LIMBURGS
UNIVERSITAIR
CENTRUM

Faculteit Wetenschappen

Local polynomial smoothing of sparse multinomial data

Proefschrift voorgelegd tot het behalen van de graad van
doctor in de wetenschappen, groep wiskunde
aan het Limburgs Universitair Centrum
te verdedigen door

Ilse Augustyns

Promotor : Prof. dr. P. Janssen
Copromotor : Prof. dr. M. Aerts

1997

To Stefan, Tom and ...

Dankwoord

Zoals vaak in het leven, is het realiseren van iets niet toe te schrijven aan het werk van één persoon. Dit geldt natuurlijk ook bij het tot stand brengen van dit proefschrift. In de eerste plaats denk ik hierbij aan mijn promotor Paul Janssen. Als lesgever was hij reeds sinds mijn kandidatuursjaren een voorbeeld voor mij. Toen ik de mogelijkheid kreeg om bij hem te doctoreren, twijfelde ik dan ook geen moment. Door zijn enthousiasme voor het wetenschappelijk onderzoek wist hij me steeds te motiveren. Ook van mijn copromotor Marc Aerts heb ik veel hulp en aanmoedigingen gekregen. Hij was steeds bereid om mee te zoeken naar antwoorden op grote en minder grote vragen. Verder heb ik onder zijn begeleiding mijn eerste stappen in de wereld van simulaties gezet. Ik ben hen beiden dankbaar voor de aangename samenwerking gedurende de voorbije jaren. In heel de onderzoeksgroep Statistiek is trouwens een aangename werksfeer aanwezig, zowel tussen de proffen als de assistenten. Dit maakt dat ik hen allen niet alleen maar als collega's, maar ook als vrienden beschouw. Natuurlijk verdienen ook Stefan en Tom een plaats in dit dankwoordje. Stefan vooral omdat hij steeds mijn steun en toeverlaat was, al was het soms alleen via email. Zonder dat Tom het beseft heb ik ook steun van hem gekregen. De aangename kanten van het moederschap zijn immers een extra stimulans om alles op een positieve manier te bekijken.

Contents

Samenvatting

In dit proefschrift richten we onze aandacht naar schaars gevulde multinomiale gegevens. Wij beperken ons tot het bestuderen van schatters voor de ongekende celkansen in een welbepaald asymptotisch kader. In de klassieke asymptotiek voor multinomiale gegevens laat men de steekproefgrootte n oneindig groot worden terwijl het aantal cellen in de tabel, en de bijhorende celkansen, ongewijzigd blijven. Dit impliceert automatisch dat het verwacht aantal observaties in elke cel van de tabel oneindig groot wordt, zodat zo'n asymptotisch kader niet geschikt is om schaars gevulde tabellen te bestuderen. In het 'schaars' asymptotisch kader daarentegen, laten we naast de steekproefgrootte ook het aantal cellen toenemen, zodat het verwacht aantal observaties in elke cel van de tabel niet noodzakelijk oneindig groot wordt.

De meest gebruikte schatters voor de celkansen zijn de frequentieschatters, voor multinomiale gegevens zijn dit immers de maximum likelihood schatters. Naast deze schatters werden in de literatuur ook alternatieve schatters voor de celkansen bestudeerd. Voorbeelden hiervan kunnen o.a. gevonden worden in Good (1965), Fienberg, Bishop en Holland (1975, Hoofdstuk 4), Aitchison en Aitken (1976), Hall (1981) en Bowman, Hall en Titterington (1984). Het voornaamste dat we uit deze werken kunnen besluiten is dat, in het klassieke asymptotische kader, deze alternatieve schatters equivalent zijn met met de frequentieschatters (in die zin dat ze dezelfde asymptotische verdeling hebben). In Fienberg, Bishop en Holland (1975, p. 416) en Titterington en Bowman (1985) wordt er aan de hand van eindige steekproefresultaten geïllustreerd dat hun schatters toch een kleinere gemiddelde som van kwadratische afwijkingen (MSSE) kunnen hebben dan de frequentieschatters.

Fienberg, Bishop en Holland (1975, Sectie 12.3.1) bestuderen voor hun schatters ook de MSSE in het schaars asymptotische kader. Zij tonen aan dat binnen deze asymptotiek de leidende term van de MSSE van de frequentieschatters groter is dan die van hun schatters. Deze laatsten zijn geïnspireerd op Bayesiaanse ideeën en kunnen gebruikt worden zowel voor geordende als voor niet-geordende multinomiale gegevens. De voorstellen van Hall (1981) en Bowman, Hall en Titterington (1984) daarentegen zijn gebaseerd op ideeën van kernschatters. Deze methode leent informatie bij de "buren" om schatters te definiëren. Hierdoor is deze techniek alleen zinvol voor geordende multinomiale gegevens.

Hall en Titterington (1987) en Burman (1987a) bestuderen de MSSE van kernschatters in het schaars asymptotische kader. Bovendien tonen Hall en Titterington (1987) aan dat er, onder bepaalde regulariteitsvoorwaarden, een optimale convergentiesnelheid naar nul bestaat voor de MSSE van schatters voor de celkansen. De frequentieschatters en de schatters van Fienberg, Bishop en Holland (1975) halen deze optimale convergentiesnelheid niet, terwijl zowel de kernschatters van Hall en Titterington (1987) als die van Burman (1987a) die convergentiesnelheid wel halen. Om deze snelheid te bereiken hebben beiden echter extra voorwaarden nodig op het gedrag van de ongekende celkansen aan de rand van de tabel. Ditzelfde probleem is aanwezig bij de kernschattingsmethode in andere situaties, bijvoorbeeld in de regressiecontext. Om deze zogenaamde randproblemen op te lossen zijn er een aantal methoden beschikbaar. Eén ervan is door gebruik te maken van randgecorrigeerde kernen. Dit zijn speciaal geconstrueerde kernfuncties die alleen bij het schatten in het randgebied moeten gebruikt worden. Dong en Simonoff (1994) bestuderen deze techniek in de context van schaarse multinomiale gegevens en tonen aan dat de optimale convergentiesnelheid van de MSSE inderdaad bereikt wordt zonder extra randvoorwaarden.

Een andere methode die de randproblemen van kernschatters oplost, is gebaseerd op lokale veeltermbenaderingen. Een bandbreedteparameter bepaalt hoeveel buren er lokaal een bijdrage kunnen leveren tot de uiteindelijke schatters, en speelt een cruciale rol in de hele procedure. In Sectie 1.2.2 geven we een beknopt overzicht van relevante eigenschappen van de lokale veeltermbenaderingsmethode in de regressiecontext en we vergelijken deze met analoge resultaten voor de klassieke kernschatters. Eén van de attractieve eigenschappen van de lokale veeltermbenaderingsmethode is dat het de randproblemen van de klassieke kernschatters automatisch opvangt, d.w.z. de vorm van de schatter past zichzelf aan wanneer er in het randgebied geschat wordt.

Het is deze techniek van lokale veeltermbenaderingen die wij in Sectie 1.3 gebruiken om onze schatters voor de celkansen in een schaarse tabel te definiëren. Zoals hierboven vermeld, is de idee van lokale veeltermbenaderingen gebaseerd op voldoende gladheid van de te schatten ongekende "functie", hier de vector $\boldsymbol{p} = (p_1, \ldots, p_k)^T$ van celkansen. In de schaarse asymptotiek neemt het aantal cellen in de tabel (k) toe bij een toename van de steekproefgrootte n, d.w.z. we beschouwen $k \equiv k(n)$ als functie van n. Ook de vector $\boldsymbol{p}$ is dan, via k, een functie van n. In de schaars asymptotische studie van onze schatters vereenvoudigen wij deze structuur voor $\boldsymbol{p}$ door het bestaan van een onderliggende dichtheid $f(\cdot)$ te veronderstellen, die de celkansen bepaalt door

$$p_i = \int_{\frac{i-1}{k}}^{\frac{i}{k}} f(u)\,du, \qquad i = 1, \ldots, k,$$

zodat we steeds onze veranderende vector $\boldsymbol{p}$ kunnen uitdrukken m.b.v. een vaste functie $f(\cdot)$ (zie ook Santner en Duffy (1989, p. 60)). Bovendien kunnen we de vereiste gladheid voor de vector $\boldsymbol{p}$ bekomen door gladheidsvoorwaarden te veronderstellen op deze onderliggende dichtheid $f(\cdot)$. Naast het definiëren van onze lokale veeltermschatters voor de celkansen, bestuderen we in Sectie 1.3 reeds hun schaars asymptotische vertekening en variantie.

In Hoofdstuk 2 onderzoeken we schaarse consistentie van zowel de frequentieschatters als van onze lokale veeltermschatters. Van schatters $\boldsymbol{P}^* = (P_1^*, \ldots, P_k^*)^T$ voor $\boldsymbol{p} = (p_1, \ldots, p_k)^T$ wordt gezegd dat ze schaars consistent zijn als

$$\sup_{1 \leq i \leq k} \left| \frac{P_i^*}{p_i} - 1 \right| \xrightarrow{b.o.} 0, \quad \text{als } n \to \infty.$$

We illustreren eerst aan de hand van een voorbeeld in welke situatie de frequentieschatters niet schaars consistent zijn, en gaan vervolgens op zoek naar voldoende voorwaarden voor schaarse consistentie van de frequentieschatters. In concrete voorbeelden worden deze voldoende voorwaarden geïnterpreteerd in termen van schaarstegraad (d.i. hoe we k bekijken als functie van n).

Ook voor onze lokale veeltermschatters gaan we op zoek naar voldoende voorwaarden voor schaarse consistentie, welke in dezelfde concrete voorbeelden als voor de frequentieschatters worden vertaald naar voorwaarden op de schaarstegraad. Uit deze studie blijkt dat onze schatters wel schaars consistent zijn in die situatie waar de frequentieschatters het niet zijn. In een andere situatie echter, kunnen we voor de frequentieschatters in eerste instantie een hogere schaarstegraad toelaten dan voor de lokale veeltermschatters. Voor een specifieke familie van onderliggende dichtheden $f(\cdot)$ kunnen we het resultaat voor de lokale veeltermschatters verscherpen, zodat toch dezelfde schaarstegraad als voor de frequentieschatters kan toegelaten worden. In Hoofdstuk 2 wordt een inzicht gegeven omtrent schaarse consistentie, maar het voorziet niet in een volledige karakterisatie van schaarse consistentie. De belangrijke vraag of er een optimale convergentiesnelheid voor schaarse consistentie van schatters voor celkansen bestaat, en of zo'n resultaat de bevindingen van Hoofdstuk 2 kan verklaren, blijven nog onopgelost.

Zoals reeds vroeger vermeld, bestaat er wel een optimaliteitsresultaat voor de convergentiesnelheid van de MSSE van schatters voor de celkansen. In Hoofdstuk 3 bespreken we dit resultaat uitvoerig en bestuderen we de MSSE van de lokale veeltermschatters. Volledig in overeenstemming met de gekende resultaten in de regressiecontext, blijkt dat de optimale convergentiesnelheid bereikt wordt wanneer de graad van de lokale veeltermbenadering oneven wordt genomen. Bij een even graad kunnen enkel bijkomende randvoorwaarden op de onderliggende dichtheid $f(\cdot)$ het behalen van deze convergentiesnelheid garanderen. Aan de hand van een kleine simulatiestudie illustreren we de asymptotische MSSE resultaten. We gaan ook dieper in op het belangrijke bandbreedtekeuzeprobleem, waarbij we twee veel gebruikte selectiemethoden bespreken en illustreren.

In Hoofdstuk 4 bestuderen we een centrale limietstelling voor de statistische grootheid $\mathrm{SSE}(\boldsymbol{P}^*) = \sum_{i=1}^{k}(P_i^* - p_i)^2$ in het geval $\boldsymbol{P}^*$ de frequentieschatters of de lokale veeltermschatters zijn. Voor deze laatsten moet het resultaat opgesplitst worden volgens de snelheid waarmee de bandbreedte naar nul convergeert. We willen vooral de nadruk leggen op het feit dat deze opsplitsing alle mogelijkheden dekt. Dit in contrast met het analoge resultaat van Burman (1987b) voor de klassieke kernschatters, waar de bandbreedte aan een specifieke voorwaarde moet voldoen, welke niet geldig is voor de optimale bandbreedte.

Tot nu toe zijn de lokale veeltermschatters voor de celkansen enkel besproken voor één-dimensionale multinomiale gegevens. In Hoofdstuk 5 geven we een uitbreiding naar meerdere dimensies. Terwijl in Hoofdstukken 1–4 de graad van de lokale veeltermbenadering vrij mocht gekozen worden (maar bij voorkeur oneven), beperken we ons nu tot lokale lineaire schatters. In dit hoofdstuk staat vooral de uitbreiding van de bandbreedteparameter centraal. In één dimensie heeft een cel alleen maar buren links en rechts waar informatie kan geleend worden, terwijl in meerdere dimensies een cel in verschillende richtingen buren heeft. Er zijn verscheidene opties mogelijk om de richtingen volgens welke deze buren geselecteerd worden, alsook de hoeveelheid buren in elke richting, te beschrijven. In het eenvoudigste geval vallen de richtingen waardoor de buren bepaald worden samen met de richtingen, bepaald door de coördinaatsassen waarin we het probleem bekijken, en gebruiken we in elke richting eenzelfde aantal buren. Een eerste veralgemening bestaat erin om in de verschillende richtingen een verschillend aantal buren te gebruiken. Wij beschouwen de veralgemening waar de richtingen niet noodzakelijk samenvallen met de richtingen bepaald door de coördinaatsassen. We bestuderen de asymptotische benadering van de MSSE van de lokale lineaire schatter, en illustreren m.b.v. simulaties de winst die bekomen wordt door de buren op zo'n algemeen mogelijke manier te kiezen. Verder breiden we de centrale limietstelling voor lokale veeltermschatters van Hoofdstuk 4 ook uit voor de meer-dimensionale lokale lineaire schatters.

In Hoofdstuk 6 keren we terug naar één-dimensionale tabellen. Gebaseerd op een heuristische argument, doen we een voorstel om de gegevens zelf een geschikte bandbreedte te laten selecteren. We willen hierbij vooral de nadruk leggen op het feit dat we ons zo weinig mogelijk door asymptotische overwegingen willen laten leiden. Na een eerste simulatiestudie kunnen we een positieve balans opmaken voor ons voorstel, wat verder onderzoek naar de methode aanmoedigt.

Chapter 1

Sparse tables and local polynomial smoothing

1.1 Introduction

It is a widely used approach to present and record data in the form of a table. The cells in the contingency table represent the different cross-classifications of the recorded categorical variables. The numbers in the cells are the frequency counts of the outcomes. These counts are considered as random variables having a certain sampling distribution. For categorical data one usually assumes a Poisson, multinomial or product multinomial sampling scheme (see Chapter 3 in Agresti (1990) for more details). We now give some typical examples of multinomial data (one and multi-dimensional).

Table 1.1 displays data on the number of boys among the first four children in 3343 Swedish families having at least four children (data source: Edwards and Fraccaro (1960)).

Number of boys	0	1	2	3	4
Number of families	183	789	1250	875	246

Table 1.1: *Swedish family data.*

Table 1.2 contains data on the monthly salary of 147 nonsupervisory female employees holding a Bachelors (but no higher) degree who were practicing mathematics

i	*Salary*	n_i	i	*Salary*	n_i	i	*Salary*	n_i
1	951-1050	5	11	1951-2050	6	20	2851-2950	5
2	1051-1150	1	12	2051-2150	9	21	2951-3050	4
3	1151-1250	0	13	2151-2250	5	22	3051-3150	2
4	1251-1350	5	14	2251-2350	12	23	3151-3250	1
5	1351-1450	2	15	2351-2450	7	24	3251-3350	2
6	1451-1550	10	16	2451-2550	3	25	3351-3450	0
7	1551-1650	5	17	2551-2650	10	26	3451-3550	1
8	1651-1750	10	18	2651-2750	4	27	3551-3650	1
9	1751-1850	10	19	2751-2850	6	28	3651-3750	1
10	1851-1950	20						

Table 1.2: *Salary data.*

i	*Days*	n_i	i	*Days*	n_i	i	*Days*	n_i
1	0-30	18	20	571-600	1	38	1111-1140	0
2	31-60	14	21	601-630	0	39	1141-1170	0
3	61-90	9	22	631-660	0	40	1171-1200	0
4	91-120	8	23	661-690	1	41	1201-1230	1
5	121-150	6	24	691-720	0	42	1231-1260	0
6	151-180	4	25	721-750	0	43	1261-1290	0
7	181-210	6	26	751-780	1	44	1291-1320	1
8	211-240	7	27	781-810	0	45	1321-1350	0
9	241-270	1	28	811-840	0	46	1351-1380	1
10	271-300	6	29	841-870	0	47	1381-1410	0
11	301-330	7	30	871-900	0	48	1411-1440	0
12	331-360	5	31	901-930	1	49	1441-1470	0
13	361-390	5	32	931-960	0	50	1471-1500	0
14	391-420	0	33	961-990	0	51	1501-1530	0
15	421-450	0	34	991-1020	0	52	1531-1560	0
16	451-480	2	35	1021-1050	0	53	1561-1590	0
17	481-510	1	36	1051-1080	0	54	1591-1620	1
18	511-540	1	37	1081-1110	0	55	1621-1650	1
19	541-570	1						

Table 1.3: *Mine explosions data.*

or statistics in 1981 (data source: Department of Energy (1982)). Despite the continuous nature of the response variable, these data were given in a discretized form in the original data source.

Table 1.3 gives counts that correspond to a discretization into 55 cells of 109 time intervals between explosions in mines involving more than ten men killed in Great-Britain from December 8, 1875 to May 29, 1951 (data source: original form Maguire, Pearson and Wynn (1952), in discretized form: Simonoff (1983)).

Table 1.4 displays the salary data of Table 1.2 in a 12 × 10 contingency table, where the 147 respondents are now cross-classified according to their salary and the number of years since receiving their degree (data source : Simonoff (1987)).

Table 1.5 is a 7 × 7 cross-classification of the responses of 55 first year students, at New York University's Stern School of Business in 1991, to questions about the importance of Statistics and Economics in business education. Responses were coded on a 7-point scale ranging from "completely useless"to "absolutely crucial"(note that no students rated Statistics "completely useless")(data source: Simonoff (1995a)).

	Years since degree									
Salary	0-	3-	6-	9-	12-	15-	18-	24-	30-	>
	2	5	8	11	14	17	23	29	35	35
951-1150	5	0	1	0	0	0	0	0	0	0
1151-1350	2	1	0	0	0	0	0	2	0	0
1351-1550	5	1	3	2	0	0	1	0	0	0
1551-1750	5	5	2	1	0	1	0	1	0	0
1751-1950	9	9	5	0	2	2	1	1	1	0
1951-2150	3	5	2	1	2	0	1	0	0	1
2151-2350	0	1	4	3	2	1	3	0	2	1
2351-2550	0	0	4	0	1	2	2	0	0	1
2551-2750	0	0	2	2	0	5	1	2	1	1
2751-2950	0	0	1	0	0	1	4	0	2	3
2951-3150	1	0	1	0	1	1	1	0	1	0
3151-3750	0	0	0	0	0	0	5	0	0	1

Table 1.4: *Salary data.*

		Economics						
		1	2	3	4	5	6	7
Statistics	1	0	0	0	0	0	0	0
	2	0	1	0	0	0	1	0
	3	0	0	0	1	0	0	0
	4	0	0	3	6	4	0	0
	5	0	0	1	4	7	4	0
	6	1	0	0	2	6	10	1
	7	0	0	0	0	0	2	1

Table 1.5: *MBA survey data.*

		Urban Region			
		Years of Schooling			
Sex	Age	Less than 4	4-7	8-10	More than 10
Male	16-29	631571	4381454	7557356	915307
	30-49	1176623	4174797	5384310	1055949
	50 or over	2376214	1807879	1187537	501741
Female	16-29	664308	2843925	9180042	1274867
	30-49	3825831	4481894	6873450	1070910
	50 or over	7432953	1888644	1312793	275948

		Rural Region			
		Years of Schooling			
Sex	Age	Less than 4	4-7	8-10	more than 10
Male	16-29	1719046	5266179	5718852	153669
	30-49	2386652	4297528	3050642	316853
	50 or over	5378665	1862132	376084	68963
Female	16-29	2536078	4246961	5886728	232464
	30-49	7181386	4653138	3089527	241507
	50 or over	13089113	1071037	207498	24283

Table 1.6: *Soviet Union population data.*

Table 1.6 gives a classification of the population in the Soviet Union in 1959 according to:

- schooling : 4 categories
- sex : 2 categories
- age : 3 categories
- region : 2 categories

(data source: Selesnick (1970)).

The data sets presented in Tables 1.2–1.5 have in common that the total number of observations is moderate in comparison to the total number of cells in the table. As a result the counts are quite sparsely distributed among the cells, and even empty cells arise. In this thesis the focus is on such sparse tables and the main objectives are:

(i) to propose estimators for the multinomial cell probabilities

(ii) to study asymptotic properties of these estimators

(iii) to illustrate the finite sample performance based on simulations and real data sets.

In the subsequent chapters we restrict attention to the basic problem of finding good estimators for cell probabilities in sparse tables. Other statistical questions, e.g. testing independence between categorical variables, are not discussed in this thesis.

It is well known that frequency estimators are maximum likelihood estimators and that they are optimal, i.e., they are unique minimum variance unbiased estimators. For sparse tables, however, the visual impression of the frequency estimates shows some roughness that we do not expect to be there. We therefore want to smooth the observed counts. We will make an asymptotic comparison of the proposed estimators to the frequency estimators.

To develop asymptotic results one should realize that sparseness requires an appropriate description of the type of asymptotics we have in mind. The usual asymptotic approach to handle multinomial data is to consider the number of cells in the table as being fixed and to let the sample size n tend to infinity. In this classical approach the expected number of counts in each cell becomes large as n increases. Hence, this type of asymptotics leaves no room for sparse tables, since they typically have small cell counts. A way to incorporate small cell counts in an asymptotic study is to say that the number of cells in the table grows with the number of observations, i.e., we assume $k \equiv k(n) \to \infty$ as $n \to \infty$. We refer to this approach as the sparse asymptotic framework. Fienberg and Holland (1973) and Fienberg et al. (1975, Chapter 12) introduced the idea of sparse asymptotics.

Not only in the context of sparse tables, alternatives to the frequency estimators for the cell probabilities have been proposed. In the general context of multinomial cell probability estimation, Good (1965) introduced the concept of smoothing. From then on, several smoothing methods have been investigated. In the remaining part of this section we mention some of these proposals. For a detailed survey see Simonoff (1995a). The asymptotic discussion in the first part of the overview will be in the usual asymptotic framework (k fixed and $n \to \infty$).

Fienberg et al. (1975, p. 404–410) use Bayesian ideas to introduce estimators for the cell probabilities, and they also propose pseudo Bayesian estimators. Kernel based estimators in the context of multinomial cell probability estimation were first proposed by Aitchison and Aitken (1976). Their estimator has the form of a weighted average of the frequency estimators, a high weight is given to the frequency of the cell one wants to estimate, and a low, constant weight to all other frequencies. This has the effect of shrinking the estimators towards a uniform distribution. Brown and Rundell (1985) have shown that a shrinkage factor exists such that the mean sum of squared errors (MSSE) of the kernel estimator is smaller than that of the frequency estimator. Since in the construction of the estimator only two weight factors are used, this method applies to multinomial data where the categories are not necessarily ordered.

For ordered tables a more sophisticated way of smoothing is possible, more specific, a smoothing approach can take the ordering into account by borrowing information from "neighboring"cells. Hall (1981), Bowman et al. (1984) introduce kernel smoothers that minimize the MSSE.

Within the framework of standard asymptotics it has been shown that frequency estimators, Bayesian estimators and kernel smoothers are asymptotically equivalent, in the sense that they have the same asymptotic distribution. In finite samples, however, it is possible for the smoothed estimators to have better performance. Bishop et al. (1975, p. 416) present small sample results for problems in which (pseudo) Bayesian estimators for cell probabilities have smaller MSSE than the frequency estimators. Titterington and Bowman (1985) performed a small Monte Carlo study to compare different smoothing techniques for multinomial data. This study also illustrates the benefit of smoothing the frequency counts in order to reduce the MSSE.

In the sparse asymptotic framework one can define estimators that have better performance than the frequency estimators. Two types of results are available in the sparse asymptotic setup. One type of results concerns consistency. A generic sequence of estimators $\boldsymbol{P}^* = (P_1^*, \dots, P_k^*)^T$ for the cell probabilities $\boldsymbol{p} = (p_1, \dots, p_k)^T$ is defined to be sparse consistent if

$$\sup_{1 \leq i \leq k} \left| \frac{P_i^*}{p_i} - 1 \right| \xrightarrow{a.s.} 0, \text{ as } n \to \infty.$$

In general the frequency estimators may fail to be sparse consistent. A simple illustration of this will be given in Chapter 2. Simonoff (1983) proves that, under smoothness conditions on the sequence of true underlying cell probabilities, the maximum penalized likelihood estimators are sparse consistent. For more details we refer to Section 2.2.

A second type of results concerns asymptotic approximations for the MSSE of the estimators. Bishop et al. (1975, p. 410–413) show that, in the sparse asymptotic framework, their pseudo Bayesian estimator has smaller leading term for the MSSE than the frequency estimators and than the generally accepted practice of adding 1/2 to the count in each cell. Kernel smoothing methods for ordered multinomial data are investigated by Burman (1987a) and Hall and Titterington (1987). The latter obtain an optimal rate of convergence to zero for the MSSE, under smoothness conditions on the underlying cell probabilities. This optimal rate of convergence is neither achieved by the frequency estimators, nor by the pseudo Bayesian estimators of Bishop et al. (1975). Hall and Titterington (1987) introduce kernel smoothers for which this optimal rate is achieved, but they require rather restrictive conditions on the behavior of the underlying cell probabilities at the boundaries of the table. Burman (1987a) shows that these boundary conditions can be somewhat weakened. Dong and Simonoff (1994) show that the stringent conditions in Burman (1987a) become superfluous if boundary corrected kernel estimators for the cell probabilities are used.

A smoothing technique in regression that received much attention in recent years is local polynomial smoothing. One reason for its popularity is that the local polynomial estimator corrects for boundary problems in an automatic way. We will investigate this local polynomial smoothing method in the context of the estimation of cell probabilities for sparse tables. In Section 1.2 we first give a short discussion on kernel smoothing in the regression context. The local polynomial estimators for the cell probabilities of a sparse one-dimensional table are defined in Section 1.3. Some of its basic properties are presented there as well.

The main structure of this thesis is as follows. After the introduction of our estimators for the cell probabilities, we give in Chapter 2 a detailed discusssion of the sparse asymptotic consistency of frequency estimators and local polynomial estimators. The behavior of the mean sum of squared errors of the local polynomial cell probability estimator is studied in Chapter 3. In Chapter 4 we prove a central limit result for the sum of squared errors for the local polynomial smoother, and for the frequency estimator. Generalizations for the local polynomial smoothers to multi-dimensional tables are given in Chapter 5. Finally, in Chapter 6 we propose a method to select the bandwidth, a smoothing parameter which occurs in the construction of the estimators.

1.2 Kernel smoothing in regression

In regression theory one is interested in determining an appropriate functional relationship between the mean of a response variable Y and a predictor x, based on observations $(x_1, Y_1), \ldots, (x_n, Y_n)$. A typical regression model is

$$Y_i = m(x_i) + \sigma(x_i)\epsilon_i, \quad i = 1, \ldots, n,$$

where the errors ϵ_i are i.i.d. random variables having mean zero and variance one.

Parametric regression assumes that the functional form of the regression function $m(\cdot)$ is known, e.g., the simple linear regression model with $m(x) = \beta_0 + \beta_1 x$. The problem of finding $m(\cdot)$ then reduces to the estimation of a finite number of parameters. The choice of the parametric model depends on the situation, and can be based on scientific reasons or on previous experience with data sets of similar type. If one chooses a parametric form that does not fit the data properly, the conclusions obtained from the analysis are not reliable.

Nonparametric regression removes the restriction that the true regression function belongs to some parametric family. By making relatively weak assumptions on the smoothness of the regression function it is possible to let the data tell what the pattern is. Many methods for obtaining nonparametric estimators exist. Here we will restrict attention to kernel methods.

1.2.1 Classical kernel estimators

Under the sole assumption that $m(\cdot)$ is a smooth curve, the points in the neighborhood of x carry information about $m(x)$. Since points close to x are more informative it seems natural to consider a weighted average as estimator. Typically the weights are choosen in such a way that they are higher for points closer to x. The general form of a locally weighted average is

$$\hat{m}(x) = \sum_{i=1}^{n} w_{ni}(x; x_1, \ldots, x_n) Y_i = \sum_{i=1}^{n} w_{ni}(x) Y_i, \tag{1.1}$$

with $\sum_{i=1}^{n} w_{ni}(x) = 1$. Note that (1.1) is a linear combination of the responses Y_i, we therefore say that a weighted average is a linear smoother.

The Nadaraya-Watson weights, introduced independently by Nadaraya (1964) and Watson (1964), are given by

$$w_{ni}(x) = \frac{h^{-1}K((x_i - x)/h)}{\sum\limits_{j=1}^{n} h^{-1}K((x_j - x)/h)},$$

where $K(\cdot)$ is a symmetric density function, called a kernel function, and $h > 0$ a bandwidth controling the width of the local neighborhood. Since the magnitude of the bandwidth determines the smoothness of the resulting estimated function $\hat{m}(x)$, one often uses the term smoothing parameter as well. This smoothing parameter typically decreases with n. Finding the proper amount of smoothing is a major problem in nonparametric smoothing. In Section 3.3 and Chapter 6 we pursue this problem further. To simplify notation we suppress the dependence of the bandwidth sequence on n and write h for $h(n)$.

Assuming the data are sorted according to the x-variable, the Gasser-Müller weights (Gasser and Müller, 1979) are

$$w_{ni}(x) = \int\limits_{s_{i-1}}^{s_i} \frac{1}{h} K\left(\frac{u - x}{h}\right) du,$$

where $s_i = \frac{x_i + x_{i+1}}{2}, x_0 = -\infty, x_{n+1} = +\infty$.

Figures 1.1 and 1.2 illustrate on a tutorial example how both methods assign their weights to each design point. These tutorial figures are taken from Seifert and Gasser (1996). The data consist of 4 collinear observations. The Epanechnikov kernel $K(u) = 0.75(1 - u^2)\,\mathbb{1}\{|u| \leq 1\}$ is used to construct the estimators. For both systems of weights the kernel function is first transformed to the interval $[x - h, x + h]$, i.e., use the rescaled and shifted kernel function $K_{x,h}(u) = h^{-1}K\left((u - x)/h\right)$. The middle part of Figure 1.1 visualizes the construction of the Nadaraya-Watson weights. First the transformed kernel function is evaluated at the design points in the interval $[x - h, x + h]$, i.e., compute $K_{x,h}(x_i)$ (the vertical lines in the middle part of the

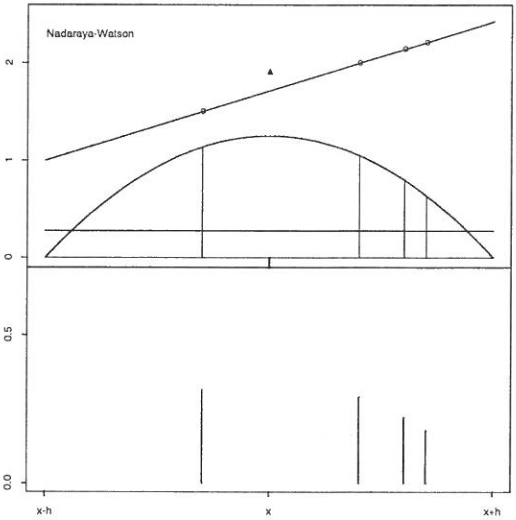

Figure 1.1: *Tutorial example of the Nadaraya-Watson estimator. The data are indicated by* ○ *and the resulting estimate by* ▲. *The middle part illustrates the construction of the weights while the bottom part shows the actual weights.*

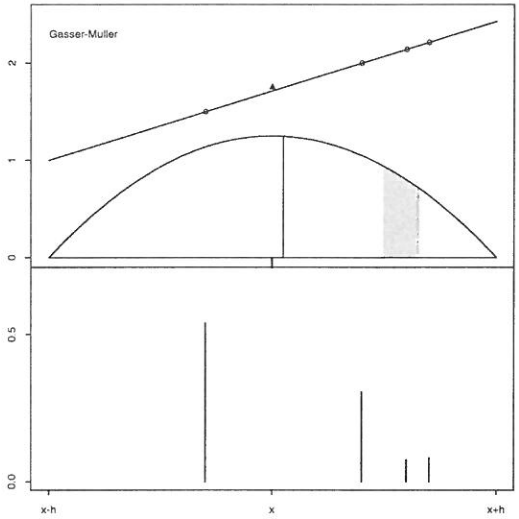

Figure 1.2: *Tutorial example of the Gasser-Müller estimator. The data are indicated by* ○ *and the resulting estimate by* ▲. *The middle part illustrates the construction of the weights while the bottom part shows the actual weights.*

figure), then rescale by the normalizing factor $\left(\sum_{i=1}^{n} K_{x,h}(x_i)\right)^{-1}$ (indicated by the horizontal line). The actual weights are shown at the bottom part of the figure. The resulting estimate at x is indicated by the triangle in the figure.

The weights of the Gasser-Müller estimator (Figure 1.2) are obtained by integrating the transformed kernel function $K_{x,h}(u)$ between averages of subsequent design points (the middle part of the plot). The actual weights are shown at the bottom part of the figure. The resulting estimate at x is indicated by the triangle.

For an asymptotic investigation of the kernel type regression estimators one has to be more specific about the nature of the design of the covariate. There is the setting of a random design, where the covariate is considered to be a random variable X with density $f_X(\cdot)$. This design is suitable for cases where the covariates are beyond the control of the experimenter. In other cases, the design points are prescribed by data analysts, and are treated as fixed points. To present the asymptotic results for fixed designs, one often uses the following device of Sacks and Ylvisaker (1970) (see also Müller (1988, p. 9)). The design points are defined through a "design density" $f_X(\cdot)$ in the following way,

$$x_i = F_X^{-1}(i/n) \quad \text{with} \quad F_X(x) = \int_{-\infty}^{x} f_X(y)\, dy. \tag{1.2}$$

We will consider this approach to present the asymptotic results in this section. Denote $\hat{m}_{NW}(x)$ for the Nadaraya-Watson and $\hat{m}_{GM}(x)$ for the Gasser-Müller estimator.

Theorem 1.1 (Müller (1988 p. 29), Chu and Marron (1991))
Assume the following conditions

(i) $K(\cdot)$ *is a symmetric kernel with support [-1,1],*
$K(\cdot)$ *is bounded above 0 on [-1/2,1/2] and* $K(\cdot)$ *has bounded derivative,*

(ii) $h \to 0, nh^3 \to \infty$, *as* $n \to \infty$,

(iii) $m''(\cdot)$ *is continuous at* x,

(iv) $\sigma^2(\cdot) \equiv \sigma^2$,

(v) $f_X(x) > 0$, $f_X(\cdot)$ *is Lipschitz continuous and* $f'_X(\cdot)$ *is continuous at* x.

Then, with $\mu_2(K) = \int_{-1}^{1} u^2 K(u)\, du$ *and* $R(K) = \int_{-1}^{1} K^2(u)\, du$, *we have*

$$E\left(\hat{m}_{NW}(x)\right) - m(x) = \left(m''(x) + 2m'(x)\frac{f'_X(x)}{f_X(x)}\right)\frac{h^2}{2}\mu_2(K) + o(h^2),$$
$$E\left(\hat{m}_{GM}(x)\right) - m(x) = m''(x)\frac{h^2}{2}\mu_2(K) + o(h^2)$$

and

$$\text{Var}(\hat{m}_{NW}(x)) = \text{Var}(\hat{m}_{GM}(x)) = \frac{\sigma^2}{f_X(x)nh}R(K) + o\left(\frac{1}{nh}\right).$$

So far we assumed implicitly that $m(\cdot)$ and $f_X(\cdot)$ have unbounded domain. Asymptotic properties of both estimators, however, change when this unboundedness is not satisfied. Assume w.l.o.g. that $f_X(\cdot)$ is defined on [0,1]. For points in the interior region, $x \in [h, 1-h]$, the above theorem remains valid. For boundary points, $x \in [0, h)$ or $x \in (1-h, 1]$, the asymptotic expressions are different. Main reason for this is that for left (resp. right) boundary points the local neighborhood

$[x-h, x+h]$ around x only contains design points in $[0, x+h]$ (resp. $[x-h, 1]$) and the weight assignment mechanism is not able to adapt to this. The main implication is that in the boundary region the order of the bias becomes $O(h)$, while it is $O(h^2)$ in the interior. The theoretical result describing this phenomenon for the Nadaraya-Watson estimator is given in Section 1.2.2, for the Gasser-Müller estimator see Müller (1988).

Several proposals have been made to solve this boundary problem. Two well known approaches are boundary corrected kernel methods and the reflection principle. Boundary corrected kernel methods modify the kernel in the estimation procedure in the boundary region in order to reduce the bias. Gasser and Müller (1979) and Gasser, Müller and Mammitzsch (1985) discuss this approach. The reflection method is introduced by Schuster (1985) in the density estimation context. Artificial data are created beyond the support by reflecting the actual data points. Their estimator is then based on this augmented data set. Hall and Wehrly (1991) discuss a modified version in the regression context.

In the next section we will introduce local polynomial estimators. For a general study of local polynomial estimators see Ruppert and Wand (1994) and Fan and Gijbels (1996).

1.2.2 Local polynomial estimators

Local polynomial regression estimators are obtained by assuming, locally around x, a polynomial regression model. Suppose the regression function $m(\cdot)$ has derivatives up to a certain order p. By a Taylor approximation we then have

$$m(z) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j \equiv \sum_{j=0}^{p} \beta_j(x)(z-x)^j,$$

for z in a neighborhood of x. We then use, locally, the weighted least squares method to obtain estimators for $m(x)$, i.e., with $\boldsymbol{\beta}_x = (\beta_0(x), \ldots, \beta_p(x))^T$,

$$\underset{\boldsymbol{\beta}_x}{\text{minimize}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j(x)(x_i - x)^j \right)^2 w_{ni}(x).$$

Local regression is a natural local application of parametric fitting, so natural that, already in the 19th century, it arose independently at different points in time and in different countries. A nice historical review can be found in Cleveland and

Loader (1996). References therein go back to 1870 and even work dating from 1829 is mentioned.

We consider a local polynomial regression model with kernel weights

$$w_{ni}(x) = \frac{1}{h} K\left(\frac{x_i - x}{h}\right),$$

where $K(\cdot)$ is the kernel function and $h > 0$ is the bandwidth. The minimization problem then becomes:

$$\underset{\boldsymbol{\beta}_x}{\text{minimize}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j(x)(x_i - x)^j\right)^2 \frac{1}{h} K\left(\frac{x_i - x}{h}\right). \tag{1.3}$$

Local polynomial regression using kernel weights was introduced by Stone (1977), Cleveland (1979) and Katkovnik (1979). In the last few years the idea of local least squares regression has received a lot of attention. Work on this include papers by Fan (1992,1993), Fan and Gijbels (1992), Hastie and Loader (1993), Ruppert and Wand (1994) and many others. A good textbook reference is Fan and Gijbels (1996).

Let $\hat{\boldsymbol{\beta}}_x = (\hat{\beta}_0(x), \ldots, \hat{\beta}_p(x))^T$ denote the minimizer of (1.3). The estimator for $m(x)$ then becomes $\hat{m}(x; p, h) = \hat{\beta}_0(x)$. The other parameters $\hat{\beta}_j(x)(j = 1, \ldots, p)$ provide estimators for the derivatives of the regression function $m(\cdot)$ at x up to order p. Since the polynomial approximation only applies locally, the estimation procedure is also local and must be redone when estimating $m(\cdot)$ at another point. Because of this local modeling, the degree p of the polynomial approximation can be kept small, in contrast with the global modeling, where higher order polynomials are required to control the bias (if possible at all, see the illustrations in Fan and Gijbels (1996, p. 2-5)).

The solution to the minimization problem (1.3) is obtained from weighted least squares theory. Denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{W}_x = \text{diag}_{1 \le i \le n}\left(h^{-1} K((x_i - x)/h)\right)$ and

$$\boldsymbol{X}_x = \begin{pmatrix} 1 & x_1 - x & \ldots & (x_1 - x)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n - x & \ldots & (x_n - x)^p \end{pmatrix}$$

the $n \times (p+1)$ design matrix. With this notation the least squares problem (1.3) can be rewritten as

$$\underset{\boldsymbol{\beta}_x}{\text{minimize}} (\boldsymbol{Y} - \boldsymbol{X}_x \boldsymbol{\beta}_x)^T \boldsymbol{W}_x (\boldsymbol{Y} - \boldsymbol{X}_x \boldsymbol{\beta}_x) \tag{1.4}$$

with solution

$$\hat{\boldsymbol{\beta}}_x = (\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y}. \tag{1.5}$$

The estimator for $m(x)$ is

$$\hat{m}(x; p, h) = \hat{\beta}_0(x) = \boldsymbol{e}_1^T (\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y}, \tag{1.6}$$

where $\boldsymbol{e}_1^T$ is the $(p+1)$ vector $(1, 0, \ldots, 0)$. From this expression it is clear that local polynomial estimators are linear smoothers of the form (1.1). The coefficients in the linear combination also depend on the degree p of the polynomial approximation. For $p = 0$ the estimator reduces to the Nadaraya-Watson estimator, i.e., the Nadaraya-Watson estimator can be seen as a local constant approximation to the regression function. For $p = 1$, a more explicit formula for (1.6) is given by

$$\hat{m}(x; 1, h) = \frac{1}{nh} \sum_{i=1}^{n} \frac{s_2(x,h) - s_1(x,h)(x_i - x)}{s_2(x,h)s_0(x,h) - s_1^2(x,h)} K\left(\frac{x_i - x}{h}\right) Y_i, \tag{1.7}$$

where $s_r(x,h) = (nh)^{-1} \sum_{i=1}^{n} (x_i - x)^r K\left((x_i - x)/h\right).$
It turns out that polynomials of odd degree have more desirable boundary properties (see the theoretical results later in this section).

In Figure 1.3 the weight assignment mechanism for local linear estimators is shown for the tutorial example used to illustrate the mechanism for the Nadaraya-Watson and Gasser-Müller estimators (Figures 1.1 and 1.2).

Similar to the Nadaraya-Watson estimator the local linear estimator first uses the transformed kernel function $K_{x,h}(u) = h^{-1}K\left((u - x)/h\right)$, i.e., compute $K_{x,h}(x_i)$ (the vertical lines in the middle part of the figure). These weights are then rescaled by $\left(s_2(x,h) - s_1(x,h)(x_i - x)\right) / \left(s_2(x,h)s_0(x,h) - s_1^2(x,h)\right)$, not by a constant factor as for the NadaryaWatson estimator. The actual weights are shown at the bottom part of the figure. The resulting estimate at x is indicated by the triangle in the figure.

We now introduce more notation which will be helpfull in the presentation of the asymptotic properties for the local polynomial smoothers. As noted earlier, we have to distinguish between interior points, i.e., $x \in [h, 1-h]$, and boundary points, i.e., $x \in [0, h)$ or $x \in (1-h, 1]$. The results, which we will describe below in Theorems 1.2 and 1.3, are obtained by Ruppert and Wand (1994) for the random design case, where asymptotic expressions are given for the conditional bias and variance of the estimator. Fan and Gijbels (1996, p. 68) claim that the results remain valid for fixed designs (for explicit results in the local linear case see Fan and Gijbels (1991)).

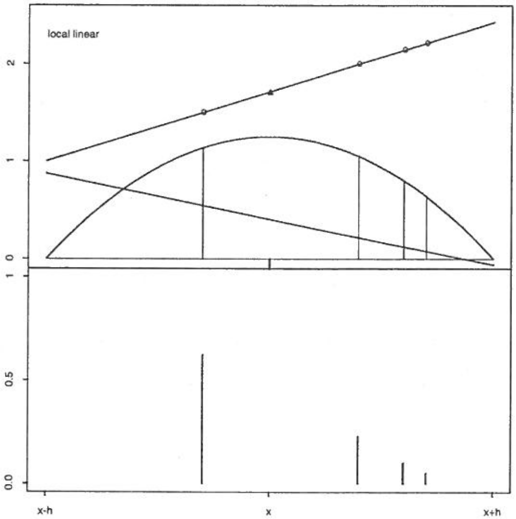

Figure 1.3: *Tutorial example of the local linear estimator. The data are indicated by ○ and the resulting estimate by ▲. The middle part illustrates the construction of the weights while the bottom part shows the actual weights.*

Interior points

Let $\mu_i(K)$ denote the i–th moment of the kernel $K(\cdot)$, i.e., $\int_{-1}^{1} u^i K(u)\,du$. Let $\boldsymbol{N}_p$ be the $(p+1)\times(p+1)$ matrix with (i,j)-th entry $\mu_{i+j-2}(K)$, and let $\boldsymbol{M}_p(u)$ be the same as $\boldsymbol{N}_p$, but with the first column replaced by $(1,u,\ldots,u^p)^T$. Define $K_{(p)}(u) = \{|\boldsymbol{M}_p(u)|/|\boldsymbol{N}_p|\}K(u)$. Then $K_{(p)}(\cdot)$ is a higher order kernel of order $(p+1)$ when p is odd, and of order $(p+2)$ when p is even. A function $L(\cdot)$, with support [-1,1], is called a higher order kernel of order r if

$$\mu_0(L) = \int_{-1}^{1} L(u)\,du = 1$$

$$\mu_\ell(L) = \int_{-1}^{1} u^\ell L(u)\,du = 0 \qquad \ell = 1,\ldots,r-1$$

and

$$\mu_r(L) = \int_{-1}^{1} u^r L(u)\,du \neq 0.$$

A higher order kernel $L(\cdot)$ can only be a proper kernel function when $r=2$. For $r>2$, $L(\cdot)$ must become negative somewhere in order to satisfy the moment conditions. This implies that the function $K_{(p)}(\cdot)$ takes negative values as soon as $p>1$.

