50 - 24\gamma$, we have for $K = 0$, $E(Z) = -\gamma = H_1$; for $K = 1$, $E(Z) = a_0^2 H_1 + 2a_0 a_1 H_2 + a_1^2 H_3$; and for $K = 2$, $E(Z) = a_0^2 H_1 + 2a_0 a_1 H_2 + (2a_0 a_2 + a_1^2) H_3 + 2a_1 a_2 H_4 + a_2^2 H_5$.

In fitting this model, one may include $\boldsymbol{\alpha}$ as an additional parameter to be estimated; typically, $\boldsymbol{\alpha}$ will be of low dimension (1 or 2). As noted by Stare et al. (2000), graphical displays that are standard diagnostic tools for detecting heteroscedasticity in ordinary uncensored regression (Carroll and Ruppert, 1988) can be misleading, so it is not prudent to rely on such techniques to suggest starting values. As we propose "working" variance models such as the exponential model for which $\boldsymbol{\alpha} = \mathbf{0}$ corresponds to no heterogeneity, we suggest using $\boldsymbol{\alpha} = \mathbf{0}$ as the starting value in the "wave" of fits across the grid of $\boldsymbol{\phi}$. Upon inspection of the results, a second "wave" may be undertaken using a new starting value for $\boldsymbol{\alpha}$. This process may be iterated until the analyst feels confident that the procedure has "zeroed in" on a reasonable fit.

Of course, (C.1) no longer has the usual AFT property that time is simply rescaled relative to baseline by a function of covariates. See Hsieh (1996) for an interpretation of (C.1) when $\boldsymbol{X}$ is a vector of treatment indicators and $v(\boldsymbol{x}, \boldsymbol{\alpha}) = \exp(\boldsymbol{x}^T\boldsymbol{\alpha})$, allowing different location and scale for each treatment, and the goal is to test for homogeneity of scale, corresponding here to $\boldsymbol{\alpha} = \boldsymbol{0}$. More generally, the SNP-based model offers a convenient framework for detecting heterogeneity, alerting the analyst that standard methods may be inappropriate.

We carried out a small simulation (100 data sets for each scenario) to demonstrate its value for accommodating and detecting heterogeneity of the "errors" in the AFT model using (C.1). For each data set with $n = 200$, iid Z_i were generated from the (bimodal) normal mixture $0.3\mathcal{N}(0.21, 0.36) + 0.7\mathcal{N}(-0.9, 0.36)$; X_i were generated as uniform on $(0, 1)$ as in Section 4 of the main paper; and T_i were generated from either (C.1) or (C.1) with $\mu = -0.9$ and $\beta = 2.0$, subject to independent uniform 30% right censoring. In scenario I, T_i were generated from the usual AFT model (C.1) with $\sigma = 1$, and (C.1) was fitted via SNP. Scenario II was the same as I, except we fitted (C.1) with $v(x, \alpha) = \exp(x\alpha)$. In scenarios III and IV, data were generated from (C.1) with $\sigma = 0.4$ and $v(x, \alpha) = \exp(x)$ $(\alpha = 1.0)$; (C.1) was fitted in III and (C.1) with $v(x, \alpha) = \exp(x\alpha)$ was fitted in IV. In scenario V, T_i were from (C.1) with $\sigma = 1$ and $v(x, \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 x$, $\boldsymbol{\alpha} = (0.4, 0.7)^T$, but (C.1) was fitted with $v(x, \alpha) = \exp(x\alpha)$ as in IV, so misspecifying v. In II, IV, and V, α was estimated along with β and $\boldsymbol{\theta}$. Table D.1 shows the results. I and IV show that the SNP method yields reliable performance when the correct model is fitted, while II shows that departures from homogeneity may be detected. III shows that failure to take account of heterogeneity has dire consequences, and V demonstrates that the exponential model can detect heterogeneity even if the working variance model is not of the correct functional form.

Table D.1: *Simulation results based on 100 Monte Carlo data sets under different scenarios involving possible "heteroscedastic" errors in the AFT model (C.1). Scenarios I–V are described in the text; table entries are as described in the tables in the main paper for estimation of each of the parameters* μ, β, α.

