

Samenvatting

Het niet-parametrisch schatten van dichtheids- en regressiefuncties vormt reeds lang het onderwerp van diepgaand onderzoek, wat leidde tot een grote variëteit aan methoden. De niet-parametrische benadering steunt op het idee dat weinig of geen veronderstellingen worden gemaakt betreffende de onderliggende verdeling van de variabelen. Dit in tegenstelling tot de parametrische benadering, waarbij verondersteld wordt dat de verdeling van de variabelen een gekend parametrisch model volgt. Het probleem herleidt zich in dit geval tot het schatten van een eindig aantal parameters die dit model beschrijven. Het spreekt voor zich dat de toepasbaarheid van niet-parametrische modellen, wegens de beperkte opgelegde veronderstellingen, veel breder is dan die van het overeenkomstig parametrisch model.

Een welbekende en algemeen gebruikte klasse van schatters bestaat uit de zogenaamde *kernschatters*, die onder andere toelaten om zowel dichtheids- als regressiefuncties te schatten. Een typische kernschatter gebaseerd op onafhankelijke variabelen $(X_1, Y_1), \dots, (X_n, Y_n)$ met waarden in $\mathbb{R}^d \times \mathbb{R}^r$, is gedefinieerd als

$$\hat{\varphi}_{n,h}(t) = \frac{1}{nh^d} \sum_{i=1}^n \varphi(Y_i) K\left(\frac{t - X_i}{h}\right), \quad t \in \mathbb{R}^d,$$

waar K een kernfunctie is, $0 < h < 1$ een bandbreedte en $\varphi : \mathbb{R}^r \rightarrow \mathbb{R}$ een geschikte meetbare functie. Men kan $\hat{\varphi}_{n,h}(t)$ beschouwen als een gewogen gemiddelde van de variabelen $\varphi(Y_i)$, waar de kernfunctie de vorm van de gewichten bepaalt die aan een omgeving van X_i worden toegewezen, en waar de bandbreedte de grootte van deze omgeving onder controle houdt. Naast de meetbaarheid en de veronderstelling dat $\int K(x)dx = 1$ zijn er geen fundamentele beperkingen op de keuze van de kernfunctie. Aangezien de kernschatter de gladheid van de kernfunctie erft, wordt voor een gladde

of minder gladde kernfunctie gekozen afhankelijk van hoe glad men de kernschatting $\hat{\varphi}_{n,h}(t)$ wenst.

De keuze van de bandbreedte daarentegen moet zorgvuldig gebeuren en is cruciaal voor de consistentie van de resulterende kernschatting, die bijzonder gevoelig is aan deze keuze. In het geval van de Nadaraya–Watson schatting (de meest gebruikte niet–parametrische regressieschatting) zal een te kleine bandbreedte leiden tot iets dat hoofdzakelijk op een interpolatie van de data lijkt, terwijl een schatting met een te grote bandbreedte min of meer een horizontale lijn zal zijn. De keuze van de bandbreedte heeft ook een grote invloed op de zogenaamde *bias*. De bias is het verschil tussen de verwachtingswaarde van de schatting en de geschatte grootte. In het algemeen geeft een kleine bandbreedte een schatting met een beperkte bias, maar grotere variantie, terwijl een grotere bandbreedte eerder leidt tot een schatting met een kleine variantie, maar een grotere bias. Men dient dus een aangewezen bandbreedte te bepalen die een schatting zal voortbrengen met een goed evenwicht tussen bias en variantie. Meestal zal de optimale bandbreedte echter verschillen naargelang de situatie, en zal deze afhankelijk zijn van de beschikbare data. Dit betekent dat men het gedrag van zulke schattingen niet meer kan bestuderen aan de hand van “klassieke” resultaten voor schattingen met deterministische bandbreedtes, maar dat men in plaats daarvan nieuwe methodes moet ontwikkelen om de studie van schattingen met data–afhankelijke bandbreedtes mogelijk te maken.