Theorem 1.2 (Ruppert and Wand (1994))
Assume x is an interior point and

(i) *$K(\cdot)$ is a symmetric, continuous kernel with support [-1,1],*

(ii) *$h\to 0$, $nh\to\infty$, as $n\to\infty$,*

(iii) *$m^{(p+1)}(\cdot)$ is continuous at x,*

(iv) *$\sigma^2(x)>0$ and $\sigma^2(\cdot)$ is continuous at x,*

(v) *$f_X(x)>0$ and $f'_X(\cdot)$ is continuous at x.*

The asymptotic variance is given by

$$Var(\hat{m}(x;p,h)) = \frac{\sigma^2(x)}{nhf_X(x)}R(K_{(p)}) + o\left(\frac{1}{nh}\right).$$

The asymptotic bias for p odd is

$$E\left(\hat{m}(x;p,h)\right) - m(x) = m^{(p+1)}(x)\frac{h^{p+1}}{(p+1)!}\mu_{p+1}(K_{(p)}) + o(h^{p+1}).$$

The asymptotic bias for p even is

$$E\left(\hat{m}(x;p,h)\right) - m(x) = \left\{m^{(p+2)}(x) + (p+2)m^{(p+1)}(x)\frac{f_X'(x)}{f_X(x)}\right\}\frac{h^{p+2}}{(p+2)!}\mu_{p+2}(K_{(p)}) + o(h^{p+2})$$

provided that $m^{(p+2)}(\cdot)$ is continuous at x.

Boundary points

We restrict attention to left boundary points, i.e., $0 \le x < h$. A similar analysis is possible for right boundary points. It is convenient to treat boundary points as $x = \alpha h$, where $0 \le \alpha < 1$. Incomplete moments of the kernel $K(\cdot)$ are denoted by $\mu_{\ell,\alpha}(K) = \int_{-\alpha}^{1} u^\ell K(u)\,du$. For boundary points we modify the notation for $K_{(p)}(\cdot)$ in the following way : let $\boldsymbol{N}_p(\alpha)$ be the $(p+1)\times(p+1)$ matrix with (i,j)-th entry $\mu_{i+j-2,\alpha}(K)$, and let $\boldsymbol{M}_p(u,\alpha)$ be the same as $\boldsymbol{N}_p(\alpha)$, but with the first column replaced by $(1,u,\ldots,u^p)^T$. Define $K_{(p)}(u,\alpha) = (|\boldsymbol{M}_p(u,\alpha)|/|\boldsymbol{N}_p(\alpha)|)K(u), -\alpha < u < 1$.

Theorem 1.3 (Ruppert and Wand (1994))
Assume x is a left boundary point of the form $x = \alpha h$. Further assume conditions (i)–(v) of Theorem 1.2. The asymptotic bias is

$$E(\hat{m}(x;p,h)) - m(x) = m^{(p+1)}(x)\frac{h^{p+1}}{(p+1)!}\int_{-\alpha}^{1} u^{p+1}K_{(p)}(u,\alpha)\,du + o(h^{p+1})$$

and the asymptotic variance is

$$\mathrm{Var}(\hat{m}(x;p,h)) = \frac{\sigma^2(x)}{f_X(x)nh}\int_{-\alpha}^{1} K_{(p)}^2(u,\alpha)\,du + o\left(\frac{1}{nh}\right).$$

Remark 1.1
Assume p even and $m^{(p+2)}(\cdot)$ is continuous at x. Based on Theorems 1.2 and 1.3 we can compare the performance of the local polynomial smoothers $\hat{m}(x;p,h)$ and

$\hat{m}(x;p+1,h)$ both at interior and boundary points. First note that for $\hat{m}(x;p,h)$ the order of the bias drops from h^{p+2} in the interior to h^{p+1} in the boundary, while for $\hat{m}(x;p+1,h)$ the order is h^{p+2}, regardless of the position of x. Furthermore, for interior points, the bias of $\hat{m}(x;p,h)$ has a more complicated structure than the bias of $\hat{m}(x;p+1,h)$, depending on the design density. The term $\{f'_X(x)/f_X(x)\}m^{(p+1)}(x)$ can cause $\hat{m}(x;p,h)$ to have a large bias. When $|m^{(p+1)}(x)|$ is large, so is the bias of the estimator. Thus, even in situations where the true regression function is a polynomial of degree $p+1$, the estimator $\hat{m}(x;p,h)$ can have a large bias, while the estimator $\hat{m}(x;p+1,h)$ is unbiased (see below). Also, in highly clustered designs, where $|f'_X(x)/f_X(x)|$ is large, the bias of $\hat{m}(x;p,h)$ is large. Hence, the local polynomial estimator with p even cannot adapt to highly clustered designs. This bias problem disappears when using the local polynomial estimator with p odd.

Figure 1.4 illustrates the superior boundary behavior of the local linear estimator to the Nadaraya-Watson estimator. We consider the same tutorial data set as in Figures 1.1 and 1.3. It shows the unbiasedness of the local linear estimator when the true regression curve is linear (for reasons of this unbiasedness we refer to Remark 1.6).

The Gasser-Müller estimator does not have the drawback that the asymptotic bias depends on the design density $f_X(\cdot)$ (see Theorem 1.1). However, the estimator has another design problem. For random designs the variance of the Gasser-Müller estimator becomes 1.5 times that of the Nadaraya-Watson estimator (see Mack and Müller (1988)).

Fan (1992) refers to the local linear estimators as being design adaptive, since the method adapts to both fixed and random designs, and to both interior and boundary points. From the discussion above, it is clear that this design adaptivity property holds in general for local polynomial estimators with p odd.

The discussion on boundary problems in Remark 1.1 was completely concentrated on the performance of the bias. In terms of rates of convergence of the bias, no boundary adjustment is necessary for p odd. This is seen to be a great advantage of local linear fitting over the Nadaraya-Watson and other classical kernel estimators. Naturally, bias correction comes with a price, that is, increased variance. The rate of convergence of the variance does not depend on the degree p, but the constant term causes the variance to be larger for larger p. For example, for the biweight kernel $K(u) = \frac{15}{16}(1-u^2)^2\,\mathbb{1}\{|u| \leq 1\}$, the asymptotic variance of the local linear estimator is about 3.58 times that of the Nadaraya-Watson estimator when estimating $m(\cdot)$ at 0, if the same bandwidth is used for each (see e.g. Wand and Jones (1995, p. 129)).

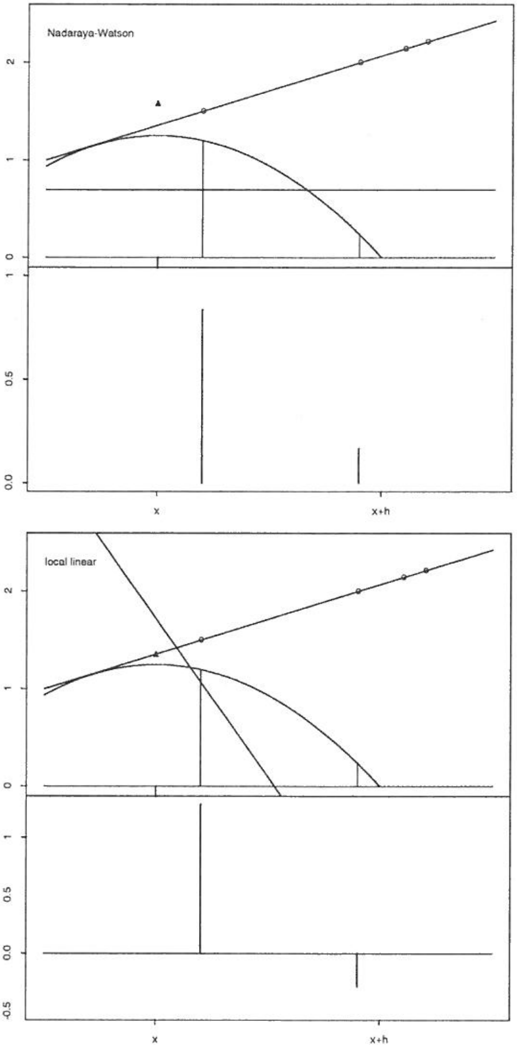

Figure 1.4: *Tutorial example of the Nadaraya-Watson and local linear estimators at a boundary point. The data are indicated by* ○ *and the resulting estimates by* ▲. *The interpretation and construction of these figures is the same as for Figures 1.1 and 1.3.*

1.3 Local polynomial estimators for sparse multinomials

In the previous section we considered local polynomial smoothing in a regression context. In this section we will apply local polynomial smoothing method to ordered sparse multinomial data (one-dimensional).

Let $\boldsymbol{p} = (p_1, \dots, p_k)^T$ be the vector of cell probabilities we want to estimate. Let $\boldsymbol{N} = (N_1, \dots, N_k)^T$ be cell counts generated from the multinomial distribution with cell probabilities $\boldsymbol{p}$ and with total sample size $n = \sum_{i=1}^k N_i$. The frequency estimators are denoted by $\overline{\boldsymbol{P}} = (\overline{P}_1, \dots, \overline{P}_k)^T$. To define local polynomial estimators for the cell probabilities we look at the data as regression type data. Let $x_i = (i-1/2)/k, i = 1, \dots, k$, be fixed equidistant design points on [0,1]. The multinomial data can be represented as $(x_i, \overline{P}_i), i = 1, \dots, k$.

We define the local p-th degree polynomial estimator for cell probability p_i as

$$\widehat{P}_i = \boldsymbol{e}_1^T (\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i^T \boldsymbol{W}_i \overline{\boldsymbol{P}} \tag{1.8}$$

where $\boldsymbol{e}_1^T$ is the $(p+1)$-vector $(1, 0, \dots, 0)$, $\boldsymbol{W}_i = \text{diag}_{1 \leq j \leq k} \left(h^{-1} K((x_j - x_i)/h)\right)$ and

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & x_1 - x_i & \dots & (x_1 - x_i)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_k - x_i & \dots & (x_k - x_i)^p \end{pmatrix}$$

the $k \times (p+1)$ design matrix, with p the order of the local polynomial approximation.

The main differences with the regression problem of Section 1.2 are that the frequency estimators are not independent and their variance depends on the cell probabilities $\boldsymbol{p}$. Since we aim at the study of the local polynomial smoothers for sparse tables we work in the sparse asymptotic framework, i.e., $k \equiv k(n) \to \infty$ as $n \to \infty$. Therefore, the study of the sparse asymptotic properties of the estimator is somewhat more complex if compared to classical asymptotic properties in the regression context, where only the sample size n tends to infinity.

The idea of local polynomial smoothing is based on a smoothness assumption of the true regression curve. Similarly, to study local polynomial smoothers in the ordered multinomial data context, we will need some smoothness criteria for the vector of true cell probabilities $\boldsymbol{p}$. Since the dimension k of the table increases, we must consider an infinite sequence of probability vectors $\boldsymbol{p}$ whose dimensions increase without bound. We simplify this structure through the following device :

we assume there exists a density function $f(\cdot)$ on [0,1] that generates the probability vector $\boldsymbol{p}$ via the relation

$$p_i = \int_{\frac{i-1}{k}}^{\frac{i}{k}} f(u)\,du, \quad i = 1, \ldots, k \tag{1.9}$$

(see also Bishop et al. (1975, p. 411) and Santner and Duffy (1989, p. 60)). Note that in this way the sequence of probability vectors $\boldsymbol{p}$ is linked to a single function $f(\cdot)$. We will refer to (1.9) as the latent density assumption. Via smoothness conditions on this latent density $f(\cdot)$, it is guaranteed that the vector of cell probabilities smoothly varies as the dimension of the table increases.

In the sequel of this section we will first rewrite the local polynomial smoother. As already noted in the previous section, local polynomial estimators are linear smoothers. In this section we will derive an explicit expression for the coefficients in the linear combination. Next, we will give some technical results which will be helpful in later chapters. Finally, we will derive the asymptotic bias and variance expressions for the local polynomial estimator $\widehat{P}_i$. The discussion on the superior boundary behavior of the local polynomial estimator compared to the classical kernel estimator, is delayed until Chapter 3.

Linear representation of the local polynomial smoother

The arguments to derive the explicit linear combination for the local polynomial smoother come from Ruppert and Wand (1994). The derivation is based on standard matrix calculus (see e.g. Searle (1982)).

If the matrix $\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i$ is invertible, the inverse can be written as

$$\left(\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1} = \left\{\det\left(\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right)\right\}^{-1} \operatorname{adj}\left(\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right),$$

where adj(A) represents the adjoint matrix of a square matrix A. Hence, the j-th element of the $(1 \times k)$–vector $\boldsymbol{e}_1^T \left(\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1} \boldsymbol{X}_i^T \boldsymbol{W}_i$ is given by

$$\begin{aligned}&\left\{\boldsymbol{e}_1^T \left(\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1} \boldsymbol{X}_i^T \boldsymbol{W}_i\right\}_j \\ &= \frac{1}{h} K\left(\frac{x_j - x_i}{h}\right) \left\{\det\left(\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right)\right\}^{-1} \sum_{\ell=1}^{p+1} \left\{\operatorname{adj}\left(\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right)\right\}_{1\ell} (x_j - x_i)^{\ell-1}.\end{aligned}$$

The term $\sum_{\ell=1}^{p+1} \left\{\operatorname{adj}\left(\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right)\right\}_{1\ell} (x_j - x_i)^{\ell-1}$ can be seen as the determinant of the $(p+1) \times (p+1)$-matrix $\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i$ with the first column replaced by

$(1, x_j - x_i, \ldots, (x_j - x_i)^p)^T$. Further note that

$$\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i = k \operatorname{diag}(1, h, \ldots, h^p)\, \boldsymbol{N}_{i,p} \operatorname{diag}(1, h, \ldots, h^p),$$

where $\boldsymbol{N}_{i,p}$ is the $(p+1) \times (p+1)$-matrix having the (r, s) entry equal to

$$(\boldsymbol{N}_{i,p})_{rs} = m_{k,r+s-2}(x_i) \tag{1.10}$$

with

$$m_{k,\ell}(x) = \frac{1}{kh} \sum_{j=1}^{k} \left(\frac{x_j - x}{h} \right)^{\ell} K\left(\frac{x_j - x}{h} \right).$$

Combining these results yields,

$$\widehat{P}_i = \frac{1}{kh} \sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h} \right) \overline{P}_j \tag{1.11}$$

where

$$L_{i,p}(u) = \frac{|\boldsymbol{M}_{i,p}(u)|}{|\boldsymbol{N}_{i,p}|} K(u). \tag{1.12}$$

From (1.12) we have

$$\begin{aligned} C_{k,\ell}(x_i) &= \frac{1}{kh} \sum_{j=1}^{k} \left(\frac{x_j - x_i}{h} \right)^{\ell} L_{i,p}\left(\frac{x_j - x_i}{h} \right) \\ &= \frac{\begin{vmatrix} m_{k,\ell}(x_i) & m_{k,1}(x_i) & \ldots & m_{k,p}(x_i) \\ \vdots & \vdots & & \vdots \\ m_{k,\ell+p}(x_i) & m_{k,p+1}(x_i) & \ldots & m_{k,2p}(x_i) \end{vmatrix}}{|\boldsymbol{N}_{i,p}|}. \end{aligned} \tag{1.13}$$

From this relation it is immediate that

$$C_{k,0}(x_i) = 1 \qquad \text{and} \qquad C_{k,\ell}(x_i) = 0 \qquad \text{for } \ell = 1, \ldots, p. \tag{1.14}$$

The fact that (1.14) is an exact (i.e., non asymptotic) relation that holds for all design points x_i, is of major relevance for the behavior of the local polynomial estimator. We will discuss this in more detail in Remark 1.6 and in Chapter 3.

We can think about $C_{k,\ell}(x_i)$ as a kind of "discrete moment"of the function $L_{i,p}(\cdot)$. Furthermore, relation (1.14) states that these "moments", up to order p, are equal to zero. We therefore say that $L_{i,p}(\cdot)$ is a discrete higher order kernel of order $p+1$. The immediate consequence of working with a higher order kernel function $L_{i,p}(\cdot)$ is that the weights $L_{i,p}((x_j - x_i)/h))$ can become negative as soon as $p > 1$. It turns out that even for $p = 1$ the function $L_{i,p}(\cdot)$ can take negative values, but only for design points x_i in the boundary region (see Section 3.2 for more details).

Some graphical illustrations

In Section 1.2 we illustrated on a tutorial data set how local constant and local linear estimators are constructed when estimating a function at a single point. In this paragraph, we illustrate for a sparse multinomial data set the complete picture of the estimated cell probabilities. As data set we consider the mine explosions data presented in Table 1.3. Figure 1.5 shows the estimated cell probabilities based on the local constant estimation procedure, while Figure 1.6 is for the local linear estimators.

To construct the estimators we have chosen three different bandwidths and as can be seen from the figures, the bandwidth has a crucial effect on the visual impression of the resultant cell probability estimates. For too small bandwidths there is almost no difference between the local constant, local linear and frequency estimates, since so little information is borrowed from the neighbors to compute the smoothed estimators (e.g., with a bandwidth equal to 0.02 we use information from at most 2 neighbors). For too large bandwidths on the other hand, almost all structure that is present in the data is smoothed away, because of the fact that too many neighbors have an influence on the smoothed estimators (e.g., a bandwidth of 0.2 corresponds to smoothing based on at most 22, and at least 11 neighbors). Both extreme levels of smoothing are respectively called under- and oversmoothing. This illustration makes clear that bandwidth selection is an important issue in the nonparametric smoothing literature. In Section 3.3 we discuss this problem in more depth.

The figures make also clear that local constant estimators suffer from boundary bias problems, which can be seen from the severe underestimates at the left boundary cells. Also this topic we will discuss in more detail in Chapter 3.

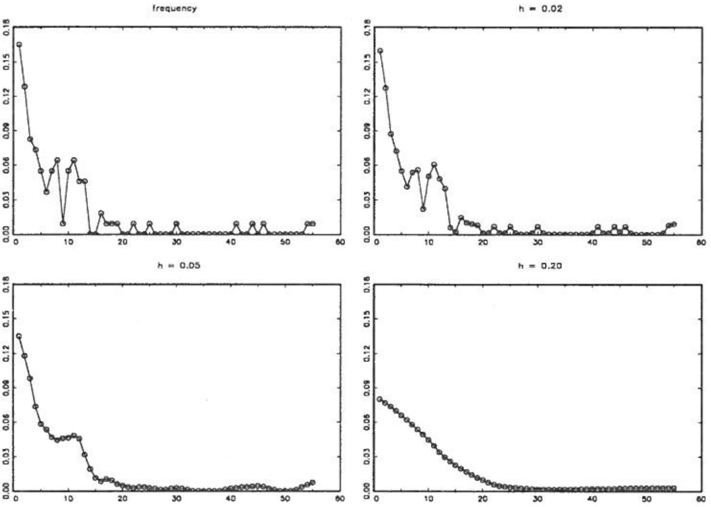

Figure 1.5: *Frequency estimators and local constant estimators based on three different bandwidths for the mine explosions data.*

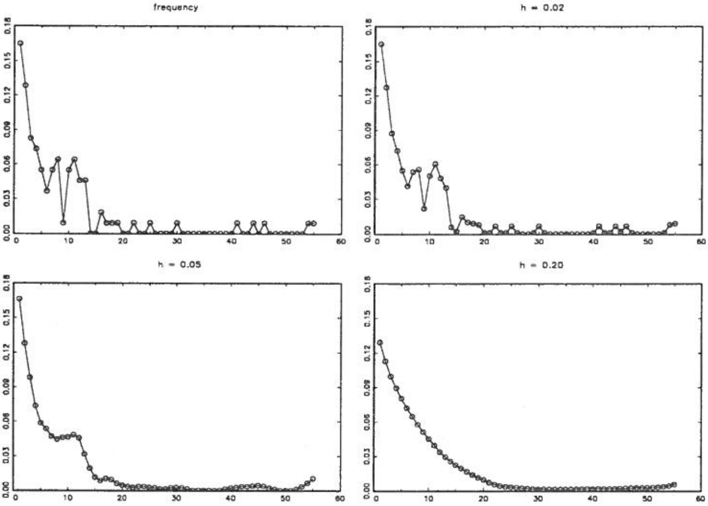

Figure 1.6: *Frequency estimators and local linear estimators based on three different bandwidths for the mine explosions data.*

Useful technical results

In this paragraph we will give two lemmas. In the first one we describe how we can approximate summations by integrals when doing sparse asymptotics. The second result states that the weight function $L_{i,p}(\cdot)$ is bounded. Note that both results are valid uniformly in the i-index, which will be important in Chapters 2 and 3, since there we will investigate global measures of performance (sparse consistency and MSSE). The rather technical proofs of both results are given in Section 1.4.

We will need the following conditions on the kernel function $K(\cdot)$ and the bandwidth h:

(C.1) $K(\cdot)$ is a symmetric, continuous kernel with bounded support [-L,L],

(C.2) $h \to 0,\ hk \to \infty$ as $n \to \infty$.

Lemma 1.1
Let $G(\cdot)$ be a continuous function on a compact support. Define $\alpha_i = x_i/h$ and $\beta_i = (1-x_i)/h$. Assume (C.2), we then have, uniformly in i, for $i = 1, \ldots, k$,

$$\frac{1}{kh}\sum_{j=1}^{k} G\left(\frac{x_j - x_i}{h}\right) = \int_{-\alpha_i}^{\beta_i} G(v)\,dv + o(1).$$

Remark 1.2
For $G(u) = u^{\ell}K(u)$ the conditions of Lemma 1.1 are satisfied under (C.1) and (C.2). Define

$$\mu_\ell(x_i) = \int_{-\alpha_i}^{\beta_i} v^{\ell}K(v)\,dv. \tag{1.15}$$

For $m_{k,\ell}(x_i)$, the entries of the matrices $\boldsymbol{N}_{i,p}$ and $\boldsymbol{M}_{i,p}(\cdot)$, we then have,uniformly in the i-index,

$$m_{k,\ell}(x_i) = \mu_\ell(x_i) + o(1). \tag{1.16}$$

Lemma 1.2
Assume (C.1) and (C.2). If at least (p+1) weights in the matrix $\boldsymbol{W}_i$ are strictly positive, then the function $L_{i,p}(\cdot)$ is bounded. This bound is uniform in the i-index.

Remark 1.3

Since $K(\cdot)$ is defined on the support $[-L, L]$, $K\left((x_j - x_i)/h\right) = 0$ when $|x_j - x_i| > Lh$. This implies that the number of positive weights is of the order $O(kh)$. Hence, under condition (C.2) the requirement of $p + 1$ positive weights is satisfied for k large enough.

Remark 1.4

The weight function $L_{i,p}(\cdot)$ has the same support $[-L, L]$ as $K(\cdot)$, and is bounded on this support. Denote C_L the bound of $|L_{i,p}(\cdot)|$. Therefore, we have that expressions of the form

$$\frac{1}{kh}\sum_{j=1}^{k} g_1\left(\frac{x_j - x_i}{h}\right) g_2\left(L_{i,p}\left(\frac{x_j - x_i}{h}\right)\right)$$

are bounded, uniformly in the i-index, where $g_1(\cdot)$ is a continuous function on $[-L, L]$, and $g_2(\cdot)$ a continuous function on $[-C_L, C_L]$ with $g_2(0) = 0$.

Asymptotic bias and variance expressions for the local polynomial smoother

The main result of this section is given in the next theorem.

Theorem 1.4

Assume (C.1), (C.2) and

(C.3) $f^{(p+1)}(\cdot)$ is continuous on $[0, 1]$.

We then have, uniformly in the i-index

$$E\widehat{P}_i - p_i = \frac{f^{(p+1)}(x_i)}{(p+1)!}\frac{h^{p+1}}{k}C_{k,p+1}(x_i) + o\left(\frac{h^{p+1}}{k}\right) \tag{1.17}$$

and

$$\text{Var}\widehat{P}_i = \frac{f(x_i)}{nk^2h}\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j - x_i}{h}\right) + o\left(\frac{1}{nk^2h}\right). \tag{1.18}$$

Proof

To obtain the expression for the bias, we need the following two analytical facts, which are based on (C.3) : a Taylor expansion yields, uniformly in the j-index,

$$p_j = \int\limits_{(j-1)/k}^{j/k} f(u)du = \sum_{\ell \text{ even}}^{\ell=0,p+1} \frac{f^{(\ell)}(x_j)}{(\ell+1)!2^\ell k^{\ell+1}} + o\left(\frac{1}{k^{p+2}}\right) \tag{1.19}$$

and, uniformly in the i and j-index,

$$f^{(\ell)}(x_j) = \sum_{r=0}^{p+1-\ell} \frac{f^{(\ell+r)}(x_i)}{r!}(x_j - x_i)^r + o(h^{p+1-\ell}) \quad \text{for } |x_j - x_i| \leq Lh. \quad (1.20)$$

Further note that $L_{i,p}((x_j - x_i)/h) = 0$ for $|x_j - x_i| > Lh$. Based on these facts and using the orthogonality relation (1.14) and $kh \to \infty$ we obtain

$$\begin{aligned}
E\widehat{P}_i &= \frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right) p_j \\
&= \sum_{\ell \text{ even}}^{\ell=0,p+1} \frac{1}{(\ell+1)!2^\ell k^{\ell+1}} \left\{ \frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right) \left[\sum_{r=0}^{p+1-\ell} \frac{f^{(\ell+r)}(x_i)}{r!}(x_j - x_i)^r\right]\right\} \\
&\quad + \sum_{\ell \text{ even}}^{\ell=0,p+1} o\left(\frac{h^{p+1-\ell}}{k^{\ell+1}}\right) + o\left(\frac{1}{k^{p+2}}\right) \\
&= \sum_{\ell \text{ even}}^{\ell=0,p+1} \frac{f^{(\ell)}(x_i)}{(\ell+1)!2^\ell k^{\ell+1}} + \frac{1}{k}\frac{f^{(p+1)}(x_i)}{(p+1)!} C_{k,p+1}(x_i) + o\left(\frac{h^{p+1}}{k}\right) \\
&= p_i + \frac{f^{(p+1)}(x_i)}{(p+1)!}\frac{h^{p+1}}{k} C_{k,p+1}(x_i) + o\left(\frac{h^{p+1}}{k}\right).
\end{aligned}$$

To obtain expression (1.18) for the variance, we will use the following facts :
from (1.19) we have, uniformly in the j-index,

$$p_j = \frac{f(x_j)}{k} + O\left(\frac{1}{k^3}\right). \quad (1.21)$$

Note that this relation is already valid if $f''(\cdot)$ is bounded on $[0,1]$.
From (1.20) we also have, for $|x_j - x_i| \leq Lh$, uniformly in the i and j-index,

$$f(x_j) = f(x_i) + O(h) \quad (1.22)$$

and $L_{i,p}((x_j - x_i)/h) = 0$ for $|x_j - x_i| > Lh$.
Based on these facts and using $\text{Cov}(\overline{P}_j, \overline{P}_\ell) = (p_j(\delta_{j\ell} - p_\ell))/n$ we have

$$\text{Var}(\widehat{P}_i) = \frac{1}{nk^2h^2}\sum_{j=1}^{k}\sum_{\ell=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right) L_{i,p}\left(\frac{x_\ell - x_i}{h}\right) p_j(\delta_{j\ell} - p_\ell)$$

$$= \frac{1}{nk^2h^2}\left\{\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j-x_i}{h}\right)p_j - \left(\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j-x_i}{h}\right)p_j\right)^2\right\}$$

$$= \frac{f(x_i)}{nk^2h}\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j-x_i}{h}\right) + O\left(\frac{1}{nk^2}\right)\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j-x_i}{h}\right) + O\left(\frac{1}{nk^2}\right).$$

Since $\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j-x_i}{h}\right) = O(1)$ uniformly in the i-index (see Remark 1.4) we obtain, uniformly in the i-index, (1.17). ■

Remark 1.5

If instead of condition (C.3) we only have

(C.3') $f^{(p+1)}(\cdot)$ is bounded on $[0,1]$,

the expression for the variance remains valid (assume $p \geq 1$). For the bias, only the order remains, uniformly in the i-index,

$$E\widehat{P}_i - p_i = O\left(\frac{h^{p+1}}{k}\right). \tag{1.23}$$

In Chapter 2 we will use this version of the asymptotic bias expression to present the results on the sparse consistency of the local polynomial estimator. For the discussion on the MSSE of the local polynomial smoother (Chapter 3) the stronger condition (C.3) is needed.

Note that (1.23) follows from the proof of (1.17), where the Taylor arguments leading to (1.19) and (1.20) are replaced by Young's form of Taylor, which yields, for $|x_j - x_i| \leq Lh$,

$$p_j = \int_{(j-1)/k}^{j/k} f(u)du = \sum_{\ell \text{ even}}^{\ell=0,p+1} \frac{f^{(\ell)}(x_j)}{(\ell+1)!2^\ell k^{\ell+1}} + O\left(\frac{1}{k^{p+2}}\right)$$

and

$$f^{(\ell)}(x_j) = \sum_{r=0}^{p+1-\ell} \frac{f^{(\ell+r)}(x_i)}{r!}(x_j - x_i)^r + O(h^{p+1-\ell})$$

uniformly in the i and j-index.

Remark 1.6
An important consequence of property (1.14) is that local polynomial smoothers are exactly unbiased when $f(\cdot)$ is a polynomial of degree p. This can be seen from the proof of (1.17), if we note that relations (1.19) and (1.20) reduce to, since $f^{(p+1)}(\cdot) \equiv 0$

$$p_j = \int\limits_{(j-1)/k}^{j/k} f(u)du = \sum_{\ell \text{ even}}^{\ell=0,p} \frac{f^{(\ell)}(x_j)}{(\ell+1)!2^\ell k^{\ell+1}}$$

and

$$f^{(\ell)}(x_j) = \sum_{r=0}^{p-\ell} \frac{f^{(\ell+r)}(x_i)}{r!}(x_j - x_i)^r.$$

1.4 Proofs

Proof of Lemma 1.1
By a simple substitution we obtain

$$\begin{aligned}
&\left| \frac{1}{kh} \sum_{j=1}^{k} G\left(\frac{x_j - x_i}{h}\right) - \int\limits_{-\alpha_i}^{\beta_i} G(v)\, dv \right| \\
&= \left| \frac{1}{kh} \sum_{j=1}^{k} G\left(\frac{x_j - x_i}{h}\right) - \frac{1}{h} \int\limits_{0}^{1} G\left(\frac{y - x_i}{h}\right)\, dy \right| \\
&= \left| \frac{1}{h} \sum_{j=1}^{k} \int\limits_{(j-1)/k}^{j/k} \left(G\left(\frac{x_j - x_i}{h}\right) - G\left(\frac{y - x_i}{h}\right)\right)\, dy \right| \\
&\le \frac{1}{h} \sum_{j=1}^{k} \int\limits_{(j-1)/k}^{j/k} \left| G\left(\frac{x_j - x_i}{h}\right) - G\left(\frac{y - x_i}{h}\right)\right|\, dy \\
&= \frac{1}{h} \left(\sum_{j \in I_i} + \sum_{j \in II_i} + \sum_{j \in III_i} \right) \int\limits_{(j-1)/k}^{j/k} \left| G\left(\frac{x_j - x_i}{h}\right) - G\left(\frac{y - x_i}{h}\right)\right|\, dy, \quad (1.24)
\end{aligned}$$

where I_i, II_i, III_i is the following partition of $\{1,\ldots,k\}$. With $C_j = [(j-1)/k, j/k]$, denote $(C_j - x_i)/h = \{(x - x_i)/h \ : \ x \in C_j\}$ and define

$$\begin{aligned} I_i &= \{j \ : \ (C_j - x_i)/h \ \subset \ [a,b]\} \\ II_i &= \{j \ : j \notin I_i \text{ and } (C_j - x_i)/h \ \cap \ [a,b] \neq \emptyset\} \\ III_i &= \{j \ : \ (C_j - x_i)/h \ \cap \ [a,b] = \emptyset\}\,, \end{aligned}$$

where $[a,b]$ is the support of $G(\cdot)$. First note that the sum over III_i has no contribution to (1.24), and that for $j \in II_i$ we don't know whether $(x_j - x_i)/h \in [a,b]$. Further, for $j \in I_i$ we have, since $G(\cdot)$ is continuous on a compact support,

$$G\left(\frac{x_j - x_i}{h}\right) - G\left(\frac{y - x_i}{h}\right) = o(1),$$

and for $j \in II_i$ we certainly have

$$G\left(\frac{x_j - x_i}{h}\right) - G\left(\frac{y - x_i}{h}\right) = O(1),$$

where both order bounds are uniform in the i and j-index. Since we work on an equidistant design, we have $\#I_i \leq (b-a)kh$ and $\#II_i \leq 2$. Combine these facts into (1.24) to end the proof. ■

Proof of Lemma 1.2

From the definition of $L_{i,p}(u)$, and the fact that $K(\cdot)$ is continuous on the compact support [-L,L], it suffices to show that $|\boldsymbol{M}_{i,p}(u)|$ is bounded uniformly in the i-index for bounded u and that $|\boldsymbol{N}_{i,p}|$ is uniformly bounded away from zero. The uniform boundedness of $|\boldsymbol{M}_{i,p}(u)|$ for bounded u follows immediately from (1.16).

The matrices $\boldsymbol{N}_{i,p}$ are positive definite if there are at least $p+1$ design points with positive weights. To see this, let $\boldsymbol{z} = (z_1, \ldots, z_{p+1})^T \neq \boldsymbol{0}$ and write

$$\begin{aligned} \boldsymbol{z}^T \boldsymbol{N}_{i,p} \boldsymbol{z} &= \sum_{r=1}^{p+1} \sum_{s=1}^{p+1} m_{r+s-2}(x_i) z_r z_s \\ &= \frac{1}{kh} \sum_{j=1}^{k} \left(\sum_{r=1}^{p+1} \left(\frac{x_j - x_i}{h}\right)^{r-1} z_r \right)^2 K\left(\frac{x_j - x_i}{h}\right). \qquad (1.25) \end{aligned}$$

If there are $p+1$ positive weights in the matrix $\boldsymbol{W}_i$ then, for all possible $\boldsymbol{z} \neq \boldsymbol{0}$, there is at least one j for which the contribution to (1.25) is strictly positive. Therefore,

$\boldsymbol{z}^T \boldsymbol{N}_{i,p} \boldsymbol{z}$ is strictly positive for all $\boldsymbol{z} \neq \boldsymbol{0}$, and hence the matrix $\boldsymbol{N}_{i,p}$ has a strictly positive determinant.

Moreover, $|\boldsymbol{N}_{i,p}| \geq |\boldsymbol{N}_{1,p}|, 1 \leq i \leq k$. This can be seen as follows. Since the matrices $\boldsymbol{N}_{i,p}$ and $\boldsymbol{N}_{1,p}$ are symmetric and positive definite, there exists a non-singular matrix $\boldsymbol{Q}$ such that

$$\boldsymbol{N}_{i,p} = \boldsymbol{Q}^T \boldsymbol{\Lambda}_i \boldsymbol{Q}$$
$$\boldsymbol{N}_{1,p} = \boldsymbol{Q}^T \boldsymbol{\Lambda}_1 \boldsymbol{Q}$$

where $\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Lambda}_1$ are diagonal matrices (which are also positive definite). From (1.25), and the fact the design points are equidistant, it is clear that $\boldsymbol{z}^T \boldsymbol{N}_{i,p} \boldsymbol{z} \geq \boldsymbol{z}^T \boldsymbol{N}_{1,p} \boldsymbol{z}$, for $i = 1, \ldots, k$. This yields that the matrix $\boldsymbol{N}_{i,p} - \boldsymbol{N}_{1,p}$ is non-negative definite, and hence

$$0 \leq |\boldsymbol{N}_{i,p} - \boldsymbol{N}_{1,p}| = |\boldsymbol{Q}|^2 |\boldsymbol{\Lambda}_i - \boldsymbol{\Lambda}_1|.$$

We also have

$$0 < |\boldsymbol{N}_{i,p}| = |\boldsymbol{Q}|^2 |\boldsymbol{\Lambda}_i|,$$
$$0 < |\boldsymbol{N}_{1,p}| = |\boldsymbol{Q}|^2 |\boldsymbol{\Lambda}_1|.$$

Therefore, to show $|\boldsymbol{N}_{i,p}| \geq |\boldsymbol{N}_{1,p}|$, it suffices to prove $|\boldsymbol{\Lambda}_i| \geq |\boldsymbol{\Lambda}_1|$. From the fact that the matrix $\boldsymbol{N}_{i,p} - \boldsymbol{N}_{1,p}$ is non-negative definite we also know this for the matrix $\boldsymbol{\Lambda}_i - \boldsymbol{\Lambda}_1$. Hence, the elements of $\boldsymbol{\Lambda}_i - \boldsymbol{\Lambda}_1$ are positive. Furthermore, from the positive definitness of the matrices $\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Lambda}_1$, we also know that the elements on the diagonal of those matrices are strictly positive. Using these arguments we can show, by induction on the dimensions of the matrices, that

$$0 \leq |\boldsymbol{N}_{i,p} - \boldsymbol{N}_{1,p}| \leq |\boldsymbol{N}_{i,p}| - |\boldsymbol{N}_{1,p}|\,.$$

So, now we know that $\inf_{1 \leq i \leq k} |\boldsymbol{N}_{i,p}| = |\boldsymbol{N}_{1,p}| > 0$. We will show that $|\boldsymbol{N}_{1,p}|$ is bounded away from zero when $k \to \infty$. For the elements $m_{k,\ell}(x_1)$ of the matrix $\boldsymbol{N}_{1,p}$ we know from (1.16) that $m_{k,\ell}(x_1) = \mu_\ell(x_1) + o(1)$, as $n \to \infty$. It is clear that $\mu_\ell(x_1) = \mu_\ell(0) + o(1)$ as $n \to \infty$,where $\mu_\ell(0) = \int_0^L v^\ell k(v)\, dv$. This yields for the determinant of $\boldsymbol{N}_{1,p}$

$$|\boldsymbol{N}_{1,p}| = \begin{vmatrix} \mu_0(0) & \mu_1(0) & \ldots & \mu_p(0) \\ \vdots & \vdots & & \vdots \\ \mu_p(0) & \mu_{p+1}(0) & \ldots & \mu_{2p}(0) \end{vmatrix} + o(1).$$

The determinant on the right hand side is strictly positive since the corresponding matrix is positive definite (similar proof as for the matrix $\boldsymbol{N}_{i,p}$).
Hence, $\inf_{1\leq i\leq k} |\boldsymbol{N}_{i,p}| > 0$ as $n \to \infty$, i.e., $|\boldsymbol{N}_{i,p}|$ is uniformly bounded away from zero. ■

Chapter 2

Sparse consistency rates

In this chapter we investigate sparse asymptotic consistency of the frequency estimators and the local polynomial smoothers. Recall from Chapter 1 that sparse asymptotic consistency of a generic sequence of estimators $\boldsymbol{P}^* = (P_1^*, \ldots, P_k^*)^T$ for the cell probabilities $\boldsymbol{p} = (p_1, \ldots, p_k)^T$ is defined as

$$\sup_{1 \leq i \leq k} \left| \frac{P_i^*}{p_i} - 1 \right| \xrightarrow{a.s.} 0, \quad \text{as } n \to \infty. \tag{2.1}$$

In general, the frequency estimators may fail to be sparse consistent. First we give, in Section 2.1, a simple example where the frequency estimators are not sparse consistent. Next, we give in Theorem 2.1 sufficient conditions under which they are sparse consistent. Since the theorem is in terms of rates of convergence, it is possible to investigate what degree of sparseness (i.e., the rate at which k tends to infinity) we can allow to guarantee sparse consistency. This will be illustrated in two examples.

In Section 2.2 we study sparse consistency for local polynomial estimators for cell probabilities. Sufficient conditions are given to guarantee sparse consistency. In the result we obtain information on the rate of convergence, as we did for the frequency estimators. Based on the same examples as in Section 2.1 we compare the performance of both estimators, in terms of degree of sparseness and sparse consistency rates. We show that the smoothed estimators obey the sparse consistency property for degrees of sparseness at which the frequency estimators are not sparse consistent.

Simonoff (1983) studies sparse consistency rates for penalized likelihood estimators for cell probabilities. We will compare his results to the ones we obtain for local polynomial smoothers. The fastest rate in Simonoff (1983), coincides with the rate we get for the local linear smoother. However, to obtain his result he needs regularity conditions that are more stringent than our assumptions. A detailed comparison is given in Section 2.2.

To prove our main results (Theorems 2.1 and 2.2) we rewrite the frequency estimators in terms of independent zero-one random variables, and we rely on the Bernstein inequality.

2.1 Sparse consistency of frequency estimators

As noted in the introduction, the frequency estimators may fail to be sparse consistent. To illustrate this we consider the following example.

Example 2.1
Assume k and n are related as follows, $n = ck$, where $c \in I\!N_0$ is a constant. This means that k and n grow at the same rate. Further assume uniform cell probabilities, i.e., $p_i = k^{-1}, i = 1, \dots, k$.
For $0 < \varepsilon < 1/c$ we have

$$
\begin{aligned}
& I\!P\left\{\sup_{1\le i\le k}\left|\frac{\overline{P}_i}{p_i} - 1\right| < \varepsilon\right\} = I\!P\left\{\sup_{1\le i\le k} |N_i - c| < \varepsilon c\right\} \\
& = I\!P\{N_i = c, i = 1,\dots,k\} = \frac{n!}{(c!)^k}\left(\frac{1}{k}\right)^n = \frac{n!}{n^n}\left(\frac{c^c}{c!}\right)^{n/c} \\
& = \frac{\sqrt{2\pi} n^{1/2}\exp(-n)\exp\left((12n)^{-1}\theta_n\right)}{\left(\sqrt{2\pi}c^{1/2}\exp(-c)\exp\left((12c)^{-1}\theta_c\right)\right)^{n/c}} \qquad (2.2)
\end{aligned}
$$

with $0 < \theta_n, \theta_c < 1$ (Stirling's formula). Since (2.2) converges to zero as $n \to \infty$, even in probability, (2.1) is not valid.

In Theorem 2.1 we give a sufficient condition under which the frequency estimators are sparse consistent. Denote

$$m_n = \inf_{1\le i\le k} p_i \quad \text{and} \quad M_n = \sup_{1\le i\le k} p_i.$$

Theorem 2.1
If

$$\frac{\ln n + \ln k}{n m_n} \to 0, \;\; \textit{as } n \to \infty$$

then

$$\sup_{1\leq i\leq k}\left|\frac{\overline{P}_i}{p_i}-1\right| \overset{a.s.}{=} O\left(\sqrt{\frac{\ln n+\ln k}{nm_n}}\right).$$

Proof

Let, for $\ell=1,\ldots,n$, $\boldsymbol{Y}_{n\ell}=(Y_{\ell 1},\ldots,Y_{\ell k})^T$ where, for $i=1,\ldots,k$,

$$Y_{\ell i}=\begin{cases}1 & \text{if the } \ell\text{-th observation is in cell } i\\ 0 & \text{otherwise.}\end{cases} \tag{2.3}$$

Note that $\boldsymbol{Y}_{n\ell}$ is a triangular array, and that, for each fixed n, $\boldsymbol{Y}_{n1},\ldots,\boldsymbol{Y}_{nn}$ are independent. In terms of these variables we can write $\overline{P}_i=\dfrac{1}{n}\sum\limits_{\ell=1}^{n}Y_{\ell i}$.

By using a Bonferroni type inequality we have

$$I\!P\left\{\sup_{1\leq i\leq k}\left|\frac{\overline{P}_i-p_i}{p_i}\right|>\varepsilon\right\}\leq k\sup_{1\leq i\leq k} I\!P\left\{\left|\sum_{\ell=1}^{n}\widetilde{Y}_{\ell i}\right|>\varepsilon n\right\},$$

where $\widetilde{Y}_{\ell i}=(Y_{\ell i}-p_i)/p_i$.

Bernstein's inequality (see e.g. Pollard (1984, p. 193)) states that for independent random variables $S_1,\ldots,S_n$ with $E(S_\ell)=0, |S_\ell|\leq b$ and $\sum\limits_{\ell=1}^{n}\mathrm{Var}(S_\ell)\leq v$

$$I\!P\left\{\left|\sum_{\ell=1}^{n}S_\ell\right|\geq x\right\}\leq 2\exp\left\{\frac{-x^2}{(2v+\frac{2}{3}bx)}\right\}$$

for all $x>0$.

It is clear that the variables $\widetilde{Y}_{\ell i}$ have zero mean and

$$|\widetilde{Y}_{\ell i}|\leq\frac{1}{p_i}\leq\frac{1}{m_n}$$
$$\mathrm{Var}(\widetilde{Y}_{\ell i})=\frac{p_i(1-p_i)}{p_i^2}\leq\frac{1}{m_n}.$$

An application of Bernstein's inequality results in

$$I\!P\left\{\left|\sum_{\ell=1}^{n}\widetilde{Y}_{\ell i}\right|>\varepsilon n\right\}\leq 2\exp\left\{\frac{-\varepsilon^2 nm_n}{2+\frac{2}{3}\varepsilon}\right\}.$$

For $0 < \varepsilon < 3$ we then have

$$\mathbb{P}\left\{\left|\sum_{\ell=1}^{n} \widetilde{Y}_{\ell i}\right| > \varepsilon n\right\} \leq 2\exp\left\{-\frac{1}{4}n\varepsilon^2 m_n\right\}.$$

Now, take $\varepsilon_n = \sqrt{\dfrac{4[(1+\delta)\ln n + \ln k]}{nm_n}}$ with $\delta > 0$. For n sufficiently large we have $0 < \varepsilon_n < 3$, since

$$\frac{\ln n + \ln k}{nm_n} \to 0, \text{ as } n \to \infty.$$

Therefore, we obtain, for N sufficiently large,

$$\begin{aligned}
&\sum_{n=N}^{\infty} \mathbb{P}\left\{\sup_{1\leq i\leq k}\left|\frac{\overline{P}_i - p_i}{p_i}\right| > \varepsilon_n\right\} \\
&\leq \sum_{n=N}^{\infty} k \sup_{1\leq i\leq k} \mathbb{P}\left\{\left|\sum_{\ell=1}^{n} \widetilde{Y}_{\ell i}\right| > \varepsilon_n n\right\} \\
&\leq 2\sum_{n=1}^{\infty} k\exp(-(1+\delta)\ln n)\exp(-\ln k) \\
&= 2\sum_{n=1}^{\infty} n^{-1-\delta} < \infty.
\end{aligned}$$

This complete convergence result implies the lemma. ■

Remark 2.1

In Aerts et al. (1997a) sparse consistency of frequency estimators is studied using a natural link between $\overline{P}_i - p_i$ and the oscillation of an appropriate empirical process. This approach requires more stringent conditions on the vector of cell probabilities. Specifically, they assume that the cell probabilities are generated by a latent density function $f(\cdot)$ on [0,1] through relation (1.9), which implies smoothness of the vector of cell probabilities $\boldsymbol{p}$. For frequency estimators this condition is superfluous.

