Semiparametric Additive Transformation Models under General Censorship

Guang Cheng

Department of Statistics, Purdue University

ICSA 2011 Applied Statistics Symposium
June 28th, 2011
Collaborator: Xiao Wang at Purdue

Outline

Semiparametric Additive Transformation (SAT) Models

We consider the efficient estimation of the following SAT models:

$$H(U) = Z'\beta + \sum_{j=1}^{d} h_j(W_j) + \epsilon,$$

where H is a monotone transformation function, h_j's are smooth regression function (with possibly different degrees of smoothness) and ϵ has a known distribution function F.

A wide range of survival models and econometric models can be incorporated into this general transformation framework. For example,

A wide range of survival models and econometric models can be incorporated into this general transformation framework. For example,

- Let T be the event time and $U = \log T$;

A wide range of survival models and econometric models can be incorporated into this general transformation framework. For example,

- Let T be the event time and $U = \log T$;
- If we assume $F(s) = 1 - \exp(-e^s)$ and $H(u) = \log A(e^u)$, then SAT becomes the partly linear additive Cox model (Huang, 1999) and A is just the cumulative hazard function;

A wide range of survival models and econometric models can be incorporated into this general transformation framework. For example,

- Let T be the event time and $U = \log T$;
- If we assume $F(s) = 1 - \exp(-e^s)$ and $H(u) = \log A(e^u)$, then SAT becomes the partly linear additive Cox model (Huang, 1999) and A is just the cumulative hazard function;
- If we assume $F(s) = \exp(s)/(1 + \exp(s))$, then SAT becomes the partly linear additive proportional odds model.

Introduction: Mixed Case Interval Censored (MIC) Data

In view of the connection between SAT and survival models, we consider the survival data in which U cannot be observed directly.

Introduction: Mixed Case Interval Censored (MIC) Data

In view of the connection between SAT and survival models, we consider the survival data in which U cannot be observed directly.

- In Case 1 IC data (current status data), we observe V and $\Delta = 1\{U \leq V\}$;

Introduction: Mixed Case Interval Censored (MIC) Data

In view of the connection between SAT and survival models, we consider the survival data in which U cannot be observed directly.

- In Case 1 IC data (current status data), we observe V and $\Delta = 1\{U \leq V\}$;
- In Case 2 IC data, we observe (L, R), $\Delta_1 = 1\{U \leq L\}$ and $\Delta_2 = 1\{L < U \leq R\}$;

Introduction: Mixed Case Interval Censored (MIC) Data

In view of the connection between SAT and survival models, we consider the survival data in which U cannot be observed directly.

- In Case 1 IC data (current status data), we observe V and $\Delta = 1\{U \leq V\}$;
- In Case 2 IC data, we observe (L, R), $\Delta_1 = 1\{U \leq L\}$ and $\Delta_2 = 1\{L < U \leq R\}$;
- Note that the Case k IC data for $k > 2$ can be always reduced to a Case 2 IC with L and R determined by U and $(V_1, \ldots, V_k)$ jointly. Therefore, the key assumption that (L, R) is independent of U is not valid in practice.

- The above assumption violation was observed in Schick and Yu (2000). This motivates them to propose the more realistic mixed case IC data. In MIC data, we observe a vector of ordered random examination times $V_K = (V_{K,1}, \ldots, V_{K,K})$ and $\Delta_K = (\Delta_{K,1}, \ldots, \Delta_{K,K+1})$, where $\Delta_{K,j} = 1\{V_{K,j-1} < U \leq V_{K,j}\}$ (with $V_{K,0} = -\infty$ and $V_{K,K+1} = +\infty$) and K is random;

- The above assumption violation was observed in Schick and Yu (2000). This motivates them to propose the more realistic mixed case IC data. In MIC data, we observe a vector of ordered random examination times $V_K = (V_{K,1}, \ldots, V_{K,K})$ and $\Delta_K = (\Delta_{K,1}, \ldots, \Delta_{K,K+1})$, where $\Delta_{K,j} = 1\{V_{K,j-1} < U \leq V_{K,j}\}$ (with $V_{K,0} = -\infty$ and $V_{K,K+1} = +\infty$) and K is random;
- Now, it is reasonable to assume that V_K is independent of U.

