DIFFERENTIAL CUMULANTS, HIERARCHICAL MODELS AND MONOMIAL IDEALS

Daniel Bruynooghe
Department of Statistics

Dissertation

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

London School of Economics

2011

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work.

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of approximately 30000 words.

Abstract

This thesis studies hierarchical models for the joint density f_X of a random vector $X = (X_1, ..., X_d)$, i.e. models characterised by the fact that interaction within a set of variables implies interaction within any of its subsets. A hierarchical model on a simplicial complex $\mathcal{S}$ can be written in exponential form as

$$f_X(x) = \exp\left\{ \sum_{J \in \mathcal{S}} h_J(x_J) \right\}.$$

The statistical implications of the choice of $\mathcal{S}$ are studied. Associated to a simplicial complex $\mathcal{S}$ is its Stanley-Reisner ideal $I_{\mathcal{S}}$. Hence, hierarchical models can be identified uniquely with monomial ideals. This isomorphism bridges the fields of statistics and commutative algebra. Simplicial complexes holding sets with at most two elements can be illustrated with graphs, thus leading to graphical models. The missing edges of a graph represent two-element sets excluded from $\mathcal{S}$ implying conditional independence. It is shown that sets excluded from $\mathcal{S}$ imply differential conditions on the log-density and that these differentials arise naturally as cumulants in an infinitesimally small neighbourhood around a given $x_0 \in \mathbb{R}^d$. A new bootstrap test for conditional independence is constructed based on the notion that certain differential cumulants are zero everywhere under conditional independence.

To Vlad, who is sadly missed.

Acknowledgments

I would like to express my gratitude to my supervisor, Henry Wynn, whose expertise, creativity and humor made this a most enjoyable graduate experience. To plagiarise Michael Atiyah, Henry peered around a lot of corners. He also taught me a lot about life in general. Henry's dedication and commitment was outstanding. I could not have hoped for better supervision.

I would like to thank Lenny Smith for his continued support throughout the project. He taught me to remain sceptical.

Eduardo Sáenz de Cabezón was kind enough to let me draw on his vast amount of algebraic knowledge. His guidance proved invaluable.

The Department of Statistics has provided a great environment for PhD research. Special thanks go to Pauline Barrieu for her general support throughout and to Ian Marshall for taking care of all administrative issues.

IT Services has been a great help. Vladimir Konrad and Kuldip Purewal have done a fantastic job in setting up and maintaining an infrastructure which allowed me to work efficiently and remotely.

I would also like to thank my family. My parents and my parents-in-law have provided fantastic support, each in their very own way. I can truly they that I would not have finished this thesis without their help. Special thanks goes to my parents for the way they have raised and supported me throughout my entire life.

In particular, I am grateful to my wonderful wife Lise. Her love, encouragement and support are invaluable. I am very lucky to have her at my side.

Financial and other support from Studienstiftung des Deutschen Volkes, Lloyds of London, the EPSRC and the Grantham Research Institute on Climate Change and the Environment is very gratefully acknowledged.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The primary object studied in this thesis is the class of hierarchical models for the joint density f_X of a random vector $X = (X_1, ..., X_d)$. Hierarchical models are characterised by the fact that interaction within a set of variables implies interaction within any of its subsets. By contraposition, a lack of interaction in a set implies the lack of interactions in all sets it is contained in.

Let $\mathcal{S}$ be a simplicial complex on $[d] = \{1, ..., d\}$, i.e. a collection of subsets of $[d]$ closed under taking subsets. Assuming, as we do throughout, that f_X is strictly positive everywhere, we may model f_X in exponential form as

$$f_X(x) = \exp\left\{\sum_{J \in \mathcal{S}} h_J(x_J)\right\},$$

where each function h_J is operating on x_J, the subset of indeterminates indexed by J.

Much of this thesis is about the choice of $\mathcal{S}$ and its statistical implications. Typically, the sets excluded from $\mathcal{S}$ induce interesting statistical structures. For instance, the exclusion of a two-element subset $\{i, j\}$ from $\mathcal{S}$ implies that no function h_J must be a function of both x_i and x_j. Consequently, f_X factorises and we obtain conditional independence of X_i and X_j given the remaining variables.

A natural place to start with is a simplicial complex which holds only sets with one or two elements or sets which are entirely determined by the two-element sets via additional restrictions. Such a complex can be associated to a graph, whose vertices represent the random variables $X_1, ..., X_d$ and whose edges represent pairwise interactions. This makes the statistical problem accessible to graph-theoretic

considerations. For instance, it is well known that the graph-theoretic concept of decomposability leads to models with closed form maximum likelihood estimators (Lauritzen, 1996). The corresponding subclass of models is referred to as graphical models.

As sets with more than two elements are allowed into $\mathcal{S}$ an algebraic treatment becomes essential. The link between statistics and commutative algebra has been investigated primarily for discrete probability models (Pistone et al., 2001; Geiger et al., 2006). One of the key contributions of this thesis is to bridge the two fields in the continuous case. Associated to a simplicial complex $\mathcal{S}$ is its Stanley-Reisner ideal in the polynomial ring $k[x_1, ..., x_d]$. This is the ideal generated by the minimal sets which are excluded from $\mathcal{S}$ allowing us to uniquely identify hierarchical models with monomial ideals.

Another key observation is that a function h_J is excluded from the model as the derivative of the log-density with respect to the indeterminates x_J is set to zero everywhere. For this reason the differentials of the log-density have been called mixed interaction terms (Whittaker, 1990). By showing that these differentials arise naturally as cumulants in an infinitesimally small neighbourhood around a given $x \in \mathbb{R}^d$, we offer a new interpretation of mixed interaction terms as *differential cumulants*.

Differential cumulants with respect to two variables X_i and X_j take the form

$$\frac{\partial^2 \log f_X(x)}{\partial x_i \partial x_j}. \tag{1.1}$$

As mentioned, setting these to zero everywhere annihilates associated h_J functions and expresses conditional independence of X_i and X_j given the remaining $d-2$ variables. Differential cumulants are estimable using kernel estimators for f_X and its first two derivatives. Thus, we can construct a bootstrap hypotheses test for conditional independence.

Summarising, this thesis deals with hierarchical models which are, to a large

extent, identified through what they do not have: interaction between subsets of variables and, equivalently, edges or faces in the associated graphs or simplicial complexes. Phrased positively, they do have differential cumulants which vanish everywhere and they do have generators in their Stanley-Reisner ideals. In the case that they do not have interaction between just a pair of variables they also have conditional independence structures attached to them. Explaining all these links carefully is the challenge that lies ahead.

Chapter 2 starts with the problem of expressing moments in terms of cumulants and vice versa. Based on a multivariate chain rule, a formula is provided which makes the combinatorial aspects explicit. The second part of the chapter introduces local analogues to moments and cumulants. *Local* moments will be defined as the conditional moment in a sufficiently small neighbourhood of a point $x \in \mathbb{R}^d$. Local cumulants are defined in terms of local moments through the ex-log-relationship induced by their generating functions. The limiting process is considered and the remarkably simple forms of *differential* moments and cumulants are derived.

Chapter 3 explains the relations between sets of *zero-cumulants*, conditional independence statements and hierarchical models. Naturally, many of the results linking graphical models to conditional independence associations are well established. The novelty of this chapter is to demonstrate how particular model classes can be obtained through imposing restrictions on differential cumulants.

Chapter 4 investigates the link between hierarchical models based on a simplicial complex $\mathcal{S}$ and the algebra via monomial ideals. The subclass of decomposable models is characterised through algebraic properties of the Stanley-Reisner ideal of $\mathcal{S}$. Furthermore, models derived from the so called Ferrer ideals are presented as an example of how the algebra can lead us to interesting classes of statistical models. Finally, the algebraic concept of shellability is introduced, which is closely related to decomposability.

Chapter 5 develops a nonparametric hypothesis test for conditional independence. As this is equivalent to certain differential cumulants vanishing everywhere, a squared version of them, integrated over $\mathbb{R}^d$, should be close to zero under conditional independence. The density f_X and its derivatives are estimated through kernel estimators. The test statistic is based on an expansion of (1.1). It takes the form

$$\int_{\mathbb{R}^d} \left(\frac{1}{\hat{f}(x)} \frac{\partial^2 \hat{f}(x)}{\partial x_i \partial x_j} - \frac{1}{\hat{f}^2(x)} \frac{\partial \hat{f}(x)}{\partial x_i} \frac{\partial \hat{f}(x)}{\partial x_j} \right)^2 dx.$$

We suggest a bootstrap hypotheses test and demonstrate its validity through simulations.

Chapter 6 describes the estimation of a functional of local moments. It is based on Chapter 2 and is not related to intermediate chapters. Local moments are functions of the density f_X and its derivative. The key idea of this chapter is to demonstrate two alternative views on how to exploit this relation for estimation. The first view takes sample analogues to local moments and uses them to estimate densities. The second view takes density estimators and uses them to estimate local moments.

Finally, we give conclusions and list some topics which might naturally have been included or which are thought to be promising future research topics.

Chapter 2

Differential moments and cumulants

2.1 Introduction

This chapter investigates the relationship between multivariate moments and cumulants and their localised counterparts. The first part is devoted to the *ex-log* relationship through which moments can be expressed as functions of cumulants and vice versa.

A well established approach to computing cumulants from moments is to compare the coefficients in the formal power series expansions of the moment and cumulant generating functions. This leads to a formula for moments in terms of cumulants. A subsequent application of a *Möbius inversion* yields an expression for cumulants in terms of moments (Barndorff-Nielsen and Cox, 1989).

Higher order cumulants can be calculated as we identify formerly distinct random variables. An example: In fairly standard notation, which is also explained below, the cumulant κ_{120} can be treated similarly to the cumulant κ_{111} as we identify X_2 with X_3. This identification introduces extra factors, which are not always particularly easy to calculate (Speed, 1983). Other contributions come from McCullagh (1984) and Stuart and Ord (1994).

An alternative to coefficient comparison is to compute cumulants directly as derivatives of the logarithm of the moment generating function evaluated at the origin. This approach requires us to consider higher order derivatives of composite functions, where the inner function is multivariate. Once such a formula

is established, multivariate cumulants of arbitrary order can be computed readily. We refer to the combinatorial quantity which explicitly accounts for multiple identifications as the *collapse number* of a partition.

Whilst this particular multivariate extension of the chain rule has been provided by Hardy (2006), it has, to the best of our knowledge, not been applied to the statistical problem of expressing cumulants through moments. The first part of this chapter thus explains the combinatorics of this identification carefully. Numerous examples are provided.

The second part of the chapter introduces local analogues of moments and cumulants. A *local* moment will be defined as a conditional moment in a sufficiently small neighbourhood of a point $\xi \in \mathbb{R}^d$. Local cumulants are defined in terms of local moments through the ex-log-relationship discussed in the first part of the chapter. The limiting process is considered and the remarkably simple forms of *differential* moments and cumulants are derived.

Of particular interest throughout the thesis are *square-free* cumulants, i.e. cumulants of binary order. A uniquely characterising property of square-free cumulants is proved before the chapter is concluded.

2.2 Moments and cumulants

Let $X = (X_i)_{1 \leq i \leq d}$ be a random vector whose components are defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For the most part, we consider the cases $d = 2$ or $d = 3$, which make the exposition and notation tractable whilst allowing us to illustrate multivariate phenomena. We assume that X is real-valued and its distribution function F_X is absolutely continuous and allows a $d+1$ times continuously differentiable density f_X which is strictly positive everywhere. We further require X to have at least the first two moments.

Definition 1 (Monomial)**.** Let $x \in \mathbb{R}^d$. A monomial in $x_1, ..., x_d$ is a product

$\prod_{i=1}^{d} x_i^{k_i}$ for some $k \in \mathbb{N}_0^d$. We set $x^k := \prod_{i=1}^{d} x_i^{k_i}$.

Monomials naturally occur in defining moments of X. Since we deal with non-central moments, moments will be denoted by m. Let $\mathbb{E}$ denote the expectation operator. We use the following indexing convention:

$$m_{k_1 \dots k_d} = \mathbb{E}(X^k) = \mathbb{E}\left(\prod_{i=1}^{d} X_i^{k_i} \right).$$

Example 1. For $X \in \mathbb{R}^3$ the non-centralised moment of order $(0, 2, 1)$ is given by $\mathbb{E}\left(X_2^2 X_3\right)$. It is denoted by m_{021}.

Let e_i denote the i-th unit vector. The first moment of X_i is thus represented by m_{e_i} and the covariance matrix of X can be expressed as

$$\operatorname{cov}(X) = (m_{e_i+e_j})_{1 \le i,j \le d} - m_{e_i} m'_{e_j},$$

where we use the notation $(a_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}}$ to denote an m by n matrix with (i, j)-th entry a_{ij}.

Example 2. The covariance between two random variables is given by $m_{11} - m_{10} m_{01}$.

An alternative tensor representation is suggested by McCullagh (1984), which may generalise better in higher dimensions.

Let $M_X : \mathbb{R}^d \longrightarrow \mathbb{R}$ denote the moment generating function of X:

$$M_X(t) := E(e^{t'X}) = \int_{\mathbb{R}^d} e^{\sum_{i=1}^{d} t_i x_i} dF_X(x).$$

The moment generating function exists, whenever the integral on the right hand side is absolutely convergent. The existence of moments of all orders is not a sufficient condition for this to be the case, as can be demonstrated through the lognormal distribution. The moment generating function derives its name from the fact that its derivatives evaluated at the origin correspond to the moments of X.

Let the cumulant generating function $K_X(t) : \mathbb{R}^d \longrightarrow \mathbb{R}$ be defined as the natural logarithm of the moment generating function $K_X(t) := \log(M_X(t))$. We may surpress the subscript X where no ambiguity is expected. Provided that $K_X(t)$ has a Taylor representation about the origin, we can write

$$K_X(t) := 1 + \kappa_1' t + \frac{1}{2!} t' \kappa_2 t + ...,$$

where

$$\kappa_1 := (\kappa_{e_i})_{1 \leq i \leq d}, \quad \kappa_2 := (\kappa_{e_i + e_j})_{1 \leq i,j \leq d} \text{ etc.}$$

The coefficients of this representation define the cumulants. Note that $\kappa_k \in \mathbb{R}^{d^k}$. Cumulants are of considerable theoretical interest for, in the words of Speed (1983),

> 'In a sense which is hard to make precise, all of the important aspects of (joint) distributions seem to be simpler functions of cumulants than of anything else.'

The definition of the cumulant generating function as the logarithm of the moment generating function implicitly defines a bijective mapping from moments to cumulants. This mapping can be recovered by a coefficient comparision of the respective Taylor expansions (see for instance Stuart and Ord, 1994).

Example 3. Table 2.1 lists the first few cumulants in terms of moments for the case $d = 3$.

An alternative to coefficient comparison is to compute the cumulants via differentiation of K_X. This is demonstrated through a univariate example.

Example 4. To find the cumulant κ_2 consider the second derivative of $K(t)$:

$$\begin{aligned} \frac{d^2 K(t)}{dt^2} &= \frac{d}{dt} \left\{ \frac{d \log(M(t))}{dM(t)} \frac{dM(t)}{dt} \right\} \\ &= \frac{d^2 \log(M(t))}{dM(t)^2} \left(\frac{dM(t)}{dt} \right)^2 + \frac{d \log(M(t))}{dM(t)} \frac{d^2 M(t)}{dt^2}. \end{aligned}$$

Cumulants	Moments
κ_{100}	m_{100}
κ_{200}	$m_{200} - m_{100}^2$
κ_{110}	$m_{110} - m_{100}m_{010}$
κ_{300}	$m_{300} - 3m_{200}m_{100} + 2m_{100}^3$
κ_{210}	$m_{210} - m_{200}m_{010} - 2m_{100}m_{110} + 2m_{100}^2 m_{010}$
κ_{111}	$m_{111} - m_{110}m_{001} - m_{101}m_{010} - m_{011}m_{100} + 2m_{100}m_{010}m_{001}$

Table 2.1 – Cumulants in terms of moments.

Evaluated at the origin, the differentials of M are the moments:

$$\left.\frac{dM(t)}{dt}\right|_{t=0} = m_1 \quad \text{and} \quad \left.\frac{d^2M(t)}{dt^2}\right|_{t=0} = m_2.$$

Evaluating the differentials of $K = \log M(t)$ at the origin yields $\frac{d\log(M(t))}{dM(t)} = \frac{1}{M(t)} = 1$ and $\frac{d^2\log(M(t))}{dM(t)^2} = \frac{d}{dt}\frac{1}{M(t)} = -\frac{1}{M(t)^2} = -1$ since $M(0) = 1$. This identifies the cumulant κ_2 as the variance:

$$\kappa_2 = \left.\frac{d^2K(t)}{dt^2}\right|_{t=0} = m_2 - m_1^2.$$

Extending Example 4 to univariate higher order cumulants is achieved via Faa Di Bruno's formula for higher order derivatives of composite functions (di Bruno, 1855). It states that, for two functions g and h from $\mathbb{R}$ to $\mathbb{R}$, it holds that

$$\frac{d^k}{dx^k}g(h(x)) = \sum \frac{k!}{m_1!1!^{m_1}m_2!2!^{m_2}\cdots m_k!k!^{m_k}} g^{(m_1+\cdots+m_k)}(h)\prod_{j=1}^{n}\left(h^{(j)}(x)\right)^{m_j},$$

where the summation is over all k-tuples $(m_1, ..., m_k)$ of non-negative integers satisfying the constraint $\sum_{i=1}^{k} m_i\, i = k$.

Example 5. We consider the third derivative of $K(t)$. The three m-tuples satis-

fying the above constraint are given by $(3,0,0),(1,1,0)$ and $(0,0,1)$. Hence,

$$\begin{aligned}\frac{d^3}{dt^3}\log(M(t)) = K^{(3)}(M)\ M'(t) \\ + 3K^{(2)}(M)\ M^{(2)}(t)M'(t) \\ + K'(M)\ M^{(3)}(t),\end{aligned}$$

from which we infer that $\kappa_3 = 2m_1 - 3m_1m_2 + m_3$.

Multivariate cumulants can also be obtained through differentiating the cumulant generating function, as Example 6 illustrates.

Example 6. The covariance between two random variables is given by the cumulant of order $(1,1)$ since

$$\kappa_{11} = \frac{\partial^2 K_X(t)}{\partial t_1 \partial t_2}\bigg|_{t=0} = \left(\frac{\partial^2 M_X(t)}{\partial t_1 \partial t_2} - \frac{\partial M_X(t)}{\partial t_1}\frac{\partial M_X(t)}{\partial t_2}\right)\bigg|_{t=0} = m_{11} - m_{10}m_{01}.$$

Similarly to the univariate case, we seek a formula which allows us to compute higher order cumulants directly. In order to extend Example 5 to multivariate random variables we need a generalisation of Faa di Bruno's formula for computing arbitrary derivatives of composite functions, where the inner function maps from $\mathbb{R}^d$ to $\mathbb{R}$. To the best of our knowledge, this formula was first derived by Hardy in a way that makes the combinatorial aspects explicit. Much of the following is based on Hardy (2006).

Before we can state the multivariate chain rule, we need to introduce some notation and set-related quantities such as *multisets*, *partitions* of them and *collapse numbers*. The collapse number of a partition is a combinatorial quantity which, roughly speaking, counts the number of partitions which become indistinguishable as elements of a set become indistinguishable. The collapse number will turn out to play a key role in the multivariate chain rule. We provide numerous examples following the definitions.

Definition 2 (Multiset, multiplicity, size)**.** A *multiset* M is a set which may hold multiple copies of its elements. The *population set* of a multiset is the set of elements M can hold copies of. The *multiplicity of an element* is the number of occurrences of that element in the multiset. The *multiplicity of a multiset* is the vector of multiplicities of its elements, denoted by ν_M. The total number of not necessarily distinct elements $|M|$ in M is the *size* of M. A multiset which is a set is called *degenerate*, i.e. degenerate multisets hold exactly one copy of each element.

Example 7 (Multiset)**.** The multiset $M_1 = \{x_1, x_3, x_3\}$ holds one copy of x_1, no copy of x_2 and two copies of x_3. The multiplicity is $\nu_{M_1} = (1, 0, 2)$. The set $M_2 = \{x_1, x_2, x_3\}$ is a degenerate multiset since no element occurs more than once.

Multisets are hybrids between vectors and sets since, like vectors, multiple occurences of a member are regarded as different entities whereas, like sets, the order of elements does not matter. Multisets with same multiplicity are isomorphic. Hence, we may choose a multiset to hold integers since its properties are not affected by the names of the elements.

As in Example 7, we will not explicitly mention the population set, which typically consists of the set of variates $\{x_1, ..., x_d\}$ or the set of integers $\{1, ..., d\}$. The population set only affects the zeros in the multiplicity of a multiset, which do not affect our results.

Definition 3 (Partition of a multiset)**.** Let I be some index set. A *partition* π of a multiset M is a multiset of multisets $\{(M_i)_{i\in I}\}$ such that $\nu_M = \sum_{i\in I} \nu_{M_i}$, where $(\nu_{M_i})_{i\in I}$ is the family of multiplicities associated with π. We denote the multiplicity of π by ν_π and adopt the shorthand notation $\pi = \{M_1|M_2| \cdots |M_{|I|}\}$.

A partition of a multiset is a regrouping into smaller multisets such that every copy of every element is put inside exactly one of the smaller multisets. Being a

multiset itself, a partition can hold multiple copies of one or more multisets.

Example 8 (Partition of a multiset)**.** The multiset $\pi = \{\{x_1, x_3\}, \{x_1, x_3\}, \{x_3\}\} = \{x_1x_3|x_1, x_3|x_3\}$ has the associated family of multiplicities $\big((1,0,1),(1,0,1),(0,0,1)\big)$. It is a partition of $M = \{x_1, x_1, x_3, x_3, x_3\}$ since $(1,0,1)+(1,0,1)+(0,0,1) = (2,0,3) = \nu_M$. The multiplicity of π is $(2,1)$, as π holds two copies of $\{x_1, x_3\}$ and one copy of $\{x_3\}$.

We use multisets in the current context as we can identify orders of derivatives with them. For a vector $\alpha \in \mathbb{N}_0^d$ we set

$$D^\alpha f(x) := \frac{\partial^{|\alpha|}}{\prod_{i=1}^d \partial x_i^{\alpha_i}} f(x),$$

where $|\alpha| := \sum_{i=1}^n \alpha_i$. By convention $D^0 f(x) := f(x)$. We refer to α as the order of derivative and $|\alpha|$ as the total degree. The D-operator notation makes the close link between multisets and derivatives obvious since the order of the derivative operator is identical to the multiplicity of the associated multiset. Again, this is best illustrated with an example.

Example 9 (Partial derivative and multiset)**.** Given the partial derivative $D^{102} = \frac{\partial^3}{\partial x_1 \partial x_3^2} f(x)$ the differentiation is once with respect to x_1 and twice with respect to x_3. This differential operation can be associated with the multiset $\{1,3,3\}$ with multiplicity $(1,0,2)$.

The following formula provides a generalisation of the chain rule for composite functions, when the inner function is from $\mathbb{R}^d$ to $\mathbb{R}$ and the outer function is from $\mathbb{R}$ to $\mathbb{R}$. Suitable differentiability conditions are assumed.

$$D^{(1\cdots1)} g(h(x)) = \sum_{\pi \in \Pi(k)} \frac{d^{|\pi|} g(h)}{dh^{|\pi|}} \prod_{j=1}^{|\pi|} D^{\nu_{M_j}} h(x), \tag{2.1}$$

where $\Pi(k)$ is the set of all partitions of a multiset with multiplicity k and M_j is the j-th multiset in the partition π. It can be proved by induction on the number of variates.

Formula (2.1) allows us to compute higher order derivatives of composite functions when the differentiation is taken with respect to each variable once. What we seek is a generalisation to derivatives of arbitrary order. The key insight is that repeated differentiation with respect to one variable can be treated as a special case of differentiation with respect to several variables, where some of them are indistinguishable. For instance, $D^{(42)}g(h(x_1, x_2))$ can be thought of as a special case of $D^{(111111)}g(h(x_1, ..., x_6))$, where $x_1 = x_2 = x_3 = x_4$ and $x_5 = x_6$. As we identify derivative operators with multisets, the combinatorics of multiset partitions need to be taken into account.

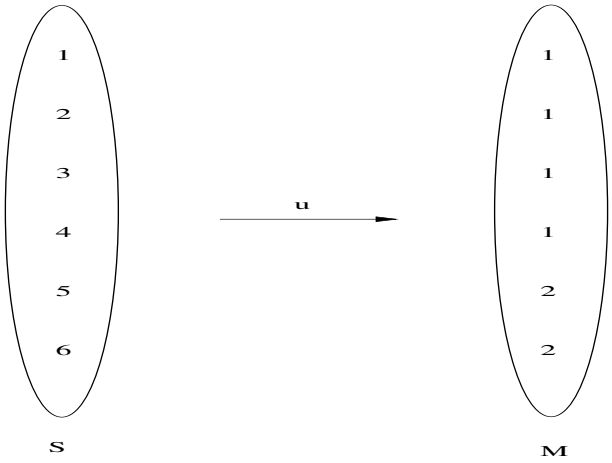

Figure 2.1 – The diagram shows the collapsing mapping u.

For a multiset M with size $|M|$, define $S := \{1, 2, ..., |M|\}$. The set S is the equivalent of the differential operator without repeated differentiation; or D^{111111} in the above example. By choice of S, M and S are of same size. We consider the class of surjective mappings $u : S \longrightarrow M$, such that every element of M has a pre-image under u. As will become clear below, all mappings of this kind induce the same collapse number and there is no need to specify the details of the mapping u.

If M is non-degenerate, i.e. M holds multiple copies of at least one if its elements, then M holds fewer distinct elements than S. Hence, some of the elements

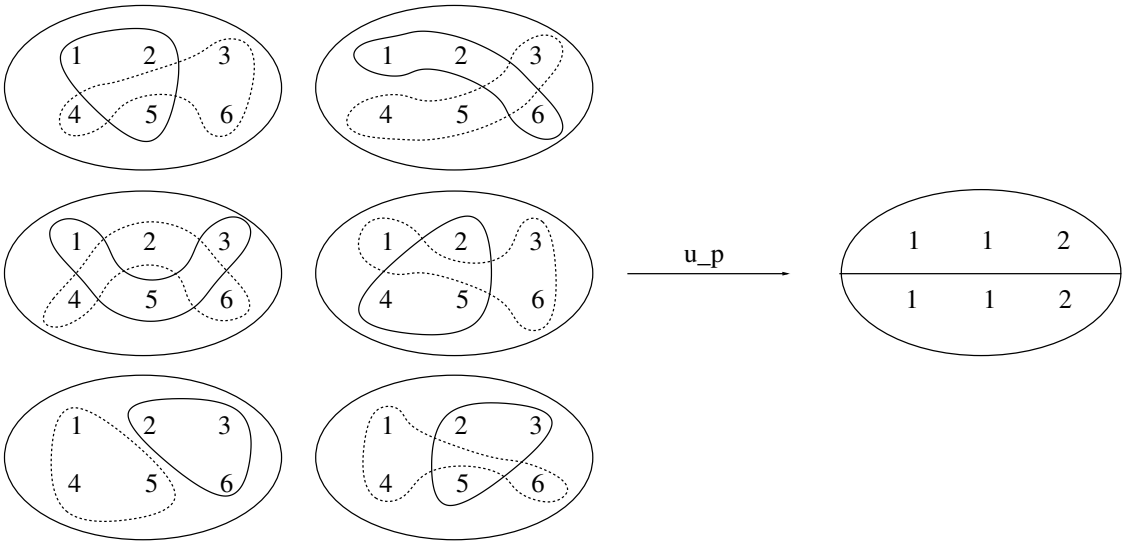

Figure 2.2 – The diagram shows the partitions of S which get mapped onto partition $\{112|112\}$ of M under the collapsing mapping u_P.

of S become indistinguishable under u. We say that S collapses onto M.

Example 10 (Collapsing mapping of a multiset)**.** Consider the multiset $M = \{1, 1, 1, 1, 2, 2\}$. We set $S := \{1, 2, 3, 4, 5, 6\}$. A *collapsing mapping* is defined through

$$u : S \to M$$
$$u(1) = u(2) = u(3) = u(4) = 1 \text{ and } u(5) = u(6) = 2,$$

as depicted in Figure 2.1.

The collapsing mapping u induces a canonical mapping u_P from the set of partitions of S into the set of partitions of M. The pre-image of a partition π of a multiset M under u_P is the set of all partitions of S, which are mapped on to π.

Example 11 (Collapsing of partitions)**.** Consider again the multiset $M = \{1, 1, 1, 1, 2, 2\}$ with partition $\pi = \{112|112\}$ and collapsing mapping u as defined

in Example 10. The pre-image of π under u_P is given by the six partitions

$$\{125|346\}, \{135|246\}, \{145|236\}, \{235|146\}, \{245|136\}, \{345|126\}$$

since each of them gets mapped onto π under u_P. For instance,

$$\begin{aligned} u_P(\{125|346\}) &= \{u(1)u(2)u(5)|u(3)u(4)u(6)\} \\ &= \{112|112\} \\ &= \pi. \end{aligned}$$

This collapsing of the partitions is illustrated in Figure 2.2.

Of particular interest is the number of elements in the pre-image, the collapse number.

Definition 4 (Collapse number of a partition)**.** The *collapse number* $c(\pi)$ is the size of the pre-image of π under u_P.

The collapse number of a partition π can be interpreted as the number of partitions that would exist if multiple copies of elements of M were distinguishable.

Let the factorial sign following a vector denote the product of its entries, i.e. for $x \in \mathbb{N}_0^d$ we set $x! := \prod_{i=1}^d x_i$. Lemma 1 states a formula for computing collapse numbers. Note that the collapse number only depends on the multiplicities of the multiset M, the multiplicity of the partition π and the family of multiplicities associated with π.

Lemma 1 (Collapse number of a partition)**.** *Let $\pi = \{(M_i)_{i\in I}\}$ be a partition of a multiset M. Let $\nu_\pi, (\nu_{M_i})_{i\in I}$ and ν_M denote the multiplicities of $\pi, (M_i)_{i\in I}$ and M respectively. Then the collapse number $c(\pi)$ is given by*

$$c(\pi) := \frac{\nu_M!}{\prod_{i\in I} \nu_{M_i}! \nu_\pi!}. \tag{2.2}$$

Proof. The numerator in (2.2) is a count of the permutations of copies of the elements of M, all of which correspond to the same partition once collapsed but to different partitions prior to collapsing. This number needs to be qualified by $\prod_{i\in I}\nu_{M_i}!$, the number of permutations within partition blocks, and $(\nu_\pi!)$, the number of permutations of partition blocks. □

Example 12 (Collapse number of a partition)**.** Consider the multiset $M = \{1,1,1,1,2,2\}$ with partition $\pi = \{112|112\}$ from previous examples. In order to apply Lemma 1, we note that $\nu_M = (4,2), \nu_{M_1} = \nu_{M_2} = (2,1)$ and $\nu_\pi = 2$. The collapse number is given by

$$c(\pi) := \frac{4!2!}{2!2!2!} = 6,$$

as was verifed before in Example 11 by counting sets.

The next theorem states a formula for higher order derivatives. We will refer to it as *Hardy's Theorem.*

Theorem 1 (Higher order derivative of chain functions)**.** *Let $k \in \mathbb{N}_0^d$ be an order of a derivative and g and h be functions from $\mathbb{R}$ to $\mathbb{R}$ and $\mathbb{R}^d$ to $\mathbb{R}$ respectively which are at least $|k|$ times differentiable in x. Then it holds for the $k-$th derivative of the composite function that*

$$D^k g(h(x)) = \sum_{\pi\in\Pi(k)} c(\pi)\frac{d^{|\pi|}g(h)}{dh^{|\pi|}}\prod_{j=1}^{|\pi|} D^{\nu_{M_j}}h(x),$$

where $\Pi(k)$ is the set of all partitions of a multiset with multiplicity k and M_j is the j-th multiset in the partition π.

Proof. The core of the proof is the identification of derivatives with multisets, equation (2.1) and the notion that repeated differentation with respect to one variable can be treated as a special case of (2.1) with some variables being indistinguishable. All we are required to establish is the number of differentials

that become indistinguishable for a given partition. This, however, is exactly the collapse number as the derivative operator collapses from $D^{\overbrace{1 \cdots 1}^{|k|}}$ to D^k. □

As a corollary, we obtain a general formula for computing multivariate cumulants from moments.

Corollary 1 (Cumulants as functions of moments)**.** *Let κ_k be the k-th cumulant. Then*

$$\kappa_k = \sum_{\pi \in \Pi(k)} c(\pi)(-1)^{(|\pi|-1)}(|\pi|-1)! \prod_{j=1}^{|\pi|} m_{\nu_{M_j}}, \tag{2.3}$$

where $\Pi(k)$ is the set of all partitions of a multiset with multiplicity k, M_j is the $j-$th multiset in partition π with associated multiplicity ν_{M_j} and $m_{\nu_{M_j}}$ is the moment of order ν_{M_j}.

Proof. Set $g(h) = \log(h)$ and $h(t) = M_X(t) = \mathbb{E}e^{t'X}$, apply Hardy's Theorem and evaluate at the origin. Note that $M_X(0) = 1$. Hence, neither $M_X(t)$ nor any of its powers appear in (2.3). □

Example 13 (Higher order cumulants)**.** Consider the partial derivative

$$\frac{\partial^3}{\partial x \partial z^2} g(h(x,y,z))$$

from Example 9. The associated multiset $\{1,3,3\}$ has partitions $\{133\}$, $\{13|3\}$, $\{1|33\}$, $\{1|3|3\}$. By Hardy's Theorem,

$$\begin{aligned} D^{102}g(h(x,y,z)) = {} & DgD^{102}h \\ & + 2D^2gD^{101}hD^{001}h \\ & + D^2gD^{100}hD^{002}h \\ & + D^3gD^{100}h(D^{001}h)^2, \end{aligned}$$

where function arguments have been supressed to avoid a cluttered notation. For the particular case that $g(\cdot) = \log(\cdot)$ and $h(t) = M_X(t)$ we obtain: $\left.\frac{d\log M_X(t)}{dM_X(t)}\right|_{t=0} =$

1 , $\left.\frac{d^2 \log M_X(t)}{dM_X(t)^2}\right|_{t=0} = -1$ and $\left.\frac{d^3 \log M_X(t)}{dM_X(t)^3}\right|_{t=0} = 2$. We may conclude that

$$\kappa_{102} = m_{102} - 2m_{101}m_{001} - m_{100}m_{002} + 2m_{100}m_{001}^2,$$

as claimed in Table 2.1 up to relabelling of the variables.

If the index vector k is binary the multiset associated with k is degenerate and $c(\pi) = 1$. Equation (2.3) simplifies to

$$\kappa_k = \sum_{\pi \in \Pi(k)} (-1)^{(|\pi|-1)} (|\pi| - 1)! \prod_{j=1}^{|\pi|} m_{\nu_{M_j}}.$$

Example 14 (Square free cumulants)**.** Consider the cumulants in Table 2.1 with binary index vector: $\kappa_{100}, \kappa_{110}$ and κ_{111}.

2.3 Differential moments and differential cumulants

This section investigates properties of moments and cumulants in local neighbourhoods. One reason for studying localised properties of random variables is that global measures do not exist or, if existent, give a misleading picture. A prime example is the linear correlation between two variables which have a non-linear dependence globally.

Another reason is that one may only be interested in particular events rather than the whole distribution or some global measure of it like the expectation of a random variable. Insurance contracts, for instance, typically specify that one party receives tail risks in exchange for regular payments contingent on tail events not happening. Insurers take a natural interest in the tails of distributions and, in particular, in the local dependence of several tail events.

Local moments have been suggested by Mueller and Yan (2001) who also consider a local covariance. They provide formulae for moments, when the order

vector k holds no more than two odd components. We extend their approach to multivariate moments and cumulants of any order. Formulae for local moments and cumulants are derived and a Taylor expansion of the local moment generating function is given provided its global counterpart exists.

The key quantity we derive are *differential cumulants* at a specific point $\xi \in \mathbb{R}^d$. In the bivariate case, $d = 2$, they take the form

$$\frac{\partial^2 \log f(x, y)}{\partial x \partial y}.$$

This quantity was first investigated by Holland and Wang (1987). Later, Jones et al. (1996) referred to it as the *local dependence function*. The local dependence function has some remarkable properties. It vanishes if and only if X and Y are independent, it is constant if f is the bivariate normal density, and it is margin-free in a sense that multiplying f by the ratio of one marginal density over another marginal density leaves the local dependence function unaffected.

For a strictly positive edge length $\varepsilon \in \mathbb{R}$, let $A(\xi, \varepsilon) := \prod_{i=1}^{d} [\xi_i - \frac{\varepsilon}{2}, \xi_i + \frac{\varepsilon}{2}]$ denote the hyper cube centralised at ξ. Let $|A| = \varepsilon^d$ denote its volume. The density of a random variable $X \in \mathbb{R}^d$ conditional on being in A is given by

$$f_X^A(x) = \frac{f_X(x) \mathbb{1}_A(x)}{P(X \in A)}.$$

We define local moments as moments of X conditional on X being in A:

Definition 5 (Local moment)**.** Given a point $\xi \in \mathbb{R}^d$ with neighbourhood A, the local moment $m^A_{k_1 \dots k_d}$ of order k is defined as

$$m^A_{k_1 \dots k_d} = \mathbb{E}\left(\prod_{i=1}^{d} (X_i - \xi_i)^{k_i} \,\middle|\, X \in A \right).$$

The centralisation about ξ implies that the local moment captures the direction of near data rather than their absolute value. This is necessary since we are ultimately interested in the limiting process as the size of the window A approaches

zero. Without the centralisation the conditional density would collapse to the Dirac measure with all probability mass at the trivial local moment ξ.

Let $\mathbb{N}$ denote the integers strictly greater than zero and $2\mathbb{N} := \{n \in \mathbb{N} | \exists m \in \mathbb{N} : n = 2m\}$ and $2\mathbb{N}+1 := \{n \in \mathbb{N} | \exists m \in \mathbb{N} : n = 2m+1\}$ the set of positive even and odd integers respectively. Note that zero is excluded from the even integer set $2\mathbb{N}$.

For symmetry reasons, even and odd elements of the order vector k have different effects on local moments. This motivates the following definition:

$$|\alpha|^+ := |\alpha| + \sum_{i=1}^{d} \mathbb{1}(\alpha_i \in 2\mathbb{N}+1).$$

The operator $|\cdot|^+$ increments the total sum of the components of a vector by one for each odd component.

It will be a useful convention, to define the product over an empty set as 1. For instance,

$$\prod_{\substack{i=1,\\ k_i \in 2\mathbb{N}}}^{d} \frac{1}{k_i+1} = 1$$

if k holds odd order terms only.

Theorem 2 (Local moments)**.** *Let $X \in \mathbb{R}^d$ be an absolutely continuous random vector with density f_X which is d times differentiable in $\xi \in \mathbb{R}^d$. Let $k \in \mathbb{N}^d$ determine the order of moment. Then, for $|A|$ sufficiently small, X has local moment*

$$m^A_{k_1 \dots k_d} = r(\varepsilon, k)\left(\frac{D^\alpha f_X(\xi)}{f_X(\xi)} + O(\varepsilon^2)\right) \tag{2.4}$$

where $r(\varepsilon, k) := \varepsilon^{|k|^+} \prod_{\substack{i=1,\\ k_i \in 2\mathbb{N}}}^{d} \frac{1}{k_i+1} \prod_{\substack{i=1,\\ k_i \in 2\mathbb{N}+1}}^{d} \frac{1}{k_i+2}$ *and* $\alpha := \sum_{\substack{i=1,\\ k_i \in 2\mathbb{N}+1}}^{d} e_i$.

Proof. Consider

$$m^A_{k_1 \dots k_d} = \frac{\int_A \prod_{i=1}^d (x_i - \xi_i)^{k_i} f_X(x) dx}{\int_A f_X(x) dx}. \tag{2.5}$$

Since f_X is d times differentiable in ξ, it has a Taylor expansion of order d about ξ:

$$f_X(x) = f_X(\xi) + \sum_{i=1}^{d}(x_i - \xi_i)\left.\frac{\partial f_X(x)}{\partial x_i}\right|_{x=\xi} + \sum_{i=1}^{d}\sum_{j>i}^{d}(x_i - \xi_i)(x_j - \xi_j)\left.\frac{\partial^2 f_X(x)}{\partial x_i \partial x_j}\right|_{x=\xi} + \frac{1}{2}\sum_{i=1}^{d}(x_i - \xi_i)^2\left.\frac{\partial^2 f_X(x)}{\partial x_i^2}\right|_{x=\xi} + \ldots + o(|\varepsilon|^d).$$

Hence, we may expand f_X in (2.5) and change the order of summation and integration. A term involving an odd order in at least one component is point symmetric in that component about the origin, so that the integral vanishes.

