Feature Abberation at Surivaltimes Statistics

Jan Weimer

26 Januar, 2015

Was ist Sure Independent Screening

- Ziel : Das Herausfiltern einer Obermenge der aktiven Menge
- Univariate Auswertung
- Anfälligkeit gegen Korrelationen

Inhaltsverzeichnis

Das Modell

- Z p-dimensionales Kovariat
- T Ausfallzeit, C Abbruchzeit
- $N(t) = \mathbf{1}_{\{T \wedge C \leq t, T \leq C\}}$ der Ausfallprozess
- $Y(t) = \mathbf{1}_{\{T \wedge C \geq t\}}$ der Aktivitätsprozess

Davon betrachten wir nun n u.i.v. Kopien $(Z_i, T_i, C_i)_{i \in \{1,\dots,n\}}$

Das Modell

Das Lin-Ying-Modell hat eine Hazard Rate der Form:

$$\lambda(t) = \lambda_0(t) Z^T \cdot \alpha_0$$

Die Ausfallzeit T ist also exponential verteilt mit Parameter λ.

Die FAST Statistik

Wir definieren die FAST Statistik d

$$d := n^{-1} \int_0^\tau \sum_{i=1}^n (Z_i - \bar{Z}(t)) dN_i(t),$$

wobei $\bar{Z}$ als Durchschnitt der aktiven Kovariate definiert ist.

$$\bar{Z}(t) = \sum_{i=1}^n Z_i Y_i / \sum_{i=1}^n Y_i$$

Die FAST Statistik

Wir definieren die FAST Statistik d

$$d := n^{-1} \int_0^{\tau} \sum_{i=1}^{n} (Z_i - \bar{Z}(t)) dN_i(t) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{T_i < C_i\}} (Z_i - \bar{Z}(T \wedge C)),$$

wobei $\bar{Z}$ als Durchschnitt der aktiven Kovariate definiert ist.

$$\bar{Z}(t) = \sum_{i=1}^{n} Z_i Y_i / \sum_{i=1}^{n} Y_i = \frac{1}{\# A_t} \sum_{i \in A_t} Z_i Y_i \quad A_t := \{i : Y_i = 1\}$$

Die FAST Statistik als Schätzer

$$d :== n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{T_i < C_i < \tau\}} \left(Z_i - \bar{Z} \left(T \wedge C \right) \right)$$

Ist $\alpha_k > 0$, so fallen Objekte mit großem k-ten Merkmal schneller als solche mit kleinem.
Also ist üblicherweise $Z_{i,k}$ größer als $\bar{Z}_k(T_i)$ und d_k nimmt Werte größer Null an.

Die FAST Statistik als Schätzer

$$d :== n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{T_i < C_i < \tau\}} \left(Z_i - \bar{Z}\left(T \wedge C\right)\right)$$

Ist $\alpha_k < 0$ so fallen Objekte mit kleinem k-ten Merkmal schneller als solche mit kleinem.
Also ist üblicherweise $Z_{i,k}$ kleiner als $\bar{Z}_k(T_i)$ und d_k nimmt Werte kleiner Null an.

Die FAST Statistik als Schätzer

$$d :== n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\{T_i < C_i < \tau\}} \left(Z_i - \bar{Z} \left(T \wedge C \right) \right)$$

Ist $\alpha_k = 0$ so fallen Objekte mit großem k-ten Merkmal genauso schnell aus, wie solche mit kleinem.
Also ist üblicherweise $Z_{i,k}$ gleich $\bar{Z}_k(T_i)$ und d_k konvergiert nach dem Gesetz der großen Zahlen gegen 0.

Die empirische Design-Matrix

Wir definieren das empirische Äquivalent zur Designmatrix

$$D = n^{-1} \sum_{i=0}^{n} \int_{0}^{\tau} Y_i(t) \left\{ Z_i - \bar{Z}(t) \right\}^{\otimes 2} dt.$$

Schätzer für α_0

Betrachtet man nun die Doob-Meyer Zerlegung von N_i, erhält man

$$N_i = M_i + Y(t)_i \left(\lambda_0 (t) + Z^T \alpha_0 \right).$$

Integration von $\sum_{i=1}^n Z_i - \bar{Z}(t)$ ergibt:

$$\sum_{i=1}^n \int_0^\tau Z_i - \bar{Z}(t) dN_i(t) = \left[\sum_{i=1}^n \int_0^\tau \left(Z_i - \bar{Z}(t) \right) * \left(Z_i - \bar{Z}(t) \right)^T dt \right] \alpha_0$$
$$+ \sum_{i=1}^n \int_0^\tau Z_i - \bar{Z}(t) \, dM_i(t).$$

Der FAST Algorithmus

Schritt 1 Berechne anhand der Daten die FAST Statistik d.