- The above assumption violation was observed in Schick and Yu (2000). This motivates them to propose the more realistic mixed case IC data. In MIC data, we observe a vector of ordered random examination times $V_K = (V_{K,1}, \ldots, V_{K,K})$ and $\Delta_K = (\Delta_{K,1}, \ldots, \Delta_{K,K+1})$, where $\Delta_{K,j} = 1\{V_{K,j-1} < U \le V_{K,j}\}$ (with $V_{K,0} = -\infty$ and $V_{K,K+1} = +\infty$) and K is random;
- Now, it is reasonable to assume that V_K is independent of U.
- When K degenerates at 1 and 2, MIC becomes the Case 1 and Case 2 IC data, respectively;

- For simplicity, we first focus on the current status data, i.e., $K = 1$, and thus observe (V, Δ, Z, W). In the end of this talk, we will mention how to extend the results to the MIC data;

- For simplicity, we first focus on the current status data, i.e., $K = 1$, and thus observe (V, Δ, Z, W). In the end of this talk, we will mention how to extend the results to the MIC data;
- The major contribution of this talk is to consider the efficient estimation in a general class of transformation models under general censorship, i.e., MIC data.

The log-likelihood of SAT model is written as

$$\begin{aligned}
& \ell(\beta, h_1, \ldots, h_d, H) \\
= \; & \delta \log \left\{ F\left[H(v) + \beta' z + \sum_{j=1}^{d} h_j(w_j) \right] \right\} \\
& +(1-\delta) \log \left\{ 1 - F\left[H(v) + \beta' z + \sum_{j=1}^{d} h_j(w_j) \right] \right\}.
\end{aligned}$$

Note that the semiparametric binary model, i.e., $P(\Delta = 1|Z, W, V) = F(\beta' Z + \sum_{j=1}^{d} h_j(W_j) + H(V))$, has the same form of log-likelihood.

- Let $g(\cdot) = \log \dot{H}(\cdot)$. Assuming that g and h_j's are smooth functions, we can approximate them by the B-splines:

$$\begin{aligned} g(v) &\approx \gamma_0' \mathbf{B}_0(v), \\ h_j(w_j) &\approx \gamma_j' \mathbf{B}_j(w_j) \text{ for } j = 1, \ldots, d, \end{aligned}$$

where $\mathbf{B}_j = (B_{j1}, \ldots, B_{jK_j})'$ is a K_j-vector of smooth basis functions.

- Let $g(\cdot) = \log \dot{H}(\cdot)$. Assuming that g and h_j's are smooth functions, we can approximate them by the B-splines:

$$\begin{aligned} g(v) &\approx \gamma_0' \mathbf{B}_0(v), \\ h_j(w_j) &\approx \gamma_j' \mathbf{B}_j(w_j) \text{ for } j = 1, \ldots, d, \end{aligned}$$

  where $\mathbf{B}_j = (B_{j1}, \ldots, B_{jK_j})'$ is a K_j-vector of smooth basis functions.

- By approximating $\log \dot{H}$ with the B-spline, we can avoid the monotonicity constraint on H in the implementation.

- Denote $\alpha = (\beta', g, h_1, \ldots, h_d)'$. We obtain the B-spline estimate as

$$\widehat{\alpha} = (\widehat{\beta}', \widehat{\gamma}_0'\mathbf{B}_0, \ldots, \widehat{\gamma}_d'\mathbf{B}_d) = \arg \sup_{\beta, \gamma_0, \ldots, \gamma_d} \sum_{i=1}^{n} \ell_i(\alpha),$$

where $\ell_i(\alpha) = \ell(\beta, \gamma_1'\mathbf{B}_1, \ldots, \gamma_d'\mathbf{B}_d, \int \exp(\gamma_0'\mathbf{B}_0) ds)$ at the observation i.

- Denote $\alpha = (\beta', g, h_1, \ldots, h_d)'$. We obtain the B-spline estimate as

$$\widehat{\alpha} = (\widehat{\beta}', \widehat{\gamma}_0' \mathbf{B}_0, \ldots, \widehat{\gamma}_d' \mathbf{B}_d) = \arg \sup_{\beta, \gamma_0, \ldots, \gamma_d} \sum_{i=1}^{n} \ell_i(\alpha),$$

  where $\ell_i(\alpha) = \ell(\beta, \gamma_1' \mathbf{B}_1, \ldots, \gamma_d' \mathbf{B}_d, \int \exp(\gamma_0' \mathbf{B}_0) ds)$ at the observation i.