The smallest non-vanishinig order term in the expansion of f_X is $D^\alpha f_X(\xi)$ since $x_i^{k_i}$ is multiplied by x_i whenever k_i is odd, and $x_i^{k_i}$ is multiplied by one whenever k_i is even. The expression for $r(\varepsilon, k)$ follows from polynomial integration. Thus, the numerator of (2.5) can be written as:

$$\left(\prod_{\substack{i=1,\\ k_i\in 2\mathbb{N}}}^{d}\frac{1}{k_i+1}\prod_{\substack{i=1,\\ k_i\in 2\mathbb{N}+1}}^{d}\frac{1}{k_i+2}\right)\varepsilon^d\varepsilon^{|k|^+}(D^\alpha f_X(x) + O(\varepsilon^2)).$$

The denominator of (2.5) can be interpreted as the particular moment with $k = 0 \in \mathbb{N}_0^d$. It simplifies to $\varepsilon^d(f_X(\xi) + O(\varepsilon^2))$, which completes the proof. □

Example 15 (Local moment m_{120})**.** Consider a trivariate random variable X with local moment $m_{120}^A = E((X_1 - \xi_1)(X_2 - \xi_2)^2 | X \in A)$. Then $r(\varepsilon, k) = \frac{\varepsilon^4}{9}, \alpha := (1, 0, 0)'$ and we obtain

$$m_{120}^A = \frac{\varepsilon^4}{9}\left.\frac{\frac{\partial f(x_1,x_2,x_3)}{\partial x_1}}{f(x_1, x_2, x_3)}\right|_{x=\xi} + O(\varepsilon^6).$$

A natural way to extend the concept of a local moment is to consider the limiting case that $|A| \to 0$. Meaningful convergence occurs only if the limiting process is rate adjusted. This leads to the definition of a differential moment.

Definition 6 (Differential moment)**.** The differential moment of an absolutely continuous random vector $X \in \mathbb{R}^d$ at ξ is defined as:

$$m^{\xi}_{k_1 \dots k_d} := \lim_{\varepsilon \to 0} \frac{m^A_{k_1 \dots k_d}}{r(\varepsilon, k)},$$

where $r(\varepsilon, k)$ as defined in (2.4).

This definition of a *differential* moment coincides with the quantity Mueller and Yan (2001) term a *local* moment up to the normalising constant

$$\prod_{\substack{i=1,\\ k_i \in 2\mathbb{N}}}^{d} \frac{1}{k_i + 1} \prod_{\substack{i=1,\\ k_i \in 2\mathbb{N}+1}}^{d} \frac{1}{k_i + 2}.$$

The next corollary is an immediate consequence of Theorem 2. It shows the remarkably simple form of differential moments.

Corollary 2 (Differential moment)**.** *For the differential moment of order $k \in \mathbb{N}^d$ at $\xi \in \mathbb{R}^d$, it holds that*

$$m^{\xi}_{k_1 \dots k_d} = \frac{D^{\alpha} f_X(\xi)}{f_X(\xi)}.$$

Proof. According to Theorem 2, the local moment $m^A_{k_1 \dots k_d}$ can be written as

$$m^A_{k_1 \dots k_d} = r(\varepsilon, k) \left(\frac{D^{\alpha} f_X(\xi)}{f_X(\xi)} + O(\varepsilon^2) \right).$$

The differential moment was defined, so that the $r(\varepsilon, k)$ terms disappear:

$$m^{\xi}_{k_1 \dots k_d} := \lim_{\varepsilon \to 0} \frac{m^A_{k_1 \dots k_d}}{r(\varepsilon, k)} = \lim_{\varepsilon \to 0} \left(\frac{D^{\alpha} f_X(\xi)}{f_X(\xi)} + O(\varepsilon^2). \right)$$

The remainder term of order $O(\varepsilon^2)$ disappears as ε, the edge length of the window A, approaches zero. □

Remark 1 (Interpretation of differential moments). Corollary 2 entails an interesting interpretation of the differential moment, based on elementary calculus: It

is the relative change of the density f_X as variables change about ξ. The differential moment tells us how the relative probability mass changes near ξ. If $m^{\xi}_{k_1 \dots k_d}$ is positive (negative), then relatively more probability mass can be found in the (opposite) direction of the changing variables. If $m^{\xi}_{k_1 \dots k_d}$ is high (low) in absolute value, then the probability mass changes more rapidly (slowly).

From (2.4) it is clear that the choice α in the derivative $D^{\alpha} f_X$ depends only on the pattern of odd and even components of the moment. To be precise, α holds a one corresponding to odd components and a zero corresponding to even components. Consequently, the differential moment $m^{\xi}_{k_1 \dots k_d}$ depends on k only in as much as the pattern of odd and even in k is concerned.

This suggests defining an equivalence relation meaning *same differential moment* on $\mathbb{N}^d \times \mathbb{N}^d$: For $u, k \in \mathbb{N}^d$ set

$$u \sim_m k \iff m_{u_1 \cdots u_d} = m_{k_1 \cdots k_d}.$$

The relation $\sim_m$ partitions the product space $\mathbb{N}^d \times \mathbb{N}^d$ into 2^d equivalence classes of same differential moments. The graph corresponding to $\sim_m$ in the bivariate case is depicted in Figure 2.3. The axes show the order of the moments. Each equivalence class is depicted with a different symbol. For instance, $(2,2) \sim_m (4,2)$ since $m^{\xi}_{22} = m^{\xi}_{42}$. Note that $u \sim_m k \iff |u - k| \in 2\mathbb{N}$.

We can define a local moment generating function as

$$M^{A}_{X}(t) := \mathbb{E}(e^{t'X} | X \in A).$$

As an integral of a continuous function over a closed hypercubed the local moment

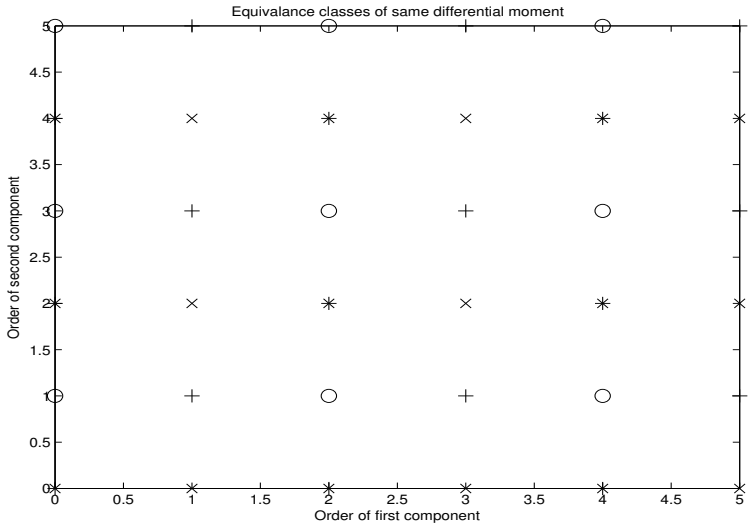

Figure 2.3 – The diagram shows the graph of the equivalence relation $\sim_m$ for the bivariate case.

generating function always exists. We have the following expansion:

$$
\begin{aligned}
M_X^A(t) &= \frac{1}{P(X \in A)} \int_A e^{\sum_{i=1}^d t_i x_i} f_X(x) dx \\
&= \frac{1}{P(X \in A)} \int_A \left(\sum_{j=0}^{\infty} \frac{(\sum_{i=1}^d t_i x_i)^j}{j!} \right) \\
&\left(f_X(\xi) + \sum_{i=1}^d (x_i - \xi_i) \left. \frac{\partial f_X(x)}{\partial x_i} \right|_{x=\xi} + \sum_{i=1}^d \sum_{j>i}^d (x_i - \xi_i)(x_j - \xi_j) \left. \frac{\partial^2 f_X(x)}{\partial x_i \partial x_j} \right|_{x=\xi} \right. \\
&\left. + \frac{1}{2} \sum_{i=1}^d (x_i - \xi_i)^2 \left. \frac{\partial^2 f_X(x)}{\partial x_i^2} \right|_{x=\xi} + \ldots + O(\varepsilon^d) \right) dx \\
&= 1 + \sum_{i=1}^d t_i \left. \frac{\partial f_X(x)}{\partial x_i} \right|_{x=\xi} \left(\frac{\varepsilon^2}{3 f_X(\xi)} + O(\varepsilon^4) \right) \\
&+ \sum_{i=1}^d t_i^2 \left. \frac{\partial^2 f_X(x)}{\partial x_i^2} \right|_{x=\xi} \left(\frac{\varepsilon^2}{6 f_X(\xi)} + O(\varepsilon^4) \right) \\
&+ \sum_{i=1}^d \sum_{j>i}^d t_i t_j \left. \frac{\partial^2 f_X(x)}{\partial x_i \partial x_j} \right|_{x=\xi} \left(\frac{\varepsilon^4}{9 f_X(\xi)} + O(\varepsilon^6) \right) + O(\varepsilon^4 \left| t^3 \right|).
\end{aligned}
$$

The local moments can be computed from the local moment generating function

via differentiation to appropriate order and evaluation at $t = 0$ as was demonstrated in Example 15. The natural logarithm of the local moment generating function defines the local cumulant generating function $K_X^A(t) : \mathbb{R}^d \longrightarrow \mathbb{R}$:

$$K_X^A(t) := \log(M_X^A(t)).$$

Corollary 3 (Local cumulants)**.** *Under the conditions of Theorem 2, it holds for the local cumulants that*

$$\kappa_k^A = \sum_{\pi \in \Pi(k)} c(\pi)(-1)^{(|\pi|-1)}(|\pi| - 1)! \prod_{j=1}^{|\pi|} m_{\nu_{M_j}}^A,$$

Proof. The proof is the same as in Corollary 1. □

We next define differential cumulants at ξ. There are two natural ways of doing this. We may define a differential cumulant as the limiting quantity of a local cumulant as the size of the conditioning window A approaches zero. Alternatively, we may take a series of differential moments and require that the ex-log-relation between moments and cumulants is preserved in the differential case. As is demonstrated in Theorem 4, the limiting local cumulant is in general not equal to the ex-log relation induced counterpart of the differential moments. They coincide exactly in the square-free case. In order to maintain the defining relation between cumulants and moments, we define differential cumulants in terms of differential moments.

Definition 7 (Differential cumulant)**.** For an index vector $k \in \mathbb{N}^d$, the differential cumulant at $\xi \in \mathbb{R}^d$ is defined as

$$\kappa_k^\xi := \sum_{\pi \in \Pi(k)} c(\pi)(-1)^{(|\pi|-1)}(|\pi| - 1)! \prod_{j=1}^{|\pi|} m_{\nu_{M_j}}^\xi.$$

The next theorem shows the remarkably simple form of the differential cumulants.

Theorem 3 (Differential cumulant). *For a differential cumulant at $\xi \in \mathbb{R}^d$ of order $k \in \mathbb{N}^d$ it holds that*

$$\kappa^{\xi}_{k_1 \dots k_d} = D^{\alpha} \log(f_X(\xi)).$$

Proof. By Hardy's formula

$$\begin{aligned} D^{\alpha} \log(f_X(\xi)) &= \sum_{\pi \in \Pi(k)} c(\pi)(-1)^{(|\pi|-1)}(|\pi|-1)! \frac{1}{f_X^{(|\pi|)}(\xi)} \prod_{j=1}^{|\pi|} D^{\nu_{M_j}} f_X(\xi) \\ &= \sum_{\pi \in \Pi(k)} c(\pi)(-1)^{(|\pi|-1)}(|\pi|-1)! \prod_{j=1}^{|\pi|} \frac{D^{\nu_{M_j}} f_X(\xi)}{f_X(\xi)} \\ &= \sum_{\pi \in \Pi(k)} c(\pi)(-1)^{(|\pi|-1)}(|\pi|-1)! \prod_{j=1}^{|\pi|} m^{\xi}_{\nu_{M_j}} \\ &= \kappa^{\xi}_{k_1 \dots k_d}. \end{aligned}$$

Note that, in the first line, $f_X^{(|\pi|)}(\xi)$ denotes $f_X(\xi)$ raised to the power of $|\pi|$ rather than the derivative of order $|\pi|$. □

The next theorem relates a differential cumulant to the limit of a local cumulant.

Theorem 4 (Differential and limiting local cumulant). *A differential cumulant κ^{ξ}_k is the limit of the local cumulant $lim_{\varepsilon \to 0} \frac{1}{r(\varepsilon,k)} \kappa^A_k$ if and only if k is binary, i.e. κ_k is a square-free cumulant.*

Proof. First, let $k \in \{0,1\}^d$. Let π be a partition of the lattice corresponding to k. The key is to show that the contribution from π converges at the same rate as the local cumulant, i.e $r(\varepsilon, k) = \prod_{j=1}^{|\pi|} r(\varepsilon, \nu_{M_j})$.

Since k is binary, so is the family of multiplicities $(\nu_{M_j})_{1 \le j \trianglelefteq |\pi|}$ corresponding to π. Since π is a partition, $k = \sum_{j=1}^{|\pi|} \nu_{M_j}$. Hence,

$$k_i = 1 \Leftrightarrow \text{there exists exactly one } j \text{ such that } \nu_{M_j}(i) = 1. \tag{2.6}$$

Let $M := \sum_{i=1}^{d} \mathbb{1}(k_i = 1)$ be the number of odd components in k. By (2.6) $M = \sum_{j=1}^{|\pi|} \sum_{i=1}^{d} \mathbb{1}(\nu_{M_j}(i) = 1)$, that is, the total number of odd components is not changed through a partition of a binary vector. This allows us to write

$$\begin{aligned} |k|^+ &= \left| \sum_{j=1}^{|\pi|} \nu_{M_j} \right|^+ = \left| \sum_{j=1}^{|\pi|} \nu_{M_j} \right| + \sum_{i=1}^{d} \mathbb{1}\left(\left(\sum_{j=1}^{|\pi|} \nu_{M_j} \right)_i = 1 \right) \\ &= \sum_{j=1}^{|\pi|} \left| \nu_{M_j} \right| + M = \sum_{j=1}^{|\pi|} \left| \nu_{M_j} \right|^+ . \end{aligned} \tag{2.7}$$

Similarly,

$$\begin{aligned} \prod_{\substack{i=1,\\ k_i=0}}^{d} \frac{1}{k_i+1} \prod_{\substack{i=1,\\ k_i=1}}^{d} \frac{1}{k_i+2} &= 3^{-M} \\ &= \prod_{j=1}^{|\pi|} \Bigg(\prod_{\substack{i=1,\\ \sum_{j=1}^{|\pi|} \nu_{M_j}(i)=0}}^{d} \frac{1}{\sum_{j=1}^{|\pi|} \nu_{M_j}(i) + 1} \prod_{\substack{i=1,\\ \sum_{j=1}^{|\pi|} \nu_{M_j}(i)=1}}^{d} \frac{1}{\sum_{j=1}^{|\pi|} \nu_{M_j}(i) + 2} \Bigg) \end{aligned} \tag{2.8}$$

Together, (2.7) and (2.8) imply that $r(\varepsilon, k) = \prod_{j=1}^{|\pi|} r(\varepsilon, \nu_{M_j})$. We thus have

$$\begin{aligned} \frac{1}{r(\varepsilon,k)} \kappa_k^A &= \frac{1}{r(\varepsilon,k)} \sum_{\pi \in \Pi(k)} c(\pi) (-1)^{(|\pi|-1)} (|\pi| - 1)! \prod_{\substack{j=1,\\ M_j \in \pi}}^{|\pi|} r(\varepsilon, \nu_{M_j}) \left(\frac{D^{\nu_{M_j}} f_X(\xi)}{f_X(\xi)} + O(\varepsilon^2) \right) \\ &= \sum_{\pi \in \Pi(k)} (-1)^{(|\pi|-1)} (|\pi| - 1)! \prod_{\substack{j=1,\\ M_j \in \pi}}^{|\pi|} \frac{D^{\nu_{M_j}} f_X(\xi)}{f_X(\xi)} + O(\varepsilon^2). \end{aligned} \tag{2.9}$$

Now take limits as $\varepsilon \to 0$ to obtain

$$\begin{aligned} \lim_{\varepsilon \to 0} \frac{1}{r(\varepsilon,k)} \kappa_k^A &= \sum_{\pi \in \Pi(k)} c(\pi) (-1)^{(|\pi|-1)} (|\pi| - 1)! \prod_{j=1}^{|\pi|} m^{\xi}_{\nu_{M_j}} \\ &= \kappa_k^{\xi}, \end{aligned}$$

which shows that the local cumulant κ_k^A converges to the differential cumulant κ_k^{ξ} if k is binary.

Conversely, suppose k is not binary. We show that $\frac{1}{r(\varepsilon,k)}\kappa_k^A$ converges to a quantity which is not a linear combination of products of differential moments. First note that differential moments are proportional to $D^\alpha f_X(\xi)$ for some binary α. Take π to be the degenerate partition, i.e. $|\pi| = 1$. Then π holds only one set with multiplicity $b = k$. The associated quantity with this partition in (2.9) converges to $c\frac{D^b f_X(\xi)}{f_X(\xi)}$ for some constant $c \in \mathbb{R}$. The multiplicity b not being binary, this cannot be a local moment. □

Example 16 (Square-free differential cumulant)**.** Consider the square-free differential cumulant κ^ξ_{11011} and take, for illustration purposes, the partition $\{10001|01000|00010\}$. Table 2.2 shows the computation of $r(\varepsilon, k)$ and $\prod_{j=1}^3 r(\varepsilon, \nu_{M_j})$, where $M_1 = \{x_1, x_5\}, M_2 = \{x_2\}, M_3 = \{x_4\}$. Since k is binary, the associated multiset $M = \{x_1, x_2, x_4, x_5\}$ is degenerate and each variable in M appears exactly once in either M_1, M_2 or M_3. This implies that each one in k corresponds to exactly one unity in ν_{M_1}, ν_{M_2} or ν_{M_3}. Simple calculations show that $r(\varepsilon, k) = \frac{\varepsilon^8}{81} = \prod_{j=1}^3 r(\varepsilon, \nu_{M_j})$. Incidentally, this partition adds the term $2m_{10001}m_{01000}m_{00010}$ to κ_{11011}. Summing over all partitions of M in similar manner, the expression for κ_{11011} is readily verified as:

$$
\begin{aligned}
\kappa^\xi_{11011} = {} & m^\xi_{11011} \\
& - m^\xi_{11010}m^\xi_{00001} - m^\xi_{11001}m^\xi_{00010} - m^\xi_{10011}m^\xi_{01000} - m^\xi_{01011}m^\xi_{10000} \\
& - m^\xi_{11000}m^\xi_{00011} - m^\xi_{10010}m^\xi_{01001} - m^\xi_{10001}m^\xi_{01010} \\
& + 2(m^\xi_{11000}m^\xi_{00010}m^\xi_{00001} + m^\xi_{10010}m^\xi_{01000}m^\xi_{00001} + m^\xi_{10001}m^\xi_{01000}m^\xi_{00010} \\
& \quad + m^\xi_{01010}m^\xi_{10000}m^\xi_{00001} + m^\xi_{01001}m^\xi_{10000}m^\xi_{00010} + m^\xi_{00011}m^\xi_{10000}m^\xi_{01000}) \\
& - 6m^\xi_{10000}m^\xi_{01000}m^\xi_{00010}m^\xi_{00001}.
\end{aligned}
$$

Multiplicity	x_1	x_2	x_3	x_4	x_5	$\lvert\cdot\rvert^+$	$\prod_{\substack{i=1,\\ k_i\in 2\mathbb{N}}}^{d} \frac{1}{k_i+1} \prod_{\substack{i=1,\\ k_i=1}}^{d} \frac{1}{k_i+2}$	$r(\varepsilon,\cdot)$
k	1	1	0	1	1	8	$\frac{1}{3}^4$	$\frac{\varepsilon^8}{81}$
ν_{M_1}	1	0	0	0	1	4	$\frac{1}{3}^2$	
ν_{M_2}	0	1	0	0	0	2	$\frac{1}{3}$	$\frac{\varepsilon^8}{81}$
ν_{M_3}	0	0	0	1	0	2	$\frac{1}{3}$	

Table 2.2 – Convergence rates of a binary differential cumulant.

2.4 Conclusion

This chapter shed some light on the relation between moments and cumulants. A multivariate higher order derivative chain function was introduced which allows us to calculate multivariate cumulants of any order from moments. Differential cumulants were defined via differential moments. It was shown that the differential cumulant of order k at ξ takes the form

$$\kappa_k^{\xi} = D^k \log f_X(\xi).$$

It is the single most important quantity of this thesis. The next chapter shows that conditional independence statements can be captured through *pairwise zero-cumulants*, binary differential cumulants which hold exactly two ones and vanish everywhere.

Chapter 3

Conditional independence and hierarchical models

3.1 Introduction

In the previous chapter the concept of differential cumulants was introduced. This chapter explains the relations between sets of *zero-cumulants*, conditional independence statements and hierarchical models, a subclass of which is the class of graphical models. The first part of the chapter covers a brief introduction to conditional independence and graphical models. Naturally, many of the results, or similar versions of them, are well established. Textbook references are Whittaker (1990) and Lauritzen (1996). The second part of the chapter covers novel ideas. It is dedicated to specific model classes which are identified via sets of zero-cumulants.

Pairwise zero-cumulants are binary differential cumulants which hold exactly two ones and vanish everywhere. Section 3.2 relates pairwise zero-cumulants to conditional independence statements. Lemma 3 shows that a vanishing pairwise cumulant implies the factorisation of the density and vice versa. To the best of our knowledge, this link has not yet been exploited for nonparametric estimation of conditional independence as we suggest in Chapter 5. It is even more powerful as arbitrary conditional independence structures can be expressed via sets of pairwise zero-cumulants, as Lemma 4 demonstrates.

Conditional independence statements are well known to underlie the theory of graphical models which are introduced in Section 3.3. Graphical models incorporate conditional independence statements via Markov properties of the density. This ensures that graph separation is isomorphic to conditional independence statements. This thesis only deals with undirected graphs. Conditional independence models in directed graphs have, for instance, been investigated by Spiegelhalter et al. (1993) and Settimi and Smith (2000).

Graphical models have been most studied in the continuous case when the density is multivariate normal. In the normal case the entire conditional independence structure is captured in the inverse covariance matrix. Consequently, continuous graphical models have been termed covariance selection models (Dempster, 1972). In contrast, our approach makes no specific distribution assumption. As a result, we are not concerned with parameter estimation and our only goal is to model the interaction between random variables.

Graphical models form a subclass of the class of hierarchical models. Hierarchical models are characterised by the fact that interaction at lower levels implies interaction at higher levels. They are explained in Section 3.4. It is shown how particular model classes can be obtained through imposing restrictions on the differential cumulants.

3.2 Conditional independence

Conditional independence is a statistical concept applicable to a minimum of three variables. As in Chapter 2, we assume that the $d-$variate random variable X has a density f_X which is strictly positive everywhere. Let the integer sets I, J and K partition $[d] := \{1, ..., d\}$. We write

$$X_I \perp\!\!\!\perp X_J | X_K$$

to denote that X_I is independent of X_J for any given value that X_K may take. We may express conditional independence in terms of densities through

$$f_{X_I|X_J,X_K} = f_{X_I|X_K}, \tag{3.1}$$

where the equality is understood to hold for any value $x \in \mathbb{R}^d$. Intuitively, (3.1) says that, once X_K is fixed, the density of X_I is independent of X_J. Hence, the probabilty of X_I falling into a measurable set is the same for all values of X_J given X_K.

The importance of modelling conditional associations can be demonstrated through Simpson's paradox. Simpson's paradox prevails in situations where correlations between a random variable Y and two subgroups X_1 and X_2 are reversed when X_1 and X_2 are combined. A classic example is the Berkeley admission paradox (Bickel et al., 1975). It is so instructive that we reproduce the key data and findings in Example 17.

Example 17 (Simpson's paradox)**.** Admission numbers to Berkeley Graduate School in 1973 were 8442 and 4321 for males and females respectively. These corresponded to admission rates of 44 per cent for male applicants and 35 per cent for female applicants. Given the sample sizes a statistically significant discrimination based on gender seems apparent.

The admission rates, however, differ greatly between different departments independent of gender. Indeed, the apparent discrimination is reversed at the departmental level with females having statistically significantly better chances of admission in many departments.

Table 3.1 shows the admission rates and number of applicants for the six largest departments. The paradox can be resolved when the number of applications to different departments is taken into account. A much higher proportion of male candidates applied to departments with high admission rates. This resulted in a higher overall admission rate for males compared to females. The effect could

Department	Male		Female	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6 %	341	7 %

Table 3.1 – Admission rates to the Berkeley Graduate School (1973) for the six largest departments.

have occured even if male admission rates had been lower than female admission rates in every single department.

Simpson's paradox illustrates how modelling conditional associations can reveal relations that are hidden in marginal observations. In the example, females seem to have higher admission rates conditional on considering some particular departments. The marginalisation process of summing up admission numbers across all departments results in a misleading picture of a gender bias in favour of male applicants.

In this section, we connect the theory of differential cumulants with conditional independence structures. It is a key observation that setting *pairwise differential cumulants* equal to zero everywhere allows us to express conditional and unconditional dependency structures.

Definition 8 (Pairwise cumulant)**.** A cumulant κ_k is *pairwise* if k is binary and holds exactly two ones, i.e. pairwise cumulants take the form κ_k, $k = e_i + e_j$ for some $(i, j) \in [d] \times [d], i \neq j$.

Definition 9 (Zero cumulant)**.** A *zero-cumulant* is a differential cumulant which vanishes everywhere.

Lemma 2 (Independence in the bivariate case). *Let* $X \in \mathbb{R}^2$. *Then* $X_1 \perp\!\!\!\perp X_2 \iff \kappa^x_{11} = 0, \quad \forall x \in \mathbb{R}^2$.

Proof. The proof follows from straightforward integration:

$$0 = \kappa^x_{11} = \frac{\partial^2}{\partial x_1 \partial x_2} \log(f_{X_1,X_2}(x_1, x_2)) \iff f_{X_1,X_2}(x_1, x_2) = e^{h_1(x_1)+h_2(x_2)}$$

for some functions $h_1, h_2 : \mathbb{R} \to \mathbb{R}$. □

In the multivariate case we can express pairwise conditional independence given the remaining variables through setting the associated pairwise differential cumulants equal to zero everywhere.

Lemma 3 (Conditional independence of two random variables). *Let* $X \in \mathbb{R}^d$. *Then*

$$X_i \perp\!\!\!\perp X_j | X_{-ij} \iff \kappa^x_{e_i+e_j} = 0, \quad \forall x \in \mathbb{R}^d,$$

where

$$X_{-ij} := (X_1, ..., X_{i-1}, X_{i+1}, ..., X_{j-1}, X_{j+1}, ..., X_d).$$

Proof. The proof proceeds in analogy to the bivariate case and is omitted. □

Setting several pairwise differential cumulants to zero simultaneously allows us to express arbitrary conditional independence statements.

Lemma 4 (Multivariate conditional independence). *Given three index sets* I, J, K *which partition* $[d]$, *let* $S = \{e_i + e_j, i \in I, j \in J\}$. *Then*

$$X_I \perp\!\!\!\perp X_J | X_K \iff \kappa^x_k = 0 \text{ for all } k \in S \text{ and for all } x \in \mathbb{R}^d.$$

Proof. From Lemma 3 it is clear, that this is equivalent to the conditional independence statement

$$X_I \perp\!\!\!\perp X_J | X_K \iff X_i \perp\!\!\!\perp X_j | X_{-ij} \; \forall (i, j) \in I \times J.$$

Sufficiency ($\Rightarrow$) and necessity ($\Leftarrow$) are semi-graphoid and graphoid axioms referred to as weak union and intersection respectively. Both hold true for strictly positive densities (see for instance Cozman and Walley, 2005). □

Example 18. Consider the random variables X_1, X_2, X_3 and X_4. Let $I = \{1,2\}, J = \{4\}, K = \{3\}$. Lemma 4 states that

$$(X_1, X_2) \perp\!\!\!\perp X_4 | X_3 \iff \kappa^x_{1001} = 0 \text{ and } \kappa^x_{0101} = 0 \text{ everywhere.}$$

By Lemma 3

$$\kappa^x_{1001} = 0, \text{ for all } x \in \mathbb{R}^4 \iff X_1 \perp\!\!\!\perp X_4 | (X_2, X_3)$$

and

$$\kappa^x_{0101} = 0, \text{ for all } x \in \mathbb{R}^4 \iff X_2 \perp\!\!\!\perp X_4 | (X_1, X_3).$$

The weak union property of conditional independence states that

$$(X_1, X_2) \perp\!\!\!\perp X_4 | X_3 \implies X_1 \perp\!\!\!\perp X_4 | (X_2, X_3)$$

and, by symmetry,

$$(X_1, X_2) \perp\!\!\!\perp X_4 | X_3 \implies X_2 \perp\!\!\!\perp X_4 | (X_1, X_3).$$

which proves sufficiency. The intersection property states that

$$X_2 \perp\!\!\!\perp X_4 | (X_1, X_3) \text{ and } X_1 \perp\!\!\!\perp X_4 | (X_2, X_3) \implies (X_1, X_2) \perp\!\!\!\perp X_4 | X_3, \tag{3.2}$$

proving necessity.

We prove the intersection property directly for this example. The left hand side of (3.2) can be translated into the density statements

$$\frac{f_{1234}}{f_{134}} = \frac{f_{123}}{f_{13}} \quad \text{and} \quad \frac{f_{1234}}{f_{234}} = \frac{f_{123}}{f_{13}}. \tag{3.3}$$

By equating the two expressions and integrating out X_2, we obtain $\frac{f_{134}}{f_{34}} = \frac{f_{13}}{f_3}$. Substituting this expression back into (3.3) yields

$$\frac{f_{1234}}{f_{123}} = \frac{f_{34}}{f_3}, \quad \text{or} \quad (X_1, X_2) \perp\!\!\!\perp X_4 | X_3$$

as required. The key insight of the proof is that the intersection property depends on the positivity assumption of the density.

Lemma 4 allows us to reduce complex conditional independence statements to a joint set of pairwise zero-cumulants. This reduction of complexity plays an important role in the estimation of conditional independence structures. Chapter 5 illustrates how nonparametric estimation techniques can be used to estimate conditional independence pairwise. Lemma 4 shows that this is sufficient in order to estimate arbitrary conditional independence statements.

The next theorem shows that the random variables $X_1, ..., X_d$ are independent if and only if pairwise conditional independence holds for every pair. Put differently, independence holds if and only if any permutation of $k = (1, 1, 0, ..., 0)$ leads to a zero-cumulant κ_k.

Theorem 5 (All-pairwise conditional independence if and only if independence)**.** *The random variables $X_1, ..., X_d$ are independent if and only if $\kappa^x_{e_i+e_j} = 0$, for all $(i, j) \in [d]^2$, $i \neq j$, and for all $x \in \mathbb{R}^d$.*

Proof. Sufficiency ($\Rightarrow$) follows from differentiation of the log-density. Necessity ($\Leftarrow$) can be proved by induction on the number of variables n. The statement is true for $n = 2$ by Lemma 2. Let the statement be true for n and let the $\binom{n+1}{2}$ differential cumulants $\kappa^x_{e_i+e_j}$ vanish everywhere, where e_i and e_j are unit vectors in $\mathbb{R}^{n+1}$. We show that X_2 is independent of X_{-2}. This completes the proof since the variables X_{-2} are independent by induction assumption.

Consider $\kappa_{e_1+e_2} = 0$. Integration with respect to x_1 and x_2 yields

$$f_{X_1,...,X_{n+1}}(x_1, ..., x_{n+1}) = e^{h_1(x_{-1})+h_2(x_{-2})} \tag{3.4}$$

for some functions $h_1 : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $h_2 : \mathbb{R}^n \longrightarrow \mathbb{R}$. Now integrate again with respect to x_1 to obtain

$$f_{X_{-1}}(x_{-1}) = e^{h_1(x_{-1})} \int_{\mathbb{R}} e^{h_2(x_{-2})} dx_1.$$

The left hand side is a n-dimensional marginal density which factorises into n marginals by induction assumption: $f_{X_{-1}}(x_{-1}) = \prod_{i=2}^{n+1} f_{X_i}(x_i)$. Thus, $h_1(x_{-1})$ can be split into a sum of two functions, $g_1 : \mathbb{R}^{n-1} \longrightarrow \mathbb{R}$ and $g_2 : \mathbb{R} \longrightarrow \mathbb{R}$, where the latter is a function of x_2 only, i.e. $h_1(x_{-1}) = g_1(x_{-12}) + g_2(x_2)$. Considering (3.4) again, we see that the density $f_{X_1,\ldots,X_{n+1}}$ factorises since

$$f_{X_1,\ldots,X_{n+1}}(x_1, \ldots, x_{n+1}) = e^{g_2(x_2)+g_1(x_{-12})+h_2(x_{-2})}.$$

Hence $X_2 \perp\!\!\!\perp X_{-2}$ as required. □

3.3 Graphical models

The analysis of the last section makes clear that pairwise zero-cumulants are equivalent to independence or conditional independence statements. Conditional independence structures can be represented through graphical models. This section investigates the link between sets of zero-cumulants and graphical models.

Definition 10 (Graph and graph related concepts)**.** A *graph* $\mathcal{G} = (V, E)$ is a pair of a vertex set V and a set of undirected edges E. The vertices represent random variables and the edges represent interactions. A graph is *complete* if all vertices are mutually connected by an edge. A *clique* is a subset of vertices which induces a complete subgraph, i.e. all vertices are mutually connected by an edge. A clique is *maximal* if no further vertex can be added such that the extended set is still a clique.

We restrict our attention to finite, simple graphs. In particular, we do not allow multiple edges between two vertices or an edge between a vertex and itself.

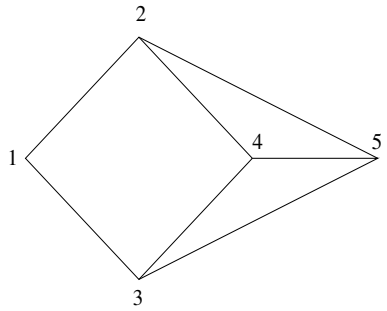

Figure 3.1 – Probabilistic graphical model with three conditional independencies.

Example 19. The graph shown in Figure 3.1 has maximal cliques $\{12, 245, 345, 13\}$. Note that some authors require maximality in their definition of a clique.

A graphical model combines a graph and a collection of random variables in such a way that graph separation corresponds to conditional independence. This can be expressed formally through the *global Markov property*: If a triple (I, J, K) of disjoint subsets of V is such that K separates I from J in $\mathcal{G}$, then it must hold that $X_I \perp\!\!\!\perp X_J | X_K$. Separation means that any path from I to J passes through K. We sometimes use the short notation $I \perp\!\!\!\perp J | K$. We assume for any density/graph pair $(f_X, \mathcal{G})$ that f_X has the global Markov property with respect to $\mathcal{G}$. An important theorem due to Hammersley and Clifford says that f_X, which is assumed to be strictly positive everywhere, factorises over the maximal cliques of $\mathcal{G}$ if it has the global Markov property with respect to $\mathcal{G}$. In that case it can be written as

$$f_X(x) = \prod_{J \in \mathcal{C}} h_J(x_J),$$

where $\mathcal{C}$ denotes the set of maximal cliques of $\mathcal{S}$.

Example 20 (Probabilistic graphical model)**.** Consider a normally distributed random vector $X \in \mathbb{R}^5$ with covariance matrix Σ. It can be shown that $X_i \perp\!\!\!\perp X_j | X_{-ij}$ if and only if $\Sigma^{-1}_{ij} = 0$. The matrix Σ^{-1} is referred to as precision or

influence matrix. Suppose Σ^{-1} is given by

$$\Sigma^{-1} = \begin{pmatrix} * & * & * & 0 & 0 \\ * & * & 0 & * & * \\ * & 0 & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \end{pmatrix},$$

where a star denotes a non-zero entry. This influence matrix entails three conditional independence statements:

$$\begin{aligned} X_1 &\perp\!\!\!\perp X_5 | (X_2, X_3, X_4), \\ X_1 &\perp\!\!\!\perp X_4 | (X_2, X_3, X_5), \\ X_2 &\perp\!\!\!\perp X_3 | (X_1, X_4, X_5). \end{aligned} \tag{3.5}$$

The corresponding graph is depicted in Figure 3.1.

At the core of a graphical model is the edge structure of its graph $\mathcal{G}$. Edges define the set of maximal cliques $\mathcal{C}$, and f_X factorises over the maximal cliques by the Hammersley-Clifford Theorem. In order to expand the class of permissible densities, the next section introduces the concept of a hierarchical model.

3.4 Hierarchical models

3.4.1 Introduction

The class of hierarchical models derives its name from the fact that interaction within a set of random variables implies interaction in any subset, implying a hierarchical interaction structure. For example, a model for three random variables X_1, X_2, X_3 which specifies an interaction between X_1, X_2 and X_3 is hierarchical, if it also has all two way interactions $X_1 - X_2, X_1 - X_3, X_2 - X_3$.

The key theorem of this section shows that the class of hierarchical models is isomorph to certain sets of zero-cumulants. By the nature of isomorphisms, this allows a two-fold interpretation. On the one hand, we may take an arbitrary set of zero-cumulants and investigate the hierarchical model implied by them. On the other hand, we may take a hierarchical model and represent it in terms of zero-cumulants.

The full virtue of this isomorphism unfolds in the next chapter. There, we identify the ideal generated by zero-cumulants with the ideal generated through monomials. Together this gives a bijective representation of hierarchical models through monomial ideals. This bijection bridges the gap between statistics and commutative algebra for continuous densities.

3.4.2 The duality with zero differential cumulants

Let $[d]$ be the vertex set representing the random variables $X_1, ..., X_d$. Hierarchical models were loosely characterised by the fact that interaction within a set of variables implies interaction in any of its subsets. Graphs can only express whether or not two variables are conditionally independent. In order to express higher order interactions, we need to generalise the concept of a graph to *abstract simplicial complexes*.

Definition 11 (Abstract simplicial complex)**.** A collection of subsets of $[d]$ is an *abstract simplicial complex* $\mathcal{S}$ if it is closed under taking subsets, i.e. if $J \in \mathcal{S}$ and $K \subseteq J$ then $K \in \mathcal{S}$.

Definition 12 (Hierarchical model)**.** Given a simplicial complex $\mathcal{S}$ over an index set $[d]$, a *hierarchical model* for the joint distribution function $f_X(x)$ takes the form

$$f_X(x) = \exp\left\{\sum_{J\in\mathcal{S}} h_J(x_J)\right\}, \tag{3.6}$$

where $h_J : \mathbb{R}^J \longrightarrow \mathbb{R}$ and $x_J \in \mathbb{R}^J$ is the canonical projection of $x \in \mathbb{R}^d$ onto the subspace associated with the index set J.

Let $g(x) := \log f_X(x)$ denote the log-density. The hierarchical model for f_X is equivalent to a quasi-additive model for $g(x) = \sum_{J \in \mathcal{S}} h_J(x_J)$, and we also refer to this model for g as being hierarchical.

We introduce a few more quantities related to simplicial complexes. Some of them will only be needed in Chapter 4.

Definition 13 (Face, nonface, facet)**.** The elements of a simplicial complex $\mathcal{S}$ on $[d]$ are called *faces*. *Nonfaces* are subsets of $[d]$ which are not in $\mathcal{S}$. The set of nonfaces is denoted by $\bar{\mathcal{S}}$. Maximal faces under inclusion are called *facets*.

Definition 14 (Dimension of a face)**.** The *dimension of a face* $F \in \mathcal{S}$ is the number of elements of F minus one:

$$\dim F := |F| - 1.$$

Definition 15 (Dimension of a simplicial complex)**.** The *dimension of a simplical complex* $\mathcal{S}$ is the maximum of the dimensions of its faces:

$$\dim \mathcal{S} := \max\{\dim F | F \in \mathcal{S}\}.$$

Definition 16 (Flag simplicial complex)**.** A simplicial complex $\mathcal{S}$ is *flag* if every minimal nonface of $\mathcal{S}$ is a 2-elements subset of $[d]$.

Definition 17 (Pure simplicial complex)**.** A simplicial complex $\mathcal{S}$ is *pure* if all facets are of the same dimension.