Further note that the quantity M_n does not appear in Theorem 2.1. The rate at which m_n tends to zero, together with the degree of sparseness of the table, are the only two factors that determine the sparse consistency rate of the frequency estimators. Compared to the result in Aerts et al. (1997a), Theorem 2.1 allows, under weaker conditions, a slightly higher degree of sparseness and still guarantees sparse consistency (see Example 2.3).

As an illustration of Theorem 2.1, we now discuss for two examples the interplay between degree of sparseness and sparse consistency. We include information on the rate of convergence.

Example 2.2
Assume the cell probabilities satisfy

$$0 < \frac{\gamma_1}{k} \leq p_i \leq \frac{\gamma_2}{k} < 1 \quad i = 1, \ldots, k. \tag{2.4}$$

This condition on the cell probabilities is imposed by Simonoff (1983) (see also Theorem 2.3 below). For this example the condition of Theorem 2.1 is equivalent to $k(\ln k + \ln n)/n \to 0$. Hence, it is obvious that in this example, the frequency estimators are sparse consistent for tables with $k \propto n^q, 0 < q < 1$, where the notation $a_n \propto b_n$ means that a_n and b_n are of the same order, i.e.,

$$\frac{a_n}{b_n} \to C, \qquad \text{as } n \to \infty,$$

with C a nonzero, finite constant. The sparse consistency rate for this degree of sparseness is $O(n^{(q-1)/2}(\ln n)^{1/2})$.
Note that in Example 2.1 (2.4) is satisfied with $\gamma_1 = \gamma_2 = 1$. There we have shown that the frequency estimators are not sparse consistent when $k \propto n$. Therefore, the degree of sparseness where the consistency for the frequency estimators breaks down must be somewhere between $k \propto n^q$, $0 < q < 1$, and $k \propto n$. When we take a closer look at the condition guaranteeing consistency, we see that this is also fulfilled for $k \propto \frac{n}{(\ln n)^{1+\varepsilon}}$ with $\varepsilon > 0$, but not for $k \propto \frac{n}{\ln n}$. This suggests $k \propto \frac{n}{\ln n}$ as a possible rate for the breakdown of the sparse consistency of the frequency estimators. The sparse consistency rate of the frequency estimators when $k \propto \frac{n}{(\ln n)^{1+\varepsilon}}$ with $\varepsilon > 0$, is $O((\ln n)^{-\varepsilon/2})$.

Example 2.3
Condition (2.4) is equivalent to $m_n \propto k^{-1}$ and $M_n \propto k^{-1}$. Now consider cell probabilities where the smallest cell probability converges faster to 0 than k^{-1}. Take e.g. a vector of cell probabilities $\boldsymbol{p}$ with $m_n \propto k^{-\alpha}$, $\alpha > 1$. For specific examples see Example 2.5. The condition of Theorem 2.1 then becomes $k^\alpha(\ln k + \ln n)/n \to 0$. This condition is fulfilled for $k \propto \left(\frac{n}{(\ln n)^{1+\varepsilon}}\right)^{1/\alpha}$ with $\varepsilon > 0$, and the corresponding rate of convergence is $O((\ln n)^{-\varepsilon/2})$. This example shows that, for cell probabilities

violating $0 < \frac{\gamma_1}{k} \leq p_i \leq \frac{\gamma_2}{k} < 1$, $i = 1, \dots, k$, the frequency estimators are still sparse consistent provided that the table is not too sparse. Variations on this theme are of course possible.

2.2 Sparse consistency of local polynomial estimator

Denote the local polynomial smoothers for the cell probabilities $\boldsymbol{p} = (p_1, \dots, p_k)^T$ by $\widehat{\boldsymbol{P}} = (\widehat{P}_1, \dots, \widehat{P}_k)^T$. We have the following sparse consistency result.

Theorem 2.2
Assume that the cell probabilities are generated by an underlying latent density $f(\cdot)$ on [0,1] through relation (1.9). Further assume,

(C.1) $K(\cdot)$ is a symmetric, continuous kernel with bounded support [-L,L],

(C.2) $h \to 0$, $hk \to \infty$ as $n \to \infty$,

(C.3') $f^{(p+1)}(\cdot)$ is bounded on $[0, 1]$,

and

$$(i)\ A_n^2 = \frac{\ln n + \ln k}{n}\left(\frac{1}{m_n hk} + \frac{1}{m_n^2 k^2}\right) \to 0 \text{ as } n \to \infty,$$

$$(ii)\ B_n = \frac{h^{p+1}}{m_n k} \to 0.$$

The vector $\widehat{\boldsymbol{P}}$ is a sparse consistent estimator for $\boldsymbol{p}$ and the consistency rate is

$$\sup_{1 \leq i \leq k} \left| \frac{\widehat{P}_i}{p_i} - 1 \right| \overset{a.s.}{=} O(A_n + B_n).$$

Proof
We decompose $\frac{\widehat{P}_i}{p_i} - 1$ into a stochastic part $\frac{\widehat{P}_i - E\widehat{P}_i}{p_i}$ and a deterministic part

$\dfrac{E\widehat{P}_i - p_i}{p_i}$. From the bias expression (1.23) for the local polynomial smoother (see Remark 1.5 after Theorem 1.4), we immediately obtain

$$\sup_{1\le i\le k}\left|\frac{E\widehat{P}_i - p_i}{p_i}\right| = O\left(B_n\right). \tag{2.5}$$

For the stochastic component we will show

$$\sup_{1\le i\le k}\left|\frac{\widehat{P}_i - E\widehat{P}_i}{p_i}\right| \overset{a.s.}{=} O(A_n). \tag{2.6}$$

Combining relations (2.5) and (2.6) yields the desired result.
Let, for $\ell = 1,\ldots,n$, $\boldsymbol{X}_{n\ell} = (X_{\ell 1},\ldots,X_{\ell k})^T$ be the triangular array where

$$X_{\ell i} = \frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right)(Y_{\ell j} - p_j), \tag{2.7}$$

with $\boldsymbol{Y}_{n\ell} = (Y_{\ell 1},\ldots,Y_{\ell k})^T$ defined in (2.3). Note that, for each fixed n, $\boldsymbol{X}_{n1},\ldots,\boldsymbol{X}_{nn}$ are independent. We then can write $\widehat{P}_i - E\widehat{P}_i$ as a sum of independent random variables, i.e., $\widehat{P}_i - E\widehat{P}_i = \dfrac{1}{n}\sum_{\ell=1}^{n} X_{\ell i}$.
By using a Bonferroni type inequality we have

$$I\!P\left\{\sup_{1\le i\le k}\left|\frac{\widehat{P}_i - E\widehat{P}_i}{p_i}\right| > \varepsilon\right\} \le k\sup_{1\le i\le k} I\!P\left\{\left|\sum_{\ell=1}^{n}\widetilde{X}_{\ell i}\right| > \varepsilon n\right\},$$

where $\widetilde{X}_{\ell i} = X_{\ell i}/p_i$. It is clear that the variables $\widetilde{X}_{\ell i}$ have mean zero. From its definition and Lemma 1.2 we obtain, with $C_1 > 0$,

$$|\widetilde{X}_{\ell i}| \;\le\; \frac{2}{m_n}\frac{1}{kh}\sup_{1\le i,j\le k}\left|L_{i,p}\left(\frac{x_j - x_i}{h}\right)\right| \le \frac{C_1}{m_n kh}. \tag{2.8}$$

For the variance of $\widetilde{X}_{\ell i}$ we have

$$\begin{aligned}\operatorname{Var}(\widetilde{X}_{\ell i}) &= \frac{1}{p_i^2}\frac{1}{k^2h^2}\left\{\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j - x_i}{h}\right)p_j - \left(\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right)p_j\right)^2\right\} \qquad (2.9)\\ &\le \frac{1}{p_i}\frac{1}{k^2h^2}\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j - x_i}{h}\right)\frac{p_j}{p_i}.\end{aligned}$$

For j with $|x_j - x_i| > Lh$ we have $L_{i,p}\left((x_j - x_i)/h\right) = 0$, and for j with $|x_j - x_i| \leq Lh$ use (1.21) and (1.22) in the proof of Theorem 1.4 to get, with $C > 0$,

$$\frac{p_j}{p_i} \leq 1 + C\frac{h}{kp_i}. \tag{2.10}$$

By using Remark 1.4 we now obtain, with $C_2, C_3 > 0$,

$$\mathrm{Var}(\widetilde{X}_{\ell i}) \;\leq\; \frac{C_2}{m_n kh} + \frac{C_3}{m_n^2 k^2}. \tag{2.11}$$

Use (2.8), (2.11) and Bernstein's inequality to obtain

$$\begin{aligned} & I\!\!P\left\{\left|\sum_{\ell=1}^{n} \widetilde{X}_{\ell i}\right| > \varepsilon n\right\} \\ & \leq 2\exp\left\{-\frac{1}{2}\varepsilon^2 n^2 \left(\frac{nC_2}{m_n kh} + \frac{nC_3}{m_n^2 k^2} + \frac{C_1 n\varepsilon}{3m_n kh}\right)^{-1}\right\} \\ & = 2\exp\left\{-\frac{1}{2}\varepsilon^2 n m_n kh \left(C_2 + \frac{1}{3}C_1\varepsilon + \frac{C_3 h}{m_n k}\right)^{-1}\right\}. \end{aligned}$$

For $0 < \varepsilon < \dfrac{3C_2}{C_1}$ we then have

$$I\!\!P\left\{\left|\sum_{\ell=1}^{n} \widetilde{X}_{\ell i}\right| > \varepsilon n\right\} \leq 2\exp\left\{-\frac{1}{2}n\varepsilon^2 m_n kh \left(2C_2 + \frac{C_3 h}{m_n k}\right)^{-1}\right\}.$$

Now, take $\varepsilon_n = \sqrt{2\dfrac{(1+\delta)\ln n + \ln k}{n}\left(\dfrac{2C_2}{m_n kh} + \dfrac{C_3}{m_n^2 k^2}\right)}$ with $\delta > 0$. For n sufficiently large we have $0 < \varepsilon_n < \dfrac{3C_2}{C_1}$ by (i). Therefore, we obtain, for N sufficiently large,

$$\begin{aligned}
&\sum_{n=N}^{\infty} I\!P\left\{\sup_{1\leq i\leq k}\left|\frac{\widehat{P}_i - E\widehat{P}_i}{p_i}\right| > \varepsilon_n\right\} \\
&\leq \sum_{n=N}^{\infty} k \sup_{1\leq i\leq k} I\!P\left\{\left|\sum_{\ell=1}^{n}\widetilde{X}_{\ell i}\right| > \varepsilon_n n\right\} \\
&\leq 2\sum_{n=1}^{\infty} k\exp(-(1+\delta)\ln n)\exp(-\ln k) \\
&= 2\sum_{n=1}^{\infty} n^{-1-\delta} < \infty.
\end{aligned}$$

This complete convergence result implies (2.6). ■

We compare our sparse consistency result for local polynomial smoothers to the result obtained by Simonoff (1983) for maximum penalized likelihood estimators. We first give a short discussion on maximum penalized likelihood estimators, for more details see Tapia and Thompson (1978, Chapters 4 and 5).

Maximum penalized likelihood estimators

The maximum penalized likelihood method is a nonparametric technique to define estimators for a density function $f(\cdot)$. Let $X_1, \ldots, X_n$ represent a random sample from the unknown density $f(\cdot)$. This density $f(\cdot)$ is considered to belong to a certain class H of functions, where H is usually defined in terms of smoothness conditions. An estimator $\hat{f}(\cdot)$ for $f(\cdot)$ is defined as the function that maximizes

$$L \;=\; \sum_{\ell=1}^{n} \ln f(X_l) - \beta\Phi(f)$$

subject to the constraints

$$f(\cdot) \in H \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1 \qquad \text{and} \qquad f(\cdot) \geq 0.$$

The term $\Phi(f)$ is a nonnegative roughness penalty that becomes smaller as $f(\cdot)$ becomes smoother, β is a smoothing parameter and the object function L is called the penalized log likelihood. A very natural penalty is $\Phi(f) = \|f\|^2$, with $\|\cdot\|$ a norm on H. Any solution to the optimization problem is called a maximum penalized likelihood estimator. The maximum penalized likelihood method has been proposed by Good and Gaskins (1971). They don't consider $f(\cdot)$ to belong to a class H, and in order to avoid the constraint $f(\cdot) \geq 0$ in the maximization procedure they take

$$\Phi(f) = \int_{-\infty}^{\infty} \left(\frac{df^{1/2}(x)}{dx} \right)^2 dx,$$

which requires $f(\cdot) \geq 0$ implicitely.

Existence and uniqueness of the maximum penalized likelihood estimator are obtained by de Montricher, Tapia and Thompson (1975) for different choices of the roughness penalty $\Phi(f)$, including for the proposal of Good and Gaskins (1971).

For computational considerations Scott, Tapia and Thompson (1980) introduce a discrete version of the maximum penalized likelihood method (based on the natural penalty), by discretizing the continuous data into subsequent intervals. This formulation of the problem is closely related to the problem of estimating cell probabilities for multinomial data. Their proposal is to consider the following constrained optimization problem (in terms of our notation)

$$\text{maximize} \qquad L = \sum_{i=1}^{k} N_i \ln p_i - \beta k^3 \sum_{i=1}^{k} (p_i - p_{i+1})^2$$

subject to the $\sum_{i=1}^{k} p_i = 1$ and $p_i \geq 0$, $i = 1, \ldots, k$, where $p_{k+1} = 0$. Scott et al. (1980) established the consistency of the discrete maximum penalized likelihood estimator if $k \propto n^q$ with $0 < q < 1/4$.

Simonoff (1983) considers the following optimization problem

$$\text{maximize} \qquad L = \sum_{i=1}^{k} N_i \ln p_i - \beta \sum_{i=1}^{k-1} (\ln p_i - \ln p_{i+1})^2, \tag{2.12}$$

subject to the constraint $\sum_{i=1}^{k} p_i = 1$. Note that the motivation to consider the penalty function in terms of logarithms is similar to the motivation of Good and Gaskins (1971). Let $\widehat{\boldsymbol{P}}^M = (\widehat{P}_1^M, \ldots, \widehat{P}_k^M)^T$ denote the solution to (2.12). Simonoff shows that this estimator is uniquely defined, but there is no closed form expression for $\widehat{\boldsymbol{P}}^M$. The smoothing parameter β in (2.12) does not play its usual role, i.e., $\beta \to 0$, since k is absorbed into its definition. The next theorem states that this maximum penalized likelihood estimator is sparse consistent when k and n grow at the same rate.

Theorem 2.3 (Simonoff (1983, 1995a))
Let $n, k, \beta \to \infty$ such that k and n grow at the same rate (i.e., $k \propto n$) and $k^{4/3}(\ln k)^{2/3}\beta^{-1} \to 0$ and $\beta k^{-2} \to 0$. Assume the cell probabilities satisfy

$$0 < \frac{\gamma_1}{k} \leq p_i \leq \frac{\gamma_2}{k} < 1 \quad i = 1, \ldots, k. \tag{2.13}$$

Further assume the smoothness constraint

$$\sup_{1 \leq i \leq k} \left| \ln \left(\frac{p_{i-1}p_{i+1}}{p_i^2} \right) \right| = O(k^{-2}) \tag{2.14}$$

and the boundary conditions

$$\left| \ln \left(\frac{p_1}{p_2} \right) \right| = O(k^{-2})$$

and

$$\left| \ln \left(\frac{p_{k-1}}{p_k} \right) \right| = O(k^{-2}).$$

Then

$$\sup_{1 \leq i \leq k} \left| \frac{\widehat{P}_i^M}{p_i} - 1 \right| = O_P(\beta^{-1/4}(\ln k)^{1/2} + \beta k^{-2}).$$

Taking β of the order $k^{8/5}(\ln k)^{2/5}$ results in the convergence rate

$$\sup_{1 \leq i \leq k} \left| \frac{\widehat{P}_i^M}{p_i} - 1 \right| = O_P(k^{-2/5}(\ln k)^{2/5}).$$

First note that this is a consistency, in probability, result. From the proof in Simonoff (1983) it is easy to show that the result is in fact an almost sure result.

Remark 2.2
In order to compare our result for local polynomial smoothers (Theorem 2.2) to that for maximum penalized likelihood estimators (Theorem 2.3), we first give a short discussion on the assumptions of both theorems.

In Theorem 2.2 the smoothness assumption on the cell probability vector $\boldsymbol{p}$ is stated in terms of the underlying latent density $f(\cdot)$. In Theorem 2.3 smoothness is expressed in terms of consecutive cell probabilities p_{i-1}, p_i, p_{i+1}. However, this assumption (2.14) is roughly equivalent to the existence of an underlying latent density $f(\cdot)$ on [0,1] with a bounded second derivative, which is condition $(C.3')$ for $p = 1$. To see this, note that the condition

$$f''(\cdot) \text{ is bounded on } [0,1] \tag{2.15}$$

together with (2.13) imply smoothness condition (2.14). Indeed, use (2.13), $\ln x = x - 1 + O((x-1)^2)$ and a Taylor expansion of p_{i-1}, p_i and p_{i+1} around x_i to obtain

$$\begin{aligned}\ln \frac{p_{i-1}}{p_i} = & \\ & \frac{1}{p_i}\left(\frac{-f'(x_i)}{k^2} + \int_{\frac{i-2}{k}}^{\frac{i-1}{k}} \frac{f''(\xi_{i-1}(x))}{2}(x-x_i)^2\,dx - \int_{\frac{i-1}{k}}^{\frac{i}{k}} \frac{f''(\xi_i(x))}{2}(x-x_i)^2\,dx\right) \\ & + O\left(\frac{1}{k^2}\right)\end{aligned} \tag{2.16}$$

and

$$\begin{aligned}\ln \frac{p_{i+1}}{p_i} = & \\ & \frac{1}{p_i}\left(\frac{f'(x_i)}{k^2} + \int_{\frac{i}{k}}^{\frac{i+1}{k}} \frac{f''(\xi_{i+1}(x))}{2}(x-x_i)^2\,dx - \int_{\frac{i-1}{k}}^{\frac{i}{k}} \frac{f''(\xi_i(x))}{2}(x-x_i)^2\,dx\right) \\ & + O\left(\frac{1}{k^2}\right),\end{aligned} \tag{2.17}$$

where $\xi_{i-1}(x)$, $\xi_i(x)$ and $\xi_{i+1}(x)$ are points between x and x_i. Combining (2.16) and (2.17) yields

$$\begin{aligned}\ln\left(\frac{p_{i-1}p_{i+1}}{p_i^2}\right) = & \\ & \frac{1}{p_i}\left(\int_{\frac{i-2}{k}}^{\frac{i-1}{k}} \frac{f''(\xi_{i-1}(x))}{2}(x-x_i)^2\,dx + \int_{\frac{i}{k}}^{\frac{i+1}{k}} \frac{f''(\xi_{i+1}(x))}{2}(x-x_i)^2\,dx \right. \qquad (2.18) \\ & \left. -2\int_{\frac{i-1}{k}}^{\frac{i}{k}} \frac{f''(\xi_i(x))}{2}(x-x_i)^2\,dx\right) + O\left(\frac{1}{k^2}\right).\end{aligned}$$

Using (2.13) and (2.15) we then obtain

$$\sup_{1\leq i\leq k}\left|\ln\left(\frac{p_{i-1}p_{i+1}}{p_i^2}\right)\right| = O(k^{-2}).$$

Further, if (2.15) is not valid also condition (2.14) is not satisfied. To prove this, assume (2.13) and $f''(\cdot)$ continuous on (0,1), but not bounded at zero. We can repeat the arguments leading to (2.18) and from thereon, using the continuity of $f''(\cdot)$ and (2.13), we obtain

$$\ln\left(\frac{p_{i-1}p_{i+1}}{p_i^2}\right) = \quad \frac{1}{p_i}\left(\frac{f''(x_{i-1})}{k^3}+O(k^{-3})\right)+O\left(\frac{1}{k^2}\right).$$

Since $f''(x_1)\to\infty$ as $k\to\infty$ and (2.13) we do not have

$$\sup_{1\leq i\leq k}\left|\ln\left(\frac{p_{i-1}p_{i+1}}{p_i^2}\right)\right| = O(k^{-2}).$$

We now formulate the boundary conditions in Theorem 2.3 in terms of latent density assumptions. Condition (2.13) is equivalent to $f(\cdot)$ being bounded away from zero (use the continuity of $f(\cdot)$). Assuming (2.13) and (2.15) it becomes obvious from (2.16) and (2.17) that the boundary conditions in Theorem 2.3 are equivalent to $f'(0) = f'(1) = 0$. This last condition is typical for avoiding boundary problems. In the study of the rate of convergence of the mean sum of squared errors of classical kernel estimators for cell probabilities, Hall and Titterington (1987) and Burman (1987a) also need this boundary condition (see Section 3.1 for more details).

Based on this discussion we conclude that the conditions on the latent density $f(\cdot)$ in Theorem 2.2 are less restrictive than in Theorem 2.3. Moreover, we do not need in our Theorem 2.2 the sparseness condition $k\propto n$.

Example 2.4
We reconsider Example 2.2, i.e., we assume the cell probabilities satisfy $0<\dfrac{\gamma_1}{k}\leq p_i\leq\dfrac{\gamma_2}{k}<1$, $i=1,\ldots,k$. For $k\propto n^q, q>0$, the conditions in Theorem 2.2 reduce to $hn^q\to\infty$, $h\to 0$ and $\ln n/(nh)\to 0$, and we obtain $\sup\limits_{1\leq i\leq k}\left|\dfrac{\widehat{P}_i}{p_i}-1\right| \overset{a.s.}{=} O\left(\sqrt{\dfrac{\ln n}{nh}}+h^{p+1}\right)$. The best rate we can obtain with this result is by taking $h\propto\left(n^{-1}\ln n\right)^{\frac{1}{2p+3}}$, and the rate becomes $O\left(\left(n^{-1}\ln n\right)^{\frac{p+1}{2p+3}}\right)$. Note that this best

rate can only be obtained when $q \geq 1/(2p+3)$, since the condition $hn^q \to \infty$ needs to be satisfied. This illustrates that local polynomial smoothers have a faster rate of convergence than the frequency estimators (see Example 2.2), for tables with degree of sparseness $k \propto n^q$, $1/(2p+3) \leq q < 1$.

For less sparse tables, i.e., $q < 1/(2p+3)$, the rate of convergence for the local polynomial smoothers becomes $O\left(h^{p+1}\right)$, which is slower than the corresponding rate for the frequency estimators. In Chapter 3 we will see that the benefit of smoothing, in terms of rates of convergence of mean sum of squared errors, starts at the same degree of sparseness.

For $p = 0$ (the local polynomial smoother is now the classical kernel estimator) the convergence rate is $O(n^{-1/3}(\ln n)^{1/3})$, a result in correspondence with rate of convergence results in nonparametric regression estimation (see e.g. Stone (1988)).

For $p = 1$ (the local linear smoother) we obtain $O(n^{-2/5}(\ln n)^{2/5})$, this rate is in correspondence with the rate obtained in Theorem 2.3, but under weaker assumptions (see Remark 2.2).

Further note that the degree of sparseness we can allow in this example, $k \propto n^q$ for any $q > 0$, even includes supersparseness (i.e. $k/n \to \infty$). Compared to the findings in Examples 2.1 and 2.2 this clearly demonstrates the beneficial effect of smoothing.

Example 2.5

We reconsider Example 2.3, where $p_i \propto k^{-1}$ for $i = 1, \ldots, k$ is no longer valid. Instead, we assume $m_n \propto k^{-\alpha}$. Since, to apply the result in Theorem 2.2, also condition $(C.3')$ needs to be satisfied, we can only consider cell probabilities $\boldsymbol{p}$ for which $m_n \propto k^{-\alpha}$ for $\alpha \geq 2$. An example of a family of latent densities generating such $\boldsymbol{p}$ are $f(u) = \alpha u^{\alpha-1} \mathbb{1}\{0 \leq u \leq 1\}$, $\alpha \geq 2$.

The conditions in Theorem 2.2 reduce to $h \to 0$, $kh \to \infty$, $k^{\alpha-1}h^{p+1} \to 0$ and

$$\frac{\ln n + \ln k}{n} k^{2\alpha-2} \to 0.$$

It readily follows that strong consistency is guaranteed if we take

$$k \propto \left(\frac{n}{(\ln n)^{1+\varepsilon}}\right)^{\frac{1}{2\alpha-2}}, \text{ for any } \varepsilon > 0,$$

and h such that $kh \to \infty$ and $k^{\alpha-1}h^{p+1} \to 0$. The rate becomes $O\left((\ln n)^{-\varepsilon/2}\right)$.

For $h \propto n^{-\frac{1}{2p+2}}$ the condition $k^{\alpha-1}h^{p+1} \to 0$ is satisfied and $kh \to \infty$ is equivalent to $p > \alpha - 2$.

Let us compare the degree of sparseness obtained for the local polynomial smoothers to the one obtained for the frequency estimators (Example 2.3). We note that the frequency estimators are guaranteed to be sparse consistent for a higher degree of sparseness than the local polynomial smoothers. Does this mean that smoothing is not beneficial in tables where the smallest cell probability tends to zero faster than k^{-1}; or can the result of Theorem 2.2 be improved? Indeed, the next example shows that for a specific family of densities this result can be sharpened.

Example 2.6
Consider the family of latent densities

$$f(u) = \alpha u^{\alpha-1}\mathbb{1}\{0 \le u \le 1\}, \ \alpha \in I\!N \text{ and } \alpha \ge 2.$$

Note that the term $(m_n^2 k^2)^{-1}$ in condition (i) of Theorem 2.2 is the reason for restricting the degree of sparseness in Example 2.5. Recall from the proof of Theorem 2.2 that this term originates from the approximation (2.11) for the variance of the variables $\widetilde{X}_{\ell i}$. In this example we will derive a more accurate approximation.

For our specific family of latent densities we have, for $0 \le u \le 1$,

$$f^{(\ell)}(u) = \frac{\alpha!}{(\alpha-1-\ell)!}u^{\alpha-1-\ell} \qquad \ell = 1,\dots,\alpha-1$$

and $f^{(\alpha)}(u) = 0$. A Taylor expansion yields,

$$p_j = \sum_{\ell \text{ even}}^{\ell=0,\alpha-1} \frac{f^{(\ell)}(x_j)}{(\ell+1)!2^\ell k^{\ell+1}}$$

and

$$f^{(\ell)}(x_j) = \sum_{r=0}^{\alpha-1-\ell} \frac{f^{(\ell+r)}(x_i)}{r!}(x_j - x_i)^r.$$

Note that, compared to (1.19) and (1.20), these relations are now exact (see also Remark 1.6). Use these relations in expression (2.9) for the variance of $\widetilde{X}_{\ell i}$, to obtain (for $p \ge \alpha - 1$)

$$\begin{aligned}\mathrm{Var}(\widetilde{X}_{\ell i}) \ &= \ \frac{1}{p_i k h}\left\{\frac{1}{kp_i}\sum_{\ell \text{ even}}^{\ell=0,\alpha-1}\frac{1}{(\ell+1)!2^\ell k^\ell}\left(\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j-x_i}{h}\right)\times\right.\right.\\ &\qquad \left.\left.\sum_{r=0}^{\alpha-1-\ell}\frac{f^{(\ell+r)}(x_i)}{r!}(x_j-x_i)^r\right)\right\} - 1.\end{aligned}$$

Since $kp_i \geq f(x_i)$ and

$$\frac{f^{(\ell+r)}(x_i)}{f(x_i)} = \frac{(\alpha-1)!}{(\alpha-1-(\ell+r))!}\frac{1}{x_i^{\ell+r}} \leq \frac{(\alpha-1)!}{(\alpha-1-(\ell+r))!}2^{\ell+r}k^{\ell+r}$$

we obtain

$$\begin{aligned}
\mathrm{Var}(\widetilde{X}_{\ell i}) &\leq \frac{k^\alpha}{kh}\left\{\sum_{\ell \text{ even}}^{\ell=0,\alpha-1} \frac{1}{(\ell+1)!}\left(\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j-x_i}{h}\right)\times\right.\right.\\
&\qquad\left.\left.\sum_{r=0}^{\alpha-1-\ell}\frac{(\alpha-1)!}{(\alpha-1-(\ell+r))!}2^r k^r (x_j-x_i)^r\right)\right\}-1\\
&= O\left(\frac{k^\alpha}{kh}(kh)^{\alpha-1}\right) = O\left(k^{2\alpha-2}h^{\alpha-2}\right).
\end{aligned}$$

Use this relation instead of (2.11) in the sequel of the proof of Theorem 2.2. Theorem 2.2 remains valid if we replace A_n^2 in condition (i) by

$$A_n^2 = \frac{\ln n + \ln k}{n}k^{2\alpha-2}h^{\alpha-2}.$$

We will now investigate which degree of sparseness this new result allows. Consider

$$k \propto \left\{\frac{n}{(\ln n)^{1+\varepsilon}}\right\}^{1/\alpha} \text{ and } h \propto \left\{\frac{(\ln n)^{1+\delta}}{n}\right\}^{1/\alpha},$$

with ε, $\delta > 0$. The condition $A_n \to 0$ is then equivalent to $(\alpha-2)\delta < (2\alpha-2)\varepsilon$, while $kh \to \infty$ to $\delta > \varepsilon$. In order to have $k^{\alpha-1}h^{p+1} \to 0$ we need $p > \alpha - 2$. The convergence rate is $O\left((\ln n)^{-((2\alpha-2)\varepsilon-(\alpha-2)\delta)/(2\alpha)}\right)$.

So, through this example we have shown, for this special class of densities, that we can allow the same degree of sparseness for the local polynomial smoothers as for the frequency estimators, but the rate of convergence for the local polynomial smoothers is slower.

The degree of sparseness we can allow to guarantee sparse asymptotic consistency depends on the structural behavior of the density. We typically have that the degree of sparseness is a decreasing function of α. Moreover the degree of the local polynomial needed to guarantee the sparse consistency increases with α. Based on the examples discussed above, it is clear that a complete characterization for sparse consistency is not yet obtained. Two important open questions are :

> "Is it possible to obtain, for a well defined class $\mathcal{C}_f(p, \alpha)$ of latent densities, an optimal sparse consistency rate?"

> "Is such a result an explanation for the fact that frequency estimators behave better than local polynomial smoothers if $m_n \propto k^{-\alpha}$, $\alpha \geq 2$?"

For mean sum of squared error rates of convergence such an optimality result is available (see Section 3.1). Moreover, it turns out that the local polynomial smoothers are optimal in the sense that they achieve the optimal MSSE rate (see Section 3.2). Anologue results for sparse consistency would be desirable.

Chapter 3

Mean sum of squared errors rates

For k-cell multinomial data, where k is large, the main interest is in the global trend of the p_i's rather than in the individual behavior of cell probabilities. A well established global measure of performance is the sum of squared errors $\mathrm{SSE}(\boldsymbol{P}^*) = \sum_{i=1}^{k}(P_i^* - p_i)^2$, where $\boldsymbol{P}^* = (P_1^*, \ldots, P_k^*)^T$ is an estimator for $\boldsymbol{p} = (p_1, \ldots, p_k)^T$. In this chapter we study the rate of convergence of the mean sum of squared errors, $\mathrm{MSSE}(\widehat{\boldsymbol{P}}) = E(\mathrm{SSE}(\widehat{\boldsymbol{P}}))$, of the local polynomial smoothers $\widehat{\boldsymbol{P}}$ for the cell probabilities in an ordered sparse table (one-dimensional).

In Section 3.1 we give an overview of some work presented in earlier literature about sparse asymptotic MSSE results of classical kernel estimators. To summarize, Hall and Titterington (1987) obtain a general theorem on the optimal rate of convergence to zero of the MSSE. For tables with a certain degree of sparseness, the MSSE of the frequency estimators does not achieve this rate. Further they study kernel type estimators, for which they show that the MSSE of these estimators achieves this optimal rate. Hall and Titterington require rather stringent conditions on the behavior of the vector of true cell probabilities at the boundaries of the table. Burman (1987a) shows that this optimal rate can be achieved under less restrictive boundary conditions. In Section 3.1 we discuss these boundary conditions in more detail.

In Section 3.2 we investigate the MSSE convergence rate for the local polynomial estimators for the cell probabilities. It turns out that the optimal rate is achieved without boundary conditions, when p, the degree of the local polynomial approximation, is odd.

Section 3.3 presents a small simulation study, where the local constant estimator is compared to the local linear and the local cubic smoothers. Further, we briefly discuss two bandwidth selection methods, that are used to illustrate the local linear smoother on real data sets.

3.1 Optimal rate of convergence for MSSE

Hall and Titterington (1987) formulate a general theorem which states that there exists an optimal rate of convergence for the MSSE (see Theorem 3.1 below). For the ease of reading, we present their theorem in the notation used throughout this thesis.

They extend the vector of multinomial cell probabilities $\boldsymbol{p} = (p_1, \ldots, p_k)^T$ to a discrete function $p(\cdot)$ on $\mathbb{Z}$ through

$$p(i) = \begin{cases} p_i & i \in \{1, \ldots, k\} \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

For this discrete function $p(\cdot)$ they assume, for some positive constant C,

$$\sup_{i \in \mathbb{Z}} p(i) \leq \frac{C}{k} \tag{3.2}$$

and, for all $j \in \mathbb{Z}$,

$$\sup_{i \in \mathbb{Z}} \left| p(i+j) - \sum_{\ell=0}^{p} \binom{j}{\ell} \Delta^\ell p(i) \right| \leq \frac{C}{k} \left| \frac{j}{k} \right|^{p+1}, \tag{3.3}$$

where $\Delta p(i) = p(i+1) - p(i)$ and $\Delta^\ell = \Delta \Delta^{\ell-1}$, $\ell \geq 2$. Note that (3.3) is a discrete version of the assumption that $p(\cdot)$ has "derivatives"up to order $p+1$.

The vector of cell probabilities $\boldsymbol{p}$ belongs to the smoothness class $\mathcal{P}_{p+1}(k, C)$, if, with (3.1), assumptions (3.2) and (3.3) are satisfied.

Theorem 3.1 (Hall and Titterington, 1987)
The optimal rate of MSSE of any estimator $\boldsymbol{P}^*$ for $\boldsymbol{p}$, with $\boldsymbol{p}$ in the smoothness class $\mathcal{P}_{p+1}(k, C)$, equals

$$MSSE(\boldsymbol{P}^*) = \begin{cases} O(n^{-1}) & \text{if } n^{-1/(2p+3)}k \to \lambda,\ 0 \leq \lambda < \infty, \\ O(n^{-\frac{2p+2}{2p+3}} k^{-1}) & \text{if } n^{-1/(2p+3)}k \to \infty. \end{cases}$$

First note that the MSSE of the frequency estimators is

$$\begin{aligned}\mathrm{MSSE}(\overline{\boldsymbol{P}}) &= \frac{1}{n}\sum_{i=1}^{k} p_i(1-p_i) \\ &= \frac{1}{n}\left(1-\sum_{i=1}^{k} p_i^2\right)\end{aligned}$$

and hence $(1-\sup_i p_i)/n \le \mathrm{MSSE}(\overline{\boldsymbol{P}}) \le 1/n$. This yields, by (3.2), that the frequency estimators achieve the rate $O(n^{-1})$ and that no faster rate of convergence can be achieved. Therefore, we interpret Theorem 3.1 as follows.

For situations, where k is such that $n^{-1/(2p+3)}k \not\to \infty$ (i.e., the multinomial data are not too sparse) no estimator can improve the frequency estimators, in terms of faster rates of convergence of MSSE. For multinomial data with a higher degree of sparseness (i.e., $n^{-1/(2p+3)}k \to \infty$) it could be possible that there exist estimators, which have a better MSSE convergence rate than that of the frequency estimators, but no faster rate than $O(n^{-(2p+2)/(2p+3)}k^{-1})$ can be achieved. Further, Hall and Titterington (1987) show that an estimator, with optimal MSSE convergence rate exists.

Hall and Titterington (1987) propose a kernel estimator of the form

$$\widehat{P}_i^{HT} = \frac{1}{kh}\sum_{j=1}^{k} K_{kh}\left(\frac{j-i}{kh}\right)\overline{P}_j,$$

where $K_{kh}(\cdot/(kh))$ is a slight modification of a kernel function $K(\cdot)$. They show that this estimator achieves the optimal MSSE convergence rate, when using a kernel $K(\cdot)$ of order $(p+1)$. Next, they study, for smoothness class $\mathcal{P}_{p+1}(k,C)$, the asymptotic expansion of the MSSE of their estimator. From hereon, they assume that the cell probabilities are generated by a latent density $f(\cdot)$ on [0,1] (see also (1.9)) and that $f''(\cdot)$ is continuous on [0,1]. Further, they require rather strong conditions on the boundary behavior of $f(\cdot)$, i.e.,

$$f(0) = f''(0) = f(1) = f''(1) = 0 \tag{3.4}$$

and

$$f'(0) = f'(1) = 0. \tag{3.5}$$

Burman(1987a) shows that, when using an estimator of Nadaraya-Watson type, and assuming $f''(\cdot)$ continuous on [0,1], (3.5) is sufficient to obtain the optimal convergence rate for MSSE.

The boundary conditions needed for the classical kernel estimators to achieve optimal MSSE rate are often not satisfied. Dong and Simonoff (1994) show that the optimal rate can be attained without any boundary condition on $f(\cdot)$, if boundary kernels (as developed in Gasser and Müller (1979)) are used. In the interior region, a kernel $K(\cdot)$ of order 2 is used, while in the boundary region a specially constructed kernel, based on $K(\cdot)$, is used. In order to satisfy appropriate moment conditions, these boundary kernels take negative values at a small region of their support. The consequence is that negative estimates for boundary cell probabilities cannot be excluded. Furthermore, every choice of kernel function $K(\cdot)$ requires its own boundary corrected version (see Table 1 in Dong and Simonoff (1994)).

Rajagopalan and Lall (1995) define a discrete kernel estimator for the cell probabilities. Their estimator has the form $\widehat{P}_i^{RL} = \sum_{j=1}^{k} K((i-j)/s)\overline{P}_j$, where the smoothing parameter s is an integer. They study the special case where the kernel function is of the form $K(u) = au^2 + b$ for $|u| \leq 1$. The coefficients a and b of this kernel are then chosen in such a way that the weights $K((i-j)/s)$ satisfy appropriate discrete moment conditions, which make boundary conditions on the vector of true cell probabilities superfluous. For other choices of the kernel function $K(\cdot)$, the derivations to obtain the coefficients need to be redone. Rajagopalan and Lall (1995) do not investigate the asymptotic behavior of the MSSE of their estimator.

Dong and Simonoff (1995) propose the geometric combination estimator (we only present its simplest form, which they recommend)

$$\widehat{P}_i^{GC}(h) = \left(\widehat{P}_i^{BC}(h)\right)^{4/3} \left(\widehat{P}_i^{BC}(2h)\right)^{-1/3},$$

where $\widehat{P}_i^{BC}(h)$ is denoted for the boundary corrected estimator (as defined in Dong and Simonoff (1994)) based on bandwidth h. Since $\widehat{P}_i^{BC}(2h)$ can become negative for boundary cells i, this is also true for $\widehat{P}_i^{GC}(h)$ (due to the exponent $-1/3$). They study convergence in probability of the sum of squared errors, $\mathrm{SSE}(\widehat{\boldsymbol{P}}^{GC})$, and obtain, under the assumption that $f(\cdot)$ has continuous fourth derivative, $\mathrm{SSE}(\widehat{\boldsymbol{P}}^{GC}) = O_P(n^{-8/9}k^{-1})$. For a stronger MSSE rate result, they cannot rely on boundary kernels (due to a technical reason). This implies that boundary conditions on $f(\cdot)$ are required to achieve the optimal MSSE convergence rate.

In the next section we show that local p-th degree polynomial smoothers, with p odd, achieve the optimal rate of convergence of the MSSE if $f^{(p+1)}(\cdot)$ is continuous on $[0,1]$. From the further discussion, it will be seen that, for p even, this optimal rate is attained under the extra boundary condition $f^{(p+1)}(0) = f^{(p+1)}(1) = 0$.

3.2 MSSE of local polynomial smoothers

In order to give a profound discussion on the local polynomial smoothers for cell probabilities it is needed to get an insight in the behavior of the local polynomial kernel function $L_{i,p}(\cdot)$. Recall from Section 1.2.2 that the notions boundary and interior region are important. The boundary region consists of those points whose local neighborhood partially lies outside the design region. In our problem the design region is [0,1] and the design points are $(i-1/2)/k$ which yields that the set I of interior indices and B of boundary indices are :

$$I = \left\{ Lhk + \frac{1}{2} \le i \le (1-Lh)k + \frac{1}{2} \right\} \tag{3.6}$$

and

$$B = \left\{ 1 \le i < Lhk + \frac{1}{2} \right\} \cup \left\{ (1-Lh)k + \frac{1}{2} < i \le k \right\} = B_L \cup B_R. \tag{3.7}$$

A key result in the analysis of local polynomial estimators is the discrete higher order property of the function $L_{i,p}(\cdot)$ (see (1.14)) and its uniform boundedness (see Lemma 1.2). These results hold both for interior and boundary indices.

For interior points x_i (i.e., $i \in I$) some further properties of the weight function $L_{i,p}(\cdot)$ are available.

Lemma 3.1
Assume that $K(\cdot)$ satisfies (C.1). For $i \in I$ we have

(i) $m_{k,\ell}(x_i) = 0$ for ℓ odd,

(ii) $L_{i,p}(u)$ is a symmetric kernel function,

(iii) $C_{k,\ell}(x_i) = 0$ for ℓ odd.

Proof
(i) Since x_i is an interior point, the design points are equidistant and $K(\cdot)$ is defined on $[-L, L]$, we can write

$$\begin{aligned} m_{k,\ell}(x_i) &= \sum_{j=-[Lhk]}^{[Lhk]} \left(\frac{x_{i+j} - x_i}{h} \right)^\ell K\left(\frac{x_{i+j} - x_i}{h} \right) \\ &= \sum_{j=-[Lhk]}^{[Lhk]} \left(\frac{j}{kh} \right)^\ell K\left(\frac{j}{kh} \right) \end{aligned} \tag{3.8}$$

where $[a]$ denotes the integer part of a. By symmetry of $K(\cdot)$, (3.8) is zero for ℓ odd.
(ii) From the definition of $L_{i,p}(u)$ (see (1.12)) it is clear that, to prove (ii), it suffices to show that $|\boldsymbol{M}_{i,p}(u)|$ is symmetric. $\boldsymbol{M}_{i,p}(-u)$ is the same as $\boldsymbol{M}_{i,p}(u)$ but with first column $(1, -u, \ldots, (-u)^p)^T$ instead of $(1, u, \ldots, u^p)^T$. Since (by (i)) the elements of the even rows of $\boldsymbol{M}_{i,p}(-u)$ are zero at the odd places (except for the first), multiplying these even rows by -1 affects only the first and the even columns. The same argument applies for multiplying the even columns by -1. This last operation changes the even columns back to their original values. Hence, we have performed an even number of multiplications by -1 on the rows and columns to transform $\boldsymbol{M}_{i,p}(-u)$ into $\boldsymbol{M}_{i,p}(u)$. This yields $|\boldsymbol{M}_{i,p}(-u)| = |\boldsymbol{M}_{i,p}(u)|$.
(iii) The same argument as in the proof of (i), but now based on the symmetry of the kernel function $L_{i,p}(\cdot)$, can be used to show (iii). ■

Remark 3.1

Note that Lemma 3.1 only holds for interior points, not for boundary points, and that the equidistant design is crucial to prove the result. Further, from (3.8) we can conclude that $m_{k,\ell}(x_i) = m_{k,\ell}(x_j)$, when i and j are both interior indices, which also yields

$$L_{i,p}(u) = L_{j,p}(u). \tag{3.9}$$

Remark 3.2

Property (iii) of Lemma 3.1 plays an important role in the boundary issue of kernel estimation with bounded design support. Recall that the bias of the local polynomial estimator is (see Theorem 1.4)

$$E\widehat{P}_i - p_i = \frac{f^{(p+1)}(x_i)}{(p+1)!} C_{k,p+1}(x_i) \frac{h^{p+1}}{k} + o\left(\frac{h^{p+1}}{k}\right).$$

When p is even, $C_{k,p+1}(x_i) = 0$ for interior points, but not for boundary points. The consequence is that the order of the bias differs for interior and boundary points, which will have an effect on the rate of convergence of the MSSE (see Theorem 3.2 and Remark 3.5 for more details).

Remark 3.3

The local linear estimator (i.e., $p = 1$) can be written as

$$\widehat{P}_i^{LL} = \sum_{j=1}^{k} \frac{m_{k,2}(x_i) - \left(\frac{x_j - x_i}{h}\right) m_{k,1}(x_i)}{m_{k,2}(x_i) m_{k,0}(x_i) - m_{k,1}^2(x_i)} K\left(\frac{x_j - x_i}{h}\right) \overline{P}_j.$$

The factor $LL(x_i,x_j) = (m_{k,2}(x_i) - \left(\frac{x_j-x_i}{h}\right) m_{k,1}(x_i))/(m_{k,2}(x_i)m_{k,0}(x_i) - m_{k,1}^2(x_i))$ reduces, by Lemma 3.1(i), to $m_{k,0}^{-1}(x_i)$ for interior indices, but not for boundary indices. Therefore, for interior points, the local linear estimator is the classical Nadaraya-Watson estimator (i.e., $p = 0$). In general, the local linear smoother can be seen as a particular form of a boundary corrected kernel estimator. See Section 3 in Jones (1993) (in the context of density estimation) and Dong and Simonoff (1994) for a more detailed discussion on boundary corrected methods.
Further note that, for boundary indices i, the factor $LL(x_i,x_j)$ can become negative for some j, which is typical when using boundary corrected kernels.

Asymptotic results of the kernel function $L_{i,p}(\cdot)$ are also useful for the asymptotic investigation of the behavior of MSSE($\widehat{\boldsymbol{P}}$). In Section 1.3 we have seen that, uniformly in the i-index, $m_{k,\ell}(x_i) = \mu_\ell(x_i) + o(1)$ where $\mu_\ell(x_i) = \int_{-\alpha_i}^{\beta_i} v^\ell K(v)\,dv$ with $\alpha_i = x_i/h$, $\beta_i = (1-x_i)/h$.

