University of Illinois Department of Economics
Spring 2017 Econ 574 Roger Koenker

Lecture 18
"Logspline Density & Hazard Estimation"

In the previous lecture we considered using smoothing splines for non-parametric regression. We found that the roughness penalty (*aka* smoothness prior) $\int (g''(x))^2 dx$ implied that solutions were cubic splines, i.e., twice continuously differentiable functions on $\Re$ or [0, 1] such that on each segment $[x_i, x_{i+1})$ we may write $s(x)$ as a cubic polynomial.

For smoothing splines we introduce knots at each x_i value and rely on the penalty to shrink the $\hat{g}(x_i) = a_i$'s toward the linear fit. An alternative to this approach is simply to delete knots and rely on the knot deletion to accomplish what the shrinkage does.

More formally, let

$$\mathcal{S}_0 = \{s : \Re \to \Re | s \in C^2, s \text{ is a cubic polynominal on the intervals } (-\infty, t_1], (t_1, t_2], \ldots, (t_k, \infty)\}$$

There are $(K+1)$ segments (intervals), at each $t_i, i = 1, \ldots, K$ we have 3 linear continuity constraints so the dimension of $\mathcal{S}_0$ is $4(K+1) - 3K = K + 4$. Define the K-dim subspace $\mathcal{S}$ of $\mathcal{S}_0$ consisting of $s \in \mathcal{S}_0$ such that the extreme intervals are linear. Now construct a basis for $\mathcal{S}$ of the form $\{1, B_1(x), \ldots, B_p(x)\}$ where $p = K - 1$.
[Require that B_1 is linear on $(-\infty, t_1)$ while $B_2, \ldots, B_p$ vanish there, and B_p is linear on (t_K, ∞) and $B_1, \ldots, B_{p-1}$ vanish there *and* B_1 has a negative slope, and B_p a positive one]
Some typical B-spline basis functions are illustrated in the following figure adapted from Hastie and Tibshirani (1990). The figure can be easily reproduced at home by the R commands:

```
library(splines)
u <- 1:1000/100
matplot(u,bs(u,knots=c(2,5,7)),type="l")
```

Now consider densities of the form,

$$f(y, \theta) = \exp\{\sum_{i=1}^{p} \theta_i B_i(y) - c(\theta)\}$$

where $c(\theta)$ is chosen so that $\int f(y, \theta) dy = 1$, i.e.,

$$c(\theta) = \log\{\int_{\Re} \exp\{\sum \theta_i B_i(y)\} dy\}.$$

These are just the densities whose logarithm is a piecewise cubic function. An important special case is the normal density which has a globally quadratic log density.

Note that $c(\theta)$ is just a constant of integration which assures that

$$\int f(y, \theta) dy = 1.$$

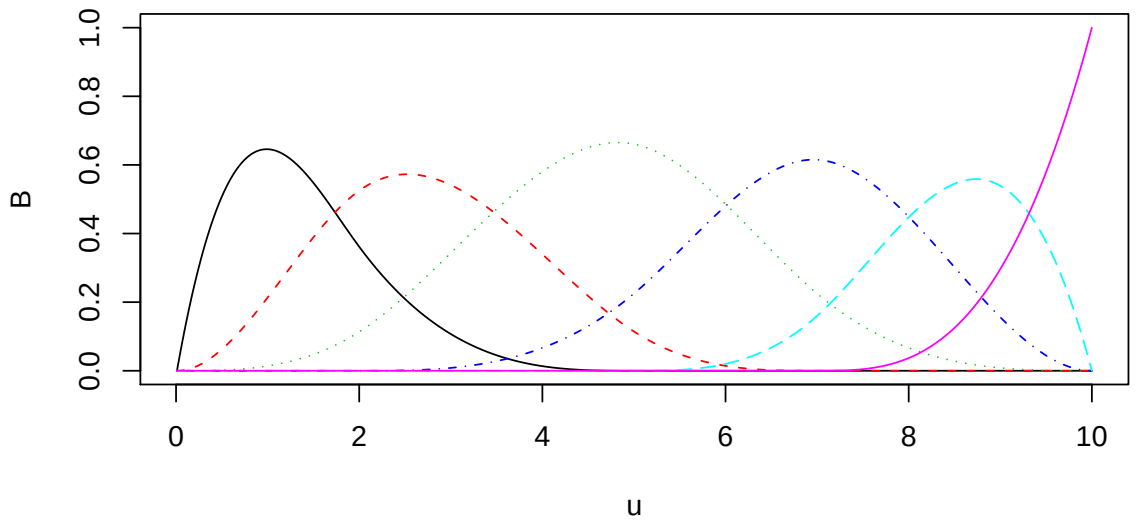

Figure 1: This figure illustrates some typical cubic B-spline basis functions. They are evaluated on an equally spaced grid from 0 to 10, interior knots are located at 2,5,7. This figure can be reproduced in R by

