Essays in Econometrics

Vitaliy Oryshchenko

Girton College and Faculty of Economics
University of Cambridge

This dissertation is submitted for
the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation, including tables, footnotes, bibliography and appendices, does not exceed the permitted word limit.

Section 1.2 and Appendix 1.A to Chapter 1 include material from the paper Oryshchenko, V. (2010): "Does Foreign Ownership Matter for Enterprise Training? Empirical Evidence from Transition Countries," published as Chapter 10 in *Global Exchange And Poverty: Trade, Investment and Migration*, ed. by R. E. B. Lucas, L. Squire, and T. N. Srinivasan, Edward Elgar, pp. 269–290. This paper also contains materials from the M.Phil dissertation I submitted in August 2006.

Chapter 2 is based in part on the joint paper "Kernel Density Estimation for Time Series Data" co-authored with Andrew Harvey and forthcoming in the *International Journal of Forecasting.* Results in section 2.3.4 and Appendix 2.6.3 are new.

V. Oryshchenko
Cambridge, 2011

Acknowledgements

It is a great pleasure to have the opportunity to thank my supervisor, Richard Smith, for his support, encouragement and guidance. I would also like to thank Andrew Harvey, M. Hashem Pesaran and my examiners, Melvyn Weeks and James Davidson for their constructive advice.

I would like to thank Prema-Chandra Athukorala, Tom Coupé, Jeff Dayton-Johnson, Ana Perez, Neil Shephard and participants of the Seventh and Eighth Annual Global Development Conferences (St Petersburg, January 2006 and Beijing, January 2007), the LACEA–LAMES Annual Meeting (Bogotá, October 2007), the cemmap/ ESRC Econometric Study Group Workshop on Quantile Regression (London, June 2009), the ESRC Econometric Study Group Annual Conference (Bristol, July 2009), the 7th OxMetrics User Conference (London, September 2009), the 'Stats in the Château' summer school (Jouy-en-Josas, France, September 2009), the Econometrics Seminar (Cambridge, October 2009), the Student Econometrics Workshop (Oxford-Man Institute, Oxford, September 2010) and the Econometrics Workshop (Cambridge, November 2010) for their helpful comments.

I would also like to thank the European Bank for Reconstruction and Development and personally Erik Berglöf and Utku Teksöz for providing the pre-release data from the 2005 round of *the Business Environment and Enterprise Performance Survey.*

I am grateful to the Cambridge Trusts, the Open Society Institute, the Foreign and Commonwealth Office and the TNK-BP for their financial support. Financial support from the Global Development Network and the ESRC under the grant Time-Varying Quantiles, RES-062-23-0129, is also gratefully acknowledged.

Essays in Econometrics

Vitaliy Oryshchenko

Summary

This dissertation contributes to the theoretical understanding and practical application of non- and semi-parametric methods in econometrics. It consists of three chapters.

The first chapter advocates the use of unsupervised statistical learning (clustering) techniques to group observations from a series of repeated cross-sections to create a pseudo-panel of group averages. This clustering method is based on features of the data space and does not require external grouping variables unlike many other methods. Using a model of enterprise training as an example, fixed effects panel data model is estimated using a pseudo-panel of cluster centers.

Chapters 2 and 3 extend univariate kernel methods to the estimation of time-varying distributions and densities subject to moment constraints.

Chapter 2 proposes a weighted kernel density estimator for a time-varying probability density function and the corresponding cumulative distribution function. Time-varying quantiles are estimated by inverting an estimate of the cumulative distribution function. Weighting schemes are derived from those used in time series modelling. Parameters, including the bandwidth, may be estimated by maximum likelihood or cross-validation. Diagnostic checks are constructed based on residuals given by the predictive cumulative distribution function.

Chapter 3 considers a set-up where additional information concerning the distribution of random variables is available in the form of moment conditions. A weighted kernel density estimate reflecting the extra information is constructed by replacing the uniform weights associated with standard kernel density estimator by generalised empirical likelihood implied probabilities. This chapter shows that the resulting density estimator provides an improved approximation to the moment conditions. Moreover, a reduction in variance is achieved due to the systematic use of the extra moment information.

Journal of Economic Literature Codes: C14, C22, C23, C45, F21.

Contents

List of Figures

List of Tables

List of abbreviations

ACF	— autocorrelation function
AMISE	— asymptotic mean integrated squared error
ARMA	— autoregressive–moving average
AR(p)	— autoregression (of order p)
BEEPS	— the Business Environment and Enterprise Performance Survey
CAViaR	— conditional autoregressive value at risk
cdf	— cumulative distribution function
CLT	— Central Limit Theorem
CUE	— continuously updating estimator
CV	— cross-validation
EL	— empirical likelihood
ET	— exponentially tilting (estimator)
EVE	— errors-in-variables estimator
EW	— exponential weighting
EWMA	— exponentially weighted moving average
FDI	— foreign direct investment
FE	— fixed effects
GARCH	— generalised autoregressive heteroscedasticity
GEL	— generalised empirical likelihood
GELKDE	— generalised empirical likelihood-based kernel density estimator
GLS	— generalised least squares
GMM	— generalised method of moments
IQR	— interquartile range
ISB	— integrated squared bias
IV	— instrumental variable
IVar	— integrated variance
JIVE	— jackknife instrumental variables estimator

K-S	— Kolmogorov-Smirnov (test)
KDE	— kernel density estimator
LLN	— Law of Large Numbers
MA(q)	— moving average (of order q)
M(C)AR	— missing (completely) at random
MISE	— mean integrated squared error
ML	— maximum likelihood
MMSE	— minimum mean squared error
MNE	— multinational enterprise
MSE	— mean squared error
OLS	— ordinary least squares
pdf	— probability density function
PIT	— probability integral transform
RCS	— repeated cross-sections
RPKDE	— Rosenblatt-Parzen kernel density estimator
SOM	— self-organising (feature) map
WLLN	— Weak Law of Large Numbers
2SLS	— two-stage least squares

Chapter 1

Effect of foreign direct investment on training: Empirical evidence from transition countries

This chapter discusses estimation of panel data models and inference with pseudo-panels of group averages when the data is a series of repeated cross-sections. Under certain conditions valid inference is possible with pseudo-panels constructed by averaging individual observations in available cross-sections over members of prespecified groups and treating the resulting data as a genuine panel. In particular, consideration is given to methods of cluster analysis (unsupervised learning) that can be used to group observations based on the features of the data space and do not require external grouping variables.

The methods are applied to a model of enterprise training estimated using data from two rounds of the Business Environment and Enterprise Performance Survey.

Contents

1.1 Introduction

The modern global economy transcends regional boundaries with the distinction between national and global matters becoming increasingly fluid. The activities of multinational enterprises (MNE)—a central characteristic of the new order—have changed the pattern of international production and trade. These broad patterns of integration contrast with the less fluid traits of national labour markets, in which labour mobility still remains largely restricted. Analysts do not agree whether foreign firms create or deprive host economies of their skilled labour (Teitel, 2005; Dore, 2001; Barba Navaretti and Venables, 2004). In particular, there is an intense debate about whether positive externalities emanating from foreign firms can spill over to indigenous businesses via learning, imitation and other routes, with empirical results on spillovers being quite mixed (Blomström and Kokko, 1998; Görg and Greenaway, 2001; Moran, Graham, and Blomström, 2005; Hu, 2004; Singh, 2004).

The possibility of knowledge spillovers from multinationals to indigenous enterprises is tightly linked to the type of training offered to employees of those firms. Local enterprises may benefit from knowledge spillovers as trained employees move from foreign to local firms or establish their own businesses. Both theoretical and empirical studies have examined whether it is possible for a firm to extract a part of an increased marginal product of trained workers which drives personnel training (Acemoglu and Pischke, 1998). Several theoretical models have been proposed to explain why multinational enterprises might have higher training incidence as compared with domestic companies (Campbell and Vousden, 2003; Gersbach and Schmutzler, 2003; Fosfuri, Motta, and Ronde, 2001). Relevant hypotheses derive from the idea that multinationals possess an advantageous technology[1] and, in the process of reallocating their production facilities into (predominantly low-wage) host countries, have to train local workers. The bulk of evidence from empirical studies often documents higher training intensity in foreign-owned firms[2].

This chapter offers a model of enterprise training based on both theoretical predictions and results of earlier empirical studies. Section 1.2 discusses the choice of variables included in the model. We hypothesize that there is a positive relationship between foreign ownership and the amount of training provided by the firm.

The model is estimated using the Business Environment and Enterprise Performance

[1] This should be understood to include also managerial practices and 'business culture' which increases general labour productivity.

[2] For instance, such evidence has been documented by Yadapadithaya (2001) for Indian firms, Walsh (2001) for Australian and Parker and Coleman (1999) for the United Kingdom; Bangert and Poor (1993) provide related evidence for Hungarian economy. See also Blomström and Kokko (2003) for an overview of the relevant literature.

Survey data which come as two independent cross-sections. Results of cross-sectional analysis are presented in Appendix 1.A. However, it is likely that there are many firm characteristics which influence both a firm's decision to train and to attract foreign direct investment (FDI) thus rendering cross-sectional estimates inconsistent.

Ideally, we would like to have access to a genuine panel dataset to be able to control for the unobserved time-invariant heterogeneity by estimating a 'fixed effects' model. As panel data in unavailable, we estimate the model with a pseudo-panel of group averages constructed from the available series of repeated cross-sections (RCS). Section 1.3 reviews existing methods of estimation with pseudo-panels of repeated cross-sections and section 1.4 presents estimation results with grouping variables constructed from available firm characteristics.

Using RCS to identify parameters of a genuine panel data model does not come without costs the most important of which is the need to find a grouping variable valid in a sense to be discussed further. In many empirical papers studies concerned with household survey data, the age of the head of household is typically chosen to serve as such a variable. Whilst the age of an individual could be thought of as a reasonable proxy to use as grouping variable, it is far harder to find such variables for models concerned with firm survey data.

Group design is a critical issue for estimation with RCS; some of the problems are discussed in section 1.4.1. However, if a grouping variable is not readily available, it may be possible to identify groups based on features of the data space itself using unsupervised statistical learning (clustering) techniques, *nonparametric* in nature. For example, Cottrell and Gaubert (2007) applies Kohonen's self-organising map (SOM) algorithm to construct a pseudo-panel from several cross-sectional survey datasets.

The benefit of using clustering techniques is that no external grouping variable is necessary. However, if there exist groups such that firms belonging to a same group behave similarly, whereas firms belonging to different groups differ in their behaviour, this should be reflected in observable firm characteristics. It may then be possible to *uncover* the grouping based on those characteristics only.

Section 1.5 illustrates the use of k-means and SOM clustering techniques to construct a pseudo-panel; a brief general overview of clustering methods is given in Appendix 1.B. Pseudo-panel estimates are contrasted with those obtained using cross-sectional analysis.

The drawback of using clustering algorithms to construct pseudo-panels is that the algorithms are often 'black boxes' in that little is known about their theoretical properties. This is particularly true about the SOM algorithm. Other limitations of classical clustering algorithms are discussed in section 1.6 which concludes.

1.2 The model of enterprise training

The basic premise which drives studies of the relationship between training incidence and firm ownership is the idea that multinational enterprises possess a certain advantageous technology or other relevant information and use it to produce goods and services. Thinking in terms of a formal model, multinationals want to sell their products in the foreign market and, therefore, have to decide whether to export or to establish affiliates in the foreign country via foreign direct investment. If chosen, FDI requires multinational enterprises (MNEs) to transfer their technology to subsidiaries, being achieved in several ways including oral communication and on-the-job training. Thus foreign-owned companies should provide more training to their employees. A number of empirical studies have verified this conjecture, documenting some evidence of a positive correlation between foreign ownership and training; see e.g. Blomström and Kokko (2003).

Along with empirical studies, several theoretical models have tried to explain firm behavior regarding personnel training. These models normally emphasise the low quality of the labour force in the target country and the competitive environment both at the intra- and interstate levels. Drawing from numerous empirical studies, Fosfuri et al. (2001) present a theoretical framework which rationalises the importance of spillovers from MNE personnel training. In particular, they argue that competition and the costs of transferring technology are the factors which influence MNE decisions to establish affiliates in the foreign country. They identify values for these two variables where FDI will lead to personnel training and, possibly, to knowledge spillovers (e.g., if competition is low and technology could be easily transferred, then knowledge spillovers are likely to occur).

Perhaps the most obvious factor determining the need for training is the low quality of the workforce in the host countries. That is, the lower the quality of workers in the country, the more training should be provided by a foreign firm to raise the skills of its employees to meet the requirements of the existing advance technology. At the same time, Blomström and Kokko (2003), stressing the importance of labour-force quality in the host country as a determinant of training, note that if local workers are already highly qualified, it is less costly to train them further, so an employer would benefit more from training these workers than if his employees were unskilled. However, Frazis, Gittleman, and Joyce (2000) provide empirical support for education being positively related to the receipt and intensity of formal training. Therefore, there needs to be a distinction between the general quality of the workforce in the country and the quality of employees at a particular enterprise. While national education levels should be negatively related to training, the effect of training at the enterprise level is likely to be positive (Harris, 1999).

Competition is one of the most important factors driving enterprise training. Firms in competitive markets must maintain their positions by advancing the production process through developing the productive skills of their personnel. But firms protected from competition are less likely to engage in costly training. Both domestic and international competition are also likely to affect firm decisions via different channels, suggesting another determinant of training—whether a firm is oriented toward export or domestic markets. Export-oriented firms are more likely to be affected by international competition,[3] while non-export-oriented ones are more sensitive to domestic competitors. Training is linked to firm performance, because non-profitable companies might tighten their budgets by reducing their training expenditures. At the same time, performance should naturally depend on training, otherwise there would have been no point to spend resources on training. Several empirical studies have already documented the positive correlation between firm performance and personnel training (Aragon-Sanchez, Barba-Aragon, and Sanz-Valle, 2003).

The empirical literature on enterprise training has identified a number of other 'conventional' factors used to specify a model, factors which are often determined by the variables available to researchers from a particular survey dataset. Using workplace characteristics as determinants of training, Sutherland (2004) found that, inter alia, age, educational qualifications, occupation and the size of the workplace are important determinants of the probability that an individual receives training. Based on case studies of 42 individual enterprises in five industry sectors, Smith and Hayton (1999) define a set of factors perceived as important for firms when making decisions on personnel training. They found, for example, that the size of the organisation and industry sector have strong positive relationships with training, that investments in new products or technology influence training positively but to a smaller extent and that enterprise ownership (Australian versus multinational) turns out to have no significant effect.

In this literature, firm size is usually positively associated with training. One possible explanation is that training implies economies of scale: early empirical studies had found relatively little training in small firms with less than 50 employees (Frazis et al., 2000). Also, as Harris (1999) notes, 'Large employers actually take a different approach to small employers with regard to the riskiness of investing in their employees', thus large firms tend to provide more training. Finally, it is natural to suggest that general labour market conditions should influence enterprise training arrangements (Acemoglu and Pischke,

[3] As the survey by Yadapadithaya (2001) reveals, 100% of respondents in the MNE group consider global competition and pressure for increased quality, innovation, and productivity as driving forces for providing personnel training.

1997). As Blomström and Kokko (2003) summarise, the amount of training provided to MNE employees 'var[ies] depending on industry, mode of entry, size and time horizon of investment, type of operations and local conditions'.

The data used here come from *the Business Environment and Enterprise Performance Survey* (BEEPS) jointly conducted by the European Bank for Reconstruction and Development and the World Bank in 2002 (BEEPS–II) and 2005 (BEEPS–III). The project surveys managers and firm owners in the countries of Eastern Europe, the former Soviet Union and Turkey (27 countries in total). For comparability reasons we have discarded the data on those firms surveyed in 2005 that began their operations after 1999 and were not covered by the survey round conducted in 2002. Firms established in 2000–2002 were not surveyed in the 2002 round due to survey design, and in the 2005 survey round only firms that had begun operations before 2002 were surveyed. Thus, 1708 observations from the BEEPS–III dataset were dropped. The resulting dataset has 14606 observations: 6667 observations from BEEPS–II and 7939 observations from BEEPS–III.

Previous studies have argued that enterprise training can be endogenously determined together with foreign ownership, in the sense that there may be some factors simultaneously influencing both variables. It is problematic, if possible, to find any valid instruments from the available choice set. Hence, the results of cross-sectional analysis[4] are likely to be highly biased. However, it may be possible to identify the parameters of interest under the assumption of a fixed effects model, that seems to be appropriate for this setting with fixed effects given an interpretation as unobserved factors influencing both a firm's decision to train and to attract foreign investment (or to locate production in a particular host country, in terms of decisions made by multinational parent companies).

Formally, we seek to estimate a linear fixed effect model

$$y_{it} = \alpha + \mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta} + \psi_i + \varepsilon_{it}, \qquad i = 1, \dots, N, \ t = 1, 2, \tag{1.1}$$

where the dependent variable y_{it} is a measure of training intensity, $\mathbf{x}_{it}$ is a vector of explanatory variables including driving and mediating factors and controls as detailed in Table 1.1; α is an intercept term, ψ_i's are the fixed effects, and ε_{it} is the idiosyncratic error term uncorrelated with explanatory variables.

As the data comes as two independent cross-sections, this model cannot be estimated

[4] Results of cross-sectional analysis are presented in Appendix 1.A for completeness. Table 1.5 summarises the set of variables, and estimation results are presented in Tables 1.6–1.8. The main message of the cross-sectional analysis is that there is positive correlation between foreign ownership and training.

Table 1.1: Summary of variables used in the model

Name	Question numbers[a]		Short description
	BEEPS–II	BEEPS–III	
Dependent variable			
Training	q96a3-q96a5, q96b3-q96b5, q92c-q92e	q71a1-q71a3, q71b1-q71b3, q68a3-q68a5	A weighted index for training intensity. Three categories of employees are included in calculation of the index: skilled, unskilled, and support workers.
Driving factors			
Innovativeness	q85a1-q85a4, q85a7-q85a11, q85b1-q85b4, q85b7-q85b11	q60a1-q60a8, q60b1-q60b8, q61a	Proxy for innovativeness (weighted index)
Exports	q14a1-q14a3	q7b, q7c	Share of firm's sales that are exported.
Competition from imports	q19	q10	Dummy for subjective importance of competition from imports (takes value of 1 if the competition from imports in the market for main product/services is very or extremely important).
Skills of available workers	q80l	q54m	Characteristics of skills and education of available workforce (categorical: ranges from 'major obstacle' to 'no obstacle' for business).
Education of firm workforce	q94a-q94f	q69a1-q69a4	Weighted index characterising education of firms' workforce.
Mediating factors/controls			
Foreign ownership	s4c	s5b	Share of firm assets owned by private foreign company/organisation.
Performance	q81a1, q81a2, q81b1	q55a1, q55b1	Relative change in firm's sales since 1998, in real terms.
Monopolisation	q18a	q12ba, q13ba	Dummy for monopolistic/oligopolistic position of a firm.
Legal organisation	s2a	s2a	Dummy for privately owned company.
Full-time employment	q91a1cat	q66acat	Number of full-time employees, hundreds (categorical; treated as continuous with midpoints substituted for appropriate range categories).
Labour regulations	q80k	q54l	Subjective measure of the effects of labour regulations (categorical: ranges from 'major obstacle' to 'no obstacle').
Regional dummies	country	country	Dummies for country.
Industry	q2a-q2h	q2a-q2h	The shares of firms' sales coming from specified sectors of the economy where it operates.

[a] Original codes as used in survey questionnaires.

by conventional panel data methods. Instead, we estimate the model with a pseudo-panel of group averages constructed from the available RCS. Relevant estimators are reviewed in the next section. Group design and estimation results are presented in section 1.4.

1.3 Estimation with RCS: An overview

One advantage of panel data perhaps the most attractive is the possibility of controlling for ‘fixed effects’ (FE): any unobserved time-invariant heterogeneity that is possibly correlated with explanatory variables. However, in many cases where one would like to exploit panel data they may not be available. In such cases estimation methods based on pseudo-panels (sometimes also referred to as ‘synthetic panels’) of group averages have proved to be useful.

Moreover, even when genuine panel data is available, they may be inferior to series of repeated cross-sections due to problems of attrition or insufficient sample size, etc. Attrition is not a concern with RCS data as a new sample is drawn every time a survey is conducted. For instance, the US Current Population Survey, which is a rotating cross-section survey, is used to create short panels by matching individuals across consecutive cross-sections. Peracchi and Welch (1995) investigate attrition problems in the longitudinal data created from the Current Population Survey and show that matching failures are often related to some important household characteristics, and that labour market outcomes are often related to the process that determines attrition. Although explicitly controlling for a number of household characteristics may soften the negative consequences of attrition, it is still questionable whether behavioural relationships estimated with such datasets are representative of population relationships.

Another apparent advantage of RCS data lies in the possibility of explicitly controlling for measurement error: as micro data are available, measurement error variances can be consistently estimated and used to obtain error-corrected estimates (Deaton, 1985; Carraro, Peracchi, and Weber, 1993).

Since the pioneering paper Deaton (1985), a number of studies have been completed dealing with issues of inference using RCS data. These cover both static and dynamic linear FE models and extend to nonlinear models with binary dependent variables.

This section reviews existing methods of estimation with pseudo-panels of repeated cross-sections and discusses relevant identification assumptions. A brief summary of the relevant literature can also be found in Verbeek (2006) and Ridder and Moffitt (2007).

Consider the linear fixed effects model

$$y_{it} = \alpha + \mathbf{x}_{it}^{\mathsf{T}}\boldsymbol{\beta} + \psi_i + \varepsilon_{it}, \quad t = 1, \ldots, T, \quad i = 1, \ldots, N, \tag{1.2}$$

where t indexes cross-sections over time and i indexes individuals, α is an intercept term. It is assumed that regressors, $\mathbf{x}_{it}$, are uncorrelated with the idiosyncratic error term, ε_{it}, but can be correlated with fixed effects, ψ_i, i.e. $\mathbb{E}\left\{\varepsilon_{it}|\psi_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}\right\} = 0$ for all t.

When RCS data are available, (1.2) may be written as

$$y_{i(t)t} = \alpha + \mathbf{x}_{i(t)t}^{\mathsf{T}}\boldsymbol{\beta} + \psi_{i(t)} + \varepsilon_{i(t)t}, \quad t = 1, \ldots, T, \quad i(t) = 1, \ldots, N(t), \tag{1.3}$$

where conventionally the notation $i(t)$ rather than i is used to emphasise that individuals are (possibly) different in each cross-section. Even when data for the same individuals is available in more than one cross-section, their identities remain unknown and entries cannot be matched by i. Therefore, conventional panel data methods such as within estimation or first-differencing cannot be applied.

If the *unobserved* fixed effects are uncorrelated with regressors, one could simply pool all available cross-sections and apply either a least squares estimator on a pooled sample or a random-effects estimator (the latter being efficient in this case). However, $\psi_{i(t)}$ is often likely to be correlated with right-hand side variables, in which case the vector of coefficients, $\boldsymbol{\beta}$, in (1.3) cannot be consistently estimated by these methods due to omitted variables bias. A transformation which removes $\psi_{i(t)}$'s, such as first differencing or within transformation, could be applied if genuine panel data were available, but not with RCS. This motivates the quest for a robust estimator of $\boldsymbol{\beta}$ in (1.3).

Wald estimator

Suppose there is a single regressor and two RCS. (1.3) then becomes

$$y_{i(t)t} = \alpha + \beta x_{i(t)t} + \psi_{i(t)} + \varepsilon_{i(t)t}, \qquad t = 1, 2, \; i(t) = 1, \ldots, N(t). \tag{1.4}$$

Assume there exists a grouping variable, g, which takes G different values, $1, \ldots, G$, such that every individual can be unambiguously identified as a member of a certain group. Assume further that the population belonging to each group is *fixed* through time. Upon taking expectations conditional on g we obtain

$$y_{gt} = \alpha + \beta x_{gt} + \psi_g, \qquad t = 1, 2, \quad g = 1, \ldots, G, \tag{1.5}$$

where $z_{gt} = \mathbb{E}\left\{z_{i(t)t}|\text{object } i(t) \text{ belongs to group } g\right\}$.

To identify β it is sufficient to set $g = t$, i.e. to ignore the cross-section dimension. Then (1.5) becomes $y_{\cdot t} = \alpha + \beta x_{\cdot t} + \psi_{\cdot}$, $t = 1, 2$, where $z_{\cdot t} = \mathbb{E}\left\{z_{i(t)t}|t\right\}$. Taking first differences yields a natural analogue estimator of β,

$$\hat{\beta}_W = \frac{\frac{1}{N(2)}\sum_{i(2)=1}^{N(2)} y_{i(2)2} - \frac{1}{N(1)}\sum_{i(1)=1}^{N(1)} y_{i(1)1}}{\frac{1}{N(2)}\sum_{i(2)=1}^{N(2)} x_{i(2)2} - \frac{1}{N(1)}\sum_{i(1)=1}^{N(1)} x_{i(1)1}}, \tag{1.6}$$

where division by $N(t)$ is necessary as the number of observations may differ across cross-sections. This estimator was first proposed by Wald (1940). It is immediately apparent that if the population is *not* fixed through time, then in general ψ_g will vary with t and the above estimator will not be consistent due to so-called 'survival bias'.

It is useful to note that the above estimator is equivalent to an instrumental variables (IV) estimator with group dummies used as instruments[5]; see Pakes (1982) for the asymptotic properties of these Wald-type estimators.

If more than two time periods are available, several estimators of β can be computed by taking differences in time means between first and second time periods, second and third, and so on. Differences in estimates obtained in this manner can be tested for significance and, if found to be insignificant, alternative Wald estimators can be combined to give an efficient estimator for β. It turns out that a minimum-variance linear combination of any full set of linearly independent pairwise Wald estimates is the Prais and Aitchison (1954) Generalised Least Squares (GLS) estimate for grouped data (Angrist, 1991, Proposition 1); for a proof see Angrist (1988). Asymptotic properties of estimators for RCS data based on the IV approach of Angrist (1991) have been explored in Verbeek and Vella (2005) and Moffitt (1993).

Clearly, in a model with more than one regressor, one would require more time periods than the number of included regressors. When this is not the case, one might seek some other categorical variable that can serve as a valid instrument and base grouping on this variable interacted with time dummies. It will often be the case that a researcher will have certain flexibility in how to choose a grouping variable, thus producing different numbers of groups, and hence, different numbers of simple Wald estimators as above. An overidentification test can then be constructed; see Angrist (1991).

[5] Bartlett (1949) suggested dropping the middle third of observations when estimating the slope coefficient; see also Reiersøl (1950), Mallios (1969), and Neyman and Scott (1951) for relevant discussions, and Madansky (1959) for an early overview of the relevant literature.

Errors-in-Variables Estimators

The literature on pseudo-panels originated in Deaton (1985). The idea is that if one can group individuals into cohorts, it would be possible to track cohorts over time and, 'if there are additive individual fixed effects, there will be corresponding additive cohort fixed effects for the cohort population'. His errors-in-variables (EVE) estimator is motivated by viewing sample cohort averages as consistent estimates of the population cohort means but observed with error. The availability of individual level data allows estimates of variances and covariances of cohort means to be computed and then used to correct the estimator for measurement error.

Averaging observations for each t over those $i(t)$ in group g observed in the survey taken at t, we can write (1.3) in terms of the observed sample group averages as

$$\bar{y}_{gt} = \alpha + \bar{\mathbf{x}}_{gt}^{\mathsf{T}}\boldsymbol{\beta} + \bar{\psi}_{gt} + \bar{\varepsilon}_{gt}, \qquad t = 1, \dots, T, \quad g = 1, \dots, G, \tag{1.7}$$

where the average of the fixed effects for every group g is now *not* constant over time. Deaton (1985) proposes to consider a version of equation (1.7) in population group means, viz.

$$y_{gt}^{\star} = \alpha + \mathbf{x}_{gt}^{\star\mathsf{T}}\boldsymbol{\beta} + \psi_g^{\star} + \varepsilon_{gt}, \qquad t = 1, \dots, T, \quad g = 1, \dots, G. \tag{1.8}$$

If the population belonging to each group is fixed through time, $\psi_g^{\star}$ is time-invariant and can be replaced with group dummies in the above equation. However, unless the cohort size is very large, $\bar{\psi}_{gt}$ cannot be employed as a good approximation for $\psi_g^{\star}$.

Extending the above idea, Verbeek and Nijman (1993) propose indexing the class of errors-in-variables estimators by the proportion $\gamma \in [0, 1]$ of error variance to be eliminated. In particular, sample group averages $\bar{y}_{gt}$ and $\bar{\mathbf{x}}_{gt}$ are considered as error-ridden measurements of corresponding population group means $y_{gt}^{\star}$ and $\mathbf{x}_{gt}^{\star}$ where measurement errors are assumed to be normally distributed, viz.

$$\begin{pmatrix} \bar{y}_{gt} - y_{gt}^{\star} \\ \bar{\mathbf{x}}_{gt} - \mathbf{x}_{gt}^{\star} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} \sigma_y & \boldsymbol{\sigma}^{\mathsf{T}} \\ \boldsymbol{\sigma} & \boldsymbol{\Sigma} \end{pmatrix} \right).$$

The errors-in-variables estimator for $\boldsymbol{\beta}$ indexed by the values of $\gamma \in [0, 1]$ takes the following form:

$$\widehat{\boldsymbol{\beta}}(\gamma) = (M_{xx} - \gamma\Sigma)^{-1}(m_{xy} - \gamma\boldsymbol{\sigma}), \tag{1.9}$$

where $M_{xx} = (GT)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}$, $m_{xy} = (GT)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$, $\mathbf{X}$ and $\mathbf{y}$ are vertically stacked and demeaned data; see equations (5)–(7) in Verbeek and Nijman (1993). With $\gamma = 1$, estimator (1.9), $\widehat{\boldsymbol{\beta}}(1)$, is Deaton's EVE, whereby all error variance is eliminated; whereas with $\gamma = 0$, $\widehat{\boldsymbol{\beta}}(0)$ is the within estimator applied to the pseudo-panel of group averages whilst ignoring measurement errors.

Under the assumption that the number of cohorts tends to infinity, Verbeek and Nijman (1993) show that

$$\underset{G\to\infty}{\text{plim}}\widehat{\boldsymbol{\beta}}(\gamma) = \left(\Sigma^\star + (\tau-\gamma)\Sigma\right)^{-1}\left(\Sigma^\star\boldsymbol{\beta} + (\tau-\gamma)\boldsymbol{\sigma}\right),$$

where $\Sigma^\star$ is the asymptotic within variation in the true group means of $\mathbf{x}$'s, and $\tau = (T-1)/T$.

It is then easily verifiable that if unobserved fixed effects are correlated with $\mathbf{x}$'s, this estimator is consistent for finite T only if $\gamma = \tau$, and hence, Deaton's estimator is inconsistent unless $T \to \infty$, in which case $\tau \to 1$. Furthermore, they show that in finite samples the minimum mean squared error (MMSE) estimator is characterised by $\gamma^{optimal} < \tau$. Numerical values of MSEs of competing estimators suggest it is never optimal to choose $\gamma = 1$. The performance of estimators with $\gamma = \gamma^{optimal}$ and $\gamma = \tau$ is very close and the differences become negligible as group size gets large enough[6], say, greater than 50. Ultimately, the best choice in most practical applications is to set $\gamma = (T-1)/T$.

Finally, Devereux (2003) shows that Deaton's EVE is exactly equivalent to the Jackknife Instrumental Variables Estimator (JIVE) with the set of group dummies as instruments; see inter alia Phillips and Hale (1977), Angrist, Imbens, and Krueger (1995), Angrist, Imbens, and Krueger (1999), and Blomquist and Dahlberg (1999).

To conclude which particular estimator for RCS data is chosen depends largely on which of the possible assumptions about data dimension is most relevant. With data being aggregated over members of predefined groups, new asymptotic considerations emerge: apart from conventional arguments in terms of the number of time periods or the number of individuals in each time period approaching infinity, asymptotics in terms of the number of groups or the number of individuals per group, or some combinations of these can be considered (see McKenzie (2006) for an example of sequential and diagonal path asymptotic arguments).

It is often the case that the number of time periods is usually too small to rely on $T \to \infty$ asymptotics. Even when one has access to a long panel, it is questionable

[6] Although the bias of the within estimator applied to a synthetic panel is likely to be small if the number of observations per group is sufficiently large, the higher the number of observations per group, the smaller the number of observations in the synthetic panel and, hence, the higher the variance of the within estimator. The cell size—group number trade-off is an important aspect of any applied work using RCS data; see also Verbeek and Nijman (1992).

whether the assumption of a fixed population can be sensibly maintained. Assuming $G \to \infty$ is quite problematic too, as there is often a physical limit beyond which the number of groups cannot be increased; see e.g. Hsiao (2003). Ultimately, in short panels, the only realistic assumption may be that $N \to \infty$ such that the number of groups stays constant. In this case, if the number of observations per group is large, correcting for measurement error is unnecessary for consistency, but may still be made for efficiency reasons.

Empirical applications

An initial application using RCS is Browning, Deaton, and Irish (1985) which uses the Deaton (1985) estimator in the context of an empirical analysis of family labour supply and consumption based on household expenditures data from British Family Expenditure Surveys conducted in 1970-77. The age of the head of household is used to identify cohorts creating a pseudo-panel of cohort averages grouping over five-year age bands subdivided as to whether the head of household is a manual or non-manual worker (resulting in 16 groups: 8 age bands interacted with two categories, tracked over seven years). Sample cohort averages are treated as population cohort means. Blundell, Meghir, and Neves (1993) studies an intertemporal model for labour supply and consumption using the same survey data over a longer period (up to the mid-80s) and constructs an exactly aggregated pseudo-panel of year-of-birth cohorts following Browning et al. (1985). Later, Moffitt (1993) applies the Browning et al. (1985) linear fixed effects model to the US Current Population Survey.

The empirical literature originated by these studies has a long tradition of using year-of-birth cohorts to create pseudo-panels of cohort averages which are then usually treated as a genuine panel; correction for measurement error is rarely made. An intensively used RCS dataset is the UK Family Expenditure Survey, being used to create a pseudo-panel identifying groups by the year-of-birth of the head of household with five- (Alessie, Devereux, and Weber, 1997; Banks, Blundell, and Preston, 1994; Blundell, Browning, and Meghir, 1994; Dargay, 2001), four- (Banks, Blundell, and Tanner, 1995), and two-year age bands (Gassner, 1998); age bands interacted with education level (Blundell, Duncan, and Meghir, 1998) and residential location (Dargay, 2002; Propper, Rees, and Green, 2001). A number of studies using US datasets employ year-of-birth grouping to construct pseudo-panels based on the US Consumer Expenditure Survey (Attanasio, 1993) and the Current Population Survey (Card and Lemieux, 1996; Chay and Lee, 2000). Other studies employ year-of-birth groupings using German Income and Expenditure Survey (Börsch-Supan, Reil-Held, Rodepeter, Schnabel, and Winter, 2001) and

Taiwanese Survey of Personal Income Distribution (Levenson, 1996; McKenzie, 2006). Gardes, Langlois, and Richaudeau (1996) group over interactions of five income classes, three groups for the age of the head of household, and two classes for education using the Canadian Households Expenditures Survey, whereas Robertson (2003) identifies groups with four education levels, four age groupings, and five industries based on the Mexican National Urban Employment Survey.

Few studies use population surveys identifying groups of individuals based on criteria other than the year-of-birth. For instance, Beine, Bismans, Docquier, and Laurent (2001) using the US Consumer Expenditure Survey identifies cohorts by interactions of the highest education level in the household, the race of the head of household (white and non-white), and geographical location; DiNardo (1993) combines two surveys; Drug Enforcement Administration's STRIDE (System to Retrieve Information from Drug Evidence) and MTF (Monitoring the Future), and defines state interacted with year groups.

To the best of our knowledge, the only empirical study that uses firm-level RCS data is that by Morrison Paul and Nehring (2005) which investigates the US Department of Agriculture farm survey and defines 130 groups as interactions of 13 'cohorts' with 10 states included in the dataset, the 13 'cohorts' being defined using the farm typology developed by the US Department of Agriculture Economic Research Service (annual sales and other factors e.g. family and non-family farms).

Most empirical papers group the data into a small number of cohorts with a fairly large number of observations per cohort so that measurement error can arguably be ignored and the pseudo-panel treated as a genuine panel. The cell size conventionally adopted is about 100 observations per cell, which it is argued is large enough to serve as a good approximation to the population group mean. Some researchers use unequally spaced bands to obtain approximately equal cell sizes for construction of a pseudo-panel. Others exclude those cells with too few observations available as it is problematic to treat averages for those cells as population group means. It is questionable, however, whether such a practice leads to consistent estimation of model parameters as the sample remaining after these small cells are deleted may not then be random even if the original sample was.

Asymptotic results used in applied work are usually determined by the dimensions of the available dataset or the estimator applied. Most empirical papers reviewed above assume either the number of groups or the number of observations per group to be large. Few studies use long pseudo-panels with large T asymptotics.

1.4 Pseudo-panel analysis

1.4.1 Group design

A critical issue for estimation with RCS is the design of the groups used to construct a pseudo-panel of group averages. First, there is always a trade-off between the number of groups (and hence the number of observations in the resulting pseudo-panel) and the size of a group. While more groups give more reliable estimates, reducing the size of a group affects the reliability of sample group averages as consistent estimates of the corresponding population group means. Furthermore, often a problem arises with groups very different in size or discovering empty group-time cells for some variables (i.e., missing data points in the constructed pseudo-panel). Combining such groups does not seem to be good practice, as it is likely to introduce a further distortion. Obviously, different weighting schemes may be employed, but how many observations any given group should contain for a reliable approximation of the corresponding population quantity remains an open question. Second, and more important, defined grouping variables should be exogenous in the model (just as valid instrumental variables are); otherwise computed statistics will be inconsistent. This problem is endemic. A credible group design valid for consistent estimation of the parameters requires considerable judgment on the part of the researcher.

Based on the available data, we use the following variables as a basis for grouping; (apart from country, all variables have been redefined to result in approximately equally sized categories):

- Year firm began operations in a particular country (15 dates, five unequally spaced bands covering 1800–1989 inclusive, then ten one-year categories covering 1990–1999, both dates inclusive).

- Country (27 countries).

- Sector of the economy in which firm operates (four categories: mining and quarrying, construction, and transport, storage and communication; manufacturing; wholesale, retail, and repairs; and real estate, renting and business services, hotels and restaurants, and other sectors).

- Legal status of firm (four categories: single proprietorship; partnership or cooperative; corporation or other private sector; and state or municipal-owned, corporatised state-owned or other state-owned).

We compare three alternative group designs based on the above four variables, that result in a significantly larger number of observations in the constructed pseudo-panel; see Table 1.2.

Table 1.2: Characteristics of alternative group designs

No.	Description	No. of groups		Group-time cell size[a]		
		By construction	Effective[b]	Min	Max	Average
1	Interactions of country and sector	108	108	12	476	65.7
2	Interactions of year, sector, and legal status	240	238	1	126	29.8
3	Interactions of country and legal status	108	107	1	451	66.2

[a]Only nonempty cells are taken into account.

[b]After excluding groups for which at least one group-time cell has no observations.

1.4.2 Estimation results

The model is estimated using the EVE (1.9) with Σ and $\boldsymbol{\sigma}$ replaced by the sample estimates. Table 1.3 reports estimated coefficients and their respective t-ratios for the enterprise-training model where the dependent variable is the index of training intensity. Since all proposed group designs result in groups of very different sizes, we report only those estimates which were obtained by weighting observations by the square root of the corresponding group size. Four sets of estimates are reported for each group design: conventional within estimates which ignore the measurement error problem and three sets of errors-in-variables estimates which subtract different proportions of the estimated measurement error variance (this proportion is given by the parameter γ in eq. (1.9)).

All four estimators yield reasonably close estimates within each of the three sets defined by the alternative group designs. However, there are major differences across alternative group designs: for some variables coefficient estimates switch signs and their significance moves above/below the threshold (kept at the conventional 5% level). For instance, competition from imports has a significant effect on training if we consider group designs 3 and 1 (marginally significant), but appears insignificant with group design 2. Similar patterns can be found for other coefficient estimates.

The most striking difference, however, is between the estimates reported for the FE model and those obtained for linear regressions estimated with cross-sectional datasets (see Table 1.8 in Appendix 1.A). Innovativeness, highly significant in the model estimated

Table 1.3: Estimation results for the model of enterprise training: FE specification

Variables	Coefficient estimates[a] of the linear FE model for training intensity											
	Group design 1				Group design 2				Group design 3			
	Within	EVE; $\gamma =$			Within	EVE; $\gamma =$			Within	EVE; $\gamma =$		
	$(\gamma = 0)$	$\frac{G-K-1}{G}$	$\frac{T-1}{T}$	1	$(\gamma = 0)$	$\frac{G-K-1}{G}$	$\frac{T-1}{T}$	1	$(\gamma = 0)$	$\frac{G-K-1}{G}$	$\frac{T-1}{T}$	1
Innovativeness	0.2550	0.2295	0.2584	0.2063	-0.0062	-0.0191	-0.0126	-0.0198	0.1203	-0.0870	0.0733	-0.2175
	(1.97)	(0.84)	(1.34)	(0.66)	(-0.06)	(-0.17)	(-0.12)	(-0.18)	(0.78)	(-0.26)	(0.37)	(-0.44)
Exports	0.1747	0.3955	0.2528	0.4647	-0.0856	-0.0964	-0.0909	-0.0970	0.1771	0.6867	0.2982	1.0065
	(1.24)	(1.11)	(1.16)	(1.07)	(-1.03)	(-0.97)	(-0.98)	(-0.97)	(1.13)	(1.00)	(1.00)	(0.92)
Competition from imports	0.1321	0.1929	0.1594	0.2045	0.0158	0.0164	0.0162	0.0165	0.1523	0.2886	0.2017	0.3349
	(2.21)	(1.85)	(1.97)	(1.80)	(0.40)	(0.36)	(0.37)	(0.36)	(2.21)	(2.10)	(2.18)	(1.92)
Skills of available workers	0.0346	0.0563	0.0435	0.0617	-0.0001	0.0004	0.0001	0.0005	0.0162	0.0400	0.0245	0.0491
	(1.55)	(1.52)	(1.48)	(1.51)	(-0.01)	(0.02)	(0.01)	(0.02)	(0.72)	(0.96)	(0.76)	(0.99)
Education of firm workforce	-0.3414	-0.5239	-0.4103	-0.5751	-0.5607	-0.5947	-0.5777	-0.5965	-0.2470	-0.5043	-0.3181	-0.6403
	(-2.00)	(-1.78)	(-1.86)	(-1.72)	(-3.84)	(-3.55)	(-3.65)	(-3.54)	(-1.27)	(-1.35)	(-1.29)	(-1.25)
Foreign ownership	-0.3379	-0.5013	-0.4063	-0.5387	-0.0336	-0.0322	-0.0330	-0.0321	-0.2058	-0.1710	-0.2037	-0.1457
	(-3.64)	(-2.91)	(-3.30)	(-2.75)	(-0.40)	(-0.39)	(-0.42)	(-0.39)	(-2.27)	(-1.06)	(-1.68)	(-0.76)
Performance	-0.0985	-0.1106	-0.1066	-0.1098	-0.0115	-0.0062	-0.0090	-0.0059	-0.0140	0.0291	-0.0016	0.0509
	(-2.17)	(-1.72)	(-2.09)	(-1.56)	(-0.38)	(-0.16)	(-0.25)	(-0.16)	(-0.32)	(0.37)	(-0.03)	(0.50)
Monopolisation	0.1139	0.0721	0.1024	0.0563	0.1869	0.2007	0.1938	0.2014	0.2063	0.2695	0.2368	0.2751
	(2.18)	(0.69)	(1.34)	(0.47)	(4.78)	(4.36)	(4.42)	(4.35)	(3.13)	(2.31)	(2.80)	(1.98)
Full-time employment	0.0048	0.0110	0.0076	0.0118	0.0021	0.0024	0.0022	0.0024	-0.0362	-0.0994	-0.0561	-0.1276
	(0.30)	(0.37)	(0.37)	(0.35)	(0.24)	(0.21)	(0.21)	(0.21)	(-1.60)	(-1.86)	(-2.11)	(-1.57)
Labour regulations	0.0293	0.0435	0.0366	0.0453	-0.0312	-0.0307	-0.0310	-0.0307	0.0309	0.0654	0.0448	0.0744
	(1.20)	(1.39)	(1.38)	(1.37)	(-1.48)	(-1.23)	(-1.32)	(-1.22)	(1.29)	(1.53)	(1.39)	(1.48)
Number of obs.				14606				14594				14585

[a] All the estimators employed observations weighted by the square root of the corresponding group size. t-ratios reported for the within estimator are based on standard errors that allow for a serially correlated error variance structure; t-ratios for errors-in-variables estimators are based on asymptotic standard errors; the estimator of the covariance matrix is given in equation (38) in Deaton (1985); see also Devereux (2007).

with cross-sections, appears insignificant in the FE model estimated with pseudo-panels. Furthermore, the previously positive and significant effect of foreign ownership now turns negative wherever it is significant.

Being a monopoly appears to be significantly positive with a rather high magnitude (also positive and significant in cross-section regressions, although of a smaller magnitude). At the same time, competition from imports has a positive and significant effect on training, suggesting that it is global rather than local competition that drives training. Finally, the effect of formal qualifications (education) of firm employees, that proxy cognitive skills and ability, is negative, meaning that if firm employees have better education, less training will be provided. This contradicts the hypothesis that firms train highly educated employees more because in this way training yields a higher return at lower costs. Other variables, with some occasional exceptions, appear to have no significant effect on enterprise training.

It should be kept in mind that these estimates are quite sensitive to group design, especially if some group designs result in grouping variables which are endogenous to the model. There is no way, as yet, to discriminate between the alternative group designs, and this is an issue deserving further attention.

1.5 Estimation with ‘self-organised’ pseudo-panels

Pseudo-panel estimation relies heavily on the availability of an external grouping variable used to combine RCS into a pseudo-panel of group averages. Such grouping variables may be of dubious quality or may not be readily available. However, if is it hypothesized that there exist groups such that firms belonging to a same group behave similarly, whereas firms belonging to different groups differ in their behaviour, this should be reflected in observable firm characteristics. It may then be possible to uncover the grouping based on those characteristics only using clustering techniques.

A brief review of clustering techniques is given in Appendix 1.B. For the purposes of this chapter we will concentrate on two algorithms: the k-means and SOM. The attraction of the latter algorithm is the ability to provide a two-dimensional visualisation of the data. Alternatives are discussed in section 1.6.

1.5.1 k-means clustering of the BEEPS dataset

The k-means algorithm can be used to find the optimal number of clusters by setting up what is essentially a double optimisation program. At the first stage, the k-means is run on a dataset for all possible values of k, the number of clusters. In practice, however, with large datasets there will be a need to stop the search at some chosen number of clusters although, in principle, one can run k-means for all values from two to the number of observations. Each such run consists of several replicas from which the 'best' is chosen based on the adopted criterion.

The k-means algorithm seeks to minimise the sum-of-squares criterion. To be specific, let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be the data set and let $\mathscr{C} = (C_1, \ldots, C_k)$ be its clustering into k clusters. Let $d(\mathbf{x}_j, \mathbf{x}_l)$ be the distance between $\mathbf{x}_j$ and $\mathbf{x}_l$. The objective function is

$$J = \sum_{j=1}^{k} \sum_{i \in C_j} d(\mathbf{x}_i, c_j),$$

where c_j is the geometric centroid of the data point in cluster C_j. The distance measure used in the classical k-means algorithm is the Euclidean distance.

A criterion is then computed to characterise the 'goodness' of the resulting clustering. Several such criteria are proposed in the literature; we use the Davies-Bouldin index[7] (Davies and Bouldin, 1979) defined as

$$DB(\mathscr{C}) = \frac{1}{k} \sum_{j=1}^{k} \max_{l \neq j} \left[\frac{S_k(C_j) + S_k(C_l)}{d(c_j, c_l)} \right],$$

where $S_k(C_j)$ is the average distance of all objects from cluster j to their cluster centre and $d(c_j, c_l)$ is the distance between cluster centers.

At the second stage, clustering is chosen which minimises the Davies-Bouldin index over all possible clusterings. This is then treated as the optimal clustering.

Figure 1.1 shows the Davies-Bouldin index for k-means clusterings of the BEEPS dataset[8] where the algorithm was searching over clusterings with 2 to 200 groups. The Davies-Bouldin index has its minimum of ≈ 1.52 corresponding to clustering with 44 groups. The associated sum of squared errors is shown in the lower panel. It is unclear why the Davies-Bouldin index jumps at a number of points. The optimal clustering has group sizes ranging from 23 to 649 objects with median size 257.5; 25-th and 75-th percentiles are, respectively, 129.5 and 379 observations.

[7] The index measures the average similarity of each cluster with its most similar cluster. It does not depend on the clustering algorithm employed and requires the distance function and the dispersion measure to be specified along with the rule to choose the representative vector from each cluster. Comparing several popular cluster validity indices, Kim and Ramakrishna (2005) conclude in favour of the Davies-Bouldin index as having the best performance in their experiments.

[8] The variables used in clustering are the regressors in (1.1).

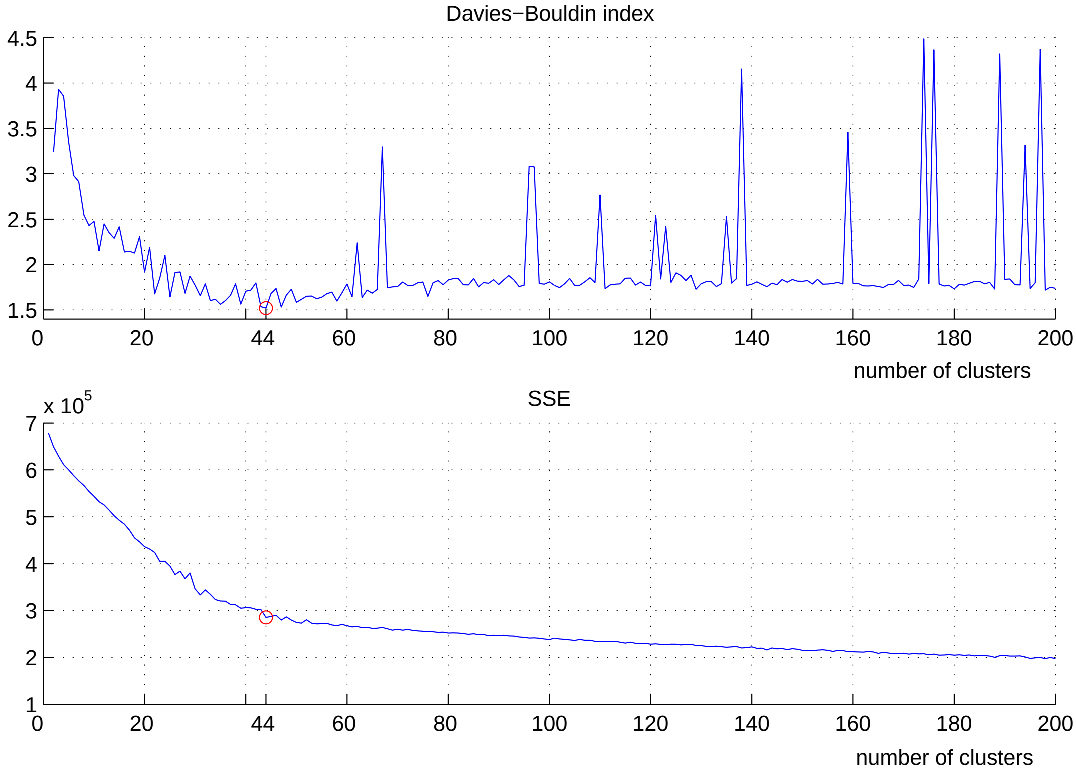

Figure 1.1: Davies-Bouldin index for k-means clustering of the data

1.5.2 Cluster analysis with the SOM algorithm

The Kohonen self-organising feature map algorithm, primarily designed for the preprocessing of patterns for their recognition and for visualisation of high-dimensional object spaces on a two-dimensional display, can be used to perform unsupervised classification of the data space. See Kohonen (1997) for details of the SOM method; Deboeck and Kohonen (1998) present a collection of recent applications of the SOM method in finance.