Example 21 (Simplicial complex and related quantities)**.** Consider the simplicial complex shown in Figure 3.2. The shading indicates that the face $\{1, 2, 3\}$ is included in $\mathcal{S}$. The full simplicial complex is given by

$$\mathcal{S} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{3, 4\}, \{1, 2, 3\}\}. \tag{3.7}$$

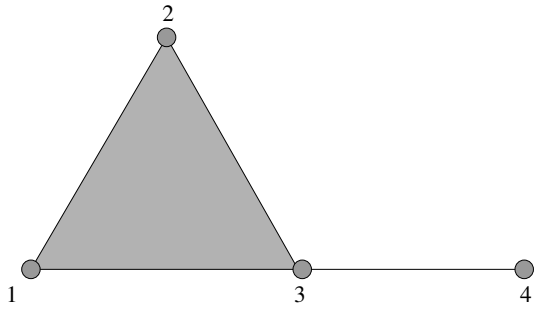

Figure 3.2 – A simplicial complex.

The faces of $\mathcal{S}$ are the elements of $\mathcal{S}$ listed above. Since $\mathcal{S}$ is closed under taking subsets, (3.7) holds redundancies. The generators of $\mathcal{S}$ are the maximal cliques, or facets, $\{1,2,3\}$ and $\{3,4\}$. The definition of a simplicial complex implies that it is completely characterised by its facets.

To simplify the notation, we use $\{123,34\}$ instead of $\{\{1,2,3\},\{3,4\}\}$ whenever we have a collection of at least two sets. Furthermore, we write $\mathcal{S} = \{123,34\}$ to mean that $\mathcal{S}$ is the simplicial complex with facets $\{1,2,3\}$ and $\{3,4\}$. No ambiguity should arise from this convention since $\mathcal{S}$ always denotes a simplicial complex. By definition it holds all subsets of its facets and so the notation cannot be mistaken with a collection of the two sets $\{1,2,3\}$ and $\{3,4\}$.

Vertices have dimension zero, edges have dimension one and the face $\{1,2,3\}$ has dimension two. The dimension of $\mathcal{S}$ is two since the facet $\{1,2,3\}$ has dimension two which is maximal. The set of nonfaces of $\mathcal{S}$ is given by

$$\bar{\mathcal{S}} = \{1234, 234, 134, 124, 14, 24\}. \tag{3.8}$$

Note immediately that $\bar{\mathcal{S}}$ is closed under unions. The minimal nonfaces are $\{1,4\}$ and $\{2,4\}$ since every nonface contains either $\{1,4\}$ or $\{2,4\}$ or both. These particular minimal nonfaces have two elements each, implying that $\mathcal{S}$ is flag. Similarly to writing a simplicial complex $\mathcal{S}$ in terms of its facets, we will write $\bar{\mathcal{S}}$ in terms of the minimal nonfaces only. It is understood that $\bar{\mathcal{S}}$ holds all subsets of $[d]$ of which at least one nonface is a subset. For instance, if $d = 4$ then $\bar{\mathcal{S}} = \{14, 24\}$

implies that any set from (3.8) is also contained in $\bar{\mathcal{S}}$. If the set $\{1, 2, 3\}$ was excluded from $\mathcal{S}$, then it would be a minimal nonface and the resulting complex would no longer be flag.

Flag complexes are completely characterised by the 2-element subsets which are excluded. This is because the definition of flagness implies that no set must be excluded from the complex unless at least one of its subsets with two elements is also excluded. This allows us to uniquely identify flag complexes with graphs. To be precise, a simplicial complex $\mathcal{S}$ is flag if and only if there exists a graph $\mathcal{G}$ such that the faces of $\mathcal{S}$ are the cliques of $\mathcal{G}$. In that case $\mathcal{S}$ is referred to as the *clique complex* of $\mathcal{G}$ and often written as $\Delta(\mathcal{G})$.

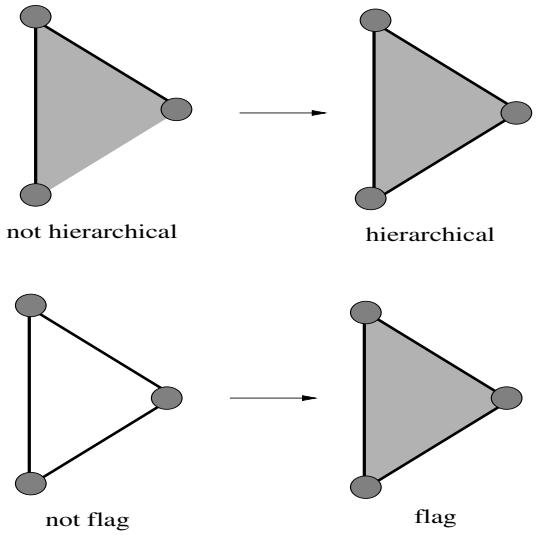

Figure 3.3 – Illustration of hierarchical models and flag simplicial complexes.

Flag simplicial complexes are important from a statistical viewpoint because they naturally lead us to the class of *graphical interaction models*.

Definition 18 (Graphical interaction model)**.** A graphical interaction model is a hierarchical model based on a flag simplicial complex.

Graphical interaction models are formed exclusively from the unique set of maximal cliques of a graph. Thus, unlike most hierarchical models, they can be fully

described by a graph. They will play an important role in Chapter 4.

Figure 3.3 illustrates the definitions of a hierarchical model and of a flag complex: Hierarchical models cannot have higher order interaction terms unless lower order interaction terms are included. Models based on flag complexes (graphical interaction models) specify that cliques must have all interaction terms. Note that flagness is an additional requirement further to being a simplicial complex, so that graphical interaction models are necessarily hierarchical.

We now turn to the main idea of this section which is to relate hierarchical models to sets of square free zero-cumulants. Associated to an index set $K \subseteq [d]$ is a differential operator D^k, where k is the multiplicity of K. Recall from Definition 2 that the multiplicity of a set indicates which elements are in the set. We may write $k = \sum_{i \in K} e_i \in \{0,1\}^d$. Then k holds ones for every member of K and zeros otherwise. In the following, we overload the differential operator by allowing it to be superscripted by a set or by a vector. Thus, for an index set K we set $D^K := D^k$ and similarly $\kappa^x_K := \kappa^x_k$. D^K returns the differential cumulant κ^x_K, when applied to $g(x)$.

Example 22. Let $K = \{2,4,6\}$. We obtain $k = (0,1,0,1,0,1)$ and $D^K g(x) = \kappa^x_K = \kappa^x_k = \frac{\partial^3}{\partial x_2 \partial x_4 \partial x_6} g(x)$.

It is a main point of this section that there is a duality between setting collections of mixed differential cumulants equal to zero and a general hierarchical model:

Theorem 6. *Given a simplicial complex $\mathcal{S}$ on an index set $[d]$, a model g is hierarchical, based on $\mathcal{S}$ if and only if all differential cumulants on $\bar{\mathcal{S}}$ vanish everywhere that is*

$$\kappa^x_K = 0, \quad \textit{for all } x \in \mathbb{R}^d \textit{ and for all } K \in \bar{\mathcal{S}}.$$

Proof. First, let g be hierarchical with respect to $\mathcal{S}$, that is g is a log-density with representation $g(x) = \sum_{J \in \mathcal{S}} h_J(x_J)$. Then, for $K \in \bar{\mathcal{S}}$, the associated differential operator D^K annihilates any term h_J in g, since $K \not\subseteq J$ for any $J \in S$.

Conversely, suppose $\kappa_K^x = 0$ for all $x \in \mathbb{R}^d$ and for all $K \in \bar{\mathcal{S}}$. Then, by Lemma 4, f_X is pairwise Markov with respect to $\mathcal{S}$ and hence factorises over maximal cliques of $\mathcal{S}$ by the Hammersley-Clifford Theorem. □

3.4.3 Special model classes

The terms $h_J(x_J)$ which appear in the definition of a hierarchical model have not been given any special form. Certain classes of hierarchical models can however be obtained by imposing further differential conditions. The following lemma shows that the log-density is polynomial if we impose univariate derivative restrictions.

Lemma 5. *If in addition to the differential conditions in Theorem 6 we impose conditions of the form*

$$\frac{\partial^{n_i}}{\partial x_i^{n_i}} g(x) = 0, \; n_i \in \mathbb{N}, \; \textit{for all } 1 \leq i \leq d, \tag{3.9}$$

then the h functions in the corresponding hierarchical model are polynomials in which the degree of x_i does not exceed $n_i - 1$, for all $1 \leq i \leq d$.

Proof. Repeated integration with respect to x_i shows that g is indeed a polynomial in x_i of degree less than n_i, when the other variable are fixed. Since this holds for all $1 \leq i \leq d$, the result follows. □

The simultaneous inclusions of derivative operators with respect to one indeterminate in (3.9) constitutes an algebraic operation known as *Artinian closure* (Sáenz-De-Cabezón Irigaray, 2008).

3.4.3.1 The multivariate conditional exponential distribution

Perhaps the simplest case of obtaining polynomials in the h functions is to force all second order terms in one variable to zero. This yields the so called multivariate

conditional exponential distribution. We start with the bivariate case. Thus, suppose X is bivariate and we impose the symmetric conditions

$$\frac{\partial^2}{\partial x_i^2} g(x_1, x_2) = 0, \quad \text{for } i = 1, 2.$$

Then integration yields

$$g(x_1, x_2) = x_1 h_1(x_2) + h_2(x_2)$$

and

$$g(x_1, x_2) = x_2 h_3(x_1) + h_4(x_1).$$

A comparison of these functionals identifies $h_1(x_2) = a_3 x_2 + a_1$, $h_2(x_2) = a_0 + a_2 x_2$, $h_3(x_1) = a_3 x_1 + a_2$, $h_4(x_1) = a_1 x + a_0$, for some $a_i \in \mathbb{R}$ for all $0 \leq i \leq 3$, so that $g(x_1, x_2)$ can be written as

$$g(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2. \tag{3.10}$$

It can be shown that X_1 is distributed exponentially conditional on $X_2 = x_2$ for all $x_2 > 0$ and vice versa (Arnold and Strauss, 1988). A distributions with this property is called bivariate exponential conditionals (BEC) distribution. BEC distributions are completely described by g in the sense that any BEC density is of the form (3.10). In particular, the independence case is included if we force $a_3 = 0$ by imposing the additional restriction

$$\frac{\partial^2}{\partial x_1 x_2} g(x_1, x_2) = 0.$$

This also confirms Lemma 2 for this particular example.

The example extends readily into higher dimension. We call a distribution multivariate exponential conditionals (MEC) distribution if X_j is distributed exponentially conditional on $X_{-j} = x_{-j}$ for all $1 \leq j \leq d$. We capture the extension to the d-dimensional case in the following lemma:

Lemma 6 (MEC distributions and Artinian closure)**.** *The following statements are equivalent:*

1. *A distribution belongs to the class of MEC distributions.*
2. *The log-density g is multi-linear, i.e there exist 2^d indices $a_s \in \mathbb{R}$ such that $g = \sum_{s\in\zeta} a_s x^s$, where $\zeta = \{0,1\}^d$ denotes the set of d-dimensional binary vectors.*
3. $\frac{\partial^2}{\partial x_i^2} g(x) = 0$, *for all* $1 \leq i \leq d$.

Proof. For a proof of (1) $\iff$ (2) see Arnold and Strauss (1988). The proof of (2) $\iff$ (3) follows the lines of the example. □

3.4.3.2 The multivariate normal distribution

Another case of considerable interest is the Gaussian distribution. Here the maximal cliques are of degree two. The latter condition is partly obtained with an Artinian closure with $n_i = 3,\ i = 1,\ldots,p$. However, more is required. We can guess, from the fact that for a normal distribution all (ordinary) cumulants of degree three and above are zero that if we impose all degree three *differential* cumulant to be zero we have polynomial terms of maximum degree 2. This is, in fact, the correct set of conditions to make the model's terms of degree at most two. In the α-notation the conditions are

$$D^\alpha g = 0, \text{ for all } \alpha \in \mathbb{N}^d \text{ with } |\alpha| = 3.$$

This includes the Artinian closure conditions. The corresponding ideal is generated by all polynomials of degree three. For a non-singular multivariate Gaussian, we also require non-negative definiteness of the degree two part of the model, considered as a quadratic form.

The hierarchical model is given by additional restrictions which are equivalent to removing certain terms of the form $x_i x_j$, $i \neq j$. This is the same as setting the corresponding (i, j)-th entry in the influence matrix equal to zero.

3.4.3.3 The multivariate von Mises distribution

The von Mises distribution is used to model angular variables. In the univariate case it is supported on the unit circle $[0, 2\pi]$ and has density:

$$f(x, \kappa) = \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\theta)}, \; x \in [0, 2\pi],$$

where μ and θ are location and scale parameters and $I_0(\theta)$ is the Bessel function or order θ.

Singh et al. (2002) and Mardia et al. (2008) generalise the distribution to the bivariate and multivariate case respectively and, in a research report, Razavian et al. (2011), introduce an undirected von Mises graphical model. Similarly, we will generalise the univariate distribution to various forms of multivariate distributions, taking into account the principles of this chapter.

The natural Fourier expansion in one dimension takes the form

$$g(x) = \theta_0 + \sum_{j=1}^{\infty} \theta_j \sin(jx) + \phi_j \cos(jx).$$

If we truncate at $j = 1$ we obtain for the exponentiated model

$$f(x) = \exp(\theta_0 + \theta_1 \sin(x) + \phi_1 \cos(x)).$$

Setting $\theta_0 = -\log\{2\pi I_0(\theta)\}, \theta_1 = \kappa \sin(\mu)$ and $\phi_1 = \kappa \cos(\mu)$ shows that the univariate von Mises distribution is just a first order Fourier exponential family:

$$\begin{aligned} f(x, \kappa) &= \frac{e^{\kappa \cos(x-\mu)}}{2\pi I_0(\theta)} \\ &= \exp\{\kappa(\cos(x)\cos(\mu) + \sin(x)\sin(\mu) - \log\{2\pi I_0(\theta)\}\} \\ &= \exp\{\theta_0 + \theta_1 \sin(x) + \phi_1 \cos(x)\}, \end{aligned}$$

where the second line follows from the angle difference identity

$$\cos(x-\mu) = \cos(x)\cos(\mu) + \sin(x)\sin(\mu).$$

A two-dimensional version which is supported on the torus $[0, 2\pi]^2$ starts, naturally,

$$g(x_1, x_2) = \theta_{00} + \theta_{10}\sin(x_1) + \phi_{10}\cos(x_1) + \theta_{01}\sin(x_2) + \phi_{01}\cos(x_2). \tag{3.11}$$

This is clearly represented as two independent von Mises distributions.

The first difficulty in two dimensions is how to obtain a correlated case. Our approach uses the natural two-dimensional Fourier series, but with higher order terms. We include terms in

$$\sin(x_1+x_2), \cos(x_1+x_2), \sin(x_1-x_2), \cos(x_1-x_2).$$

Thus, let

$$g(x_1, x_2) = \theta_0 + \theta_1\sin(x_1) + \phi_1\cos(x_1) + \theta_2\sin(x_2) + \phi_2\cos(x_2) \tag{3.12}$$

$$+ \theta_{12}\sin(x_1+x_2) + \phi_{12}\cos(x_1+x_2) \tag{3.13}$$

$$+ \theta'_{12}\sin(x_1-x_2) + \phi'_{1,2}\cos(x_1-x_2). \tag{3.14}$$

Note that this does not include frequency two terms in x_1 or x_2. We can convert this model to actual multilinear terms in sin and cos using the angle sum and difference identities such as

$$\sin(a+b) = \sin(a)\sin(b) - \cos(a)\cos(b).$$

Other bivariate von Mises distributions in the literature take this form, but it seems easier to integrate our form with the ideas of this chapter.

Given that we have defined the interaction terms for two variables above, it is straightforward to write down a general multivariate von Mises distribution using

only first order interactions: $f(x_1, \ldots, x_d) = \exp\{g(x_1, \ldots, x_d)\}$, where:

$$\begin{aligned} g(x) = \theta_0 &+ \sum_{i=1}^{d} \theta_i \sin(x_i) + \sum_{i=1}^{d} \phi_i \cos(x_i) \\ &+ \sum_{i=1}^{d} \sum_{j>i}^{d} \left(\theta_{ij} \sin(x_i + x_j) + \phi_{ij} \cos(x_i + x_j)\right) \\ &+ \sum_{i=1}^{d} \sum_{j>i}^{d} \left(\theta'_{ij} \sin(x_i - x_j) + \phi'_{ij} \cos(x_i - x_j)\right). \end{aligned} \tag{3.15}$$

The obvious extension to the three-dimensional case is to include the eight terms:

$$\sin(x_1 \pm x_2 \pm x_3) \text{ and } \cos(x_1 \pm x_2 \pm x_3).$$

Recall, from Definition 12, the general form of a hierarchical model

$$f_X(x) = \exp\left\{\sum_{J \in \mathcal{S}} h_J(x_J)\right\}.$$

In order to generate hierarchical models of von Mises type, we set the h_J functions as sin and cos functions. The arguments of the trigonometric functions are sums and differences in the set of variables $\{x_j | j \in J\}$.

The base model, corresponding to the independence case (3.11), has a constant term plus sin and cos terms in all variables $x_1, ..., x_d$, one at a time. The von Mises base model is similar to including only main effects in log-linear models. The associated graph is a set of vertices without edges.

The following steps outline how to build increasingly complex hierarchical models based on von Mises distributions supported on $[0, 2\pi]^d$:

1. Start with the independence case.

2. Corresponding to the edges of a graph, any h_J is modelled by all two-at-a-time terms, as explained in equation (3.15), for all i, j in J. Unless the graph has no cliques of length greater than two, this model is not a graphical interaction model.

3. As above, but also including all degree 2 terms

$$\sin(2x_i), \cos(2x_i), \quad i \in J.$$

4. Include all 'multilinear' terms:

$$\sin(x_i \pm x_j \pm x_k \pm \ldots), \cos(x_i \pm x_j \pm x_k \pm \ldots), \; i < j < k \in J.$$

5. Include all degree J terms

$$\sin\left(\sum_{i \in J: \sum n_i \leq |J|} \pm n_i x_i)\right), \quad \cos\left(\sum_{i \in J: \sum n_i \leq |J|} \pm n_i x_i\right).$$

There is redundancy of terms in part 4 which we have ignored. Thus we have some ways to express hierarchical models in suitable generalised von Mises distribution.

As an example we write down a simple conditional independence model, $X_1 \perp\!\!\!\perp X_2 | X_3$:

$$\begin{aligned} g(x_1, x_2, x_3) \;=\; & \theta_0 + \theta_1 \sin(x_1) + \phi_1 \cos(x_1) + \theta_2 \sin(x_2) + \phi_2 \cos(x_2) \\ & + \theta_3 \sin(x_3) + \phi_3 \cos(x_3) \\ & + \theta_{13} \sin(x_1 + x_3) + \phi_{13} \cos(x_1 + x_3) \\ & + \theta'_{13} \sin(x_1 - x_3) + \phi'_{13} \cos(x_1 - x_3) \\ & + \theta_{23} \sin(x_2 + x_3) + \phi_{23} \cos(x_2 + x_3) \\ & + \theta'_{23} \sin(x_2 - x_3) + \phi'_{23} \cos(x_2 - x_3). \end{aligned}$$

Unlike in the MEC and Gaussian case, the multivariate von Mises distribution cannot be embedded easily into a pure framework of differential cumulants. In each case square-free cumulants can achieve a hierarchical model structure. Being exponentials of polynomials, the MEC and Gaussian distributions can be fully described through the imposition of further constraints on higher order cumulants. Such a simple embedding fails for the von Mises distribution. Consider the univariate case first:

$$g(x) = \theta_1 \sin(x) + \phi_1 \cos(x).$$

The natural differential condition associated to this model is

$$(D^2 + 1)g(x) = 0, \tag{3.16}$$

that is the model cannot be expressed in terms of vanishing differential cumulants. Another complication becomes apparent as we move from the homogenous to the inhomogenous case where the right hand side of (3.16) is not zero. The simplest case arises as we include a constant term θ_0,

$$g(x) = \theta_0 + \theta_1 \sin(x) + \phi_1 \cos(x)$$

which leads to

$$(D^2 + 1)g(x) = \theta_0.$$

Expressing multivariate models in differential form does not introduce conceptual differences. For instance, the bivariate model g described in (3.14) has associated differential form:

$$(D^{20} + D^{02} + D^{22} + 1)g(x) = \theta_0.$$

It is now clear that the difficulties do not lie in expressing the model g in differential form but rather in the increasingly complex interpretation in terms of differential cumulants. Furthermore, unlike in the Gaussian or the MEC case, the differential conditions of the von Mises distribution cannot be mapped easily to the ideal theory explained in the next chapter. There, sets of zero cumulants are mapped to monomial ideals. The ideals generated by the von Mises distribution are generated by polynomials rather than monomials. For instance, $x^2 + y^2 + x^2y^2 + 1$ is the generator associated to the homogenous version of (3.14). In order to analyse the von Mises distribution, polynomial ideal theory needs to be invoked which is beyond the scope of this thesis.

3.5 Conclusion

This chapter provided the theoretical foundation underlying our statistical modelling approach. It was shown how graphical models can be used to represent conditional independence statements. Furthermore, hierarchical models and graphical interaction models were introduced.

This chapter revealed that arbitrary conditional indendence statements can be expressed in terms of pairwise zero-cumulants. This result allows us to describe and estimate the structure of graphical models without explicit reference to the covariance matrix. Hence, the need for a Gaussian distribution assumption is eliminated.

Conditional independence is obtained through setting pairwise binary differential cumulants to zero. The imposition of further differential conditions leads to specific model classes. The multivariate exponential conditional, the multivariate normal and the mulitvariate von Mises distribution are three examples considered. Exploring further such model classes remains for future research.

Chapter 4

Hierarchical models and monomial ideals

4.1 Introduction

The growing area of algebraic statistics makes use of computational commutative algebra particularly for discrete probability models, notably poisson and multinomial log-linear models. Diaconis and Sturmfels (1998) constructed Markov chain algorithms for the discrete case conditional on a sufficient statistic. Other notable contributions are Pistone and Wynn (1996, 1999, 2006). A textbook reference is Pistone et al. (2001) and Riccomagno (2009) gives a brief overview of recent developments in the field. Work connecting the algebraic methods to continuous probability models is sparser although considerable process has been made in the Gaussian case (Drton and Xiao, 2009; Drton et al., 2009).

Section 4.2 explains our link to the algebra via monomial ideals. The previous chapter defined a hierarchical model in terms of a quasi-additive model of the log-density g over a simplicial complex $\mathcal{S}$. One-to-one associated to $\mathcal{S}$ is its so called Stanley-Reisner ideal $I_{\mathcal{S}}$ which will turn out to be the most important quantity of this chapter.

Section 4.3 investigates the class of decomposable models, which is a subclass of the graphical interaction models, also introduced in the previous chapter. Decomposable models are very well studied in the statistical literature (Lauritzen and

Wermuth, 1989; Lauritzen, 1996). They allow for exact computations of maximum likelihood estimators and facilitate causal inference. Our aim is to help characterise them from an algebraic viewpoint.

Section 4.4 gives an example of how the link between the algebra and statistics can be exploited. The example is based on the so called Ferrer ideal, a well studied algebraic quantity. We show how Ferrer ideals naturally lend themselves to modelling data which is characterised by two subgroups within which all variables interact mutually. Section 4.5 introduces the algebraic concept of shellability, which is similar to the statistical concept of decomposability.

4.2 The duality with monomial ideals

A monomial in $x_1, ..., x_d$ is a product of the form $x^\alpha = \prod_{j=1}^d x_j^{\alpha_j}$, where $\alpha \in \mathbb{N}^d$. A monomial ideal I is a subset of a polynomial ring $k[x_1, ..., x_d]$ such that any $m \in I$ can be written as a finite polynomial combination $m = \sum_{k \in K} h_k x^{\alpha_k}$, where $h_k \in k[x_1, ..., x_d]$ and $\alpha_k \in \mathbb{N}^d$ for all $k \in K$. We write $I =< x^{\alpha_1}, ..., x^{\alpha_K} >$ to express that I is generated by the family of monomials $(x^{\alpha_k})_{k \in K}$.

The full set M of monomials contained in the monomial ideal I has the hierarchical structure:

$$x^\alpha \in M \Rightarrow x^{\alpha+\gamma} \in M, \tag{4.1}$$

for any index set $\gamma \in \mathbb{N}^d$. A monomial ideal is square-free if its generators $(x^{\alpha_k})_{1 \leq k \leq K}$ are square free, i.e. $\alpha_k \in \{0, 1\}^d$ for all $1 \leq k \leq K$.

The following discussion, which is one of the main developments in this thesis, shows that there is complete duality between the structure of square-free monomial ideals and hierarchical models. One-to-one associated with a simplicial complex $\mathcal{S}$ is its *Stanley-Reisner ideal* $I_\mathcal{S}$. This is the ideal generated by all square-free monomials in $\bar{\mathcal{S}}$:

Definition 19 (Stanley-Reisner ideal)**.** For a face $K \in \bar{\mathcal{S}}$ let $m_K(x) := \prod_{k \in K} x_k$ denote the associated square-free monomial. Then

$$I_{\mathcal{S}} :=< (m_K)_{K \in \bar{\mathcal{S}}} > .$$

As an example, the Stanley-Reisner ideal $I_{\mathcal{S}}$ associated to the simplicial complex shown in Figure 3.2 is the ideal $< 14, 24 >$ generated by the minimal nonfaces corresponding to the missing one-dimensional faces. Table 4.1 lists several more examples of simplicial complexes and their Stanley-Reisner ideals. We will encounter them further below.

	Facets of $\mathcal{S}$	Stanley-Reisner ideal $I_{\mathcal{S}}$	Figure	Page
Model 1	$\{12, 34, 34, 14\}$	$< 13, 24 >$	Figure 4.2	68
Model 2	$\{123, 234, 345\}$	$< 14, 15, 25 >$	Figure 4.3	70
Model 3	$\{125, 235, 345, 145\}$	$< 13, 24 >$	Figure 4.16	88
Model 4	$\{123, 124, 134, 234, 235, 15\}$	$< 45, 125, 135, 1234 >$	Figure 4.5	76
Model 5	$\{123, 234, 345, 456\}$	$< 14, 15, 16, 25, 26, 36 >$	Figure 4.7	78

Table 4.1 – Facets of $\mathcal{S}$ and generators $I_{\mathcal{S}}$ for some example models.

Having linked a simplicial complex $\mathcal{S}$ to its Stanley-Reisner ideal $I_{\mathcal{S}}$, the second step is to associate the differential operator D^K with the monomial $m_K(x)$. We need only confirm that the hierarchical structure implied by (4.1) is consistent with differential conditions of Theorem 6. Without loss of generality include all differential operators which are obtained by continued differentiation. Then, (4.1) is mapped exactly to

$$D^{\alpha} g(x) = 0, \text{ for all } x \in \mathbb{R}^d \Rightarrow D^{\alpha+\gamma} g(x) = 0, \text{ for all } x \in \mathbb{R}^d \text{ and for all } \gamma \in \mathbb{N}^d.$$

simply by continued differentiation. This bijective mapping from monomial ideals into differential operators is sometimes referred to as a *polarity* and differential ideal theory has its origins in Seidenberg's differential nullstellensatz (Seidenberg,

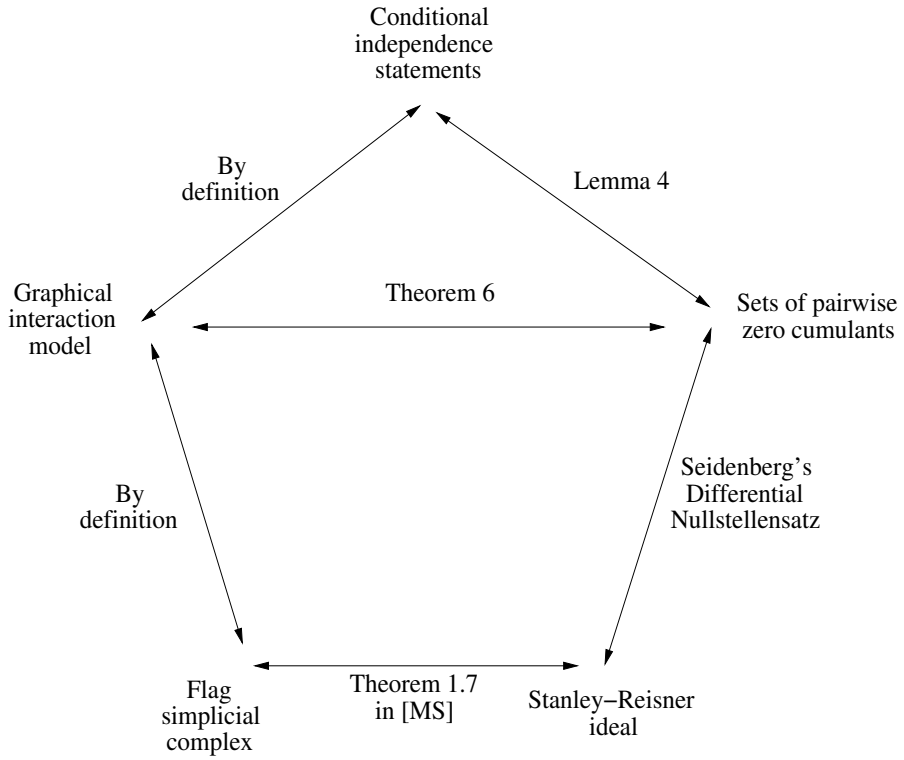

Figure 4.1 – Isomorphisms relating Chapters 2, 3 and 4.

1956). It allows us to map sets of zero-cumulants to monomial ideals. It is beyond this thesis to explore this link rigorously, but we note that for the simplicial complex $\mathcal{S}$ of a hierarchical model the Stanley-Reisner ideal is generated by all monomial terms arising from the polarity. This is the formal Stanley-Reisner ideal $I_{\mathcal{S}}$ of $\mathcal{S}$ (Miller and Sturmfels, 2005).

The above discussion makes clear that sets of pairwise zero-cumulants are isomorphic to flag simplicial complexes and their Stanley-Reisner ideals. These two links close the pentagon of ideas upon which this thesis is built. Figure 4.1 shows the isomorphisms between sets of zero-cumulants, conditional independence statements, graphical interaction models, flag simplicial complexes and their Stanley-Reisner ideals. The previous two chapters discussed the top of Figure 4.1. The rest of this chapter is primarily concerned with the links between the top and

the bottom. It can be considered as a prelude to a wider study of the implications of the equivalences and in particular of algebraic concepts which may lead to interesting statistical properties.

4.3 Decomposable models

4.3.1 Graph-theoretic characterisation of decomposable models

One of the main conditions discussed in the theory of hierarchical models in statistics is the decomposability of a joint density function into a product of certain marginal probabilities. Simple conditional probability is a canonical case. Thus with $p = 3$ the conditional independence $X_1 \perp\!\!\!\perp X_2 | X_3$ is represented by the graph $1 - 3 - 2$. In this case the graph has the model complex: $\mathcal{S} = \{13, 23\}$. The Stanley-Reisner ideal is $< x_1 x_2 >$. There is a factorisation:

$$f_{X_1,X_2,X_3}(x_1, x_2, x_3) = \frac{f_{X_1,X_3}(x_1, x_3) f_{X_2,X_3}(x_2, x_3)}{f_{X_3}(x_3)}.$$

Decomposable graphical models, discussed below, are a generalisation of this simple case. There are other cases, however, where one or more factorisations are associated with the same simplicial complex. An example is the 4-cycle shown in Figure 4.2: $\mathcal{S} = \{12, 23, 34, 41\}$. Any hierarchical model of the form

$$g = h_{1,2} + h_{2,3} + h_{3,4} + h_{4,1}$$

has a factorisation representing the four-cycle. The h functions do not, however, represent marginal densities. The Stanley-Reisner ideal of the four-cycle is given by $I_{\mathcal{S}} =< x_1 x_3, x_2 x_4 >$. Although this ideal is rather simple from an algebraic point of view, the four-cycle from a statistical point of view is rather complex (Whittaker, 1990; Drton et al., 2009).

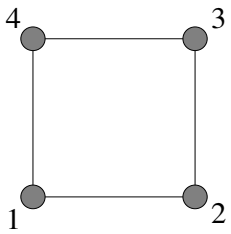

Figure 4.2 – Graph of model 1, the four-cycle. No factorisation is possible which reflects all conditional independencies.

Furthermore, the structure of $\mathcal{S}$ may suggest factorisations even when they are not natural from a statistical viewpoint. Perhaps the first such case is the 3-cycle: $\mathcal{S} = \{12, 13, 23\}$. The Stanley-Reisner ideal is $< x_1x_2x_3 >$. The maximal clique log-density representation has no three-way interaction:

$$g(x_1, x_2, x_3) = h_{12}(x_1, x_2) + h_{13}(x_1, x_3) + h_{14}(x_1, x_4).$$

This might suggest the factorisation

$$f_{X_1,X_2,X_3}(x_1, x_2, x_3) = \frac{f_{X_1,X_2}(x_1, x_2)f_{X_1,X_3}(x_1, x_3)f_{X_2,X_3}(x_2, x_3)}{f_{X_1}(x_1)f_{X_2}(x_2)f_{X_3}(x_3)}. \tag{4.2}$$

A factorisation of this kind is the continuous analogue to a perfect three-dimensional table in the discrete case (Darroch, 1962). However, except when X_1, X_2, X_3 are independent we have not been able to provide a density for which (4.2) holds.

As mentioned, one class of models with particular nice properties is the class of decomposable models. This class of models is graphical in a sense of being fully described by a graph. Whether or not a model is decomposable depends on the associated graph.

Definition 20 (Decomposition of a graph)**.** A partition (I, J, K) of the vertex V of a graph $\mathcal{G}$ decomposes $\mathcal{G}$ if

1. K separates I from J, i.e. any path from I to J must pass through K.

2. K is a clique of $\mathcal{G}$.

We can now define a decomposable graph recursively:

Definition 21 (Decomposable graphs)**.** A graph $\mathcal{G}$ is decomposable if it is complete or if there exists a decomposition (I, J, K) into decomposable subgraphs $\mathcal{G}_{I\cup K}$ and $\mathcal{G}_{J\cup K}$.

Consider the left diagram of Figure 4.3. The graph $\mathcal{G}$ has maximal cliques $\{123, 234, 345\}$. The partition $I = \{1, 2\}, J = \{5\}$ and $K = \{3, 4\}$ of $\mathcal{G}$ induces the subgraphs $\mathcal{G}_{1234}$ and $\mathcal{G}_{345}$. The subgraph $\mathcal{G}_{345}$ is complete since all vertices are mutually connected. The subgraph $\mathcal{G}_{1234}$ decomposes into two complete subgraphs $\mathcal{G}_{123}$ and $\mathcal{G}_{234}$. Hence, $\mathcal{G}$ is decomposable.

Definition 22 (Decomposable simplicial complex and decomposable model)**.** A simplicial complex $\mathcal{S}$ is decomposable if it is the simplicial complex of a decomposable graph or if it is the clique complex of a decomposable graph. A hierarchical model over a decomposable complex $\mathcal{S}$ is decomposable.

If $\mathcal{S}$ represents a decomposable graph $\mathcal{G}$, then it is necessarily one-dimensional and does not include any interaction terms beyond the edges of $\mathcal{G}$.

Decomposable models have a factorisation

$$f_V(x_V) = \frac{\prod_{J\in C} f_J(x_J)}{\prod_{K\in S} f_K(X_K)},$$

where the numerator on the right hand side corresponds to maximal cliques and the denominator to separators which arise in the continued factorisation under the definition. This factorisation does not depend on the order in which the graph is decomposed recursively. Consider again the example shown in Figure 4.3. Taking $K = \{2, 3\}$ gives the factorisation

$$f_{12345} = \frac{f_{123} f_{2345}}{f_{23}}$$

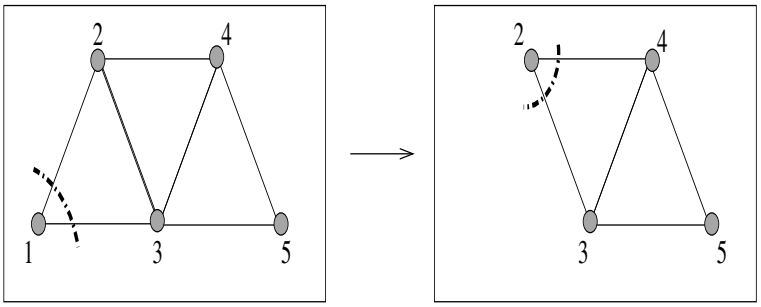

Figure 4.3 – Repeated factorisation and marginalisation of a hierarchical model based on a decomposable graph.

and, using $K = \{3, 4\}$ as the separating set for the induced subgraph $\mathcal{G}_{2345}$, we obtain

$$f_{12345} = \frac{f_{123} f_{234} f_{345}}{f_{23} f_{34}}.$$

The same factorisation would have been achieved with a first stage separation through $K = \{3, 4\}$, followed by a second stage separation through $K = \{2, 3\}$.

It is important to realise that in order to proceed with the factorisation at each stage a marginalisation is required. This is clear from the exponential expression of the model:

$$f_{12345} = \exp\left\{h_{123}(x_1, x_2, x_3) + h_{234}(x_2, x_3, x_4) + h_{345}(x_3, x_4, x_5)\right\}.$$

Integrating with respect to x_1 we obtain a hierarchical model for the marginal joint distribution of (X_2, X_3, X_4, X_5). This marginalisation is possible because x_1 appears only in the single clique $\{1, 2, 3\}$.

The marginalisation has implications for the polynomial rings of which the associated Stanley-Reisner ideals are subsets of. The Stanley-Reisner ideal of the model, $< x_1x_4, x_1x_5, x_2x_5 >$, is a subset of $k[x_1, x_2, x_3, x_4, x_5]$. The factorisation of f_{2345} is, however, mapped into the monomial ideal $< x_2x_5 >$, a subset of $k[x_2, x_3, x_4, x_5]$. A marginalisation has allowed us to drop from five dimensions to four. Here we have an interesting relationship between the statistical and algebraic

formulation: in order to reduce the dimensionality and obtain the Stanley-Reisner ideal for a reduced set of variables, we must first perform a marginalisation, which is a non-algebraic operation in the sense of polynomial ideal theory. We capture this in the following lemma:

Lemma 7 (Marginalisation of rest-graphs)**.** *Suppose the graph $\mathcal{G}$ with vertex set $[d]$ is not complete and has vertex subsets J and K such that $J \cup K$ is a maximal clique and K separates $J \cup K$ from $[d] \setminus (J \cup K)$. Then the marginal model for $[d] \setminus J$ is based on the subgraph $\mathcal{G}_{[d]\setminus J}$. Moreover, the monomial ideal for the marginal representation is obtained by deleting any generators containing elements of J and is in the ring in $x_{[d]\setminus J}$.*

Proof. This follows the lines of the example. The exponential expression for the density will hold a unique term $\exp(h_{J\cup K}(x_{J\cup K}))$ in which x_J appears. Integrating with respect to x_J to obtain the marginal distribution for $X_{[d]\setminus J}$ gives the reduced model. The monomial ideal representation follows accordingly. □

An important class of graphs are the *triangulated* or *chordal* graphs.

Definition 23 (Chordal graph)**.** A *triangulated* or *chordal* graph is a graph with the property that every cycle of length greater than three possesses a chord, i.e. two non-consecutive vertices that are neighbours.

Lemma 8 (Triangulated graphs are decomposable)**.** *A graph is decomposable if and only if it is triangulated.*

Proof. This is a standard result, see for instance Lauritzen (1996). □

The smallest graph which is non-chordal is the four-cycle shown in Figure 4.2. In practice, it is often easier to check triangulation than to check decomposability from first principles. Remarkably, the concept of a chordless graph and the links to various algebraic conditions is known in the algebraic literature. We now sketch these.

4.3.2 Algebraic characterisations of decomposable models

The aim of this section is to give two equivalent algebraic characterisations of a decomposable graph. Dirac's Theorem, see page 81, relates decomposability to properties of the *minimal free resolution* of the Stanley-Reisner ideal $I_{\mathcal{S}}$ and to the *projective dimension* of the Stanley-Reisner ideal of the *Alexander dual* $\mathcal{S}^\star$ of a simplicial complex $\mathcal{S}$. The following introduces the concepts of a minimal free resolution and Alexander duality. We follow Cox et al. (2005) and He (2006).

