

EEN VALKUIL BIJ HET TOEPASSEN VAN DE JACKKNIFE

door

WILBERT C.M. KALLENBERG¹

SAMENVATTING

De jackknife methode is een procedure om onzuiverheid van schatters te reduceren. In dit artikel wordt een voorbeeld gepresenteerd waarin het rechtstreeks toepassen van de jackknife methode leidt tot een aanzienlijk onzuiverder resultaat dan de oorspronkelijke maximum likelihood methode. De conclusie is dat de jackknife methode met de nodige zorgvuldigheid toegepast dient te worden.

1 Faculteit Toegepaste Wiskunde, Universiteit Twente, Postbus 217,
7500 AE Enschede, tel. 053 - 893374

INLEIDING

De jackknife methode werd in 1949 geïntroduceerd door Quenouille als een techniek om de onzuiverheid van een schatter voor de seriële correlatiecoëfficiënt te verminderen. Dit idee generaliseerde hij in 1956 voor algemene situaties. Tukey (1958) gaf aan dat de jackknife methode ook gebruikt kan worden om de variantie te schatten en om betrouwbaarheidsintervallen te construeren. Voor een introductie en een historisch overzicht zij verwezen naar Miller (1974).

Rechtstreekse toepassing van de jackknife kan onder bepaalde omstandigheden een forse extra onzuiverheid introduceren. In dit artikel wordt een voorbeeld uit de bacteriologie gegeven, waarin zich deze situatie voordoet. Tevens wordt uiteengezet hoe de valkuil vermeden kan worden door de jackknife methode op de juiste schatter toe te passen.

EEN VOORBEELD WAARIN DE JACKKNIFE PROCEDURE FAALT

We bekijken de situatie waarbij het aantal bacteriën in verontreinigd water per eenheid volume geschat moeten worden. Een gebruikelijke aanname hierbij is dat het aantal bacteriën een Poisson verdeling heeft met parameter $\mu > 0$. In de praktijk is het niet mogelijk om het aantal bacteriën rechtstreeks te tellen. Een procedure om het aantal bacteriën te bepalen is om de volume eenheid van het verontreinigde water op te delen in n gelijke hoeveelheden. Voor elk deel kan geregistreerd worden of er wel of niet een colonie bacteriën zich heeft ontwikkeld hetgeen betekent dat wel of niet tenminste 1 bacterie in dat deel aanwezig was.

Dit leidt tot de volgende formulering van het probleem. Zij gegeven de waarnemingen $X_1, \dots, X_n$ met

$$X_i = \begin{cases} 0 & \text{als het } i\text{-de deel geen bacteriën bevat} \\ 1 & \text{als het } i\text{-de deel tenminste 1 bacterie bevat.} \end{cases}$$

Deze $X_1, \dots, X_n$ zijn onderling onafhankelijk en identiek verdeelde stochastische grootheden met $P_\mu(X_i=1) = 1 - \exp(-\mu/n)$, $i=1, \dots, n$. Noteren we het gemiddelde van de waarnemingen met $\bar{X}_n$ dan is de maximum likelihood schatter $\hat{\mu}_n$ gelijk aan $-\log(1-\bar{X}_n)$ als $\bar{X}_n < 1$. Indien $\bar{X}_n = 1$ dan bestaat de maximum likelihood schatter niet. In dat geval stellen we $\hat{\mu}_n = n \log n$. Merk op dat $P_\mu(\bar{X}_n=1) = \{1 - \exp(-\mu/n)\}^n \leq (\mu/n)^n$. Voor grote waarden van n is deze kans bijzonder klein. Hieruit volgt dat een redelijke schatter voor μ gedefinieerd wordt door:

$$\hat{\mu}_n = \begin{cases} -\log(1-\bar{X}_n) & , \text{als } \bar{X}_n < 1, \\ n \log n & , \text{als } \bar{X}_n = 1. \end{cases} \quad (1)$$

Deze schatter $\hat{\mu}_n$ is onzuiver. Er bestaat namelijk geen enkele zuivere schatter van μ op grond van $X_1, \dots, X_n$, zoals het volgende lemma met $Y = (X_1, \dots, X_n)$, $\theta = \mu$ en $g(\theta) = \theta$ laat zien.