	μ (true=-0.9)				β (true=2.0)				α			
	Mean	SD	Ave SE	CP	Mean	SD	Ave SE	CP	Mean	SD	Ave SE	CP
I	-0.92	0.15	0.14	0.95	2.03	0.18	0.18	0.94				
II	-0.91	0.14	0.15	0.94	2.01	0.22	0.23	0.96	-0.03	0.18	0.17	0.95
III	-0.75	0.11	0.09	0.66	1.51	0.22	0.12	0.08				
IV	-0.90	0.06	0.07	0.96	2.01	0.15	0.16	0.93	0.97	0.15	0.16	0.94
V	-0.91	0.06	0.07	0.97	2.03	0.16	0.18	0.95	0.99	0.16	0.17	

Web Appendix E: Extension of the AFT Model to Time-Dependent Covariates

Time-to-event regression analyses involving time-dependent covariates are commonplace in practice; see Kalbfleisch and Prentice (2002, sec. 6.3) for a discussion of the care that must be taken in this setting. Due to ease of implementation, analysts routinely default to the Cox model (which no longer has proportional hazards); however, alternative models are available, but are rarely used. Cox and Oakes (1984, sec. 5.2) define an AFT model in this case, which we describe for scalar such covariate $X(t)$; see also Robins and Tsiatis (1992). For a subject with covariate $X(t)$ and event time T, the model assumes that time evolves relative to the time T_0 the subject would have had if $X(t) \equiv 0$ according to a monotone transformation $T_0 = \int_0^T \exp\{\beta X(u)\}du = \Psi\{\overline{X}(T), \beta\}$, where $\overline{X}(t) = \{X(s), 0 \leq s \leq t\}$ is the covariate history to t, assumed independent of T_0. If T_0 has survival function $S_0(t)$ with density $f_0(t)$ and hazard function $\lambda_0(t)$, it is conventional to express the model in terms of the hazard for T given the covariate history, which we denote in obvious notation as

$$\lambda\{t|\overline{X}(t)\} = \lambda_0[\Psi\{\overline{X}(t), \beta\}]\dot{\Psi}\{\overline{X}(t), \beta\} = \lambda_0\left[\int_0^t \exp\{\beta X(u)\}du\right]\exp\{\beta X(t)\}, \qquad \text{(E.1)}$$

where $\dot{\Psi}(u,\beta) = d\Psi(u,\beta)/du$. Ordinarily, $\lambda_0(t)$ is left completely unspecified (e.g., Robins and Tsiatis, 1992; Lin and Ying, 1995). If the analyst is willing to assume $f_0(t)$ is "smooth," so that it and $S_0(t)$ may be represented by SNP as in (3) or (4) of the main paper, it should be clear that the conditional hazard in (E.1) may be approximated by $\lambda_{0,K}(t;\boldsymbol{\theta}) = f_{0,K}(t;\boldsymbol{\theta})/S_{0,K}(t;\boldsymbol{\theta})$. In the case of right-censored data, then, where now L is a right-censoring time if $\Delta = 0$ and an event time if $\Delta = 1$, with iid data $\{L_i, \Delta_i, \overline{X}_i(L_i)\}$, $i =, 1, \ldots, n$, the loglikelihood for fixed K and base density, $\ell_K(\beta,\boldsymbol{\theta})$, for $(\beta,\boldsymbol{\theta})$ conditional on covariate history satisfies

$$\exp\{\ell_K(\beta,\boldsymbol{\theta})\} = \prod_{i=1}^{n} \Big(\lambda_{0,K}[\Psi\{\overline{X}_i(V_i),\beta\};\boldsymbol{\theta}]\dot{\Psi}_i\{\overline{X}_i(V_i),\beta\}\Big)^{\Delta_i} \exp\left\{-\int_0^{\Psi\{\overline{X}_i(V_i),\beta\}} \lambda_{0,K}(u;\boldsymbol{\theta})\,du\right\}. \tag{E.2}$$

Extension to multivariate $\boldsymbol{X}(t)$ and time-independent covariates $\boldsymbol{Z}$; i.e., $\Psi\{\boldsymbol{X}^T(T), \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\delta}\} = \int_0^T \exp\{\boldsymbol{X}^T(u)\boldsymbol{\beta} + \boldsymbol{Z}^T\boldsymbol{\delta}\}du$, is straightforward. A similar formulation holds for the PH model.