4.3.2.1 Minimal free resolution

Recall from Definition 19 that the Stanley-Reisner ideal $I_{\mathcal{S}}$ is the square-free monomial ideal generated by the monomials corresponding to the nonfaces of $\mathcal{S}$. An important homological object through which $I_{\mathcal{S}}$ can be studied is its *free resolution*. Heuristically, the free resolution can be thought of as a sequence of matrices determined by the generators of $I_{\mathcal{S}}$ with the defining property that the product of any two consecutive matrices is zero. In order to construct and define the free resolution formally, it is necessary to introduce a minimum of algebraic topology.

Consider a sequence of R-modules and homomorphisms

$$0 \longrightarrow M_l \xrightarrow{\varphi_l} \cdots \longrightarrow M_{i+1} \xrightarrow{\varphi_{i+1}} M_i \xrightarrow{\varphi_i} \cdots \xrightarrow{\varphi_1} M_0 \longrightarrow 0. \tag{4.3}$$

In this notation, the homomorphism φ_{i+1} maps from M_{i+1} to M_i, φ_i maps from M_i to M_{i-1} etc. The first homomorphism, $0 \longrightarrow M_l$, maps 0 to the additive identity of M_l. The last homomorphism, $M_0 \longrightarrow 0$, maps any element in M_0 to 0. Let $\mathrm{im}(\varphi)$ and $\ker(\varphi)$ denote the image and the kernel of φ respectively. The sequence (4.3) is *exact*, if $\mathrm{im}(\varphi_{i+1}) = \ker(\varphi_i)$ for all $i = 1, ..., l$.

It is the R-module M_0 which determines the family of homomorphisms $(\varphi_i)_{1 \leq i \leq l}$ and the family of modules $(M_i)_{1 \leq i \leq l}$. This is the reason why some authors reverse the above chain and put M_0 to the front. Let $f_1, ..., f_t$ denote the generators of

M_0. In our context, R will be the polynomial ring in $x_1, ..., x_d$ over some field K, M_0 will be the Stanley-Reisner ideal $I_{\mathcal{S}}$ and $f_1, ..., f_t$ will be the generators of $I_{\mathcal{S}}$, i.e. the missing edges of the associated graph. Define a homomorphism

$$\begin{array}{rccc} \varphi_1 & : & R^t & \longrightarrow & M_0 \\ & & e_i & \longmapsto & f_i \end{array}, \tag{4.4}$$

where e_i denotes the i-th standard vector in R^t. One can think of φ_1 as the inner product operator between its argument and the vector of generators $(f_1, ..., f_t)$.

A homomorphism $\varphi : M \longrightarrow N$ is surjective if and only if the sequence

$$M \xrightarrow{\varphi} N \longrightarrow 0$$

is exact since then $\text{im}(\varphi) = \ker(N \longrightarrow 0) = N$.

The homomorphism φ_1 defined in (4.4) can be shown to be surjective. Hence, we can identify the generators $f_1, ..., f_t$ of $I_{\mathcal{S}}$ with the exact sequence

$$R^t \xrightarrow{\varphi_1} M_0 \longrightarrow 0.$$

Moreover, the kernel of φ_1 can be shown to be a finitely generated R-submodule. Since $\ker(\varphi)$ has a finite set of generators, we can repeat the above procedure which started with a set of generators, defined the homomorphism φ_1 and considered its kernel $\ker(\varphi_1)$. This results in a new homomorphism φ_2 with kernel $\ker(\varphi_2)$. Unlike φ_1 the homomorphism φ_2 will in general be a matrix since the image is t-dimensional.

We can continue the procedure repeatedly. The Hilbert Syzygy Theorem guarantees that, dealing with the polynomial ring $k[x_1, ..., x_d]$, we only have to do this finitely many times. The process stops with the first injective homomorphism. Figure 4.4 shows a flow chart of the algorithm which determines the sequence of homomorphisms and R-modules. We can now define the free resolution formally:

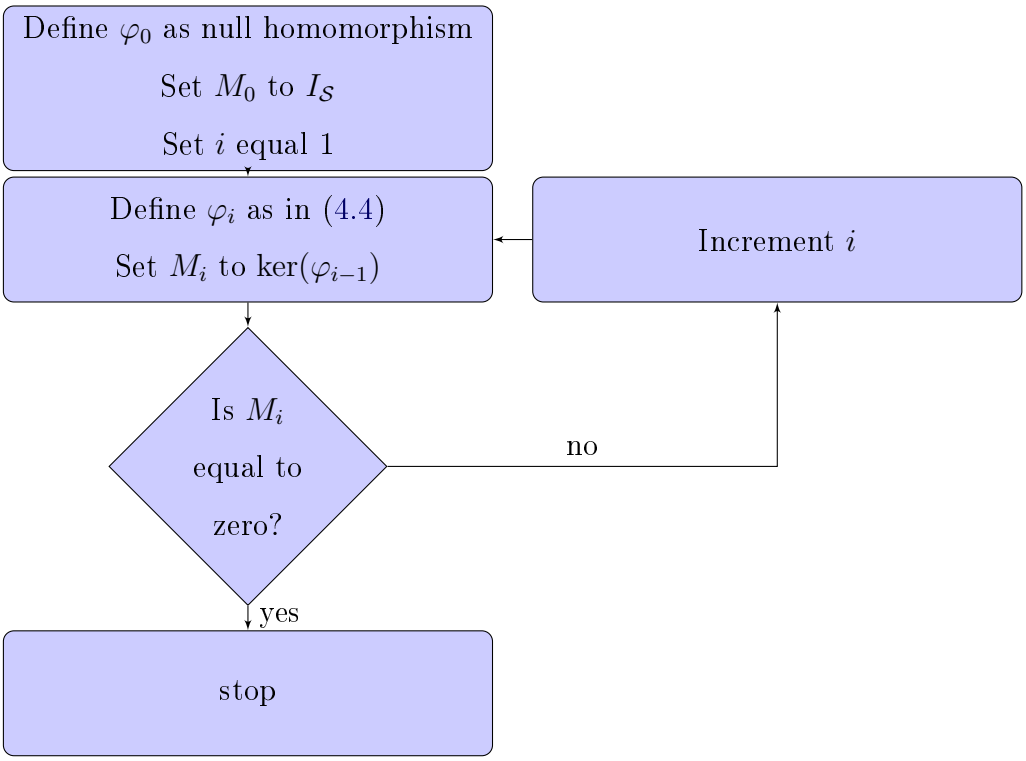

Figure 4.4 – Flow chart of the algorithm for constructing the free resolution of $I_{\mathcal{S}}$.

Definition 24 (Free resolution in polynomial rings)**.** Let $R = k[x_1, ..., x_d]$ and M be a R-module. A *free resolution* of M is an exact sequence of the form

$$0 \longrightarrow M_I \xrightarrow{\varphi_I} \cdots \longrightarrow M_{i+1} \xrightarrow{\varphi_{i+1}} M_i \xrightarrow{\varphi_i} \cdots \xrightarrow{\varphi_1} M_0 \longrightarrow 0. \tag{4.5}$$

The *length* of a free resolution is given by the number of homomorphisms I in (4.5). A free resolution with shortest length is *minimal*. The length of the minimal free resolution is the *projective dimension*, projdim(M). A free resolution is *k-linear* if its associated matrices are linear forms and all generators of M have degree k.

From the construction outlined above it is clear that, in general, the free resolution of a module contains more information than the set of its generators. It is beyond the scope of this thesis to go into the details of the extra information gained. We will, however, sketch one interesting interpretation of the minimal

free resolution of $I_{\mathcal{S}}$ as a sequence of lowest common multiple operations on the generators of $I_{\mathcal{S}}$.

Example 23 (Minimal free resolution)**.** Consider model 4 from Table 4.1, which is essentially Example 1.14 of Miller and Sturmfels (2005). The simplicial complex is shown in Figure 4.5. This model is not a graphical interaction model since the Stanley-Reisner ideal holds generators with more than two elements. This implies that the model is not decomposable. The free resolution is given by the three matrices:

$$A' = \begin{bmatrix} 45 \\ 125 \\ 135 \\ 1234 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & -12 & -13 & -123 \\ 3 & 4 & 0 & 0 \\ -2 & 0 & 4 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}, \qquad C = \begin{bmatrix} -4 \\ 4 \\ -2 \\ 0 \end{bmatrix}.$$

In the interest of readibility, we have displayed only the integer subscripts and we will continue to do so. We keep in mind that 45, for instance, is short for x_4x_5. The free resolution is written

$$0 \longrightarrow S \xrightarrow{C} S^4 \xrightarrow{B} S^4 \xrightarrow{A} S \longrightarrow 0.$$

It is easily checked that

$$AB = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad (BC)' = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix},$$

confirming the exactness property of the free resolution. Note that the matrices A and B are not linear for they hold products of indeterminates.

As mentioned, free resolutions are related to repeated lowest common multiple operations on the generators of $I_{\mathcal{S}}$. The aim is to construct a new simplicial complex based on $I_{\mathcal{S}}$. Take the generators as vertices. Edges are obtained using the lowest common multiple (LCM) of vertices.

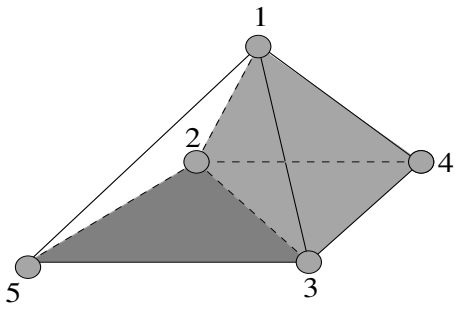

Figure 4.5 – The simplicial complex of model 4 in Table 4.1.

Returning to the example, $x_1x_2x_3$ and $x_1x_3x_5$ belong to the set of generators of $I_{\mathcal{S}}$. Let the symbol $\wedge$ represents the LCM operator. Then $x_1x_2x_3x_5$ forms the edge between the vertices $x_1x_2x_3$ and $x_1x_3x_5$ since

$$x_1x_2x_3 \wedge x_1x_3x_5 = x_1x_2x_3x_5.$$

Continuing this way and joining the edges to form two dimensional faces we construct the entire simplicial complex from LCM operations. It is shown in Figure 4.6. The free resolution captures the mapping between successive levels of the complex as we progress from

$$\{x_4x_5, x_1x_2x_5, x_1x_3x_5, x_1x_2x_3x_4\} \text{ to } \{x_1x_2x_3x_4x_5\}.$$

To find out which vertices are joined by edges, the matrix multiplication in the resolution has to be taken into account. For instance, the first column of B holds two non-zero entries in the second and third row. This allows us to conclude that the vertices $\{1, 2, 5\}$ and $\{1, 3, 5\}$ will be joined through an edge since they are in the second and third column of A. The edge itself is given by either of the products. For instance, $\{1, 2, 5\}$ is multiplied into 3, so that the edge will represent the face $\{1, 2, 3, 5\}$.

We next consider the decomposable model 5 shown in Figure 4.7. It has facets $\{123, 234, 345, 456\}$ and Stanley-Reisner ideal $< 14, 15, 16, 25, 26, 36 >$. Using the resolution function of the commutative algebra package CoCoA (CoCoATeam,

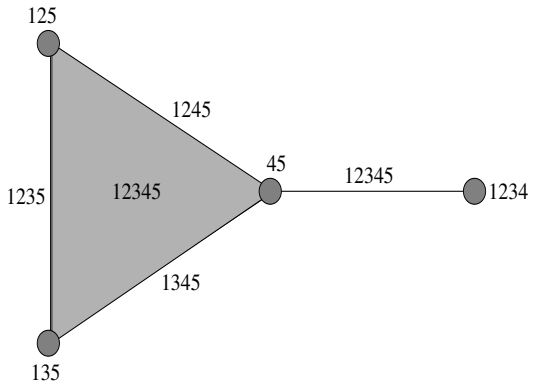

Figure 4.6 – Simplicial complex of the Stanley-Reisner ideal $I_{\mathcal{S}}$. Vertices are ideal generators, edges are obtained through the lowest common multiple operation on the vertices.

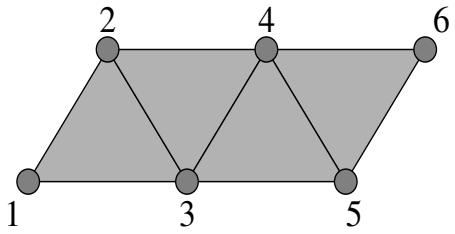

Figure 4.7 – Simplicial complex of model 5.

2004) we have the matrices

$$A' = \begin{bmatrix} 14 \\ 15 \\ 16 \\ 25 \\ 26 \\ 36 \end{bmatrix}, \; B = \begin{bmatrix} 0 & 0 & -6 & 0 & 0 & 0 & -5 & 0 \\ 0 & -6 & 0 & 0 & 0 & 0 & 4 & -2 \\ 0 & 5 & 4 & 0 & -3 & -2 & 0 & 0 \\ -6 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 5 & 0 & 0 & -3 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 \end{bmatrix}, \; C = \begin{bmatrix} 0 & -1 & 0 \\ -4 & 2 & 0 \\ 5 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 2 \\ 0 & 5 & -3 \\ -6 & 0 & 0 \\ 0 & -6 & 0 \end{bmatrix}.$$

The matrices A, B and C are linear in the indeterminates. Again, the simplical complex shown in Figure 4.8 can be constructed from the resolution. An impor-

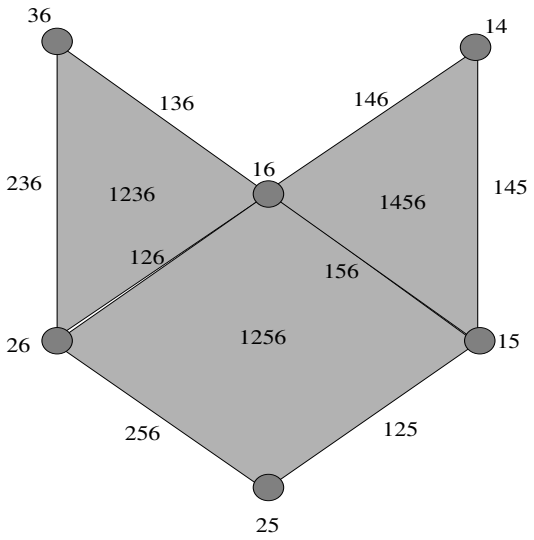

Figure 4.8 – Simplicial complex of $I_{\mathcal{S}}$ for model 5.

tant point is that we do not at this point continue to build the complex up to $\{x_1x_2x_3x_4x_5x_6\}$. This point is somewhat beyond the remit we have adopted for this thesis. A heuristic reason is that a full description of the complexity of the ideal is given and going deeper inside the ideal does not achieve a finer description.

There is a nice interpretation in terms of inclusion-exclusion. The set of monomials in the Stanley-Reisner ideal $I_{\mathcal{S}}$ is the union of sets generated by the individual generators of $I_{\mathcal{S}}$:

$$< x_1x_4, x_1x_5, x_1x_6, x_2x_5, x_2x_6, x_3x_6 >=< x_1x_4 > \cup \cdots \cup < x_3x_6 > .$$

A natural way to express this is via inclusion-exclusion where

$$< x_1x_4 > \cap < x_1x_5 > \; = \; < x_1x_4 \wedge x_1x_5 > \; = \; < x_1x_4x_5 >$$

etc. This is illustrated in Figure 4.9. Suppose we are given a Stanley-Reisner ideal $I_{\mathcal{S}} =< x_1^4, x_1^2x_2^2, x_2^3 >$ and we are to obtain all monomials outside the Stanley-Reisner ideal represented by the shaded area. Black dots represent monomials outside the Stanley-Reisner ideal. They can be counted as the sums and differences of elements in shifted orthants. Roman numbers indicate how many times an

orthant gets added and subtracted. We start with the entire positive orthant. Next we subtract all monomials inside the Stanley-Reisner ideal $< x_1^4 >$, inside the Stanley-Reisner ideal $< x_1^2 x_2^2 >$ and inside the Stanley-Reisner ideal $< x_2^3 >$. This subtracts some monomials in the shaded area more than once, so we need to add back the monomials inside the Stanley-Reisner ideals $< x_1^4 > \cap < x_1^2 x_2^2 >$, $< x_1^4 > \cap < x_2^3 >$ and $< x_1^2 x_2^2 > \cap < x_2^3 >$. Finally subtract the interjection of all three prime ideals, $< x_1^4 x_2^3 >$.

The standard inclusion-exclusion procedure adds and subtracts many terms which cancel. The minimal free resolution gives, in some sense, the shortest identity of inclusion-exclusion type typically without needing to perform LCM to the deepest possible level. In the example above, one can proceed to add further terms leading to $x_1 x_2 x_3 x_4 x_5 x_6$ but the resolution would no longer be minimal.

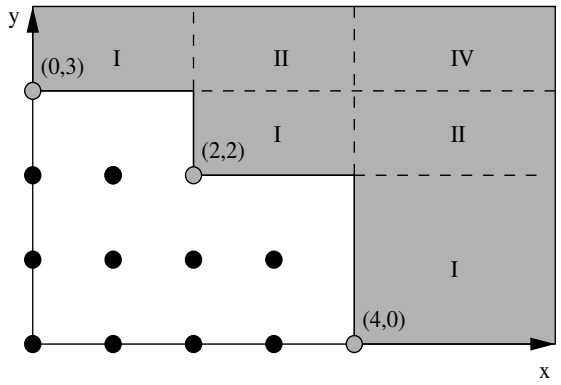

Figure 4.9 – Inclusion-exclusion interpretation of the free resolution.

4.3.2.2 Alexander duality

The second quantity of interest is the Alexander dual of a simplicial complex $\mathcal{S}$ (Miller and Sturmfels, 2005, Definition 1.35). Recall that $\bar{\mathcal{S}}$ holds the nonfaces of $\mathcal{S}$.

Definition 25 (Alexander dual)**.** The Alexander dual of a simplicial complex $\mathcal{S}$

on $[d]$ is defined as the collection of set compliments of the nonfaces of $\mathcal{S}$:

$$S^\star := \{[d] \setminus F : F \in \bar{\mathcal{S}}\}.$$

Example 24. Consider the model complex on $d = [4]$ formed by the cliques $\{123, 234\}$. Then $\bar{S}$ is generated by the nonfaces $\{14, 124, 134, 1234\}$. The collection of compliments of $\bar{\mathcal{S}}$ is the Alexander dual $\mathcal{S}^\star = \{23, 3, 2, \emptyset\}$.

4.3.2.3 Dirac's Theorem

Definition 24 introduced the minimal free resolution, the projective dimension and the concept of k-linearity of a minimal free resolution. Section 4.3.2.2 explained the Alexander dual of a simplicial complex $\mathcal{S}$. We can now relate decomposable models to algebraic properties of the Stanley Reisner ideals of $\mathcal{S}$ and $\bar{\mathcal{S}}$. The following is referred to as Dirac's Theorem.

Theorem 7 (Algebraic characterisation of decomposability)**.** *Given a finite, non-complete graph $\mathcal{G}$ on d vertices with clique complex $\mathcal{S}$ the following are equivalent:*

1. *$\mathcal{G}$ is chordal.*
2. *$I_{\mathcal{S}}$ has a 2-linear resolution.*
3. *The projective dimension of $I_{\mathcal{S}^\star}$ is 1.*

Proof. See Herzog and Hibi (2011, Theorem 9.2.12). Note that the authors only give linearity of $I_{\mathcal{S}}$ as their second equivalent condition. The fact that $I_{\mathcal{S}}$ is indeed 2-linear is, however, implied from their fifth condition stating that $\mathcal{S}$ is a quasi-forest and hence flag. A different proof for 2-linearity is provided by Petrovic and Stokes (2010). □

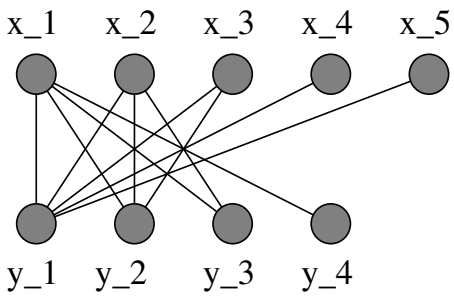

Figure 4.10 – Ferrer graph.

4.4 Ferrer ideals

This subsection gives a particular example that illustrates the research potential that the duality between hierarchical models and monomial ideals offers. The example starts in the algebraic space. It takes a particular class of ideals, the Ferrer ideals, and shows that this ideal class corresponds to statistical models which are decomposable. Models generated by Ferrer ideals, to be defined below, are appropriate if subgroups of variables can be identified which have either no interaction amongst themselves or complete mutual interaction. Ferrer ideals are based on Ferrer graphs, a special class of bipartite graphs.

Definition 26. A bipartite graph $\mathcal{G}$ on $[d]$ is characterised by a partition $[d] = V_1 \cup V_2$ such that every edge of $\mathcal{G}$ is of the form $\{i, j\}$ where $i \in V_1$ and $j \in V_2$.

Figure 4.10 shows a bipartite graph. It has no edges between vertices of the same vertex set.

Definition 27. A Ferrer graph is a bipartite graph on two distinct vertex sets $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_m\}$ such that if (x_i, y_j) is an edge of $\mathcal{G}$, then so is (x_p, y_q) for all $1 \leq p \leq i$ and for all $1 \leq q \leq j$. In addition, (x_1, y_m) and (x_n, y_1) are required to be edges of $\mathcal{G}$.

Bipartite graphs are characterised by the lack of edges within each vertex set. Ferrer graphs impose an extra restriction regarding the edges between the sub-

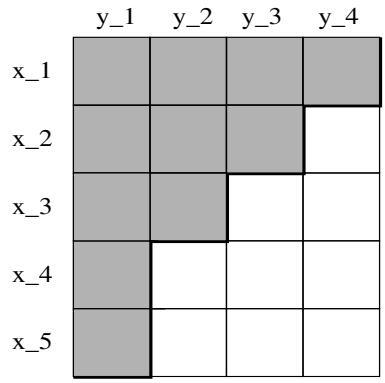

Figure 4.11 – Ferrer tableau exhibiting the characteristic staircase. The bold boundary is sufficient to describe a Ferrer graph.

groups. For this reason, a Ferrer graph can be written in a Ferrer tableau which exhibits an inverse staircase structure. This is shown in Figure 4.11.

Models based directly on Ferrer graphs are applicable when two subgroups of variables exists which have zero interaction within either. Perhaps more likely are scenarios where the interaction within a subgroup is complete. In such a case we should reverse the so far adopted convention that edges of the graph correspond to interaction. Instead, edges should indicate missing interactions and, unlike graphs we have seen before, the Ferrer graph does no longer correspond to the model complex.

The apparent advantage of this reversed approach is a much more efficient encoding of information. If *all* variables within a subgroup interact, there is no need for a large amount of edges expressing exactly that. In fact, doing so may hide otherwise visible structure. Comparing Figure 4.12 with Figure 4.10 illustrates this point.

In the following we assume that edges of the Ferrer graph $\mathcal{G}$ give missing interactions in the model. This leads us to study the *edge ideal* $I(\mathcal{G})$. It is, as the name suggests, generated by the edges of $\mathcal{G}$ and it is the ideal which we refer to as the Ferrer ideal. Of course, the Ferrer ideal is still the Stanley-Reisner ideal of

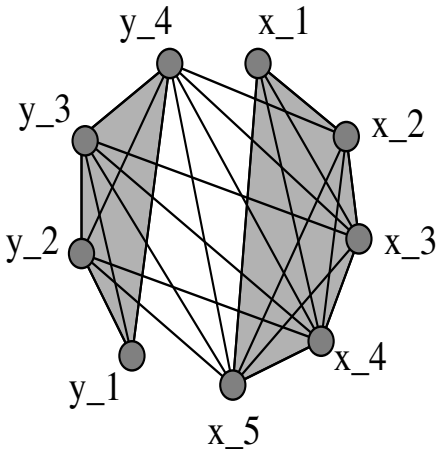

Figure 4.12 – The model complex where edges indicate interaction.

the model simplicial complex $\mathcal{S}$. However, having started with the Ferrer graph, $\mathcal{S}$ is still unkwown to us. Whilst we can retrieve it from $\bar{\mathcal{S}}$ there is no particular reason why we should do so.

Example 25 (Ferrer ideal)**.** As an example consider the graph $\mathcal{G}$ shown in Figure 4.10. The edge ideal $I(\mathcal{G})$ is subset of $k[x_1, ..., x_5, y_1, ..., y_4]$ and is generated by the edges of $\mathcal{G}$. These correspond to the shaded squares in Figure 4.11:

$$I(\mathcal{G}) =< x_1y_1, x_1y_2, x_1y_3, x_1y_4, x_2y_1, x_2y_2, x_2y_3, x_3y_1, x_3y_2, x_4y_1, x_5y_1 > .$$

Corso and Nagel (2009) show that Ferrer ideals have 2-linear resolutions which, by Theorem 7, makes the associated model decomposable. It is straightforward to exhibit the decomposition directly, following Lemma 7. Since we use shaded areas to imply interaction, we consider the complement staircase shown in Figure 4.13. Importantly, both vertex sets are now fully connected inside.

We start with the top row of the complement staircase and note that x_1 has no connection to any y-variable. This identifies $\{x_1, ..., x_5\}$ as the first maximal clique and the only maximal clique to contain x_1. We may integrate out x_1. x_2 interacts with y_4 and hence so do x_3, x_4 and x_5 by the defining property of the Ferrer ideal. x_2 does not interact with y_3 so the next maximal clique, $\{x_2, ..., x_5, y_4\}$ is found.

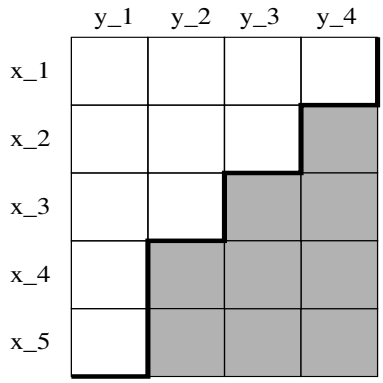

Figure 4.13 – The complement tableau allows us to associate shaded squares to variable interaction. Furthermore, all variables within a vertex set interact.

We continue this process along the boundary of the shaded area in the complement staircase of Figure 4.13. The set of maximal cliques is given by

$$\{x_1x_2x_3x_4x_5, x_2x_3x_4x_5y_4, x_3x_4x_5y_4y_3, x_4x_5y_4y_3y_2, y_1y_2y_3y_4\}.$$

The set of separators is

$$\{x_2x_3x_4x_5, x_3x_4x_5y_4, x_4x_5y_4y_3, y_4y_3y_2\}.$$

This procedure confirms, without invoking the more general Theorem 7, that hierarchical models generated by Ferrer ideals are decomposable.

4.5 Shellability

As implied by the sections above a test of whether the algebraic representation yields new ideas for hierarchical models is where new structures are contributed. We have seen this to some extent with the Ferrer ideals, as yielding a nice subclass of decomposable models.

An important algebraic structure is that of a *shellable simplicial complex*. It is similar, though not identical, to the graph-theoretical concept of decomposability.

We shall see (i) that the concepts overlap, but one does not imply the other, (ii) that being shellable has several algebraic consequences and (iii) that shellability leads to factorisations of the associated density.

We found the lecture notes by He (2006) and the timely book by Herzog and Hibi (2011) and Björner (1995) particularly useful.

Given a set of faces $\{G_1, ..., G_s\}$ of $\mathcal{S}$, we denote by $< G_1, ..., G_s >$ the subcomplex of $\mathcal{S}$ consisting of those faces of $\mathcal{S}$ which are contained in some $G_i, 1 \leq i \leq s$. For instance, $S =< F_1, ..., F_m >$ if $F_1, ..., F_m$ are the facets of $\mathcal{S}$.

Definition 28 (Boundary of a facet)**.** The boundary $\delta(F)$ of a facet F is the union of subsets of F which have one vertex less than F. Formally,

$$\delta(F) := \{f \subset F | \dim(f) = \dim(F) - 1\}.$$

For instance, the boundary of the facet $\{1, 2, 3\}$ representing a triangle is the set of edges $\{12, 13, 23\}$.

Definition 29 (Shellable complex)**.** A simplicial complex $\mathcal{S}$ is shellable if its facets can be ordered $F_1, F_2, ..., F_m$ such that, for all $2 \leq j \leq m$, the subcomplex

$$< \delta(F_j) \cap \cup_{k=1}^{j-1} \delta(F_k) > \tag{4.6}$$

is pure of dimension $\dim F_j - 1$. An order of the facets satisfying these conditions is called a shelling order. To say that $F_1, ..., F_m$ is a shelling order of $\mathcal{S}$ is equivalent to saying that for all i and all $j < i$, there exists $l \in F_i \setminus F_j$ and $k < i$ such that $F_i \setminus F_k = \{l\}$.

The subcomplex in (4.6) is generated by the intersection of boundaries of facet F_j and the union of facets $F_1, ..., F_{j-1}$. The requirement that it has dimension $F_j - 1$ says that, as a new facet it added to the union of facets already considered under the shelling order, it must contribute exactly one new vertex. Hence, in a non-technical sense, the facets of a shellable simplicial complex can be ordered so that they are dense.

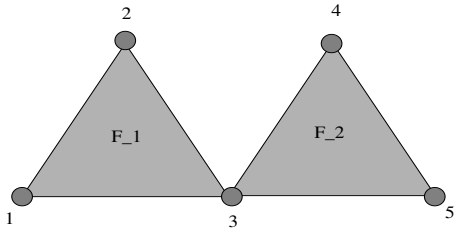

Figure 4.14 – A decomposable graph which complex is not shellable.

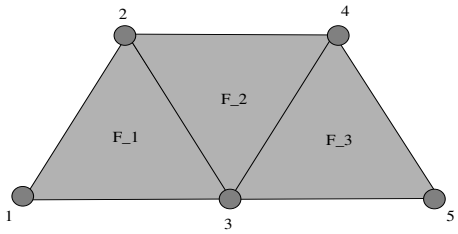

Figure 4.15 – A shellable and decomposable simplicial complex.

We now give a few examples to illustrate the concept and, at the same time, distinguish it from decomposability.

Example 26 (Decomposable complex which is not shellable)**.** Consider the complex depicted in Figure 4.14. It is the clique complex of a triangulated graph and hence decomposable. The associated complex is not shellable since the intersection of F_1 and F_2 is the set $\{3\}$, which contains just a single vertex. The vertex $\{3\}$ has dimension zero whereas $\dim F_2 = 2$. For a clique complex based on a triangulated graph to be shellable, it is necessary that every triangle shares at least one common edge with one other triangle.

Example 27 (Decomposable complex with shellable complex)**.** The simplicial complex depicted in Figure 4.15. is a clique complex of a triangulated graph and hence decomposable. It is also shellable: we can build up by attaching each new triangle by an edge.

Example 28 (Non-decomposable, shellable complex)**.** The four-cycle shown in Figure 4.2 is the prime example of a non-decomposable complex. The cliques

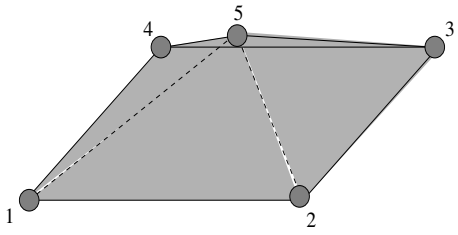

Figure 4.16 – A shellable and non-decomposable simplicial complex.

are formed by the edges. With the ordering $F_1 = \{12\}, F_2 = \{23\}, F_3 = \{34\}$ and $F_4 = \{14\}$ the relevant intersections are all zero-dimensional. Hence, the four-cycle is shellable.

Building on the four-cycle, we can create an example in dimension 2. The simplicial complex depicted in Figure 4.16 includes the four-cycle $\{1234\}$ so it is non-decomposable. It is generated by $\{125, 145, 235, 345\}$. As each of the two-dimensional facets shares two of its edges with other facets, the complex is shellable.

Good intuition for shellability is to think of playing simplicial childrens' bricks with the rule that any new brick must be placed with its maximal faces lying along the maximal face of one or more already placed brick. The definition of shellability and its intuitive representation leads naturally to a class of conditional independence statements.

Lemma 9. *Consider a simplicial complex with vertex set V. Let $F_1 \leq \cdots \leq F_n$ be its ordered facets. For $1 < k < n$ define three sets of vertices:*

$$K = \{v : v \in (\cup_{i=1}^{k} \delta(F_i) \cap \cup_{i=k+1}^{n} \delta(F_i))\}$$
$$I = \{v : v \in \cup_{i=1}^{k} F_i\}$$
$$J = \{v : v \in \cup_{i=k+1}^{n} F_i\}.$$

Assume that $I \setminus K$ and $J \setminus K$ are non-empty. Then

$$X_{I \setminus K} \perp\!\!\!\perp X_{J \setminus K} | X_K.$$

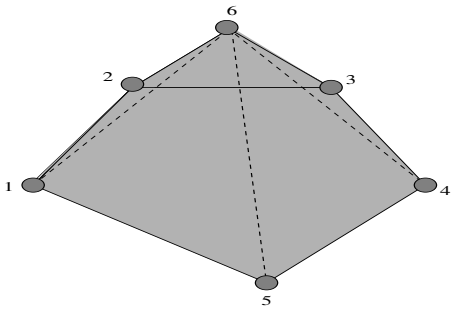

Figure 4.17 – The shellable complex is not decomposable, but the associated model includes conditional independence relations.

Proof. This follows because K is easily seen to separate $I \setminus K$ from $J \setminus K$. □

Example 29 (Conditional independence implied by shellability)**.** Consider the graph shown in Figure 4.17. For $k = 4$ the partitioning sets are $K = \{1, 4, 6\}, I = \{2, 3\}$ and $J = \{5\}$ leading to

$$(X_2, X_3) \perp\!\!\!\perp X_5 | (X_1, X_4, X_6).$$

Decomposable models can be characterized uniquely through the running intersection property (Lauritzen, 1996). Whilst the running intersection property may look similar at first sight, the two concepts have important differences. As one lists the cliques or facets in a shelling order, a new facet must contain exactly one new vertex not contained in the previous facet for the complex to be shellable. Thus, shellability is primarily about the dimension of the intersection. The running intersection property intersects a new clique with the union of cliques lower in the order and tests whether the intersection is fully contained in one clique. Rather than with the number of vertices added outside existing cliques, the running intersection property is concerned with where in the union of existing cliques old vertices are placed. Given that, in general, shellable simplicial complexes do not necessarily correspond to decomposable models and vice versa, no concept can be interpreted as the weakening or strengthening of the other.

4.6 Conclusion

This chapter has linked hierarchical models to monomial ideals and has demonstrated some of the potential of bringing together the worlds of algebra and statistics. Models based on decomposable graphs were shown to be particularly well suited for algebraic analysis.

There are various ways forward. These include models based on geometric constructions where abstract complexes are determined by d-dimensional balls centred about the vertices. Further algebraic quantities such as the Krull dimension or the projective dimension can be linked to hierarchical models. The aim of this chapter has been to demonstrate the large potential of linking the algebra and statistics.

Chapter 5

Nonparametric estimation of conditional independence relations

5.1 Introduction

In previous chapters, the mixed partial derivatives of the log-density were identified as differential cumulants and their intrinsic relation to conditional independence structures was described. This chapter develops a nonparametric hypothesis test for conditional independence based on this condition.

The test is partially based on Proposition 4 of Chapter 3 which suggests that two sets of random variables X_I and X_J are conditionally independent of a third set X_K if and only if all pairwise differential cumulants between X_I and X_J vanish. Graphically, no vertex in I must be joined with any vertex in J for X_I and X_J to be conditionally independent.

By definition, a differential cumulant takes the form

$$\kappa^x_{e_i+e_j} = \frac{1}{f(x)}\frac{\partial^2 f(x)}{\partial x_i \partial x_j} - \frac{1}{f^2(x)}\frac{\partial f(x)}{\partial x_i}\frac{\partial f(x)}{\partial x_j}$$

for some $d-$dimensional unit vectors e_i and e_j. A plug-in estimator $\hat{\kappa}^x_{e_i+e_j}$ replaces densities and their derivatives with kernel estimators. Conditional independence is linked to differential cumulants vanishing everywhere. Hence, we focus on the squared integrated version of $\kappa^2_{e_i+e_j}(x)$, which we denote by $\theta_{e_i+e_j}$:

$$\theta_{e_i+e_j} := \int_{\mathbb{R}^d} \kappa^2_{e_i+e_j}(x)dx.$$

Section 5.2 describes an estimator $\hat{\theta}_{e_i+e_j}$ based on numerical integration of $\hat{\kappa}^2_{e_i+e_j}(x)$.

Kernel estimators are subject to the choice of a smoothing parameter. As the multivariate density f_X, its gradient and its Hessian matrix need to be estimated, the smoothing parameter takes the form of three bandwidth matrices. Their choice is discussed in Section 5.3, where we argue for a sufficiently flexible choice derived via the so called *normal reference rule*.

Section 5.4 describes a bootstrap test for the null hypothesis $H_0 : \theta_{e_i+e_j} = 0$ against the alternative $H_1 : \theta_{e_i+e_j} > 0$. The validity of the test is demonstrated through simulations in Section 5.5. Section 5.6 discusses the choice of the bandwidth matrices when just a single differential cumulant needs to be estimated.

5.2 Description of the estimator

Let $(X_1, \ldots, X_d)$ be a d-variate sample. The curse of dimensionality makes higher-dimensional nonparametric estimation difficult (Silverman, 1986, page 91). In practice, the methodology we propose would be best applicable to the case $d \leq 4$.

We maintain the notation from Chapter 2 and denote by κ^x_k the differential cumulant in $x \in \mathbb{R}^d$ of order k. In Chapter 2 we considered the entire class of square-free cumulants. Here we consider pairwise cumulants only. They correspond to the edges of the graphical model. Hence, k is restricted to hold exactly 2 ones and $d-2$ zeros. Thus,

$$\kappa^x_k := \frac{\partial^2}{\partial x_i \partial x_j} \log f(x),$$

for some (i, j) in $[d] \times [d]$, $i \neq j$, $k = e_i + e_j$. Consider $\kappa^x_{e_i+e_j}$ in expanded form:

$$\begin{aligned} \kappa^x_{e_i+e_j} &= \frac{\partial^2 \log f(x)}{\partial x_i \partial x_j} = \frac{1}{f(x)} \frac{\partial^2 f(x)}{\partial x_i \partial x_j} - \frac{1}{f^2(x)} \frac{\partial f(x)}{\partial x_i} \frac{\partial f(x)}{\partial x_j} \\ &= f(x)^{-1} D^{e_i+e_j} f(x) - f(x)^{-2} D^{e_i} f(x) D^{e_j} f(x). \end{aligned} \tag{5.1}$$

A plug-in estimator replaces each term in (5.1) by kernel estimators.

The quantities we are ultimately interested in are the integrated squared differential cumulants:

$$\theta_k := \int_{\mathbb{R}^d} \kappa_k^2(x)dx \tag{5.2}$$

since zero-cumulants of order k have vanishing θ_k.

This section describes a nonparametric estimator of differential cumulants, which we denote by $\hat{\kappa}_k^x$. Given $\hat{\kappa}_k^x$ we can estimate θ_k as

$$\widehat{\theta}_k := \int_{\mathbb{R}^d} \hat{\kappa}_k^2(x)dx.$$

We suggest a plug-in estimator for $\widehat{\theta}_k$, which numerically integrates the squared estimates of κ_k^x over a bounded region in $\mathbb{R}^d$. Since no asssumption is placed on the distribution of $X_1, ..., X_d$, the distribution of this estimator cannot be obtained. This is why the hypothesis test described in Section 5.4 is based on the bootstrap approach.

In order to estimate $\kappa_{e_i+e_j}^x$ via its representation in (5.1) we require estimators of $f(x), D^{e_i}f(x), D^{e_j}f(x)$ and $D^{e_i+e_j}f(x)$. Since our aim is to estimate all conditional independence relations pairwise, a total of $\frac{d(d-1)}{2}$ estimators $\hat{\theta}_k$ need to be tested for zeros. Consequently, estimators of the density f_X, its entire gradient and the upper-diagonal entries of its Hessian are required.

Example 30. Suppose the data is three-dimensional. Three pairwise cumulants exist: $\kappa_{110}^x, \kappa_{101}^x$ and κ_{011}^x. Table 5.1 shows the individual quantities these cumulants are composed of: The density f_X, the gradient $\nabla f_X = (D^{100}f, D^{010}f, D^{001}f)'$ and the upper-diagonal entries of the Hessian matrix: $D^{110}f, D^{101}f$ and $D^{011}f$.

In order to estimate f_X and its derivatives, we apply a multivariate kernel density approach. Our description follows Wand and Jones (1995), Chacón et al. (2011) and Chacón and Duong (2010). Chacón and Duong (2010) have introduced a vectorised treatment of higher order derivatives based on Kronecker products.