The SOM is a type of artificial neural network that is used to produce a (typically) two-dimensional discretized representation of the input space. In Euclidean space, the SOM defines a mapping from the input space onto a two-dimensional array of nodes, each having an associated reference vector in the input space. The nodes are connected to adjacent nodes by a neighbourhood relation which dictates the topology of the map. The resulting 'elastic net' adjusts during the learning process to best cover the 'data cloud'. In each training step, one vector $\mathbf{x}$ from the input space is chosen randomly and a node is found whose reference vector in the input space is closest to $\mathbf{x}$. This node is called a best matching unit (BMU). The weight vectors of SOM are then updated in such a way that BMU is moved closer to $\mathbf{x}$ in the input space. As the nodes are connected by a neighbourhood relation, the adjacent nodes are updated as well. The updating is illustrated in Figure 1.2.

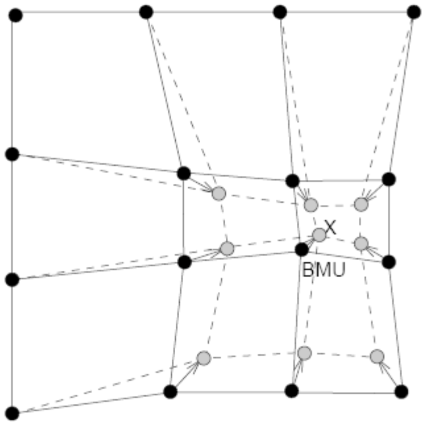

Figure 1.2: Updating the BMU and its neighbours towards the input vector

Source: Vesanto, Himberg, Alhoniemi, and Parhankangas (2000, Fig. 3).
The input vector is marked with X. The solid and dashed lines correspond to situation before and after updating, respectively.

After the learning stage, the SOM net can be used to visualise the data. Figure 1.3 shows the U-matrix representation of the SOM[9] which captures the relative distances between the map units. The U-matrix value of a particular node is the average distance between the node and its closest neighbors. In Figure 1.3 darker colours correspond to the weight vectors (nodes) that are farther away from each other, and hence give cluster borders; lighter colours represent clusters themselves. It is clear from this map that the data space is clustered into a number of unequally-sized groups.

The SOM map can further be analysed by the k-means method to obtain the optimal number of clusters (see section 1.5.1 for a description of the k-means algorithm). The Davies-Bouldin index for SOM clustering with the k-means algorithm is minimised at 35 clusters, and the resulting clustered U-matrix representation of the SOM is shown in Figure 1.4.

As can be seen from the clustered map, some clusters appear to be rather small; and indeed, when asked to classify the dataset with 35 groups, SOM results in 11 empty clusters. The 24 non-empty clusters have sizes ranging from 90 to 2706 observations with 25-th, 50-th, and 75-th percentiles being 253, 387, and 754.5 observations respectively.

[9] The SOM learning algorithm has been initialised on the variance-normalised dataset to perform a sequential learning process. Computations were performed using the SOM toolbox for Matlab; see Vesanto et al. (2000).

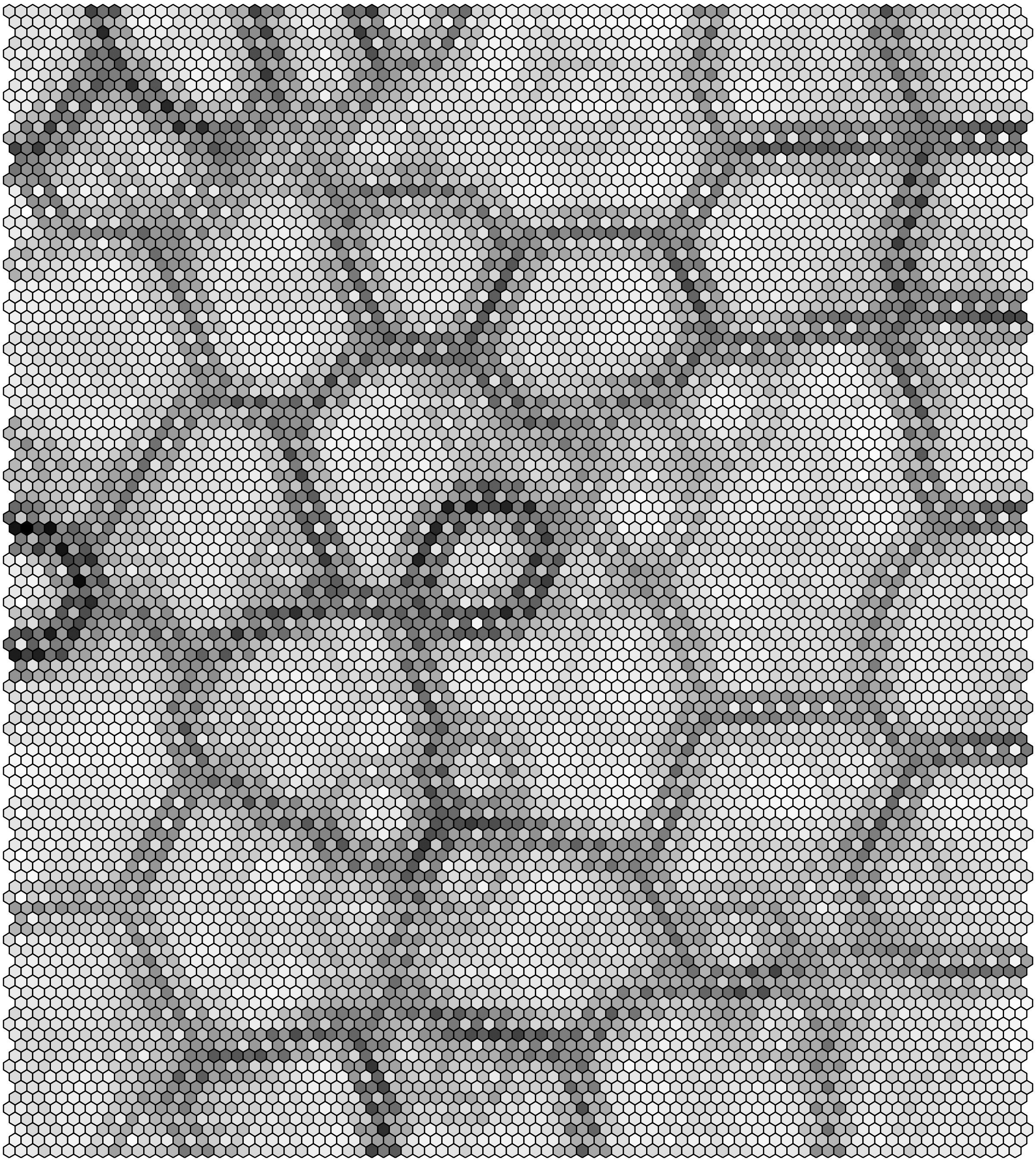

Figure 1.3: U-matrix representation of SOM

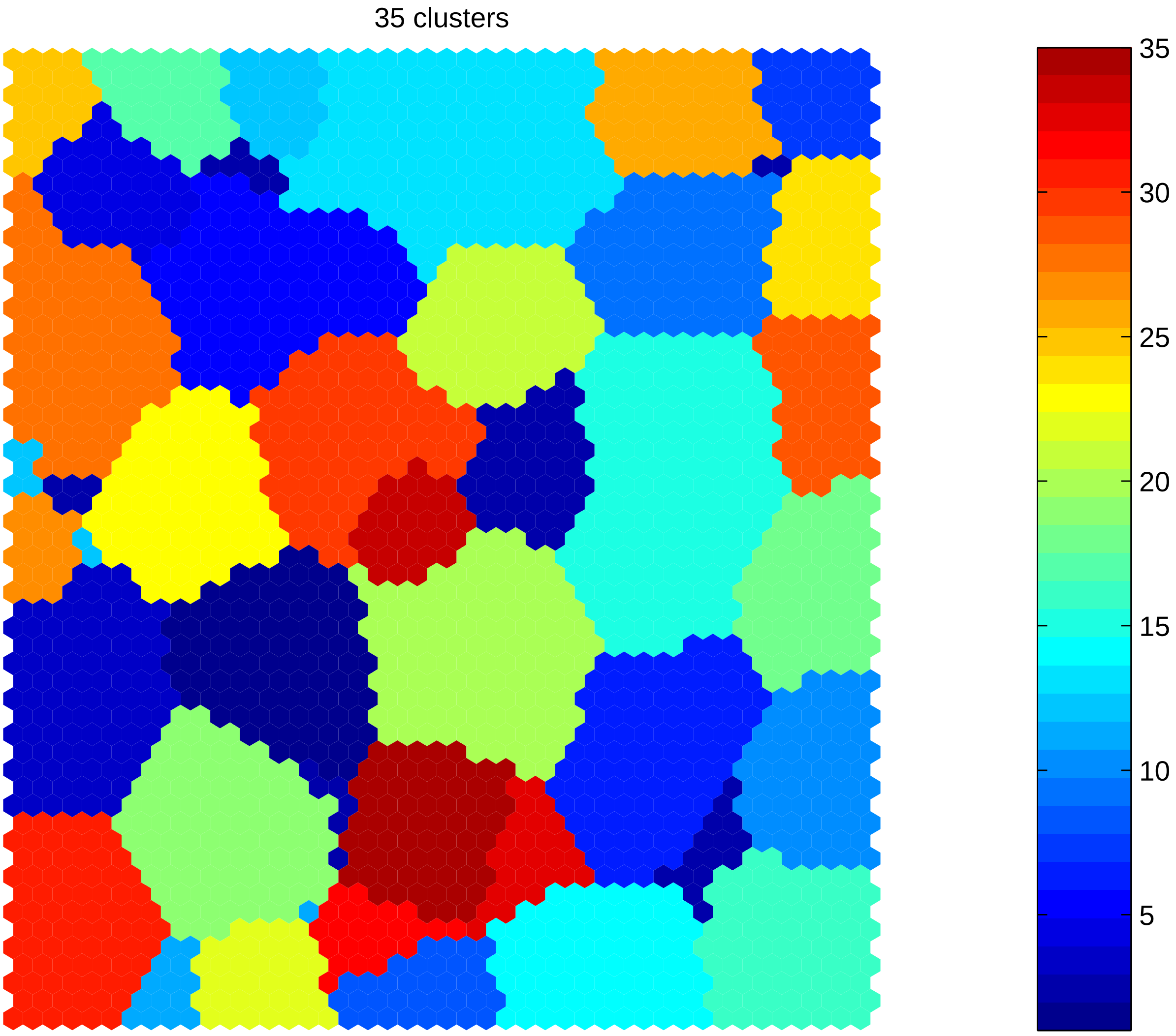

Figure 1.4: Clustered SOM net

1.5.3 Estimation with a pseudo-panel of 'self-clustered' data

The clustering produced in sections 1.5.1 and 1.5.2 can be used to construct a pseudo-panel of cluster averages and, hence, to estimate the model considered in this chapter following the same steps as in section 1.4.

Before considering the estimation results, it is interesting to compare two clusterings. Figure 1.5 visualises a cross-tabulation of the clusterings produced by k-means and SOM algorithms. The columns of the matrix correspond to the 44 groups resulting from k-means clustering, whilst the rows represent the 23 clusters from SOM. One cluster of the SOM is not included in the matrix and contains observations with missing data cells which were not used in the k-means classification. It is a somewhat alarming drawback of the SOM algorithm that all these observations were mapped into one group which, moreover, contains no complete observations and, hence, has no observations in common with groups produced by complete-case k-means classification. The colour codings should be read as follows: white areas represent cells with no observations, whilst grey areas show cells with some observations in them; the darkest, black, cells correspond to the maximum value of the cross-tabulation matrix. The few points to note about this picture is that the majority of cells are empty and that there is a rather significant number of dark cells which shows that the two algorithms, unsurprisingly, resulted in similar clusterings.

Table 1.4 reports estimates for the model of enterprise training examined in section 1.2. For ease of comparison, selected results from section 1.4.2 are reproduced in the last panel together with the column of cross-sectional (CS) estimates.

It is not surprising that only a few coefficients turn out to be significant as the number of observations in the resulting pseudo-panel is relatively small. However, there is a high degree of concordance between the various pseudo-panel estimates based on different groupings and different regressors included. Hence, most of the comments in section 1.4.2 apply.

What is worth emphasising though, is the effect of foreign ownership which, similarly to earlier pseudo-panel estimates but in drastic contrast to the results of cross-sectional analysis, appears *negative* and significant with its magnitude going as high as 0.93. Whether this result is reassuring or not depends on one's belief in the consistency of pseudo-panel estimates.

1.6 Conclusions

As estimation of a fixed effects model gives researchers the possibility of identifying causal relationships with observational data, the importance of the availability of genuine

Table 1.4: The model of enterprise training estimated with 'self-organised' pseudo-panel

	Coefficient estimates[a] and t-ratios of the linear FE model for training intensity					
Variables	SOM	k-means	Replicated earlier estimates			
	clustering	clustering	Pseudo-panel (sec. 1.4.2)			CS (App. 1.A)
Innovativeness	-0.3633	0.2527	0.2550	-0.0062	0.1203	0.1902
	(-1.18)	(1.00)	(1.97)	(-0.06)	(0.78)	(14.72)
Exports	0.5864	-0.2684	0.1747	-0.0856	0.1771	-0.0271
	(0.93)	(-0.66)	(1.24)	(-1.03)	(1.13)	(-2.26)
Competition from imports	0.1081	0.1743	0.1321	0.0158	0.1523	0.0026
	(0.66)	(1.48)	(2.21)	(0.40)	(2.21)	(0.48)
Skills of available workers	0.0979	0.0206	0.0346	-0.0001	0.0162	-0.0037
	(1.98)	(0.47)	(1.55)	(-0.01)	(0.72)	(-1.44)
Education of firm workforce	-0.7398	-0.2562	-0.3414	-0.5607	-0.2470	-0.0067
	(-2.43)	(-0.96)	(-2.00)	(-3.84)	(-1.27)	(-0.30)
Foreign ownership	-0.9296	-0.3846	-0.3379	-0.0336	-0.2058	0.0546
	(-3.43)	(-1.93)	(-3.64)	(-0.40)	(-2.27)	(5.99)
Performance	-0.0338	-0.0057	-0.0985	-0.0115	-0.0140	0.0106
	(-0.39)	(-0.08)	(-2.17)	(-0.38)	(-0.32)	(2.56)
Monopolisation	0.1022	0.3090	0.1139	0.1869	0.2063	0.0082
	(0.53)	(2.90)	(2.18)	(4.78)	(3.13)	(1.37)
Full-time employment	-0.0722	-0.0497	0.0048	0.0021	-0.0362	0.0063
	(-1.94)	(-1.53)	(0.30)	(0.24)	(-1.60)	(4.80)
Labour regulations	0.1630	0.0816	0.0293	-0.0312	0.0309	-0.0003
	(3.75)	(1.93)	(1.20)	(-1.48)	(1.29)	(-0.11)
Number of obs.	11900	11900	14606	14594	14585	11047
Number of groups	23	44	108	238	107	x

[a] All the reported estimators employed observations weighted by the square root of the corresponding group size. t-ratios are based on standard errors that allow for a serially correlated error variance structure.

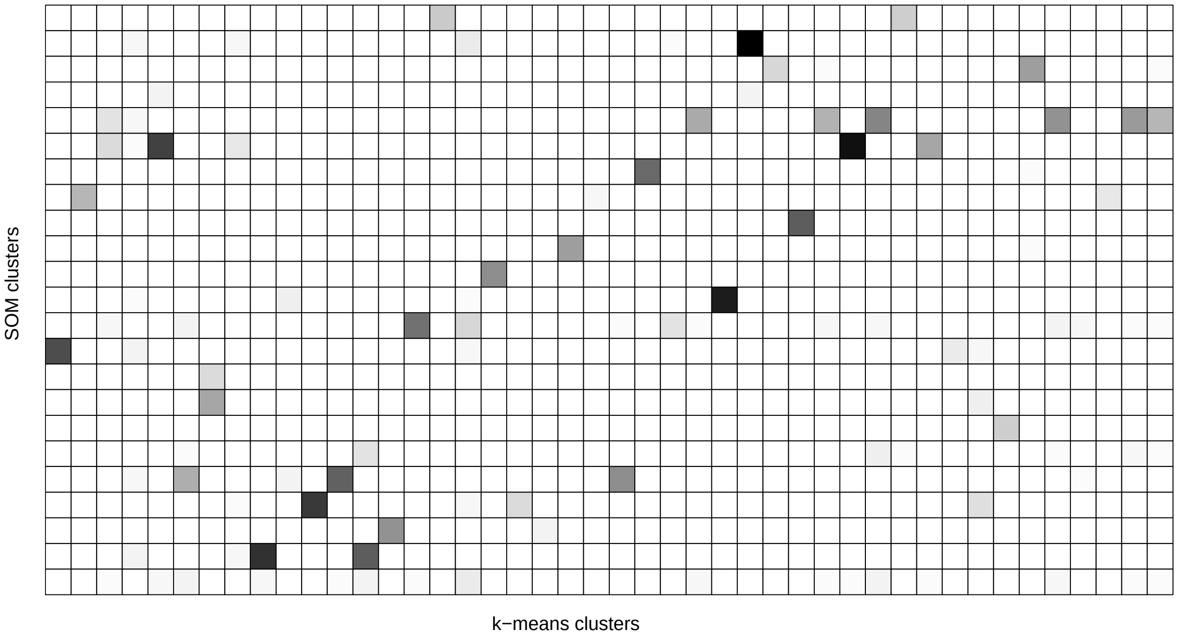

Figure 1.5: Comparison of SOM and k-means clustering

Based on 11900 observations used in k-means clustering. 2706 observations with missing data cells were used in SOM clustering and resulted in a single cluster of their own.

panel data cannot be understated. Yet in some cases when the true panel data is not available, RCS can be used to identify the parameters of interest by averaging individual observations and creating a pseudo-panel of sample group averages. This approach can produce consistent estimators with either the number of groups or the number of individual observations per group asymptotics (the latter assumption is arguably more realistic in empirical applications).

In this paper we have estimated a model of enterprise training and contrasted pseudo-panel estimates with conventional cross-sectional estimates. As the results suggest, the effect of foreign ownership—the main factor of interest in many related studies—turns out to be negative. This contradicts earlier studies that documented a positive *correlation* between foreign ownership and training. However, it should be emphasised that whilst cross-sectional analysis can at best measure correlation (as it is very hard to find a valid set of instrumental variables), the estimates reported for the fixed effects model are capable of revealing *causal* relationships; and it may be the case that the effect of unobserved factors is strong enough to change the signs of coefficients obtained in cross-sectional analysis as compared to (pseudo) panel analysis. It should be kept in mind that estimates are quite sensitive to group design, a problem that is expected to arise if some group designs result in grouping variables being endogenous to the model.

Despite being intuitively appealing, the pseudo-panel approach suffers from a number of complications which may invalidate the results, and the conditions under which estimators are consistent may be too strong or unreasonable. Furthermore, as with the validity of instruments in an exactly identified model, the validity of the grouping approach is untestable and is thus part of the maintained hypothesis.

There are many issues remaining. In particular, identification and efficiency questions, and the relaxation of the assumption of a closed population are important topics that merit further research. The relative efficiency of estimators using genuine panel data versus RCS data as well as the question of whether the identification conditions for RCS data hold can only be investigated with true panel data that include information on potential grouping variables. One attempt to investigate the effect of treating the true panel data as RCS is made in Verbeek and Nijman (1992) that considers an empirical example using Dutch data on household expenditures (a true panel). A within estimator for a synthetic panel created by grouping individuals based on the date-of-birth of the head of household is used. The results suggest that with large enough cohort sizes (say, more than 100 or 200 observations per cohort) the model parameters are identified and the bias of the within estimator may be ignored.

One common problem with survey data is that variables often assume categorical values which may or may not have a natural ordering. Even if a natural ordering of values exists, it may not always be possible to recover the original metric associated with those values, for instance, attitude questions usually recorded with values like ‘disagree’, ‘agree’, ‘strongly agree’, etc., which can have arbitrary distances between them. Even for genuinely continuous variables some kind of rounding and clustering almost always takes place. Despite it being common in the social science literature to assume any variable taking a sufficiently large number (say, more than ten) of distinct values to be continuous, such practice may result in certain complications when one tries to cluster the data. In particular, a clustering algorithm can be misled to group the observations around points corresponding to typical values that such variables take, and hence a spurious structure is effectively imposed on the data. Increasing the dimension of the feature space should, in principle, reduce the risk of false classification, but will often be infeasible as there is usually only a limited number of variables recorded in any given survey. Increasing the precision with which data is recorded is again desirable for statistical analysis but exacerbates problems of limited recall and nonresponse; see e.g. Tourangeau, Rips, and Rasinski (2000) for a detailed account of survey analysis in relation to factors affecting response.

A limitation of many classical clustering algorithms is the assumption that the data space is Euclidean. As this is almost never true with survey data, it would be more realistic to use some general (dis)similarity coefficient instead of Euclidean distance. For instance, a similarity coefficient proposed by Gower (1971) accommodates features of different nature as long as an appropriate (dis)similarity measure is defined for each feature. The use of this coefficient transformed to measure dissimilarities rather than similarities is advocated by Kaufman and Rousseeuw (2005). The resulting dissimilarity matrix can be clustered using relational clustering methods such as those described in Runkler (2007) and Weber (2007).

Appendix 1.A Cross-sectional analysis of the model of enterprise training

This appendix presents results of cross-sectional analysis of the model of enterprise training discussed in section 1.2. A brief summary of the variables included into the model is given in Table 1.5.

Information on the dependent variable is available in two forms, i.e. a binary variable stating whether a firm offers training for a particular type of employees and a continuous variable recording the percentage of employees actually offered training in each category. Two sets of estimates are correspondingly presented using *probit* and ordinary least squares (OLS) estimation. Standard errors are estimated using an empirical distribution function (EDF) bootstrap (often called a *nonparametric bootstrap*); 1,800 bootstrap replications were performed in each case as estimated using the method of Andrews and Buchinsky (2000).

Finally, estimates obtained from the complete-case analysis (CC-analysis), where missing observations are deleted case-wise, are presented alongside estimates obtained using the complete data set with missing observations imputed using the method of multiple imputations by chained equations ('MICE data'). The latter method does not require a multivariate joint distribution assumption and may be used for simultaneous imputation of different types of variables; see e.g. Cameron and Trivedi (2005) and Little (1992).

Table 1.6 reports estimated coefficients, t-ratios and marginal effects for probit estimation of the model describing firms' incidence of training managerial personnel. With two exceptions where coefficients are insignificant, all four regressions give the same directal effect for all reported regressors. This is unsurprising, as it involves the same relationship using different information on the measure of training incidence. It is nonetheless reassuring, showing some robustness for the model. Second, the coefficient estimates change in significance when we move from CC-analysis to regressions with multiply-imputed data; most of the coefficients become more significant with MICE data.

Table 1.7 reports OLS estimates for the five regressions corresponding to five groups of employees as defined above (the column for managers is repeated here for convenience). These estimates were obtained from MICE data and reflect the best information on training available in the dataset.

This study seeks to estimate whether foreign ownership matters for the incidence of enterprise training. As expected, the effect of foreign ownership is significant for training in each of the five regressions considered. Importantly foreign ownership has the biggest

Table 1.5: Summary of variables used in the model

Name	Description
	Dependent variable
Training	A binary variable (training offered or not) and a continuous variable representing the share (%) of employees in each category who received training.
	Factors
Foreign ownership	Percentage of firm assets owned by private foreign company/organisation
Performance	Percentage change in firm's sales since 1998, in real terms
Monopolisation	Dummy for monopolistic/oligopolistic position of a firm
Advanced technology	Dummy subjective characterisation of firm's technology being more advanced than that of the main competitor
Competition from imports	Dummy for subjective importance of competition from imports
Innovativeness	Proxy for innovativeness (weighted index)
Skills of available workers	Characteristics of skills and education of available workforce ('-3' 'major obstacle', '0' 'no obstacle' for the operation and growth of a firm)
Education of firm workforce	Index characterising education of firm's workforce (values between zero and one with a larger value meaning better education)
Exports	Share of firm's sales that are exported (lies between zero and one)
Full-time employment	Number of full-time employees
Part-time employment	Number of part-time employees
Labour regulations	Subjective measure of effects of labour regulations ('-3' 'major obstacle', '0' 'no obstacle' for the operation and growth of a firm)
	Controls
Structure of permanent full-time workforce	Percentage of full-time workers in corresponding group
Legal organisation of the firm	Dummy for privately owned company
Regional dummies	Dummies for country
Industry dummies	Dummies for industry

Table 1.6: Estimation results for the model of training managers

Variables[a]	LS estimates				Probit estimates			
	CC-analysis		MICE data		CC-analysis		MICE data	
Foreign ownership	0.1469	(7.17)	0.1396	(8.07)	0.0052	(7.06) [0.0017]	0.0051	(8.04) [0.0016]
Performance	0.0101	(1.49)	0.0132	(2.19)	0.0008	(2.70) [0.0002]	0.0007	(2.98) [0.0002]
Monopolisation	2.4690	(1.75)	3.4140	(2.78)	0.1530	(2.70) [0.0512]†	0.1642	(3.22) [0.0544]†
Advanced technology	5.0726	(4.23)	4.1816	(3.97)	0.1867	(3.82) [0.0621]†	0.1709	(3.91) [0.0561]†
Competition from imports	1.1571	(1.06)	2.0845	(2.18)	0.0741	(1.55) [0.0243]†	0.0916	(2.15) [0.0297]†
Innovativeness	29.0202	(9.18)	31.0863	(11.36)	1.4056	(11.00) [0.4566]	1.4799	(13.12) [0.4748]
Skills of available workers	-0.5443	(-1.02)	-0.9803	(-2.08)	-0.0329	(-1.40) [-0.0107]	-0.0521	(-2.57) [-0.0167]
Education of firms' workforce	35.1526	(8.18)	31.4749	(8.34)	1.7730	(8.21) [0.5759]	1.6916	(8.92) [0.5427]
Exports	-2.9887	(-1.20)	-4.5116	(-2.43)	0.0609	(0.60) [0.0198]	-0.0390	(-0.48) [-0.0125]
Full-time employment	0.0151	(5.44)	0.0159	(6.60)	0.0010	(9.34) [0.0003]	0.0010	(11.22) [0.0003]
Part-time employment	0.0032	(0.64)	-0.0019	(-0.49)	0.0003	(1.40) [0.0001]	0.0001	(0.45) [0.0000]
Labour regulations	-0.8488	(-1.42)	-0.8154	(-1.53)	-0.0551	(-2.14) [-0.0179]	-0.0574	(-2.50) [-0.0184]
'Private' (dummy)	-1.0047	(-0.68)	-1.6724	(-1.23)	-0.0719	(-1.11) [-0.024]†	-0.1148	(-1.98) [-0.038]†
[Pseudo-]R^2	0.19		0.19		0.18		0.19	
Adjusted R^2	0.18		0.18					
χ^2-statistics	956.5		1290.3		863.8		1431.9	
Prob($\chi^2 > \chi^2_{crit.}$)	(0.0000)		(0.0000)		(0.0000)		(0.0000)	

t-ratios ($t = \hat{\beta}_j/\widehat{se}_{\hat{\beta}_j,B}$) are given in brackets and marginal effects are given in square brackets. † stands for change for dummy variable from 0 to 1.

[a] Coefficient estimates for country dummies, industry dummies, and a variable controlling for the structure of permanent full-time workforce are omitted.

Table 1.7: LS estimates for MICE data

Variables[a]	Employee group				
	Managers	Profes-sionals	Skilled workers	Unskilled workers	Support workers
Foreign ownership	0.1396	0.1081	0.0682	0.0384	0.0611
	(8.07)	(6.47)	(4.24)	(3.07)	(4.44)
Performance	0.0132	0.0183	0.0102	0.0058	0.0075
	(2.19)	(2.90)	(1.76)	(1.35)	(1.65)
Monopolisation	3.4140	2.7648	2.0977	-0.2607	-0.1380
	(2.78)	(2.25)	(1.86)	(-0.33)	(-0.16)
Advanced technology	4.1816	3.0600	4.9981	0.8157	2.3858
	(3.97)	(3.07)	(5.15)	(1.20)	(3.04)
Competition from imports	2.0845	2.0951	1.2449	0.9106	0.6819
	(2.18)	(2.21)	(1.34)	(1.37)	(0.93)
Innovativeness	31.0863	27.4601	27.6085	12.4374	13.0514
	(11.36)	(10.35)	(10.57)	(6.29)	(6.17)
Skills of available workers	-0.9803	-0.5745	-0.8931	0.0007	0.3828
	(-2.08)	(-1.26)	(-1.95)	(0.00)	(1.09)
Education of firms' workforce	31.4749	22.8733	12.6752	1.5240	8.1968
	(8.34)	(5.95)	(3.53)	(0.52)	(2.93)
Exports	-4.5116	-1.0618	-2.1105	-1.4512	-1.9850
	(-2.43)	(-0.56)	(-1.12)	(-1.10)	(-1.41)
Full-time employment	0.0159	0.0129	0.0073	0.0087	0.0088
	(6.60)	(5.57)	(3.33)	(4.82)	(4.75)
Part-time employment	-0.0019	0.0006	0.0017	-0.0000	0.0047
	(-0.49)	(0.17)	(0.47)	(-0.00)	(1.42)
Labour regulations	-0.8154	-1.0385	-0.6863	-0.0897	-0.4864
	(-1.53)	(-2.00)	(-1.34)	(-0.25)	(-1.24)
'Private' (dummy)	-1.6724	-1.4575	-1.5521	-0.5216	0.1623
	(-1.23)	(-1.13)	(-1.24)	(-0.57)	(0.17)
Adjusted R^2	0.18	0.20	0.17	0.10	0.12
χ^2-statistics	956.5	1338.1	1184.2	563.0	702.2
(Prob($\chi^2 > \chi^2_{crit.}$))	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)

t-ratios ($t = \hat{\beta}_j / \widehat{se}_{\hat{\beta}_j,B}$) are given in brackets.

[a] Coefficient estimates for country dummies, industry dummies, and a variable controlling for the structure of permanent full-time workforce are omitted.

effect on how managers are trained. The estimated coefficient implies that a 1% increase in the foreign ownership of a firm generates 0.14% more managers receiving training. The effect on professional workers' training is lower but still substantial—about 0.11%, while for other groups of employees the effects of foreign ownership are much smaller.

Apart from the effect of foreign ownership, the model captures several other interesting dependencies. For example, two highly significant factors are innovativeness and firm size, measured by the number of full-time workers, while the number of part-time workers appears to have no effect on training. If the index of innovativeness changes by one unit (i.e., moving from absolutely non-innovative to highly innovative), the share of managers who receive training increases by 31.1%. This effect is significant and maintains a high amplitude for all other groups of workers as well. This result is in line with the hypothesis of innovation-driven training.

Competition from imports is important only for training managers and professional workers, but the magnitude of the effect is quite low—about 2.1%. At the same time, monopolies are more likely to provide training to their managers and professional workers (and, with marginal significance, to skilled workers), contrary to the expected competition effect. One possible explanation may be that in transition economies firms with only a few competitors are much stronger and thus can actually engage in personnel training, while firms operating in emerging competitive markets remain very fragile and simply unable to develop long-run training strategies.

Another interesting feature of the model is that it is able to capture the effect of labour-force quality proxied by education of the existing personnel and by the subjective measure of skills of workers available in the labour market. There are two different effects of labour-force quality. First, the general skills of the country's labour force are negatively related to training. Secondly, as expected, the quality of existing personnel relates positively to the amount of training (the effect is significant for all categories of employees except unskilled workers). Finally, three factors appear to have generally no effect on training: a firm's trade orientation, labour regulations and—surprisingly—the legal organization of a firm.

To provide a benchmark, Table 1.8 reports estimation results using the two cross-section survey datasets (BEEPS–II and BEEPS–III). Estimates are given for each dataset taken separately and for the pooled dataset. The four sets of OLS estimates correspond to the four dependent variables: three measuring the share of employees in the respective category who received training and one variable measuring an overall training intensity (a weighted index of the previous three variables).

Table 1.8: Estimation results for the model of enterprise training: Cross-section analysis

Variables[a]	Dataset/Dependent variable											
	BEEPS–II				BEEPS–III				Pooled			
	Skilled workers	Unskilled workers	Support workers	Overall index	Skilled workers	Unskilled workers	Support workers	Overall index	Skilled workers	Unskilled workers	Support workers	Overall index
Innovativeness	0.2630	0.1051	0.1200	0.1552	0.3381	0.1959	0.2242	0.2150	0.3085	0.1510	0.1663	0.1902
	(11.09)	(6.13)	(6.40)	(9.12)	(12.45)	(6.40)	(7.18)	(11.15)	(17.09)	(9.20)	(9.69)	(14.72)
Exports	-0.0350	-0.0233	-0.0372	-0.0286	-0.0485	-0.0551	-0.0123	-0.0292	-0.0380	-0.0302	-0.0197	-0.0271
	(-1.60)	(-1.47)	(-2.18)	(-1.82)	(-1.96)	(-2.02)	(-0.45)	(-1.65)	(-2.28)	(-2.02)	(-1.28)	(-2.26)
Competition	0.0112	0.0081	0.0062	-0.0001	0.0131	0.0068	-0.0163	0.0027	0.0141	0.0090	-0.0011	0.0026
from imports	(1.14)	(1.15)	(0.80)	(-0.01)	(1.19)	(0.51)	(-1.25)	(0.34)	(1.89)	(1.31)	(-0.15)	(0.48)
Skills of available	-0.0094	0.0010	0.0042	-0.0038	-0.0088	-0.0057	-0.0108	-0.0038	-0.0087	-0.0018	-0.0036	-0.0037
workers	(-1.96)	(0.30)	(1.11)	(-1.10)	(-1.65)	(-0.92)	(-1.70)	(-1.00)	(-2.40)	(-0.53)	(-1.03)	(-1.44)
Education of firm	0.1063	-0.0156	0.1259	-0.0157	0.3244	-0.0373	0.2712	-0.0137	0.2067	-0.0064	0.1951	-0.0067
workforce	(2.72)	(-0.56)	(4.12)	(-0.56)	(6.42)	(-0.60)	(4.68)	(-0.40)	(6.57)	(-0.22)	(6.63)	(-0.30)
Foreign	0.0689	0.0412	0.0711	0.0547	0.1151	0.0598	0.0722	0.0586	0.0870	0.0414	0.0682	0.0546
ownership	(4.33)	(3.60)	(5.77)	(4.86)	(5.55)	(2.63)	(3.36)	(4.07)	(6.72)	(3.60)	(5.90)	(5.99)
Performance	0.0121	0.0088	0.0095	0.0090	0.0155	0.0609	0.0419	0.0197	0.0114	0.0172	0.0129	0.0106
	(2.01)	(1.98)	(2.02)	(2.11)	(1.17)	(4.04)	(2.63)	(2.12)	(1.94)	(3.33)	(2.45)	(2.56)
Monopolisation	0.0241	0.0042	-0.0033	0.0098	0.0131	0.0404	0.0156	0.0082	0.0191	0.0187	0.0010	0.0082
	(2.04)	(0.50)	(-0.36)	(1.15)	(1.08)	(2.94)	(1.13)	(0.96)	(2.25)	(2.42)	(0.13)	(1.37)
‘Private’	-0.0001	-0.0043	0.0057	0.0097	-0.0146	-0.0107	-0.0074	-0.0058	-0.0101	-0.0071	0.0009	0.0021
(dummy)	(-0.01)	(-0.45)	(0.54)	(1.00)	(-0.81)	(-0.57)	(-0.39)	(-0.46)	(-0.91)	(-0.75)	(0.09)	(0.27)
Full-time	0.0082	0.0089	0.0096	0.0074	0.0009	0.0014	0.0051	0.0046	0.0051	0.0057	0.0082	0.0063
employment	(3.65)	(5.63)	(5.44)	(4.55)	(0.32)	(0.49)	(1.76)	(2.28)	(2.84)	(3.75)	(5.07)	(4.80)
Labour	-0.0048	-0.0033	-0.0051	-0.0033	-0.0001	0.0144	0.0055	-0.0001	-0.0012	0.0053	0.0015	-0.0003
regulations	(-0.89)	(-0.84)	(-1.20)	(-0.85)	(-0.01)	(2.12)	(0.81)	(-0.03)	(-0.32)	(1.45)	(0.39)	(-0.11)
Number of obs.	4812	3982	4046	5012	5500	2931	3137	6035	10312	6913	7183	11047
Adjusted R^2	0.1592	0.0770	0.0864	0.1439	0.1551	0.1383	0.1151	0.1406	0.1647	0.1272	0.1233	0.1469
F-statistics	21.6964	8.5438	9.6974	20.1502	23.9379	11.6872	10.2747	23.4446	46.1647	23.3933	23.4393	43.2747

[a] Coefficient estimates for country dummies and industry dummies are omitted.

Appendix 1.B Overview of clustering techniques

Cluster analysis[10] has been applied in many disparate areas including astronomy, taxonomy, biology, psychology, linguistics, cryptology and archeology. Originally motivated by the need to summarise large and possibly multivariate datasets, it has since been applied in other areas including pattern recognition and market research. The basic idea is as follows: having a (large) set of objects each being described by a number of characteristics, we ask whether there exists a certain (smaller than the number of objects) number of groups, often called classes or clusters, such that objects within each group exhibit 'similarity' whilst objects belonging to different groups are 'dissimilar' or, (Gordon, 1981), the two possible desiderata for a cluster are internal cohesion and external isolation.

It is important to note the distinction between clustering and assignment or identification (Gordon, 1981), the latter referring to a procedure whereby an object has to be assigned to one from a *known* number of existing classes. In clustering the number of classes is *unknown* a priori. Hence, the aim is to *uncover* the structure of the data.

From a statistical point of view, classification can be regarded as methods for the exploratory analysis of multivariate data. Classification methods can broadly be divided into partitioning, hierarchical, and clumping (allowing overlapping groups) methods which together constitute a group of clustering methods, and geometrical methods. The latter are mostly suitable for visualisation of complex datasets and preliminary analysis; Chambers and Kleiner (1982) discuss several techniques for graphical analysis of multivariate data. Development of most 'classical' clustering methods dates back to the early fifties; Hartigan (1975) describes a number of clustering techniques developed by early seventies and gives precise algorithms for all the methods considered; Breiman, Friedman, Olshen, and Stone (1984) discusses the tree methodology. A neat account of hierarchical clustering methods—of which perhaps the most widely known one is the single linkage or nearest neighbour method—can be found in Everitt et al. (2001); see also Banks, House, McMorris, Arabie, and Gaul (2004) for a recent account of the new methods in clustering.

At its basic level, classification may be seen as simply a method to describe a large set of objects by means of allocating them to a smaller set of homogeneous groups; however, one can often be interested in classification as a tool for revealing more fundamental

[10] Classification or cluster analysis is known under different names in different fields of science: it is referred to as numerical taxonomy in biology, Q-analysis in psychology, unsupervised pattern recognition in artificial intelligence (on the contrary, discrimination and assignment methods are known under the term supervised learning), or segmentation in market research; see e.g. Everitt, Landau, and Leese (2001).

properties of the data, understanding causalities and dependencies amongst variables. For a univariate data set, it is natural to argue that if the data have a distribution which is unimodal, then they correspond to a homogeneous unclustered population; multimodal distributions, on the contrary, are thought of as characterising a clustered population with the number of modes representing the number of clusters. In a one- or two-variate case it is then possible to use graphical methods to reveal potential clusters in the data.

A significant improvement upon hierarchical clustering methods is optimisation clustering techniques. These produce a partition of the set of objects by optimising some prespecified criterion, and hence, result in an 'optimal' partition. The criterion is normally chosen to measure the 'adequacy' of a partition with a given number of groups, and the optimal number of groups is then delivered by the optimisation procedure; typically this either minimises the lack of homogeneity or maximises the separation of groups. A number of different such criteria are available. One popular criterion uses the decomposition of the dispersion matrix into the sum of within-group and between-group dispersions and minimises the trace of the within-group dispersion matrix. Other suggested procedures are the minimisation of the determinant of the within-group dispersion matrix and maximisation of the trace of the product of the between-group dispersion matrix and the inverse of the within-group dispersion matrix; several modifications drawing on these ideas have also been proposed.

In principle, optimisation of any chosen criterion should be taken over all possible partitions of a given set of objects into any possible number of clusters. However, this is extremely computationally intensive, and various algorithms exist that search for the optimum value over a small subset of all such partitions; hill-climbing algorithms provide one example. One of the earliest but still very popular hill-climbing algorithm is the so-called *k-means* algorithm[11] which is widely available in classification software. k-means algorithm defines a partition of the feature space by the Dirichlet tessellation of the cluster representatives (also known as Voronoi tessellation, especially if the space in question is $\mathbb{R}^2$).

More recent methods for cluster analysis include parametric models based on finite mixture densities, the application of which involves estimating parameters of the assumed mixture and implied probabilities of cluster membership, density search methods based on identifying the most 'dense' regions in the input space, and a number of methods that allow for overlapping clusters; see e.g. Everitt et al. (2001) for an overview of these methods.

[11] k-means is an L_2 method; an L_1 sibling of k-means, the k-medoid method, is more robust to outliers as is usual with L_1 methods.

Fuzzy clustering algorithms and artificial neural networks methods deserve a special mention. Fuzzy clustering can be thought of as a generalisation of partitioning which allows for some ambiguity in the data. In particular, for each object the degree of belonging to each group is estimated by membership coefficients ranging from zero to one, so that, for instance, an object can belong 70% to one group and 30% to some other group, etc. Fuzzy clustering thus provides much more detailed information about the structure of the data set than crisp clustering does, and can be particularly suitable for estimation with RCS where the vectors of group membership coefficients can be used as instruments instead of a set of zero-one membership indicators. However, at their present level of development, fuzzy clustering algorithms require the number of groups to be known a priori, and an external procedure is required to determine this number.

Despite the bulk of applications of neural networks being for supervised learning tasks, a number of unsupervised methods exists; see Ripley (1996) for a comprehensive account of neural methods. One successful example is the Kohonen Self-Organising Feature Map (SOM) algorithm, which is a special type of clustering algorithm that assigns objects to clusters arranged on a regular one- or two-dimensional grid; see Kohonen (1997). The SOM algorithm can be regarded as 'a spatially smooth version of k-means', and a batch version of SOM is an adaptation of the latter (Ripley, 1996). Little is known, however, about theoretical properties of the SOM algorithm: although simulation studies report organisation and convergence of the SOM, a proof exists only for the simplest one-dimensional case; see Cottrell, Fort, and Pagès (1998) for a recent review.

Chapter 2

Kernel density estimation for time series data

A time-varying probability density function, or the corresponding cumulative distribution function, may be estimated nonparametrically using a kernel function and weighting the observations using schemes derived from time series modelling. The parameters, including the bandwidth, may be estimated by maximum likelihood or cross-validation. Diagnostic checks may be carried out directly on residuals given by the predictive cumulative distribution function. Since tracking the distribution is only viable if it changes relatively slowly, the technique may need to be combined with a filter for scale and/or location. The methods are applied to data on the NASDAQ index.

Contents

2.1 Introduction

A probability density function (pdf), or the corresponding cumulative distribution function (cdf), may be estimated nonparametrically by using kernel function methods. Standard kernel density estimators (KDE) have been concerned with estimation of the stationary marginal distribution. The focus of this chapter is to allow the density to change over time.

If it is assumed that the density is 'gradually' changing with time, but the change is 'slow', then for a certain (small) period of time the observations can be thought of as having a common distribution. The changing density can then be analysed by passing a window of an appropriate size over which the density is assumed to be the same; see Hall and Patil (1994, p. 1509).

Analysing evolving densities by moving blocks of data is of course equivalent to weighting observations over time using rectangular weighting function. By weighting we entertain the possibility of a density potentially changing over time. Beside rectangular, many other weighing schemes can be contemplated; for example, triangular (linear) or quadratic weights are very simple to construct and may have their own appeal. The crucial question, however, is whether there exists a weighting scheme which is optimal in some way and, if it exists, whether there is a way to find it. This remains an open question for future research.

For the purposes of this chapter, we are going to concentrate on one of the simplest time series weighting schemes, viz. the exponentially weighted moving average (EWMA). EWMA is widely used to estimate the level of a series and hence future observations. A similar scheme may be used to estimate the conditional variance, e.g. 'Riskmetrics', but a firmer theoretical underpinning is the integrated generalised autoregressive heteroscedasticity (GARCH) model. Other models imply other weighting schemes and hence other recursions for updating parameter estimates that evolve over time. For example, changing growth rates and seasonal patterns can easily be accommodated. Recursions are usually combined with an assumption about the form of the one-step ahead predictive distribution. As a result a likelihood function can be constructed and then maximized with respect to the unknown parameters in the model. Once a model has been fitted, the one-step ahead predictions may be subjected to diagnostic checking by reference to the predictive distribution. Most commonly the predictive distribution is Gaussian and tests are carried out on standardised residuals.

This chapter demonstrates that similar ideas carry over to the nonparametric estimation of a time-varying density or distribution function. Not only can updating be carried out recursively, but a likelihood function can be constructed from *predictive* distributions.

Hence dynamic parameters, such as the discount parameter in the EWMA, may be estimated by maximum predictive likelihood (MPL). Furthermore the dynamic specification may be checked by using the residuals given by the predictive cumulative distribution function. Diagnostics are those appropriate for the probability integral transform, as described in Diebold, Gunther, and Tay (1998).

Time varying quantiles may be extracted from the cumulative distribution function. In the time-invariant case there are efficiency gains for estimating quantiles this way as compared with simply using the sample quantiles calculated from the order statistics, but the gains may be small; see Sheather and Marron (1990). There has been considerable interest in the last few years in estimating changing quantiles. The conditional autoregressive value at risk (CAViaR) approach of Engle and Manganelli (2004) models quantiles in terms of functions of past observations. De Rossi and Harvey (2009) adopt a different method, based on ideas from signal extraction and using only indicator variables. One drawback of the CAViaR approach is that, as pointed out by Gourieroux and Jasiak (2008), the quantiles may cross. This problem is circumvented if the cumulative distribution function is used.

Section 2.2 discusses linear filters and in section 2.3 filters for estimating time-varying densities are developed. Attention is focussed on EWMA and a stable filter with an extra parameter. We also explain how to estimate densities using a two-sided filter that is the equivalent of smoothing, or signal extraction, in time series and how to construct algorithms for weighting schemes associated with more general time series models. Ways in which bandwidth selection methods designed for time-invariant distributions may be adapted to deal with changing distributions are explored and estimation by maximum predictive likelihood and cross-validation is discussed. Section 2.4 describes diagnostic checking with probability integral transforms of the predictions. Section 2.5 discusses time-varying quantiles. Section 2.6 applies the methods to the NASDAQ index, while section 2.6.3 compares the results using EWMA with rectangular weighting scheme. The last section concludes.

2.2 Filters

A linear filter is a scheme for weighting current and past observations in order to estimate an unobserved component or a future value of the series. Thus an estimator of the level at time t could be written as

$$m_t = \sum_{i=0}^{t-1} w_{t,i} y_{t-i}, \qquad t = 1, \ldots, T,$$

where $w_{t,i}$ are weights. One way of putting more weight on the most recent observations is to let the weights decline exponentially. If t is large then exponential weighting (EW) sets $w_{t,i} = (1-\omega)\omega^i$, $i = 0, 1, 2, \ldots$, where ω is a discount parameter in the range $0 \leq \omega < 1$. (The weights sum to unity in the limit as $t \to \infty$). The attraction of EW is that estimates can be updated by the recursion

$$m_t = \omega m_{t-1} + (1-\omega) y_t, \qquad t = 1, \ldots, T$$

with $m_0 = 0$ or $m_1 = y_1$. This filter can also be expressed in terms of the one step ahead prediction, with m_t replaced by $m_{t+1|t}$, that is $\hat{y}_{t+1|t} = m_{t+1|t}$. Thus the recursion can be written

$$m_{t+1|t} = m_{t|t-1} + (1-\omega)\nu_t, \qquad t = 1, \ldots, T, \tag{2.1}$$

where $\nu_t = y_t - \hat{y}_{t|t-1}$ is the one-step ahead prediction error or innovation.

The EW filter may be rationalised as the steady-state solution to an unobserved components model consisting of a random walk plus noise. The model, known as the local level model, is defined by

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim NID(0, \sigma_\varepsilon^2), \quad t = 1, \ldots, T, \\ \mu_t &= \mu_{t-1} + \eta_t, & \eta_t &\sim NID(0, \sigma_\eta^2). \end{aligned} \tag{2.2}$$

where the disturbances ε_t and η_t are mutually independent and the notation $NID\,(0, \sigma^2)$ denotes normally and independently distributed with mean zero and variance σ^2. The Kalman filter for the optimal estimator of μ_t based on information at time t is

$$m_{t+1|t} = (1-k_t)\, m_{t|t-1} + k_t y_t, \quad t = 1, \ldots, T, \tag{2.3}$$

where $k_t = p_{t|t-1} / \left(p_{t|t-1} + 1\right)$ is the gain, and

$$p_{t+1|t} = p_{t|t-1} - \left[p_{t|t-1}^2 / \left(1 + p_{t|t-1}\right)\right] + q, \quad t = 1, \ldots, T,$$

where $q = \sigma_\eta^2 / \sigma_\varepsilon^2$ is the signal-to-noise ratio; see Harvey (2006, 1989, p.175). The MSE of $m_{t+1|t}$ is $\sigma_\varepsilon^2 p_{t+1|t}$. The filter can be initialised using a diffuse prior, i.e. setting $m_{1|0} = 0$. Then as $p_{1|0} \to \infty$, $k_1 \to 1$ and hence $m_{2|1} = y_1$ and $p_{2|1} = 1 + q$. The steady-state solution for k_t is $1 - \omega$, where the parameter ω is a monotonic function of $q = \sigma_\eta^2 / \sigma_\varepsilon^2$. The likelihood function may be constructed from the one-step ahead prediction errors and maximised with respect to ω.

