Essays in Panel Data and Financial Econometrics

Cavit Pakel

University College

Submitted for the

Doctor of Philosophy Degree in Economics

Department of Economics

University of Oxford

Trinity Term

2012

Essays in Panel Data and Financial Econometrics

Cavit Pakel

University College

Submitted for the Doctor of Philosophy Degree at the Department of Economics

University of Oxford

Trinity 2012

Abstract

This thesis is concerned with volatility estimation using financial panels and bias-reduction in non-linear dynamic panels in the presence of dependence.

Traditional GARCH-type volatility models require large time-series for accurate estimation. This makes it impossible to analyse some interesting datasets which do not have a large enough history of observations. This study contributes to the literature by introducing the GARCH Panel model, which exploits both time-series and cross-section information, in order to make up for this lack of time-series variation. It is shown that this approach leads to gains both in- and out-of-sample, but suffers from the well-known incidental parameter issue and therefore, cannot deal with short data either. As a response, a bias-correction approach valid for a general variety of models beyond GARCH is proposed. This extends the analytical bias-reduction literature to cross-section dependence and is a theoretical contribution to the panel data literature. In the final chapter, these two contributions are combined in order to develop a new approach to volatility estimation in short panels. Simulation analysis reveals that this approach is capable of removing a substantial portion of the bias even when only 150-200 observations are available. This is in stark contrast with the standard methods which require 1,000-1,500 observations for accurate estimation. This approach is used to model monthly hedge fund volatility, which is another novel contribution, as it has hitherto been impossible to analyse hedge fund volatility, due to their typically short histories. The analysis reveals that hedge funds exhibit variation in their volatility characteristics both across and within investment strategies. Moreover, the sample distributions of fund volatilities are asymmetric, have large right tails and react to major economic events such as the recent credit crunch episode.

Acknowledgements

First and foremost, I would like to thank Neil Shephard for his guidance and support, especially at times when it was most needed. His clear and critical evaluation and judgement of my work has been invaluable to me in understanding the makings of a good researcher. It is my sincere hope that I have learned well by his example.

I am also grateful to several people in Oxford and elsewhere for useful discussions, which contributed greatly to this study. I am particularly thankful to Manuel Arellano, Debopam Bhattacharya, Steve Bond, Stephané Bonhomme, Bent Nielsen, Shin Kanaya, Arnaud Lionnet (my first and best mathematics teacher!), Kasper Lund-Jensen, Sophocles Mavroeidis, Andrew Patton, Anders Rahbek, Tarun Ramadorai, Olivier Scaillet, Enrique Sentana, Kevin Sheppard, Marianne Simonsen and Michael Streatfield.

In February-March 2011, I was a visiting student at CEMFI and their unbounded hospitality is greatly acknowledged. I would like to note here that the comments I received from Manuel and Stephané during my visit had a major impact on the direction of my thesis and my understanding of panel data models.

I must also thank my good friends Alex, Arnaud, Bahman, Diaa, Irem, Joel, Kadu, Kasper, Khatchig, Salvatore, Sriram and Sylvestre for their friendship that made life a lot more pleasant, especially when research was not going as it was supposed to be.

Financial support from the Department of Economics and the Oxford-Man Institute of Quantitative Finance are greatly acknowledged.

All virtues of this study reflect the contributions of the aforementioned people. The responsibility for all shortcomings and errors is entirely mine.

Finally, my most heartfelt thanks go to my mother Nilgün Çehreli, my late grandmother Sabahat Çehreli and Fitnat Banu Demir. My mother and grandmother have managed to weather all my ups and downs through life most heroically and efficiently. Banu has willingly accepted to be the next one to weather future storms and has since been an excellent chef, a fierce critic, but most importantly, a patient friend and my dear companion. I am grateful to have been blessed with their presence in my life. This thesis is dedicated to them.

Cavit Pakel
Oxford
May 2012

Word Count

The body of this thesis consists of 193 pages with a typical page containing 354 words. Therefore, excluding the bibliography, the thesis is 68,332 words long.

This thesis was typed using Scientific Word, which was also used to create all tables. Matlab was used to conduct all simulations and estimations and to generate the related figures.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is concerned with three inter-connected objectives. The first objective is to introduce a novel panel data based volatility estimation approach using the well-known Generalised Autoregressive Conditional Heteroskedasticity (GARCH) type models due to Engle (1982) and Bollerslev (1986). To this end, Chapter 2 introduces the GARCH Panel model.

The second objective is a theoretical investigation of the small sample bias and its correction in non-linear and dynamic panel data models in the presence of both time-series and cross-section dependence. This is related to the first objective, as GARCH panels are essentially non-linear dynamic panels. Moreover, the GARCH Panel model is mainly concerned with modelling the volatility of macroeconomic and financial panels, which will almost certainly exhibit some degree of dependence in both dimensions. Importantly, simulation based investigation of the GARCH Panel model in Chapter 2 reveals that this approach is not well-suited for datasets that are inherently short, such as monthly hedge fund returns, as it suffers from the well-known incidental parameter bias, first observed by Neyman and Scott (1948).[1] This is the main empirical motivation behind the theoretical analysis on bias-reduction conducted in Chapter 3. Nevertheless, it must be underlined that Chapter 3 considers a general likelihood-based framework with no specific model or application in mind. Indeed, Chapter 3 can easily be read separately from the rest of the thesis. Consequently, its results are generally applicable to models and applications beyond GARCH modelling of volatility.

[1] Here and elsewhere, the term "short" used in reference for panels or datasets means that the time-series dimension is not large enough for the small sample bias to vanish.

The final objective is to propose a novel approach for volatility estimation in small samples. GARCH-type models are quite successful in modelling volatility in samples of $1,000 - 2,000$ time-series observations. However, they yield inaccurate results when there are not more than a few hundred observations. This makes it virtually impossible to model some interesting datasets which are inherently short such as inflation and hedge fund returns, due to being characteristically observed at low frequency. The main empirical motivation of Chapter 4 then is to carry GARCH type modelling to such datasets. To do so, Chapter 4 links the contributions of Chapters 2 and 3 by employing the bias-reduction mechanism proposed in Chapter 3 to estimate the GARCH Panel model in panels with around 200 time-series observations.

Consequently, this thesis contributes to both the panel data and financial econometrics literatures. The main contribution to the panel data literature is Chapter 3, which extends the analytical bias reduction literature to cross-section dependence, which has not been considered before. Chapters 2 and 4, on the other hand, contribute to the financial econometrics literature both theoretically and empirically. Below, a summary of the main ideas is discussed as an introduction to this study. The discussion is meant to be intuitive and a minimal number of references is used. The main aim is to provide a unifying overview of the key ideas that lie at the core of this study. Almost all of the points discussed will again be mentioned in the thesis as they become relevant and will be treated formally.

1.1 Overview of the Thesis

1.1.1 Chapter 2

GARCH-type models are highly popular and are included as a standard option in many commercial econometric software packages. The main idea behind this strand of models is to exploit the dependence structure in the conditional variance of data. For instance, although daily stock returns do not exhibit a distinctive pattern, squared returns offer some predictability across time. This is illustrated in Figure 1.1. For some zero-mean financial returns series r_t, $t = 1, ..., T$, the basic GARCH(1,1) model is given by

$$r_t|\mathcal{F}_{t-1} \sim F(0, \sigma_t^2), \quad \sigma_t^2 = \lambda(1 - \alpha - \beta) + \alpha r_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{1.1}$$

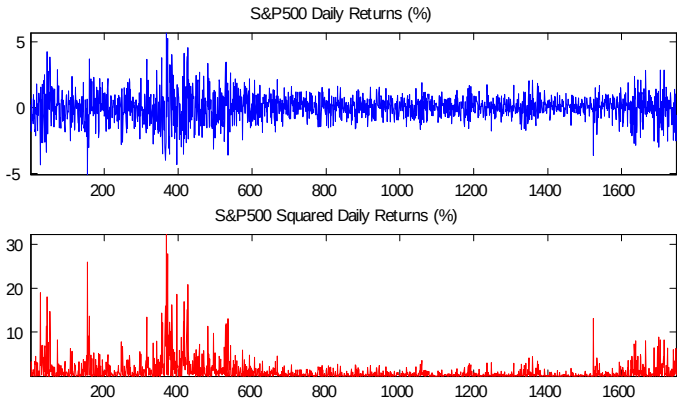

Figure 1.1: Daily returns and squared returns on the S&P500 index from 2 February 2001 to 16 January 2008.

where $\mathcal{F}_{t-1}$ is the information set at time t and F is some distribution with variance σ_t^2. Clearly, the intuition is that a large (small) shock and/or a large conditional variance yesterday will cause a large (small) conditional variance today. The parameters (α, β) measure this reaction while λ is related to the long-run variance. This simple model is highly successful in modelling conditional volatility. However, it also requires a large number of observations, possibly around $1,000 - 2,000$. This is illustrated by a simple simulation analysis. Figure 1.2 gives the sample distributions of $\hat{\alpha}$ and $\hat{\beta}$ estimated for $2,500$ replications of the GARCH process in (1.1) for varying T. This illustrates two important points. First, until T is around $1,000$, a substantial portion of $\hat{\alpha}$ are arbitrarily close to zero. This is a serious problem as $\alpha = 0$ implies that β is not identified. Second, especially in the case of $\hat{\beta}$, estimates exhibit high dispersion unless T is large enough. As far as modelling of daily data is concerned, this is not an important issue as usually long histories of data are available for many financial series. However, some interesting datasets, such as monthly hedge fund returns, are inherently scarce as the available datasets are characteristically short. Something else is needed to model such datasets.

Clearly, if the time-series variation is not sufficient to estimate the parameters, one possibility is to open up a new source of variation; the cross-sectional variation. This is supported by the empirical observation that $\hat{\alpha}$ and $\hat{\beta}$ usually cluster around similar values when the model is fit to different series. This can again be illustrated by using daily returns from S&P100. Figure 1.3 is based on the GARCH parameter estimates for 94 stocks traded in S&P100 between 20 October 2000 and 25 March 2009. The result is clear: $\hat{\alpha}$ and $\hat{\beta}$ are

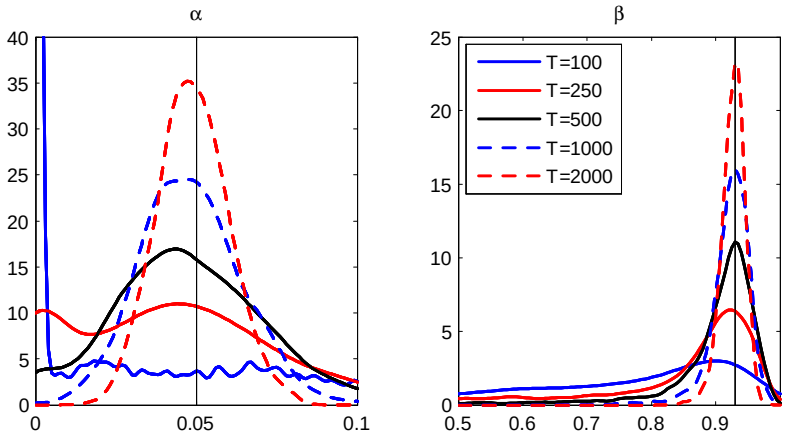

Figure 1.2: Sample distributions of $\hat{\alpha}$ (left hand panel) and $\hat{\beta}$ (right hand panel) for GARCH processes of varying length. Based on $2,500$ replications.

concentrated around a small section of the parameter space and take on similar values for all stocks (given by the triangle bordered by the axis and the red line) while the estimates of λ (presented in terms of annualised volatility) are dispersed. Motivated by such empirical observations, Chapter 2 proposes a panel data based volatility model, the *GARCH Panel model*. The model is given by

$$r_{it}|\mathcal{F}_{i,t-1} \sim F(0, \sigma_{it}^2), \quad \sigma_{it}^2 = \lambda_i(1-a-\beta) + \alpha r_{i,t-1}^2 + \beta\sigma_{i,t-1}^2,$$

where $\mathcal{F}_{i,t-1}$ is the information set for asset i at time $t-1$. This allows for heterogeneous long-run variances, but assumes homogeneity of α and β. The advantages of this approach can be summarised as follows:

- Estimation of (α, β) is based on a richer information set as it exploits both time-series and cross-sectional information.
- The availability of an extra source of information can possibly make up for the lack of time-series variation in smaller samples.

At first sight, however, there is an implicit difficulty. Estimation is conducted using the maximum likelihood method, as is standard for GARCH modelling. Given the panel structure, a joint density function has to be specified, implying that the cross-section dependence has to be modelled. As will be discussed in more detail in Chapter 2, this leads to two problems. The statistical problem is that now one has to model the parameters

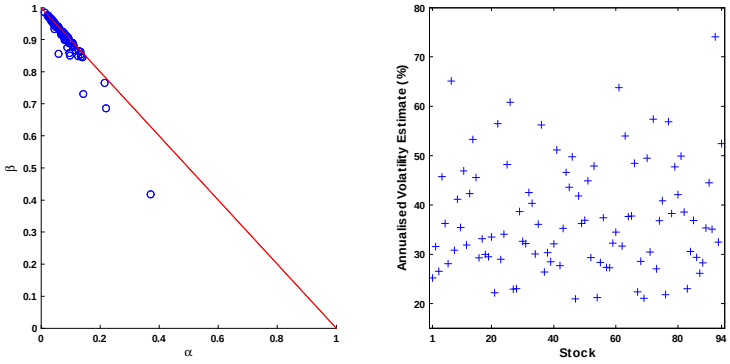

Figure 1.3: Parameter estimates for the GARCH model for returns on 94 stocks traded in S&P100. Estimation is based on daily returns from 20 October 2000 to 25 March 2009 (2,117 observations). Estimates for (α, β) are illustrated on the left panel while estimates for λ are given on the right panel, in terms of annualised volatility.

pertaining the dependence structure, in addition to $(\lambda_1, ..., \lambda_N, \alpha, \beta)$. This can possibly increase the dimension of the parameter space dramatically, leading to curse of dimensionality. Moreover, GARCH estimation is conducted by numerical optimisation methods, and a large dimensional dependence parameter will make numerical optimisation practically infeasible. These issues are side-stepped by using the composite likelihood method, which provides a justification for approximating the joint density by averaging the marginal densities. This virtually is the same as assuming cross-section independence. However, under regularity conditions, this still ensures consistency. This idea is implemented in this study by adapting the framework of Engle, Shephard and Sheppard (2008) who introduce this idea to the financial econometrics literature.

The simulation and empirical analysis results indicate that this approach leads to significant gains over the standard time-series based estimation approach, both and in- and out-of-sample. Indeed, good small sample properties are obtained even when T is as small as 500. However, at the same time, it turns out that the panel approach is not a panacea guaranteeing accurate results at any sample size, and certainly not when T is around 200. The GARCH panel is, at the end of the day, a nonlinear dynamic panel data model characterised by a common set of parameters and many individual-specific parameters. Hence, it suffers from the same problem that all models of this type do: the incidental-parameter issue (Neyman and Scott (1948)). As far as this study is concerned, intuitively this is a finite-sample time-series bias.

The statistics and, more recently, microeconometrics literatures contain influential studies on the incidental parameter issue and possible bias-reduction methods. The clear message then is that one can simply pick an adequate bias-correction method and apply it to GARCH panels. However, remember that GARCH panels are characterised by both time-series and, more importantly, cross-section dependence. Unfortunately, there is no study on correction of the incidental parameter bias in the presence of cross-section dependence. This motivates the next Chapter.

1.1.2 Chapter 3

In the panel data literature, the incidental parameter issue generally arises due to the presence of unobserved heterogeneity, which is a prominent theme in this literature. To explain, consider the following standard examples.

Example 1.1.1 *The static linear model given by*

$$y_{it} = x_{it}\theta + \varepsilon_{it}, \qquad \textit{where } \varepsilon_{it} = \lambda_i + u_{it} \textit{ and } \mathbb{E}[x_{it}\lambda_i] \neq 0.$$

Hence, $\mathbb{E}[x_{it}\varepsilon_{it}] \neq 0$.

Example 1.1.1 considers the simple static linear model with a common parameter of interest, θ. The economic intuition behind the error term specification is that economic agents possess different characteristics which are unobserved; hence unobserved heterogeneity. This is captured by the individual-specific term λ_i. However, these unobserved characteristics are almost certainly correlated with agents' economic behaviour, given by x_{it}. The implication is that the exogeneity assumption is violated and standard OLS estimation is not valid. In this example, the solution is very simple. First-differencing yields $\Delta y_{it} = \theta \Delta x_{it} + \Delta \varepsilon_{it}$, where $\Delta y_{it} = y_{it} - y_{i,t-1}$. Then, $\mathbb{E}[\Delta x_{it} \Delta \varepsilon_{it}] = 0$ since $\Delta \varepsilon_{it} = \Delta u_{it}$ and λ_i is gone.

Example 1.1.2 *The dynamic autoregressive model is given by*

$$y_{it} = \theta y_{i,t-1} + (\lambda_i + u_{it}).$$

In Example 1.1.2, first-differencing is not sufficient to solve the problem as $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$ is still correlated with the error term $\Delta u_{it} = u_{i,t} - u_{i,t-1}$, due to the dynamic

structure of the model. The dominant solution in the literature is to use valid instruments and conduct GMM estimation. For example, in the case of $\Delta y_{i3} = \theta \Delta y_{i2} + \Delta u_{i3}$, y_{i1} is a suitable instrument for Δy_{i2}.

Unfortunately, such smart and exact solutions do not exist for all possible models. In addition, the GMM based approach inherently depends on the first-differencing concept which is not straightforward for non-linear models. Most importantly, such methods are suited for panels characterised by a few time-series observations and large cross-section dimensions; the so called large-N fixed-T framework. In large-T large-N settings, which are becoming increasingly more relevant due to an increase in the availability of large panels, some popular methods are known to suffer from asymptotic bias. Therefore, recently the approach has been to consider λ_i as an individual-specific parameter instead and estimate it along with θ using likelihood methods. The problem with this approach is that, in short panels estimation suffers from the incidental parameter issue.

The problem is very intuitive. Suppose the likelihood function is given by $\ell(\theta, \lambda_1, ..., \lambda_N)$. Estimation of the individual-specific parameters relies on individual-specific information, only. As such, if a sufficiently long history of data is not available, λ_i is estimated inaccurately. This inaccuracy is accumulated across $\hat{\lambda}_i$ and contaminates the likelihood function $\ell(\theta, \hat{\lambda}_1, ..., \hat{\lambda}_N)$, which is used to estimate θ. Accordingly, $\hat{\theta}$ inherits the accumulate bias. The solution is to analytically derive the effect of the inaccuracy of $(\hat{\lambda}_1, ..., \hat{\lambda}_N)$ on θ, and correct for this bias.[2] Generally, in the standard case of iid panels, one can characterise this small sample bias in terms of increasing orders of $1/T$ by asymptotic expansions:

$$\mathbb{E}[\hat{\theta} - \theta] = \frac{A}{T} + O\left(\frac{1}{T^2}\right)$$

The focus of the literature is on the first order bias, A/T, while higher order bias terms are considered negligible. Clearly, the first order bias can successfully be removed by adjusting $\hat{\theta}$ accordingly, once a characterisation for A is obtained.

At this point, it is important to illustrate the link between the seemingly unrelated panel data and financial econometrics literatures. To see the connection, it is central to

[2] An elegant exact solution is based on using a sufficient statistic. For example, in the case of the binary logit model, the incidental parameter issue can be solved by using a sufficient statistics (Andersen (1970) and Chamberlain (1980)). Unfortunately, this statistic does not necessarily exist for all models of interest.

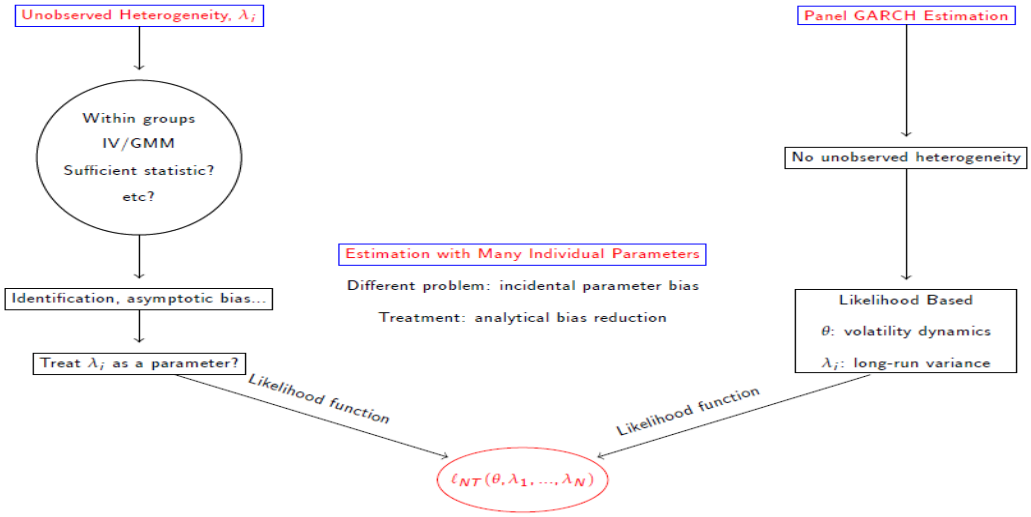

Figure 1.4: Overview of the connection between literatures.

observe that the incidental parameter issue is caused entirely by the presence of individual-specific parameters. The exact nature of why these individual specific parameters exist in the first place is of no consequence. In Chapter 2, individual specific parameters arise due to heterogeneous long-run variance parameters. In the motivating examples considered here, on the other hand, these parameters arise due to unobserved heterogeneity. Either way, the existence of such parameters leads to the statistical problem of incidental parameter bias. In addition, remember that in both cases estimation is conducted by the maximum likelihood method. Therefore, essentially both research questions boil down to maximisation of some likelihood function given by $\ell(\theta, \lambda_1, ..., \lambda_N)$. Seen from this perspective, these two different research questions from two separate literatures are actually two instances of the much more general incidental parameter issue. The treatment of this issue is analytical bias reduction. These connections are visualised in Figure 1.4.

As mentioned in the previous Section, the analytical bias reduction literature is abundant with different bias correction approaches. However, the literature is invariably based on the assumption of cross-section independence. Therefore, the main objective of Chapter 3 is extension of the literature to the case of cross-section dependence. The particular approach studied is the integrated likelihood method, which is based on

$$\frac{1}{T} \ln \int T[\exp \ell_i(\theta, \lambda_i)] \pi_i(\lambda_i|\theta) d\lambda_i,$$

where ℓ_i is the normalised log-likelihood function for individual i. The bias-reduction mechanism is based on choosing an appropriate weight function, $\pi_i(\lambda_i|\theta)$, such that the integrated likelihood function is robust to the incidental parameter bias.

The difficulty is finding a convenient way of integrating cross-section dependence into the analysis such that double-asymptotic expansions as both N and T go to ∞ can be obtained. This is achieved by adapting a flexible structure for the double-asymptotic convergence rates. To illustrate this, consider some random variable X_{it} such that

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it} = O_p\left(\frac{1}{N^{\rho}T}\right). \tag{1.2}$$

Under cross-section independence, subject to standard regularity conditions $\rho = 1$, implying the usual $\sqrt{NT}$-convergence. However, theoretically, cross-section dependence can be so strong that cross-sectional information does not contribute to convergence. Then, $\rho = 0$ and $\sqrt{T}$-convergence ensues. Hence, cross-section dependence is conceptualised in this study in the same fashion as (1.2), somewhere in between these two polar cases. Bias reduction using the integrated likelihood method has recently been analysed by Arellano and Bonhomme (2009) under time-series and cross-section independence. Using this framework, Chapter 3 extends their work to dependence in both dimensions and finds that

- time-series dependence leads to an extra $O(T^{-3/2})$ incidental parameter bias term,
- cross-section dependence leads to an extra bias term of a different nature, the order of which depends on the strength of dependence.

Naturally, the strength of dependence is quantified by ρ. The extra bias due to temporal dependence is $O(T^{-3/2})$ and so is negligible. Therefore, under time-series dependence, the Arellano-Bonhomme method is still valid. It is cross-section dependence which potentially can render their method inappropriate. Therefore, the assumptions under which the Arellano-Bonhomme method is still valid under cross-section dependence is also provided.

Finally, a spatial dependence/clustering based specific dependence structure is analysed. This is interesting as it makes it possible to analyse cross-section dependence without making a high-level assumption on the double-asymptotic convergence rates. Indeed, convergence rates are implied indirectly by the spatial dependence structure.

1.1.3 Chapter 4

The findings obtained in the first two Chapters are combined in Chapter 4 in order to achieve the ultimate objective of panel based volatility estimation in samples of around 150−200 observations. The major contribution of this chapter is to the financial econometrics literature, opening up the possibility of using the GARCH methodology to model volatility of financial series that are not large enough to use standard estimation approaches. Put differently, the results of this Chapter potentially make GARCH modelling available to a wider selection of applications and datasets.

The study in this part is based entirely on simulation exercises and empirical studies. It would be interesting from a theoretical point of view to fully investigate whether the high-order assumptions made in Chapter 3 indeed hold for GARCH models. However, this sort of a study is liable to become complicated very quickly and turn into a major research project. Therefore, it is not pursued here.

The simulation analysis investigates two different strategies for obtaining the bias-corrected estimators. These differ in the way the initial values are treated. In order to construct the GARCH likelihood function, an initial value for conditional variance, $\sigma_{i,0}^2$ has to be specified. Unfortunately, this value cannot be obtained from data because conditional volatility is latent. Therefore, it can never be observed, even ex-post. One option considered frequently is to use an estimate of $\mathbb{E}[\sigma_{i,0}^2]$, the unconditional variance, which can be shown to be equal to λ_i. In this case, it is possible to incorporate initial value selection into the bias-reduction framework, by asking the integrated likelihood function to integrate the initial value out, as well. Simulation results reveal that this improves the bias performance of the estimator greatly. In addition, the sample distribution of the bias-reduced estimator closely matches that of the infeasible estimator based on the true parameter values of λ_i.

The empirical analysis considers two illustrations. The first one is a comparison of predictive ability using daily stock market returns. Test results reveal that the bias-reduced estimators attain superior forecasting ability compared to the non-bias-corrected panel estimator and the standard time-series based GARCH estimator. In the final illustration, this novel method is used to investigate the conditional volatility of monthly hedge fund returns. This is another important contribution of this thesis. Due to the inherently small

sample sizes of monthly hedge fund returns, their volatility modelling using the GARCH methodology has hitherto been impossible. Therefore, the empirical analysis presented here is novel. The main observations are as follows. First, funds are characterised by different levels of average volatility, both within and across different strategies. Moreover, sample distributions of fund volatilities are asymmetric and skewed to right. The right tail tends to become larger during important financial events such as the burst of the dotcom bubble and the recent credit crunch episode. However, when adjusted by median volatility, volatility sample distributions tends to be symmetric and the right tail behaviour becomes more stable.

Chapter 2

Volatility Estimation Using Panels of Financial Data: The GARCH Panel Model

2.1 Introduction

Volatility modelling has been a popular area both in academia and industry since the seminal papers of Robert Engle (1982) and Tim Bollerslev (1986), which introduced GARCH-type (Generalised Autoregressive Conditional Heteroskedasticity) models. After three decades, it still remains an active research area, which testifies to the importance of the subject matter.

Broadly speaking, GARCH-type models focus on parametric modelling of univariate and multivariate volatility. The former corresponds to modelling the volatility process for each time-series individually while the latter approach is concerned with modelling time-varying covariance matrices for a collection of financial variables. This study contributes to the univariate modelling literature by introducing the GARCH Panel model, as a novel approach for panel estimation of volatility. A GARCH panel is simply a collection of financial time-series which have GARCH dynamics. To motivate and explain this approach, consider some financial variable of interest y_t (e.g. daily stock returns) observed at time t, $t = 1, ..., T$, such that

$$y_t|\mathcal{F}_{t-1} \sim F(0, \sigma_t^2),$$

where $\mathcal{F}_{t-1}$ is the information set at time $t-1$ and F is some zero-mean distribution with variance σ_t^2.[1] The standard GARCH(1,1) model (or any GARCH-type model) is a statement

[1] For clarity of exposition, y_t is implicitly assumed to be conditionally mean zero. This is a standard assumption for daily returns, although one can also model the conditional mean process, in addition to

about the specification of σ_t^2, the conditional variance:

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2.$$

Here, α and β govern the evolution of volatility dynamics, while ω is related to the long-run, or unconditional variance. The standard approach is to obtain an individual estimate $(\hat{\omega}, \hat{\alpha}, \hat{\beta})$ for each asset.

A common observation in the literature is that, when the model is fitted to, for example, daily returns for different stocks, estimates of α and β for different assets tend to be very close. This clustering of volatility dynamics then provides a strong motivation to assume parameter homogeneity for α and β and estimate these in a panel framework by pooling observations for all assets.[2] More specifically, given N assets indexed by $i = 1, ..., N$, which all follow GARCH(1,1) dynamics, homogeneity of the slope parameters implies that

$$y_{it}|\mathcal{F}_{i,t-1} \sim F(0, \sigma_{it}^2) \quad \text{and} \quad \sigma_{it}^2 = \omega_i + \alpha y_{i,t-1}^2 + \beta \sigma_{i,t-1}^2,$$

where $\mathcal{F}_{i,t-1}$ is the information set for individual i at time t.

The panel estimation framework introduced here has two objectives. The first objective is to obtain a superior estimator for (α, β) by using a richer information set. This is made possible by the empirical observation of parameter clustering as this implies parameter homogeneity. Consequently, a panel structure can be used to exploit both time-series and cross-sectional information. This is in contrast with the standard estimation approach, which utilises time-series information only. The second objective is a pragmatic one and is based on an empirical motivation. Accurate estimation of GARCH parameters by standard methods requires a large number of observations (potentially as large as 1,000-1,500 observations), due to the nonlinear dynamics of the GARCH model and the high levels of persistence exhibited by the conditional variance.[3] Then, a pragmatic question is whether, in small samples, it would be possible to make up for the lack of time-series variation/information by using a second source of information: cross-section variation. Es-

conditional volatility.

[2] Examples of approaches in this vein include Bauwens and Rombouts (2007), Barigozzi, Brownlees, Gallo and Veredas (2010), Brownlees (2010) and Aielli and Caporin (2011).

[3] For example, GARCH parameter estimates for stock market volatility usually imply high levels of persistence in conditional variance, close to being unit-root (Nelson (1991))

pecially for daily financial data, usually a long history of observations is available and so, lack of time-series information is non-existent. However, some interesting datasets are inherently short. One example is hedge fund returns. These are recorded almost invariably at monthly frequency and given that most datasets start in 1994, the maximum number of observations is around 200 (and usually much less than that). This then makes it virtually impossible to use standard GARCH tools to model hedge fund volatility. In addition, even when the interest in on variables with a long history of observations, the researcher might still have to work with a limited amount of data due to a recent structural break. Hence, a strong motivation for the analysis here is to propose a GARCH-based volatility modelling framework for datasets which are not large enough for the standard GARCH methodology to deliver accurate results.

The transition from time-series to panel modelling requires a parallel transition from modelling the marginal likelihood function to modelling the joint likelihood function. Unfortunately, this leads to both statistical and computational complications. In the case of GARCH estimation, the main cause for these issues is the presence of a large dimensional covariance matrix. As such, this matrix is difficult to model due to curse of dimensionality and leads to complications in numerical optimisation. The composite likelihood (CL) method provides a convenient approximation to the joint density and side-steps both issues. This method is based on the idea that, under regularity conditions, it is possible to approximate the joint density by the average of the marginal likelihoods and still obtain consistent estimators. Of course, some efficiency loss is expected due to misspecification, depending on the dependence characteristics of data.

The CL method offers a fast and effective way to exploit information contained in both dimensions, without making the analysis complicated. This chapter first develops the large sample theory of CL estimation for GARCH panels. The relevant asymptotic theory for a more general case has already been established by Engle, Shephard and Sheppard (2008), who introduced this method to financial econometrics. Therefore, the theoretical analysis here is largely a special case of their results. Then, small sample properties of this approach are analysed in a simulation exercise. In order to make a comparison, a recent alternative method by Engle (2009) is also considered. Simulation results reveal that the CL method performs well in samples with around 500 time-series observations. This is an

important improvement over the standard GARCH estimation approach. Simulation results also document that GARCH panels are subject to the well-known nuisance parameter issue (Neyman and Scott (1948)), observed commonly in the panel data literature.[4] The reason is estimation of ω_i which still only relies on time-series information and is far from accurate when T is small. Finally an empirical analysis of predictive ability using daily stock market data is considered. Results indicate that CL delivers more accurate forecasts compared to its alternatives and, more interestingly, stands out as a legitimate alternative to the standard time-series estimation approach when $T = 1,000$. This strongly suggests that even when the sample size is large enough for the standard GARCH estimation tools to work, there is still virtue in exploiting the panel structure.

The rest of this chapter is organised as follows: Section 2.2 provides a short introduction to GARCH-type models. Section 2.3 develops the estimation approach and the large sample theory for GARCH Panels. In addition, the alternative "MacGyver" method of Engle (2009) is discussed. The simulation and empirical analyses are conducted in Sections 2.4 and 2.5, respectively. Section 2.6 concludes. All proofs are given in the Mathematical Appendix.

2.2 A Short Introduction to GARCH Modelling

This section gives a short introduction to GARCH-type models, with a special focus on the tools that will be utilised frequently in this study. Although the purpose of this chapter is to analyse the application of the CL method in GARCH panels only, it is nevertheless necessary to provide a modest catalogue of the most widely used volatility models. A complete analysis of all models is unfortunately an impossible task. Nevertheless, there are many references that offer a more detailed discussion. An excellent, clear introduction to univariate and multivariate volatility modelling can be found in Sheppard (2010). Most of the treatment below is based on this reference. A detailed and more technical textbook treatment is available by Francq and Zakoïan (2010). Although perhaps outdated, a good starting point could be the general overviews by Bollerslev, Chou and Kroner (1992) and Bollerslev, Engle and Nelson (1994). Poon and Granger (2003) and Andersen, Bollerslev, Christoffersen and Diebold (2006) provide a more general review of volatility modelling and forecasting, while

[4] This issue is the central topic of Chapter 3, where it will be discussed in detail. In the financial econometrics literature, it has recently been observed by Engle and Sheppard (2001) and Engle, Shephard and Sheppard (2008).

Andersen, Davis, Kreiss and Mikosch (2009) includes many chapters devoted to different aspects of GARCH modelling. For other textbook treatments, see Gourieroux and Jasiak (2001), Taylor (2005), Tsay (2005) and Enders (2010). Also, a collection of key published articles can be found in Engle (1995).

2.2.1 Some Stylized Facts on Financial Data

Financial time-series share some important stylised facts. Perhaps the most obvious property is that their volatilities exhibit a predictable pattern over time. Generally, periods of high (low) volatility tend to be followed by periods of high (low) volatility. This is also known as volatility clustering. This is in part due to the high persistence in square returns. Interestingly, this type of high persistence is not observed in the returns series themselves. Hence, this predictability is one of the main motivations for modelling conditional volatility.

Distributions of daily returns also generally have fatter tails than that of the Normal distribution. Distributions with this property are called "leptokurtic". One interpretation of this is that extreme events are more likely to happen than the Normal distribution would predict. A further observation is that, distributions of returns can sometimes be skewed. This is especially relevant to co-movements (or, covariances) between two series, as financial returns are more likely to decrease together (such as, during a financial crisis) than to increase together. Finally, another well-observed characteristic of financial returns is that, large negative changes in time series, such as large decreases in stock prices, have a bigger impact on volatility than a large positive change. Moreover, the direction and the speed of change in volatility also depend on the direction of the shock. In general, bad news lead to higher volatility while good news cause a decline in volatility. This is called the "statistical leverage effect." For a more detailed discussion of these observations, see also Taylor (2005, Chapter 4) and Enders (2010).

2.2.2 The Autoregressive Conditional Heteroskedasticity (ARCH) Model

The ARCH(P) model is due to Engle (1982). Focusing on observations from a time series r_t, $t = 1, ..., T$, the ARCH(P) model is specified as[5]

$$r_t = \mu_t + \varepsilon_t, \quad \mu_t = \mathbb{E}_{t-1}[r_t], \quad \mathbb{E}_{t-1}[\varepsilon_t] = 0, \quad Var_{t-1}(\varepsilon_t) = \sigma_t^2, \tag{2.1}$$

$$\sigma_t^2 = \gamma + \sum_{p=1}^{P} \alpha_p \varepsilon_{t-p}^2, \quad \text{where} \quad \sum_{p=1}^{P} \alpha_p < 1, \ \alpha_p \geq 0 \ \forall p \quad \text{and} \quad \gamma > 0. \tag{2.2}$$

The setting given by (2.1) is standard in the literature. The difference comes from the way the conditional variance is specified. Here, ε_t can be interpreted as a shock variable. For example, assuming r_t are stock returns and μ_t are the conditional expected returns, ε_t give the unanticipated change in (and hence the shock to) the stock price at t. $\sum_{p=1}^{P} \alpha_p < 1$ is necessary to establish covariance stationarity, while other parameter restrictions ensure that conditional variances are positive. It must be underlined that, at this stage no distributional assumptions are necessary.

That σ_t^2 is the conditional variance for r_t can be shown easily:

$$Var_{t-1}(r_t) = \mathbb{E}_{t-1}[(r_t - \mu_t)^2] = \mathbb{E}_{t-1}[\varepsilon_t^2] = \sigma_t^2,$$

since ε_t is a zero-mean random variable. Although ARCH(P) has been empirically very successful, one caveat is that it requires a large number of lags in the conditional covariance specification, to work properly. As a possible solution, Bollerslev (1986) provides the Generalised ARCH (GARCH) model.

2.2.3 The Generalised Autoregressive Conditional Heteroskedasticity (GARCH) Model

GARCH manages to include infinitely many lags of ε_t^2 at the cost of a single extra parameter only. Specifically, the GARCH(P,Q) model is given by

$$\sigma_t^2 = \gamma + \sum_{p=1}^{P} \alpha_p \varepsilon_{t-p}^2 + \sum_{q=1}^{Q} \beta_q \sigma_{t-q}^2 \ . \tag{2.3}$$

[5] In the remainder of this work, $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{t-1}]$ and $Var_{t-1}(\cdot) = Var(\cdot|\mathcal{F}_{t-1})$.

Derivation of parameter restrictions is very involved for the general GARCH(P,Q) case. However, for GARCH(1,1) it can be shown that these conditions are

$$\gamma > 0, \quad \alpha, \beta \in [0, 1) \quad \text{and} \quad \alpha + \beta < 1. \tag{2.4}$$

Parameters are restricted to be non-negative as a negative parameter can feed a negative effect into the system which might lead to negative conditional variances. Moreover, neither α, nor β and nor their sum can exceed 1 as this will lead to non-stationarity. This will become more obvious once the ARMA(1,1) interpretation of GARCH(1,1) is derived. Perhaps most importantly, β is not identified when $\alpha = 0$. For an intuitive explanation, first observe that when $\alpha = 0$, the model becomes $\sigma_t^2 = \gamma + \beta\sigma_{t-1}^2$. As will be derived below, $\sigma_t^2 = \sum_{j=0}^{\infty} \beta^j \gamma = \gamma/(1-\beta)$. Hence, for any selection of γ, a corresponding selection of β can be found which will yield the same σ_t^2. Therefore, β cannot be identified.

It is obvious that GARCH(P,Q) implies similar dynamics as ARCH(P), except that the former's dynamics are based on past observations of σ_t^2, as well. Nevertheless, some further algebraic manipulations are necessary to fully see the connection between the two models. For expositional simplicity, the following analysis focuses on GARCH(1,1) only. However, it can be extended to the more general GARCH(P,Q) case, at the cost of more involved algebra.

First, it can be shown by backward substitution that

$$\begin{aligned}
\sigma_t^2 &= \gamma + \alpha\varepsilon_{t-1}^2 + \beta\left(\gamma + \alpha\varepsilon_{t-2}^2 + \beta\sigma_{t-2}^2\right) \\
&= (1+\beta)\gamma + \alpha\left(\varepsilon_{t-1}^2 + \beta\varepsilon_{t-2}^2\right) + \beta^2\left(\gamma + \alpha\varepsilon_{t-3}^2 + \beta\sigma_{t-3}^2\right) \\
&\vdots \\
&= \lim_{k\to\infty} \beta^k \sigma_{t-k}^2 + \sum_{j=0}^{\infty} \beta^j \gamma + \sum_{j=0}^{\infty} \alpha\beta^j \varepsilon_{t-j-1}^2.
\end{aligned}$$

Since $\beta \in [0, 1)$ by (2.4), $\lim_{k\to\infty} \beta^k \sigma_{t-k}^2 = 0$. Thus,

$$\sigma_t^2 = \sum_{j=0}^{\infty} \beta^j \gamma + \sum_{j=0}^{\infty} \delta_j \varepsilon_{t-j-1}^2, \ \delta_j = \alpha\beta^j. \tag{2.5}$$

Equation (2.5) reveals that GARCH(1,1) is in fact equivalent to ARCH(∞). Clearly, this

solves the problem of having to include a large number of lags in the conditional variance specification when using ARCH.

Estimation

Estimation by the maximum likelihood (ML) method requires a distributional assumption for ε_t and a common choice is the Normal distribution. In this case, it is common to use a more convenient specification;

$$\varepsilon_t|\mathcal{F}_{t-1} \sim \mathcal{N}\left(0, \sigma_t^2\right), \quad \text{where } \varepsilon_t = \sigma_t e_t \quad \text{and} \quad e_t \overset{iid}{\sim} \mathcal{N}\left(0,1\right). \tag{2.6}$$

This implies that,

$$\begin{aligned} \ell_t(\alpha,\beta,\gamma) &= \log L(\alpha,\beta,\gamma; r_t|\mathcal{F}_{t-1}) = \log f\left(r_t|\mathcal{F}_{t-1};\alpha,\beta,\gamma\right) \\ &= -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(\sigma_t^2\right) - \frac{\left(r_t-\mu_t\right)^2}{2\sigma_t^2}, \end{aligned}$$

where $f\left(r_t|\mathcal{F}_{t-1};\alpha,\beta,\gamma\right)$ is the conditional distribution for r_t and $L(\alpha,\beta,\gamma;r_t|\mathcal{F}_{t-1})$ is the corresponding quasi-likelihood function. For GARCH(1,1),

$$\frac{\partial \ell_t(\alpha,\beta,\gamma)}{\partial \gamma} = \frac{\partial \ell_t(\alpha,\beta,\gamma)}{\partial \sigma_t^2}\frac{\partial \sigma_t^2}{\partial \gamma} = \frac{\partial \ell_t(\alpha,\beta,\gamma)}{\partial \sigma_t^2}\left(1+\beta\frac{\partial \sigma_{t-1}^2}{\partial \gamma}\right), \tag{2.7}$$

$$\frac{\partial \ell_t(\alpha,\beta,\gamma)}{\partial \alpha} = \frac{\partial \ell_t(\alpha,\beta,\gamma)}{\partial \sigma_t^2}\frac{\partial \sigma_t^2}{\partial \alpha} = \frac{\partial \ell_t(\alpha,\beta,\gamma)}{\partial \sigma_t^2}\left(\varepsilon_{t-1}^2+\beta\frac{\partial \sigma_{t-1}^2}{\partial \alpha}\right), \tag{2.8}$$

$$\frac{\partial \ell_t(\alpha,\beta,\gamma)}{\partial \beta} = \frac{\partial \ell_t(\alpha,\beta,\gamma)}{\partial \sigma_t^2}\frac{\partial \sigma_t^2}{\partial \beta} = \frac{\partial \ell_t(\alpha,\beta,\gamma)}{\partial \sigma_t^2}\left(\sigma_{t-1}^2+\beta\frac{\partial \sigma_{t-1}^2}{\partial \beta}\right). \tag{2.9}$$

The score matrices for ARCH(P) and GARCH(P,Q) are provided by Engle (1982) and Bollerslev (1986), respectively. Due to the recursive structure in (2.7)-(2.9), estimation is usually conducted by numerical optimisation.

An immediate issue is the validity of the distributional assumption. This is an important concern as financial time series are not necessarily conditionally Normal. They are likely to be skewed in one direction or have fatter tails, as mentioned previously. For this reason, alternative distributions such as the Student's t for fatter tails, Hansen's skewed t (Hansen (1994)) or the Generalised Error Distribution (GED) can be considered, as well.

Incorrect distributional assumptions will most likely lead to inefficient estimators. How-

ever, Bollerslev and Wooldridge (1992) show for GARCH that as long as the conditional mean and covariance specifications are correct, estimators will still be consistent under the Normality assumption, even when this assumption is violated. Moreover, their simulation analysis reveals that the efficiency loss when the underlying distribution is the t-distribution is little.[6] This is an application of the well-known quasi maximum likelihood estimation (QMLE) principle (see, for example, White (1982) and Gourieroux, Monfort and Trognon (1984)). Bollerslev and Wooldridge (1992) also develop asymptotic standard errors and a simple Lagrange Multiplier test procedure both of which are robust to the violation of the normality assumption. Following their results, the rest of this section is based on (2.6).

Unconditional Variance and Variance Targeting

The unconditional variance $\sigma^2 = \mathbb{E}[\sigma_t^2]$ can be derived as follows:

$$\begin{aligned}
\mathbb{E}[\sigma_t^2] &= \mathbb{E}[\varepsilon_t^2] = \mathbb{E}\{\mathbb{E}_{t-1}[\varepsilon_t^2]\} \\
&= \mathbb{E}\left[\gamma + \sum_{p=1}^{P} \alpha_p \varepsilon_{t-p}^2 + \sum_{q=1}^{Q} \beta_q \sigma_{t-q}^2\right] = \gamma + \sum_{p=1}^{P} \alpha_p \mathbb{E}[\varepsilon_{t-p}^2] + \sum_{q=1}^{Q} \beta_q \mathbb{E}[\sigma_{t-q}^2] \\
&= \gamma + \sum_{p=1}^{P} \alpha_p \mathbb{E}[\sigma_{t-p}^2] + \sum_{q=1}^{Q} \beta_q \mathbb{E}[\sigma_{t-q}^2] = \gamma + \sum_{p=1}^{P} \alpha_p \sigma^2 + \sum_{q=1}^{Q} \beta_q \sigma^2, \\
\text{implying,} \quad \sigma^2 &= \frac{\gamma}{1 - \sum_{p=1}^{P} \alpha_p - \sum_{q=1}^{Q} \beta_q}. \qquad (2.10)
\end{aligned}$$

The first step uses the Law of Iterated Expectations, while in the third step, $\mathbb{E}[\varepsilon_{t-p}^2] = \mathbb{E}[\sigma_{t-p}^2 e_{t-p}^2] = \mathbb{E}[\sigma_{t-p}^2]\mathbb{E}[e_{t-p}^2]$. This is because, σ_{t-p}^2 are determined by lagged observations of ε_{t-p}^2 (or e_{t-p}^2, for that matter) and since e_{t-p} are distributed independently, σ_{t-p}^2 and e_{t-p}^2 are also independent for all t and p. Lastly, using $\sigma^2 = \mathbb{E}[\sigma_t^2]$ leads to the final result. Clearly, $\sum_{p=1}^{P} \alpha_p + \sum_{q=1}^{Q} \beta_q < 1$ is a necessary condition for the existence of a non-negative σ^2.

Observing (2.10), the following specification simplifies estimation of the intercept parameter greatly:

$$\sigma_t^2 = \gamma(1 - \sum_{p=1}^{P} \alpha_p - \sum_{q=1}^{Q} \beta_q) + \sum_{p=1}^{P} \alpha_p \varepsilon_{t-p}^2 + \sum_{q=1}^{Q} \beta_q \sigma_{t-q}^2, \qquad (2.11)$$

[6] However, they also note that it can be high when the true distribution is asymmetric.

$$\text{where} \quad \sum_{p=1}^{P} \alpha_p + \sum_{q=1}^{Q} \beta_q < 1 \quad \text{and} \quad \gamma > 0,$$

$$\text{implying} \quad \sigma^2 = \frac{\lambda(1 - \sum_{p=1}^{P} \alpha_p - \sum_{q=1}^{Q} \beta_q)}{1 - \sum_{p=1}^{P} \alpha_p - \sum_{q=1}^{Q} \beta_q} = \lambda. \tag{2.12}$$

Equation (2.12) follows from (2.10), using the new intercept parameter, $\lambda \left(1 - \sum_{p=1}^{P} \alpha_p - \sum_{q=1}^{Q} \beta_q\right)$. All of these results can be obtained for ARCH(P) by assuming $\beta_q = 0$. This specification is called the "variance targeting" specification and has been introduced by Engle and Mezrich (1996).[7]

In the following chapters, as the focus is on the conditional covariance dynamics, μ_t will be assumed to be equal to 0. Then, $\sigma^2 = \mathbb{E}[\varepsilon_t^2] = \mathbb{E}[r_t^2]$ and, λ can be estimated easily by using a method of moments estimator:

$$\tilde{\lambda} = \frac{1}{T} \sum_{t=1}^{T} r_t^2. \tag{2.13}$$

Hence, using (2.11) decreases the number of simultaneously estimated parameters from three to two. Consequently, only (2.8) and (2.9) are relevant for QMLE.[8]

Autoregressive (AR) and Autoregressive Moving Average (ARMA) Representations

ARCH and GARCH have AR and ARMA representations, respectively, as well. To see this, for ARCH

$$\begin{aligned} \sigma_t^2 &= \lambda + \alpha\varepsilon_{t-1}^2 = \lambda + \alpha\varepsilon_{t-1}^2 + (\varepsilon_t^2 - \sigma_t^2) - (\varepsilon_t^2 - \sigma_t^2) \\ \varepsilon_t^2 &= \lambda + \alpha\varepsilon_{t-1}^2 + \sigma_t^2 \left(e_t^2 - 1\right) = \lambda + \alpha\varepsilon_{t-1}^2 + v_t, \quad \text{where} \quad v_t = \sigma_t^2 \left(e_t^2 - 1\right), \end{aligned} \tag{2.14}$$

while for GARCH

$$\begin{aligned} \sigma_t^2 &= \lambda + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + \left(\varepsilon_t^2 - \sigma_t^2\right) - \left(\varepsilon_t^2 - \sigma_t^2\right) \\ \varepsilon_t^2 &= \lambda + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + \left(\varepsilon_t^2 - \sigma_t^2\right) - \beta\varepsilon_{t-1}^2 + \beta\varepsilon_{t-1}^2 \end{aligned}$$

[7] See also a recent paper by Francq, Horváth and Zakoïan (2011) who provide a theoretical and empirical comparison of the standard quasi maximum likelihood and variance targeting based approaches.

[8] Note that $\hat{\theta}$ is not efficient, since the estimation method is not the maximum likelihood method anymore, as $\tilde{\lambda}$ is not an ML but a method of moments estimator.

$$\varepsilon_t^2 \;=\; \lambda + (\alpha+\beta)\,\varepsilon_{t-1}^2 - \beta v_{t-1} + v_t, \qquad \text{where} \qquad v_t = \sigma_t^2\left(e_t^2 - 1\right). \tag{2.15}$$

In both cases, v_t is a zero-mean iid process $\forall t$ since $e_t \overset{iid}{\sim} N(0,1)$. Equations (2.14) and (2.15) show why $\alpha < 1$ and $(\alpha+\beta) < 1$ are necessary to establish covariance stationarity, for ARCH and GARCH, respectively. As these are the coefficients for the lagged squared error, they also serve as a measure of persistence.