Schritt 2 Wähle einen Schwellwert γ und erkläre alle Features k mit $|d_k| > \gamma$ als relevant.

Der FAST Algorithmus

Schritt 1 Berechne anhand der Daten die FAST Statistik d.

Schritt 2 Wähle einen Schwellwert γ und erkläre alle Features k mit $|d_k| > \gamma$ als relevant.

Theoretisch kann man den Algorithmus in zwei Probleme aufteilen. Hierfür definieren wir den Erwartungswert von d für n gegen unendlich

$$\delta = E\left[\int_0^\tau \left(Z_1 - \frac{E[Z_1 Y_1]}{E[Y_1]}\right) \mathbf{1}_{\{T_1 \leq t \wedge \tau\}}\right].$$

Notation

Wir definieren drei Mengen von Merkmalen:

$M^n := \{1 \leq k \leq p_n : \alpha_k^0 \neq 0\}$ die aktiven Merkmale

$M_\delta^n := \{1 \leq k \leq p_n : \delta_k \neq 0\}$ die FAST-aktiven Merkmale

$\hat{M}_d^n := \{1 \leq k \leq p_n : |d_k| > \gamma_n\}$ die aktiv geschätzten Merkmale

Notation

Wir definieren drei Mengen von Merkmalen:

$M^n := \{1 \leq k \leq p_n : \alpha_k^0 \neq 0\}$ die aktiven Merkmale

$M_\delta^n := \{1 \leq k \leq p_n : \delta_k \neq 0\}$ die FAST-aktiven Merkmale

$\hat{M}_d^n := \{1 \leq k \leq p_n : |d_k| > \gamma_n\}$ die aktiv geschätzten Merkmale

Fragestellung 1: Unter welchen Voraussetzungen ist die Menge M^n der aktiven Merkmale Teilmenge der FAST-aktiven Merkmale M_δ^n.

Fragestellung 2: Unter welchen Voraussetzungen ist die Menge M_δ^n der FAST-aktiven Merkmale Teilmenge der aktiv geschätzten Merkmale $\hat{M}_d^n$.

Annahmen 1

1) Die Überlebensfunktion $\exp\left(-\int_0^t \lambda(t,\cdot)ds\right)$ ist stetig differenzierbar
2) Es gibt ein $c \in \mathbb{R}^p : E[Z_1|Z_1^T\alpha^0] = c \cdot Z_1^T\alpha_0$
3) Die Abbruchzeit hängt von T_1 und Z_1 nur durch $Z_{1,k}$, $k \notin M^n$ ab
4) $Z_{1,k}$ $k \in M^n$ist unabhängig von $Z_{1,j}, j \notin M^n$

Annahmen 2

a) $E[Z_1, k] = 0, V[Z_1, k] = 1$

b) $P\left(Y_1(\tau) = 1\right) > 0$

c) $Var[Z_1^T \alpha^0]$ ist gleichmäßig beschränkt

d) $P(|Z_{1,k}| > s) \leq l_0 \exp(-l_1 s^\eta)$ für ein $l_0, l_1, \eta > 0$ und s genügend groß.

Lemma

Gelten Annahmen a) und b), dann gibt es $C_1, C_2 > 0$, so dass für alle $K > 0$ und $k \in \{1, ..., p_n\}$ gilt:

$$
\begin{aligned}
P[n^{\frac{1}{2}}|d_k - \delta_k| > C_1(1+t)] \leq & 10 exp(-t^2/(8K^2)) \\
& + C_2 exp(-n/2) \\
& + nP[|Z_{1,k}| > K]
\end{aligned}
$$

Massard Concentration Theoreme

Seien $X_1, ..., X_n$ unabhängig mit Werten in $\mathbb{R}^N$ so dass $a_{i,t} < X_{i,t} < b_{i,t}$. Definiere

$$L^2 := \sup_{1 \leq t \leq N} \sum_{i=1}^{n} (b_{i,t} - a_{i,t})^2 \qquad Z := \sup_{t \in \{1,...,N\}} \sum_{i=1}^{n} X_{i,t}$$

dann gilt

$$P\left[Z \geq E[Z] + x\right] < \exp\left(-x^2/2L^2\right).$$

Massard Concentration Theoreme

Für $\tilde{Z} = \frac{1}{n} \sup \sum_{i=1}^{n} X_{i,t}$ und $a_{i,t} = -K$ und $b_{i,t} = K$ gilt

$$P\left[n\tilde{Z} \geq nE[\tilde{Z}] + x\right] < \exp\left(\frac{-x^2}{2n(2K)^2}\right)$$

$$\Longleftrightarrow$$

$$P\left[n^{\frac{1}{2}}\tilde{Z} \geq n^{\frac{1}{2}}E[\tilde{Z}] + \tilde{x}\right] < \exp\left(\tilde{x}^2/8K^2\right)$$