For interior points, $\alpha_i \geq L$ and $\beta_i \geq L$, which gives $\mu_\ell(x_i) = \mu_\ell = \int_{-L}^{L} v^\ell K(v)\,dv$. Therefore, it is immediate that for interior points

$$L_{i,p}(u) = L_{(p)}(u) + o(1), \tag{3.10}$$

with

$$L_{(p)}(u) = \frac{|\boldsymbol{M}_p(u)|}{|\boldsymbol{N}_p|} K(u)$$

where the $(p+1)\times(p+1)$-matrix $\boldsymbol{N}_p$ has the (r,s) entry equal to μ_{r+s-2} and $\boldsymbol{M}_p(u)$ is the same as $\boldsymbol{N}_p$ but with the first column replaced by $(1,u,\ldots,u^p)^T$.
Note that the order bound in (3.10) is uniformly in the i-index and in u. Therefore, we also have, uniformly in i and j,

$$L_{i,p}\left(\frac{x_j-x_i}{h}\right) = L_{(p)}\left(\frac{x_j-x_i}{h}\right) + o(1). \tag{3.11}$$

Further, from (1.13) we obtain, uniformly in interior indices i,

$$C_{k,\ell}(x_i) \to C_\ell \ = \frac{\begin{vmatrix} \mu_\ell & \mu_1 & \cdots & \mu_p \\ \vdots & \vdots & & \vdots \\ \mu_{p+\ell} & \mu_{p+1} & \cdots & \mu_{2p} \end{vmatrix}}{|\boldsymbol{N}_p|} = \int_{-L}^{L} v^\ell L_{(p)}(v)\,dv. \tag{3.12}$$

The results for the weight function $L_{i,p}(\cdot)$ ((1.14), Lemma 1.2 and Lemma 3.1) carry over to the kernel function $L_{(p)}(\cdot)$. Further note that $L_{(p)}(\cdot)$ is the Lejeune and Sarda (1992) kernel (a kernel of order $p+1$ for p odd and order $p+2$ for p even (see also Ruppert and Wand (1994)).

For (left) boundary points we have $\alpha_i < L$ and $\beta_i \geq L$, which gives

$$\mu_\ell(x_i) = \mu_\ell(\alpha_i) = \int_{-\alpha_i}^{L} v^\ell K(v)\, dv.$$

Therefore, for (left) boundary points we have

$$L_{i,p}(u) \to L_{(p,\alpha_i)}(u) = \frac{|\boldsymbol{M}_{(p,\alpha_i)}(u)|}{|\boldsymbol{N}_{(p,\alpha_i)}|} K(u) \tag{3.13}$$

where the matrices $\boldsymbol{M}_{(p,\alpha_i)}(u)$ and $\boldsymbol{N}_{(p,\alpha_i)}$ are now based on the incomplete moments $\mu_\ell(\alpha_i)$. A similar treatment is possible for the right boundary region.

The next lemma shows how to approximate sums by integrals. The proof is omitted, since it is similar to that of Lemma 1.1.

Lemma 3.2
Let $g_1(\cdot)$ be a continuous function on [0,1], $g_2(\cdot)$ a continuous function on [0,L] and $S \subset \{1, \ldots, k\}$ with $\#S^C = o(k)$. If (C.2), we have

$$\frac{1}{k} \sum_{i \in S} g_1(x_i) = \int_0^1 g_1(u)\, du + o(1)$$

and

$$\frac{1}{kh} \sum_{i \in B_L} g_2\left(\frac{x_i}{h}\right) = \int_0^L g_2(u)\, du + o(1).$$

The main result of this section reads as follows.

Theorem 3.2
Assume (C.1)–(C.3).

For p odd, we have

$$\begin{aligned} MSSE(\widehat{\boldsymbol{P}}) &= \frac{h^{2p+2}}{k}\left(\int_{-L}^{L} v^{p+1}L_{(p)}(v)\,dv\right)^2 \frac{\int_0^1 (f^{(p+1)}(u))^2\,du}{((p+1)!)^2} \\ &\quad + \frac{1}{nhk}\int_{-L}^{L} L^2_{(p)}(v)dv + o\left(\frac{h^{2p+2}}{k}\right) + o\left(\frac{1}{nhk}\right). \end{aligned} \tag{3.14}$$

For p even and $f^{(p+2)}(\cdot)$ bounded, we have

$$\begin{aligned} MSSE(\widehat{\boldsymbol{P}}) &= \frac{h^{2p+3}}{k}\int_0^L \left(\int_{-\alpha}^{L} v^{p+1}L_{(p,\alpha)}(v)\,dv\right)^2 d\alpha \left\{\frac{\left(f^{(p+1)}(0)\right)^2 + \left(f^{(p+1)}(1)\right)^2}{((p+1)!)^2}\right\} \\ &\quad + \frac{1}{nhk}\int_{-L}^{L} L^2_{(p)}(v)dv + o\left(\frac{h^{2p+3}}{k}\right) + o\left(\frac{1}{nhk}\right). \end{aligned} \tag{3.15}$$

Remark 3.4

Assume p odd, for a note on p even we refer to Remark 3.5. By balancing the leading terms in (3.14), it follows that the asymptotic optimal choice of the bandwidth is $h = Cn^{-1/(2p+3)}$, C a positive constant. With this optimal choice of the bandwidth, the corresponding rate for the MSSE becomes $\text{MSSE}(\widehat{\boldsymbol{P}}) = O\left(n^{-(2p+2)/(2p+3)}k^{-1}\right)$. Since (C.2) needs to be satisfied, this choice of the the bandwidth restricts to situations where $n^{-1/(2p+3)}k \to \infty$. Therefore, local polynomial fitting with p odd, yields the optimal rate of convergence (see Theorem 3.1).

For local linear smoothers, denoted by $\widehat{\boldsymbol{P}}^{LL}$, with $h = Cn^{-1/5}$ we have $\text{MSSE}(\widehat{\boldsymbol{P}}^{LL}) = O(n^{-4/5}k^{-1})$. Therefore these estimators provide competitors for the boundary corrected kernel estimators studied in Dong and Simonoff (1994). For local cubic smoothers, denoted by $\widehat{\boldsymbol{P}}^{LC}$, with $h = Cn^{-1/9}$ we have $\text{MSSE}(\widehat{\boldsymbol{P}}^{LC}) = O(n^{-8/9}k^{-1})$, such that these estimators provide alternatives for the geometric combination estimators for the cell probabilities studied in Dong and Simonoff (1995).

As noted in Section 1.3, the weights $L_{i,p}((x_j - x_i)/h)$ can become negative in the interior region as soon as $p > 1$. The consequence is that the resulting estimators are not guaranteed to be nonnegative. For $p = 1$ this problem is restricted to the boundary region (see Remark 3.3). Also the boundary corrected kernel method of Dong and Simonoff (1994) suffers from this problem (see Section 3.1). The geometric combination estimator, based on the boundary corrected kernel estimator, can result in negative estimates, but only in the boundary region. This can be seen as an advantage of the geometric combination estimator compared to the local cubic estimator. But, recall from Section 3.1, that for the geometric combination estimator boundary conditions are required in order to achieve the optimal MSSE convergence rate.

In Remark 3.6 we suggest some estimators that are positive by construction. Their theoretical performance has yet to be studied.

Proof of Theorem 3.2
The mean sum of squared errors can be decomposed as

$$\begin{aligned}\text{MSSE} &= B_I^2 + B_B^2 + V_I + V_B \\ &= \sum_{i\in I}(E\widehat{P}_i - p_i)^2 + \sum_{i\in B}(E\widehat{P}_i - p_i)^2 + \sum_{i\in I}\text{Var}\widehat{P}_i + \sum_{i\in B}\text{Var}\widehat{P}_i\end{aligned} \tag{3.16}$$

with I and B the sets defined in (3.6) and (3.7). In Theorem 1.4 we obtained, uniformly in the i-index, the following asymptotic expansions for the bias and the variance of the local polynomial estimator :

$$E\widehat{P}_i - p_i = \frac{f^{(p+1)}(x_i)}{(p+1)!}C_{k,p+1}(x_i)\frac{h^{p+1}}{k} + o\left(\frac{h^{p+1}}{k}\right) \tag{3.17}$$

and

$$\text{Var}\widehat{P}_i = \frac{f(x_i)}{nk^2h}\frac{1}{kh}\sum_{j=1}^{k}L_{i,p}^2\left(\frac{x_j - x_i}{h}\right) + o\left(\frac{1}{nk^2h}\right).$$

For the contribution of the interior region to the squared bias term, we obtain from (3.12) and Lemma 3.2

$$B_I^2 \;=\; \frac{h^{2p+2}}{k}\left(\int_{-L}^{L} v^{p+1}L_{(p)}(v)dv\right)^2 \frac{\int_0^1 (f^{(p+1)}(u))^2du}{((p+1)!)^2} + o\left(\frac{h^{2p+2}}{k}\right). \tag{3.18}$$

Further, use (3.11) and Lemma 1.1, to obtain for interior indices i

$$\frac{1}{kh}\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j - x_i}{h}\right) = \int_{-L}^{L} L_{(p)}^2(v)\,dv + o(1),$$

which, together with the expression for the variance and Lemma 3.2, yields

$$V_I = \left(\int_{-L}^{L} L_{(p)}^2(v)\,dv\right)\frac{1}{nhk} + o\left(\frac{1}{nhk}\right). \tag{3.19}$$

The cardinality of the boundary region is $\#B \leq 2Lhk$ and, by Remark 1.4 (see Section 1.3), we obtain

$$B_B^2 = O\left(\frac{h^{2p+3}}{k}\right) = o\left(\frac{h^{2p+2}}{k}\right) \tag{3.20}$$

and

$$V_B = O\left(\frac{1}{nk}\right) = o\left(\frac{1}{nhk}\right). \tag{3.21}$$

The result now is immediate from (3.18)–(3.21).

For p even, the term $C_{k,p+1}(x_i)$ in (3.17) is zero for interior points, but not for boundary points (see Lemma 3.1). Therefore, we need a more accurate expansion for the bias at interior points when p is even. Under the assumption that $f^{(p+2)}(\cdot)$ is continuous, this becomes, uniformly in the i-index,

$$\begin{aligned} E\widehat{P}_i - p_i &= \frac{f^{(p+2)}(x_i)}{(p+2)!}C_{k,p+2}(x_i)\frac{h^{p+2}}{k} + o\left(\frac{h^{p+2}}{k}\right) \\ &= O\left(\frac{h^{p+2}}{k}\right), \end{aligned} \tag{3.22}$$

where the last order bound is already valid if $f^{(p+2)}(\cdot)$ is bounded. This yields

$$B_I^2 = O\left(\frac{h^{2p+4}}{k}\right). \tag{3.23}$$

For the (left) boundary points the bias is

$$\begin{aligned} E\widehat{P}_i - p_i &= \frac{f^{(p+1)}(x_i)}{(p+1)!} C_{k,p+1}(x_i) \frac{h^{p+1}}{k} + O\left(\frac{h^{p+2}}{k}\right) \\ &= \frac{f^{(p+1)}(0)}{(p+1)!} C_{k,p+1}(x_i) \frac{h^{p+1}}{k} + O\left(\frac{h^{p+2}}{k}\right). \end{aligned} \tag{3.24}$$

Similar to the derivation of (3.12) we have, for (left) boundary points

$$C_{k,p+1}(x_i) = \int_{-\alpha_i}^{L} v^{p+1} L_{(p,\alpha_i)}(v)\, dv + o(1), \tag{3.25}$$

with $\alpha_i = x_i/h$. Use Lemma 3.2, with $g_2(\alpha) = \left(\int_{-\alpha}^{L} v^{p+1} L_{(p,\alpha)}(v)\, dv\right)^2$, to obtain, uniformly in the i-index,

$$\frac{1}{kh} \sum_{i \in B_L} C^2_{k,p+1}(x_i) = \int_0^L \left(\int_{-\alpha}^{L} v^{p+1} L_{(p,\alpha)}(v)\, dv \right)^2 d\alpha + o(1).$$

This results, for left boundary points, in

$$B^2_{B_L} = \left(\frac{f^{(p+1)}(0)}{(p+1)!}\right)^2 \int_0^L \left(\int_{-\alpha}^{L} v^{p+1} L_{(p,\alpha)}(v)\, dv \right)^2 d\alpha \, \frac{h^{2p+3}}{k} + O\left(\frac{h^{2p+5}}{k}\right). \tag{3.26}$$

A similar analysis is possible for right boundary points. From (3.23) and (3.26) we get the desired result. ■

Remark 3.5

For p even, the bias contribution is completely dominated by the behavior of the latent density at the boundary points 0 and 1. Under the extra boundary condition $f^{(p+1)}(0) = f^{(p+1)}(1) = 0$ and continuity of $f^{(p+2)}(\cdot)$, this dominating effect of the boundary drops out (use (3.22) and(3.26)). The MSSE has the same expression as (3.14) but with p replaced by $p+1$. This is essentialy the argument used by Burman (1987a) in the case $p = 0$. The use of boundary corrected kernels also circumvents the dominating boundary bias, again in the case $p = 0$ this approach is followed by Dong and Simonoff (1994). By using an odd degree local polynomial this boundary problem is avoided in an automatic way.

Remark 3.6
The fact that some of the components of $\widehat{\boldsymbol{P}}$ can become negative is an unfortunate property of the local polynomial estimator. For local linear estimators (and boundary corrected kernel estimators) this problem restricts to components of $\widehat{\boldsymbol{P}}$ from the boundary region (see Remark 3.3). Further, we typically have $\sum_{i=1}^{k} \widehat{P}_i \neq 1$. A simple suggestion might be to work with the following rescaled version of $\widehat{P}_i$, $i = 1, \ldots, k$:

$$\widehat{P}_i^R = \frac{\widehat{P}_i \mathbb{1}\{\widehat{P}_i > 0\}}{\sum\limits_{\ell=1}^{k} \widehat{P}_\ell \mathbb{1}\{\widehat{P}_\ell > 0\}}.$$

The performance of this rescaled version is not investigated theoretically.

Jones and Foster (1996) propose a method to avoid negative estimates for boundary corrected kernel methods in the density estimation context. Let $f(\cdot)$ denote the density which one wants to estimate, $\tilde{f}(x)$ the classical (nonnegative) kernel density estimator for $f(x)$ and $\hat{f}(x)$ a boundary corrected kernel density estimator (which can be negative). They propose a modified boundary corrected estimator by

$$\hat{f}_P(x) = \tilde{f}(x) \exp\left\{\frac{\hat{f}(x)}{\tilde{f}(x)} - 1\right\}$$

which is guaranteed to be nonnegative. They investigate the theoretical performance of the proposed estimator and show that the bias of $\hat{f}_P(x)$ and $\hat{f}(x)$ have the same order of magnitude, if $\lim_{x \to 0} f'^2(x)/f(x) \neq \infty$. The asymptotic variances of $\hat{f}_P(x)$ and $\hat{f}(x)$ are equal, to first order. They also mention that similar ideas could be used to nonnegativise higher order kernels. It would be interesting to study this method on local polynomial smoothers for estimating cell probabilities.

3.3 Local polynomial smoothers in action

The practical implementation of the local polynomial smoother requires the specification of the order p of the polynomial approximation, the kernel function $K(\cdot)$ and the bandwidth h.

From the results in Section 1.2 and 3.2 we know that the asymptotic performance of local polynomial smoothers improves for higher values of p and that odd degree fits are preferable. On the other hand, the variance of the estimator becomes larger for higher p (see also Remark 1.1), and large samples may be required to see a substantial improvement in practical performance. Based on these facts, Wand and Jones (1995, p. 126) suggest to use either $p = 1$ or $p = 3$. Fan and Gijbels (1995) propose a locally adaptive order selection procedure to adjust the order of the polynomial approximation to the local curvature of the unknown regression fuction. In Fan and Gijbels (1996, p. 77) it is mentioned that in many applications the choice $p = 1$ or $p = 3$ suffices, and that an order selection procedure is mainly proposed for recovering spatially inhomogeneous curves. Based on these arguments we decided to demonstrate the practical performance of local linear and local cubic smoothers.

As known from kernel density and regression estimation the bandwidth h is a crucial parameter for the practical performance of kernel-type estimators. In terms of MSSE performance the optimal amount of smoothing is defined as the bandwidth h that minimizes the mean sum of squared errors. Since the exact expression for MSSE depends on the bandwidth in a complicated way we do not have a closed form expression for the optimal bandwidth. The asymptotic approximation for MSSE given in Theorem 3.2 has a very simple expression that allows to derive an asymptotically optimal bandwidth. Let

$$\text{AMSSE} = \frac{h^{2p+2}}{k} \frac{\mu_{p+1}^2(L_{(p)})}{((p+1)!)^2} R(f^{(p+1)}) + \frac{R(L_{(p)})}{nkh} \tag{3.27}$$

with, for a function $g(\cdot)$, $\mu_{p+1}(g) = \int u^{p+1} g(u)\,du$ and $R(g) = \int g^2(u)\,du$, where integration is over the support of $g(\cdot)$.

We use AMSSE as notation for the first order asymptotic approximation to MSSE. From (3.27) it is clear that using smaller bandwidths would decrease the leading bias term, but at the same time increase the variance. This phenomenon is known as the bias-variance trade-off of the smoothing parameter. Minimizing AMSSE w.r.t. h gives

$$h_{\text{opt}} = \left\{ n^{-1} C_p(K) R(f^{(p+1)})^{-1} \right\}^{1/(2p+3)} \tag{3.28}$$

where

$$C_p(K) = \frac{R(L_{(p)})}{\mu_{p+1}^2(L_{(p)})} \frac{((p+1)!)^2}{2p+2}.$$

Given the degree p and the kernel function $K(\cdot)$, the constant $C_p(K)$ is obtained by direct calculation, e.g. for the Epanechnikov kernel $K(u) = 0.75(1-u^2)\mathbb{1}\{|u| < 1\}$ $C_1(K) = 1.719$ and $C_3(K) = 3.243$ (see Fan and Gijbels (1996, p. 67)).

The asymptotically optimal bandwidth (3.28) also depends on the unknown quantity $R(f^{(p+1)})$. In a simulation study this quantity is known so that h_{opt} can be calculated. In real data applications further work is needed to achieve a practical bandwidth selection rule. We will come back to this topic further in this section.

We also need to choose the kernel function $K(\cdot)$. For the optimal bandwidth (3.28), AMSSE becomes

$$AMSSE = (2p+3)k^{-1}n^{-\frac{2p+2}{2p+3}}\left(\frac{R(L_{(p)})^{2p+2}\mu_{p+1}^2(L_{(p)})R(f^{(p+1)})}{((p+1)!)^2(2p+2)^{2p+2}}\right)^{1/(2p+3)},$$

which depends on the kernel function $K(\cdot)$ through

$$T_p(K) = R(L_{(p)})^{2p+2}\mu_{p+1}^2(L_{(p)}).$$

Fan et al. (1997) have shown that the Epanechnikov kernel is the optimal kernel, in the sense that it minimizes $T_p(K)$ over all nonnegative, symmetric and Lipschitz continuous functions.

We performed a small simulation study to verify whether the superior theoretical performance of the local linear and local cubic estimator is noticeable for moderate-sized sparse tables.

We considered two underlying latent densities to form the cell probabilities in the table, namely the $B(3,3)$-density

$$f_B(u) = 30u^2(1-u)^2\mathbb{1}\{0 \le u \le 1\}$$

and the exponential-like density

$$f_E(u) = 5(1-e^{-5})^{-1}e^{-5u}\mathbb{1}\{0 \le u \le 1\}.$$

In Figures 3.1 and 3.2 we compare the SSE performance of the local constant smoother to the local linear smoother for the latent densities $f_B(\cdot)$ and $f_E(\cdot)$. For the bandwidth we use formula (3.28) with $p = 1$. The figures show boxplots representing the differences $n(\text{SSE}(\widehat{\boldsymbol{P}}^{NW}) - \text{SSE}(\widehat{\boldsymbol{P}}^{LL}))$, where $\widehat{\boldsymbol{P}}^{NW}$ is the local constant smoother, and $\widehat{\boldsymbol{P}}^{LL}$ the local linear smoother. Positive values of the difference indicate that the local linear smoother is globally more accurate than the local constant smoother. We considered tables with $k = 10,\ 20,\ 50$ and 100 and n such that $n/k = 1,\ 2,\ 5$.

Both Figures 3.1 and 3.2 clearly demonstrate the superior SSE performance of local linear to local constant fits.

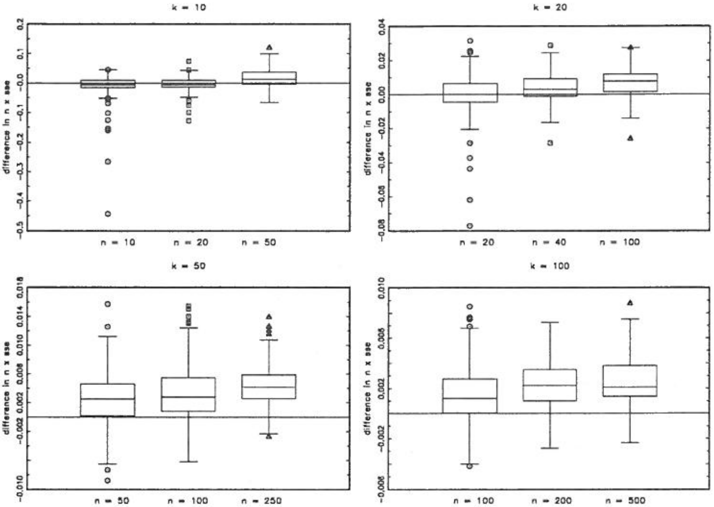

Figure 3.1: *For latent density $f_B(\cdot)$, boxplots of $n(SSE(\widehat{\boldsymbol{P}}^{NW}) - SSE(\widehat{\boldsymbol{P}}^{LL}))$ based on 100 simulations. Positive values of the difference indicate that the local linear smoother $\widehat{\boldsymbol{P}}^{LL}$ is globally more accurate than the local constant smoother $\widehat{\boldsymbol{P}}^{NW}$.*

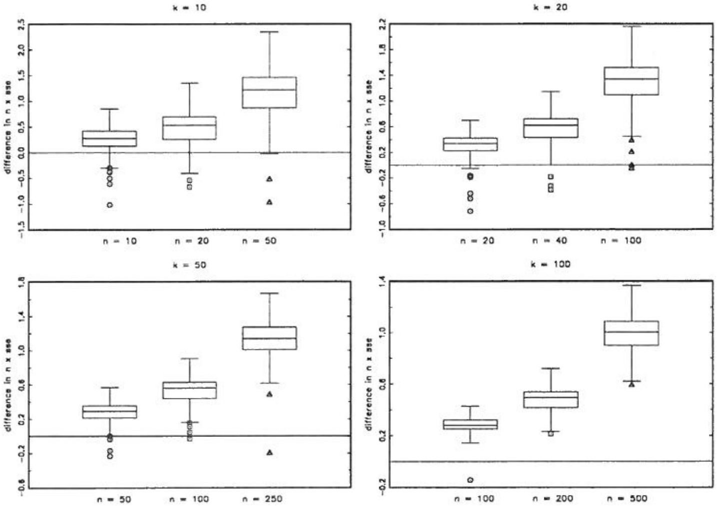

Figure 3.2: *For latent density $f_E(\cdot)$, boxplots of $n(SSE(\widehat{\boldsymbol{P}}^{NW}) - SSE(\widehat{\boldsymbol{P}}^{LL}))$ based on 100 simulations. Positive values of the difference indicate that the local linear smoother $\widehat{\boldsymbol{P}}^{LL}$ is globally more accurate than the local constant smoother $\widehat{\boldsymbol{P}}^{NW}$.*

Figures 3.3 and 3.4 illustrate the performance at individual cells. For $k = 10$ the real cell probabilities (the solid vertical lines) are shown. For 100 simulations runs with $n = 20$ we obtain $\widehat{P}_{i(j)}$, the local polynomial smoothers based on the j-th simulation. The plus signs (+) represent the means $\widehat{P}_{i\cdot}$ of the $\widehat{P}_{i(j)}$'s for $p = 1$ (the local linear smoothers) and the squares (■) represent these means for $p = 3$ (the local cubic smoothers), $i = 1, \ldots, 10$. In Figure 3.3 the circles (●) represent the means for $p = 0$ (the local constant smoothers). For latent density $f_B(\cdot)$ the boundary conditions $f'(0) = f'(1) = 0$ are satisfied. Therefore (based on Remark 3.5), expression (3.28) for the asymptotically optimal bandwidth with $p = 1$ is valid. Since this is not the case for latent density $f_E(\cdot)$, we dropped the local constant smoother in Figure 3.4.

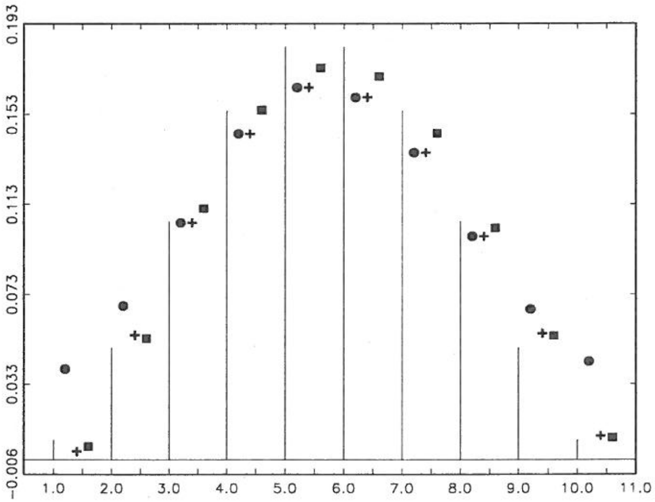

Figure 3.3: *Comparison of the mean of the $\widehat{P}_i$'s for $p = 0(\bullet), 1(+), 3(\blacksquare)$ for a multinomial with cell probabilities generated from $f_B(\cdot)$.*

Figures 3.3 and 3.4 both illustrate that increasing the degree of the local polynomial approximation reduces the bias, and Figure 3.3 demonstrates the superior boundary behavior of local linear fits to local constants fits.

Tables 3.1 and 3.2 show for $f_B(\cdot)$ the exact bias

$$E\widehat{P}_i - p_i = \frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right) p_j - p_i$$

and the exact variance

$$\mathrm{Var}(\widehat{P}_i) = \frac{1}{nk^2h^2}\left\{\sum_{j=1}^{k} L_{i,p}^2\left(\frac{x_j - x_i}{h}\right) p_j - \left(\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right) p_j\right)^2\right\}.$$

These tables also demonstrate that higher degree fits reduce the bias but at the cost of increasing the variance. From Table 3.1 we see that for the boundary cells the local constant smoother has remarkably large bias compared to the local linear smoother. Since $f_B(\cdot)$ satisfies the boundary condition $f'(0) = f'(1) = 0$, asymptotically the local constant estimator has the same performance as the local linear smoother (see Remark 3.5). Therefore, we expect that, for local constant smoothers, increasing the values of k and n will lead to a less pronounced bias at the boundary. Indeed, Table 3.2 ($k = 40$, $n = 80$) provides an illustration of this fact.

Further, note from Tables 3.1 and 3.2 that the exact squared bias and variance contribution to MSSE are not in balance, which they should be asymptotically. Dong and Ye (1996) note that this is typical for tables with $10 \leq k \leq 500$. They advocate that the uniform kernel function $K(\cdot) = 0.5\mathbb{1}\{|u| \leq 1\}$ should be used in order to reduce this phenomenon. Their motivation is that the uniform kernel minimizes the asymptotic variance term $R(L_{(p)})$.

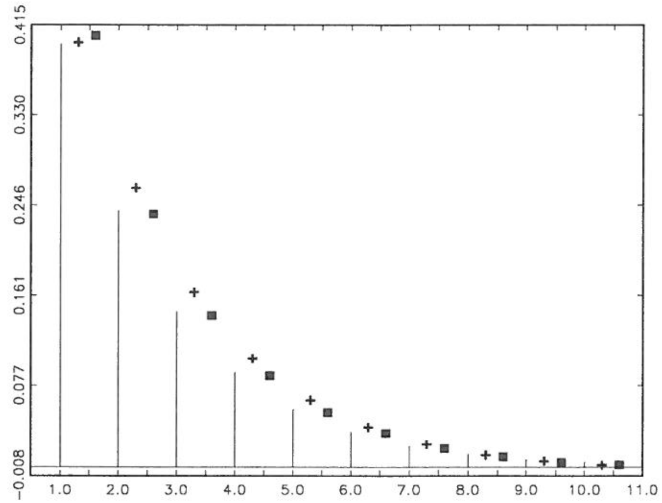

Figure 3.4: *Comparison of the mean of the* $\widehat{P}_i$*'s for* $1(+), 3(■)$ *for a multinomial with cell probabilities generated from* $f_E(\cdot)$*.*

REAL BIAS			REAL VARIANCE		
$p=0$	$p=1$	$p=3$	$p=0$	$p=1$	$p=3$
0.03188	-0.00179	-0.00034	0.00051	0.00054	0.00050
0.01703	0.00496	0.00271	0.00059	0.00050	0.00159
-0.00372	-0.00372	0.00049	0.00066	0.00066	0.00159
-0.01355	-0.01355	-0.00434	0.00067	0.00067	0.00094
-0.01847	-0.01847	-0.00960	0.00059	0.00059	0.00047

Table 3.1: *Latent density* $f_B(\cdot)$*,* $k=10$*,* $n=20$*. We only give the first 5 cells (then use symmetry).*

REAL BIAS			REAL VARIANCE		
$p = 0$	$p = 1$	$p = 3$	$p = 0$	$p = 1$	$p = 3$
0.00474	-0.00129	-0.00019	0.00001	0.00001	0.00001
0.00513	-0.00003	0.00004	0.00001	0.00001	0.00000
0.00494	0.00078	0.00020	0.00001	0.00001	0.00001
0.00432	0.00123	0.00029	0.00001	0.00001	0.00002
0.00341	0.00137	0.00031	0.00001	0.00001	0.00002
0.00237	0.00128	0.00028	0.00001	0.00001	0.00002
0.00136	0.00099	0.00021	0.00001	0.00001	0.00002
0.00057	0.00057	0.00010	0.00001	0.00001	0.00003
0.00007	0.00007	-0.00005	0.00001	0.00001	0.00002
-0.00039	-0.00039	-0.00021	0.00002	0.00002	0.00002
-0.00081	-0.00081	-0.00039	0.00002	0.00002	0.00002
-0.00118	-0.00118	-0.00058	0.00002	0.00002	0.00002
-0.00151	-0.00151	-0.00077	0.00002	0.00002	0.00002
-0.00181	-0.00181	-0.00096	0.00002	0.00002	0.00002
-0.00206	-0.00206	-0.00115	0.00002	0.00002	0.00001
-0.00226	-0.00226	-0.00132	0.00002	0.00002	0.00001
-0.00243	-0.00243	-0.00147	0.00002	0.00002	0.00001
-0.00255	-0.00255	-0.00158	0.00002	0.00002	0.00001
-0.00264	-0.00264	-0.00164	0.00002	0.00002	0.00001
-0.00268	-0.00268	-0.00164	0.00002	0.00002	0.00001

Table 3.2: *Latent density $f_B(\cdot)$, $k = 40$, $n = 80$. We only give the first 20 cells (then use symmetry).*

The expression for the asymptotically optimal bandwidth given by (3.28), contains $f^{(p+1)}(\cdot)$. Therefore, for real data applications, the optimal bandwidth needs to be estimated. This problem received a lot of attention in the recent statistical literature, especially within the context of density and regression estimation. Two well known methods are cross-validation and plug-in methods. Below we describe these methods, and further we demonstrate the methods in action on two sparse multinomial data sets.

Least squares cross validation

A widely used bandwidth selection procedure is the least squared cross validation method (LSCV) proposed by Rudemo (1982) and Bowman (1984). The idea of LSCV is quite general and can be used for a variety of nonparametric estimation problems. We will present the main ideas in our context of smoothing sparse multinomial data. A decomposition of MSSE is given by

$$\text{MSSE} = E\left(\sum_{i=1}^{k}\widehat{P}_i^2 - 2\sum_{i=1}^{k}p_i\widehat{P}_i\right) + \sum_{i=1}^{k}p_i^2$$

where the last term does not depend on h. So minimization of MSSE w.r.t. h is equivalent to minimization of

$$\text{MSSE} - \sum_{i=1}^{k}p_i^2 = E\left(\sum_{i=1}^{k}\widehat{P}_i^2 - 2\sum_{i=1}^{k}p_i\widehat{P}_i\right).$$

The right-hand side is unknown since it depends on $\boldsymbol{p}$. An unbiased estimator of this quantity is

$$\text{CV}(h) = \sum_{i=1}^{k}\widehat{P}_i^2 - \frac{2}{n}\sum_{\ell=1}^{n}\widehat{P}_{X_\ell}^{(-\ell)} \tag{3.29}$$

where $\widehat{P}_i^{(-\ell)}$ is the estimator for the i-th cell probability p_i, based on the data with the ℓ-th observation being deleted, and X_ℓ denote the random variables with $X_\ell = i$ if the ℓ-th observation falls into cell i. First note that the random variable X_ℓ is directly related to the triangular array $\mathbf{Y}_{n\ell}$ defined in (2.3), such that, for a fixed n, $X_1, \ldots, X_n$ are independent. In terms of these variables we can write $\overline{P}_i = n^{-1}\sum_{\ell=1}^{n}\mathbb{1}\{X_\ell = i\}$. Further, to see that the estimator is unbiased, it suffices to show $E(\widehat{P}_{X_\ell}^{(-\ell)}) = \sum_{i=1}^{k}p_iE(\widehat{P}_i)$. First, from

$$\widehat{P}_i = \frac{1}{nkh}\sum_{m=1}^{n}\sum_{j=1}^{k}L_{i,p}\left(\frac{x_j - x_i}{h}\right)\mathbb{1}\{X_m = j\}$$

we obtain

$$\widehat{P}_i^{(-\ell)} = \frac{1}{(n-1)kh} \sum_{\substack{m=1 \\ m\neq\ell}}^{n} \sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right) \mathbb{1}\{X_m = j\}$$

and

$$\widehat{P}_{X_\ell}^{(-\ell)} = \frac{1}{(n-1)kh} \sum_{\substack{m=1 \\ m\neq\ell}}^{n} \sum_{i=1}^{k} \sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right) \mathbb{1}\{X_\ell = i\}\mathbb{1}\{X_m = j\}.$$

Since X_ℓ and X_m, $m \neq \ell$, are independent we obtain $E(\widehat{P}_{X_\ell}^{(-\ell)}) = \sum_{i=1}^{k} p_i E(\widehat{P}_i)$.

The cross-validation bandwidth selector is defined as the minimizer of the cross-validation function (3.29), i.e., $\hat{h}_{\text{LSCV}} = \operatorname{argmin}_{h>0} \text{CV}(h)$. For local polynomial smoothers (3.29) can be rewritten as

$$\text{CV}(h) = \sum_{i=1}^{k} \widehat{P}_i^2 - 2\frac{n}{n-1}\sum_{i=1}^{k} \overline{P}_i \widehat{P}_i + \frac{2}{(n-1)kh}\sum_{i=1}^{k} L_{i,p}(0)\overline{P}_i,$$

which is computationally faster than its original definition (3.29).

In the density estimation context it was shown that LSCV gives asymptotically the proper amount of smoothing in the sense that, under appropriate regularity conditions,

$$\lim_{n\to\infty} \frac{\text{MISE}(\hat{h}_{\text{LSCV}})}{\text{MISE}(h_0)} = 1$$

(Stone (1984)), where $\hat{h}_{\text{LSCV}}$ is the minimizer of the appropriate cross-validation function, $h_0 = \operatorname{argmin}_{h>0} \text{MISE}$, and MISE is an abbreviation for mean integrated squared error. Also in the context of sparse tables Hall and Titterington (1987) have shown that least squares cross-validation works (see their Theorem 3.2).

Studies have shown that both the theoretical and practical performance of LSCV are somewhat disappointing. Hall and Marron (1987a) show that the bandwidth selector $\hat{h}_{\text{LSCV}}$ has the slow rate

$$\frac{\hat{h}_{\text{LSCV}}}{h_0} = 1 + O_P(n^{-1/10}),$$

which translates into high variability of $\hat{h}_{\text{LSCV}}$. This has been noted in Monte Carlo simulation studies (see e.g., the survey papers Park and Marron (1990) and Jones, Marron and Sheather (1996)).

Plug-in bandwidth selectors

A popular family of selection methods are the so-called "plug-in" type bandwidth selectors. The method is based on the expression (3.28) for asymptotically optimal bandwidth, more precisely, the idea is to substitute the unknown quantity $R(f^{(p+1)})$ by a "pilot" estimate.

A nonparametrically natural way to define an estimator for this quantity is to consider kernel-type estimators based on a bandwidth g, the so-called pilot bandwidth. Hall and Marron (1987b) and Jones and Sheather (1991) propose an estimator for $R(f^{(p+1)})$ when $f(\cdot)$ has unbounded support and decreases to zero sufficiently fast at infinity.

Ruppert, Sheather and Wand (1995) propose in the regression context the use of local polynomial smoothers to define estimators for functionals of the form $\int m^{(r)}(x)m^{(s)}(x)\,dx$. The case $r = s = 2$, with the estimator defined through local cubic based smoothers for the second derivative, is studied in detail. Cheng (1996, 1997) studies similar estimators in the context of binned density estimation, which is closely related to the setting of sparse multinomials. In Remark 3.4 in Aerts et al. (1997b) the similarities and differences between both settings are discussed in some detail. We will restrict attention to the optimal bandwidth for the local linear smoother, i.e., take $p = 1$ in (3.28), and we explain the method in the density estimation context. Denote $\theta_{rs} = \int f^{(r)}(x)f^{(s)}(x)\,dx$.

Direct plug-in

The unknown quantity in the optimal bandwidth expression is θ_{22}. An estimator for this quantity is defined through local cubic smoothers for the second derivative of $f(\cdot)$, based on a bandwidth g. Replacement of θ_{22} by $\hat{\theta}_{22}(g)$ leads to the direct plug-in rule

$$\hat{h}_{\mathrm{DPI}} = \left\{ n^{-1}C_1(K)\left(\hat{\theta}_{22}(g)\right)^{-1} \right\}^{1/5}. \tag{3.30}$$

This rule is not fully automatic yet, since it depends on the choice of the pilot bandwidth g. A way of choosing g is to appeal to the asymptotic MSE optimal bandwidth for estimation of $\hat{\theta}_{22}(g)$. In Cheng (1997) it is shown that the bandwidth g that minimizes AMSE is given by

$$g_{\mathrm{opt}} = \left(\frac{24\chi R(K_2^*)}{n\theta_{24}\mu_4(K_2^*)} \right)^{1/7} \tag{3.31}$$

where

$$\chi = \begin{cases} -1 & \text{if } \theta_{24} < 0 \\ \frac{5}{2} & \text{if } \theta_{24} > 0 \end{cases}$$

and

$$K_2^*(u) = \frac{|\boldsymbol{M}_{(2,3)}(u)|}{|\boldsymbol{N}_3|} K(u)$$

where $\boldsymbol{M}_{(2,3)}(u)$ is the same as $\boldsymbol{N}_3$, except that the third column is replaced by $(1, u, u^2, u^3)^T$.

The optimal pilot bandwidth (3.31) depends on θ_{24}, again some unknown quantity involving derivatives of $f(\cdot)$. One could once more propose a kernel-based estimator for this quantity, but this would lead to further bandwidth selection problems.

The usual strategy to overcome this problem is to estimate the unknown quantity θ_{24} by some "quick and simple" estimate. A widely used approach in the density estimation context is to estimate it via the so-called normal reference rule, i.e., pretend as if $f(\cdot)$ is a normal density with standard deviation σ, $f(\cdot) = \phi_\sigma(\cdot)$. Next, estimate the scale parameter from the data. Based on this estimate use $\theta_{24}(\phi_{\hat{\sigma}})$ as an estimator for θ_{24}.

To summarize, the algorithm looks like :

Step 1 Estimate θ_{24} using the normal reference rule, i.e., use the estimator $\theta_{24}(\phi_{\hat{\sigma}})$, where $\hat{\sigma}$ is a scale estimator.

Step 2 Estimate θ_{22} using the local cubic based second derivative estimator $\hat{\theta}_{22}(\hat{g})$ where

$$\hat{g} = \left(\frac{24\hat{\chi}R(K_2^*)}{n\theta_{24}(\phi_{\hat{\sigma}})\mu_4(K_2^*)} \right)^{1/7}$$

and $\hat{\chi}$ is determined by the sign of $\theta_{24}(\phi_{\hat{\sigma}})$.

Step 3 The selected bandwidth is

$$\hat{h}_{\text{DPI}} = \left\{ n^{-1} C_1(K) \left(\hat{\theta}_{22}(\hat{g}) \right)^{-1} \right\}^{1/5} .$$

Solve-the-equation

Solve-the-equation methods (STE) (Scott, Tapia and Thompson (1977), Sheather

(1986), Sheather and Jones (1991)) have a close connection to the direct plug-in approach. Also motivated by the formula for the asymptotically optimal bandwidth, solve-the-equation rules link expressions (3.28) and (3.31) to obtain the relationship

$$g_{\text{opt}}(h) = \left(\frac{24\chi R(K_2^*)\mu_2^2(K)R(f^{(2)})}{R(K)\mu_4(K_2^*)\theta_{24}} \right)^{1/7} h_{\text{opt}}^{5/7}. \tag{3.32}$$

The solve-the-equation algorithm to select the local linear optimal bandwidth is : (Cheng (1996))

Step 1 Estimate $R(f^{(2)})$ and θ_{24} by some reference rule, such as the normal reference rule, i.e., use the estimators $R(\phi_{\hat{\sigma}}^{(2)})$ and $\theta_{24}(\phi_{\hat{\sigma}})$, where $\hat{\sigma}$ is a scale estimator.

Step 2 Estimate θ_{22} using the local cubic based second derivative estimator $\hat{\theta}_{22}(g(h))$ where

$$g(h) = \left(\frac{24\hat{\chi} R(K_2^*)\mu_2^2(K)R(\phi_{\hat{\sigma}}^{(2)})}{R(K)\mu_4(K_2^*)\theta_{24}(\phi_{\hat{\sigma}})} \right)^{1/7} h^{5/7}$$

and $\hat{\chi}$ is determined by the sign of $\theta_{24}(\phi_{\hat{\sigma}})$.

Step 3 The selected bandwidth $\hat{h}_{\text{STE}}$ is the solution to the equation

$$h = \left\{ n^{-1}C_1(K) \left(\hat{\theta}_{22}(g(h)) \right)^{-1} \right\}^{1/5}.$$

Step 3 of the algorithm requires a root-finding numerical algorithm to implement the bandwidth selector. This is usually done by searching the root based on a grid of h-values. The direct plug-in approach does not require such a grid search.

The theoretical performance of both the direct plug-in and the solve-the-equation bandwidth selection methods is, under some regularity conditions,

$$\frac{\hat{h}}{h_{\text{opt}}} - 1 = O_P(n^{-\alpha}),$$

where

$$\alpha = \begin{cases} 5/14 & \text{if } \theta_{24} < 0 \\ 4/14 & \text{if } \theta_{24} > 0 \end{cases}$$

(see Ruppert, Sheather and Wand (1995) and Cheng (1997) for more details). Note that the obtained rate is a big improvement on that of LSCV.

We illustrate local constant and local linear smoothers on two real data sets. Figure 3.5 is for the mine explosions data, presented in Table 1.3, while Figure 3.6 is for the salary data, given in Table 1.2. In both figures LSCV is used to select the bandwidth for the local constant and local linear smoother. Also STE is used for the local linear smoother. The DPI bandwidth selector, not presented here, gave virtually the same answer as STE. In Figure 3.5 the better boundary performance of the local linear smoother compared to the classical kernel estimator is clearly perceptible. In Figure 3.6 there seems to be no need to use a method that corrects at the boundaries, especially since for this data the local linear smoother gives some negative estimated cell probabilities.

At this moment it is hard to comment on the different bandwidth selectors for the local linear estimator, since, for a real data set, we don't know what the truth is. For a simulation study we refer to Chapter 6, where we will compare LSCV, DPI to a newly proposed bandwidth selection method.

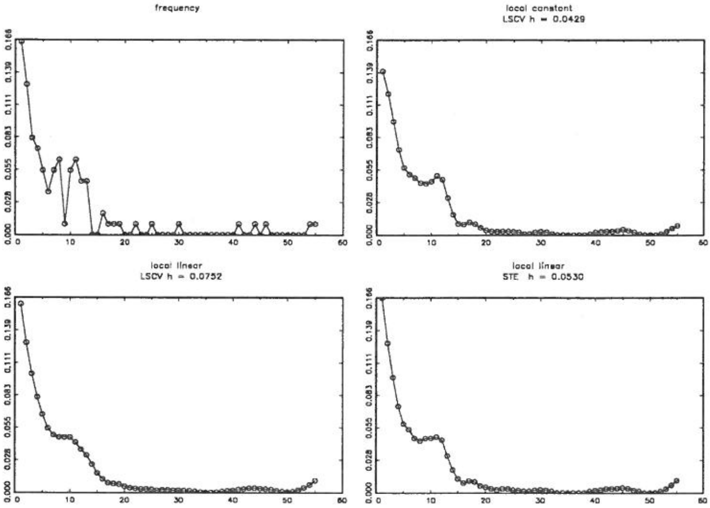

Figure 3.5: *Mine explosions data. Top left: frequency estimates. Top right: local constant estimates with bandwidth chosen by LSCV. Bottom left: local linear estimates with bandwidth chosen by LSCV. Bottom right: local linear estimates with bandwidth chosen by STE.*

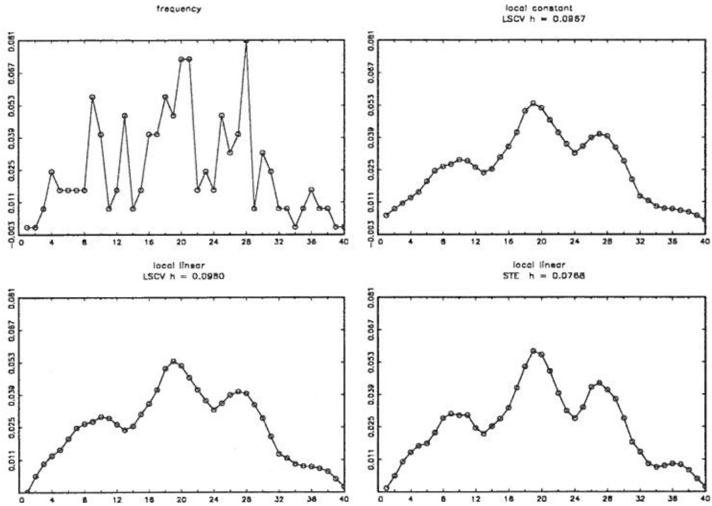

Figure 3.6: *Salary data. Top left: frequency estimates. Top right: local constant estimates with bandwidth chosen by LSCV. Bottom left: local linear estimates with bandwidth chosen by LSCV. Bottom right: local linear estimates with bandwidth chosen by STE.*

Chapter 4

Central limit theorem for SSE

In Chapter 3 we have studied the rate of convergence of the MSSE for the local polynomial smoother for the cell probabilities in sparse multinomial data (one-dimensional case). In this chapter we obtain a central limit result for the statistic $\mathrm{SSE}(\boldsymbol{P}^*) = \sum_{i=1}^{k} (P_i^* - p_i)^2$, where we consider $\boldsymbol{P}^* = \overline{\boldsymbol{P}}$, the frequency estimators, and $\boldsymbol{P}^* = \widehat{\boldsymbol{P}}$, the local polynomial smoothers.