Forecasting

Here, the focus is again on the GARCH(1,1) model, though the same mechanism can be employed for other lag lengths, as well. Starting with the one-period ahead forecast, (2.3) reveals that, if the true parameter values are known, then the forecast is also known, since ε_t and σ_t belong to $\mathcal{F}_t$. Taking the conditional expectation,

$$\mathbb{E}_t[\sigma_{t+1}^2] = \mathbb{E}_t[\lambda + \alpha\varepsilon_t^2 + \beta\sigma_t^2] = \lambda + \alpha\varepsilon_t^2 + \beta\sigma_t^2.$$

Similarly,

$$\begin{aligned}
\mathbb{E}_t[\sigma_{t+2}^2] &= \mathbb{E}_t[\lambda + \alpha\varepsilon_{t+1}^2 + \beta\sigma_{t+1}^2] = \lambda + \alpha\mathbb{E}_t[\sigma_{t+1}^2]\mathbb{E}_t[e_{t+1}^2] + \beta\mathbb{E}_t[\sigma_{t+1}^2] \\
&= \lambda + (\alpha+\beta)\mathbb{E}_t[\sigma_{t+1}^2] = \lambda + (\alpha+\beta)(\lambda + \alpha\varepsilon_t^2 + \beta\sigma_t^2) \\
&= (1+\alpha+\beta)\lambda + (\alpha+\beta)\alpha\varepsilon_t^2 + (\alpha+\beta)\beta\sigma_t^2, \\
\text{and } \mathbb{E}_t[\sigma_{t+3}^2] &= \mathbb{E}_t[\lambda + \alpha\varepsilon_{t+2}^2 + \beta\sigma_{t+2}^2] = \lambda + (\alpha+\beta)\mathbb{E}_t[\sigma_{t+2}^2] \\
&= \lambda + (\alpha+\beta)\lambda + (\alpha+\beta)^2\lambda + (\alpha+\beta)^2\alpha\varepsilon_t^2 + (\alpha+\beta)^2\beta\sigma_t^2.
\end{aligned}$$

This pattern leads to the h-step ahead forecast,

$$\mathbb{E}_t[\sigma_{t+h}^2] = \left[\sum_{i=0}^{h-1}(\alpha+\beta)^i\lambda\right] + (\alpha+\beta)^{h-1}(\alpha\varepsilon_t^2 + \beta\sigma_t^2). \tag{2.16}$$

Note that, it can be shown that the h-step ahead forecast for ARCH(1) can be obtained from (2.16) by using $\beta = 0$. The same goes for ARCH(P) and GARCH(P,Q), as well, using $\beta_q = 0$.

Kurtosis

Lastly, it can be shown that the kurtosis for ε_t under ARCH and GARCH are equal to

$$K_{ARCH} = \frac{3(1-\alpha^2)}{(1-3\alpha^2)} > 3 \quad \text{and} \quad K_{GARCH} = \frac{3(1+\alpha+\beta)(1-\alpha-\beta)}{1-2\alpha\beta-3\alpha^2-\beta^2} > 3,$$

respectively; see, for example, Tsay (2005). Although ε_t are based on the Normally distributed innovation variables e_t, the kurtosis is larger than that implied by the standard normal distribution. This is because these shocks are all weighted by σ_t^2. Therefore, ARCH/GARCH shocks and the time-series that depend on them have fat-tailed distributions, satisfying one of the stylised facts of financial time-series.

2.2.4 Further Specifications

GARCH(1,1) is widely used in financial applications and stands out as one of the benchmark models of financial econometrics. Indeed, Hansen and Lunde (2005) analyse out-of-sample forecasting performance of 330 ARCH-type models using DM/USD exchange rate and IBM returns data. They find that GARCH(1,1) is not outperformed when the analysis focuses on the exchange rate data, while models that account for leverage effect clearly perform better when stock market data is used. As explained before, leverage effect is a very common observation in financial data, which has led to further specifications of conditional volatility. Some of these are mentioned here.

Two important shortcomings of GARCH can be identified. Firstly, GARCH requires parameter restrictions to ensure positive conditional variances. These may become increasingly difficult to both specify and maintain as the number of lags increases. Secondly, GARCH generally fails to account for the leverage effect observed in financial data, since it uses squared errors. Thus, it is impossible to model asymmetric effects for shocks.

As a possible solution, Nelson (1991) proposes the Exponential GARCH (EGARCH) model, which models the natural logarithm of conditional variance. Therefore, no parameter restrictions are necessary anymore. Moreover, leverage effects are introduced by using a weighted innovation term, instead of using e_t^2 only. As a result, the contribution of negative shocks to log variance is different than that of positive shocks. A different approach leads to the GJR-GARCH model suggested by Glosten, Jagannathan and Runkle (1993) who

use an indicator function to provide an extra parameter for negative shocks. However, this model still involves parameter restrictions, since the focus is again on conditional variance. As distinct from these two contributions, Zakoïan (1994) focuses on the specification of the conditional standard deviation rather than the conditional variance (the Threshold-ARCH (TARCH) model). This enables using the absolute values of shocks which are less responsive to changes than the squared shocks. This model is also known in the literature as the ZARCH model, after Zakoïan (1994) and the Absolute Value GARCH (AVGARCH) model, which is a special case of the full model. Another interesting approach is due to Ding, Engle and Granger (1993) who, instead of focusing on σ_t or σ_t^2, choose to estimate the power of σ_t, as well. This is known as the Asymmetric Power ARCH (APARCH) model. It can be shown that their specification nests not only the GJR-GARCH and TARCH models, but also the ARCH and GARCH models. Consequently, this is a more flexible model.

Finally, a further strand of volatility modelling is multivariate volatility modelling. This approach focuses on a collection of assets in order to model not only the conditional variance but also the conditional covariance. Parallel to the univariate analysis, ensuring positive definite covariance matrices is a key requirement in multivariate modelling. The other central issue is parsimony. Generally, although models with a smaller number of parameters (i.e., parsimonious models) are very effective, in some cases estimation may require days, even for a modest number of assets. Some of the better known examples are the VEC (Bollerslev, Engle and Wooldridge (1988)), BEKK (Engle and Kroner (1995)), Constant Conditional Correlation (Bollerslev (1990)) and Dynamic Conditional Correlation (Engle and Sheppard (2001) and Engle (2002)) models. Moreover, modelling of asymmetric effects has also been adopted in multivariate modelling by, for example, Kroner and Ng (1998), Cappiello, Engle and Sheppard (2006) and Kawakatsu (2006).[9] For comprehensive surveys, see Bauwens, Laurent and Rombouts (2006) and Silvennoinen and Teräsvirta (2009).

2.3 The GARCH Panel and the Composite Likelihood Method

This section discusses the composite likelihood (CL) method and establishes the large sample theory for the special case of GARCH panels. A related estimation method proposed by

[9] These models are the Asymmetric Dynamic Covariance Matrix, Asymmetric Dynamic Conditional Correlation and Matrix Exponential models, respectively.

Engle (2009), the MacGyver (MG) method, is an important natural alternative to the CL method. These methods will be compared to each other in the simulation and empirical analyses. Therefore, a quick overview of MG is also given.

The large sample theory for the general case is due to Engle, Shephard and Sheppard (2008) and the asymptotic analysis presented here is basically an adaptation of their results to the specific case at hand.

2.3.1 The GARCH Panel

The GARCH panel is essentially a collection of financial time-series that are assumed to have GARCH dynamics and share a common set of parameters, $\theta = (\alpha, \beta)$, while the intercept parameters, λ_i, are allowed to be asset-dependent. Using panel data terminology, $\theta = (\alpha, \beta)$ is the common parameter of interest, while λ_i are the incidental parameters. The respective (pseudo) true parameters are given by λ_{i0}, α_0 and β_0. Also, $\lambda = (\lambda_1, ..., \lambda_N)$ and $\lambda_0 = (\lambda_{10}, ..., \lambda_{N0})$. The parameter homogeneity assumption is based on the empirical observation that estimates of θ for different assets cluster around similar values, as discussed in the Introduction. This is a crucial assumption because it allows for information pooling across asset panels to obtain a single estimator for θ rather than N different ones.

Let y_{it} be the return on asset i at time t where $i = 1, ..., N$ and $t = 1, ..., T$. Throughout, large-T large-N asymptotics are assumed; hence, $N/T \to c$ as $N, T \to \infty$ where $0 < c < \infty$. In addition, asset returns are allowed to display cross-section dependence of unknown form. The asset panel is given by $y = (y_1, ..., y_T)'$ where $y_t = (y_{1t}, y_{2t}, ..., y_{Nt})'$. Moreover,

$$y_{it} = \mu_{it} + \varepsilon_{it}, \quad \mu_{it} = \mathbb{E}[y_{it}|\mathcal{F}_{i,t-1}], \quad \varepsilon_{it}|\mathcal{F}_{i,t-1} \sim F(0, \sigma_{it}^2),$$

where F is some distribution and $\mathcal{F}_{i,t-1}$ is the information set for asset i at time $t-1$. Formally, $\mathcal{F}_{i,t-1} = \sigma(\sigma_{it}, y_{i,t-1}, \sigma_{i,t-1}, y_{i,t-2}, \sigma_{i,t-2}, ...)$ and $\mathcal{F}_{t-1} = \sigma(\sigma_t, y_{t-1}, \sigma_{t-1}, y_{t-2}, \sigma_{t-2}, ...)$, where $\sigma_t = (\sigma_{1t}, ..., \sigma_{Nt})$.[10] As the focus is on modelling conditional variance, hereafter it is simply assumed that $\mu_{it} = 0 \ \forall i, t$. This is a reasonable setting for, for example, daily returns which are generally zero-mean.

[10] Note that $\sigma(\cdot)$ denotes the sigma-algebra and is not related to conditional volatility, σ_t, despite similarity in notation.

The relevant GARCH(1,1) specification is

$$\sigma_{it}^2 = \lambda_i(1-\alpha-\beta) + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2, \tag{2.17}$$

$$\text{where} \quad \lambda_i > 0 \ \forall i; \quad \alpha, \beta \in [0,1) \quad \text{and} \quad \alpha+\beta < 1. \tag{2.18}$$

It was shown previously that this specification, often called variance targeting (Engle and Mezrich (1996)), implies that

$$\mathbb{E}[y_{it}^2] = \lambda_i.$$

Therefore, λ_i can be estimated separately from θ by using the method of moments, which can then be used to obtain a likelihood-based estimator of θ. Estimation of θ by the maximum likelihood (ML) method still requires the researcher to make a parametric assumption concerning the unknown distribution, F. By the quasi maximum likelihood (QML) argument of Bollerslev and Wooldridge (1992), as long as the conditional mean and variance are correctly specified, one can assume that F is the Normal distribution and still obtain consistent estimators. Hence, also given the convenience of working with the Normal distribution, in the remainder F will be assumed to be the Normal distribution.

Conventionally, estimation of θ can be conducted for each asset individually, that is by obtaining a separate $\hat{\theta}$ for each asset in the panel. However, this would exploit time-series information only. Under parameter homogeneity, a better option would be to pool all information available in the panel and use both cross-section and time-series information. This is made possible by the CL method.

2.3.2 The Composite Likelihood Method

The composite likelihood method can be traced back to Besag (1974) and Lindsay (1988). More recently, it has been studied by Cox and Reid (2004), Varin and Vidoni (2005) and Varin (2008), among others. A survey is given by Varin, Reid and Firth (2011). This method has been introduced to the financial econometrics literature by Engle, Shephard and Sheppard (2008), who utilise it to model multivariate volatility in the presence of incidental parameters.

Let $f(y_{it}|\mathcal{F}_{i,t-1};\theta,\lambda_i)$ be the univariate density for the return on asset i at time t, and

$f(y_t|\mathcal{F}_{t-1};\theta,\lambda,\Sigma_t)$ be the joint density for all asset returns at time t, where Σ_t is the set of parameters that govern the cross-section dependence structure. Moreover, $\theta \in \Theta$ and $\lambda_i \in \Lambda_i$ with $\lambda = (\lambda_1, ..., \lambda_N) \in \Lambda = \Lambda_1 \times ... \times \Lambda_N$. The parameter spaces satisfy the conditions set in (2.18) and are assumed to be compact. Concisely, $\psi_i = (\theta, \lambda_i) \in \Psi_i = \Theta \times \Lambda_i$ and $\psi = (\theta, \lambda) \in \Psi = \Theta \times \Lambda$.

Let the (normalised) true joint density be given by[11]

$$\tilde{\ell}_{NT}(\psi, \Sigma_1, ..., \Sigma_T) = \frac{1}{NT} \sum_{t=1}^{T} \log f(y_t|\mathcal{F}_{t-1};\psi,\Sigma_t). \tag{2.19}$$

Although theoretically it is always more desirable to base estimation on the true density, which would guarantee efficiency, in reality a convenient specification (or maybe any specification) for the joint density may not be available. In the GARCH panel case, the "inconvenience" is manifested in terms of statistical and computational problems. To illustrate these, consider the following example. Let y_t be mean zero multivariate Normal with the $(N \times N)$ covariance matrix Σ_t. Then,

$$\log f(y_t|\mathcal{F}_{t-1};\psi,\Sigma_t) \propto -\frac{1}{2} \ln |\Sigma_t| - \frac{1}{2} y_t' \Sigma_t^{-1} y_t.$$

The statistical issue is the curse of dimensionality in modelling the covariance matrix Σ_t, which involves $O(N^2)$ free parameters. One could of course employ some smart parameter reduction method, such as factor modelling, to overcome this issue. Indeed, without doubt, there are important virtues in understanding the dependence structure. However, the aim of this paper is to obtain a fast and consistent estimator of θ by using as much information as possible. As far as this objective is concerned, the underlying dependence structure is of no interest and, consequently, adding an extra layer of statistical modelling by incorporating Σ_t into the analysis is simply not desirable.

Even if the dependence structure could be modelled without much complication, there is still a second, computational problem. Remember that maximisation of the likelihood function for GARCH estimation is done numerically, as closed form expressions for likelihood derivatives do not exist (see the discussion in Section 2.2.3). This requires inversion of Σ_t,

[11] Here and elsewhere, all densities are normalised by the relevant sample size. For example, $(NT)^{-1} \sum_{t=1}^{T} \log f(y_t|\mathcal{F}_{t-1};\psi,\Sigma_t)$ rather than $\sum_{t=1}^{T} \log f(y_t|\mathcal{F}_{t-1};\psi,\Sigma_t)$ is used.

an $(N \times N)$ matrix, for each iteration of the optimisation process.[12] Even for moderate N, this would be a time-consuming task, making it virtually impossible to consider panels with more than $40-50$ assets. This provides another strong motivation to eliminate Σ_t from the analysis.

The composite likelihood (CL) method provides a convenient solution to these complications. The main idea is to use an approximation to the joint density in order to obtain a less complicated pseudo-likelihood function. This approximation can, for example, be a function of univariate marginal and/or bivariate joint densities[13], assuming that they can be specified conveniently. For notational simplicity, define

$$\ell_{it}\left(\theta, \lambda_i\right) = \log f(y_{it}|\mathcal{F}_{i,t-1}; \theta, \lambda_i).$$

Then, following Cox and Reid (2004), the composite likelihood function based on univariate marginal densities, which approximates (2.19), is given by

$$\ell_{NT}(\psi) = \frac{1}{NT}\sum_{t=1}^{T}\ell_{tN}(\psi) \quad \text{where} \quad \ell_{tN}(\psi) = \frac{1}{N}\sum_{i=1}^{N}\ell_{it}\left(\theta, \lambda_i\right). \tag{2.20}$$

Thus, using (2.20) for maximum likelihood estimation virtually corresponds to assuming that data are cross-sectionally independent. The cost of this is the inefficiency due to ignoring the Copula and misspecifying the joint density. Still, the degree of inefficiency will depend on the particular dependence structure and will not necessarily be severe. It is possible to construct the composite likelihood function by using bivariate densities, as well, which would automatically take co-movements into account. This approach would be more natural for multivariate volatility analysis and is not pursued here.[14]

[12] Such methods evaluate the likelihood function for different selections of $(\alpha, \beta, \lambda_1, ..., \lambda_N)$ based on some search algorithm and so, Σ_t has to be inverted for each such selection until the optimisation process decides that the maximum has been found.

[13] In this case, a bivariate joint density for assets i and j would be given by $-\frac{1}{2}\ln|\Sigma_{ij,t}| - \frac{1}{2}(y_{i,t}, y_{j,t})\Sigma_{ij,t}^{-1}(y_{i,t}, y_{j,t})'$, where $\Sigma_{ij,t} = Cov[(y_{i,t}, y_{j,t})']$.

[14] In the case of multivariate volatility analysis, conditional covariances are also of direct interest. Therefore, co-movements have to be accounted for. Approximating the joint density using bivariate marginal densities decreases computation time greatly as the composite-likelihood will involve many (2×2) matrices, instead of the full $(N \times N)$ covariance matrix. See Cox and Reid (2004) for more discussion on approximation by bivariate densities and Engle, Shephard and Sheppard (2008) for the application of this idea to multivariate volatility modelling.

2.3.3 Estimation

The standard parametric estimation approach for the given parameter structure would be concentrated likelihood estimation where

$$\hat{\lambda}_i(\theta) = \arg\max_{\lambda_i} \frac{1}{T} \sum_{t=1}^{T} \ell_{it}(\theta, \lambda_i), \tag{2.21}$$

$$\text{and} \quad \hat{\theta} = \arg\max_{\theta} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \ell_{it}(\theta, \hat{\lambda}_i(\theta)). \tag{2.22}$$

Therefore, first λ_i is estimated for each asset, given some fixed θ. Then the composite likelihood function is concentrated with respect to these estimators. Finally, $\hat{\theta}$ is obtained by maximising the "concentrated composite likelihood" function.[15] The potential problem with this approach is that estimation is a $N + 2$ dimensional problem. When N is large, numerical optimisation will become both time-consuming and, possibly, unstable.

Fortunately, there is a much more convenient alternative. Remember that the specification in (2.17) implies that $\mathbb{E}[y_{it}^2] = \lambda_i$. Then, λ_i can be estimated by method of moments in a first step, to obtain $\tilde{\lambda}_i$. These can then be used to construct $(NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \ell_{it}(\theta, \tilde{\lambda}_i)$ and estimate θ in a second step based on this objective function. By this framework, estimation of ψ is based on a two-step estimation procedure. A detailed exposition of the theory for two-step estimation is provided by Newey and McFadden (1994), who analyse this method under the generalised method of moments (GMM) framework. However, as they keep N fixed, standard results do not apply to the current case. Consequently, the large sample theory for this particular case is developed in Section 2.3.5.[16]

To discuss the estimation method in more detail, start with the population moment condition for λ_i,

$$\mathbb{E}[m_N(y_t, \lambda_0)] = 0, \quad \text{where} \quad m_N(y_t, \lambda) = \begin{bmatrix} y_{1t}^2 - \lambda_1 \\ \vdots \\ y_{Nt}^2 - \lambda_N \end{bmatrix}. \tag{2.23}$$

[15] For a more detailed discussion on concentrated likelihood estimation and other related concepts, see Barndorff-Nielsen and Cox (1994) and Severini (2000).

[16] It is important to note that although the outlined two-step procedure resembles of the concentrated likelihood method, this is not the case. This is because the first-step estimator $\tilde{\lambda}_i$ is not a likelihood based estimator.

For θ, assuming that the underlying distribution is the Normal distribution (Bollerslev and Wooldridge (1992)), the conditional score for the composite likelihood function at time t is

$$g(y_t, \theta, \lambda) \propto \frac{\partial}{\partial \theta} \frac{1}{N} \sum_{i=1}^{N} \left(-\frac{1}{2} \log \sigma_{it}^2 - \frac{1}{2} \frac{\varepsilon_{it}^2}{\sigma_{it}^2} \right), \tag{2.24}$$

up to a constant. Equation (2.24) can be interpreted from a method of moments perspective, as well, by noting that,

$$\mathbb{E}[g(y_t, \theta_0, \lambda_0)] = 0,$$

which gives the second population moment condition. Finally, the estimators are implied by the sample moment conditions,

$$\frac{1}{T} \sum_{t=1}^{T} m_N(y_t, \tilde{\lambda}) = 0 \quad \text{and} \quad \frac{1}{T} \sum_{t=1}^{T} g(y_t, \tilde{\lambda}, \hat{\theta}) = 0.$$

2.3.4 The MacGyver Method

An alternative method is the MacGyver method by Engle (2009). This method is based on "blending" an already available collection of estimates of a parameter of interest to obtain a new estimate of this parameter. Obviously, the key assumption, again, is that θ is the same for all assets.

Let $\hat{\theta}_k$, $k = 1, ..., K$, be K different estimates of θ. These may be obtained by using different methods, models or data sets. For example, individual maximum likelihood estimates for each asset in the panel yields such a collection of $\hat{\theta}_k$. These estimates are then combined using a "blend function", $b(\cdot)$, to obtain the final estimate $\hat{\theta}_{MG}$:

$$\hat{\theta}_{MG} = b(\hat{\theta}_1, ..., \hat{\theta}_K).$$

Following Engle (2009), the three obvious blend functions are the mean, median and the mean of a trimmed set constructed by eliminating the highest and lowest 5% of the estimates. The latter two blending functions serve the purpose of discarding outliers, which could otherwise introduce bias.

For the case of GARCH panels, θ_k and λ_k can now conveniently be estimated simulta-

neously as the dimension of the parameter space is down to 3 rather than $N + 2$. Hence,

$$(\hat{\theta}_k, \hat{\lambda}_k) = \arg\max_{\theta,\lambda_k} \frac{1}{T} \sum_{t=1}^{T} \ell_{it}(\theta, \lambda_k), \qquad k = 1, ..., N.$$

Of course, the two-step estimation approach outlined previously is still valid and can be employed to generate the pool of estimates, as well.

2.3.5 Large Sample Theory for GARCH Panels

Asymptotic properties of both the quasi maximum likelihood method and method of moments, as well as the two-step estimators, have been analysed extensively in the literature. Relevant references include Newey and McFadden (1994), White (1994) and Hall (2005). The setting in this paper can be summarised as composite likelihood estimation in the presence of nuisance parameters in a large-T large-N asymptotic framework. This framework implies that as the sample size increases, the time-series and cross-section dimensions will be of similar magnitudes.[17] In addition, cross-section dependence of unknown form is allowed, as would be the case with financial and macro panels.

The theory presented below is simply an adaptation of the large sample theory developed by Engle, Shephard and Sheppard (2008) to the special case of composite likelihood functions based on univariate marginal densities. Their objective is to model *multivariate* volatility by using many composite likelihoods based on bivariate joint densities, rather than considering the joint density of all assets. Although this study focuses on panel estimation of *univariate* volatility, their results immediately specialise to the case at hand when bivariate joint densities are replaced by univariate marginal densities. In addition, although this chapter uses the two-step estimation approach based on the first-step method of moments estimator, their theory is general enough to also encompass the concentrated likelihood estimation approach.

The next assumption formally restates the underlying asymptotic framework.

Assumption 2.3.1 *As* $N, T \to \infty$, $N/T \to c$ *where* $0 < c < \infty$.

The following Assumptions are used to establish the consistency of $\hat{\theta}$.

[17] This does not necessarily mean that both dimensions should be of equal size. It is still possible to consider, for example, a (20×200) or (200×20) panel. However, for instance, a (4×200) panel would not belong to this asymptotic framework.

Assumption 2.3.2 *$\ell_{it}(\theta, \lambda_i)$ is continuously differentiable in $\lambda_i \in \Lambda_i$ $\forall i, t$ and θ.*

Assumption 2.3.3 *For all i, there exists a (pseudo) true $\lambda_{i0} \in \Lambda_i$ such that $\mathbb{E}[m_i(y_t, \lambda_i)] = 0$ is uniquely solved by λ_{i0}.*

Assumption 2.3.4 *There exists a (pseudo) true $\theta_0 \in \Theta$ where*

$$\arg\max_{\theta \in \Theta} \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \ell_{it}\left(\theta, \lambda_{i0}\right) \xrightarrow{p} \theta_0.$$

Assumption 2.3.5 $\max_{i \in \{1,\ldots,N\}} \left|(\tilde{\lambda}_i - \lambda_{i0})\right| = o_p(1)$.

Assumption 2.3.6 $(NT)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N} \sup_{\theta \in \Theta, \lambda_i \in \Lambda_i} |\partial \ell_{it}(\theta, \lambda_i)/\partial \lambda_i|$ *obeys a Weak Law of Large Numbers as $N, T \to \infty$.*

Assumption 2.3.2 is a standard regularity condition. Assumptions 2.3.3 and 2.3.4 ensure that the population moment condition uniquely identifies the (pseudo) true λ_{i0} and that when λ_{i0} is known, the composite likelihood function is consistent for the (pseudo) true θ_0. The last two Assumptions are technical conditions which ensure that $\hat{\theta}$ converges when λ_{i0} is unknown and has to be estimated as an incidental parameter. The theorem for the consistency of $\hat{\theta}$ follows.

Theorem 2.3.1 *Under Assumptions 2.3.1-2.3.6,*

$$\hat{\theta} \xrightarrow{p} \theta_0.$$

Remark 2.3.1 *Notice that the consistency result can easily be adapted to concentrated likelihood estimation by assuming (instead of Assumption 2.3.3) that for all θ there exists a (pseudo) true $\lambda_{i0}(\theta)$ such that $\lim_{T\to\infty} \mathbb{E}[T^{-1} \sum_{t=1}^{T} \partial \ell_{it}(\theta, \lambda_i)/\partial \lambda_i]$ is uniquely maximised at $\lambda_{i0}(\theta)$ for all i.*

For asymptotic normality, first define the following:

$$F_{iT} = -\frac{1}{T} \sum_{t=1}^{T} \frac{\partial^2 \ell_{it}(\theta, \lambda_i)}{\partial \theta \partial \lambda_i}, \qquad Z_{t,NT} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\partial \ell_{it}(\theta, \lambda_i)}{\partial \theta} - m(y_{it}, \lambda_i) F_{iT} \right],$$

$$D_{\theta\theta,iT} = \frac{1}{T} \sum_{t=1}^{T} \frac{\partial^2 \ell_{it}(\theta, \lambda_i)}{\partial \theta \partial \theta'}, \qquad D_{\theta\theta,NT} = \frac{1}{N} \sum_{i=1}^{N} D_{\theta\theta,iT} \quad \text{and} \quad m\left(y_{it}, \lambda_i\right) = y_{it}^2 - \lambda_i.$$

The following set of Assumptions ensures asymptotic normality of the composite likelihood estimator.

Assumption 2.3.7 *For all i, λ_i and θ belong to the interior of Λ_i and Θ, respectively.*

Assumption 2.3.8 *$\ell_{it}(\theta, \lambda_i)$ is twice and $m(y_t, \lambda_i)$ is once continuously differentiable $\forall i, t$.*

Assumption 2.3.9 *As $N, T \to \infty$, $T^{-1/2} \sum_{t=1}^{T} Z_{t,NT} \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\theta\theta})$, where $\mathcal{I}_{\theta\theta}$, the asymptotic covariance matrix, is positive definite and has diagonal elements bounded from above.*

Assumption 2.3.10 *As $N, T \to \infty$, $D_{\theta\theta,NT} \xrightarrow{p} \mathcal{D}_{\theta\theta}$ where $\mathcal{D}_{\theta\theta}$ is a positive definite, non-singular matrix.*

The next theorem establishes asymptotic normality.

Theorem 2.3.2 *If $\hat{\theta}$ is a consistent estimator for θ_0, then under Assumptions 2.3.1 and 2.3.7-2.3.10*

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{D}_{\theta\theta}^{-1} \mathcal{I}_{\theta\theta} \mathcal{D}_{\theta\theta}^{-1}).$$

Assumptions 2.3.7 and 2.3.8 are standard regularity conditions. Assumption 2.3.9 is probably the most important assumption of this chapter as it implicitly characterises the nature of cross-section dependence. Under cross-section independence, the convergence rate for such a term would generally be $\sqrt{NT}$. Instead, assumption 2.3.9 stipulates that the convergence rate is $\sqrt{T}$. In a way, this implies that, although the cross-sectional averages (normalised by N) are bounded, there is no convergence in the cross-section. This is clearly a worst-case scenario where dependence is allowed to be so large that there are no gains from the cross-sectional information as far as asymptotic normality is concerned. A milder version of this assumption, allowing for more contribution from the cross-sectional information, would be possible, but at the cost of putting a more specific structure on the cross-section dependence. This will be discussed again in Chapter 4, where cross-section dependence will have a more significant impact on the asymptotic results. Finally, Assumption 2.3.10 is simply a Law of Large Numbers. Clearly, more primitive conditions can be found. However, the sole aim of this Chapter is to introduce the idea of panel estimation of volatility by adopting the analysis of Engle, Shephard and Sheppard (2008). Providing refinements

on their Assumptions is beyond the scope of this study, which could nevertheless be an interesting future research subject.

By Assumption 2.3.10, $D^{-1}_{\theta\theta,NT}$ can be used to estimate $\mathcal{D}^{-1}_{\theta\theta}$ in order to obtain the sandwich estimator of the asymptotic covariance matrix. Also, an estimator for

$$p \lim_{N,T\to\infty} Var(T^{-1/2} \sum_{t=1}^{T} Z_{t,NT})$$

is necessary in order to estimate $\mathcal{I}_{\theta\theta}$. When evaluated at the true parameter values, this will be equal to $p\lim_{N,T\to\infty} T^{-1} \sum_{t=1}^{T} Var(Z_{t,NT})$ since $Cov(Z_{s,NT}, Z_{t,NT}) = 0$ for $s \neq t$ by a Martingale Difference Sequence (MDS) argument. However, this property does not necessarily hold at other parameter values. Hence a "Heteroskedasticity and Autocorrelation Consistent" (HAC) covariance estimator has to be used. The simulation and empirical exercises in this study are based on the Newey and West (1987) HAC estimator.[18]

2.3.6 Discussion

Both CL and MG are based on the idea of pooling information. However, they pool information at different stages of estimation. MG pools information by using "blend" functions to combine available estimates of parameters of interest. CL, on the other hand, pools information to form a single composite-likelihood function that characterises all assets in the asset panel. Therefore, it can be argued that CL fully exploits the information in both dimensions, while MG is still largely limited to time-series information.

A further difference between the two methods is related to their theoretical properties. Although MG provides a straight-forward method for estimation using GARCH panels, a major drawback is that there is no asymptotic theory for this method. For some blend functions, such as median, developing a Central Limit Theorem is a non-trivial task. CL, on the other hand, has a complete large sample theory which allows for statistical inference, as established by Engle, Shephard and Sheppard (2008). This is an important advantage of CL over MG.

Another issue arises when the sample size is not large enough. In such samples, due to the small sample (or small-T) bias, optimisation has a tendency to fail and simply return the

[18] See also Andrews (1991), Andrews and Monahan (1992), Newey and West (1994) and den Haan and Levin (1996)

initial values used in optimisation as the parameter estimates.[19] A quick and inexpensive solution is to exclude such cases from the calculation of the final estimate. Furthermore, again when T is not large enough, $\hat{\alpha}$ will generally be biased towards zero yielding many cases where $\hat{\alpha}$ is arbitrarily close to zero. As explained previously, β is not identified when $\alpha = 0$. Consequently, this study also ignores sets of estimates where $\hat{\alpha}$ is less than 0.0025.[20] This particular choice of the cut-off value and the elimination of non-converging cases reflect the ad-hoc nature of MG. However, it must be underlined that the aim of MG is not to obtain a set of very good estimates, but rather to find a blend function that yields a good estimate out of a pool of a large number of estimates. Of course, both of these practical issues will lose their gravity as the sample size increases. The more serious problem is the absence of a large sample theory.

2.4 Simulation Analysis

This section provides an analysis of the small sample characteristics of the CL and MG methods. Two types of simulations are considered. In the first one T is fixed to $2,000$ for all assets, while the second one considers varying samples sizes. To analyse the effect of estimation of the individual-specific parameters, a further simulation analysis is considered, where the true values of the individual-specific parameters are used to estimate the parameters of interest, eliminating any possible incidental parameter issues. Results show that this has serious effects on the finite sample performance of CL.

2.4.1 The Data Generating Process

The asset panel is generated using the variance targeting specification for GARCH(1,1), where

$$y_{it} = \mu_{it} + \varepsilon_{it}, \quad \mu_{it} = \mathbb{E}[y_{it}|\mathcal{F}_{i,t-1}] = 0, \quad \varepsilon_{it}|\mathcal{F}_{t-1} \sim N\left(0, \sigma_{it}^2\right),$$
$$\sigma_{it}^2 = \lambda_i(1-\alpha-\beta) + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2.$$

[19] The optimisation procedure used for this study starts at pre-specified starting values and searches for an optimum. If optimisation fails to find an optimum, then the starting values are given as the parameter estimates.

[20] In a simulation analysis not presented here, among $2,500$ replications of a GARCH process with $T = 100$, estimation failed to converge in more than $1,400$ replications, while around 100 replications produced $\hat{\alpha} = 0$.

To avoid any potential computational bias in the initial portion of the sample, 500 extra observations are generated which are then discarded. Assuming that a range between 15% and 80% will be representative of annual volatility for most stock returns, λ_i are generated using $\lambda_i \sim U(\underline{\sigma}^2, \overline{\sigma}^2)$ where U is the uniform distribution, $\underline{\sigma} = \sqrt{(15\%)^2/252}$ and $\overline{\sigma} = \sqrt{(80\%)^2/252}$.[21] Cross-section dependence is generated by a single-factor model where,

$$\varepsilon_{it} = \sigma_{it}\eta_{it}, \qquad \eta_{it} = \rho_i u_t + \sqrt{1-\rho_i^2}\tau_{it},$$
$$u_t \overset{iid}{\sim} \mathcal{N}(0,1) \quad \text{and} \quad \tau_{it} \overset{iid}{\sim} \mathcal{N}(0,1).$$

This implies that, η_{it} is zero-mean for all i, t and

$$Cov(\eta_{is}, \eta_{jt}) = \begin{cases} \rho_i\rho_j & \forall i \neq j \text{ if } s = t \\ 0 & \forall t \neq s \text{ and any } i, j \end{cases}.$$

Therefore, only contemporaneous correlation is allowed. The correlation parameters, ρ_i, are drawn randomly from a Uniform distribution such that $\rho_i \sim U(0.5, 0.9)$, which implies that the lowest and highest possible correlation between any two assets will be equal to 0.25 and 0.81, respectively.

Finally, as in Engle, Shephard and Sheppard (2008), α and β are chosen from three different alternatives which cover a wide range of parameter values,

$$(\alpha, \beta) \in \left\{\theta^{(1)}, \theta^{(2)}, \theta^{(3)}\right\} = \{(0.02, 0.97), (0.05, 0.93), (0.10, 0.80)\}. \tag{2.25}$$

2.4.2 Simulation Results

Fixed Sample Size

In this section, T is fixed to 2,000 while $N \in \{3, 10, 50, 100\}$. All results are based on $2,500$ replications for each parameter selection in (2.25). Results are presented in Tables 2.1 to 2.3. Estimators by the available methods are compared on the basis of their average values, Monte Carlo standard deviations (MCSD) and root mean squared errors. In addition, to investigate whether the theoretical large sample properties of CL hold in finite samples, average asymptotic standard deviations (ASD) are also provided. These are defined as

[21] This is a standard method for converting annual volatility to daily volatility.

follows:

$$\begin{aligned} MCSD &: \bar{\sigma}_{\hat{\kappa}} = \sqrt{\frac{1}{Z}\sum_{z=1}^{Z}\left(\hat{\kappa}_z - \frac{1}{Z}\sum_{z=1}^{Z}\hat{\kappa}_z\right)^2}, \\ ASD &: \hat{\sigma}_{\hat{\kappa}} = \frac{1}{Z}\sqrt{\sum_{z=1}^{Z}\hat{\sigma}^2_{\hat{\kappa}z}}, \\ RMSE &: \mathcal{R}_{\hat{\kappa}} = \sqrt{\frac{1}{Z}\left(\hat{\kappa}_z - \kappa_0\right)^2}, \end{aligned}$$

where Z is the number of replications, $\hat{\alpha}_z$ and $\hat{\beta}_z$ are the estimates for replication z, $z = 1, ..., Z$, $\hat{\kappa}_z \in \{\hat{\alpha}_z, \hat{\beta}_z\}$ and $\kappa_0 \in \{\alpha_0, \beta_0\}$. $\hat{\sigma}^2_{\hat{\kappa},z}$ is the estimated asymptotic variance for $\hat{\kappa}_z$, obtained by using Theorem 2.3.5. ASD serves as an average measure of the asymptotic standard deviation across all replications. As a further measure to evaluate the finite sample performance of CL, sample confidence interval coverage rates (CI) are also provided. This statistic gives the percentage of replications for which the sample confidence intervals, constructed by using the asymptotic theory in Theorem 2.3.5, contain the true parameter values. For $\hat{\kappa}_z$, this interval is given by

$$[\hat{\kappa}_z \pm \hat{\sigma}_{\hat{\kappa},z}\mathcal{Z}_{\alpha/2}] \quad \text{where} \quad \Pr[-\mathcal{Z}_{\alpha/2} < Z < \mathcal{Z}_{\alpha/2}] = 1 - \alpha \quad \text{and} \quad Z \sim \mathcal{N}(0,1).$$

All results are calculated for 95% confidence intervals. Note that, as explained previously in Section 2.3.4, when calculating the MG estimates, cases where optimisers do not converge and where $\hat{\alpha} \leq 0.0025$ are discarded.

Table 2.1 reveals that all estimators perform very well and are subject to some negligible bias only, which might be due to computational reasons.[22] Only when $\theta = (0.02, 0.97)$ is MG-Mean subject to some more bias than the other MG methods but this is still not substantial. In line with Engle's (2009) findings, median generally leads to lowest average bias within the group of MG estimators, although the differences in performance are negligible. In any case, estimation performance is so satisfactory that there is not much point in comparing the different methods on that front.

Table 2.2 gives the MCSD statistics. Clearly, across all methods, the average sample standard deviation decreases as N increases. Importantly, the speed of decline of MCSD falls

[22] Note that MG-Mean and MG-Trimmed have identical results when the cross-section size is equal to 3 and 10, This is because the 5% trimmed sample is simply the same as the original sample in these cases.

	CL		MG-Mean		MG-Median		MG-Trimmed	
			$\alpha = 0.02, \quad \beta = 0.97$					
N	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
3	.0196	.9659	.0207	.9503	.0201	.9617	.0207	.9503
10	.0194	.9682	.0207	.9489	.0198	.9661	.0207	.9489
50	.0193	.9685	.0207	.9487	.0197	.9666	.0202	.9610
100	.0193	.9685	.0207	.9483	.0197	.9666	.0201	.9619
			$\alpha = 0.05, \quad \beta = 0.93$					
N	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
3	.0490	.9260	.0491	.9268	.0489	.9278	.0491	.9268
10	.0490	.9295	.0491	.9267	.0489	.9281	.0491	.9267
50	.0490	.9297	.0491	.9265	.0489	.9282	.0494	.9279
100	.0489	.9298	.0491	.9265	.0488	.9283	.0490	.9274
			$\alpha = 0.10, \quad \beta = 0.80$					
N	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
3	.0991	.7974	.0997	.7930	.0994	.7957	.0997	.7930
10	.0990	.7986	.0997	.7929	.0992	.7962	.0997	.7929
50	.0990	.7989	.0998	.7926	.0992	.7966	.1003	.7959
100	.0989	.7990	.0997	.7927	.0991	.7967	.0995	.7944

Table 2.1: Average parameter estimates across replications. T=2,000. Based on 2,500 replications.

	CL		MG-Mean		MG-Median		MG-Trimmed	
			$\alpha = 0.02, \quad \beta = 0.97$					
N	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
3	.0043	.0350	.0052	.0624	.0052	.0486	.0052	.0624
10	.0028	.0048	.0033	.0370	.0031	.0062	.0033	.0370
50	.0021	.0035	.0022	.0221	.0022	.0040	.0022	.0132
100	.0020	.0033	.0020	.0204	.0020	.0037	.0020	.0107
			$\alpha = 0.05, \quad \beta = 0.93$					
N	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
3	.0067	.0103	.0068	.0113	.0076	.0123	.0068	.0113
10	.0044	.0067	.0045	.0073	.0050	.0076	.0045	.0073
50	.0034	.0050	.0034	.0055	.0037	.0055	.0034	.0052
100	.0032	.0047	.0032	.0050	.0034	.0051	.0032	.0050
			$\alpha = 0.10, \quad \beta = 0.80$					
N	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
3	.0122	.0275	.0124	.0295	.0141	.0314	.0124	.0295
10	.0082	.0179	.0083	.0193	.0091	.0203	.0083	.0193
50	.0062	.0133	.0062	.0141	.0066	.0146	.0063	.0138
100	.0058	.0125	.0058	.0131	.0062	.0135	.0059	.0130

Table 2.2: Monte Carlo (sample) standard deviations. T=2,000. Based on 2,500 replications.

as N increases. This suggests that the underlying data generating process indeed implies less than $\sqrt{NT}$-convergence. Comparing the two methods, CL generally leads to lower MCSD, but the difference with the MG method vanishes as N increases. CL's superiority in MCSD is distinctive for $\hat{\beta}$, and again, when N is small. This is not surprising. For CL, a larger N means a more informative information set, while for MG it simply means a larger pool of estimates of the same kind. Within the MG estimators, there is no clear preference. Although MG-Median performs best for $\theta = (0.02, 0.97)$, it leads to higher MCSD for the remaining parameter selections. The case of $\theta = (0.02, 0.97)$ is interesting in the sense that the average standard deviations for $\hat{\beta}$ are quite large for MG-Mean and MG-Trimmed, while MG-Median delivers significant gains in terms of variance. Such a sharp contrast is not observed for the other parameter combinations. All in all, it can be concluded that as far as sample standard deviation performance is concerned, all methods perform well, while CL gives slightly better results.

Lastly, Table 2.3 presents the ASD, RMSE and CI statistics. MCSDs for CL are also reproduced for quick reference. For all three parameters sets, the MCSD and ASD statistics for both $\hat{\alpha}$ and $\hat{\beta}$ are generally very close to each other (one notable exception is observed for $\hat{\beta}$ when $\theta = (0.02, 0.97)$ and $N = 3$). This implies that simulation results are in line with the theory developed in Section 2.3.5, as sample results are, on average, close to those implied by theory. Root mean squared errors are, in some cases, slightly different from the average sample standard deviations owing to the small bias observed in average parameter estimates. The confidence interval coverage rates are satisfactory for $\hat{\beta}$, ranging between 93.76% and 95.80%. However, in most cases the significance test for $\hat{\alpha}$ has a tendency to over-reject. Interestingly, the coverage rates are reasonable for $\theta = (0.10, 0.80)$. This might be due to the true parameter values being away from the boundaries of the parameter set.

In light of these results several conclusions can be reached. Firstly, both CL and MG stand as valid estimation methods. Secondly, simulation results establish that CL is marginally better than MG, both in terms of average bias and sample standard deviation performance. However, when the asymptotic theory for CL is concerned, Table 2.3 confirms that in samples of reasonable size, finite sample results are in accordance with results provided by the large sample theory. This is an important advantage. Given that MG lacks a complete asymptotic theory, these results suggest that CL should be preferred to MG.

$\alpha = 0.02, \quad \beta = 0.97$								
N	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
3	.0043	.0350	.0044	.0098	.0043	.0352	.9204	.9400
10	.0028	.0048	.0027	.0051	.0028	.0052	.9144	.9580
50	.0021	.0035	.0020	.0037	.0022	.0038	.8996	.9560
100	.0020	.0033	.0019	.0035	.0021	.0037	.8876	.9540
$\alpha = 0.05, \quad \beta = 0.93$								
N	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
3	.0067	.0103	.0064	.0106	.0068	.0104	.9176	.9376
10	.0044	.0067	.0042	.0067	.0045	.0067	.9116	.9428
50	.0034	.0050	.0032	.0050	.0035	.0050	.8988	.9420
100	.0032	.0047	.0030	.0047	.0034	.0047	.8936	.9440
$\alpha = 0.10, \quad \beta = 0.80$								
N	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
3	.0122	.0275	.0120	.0279	.0122	.0276	.9360	.9444
10	.0082	.0179	.0080	.0180	.0082	.0180	.9304	.9424
50	.0062	.0133	.0059	.0132	.0062	.0133	.9256	.9432
100	.0058	.0125	.0057	.0126	.0059	.0125	.9236	.9492

Table 2.3: MCSD, ASD, RMSE and sample confidence interval (CI) statistics. T=2,000. Based on 2,500 replications. All results are for the composite likelihood method.

However, it must also be noted that the confidence interval coverage rates are not accurate for $\hat{\alpha}$. Therefore, inference results for $\hat{\alpha}$ should be treated carefully.

It must be underlined that this exercise uses panels with $2,000$ time-series observations, which is large enough for the asymptotic results to work well in finite samples. A more interesting analysis is that of smaller sample sizes where estimation would be expected to be more problematic. This is considered next.

Varying Sample Sizes

This section analyses the implications of varying both the number of assets and observations per asset, by using $N \in \{10, 50, 100\}$ and $T \in \{100, 250, 500, 1000, 2000\}$. Again, all results are based on $2,500$ replications. For space considerations, the analysis is conducted for $\theta = (0.05, 0.93)'$ only. Results are presented in Tables 2.4-2.6.

Clearly, all methods suffer from bias as T decreases (see Table 2.4). For CL, the bias is most acute when $T \leq 250$. Within the MG estimators, there is some variation in performance across different blend functions. For example, MG-Median has a better performance in estimation of β, and the small sample bias reaches reasonable levels around $T = 500$. On the other hand, MG-Trimmed generally performs best in estimating α. Nevertheless, the

	CL		MG-Mean		MG-Median		MG-Trimmed	
				$N = 100$				
T	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
$2,000$	.0489	.9298	.0491	.9265	.0488	.9283	.0490	.9274
$1,000$	.0479	.9292	.0490	.9172	.0480	.9259	.0485	.9231
500	.0462	.9261	.0516	.8808	.0482	.9176	.0499	.9014
250	.0417	.9175	.0601	.8098	.0518	.8920	.0561	.8362
100	.0185	.8821	.0880	.6689	.0703	.7757	.0804	.6868
				$N = 50$				
T	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
$2,000$	.0490	.9297	.0491	.9265	.0489	.9282	.0494	.9279
$1,000$	.0479	.9291	.0490	.9172	.0480	.9258	.0490	.9238
500	.0463	.9259	.0516	.8810	.0483	.9174	.0495	.8978
250	.0416	.9172	.0601	.8101	.0518	.8908	.0556	.8333
100	.0190	.8821	.0882	.6691	.0709	.7668	.0808	.6774
				$N = 10$				
T	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
$2,000$	.0490	.9295	.0491	.9267	.0489	.9281	.0491	.9267
$1,000$	.0480	.9286	.0490	.9171	.0481	.9253	.0490	.9171
500	.0463	.9248	.0517	.8808	.0486	.9153	.0517	.8808
250	.0419	.9122	.0599	.8103	.0530	.8782	.0599	.8103
100	.0222	.8433	.0879	.6741	.0770	.7294	.0879	.6741

Table 2.4: Average parameter estimates across replications. $\alpha = 0.05$, $\beta = 0.93$, based on 2,500 replications.

	CL		MG-Mean		MG-Median		MG-Trimmed	
				$N = 100$				
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
2,000	.0032	.0047	.0032	.0050	.0034	.0051	.0032	.0050
1,000	.0044	.0066	.0044	.0100	.0046	.0073	.0045	.0077
500	.0063	.0098	.0066	.0309	.0066	.0118	.0065	.0240
250	.0094	.0157	.0103	.0624	.0095	.0297	.0099	.0623
100	.0182	.0692	.0184	.1055	.0180	.1168	.0183	.1157
				$N = 50$				
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
2,000	.0034	.0050	.0034	.0055	.0037	.0055	.0034	.0052
1,000	.0047	.0071	.0047	.0115	.0050	.0079	.0048	.0081
500	.0067	.0105	.0072	.0342	.0073	.0130	.0072	.0283
250	.0100	.0174	.0114	.0687	.0106	.0353	.0106	.0689
100	.0197	.0759	.0215	.1151	.0211	.1362	.0210	.1244
				$N = 10$				
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
2,000	.0044	.0067	.0045	.0073	.0050	.0076	.0045	.0073
1,000	.0062	.0096	.0065	.0207	.0071	.0114	.0065	.0207
500	.0091	.0148	.0105	.0529	.0104	.0216	.0105	.0529
250	.0142	.0519	.0176	.0974	.0166	.0752	.0176	.0974
100	.0266	.1443	.0392	.1706	.0395	.2055	.0392	.1706

Table 2.5: Monte Carlo (sample) standard deviations. $\alpha = 0.05$, $\beta = 0.93$, based on 2,500 replications.

small sample bias is also large for MG estimators when $T \leq 250$. Interestingly, when the sample size is large, in some cases MG performs better than CL in estimating α. Another observation is that, as N increases, there is a slight tendency for the bias of α to increase and for the bias of β to decrease. Nevertheless, the change is not significant and only observed when T is very small. In general, CL delivers the best results in estimating β while MG-Trimmed attains good performance in estimating α.

Sample standard deviations are presented in Table 2.5. In general, when $\hat{\alpha}$ is concerned, CL estimates are slightly less dispersed. This becomes more pronounced when both N and T are small. CL's superiority is more significant for $\hat{\beta}$. The difference between the two methods is clear when $T \leq 250$ where CL delivers distinctively better results. Considering the MG estimators, MG-Median almost always delivers the smallest dispersion. Interestingly, for $T = 100$, it is beaten by the other blend functions. This suggests that the location of the median of the pool varies much more than the mean. As expected, the sample standard deviation uniformly decreases for all methods and parameters as N increases.

$N = 100$								
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
2,000	.0032	.0047	.0030	.0047	.0034	.0047	.894	.944
1,000	.0044	.0066	.0042	.0068	.0049	.0067	.861	.952
500	.0063	.0098	.0060	.0103	.0074	.0105	.804	.954
250	.0094	.0157	.0088	.0178	.0125	.0201	.686	.946
100	.0182	.0692	.0222	.2750	.0364	.0841	.426	.823
$N = 50$								
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
2,000	.0034	.0050	.0032	.0050	.0035	.0050	.899	.942
1,000	.0047	.0071	.0045	.0073	.0051	.0071	.867	.953
500	.0067	.0105	.0063	.0110	.0077	.0113	.815	.954
250	.0100	.0174	.0094	.0193	.0130	.0216	.704	.941
100	.0197	.0759	.0548	.4143	.0368	.0914	.455	.832
$N = 10$								
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
2,000	.0044	.0067	.0042	.0067	.0045	.0067	.912	.943
1,000	.0062	.0096	.0060	.0097	.0065	.0097	.897	.946
500	.0091	.0148	.0085	.0157	.0098	.0157	.859	.946
250	.0142	.0519	.0135	.0548	.0163	.0548	.782	.932
100	.0266	.1443	.1459	.1684	.0385	.1684	.725	.838

Table 2.6: MCSD, ASD, RMSE and sample confidence interval (CI) statistics. $\alpha = 0.05$, $\beta = 0.93$, based on 2,500 replications. All results are for the composite likelihood method.

Table 2.6, reveals that the large sample theory developed in Section 2.3.5, on average, delivers reasonably accurate estimators of the sample standard deviation when $T \geq 500$. However, the accuracy deteriorates quickly as T decreases below 500, as reflected by the discrepancy between MCSD and ASD. The sample confidence interval coverage rates are very interesting. For $\hat{\beta}$, except for $T = 100$ the results are excellent as the coverage rates are very close to 95%. However, for $\hat{\alpha}$, the picture is not as optimistic. Only when T is really large are the coverage rates around 90%. Otherwise, the coverage rates suggest that the significance tests tend to over-reject. However, this does not necessarily imply that there is a serious problem in the estimation of asymptotic standard errors. Indeed, slight inaccuracies in standard error estimation can possibly lead to significant changes in the test-statistics. In addition, the discrepancy is likely to be due to inaccurate estimation of the incidental parameters. This latter possibility is investigated in the next section.