	Hessian	Gradient		Density
κ_{110}	$D^{110}f$	$D^{100}f$	$D^{010}f$	f
κ_{101}	$D^{101}f$	$D^{100}f$	$D^{101}f$	f
κ_{011}	$D^{011}f$	$D^{010}f$	$D^{001}f$	f

Table 5.1 – Composition of pairwise cumulants in terms of f_X, its derivatives and its Hessian.

The elegance of their notation becomes apparent when the choice of bandwidth matrices is discussed.

What follows is an exact reproduction of the notational introduction of Chacón et al. (2011): For a matrix A let

$$A^{\otimes r} := \otimes_{i=1}^{r} A = A \otimes \cdots \otimes A$$

denote the r-th Kronecker power of A. If $A \in \mathcal{M}_{m\times n}$, then $A^{\otimes r} \in \mathcal{M}_{m^r \times n^r}$, with the conventions $A^{\otimes 1} = A$ and $A^{\otimes 0} = 1 \in \mathbb{R}$. Let $D^{\otimes r} f(x) \in \mathbb{R}^{d^r}$ be the vector containing the partial derivatives of order r of f at x, arranged so that

$$D^{\otimes r} f = \frac{\partial f}{(\partial x)^{\otimes r}} \in \mathbb{R}^{d^r}.$$

Thus, we write the r-th derivative of f as a vector of length d^r, and not as an r-fold tensor. Each entry of $D^{\otimes r} f$ is a partial derivative $D^{\alpha} f$, where $|\alpha| = r$.

We have $D(D^{\otimes r} f) = D^{\otimes (r+1)} f$. The gradient and the vectorised Hessian of f can be written as $\nabla f = D^{\otimes 1} f$ and $\operatorname{vec}(\frac{\partial^2 f}{\partial x \partial x'}) = D^{\otimes 2} f$ respectively. The isomorphic operator vec converts a matrix A into a column vector $\operatorname{vec}(A)$. Specifically, if

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

then $\text{vec}'(A) = (a_{11}, \ldots, a_{m1}, \ldots, a_{1n}, \ldots, a_{mn})$. We can express the differential cumulant $\kappa^x_{e_i+e_j}$ as

$$\kappa^x_{e_i+e_j} = \frac{e'_{i+(j-1)d} D^{\otimes 2} f(x)}{f(x)} - \frac{e'_i D^{\otimes 1} f(x)}{f(x)} \frac{e'_j D^{\otimes 1} f(x)}{f(x)}.$$

Remark 2. The vector $e_{i+(j-1)d}$ represents a unit vector in $\mathbb{R}^{d^2}$. It selects the $(i + (j-1)d)$-th element of $D^{\otimes 2} f_X$ corresponding to the (i, j)-th entry of the Hessian matrix. The vectors e_i and e_j are d-dimensional unit vectors. For the inner product to be well defined, the dimension of a unit vector needs to match the dimension of the Kronecker derivative being multiplied. Hence, there is no ambiguity regarding the dimensionality of unit vectors and no additional indexing is necessary.

Let $K(u) : \mathbb{R}^d \longrightarrow \mathbb{R}$ be a kernel, i.e. a non-negative function which is symmetric about the origin and integrates to one over its domain. Let H be a generic notation denoting a symmetric and positive definite bandwidth matrix in $\mathbb{R}^{d \times d}$. The notation is generic in the following sense: In general, distinct multivariate kernel density estimators will have distinct bandwidth matrices attached to it. For instance, (5.3) below defines a kernel density estimator and (5.4) defines several kernel density derivative estimators, one for each order of derivative. Each of these estimators has a different bandwidth matrix attached to it. All of them will be denoted by H since the context will make it clear which estimator they belong to. Furthermore, H will normally be considered to be a function of n, the size of the data. Again, we surpress this dependence to avoid too many iterated indices.

Setting $K_H(u) := |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}} u)$, an estimator for the density $f_X(x)$ is given by

$$\hat{f}(x; H) = n^{-1} \sum_{i=1}^{n} K_H(x - X_i), \tag{5.3}$$

where $x \in \mathbb{R}^d$ and $X_i \in \mathbb{R}^d$ for all $i = 1, ..., d$.

Derivatives of $f_X(x)$ can be estimated via a derivative kernel:

$$\begin{aligned}\widehat{D^{\otimes r} f}(x;H) &= D^{\otimes r}\hat{f}(x;H)\\ &= n^{-1}\sum_{i=1}^{n} D^{\otimes r} K_H(x-X_i)\\ &= n^{-1}(H^{-\frac{1}{2}})^{\otimes r}\sum_{i=1}^{n}(D^{\otimes r}K)_H(x-X_i)\\ &= n^{-1}|H|^{-\frac{1}{2}}\,(H^{-\frac{1}{2}})^{\otimes r}\sum_{i=1}^{n} D^{\otimes r}K(H^{-\frac{1}{2}}(x-X_i)).\end{aligned} \tag{5.4}$$

Equations (5.3) and (5.4) provide all estimators required. We can now express the estimator for $\kappa^x_{e_i+e_j}$ as

$$\hat{\kappa}^x_{e_i+e_j} = \frac{e'_{i+(j-1)d} D^{\otimes 2}\hat{f}(x)}{\hat{f}(x)} - \frac{e'_i D^{\otimes 1}\hat{f}(x)}{\hat{f}(x)}\frac{e'_j D^{\otimes 1}\hat{f}(x)}{\hat{f}(x)}. \tag{5.5}$$

As Example 30 illustrates, it is necessary to estimate the density, the gradient and the Hessian matrix of f_X. Each of these estimators requires its own bandwidth matrix. Their choice will be discussed in the following section.

5.3 Choice of the bandwidth matrices

The choice of the bandwidth parameter h in the univariate setting has been studied by various authors. Notable contributions include Sheather and Jones (1991), Hall and Marron (1991) and Hall et al. (1991). Jones et al. (1996) present a literature survey.

Less progress has been made in the multivariate case. Early work (see e.g. Härdle et al., 1990) concentrated on the rather restricted version of the bandwidth matrix, requiring that H can be written as $H = h^2 I_d$. Only one bandwidth parameter is chosen for all variates and associated kernels are spherically symmetric. Whilst being as parsimonious as possible, it might be overly restrictive for densities with high curvatures (Wand and Jones, 1993).

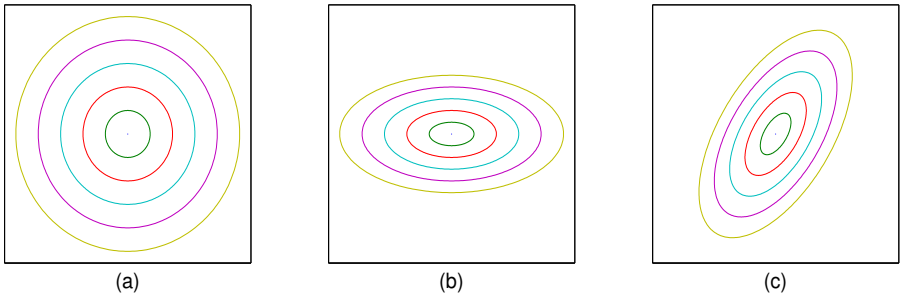

Figure 5.1 – Contour plots of bivariate kernels for different bandwidth matrices: spherical (a), elliptical (b), rotated elliptical (c).

.

A more flexible approach allows independent smoothing in each principal direction. In that case H belongs to the class of diagonal bandwidth matrices, i.e. it can be written as $H = \text{diag}(h_1^2, \ldots, h_d^2)$ for some $(h_1, ..., h_d) \in \mathbb{R}^d$. This allows the kernel to take elliptical contours along the principal axes.

In the most general version, which we adopt, H is only required to be symmetric and positive definite. It allows the shape of the kernel to stretch independently into any direction. The added flexibility comes at the price of $d^2 - d$ additional parameters compared to the diagonal case. Figure 5.1 shows contour plots of kernels parametrised by the three types of bandwidth matrices. The theory for unconstrained multivariate bandwidth matrices has been progressing quickly in recent years. We draw on Chacón (2009), Chacón et al. (2011) and Chacón and Duong (2010).

Various methods have been proposed for estimating the bandwidth matrix H. Common to most of them is a loss-function which is to be minimised. We restrict our attention to the asymptotic mean integrated squared error (AMISE).

For an estimator $\hat{\theta}$ of a vector θ the mean square error (MSE) of $\hat{\theta}$ is a measure of estimator quality. The MSE is defined as

$$MSE(\hat{\theta}) = \mathbb{E}\left\|\hat{\theta} - \theta\right\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. It can be shown that the multivariate bias-variance decomposition of the mean-square-error holds:

$$MSE(\hat{\theta}) = B^2(\hat{\theta}) + V(\hat{\theta}),$$

where $B^2(\hat{\theta}) = \left\|\mathbb{E}\hat{\theta} - \theta\right\|^2$ and $V(\hat{\theta}) = \mathbb{E}\left\|\hat{\theta} - \mathbb{E}\hat{\theta}\right\|^2$. A global measure for the quality of a density estimators is the asymptotic mean integrated squared error:

$$\text{AMISE} := \lim_{n\longrightarrow\infty} \mathbb{E}\left(\int_{\mathbb{R}^d} \left\|\hat{f}(x) - f(x)\right\|^2 dx \right).$$

AMISE can be interpreted as a function of the bandwidth matrix H and the second derivative of the unknown density f_X (see Wand and Jones, 1995). Hence, we would need to know f_X in order to estimate H and H in order to estimate f_X. Various methods have been suggested to overcome this apparent dilemma.

The so-called 'rule-of-thumb' estimators replace f_X by an arbitrary pilot density in order to determine H. Typically, the pilot density is multivariate Gaussian and the procedure is referred to as the *normal reference rule*. The bandwidth matrix H can then be used to estimate f_X. The normal reference rule is easy to implement and works well for smooth densities. It is the method by which we estimate H. Other selection methods include plug-in bandwidth selection (Duong and Hazelton, 2003), *cross-validation* or biased cross validation (Duong and Hazelton, 2005).

Chacón et al. (2011, Theorem 6) determine the AMISE optimal bandwidth matrix according to the normal reference rule. As this is the estimation method we employ, we state their result here for completeness.

Theorem 8 (Normal reference bandwidth matrix)**.** *Assume that H is a symmetric and positive definite bandwidth matrix, and such that every element of $H \longrightarrow 0$ and $n^{-1}|H|^{-\frac{1}{2}}(H^{-1})^{\otimes r} \longrightarrow 0$ as $n \longrightarrow \infty$. Further assume that f_X is a normal density with variance Σ and K is the normal kernel. Then, the bandwidth which*

minimises AMISE$\{\widehat{D^{\otimes r}f}\}$ *is given by:*

$$H = \left(\frac{4}{d+2r+2}\right)^{\frac{2}{d+2r+4}} \Sigma n^{\frac{-2}{d+2r+4}}.$$

Proof. See Chacón et al. (2011). □

5.4 A bootstrap hypotheses test

Given the variables, $X_i, ..., X_d$, the goal of this section is to identify all partitions (I, J, K) of $\{1, ..., d\}$ such that $X_I \perp\!\!\!\perp X_J | X_K$. These conditional independence relations can be of interest in their own right. They can also be used to determine the interaction terms to be excluded from a hierarchical model. We describe a hypotheses test for conditional dependence based on a nonparametric bootstrap approach (Hall and Wilson, 1991; Davison and Hinkley, 1997; Efron and Tibshirani, 1994).

Lemma 4 showed that $X_I \perp\!\!\!\perp X_J | X_K$ if and only if $X_i \perp\!\!\!\perp X_j | X_K$ for all $i \in I$ and $j \in J$. The importance of this result becomes clear now, as it allows us to restrict our attention to estimating conditional independence pairwise.

Recall the definition of $\theta_{e_i+e_j}$ from (5.2):

$$\theta_{e_i+e_j} := \int_{\mathbb{R}^d} \kappa^2_{e_i+e_j}(x)dx.$$

Conditional independence of X_i and X_j implies $\theta_{e_i+e_j} = 0$. An estimator $\hat{\theta}_{e_i+e_j}$ can be constructed by replacing $\kappa^x_{e_i+e_j}$ by $\hat{\kappa}^x_{e_i+e_j}$ for various values of x, squaring and numerically integrating. The distribution of $\hat{\theta}_{e_i+e_j}$, however, is unavailable since no distribution assumption for the random variables $X_1, ..., X_d$ is made. The nonparametric bootstrap approach is suitable since it does not assume anything about F_X other than its existence.

The key idea behind the bootstrap is to resample the original data set with replacement in order to gain information about the variability of an estimator. If

the data are independently and identically distributed, any *bootstrap replication* $X_1^*, ..., X_d^*$ of the data set could have arisen under F_X. This provides an intuitive justification for the bootstrap approach.

In the following, we fix i and j and use the generic expression θ instead of $\theta_{e_i+e_j}$. Set $\theta_0 = 0$. A natural null hypothesis is

$$H_0 : \theta = \theta_0$$

against the alternative

$$H_1 : \theta > \theta_0.$$

By the design of the test, we conclude that X_i and X_j are not conditionally independent if $\hat{\theta}$ differs significantly from zero. What exactly it means to be 'significantly different from zero' is determined by the following bootstrap procedure: First choose a significance level α. Then draw R bootstrap replications $X_1^*, ..., X_d^*$ with replacement. For each of them compute the difference between the bootstrap estimator $\hat{\theta}^*$ and $\hat{\theta}$. The critial value $\hat{t}$ is chosen such that

$$\mathrm{P}\left(\hat{\theta}^* - \hat{\theta} > \hat{t}\right) = \alpha,$$

where the probability distribution is obtained from the bootstrap replications. Finally, reject H_0 if $\hat{\theta} - \theta_0 > \hat{t}$.

5.5 Simulation results

This section illustrates the nonparametric method for estimating conditional independence described in this chapter through a simulation of three Gaussian distributed random variables. Thus, let the true data generating system be

$$X \sim \mathcal{N}(0, \Sigma),$$

where the covariance and the precision matrix are respectively given by

$$\Sigma = \begin{pmatrix} 1 & -1 & -\frac{1}{2} \\ -1 & \frac{4}{3} & \frac{2}{3} \\ -\frac{1}{2} & \frac{2}{3} & \frac{7}{12}. \end{pmatrix} \quad \text{and} \quad \Sigma^{-1} = \begin{pmatrix} 4 & 3 & 0 \\ 3 & 4 & -2 \\ 0 & -2 & 4 \end{pmatrix}. \tag{5.6}$$

The $(1,3)$ entry of Σ^{-1} is zero. Hence, X_1 is conditionally independent of X_3 given X_2.

Let $\rho_{i,j}$ denote the correlation coefficient between X_i and X_j. Differentiation of the log-density shows that

$$\begin{aligned} \kappa_{110} &= -\frac{-\rho_{1,3}\rho_{2,3} + \rho_{1,2}}{\sigma_1\sigma_2(-2\,\rho_{1,3}\rho_{1,2}\rho_{2,3} + {\rho_{1,3}}^2 + {\rho_{1,2}}^2 - 1 + {\rho_{2,3}}^2)}, \\ \kappa_{101} &= \frac{\rho_{1,2}\rho_{2,3} - \rho_{1,3}}{\sigma_1\sigma_3(-2\,\rho_{1,3}\rho_{1,2}\rho_{2,3} + {\rho_{1,3}}^2 + {\rho_{1,2}}^2 - 1 + {\rho_{2,3}}^2)}, \end{aligned} \tag{5.7}$$

and

$$\kappa_{011} = \frac{-\rho_{2,3} + \rho_{1,3}\rho_{1,2}}{\sigma_2\sigma_3(-2\,\rho_{1,3}\rho_{1,2}\rho_{2,3} + {\rho_{1,3}}^2 + {\rho_{1,2}}^2 - 1 + {\rho_{2,3}}^2)}.$$

The multivariate normal density holds a quadratic form $x'\Sigma^{-1}x$ in the exponent. The second derivative of the quadratic form with respect to x_i and x_j is the (i,j)-th entry of Σ^{-1}. Differential cumulants can be evaluated as the differentials of the log-density. Hence, in the Gaussian case, they are equal to the negative of the entries of the precision matrix.

Applied to the current case with covariance matrix Σ as in (5.6) the differential cumulants take the values

$$\kappa_{110} = -3,$$

$$\kappa_{101} = 0$$

and

$$\kappa_{011} = 2.$$

Figure 5.2 shows the pairwise scatterplots of X_2 against X_1, X_3 against X_1 and X_3 against X_2 in that order. Note that, just by looking at the marginal associations, the conditional independence between X_1 and X_3 is completely hidden.

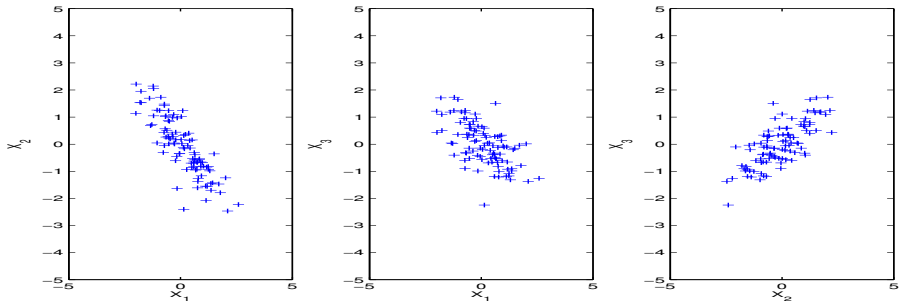

Figure 5.2 – Pairwise scatterplots of normally distributed random variables.

The simulation and estimation proceeds as follows:

1. Simulation of n instances of the random vector $(X_1, X_2, X_3)'$.
2. Estimation of κ^x_{110}, κ^x_{101} and κ^x_{011} for a given $x \in \mathbb{R}^3$.
3. Numerical integration over $\mathbb{R}^3$ of $\hat{\kappa}^2_{110}(x)$, $\hat{\kappa}^2_{101}(x)$ and $\hat{\kappa}^2_{011}(x)$.
4. Bootstrapping of the procedure through repeated sampling of the data with replacement.

The first two steps are straightforward. Steps three and four deserve further attention. In order to establish conditional independence, we would have to show that the differential cumulant κ^x_{101} vanishes everywhere. This task, however, is impossible with finite data, since the variance of the estimator is unbounded in regions where the data is sparse. In order to overcome sparsity issues, we bound the region of integration.

Our approach to this problem is pragmatic. Since the multivariate normal distribution is unimodal, we integrate over the ellipsoid about the sample mean which holds the closest $(\gamma \times 100)$ per cent of the data. The metric we apply is the Mahalanobis distance. Informally, the Mahalanobis distance between a point x_0 and a data set corresponds to the Euclidean distance between the mean of the data and x_0, once the co-ordinate system has been rotated and scaled according to

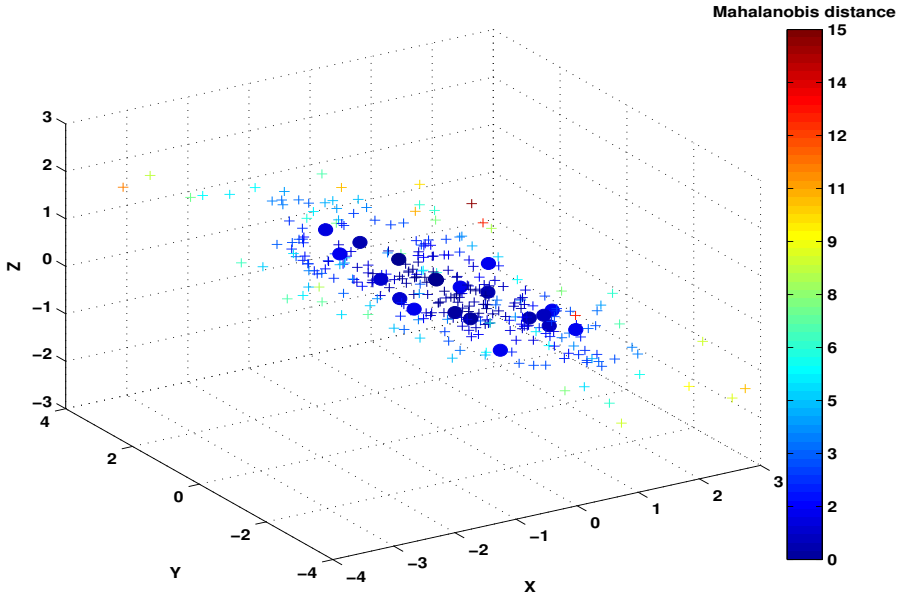

Figure 5.3 – Three dimensional scatterplot of the simulated data (+) and estimation grid points (•). The uniformly drawn grid points are close in Mahalanobis distance to the centroid of the data.

the eigendecomposition of the sample covariance matrix of the data. Formally, for a vector $x_0 \in \mathbb{R}^d$ and a data set X with sample covariance matrix S and sample mean vector $\bar{X}$, the Mahalanobis distance is defined as

$$D_M(x_0) = \sqrt{(x_0 - \bar{X})' S^{-1} (x_0 - \bar{X})}.$$

No optimizing criteria for the choice of γ have been investigated. A more complex simulation study could explore the trade-off between discarding information (lowering γ) and deterioration in estimator performance due to the curse of dimensionality (increasing γ).

Figure 5.3 shows a three-dimensional scatterplot of a simulated data set ($n = 300, \gamma = 0.5$). The data is depicted through pluses. The colour encodes the distance to the mean, where warmer colours represent larger Mahalanobis distances. Outliers have warm colours even if they are close to the centroid in Euclidean distance. We draw N grid points randomly from a uniform distribution over the

γ-ellipsoid. The Mahalanobis distance corresponding to $\gamma = 0.5$ is 2.4, which acts as a cut-off value. This corresponds to a dark blue in the colour bar, which is reflected in the colour of the grid points.

If the data is known to be Gaussian, the differential cumulants can be shown to be constant. It is then meaningful to compare the histogram of $\hat{\kappa}^2_{101}$ with histograms of $\hat{\kappa}^2_{110}$ and $\hat{\kappa}^2_{011}$.

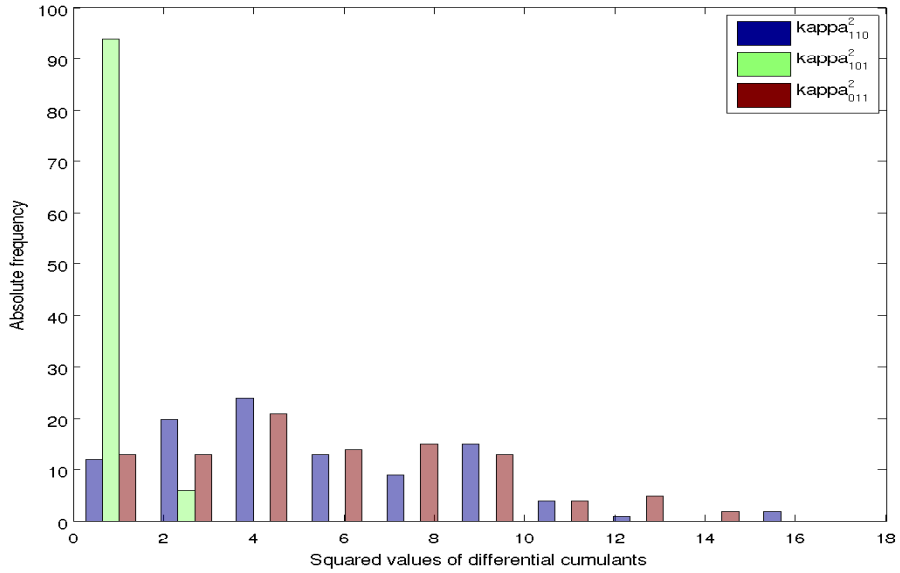

Figure 5.4 – Histogram of squared differential cumulants. The empirical distribution of κ^2_{101} is more concentrated near the origin than the distributions of κ^2_{110} and κ^2_{011}.

Figure 5.4 shows the histogram of the three squared differential cumulants $\hat{\kappa}^2_{101}$, $\hat{\kappa}^2_{110}$ and $\hat{\kappa}^2_{011}$ across $N = 100$ randomly chosen grid points ($n = 500, \gamma = 0.5$). The grid point distribution of κ^2_{101} is clearly left of the grid point distribution of the other differential cumulants. This methodology is only sensible in the Gaussian case and it is used primarily for demonstration purposes.

Figure 5.5 plots $\hat{\kappa}^2_{101}$ integrated over small cubes. The colour of a cube represents the value that the integrated estimator of κ^2_{101} takes in the cube. The centre of the ellipsoid is predominantly blue, whereas both warm and cold colours can be

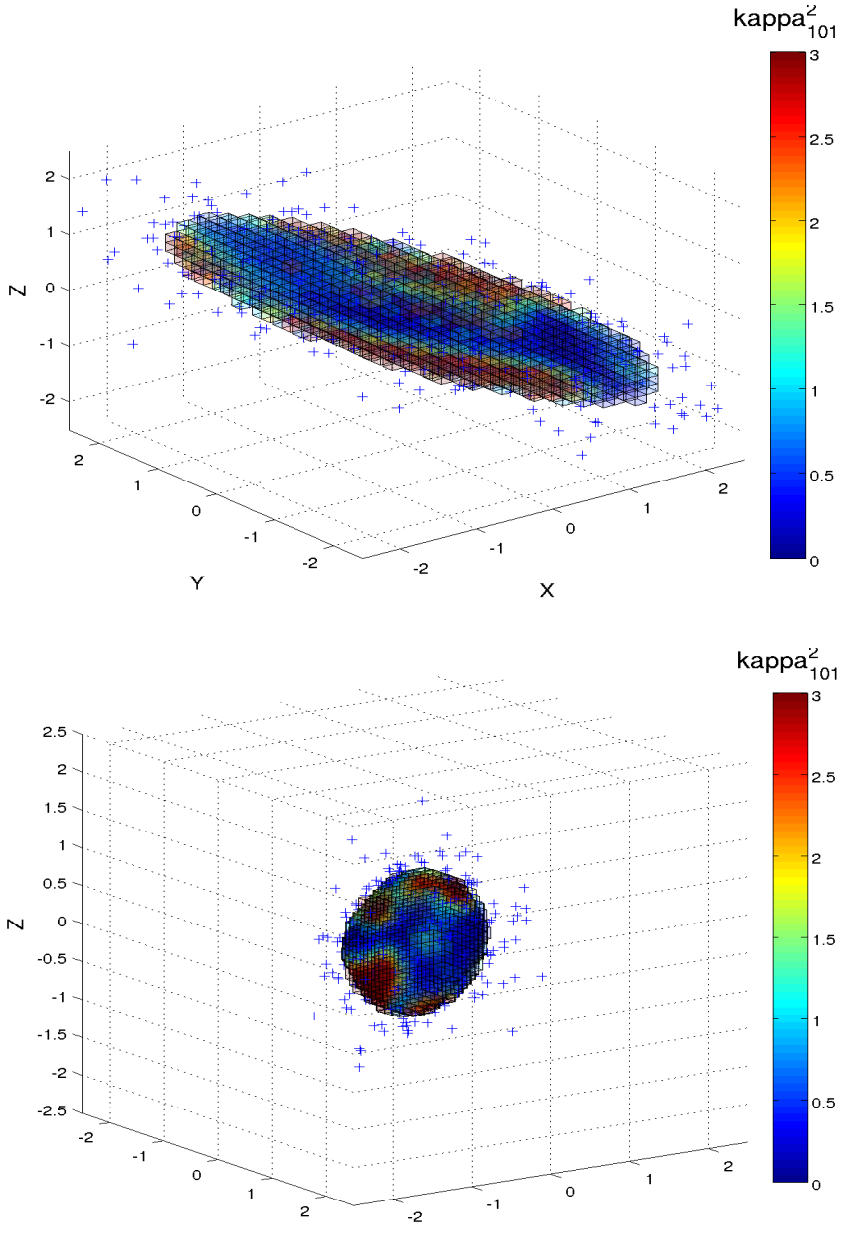

Figure 5.5 – Estimates of $\int \kappa_{101}^2$ over small cubes.

found near the edges. This reflects the increased variability of the estimator in regions with sparse data. Both figures show exactly the same experiment from different angles.

Figure 5.6 shows a comparison between the bootstrap density of $\hat{\theta}$ (dashed) and the density obtained from $\hat{\kappa}^2$ computed at randomly sampled points from the γ-ellipsoid (solid). These are rather different concepts. $\hat{\theta}$ represent the integrated version and its density is obtained from bootstrap replications. $\hat{\kappa}^2$ is not integrated. Furthermore, its distribution is obtained from computing estimates at the N random grid points without bootstrap replications. As mentioned, this is only meaningful in the Gaussian case where the differential cumulants are constant over $\mathbb{R}^3$. We can clearly see the integration effect which moves the mode of the bootstrap distribution further outside compared to the grid point density. At the same time the tail is significantly shortened. Figure 5.7 shows the corresponding cumulative distribution function.

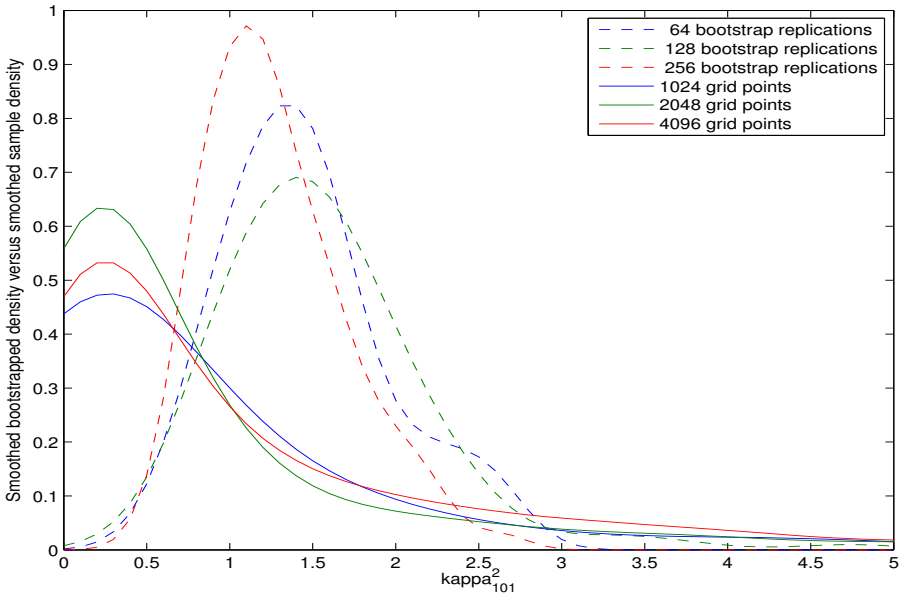

Figure 5.6 – Smoothed bootstrap density of $\hat{\kappa}^2_{101}$ $(n = 1024, \gamma = 0.7)$.

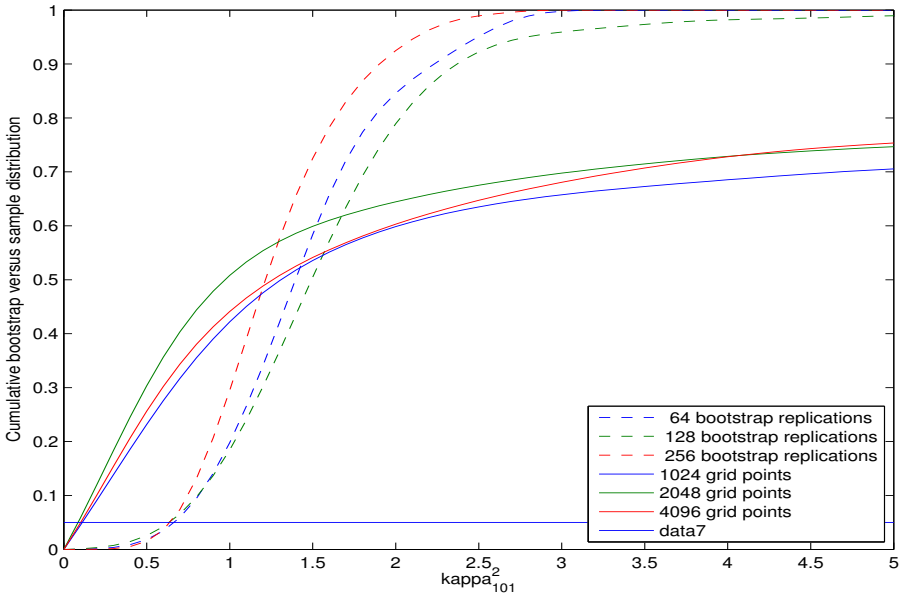

Figure 5.7 – Cumulative distribution for densities. The solid horizontal line (data7) shows the 5-percentile.

Finally, we present the outcome of the bootstrap hypothesis test. We draw 1000 instances from the multivariate normal model with precision matrix Σ^{-1} as in (5.6). θ is estimated over an γ-ellipsoid which holds 70 per cent of the data. The number of bootstrap replication is set to 200. The vector of estimates of θ_{110}, θ_{101} and θ_{011} is

$$\hat{\theta}' = \begin{pmatrix} 6.73 & 0.64 & 1.95 \end{pmatrix}.$$

At a significance level of five per cent, the vector of critical values for the bootstrap distributions is

$$\hat{t}' = \begin{pmatrix} 3.52 & 1.63 & 1.68 \end{pmatrix}. \tag{5.8}$$

This leads us to reject $X_1 \perp\!\!\!\perp X_2|X_3$ and $X_2 \perp\!\!\!\perp X_3|X_1$, whilst failing to reject $X_1 \perp\!\!\!\perp X_3|X_2$.

If the data is known to be normal, one can employ a maximal likelood test for conditional independence (Edwards, 2000). Let $|\hat{\Sigma}_0|$ denote the maximum likelihood estimate of the covariance matrix under the restriction that its inverse has a zero as appropriate entry. Let $|\hat{\Sigma}|$ be the unrestricted maximum likelihood estimate. It can be shown that the deviance test statistic $d = n \log\left(\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}|}\right)$ is asymptotically Chi-squared distributed on 1 degrees of freedom under the null hypotheses that the restricted model is valid (Hojsgaard et al., 2012). For the above data, Table 5.2 reports the results from the maximum likelihood test. The p-values suggest to reject $X \perp\!\!\!\perp Y|Z$ and $Y \perp\!\!\!\perp Z|X$ whilst failing to reject $X \perp\!\!\!\perp Z|Y$.

Null hypotheses	d	p-value
$X \perp\!\!\!\perp Y\|Z$	867.266	0.0000
$X \perp\!\!\!\perp Z\|Y$	0.181	0.6709
$Y \perp\!\!\!\perp Z\|X$	287.057	0.0000

Table 5.2 – Deviance and p-values from maximmum likelihood test.

5.6 Choice of H in a single zero-cumulant test

This section discusses the choice of the bandwidth matrix when the aim is to estimate a single $k_{e_i+e_j}$ as opposed to the total $\frac{d(d-1)}{2}$ pairs. Without loss of generality, we may take $i = 1$ and $j = 2$. We reproduce the estimator for $\kappa^x_{e_1+e_2}$ from (5.5) for convenience:

$$\hat{\kappa}^x_{e_1+e_2} = \frac{e_2' D^{\otimes 2} \hat{f}(x)}{\hat{f}(x)} - \frac{e_1' D^{\otimes 1} \hat{f}(x)}{\hat{f}(x)} \frac{e_2' D^{\otimes 1} \hat{f}(x)}{\hat{f}(x)},$$

where $\hat{f}$ and $D^{\otimes r}\hat{f}$ as defined in (5.3) and (5.4) respectively.

In the preceeding sections the target was to estimate all pairs of cumulants. This rendered it necessary to estimate the entire gradient of f_X as well as the off-diagonal elements of the Hessian. The bandwidth matrices were chosen to minimise the AMISE of the vector valued estimators $D^{\otimes 1}\hat{f}$ and $D^{\otimes 2}\hat{f}$.

This optimisation criterion may be a poor choice when we are only interested in one or two entries of the vector valued estimator. As an example, $D^{\otimes 2}\hat{f}$ holds d^2 entries. The second of these entries corresponds to the $(1, 2)$-entry of the Hessian and is the only entry of $D^{\otimes 2}\hat{f}$ used to estimate $\hat{\kappa}^x_{e_1+e_2}$. The bandwidth matrix which minimises the AMISE of the estimator of the entire Hessian may be a poor choice when d is large and when the smoothness of the second derivative varies greatly across different dimensions.

In this section we derive the AMISE of those parts of the estimator which are used to estimate $\hat{\kappa}^x_{e_1+e_2}$. No minimising bandwidth matrix could be obtained. Hence, no equivalent to the normal reference rule of Theorem 8 can be given. In practice, one would have to resort to numerical minimisation of the AMISE expressions provided below.

As before, we propose to estimate three separate bandwidth matrices: one for $\hat{f}$, one jointly for $e_1' D^{\otimes 1}\hat{f}$ and $e_2' D^{\otimes 1}\hat{f}$ and one for $e_2' D^{\otimes 2}\hat{f}$. Accordingly, the three bandwidth matrices should be chosen to minimise $\mathrm{AMISE}(\hat{f})$, $\mathrm{AMISE}(e_1' D^{\otimes 1}\hat{f}) + \mathrm{AMISE}(e_2' D^{\otimes 1}\hat{f})$ and $\mathrm{AMISE}(e_2' D^{\otimes 2}\hat{f})$ respectively.

The AMISE of the kernel density estimator described in (5.3) is well known (Wand and Jones, 1995). We present a more general theorem due to Chacón et al. (2011) which specifies the AMISE of an estimator of a derivative of arbitrary order r as described in (5.4). Setting r equal to zero includes the density estimator.

Theorem 9 (AMISE$\{\widehat{D^{\otimes r} f}(x;H)\}$). *Assume that H is a symmetric and positive definite bandwidth matrix, and such that every element of $H \longrightarrow 0$ and $n^{-1}|H|^{-\frac{1}{2}}(H^{-1})^{\otimes r} \longrightarrow 0$ as $n \longrightarrow \infty$. Let f_X be a density with square integrable partial derivatives up to order r and square integrable, bounded and continuous partial derivatives up to order $(r+2)$. Let K be a square integrable kernel with square integrable derivatives of order r. Then it holds that*

$$\begin{aligned}\text{AMISE}\{\widehat{D^{\otimes r} f}(x;H)\} = n^{-1}|H|^{-\frac{1}{2}} \operatorname{tr}\left((H^{-1})^{\otimes r} R(D^{\otimes r} K)\right) \\ + \frac{\mu_2(K)^2}{4} \operatorname{tr}\left((I_{d^r} \otimes \operatorname{vec}' H) R(D^{\otimes (r+2)} f)(I_{d^r} \otimes \operatorname{vec} H)\right),\end{aligned}$$

where $\mu_j(K) := \int_{\mathbb{R}} z^j K(z) dz$ and, for a function $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, $R(g)$ is defined as

$$R(g) := \int_{\mathbb{R}^d} g(x) g(x)' dx \in \mathbb{R}^{d \times d}. \tag{5.9}$$

Proof. See Chacón et al. (2011). Some details omitted in their proof appear in the appendix. □

We can apply Theorem 9 to obtain the AMISE$(\hat{f}(x;H))$ as we set r to zero:

$$\text{AMISE}\{\hat{f}(x;H)\} = \frac{\mu_2^2(K)}{4} \operatorname{vec}' H R(D^{\otimes 2} f) \operatorname{vec} H + n^{-1}|H|^{-\frac{1}{2}} R(K). \tag{5.10}$$

Using the fact that for two real matrices A, B of same dimensions it holds that

$$\operatorname{tr}(A'B) = \operatorname{vec}' A \operatorname{vec} B,$$

(5.10) can be equally expressed as

$$\text{AMISE}\{\hat{f}(x;H)\} = \frac{\mu_2^2(K)}{4} \int_{\mathbb{R}^d} \operatorname{tr}^2\{H D^2 f(x)\} dx + n^{-1}|H|^{-\frac{1}{2}} R(K),$$

which coincides with Wand and Jones (1995, page 97).

The next theorem gives an expression for the AMISE of a single component of $\widehat{D^{\otimes r} f}$. Without loss of generality, we may take this to be the first component. It turns out that AMISE$\{\widehat{D^{\otimes r} f}(x; H)\}$ and AMISE$(e_1' \widehat{D^{\otimes r} f})$ differ in that the latter replaces trace expressions by the first component of the respective traces.