To illustrate the feasibility of implementing of the AFT model with time-dependent covariates using the SNP approach, we conducted a simulation with 1000 MC data sets and $n = 200$ generated to mimic a heart transplant scenario (e.g., Lin and Ying, 1995). For each i, a $U(0,600)$ waiting time W_i was generated, and T_{0i} was generated independently from a gamma distribution with shape 10 and scale 40. With $X_i(t) = 0$ for $t < W_i$ and $X_i(t) = 1$ for $t \geq W_i$, the event time T_i was computed according to the transformation $T_{0i} = \int_0^{T_i} \exp\{\beta X_i(u)\}du$ with $\beta = -1.0$ and was possibly right censored by an independently generated $U(0,600)$ censoring time, yielding about 30% censoring. Maximizing the SNP-based loglikelihood (E.2) yielded MC mean estimated β of -1.00, with MC standard deviation and average of estimated delta method standard errors both equal to 0.08, and MC coverage of the nominal 95% Wald interval for β of 93.0%.

It is worth noting that other models, e.g. for interval censored data with time dependent covariates (Sparling, Younes, and Lachin, 2006) may also be placed in the SNP framework.

Web Appendix F: More Complex Models

At the end of Section 4 of the main paper, we report on simulations involving several covariates for which the AFT model is the true model. These show that, although the estimator for $\boldsymbol{\beta}$ is approximately unbiased, the use of delta method standard errors and associated Wald confidence intervals may be suspect in some cases and that a nonparametric bootstrap may be used to compute alternative, more reliable standard errors and Wald intervals. Here, we show results of two representative, analogous simulations when the true model is the PH or PO model. Each is based on 1000 MC data sets with $n = 200$ and 25% independent uniform right censoring. In each case, $\boldsymbol{X} = (X_1, X_2, X_3)^T$ were generated as in the main paper.

In the first scenario, data were generated from the PH model in (5) of the main paper with $\boldsymbol{\beta} = (1.2, 1.0, 0.2)^T$ and with true $\lambda_0(t)$ corresponding to a gamma with with shape 2.0 and scale 6.0. In the second scenario, data were generated from the PO model in (9) of the main paper with $\boldsymbol{\beta} = (1.2, -1.0, 0.2)^T$ and with true $f_0(t)$ a log-mixture of normals $0.3\mathcal{N}(10, , 0.6) + 0.7\mathcal{N}(8, 0.6)$. Table D.2 shows the results, where the PH model was also fitted using PL, which are qualitatively similar to those with a single covariate reported in Section 4 of the main paper for these models.

As noted in Section 5.1 of the main paper, for demonstration of analysis under a more complex model in practice, we fit PH, PO, and AFT models to the CALGB 8541 data involving a linear predictor in several covariates $\boldsymbol{X}$. As the primary analysis found no difference between the high and moderate doses of CAF, with both superior to the low dose, we considered the treatment indicator $X_1 = 1$ if high-moderate dose and 0 if low dose. We also included $X_2 = 1$ if the woman was ER-positive, $= 0$ otherwise; $X_3 = 1$ if the woman was post-menopausal, $= 0$ otherwise; $X_4 =$ tumor size (cm); and $X_5 =$ number of histologically positive lymph nodes found. Letting $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)^T$, we fit the SNP-based PH, PO, and AFT models to the data from the 1429 subjects for whom all five covariates are

Table D.2: *Simulation results based on 1000 Monte Carlo data sets when the true model is PH or PO. Table entries are as described in the tables in the main paper for estimation of* $\boldsymbol{\beta}$ (3×1).

$f_0(t)$	n	Cens. rate	Method	True $\boldsymbol{\beta}$	Mean	SD	SE	CP
				PH model				
gamma	200	25%	SNP	1.2	1.24	0.31	0.30	94.7
				1.0	1.05	0.18	0.18	94.2
				0.2	0.21	0.09	0.09	94.9
			$(N_0 = 137,\ N_1 = 8,\ N_2 = 31,\ E_0 = 681,\ E_1 = 115,\ E_2 = 28)$					
			PL	1.2	1.22	0.31	0.30	94.8
				1.0	1.03	0.18	0.18	95.0
				0.2	0.21	0.09	0.09	95.4
				PO model				
log-mixture	200	25%	SNP	1.2	1.24	0.24	0.23	95.4
				-1.0	-1.02	0.27	0.27	95.3
				0.2	0.21	0.13	0.13	95.4
			$(N_0 = 5,\ N_1 = 125,\ N_2 = 850,\ E_0 = 0,\ E_1 = 0,\ E_2 = 20)$					

Table D.3: *Fits to the CALGB data. Base density-K shows the combination chosen by the HQ criterion for the indicated model, and HQ gives the value of the criterion for the preferred choice. Est denotes the estimate of the corresponding component of $\boldsymbol{\beta}$, and SE denotes either delta method (SNP) or usual (PL, likelihood) standard errors.*