Lemma

Gelten Annahmen *a*) und *b*), dann gibt es $C_1, C_2 > 0$, so dass für alle $K > 0$ und $k \in \{1, ..., p_n\}$ gilt:

$$\begin{aligned} P[n^{\frac{1}{2}}|d_k - \delta_k| > C_1(1+t)] \leq \quad & 10 \exp(-t^2/(8K^2)) \\ & + C_2 \exp(-n/2) \\ & + nP[|Z_{1,k}| > K] \end{aligned}$$

Gilt nun außerdem Annahme *d*),also $P(|Z_{1,k}| > s) \leq l_0 \exp(-l_1 s^\eta)$ für ein $l_0, l_1, \eta > 0$ so erhält man

$$nP[|Z_{1,k}| > K] \leq l_0 \exp\left(-l_1 * nK^\eta\right)$$

und kann die drei exponentiellen Raten zusammenfassen.

Satz 1

Gelten die Annahmen 1-3 sowie a-dund außerdem

$$|cov(Z_{1_k}), Z_1^T\alpha^0| \geq c_1 n^{-\kappa} \forall k \in M^n$$

für ein $c_1 > 0$ und $\kappa \in \left(0, \frac{1}{2}\right)$, dann gilt

$$M^n \subset M^n_\delta.$$

Satz 1

Gelten die Annahmen 1-3 sowie a-dund außerdem

$$|cov(Z_{1_k}), Z_1^T \alpha^0| \geq c_1 n^{-\kappa} \forall k \in M^n$$

für ein $c_1 > 0$ und $\kappa \in \left(0, \frac{1}{2}\right)$, dann gilt

$$M^n \subset M_\delta^n.$$

Wählt man nun $\gamma_n = c_2 n^{-\kappa}$ für ein $c_2 \in [0, \frac{c_1}{2}]$ und ist $log(p_n) = o(n^{(1-2\kappa)\eta/(\eta+2)})$ dann gilt auch die SIS Eigenschaft

$$\lim_{n\to\infty} P(M^n \subset \hat{M}_d^n) = 1.$$

Satz 2

Unter Annahmen 1-4 , a-d und $\gamma_n = \tilde{c_2} n^{-\kappa}$ für $\kappa < \frac{1}{2}, \tilde{c_2} > 0$, dann gibt es ein $c_3 > 0$, sodass

$$P\left[\hat{M}_d^n \geq O\left(n^{2\kappa}\lambda_{max}\left(\Sigma\right)\right)\right] \leq 1 - O\left(p_n exp\left(-c_5 n^{(1-2\kappa)\eta/(eta+2)}\right)\right)$$

Skalieren der FAST Statistik

Problem: Merkmale sind nicht immer zentriert und standardisiert.
Deshalb kann man verschiedene Normierungen vornehmen:

Ling-Ying $d_k^L Y := d_k D_{k,k}^{-1}$

FAST-Loss $d_k^L Y := d_k D_{k,k}^{-\frac{1}{2}}$

Z-FAST $d_k^L Y := d_k B_{k,k}^{-\frac{1}{2}}$

Satz 4

Folge T einem Modell mit einer Ausfallrate der Form $\lambda(t, Z^T\alpha^0)$. Es gelte außerdem Annahme 2 und, dass C unabhängig von T und Z ist. Verwendet man nun ein Ling Ying Modell und löst die Gleichung $\Lambda\beta^0 = \delta$ auf und ist Λ nicht singulär ,so gibt es eine Konstante $\nu(\mathbb{L}(Z_1^T\alpha^0), \mathbb{L}(C))) \neq 0$,sodass

$$\beta^0 = \nu\alpha^0.$$

Iterated SIS (ISIS)

Es folgt ein Standard ISIS Algorithmus:

Schritt 1 Berechne die FAST Statistik d und skaliere sie mit D_{kk}^{-1}. Sei A_0 die Menge der k_0 relevantesten Merkmale.

Schritt 2 Führe mit den ausgewählten Parametern eine Regression durch (z.B. Lasso, SCAD, SICA) und schreibe die übrig gebliebenen Merkmale in B_0.

Schritt 3 Falls keine Merkmale entfernt wurden, d.h. $B_r = B_{r-1}$ oder $r = r_{max}$, gebe B_r aus, sonst gehe zu Schritt 4.

Schritt 4 (Neurekrutierung) Evaluiere die Relevanz der Merkmale, die nicht Ausgewählt wurden erneut, in dem Ling-Ying Modell ohne Penalty-Funktion, welches nur die Merkmale $i \notin M \setminus B_R$ betrachtet.

FAST Statistik

Quelle : Anders Gorst-Rasmussen, Thomas Scheike, Independent screening for single-index hazard rate models with ultrahigh dimensional features.