Theorem 10 (AMISE$(e_1' \widehat{D^{\otimes r} f})$). *Under the conditions of Theorem 9 it holds that*

$$\text{AMISE}(e_1' \widehat{D^{\otimes r} f}(x; H)) = \text{AIB}^2(e_1' \widehat{D^{\otimes r} f}) + \text{AIV}(e_1' \widehat{D^{\otimes r} f}),$$

where

$$\begin{aligned} \text{AIB}^2(e_1' \widehat{D^{\otimes r} f}) &:= \frac{\mu_2^2(K)}{4} e_1'(I_{d^r} \otimes \text{vec}'\, H) R(D^{\otimes (r+2)} f)(I_{d^r} \otimes \text{vec}\, H) e_1, \\ \text{AIV}(e_1' \widehat{D^{\otimes r} f}) &:= n^{-1} |H|^{-\frac{1}{2}}\, e_1'(H^{-\frac{1}{2}})^{\otimes r} R(D^{\otimes r} K)(H^{-\frac{1}{2}})^{\otimes r} e_1, \end{aligned}$$

and $R(\cdot)$ as defined in (5.9).

Proof. The proof is deferred to Appendix 5.A.2. □

As corollaries we obtain the desired AMISE expressions:

Corollary 4 (AMISE$(e_1' D^{\otimes 1} \hat{f})$ + AMISE$(e_2' D^{\otimes 1} \hat{f})$). *Under the conditions of Theorem 9 it holds that*

$$\text{AMISE}(e_1' D^{\otimes 1} \hat{f}) + \text{AMISE}(e_2' D^{\otimes 1} \hat{f}) = \sum_{i=1}^{2} \text{AIB}^2(e_i' D^{\otimes 1} \hat{f}) + \sum_{i=1}^{2} \text{AIV}(e_i' D^{\otimes 1} \hat{f}),$$

where

$$\text{AIB}^2(e_i' D^{\otimes 1} \hat{f}) = \frac{\mu_2^2(K)}{4} e_i'(I_d \otimes \text{vec}'\, H) R(D^{\otimes 3} f)(I_d \otimes \text{vec}\, H) e_i$$

and

$$\text{AIV}(e_i' D^{\otimes 1} \hat{f}) = n^{-1} |H|^{-\frac{1}{2}}\, e_i' H^{-\frac{1}{2}} R(\nabla K) H^{-\frac{1}{2}} e_i.$$

Corollary 5 (AMISE$(e_2' D^{\otimes 2} \hat{f})$). *Under the conditions of Theorem 9 it holds that*

$$\text{AMISE}(e_2' D^{\otimes 2} \hat{f}) = \text{AIB}^2(e_2' D^{\otimes 2} \hat{f}) + \text{AIV}(e_2' D^{\otimes 2} \hat{f}),$$

where

$$\text{AIB}^2(e_2' D^{\otimes 2} \hat{f}) = \frac{\mu_2^2(K)}{4} e_2'(I_{d^2} \otimes \text{vec}' H) R(D^{\otimes 4} f)(I_{d^2} \otimes \text{vec}\, H) e_2 \tag{5.11}$$

and

$$\text{AIV}(e_2' D^{\otimes 2} \hat{f}) = n^{-1} |H|^{-\frac{1}{2}} e_2'(H^{-\frac{1}{2}})^{\otimes 2} R(D^{\otimes 2} K)(H^{-\frac{1}{2}})^{\otimes 2} e_2. \tag{5.12}$$

The next two theorems provide explicit AMISE expression for estimating a normal density with a normal kernel. Without loss of generality, we may assume that the random variables $X_1, ..., X_d$ have zero mean. Denote by ϕ_Σ the density of a d-variate normal distribution with zero mean and covariance matrix Σ. If Σ is surpressed, ϕ denotes the density of a d-variate standard normal distribution. In short, we assume that $f_X = \phi_\Sigma$ and $K = \phi$.

Define the matrix $I_\alpha^{[\beta]}$ as an α by α diagonal matrix which holds ones on the diagonal up to row β and zeroes otherwise. Formally, $I_\alpha^{[\beta]} = (a_{ij})$, where $a_{ij} = 1$ if $1 \leq i = j \leq \beta$ and $a_{ij} = 0$ otherwise:

$$I_\alpha^{[\beta]} := \left(\begin{array}{c:c} I_\beta & 0 \\ \hdashline 0 & 0 \end{array} \right) \Bigg\} \alpha.$$

If multiplied from the right, $I_\alpha^{[\beta]}$ leaves the first β columns of a matrix unchanged whilst sending the others to zero. If multiplied from the left, $I_\alpha^{[\beta]}$ leaves the first β rows of a matrix unchanged whilst sending the others to zero.

Theorem 11 (AMISE$(e_1' \widehat{D^{\otimes 1} \phi_\Sigma}) +$ AMISE$(e_2' \widehat{D^{\otimes 1} \phi_\Sigma})$). *Assume that the conditions of Theorem 8 are met. Define the auxilliary matrices* $B = \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}}$, $C =$

$2^{-(d+1)}\pi^{-\frac{d}{2}}$ *and* $D = \Sigma^{-\frac{1}{2}} I_d^{[2]} \Sigma^{-\frac{1}{2}}$. *It then holds that*

$$\mathrm{AMISE}(e_1'\widehat{D^{\otimes 1}\phi_\Sigma}) + \mathrm{AMISE}(e_2'\widehat{D^{\otimes 1}\phi_\Sigma}) = \sum_{i=1}^{2} \mathrm{AIB}^2(e_i' D^{\otimes 1}\hat{f}) + \sum_{i=1}^{2} \mathrm{AIV}(e_i' D^{\otimes 1}\hat{f}), \tag{5.13}$$

where

$$\sum_{i=1}^{2} \mathrm{AIV}(e_i' D^{\otimes 1}\hat{f}) = n^{-1}|H|^{-\frac{1}{2}}\, C(H_{11}^{-1} + H_{22}^{-1})$$

and

$$\begin{aligned}\sum_{i=1}^{2} \mathrm{AIB}^2(e_i' D^{\otimes 1}\hat{f}) = \frac{1}{16} C|\Sigma|^{-\frac{1}{2}} \Big[&\operatorname{tr}(D)\operatorname{tr}^2(B) + 2\big(\operatorname{tr}(D)\operatorname{tr}(B^2) \\ &+ 2\operatorname{tr}(B)\operatorname{tr}(DB)\big) + 8\operatorname{tr}(DB^2)\Big].\end{aligned}$$

Proof. The proof is deferred to Appendix 5.A.3. □

Theorem 12 ($\mathrm{AMISE}(e_2'\widehat{D^{\otimes 2}\phi_\Sigma})$)**.** *Assume that the conditions of Theorem 8 are met. It then holds that*

$$\mathrm{AMISE}(e_2'\widehat{D^{\otimes 2}\phi_\Sigma}) = \mathrm{AIB}^2(e_2' D^{\otimes 2}\hat{f}) + \mathrm{AIV}(e_2' D^{\otimes 2}\hat{f}),$$

where

$$\mathrm{AIV}(e_2' D^{\otimes 2}\hat{f}) = C(4n)^{-1}|H|^{-\frac{1}{2}} \left[H_{11}^{-1}H_{22}^{-1} + 2H_{21}^{-1}H_{12}^{-1} \right],$$

$$\begin{aligned}\mathrm{AIB}^2(e_2' D^{\otimes 2}\hat{f}) = C \Bigg\{ 2^{-5}|\Sigma|^{-\frac{1}{2}} \Big[&\operatorname{tr}(F)\operatorname{tr}(G)\operatorname{tr}^2(B) \\ &+ 2\big(\operatorname{tr}(F)\operatorname{tr}(G)\operatorname{tr}(B^2) + 2\operatorname{tr}(F)\operatorname{tr}(B)\operatorname{tr}(BG) \\ &\quad + 2\operatorname{tr}(G)\operatorname{tr}(B)\operatorname{tr}(FB) + \operatorname{tr}(FG)\operatorname{tr}^2(B)\big) \\ &+ 8\big(\operatorname{tr}(F)\operatorname{tr}(GB^2) + \operatorname{tr}(G)\operatorname{tr}(FB^2) + 2\operatorname{tr}(B)\operatorname{tr}(FBG)\big) \\ &+ 4\big(\operatorname{tr}(FG)\operatorname{tr}(B^2) + 2\operatorname{tr}(FB)\operatorname{tr}(BG)\big) \\ &+ 16\big(2\operatorname{tr}(FGB^2) + \operatorname{tr}(FBGB)\big)\Big]\Bigg\},\end{aligned}$$

$C = 2^{-(d+1)}\pi^{-\frac{d}{2}}$, $F = \Sigma^{-\frac{1}{2}} I_d^1 \Sigma^{-\frac{1}{2}}$, $G = \Sigma^{-\frac{1}{2}} I_d^2 \Sigma^{-\frac{1}{2}}$ *and* $B = \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}}$.

Proof. The proof is deferred to Appendix 5.A.4. □

5.7 Conclusion

This chapter demonstrated how conditional dependency structures can be detected through a nonparametric bootstrap test on pairwise differential cumulants. Simulation results from the normal distribution indicated that the methodology works. The research can be extended in several ways. One interesting characteristic of a hypothesis test is its power, i.e. the probability to correctly reject the null hypothesis when it is false. It can be assessed through numerical simulation. In the Gaussian case, the procedure is straighforward since a closed form solution of the differential cumulants is known. From (5.7)

$$\kappa_{110} = -\frac{-\rho_{1,3}\rho_{2,3} + \rho_{1,2}}{\sigma_1\sigma_2(-2\,\rho_{1,3}\rho_{1,2}\rho_{2,3} + {\rho_{1,3}}^2 + {\rho_{1,2}}^2 - 1 + {\rho_{2,3}}^2)}. \tag{5.14}$$

We may, for instance, set $\rho_{1,3} = \rho_{2,3} = 0$ and $\sigma_1 = \sigma_2 = 1$ so that the differential cumulant simplifies to

$$\kappa_{110} = \frac{\rho_{1,2}}{1 - \rho_{1,2}^2}. \tag{5.15}$$

This is a quadratic equation in the correlation coefficient $\rho_{1,2}$. Any desired level of κ_{110} can hence be epressed in terms of $\rho_{1,2}$ and the power of the test be evaluated through Monte-Carlo simulations, where the test is carried out repeatedly and the power is estimated as the fraction of replications for which the test rejected the null hypothesis. Similarly the size of the test can be estimated as the fraction of replications where the test rejected the null hypothesis when the test is carried out repeatedly for a precision matrix which holds a zero in the appropriate position. The maximum likelihood and nonparametric bootstrap test can then be compared in terms of their size and power for different sample sizes.

When the true distribution is not Gaussian, the analysis of the test complicates. We may, however, exploit the fact that differential cumulants are invariant under

marginal transformation (Jones et al., 1996) in order to study special cases. In particular, the Gaussian copula allows us to investigate power and size of the bootstrap test for random variables which are distributed uniformly on a unit hypercube. Since the differential cumulants are unaffected by marginal transforms, they can indeed be set through the precision matrix as in the Gaussian case and the study would proceed along the same lines.

The properties of the bootstrap test are harder to study for arbitrary multivariate distributions. The challenge lies in the fact, that the differential cumulants are not necessarily simple functions of the parameters that allow to systematically sample from the distribution for a given differential cumulant. A non-systematic approach is however feasible where the parameters are changed, the differential cumulant is computed, the data sampled and the test carried out. The power of the test can then be approximated over bins as the fraction of tests which rejected the null hypothesis.

Once the power and the size are estimated a natural extension is to investigate how they change with the sample size, the choice of grid points or the numerical integration procedure. The approach adopted here is to numerically integrate through simple averaging over hypercubes where no particular attention is paid to the choice of grid points and the weighting is uniform. More sophisticated approaches such as Gaussian quadrature optimize the choice of grid points and the weighting associated to them in some optimal way. Investigating how the numerical integration scheme and the sample size affect the power of the test remains for future research. Finally, a challenge is to apply the hypotheses test to real data sets.

5.A Proofs

5.A.1 Preliminaries

We state, without proof, two standard results from multivariate calculus, a useful property of the convolution operator and a fact relating the vec operator to the Kronecker product.

Lemma 10 (Multivariate integration by substitution for a linear change of variables)**.** *Let A be invertible and g be a real-valued function with compact support. Then it holds that*

$$\int_{\mathbb{R}} g(y)dy = |A| \int_{\mathbb{R}} g(Ax)dx.$$

Proof. The proof follows from the fundamental theorem of calculus and is omitted. □

Lemma 11 (Multivariate Taylor expansion using Kronecker notation)**.** *Let $f : \mathbb{R}^d \longrightarrow \mathbb{R}^p$ have the property that every entry of $D^{\otimes r} f(x)$ is piecewise continuous. Then f has Taylor expansion:*

$$f(x+h) = \sum_{r=0}^{q} \frac{1}{r!} [I_p \otimes (h')^{\otimes r}] D^{\otimes r} f(x) + o(\|h\|^q) 1_p, \quad x, h \in \mathbb{R}^d.$$

Proof. See Baxandall and Liebeck (1986). □

Let $*$ denote the convolution operator, i.e.

$$(g * f)(x) := \int g(y) f(x-y) dy.$$

We make use of the fact that

$$(D^{\otimes r} g * f)(x) = (g * D^{\otimes r} f)(x).$$

Finally, the following relation between the vec operator and the Kronecker product shall be useful:

$$\operatorname{vec}(ABC) = (C' \otimes A) \operatorname{vec} B. \tag{5.16}$$

Setting $B = I_d$ shows that it holds for a column vector $x \in \mathbb{R}^d$ that

$$x \otimes x = \operatorname{vec}(xx').$$

5.A.2 Proof of Theorem 10

The proof of Theorem 10 is a modified version of Theorem 2 of Chacón et al. (2011). As mentioned, Chacón et al. (2011) derive the AMISE of the full derivative $D^{\otimes r}\hat{f}$ which naturally leads them to consider Euclidean norms. They derive expressions for the norms in terms of traces of matrices through the relationship $\|X\|^2 = \operatorname{tr}(XX')$. We show the equivalent result for the single component of the r-th derivative. The proof idea is standard in nonparametric asymptotic theory: It decomposes the AMISE into a bias and a variance term, applies Taylor expansions and shows that the remainder terms are of vanishing order.

Without loss of generality, we derive the AMISE of the first component of the r-th derivative. Our aim is to show that

$$\begin{aligned}\operatorname{AMISE}(e_1'\widehat{D^{\otimes r}}f(x;H)) = n^{-1}|H|^{-\frac{1}{2}}\, e_1'(H^{-\frac{1}{2}})^{\otimes r} R(D^{\otimes r}K)(H^{-\frac{1}{2}})^{\otimes r} e_1 \\ + \frac{\mu_2^2(K)}{4} e_1'(I_{d^r} \otimes \operatorname{vec}' H) R(D^{\otimes(r+2)}f)(I_{d^r} \otimes \operatorname{vec} H) e_1.\end{aligned}$$

We consider first $\operatorname{MISE}(e_1' D^{\otimes r}\hat{f}(x))$ before taking limits:

$$\begin{aligned}\operatorname{MISE}(e_1' D^{\otimes r}\hat{f}(x)) = \int_{\mathbb{R}^d} \underbrace{\left[e_1'\left(\mathbb{E}\widehat{D^{\otimes r}f}(x) - D^{\otimes r}f(x)\right)\right]^2}_{B^2(e_1' D^{\otimes r}\hat{f}(x))} \\ + \underbrace{\mathbb{E}(e_1'\widehat{D^{\otimes r}f}(x))^2 - \left(\mathbb{E}e_1'\widehat{D^{\otimes r}f}(x)\right)^2}_{\operatorname{var}(e_1'\widehat{D^{\otimes r}f}(x))} dx, \qquad (5.17)\end{aligned}$$

where $e_1 \in \mathbb{R}^d$. We consider the squared bias component first. Applying Lemma

10 we may write

$$\begin{aligned}\mathbb{E}\big(\widehat{D^{\otimes r}f}(x)\big) &= \mathbb{E}\big(n^{-1}\sum_{i=1}^{n} D^{\otimes r}K_H(x-X_i)\big)\\ &= \mathbb{E}\big(D^{\otimes r}K_H(x-X_1)(x)\big)\\ &= \int_{\mathbb{R}^d} D^{\otimes r}K_H(x-u)f(u)du\\ &= \int_{\mathbb{R}^d} K_H(x-u)D^{\otimes r}f(u)du\\ &= \int_{\mathbb{R}^d} K(z)D^{\otimes r}f(x-H^{-\frac{1}{2}}z)dz.\end{aligned}$$

An application of a Lemma 11 yields:

$$\begin{aligned}D^{\otimes r}f(x-H^{-\frac{1}{2}}z) = D^{\otimes r}f(x) - [I_{d^r}\otimes(z'H^{\frac{1}{2}})]D^{\otimes r+1}f(x)\\ + \frac{1}{2}[I_{d^r}\otimes(z'H^{\frac{1}{2}})^{\otimes 2}]D^{\otimes(r+2)}f(x) + o(\operatorname{tr}H)1_{d^r}.\end{aligned}$$

Since K is symmetric about the origin, the bias is of second order:

$$\begin{aligned}\mathbb{E}\big(\widehat{D^{\otimes r}f}(x)\big) - D^{\otimes r}f(x) &= \int_{\mathbb{R}^d} K(z)\frac{1}{2}[I_{d^r}\otimes(z'H^{\frac{1}{2}})^{\otimes 2}]D^{\otimes(r+2)}f(x) + o(\operatorname{tr}H)1_{d^r}dz\\ &= \frac{1}{2}\int_{\mathbb{R}^d} K(z)[I_{d^r}\otimes \operatorname{vec}'(H^{\frac{1}{2}}zz'H^{\frac{1}{2}})]D^{\otimes(r+2)}f(x) + o(\operatorname{tr}H)1_{d^r}dz\\ &= \frac{1}{2}\int_{\mathbb{R}^d} K(z)[I_{d^r}\otimes \operatorname{vec}'(zz')(H^{\frac{1}{2}})^{\otimes 2}]D^{\otimes(r+2)}f(x) + o(\operatorname{tr}H)1_{d^r}dz\\ &= \frac{1}{2}I_{d^r}\otimes\int_{\mathbb{R}^d} K(z)[\operatorname{vec}'(zz')dz(H^{\frac{1}{2}})^{\otimes 2}]D^{\otimes(r+2)}f(x) + o(\operatorname{tr}H)1_{d^r}\\ &= \frac{\mu_2(K)}{2}I_{d^r}\otimes[vec'(I_d)(H^{\frac{1}{2}})^{\otimes 2}]D^{\otimes(r+2)}f(x) + o(\operatorname{tr}H)1_{d^r}\\ &= \frac{\mu_2(K)}{2}(I_{d^r}\otimes \operatorname{vec}'H)D^{\otimes(r+2)}f(x) + o(\operatorname{tr}H)1_{d^r}.\end{aligned}$$

Thus, for the squared bias component of the MISE it holds that

$$\int_{\mathbb{R}^d} B^2(e_1'D^{\otimes r}\hat{f}(x))dx = \frac{\mu_2^2(K)}{4}e_1'(I_{d^r}\otimes \operatorname{vec}'H)R(D^{\otimes(r+2)}f(x))(I_{d^r}\otimes \operatorname{vec}H)e_1 + o(\operatorname{tr}^2\{H\})$$

where $R(D^{\otimes(r+2)}f(x)) := \int_{\mathbb{R}^d} D^{\otimes(r+2)}f(x)(D^{\otimes(r+2)}f(x))'dx$.

Consider next the integrated variance term in (5.17):

$$\int_{\mathbb{R}^d}\operatorname{var}(e_1'\widehat{D^{\otimes r}f}(x))dx = \int_{\mathbb{R}^d}\mathbb{E}(e_1'\widehat{D^{\otimes r}f}(x))^2 - \big(\mathbb{E}e_1'\widehat{D^{\otimes r}f}(x)\big)^2dx.$$

For the second moment term it holds that

$$\begin{aligned}
\int_{\mathbb{R}^d} \mathbb{E}(e_1' \widehat{D^{\otimes r} f}(x))^2 dx &= n^{-1} \int_{\mathbb{R}^d} E(e_1' D^{\otimes r} K_H(x - X_1))^2 dx \\
&= n^{-1} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (e_1' D^{\otimes r} K_H(x - y))^2 f(y) dx dy \\
&= n^{-1} \int_{\mathbb{R}^d} (e_1' D^{\otimes r} K_H(x))^2 dx \\
&= n^{-1} \int_{\mathbb{R}^d} (e_1' |H|^{-\frac{1}{2}} (H^{-\frac{1}{2}})^{\otimes r} D^{\otimes r} K(H^{-\frac{1}{2}} x))^2 dx \\
&= n^{-1} |H|^{\frac{1}{2}} \int_{\mathbb{R}^d} (e_1' |H|^{-\frac{1}{2}} (H^{-\frac{1}{2}})^{\otimes r} D^{\otimes r} K(z))^2 dz \\
&= n^{-1} |H|^{-\frac{1}{2}} e_1' (H^{-\frac{1}{2}})^{\otimes r} \int_{\mathbb{R}^d} D^{\otimes r} K(z) (D^{\otimes r} K(z))' dz (H^{-\frac{1}{2}})^{\otimes r} e_1 \\
&= n^{-1} |H|^{-\frac{1}{2}} e_1' (H^{-\frac{1}{2}})^{\otimes r} R(D^{\otimes r} K)(H^{-\frac{1}{2}})^{\otimes r} e_1.
\end{aligned}$$

A Taylor expansion shows that the squared moment term $\int_{\mathbb{R}^d} \left(\mathbb{E} e_1' \widehat{D^{\otimes r} f}(x)\right)^2 dx$ is of order $O(n^{-1})$, so that the integrated variance term can be written as:

$$\int_{\mathbb{R}^d} \operatorname{var}(e_1' \widehat{D^{\otimes r} f}(x)) dx = n^{-1} |H|^{-\frac{1}{2}} e_1' (H^{-\frac{1}{2}})^{\otimes r} R(D^{\otimes r} K)(H^{-\frac{1}{2}})^{\otimes r} e_1 + o(n^{-1} |H|^{-\frac{1}{2}}).$$

This completes the proof as one considers the limit as n goes to infinity.

5.A.3 Proof of Theorem 11

Theorem 11 is a particular case of Corollary 4, where $f_X = \phi_\Sigma$ and $K = \phi$. We need to proof that

$$\operatorname{AMISE}(e_1' \widehat{D^{\otimes 1} \phi_\Sigma}) + \operatorname{AMISE}(e_2' \widehat{D^{\otimes 1} \phi_\Sigma}) = \sum_{i=1}^{2} \operatorname{AIB}^2(e_i' D^{\otimes 1} \hat{f}) + \sum_{i=1}^{2} \operatorname{AIV}(e_i' D^{\otimes 1} \hat{f}),$$

where

$$\sum_{i=1}^{2} \operatorname{AIV}(e_i' D^{\otimes 1} \hat{f}) = n^{-1} |H|^{-\frac{1}{2}} C(H_{11}^{-1} + H_{22}^{-1})$$

and

$$\sum_{i=1}^{2} \mathrm{AIB}^2(e_i' D^{\otimes 1} \hat{f}) = \frac{1}{16} C |\Sigma|^{-\frac{1}{2}} \big[\operatorname{tr}(D) \operatorname{tr}^2(B) + 2\big(\operatorname{tr}(D) \operatorname{tr}(B^2) + 2 \operatorname{tr}(B) \operatorname{tr}(DB) \big) + 8 \operatorname{tr}(DB^2) \big].$$

Furthermore, the auxilliary matrices were defined as $B = \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}}$, $C = 2^{-(d+1)} \pi^{-\frac{d}{2}}$ and $D = \Sigma^{-\frac{1}{2}} I_d^{[2]} \Sigma^{-\frac{1}{2}}$.

We compute expressions for the asymptotic integrated variance and squared bias terms separately. We employ the fact that for any positive integer r it holds that

$$R(D^{\otimes r} \phi_\Sigma) = 2^{-(d+r)} \pi^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} (\Sigma^{-\frac{1}{2}} \mathbb{E}(zz') \Sigma^{-\frac{1}{2}})^{\otimes r}, \tag{5.18}$$

where $R(\cdot)$ as defined in (5.9) (Chacón et al., 2011).

In order to compute the expression for the squared bias in (5.13) we make use of (5.16) and (5.18).

The squared bias expression $\sum_{i=1}^{2} \mathrm{AIB}^2(e_i' D^{\otimes 1} \hat{f})$ is given by

$$
\begin{aligned}
&\operatorname{tr}\left[I_d^{[2]}(I_d \otimes \operatorname{vec}' H) R (D^{\otimes 2}\phi_\Sigma)(I_d \otimes \operatorname{vec} H)\right] \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \operatorname{tr}\left[I_d^{[2]}\big(I_d \otimes \operatorname{vec}' H\big)(\Sigma^{-\frac{1}{2}})^{\otimes 3}\mathbb{E}[(zz')^{\otimes 2}](\Sigma^{-\frac{1}{2}})^{\otimes 3}\big(I_d \otimes \operatorname{vec} H\big)\right] \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \operatorname{tr}\left[(\Sigma^{-\frac{1}{2}})^{\otimes 3}\big(I_d \otimes \operatorname{vec} H\big) I_d^{[2]}\big(I_d \otimes \operatorname{vec}' H\big)(\Sigma^{-\frac{1}{2}})^{\otimes 3}\mathbb{E}[(zz')^{\otimes 3}]\right] \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \operatorname{tr}\left[(\Sigma^{-\frac{1}{2}}) \otimes ((\Sigma^{-\frac{1}{2}})^{\otimes 2} \operatorname{vec} H) I_d^{[2]} (\Sigma^{-\frac{1}{2}}) \otimes (\operatorname{vec}' H (\Sigma^{-\frac{1}{2}})^{\otimes 2})\mathbb{E}[(zz')^{\otimes 3}]\right] \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \operatorname{tr}\left[(\Sigma^{-\frac{1}{2}}) I_d^{[2]} (\Sigma^{-\frac{1}{2}}) \otimes (\Sigma^{-\frac{1}{2}})^{\otimes 2} \operatorname{vec} H \operatorname{vec}' H (\Sigma^{-\frac{1}{2}})^{\otimes 2}\mathbb{E}[(zz')^{\otimes 3}]\right] \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \operatorname{tr}\left[(\Sigma^{-\frac{1}{2}}) I_d^{[2]} (\Sigma^{-\frac{1}{2}}) \otimes \operatorname{vec} B \operatorname{vec}' B \mathbb{E}[(zz')^{\otimes 3}]\right] \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \mathbb{E}\Big(\operatorname{tr}\left[(\Sigma^{-\frac{1}{2}}) I_d^{[2]} (\Sigma^{-\frac{1}{2}})(zz') \otimes \operatorname{vec} B \operatorname{vec}' B (zz')^{\otimes 2}\right]\Big) \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \mathbb{E}\Big(\operatorname{tr}\left[(\Sigma^{-\frac{1}{2}}) I_d^{[2]} (\Sigma^{-\frac{1}{2}})(zz')\right] \operatorname{tr}\left[\operatorname{vec} B (\operatorname{vec}' B \mathbb{E}[(zz')^{\otimes 2}])\right]\Big) \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \mathbb{E}\Big((z'\Sigma^{-\frac{1}{2}} I_d^{[2]} \Sigma^{-\frac{1}{2}} z) \operatorname{tr}\left[\operatorname{vec} B \operatorname{vec}'(zz' B zz')\right]\Big) \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \mathbb{E}\Big((z' D z)\left[\operatorname{vec}'(zz' B zz') \operatorname{vec} B\right]\Big) \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \mathbb{E}\Big((z' D z)\big[\sum_i \sum_j \sum_k \sum_l z_i z_j z_k z_l B_{ij} B_{kl}\big]\Big) \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \mathbb{E}\Big((z' D z)(z' B z)^2\Big) \\
&= \frac{1}{16}|\Sigma|^{-\frac{1}{2}}\, C \Big(\operatorname{tr}(D)\operatorname{tr}^2(B) + 2\big(\operatorname{tr}(D)\operatorname{tr}(B^2) + 2\operatorname{tr}(B)\operatorname{tr}(DB)\big) + 8\operatorname{tr}(DB^2)\Big)
\end{aligned}
$$

where $B = \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}}, C = 2^{-(d+1)}\pi^{-\frac{d}{2}}$ and $D = \Sigma^{-\frac{1}{2}} I_d^{[2]} \Sigma^{-\frac{1}{2}}$.

Reproducing from Corollary 4

$$\sum_{i=1}^{2} \mathrm{AIV}(e_i' D^{\otimes 1} \hat{f}) = \sum_{i=1}^{2} n^{-1} |H|^{-\frac{1}{2}} e_i' H^{-\frac{1}{2}} R(\nabla K) H^{-\frac{1}{2}} e_i. \tag{5.19}$$

By setting $r = 1$ and $\Sigma = I_d$ in (5.18) we obtain

$$\begin{aligned}
\sum_{i=1}^{2} e_i'(H^{-\frac{1}{2}}) R(\nabla\phi) H^{-\frac{1}{2}} e_i &= \operatorname{tr}\left(I_d^{[2]} H^{-\frac{1}{2}} R(\nabla\phi) H^{-\frac{1}{2}}\right) \\
&= C\mathbb{E}\left(\operatorname{tr}\left(I_d^{[2]} H^{-\frac{1}{2}} z z' H^{-\frac{1}{2}}\right)\right) \\
&= C\mathbb{E}\left(z' H^{-\frac{1}{2}} I_d^{[2]} H^{-\frac{1}{2}} z \right) \\
&= C \operatorname{tr}\left(H^{-\frac{1}{2}} I_d^{[2]} H^{-\frac{1}{2}}\right) \\
&= C \operatorname{tr}\left(I_d^{[2]} H^{-1}\right) \\
&= C(H_{11}^{-1} + H_{22}^{-1}),
\end{aligned} \tag{5.20}$$

where $C = 2^{-(d+1)}\pi^{-\frac{d}{2}}$ and H_{ij}^{-1} is the (i,j)-th element of the inverse bandwidth matrix H^{-1}. Substituting (5.20) into (5.19) we obtain

$$\sum_{i=1}^{2} \mathrm{AIV}(e_i' D^{\otimes 1} \hat{f}) = n^{-1} |H|^{-\frac{1}{2}} C(H_{11}^{-1} + H_{22}^{-1}).$$

5.A.4 Proof of Theorem 12

We are interested in the AMISE expression for the cross derivative with respect to the first two variables, which is the second entry of $\widehat{D^{\otimes 2} f}$. This leads us to consider the second contribution of the trace expressions, rather than the whole traces - both in the variance and in the squared bias part.

Define the matrix I_α^β as an α by α elementary matrix which holds a one as the β's diagonal entry and zeroes otherwise. Formally, $I_\alpha^\beta = (a_{ij})$, where $a_{ij} = 1$ if $i = j = \beta$ and $a_{ij} = 0$ otherwise. It holds that $I_{d^2}^2 = I_d^1 \otimes I_d^2$. The I_α^β matrix allows us to pick an arbitrary diagonal entry of a matrix A by computing the trace

of the product of I_α^β and A: $a_{\beta\beta} = \operatorname{tr}(I_\alpha^\beta A)$. Recall from (5.11)

$$\operatorname{AIB}^2(e_2' D^{\otimes 2} \hat{f}) = \frac{\mu_2^2(K)}{4} e_2'(I_{d^2} \otimes \operatorname{vec}' H) R (D^{\otimes 4} f)(I_{d^2} \otimes \operatorname{vec} H) e_2.$$

Substituting for f and K, the squared bias expressions becomes:

$$\begin{aligned}
&\operatorname{AIB}^2(e_2' \widehat{D^{\otimes 2}\phi_\Sigma}) \\
&= \operatorname{tr}\left[I_{d^2}^2 (I_{d^2} \otimes \operatorname{vec}' H) R (D^{\otimes 4}\phi_\Sigma)(I_{d^2} \otimes \operatorname{vec} H)\right] \\
&= 2^{-5}|\Sigma|^{-\frac{1}{2}}\, C \operatorname{tr}\left[(I_d^1 \otimes I_d^2)\big(I_{d^2} \otimes \operatorname{vec}' H\big)(\Sigma^{-\frac{1}{2}})^{\otimes 4}\mathbb{E}[(zz')^{\otimes 4}](\Sigma^{-\frac{1}{2}})^{\otimes 4}\big(I_{d^2} \otimes \operatorname{vec} H\big)\right] \\
&= 2^{-5}|\Sigma|^{-\frac{1}{2}}\, C \operatorname{tr}\left[(\Sigma^{-\frac{1}{2}})^{\otimes 4}\big(I_{d^2} \otimes \operatorname{vec} H\big)(I_d^1 \otimes I_d^2)\big(I_{d^2} \otimes \operatorname{vec}' H\big)(\Sigma^{-\frac{1}{2}})^{\otimes 4}\mathbb{E}[(zz')^{\otimes 4}]\right] \\
&= 2^{-5}|\Sigma|^{-\frac{1}{2}}\, C \operatorname{tr}\left[(\Sigma^{-\frac{1}{2}})^{\otimes 2} \otimes (\Sigma^{-\frac{1}{2}})^{\otimes 2} \operatorname{vec} H (I_d^1 \otimes I_d^2)(\Sigma^{-\frac{1}{2}})^{\otimes 2} \otimes \operatorname{vec}' H (\Sigma^{-\frac{1}{2}})^{\otimes 2}\mathbb{E}[(zz')^{\otimes 4}]\right] \\
&= 2^{-5}|\Sigma|^{-\frac{1}{2}}\, C \operatorname{tr}\left[(\Sigma^{-\frac{1}{2}})^{\otimes 2}(I_d^1 \otimes I_d^2)(\Sigma^{-\frac{1}{2}})^{\otimes 2} \otimes (\Sigma^{-\frac{1}{2}})^{\otimes 2} \operatorname{vec} H \operatorname{vec}' H (\Sigma^{-\frac{1}{2}})^{\otimes 2}\mathbb{E}[(zz')^{\otimes 4}]\right] \\
&= 2^{-5}|\Sigma|^{-\frac{1}{2}}\, C \operatorname{tr}\left[(\Sigma^{-\frac{1}{2}})^{\otimes 2}(I_d^1 \otimes I_d^2)(\Sigma^{-\frac{1}{2}})^{\otimes 2} \otimes \operatorname{vec} F \operatorname{vec}' F \mathbb{E}[(zz')^{\otimes 4}]\right] \\
&= 2^{-5}|\Sigma|^{-\frac{1}{2}}\, C\mathbb{E} \operatorname{tr}\left[(\Sigma^{-\frac{1}{2}})^{\otimes 2}(I_d^1 \otimes I_d^2)(\Sigma^{-\frac{1}{2}})^{\otimes 2}[(zz')^{\otimes 2}] \otimes \operatorname{vec} F \operatorname{vec}' F [(zz')^{\otimes 2}]\right] \\
&= 2^{-5}|\Sigma|^{-\frac{1}{2}}\, C\mathbb{E}\big(\operatorname{tr}\left[(I_d^1 \otimes I_d^2)(\Sigma^{-\frac{1}{2}}(zz')\Sigma^{-\frac{1}{2}})^{\otimes 2}\right](z'Fz)^2\big) \\
&= 2^{-5}|\Sigma|^{-\frac{1}{2}}\, C\mathbb{E}\big((z'Fz)(z'Gz)(z'Bz)^2\big),
\end{aligned}$$

where $F = \Sigma^{-\frac{1}{2}} I_d^1 \Sigma^{-\frac{1}{2}}$, $\quad G = \Sigma^{-\frac{1}{2}} I_d^2 \Sigma^{-\frac{1}{2}}$ and $B = \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}}$.

The last expression can be computed explicitly by an application of Theorem 5.1 of Magnus (1978):

$$\begin{aligned}
\operatorname{AIB}^2(e_2' \widehat{D^{\otimes 2}\phi_\Sigma}) = 2^{-5}|\Sigma|^{-\frac{1}{2}}\, C\Big(&\operatorname{tr}(F)\operatorname{tr}(G)\operatorname{tr}^2(B) \\
&+ 2\Big[\operatorname{tr}(F)\operatorname{tr}(G)\operatorname{tr}(B^2) + 2\operatorname{tr}(F)\operatorname{tr}(B)\operatorname{tr}(BG) \\
&\quad + 2\operatorname{tr}(G)\operatorname{tr}(B)\operatorname{tr}(FB) + \operatorname{tr}(FG)\operatorname{tr}^2(B)\Big] \\
&+ 8\Big[\operatorname{tr}(F)\operatorname{tr}(GB^2) + \operatorname{tr}(G)\operatorname{tr}(FB^2) + 2\operatorname{tr}(B)\operatorname{tr}(FBG)\Big] \\
&+ 4\Big[\operatorname{tr}(FG)\operatorname{tr}(B^2) + 2\operatorname{tr}(FB)\operatorname{tr}(BG)\Big] \\
&+ 16\Big[2\operatorname{tr}(FGB^2) + \operatorname{tr}(FBGB)\Big]\Big).
\end{aligned}$$

For the aymptotic variance term we obtain the expression

$$\begin{aligned}
\mathrm{AIV}(e_2' D^{\otimes 2} \hat{f}) &= n^{-1}|H|^{-\frac{1}{2}}\, e_2'(H^{-\frac{1}{2}})^{\otimes 2} R(D^{\otimes 2}\phi)(H^{-\frac{1}{2}})^{\otimes 2} e_2 \\
&= \operatorname{tr}\left(I_{d^2}^2 (H^{-\frac{1}{2}})^{\otimes 2} R(D^{\otimes 2}(\phi))(H^{-\frac{1}{2}})^{\otimes 2}\right) \\
&= \frac{1}{4} C \mathbb{E}\left(\operatorname{tr}\left(I_{d^2}^2 (H^{-\frac{1}{2}} z z' H^{-\frac{1}{2}})^{\otimes 2}\right)\right) \\
&= \frac{1}{4} C \mathbb{E}\left(\operatorname{tr}\left(I_d^1 \otimes I_d^2 (H^{-\frac{1}{2}} z z' H^{-\frac{1}{2}})^{\otimes 2}\right)\right) \\
&= \frac{1}{4} C \mathbb{E}\left(\operatorname{tr}\left((I_d^1 H^{-\frac{1}{2}} z z' H^{-\frac{1}{2}}) \otimes (I_d^2 H^{-\frac{1}{2}} z z' H^{-\frac{1}{2}})\right)\right) \\
&= \frac{1}{4} C \mathbb{E}\left(\operatorname{tr}\left(I_d^1 H^{-\frac{1}{2}} z z' H^{-\frac{1}{2}}\right) \operatorname{tr}\left(I_d^2 H^{-\frac{1}{2}} z z' H^{-\frac{1}{2}}\right)\right) \\
&= \frac{1}{4} C \mathbb{E}\left(\operatorname{tr}\left(I_d^1 H^{-\frac{1}{2}} z z' H^{-\frac{1}{2}}\right) \operatorname{tr}\left(I_d^2 H^{-\frac{1}{2}} z z' H^{-\frac{1}{2}}\right)\right) \\
&= \frac{1}{4} C E\big((z' D z)(z' E z)\big) \\
&= \frac{1}{4} C\big(\operatorname{tr}(D)\operatorname{tr}(E) + 2\operatorname{tr}(DE)\big) \\
&= \frac{1}{4} C\big(H_{11}^{-1} H_{22}^{-1} + 2 H_{21}^{-1} H_{12}^{-1}\big),
\end{aligned}$$

where $C = 2^{-(d+1)}\pi^{-\frac{d}{2}}$, $D = H^{-1} I_d^1$ and $E = H^{-1} I_d^2$. The second to last line can be computed by hand or, more elegantly, by Theorem 5.1 in Magnus (1978).

Chapter 6

Estimation of a functional of differential moments

6.1 Introduction

This chapter describes the estimation of a functional of local moments. It is built on Chapter 2 and is not related to intermediate chapters. Informally, we investigate by how much a differential moment changes as additional information through a related variable is taken into account. Does the inclusion of an associated variable Y significantly change our prediction for where, in a given interval, X is likely to fall?

Suppose we are interested in the local moment of a random variable X given that X falls in a certain intervall $A_X := [x_0 \pm \varepsilon]$. From Remark 1 in Chapter 2 we know that the local moment tells us how the probability mass changes in a local environment of x_0. For instance, if m_1^A is positive, we expect to see more of the realisations in the intervall $[x_0, x_0 + \varepsilon]$ than in the interval $[x_0 - \varepsilon, x_0]$.

Suppose further that another variable Y exists which is not independent of X. By how much does the conditional mean of X change as we further condition on Y falling into the intervall $A_Y := [y_0 \pm \varepsilon]$? Formally, we would like to quantify the difference between the differential moment of X given that both X and Y are in $A := A_X \times A_Y$ and the expected value of X given that X is in A_X and Y is anywhere.