- The Hessian matrix of $\ell_i(\alpha)$ w.r.t. $(\beta', \gamma_0', \ldots, \gamma_d')'$ is negative semidefinite under mild conditions. This implies the existence of $\widehat{\alpha}$.

- Now, we have translated the semiparametric estimation with multiple nonparametric functions into the parametric estimation with increasing dimension.

- Now, we have translated the semiparametric estimation with multiple nonparametric functions into the parametric estimation with increasing dimension.

- Our asymptotic theory will show that (i) $\widehat{\beta}$ is semiparametric efficient after such parametric approximations; (ii) such parametric approximation yields a consistent B-spline estimate for the asymptotic variance of $\widehat{\beta}$ under reasonable conditions.

Assumptions

M1. Regularity Conditions: U and V are independent given (Z, W); $E(Z - E(Z|V, W))^{\otimes 2}$ is positive definite; $V \in [l_v, u_v]$....

Assumptions

M1. Regularity Conditions: U and V are independent given (Z, W); $E(Z - E(Z|V, W))^{\otimes 2}$ is positive definite; $V \in [l_v, u_v]$....

M2. Conditions on the residual error distribution $F(\cdot)$: standard normal (probit model); Pareto distribution (odds-rate model); extreme value distribution (log-log transformation model); logistic distribution (logit transformation model).

Assumptions

M1. Regularity Conditions: U and V are independent given (Z, W); $E(Z - E(Z|V, W))^{\otimes 2}$ is positive definite; $V \in [l_v, u_v]$....

M2. Conditions on the residual error distribution $F(\cdot)$: standard normal (probit model); Pareto distribution (odds-rate model); extreme value distribution (log-log transformation model); logistic distribution (logit transformation model).

M3. Parameter space conditions: we assume that $g \in \mathbf{H}_{c_0}^{r_0}[l_v, u_v]$ and $h_j \in \mathbf{H}_{c_j}^{r_j}[0, 1]$ for $j = 1, \ldots, d$, where $\mathbf{H}_c^r$ denotes the Hölder ball with smoothness r and norm c.

Theorem 1. (Convergence Rate)

Let $d(\alpha, \alpha_0) = \|\beta - \beta_0\| + \|H - H_0\|_2 + \sum_{j=1}^{d} \|h_j - h_{j0}\|_2$. Under Conditions M1-M3, we have

$$d(\widehat{\alpha}, \alpha_0) = O_P(n^{-r/(2r+1)}),$$

where $r = \min_{0 \leq j \leq d}\{r_j\}$, if we require that $K_j \asymp n^{1/(2r_j+1)}$.

Remark:

Theorem 1. (Convergence Rate)

Let $d(\alpha, \alpha_0) = \|\beta - \beta_0\| + \|H - H_0\|_2 + \sum_{j=1}^{d} \|h_j - h_{j0}\|_2$. Under Conditions M1-M3, we have

$$d(\widehat{\alpha}, \alpha_0) = O_P(n^{-r/(2r+1)}),$$

where $r = \min_{0 \le j \le d}\{r_j\}$, if we require that $K_j \asymp n^{1/(2r_j+1)}$.

Remark:

- The B-spline estimates $\widehat{H}$ and $\widehat{h}_j$ are also uniformly consistent;

Theorem 1. (Convergence Rate)

Let $d(\alpha, \alpha_0) = \|\beta - \beta_0\| + \|H - H_0\|_2 + \sum_{j=1}^{d} \|h_j - h_{j0}\|_2$. Under Conditions M1-M3, we have

$$d(\widehat{\alpha}, \alpha_0) = O_P(n^{-r/(2r+1)}),$$

where $r = \min_{0 \le j \le d}\{r_j\}$, if we require that $K_j \asymp n^{1/(2r_j+1)}$.

Remark:

- The B-spline estimates $\widehat{H}$ and $\widehat{h}_j$ are also uniformly consistent;
- Interesting convergence interfere phenomenon: the convergence rate for each B-spline estimate is forced to equal the slowest one. How to solve this issue?