Smoothed estimates for the Gaussian local level model (2.2) can be computed by saving the innovations and Kalman gains from the filter (2.3) and using them in the backward recursions

$$r_{t-1} = (1 - k_t)r_t + (1 - k_t)\nu_t, \qquad t = T, \ldots, 2,$$

where $\nu_t = y_t - m_{t|t-1}$ and $r_T = 0$, and

$$\begin{aligned} m_{t|T} &= m_{t|t-1} + p_{t|t-1}r_{t-1}, \qquad t = 1, \ldots, T, \\ &= m_{t|t-1} + k_t(r_t + \nu_t) \end{aligned}$$

Since $r_0 = (1-k_1)r_1+(1-k_1)\nu_1$, initializing with a diffuse prior gives $m_{1|T} = (p_{1|0}/(p_{1|0}+1))(r_1 + y_1)$ which tends to $r_1 + y_1$ as $p_{1|0}$ approaches to infinity. The following forward recursion can also be used

$$m_{t+1|T} = m_{t|T} + qr_t, \qquad t = 1, \ldots, T - 1,$$

with $m_{1|T} = r_1 + y_1$; see Koopman (1993).

The weights implicitly used in the smoother, that is, the weights in

$$m_{t|T} = \sum_{t=1}^{T} w_{t,i}y_i, \quad t = 1, \ldots, T,$$

may be computed using the algorithm of Koopman and Harvey (2003).

In the middle of a semi-infinite sample, the weights decline symmetrically and exponentially (Harvey and De Rossi, 2006, eq. 2.13), viz.

$$w_{t,i} \approx \frac{1 - \omega}{1 + \omega}\omega^{|t-i|}, \qquad i = 1, \ldots, T. \tag{2.4}$$

The weights in (2.4) do not sum to one in finite samples (cf. equation (2.15)) but provide a good approximation if both t and T are large. Although these formulae are not used in our computations, they are useful in showing the nature of the weighting patterns.

The random walk in (2.2) may be replaced by a stationary first-order autoregressive process. More complex models, perhaps with slopes and seasonals, may be set up and the appropriate filters derived by putting the model in state space form. Again the likelihood function may be constructed from the one-step ahead prediction errors given by the Kalman filter and the implicit weights for filtering and smoothing obtained from the algorithm of Koopman and Harvey (2003).

A nonlinear class of models may be constructed by applying the linear filters obtained from unobserved component models to transformations of the observations that reflect quantities of interest. For example, if the mean is fixed at zero, but the variance changes we might consider the filter

$$\sigma^2_{t+1|t} = (1-\omega)y_t^2 + \omega\sigma^2_{t|t-1} = \sigma^2_{t|t-1} + (1-\omega)(y_t^2 - \sigma^2_{t|t-1}), \quad t = 1, \ldots, T,$$

where the notation $\sigma^2_{t+1|t}$ accords with that used by Andersen, Bollerslev, Christoffersen, and Diebold (2006) for the variance in a GARCH model. This scheme is an EWMA in squares, with $y_t^2 - \sigma^2_{t|t-1}$ playing a similar role to the innovation in (2.1). It corresponds to integrated GARCH, where the predictive distribution in the Gaussian case is $y_t \mid Y_{t-1} \sim N(0, \sigma^2_{t|t-1})$. The more general filter

$$\sigma^2_{t+1|t} = (1-\omega^* - \omega)\sigma^2 + \omega^* y_t^2 + \omega\sigma^2_{t|t-1}, \quad t = 1, \ldots, T,$$

is stable when $\omega^* + \omega < 1$ and hence is able to generate a stationary series. Estimation may be simplified by setting σ^2 equal to the (unconditional) variance in the sample; this is known as 'variance targeting', as in Laurent (2007, p. 25).

If the above filtering schemes are viewed as approximations to an unobserved variance, the smoother that would correspond to the filter in a linear unobserved components model may be useful as a descriptive device.

The next section shows how filters may be applied to the whole distribution, rather than to selected moments.

2.3 Dynamic kernel density estimation

Using a sample of T observations drawn from a distribution with cdf $F(y)$ with a corresponding pdf $f(y)$, a kernel estimator of $f(y)$ at point y is given by

$$\hat{f}_T(y) = \frac{1}{Th}\sum_{i=1}^{T} K\left(\frac{y - y_i}{h}\right), \tag{2.5}$$

where $K(\cdot)$ is the kernel and h the bandwidth. The kernel, $K(\cdot)$, is a bounded pdf which is symmetric about the origin; see also discussion in Chapter 3.

The kernel estimator of the cumulative distribution function is given by

$$\widehat{F}_T(y) = \frac{1}{T}\sum_{i=1}^{T} H\left(\frac{y - y_i}{h}\right),$$

where $H(\cdot)$ is a kernel which now takes the form of a cdf. A kernel of this form may be obtained by integrating the kernel in (2.5).

2.3.1 Filtering and smoothing

In order to estimate a time varying density, a weighting scheme may be introduced into the kernel estimator (2.5), that is,

$$\hat{f}_t(y) = \frac{1}{h}\sum_{i=1}^{t} K\left(\frac{y-y_i}{h}\right) w_{t,i}, \quad t = 1, \dots, T, \tag{2.6}$$

while, for the distribution function,

$$\widehat{F}_t(y) = \sum_{i=1}^{t} H\left(\frac{y-y_i}{h}\right) w_{t,i}. \tag{2.7}$$

In both cases, $\sum_{i=1}^{t} w_{t,i} = 1$, $t = 1, \dots, T$. The weights, $w_{t,i}$, $i = 1, \dots, t$, $t = 1, \dots, T$, may change over time, although in the steady-state, $w_{t,i} = w_{t-i}$.

Similarly for smoothing

$$\hat{f}_{t|T}(y) = \frac{1}{h}\sum_{i=1}^{T} K\left(\frac{y-y_i}{h}\right) w_{t,i}, \quad t = 1, \dots, T,$$

and

$$\widehat{F}_{t|T}(y) = \sum_{i=1}^{T} H\left(\frac{y-y_i}{h}\right) w_{t,i}, \tag{2.8}$$

with $\sum_{i=1}^{T} w_{t,i} = 1$, $t = 1, \dots, T$.

2.3.2 Recursions

Simple exponential weighting gives recursions similar to those of section 2.2. Thus for the cdf

$$\widehat{F}_t(y) = \omega\widehat{F}_{t-1}(y) + (1-\omega)H\left(\frac{y-y_t}{h}\right), \quad t = 1, \dots, T.$$

Schemes of this kind are not new; see, for example, Wegman and Davies (1979).

The above recursion can be re-written with $\widehat{F}_{t+1|t}(y)$ and $\widehat{F}_{t|t-1}(y)$ replacing $\widehat{F}_t(y)$ and $\widehat{F}_{t-1}(y)$ respectively. A simple re-arrangement then gives

$$\widehat{F}_{t+1|t}(y) = \widehat{F}_{t|t-1}(y) + (1-\omega)V_t(y), \quad 0 \leq \omega < 1, \quad t = 1, \dots, T,$$

where

$$V_t(y) = H\left(\frac{y - y_t}{h}\right) - \widehat{F}_{t|t-1}(y) \tag{2.9}$$

plays a similar role to the innovation[1] in (2.1). However, $V_t(y) < 0$ when $y_t > y$. Note also that $-\widehat{F}_{t|t-1}(y) \leq V_t(y) \leq 1 - \widehat{F}_{t|t-1}(y)$.

An analogous recursion can be written down for the pdf. To be specific

$$\widehat{f}_{t+1|t}(y) = \widehat{f}_{t|t-1}(y) + (1-\omega)\nu_t(y), \quad 0 \leq \omega < 1, \quad t = 1, \ldots, T,$$

where the innovation is

$$\nu_t(y) = \frac{1}{h}K\left(\frac{y - y_t}{h}\right) - \widehat{f}_{t|t-1}(y), \tag{2.10}$$

with $-\widehat{f}_{t|t-1}(y) \leq \nu_t(y) \leq h^{-1}K(0)$.

The filter can be initialized with $\hat{f}_{1|0}(y) = 0$ and, in order to ensure that the weights discounting past observations sum to unity, ω may be set to $1 - k_t$, where k_t is defined in (2.3), until such time, $t = m$, as the filter is deemed to have converged. Alternatively $\hat{f}_{m+1|m}(y)$ may be computed directly from (2.6). The cdf recursion for $\widehat{F}_{t+1|t}(y)$ may be similarly initialized from the first m observations.

The stable filter is

$$\widehat{F}_{t+1|t}(y) = (1 - \omega^* - \omega)\overline{F}(y) + \omega^* H\left(\frac{y - y_t}{h}\right) + \omega\widehat{F}_{t|t-1}(y), \quad t = 1, \ldots, T, \tag{2.11}$$

where $\overline{F}(y)$ is the unconditional kernel density for the whole sample. Setting the initial condition as $\widehat{F}_{1|0}(y) = \overline{F}(y)$ means that the weight attached to $\overline{F}(y)$ at time t is $(1-\omega^*)$, and it gradually tends to $(1 - \omega^* - \omega)$. We can also write

$$\widehat{F}_{t+1|t}(y) = (1 - \omega^* - \omega)\overline{F}(y) + (\omega^* + \omega)\widehat{F}_{t|t-1}(y) + \omega^* V_t, \quad t = 1, \ldots, T.$$

More complex weighting schemes, derived from unobserved components models, may also be adopted. For example an integrated random walk plus trend yields a cubic spline with the Kalman filter reduced to a single equation recursion which for the cdf is

$$\widehat{F}_{t+1|t}(y) = 2\widehat{F}_{t|t-1}(y) - \widehat{F}_{t-1|t-2}(y) + k_1\omega^* H\left(\frac{y - y_t}{h}\right) + k_2\omega^* H\left(\frac{y - y_{t-1}}{h}\right),$$

where k_1 and k_2 are parameters that depend on a signal-to-noise ratio in the original

[1] In a Gaussian model, $H(y_t) = y_t$ and $\widehat{F}_{t|t-1}(y) = \widehat{y}_{t|t-1}$. The only impact is on location and ν_t is a scalar.

unobserved components model.

Finally, other weighting schemes can be used that are not necessarily motivated in terms of an underlying model. The simplest scheme, perhaps, is the one using rectangular weights (i.e. analysing moving blocks of data); this is illustrated in section 2.6.3.

2.3.3 Estimation

The recursive nature of the filter leads naturally to maximum predictive likelihood (MPL) estimation of the bandwidth, h, and any parameters governing the dynamics, such as the discount factor, ω, in exponential weighting. The predictive log-likelihood function, normalized by the sample size, is

$$\begin{aligned}\ell(\omega,h) &= \frac{1}{T-m}\sum_{t=m}^{T-1}\ln \hat{f}_{t+1|t}(y_{t+1}) \\ &= \frac{1}{T-m}\sum_{t=m}^{T-1}\ln\left[\frac{1}{h}\sum_{i=1}^{t}K\left(\frac{y_{t+1}-y_i}{h}\right)w_{t,i}(\omega)\right],\end{aligned} \tag{2.12}$$

where $w_{t,i}(\omega)$ are the weights, which may be obtained as described in section 2.2, and m is some preset number of observations used to initialise the procedure. The value of m will depend on the sample at hand, but it may not be unreasonable to suggest setting $m = 50$ or 100 if the sample size is large. The main consideration is that the predictions should be meaningful.

The predictive log-likelihood (2.12) can be maximized subject to $\omega \in (0,1]$ and $h > 0$ using constrained maximization with numerical derivatives obtained via finite differencing. Using a non-negative kernel with unbounded support, such as a Gaussian kernel, theoretically guarantees that $\widehat{f}_{t+1|t}(y_{t+1}) > 0$ for all $t = m, \dots, T-1$. A problem arises when the density is evaluated at outlier points for which the estimate is numerically zero. In these cases $\widehat{f}_{t+1|t}(\cdot)$ can be set equal to a very small positive number.

From a theoretical point of view, it is interesting to note that as in a linear Gaussian model, such as (2.2), the predictive likelihood can be written in terms of the innovations since, from (2.10), $\widehat{f}_{t+1|t}(y_{t+1}) = \widehat{f}_{t|t-1}(y_{t+1}) + (1-\omega)\nu_t(y_{t+1})$ for $t = m, \dots, T-1$. Thus, instead of re-computing the density estimate at each t using the data up to $t-1$ inclusive, the recursive formulae given in section 2.3 can, in principle, be used. However, in order to evaluate the log-likelihood (2.12), the grid for the recursion will need to include all the sample values, $y_1, \dots, y_T$.

For smoothing, the parameters can be estimated by maximizing the likelihood cross-validation (CV) criterion

$$CV(\omega, h) = \frac{1}{T}\sum_{t=1}^{T} \ln \widehat{f}_{(-t)|T}(y_t) = \frac{1}{T}\sum_{t=1}^{T} \ln \left[\frac{1}{h} \sum_{\substack{i=1 \\ i \neq t}}^{T} K\left(\frac{y_t - y_i}{h}\right) w_{t,T,i}(\omega) \right], \tag{2.13}$$

where $w_{t,T,i}(\omega)$ is given by a two-sided smoothing filter such as (2.4).

Alternatively, one can simply use the same parameters as for filtering.

The number of parameters to be estimated can be reduced by setting the bandwidth according to a rule of thumb, $h = cT^{-1/5}$, where the constant c depends on the spread of the data[2] and T is set equal to the effective sample size[3], $T(\omega)$, a function of ω only. In this case the likelihood and CV criterion are maximized only with respect to ω. In the steady-state of the local level model, the mean square error (MSE) of the contemporaneous filtered estimator, m_t, of the level is $\sigma_\varepsilon^2(1-\omega)$. If the level were fixed, the MSE of the sample mean would be σ_ε^2/T. This suggests an effective sample size for filtering of $T(\omega) = 1/(1-\omega)$. For smoothing the suggestion is $T(\omega) = (1+\omega)/(1-\omega) \approx 2/(1-\omega)$, provided that t is not too close to the beginning or end of the sample. Thus when the bandwidth selection criterion is proportional to $T^{-1/5}$, the bandwidth for filtering will be larger by a factor of approximately $2^{1/5} = 1.15$.

In our examples, the values of the bandwidth chosen by maximising the predictive likelihood or the CV criterion were usually close to the normal reference rule bandwidth with the effective sample size, $T(\omega) = 1\,/(1-\omega)$, in place of T.

[2] The constant in the asymptotically optimal bandwidth, $h = cT^{-1/5}$, depends on the kernel and the curvature of the true density; see equation (3.4) and discussion in section 3.2. As the true density is unknown, the idea behind the rule of thumb (or plug-in) bandwidth is to construct a simple rule for choosing c which performs well in practice. For instance, if the kernel is the Gaussian density, and the underlying distribution is normal with variance σ^2, the constant in the asymptotically optimal bandwidth is $c = 1.06\sigma$, which gives the normal reference rule bandwidth $h = 1.06\sigma T^{-1/5}$. If the density is close to normal, this bandwidth usually perform well, but often results in oversmoothing; see e.g. Jones, Marron, and Sheather (1996). In the presence of outliers, a bandwidth choice based on robust measures of spread may perform better. One popular choice is $c = 1.06 \min\left(\hat{\sigma}, \widehat{IQR}/1.34\right)$, where $\widehat{IQR}$ is the sample interquartile range; see Silverman (1986). See *inter alia* Wand and Jones (1995), Silverman (1986), Pagan and Ullah (1999), Li and Racine (2007) and Sheather (2004) for a general discussion of bandwidth selection.

[3] Effective sample size is a measure of the 'weighting effect'. It is obtained by comparing the variances of the weighted and unweighted estimates. To illustrate this with a simple example, let $\bar{x}_w = \sum_{t=1}^{T} w_t x_t$ be a weighted average of T independent observations, $x_1, \ldots, x_T$, drawn from a population with the variance σ^2, and w_t's are the weights which are non-negative and normalised to sum to 1. The variance of $\bar{x}_w$ is $\mathbb{Var}\,\{\bar{x}_w\} = \sigma^2 \sum_{t=1}^{T} w_t^2 = \sigma^2/b$, where $b = 1 \Big/ \sum_{t=1}^{T} w_t^2$. When there is no weighting, i.e. all w_t's are 1, the variance is $\mathbb{Var}\,\{\bar{x}_w\} = \sigma^2/T$. Hence, b measures the effective sample size; it is less than the actual sample size, T. (A related notion is that of 'weighting efficiency' which measures how much data has been retained.)

Although the idea of choosing bandwidth by likelihood cross validation (also known as Kullback-Leibler cross validation) is not new (see e.g Pagan and Ullah, 1999, sec. 2.7), in the present contents the properties of the estimators for h obtained by maximising the MPL or CV criterion are as yet unknown and are subject to future research.

The estimation procedure thus involves first maximizing the likelihood function (2.12) or the CV criterion (2.13) to obtain estimates of the smoothing parameter ω and the bandwidth h. The estimates are then used to compute the estimates of the pdf, cdf and quantiles. cdf (filtered or smoothed) can be computed by applying formulae (2.7) and (2.8) directly. Quantile functions can be obtained by inverting estimated cdfs as described in section 2.5.1 below.

In our computations, simple EWMA weights are used. To be precise, the weights for filtering are given by

$$w_{t,i} = \begin{cases} \frac{1-\omega}{1-\omega^t}\omega^{(t-i)} & \text{if} \quad \omega \in (0,1) \\ 1/t & \text{if} \quad \omega = 1, \end{cases} \qquad i = 1, \ldots, t, \ t = m, \ldots, T, \tag{2.14}$$

where $\omega \in (0;1]$ is a smoothing parameter. These weights are positive and sum to unity over i by construction. Two-sided EWMA weights take the form

$$w_{t,T,i} = \begin{cases} \frac{1-\omega}{1+\omega-\omega^t-\omega^{T-t+1}}\omega^{|t-i|} & \text{if} \quad \omega \in (0,1) \\ 1/T & \text{if} \quad \omega = 1, \end{cases} \qquad i, t = 1, \ldots, T, \tag{2.15}$$

where, as before, $\omega \in (0;1]$ is a smoothing constant. The weights for cross-validation are given by

$$w^{\text{CV}}_{t,T,i} = \begin{cases} \frac{1-\omega}{2\omega-\omega^t-\omega^{T-t+1}}\omega^{|t-i|} & \text{if} \quad \omega \in (0,1) \\ 1/(T-1) & \text{if} \quad \omega = 1. \end{cases} \qquad i, t = 1, \ldots, T, \ i \neq t. \tag{2.16}$$

2.3.4 Sequential estimation

In section 2.3.3 only one set of parameters is estimated per series, as is common in the time series literature. However, if the purpose lies in forecasting, issuing a filtered density at time t as a forecast at time $t+1$ will result in over-optimism[4] as in practice only observations up to time t will be available. This is due to the fact that in (2.12) the same data is used to fit the model and assess its error.

[4] Optimism is the expected difference between the in-sample prediction error and the training error, which is typically positive. Formally, let Y be a target variable, X a vector of predictors, and $\hat{g}(X)$ be a prediction model estimated on a training sample (y_i, x_i), $i = 1, \ldots, n$. Let $L\left(Y, \hat{g}(X)\right)$ be the loss

Hence, at time t, instead of (2.12), the predictive log-likelihood should be given by

$$\begin{aligned}\ell(\omega_t, h_t) &= \frac{1}{t-m_0}\sum_{s=m_0}^{t-1} \ln \hat{f}_{s+1|s}(y_{s+1}) \\ &= \frac{1}{t-m_0}\sum_{s=m_0}^{t-1} \ln \left[\frac{1}{h}\sum_{i=1}^{s} K\left(\frac{y_{s+1}-y_i}{h}\right) w_{s,i}(\omega)\right], \quad T \geq t \geq m_1 > m_0, \quad (2.17)\end{aligned}$$

where m_1 is the number observations which are used for initialization and m_0 is the number of observations which allows a sensible density estimate to be computed; $m_1 - m_0$ observations will be used to obtain the first estimates of the parameters. If the sample is large, we suggest setting $m_0 = 50$ and $m_1 = 100$ or more. Maximizing $\ell(\omega_t, h_t)$ subject to $\omega_t \in (0, 1]$ and $h_t > 0$ for each t gives a sequence of estimates $\{\hat{\omega}_t, \hat{h}_t\}_{t=m_1}^{T}$ which are then used to obtain filtered estimates of the pdf, cdf and quantiles.

Note that (2.17) is the prequential likelihood of Dawid (1984, p. 287). It avoids the over-optimism of (2.12) in which future values of y are used to estimate the parameters entering the forecast at time t. Maximizing (2.17) is also equivalent to minimizing the 'ignorance' of a forecaster, which is a strictly proper scoring rule in that the expected ignorance has a single minimum when the forecast density is the same as the true density; see Roulston and Smith (2002, sec. 2).

2.4 Specification and diagnostic checking

The probability integral transform (PIT) of an observation from a given distribution has a uniform distribution on the range $[0, 1]$. Hence the hypothesis that a set of observations follow a particular parametric distribution can be tested. One possibility is to use a Kolmogorov-Smirnov test.

PITs are often used to assess forecasting schemes; see Dawid (1984) or Diebold et al. (1998). Here the PIT is given directly by the predictive kernel cdf, that is the PIT of the t-th observation is $\widehat{F}_{t|t-1}(y_t)$, $t = m+1, \ldots, T$. As with the evaluation of $\widehat{f}_{t|t-1}(y_t)$ in the likelihood function, the calculation at each point in time need only be done for $y = y_t$.

function. The test (or generalisation) error is the expected prediction error over an independent test sample, i.e. $\text{Err} = \mathbb{E}\left\{L\left(Y, \hat{g}(X)\right)\right\}$. *Training error* is the average loss over the training sample, i.e. $\overline{\text{err}} = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, \hat{g}(x_i)\right)$. Finally, the *in-sample error* is defined as $\text{Err}_{in} = \frac{1}{n}\sum_{i=1} \mathbb{E}_{Y^{new}}\left\{L\left(Y_i^{new}, \hat{g}(x_i)\right)\right\}$, where Y^{new} denotes n new responses observed at the values x_i, $i = 1, \ldots, n$. *Optimism* is then defined as $\text{Err}_{in} - \mathbb{E}_{\mathbf{y}}\left\{\overline{\text{err}}\right\}$; see e.g. Hastie, Tibshirani, and Friedman (2001, Sec. 7.4).

PITs may be expressed in terms of innovations. Specifically,

$$\widehat{F}_{t|t-1}(y_t) = H(0) - V_{t|t-1}(y_t) = 0.5 - V_{t|t-1}(y_t).$$

Hence $E(V_{t|t-1}(y_t)) = 0$ as $E(F_t(y_t)) = 0.5$.

If PITs are not uniformly distributed, their shape can still be informative. For example, a humped distribution indicates that forecasts are too narrow and that tails are not adequately accounted for; see Laurent (2007, p. 98). Plots of the autocorrelation functions (ACFs) of the PITs, and of absolute values[5] and powers of demeaned PITs, may indicate the source of serial dependence. Tests statistics for detecting serial correlation, such as Box-Ljung and stationarity test statistics may be used, but it should be noted that their asymptotic distribution is unknown. There may sometimes be advantages in transforming to normality as in Berkowitz (2001).

2.5 Time-varying quantiles

Visualising a time-varying density requires either a three-dimensional plot or a movie depicting the changes. One way to create a two-dimensional static display is to focus on selected quantiles: a plot showing how quantiles have evolved over time provides a good visual impression of the changing distribution.

Selected quantiles are also of independent interest and considerable practical importance. For example, value at risk (VaR)—the standard measure of market risk used in finance—is a particular quantile of future portfolio values (Engle and Manganelli, 2004). Also, predicting wind power by issuing forecasts on a number of quantiles is an important tool in the daily management of wind generation (Pinson, Nielsen, Møller, Madsen, and Kariniotakis, 2007).

Quantiles can be obtained by inverting an estimate of the cumulative distribution function as described in the first sub-section below.

The second sub-section reviews some of the procedures for direct estimation of time-varying quantiles by formulating a model for a particular quantile and contrasts them with the nonparametric approach proposed in this chapter.

2.5.1 Kernel-based estimation

When the distribution is constant, the τ-quantile, $\xi(\tau)$, $0 < \tau < 1$, can be estimated from the distribution function by solving $\widehat{F}(y) = \tau$, i.e. $\widehat{F}^{-1}(\tau) = \widehat{\xi}(\tau)$. Nadaraya

[5] The absolute value of a demeaned PIT is also uniformly distributed, unlike its square.

(1964) shows that $\widehat{\xi}(\tau)$ is consistent and asymptotically normal with the same asymptotic distribution as the sample quantile. Azzalini (1981) proposes the use of a Newton-Raphson procedure for obtaining $\widehat{\xi}(\tau)$.

Filtered and smoothed estimators of changing quantiles can be similarly computed from time-varying cdf's. Thus, for filtering, $\widehat{\xi}_{t|t-1}(\tau) = \widehat{F}^{-1}_{t|t-1}(\tau)$, for $t = m, \ldots, T$. The iterative procedure to calculate $\widehat{\xi}_{t|t-1}(\tau)$ is based on the direct evaluation of $\widehat{F}_{t|t-1}(y)$ in the vicinity of the quantile. To reduce computational time, a good starting value can be obtained from a preliminary estimate of a cdf by (linear) interpolation[6]. Alternatively, for $t = m+1, \ldots, T$, the estimate in the previous time period may be used as a starting value.

The estimates of bandwidth obtained by MPL or CV suffer from the drawback that the asymptotically optimal choice of bandwidth for a kernel estimator of a cdf is proportional to $T^{-\frac{1}{3}}$, whilst the optimal bandwidth for a pdf is proportional to $T^{-\frac{1}{5}}$; see, for example, Azzalini (1981). A bandwidth for a kernel estimator of a cdf can be found by CV, as in Bowman, Hall, and Prvan (1998), or by a rule of thumb approach, as in Altman and Léger (1995). It may be worth experimenting with these bandwidth selection criteria for quantile estimation. Similar considerations may apply to the computation of the PITs.

2.5.2 Direct estimation of individual quantiles

Yu and Jones (1998) adopt a nonparametric approach. Their (smoothed) estimate, $\widehat{\xi}_t(\tau)$, of the τ-quantile is obtained by (iteratively) solving

$$\sum_{j=-h}^{h} K(\frac{j}{h}) IQ(y_{t+j} - \widehat{\xi}_t) = 0,$$

where $\widehat{\xi}_t = \widehat{\xi}_t(\tau)$, $K(\cdot)$ is a weighting kernel (applied over time), h is a bandwidth and $IQ(\cdot)$ is the quantile indicator function

$$IQ(y_t - \xi_t) = \begin{cases} \tau - 1, & \text{if} \quad y_t < \xi_t, \\ \tau, & \text{if} \quad y_t > \xi_t, \end{cases} \qquad t = 1, \ldots, T.$$

[6] To be precise, in our code, the cdf is first estimated on a grid of K points $\xi_1, \ldots, \xi_K$, and the initial estimate of ξ_t is obtained by finding $\xi_{lo} = \max_j \left(\xi_j : \widehat{F}_t(\xi_j) \leq \tau\right)$ and $\xi_{up} = \min_j \left(\xi_j : \widehat{F}_t(\xi_j) \geq \tau\right)$, and linearly interpolating between them. This is then used as a starting value in solving $\widehat{F}_t(\xi_t) = \tau$ for ξ_t. The final solution can usually be found in just a few iterations (we used the Matlab routine `fzero`). In fact, with large K, the precision of the initial estimate of ξ_t will be sufficient for all practical purposes.

$IQ(0)$ is not determined, but in the present context we can set $IQ(0) = 0$. Adding and subtracting $\widehat{\xi}_t$ to each of the $IQ(y_{t+j} - \widehat{\xi}_t)$ terms in the sum leads to the alternative expression

$$\widehat{\xi}_t = \frac{1}{\sum_{j=-h}^{h} K(j/h)} \sum_{j=-h}^{h} K(\frac{j}{h})[\widehat{\xi}_t + IQ(y_{t+j} - \widehat{\xi}_t)]. \tag{2.18}$$

De Rossi and Harvey (2006, 2009) estimate time-varying quantiles by smoothing with weighting patterns derived from linear models for signal extraction. These quantiles have no more than $T\tau$ observations below and no more than $T(1-\tau)$ above. The weighting scheme derived from the local level model gives

$$\widetilde{\xi}_t = \frac{1-\omega}{1+\omega} \sum_{j=-\infty}^{\infty} \omega^{|j|}[\widetilde{\xi}_t + IQ(y_{t+j} - \widetilde{\xi}_{t+j})],$$

in a doubly infinite sample; cf. (2.4). The nonparametric kernel $K(j/h)$ in (2.18) is replaced by $\omega^{|j|}$ so giving an exponential decay. Note that the smoothed estimate, $\widehat{\xi}_{t+j}$, is used instead of $\widehat{\xi}_t$ when j is not zero. The time series model determines the shape of the kernel while the signal-to-noise ratio plays a role similar to that of the bandwidth.

The smoothed estimate of a quantile at the end of the sample is the filtered estimate. The model-based approach automatically determines a weighting pattern at the end of the sample. For the EWMA scheme derived from the local level model, the filtered estimator must satisfy

$$\widetilde{\xi}_{t|t} = (1-\omega) \sum_{j=0}^{\infty} \omega^j [\widetilde{\xi}_{t-j|t} + IQ(y_{t-j} - \widetilde{\xi}_{t-j|t})].$$

Thus $\widetilde{\xi}_{t|t}$ is an EWMA of the synthetic observations, $\widetilde{\xi}_{t-j|t} + IQ(y_{t-j} - \widetilde{\xi}_{t-j|t})$. As new observations become available, the smoothed estimates need to be revised. However, filtered estimates could be used instead, so

$$\widehat{\xi}_{t+1|t}(\tau) = \widehat{\xi}_{t|t-1}(\tau) + (1-\omega)\nu_t(\tau), \tag{2.19}$$

where $\nu_t(\tau) = IQ(y_t - \widehat{\xi}_{t|t-1}(\tau))$ is an indicator that plays an analogous role to that of the innovation in the Kalman filter. Such a scheme would belong to the class of CAViaR models proposed by Engle and Manganelli (2004) in the context of tracking value at risk. In CAViaR, the conditional quantile is

$$\widehat{\xi}_{t+1|t}(\tau) = \alpha_0 + \sum_{i=1}^{q} \beta_i \widehat{\xi}_{t+1-i|t-i}(\tau) + \sum_{j=1}^{r} \alpha_j g(y_{t-j}),$$

where $g(y_t)$ is a function of y_t. Suggested forms include the *adaptive* model

$$\xi_t(\tau) = \xi_{t-1}(\tau) + \gamma\{[1 + \exp(\delta[y_{t-1} - \xi_{t-1}(\tau)])]^{-1} - \tau\}, \tag{2.20}$$

where δ is a positive parameter. The recursion in (2.19) has the same form as the limiting case ($\delta \to \infty$) of (2.20). Other CAViaR specifications, which are based on actual values, rather than indicators, may suffer from a lack of robustness to additive outliers. That this is the case is clear from an examination of Fig. 1 in Engle and Manganelli (2004, p. 373). More generally, recent evidence on predictive performance in Kuester, Mittnik, and Paolella (2006, pp. 80–81) indicates a preference for the adaptive specification.

The advantage of fitting individual quantiles is that different parameters may be estimated for different quantiles. The disadvantage of having different parameters is that the quantiles may cross; see Gourieroux and Jasiak (2008). If the parameters across quantiles are restricted to be the same to prevent quantiles crossing, the ability to have different models for different quantiles loses much of its appeal.

2.6 Empirical application: NASDAQ index

Data on the NASDAQ index was obtained from Yahoo-Finance (`http://uk.finance.yahoo.com`). The sample starts on 5th February 1971 and ends on 20th February 2009, thus covering 13,896 days. Once weekends and holidays are excluded, there are 9,597 observations. As is usually the case with financial series, there is clear volatility clustering and the correlograms of the absolute values and squares of demeaned returns are large and slowly decaying; see Fig. 2.1. Some of the sample autocorrelations for the actual returns and their cubes also lie outside ±2 standard deviations from the horizontal axis. The distribution of returns is heavy-tailed and asymmetric.

2.6.1 Time-varying kernel

Fig. 2.2 shows filtered (upper panel) and smoothed (lower panel) time-varying quantiles of NASDAQ returns for $\tau = 0.05, 0.25, 0.50, 0.75, 0.95$. Exponential weights and an Epanechnikov kernel are used throughout. The discount parameters for filtering and smoothing are estimated by maximizing the log-likelihood and likelihood CV criterion respectively. MPL estimates of the discount parameter and bandwidth are, respectively, $\widetilde{\omega} = 0.9928$ and $\widetilde{h} = 0.4286$. CV estimates (for smoothing) are $\widehat{\omega} = 0.9928$ and $\widehat{h} = 0.2555$.

The quantiles, plotted in Fig. 2.2, seem to track the changing distribution well.

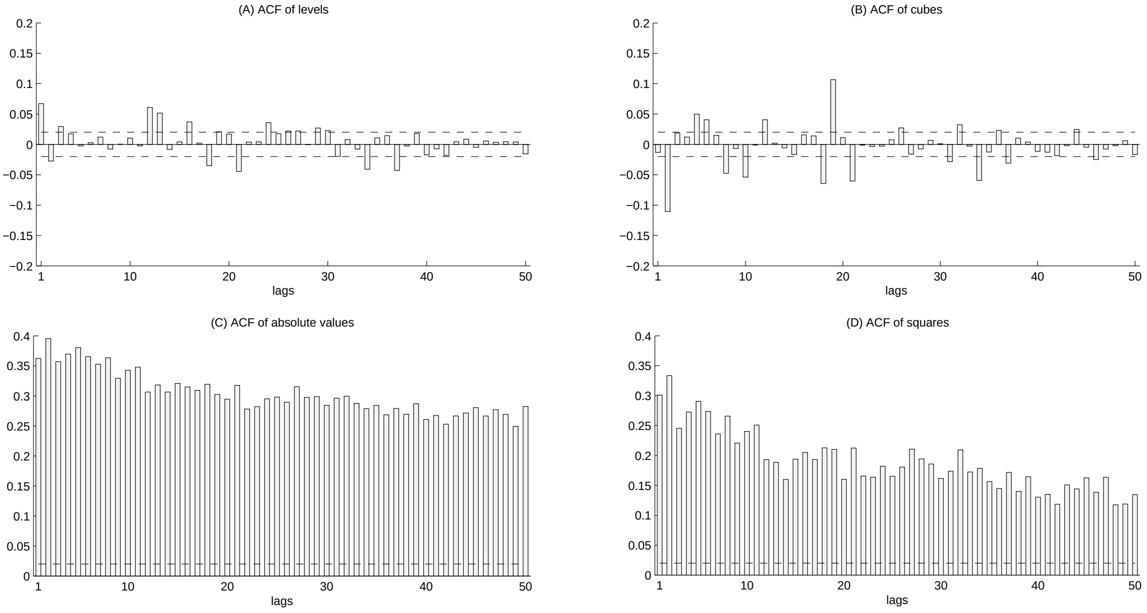

Figure 2.1: ACFs of NASDAQ returns.

Panel A: ACF of returns y_t. Panels B, C and D: ACFs of $(y_t - \bar{y})^3$, $|y_t - \bar{y}|$ and $(y_t - \bar{y})^2$ respectively. Lines parallel to the horizontal axis are ±2 standard deviations (i.e. $2/\sqrt{T}$).

However, as Fig. 2.3 shows, there is still some residual serial correlation in absolute values and squares of the PITs. With raw data, changing volatility tends to show up more in absolute values than in squares, as in Fig. 2.1. One reason for this is that sample autocorrelations are less sensitive to outliers when constructed from absolute values rather squares. However, the PITs do not have heavy tails, and the absolute value sample autocorrelations are, in most cases, slightly less than the corresponding sample autocorrelations computed from squares.

The first-order sample autocorrelation in the raw returns is rather high. It is even higher in the PITs. This may be partly a consequence of the transformation, though the higher order autocorrelations are, if anything, smaller than the corresponding autocorrelations for the raw returns.

The sample autocorrelations of the third and fourth powers of the demeaned PITs (not shown here) are, like those of the absolute values, small but persistent.

The histogram of PITs, shown in Fig. 2.3 panel D, is too high in the middle and too low at the ends, showing departures from uniformity and hence imperfections in the forecasting scheme. The hump-shaped distribution of the PITs indicates that tail behaviour is not adequately captured. The problem could be caused by the bandwidth being too wide, resulting in a degree of oversmoothing. Forecasting performance might

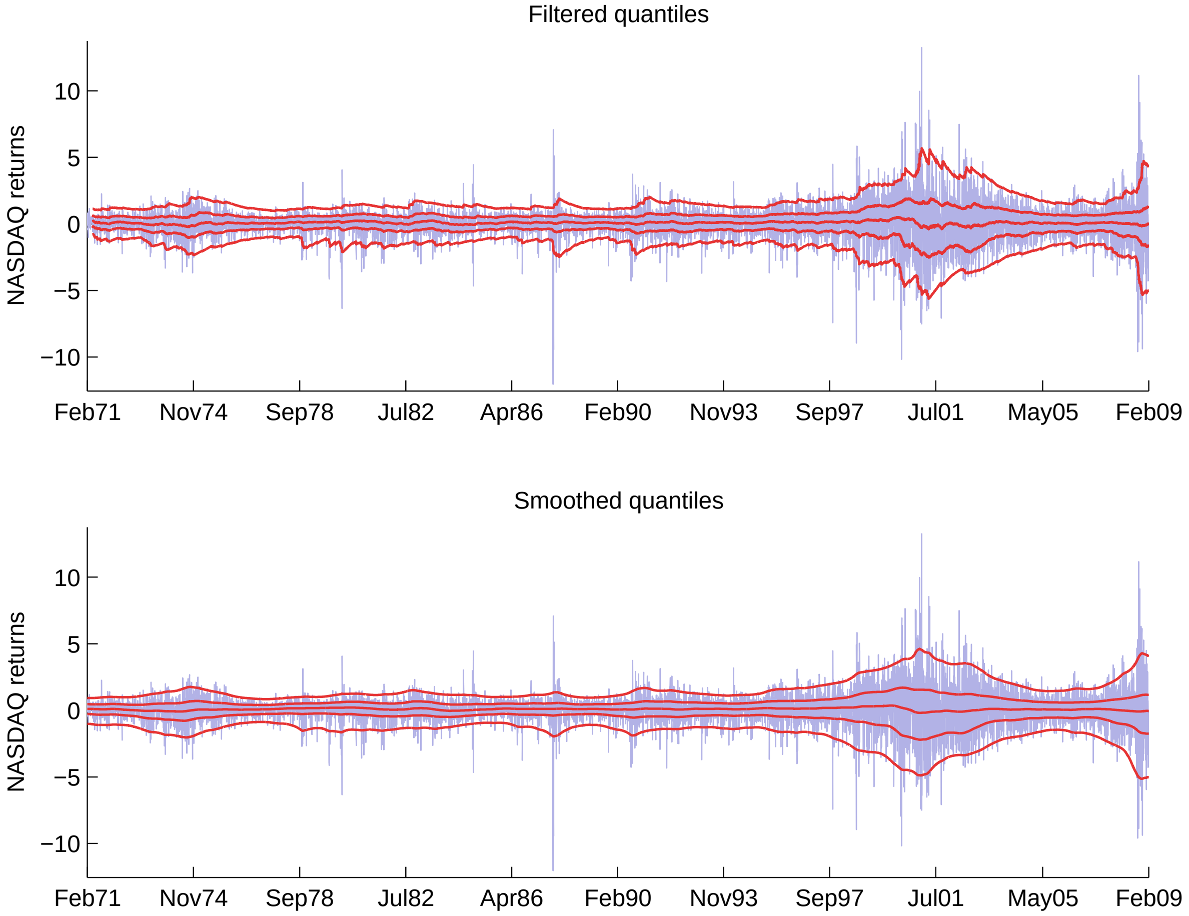

Figure 2.2: Filtered (upper panel) and smoothed (lower panel) time-varying quantiles of NASDAQ returns.

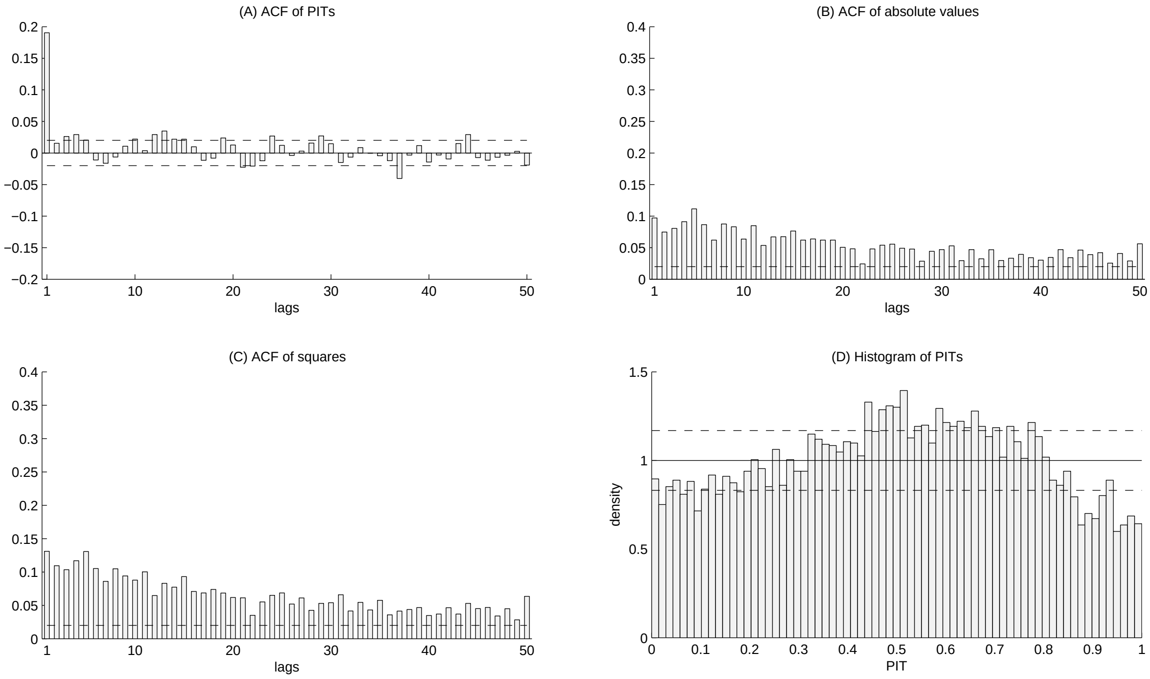

Figure 2.3: ACFs and histogram of PITs.

Panels A, B and C: ACFs of PITs, z_t, absolute values, $|z_t - \bar{z}|$, and squares of the demeaned PITs, $(z_t - \bar{z})^2$, respectively; lines parallel to the horizontal axis are ±2 standard deviations (i.e. $2/\sqrt{T}$). Panel D: histogram of PITs; dashed lines are ±2 standard deviations (i.e. $2\sqrt{(k-1)/T}$, where k is the number of bins).

be improved by using different bandwidths for the tails and middle of the distribution.

Changing the basis for bandwidth selection is unlikely to correct the failure to pick up short term serial correlation (at lag one) or to remove all the movements in volatility. The reason is that a time-varying kernel can really only pick up long-term changes. Hence there may be a case for pre-filtering.

2.6.2 ARMA-GARCH residuals

To pick up trending and/or seasonal movements the level can be modelled separately, for example by formulating a state space model. Short term serial correlation may be similarly handled by fitting an autoregressive–moving average (ARMA) model. The most straightforward method for dealing with short-term movements in variance is to fit a GARCH model for the conditional variance. Dynamic kernel estimation can then be applied to the innovations. As the following example shows, such analysis can pick up time variation in the features of the distribution not captured by the parametric model used to pre-filter the data.

First order serial correlation and conditional volatility on NASDAQ returns can be modelled parametrically by an MA(1) model with a GARCH(1,1)-t conditional variance equation. The model was fitted using the G@RCH 5 program of Laurent (2007). GARCH parameters are estimated to be 0.0979 (the coefficient of the lagged squared observation) and 0.9010, so the sum is close to the IGARCH boundary. The estimated MA(1) parameter is 0.2102, while the degrees of freedom of the t-distribution is estimated to be 7.04.

Fitting a time-varying kernel to the GARCH residuals gives MPL estimates of $\widetilde{\omega} = 0.9996$ and $\widetilde{h} = 0.3595$, and CV estimates $\widehat{\omega} = 0.9991$ and $\widehat{h} = 0.3339$. The discount parameters are bigger than those estimated for the raw data and since they are closer to one there is less scope for picking up time variation, as can be seen from the quantiles in Fig. 2.4 (quantiles are shown for $\tau = 0.01, 0.05, 0.25, 0.50, 0.75, 0.95, 0.99$). As might be anticipated, pre-filtering effectively renders the median and inter-quartile range constant. Any remaining time variation is to be found in the high and low quantiles.

Some notion of the way in which tail dispersion changes can be obtained by plotting the ratio of the τ to $1-\tau$ range, for small τ, to the interquartile range, that is

$$\widetilde{\alpha}_t(\tau) = \frac{\widetilde{\xi}_t(1-\tau) - \widetilde{\xi}_t(\tau)}{\widetilde{\xi}_t(0.75) - \widetilde{\xi}_t(0.25)}, \quad \tau < 0.25,$$

where $\widetilde{\xi}_t(\tau)$ is an estimator obtained by filtering or smoothing. Fig. 2.5 plots $\widetilde{\alpha}_t(\tau)$ for

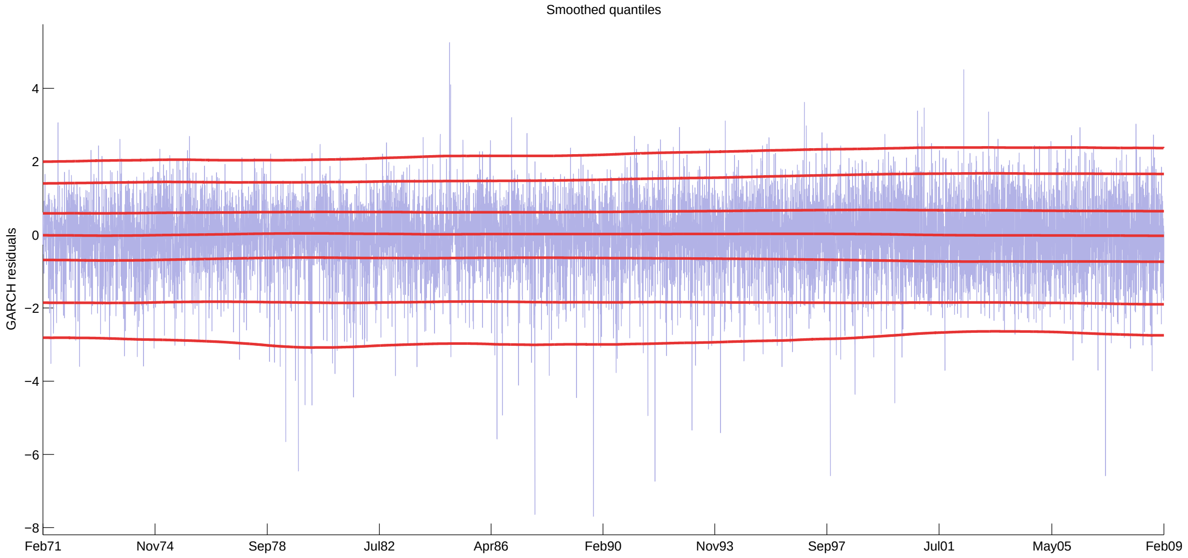

Figure 2.4: Smoothed time-varying quantiles of GARCH residuals.

$\tau = 0.01$ and 0.05 computed using smoothed quantiles. Note that $\alpha(0.05)$ is 2.44 for a normal distribution and 2.66 for t_7; the corresponding figures for $\alpha(0.01)$ are 3.45 and 4.22 respectively.

For a symmetric distribution $\xi(\tau) + \xi(1-\tau) - 2\xi(0.5)$ is zero for all $t = 1, \ldots, T$. Hence a plot of the skewness measure

$$\widetilde{\beta}_t(\tau) = \frac{\widetilde{\xi}_t(1-\tau) + \widetilde{\xi}_t(\tau) - 2\widetilde{\xi}_t(0.5)}{\widetilde{\xi}_t(1-\tau) - \widetilde{\xi}_t(\tau)}, \quad \tau < 0.5,$$

shows how asymmetry captured by the complementary quantiles, $\xi_t(\tau)$ and $\xi_t(1-\tau)$, changes over time. The statistic $\beta(0.25)$ was originally proposed by Bowley in 1920; see Groeneveld and Meeden (1984) for a detailed discussion. The maximum value of $\widetilde{\beta}_t(\tau)$ is one, representing extreme right (positive) skewness and the minimum value is minus one, representing extreme left skewness. Fig. 2.5 plots $\widetilde{\beta}_t(\tau)$ for $\tau = 0.01$, 0.05 and 0.25 using the smoothed quantiles. There is substantial time variation in skewness: it is high in the late 70s, whereas around 2002–2005, the distribution is almost symmetric. It is unclear why this is occurring and these features may be worthy of further investigation.

The ACFs of the PITs, their squares and absolute values are shown in Fig. 2.6. There is far less serial correlation than in the corresponding correlograms in Fig. 2.3. The histogram of PITs from a time-varying kernel fitted to ARMA-GARCH residuals, shown in Fig. 2.6, displays the same hump-shaped pattern as was evident in the PITs from the raw data, but arguably to a lesser extent.