Let Z denote the quantity of interest. We will refer to Z as the 'information gain in X through Y' or just 'information gain'. Z can be written as

$$Z := \lim_{\varepsilon \to 0} \frac{3}{\varepsilon^2} \big[E(X - x_0 | (X,Y) \in A) - E(X - x_0 | X \in A_X) \big]. \tag{6.1}$$

We suggest a histogram-type estimator based on the sample counterparts of the population moments and a local regression estimator. Asymptotic results are derived.

6.2 The local sample moment approach

The local sample moment approach has been studied in some detail by Mueller and Yan (2001). They show that local sample moments converge to their population counterparts in probability and that the limiting distribution is normal provided some regularity conditions are met. For the sake of completeness, we state their main results regarding the limiting distribution and the AMSE optimal choice of the window size.

One of their suggested applications is to estimate densities and derivatives through local sample moments. In contrast, we employ density and density derivative estimators in order to estimate a functional of local moments.

Given a sample $(x_1, y_1), ..., (y_n, y_n)$, a naive estimator of Z can be constructed from (6.1) by replacing population with sample moments. The differential moment of order k at ξ was defined in Definition 6 as

$$m^{\xi}_{k_1 \dots k_d} = \lim_{\varepsilon \to 0} \frac{1}{r(\varepsilon, k)} \mathbb{E}\left(\prod_{j=1}^{d} (X_j - \xi_j)^{k_j} \,\middle|\, X \in A \right),$$

where $r(\varepsilon, k)$ was defined as $r(\varepsilon, k) = \varepsilon^{|k|^+} \prod_{\substack{i=1, \\ k_i \in 2\mathbb{N}}}^{d} \frac{1}{k_i + 1} \prod_{\substack{i=1, \\ k_i \in 2\mathbb{N}+1}}^{d} \frac{1}{k_i + 2}$.

Their sample counterparts are:

$$\hat{m}^{\xi}_{k_1 \dots k_d} = \frac{1}{r(\varepsilon, k)} \frac{\sum_{i=1}^{n} \prod_{j=1}^{d} (x_{i,j} - \xi_j)^{k_j} \mathbb{1}(x_i \in A)}{\sum_{i=1}^{n} \mathbb{1}(x_i \in A)}.$$

In order to estimate Z we only need the sample moments

$$\hat{m}_{10}^{(x_0,y_0)} = \frac{3}{\varepsilon^2} \frac{\sum_{i=1}^n (x_i - x_0)\, \mathbb{1}((x_i, y_i) \in A)}{\sum_{i=1}^n \mathbb{1}((x_i, y_i) \in A)}$$

and

$$\hat{m}_1^{x_0} = \frac{3}{\varepsilon^2} \frac{\sum_{i=1}^n (x_i - x_0)\, \mathbb{1}(x_i \in A_X)}{\sum_{i=1}^n \mathbb{1}(x_i \in A_X)}.$$

The naive estimator is then given by

$$\hat{Z}_N = \hat{m}_{10}^{(x_0,y_0)} - \hat{m}_1^{x_0}, \tag{6.2}$$

where the subscript N is chosen to indicate that this is the naive estimator. It computes the sample average of the x-values for which the pair (x_i, y_i) is in A and subtracts the sample average of x-values for which x_i is in A_X.

We state two specialised versions of Theorem 3.1 of Mueller and Yan (2001). Let $\xrightarrow{D}$ and $\xrightarrow{P}$ denote convergence in distribution and in probability respectively.

Theorem 13 (Convergence of local sample moments)**.**

1. *Univariate case: Assume that f_X is twice continuously differentiable in x_0, that $n\varepsilon^3 \to \infty$ and $n\varepsilon^7 \to 0$ as $n \to \infty$ then*

$$\sqrt{n}\varepsilon^{\frac{3}{2}}(\hat{m}_1^{x_0} - m_1^{x_0}) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{6 f_X(x_0)}\right)$$

   and hence $\hat{m}_1^{x_0} \xrightarrow{P} m_1^{x_0}$.

2. *Bivariate case: Assume that $f_{X,Y}$ is twice continuously differentiable in (x_0, y_0), that $n\varepsilon^4 \to \infty$ and $n\varepsilon^8 \to 0$ as $n \to \infty$ then*

$$\sqrt{n}\varepsilon^2(\hat{m}_{10}^{(x_0,y_0)} - m_{10}^{(x_0,y_0)}) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{12 f_{X,Y}(x_0, y_0)}\right)$$

   and hence $\hat{m}_{10}^{(x_0,y_0)} \xrightarrow{P} m_{10}^{(x_0,y_0)}$.

Proof. This is Theorem 3.1 of Mueller and Yan (2001). □

Corollary 6 (Asymptotic distribution of Z_N). *Assume that $f_{X,Y}$ is twice continuously differentiable in (x_0, y_0), that $n\varepsilon^4 \to \infty$ and $n\varepsilon^7 \to 0$ as $n \to \infty$.*

Then Z_N is consistent and

$$\sqrt{n}\varepsilon^2(\hat{Z}_N - Z) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{12 f_{X,Y}(x_0, y_0)}\right).$$

Proof. According to (6.2), $\hat{Z}_N$ can be written as $\hat{Z}_N = \hat{m}_{10}^{(x_0,y_0)} - \hat{m}_1^{x_0}$. Since $\hat{m}_{10}^{(x_0,y_0)}$ converges slower than $\hat{m}_1^{x_0}$ it dominates the asymptotic behaviour of $\hat{Z}_N$ and the result follows. □

6.3 The conditional density approach

6.3.1 Introduction

A local polynomial estimator makes use of a different representation of Z in terms of conditional densities. Using Corrolary 2, we may write Z as

$$\begin{aligned} Z &= \frac{\frac{\partial}{\partial x} f_{X,Y}(x,y)}{f_{X,Y}(x,y)} - \frac{\frac{d}{dx} f_X(x)}{f_X(x)} \\ &= \frac{\frac{\partial}{\partial x} f_{Y|X}(y|x)}{f_{Y|X}(y|x)}. \end{aligned} \tag{6.3}$$

Fan et al. (1996b) show how conditional densities and their derivatives with respect to the conditioning variable can be estimated through local polynomial regression. This section shows how an estimator of Z can be based on their approach. Asymptotic normality is proved for a joint estimator of information gains in X through Y and Y through X.

6.3.2 Motivation and description of the estimator

The description of the estimator is rather technical, so we first give an overview. We then present the estimator for the information gain of X through Y. The estimator for the information gain of Y through X is symmetric. We spend the

last part of this subsection on this reverse regression in order to introduce the necessary notation.

Recall from (6.3) that Z can be written as

$$Z = \frac{\frac{\partial}{\partial x} f_{Y|X}(y|x)}{f_{Y|X}(y|x)}.$$

A conditional density estimator of Z estimates the conditional density of Y given X and its derivative with respect to x and divides the latter by the former. Hence, the aim is to estimate $f^{(j)}_{Y|X}(y|x)$ for $j = 0$ and $j = 1$, where differentiation is with respect to the conditioning variable.

In general, local polynomial regression assumes a regression relation

$$g(Y_i) = m(X_i) + \sigma(X_i)\varepsilon_i, \quad \forall\, 1 \leq i \leq n,$$

where g and m are functions from $\mathbb{R}$ to $\mathbb{R}$, $E(\varepsilon_i) = 0, Var(\varepsilon_i) = 1$ and ε_i is independent of X_i for all $1 \leq i \leq n$. There are two key ideas to the approach. The first idea is to choose an appropriate function for g which approximately turns the dependent variable into the conditional density $f(y|x)$. The second idea is to choose the regression function m to be a polynomial in x. This is because the coefficent of a polynomial are multiples of its derivatives, e.g. if $y = a_0 + a_1 x + ... + a_p x^p$ then $a_j = \frac{1}{j!}\frac{d^j y}{dx^j}$. The coefficients of the polynomial, however, are exactly what is estimated through least square regression.

In ordinary least square regression the explanation and prediction of the dependent variable is often the primary goal and the least square estimators are useful tools to achieve this aim. In local polynomial regression the estimation of the coefficients is the final target.

Let W be a kernel function. We maintain the standard notation $W_h(\cdot) := \frac{1}{h}W(\frac{\cdot}{h})$. For polynomial regression of order p, the quanity to be minimised as a function of $b \in \mathbb{R}^{p+1}$ takes the form:

$$\sum_{i=1}^{n} \left(g(Y_i) - \sum_{j=0}^{p} b_j (X_i - x)^j \right)^2 W_{h_X}(X_i - x).$$

Setting $g(Y_i) := K_{h_Y}(Y_i - y)$ allows us to estimate conditional densities and their derivatives. A Taylor expansion and a change of variable show that

$$E\{K_{h_Y}(Y-y)|X=x\} = f_{Y|X}(y|x) + \frac{1}{2}h_Y^2\tilde{\mu}_2\frac{\partial^2}{\partial y^2}f_{Y|X}(y|x) + o(h_Y^2), \tag{6.4}$$

where $\tilde{\mu}_j := \int z^j K(z)dz$ denotes the j-th central moment of the kernel function K. From this it follows that $E\{K_{h_Y}(Y-y)|X=x\} \xrightarrow{P} f(y|x)$ as $h_Y \to 0$, which makes it suitable as a regression target. A Taylor expansion of $f_{Y|X}(y|X)$ about x shows that the target is linear in the polynomials of $(x - X)$:

$$f_{Y|X}(y|X) = f_{Y|X}(y|x) + \sum_{j=1}^{p}\frac{f_{Y|X}^{(j)}(y|x)}{j!}(X-x)^j + o_P\{(X-x)^p\}, \tag{6.5}$$

where $f_{Y|X}^{(j)}(y|x) := \frac{\partial^j}{\partial x^j}f_{Y|X}(y|x)$ is the j-th derivative of the conditional density $f_{Y|X}$ with respect to the conditioning variable.

It is convenient to define the following quantities: Let $X_x \in \mathbb{R}^{n\times(p+1)}$ be the design matrix:

$$X_x = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1-x)^p \\ \vdots & \ddots & & \vdots \\ 1 & (X_n - x) & \cdots & (X_n-x)^p \end{pmatrix}. \tag{6.6}$$

Let $\mathcal{W} \in \mathbb{R}^{n\times n} = \mathrm{diag}(W_{h_X}(x - X_i)_{1\le i\le n})$ be a diagonal matrix such that

$$\mathcal{W}_{i,j} = \begin{cases} W_{h_X}(X_i - x) & \text{if } i = j\ , \\ 0 & \text{otherwise.} \end{cases} \tag{6.7}$$

Furthermore, define $\mathcal{Y} := (K_{h_Y}(Y_i - y))_{1\le i\le n}$. Using these definition, the estimator of $\left(\frac{f_{Y|X}^{(j)}(y|x)}{j!}\right)_{0\le j\le p}$ is given by

$$\hat{\beta} = \mathrm{argmin}_b(\mathcal{Y} - X_x b)'\mathcal{W}(\mathcal{Y} - X_x b).$$

We assume that $X_x'\mathcal{W}X_x$ is of full rank. Then, we can write

$$\hat{\beta} = (X_x'\mathcal{W}X_x)^{-1}X_x'\mathcal{W}\mathcal{Y}. \tag{6.8}$$

It is common to start indexing $\hat{\beta}$ at zero. An estimator of the conditional density $f_{Y|X}(y|x)$ is given by $\hat{\beta}_0$. The j-th derivative of $f_{Y|X}(y|x)$ with respect to the conditioning variable can be estimated through $j!\hat{\beta}_j$.

Remark 3. The estimator $\hat{\beta}$ is biased from two different sources: The kernel estimation introduces a bias of order h_Y^2 as seen in (6.4) and the polynomial approximation introduces a bias of order $\sup_{1\leq i\leq n}\{(X_i - x)\}$ as seen in (6.5). How well the approximation (6.4) holds, depends in particular on the local smoothness of $f_{Y|X}$, for smooth functions are characterised by small second derivatives in absolute value.

Fan et al. (1996a) investigate the bias / variance trade-off for different orders of polynomial fit. Let j be the order of derivative being estimated and p be the order of polynomial fit. The order of the fit shall be defined as $p - j$. In general, a larger p is associated with a smaller asymptotic bias and a larger asymptotic variance. Yet, when passing from an even order fit to the consecutive odd order fit, the asymptotic variance does not increase. Odd order fits are hence superior. It is recommended to choose the parsimonious $p = j + 1$ (Fan and Yao, 2003).

Remark 4. The above approach includes the famous Rosenblatt estimator if $p = 0$ (Rosenblatt, 1969). The local-linear estimator ($p = 1$) was, for instance, investigated by Fan (1993). Whilst it has the advantage of a smaller bias compared to the Rosenblatt estimator, the estimated density function is neither restricted to be non-negative nor to integrate to 1.

With β_1 and β_0 as defined in (6.8) a conditional density estimator of Z is given by

$$\hat{Z}_{CD} = \frac{\beta_1}{\beta_0}.$$

The subscript CD stands for conditional density. As we estimate first derivatives, we consider a local quadratic estimator, i.e. we set $p = 2$ in (6.6). Given Remark 3 this is bias efficient compared to a local linear estimator.

We are interested in jointly estimating the information gain in X through Y and Y through X and denote the latter by $\tilde{Z}_{CD}$. The rest of this subsection is devoted to the reverse estimation of $W_{h_X}(X-x)$ on $(Y-y)$.

For simplicity, we consider the independent case, i.e. X_i is independent of X_j and Y_j whenever $i \neq j$ and similarly for Y. We assume that the same bandwidth parameter h_X is used in both regressions and equally for h_Y.

Some additional notation is needed for the weighted regression of $W_{h_X}(X-x)$ on polynomials in $(Y-y)$: Let $Y_y \in \mathbb{R}^{n\times 3}$ and $\mathcal{K} \in \mathbb{R}^{n\times n}$ be defined analogously to (6.6) and (6.7) as the design matrix for the regression of $\mathcal{X} := (W_{h_X}(X_i-x))_{1\leq i\leq n}$ on $(Y-y)$

$$Y_y = \begin{pmatrix} 1 & (Y_1-y) & (Y_1-y)^2 \\ \vdots & \vdots & \vdots \\ 1 & (Y_n-y) & (Y_n-y)^2 \end{pmatrix}$$

and the diagonal kernel matrix $\mathcal{K} = \operatorname{diag}(K_{h_Y}(y-Y_i)_{1\leq i\leq n})$. A tilde version of a variable defined in the regression setting for $K_{h_Y}(Y-y)$ on $(X-x)$ denotes its equivalent in the regression of $W_{h_X}(X-x)$ on $(Y-y)$. Hence, assuming full rank of $Y_y\mathcal{K}Y_y$, the local polynomial estimator for the regression of $W_{h_X}(X-x)$ on $(Y-y)$ is given by:

$$\hat{\tilde{\beta}} = (Y_y\mathcal{K}Y_y)^{-1}Y_y\mathcal{K}\mathcal{X}. \tag{6.9}$$

With $\hat{\beta}$ and $\hat{\tilde{\beta}}$ as defined in equations (6.8) and (6.9) respectively, we can define the vectors θ and $\hat{\theta} \in \mathbb{R}^4$ which hold the four quantities of interest and their estimators respectively:

$$\theta = \begin{pmatrix} f(y|x) \\ \frac{\partial f(y|x)}{\partial x} \\ f(x|y) \\ \frac{\partial f(x|y)}{\partial y} \end{pmatrix} \qquad \hat{\theta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\tilde{\beta}}_0 \\ \hat{\tilde{\beta}}_1 \end{pmatrix}. \tag{6.10}$$

The vector of information gains can be expressed as

$$\begin{pmatrix} \hat{Z}_{CD} \\ \hat{\tilde{Z}}_{CD} \end{pmatrix} = \begin{pmatrix} \frac{\hat{\theta}_2}{\hat{\theta}_1} \\ \frac{\hat{\theta}_4}{\hat{\theta}_3} \end{pmatrix}.$$

To avoid a cluttered notation, we write $\hat{Z}_1 := (\hat{Z}_{CD}, \hat{\tilde{Z}}_{CD})$ and define the function $g : \mathbb{R}^4 \to \mathbb{R}^2$

$$g(\theta) = \begin{pmatrix} \frac{\theta_2}{\theta_1} \\ \frac{\theta_4}{\theta_3} \end{pmatrix}, \tag{6.11}$$

so that $\hat{Z}_1 = g(\hat{\theta})$. The Jacobian matrix of g is given by

$$J_g(\theta) = \begin{pmatrix} -\frac{\theta_2}{\theta_1^2} & \frac{1}{\theta_1} & 0 & 0 \\ 0 & 0 & -\frac{\theta_4}{\theta_3^2} & \frac{1}{\theta_3} \end{pmatrix}. \tag{6.12}$$

It will be needed at a later stage, when we infer the asymptotic distribution of $\hat{Z}_1$ from the asymptotic distribution of $\hat{\theta}$.

6.3.3 Asymptotic properties

This section deals with the properties of $\hat{Z}_1$ when the sample size n goes to infinity and the associated bandwidth processes h_X and h_Y go to zero. A standard assumption in univariate kernel density estimation is to require that

$$nh \longrightarrow \infty \text{ as } n \longrightarrow \infty \text{ and } h \longrightarrow 0$$

for some bandwidth process h. Similarly, we need to specify the rate at which h_X and h_Y converge to zero. This turns out to be a rather important determinant of the exact asymptotic distribution of Z_1. We will assume throughout that

$$nh_X^3 h_Y \longrightarrow \infty, \quad nh_X h_Y^3 \longrightarrow \infty, \quad h_X, h_Y \longrightarrow 0$$
$$\text{and that there exists a } C \in \mathbb{R} \text{ such that } \frac{h_X}{h_Y} \longrightarrow C.$$

The last condition ensures that neither bandwidth process dominates the other.

Under suitable further regularity conditions, Fan and Gijbels (1996) prove that local polynomial estimators for conditional densities and their derivatives are consistent. We can get some first insight into the asymptotics of $\hat{Z}_1$ by the continuous mapping theorem:

Theorem 14 (Continuous mapping theorem)**.** *Let $g : \mathbb{R}^k \to \mathbb{R}^l$ be a continuous function and $\{X_n\}$ be a sequence of random variables taking values in $\mathbb{R}^k$. Then it holds that*

$$X_n \xrightarrow{D} X \Longrightarrow g(X_n) \xrightarrow{D} g(X).$$
$$X_n \xrightarrow{P} X \Longrightarrow g(X_n) \xrightarrow{P} g(X).$$

Proof. See for instance White (2000). □

The function g, as defined in (6.11), is continuous everywhere since densities are assumed to be strictly positive. Consequently, $\hat{Z}_1$ is consistent. We devote the rest of this subsection to the asymptotic distribution of $\hat{Z}_1$. Similarly to Fan and Yao (2003), we proof joint asymptotic normality of $(\hat{\beta}, \hat{\tilde{\beta}})'$. Since our results differ with respect to the expression for the asymptotic variance, and, since we have an extended covariance matrix to compute, we provide a rather detailed proof. As an immediate consequence we obtain the asymptotic distribution of $\hat{\theta}$.

We then employ the 'Delta-method', which allows us to derive the asymptotic distribution of an estimator which can be expressed as a continuous transformation of an estimator with known asymptotic distribution. It is based on a Taylor expansion of g coupled with an application of the continuous mapping theorem for weak convergence and Slutsky's lemma.

Finally, we state, without proof, the asympotic distribution of $\hat{\theta}$ when the local linear fit is employed instead of the local quadratic fit. This corresponds to setting $p = 1$ in (6.6). The bias of $\hat{\theta}$ will be shown to increase compared to the quadratic fit, as is expected given the discussion in Remark 3.

Regression $K_{h_Y}(Y-y)$ on X	Regression $W_{h_X}(X-x)$ on Y
$\mu_j := \int z^j W(z)dz$	$\tilde{\mu}_j := \int z^j K(z)dz$
$\nu_j := \int z^j W^2(z)dz$	$\tilde{\nu}_j := \int z^j K^2(z)dz$
$S := (\mu_{j+l})_{0\le j,l\le p} \in \mathbb{R}^{3\times 3}$	$\tilde{S} := (\tilde{\mu}_{j+l})_{0\le j,l\le p} \in \mathbb{R}^{3\times 3}$
$S^\star := (\nu_{j+l})_{0\le j,l\le p} \in \mathbb{R}^{3\times 3}$	$\tilde{S}^\star := (\tilde{\nu}_{j+l})_{0\le j,l\le p} \in \mathbb{R}^{3\times 3}$
$H_X := \mathrm{diag}(1, h_X, h_X^2) \in \mathbb{R}^{3\times 3}$	$H_Y := \mathrm{diag}(1, h_Y, h_Y^2) \in \mathbb{R}^{3\times 3}$
$m(x,y) := E\{K_{h_y}(Y-y)\|X=x\} \in \mathbb{R}$	$\tilde{m}(x,y) := E\{W_{h_X}(X-x)\|Y=y\} \in \mathbb{R}$
$m_j(x,y) := \begin{cases} m(x,y) & j=0 \\ \frac{1}{j!}\frac{\partial^j}{\partial x^j} m(x,y) & j=1,2 \end{cases}$	$\tilde{m}_j(x,y) := \begin{cases} \tilde{m}(x,y) & j=0 \\ \frac{1}{j!}\frac{\partial^j}{\partial x^j} \tilde{m}(x,y) & j=1,2 \end{cases}$
$\beta := (m_j(x,y))_{0\le j\le 2} \in \mathbb{R}^3$	$\tilde{\beta} := (\tilde{m}_j(x,y))_{0\le j\le 2} \in \mathbb{R}^3$

Table 6.1 – Definitions of quantities appearing in the asymptotic expression of $\hat{Z}_1$, the joint conditional density estimator of prediction gains.

Table 6.1 defines various quantities that appear in the expression of the asymptotic distribution of $\hat{Z}_1$. We further need the diagonal matrix $H_{X,Y} \in \mathbb{R}^{6\times 6}$:

$$H_{X,Y} := \begin{pmatrix} H_X & 0 \\ 0 & H_Y \end{pmatrix}.$$

Finally, let $r_n := \sqrt{nh_Xh_Y}$.

In this section we refer to the following two assumptions as the standard regularity conditions:

1. The kernel functions K and W are symmetric with bounded support.

2. The conditional densities $f_{X|Y}$ and $f_{Y|X}$ have bounded continuous third order derivatives with respect to x and y respectively at (x,y).

The following theorem states the asymptotic distribution of the conditional density estimators under the assumption that neither bandwidth process is dominant.

This assumption is not necessary for asymptotic normality and only effects the asymptotic bias and covariance. The relevant matrices are given explicitly in Appendix 6.B.

Theorem 15 (Asymptotic distribution of $(\hat{\beta}, \hat{\tilde{\beta}})'$)**.** *Assume the existence of a constant $C \in \mathbb{R}$ such that $\frac{h_X}{h_Y} \longrightarrow C$. It then holds, under the standard regularity conditions, that the conditional density estimator $(\hat{\beta}, \hat{\tilde{\beta}})'$ is asymptotically normal:*

$$r_n H_{X,Y} \left\{ \begin{pmatrix} \hat{\beta} \\ \hdashline \hat{\tilde{\beta}} \end{pmatrix} - \begin{pmatrix} \beta \\ \hdashline \tilde{\beta} \end{pmatrix} - \Sigma_1^{-1} b \right\} \xrightarrow{D} \mathcal{N}\Big(0, \Sigma_\beta\Big),$$

where $r_n := \sqrt{n h_X h_Y}$, $\Sigma_\beta := f_{X,Y}(x,y)\Sigma_1^{-1} \left(\begin{array}{c:c} \tilde{\nu}_0 S^\star & \Sigma \\ \hdashline \Sigma & \nu_0 \tilde{S}^\star \end{array}\right) \Sigma_1^{-1}$,

$\Sigma := (\tilde{\nu}_j \nu_k)_{0 \le j,k \le 2}$, $\Sigma_1 := \left(\begin{array}{c:c} f_X(x) S & 0 \\ \hdashline 0 & f_Y(y)\tilde{S} \end{array}\right)$,

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \frac{1}{6} h_X^3 f_X(x) \begin{pmatrix} h_X \frac{\partial^4}{\partial x^4} f_{Y|X}(y|x)\mu_4 \\ \frac{\partial^3}{\partial x^3} f_{Y|X}(y|x)\mu_4 \\ h_X \frac{\partial^4}{\partial x^4} f_{Y|X}(y|x)\mu_6 \end{pmatrix} \{1 + o_P(1)\}$$

and

$$\begin{pmatrix} b_4 \\ b_5 \\ b_6 \end{pmatrix} = \frac{1}{6} h_Y^3 f_Y(y) \begin{pmatrix} h_Y \frac{\partial^4}{\partial y^4} f_{X|Y}(x|y)\tilde{\mu}_4 \\ \frac{\partial^3}{\partial y^3} f_{X|Y}(x|y)\tilde{\mu}_4 \\ h_Y \frac{\partial^4}{\partial y^4} f_{X|Y}(x|y)\tilde{\mu}_6 \end{pmatrix} \{1 + o_P(1)\}. \tag{6.13}$$

Proof. The proof is deferred to Appendix 6.A.1. □

The entries $\hat{\beta}_3$ and $\hat{\beta}_6$ hold the derivative estimators for the second derivatives of $f_{Y|X}$ and $f_{X|Y}$ with respect to the conditioning variable respectively. We had chosen to include them into the estimation in view of Remark 3. Their inclusion lowers the bias of $\hat{\beta}_1$, whilst leaving the variance unchanged. However, since we

are not interested in second order derivatives of the conditional density, it will be convenient to filter out the relevant first two entries of $\hat{\beta}$ and $\hat{\tilde{\beta}}$. To this end, define E' as the projection matrix from $\mathbb{R}^3$ to $\mathbb{R}^2$ which chooses the first two rows:

$$E' := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

With that notation, we can write:

$$\hat{\theta} := \left(\begin{array}{c:c} E' & 0 \\ \hdashline 0 & E' \end{array}\right) \begin{pmatrix} \hat{\beta} \\ \hdashline \hat{\tilde{\beta}} \end{pmatrix}$$

The following corollary to Theorem 15 gives the asymptotic distribution of the $(1, 2, 4, 5)$ entries of $\hat{\beta}$ which form $\hat{\theta}$. Furthermore, a more explicit expression for the bias vector is provided. The vectors β and $\tilde{\beta}$ appearing in Theorem 15 do not hold the conditional densities and their derivatives but instead the approximating quantities $E\{K_{h_y}(Y - y)|X = x\}, E\{W_{h_X}(X - x)|Y = y\}$ and their derivatives. This introduces a second source of bias as explained in Remark 3. As it turns out, under the assumption that the bandwidth processes h_X and h_Y converge to zero at the same rate, this part of the bias dominates the bias $\Sigma_1^{-1}b$ which arises from the polynomial approximation.

Corollary 7 (Distribution of $\hat{\theta}$)**.** *Assume the standard regularity conditions hold. Define* $H := diag(1, h_X, 1, h_Y)$*. Then* $\hat{\theta}$ *as defined in* (6.10) *is asymptotically normal and it holds that*

$$r_n H(\hat{\theta} - \theta - \zeta) \xrightarrow{D} N(0, \Sigma_\theta),$$

where

$$\zeta := \begin{pmatrix} \frac{1}{2}h_Y^2\tilde{\mu}_2\frac{\partial^2 f_{Y|X}(y|x)}{\partial y^2} \\ \frac{1}{2}h_Y^2\tilde{\mu}_2\frac{\partial^3 f_{Y|X}(y|x)}{\partial x\partial y^2} + \frac{1}{6}h_X^2\frac{\mu_4}{\mu_2}\frac{\partial^3 f_{Y|X}(y|x)}{\partial x^3} \\ \frac{1}{2}h_X^2\mu_2\frac{\partial^2 f_{X|Y}(x|y)}{\partial x^2} \\ \frac{1}{2}h_X^2\mu_2\frac{\partial^3 f_{X|Y}(x|y)}{\partial y\partial x^2} + \frac{1}{6}h_Y^2\frac{\tilde{\mu}_4}{\tilde{\mu}_2}\frac{\partial^3 f_{X|Y}(x|y)}{\partial y^3}f_Y(y) \end{pmatrix} \tag{6.14}$$

and

$$\Sigma_\theta := f_{X,Y}(x,y)\left(\begin{array}{c:c} \tilde{\nu}_0 f_X^{-2}(x)E'S^{-1}\tilde{S}^\star S^{-1}E & f_X^{-1}(x)f_Y^{-1}(y)E'S^{-1}\Sigma\tilde{S}^{-1}E \\ \hdashline f_X^{-1}(x)f_Y^{-1}(y)E'\tilde{S}^{-1}\Sigma'S^{-1}E & \nu_0 f_Y^{-2}(y)E'\tilde{S}^{-1}S^\star\tilde{S}^{-1}E \end{array}\right). \tag{6.15}$$

Proof. The proof is deferred to Appendix 6.A.2. The covariance matrix Σ_θ is computed explicitly in Appendix 6.B.1.2. □

Corollary 7 shows that the estimators of the conditional densities converge quicker than the estimators of the derivatives by a factor of order h_X. Hence, we would expect the limiting distribution of $\hat{Z}_1$ to depend only on the limiting distributions of the derivative estimators. The next theorem states the asymptotic distribution of $\hat{Z}_1$ and confirms that this is indeed the case.

Theorem 16 (Asymptotic distribution of $\hat{Z}_1$)**.** *Let ζ be defined as in (6.14). Assume that g, as defined in (6.11), is continuously differentiable about $\xi := \theta + \zeta \in \mathbb{R}^4$ and that there exists a constant $C \in \mathbb{R}$ such that $\frac{h_X}{h_Y} \longrightarrow C$. It then holds, under the standard regularity conditions, that $\hat{Z}_1$ is normally distributed asymptotically. Specifically,*

$$r_n h_X(\hat{Z}_1 - g(\xi)) \xrightarrow{D} N(0, \Sigma_{\hat{Z}_1}),$$

where

$$\Sigma_{\hat{Z}_1} := \begin{pmatrix} \frac{1}{\xi_1^2}\frac{f_{Y|X}(x,y)}{f_X(x)}\frac{\nu_2\tilde{\nu}_0}{\mu_2^2} & 0 \\ 0 & \frac{C^2}{\xi_3^2}\frac{f_{X|Y}(x,y)}{f_Y(y)}\frac{\nu_0\tilde{\nu}_2}{\tilde{\mu}_2^2} \end{pmatrix}.$$

Proof. The proof is deferred to Appendix 6.A.3. □

Note that the asymptotic covariance is zero. The proof of Theorem 16 demonstrates that this due to the symmetry of the kernel functions. This last section states, without proof, the results for the local linear estimator for the case that neither bandwidth process dominates the other.

Theorem 17 (Distribution of $\hat{\theta}$)**.** *Assume the standard regularity conditions hold. Let* $r_n := \sqrt{nh_Xh_Y}$ *and* $H := diag(1, h_X, 1, h_Y)$*. Then* $\hat{\theta}$ *is asymptotically normal in accordance with*

$$r_n H(\hat{\theta} - \theta - \zeta) \xrightarrow{D} N(0, \Sigma_\theta)$$

where

$$\zeta := \begin{pmatrix} \frac{1}{2}h_Y^2\tilde{\mu}_2\frac{\partial^2 f_{Y|X}(y|x)}{\partial y^2} + \frac{1}{2}h_X^2\mu_2\frac{\partial^2 f_{Y|X}(y|x)}{\partial x^2} \\ \frac{1}{2}h_Y^2\tilde{\mu}_2\frac{\partial^3 f_{Y|X}(y|x)}{\partial x\partial y^2} + \frac{1}{6}h_X^2\frac{\mu_4}{\mu_2}\frac{\partial^3}{\partial x^3}f_{Y|X}(y|x) \\ \frac{1}{2}h_X^2\mu_2\frac{\partial^2 f_{X|Y}(x|y)}{\partial x^2} + \frac{1}{2}h_Y^2\tilde{\mu}_2\frac{\partial^2 f_{X|Y}(x|y)}{\partial y^2} \\ \frac{1}{2}h_X^2\mu_2\frac{\partial^3 f_{X|Y}(x|y)}{\partial y\partial x^2} + \frac{1}{6}h_Y^2\frac{\tilde{\mu}_4}{\tilde{\mu}_2}\frac{\partial^3}{\partial y^3}f_{X|Y}(x|y) \end{pmatrix}$$

and

$$\Sigma_\theta = \begin{pmatrix} \frac{f_{Y|X}(y|x)\nu_0\tilde{\nu}_0}{f_X(x)} & 0 & \frac{f_{X,Y}(x,y)\nu_0\tilde{\nu}_0}{f_X(x)f_Y(y)} & 0 \\ 0 & \frac{f_{Y|X}(y|x)\tilde{\nu}_0\nu_2}{f_X(x)\mu_2^2} & 0 & 0 \\ \frac{f_{X,Y}(x,y)\nu_0\tilde{\nu}_0}{f_X(x)\,f_Y(y)} & 0 & \frac{f_{X|Y}(x|y)\nu_0\tilde{\nu}_0}{f_Y(y)} & 0 \\ 0 & 0 & 0 & \frac{f_{X|Y}(x|y)\nu_0\tilde{\nu}_2}{f_Y(y)\tilde{\mu}_2^2} \end{pmatrix}.$$

By definition, $\hat{\theta}_1$ and $\hat{\theta}_3$ hold the density estimators, and $\hat{\theta}_2$ and $\hat{\theta}_4$ the estimators of their derivatives. The density estimators have a higher bias if the linear fit is applied. As mentioned earlier, the bias arises from two sources: The bias in Y is due to the approximation of $f_{Y|X}(y|x)$ through the expectation of the kernel K: $E(K_{h_Y}(Y - y)|X = x)$, which differs from the conditional density $f_{Y|X}(y|x)$

by an $O(h_Y^2)$-term. The bias in X is a result of the polynomial approximation. As we approximate through second order polynomials, the remainder term is of order $o(h_X^2)$. By assumption, no bandwidth process dominates the other. Hence, the bias in X is dominated by the bias in Y asymptotically. For the linear fit, the bias in X is of order $O(h_X^2)$ and, no longer dominated by the bias in Y, appears in the the expression for ζ.

6.4 Conclusion

This chapter has demonstrated how a functional of differential moments can be estimated either through its local counterpart or via a local polynomial kernel estimator.

An interesting aspect is that the two approaches are representative for two different modelling strategies. The local sample moment approach uses the local analogues of differential moments in order to estimate functionals of densities. The density estimation approach uses kernel estimators of densities and their derivatives in order to estimate differential moments.

6.A Proofs of section 6.3

6.A.1 Proof of Theorem 15

The idea of the proof is to decompose $(\hat{\beta}, \tilde{\hat{\beta}})'$ into a bias vector $b \in \mathbb{R}^6$, which converges in probability, and a centralised vector $t \in \mathbb{R}^6$ of partial sums, which is asymptotically Gaussian. The two convergence results are proved in separate

lemma. Define

$$
\begin{aligned}
S_{n,j} &:= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{X_i - x}{h_X}\Big)^j W_{h_X}(X_i - x) \in \mathbb{R} \\
S_n &:= (S_{n,j+l})_{0\le j,l\le 2} = \frac{1}{n}(H_X^{-1} X_x' \mathcal{W} X_x H_X^{-1}) \in \mathbb{R}^{3\times 3} \\
\mathcal{M} &:= (m(X_i, y))_{1\le i\le n} \in \mathbb{R}^n
\end{aligned}
$$

and the 'tilde versions'

$$
\begin{aligned}
\tilde{S}_{n,j} &:= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{Y_i - y}{h_Y}\Big)^j K_{h_Y}(Y_i - y) \in \mathbb{R} \\
\tilde{S}_n &:= \frac{1}{n}(H_Y^{-1} Y_y' \mathcal{K} Y_y H_Y^{-1}) \in \mathbb{R}^{3\times 3} \\
\tilde{\mathcal{M}} &:= (\tilde{m}(x, Y_i))_{1\le i\le n} \in \mathbb{R}^n.
\end{aligned}
$$

We have have the following decomposition:

$$
\begin{aligned}
H_{X,Y}\left\{ \begin{pmatrix} \hat{\beta} \\ \hat{\tilde{\beta}} \end{pmatrix} - \begin{pmatrix} \beta \\ \tilde{\beta} \end{pmatrix} \right\} &= \left(\begin{array}{c|c} S_n & 0 \\ \hline 0 & \tilde{S}_n \end{array}\right)^{-1} \left(\begin{array}{c} \frac{1}{n} H_X^{-1} X_x' \mathcal{W} \mathcal{Y} \\ \hline \frac{1}{n} H_Y^{-1} Y_y' \mathcal{K} \mathcal{X} \end{array}\right) \\
&= \left(\begin{array}{c|c} S_n & 0 \\ \hline 0 & \tilde{S}_n \end{array}\right)^{-1} (b + t),
\end{aligned}
$$

where

$$
b := \left(\begin{array}{c} \frac{1}{n} H_X^{-1} X_x' \mathcal{W}\{\mathcal{M} - X_x\beta\} \\ \hline \frac{1}{n} H_Y^{-1} Y_y' \mathcal{K}\{\tilde{\mathcal{M}} - Y_y\tilde{\beta}\} \end{array}\right) \tag{6.16}
$$

and

$$
t := \left(\begin{array}{c} \frac{1}{n} H_X^{-1} X_x' \mathcal{W}\{\mathcal{Y} - \mathcal{M}\} \\ \hline \frac{1}{n} H_Y^{-1} Y_y' \mathcal{K}\{\mathcal{X} - \tilde{\mathcal{M}}\} \end{array}\right). \tag{6.17}
$$

Fan and Gijbels (1996) show that

$$
\left(\begin{array}{c|c} S_n & 0 \\ \hline 0 & \tilde{S}_n \end{array}\right) = \left(\begin{array}{c|c} f_X(x) S & 0 \\ \hline 0 & f_Y(y)\tilde{S} \end{array}\right) \{1 + o_P(1)\}. \tag{6.18}
$$

In Lemma 12 it is shown that b converges to the expression claimed in (6.13). Finally, Lemma 13 shows that

$$r_n t \xrightarrow{D} \mathcal{N}\left(0, f_{X,Y}(x,y) \left(\begin{array}{c|c} \tilde{\nu}_0 S^\star & \Sigma \\ \hline \Sigma & \nu_0 \tilde{S}^\star \end{array}\right)\right). \tag{6.19}$$

This completes the proof of Theorem 15 since Slutsky's lemma then entails that $(\hat{\beta}', \tilde{\hat{\beta}}')'$ is asymptotically Gaussian with the required moments.