The result for the frequency estimators is obtained in Section 4.1 as a special case of a result obtained by Burman (1987b), who studies central limit theorems of various statistics in sparse tables. These statistics also include $\mathrm{SSE}(\widehat{\boldsymbol{P}}_0)$, where $\widehat{\boldsymbol{P}}_0$ is the local constant smoother. To prove a central limit result on this local constant smoother, he needs stringent boundary conditions on $f(\cdot)$. Moreover, for the bandwidth he requires the condition $h = o(n^{-2/9})$, while the optimal bandwidth is $h = Cn^{-1/5}$ for local constant smoothers when the boundary conditions are satisfied (see Chapter 3).

In Section 4.2, we study the asymptotic distribution of the sum of squared errors of the local polynomial smoothers. We restrict attention to local polynomial smoothers with p odd, since it is clear from Chapter 3 that in this case we will need no boundary conditions. In addition, our result includes the optimal bandwidth case. To proof our central limit theorem for the local polynomial smoothers we use a result obtained by Hall (1984), who studies the asymptotic distribution of the integrated squared error (ISE) of kernel-type density estimators.

Essentially, the proofs by Hall (1984) and Burman (1987b) are based on the same technique. More precisely, the statistic can be written as a quadratic form, which on its turn can be written as a martingale. Next, a martingale central limit theorem (McLeish (1974)) is applied. Hall pays special attention to include the case of the optimal bandwidth.

As noted in Hall (1984) (and also in Burman (1987b)) the martingale central limit technique is also suitable for the multivariate case. In Chapter 5 we will give the central limit result for the local linear smoother for estimating the cell probabilities in a multi-dimensional sparse table.

4.1 CLT for frequency estimators

Theorem 4.1
Assume $n \to \infty$, $k \to \infty$, $2k \sum_{i=1}^{k} p_i^2 + k/n \to \sigma_0^2$, *as* $n \to \infty$ *and* $p_i \leq C/k$, $i = 1, \dots, k$. *Then we have*

$$n\sqrt{k}\left(SSE(\overline{\boldsymbol{P}}) - MSSE(\overline{\boldsymbol{P}})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_0^2).$$

Proof
This normality result is a special case of Theorem 3.2 in Burman (1987b). Using his notation, with

$$\begin{aligned} W_{n,i}(x) &= \begin{cases} n^{-1} & x = i \\ 0 & x \neq i \end{cases} \\ a_n(x) &\equiv n\sqrt{k} \\ \mu_n(x) &= x, \end{aligned}$$

we have the following relation between his T_{2n} and SSE($\overline{\boldsymbol{P}}$):

$$T_{2n} = n\sqrt{k}\mathrm{SSE}(\overline{\boldsymbol{P}}).$$

It therefore suffices to show his conditions (C1)-(C8) and his condition (2.1) with $\sigma^2 = \sigma_0^2$ and

$$C_{j_1 j_2} = \begin{cases} \dfrac{\sqrt{k}}{n} & j_1 = j_2 \\ 0 & j_1 \neq j_2 \end{cases}$$

to obtain

$$T_{2n} - E(T_{2n}) = n\sqrt{k}\left(\mathrm{SSE}(\overline{\boldsymbol{P}}) - \mathrm{MSSE}(\overline{\boldsymbol{P}})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_0^2).$$

His condition (2.1) becomes

$$2k \sum_{i=1}^{k} p_i^2 + \frac{k}{n} \to \sigma_0^2, \qquad 0 < \sigma_0^2 < \infty$$

which is valid by assumption. From the condition $p_i \leq C/k,\ i = 1, \dots, k$ and the definition of $C_{j_1 j_2}$ his conditions (C1)–(C8) are easily shown to be satisfied. ■

Remark 4.1
First note that we restrict to the case $p_i \leq C/k$ for simplicity. A central limit theorem could be obtained in the more general situation, with conditions on $M_n = \sup_{1 \leq i \leq k} p_i$. We do not investigate this situation, since our main interest is the comparison of Theorem 4.1 to Theorem 4.2, the central limit theorem for the local polynomial smoothers, for which, by boundedness of the latent density $f(\cdot)$, the condition $p_i \leq C/k$ is satisfied.

As noted in the introduction, the results obtained by Hall (1984) and Burman (1987b) are essentially based on the same technique (i.e., a martingale central limit theorem). For the proof of Theorem 4.1 we rely on the result by Burman (1987b) for two reasons. The first one is that checking the conditions (C1)–(C8) of Burman (1987b) is easier than checking the conditions in Hall (1984). The second reason is that, when applying Hall's result, we need a stronger assumption on k, i.e., $k/n \to 0$. The assumption $2k \sum_{i=1}^{k} p_i^2 + k/n \to \sigma_0^2$, in Theorem 4.1 only requires $k/n \to c$, with c a constant. This difference in assumptions comes from the fact that, although Hall (1984) and Burman (1987b) rely on the same martingale central limit theorem, the sufficient conditions in their results differ.

4.2 CLT for local polynomial estimators

The next theorem is the main result of this chapter. Before we prove this result, we first compare it to Theorem 4.1. The complete proof of Theorem 4.2 is rather technical. Therefore, we first give an outline, and present the key steps in different lemmas.

Theorem 4.2
Assume (C.1)–(C.3), p odd and $nh \to \infty$, then

$$d(n)\left(SSE(\widehat{\boldsymbol{P}}) - MSSE(\widehat{\boldsymbol{P}})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where

$$d(n) = \begin{cases} \dfrac{\sqrt{n}k}{h^{p+1}} & nh^{2p+3} \to \infty \\ n^{\frac{4p+5}{2(2p+3)}} k & nh^{2p+3} \to \lambda \\ nk\sqrt{h} & nh^{2p+3} \to 0 \end{cases}$$

and

$$\sigma^2 \equiv \sigma^2(f, L_{(p)}) = \begin{cases} 4\sigma_1^2 & nh^{2p+3} \to \infty \\ 4\sigma_1^2 \lambda^{\frac{2(p+1)}{2p+3}} + 2\sigma_2^2 \lambda^{\frac{-1}{2p+3}} & nh^{2p+3} \to \lambda \\ 2\sigma_2^2 & nh^{2p+3} \to 0 \end{cases}$$

with σ_1^2 and σ_2^2 given in Lemma 4.1 and Lemma 4.2 respectively.

Remark 4.2

In terms of rates of convergence we can present Theorems 4.1 and 4.2 as follows :

$$\mathrm{SSE}(\overline{\boldsymbol{P}}) - \mathrm{MSSE}(\overline{\boldsymbol{P}}) = O_P\left(\frac{1}{n\sqrt{k}}\right)$$
$$\mathrm{SSE}(\widehat{\boldsymbol{P}}) - \mathrm{MSSE}(\widehat{\boldsymbol{P}}) = O_P\left(\frac{1}{d(n)}\right),$$

with $d(n)$ as in Theorem 4.2.

We now study when the local polynomial smoothers have a faster rate of convergence than the frequency estimators, i.e., we investigate under what degree of sparseness

$$\frac{n\sqrt{k}}{d(n)} \to 0. \tag{4.1}$$

Consider tables where $k \propto n^q$ with $q > 0$. The three different levels of smoothing become :

Case 1 : $nh^{2p+3} \to \infty$, i.e., the case of oversmoothing. In this situation (4.1) is equivalent to $n^{(1-q)/(2p+2)}h \to 0$. Therefore $q > 1/(2p+3)$ is necessary, since also $nh^{2p+3} \to \infty$ needs to be satisfied. Furthermore, $n^{(1-q)/(2p+2)}h \to 0$ means that the bandwidth h is not allowed to oversmooth too much.

Case 2 : $nh^{2p+3} \to \lambda$, i.e., the case of optimal smoothing. Condition (4.1) is equivalent to $n^{1/(2p+3)-q} \to 0$, which is in this situation equivalent to $q > 1/(2p+3)$.

Case 3 : $nh^{2p+3} \to 0$, i.e., the case of undersmoothing. Now, (4.1) reduces to $kh \to \infty$, and, since also $nh^{2p+3} \to 0$ needs to be satisfied, $q > 1/(2p+3)$ is necessary. Similar to the case of oversmoothing, the bandwidth is not allowed to undersmooth too much (since $n^q h \to \infty$).

So, we can conclude once more that, for sparse tables with $kn^{-1/(2p+3)} \to \infty$, the local polynomial smoothers have a better performance than the frequency estimators.

Remark 4.3

Theorem 4.2 can be used in testing situations. Under a given hypothesis on the cell probabilities (in terms of the latent density, e.g., $f(\cdot)$ is uniform), the asymptotic distribution of $\mathrm{SSE}(\widehat{\boldsymbol{P}}) - \mathrm{MSSE}(\widehat{\boldsymbol{P}})$ is completely known, and can be used to compute p-values. A practical investigation of this test is not yet studied. Further, it would also be desirable to study a relative statistic such as $\sum_{i=1}^{k}(\widehat{P}_i - p_i)^2/p_i$ as an alternative to the Pearson statistic. Note that for the special case of uniform cell probabilities, i.e., $p_i = k^{-1}$, this statistic reduces to $k\mathrm{SSE}(\widehat{\boldsymbol{P}})$, for which Theorem 4.2 can be applied.

Outline of the Proof

We use the following decomposition :

$$\mathrm{SSE}(\widehat{\boldsymbol{P}}) = \sum_{i=1}^{k}(\widehat{P}_i - E\widehat{P}_i)^2 + \sum_{i=1}^{k}(E\widehat{P}_i - p_i)^2 + 2\sum_{i=1}^{k}(E\widehat{P}_i - p_i)(\widehat{P}_i - E\widehat{P}_i). \qquad (4.2)$$

In Chapter 2 we have seen that we can write $\widehat{P}_i - E\widehat{P}_i$ as a sum of independent identically distributed random variables (see (2.7)). To summarize, this was done through the triangular arrays $\boldsymbol{Y}_{n\ell} = (Y_{\ell 1}, \ldots, Y_{\ell k})^T$ and $\boldsymbol{X}_{n\ell} = (X_{\ell 1}, \ldots, X_{\ell k})^T$, $\ell = 1, \ldots, n$ where, for $i = 1, \ldots, k$,

$$Y_{\ell i} = \begin{cases} 1 & \text{if the } \ell\text{-th observation is in cell } i \\ 0 & \text{otherwise,} \end{cases}$$

and

$$X_{\ell i} = \frac{1}{kh}\sum_{j=1}^{k} L_{i,p}\left(\frac{x_j - x_i}{h}\right)(Y_{\ell j} - p_j).$$

Further, for a fixed n, $\boldsymbol{Y}_{n1}, \ldots, \boldsymbol{Y}_{nn}$ are i.i.d., and hence also $\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{nn}$, and

$$\widehat{P}_i - E\widehat{P}_i = \frac{1}{n}\sum_{\ell=1}^{n} X_{\ell i}.$$

This notation for $\widehat{P}_i - E\widehat{P}_i$ enables us to decompose the first term of (4.2) into

$$\sum_{i=1}^{k}(\widehat{P}_i - E\widehat{P}_i)^2 = \frac{1}{n^2}\sum_{\ell=1}^{n}\sum_{i=1}^{k} X_{\ell i}^2 + \frac{2}{n^2}\sum\sum_{1\le \ell_1 < \ell_2 \le n}\sum_{i=1}^{k} X_{\ell_1 i}X_{\ell_2 i}. \qquad (4.3)$$

The second term on the r.h.s. has a U-statistic stucture with symmetric kernel

$$H_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}) = \sum_{i=1}^{k} X_{1i}X_{2i}. \qquad (4.4)$$

Since $\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{nn}$ are i.i.d. with mean zero, $E\left(H_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2})|\boldsymbol{X}_{n1}\right) = 0$, so that this "U-statistic" is degenerate.

Combining (4.2) and (4.3) gives the following decomposition

$$\mathrm{SSE}(\widehat{\boldsymbol{P}}) - \mathrm{MSSE}(\widehat{\boldsymbol{P}}) = 2I_n + \frac{n-1}{n} U_n + R_n - E(R_n) \tag{4.5}$$

where

$$\begin{aligned} I_n &= \sum_{i=1}^{k} (E\widehat{P}_i - p_i)(\widehat{P}_i - E\widehat{P}_i) \\ U_n &= \binom{n}{2}^{-1} \sum\sum_{1 \le \ell_1 < \ell_2 \le n} H_n(\boldsymbol{X}_{n\ell_1}, \boldsymbol{X}_{n\ell_2}) \\ R_n &= \frac{1}{n^2} \sum_{\ell=1}^{n} H_n(\boldsymbol{X}_{n\ell}, \boldsymbol{X}_{n\ell}). \end{aligned} \tag{4.6}$$

Each term is studied separately in Lemmas 4.1-4.3 (see further), and from these results we have

$$\begin{aligned} I_n &= O_P\left(\frac{h^{p+1}}{\sqrt{n}k}\right) \\ U_n &= O_P\left(\frac{1}{nk\sqrt{h}}\right) \\ R_n - E(R_n) &= O_P\left(\frac{1}{\sqrt{n^3k^2h^2}}\right) = o_P\left(\frac{1}{nk\sqrt{h}}\right). \end{aligned}$$

Note that the condition $nh \to \infty$ is used to obtain the last order bound.
Hence $R_n - E(R_n)$ is always of smaller order than U_n.
If $nh^{2p+3} \to \infty$, I_n dominates U_n and the result follows from Lemma 4.1.
If $nh^{2p+3} \to 0$, U_n dominates I_n and the result follows from Lemma 4.2.
If $nh^{2p+3} \to \lambda$, the situation is more difficult since now I_n and U_n have the same order of magnitude. The result follows from Lemma 4.4. ■

We will first introduce all the lemmas referred to in the outline of the proof. Next, we prove these results, but the more technical calculations are given in Section 4.3.

Lemma 4.1
Assume (C.1)–(C.3) and p odd. We have

$$\frac{\sqrt{n}k}{h^{p+1}} I_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_1^2)$$

where

$$\sigma_1^2 \equiv \sigma_1^2(f, L_{(p)}) = \frac{\mu_{p+1}^2(L_{(p)})}{((p+1)!)^2} \left\{ \int_0^1 (f^{(p+1)}(x))^2 f(x)\, dx - \left(\int_0^1 f^{(p+1)}(x) f(x)\, dx \right)^2 \right\}.$$

Lemma 4.2
Assume (C.1)–(C.3), p odd and $nh \to \infty$. *We have*

$$nk\sqrt{h}U_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\sigma_2^2)$$

where $\sigma_2^2 \equiv \sigma_2^2(f, L_{(p)}) = \int\limits_0^1 f^2(x)\, dx \int\limits_{-2L}^{2L} \left[\int\limits_{-L}^{L} L_{(p)}(u) L_{(p)}(u+v)\, du \right]^2 dv.$

Lemma 4.3
Assume (C.1)–(C.3) and p odd. We have

$$R_n - E(R_n) = O_p\left(\frac{1}{\sqrt{n^3 k^2 h^2}} \right).$$

Lemma 4.4
Assume (C.1)–(C.3), p odd and $nh \to \infty$. *If* $\mathrm{Var}(I_n) \propto \mathrm{Var}(U_n)$, *i.e.* $nh^{2p+3} \to \lambda$, *then* $aI_n + bU_n$ *is asymptotically normally distributed with mean zero and variance* $a^2\mathrm{Var}(I_n) + b^2\mathrm{Var}(U_n)$.
For $a = 2$ *and* $b = 1$ *this result reduces to*

$$n^{\frac{4p+5}{2(2p+3)}} k\,(2I_n + U_n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 4\sigma_1^2 \lambda^{\frac{2(p+1)}{2p+3}} + 2\sigma_2^2 \lambda^{\frac{-1}{2p+3}}).$$

To simplify the notation in the proofs we define the following shorthand notation

$$\begin{aligned} b_i &= E\widehat{P}_i - p_i \\ S_{ij} &= \frac{1}{kh} L_{i,p}\left(\frac{x_j - x_i}{h} \right) \\ D_{j_1 j_2} &= \sum_{i=1}^{k} S_{ij_1} S_{ij_2} \\ \varepsilon_{\ell i} &= Y_{\ell i} - p_i. \end{aligned} \qquad (4.7)$$

Proof of Lemma 4.1

With the shorthand notation we can write

$$I_n = \frac{1}{n}\sum_{\ell=1}^{n}\sum_{i=1}^{k} b_i X_{\ell i} \overset{def.}{=} \frac{1}{n}\sum_{\ell=1}^{n} T_{n\ell} \tag{4.8}$$

$$X_{\ell i} = \sum_{j=1}^{k} S_{ij}\varepsilon_{\ell j}. \tag{4.9}$$

For each n, the random variables $T_{n\ell}, \ell = 1, \dots, n$, are i.i.d. with mean zero and therefore the Lyapounov condition

$$\{\mathrm{Var}(I_n)\}^{-\frac{(2+\delta)}{2}} \sum_{\ell=1}^{n} E\left|\frac{1}{n}T_{n\ell}\right|^{2+\delta} \to 0,$$

for some $\delta > 0$, is sufficient to guarantee

$$\{\mathrm{Var}(I_n)\}^{-\frac{1}{2}}\, I_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

We will check this Lyapounov condition for $\delta = 2$.
Since $\mathrm{Var}(I_n) = n^{-1}E(T_{n1}^2)$ we first need to find an expression for $E(T_{n1}^2)$ and an order bound for $E(T_{n1}^4)$. In Section 4.3 we show

$$E(T_{n1}^2) = \frac{h^{2(p+1)}}{k^2}\sigma_1^2 + o\left(\frac{h^{2(p+1)}}{k^2}\right) \tag{4.10}$$

and

$$E(T_{n1}^4) = O\left(\frac{h^{4(p+1)}}{k^4}\right). \tag{4.11}$$

These relations immediately imply that the Lyapounov condition with $\delta = 2$ is satisfied. The leading term in the asymptotic expression for $n^{-1}E(T_{n1}^2)$ determines the asymptotic variance of I_n. ■

To prove Lemma 4.2 we will rely on Lemma 4.5 which is closely related to a result obtained by Hall (1984).

Lemma 4.5
With $U_n = \binom{n}{2}^{-1} \sum\sum_{1\le\ell_1<\ell_2\le n} H_n(\boldsymbol{X}_{n\ell_1}, \boldsymbol{X}_{n\ell_2})$ *the degenerate "U-statistic"defined in (4.6), define*

$$G_n(\boldsymbol{x},\boldsymbol{y}) = E(H_n(\boldsymbol{X}_{n1},\boldsymbol{x})H_n(\boldsymbol{X}_{n1},\boldsymbol{y})). \tag{4.12}$$

If

$$\left(E(G_n^2(\boldsymbol{X}_{n1},\boldsymbol{X}_{n2})) + n^{-1}E(H_n^4(\boldsymbol{X}_{n1},\boldsymbol{X}_{n2}))\right) / \left(E(H_n^2(\boldsymbol{X}_{n1},\boldsymbol{X}_{n2}))\right)^2 \to 0, \tag{4.13}$$

then U_n *is asymptotically normally distributed with mean zero and variance given by* $2n^{-2}E(H_n^2(\boldsymbol{X}_{n1},\boldsymbol{X}_{n2}))$.

The difference between Lemma 4.5 and the result by Hall (1984, Theorem 1) is that we have triangular arrays $\boldsymbol{X}_{n1},\ldots,\boldsymbol{X}_{nn}$, which are i.i.d. for a fixed n, while Hall considers i.i.d. random variables $X_1,\ldots,X_n$. The proof of his result can be used without any modification. The reason is based on the fact that the key tool in the proof of Hall's Theorem 1 is the martingale difference array structure. Although we consider a somewhat more complicated situation, this martingale difference array structure is still true, so that the proof by Hall remains valid. For more details, we refer to the proof of Lemma 4.4, which is based on the same principle.

Proof of Lemma 4.2
In Section 4.3 we show

$$\begin{aligned} E(H_n^2(\boldsymbol{X}_{n1},\boldsymbol{X}_{n2})) &= \frac{1}{k^2h}\sigma_2^2 + o\left(\frac{1}{k^2h}\right) && (4.14)\\ E(H_n^4(\boldsymbol{X}_{n1},\boldsymbol{X}_{n2})) &= O\left(\frac{1}{k^4h^3}\right) && (4.15)\\ E(G_n^2(\boldsymbol{X}_{n1},\boldsymbol{X}_{n2})) &= O\left(\frac{1}{k^4h}\right). && (4.16)\end{aligned}$$

Since $h \to 0$ and $nh \to \infty$ as $n \to \infty$ these properties imply (4.13) in Lemma 4.5. So, U_n is asymptotically normal with mean zero and variance $2\sigma_2^2/(n^2k^2h)$. ■

Proof of Lemma 4.3
By (4.4) and (4.6) it is obvious that

$$E(R_n) = \frac{1}{n}E(H_n(\boldsymbol{X}_{n1},\boldsymbol{X}_{n1}))$$
$$\mathrm{Var}(R_n) = \frac{1}{n^3}\mathrm{Var}(H_n(\boldsymbol{X}_{n1},\boldsymbol{X}_{n1})).$$

From (4.3) and from (3.14) in Theorem 3.2 it is clear that

$$E(R_n) \;=\; E\sum_{j=1}^{k}(\widehat{P}_j - E\widehat{P}_j)^2 = \sum_{j=1}^{k}\mathrm{Var}(\widehat{P}_j) = \frac{R(L_{(p)})}{nkh} + o\left(\frac{1}{nkh}\right).$$

So $E(H_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n1})) = O((kh)^{-1})$ and we will show

$$E(H_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n1})) = O((kh)^{-2}) \tag{4.17}$$

in Section 4.3. We then have

$$\begin{aligned} R_n - E(R_n) \;&=\; O_P\left(\sqrt{\mathrm{Var}(R_n)}\right) \\ &=\; O_P\left(\frac{1}{\sqrt{n^3k^2h^2}}\right). \end{aligned}$$

■

Proof of Lemma 4.4

The proof is based on that of Theorem 1 in Hall (1984). From the technical details below (martingale difference array structure) it becomes clear why the proof of Theorem 1 in Hall can be used without any modification to show the validity of Lemma 4.5. Write

$$aI_n = \frac{a}{n}\sum_{\ell=1}^{n} T_{n\ell} \overset{def.}{=} \sum_{\ell=1}^{n}\widetilde{T}_{n\ell}$$

and

$$bU_n = b\binom{n}{2}^{-1}\sum\sum_{1\le\ell_1<\ell_2\le n} H_n(\boldsymbol{X}_{n\ell_1}, \boldsymbol{X}_{n\ell_2}) \overset{def.}{=} \sum\sum_{1\le\ell_1<\ell_2\le n}\widetilde{H}_n(\boldsymbol{X}_{n\ell_1}, \boldsymbol{X}_{n\ell_2}) = \sum_{\ell=1}^{n}\xi_{n\ell},$$

where $\xi_{n\ell} = \sum_{\ell_1=1}^{\ell-1}\widetilde{H}_n(\boldsymbol{X}_{n\ell_1}, \boldsymbol{X}_{n\ell})$, $\ell \ge 2$ and $\xi_{n1} = 0$ and define

$$\widetilde{G}_n(\boldsymbol{x}, \boldsymbol{y}) = E(\widetilde{H}_n(\boldsymbol{X}_{n1}, \boldsymbol{x})\widetilde{H}_n(\boldsymbol{X}_{n2}, \boldsymbol{y})).$$

The sequence $\left\{\widetilde{T}_{n\ell} + \xi_{n\ell} \overset{def.}{=} Z_{n\ell},\ \mathcal{F}_{n\ell} = \sigma(\boldsymbol{Y}_{n1}, \ldots, \boldsymbol{Y}_{n\ell}),\ \ell = 1, \ldots, n\right\}$ is a martingale difference array. In order to show asymptotic normality of $\sum_{\ell=1}^{n}(\widetilde{T}_{n\ell} + \xi_{n\ell}) =$

$aI_n + bU_n$, the following conditions are sufficient (McLeish (1974) or Theorem 1 in Pollard (1984, p. 171))

$$\forall \varepsilon > 0 \quad s_n^{-2} \sum_{\ell=1}^{n} E\left(Z_{n\ell}^2 I(|Z_{n\ell}| > \varepsilon s_n)\right) \to 0, \text{ as } n \to \infty \tag{4.18}$$

and

$$s_n^{-2} W_n^2 \xrightarrow{P} 1, \tag{4.19}$$

where $s_n^2 = E\left((aI_n + bU_n)^2\right)$ and $W_n^2 = \sum_{\ell=1}^n E(Z_{n\ell}^2 | \boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{n\ell-1})$.
First note that $\xi_{n1}, \ldots, \xi_{nn}$ are (pairwise) uncorrelated and that $E(\xi_{n\ell_1} \widetilde{T}_{n\ell_2}) = 0$, $\ell_1, \ell_2 = 1, \ldots, n$, since

$$E(\widetilde{H}_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}) \widetilde{H}_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n3})) = E(\widetilde{H}_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}) \widetilde{H}_n(\boldsymbol{X}_{n3}, \boldsymbol{X}_{n4})) = 0$$

and

$$E(\widetilde{H}_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}) \widetilde{T}_{n3}) = E(\widetilde{H}_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}) \widetilde{T}_{n2}) = 0.$$

Therefore, $\text{Var}(U_n) = \sum_{\ell=1}^n E(\xi_{n\ell}^2) = \binom{n}{2} E(\widetilde{H}_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}))$ and

$$E(W_n^2) = s_n^2 = a^2 \text{Var}(I_n) + b^2 \text{Var}(U_n) \overset{def.}{=} s_{n1}^2 + s_{n2}^2.$$

A sufficient condition for (4.18) is

$$s_n^{-4} \sum_{\ell=1}^{n} E(Z_{n\ell}^4) \to 0, \tag{4.20}$$

since, by Hölder's inequality,

$$\begin{aligned} &\sum_{\ell=1}^{n} E\left(Z_{n\ell}^2 I(|Z_{n\ell}| > \varepsilon s_n)\right) \\ &\qquad \le \sum_{\ell=1}^{n} \sqrt{E(Z_{n\ell}^4) I\!P\left(|Z_{n\ell}| > \varepsilon s_n\right)} \\ &\qquad \le \sum_{\ell=1}^{n} \frac{E(Z_{n\ell}^4)}{\varepsilon^2 s_n^2}. \end{aligned}$$

Similar to $E(\xi_{n\ell_1}\widetilde{T}_{n\ell_2}) = 0$, $\ell_1, \ell_2 = 1, \ldots, n$, also $E(\xi_{n\ell}\widetilde{T}^3_{n\ell}) = 0$, $\ell = 1, \ldots, n$, so that, by Hölder's inequality,

$$\begin{aligned} E(Z^4_{n\ell}) &= E\left((\xi_{n\ell} + \widetilde{T}_{n\ell})^4\right) \\ &\leq E(\xi^4_{n\ell}) + 4\sqrt[4]{(E(\xi^4_{n\ell}))^3\, E(\widetilde{T}^4_{n\ell})} + 6\sqrt{E(\xi^4_{n\ell})E(\widetilde{T}^4_{n\ell})} + E(\widetilde{T}^4_{n\ell}). \end{aligned}$$

An application of a discrete version of Hölder's inequality (see Chow and Teicher (1978, p. 107)) yields

$$\begin{aligned} s_n^{-4}\sum_{\ell=1}^n E(Z^4_{n\ell}) &\leq s_n^{-4}\sum_{\ell=1}^n E(\xi^4_{n\ell}) + 4\sqrt[4]{\left(s_n^{-4}\sum_{\ell=1}^n E(\xi^4_{n\ell})\right)^3}\sqrt[4]{s_n^{-4}\sum_{\ell=1}^n E(\widetilde{T}^4_{n\ell})} \\ &\quad + 6\sqrt{s_n^{-4}\sum_{\ell=1}^n E(\xi^4_{n\ell})}\sqrt{s_n^{-4}\sum_{\ell=1}^n E(\widetilde{T}^4_{n\ell})} + s_n^{-4}\sum_{\ell=1}^n E(\widetilde{T}^4_{n\ell}). \end{aligned}$$

From the proof of Hall's Theorem 1 it is clear that condition (4.13) implies

$$s_{n2}^{-4}\sum_{\ell=1}^n E(\xi^4_{n\ell}) \to 0.$$

and in Lemma 4.2 we have shown the vailidity of (4.13). Further, in the proof of Lemma 4.1 we have shown

$$s_{n1}^{-4}\sum_{\ell=1}^n E(\widetilde{T}^4_{n\ell}) \to 0.$$

Since $s_n^2 \propto s_{n1}^2 \propto s_{n2}^2$ these results imply (4.20).

Condition (4.19) is implied by

$$s_n^{-4}E((W_n^2 - s_n^2)^2) = s_n^{-4}\mathrm{Var}(W_n^2) \to 0, \tag{4.21}$$

which we will check. Note that

$$\begin{aligned} w_{n\ell} &\equiv E(Z^2_{n\ell}|\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{n\ell-1}) \\ &= E(\xi^2_{n\ell}|\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{n\ell-1}) + E(\widetilde{T}^2_{n\ell}|\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{n\ell-1}) + 2E(\widetilde{T}_{n\ell}\xi_{n\ell}|\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{n\ell-1}) \\ &= v_{n\ell} + n^{-1}s^2_{n1} + 2E(\widetilde{T}_{n\ell}\xi_{n\ell}|\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{n\ell-1}) \end{aligned}$$

where $v_{n\ell}$ is studied in the proof of Hall. With $\widetilde{Q}_n(\boldsymbol{x}) = E(\widetilde{H}_n(\boldsymbol{X}_{n1}, \boldsymbol{x})\widetilde{T}_{n1})$ we have

$$E(\widetilde{T}_{n\ell}\xi_{n\ell}|\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{n\ell-1}) = \sum_{\ell_2=1}^{\ell-1} \widetilde{Q}_n(\boldsymbol{X}_{n\ell_2}).$$

Since $\sum_{\ell=1}^{n} E(v_{n\ell}) = s_{n2}^2$ and $E(\widetilde{Q}_n(\boldsymbol{X}_{n1})) = 0$ we obtain

$$\begin{aligned}\sum_{\ell_1=1}^{n}\sum_{\ell_2=1}^{n} E(w_{n\ell_1}w_{n\ell_2}) &= \sum_{\ell_1=1}^{n}\sum_{\ell_2=1}^{n} E(v_{n\ell_1}v_{n\ell_2}) + 2s_{n1}^2 s_{n2}^2 + s_{n1}^4 \\ &\quad +4\sum_{\ell_1=1}^{n}\sum_{\ell_2=1}^{n}\sum_{\ell_3=1}^{\min(\ell_1,\ell_2)-1} E(\widetilde{Q}_n^2(\boldsymbol{X}_{n\ell_3})) + 4\sum_{\ell_1=1}^{n}\sum_{\ell_2=1}^{n}\sum_{\ell_3=1}^{\ell_2-1} E(v_{n\ell_1}\widetilde{Q}_n(\boldsymbol{X}_{n\ell_3})).\end{aligned}$$

Now

$$E(v_{n\ell_1}\widetilde{Q}_n(\boldsymbol{X}_{n\ell_3})) = \sum_{r_1=1}^{\ell_1-1}\sum_{r_2=1}^{\ell_1-1} E(\widetilde{G}_n(\boldsymbol{X}_{nr_1}, \boldsymbol{X}_{nr_2})\widetilde{Q}_n(\boldsymbol{X}_{n\ell_3}))$$

and

$$E(\widetilde{G}_n(\boldsymbol{X}_{nr_1}, \boldsymbol{X}_{nr_2})\widetilde{Q}_n(\boldsymbol{X}_{n\ell_3})) = \begin{cases} E(\widetilde{G}_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n1})\widetilde{Q}_n(\boldsymbol{X}_{n1})) & r_1 = r_2 = \ell_3 \\ 0 & \text{otherwise.}\end{cases}$$

Therefore

$$\begin{aligned}E(W_n^4) - s_n^4 &= \sum_{\ell_1=1}^{n}\sum_{\ell_2=1}^{n} E(v_{n\ell_1}v_{n\ell_2}) - s_{n2}^4 \\ &\quad +4\sum_{\ell_1=1}^{n}\sum_{\ell_2=1}^{n}(\min(\ell_1,\ell_2)-1)E(\widetilde{Q}_n^2(\boldsymbol{X}_{n1})) \\ &\quad +4\sum_{\ell_1=1}^{n}\sum_{\ell_2=1}^{n}(\min(\ell_1,\ell_2)-1)E(\widetilde{G}_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n1})\widetilde{Q}_n(\boldsymbol{X}_{n1})).\end{aligned}$$

The term $\sum_{\ell_1=1}^{n}\sum_{\ell_2=1}^{n} E(v_{n\ell_1}v_{n\ell_2}) - s_{n2}^4$ is studied in the proof of Theorem 1 in Hall (1984), for which he obtains

$$\sum_{\ell_1=1}^{n}\sum_{\ell_2=1}^{n} E(v_{n\ell_1}v_{n\ell_2}) - s_{n2}^4 = O\left(n^4 E(\widetilde{G}_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2})) + n^3 E(\widetilde{H}_n^4(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}))\right),$$

so that

$$\begin{aligned} s_n^{-4}(E(W_n^4) - s_n^4) &= O\Big(s_n^{-4}n^4E(\widetilde{G}_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2})) + s_n^{-4}n^3E(\widetilde{H}_n^4(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2})) \\ &\qquad + s_n^{-4}n^3E(\widetilde{Q}_n^2(\boldsymbol{X}_{n1})) + s_n^{-4}n^3E(\widetilde{G}_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n1})\widetilde{Q}_n(\boldsymbol{X}_{n1}))\Big). \end{aligned}$$

Condition (4.13), and the fact that $s_n^2 \propto s_{n2}^2 \propto n^2E(\widetilde{H}_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}))$, imply that the first 2 terms converge to zero. For the last term note that, by Hölder's inequality,

$$E(\widetilde{G}_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n1})\widetilde{Q}_n(\boldsymbol{X}_{n1})) \leq \sqrt{E(\widetilde{H}_n^4(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}))E(\widetilde{Q}_n^2(\boldsymbol{X}_{n1}))}.$$

So, to show the validity of (4.21), it now suffices to prove $s_n^{-4}n^3E(\widetilde{Q}_n^2(\boldsymbol{X}_{n1})) \to 0$. By Hölder's inequality we have

$$s_n^{-4}n^3E(\widetilde{Q}_n^2(\boldsymbol{X}_{n1})) \leq \sqrt{s_n^{-4}n^4E(\widetilde{G}_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}))} \times s_n^{-2}nE(\widetilde{T}_{n1}^2)$$

which converges to zero, since the first part converges to zero and the second part converges to a constant. ■

4.3 Proofs

Proof of (4.10)

Using (4.8) and (4.9) we combine

$$\begin{aligned} E(T_{n1}^2) &= \sum_{i_1=1}^{k}\sum_{i_2=1}^{k} b_{i_1}b_{i_2}E(X_{1i_1}X_{1i_2}), \\ E(X_{1i_1}X_{1i_2}) &= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k} S_{i_1j_1}S_{i_2j_2}E\left(\varepsilon_{1j_1}\varepsilon_{1j_2}\right) \end{aligned}$$

and $E\left(\varepsilon_{1j_1}\varepsilon_{1j_2}\right) = p_{j_1}(\delta_{j_1j_2} - p_{j_2})$ to obtain

$$E(T_{n1}^2) = \sum_{j=1}^{k} p_j\left(\sum_{i=1}^{k} b_iS_{ij}\right)^2 - \left(\sum_{j=1}^{k} p_j\sum_{i=1}^{k} b_iS_{ij}\right)^2. \tag{4.22}$$

Define the set $I' = \left\{j : 2Lhk + \frac{1}{2} \leq j \leq (1-2Lh)k + \frac{1}{2}\right\}$ of "purely"interior points. Since $S_{ij} = 0$ for $|x_i - x_j| > Lh$ we only have to consider those indices i such

that $|x_i - x_j| \leq Lh$. If $j \in I'$ such i is an interior index.
From expression (3.17) for the bias of the local polynomial, the continuity of $f^{(p+1)}(\cdot)$ and (3.12) we obtain, uniformly in i and j, with $j \in I'$,

$$b_i = \frac{h^{p+1}}{k}\frac{f^{(p+1)}(x_j)}{(p+1)!}\mu_{p+1}(L_{(p)}) + o\left(\frac{h^{p+1}}{k}\right),$$

so that, uniformly in $j \in I'$,

$$\sum_{i=1}^{k} b_i S_{ij} = \frac{h^{p+1}}{k}\frac{f^{(p+1)}(x_j)}{(p+1)!}\mu_{p+1}(L_{(p)})\sum_{i=1}^{k} S_{ij} + o\left(\frac{h^{p+1}}{k}\right).$$

Since both i and j are interior indices $S_{ij} = S_{ji}$ (by (3.9) and symmetry of the kernel function $K(\cdot)$). Therefore we have for $j \in I'$ $\sum_{i=1}^{k} S_{ij} = \sum_{j=1}^{k} S_{ij} = 1$ (by (1.14)). These facts result in

$$\sum_{j\in I'} p_j\left(\sum_{i=1}^{k} b_i S_{ij}\right)^2 - \left(\sum_{j\in I'} p_j \sum_{i=1}^{k} b_i S_{ij}\right)^2 =$$

$$\frac{h^{2(p+1)}}{k^2}\frac{\mu_{p+1}^2(L_{(p)})}{((p+1)!)^2}\left\{\sum_{j\in I'} p_j (f^{(p+1)})^2(x_j) - \left(\sum_{j\in I'} p_j f^{(p+1)}(x_j)\right)^2\right\} + o\left(\frac{h^{2(p+1)}}{k^2}\right).$$

Now use $p_j = f(x_j)/k + O(k^{-3})$, uniformly in j, (see (1.21)) and Lemma 3.2 to obtain

$$\sum_{j\in I'} p_j\left(\sum_{i=1}^{k} b_i S_{ij}\right)^2 - \left(\sum_{j\in I'} p_j \sum_{i=1}^{k} b_i S_{ij}\right)^2 = \frac{h^{2(p+1)}}{k^2}\frac{\mu_{p+1}^2(L_{(p)})}{((p+1)!)^2} \times$$

$$\left\{\int_0^1 f(x)(f^{(p+1)}(x))^2\,dx - \left(\int_0^1 f(x)f^{(p+1)}(x)\,dx\right)^2\right\} + o\left(\frac{h^{2(p+1)}}{k^2}\right). \quad (4.23)$$

Since, uniformly in i and j, $p_j = O(k^{-1})$, $b_i = O(h^{p+1}/k)$, $S_{ij} = O((kh)^{-1})$ (by Lemma 1.2), $S_{ij} = 0$ when $|i-j| > 2Lkh$ and $\#(\{1,\ldots,k\}\backslash I') = O(kh)$, we have

$$\sum_{j\notin I'} p_j\left(\sum_{i=1}^{k} b_i S_{ij}\right)^2 - \left(\sum_{j\notin I'} p_j \sum_{i=1}^{k} b_i S_{ij}\right)^2 = o\left(\frac{h^{2(p+1)}}{k^2}\right). \quad (4.24)$$

Now (4.10) follows from (4.22)-(4.24). ■

Proof of (4.11)
From (4.8) and (4.9) we have

$$E(T_{n1}^4) = \sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} b_{i_1}b_{i_2}b_{i_3}b_{i_4}E(X_{1i_1}X_{1i_2}X_{1i_3}X_{1i_4})$$

and

$$E(X_{1i_1}X_{1i_2}X_{1i_3}X_{1i_4}) = \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\sum_{j_3=1}^{k}\sum_{j_4=1}^{k} S_{i_1j_1}S_{i_2j_2}S_{i_3j_3}S_{i_4j_4}E(\varepsilon_{1j_1}\varepsilon_{1j_2}\varepsilon_{1j_3}\varepsilon_{1j_4})$$

where $\varepsilon_{\ell j}$ is defined in (4.7). It is easy to see that

$$E(\varepsilon_{1j_1}\varepsilon_{1j_2}\varepsilon_{1j_3}\varepsilon_{1j_4}) = \begin{cases} O(\frac{1}{k}) & j_1 = j_2 = j_3 = j_4 \\ O(\frac{1}{k^2}) & j_1 = j_2 = j_3 \neq j_4 \\ O(\frac{1}{k^3}) & j_1 = j_2 \neq j_3 = j_4 \\ O(\frac{1}{k^3}) & j_1 \neq j_2 \neq j_3 = j_4 \\ O(\frac{1}{k^4}) & j_1 \neq j_2 \neq j_3 \neq j_4 \end{cases}. \tag{4.25}$$

For a fixed index j the number of indices i such that $S_{ij} \neq 0$ is $O(kh)$ and $S_{ij} = O((kh)^{-1})$. Further, $b_i = O(h^{p+1}k^{-1})$ uniform in the i-index (see (3.17)). Now it is just a matter of counting the number of indices that make a contribution to the sum and to use (4.25) to see that (4.11) holds. Consider e.g., the situation $j_1 = j_2 = j_3 = j_4$, for which the contribution to $E(T_{n1}^4)$ is

$$\sum_{j_1=1}^{k}\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} b_{i_1}b_{i_2}b_{i_3}b_{i_4}S_{i_1j_1}S_{i_2j_1}S_{i_3j_1}S_{i_4j_1}E(\varepsilon_{1j_1}^4)$$
$$= O\left(k \times (kh)^4 \times \left(\frac{h^{p+1}}{k}\right)^4 \times \frac{1}{(kh)^4} \times \frac{1}{k}\right) = O\left(\frac{h^{4(p+1)}}{k^4}\right).$$

The other combinations can be treated similarly and are of order $o(h^{4(p+1)}k^{-4})$. ■

Proof of (4.14)
From (4.4), the fact that $\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{nn}$ are i.i.d. and (4.9) it is clear that

$$\begin{aligned} E(H_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2})) &= \sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\left(E(X_{1i_1}X_{1i_2})\right)^2 \\ &= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\sum_{j_3=1}^{k}\sum_{j_4=1}^{k} D_{j_1j_3}D_{j_2j_4}E(\varepsilon_{1j_1}\varepsilon_{1j_2})E(\varepsilon_{1j_3}\varepsilon_{1j_4}). \quad (4.26) \end{aligned}$$

It turns out that the situation $j_1 = j_2$ and $j_3 = j_4$ is the term that determines the order of $E(H_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}))$. We first discuss this situation and return to the other terms later. Since $E(\varepsilon_{1j_1}\varepsilon_{1j_2}) = p_{j_1}(\delta_{j_1j_2} - p_{j_2})$ this leading term has the form

$$\begin{aligned} &\sum_{j_1=1}^{k}\sum_{j_2=1}^{k} D_{j_1j_2}^2 p_{j_1}(1-p_{j_1})p_{j_2}(1-p_{j_2}) \\ &= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k} D_{j_1j_2}^2 p_{j_1}p_{j_2} + \sum_{j_1=1}^{k}\sum_{j_2=1}^{k} D_{j_1j_2}^2 p_{j_1}p_{j_2}\left(p_{j_1}p_{j_2} - p_{j_1} - p_{j_2}\right), \quad (4.27) \end{aligned}$$

and its major contribution is the term $\sum_{j_1=1}^{k}\sum_{j_2=1}^{k} D_{j_1j_2}^2 p_{j_1}p_{j_2}$. From the fact that $S_{ij} = 0$ if $|i-j| > Lkh$ it is easy to see that also

$$D_{j_1j_2} = 0 \quad \text{if} \quad |j_1 - j_2| > 2Lkh \qquad (4.28)$$

which implies that for a fixed index j_1 the number of indices j_2 such that $D_{j_1j_2} \neq 0$ is $O(kh)$. From $S_{ij} = O((kh)^{-1})$, uniformly in i and j, and the definition of $D_{j_1j_2}$ follows that also $D_{j_1j_2} = O((kh)^{-1})$, uniformly in j_1 and j_2.

Now, uniformly in j_1, $p_{j_1} = f(x_{j_1})/k + O(k^{-3})$, and, uniformly in j_1 and j_2 with $|j_1 - j_2| \leq 2Lkh$, $p_{j_2} = f(x_{j_1})/k + O(k^{-3})$, which yields

$$\sum_{j_1=1}^{k}\sum_{j_2=1}^{k} D_{j_1j_2}^2 p_{j_1}p_{j_2} = \sum_{j_1=1}^{k}\frac{f^2(x_{j_1})}{k^2}\sum_{j_2=1}^{k} D_{j_1j_2}^2 + O\left(\frac{1}{k^4h}\right) \qquad (4.29)$$

Recall that $D_{j_1j_2} = \sum_{i=1}^{k} S_{ij_1}S_{ij_2}$. If we consider $j_1 \in I'$, with I' defined in the proof of (4.10), the indices i that have a contribution to $D_{j_1j_2}$ are interior indices, therefore, uniformly in j_1 and j_2, (by (3.11))

$$D_{j_1j_2} = \frac{1}{(kh)^2}\sum_{i=1}^{k} L_{(p)}\left(\frac{x_{j_1} - x_i}{h}\right)L_{(p)}\left(\frac{x_{j_2} - x_i}{h}\right) + o\left(\frac{1}{kh}\right)$$

$$= \frac{1}{(kh)^2}\sum_{i=1}^{k} L_{(p)}\left(\frac{x_{j_1}-x_i}{h}\right) L_{(p)}\left(\frac{x_{j_1}-x_i+(x_{j_2}-x_{j_1})}{h}\right) + o\left(\frac{1}{kh}\right).$$

We have, uniformly in j_1 and j_2, with $j_1 \in I'$,

$$D_{j_1 j_2} = \frac{1}{kh}\left(\int_{-L}^{L} L_{(p)}(u)L_{(p)}\left(u+\frac{x_{j_2}-x_{j_1}}{h}\right) du\right) + o\left(\frac{1}{kh}\right),$$

for which the proof is analogous to that of Lemma 1.1. Similarly, we have, uniformly in $j_1 \in I'$,

$$\sum_{j_2=1}^{k} D^2_{j_1 j_2} = \frac{1}{kh}\int_{-2L}^{2L}\left[\int_{-L}^{L} L_{(p)}(u)L_{(p)}(u+v)\,du\right]^2 dv + o\left(\frac{1}{kh}\right).$$

Combine all this to obtain

$$\sum_{j_1\in I'} \frac{f^2(x_{j_1})}{k^2}\sum_{j_2=1}^{k} D^2_{j_1 j_2} = \frac{1}{k^2 h}\sigma_2^2 + o\left(\frac{1}{k^2 h}\right). \tag{4.30}$$

Also for $j_1 \notin I'$, $\sum_{j_2=1}^{k} D^2_{j_1 j_2} = O((kh)^{-1})$, but $\#\{1,\ldots,k\}\setminus I' = O(kh)$, which yields

$$\sum_{j_1\notin I'} \frac{f^2(x_{j_1})}{k^2}\sum_{j_2=1}^{k} D^2_{j_1 j_2} = O\left(\frac{1}{k^2}\right) = o\left(\frac{1}{k^2 h}\right). \tag{4.31}$$

Combine (4.29), (4.30) and (4.31) to obtain

$$\sum_{j_1=1}^{k}\sum_{j_2=1}^{k} D^2_{j_1 j_2} p_{j_1} p_{j_2} = \frac{1}{k^2 h}\sigma_2^2 + o\left(\frac{1}{k^2 h}\right). \tag{4.32}$$

For the second term on the r.h.s. of (4.27) we have

$$\sum_{j_1=1}^{k}\sum_{j_2=1}^{k} D^2_{j_1 j_2} p_{j_1} p_{j_2}\left(p_{j_1}p_{j_2} - p_{j_1} - p_{j_2}\right) = O\left(k \times kh \times \frac{1}{(kh)^2} \times \frac{1}{k^3}\right) = o\left(\frac{1}{k^2 h}\right),$$

where the order bound is obtained by counting the number of indices that have a contribution, based on the properties of $D_{j_1 j_2}$, and by $p_i = O(k^{-1})$, uniformly in i.