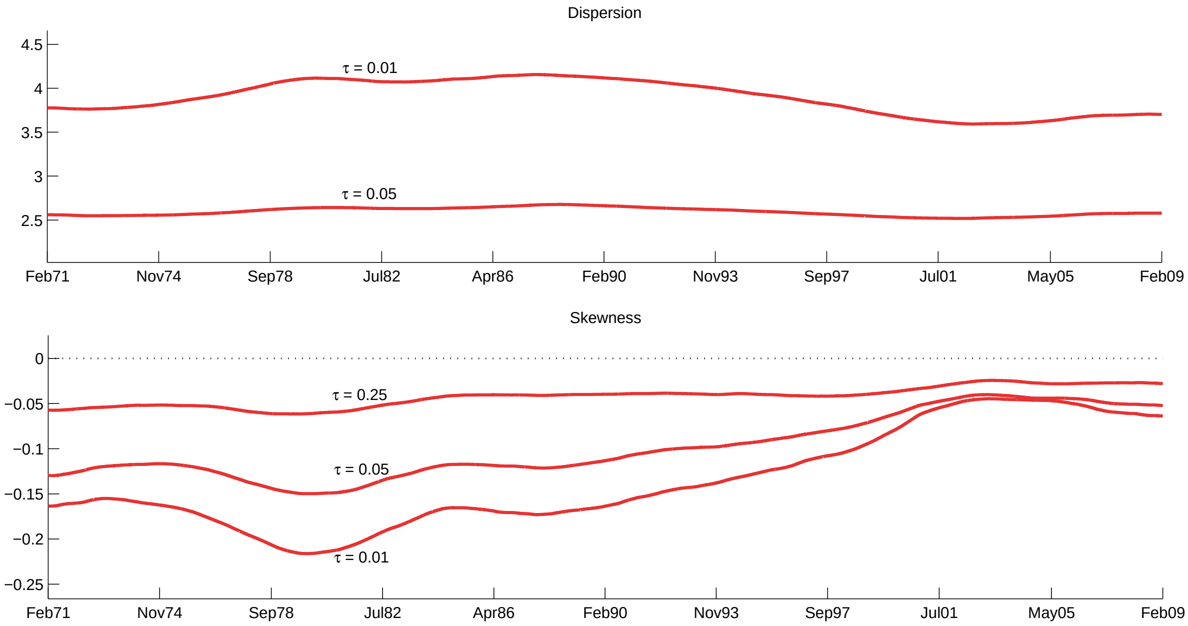

Figure 2.5: Changing tail dispersion and skewness for GARCH residuals.

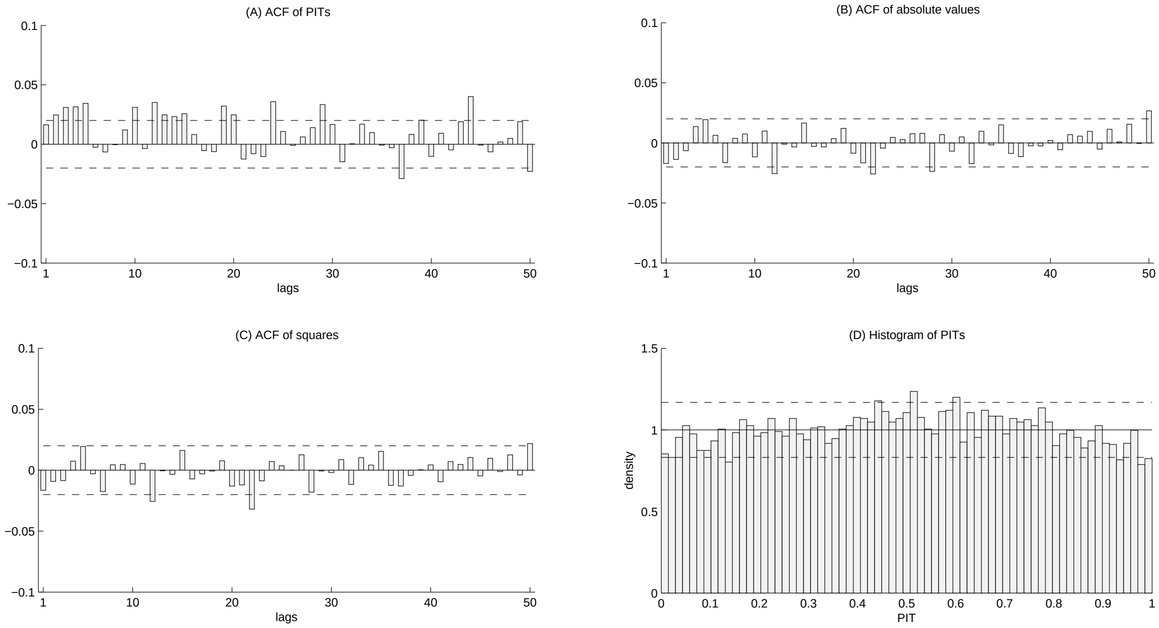

Figure 2.6: ACFs and histogram of PITs of GARCH residuals.

Panels A, B and C: ACFs of PITs, z_t, absolute values, $|z_t - \bar{z}|$, and squares of the demeaned PITs, $(z_t - \bar{z})^2$, respectively; lines parallel to the horizontal axis are ± 2 standard deviations (i.e. $2/\sqrt{T}$). Panel D: histogram of PITs; dashed lines are ± 2 standard deviations (i.e. $2\sqrt{(k-1)/T}$, where k is the number of bins).

Pre-filtering the data with a GARCH model thus moves the focus away from the dynamics of conditional volatility (well captured by GARCH) and towards a finer features of the distribution discernible by analysing high quantiles. Such analysis may also be used as a part of a model building procedure; for example, if a parametric model is sought for NASDAQ returns, it should at least accommodate changes in skewness.

The disadvantage of pre-filtering is that the treatment of location and scale becomes decoupled from the estimation of the distribution as a whole.

2.6.3 Alternative weighting schemes

As has been pointed out before, although EWMA weights arise naturally in a class of models (section 2.2), it is not known whether they possess any optimality properties in the present context. It is thus of interest to compare the results employing alternative weighting schemes.

The simplest way to analyse evolving densities is by passing a window of a certain size, m, through the series; that is, using rectangular weights. This suggestion appears in, *inter alia*, Hall and Patil (1994), but finding the optimal size of the window is left to the subjective judgement of the user. We propose estimating m in the same way as the discount parameter in the EWMA weighting scheme, viz. by maximising the predictive likelihood and the likelihood CV criterion for one- and two-sided filtering, respectively.

For example, for NASDAQ returns, maximising the CV criterion gives the optimal window size $\hat{m} = 633$ (that is, 316 observations are used on either side of t) with the optimal bandwidth estimated as $\hat{h} = 0.3631$. The resulting smoothed quantiles are shown as solid lines in Fig. 2.7; quantiles obtained using exponential weights are replicated for ease of comparison (dashed lines). Qualitatively, both weighting schemes deliver similar results, with the rectangular weighting resulting in a somewhat more rugged pattern.

Finally, other simple weighting patterns—such as linear (triangular) or quadratic—can be used. Estimation results (not reported) using these weighting schemes, however, appear to be inferior to EWMA and rectangular weighting.

2.7 Conclusions

We have proposed a modification of kernel density estimation that allows changes in the density, and hence quantiles, to be captured by weighting observations using schemes derived from time series models. The paper shows how the implied recursive procedures are of a similar form to those used for filtering time series observations to extract evolving means or variances. Associated smoothing schemes are obtained in the same way.

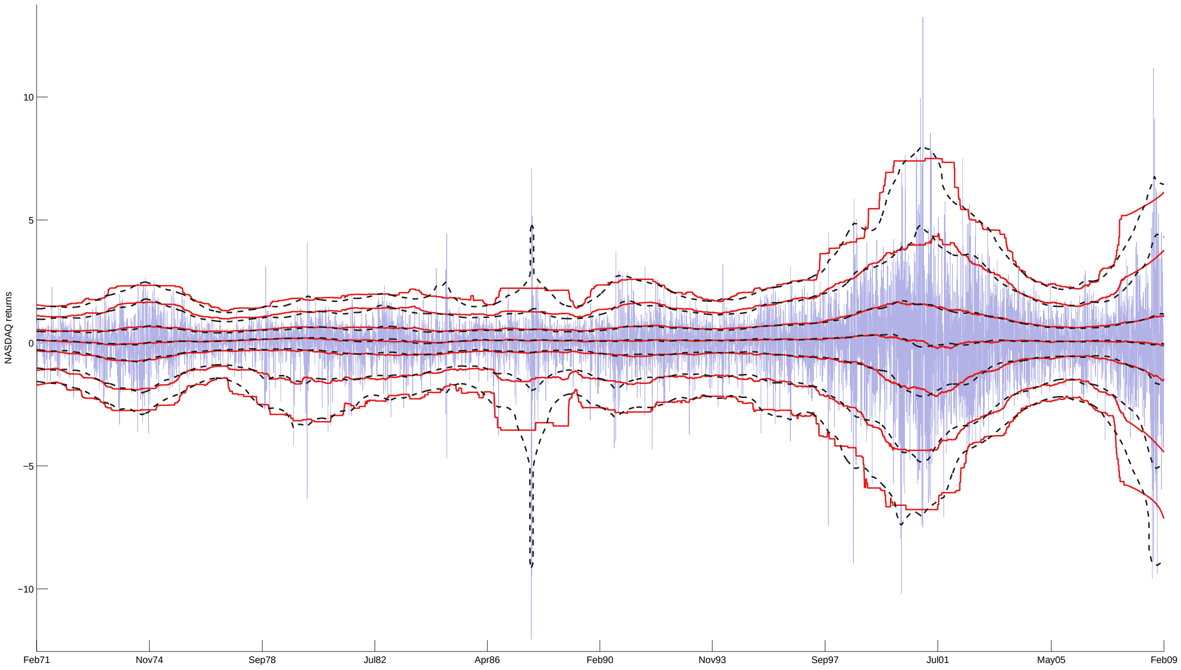

Figure 2.7: Smoothed time-varying quantiles of NASDAQ returns.

As is the case for many time series models, the likelihood function may be obtained from the predictive distribution. Hence the parameters governing the dynamics of the kernel can be estimated, together with the bandwidth, by MPL. Estimates for smoothing may be obtained by CV. The innovations produced by the predictive cdf are PITs and can be used for diagnostic checking. If there is time variation in medians, asymmetry and the tails of distributions, tracking the changes in the whole distribution, or in a limited number of quantiles or quantile contrasts, may be informative.

Attention has been focussed on discounting past observations using EW. Exponential weighting is very simple to apply. However, generalizations to other weighting schemes are not difficult because the filters can be obtained from the state space forms of appropriate time series models. One scheme that certainly warrants future investigation is the stable filter corresponding to the standard stationary GARCH model.

The techniques are illustrated on NASDAQ stock market index. These applications show the advantages of the proposed methods, but also expose their limitations. In particular the methods are only appropriate for monitoring distributions that change relatively slowly over time, since otherwise the effective sample size is too small. Short bursts of volatility may have to be accommodated by fitting a GARCH model.

A second limitation is that the bandwidth chosen by maximising the likelihood function or the likelihood CV criterion appears to result in a degree of oversmoothing, which manifests itself in the hump-shaped histogram of the PITs. It may be possible to mitigate this effect by letting the bandwidth vary over the support of the distribution, but the fundamental problem is that there is not enough information to provide an accurate description of tail behaviour. Modifications, such as combining kernel estimators with extreme value distributions for the tails, as in Markovich (2007, pp. 101–111), may be worth exploring.

Further research is required to assess the relative merits of choosing the bandwidth by maximising MPL and CV criterion or by a rule of thumb or other methods.

Chapter 3

Generalised empirical likelihood–based kernel density estimation

If additional information about the distribution of a random variable is available in the form of moment conditions, a weighted kernel density estimate reflecting the extra information can be constructed by replacing the uniform weights with the generalised empirical likelihood probabilities. It is shown that the resultant density estimator provides an improved approximation to the moment constraints. Moreover, a reduction in variance is achieved due to the systematic use of the extra moment information.

Contents

3.1 Introduction

Nonparametric density estimation is an important tool in applied econometrics, finance, and many other areas, where it is often used for exploratory data analysis or as a part of another estimator; see e.g. Pagan and Ullah (1999), Wand and Jones (1995), Silverman (1986) and Li and Racine (2007).

The simplest case of nonparametric density estimation can be stated as follows. Let X be a univariate random variable which has a continuous probability density function f, and let $\{X_1, \ldots, X_n\}$ denote a random sample of size n. The goal is then to estimate f based on the observed sample.

The kernel density estimator (KDE) of f at an arbitrary point x is given by

$$\hat{f}(x; h_n) = \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(X_i - x), \tag{3.1}$$

where $K_{h_n}(z) = K(z/h_n)/h_n$, $K(\cdot)$ is the kernel function, and $h_n > 0$ is the smoothing parameter known as bandwidth. This estimator was proposed by Rosenblatt (1956) and Parzen (1962) and can be motivated as a smoothed version of a histogram. We will write $\hat{f}(x)$ for $\hat{f}(x; h_n)$ and h for h_n; the dependence of $\hat{f}$ on bandwidth and of h on sample size being implicit.

In some applications it may be necessary to construct an estimator of a probability density function (pdf) which obeys certain constraints. For instance, the mean of X may be known or there may be a known relationship between the moments, perhaps implied by estimating equations. Extra distributional information may be due to a certain physical law as in the example considered in Chen (1997) where according to the line transect theory the distribution of the perpendicular sighting distances in an aerial line transect survey should have mean zero.

Assumptions about the relationship between the mean and variance of the observations underlies the standard quasi-likelihood estimation; see Wedderburn (1974) and Godambe and Thompson (1989). If the variance of X is a known function of the (unknown) mean, μ, the information about f can be expressed in the form of two moment conditions, viz. $\mathbb{E}\{X\} = \mu$ and $\mathbb{E}\left\{(X-\mu)^2\right\} = g(\mu)$, where $g(\cdot)$ is a known function. The method presented in this chapter allows such information to be incorporated into an estimate of f.

Incorporating auxiliary population information is also of interested when using survey data; see e.g. Qin and Lawless (1994, p. 301) and Chen and Qin (1993). For example, one may be interested in estimating the density of household income based on a survey data. If the average income is known from, say, census data, it can be treated as a known population mean and incorporated into the estimate.

In principle, representing (3.1) as a maximum smoothed likelihood estimator, provides a way to incorporate extra information by solving a constrained optimisation problem instead, but the latter may be difficult or even impossible; see Eggermont and LaRiccia (2001).

This chapter considers the case when the extra information can be formulated in the form of moment conditions on X. This case has been examined by Chen (1997) who proposes re-weighting the Rosenblatt-Parzen KDE (RPKDE) using *empirical likelihood* weights instead of equal probability weights, n^{-1}, placed at every data point. A similar approach has been applied by Hall and Presnell (1999).

Specifically, suppose that additional information about f is available in the form

$$\mathbb{E}_f \left\{ \boldsymbol{\psi} \left(X; \boldsymbol{\beta}_0 \right) \right\} = 0, \tag{3.2}$$

where $\boldsymbol{\psi}\left(x;\boldsymbol{\beta}\right) = \left[\psi_1(x;\boldsymbol{\beta}), \ldots, \psi_q(x;\boldsymbol{\beta})\right]^{\mathsf{T}}$ is a *known* real vector-valued function representing q moment conditions, $\boldsymbol{\beta} \in \mathscr{B} \subseteq \mathbb{R}^p$ is a $p \times 1$ vector of unknown parameters, $p \leq q$, and expectation is taken with respect to the distribution of X.

In this paper, we seek an estimator of f, $\tilde{f}(\cdot)$, which satisfies constraints (3.2) in the sense that $\int \psi_l(u;\boldsymbol{\beta}_0)\tilde{f}(u)du = 0$, $l = 1, \ldots, q$. As shown in section 3.2 RPKDE will not in general possess this property. The reweighted KDE defined in section 3.4 will better satisfy conditions (3.2).

This work extends the previous analysis by allowing parameters in the moment conditions to be *estimated* using *generalised* empirical likelihood (GEL) estimation, described in section 3.3.

Prior to computing an estimate with the constraints imposed, one should test whether the constraints are consistent with the data. For example, in a simple case when the mean is hypothesized to be known a standard t-test can be employed. GEL-based tests can be used as described in section 3.3. As GEL estimation is part of the proposed procedure, such test statistics can be computed at no extra cost.

Properties of the GEL-based estimator are presented in section 3.4. In particular, it is shown that, provided moment conditions contain some overidentifying information, a reweighted estimate will have smaller variance than the standard kernel estimate. We show that the reduction in the variance occurs in the second order term and is the *same* for all members of the GEL family.

Section 3.5 analyses the performance of the proposed density estimator in small and medium samples via a Monte-Carlo study. The final section concludes.

3.2 Rosenblatt–Parzen kernel density estimator

The Rozenblatt-Parzen KDE has been studied extensively and its properties are well-documented. Thus its mean and variance are $\mathbb{E}\left\{\hat{f}(x)\right\} = (K_h * f)(x)$, and

$$\mathbb{Var}\left\{\hat{f}(x)\right\} = n^{-1}\left[\left(K_h^2 * f\right)(x) - \left(K_h * f\right)^2(x)\right]$$

respectively, where $*$ denotes convolution, i.e. $(f * g)(x) = \int f(x-y)g(y)dy$.

The kernel $K(\cdot)$ is assumed to be a bounded probability density function symmetric about the origin, i.e. $K(-z) = K(z)$ and $\int K(z)dz = 1$. Let $\mu_j(K) = \int_{\mathbb{R}} z^j K(z)dz$ be the j-th moment of $K(\cdot)$. Then $K(\cdot)$ is said to be a k-th order kernel if $\mu_0(K) = 1$, $\mu_j(K) = 0$ for $j = 1, \ldots, k-1$, and $\mu_k(K) \neq 0$. Due to symmetry, only even orders need to be considered, and the choice is usually restricted to second order kernels as kernels of order higher than two take negative values, which implies that the resulting density estimate can take negative values and hence is not a density itself. The optimal[1] second-order kernel is the truncated quadratic kernel of Epanechnikov (1969), but as the efficiency loss from using suboptimal kernels is small, the Gaussian kernel is commonly used in practice.

Asymptotic approximations to the mean integrated squared error (MISE) of $\hat{f}$ can be obtained under additional assumptions. In particular, we assume

Assumption 1

(a) $K(\cdot)$ is a symmetric second order kernel and $\mu_4(K) < \infty$.

(b) The bandwidth $h = h_n$ is a non-random sequence such that $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} hn = \infty$.

(c) f possesses a fourth derivative which is continuous and square integrable.

Let $f^{(j)}(x) = \partial^j f(x)/\partial x^j$ denote the j-th derivative of $f(x)$. Then the following asymptotic expansion for the expectation obtains (see e.g. Wasserman, 2006, Theorem 6.28):

$$\mathbb{E}\left\{\hat{f}(x)\right\} = f(x) + \frac{1}{2}h^2\mu_2(K)f^{(2)}(x) + \frac{1}{24}h^4\mu_4(K)f^{(4)}(x) + o\left(h^4\right),$$

where the terms involving odd powers of h vanish due to the symmetry of the kernel.

[1] The truncated quadratic (Epanechnikov) kernel is optimal if the choice is restricted to nonnegative symmetric density functions and the optimality criterion is asymptotic mean integrated squared error; cf. Tsybakov (2009, Ch. 1).

The asymptotic variance of $\hat{f}(x)$ is

$$\mathbb{V}\mathrm{ar}\left\{\hat{f}(x)\right\} = \frac{1}{nh}R(K)f(x) + o\left(n^{-1}\right),$$

where $R(g) = \int_{\mathbb{R}} g^2(x)dx$ for any square-integrable function g. Thus the mean squared error (MSE) is

$$\mathbb{MSE}\left\{\hat{f}(x)\right\} = \frac{1}{nh}R(K)f(x) + \frac{1}{4}h^4\mu_2^2(K)\left(f^{(2)}(x)\right)^2 + o\left(n^{-1}\right) + o\left(h^6\right),$$

where the first term is the variance and the second is the squared bias. Integrating over the range of X gives the mean integrated squared error (MISE) of $\hat{f}$, viz.

$$\mathbb{MISE}\left\{\hat{f}(\cdot)\right\} = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2^2(K)R\left(f^{(2)}\right) + o\left(n^{-1}\right) + o\left(h^6\right). \tag{3.3}$$

The first two terms in (3.3) give the asymptotic MISE (AMISE) of $\hat{f}$. AMISE provides a useful large-sample approximation to MISE. Note that the bias term is of order h^4, whereas the variance term is of order $(nh)^{-1}$. Hence the bandwidth is to be chosen to balance the bias-variance trade-off: smaller values of h reduce bias but increase variance. Differentiating (3.3) with respect to h and setting the result to zero gives the asymptotically optimal (AMISE-minimising) bandwidth,

$$h_{\text{AMISE}} = \left[\frac{R(K)}{\mu_2^2(K)R\left(f^{(2)}\right)}\right]^{1/5} n^{-1/5}. \tag{3.4}$$

With the optimal bandwidth both terms in AMISE become of the same order, $n^{-4/5}$. In practice, the choice of bandwidth is very important; see Sheather (2004) for a recent review and the references given above.

In general, the RPKDE will not satisfy conditions (3.2).

Example 1 Note that $\mathbb{E}_{\hat{f}}\left\{X^j\right\} = n^{-1}\sum_{i=1}^n \int (x_i+zh)^j K(z)dz$. Since $K(\cdot)$ is a symmetric density function, $\mathbb{E}_{\hat{f}}\left\{X\right\} = n^{-1}\sum_{i=1}^n x_i$, the sample average. Hence the constraint that the mean is μ, say, will not generally be satisfied in finite samples. Also

$$\begin{aligned}
\mathbb{E}_{\hat{f}}\left\{X^2\right\} &= n^{-1}\sum_{i=1}^{n} x_i^2 + h^2\mu_2(K), \\
\mathbb{E}_{\hat{f}}\left\{X^3\right\} &= n^{-1}\sum_{i=1}^{n} x_i^3 + 3h^2\mu_2(K)n^{-1}\sum_{i=1}^{n} x_i \qquad \text{and} \\
\mathbb{E}_{\hat{f}}\left\{X^4\right\} &= n^{-1}\sum_{i=1}^{n} x_i^4 + 6h^2\mu_2(K)n^{-1}\sum_{i=1}^{n} x_i^2 + h^4\mu_4(K).
\end{aligned}$$

■

Let $\psi_l^{(j)} = \partial^j \psi_l(x;\boldsymbol{\beta})/\partial x^j$ denote the j-th derivative of $\psi_l(x;\boldsymbol{\beta})$ with respect to x, $l = 1,\ldots,q$. Suppose that $\boldsymbol{\psi}(\cdot)$ satisfies the following conditions:

Assumption 2

(a) $\psi_l(x;\boldsymbol{\beta})$ is four times continuously differentiable in x with a square integrable fourth derivative for all $\boldsymbol{\beta} \in \mathrm{B}_r(\boldsymbol{\beta}_0)$, an open ball around $\boldsymbol{\beta}_0$, $l = 1,\ldots,q$.

(b) $\boldsymbol{\psi}(x;\boldsymbol{\beta})$ is twice continuously differentiable with respect to $\boldsymbol{\beta}$ in a neigbourhood $\mathrm{B}_r(\boldsymbol{\beta}_0)$ of $\boldsymbol{\beta}_0$, with a square integrable second derivative.

Then for general $\psi_l(x_i;\boldsymbol{\beta})$ and $\boldsymbol{\beta} \in \mathrm{B}_r(\boldsymbol{\beta}_0)$, for $l = 1,\ldots,q$,

$$\mathbb{E}_{\hat{f}}\left\{\psi_l(X;\boldsymbol{\beta})\right\} = \frac{1}{n}\sum_{i=1}^{n}\psi_l(x_i;\boldsymbol{\beta}) + \frac{1}{2}h^2\mu_2(K)\frac{1}{n}\sum_{i=1}^{n}\psi_l^{(2)}(x_i;\boldsymbol{\beta}) + O_p\left(h^4\right). \qquad (3.5)$$

See Appendix 3.A.1 for a proof.

As shown in section 3.4, the reweighted estimator provides an improved approximation to the moment conditions; in particular, the first term in (3.5) is zero.

3.3 Generalised empirical likelihood

Implied probabilities, obtained as a by-product of the GEL estimation, can be used to reweight the RPKDE so that the resultant density estimator better approximates conditions (3.2).

GEL is an estimation method for models based on moment conditions of the form (3.2); see inter alia Smith (1997), Imbens (2002) and Newey and Smith (2004), NS. To give a brief overview of GEL, introduce the *carrier* function $\rho(\cdot) : \mathscr{V} \to \mathbb{R}$, a *concave* real-valued scalar function defined on an open interval $\mathscr{V} \subseteq \mathbb{R}$ containing zero. Let

$\rho^{(k)}(v) = \partial^k \rho(v)/\partial v^k$ denote the k-th derivative of $\rho(\cdot)$, $k = 0, 1, 2, \ldots$. It will be convenient to impose the innocuous normalisation $\rho^{(1)}(0) = \rho^{(2)}(0) = -1$.

Special cases of GEL include empirical likelihood (EL), exponentially tilting (ET) and continuously updating estimators (CUE). These correspond to $\rho(v) = \ln(1-v)$ for $v < 1$, $\rho(v) = -\exp(v)$ and $\rho(v) = -v^2/2 - v$ respectively, all of which are members of the Cressie and Read (1984) family, $\rho(v) = \frac{-1}{\gamma+1}(1+\gamma v)^{\frac{\gamma+1}{\gamma}}$; see also NS.

Assume further that

Assumption 3

(a) $\boldsymbol{\beta}_0 \in \mathscr{B}$ is the unique solution to $\mathbb{E}_f\{\boldsymbol{\psi}(X_i;\boldsymbol{\beta})\} = 0$, $\mathscr{B}$ is compact and $\boldsymbol{\beta}_0$ is in the interior of $\mathscr{B}$.

(b) Matrix $\mathbf{V}_{\boldsymbol{\psi}} = \mathbb{E}_f\left\{\boldsymbol{\psi}(X_i;\boldsymbol{\beta}_0)\,\boldsymbol{\psi}(X_i;\boldsymbol{\beta}_0)^{\mathsf{T}}\right\}$ is positive definite.

(c) Matrix $\mathbb{E}_f\left\{\partial\boldsymbol{\psi}(X_i;\boldsymbol{\beta}_0)/\,\partial\boldsymbol{\beta}^{\mathsf{T}}\right\}$ has rank p.

(d) $\rho(v)$ is four times continuously differentiable in a neighbourhood of zero.

The class of GEL criteria considered here is defined as

$$P_n(\boldsymbol{\lambda},\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left[\rho\left(\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{\psi}(x_i;\boldsymbol{\beta})\right) - \rho(0)\right] \tag{3.6}$$

The estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$, solves the saddle point problem

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathscr{B}}{\operatorname{argmin}}\ \sup_{\boldsymbol{\lambda}\in\Lambda_n(\boldsymbol{\beta})} P_n(\boldsymbol{\lambda},\boldsymbol{\beta}), \tag{3.7}$$

where $\Lambda_n(\boldsymbol{\beta}) = \left\{\boldsymbol{\lambda} : \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{\psi}_i \in \mathscr{V}, i = 1,\ldots,n\right\}$. For given $\boldsymbol{\beta}$, the vector of auxiliary parameters (Lagrange multipliers), $\widehat{\boldsymbol{\lambda}} = \widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta})$, solves the first-order conditions

$$Q_{\lambda,n}(\widehat{\boldsymbol{\lambda}}(\boldsymbol{\beta})) = \frac{1}{n}\sum_{i=1}^{n}\rho^{(1)}\left(\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\boldsymbol{\psi}(x_i;\boldsymbol{\beta})\right)\boldsymbol{\psi}(x_i;\boldsymbol{\beta}) = \mathbf{0}. \tag{3.8}$$

The *implied probabilities* are then defined as

$$\hat{\pi}_i = \rho^{(1)}\left(\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\boldsymbol{\psi}\left(x_i,\widehat{\boldsymbol{\beta}}\right)\right)\Big/\sum_{j=1}^{n}\rho^{(1)}\left(\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\boldsymbol{\psi}\left(x_i,\widehat{\boldsymbol{\beta}}\right)\right). \tag{3.9}$$

By construction, the $\hat{\pi}_i$'s sum to unity over $i = 1,\ldots,n$. Furthermore, the first order conditions imply that $\sum_{i=1}^{n}\hat{\pi}_i\widehat{\boldsymbol{\psi}}_i = \mathbf{0}$, where $\widehat{\boldsymbol{\psi}}_i = \boldsymbol{\psi}\left(x_i;\widehat{\boldsymbol{\beta}}\right)$. It is this latter property that eliminates the first term in (3.5) when the expectation is taken over the reweighted density estimator.

As shown in NS, $\widehat{\boldsymbol{\beta}}$ is a consistent and asymptotically normal estimator of $\boldsymbol{\beta}_0$, the solution to the inner optimisation in (3.7) when $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ exists with probability approaching one, and $\widehat{\boldsymbol{\lambda}} = o_p\left(n^{-1/2}\right)$. The latter result of course holds when $\boldsymbol{\beta}_0$ is known; a proof is given in Appendix 3.A.2.

If $\boldsymbol{\beta}_0$ is known, then only the inner optimisation in (3.7) has to be carried out, and the implied probabilities are defined by (3.9) with $\boldsymbol{\beta}_0$ replacing $\widehat{\boldsymbol{\beta}}$.

As with the maximum likelihood estimation, GEL allows the construction of the tests for overidentifying restrictions that are similar to the classical likelihood ratio, Wald, and Lagrange multiplier tests. As the focus of this chapter is not on testing, we will only note that the normalised GEL criterion evaluated at the estimated parameters $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\lambda}}$, $2n\widehat{P}_n\left(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}\right)$, possesses a chi-square limiting distribution with $q-p$ degrees of freedom, χ^2_{q-p}. From the computational point of view, this statistic is the easiest as it is automatically produced by the optimisation routine. Other test statistics can be constructed as described in *inter alia* Smith (1997, pp. 510–514), Kitamura and Stutzer (1997, pp. 867–868) and Ramalho and Smith (2005).

Asymptotic expansions

Let $\hat{v}_i$ denote $\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\boldsymbol{\psi}\left(x_i, \widehat{\boldsymbol{\beta}}\right)$. As shown in Appendix 3.A.3, expanding the implied probabilities (3.9) around $\widehat{\boldsymbol{\lambda}} = 0$ gives

$$\hat{\pi}_i = \frac{1}{n} + \frac{1}{n}\left[\hat{v}_i - \frac{\rho^{(3)}(0)}{2}\hat{v}_i^2\right] + \frac{1}{n}\left[-\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}\left(x_i; \widehat{\boldsymbol{\beta}}\right) + \frac{\rho^{(3)}(0)}{2}\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{V}_{\boldsymbol{\psi}}\widehat{\boldsymbol{\lambda}}\right] + R_n^{[\pi]}, \quad (3.10)$$

where the remainder term is

$$R_n^{[\pi]} = \left[\hat{v}_i - \frac{\rho^{(3)}(0)}{2}\hat{v}_i^2\right] O_p\left(n^{-2}\right) - \frac{\rho^{(4)}(0)}{6}\hat{v}_i^3\left(1 + o_p(1)\right)\left[n^{-1} + O_p\left(n^{-2}\right)\right] + O_p\left(n^{-5/2}\right).$$

If $\boldsymbol{\beta}_0$ is known, expansion (3.10) is valid with $\boldsymbol{\beta}_0$ replacing $\widehat{\boldsymbol{\beta}}$ throughout, so that $v_i = \widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\boldsymbol{\psi}\left(x_i, \boldsymbol{\beta}_0\right)$ replaces $\hat{v}_i$.

To obtain an expansion for $\widehat{\boldsymbol{\lambda}}$ it will be convenient to introduce the transformation $\mathbf{w}_i = \mathbf{V}_{\boldsymbol{\psi}}^{-1/2}\boldsymbol{\psi}\left(x_i; \boldsymbol{\beta}_0\right)$, so that $\mathbb{E}\left\{\mathbf{w}_i\mathbf{w}_i^{\mathsf{T}}\right\} = I_q$, a $q \times q$ identity matrix. Further, let

$\boldsymbol{\theta}^{\mathsf{T}} = \widehat{\boldsymbol{\lambda}}^{\mathsf{T}} \mathbf{V}_{\boldsymbol{\psi}}^{1/2}$, and define

$$\alpha^{j_1 \dots j_k} = \mathbb{E}\left\{ \mathbf{w}_i^{j_1} \dots \mathbf{w}_i^{j_k} \right\} \qquad \text{and} \qquad A^{j_1 \dots j_k} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i^{j_1} \dots \mathbf{w}_i^{j_k} - \alpha^{j_1 \dots j_k},$$

where superscripts denote elements of the respective vector, e.g. $\mathbf{z}^j$ denotes the j-th element of vector $\mathbf{z}$, and the convention is used that if a superscript is repeated, a summation over that superscript is understood[2]. Note that $\alpha^j = 0$ and $\alpha^{jk} = \delta^{jk}$, where δ is the Kronecker delta.

In this notation, $v_i = \boldsymbol{\theta}^{\mathsf{T}} \mathbf{w}_i = \boldsymbol{\theta}^j \mathbf{w}_i^j$, $\widehat{\boldsymbol{\lambda}}^{\mathsf{T}} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}(x_i; \boldsymbol{\beta}_0) = \boldsymbol{\theta}^j A^j$ and $\widehat{\boldsymbol{\lambda}}^{\mathsf{T}} \mathbf{V}_{\boldsymbol{\psi}} \widehat{\boldsymbol{\lambda}} = \boldsymbol{\theta}^j \boldsymbol{\theta}^j$. An expansion for $\boldsymbol{\theta}$ is given in Propositions 1 and 2 for the cases where $\boldsymbol{\beta}_0$ is known and estimated, respectively; see also equation 3.1 in NS.

Proposition 1 Under Assumptions 2 and 3, if $\boldsymbol{\beta}_0$ is known, the vector of auxiliary parameters, $\boldsymbol{\theta}$, admits the following expansion

$$\begin{aligned}
\boldsymbol{\theta}^j = & -A^j + A^{jk} A^k + \frac{\rho^{(3)}(0)}{2} \alpha^{jkl} A^k A^l - A^{jk} A^{kl} A^l + \frac{\rho^{(3)}(0)}{2} A^{jkl} A^k A^l \\
& - \frac{\rho^{(3)}(0)}{2} A^{jk} \alpha^{klm} A^l A^m - \rho^{(3)}(0) \alpha^{jkl} A^{lm} A^k A^m \\
& - \frac{\left(\rho^{(3)}(0)\right)^2}{2} \alpha^{jkl} \alpha^{lmn} A^k A^m A^n - \frac{\rho^{(4)}(0)}{6} \alpha^{jklm} A^k A^l A^m + O_p\left(n^{-2}\right),
\end{aligned} \tag{3.11}$$

where $j, k, l, m, n \in \{1, \dots, q\}$. Proof is given in Appendix 3.A.4.

Note that for EL, with $\rho(v) = \ln(1-v)$, $\rho^{(j)}(v) = -(j-1)!(1-v)^{-j}$ and $\rho^{(j)}(0) = -(j-1)!$, (3.11) becomes

$$\begin{aligned}
\boldsymbol{\theta}^j = & -A^j + A^{jk} A^k - \alpha^{jkl} A^k A^l - A^{jk} A^{kl} A^l - A^{jkl} A^k A^l + A^{jk} \alpha^{klm} A^l A^m \\
& + 2\alpha^{jkl} A^{lm} A^k A^m - 2\alpha^{jkl} \alpha^{lmp} A^k A^m A^p + \alpha^{jklm} A^k A^l A^m + O_p\left(n^{-2}\right).
\end{aligned}$$

A similar expansion was obtained by DiCiccio, Hall, and Romano (1991) for EL for the mean; see also DiCiccio, Hall, and Romano (1988, sec. 3). Chen and Cui (2007) give analogous expansions for EL for generalised moment restrictions.

[2] For example, let $\mathbf{z}$ and $\mathbf{w}$ be p-dimensional vectors, then for $j, k \in \{1, \dots, p\}$ $\mathbf{z}^j \mathbf{w}^k$ is simply a product of the j-th element of $\mathbf{z}$ and k-th element of $\mathbf{w}$, whereas in expression $\mathbf{z}^j \mathbf{w}^j$ superscript j is repeated, hence a summation over j is understood: $\mathbf{z}^j \mathbf{w}^j = \mathbf{z}_1 \mathbf{w}_1 + \mathbf{z}_2 \mathbf{w}_2 + \dots + \mathbf{z}_p \mathbf{w}_p$. Two- and higher-dimensional arrays are indexed by an appropriate number of superscripts. For example, if A is a $q \times p$ matrix, expression $A^{lk} \mathbf{z}^k$, $l \in \{1, \dots, q\}$, represents the l-th element of the $q \times 1$ vector $A\mathbf{z}$, etc.

To obtain an expansion for $\boldsymbol{\theta}$ when $\boldsymbol{\beta}$ is estimated, let

$$\gamma^{j,k_1\ldots k_l} = \mathbb{E}\left\{\frac{\partial^l \mathbf{w}_i^j}{\partial\beta^{k_1}\cdots\partial\beta^{k_l}}\right\}, \quad \text{and} \quad \Gamma^{j,k_1\ldots k_l} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^l \mathbf{w}_i^j}{\partial\beta^{k_1}\cdots\partial\beta^{k_l}} - \gamma^{j,k_1\ldots k_l}.$$

Let $[\gamma^{j,r}]$ denote a $q \times p$ matrix with elements $\gamma^{j,r}$ and $\Omega = \left([\gamma^{j,r}]^{\mathsf{T}}[\gamma^{j,r}]\right)^{-1}$ be a $p \times p$ matrix with elements ω^{rs}.

As will be seen later, a contribution from the terms of order $\mathcal{O}_p\left(n^{-3/2}\right)$ in the expansion for $\boldsymbol{\theta}$ will be of smaller order than is of interest. Thus, Proposition 2 does not list the $\mathcal{O}_p\left(n^{-3/2}\right)$ terms in the expansion for $\boldsymbol{\theta}$ when $\boldsymbol{\beta}_0$ is estimated.

Proposition 2 Under Assumptions 2 and 3, if $\boldsymbol{\beta}_0$ is estimated, the vector of auxiliary parameters, $\boldsymbol{\theta}$, admits the following expansion

$$\begin{aligned}
\boldsymbol{\theta}^j = {} & -A^j + \gamma^{j,r}\gamma^{k,s}\omega^{rs}A^k \\
& + A^{jk}A^k - \gamma^{k,r}\gamma^{l,s}\omega^{rs}A^{jk}A^l - \gamma^{j,r}\gamma^{k,s}\omega^{rs}A^{kl}A^l + \gamma^{j,r}\gamma^{k,t}\gamma^{l,u}\gamma^{m,s}\omega^{tu}\omega^{rs}A^{mk}A^l \\
& + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}A^kA^l - \frac{\rho^{(3)}(0)}{2}\alpha^{mkl}\gamma^{j,r}\gamma^{m,s}\omega^{rs}A^kA^l \\
& \quad + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\gamma^{k,r}\gamma^{m,t}\gamma^{l,s}\gamma^{n,u}\omega^{su}\omega^{rt}A^mA^n \\
& \quad - \frac{\rho^{(3)}(0)}{2}\alpha^{okl}\gamma^{j,r}\gamma^{k,s}\gamma^{m,u}\gamma^{l,w}\gamma^{n,v}\gamma^{o,t}\omega^{wv}\omega^{rt}\omega^{su}A^mA^n \\
& \quad + \rho^{(3)}(0)\,\alpha^{nkl}\gamma^{j,r}\gamma^{l,s}\gamma^{m,v}\gamma^{n,t}\omega^{sv}\omega^{rt}A^kA^m - \rho^{(3)}(0)\,\alpha^{jkl}\gamma^{l,r}\gamma^{m,s}\omega^{rs}A^kA^m \\
& + \frac{1}{2}\gamma^{m,sv}\gamma^{j,r}\gamma^{k,u}\gamma^{l,w}\gamma^{m,t}\omega^{vw}\omega^{rt}\omega^{su}A^kA^l - \frac{1}{2}\gamma^{j,rs}\gamma^{k,t}\gamma^{l,u}\omega^{su}\omega^{rt}A^kA^l \\
& \quad - \gamma^{l,tv}\gamma^{j,r}\gamma^{k,u}\omega^{rt}\omega^{vu}A^kA^l + \gamma^{m,tv}\gamma^{j,r}\gamma^{m,s}\gamma^{k,u}\gamma^{l,w}\omega^{vw}\omega^{rt}\omega^{su}A^kA^l \\
& + \gamma^{k,s}\omega^{rs}\Gamma^{j,r}A^k + \gamma^{j,r}\omega^{rs}\Gamma^{k,s}A^k \\
& \quad - \gamma^{j,r}\gamma^{k,u}\gamma^{l,t}\omega^{rt}\omega^{su}\Gamma^{l,s}A^k - \gamma^{j,r}\gamma^{l,s}\gamma^{k,u}\omega^{rt}\omega^{su}\Gamma^{l,t}A^k + \mathcal{O}_p\left(n^{-3/2}\right),
\end{aligned} \tag{3.12}$$

where $j, k, l, m, n, o \in \{1, \ldots, q\}$ and $r, s, t, u, v, w \in \{1, \ldots, p\}$. See Appendix 3.A.5 for a proof.

Computational aspects

It should be noted that the solution to (3.8) does not always exist. In particular, there is no solution when zero is not in $\mathrm{CH}(\Psi_n(\boldsymbol{\beta}))$, the convex hull of $\Psi_n(\boldsymbol{\beta}) = \{\boldsymbol{\psi}(x_1;\boldsymbol{\beta}), \ldots, \boldsymbol{\psi}(x_n;\boldsymbol{\beta})\}$; see e.g. Kitamura (2006, sec. 8.1). When $\boldsymbol{\beta}_0$ is known, it is only required that $\mathbf{0} \in \mathrm{CH}\left(\Psi_n\left(\boldsymbol{\beta}_0\right)\right)$, but when $\boldsymbol{\beta}$ is estimated, zero must be in the convex hull of $\Psi_n\left(\boldsymbol{\beta}\right)$ for all $\boldsymbol{\beta}$ at which the GEL criterion is evaluated.

Example 2 Let $X_i \overset{\text{iid}}{\sim} N(0,1)$ and $\psi(x_i) = x_i$; i.e. we impose the constraint that the mean is zero. Then with probability 2^{-n+1} in a sample of size n, the x_i's will be either all positive or all negative, and there will be no solution to (3.8); see also Qin and Lawless (1994, example 2).

It is interesting to note that when all the sample values are positive, $P_n(\lambda)$ is a decreasing function of λ for both EL and ET, and the maximum is achieved at $\hat{\lambda} = -\infty$. (A similar argument applies to the case when all sample values are negative). The EL probabilities then become

$$\lim_{\substack{\lambda \to -\infty \\ \lambda \in \Lambda_n}} \pi_i = \frac{1}{x_i \sum_{j=1}^{n} \frac{1}{x_j}},$$

and $\sum_{i=1}^{n} \pi_i x_i = n \Big/ \sum_{j=1}^{n} \frac{1}{x_j} = H_x$, the harmonic average of x_i's. The harmonic average is greater than zero but smaller than the arithmetic average; i.e. $H_x \leq \bar{x}$, with equality if all x_i's are the same. (The harmonic average is also less than the geometric average).

ET in this case assigns weight one to the smallest observation (assuming no ties in the data) and zero to all other data points. CUE avoids this problem, but at a cost that some of the implied probabilities are *negative*.

■

One possibility then is to use *adjusted GEL*, whereby an artificial observation, $\boldsymbol{\psi}_{n+1}$, is added to the data such that zero is in the convex hull of $\Psi_n\left(\boldsymbol{\beta}\right) \cup \boldsymbol{\psi}_{n+1}\left(\boldsymbol{\beta}\right)$. In particular, adding $\boldsymbol{\psi}_{n+1} = -a_n \overline{\boldsymbol{\psi}}_n$, where $\overline{\boldsymbol{\psi}}_n = n^{-1} \sum_{i=1}^{n} \boldsymbol{\psi}\left(x_i; \boldsymbol{\beta}\right)$ and $a_n > 0$ ensures that $\mathbf{0}_q \in \mathrm{CH}\left(\Psi_n\left(\boldsymbol{\beta}\right) \cup \boldsymbol{\psi}_{n+1}\right)$; see Chen, Variyath, and Abraham (2008) and Liu and Chen (2010, sec. 3). Their suggestion is to set $a_n = \max(1, \ln(n)/2)$ and to use a trimmed mean of the $\boldsymbol{\psi}\left(x_i; \boldsymbol{\beta}\right)$'s in place of $\overline{\boldsymbol{\psi}}_n$ if desired.

The approach employed in our computations can be summarised as follows.

1. $\boldsymbol{\beta}_0$ known.

   IF $\mathbf{0}_q \in \mathrm{CH}\left(\Psi_n\left(\boldsymbol{\beta}_0\right)\right)$ use unadjusted GEL;

   ELSE use adjusted GEL.

2. $\boldsymbol{\beta}_0$ unknown.

   - Obtain a preliminary estimate of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{init}$, by GMM or another appropriate method;

```
        IF $\mathbf{0}_q \notin \text{CH}\left(\Psi_n(\widehat{\boldsymbol{\beta}}_{init})\right)$ use adjusted GEL.
    ELSE try estimation using unadjusted GEL;
            IF unadjusted GEL fails, use adjusted GEL.
```

Finally, the outer optimisation can also be challenging as several local minima may exist; see e.g. Guggenberger (2008). Whilst in low dimensions, a grid search over $\boldsymbol{\beta}$ may be feasible, as the dimension of $\boldsymbol{\beta}$ becomes large, stochastic optimisation methods such as simulated annealing can be used, perhaps combined with a direct search near the final value.

3.4 GEL-based KDE

The GEL-based KDE (GELKDE) is obtained by replacing the empirical probabilities n^{-1} by the implied probabilities (3.9), i.e.

$$\tilde{f}_\rho(x) = \sum_{i=1}^{n} \hat{\pi}_i K_h(x - x_i). \tag{3.13}$$

Because the GEL weights, $\hat{\pi}_i$, are not always non-negative, $\tilde{f}_\rho(x)$ may also take negative values (typically, in the tails of the distribution). In this case, one can 'shrink' the implied probabilities, for example, by transforming to

$$\begin{aligned} \hat{\pi}_i^\star &= \frac{1}{1+\epsilon_n}\hat{\pi}_i + \frac{\epsilon_n}{1+\epsilon_n} \cdot \frac{1}{n}, \\ &= \frac{\hat{\pi}_i + \epsilon_n/n}{\sum_{i=1}^n (\hat{\pi}_i + \epsilon_n/n)}, \qquad\qquad \text{where} \;\; \epsilon_n = -n \min\left[\min_{1\le i\le n} \hat{\pi}_i \,, 0\right]; \end{aligned}$$

see Smith (2010) and Antoine, Bonnal, and Renault (2007). Consequently, $\hat{\pi}_i^\star \geq 0$ and sum to one by construction, thus ensuring that $\tilde{f}_\rho(\cdot)$ is a proper density.

Alternatively, one can simply take a positive part of $\tilde{f}_\rho(x)$, $\tilde{f}_\rho^+(x) = \max\left(\tilde{f}_\rho(x),\ 0\right)$, to be the final estimate. In this case $\tilde{f}_\rho^+(\cdot)$ should be renormalized to ensure it integrates to one. However, as the latter is computationally difficult, we prefer to shrink the implied probabilities as detailed above if any are negative.

Because $\sum_{i=1}^n \hat{\pi}_i \widehat{\boldsymbol{\psi}}_i = \mathbf{0}$, we will see that GELKDE approximates the constraints (3.2) better than RPKDE. Since $\mathbb{E}_{\tilde{f}_\rho}\{\psi_l(X;\boldsymbol{\beta})\} = \sum_{i=1}^n \hat{\pi}_i \int \psi_l(x_i + zh)K(z)dz$, when the mean is known to be μ, $\psi_l(x_i) = x_i - \mu$, $\mathbb{E}_{\tilde{f}_\rho}\{X\} = \sum_{i=1}^n \hat{\pi}_i x_i = \mu$, i.e. the constraint is satisfied exactly (provided the solution to (3.8) exists). Note also that $\mathbb{E}_{\tilde{f}_\rho}\{X^2\} = \sum_{i=1}^n \hat{\pi}_i x_i^2 + h^2\mu_2(K)$. Hence if the constraint is $\mathbb{E}\{X^2\} = m^2$, say, although it will not be met exactly, $\mathbb{E}_{\tilde{f}_\rho}\left\{X^2\right\} = m^2 + h^2\mu_2(K)$, the GELKDE approximates this constraint better than RPKDE; cf. Example 1.

For general $\psi_l(x_i;\boldsymbol{\beta})$, $l = 1, \ldots, q$, $\boldsymbol{\beta} \in \mathrm{B}_r\left(\boldsymbol{\beta}_0\right)$, we obtain

$$\mathbb{E}_{\tilde{f}_\rho}\left\{\psi_l(X;\boldsymbol{\beta})\right\} = \left(\frac{1}{2}h^2\mu_2(K)\frac{1}{n}\sum_{i=1}^{n}\psi_l^{(2)}(x_i;\boldsymbol{\beta})\right)\left(1 + o_p\left(1\right)\right) + O_p\left(h^4\right). \tag{3.14}$$

See Appendix 3.A.1 for a proof. Note that the first term in (3.14) is the same as the second term in (3.5), whereas the first term in (3.5) vanishes. Hence in general GELKDE provides a better approximation to moment conditions than RPKDE.

Bias and variance

Since the GEL estimator is defined implicitly, the exact expectation of GELKDE cannot be obtained[3]. Hence, an asymptotic analysis is required.

As shown in Appendix 3.A.6, using expansions for the implied probabilities and auxiliary parameters, an asymptotic approximation to the expectation of GELKDE up to an order $O(n^{-1}h^2)$ is given by

$$\begin{aligned}\mathbb{E}\left\{\tilde{f}_\rho(x)\right\} = \mathbb{E}\left\{\hat{f}(x)\right\} &+ n^{-1}k_\rho\left[-\mathbf{w}^j(x)\mathbf{w}^j(x) + \alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) + q\right]f(x)\\ &+ n^{-1}h^2\frac{1}{2}\mu_2(K)k_\rho\left[-C_1 + \alpha^{jkl}\delta^{kl}C_2^j + qf^{(2)}(x)\right] + O\left(n^{-1}h^4\right),\end{aligned} \tag{3.15a}$$

where $k_\rho = 1 + \rho^{(3)}\left(0\right)/\,2$, $C_1 = \frac{d^2}{dv^2}\left[\mathbf{w}(v)^j\mathbf{w}^j(v)f(v)\right]\Big|_{v=x}$, $C_2^j = \frac{d^2}{dv^2}\left[\mathbf{w}^j(v)f(v)\right]\Big|_{v=x}$ and $\mathbf{w}(x) = \mathbf{w}(x;\boldsymbol{\beta}_0)$.

Note that for any carrier function with $\rho^{(3)}\left(0\right) = -2$, e.g. EL, $k_\rho = 0$ and thus the n^{-1} bias term in (3.15a) vanishes. Note that for ET $k_\rho = 1/2$, whereas for CUE k_ρ is unity. The derivations indicate that under sufficient smoothness EL-based KDE will have the same expectation as $\hat{f}$, asymptotically, to a higher order than $O(n^{-1}h^2)$.