Als uitgangspunt beschouwen we de oplossing die door Einmahl en Mason (2005) werd voorgesteld en die aan de bestaande consistentie–resultaten een supremum over een hele waaier van bandbreedtes toevoegt. Tegenwoordig verwijzen we naar zulke resultaten als “uniform in bandbreedte” resultaten, en deze zijn typisch van de vorm

$$\limsup_{n \rightarrow \infty} \sup_{a_n \leq h \leq b_0} \sup_{\varphi \in \mathcal{F}} \sup_{t \in I} \frac{\sqrt{nh^d} |\hat{\varphi}_{n,h}(t) - \mathbb{E}\hat{\varphi}_{n,h}(t)|}{\sqrt{|\log h| \vee \log \log n}} < \infty, \quad \text{b.o.}, \quad (\star)$$

waar I een compact interval in \mathbb{R}^d is, $b_0 < 1$ een positieve constante en a_n een rij niet–random getallen die met een zekere snelheid naar nul convergeert. Dit extra supremum laat onder andere toe om kernschattingen te beschouwen die gebaseerd zijn op bandbreedtes die functies zijn van de data. Inderdaad, als $\hat{h}_n = H(X_1, \dots, X_n) \in [a_n, b_n]$ zulk een rij “data–driven” bandbreedtes is, impliceert (\star) onmiddellijk dat met kans 1, $|\hat{\varphi}_{n,\hat{h}_n}(t) - \mathbb{E}\hat{\varphi}_{n,\hat{h}_n}(t)| \rightarrow 0$, uniform op compacten $I \subseteq \mathbb{R}^d$. Bovendien kan

men ook convergentiesnelheden bepalen.

Het voornaamste doel van deze thesis is aan te tonen hoe dergelijke consistentieresultaten voor een grote verscheidenheid aan schatters kunnen worden verkregen. Onze methodologie is gebaseerd op technieken die door Einmahl en Mason (2005) ontwikkeld werden en waarvan de belangrijkste stappen in Section 3.3 samengevat worden. We zullen naar deze methode verwijzen als “Standard Methodology”, of kort [SM]. Hoofdzakelijk steunt het op de algemene theorie van empirische processen. Enkele specifieke exponentiële afwijkings- en momentongelijkheden vormen de belangrijkste hulpmiddelen. Doorheen de verschillende hoofdstukken van deze thesis zullen we deze methode meerdere malen toepassen om de “uniform in bandbreedte consistentie” van enkele specifieke klassen van kernschatters te bekomen.

Een eerste toepassing van [SM] resulteert in de uitbreiding van het consistentieresultaat (\star) voor de kern dichtheidsschatter $\hat{f}_{n,h}(t)$ ten opzichte van gewogen supremumnormen gebaseerd op een gepaste gewichtsfunctie ψ . Steunende op een aantal ideeën die in Giné, Koltchinskii en Zinn (2004) staan uitgewerkt, worden onder enkele extra veronderstellingen nodige en voldoende voorwaarden verstrekt opdat

$$\sup_{a_n \leq h \leq b_n} \sup_{t \in \mathbb{R}^d} \frac{\sqrt{nh^d} |\psi(t)(\hat{f}_{n,h}(t) - \mathbb{E}\hat{f}_{n,h}(t))|}{\sqrt{|\log h|}}$$

stochastisch en bijna overal begrensd zou zijn. Hierin is ψ een gepaste gewichtsfunctie en a_n en b_n zijn regulier variërende rijen met negatieve index. Een gedetailleerd bewijs van dit resultaat wordt verder uitgewerkt in Chapter 4.

In Chapter 5 veronderstellen we dat de regressiefunctie $m(t) = \mathbb{E}[Y|X = t]$, $p + 1$ maal differentieerbaar is in $x_0 \in \mathbb{R}$, en we beschouwen een grotere klasse kernschatters voor $m(x_0)$. Deze klasse bestaat uit de zogenaamde “locaal polynomiale regressieschatters”, gedefinieerd als oplossingen van een gewogen kleinste kwadraten probleem waarbij de gewichten gegeven worden door $K((x_0 - X_i)/h)$. In het bijzonder behoort de Nadaraya–Watson schatter tot deze klasse, daar deze een oplossing is van het kleinste kwadraten probleem met $p = 0$. Het is gebleken dat de methode beschreven in [SM] en gebaseerd op de theorie van empirische processen, ook toepasbaar is om de consistentie te bewijzen van de lokaal polynomiale schatters,

uniform in bandbreedte. Als we de lokaal polynomiale schatter van graad $p \geq 0$ in x_0 noteren als $\hat{m}_{n,h}^{(p)}(x_0)$, tonen we in Section 5.2 dat

$$\sup_{(\frac{c \log n}{n})^\gamma \leq h \leq b_n} \sup_{x_0 \in I} |\hat{m}_{n,h}^{(p)}(x_0) - m(x_0)| \longrightarrow 0, \quad \text{b.o.},$$

waar b_n een willekeurige rij is die naar nul convergeert en waar $\gamma = 1$ of $\gamma = 1 - 2/q$ afhankelijk van het feit of Y begrensd is of een eindig q -de moment heeft.