To consider the small sample properties of the estimation methods under scrutiny from a different angle, sample distribution graphs for estimators are presented in Figures 2.1 and 2.2. Notice that when $N = 10$, MG-Mean and MG-Trimmed are identical. Hence, both have

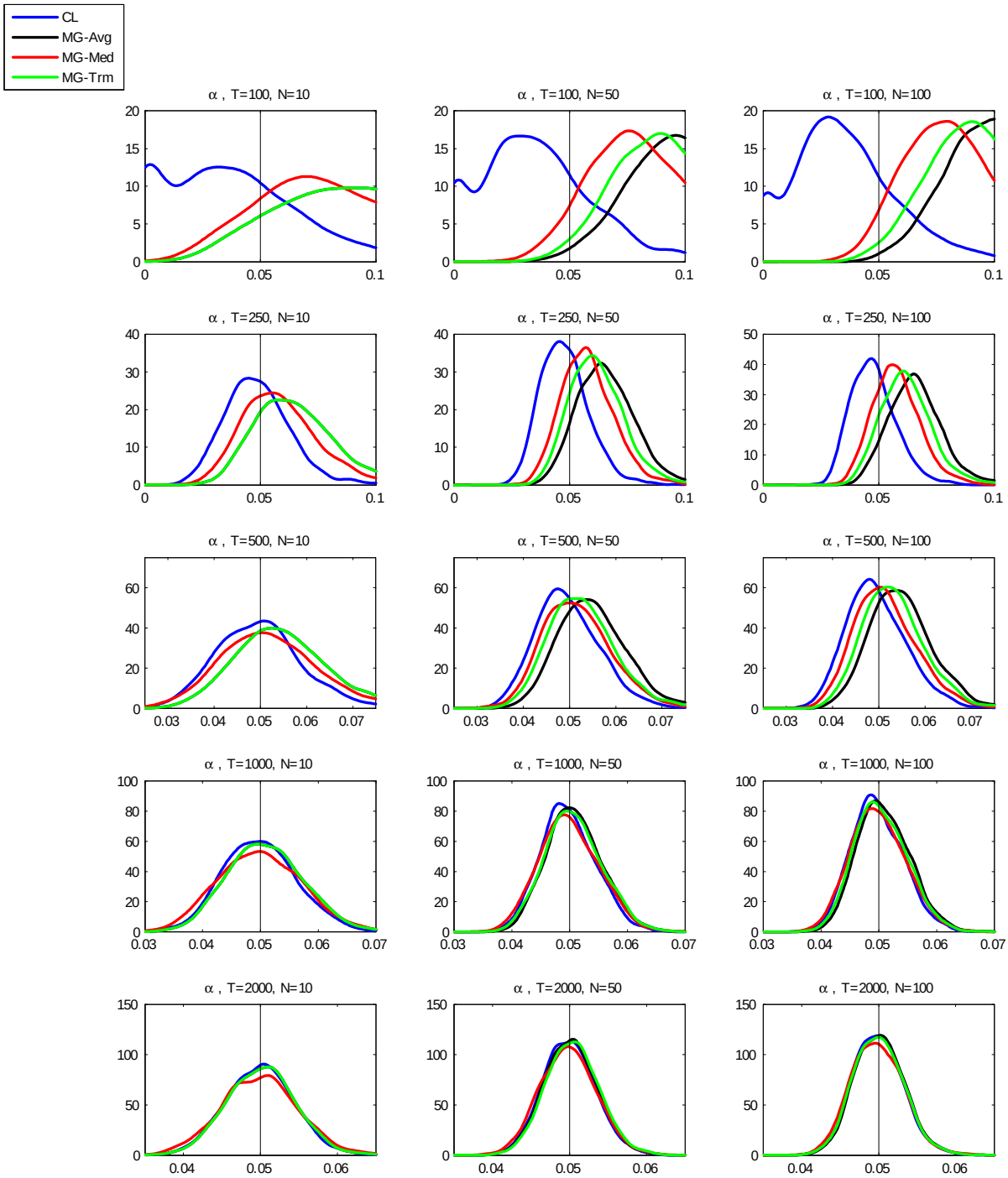

Figure 2.1: Sample distribution graphs for $\hat{\alpha}$. $\alpha = 0.05, \beta = 0.93$, based on 2,500 replications. MG-Avg, MG-Med and MG-Trm stand for MG-Mean, MG-Median and MG-Trimmed, respectively. The vertical line is drawn at $\alpha = 0.05$.

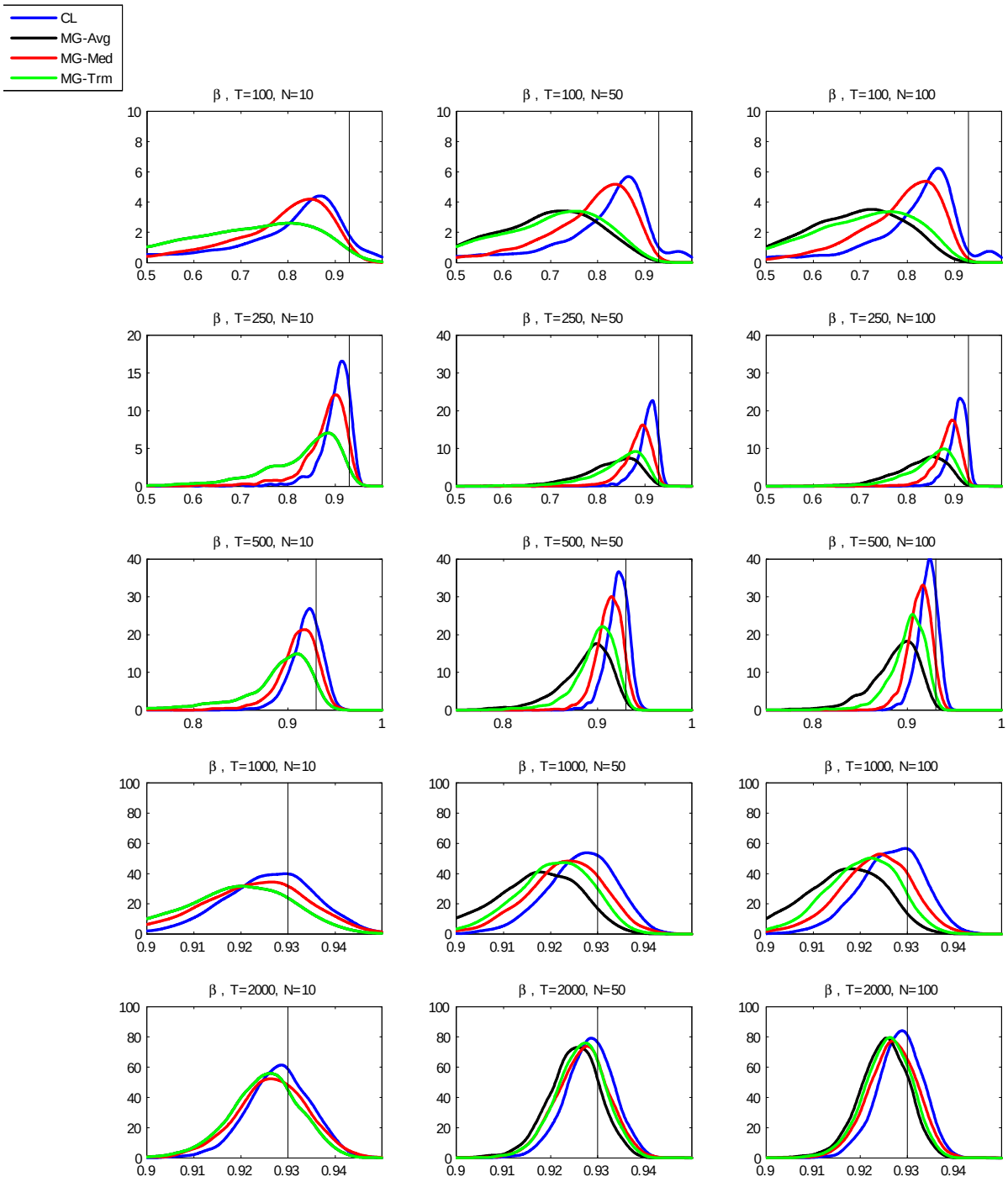

Figure 2.2: Sample distribution graphs for $\hat{\beta}$. $\alpha = 0.05, \beta = 0.93$, based on 2,500 replications. MG-Avg, MG-Med and MG-Trm stand for MG-Mean, MG-Median and MG-Trimmed, respectively. The vertical line is drawn at $\beta = 0.93$.

	CL		MG-Mean		MG-Median		MG-Trimmed	
				$N = 100$				
T	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
2,000	.0502	.9296	.0504	.9272	.0501	.9287	.0503	.9279
1,000	.0503	.9292	.0514	.9212	.0503	.9272	.0509	.9250
500	.0509	.9282	.0564	.8956	.0520	.9221	.0540	.9118
250	.0518	.9267	.0702	.8407	.0574	.9080	.0640	.8677
100	.0551	.9216	.1096	.7365	.0768	.8488	.0965	.7594
				$N = 50$				
T	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
2,000	.0502	.9295	.0505	.9272	.0502	.9286	.0507	.9284
1,000	.0504	.9292	.0514	.9211	.0504	.9270	.0514	.9258
500	.0509	.9281	.0564	.8956	.0521	.9219	.0545	.9115
250	.0518	.9266	.0701	.8408	.0574	.9076	.0631	.8618
100	.0553	.9204	.1096	.7364	.0773	.8457	.0952	.7551
				$N = 10$				
T	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
2,000	.0503	.9294	.0505	.9273	.0502	.9285	.0505	.9273
1,000	.0505	.9288	.0514	.9211	.0504	.9268	.0514	.9211
500	.0512	.9272	.0566	.8954	.0526	.9204	.0566	.8954
250	.0522	.9248	.0700	.8409	.0589	.9005	.0700	.8409
100	.0587	.9038	.1097	.7378	.0827	.8215	.1097	.7378

Table 2.7: Average parameter estimates across replications. $\alpha = 0.05$, $\beta = 0.93$, based on 2,500 replications. All estimations are conducted by using the true nuisance parameter values. In other words, nuisance parameter estimation has been by-passed and only the parameters of interest, α and β are estimated using the true values of the nuisance parameters.

identical sample distributions. Considering estimation of α first, when $T \geq 500$, all methods have more or less identical sample distributions with their modes around $\alpha = 0.05$. Another immediate observation is that CL performs poorly when $T = 100$. The large proportion of cases where $\hat{\alpha}$ is close to zero indicate that there are many cases where β cannot be identified. This is not observed for MG methods because, as explained previously, all cases where $\hat{\alpha} \leq 0.0025$ are eliminated from the pool of estimates. Still, even then the sample distributions for the MG methods do not look satisfactory, either. For all methods, a substantial improvement is observed as soon as T increases to 250. In accordance with the previous observations, sample distributions tend to become less dispersed as N increases. In addition, as expected, their modes move towards 0.05 as T increases. In general, it can be deduced that CL has a tendency to underestimate α, whereas MG is more likely to overestimate.

Sample distributions for $\hat{\beta}$ lead to the similar observations. One difference is that sample

distributions more or less overlap when $T = 2000$, but not for smaller T. In fact, as far as the distance of the mode to 0.93 is concerned, there is a clear ranking where CL is followed by MG-Median, MG-Trimmed and MG-Mean. Finally, for all methods, there is a tendency to underestimate β.

The general message of the simulation results is that CL generally performs well in samples of around at least 500 observations. Therefore, the sample can be considered "large" when $T \geq 500$. In addition, the analysis also confirms that bias is directly related to T and not N. As the sample size decreases, the performances of all methods deteriorate. Interestingly, for $\hat{\beta}$ there is a good asymptotic coverage rate even when T is as low as 250, whereas for $\hat{\alpha}$ the coverage rates are small except when T is around $1,000 - 2,000$.

Even when T is small, due to the panel structure the CL estimators for $\theta = (\alpha, \beta)$ are still based on a large number of observations. This would suggest that $\hat{\theta}$ should still have good properties when T is small. However, crucially, even in a panel structure $(\lambda_1, ..., \lambda_N)$ are still estimated using time-series information only. The inaccuracy in the estimation of these parameters, when T is small, can possibly feed into the estimation of θ. In the statistics and panel data literature, this is known as the incidental parameter issue (Neyman and Scott (1948)). This possibility is investigated next.

2.4.3 Nuisance Parameters and the Incidental Parameter Issue

As outlined in Section 2.3.3, the GARCH panel model is estimated in two steps. Although the common parameter estimator uses both time-series and cross-section information, the first-step estimator of $(\lambda_1, ..., \lambda_N)$ still relies on time-series information only. It is highly possible that this structure, characterised by some common parameter and many individual specific parameters, suffers from the incidental parameter bias. The mechanism that generates bias can be summarised as follows. When the time-series dimension is not large enough, the first-step estimator, $(\tilde{\lambda}_1, ..., \tilde{\lambda}_N)$, will be subject to small-sample bias. Crucially, the second-step estimator is based on the first-step estimator. Therefore, the small sample bias accumulated by $(\tilde{\lambda}_1, ..., \tilde{\lambda}_N)$ will contaminate $\hat{\theta}$, which will also be subject to bias, independent of the fact that $\hat{\theta}$ utilises a larger source of information.[23] This issue

[23] In fact, in the standard case considered in the literature, the asymptotic ratio of N to T is central in determining whether this bias mechanism will be present or not. This will be discussed in more detail in Chapter 3.

	CL		MG-Mean		MG-Median		MG-Trimmed	
				$N = 100$				
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
2,000	.0030	.0047	.0031	.0049	.0032	.0050	.0031	.0049
1,000	.0042	.0065	.0044	.0086	.0045	.0072	.0044	.0073
500	.0060	.0095	.0071	.0232	.0065	.0110	.0066	.0169
250	.0088	.0138	.0139	.0486	.0101	.0207	.0123	.0455
100	.0157	.0323	.0311	.0897	.0222	.0772	.0293	.0961
				$N = 50$				
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
2,000	.0032	.0050	.0033	.0053	.0035	.0055	.0033	.0052
1,000	.0045	.0070	.0047	.0099	.0049	.0079	.0047	.0078
500	.0065	.0102	.0078	.0265	.0071	.0124	.0075	.0197
250	.0093	.0148	.0151	.0541	.0110	.0238	.0133	.0526
100	.0169	.0402	.0334	.0962	.0250	.0878	.0305	.1037
				$N = 10$				
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$
2,000	.0042	.0066	.0043	.0072	.0047	.0075	.0043	.0072
1,000	.0060	.0094	.0065	.0168	.0069	.0111	.0065	.0168
500	.0088	.0142	.0119	.0434	.0105	.0187	.0119	.0434
250	.0130	.0227	.0229	.0813	.0176	.0482	.0229	.0813
100	.0292	.0980	.0481	.1325	.0404	.1438	.0481	.1325

Table 2.8: Monte Carlo (sample) standard deviations. $\alpha = 0.05$, $\beta = 0.93$, based on 2,500 replications. All estimations are conducted by using the true nuisance parameter values. In other words, nuisance parameter estimation has been by-passed and only the parameters of interest, α and β are estimated using the true values of the nuisance parameters.

will be considered in detail in the next chapter. Here, a simulation exercise is considered in order to demonstrate that the GARCH panel model is also subject to the incidental parameter issue.[24] In order to by-pass the first-step estimation (and any small-sample bias it carries), estimation is based on the true values, $(\lambda_{10}, ..., \lambda_{N0})$. MG estimators are also based on these parameter values, in order to investigate the benefits of the panel approach over the time-series approach free from any first-step estimation related problem.

Results are presented in Tables 2.7-2.9 and Figures 2.3-2.4. Clearly, the small sample performance of CL improves greatly, even when T is as small as 250. In addition, there is still some bias when $T = 100$, but the situation is much better compared to the previous simulation results. For MG, the picture is different. Estimation of β gives slightly better

[24] Notice that, traditionally, the incidental parameter issue is a likelihood-related problem, where the "first-step" estimator is actually a concentrated likelihood estimator, as in (2.21) and (2.22). Here, the first-step estimator is the non-parametric method of moments estimator. As is generally known, non-parametric estimators converge slowly. For this reason, this fast estimation method might be causing a larger small-sample bias.

$N = 100$								
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
$2,000$	.0030	.0047	.0030	.0046	.0030	.0047	.946	.945
$1,000$	.0042	.0065	.0042	.0066	.0042	.0066	.938	.945
500	.0060	.0095	.0060	.0098	.0061	.0097	.935	.947
250	.0088	.0138	.0088	.0153	.0090	.0142	.923	.941
100	.0157	.0323	.0162	.0350	.0165	.0334	.936	.938
$N = 50$								
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
$2,000$	.0032	.0050	.0032	.0049	.0032	.0050	.942	.939
$1,000$	.0045	.0070	.0045	.0070	.0045	.0070	.944	.949
500	.0065	.0102	.0064	.0104	.0065	.0104	.936	.944
250	.0093	.0148	.0094	.0165	.0094	.0152	.926	.948
100	.0169	.0402	.0176	.0404	.0178	.0413	.930	.934
$N = 10$								
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\hat{\sigma}_{\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}}$	$\mathcal{R}_{\hat{\alpha}}$	$\mathcal{R}_{\hat{\beta}}$	$CI_{\hat{\alpha}}$	$CI_{\hat{\beta}}$
$2,000$	.0042	.0066	.0042	.0066	.0042	.0066	.945	.944
$1,000$	.0060	.0094	.0060	.0096	.0060	.0095	.939	.943
500	.0088	.0142	.0088	.0147	.0089	.0145	.935	.934
250	.0130	.0227	.0138	.0265	.0132	.0233	.920	.926
100	.0292	.0980	.0371	.1033	.0305	.1015	.908	.921

Table 2.9: MCSD, ASD, RMSE and sample confidence interval (CI) statistics. $\alpha = 0.05$, $\beta = 0.93$, based on 2,500 replications. All results are for the Composite Likelihood Method. All estimations are conducted by using the true nuisance parameter values. In other words, nuisance parameter estimation has been by-passed and only the parameters of interest, α and β are estimated using the true values of the nuisance parameters.

results, but now the average bias of $\hat{\alpha}$ is larger than before when T is small. In addition, CL is uniformly superior in terms of average bias. This is a very important result in favour of CL. This clearly demonstrates that exploiting the information across the whole panel is superior to using time-series information only. In addition the sample standard deviation results are also generally in favour of CL.

Sample distributions of estimators in Figures 2.3 and 2.4 confirm the previous observations. For both $\hat{\alpha}$ and $\hat{\beta}$ it is encouraging to observe that modes of the sample distributions for CL are always either on or very close to the real parameter value, even when $T = 100$. Similar to previous results, larger T decreases bias while increasing N leads to higher precision.

Table 2.9 reveals that the large sample theory is also spot on when $T \geq 250$. The average estimated standard deviations are generally close to the sample standard deviations. In addition, unlike the previous analysis, the confidence interval coverage rates are very good

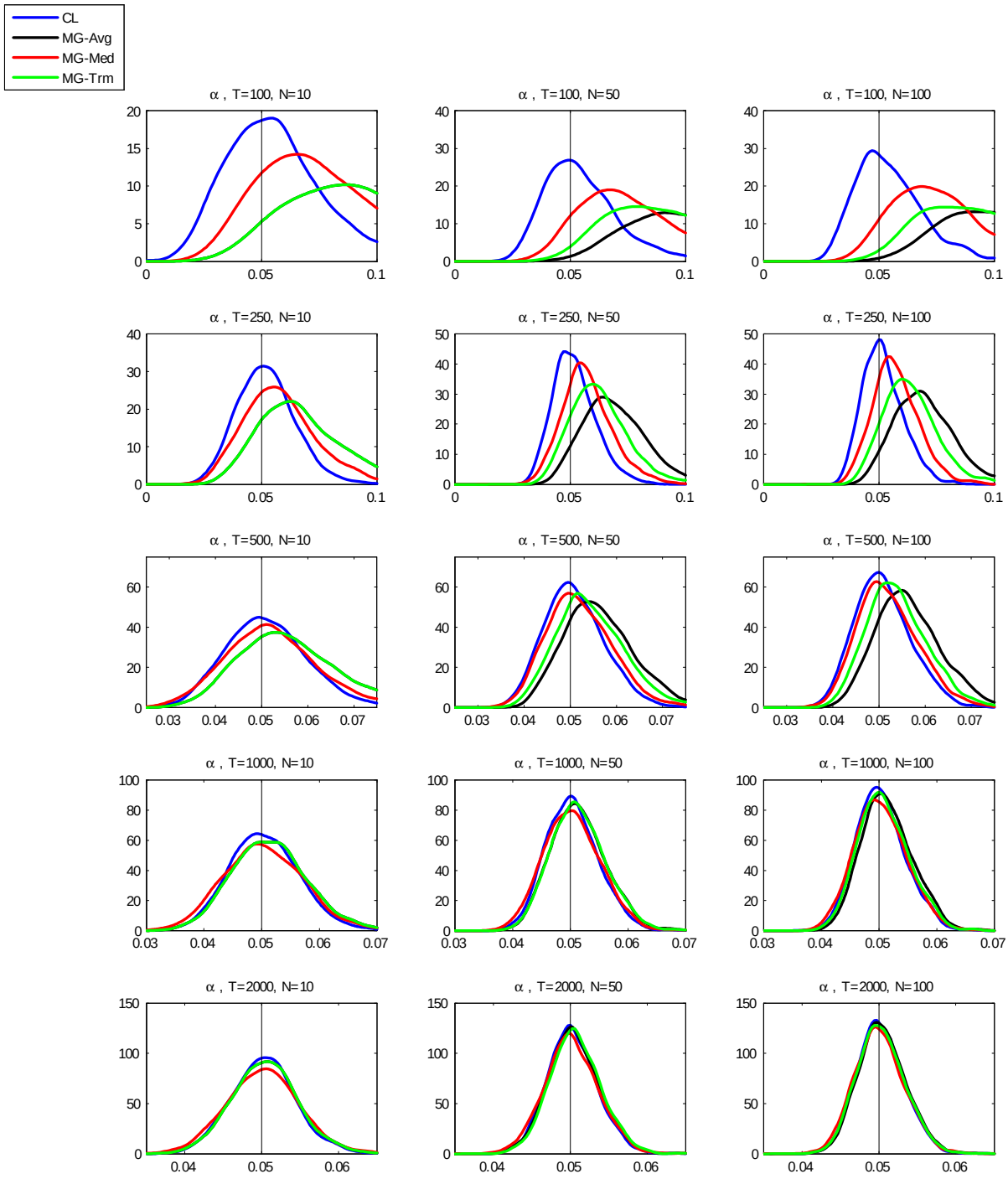

Figure 2.3: Sample distribution graphs for $\hat{\alpha}$. $\alpha = 0.05, \beta = 0.93$, based on 2,500 replications. MG-Avg, MG-Med and MG-Trm stand for MG-Mean, MG-Median and MG-Trimmed, respectively. True values of the individual-specific parameters are used when estimating $\hat{\alpha}$. The vertical line is drawn at $\alpha = 0.05$.

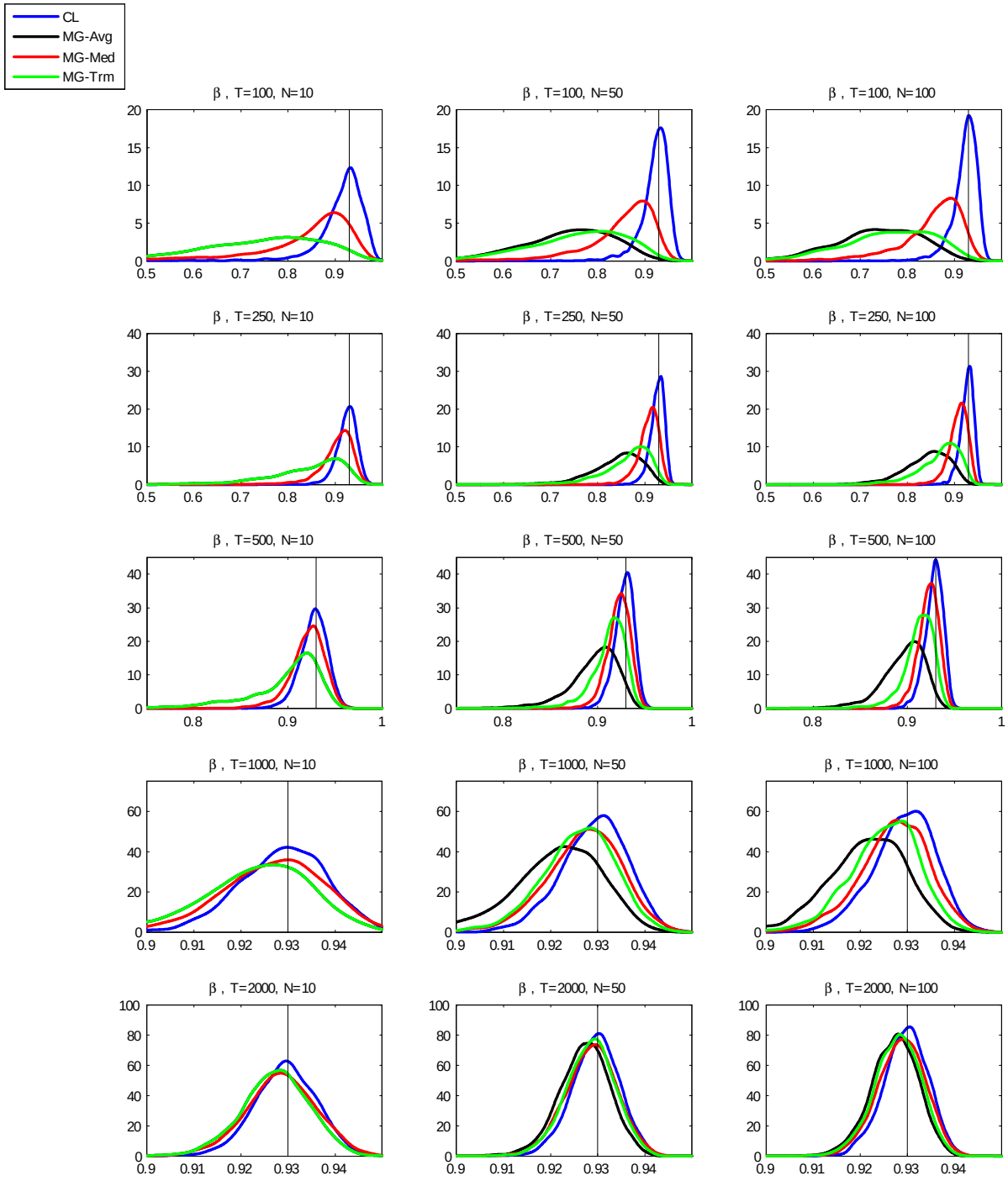

Figure 2.4: Sample distribution graphs for $\hat{\beta}$. $\alpha = 0.05, \beta = 0.93$, based on 2,500 replications. MG-Avg, MG-Med and MG-Trm stand for MG-Mean, MG-Median and MG-Trimmed, respectively. True values of the individual-specific parameters are used when estimating $\hat{\alpha}$. The vertical line is drawn at $\beta = 0.93$.

ranging between 90.8% and 94.9%. This is a very encouraging result as it implies that the large sample theory is actually accurate and the less than satisfactory confidence interval coverage rates for $\hat{\alpha}$ in the previous analysis are most likely due to the incidental parameter issue.

These results suggest that the incidental parameter issue also applies to CL estimation of GARCH panels. In addition, when the incidental parameter issue is bypassed, there are clear gains in using the panel rather than the time-series approach, as the comparison between CL and MG reveals. Clearly, in real life cases there is no way of obtaining an accurate estimate of the individual-specific parameters in short panels. However, it is possible to analytically characterise the effect of the small sample bias caused by estimation of the incidental parameters and correct for its effect on $\hat{\theta}$. If successful, this would imply that the GARCH methodology can be used in panels that are not long enough, such as $T = 200$, for the standard estimation approaches to work. Indeed, the next two Chapters are entirely dedicated to this idea.

2.5 Empirical Analysis

This section analyses the empirical performances of the CL, MG, and quasi maximum likelihood (QML) methods. Here, QML corresponds to estimating parameters of interest for each asset individually.

Although an obvious comparison of interest is that of CL against MG, it is also equally intriguing to compare the performance of the information pooling methods to the performance of QML. Unlike CL and MG, QML is not based on the parameter homogeneity assumption and so it can be considered a more flexible approach. However, while this assumption is likely to be violated, the dispersion of the true parameter values may not necessarily be very high, as illustrated in Figure 1.3 in Chapter 1. In addition, QML is not likely to perform well when there are only a few hundred observations.

It is not difficult to imagine a scenario where the forecaster has to use a very small sample. A recent structural break is one possibility. Assuming the break occurred a year ago, corresponding to around 250 observations following the break, QML and MG, which is based on QML, will most likely deliver a poor performance in modelling the conditional

heteroskedasticity. CL, on the other hand, has the potential to work well in this sample as it uses more information to estimate the GARCH parameters. This would enable the forecaster to use more recent data rather than having to disregard the structural break altogether.

Clearly, there is no guarantee that some or all assets follow the GARCH process, although this model has been found to be very successful in empirical analysis. Furthermore, even if it were known with certainty that all assets have GARCH dynamics, they may not have common and/or time-invariant parameters. However, these are of no direct interest. Instead, the questions of interest are highly pragmatic: can the data pooling approach attain better forecasting performance compared to QML? And, more importantly, does CL have an advantage over the other methods, especially in very small samples where QML is expected to perform poorly?

The analysis is conducted using stock-market data from S&P100. Note that in this study the comparison is between different methods (CL, MG and QML) that are used to estimate the parameters of the same model (GARCH). This is in contrast to comparison of two different models (e.g. GARCH and EGARCH). Therefore, the Giacomini and White (2006) test is well suited to the problem at hand as it allows comparison of the predictive ability of different methods as opposed to different models. A brief discussion of the methodology is provided next, which is followed by the analysis of empirical results.

2.5.1 Methodology

The Giacomini-White Test

Two seminal contributions in predictive ability comparison methodology are due to Diebold and Mariano (1995) and West (1996). Their approach is generally called the Diebold-Mariano-West (DMW) approach.[25] Under scrutiny are two competing τ-period ahead forecasts obtained at time t, $\hat{Y}_{1,t+\tau}$ and $\hat{Y}_{2,t+\tau}$, for a variable of interest, $Y_{t+\tau}$. The accuracy of forecasts is measured using loss functions. "Loss", in the forecast comparison sense, occurs due to the distance between the forecast and the true value of the variable of inter-

[25] Examples of other important contributions include McCracken (2000), White (2000), Chao, Corradi and Swanson(2001), Clark and McCracken (2001) and Corradi, Swanson and Olivetti (2001). Recent general surveys include Clements (2005), Corradi and Swanson (2006) and West (2006).

est. Therefore, the loss function provides a criterion to assess how well $\hat{Y}_{t+\tau}$ predicts $Y_{t+\tau}$. Formally, the loss due to $\hat{Y}_{t+\tau}$ is defined as

$$L_{t+\tau}(Y_{t+\tau}, \hat{Y}_{t+\tau}). \tag{2.26}$$

Examples of loss functions used in the literature are many.[26] Two prominent ones are the mean-squared error (MSE) and QLIKE loss functions:

$$\begin{aligned} MSE &: L_{t+\tau}(Y_{t+\tau}, \hat{Y}_{t+\tau}) = (Y_{t+\tau} - \hat{Y}_{t+\tau})^2, \\ QLIKE &: L_{t+\tau}(Y_{t+\tau}, \hat{Y}_{t+\tau}) = \log \hat{Y}_{t+\tau} + \frac{Y_{t+\tau}}{\hat{Y}_{t+\tau}}. \end{aligned}$$

The key feature of the DMW framework for the case at hand is that the null hypothesis is based on the probability limits of the estimators, rather than the estimates themselves. For example, in West (1996) this is because both the in-sample and out-of-sample sizes are allowed to increase asymptotically.[27] To see the implications of this, consider the null hypothesis under the DMW framework,

$$H_0 : E\left[L_{t+\tau}(Y_{t+\tau}, f_t(\beta_{1t}^*)) - L_{t+\tau}(Y_{t+\tau}, g_t(\beta_{2t}^*))\right] = 0, \tag{2.27}$$

where $f_t(\cdot)$ and $g_t(\cdot)$ are two different forecasting models with the parameters β_{1t} and β_{2t}, respectively. In addition, define the respective (pseudo) true parameters β_{1t}^* and β_{2t}^*. The loss function is based on β_{1t}^* and β_{2t}^* since, assuming consistency, the estimators will converge to these values when the in-sample size grows to infinity. For this reason, the DMW approach cannot be used to compare two different methods (e.g. CL and MG) which are used to estimate the same model. This is because both methods are estimating the same model/parameter and, therefore, asymptotically (2.27) is identical to zero. This issue is solved by Giacomini and White (2006) (GW) who provide a suitable testing framework for comparing two methods used to estimate the same model. This is achieved by not allowing the in-sample size to increase asymptotically. As such, the null hypothesis is based

[26] See Patton (2011) for a more detailed study of implications of using different loss functions.

[27] The in-sample denotes the portion of the sample which is used to estimate the parameters or the model. Out-of-sample, on the other hand, is the portion of the sample which is being forecast.

on parameter estimates and not their probability limits:

$$H_0 : E[L_{t+\tau}(Y_{t+\tau}, f_t(\hat{\beta}_{1t})) - L_{t+\tau}(Y_{t+\tau}, g_t(\hat{\beta}_{2t}))|\mathcal{G}_t] = 0, \tag{2.28}$$

where $\mathcal{G}_t$ is the information set at time t.

Two differences between the GW and DMW tests are revealed by (2.28). First, as expected, the null hypothesis uses sample values ($\hat{\beta}_{it}$) instead of population values (β^*_{it}). This is a crucial difference, accounting for factors that influence forecasts such as data selection, method of estimation, estimation window and parametric assumptions, among others. In applications, this is achieved by considering limited memory estimators, such as the rolling window estimator.[28] The second difference is that the GW test allows for a conditional, as well as an unconditional approach (by letting $\mathcal{G}_t = \{\oslash, \Omega\}$). The latter compares the average performances of two forecasting methods while the former analyses whether past information can be used to predict which method will provide a better forecast for a particular future date. Details of the GW test are provided in Appendix 2.B.

Latent Variables and Choice of Proxy

Equation (2.26) assumes that the true value at $t+\tau$, $Y_{t+\tau}$, is observable. However, volatility is latent, in the sense that its values are never observed, even ex-post. Therefore, a proxy is needed. In this study the squared return, r_t^2, is used as proxy, which is a very common choice.

An important criticism against using the squared returns as a proxy for conditional variance is that it is a very noisy estimator. Instead, realised volatility is recommended as a better proxy.[29] For example, Hansen and Lunde (2006) show that the use of noisy proxies such as r_t^2 may lead to inconsistent ranking of volatility models, whereby the empirical ranking may not necessarily be the same as the true ranking. As a possible remedy, Patton

[28] The rolling window estimator uses a fixed number of observations for estimation at each point in time. In other words, the size of the in-sample remains the same over time. Therefore, even asymptotically, $\hat{\beta}_{kt}$ does not converge to β^*_{kt}, $k = 1, 2$.

[29] Very briefly, realised volatility is the sum of squared high-frequency intra-daily returns. Prominent examples of the application of this idea to econometrics include Andersen, Bollerslev, Diebold and Labys (2001 and 2003), Barndorff-Nielsen and Shephard (2002, 2004) and Meddahi (2002) while examples of related analysis on volatility forecasting are provided by Andersen and Bollerslev (1998), Andersen, Bollerslev and Lange (1999), Andersen, Bollerslev and Meddahi 2004). See Andersen and Benzoni (2009) for a recent survey.

(2011) considers loss functions that are robust to the choice of the volatility proxy, in the sense that the empirical ranking implied by these loss functions are the same independent of which proxy is used. He provides a family of homogeneous and robust loss functions that contains MSE and QLIKE, as well. In addition, the simulation analysis by Patton and Sheppard (2009) indicates that QLIKE has the best power performance. Motivated by these results, this study uses QLIKE as the loss function.

2.5.2 Empirical Results

The empirical analysis is based on the daily returns for 94 stocks from S&P100 for the period from 3 March 2000 to 12 December 2008, which gives $2,200$ time-series observations for each stock.[30] Data were obtained from DataStream. The forecast comparison analysis is based on one-step ahead forecasts obtained by using the CL, MG and QML methods. In order to keep the exposition clear, only the median blend function is used in this part. In the original paper of Engle (2009), median is generally superior to other blend functions. In addition, simulation results of the previous section also indicate that median is a good choice. To cover a variety of cases, forecasts are based on samples of size 250, 500, 750 and $1,000$. The first two consider the case where the time-series variation is not sufficient for QML to deliver accurate estimates. The remaining sample sizes are, on the other hand, large enough for standard time-series methods to perform well. However, interestingly, the test results will reveal that even in large samples, where there is not much need for additional methods to extract information, information pooling still emerges as a valid competitor to the standard QML method.

Results for both the conditional and unconditional tests are reported. As discussed in Appendix 2.B, a "test function" is required for the conditional test. This study employs the same test function as Giacomini and White (2006), which consists of a constant and the previous period's loss-difference. This is given by $h_t = (1, \Delta L_{m,t-1+\tau})'$. This is a reasonable choice: there may be some time-independent difference in the predictive abilities of the two methods at any point in time. This is reflected by the constant. Moreover, past comparisons of methods can also give an idea about their relative future performances, since a method

[30] Data for six firms has been discarded as the stocks for these firms were not traded in part of the period considered in the analysis. These firms are Covidien, Google, Kraft Foods, Mastercard, NYSE Euronext and Philip Morris International.

Unconditional 1-Step Ahead

n	m	CL vs Md		CL vs QML		Md vs QML	
1,200	1,000	16 (.94)	1 (.06)	15 (.58)	11 (.42)	12 (.48)	13 (.52)
1,450	750	36 (1.0)	0 (.00)	18 (.62)	11 (.38)	12 (.44)	15 (.56)
1,700	500	37 (1.0)	0 (.00)	18 (.75)	6 (.25)	5 (.31)	11 (.69)
1,950	250	25 (.96)	1 (.04)	17 (.81)	4 (.19)	11 (.61)	7 (.39)

Conditional 1-Step Ahead

n	m	CL vs Md		CL vs QML		Md vs QML	
1,200	1,000	20 (.95)	1 (.05)	14 (.52)	13 (.48)	11 (.42)	15 (.58)
1,450	750	38 (.95)	2 (.05)	18 (.60)	12 (.40)	8 (.32)	17 (.68)
1,700	500	37 (1.0)	0 (.00)	15 (.63)	9 (.37)	8 (.40)	12 (.60)
1,950	250	30 (1.0)	0 (.00)	16 (.84)	3 (.16)	7 (.54)	6 (.46)

Table 2.10: Test results for 1-step ahead forecasts. Top panel gives the results for the unconditional test while the results for the conditional test are given in the bottom panel. The in- and out-of-sample sizes are given by m and n, respectively. CL, Md and QML stand for 'composite likelihood', 'MacGyver-Median' and 'quasi maximum likelihood'. When calculating the Md forecasts, cases where estimates did not converge during optimisation and/or cases where $\hat{\alpha} \leq 0.0025$ have not been removed from the pool of estimates. Each comparison is based on individual Giacomini-White tests of predictive ability for each of the 94 stocks. For each comparison, the number of assets where the Giacomini-White test decides in favour of a given method is given in columns 3-8, under the name of the respective method. The proportion of cases where a given method is favoured by the test procedure is given in parantheses. For example, in the comparison of 1-step ahead predictive ability between CL and QML, when $m = 500$ and $n = 1,700$, the unconditional Giacomini-White test finds that in 24 out of 94 cases, equal predictive ability is rejected, while for the remaining cases the test is inconclusive. In 18 (75%) of these 24 cases, the test decides in favour of the CL method.

that has been superior in the past is likely to remain so in the future. This is reflected by the past loss difference. Lastly, all tests are conducted on an asset-by-asset basis; that is, comparison of predictive ability is conducted for each asset individually. The test results are presented in Tables 2.10 and 2.11. All results are for 5% level of significance.

Table 2.10 presents results of the conditional and unconditional tests, where the pool of estimates used to calculate the MG estimator contains cases where $\hat{\alpha} \leq 0.0025$ and/or where the estimator does not converge during optimisation. One might argue that, in practice the researcher would not hesitate to eliminate such cases from the pool, as, for example, $\alpha = 0$ implies that β is not identified. However, it must be remembered that for more complex models, identification conditions might not be as straight-forward as here. In other words, MG has the additional difficulty that the researcher has to characterise cases where a given

Unconditional 1-Step Ahead

n	m	CL vs Md		Md vs QML	
$1,200$	$1,000$	14 (.93)	1 (.07)	11 (.46)	13 (.54)
$1,450$	750	28 (1.0)	0 (.00)	13 (.48)	14 (.52)
$1,700$	500	11 (.92)	1 (.08)	9 (.60)	6 (.40)
$1,950$	250	1 (.06)	15 (.94)	16 (.94)	1 (.06)

Conditional 1-Step Ahead

n	m	CL vs Md		Md vs QML	
$1,200$	$1,000$	14 (.82)	3 (.18)	10 (.40)	15 (.60)
$1,450$	750	30 (.94)	2 (.06)	10 (.38)	16 (.62)
$1,700$	500	11 (.79)	3 (.21)	11 (.61)	7 (39)
$1,950$	250	2 (.15)	11 (.85)	13 (.87)	2 (.13)

Table 2.11: Test results for 1-step ahead forecasts. When calculating the Md forecasts, cases where estimates did not converge during optimisation and/or cases where $\hat{\alpha} \leq 0.0025$ have been removed from the pool of estimates. See Table 2.11 for more information.

estimate should be eliminated from the pool. Therefore, to reflect this issue and the ad-hoc nature of MG, such problematic cases have not been removed from the pool of estimates in the forecast calculations underlying Table 2.10. To illustrate the impact of this elimination process, Table 2.11 presents the results when the problematic cases are eliminated from the pool.

Table 2.10 reveals that both conditional and unconditional approaches exhibit similar patterns. Comparing the two pooling methods, it is evident that whenever their predictive abilities can be distinguished, the GW test decides in favour of CL almost all the time. Results for the same comparison in Table 2.11 show that excluding problematic cases leads to a decrease in the number of rejections of equal predictive ability. More importantly, now MG is clearly superior to CL when $m = 250$. Clearly, excluding problematic estimates does help MG to achieve better performance. Neither CL nor QMLE enjoy such a luxury as neither is based on a pool of estimates. Without much doubt, when m is as small as 250 both CL and QMLE perform poorly in estimating the parameters of interest. However, by eliminating the very worst estimates, MG attains better performance. Moreover, a parallel improvement in MG's performance against QMLE can be observed, as well. Nevertheless, it must be noted that this advantage of MG crucially depends on the ability to characterise

"problematic" cases, which will be more difficult when working with complex models.

Another observation for the comparison between CL and MG is that the highest number of rejections of equal predictive ability is achieved for $m \in \{500, 750\}$, while the number of rejections fall when $m = 1,000$. There are two messages here. First, it can be argued that as the in-sample size increases, both methods start to perform equally well, and hence, it is more difficult to distinguish between them. Second, although it is now more difficult to distinguish between the two methods, CL still is preferred to MG.

Comparison of the performances of CL and QMLE shows that, not surprisingly, rejections in favour of CL increase in percentage as m decreases. When m is small, QMLE is not expected to perform well as the number of observations is not large enough to estimate the parameters accurately. CL, on the other hand, utilises cross-sectional information, as well, which would explain its higher success rate when m is small. However, as m increases, CL's relative superiority over QMLE lessens, as QMLE has more data to estimate the parameters of interest. Perhaps the most encouraging result is that, interestingly, even when QMLE is expected to work well (when, for example, $m = 1,000$), CL is still preferred to QMLE more than 50% of the time. This is very important. This suggests that the GARCH panel model is not confined to estimation when the time-series information is insufficient. Instead, in this particular exercise, it emerges as a valid alternative to QML even when the time-series information is supposed to be rich enough. As such, clearly there are gains in pooling cross-sectional information, even when the time-series information can be deemed sufficient.

The performance of MG against QMLE is relatively worse compared to that of CL against QMLE, especially as m decreases. Comparison of Tables 2.10 and 2.11 again reveals the remarkable effect of removing problematic estimates from the pool. This leads to a dramatic improvement in MG's performance against QMLE at small in-sample sizes. However, it is interesting that even when such cases are not eliminated, MG is still doing well against QMLE at $m = 250$. At first sight, it might seem strange that an estimator based on a pool of QMLE estimates, which are expected to be very poor when $m = 250$, performs distinctively better than QMLE. However, this is not surprising: the sample distribution of QMLE estimates is so dispersed that although the estimates are individually very poor in general, their median still delivers a reasonable value, instead of $\hat{\alpha} = 0$.

Hence, empirical analysis results suggest that CL delivers better forecasting performance than MG at all in-sample sizes, while it stands out as a good alternative to QMLE even when m is large. Although it can be argued that MG is not decisively beaten by QMLE either, it is clear that CL's performance against QMLE is superior compared to that of MG against QMLE. This is an important result in favour of CL.

To conclude, empirical analysis results are in support of panel estimation of volatility using CL in both small and large samples. This is a strong motivation to consider further improvements to the CL method, especially in order to solve the incidental parameter problem, which can lead to an even better performance.

2.6 Conclusion

This chapter has introduced and studied the theoretical and empirical properties of a novel volatility modelling approach using panels of financial data. Based on the work of Engle, Shephard and Sheppard (2008), the large sample theory of composite likelihood estimation of GARCH panels has been developed. Also, simulation and empirical analyses have been conducted in order to assess the finite sample properties of both information pooling methods.

Simulation results reveal that CL has good finite sample properties and delivers accurate estimators in panels with around or more than 500 time-series observations. This is an improvement over the standard time-series based estimation approach, which potentially requires around $1,000$ or more observations for accurate parameter estimation. Simulation results also confirm that CL suffers severely from the incidental parameter issue in smaller samples. This is natural given that the GARCH Panel model is essentially a non-linear dynamic panel data model, and such models are well-documented to suffer from this bias in the presence of individual-specific parameters. Finally, it has been shown that once this issue is solved, CL delivers a highly satisfactory performance using as little as 250 time-series observations. These observations are very encouraging as they imply that this new approach can make it possible to use the GARCH methodology in samples that are typically short and are, therefore, beyond the reach of standard GARCH estimation approaches. Finally, the forecast comparison analysis confirms that the panel estimation method attains superior

forecasting performance in small samples against the standard quasi maximum likelihood method (QML). More interestingly, even in larger samples, CL still stands out as a legitimate alternative to QML.

Given that the model is capable of modelling volatility accurately in samples of previously unimaginable sizes once the incidental parameter issue is solved, a highly interesting question is whether this problem can indeed be solved. This is investigated in the next chapter, within the greater framework of non-linear and dynamic panel data models.