It is useful to note that in terms of the original $\boldsymbol{\psi}\left(x\right) = \boldsymbol{\psi}\left(x;\boldsymbol{\beta}_0\right) = \mathbf{V}_{\boldsymbol{\psi}}^{1/2}\mathbf{w}(x)$, the n^{-1} bias terms can be written as $\mathbf{w}^j(x)\mathbf{w}^j(x) = \boldsymbol{\psi}^\mathsf{T}\left(x\right)\mathbf{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\psi}\left(x\right)$ and $\alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) = \boldsymbol{\psi}^\mathsf{T}\left(x\right)\mathbf{V}_{\boldsymbol{\psi}}^{-1}\mathbb{E}\left\{\boldsymbol{\psi}_i\boldsymbol{\psi}_i^\mathsf{T}\mathbf{V}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\psi}_i\right\}$. These expressions may be easier to implement computationally as they avoid taking a square root of the variance matrix $\mathbf{V}_{\boldsymbol{\psi}}$.

[3] Even in the simplest case with only one constraint and quadratic carrier function expression for implied probabilities involves a ratio; see Appendix 3.B.

The asymptotic variance of GELKDE is given by

$$\mathbb{Var}\left\{\tilde{f}_\rho(x)\right\} = \mathbb{Var}\left\{\hat{f}(x)\right\} - n^{-1}\mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x) \\ - n^{-1}h^2\mu_2(K)\mathbf{w}^j(x)C_2^j f(x) + o\left(n^{-1}h^4\right). \quad (3.15b)$$

As $\mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x)$ is non-negative, there is always an $1/n$ *reduction* in variance, which does *not* depend on the GEL carrier function.

From (3.15a) and (3.15b), we obtain the expressions for the integrated squared bias (ISB) and integrated variance (IVar):

$$\mathbb{ISB}\left\{\tilde{f}_\rho(\cdot)\right\} = \mathbb{ISB}\left\{\hat{f}(\cdot)\right\} + n^{-1}h^2\mu_2(K)k_\rho I_1 + o\left(n^{-1}h^4\right), \\ \text{where} \quad I_1 = \int_{\mathbb{R}} \left[-\mathbf{w}^j(x)\mathbf{w}^j(x) + \alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) + q\right] f^{(2)}(x)f(x)dx, \quad (3.15c)$$

$$\text{and} \quad \mathbb{IVar}\left\{\tilde{f}_\rho(\cdot)\right\} = \mathbb{IVar}\left\{\hat{f}(\cdot)\right\} - n^{-1}\int_{\mathbb{R}} \mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x)dx \\ - n^{-1}h^2\mu_2(K)I_2 + o\left(n^{-1}h^4\right), \quad (3.15d)$$

where $I_2 = \int_{\mathbb{R}} \mathbf{w}^j(x)C_2^j f(x)dx$. Thus, asymptotically the effect entering via variance dominates and GELKDE enjoys a $1/n$ reduction in mean integrated squared error, viz.

$$\mathbb{MISE}\left\{\tilde{f}_\rho(\cdot)\right\} = \mathbb{MISE}\left\{\hat{f}(\cdot)\right\} - n^{-1}\int_{\mathbb{R}} \mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x)dx \\ + n^{-1}h^2\mu_2(K)\left[k_\rho I_1 - I_2\right] + o\left(n^{-1}h^4\right). \quad (3.15e)$$

Formally, the following proposition can be stated.

Proposition 3 If $\boldsymbol{\beta}_0$ is known, the mean, variance, integrated squared bias, integrated variance, and mean integrated squared error of GELKDE are given by equations (3.15a)–(3.15e).

See Appendix 3.A.6 for a proof.

Example 3 Suppose $X_i \overset{\text{iid}}{\sim} N(0,1)$. Let $\phi(x)$ denote the standard normal density. Since $d^2\phi(x)/\,dx^2 = (x^2-1)\phi(x)$, it is straightforward to compute the leading constants in the integrated squared bias and integrated variance directly. For the variance, the $1/n$ term does not depend on the kernel or the carrier function. The $n^{-1}h^2$ term in the integrated squared bias is $\mu_2(K)k_\rho I_1$. Assuming that a Gaussian kernel is used, $\mu_2(K) = 1$. Also, k_ρ

is known for a given choice of carrier function. The following table presents the leading constants for three examples.

Moment constraints	$\boldsymbol{\psi}(x) =$	Leading constants in	
		$\mathbb{ISB}\left\{\tilde{f}_\rho(\cdot)\right\}$, I_1	$\mathbb{IV}\mathrm{ar}\left\{\tilde{f}_\rho(\cdot)\right\}$, $-\int_{\mathbb{R}} \mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x)dx$
1. Known mean	x	$-\frac{3}{8\sqrt{\pi}} \approx -0.2116$	$-\frac{1}{4\sqrt{\pi}} \approx -0.1410$
2. Known mean and variance	$\left[x, x^2-1\right]^\mathsf{T}$	$+\frac{15}{32\sqrt{\pi}} \approx +0.2645$	$-\frac{7}{16\sqrt{\pi}} \approx -0.2468$
3. Known mean and third moment	$\left[x, x^3\right]^\mathsf{T}$	$-\frac{23}{64\sqrt{\pi}} \approx -0.2028$	$-\frac{13}{32\sqrt{\pi}} \approx -0.2292$

Note that in case 2 $\mathbb{ISB}\left\{\tilde{f}_\rho(\cdot)\right\} \geq \mathbb{ISB}\left\{\hat{f}(\cdot)\right\}$.

The expectation of the difference between GELKDE and RPKDE in these cases is of the form $n^{-1}k_\rho P_m(x)\phi(x)$, where $P_m(x)$ is a polynomial in x. In case 1, the polynomial is $-x^2+1$. Figure 3.1 shows the simulated difference between $\mathbb{E}\left\{\tilde{f}_{cue}\right\}$ for case 1, incorporating the know mean constraint, and $\mathbb{E}\left\{\hat{f}\right\}$ scaled up by the sample size, $n = 1,000$, (solid line) and the curve $(-x^2+1)\phi(x)$ (dashed line) to which this difference should converge as n approaches infinity. The two curves are quite close agreeing with our theoretical results.

■

An immediate consequence of Proposition 3 is that the asymptotically optimal bandwidth, h_{AMISE}, given in (3.4), remains unchanged. Recall that h_{AMISE} minimises the two leading terms in (3.3), which are also the leading terms in (3.15e). Thus setting $h = cn^{-1/5}$ the first two terms in $\mathbb{MISE}\left\{\hat{f}(\cdot)\right\}$ are of the same order, $n^{-4/5}$, with the next term of order $1/n$, which is only moderately faster than $n^{-4/5}$. In fact, the $\mathcal{O}(n^{-1})$ term in (3.3) is $n^{-1}R(f)$, where $R(f) = \int_{\mathbb{R}} f^2(x)dx$. Hence, in small and moderate samples, the reduction in variance can be substantial; however, this may be offset by the effect on ISB, which is of order $n^{-1}h^2$. Simulation evidence presented in the next section suggests that for moderate and large sample sizes the reduction in variance dominates, but in very small samples MISE may increase.

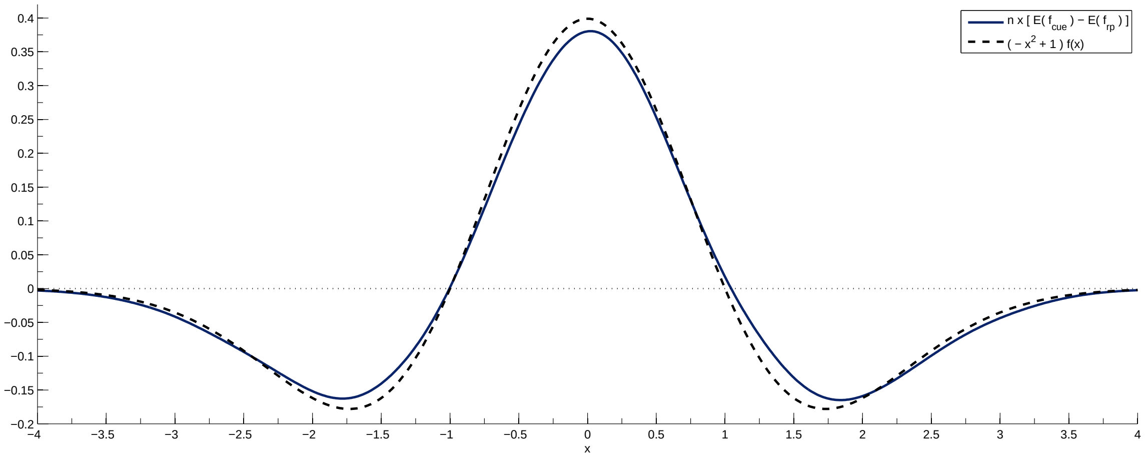

Figure 3.1: $1/n$ bias term of the CUE-based KDE with $\boldsymbol{\psi}\left(x\right) = x$

$\mathbb{E}\left\{\tilde{f}_{cue}(x)\right\}$ is simulated using one million replications and the optimal bandwidth, $h_{MISE} = 0.2723$; $\mathbb{E}\left\{\hat{f}(x)\right\}$ is exact.

Estimated parameters

If the vector of parameters, $\boldsymbol{\beta}_0$, is *estimated*, the MSE of GELKDE can be obtained following the same steps as above, the only difference being that extra terms related to the estimation of $\boldsymbol{\beta}_0$ enter.

Thus, the expectation of GELKDE is given by

$$\mathbb{E}\left\{\tilde{f}_\rho(x)\right\} = \mathbb{E}\left\{\hat{f}(x)\right\} + n^{-1}B_1(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)B_2(x) + O\left(n^{-3/2}\right), \quad (3.16a)$$

where the expressions for $B_1(x)$ and $B_2(x)$ are given in equations (3.29) and (3.30) in Appendix 3.A.7. $B_1(x)$ and $B_2(x)$ contain all of the terms as those in (3.15a) plus extra terms due to estimation of the unknown $\boldsymbol{\beta}_0$. However, unlike the known $\boldsymbol{\beta}_0$ case, in general it is no longer true that the n^{-1} term may be set to zero for a particular choice of carrier function.

The variance of GELKDE is

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left\{\tilde{f}_\rho(x)\right\} &= \mathbb{V}\mathrm{ar}\left\{\hat{f}(x)\right\} - n^{-1}\mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x) + n^{-1}\gamma^{j,s}\gamma^{k,r}\omega^{rs}\mathbf{w}^j(x)\mathbf{w}^k(x)f^2(x) \\ &- n^{-1}h^2\mu_2(K)\mathbf{w}^j(x)C_2^j(x)f(x) + n^{-1}h^2\mu_2(K)\gamma^{j,s}\gamma^{k,r}\omega^{rs}\mathbf{w}^j(x)C_2^k(x)f(x) + O\left(n^{-3/2}\right).\end{aligned} \quad (3.16b)$$

It immediately follows that

$$\mathbb{ISB}\left\{\tilde{f}_\rho(\cdot)\right\} = \mathbb{ISB}\left\{\hat{f}(\cdot)\right\} + n^{-1}h^2\mu_2(K)\int_{\mathbb{R}} B_1(x)f^{(2)}(x)f(x)dx + O\left(n^{-3/2}\right), \quad (3.16c)$$

and

$$\begin{aligned}\mathbb{IVar}\left\{\tilde{f}_\rho(\cdot)\right\} &= \mathbb{IVar}\left\{\hat{f}(\cdot)\right\} - n^{-1}\int_{\mathbb{R}} \mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x)dx \\ &+ n^{-1}\gamma^{j,s}\gamma^{k,r}\omega^{rs}\int_{\mathbb{R}} \mathbf{w}^j(x)\mathbf{w}^k(x)f^2(x)dx - n^{-1}h^2\mu_2(K)\int_{\mathbb{R}} \mathbf{w}^j(x)C_2^j(x)f(x)dx \\ &+ n^{-1}h^2\mu_2(K)\gamma^{j,s}\gamma^{k,r}\omega^{rs}\int_{\mathbb{R}} \mathbf{w}^j(x)C_2^k(x)f(x)dx + O\left(n^{-3/2}\right). \quad (3.16d)\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{MISE}\left\{\tilde{f}_\rho(\cdot)\right\} &= \mathbb{MISE}\left\{\hat{f}(\cdot)\right\} - n^{-1}\int_{\mathbb{R}} \mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x)dx \\ &+ n^{-1}\gamma^{j,s}\gamma^{k,r}\omega^{rs}\int_{\mathbb{R}} \mathbf{w}^j(x)\mathbf{w}^k(x)f^2(x)dx - n^{-1}h^2\mu_2(K)\int_{\mathbb{R}} \mathbf{w}^j(x)C_2^j(x)f(x)dx \\ &+ n^{-1}h^2\mu_2(K)\gamma^{j,s}\gamma^{k,r}\omega^{rs}\int_{\mathbb{R}} \mathbf{w}^j(x)C_2^k(x)f(x)dx + n^{-1}h^2\mu_2(K)\int_{\mathbb{R}} B_1(x)f^{(2)}(x)f(x)dx \\ &+ O\left(n^{-3/2}\right). \quad (3.16e)\end{aligned}$$

Proposition 4 If $\boldsymbol{\beta}_0$ is unknown, the mean, variance, integrated squared bias, integrated variance, and mean integrated squared error of GELKDE are given by equations (3.16a)–(3.16e).

The proof follows the same steps as the proof of Proposition 3 and is given in Appendix 3.A.7.

Bias correction

Although the contribution from the $1/n$ bias terms in (3.15a) and (3.16a) to the MISE of GELKDE is of order $O(n^{-1}h^2)$, of a lower order than the contribution from the variance, in small samples the bias effect can be substantial. As the direction of the bias is not known *a priori*, unless the true density is known, it may be advisable to bias-correct GELKDE by estimating and subtracting the $1/n$ bias term. To be specific, the bias-corrected GELKDE is defined as

$$\tilde{f}_\rho^{bc}(x) = \tilde{f}_\rho(x) - n^{-1}\widetilde{B}_1(x)\tilde{f}_\rho(x), \quad (3.17)$$

where $\widetilde{B}_1(x)$ is a suitable estimate of $B_1(x)$. In the case when $\boldsymbol{\beta}_0$ is known, the bias correction is an estimate of $k_\rho\left[-\mathbf{w}^j(x)\mathbf{w}^j(x)+\alpha^{jkl}\delta^{kl}\mathbf{w}^j(x)+q\right]$. We suggest using implied probabilities to obtain weighted estimators for $\mathbf{V}_{\boldsymbol{\psi}}$ and other moments entering $B_1(x)$; see e.g. Smith (2010, sec. 3). In particular, a plug-in estimator of $\mathbf{w}^j(x)\mathbf{w}^j(x)$ is $\boldsymbol{\psi}^\mathsf{T}(x)\,\widetilde{\mathbf{V}}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\psi}(x)$, where $\widetilde{\mathbf{V}}_{\boldsymbol{\psi}}$ is a weighted sample covariance matrix; and a plug-in estimate of $\alpha^{jkl}\delta^{kl}\mathbf{w}^j(x)$ can be computed as $\boldsymbol{\psi}^\mathsf{T}(x)\,\widetilde{\mathbf{V}}_{\boldsymbol{\psi}}^{-1}\left(\frac{1}{n}\sum_{i=1}^n\hat{\pi}_i\boldsymbol{\psi}_i\boldsymbol{\psi}_i^\mathsf{T}\widetilde{\mathbf{V}}_{\boldsymbol{\psi}}^{-1}\boldsymbol{\psi}_i\right)$. To ensure that the bias-corrected estimate is a density, any negative values can be set to zero and renormalised as necessary.

3.5 Monte-Carlo study

3.5.1 Known parameters

Consider first the case where $X_i\overset{\text{iid}}{\sim}N(0,1)$, $i=1,\dots,n$, and a Gaussian kernel, $K(z)=\phi(z)$, is used, where $\phi(\cdot)$ denotes the standard normal density. Although very simple, this setup is appealing because the integrated mean squared error of the unweighted KDE can be evaluated analytically and is given by

$$\begin{aligned}\mathbb{MISE}\left\{\hat{f}(x)\right\}&=\frac{1}{2\sqrt{\pi}}\left[\frac{1}{\sqrt{1+h^2}}-\frac{2\sqrt{2}}{\sqrt{2+h^2}}+1\right]+\frac{1}{2\sqrt{\pi}}\frac{1}{n}\left[\frac{1}{h}-\frac{1}{\sqrt{1+h^2}}\right]\\&=\frac{1}{2\sqrt{\pi}}\left[\frac{1}{nh}+\frac{n-1}{n\sqrt{1+h^2}}-\frac{2\sqrt{2}}{\sqrt{2+h^2}}+1\right],\end{aligned}\tag{3.18}$$

where the first summand is the ISB and the second the IVar; see Fryer (1976).

The asymptotically optimal bandwidth in this case is $h_{AMISE}=(4/3)^{1/5}n^{-1/5}$, and the optimal AMISE is $\frac{3(4/3)^{4/5}}{32\sqrt{\pi}}n^{-4/5}+\frac{(4/3)^{-1/5}}{2\sqrt{\pi}}n^{-4/5}$, where the first term is the asymptotic ISB and the second the IVar. The exact MISE-minimising bandwidth h_{MISE} is obtained by minimising (3.18) with respect to h for a given sample size, n. The setup is thus the most favourable for PRKDE.

It is interesting to note that the MISE-minimising bandwidth approaches its asymptotic value from above. Even when $n=1,000,000$ the exact MISE-minimising bandwidth is still approximately 0.16% greater than the asymptotically optimal value. The top panel of Figure 3.2 shows h_{MISE} inflated by $n^{1/5}$; the horisontal dashed line is drawn at the level of the constant in h_{AMISE}, $(4/3)^{1/5}\approx1.0592$. The bottom panel depicts the behaviour of the optimal MISE and its components[4].

[4] Recall that the second term in the asymptotic IVar is of order n^{-1}, in this case $n^{-1}/(2\sqrt{\pi})$ and when $n=1,000,000$ equals 0.0178, approximately the discrepancy between $n^{-4/5}\times$ IVar and

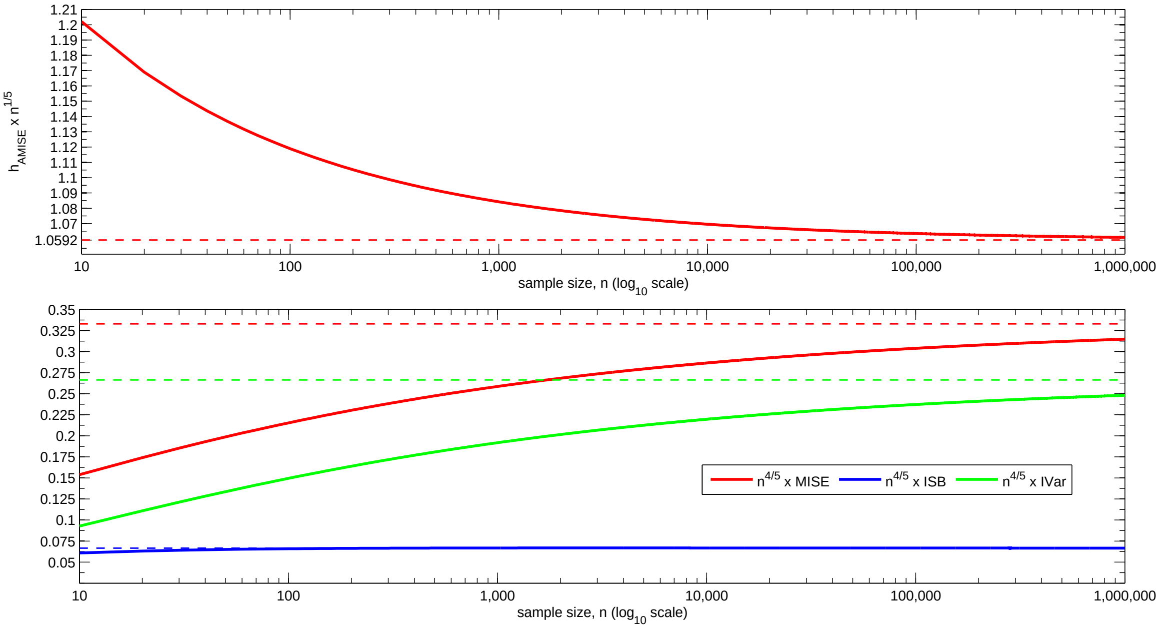

Figure 3.2: Exact MISE-minimising bandwidth, ISB, IVar and MISE

The moment conditions studied here are those considered in Example 3, viz.

1. Known mean. $\mathbb{E}\left\{X\right\} = 0$.
2. Known mean and variance. $\mathbb{E}\left\{X\right\} = 0$, $\mathbb{Var}\left\{X\right\} = 1$.
3. Known mean and third moment. $\mathbb{E}\left\{X\right\} = 0$, $\mathbb{E}\left\{X^3\right\} = 0$.

For each case, the performance of the unweighted estimator is compared to the GEL-based estimators (3.13) using three popular carrier functions: $\rho(v) = \ln(1-v)$ (EL), $\rho(v) = -\exp(v)$ (ET) and $\rho(v) = -\frac{1}{2}v^2 - v$ (CUE). Unless stated otherwise, all the results presented below are based on $100,000$ replications; multiple-segment trapezoidal rule numerical integration is used to obtain the ISB, IVar and MISE of GELKDE.

Figure 3.3 shows the relative performance of GELKDE for small and moderate samples. In this and the subsequent figures **red** lines correspond to the quadratic carrier function (CUE), **blue** lines—exponential (ET) and **green**—logarithmic carrier functions (EL). Solid lines show the performance of the original GELKDE[5], whereas dashed lines represent bias-corrected estimates, see (3.17). Since there is no $1/n$ term for EL, these two lines coincide.

$(4/3)^{-1/5}\Big/(2\sqrt{\pi}) \approx 0.2663$. The second term in the asymptotic ISB is $\left(-7/\left(128\sqrt{\pi}\right)\right)h^6$. Hence $n^{-4/5} \times$ ISB is approximately $-0.0436n^{-2/5}$ away from the asymptotic value, $3\,(4/3)^{4/5}\Big/(32\sqrt{\pi}) \approx 0.0666$.

[5] In all cases, implied probabilities are shrunk where necessary to ensure $\tilde{f}$ is nonnegative.

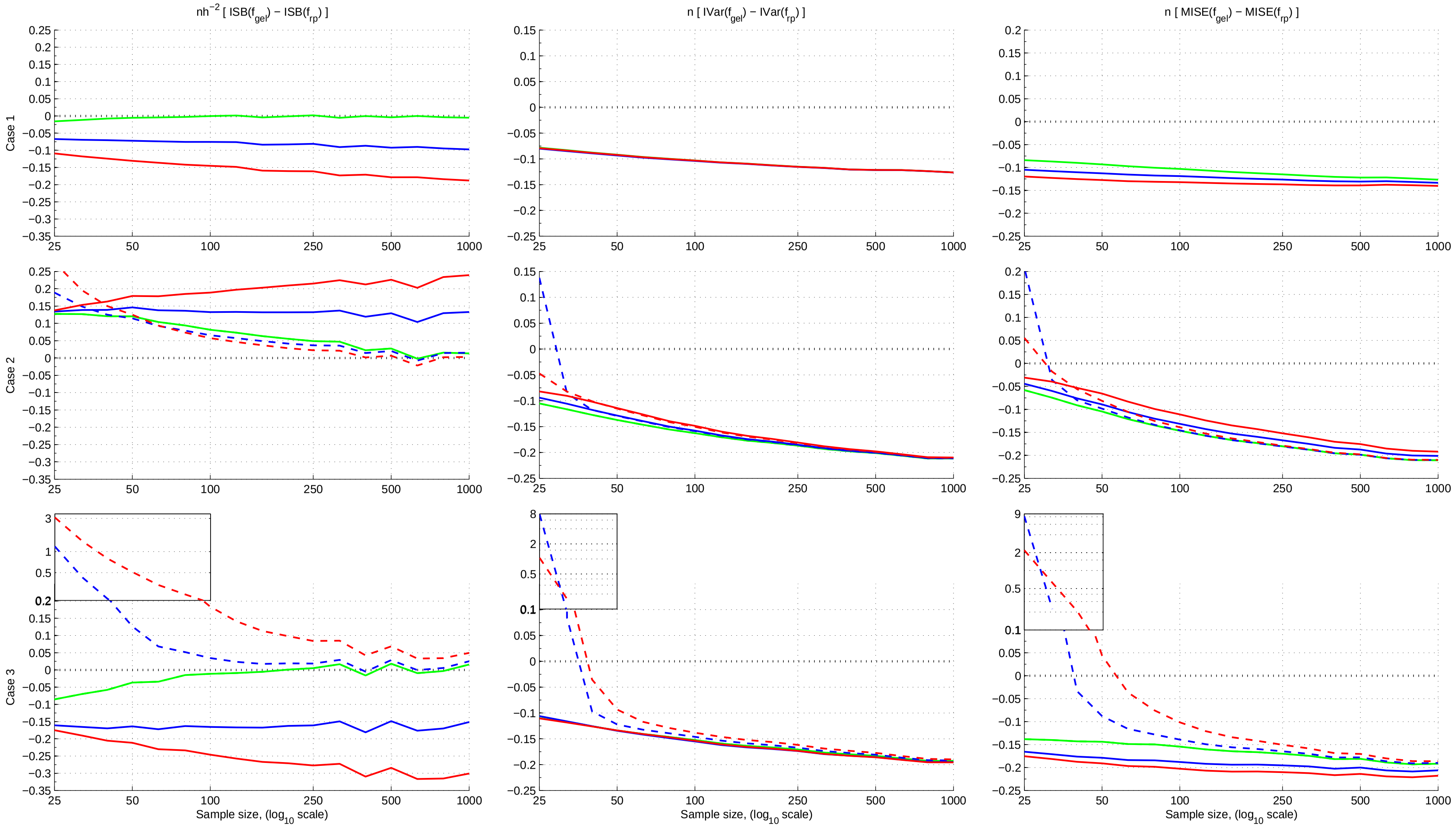

Figure 3.3: Performance of GELKDE with known parameters.

Horizontal axes (sample size) are shown on a common logarithm scale. The columns show the difference between the integrated squared bias, variance and mean squared error of GELKDE and RPKDE scaled by sample size in the case of IVar and MISE, and by nh^{-2} in the case of ISB. Thus, the lines should tend to the respective constants computed in Example 3 as sample size increases. It should be noted that the remainder term in all graphs is of order $h^2 \overset{asy}{\sim} n^{-1/5}$, which may be substantial relative to main constants for small sample sizes.

The results confirm the conclusions of the previous section. For small sample sizes the reduction in MISE is smaller than the asymptotic value. In this example, an increase in bias for case 2 is not big enough to offset the reduction in variance even when only 25 observations are available, but as shown below, this is not always the case.

The jagged appearance of the lines showing the difference between ISBs of GELKDE and RPKDE is largely due to simulation error. As the differences in ISBs are being inflated by nh^{-2}, which for $n = 1,000$ is about $13,500$, and the quantities themselves are small, they need to be estimated with a very high precision, which is costly in terms of computing time. The results presented for case 1 are obtained with $300,000$ replications, and it can be seen that the lines are almost perfectly smooth.

Some departures from normality

We also examine the performance of GELKDE when the distribution of the data is non-normal. The only additional information used by GEL-based estimators is that the mean is known.

Figure 3.4 shows simulation results when X_i, $i = 1, \ldots, n$, are drawn from a Student's t-distribution with degrees of freedom $\nu = 16, 8, 4$, and 2 (from top to bottom, respectively). The asymptotically optimal bandwidth is used as—to the best of our knowledge—the exact MISE cannot be obtained analytically in this case. Qualitatively, the performance of GELKDE is similar to the case when the data is Gaussian. However, as the tails become heavier, the reduction in variance is smaller.

To examine other departures from normality, we consider mixtures of normal densities which provide a powerful tool to study the performance of kernel estimators as they can approximate many interesting densities. An additional attraction is that if the kernel function is the standard normal pdf, the exact MISE of the unweighted KDE can be computed analytically. Marron and Wand (1992) derive an expression for the exact MISE and construct fifteen examples of mixture densities which have since been widely used in the literature. The three densities selected here are the skewed unimodal density (#2), the strongly skewed density (#3) and the outlier density (#5). The mixtures are

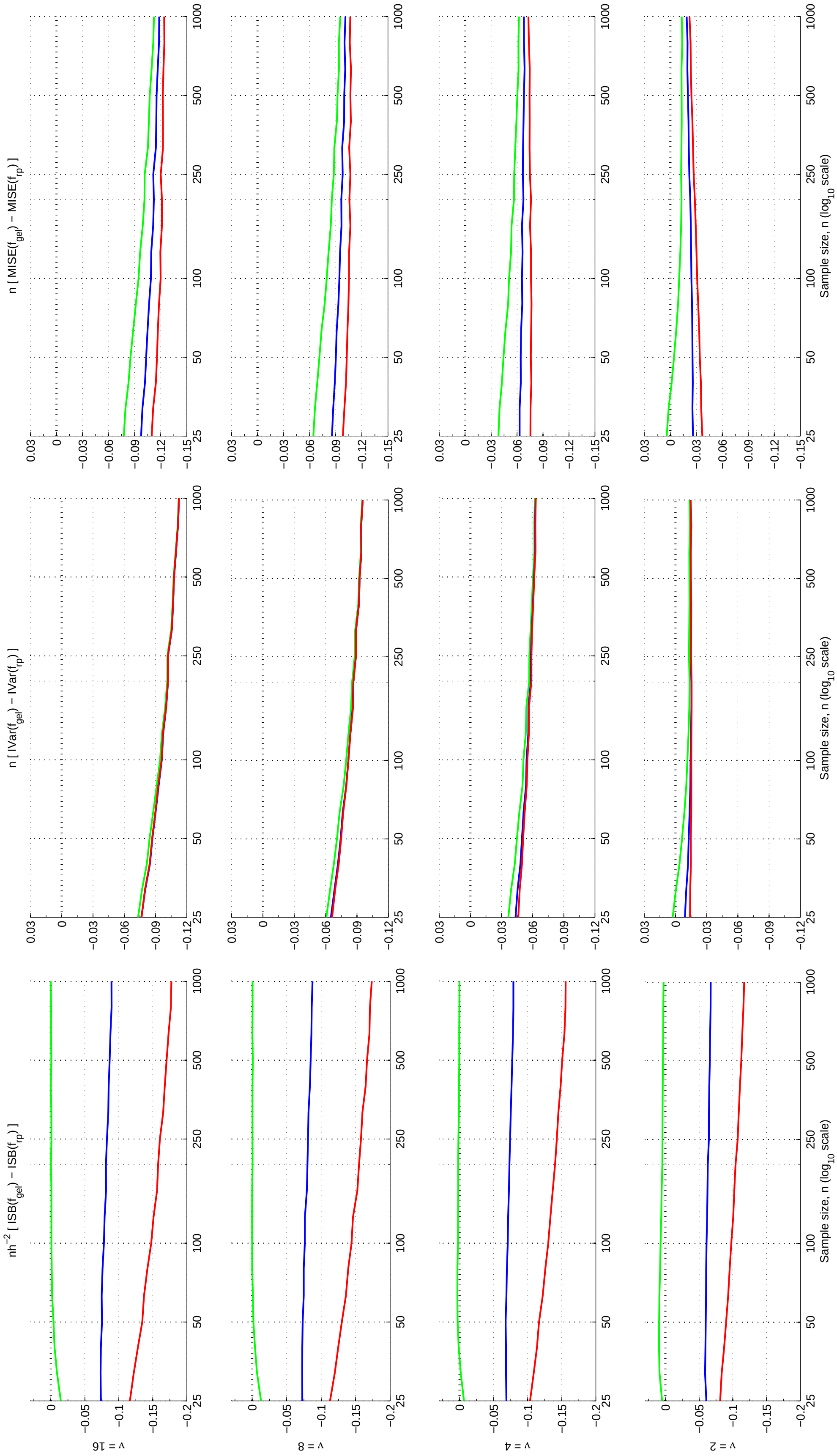

Figure 3.4: GELKDE with heavy-tailed data.

constructed as $\frac{1}{5}N\left(0,1\right)+\frac{1}{5}N\left(\frac{1}{2},\left(\frac{2}{3}\right)^2\right)+\frac{3}{5}N\left(\frac{13}{12},\left(\frac{5}{9}\right)^2\right)$, $\sum_{k=0}^{7}\frac{1}{8}N\left(3\left[\left(\frac{2}{3}\right)^k-1\right],\left(\frac{2}{3}\right)^{2k}\right)$ and $\frac{1}{10}N\left(0,1\right)+\frac{9}{10}N\left(0,\left(\frac{1}{10}\right)^2\right)$ respectively. The upper three subplots of Figure 3.5 show these densities alongside the normal density with the same mean and variance (dashed line) plotted for reference. Densities #2 and #3 are constructed to resemble the extreme value and the lognormal densities respectively. The outlier density is similar to the normal but with 10% of observations being strong outliers. The exact MISE-minimising bandwidth and the resulting optimal ISB, IVar and MISE are shown in the middle and bottom subplots respectively.

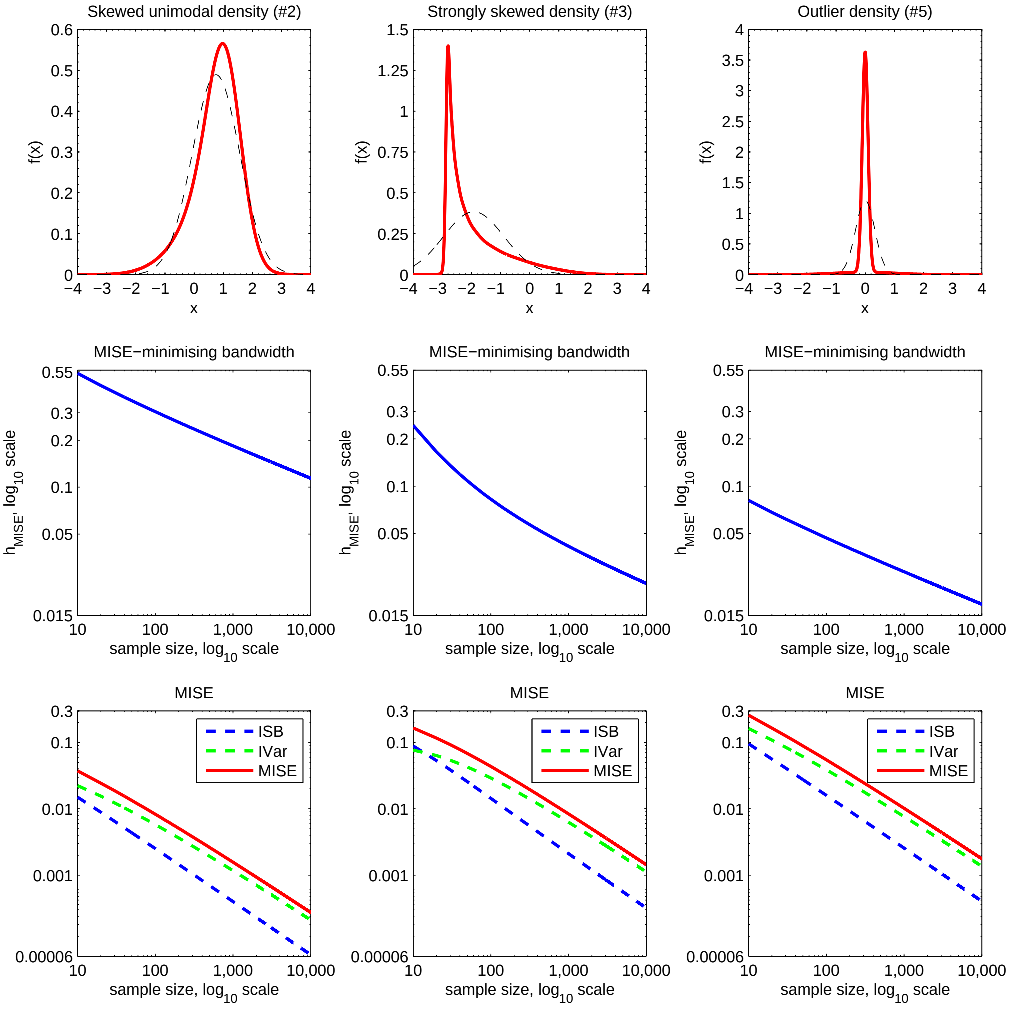

Figure 3.5: Selected normal mixture densities.

Figure 3.6 shows the MISE of GELKDE incorporating the extra information of a

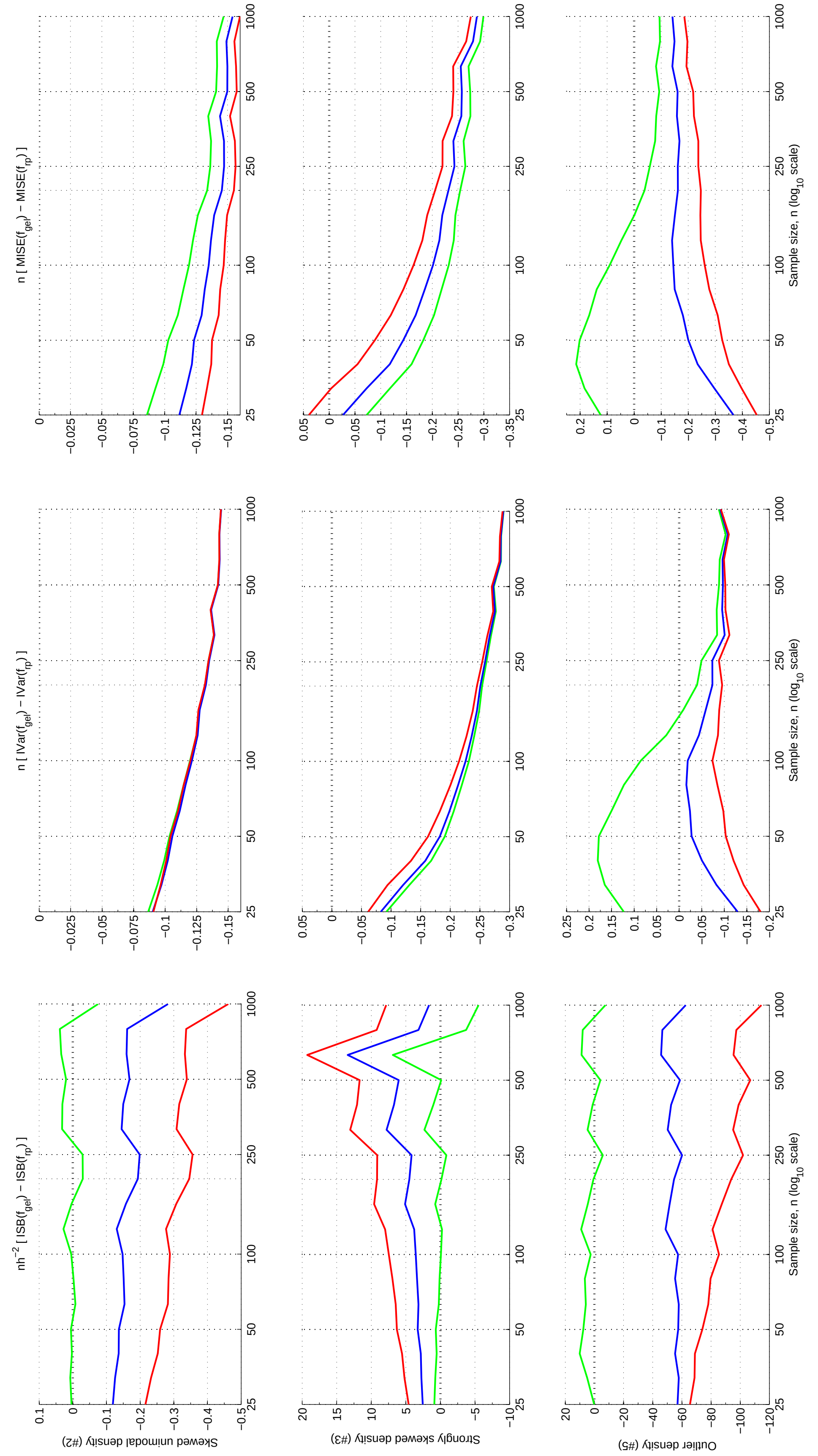

Figure 3.6: GELKDE with skewed and contaminated data.

known mean. For the mildly skewed density (#2) the results are qualitatively similar to case 1 in Figure 3.3. For the strongly skewed density, the MISE of GELDKE overshoots the MISE of RPKDE for small sample sizes if CUE is used. Otherwise, the results are consistent with those suggested by the asymptotic analysis.

The most interesting results are for the outlier density. Relative to RPKDE, the EL-based estimator has a significantly larger variance and MISE for small and moderate sample sizes. CUE and ET estimators, however, perform remarkably well in reducing both bias and variance.

3.5.2 Estimated parameters

Finally, we consider the case when parameters are estimated. The two examples of moment conditions are:

4. Unknown mean and known variance, $\mathbb{V}\mathrm{ar}\left\{X\right\} = 1$.
   $\psi_1(x_i) = x_i - \beta$, $\psi_2(x_i) = (x_i - \beta)^2 - 1$.

5. Unknown mean and known third central moment, $\mathbb{E}\left\{(X - \mathbb{E}\left\{X\right\})^3\right\} = 0$.
   $\psi_1(x_i) = x_i - \beta$, $\psi_2(x_i) = (x_i - \beta)^3$.

These examples extend cases 2 and 3 of the previous subsection. Here the mean is estimated rather than set to zero; otherwise, the setup is the same. Bias-corrected ET and CUE estimators are not considered. Results are presented in Figure 3.7.

The reduction in variance is now smaller than in cases 2 and 3 above as the extra information used is less and, additionally, estimation error now contributes to the variance term. In case 5, for small sample sizes, the variance of GELKDE exceeds that of PRKDE. In moderate and large samples, however, there is a $1/n$ reduction in MISE. In case 4, as in case 2, bias increases, but now the increase is great enough to outweigh the reduction in variance for sample sizes below about 100. EL performs better as its bias goes to zero faster than the bias of the other two estimators.

3.6 Conclusions

Additional information concerning the distribution of a random variable formulated in terms of moment conditions depending on a finite-dimensional parameter vector, which may or may not be known, can be incorporated by reweighting a kernel density estimate using implied GEL probabilities.

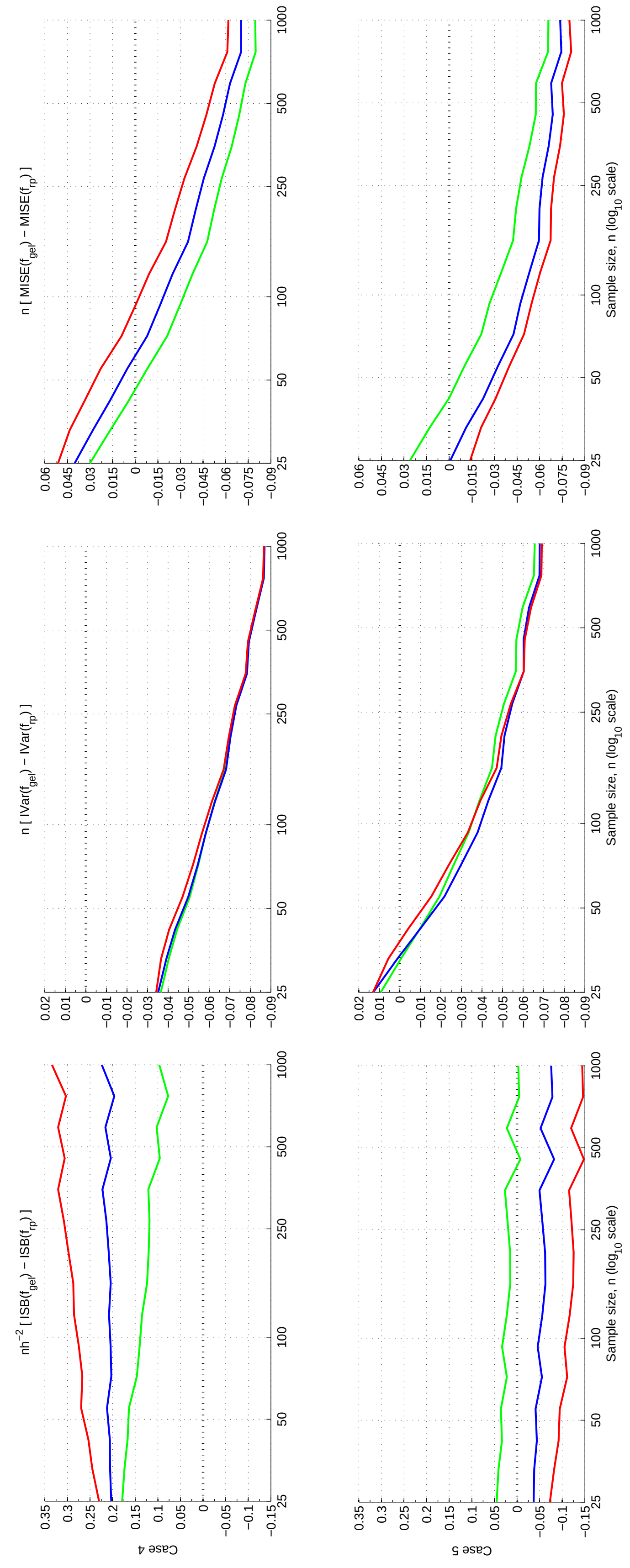

Figure 3.7: Performance of GELKDE with estimated parameters.

The resultant density estimator better approximates the moment conditions than the unweighted, Rosenblatt-Parzen, estimator. Furthermore, a reduction in variance is achieved due to the use of the extra moment information, provided that, if the parameter vector is unknown, it is overidentified. The effect on variance does not depend on the GEL carrier function and dominates the bias effect asymptotically. Simulation evidence suggests that the above conclusions hold in moderate and large samples, whereas in small samples bias can increase and dominate the reduction in variance. The bias of GELKDE depends on the carrier function; however, bias-corrected estimators may be formulated to eliminate the bias.

Extending the above results to the multivariate case is a straightforward exercise, but as performance of kernel density estimators deteriorates in higher dimensions, such an extension may not be of much practical use. However, an extension of these methods for dependent processes may be of interest in economics and finance. Preliminary simulation evidence presented in Appendix 3.C suggests that incorporating information about the dependence structure gives a reduction in variance and mean integrated squared error as compared with RPKDE.

GEL methods need to be modified appropriately to deal with dependent data. One possibility is to use a version of GEL defined via smoothed moment indicators, developed in Smith (2010), which extends this class of estimators to weakly dependent data. Extensions to long-range dependence may be possible using frequency domain empirical likelihood; see e.g. Nordman and Lahiri (2006).

Furthermore, GEL methods can be coupled with penalisation methods thus combining model selection and estimation steps; see inter alia Otsu (2007) and Shahidi (2009). This may be of particular relevance for dependent data when the dependence structure is unknown.

Other possible extensions include the estimation of *conditional* densities and nonparametric regression with extra moment conditions. De Gooijer and Zerom (2003) propose an ad hoc reweighting of a Nadaraya-Watson estimator of a conditional density which is an improvement over the unweighted case and enjoys superior bias properties of the local linear smoother. In particular, EL is used to make the Nadaraya-Watson weights more resemble local linear weights. The encouraging results of this paper suggest that further extensions may be developed for the estimation of conditional densities.

Appendices

Appendix 3.A Proofs

3.A.1 Equations (3.5) and (3.14)

Changing variables such that $z = (x_i - x)/h$, and using the Young's form of Taylor's Theorem to expand $\psi_l(x_i + hz; \boldsymbol{\beta})$ around x_i for given $\boldsymbol{\beta} \in \mathrm{B}_r(\boldsymbol{\beta}_0)$ gives

$$\begin{aligned}\int_{\mathbb{R}} \psi_l(x;\boldsymbol{\beta})\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \psi_l(x_i + hz;\boldsymbol{\beta})K(z)dz \\ &= \frac{1}{n}\sum_{i=1}^{n}\Bigg[\psi_l(x_i;\boldsymbol{\beta})\int_{\mathbb{R}} K(z)dz + \psi_l^{(1)}(x_i;\boldsymbol{\beta})h\int_{\mathbb{R}} zK(z)dz \\ &\qquad + \frac{1}{2}\psi_l^{(2)}(x_i;\boldsymbol{\beta})h^2\int_{\mathbb{R}} z^2K(z)dz + \frac{1}{6}\psi_l^{(3)}(x_i;\boldsymbol{\beta})h^3\int_{\mathbb{R}} z^3K(z)dz \\ &\qquad + \frac{1}{24}\left(\psi_l^{(4)}(x_i;\boldsymbol{\beta})h^4 + o_p\left(h^4\right)\right)\int_{\mathbb{R}} z^4K(z)dz\Bigg] \\ &= \frac{1}{n}\sum_{i=1}^{n}\psi_l(x_i;\boldsymbol{\beta}) + \frac{1}{2}h^2\mu_2(K)\frac{1}{n}\sum_{i=1}^{n}\psi_l^{(2)}(x_i;\boldsymbol{\beta}) + O_p\left(h^4\right).\end{aligned}$$

By Assumption 1(a), $\mu_0(K) = 1$, $\mu_1(K) = \mu_3(K) = 0$ and $\mu_4(K) < \infty$. Hence the terms involving odd powers of h are zero, and the remainder term is of order $O_p\left(h^4\right)$ by the Weak Law of Large Numbers (WLLN) applied to averages of $\psi_l^{(j)}(x_i;\boldsymbol{\beta})$ in view of Assumption 2(a).

Equation (3.14) obtains since $\sum_{i=1}^{n}\hat{\pi}_i = 1$ and $\sum_{i=1}^{n}\hat{\pi}_i\psi_l(x_i;\boldsymbol{\beta}) = 0$. We have that for a GEL-based estimator,

$$\begin{aligned}\int \psi_l(x;\boldsymbol{\beta})\tilde{f}_\rho(x)dx = \int \psi_l(x;\boldsymbol{\beta})\sum_{i=1}^{n}\hat{\pi}_iK\left(\frac{x-x_i}{h}\right)\frac{dx}{h} = \sum_{i=1}^{n}\hat{\pi}_i\int\psi_l(x_i+hz;\boldsymbol{\beta})K(z)\,dz \\ = \sum_{i=1}^{n}\hat{\pi}_i\psi_l(x_i;\boldsymbol{\beta}) + \frac{1}{2}h^2\mu_2(K)\sum_{i=1}^{n}\hat{\pi}_i\psi_l^{(2)}(x_i;\boldsymbol{\beta}) + \sum_{i=1}^{n}\hat{\pi}_i\left[\frac{1}{24}h^4\psi_l^{(4)}(x_i;\boldsymbol{\beta}) + o_p\left(h^4\right)\right]\mu_4(K) \\ = \frac{1}{2}h^2\mu_2(K)\sum_{i=1}^{n}\hat{\pi}_i\psi_l^{(2)}(x_i;\boldsymbol{\beta}) + O_p\left(h^4\right).\end{aligned}$$

By writing $\hat{\pi}_i = \frac{1}{n}\left(1 + o_p(1)\right)$, uniformly i, we can write

$$\frac{1}{2}h^2\mu_2(K)\sum_{i=1}^{n}\hat{\pi}_i\psi_l^{(2)}(x_i;\boldsymbol{\beta}) = \left(\frac{1}{2}h^2\mu_2(K)\frac{1}{n}\sum_{i=1}^{n}\psi_l^{(2)}(x_i;\boldsymbol{\beta})\right)\left(1 + o_p(1)\right),$$

where the first term is now the same as the second term in (3.5).