To see that the situations other than $j_1 = j_2$ and $j_3 = j_4$ are of lower order, consider e.g., the situation $j_1 \neq j_2$ and $j_3 = j_4$, whose contribution to (4.27) is

$$\sum_{j_1=1}^{k} \sum_{\substack{j_2=1 \\ j_2 \neq j_1}}^{k} \sum_{j_3=1}^{k} D_{j_1 j_3} D_{j_2 j_3} p_{j_1} p_{j_2} p_{j_3} (1 - p_{j_3})$$
$$= O\left(k \times (kh)^2 \times \frac{1}{(kh)^2} \times \frac{1}{k^3}\right) = O\left(\frac{1}{k^2}\right) = o\left(\frac{1}{k^2 h}\right),$$

where the order bound is obtained in a similar way as above. The other situations can be treated similarly and are of order $o((k^2 h)^{-1})$. ■

Proof of (4.15)
From (4.4) it is clear that

$$E(H_n^4(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2})) \;=\; \sum_{i_1=1}^{k} \sum_{i_2=1}^{k} \sum_{i_3=1}^{k} \sum_{i_4=1}^{k} (E(X_{1i_1} X_{1i_2} X_{1i_3} X_{1i_4}))^2$$

and by (4.25)

$$E(X_{1i_1} X_{1i_2} X_{1i_3} X_{1i_4}) = \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} \sum_{j_3=1}^{k} \sum_{j_4=1}^{k} S_{i_1 j_1} S_{i_2 j_2} S_{i_3 j_3} S_{i_4 j_4} E(\varepsilon_{1j_1} \varepsilon_{1j_2} \varepsilon_{1j_3} \varepsilon_{1j_4})$$

$$= \sum_{j_1=1}^{k} S_{i_1 j_1} S_{i_2 j_1} S_{i_3 j_1} S_{i_4 j_1} E(\varepsilon_{1j_1}^4)$$

$$+\text{const} \sum\sum_{j_1 \neq j_2}^{k} S_{i_1 j_1} S_{i_2 j_1} S_{i_3 j_1} S_{i_4 j_2} E(\varepsilon_{1j_1}^3 \varepsilon_{1j_2})$$

$$+\text{const} \sum\sum_{j_1 \neq j_2}^{k} S_{i_1 j_1} S_{i_2 j_1} S_{i_3 j_2} S_{i_4 j_2} E(\varepsilon_{1j_1}^2 \varepsilon_{1j_2}^2)$$

$$+\text{const} \sum\sum\sum_{j_1 \neq j_2 \neq j_3}^{k} S_{i_1 j_1} S_{i_2 j_2} S_{i_3 j_3} S_{i_4 j_3} E(\varepsilon_{1j_1} \varepsilon_{1j_2} \varepsilon_{1j_3}^2)$$

$$+\text{const} \sum\sum\sum\sum_{j_1 \neq j_2 \neq j_3 \neq j_4}^{k} S_{i_1 j_1} S_{i_2 j_2} S_{i_3 j_3} S_{i_4 j_4} E(\varepsilon_{1j_1} \varepsilon_{1j_2} \varepsilon_{1j_3} \varepsilon_{1j_4})$$

$$\overset{def.}{=} A^{[1]}_{i_1 i_2 i_3 i_4} + A^{[2]}_{i_1 i_2 i_3 i_4} + A^{[3]}_{i_1 i_2 i_3 i_4} + A^{[4]}_{i_1 i_2 i_3 i_4} + A^{[5]}_{i_1 i_2 i_3 i_4},$$

where the constants are ignored in the definition of the A–terms. By (4.7), $E(H_n^4(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}))$ can be written in terms of $D_{j_1j_2}$'s. Further, based on (4.25), $D_{j_1j_2} = 0$ if $|j_1 - j_2| > 2Lkh$ and $D_{j_1j_2} = O((kh)^{-1})$, we can see that (4.15) is valid. Indeed, consider e.g.,

$$\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k}\left(A^{[1]}_{i_1i_2i_3i_4}\right)^2 = \sum_{j_1=1}^{k}\sum_{j_2=1}^{k} E(\varepsilon_{1j_1}^4)E(\varepsilon_{1j_2}^4)D_{j_1j_2}^4$$
$$= O(k \times kh \times \frac{1}{k^2} \times \frac{1}{(kh)^4}) = O\left(\frac{1}{k^4h^3}\right)$$

and

$$\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} A^{[1]}_{i_1i_2i_3i_4}A^{[2]}_{i_1i_2i_3i_4} = \sum_{j_1=1}^{k}\sum^{k}\sum_{j_2 \neq j_3} E(\varepsilon_{1j_1}^4)E(\varepsilon_{1j_2}^3\varepsilon_{1j_3})D_{j_1j_2}^3 D_{j_1j_3}$$
$$= O(k \times (kh)^2 \times \frac{1}{k^3} \times \frac{1}{(kh)^4}) = O\left(\frac{1}{k^4h^2}\right) = o\left(\frac{1}{k^4h^3}\right).$$

All the other terms can be treated similarly and are $o((k^4h^3)^{-1})$. ■

Proof of (4.16)
From (4.12) and (4.4) it is clear that

$$E(G_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2})) = E\left(H_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n3})H_n(\boldsymbol{X}_{n2}, \boldsymbol{X}_{n3})H_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n4})H_n(\boldsymbol{X}_{n2}, \boldsymbol{X}_{n4})\right)$$
$$= \sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} E(X_{1i_1}X_{3i_1}X_{2i_2}X_{3i_2}X_{1i_3}X_{4i_3}X_{2i_4}X_{4i_4})$$
$$= \sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} E(X_{1i_1}X_{1i_3})E(X_{2i_2}X_{2i_4})E(X_{3i_1}X_{3i_2})E(X_{4i_3}X_{4i_4}).$$

Now,

$$E(X_{1i_1}X_{1i_2}) =$$
$$\sum_{j_1=1}^{k} S_{i_1j_1}S_{i_2j_1}p_{j_1} - \left\{\sum_{j_1=1}^{k} S_{i_1j_1}p_{j_1}\right\}\left\{\sum_{j_2=1}^{k} S_{i_2j_2}p_{j_2}\right\} \overset{def.}{=} A_{i_1i_2} - B_{i_1i_2}.$$

Then

$$\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} A_{i_1i_3}A_{i_2i_4}A_{i_1i_2}A_{i_3i_4}$$

$$
\begin{aligned}
&= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\sum_{j_3=1}^{k}\sum_{j_4=1}^{k}\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} S_{i_1j_1}S_{i_3j_1}S_{i_2j_2}S_{i_4j_2}S_{i_1j_3}S_{i_2j_3}S_{i_3j_4}S_{i_4j_4}p_{j_1}p_{j_2}p_{j_3}p_{j_4}\\
&= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\sum_{j_3=1}^{k}\sum_{j_4=1}^{k} D_{j_1j_3}D_{j_2j_3}D_{j_1j_4}D_{j_2j_4}p_{j_1}p_{j_2}p_{j_3}p_{j_4}\\
&= O\left(k^4h^3 \times \frac{1}{(kh)^4} \times \frac{1}{k^4}\right) = O\left(\frac{1}{k^4h}\right),
\end{aligned}
$$

where the order relation follows from counting the number of indices that make a contribution, based on the properties of $D_{j_1j_2}$. Similarly,

$$
\begin{aligned}
&\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} A_{i_1i_3}A_{i_2i_4}A_{i_1i_2}B_{i_3i_4}\\
&= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\sum_{j_3=1}^{k}\sum_{j_4=1}^{k}\sum_{j_5=1}^{k}\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} S_{i_1j_1}S_{i_3j_1}S_{i_2j_2}S_{i_4j_2}S_{i_1j_3}S_{i_2j_3}S_{i_3j_4}S_{i_4j_5}\prod_{\ell=1}^{5}p_{j_\ell}\\
&= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\sum_{j_3=1}^{k}\sum_{j_4=1}^{k}\sum_{j_5=1}^{k} D_{j_1j_3}D_{j_2j_3}D_{j_1j_4}D_{j_2j_5}p_{j_1}p_{j_2}p_{j_3}p_{j_4}p_{j_5}\\
&= O\left(k^5h^4 \times \frac{1}{(kh)^4} \times \frac{1}{k^5}\right) = O\left(\frac{1}{k^4}\right) = o\left(\frac{1}{k^4h}\right).
\end{aligned}
$$

The other combinations can be treated similarly and are of order $o((k^4h)^{-1})$. ■

Proof of (4.17)

From (4.4) and (4.9) we have

$$
\begin{aligned}
E(H_n^2(\boldsymbol{X}_{n1},\boldsymbol{X}_{n1})) &= \sum_{i_1=1}^{k}\sum_{i_2=1}^{k} E(X_{1i_1}^2X_{1i_2}^2)\\
&= \sum_{i_1=1}^{k}\sum_{i_2=1}^{k} E\left(\left\{\sum_{j_1=1}^{k} S_{i_1j_1}\varepsilon_{1j_1}\right\}^2\left\{\sum_{j_2=1}^{k} S_{i_2j_2}\varepsilon_{1j_2}\right\}^2\right)\\
&= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\sum_{j_3=1}^{k}\sum_{j_4=1}^{k} D_{j_1j_2}D_{j_3j_4}E(\varepsilon_{1j_1}\varepsilon_{1j_2}\varepsilon_{1j_3}\varepsilon_{1j_4})\\
&= \sum_{j_1=1}^{k} D_{j_1j_1}^2E(\varepsilon_{1j_1}^4)
\end{aligned}
$$

$$
\begin{aligned}
&+\text{const} \mathop{\sum\sum}_{j_1 \neq j_2}^{k} D_{j_1 j_1} D_{j_1 j_2} E(\varepsilon_{1j_1}^3 \varepsilon_{1j_2}) \\
&+\text{const} \mathop{\sum\sum}_{j_1 \neq j_2}^{k} D_{j_1 j_1} D_{j_2 j_2} E(\varepsilon_{1j_1}^2 \varepsilon_{1j_2}^2) \\
&+\text{const} \mathop{\sum\sum\sum}_{j_1 \neq j_2 \neq j_3}^{k} D_{j_1 j_2} D_{j_3 j_3} E(\varepsilon_{1j_1} \varepsilon_{1j_2} \varepsilon_{1j_3}^2) \\
&+\text{const} \mathop{\sum\sum\sum\sum}_{j_1 \neq j_2 \neq j_3 \neq j_4}^{k} D_{j_1 j_2} D_{j_3 j_4} E(\varepsilon_{1j_1} \varepsilon_{1j_2} \varepsilon_{1j_3} \varepsilon_{1j_4}) \\
\overset{def.}{=}\ & A_1 + A_2 + A_3 + A_4 + A_5.
\end{aligned}
$$

Now use (4.25) and count the number of indices that make a contribution, based on the properties of $D_{j_1 j_2}$, to see that

$$A_1 = O\left(k \times \frac{1}{(kh)^2} \times \frac{1}{k}\right) = O\left(\frac{1}{(kh)^2}\right).$$

Further,

$$A_2 = O\left(k^2 h \times \frac{1}{(kh)^2} \times \frac{1}{k^2}\right) = O\left(\frac{h}{(kh)^2}\right) = o\left(\frac{1}{(kh)^2}\right).$$

The other terms can be treated similarly and are $o((kh)^{-2})$. ■

Chapter 5

Multi-dimensional tables

This chapter deals with a generalization of local polynomial estimators for cell probabilities introduced in Chapter 1 to d-dimensional tables. We consider contingency tables with k_j ordered cells in the j-th dimension, $j = 1, \ldots, d$. We investigate the sparse asymptotic behavior of the mean sum of squared errors of the multi-dimensional local linear smoothers for the cell probabilities, and show that the MSSE converges to zero at a faster rate than for the frequency estimators. Our sparse asymptotic framework is of a form in which the dimension d of the table is fixed, but in each dimension the number of cells k_j, $j = 1, \ldots, d$, tends to infinity. Grund and Hall (1993) study kernel smoothing in multi-dimensional sparse tables, where the number of cells in each dimension is fixed ($k_j = 2$ in their case) and the dimension d tends to infinity.

An important parameter in the construction of kernel-type estimators is the bandwidth. This parameter defines the local neighborhood of a cell, and, for kernels with compact support, only these neighbors have a contribution to the estimator of the cell probability. There are several levels of options to parameterize this local neighborhood in the multi-dimensional setting. The simplest multivariate kernel estimator uses a single smoothing parameter h, as in the one-dimensional case. This means that the amount of smoothing is the same in all directions. A simple generalization is to use d different bandwidth parameters $h_1, \ldots, h_d$ which allow different levels of smoothing in each of the coordinate directions. A general way is to work with a bandwidth parameter that allows for smoothing along directions different from the coordinate axes. A general $d \times d$ bandwidth matrix permits this kind of smoothing. The idea of a general bandwidth matrix goes back to Deheuvels (1977). He introduces general bandwidth matrices in the context of multivariate density estimation. Wand and Jones (1993) illustrate in the bivariate density estimation context the beneficial effect of a general bandwidth parameterization. This is our main motivation to consider a general bandwidth matrix in the definition of our multi-dimensional estimators for the cell probabilities.

Ruppert and Wand (1994) investigate local linear and quadratic regression estimation in d dimensions. We adapt these results for local linear estimation to our sparse table problem, which is reformulated as a fixed design multiple regression problem, in the same way as in the one-dimensional situation.

Since in higher dimensional tables the boundary region can be quite large it is desirable to consider estimators which do not suffer from boundary problems. From Chapters 1 and 3 we know that local polynomial smoothers with odd degree have this property in the one-dimensional case. We investigate in Section 5.1 the MSSE performance of local linear smoothers for cell probabilities based on a general $d \times d$ bandwidth matrix. For notational and technical simplicity we restrict to the local linear smoothers. In Section 5.2 we illustrate through simulations the benefit of working with a general $d \times d$ bandwidth matrix. In Section 5.3 we obtain for the multi-dimensional local linear smoothers a generalization of the central limit result given in Chapter 4. Based on the material we present in Section 5.1 and the technique used in Chapter 2 it is also possible to show that the multi-dimensional local linear smoothers are sparse asymptotic consistent.

5.1 Local linear estimators for the cell probabilities

For one-dimensional tables we considered cell probabilities generated by an underlying latent density on $[0,1]$. Each cell i in the one-dimensional table corresponded to the interval $I_i = [(i-1)/k, i/k]$. Extension of this to the d-dimensional case is straightforward. Consider the unit cube $[0,1]^d$, partition the $[0,1]$-interval in the ℓ-th dimension in k_ℓ subsequent intervals $I_{\ell i}$ of equal length with midpoints $x_{\ell i} = (i-\frac{1}{2})/k_\ell$, $i = 1,\ldots,k_\ell$. Link each cartesian product $I_{i_1,\ldots,i_d} = I_{1i_1} \times I_{2i_2} \times \ldots \times I_{di_d}$ with the cell in the contingency table having multiple index $(i_1,\ldots,i_d)$. To avoid the use of multiple indices we relabel, in an arbitrary but fixed manner, $I_{i_1,\ldots,i_d}$ as C_j with midpoint $\boldsymbol{x}_j$,where $j = 1,\ldots,k = \prod_{\ell=1}^d k_\ell$.

Let $\boldsymbol{p}^T = (p_1,\ldots,p_k)$ be the vector of cell probabilities and $\overline{\boldsymbol{P}}^T = (\overline{P}_1,\ldots,\overline{P}_k)$ the vector of frequency estimators.

Denote, for $i = 1,\ldots,k$, $\boldsymbol{W}_i$ the weight matrix given by

$$\boldsymbol{W}_i = \mathrm{diag}(K_{\boldsymbol{H}}(\boldsymbol{x}_1 - \boldsymbol{x}_i),\ldots,K_{\boldsymbol{H}}(\boldsymbol{x}_k - \boldsymbol{x}_i))$$

with $K_{\boldsymbol{H}}(\cdot) = \frac{1}{|\boldsymbol{H}|^{1/2}} K(\boldsymbol{H}^{-1/2}(\cdot))$, where $K(\cdot)$ is a d-dimensional kernel function and $\boldsymbol{H}$ a positive definite and symmetric $d \times d$ bandwidth matrix. The multi-dimensional local polynomial smoother for the cell probabilities is defined as

$$\widehat{P}_i = \boldsymbol{e}_1^T(\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i)^{-1}(\boldsymbol{X}_i^T \boldsymbol{W}_i \overline{\boldsymbol{P}}) \tag{5.1}$$

where $\boldsymbol{e}_1^T$ is the $(d+1)$-vector $(1, 0, \ldots, 0)$ and $\boldsymbol{X}_i$ an appropriate design matrix.

When the design matrix is taken to be $\boldsymbol{X}_i = (1, \ldots, 1)^T$ the resulting estimator is the local constant smoother. As one can expect from the one-dimensional situation, this estimator suffers from boundary bias problems. Therefore, the optimal rate for the MSSE can only be attained under extra boundary conditions on $f(\ cdot)$. Burmann (1987a) studies some version of the local constant smoother with a diagonal bandwidth matrix.

To avoid the boundary conditions, boundary corrected kernels can be used. Dong and Simonoff (1995) consider multi-dimensional boundary corrected kernel estimators based on a diagonal bandwidth matrix. Local linear smoothing is an alternative method to avoid boundary conditions, and is such that no special boundary correction is needed, i.e., the local linear estimator adapts automatically in the boundary region.

We obtain the local linear estimator if we take the $k \times (d+1)$ design matrix

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & (\boldsymbol{x}_1 - \boldsymbol{x}_i)^T \\ \vdots & \vdots \\ 1 & (\boldsymbol{x}_k - \boldsymbol{x}_i)^T \end{pmatrix}.$$

It is clear how to define local polynomial smoothers of order $p > 1$. We restrict attention to local linear estimators, for notational and technical simplicity. From Chapter 3 it is clear that local quadratic estimators will suffer from boundary bias problems (see also Ruppert and Wand (1994) who study, in the regression context, local quadratic smoothers in detail). We restrict attention to local linear smoothers.

Similar as in Section 1.3 we can rewrite the local linear estimator as a usual kernel type estimator. To see this write

$$\frac{1}{k}(\boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i) = \begin{pmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{H}^{1/2} \end{pmatrix} \boldsymbol{N}_i \begin{pmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{H}^{1/2} \end{pmatrix}$$

with $\boldsymbol{N}_i$ the $(d+1)\times(d+1)$ matrix

$$\frac{1}{k}\begin{pmatrix} \sum_{j=1}^{k} K_{\boldsymbol{H}}(\boldsymbol{x}_j-\boldsymbol{x}_i) & \sum_{j=1}^{k} K_{\boldsymbol{H}}(\boldsymbol{x}_j-\boldsymbol{x}_i)(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i))^T \\ \sum_{j=1}^{k} K_{\boldsymbol{H}}(\boldsymbol{x}_j-\boldsymbol{x}_i)\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i) & \sum_{j=1}^{k} K_{\boldsymbol{H}}(\boldsymbol{x}_j-\boldsymbol{x}_i)\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i))^T \end{pmatrix}$$

The local linear estimator then becomes

$$\widehat{P}_i \;=\; \boldsymbol{e}_1^T \boldsymbol{N}_i{}^{-1}\begin{pmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{H}^{-1/2}\end{pmatrix}\frac{1}{k}\boldsymbol{X}_i^T\boldsymbol{W}_i\overline{\boldsymbol{P}}.$$

The j-the column of the $(d+1)\times k$-matrix

$$\begin{pmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{H}^{-1/2}\end{pmatrix}\frac{1}{k}\boldsymbol{X}_i^T\boldsymbol{W}_i$$

is $(1\ (\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i))^T)^T K_{\boldsymbol{H}}\,(\boldsymbol{x}_j-\boldsymbol{x}_i)$, so that

$$\left(\boldsymbol{e}_1^T \boldsymbol{N}_i{}^{-1}\begin{pmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{H}^{-1/2}\end{pmatrix}\frac{1}{k}\boldsymbol{X}_i^T\boldsymbol{W}_i\right)_j =$$
$$\frac{1}{k|\boldsymbol{H}|^{1/2}}\frac{1}{|\boldsymbol{N}_i|}(\text{adj}\boldsymbol{N}_i)_{.1}^T\begin{pmatrix} 1 \\ \boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\end{pmatrix} K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\right),$$

where $\boldsymbol{A}_{.1}$ denotes the first column of the matrix $\boldsymbol{A}$.
The term $(\text{adj}\boldsymbol{N}_i)_{.1}^T(1\ (\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i))^T)^T$ can be seen as the determinant of the $(d+1)\times(d+1)$ matrix $\boldsymbol{N}_i$ with the first column replaced by $(1\ (\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i))^T)^T$. Combining these results yields

$$\widehat{P}_i \;=\; \frac{1}{k|\boldsymbol{H}|^{1/2}}\sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i))\overline{P}_j \qquad (5.2)$$

with $L_i(\boldsymbol{u}) = \dfrac{|\boldsymbol{M}_i(\boldsymbol{u})|}{|\boldsymbol{N}_i|}K(\boldsymbol{u})$, where $\boldsymbol{M}_i(\boldsymbol{u})$ is the same as $\boldsymbol{N}_i$ but with the first column replaced by $(1\ \boldsymbol{u}^T)^T$. Further, it is easy to see that, for all $i \in \{1, \dots, k\}$,

$$\begin{aligned} \frac{1}{k|\boldsymbol{H}|^{1/2}} \sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)) &= 1 \\ \sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))(\boldsymbol{x}_j - \boldsymbol{x}_i) &= \mathbf{0}. \end{aligned} \tag{5.3}$$

This illustrates that each $L_i(\cdot)$ can be considered as a multivariate discrete kernel of order 2, which is a multi-dimensional version of the discrete order property (1.14) for $p = 1$. Similar to the one-dimensional case, this property ensures that the local linear estimator does not suffer from boundary bias problems. The weight function $L_i(\cdot)$ has properties similar to the weight function in the one-dimensional case. For details see Lemmas 5.2 and 5.3 and the remarks following these lemmas.

In order to study the MSSE for the local linear estimator we will assume that the cell probabilities are generated by a density on $[0,1]^d$ through the relation

$$p_i = \int\limits_{C_i} f(\boldsymbol{x})\, d\boldsymbol{x}, \qquad i = 1, \dots, k,$$

which is an immediate generalization of the latent density assumption (1.9). Further we will assume the following conditions hold:

(C.1) *$f(\cdot)$ has continuous second order partial derivatives on $[0,1]^d$,*

(C.2) *K is continuous with compact and convex support supp(K),*

(C.3) *$\int u_i K(\boldsymbol{u})\, d\boldsymbol{u} = 0$, $i = 1, \dots, d$ and $\int \boldsymbol{u}\boldsymbol{u}^T K(\boldsymbol{u})\, d\boldsymbol{u} = \mu_2(K)\boldsymbol{I}_d$ with $\mu_2(K) = \int u_i^2 K(\boldsymbol{u})\, d\boldsymbol{u}$ a strictly positive scalar and $\boldsymbol{I}_d$ the $d \times d$ identity matrix,*

(C.4) *$tr(\boldsymbol{H}) \to 0$ and $k^{2/d} tr(\boldsymbol{H}) \to \infty$,*

(C.5) *there exists a fixed constant L such that the condition number of $\boldsymbol{H}$ (the ratio of its largest to its smallest eigenvalue) is at most L for all n,*

(C.6) *$k_j \propto k^{1/d}$.*

Remark 5.1
(i) For $d = 1$ and $H = h^2$, the conditions on the bandwidth matrix $\boldsymbol{H}$ reduce to $h \to 0$ and $kh \to \infty$, the usual condition in the one-dimensional case.

(ii) Consider the special bandwidth matrix $\boldsymbol{H} = h^2\boldsymbol{I}_d$. Condition (C.4) reduces to $h \to 0$ and $kh^d \to \infty$, which is by (C.6), equivalent to $k_j h \to \infty$, $j = 1, \ldots, d$. Since $\boldsymbol{H} = h^2\boldsymbol{I}_d$ only has one eigenvalue, h^2, condition (C.5) is satisfied.
For the bandwidth matrix $\boldsymbol{H} = \text{diag}(h_1^2, \ldots, h_d^2)$ condition (C.4) is equivalent to $h_j \to 0$, $j = 1, \ldots, d$ and $k^{2/d}(h_1^2 + \ldots h_d^2) \to \infty$. Condition (C.5) implies that all bandwidths are of the same order, such that, by (C.6), (C.4) reduces to $h_j \to 0$ and $k_j h_j \to \infty$, $j = 1, \ldots, d$.
(iii) Spherically symmetric kernels with support a sphere with center $\mathbf{0}$ and product kernels based on symmetric univariate kernels on a compact support satisfy condition (C.3). In these cases all odd order moments of $K(\cdot)$ vanish, that is $\int_{supp(K)} u_1^{\ell_1} \ldots u_d^{\ell_d} K(\boldsymbol{u})\, d\boldsymbol{u} = 0$ for all non-negative integers $\ell_1, \ldots, \ell_d$ such that their sum is odd. This is an assumption which is needed to study higher-degree polynomial fitting.

Before we give our main result, we first collect some technical results. Remark 5.2 contains a note on the number of indices that have a contribution to $(k|\boldsymbol{H}|^{1/2})^{-1} \sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))\overline{P}_j$. Lemma 5.1 shows how to replace sums by integrals in the multi-dimensional situation. An immediate application yields a property of the weight function $L_i(\cdot)$, and Lemma 5.3 says that this function is bounded. The proofs of these results are given in Section 5.4.

Remark 5.2
Let $supp\left(K_{\boldsymbol{H}}(\cdot - \boldsymbol{x}_i)\right) = \{\boldsymbol{y} : \boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i) \in supp(K)\}$.
Since the design is equidistant on $[0,1]^d$, we have that

$$\begin{aligned} \#\{j \ : \ \boldsymbol{x}_j \in supp\left(K_{\boldsymbol{H}}(\cdot - \boldsymbol{x}_i)\right)\} &= O\left(k\text{vol}\left(K_{\boldsymbol{H}}(\cdot - \boldsymbol{x}_i)\right)\right) \\ &= O\left(k|\boldsymbol{H}|^{1/2}\text{vol}\left(supp(K)\right)\right). \end{aligned}$$

By the fact that $supp(K)$ is bounded we obtain that for a fixed index i, $O(k|\boldsymbol{H}|^{1/2})$ indices j have $K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)) \neq 0$.
Since $L_i(u) = (|\boldsymbol{M}_i(u)|/|\boldsymbol{N}_i|)K(u)$, this weight function has the same support as the kernel function $K(\cdot)$. Therefore, we also have that for a fixed index i, $O(k|\boldsymbol{H}|^{1/2})$ indices j have $L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)) \neq 0$. This means that only $O(k|\boldsymbol{H}|^{1/2})$ indices j have contribute to $(k|\boldsymbol{H}|^{1/2})^{-1} \sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))\overline{P}_j$.

Lemma 5.1
(i) Assume (C.4) - (C.6) and let $G(\cdot)$ be a real-valued continuous function defined

on a compact and convex subset of $I\!R^d$. Then, uniformly in the i-index,

$$\frac{1}{k|\boldsymbol{H}|^{1/2}}\sum_{j=1}^{k} G(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)) = \frac{1}{|\boldsymbol{H}|^{1/2}}\int\limits_{[0,1]^d} G(\boldsymbol{H}^{-1/2}(\boldsymbol{y}-\boldsymbol{x}_i))\,d\boldsymbol{y} + o(1).$$

(ii) Assume $g(\cdot)$ is a real-valued continuous function defined on $[0,1]^d$. Let $S \subset \{1,\ldots,k\}$ with $\{1,\ldots,k\}\setminus S = o(k)$. If $k\to\infty$, then

$$\frac{1}{k}\sum_{i\in S} g(\boldsymbol{x}_i) = \int\limits_{[0,1]^d} g(\boldsymbol{x})\,d\boldsymbol{x} + o(1).$$

We now define, for multi-dimensional tables, what we mean by interior and boundary points. A cell midpoint $\boldsymbol{x}_i$ is called an interior point if $supp\left(K(\boldsymbol{H}^{-1/2}(\cdot-\boldsymbol{x}_i))\right) \subset [0,1]^d$ and the set of interior indices is defined as $I = \{i : \boldsymbol{x}_i \text{ is an interior point}\}$. The cell midpoints that are not interior are called boundary points and $B = \{i : \boldsymbol{x}_i \text{ is a boundary point}\}$ is the set of boundary indices. An important remark is that the number of boundary points is of smaller order than the number of interior points. More precisely, the number of boundary points is

$$\#B = O\left(k\sqrt{tr(\boldsymbol{H})}\right), \tag{5.4}$$

while the number of interior points is $\#I = O(k)$. The proof of (5.4) will be given in Section 5.4.

An application of Lemma 5.1(i) gives, for interior points, the following result.

Lemma 5.2
Assume (C2) - (C6). Let $\boldsymbol{J}$ be a $d\times d$ matrix of ones. If $\boldsymbol{x}_i$ is an interior point, then, uniformly in the i-index

(i) $L_i(\boldsymbol{u}) = K(\boldsymbol{u}) + o(1)$ uniformly in $\boldsymbol{u}$

(ii) $$\frac{1}{k|\boldsymbol{H}|^{1/2}}\sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i))(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i))(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i))^T = \mu_2(K)\boldsymbol{I}_d + o(\boldsymbol{J}).$$

Remark 5.3
Lemma 5.2(i) implies that in the interior region the local linear estimator is asymptotically equivalent to the classical kernel estimator. Further, since the result is uniformly in $\boldsymbol{u}$ and i, we have, uniformly in the i and j-index,

$$L_i\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\right)=K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\right)+o(1). \tag{5.5}$$

Lemma 5.3
Assume (C.2)–(C.6). We have that $L_i(\cdot)$ is bounded, uniformly in the i-index.

Remark 5.4
The weight function $L_i(\cdot)$ has the same support as $K(\cdot)$, and is bounded on this support by Lemma 5.3. Denote C_L the bound of $|L_i(\cdot)|$. By Remark 5.2 we have

$$\frac{1}{k|\boldsymbol{H}|^{1/2}}\sum_{j=1}^{k} g_1\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\right)g_2\left(L_i\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\right)\right)=O(1),$$

uniformly in the i-index, where $g_1(\cdot)$ is a continuous function with support $supp(K)$, and $g_2(\cdot)$ a continuous function on $[-C_L, C_L]$ with $g_2(0)=0$.

The main result of this section reads as follows.

Theorem 5.1
Assume (C.1)-(C.6). Denote by $\boldsymbol{\mathcal{H}}_f(\boldsymbol{x})$ the $(d\times d)$ Hessian matrix of $f(\cdot)$ at $\boldsymbol{x}$ and $R(K)=\int\limits_{supp(K)} K^2(\boldsymbol{u})\,d\boldsymbol{u}$. The asymptotic expansion for the MSSE of the local linear estimator is given by

$$\begin{aligned} MSSE &= \frac{\mu_2^2(K)}{4k}\int\limits_{[0,1]^d} tr^2(\boldsymbol{H}\boldsymbol{\mathcal{H}}_f(\boldsymbol{x}))\,d\boldsymbol{x}+\frac{1}{nk|\boldsymbol{H}|^{1/2}}R(K)\\ &\quad +o\left(\frac{tr^2(\boldsymbol{H})}{k}\right)+o\left(\frac{1}{nk|\boldsymbol{H}|^{1/2}}\right). \end{aligned}$$

Remark 5.5
The first term in the expansion for MSSE, which is the leading squared bias term, can be written as $\sum_{i\leq j}\sum_{k<l} c_{ijkl}(\boldsymbol{H})_{ij}(\boldsymbol{H})_{kl}$ (for some constants c_{ijkl} depending on second derivatives of $f(\cdot)$). Balancing each term in this sum with the leading variance term implies that all entries of $\boldsymbol{H}$ are of the same order of magnitude, say $\boldsymbol{H}=h^2\boldsymbol{C}$ where h is a scalar bandwidth parameter and $\boldsymbol{C}$ is a $d\times d$ matrix of

(unknown) constants. The choice $h \sim n^{-\frac{1}{4+d}}$ balances the squared bias part and the variance part in the MSSE. We therefore redefine $\boldsymbol{C}$ so that

$$\boldsymbol{H} = n^{-\frac{2}{4+d}}\boldsymbol{C}.$$

For this choice of $\boldsymbol{H}$,

$$\text{MSSE} = O(n^{-\frac{4}{4+d}}k^{-1})$$

which is the optimal rate assuming the existence of second order partial derivatives of $f(\cdot)$ (Burman (1987a)). A further relevant reference is Cristóbal and Alcalá (1996). They study, in the regression estimation context, the decomposition $\boldsymbol{H} = h^2\boldsymbol{C}$ in greater detail.

For the frequency estimators we have (see Section 3.1) MSSE($\overline{\boldsymbol{P}}$) $= O(n^{-1})$. Therefore, it is easy to see that local linear smoothing becomes beneficial, in terms of faster MSSE convergence rate, for tables with degree of sparseness such that $kn^{-4/(d+4)} \to \infty$. This means that the higher the dimension of the table, the more total number of cells the table needs to have, before smoothing becomes beneficial, and the MSSE convergence rate decreases with the dimension d.

Proof of Theorem 5.1

Use the following decomposition for MSSE :

$$\begin{aligned} \text{MSSE} &= B_I^2 + B_B^2 + V_I + V_B \\ &= \sum_{i \in I}(E\widehat{P}_i - p_i)^2 + \sum_{i \in B}(E\widehat{P}_i - p_i)^2 + \sum_{i \in I}\text{Var}(\widehat{P}_i) + \sum_{i \in B}\text{Var}(\widehat{P}_i). \end{aligned}$$

First we will derive asymptotic expansions for the bias and variance of the local linear estimator. It is clear that

$$E\widehat{P}_i = \frac{1}{k|\boldsymbol{H}|^{1/2}}\sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))p_j.$$

By a one-term Taylor expansion on $f(\cdot)$ and condition (C.1) we have, uniformly in the j-index,

$$p_j = \frac{f(\boldsymbol{x}_j)}{k} + O\left(\frac{1}{k}\left(\frac{1}{k_1^2} + \ldots + \frac{1}{k_d^2}\right)\right). \tag{5.6}$$

For $\boldsymbol{x}_j$'s satisfying $\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i) \in supp(K)$, use Young's form of Taylor's theorem and the fact that $\|\boldsymbol{x}_j - \boldsymbol{x}_i\|^2 = \|\boldsymbol{H}^{1/2}\|^2 \|\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\|^2 \leq R^2 tr(\boldsymbol{H})$, where R is the radius of the support of K, to obtain

$$f(\boldsymbol{x}_j) = f(\boldsymbol{x}_i) + D_f^T(\boldsymbol{x}_i)(\boldsymbol{x}_j - \boldsymbol{x}_i) + \frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{x}_i)^T \boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_i)(\boldsymbol{x}_j - \boldsymbol{x}_i) + o(tr(\boldsymbol{H}))$$

with $D_f(\boldsymbol{x})$ the gradient of $f(\cdot)$ at $\boldsymbol{x}$. Now use the discrete order property (5.3) to get

$$E\widehat{P}_i - p_i =$$

$$\frac{1}{2k^2|\boldsymbol{H}|^{1/2}} \sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))^T \boldsymbol{H}^{1/2} \boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_i) \boldsymbol{H}^{1/2} (\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i$$

$$+ o\left(\frac{tr(\boldsymbol{H})}{k}\right) + O\left(\frac{1}{k}\left(\frac{1}{k_1^2} + \ldots + \frac{1}{k_d^2}\right)\right).$$

Note that

$$(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))^T \boldsymbol{H}^{1/2} \boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_i) \boldsymbol{H}^{1/2} (\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))$$

$$= tr(\boldsymbol{H}^{1/2} \boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_i) \boldsymbol{H}^{1/2} (\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))^T)$$

and conditions (C.4) and (C.6) imply $O\left(\frac{1}{k}\left(\frac{1}{k_1^2} + \ldots + \frac{1}{k_d^2}\right)\right) = o\left(\frac{tr(\boldsymbol{H})}{k}\right)$. Hence, the asymptotic expansion for the bias becomes

$$E\widehat{P}_i - p_i =$$

$$\frac{1}{2k} tr\left\{\frac{\boldsymbol{H}^{1/2} \boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_i) \boldsymbol{H}^{1/2}}{k|\boldsymbol{H}|^{1/2}} \sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))^T\right\}$$

$$+ o\left(\frac{tr(\boldsymbol{H})}{k}\right).$$

For $\boldsymbol{x}_i$ an interior point, using Lemma 5.2(ii), (5.7) reduces to

$$E\widehat{P}_i - p_i = \frac{1}{2k}\mu_2(K) tr(\boldsymbol{H}\boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_i)) + o\left(\frac{tr(\boldsymbol{H})}{k}\right) \tag{5.8}$$

and for boundary points, by Remark 5.4, the order of the bias is

$$E\widehat{P}_i - p_i = \frac{1}{2k} tr(\boldsymbol{H}^{1/2}\boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_i)\boldsymbol{H}^{1/2}O(\boldsymbol{J})) = O\left(\frac{tr(\boldsymbol{H})}{k}\right). \tag{5.9}$$

Note that the order bounds are uniformly in the i-index.
A calculation, similar to that of Lemma 5.1(ii) yields

$$B_I^2 = \frac{1}{4k}\mu_2^2(K) \int\limits_{[0,1]^d} tr^2(\boldsymbol{H}\boldsymbol{\mathcal{H}}_f(\boldsymbol{x}))d\boldsymbol{x} + o\left(\frac{tr^2(\boldsymbol{H})}{k}\right). \tag{5.10}$$

Since $\#B = O\left(k\sqrt{tr(\boldsymbol{H})}\right)$ we have from (5.9)

$$B_B^2 = \frac{1}{4k^2}\sum_{i\in B} O(tr^2(\boldsymbol{H})) = o\left(\frac{tr^2(\boldsymbol{H})}{k}\right). \tag{5.11}$$

To obtain an expression for $\mathrm{Var}(\widehat{P}_i)$ note that for $\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i) \in supp(K)$ we have uniformly in the i and j-index

$$f(\boldsymbol{x}_j) = f(\boldsymbol{x}_i) + O(\|\boldsymbol{x}_j - \boldsymbol{x}_i\|) = f(\boldsymbol{x}_i) + O\left(\sqrt{tr(\boldsymbol{H})}\right).$$

This and (5.6) (remember that $O\left(\frac{1}{k}\left(\frac{1}{k_1^2} + \ldots + \frac{1}{k_d^2}\right)\right) = o\left(\frac{tr(\boldsymbol{H})}{k}\right)$) lead to, uniform in the i-index,

$$\begin{aligned}
&\mathrm{Var}(\widehat{P}_i) \\
&= \frac{1}{nk^2|\boldsymbol{H}|}\left\{\sum_{j=1}^{k} L_i^2(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))p_j - \left(\sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))p_j\right)^2\right\} \\
&= \frac{1}{nk^2|\boldsymbol{H}|}\sum_{j=1}^{k} L_i^2(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))\frac{f(\boldsymbol{x}_i) + O(\sqrt{tr(\boldsymbol{H})})}{k} \\
&\quad -\frac{1}{n}\left\{\frac{1}{k|\boldsymbol{H}|^{1/2}}\sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))\frac{f(\boldsymbol{x}_i) + O(\sqrt{tr(\boldsymbol{H})})}{k}\right\}^2 \\
&= \frac{1}{n}\frac{f(\boldsymbol{x}_i)}{k^2|\boldsymbol{H}|^{1/2}}\frac{1}{k|\boldsymbol{H}|^{1/2}}\sum_{j=1}^{k} L_i^2(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)) + o\left(\frac{1}{nk^2|\boldsymbol{H}|^{1/2}}\right).
\end{aligned} \tag{5.12}$$

For $\boldsymbol{x}_i$ an interior point (5.12) reduces, by (5.5) and Lemma 5.1(i), to

$$\mathrm{Var}(\widehat{P}_i) = \frac{1}{n}\frac{f(\boldsymbol{x}_i)}{k^2|\boldsymbol{H}|^{1/2}}R(K) + o\left(\frac{1}{nk^2|\boldsymbol{H}|^{1/2}}\right). \tag{5.13}$$

This yields, by Lemma 5.1(ii),

$$\begin{aligned} V_I &= \frac{R(K)}{nk|\boldsymbol{H}|^{1/2}}\frac{1}{k}\sum_{i\in I} f(\boldsymbol{x}_i) + o\left(\frac{1}{nk|\boldsymbol{H}|^{1/2}}\right) \\ &= \frac{R(K)}{nk|\boldsymbol{H}|^{1/2}} + o\left(\frac{1}{nk|\boldsymbol{H}|^{1/2}}\right). \end{aligned} \tag{5.14}$$

A similar argument as the one leading to (5.11) results in

$$V_B = o\left(\frac{1}{nk|\boldsymbol{H}|^{1/2}}\right). \tag{5.15}$$

Combining (5.10),(5.11),(5.14) and (5.15) yields the desired expansion for MSSE. ■

5.2 Simulation study

Based on a simulation study we illustrate the benefit of working with general bandwith matrix. For two specific latent densities we consider contingency tables of size 10×10 with $n = 250$. The first one is

$$f(x_1, x_2) = \frac{\alpha\{(\alpha-1)(x_1+x_2-2x_1x_2)+1\}}{\{[1+(\alpha-1)(x_1+x_2)]^2 - 4\alpha(\alpha-1)x_1x_2\}^{3/2}} \tag{5.16}$$

for $0 \leq x_1, x_2 \leq 1$ and zero otherwise. This is the bivariate Plackett density with uniform marginals. We take $\alpha = 10$. See Mardia (1970) for a further discussion on the properties of this contingency-type distribution. The second one is a polynomial density

$$f(x_1, x_2) = L_f\{20x_1^2x_2^2 + x_1^2 + 11x_2^2 + 1\} \tag{5.17}$$

for $0 \leq x_1$, $x_2 \leq 1$ and zero otherwise with L_f a normalizing constant. See Figure 5.1 for a graphical display of the cell probabilities generated from these densities. Note that both densities satisfy (C.1) but not the boundary condition required in Burman (1987a).

In Remark 5.5 we have noted that the bandwidth matrix that minimizes the leading terms in the asymptotic expansion of MSSE is of the form $\boldsymbol{H} = n^{\frac{-2}{4+d}}\boldsymbol{C}$ with $\boldsymbol{C}$ a $d \times d$ matrix of constants. To derive the optimal matrix $\boldsymbol{C}$ first rewrite the leading bias-squared term in Theorem 5.1. Denote $\boldsymbol{\Psi}_f$ the $\frac{d(d+1)}{2} \times \frac{d(d+1)}{2}$ matrix given by

$$\boldsymbol{\Psi}_f = \int\limits_{[0,1]^d} \text{vech}\{2\mathcal{H}_f(\boldsymbol{x}) - dg\mathcal{H}_f(\boldsymbol{x}))\text{vech}\{2\mathcal{H}_f(\boldsymbol{x}) - dg\mathcal{H}_f(\boldsymbol{x})\}^T d\boldsymbol{x}$$

where $\mathcal{H}_f(\boldsymbol{x})$ denotes the $(d \times d)$ Hessian matrix of $f(\cdot)$ at $\boldsymbol{x}$ and dg denotes the diagonal matrix formed by replacing all off-diagonal entries by zeroes. For a $d \times d$ symmetric matrix $\boldsymbol{A}$, vech$(\boldsymbol{A})$ is the $\frac{1}{2}d(d+1)$ column vector created by stacking the columns of $\boldsymbol{A}$, one under the other to form a single column, but with entries above the main diagonal omitted. For example, if $d = 2$

$$\boldsymbol{\Psi}_f = \begin{pmatrix} \Psi_{11} & 2\Psi_1^1 & \Psi_{12} \\ 2\Psi_1^1 & 4\Psi^2 & 2\Psi_2^1 \\ \Psi_{12} & 2\Psi_2^1 & \Psi_{22} \end{pmatrix},$$

where, for $i, j = 1, 2$,

$$\Psi_{ij} = \int_0^1\int_0^1 \frac{\partial^2 f(x_1,x_2)}{\partial x_i^2}\frac{\partial^2 f(x_1,x_2)}{\partial x_j^2}\,dx_1dx_2$$

$$\Psi_i^1 = \int_0^1\int_0^1 \frac{\partial^2 f(x_1,x_2)}{\partial x_i^2}\frac{\partial^2 f(x_1,x_2)}{\partial x_1\partial x_2}\,dx_1dx_2$$

$$\Psi^2 = \int_0^1\int_0^1 \left(\frac{\partial^2 f(x_1,x_2)}{\partial x_1\partial x_2}\right)^2 dx_1dx_2.$$

By direct calculations one can show that (see e.g. Wand (1992))

$$\frac{\mu_2^2(K)}{4k}\int\limits_{[0,1]^d} tr^2(\boldsymbol{H}\mathcal{H}_f(\boldsymbol{y}))d\boldsymbol{y} = \frac{\mu_2^2(K)}{4k}(\text{vech}\boldsymbol{H})^T\boldsymbol{\Psi}_f(\text{vech}\boldsymbol{H}).$$

The optimal matrix $\boldsymbol{C}$ can then be determined by minimizing

$$\frac{\mu_2^2(K)(\mathrm{vech}\boldsymbol{C})^T\boldsymbol{\Psi}_f(\mathrm{vech}\boldsymbol{C})}{4}+\frac{R(K)}{|\boldsymbol{C}|^{1/2}} \tag{5.18}$$

in $\boldsymbol{C}$. In general this minimization can only be performed numerically. Wand (1992) derives the necessary formulae for Newton-Raphson. Significant simplification can be made by taking $\boldsymbol{C}=\mathrm{diag}(C_1,\ldots,C_d)$. This corresponds to bandwidth parametrization $\boldsymbol{H}=\mathrm{diag}(h_1^2,\ldots,h_d^2)$. For $d=2$, the expression (5.18) can be minimized analytically, and the solution is

$$\begin{aligned} C_1 &= \left(\frac{\Psi_{22}^{3/4}R(K)}{\mu_2^2(K)\Psi_{11}^{3/4}(\Psi_{12}+\Psi_{11}^{1/2}\Psi_{22}^{1/2})}\right)^{1/3} \\ C_2 &= \left(\frac{\Psi_{11}}{\Psi_{22}}\right)^{1/2} C_1. \end{aligned} \tag{5.19}$$

If in each direction $i=1,\ldots,d$ the same amount of smoothing is used ($C_1=\ldots=C_d$), i.e., $\boldsymbol{H}=h^2\boldsymbol{I}_d$, then the optimal constant is given by

$$C_1=\left(\frac{dR(K)}{\mu_2^2(K)\sum^*\Psi_{ij}}\right)^{2/(d+4)}$$

where $\sum^*$ is the sum over all integrals Ψ_{ij} in $\boldsymbol{\Psi}_f$ not involving mixed partial derivatives.