2.A Mathematical Appendix

2.A.1 Proof of Theorem 2.3.1

Proof. Define,

$$\tilde{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \ell_{it}(\theta, \lambda_{i0}).$$

By Assumption 2.3.2,

$$\tilde{\theta} \xrightarrow{p} \theta_0. \tag{2.29}$$

Note that an Array Law of Large Numbers is necessary, as both N and T go to ∞. Importantly, $\tilde{\theta}$ is the maximiser when the pseudo-true value of the nuisance parameter, λ_{i0}, is known for all i. As such, convergence of $\tilde{\theta}$ to θ_0 does not necessarily imply convergence of $\hat{\theta}$ to θ_0. Remember that the composite likelihood function is given by

$$\ell(\theta, \lambda) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \ell_{it}(\theta, \lambda_i).$$

To show that $\hat{\theta} \xrightarrow{p} \theta_0$, first observe that, since $\ell(\theta, \tilde{\lambda}) = \ell(\theta, \lambda_0) + \ell(\theta, \tilde{\lambda}) - \ell(\theta, \lambda_0)$ by definition,

$$\arg\max_{\theta\in\Theta} \ell(\theta, \tilde{\lambda}) = \arg\max_{\theta\in\Theta} [\ell(\theta, \lambda_0) + \ell(\theta, \tilde{\lambda}) - \ell(\theta, \lambda_0)]. \tag{2.30}$$

Now,

$$\begin{aligned}
\sup_{\theta\in\Theta} \left|\ell(\theta, \tilde{\lambda}) - \ell(\theta, \lambda_0)\right| &= \sup_{\theta\in\Theta} \left| \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \left[\ell_{it}(\theta, \tilde{\lambda}_i) - \ell_{it}(\theta, \lambda_{i0})\right]\right| \\
&\leqslant \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \sup_{\theta\in\Theta} \left|\ell_{it}(\theta, \tilde{\lambda}_i) - \ell_{it}(\theta, \lambda_{i0})\right|.
\end{aligned}$$

By the Mean Value Theorem and using Assumption 2.3.3,

$$\ell_{it}(\theta, \tilde{\lambda}_i) = \ell_{it}(\theta, \lambda_{i0}) + (\tilde{\lambda}_i - \lambda_{i0}) \frac{\partial \ell_{it}(\theta, \lambda_i)}{\partial \lambda_i}\Big|_{\lambda_i = \bar{\lambda}_i},$$

where $\bar{\lambda}_i \in [\min(\tilde{\lambda}_i, \lambda_{i0}), \max(\tilde{\lambda}_i, \lambda_{i0})]$. So,

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \sup_{\theta\in\Theta} \left|\ell_{it}(\theta, \tilde{\lambda}_i) - \ell_{it}(\theta, \lambda_{i0})\right|$$

$$
\begin{aligned}
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sup_{\theta\in\Theta}\left|(\tilde{\lambda}_i-\lambda_{i0})\frac{\partial\ell_{it}(\theta,\lambda_i)}{\partial\lambda_i}\Big|_{\lambda_i=\bar{\lambda}_i}\right| && (2.31)\\
&\leqslant \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{\lambda}_i-\lambda_{i0}\right|\sup_{\theta\in\Theta}\left|\frac{\partial\ell_{it}(\theta,\lambda_i)}{\partial\lambda_i}\Big|_{\lambda_i=\bar{\lambda}_i}\right|\\
&\leqslant \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left|\tilde{\lambda}_i-\lambda_{i0}\right|\sup_{\theta\in\Theta,\lambda_i\in\Lambda_i}\left|\frac{\partial\ell_{it}(\theta,\lambda_i)}{\partial\lambda_i}\right|\\
&\leqslant \max_{i\in\{1,\ldots,N\}}\left|\tilde{\lambda}_i-\lambda_{i0}\right|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sup_{\theta\in\Theta,\lambda_i\in\Lambda_i}\left|\frac{\partial\ell_{it}(\theta,\lambda_i)}{\partial\lambda_i}\right|\\
&= o_p(1). && (2.32)
\end{aligned}
$$

The last line follows by Assumptions 2.3.5 and 2.3.6 which ensure that, $\max_{i\in\{1,\ldots,N\}}\left|\tilde{\lambda}_i-\lambda_{i0}\right| = o_p(1)$ and $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sup_{\theta\in\Theta,\lambda_i\in\Lambda_i}|\partial\ell_{it}(\theta,\lambda_i)/\partial\lambda_i| = O(1)$, respectively. Therefore,

$$
\sup_{\theta\in\Theta}\left|\ell(\theta,\tilde{\lambda})-\ell(\theta,\lambda_0)\right|\xrightarrow{p}0, \tag{2.33}
$$

and, consequently

$$
\hat{\theta} := \arg\max_{\theta\in\Theta}\ell(\theta,\tilde{\lambda}) = \arg\max_{\theta\in\Theta}\ell(\theta,\lambda_0) =: \tilde{\theta} \text{ as } N,T\to\infty.
$$

Consistency of $\hat{\theta}$ then follows from (2.29). ■

2.A.2 Proof of Theorem 2.3.2

Proof. First, define

$$
\tilde{g}_{tN}(y_t,\lambda,\theta) = \begin{bmatrix} (y_{1t}^2-\lambda_1) \\ \vdots \\ (y_{Nt}^2-\lambda_N) \\ \frac{\partial}{\partial\theta}\frac{1}{N}\sum\limits_{i=1}^{N}\left(-\frac{1}{2}\log\sigma_{it}^2-\frac{1}{2}\frac{\varepsilon_{it}^2}{\sigma_{it}^2}\right) \end{bmatrix}.
$$

Notice that $T^{-1}\sum_{t=1}^{T}\tilde{g}_{tN}(y_t,\tilde{\lambda},\hat{\theta}) = \mathbf{0}_{N+2}$, where $\mathbf{0}_{N+2}$ is an $((N+2)\times 1)$ vector of zeroes. Then, by a first-order Taylor approximation,

$$
\frac{1}{T}\sum_{t=1}^{T}\tilde{g}_{tN}(y_t,\hat{\psi}) = \frac{1}{T}\sum_{t=1}^{T}\tilde{g}_{tN}(y_t,\psi_0) + \left[\frac{1}{T}\sum_{t=1}^{T}\nabla_{\psi}\tilde{g}_{tN}(y_t,\psi_0)\right](\hat{\psi}-\psi_0) + O_p(||\hat{\psi}-\psi_0||^2),
$$

where $||\cdot||$ is the Euclidian norm, $\hat{\psi} = (\hat{\theta}, \tilde{\lambda})$, $\psi_0 = (\theta_0, \lambda_0)$ and

$$\nabla_\psi \tilde{g}_{tN}(y_t, \psi_0) = \frac{\partial \tilde{g}_{tN}(y_t, \psi)}{\partial \psi'}\Big|_{\psi=\psi_0}.$$

It then follows from the sample moment condition $T^{-1}\sum_{t=1}^{T} \tilde{g}_{tN}(y_t, \hat{\psi}) = \mathbf{0}_{N+2}$ that

$$(\hat{\psi} - \psi_0) = -\left[\frac{1}{T}\sum_{t=1}^{T} \nabla_\psi \tilde{g}_{tN}(y_t, \psi_0)\right]^{-1} \frac{1}{T}\sum_{t=1}^{T} \tilde{g}_{tN}(y_t, \psi_0) + o_p(1).$$

Now, $T^{-1}\sum_{t=1}^{T} \nabla_\psi \tilde{g}_{tN}(y_t, \psi)$ can be partitioned as follows:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \nabla_\psi \tilde{g}_{tN}(y_t, \psi) &= \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \\
\underset{(N\times N)}{a} &= \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} \frac{\partial m(y_{1t},\lambda_1)}{\partial \lambda_1} & \cdots & \frac{\partial m(y_{1t},\lambda_1)}{\partial \lambda_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial m(y_{Nt},\lambda_N)}{\partial \lambda_1} & \cdots & \frac{\partial m(y_{Nt},\lambda_N)}{\partial \lambda_N} \end{bmatrix} \\
&= -I_N, \text{ the identity matrix of dimension } N,
\end{aligned}$$

$$\begin{aligned}
\underset{(N\times 2)}{b} &= \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} \frac{\partial m(y_t,\lambda_1)}{\partial \alpha} & \frac{\partial m(y_t,\lambda_1)}{\partial \beta} \\ \vdots & \vdots \\ \frac{\partial m(y_t,\lambda_N)}{\partial \alpha} & \frac{\partial m(y_t,\lambda_N)}{\partial \beta} \end{bmatrix} = \mathbf{0}_{N\times 2}, \\
\underset{(2\times N)}{c} &= \frac{1}{NT}\sum_{t=1}^{T} \begin{bmatrix} \frac{\partial^2 \ell_{1t}(y_t,\theta,\lambda_1)}{\partial \alpha \partial \lambda_1} & \cdots & \frac{\partial^2 \ell_{Nt}(y_t,\theta,\lambda_N)}{\partial \alpha \partial \lambda_N} \\ \frac{\partial^2 \ell_{1t}(y_t,\theta,\lambda_1)}{\partial \beta \partial \lambda_1} & \cdots & \frac{\partial^2 \ell_{Nt}(y_t,\theta,\lambda_N)}{\partial \beta \partial \lambda_N} \end{bmatrix}, \\
\underset{(2\times 2)}{d} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \begin{bmatrix} \frac{\partial^2 \ell_{it}(\theta,\lambda_i)}{\partial \alpha^2} & \frac{\partial^2 \ell_{it}(\theta,\lambda_i)}{\partial \alpha \partial \beta} \\ \frac{\partial^2 \ell_{it}(\theta,\lambda_i)}{\partial \beta \partial \alpha} & \frac{\partial^2 \ell_{it}(\theta,\lambda_i)}{\partial \beta^2} \end{bmatrix},
\end{aligned}$$

where $\mathbf{0}_{N\times 2}$ is an $(N \times 2)$ matrix of zeroes. Note that, for a matrix A partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

one has

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}X^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}X^{-1} \\ -X^{-1}A_{21}A_{11}^{-1} & X^{-1} \end{bmatrix}$$

assuming that both A and $X = A_{22} - A_{21}A_{11}^{-1}A_{12}$ are non-singular (see, for example, Magnus and Neudecker (1988, p.11)). Then,

$$\left[\frac{1}{T}\sum_{t=1}^{T}\nabla_{\psi}\tilde{g}_{tN}(y_t,\psi)\right]^{-1} = \begin{bmatrix} \cdot & \cdot \\ d^{-1}c & d^{-1} \end{bmatrix}. \tag{2.34}$$

As the analysis focuses on $\hat{\theta}$, only the lower block of (2.34) is considered. Therefore,

$$\begin{aligned}
(\hat{\theta}-\theta_0) &\simeq -\begin{bmatrix} d^{-1}c & d^{-1} \end{bmatrix}\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} m(y_{1t},\lambda_{10}) \\ \vdots \\ m(y_{Nt},\lambda_{N0}) \\ \frac{\partial}{\partial\theta}\frac{1}{N}\sum_{i=1}^{N}\left(-\frac{1}{2}\log\sigma_{it}^2 - \frac{1}{2}\frac{\varepsilon_{it}^2}{\sigma_{it}^2}\right) \end{bmatrix} \\
&= -D_{\theta\theta,NT}^{-1}\left\{\frac{1}{NT}\sum_{i=1}^{N}\left[m(y_{it},\lambda_{i0})\sum_{t=1}^{T}\frac{\partial^2\ell_{it}(\theta,\lambda_i)}{\partial\theta\partial\lambda_i}\Big|_{\psi=\psi_0}\right] + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial\ell_{it}(\theta,\lambda_i)}{\partial\theta}\Big|_{\psi=\psi_0}\right\} \\
&= -D_{\theta\theta,NT}^{-1}\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[\frac{\partial\ell_{it}(\theta,\lambda_i)}{\partial\theta}\Big|_{\psi=\psi_0} - m(y_{it},\lambda_{i0})F_{iT}\right]\right\} \qquad (2.35) \\
&= -D_{\theta\theta,NT}^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}Z_{t,NT}\right\} \qquad (2.36)
\end{aligned}$$

where (2.35) and (2.36) follow from

$$\begin{aligned}
F_{iT} &= -\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2\ell_{it}(\theta,\lambda_i)}{\partial\theta\partial\lambda_i}\Big|_{\psi=\psi_0} \\
\text{and } Z_{t,NT} &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial\ell_{it}(\theta,\lambda_i)}{\partial\theta}\Big|_{\psi=\psi_0} - m(y_{it},\lambda_i)F_{iT}\right],
\end{aligned}$$

respectively. Also, the existence of derivatives is ensured by Assumption 2.3.8. Then,

$$\sqrt{T}(\hat{\theta}-\theta_0) = -D_{\theta\theta,NT}^{-1}\left(\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}Z_{t,NT}\right) + o_p(1)$$

and, finally, using Assumptions 2.3.9 and 2.3.10

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{D}_{\theta\theta}^{-1}\mathcal{I}_{\theta\theta}\mathcal{D}_{\theta\theta}^{-1}),$$

by Slutsky's Theorem. ■

2.B An Overview of the Giacomini-White Test

The details of the GW Test are provided here, for the sake of completeness. The discussion in this section is not novel and essentially is a reproduction of the results of Giacomini and White (2006).

First, several assumption are presented to describe the setting.

Assumption 2.B.1 *Let $(\Omega, \mathcal{F}, P_0)$ define a complete probability space. The stochastic process W and the σ-field $\mathcal{F}_t$ are defined as*

$$W = \{W_t : \Omega \to \mathbb{R}^{k+1}, k \in \mathbb{N}, t = 1, 2, ..., T\} \ , \ W_t = (X_t, Y_t) \tag{2.37}$$

$$\text{and } \mathcal{F}_t = \sigma(W_t', W_{t-1}', ..., W_1'). \tag{2.38}$$

where $X_t : \Omega \to \mathbb{R}^k$ are the predictor variables and $Y_t : \Omega \to \mathbb{R}$ is the variable of interest.

In this study, there are no exogenous predictors. Therefore $W_t = Y_t$.

Assumption 2.B.2 *Two competing τ-step ahead forecasts, calculated at time t, of a variable of interest are given by*

$$\hat{f}_{t+\tau,m_f} = f(W_t, W_{t-1}, ..., W_{t-m_f+1}; \hat{\beta}_{t,m_f}) \text{ and } \hat{g}_{t+\tau,m_g} = f(W_t, W_{t-1}, ..., W_{t-m_g+1}; \hat{\beta}_{t,m_g}),$$

where m_f and m_g are the sizes of the in-sample used to obtain $\hat{f}_{t+\tau,m_f}$ and $\hat{g}_{t+\tau,m_g}$, respectively. Similarly, $\hat{\beta}_{t,m_f}$ and $\hat{\beta}_{t,m_g}$ are the parameter estimates used by the respective forecasting methods.

This framework implies a total of $n = T - \tau - m_i + 1$ out-of-sample predictions, where T is the total number of observations in the whole sample, τ is the step-size for forecasting and $i \in \{m, n\}$. Note that, both m_f and m_g can be chosen to be time-dependent. However, this option is not considered here. Also, in what follows, it is assumed for simplicity that $m_f = m_g$.

2.B.1 The Test of Conditional Predictive Ability

Under Assumptions 2.B.1 and 2.B.2, and using an appropriate loss function, the losses associated with the two forecasts are given by $L_{t+\tau}(Y_{t+\tau}, \hat{f}_{t+\tau,m})$ and $L_{t+\tau}(Y_{t+\tau}, \hat{g}_{t+\tau,m})$,

respectively. For $\mathcal{G}_t = \mathcal{F}_t$, the null hypothesis of equal conditional predictive ability stated in equation (2.28) can now be formulated as

$$\begin{aligned} H_0 \quad &: \quad \mathbb{E}\left[\Delta L_{m,t+\tau} | \mathcal{F}_t\right] = 0, \\ \text{where} \quad \Delta L_{m,t+\tau} \quad &= \quad L_{t+\tau}(Y_{t+\tau}, \hat{f}_{t+\tau,m}) - L_{t+\tau}(Y_{t+\tau}, \hat{g}_{t+\tau,m}). \end{aligned} \tag{2.39}$$

An unconditional version of (2.39) can easily be obtained using any $\mathcal{F}_t$-measurable function h_t and the Law of Iterated Expectations, which yields

$$H_{0,h} : \mathbb{E}\left[h_t \Delta L_{m,t+\tau}\right] = 0.$$

Here, h_t is the "test function" and the null hypothesis is defined based on the particular choice of h_t.[31]

Using a Wald-type test statistic, Giacomini and White (2006) show in their Theorem 3 that under H_0 given in (2.39)

$$\begin{aligned} T^h_{m,n,\tau} &= n \left(n^{-1} \sum_{t=m}^{T-\tau} h_t \Delta L_{m,t+\tau} \right)' \tilde{\Omega}_n^{-1} \left(n^{-1} \sum_{t=m}^{T-\tau} h_t \Delta L_{m,t+\tau} \right) \\ &= n \bar{Z}'_{m,n} \tilde{\Omega}_n^{-1} \bar{Z}_{m,n} \xrightarrow{p} \chi^2_{(q)} \quad \text{as } n \to \infty \\ \text{where,} \quad \bar{Z}_{m,n} &= n^{-1} \sum_{t=m}^{T-\tau} Z_{m,t+\tau}, \; Z_{m,t+\tau} = h_t \Delta L_{m,t+\tau}, \\ \tilde{\Omega}_n^{-1} &= n^{-1} \sum_{t=m}^{T-\tau} Z_{m,t+\tau} Z'_{m,t+\tau} \\ &\quad + n^{-1} \sum_{j=1}^{\tau-1} w_{n,j} \sum_{t=m+j}^{T-\tau} \left[Z_{m,t+\tau} Z'_{m,t+\tau-j} + Z_{m,t+\tau-j} Z'_{m,t+\tau} \right]. \end{aligned} \tag{2.40}$$

and h_t is a $q \times 1$ $\mathcal{F}_t$-measurable test function. In (2.40), $w_{n,j}$ is a weight function such that $w_{n,j} \to 1$ as $n \to \infty$ for each $j = 1, ..., \tau - 1$. An appropriate weight function can be chosen in the same spirit as for HAC estimators. Moreover, under the alternative hypothesis

$$H_{A,h} \quad : \quad E\left[\bar{Z}'_{m,n}\right] E\left[\bar{Z}_{m,n}\right] > 0 \quad \text{for all } n \text{ sufficiently large} \tag{2.41}$$

[31] Strictly speaking, Giacomini and White (2006) focus on a subset of $\tilde{h}_t$, which is *any* $\mathcal{F}_t$-measurable function. Therefore, the test function must be 'hand-picked' by the forecaster, to include variables that can predict forecast performance.

$$\text{and} \quad P\left[T_{m,n}^{h} > c\right] \quad \rightarrow \quad 1 \quad \text{as } n \rightarrow \infty,$$

for any constant $c \in \mathbb{R}$.

2.B.2 The Test of Unconditional Predictive Ability

The unconditional case corresponds to choosing $\mathcal{G}_t = \{\varnothing, \Omega\}$, implying that the information set includes all available information. In this case, the expectation in (2.39) will by definition be unconditional. Therefore, the focus will be on the test of equal predictive ability on average. The relevant null and alternative hypotheses are

$$\begin{aligned} H_0 \quad &: \quad \mathbb{E}\left[\Delta L_{m,t+\tau}\right] = 0 \quad \text{for } t = 1, ..., T \\ \text{and} \quad H_A \quad &: \quad \left|\mathbb{E}\left[\Delta \bar{L}_{m,n}\right]\right| \geq \delta > 0 \quad \text{for all } n \text{ sufficiently large,} \\ \text{where} \quad \Delta \bar{L}_{m,n} \quad &= \quad \frac{1}{n} \sum_{t=m}^{T-\tau} \Delta L_{m,t+\tau}. \end{aligned}$$

The relevant test statistic is

$$t_{m,n,\tau} = \frac{\Delta \bar{L}_{m,n}}{\hat{\sigma}_n / \sqrt{n}},$$

where $\hat{\sigma}_n$ is an estimator for $\sigma_n^2 = var\left[\sqrt{n}\Delta \bar{L}_{m,n}\right]$ and has to be obtained by using a suitable HAC estimator.

Chapter 3

Bias Reduction in Nonlinear and Dynamic Panels in the Presence of Cross-Section Dependence

3.1 Introduction

A substantial body of research in econometrics has been dedicated to controlling for unobserved heterogeneity (see Chamberlain (1984) and Arellano and Honoré (2001) for surveys). In the simple case of linear static models, the endogeneity issue caused by unobserved heterogeneity can be dealt with by first-differencing and thereby eliminating the time-invariant heterogeneity. In dynamic and non-linear models, however, such inexpensive solutions are largely model-specific and not widely available (see Andersen (1970), Honoré (1992), Honoré and Kyriazidou (2000) and Horowitz and Lee (2004) for examples). In addition to inconsistency, a further potential statistical problem in this literature is identification of the common parameter, as mentioned by Arellano and Hahn (2007) and Arellano and Bonhomme (2011).

Originally the interest has mainly been on data characterised by a few time-series and a large number of cross-sectional observations, i.e. fixed-T large-N panels. Nevertheless, increasing availability of datasets with comparable time-series and cross-section dimensions makes large-T large-N asymptotics equally relevant.[1] There is now a growing literature where, in order to deal with the heterogeneity issue under large-T large-N asymptotics, the individual-effects are considered as individual-specific parameters to be estimated in a maximum likelihood framework.[2] However, this approach is known to be subject to the

[1] Examples of such datasets are cross-country data (Islam (1995)), growth data (Caselli, Esquivel and Lefort (1996)), firm data (e.g. studies of insider trading activity (Bester and Hansen (2009)), earnings studies (Carro (2007), Fernández-Val (2009), Hospido (2010)) and data on hedge fund returns.

[2] See Hahn and Kuersteiner (2002, 2011), Hahn and Newey (2004), Arellano and Hahn (2006), and Arellano and Bonhomme (2009).

incidental parameter issue, first studied by Neyman and Scott (1948) (see also the excellent survey by Lancaster (2000)). The main cause of this issue is that in short panels the time-series information is insufficient to accurately estimate the individual specific parameters. This inaccuracy contaminates the likelihood function and leads to a small-T bias for the estimator of the common parameter. Indeed, Arellano and Hahn (2007) note that for large-T large-N panels "*it is not less natural to talk of time-series finite sample bias than of fixed-T inconsistency or underidentification.*" This paper is in the same spirit.

To motivate the discussion, let x_{it} be the observation on some random variable of interest for individual i at time t, $i = 1, ..., N$ and $t = 1, ..., T$. For simplicity of exposition, assume for the moment that x_{it} are cross-sectionally independent. Let $L_i(\theta, \lambda_i) = L_i(\theta, \lambda_i; x_i)$ be the likelihood function for the i^{th} individual, where $x_i = (x_{i1}, ..., x_{iT})'$, θ is the common parameter and $\lambda_1, ..., \lambda_N$ are the individual-specific parameters. The concentrated likelihood estimator of λ_i is

$$\hat{\lambda}_i(\theta) = \arg\max_{\lambda_i} \ln L_i(\theta, \lambda_i).$$

If T is not sufficiently large, in the sense that the time-series information is not sufficient to estimate λ_i accurately, $\hat{\lambda}_i(\theta)$ will be subject to estimation error. This estimation error will be inherited by the corresponding concentrated likelihood function,

$$\ln L(\theta, \hat{\lambda}_1(\theta), ..., \hat{\lambda}_N(\theta)) = \sum_{i=1}^{N} \ln L_i(\theta, \hat{\lambda}_i(\theta))$$

which will be incorrectly centred. Consequently, the resulting fixed-T large-N estimator $\hat{\theta}_T = \arg\max_\theta p\lim_{N\to\infty} \ln L(\theta, \hat{\lambda}_1(\theta), ..., \hat{\lambda}_N(\theta))$ will also be biased. More importantly, even in a large-T large-N setting, this incidental parameter bias will not vanish if T is small relative to N. Formally, if $N/T \to c$ as $N, T \to \infty$ where $0 < c < \infty$, then for the large-T large-N estimator, $\hat{\theta}$, of the true parameter, θ_0, one generally has

$$\sqrt{NT}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(\sqrt{c}\beta, \Omega),$$

where β is some bounded term and Ω is some asymptotic covariance matrix. The solution then is either to ensure that N is of a negligible magnitude compared to T or to employ analytical bias reduction.

The solution offered by the analytical bias-reduction literature is based on characterising the finite-sample bias of the concentrated likelihood estimator $\hat{\theta}$ in increasing orders of $1/T$ and removing the leading $O(1/T)$ bias term. Specifically, suppose

$$\mathbb{E}[\hat{\theta} - \theta_0] = \frac{A}{T} + \frac{B}{T^2} + O\left(\frac{1}{T^3}\right),$$

for some bounded terms A and B. The leading bias term, A/T is called the first-order bias. Higher order bias terms such as B/T^2 are expected to be negligibly small as T^2 will be quite large for moderate T. However, A/T will be non-negligible even for moderate T, and therefore the objective of analytical bias reduction is to correct for this bias term and obtain a $o(T^{-1})$ small sample bias. Once A is characterised, a consistent estimator of this bias, $\hat{A} = A + o_p(1)$, can be used to obtain a first-order unbiased estimator, $\tilde{\theta} = \hat{\theta} - \hat{A}/T$. This is because,

$$\mathbb{E}\left[\hat{\theta} - \frac{\hat{A}}{T} - \theta_0\right] = o\left(\frac{1}{T}\right).$$

Important examples of previous research on correcting the first-order bias of the estimator $\hat{\theta}$ are given by Hahn and Kuersteiner (2002, 2011), Hahn and Newey (2004), Hahn and Moon (2006) and Fernández-Val (2009). Similarly, it is possible instead to bias-correct the likelihood/the objective function (Arellano and Hahn (2006), Arellano and Bonhomme (2009), Bester and Hansen (2009) and Kristensen and Salanie (2010)); or the score function/estimating equation (Woutersen (2002), Arellano (2003), Carro (2007) and Dhaene and Jochmans (2011)).[3] Of course, independent of the method used, the resulting bias-corrected estimators will be equivalent to the first order. For reviews, see Arellano and Hahn (2007) and, more recently, Arellano and Bonhomme (2011).

In a recent study, Arellano and Bonhomme (2009) consider the integrated likelihood

[3]It must be noted that analytical bias-correction methods constitute part of the literature only. The statistics literature includes many influential studies of the incidental parameter issue and possible bias-reduction methods. Two mile-stones in this area are the works by Barndorff-Nielsen (1983) and Cox and Reid (1987) who consider the modified profile and approximate conditional likelihoods, respectively. Examples of other notables contributions in the statistics literature are McCullagh and Tibshirani (1990), Firth (1993), Waterman and Lindsay (1996), Li, Lindsay and Waterman (2003) and Sartori (2003), Moreover, numerical, as opposed to analytical, corrections, such as the panel jackknife and bootstrap adjustment, can also be employed. See, for example, Hahn and Newey (2004) and Dhaene and Jochmans (2010). More recently, Bonhomme (2011) introduces the "functional differencing" approach. This is based on obtaining moment restrictions involving the parameter of interest only, which are then used to estimate θ in a GMM framework. However, compared to the study in this chapter, this last contribution is in a different spirit because it considers a fixed-T large-N framework and it is a GMM- rather than likelihood-based method.

function as a unifying framework for likelihood-based estimation. This is given by

$$\ell_i^I(\theta) = \frac{1}{T} \log \int_{\lambda_i \in \Lambda_i} L_i(\theta, \lambda_i)\, \pi_i(\lambda_i|\theta)\, d\lambda_i, \tag{3.1}$$

where $\pi_i(\lambda_i|\theta)$ is some weight or, from a Bayesian perspective, prior function. For example, if $\pi_i(\lambda_i|\theta) = 1$ for $\lambda_i = \hat{\lambda}_i(\theta)$ and zero otherwise, the resulting function is the concentrated likelihood function. Similarly, one can also obtain the random effects or Bayesian type likelihoods (see Arellano and Bonhomme (2009)). Under time-series and cross-section independence, they propose a class of weights/priors that removes the first-order bias of the score of this likelihood function; the *robust priors*. In this chapter, their analysis is extended to study the bias properties of (3.1) under serial and cross-section dependence to obtain likelihood-based general characterisations of extra bias terms. The latter type of dependence has not been considered in the analytical bias reduction literature previously. The theoretical analysis reveals that time-series dependence leads to an extra $O(T^{-3/2})$ incidental parameter bias term which is not present under serial independence. Then, without specifying an explicit structure for cross-section dependence, a flexible setting of $\sqrt{N^\rho T}$- rather than $\sqrt{NT}$-convergence is considered where $0 < \rho < 1$. In other words, double asymptotic convergence rates are assumed to be somewhere between the two polar cases of cross-section independence ($\sqrt{NT}$-convergence) and strong dependence ($\sqrt{T}$-convergence). This allows for a flexible analysis of the first-order bias. Consequently, specific conditions on ρ, under which a second type of "cross-section induced" bias of order $O(T^{-1})$ emerges, are given. This extra bias is not related to the incidental parameter issue and so, it has to be corrected for separately. It is also shown that, if the cross-section dimension is allowed to contribute to convergence at a mild rate, the cross-section dependence bias becomes $O(T^{-3/2})$. Then, crucially, it can be shown that the $O(T^{-1})$ bias is identical to the one characterised by Arellano and Bonhomme (2009) and their robust priors can be used to reduce the magnitude of the small-sample bias to $O(1/T^{3/2})$. The analysis in this chapter is the main contribution of this thesis to the panel data literature. Put differently, this theoretical analysis finds the conditions under which the Arellano-Bonhomme robust priors can still be used in the presence of time-series and cross-section dependence.

Characterising cross-section dependence in terms of slower double-asymptotic conver-

gence rates is technically convenient. However, from an applied perspective, it would be desirable to extend the analysis to standard dependence structures used in the literature. Therefore, in the final part of this chapter, an initial step in this direction is taken by considering a spatial dependence/clustering framework. This setting is based on the idea that individuals are weakly dependent across cross-section, in a similar fashion to mixing sequences in the time-series literature. It is shown that under appropriate assumptions on the growth rate of cluster sizes as the sample size grows, cross-section dependence does not lead to extra small-sample bias. This suggests interesting avenues for future research.

The rest of this study is organised as follows: Section 3.2 introduces the notation and briefly discusses relevant concepts. Key assumptions and main theoretical results are given in Section 3.3. The analysis of bias under time-series dependence is treated in Section 3.4 while Section 3.5 extends the analysis to cross-section dependence. Section 3.6 considers a short analysis of a spatial dependence/clustering based extension. Section 5 concludes. All proofs are given in the Mathematical Appendix.

3.2 Main Concepts and Notation

Following the convention as in e.g. Arellano and Hahn (2006) and Hahn and Kuersteiner (2011), define some random variable x_{it} indexed by individuals, i, and time, t where $i = 1, ..., N$ and $t = 1, ..., T$. Let θ be the P-dimensional common parameter of interest and λ_i be the scalar individual-specific parameter for the i^{th} individual. The corresponding (pseudo) true parameter values are given by θ_0 and λ_{i0}. Let, furthermore, $\varphi_{it}(\theta, \lambda_i) = \varphi(\theta, \lambda_i; x_{it})$ be some criterion function. This setting is general in the sense that one can consider a variable of interest y_{it} such that $x_{it} = y_{it}$ and $\varphi_{it}(\theta, \lambda_i) = \ell(\theta, \lambda_i; x_{it}) = \ell(\theta, \lambda_i; y_{it})$ or $x_{it} = (y_{it}, y_{i,t-1}, ..., y_{i,t-q})$ and $\varphi_{it}(\theta, \lambda_i) = \ell(\theta, \lambda_i; y_{it}|y_{i,t-1}, ..., y_{i,t-q})$ where $\ell(\cdot)$ is the (conditional) log-likelihood function (Arellano and Hahn (2006)).

Under scrutiny is the following estimator:

$$(\hat{\theta}, \hat{\lambda}_1, ..., \hat{\lambda}_N) = \arg \max_{\theta, \lambda_1, ..., \lambda_n} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \varphi_{it}(\theta, \lambda_i). \tag{3.2}$$

The first and foremost assumption is that $\varphi_{it}(\cdot)$ is an appropriate function in the sense that for $T \to \infty$ and N fixed, $(\hat{\theta}, \hat{\lambda}_1, ..., \hat{\lambda}_N)$ is consistent for $(\theta_0, \lambda_{10}, ..., \lambda_{N0})$. In other words, the

researcher already has a valid, consistent estimator available. The problem is that the sample is not large enough and the estimator is prone to small-sample bias. Importantly, in the case of cross-section dependence, (3.2) has a composite likelihood function interpretation, where the joint density is approximated by the average of marginal densities. As explained in Remark 2.3.1, the consistency result developed for the two-step estimator of Chapter 2 can easily be generalised to the framework of this chapter. Therefore, the analysis of this chapter can also be considered as an investigation of the incidental parameter bias under neglected cross-section dependence.

This setting has a general scope because it is based on a generic (possibly pseudo) likelihood function, with no particular model in mind. As such the objective function may exhibit non-linearities and/or a dynamic structure. Therefore, the bias characterisations and asymptotic expansions derived in this paper potentially apply to a wide array of non-linear dynamic panel models and provide important insights into bias reduction under cross-section dependence. Of course, when the objective function is based on a pseudo-likelihood, some careful thinking might be required on a case-by-case basis. For example, the quasi maximum likelihood theory for the GARCH model is well-established (Bollerslev and Wooldridge (1992)) but for a different non-linear dynamic model there may be issues.

The rest of the analysis will be based on the likelihood notation, without loss of generality. Define

$$\begin{aligned} \ell_{iT}(\theta,\lambda_i) &= \frac{1}{T}\sum_{t=1}^{T}\ell_{it}(\theta,\lambda_i), \quad \ell_{NT}(\theta,\lambda)=\frac{1}{N}\sum_{i=1}^{N}\ell_{iT}(\theta,\lambda_i), \\ \ell_{iT}^{\lambda}(\theta,\lambda_i) &= \frac{\partial\ell_{iT}(\theta,\lambda_i)}{\partial\lambda_i}, \quad \ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)=\frac{\partial^2\ell_{iT}(\theta,\lambda_i)}{\partial\lambda_i^2} \quad \text{etc.} \end{aligned}$$

Hence, λ appearing as a superscript denotes differentiation with respect to λ_i. The operator $\nabla_{\theta^{(k)}}$ is used to take the k^{th} order total derivative with respect to θ. For example,

$$\nabla_{\theta^{(2)}}\ell_{iT}(\theta,\lambda_i)=\frac{d^2\ell_{iT}(\theta,\lambda_i)}{d\theta d\theta'}, \quad \nabla_{\theta^{(2)}}\ell_{iT}^{\lambda}(\theta,\lambda_i)=\frac{d^2\ell_{iT}^{\lambda}(\theta,\lambda_i)}{d\theta d\theta'} \quad \text{etc.}$$

The centred likelihood derivatives with respect to λ_i are defined as

$$V_{iT}^{\lambda\lambda}(\theta,\lambda_i)=\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)-\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)], \quad V_{iT}^{\lambda\lambda\lambda}(\theta,\lambda_i)=\ell_{iT}^{\lambda\lambda\lambda}(\theta,\lambda_i)-\mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}(\theta,\lambda_i)] \quad \text{etc.}$$

Unless otherwise noted, all expectations are taken with respect to the underlying true density evaluated at $(\theta_0, \lambda_{10}, ..., \lambda_{N0})$.

The bias-reduction analysis utilises three different likelihood functions. These are the concentrated, integrated and target likelihood functions. The most familiar of these is the concentrated likelihood, given by

$$\ell^c_{iT}(\theta) = \ell_{iT}(\theta, \hat{\lambda}_i(\theta)),$$

$$\text{where} \quad \hat{\lambda}_i(\theta) = \arg\max_{\lambda_i} \sum_{t=1}^{T} \ell_{it}(\theta, \lambda_i) \quad \text{and} \quad \hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T} \ell_{it}(\theta, \hat{\lambda}_i(\theta)).$$

The main idea is to centre the likelihood function at the likelihood estimator for λ_i, for some given value of θ. In large samples this estimator has good properties.[4] However, when T is not sufficiently large, $\hat{\lambda}_i(\theta)$ is estimated with error. As a result, the likelihood is concentrated with respect to a biased value for λ_{i0}. Crucially, the estimation error (or the small-sample bias) in $\hat{\lambda}_i(\theta)$ is accumulated across strata, and contaminates the estimation of θ_0 (see, e.g., McCullagh and Tibshirani (1990) and Sartori (2003)). Consequently, $\hat{\theta}$ is inconsistent for θ_0. More formally, $\hat{\theta}_T = \arg\max_\theta p\lim_{N\to\infty}(NT)^{-1}\ell_{NT}(\theta, \hat{\lambda}_i(\theta)) \neq \theta_0$. This is the well-known incidental parameter issue, first investigated by Neyman and Scott (1948). In the case of the dynamic autoregressive panel model, this is also widely known as the Nickell bias due to Nickell (1981).

A possible solution is to integrate λ_i out from the density function and to obtain a new density, free of the nuisance parameter. This is the integrated likelihood approach which, for a given weighting scheme $\pi_i(\lambda_i|\theta)$, returns

$$\ell^I_{iT}(\theta) = \frac{1}{T} \ln \int \exp\left[T\ell_{iT}(\theta, \lambda_i)\right] \pi_i(\lambda_i|\theta)\, d\lambda_i.$$

The choice of weights/priors, $\pi_i(\lambda_i|\theta)$, is key to successfully removing the incidental parameter bias. Following Arellano and Bonhomme (2009), who investigate this method in the case of non-linear dynamic panel models under time-series and cross-section independence, a *robust prior* is defined as the prior that removes the first-order bias of the profile score. Specification of these robust priors is the essence of this study.

[4] See Barndorff-Nielsen and Cox (1994) and Severini (2000) for excellent textbook treatments of likelihood based estimation.

One might be tempted to think that, since $\ell_{iT}^{I}(\theta)$ is not a function of λ_i anymore, it does not suffer from the incidental parameter bias. However, if the correct, or robust, weighting scheme is not employed, then the resulting likelihood function will still be wrongly centred. To give an example, observe that if

$$\pi_i\left(\lambda_i|\theta\right) = \begin{cases} 1 & \text{for} \quad \lambda_i = \hat{\lambda}_i(\theta) \\ 0 & \text{for} \quad \lambda_i \neq \hat{\lambda}_i(\theta) \end{cases},$$

then $\ell_{iT}^{I}(\theta)$ is still free of λ_i but it coincides with $\ell_{iT}(\theta, \hat{\lambda}_i(\theta))$, the concentrated likelihood function, which is incorrectly centred.

It must be underlined that this study follows a frequentist approach in the sense that the specification of the robust prior depends entirely on the characterisation of the incidental parameter bias. Therefore, no subjective prior has to be specified and one can indeed refer to the robust prior as a robust weighting scheme. By way of analogy, $\pi_i\left(\lambda_i|\theta\right)$ is used as a tool (or as a "vacuum cleaner") to mop up the first-order bias of the integrated likelihood function. It is of course also possible to use a subjective prior, but this approach is not pursued here. Indeed, bias correction by integrated likelihood is a common approach in the Bayesian literature. More importantly, recent research reveals that there are important links between integrated likelihood estimation and the traditional bias reduction approaches employed in the frequentist literature. In particular, Severini (1999) shows that the adjusted profile likelihood function (Cox and Reid (1987)) is third-order asymptotically Bayes. Moreover, he also mentions that since the profile log-likelihood and the modified profile log-likelihood (Barndorff-Nielsen (1983)) functions are locally equivalent to second order, the latter is asymptotically Bayesian to second order, as well. The adjusted and modified profile log-likelihood functions are important contributions in the frequentist bias reduction literature and therefore, these observations imply a natural connection between the frequentist and Bayesian approaches. In addition, Severini (2007) also provides an analysis of the issue of selecting the priors that would ensure that the integrated likelihood is appropriate under the frequentist approach, as well. Finally, in a recent work, Severini (2010) investigates the integrated log-likelihood ratio statistic and compares it to the standard log-likelihood ratio statistic. All these contributions provide strong theoretical justification

for the use of the integrated likelihood method within the frequentist framework.

The aim then is to correct the bias of the integrated likelihood function with respect to an appropriate benchmark function. This benchmark is given by the target likelihood function,

$$\ell_{iT}(\theta, \bar{\lambda}_{iT}(\theta)), \quad \text{where} \quad \bar{\lambda}_{iT}(\theta) = \arg\max_{\lambda_i} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\theta_0, \lambda_{i0}}[\ell_{it}(\theta, \lambda_i)] \quad \text{for some fixed } \theta.$$

Here, $\mathbb{E}_{\theta_0, \lambda_{i0}}[\cdot]$ is the expectation based on the underlying density evaluated at θ_0 and λ_{i0}. This is an appropriate benchmark as the curve defined by $\left(\theta, \bar{\lambda}_{iT}(\theta)\right)$ is referred to as the "least favourable curve" in the parameter space, after Stein (1956). In the likelihood setting, this is because the expected information for θ, obtained by using $\ell_{it}(\theta_0, \bar{\lambda}_{iT}(\theta_0))$, is equal to the partial expected information. The latter, in turn, coincides with the inverse of the Cramér-Rao lower bound. Hence, the target likelihood used here is the "least favourable" benchmark to compare the concentrated likelihood to.[5] Importantly, this is an infeasible benchmark as $\bar{\lambda}_{iT}(\theta)$ is based on θ_0 and λ_{i0} (through calculation of the expectation), as well as θ. Nevertheless, it still is a useful benchmark to analyse the theoretical properties of the incidental parameter bias.

In what follows, the following notational convention will be used for sake of conciseness: $\hat{\lambda}_i$ and $\bar{\lambda}_i$ are used as shorthand for $\hat{\lambda}_i(\theta)$ and $\bar{\lambda}_{iT}(\theta)$; therefore, the dependence of $\bar{\lambda}_{iT}(\theta)$ on T will be implicit. Moreover, whenever a likelihood function is evaluated at $(\theta, \bar{\lambda}_i(\theta))$, the argument will be omitted. Also, if the likelihood is evaluated at $(\psi, \bar{\lambda}_i(\psi))$ for some $\psi \neq \theta$, then the likelihood is written as a function of ψ only. Specifically,

$$\ell_{it} = \ell_{it}(\theta, \bar{\lambda}_i(\theta)), \quad \ell_{iT} = \ell_{iT}(\theta, \bar{\lambda}_i(\theta)), \quad \ell_{NT} = \ell_{NT}(\theta, \bar{\lambda}(\theta)),$$
$$\ell_{it}(\theta_0) = \ell_{it}(\theta_0, \bar{\lambda}_i(\theta_0)), \quad \ell_{iT}(\theta_0) = \ell_{iT}(\theta_0, \bar{\lambda}_i(\theta_0)), \quad \ell_{NT}(\theta_0) = \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0)),$$

where $\bar{\lambda}(\theta) = (\bar{\lambda}_1(\theta), ..., \bar{\lambda}_N(\theta))$. The same applies to functions such as $V_{iT}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta))$, $\ell_{NT}^{\lambda}(\theta, \bar{\lambda}_i(\theta))$ etc. Lastly, $\mathbb{E}[\cdot]$ and $Var(\cdot)$ are used as shorthand for $\mathbb{E}_{\theta_0, \lambda_{i0}}[\cdot]$ and $Var_{\theta_0, \lambda_{i0}}(\cdot)$, the expectation and variance evaluated at the true parameter values, respectively.

[5] See Severini and Wong (1992) and Severini (2000, Chapter 4). A lucid discussion is given by Pace and Salvan (2006). In particular, the target likelihood is a proper likelihood and is maximised at θ_0.

3.3 Assumptions

As outlined in Section 3.1, the bias reduction strategy is based on obtaining an analytical expression for the incidental parameter bias of order $O(T^{-1})$. To do this, first the small sample bias of the integrated likelihood function with respect to the benchmark infeasible target likelihood functions is derived, by using large-T asymptotic expansions. Based on this, it is straightforward to obtain the characterisation of the robust prior which removes the first-order bias of the score function. The eventual objective, of course, is to correct the bias of the integrated likelihood estimator. In the cross-section independence case, it is known that both bias-corrected likelihood and bias-corrected score functions imply bias corrected estimators (Arellano and Hahn (2007)). However, as will be shown in Section 3.5, it is possible that this does not hold anymore in the presence of cross-section dependence. Nevertheless, from an intuitive perspective, it makes more sense to attack the bias of the score function first (rather than the bias of the estimator itself) as this is the process which produces the estimator.

Arellano and Bonhomme (2009) have already studied the time-series and cross-section independence case, so results presented in this section extend their analysis to time-series dependence, which is assumed to be of mixing type. The definitions for two common mixing types are given next.

Definition 3.3.1 (α- and ϕ-mixing) *Define the σ-fields, $\mathcal{G}_t^i = \sigma(x_{it}, x_{i,t-1}, ...)$ and $\mathcal{H}_t^i = \sigma(x_{it}, x_{i,t+1}, ...)$. Define also,*

$$\begin{aligned}
\alpha_i(m) &= \alpha_i(\mathcal{G}_t^i, \mathcal{H}_{t+m}^i) = \sup_t \sup_{G \in \mathcal{G}_t^i \text{ and } H \in \mathcal{H}_{t+m}^i} |P(G \cap H) - P(G)P(H)|, \\
\phi_i(m) &= \phi_i(\mathcal{G}_t^i, \mathcal{H}_{t+m}^i) = \sup_t \sup_{G \in \mathcal{G}_t^i,\ P(G)>0 \text{ and } H \in \mathcal{H}_{t+m}^i} |P(H|G) - P(H)|.
\end{aligned}$$

Then, the sequence of random vectors $(x_{it}, x_{i,t-1}, x_{i,t-2}, ...)$ is called

$$\begin{aligned}
\alpha\text{-mixing if } \alpha(m) &\to 0 \text{ as } m \to \infty, \\
\phi\text{-mixing if } \phi(m) &\to 0 \text{ as } m \to \infty.
\end{aligned}$$

Moreover, for $s \in \mathbb{R}$, if $\alpha(m) = O(m^{-s-\epsilon})$ for some $\epsilon > 0$, then s is said to be of size

$-s$ (and similarly for ϕ).

The mixing concept is commonly used to impose weak dependence on economic time-series.[6] Intuitively, a mixing sequence resembles an independently distributed time-series, as the observations become more and more distant across time. For example, although the returns on a given stock observed on two consecutive days will exhibit considerable dependence, the dependence between observations one week apart will usually be much weaker.

A convenient property of mixing sequences is given next, which is used frequently in the proofs.

Theorem 3.3.2 (Theorem 14.1 (Davidson 1994), Theorem 3.49 (White 2001)) *Let $g(\cdot)$ be a measurable function and let $y_{it} = g(x_{it}, x_{i,t-1}, ..., x_{i,t-\tau})$ for finite τ. If the sequence $(x_{it}, x_{i,t-1}, x_{i,t-2}, ...)$ is α-mixing (ϕ-mixing) of size $-s$, $s > 0$, then $(y_{it}, y_{i,t-1}, y_{i,t-2}, ...)$ is also α-mixing (ϕ-mixing) of size $-s$.*

The assumptions are given next. In what follows, $j_1, ..., j_k \in \{1, 2, ..., P\}$.

Assumption 3.3.1 *$N, T \to \infty$ jointly and, for $0 < c < \infty$, $N/T \to c$.*

Assumption 3.3.2 *$\{x_{it}\}$ is an $\alpha-$mixing sequence for each i. Moreover, for all i and t the mixing coefficients are of size $-r/(r-2)$ for some $r > 2$.*

Assumption 3.3.3 *$\ell_{it}(\theta, \lambda_i) \in \mathcal{C}^8$ for all i, t, where $\mathcal{C}^c$ is the class of functions whose derivatives up to and including order c are continuous.*

Assumption 3.3.4 *The support of $\pi_i(\lambda_i|\theta)$ contains an open neighbourhood of the true parameters λ_{i0} and θ_0.*

Assumption 3.3.5 *θ and λ_i belong to the interior of Θ and Λ_i, respectively, where $\Theta \subseteq \mathbb{R}^P$ and $\Lambda_i \subseteq \mathbb{R}$ are compact and convex parameter spaces.*

Assumption 3.3.6 *For each $\theta \in \Theta$, $\ell_{iT}(\theta, \lambda_i)$ has a unique maximum at $\hat{\lambda}_i(\theta)$ for all i.*

[6] See Davidson (1994) and White (2001) for a detailed treatment of mixing sequences from an econometric perspective. A recent survey, which includes many other types of mixing processes, is given by Bradley (2005). The classical reference is Doukhan (1994).

Assumption 3.3.7 *As* $T \to \infty$, $\sup_i \sup_{\theta \in \Theta} \left| \hat{\lambda}_i(\theta) - \bar{\lambda}_{iT}(\theta) \right| = O_p(T^{-1/2})$.

Assumption 3.3.8 *Define*

$$\mathcal{Z}_{it}^{m,k}(\theta, \lambda_i) = \frac{d^{(m+k)}}{d\lambda_i^m d\theta_{j_1} ... d\theta_{j_k}} \ell_{it}(\theta, \lambda_i).$$

For all combinations of $m \in \{0, 1, 2, 3\}$ *and* $k \in \{0, 1, 2, 3, 4, 5\}$, *there exist individual functions* $M_{it}(\theta)$, *possibly dependent on* x_{it} *and* θ, *such that*

$$\left| \mathcal{Z}_{it}^{m,k}(\theta, \bar{\lambda}_i(\theta)) \right| \leq M_{it}(\theta),$$

where $\sup_{i,t} \sup_{\theta \in \Theta} M_{it}(\theta) < \infty$. *Moreover,* $\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta))]$ *and* $\mathbb{E}[\nabla_{\theta^{(2)}} \ell_{NT}(\theta, \bar{\lambda}_i(\theta))]$ *are non-singular for all* i, T, N *and* $\theta \in \Theta$.

Assumption 3.3.9 *For all combinations of* $m \in \{0, 1\}$ *and* $k \in \{0, 1, 2, 3, 4\}$ *there exist individual functions* $H_{i,T}(\theta)$, *possibly dependent on* $\{x_{i1}, ..., x_{iT}\}$ *and* θ, *such that*

$$\left| \frac{d^{(m+k)}}{d\lambda_i^m d\theta_{j_1} ... d\theta_{j_k}} \ln \pi_i(\bar{\lambda}_{iT}(\theta)|\theta) \right| \leq H_{i,T}(\theta),$$

where $\sup_{i,T} \sup_{\theta \in \Theta} H_{i,T}(\theta) < \infty$.

Assumption 3.3.10 *Define the zero-mean random variables* $\bar{\mathcal{Z}}_{it}^{m,k}(\theta, \lambda_i) = \mathcal{Z}_{it}^{m,k}(\theta, \lambda_i) - \mathbb{E}[\mathcal{Z}_{it}^{m,k}(\theta, \lambda_i)]$. *For all combinations of* $m \in \{1, 2\}$ *and* $k \in \{0, 1, 2, 3, 4\}$

$$\sup_{i,t} \sup_{\theta \in \Theta} \mathbb{E} \left| \mathcal{Z}_{it}^{m,k}(\theta, \bar{\lambda}_i(\theta)) \right|^r < \infty,$$

and

$$\inf_i \inf_\theta Var \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathcal{Z}_{it}^{m,k}(\theta, \bar{\lambda}_i(\theta)) \right) > 0 \quad \text{as } T \to \infty.$$

The same also holds for $(m, k) = (3, 0)$.

Assumption 3.3.11 *As* $N, T \to \infty$,

$$\nabla_\theta \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0)) = O_p\left(\frac{1}{\sqrt{N^{\rho_1} T}} \right),$$
$$\nabla_{\theta^{(2)}} \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0)) - \mathbb{E}[\nabla_{\theta^{(2)}} \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0))] = O_p\left(\frac{1}{\sqrt{N^{\rho_2} T}} \right),$$

$$\nabla_{\theta^{(3)}}\ell_{NT}(\theta_0,\bar{\lambda}(\theta_0)) - \mathbb{E}[\nabla_{\theta^{(j)}}\ell_{NT}(\theta_0,\bar{\lambda}(\theta_0))] = O_p\left(\frac{1}{\sqrt{N^{\rho_3}T}}\right),$$

where $0 \leq \rho_1, \rho_2, \rho_3 \leq 1$.

Assumption 3.3.1 implies that N and T converge to infinity at the same rate, hence a large-T large-N setting is considered. This would, for example, be appropriate for financial panels where T and N are of comparable magnitudes. In addition, as mentioned previously, this setting has also been considered frequently for microeconometric applications.

Assumption 3.3.2 characterises the structure of time-series dependence. As stated in Theorem 3.3.2, any measurable function of a finite sequence of x_{it} will retain all mixing and size properties of x_{it}. Moreover, it is also well-known that continuous functions are measurable (see, for example, Theorem 13.2 in Billingsley (1995)). By Assumption 3.3.3, all derivatives of the objective function up to and including the eighth order are guaranteed to exist and to be continuous.[7] Consequently, Assumptions 3.3.2 and 3.3.3 together imply that all such likelihood derivatives are α-mixing and are of the same size as $\{x_{it}\}$. The virtue of this result is that mixing LLNs and CLTs can be applied to (properly normalised) averages of the derivatives of the objective function, which appear naturally in asymptotic expansions. The existence of these LLNs and CLTs is ensured by Assumption 3.3.10 which states that the necessary moment conditions hold. The proofs for the results given in Section 3.4 use this property heavily. It is important to underline that this idea would generalise to any type of mixing, although different types of mixing would require different moment conditions. Hence, α-mixing is assumed for illustration purposes only and the proofs can be adapted to a different type of mixing.

Assumption 3.3.4 rules out cases where the prior is not defined at the true parameter values, (λ_{i0}, θ_0). In other words, the possibility of the integrated likelihood not being defined at the true parameter values is assumed away. Assumption 3.3.5 is a standard regularity condition on the parameter space. Assumption 3.3.6 is required for the existence of a Laplace approximation to the integrated likelihood function and is a mild condition. The Laplace approximation is used to obtain a linear approximation to the integral. Assumption 3.3.7 controls the convergence rate of $\hat{\lambda}_i(\theta)$ to $\bar{\lambda}_{iT}(\theta)$, the benchmark "least-favourable" estimator.

[7] This is standard and would generally be assumed implicitly.

Intuitively, this ensures that the concentrated likelihood estimator is never "too" far away from the benchmark estimator. Assumption 3.3.8 simply guarantees that the likelihood derivatives that appear in the expansions exist and are finite. The parallel conditions regarding the prior function are given in Assumption 3.3.9.

Assumption 3.3.11 is the most important assumption that implicitly characterises the nature of cross-section dependence. Therefore, this is a good moment to elaborate on how cross-section dependence is integrated into the analysis. First, observe that all expressions in Assumption 3.3.11 are likelihood derivatives and are zero mean (either by themselves or by centering). By the previous heuristic discussion, the time-series averages of the considered likelihood derivatives will all be $O_p(T^{-1/2})$ due to existence of mixing CLTs. However, when these terms are averaged across cross-section as well, it is not clear what the convergence rate will be. One idea is to consider two polar cases:

Case 3.3.3 *Cross-section independence, which, under standard regularity conditions, implies $\sqrt{NT}$-convergence.*

Case 3.3.4 *Cross-section dependence of strong form, such that there is no gain from cross-section size. Hence, $\sqrt{T}$-convergence.*

Case 3.3.3 is the setting considered invariably in the analytical bias reduction literature. Case 3.3.4, on the other hand, is the worst-case scenario, where cross-section dependence can be so strong that the cross-section size does not contribute to rate of convergence. A simple example is the extreme case where all individuals in the panel are identical. Clearly, in terms of the information it contains, this panel is identical to a single time-series implying $\sqrt{T}$ convergence. In reality this will not be the case, but cross-section dependence can still be so strong that each new individual brings a minimal amount of new information. Following Engle, Shephard and Sheppard (2008), who also use this approach to characterise cross-section dependence, a different way to put this would be to say that there is no LLN in the cross-section. Of course, in practice one would expect at least some contribution from N. However, $\sqrt{T}$-convergence provides a good benchmark to understand whether and how cross-section dependence might affect the first-order bias if worst comes to worst. This idea is integrated into the analysis by allowing the convergence rates for the cross-sectional averages of likelihood terms to be defined in a flexible way, where $0 \leq \rho_1, \rho_2, \rho_3 \leq 1$. On the

one hand, ρ_1, ρ_2 and ρ_3 can be equal to zero, implying $\sqrt{T}$-convergence for all terms. On the other hand, the usual $\sqrt{NT}$ rate is achieved for these terms when $\rho_1 = \rho_2 = \rho_3 = 1$. A discussion of other possibilities for modelling cross-section dependence is given in Section 3.5.