Proof:

$$\begin{aligned} 1 &= \int f(y,\theta)dy = \frac{\int \exp\{\sum \theta_i B_i(y)\}dy}{\exp c(\theta)} \\ &\Rightarrow \quad c(\theta) = \log(\int \exp\{\quad\}dy) \end{aligned}$$

Note that the information matrix

$$-E\nabla_\theta^2 \log f = \nabla_\theta^2 c(\theta)$$

which is positive definite so the model has a concave likelihood, hence unique mle, etc.

Given a random sample $y_1, \ldots, y_n$ from f we have sufficient statistics: $b_j = \sum B_j(Y_i)$, and we can employ the Newton method to compute mle. (Some modifications are required to insure that the slopes have the right signs in the tails). Note that the normal model is nested in this class.

The trick here is to choose the knots properly, cleverly, fortuitously. Ref [1] discusses this in detail. Akaike, Schwarz, and various other chicanery can be used.

Instead of more theory about this, I'll discuss an example. In Figure 1 I have illustrated the outcome of a contest that I ran in 478 in 1993. I generated 200 observations from a mixture of 3, 3-parameter lognormal, densities. This target density is illustrated in each panel by the dotted curve. I asked each student to write an *S*-function to compute an estimated density using whatever technique they wanted. I illustrate some of their estimates here. Most students used kernels. In the upper right panel I illustrate *my* best kernal estimate based on Hall,Sheather, Jones, and Marron's preliminary bandwidth selection and Silverman's adaptive kernel approach. In the lower left panel I illustrate the Gallant and Nychka Hermite series estimate. Obviously, all these do poorly. In contrast, Frank Shorfheide's entrant – the Stone and Kooperberg logspline method with default settings – does spectacularly well. Frank was at the time an undergraduate German exchange student, subsequently he was a graduate student at Yale and is now on the faculty at Penn.

For more details on the R implementation of this try `library(logspline); help(logspline.fit)` on ragnar. If you would like to try this yourself you can use the function `rlambda` to generate the data:

```
> rlambda <- function(n, mu = c(0.5, 1.1, 2.6), sigma = c(0.2, 0.3, 0.2),
      alpha = c(0.4, 1.2, 2.4), w = c(0.33, 0.33, 0.34))
{
#mixture of lognormals -- random numbers
#No error checking!  w is a weight vector which should add to one.
      m <- length(w)
      w <- cumsum(w)
      U <- runif(n)
      W <- matrix(0, n, m)
      W[, 1] <- U < w[1]
      for(i in 2:m) {
            W[, i] <- (U < w[i]) & (U >= w[i - 1])
      }
```

