

Lab2: Using Likelihood-based cross-validation for model selection when estimating the survival function using targeted MLE - due Oct. 30 in my mailbox in Haviland

October 16, 2006

1 Introduction

Like Lab 1, this lab involves using a targeted MLE approach to estimating the survival distribution using an initial smooth density estimate. The goal is to implement, in R , an algorithm that.

- Provides an initial density estimate using kernel density estimation with bandwidth, h , $p_{n,h}^0$.
- Derive the estimate of the survival function at several (say 10) points (t) using numerical integration $S_n^0(t) = \int I(u > t)p_{n,h}^0(u)du$.
- Derive a targeted MLE, by first posing one-dimensional sub-models: $p_{n,h}^0(\epsilon)(u) = [1 + \epsilon^T D(p_{n,h}^0(u))]p_{n,h}^0(u)$ and find the MLE for ϵ , say $\epsilon_{n,h}$. Use the following closed form estimator, $\epsilon_{n,h} = \Sigma(p_{n,h}^0)^{-1}P_{n,h}D(p_{n,h}^0)$, where $\Sigma(p_{n,h}^0) = E_{p_{n,h}^0} D(p_{n,h}^0)D(p_{n,h}^0)^T$ and $D(p_{n,h}^0)(t) = I(Y > t) - S_{n,h}^0(t)$.
- Find the new targeted MLE of the survival function as: $S_{n,h}^0(\epsilon_n)(t) = \int I(u > t)p_{n,h}^0(\epsilon_n)(u)du$.

The addition to last week's lab is to use V-fold likelihood-based cross-validation to find the "optimal" bandwidth. Divide the data up into $V = 5$

parts, then for each unique (left-out) validation set, fit the same target MLE procedure on the remaining training sample for each of 10 candidate bandwidths (I suggest choosing from 0.40 to 0.90). For the left out validation sample, get the sum of the log(densities) (i.e., log-likelihood), using the density estimate from the training sample. Note, this means getting $p_{n,h}^0(Y)$ for all Y in the validation sample and then getting $p_{n,h}^0(\epsilon)(Y) = [1 + \epsilon^T D(p_{n,h}^0(Y))] p_{n,h}^0(Y)$ for the validation sample, where $D(p_{n,h}^0(Y)) = I(Y > t) - S_{n,h}^0(t)$, again for the corresponding Y in the validation sample, but with $S_{n,h}^0(t)$ estimated on the training sample. Find the bandwidth that provides the maximum log-likelihood as summed up over all the validation samples.

To turn in, show a plot of bandwidth (h) versus objective function (log-likelihood) as well as showing the original density and targeted MLE at the optimal h . In addition, provide a plot that shows the survival function at all original data points from the original smoothed density ($S_{n,h}^0(t)$), the same but based on the targeted estimate ($S_{n,h}^0(\epsilon_n)(t)$), both at the optimal bandwidth, and the empirical survival distribution. Please provide the code you used as well as a short write-up that describes the results.