3.A.2 Lagrange multipliers when parameters are known

Consider a population version of the GEL criterion (3.6), $P(\boldsymbol{\lambda}) = \mathbb{E}_f\left\{\rho\left(\boldsymbol{\lambda}^\intercal\boldsymbol{\psi}(X)\right)\right\} - \rho(0)$. Since $\rho(\cdot)$ is globally concave, it follows by Jensen's inequality that $\mathbb{E}_f\left\{\rho\left(\boldsymbol{\lambda}^\intercal\boldsymbol{\psi}(X)\right)\right\} \le \rho\left(\mathbb{E}_f\left\{\boldsymbol{\lambda}^\intercal\boldsymbol{\psi}(X)\right\}\right)$, and that

$$P(\boldsymbol{\lambda}) = \mathbb{E}_f\left\{\rho\left(\boldsymbol{\lambda}^\intercal\boldsymbol{\psi}(X)\right)\right\} - \rho(0) \le \rho\left(\mathbb{E}_f\left\{\boldsymbol{\lambda}^\intercal\boldsymbol{\psi}(X)\right\}\right) - \rho(0) = \rho\left(\boldsymbol{\lambda}^\intercal\mathbb{E}_f\left\{\boldsymbol{\psi}(X)\right\}\right) - \rho(0) = 0,$$

where the last equality is implied by the moment conditions $\mathbb{E}_f\left\{\boldsymbol{\psi}\left(X\right)\right\}=0$. Hence the maximum value of $P(\boldsymbol{\lambda})$ is zero, i.e. $P(\mathbf{0})=0$. Indeed, $\boldsymbol{\lambda}=\mathbf{0}$ is a local maximum of $P(\boldsymbol{\lambda})$ as

$$\nabla_{\boldsymbol{\lambda}}P(\boldsymbol{\lambda})\big|_{\boldsymbol{\lambda}=\mathbf{0}}=\mathbb{E}_f\left\{\rho^{(1)}\left(0\right)\boldsymbol{\psi}\left(X\right)\right\}=-\mathbb{E}_f\left\{\boldsymbol{\psi}\left(X\right)\right\}=\mathbf{0}$$

$$\text{and}\qquad \nabla_{\boldsymbol{\lambda}}^{\mathsf{T}}\nabla_{\boldsymbol{\lambda}}P(\boldsymbol{\lambda})\big|_{\boldsymbol{\lambda}=\mathbf{0}}=\mathbb{E}_f\left\{\rho^{(2)}\left(0\right)\boldsymbol{\psi}\left(X\right)\boldsymbol{\psi}\left(X\right)^{\mathsf{T}}\right\}=-\mathbb{E}_f\left\{\boldsymbol{\psi}\left(X\right)\boldsymbol{\psi}\left(X\right)^{\mathsf{T}}\right\}=-\mathbf{V}_{\boldsymbol{\psi}},$$

a negative definite matrix.

Moreover, since $\rho\left(\cdot\right)$ is *concave*, its second derivative, $\rho^{(2)}\left(v\right)$, is non-positive for all v. Hence, by the Lemma and the proof of Theorem 1 in Chesher and Smith (1997, p. 643), $\nabla_{\boldsymbol{\lambda}}^{\mathsf{T}}\nabla_{\boldsymbol{\lambda}}P(\boldsymbol{\lambda})$ is negative definite, and $\boldsymbol{\lambda}=\mathbf{0}$ is a *unique* maximum of $P(\boldsymbol{\lambda})$.

Since $\widehat{\boldsymbol{\lambda}}$ is an M-estimator, $\widehat{\boldsymbol{\lambda}}\xrightarrow{p}\mathbf{0}$, the maximum of $P(\boldsymbol{\lambda})$. Moreover, by Theorem 5.23 of van der Vaart (1998), $\widehat{\boldsymbol{\lambda}}=\mathcal{O}_p\left(n^{-1/2}\right)$.

3.A.3 Expansion for implied probabilities

Expanding $\rho^{(1)}\left(\hat{v}_i\right)$ around zero gives

$$\rho^{(1)}\left(\hat{v}_i\right)=-1-\hat{v}_i+\frac{1}{2}\rho^{(3)}\left(0\right)\hat{v}_i^2+\frac{1}{6}\rho^{(4)}\left(\dot{v}_i\right)\hat{v}_i^3,$$

where $\hat{v}_i=\widehat{\boldsymbol{\lambda}}\boldsymbol{\psi}\left(x_i;\widehat{\boldsymbol{\beta}}\right)$ and $\dot{v}_i=\dot{\boldsymbol{\lambda}}\boldsymbol{\psi}\left(x_i;\widehat{\boldsymbol{\beta}}\right)$ for some $\dot{\boldsymbol{\lambda}}$ on the line joining $\widehat{\boldsymbol{\lambda}}$ and zero. By Lemma A1 of NS,

$$\sup_{\boldsymbol{\beta}\in\mathscr{B},\ \boldsymbol{\lambda}\in\Lambda_n(\boldsymbol{\beta}),\ 1\le i\le n}\left|\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{\psi}\left(x_i;\boldsymbol{\beta}\right)\right|\xrightarrow{p}0,$$

and hence $\rho^{(4)}\left(\dot{v}_i\right)\hat{v}_i^3=\rho^{(4)}\left(0\right)\hat{v}_i^3\left(1+o_p\left(1\right)\right)$.

Expanding the denominator gives

$$\left[\sum_{j=1}^{n}\rho^{(1)}\left(\hat{v}_j\right)\right]^{-1}=-\frac{1}{n}\left[1-\widehat{\boldsymbol{\lambda}}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}\left(x_i;\widehat{\boldsymbol{\beta}}\right)+\frac{\rho^{(3)}\left(0\right)}{2}\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{V}_{\boldsymbol{\psi}}\widehat{\boldsymbol{\lambda}}+\mathcal{O}_p\left(n^{-3/2}\right)\right],$$

where we used the fact that, as shown in NS, $\widehat{\boldsymbol{\lambda}}=\mathcal{O}_p\left(n^{-1/2}\right)$ and $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0=\mathcal{O}_p\left(n^{-1/2}\right)$, thus

$$\frac{1}{n}\sum_{i=1}^{n}\hat{v}_i^2=\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}\left(x_i;\widehat{\boldsymbol{\beta}}\right)\boldsymbol{\psi}\left(x_i;\widehat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\right)\widehat{\boldsymbol{\lambda}}=\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{V}_{\boldsymbol{\psi}}\widehat{\boldsymbol{\lambda}}+\mathcal{O}_p\left(n^{-3/2}\right).$$

Combing the two expansions gives

$$\begin{aligned}\hat{\pi}_i&=\frac{1}{n}\left[1+\hat{v}_i-\frac{1}{2}\rho^{(3)}\left(0\right)\hat{v}_i^2-\frac{1}{6}\rho^{(4)}\left(0\right)\hat{v}_i^3\left(1+o_p\left(1\right)\right)\right]\times\cdots\\&\qquad\cdots\times\left[1-\widehat{\boldsymbol{\lambda}}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}\left(x_i;\widehat{\boldsymbol{\beta}}\right)+\frac{\rho^{(3)}\left(0\right)}{2}\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{V}_{\boldsymbol{\psi}}\widehat{\boldsymbol{\lambda}}+\mathcal{O}_p\left(n^{-3/2}\right)\right]\\&=\frac{1}{n}+\frac{1}{n}\left[\hat{v}_i-\frac{\rho^{(3)}\left(0\right)}{2}\hat{v}_i^2\right]-\frac{1}{n}\frac{\rho^{(4)}\left(0\right)}{6}\hat{v}_i^3\left(1+o_p\left(1\right)\right)+\frac{1}{n}\left[-\widehat{\boldsymbol{\lambda}}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}\left(x_i;\widehat{\boldsymbol{\beta}}\right)+\frac{\rho^{(3)}\left(0\right)}{2}\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{V}_{\boldsymbol{\psi}}\widehat{\boldsymbol{\lambda}}\right]\\&\qquad+\left[\hat{v}_i-\frac{1}{2}\rho^{(3)}\left(0\right)\hat{v}_i^2-\frac{1}{6}\rho^{(4)}\left(0\right)\hat{v}_i^3\left(1+o_p\left(1\right)\right)\right]\mathcal{O}_p\left(n^{-2}\right)+\mathcal{O}_p\left(n^{-5/2}\right).\end{aligned}$$

3.A.4 Proof of Proposition 1

Using the transformation introduced in section 3.3, the first-order conditions for $\boldsymbol{\theta}$ can be written as

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \rho^{(1)}\left(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{w}_i\right)\mathbf{w}_i = \mathbf{0}.$$

Provided $\rho(\cdot)$ possesses enough derivatives, expanding $\rho^{(1)}\left(\boldsymbol{\theta}^{\mathsf{T}}w_i\right)$ around zero yields

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[-1 - v_i + \frac{\rho^{(3)}(0)}{2}v_i^2 + \frac{\rho^{(4)}(0)}{6}v_i^3 + \mathcal{O}_p\left(v_i^4\right)\right]\mathbf{w}_i, \tag{3.19}$$

where by normalisation $\rho^{(1)}(0) = \rho^{(2)}(0) = -1$, and $v_i = \boldsymbol{\theta}^{\mathsf{T}}\mathbf{w}_i = \boldsymbol{\theta}^j\mathbf{w}_i^j$.

Given that $\boldsymbol{\theta} \sim \mathcal{O}_p\left(n^{-1/2}\right)$, the j-th equation in (3.19) can be rewritten as

$$Q_n^j(\boldsymbol{\theta}) = -A^j - \boldsymbol{\theta}^j - A^{jk}\boldsymbol{\theta}^k + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\boldsymbol{\theta}^k\boldsymbol{\theta}^l + \frac{\rho^{(3)}(0)}{2}A^{jkl}\boldsymbol{\theta}^k\boldsymbol{\theta}^l + \frac{\rho^{(4)}(0)}{6}\alpha^{jklm}\boldsymbol{\theta}^k\boldsymbol{\theta}^l\boldsymbol{\theta}^m + \mathcal{O}_p\left(n^{-2}\right). \tag{3.20}$$

Solving $Q_n(\boldsymbol{\theta}) = \mathbf{0}$ for $\boldsymbol{\theta}$ gives (3.11). To be specific, first note from (3.20) that

$$\boldsymbol{\theta}^j = -A^j - A^{jk}\boldsymbol{\theta}^k + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\boldsymbol{\theta}^k\boldsymbol{\theta}^l + \frac{\rho^{(3)}(0)}{2}A^{jkl}\boldsymbol{\theta}^k\boldsymbol{\theta}^l + \frac{\rho^{(4)}(0)}{6}\alpha^{jklm}\boldsymbol{\theta}^k\boldsymbol{\theta}^l\boldsymbol{\theta}^m + \mathcal{O}_p\left(n^{-2}\right).$$

Considering each term on the right-hand side in turn gives:

$$\begin{aligned}
A^{jk}\boldsymbol{\theta}^k &= -A^{jk}A^k - A^{jk}A^{kl}\boldsymbol{\theta}^l + \frac{\rho^{(3)}(0)}{2}A^{jk}\alpha^{klm}\boldsymbol{\theta}^l\boldsymbol{\theta}^m + \mathcal{O}_p\left(n^{-2}\right)\\
&= -A^{jk}A^k + A^{jk}A^{kl}A^l + \frac{\rho^{(3)}(0)}{2}A^{jk}\alpha^{klm}A^lA^m + \mathcal{O}_p\left(n^{-2}\right);\\
\alpha^{jkl}\boldsymbol{\theta}^k\boldsymbol{\theta}^l &= \alpha^{jkl}A^kA^l + 2\alpha^{jkl}A^kA^{lm}\boldsymbol{\theta}^m - \rho^{(3)}(0)\,\alpha^{jkl}\alpha^{lmp}A^k\boldsymbol{\theta}^m\boldsymbol{\theta}^p + \mathcal{O}_p\left(n^{-2}\right)\\
&= \alpha^{jkl}A^kA^l - 2\alpha^{jkl}A^kA^{lm}A^m - \rho^{(3)}(0)\,\alpha^{jkl}\alpha^{lmp}A^kA^mA^p + \mathcal{O}_p\left(n^{-2}\right);\\
A^{jkl}\boldsymbol{\theta}^k\boldsymbol{\theta}^l &= A^{jkl}A^kA^l + \mathcal{O}_p\left(n^{-2}\right);\\
\alpha^{jklm}\boldsymbol{\theta}^k\boldsymbol{\theta}^l\boldsymbol{\theta}^m &= -\alpha^{jklm}A^kA^lA^m + \mathcal{O}_p\left(n^{-2}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\boldsymbol{\theta}^j = &-A^j + A^{jk}A^k + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}A^kA^l - A^{jk}A^{kl}A^l + \frac{\rho^{(3)}(0)}{2}A^{jkl}A^kA^l - \frac{\rho^{(3)}(0)}{2}A^{jk}\alpha^{klm}A^lA^m\\
&- \rho^{(3)}(0)\,\alpha^{jkl}A^{lm}A^kA^m - \frac{\left(\rho^{(3)}(0)\right)^2}{2}\alpha^{jkl}\alpha^{lmp}A^kA^mA^p - \frac{\rho^{(4)}(0)}{6}\alpha^{jklm}A^kA^lA^m + \mathcal{O}_p\left(n^{-2}\right).
\end{aligned}$$

3.A.5 Proof of Proposition 2

Let $\widehat{\mathbf{w}}_i = \mathbf{V}_{\boldsymbol{\psi}}^{-1/2}\boldsymbol{\psi}\left(x_i;\widehat{\boldsymbol{\beta}}\right)$; then the first order conditions for $\boldsymbol{\theta}$ are

$$Q_{\theta,n}^j(\boldsymbol{\theta},\widehat{\boldsymbol{\beta}}) = -\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{w}}_i^j - \frac{1}{n}\sum_{i=1}^{n}v_i\widehat{\mathbf{w}}_i^j + \frac{\rho^{(3)}(0)}{2}\frac{1}{n}\sum_{i=1}^{n}v_i^2\widehat{\mathbf{w}}_i^j + \frac{1}{n}\sum_{i=1}^{n}\mathcal{O}_p\left(|\hat{v}_i|^3\right)\widehat{\mathbf{w}}_i^j, \quad j \in \{1,\ldots,q\}.$$

Let

$$\Gamma_i^{j,j_1\cdots j_l} = \frac{\partial^l\mathbf{w}_i^j}{\partial\beta^{j_1}\cdots\partial\beta^{j_l}} - \gamma^{j,j_1\cdots j_l}.$$

Taking a second order expansion of $\widehat{\mathbf{w}}_i^j$ around $\boldsymbol{\beta}_0$ gives

$$\widehat{\mathbf{w}}_i^j = \mathbf{w}_i^j + \gamma^{j,r}\tilde{\beta}^r + \Gamma_i^{j,r}\tilde{\beta}^r + \frac{1}{2}\gamma^{j,rs}\tilde{\beta}^r\tilde{\beta}^s + \frac{1}{2}\Gamma_i^{j,rs}\tilde{\beta}^r\tilde{\beta}^s + R_n, \tag{3.21}$$

where R_n is the remainder term. Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{w}}_i^j = A^j + \gamma^{j,r}\tilde{\beta}^r + \Gamma^{j,r}\tilde{\beta}^r + \frac{1}{2}\gamma^{j,rs}\tilde{\beta}^r\tilde{\beta}^s + O_p\left(n^{-3/2}\right).$$

Furthermore,

$$\frac{1}{n}\sum_{i=1}^{n} v_i\widehat{\mathbf{w}}_i^j = A^{jk}\boldsymbol{\theta}^k + \alpha^{jk}\boldsymbol{\theta}^k + O_p\left(n^{-3/2}\right), \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} v_i^2\widehat{\mathbf{w}}_i^j = \alpha^{jkl}\boldsymbol{\theta}^k\boldsymbol{\theta}^l + O_p\left(n^{-3/2}\right).$$

Combining terms and noting that $\alpha^{jk} = \delta^{jk}$ gives

$$Q_{\theta,n}^j(\boldsymbol{\theta},\widehat{\boldsymbol{\beta}}) = -\boldsymbol{\theta}^j - A^j - A^{jk}\boldsymbol{\theta}^k + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\boldsymbol{\theta}^k\boldsymbol{\theta}^l - \gamma^{j,r}\tilde{\beta}^r - \Gamma^{j,r}\tilde{\beta}^r - \frac{1}{2}\gamma^{j,rs}\tilde{\beta}^r\tilde{\beta}^s + O_p\left(n^{-3/2}\right), \tag{3.22}$$

where $\tilde{\beta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$.

Solving $Q_{\theta,n}^j(\boldsymbol{\theta},\widehat{\boldsymbol{\beta}}) = 0$ for $\boldsymbol{\theta}(\widehat{\boldsymbol{\beta}})$ gives

$$\boldsymbol{\theta}^j = -A^j - A^{jk}\boldsymbol{\theta}^k + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\boldsymbol{\theta}^k\boldsymbol{\theta}^l - \gamma^{j,r}\tilde{\beta}^r - \Gamma^{j,r}\tilde{\beta}^r - \frac{1}{2}\gamma^{j,rs}\tilde{\beta}^r\tilde{\beta}^s + O_p\left(n^{-3/2}\right).$$

Substituting for $\boldsymbol{\theta}$ gives

$$\begin{aligned} A^{jk}\boldsymbol{\theta}^k &= -A^{jk}A^k - A^{jk}\gamma^{k,r}\tilde{\beta}^r + O_p\left(n^{-3/2}\right), \\ \alpha^{jkl}\boldsymbol{\theta}^k\boldsymbol{\theta}^l &= \alpha^{jkl}A^kA^l + 2\alpha^{jkl}A^k\gamma^{l,r}\tilde{\beta}^r + \alpha^{jkl}\gamma^{k,r}\tilde{\beta}^r\gamma^{l,s}\tilde{\beta}^s + O_p\left(n^{-3/2}\right). \end{aligned}$$

Hence,

$$\begin{aligned} \boldsymbol{\theta}^j = &-A^j + A^{jk}A^k + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}A^kA^l - \gamma^{j,r}\tilde{\beta}^r + A^{jk}\gamma^{k,r}\tilde{\beta}^r \\ &+ \rho^{(3)}(0)\,\alpha^{jkl}A^k\gamma^{l,r}\tilde{\beta}^r + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\gamma^{k,r}\tilde{\beta}^r\gamma^{l,s}\tilde{\beta}^s - \Gamma^{j,r}\tilde{\beta}^r - \frac{1}{2}\gamma^{j,rs}\tilde{\beta}^r\tilde{\beta}^s + O_p\left(n^{-3/2}\right). \end{aligned} \tag{3.23}$$

$\widehat{\boldsymbol{\beta}}$ solves the first-order conditions

$$\begin{aligned} Q_{\beta,n}^r(\boldsymbol{\theta},\widehat{\boldsymbol{\beta}}) &= \frac{1}{n}\sum_{i=1}^{n}\rho^{(1)}\left(\widehat{\boldsymbol{\lambda}}^{\mathsf{T}}\boldsymbol{\psi}\left(x_i;\widehat{\boldsymbol{\beta}}\right)\right)\frac{\partial\widehat{\boldsymbol{\psi}}_i^j}{\partial\widehat{\boldsymbol{\beta}}^r}\widehat{\boldsymbol{\lambda}}^j = \frac{1}{n}\sum_{i=1}^{n}\rho^{(1)}\left(\boldsymbol{\theta}^{\mathsf{T}}\widehat{\mathbf{w}}_i\right)\frac{\partial\widehat{\mathbf{w}}_i^j}{\partial\widehat{\boldsymbol{\beta}}^r}\boldsymbol{\theta}^j \\ &= -\boldsymbol{\theta}^j\gamma^{j,r} - \boldsymbol{\theta}^j\Gamma^{j,r} - \boldsymbol{\theta}^j\gamma^{j,rs}\tilde{\beta}^s + O_p\left(n^{-3/2}\right), \quad j \in \{1,\ldots,q\},\ r,s \in \{1,\ldots,p\}; \end{aligned} \tag{3.24}$$

where the third equality is obtained by expanding $\rho^{(1)}(\cdot)$ around zero and $\partial\widehat{\mathbf{w}}_i^j \big/ \partial\widehat{\boldsymbol{\beta}}^r$ around $\boldsymbol{\beta}_0$.

Substituting for $\boldsymbol{\theta}$ from (3.23) into (3.24) gives

$$\begin{aligned}\boldsymbol{\theta}^j\gamma^{j,t} &= -A^j\gamma^{j,t} + A^{jk}A^k\gamma^{j,t} + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}A^kA^l\gamma^{j,t} - \gamma^{j,r}\tilde{\beta}^r\gamma^{j,t} + A^{jk}\gamma^{k,r}\tilde{\beta}^r\gamma^{j,t} - \Gamma^{j,r}\tilde{\beta}^r\gamma^{j,t} \\ &\quad + \rho^{(3)}(0)\,\alpha^{jkl}A^k\gamma^{l,r}\tilde{\beta}^r\gamma^{j,t} + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\gamma^{k,r}\tilde{\beta}^r\gamma^{l,s}\tilde{\beta}^s\gamma^{j,t} - \frac{1}{2}\gamma^{j,rs}\tilde{\beta}^r\tilde{\beta}^s\gamma^{j,t} + \mathcal{O}_p\left(n^{-3/2}\right), \\ \boldsymbol{\theta}^j\Gamma^{j,t} &= -A^j\Gamma^{j,t} - \gamma^{j,r}\tilde{\beta}^r\Gamma^{j,t} + \mathcal{O}_p\left(n^{-3/2}\right), \qquad \text{and} \\ \boldsymbol{\theta}^j\gamma^{j,tv}\tilde{\beta}^v &= -A^j\gamma^{j,tv}\tilde{\beta}^v - \gamma^{j,r}\tilde{\beta}^r\gamma^{j,tv}\tilde{\beta}^v + \mathcal{O}_p\left(n^{-3/2}\right).\end{aligned}$$

Combining terms, equating $Q^r_{\beta,n}(\boldsymbol{\theta},\widehat{\boldsymbol{\beta}})$ to zero and rearranging gives

$$\begin{aligned}\gamma^{j,r}\tilde{\beta}^r\gamma^{j,t} &= -A^j\gamma^{j,t} + A^{jk}A^k\gamma^{j,t} + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}A^kA^l\gamma^{j,t} + A^{jk}\gamma^{k,r}\tilde{\beta}^r\gamma^{j,t} - \Gamma^{j,r}\tilde{\beta}^r\gamma^{j,t} \\ &\quad + \rho^{(3)}(0)\,\alpha^{jkl}A^k\gamma^{l,r}\tilde{\beta}^r\gamma^{j,t} + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\gamma^{k,r}\tilde{\beta}^r\gamma^{l,s}\tilde{\beta}^s\gamma^{j,t} - \frac{1}{2}\gamma^{j,rs}\tilde{\beta}^r\tilde{\beta}^s\gamma^{j,t} \\ &\quad - A^j\Gamma^{j,t} - \gamma^{j,r}\tilde{\beta}^r\Gamma^{j,t} - A^j\gamma^{j,tv}\tilde{\beta}^v - \gamma^{j,r}\tilde{\beta}^r\gamma^{j,tv}\tilde{\beta}^v + \mathcal{O}_p\left(n^{-3/2}\right).\end{aligned}$$

By definition, $\omega^{st}\gamma^{j,t}\gamma^{j,r} = \delta^{sr}$. Premultiplying the above equation by ω^{st} gives

$$\begin{aligned}\tilde{\beta}^s &= -\omega^{st}A^j\gamma^{j,t} + \omega^{st}A^{jk}A^k\gamma^{j,t} + \frac{\rho^{(3)}(0)}{2}\omega^{st}\alpha^{jkl}A^kA^l\gamma^{j,t} + \omega^{st}A^{jk}\gamma^{k,r}\tilde{\beta}^r\gamma^{j,t} - \omega^{st}\Gamma^{j,r}\tilde{\beta}^r\gamma^{j,t} \\ &\quad + \rho^{(3)}(0)\,\omega^{st}\alpha^{jkl}A^k\gamma^{l,r}\tilde{\beta}^r\gamma^{j,t} + \frac{\rho^{(3)}(0)}{2}\omega^{st}\alpha^{jkl}\gamma^{k,r}\tilde{\beta}^r\gamma^{l,s}\tilde{\beta}^s\gamma^{j,t} - \frac{1}{2}\omega^{st}\gamma^{j,rs}\tilde{\beta}^r\tilde{\beta}^s\gamma^{j,t} \\ &\quad - \omega^{st}A^j\Gamma^{j,t} - \omega^{st}\gamma^{j,r}\tilde{\beta}^r\Gamma^{j,t} - \omega^{st}A^j\gamma^{j,tv}\tilde{\beta}^v - \omega^{st}\gamma^{j,r}\tilde{\beta}^r\gamma^{j,tv}\tilde{\beta}^v + \mathcal{O}_p\left(n^{-3/2}\right).\end{aligned}$$

Solving for $\tilde{\beta}$ yields

$$\begin{aligned}\tilde{\beta}^s &= -\omega^{st}A^j\gamma^{j,t} + \omega^{st}A^{jk}A^k\gamma^{j,t} + \frac{\rho^{(3)}(0)}{2}\omega^{st}\alpha^{jkl}A^kA^l\gamma^{j,t} - \omega^{st}A^{jk}\gamma^{k,r}\omega^{rv}A^l\gamma^{l,v}\gamma^{j,t} - \omega^{st}A^j\Gamma^{j,t} \\ &\quad - \rho^{(3)}(0)\,\omega^{st}\alpha^{jkl}A^k\gamma^{l,r}\omega^{rv}A^m\gamma^{m,v}\gamma^{j,t} + \frac{\rho^{(3)}(0)}{2}\omega^{st}\alpha^{jkl}\gamma^{k,r}\omega^{ru}A^m\gamma^{m,u}\gamma^{l,w}\omega^{wv}A^n\gamma^{n,v}\gamma^{j,t} \\ &\quad + \omega^{st}\Gamma^{j,r}\omega^{ru}A^k\gamma^{k,u}\gamma^{j,t} - \frac{1}{2}\omega^{st}\gamma^{j,rv}\omega^{ru}A^k\gamma^{k,u}\omega^{vw}A^l\gamma^{l,w}\gamma^{j,t} + \omega^{st}\gamma^{j,r}\omega^{ru}A^k\gamma^{k,u}\Gamma^{j,t} \\ &\quad + \omega^{st}A^j\gamma^{j,tv}\omega^{vu}A^k\gamma^{k,u} - \omega^{st}\gamma^{j,r}\omega^{ru}A^k\gamma^{k,u}\gamma^{j,tv}\omega^{vw}A^l\gamma^{l,w} + \mathcal{O}_p\left(n^{-3/2}\right).\end{aligned} \tag{3.25}$$

Finally, substituting this back into (3.23) reproduces equation (3.12) in Proposition 2.

3.A.6 Proof of Proposition 3

Using (3.10) with $v_i = \boldsymbol{\theta}^j\mathbf{w}^j_i$ write

$$\tilde{f}_\rho(x) = \hat{f}(x) + T_1 - \frac{\rho^{(3)}(0)}{2}T_2 + T_3 + \sum_{i=1}^{n} R^{[\pi]}_n K_h\left(X_i - x\right),$$

$$\text{where}\quad T_1 = n^{-1}\sum_{i=1}^{n} v_i K_h\left(X_i - x\right), \quad T_2 = n^{-1}\sum_{i=1}^{n} v_i^2 K_h\left(X_i - x\right),$$
$$T_3 = -n^{-1}\boldsymbol{\theta}^j A^j \sum_{i=1}^{n} K_h\left(X_i - x\right) + n^{-1}\frac{\rho^{(3)}\left(0\right)}{2}\boldsymbol{\theta}^j\boldsymbol{\theta}^j \sum_{i=1}^{n} K_h\left(X_i - x\right),$$

and the reminder term $R_n^{[\pi]}$ is defined below (3.10).

As shown in Appendix 3.A.6.1, provided $n^{-1}h^{-4}$ goes to zero as $n \to \infty$,

$$\begin{aligned}\mathbb{E}\left\{T_1\right\} &= -n^{-1}\mathbf{w}^j(x)\mathbf{w}^j(x)f(x) + n^{-1}k_\rho\alpha^{jkl}\delta^{kl}\mathbf{w}^j(x)f(x)\\ &\quad + \frac{1}{2}\mu_2(K)n^{-1}h^2\left[-C_1 + k_\rho\alpha^{jkl}\delta^{kl}C_2^j\right] + o\left(n^{-1}h^4\right),\\ \text{and}\quad \mathbb{E}\left\{T_2\right\} &= n^{-1}\mathbf{w}^j(x)\mathbf{w}^j(x)f(x) + \frac{1}{2}C_1\mu_2(K)n^{-1}h^2 + o\left(n^{-1}h^4\right),\end{aligned}$$

where $k_\rho = 1 + \rho^{(3)}\left(0\right)/\,2$, $C_1 = \frac{d^2}{dv^2}\left[\mathbf{w}^j(v)\mathbf{w}^j(v)f(v)\right]\Big|_{v=x}$ and $C_2^j = \frac{d^2}{dv^2}\left[\mathbf{w}^j(v)f(v)\right]\Big|_{v=x}$. It is then easy to see that

$$\mathbb{E}\left\{T_3\right\} = n^{-1}k_\rho q f(x) + n^{-1}h^2 k_\rho\frac{\mu_2(K)}{2}qf^{(2)}(x) + o\left(n^{-1}h^4\right),$$

and the contribution from the remainder term, $R_n^{[\pi]}$, is of order $o\left(n^{-2}\right)$. Combining the terms gives equation (3.15a).

The expression for the integrated squared bias, (3.15c), is obtained immediately from (3.15a) noting that $\mathbb{E}\left\{\hat{f}(x)\right\} = f(x) + \frac{1}{2}h^2 f^{(2)}(x)\mu_2(K) + o\left(h^4\right)$, and the leading term in the integrated squared bias of $\hat{f}$ is of order h^4.

To obtain the variance, first note that from (3.15a),

$$\begin{aligned}\left[\mathbb{E}\left\{\tilde{f}_\rho(x)\right\}\right]^2 &= \left[\mathbb{E}\left\{\hat{f}(x)\right\}\right]^2 + 2n^{-1}k_\rho\left[-\mathbf{w}^j(x)\mathbf{w}^j(x) + \alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) + q\right]f^2(x)\\ &\quad + n^{-1}h^2\mu_2(K)k_\rho\left[-\mathbf{w}^j(x)\mathbf{w}^j(x) + \alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) + q\right]f^{(2)}(x)f(x)\\ &\quad + n^{-1}h^2\mu_2(K)k_\rho\left[-C_1 + \alpha^{jkl}\delta^{kl}C_2^j + qf^{(2)}(x)\right]f(x) + o\left(n^{-1}h^4\right).\end{aligned}$$

As shown in Appendix 3.A.6.2,

$$\begin{aligned}\mathbb{E}\left\{\tilde{f}_\rho^2(x)\right\} &= \mathbb{E}\left\{\hat{f}^2(x)\right\} - n^{-1}\mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x)\\ &\quad + 2n^{-1}k_\rho\left[-\mathbf{w}^j(x)\mathbf{w}^j(x) + \alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) + q\right]f^2(x)\\ &\quad + n^{-1}h^2\mu_2(K)k_\rho\left[-\mathbf{w}^j(x)\mathbf{w}^j(x) + \alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) + q\right]f^{(2)}(x)f(x)\\ &\quad + n^{-1}h^2\mu_2(K)k_\rho\left[-C_1 + \alpha^{jkl}\delta^{kl}C_2^j + qf^{(2)}(x)\right]f(x)\\ &\quad - n^{-1}h^2\mu_2(K)\mathbf{w}^j(x)C_2^j f(x) + o\left(n^{-1}h^4\right).\end{aligned}$$

Subtracting $\left[\mathbb{E}\left\{\tilde{f}_\rho(x)\right\}\right]^2$ from $\mathbb{E}\left\{\tilde{f}_\rho^2(x)\right\}$ gives the variance expression (3.15b). Equations (3.15d) and (3.15e) then follow immediately.

3.A.6.1 Expectation of GELKDE

Recall that $v_i = \widehat{\boldsymbol{\lambda}}^{\mathsf{T}} \boldsymbol{\psi}_i = \boldsymbol{\theta}^j \mathbf{w}_i^j$. Thus from (3.11) we obtain

$$\begin{aligned} v_i = &-A^j \mathbf{w}_i^j + A^{jk} A^k \mathbf{w}_i^j + \frac{\rho^{(3)}(0)}{2} \alpha^{jkl} A^k A^l \mathbf{w}_i^j - A^{jk} A^{kl} A^l \mathbf{w}_i^j + \frac{\rho^{(3)}(0)}{2} A^{jkl} A^k A^l \mathbf{w}_i^j \\ &- \frac{\rho^{(3)}(0)}{2} A^{jk} \alpha^{klm} A^l A^m \mathbf{w}_i^j - \rho^{(3)}(0)\, \alpha^{jkl} A^{lm} A^k A^m \mathbf{w}_i^j - \frac{\left(\rho^{(3)}(0)\right)^2}{2} \alpha^{jkl} \alpha^{lmp} A^k A^m A^p \mathbf{w}_i^j \\ &- \frac{\rho^{(4)}(0)}{6} \alpha^{jklm} A^k A^l A^m \mathbf{w}_i^j + O_p\left(n^{-2}\right) \mathbf{w}_i^j. \end{aligned} \tag{3.26}$$

Substituting from (3.26) into T_1 and taking expectations we obtain (T_{1m} stands for T_1 with the m-th term from (3.26) substituted for v_i).

$$\begin{aligned} n\mathbb{E}\left\{T_{11}\right\} &= -\mathbb{E}\left\{\sum_{i=1}^n A^j \mathbf{w}_i^j K_h\left(X_i - x\right)\right\} = -n^{-1}\mathbb{E}\left\{\sum_{i=1}^n \sum_{s=1}^n \mathbf{w}_s^j \mathbf{w}_i^j K_h\left(X_i - x\right)\right\} \\ &= -\mathbb{E}\left\{\mathbf{w}_1^j \mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} = -\mathbf{w}^j(x)\mathbf{w}^j(x) f(x) - \frac{1}{2} C_1 \mu_2(K) h^2 + O\left(h^4\right), \end{aligned}$$

where $\mathbf{w}(x) = \mathbf{w}(x; \boldsymbol{\beta}_0)$; $\mathbf{w}^j(x)\mathbf{w}^j(x) = \mathbf{w}(x)^{\mathsf{T}}\mathbf{w}(x)$ and

$$\begin{aligned} \mathbb{E}\left\{\mathbf{w}(X_i)^{\mathsf{T}}\mathbf{w}(X_i) K_h\left(X_i - x\right)\right\} &= \int_{\mathbb{R}} \mathbf{w}(u)^{\mathsf{T}}\mathbf{w}(u) K_h\left(u - x\right) f(u) du \\ &= \int_{\mathbb{R}} \mathbf{w}(x+hz)^{\mathsf{T}}\mathbf{w}(x+hz) K\left(z\right) f(x+hz) dz \quad \text{(by change of variables: } u = x + hz\text{)} \\ &= \mathbf{w}^j(x)\mathbf{w}^j(x) f(x) + \frac{1}{2} C_1 \mu_2(K) h^2 + O\left(h^4\right) \qquad \text{(expanding around } x \text{ and integrating).} \end{aligned}$$

Here $C_1 = \frac{d^2}{dv^2}\left[\mathbf{w}(v)^{\mathsf{T}}\mathbf{w}(v) f(v)\right]\Big|_{v=x}$, and the $O\left(h^4\right)$ term is $\frac{1}{24} D_1 \mu_4(K) h^4$ with $D_1 = \frac{d^4}{dv^4}\left[\mathbf{w}(v)^{\mathsf{T}}\mathbf{w}(v) f(v)\right]$ evaluated at a point between x and $x + hz$.

Writing

$$A^{jk} A^k \mathbf{w}_i^j = \left[n^{-1} \sum_{h=1}^n \mathbf{w}_h^j \mathbf{w}_h^k - \alpha^{jk}\right]\left[n^{-1} \sum_{s=1}^n \mathbf{w}_s^k\right] \mathbf{w}_i^j = n^{-2} \sum_{h=1}^n \sum_{s=1}^n \mathbf{w}_h^j \mathbf{w}_h^k \mathbf{w}_s^k \mathbf{w}_i^j - n^{-1} \sum_{s=1}^n \mathbf{w}_s^j \mathbf{w}_i^j,$$

and assuming $n^{-1}h^{-4} \to 0$ as $n \to \infty$, we obtain

$$\begin{aligned} n\mathbb{E}\left\{T_{12}\right\} &= \mathbb{E}\left\{\sum_{i=1}^n A^{jk} A^k \mathbf{w}_i^j K_h\left(X_i - x\right)\right\} \\ &= n^{-2}\mathbb{E}\left\{\sum_{i=1}^n \sum_{h=1}^n \sum_{s=1}^n \mathbf{w}_h^j \mathbf{w}_h^k \mathbf{w}_s^k \mathbf{w}_i^j K_h\left(X_i - x\right)\right\} - n^{-1}\mathbb{E}\left\{\sum_{i=1}^n \sum_{s=1}^n \mathbf{w}_s^j \mathbf{w}_i^j K_h\left(X_i - x\right)\right\} \\ &= \mathbb{E}\left\{\mathbf{w}_2^j \mathbf{w}_2^k \mathbf{w}_1^k \mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} + \mathbb{E}\left\{\mathbf{w}_2^j \mathbf{w}_2^k \mathbf{w}_2^k \mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} \\ &\qquad - \mathbb{E}\left\{\mathbf{w}_1^j \mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} + O\left(n^{-1}\right) \\ &= \mathbb{E}\left\{\delta^{jk} \mathbf{w}_1^k \mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} + \mathbb{E}\left\{\alpha^{jkl} \delta^{kl} \mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} \\ &\qquad - \mathbb{E}\left\{\mathbf{w}_1^j \mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} + O\left(n^{-1}\right) \qquad (\star) \\ &= \alpha^{jkl} \delta^{kl} \mathbf{w}^j(x) f(x) + \frac{1}{2} \alpha^{jkl} \delta^{kl} C_2^j \mu_2(K) h^2 + O\left(h^4\right). \end{aligned}$$

Note that the first and third terms in line ($\star$) cancel, and

$$\begin{aligned}\mathbb{E}\left\{\mathbf{w}^j(X_i)K_h\left(X_i-x\right)\right\} &= \int_{\mathbb{R}} \mathbf{w}^j(u)K_h\left(u-x\right)f(u)du \\ &= \int_{\mathbb{R}} \mathbf{w}^j(x+hz)K\left(z\right)f(x+hz)dz \qquad \text{(by change of variables: } u=x+hz\text{)} \\ &= \mathbf{w}^j(x)f(x)+\frac{1}{2}C_2^j\mu_2(K)h^2+O\left(h^4\right) \qquad \text{(expanding around } x \text{ and integrating),}\end{aligned}$$

where $C_2^j = \frac{d^2}{dv^2}\left[\mathbf{w}^j(v)f(v)\right]\Big|_{v=x}$, and the $O\left(h^4\right)$ term is $\frac{1}{24}D_2^j\mu_4(K)h^4$ with $D_2^j = \frac{d^4}{dv^4}\left[\mathbf{w}^j(v)f(v)\right]$ evaluated at a point between x and $x+hz$.

$$\begin{aligned}n\mathbb{E}\left\{T_{13}\right\} &= \frac{\rho^{(3)}\left(0\right)}{2}\mathbb{E}\left\{\sum_{i=1}^{n}\alpha^{jkl}A^kA^l\mathbf{w}_i^jK_h\left(X_i-x\right)\right\} \\ &= \frac{\rho^{(3)}\left(0\right)}{2}n^{-2}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{h=1}^{n}\sum_{s=1}^{n}\alpha^{jkl}\mathbf{w}_h^k\mathbf{w}_s^l\mathbf{w}_i^jK_h\left(X_i-x\right)\right\} \\ &= \frac{\rho^{(3)}\left(0\right)}{2}\mathbb{E}\left\{\alpha^{jkl}\mathbf{w}_2^k\mathbf{w}_2^l\mathbf{w}_1^jK_h\left(X_1-x\right)\right\}+O\left(n^{-1}\right) \\ &= \frac{\rho^{(3)}\left(0\right)}{2}\alpha^{jkl}\delta^{kl}\mathbf{w}^j(x)f(x)+\frac{\rho^{(3)}\left(0\right)}{4}\alpha^{jkl}\delta^{kl}C_2^j\mu_2(K)h^2+O\left(h^4\right).\end{aligned}$$

Noting that

$$\begin{aligned}A^{jk}A^{kl}A^l\mathbf{w}_i^j &= \left[n^{-1}\sum_{h=1}^{n}\mathbf{w}_h^j\mathbf{w}_h^k-\delta^{jk}\right]\left[n^{-1}\sum_{s=1}^{n}\mathbf{w}_s^k\mathbf{w}_s^l-\delta^{kl}\right]\left[n^{-1}\sum_{t=1}^{n}\mathbf{w}_t^l\right]\mathbf{w}_i^j \\ &= n^{-3}\sum_{h=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{w}_h^j\mathbf{w}_h^k\mathbf{w}_s^k\mathbf{w}_s^l\mathbf{w}_t^l\mathbf{w}_i^j-n^{-2}\sum_{h=1}^{n}\sum_{t=1}^{n}\delta^{kl}\mathbf{w}_h^j\mathbf{w}_h^k\mathbf{w}_t^l\mathbf{w}_i^j \\ &\quad -n^{-2}\sum_{s=1}^{n}\sum_{t=1}^{n}\delta^{jk}\mathbf{w}_s^k\mathbf{w}_s^l\mathbf{w}_t^l\mathbf{w}_i^j+n^{-1}\sum_{t=1}^{n}\delta^{jk}\delta^{kl}\mathbf{w}_t^l\mathbf{w}_i^j\end{aligned}$$

we obtain

$$
\begin{aligned}
n\mathbb{E}\left\{T_{14}\right\} &= -\mathbb{E}\left\{\sum_{i=1}^{n} A^{jk}A^{kl}A^{l}\mathbf{w}_i^j K_h\left(X_i - x\right)\right\} \\
&= -n^{-3}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{h=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n} \mathbf{w}_h^j\mathbf{w}_h^k\mathbf{w}_s^k\mathbf{w}_s^l\mathbf{w}_t^l\mathbf{w}_i^j K_h\left(X_i - x\right)\right\} \\
&\quad + n^{-2}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{h=1}^{n}\sum_{t=1}^{n} \delta^{kl}\mathbf{w}_h^j\mathbf{w}_h^k\mathbf{w}_t^l\mathbf{w}_i^j K_h\left(X_i - x\right)\right\} \\
&\quad + n^{-2}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n} \delta^{jk}\mathbf{w}_s^k\mathbf{w}_s^l\mathbf{w}_t^l\mathbf{w}_i^j K_h\left(X_i - x\right)\right\} \\
&\quad - n^{-1}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{t=1}^{n} \delta^{jk}\delta^{kl}\mathbf{w}_t^l\mathbf{w}_i^j K_h\left(X_i - x\right)\right\} \\
&= -\mathbb{E}\left\{\mathbf{w}_3^j\mathbf{w}_3^k\mathbf{w}_2^k\mathbf{w}_2^l\mathbf{w}_1^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} - \mathbb{E}\left\{\mathbf{w}_3^j\mathbf{w}_3^k\mathbf{w}_2^k\mathbf{w}_2^l\mathbf{w}_2^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} \\
&\quad - \mathbb{E}\left\{\mathbf{w}_3^j\mathbf{w}_3^k\mathbf{w}_2^k\mathbf{w}_2^l\mathbf{w}_3^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} + \mathbb{E}\left\{\delta^{kl}\mathbf{w}_2^j\mathbf{w}_2^k\mathbf{w}_1^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} \\
&\quad + \mathbb{E}\left\{\delta^{kl}\mathbf{w}_2^j\mathbf{w}_2^k\mathbf{w}_2^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} + \mathbb{E}\left\{\delta^{jk}\mathbf{w}_2^k\mathbf{w}_2^l\mathbf{w}_1^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} \\
&\quad + \mathbb{E}\left\{\delta^{jk}\mathbf{w}_2^k\mathbf{w}_2^l\mathbf{w}_2^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} - \mathbb{E}\left\{\delta^{jk}\delta^{kl}\mathbf{w}_1^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} + \mathcal{O}\left(n^{-1}\right) \\
&= -\mathbb{E}\left\{\delta^{jk}\delta^{kl}\mathbf{w}_1^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} - \mathbb{E}\left\{\delta^{jk}\alpha^{klm}\delta^{lm}\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} \\
&\quad - \mathbb{E}\left\{\alpha^{jkl}\delta^{kl}\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} + \mathbb{E}\left\{\delta^{kl}\delta^{jk}\mathbf{w}_1^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} \\
&\quad + \mathbb{E}\left\{\delta^{kl}\alpha^{jkl}\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} + \mathbb{E}\left\{\delta^{jk}\delta^{kl}\mathbf{w}_1^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} \\
&\quad + \mathbb{E}\left\{\delta^{jk}\alpha^{klm}\delta^{lm}\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} - \mathbb{E}\left\{\delta^{jk}\delta^{kl}\mathbf{w}_1^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} + \mathcal{O}\left(n^{-1}\right) \\
&= \mathcal{O}\left(n^{-1}\right).
\end{aligned}
$$

We proceed in a similar fashion to verify each subsequent term.

$$
\begin{aligned}
A^{jkl}A^{k}A^{l}\mathbf{w}_i^j &= \left[n^{-1}\sum_{h=1}^{n}\mathbf{w}_h^j\mathbf{w}_h^k\mathbf{w}_h^l - \alpha^{jkl}\right]\left[n^{-1}\sum_{s=1}^{n}\mathbf{w}_s^k\right]\left[n^{-1}\sum_{t=1}^{n}\mathbf{w}_t^l\right]\mathbf{w}_i^j \\
&= n^{-3}\sum_{h=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{w}_h^j\mathbf{w}_h^k\mathbf{w}_h^l\mathbf{w}_s^k\mathbf{w}_t^l\mathbf{w}_i^j - n^{-2}\sum_{s=1}^{n}\sum_{t=1}^{n}\alpha^{jkl}\mathbf{w}_s^k\mathbf{w}_t^l\mathbf{w}_i^j.
\end{aligned}
$$

$$\begin{aligned}
n\mathbb{E}\left\{T_{15}\right\} &= \frac{\rho^{(3)}\left(0\right)}{2}\mathbb{E}\left\{\sum_{i=1}^{n} A^{jkl}A^{k}A^{l}\mathbf{w}_i^j K_h\left(X_i - x\right)\right\} \\
&= \frac{\rho^{(3)}\left(0\right)}{2}\left[n^{-3}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{h=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{w}_h^j\mathbf{w}_h^k\mathbf{w}_h^l\mathbf{w}_s^k\mathbf{w}_t^l\mathbf{w}_i^j K_h\left(X_i - x\right)\right\}\right. \\
&\quad \left. -n^{-2}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\alpha^{jkl}\mathbf{w}_s^k\mathbf{w}_t^l\mathbf{w}_i^j K_h\left(X_i - x\right)\right\}\right] \\
&= \frac{\rho^{(3)}\left(0\right)}{2}\left[\mathbb{E}\left\{\mathbf{w}_3^j\mathbf{w}_3^k\mathbf{w}_3^l\mathbf{w}_2^k\mathbf{w}_2^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} - \mathbb{E}\left\{\alpha^{jkl}\mathbf{w}_2^k\mathbf{w}_2^l\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\}\right] + \mathcal{O}\left(n^{-1}\right) \\
&= \frac{\rho^{(3)}\left(0\right)}{2}\left[\mathbb{E}\left\{\alpha^{jkl}\delta kl\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\} - \mathbb{E}\left\{\alpha^{jkl}\delta^{kl}\mathbf{w}_1^j K_h\left(X_1 - x\right)\right\}\right] + \mathcal{O}\left(n^{-1}\right) \\
&= \mathcal{O}\left(n^{-1}\right).
\end{aligned}$$

$$\begin{aligned}
A^{jk}\alpha^{klm}A^{l}A^{m}\mathbf{w}_i^j &= \left[n^{-1}\sum_{h=1}^{n}\mathbf{w}_h^j\mathbf{w}_h^k - \delta^{jk}\right]\alpha^{klm}\left[n^{-1}\sum_{s=1}^{n}\mathbf{w}_s^l\right]\left[n^{-1}\sum_{t=1}^{n}\mathbf{w}_t^m\right]\mathbf{w}_i^j \\
&= n^{-3}\sum_{h=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\alpha^{klm}\mathbf{w}_h^j\mathbf{w}_h^k\mathbf{w}_s^l\mathbf{w}_t^m\mathbf{w}_i^j - n^{-2}\sum_{s=1}^{n}\sum_{t=1}^{n}\alpha^{klm}\delta^{jk}\mathbf{w}_s^l\mathbf{w}_t^m\mathbf{w}_i^j;
\end{aligned}$$

$$n\mathbb{E}\left\{T_{16}\right\} = -\frac{\rho^{(3)}\left(0\right)}{2}\mathbb{E}\left\{\sum_{i=1}^{n}A^{jk}\alpha^{klm}A^{l}A^{m}\mathbf{w}_i^j K_h\left(X_i - x\right)\right\} = \mathcal{O}\left(n^{-1}\right).$$

$$\begin{aligned}
\alpha^{jkl}A^{lm}A^{k}A^{m}\mathbf{w}_i^j &= \alpha^{jkl}\left[n^{-1}\sum_{h=1}^{n}\mathbf{w}_h^l\mathbf{w}_h^m - \delta^{lm}\right]\left[n^{-1}\sum_{s=1}^{n}\mathbf{w}_s^k\right]\left[n^{-1}\sum_{t=1}^{n}\mathbf{w}_t^m\right]\mathbf{w}_i^j \\
&= n^{-3}\sum_{h=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\alpha^{jkl}\mathbf{w}_h^l\mathbf{w}_h^m\mathbf{w}_s^k\mathbf{w}_t^m\mathbf{w}_i^j - n^{-2}\sum_{s=1}^{n}\sum_{t=1}^{n}\alpha^{jkl}\delta^{lm}\mathbf{w}_s^k\mathbf{w}_t^m\mathbf{w}_i^j;
\end{aligned}$$

$$n\mathbb{E}\left\{T_{17}\right\} = -\rho^{(3)}\left(0\right)\mathbb{E}\left\{\sum_{i=1}^{n}\alpha^{jkl}A^{lm}A^{k}A^{m}\mathbf{w}_i^j K_h\left(X_i - x\right)\right\} = \mathcal{O}\left(n^{-1}\right).$$

$$\begin{aligned}
\alpha^{jkl}\alpha^{lmp}A^{k}A^{m}A^{p}\mathbf{w}_i^j &= \alpha^{jkl}\alpha^{lmp}\left[n^{-1}\sum_{h=1}^{n}\mathbf{w}_h^k\right]\left[n^{-1}\sum_{s=1}^{n}\mathbf{w}_s^m\right]\left[n^{-1}\sum_{t=1}^{n}\mathbf{w}_t^p\right]\mathbf{w}_i^j \\
&= n^{-3}\sum_{h=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\alpha^{jkl}\alpha^{lmp}\mathbf{w}_h^k\mathbf{w}_s^m\mathbf{w}_t^p\mathbf{w}_i^j;
\end{aligned}$$

$$n\mathbb{E}\left\{T_{18}\right\} = -\frac{\left(\rho^{(3)}\left(0\right)\right)^2}{2}\mathbb{E}\left\{\sum_{i=1}^{n}\alpha^{jkl}\alpha^{lmp}A^{k}A^{m}A^{p}\mathbf{w}_i^j K_h\left(X_i - x\right)\right\} = \mathcal{O}\left(n^{-1}\right).$$

$$\alpha^{jklm}A^{k}A^{l}A^{m}\mathbf{w}_i^j = n^{-3}\sum_{h=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\alpha^{jklm}\mathbf{w}_h^k\mathbf{w}_s^l\mathbf{w}_t^m\mathbf{w}_i^j;$$

$$n\mathbb{E}\left\{T_{19}\right\} = -\frac{\rho^{(4)}\left(0\right)}{6}\mathbb{E}\left\{\sum_{i=1}^{n}\alpha^{jklm}A^{k}A^{l}A^{m}\mathbf{w}_i^j K_h\left(X_i - x\right)\right\} = \mathcal{O}\left(n^{-1}\right).$$

Therefore

$$\begin{aligned} n\mathbb{E}\left\{T_1\right\} &= -\mathbf{w}^j(x)\mathbf{w}^j(x)f(x) + \left(1+\frac{\rho^{(3)}\left(0\right)}{2}\right)\alpha^{jkl}\delta^{kl}\mathbf{w}^j(x)f(x) \\ &\quad + \frac{1}{2}\mu_2(K)h^2\left[-C_1 + \left(1+\frac{\rho^{(3)}\left(0\right)}{2}\right)\alpha^{jkl}\delta^{kl}C_2^j\right] + \mathcal{O}\left(h^4\right). \end{aligned}$$

Note that the contribution from terms of order $\mathcal{O}_p\left(n^{-3/2}\right)$ in expansion (3.11) for $\boldsymbol{\theta}$ is of order less than $1/n$.