3.4 Bias of the Integrated Likelihood

The first main result of this study, which characterises the bias of the integrated likelihood function under time-series dependence is presented below.

Theorem 3.4.1 *Under Assumptions 3.3.1-3.3.8,*

$$\mathbb{E}_{\theta_0,\lambda_{i0}}\left[\ell_{iT}^{I}(\theta) - \bar{\ell}_{iT}(\theta)\right] = C + \frac{\mathcal{B}_{iT}^{(1)}(\theta)}{T} + \frac{\mathcal{B}_{iT}^{(2)}(\theta)}{T^{3/2}} + O\left(\frac{1}{T^2}\right), \tag{3.3}$$

where

$$\begin{aligned}
\mathcal{B}_{iT}^{(1)}(\theta) &= \frac{1}{2}\{\mathbb{E}_{\theta_0,\lambda_{i0}}[-\ell_{iT}^{\lambda\lambda}]\}^{-1}\mathbb{E}_{\theta_0,\lambda_{i0}}[T(\ell_{iT}^{\lambda})^2] \\
&\quad -\frac{1}{2}\ln\mathbb{E}_{\theta_0,\lambda_{i0}}[-\ell_{iT}^{\lambda\lambda}] + \ln\pi_i(\bar{\lambda}_i|\theta), \qquad (3.4)\\
\mathcal{B}_{iT}^{(2)}(\theta) &= T^{3/2}\frac{1}{2}\frac{\mathbb{E}_{\theta_0,\lambda_{i0}}\left[V_{iT}^{\lambda\lambda}(\ell_{iT}^{\lambda})^2\right]}{\{\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda}]\}^2} - T^{3/2}\frac{1}{6}\frac{\mathbb{E}_{\theta_0,\lambda_{i0}}[(\ell_{iT}^{\lambda})^3]\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda\lambda}]}{\{\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda}]\}^3}, \qquad (3.5)
\end{aligned}$$

and $C = (2T)^{-1}\ln\left(2\pi T^{-1}\right)$.

Several remarks are in order. Notice that by standard arguments $\mathcal{B}_{i}^{(1)}(\theta)$ and $\mathcal{B}_{i}^{(2)}(\theta)$ are both $O(1)$. Theorem 3.4.1 is different from the corresponding Theorem 1 in Arellano and Bonhomme (2009) in that the bias of the integrated likelihood includes an extra $O(T^{-3/2})$ term given by $\mathcal{B}_{iT}^{(2)}(\theta)/T^{3/2}$. This is due to the presence of time-series dependence. If, on the other hand, the likelihood derivatives are independent across t, then, $\mathcal{B}_{iT}^{(2)}(\theta)/T^{3/2}$ is actually $O(T^{-2})$.[8] Another contribution is the derivation of a likelihood based characterisation of this extra bias term, given in (3.5).

It must be underlined that, as far as first-order bias reduction is concerned, any $o(T^{-1})$ bias term would be considered negligible in the literature. So, presence of the extra $O(T^{-3/2})$ does not pose extra difficulties in terms of bias reduction. However, for higher order bias

[8] A proof of this result, which is not novel, is given in Proposition 3.A.3 in the Mathematical Appendix in Section 3.A.3.

correction, these results would be very useful. Note that all bias terms involve expectations calculated at the true parameter values, which can easily be estimated by using the sample means.

The more interesting feature of Theorem 3.4.1 is that the first order bias term, $\mathcal{B}_{iT}^{(1)}(\theta)/T$ is identical to the one found by Arellano and Bonhomme (2009). Then, given that the sole interest is in correcting the $O(T^{-1})$ bias, this result implies that the robust priors suggested by Arellano and Bonhomme (2009) are still valid under time-series dependence, assuming cross-section independence for the moment.

The robust priors are obtained by choosing $\pi_i(\lambda_i|\theta)$ in such a way that it cancels the other bias terms. By analogy, it can be likened to a "vacuum cleaner" which is used to "clean" the bias terms. By taking the derivative of (3.3) with respect to θ, one can derive an expression for $\pi_i(\bar{\lambda}_i|\theta)$ that removes the first order bias of the score. The specifications of the bias-reducing priors that correct the $O(T^{-1})$ bias term only and both the $O(T^{-1})$ and $O(T^{-3/2})$ bias terms are given below.

Corollary 3.4.2 *The robust prior that cancels the bias term of order $O(T^{-1})$ only is given by*

$$\pi_i^R(\lambda_i|\theta) \propto \widehat{\mathbb{E}}[-\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)]\left(\widehat{\mathbb{E}}\{[\ell_{iT}^{\lambda}(\theta,\lambda_i)]^2\}\right)^{-1/2} \tag{P1}$$

which is valid in a likelihood setting while

$$\begin{aligned}\pi_i^R(\lambda_i|\theta) \quad \propto \quad & \{\widehat{\mathbb{E}}[-\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)]\}^{1/2} \\ & \times \exp\left(-\frac{T}{2}\{\widehat{\mathbb{E}}[-\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)]\}^{-1}\widehat{\mathbb{E}}\{[\ell_{iT}^{\lambda}(\theta,\lambda_i)]^2\}\right),\end{aligned} \tag{P2}$$

is valid in pseudo-likelihood settings, as well. Under the same assumptions, the specification of the robust prior that cancels bias terms of order both $O(T^{-1})$ and $O(T^{-3/2})$ is given by

$$\begin{aligned}\pi_i^R(\lambda_i|\theta) \quad \propto \quad & \{\widehat{\mathbb{E}}[-\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)]\}^{1/2} \\ & \times \exp\left[-\frac{T}{2}\left(\frac{\widehat{\mathbb{E}}\{[\ell_{iT}^{\lambda}(\theta,\lambda_i)]^2\}}{\widehat{\mathbb{E}}[-\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)]}\right.\right. \\ & \left.\left.+\frac{\sqrt{T}\widehat{\mathbb{E}}\left[V_{iT}^{\lambda\lambda}(\ell_i^{\lambda})^2\right]}{\{\widehat{\mathbb{E}}[-\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)]\}^2}-\frac{1}{3}\frac{\sqrt{T}\widehat{\mathbb{E}}[(\ell_{iT}^{\lambda})^3]\widehat{\mathbb{E}}[\ell_{iT}^{\lambda\lambda\lambda}]}{\{\widehat{\mathbb{E}}[-\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)]\}^3}\right)\right].\end{aligned} \tag{P2*}$$

Priors (P1), (P2) and (P2*) follow directly from Theorem 3.4.1. In particular, the derivation of Priors (P1) and (P2) is given in Arellano and Bonhomme (2009). In addition, proofs of (P2) and (P2*) are analogous and follow by simple inspection. See Arellano and Bonhomme (2009) for the proofs. Derivation of Prior (P1) is slightly more involved as it relies on a simplification by Pace and Salvan (1996) based on the information equality, which holds under correct parametric assumptions only. Therefore, Prior (P1) is valid in a likelihood setting while Priors (P2) and (P2*) are more suitable for empirical analysis where parametric assumptions are not guaranteed to be correct.

Finally, note that (3.3) is a large-T expansions for fixed i. Therefore, cross-section dependence has not come into play yet. To obtain the first-order bias of the integrated likelihood estimator, a double asymptotic expansions letting $N \to \infty$, as well, is needed. This is done next.

3.5 Bias of the Integrated Likelihood Estimator under Cross-Section Dependence

The case of cross-section dependence has so far not been analysed in the analytical bias-reduction literature.[9] The question of interest is whether cross-section dependence introduces extra bias terms in addition to $\mathcal{B}_{iT}^{(1)}(\theta)/T$. As discussed in Section 3.3, under cross-section independence one would, under regularity conditions, usually have $\sqrt{NT}$-convergence. However, when cross-section dependence is present, convergence will most likely be at a slower rate. This idea manifests itself in Assumption 3.3.11 in the form of zero-mean likelihood derivatives converging at (possibly) slower rates.

The implication of slower convergence rates on the mechanics of bias derivations is that higher order expansions are required in order to characterise the bias. More specifically, under $\sqrt{NT}$-convergence, at most second order Taylor expansions are sufficient for bias derivations. Here, on the other hand, fourth order expansions for estimators of both λ_i and θ have to be obtained in order to characterise the first-order bias. This is because the terms that appear in expansions converge at a slower rate and so higher order expansions are required to characterise the small sample behaviour up to the $O(T^{-2})$ remainder term.

[9] One important exception is the work by Phillips and Sul (2007) who consider the specific case of a dynamic autoregressive panel model under neglected cross-section dependence and calculate the probability limit of the dynamic parameter. Hence, their analysis extends the Nickell (1981) bias.

An indirect contribution of this study then is derivation of higher order likelihood-based expansions, which can be used for purposes other than bias reduction.

Remember that θ is P-dimensional. This introduces no conceptual difficulties, but it makes the algebra of the asymptotic expansions more complicated. This is because likelihood derivatives with respect to θ are now possibly multi-dimensional arrays. To overcome this issue, the proofs are based on the index notation and the Einstein summation convention. Basically, these notational conventions allow multi-dimensional arrays to be manipulated algebraically in the same way as scalars. The final result then can be translated into matrix notation. An overview of these techniques is provided in the Mathematical Appendix.

The following definitions are used in characterising the bias of the integrated likelihood estimator:

$$\hat{\theta}_{IL} = \arg\max_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \ell_{iT}^{I}(\theta), \qquad S = \nabla_\theta \ell_{NT}(\theta_0) = \left\{ \ell_{NT}^{\theta}(\theta_0) - \frac{\mathbb{E}[\ell_{NT}^{\lambda\theta}(\theta_0)]}{\mathbb{E}[\ell_{NT}^{\lambda\lambda}(\theta_0)]} \ell_{NT}^{\lambda}(\theta_0) \right\},$$

$$H = \nabla_{\theta\theta}\ell_{NT}(\theta_0), \qquad \nu = \mathbb{E}[H],$$

$$Z_j = \mathbb{E}\left[\nabla_{\theta\theta} \frac{d\ell_{NT}(\theta_0)}{d\theta_j}\Big|_{\theta=\theta_0} \right], \quad \text{and} \quad M = \begin{bmatrix} S'\nu^{-1}Z_1\nu^{-1}S \\ \vdots \\ S'\nu^{-1}Z_D\nu^{-1}S \end{bmatrix},$$

where $j \in \{1, ..., P\}$. Elsewhere in the literature, S is also referred to as the projected score. H is the Hessian matrix with respect to θ while Z is related to the third-order derivatives. The bias of the integrated likelihood estimator is given in the next theorem, which is the second main theoretical contribution of this study.

Theorem 3.5.1 *Under Assumptions 3.3.1-3.3.11*

$$\begin{aligned}
(\hat{\theta}_{IL} - \theta_0) \;=\; & -\nu^{-1}S \\
& -\nu^{-1}\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta \left\{ \frac{1}{T}\ln \mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta))] \right. \\
& \left. + \frac{1}{T}\ln \pi_i(\bar{\lambda}_i(\theta)|\theta) - \frac{\left[\ell_{iT}^{\lambda}(\theta, \bar{\lambda}_i(\theta))\right]^2}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta))]} \right\}\Big|_{\theta=\theta_0} \\
& + \nu^{-1}(H-\nu)S\nu^{-1} - \frac{1}{2}\nu^{-1}M + O_p\left(\frac{1}{T^{3/2}}\right). \qquad (3.6)
\end{aligned}$$

The first term on the right-hand side of (3.6) is the average projected score with respect to θ and, if a CLT for this term exists, then this term generates the convergence in distribution result for $\hat{\theta}_{IL}$. Whether a CLT exists or not is crucial for inference, but not for bias reduction; all that matters is the convergence rate. The second term contains the now familiar term of the first-order incidental parameter bias of the score. Remember that if the robust prior $\pi_i^R(\cdot)$ is used to construct $T^{-1}\ln\pi_i(\bar{\lambda}_i(\theta)|\theta)$, then by the definition of the robust priors, the expectation of this term is $O(T^{-3/2})$. The orders of magnitude of the third and fourth terms are determined by Assumption 3.3.11 and these are $O_p(N^{-(\rho_1+\rho_2)/2}T^{-1})$ and $O_p(N^{-\rho_1}T^{-1})$, respectively. The below corollary follows immediately.

Corollary 3.5.2 *When the robust prior,* $\pi_i^R\left(\bar{\lambda}_i(\theta_0)|\theta_0\right)$, *is used,*

$$\begin{aligned}\mathbb{E}[\hat{\theta}_{IL}-\theta_0] &= \nu^{-1}\mathbb{E}[(H-\nu)\,S]\nu^{-1}-\frac{1}{2}\nu^{-1}\mathbb{E}[M]+O\left(\frac{1}{T^{3/2}}\right),\\ &= O\left(\frac{1}{N^{(\rho_1+\rho_2)/2}T}\right)+O\left(\frac{1}{N^{\rho_1}T}\right)+O\left(\frac{1}{T^{3/2}}\right).\end{aligned}$$

Theorem 3.5.1 and Corollary 3.5.2 reveal important insights about the first order bias under both time-series and cross-section dependence. First and foremost, there are two types of small sample bias. The first type is the standard incidental parameter bias which is captured by the second and third lines in (3.6) and is corrected by the robust prior. The (potential) second type of bias is due to the third and fourth terms in (3.6). Whether this second type of bias will matter directly depends on ρ_1 and ρ_2, or equivalently, on the contribution of N to the rate of convergence of the score and Hessian with respect to θ. Crucially, this is not an incidental parameter bias term. Instead, this type of bias arises only due to the (possibly) slower rates of convergence. However, depending on the particular values of ρ_1 and ρ_2 this term may not matter after all. This is summarised in the next corollary.

Corollary 3.5.3 *If* ρ_1 *and* ρ_2 *are assumed to be such that*

$$1/2\leq\rho_1\leq 1,\quad 0\leq\rho_2\leq 1\quad \textit{and}\quad 1\leq\rho_1+\rho_2\leq 2, \tag{3.7}$$

then, using the robust prior gives

$$\mathbb{E}[\hat{\theta}_{IL} - \theta_0] = O\left(\frac{1}{T^{3/2}}\right).$$

Hence, (3.7) characterises the setting in which the robust priors of Arellano and Bonhomme (2009) are still valid, despite the presence of time-series and cross-section dependence.

If, in addition, a Central Limit Theorem for $S = \nabla_\theta \ell_{NT}(\theta_0)$ *exists such that,*

$$\sqrt{N^{\rho_1}T}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \ell_{it}(\theta_0, \bar{\lambda}_i(\theta_0)) \xrightarrow{d} \mathcal{N}(0, \Omega),$$

where Ω *is some asymptotic covariance matrix, then by (3.6)*

$$\sqrt{N^{\rho_1}T}(\hat{\theta}_{IL} - \theta_0) \xrightarrow{d} \mathcal{N}(\hat{c}\hat{\mathcal{B}}, \tilde{\Omega}),$$

where $\hat{c} = \lim_{N,T\to\infty}\sqrt{N^{\rho_1}/T^2}$, $\hat{\mathcal{B}} = O(1)$ *and* $\tilde{\Omega}$ *is some asymptotic covariance matrix.*

Corollary 3.5.3 can be confirmed by inspection, using the results of Theorem 3.5.1 and Corollary 3.5.2. The first part of Corollary 3.5.3 gives the conditions under which the Arellano-Bonhomme robust priors are still valid, even under time-series and cross-section dependence. The second part, where the existence of a CLT is assumed, reveals that the asymptotic distribution will correctly be centred at zero if $\lim_{N,T\to\infty}\sqrt{N^{\rho_1}/T^2} = 0$. For example, for $\rho_1 = 1/2$, this suggests that N can grow at the same rate as T^3, a realistic case for financial panels.

It is also worth mentioning intuitively that when $\rho_1 = 0$, if a CLT exists for S such that $\sqrt{T}S \xrightarrow{d} \mathcal{N}(0, \cdot)$, then

$$\sqrt{T}(\hat{\theta}_{IL} - \theta_0) \xrightarrow{d} \mathcal{N}(\tilde{c}\tilde{\mathcal{B}}, \cdot),$$

where again $\tilde{\mathcal{B}} = O(1)$ but this time $\tilde{c} = \lim_{N,T\to\infty}\sqrt{1/T}$. In other words, $p\lim_{T\to\infty}$ $\mathbb{E}[\hat{\theta}_{IL} - \theta_0] = 0$ independent of at what rate N and T go to infinity. This is a very strange implication, because it suggests that the rate at which N and T go to infinity does not matter at all and the small-sample bias is now not an incidental parameter issue but purely

a time-series problem.[10] Nevertheless, it must be underlined that this is more of a thought experiment as, in practice, some contribution from N to the rate of convergence will be expected.

3.5.1 Discussion on Modelling Cross-Section Dependence

One implication of Theorem 3.5.1 is that the specification of the robust prior can be modified in such a way that it corrects both the incidental parameter bias and the cross-section dependence induced bias. However, this is not a trivial task, requiring more specific knowledge on the nature of cross-section dependence. To see this, notice that unless the exact values of ρ_1 and ρ_2 are unknown, it is impossible to know the magnitudes of $\nu^{-1}\mathbb{E}[(H-\nu)\,S]\nu^{-1}$ and $(2\nu)^{-1}\mathbb{E}[M]$. For example, one might mistakenly assume that both of these terms are $O(T^{-1})$ while in reality they are $O(T^{-3/2})$. Then by attempting to remove these terms, one will actually introduce a $O(T^{-1})$ bias into the system. Therefore, correction of the cross-section dependence requires exact knowledge of the magnitudes of the extra terms, which is impossible under the current flexible setting.

In addition, it may be argued that, although convenient to work with, Assumption 3.3.11 does not provide much economic intuition. This is relevant especially for applied research, as it is important to have some concrete idea about the setting within which the theoretical results make sense. An interesting related question is whether one can identify specific dependence structures, under which cross-section dependence does not lead to an extra bias term.

These issues can possibly be solved by modelling dependence explicitly. The rest of this section discusses potential ideas for future research in this direction. One particular approach will be illustrated in the next section.

One popular method is factor modelling, where variables are assumed to be affected by some common time-varying process, the common factor. To illustrate by a simple example,

[10] In the absence of cross-section dependence, the classical result for a non-bias-corrected estimator is

$$\sqrt{NT}(\hat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(\bar{c}\bar{\mathcal{B}},\cdot),$$

where $\bar{\mathcal{B}}=O(1)$, $\hat{\theta}$ is a standard non-bias-corrected estimator (such as the concentrated likelihood estimator) and $\bar{c}=\lim_{N,T\to\infty}\sqrt{N/T}$. As such, the incidental parameter bias depends on the asymptotic ratio of N and T.

let ε_{it} be the variable of interest and consider

$$\varepsilon_{it} = \delta_i \phi_t + \zeta_{it}, \quad \phi_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\phi^2) \quad \text{and} \quad \zeta_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\zeta^2). \tag{3.8}$$

Then, ε_{it} exhibits contemporaneous cross-section dependence due to the presence of the common factor, ϕ_t. Hence, $Cov(\varepsilon_{it}, \varepsilon_{jt}) = \delta_i \delta_j \sigma_\phi^2$ for all $i \neq j$. The factor loading term, δ_i, ensures that the common factor impacts individuals differently.[11] This approach might be difficult in the context of this study for several reasons. First, the entire analysis is based on the likelihood function, rather than some random variable of interest. One could still assume a factor structure for the variable of interest, and investigate the implications of this on the dependence structure of the likelihood and its derivatives. However, given that the analysis considers higher order derivatives of the likelihood function, this approach can become tedious very quickly. In addition, for some models there might be inherent complications with the likelihood function. For example, for the GARCH(1,1) model of Bollerslev (1986), likelihood derivatives do not exist in closed form, due to the recursive structure of the likelihood function. In such cases, keeping track of the factor structure as higher order derivatives of the likelihood function are taken will be very difficult.

A better alternative would be to consider tractable models and then analyse bias reduction in the presence of a factor structure. For example, Phillips and Sul (2007) consider the Dynamic AR(1) model in the presence of neglected cross-section dependence and analyse the Nickell (1981) bias. Also, Bai (2009) considers a linear regression model in the presence of interactive fixed effects and proposes a bias-corrected estimator. The advantage of considering a specific tractable model is that this would most likely enable derivation of a closed form expression of the small-sample bias. When a closed form expression is not available, numerical methods have to be used to calculate the small sample bias. However, numerical calculations usually are more time-consuming and, as the simulation results in the bias reduction literature reveal, do not lead to better results compared to exact calculations. Nevertheless, this price is unavoidable when the underlying model is complicated. One way or another, it is highly likely that eventually a full analysis of bias reduction for non-linear

[11] Important recent contributions in this literature include, but are certainly not limited to, Bai (2003, 2009), Bai and Ng (2002, 2004, 2006), Chudik, Pesaran and Tosetti (2011), Kapetanios, Pesaran and Yamagata (2011), Moon and Perron (2004), Pesaran (2006), Pesaran and Tosetti (2011) and Phillips and Sul (2003). See also Wansbeek and Meijer (2000).

and dynamic panel data models in the presence of a factor structure will be developed.

The other possibility is to consider a spatial dependence framework. Here, the idea is that the degree of dependence between individuals is related to their "distance."[12] This approach is applicable to different fields, including urban, agricultural, development and labour economics and economic growth. A prominent issue about modelling cross-section dependence in this fashion is that, unlike in the time-series case, there is no natural ordering of data. For time-series data, one naturally expects the dependence to decrease as observations further apart in time are considered. In other words, data are already naturally ordered as they arrive. In addition there is a natural sense of distance in time-series data: if today's observation is one unit apart from yesterday's observation, then weekly observations are seven units apart from each other. Such a natural ordering and distance does not exist for cross-section data, which is why finding a meaningful notion of distance, or more formally a distance metric, is important in setting up the spatial dependence framework.

Naturally, the definition of distance will depend on the application under consideration. This can be an important issue in applications, as the researcher already has to possess a meaningful notion of distance between the variables of interest.[13] In terms of the particular example of volatility modelling, it is not clear how this distance can be defined. For example, what is distance between the daily returns on IBM and Microsoft stocks on a given day? Nevertheless, whenever meaningful concepts of distance already exist, spatial dependence is a common way to model cross-section dependence. Of course, whether the employed distance concept is relevant or not is also a question of applied econometric analysis.

Interestingly, it appears that Assumption 3.3.11 can be connected to a clustering/spatial dependence framework. This is important because it makes it possible to analyse the contributions of this paper within a concrete and intuitive dependence framework. Moreover, the possibility to connect the bias reduction and clustering literatures is by itself very interesting, opening the way for applications in clustered samples. A first attempt in this direction is made in the next section.

[12] Recent important theoretical contributions in this area include Conley (1999), Kelejian and Prucha (2007), Lee (2004, 2007), Jenish and Prucha (2009, 2010) and Bester, Conley and Hansen (2011).

[13] See Section 1 in Conley (1999) for a detailed discussion on defining a meaningful "economic distance" for different types of economic applications.

3.6 A Clustering/Spatial Dependence Framework

This section considers a spatial dependence/clustering based approach to model cross-section dependence and analyse the first-order bias properties of the integrated likelihood estimator. It turns out that the cross-section dependence assumptions fit naturally into this framework. The theoretical contribution of this section is establishing a connection between the bias-reduction analysis conducted in this study and the clustering/spatial dependence literatures. It must be underlined that the main objective of the clustering literature is to obtain covariance matrices that are robust to clustering in cross-section. The approach of this study, on the other hand, is entirely pragmatic. The interest is not in modelling robust covariance matrices, but to use their framework and tools.

The analysis here does not in the least claim to provide a complete treatment of bias-reduction in clustered samples. Instead, the focus is on a specific setting, characterised by an increasing number of clusters and an increasing number of members in each cluster, as the sample size goes to infinity. Considering other settings could be an interesting project but this is beyond the scope of this study and is left for future research.

The setting is assumed to be such that the available dataset has already been grouped into clusters. Of course, the researcher will still have to define a metric in order to measure distances. However, this study abstracts away from such. Instead, the aim is to use already developed theoretical results in order to illustrate one possible example where bias reduction under cross-section dependence fits naturally.

The theoretical framework is based on recent work by Jenish and Prucha (2009) and Bester, Conley and Hansen (2011). The main idea is to model cross-section dependence in terms of the mixing concept, in a similar way to the time-series case. This is in contrast to the common assumption that individuals in different clusters are independent, which is a more restrictive setting.[14] Instead, it is possible to assume that observations are spatially weakly dependent, in the sense that the farther apart they are from each other, the closer they are to being independent.

[14] See, for example, Liang and Zeger (1986), Arellano (1987), Bertrand, Duflo and Mullainathan (2004) and Hansen (2007) for important examples. Wooldridge (2003) and Cameron and Miller (2011) provide surveys while a textbook treatment is available in Chapter 20 in Wooldridge (2010).

3.6.1 Notation and the Dependence Setting

Consider some zero-mean random variable Z_{it} and define

$$Z_{iT} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_{it}.$$

It is assumed that $\{Z_{it}\}$ already has a CLT in the time-series dimension for each i. As such, $Z_{iT} = O(1)$. The object of interest is the stochastic order of magnitude of

$$\frac{1}{N} \sum_{i=1}^{N} Z_{iT},$$

as $N, T \to \infty$, where it is known that Z_{it} exhibit cross-section dependence, which will be formalised below. One possible way to obtain the order of magnitude is to find a bound on $Var\left(N^{-1}\sum_{i=1}^{N} Z_{iT}\right)$, which, by Markov's inequality, implies a bound on $N^{-1}\sum_{i=1}^{N} Z_{iT}$.[15] When Z_{iT} is replaced by the relevant centred log-likelihood derivatives, the relevance of this approach in relation with Assumption 3.3.11 becomes immediately obvious. The objective then is to investigate the bound on

$$Var\left(\frac{1}{N} \sum_{i=1}^{N} Z_{iT}\right)$$

As mentioned, individuals are assumed to have already been grouped into clusters by the researcher. Both the number of groups/clusters,[16] G_N, and the number of members of each group, L_N are assumed to grow with N as $N \to \infty$. It is also implicitly assumed that the number of members is the same for all groups. Hence, $L_N G_N = N$. Denote the index set for each group by $\mathcal{G}_g$ where $g = 1, ..., G$ denotes a group and consequently, $\mathcal{G}_g \in \{\mathcal{G}_1, ..., \mathcal{G}_g\}$. $|\cdot|$ gives the number of elements in a given set, e.g. $|\mathcal{G}_g| = L_N$.

Next, a mixing-type dependence concept for spatial processes is defined. The following discussion closely follows Section 2 of Jenish and Prucha (2009) and Section 3 of Bester,

[15] To see this, assume $Var\left(N^{-1}\sum_{i=1}^{N} Z_{iT}\right) = O(N^{\rho})$, for some $-1 < \rho < 0$. Then, by Markov's inequality,

$$P\left[\left(N^{-1}\sum_{i=1}^{N} Z_{iT}\right)^2 > \eta\right] < \eta^{-1} Var\left(N^{-1}\sum_{i=1}^{N} Z_{iT}\right),$$

implying that $\left(N^{-1}\sum_{i=1}^{N} Z_{iT}\right)^2 = O_p(N^{\rho})$ and $N^{-1}\sum_{i=1}^{N} Z_{iT} = O_p(N^{\rho/2})$.

[16] In the remainder, the concepts of group and cluster are used interchangeably.

Conley and Hansen (2011). First, a general d-dimensional discussion is provided.[17] The final result will then be applied to the one-dimensional (cross-sectional) case of this study.

Indices are assumed to be located on an integer lattice $D \subseteq \mathbb{Z}^d$, where $d > 0$, which is a standard setting. If $d = 1$, indices are integers on a line; if $d = 2$, the indices are on a plane etc. The distance metric used in what follows is the *maximum coordinatewise distance metric*, given by

$$\rho(i, j) = \max_{l \in \{1,\ldots,m\}} |j_l - i_l| .$$

Here i_l is the l^{th} component of i. It is also necessary to have a notion of distance between subsets of D. This is given next:

$$\rho(D', D'') = \inf\{\rho(i, j) : i \in D' \text{ and } j \in D''\}, \quad \text{for any } D', D'' \subseteq D.$$

The distance metric enables to measure the distance between two indices, or more intuitively, between two locations. The distance between subsets on the other hand is useful when considering the distance between two clusters, e.g. between two collections of locations. For example, all cities in Germany can be one subset while all cities in France can be another subset. Finally, define the boundary of an index set,

$$\partial\mathcal{G}_g = \{i \in \mathcal{G}_g : \exists j \notin \mathcal{G}_g \text{ such that } \rho(i, j) = 1\}.$$

This simply is the collection of locations that sit on the boundary of group g.

The analysis is based on the assumption that the *sample space* grows to infinity, which ensures that the *sample size* grows to infinity. This is defined as *increasing domain asymptotics*. The alternative is *infill asymptotics* where the sample space remains fixed; consequently, as the sample size grows, observations have to be located more densely.[18] Intuitively, observations are assumed to be located no nearer than some pre-specified distance as the sample size grows. Since location indices are located on an integer lattice, they are all at least one unit away from each other by default, so infill asymptotics is assumed away by definition.

[17] Clustering in many dimensions is not uncommon. For example, in the international trade literature, observations can be clustered by destination and product, by firm and destination or by firm and product. See, e.g., Manova and Zhang (2009).

[18] This is formalised in Assumption 1 of Jenish and Prucha (2009).

A definition of α-mixing for random fields can now be given.

Definition 3.6.1 *For $D'_N \subseteq D$ and $D''_N \subseteq D$, define the random fields $Y' = \{Y_{iT,N} : i \in D'_N\}$ and $Y'' = \{Y_{iT,N} : i \in D''_N\}$. Define further the respective σ-fields as $\mathcal{A}_{T,N} = \sigma(Y_{iT,N} : i \in D'_N)$ and $\mathcal{B}_{T,N} = \sigma(Y_{iT,N} : i \in D'_N)$. Then, the α-mixing coefficient is given by*

$$\begin{aligned} \alpha_{k,l,T,N}(m) &= \sup_{\mathbb{S}} |P(A \cap B) - P(A)P(B)|, \\ \textit{where} \quad \mathbb{S} &= \{A, B : A \in \mathcal{A}_{T,N},\ B \in \mathcal{B}_{T,N},\ \left|D'_N\right| \leq k,\ \left|D''_N\right| \leq l,\ \rho(D'_N, D''_N) \geq m\}. \end{aligned}$$

This definition is different from the standard time-series definition in several ways. First, the cardinalities, or the number of elements, of the index sets D'_N and D''_N do matter. Jenish and Prucha (2011) explain that this is due to the fact that, given a fixed distance between two sets, the dependence between larger sets will be at least as high as the dependence between smaller sets, due to accumulation of dependence. This in turn leads to possibly greater dependence between the related σ-algebras. Consequently, one would, for instance, expect $\alpha_{k,l,T,N}(m) \geq \alpha_{\tilde{k},\tilde{l},T,N}(m)$ when $\tilde{k} > k$ and $\tilde{l} > l$. Here, m is a measure of distance between the two index sets. Consequently, if the α-mixing coefficient vanishes as $m \to \infty$, the underlying random field will be α-mixing. The index sets naturally depend on N because as N increases, the sample space expands. Dependence on T is a modification introduced here. This is necessary because the random fields are constructed using Z_{iT}. Since these depend on T, the σ-algebras generated by these observations will also depend on T. This is mentioned to make the analysis complete; in the remainder the focus will be on

$$\alpha_{k,l}(m) = \sup_{T,N} \alpha_{k,l,T,N},$$

in any case, so dependence on T and N will not be an explicit problem. Conley (1999) and Bester, Conley and Hansen (2011), for example, do not include dependence on N in their definition of α-mixing for random fields at all.

The following assumptions are adapted from Bester, Conley and Hansen (2011). Some of these have already been mentioned, but are nevertheless listed below for sake of completeness.

Assumption 3.6.1 *D grows uniformly in d non-opposing directions as* $N \to \infty$. *In addition,* G_N,$L_N \to \infty$ *as* $N \to \infty$. *Also, for all groups* $g \in \{1, ..., G\}$, $|\mathcal{G}_g| = L_N$.

Assumption 3.6.2 *For all* $g \in \{1, ..., G\}$, $|\partial\mathcal{G}_g| < CL_N^{(d-1)/d}$, *where C is some constant.*

Assumption 3.6.3 *Groups are mutually exclusive, exhaustive and contiguous in the maximum coordinatewise distance metric.*

Assumption 3.6.4 $\sup_{i,T} \mathbb{E}[|Z_{iT}|^{\tilde{\varepsilon}}] < \infty$ *where* $\tilde{\varepsilon} > 2 + \delta$ *for some* $\delta > 0$.

Assumption 3.6.5 *(a)* $\sum_{m=1}^{\infty} m^d \alpha_{1,1}(m)^{\delta/(2+\delta)} < \infty$; *(b)* $\sum_{m=1}^{\infty} m^{d-1} \alpha_{k,l}(m) < \infty$ *for* $k + l \leq 4$; *(c)* $\alpha_{1,\infty}(m) = O(m^{-d-\varepsilon})$ *for some* $\varepsilon > 0$.

Assumption 3.6.6 $\liminf_{N\to\infty} Var\left(L_N^{-1/2} \sum_{i\in\mathcal{G}_g} Z_{iT}\right) > 0$, *for all g.*

Assumptions 3.6.1-3.6.3 determine the characteristics of the cluster structure. These ensure that the clustering cannot be considered in less than d dimensions, that the number of groups and the number of members of a given group increase with N and that all groups have an equal number of members. In addition, the border size of a given cluster is bounded above by $CL_N^{(d-1)/d}$. This precludes, for example, "narrow" yet "very long" clusters. Essentially, this assumption is used to limit the dependence between clusters. Finally, it is assumed that all members of a sample are assigned to one and only one cluster. Contiguity ensures that a given group does not have disjoint components. Assumptions 3.6.4-3.6.6 are technical conditions for mixing processes which are used to invoke Theorem 1 of Jenish and Prucha (2009).

3.6.2 Cross-Section Dependence Revisited: Clustering Perspective

Under the notation introduced in the previous section,

$$\begin{aligned} Var\left(\frac{1}{N}\sum_{i=1}^{N} Z_{iT}\right) &= \frac{1}{N^2}\sum_{i=1}^{N} Var\left(Z_{iT}\right) + \frac{1}{N^2}\sum_{i\neq j}^{N}\sum^{N} Cov(Z_{iT}, Z_{jT}) \\ &= \frac{1}{N^2}\sum_{g=1}^{G}\sum_{i\in\mathcal{G}_g} Cov\left(Z_{iT}, Z_{jT}\right) \qquad (3.9) \\ &\quad + \frac{1}{N^2}\sum_{g\neq h}^{G}\sum^{G}\sum_{i\in\mathcal{G}_g}\sum_{j\in\mathcal{G}_h} Cov(Z_{iT}, Z_{jT}). \qquad (3.10) \end{aligned}$$

This gives the variance of the cross-section average as the sum of two parts: (i) the normalised sum of covariances of all pairs from the same cluster (3.9) and (ii) the normalised sum of covariances of all pairs from different clusters (3.10). The main result now follows.

Theorem 3.6.2 *Under Assumptions 3.6.1-3.6.6,*

$$Var\left(\frac{1}{N}\sum_{i=1}^{N} Z_{iT}\right) = O\left(\frac{1}{N}\right) + O\left(\frac{1}{L_N^{(d+1)/d}}\right).$$

The idea behind the proof, given in the Mathematical Appendix, is to attack the two terms separately. The first term is dealt with by using the CLT given in Theorem 1 in Jenish and Prucha (2009), using Assumptions 3.6.4-3.6.6. The bound for the second term is found by employing the method used by Bester, Conley and Hansen (2011) in the proof of their Lemma 1. The main idea is to first find a bound on the maximum number of pairs $\{i, j : i \in \mathcal{G}_g, j \in \mathcal{G}_h, g \neq h; g, h = 1, ..., G\}$ that will be considered. This is where Assumptions 3.6.2 and 3.6.3 are used. A bound on the covariances is already available due to Bolthausen (1982). Combining these two bounds results in the bound given in Theorem 3.6.2.

To illustrate the significance of Theorem 3.6.2, first notice that in this study $d = 1$. Then, the following corollary yields the updated convergence rates for Assumption 3.3.11.

Corollary 3.6.3 *Assume that $L_N = O(\sqrt{N})$. If the sampling setting and the three likelihood derivatives considered in Assumption 3.3.11 satisfy Assumptions 3.6.1-3.6.6, then*

$$\begin{aligned}
&\nabla_\theta \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0)) = O_p\left(\frac{1}{\sqrt{NT}}\right), \\
&\nabla_{\theta^{(2)}} \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0)) - \mathbb{E}[\nabla_{\theta^{(2)}} \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0))] = O_p\left(\frac{1}{\sqrt{NT}}\right), \\
&\nabla_{\theta^{(3)}} \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0)) - \mathbb{E}[\nabla_{\theta^{(j)}} \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0))] = O_p\left(\frac{1}{\sqrt{NT}}\right),
\end{aligned}$$

as $N, T \to \infty$.

The results follow by observing that, for example,

$$TVar\left[\nabla_\theta \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0))\right] = O\left(\frac{1}{N}\right) + O\left(\frac{1}{L_N^2}\right) = O\left(\frac{1}{N}\right)$$

$$\Rightarrow \quad \nabla_\theta \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0)) = O_p\left(\frac{1}{\sqrt{NT}}\right),$$

where the bound on $\nabla_\theta \ell_{NT}(\theta_0, \bar{\lambda}(\theta_0))$ follows from Markov's inequality.

This approach then suggests that it is possible to achieve a faster convergence rate by assuming weak dependence across both time and cross-section. This in turn implies that the cross-section dependence induced bias term will be negligible, as Corollary 3.6.3 indicated that $\rho_1 = \rho_2 = 1$ in Assumption 3.3.11. Needless to say, this result is based on high level assumptions and the convenient presupposition that data have already been grouped into well-defined clusters. This latter task is not always an easy one. Be that as it may, as explained previously, the main objective of this section is to point towards an interesting and promising approach to linking the analytical bias reduction and spatial dependence literatures. A more detailed analysis of the related technical and empirical details would certainly be an interesting research avenue.

3.7 Conclusion

This chapter has analysed the properties of first-order bias correction by the integrated likelihood method in a non-linear dynamic panel estimation setting under both serial and cross-section dependence. Analysis of bias-reduction of the integrated likelihood method has been extended to both time-series and cross-section dependence. The extension to the case of cross-section dependence is the major theoretical contribution of this chapter as this has not yet been analysed in the analytical bias reduction literature. Considering that many macroeconomic and financial variables are highly likely to exhibit cross-section dependence of some degree, this extension has many potential uses.

The theoretical analysis reveals that time-series dependence leads to an extra bias term. However, this term is shown to be of a negligible order and is, therefore, not relevant for first-order bias correction. This finding would however be useful for higher-order bias correction. The analysis further reveals that cross-section dependence leads to a second and different kind of bias. This is not an incidental parameter bias and is entirely due to cross-section dependence. It has been shown that the exact order of this extra term depends on the level of cross-section dependence, through double asymptotic convergence rates. Based

on these results, the particular double-asymptotic convergence rates which ensure that the Arellano-Bonhomme prior is still valid in the presence of time-series and cross-section dependence have been derived. Therefore, this study extends the analysis of Arellano and Bonhomme (2009) to panels with dependence in both dimensions. In addition, likelihood-based analytical expressions for all extra bias terms have been derived. The asymptotic expansions and bias characterisations provided here can be used in a general setting, not necessarily confined to the particular application of interest in this thesis. Therefore, the results of this chapter appeal to a wide audience in econometrics. A final application considers bias-reduction in the specific setting of spatial dependence for a clustered sample. The analysis reveals that, under certain assumptions on the clustering characteristics, specific double-asymptotic convergence rates can be obtained which ensure that cross-section dependence does not lead to an extra bias term.

Within the context of this thesis, this chapter developed the necessary tools to obtain bias corrected estimators for GARCH panels. The next chapter combines the findings of Chapters 2 and 3 in order to investigate the possibility of modelling volatility in short panels characterised by as little as 150-200 time-series observations.

3.A Mathematical Appendix

Remark 3.A.1 *In what follows, E_{iT} is used as shorthand for $\mathbb{E}[\ell_{iT}^{\lambda\lambda}]$.*

3.A.1 A Preliminary Lemma

The following lemma will be useful in proving some of the results mentioned in this study.

Lemma 3.A.1 *Under Assumption 3.3.7,*

$$\begin{aligned}
\delta_i &= \frac{1}{E_{iT}}\Bigg\{-\ell_{iT}^{\lambda}+\frac{V_{iT}^{\lambda\lambda}\ell_{iT}^{\lambda}}{E_{iT}}-\frac{1}{2}\frac{\ell_{iT}^{\lambda\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^2}+\frac{-V_{iT}^{\lambda\lambda}}{E_{iT}}\left[\frac{V_{iT}^{\lambda\lambda}\ell_{iT}^{\lambda}}{E_{iT}}-\frac{1}{2}\frac{\ell_{iT}^{\lambda\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^2}\right] \\
&\quad -\frac{1}{2}\frac{\ell_{iT}^{\lambda\lambda\lambda}}{E_{iT}^2}\left[-2\frac{V_i^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}}+\frac{\ell_{iT}^{\lambda\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^3}{E_{iT}^2}\right]-\frac{1}{6}\frac{\ell_{iT}^{\lambda\lambda\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^3}{E_{iT}^3}\Bigg\}+O_p(T^{-2}), \quad (3.11)\\
\delta_i^2 &= \frac{1}{E_{iT}^2}\left[\left(\ell_{iT}^{\lambda}\right)^2-2\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}}+\frac{\ell_{iT}^{\lambda\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^3}{E_{iT}^2}\right]+O_p(T^{-2}), \quad (3.12)\\
\delta_i^3 &= -\frac{1}{E_{iT}^3}\left(\ell_{iT}^{\lambda}\right)^3+O_p(T^{-2}). \quad (3.13)
\end{aligned}$$

Proof of Lemma 3.A.1. Expanding $\ell_{iT}^{\lambda}(\theta, \hat{\lambda}_i(\theta))$ around $\hat{\lambda}_i(\theta) = \bar{\lambda}_i(\theta)$ yields

$$\begin{aligned}\ell_{iT}^{\lambda}(\theta, \hat{\lambda}_i(\theta)) &= \ell_{iT}^{\lambda} + \ell_{iT}^{\lambda\lambda}\delta_i + \frac{1}{2}\ell_{iT}^{\lambda\lambda\lambda}\delta_i^2 + \frac{1}{6}\ell_{iT}^{\lambda\lambda\lambda\lambda}\delta_i^3 + O_p(T^{-2}) \\ &= \ell_{iT}^{\lambda} + V_{iT}^{\lambda\lambda}\delta_i + E_{iT}\delta_i + \frac{1}{2}\ell_{iT}^{\lambda\lambda\lambda}\delta_i^2 + \frac{1}{6}\ell_{iT}^{\lambda\lambda\lambda\lambda}\delta_{iT}^3 + O_p(T^{-2}).\end{aligned}$$

Then

$$\begin{aligned}\delta_i &= \frac{1}{E_{iT}}\left[-\ell_{iT}^{\lambda} - \frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\left(-\ell_{iT}^{\lambda} - V_{iT}^{\lambda\lambda}\delta_i - \frac{1}{2}\ell_{iT}^{\lambda\lambda\lambda}\delta_{iT}^2\right) - \frac{1}{2}\ell_{iT}^{\lambda\lambda\lambda}\delta_i^2 - \frac{1}{6}\ell_{iT}^{\lambda\lambda\lambda\lambda}\delta_i^3\right] + O_p(T^{-2}) \\ &= \frac{1}{E_{iT}}\Bigg\{-\ell_{iT}^{\lambda} - \frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\left[-\ell_{iT}^{\lambda} - \frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\left(-\ell_{iT}^{\lambda}\right) - \frac{1}{2}\ell_{iT}^{\lambda\lambda\lambda}\delta_i^2\right] \\ &\quad -\frac{1}{2}\ell_{iT}^{\lambda\lambda\lambda}\delta_i^2 - \frac{1}{6}\ell_{iT}^{\lambda\lambda\lambda\lambda}\delta_i^3\Bigg\} + O_p(T^{-2}) \qquad (3.14)\end{aligned}$$

Similarly,

$$\begin{aligned}\delta_i^2 &= \frac{1}{E_{iT}^2}\left[\left(\ell_{iT}^{\lambda}\right)^2 + 2\ell_{iT}^{\lambda}V_{iT}^{\lambda\lambda}\delta_i + \ell_{iT}^{\lambda}\ell_{iT}^{\lambda\lambda\lambda}\delta_i^2\right] + O_p(T^{-2}) \\ &= \frac{1}{E_{iT}^2}\left[\left(\ell_{iT}^{\lambda}\right)^2 - 2\frac{\left(\ell_{iT}^{\lambda}\right)^2 V_{iT}^{\lambda\lambda}}{E_i} + \ell_{iT}^{\lambda}\ell_{iT}^{\lambda\lambda\lambda}\delta_i^2\right] + O_p(T^{-2}), \qquad (3.15)\end{aligned}$$

where (3.14) is used to obtain (3.15). Substituting δ_i^2 back into (3.15) yields (3.12), while observing $\delta_i^3 = \delta_i\delta_i^2$ gives (3.13). Finally, using (3.12) and (3.13) in (3.14), (3.11) follows. ■

For a more detailed treatment of similar expansions, see, among others, McCullagh (1987) and Pace and Salvan (1997).

3.A.2 Proof of Theorem 3.4.1

This theorem will be proved by using a series of results. The objective is to find an expression for

$$\mathbb{E}[\ell_{iT}^{I}(\theta) - \ell_{iT}(\theta)],$$

which will be done in two steps by deriving first $\mathbb{E}[\ell_{iT}^{I}(\theta) - \ell_{iT}^{c}(\theta)]$ and then $E[\ell_{iT}^{c}(\theta) - \ell_{iT}(\theta)]$.

Lemma 3.A.2

$$\ell_{iT}^{I}(\theta) - \ell_{iT}^{c}(\theta) = \frac{1}{2T}\ln\left(\frac{2\pi}{T}\right) - \frac{1}{2T}\ln[-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))] + \frac{1}{T}\ln \pi_i(\hat{\lambda}_i(\theta)) + O\left(\frac{1}{T^2}\right). \quad (3.16)$$

Proof. This proof is closely based on the exposition in Pace and Salvan (1997). The final expression is the same as in Tierney, Kass and Kadane (1989). See also Davison (2003), Erdélyi (1956) and Severini (2005). Define

$$\begin{aligned} g_i &= -\ell_{iT}(\theta, \lambda_i), \qquad h_i = \pi_i(\lambda_i|\theta), \\ \hat{g}_i &= -\ell_{iT}(\theta, \hat{\lambda}_i(\theta)), \qquad \hat{h}_i = \pi_i(\hat{\lambda}_i(\theta)|\theta), \\ \hat{\delta}_i &= \lambda_i - \hat{\lambda}_i(\theta), \\ \hat{g}_i' &= \frac{\partial \ell_{iT}(\theta, \lambda_i)}{\partial \lambda_{iT}}\Big|_{\lambda_i = \hat{\lambda}_i(\theta)}, \qquad \hat{h}_i' = \frac{\partial \pi_i(\lambda_i|\theta)}{\partial \lambda_i}\Big|_{\lambda_i = \hat{\lambda}_i(\theta)}, \end{aligned}$$

and likewise for higher order derivatives. Then, expanding $\ell_i(\theta, \lambda_i)$ and $\pi_i(\lambda_i|\theta)$ around $\hat{\lambda}_i(\theta)$ and using Assumption 3.3.6, one gets

$$\begin{aligned} g_i &= \hat{g}_i + \frac{1}{2}\hat{\delta}_i^2\hat{g}_i'' + \frac{1}{6}\hat{\delta}_i^3\hat{g}_i''' + \frac{1}{24}\hat{\delta}_i^4\hat{g}_i'''' + O(\hat{\delta}_i^5), \\ h_i &= \hat{h}_i + \hat{\delta}_i\hat{h}_i' + \hat{\delta}_i^2\hat{h}_i'' + O(\hat{\delta}_i^3). \end{aligned}$$

Now,

$$\begin{aligned} L_i^I(\theta) &= \int \exp[T\ell_i(\theta, \lambda_i)]\pi_i(\lambda_i|\theta)d\lambda_i \\ &= \int \exp[-Tg_i]\pi_i(\lambda_i|\theta)d\lambda_i \\ &= \int \exp\left[-T\hat{g}_i - \frac{1}{2}\hat{\delta}_i^2 T\hat{g}_i'' - \frac{1}{6}\hat{\delta}_i^3 T\hat{g}_i''' - \frac{1}{24}\hat{\delta}_i^4 T\hat{g}_i'''' + O(T\hat{\delta}_i^5)\right] h_i d\lambda_i. \end{aligned}$$

Changing the variable to $z_i = (\lambda_i - \hat{\lambda}_i(\theta))\sqrt{T\hat{g}_i''}$ and multiplying and dividing by $\sqrt{T\hat{g}_i''/(2\pi)}$ yields[19]

$$L_i^I(\theta) = \frac{\sqrt{2\pi}\exp(-T\hat{g}_i)}{\sqrt{T\hat{g}_i''}} \int \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{z_i^2}{2}\right]$$

[19] Notice that π here is the pi number and not some prior.

$$\times \exp\left[-\frac{z_i^3\hat{g}_i'''}{6\sqrt{T}(\hat{g}_i'')^{3/2}} - \frac{z_i^4\hat{g}_i''''}{24T(\hat{g}_i'')^2} + O\left(\frac{1}{T^{3/2}}\right)\right] h_i dz_i.$$

Notice that $\phi(z_i) = (2\pi)^{-1/2}\exp\left(-z_i^2/2\right)$ is the Standard Normal density for z_i. Since $\exp x = 1 + x + x^2/2 + x^3/6 + ...$,

$$\begin{aligned}
L_i^I(\theta) &= \frac{\sqrt{2\pi}\exp(-T\hat{g}_i)}{\sqrt{T\hat{g}_i''}} \\
&\quad \times \int \left[1 - \frac{z_i^3\hat{g}_i'''}{6\sqrt{T}(\hat{g}_i'')^{3/2}} - \frac{z_i^4\hat{g}_i''''}{24T(\hat{g}_i'')^2} + \frac{1}{2}\frac{(\hat{g}_i''')^2}{36T(\hat{g}_i'')^3}z_i^6 + O\left(\frac{1}{T^{3/2}}\right)\right] h_i\phi(z_i)dz_i \\
&= \frac{\sqrt{2\pi}\exp(-T\hat{g}_i)}{\sqrt{T\hat{g}_i''}} \\
&\quad \times \int \left[1 - \frac{\hat{g}_i'''}{6\sqrt{T}(\hat{g}_i'')^{3/2}}z_i^3 - \frac{\hat{g}_i''''}{24T(\hat{g}_i'')^2}z_i^4 + \frac{1}{2}\frac{(\hat{g}_i''')^2}{36T(\hat{g}_i'')^3}z_i^6 + O\left(\frac{1}{T^{3/2}}\right)\right] \\
&\quad \times \left[\hat{h}_i + \frac{\hat{h}_i'}{\sqrt{T\hat{g}_i''}}z_i + \frac{\hat{h}_i''}{T\hat{g}_i''}z_i^2 + O\left(\frac{1}{T^{3/2}}\right)\right]\phi(z_i)dz_i \\
&= \frac{\sqrt{2\pi}\exp(-T\hat{g}_i)}{\sqrt{T\hat{g}_i''}} \int \left[\hat{h}_i + \frac{\hat{h}_i'}{\sqrt{T\hat{g}_i''}}z_i - \frac{\hat{g}_i'''\hat{h}_i}{6\sqrt{T}(\hat{g}_i'')^{3/2}}z_i^3 - \frac{\hat{g}_i''''\hat{h}_i}{24T(\hat{g}_i'')^2}z_i^4\right. \\
&\quad \left. + \frac{1}{2}\frac{(\hat{g}_i''')^2\hat{h}_i}{36T(\hat{g}_i'')^3}z_i^6 - \frac{\hat{h}_i'\hat{g}_i'''}{6T(\hat{g}_i'')^2}z_i^4 + \frac{\hat{h}_i''}{T\hat{g}_i''}z_i^2 + O\left(\frac{1}{T^{3/2}}\right)\right]\phi(z_i)dz_i \\
&= \frac{\sqrt{2\pi}\exp(-T\hat{g}_i)}{\sqrt{T\hat{g}_i''}}\left[\hat{h}_i - \frac{1}{8}\frac{\hat{g}_i''''\hat{h}_i}{T(\hat{g}_i'')^2} + \frac{5}{24}\frac{(\hat{g}_i''')^2\hat{h}_i}{T(\hat{g}_i'')^3} - \frac{1}{2}\frac{\hat{h}_i'\hat{g}_i'''}{T(\hat{g}_i'')^2} + \frac{\hat{h}_i''}{T\hat{g}_i''} + O\left(\frac{1}{T^2}\right)\right],
\end{aligned}$$

where the last line follows from the fact that for standard normal random variables odd moments are equal to zero while even moments of order n are equal to $\prod_{j=1}^{n}(n-2j+1)$. Moreover, it can be checked that all $O(T^{-3/2})$ terms involve odd powers of z_i implying that their expectations will all be $O(T^{-2})$. Hence,

$$\begin{aligned}
\ell_{iT}^I(\theta) - \ell_{iT}^c(\theta) &= \frac{1}{T}\ln\int \exp[T\ell_{iT}(\theta,\lambda_i)]\pi_i(\lambda_i|\theta)d\lambda_i - \hat{\ell}_{iT}(\theta,\hat{\lambda}_i(\theta) \\
&= \frac{1}{T}\ln\left\{\frac{\sqrt{2\pi/T}\exp\left[T\hat{\ell}_{iT}(\theta,\hat{\lambda}_i(\theta))\right]}{\sqrt{\hat{\ell}_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta))}}\left[\pi_i(\hat{\lambda}_i(\theta)|\theta) + O\left(\frac{1}{T}\right)\right]\right\} \\
&\quad -\hat{\ell}_{iT}(\theta,\hat{\lambda}_i(\theta) \\
&= \frac{1}{2T}\ln\frac{2\pi}{T} - \frac{1}{2T}\ln\hat{\ell}_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta)) + \ln\pi_i(\hat{\lambda}_i(\theta)|\theta) + O\left(\frac{1}{T^2}\right).
\end{aligned}$$

■

Next, a series of Taylor approximations will be used to derive an expression for $E[\ell_{iT}^{I}(\theta) - \ell_{iT}^{c}(\theta)]$ using (3.16). All asymptotic expansions in this paper heavily make use the fact that the likelihood function and its derivatives are mixing processes, as detailed in the set of Assumptions in Section 3.3. This property, along with the moment conditions in Assumption 3.3.10, ensures that there exist Laws of Large Numbers and Central Limit Theorems for the relevant properly normalised likelihood terms, by, for example, Corollary 3.48 and Theorem 5.20 in White (2001).