Lemma 12 (Fan and Gijbels (1996) - Convergence of b)**.** *For the bias vector b as defined in (6.16) it holds that*

$$(b_k)_{1\leq k\leq 3} = \frac{1}{6} h_X^3 f_X(x) \begin{pmatrix} 4h_X \frac{\partial^4}{\partial x^4} f_{Y|X}(y|x)\mu_4 \\ \frac{\partial^3}{\partial x^3} f_{Y|X}(y|x)\mu_4 \\ 4h_X \frac{\partial^4}{\partial x^4} f_{Y|X}(y|x)\mu_6 \end{pmatrix} \{1 + o_P(1)\}$$

and

$$(b_k)_{4\leq k\leq 6} = \frac{1}{6} h_Y^3 f_Y(y) \begin{pmatrix} 4h_Y \frac{\partial^4}{\partial y^4} f_{X|Y}(x|y)\tilde{\mu}_4 \\ \frac{\partial^3}{\partial y^3} f_{X|Y}(x|y)\tilde{\mu}_4 \\ 4h_Y \frac{\partial^4}{\partial y^4} f_{X|Y}(x|y)\tilde{\mu}_6 \end{pmatrix} \{1 + o_P(1)\}.$$

Proof. We provide the full proof since Fan and Gijbels (1996) only give a sketch. By definition, $b = \left(\begin{array}{c} \frac{1}{n} H_X^{-1} X_x' \mathcal{W}\{\mathcal{M} - X_x\beta\} \\ \hline \frac{1}{n} H_Y^{-1} Y_y' \mathcal{K}\{\tilde{\mathcal{M}} - Y_y\tilde{\beta}\} \end{array}\right)$. By a Taylor expansion of $\mathcal{M}$ about (x, y)

$$\begin{aligned} (b)_{1\leq j\leq 3} &= \frac{1}{n} H_X^{-1} X_x' \mathcal{W}\{\mathcal{M} - X_x\beta\} \\ &= \frac{1}{n} H_X^{-1} X_x' \mathcal{W} \left\{ \left(m_3(x,y)(X_i - x)^3 + m_4(x,y)(X_i - x)^4 + o_P\{(X_i - x)^4\} \right)_{1\leq i\leq n} \right\}. \end{aligned} \tag{6.20}$$

The first term can be written as:

$$\frac{1}{n}H_X^{-1}X_x'\mathcal{W}\Big(m_3(x,y)(X_i-x)^3\Big)_{1\le i\le n} = m_3(x,y)h_X^3(S_{n,j})_{3\le j\le 5}$$
$$= m_3(x,y)h_X^3 f_X(x)(\mu_j)_{3\le j\le 5}\{1+o_P(1)\}. \tag{6.21}$$

By the symmetry of the kernel W, $\mu_j = 0$ whenever j is odd, which is why the expansion (6.20) includes the term involving $m_4(x,y)$. Applying yet another Taylor expansion, we have

$$m_j(x,y) := \frac{1}{j!}\frac{\partial^j}{\partial x^j}E\{K_{h_y}(Y-y)|X=x\}$$
$$= \frac{1}{j!}\frac{\partial^j}{\partial x^j}\Big(f_{Y|X}(y|x)+o_P(1)\Big). \tag{6.22}$$

Substituting (6.22), with $j=3$, into (6.21) yields:

$$\frac{1}{n}H_X^{-1}X_x'\mathcal{W}\Big(m_3(x,y)(X_i-x)^3\Big)_{1\le i\le n} = \frac{1}{6}h_X^3 f_X(x)\frac{\partial^3}{\partial x^3}f_{Y|X}(y|x)\begin{pmatrix}0\\ \mu_4\\ 0\end{pmatrix}\{1+o_P(1)\}.$$

Similarly,

$$\frac{1}{n}H_X^{-1}X_x'\mathcal{W}\Big(m_4(x,y)(X_i-x)^4\Big)_{1\le i\le n} = \frac{1}{4!}h_X^4 f_X(x)\frac{\partial^4}{\partial x^4}f_{Y|X}(y|x)\begin{pmatrix}\mu_4\\ 0\\ \mu_6\end{pmatrix}\{1+o_P(1)\}.$$

The remainder term in (6.20) is of order h_X^4:

$$\frac{1}{n}H_X^{-1}X_x'\mathcal{W}(o_P\{(X_i-x)^4\})_{1\le i\le n} = \frac{1}{n}\Big(\sum_{i=1}^n(\frac{X_i-x}{h_x})^j W_{h_X}(X_i-x)o_P\{(X_i-x)^4\}\Big)_{0\le j\le 2}$$
$$= (S_{n,j})_{0\le j\le 2}\; o_p(h_X^4)$$
$$= o_p(h_X^4), \tag{6.23}$$

where (6.23) follows from the boundedness of the kernel W. We have for the first three entries of b:

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \frac{1}{6} h_X^3 \frac{\partial^3}{\partial x^3} f_{X,Y}(x,y) \begin{pmatrix} 4h_X\mu_4 \\ \mu_4 \\ 4h_X\mu_6 \end{pmatrix} \{1 + o_p(1)\}.$$

Similar arguments hold for the entries $(b_k)_{4\leq k\leq 6}$, which completes the proof of Lemma 12. □

Lemma 13 (Asymptotic normality of t)**.** *The centralised vector t, as defined in (6.17), is asymptotically Gaussian. In particular it holds that:*

$$t \xrightarrow{D} N(0, \Sigma_t),$$

where

$$\Sigma_t = \frac{1}{nh_Xh_Y} f_{X,Y}(x,y) \left(\begin{array}{c|c} \tilde{\nu}_0 S^\star & \Sigma \\ \hline \Sigma & \nu_0 \tilde{S}^\star \end{array}\right).$$

Proof. By the Cramer-Wold, it is sufficient to show that the linear combination $\lambda' t \in \mathbb{R}$ is asymptotically Gaussian distributed for an arbitrary $\lambda \in \mathbb{R}^6$.

$$\begin{aligned} \lambda' t &= \lambda' \left(\begin{array}{c} \frac{1}{n} H_X^{-1} X_x' \mathcal{W}\{\mathcal{Y} - \mathcal{M}\} \\ \hline \frac{1}{n} H_Y^{-1} Y_y' \mathcal{K}\{\mathcal{X} - \tilde{\mathcal{M}}\} \end{array}\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{j=0}^{2} \lambda_j \left(\frac{X_i - x}{h_X}\right)^j W_{h_X}(X_i - x)(K_{h_Y}(Y_i - y) - \mathcal{M}_i) \right. \\ &\quad \left. + \sum_{k=0}^{2} \lambda_{k+3} \left(\frac{Y_i - y}{h_Y}\right)^k K_{h_Y}(Y_i - y)(W_{h_X}(X_i - x) - \tilde{\mathcal{M}}_i) \right\}. \end{aligned}$$

Asymptotic normality of t follows from the Central Limit Theorem for i.i.d. sequences. The same result can be proved for the dependent case when certain mixing conditions are met and the interested reader is referred to Fan and Gijbels (1996) and Fan et al. (1996b). It remains to derive the mean and covariance matrix of the asymptotic distribution.

We first note that t has zero mean: Conditioning on $(X_i)_{1\le i\le n}$ and applying the law of total expectation yields:

$$\begin{aligned} E\{(t_k)_{1\le k\le 3}\} &= E\{E(n^{-1}H_X^{-1}X_x'\mathcal{W}\{\mathcal{Y}-\mathcal{M}\}|(X_i)_{1\le i\le n})\} \\ &= E\{E(n^{-1}H_X^{-1}X_x'\mathcal{W}\{\mathcal{M}-\mathcal{M}\}|(X_i)_{1\le i\le n})\} \\ &= 0\in\mathbb{R}^3. \end{aligned} \tag{6.24}$$

By conditioning on $(Y_i)_{1\le i\le n}$, we find that $E\{(t_k)_{4\le k\le 6}\}=0$.

We derive the asymptotic covariance of t. Define

$$C(u) := \sum_{j=0}^{2}\lambda_j u^j W(u) \qquad \tilde{C}(u) := \sum_{j=0}^{2}\lambda_{j+3}u^j K(u)$$

$$C_{h_X}(u) := \frac{1}{h_X}C(\frac{u}{h_X}) \qquad \tilde{C}_{h_Y}(u) := \frac{1}{h_Y}\tilde{C}(\frac{u}{h_Y}).$$

We can write $\lambda' t = \frac{1}{n}\sum_{i=1}^n Z_i$, where

$$Z_i := C_{h_X}(X_i-x)(K_{h_Y}(Y_i-y)-\mathcal{M}_i)+\tilde{C}_{h_Y}(Y_i-y)(W_{h_X}(X_i-x)-\tilde{\mathcal{M}}_i).$$

By independence and identical distribution,

$$\operatorname{var}(\lambda' t) = \frac{1}{n}\operatorname{var}(Z_1) = \frac{1}{n}E(Z_1^2), \tag{6.25}$$

since $E(Z_1)=0$ by (6.24).

We decompose the variance of Z_1. The expression we obtain for $\operatorname{var}\{C_{h_X}(X_1-x)(K_{h_Y}(Y_1-y)-\mathcal{M}_1)\}$ differs from Fan and Gijbels (1996), Fan et al. (1996b) and Fan and Yao (2003) by a kernel moment term, so that we derive it explicitly.

$$\begin{aligned} \operatorname{var}(Z_1) &= E(Z_1^2)-\{E(Z_1)\}^2 = E(Z_1^2) \\ &= E\Big\{C_{h_X}(X_1-x)\{K_{h_Y}(Y_1-y)-\mathcal{M}_1\}\Big\}^2 \\ &\quad + 2E\Big\{C_{h_X}(X_1-x)\{K_{h_Y}(Y_1-y)-\mathcal{M}_1\}\tilde{C}_{h_Y}(Y_1-y)\{W_{h_X}(X_1-x)-\tilde{\mathcal{M}}_1\}\Big\} \\ &\quad + E\Big\{\tilde{C}_{h_Y}(Y_1-y)\{W_{h_X}(X_1-x)-\tilde{\mathcal{M}}_1\}\Big\}^2 \\ &= A+2B+\tilde{A}, \end{aligned} \tag{6.26}$$

where

$$A := E\Big\{C_{h_X}(X_1 - x)\{K_{h_Y}(Y_1 - y) - \mathcal{M}_1\}\Big\}^2 \tag{6.27}$$
$$B := E\Big\{C_{h_X}(X_1 - x)\{K_{h_Y}(Y_1 - y) - \mathcal{M}_1\}\tilde{C}_{h_Y}(Y_1 - y)\{W_{h_X}(X_1 - x) - \tilde{\mathcal{M}}_1\}\Big\}$$
$$\tilde{A} := E\Big\{\tilde{C}_{h_Y}(Y_1 - y)\{W_{h_X}(X_1 - x) - \tilde{\mathcal{M}}_1\}\Big\}^2.$$

We derive A and B explicitly. By the law of total expectation:

$$\begin{aligned} A &= E\Big\{E\Big(C_{h_X}^2(X_1 - x)\{K_{h_Y}(Y_1 - y) - \mathcal{M}_1\}^2 \,\Big|\, (X_i)_{1\le i\le n}\Big)\Big\} \\ &= E\Big\{C_{h_X}^2(X_1 - x)E\Big(\{K_{h_Y}(Y_1 - y) - \mathcal{M}_1\}^2 \,\Big|\, (X_i)_{1\le i\le n}\Big)\Big\} \end{aligned}$$

With K and W bounded we may assume without loss of generality that $(X_1 - x) = O_P(h_X)$ and $(Y_1 - y) = O_P(h_Y)$.

$$\begin{aligned} E\Big(\{K_{h_Y}(Y_1 - y) - \mathcal{M}_1\}^2 \,\Big|\, (X_i)_{1\le i\le n}\Big) &= \operatorname{var}\Big(K_{h_Y}(Y_1 - y) \,\Big|\, X_1\Big) \\ &= \int K_{h_Y}^2(u - y) f_{Y|X}(u|X_1)du \\ &\quad - \Big(\int K_{h_Y}(u - y) f_{Y|X}(u|X_1)du\Big)^2 \\ &= \frac{1}{h_Y}\int K^2(u)\{f_{Y|X}(u|X_1) + o_P(1)\}du \\ &\quad - \{f_{Y|X}(u|X_1) + o_P(1)\}^2 \\ &= \frac{1}{h_Y}\tilde{\nu}_0 f_{Y|X}(y|X_1) + o_p(h_Y). \end{aligned}$$

Substituting this expression into (6.27) yields:

$$\begin{aligned} A &= E\Big\{C_{h_X}^2(X_1 - x)(\frac{1}{h_Y}\tilde{\nu}_0 f_{Y|X}(y|X_1) + o_p(h_Y))\Big\} \\ &= \frac{1}{h_Y}\tilde{\nu}_0 \int C_{h_X}^2(u - x) f_{Y|X}(y|u) f_X(u)du\} + o_p(h_Y) \\ &= \frac{1}{h_X h_Y} f_{Y,X}(y|x) f_X(x)\tilde{\nu}_0 \int C^2(z)dz\{1 + o_p(1)\}. \end{aligned}$$

Finally, noting that $\int C^2(z)dz = (\lambda_j)'_{0\leq j\leq 2} S^\star(\lambda_j)_{0\leq j\leq 2}$, we arrive at:

$$A = \frac{1}{h_X h_Y} f_{X,Y}(x,y)\tilde{\nu}_0(\lambda_j)'_{0\leq j\leq 2} S^\star(\lambda_j)_{0\leq j\leq 2}\{1+o_p(1)\}. \tag{6.28}$$

By similar arguments we have

$$\tilde{A} = \frac{1}{h_X h_Y} f_{X,Y}(x,y)\nu_0(\lambda_j)'_{3\leq j\leq 5} \tilde{S}^\star(\lambda_j)_{3\leq j\leq 5}\{1+o_p(1)\}. \tag{6.29}$$

We apply the law of total expectation to B as well:

$$\begin{aligned} B &:= E\Big\{C_{h_X}(X_1-x)\{K_{h_Y}(Y_1-y)-\mathcal{M}_1\}\tilde{C}_{h_Y}(Y_1-y)\{W_{h_X}(X_1-x)-\tilde{\mathcal{M}}_1\}\Big\} \\ &= E_X\Big\{E\Big(C_{h_X}(X_1-x)\{K_{h_Y}(Y_1-y)-\mathcal{M}_1\}\tilde{C}_{h_Y}(Y_1-y)\{W_{h_X}(X_1-x)-\tilde{\mathcal{M}}_1\}\,\Big|\,X_1\Big)\Big\} \\ &= E_X\Big\{C_{h_X}(X_1-x)\{W_{h_X}(X_1-x)-\tilde{\mathcal{M}}_1\}E\Big(\{K_{h_Y}(Y_1-y)-\mathcal{M}_1\}\tilde{C}_{h_Y}(Y_1-y)\,\Big|\,X_1\Big)\Big\}. \end{aligned} \tag{6.30}$$

Upon multiplying out (6.30), we find that

$$E_X\Big\{C_{h_X}(X_1-x)\{W_{h_X}(X_1-x)\}E\Big(\{K_{h_Y}(Y_1-y)\}\tilde{C}_{h_Y}(Y_1-y)\,\Big|\,X_1\Big)\Big\} = O\Big(\frac{1}{h_X h_Y}\Big) \tag{6.31}$$

is the dominating term. This holds true since we assume $\frac{h_X}{h_Y}$ converges to some constant and other terms are of order $O(h_X^{-1}), O(h_Y^{-1})$ and $O(1)$, hence converge at a faster rate. Consider first

$$\begin{aligned} E\Big(\{K_{h_Y}(Y_1-y)\}\tilde{C}_{h_Y}(Y_1-y)\,\Big|\,X_1\Big) &= \int \tilde{C}_{h_Y}(u-y)K_{h_Y}(u-y)f_{Y|X}(u|X_1)du \\ &= \frac{1}{h_Y} f_{Y|X}(y|x)\int \tilde{C}(z)K(z)dz\{1+o_P(1)\}. \end{aligned}$$

The integral $\int \tilde{C}(z)K(z)dz$ has the simple form:

$$\begin{aligned} \int \tilde{C}(z)K(z)dz &= \int \sum_{j=0}^{2} \lambda_{j+3} z^j K(z)K(z)dz \\ &= \sum_{j=0}^{2} \lambda_{j+3}\int z^j K^2(z)dz \\ &= \sum_{j=0}^{2} \lambda_{j+3}\tilde{\nu}_j. \end{aligned} \tag{6.32}$$

Substituting (6.32) into (6.31) and using another change of variable argument and a Taylor expansion we obtain:

$$B = \frac{1}{h_X h_Y} f_{X,Y}(x,y) \left(\sum_{j=0}^{2} \lambda_{j+3} \tilde{\nu}_j \right) \left(\sum_{j=0}^{2} \lambda_j \nu_j \right) \{1 + o_P(1)\}. \tag{6.33}$$

Substituting (6.28), (6.29) and (6.33) into (6.26) we obtain for the variance of Z_1 :

$$\operatorname{var}(Z_1) = \frac{1}{h_X h_Y} f_{X,Y}(x,y) \lambda' \left(\begin{array}{c|c} \tilde{\nu}_0 S^\star & \Sigma \\ \hline \Sigma' & \nu_0 \tilde{S}^\star \end{array} \right) \lambda \{1 + o_P(1)\},$$

where Σ was defined as $\Sigma := (\tilde{\nu}_j \nu_k)_{0 \le j,k \le 2}$. Since $\operatorname{var}(\lambda' t) = \frac{1}{n} \operatorname{var}(Z_1)$ the variance of t is given by:

$$\operatorname{var}(t) = \frac{1}{n h_X h_Y} f_{X,Y}(x,y) \left(\begin{array}{c|c} \tilde{\nu}_0 S^\star & \Sigma \\ \hline \Sigma' & \nu_0 \tilde{S}^\star \end{array} \right) \{1 + o_P(1)\}.$$

This completes the proof of Lemma 13. □

6.A.2 Proof of Corollary 7

Let ζ denote the bias term, which is the only part which deserves explanation. We need to show that ζ is indeed equal to the expression claimed in (6.14). The bias can be decomposed into two components:

$$\zeta = \left(\begin{array}{c|c} E' & 0 \\ \hline 0 & E' \end{array} \right) \Sigma_1^{-1} b + \left(\begin{array}{c|c} E' & 0 \\ \hline 0 & E' \end{array} \right) \left(\begin{array}{c} \beta \\ \hline \tilde{\beta} \end{array} \right) - \theta.$$

The first part corresponds to the relevant entries of $\Sigma_1^{-1} b$ in the preceeding theorem. They are of order $o(h_X^3)$. The second part represents the bias introduced by the approximation of $f_{Y|X}$ through $E\{K_{h_y}(Y - y)|X = x\}$ and similarly for $f_{Y|X}$. Considering the Taylor expansions

$$m_j(x,y) = \frac{\partial^j f_{Y|X}(y|x)}{\partial x^j} + \frac{1}{2} h_Y^2 \tilde{\mu}_2 \frac{\partial^{j+2} f_{Y|X} y|x}{\partial x^j \partial y^2} + o(h_Y^2)$$

$$\tilde{m}_j(x,y) = \frac{\partial^j f_{X|Y}(x|y)}{\partial y^j} + \frac{1}{2} h_X^2 \mu_2 \frac{\partial^{j+2} f_{X|Y} x|y}{\partial y^j \partial x^2} + o(h_X^2)$$

and comparing the order of the summands yields the expression for ζ.

6.A.3 Proof of Theorem 16

The key of the proof is an application of the so called *Delta method*, which allows to compute the asymptotic distribution of a derived statistic from one with known asymptotic distribution. Some care has to be taken, to accomodate the fact that the components of the estimators have different convergence rates.

Consider the Taylor expansion of $g(\hat{\theta}) = \left(\frac{\hat{\theta}_2}{\hat{\theta}_1}, \frac{\hat{\theta}_4}{\hat{\theta}_3}\right)$ about ξ:

$$g(\hat{\theta}) = g(\xi) + J_g(u)\Big|_{u=\xi} (\hat{\theta} - \xi) + o_P\Big(\big\|\hat{\theta} - \xi\big\|\Big) \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where $J_g(u)\Big|_{u=\xi}$ denotes the Jacobian matrix of g evaluated at ξ, see (6.12). Multiply through by $r_n h_X$ to obtain:

$$r_n h_X(g(\hat{\theta}) - g(\xi)) = r_n h_X \, J_g(u)\Big|_{u=\xi} (\hat{\theta} - \xi) + o_P\Big(r_n h_X \big\|\hat{\theta} - \xi\big\|\Big) \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

We first proof that the remainder term converges to zero. The sequence $r_n h_X(\hat{\theta} - \xi)$ converges weakly and is thus bounded in probability. If ξ was fixed, this would be sufficient to show that the remainder term converges to zero in probability. ξ is, however, a function of the bandwidth processes h_X and h_Y, which again are functions of n. Van der Vaart (2000, Theorem 3.8, page 33) shows that the same result obtains even if ξ varies with n provided g is continuously differentiable in a neighbourhood of ξ as assumed.

The term $r_n h_X \, J_g(u)\Big|_{u=\xi} (\hat{\theta} - \xi)$ holds terms of different convergence rates. The terms $r_n(\hat{\theta}_1 - \xi_1)$ and $r_n(\hat{\theta}_3 - \xi_3)$ converge in distribution to a Gaussian by Corollary 7. Since $h_X \to 0$, Slutsky's lemma entails that $r_n h_X(\hat{\theta}_1 - \xi_1)$ and $r_n h_X(\hat{\theta}_3 - \xi_3)$ converge to zero in probability. We thus have

$$r_n h_X(g(\hat{\theta}) - g(\xi)) = r_n h_X \begin{pmatrix} \frac{(\hat{\theta}_2 - \xi_2)}{\xi_1} \\ \frac{(\hat{\theta}_2 - \xi_4)}{\xi_3} \end{pmatrix} + o_P(1) \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

By Corollary 7 the first summand on the right hand side is normal asymptotically:

$$\begin{pmatrix} r_n h_X(\hat{\theta}_2 - \xi_2) \\ r_n h_Y(\hat{\theta}_4 - \xi_4) \end{pmatrix} \xrightarrow{D} \mathcal{N}(0, \Sigma_2),$$

where

$$\Sigma_2 := \begin{pmatrix} \frac{f_{Y|X}(x,y)}{f_X(x)} \frac{\tilde{\nu}_0 \nu_2}{\mu_2^2} & 0 \\ 0 & \frac{f_{X|Y}(x,y)}{f_Y(y)} \frac{\nu_0 \tilde{\nu}_2}{\tilde{\mu}_2^2} \end{pmatrix}.$$

An application of Slutsky's lemma completes the proof.

6.B Matrices

This appendix lists a few of the matrices explicitly.

6.B.1 The quadratic case

6.B.1.1 Matrices appearing in Theorem 15

The kernel moments and associated matrices were defined on page 135. Odd kernel moments are zero by the symmetry of K and W. This property is inherited by K^2 and W^2. The covariance matrix Σ_β is given by

$$\begin{aligned} \Sigma_\beta &= f_{X,Y}(x,y) \Sigma_1^{-1} \left(\begin{array}{c|c} \tilde{\nu}_0 S^\star & \Sigma \\ \hline \Sigma' & \nu_0 \tilde{S}^\star \end{array}\right) \Sigma_1^{-1} \\ &= f_{X,Y}(x,y) \left(\begin{array}{c|c} \tilde{\nu}_0 f_X^{-2}(x) S^{-1} S^\star S^{-1} & f_X^{-1}(x) f_Y^{-1}(y) S^{-1} \Sigma \tilde{S}^{-1} \\ \hline f_X^{-1}(x) f_Y^{-1}(y) \tilde{S}^{-1} \Sigma' S^{-1} & \nu_0 f_Y^{-2}(y) \tilde{S}^{-1} \tilde{S}^\star \tilde{S}^{-1} \end{array}\right), \end{aligned}$$

where

$$S^{-1}\tilde{S}^{\star}S^{-1} = \begin{pmatrix} \frac{\mu_4^2\nu_0-2\mu_4\mu_2\nu_2+\mu_2^2\nu_4}{(\mu_2^2-\mu_4)^2} & 0 & \frac{-\mu_2\mu_4\nu_0+\mu_2^2\nu_2+\mu_4\nu_2-\mu_2\nu_4}{(\mu_2^2-\mu_4)^2} \\ 0 & \frac{\nu_2}{\mu_2^2} & 0 \\ \frac{-\mu_2\mu_4\nu_0+\mu_2^2\nu_2+\mu_4\nu_2-\mu_2\nu_4}{(\mu_2^2-\mu_4)^2} & 0 & \frac{\mu_2^2\nu_0-2\mu_2\nu_2+\nu_4}{(\mu_2^2-\mu_4)^2} \end{pmatrix}$$

$$S^{-1}\Sigma\tilde{S}^{-1} = \begin{pmatrix} \frac{(\tilde{\mu}_2\tilde{\nu}_2-\tilde{\mu}_4\tilde{\nu}_0)(\mu_2\nu_2-\mu_4\nu_0)}{(\mu_2^2-\mu_4)(\tilde{\mu}_2^2-\tilde{\mu}_4)} & 0 & \frac{(-\tilde{\mu}_4\tilde{\nu}_0+\tilde{\mu}_2\tilde{\nu}_2)(\mu_2\nu_0-\nu_2)}{(\mu_2^2-\mu_4)(\tilde{\mu}_2^2-\tilde{\mu}_4)} \\ 0 & 0 & 0 \\ -\frac{(\mu_2\nu_0-\nu_2)(\tilde{\mu}_4\tilde{\nu}_0-\tilde{\mu}_2\tilde{\nu}_2)}{(\mu_2^2-\mu_4)(\tilde{\mu}_2^2-\tilde{\mu}_4)} & 0 & \frac{(\mu_2\nu_0-\nu_2)(\tilde{\mu}_2\tilde{\nu}_0-\tilde{\nu}_2)}{(\mu_2^2-\mu_4)(\tilde{\mu}_2^2-\tilde{\mu}_4)} \end{pmatrix}$$

$$\tilde{S}^{-1}\Sigma' S^{-1} = \begin{pmatrix} \frac{(\mu_2\tilde{\nu}_2-\mu_4\tilde{\nu}_0)(\tilde{\mu}_2\nu_2-\tilde{\mu}_4\nu_0)}{(\tilde{\mu}_2^2-\tilde{\mu}_4)(\mu_2^2-\mu_4)} & 0 & -\frac{(\mu_2\tilde{\nu}_0-\tilde{\nu_2})(\tilde{\mu}_4\nu_0-\tilde{\mu}_2\nu_2)}{(\tilde{\mu}_2^2-\tilde{\mu}_4)(\mu_2^2-\mu_4)} \\ 0 & 0 & 0 \\ \frac{(-\mu_4\tilde{\nu}_0+\mu_2\tilde{\nu}_2)(\tilde{\mu}_2\nu_0-\nu_2)}{(\tilde{\mu}_2^2-\tilde{\mu}_4)(\mu_2^2-\mu_4)} & 0 & \frac{(\tilde{\mu}_2\nu_0-\nu_2)(\mu_2\tilde{\nu}_0-\tilde{\nu}_2)}{(\tilde{\mu}_2^2-\tilde{\mu}_4)(\mu_2^2-\mu_4)} \end{pmatrix}$$

$$\tilde{S}^{-1}\tilde{S}^{\star}\tilde{S}^{-1} = \begin{pmatrix} \frac{\tilde{\mu}_4^2\tilde{\nu}_0-2\tilde{\mu}_4\tilde{\mu}_2\tilde{\nu}_2+\tilde{\mu}_2^2\tilde{\nu}_4}{(\tilde{\mu}_2^2-\tilde{\mu}_4)^2} & 0 & \frac{-\tilde{\mu}_2\tilde{\mu}_4\tilde{\nu}_0+\tilde{\mu}_2^2\tilde{\nu}_2+\tilde{\mu}_4\tilde{\nu}_2-\tilde{\mu}_2\tilde{\nu}_4}{(\tilde{\mu}_2^2-\tilde{\mu}_4)^2} \\ 0 & \frac{\tilde{\nu}_2}{\tilde{\mu}_2^2} & 0 \\ \frac{-\tilde{\mu}_2\tilde{\mu}_4\tilde{\nu}_0+\tilde{\mu}_2^2\tilde{\nu}_2+\tilde{\mu}_4\tilde{\nu}_2-\tilde{\mu}_2\tilde{\nu}_4}{(\tilde{\mu}_2^2-\tilde{\mu}_4)^2} & 0 & \frac{\tilde{\mu}_2^2\tilde{\nu}_0-2\tilde{\mu}_2\tilde{\nu}_2+\tilde{\nu}_4}{(\tilde{\mu}_2^2-\tilde{\mu}_4)^2} \end{pmatrix}.$$

The above matrices were computed using

$$S := (\mu_{j+l})_{0\le j,l\le 2} = \begin{pmatrix} 1 & 0 & \mu_2 \\ 0 & \mu_2 & 0 \\ \mu_2 & 0 & \mu_4 \end{pmatrix} \qquad S^{-1} := \begin{pmatrix} \frac{-\mu_4}{\mu_2^2-\mu_4} & 0 & \frac{\mu_2}{\mu_2^2-\mu_4} \\ 0 & \frac{1}{\mu_2} & 0 \\ \frac{-\mu_2}{\mu_2^2-\mu_4} & 0 & \frac{-1}{\mu_2^2-\mu_4} \end{pmatrix}$$

$$\tilde{S} := (\tilde{\mu}_{j+l})_{0\le j,l\le 2} = \begin{pmatrix} 1 & 0 & \tilde{\mu}_2 \\ 0 & \tilde{\mu}_2 & 0 \\ \tilde{\mu}_2 & 0 & \tilde{\mu}_4 \end{pmatrix} \qquad \tilde{S}^{-1} := \begin{pmatrix} \frac{-\tilde{\mu}_4}{\tilde{\mu}_2^2-\tilde{\mu}_4} & 0 & \frac{\tilde{\mu}_2}{\tilde{\mu}_2^2-\tilde{\mu}_4} \\ 0 & \frac{1}{\tilde{\mu}_2} & 0 \\ \frac{-\tilde{\mu}_2}{\tilde{\mu}_2^2-\tilde{\mu}_4} & 0 & \frac{-1}{\tilde{\mu}_2^2-\tilde{\mu}_4} \end{pmatrix}$$

$$S^\star = (\nu_{j+l})_{0\le j,l\le 2} = \begin{pmatrix} \nu_0 & 0 & \nu_2 \\ 0 & \nu_2 & 0 \\ \nu_2 & 0 & \nu_4 \end{pmatrix} \qquad \tilde{S}^\star := (\tilde{\nu}_{j+l})_{0\le j,l\le 2} = \begin{pmatrix} \tilde{\nu}_0 & 0 & \tilde{\nu}_2 \\ 0 & \tilde{\nu}_2 & 0 \\ \tilde{\nu}_2 & 0 & \tilde{\nu}_4 \end{pmatrix}$$

$$\Sigma := (\tilde{\nu}_j \nu_l)_{0\le j,l\le 2} = \begin{pmatrix} \tilde{\nu}_0\nu_0 & 0 & \tilde{\nu}_2\nu_0 \\ 0 & 0 & 0 \\ \tilde{\nu}_0\nu_2 & 0 & \tilde{\nu}_2\nu_2 \end{pmatrix} \qquad \Sigma_1 = \left(\begin{array}{c|c} f_X(x)S & 0 \\ \hline 0 & f_Y(y)\tilde{S} \end{array}\right).$$

6.B.1.2 Matrix Σ_θ in Corollary 7

The covariance matrix Σ_θ is a submatrix of Σ_β:

$$\Sigma_\theta = \left(\begin{array}{c|c} E' & 0 \\ \hline 0 & E' \end{array}\right) \Sigma_\beta \left(\begin{array}{c|c} E & 0 \\ \hline 0 & E \end{array}\right)$$

$$= \begin{pmatrix} \frac{f_{Y|X}(x,y)}{f_X(x)}\tilde{\nu}_0 \frac{\mu_4^2\nu_0 - 2\mu_4\mu_2\nu_2 + \mu_2^2\nu_4}{(\mu_2^2-\mu_4)^2} & 0 & \frac{f_{Y|X}(x,y)}{f_Y(y)} \frac{(\mu_2\nu_2-\mu_4\nu_0)(\tilde{\mu}_2\tilde{\nu}_2-\tilde{\mu}_4\tilde{\nu}_0)}{(\mu_2^2-\mu_4)(\tilde{\mu}_2^2-\tilde{\mu}_4)} & 0 \\ 0 & \frac{f_{Y|X}(x,y)}{f_X(x)}\frac{\tilde{\nu}_0\nu_2}{\mu_2^2} & 0 & 0 \\ \frac{f_{X|Y}(x,y)}{f_X(x)} \frac{(\tilde{\mu}_2\nu_2-\tilde{\mu}_4\nu_0)(\mu_2\tilde{\nu}_2-\mu_4\tilde{\nu}_0)}{(\tilde{\mu}_2^2-\tilde{\mu}_4)(\mu_2^2-\mu_4)} & 0 & \frac{f_{X|Y}(x,y)}{f_Y(y)}\tilde{\nu}_0 \frac{\tilde{\mu}_4^2\tilde{\nu}_0 - 2\tilde{\mu}_4\tilde{\mu}_2\tilde{\nu}_2 + \tilde{\mu}_2^2\tilde{\nu}_4}{(\tilde{\mu}_2^2-\tilde{\mu}_4)^2} & 0 \\ 0 & 0 & 0 & \frac{f_{X|Y}(x,y)}{f_Y(y)}\frac{\nu_0\tilde{\nu}_2}{\tilde{\mu}_2^2} \end{pmatrix}$$

6.B.2 The linear case

The relevant matrices which appear in Theorem 17 are rather simple: The covariance matrix Σ_θ is given by

$$\Sigma_\theta = f_{X,Y}(x,y)\Sigma_1^{-1}\left(\begin{array}{c|c} \tilde{\nu}_0 S^\star & \Sigma \\ \hline \Sigma' & \nu_0\tilde{S}^\star \end{array}\right)\Sigma_1^{-1}$$

$$= \begin{pmatrix} \frac{f_{Y|X}(y|x)\nu_0\tilde{\nu}_0}{f_X(x)} & 0 & \frac{f_{X,Y}(x,y)\nu_0\tilde{\nu}_0}{f_X(x)f_Y(y)} & 0 \\ 0 & \frac{f_{Y|X}(y|x)\tilde{\nu}_0\nu_2}{f_X(x)\mu_2^2} & 0 & 0 \\ \frac{f_{X,Y}(x,y)\nu_0\tilde{\nu}_0}{f_X(x)\,f_Y(y)} & 0 & \frac{f_{X|Y}(x|y)\nu_0\tilde{\nu}_0}{f_Y(y)} & 0 \\ 0 & 0 & 0 & \frac{f_{X|Y}(x|y)\nu_0\tilde{\nu}_2}{f_Y(y)\tilde{\mu}_2^2} \end{pmatrix},$$

since

$$S = \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} \qquad S^{-1} := \begin{pmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{pmatrix}$$

$$\tilde{S} = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{\mu}_2 \end{pmatrix} \qquad \tilde{S}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{\mu}_2^{-1} \end{pmatrix}$$

$$S^\star = \begin{pmatrix} \nu_0 & 0 \\ 0 & \nu_2 \end{pmatrix} \qquad \tilde{S}^\star = \begin{pmatrix} \tilde{\nu}_0 & 0 \\ 0 & \tilde{\nu}_2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \tilde{\nu}_0\nu_0 & 0 \\ 0 & 0 \end{pmatrix} \qquad \Sigma_1 = \left(\begin{array}{c|c} f_X(x)S & 0 \\ \hline 0 & f_Y(y)\tilde{S} \end{array}\right).$$

Bibliography

B.C. Arnold and D. Strauss. Bivariate distributions with exponential conditionals. *Journal of the American Statistical Association*, 83:522–527, 1988.

O.E. Barndorff-Nielsen and D.R. Cox. *Asymptotic techniques for use in statistics*. Chapman and Hall, London, UK, 1989.

P.R. Baxandall and H. Liebeck. *Vector calculus*. Oxford University Press, USA, 1986.

P.J. Bickel, E.A. Hammel, and J.W. O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398, 1975.

A. Björner. *Topological methods*, volume 2. North-Holland, Amsterdam, 1995.

J.E. Chacón. Data-driven choice of the smoothing parametrization for kernel density estimators. *Canadian Journal of Statistics*, 37:249–265, 2009.

JE Chacón and T. Duong. Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, 19:375–398, 2010.

J.E. Chacón, T. Duong, and MP Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21:807–840, 2011.

CoCoATeam. Cocoa: a system for doing computations in commutative algebra. available at http://cocoa.dima.unige.it, 2004.

A. Corso and U. Nagel. Monomial and toric ideals associated to Ferrers graphs. *Transactions of The American Mathematical Society*, 361:1371–1395, 2009.

D.A. Cox, J.B. Little, and D. O'Shea. *Using algebraic geometry*, volume 185. Springer, London, Heidelberg, New York, 2005.

F.G. Cozman and P. Walley. Graphoid properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence*, 45:173–195, 2005.

J.N. Darroch. Interactions in multi-factor contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24:251–263, 1962.

A.C. Davison and D.V. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, Cambridge, UK, 1997.

A.P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

C.F.F. di Bruno. Note sur une nouvelle formule de calcul différentiel. *Quart. J. Math*, 1:359–360, 1855.

P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of statistics*, 26:363–397, 1998.

M. Drton and H. Xiao. Smoothness of Gaussian conditional independence models. *Algebraic Methods in Statistics and Probability II*, 516:155–177, 2009.

M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics*. Birkhauser, Basel, 2009.

T. Duong and M. Hazelton. Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, 15:17–30, 2003.

T. Duong and M.L. Hazelton. Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis*, 93:417–433, 2005.

D. Edwards. *Introduction to graphical modelling*. Springer Verlag, 2000.

B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1994.

J. Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21:196–216, 1993.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Chapman and Hall, New York, 1996.

J. Fan and Q. Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer, London, Heidelberg, New York, 2003.

J. Fan, I. Gijbels, T.C. Hu, and L.S. Huang. A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, 6:113–128, 1996a.

J. Fan, Q. Yao, and H. Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83:189–206, 1996b.

D. Geiger, C. Meek, and B. Sturmfels. On the toric algebra of graphical models. *The Annals of Statistics*, 34:1463–1492, 2006.

P. Hall and JS Marron. Lower bounds for bandwidth selection in density estimation. *Probability Theory and Related Fields*, 90:149–173, 1991.

P. Hall and S.R. Wilson. Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47:757–762, 1991.

P. Hall, S.J. Sheather, MC Jones, and JS Marron. On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78:263–269, 1991.

W. Härdle, JS Marron, and MP Wand. Bandwidth choice for density derivatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52:223–232, 1990.

M. Hardy. Combinatorics of partial derivatives. *Electronic Journal of Combinatorics*, 13, 2006.

Jing He. Lecture notes on combinatorial commutative algebra, 2006. http://flash.lakeheadu.ca/~avantuyl/courses/oldcourses/2006_winter_seminar.html, accessed on 19.12.2011.

J. Herzog and T. Hibi. *Monomial ideals*. Springer, London, Heidelberg, New York, 2011.

S. Hojsgaard, D. Edwards, and S. Lauritzen. *Graphical Models with R*. Springer Verlag, 2012.

P.W. Holland and Y.J. Wang. Dependence function for continuous bivariate densities. *Communications in Statistics-Theory and Methods*, 16:863–876, 1987.

M.C. Jones, J.S. Marron, and S.J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91:401–407, 1996.

S.L. Lauritzen. *Graphical models*. Oxford University Press, USA, 1996.

S.L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of statistics*, 17:31–57, 1989.

J.R. Magnus. The moments of products of quadratic forms in normal variables. *Statistica Neerlandica*, 32:201–210, 1978.

K.V. Mardia, G. Hughes, C.C. Taylor, and H. Singh. A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics*, 36:99–109, 2008.

P. McCullagh. Tensor notation and cumulants of polynomials. *Biometrika*, 71: 461–476, 1984.

E. Miller and B. Sturmfels. *Combinatorial commutative algebra*. Springer, London, Heidelberg, New York, 2005.

H.G. Mueller and X. Yan. On local moments. *Journal of Multivariate Analysis*, 76:90–109, 2001.

S. Petrovic and E. Stokes. Betti numbers of Stanley-Reisner rings determine hierarchical Markov degrees. 2010. arXiv:0910.1610v2.

G. Pistone and H.P. Wynn. Generalised confounding with Gröbner bases. *Biometrika*, 83:653–666, 1996.

G. Pistone and H.P. Wynn. Finitely generated cumulants. *Statistica Sinica*, 9: 1029–1052, 1999.

G. Pistone and H.P. Wynn. Cumulant varieties. *Journal of symbolic computation*, 41:210–221, 2006.

G. Pistone, E. Riccomagno, and H.P. Wynn. Computational commutative algebra in discrete statistics. *Contemporary mathematics*, 287:267–282, 2001.

N.S. Razavian, H. Kamisetty, and C.J. Langmead. The von Mises graphical model: Structure learning. 2011.

E. Riccomagno. A short history of algebraic statistics. *Metrika*, 69:397–418, 2009.

M. Rosenblatt. Conditional probability density and regression estimators. *Multivariate analysis ii*, 25:31, 1969.

E. Sáenz-De-Cabezón Irigaray. Combinatorial koszul homology, computations and applications. 2008.

A. Seidenberg. Some remarks on Hilbert's Nullstellensatz. *Archiv der Mathematik*, 7:235–240, 1956.

R. Settimi and J.Q. Smith. Geometry, moments and conditional independence trees with hidden variables. *Annals of Statistics*, 28:1179–1205, 2000.

S.J. Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:683–690, 1991.

BW Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, UK, 1986.

H. Singh, V. Hnizdo, and E. Demchuk. Probabilistic model for two dependent circular variables. *Biometrika*, 89:719–723, 2002.

TP Speed. Cumulants and partition lattices 1. *Australian & New Zealand Journal of Statistics*, 25:378–388, 1983.

D.J. Spiegelhalter, A.P. Dawid, S.L. Lauritzen, and R.G. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8:219–247, 1993.

A. Stuart and J.K. Ord. *Kendall's advanced theory of statistics. Vol. 1: Distribution theory*. Charles Griffin and Co., London, 1994.

Aad Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

MP Wand and MC Jones. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, 88: 520–528, 1993.

M.P. Wand and M.C. Jones. *Kernel smoothing*. Chapman and Hall London, 1995.

H. White. *Asymptotic theory for econometricians*. Academic Press New York, 2000.

J. Whittaker. *Graphical models in applied multivariate statistics*, volume 16. Wiley New York, 1990.