LEMMA: Veronderstel dat de stochastische vector Y slechts eindig veel waarden aanneemt. Laat de verdeling van Y afhankelijk zijn van een onbekende parameter $\theta \in \Theta$, de parameterruimte. Zij g een reëelwaardige functie op Θ met $\sup\{|g(\theta)| : \theta \in \Theta\} = \infty$. Dan bestaat er geen zuivere schatter van $g(\theta)$ op grond van Y .

BEWIJS: Zij $V=V(Y)$ een schatter van $g(\theta)$. Noteer de mogelijke waarden van Y met $\{y_1, \dots, y_k\}$. Voor alle $\theta \in \Theta$ geldt

$$m = \min_{1 \leq i \leq k} V(y_i) \leq E_\theta V(Y) \leq \max_{1 \leq i \leq k} V(y_i) = M.$$

Als $|g(\theta_0)| > \max\{|m|, M\}$ is dus $E_{\theta_0} V(Y) \neq g(\theta_0)$. Omdat

$\sup\{|g(\theta)| : \theta \in \Theta\} = \infty$, bestaat zo'n θ_0 en is derhalve $V(Y)$ geen zuivere schatter van $g(\theta)$.

□

Opmerking: Met dit lemma kan op een eenvoudige en algemene wijze het niet bestaan van zuivere schatters bewezen worden van bijvoorbeeld p^{-1} of $p/(1-p)$ op grond van de steekproef $X_1, \dots, X_n$ van Bernoulli variabelen met succeskans p .

Aangetoond kan worden (zie Kallenberg (1984)) dat

$$E_{\mu} \hat{\mu}_n = \mu + \frac{\mu}{2n} + \frac{4\mu+3\mu^2}{12n^2} + O(n^{-3}), \text{ als } n \rightarrow \infty.$$

Aangezien de schatter onzuiver is, kan de jackknife procedure aangewend worden om deze onzuiverheid te reduceren. In het kort zullen wij de jackknife procedure beschrijven:

Stel dat men een parameter θ wil schatten uitgaande van $Y_1, \dots, Y_n$. Beschouw een schatter $T_n = T_n(Y_1, \dots, Y_n)$ dan wordt de i -de pseudo-waarde gedefiniëerd door

$$T_{ni}^* = nT_n - (n-1)T_{n-1}(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n) \quad , i=1, \dots, n.$$

De jackknife schatter T_n^* gedefiniëerd door

$$T_n^* = n^{-1} \sum_{i=1}^n T_{ni}^*$$

reduceert in veel gevallen de onzuiverheid. De reden hiervan is als volgt: als

$$E_{\theta} T_n = \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + O(n^{-3}), \text{ als } n \rightarrow \infty,$$

voor constanten a_1 en a_2 dan geldt veelal dat

$$E_{\theta} T_n^* = \theta + O(n^{-2}), \text{ als } n \rightarrow \infty.$$

Aangezien de schatter $\hat{\mu}_n$ in (1) onzuiver is, kunnen wij proberen bovenstaande jackknife procedure toe te passen. Echter het resultaat is desastreus omdat aangetoond kan worden (zie Kallenberg (1984)) dat voor de jackknife schatter $\hat{\mu}_n^*$ geldt dat

$$E_{\mu} \hat{\mu}_n^* = 2\mu - \frac{\mu}{2n} + O(n^{-2}), \text{ als } n \rightarrow \infty. \quad (2)$$