Lemma 3.A.3

$$\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta) = \frac{A_i}{\sqrt{T}} + O_p\left(\frac{1}{T}\right), \tag{3.17}$$

where $A_i = -\sqrt{T}\ell_{iT}^{\lambda}\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\}^{-1}$, $\mathbb{E}[A_i] = 0$ *and* $A_i = O_p(1)$ $\forall i$.

Proof. By expanding $\ell_{iT}^{\lambda}(\theta, \hat{\lambda}_i(\theta))$ around $\hat{\lambda}_i(\theta) = \bar{\lambda}_i(\theta)$.

$$\begin{aligned}
\ell_{iT}^{\lambda}(\theta, \hat{\lambda}_i(\theta)) &= \ell_{iT}^{\lambda} + (\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta))\mathbb{E}[\ell_{iT}^{\lambda\lambda}] + (\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta))V_{iT}^{\lambda\lambda} + O_p(T^{-1}) \\
&= \ell_{iT}^{\lambda} + (\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta))\mathbb{E}[\ell_{iT}^{\lambda\lambda}] + O_p(T^{-1}).
\end{aligned}$$

Since $\ell_{iT}^{\lambda}(\theta, \hat{\lambda}_i(\theta)) = 0$,

$$\hat{\lambda}_i(\theta) - \bar{\lambda}_i = -\frac{\ell_{iT}^{\lambda}}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]} + O_p(T^{-1}).$$

By definition, $\mathbb{E}[\ell_{iT}^{\lambda}(\theta, \bar{\lambda}_i(\theta))] = 0$. Hence, defining $A_{iT} = -\sqrt{T}\ell_{iT}^{\lambda}(\theta, \bar{\lambda}_i(\theta))\{\mathbb{E}\left[\ell_{iT}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta))\right]\}^{-1}$ and noting that $\mathbb{E}[A_i] = 0$ and $A_i = O_p(1)$ gives the desired result. ■

Lemma 3.A.4

$$\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta)) = \ell_{iT}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta)) + \frac{B_i}{\sqrt{T}} + O_p\left(\frac{1}{T}\right), \tag{3.18}$$

where $B_i = A_i T^{-1}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda\lambda}]$, $\mathbb{E}[B_i] = 0$ *and* $B_i = O_p(1)$.

Proof. Since $\ell_{iT}^{\lambda\lambda\lambda}$ and $\ell_{iT}^{\lambda\lambda\lambda\lambda}$ are both $O_p(1)$, expanding $\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))$ around $\hat{\lambda}_i(\theta) = \bar{\lambda}_i(\theta)$, and using (3.17) yields

$$\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta)) = \ell_{iT}^{\lambda\lambda} + \left[\frac{A_i}{\sqrt{T}} + O_p(T^{-1})\right]\ell_{iT}^{\lambda\lambda\lambda} + O_p(T^{-1}) = \ell_{iT}^{\lambda\lambda} + \frac{A_i}{\sqrt{T}}\ell_{iT}^{\lambda\lambda\lambda} + O_p(T^{-1}).$$

Then,

$$\begin{aligned}\ell_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta)) &= \ell_{iT}^{\lambda\lambda}+\frac{A_i}{\sqrt{T}}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda\lambda}]+O_p(T^{-1})\\ &= \ell_{iT}^{\lambda\lambda}+\frac{B_i}{\sqrt{T}}+O_p(T^{-1}),\end{aligned}$$

since, $B_i = A_i T^{-1}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda\lambda}]$. Moreover, $E[B_i]=0$ and $B_i = O_p(1)$, since A_i and $T^{-1}\sum_{t=1}^{T}E[\ell_{it}^{\lambda\lambda\lambda}]$ are both $O_p(1)$ and

$$\mathbb{E}[B_i]=\mathbb{E}\left\{A_i\frac{1}{T}\sum_{t=1}^{T}E[\ell_{it}^{\lambda\lambda\lambda}]\right\}=\mathbb{E}[A_i]\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda\lambda}]=0.$$

■

Lemma 3.A.5

$$\ell_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta))=\frac{C_i}{\sqrt{T}}+\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda}(\theta,\bar{\lambda}_i(\theta))]+O_p\left(\frac{1}{T}\right), \tag{3.19}$$

where $C_i = B_i+\sqrt{T}\left\{\ell_{iT}^{\lambda\lambda}-T^{-1}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda}]\right\}$, $\mathbb{E}[C_i]=0$ *and* $C_i=O_p(1)$.

Proof. Using (3.18),

$$\begin{aligned}\ell_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta)) &= \frac{\sqrt{T}V_{iT}^{\lambda\lambda}}{\sqrt{T}}+\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda}]+\frac{B_i}{\sqrt{T}}+O_p(T^{-1}),\\ &= \frac{C_i}{\sqrt{T}}+\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda}]+O_p(T^{-1}).\end{aligned}$$

So,

$$\mathbb{E}[C_i]=\mathbb{E}[B_i]+\sqrt{T}\mathbb{E}[V_{iT}^{\lambda\lambda}]=0 \quad \text{and} \quad C_i=O_p(1).$$

■

Lemma 3.A.6

$$\mathbb{E}_{\theta_0,\lambda_{i0}}\left\{\ln[-\ell_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta))]\right\}=\ln\{-\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta,\bar{\lambda}_i(\theta))]\}+O\left(\frac{1}{T}\right). \tag{3.20}$$

Proof. Using (3.19),

$$\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta)) = T^{-1}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda}] + \frac{C_i}{\sqrt{T}} + O_p(T^{-1}) = \mathbb{E}[\ell_{iT}^{\lambda\lambda}] + \frac{C_i}{\sqrt{T}} + O_p(T^{-1}),$$

Then

$$\frac{\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]} = 1 + \frac{C_i}{\sqrt{T}\mathbb{E}[\ell_{iT}^{\lambda\lambda}]} + O_p(T^{-1}),$$

and

$$\ln\left\{\frac{\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))}{E[\ell_{iT}^{\lambda\lambda}]}\right\} = \ln\left\{1 + \frac{C_i}{\sqrt{T}\mathbb{E}[\ell_{iT}^{\lambda\lambda}]} + O_p(T^{-1})\right\}.$$

Expanding $\ln(1+x)$ around $1+\tilde{x}$ where $x = C_i\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\sqrt{T}\}^{-1} + O_p(T^{-1})$ and $\tilde{x} = 0$,

$$\ln\left\{1 + \frac{C_i}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\sqrt{T}} + O_p(T^{-1})\right\} = \frac{C_i}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\sqrt{T}} + O_p(T^{-1}).$$

Hence,

$$\ln\left\{\frac{-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))}{-\mathbb{E}[\ell_{iT}^{\lambda\lambda}]}\right\} = \frac{C_i}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\sqrt{T}} + O_p(T^{-1}),$$

and

$$\ln\{-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))\} = \ln\{-\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\} + \frac{C_i}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\sqrt{T}} + O_p(T^{-1}), \tag{3.21}$$

implying

$$\begin{aligned}\mathbb{E}[\ln\{-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))\}] &= \mathbb{E}[\ln\{-\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\}] + \frac{\mathbb{E}[C_i]}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\sqrt{T}} + E[O_p(T^{-1})] \\ &= \ln\{-\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\} + O\left(T^{-1}\right).\end{aligned}$$

■

Lemma 3.A.7

$$\mathbb{E}_{\theta_0,\lambda_{i0}}\left[\ln \pi_i(\hat{\lambda}_i(\theta)|\theta)\right] = \ln \pi_i(\bar{\lambda}_i(\theta)|\theta) + O\left(\frac{1}{T}\right). \tag{3.22}$$

Proof. Expanding $\ln \pi_i(\hat{\lambda}_i(\theta)|\theta)$ around $\hat{\lambda}_i(\theta) = \bar{\lambda}_i(\theta)$,

$$\ln \pi_i(\hat{\lambda}_i(\theta)|\theta) = \ln \pi_i(\bar{\lambda}_i(\theta)|\theta) + \frac{\partial \ln \pi_i(\bar{\lambda}_i(\theta)|\theta)}{\partial \lambda_i}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta)) + O_p(T^{-1}), \tag{3.23}$$

which implies that,

$$\begin{aligned}\mathbb{E}[\ln \pi_i(\hat{\lambda}_i(\theta)|\theta)] &= \mathbb{E}\left[\ln \pi_i(\bar{\lambda}_i(\theta)|\theta)\right] + \frac{\partial \ln \pi_i(\bar{\lambda}_i(\theta)|\theta)}{\partial \lambda_i}\mathbb{E}[\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta)] + O(T^{-1}) \\ &= \ln \pi_i(\bar{\lambda}_i(\theta)|\theta) + O\left(T^{-1}\right).\end{aligned}$$

■

Using the results so far, an expression for $\mathbb{E}_{\theta_0,\lambda_{i0}}\left[\ell_i^I(\theta) - \ell_i^c(\theta)\right]$ is given in the next proposition.

Proposition 3.A.1

$$\begin{aligned}\mathbb{E}_{\theta_0,\lambda_{i0}}\left[\ell_i^I(\theta) - \ell_i^c(\theta)\right] &= C - \frac{1}{2T}\ln\left\{-T^{-1}\sum\nolimits_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda}]\right\} \\ &\quad + \frac{1}{T}\ln \pi_i\left(\bar{\lambda}_i(\theta)|\theta\right) + O\left(\frac{1}{T^2}\right). \qquad (3.24)\end{aligned}$$

Proof. Taking the expectation of (3.16) gives

$$\mathbb{E}\left[\ell_{iT}^I(\theta) - \ell_{iT}^c(\theta)\right] = \frac{1}{2T}\ln\left(\frac{2\pi}{T}\right) - \frac{1}{2T}\mathbb{E}\{\ln[-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))]\} + \frac{1}{T}\mathbb{E}[\ln \pi_i(\hat{\lambda}_i(\theta)|\theta)] + O\left(\frac{1}{T^2}\right).$$

Using $C = (2T)^{-1}\ln\left(2\pi T^{-1}\right)$ and substituting (3.20) and (3.22), (3.24) follows. ■

Proposition 3.A.2 *The first-order bias of the concentrated likelihood with respect to the target likelihood is given by*

$$\begin{aligned}\mathbb{E}\left[\ell_{iT}^c(\theta, \hat{\lambda}_i) - \ell_{iT}(\theta, \bar{\lambda}_i)\right] &= -\frac{1}{2}\frac{\mathbb{E}[(\ell_{iT}^{\lambda})^2]}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]} + \frac{1}{2}\frac{\mathbb{E}[V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^2} \\ &\quad - \frac{1}{6}\frac{\mathbb{E}[(\ell_{iT}^{\lambda})^3]\mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^3} + O_p(\frac{1}{T^2}). \qquad (3.25)\end{aligned}$$

Proof. Expanding $\ell_{iT}(\theta, \hat{\lambda}_i(\theta))$ around $\hat{\lambda}_i(\theta) = \bar{\lambda}_i(\theta)$ gives

$$\ell_{iT}(\theta, \hat{\lambda}_i(\theta)) - \ell_i = \ell_{iT}^{\lambda}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta)) + \frac{1}{2}\ell_{iT}^{\lambda\lambda}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta))^2 + \frac{1}{6}\ell_{iT}^{\lambda\lambda\lambda}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta))^3 + O_p(T^{-2}).$$

Using Lemma 3.A.1,

$$\ell_{iT}^{\lambda}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta)) = -\frac{\left(\ell_{iT}^{\lambda}\right)^2}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]} + \frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^2} - \frac{1}{2}\frac{\mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}]\left(\ell_{iT}^{\lambda}\right)^3}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^3} + O_p(T^{-2}),$$

$$
\begin{aligned}
\ell_{iT}^{\lambda\lambda}(\hat{\lambda}_i(\theta)-\bar{\lambda}_i(\theta))^2 &= \frac{\left(\ell_{iT}^{\lambda}\right)^2}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]}-\frac{\left(\ell_{iT}^{\lambda}\right)^2 V_{iT}^{\lambda\lambda}}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^2}+\frac{\left(\ell_{iT}^{\lambda}\right)^3 \mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^3}+O_p(T^{-2}),\\
\ell_{iT}^{\lambda\lambda\lambda}(\hat{\lambda}_i(\theta)-\bar{\lambda}_i(\theta))^3 &= -\frac{\left(\ell_{iT}^{\lambda}\right)^3 E[\ell_{iT}^{\lambda\lambda\lambda}]}{\left\{E[\ell_{iT}^{\lambda\lambda}]\right\}^3}+O_p(T^{-2}),
\end{aligned}
$$

which implies that

$$
\begin{aligned}
\mathbb{E}[\ell_{iT}(\theta,\hat{\lambda}_i(\theta))-\ell_{iT}] &= -\frac{\mathbb{E}[(\ell_{iT}^{\lambda})^2]}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]}+\frac{\mathbb{E}[V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^2}-\frac{1}{2}\frac{\mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}]\mathbb{E}[(\ell_{iT}^{\lambda})^3]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^3}\\
&\quad+\frac{1}{2}\frac{\mathbb{E}[(\ell_{iT}^{\lambda})^2]}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]}-\frac{1}{2}\frac{\mathbb{E}[V_{iT}^{\lambda\lambda}(\ell_{iT}^{\lambda})^2]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^2}+\frac{1}{2}\frac{\mathbb{E}[(\ell_{iT}^{\lambda})^3]\mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^3}\\
&\quad-\frac{1}{6}\frac{\mathbb{E}[(\ell_{iT}^{\lambda})^3]\mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^3}+O_p(T^{-2})\\
&= -\frac{1}{2}\frac{\mathbb{E}[(\ell_{iT}^{\lambda})^2]}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]}+\frac{1}{2}\frac{\mathbb{E}[V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^2}-\frac{1}{6}\frac{\mathbb{E}[(\ell_{iT}^{\lambda})^3]\mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^3}+O_p(T^{-2}).
\end{aligned}
$$

■

Finally, the proof of Theorem 3.4.1 follows.

Proof. (Theorem 3.4.1) Using (3.24) and (3.25) gives (3.3), (3.4) and (3.5). ■

3.A.3 Higher Order Bias of the Integrated Likelihood under the iid Assumption

When the error terms are iid, the bias term given by $\mathcal{B}_i^{(2)}(\theta)/T^{3/2}$ can be shown to be $O(T^{-2})$ rather than $O(T^{-3/2})$. A proof of this result is given below. The idea for the proof is standard; see also Pace and Salvan (1997) and Severini (2005, Theorem 4.18).

Proposition 3.A.3 *Assume that both $\ell_{it}^{\lambda}(\theta,\bar{\lambda}_i(\theta))$ and $V_{it}^{\lambda\lambda}(\theta,\bar{\lambda}_i(\theta))$ are iid across t for all i. Then $\mathbb{E}[(\ell_{iT}^{\lambda})^3]=O(T^{-2})$ and $\mathbb{E}\left[(\ell_{iT}^{\lambda})^2 V_i^{\lambda\lambda}\right]=O(T^{-2})$.*

Proof. Remember that ℓ_i^{λ} are $V_i^{\lambda\lambda}$ both zero-mean likelihood derivatives. Define

$$
\tilde{\ell}_{iT}^{\lambda}=\frac{\ell_{iT}^{\lambda}}{\sqrt{Var(\ell_{it}^{\lambda})}} \quad \text{and} \quad \tilde{V}_i^{\lambda\lambda}=\frac{V_{iT}^{\lambda\lambda}}{\sqrt{Var(V_{it}^{\lambda\lambda})}}.
$$

Under the assumption that all likelihood derivatives are iid, both $\sqrt{T}\tilde{\ell}^{\lambda}_{iT}$ and $\sqrt{T}\tilde{V}^{\lambda\lambda}_{iT}$ converge in distribution to the standard normal distribution. Define also

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \sqrt{T} \begin{bmatrix} \tilde{\ell}^{\lambda}_{iT} \\ \tilde{V}^{\lambda\lambda}_{iT} \end{bmatrix}.$$

For $s_j \in \{1, 2\}$, where $j = 1, ..., J$, let $cum(X_{s_1}, ..., X_{s_j})$ denote the generic J^{th} order cumulant and assume that for $J \leq 3$ all cumulants exist. Theorem 2.1 in Hall (1992, pp. 53-54) establishes that

$$cum(X_{s_1}, ..., X_{s_j}) = T^{-(J-2)/2}(k_{J,1} + T^{-1}k_{J,2} + T^{-2}k_{J,3} + ...), \tag{3.26}$$

where $k_{J,r}$ $(r = 1, 2, 3, ...)$ are constants independent of T. Now, $\mathbb{E}[(\ell^{\lambda}_{iT})^3]$ and $\mathbb{E}\left[(\ell^{\lambda}_{iT})^2 V^{\lambda\lambda}_{iT}\right]$ can be normalised appropriately to obtain $\mathbb{E}[(\tilde{\ell}^{\lambda}_{iT})^3]$ and $\mathbb{E}[(\tilde{\ell}^{\lambda}_{iT})^2 \tilde{V}^{\lambda\lambda}_{iT}]$. Using the relationship between moments and cumulants,

$$\begin{aligned}
\mathbb{E}[(\tilde{\ell}^{\lambda}_{iT})^3] &= cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT})cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT})cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT}) \\
&\quad + 3cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT})cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{\ell}^{\lambda}_{iT}) \\
&\quad + cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{\ell}^{\lambda}_{iT}) \\
&= cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{\ell}^{\lambda}_{iT}) \\
&= O(T^{-1/2}),
\end{aligned}$$

where the third line follows from $cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT}) = \mathbb{E}[\sqrt{T}\tilde{\ell}^{\lambda}_{iT}] = 0$ and the final result is due to (3.26). This implies that, $\mathbb{E}[(\ell^{\lambda}_{iT})^3] = O(T^{-2})$. Similarly,

$$\begin{aligned}
\mathbb{E}\left[(\sqrt{T}\tilde{\ell}^{\lambda}_{iT})^2 \sqrt{T}\tilde{V}^{\lambda\lambda}_{iT}\right] &= cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT})cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT})cum(\sqrt{T}\tilde{V}^{\lambda\lambda}_{iT}) \\
&\quad + cum(\sqrt{T}\tilde{V}^{\lambda\lambda}_{iT})cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{\ell}^{\lambda}_{iT}) \\
&\quad + 2cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT})cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{V}^{\lambda\lambda}_{iT}) \\
&\quad + cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{V}^{\lambda\lambda}_{iT}) \\
&= cum(\sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{\ell}^{\lambda}_{iT}, \sqrt{T}\tilde{V}^{\lambda\lambda}_{iT}) \\
&= O(T^{-1/2}),
\end{aligned}$$

implying that $\mathbb{E}\left[(\ell_{iT}^{\lambda})^2 V_{iT}^{\lambda\lambda}\right] = O(T^{-2})$. ■

3.A.4 Proof of Theorem 3.5.1

The proof is based on a fourth-order Taylor expansion of the integrated likelihood functions at $\hat{\theta}_{IL} = \theta_0$. As θ is a $P \times 1$ vector, such an expansion can get complicated and intractable very quickly. For that reason, this proof will heavily be based on the index notation. The main advantage of this notation is that it enables working on multi-dimensional arrays in almost the same fashion as scalars. Before the proof, a short overview of this convention is given.

A Short Overview of Index Notation

A convenient method to do algebraic manipulations with high dimensional arrays is to use the *index notation* utilised for e.g. tensors. This is a concise way of displaying arrays. For example take some P-dimensional vector, $\nu = (\nu_1, ..., \nu_P)'$. Using the index notation, this vector can also be written as $[\nu_r]$, $r = 1, ..., P$. Similarly, for a $P \times Q$ matrix A, where the row i column j entry is denoted by A_{ij} ($i = 1, ..., P$ and $j = 1, ..., Q$), the index notation representation is given by $[A_{ij}]$. Although the convenience of this notation is not immediately obvious for one- or two-dimensional arrays, it is very useful for cases where higher order arrays are considered. For a detailed explanation, see McCullagh (1984) and McCullagh (1987), which is a classical reference. Pace and Salvan (1997, Chapter 9) provide a more approachable treatment and illustrate many important asymptotic expansions for the multivariate case.

In the case at hand, $\theta = [\theta_r]$ where $r \in \{1, ..., P\}$. In the following, to make the notation less cumbersome, indices and subscripts are dropped whenever variables can be distinguished by context. For example, instead of $V_{iT}^{\lambda\lambda}$, simply V is used. Also, for a given function $f(\phi)$, and a P dimensional parameter vector $\phi = [\phi_p]$, $p = 1, ..., P$, define the generic m^{th} order derivative as

$$f_{r_1,...,r_m} = \frac{d^m f(\phi)}{d\phi_{r_1} d\phi_{r_2} ... d\phi_{r_m}} \quad \text{where } r_1, r_2, ..., r_m \in \{1, ..., P\}$$

Then, for example,

$$\frac{d^m V_{iT}^{\lambda\lambda}}{d\theta_{r_1}...d\theta_{r_m}} = \frac{d^m V}{d\theta_{r_1}...d\theta_{r_m}} = V_{r_1,...,r_m},$$

gives an m-dimensional array.

Another convention used here is the *Einstein summation convention*. The idea is to write summations implicitly by observing that, when an index appears twice in a product of arrays, the product is summed across that index. For example, for two arrays x^p and y_p^q, where $p, q = 1, ..., P$, the summation $\sum_{p=1}^{P} x^p y_p^q$ is implicit in $x^p y_p^q$ as p appears twice in the same product. Indices that are not repeated within the same product are called free indices, and the number of these indices determines the dimension of the resulting array. Indices that are repeated, on the other hand, are called dummy indices. As such, $x^p y_p^q$ is a vector (one free index, q), while $x_{rst}^p y_p^q z^{rt}$ is a matrix (two free indices, q and s). Note that the notation for the indices can be changed freely as long their relationship is left intact. For example, $x_q^p y_r^q$ is identical to $x_p^q y_r^p$; but of course $x_q^p y_p^r$ is a different object.

Again, to keep notation simple, the following definitions will be used.

$$\ell = \ell_{iT}(\theta_0, \bar{\lambda}_i(\theta_0)), \quad \ell^r = \frac{d\ell_{iT}(\theta, \bar{\lambda}_i(\theta))}{d\theta^r}\Big|_{\theta=\theta_0}, \quad \ell^{r,s} = \frac{d^2\ell_{iT}(\theta, \bar{\lambda}_i(\theta))}{d\theta^r d\theta^s}\Big|_{\theta=\theta_0}, \quad \text{etc.}$$

$$\tilde{\ell} = \frac{1}{N}\sum_{i=1}^{N} \ell; \quad \tilde{\ell}_a = \frac{1}{N}\sum_{i=1}^{N} \ell_a; \quad \tilde{\ell}_{a,b} = \frac{1}{N}\sum_{i=1}^{N} \ell_{a,b} \quad \text{etc.}$$

$$\nu_{a,b} = \mathbb{E}[\tilde{\ell}_{a,b}]; \quad \nu_{a,b,c} = \mathbb{E}[\tilde{\ell}_{a,b,c}] \quad \text{etc.}$$

$$\mathcal{H}_{a,b} = \tilde{\ell}_{a,b} - \nu_{a,b}; \quad \mathcal{H}_{a,b,c} = \tilde{\ell}_{a,b,c} - \nu_{a,b,c} \quad \text{etc.}$$

where $r, s \in \{1, ..., P\}$. In addition,

$$U_{iT} = \ell_{iT}^{\lambda}(\theta_0, \bar{\lambda}_i(\theta_0)), \quad E_{iT} = \mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta_0, \bar{\lambda}_i(\theta_0))], \quad F_{iT} = \mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}(\theta_0, \bar{\lambda}_i(\theta_0))],$$

$$\Pi_{iT} = \ln \pi_i(\bar{\lambda}_{iT}(\theta_0)|\theta_0) \quad \text{and} \quad \bar{\Pi}_{iT} = \frac{\partial \ln \pi_i(\bar{\lambda}_{iT}(\theta)|\theta)}{\partial \lambda_i}\Big|_{\theta=\theta_0}.$$

Lastly, define

$$\delta_I^r = (\hat{\theta}_{IL} - \theta_0)^r \quad \text{where } r \in \{1, ..., P\}$$

and $(\hat{\theta}_{IL} - \theta_0)^r$ is the r^{th} entry of the vector $(\hat{\theta}_{IL} - \theta_0)$. Notice that δ_I^r here does not mean the r^{th} power of δ_I.

A Preliminary Lemma

The following lemma (the proof of which is given at the end of the next section) will be useful in proving Proposition 3.5.1. Remember that, for notational conciseness, all subscripts such as iT and superscripts denoting derivatives are dropped. Index notation is used to denote derivatives with respect to θ. Hence, for example, V_{r_1} is used shorthand for $\nabla_{\theta_{r_1}} V_{iT}^{\lambda\lambda}(\theta_0)$. In this particular case, since $V_{iT}^{\lambda\lambda\lambda}$ does not appear in the derivations below, this notation is not confusing.

Lemma 3.A.8

$$
\begin{aligned}
\nabla_\theta \ln(-E_{iT}) &= -\frac{E_{r_1}}{E} = O(1), \\
\nabla_{\theta\theta} \ln(-E_{iT}) &= -\frac{E_{r_1,r_2}}{E} + \frac{E_{r_1}E_{r_2}}{E^2} = O(1), \\
\nabla_\theta \left(\frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\right) &= \frac{V_{r_1}}{E} - \frac{VE_{r_1}}{E^2} = O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_{\theta\theta} \left(\frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\right) &= \frac{V_{r_1,r_2}}{E} - \frac{V_{r_1}E_{r_2}[2] + VE_{r_1,r_2}}{E^2} + 2\frac{VE_{r_1}E_{r_2}}{E^3} = O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_\theta \left(\frac{\ell_{iT}^{\lambda} F_{iT}}{E_{iT}^2}\right) &= \frac{U_{r_1}F + UF_{r_1}}{E^2} - 2\frac{UFE_{r_1}}{E^3} = O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_{\theta\theta} \left(\frac{\ell_{iT}^{\lambda} F_{iT}}{E_{iT}^2}\right) &= \frac{U_{r_1,r_2}F + U_{r_1}F_{r_2}[2] + UF_{r_1,r_2}}{E^2} - 2\frac{U_{r_1}E_{r_2}F[2] + UE_{r_2}F_{r_1}[2] + UFE_{r_1,r_2}}{E^3} \\
&\quad +6\frac{UFE_{r_1}E_{r_2}}{E^4} \\
&= O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_\theta \left(\frac{\ell_{iT}^{\lambda} \bar{\Pi}_{iT}}{E_{iT}}\right) &= \frac{U_{r_1}\bar{\Pi} + U\bar{\Pi}_{r_1}}{E} - \frac{U\bar{\Pi}E_{r_1}}{E^2} = O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_{\theta\theta} \left(\frac{\ell_{iT}^{\lambda} \bar{\Pi}_{iT}}{E_{iT}}\right) &= \frac{U_{r_1,r_2}\bar{\Pi} + U_{r_1}\bar{\Pi}_{r_2}[2] + U\bar{\Pi}_{r_1,r_2}}{E} - \frac{U_{r_1}\bar{\Pi}E_{r_2}[2] + U\bar{\Pi}_{r_1}E_{r_2}[2] + U\bar{\Pi}E_{r_1,r_2}}{E^2} \\
&\quad +2\frac{U\bar{\Pi}E_{r_1}E_{r_2}}{E^3} \\
&= O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_\theta \left[\frac{\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}}\right] &= 2\frac{UU_{r_1}}{E} - \frac{U^2E_{r_1}}{E^2} = O_p\left(\frac{1}{T}\right), \\
\nabla_{\theta\theta} \left[\frac{\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}}\right] &= 2\frac{U_{r_2}U_{r_1} + UU_{r_1,r_2}}{E} - \frac{2UU_{r_2}E_{r_1}[2] + U^2E_{r_1,r_2}}{E^2} + 2\frac{U^2E_{r_1}E_{r_2}}{E^3} = O_p\left(\frac{1}{T}\right), \\
\nabla_\theta \left[\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^2}\right] &= \frac{V_{r_1}U^2 + 2VUU_{r_1}}{E^2} - 2\frac{VU^2E_{r_1}}{E^3} = O_p\left(\frac{1}{T^{3/2}}\right),
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\theta\theta}\left[\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^{2}}{E_{iT}^{2}}\right] &= \frac{V_{r_1,r_2}U^2+2V_{r_1}UU_{r_2}[2]+2VU_{r_2}U_{r_1}+2VUU_{r_1,r_2}}{E^2} \\
&\quad -2\frac{V_{r_1}U^2E_{r_2}[2]+2VUU_{r_1}E_{r_2}[2]+VU^2E_{r_1,r_2}}{E^3}+6\frac{VU^2E_{r_1}E_{r_2}}{E^4} \\
&= O_p\left(\frac{1}{T^{3/2}}\right), \\
\nabla_{\theta}\left[\frac{(\ell_{iT}^{\lambda})^3F_{iT}}{E_{iT}^{3}}\right] &= \frac{3U^2U_{r_1}F+U^3F_{r_1}}{E^3}-3\frac{U^3FE_{r_1}}{E^4}=O_p\left(\frac{1}{T^{3/2}}\right), \\
\nabla_{\theta\theta}\left[\frac{(\ell_{iT}^{\lambda})^3F_{iT}}{E_{iT}^{3}}\right] &= \frac{6UU_{r_2}U_{r_1}F+3U^2U_{r_1,r_2}F+3U^2U_{r_1}F_{r_2}[2]+U^3F_{r_1,r_2}}{E^3} \\
&\quad -3\frac{3U^2U_{r_1}FE_{r_2}[2]+U^3F_{r_1}E_{r_2}[2]+U^3FE_{r_1,r_2}}{E^4}+12\frac{U^3FE_{r_1}E_{r_2}}{E^5} \\
&= O_p\left(\frac{1}{T^{3/2}}\right).
\end{aligned}
$$

Moreover, the third and fourth derivatives satisfy,

$$
\begin{aligned}
\nabla_{\theta\theta\theta}\ln(-E_{iT}) &= O(1), & \nabla_{\theta\theta\theta\theta}\ln(-E_{iT}) &= O(1), \\
\nabla_{\theta\theta\theta}\left(\frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\right) &= O_p\left(\frac{1}{\sqrt{T}}\right), & \nabla_{\theta\theta\theta\theta}\left(\frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\right) &= O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_{\theta\theta\theta}\left(\frac{\ell_{iT}^{\lambda}F_{iT}}{E_{iT}^{2}}\right) &= O_p\left(\frac{1}{\sqrt{T}}\right), & \nabla_{\theta\theta\theta\theta}\left(\frac{\ell_{iT}^{\lambda}F_{iT}}{E_{iT}^{2}}\right) &= O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_{\theta\theta\theta}\left(\frac{\ell_{iT}^{\lambda}\bar{\Pi}_{iT}}{E_{iT}}\right) &= O_p\left(\frac{1}{\sqrt{T}}\right). & \nabla_{\theta\theta\theta\theta}\left(\frac{\ell_{iT}^{\lambda}\bar{\Pi}_{iT}}{E_{iT}}\right) &= O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_{\theta\theta\theta}\left[\frac{\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}}\right] &= O_p\left(\frac{1}{T}\right), & \nabla_{\theta\theta\theta\theta}\left[\frac{\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}}\right] &= O_p\left(\frac{1}{T}\right), \\
\nabla_{\theta\theta\theta}\left[\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^{2}}\right] &= O_p\left(\frac{1}{T^{3/2}}\right), & \nabla_{\theta\theta\theta\theta}\left[\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^{2}}\right] &= O_p\left(\frac{1}{T^{3/2}}\right), \\
\nabla_{\theta\theta\theta}\left\{\frac{(\ell_{iT}^{\lambda})^3F_{iT}}{E_{iT}^{3}}\right\} &= O_p\left(\frac{1}{T^{3/2}}\right), & \nabla_{\theta\theta\theta\theta}\left\{\frac{(\ell_{iT}^{\lambda})^3F_{iT}}{E_{iT}^{3}}\right\} &= O_p\left(\frac{1}{T^{3/2}}\right).
\end{aligned}
$$

The Proof

The starting point is

$$
\begin{aligned}
\ell_{iT}^{I}(\theta) &= \ell_{iT}(\theta,\bar{\lambda}_i(\theta))+C-\frac{1}{2T}\left[\ln(-E_{iT})+\frac{V_{iT}^{\lambda\lambda}}{E_{iT}}-\frac{\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta))F_{iT}}{E_{iT}^{2}}\right] \\
&\quad +\frac{1}{T}\left[\ln\pi_{iT}(\bar{\lambda}_{iT}(\theta)|\theta)-\frac{\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta))}{E_{iT}}\frac{\partial\ln\pi_{iT}(\lambda_i|\theta)}{\partial\lambda_i}\Big|_{\lambda_i=\bar{\lambda}_{iT}(\theta)}\right] \\
&\quad -\frac{1}{2}\frac{\left(\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta))\right)^2}{E_{iT}}+\frac{1}{2}\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta))\right)^2}{(E_{iT})^2}
\end{aligned}
$$

$$-\frac{1}{6}\frac{(\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta)))^3 F_{iT}}{E_{iT}^3}+O_p(T^{-2}), \tag{3.27}$$

where $C=(2T)^{-1}\ln(2\pi/T)$. The first line follows from using (3.21) and (3.23) to substitute for $\ln[-\ell_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta))]$ and $\ln\pi_i(\hat{\lambda}_i(\theta))$, respectively, in (3.16). Notice that the expression given by (3.23) is a function of $[\hat{\lambda}_i(\theta)-\bar{\lambda}_i(\theta)]$, which has to be substituted by (3.11), up to a $O_p(T^{-2})$ term. The last two lines are obtained by adding $\ell_{iT}(\theta,\hat{\lambda}_i(\theta))-\ell_{iT}(\theta,\bar{\lambda}_i(\theta))$, which is calculated by using the arguments in the Proof of Proposition 3.A.2.

By a multivariate Taylor expansion of $\ell_{r_1}^I(\hat{\theta}_{IL})$ around $\hat{\theta}_{IL}=\theta_0$,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\ell_{r_1}^I(\hat{\theta}_{IL}) &= \frac{1}{N}\sum_{i=1}^{N}\ell_{r_1}^I(\theta_0)+\left[\frac{1}{N}\sum_{i=1}^{N}\ell_{r_1,r_2}^I(\theta_0)\right]\delta_I^{r_2}+\frac{1}{2}\left[\frac{1}{N}\sum_{i=1}^{N}\ell_{r_1,r_2,r_3}^I(\theta_0)\right]\delta_I^{r_2}\delta_I^{r_3}\\ &+\frac{1}{6}\left[\frac{1}{N}\sum_{i=1}^{N}\ell_{r_1,r_2,r_3,r_4}^I(\theta_0)\right]\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\\ &+\frac{1}{24}\left[\frac{1}{N}\sum_{i=1}^{N}\ell_{r_1,r_2,r_3,r_4,r_5}^I(\bar{\theta})\right]\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\delta_I^{r_5},\end{aligned} \tag{3.28}$$

where $r_1,...,r_5\in\{1,...,P\}$ and defining the j^{th} entry of $\bar{\theta}$ as θ_j, $\theta_j\in[\min(\hat{\theta}_{IL,j},\theta_{0,j}),\max(\hat{\theta}_{IL,j},\theta_{0,j})]$. In the worst case of $\sqrt{T}$-convergence (rather than the$\sqrt{NT}$-convergence observed under cross-section dependence) it is expected that

$$\ell_{r_1,r_2,r_3,r_4,r_5}^I(\bar{\theta})\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\delta_I^{r_5}=O_p(T^{-2}).$$

Notice that the expansion gives a vector.

The integrated likelihood is not a familiar concept. Instead, the concentrated likelihood would be much more convenient intuitive to work with. This is made possible by using (3.27) to obtain target-likelihood based approximations for integrated-likelihood derivatives appearing on the right-hand side of (3.28). These approximations are then substituted for relevant integrated likelihood derivatives in (3.28). This leads to the next lemma.

Lemma 3.A.9

$$\begin{aligned}-\delta_I^{r_2}\nu_{r_1,r_2} &= \tilde{\ell}_{r_1}+\mathcal{D}_{1;r_1}+\delta_I^{r_2}\mathcal{H}_{r_1,r_2}+\frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\nu_{r_1,r_2,r_3}+\mathcal{D}_{3;r_1}\\ &+\delta_I^{r_2}\mathcal{D}_{2;r_1,r_2}+\frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\mathcal{H}_{r_1,r_2,r_3}+\frac{1}{6}\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\nu_{r_1,r_2,r_3,r_4}+O_p\left(\frac{1}{T^2}\right)\end{aligned} \tag{3.29}$$

where

$$\begin{aligned}
\mathcal{D}_{1;r_1} &= \frac{1}{TN}\sum_{i=1}^{N}\frac{E_{r_1}}{2E}+\frac{1}{TN}\sum_{i=1}^{N}\Pi_{r_1}-\frac{1}{N}\sum_{i=1}^{N}\frac{UU_{r_1}}{E}+\frac{1}{N}\sum_{i=1}^{N}\frac{U^2E_{r_1}}{2E^2}=O_p\left(\frac{1}{T}\right),\\
\mathcal{D}_{2;r_1,r_2} &= \frac{1}{TN}\sum_{i=1}^{N}\frac{E_{r_1,r_2}}{2E}-\frac{1}{TN}\sum_{i=1}^{N}\frac{E_{r_1}E_{r_2}}{2E^2}+\frac{1}{TN}\sum_{i=1}^{N}\Pi_{r_1,r_2}-\frac{1}{N}\sum_{i=1}^{N}\frac{U_{r_2}U_{r_1}+UU_{r_1,r_2}}{E}\\
&\quad+\frac{1}{N}\sum_{i=1}^{N}\frac{2U\left(U_{r_1}E_{r_2}+U_{r_2}E_{r_1}\right)+U^2E_{r_1,r_2}}{2E^2}-\frac{1}{N}\sum_{i=1}^{N}\frac{U^2E_{r_1}E_{r_2}}{E^3}\\
&= O_p\left(\frac{1}{T}\right),\\
\mathcal{D}_{3;r_1} &= \frac{1}{TN}\sum_{i=1}^{N}\frac{VE_{r_1}+U_{r_1}F+UF_{r_1}+U\bar{\Pi}E_{r_1}}{2E^2}-\frac{1}{TN}\sum_{i=1}^{N}\frac{UFE_{r_1}}{E^3}\\
&\quad-\frac{1}{TN}\sum_{i=1}^{N}\frac{V_{r_1}+U_{r_1}\bar{\Pi}+U\bar{\Pi}_{r_1}}{E}+\frac{1}{N}\sum_{i=1}^{N}\frac{V_{r_1}U^2+2VUU_{r_1}}{2E^2}\\
&\quad-\frac{1}{N}\sum_{i=1}^{N}\frac{3U^2U_{r_1}F+U^3F_{r_1}+VU^2E_{r_1}}{6E^3}+\frac{1}{N}\sum_{i=1}^{N}\frac{U^3FE_{r_1}}{2E^4}\\
&= O_p\left(\frac{1}{T^{3/2}}\right).
\end{aligned}$$

Proof. First, derivatives of (3.27) with respect to θ have to obtained. This is achieved by simply substituting the results given in Lemma 3.A.8 as necessary. Then,

$$\begin{aligned}
\ell^I_{r_1}(\theta_0) &= \ell_{r_1}(\theta_0)+\frac{1}{T}\left[\frac{E_{r_1}}{2E}+\Pi_{r_1}\right]-\frac{UU_{r_1}}{E}+\frac{U^2E_{r_1}}{2E^2}\\
&\quad+\frac{1}{T}\left[\frac{VE_{r_1}}{2E^2}-\frac{V_{r_1}}{2E}+\frac{U_{r_1}F+UF_{r_1}}{2E^2}-\frac{UFE_{r_1}}{E^3}+\frac{U\bar{\Pi}E_{r_1}}{E^2}-\frac{U_{r_1}\bar{\Pi}+U\bar{\Pi}_{r_1}}{E}\right]\\
&\quad+\frac{V_{r_1}U^2+2VUU_{r_1}}{2E^2}-\frac{VU^2E_{r_1}}{E^3}-\frac{3U^2U_{r_1}F+U^3F_{r_1}}{6E^3}+\frac{U^3FE_{r_1}}{2E^4}\\
&\quad+O_p\left(\frac{1}{T^2}\right),\\
\ell^I_{r_1,r_2}(\theta_0) &= \ell_{r_1,r_2}(\theta_0)+\frac{1}{T}\left[\frac{E_{r_1,r_2}}{2E}-\frac{E_{r_1}E_{r_2}}{2E^2}+\Pi_{r_1,r_2}\right]-\frac{U_{r_2}U_{r_1}+UU_{r_1,r_2}}{E}\\
&\quad+\frac{2U\left(U_{r_1}E_{r_2}+U_{r_2}E_{r_1}\right)+U^2E_{r_1,r_2}}{2E^2}-\frac{U^2E_{r_1}E_{r_2}}{E^3}+O_p\left(\frac{1}{T^{3/2}}\right),\\
\ell^I_{r_1,r_2,r_3}(\theta_0) &= \ell_{r_1,r_2,r_3}(\theta_0)+O_p\left(\frac{1}{T}\right),\\
\ell^I_{r_1,r_2,r_3,r_4}(\theta_0) &= \ell_{r_1,r_2,r_3,r_4}(\theta_0)+O_p\left(\frac{1}{T}\right).
\end{aligned}$$

Substituting these expansions for the integrated likelihood derivatives into (3.28) gives

$$\begin{aligned}
\tilde{\ell}^I_{r_1}(\hat{\theta}_{IL}) &= \tilde{\ell}_{r_1}\left(\theta_0, \bar{\lambda}_i(\theta_0)\right) + \frac{1}{T}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{E_{r_1}}{2E} + \frac{1}{N}\sum_{i=1}^{N}\Pi_{r_1}\right] - \frac{1}{N}\sum_{i=1}^{N}\frac{UU_{r_1}}{E} \\
&+ \frac{1}{N}\sum_{i=1}^{N}\frac{U^2E_{r_1}}{2E^2} + \frac{1}{T}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{VE_{r_1}}{2E^2} - \frac{1}{N}\sum_{i=1}^{N}\frac{V_{r_1}}{2E} + \frac{1}{N}\sum_{i=1}^{N}\frac{U_{r_1}F + UF_{r_1}}{2E^2}\right. \\
&\left. - \frac{1}{N}\sum_{i=1}^{N}\frac{UFE_{r_1}}{E^3} + \frac{1}{N}\sum_{i=1}^{N}\frac{U\Pi E_{r_1}}{E^2} - \frac{1}{N}\sum_{i=1}^{N}\frac{U_{r_1}\bar{\Pi} + U\bar{\Pi}_{r_1}}{E}\right] \\
&+ \frac{1}{N}\sum_{i=1}^{N}\frac{V_{r_1}U^2 + 2VUU_{r_1}}{2E^2} - \frac{1}{N}\sum_{i=1}^{N}\frac{VU^2E_{r_1}}{E^3} - \frac{1}{N}\sum_{i=1}^{N}\frac{3U^2U_{r_1}F + U^3F_{r_1}}{6E^3} \\
&+ \frac{1}{N}\sum_{i=1}^{N}\frac{U^3FE_{r_1}}{2E^4} + \left\{\tilde{\ell}_{r_1,r_2} + \frac{1}{T}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{E_{r_1,r_2}}{2E} - \frac{1}{N}\sum_{i=1}^{N}\frac{E_{r_1}E_{r_2}}{2E^2} + \frac{1}{N}\sum_{i=1}^{N}\Pi_{r_1,r_2}\right]\right. \\
&- \frac{1}{N}\sum_{i=1}^{N}\frac{U_{r_2}U_{r_1} + UU_{r_1,r_2}}{E} + \frac{1}{N}\sum_{i=1}^{N}\frac{2U\left(U_{r_1}E_{r_2} + U_{r_2}E_{r_1}\right) + U^2E_{r_1,r_2}}{2E^2} \\
&\left. - \frac{1}{N}\sum_{i=1}^{N}\frac{U^2E_{r_1}E_{r_2}}{E^3}\right\}\delta_I^{r_2} + \frac{1}{2}\tilde{\ell}_{r_1,r_2,r_3}\delta_I^{r_2}\delta_I^{r_3} + \frac{1}{6}\tilde{\ell}_{r_1,r_2,r_3,r_4}\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4} \\
&+ O_p\left(\frac{1}{T^2}\right)
\end{aligned}$$

Noting that $\tilde{\ell}^I_{r_1}(\hat{\theta}_{IL}) = 0$ for $r_1 \in \{1, ..., P\}$ and rearranging terms according to their stochastic orders of magnitude yields

$$\begin{aligned}
0 &= \tilde{\ell}_{r_1}\left(\theta_0, \bar{\lambda}_i(\theta_0)\right) + \delta_I^{r_2}\tilde{\ell}_{r_1,r_2} \\
&+ \frac{1}{T}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{E_{r_1}}{2E} + \frac{1}{N}\sum_{i=1}^{N}\Pi_{r_1}\right] - \frac{1}{N}\sum_{i=1}^{N}\frac{UU_{r_1}}{E} + \frac{1}{N}\sum_{i=1}^{N}\frac{U^2E_{r_1}}{2E^2} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\tilde{\ell}_{r_1,r_2,r_3} \\
&+ \frac{1}{T}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{VE_{r_1} + U_{r_1}F + UF_{r_1} + U\bar{\Pi}E_{r_1}}{2E^2} - \frac{1}{N}\sum_{i=1}^{N}\frac{V_{r_1}}{2E} - \frac{1}{N}\sum_{i=1}^{N}\frac{UFE_{r_1}}{E^3}\right. \\
&\left. - \frac{1}{N}\sum_{i=1}^{N}\frac{U_{r_1}\bar{\Pi} + U\bar{\Pi}_{r_1}}{E}\right] + \frac{1}{N}\sum_{i=1}^{N}\frac{V_{r_1}U^2 + 2VUU_{r_1}}{2E^2} - \frac{1}{N}\sum_{i=1}^{N}\frac{3U^2U_{r_1}F + U^3F_{r_1} + VU^2E_{r_1}}{6E^3} \\
&+ \frac{1}{N}\sum_{i=1}^{N}\frac{U^3FE_{r_1}}{2E^4} + \delta_I^{r_2}\left\{\frac{1}{TN}\sum_{i=1}^{N}\frac{E_{r_1,r_2}}{2E} - \frac{1}{TN}\sum_{i=1}^{N}\frac{E_{r_1}E_{r_2}}{2E^2} + \frac{1}{TN}\sum_{i=1}^{N}\Pi_{r_1,r_2}\right. \\
&- \frac{1}{N}\sum_{i=1}^{N}\frac{U_{r_2}U_{r_1} + UU_{r_1,r_2}}{E} + \frac{1}{N}\sum_{i=1}^{N}\frac{2U\left(U_{r_1}E_{r_2} + U_{r_2}E_{r_1}\right) + U^2E_{r_1,r_2}}{2E^2} - \frac{1}{N}\sum_{i=1}^{N}\frac{U^2E_{r_1}E_{r_2}}{E^3}\Bigg\} \\
&+ \frac{1}{6}\tilde{\ell}_{r_1,r_2,r_3,r_4}\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4} + O_p\left(\frac{1}{T^2}\right).
\end{aligned}$$

Then, using Assumption 3.3.11,

$$
\begin{aligned}
-\delta_I^{r_2}\nu_{r_1,r_2} &= \tilde{\ell}_{r_1}\left(\theta_0,\bar{\lambda}_i(\theta_0)\right) \\
&\quad +\delta_I^{r_2}\mathcal{H}_{r_1,r_2}+\frac{1}{T}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{E_{r_1}}{2E}+\frac{1}{N}\sum_{i=1}^{N}\Pi_{r_1}\right]-\frac{1}{N}\sum_{i=1}^{N}\frac{UU_{r_1}}{E} \\
&\quad +\frac{1}{N}\sum_{i=1}^{N}\frac{U^2E_{r_1}}{2E^2}+\frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\nu_{r_1,r_2,r_3}+\delta_I^{r_2}\delta_I^{r_3}\frac{1}{2}\mathcal{H}_{r_1,r_2,r_3} \\
&\quad +\frac{1}{6}\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\nu_{r_1,r_2,r_3,r_4}+\frac{1}{T}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{VE_{r_1}+U_{r_1}F+UF_{r_1}+U\bar{\Pi}E_{r_1}}{2E^2}\right. \\
&\quad \left.-\frac{1}{N}\sum_{i=1}^{N}\frac{V_{r_1}+U_{r_1}\bar{\Pi}+U\bar{\Pi}_{r_1}}{2E}-\frac{1}{N}\sum_{i=1}^{N}\frac{UFE_{r_1}}{E^3}\right] \\
&\quad +\frac{1}{N}\sum_{i=1}^{N}\frac{V_{r_1}U^2+2VUU_{r_1}}{2E^2}-\frac{1}{N}\sum_{i=1}^{N}\frac{3U^2U_{r_1}F+U^3F_{r_1}+VU^2E_{r_1}}{6E^3} \\
&\quad +\frac{1}{N}\sum_{i=1}^{N}\frac{U^3FE_{r_1}}{2E^4}+\delta_I^{r_2}\left\{\frac{1}{TN}\sum_{i=1}^{N}\frac{E_{r_1,r_2}}{2E}-\frac{1}{TN}\sum_{i=1}^{N}\frac{E_{r_1}E_{r_2}}{2E^2}\right. \\
&\quad +\frac{1}{TN}\sum_{i=1}^{N}\Pi_{r_1,r_2}-\frac{1}{N}\sum_{i=1}^{N}\frac{U_{r_2}U_{r_1}+UU_{r_1,r_2}}{E} \\
&\quad \left.+\frac{1}{N}\sum_{i=1}^{N}\frac{2U\left(U_{r_1}E_{r_2}+U_{r_2}E_{r_1}\right)+U^2E_{r_1,r_2}}{2E^2}-\frac{1}{N}\sum_{i=1}^{N}\frac{U^2E_{r_1}E_{r_2}}{E^3}\right\} \\
&\quad +O_p\left(\frac{1}{T^2}\right),
\end{aligned}
$$

or, more concisely,

$$
\begin{aligned}
-\delta_I^{r_2}\nu_{r_1,r_2} &= \tilde{\ell}_{r_1}\left(\theta_0,\bar{\lambda}_i(\theta_0)\right)+\mathcal{D}_{1;r_1}+\delta_I^{r_2}\mathcal{H}_{r_1,r_2}+\delta_I^{r_2}\delta_I^{r_3}\frac{1}{2}\nu_{r_1,r_2,r_3} \\
&\quad +\mathcal{D}_{3;r_1}+\delta_I^{r_2}\mathcal{D}_{2;r_1,r_2}+\frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\mathcal{H}_{r_1,r_2,r_3}+\frac{1}{6}\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\nu_{r_1,r_2,r_3,r_4}+O_p\left(\frac{1}{T^2}\right),
\end{aligned}
$$

which is the desired result. ■

Notice that, by definition, $(\hat{\theta}_{IL}-\theta_0)=[\delta_I^{r_2}]$, where $r_2\in\{1,...,P\}$. The expansion given by (3.29) is, intuitively, a polynomial of $(\hat{\theta}_{IL}-\theta_0)$. To obtain an expansion for $(\hat{\theta}_{IL}-\theta_0)$ that is not a function of itself, (3.29) has to be inverted using the iterative substitution method. This is achieved by repeatedly substituting for $\delta_I^{r_2}$, $\delta_I^{r_3}$ and $\delta_I^{r_4}$.