We next study the weak convergence of $\widehat{\beta}$ in the presence of multiple nonparametric nuisance functions.

We next study the weak convergence of $\widehat{\beta}$ in the presence of multiple nonparametric nuisance functions.

- Denote the efficient information matrix as $\widetilde{I}_0 = E\widetilde{\ell}_0\widetilde{\ell}_0'$. It is well known that $\widetilde{I}_0^{-1}$ represents the minimal asymptotic variance bound for the estimate of β.

We next study the weak convergence of $\widehat{\beta}$ in the presence of multiple nonparametric nuisance functions.

- Denote the efficient information matrix as $\widetilde{I}_0 = E\widetilde{\ell}_0\widetilde{\ell}_0'$. It is well known that $\widetilde{I}_0^{-1}$ represents the minimal asymptotic variance bound for the estimate of β.
- We derive $\widetilde{I}_0$ by taking the two-stage projection approach which is needed due to the existence of multiple nonparametric functions, i.e., the projection onto the nonorthogonal sumspace.

We next study the weak convergence of $\widehat{\beta}$ in the presence of multiple nonparametric nuisance functions.

- Denote the efficient information matrix as $\widetilde{I}_0 = E\widetilde{\ell}_0\widetilde{\ell}_0'$. It is well known that $\widetilde{I}_0^{-1}$ represents the minimal asymptotic variance bound for the estimate of β.
- We derive $\widetilde{I}_0$ by taking the two-stage projection approach which is needed due to the existence of multiple nonparametric functions, i.e., the projection onto the nonorthogonal sumspace.
- We assume some model assumptions M4 implying that the *abstract* least favorable directions belong to some Hölder balls so that they can be well approximated by the B-splines.

Theorem 2. (Semiparametric Efficient Estimation)

Under Conditions M1-M4, we have

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} N(0, \widetilde{I}_0^{-1})$$

if we require $K_j \asymp n^{1/(2r_j+1)}$ and $\widetilde{I}_0$ is invertible.

The efficient information $\widetilde{I}_0$ is related to an infinite dimensional optimization problem. However, we can give an explicit B-spline estimate for $\widetilde{I}_0$ by treating the SAT as if it were a parametric model indexed by $(\beta', \gamma_0', \ldots, \gamma_d')' \equiv (\beta', \eta')'$.

The efficient information $\widetilde{I}_0$ is related to an infinite dimensional optimization problem. However, we can give an explicit B-spline estimate for $\widetilde{I}_0$ by treating the SAT as if it were a parametric model indexed by $(\beta', \gamma_0', \ldots, \gamma_d')' \equiv (\beta', \eta')'$.

- Denote the observed information for $(\beta', \eta')'$ as

$$\widehat{J} = \begin{pmatrix} \widehat{I}_{\beta\beta} & \widehat{I}_{\beta\eta} \\ \widehat{I}_{\eta\beta} & \widehat{I}_{\eta\eta} \end{pmatrix}_{(p+\sum_{j=0}^{d} K_j)\times(p+\sum_{j=0}^{d} K_j)},$$

where $\widehat{I}_{jk} = \sum_{i=1}^{n} A_j(X_i;\widehat{\alpha})A_k'(X_i;\widehat{\alpha})/n$, and

$$\begin{aligned} A_\beta(X;\alpha) &= \dot{\ell}_\beta(X;\alpha), \\ A_\eta(X;\alpha) &= \left(\dot{\ell}_g[B_{01}], \ldots, \dot{\ell}_g[B_{0K_0}], \dot{\ell}_{h_1}[B_{11}], \ldots, \dot{\ell}_{h_d}[B_{dK_d}]\right)'. \end{aligned}$$

- The parametric inferences imply that the information estimator for β is of the form

$$\widehat{I} = \widehat{I}_{\beta\beta} - \widehat{I}_{\beta\eta}\widehat{I}_{\eta\eta}^{-1}\widehat{I}_{\eta\beta}.$$

- The parametric inferences imply that the information estimator for β is of the form

$$\widehat{I} = \widehat{I}_{\beta\beta} - \widehat{I}_{\beta\eta}\widehat{I}_{\eta\eta}^{-1}\widehat{I}_{\eta\beta}.$$