Model	Method	Base density-K	HQ		X_1	X_2	X_3	X_4	X_5
PH	SNP	normal-1	10033	Est	-0.234	-0.269	-0.104	0.181	0.058
				SE	0.091	0.090	0.089	0.036	0.006
	PL			Est	-0.239	-0.271	-0.111	0.182	0.057
				SE	0.091	0.090	0.089	0.036	0.006
PO	SNP	normal-0	10016	Est	-0.303	-0.402	-0.177	0.231	0.090
				SE	0.114	0.113	0.111	0.046	0.010
AFT	SNP	normal-1	10019	Est	0.185	0.339	0.146	-0.140	-0.058
				SE	0.074	0.072	0.071	0.030	0.007
	lognormal ML			Est	0.206	0.292	0.116	-0.148	-0.057
				SE	0.076	0.074	0.073	0.031	0.007

available; for comparison, we also fit the PH model via PL using SAS `proc phreg`, and the AFT model assuming $f_0(t)$ is lognormal using SAS `proc lifereg`. The results are shown in Table D.3. Note that for the AFT model, HQ chooses the normal base density but with $K = 1$, suggesting that, if one assumes this model, the parametric lognormal model is not appropriate. Estimates and standard errors for the SNP-based (via the delta method) and traditional fits of the PH and AFT models are comparable. Looking across models, the HQ criterion indicates support for the PO model, with normal baseline density $f_0(t)$, over the other two models.

Note: References given in the main paper are not repeated in the list here.

Additional References

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* London: Chapman and Hall.

Chen, J., Zhang, D., and Davidian, M. (2002) Generalized linear mixed models with flexible distributions of random effects for longitudinal data. *Biostatistics* **3,** 347–360.

Coppejans, M. and Gallant, A. R./ (2002). Cross-validated SNP density estimates. *Journal of Econometrics* **110,** 27–65.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

Davidian, M. and Gallant, A.R. (1992) Smooth nonparametric maximum likelihood for population pharmacokinetics, with application to quinidine. *Journal of Pharmacokinetics and Biopharmaceutics* **20,** 529–556.

Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80,** 475–488.

Eastwood, B. J. and Gallant, A. R. (1991). Adaptive rules for seminonparametric estimators that achieve asymptotic normality. *Econometric Theory* **7,** 307–340.

Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics* **29,** 153–193.

Fenton, V. M. and Gallant, A. R. (1996b). Convergence rates of SNP density estimators. *Econometrica* **64,** 719–727.

Gabler, S., Laisney, F., and Lechner, M. (1993) Seminonparametric estimation of binary choice models with an application to labor-force participation. *Journal of Business and Economic Statistics* **11,** 61–80.

Gallant, A.R., Hansen, L.P., and Tauchen, G.E. (1990) Using conditional moments of asset payoffs to infer volatility of intertemporal marginal rates of substitution. *Journal of Econometrics* **45,** 141–180.

Gallant, A.R., Rossi, P.E., and Tauchen, G.E. (1993) Nonlinear dynamic structures. *Econometrica* **61,** 871–907.

Hsieh, F. (1996). Empirical process approach in a two-sample location-scale model with censored data. *Annals of Statistics* **24,** 2705–2719.

Huang, J., Ma, S., and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates.*Biometrics* **62,** 813–820.

Kim, K. I. (2007). Uniform convergence rate of the seminonparametric density estimator and testing for similarity of two unknown densities. *Econometrics Journal* **10,** 1–34.

Lin, D. Y. and Ying, Z. (1995). Semiparametric inference for the accelerated failure life model with time-dependent covariates. *Journal of Statistical Planning and Inference* **44,** 47–63.

Robins, J. and Tsiatis, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* **79,** 311–319.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). A semiparametric likelihood approach for joint modeling of longitudinal and time-to-event data. *Biometrics* **58,** 742–753.

Sparling, Y. H., Younes, N., and Lachin, J. M. (2006). Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics* **7,** 599–614.

Stare, J., Heinzl, H., and Harrell, F. (2000). On the use of Buckley and James least squares regression for survival data. In *New Approaches in Applied Statistics*, A. Ferilgoj and A. Mrvar (eds). *Metodološki zveszki* **16,** Ljubljana: Faculty of Social Sciences, pp. 125–134.

Stone, C. J. (1990). Large sample inference for log-spline models. *Annals of Statistics* **18,** 717-741.