Lemma 3.A.10

$$\begin{aligned}\delta_I^m &= -\tilde{\ell}_a\nu^{a,m} + \tilde{\ell}_a\nu^{a,b}\mathcal{H}_{c,b}\nu^{c,m} - \mathcal{D}_{1;a}\nu^{a,m} - \frac{1}{2}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\nu_{e,b,d}\nu^{e,m} \\ &+\frac{1}{2}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\nu_{e,b,d}\nu^{e,f}\mathcal{H}_{g,f}\nu^{g,m} - \tilde{\ell}_a\nu^{a,b}\mathcal{H}_{c,b}\nu^{c,d}\mathcal{H}_{e,d}\nu^{e,m} \\ &-\frac{1}{2}\mathcal{D}_{1;a}\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\nu_{e,b,d}\nu^{e,m} + \frac{1}{2}\tilde{\ell}_a\nu^{a,b}\mathcal{H}_{c,b}\nu^{c,d}\tilde{\ell}_e\nu^{e,f}\nu_{g,d,f}\nu^{g,m} \\ &-\frac{1}{4}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\nu_{e,b,d}\nu^{e,f}\tilde{\ell}_g\nu^{g,h}\nu_{i,f,h}\nu^{i,m} - \mathcal{D}_{3;a}\nu^{a,m} \\ &+\tilde{\ell}_a\nu^{a,b}\mathcal{D}_{2;c,b}\nu^{c,m} - \frac{1}{2}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\mathcal{H}_{e,b,d}\nu^{e,m} \\ &+\frac{1}{6}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\tilde{\ell}_e\nu^{e,f}\nu_{g,b,d,f}\nu^{g,m} + O_p\left(\frac{1}{T^2}\right),\end{aligned}$$

where $a, b, c, d, e, f, g, h, i, m \in \{1, ..., P\}$.

Remark 3.A.2 *In what follows, only when* $\nu_{r_1,r_2,...}$ *is concerned, superscripts indicate the corresponding entry of the inverse of* $\nu_{r_1,r_2,...}$. *For example, if the matrix of expectations of second order likelihood derivatives with respect to* θ *is given by* $\nu'' = [\nu_{r_1,r_2}]$, *then* $(\nu'')^{-1} = [\nu^{r_1,r_2}]$.

Proof. The objective is to obtain an expression for a generic element of $(\hat{\theta}_{IL} - \theta_0)$, δ_I^m. To do this, first δ_I^m has to be isolated on the left-hand side. This cannot be done simply by replacing r_2 by m as $\delta_I^{r_2}$ appears on both sides of (3.29). However, notice that if $X^{-1} = [x^{rs}]$ is the inverse of $X = [x_{rs}]$, then

$$x^{rs}x_{st} = \kappa_t^r = \begin{cases} 1 \text{ if } r = t \\ 0 \text{ if } r \neq t \end{cases}.$$

The array κ_t^r is known as Kronecker delta, and $[\kappa_t^r]$ is the identity matrix (note that the common notation for Kronecker delta is δ_t^r; however, as δ is used elsewhere, κ is used here to avoid confusion). Hence,

$$\delta_I^{r_2}\nu_{r_1,r_2}\nu^{r_1,m} = \delta_I^{r_2}\kappa_{r_2}^m = \begin{cases} \delta_I^m \text{ if } r_2 = m \\ 0 \text{ if } r_2 \neq m \end{cases}.$$

Define the following additional notation

$$\begin{aligned}\tilde{\ell}^b &= \tilde{\ell}_a \nu^{a,b}; \quad \mathcal{H}_b^m = \mathcal{H}_{a,b}\nu^{a,m}; \quad \mathcal{H}_{b,c,d,\ldots}^m = \mathcal{H}_{a,b,c,d,\ldots}\nu^{a,m};\\ \mathcal{D}_1^m &= \mathcal{D}_{1;r_1}\nu^{r_1,m}; \quad \mathcal{D}_{2;r_2}^m = \mathcal{D}_{2;r_1,r_2}\nu^{r_1,m}; \quad \mathcal{D}_3^m = \mathcal{D}_{3;r_1}\nu^{r_1,m},\end{aligned}$$

and remember that superscripts indicate the inverse for ν only. Then, multiplying both sides of (3.29) by $\nu^{r_1,m}$

$$\begin{aligned}\delta_I^m &= -\Big[\tilde{\ell}_{r_1}\nu^{r_1,m} + \mathcal{D}_{1;r_1}\nu^{r_1,m} + \delta_I^{r_2}\mathcal{H}_{r_1,r_2}\nu^{r_1,m} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\nu_{r_1,r_2,r_3}\nu^{r_1,m} + \mathcal{D}_{3;r_1}\nu^{r_1,m}\\ &\quad + \delta_I^{r_2}\mathcal{D}_{2;r_1,r_2}\nu^{r_1,m} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\mathcal{H}_{r_1,r_2,r_3}\nu^{r_1,m} + \frac{1}{6}\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\nu_{r_1,r_2,r_3,r_4}\nu^{r_1,m}\Big]\\ &\quad + O_p\left(\frac{1}{T^2}\right). \qquad (3.30)\end{aligned}$$

The iterative substitution method can now be conducted. For convenience, write $\delta_I^{r_2}, \delta_I^{r_3}$ and $\delta_I^{r_4}$ as follows, on the basis of (3.30).

$$\begin{aligned}\delta_I^{r_2} &= -\Big[\tilde{\ell}_a\nu^{a,r_2} + \mathcal{D}_{1;a}\nu^{a,r_2} + \delta_I^b\mathcal{H}_{a,b}\nu^{a,r_2} + \frac{1}{2}\delta_I^b\delta_I^c\nu_{a,b,c}\nu^{a,r_2} + \mathcal{D}_{3;a}\nu^{a,r_2}\\ &\quad + \delta_I^b\mathcal{D}_{2;a,b}\nu^{a,r_2} + \frac{1}{2}\delta_I^b\delta_I^c\mathcal{H}_{a,b,c}\nu^{a,r_2} + \frac{1}{6}\delta_I^b\delta_I^c\delta_I^d\nu_{a,b,c,d}\nu^{a,r_2}\Big] + O_p\left(\frac{1}{T^2}\right),\\ \delta_I^{r_3} &= -\Big[\tilde{\ell}_e\nu^{e,r_3} + \mathcal{D}_{1;e}\nu^{e,r_3} + \delta_I^f\mathcal{H}_{e,f}\nu^{e,r_3} + \frac{1}{2}\delta_I^f\delta_I^g\nu_{e,f,g}\nu^{e,r_3} + \mathcal{D}_{3;e}\nu^{e,r_3}\\ &\quad + \delta_I^f\mathcal{D}_{2;e,f}\nu^{e,r_3} + \frac{1}{2}\delta_I^f\delta_I^g\mathcal{H}_{e,f,g}\nu^{e,r_3} + \frac{1}{6}\delta_I^f\delta_I^g\delta_I^h\nu_{e,f,g,h}\nu^{e,r_3}\Big] + O_p\left(\frac{1}{T^2}\right), \qquad (3.31)\\ \delta_I^{r_4} &= -\Big[\tilde{\ell}_i\nu^{i,r_4} + \mathcal{D}_{1;i}\nu^{i,r_4} + \delta_I^j\mathcal{H}_{i,j}\nu^{i,r_4} + \frac{1}{2}\delta_I^j\delta_I^k\nu_{i,j,k}\nu^{i,r_4} + \mathcal{D}_{3;i}\nu^{i,r_4}\\ &\quad + \delta_I^j\mathcal{D}_{2;i,j}\nu^{i,r_4} + \frac{1}{2}\delta_I^j\delta_I^k\mathcal{H}_{i,j,k}\nu^{i,r_4} + \frac{1}{6}\delta_I^j\delta_I^k\delta_I^l\nu_{i,j,k,l}\nu^{i,r_4}\Big] + O_p\left(\frac{1}{T^2}\right). \qquad (3.32)\end{aligned}$$

Notice that a different set of dummy indices is used in each case, to avoid confusion. Now, start by substituting for $\delta_I^{r_2}$ to obtain

$$\begin{aligned}\delta_I^m &= -\tilde{\ell}_{r_1}\nu^{r_1,m} - \mathcal{D}_{1;r_1}\nu^{r_1,m}\\ &\quad + \left(\tilde{\ell}_a\nu^{a,r_2} + \mathcal{D}_{1;a}\nu^{a,r_2} + \delta_I^b\mathcal{H}_{a,b}\nu^{a,r_2} + \frac{1}{2}\delta_I^b\delta_I^c\nu_{a,b,c}\nu^{a,r_2}\right)\mathcal{H}_{r_1,r_2}\nu^{r_1,m}\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\left(\tilde{\ell}_a\nu^{a,r_2}+\mathcal{D}_{1;a}\nu^{a,r_2}+\delta_I^b\mathcal{H}_{a,b}\nu^{a,r_2}+\frac{1}{2}\delta_I^b\delta_I^c\nu_{a,b,c}\nu^{a,r_2}\right)\delta_I^{r_3}\nu_{r_1,r_2,r_3}\nu^{r_1,m}\\
&-\mathcal{D}_{3;r_1}\nu^{r_1,m}+\tilde{\ell}_a\nu^{a,r_2}\mathcal{D}_{2;r_1,r_2}\nu^{r_1,m}+\frac{1}{2}\tilde{\ell}_a\nu^{a,r_2}\delta_I^{r_3}\mathcal{H}_{r_1,r_2,r_3}\nu^{r_1,m}\\
&+\frac{1}{6}\tilde{\ell}_a\nu^{a,r_2}\delta_I^{r_3}\delta_I^{r_4}\nu_{r_1,r_2,r_3,r_4}\nu^{r_1,m}+O_p\left(\frac{1}{T^2}\right)\\
=\;&-\tilde{\ell}_{r_1}\nu^{r_1,m}-\mathcal{D}_{1;r_1}\nu^{r_1,m}\\
&+\left(\tilde{\ell}_a\nu^{a,r_2}+\mathcal{D}_{1;a}\nu^{a,r_2}-\tilde{\ell}_w\nu^{w,b}\mathcal{H}_{a,b}\nu^{a,r_2}+\frac{1}{2}\tilde{\ell}_w\nu^{w,b}\tilde{\ell}_y\nu^{y,c}\nu_{a,b,c}\nu^{a,r_2}\right)\mathcal{H}_{r_1,r_2}\nu^{r_1,m}\\
&+\frac{1}{2}\left(\tilde{\ell}_a\nu^{a,r_2}+\mathcal{D}_{1;a}\nu^{a,r_2}-\tilde{\ell}_w\nu^{w,b}\mathcal{H}_{a,b}\nu^{a,r_2}+\frac{1}{2}\tilde{\ell}_w\nu^{w,b}\tilde{\ell}_y\nu^{y,c}\nu_{a,b,c}\nu^{a,r_2}\right)\delta_I^{r_3}\nu_{r_1,r_2,r_3}\nu^{r_1,m}\\
&-\mathcal{D}_{3;r_1}\nu^{r_1,m}+\tilde{\ell}_a\nu^{a,r_2}\mathcal{D}_{2;r_1,r_2}\nu^{r_1,m}+\frac{1}{2}\tilde{\ell}_a\nu^{a,r_2}\delta_I^{r_3}\mathcal{H}_{r_1,r_2,r_3}\nu^{r_1,m}\\
&+\frac{1}{6}\tilde{\ell}_a\nu^{a,r_2}\delta_I^{r_3}\delta_I^{r_4}\nu_{r_1,r_2,r_3,r_4}\nu^{r_1,m}+O_p\left(\frac{1}{T^2}\right).
\end{aligned}
$$

Next, (3.31) and (3.32) are substituted for $\delta_I^{r_3}$ and $\delta_I^{r_4}$, respectively, which yields

$$
\begin{aligned}
\delta_I^m \;=\;&-\tilde{\ell}_{r_1}\nu^{r_1,m}-\mathcal{D}_{1;r_1}\nu^{r_1,m}\\
&+\left(\tilde{\ell}_a\nu^{a,r_2}+\mathcal{D}_{1;a}\nu^{a,r_2}-\tilde{\ell}_w\nu^{w,b}\mathcal{H}_{a,b}\nu^{a,r_2}+\frac{1}{2}\tilde{\ell}_w\nu^{w,b}\tilde{\ell}_y\nu^{y,c}\nu_{a,b,c}\nu^{a,r_2}\right)\mathcal{H}_{r_1,r_2}\nu^{r_1,m}\\
&-\frac{1}{2}\left(\tilde{\ell}_a\nu^{a,r_2}+\mathcal{D}_{1;a}\nu^{a,r_2}-\tilde{\ell}_w\nu^{w,b}\mathcal{H}_{a,b}\nu^{a,r_2}+\frac{1}{2}\tilde{\ell}_w\nu^{w,b}\tilde{\ell}_y\nu^{y,c}\nu_{a,b,c}\nu^{a,r_2}\right)\tilde{\ell}_e\nu^{e,r_3}\nu_{r_1,r_2,r_3}\nu^{r_1,m}\\
&-\mathcal{D}_{3;r_1}\nu^{r_1,m}+\tilde{\ell}_a\nu^{a,r_2}\mathcal{D}_{2;r_1,r_2}\nu^{r_1,m}-\frac{1}{2}\tilde{\ell}_a\nu^{a,r_2}\tilde{\ell}_e\nu^{e,r_3}\mathcal{H}_{r_1,r_2,r_3}\nu^{r_1,m}\\
&+\frac{1}{6}\tilde{\ell}_a\nu^{a,r_2}\tilde{\ell}_e\nu^{e,r_3}\tilde{\ell}_i\nu^{i,r_4}\nu_{r_1,r_2,r_3,r_4}\nu^{r_1,m}+O_p\left(\frac{1}{T^2}\right).
\end{aligned}
$$

Finally, ordering terms according to the stochastic order of magnitude and redefining the dummy indices to simplify the expression, the asymptotic expansion for δ_I^m is given by,

$$
\begin{aligned}
\delta_I^m \;=\;&-\tilde{\ell}_a\nu^{a,m}+\tilde{\ell}_a\nu^{a,b}\mathcal{H}_{c,b}\nu^{c,m}-\mathcal{D}_{1;a}\nu^{a,m}-\frac{1}{2}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\nu_{e,b,d}\nu^{e,m}\\
&+\frac{1}{2}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\nu_{e,b,d'}\nu^{e,f}\mathcal{H}_{g,f}\nu^{g,m}-\tilde{\ell}_a\nu^{a,b}\mathcal{H}_{c,b}\nu^{c,d}\mathcal{H}_{e,d}\nu^{e,m}\\
&-\frac{1}{2}\mathcal{D}_{1;a}\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\nu_{e,b,d}\nu^{e,m}+\frac{1}{2}\tilde{\ell}_a\nu^{a,b}\mathcal{H}_{c,b}\nu^{c,d}\tilde{\ell}_e\nu^{e,f}\nu_{g,d,f}\nu^{g,m}\\
&-\frac{1}{4}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\nu_{e,b,d}\nu^{e,f}\tilde{\ell}_g\nu^{g,h}\nu_{i,f,h}\nu^{i,m}-\mathcal{D}_{3;a}\nu^{a,m}+\tilde{\ell}_a\nu^{a,b}\mathcal{D}_{2;c,b}\nu^{c,m}\\
&-\frac{1}{2}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\mathcal{H}_{e,b,d}\nu^{e,m}+\frac{1}{6}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\tilde{\ell}_e\nu^{e,f}\nu_{g,b,d,f}\nu^{g,m}\\
&+O_p\left(\frac{1}{T^2}\right),
\end{aligned}
$$

which proves Lemma 3.A.10. ■

Based on these results, the proof of Theorem 3.5.1 now follows.

Proof (Theorem 3.5.1). Follows directly from Lemma 3.A.10, by observing that, $-\tilde{\ell}_a \nu^{a,m}$ is $O_p(N^{-\rho_1/2}T^{-1/2})$, $\tilde{\ell}_a \nu^{a,b}\mathcal{H}_{c,b}\nu^{c,m} - \mathcal{D}_{1;a}\nu^{a,m} - \frac{1}{2}\tilde{\ell}_a \nu^{a,b}\tilde{\ell}_c \nu^{c,d}\nu_{e,b,d}\nu^{e,m}$ is $O_p(T^{-1})$ and the remaining terms up to the $O_p(T^{-2})$ remainder are all at most $O_p\left(T^{-3/2}\right)$ independent of the particular values of ρ_1, ρ_2 and ρ_3. Then, writing the first two lines in matrix notation finally gives (3.6). ■

Finally, this section ends with the proof of Lemma 3.A.8.

Proof (Lemma 3.A.8). The proof of Lemma 3.A.8 is tedious but straightforward. To save space, proofs for $V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2 (E_{iT})^{-2}$ and $\ln(-E_{iT})$ will be given only. The rest of the proofs follow along similar lines. Start with $\ln(-E_{iT})$. To keep notation simple, define $E = E_{iT}$, which is a scalar. Then,

$$\begin{aligned}
\nabla_\theta \ln(-E_{iT}) &= -\frac{E_{r_1}}{E}, \\
\nabla_{\theta\theta} \ln(-E_{iT}) &= -\frac{E_{r_1,r_2}}{E} + \frac{E_{r_1}E_{r_2}}{E^2} \\
\nabla_{\theta\theta\theta} \ln(-E_{iT}) &= -\frac{E_{r_1,r_2,r_3}}{E} + \frac{E_{r_1,r_2}E_{r_3} + E_{r_1,r_3}E_{r_2} + E_{r_2,r_3}E_{r_1}}{E^2} - \frac{2E_{r_1}E_{r_2}E_{r_3}}{E^3}, \\
&= -\frac{E_{r_1,r_2,r_3}}{E} + \frac{E_{r_1,r_2}E_{r_3}[3]}{E^2} - \frac{2E_{r_1}E_{r_2}E_{r_3}}{E^3},
\end{aligned}$$

where numbers in brackets denote all possible permutations of the free indices. For example $E_{r_1,r_2}E_{r_3}[3] = E_{r_1,r_2}E_{r_3} + E_{r_1,r_3}E_{r_2} + E_{r_2,r_3}E_{r_1}$.Then,

$$\begin{aligned}
\nabla_{\theta\theta\theta} \ln(-E_{iT}) &= -\frac{E_{r_1,r_2,r_3,r_4}}{E} + \frac{E_{r_1,r_2,r_3}E_{r_4}}{E^2} \\
&\quad + \frac{E_{r_1,r_2,r_4}E_{r_3} + E_{r_1,r_2}E_{r_3,r_4} + E_{r_1,r_3,r_4}E_{r_2}}{E^2} \\
&\quad + \frac{E_{r_1,r_3}E_{r_2,r_4} + E_{r_2,r_3,r_4}E_{r_1} + E_{r_2,r_3}E_{r_1,r_4}}{E^2} \\
&\quad - \frac{2\left(E_{r_1,r_2}E_{r_3} + E_{r_1,r_3}E_{r_2} + E_{r_2,r_3}E_{r_1}\right)E_{r_4}}{E^3} \\
&\quad - \frac{2E_{r_1,r_4}E_{r_2}E_{r_3} + 2E_{r_1}E_{r_2,r_4}E_{r_3} + 2E_{r_1}E_{r_2}E_{r_3,r_4}}{E^3} \\
&\quad + \frac{6E_{r_1}E_{r_2}E_{r_3}E_{r_4}}{E^4} \\
&= -\frac{E_{r_1,r_2,r_3,r_4}}{E} + \frac{E_{r_1}E_{r_2,r_3,r_4}[4] + E_{r_1,r_2}E_{r_3,r_4}[3]}{E^2} \\
&\quad -2\frac{E_{r_1,r_2}E_{r_3}E_{r_4}[6]}{E^3} + \frac{6E_{r_1}E_{r_2}E_{r_3}E_{r_4}}{E^4}.
\end{aligned}$$

Now, consider $V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2\left(E_{iT}\right)^{-2}$. Write it as $V\ell^2E^{-2}$. Then,

$$\begin{aligned}
\nabla_{\theta}\left[\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^2}\right] &= \frac{V_{r_1}U^2+2VUU_{r_1}}{E^2}-2\frac{VU^2E_{r_1}}{E^3},\\
\nabla_{\theta\theta}\left[\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^2}\right] &= \frac{V_{r_1,r_2}U^2+2V_{r_1}UU_{r_2}[2]+2VU_{r_2}U_{r_1}+2VUU_{r_1,r_2}}{E^2}\\
&\quad -2\frac{V_{r_1}U^2E_{r_2}[2]+2VUU_{r_1}E_{r_2}[2]+VU^2E_{r_1,r_2}}{E^3}\\
&\quad +6\frac{VU^2E_{r_1}E_{r_2}}{E^4}.
\end{aligned}$$

The third order derivative is then given by

$$\begin{aligned}
\nabla_{\theta\theta\theta}\left[\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^2}\right] &= \frac{V_{r_1,r_2,r_3}U^2+2V_{r_1,r_2}UU_{r_3}[3]+2V_{r_1}U_{r_3}U_{r_2}[3]}{E^2}\\
&\quad +2\frac{V_{r_1}UU_{r_2,r_3}[3]+VU_{r_2,r_3}U_{r_1}[3]+VUU_{r_1,r_2,r_3}}{E^2}\\
&\quad -2\frac{V_{r_1,r_2}U^2E_{r_3}[3]+2V_{r_1}UU_{r_2}E_{r_3}[6]}{E^3}\\
&\quad -2\frac{2VU_{r_2}U_{r_1}E_{r_3}[3]+2VUU_{r_1,r_2}E_{r_3}[3]}{E^3}\\
&\quad -2\frac{V_{r_1}U^2E_{r_2,r_3}[3]+2VUU_{r_1}E_{r_2,r_3}[3]+VU^2E_{r_1,r_2,r_3}}{E^3}\\
&\quad +6\frac{V_{r_1}U^2E_{r_2}E_{r_3}[3]+2VUU_{r_1}E_{r_2}E_{r_3}[3]}{E^4}\\
&\quad +6\frac{VU^2E_{r_1,r_2}E_{r_3}[3]}{E^4}-24\frac{VU^2E_{r_1}E_{r_2}E_{r_3}}{E^5}\\
&= O_p\left(\frac{1}{T^{3/2}}\right)
\end{aligned}$$

Lastly,

$$\begin{aligned}
\nabla_{\theta\theta\theta\theta}\left[\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^2}\right] &= \frac{V_{r_1,r_2,r_3,r_4}U^2+2V_{r_1,r_2,r_3}UU_{r_4}[4]+2VUU_{r_1,r_2,r_3,r_4}}{E^2}\\
&\quad +2\frac{V_{r_1,r_2}U_{r_4}U_{r_3}[6]+V_{r_1,r_2}UU_{r_3,r_4}[6]+V_{r_1}U_{r_3,r_4}U_{r_2}[12]}{E^2}\\
&\quad +2\frac{V_{r_1}UU_{r_2,r_3,r_4}[4]+VU_{r_2,r_3,r_4}U_{r_1}[4]+VU_{r_2,r_3}U_{r_1,r_4}[3]}{E^2}\\
&\quad -2\frac{V_{r_1,r_2,r_3}U^2E_{r_4}[4]+2V_{r_1,r_2}UU_{r_3}E_{r_4}[12]+2V_{r_1}U_{r_3}U_{r_2}E_{r_4}[12]}{E^3}\\
&\quad -4\frac{V_{r_1}UU_{r_2,r_3}E_{r_4}[12]+VU_{r_2,r_3}U_{r_1}E_{r_4}[12]+VUU_{r_1,r_2,r_3}E_{r_4}[4]}{E^3}\\
&\quad -2\frac{V_{r_1,r_2}U^2E_{r_3,r_4}[6]+V_{r_1}U^2E_{r_2,r_3,r_4}[4]+VU^2E_{r_1,r_2,r_3,r_4}}{E^3}
\end{aligned}$$

$$
\begin{aligned}
&-4\frac{V_{r_1}UU_{r_2}E_{r_3,r_4}[12]+VU_{r_2}U_{r_1}E_{r_3,r_4}[6]}{E^3}\\
&-4\frac{VUU_{r_1,r_2}E_{r_3,r_4}[6]+VUU_{r_1}E_{r_2,r_3,r_4}[4]}{E^3}\\
&+6\frac{V_{r_1,r_2}U^2E_{r_3}E_{r_4}[6]+2V_{r_1}UU_{r_2}E_{r_3}E_{r_4}[12]}{E^4}\\
&+6\frac{2VU_{r_2}U_{r_1}E_{r_3}E_{r_4}[6]+2VUU_{r_1,r_2}E_{r_3}E_{r_4}[6]}{E^4}\\
&+6\frac{V_{r_1}U^2E_{r_2,r_3}E_{r_4}[12]+2VUU_{r_1}E_{r_2,r_3}E_{r_4}[12]}{E^4}\\
&+6\frac{VU^2E_{r_1,r_2,r_3}E_{r_4}[4]+VU^2E_{r_1,r_2}E_{r_3,r_4}[3]}{E^4}\\
&-24\frac{V_{r_1}U^2E_{r_2}E_{r_3}E_{r_4}[4]+2VUU_{r_1}E_{r_2}E_{r_3}E_{r_4}[4]}{E^5}\\
&-24\frac{VU^2E_{r_1,r_2}E_{r_3}E_{r_4}[6]}{E^5}\\
&+120\frac{VU^2E_{r_1}E_{r_2}E_{r_3}E_{r_4}}{E^6},
\end{aligned}
$$

which is $O_p(T^{-3/2})$, as desired. ■

3.A.5 Proof of Theorem 3.6.2

Proof (First Part). The first result directly follows from Theorem 1 of Jenish and Prucha (2009). Therefore, the first part of the proof consists of verification of Assumptions 1-5 in Jenish and Prucha (2009). To avoid confusion, these will be called Assumptions JP1-JP5. As indices are assumed to be located on an integer lattice $D \subseteq \mathbb{Z}^d$, where $d > 0$, increasing domain asymptotics is implied, which verifies Assumption JP1.

Next, consider

$$\lim_{k\to\infty}\sup_{i,T}\mathbb{E}[|Z_{iT}|^{2+\delta}\,\mathbf{1}_{(Z_{iT})>k}],$$

where $\mathbf{1}_{(\cdot)}$ is the indicator function. Define $\mathbb{E}_{\mathcal{A}}[\cdot]$, the expectation taken over the set $\{Z_{iT} : |Z_{iT}| > k\}$. Then,

$$\mathbb{E}[|Z_{iT}|^{2+\delta}\,|Z_{iT}|^{\varepsilon}\,|Z_{iT}|^{-\varepsilon}\,\mathbf{1}_{(|Z_{iT}|)>k}] = \mathbb{E}_{\mathcal{A}}[|Z_{iT}|^{2+\delta}\,|Z_{iT}|^{\varepsilon}\,|Z_{iT}|^{-\varepsilon}],$$

for some $\varepsilon > 0$. Observe that for some $|Z_{iT}| > k$, $|Z_{iT}|^{-\varepsilon} > k^{-\varepsilon}$. Hence,

$$\mathbb{E}_{\mathcal{A}}[|Z_{iT}|^{2+\delta}\,|Z_{iT}|^{\varepsilon}\,|Z_{iT}|^{-\varepsilon}] \;<\; \mathbb{E}_{\mathcal{A}}[|Z_{iT}|^{2+\delta}\,|Z_{iT}|^{\varepsilon}\,k^{-\varepsilon}]$$

$$
\begin{aligned}
&= k^{-\varepsilon}\mathbb{E}_{\mathcal{A}}[|Z_{iT}|^{2+\delta}\,|Z_{iT}|^{\varepsilon}] \\
&\leq k^{-\varepsilon}\mathbb{E}[|Z_{iT}|^{2+\delta+\varepsilon}].
\end{aligned}
$$

By Assumption 3.6.4, $\sup_{i,T}\mathbb{E}[|Z_{iT}|^{\tilde{\varepsilon}}] < \infty$ for $\tilde{\varepsilon} > 2+\delta$, so $\sup_{i,T}\mathbb{E}[|Z_{iT}|^{2+\delta+\varepsilon}] < \infty$. Therefore,

$$
\sup_{i,T}\mathbb{E}_{\mathcal{A}}[|Z_{iT}|^{2+\delta}\,|Z_{iT}|^{\varepsilon}\,|Z_{iT}|^{-\varepsilon}] \leq k^{-\varepsilon}\sup_{i,T}\mathbb{E}[|Z_{iT}|^{2+\delta+\varepsilon}],
$$

and

$$
\lim_{k\to\infty}\sup_{i,T}\mathbb{E}_{\mathcal{A}}[|Z_{iT}|^{2+\delta}\,|Z_{iT}|^{\varepsilon}\,|Z_{iT}|^{-\varepsilon}] \leq \lim_{k\to\infty} k^{-\varepsilon}O(1) = 0.
$$

Hence, Z_{iT} is $L_{2+\delta}$-bounded Uniformly over i and T for some $\delta > 0$:

$$
\lim_{k\to\infty}\sup_{i,T}\mathbb{E}[|Z_{iT}|^{2+\delta}\,\mathbf{1}_{(|Z_{iT}|)>k}] = 0. \tag{3.33}
$$

In addition, Assumption 3.6.5(a) implies that $\sum_{m=1}^{\infty} m^{d-1}\alpha_{1,1}(m)^{\delta/(2+\delta)} < \infty$, which in turn implies that

$$
\sum_{m=1}^{\infty}\alpha_{1,1}(m)m^{[d(2+\delta)/\delta]-1} < \infty,
$$

as shown by Jenish and Prucha (2009). Therefore, by their Corollary 1, Assumptions JP2 and JP3(a) are also satisfied. Assumptions JP3(b)-(c) are exactly the same as Assumptions 3.6.5(b)-(c) and are directly verified. Finally, Assumption 3.6.6 corresponds to Assumption JP5, in the setting of this study. Then, by their Theorem 1,

$$
\sqrt{L_N}\frac{L_N^{-1}\sum_{i\in\mathcal{G}_g} Z_{iT}}{\sqrt{Var\left(L_N^{-1/2}\sum_{i\in\mathcal{G}_g} Z_{iT}\right)}} \xrightarrow{d} N(0,1).
$$

for all $g \in \{1,...,G\}$, implying that $Var\left(L_N^{-1}\sum_{i\in\mathcal{G}_g} Z_{iT}\right) = O\left(L_N^{-1}\right)$. Therefore,

$$
\begin{aligned}
\frac{1}{G_N^2}\sum_{g=1}^{G}\frac{1}{L_N^2}\sum_{i,j\in\mathcal{G}_g}\sum Cov(Z_{iT},Z_{jT}) &= \frac{1}{G_N^2}G_N O\left(\frac{1}{L_N}\right) \\
&= O\left(\frac{1}{G_N L_N}\right) \\
&= O\left(\frac{1}{N}\right),
\end{aligned}
$$

which proves the first result. ■

Proof (Second Part). The proof of the second result follows directly from the reasoning in the proof of Lemma 1 in Bester, Conley and Hansen (2011). The following is simply a statement of their discussion. The object of interest is

$$\frac{1}{N^2}\sum_{g\neq h}^{G}\sum^{G}\sum_{i\in\mathcal{G}_g}\sum_{j\in\mathcal{G}_h} Cov(Z_{iT}, Z_{jT}).$$

The key is to find a bound on the covariances and on the maximum number of pairs of individuals from different clusters g and h.Start by bounding the number of neighbours for any given individual. Based on the distance metric, 1-order neighbours are those individuals that lie one unit away from the selected individual. Then, 2-order neighbours are given by all points that are two units away. This generalises to m-order neighbours. The largest number of such neighbours for any individual is given by $C(d)m^{d-1}$ and naturally it depends on the order of the neighbourhood and the dimension. *But what is the number of individuals one has to consider?* Imagine each group as a collection of contour sets; that is, concentric sets starting from the boundary and moving towards the inside of the group one unit at a time. For example, the first contour set is the boundary, the second is the set of points one unit away from the boundary, the third is the set of points two points away from the boundary. For the group g, denote $\partial_1\mathcal{G}_g$ the first contour set, $\partial_2\mathcal{G}_g$ the second contour set etc. Now consider m-order neighbours from two different groups and remember that the groups are contiguous by Assumption 3.6.3. Then, these neighbours can possibly reside in m different pairs of contour sets. For instance, for two groups g and h, 3-order neighbours can be residing in the following pairs of contour sets: $(\partial_1\mathcal{G}_g, \partial_1\mathcal{G}_h), (\partial_1\mathcal{G}_g, \partial_2\mathcal{G}_h), (\partial_1\mathcal{G}_g, \partial_3\mathcal{G}_h), (\partial_2\mathcal{G}_g, \partial_1\mathcal{G}_h), (\partial_2\mathcal{G}_g, \partial_2\mathcal{G}_h), (\partial_3\mathcal{G}_g, \partial_1\mathcal{G}_h)$; notice that it is still possible to find, 3-order neighbours in two contour sets that are, say, only one unit apart from each other. Hence, the following can be determined for a given *individual*: the bound on the maximum number of m-order neighbours and the fact that these neighbours may reside on a maximum of m pairs of contours. But how many such individuals can there be in a given contour set? Observe that the largest contour set will be the boundary and there already is a bound on the number of individuals on the boundary by Assumption 3.6.2. Therefore, the maximum number of m-order neighbours for two given groups g and

h is given by

$$\kappa_d m^d L_N^{(d-1)/d} \quad \text{where} \quad \kappa_d = 2C(d)C.$$

This implies that

$$\sum_{i\in\mathcal{G}_g}\sum_{j\in\mathcal{G}_h} Cov(Z_{iT}, Z_{jT}) \leq \sum_{m=1}^{\infty} \kappa_d m^d L_N^{(d-1)/d} Cov(Z_{iT}, Z_{jT}).$$

By Lemma 1 of Bolthausen (1982), which is based on Ibragimov and Linnik (1971),

$$Cov(Z_{iT}, Z_{jT}) \leq c_\delta \alpha_{1,1}(m)^{\delta/(2+\delta)} \left|\left|Z_{iT}\right|\right|_{2+\delta} \left|\left|Z_{jT}\right|\right|_{2+\delta},$$

where c_δ is a constant depending on δ and $||\cdot||_{2+\delta} = \{\mathbb{E}[|\cdot|^{2+\delta}]\}^{1/(2+\delta)}$ is the $L_{2+\delta}$-norm. Then, by Assumption 3.6.4, $||Z_{iT}||_{2+\delta} < \infty$ for all i and T and $Cov(Z_{iT}, Z_{jT}) \leq c_\delta \alpha_{1,1}(m)^{\delta/(2+\delta)}$ leading to

$$\begin{aligned}
\sum_{i\in\mathcal{G}_g}\sum_{j\in\mathcal{G}_h} Cov(Z_{iT}, Z_{jT}) &\leq \sum_{m=1}^{\infty} \kappa_d m^d L_N^{(d-1)/d} c_\delta \alpha_{1,1}(m)^{\delta/(2+\delta)} \\
&= c_\delta \kappa_d L_N^{(d-1)/d} \sum_{m=1}^{\infty} m^d \alpha_{1,1}(m)^{\delta/(2+\delta)} \\
&= O\left(L_N^{(d-1)/d}\right),
\end{aligned}$$

where the last equality follows from Assumption 3.6.5(a). Then,

$$\begin{aligned}
\frac{1}{N^2}\sum_{g\neq h}^{G_N}\sum^{G_N}\sum_{i\in\mathcal{G}_g}\sum_{j\in\mathcal{G}_h} Cov(Z_{iT}, Z_{jT}) &\leq \frac{1}{N^2}\sum_{g\neq h}^{G_N}\sum^{G_N} O\left(L_N^{(d-1)/d}\right) \\
&= \frac{1}{N^2} O\left(G_N^2 L_N^{(d-1)/d}\right) \\
&= \frac{1}{N^2} O\left(N^2 L_N^{-(d+1)/d}\right) \\
&= O\left(L_N^{-(d+1)/d}\right),
\end{aligned}$$

which proves the second result. ■

Chapter 4

Bias Reduction in GARCH Panels

4.1 Introduction

The idea of modelling univariate volatility using panels, rather than individual time-series, of financial returns was introduced in Chapter 2. As simulation results reveal, the GARCH Panel method is subject to the incidental parameter issue. However, once the incidental parameter issue disappears, the in-sample fit of the model increases greatly, even with around 250 observations. Of course, stock returns and many other financial variables are available at daily or higher frequency so a large dataset where the incidental parameter issue will not be relevant can always be collected. However, as mentioned previously, some variables are recorded at monthly or quarterly frequency (e.g. inflation, GDP and hedge fund returns) implying that large enough datasets will not available unless there are decades of observations. In the case of hedge fund returns this is an important issue as most datasets go back to January 1994 implying around 200 monthly returns at most. When the interest is in modelling the volatility of such a dataset, a method for obtaining estimators that are robust to the incidental parameter bias is called for. Proposing such a method by combining the results of Chapters 2 and 3 is the objective of this chapter.

The bias reduction analysis of Chapter 3 was motivated from the microeconometrician's point of view, given that recent research in analytical bias reduction is largely rooted in the microeconometrics literature. However, it must be emphasized that the incidental parameter issue is not peculiar to a particular literature. Instead, it is a statistical problem which arises due to the presence of individual-specific parameters that cannot be estimated accurately. The exact nature of why these individual-specific parameters exist is of no con-

sequence. In other words, whether these parameters represent unobserved heterogeneity or are the intercepts of the GARCH(1,1) model does not make any difference statistically (other things being equal). Therefore, both applications are particular instances of the incidental parameter issue and the methods developed in the previous chapter can in principle be utilised to correct for the small-sample bias of the GARCH Panel estimators.

This chapter analyses the properties of the bias-corrected composite likelihood estimator by simulation analyses and empirical exercises. Investigating whether and under which conditions the high level assumptions of Chapter 3 hold for the GARCH Panel model is a very interesting research question in its own right. However, this analysis is not trivial and would warrant a separate study beyond the scope of this thesis. Therefore, this is not pursued here. Instead, the relevance of the bias-correction approach of Chapter 3 will be checked by using simulation analysis and a test of predictive ability.

Clearly, an interesting and obvious idea would be to extend the panel modelling approach beyond the baseline GARCH(1,1) model. As mentioned in Section 2.2, one limitation of this model is that it cannot take asymmetric effects into account as it is based on squared shocks, ε_{it}^2. Therefore, shocks of the same magnitude but different signs are assumed to effect volatility identically. However, generally volatility increases more in response to a negative shock than to a positive shock. This is also known as the "leverage effect." Two alternatives that allow for such dynamics are the Exponential GARCH (EGARCH) and the GJR-GARCH models due to Nelson (1991) and Glosten, Jagannathan and Runkle (1993), respectively. Therefore, especially from an empirical perspective, it would be natural to consider "EGARCH panels" or "GJR-GARCH panels," and their bias-corrected versions, as well. Eventually, a complete research project focusing on different GARCH-type panels and their theoretical, as well as empirical, properties would be a substantial contribution to the financial econometrics literature. Unfortunately, this is beyond the scope of this thesis and is left for future research.

The first part of the chapter discusses why the integrated likelihood method is the appropriate bias reduction approach. Also, calculation of the priors, the integrated likelihood and the estimation algorithm are outlined.

The next part of this chapter is dedicated to a simulation analysis of the performance of the Arellano-Bonhomme robust priors in GARCH panels across different panel dimensions.

In addition to the standard integrated likelihood estimation, a slightly different estimation strategy which takes initial value selection into account is also used. In addition, different dependence settings are considered in order to analyse the behaviour of the first-order bias when the cross-section independence assumption is relaxed. Under a reasonable dependence structure, it is shown that the effect of cross-section dependence is not on the first-order bias but on the variance of the estimator.

Finally two empirical illustrations of panel GARCH estimation in small samples are considered. The first illustration is concerned with volatility forecasting using daily stock market data, similar in spirit to the empirical analysis in Chapter 2. This provides an objective environment where the usefulness of the integrated likelihood method can be proved. The second illustration is an analysis of monthly hedge fund volatility. This is a novel contribution to the literature, as such analysis has hitherto been impossible using the standard GARCH methodology, due to hedge fund returns being recorded at monthly frequency. This illustration reveals that, not surprisingly, hedge fund volatility is asymmetric, skewed to right (towards high volatility observations) and its shape changes over time and in reaction to major economic events, such as the burst of the dotcom bubble and the credit crunch episode.

The rest of this study is organised as follows: Section 4.2 discusses the implementation of analytical bias reduction in GARCH panels. In particular, the algorithm used in this study for calculation of the robust priors and the integrated likelihood is detailed. Section 4.3 provides a simulation analysis to investigate the small sample properties and the bias-reduction performance of the integrated likelihood method in GARCH panels. Next, empirical applications for the bias-corrected GARCH Panel estimator are illustrated in Section 4.4. Finally, Section 4.5 concludes.

4.2 Bias Reduction for GARCH Panels

The general setup of this chapter is the same as in Chapter 2. Briefly, the random variable of interest is given by y_{it} where

$$y_{it} = \mathbb{E}[y_{it}|\mathcal{F}_{i,t-1}] + \varepsilon_{it}, \quad t = 1, ..., T \quad \text{and} \quad i = 1, ..., N.$$

As before, y_{it} can be the daily/monthly return on some financial asset i at time t. Also, again, $\mathcal{F}_{it}$ is the information set for individual i at time t and, to focus exclusively on modelling conditional volatility, it is assumed that $\mathbb{E}[y_{it}|\mathcal{F}_{i,t-1}] = 0$. Volatility is modelled using the GARCH(1,1) specification for the conditional variance where,

$$\begin{aligned} \varepsilon_{it} &= \sigma_{it}\eta_{it}, \quad \eta_{it} \sim F, \quad \mathbb{E}[\eta_{it}] = 0, \quad Var(\eta_{it}) = 1, \\ \sigma_{it}^2 &= \lambda_i(1-\alpha-\beta) + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2, \\ \lambda_i &> 0 \ \forall i; \quad \alpha, \beta \geq 0 \quad \text{and} \quad \alpha + \beta < 1, \end{aligned} \tag{4.1}$$

for some distribution, F. Unlike in the standard setting, η_{it} are dependent across i, while an iid assumption across t is maintained. Define $\theta = (\alpha, \beta)$, $\lambda = (\lambda_1, ..., \lambda_N)$ and the (pseudo) true parameter values $\theta_0 = (\alpha_0, \beta_0)$ and $\lambda_0 = (\lambda_{10}, ..., \lambda_{N0})$.

4.2.1 A Discussion on the Choice of Likelihoods

As discussed in the previous chapter, integrated likelihood based bias correction is one of several available options. A legitimate question then is why this method is of particular choice in this study. Below, the disadvantages of other available approaches are discussed briefly, which reveals why the integrated likelihood approach is better suited to GARCH estimation.

Remember that the estimation approach of Chapter 2 is based on the variance targeting specification (Engle and Mezrich (1996)) of the GARCH model, where ω_i in (4.1) is replaced by $\lambda_i(1-\alpha-\beta)$, giving

$$\begin{aligned} \sigma_{it}^2 &= \lambda_i(1-\alpha-\beta) + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2, \\ \lambda_i &> 0 \ \forall i; \quad \alpha, \beta \geq 0 \quad \text{and} \quad \alpha + \beta < 1. \end{aligned}$$

This has the advantage that the computationally burdensome problem of joint estimation of an $N+2$ dimensional parameter $(\lambda_1, ..., \lambda_N, \alpha, \beta)$ is reduced to a simpler task: moment-based estimation of N univariate parameters in an initial step using $\mathbb{E}[y_{it}^2] = \lambda_i$, followed by a likelihood-based second-step estimation of a two-dimensional parameter θ. One can then

construct the (pseudo) composite likelihood function,

$$\ell_{NT}(\theta, \tilde{\lambda}_i) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \ell_{it}(\theta, \tilde{\lambda}_i; y_{it}|\mathcal{F}_{i,t-1}),$$

and obtain

$$\tilde{\theta}(\tilde{\lambda}) = \arg\max_{\theta \in \Theta} \ell_{NT}(\theta, \tilde{\lambda}_i), \tag{4.2}$$

where Θ is the parameter space for θ. Importantly, the objective function in (4.2) is not necessarily a proper likelihood function, and certainly not a concentrated likelihood function. This is because $\tilde{\lambda}_i$ is a method of moments estimator and it is not necessarily equal to the concentrated composite likelihood estimator given by

$$\hat{\lambda}_i(\theta) = \arg\max_{\lambda_i \in \Lambda_i} \frac{1}{T} \sum_{t=1}^{T} \ell_{it}(\theta, \lambda_i; y_{it}|\mathcal{F}_{i,t-1}),$$

where, as before, Λ_i is the parameter space for λ_i.

In contrast, the analytical bias reduction literature surveyed in Chapter 3 is almost exclusively based on analysing the first-order bias of the concentrated likelihood estimator. The practical implication of this is that the analytical bias expressions are based on the concentrated likelihood function which requires estimation of λ_0 by concentrated likelihood. Hence, the statistical objective is to estimate an $N+2$ dimensional parameter. When it comes to a highly nonlinear and dynamic model such as GARCH, panel estimation of a large dimensional parameter may well suffer from computational issues, such as failure of the optimiser to converge. Therefore, despite being theoretically promising, bias reduction might practically be difficult to achieve using concentrated likelihood based reduction methods.

Another interesting possibility considered in the literature is to use a split-panel jackknife procedure, as suggested by, e.g., Dhaene and Jochmans (2010). To illustrate this approach, suppose for simplicity that T is even. Consider two partitions of the panel, given by $\mathcal{P}_1 = \{1, ..., T/2\}$ and $\mathcal{P}_2 = \{(T/2) + 1, ..., T\}$. Define $\hat{\theta}_{\mathcal{P}_1}$ and $\hat{\theta}_{\mathcal{P}_2}$ as the concentrated likelihood estimators obtained from the two subsamples and $\hat{\theta}_{\mathcal{P}}$ as the concentrated likelihood estimator based on the full sample. Under cross-section independence and other regularity assumptions, Dhaene and Jochmans (2010) show that the split-panel jackknife

bias-corrected estimator

$$2\hat{\theta}_{\mathcal{P}} - \frac{1}{2}(\hat{\theta}_{\mathcal{P}_1} - \hat{\theta}_{\mathcal{P}_2})$$

will be first-order unbiased for θ_0. It would be promising to adapt the same approach to the two-step maximum-likelihood-like estimator suggested in Chapter 2, given that the two-step estimators are estimated very quickly. However, one important requirement of the split-panel jackknife estimator is that the estimator exists in all sub-samples used in estimation. For the case illustrated above, this means that the estimator should exist both in the full sample, as well as the two sub-samples $\mathcal{P}_1$ and $\mathcal{P}_2$. Remember that the objective of this chapter is to estimate GARCH parameters with around 200 time-series observations. However, as discussed in Chapter 2, when the sample consists of as little as 100 observations, frequently $\hat{\alpha} \approx 0$ implying that β is not identified, and consequently $\hat{\theta}$ does not exist. As such, the split-panel jackknife method is not feasible for the case of interest either.

The integrated likelihood function analysed in the previous chapter, on the other hand, does not suffer from any of these issues. Remember that the integrated function is given by

$$\ell_i^I(\theta) = \frac{1}{T} \ln \int_{\lambda_i \in \Lambda_i} \exp\left[T\ell_{iT}\left(\theta, \lambda_i\right)\right] \pi_i(\lambda_i|\theta) d\lambda_i. \tag{4.3}$$

Indeed, calculation of the integral still involves λ_i. However, crucially, the task is not to estimate λ_i but simply to integrate the likelihood function with respect to λ_i. This is a computationally intensive yet straightforward task. Hence, the statistical objective is estimation of a two dimensional parameter which is much less likely to suffer from software-related optimisation issues. In addition, unlike the split-panel jackknife method, there is no requirement for the estimator to exist in any sample size considered. Indeed, as the simulation results will reveal, bias reduction by integrated likelihood indirectly solves the identification problem for β by improving the small sample performance of $\hat{\alpha}$. For these reasons, the integrated likelihood based bias reduction approach is clearly best suited to the GARCH panel application.