Het is duidelijk dat het zonder na te denken toepassen van de jackknife procedure zelfs een kleine onzuiverheid (van de orde n^{-1}) kan doen verkeren in een forse onzuiverheid (van orde een). Immers de asymptotisch zuivere schatter ($\lim_{n \rightarrow \infty} E_{\mu} \hat{\mu}_n = \mu$) verwordt tot een asymptotisch onzuivere schatter ($\lim_{n \rightarrow \infty} E_{\mu} \hat{\mu}_n^* = 2\mu > \mu$)! Waarom gaat de gebruikte jackknife procedure mis? De reden hiervoor is dat de jackknife toegepast wordt op de verkeerde schatter. Om dit te begrijpen het volgende:

$$P_{\mu}(X_i=1) = 1 - e^{-\mu/n} \approx \mu/n.$$

Dus een manier om tegen het schattingsprobleem aan te kijken is om over te gaan op onderling onafhankelijke Bernoulli variabelen $Z_1, \dots, Z_n$ met

$$P_{\mu}(Z_i=1) = \mu/n, \quad i=1, \dots, n.$$

De laatste formulering is asymptotisch equivalent met het originele probleem. Eenvoudig is in te zien dat $T_n = \sum_{i=1}^n Z_i$ een zuivere schatter is van μ . Toepassing van de jackknife procedure geeft

$$T_{ni}^* = n \sum_{j=1}^n Z_j - (n-1) \sum_{j \neq i} Z_j = T_n + (n-1)Z_i, \quad i=1, \dots, n,$$

hetgeen impliceert dat

$$T_n^* = (1 + \frac{n-1}{n})T_n$$

en

$$E_{\mu} T_n^* = (2-n^{-1})\mu.$$

De conclusie is derhalve dat uitgaande van een zuivere schatter de jackknife procedure in deze situatie leidt tot een volledig onzuivere schatter. Het is voor iedereen duidelijk dat indien men beschikt over een zuivere schatter men vervolgens niet de jackknife procedure zal gebruiken. Het wordt hier toch gepresenteerd om aan te geven waarom (2) kan optreden. Bovendien illustreert het hoe de jackknife procedure wel goed toegepast kan worden. Immers als we de jackknife methode toepassen op $\tilde{T}_n = n^{-1}T_n$ en we vermenigvuldigen de jackknife versie van $\tilde{T}_n$ achteraf met n dan is er geen enkel probleem (omdat $\tilde{T}_n^* = \tilde{T}_n$ en $n\tilde{T}_n^* = T_n$, dus de jackknife procedure levert dezelfde zuivere schatter op). Deze oplossing kunnen wij nu ook toepassen op de oorspronkelijke onzuivere schatter $\hat{\mu}_n$ in (1). Dus

$$\tilde{\mu}_n = n^{-1}\hat{\mu}_n = \begin{cases} -\log(1-\bar{X}_n) & , \text{als } \bar{X}_n < 1, \\ \log n & , \text{als } \bar{X}_n = 1. \end{cases}$$

Hieruit volgt (zie Kallenberg (1984))

$$\tilde{\mu}_n^* = \begin{cases} -n^2 \log(1-\bar{X}_n) + (n-1) \sum_{i=1}^n \log(1 - \sum_{j \neq i} X_j / (n-1)), & \text{als } \bar{X}_n \leq 1-2n^{-1}, \\ n^2 \log n - n(n-1) \log(n-1) & , \text{als } \bar{X}_n \geq 1-n^{-1}. \end{cases}$$

en $E_{\mu} n\tilde{\mu}_n^* = \mu + O(n^{-2})$ als $n \rightarrow \infty$. Dus de onzuiverheid van de oorspronkelijke schatter is door het toepassen van de jackknife procedure op de juiste grootte gereduceerd.

REFERENTIES

1. Kallenberg, W.C.M. (1984), A simple example where the jackknife fails, Rapport nr. 256, Wiskundig Seminarium der Vrije Universiteit, Amsterdam.
2. Miller, R.G. (1974), The jackknife - a review, Biometrika 61, blz. 1-15.
3. Quenouille, M.H. (1949), Approximate tests of correlation in time-series, Journal of the Royal Statistical Society Series B 11, blz. 68-84.
4. Quenouille, M.H. (1956), Notes on bias in estimation, Biometrika 43, blz 353-360.
5. Tukey, J.W. (1958), Bias and confidence in not-quite large samples (abstract), Annals of Mathematical Statistics 29, blz. 614.

Geaccepteerd: 10-05-1988