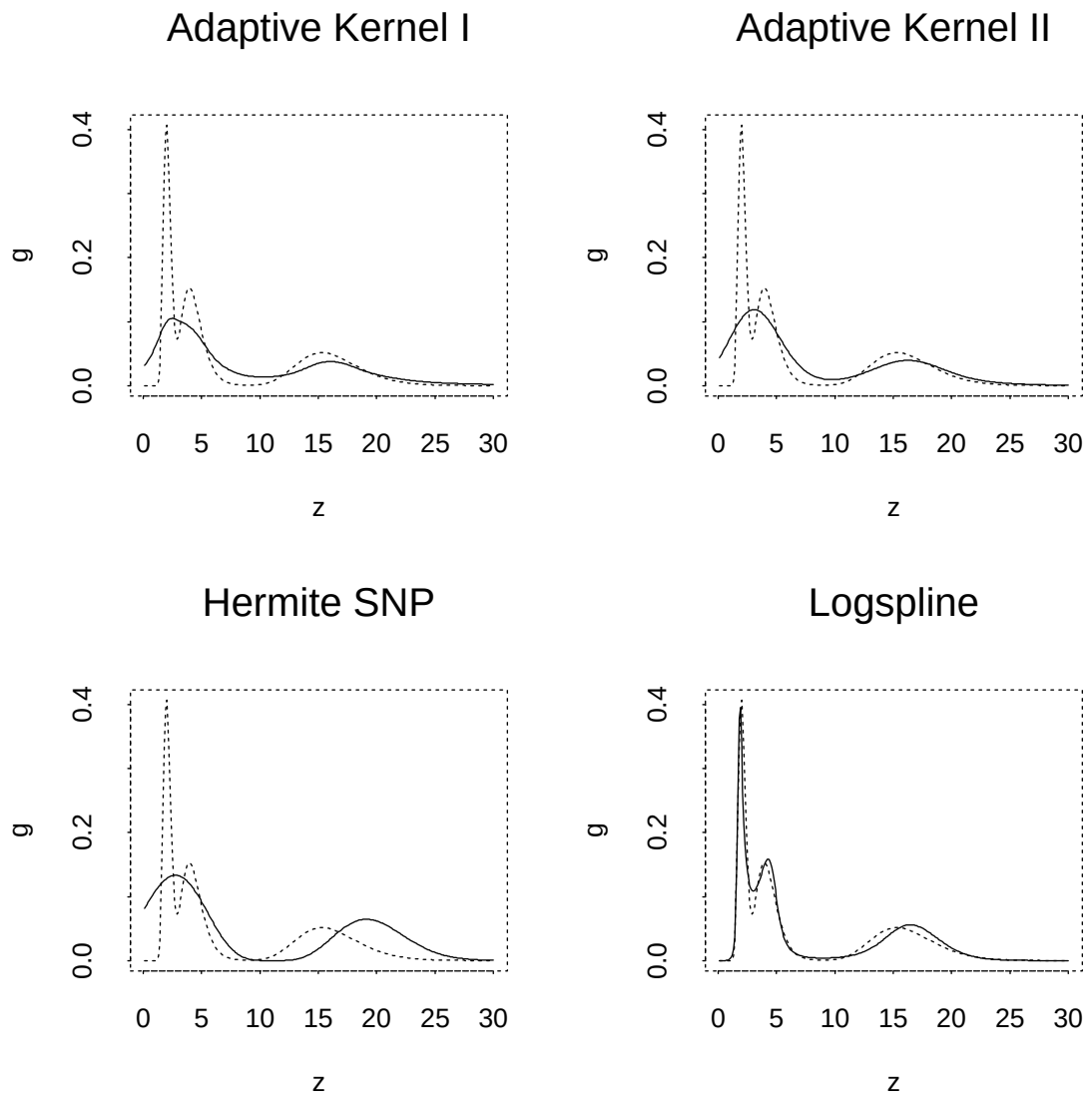

Figure 2: This figure illustrates four estimates of a mixture of lognormal density. In each panel the true density is depicted as the dotted line, and the estimate as the solid line. In the upper two panels you see two typical kernel density estimates both use the adaptive kernel method of Silverman, the former employs a naive pilot estimate and the second uses a pilot estimate based on the fixed bandwidth choice of Hall, Sheather, Jones and Marron. The third plot is the Hermite series estimate of Gallant and Nychka using their suggested default settings, and the fourth panel shows the logspline estimate of Stone and Kooperberg.

```
      z <- rep(0, n)
      for(i in 1:m) {
            z <- z + W[, i] * (alpha[i] + exp(rnorm(n, mu[i], sigma[i])))
      }
}
```

The density can be drawn using the function dlambda:

```
> dlambda <- function(z, mu = c(0.5, 1.1, 2.6),
      sigma = c(0.2, 0.3, 0.2), alpha = c(0.4, 1.2,
      2.4), w = c(0.33, 0.33, 0.34), eps = 0.0001)
{
#mixture of lognormals density function
      m <- length(w)
      f <- 0 * z
      for(i in 1:m) {
            f <- f + (w[i] * psi(log(pmax(z - alpha[i], eps)), mu[i],
                 sigma[i]))/((z - alpha[i]))
      }
}
```

Logspline Hazard Estimation

We need to recall some notation and concepts from survival analysis:

Consider $Y_1, \ldots, Y_n$ iid from F, f on $[0, \infty)$. Let

$$h = f/(1 - F)$$

denote the hazard function and

$$\lambda(\cdot) = \log h(\cdot)$$

Then

$$\begin{aligned} 1 - F(t) &= \exp\{-\int_0^t h(u)du\} = \exp\{-\int_0^t \exp(\lambda(u))du\} \\ f(t) &= \exp(\lambda(t)) \exp\{-\int_0^t \exp(\lambda(u))du\} \end{aligned}$$

and

$$\varphi(t) = \log f(t) = \lambda(t) - \int_0^t \exp\{\lambda(u)\}du$$

The idea is to adapt the approach for $\log f$ to estimating models for hazard functions. So let's write

$$\lambda(y, \theta) = \sum_{j=-1}^{p} \theta_j B_j(y)$$

where $\{B_j\}$ are again an appropriately chosen basis for the space of log hazards. To embed the problem in a familiar parametric setting we choose for some $c > 0$

$$B_{-1}(t) = \log\left(\frac{t}{t + c}\right)$$

and

$$B_0(t) = \log(t + c)$$

we can motivate these choices as follows,

1. Suppose f is Weibull

   $$f(t) = b\gamma t^{\gamma-1} \exp\{-bt^\gamma\} \qquad t > 0$$
   $$F(t) = 1 - \exp\{-bt^\gamma\}$$
   $$\lambda(t) = (\gamma - 1)\log t + \log b\gamma$$

   so for $p = 1$, so $B_1(y) \equiv 1$, we set

   $$\theta_{-1} = \theta_0 = \gamma - 1 \text{ and } \theta_1 = \log(b\gamma)$$

2. Suppose f is Pareto, i.e.,

   $$\begin{aligned} f(t) &= \frac{bc^b}{(t+c)^{b+1}} \\ F(t) &= 1 - \left(\frac{c}{t+c}\right)^b \\ \lambda(t) &= \log b - \log(t + c) \end{aligned}$$

   so $\theta_{-1} = 0, \theta_0 = -1, \theta_1 = \log(b)$ gives the Pareto case. Obviously, this requires that c is specified correctly, or estimated consistently somehow.

Principle: If you want to be non-parametric, start by nesting some reasonable standard models as the simplest special cases.

Example: For the logspline density estimator the global quadratic is the natural special case – i.e., the Gaussian model.

Estimation: Let $T_1, \ldots, T_n$ be random sample from f, F and $C_1, \ldots, C_n$ be censoring times.

$$Y_i = \min(T_i, C_i) \text{ is observed}$$

and

$$\delta_i = I(T_i < C_i) = \begin{cases} 1 & \text{if uncensored} \\ 0 & \text{if censored} \end{cases}$$

so we have data (Y_i, δ_i). Recall that the loglikelihood can be written as

$$\ell(\theta) = \sum \varphi(y_i, \delta_i, \theta)$$

where $\varphi(y_i, 1, \theta) = \log f(y_i, \theta)$ and $\varphi(y_i, 0, \theta) = \log(1 - F(y_i, \theta))$, again it can be shown that the ℓ is concave. So we have a relatively easy mle problem.

Adding Regressors To add covariates we need, simply, to make $f(t|x), h(t|x)$, etc conditional on a vector of covariates x. So now e.g.,

$$\lambda(t|x, \beta) = \sum_{j=1}^{p} \beta_j B_j(t|x)$$

and one can proceed as before if we give some structure to the basis functions $\{B_j\}$. Note if B_j's are all additively separable in t and x, we get the Cox proportional hazard (PH) model. Otherwise we get a non-PH model.

Kooperberg and Stone use linear splines and their tensor products (ineractions) to specify the $\{B_j\}$'s. So

$$K_0 \geq 1 \qquad t_k : 1 \leq k \leq K_0$$
$$B_{0k}(t) = (t_k - t)_+ \quad 1 \leq k \leq K_0$$
$$t_+ = \max(0, t).$$

$$K_m \geq -1$$
$$B_{0m}(x_m) = x_m \qquad K_m = 0$$
$$B_{km}(x_{mk}) = (x_m - x_{mk})_+ \qquad 1 \leq k \leq K_m$$

And form the basis as tensor product of these functions. This all has a nice exponential family interpretation, see Barron and Sheu (1991) for more details on the density case and Stone (1997) a complete treatment.

Kooperberg & Stone (1991). *Comp Stat & Data Anal*, pp. 327-347.

Kooperberg & Stone (1992). Logspline density estimation for censored data, *Journal of Computational and Graphical Statistics*, 301-328.

Barron & Sheu (1991). Approximation of Densities by Sequence of *exp*-families. *Annals of Stat*, 1347-1349.

Hastie and Tibshirani(1990). *Generalized Additive Models*, Chapman-Hall.

Stone, C., M.H. Hansen, C. Kooperberg and Y.K. Troung (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 Wald lecture (with discussion), *Annals of Stat*, 25, 1371-1470.