We calculated the asymptotically optimal diagonal matrix $(\boldsymbol{C}_D)$, based on formula (5.19), and full matrix $(\boldsymbol{C}_F)$,based on the Newton-Raphson procedure given in Wand (1992), for both densities. See Table 5.1 for the results.

	Plackett	Polynomial
C_F	$\begin{pmatrix} 0.0277 & -0.0054 \\ -0.0054 & 0.0277 \end{pmatrix}$	$\begin{pmatrix} 0.3652 & -0.1666 \\ -0.1666 & 0.1890 \end{pmatrix}$
C_D	$\begin{pmatrix} 0.0233 & 0 \\ 0 & 0.0233 \end{pmatrix}$	$\begin{pmatrix} 0.2058 & 0 \\ 0 & 0.1072 \end{pmatrix}$

Table 5.1: *Asymptotically optimal diagonal matrix $\boldsymbol{C}_D$ and full matrix $\boldsymbol{C}_F$ for the Plackett and polynomial density given in (5.16) and (5.17).*

For the bandwidth matrices $\boldsymbol{H}_D=n^{-1/3}\boldsymbol{C}_D$ and $\boldsymbol{H}_F=n^{-1/3}\boldsymbol{C}_F$ we obtain the exact (non-asymptotic) MSSE$_D$ and MSSE$_F$ and their asymptotic counterparts

AMSSE$_D$ and AMSSE$_F$. The difference in performance can be seen by looking at the ratios AMSSE$_F$/AMSSE$_D$ or MSSE$_F$/MSSE$_D$. From the numerical results in Table 5.2 and the discussion in Remark 5.6 the merit of a full bandwidth matrix over a diagonal one become clear.

	Plackett	Polynomial
AMSSE$_F \times n^{2/3}$	0.043894(0.014631 + 0.029263)	0.005876(0.001959 + 0.003917)
MSSE$_F \times n^{2/3}$	0.050140(0.003008 + 0.047132)	0.013723(0.000798 + 0.012924)
MSSE$^*_F \times n^{2/3}$	0.026324(0.009463 + 0.016861)	0.009114(0.002845 + 0.006268)
AMSSE$_D \times n^{2/3}$	0.051263(0.017088 + 0.034175)	0.008035(0.002678 + 0.005357)
MSSE$_D \times n^{2/3}$	0.055120(0.001713 + 0.053407)	0.015438(0.001115 + 0.014322)
MSSE$^*_D \times n^{2/3}$	0.032745(0.011546 + 0.021200)	0.010926(0.004051 + 0.006875)
AMSSE$_F$/AMSSE$_D$	0.856262	0.731237
MSSE$_F$/MSSE$_D$	0.909654	0.888904
MSSE$^*_F$/MSSE$^*_D$	0.803915	0.834132

Table 5.2: *(Within parentheses we give the contribution of the squared bias (first term) and the variance (second term) separately.)*

A further illustration is given in Figure 5.2. Based on 1000 simulation runs it presents a boxplot of SSE$_F$/SSE$_D$ with SSE$_F$ (resp. SSE$_D$) the sum of squared errors ($\sum_{i=1}^{k}(\widehat{P}_i - p_i)^2$) obtained by using the bandwidth matrix $\boldsymbol{H}_F$ (resp. $\boldsymbol{H}_D$) to calculate the actual values of the $\widehat{P}_i$'s.

Remark 5.6

First note that for the AMSSE in Table II the squared bias and the variance are of the same order of magnitude (as one expects).

The exact MSSE$_F$ and MSSE$_D$ values, obtained by using the bandwidth matrix (full and diagonal) that minimises the asymptotic expression (5.18), are slightly larger. Moreover the squared bias and variance have different order of magnitude. However calculating SSE based on bandwidth matrices that are optimal in an asymptotic sense is meaningful since typical bandwidth selectors (e.g. based on plug-in methods) are also based on the underlying asymptotics. We also computed the full and diagonal bandwidth matrix that minimizes the exact MSSE. The corresponding MSSE$^*_F$ and MSSE$^*_D$ are smaller than MSSE$_F$ and MSSE$_D$; and the squared bias and variance contributions are again in balance. All three error ratios show the benefit of using a full matrix over a diagonal matrix.

Remark 5.7
In the univariate setting there are several methods for choosing $\boldsymbol{C}$ from the data. Most of them can be extended to the multivariate case in some fashion. Wand and Jones (1994) give arguments that suggest that in the nonparametric density estimation setting, the multivariate extension of the plug-in selector of Sheather and Jones (1991) has good theoretical properties for moderate dimensional data. The unknowns in (5.18) are the integrals in the $\boldsymbol{\Psi}_f$ matrix, involving second order partial derivatives of $f(\cdot)$, and these can be replaced by "plug-in" estimates, depending on a so-called pilot bandwidth matrix $\boldsymbol{G}$. Note that Wand and Jones (1994) consider kernel-type estimators for which it is known that boundary problems are present. Since they implicitely make boundary assumptions on the unknown density, this is of no concern for them in their study. In order to define boundary-aware estimators for the unknown functionals in $\boldsymbol{\Psi}_f$, one should extend the ideas developed in Cheng (1996, 1997) to the multi-dimensional setting, which first of all requires an extension of local cubic based estimators for second derivatives. Since a plug-in bandwidth selection rule relies on the knowledge of the asymptotically optimal choice of the pilot bandwidth, first the theory should be investigated before proposals can be implemented. We think, based on similar arguments as given in Wand and Jones (1994), that also in the sparse multinomial or binned density estimation context, the investigation of such an extension of the univariate plug-in estimator is a challenging open problem.

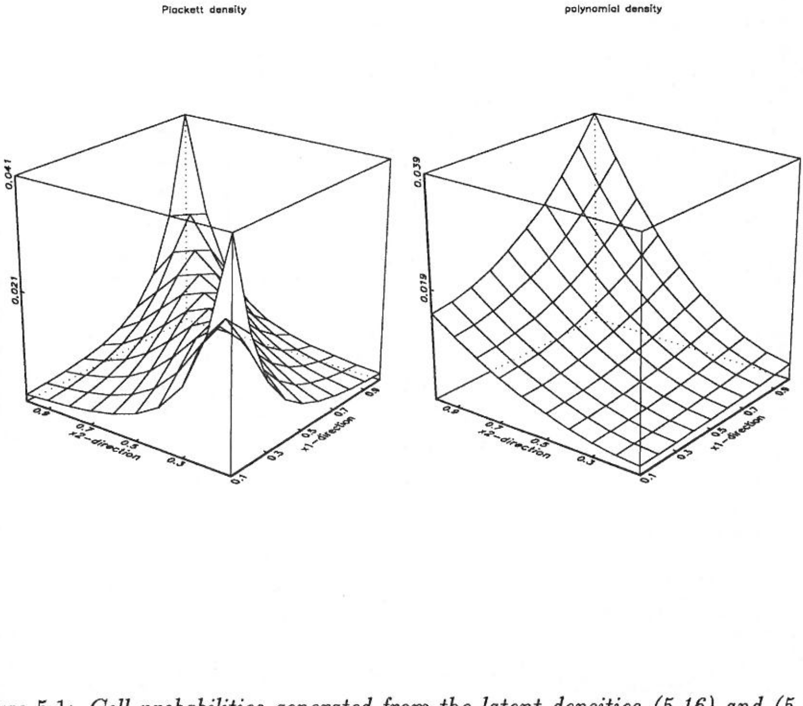

Figure 5.1: *Cell probabilities generated from the latent densities (5.16) and (5.17). Left: Plackett density. Right: Polynomial density.*

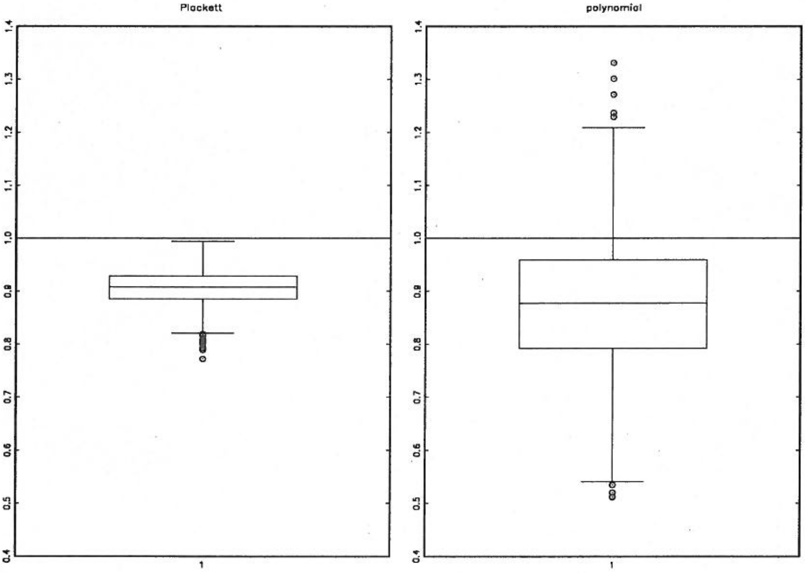

Figure 5.2: *Boxplots of SSE_F/SSE_D with SSE_F (resp. SSE_D) the sum of squared errors obtained by using the bandwidth matrix $\boldsymbol{H}_F$ (resp. $\boldsymbol{H}_D$). Values smaller than 1 indicate that the usage of a full bandwidth matrix is beneficial. Left: Plackett density. Right: Polynomial density.*

5.3 A central limit result for SSE

As noted in Hall (1984) and Chapter 4, the technique used in Chapter 4 to prove the central limit result for $\text{SSE}(\widehat{\boldsymbol{P}}) = \sum_{i=1}^{k}(\widehat{P}_i - p_i)^2$ (Theorem 4.2) can be applied to the multi-dimensional case as well. The generalization to the multi-dimensional local linear estimator is given in the next theorem. We present the central limit theorem for the local linear smoothers based on bandwidth matrices $\boldsymbol{H} = h^2\boldsymbol{C}$. From Remark 5.5 we know that the asymptotically optimal bandwidth matrix has this form. Note that this type of bandwidth matrices is still general, in the sense that it still allows for smoothing along different directions than the coordinate axes. The order of the amount of smoothing is now reduced into one parameter h, which makes the asymptotic investigation slightly easier.

Theorem 5.2
Assume (C.1)–(C.6), $\boldsymbol{H} = h^2\boldsymbol{C}$ *and* $nh^d \to \infty$. *Then*

$$d(n)\left(SSE(\widehat{\boldsymbol{P}}) - MSSE(\widehat{\boldsymbol{P}})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2)$$

where

$$d(n) = \begin{cases} \dfrac{\sqrt{n}k}{h^2} & nh^{4+d} \to \infty \\ n^{\frac{d+8}{2(d+4)}}k & nh^{4+d} \to \lambda \\ nkh^{d/2} & nh^{4+d} \to 0 \end{cases}$$

and

$$\sigma^2 \equiv \sigma^2(f,K) = \begin{cases} 4\sigma_1^2 & nh^{4+d} \to \infty \\ 4\sigma_1^2\lambda^{\frac{4}{4+d}} + 2\sigma_2^2\lambda^{\frac{-d}{4+d}} & nh^{4+d} \to \lambda \\ 2\sigma_2^2 & nh^{4+d} \to 0 \end{cases}$$

with

$$\sigma_1^2 = \frac{\mu_2^2(K)}{4}\left\{\int_{[0,1]^d} tr^2(\boldsymbol{C}\mathcal{H}_f(\boldsymbol{x}))f(\boldsymbol{x})\,d\boldsymbol{x} - \left(\int_{[0,1]^d} tr(\boldsymbol{C}\mathcal{H}_f(\boldsymbol{x}))f(\boldsymbol{x})\,d\boldsymbol{x}\right)^2\right\}$$

and

$$\sigma_2^2 = \frac{1}{|\boldsymbol{C}|^{1/2}}\int_{[0,1]^d} f^2(\boldsymbol{x})\,d\boldsymbol{x}\int_{\mathcal{S}_{0,2R}}\left[\int_{\mathcal{S}_{0,R}} K(\boldsymbol{u})K(\boldsymbol{u}+\boldsymbol{v})\,d\boldsymbol{u}\right]^2 d\boldsymbol{v},$$

with $supp(K) \subset \mathcal{S}_{0,R}$, *which is a sphere in* $I\!R^d$ *with center* $\boldsymbol{0}$ *and radius* R.

Outline of the Proof:
First we repeat some notation from Chapter 4. Denote the triangular arrays $\boldsymbol{Y}_{n\ell} = (Y_{\ell 1}, \ldots, Y_{\ell k})^T$ and $\boldsymbol{X}_{n\ell} = (X_{\ell 1}, \ldots, X_{\ell k})^T$, $\ell = 1, \ldots, n$ where, for $i = 1, \ldots, k$,

$$Y_{\ell i} = \begin{cases} 1 & \text{if the } \ell\text{-th observation is in cell } i \\ 0 & \text{otherwise,} \end{cases}$$

and

$$X_{\ell i} = \frac{1}{k|\boldsymbol{H}|^{1/2}} \sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))(Y_{\ell j} - p_j).$$

For a fixed n, $\boldsymbol{Y}_{n1}, \ldots, \boldsymbol{Y}_{nn}$ are i.i.d., and hence also $\boldsymbol{X}_{n1}, \ldots, \boldsymbol{X}_{nn}$.

An important fact in Chapter 4 is that one term in the decomposition for MSSE (vid., U_n) has a U-statistic structure with symmetric kernel

$$H_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}) = \sum_{i=1}^{k} X_{1i} X_{2i}, \tag{5.20}$$

and that the "U-statistic" is degenerate, i.e., $E\left(H_n(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}) | \boldsymbol{X}_{n1}\right) = 0$.
Also the decomposition for SSE − MSSE ((4.5) and (4.6)) remains valid in the multi-dimensional situation.

Recall from Chapter 4 that Lemmas 4.1–4.5 contain the key results to Theorem 4.2. From the proof of Lemma 4.4 it is clear that the martingale limit result can be used in the multi-dimensional case in exactly the same way as in the one-dimensional case. Therefore, since the remaining main steps in the proofs of the lemmas are equations (4.10),(4.11) and (4.14)–(4.17), we have to generalize these results to the multi-dimensional case. Note that in (4.10) and (4.14) explicit expressions for the leading terms are given; in (4.11), (4.15)–(4.17) order relations are sufficient.

Again we use the shorthand notation given in (4.7), i.e.,

$$\begin{aligned} b_i &= E\widehat{P}_i - p_i \\ S_{ij} &= \frac{1}{k|\boldsymbol{H}|^{1/2}} L_i\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right) \\ D_{j_1 j_2} &= \sum_{i=1}^{k} S_{ij_1} S_{ij_2} \\ \varepsilon_{\ell i} &= Y_{\ell i} - p_i. \end{aligned}$$

Basic, but important, properties for S_{ij} and $D_{j_1j_2}$ are (these are based on arguments explained in Remark 5.2)

$$\begin{aligned}&S_{ij}=0 \text{ for } \|\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\|>R\\ &|S_{ij}|=O((k|\boldsymbol{H}|^{1/2})^{-1})=O((kh^d)^{-1}) \text{ for } \|\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\|\le R\end{aligned}$$

and

$$\begin{aligned}&D_{j_1j_2}=0 \text{ for } \|\boldsymbol{H}^{-1/2}(\boldsymbol{x}_{j_1}-\boldsymbol{x}_{j_2})\|>2R\\ &|D_{j_1j_2}|=O((k|\boldsymbol{H}|^{1/2})^{-1})=O((kh^d)^{-1}) \text{ for } \|\boldsymbol{H}^{-1/2}(\boldsymbol{x}_{j_1}-\boldsymbol{x}_{j_2})\|\le 2R.\end{aligned}$$

Further, for a fixed design point $\boldsymbol{x}_i$, the number of design points $\boldsymbol{x}_j$ that satisfy $\|\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\|\le 2R$ is of the order $O(k|\boldsymbol{H}|^{1/2})$.
First we will derive the generalizations of (4.10) and (4.14), which are

$$E(T_{n1}^2)=\frac{h^4}{k^2}\sigma_1^2+o\left(\frac{h^4}{k^2}\right) \tag{5.21}$$

$$E(H_n^2(\boldsymbol{X}_{n1},\boldsymbol{X}_{n2}))=\frac{1}{k^2h^d}\sigma_2^2+o\left(\frac{1}{k^2h^d}\right). \tag{5.22}$$

To prove (5.21) we start from expression (4.22). Now define the set of "purely" interior points by $I'=\left\{i:\{\boldsymbol{y}:\|\boldsymbol{H}^{-1/2}(\boldsymbol{y}-\boldsymbol{x}_i)\|\le 2R\}\subset[0,1]^d\right\}$ with R the radius of the support of $K(\cdot)$. Since $S_{ij}=0$ for $\|\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\|>R$ we only have to consider those indices i such that $\|\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\|\le R$. If $j\in I'$ such i is an interior index. Since the second derivatives of $f(\cdot)$ are continuous, we have by expression (5.8) for the bias of the local linear estimator

$$\begin{aligned}b_i &= \frac{1}{2k}\mu_2(K)tr(\boldsymbol{H}\boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_j))+o\left(\frac{tr(\boldsymbol{H})}{k}\right)\\ &= \frac{h^2}{2k}\mu_2(K)tr(\boldsymbol{C}\boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_j))+o\left(\frac{h^2}{k}\right).\end{aligned}$$

Since i is an interior index we have by (5.5)

$$S_{ij}=\frac{1}{k|\boldsymbol{H}|^{1/2}}K\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j-\boldsymbol{x}_i)\right)+o\left(\frac{1}{k|\boldsymbol{H}|^{1/2}}\right),$$

such that, by Lemma 5.1(i), for $j \in I'$, $\sum_{i=1}^{k} S_{ij} = 1 + o(1)$. These facts result in

$$\sum_{j \in I'} p_j \left(\sum_{i=1}^{k} b_i S_{ij} \right)^2 - \left(\sum_{j \in I'} p_j \sum_{i=1}^{k} b_i S_{ij} \right)^2 =$$

$$\frac{h^4}{4k^2} \mu_2^2(K) \left\{ \sum_{j \in I'} p_j tr^2(\boldsymbol{C}\boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_j)) - \left(\sum_{j \in I'} p_j tr(\boldsymbol{C}\boldsymbol{\mathcal{H}}_f(\boldsymbol{x}_j)) \right)^2 \right\} + o\left(\frac{h^4}{k^2} \right).$$

An application of Lemma 5.1(ii) and the fact that $p_j = f(\boldsymbol{x}_j)/k + o(k^{-1})$ yield

$$\sum_{j \in I'} p_j \left(\sum_{i=1}^{k} b_i S_{ij} \right)^2 - \left(\sum_{j \in I'} p_j \sum_{i=1}^{k} b_i S_{ij} \right)^2 = \frac{h^4}{4k^2} \mu_2^2(K) \times$$

$$\left\{ \int\limits_{[0,1]^d} f(\boldsymbol{x}) tr^2(\boldsymbol{C}\boldsymbol{\mathcal{H}}_f(\boldsymbol{x}))\, d\boldsymbol{x} - \left(\int\limits_{[0,1]^d} f(\boldsymbol{x}) tr(\boldsymbol{C}\boldsymbol{\mathcal{H}}_f(\boldsymbol{x}))\, d\boldsymbol{x} \right)^2 \right\} + o\left(\frac{h^4}{k^2} \right). \tag{5.23}$$

Since $\#(\{1, \ldots, k\} \backslash I') = O(k\sqrt{tr(\boldsymbol{H})}) = O(kh)$ it is easy to see that

$$\sum_{j \notin I'} p_j \left(\sum_{i=1}^{k} b_i S_{ij} \right)^2 - \left(\sum_{j \notin I'} p_j \sum_{i=1}^{k} b_i S_{ij} \right)^2 = o\left(\frac{h^4}{k^2} \right). \tag{5.24}$$

From the last two expressions and (4.22), (5.21) follows.

To prove (5.22) we start from (4.26), for which the first term of (4.27) has the leading contribution. Since, uniformly in j_1, $p_{j_1} = f(\boldsymbol{x}_{j_1})/k + o(k^{-1})$, and, uniformly in j_1 and j_2 with $\|\boldsymbol{H}^{-1/2}(\boldsymbol{x}_{j_1} - \boldsymbol{x}_{j_2})\| \leq 2R$, $p_{j_2} = f(\boldsymbol{x}_{j_1})/k + o(k^{-1})$, we have, using the properties of $D_{j_1 j_2}$,

$$\sum_{j_1=1}^{k} \sum_{j_2=1}^{k} D_{j_1 j_2}^2 p_{j_1} p_{j_2} = \sum_{j_1=1}^{k} \frac{f^2(\boldsymbol{x}_{j_1})}{k^2} \sum_{j_2=1}^{k} D_{j_1 j_2}^2 + o\left(\frac{1}{k^2 h^d} \right). \tag{5.25}$$

Since $D_{j_1 j_2} = \sum_{i=1}^{k} S_{ij_1} S_{ij_2}$ and $S_{ij} = 0$ for $\|\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\| > R$ we obtain that if j_1 is a purely interior index ($j_1 \in I'$) the indices i that have a contribution to $D_{j_1 j_2}$

are also interior indices. Therefore, by (5.5),

$$D_{j_1j_2} = \frac{1}{(k|\boldsymbol{H}|^{1/2})^2}\sum_{i=1}^{k} K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_i - \boldsymbol{x}_{j_1}))K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_i - \boldsymbol{x}_{j_2})) + o\left(\frac{1}{k|\boldsymbol{H}|^{1/2}}\right).$$

A similar calculation as that of Lemma 5.1(i) yields

$$D_{j_1j_2} = \frac{1}{k|\boldsymbol{H}|^{1/2}}\int\limits_{supp(K)} K(\boldsymbol{u})K(\boldsymbol{u} + \boldsymbol{H}^{-1/2}(\boldsymbol{x}_{j_2} - \boldsymbol{x}_{j_1}))\,d\boldsymbol{u} + o\left(\frac{1}{k|\boldsymbol{H}|^{1/2}}\right). \quad (5.26)$$

An application of Lemma 5.1(i) yields, for $j_1 \in I'$,

$$\begin{aligned}\sum_{j_2=1}^{k} D_{j_1j_2}^2 &= \frac{1}{k|\boldsymbol{H}|^{1/2}}\int\limits_{\mathcal{S}_{0,2R}}\left[\int\limits_{\mathcal{S}_{0,R}} K(\boldsymbol{u})K(\boldsymbol{u}+\boldsymbol{v})\,d\boldsymbol{u}\right]^2 d\boldsymbol{v} + o\left(\frac{1}{k|\boldsymbol{H}|^{1/2}}\right)\\ &= \frac{1}{kh^d|\boldsymbol{C}|^{1/2}}\int\limits_{\mathcal{S}_{0,2R}}\left[\int\limits_{\mathcal{S}_{0,R}} K(\boldsymbol{u})K(\boldsymbol{u}+\boldsymbol{v})\,d\boldsymbol{u}\right]^2 d\boldsymbol{v} + o\left(\frac{1}{kh^d}\right)\end{aligned}$$

such that

$$\sum_{j_1\in I'}\frac{f^2(\boldsymbol{x}_{j_1})}{k^2}\sum_{j_2=1}^{k} D_{j_1j_2}^2 = \frac{1}{k^2h^d}\sigma_2^2 + o\left(\frac{1}{k^2h^d}\right). \quad (5.27)$$

By the basic properties for $D_{j_1j_2}$ we have $\sum_{j_2=1}^{k} D_{j_1j_2}^2 = O((kh^d)^{-1})$ which gives

$$\sum_{j_1\notin I'}\frac{f^2(\boldsymbol{x}_{j_1})}{k^2}\sum_{j_2=1}^{k} D_{j_1j_2}^2 = o\left(\frac{1}{k^2h^d}\right). \quad (5.28)$$

Combine (5.25), (5.27) and (5.28) to see that the first term of (4.27) in the multi-dimensional case becomes

$$\sum_{j_1=1}^{k}\sum_{j_2=1}^{k} D_{j_1j_2}^2 p_{j_1}p_{j_2} = \frac{1}{k^2h^d}\sigma_2^2 + o\left(\frac{1}{k^2h^d}\right).$$

The other terms in (4.26) are all of smaller order, which can be seen through counting the number of indices that make a contribution to the term based on the properties for $D_{j_1j_2}$.

The proofs of (4.11), (4.15)–(4.17) in the one-dimensional case are essentially based on counting how many terms make a contribution to certain sums. The basic properties of S_{ij} and $D_{j_1j_2}$, for which we have given the generalizations to the multi-dimensional case at the beginning of this proof, are important. The generalizations of (4.11), (4.15)–(4.17) to the multi-dimensional case are

$$\begin{aligned} E(T_{n1}^4) &= O\left(\frac{h^8}{k^4}\right) \\ E(H_n^4(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2})) &= O\left(\frac{1}{k^4h^{3d}}\right) \\ E(G_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2})) &= O\left(\frac{1}{k^4h^{d}}\right) \\ E(H_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n1})) &= O\left(\frac{1}{k^2h^{2d}}\right). \end{aligned}$$

As an illustration we show for $E(G_n^2(\boldsymbol{X}_{n1}, \boldsymbol{X}_{n2}))$ how the counting method works for the following specific term appearing in this expectation (see the proof of (4.16) for the notation).

$$\begin{aligned} &\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} A_{i_1i_3}A_{i_2i_4}A_{i_1i_2}A_{i_3i_4} \\ &= \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\sum_{j_3=1}^{k}\sum_{j_4=1}^{k} D_{j_1j_3}D_{j_2j_3}D_{j_1j_4}D_{j_2j_4}p_{j_1}p_{j_2}p_{j_3}p_{j_4}. \end{aligned}$$

The multi-dimensional properties of $D_{j_1j_2}$ imply

$$\begin{aligned} &\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} A_{i_1i_3}A_{i_2i_4}A_{i_1i_2}A_{i_3i_4} \\ &= O\left(k^4h^{3d} \times \frac{1}{(kh^d)^4} \times \frac{1}{k^4}\right) = O\left(\frac{1}{k^4h^d}\right). \end{aligned}$$

All the other order relations are obtained in a similar way. ∎

5.4 Proofs

Proof of Lemma 5.1

(i) Denote $\boldsymbol{H}^{-1/2}(C_j - \boldsymbol{x}_i) = \left\{\boldsymbol{H}^{-1/2}(\boldsymbol{x} - \boldsymbol{x}_i) \ : \ x \in C_j\right\}$ and define

$$\begin{aligned} I_i &= \left\{j \ : \ \boldsymbol{H}^{-1/2}(C_j - \boldsymbol{x}_i) \ \subset \ supp(G)\right\} \\ II_i &= \left\{j \ : j \notin I_i \text{ and } \boldsymbol{H}^{-1/2}(C_j - \boldsymbol{x}_i) \ \cap \ supp(G) \neq \emptyset\right\} \\ III_i &= \left\{j \ : \ \boldsymbol{H}^{-1/2}(C_j - \boldsymbol{x}_i) \ \cap \ supp(G) = \emptyset\right\}. \end{aligned}$$

We have that

$$\begin{aligned} &\left| \frac{1}{k|\boldsymbol{H}|^{1/2}} \sum_{j=1}^{k} G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right) - \frac{1}{|\boldsymbol{H}|^{1/2}} \int\limits_{[0,1]^d} G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\right) dy \right| \\ &= \left| \frac{1}{|\boldsymbol{H}|^{1/2}} \sum_{j=1}^{k} \int\limits_{C_j} \left(G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right) - G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\right)\right) dy \right| \\ &\leq \frac{1}{|\boldsymbol{H}|^{1/2}} \sum_{j=1}^{k} \int\limits_{C_j} \left|G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right) - G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\right)\right| dy \\ &= \frac{1}{|\boldsymbol{H}|^{1/2}} \left(\sum_{j \in I_i} + \sum_{j \in II_i} + \sum_{j \in III_i} \right) \\ &\qquad \int\limits_{C_j} \left|G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right) - G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\right)\right| dy, \end{aligned} \tag{5.29}$$

First note that the sum over III_i has no contribution to (5.29), and that for $j \in II_i$ we don't know whether $\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i) \in supp(G)$. Further, for $j \in I_i$ we have, since $G(\cdot)$ is continuous on a compact support,

$$G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right) - G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\right) = o(1),$$

and for $j \in II_i$ we certainly have

$$G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right) - G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\right) = O(1),$$

where both order bounds are uniformly in the i and j-index. From Remark 5.2 we know that $\#I_i = O\left(k|\boldsymbol{H}|^{1/2}\right)$, such that

$$\frac{1}{|\boldsymbol{H}|^{1/2}} \sum_{j \in I_i} \int_{C_j} \left| G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right) - G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\right) \right| \, dy = o(1).$$

Now we still have to find an order bound for $\#\mathit{II}_i$. The support of $G_{\boldsymbol{H}}(\cdot - \boldsymbol{x}_i)$ can be covered by $O(k|\boldsymbol{H}|^{1/2})$ cells. Moreover the way we cover the d-dimensional unit cube by our cells C_j implies that the projection of the support in each direction can be covered by a number of cells of strict order $O((k|\boldsymbol{H}|^{1/2})^{1/d})$. This finding and the fact that the support of $G_{\boldsymbol{H}}(\cdot - \boldsymbol{x}_i)$ is compact and convex imply that the number of cells needed to cover the boundary of $supp(G_{\boldsymbol{H}}(\cdot - \boldsymbol{x}_i))$ is of the order $O((k|\boldsymbol{H}|^{1/2})^{(d-1)/d})$.
This yields that

$$\frac{1}{|\boldsymbol{H}|^{1/2}} \sum_{j \in \mathit{II}_i} \int_{C_j} \left| G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right) - G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\right) \right| \, dy = O\left(\frac{1}{(k|\boldsymbol{H}|^{1/2})^{1/d}}\right).$$

By condition (C.5), $|\boldsymbol{H}|^{1/2} \propto tr(\boldsymbol{H})^d$, such that we have, by (C.4)

$$\frac{1}{|\boldsymbol{H}|^{1/2}} \sum_{j \in \mathit{II}_i} \int_{C_j} \left| G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\right) - G\left(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\right) \right| \, dy = o(1).$$

(ii) First note that, since $\{1, \ldots, k\} \setminus S = o(k)$ and $g(\cdot)$ bounded,

$$\frac{1}{k} \sum_{i \in S} g(\boldsymbol{x}_i) = \frac{1}{k} \sum_{i=1}^{k} g(\boldsymbol{x}_i) + o(1).$$

Further,

$$\frac{1}{k} \sum_{i=1}^{k} g(\boldsymbol{x}_i) - \int_{[0,1]^d} g(\boldsymbol{x}) \, d\boldsymbol{x} = \sum_{i=1}^{k} \int_{C_i} \left(g(\boldsymbol{x}_i) - g(\boldsymbol{x})\right) \, d\boldsymbol{x}.$$

By continuity of $g(\cdot)$ on $[0,1]^d$, and since, for $\boldsymbol{x} \in C_i$, $\|\boldsymbol{x}_i - \boldsymbol{x}\| \leq k^{-1}$, we have, uniformly in $\boldsymbol{x}$ and $\boldsymbol{x}_i$,

$$g(\boldsymbol{x}_i) - g(\boldsymbol{x}) = o(1),$$

which yields

$$\frac{1}{k}\sum_{i\in S} g(\boldsymbol{x}_i) - \int\limits_{[0,1]^d} g(\boldsymbol{x})\,d\boldsymbol{x} = o(1).$$

■

Proof of Lemma 5.2
(i) An application of Lemma 5.1(i) on each of the submatrices of $\boldsymbol{N}_i$ and $\boldsymbol{M}_i(\boldsymbol{u})$ results, uniformly in the i-index, in

$$\frac{1}{k|\boldsymbol{H}|^{1/2}}\sum_{j=1}^{k} K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)) = \frac{1}{|\boldsymbol{H}|^{1/2}}\int\limits_{[0,1]^d} K(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i))\,d\boldsymbol{y} + o(1)$$

$$\begin{aligned}&\frac{1}{k|\boldsymbol{H}|^{1/2}}\sum_{j=1}^{k} K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\\ &= \frac{1}{|\boldsymbol{H}|^{1/2}}\int\limits_{[0,1]^d} K(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)\,d\boldsymbol{y} + o(\boldsymbol{1})\end{aligned}$$

$$\begin{aligned}&\frac{1}{k|\boldsymbol{H}|^{1/2}}\sum_{j=1}^{k} K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))^T\\ &= \frac{1}{|\boldsymbol{H}|^{1/2}}\int\limits_{[0,1]^d} K(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i))\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)(\boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i))^T\,d\boldsymbol{y} + o(\boldsymbol{J}).\end{aligned}$$

Since for interior points $supp(K_{\boldsymbol{H}}(\cdot - \boldsymbol{x}_i)) \subset [0,1]^d$, the actual region for the integrals is $supp(K_{\boldsymbol{H}}(\cdot - \boldsymbol{x}_i))$. Now use the coordinate transformation $\boldsymbol{z} = \boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i)$ and (C.3) to obtain

$$\boldsymbol{N}_i = \begin{pmatrix} 1 + o(1) & \boldsymbol{0}^T + o(\boldsymbol{1}^T) \\ \boldsymbol{0} + o(\boldsymbol{1}) & \mu_2(K)\boldsymbol{I}_d + o(\boldsymbol{J}) \end{pmatrix}$$

and

$$\boldsymbol{M}_i(\boldsymbol{u}) = \begin{pmatrix} 1 & \boldsymbol{0}^T + o(\boldsymbol{1}^T) \\ \boldsymbol{u} & \mu_2(K)\boldsymbol{I}_d + o(\boldsymbol{J}) \end{pmatrix}.$$

Therefore, for all $\boldsymbol{u}$ and uniform in the i-index, $|\boldsymbol{N}_i| = |\boldsymbol{M}_i(\boldsymbol{u})| + o(1)$ which proves the result.

(ii) By (i) we have $L_i(\boldsymbol{u}) = K(\boldsymbol{u}) + o(1)$, uniformly in i and $\boldsymbol{u}$, and hence we also have,

$$\frac{1}{k|\boldsymbol{H}|^{1/2}} \sum_{j=1}^{k} L_i(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i))^T$$
$$= \mu_2(K)\boldsymbol{I}_d + o(\boldsymbol{J}).$$

■

Proof of (5.4)

The boundary region is defined as

$$B = \{i : supp(K_{\boldsymbol{H}}(\cdot - \boldsymbol{x}_i)) \not\subset [0,1]^d\}$$

with

$$\begin{aligned} supp(K_{\boldsymbol{H}}(\cdot - \boldsymbol{x}_i)) &= \{\boldsymbol{y} : \boldsymbol{H}^{-1/2}(\boldsymbol{y} - \boldsymbol{x}_i) \in supp(K)\} \\ &= \{\boldsymbol{y} : (\boldsymbol{y} - \boldsymbol{x}_i) \in \boldsymbol{H}^{1/2} supp(K)\}. \end{aligned}$$

Since the kernel function $K(\cdot)$ satisfies (C.2) and (C.3) we know $supp(K) \subset \mathcal{S}_{0,R}$, a sphere with center $\boldsymbol{0}$ and radius R. Then $\boldsymbol{H}^{1/2} supp(K) \subset \mathcal{E}_{0,H}$, an ellipsoid with center $\boldsymbol{0}$, with the direction of the axis according to the eigenvectors of the matrix $\boldsymbol{H}^{1/2}$ and the length of the axis $O(\lambda_{\max})$, where $\lambda_{\max}$ is the largest eigenvalue of $\boldsymbol{H}^{1/2}$. On his turn the ellipsoid $\mathcal{E}_{0,H} \subset \mathcal{C}_{0,H}$, a cube with center $\boldsymbol{0}$, axis along the coordinate axis with length $O(\lambda_{\max}) = O(\sqrt{tr(\boldsymbol{H})})$. All these relations imply

$$B \subset \{i \ : \ \boldsymbol{x}_i + \mathcal{C}_{0,H} \not\subset [0,1]^d\}.$$

Now it is easy to see that

$$\#\{i \ : \ \boldsymbol{x}_i + \mathcal{C}_{0,H} \not\subset [0,1]^d\} = O(k\sqrt{tr(\boldsymbol{H})})$$

which proves (5.4). ■

Proof of Lemma 5.3

From the definition of $L_i(u)$, and the fact that $K(\cdot)$ is continuous on a compact support, it suffices to show that $|\boldsymbol{M}_i(\boldsymbol{u})|$ is bounded uniform in the i-index for bounded $\boldsymbol{u}$ and that $|\boldsymbol{N}_i|$ is uniformly bounded away from zero. The uniform boundedness of $|\boldsymbol{M}_i(u)|$ for bounded $\boldsymbol{u}$ follows immediately from Remark 5.2 and the boundedness of $K(\cdot)$.

We will now show that under conditions (C.4)–(C.6) the matrices $\boldsymbol{N}_i$ are positive definite. To see this, let $\boldsymbol{z}^T = (z_1\ \boldsymbol{z}_2^T) \neq (0\ \boldsymbol{0}^T)$ be a $1 \times (d+1)$-vector. By direct calculation we have

$$\boldsymbol{z}^T \boldsymbol{N}_i \boldsymbol{z} = \\ \frac{1}{k|\boldsymbol{H}|^{1/2}} \sum_{j=1}^{k} \left((z_1\ \boldsymbol{z}_2^T) \begin{pmatrix} 1 \\ \boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i) \end{pmatrix} \right)^2 K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)). \tag{5.30}$$

For each $\boldsymbol{z} \neq \boldsymbol{0}$, the function

$$g_{(}\boldsymbol{u}) = (z_1\ \boldsymbol{z}_2^T) \begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix}$$

can only be zero on a $(d-1)$-hyperplane in $\mathbb{R}^d$. Therefore, as soon as we have $d+1$ indices j such that the points $\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)$ do not span a $(d-1)$-hyperplane and such that $K(\boldsymbol{H}^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)) \neq 0$, there is at least one index j that has a positive contribution to (5.30). Since $K(\cdot)$ has a d-dimensional support and by Remark 5.2 and conditions (C.4)–(C.6) we easily find $d+1$ such indices j, when k is large enough.

Therefore, $\boldsymbol{z}^T \boldsymbol{N}_i \boldsymbol{z}$ is strictly positive for all $\boldsymbol{z} \neq \boldsymbol{0}$, and hence the matrix $\boldsymbol{N}_i$ has a strictly positive determinant.

To show that the bound is uniformly in the i-index, we can use similar arguments as in the proof of Lemma 1.2. ■

Chapter 6

Exact double smoothing bandwidth selection

The goal of this chapter is to develop a heuristic method for selecting the bandwidth in local polynomial smoothing of ordered multinomial data. In our proposal we do not restrict to a global bandwidth, but we allow the use of local and partially local bandwidths. To explain the ideas behind our selector, we follow the recommendations of Hall, Marron and Titterington (1995) and focus on partial local smoothing, over subsets of cells, rather than "purely" local smoothing where a different bandwidth is chosen for each cell. In order to have good finite sample performance, we want to rely on asymptotic theory as less as possible. Therefore, we aim to work with finite sample estimates of risk measures. Our resultant approach can be thought of as being a non-asymptotic version of the double smoothing idea used, for example, by Müller (1985), Staniswalis (1989) and Härdle, Hall and Marron (1992).

In Section 6.1 we explain the ideas behind our proposal, and present the algorithm. In Section 6.2 we illustrate our method on a real data set. We also present a simulation study, in which we compare the performance of our newly proposed bandwidth selector to some existing bandwidth selectors.

6.1 Bandwidth selection strategy

The local polynomial estimator for the cell probabilities defined in Chapter 1 depends on the bandwidth h. Since in this chapter the bandwidth is of major interest, we stress this dependence by writing $\widehat{P}_i(h)$ instead of $\widehat{P}_i$. In matrix notation the vector of local polynomial smoothers can be written as $\widehat{\boldsymbol{P}}(h) = \boldsymbol{S}_h\overline{\boldsymbol{P}}$, where $\boldsymbol{S}_h$ is the $k \times k$

matrix, with $(\boldsymbol{S}_h)_{ij} = S_{ij}$ defined in Chapter 4.

For a subset A of the set of indices $\{1, \ldots, k\}$ we let $\widehat{\boldsymbol{P}}^A(h)$ denote the vector of estimates for the cell probabilities with index in A. We may write $\widehat{\boldsymbol{P}}^A(h) = \boldsymbol{I}_A \widehat{\boldsymbol{P}}(h)$ where $\boldsymbol{I}_A$ is an appropriate $\#A \times k$ matrix of zeros and ones, e.g., for $A = \{2, 3, 4\}$

$$\boldsymbol{I}_A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & 0 & \ldots & 0 \end{pmatrix}.$$

Also, let $\boldsymbol{S}_h^A = \boldsymbol{I}_A S_h$ so that $\widehat{\boldsymbol{P}}^A(h) = \boldsymbol{S}_h^A \overline{\boldsymbol{P}}$, i.e., $\boldsymbol{S}_h^A$ is the submatrix of $\boldsymbol{S}_h$, which contains those rows of $\boldsymbol{S}_h$ with index in the set A.

Let $\mathcal{A} = \{A_1, \ldots, A_r\}$ be a partition of the set of all indices $\{1, \ldots, k\}$. The idea of partially local bandwidth selection is to determine r bandwidths $h^{A_1}, \ldots, h^{A_r}$, all of which are optimal for their corresponding set A_i. Trivial partitions are $\mathcal{A} = \{\{1, \ldots, k\}\}$, the case of a global bandwidth, and $\mathcal{A} = \{\{1\}, \ldots, \{k\}\}$, the case of purely local bandwidths. All other partitions of $\{1, \ldots, k\}$ are examples of partially local bandwidths.

We focus on selecting the bandwidth for estimating $\boldsymbol{p}^A = \boldsymbol{I}_A \boldsymbol{p}$, for some fixed set A. The performance of the estimator $\widehat{\boldsymbol{P}}^A(h)$ is measured by the mean sum of squared errors over A

$$\text{MSSE}_A(h, \boldsymbol{p}) = \sum_{i \in A} E\left(\widehat{P}_i(h) - p_i\right)^2.$$

With respect to this risk measure, the optimal bandwidth is

$$h_0^A = \operatorname*{argmin}_{h>0}\{\text{MSSE}_A(h, \boldsymbol{p})\}.$$

We want to propose a suitable estimator for this unknown optimal bandwidth h_0^A.

In matrix notation we can write,

$$\text{MSSE}_A(h, \boldsymbol{p}) = \|(\boldsymbol{S}_h^A - \boldsymbol{I}_A)\boldsymbol{p}\|^2 - n^{-1}\|\boldsymbol{S}_h^A \boldsymbol{p}\|^2 + n^{-1}\text{tr}\{\boldsymbol{S}_h^A \text{diag}(\boldsymbol{p})(\boldsymbol{S}_h^A)^T\} \qquad (6.1)$$

where $\|v\| = (v^T v)^{1/2}$ denotes the norm of the vector v and $\text{diag}(\boldsymbol{p})$ the $k \times k$ matrix with $p_1, \ldots, p_k$ on the diagonal. See Section 6.3 for the proof. The expression (6.1) depends on the unknown vector of cell probabilities $\boldsymbol{p}$, but it gives an idea how to define an estimator for it. A double smoothing bandwidth selection strategy involves replacing the unknown $\boldsymbol{p}$ in (6.1) by a "pilot" estimate $\boldsymbol{S}_{g^A}\overline{\boldsymbol{P}}$ where g^A is another bandwidth, which we will refer to as the pilot bandwidth. Through the estimator $\mathrm{MSSE_A}(h, \boldsymbol{S}_{g^A}\overline{\boldsymbol{P}})$ for (6.1) we then define an estimator for h_0^A as

$$\hat{h}_{g^A}^A = \underset{h>0}{\operatorname{argmin}}\, \mathrm{MSSE_A}(h, \boldsymbol{S}_{g^A}\overline{\boldsymbol{P}}).$$

The main problem now becomes how to choose the pilot bandwidth g^A. The recent literature has seen a good deal of theory on the choice of the pilot bandwidth in double smoothing strategies, particularly in the density estimation context. See, for example, Jones, Marron and Park (1991), Hall, Marron and Park (1992), Härdle, Hall and Marron (1992) and Park and Marron (1992). In each case, attention has focused on the asymptotic distribution of the relative error $(\hat{h}_{g^A}^A - h_0^A)/h_0^A$, with the choice of g^A aimed at optimising the rate of convergence to a limiting normal distribution.