To obtain $\mathbb{E}\left\{T_2\right\}$, square (3.26) and keep the three leading terms only:

$$\left(\boldsymbol{\theta}^j\mathbf{w}_i^j\right)^2 = \left(A^j\mathbf{w}_i^j\right)^2 - 2A^jA^{jk}A^k\mathbf{w}_i^j\mathbf{w}_i^j - \rho^{(3)}\left(0\right)\alpha^{jkl}A^jA^kA^l\mathbf{w}_i^j\mathbf{w}_i^j + R_n. \tag{3.27}$$

Proceeding in a fashion similar to that above yields

$$\begin{aligned} n\mathbb{E}\left\{T_{21}\right\} &= \mathbb{E}\left\{\sum_{i=1}^n\left(A^j\mathbf{w}_i^j\right)^2K_h\left(X_i-x\right)\right\} = n^{-2}\mathbb{E}\left\{\sum_{i=1}^n\sum_{h=1}^n\sum_{s=1}^n\mathbf{w}_h^j\mathbf{w}_s^j\mathbf{w}_i^j\mathbf{w}_i^jK_h\left(X_i-x\right)\right\} \\ &= \mathbb{E}\left\{\mathbf{w}_2^j\mathbf{w}_2^j\mathbf{w}_1^j\mathbf{w}_1^jK_h\left(X_1-x\right)\right\} + \mathcal{O}\left(n^{-1}\right) = \mathbf{w}^j(x)\mathbf{w}^j(x)f(x) + \frac{1}{2}C_1\mu_2(K)h^2 + \mathcal{O}\left(h^4\right). \end{aligned}$$

$$\begin{aligned} A^jA^{jk}A^k\mathbf{w}_i^j\mathbf{w}_i^j &= \left[n^{-1}\sum_{h=1}^n\mathbf{w}_h^j\right]\left[n^{-1}\sum_{s=1}^n\mathbf{w}_s^j\mathbf{w}_s^k - \delta^{jk}\right]\left[n^{-1}\sum_{t=1}^n\mathbf{w}_t^k\right]\mathbf{w}_i^j\mathbf{w}_i^j \\ &= n^{-3}\sum_{h=1}^n\sum_{s=1}^n\sum_{t=1}^n\mathbf{w}_h^j\mathbf{w}_s^j\mathbf{w}_s^k\mathbf{w}_t^k\mathbf{w}_i^j\mathbf{w}_i^j - n^{-2}\sum_{h=1}^n\sum_{t=1}^n\delta^{jk}\mathbf{w}_h^j\mathbf{w}_t^k\mathbf{w}_i^j\mathbf{w}_i^j; \\ n\mathbb{E}\left\{T_{22}\right\} &= -2\mathbb{E}\left\{\sum_{i=1}^nA^jA^{jk}A^k\mathbf{w}_i^j\mathbf{w}_i^jK_h\left(X_i-x\right)\right\} \\ &= -2n^{-3}\mathbb{E}\left\{\sum_{i=1}^n\sum_{h=1}^n\sum_{s=1}^n\sum_{t=1}^n\mathbf{w}_h^j\mathbf{w}_s^j\mathbf{w}_s^k\mathbf{w}_t^k\mathbf{w}_i^j\mathbf{w}_i^jK_h\left(X_i-x\right)\right\} \\ &\quad + 2n^{-2}\mathbb{E}\left\{\sum_{i=1}^n\sum_{h=1}^n\sum_{t=1}^n\delta^{jk}\mathbf{w}_h^j\mathbf{w}_t^k\mathbf{w}_i^j\mathbf{w}_i^jK_h\left(X_i-x\right)\right\} \\ &= \mathcal{O}\left(n^{-1}\right). \\ \alpha^{jkl}A^jA^kA^l\mathbf{w}_i^j\mathbf{w}_i^j &= \alpha^{jkl}\left[n^{-1}\sum_{h=1}^n\mathbf{w}_h^j\right]\left[n^{-1}\sum_{s=1}^n\mathbf{w}_s^k\right]\left[n^{-1}\sum_{t=1}^n\mathbf{w}_t^l\right]\mathbf{w}_i^j\mathbf{w}_i^j \\ &= n^{-3}\sum_{h=1}^n\sum_{s=1}^n\sum_{t=1}^n\alpha^{jkl}\mathbf{w}_h^j\mathbf{w}_s^k\mathbf{w}_t^l\mathbf{w}_i^j\mathbf{w}_i^j; \\ n\mathbb{E}\left\{T_{23}\right\} &= -\rho^{(3)}\left(0\right)\mathbb{E}\left\{\sum_{i=1}^n\alpha^{jkl}A^jA^kA^l\mathbf{w}_i^j\mathbf{w}_i^jK_h\left(X_i-x\right)\right\} = \mathcal{O}\left(n^{-1}\right). \end{aligned}$$

Thus,

$$n\mathbb{E}\left\{T_2\right\} = \mathbf{w}^j(x)\mathbf{w}^j(x)f(x) + \frac{1}{2}C_1\mu_2(K)h^2 + \mathcal{O}\left(h^4\right).$$

3.A.6.2 Variance of GELKDE

First write

$$\tilde{f}_\rho^2(x) = \hat{f}^2(x) + 2\hat{f}(x)T_1 - \rho^{(3)}\left(0\right)\hat{f}(x)T_2 + 2\hat{f}(x)T_3 + T_1^2 + R_n,$$

where the remainder term contains T_2^2, T_3^2 and cross-products, each of which gives a contribution of order $\mathcal{O}\left(n^{-2}\right)$ or smaller. Using the results of previous subsection, we obtain

$$
\begin{aligned}
n\mathbb{E}\left\{\hat{f}(x)T_1\right\} &= -2n^{-2}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{w}_s^j\mathbf{w}_t^j K_h\left(X_i-x\right)K_h\left(X_t-x\right)\right\} \\
&\quad + n^{-3}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{w}_s^j\mathbf{w}_s^k\mathbf{w}_t^k\mathbf{w}_r^j K_h\left(X_r-x\right)K_h\left(X_i-x\right)\right\} \\
&\quad + n^{-3}\frac{\rho^{(3)}\left(0\right)}{2}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\alpha^{jkl}\mathbf{w}_s^k\mathbf{w}_t^l\mathbf{w}_r^j K_h\left(X_r-x\right)K_h\left(X_i-x\right)\right\} + R_n \\
&= -2\mathbb{E}\left\{\mathbf{w}_2^j\mathbf{w}_2^j K_h\left(X_1-x\right)K_h\left(X_2-x\right)\right\} - 2\mathbb{E}\left\{\mathbf{w}_2^j\mathbf{w}_1^j K_h\left(X_2-x\right)K_h\left(X_1-x\right)\right\} \\
&\quad + \mathbb{E}\left\{\mathbf{w}_1^j\mathbf{w}_1^k\mathbf{w}_1^k\mathbf{w}_2^j K_h\left(X_2-x\right)K_h\left(X_3-x\right)\right\} \\
&\quad + \mathbb{E}\left\{\mathbf{w}_1^j\mathbf{w}_1^k\mathbf{w}_2^k\mathbf{w}_2^j K_h\left(X_2-x\right)K_h\left(X_3-x\right)\right\} \\
&\quad + \mathbb{E}\left\{\mathbf{w}_1^j\mathbf{w}_1^k\mathbf{w}_2^k\mathbf{w}_3^j K_h\left(X_3-x\right)K_h\left(X_2-x\right)\right\} \\
&\quad + \frac{\rho^{(3)}\left(0\right)}{2}\alpha^{jkl}\mathbb{E}\left\{\mathbf{w}_3^k\mathbf{w}_3^l\mathbf{w}_2^j K_h\left(X_2-x\right)K_h\left(X_1-x\right)\right\} + \mathcal{O}\left(n^{-1}\right) \\
&= -\mathbb{E}\left\{\mathbf{w}_2^j\mathbf{w}_2^j K_h\left(X_1-x\right)K_h\left(X_2-x\right)\right\} - \mathbb{E}\left\{\mathbf{w}_2^j\mathbf{w}_1^j K_h\left(X_2-x\right)K_h\left(X_1-x\right)\right\} \\
&\quad + \left(1+\frac{\rho^{(3)}\left(0\right)}{2}\right)\mathbb{E}\left\{\alpha^{jkl}\delta^{kl}\mathbf{w}_2^j K_h\left(X_2-x\right)K_h\left(X_1-x\right)\right\} + \mathcal{O}\left(n^{-1}\right) \\
&= -\left[\mathbf{w}^j(x)\mathbf{w}^j(x)f(x) + \frac{1}{2}C_1\mu_2(K)h^2 + \mathcal{O}\left(h^4\right)\right]\left[f(x) + \frac{1}{2}h^2\mu_2(K)f^{(2)}(x) + \mathcal{O}\left(h^4\right)\right] \\
&\quad - \left[\mathbf{w}^j(x)f(x) + \frac{1}{2}C_2^j\mu_2(K)h^2 + \mathcal{O}\left(h^4\right)\right]\left[\mathbf{w}^j(x)f(x) + \frac{1}{2}C_2^j\mu_2(K)h^2 + \mathcal{O}\left(h^4\right)\right] \\
&\quad + \left(1+\frac{\rho^{(3)}\left(0\right)}{2}\right)\alpha^{jkl}\delta^{kl}\left[\mathbf{w}^j(x)f(x) + \frac{1}{2}C_2^j\mu_2(K)h^2 + \mathcal{O}\left(h^4\right)\right]\times\cdots \\
&\quad \cdots\times\left[f(x) + \frac{1}{2}h^2\mu_2(K)f^{(2)}(x) + \mathcal{O}\left(h^4\right)\right] + \mathcal{O}\left(n^{-1}\right) \\
&= -2\mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x) + \left(1+\frac{\rho^{(3)}\left(0\right)}{2}\right)\alpha^{jkl}\delta^{kl}\mathbf{w}^j(x)f^2(x) \\
&\quad - \frac{1}{2}h^2\mu_2(K)\mathbf{w}^j(x)\mathbf{w}^j(x)f^{(2)}(x)f(x) - \frac{1}{2}h^2\mu_2(K)C_1f(x) - h^2\mu_2(K)\mathbf{w}^j(x)C_2^jf(x) \\
&\quad + h^2\mu_2(K)\frac{2+\rho^{(3)}\left(0\right)}{4}\alpha^{jkl}\delta^{kl}\left[\mathbf{w}^j(x)f^{(2)}(x)f(x) + C_2^jf(x)\right] + \mathcal{O}\left(h^4\right).
\end{aligned}
$$

$$
\begin{aligned}
n\mathbb{E}\left\{\hat{f}(x)T_2\right\} &= n^{-3}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{m=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{w}_s^j\mathbf{w}_t^j\mathbf{w}_i^j\mathbf{w}_i^j K_h\left(X_i-x\right)K_h\left(X_m-x\right)\right\}\\
&\quad -2\left[n^{-4}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{m=1}^{n}\sum_{h=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{w}_h^j\mathbf{w}_s^j\mathbf{w}_s^k\mathbf{w}_t^k\mathbf{w}_i^j\mathbf{w}_i^j K_h\left(X_i-x\right)K_h\left(X_m-x\right)\right\}\right.\\
&\quad \left. -n^{-3}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{m=1}^{n}\sum_{h=1}^{n}\sum_{t=1}^{n}\delta^{jk}\mathbf{w}_h^j\mathbf{w}_t^k\mathbf{w}_i^j\mathbf{w}_i^j K_h\left(X_i-x\right)K_h\left(X_m-x\right)\right\}\right]\\
&\quad -\rho^{(3)}\left(0\right)n^{-4}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{m=1}^{n}\sum_{h=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\alpha^{jkl}\mathbf{w}_h^j\mathbf{w}_s^k\mathbf{w}_t^l\mathbf{w}_i^j\mathbf{w}_i^j K_h\left(X_i-x\right)K_h\left(X_m-x\right)\right\}\\
&= \mathbb{E}\left\{\mathbf{w}_2^j\mathbf{w}_2^j\mathbf{w}_1^j\mathbf{w}_1^j K_h\left(X_1-x\right)K_h\left(X_3-x\right)\right\}\\
&\quad -2\left[\mathbb{E}\left\{\mathbf{w}_4^j\mathbf{w}_3^j\mathbf{w}_3^k\mathbf{w}_4^k\mathbf{w}_2^j\mathbf{w}_2^j K_h\left(X_2-x\right)K_h\left(X_1-x\right)\right\}\right.\\
&\qquad \left. -\mathbb{E}\left\{\delta^{jk}\mathbf{w}_3^j\mathbf{w}_3^k\mathbf{w}_2^j\mathbf{w}_2^j K_h\left(X_2-x\right)K_h\left(X_1-x\right)\right\}\right]+\mathcal{O}\left(n^{-1}\right)\\
&= \mathbb{E}\left\{\mathbf{w}_1^j\mathbf{w}_1^j K_h\left(X_1-x\right)K_h\left(X_3-x\right)\right\}+\mathcal{O}\left(n^{-1}\right)\\
&= \mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x)+\frac{1}{2}h^2\mu_2(K)\mathbf{w}^j(x)\mathbf{w}^j(x)f^{(2)}(x)f(x)+\frac{1}{2}h^2\mu_2(K)C_1f(x)+\mathcal{O}\left(h^4\right).
\end{aligned}
$$

$$
\begin{aligned}
n\mathbb{E}\left\{T_1^2\right\} &= 4n^{-3}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\mathbf{w}_s^j\mathbf{w}_i^j\mathbf{w}_t^l\mathbf{w}_k^l K_h\left(X_i-x\right)K_h\left(X_k-x\right)\right\}\\
&\quad -4n^{-4}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{s=1}^{n}\sum_{r=1}^{n}\sum_{t=1}^{n}\mathbf{w}_s^j\mathbf{w}_i^j\mathbf{w}_r^l\mathbf{w}_r^m\mathbf{w}_t^m\mathbf{w}_k^l K_h\left(X_i-x\right)K_h\left(X_k-x\right)\right\}\\
&\quad -2\rho^{(3)}\left(0\right)n^{-4}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{s=1}^{n}\sum_{r=1}^{n}\sum_{t=1}^{n}\alpha^{mpl}\mathbf{w}_s^j\mathbf{w}_i^j\mathbf{w}_r^p\mathbf{w}_t^l\mathbf{w}_k^m K_h\left(X_i-x\right)K_h\left(X_k-x\right)\right\}\\
&\quad +n^{-5}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\sum_{q=1}^{n}\sum_{r=1}^{n}\mathbf{w}_s^j\mathbf{w}_s^k\mathbf{w}_t^k\mathbf{w}_i^j\mathbf{w}_q^l\mathbf{w}_q^m\mathbf{w}_r^m\mathbf{w}_k^l K_h\left(X_i-x\right)K_h\left(X_k-x\right)\right\}\\
&\quad +\rho^{(3)}\left(0\right)n^{-5}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\sum_{q=1}^{n}\sum_{r=1}^{n}\alpha^{mpg}\mathbf{w}_s^j\mathbf{w}_s^l\mathbf{w}_t^l\mathbf{w}_i^j\mathbf{w}_q^p\mathbf{w}_r^g\mathbf{w}_k^m K_h\left(X_i-x\right)K_h\left(X_k-x\right)\right\}\\
&\quad +\frac{\left[\rho^{(3)}\left(0\right)\right]^2}{4}n^{-5}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\sum_{q=1}^{n}\sum_{r=1}^{n}\alpha^{jkl}\alpha^{mpg}\mathbf{w}_s^k\mathbf{w}_t^l\mathbf{w}_i^j\mathbf{w}_q^p\mathbf{w}_r^g\mathbf{w}_k^m K_h\left(X_i-x\right)K_h\left(X_k-x\right)\right\}\\
&= 4\mathbb{E}\left\{\mathbf{w}_3^j\mathbf{w}_2^j\mathbf{w}_3^l\mathbf{w}_1^l K_h\left(X_2-x\right)K_h\left(X_1-x\right)\right\}\\
&\quad -4\mathbb{E}\left\{\mathbf{w}_4^j\mathbf{w}_2^j\mathbf{w}_3^l\mathbf{w}_3^m\mathbf{w}_4^m\mathbf{w}_1^l K_h\left(X_2-x\right)K_h\left(X_1-x\right)\right\}\\
&\quad +\mathbb{E}\left\{\mathbf{w}_4^j\mathbf{w}_4^k\mathbf{w}_5^k\mathbf{w}_2^j\mathbf{w}_3^l\mathbf{w}_3^m\mathbf{w}_5^m\mathbf{w}_1^l K_h\left(X_2-x\right)K_h\left(X_1-x\right)\right\}+\mathcal{O}\left(n^{-1}\right)\\
&= \mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x)+h^2\mu_2(K)\mathbf{w}^j(x)C_2^jf(x)+\mathcal{O}\left(h^4\right).
\end{aligned}
$$

$$\begin{aligned}\mathbb{E}\left\{\hat{f}(x)T_3\right\} &= -n^{-2}\mathbb{E}\left\{\boldsymbol{\theta}^j A^j \sum_{i=1}^n \sum_{t=1}^n K_h\left(X_i - x\right) K_h\left(X_t - x\right)\right\} \\ &\qquad + n^{-2}\frac{\rho^{(3)}\left(0\right)}{2}\mathbb{E}\left\{\boldsymbol{\theta}^j\boldsymbol{\theta}^j \sum_{i=1}^n \sum_{t=1}^n K_h\left(X_i - x\right) K_h\left(X_t - x\right)\right\} \\ &= n^{-1}k_\rho q f^2(x) + n^{-1}h^2\mu_2(K)k_\rho q f(x) f^{(2)}(x) + o\left(n^{-1}h^4\right).\end{aligned}$$

Combining terms yields

$$\begin{aligned}\mathbb{E}\left\{\tilde{f}^2_\rho(x)\right\} = \mathbb{E}\left\{\hat{f}^2(x)\right\} &- n^{-1}\mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x) \\ &+ 2n^{-1}k_\rho\left[-\mathbf{w}^j(x)\mathbf{w}^j(x) + \alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) + q\right] f^2(x) \\ &+ n^{-1}h^2\mu_2(K)k_\rho\left[-\mathbf{w}^j(x)\mathbf{w}^j(x) + \alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) + q\right] f^{(2)}(x) f(x) \\ &+ n^{-1}h^2\mu_2(K)k_\rho\left[-C_1 + \alpha^{jkl}\delta^{kl}C_2^j + q f^{(2)}(x)\right] f(x) \\ &- n^{-1}h^2\mu_2(K)\mathbf{w}^j(x)C_2^j f(x) + o\left(n^{-1}h^4\right).\end{aligned}$$

3.A.7 Proof of Proposition 4

Using (3.10) with $\hat{v}_i = \boldsymbol{\theta}^j\widehat{\mathbf{w}}_i^j$ write

$$\tilde{f}_\rho(x) = \hat{f}(x) + T_1' - \frac{\rho^{(3)}\left(0\right)}{2}T_2' + T_3' + \sum_{i=1}^n R_n^{[\pi]} K_h\left(X_i - x\right),$$

where

$$T_1' = n^{-1}\sum_{i=1}^n \hat{v}_i K_h\left(X_i - x\right), \qquad T_2' = n^{-1}\sum_{i=1}^n \hat{v}_i^2 K_h\left(X_i - x\right),$$

and

$$T_3' = -n^{-1}\boldsymbol{\theta}^\intercal \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{w}}_i \sum_{i=1}^n K_h\left(X_i - x\right) + n^{-1}\frac{\rho^{(3)}\left(0\right)}{2}\boldsymbol{\theta}^j\boldsymbol{\theta}^j\sum_{i=1}^n K_h\left(X_i - x\right).$$

Note that $\boldsymbol{\theta}$ is now given by (3.12). Using (3.21) and (3.25) write

$$\begin{aligned}\hat{v}_i = \boldsymbol{\theta}^j\mathbf{w}_i^j &+ A^j\gamma^{j,r}\omega^{rs}A^k\gamma^{k,s} + A^j\Gamma_i^{j,r}\omega^{rs}A^k\gamma^{k,s} \\ &- \gamma^{j,t}\gamma^{l,u}\omega^{tu}A^l\gamma^{j,r}\omega^{rs}A^k\gamma^{k,s} - \gamma^{j,t}\gamma^{l,u}\omega^{tu}A^l\Gamma_i^{j,r}\omega^{rs}A^k\gamma^{k,s} + R_n,\end{aligned}$$

Define

$$\tau^{i,r;k} = \mathbb{E}\left\{\frac{\partial\mathbf{w}_i^j}{\partial\boldsymbol{\beta}^r}\mathbf{w}_i^k\right\}, \quad \text{and } g^{j,r}(x) = \frac{\partial\mathbf{w}^j(x;\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^r}, \quad j,k\in\{1,\ldots,q\},\ r\in\{1,\ldots,p\}.$$

From the results derived in Appendix 3.A.6.1, we have

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\left\{\sum_{i=1}^{n} A^j \mathbf{w}_i^k K_h\left(X_i - x\right)\right\} &= \frac{1}{n}\mathbf{w}^j(x)\mathbf{w}^k(x)f(x) + \frac{h^2}{n}\frac{1}{2}\mu_2(K)C_1^{jk}(x) + o\left(\frac{h^4}{n}\right), && (3.28a)\\
\frac{1}{n}\mathbb{E}\left\{\sum_{i=1}^{n} A^{jk} A^l \mathbf{w}_i^m K_h\left(X_i - x\right)\right\} &= \frac{1}{n}\alpha^{jkl}\mathbf{w}^m(x)f(x) + \frac{h^2}{n}\frac{1}{2}\mu_2(K)\alpha^{jkl}C_2^m(x) + o\left(\frac{h^4}{n}\right), && (3.28b)\\
\frac{1}{n}\mathbb{E}\left\{\sum_{i=1}^{n} A^j A^k \mathbf{w}_i^l K_h\left(X_i - x\right)\right\} &= \frac{1}{n}\delta^{jk}\mathbf{w}^l(x)f(x) + \frac{h^2}{n}\frac{1}{2}\mu_2(K)\delta^{jk}C_2^l(x) + o\left(\frac{h^4}{n}\right), && (3.28c)
\end{aligned}$$

where $C_1^{jk}(x) = \frac{d^2}{dv^2}\left[\mathbf{w}^j(v)\mathbf{w}^k(v)f(v)\right]\Big|_{v=x}$ and $C_2^m(x) = \frac{d^2}{dv^2}\left[\mathbf{w}^m(v)f(v)\right]\Big|_{v=x}$.

We further obtain that

$$\begin{aligned}
\mathbb{E}\left\{n^{-1}\sum_{i=1}^{n}\Gamma^{j,r}A^k\mathbf{w}_i^l K_h\left(X_i - x\right)\right\} &= n^{-3}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{h=1}^{n}\sum_{g=1}^{n}\frac{\partial \mathbf{w}_h^j}{\partial \boldsymbol{\beta}^r}\mathbf{w}_g^k\mathbf{w}_i^l K_h\left(X_i - x\right)\right\}\\
&\quad - n^{-2}\gamma^{j,r}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{g=1}^{n}\mathbf{w}_g^k\mathbf{w}_i^l K_h\left(X_i - x\right)\right\} = n^{-1}\mathbb{E}\left\{\frac{\partial \mathbf{w}_1^j}{\partial \boldsymbol{\beta}^r}\mathbf{w}_1^k\mathbf{w}_2^l K_h\left(X_2 - x\right)\right\} + o\left(n^{-2}\right)\\
&= n^{-1}\tau^{j,r;k}\mathbf{w}^l(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\tau^{j,r;k}C_2^l(x) + o\left(n^{-1}h^4\right). \quad (3.28d)
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}\left\{n^{-1}\sum_{i=1}^{n}A^jA^k K_h\left(X_i - x\right)\right\} &= n^{-1}\delta^{jk}\mathbb{E}\left\{K_h\left(X_1 - x\right)\right\}\\
&= n^{-1}\delta^{jk}f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{jk}f^{(2)}(x) + o\left(n^{-1}h^4\right); \quad (3.28e)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left\{n^{-1}\sum_{i=1}^{n}\Gamma_i^{j,r}A^kA^l K_h\left(X_i - x\right)\right\} &= n^{-3}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{g=1}^{n}\sum_{h=1}^{n}\mathbf{w}_g^k\mathbf{w}_h^l\frac{\partial \mathbf{w}_i^j}{\partial \boldsymbol{\beta}^r}K_h\left(X_i - x\right)\right\}\\
&\quad - n^{-3}\gamma^{j,r}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{g=1}^{n}\sum_{h=1}^{n}\mathbf{w}_g^k\mathbf{w}_h^l K_h\left(X_i - x\right)\right\}\\
&= n^{-1}\delta^{kl}\mathbb{E}\left\{\frac{\partial \mathbf{w}_1^j}{\partial \boldsymbol{\beta}^r}K_h\left(X_1 - x\right)\right\} - n^{-1}\gamma^{j,r}\delta^{kl}\mathbb{E}\left\{K_h\left(X_1 - x\right)\right\}\\
&= n^{-1}\delta^{kl}g^{j,r}(x)f(x) - n^{-1}\gamma^{j,r}\delta^{kl}f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{kl}C_3^{j,r}(x) - n^{-1}h^2\frac{1}{2}\mu_2(K)\gamma^{j,r}\delta^{kl}f^{(2)}(x)\\
&\quad + o\left(n^{-1}h^4\right), \quad (3.28f)
\end{aligned}$$

where $g^{j,r}(x) = \frac{\partial \mathbf{w}^j(x;\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^r}$, $C_3^{j,r}(x) = \frac{\partial^2}{\partial v^2}\left[\frac{\partial \mathbf{w}^j(v;\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^r})f(v)\right]\Big|_{v=x}$ and

$$\begin{aligned}
\mathbb{E}\left\{\frac{\partial \mathbf{w}_1^j}{\partial \boldsymbol{\beta}^r}K_h\left(X_1 - x\right)\right\} &= \int_{\mathbb{R}}\frac{\partial \mathbf{w}^j(x_1;\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^r}K_h\left(x_1 - x\right)f(x_1)dx_1\\
&= \int_{\mathbb{R}}\frac{\partial \mathbf{w}^j(x+hz;\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^r}K\left(z\right)f(x+hz)dz = \frac{\partial \mathbf{w}^j(x;\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^r}f(x) + \frac{1}{2}h^2\mu_2(K)C_3^{j,r}(x) + o\left(h^4\right).
\end{aligned}$$

We then have

$$\mathbb{E}\left\{T_1'\right\} = n^{-1}B_1'(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)B_2'(x) + \mathcal{O}\left(n^{-1}h^4\right),$$

where

$$\begin{aligned}
B_1'(x) = &-\mathbf{w}^j(x)\mathbf{w}^j(x) + \gamma^{j,r}\gamma^{k,s}\omega^{rs}\mathbf{w}^j(x)\mathbf{w}^k(x)\\
&+ \alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) - \gamma^{k,r}\gamma^{l,s}\omega^{rs}\alpha^{jkl}\mathbf{w}^j(x) - \gamma^{j,r}\gamma^{k,s}\omega^{rs}\alpha^{klm}\delta^{lm}\mathbf{w}^j(x)\\
&\quad + \gamma^{j,r}\gamma^{k,t}\gamma^{l,u}\gamma^{m,s}\omega^{tu}\omega^{rs}\alpha^{mkl}\mathbf{w}^j(x)\\
&+ \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) - \frac{\rho^{(3)}(0)}{2}\alpha^{mkl}\gamma^{j,r}\gamma^{m,s}\omega^{rs}\delta^{kl}\mathbf{w}^j(x)\\
&\quad + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\gamma^{k,r}\gamma^{m,t}\gamma^{l,s}\gamma^{n,u}\omega^{su}\omega^{rt}\delta^{mn}\mathbf{w}^j(x)\\
&\quad - \frac{\rho^{(3)}(0)}{2}\alpha^{okl}\gamma^{j,r}\gamma^{k,s}\gamma^{m,u}\gamma^{l,w}\gamma^{n,v}\gamma^{o,t}\omega^{wv}\omega^{rt}\omega^{su}\delta^{mn}\mathbf{w}^j(x)\\
&\quad + \rho^{(3)}(0)\,\alpha^{nkl}\gamma^{j,r}\gamma^{l,s}\gamma^{m,v}\gamma^{n,t}\omega^{sv}\omega^{rt}\delta^{km}\mathbf{w}^j(x) - \rho^{(3)}(0)\,\alpha^{jkl}\gamma^{l,r}\gamma^{m,s}\omega^{rs}\delta^{km}\mathbf{w}^j(x)\\
&+ \frac{1}{2}\gamma^{m,sv}\gamma^{j,r}\gamma^{k,u}\gamma^{l,w}\gamma^{m,t}\omega^{vw}\omega^{rt}\omega^{su}\delta^{kl}\mathbf{w}^j(x) - \frac{1}{2}\gamma^{j,rs}\gamma^{k,t}\gamma^{l,u}\omega^{su}\omega^{rt}\delta^{kl}\mathbf{w}^j(x)\\
&\quad - \gamma^{l,tv}\gamma^{j,r}\gamma^{k,u}\omega^{rt}\omega^{vu}\delta^{kl}\mathbf{w}^j(x) + \gamma^{m,tv}\gamma^{j,r}\gamma^{m,s}\gamma^{k,u}\gamma^{l,w}\omega^{vw}\omega^{rt}\omega^{su}\delta^{kl}\mathbf{w}^j(x)\\
&\quad + \gamma^{k,s}\omega^{rs}\tau^{j,r;k}\mathbf{w}^j(x) + \gamma^{j,r}\omega^{rs}\tau^{k,s;k}\mathbf{w}^j(x)\\
&\quad - \gamma^{j,r}\gamma^{k,u}\gamma^{l,t}\omega^{rt}\omega^{su}\tau^{l,s;k}\mathbf{w}^j(x) - \gamma^{j,r}\gamma^{l,s}\gamma^{k,u}\omega^{rt}\omega^{su}\tau^{l,t;k}\mathbf{w}^j(x)\\
&+ \omega^{rs}\gamma^{k,s}\delta^{jk}g^{j,r}(x) - \gamma^{j,t}\gamma^{l,u}\omega^{tu}\omega^{rs}\gamma^{k,s}\delta^{kl}g^{j,r}(x)
\end{aligned}$$

and

$$\begin{aligned}
B_2'(x) = &-C_1^{jj}(x) + \gamma^{j,r}\gamma^{k,s}\omega^{rs}C_1^{kj}(x)\\
&+ \alpha^{jkl}\delta^{kl}C_2^j(x) - \gamma^{k,r}\gamma^{l,s}\omega^{rs}\alpha^{jkl}C_2^j(x) - \gamma^{j,r}\gamma^{k,s}\omega^{rs}\alpha^{klm}\delta^{lm}C_2^j(x)\\
&\quad + \gamma^{j,r}\gamma^{k,t}\gamma^{l,u}\gamma^{m,s}\omega^{tu}\omega^{rs}\alpha^{mkl}C_2^j(x)\\
&+ \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\delta^{kl}C_2^j(x) - \frac{\rho^{(3)}(0)}{2}\alpha^{mkl}\gamma^{j,r}\gamma^{m,s}\omega^{rs}\delta^{kl}C_2^j(x)\\
&\quad + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\gamma^{k,r}\gamma^{m,t}\gamma^{l,s}\gamma^{n,u}\omega^{su}\omega^{rt}\delta^{mn}C_2^j(x)\\
&\quad - \frac{\rho^{(3)}(0)}{2}\alpha^{okl}\gamma^{j,r}\gamma^{k,s}\gamma^{m,u}\gamma^{l,w}\gamma^{n,v}\gamma^{o,t}\omega^{wv}\omega^{rt}\omega^{su}\delta^{mn}C_2^j(x)\\
&\quad + \rho^{(3)}(0)\,\alpha^{nkl}\gamma^{j,r}\gamma^{l,s}\gamma^{m,v}\gamma^{n,t}\omega^{sv}\omega^{rt}\delta^{km}C_2^j(x) - \rho^{(3)}(0)\,\alpha^{jkl}\gamma^{l,r}\gamma^{m,s}\omega^{rs}\delta^{km}C_2^j(x)\\
&+ \frac{1}{2}\gamma^{m,sv}\gamma^{j,r}\gamma^{k,u}\gamma^{l,w}\gamma^{m,t}\omega^{vw}\omega^{rt}\omega^{su}\delta^{kl}C_2^j(x) - \frac{1}{2}\gamma^{j,rs}\gamma^{k,t}\gamma^{l,u}\omega^{su}\omega^{rt}\delta^{kl}C_2^j(x)\\
&\quad - \gamma^{l,tv}\gamma^{j,r}\gamma^{k,u}\omega^{rt}\omega^{vu}\delta^{kl}C_2^j(x) + \gamma^{m,tv}\gamma^{j,r}\gamma^{m,s}\gamma^{k,u}\gamma^{l,w}\omega^{vw}\omega^{rt}\omega^{su}\delta^{kl}C_2^j(x)\\
&\quad + \gamma^{k,s}\omega^{rs}\tau^{j,r;k}C_2^j(x) + \gamma^{j,r}\omega^{rs}\tau^{k,s;k}C_2^j(x)\\
&\quad - \gamma^{j,r}\gamma^{k,u}\gamma^{l,t}\omega^{rt}\omega^{su}\tau^{l,s;k}C_2^j(x) - \gamma^{j,r}\gamma^{l,s}\gamma^{k,u}\omega^{rt}\omega^{su}\tau^{l,t;k}C_2^j(x)\\
&+ \omega^{rs}\gamma^{k,s}\delta^{jk}C_3^{j,r}(x) - \gamma^{j,t}\gamma^{l,u}\omega^{tu}\omega^{rs}\gamma^{k,s}\delta^{kl}C_3^{j,r}(x).
\end{aligned}$$

Note that only three terms in $\hat{v}_i^2$ give a contribution of order $\mathcal{O}\left(n^{-1}\right)$, viz. $A^jA^k\mathbf{w}_i^j\mathbf{w}_i^k$, $-2\gamma^{l,r}\gamma^{k,s}\omega^{rs}A^jA^k\mathbf{w}_i^j\mathbf{w}_i^l$ and $\gamma^{j,r}\gamma^{k,s}\gamma^{l,t}\gamma^{m,u}\omega^{rs}\omega^{tu}A^kA^m\mathbf{w}_i^j\mathbf{w}_i^l$. Also

$$\frac{1}{n}\mathbb{E}\left\{\sum_{i=1}^{n}A^jA^k\mathbf{w}_i^l\mathbf{w}_i^m K_h\left(X_i - x\right)\right\} = n^{-1}\delta^{jk}\mathbf{w}^l(x)\mathbf{w}^m(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{jk}C_1^{lm}(x) + \mathcal{O}\left(n^{-1}h^4\right).$$

Hence

$$\mathbb{E}\left\{T_2'\right\} = n^{-1}B_1''(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)B_2''(x) + o\left(n^{-1}h^4\right),$$

where

$$B_1''(x) = \mathbf{w}^j(x)\mathbf{w}^j(x) - 2\gamma^{l,r}\gamma^{k,s}\omega^{rs}\delta^{jk}\mathbf{w}^j(x)\mathbf{w}^l(x) + \gamma^{j,r}\gamma^{k,s}\gamma^{l,t}\gamma^{m,u}\omega^{rs}\omega^{tu}\delta^{km}\mathbf{w}^j(x)\mathbf{w}^l(x)$$

and

$$B_2''(x) = C_1^{jj}(x) - 2\gamma^{l,r}\gamma^{k,s}\omega^{rs}\delta^{jk}C_1^{jl}(x) + \gamma^{j,r}\gamma^{k,s}\gamma^{l,t}\gamma^{m,u}\omega^{rs}\omega^{tu}\delta^{km}C_1^{jl}(x).$$

Finally, write

$$\begin{aligned}\boldsymbol{\theta}^{\mathsf{T}}\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{w}}_i &= \boldsymbol{\theta}^j A^j - \boldsymbol{\theta}^j\gamma^{j,r}\omega^{rs}A^k\gamma^{k,s} + o_p\left(n^{-3/2}\right)\\ &= -A^jA^j + 2A^j\gamma^{j,r}\omega^{rs}A^k\gamma^{k,s} - \gamma^{j,r}\omega^{rs}A^k\gamma^{k,s}\gamma^{j,t}\omega^{tu}A^l\gamma^{l,u} + o_p\left(n^{-3/2}\right),\end{aligned}$$

and

$$\boldsymbol{\theta}^j\boldsymbol{\theta}^j = A^jA^j - 2A^j\gamma^{j,r}\omega^{rs}A^k\gamma^{k,s} + \gamma^{j,r}\omega^{rs}A^k\gamma^{k,s}\gamma^{j,t}\omega^{tu}A^l\gamma^{l,u} + o_p\left(n^{-3/2}\right).$$

Then, as

$$\begin{aligned}&\mathbb{E}\left\{A^jA^j - 2A^j\gamma^{j,r}\omega^{rs}A^k\gamma^{k,s} + \gamma^{j,r}\omega^{rs}A^k\gamma^{k,s}\gamma^{j,t}\omega^{tu}A^l\gamma^{l,u}\right\}\\ &\quad= n^{-1}\left(q - 2\delta^{jk}\gamma^{j,r}\omega^{rs}\gamma^{k,s} + \delta^{kl}\gamma^{j,r}\omega^{rs}\gamma^{k,s}\gamma^{j,t}\omega^{tu}\gamma^{l,u}\right) = n^{-1}\left(q - 2p + p\right) = n^{-1}(q-p),\end{aligned}$$

we have

$$\mathbb{E}\left\{T_3'\right\} = n^{-1}k_\rho(q-p)\left[f(x) + \frac{1}{2}h^2\mu_2(K)f^{(2)}(x)\right] + o_p\left(n^{-3/2}\right).$$

Thus,

$$\mathbb{E}\left\{\tilde{f}_\rho(x)\right\} = \mathbb{E}\left\{\hat{f}(x)\right\} + n^{-1}B_1(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)B_2(x) + o\left(n^{-3/2}\right),$$

where

$$B_1(x) = B_1'(x) - \frac{\rho^{(3)}(0)}{2}B_1''(x) + k_\rho(q-p)$$

and

$$B_2(x) = B_2'(x) - \frac{\rho^{(3)}(0)}{2}B_2''(x) + k_\rho(q-p)f^{(2)}(x).$$

Specifically,

$$\begin{aligned}
B_1(x) = &-k_\rho \mathbf{w}^j(x)\mathbf{w}^j(x) + \gamma^{j,r}\gamma^{k,s}\omega^{rs}\mathbf{w}^j(x)\mathbf{w}^k(x) \\
&\quad + \rho^{(3)}(0)\,\gamma^{l,r}\gamma^{k,s}\omega^{rs}\delta^{jk}\mathbf{w}^j(x)\mathbf{w}^l(x) \\
&\quad - \frac{\rho^{(3)}(0)}{2}\gamma^{j,r}\gamma^{k,s}\gamma^{l,t}\gamma^{m,u}\omega^{rs}\omega^{tu}\delta^{km}\mathbf{w}^j(x)\mathbf{w}^l(x) \\
&+ k_\rho\alpha^{jkl}\delta^{kl}\mathbf{w}^j(x) - \gamma^{k,r}\gamma^{l,s}\omega^{rs}\alpha^{jkl}\mathbf{w}^j(x) - \gamma^{j,r}\gamma^{k,s}\omega^{rs}\alpha^{klm}\delta^{lm}\mathbf{w}^j(x) \\
&\quad + \gamma^{j,r}\gamma^{k,t}\gamma^{l,u}\gamma^{m,s}\omega^{tu}\omega^{rs}\alpha^{mkl}\mathbf{w}^j(x) \\
&- \frac{\rho^{(3)}(0)}{2}\alpha^{mkl}\gamma^{j,r}\gamma^{m,s}\omega^{rs}\delta^{kl}\mathbf{w}^j(x) \\
&\quad + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\gamma^{k,r}\gamma^{m,t}\gamma^{l,s}\gamma^{n,u}\omega^{su}\omega^{rt}\delta^{mn}\mathbf{w}^j(x) \\
&\quad - \frac{\rho^{(3)}(0)}{2}\alpha^{okl}\gamma^{j,r}\gamma^{k,s}\gamma^{m,u}\gamma^{l,w}\gamma^{n,v}\gamma^{o,t}\omega^{wv}\omega^{rt}\omega^{su}\delta^{mn}\mathbf{w}^j(x) \\
&\quad + \rho^{(3)}(0)\,\alpha^{nkl}\gamma^{j,r}\gamma^{l,s}\gamma^{m,v}\gamma^{n,t}\omega^{sv}\omega^{rt}\delta^{km}\mathbf{w}^j(x) - \rho^{(3)}(0)\,\alpha^{jkl}\gamma^{l,r}\gamma^{m,s}\omega^{rs}\delta^{km}\mathbf{w}^j(x) \\
&+ \frac{1}{2}\gamma^{m,sv}\gamma^{j,r}\gamma^{k,u}\gamma^{l,w}\gamma^{m,t}\omega^{vw}\omega^{rt}\omega^{su}\delta^{kl}\mathbf{w}^j(x) - \frac{1}{2}\gamma^{j,rs}\gamma^{k,t}\gamma^{l,u}\omega^{su}\omega^{rt}\delta^{kl}\mathbf{w}^j(x) \\
&\quad - \gamma^{l,tv}\gamma^{j,r}\gamma^{k,u}\omega^{rt}\omega^{vu}\delta^{kl}\mathbf{w}^j(x) + \gamma^{m,tv}\gamma^{j,r}\gamma^{m,s}\gamma^{k,u}\gamma^{l,w}\omega^{vw}\omega^{rt}\omega^{su}\delta^{kl}\mathbf{w}^j(x) \\
&\quad + \gamma^{k,s}\omega^{rs}\tau^{j,r;k}\mathbf{w}^j(x) + \gamma^{j,r}\omega^{rs}\tau^{k,s;k}\mathbf{w}^j(x) \\
&\quad - \gamma^{j,r}\gamma^{k,u}\gamma^{l,t}\omega^{rt}\omega^{su}\tau^{l,s;k}\mathbf{w}^j(x) - \gamma^{j,r}\gamma^{l,s}\gamma^{k,u}\omega^{rt}\omega^{su}\tau^{l,t;k}\mathbf{w}^j(x) \\
&+ \omega^{rs}\gamma^{k,s}\delta^{jk}g^{j,r}(x) - \gamma^{j,t}\gamma^{l,u}\omega^{tu}\omega^{rs}\gamma^{k,s}\delta^{kl}g^{j,r}(x) + k_\rho(q-p),
\end{aligned} \tag{3.29}$$

and

$$\begin{aligned}
B_2(x) = &-k_\rho C_1^{jj}(x) + \gamma^{j,r}\gamma^{k,s}\omega^{rs}C_1^{kj}(x) \\
&\quad + \rho^{(3)}(0)\,\gamma^{l,r}\gamma^{k,s}\omega^{rs}\delta^{jk}C_1^{jl}(x) - \frac{\rho^{(3)}(0)}{2}\gamma^{j,r}\gamma^{k,s}\gamma^{l,t}\gamma^{m,u}\omega^{rs}\omega^{tu}\delta^{km}C_1^{jl}(x) \\
&+ k_\rho\alpha^{jkl}\delta^{kl}C_2^j(x) - \gamma^{k,r}\gamma^{l,s}\omega^{rs}\alpha^{jkl}C_2^j(x) - \gamma^{j,r}\gamma^{k,s}\omega^{rs}\alpha^{klm}\delta^{lm}C_2^j(x) \\
&\quad + \gamma^{j,r}\gamma^{k,t}\gamma^{l,u}\gamma^{m,s}\omega^{tu}\omega^{rs}\alpha^{mkl}C_2^j(x) \\
&- \frac{\rho^{(3)}(0)}{2}\alpha^{mkl}\gamma^{j,r}\gamma^{m,s}\omega^{rs}\delta^{kl}C_2^j(x) \\
&\quad + \frac{\rho^{(3)}(0)}{2}\alpha^{jkl}\gamma^{k,r}\gamma^{m,t}\gamma^{l,s}\gamma^{n,u}\omega^{su}\omega^{rt}\delta^{mn}C_2^j(x) \\
&\quad - \frac{\rho^{(3)}(0)}{2}\alpha^{okl}\gamma^{j,r}\gamma^{k,s}\gamma^{m,u}\gamma^{l,w}\gamma^{n,v}\gamma^{o,t}\omega^{wv}\omega^{rt}\omega^{su}\delta^{mn}C_2^j(x) \\
&\quad + \rho^{(3)}(0)\,\alpha^{nkl}\gamma^{j,r}\gamma^{l,s}\gamma^{m,v}\gamma^{n,t}\omega^{sv}\omega^{rt}\delta^{km}C_2^j(x) - \rho^{(3)}(0)\,\alpha^{jkl}\gamma^{l,r}\gamma^{m,s}\omega^{rs}\delta^{km}C_2^j(x) \\
&+ \frac{1}{2}\gamma^{m,sv}\gamma^{j,r}\gamma^{k,u}\gamma^{l,w}\gamma^{m,t}\omega^{vw}\omega^{rt}\omega^{su}\delta^{kl}C_2^j(x) - \frac{1}{2}\gamma^{j,rs}\gamma^{k,t}\gamma^{l,u}\omega^{su}\omega^{rt}\delta^{kl}C_2^j(x) \\
&\quad - \gamma^{l,tv}\gamma^{j,r}\gamma^{k,u}\omega^{rt}\omega^{vu}\delta^{kl}C_2^j(x) + \gamma^{m,tv}\gamma^{j,r}\gamma^{m,s}\gamma^{k,u}\gamma^{l,w}\omega^{vw}\omega^{rt}\omega^{su}\delta^{kl}C_2^j(x) \\
&\quad + \gamma^{k,s}\omega^{rs}\tau^{j,r;k}C_2^j(x) + \gamma^{j,r}\omega^{rs}\tau^{k,s;k}C_2^j(x) \\
&\quad - \gamma^{j,r}\gamma^{k,u}\gamma^{l,t}\omega^{rt}\omega^{su}\tau^{l,s;k}C_2^j(x) - \gamma^{j,r}\gamma^{l,s}\gamma^{k,u}\omega^{rt}\omega^{su}\tau^{l,t;k}C_2^j(x) \\
&+ \omega^{rs}\gamma^{k,s}\delta^{jk}C_3^{j,r}(x) - \gamma^{j,t}\gamma^{l,u}\omega^{tu}\omega^{rs}\gamma^{k,s}\delta^{kl}C_3^{j,r}(x) + k_\rho(q-p)f^{(2)}(x).
\end{aligned} \tag{3.30}$$

We also immediately obtain that

$$\mathbb{ISB}\left\{\tilde{f}_\rho(x)\right\} = \mathbb{ISB}\left\{\hat{f}(x)\right\} + n^{-1}h^2\mu_2(K)\int_{\mathbb{R}} B_1(x)f^{(2)}(x)f(x)dx + \mathcal{O}\left(n^{-3/2}\right).$$

To obtain the variance first note that

$$\left[\mathbb{E}\left\{\tilde{f}_\rho(x)\right\}\right]^2 = \left[\mathbb{E}\left\{\hat{f}(x)\right\}\right]^2 + n^{-1}2B_1(x)f^2(x) + n^{-1}h^2\mu_2(K)B_1(x)f^{(2)}(x)f(x) + n^{-1}h^2\mu_2(K)B_2(x)f(x) + o\left(n^{-3/2}\right).$$

Analogously to the previous section, we obtain the following results.