In Chapter 6 besteden we aandacht aan het ‘‘puntsgewijs uniform in bandbreedte’’ resultaat voor $\hat{\varphi}_{n,h}(t)$, d.w.z. uniform over een zekere waaier van bandbreedtes, maar voor een vaste $t \in \mathbb{R}^d$. Natuurlijk volgt de puntsgewijze consistentie direct uit het algemeen uniform resultaat in (\star) . Maar er is gebleken dat een aantal verbeteringen kunnen worden bereikt. Er geldt namelijk dat

$$\limsup_{n \rightarrow \infty} \sup_{a_n \leq h \leq b_0} \sup_{\varphi \in \mathcal{F}} \frac{\sqrt{nh^d} |\hat{\varphi}_{n,h}(t) - \mathbb{E}\hat{\varphi}_{n,h}(t)|}{\sqrt{\log \log n}} < \infty, \quad \text{b.o.},$$

waar $a_n^d = c \log \log n/n$ of $a_n^d \geq n^{-1}(\log n)^{2/(p-2)}$ afhankelijk van het feit of de envelope functie van de klasse \mathcal{F} een eindige moment-genererende functie of een eindig p -de moment heeft. Een eerste verbetering betreft de convergentiesnelheid, die lichtjes beter is in het puntsgewijze geval. Een tweede verbetering wordt bereikt op de waaier van toelaatbare bandbreedtes, die breder is dan in het uniforme geval. Ten slotte is het verkregen resultaat algemener dan dat in het uniforme geval (\star) , hetgeen enkel toepasbaar was voor begrensde klassen of klassen waarvan de envelope functie een eindig p -de moment heeft. Wat de consistentie betreft, en in termen van convergentiesnelheden en toelaatbare bandbreedtes, impliceert dit dat begrensde klassen equivalent zijn met klassen waarvan de envelope functie een eindige moment-genererende functie heeft. Dit verrassend feit scheen nog niet bekend te zijn, zelfs niet in het klassieke geval waar de bandbreedte vast (of deterministisch) wordt genomen. Of dit ook nog waar is in de voorgaande situatie waar men uniforme convergentie op compacten beschouwt, is nog steeds een open vraag.

Een belangrijke toepassing van het bovenstaand puntsgewijs resultaat is de afleiding van de uniform in bandbreedte consistentie van de ‘‘kernel-based Hill schatter’’ voor de tail index van een Pareto-type verdeling. We

noteren deze kernschatter die gebaseerd is op n onafhankelijke Pareto-type verdeelde toevalsvariabelen met tail index τ als $\hat{\tau}_{n,h}$. In, Section 6.4 bewijzen we dat

$$\sup_{a_n \leq h \leq b_n} |\hat{\tau}_{n,h} - 1/\tau| = o_{\mathbb{P}}(1),$$

waar a_n en b_n deterministische rijen zijn die, onder andere, voldoen aan $b_n \rightarrow 0$ en $na_n \rightarrow \infty$.

Tot slot passen we [SM] toe op een veel bredere klasse van kernschatters, namelijk de klasse van de “conditionele U -statistieken”. Deze worden met $\hat{m}_{n,h,\varphi}(\mathbf{t})$ aangeduid en zijn schatters voor de “multivariate” regressiefunctie $m_{\varphi}(\mathbf{t}) = \mathbb{E}[\varphi(Y_1, \dots, Y_m) | (X_1, \dots, X_m) = \mathbf{t}]$. In Chapter 8 bewijzen we de uniform in bandbreedte consistentie van $\hat{m}_{n,h,\varphi}(\mathbf{t})$, die een gevolg is van het volgend asymptotisch resultaat :

$$\limsup_{n \rightarrow \infty} \sup_{a_n^{\gamma} \leq h < b_n} \sup_{\varphi \in \mathcal{F}} \sup_{\mathbf{t} \in I^m} \frac{\sqrt{nh^m} |\hat{m}_{n,h,\varphi}(\mathbf{t}) - \hat{\mathbb{E}} \hat{m}_{n,h,\varphi}(\mathbf{t})|}{\sqrt{|\log h| \vee \log \log n}} < \infty, \quad \text{b.o.},$$

waar $a_n = c(\log n/n)^{1/m}$, $I^m = I \times \dots \times I$ en, zoals gebruikelijk, $\gamma = 1$ of $\gamma = 1 - 2/p$ afhankelijk van het feit of de klasse \mathcal{F} begrensd is of een envelope functie heeft met een eindig p -de moment. In het bijzonder impliceert dit resultaat de uniform in bandbreedte consistentie van de Nadaraya–Watson schatter, die overeenkomt met $\hat{m}_{n,h,\varphi}(\mathbf{t})$ voor $m = 1$.