We want to follow another approach to select the pilot bandwidth, since we have the feeling that the asymptotic theory masks boundary effects. To illustrate where our feeling is based on, we present Figure 6.1. To construct this figure we consider a contingency table with 100 cells, where the cell probabilities are generated by latent density $f_E(u) = 5(1 - e^{-5})^{-1}e^{-5u}\mathbb{1}\{0 \leq u \leq 1\}$. This function (divided by 5) is shown by the dotted curve. For the partition of $\{1, \ldots, 100\}$ we consider subsets A_i of size 5, $i = 1, \ldots, 20$. The circles on the figure correspond to the exact mean sum of squared errors optimal bandwidths (with formula (6.1)). In the interior, as one moves from the right to the left in Figure 6.1 it is seen that the optimal bandwidth decreases. This is consistent with the asymptotic idea that the curvature is increasing and therefore a smaller bandwidth is optimal. However, near the boundary itself, the optimal amount of smoothing increases – reflecting the fact that near the boundary the variance of a local linear estimator increases. This indicates that the asymptotic approximation in the boundary region is not as accurate as in the interior region. It also indicates that one may wish to allow the bandwidth to vary across the cells since a bandwidth chosen for good performance in the interior is not necessarily good for estimation close to the boundary, despite the automatic boundary adjustments of local lines.

A natural alternative for selecting g^A is to use "exact" risk ideas as a guideline. One can define the optimal g^A to be the one minimising the mean squared distance between $\hat{h}_{g^A}^A$ and h_0^A

$$g_0^A = \underset{g^A>0}{\operatorname{argmin}}\, E(\hat{h}_{g^A}^A - h_0^A)^2 = \underset{g^A>0}{\operatorname{argmin}}\, \mathrm{MSE}(\hat{h}_{g^A}^A)$$

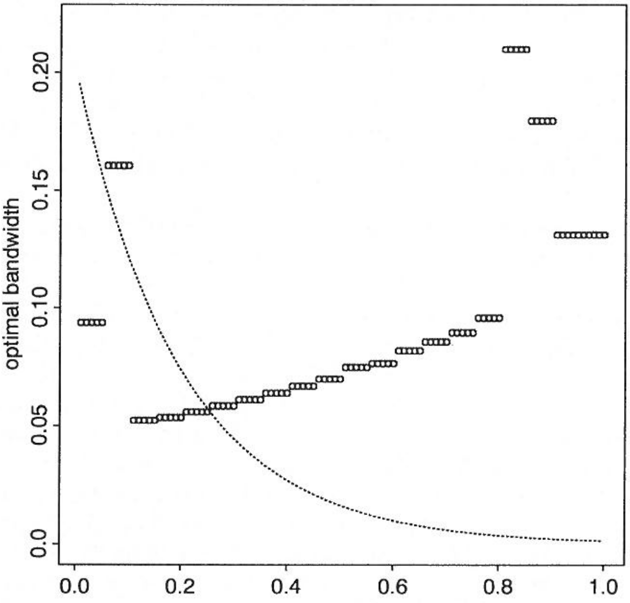

Figure 6.1: *Optimal bandwidths with respect to mean summed squared error over subsets of cells of size 5 for a table with 100 cells and cell probabilities generated by* $f_E(\cdot)$.

and use this to guide the choice of g^A. Unfortunately $\hat{h}^A_{g^A} - h^A_0$ does not have an explicit form which makes its MSE intractable. A way around this is to use the approximation

$$\hat{h}^A_{g^A} - h^A_0 \simeq -\mathrm{MSSE}^{[1]}_A(h^A_0, \boldsymbol{S}_{g^A}\overline{\boldsymbol{P}})/\mathrm{MSSE}^{[2]}_A(h^A_0, \boldsymbol{p}) \tag{6.2}$$

where $\mathrm{MSSE}^{[i]}_A(h, \boldsymbol{p}) = (\partial^i/\partial h^i)\mathrm{MSSE}_{\mathrm{A}}(h, \boldsymbol{p})$, which is based on a Taylor expansion of $\mathrm{MSSE}^{[1]}_A(h, \boldsymbol{p})$ about h^A_0. This, in turn, leads to the following approximation to g^A_0 :

$$\begin{aligned} \widetilde{g}^A_0 &= \operatorname*{argmin}_{g>0} E\left((-\mathrm{MSSE}^{[1]}_A(h^A_0, \boldsymbol{S}_g\overline{\boldsymbol{P}})/\mathrm{MSSE}^{[2]}_A(h^A_0, \boldsymbol{p}))^2\right) \\ &= \operatorname*{argmin}_{g>0} L_A(g, h^A_0) \end{aligned}$$

where

$$L_A(g, h) = E\left((\mathrm{MSSE}^{[1]}_A(h, \boldsymbol{S}_g\overline{\boldsymbol{P}}))^2\right).$$

The advantage of working with $\widetilde{g}^A_0$ is that $L_A(g, h)$ admits the following approximation (see Section 6.3) :

$$\begin{aligned} L_A(g, h) \simeq\; & \left((1 - n^{-1})\boldsymbol{p}^T \boldsymbol{D}^A_{gh}\boldsymbol{p} + n^{-1}\{\mathrm{diagonal}(\boldsymbol{D}^A_{gh})\}^T\boldsymbol{p}\right)^2 \\ & + 4n^{-1}\mathrm{tr}\left(\boldsymbol{S}'^A_h\,\mathrm{diag}(\boldsymbol{p})(\boldsymbol{S}^A_h)^T\right)\Big[(1 - n^{-1})\boldsymbol{p}^T\boldsymbol{D}^A_{gh}\boldsymbol{p} \\ & + n^{-1}\left(\mathrm{diagonal}(\boldsymbol{D}^A_{gh})\right)^T\boldsymbol{p}\Big] + \mathrm{Var}(\overline{\boldsymbol{P}}^T\boldsymbol{D}^A_{gh}\overline{\boldsymbol{P}}) \end{aligned} \tag{6.3}$$

where

$$\boldsymbol{D}^A_{gh} = \boldsymbol{S}^T_g\{(\boldsymbol{S}'^A_h)^T(\boldsymbol{S}^A_h - \boldsymbol{I}_A) + (\boldsymbol{S}^A_h - \boldsymbol{I}_A)^T\boldsymbol{S}'^A_h\}\boldsymbol{S}_g,$$

$(\boldsymbol{S}'^A_h)_{ij} = (\partial/\partial h)(\boldsymbol{S}^A_h)_{ij}$ and $\mathrm{diagonal}(\boldsymbol{D}^A_{gh})$ denotes the column vector containing the diagonal entries of the $k \times k$ matrix $\boldsymbol{D}^A_{gh}$.

Not surprisingly $L_A(g, h)$ depends on the unknown probability vector $\boldsymbol{p}$ and, in particular, $\widetilde{g}^A_0$ depends on h^A_0 – the quantity that we are aiming to estimate in the first place. One could again replace $\boldsymbol{p}$ by a pilot estimate based on yet another bandwidth, but this would lead to further bandwidth selection problems. A strategy to overcome this problem, and to make the approach workable in practice, is to use some initial estimate for $\boldsymbol{p}$. In our simulations (see Section 6.2) we used the Bayesian regression spline smoother of Smith and Kohn (1996) as initial estimator.

We refer to our bandwidth selection method as exact double smoothing (EDS) since the idea is based on "exact" (i.e., non-asymptotic) expressions at both smoothing stages. A full description of the algorithm is :

Step 1 Find $\widehat{\boldsymbol{P}}_{\text{init}}$, an initial estimate for $\boldsymbol{p}$.

Step 2 Set up a partition $\mathcal{A} = \{A_1, \ldots, A_r\}$ of the cell indices $\{1, \ldots, k\}$.

Step 3 For each $A \in \mathcal{A}$:

(i) Find $\hat{h}^A_{\text{init}} = \text{argmin}_{h>0}\, \text{MSSE}_{\text{A}}(h, \widehat{\boldsymbol{P}}_{\text{init}})$.

(ii) Find $\hat{g}^A = \text{argmin}_{g>0}\, L_A(g, \hat{h}^A_{\text{init}})$, where $\widehat{\boldsymbol{P}}_{\text{init}}$ is used as $\boldsymbol{p}$ in (6.3).

(iii) Estimate $\boldsymbol{p}$ by $\widehat{\boldsymbol{P}}_{\hat{g}^A}$.

(iv) The selected bandwidth for subset A is $\hat{h}^A = \underset{h>0}{\text{argmin}}\, \text{MSSE}_{\text{A}}(h, \widehat{\boldsymbol{P}}_{\hat{g}^A})$.

Step 4 If $r > 1$ then fit a natural cubic spline to the pairs

$$(1, \ln(\hat{h}^{A_1})), (\kappa_2, \ln(\hat{h}^{A_2})), \ldots, (\kappa_{r-1}, \ln(\hat{h}^{A_{r-1}})), (k, \ln(\hat{h}^{A_r}))$$

where κ_i is the mean of the indices in A_i. The values of the exponentiation of the spline are then used to give bandwidths for each of the k cells.

The final step overcomes the problem of non-smooth pictures at the change-over from one interval in the partition to the next. The spline is fit to the logarithms of the $\hat{h}^{A_i}$ and then exponentiated to ensure that the final bandwidths are positive.

6.2 Practical performance

In this section we demonstrate the performance of the EDS selection method for local linear smoothing of multinomial data. Figure 6.2 shows the results of applying the method to the mine data described in Section 1.1. The estimated bandwidth function is shown in Figure 6.2a, and is based on applying EDS to 5 partitions of size 11. Note that the method chooses smaller bandwidths near the boundaries, presumably because of the higher amount of curvature there. The resulting estimator is shown in Figure 6.2b. A comparison with Figure 3.5 shows that there is little difference between the two estimates in this case, so it appears that a global bandwidth may be sufficient for these data.

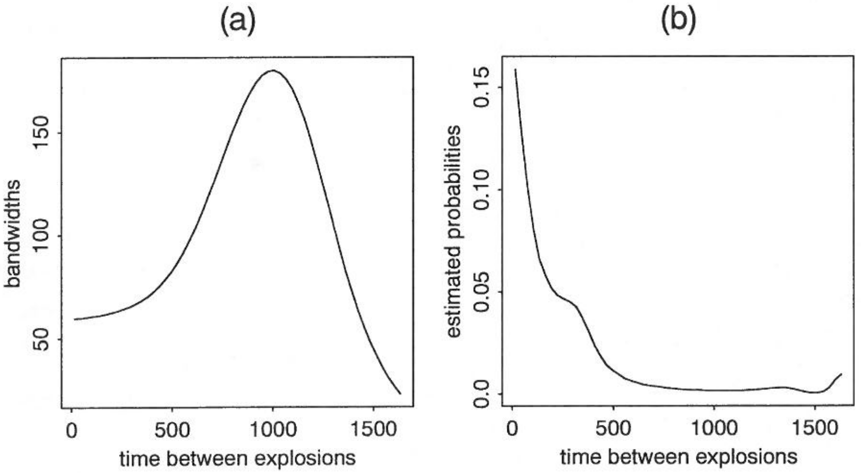

Figure 6.2: *(a) Estimated bandwidth function using the exact double smoothing algorithm and (b) resulting probability estimates for the mine data.*

We conducted a simulation study to compare the global version of our bandwidth selector to existing selection rules. In first instance we considered cross-validatory (CV) rules. Recall from Section 3.3 that the least-squares cross-validation method is based on a "leave-one-out" estimator. We implemented two versions for CV. The first one is based on treating $(x_1, \overline{\boldsymbol{P}}_1), \ldots, (x_k, \overline{\boldsymbol{P}}_k)$ as an ordinary regression-type data set. Leaving an observation out corresponds in this case to leaving out a cell $(x_i, \overline{\boldsymbol{P}}_i)$. The second version uses the fact that we are dealing with multinomial data. This version of cross-validation is described previously in Section 3.3. Leaving an observation out means working with $(N_1/(n-1), \ldots, (N_i-1)/(n-1), \ldots, N_k/(n-1))$. The two versions are denoted by $\mathrm{CV}_{\mathrm{reg}}$ and $\mathrm{CV}_{\mathrm{mnl}}$ respectively.

We also included the direct plug-in (DPI) bandwidth selector of Ruppert, Sheather and Wand (1995) in the simulation study. This selection method is developed for the use of local polynomial smoothing and uses the double smoothing strategy. At both smoothing stages, the choice of the bandwidth is guided by asymptotic considerations. The main idea behind this method is pointed out in Section 3.3. There is a difference between the algorithm presented in Section 3.3 and this DPI-bandwidth selector. Step 1 of the algorithm in Section 3.3 uses a normal reference rule to estimate θ_{24}, a quantity based on second and fourth derivatives of the unknown function $f(\cdot)$. This is a usual strategy in the density estimation context, but there is an alternative to this step in the regression estimation context, for which the DPI-selector of Ruppert, Sheather and Wand (1995) is designed. An alternative to Step 1 is to use a parametric fit at the "quick and simple" estimation step. Ruppert, Sheather and Wand (1995) propose to use a so-called "blocked" quartic fit, where the number of blocks are chosen according to a Mallow's C_p criterium. In the regression context, also the variability of the data plays a role in the asymptotically optimal bandwidth, such that the direct plug-in algorithm in that context is somewhat more elaborate (see Ruppert, Sheather and Wand (1995) for details). Note that we used the DPI method as if our multinomial problem is an ordinary regression problem.

Our intention with this simulation study is to start a first investigation of the EDS bandwidth selector. Therefore we only include two existing popular bandwidth selection methods in the simulation study. One being the cross-validatory rule, for which it is known that its behavior is somewhat disappointing, and secondly a high-level sophisticated bandwidth selector, which has in the existing literature the reputation of being quite reliable. We have chosen for the DPI bandwidth selector of Ruppert, Sheather and Wand (1995) because of practical considerations.

To generate the cell probabilities we considered the following latent densities :

(i) the exponential-like density $f_E(u) = 5(1-e^{-5})^{-1} e^{-5u} \mathbb{1}\{0 \leq u \leq 1\}$

(ii) the Beta(0.5,0.5) density

(iii) the uniform density on [0,1].

Sparse tables are generated with number of cells $k = 50$ and sample sizes $n = 50$, $n = 100$, $n = 250$. The number of replications in each simulation was 500. The normal density truncated on $[-4, 4]$ was used as kernel.

The difficulties related to the first two latent densities are the very high boundary probabilities. Because of these high boundary probabilities, we believe that for the beta latent density the setting with $k = n = 50$ is too difficult to get reasonable answers, such that we have not started simulations in this setting. The reason to include the uniform density is that the optimal bandwidth is infinity, i.e., ordinary least squares performs better than local linear regression. All bandwidth selection strategies, except DPI, used in the simulations need a minimization step. This is done by using grid search on a logarithmically-equally-spaced grid around h_0. In view of the last setting also bandwidth zero, i.e., frequency estimators, and bandwidth infinity, i.e., least squares regression, are appended to the grid.

A graphical summary of the results is given in Figure 6.3. The plots show kernel density estimates of $\ln(\hat{h}) - \ln(h_{opt})$ for each rule. The solid line represents the EDS-selector, the long dashed line the DPI-selector, the dotted line the $\mathrm{CV}_{\mathrm{mnl}}$ and the short dashed line the $\mathrm{CV}_{\mathrm{reg}}$-selector. Figure 6.3a-c are the results for the different sparseness settings for the exponential latent density, and Figure 6.3d shows the result for the beta latent density when $n = 5k$.

Numerical results of the final cell probability estimators are shown in Tables 6.1–6.4. Averages, medians and standard deviations of each of the sum of squared errors (SSE) are given. We also include the results of the Bayesian regression spline smoother (BR) which is used as the initial estimator in Step 1 of the EDS algorithm.

In order to present a kind of classification of the different estimators, on basis of SSE performance, we performed paired Wilcoxon tests. For each simulation run we retained the SSE value for each of the 5 estimation procedures considered in the simulation study. A paired Wilcoxon test is then performed on the 500 paired SSE values from 2 estimation procedures, to determine whether the median SSE's were significantly different. Estimators shared the same SSE ranking when the test showed no difference at the 0.5% level (which is an adjusted Bonferroni significance level). The results of these tests are presented in Table 6.4.

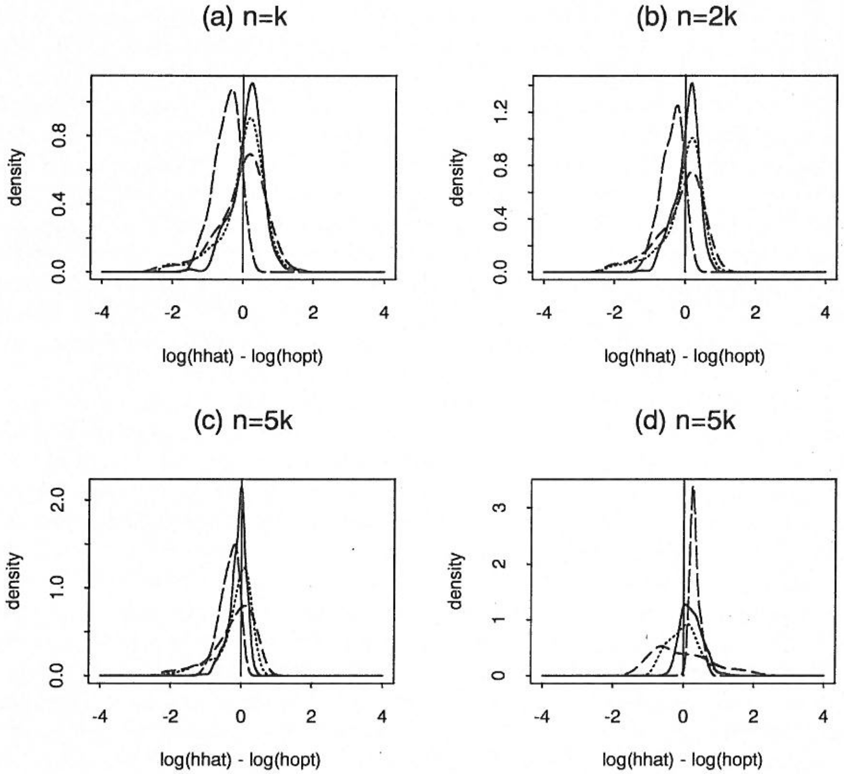

Figure 6.3: *Kernel density estimates of* $\log(\hat{h}) - \log(h_{opt})$ *for EDS-selector (solid line), DPI-selector (long dashed line),* CV_{mnl}*-selector (dotted line) and* CV_{reg}*-selector (short dashed line). (a)-(c) Exponential latent density. (d) Beta latent density.*

	$n = 50$	$n = 100$	$n = 250$
$\mathrm{SSE}_{\mathrm{EDS}}$	2.136e-3[a] (2.483e-3)[b]	1.159e-3 (1.130e-3)	5.342e-4 (4.721e-4)
	1.364e-3[c]	0.846e-3	3.906e-4
$\mathrm{SSE}_{\mathrm{CV}_{\mathrm{reg}}}$	2.713e-3 (3.523e-3)	1.522e-3 (1.637e-3)	7.244e-4 (6.829e-4)
	1.644e-3	1.022e-3	5.112e-4
$\mathrm{SSE}_{\mathrm{CV}_{\mathrm{mnl}}}$	2.615e-3 (3.296e-3)	1.536e-3 (1.626e-3)	6.938e-4 (6.851e-4)
	1.498e-3	0.977e-3	0.467e-3
$\mathrm{SSE}_{\mathrm{DPI}}$	2.286e-3 (3.008e-3)	1.2e-3 (1.294e-3)	5.571e-4 (4.98e-4)
	1.337e-3	0.794e-3	4.162e-4
$\mathrm{SSE}_{\mathrm{BR}}$	3.194e-3 (3.908e-3)	1.716e-3 (1.843e-3)	7.116e-4 (7.615e-4)
	1.908e-3	1.053e-3	4.633e-4

Table 6.1: *Numerical summary of the simulation study for the exponential latent density. Average*[a] *(standard deviation)*[b]*, and median*[c] *of SSE for each strategy.*

	$n = 100$	$n = 250$
$\mathrm{SSE}_{\mathrm{EDS}}$	4.298e-3[a] (1.827e-3)[b]	2.212e-3 (0.851e-3)
	3.934e-3[c]	2.065e-3
$\mathrm{SSE}_{\mathrm{CV}_{\mathrm{reg}}}$	5.143e-3 (2.166e-3)	2.723e-3 (1.218e-3)
	4.780e-3	2.406e-3
$\mathrm{SSE}_{\mathrm{CV}_{\mathrm{mnl}}}$	4.396e-3 (2.017e-3)	2.222e-3 (0.862e-3)
	3.950e-3	2.055e-3
$\mathrm{SSE}_{\mathrm{DPI}}$	3.75e-3 (1.637e-3)	2.15e-3 (0.722e-3)
	3.409e-3	2.05e-3
$\mathrm{SSE}_{\mathrm{BR}}$	4.515e-3 (2.152e-3)	1.85e-3 (1.076e-3)
	4.169e-3	1.597e-3

Table 6.2: *Numerical summary of the simulation study for the beta latent density. Average*[a] *(standard deviation)*[b]*, and median*[c] *of SSE for each strategy.*

It is clear from both Figure 6.3 and the tables that EDS exhibits good overall performance and offers significant improvement over the cross-validatory rules. From Figure 6.3 and Table 6.4 we conclude that the bandwidth selectors DPI and EDS have comparable behavior for the exponential latent density. For the uniform case DPI seems to perform rather poorly. Note that EDS is sometimes bettered by its initial estimator BR. Since EDS behaves rather constant over the three different latent density settings, while BR performs rather inconstant, this does not concern us too much.

	$n = 50$	$n = 100$	$n = 250$
$\mathrm{SSE}_{\mathrm{EDS}}$	3.979e-4[a] (5.857e-4)[b]	2.118e-4 (3.008e-4)	8.319e-5 (1.100e-4)
	1.722e-4[c]	1.055e-4	4.126e-5
$\mathrm{SSE}_{\mathrm{CV}_{\mathrm{reg}}}$	4.111e-4 (5.836e-4)	2.19e-4 (3.011e-4)	8.616e-5 (1.094e-4)
	1.829e-4	1.161e-4	4.605e-5
$\mathrm{SSE}_{\mathrm{CV}_{\mathrm{mnl}}}$	4.870e-4 (1.796e-4)	2.188e-4 (3.01e-4)	8.609e-5 (1.093e-4)
	1.829e-4	1.152e-4	4.605e-5
$\mathrm{SSE}_{\mathrm{DPI}}$	1.608e-3 (1.404e-3)	8.712e-4 (6.759e-4)	3.334e-4 (2.46e-4)
	1.235e-3	6.988e-4	2.822e-4
$\mathrm{SSE}_{\mathrm{BR}}$	0.609e-3 (1.3e-3)	2.439e-4 (4.638e-4)	7.389e-5 (1.392e-4)
	0.137e-3	0.625e-4	1.701e-5

Table 6.3: *Numerical summary of the simulation study for the uniform latent density. Average*[a] *(standard deviation)*[b], *and median*[c] *of SSE for each strategy.*

Uniform	$n = k$	EDS	BR	CV$_{\text{mnl}}$	CV$_{\text{reg}}$	DPI
	$n = 2k$	BR	EDS	CV$_{\text{reg}}$	CV$_{\text{mnl}}$	DPI
	$n = 5k$	BR	EDS	CV$_{\text{mnl}}$	CV$_{\text{reg}}$	DPI
Exponential	$n = k$	EDS	DPI	CV$_{\text{mnl}}$	CV$_{\text{reg}}$	BR
	$n = 2k$	DPI	EDS	CV$_{\text{mnl}}$	CV$_{\text{reg}}$	BR
	$n = 5k$	EDS	DPI	BR	CV$_{\text{mnl}}$	CV$_{\text{reg}}$
Beta	$n = 2k$	DPI	EDS	CV$_{\text{mnl}}$	BR	CV$_{\text{mnl}}$
	$n = 5k$	BR	DPI	EDS	CV$_{\text{mnl}}$	CV$_{\text{mnl}}$

Table 6.4: *The rankings based on paired Wilcoxon tests. The best performer is ranked at the left. When there is no significant difference between different methods they are underlined.*

Remark 6.1
The presented simulation study is only a first step in the investigation of the proposed bandwidth selection method EDS. Many questions about this method are still untouched, both practical and theoretical questions. The good overall performance in this simulation study encourages to do further research on the global version of EDS. For the partially local version of EDS, we encountered different kind of practical problems, which we do not have solved at this moment. The major problem we encountered is related to the possibility of negative estimated cell probabilities for the local linear estimators. This makes that the MSSE expressions used in Step 3(iv) of the algorithm become meaningless. Although the rescaled estimators, proposed in Remark 3.6, give an intuitive idea how to overcome this problem, the first results we got (only a limited number of simulations) were not too positive, in the sense that the partially local EDS method performed no better than the global version of EDS.

So, further investigation, and possibly a slight adaptation of the proposed EDS method, are necessary, before a more complete judgement of the method can be made.

6.3 Proofs

Proof of (6.1)
First note that there is a 1-1 relation between A and $\{1,\ldots,\#A\}$, vid.,

$$\begin{array}{ll} \text{for } i \in A & \exists!\ a \in \{1,\ldots,\#A\} \text{ with} \\ (\boldsymbol{S}_h)_{ij} = (\boldsymbol{S}_h^A)_{aj} & j = 1,\ldots,k. \end{array}$$

We then can write, for $i \in A$

$$\begin{aligned} E\widehat{P}_i - p_i &= \sum_{j=1}^{k} (\boldsymbol{S}_h)_{ij} p_j - p_i \\ &= \sum_{j=1}^{k} (\boldsymbol{S}_h - \boldsymbol{I}_k)_{ij} p_j \\ &= \sum_{j=1}^{k} (\boldsymbol{S}_h^A - \boldsymbol{I}_A)_{aj} p_j \\ &= \left((\boldsymbol{S}_h^A - \boldsymbol{I}_A)\boldsymbol{p}\right)_a \end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(\widehat{P}_i) &= \frac{1}{n}\sum_{j=1}^{k}(\boldsymbol{S}_h)^2_{ij}p_j - \frac{1}{n}\left(\sum_{j=1}^{k}(\boldsymbol{S}_h)_{ij}p_j\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{k}(\boldsymbol{S}_h^A)^2_{aj}p_j - \frac{1}{n}\left((\boldsymbol{S}_h^A\boldsymbol{p})_a\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{k}(\boldsymbol{S}_h^A)_{aj}\,\mathrm{diag}(\boldsymbol{p})_{jj}((\boldsymbol{S}_h^A)^T)_{ja} - \frac{1}{n}\left((\boldsymbol{S}_h^A\boldsymbol{p})_a\right)^2 \\
&= \frac{1}{n}\left(\boldsymbol{S}_h^A\,\mathrm{diag}(\boldsymbol{p})(\boldsymbol{S}_h^A)^T)\right)_{aa} - \frac{1}{n}\left((\boldsymbol{S}_h^A\boldsymbol{p})_a\right)^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{i\in A}\left(E\widehat{P}_i - p_i\right)^2 = \|(\boldsymbol{S}_h^A - \boldsymbol{I}_A)\boldsymbol{p}\|^2 \\
&\sum_{i\in A}\mathrm{Var}(\widehat{P}_i) = n^{-1}\mathrm{tr}\left(\boldsymbol{S}_h^A\,\mathrm{diag}(\boldsymbol{p})(\boldsymbol{S}_h^A)^T)\right) - n^{-1}\|\boldsymbol{S}_h^A\boldsymbol{p}\|^2.
\end{aligned}$$

■

Proof of (6.3)
In (6.1) the first term is the squared bias contribution and the second and the third term the variance contribution. The second term, which originates from the covariances among the $\overline{P}_i$'s, is, for sparse tables, of lower order than the third, which is the leading term of the variances of the p_i's. In the derivation for the L-function, this term will be omitted. Also we will only replace the $\boldsymbol{p}$ in the squared bias part of $\mathrm{MSSE_A}$ by $\boldsymbol{S}_g\overline{\boldsymbol{P}}$. This is based on the fact that estimation of the variance has a lower order effect on the performance of a double smoothing bandwidth selector. Thus, we work with the approximation

$$\mathrm{MSSE_A}(h,\boldsymbol{S}_g\overline{\boldsymbol{P}}) \simeq \|(\boldsymbol{S}_h^A - \boldsymbol{I}_A)\boldsymbol{S}_g\overline{\boldsymbol{P}}\|^2 + n^{-1}\mathrm{tr}\left\{\boldsymbol{S}_h^A\mathrm{diag}(\boldsymbol{p})(\boldsymbol{S}_h^A)^T\right\}.$$

Taking the derivative of $\mathrm{MSSE_A}(h,\boldsymbol{S}_g\overline{\boldsymbol{P}})$ with respect to h leads to

$$\mathrm{MSSE}_A^{[1]}(h,\boldsymbol{S}_g\overline{\boldsymbol{P}}) = \overline{\boldsymbol{P}}^T\boldsymbol{D}_{gh}^A\overline{\boldsymbol{P}} + c(h)$$

where $\boldsymbol{D}_{gh}^{A}$ is given in Section 6.1 and $c(h) = 2n^{-1}\mathrm{tr}\{\boldsymbol{S}_h^{\prime A}\,\mathrm{diag}(\boldsymbol{p})(\boldsymbol{S}_h^{A})^T\}$. This results in

$$\begin{aligned} E\{\mathrm{MSSE}_A^{[1]}(h, \boldsymbol{S}_g\overline{\boldsymbol{P}})^2\} &= \mathrm{Var}(\overline{\boldsymbol{P}}^T\boldsymbol{D}_{gh}^{A}\overline{\boldsymbol{P}}) + \left(E(\overline{\boldsymbol{P}}^T\boldsymbol{D}_{gh}^{A}\overline{\boldsymbol{P}})\right)^2 \\ &\quad + 2c(h)E(\overline{\boldsymbol{P}}^T\boldsymbol{D}_{gh}^{A}\overline{\boldsymbol{P}}) + c(h)^2. \end{aligned}$$

The last term does not involve g and hence can be left out in the minimization procedure. It is easy to see that

$$E(\overline{\boldsymbol{P}}^T\boldsymbol{D}_{gh}^{A}\overline{\boldsymbol{P}}) = (1-n^{-1})\boldsymbol{p}^T\boldsymbol{D}_{gh}^{A}\boldsymbol{p} + n^{-1}\{\mathrm{diagonal}(\boldsymbol{D}_{gh}^{A})\}^T\boldsymbol{p},$$

which leads to (6.3).

Before (6.3) can be implemented an explicit expression for $\mathrm{Var}(\overline{\boldsymbol{P}}^T\boldsymbol{D}_{gh}^{A}\overline{\boldsymbol{P}})$ needs to be given. This will be done for a general symmetric $k \times k$ matrix $\boldsymbol{D}$. Since the covariances between the $\overline{P}_i$'s are of lower order than the variances we will assume that the counts N_i are independent in obtaining an approximate expression for $\mathrm{Var}(\overline{\boldsymbol{P}}^T\boldsymbol{D}_{gh}^{A}\overline{\boldsymbol{P}})$. Thus, $N_i = n\overline{P}_i$, $i = 1, \ldots, k$, will be taken to be independent Binomial(n, p_i) binomial random variables.

We will use the tensor notation and results of McCullagh (1987). Let d_{ij} denote the (i,j) entry of $\boldsymbol{D}$. Generalized cumulants will be denoted using partitioned superscript notation. For example,

$$\kappa^{i,j} = \mathrm{cum}(N_i, N_j) = \mathrm{cov}(N_i, N_j) \quad \text{and} \quad \kappa^{i,jk} = \mathrm{cum}(N_i, N_jN_k).$$

It is clear that

$$\mathrm{Var}(\boldsymbol{N}^T\boldsymbol{D}\boldsymbol{N}) = \sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k}\sum_{i_4=1}^{k} d_{i_1i_2}d_{i_3i_4}\kappa^{i_1i_2,i_3i_4}.$$

A fundamental identity for generalized cumulants (see McCullagh (1987, p. 58)) states that

$$\begin{aligned} \kappa^{i_1i_2,i_3i_4} &= \kappa^{i_1,i_2,i_3,i_4} + \kappa^{i_1}\kappa^{i_2,i_3,i_4} + \kappa^{i_2}\kappa^{i_1,i_3,i_4} + \kappa^{i_3}\kappa^{i_1,i_2,i_4} + \kappa^{i_4}\kappa^{i_1,i_2,i_3} + \kappa^{i_1,i_3}\kappa^{i_2,i_4} \\ &\quad + \kappa^{i_1,i_4}\kappa^{i_2,i_3} + \kappa^{i_1}\kappa^{i_3}\kappa^{i_2,i_4} + \kappa^{i_1}\kappa^{i_4}\kappa^{i_2,i_3} + \kappa^{i_2}\kappa^{i_3}\kappa^{i_1,i_4} + \kappa^{i_2}\kappa^{i_4}\kappa^{i_1,i_3}. \end{aligned}$$

This implies that, because of the assumed mutual independence of the N_i's and the symmetry of $\boldsymbol{D}$,

$$\begin{aligned} \mathrm{Var}(\boldsymbol{N}^T\boldsymbol{D}\boldsymbol{N}) &= \sum_{i=1}^{k} d_{ii}^2\kappa^{i,i,i,i} + 4\sum_{i_1=1}^{k}\sum_{i_2=1}^{k} d_{i_1i_2}d_{i_2i_2}\kappa^{i_1}\kappa^{i_2,i_2,i_2} \\ &\quad + 2\sum_{i_1=1}^{k}\sum_{i_2=1}^{k} d_{i_1i_2}^2\kappa^{i_1,i_1}\kappa^{i_2,i_2} + 4\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\sum_{i_3=1}^{k} d_{i_1i_2}d_{i_2i_3}\kappa^{i_1}\kappa^{i_3}\kappa^{i_2,i_2}. \end{aligned}$$

Expressions for the first four cumulants of a multinomial distribution are given by,

$$\begin{aligned}
\kappa^{i} &= np_i \\
\kappa^{i,i} &= n(p_i - p_i^2) \\
\kappa^{i,i,i} &= n(p_i - 3p_i^2 + 2p_i^3) \equiv n\gamma(\boldsymbol{p})_i \\
\kappa^{i,i,i,i} &= n(p_i - 7p_i^2 + 12p_i^3 - 6p_i^4) \equiv n\tau(\boldsymbol{p})_i.
\end{aligned}$$

This results in

$$\begin{aligned}
\mathrm{Var}(\overline{\boldsymbol{P}}^T \boldsymbol{D}\overline{\boldsymbol{P}}) = n^{-3}(&\mathrm{diagonal}(\boldsymbol{D}) \odot \mathrm{diagonal}(\boldsymbol{D})^T \tau(\boldsymbol{p}) \\
&+ 4n^{-2}\boldsymbol{p}^T \boldsymbol{D}\{\mathrm{diagonal}(\boldsymbol{D}) \odot \gamma(\boldsymbol{p})\} + 2n^{-2}\mathrm{tr}(\boldsymbol{DVDV}) + 4n^{-1}\boldsymbol{p}^T\boldsymbol{DVDp}
\end{aligned}$$

where $\boldsymbol{V} = \mathrm{diag}(\boldsymbol{p} - \boldsymbol{p} \odot \boldsymbol{p})$ and $\odot$ means elementwise multiplication. ■

References

Aerts, M., Augustyns, I. and Janssen, P. (1997a) Sparse consistency and smoothing for multinomial data. *Statistics and Probability Letters*, **33**, 41–48.

Aerts, M., Augustyns, I. and Janssen, P. (1997b) Smoothing sparse multinomial data using local polynomial fitting. *Journal of Nonparametric Statistics* (to appear).

Aerts, M., Augustyns, I. and Janssen, P. (1997c) Local polynomial estimation of contingency table cell probabilities. *Statistics* (to appear).

Agresti, A. (1990) *Categorical data analysis.* J. Wiley, New York.

Aitchison, J. and Aitken, C.G.G. (1976) Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–420.

Augustyns, I. and Wand, M.P. (1997) Bandwidth selection for local polynomial smoothing of multinomial data. *Computational Statistics* (to appear).

Bowman, A.W. (1984) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–360.

Bowman, A.W. Hall, P. and Titterington, D.M. (1984) Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika*, **71**, 341–351.

Brown, P.J. and Rundell, P.W.K. (1985) Kernel estimates for categorical data. *Technometrics*, **27**, 293–299.

Burman, P. (1987a) Smoothing sparse contingency tables. *Sankhyā, Series A*, **49**, 24–36.

Burman, P. (1987b) Central limit theorem for quadratic forms for sparse tables. *Journal of Multivariate Analysis*, **22**, 258–277.

Cheng, M.Y. (1996). A bandwidth selector for local linear density estimators. *The Annals of Statistics* (to appear).

Cheng, M.Y. (1997). Boundary-aware estimators of integrated density derivative products. *Journal of the Royal Statistical Society, Series B* (to appear).

Cheng, M.-Y., Fan, J. and Marron, J.S. (1996) On automatic boundary corrections. *The Annals of Statistics* (to appear).

Chu, C.-K. and Marron, J.S. (1991) Choosing a kernel regression estimator. *Statistical Science*, **6**, 404–427.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing of scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cleveland, W.S. and Loader, C. (1996) Smoothing by local regression: principles and methods. In *Statistical Theory and Computational Aspects of Smoothing*, eds. W. Härdle and M.G. Schimeck, Physica-Verlag, Heidelberg, 10-49.

Chow, Y.S. and Teicher, H. (1978) *Probability Theory.* Springer-Verlag, New York.

Cristóbal, J.A. and Alcalá, J.T. (1996) Error process with a bandwidth matrix in multivariate local linear smoothing. Technical Report. University of Zaragoza.

Deheuvels, P. (1977) Estimation nonparamétrique de la densité par histogrammes généralisés (II). *Publications de l'Institute Statistique de l' Université de Paris*, **22**, 1–23.

de Montricher, G.M., Tapia, R.A. and Thompson, J.R. (1975) Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *The Annals of Statistics*, **3**, 1329–1348.

Dong, J. and Simonoff, J.S. (1994) The construction and properties of boundary kernels for smoothing sparse multinomials. *Journal of Computational and Graphical Statistics*, **3**, 57–66.

Dong, J. and Simonoff, J.S. (1995) A geometric combination estimator for d-dimension ordinal contingency tables. *The Annals of Statistics*, **23**, 1143–1159.

Dong, J. and Ye, Q. (1996) A minimum variance kernel and a discrete frequency polygon estimator for ordinal contingency tables. *Communications in Statistics, Theory and Methods*, **25**, 3217–3245.

Fan, J. (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

Fan, J. (1993) Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, **21**, 196–216.

Fan, J., Gasser,T., Gijbels, I., Brockmann, M. and Engel, J. (1997) Local polynomial regression: optimal kernels and asymptotic minimax efficiency. *The Annals of the Institute of Statistical Mathematics*, **49**, 79–99.

Fan, J. and Gijbels, I. (1991) Local linear smoothing in regression function estimation. Technical Report, Institute of Statistics Mimeo Series No. 2055, University of North Carolina, Chapel Hill.

Fan, J. and Gijbels, I. (1992) Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, **20**, 2008–2036.

Fan, J. and Gijbels, I. (1995) Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*,

4, 213–227.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*, Chapman and Hall, New York.

Fienberg, S.E. and Holland, P.W. (1973) Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, **68**, 683–691.

Fienberg, S.E., Bishop, Y.M. and Holland, P.W. (1975) *Discrete Multivariate Analysis. Theory and Practice.* The MIT Press, Cambridge, Ninth printing, 1988.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt, Heidelberg, Springer-Verlag, 23–68.

Gasser,T., Müller, H.-G. and Mammitzsch, V. (1985) Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B*, **47**, 238–252.

Good, I.J. and Gaskins, R.A. (1971) Nonparametric roughness penalties for probability densities. *Biometrika*, **58**, 255–277.

Grund, B. and Hall, P. (1993) On the performance of kernel estimators for high dimensional, sparse binary data. *Journal of Multivariate Analysis*, **44**, 321–344.

Hall, P. (1981) On nonparametric multivariate binary discrimination. *Biometrika*, **68**, 287–294.

Hall, P. (1984) Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis*, **14**, 1–16.

Hall, P. and Marron, J.S. (1987a) Extent to which least-squares cross-validation minimizes integrated squared error in nonparametric density estimation. *Probability Theory and Related Fields*, **74**, 567–581.

Hall, P. and Marron, J.S. (1987b) Estimation of integrated squared density derivatives. *Statistics and Probability Letters*, **6**, 109–115.

Hall, P., Marron, J.S. and Park, B.U. (1992) Smoothed cross-validation. *Probability Theory and Related Fields*, **92**,1–20.

Hall, P., Marron, J.S. and Titterington, D.M. (1995) On partial local smoothing rules for curve estimation. *Biometrika*, **82**, 575–587.

Hall, P. and Titterington, D.M. (1987) On smoothing sparse multinomial data. *Australian Journal of Statistics*, **29**, 19–37.

Hall, P. and Wehrly, T.E. (1991) A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, **86**, 665–672.

Härdle, W., Hall, P. and Marron, J.S. (1992) Regression smoothing parameters that are not far from their optimum. *Journal of the American Statistical Association*, **87**, 227–233.

Hastie, T. and Loader, C. (1993) Local regression : Automatic kernel carpentry. *Statistical Science*, **8**, 120–143.

Jones, M.C. (1993) Simple boundary correction for kernel density estimation. *Statistics and Computing*, **3**, 135–146.

Jones, M.C. and Foster, P.J. (1996) A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica*, **6**, 1005–1013.

Jones, M.C., Marron, J.S. and Park, B.U. (1991) A simple root-n bandwidth selector. *The Annals of Statistics*, **4**, 1919–1932.

Jones, M.C. Marron, J.S. and Sheather, S.J. (1996) A brief survey of bandwidth selectors for kernel density estimation. *Journal of the American Statistical Association*, **91**, 401–407.

Jones, M.C. and Sheather, S.J. (1991) Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics and Probability Letters*, **11**, 511–514.

Katkovnik, V.Y. (1979) Linear and nonlinear methods of nonparametric regression analysis. *Soviet Automatic Control*, **5**, 25–34.

Lejeune, M. and Sarda, P. (1992). Smooth estimators of distribution and density functions. *Computational Statistics and Data Analysis*, **14**, 457–471.

Mack, Y.P. and Müller, H.-G. (1989) Convolution type estimators for nonparametric regression estimation. *Statistics and Probability Letters*, **7**, 229–239.

Maguire, B.A., Pearson, E.S. and Wynn, A.H.A. (1952) The time intervals between industrial accidents. *Biometrika*, **39**, 168–180.

Mardia, K.V. (1970) *Families of bivariate distributions.* Griffin, London.

McCullagh, P. (1987) *Tensor Methods in Statistics*, Chapman and Hall, London.

McLeish, D.L. (1974) Dependent central limit theorems and invariance principles. *The Annals of Probability*, **2**, 620–628.

Müller, H.-G. (1985) Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. *Statistics and Decisions* Supplement no. **2**, 193–206.

Müller, H.-G. (1988) *Nonparametric regression analysis of longitudinal data.* Springer-Verlag, New York.

Nadaraya, E.A. (1964) On estimating regression. *Theory of Probability and Its Applications*, **10**, 186–190.

Park, B.U. and Marron, J.S. (1990) Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66–72.

Park, B.U. and Marron, J.S. (1992) On the use of pilot estimators in bandwidth selection. *Journal of Nonparametric Statistics*, **1**, 231–240.

Pollard, D. (1984) *Convergence of Stochastic Processes.* Springer-Verlag, New York.

Rajagopalan, B. and Lall, U. (1995) A kernel estimator for discrete distributions. *Journal of Nonparametric Statistics*, **4**, 409–426.

Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65–78.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1995) An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.

Ruppert, D. and Wand, M.P. (1994) Multivariate locally weighted least squares regression. *The Annals of Statistics*, **22**, 1346–1370.

Sacks, J. and Ylvisaker, D. (1970) Designs for regression problems with correlated errors III. *The Annals of Mathematical Statistics* , **41**, 2057–2074.

Santner, T.J. and Duffy, D.E. (1989) *The Statistical Analysis of Discrete Data.* Springer-Verlag, New York.

Schuster, E.F. (1985) Incorporating support constraints into nonparametric estimation of densities. *Communications in Statistics, Theory and Methods*, **14**, 1123–1126.

Scott, D.W., Tapia, R.A. and Thompson, J.R. (1977) Kernel density estimation revisited. *Nonlinear Analysis*, **1**, 339–372.

Scott, D.W., Tapia, R.A. and Thompson, J.R. (1980) Nonparametric probability density estimation by discrete maximum penalized likelihood criteria. *The Annals of Statistics*, **8**, 820–832.

Searle, S.R. (1982) *Matrix Algebra useful for Statistics.* J. Wiley, New York.

Seifert, B. and Gasser, T. (1996) Variance properties of local polynomials and ensuing modifications. In *Statistical Theory and Computational Aspects of Smoothing*, eds. W. Härdle and M.G. Schimeck, Physica-Verlag, Heidelberg, 50–79.

Sheather, S.J. (1986) An improved data-based algorithm for choosing the window width when estimating the density at a point. *Computational Statistics and Data Analysis*, **4**, 61–65.

Sheather, S.J. and Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683–690.

Simonoff, J.S. (1983) A penalty function approach to smoothing large sparse contingency tables. *The Annals of Statistics*, **11**, 208–218.

Simonoff, J.S. (1985). An improved goodness-of-fit statistic for sparse multinomials. *Journal of the American Statistical Association*, **80**, 671–677.

Simonoff, J.S. (1987) Probability estimation via smoothing in sparse contingency tables. *Statistics and Probability Letters*, **5**, 55–63.

Simonoff, J.S. (1995) Smoothing categorical data. *Journal of Statistical Planning and Inference*, **47**, 41–69.

Simonoff (1996) *Smoothing Methods in Statistics*, Springer-Verlag, New York.

Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317–344.

Staniswalis, J.G. (1989) Local bandwidth selection for kernel estimates. *Journal of the American Statistical Association*, **84**, 284–288.

Stone, C.J. (1977) Consistent nonparametric regression. *The Annals of Statistics*, **5**, 595–645.

Stone, C.J. (1984) An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, **12**, 1285–1297.

Tapia, R.A. and Thompson, J.R. (1978) *Nonparametric Probability Density Estimation.* The John Hopkins University Press, Baltimore.

Titterington, D.M. and Bowman, A.W. (1985) A comparative study of smoothing procedures for ordered categorical data. *Journal of Computational and Graphical Statistics*, **21**, 291–312.

Wand, M.P. (1992) Error analysis for general multivariate kernel estimators. *Journal of Nonparametric Statistics*, **2**, 1–15.

Wand, M.P. and Jones, M.C. (1993) Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, **88**, 520–528.

Wand, M.P. and Jones, M.C. (1994) Multivariate plug-in bandwidth selectors. *Computational Statistics*, **9**, 97–117.

Watson, G.S. (1964) Smooth regression analysis. *Sankhyā, Series A*, **26**, 359–372.