$$n^{-2}\mathbb{E}\left\{\sum_{i_1=1}^n\sum_{i_2=1}^n A^j\mathbf{w}_{i_1}^k K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\} = n^{-1}2\mathbf{w}^j(x)\mathbf{w}^k(x)f^2(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\mathbf{w}^j(x)\mathbf{w}^k(x)f^{(2)}(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)C_1^{jk}(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\mathbf{w}^j(x)C_2^k(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\mathbf{w}^k(x)C_2^j(x)f(x) + o\left(n^{-1}h^4\right). \quad (3.31a)$$

$$n^{-2}\mathbb{E}\left\{\sum_{i_1=1}^n\sum_{i_2=1}^n A^{jk}A^l\mathbf{w}_{i_1}^m K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\} = n^{-1}\alpha^{jkl}\mathbf{w}^m(x)f^2(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\alpha^{jkl}\mathbf{w}^m(x)f^{(2)}(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\alpha^{jkl}C_2^m(x)f(x) + o\left(n^{-1}h^4\right). \quad (3.31b)$$

$$n^{-2}\mathbb{E}\left\{\sum_{i_1=1}^n\sum_{i_2=1}^n A^jA^k\mathbf{w}_{i_1}^l K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\} = n^{-1}\delta^{jk}\mathbf{w}^l(x)f^2(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{jk}\mathbf{w}^l(x)f^{(2)}(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{jk}C_2^l(x)f(x) + o\left(n^{-1}h^4\right). \quad (3.31c)$$

$$n^{-2}\mathbb{E}\left\{\sum_{i_1=1}^n\sum_{i_2=1}^n \Gamma^{j,r}A^k\mathbf{w}_{i_1}^l K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\} = n^{-1}\tau^{j,r;k}\mathbf{w}^l(x)f^2(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\tau^{j,r;k}\mathbf{w}^l(x)f^{(2)}(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\tau^{j,r;k}C_2^l(x)f(x) + o\left(n^{-1}h^4\right). \quad (3.31d)$$

$$n^{-2}\mathbb{E}\left\{\sum_{i_1=1}^n\sum_{i_2=1}^n A^jA^k K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\} = n^{-1}\delta^{jk}f^2(x) + n^{-1}h^2\mu_2(K)\delta^{jk}f^{(2)}(x)f(x) + o\left(n^{-1}h^4\right). \quad (3.31e)$$

$$n^{-2}\mathbb{E}\left\{\sum_{i_1=1}^n\sum_{i_2=1}^n \Gamma_{i_1}^{j,r}A^kA^l K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\} = n^{-1}\delta^{kl}g^{j,r}(x)f^2(x) - n^{-1}\gamma^{j,r}\delta^{kl}f^2(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{kl}g^{j,r}(x)f^{(2)}(x)f(x) - n^{-1}h^2\frac{1}{2}\mu_2(K)\gamma^{j,r}\delta^{kl}f^{(2)}(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{kl}C_3^{j,r}(x)f(x) - n^{-1}h^2\frac{1}{2}\mu_2(K)\gamma^{j,r}\delta^{kl}f^{(2)}(x)f(x) + o\left(n^{-1}h^4\right). \quad (3.31f)$$

Noting that expressions (3.28a)–(3.28f) are of the form $n^{-1}\epsilon_{1j}f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\epsilon_{2j} + o\left(n^{-1}h^4\right)$, whereas expressions (3.31b)–(3.31f) are of the form $n^{-1}\epsilon_{1j}f^2(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\epsilon_{1j}f^{(2)}(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\epsilon_{2j}f(x) + o\left(n^{-1}h^4\right)$, and expression (3.31a) is of the same form with three extra terms,

$n^{-1}\mathbf{w}^j(x)\mathbf{w}^k(x)f^2(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\mathbf{w}^j(x)C_2^k(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\mathbf{w}^k(x)C_2^j(x)f(x)$, it follows that

$$\begin{aligned}\mathbb{E}\left\{\hat{f}(x)T_1'\right\} &= n^{-1}B_1'(x)f^2(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)B_1'(x)f^{(2)}(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)B_2'(x)f(x)\\ &\quad - n^{-1}\mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x) + n^{-1}\gamma^{j,r}\gamma^{k,s}\omega^{rs}\mathbf{w}^j(x)\mathbf{w}^k(x)f^2(x)\\ &\quad - n^{-1}h^2\mu_2(K)\mathbf{w}^j(x)C_2^j(x)f(x) + n^{-1}h^2\frac{1}{2}\mu_2(K)\gamma^{j,r}\gamma^{k,s}\omega^{rs}\mathbf{w}^j(x)C_2^k(x)f(x)\\ &\quad + n^{-1}h^2\frac{1}{2}\mu_2(K)\gamma^{j,r}\gamma^{k,s}\omega^{rs}\mathbf{w}^k(x)C_2^j(x)f(x) + o\left(n^{-1}h^4\right).\end{aligned}$$

Similarly, as

$$\begin{aligned}n^{-2}\mathbb{E}\left\{\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}A^jA^k\mathbf{w}_{i_1}^l\mathbf{w}_{i_1}^m K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\} &= n^{-1}\delta^{jk}\mathbf{w}^l(x)\mathbf{w}^m(x)f^2(x)\\ + n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{jk}\mathbf{w}^l(x)\mathbf{w}^m(x)f^{(2)}(x)f(x) &+ n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{jk}C_1^{lm}(x)f(x) + o\left(n^{-1}h^4\right),\end{aligned}$$

we have

$$\mathbb{E}\left\{\hat{f}(x)T_2'\right\} = \frac{1}{n}B_1''(x)f^2(x) + \frac{h^2}{n}\frac{1}{2}\mu_2(K)B_1''(x)f^{(2)}(x)f(x) + \frac{h^2}{n}\frac{1}{2}\mu_2(K)B_2''(x)f(x) + o\left(\frac{h^4}{n}\right).$$

To obtain

$$\mathbb{E}\left\{(T_1')^2\right\} = n^{-2}\mathbb{E}\left\{\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\hat{v}_{i_1}\hat{v}_{i_2}K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\},$$

note that only the $o\left(n^{-1/2}\right)$ terms in the expansion for $\boldsymbol{\theta}$ make a contribution of order $1/n$; therefore,

$$\begin{aligned}\mathbb{E}\left\{(T_1')^2\right\} &= n^{-2}\mathbb{E}\left\{\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}A^jA^k\mathbf{w}_{i_1}^j\mathbf{w}_{i_2}^k K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\}\\ &\quad - 2n^{-2}\gamma^{l,r}\gamma^{k,s}\omega^{rs}\mathbb{E}\left\{\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}A^jA^k\mathbf{w}_{i_1}^j\mathbf{w}_{i_2}^l K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\}\\ &\quad + n^{-2}\gamma^{j,t}\gamma^{m,u}\omega^{tu}\gamma^{l,r}\gamma^{k,s}\omega^{rs}\mathbb{E}\left\{\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}A^kA^m\mathbf{w}_{i_1}^j\mathbf{w}_{i_2}^l K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\} + o\left(n^{-2}\right)\\ &= n^{-1}\mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x) - 2n^{-1}\gamma^{k,r}\gamma^{j,s}\omega^{rs}\mathbf{w}^j(x)\mathbf{w}^k(x)f^2(x)\\ &\quad + n^{-1}\gamma^{j,t}\gamma^{k,u}\omega^{tu}\gamma^{l,r}\gamma^{k,s}\omega^{rs}\mathbf{w}^j(x)\mathbf{w}^l(x)f^2(x) + n^{-1}h^2\mu_2(K)\mathbf{w}^j(x)C_2^j(x)f(x)\\ &\quad - 2n^{-1}h^2\frac{1}{2}\mu_2(K)\gamma^{k,r}\gamma^{j,s}\omega^{rs}\mathbf{w}^j(x)C_2^k(x)f(x) - 2n^{-1}h^2\frac{1}{2}\mu_2(K)\gamma^{k,r}\gamma^{j,s}\omega^{rs}\mathbf{w}^k(x)C_2^j(x)f(x)\\ &\quad + n^{-1}h^2\frac{1}{2}\mu_2(K)\gamma^{j,t}\gamma^{k,u}\omega^{tu}\gamma^{l,r}\gamma^{k,s}\omega^{rs}\mathbf{w}^j(x)C_2^l(x)f(x)\\ &\quad + n^{-1}h^2\frac{1}{2}\mu_2(K)\gamma^{j,t}\gamma^{k,u}\omega^{tu}\gamma^{l,r}\gamma^{k,s}\omega^{rs}\mathbf{w}^l(x)C_2^j(x)f(x) + o\left(n^{-1}h^4\right),\end{aligned}$$

because

$$\begin{aligned}n^{-2}\mathbb{E}\left\{\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}A^jA^k\mathbf{w}_{i_1}^l\mathbf{w}_{i_2}^m K_h\left(X_{i_1}-x\right)K_h\left(X_{i_2}-x\right)\right\} &= n^{-1}\delta^{jk}\mathbf{w}^l(x)\mathbf{w}^m(x)f^2(x)\\ + n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{jk}\mathbf{w}^l(x)C_2^m(x)f(x) &+ n^{-1}h^2\frac{1}{2}\mu_2(K)\delta^{jk}\mathbf{w}^m(x)C_2^l(x)f(x) + o\left(n^{-1}h^4\right).\end{aligned}$$

Finally,

$$\mathbb{E}\left\{\hat{f}(x)T_3'\right\} = n^{-1}k_\rho(q-p)\left[f^2(x) + h^2\mu_2(K)f^{(2)}(x)f(x)\right] + \mathcal{O}\left(n^{-3/2}\right).$$

It can be seen that the contribution from other terms is of order n^{-2} or smaller. Therefore,

$$\begin{aligned}\mathbb{E}\left\{\tilde{f}_\rho^2(x)\right\} &= \mathbb{E}\left\{\hat{f}^2(x)\right\} + n^{-1}2B_1(x)f^2(x) + n^{-1}h^2\mu_2(K)B_1(x)f^{(2)}(x)f(x)\\ &+ n^{-1}h^2\mu_2(K)B_2(x)f(x) - n^{-1}\mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x) + n^{-1}\gamma^{j,t}\gamma^{k,u}\omega^{tu}\gamma^{l,r}\gamma^{k,s}\omega^{rs}\mathbf{w}^j(x)\mathbf{w}^l(x)f^2(x)\\ &- n^{-1}h^2\mu_2(K)\mathbf{w}^j(x)C_2^j(x)f(x) + n^{-1}h^2\mu_2(K)\gamma^{j,t}\gamma^{k,u}\omega^{tu}\gamma^{l,r}\gamma^{k,s}\omega^{rs}\mathbf{w}^j(x)C_2^l(x)f(x) + \mathcal{O}\left(n^{-3/2}\right).\end{aligned}$$

Thus, after simplification,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left\{\tilde{f}_\rho(x)\right\} &= \mathbb{V}\mathrm{ar}\left\{\hat{f}(x)\right\} - n^{-1}\mathbf{w}^j(x)\mathbf{w}^j(x)f^2(x) + n^{-1}\gamma^{j,s}\gamma^{k,r}\omega^{rs}\mathbf{w}^j(x)\mathbf{w}^k(x)f^2(x)\\ &- n^{-1}h^2\mu_2(K)\mathbf{w}^j(x)C_2^j(x)f(x) + n^{-1}h^2\mu_2(K)\gamma^{j,s}\gamma^{k,r}\omega^{rs}\mathbf{w}^j(x)C_2^k(x)f(x) + \mathcal{O}\left(n^{-3/2}\right).\end{aligned}$$

Expressions for the integrated variance and mean integrated squared error follow immediately.

Appendix 3.B CUE with constraint that the mean is zero

Suppose $X_i \overset{\text{iid}}{\sim} N(0,1)$ and $\psi(x_i) = x_i$, that is a known zero mean. Let the kernel be the Gaussian density, i.e. $K_h(z) = \frac{1}{h}\phi(z/h) = \frac{1}{h\sqrt{2\pi}}e^{-\frac{1}{2}\frac{z^2}{h^2}}$. The CUE criterion in this case becomes $P_n(\lambda) = -\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)\lambda^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)\lambda$, a quadratic in λ.

The first-order conditions are

$$\frac{\partial}{\partial\lambda}P_n(\lambda)\bigg|_{\lambda=\hat{\lambda}} = -\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)\hat{\lambda} - \frac{1}{n}\sum_{i=1}^n x_i = 0.$$

Hence

$$\hat{\lambda} = -\frac{\frac{1}{n}\sum_{i=1}^n x_i}{\frac{1}{n}\sum_{i=1}^n x_i^2}.$$

Also

$$\rho^{(1)}(\hat{\lambda}\psi(x_i)) = -\hat{\lambda}x_i - 1 = \frac{\frac{1}{n}\sum_{j=1}^n x_j}{\frac{1}{n}\sum_{j=1}^n x_j^2}x_i - 1 = \frac{\overline{x}}{\overline{x^2}}x_i - 1 = \frac{x_i\overline{x} - \overline{x^2}}{\overline{x^2}},$$

$$\sum_{i=1}^n \rho^{(1)}(\hat{\lambda}\psi(x_i)) = n\frac{\left(\frac{1}{n}\sum_{j=1}^n x_j\right)^2}{\frac{1}{n}\sum_{j=1}^n x_j^2} - n = n\frac{\overline{x}^2 - \overline{x^2}}{\overline{x^2}},$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2$.

Hence, the implied probabilities are

$$\hat{\pi}_i = \frac{1}{n}\frac{\overline{x^2} - x_i\overline{x}}{\overline{x^2} - (\overline{x})^2} = \frac{1}{n}\left[1 + \frac{\overline{x^2} - x_i\overline{x}}{\overline{x^2} - (\overline{x})^2} - 1\right] = \frac{1}{n} + \frac{1}{n}\frac{(\overline{x})^2 - x_i\overline{x}}{\overline{x^2} - (\overline{x})^2}.$$

The density estimate $\hat{f}_{\text{CUE}}(x)$ can be written as

$$\hat{f}_{\text{CUE}}(x) = \sum_{i=1}^n \hat{\pi}_i\frac{1}{h}\phi(\frac{x-x_i}{h}) = \hat{f}(x) + \frac{1}{n}\sum_{i=1}^n \frac{(\overline{x})^2 - x_i\overline{x}}{\overline{x^2} - (\overline{x})^2}\frac{1}{h}\phi(\frac{x-x_i}{h}),$$

where the first term is the RPKDE. Let S denote the second term. The expectation of $\hat{f}(x)$ can be obtained analytically. Specifically, note that

$$\frac{1}{h}\phi(\frac{x-u}{h})\phi(u) = \underbrace{\frac{1}{\sqrt{1+h^2}\sqrt{2\pi}}e^{-\frac{1}{2}\frac{x^2}{1+h^2}}}_{\equiv\alpha} \times \underbrace{\frac{1}{(h/\sqrt{1+h^2})\sqrt{2\pi}}e^{-\frac{1}{2}\frac{\left(u-\frac{x}{1+h^2}\right)^2}{h^2/(1+h^2)}}}_{\equiv\xi_{x,h}(u)}.$$

Hence

$$\mathbb{E}\left\{\hat{f}(x)\right\} = \int\frac{1}{h}\phi(\frac{x-u}{h})\phi(u)du = \frac{1}{\sqrt{1+h^2}\sqrt{2\pi}}e^{-\frac{1}{2}\frac{x^2}{1+h^2}}.$$

Now,

$$
\begin{aligned}
\mathbb{E}\left\{S\right\} &= \int \cdots \int S \prod_{j=1}^{n} \left(\phi(x_j) dx_j\right) \\
&= \int \cdots \int \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\left(\overline{x}\right)^2 - x_i \overline{x}}{\overline{x^2} - \left(\overline{x}\right)^2} \frac{1}{h} \phi(\frac{x - x_i}{h}) \right) \prod_{j=1}^{n} \left(\phi(x_j) dx_j\right) \\
&= \int \cdots \int \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\left(\overline{x}\right)^2 - x_i \overline{x}}{\overline{x^2} - \left(\overline{x}\right)^2} \frac{1}{h} \phi(\frac{x - x_i}{h}) \phi(x_i) \prod_{\substack{j=1 \\ j \neq i}}^{n} \phi(x_j) \right) \prod_{j=1}^{n} dx_j \\
&= \alpha \cdot \int \cdots \int \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\left(\overline{x}\right)^2 - x_i \overline{x}}{\overline{x^2} - \left(\overline{x}\right)^2} \xi_{x,h}(x_i) \prod_{\substack{j=1 \\ j \neq i}}^{n} \phi(x_j) \right) \prod_{j=1}^{n} dx_j \\
&= \alpha \cdot \frac{1}{n} \sum_{i=1}^{n} \int \cdots \int \frac{\left(\overline{x}\right)^2 - x_i \overline{x}}{\overline{x^2} - \left(\overline{x}\right)^2} \xi_{x,h}(x_i) \prod_{\substack{j=1 \\ j \neq i}}^{n} \phi(x_j) \prod_{j=1}^{n} dx_j \\
&= \alpha \cdot \int \cdots \int \frac{\left(\overline{x}\right)^2 - x_1 \overline{x}}{\overline{x^2} - \left(\overline{x}\right)^2} \xi_{x,h}(x_1) \prod_{j=2}^{n} \phi(x_j) dx_1 \prod_{j=2}^{n} dx_j \\
&= \alpha \cdot \int_{x_1} \underbrace{\left[\int \cdots \int_{x_2, \dots, x_n} \frac{\left(\overline{x}\right)^2 - x_1 \overline{x}}{\overline{x^2} - \left(\overline{x}\right)^2} \prod_{j=2}^{n} \phi(x_j) \prod_{j=2}^{n} dx_j \right]}_{\equiv I(x_1)} \xi_{x,h}(x_1) dx_1 .
\end{aligned}
$$

However, the ratio in $I(x_1)$ prevents the integral from being computed exactly (to the best of our knowledge). Approximating the denominator as $\left[\overline{x^2} - \left(\overline{x}\right)^2\right]^{-1} = \left[1 + o_p\left(n^{-1/2}\right)\right]^{-1} = 1 + o_p\left(n^{-1/2}\right)$ and writing the numerator as

$$
\begin{aligned}
\left(\overline{x}\right)^2 - x_1 \overline{x} &= \left(\frac{1}{n} \sum_{j=2}^{n} x_j + \frac{1}{n} x_1 \right)^2 - x_1 \frac{1}{n} \sum_{j=2}^{n} x_j - \frac{1}{n} x_1^2 \\
&= \overline{x}_{[-1]}^2 + 2 \frac{1}{n} x_1 \overline{x}_{[-1]} + \frac{1}{n^2} x_1^2 - x_1 \overline{x}_{[-1]} - \frac{1}{n} x_1^2,
\end{aligned}
$$

where $\overline{x}_{[-1]} \equiv \frac{1}{n} \sum_{j=2}^{n} x_j$, gives

$$
\begin{aligned}
I(x_1) &= \left(1 + o_p\left(n^{-1/2}\right)\right) \times \cdots \\
&\cdots \times \int \cdots \int_{x_2, \dots, x_n} \left[\overline{x}_{[-1]}^2 + 2 \frac{1}{n} x_1 \overline{x}_{[-1]} + \frac{1}{n^2} x_1^2 - x_1 \overline{x}_{[-1]} - \frac{1}{n} x_1^2 \right] \prod_{j=2}^{n} \phi(x_j) \prod_{j=2}^{n} dx_j \\
&= \left(1 + o_p\left(n^{-1/2}\right)\right) \left[-\frac{1}{n} x_1^2 + \frac{1}{n^2} x_1^2 + \left(2 \frac{1}{n} x_1 - x_1\right) J_1 + J_2 \right],
\end{aligned}
$$

where

$$
J_1 = \int \cdots \int_{x_2, \dots, x_n} \overline{x}_{[-1]} \prod_{j=2}^{n} \phi(x_j) \prod_{j=2}^{n} dx_j = 0
$$

and

$$\begin{aligned}
J_2 &= \int \cdots \int_{x_2,\ldots,x_n} \overline{x}^2_{[-1]} \prod_{j=2}^{n} \phi(x_j) \prod_{j=2}^{n} dx_j \\
&= \frac{1}{n^2} \int \cdots \int_{x_2,\ldots,x_n} \left(\sum_{j=2}^{n} x_j \right)^2 \prod_{j=2}^{n} \phi(x_j) \prod_{j=2}^{n} dx_j \\
&= \frac{1}{n^2} \int \cdots \int_{x_2,\ldots,x_n} \sum_{j=2}^{n} x_j^2 \prod_{j=2}^{n} \phi(x_j) \prod_{j=2}^{n} dx_j \quad \text{(+ zeros for all cross-products)} \\
&= \frac{1}{n^2}(n-1) \int_{x_j} x_j^2 \phi(x_j) dx_j = \frac{n-1}{n^2} = \frac{1}{n} - \frac{1}{n^2}.
\end{aligned}$$

Therefore

$$I(x_1) = \left(1 + \mathcal{O}_p\left(n^{-1/2}\right)\right) \left[-\frac{1}{n} x_1^2 + \frac{1}{n^2} x_1^2 + \frac{1}{n} - \frac{1}{n^2} \right] = n^{-1} \left[-x_1^2 + 1 \right] + \mathcal{O}_p\left(n^{-3/2}\right).$$

Thus,

$$\begin{aligned}
\mathbb{E}\left\{S\right\} &= \alpha n^{-1} \cdot \int_{x_1} \left[-x_1^2 + 1 \right] \xi_{x,h}(x_1) dx_1 + \mathcal{O}\left(n^{-3/2}\right) \\
&= \alpha n^{-1} \left[1 - \frac{h^2}{1+h^2} - \frac{x^2}{(1+h^2)^2} \right] + \mathcal{O}\left(n^{-3/2}\right).
\end{aligned}$$

Expanding α and the terms in square brackets we obtain that term of order n^{-1} as $(1-x^2)\phi(x)$, i.e. the same as given in equation (3.15a).

Appendix 3.C GELKDE with dependent data

This appendix presents simulation evidence that suggests GELKDE should have similar properties when applied to dependent data. The example considered here is that of the invertible first order moving average, MA(1), process

$$x_t = \varepsilon_t + \theta \varepsilon_{t-1}, \qquad \varepsilon_t \overset{\text{iid}}{\sim} N\left(0,1\right), \qquad |\theta| < 1, \qquad t = 1, \ldots, n.$$

The extra information used by GELKDE is that $\mathbb{E}\left\{X_t X_{t-1}\right\} = \theta$.

Results are presented for $\theta = -0.95$, -0.50, -0.25, 0.25, 0.50 and 0.95, and sample size n ranging from 25 to $1,000$. $100,000$ replications are performed in each case. As in section 3.5, red lines correspond to CUE, blue—ET and green—EL. The difference in ISB is scaled up by sample size n rather than by nh^{-2} as before.

In this example, the use of extra information gives a reduction in MISE of GELKDE, but its magnitude depends on the strength of dependence.

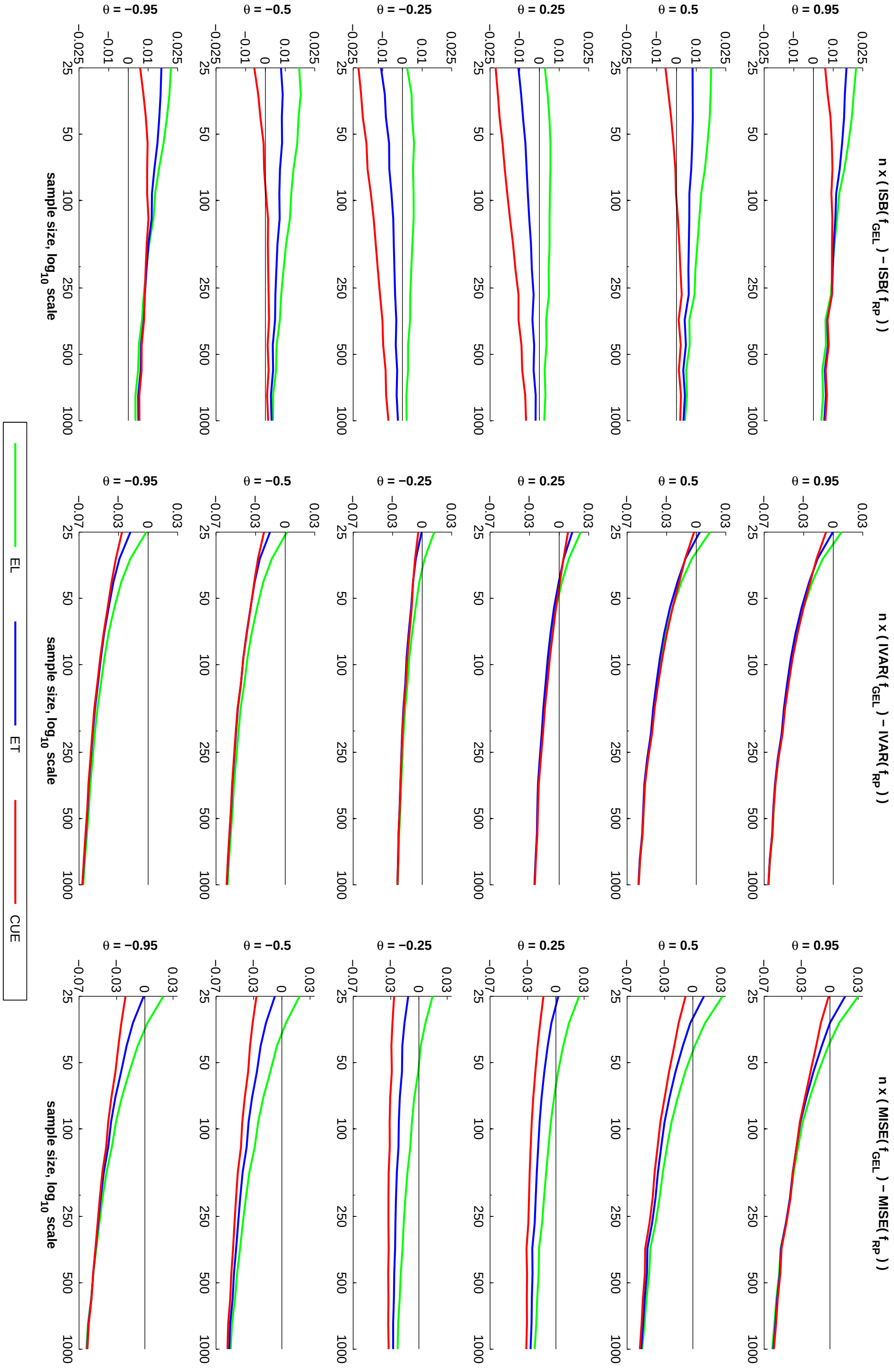

Figure 3.8: Performance of GELKDE with dependent data.

Bibliography

Acemoglu, D. and J.-S. Pischke (1997): "Why do firms train? Theory and evidence," Working paper, MIT, (Available at `http://econ-www.mit.edu/faculty/download_pdf.php?id=620`).

——— (1998): "Why Do Firms Train? Theory and Evidence," *The Quarterly Journal of Economics*, 113, 79–119.

Alessie, R., M. P. Devereux, and G. Weber (1997): "Intertemporal Consumption, Durables and Liquidity Constraints: A Cohort Analysis," *European Economic Review*, 41, 37–59.

Altman, N. and C. Léger (1995): "Bandwidth selection for kernel distribution function estimation," *Journal of Statistical Planning and Inference*, 46, 195–214.

Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold (2006): "Volatility and correlation forecasting," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. Granger, and A. Timmermann, Amsterdam: North Holland, chap. 15, 777–878.

Andrews, D. W. K. and M. Buchinsky (2000): "A Three-Step Method for Choosing the Number of Bootstrap Repetitions," *Econometrica*, 68, 23–51.

Angrist, J. D. (1988): "Grouped Data Estimation and Testing in Simple Labor Supply Models," Working Paper 234, Industrial Relations Section, Princeton University.

——— (1991): "Grouped-data Estimation and Testing in Simple Labor-Supply Models," *Journal of Econometrics*, 47, 243–266.

Angrist, J. D., G. W. Imbens, and A. Krueger (1995): "Jackknife Instrumental Variables Estimation," NBER Technical Working Papers 0172, National Bureau of Economic Research, Inc, (Available at `http://ideas.repec.org/p/nbr/nberte/0172.html`).

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67.

Antoine, B., H. Bonnal, and E. Renault (2007): "On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood," *Journal of Econometrics*, 138, 461–487.

Aragon-Sanchez, A., I. Barba-Aragon, and R. Sanz-Valle (2003): "Effects of training on business results," *International Journal of Human Resource Management*, 14, 956–980.

Attanasio, O. P. (1993): "An Analysis of Life-Cycle Accumulation of Financial Assets," *Ricerche Economiche*, 47, 323–354.

Azzalini, A. (1981): "A note on the estimation of a distribution function and quantiles by a kernel method," *Biometrika*, 68, 326–328.

Bangert, D. and J. Poor (1993): "Foreign involvement in the Hungarian economy: its impact on human resource management," *The International Journal of Human Resource Management*, 4, 817–840.

Banks, D., L. House, F. R. McMorris, P. Arabie, and W. Gaul, eds. (2004): *Classification, Clustering, and Data Mining Applications*, Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Springer-Verlag, Berlin.

Banks, J., R. Blundell, and I. Preston (1994): "Life-Cycle Expenditure Allocations and the Consumption Costs of Children," *European Economic Review*, 38, 1391–1410.

Banks, J., R. Blundell, and S. Tanner (1995): "Consumption Growth, Saving and Retirement in the UK," *Ricerche Economiche*, 49, 255–275.

Barba Navaretti, G. and A. J. Venables (2004): *Multinational Firms in the World Economy*, Princeton University Press.

Bartlett, M. S. (1949): "Fitting a Straight Line When Both Variables are Subject to Error," *Biometrics*, 5, 207–212.

Beine, M., F. Bismans, F. Docquier, and S. Laurent (2001): "Life-Cycle Behaviour of US Households: A Nonlinear GMM Estimation on Pseudopanel Data," *Journal of Policy Modeling*, 23, 713–729.

BERKOWITZ, J. (2001): "Testing Density Forecasts, with Applications to Risk Management," *Journal of Business & Economic Statistics*, 19, 465–474.

BLOMQUIST, S. AND M. DAHLBERG (1999): "Small Sample Properties of LIML and Jackknife IV Estimators: Experiments with Weak Instruments," *Journal of Applied Econometrics*, 14, 69–88.

BLOMSTRÖM, M. AND A. KOKKO (1998): "Multinational Corporations and spillovers," *Journal of Economic Surveys*, 12.

——— (2003): "Human capital and inward FDI," Working Paper 167, Stockholm School of Economics.

BLUNDELL, R., M. BROWNING, AND C. MEGHIR (1994): "Consumer Demand and the Life-Cycle Allocation of Household Expenditures," *The Review of Economic Studies*, 61, 57–80.

BLUNDELL, R., A. DUNCAN, AND C. MEGHIR (1998): "Estimating Labor Supply Responses Using Tax Reforms," *Econometrica*, 66, 827–861.

BLUNDELL, R., C. MEGHIR, AND P. NEVES (1993): "Labour Supply and Intertemporal Substitution," *Journal of Econometrics*, 59, 137–160.

BÖRSCH-SUPAN, A., A. REIL-HELD, R. RODEPETER, R. SCHNABEL, AND J. WINTER (2001): "The German Savings Puzzle," *Research in Economics*, 55, 15–38.

BOWMAN, A., P. HALL, AND T. PRVAN (1998): "Bandwidth Selection for the Smoothing of Distribution Functions," *Biometrika*, 85, 799–808.

BREIMAN, L., J. H. FRIEDMAN, R. A. OLSHEN, AND C. J. STONE (1984): *Classification and Regression Trees*, (The Wadsworth Statistics/Probability Series), Wadsworth International Group, Belmont, California.

BROWNING, M., A. DEATON, AND M. IRISH (1985): "A Profitable Approach to Labor Supply and Commodity Demands over the Life-Cycle," *Econometrica*, 53, 503–544.

CAMERON, A. C. AND P. K. TRIVEDI (2005): *Microeconometrics: Methods and Applications*, Cambridge University Press.

CAMPBELL, N. AND N. VOUSDEN (2003): "Training and technology transfer," *Australian Economic Papers*, 35–49.

Card, D. and T. Lemieux (1996): "Wage Dispertion, Returns to Skill, and Black-White Wage Differentials," *Journal of Econometrics*, 74, 319–361.

Carraro, C., F. Peracchi, and G. Weber (1993): "The Econometrics of Panels and Pseudo Panels," *Journal of Econometrics*, 59, 1–4.

Chambers, J. M. and B. Kleiner (1982): "Graphical Techniques for Multivariate Data and for Clustering," in *Handbook of Statistics*, ed. by P. R. Krishnaiah and L. N. Kanal, North-Holland Publishing Company, vol. 2, chap. 10, 209–244.

Chay, K. Y. and D. S. Lee (2000): "Changes in Relative Wages in the 1980s: Returns to Observed and Unobserved Skills and Black-White Wage Differentials," *Journal of Econometrics*, 99, 1–38.

Chen, J. and J. Qin (1993): "Empirical Likelihood Estimation for Finite Populations and the Effective Usage of Auxiliary Information," *Biometrika*, 80, 107–116.

Chen, J., A. M. Variyath, and B. Abraham (2008): "Adjusted Empirical Likelihood and its Properties," *Journal of Computational and Graphical Statistics*, 17, 426–443.

Chen, S. X. (1997): "Empirical Likelihood-based Kernel Density Estimation," *Australian and New Zealand Journal of Statistics*, 39, 47–56.

Chen, S. X. and H. Cui (2007): "On the second-order properties of empirical likelihood with moment restrictions," *Journal of Econometrics*, 141, 492–516.

Chesher, A. and R. J. Smith (1997): "Likelihood Ratio Specification Tests," *Econometrica*, 65, 627–646.

Cottrell, M., J. C. Fort, and G. Pagès (1998): "Theoretical Aspects of the SOM Algorith," *Neurocomputing*, 21, 119–138.

Cottrell, M. and P. Gaubert (2007): "Efficient Estimators: the Use of Neural Networks to Construct Pseudo Panels," mimeo, HAL, CCSD, (Available at `http://ideas.repec.org/p/hal/papers/hal-00122817_v1.html`).

Cressie, N. and T. R. C. Read (1984): "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistical Society. Series B (Methodological)*, 46, 440–464.

Dargay, J. M. (2001): "The Effect of Income on Car Ownership: Evidence of Asymmetry," *Transportation Research Part A*, 35, 807–821.

——— (2002): "Determinants of Car Ownership in Rural and Urban Areas: A Pseudo-Panel Analysis," *Transportation Research Part E*, 38, 351–366.

Davies, D. L. and D. W. Bouldin (1979): "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 224–227.

Dawid, A. P. (1984): "Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society. Series A (General)*, 147, 278–292.

De Gooijer, J. G. and D. Zerom (2003): "On Conditional Density Estimation," *Statistica Neerlandica*, 57, 159–176.

De Rossi, G. and A. C. Harvey (2006): "Time-varying quantiles," CWPE 0649, University of Cambridge.

——— (2009): "Quantiles, expectiles and splines," *Journal of Econometrics*, 152, 179–185.

Deaton, A. (1985): "Panel Data from Time Series of Cross-Sections," *Journal of Econometrics*, 30, 109–126.

Deboeck, G. J. and T. Kohonen, eds. (1998): *Visual Explorations in Finance with Self-Organizing Maps*, Springer-Verlag, London.

Devereux, P. J. (2003): "Improved Errors-in-Variables Estimators for Grouped Data," Working paper, Department of Economics, UCLA, (Available at `http://www.econ.ucla.edu/devereux/ueve.pdf`).

——— (2007): "Improved Errors-in-Variables Estimators for Grouped Data," *Journal of Business and Economic Statistics*, 25, 278–287.

DiCiccio, T., P. Hall, and J. Romano (1988): "Bartlett Adjustment for Empirical Likelihood," Technical Report 298, Department of Statistics, Stanford University.

——— (1991): "Empirical Likelihood is Bartlett-Correctable," *Annals of Statistics*, 19, 1053–1061.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998): "Evaluating Density Forecasts, with Applications to Financial Risk Management," *International Economic Review*, 39, 863–883.

DiNardo, J. (1993): "Law Enforcement, the Price of Cocaine and Cocaine Use," *Mathematical and Computer Modelling*, 17, 53–64.

DORE, R. (2001): “Making Sence of Globalisation,” CEP Occasional Papers CEPOP16, Centre for Economic Performance, London School of Economics.

EGGERMONT, P. P. B. AND V. N. LARICCIA (2001): *Maximum Penalized Likelihood Estimation. Volume I: Density Estimation*, Springer Series in Statistics, Springer.

ENGLE, R. F. AND S. MANGANELLI (2004): “CAViaR: conditional autoregressive value at risk by regression quantiles,” *Journal of Business and Economic Statistic*, 22, 367–381.

EPANECHNIKOV, V. A. (1969): “Non-parametric estimation of a multivariate probability density,” *Theory of Probability and its Applications*, 14, 153–158.

EVERITT, B. S., S. LANDAU, AND M. LEESE (2001): *Cluster Analysis*, Arnold, Hodder Headline Group, London, 4 ed.

FOSFURI, A., M. MOTTA, AND T. RONDE (2001): “Foreign direct investment and spillovers through workers’ mobility,” *Journal of International Economics*, 53, 205–222.

FRAZIS, H., M. GITTLEMAN, AND M. JOYCE (2000): “Correlates of training: an analysis using both employer and employee characteristics,” *Industrial and Labor Relations Review*, 53, 443–462.

FRYER, M. J. (1976): “Some Errors Associated with the Non-parametric Estimation of Density Functions,” *IMA Journal of Applied Mathematics*, 18, 371–380.

GARDES, F., S. LANGLOIS, AND D. RICHAUDEAU (1996): “Cross-Section Versus Time-Series Income Elasticities of Canadian Consumption,” *Economics Letters*, 51, 169–175.

GASSNER, K. (1998): “An Estimation of UK Telephone Access Demand Using Pseudo-Panel Data,” *Utilities Policy*, 7, 143–154.

GERSBACH, H. AND A. SCHMUTZLER (2003): “Endogenous technological spillovers: causes and consequences,” *Journal of Economics & Management Strategy*, 12, 179–205.

GODAMBE, V. P. AND M. E. THOMPSON (1989): “An extension of quasi-likelihood estimation,” *Journal of Statistical Planning and Inference*, 22, 137–152.

GORDON, A. D. (1981): *Classification*, (Monographs on applied probability and statistics), Chapman and Hall.

Görg, H. and D. Greenaway (2001): "Foreign Direct Investment and Intra-industry Spillovers: A Review of the Literature," Globalisation and Labour Markets Programme Research Paper 2001/37, Leverhulme Centre for Research on Globalisation and Economic Policy, Nottingham, UK.

Gourieroux, C. and J. Jasiak (2008): "Dynamic quantile models," *Journal of Econometrics*, 147, 198–205.

Gower, J. C. (1971): "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, 27, 857–871.

Groeneveld, R. A. and G. Meeden (1984): "Measuring Skewness and Kurtosis," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 33, 391–399.

Guggenberger, P. (2008): "Finite Sample Evidence Suggesting a Heavy Tail Problem of the Generalized Empirical Likelihood Estimator," *Econometric Reviews*, 27, 526–541.

Hall, P. and P. Patil (1994): "On the efficiency of on-line density estimators," *IEEE Transactions on Information Theory*, 40, 1504–1512.

Hall, P. and B. Presnell (1999): "Density Estimation under Constraints," *Journal of Computational and Graphical Statistics*, 8, 259–277.

Harris, R. I. D. (1999): "The determinants of work-related training in Britain in 1995 and the implications of employer size," *Applied Economics*, 451–463.

Hartigan, J. A. (1975): *Clustering Algorithms*, (Wiley series in probability and mathematical statistics), John Wiley & Sons.

Harvey, A. C. (1989): *Forecasting, Structural Time Series Models and Kalman Filter*, Cambridge University Press.

——— (2006): "Forecasting with unobserved components time series models," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. Granger, and A. Timmermann, Amsterdam: North Holland, chap. 7, 327–412.

Harvey, A. C. and G. De Rossi (2006): "Signal Extraction," in *Palgrave handbook of econometrics: Volume 1 Econometric Theory*, ed. by T. C. Mills and K. Patterson, Basingstoke: Palgrave Macmillan.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2001): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer series in statistics, Springer.

HSIAO, C. (2003): *Analysis of Panel Data*, Cambridge University Press, 2nd ed.

HU, A. G. (2004): "Multinational corporations, patenting, and knowledge flow: the case of Singapore," *Economic Development and Cultural Change*, 781–800.

IMBENS, G. W. (2002): "Generalized Method of Moments and Empirical Likelihood," *Journal of Business & Economic Statistics*, 20, 493–506.

JONES, M. C., J. S. MARRON, AND S. J. SHEATHER (1996): "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, 91, 401–407.

KAUFMAN, L. AND P. J. ROUSSEEUW (2005): *Finding Groups in Data: An Introduction to Cluster Analysis*, (Wiley series in probability and statistics), John Wiley & Sons.

KIM, M. AND R. RAMAKRISHNA (2005): "New indices for cluster validity assessment," *Pattern Recognition Letters*, 26, 2353–2363.

KITAMURA, Y. (2006): "Empirical Likelihood Methods in Econometrics: Theory and Practice," Cowles Foundation Discussion Paper 1569, Cowles Foundation, Yale University.

KITAMURA, Y. AND M. STUTZER (1997): "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861–874.

KOHONEN, T. (1997): *Self-Organizing Maps*, (Springer series in information sciences; 30), Springer-Verlag Berlin, 2nd ed.

KOOPMAN, S. J. (1993): "Disturbance Smoother for State Space Models," *Biometrika*, 80, 117–126.

KOOPMAN, S. J. AND A. C. HARVEY (2003): "Computing observation weights for signal extraction and filtering," *Journal of Economic Dynamics and Control*, 27, 1317–1333.

KUESTER, K., S. MITTNIK, AND M. S. PAOLELLA (2006): "Value-at-Risk Prediction: A Comparison of Alternative Strategies," *Journal of Financial Econometrics*, 4, 53–89.

LAURENT, S. (2007): *GARCH 5*, Timberlake Consultants Ltd., London.

LEVENSON, A. R. (1996): "Do Consumers Respond to Future Income Shocks? Evidence from Social Security Reform in Taiwan," *Journal of Public Economics*, 62, 275–295.

LI, Q. AND J. S. RACINE (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

LITTLE, R. J. A. (1992): "Regression With Missing X's: A Review," *Journal of the American Statistical Association*, 87, 1227–1237.

LIU, Y. AND J. CHEN (2010): "Adjusted empirical likelihood with high-order precision," *Annals of Statistics*, 38, 1341–1362.

MADANSKY, A. (1959): "The Fitting of Straight Lines When both Variables are Subject to Error," *Journal of the American Statistical Association*, 54, 173–205.

MALLIOS, W. S. (1969): "A Generalized Application of Instrumental Variable Estimation to Straight-Line Relations When Both Variables Are Subject to Error," *Technometrics*, 11, 255–263.

MARKOVICH, N. (2007): *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*, Wiley series in probability and statistics, John Wiley & Sons.

MARRON, J. S. AND M. P. WAND (1992): "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–736.

MCKENZIE, D. J. (2006): "Precautionary Savings and Consumption Growth in Taiwan," *China Economic Review*, 17, 84–101.

MOFFITT, R. (1993): "Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-Sections," *Journal of Econometrics*, 59, 99–123.

MORAN, T. H., E. GRAHAM, AND M. BLOMSTRÖM, eds. (2005): *Does Foreign Direct Investment Promote Development?*, Washington, DC: Institute for International Economics, Center for Global Development.

MORRISON PAUL, C. J. AND R. NEHRING (2005): "Product Diversification, Production Systems, and Economic Performance in U.S. Agricultural Production," *Journal of Econometrics*, 126, 525–548.

NADARAYA, E. A. (1964): "Some New Estimates for Distribution Functions," *Theory of Probability and its Applications*, 9, 497–500.

Newey, W. K. and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255.

Neyman, J. and E. L. Scott (1951): "On Certain Methods of Estimating the Linear Structural Relation," *The Annals of Mathematical Statistics*, 22, 352–361.

Nordman, D. J. and S. N. Lahiri (2006): "A frequency domain empirical likelihood for short– and long–range dependence," *The Annals of Statistics*, 34, 3019–3050.

Otsu, T. (2007): "Penalized empirical likelihood estimation of semiparametric models," *Journal of Multivariate Analysis*, 98, 1923–1954.

Pagan, A. and A. Ullah (1999): *Nonparametric Econometrics*, Themes in Modern Econometrics, Cambridge University Press.

Pakes, A. (1982): "On the Asymptotic Bias of Wald-Type Estimators of a Straight Line when Both Variables are Subject to Error," *International Economic Review*, 23, 491–497.

Parker, S. C. and J. Coleman (1999): "Training in the UK: Does National Ownership Matter?" *International Journal of Training and Development*, 3, 278–291.

Parzen, E. (1962): "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, 33, 1065–1076.

Peracchi, F. and F. Welch (1995): "How Representative Are Matched Cross-Sections? Evidence from the Current Population Survey," *Journal of Econometrics*, 68, 153–179.

Phillips, G. D. A. and C. Hale (1977): "The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems," *International Economic Review*, 18, 219–228.

Pinson, P., H. A. Nielsen, J. K. Møller, H. Madsen, and G. N. Kariniotakis (2007): "Non-parametric probabilistic forecasts of wind power: required properties and evaluation," *Wind Energy*, 10, 497–516.

Prais, S. J. and J. Aitchison (1954): "The Grouping of Observations in Regression Analysis," *Revue de l'Institut International de Statistique*, 22, 1–22.

Propper, C., H. Rees, and K. Green (2001): "The Demand for Private Medical Insurance in the UK: A Cohort Analysis," *The Economic Journal*, 111, C180–C200.

Qin, J. and J. Lawless (1994): "Empirical likelihood and general estimating equations," *The Annals of Statistics*, 22, 300–325.

Ramalho, J. J. S. and R. J. Smith (2005): "Goodness of Fit Tests for Moment Condition Models," Working Paper 2005/05, Universidade de Évora.

Reiersøl, O. (1950): "Identifiability of a Linear Relation between Variables Which Are Subject to Error," *Econometrica*, 18, 375–389.

Ridder, G. and R. Moffitt (2007): "The Econometrics of Data Combination," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, North Holland, vol. 6B, chap. 75, 5469–5547.

Ripley, B. D. (1996): *Pattern Recognition and Neural Networks*, Cambridge University Press.

Robertson, R. (2003): "Exchange Rates and Relative Wages: Evidence from Mexico," *North American Journal of Economics and Finance*, 14, 25–48.

Rosenblatt, M. (1956): "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, 27, 832–837.

Roulston, M. S. and L. A. Smith (2002): "Evaluating Probabilistic Forecasts Using Information Theory," *Monthly Weather Review*, 130, 16531660.

Runkler, T. A. (2007): "Relational Fuzzy Clustering," in *Advances in Fuzzy Clustering and Its Applications*, ed. by J. Valente de Oliveira and W. Pedrycz, John Wiley & Sons, chap. 2.

Shahidi, A. R. (2009): "Model selection for moment condition models using the penalized empirical likelihood procedure," in *Three essays on model selection, modulation estimators and herd behavior under asymmetric beliefs*, PhD dissertation, University of Pittsburgh, chap. 1.

Sheather, S. J. (2004): "Density Estimation," *Statistical Science*, 19, 588–597.

Sheather, S. J. and J. S. Marron (1990): "Kernel Quantile Estimators," *Journal of the American Statistical Association*, 85, 410–416.

Silverman, B. W. (1986): *Density Estimation*, Chapman and Hall.

SINGH, J. (2004): “Multinational firms and knowledge diffusion: evidence using patent citation data,” in *Academy of Management Proceedings: Academy of Management Best Conference Paper*, bPS: D1.

SMITH, A. AND G. HAYTON (1999): “What drives enterprise training? Evidence from Australia,” *The International Journal of Human Resource Management*, 10, 251–272.

SMITH, R. J. (1997): “Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation,” *The Economic Journal*, 107, 503–519.

——— (2010): “GEL criteria for moment condition models,” *Econometric Theory*, forthcoming.

SUTHERLAND, J. (2004): “The determinants of training,” *Economic Issues*, 9, 23–39.

TEITEL, S. (2005): “Globalization and Its Disconnects,” *The Journal of Socio-Economics*, 444–470.

TOURANGEAU, R., L. J. RIPS, AND K. RASINSKI (2000): *The Psychology of Survey Response*, Cambridge University Press.

TSYBAKOV, A. B. (2009): *Introduction to Nonparametric Estimation*, (Springer Series in Statistics), Springer.

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

VERBEEK, M. (2006): “Pseudo Panels and Repeated Cross-Sections,” in *The Econometrics of Panel Data*, ed. by L. Mátyás and P. Sevestre, Kluwer Academic Publishers, chap. 10, 3rd ed ed., (Forthcoming). Available at SSRN: http://ssrn.com/abstract=869445.

VERBEEK, M. AND T. NIJMAN (1992): “Can Cohort Data Be Treated as Genuine Panel Data?” *Empirical Economics*, 17, 9–23.

——— (1993): “Minimum MSE Estimation of a Regression Model with Fixed Effects from a Series of Cross-Sections,” *Journal of Econometrics*, 59, 125–136.

VERBEEK, M. AND F. VELLA (2005): “Estimating Dynamic Models from Repeated Cross-Sections,” *Journal of Econometrics*, 127, 83–102.

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas (2000): "SOM Toolbox for Matlab 5," Report A57, Helsinki University of Technology, (Available at `http://www.cis.hut.fi/projects/somtoolbox/`).

Wald, A. (1940): "The Fitting of Straight Lines if Both Variables are Subject to Error," *The Annals of Mathematical Statistics*, 11, 284–300.

Walsh, J. (2001): "Human resource management in foreign-owned workplaces: evidence from Australia," *International Journal of Human Resource Management*, 12, 425–444.

Wand, M. P. and M. C. Jones (1995): *Kernel Smoothing*, vol. 60 of *Monographs on Statistics and Applied Probability*, Chapman & Hall.

Wasserman, L. (2006): *All of Nonparametric Statistics*, Springer Texts in Statistics, Springer.

Weber, R. (2007): "Fuzzy Clustering in Dynamic Data Mining—Techniques and Applications," in *Advances in Fuzzy Clustering and Its Applications*, ed. by J. Valente de Oliveira and W. Pedrycz, John Wiley & Sons, chap. 15.

Wedderburn, R. W. M. (1974): "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, 61, 439–447.

Wegman, E. J. and H. I. Davies (1979): "Remarks on Some Recursive Estimators of a Probability Density," *The Annals of Statistics*, 7, 316–327.

Yadapadithaya, P. S. (2001): "Evaluating Corporate Training and Development: An Indian Experience," *International Journal of Training and Development*, 5, 261–274.

Yu, K. and M. C. Jones (1998): "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228–237.