- Some calculations further reveal that

$$\widehat{I} = \mathbb{P}_n \left[\dot{\ell}_{\widehat{\beta}} - \dot{\ell}_{\widehat{g}}[(\bar{\gamma}_0^{\dagger})'\mathbf{B}_0] - \sum_{j=1}^{d} \dot{\ell}_{\widehat{h}_j}[(\bar{\gamma}_j^{\dagger})'\mathbf{B}_j] \right]^{\otimes 2},$$

where $[\bar{\gamma}_j^{\dagger}]_{K_j \times l} = (\gamma_{j1}^{\dagger}, \ldots, \gamma_{jl}^{\dagger})$ for $j = 0, 1, \ldots, d$ and $(\gamma_{0k}^{\dagger}, \ldots, \gamma_{dk}^{\dagger})^T = \widehat{I}_{\eta\eta}^{-1}\widehat{I}_{\eta\beta}1_k$ (just the B-spline estimates of the least favorable directions).

Theorem 3. Consistency of the Efficient Information Estimate

Suppose that Conditions M1-M4 hold. If we further assume that

$$E \sup_{\gamma_0} \left[\int_{l_v}^{V} [\exp(g(s)) - \exp(g_0(s))] \gamma_0' \mathbf{B}_0(s) \right]^2 ds \leq C \|H - H_0\|^2,$$

then we have $\widehat{I} \xrightarrow{P} \widetilde{I}_0$.

Summary of Simulation Results

- In the simulations, we assume $d = 1$ for simplicity. By AIC criterion, we choose $K_0, K_1 = 5$. Based on our experiences, it is proper to choose less than ten knots to achieve reasonable approximation.

Summary of Simulation Results

- In the simulations, we assume $d = 1$ for simplicity. By AIC criterion, we choose $K_0, K_1 = 5$. Based on our experiences, it is proper to choose less than ten knots to achieve reasonable approximation.
- Our computational cost is much less than the penalized estimation approach proposed in Ma and Kosorok (2005), i.e., the cumulative sum diagram, since K_0, K_1 is usually smaller than n.

Extensions to the MIC data

Extensions to the MIC data

- Key idea: introduce a new measure;

Extensions to the MIC data

- Key idea: introduce a new measure;
- Let $G(z,w)$ be the distribution function of (Z,W). Introduce

$$\begin{aligned}\nu(B\times C) &= \int_C \sum_{k=1}^{\infty} P(K=k|Z=z,W=w) \\ &\quad \times \sum_{j=1}^{k} P(V_{K,j}\in B|K=k,Z=z,W=w)dG(z,w),\end{aligned}$$

and $\mu(B)=\nu(B\times\mathbb{R}^{d+1})$.

- Redefine $d(\alpha, \alpha_0) = \|\beta - \beta_0\| + \|H - H_0\|_{L_2(\mu)} + \sum_{j=1}^{d} \|h_j - h_{j0}\|_2$ in the MIC data.;

- Redefine $d(\alpha, \alpha_0) = \|\beta - \beta_0\| + \|H - H_0\|_{L_2(\mu)} + \sum_{j=1}^{d} \|h_j - h_{j0}\|_2$ in the MIC data.;

- Need two new assumptions:

- Redefine $d(\alpha, \alpha_0) = \|\beta - \beta_0\| + \|H - H_0\|_{L_2(\mu)} + \sum_{j=1}^{d} \|h_j - h_{j0}\|_2$ in the MIC data.;
- Need two new assumptions:
  - $P(K \leq k_0) = 1$ for some $k_0 < \infty$;

- Redefine $d(\alpha, \alpha_0) = \|\beta - \beta_0\| + \|H - H_0\|_{L_2(\mu)} + \sum_{j=1}^{d} \|h_j - h_{j0}\|_2$ in the MIC data.;

- Need two new assumptions:
  - $P(K \leq k_0) = 1$ for some $k_0 < \infty$;
  - The $V_{K,j}$'s are s_0-separated: there exists a constant $s_0 > 0$ such that $P(V_{K,j} - V_{K,j-1} \geq s_0 \;\text{ for all } j = 1, \ldots, K+1) = 1$.

Thanks for your attention....

Assistant Professor Guang Cheng
Department of Statistics, Purdue University
chengg@purdue.edu