4.2.2 Calculation of Robust Priors

In the remainder of this chapter, the notation of Chapter 3 is used. Remember that in Theorem 3.4.1 it has been shown that under time-series dependence the incidental parameter

bias of the integrated likelihood function is given by

$$\mathbb{E}_{\theta_0,\lambda_{i0}}\left[\ell_{iT}^{I}(\theta)-\bar{\ell}_{iT}(\theta)\right]=C+\frac{\mathcal{B}_{iT}^{(1)}(\theta)}{T}+\frac{\mathcal{B}_{iT}^{(2)}(\theta)}{T^{3/2}}+O\left(\frac{1}{T^2}\right),$$

where

$$\begin{aligned}\mathcal{B}_{iT}^{(1)}(\theta) &= \frac{1}{2}\{\mathbb{E}_{\theta_0,\lambda_{i0}}[-\ell_{iT}^{\lambda\lambda}]\}^{-1}\mathbb{E}_{\theta_0,\lambda_{i0}}[T(\ell_{iT}^{\lambda})^2]\\ &\quad -\frac{1}{2}\ln\mathbb{E}_{\theta_0,\lambda_{i0}}[-\ell_{iT}^{\lambda\lambda}]+\ln\pi_i(\bar{\lambda}_i|\theta),\\ \mathcal{B}_{iT}^{(2)}(\theta) &= T^{3/2}\frac{1}{2}\frac{\mathbb{E}_{\theta_0,\lambda_{i0}}\left[V_{iT}^{\lambda\lambda}(\ell_{iT}^{\lambda})^2\right]}{\{\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda}]\}^2}-T^{3/2}\frac{1}{6}\frac{\mathbb{E}_{\theta_0,\lambda_{i0}}[(\ell_{iT}^{\lambda})^3]\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda\lambda}]}{\{\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda}]\}^3}. \qquad (4.4)\end{aligned}$$

Accordingly, the two available specifications of the robust prior are given by

$$\pi_i^R(\lambda_i|\theta) \propto \widehat{\mathbb{E}}[-\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)]\left(\widehat{\mathbb{E}}\{[\ell_{iT}^{\lambda}(\theta,\lambda_i)]^2\}\right)^{-1/2}, \qquad \text{(P1)}$$

$$\pi_i^R(\lambda_i|\theta) \propto \{\widehat{\mathbb{E}}[-\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)]\}^{1/2}\exp\left(-\frac{T}{2}\{\widehat{\mathbb{E}}[-\ell_{iT}^{\lambda\lambda}(\theta,\lambda_i)]\}^{-1}\widehat{\mathbb{E}}\{[\ell_{iT}^{\lambda}(\theta,\lambda_i)]^2\}\right). \qquad \text{(P2)}$$

As explained before, (P1) and (P2) are equivalent except that (P1) is based on the information equality. Therefore, when parametric assumptions are not correct, (P1) can lead to misleading results depending on the magnitude of deviation from the information equality. For this reason, although the performance of (P1) is investigated to some extent in the simulation analysis, the empirical applications are based exclusively on (P2).

A convenient feature of the robust priors is that they are based on the population expectations. Therefore, the expectations can be estimated by using the simple method of moments estimator. Notice that, when evaluated at (θ_0,λ_{i0}), the GARCH score, $\ell_{it}^{\lambda}(\theta,\lambda_i)$, is a Martingale Difference Sequence and therefore,

$$\mathbb{E}\left\{\left[\sum_{t=1}^{T}\ell_{it}^{\lambda}(\theta_0,\lambda_{i0})\right]^2\right\}=\sum_{t=1}^{T}\mathbb{E}\{[\ell_{it}^{\lambda}(\theta_0,\lambda_{i0})]^2\}.$$

However, when evaluated at other parameter values, this property will not necessarily hold and $\ell_{it}^{\lambda}(\theta,\lambda_i)$ will be correlated across t. Therefore, a HAC estimator has to be used. Following Arellano and Hahn (2006), the estimators of the population moments are then

given by,

$$\begin{aligned}
\widehat{\mathbb{E}}[\ell_i^{\lambda\lambda}(\theta,\lambda_i)] &= \frac{1}{T}\sum_{t=1}^{T}\ell_{it}^{\lambda\lambda}(\theta,\lambda_i), \\
\widehat{\mathbb{E}}\{[\ell_i^{\lambda}(\theta,\lambda_i)]^2\} &= 2\sum_{l=0}^{m} w_{l,T}\Omega_l(\theta,\lambda_i), \\
\Omega_l(\theta,\lambda_i) &= \frac{1}{T}\sum_{t=\max(1,l+1)}^{\min(T,T+l)}\left[\ell_{it}^{\lambda}(\theta,\lambda_i)\times\ell_{i,t-l}^{\lambda}(\theta,\lambda_i)\right].
\end{aligned}$$

Here m is the bandwidth parameter and $w_{l,T}$ is the weight associated with l^{th} order autocovariance. For illustration purposes, these are chosen as $m = T^{1/3}$ and $w_{l,T} = 1-l(1+T^{1/3})^{-1}$, where the latter is also known as the Bartlett weight. Lastly, as derivatives of the log-likelihood are not available in closed form for the GARCH(1,1) model, these are calculated by using numerical differentiation methods.

4.2.3 Calculation of the Integrated Likelihood

Numerical evaluation of the integrated likelihood follows the simple quadrature method. Consider,

$$\ell_{iT}^{I}(\theta) = \frac{1}{T}\ln\int_{\underline{b}}^{\bar{b}}\exp\left[T\ell_{iT}(\theta,\lambda_i)\right]\pi_i(\lambda_i|\theta)d\lambda_i.$$

It is important that the lower and upper boundaries of this integral, $\underline{b}$ and $\bar{b}$ respectively, contain the true value of λ_i. Otherwise, the integrated likelihood will focus on an irrelevant portion of the parameter space and yield misleading results. Preferably, if no prior knowledge about the likely location of λ_i is available, the boundaries should contain the parameter space for λ_i. To achieve this, in empirical applications $\underline{b}$ and $\bar{b}$ are set to $.8\times(\min_{i,t} r_{it}^2)$ and $2\times\left(\max_{i,t} r_{it}^2\right)$, respectively, as r_{it}^2 provides a proxy for variance (using de-meaned observations). Clearly, these boundaries are chosen randomly and different choices can be used as long as the interval contains the true parameter values.

The integral is calculated by evaluating $\exp\left[T\ell_{iT}(\theta,\lambda_i)\right]\pi_i(\lambda_i|\theta)$ at 15 equidistant values of λ_i on a grid between $\underline{b}$ and $\bar{b}$. The reason for choosing 15 values for this purpose is to keep the computation time at a reasonable length. The integral can be calculated using a larger number of draws within the interval, which would give a more precise approximation. However, given that an individual numerical integration for each individual time-series is

required, the computational burden will quickly increase if a finer grid is used. Clearly, if the researcher has prior knowledge about the possible values of λ_i, it is much more desirable to concentrate the limited computation power on a tight interval and, consequently, a finer grid. This is another advantage of the integrated likelihood method as it allows for flexibility and discretion.

Finally, the estimation algorithm is as follows.

Step 1 Pick J (the number of points on the grid between $\underline{b}$ and $\bar{b}$) and B (the convergence criterion for the iterated estimator).

Step 2 Using some consistent estimator, such as the composite likelihood estimator of Chapter 2, obtain $\hat{\alpha}$ and $\hat{\beta}$. Define $\hat{\theta}^{(1)} = (\hat{\alpha}, \hat{\beta})$.

Step 3 Define each value of λ_i on the grid that is used to evaluate the integral as $\lambda_i^{(j)}$, $j = 1, ..., J$. Calculate the set of priors, $\pi_i(\lambda_i^{(j)}|\hat{\theta}^{(1)})$.

Step 4 Define $\Delta_\lambda = \lambda_i^{(j)} - \lambda_i^{(j-1)}$. Calculate the numerical approximation to the integrated likelihood function,

$$\tilde{\ell}_{iT}^{I}(\theta, \hat{\theta}^{(1)}) = \frac{1}{T} \ln \sum_{j=1}^{J} \exp[T\ell_{iT}(\theta, \lambda_i^{(j)})]\pi_i(\lambda_i^{(j)}|\hat{\theta}^{(1)})\Delta_\lambda.$$

Step 5 Obtain

$$\hat{\theta}_{IL} = \arg\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \tilde{\ell}_{iT}^{I}(\theta, \hat{\theta}^{(1)}).$$

Step 6 Calculate the set of priors, $\pi_i(\lambda_i^{(j)}|\hat{\theta}_{IL})$. Obtain

$$\hat{\theta}'_{IL} = \arg\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \tilde{\ell}_{iT}^{I}(\theta, \hat{\theta}_{IL}).$$

Step 7 Stop if $\left|\hat{\theta}'_{IL} - \hat{\theta}_{IL}\right| \leq B$ and report $\hat{\theta}_{IL}$ as the integrated likelihood estimator. Otherwise, substitute $\hat{\theta}'_{IL}$ for $\hat{\theta}_{IL}$ and return to Step 6.

Therefore $\ell_{iT}(\theta, \lambda_i)$ and $\pi_i(\lambda_i|\theta)$ are calculated at different values of θ. This is referred to as iterated updating. Doing otherwise would mean that the prior function changes with θ rather than remain fixed over θ, which is known as continuous updating. Here, B determines when the iteration stops. Alternatively, one can also limit the maximum

number of iterations and continue with iteration until convergence or the maximum number of iterations is reached, whichever happens first. This is indeed the approach taken in the rest of this chapter.

4.2.4 Higher Order Bias Terms

As discussed in Chapter 3, the analytical bias correction literature is concerned with the $O(T^{-1})$ bias while other higher-order terms are considered negligible. For this reason, the $O(T^{-3/2})$ bias given by (4.4) will not be dealt with here. Nevertheless, even if one wished to remove the higher-order bias, there is an important complication. Notice that (4.4) includes the non-standard expressions, $\mathbb{E}[(\ell_{iT}^{\lambda})^3]$, $\mathbb{E}\left[V_{iT}^{\lambda\lambda}(\ell_{iT}^{\lambda})^2\right]$ and $\mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}]$. To explain the problem, consider $\mathbb{E}\left[(\ell_{iT}^{\lambda})^3\right]$ evaluated at $(\lambda_{10}, ..., \lambda_{N0}, \alpha_0, \beta_0)$;

$$\begin{aligned}
\mathbb{E}\left\{[\ell_{iT}^{\lambda}(\theta_0, \lambda_{i0})]^3\right\} &= \frac{1}{T^3}\mathbb{E}\left[\sum_{s=1}^{T}\sum_{t=1}^{T}\sum_{q=1}^{T}\ell_{is}^{\lambda}(\theta_0, \lambda_{i0})\ell_{it}^{\lambda}(\theta_0, \lambda_{i0})\ell_{iq}^{\lambda}(\theta_0, \lambda_{i0})\right] \\
&= \frac{1}{T^3}\sum_{s\neq t\neq q}^{T}\mathbb{E}[\ell_{is}^{\lambda}(\theta_0, \lambda_{i0})\ell_{it}^{\lambda}(\theta_0, \lambda_{i0})\ell_{iq}^{\lambda}(\theta_0, \lambda_{i0})] \\
&\quad + \frac{1}{T^3}\sum_{s\neq t}^{T}\mathbb{E}[(\ell_{is}^{\lambda}(\theta_0, \lambda_{i0}))^2\ell_{it}^{\lambda}(\theta_0, \lambda_{i0})] \\
&\quad + \frac{1}{T^3}\sum_{s=1}^{T}\mathbb{E}[(\ell_{is}^{\lambda}(\theta_0, \lambda_{i0}))^3].
\end{aligned}$$

Notice that if $s \neq t \neq q$, then $\mathbb{E}[\ell_{is}^{\lambda}(\theta_0, \lambda_{i0})\ell_{it}^{\lambda}(\theta_0, \lambda_{i0})\ell_{iq}^{\lambda}(\theta_0, \lambda_{i0})] = 0$ since, for example,

$$\begin{aligned}
\mathbb{E}[\ell_{it_1}^{\lambda}(\theta_0, \lambda_{i0})\ell_{it_2}^{\lambda}(\theta_0, \lambda_{i0})\ell_{it_3}^{\lambda}(\theta_0, \lambda_{i0})|\mathcal{F}_{i,t_1-1}] &= \ell_{it_2}^{\lambda}(\theta_0, \lambda_{i0})\ell_{it_3}^{\lambda}(\theta_0, \lambda_{i0}) \\
&\quad \times\mathbb{E}[\ell_{it_1}^{\lambda}(\theta_0, \lambda_{i0})|\mathcal{F}_{i,t_1-1}] \\
&= 0, \\
\text{and} \quad \mathbb{E}\left\{\mathbb{E}[\ell_{it_1}^{\lambda}(\theta_0, \lambda_{i0})\ell_{it_2}^{\lambda}(\theta_0, \lambda_{i0})\ell_{it_3}^{\lambda}(\theta_0, \lambda_{i0})|\mathcal{F}_{i,t_1-1}]\right\} &= 0,
\end{aligned}$$

due to the MDS property of the score of the GARCH process and the Law of Iterated Expectations, where without loss of generality it is assumed that $t_1 > t_2 > t_3$. However, as discussed previously a HAC-type estimator has to be used when this expression is evaluated at other parameter values. The problem is that, this expression consists of three, rather than

the usual two, terms. Hence, the standard Newey-West (1987) type approach cannot be used and instead, a different method that accounts for the complex dependence structure between $\ell_{is}^{\lambda}(\theta_0, \lambda_{i0})$, $\ell_{it}^{\lambda}(\theta_0, \lambda_{i0})$ and $\ell_{iq}^{\lambda}(\theta_0, \lambda_{i0})$ is needed. In addition, (4.4) also contains $\mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}]$, a term composed of third-order derivatives of the GARCH pseudo-likelihood. This is another problematic term as derivatives of the GARCH likelihood have to be calculated numerically. Calculation of the first and second order derivatives is standard. However, with higher order derivatives the numerical derivation algorithm becomes complicated. Straightforward algorithms based on differentiating the second order derivative can be used, but at the possible cost of numerical inaccuracies.

Perhaps most importantly, as outlined in Section 4.2.2, correcting for the $O(T^{-1})$ bias term involves estimation of population moments. By definition, this estimation will introduce some $O(T^{-3/2})$ bias in any case. Therefore, even if the $O(T^{-3/2})$ bias in (4.4) can successfully be removed, there will nevertheless remain some $O(T^{-3/2})$ bias due to first-order bias correction. Therefore, based on these observations, correction of the $O(T^{-3/2})$ bias term is not considered here.

4.3 Simulation Analysis

4.3.1 Simulation Setup

In this section, small sample performance of the integrated likelihood method is analysed. The baseline estimation method is the composite likelihood (CL) method introduced in Chapter 2. The infeasible composite likelihood (InCL) method, where true values of λ_i are used in estimation, is used as the theoretical benchmark. The integrated likelihood methods using prior (P1) and (P2) are designated as the integrated composite likelihood (ICL) and integrated pseudo composite likelihood (IPCL) methods, respectively, as the first prior is based on the information inequality whereas the second one is not.[1]

In light of the simulation results from Chapter 2 which suggest that the incidental parameter problem is more acute when T is around or less than 250, the simulations in this section focus on $T \in \{75, 100, 150, 200, 400\}$ and $N \in \{25, 50, 100\}$. Otherwise, the

[1] Note that in empirical applications parameter assumptions do not necessarily hold and so, all estimation approaches will essentially yield "pseudo" likelihoods. However, this is not reflected in the notation, in order not to make it more cumbersome.

simulation setting is the same as in Chapter 2. To briefly summarise, data are generated by using

$$y_{it} = \mu_{it} + \varepsilon_{it}, \quad \mu_{it} = E[y_{it}|\mathcal{F}_{i,t-1}] = 0, \quad \varepsilon_{it} = \sigma_{it}\eta_{it},$$
$$\sigma^2_{it} = \lambda_i(1 - \alpha_0 - \beta_0) + \alpha_0\varepsilon^2_{i,t-1} + \beta_0\sigma^2_{i,t-1}, \quad \sigma^2_{i0} = \lambda_{i0},$$
$$\alpha_0 = 0.05 \quad \text{and} \quad \beta_0 = 0.93.$$

As before, η_{it} is assumed to be Normally distributed, which yields consistent estimators by the quasi maximum likelihood argument of Bollerslev and Wooldridge (1992). The nuisance parameters are drawn from a uniform distribution such that the corresponding annual volatility is between 15% and 80%. Cross-section dependence for η_{it} is generated using the single-factor model

$$\eta_{it} = \rho_i u_t + \sqrt{1 - \rho_i^2}\tau_{it}, \quad u_t \overset{iid}{\sim} \mathcal{N}(0,1) \quad \text{and} \quad \tau_{it} \overset{iid}{\sim} \mathcal{N}(0,1),$$

which implies contemporaneous correlation of $\rho_i\rho_j$ between any two assets. The correlation parameters will be generated in three different ways, to reflect various levels of dependence. These are

i. independence: $\rho_i = 0$,

ii. mild dependence: $\rho_i \sim U(0.5, 0.9)$ where U is the Uniform distribution,

iii. strong dependence: $\rho_i = 0.9$.

The starting values for optimisation are drawn randomly, in order to prevent possible bias due to starting value selection, using $\alpha+\beta \sim U(0.5, 0.99)$ and $\alpha/(\alpha+\beta) \sim U(0.01, 0.3)$.

The integrated likelihood is calculated using the algorithm detailed in Section 4.2.3. Two different estimation strategies will be considered. First, only the intercept parameter λ_i will be integrated out. In a second and more detailed analysis, the structure of the integrated likelihood will be exploited to integrate out the initial value, as well. This latter case will be explained in more detail below. In both cases, iteration of the estimator stops either at the tenth iteration or when the estimator converges, whichever happens first. In this illustration, convergence of iterated estimators is considered to be achieved when the difference between

two successive iterations is less than 0.003 for $\hat{\alpha}$ and 0.01 for $\hat{\beta}$. The particular choice of $(0.003, 0.01)'$ as the cut-off point is for illustration purposes and is not determined by a specific criterion.[2] Finally, all simulation results are based on 500 replications.

4.3.2 Initial Value Selection

Remember that the Gaussian log-likelihood function for $(y_{i1}, ..., y_{iT})$ is proportional to

$$-\frac{1}{2}\sum_{t=1}^{T}\log\sigma_{it}^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{\varepsilon_{it}^2}{\sigma_{it}^2}.$$

Assuming that the sample actually starts at $t = 0$, one needs to know the value of $\sigma_{i,0}^2$ in order to calculate the likelihood because

$$\sigma_{i1}^2 = \omega_i + \alpha\varepsilon_{i,0}^2 + \beta\sigma_{i,0}^2.$$

So, once the initial value is selected, conditional variances for the rest of the sample can be generated recursively. However, as volatility is latent, $\sigma_{i,0}^2$ can never be observed. Therefore, it has to be replaced by a proxy. One common approach is to use an estimator of the long-run or unconditional variance, which is given by $\sigma_i^2 = \mathbb{E}[\sigma_{it}^2]$. Due to the variance targeting specification used in (4.1), this can easily be estimated by using $T^{-1}\sum_{t=1}^{T} y_{it}^2$. Noting that the aim is to find a proxy for the "initial" value, another possibility is to focus on an initial period of the sample when estimating σ_i^2, as this will be more relevant to the initial value. Therefore, a more relevant estimator is

$$\frac{1}{\lceil T^{1/2}\rceil}\sum_{t=1}^{\lceil T^{1/2}\rceil} y_{it}^2, \tag{4.5}$$

where $\lceil T^{1/2}\rceil$ is obtained by rounding $T^{1/2}$ up to the nearest integer (see, e.g., Shephard and Sheppard (2010).). This initial value is used for the first estimation strategy.

Generally, the particular choice of the initial value is of no consequence, as the effect of the initial value on GARCH estimation is asymptotically negligible.[3] However, when the

[2] In simulations, the maximum number of iterations across all panel dimensions was six and in most cases two to four iterations were sufficient for convergence.

[3] See, for example, Francq and Zakoïan (2004) who show this in their proof of GARCH quasi maximum likelihood estimation.

sample size is small, such as those considered here, the effect is likely to be non-negligible.

In the case that the initial value is based on the long-run variance, the integrated likelihood function offers an interesting possibility to incorporate the estimation of the initial value into the analysis. Remember that $\lambda_{i0} = \mathbb{E}[\sigma_{it}^2]$. Then one can simply construct the integrated likelihood in such a way that the initial value is also integrated out, along with the intercept parameter, as both are estimating the same value. In terms of computational burden, this operation brings no extra cost, other than a simple modification of the computer code. As simulation and empirical results below will illustrate, this approach leads to marked improvement both in- and out-of-sample. Of course, the analysis of such an approach will remain incomplete without a rigorous theoretical investigation. Given the complex structure of the GARCH likelihood, such a theoretical analysis would require a separate study. However, this exercise has no claim to be a thorough treatment of the initial value selection issue. Instead, the sole aim is to take the first step towards an interesting and, possibly, promising direction.

4.3.3 Simulation Results for Estimation Strategy I

In this part, the mild dependence case only is considered for space considerations, as a detailed analysis of different dependence structures will be provided in the next section. Estimators are obtained by using the initial value selection method given in (4.5). Simulation results are presented in Tables 4.1-4.2 and Figures 4.1, 4.2 and 4.3.

Result in Table 4.1 suggest that using the integrated likelihood and the robust priors leads to substantial reductions in the bias of the CL estimator. In some cases, the reduction in bias is enormous: for example, for $T = 100$ and $N = 100$, ICL reduces 55% of the bias in $\hat{\alpha}$ due to CL, while the bias in $\hat{\beta}$ is reduced by 61%, in absolute value. Similarly, when $T = 100$ and $N = 25$, 47% of the bias in $\hat{\alpha}$ and 91% of the bias in $\hat{\beta}$ is removed by ICL. Simulation results also confirm that, in the simulation setting considered, bias is indeed related to T and not N. There is a clear downward pattern in the bias as T increases. However, no such tendency is observed in relation to N. As expected, as T increases, all methods tend to perform similar to InCL. This is intuitive for ICL and IPCL. For large T, the first order bias will be very small anyway, so the choice of the prior will have no effect.

It is also interesting to compare the bias performance in estimation of $\alpha + \beta$, which

	$\alpha = 0.05, \quad \beta = 0.93, \quad \alpha + \beta = 0.98$											
	CL			InCL			ICL			IPCL		
T	α	β	$a+\beta$	α	β	$a+\beta$	α	β	$a+\beta$	α	β	$a+\beta$
	N=100											
400	.045	.924	.969	.048	.932	.980	.046	.935	.981	.046	.935	.981
200	.038	.913	.951	.046	.932	.978	.042	.941	.983	.042	.940	.982
150	.034	.901	.935	.048	.927	.975	.041	.941	.982	.040	.940	.980
100	.017	.886	.902	.046	.925	.972	.035	.947	.981	.031	.947	.978
75	.009	.850	.860	.048	.920	.967	.032	.948	.980	.026	.950	.976
	N=50											
400	.046	.924	.969	.048	.932	.979	.046	.935	.981	.046	.935	.981
200	.039	.912	.950	.047	.930	.977	.042	.939	.982	.042	.939	.981
150	.033	.900	.933	.047	.929	.976	.040	.943	.982	.039	.942	.981
100	.019	.876	.895	.048	.923	.971	.036	.943	.978	.032	.943	.975
75	.008	.854	.863	.046	.920	.967	.029	.942	.971	.024	.943	.967
	N=25											
400	.045	.923	.969	.048	.932	.979	.046	.935	.981	.046	.935	.981
200	.039	.911	.950	.047	.931	.977	.042	.939	.981	.042	.939	.980
150	.034	.893	.927	.047	.927	.974	.040	.939	.979	.039	.938	.978
100	.020	.864	.884	.048	.923	.970	.034	.936	.970	.031	.936	.968
75	.012	.833	.844	.046	.921	.967	.030	.935	.965	.025	.936	.961

Table 4.1: Average parameter estimates for $\hat{\alpha}$ and $\hat{\beta}$ by Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL) and Integrated Pseudo CL (IPCL). Based on 500 replications of GARCH panels (exhibiting mild cross-section dependence) for varying T and N where true parameter values are given by $\alpha = 0.05$ and $\beta = 0.93$.

	CL		InCL		ICL		IPCL		CL		InCL		ICL		IPCL	
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$
	N=100															
400	.007	.011	.006	.010	.007	.011	.007	.011	.008	.013	.007	.010	.008	.012	.008	.012
200	.010	.018	.009	.013	.009	.016	.009	.016	.016	.025	.009	.013	.012	.020	.012	.019
150	.014	.026	.011	.017	.011	.022	.011	.022	.022	.039	.011	.017	.014	.024	.015	.024
100	.016	.061	.012	.019	.012	.034	.013	.035	.037	.075	.013	.019	.020	.038	.023	.039
75	.018	.113	.014	.022	.016	.026	.015	.027	.045	.138	.014	.025	.024	.032	.028	.034
	N=50															
400	.008	.012	.007	.011	.007	.011	.007	.011	.009	.014	.007	.011	.008	.012	.008	.012
200	.012	.021	.010	.014	.010	.018	.010	.017	.016	.028	.010	.014	.012	.020	.013	.020
150	.014	.036	.010	.015	.011	.021	.011	.021	.022	.047	.011	.015	.015	.024	.016	.024
100	.020	.084	.014	.021	.014	.035	.015	.035	.037	.100	.014	.023	.020	.037	.023	.038
75	.015	.090	.016	.025	.017	.046	.016	.047	.044	.118	.016	.027	.027	.048	.031	.049
	N=25															
400	.009	.014	.008	.012	.008	.014	.008	.014	.010	.016	.008	.012	.009	.014	.009	.014
200	.013	.023	.011	.016	.011	.020	.011	.020	.017	.029	.011	.016	.014	.022	.014	.022
150	.018	.076	.012	.020	.013	.038	.013	.038	.024	.084	.012	.020	.016	.039	.017	.039
100	.021	.116	.016	.026	.018	.074	.018	.074	.036	.133	.016	.027	.024	.074	.026	.075
75	.021	.130	.018	.027	.021	.085	.021	.085	.044	.163	.018	.028	.029	.085	.032	.085

Table 4.2: Sample standard deviation (left panel) and root mean squared error (right panel) for $\hat{\alpha}$ and $\hat{\beta}$ by Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL) and Integrated Pseudo CL (IPCL). Based on 500 replications of GARCH panels (exhibiting mild cross-section dependence) for varying T and N where true parameter values are given by $\alpha = 0.05$ and $\beta = 0.93$.

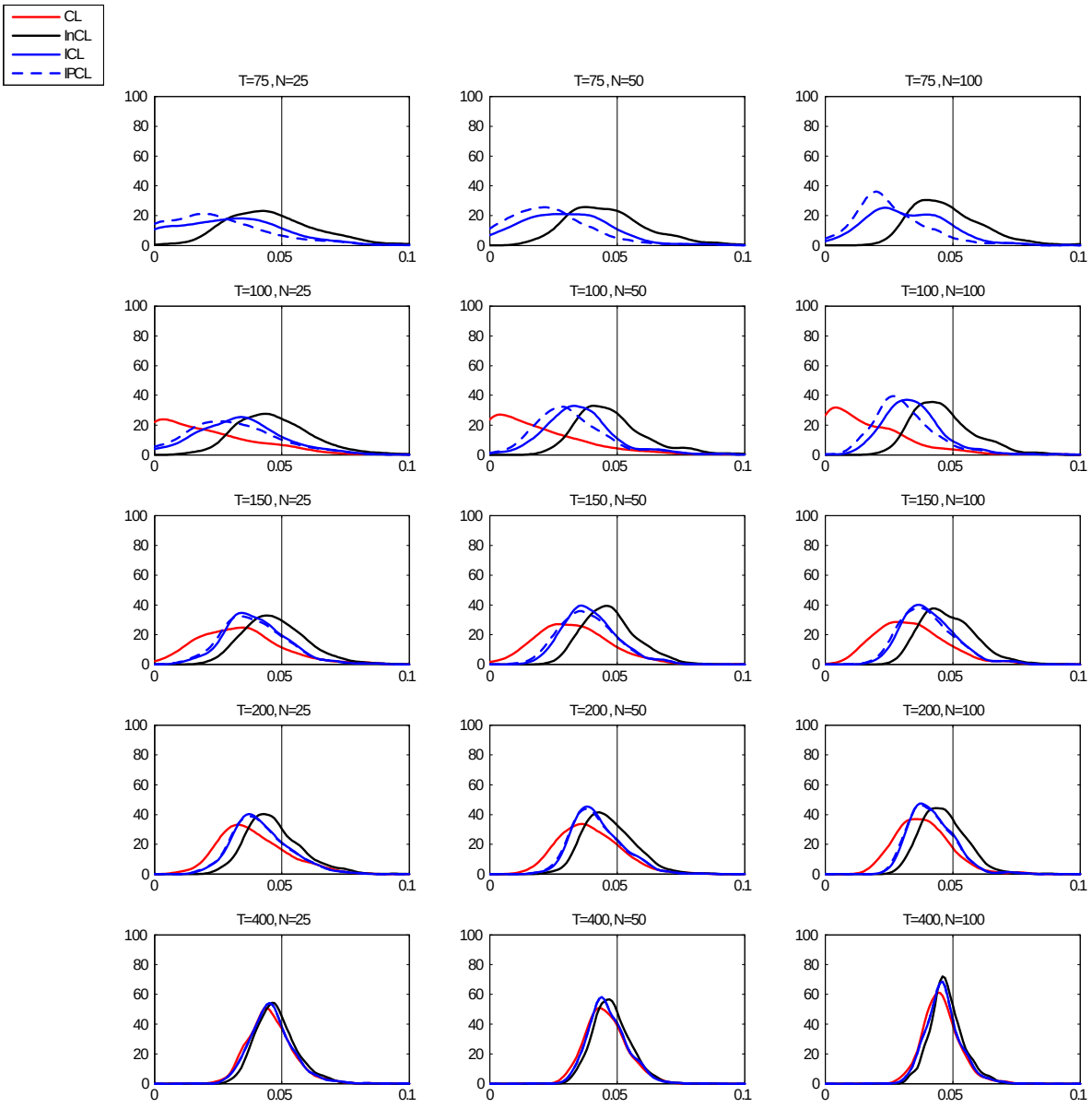

Figure 4.1: Sample distributions of $\hat{\alpha}$ using the composite likelihood (CL), infeasible CL (InCL), integrated CL (ICL), and integrated pseudo CL (IPCL) methods. The vertical line is drawn at the true parameter value. Based on 500 replications under mild cross-section dependence where $(\alpha, \beta) = (0.05, 0.93)$.

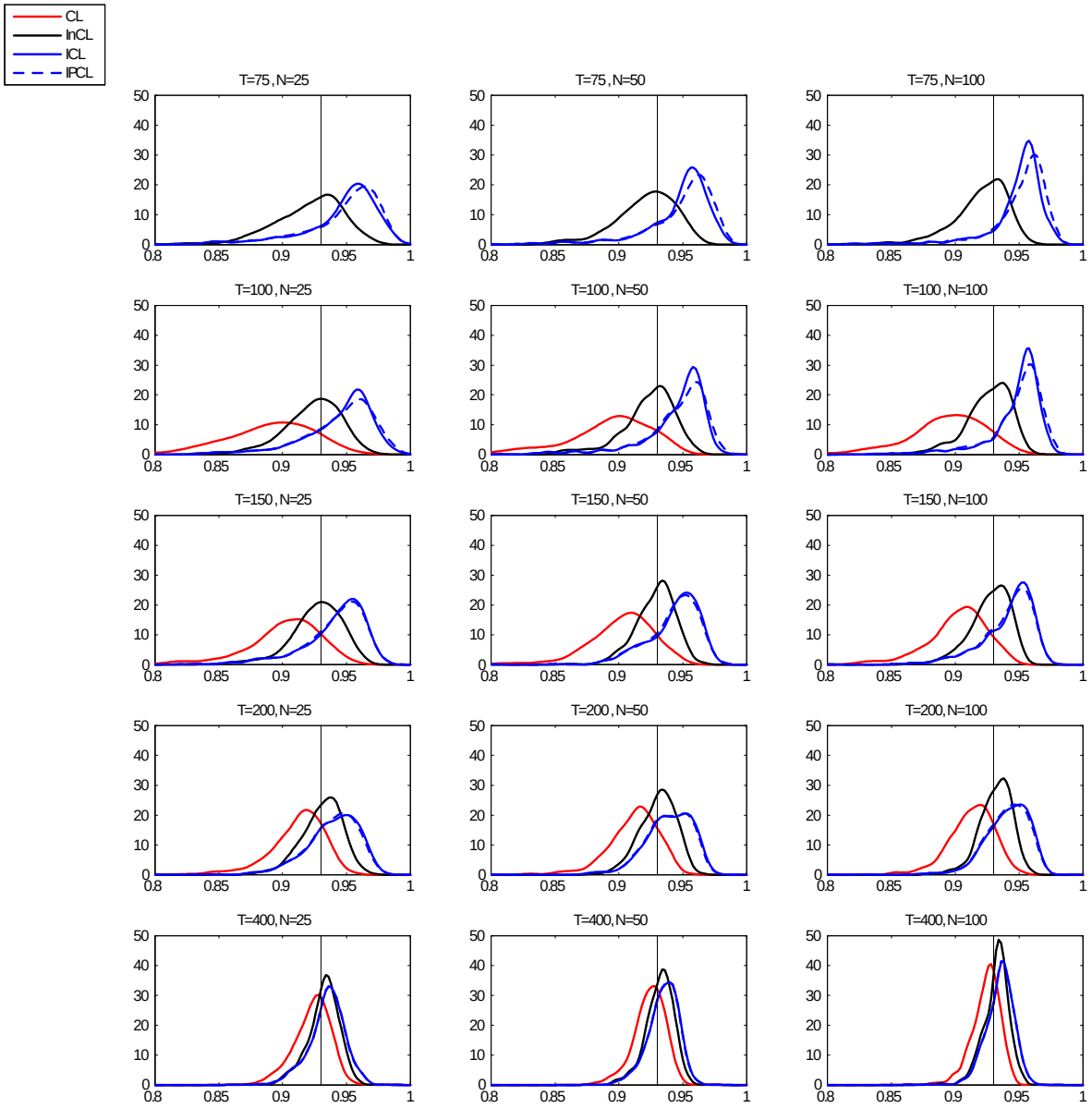

Figure 4.2: Sample distributions of $\hat{\beta}$ using the composite likelihood (CL), infeasible CL (InCL), integrated CL (ICL), and integrated pseudo CL (IPCL) methods. The vertical line is drawn at the true parameter value. Based on 500 replications under mild cross-section dependence where $(\alpha, \beta) = (0.05, 0.93)$.

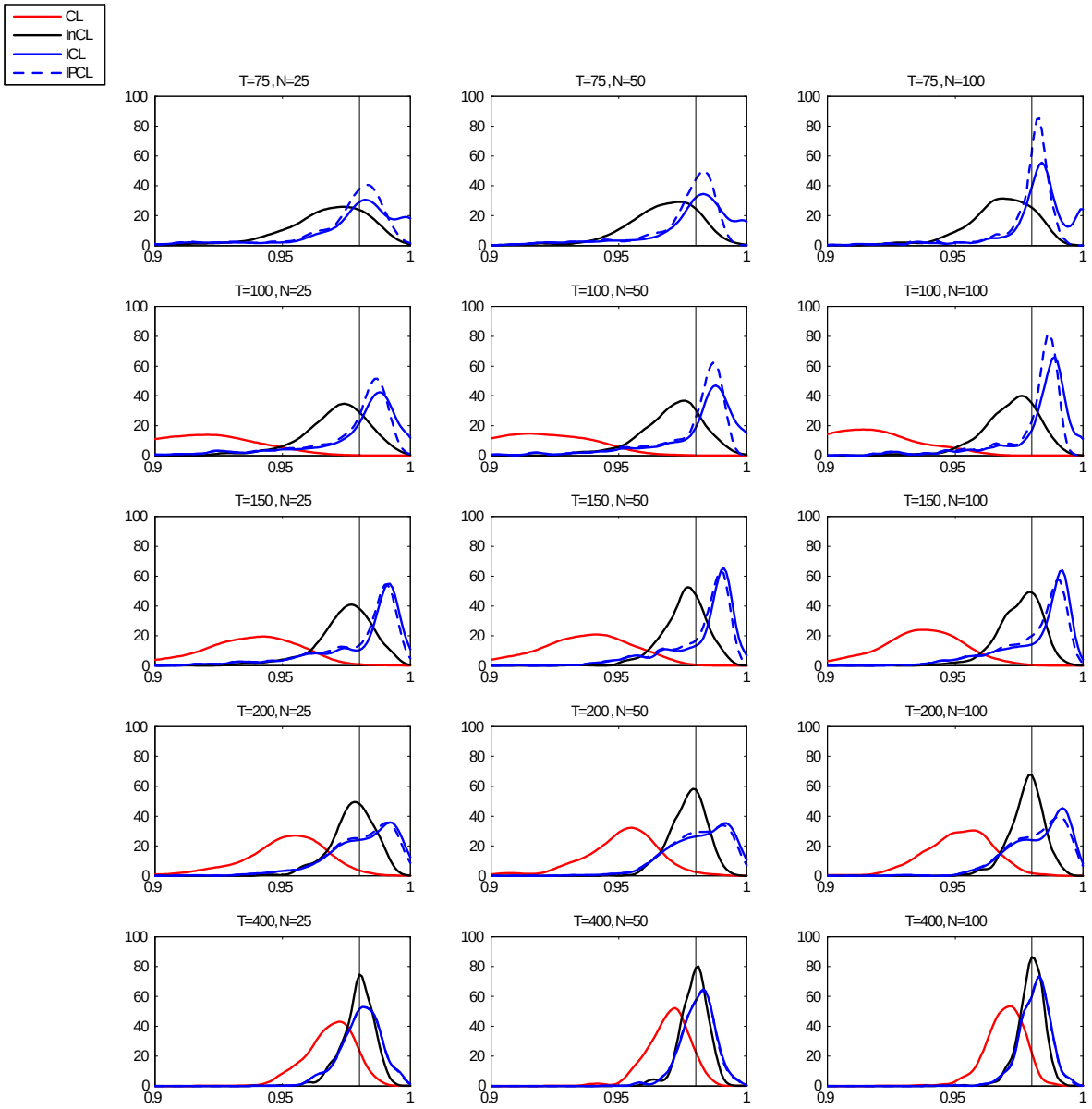

Figure 4.3: Sample distributions of $\hat{\alpha} + \hat{\beta}$ using the composite likelihood (CL), infeasible CL (InCL), integrated CL (ICL), and integrated pseudo CL (IPCL) methods. The vertical line is drawn at the true parameter value. Based on 500 replications under mild cross-section dependence where $(\alpha, \beta) = (0.05, 0.93)$.

gives the memory of the GARCH process. An intriguing observation is that the integrated likelihood tends to estimate this quantity much better, even when compared to the infeasible method. Especially for larger N, integrated likelihood estimator achieves accuracy even when T is as low as 75. CL, on the other hand, never manages to catch up, even when $T = 400$. Interestingly, performance of a similar calibre is not attained in estimating α and β separately. Therefore, one implication is that perhaps the integrated likelihood method's structure is such that essentially it estimates $\alpha + \beta$.[4] However, without further theoretical analysis, this remains a speculation.

Figures 4.1, 4.2 and 4.3 provide additional insights into the properties of the methods considered here. The locations of the modes of sample distributions imply that, independent of T and N, ICL and IPCL are more likely to underestimate α and overestimate β. These methods also overestimate $\alpha + \beta$, on average. It is also clear that the performances of the four methods in estimating α and β converge to each other as T increases. Estimation of $\alpha + \beta$ is a slightly different story where, in line with the previous discussion, CL is slow in converging to InCL. Finally, especially when $T \geq 150$, estimates by the two priors have almost overlapping sample distributions.

The left panel of Table 4.2 gives the sample standard deviations for $\hat{\alpha}$ and $\hat{\beta}$ ($\bar{\sigma}_{\hat{\alpha}}$ and $\bar{\sigma}_{\hat{\beta}}$). Clearly, bias-reduction by robust priors does not come at a cost of increased variance, in line with the observations in the rest of the literature. This is due to the fact that bias is of a smaller order than variance. Therefore, adjusting the bias term should not affect the variance. In fact, simulation results show that bias reduction instead leads to lower standard deviation in comparison to CL. Not surprisingly, for given T, larger N generally leads to lower standard deviation. The combination of superior bias and standard deviation performance of the robust priors is translated into superior root mean square error performance ($\mathcal{R}_{\hat{\alpha}}$ and $\mathcal{R}_{\hat{\beta}}$), as can be observed in the right panel of Table 4.2.

4.3.4 Simulation Results for Estimation Strategy II

In this part of the analysis, the bias-reduced estimators are obtained by integrating both the initial value and the individual specific parameter out. In a way, this can be considered as

[4] This is also a well-observed (but not much documented) general feature of the GARCH model. Usually, estimation of $\alpha + \beta$ is much more accurate compared to estimation of α and β individually.

a reduction of both the incidental parameter bias and the initial value selection/estimation bias. Note that this only affects the results for the bias-reduced estimators as the remaining estimators are still based on the standard initial value estimator in (4.5). Given the previous observation that ICL and IPCL deliver similar results for reasonable T, only IPCL is considered in the following analysis. In what follows, to avoid confusion, IPCL* denotes the bias reduction method based on integrating the initial values out.

To provide a more in-depth scrutiny of the effect of cross-section dependence, results are reported for the three dependence types listed in Section 4.3.1. Results are presented in Tables 4.3-4.8 and Figures 4.4-4.12. This yields a large collection of results, the sheer size of which seems very daunting at first sight. However, the main story told by these results boils down to a few important key observations, which are discussed next. In the first instance, IPCL* is compared to IPCL by focusing on the mild dependence case. Then, implications of different dependence settings will be analysed.

The first main result is that IPCL* achieves a remarkable improvement in estimating $\hat{\alpha}$. Table 4.5 reveals that on average α is estimated with little or no bias, across all sample sizes. At first sight, the story for $\hat{\beta}$ seems to be somewhat different: for $T \geq 150$, IPCL* achieves smaller bias in absolute value, while for smaller T it suffers from higher bias compared to IPCL. Moreover, estimation of $\alpha + \beta$ is not as good as before. Still, except in a few cases, IPCL* still removes a substantial portion of the small sample bias.

On the basis of this information alone, one might conclude that it is ambiguous whether IPCL* is the better choice or not. However, a better overview of the situation is obtained by analysing the sample distributions of the three estimators, given in Figures 4.7-4.9. The message delivered by these figures is striking: In almost all sample sizes and across all three parameters, $(\alpha, \beta, \alpha + \beta)$, the mode of the sample distribution is either at or very close to the true parameter values. Remember that this is not the case for IPCL, even when the sample size is reasonably large. In addition, in many cases IPCL* does a very good job in matching the sample distribution of the infeasible estimator. Clearly, the less satisfactory results for average $\hat{\beta}$ and $\hat{\alpha} + \hat{\beta}$ are due to the sample distributions being skewed. These observations then reveal that IPCL* estimates will *most of the time* be close to the true parameter values, while *on average* $\hat{\beta}$ and $\hat{\alpha} + \hat{\beta}$ will be downward biased.

The second interesting observation is related to sample standard deviations. Despite

	$\alpha = 0.05,$		$\beta = 0.93,$	$\alpha + \beta = 0.98,$		$\rho_i = 0$		
	CL			InCL			IPCL*		
T	α	β	$a + \beta$	α	β	$a + \beta$	α	β	$a + \beta$
	N=100								
400	.045	.926	.971	.047	.933	.981	.050	.930	.980
200	.039	.914	.954	.047	.932	.979	.050	.927	.976
150	.033	.906	.939	.048	.929	.977	.049	.924	.973
100	.018	.895	.913	.049	.924	.973	.047	.923	.970
75	.003	.892	.895	.048	.921	.970	.045	.919	.963
	N=50								
400	.045	.925	.971	.047	.933	.980	.050	.930	.980
200	.039	.914	.953	.047	.932	.979	.050	.927	.976
150	.034	.905	.938	.048	.929	.977	.049	.923	.972
100	.018	.895	.913	.048	.924	.972	.047	.920	.967
75	.005	.884	.889	.048	.921	.969	.045	.911	.957
	N=25								
400	.045	.926	.970	.047	.933	.980	.050	.930	.979
200	.038	.915	.953	.047	.932	.979	.049	.927	.976
150	.033	.904	.937	.048	.929	.976	.049	.922	.972
100	.019	.891	.911	.048	.924	.972	.048	.914	.963
75	.007	.874	.887	.049	.920	.969	.047	.896	.943

Table 4.3: Average parameter estimates for $\hat{\alpha}$ and $\hat{\beta}$ by Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL) and Integrated Pseudo CL (IPCL). Based on 500 replications of GARCH panels (exhibiting cross-section independence) for varying T and N where true parameter values are given by $\alpha = 0.05$ and $\beta = 0.93$.

	CL		InCL		IPCL*		CL		InCL		IPCL*	
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$
	N=100											
400	.002	.004	.002	.003	.002	.004	.005	.006	.004	.005	.002	.004
200	.004	.006	.003	.005	.003	.007	.012	.017	.004	.005	.003	.007
150	.005	.008	.004	.006	.004	.008	.018	.025	.004	.006	.004	.010
100	.007	.011	.005	.007	.005	.012	.033	.037	.005	.009	.006	.015
75	.005	.020	.005	.008	.007	.017	.047	.043	.006	.012	.008	.021
	N=50											
400	.003	.005	.003	.005	.003	.006	.006	.007	.004	.006	.003	.006
200	.005	.008	.004	.006	.005	.009	.012	.018	.005	.007	.005	.009
150	.006	.012	.005	.007	.005	.010	.018	.028	.005	.007	.005	.012
100	.009	.016	.006	.010	.007	.018	.033	.039	.007	.011	.008	.020
75	.007	.036	.008	.011	.009	.056	.046	.059	.008	.015	.010	.059
	N=25											
400	.005	.007	.005	.007	.005	.008	.007	.009	.005	.008	.005	.008
200	.008	.013	.006	.009	.007	.013	.014	.020	.007	.010	.007	.013
150	.009	.016	.007	.011	.008	.017	.019	.030	.007	.011	.008	.019
100	.012	.032	.009	.013	.011	.037	.033	.050	.009	.015	.011	.041
75	.010	.059	.011	.017	.013	.094	.044	.081	.011	.020	.013	.100

Table 4.4: Sample standard deviation (left panel) and root mean squared error (right panel) for $\hat{\alpha}$ and $\hat{\beta}$ by Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL) and Integrated Pseudo CL (IPCL). Based on 500 replications of GARCH panels (exhibiting cross-section independence) for varying T and N where true parameter values are given by $\alpha = 0.05$ and $\beta = 0.93$.

	$\alpha = 0.05,$		$\beta = 0.93,$		$\alpha + \beta = 0.98,$		$\rho_i \sim U(0.5, 0.9)$		
	CL			InCL			IPCL*		
T	α	β	$a + \beta$	α	β	$a + \beta$	α	β	$a + \beta$
	N=100								
400	.045	.924	.969	.048	.932	.980	.050	.929	.979
200	.038	.913	.951	.046	.932	.978	.049	.926	.976
150	.034	.901	.935	.048	.927	.975	.051	.919	.969
100	.017	.886	.902	.046	.925	.972	.049	.904	.953
75	.009	.850	.860	.048	.920	.967	.051	.868	.919
	N=50								
400	.046	.924	.969	.048	.932	.979	.050	.929	.979
200	.039	.912	.950	.047	.930	.977	.050	.924	.974
150	.033	.900	.933	.047	.929	.976	.050	.920	.969
100	.019	.876	.895	.048	.923	.971	.050	.896	.945
75	.008	.854	.863	.046	.920	.967	.050	.857	.906
	N=25								
400	.045	.923	.969	.048	.932	.979	.050	.928	.978
200	.039	.911	.950	.047	.931	.977	.050	.924	.976
150	.034	.893	.927	.047	.927	.974	.051	.915	.966
100	.020	.864	.884	.048	.923	.970	.050	.891	.941
75	.012	.833	.844	.046	.921	.967	.051	.846	.897

Table 4.5: Average parameter estimates for $\hat{\alpha}$ and $\hat{\beta}$ by Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL) and Integrated Pseudo CL (IPCL). Based on 500 replications of GARCH panels (exhibiting mild cross-section dependence) for varying T and N where true parameter values are given by $\alpha = 0.05$ and $\beta = 0.93$.

	CL		InCL		IPCL*		CL		InCL		IPCL*	
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$
	N=100											
400	.007	.011	.006	.010	.007	.011	.008	.013	.007	.010	.007	.011
200	.010	.018	.009	.013	.009	.017	.016	.025	.009	.013	.009	.017
150	.014	.026	.011	.017	.012	.042	.022	.039	.011	.017	.012	.043
100	.016	.061	.012	.019	.015	.089	.037	.075	.013	.019	.015	.093
75	.018	.113	.014	.022	.020	.162	.045	.138	.014	.025	.020	.173
	N=50											
400	.008	.012	.007	.011	.008	.012	.009	.014	.007	.011	.008	.012
200	.012	.021	.010	.014	.010	.021	.016	.028	.010	.014	.010	.022
150	.014	.036	.010	.015	.012	.039	.022	.047	.011	.015	.012	.040
100	.020	.084	.014	.021	.018	.120	.037	.100	.014	.023	.018	.125
75	.015	.090	.016	.025	.022	.175	.044	.118	.016	.027	.022	.189
	N=25											
400	.009	.014	.008	.012	.008	.014	.010	.016	.008	.012	.008	.015
200	.013	.023	.011	.016	.012	.025	.017	.029	.011	.016	.012	.026
150	.018	.076	.012	.020	.014	.055	.024	.084	.012	.020	.016	.057
100	.021	.116	.016	.026	.020	.115	.036	.133	.016	.027	.020	.121
75	.021	.130	.018	.027	.025	.202	.044	.163	.018	.028	.025	.219

Table 4.6: Sample standard deviation (left panel) and root mean squared error (right panel) for $\hat{\alpha}$ and $\hat{\beta}$ by Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL) and Integrated Pseudo CL (IPCL). Based on 500 replications of GARCH panels (exhibiting mild cross-section dependence) for varying T and N where true parameter values are given by $\alpha = 0.05$ and $\beta = 0.93$.

	$\alpha = 0.05, \quad \beta = 0.93, \quad \alpha + \beta = 0.98, \quad \rho_i = 0.9$								
	CL			InCL			IPCL*		
T	α	β	$a+\beta$	α	β	$a+\beta$	α	β	$a+\beta$
	N=100								
400	.047	.908	.955	.049	.925	.974	.052	.916	.968
200	.044	.857	.901	.050	.913	.964	.056	.880	.936
150	.038	.828	.865	.050	.907	.957	.056	.850	.906
100	.032	.778	.811	.052	.886	.938	.062	.767	.828
75	.025	.748	.773	.049	.884	.933	.058	.733	.791
	N=50								
400	.048	.914	.961	.049	.927	.976	.052	.923	.975
200	.043	.859	.902	.048	.917	.965	.054	.886	.940
150	.038	.829	.867	.050	.903	.953	.056	.840	.897
100	.028	.781	.810	.051	.884	.935	.059	.760	.819
75	.028	.753	.781	.051	.876	.927	.062	.723	.785
	N=25								
400	.046	.915	.961	.047	.930	.977	.051	.922	.973
200	.044	.862	.906	.050	.918	.967	.056	.881	.937
150	.040	.814	.854	.050	.902	.952	.058	.830	.888
100	.032	.767	.799	.050	.893	.943	.060	.751	.811
75	.029	.738	.767	.055	.865	.920	.063	.714	.776

Table 4.7: Average parameter estimates for $\hat{\alpha}$ and $\hat{\beta}$ by Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL) and Integrated Pseudo CL (IPCL). Based on 500 replications of GARCH panels (exhibiting strong cross-section dependence) for varying T and N where true parameter values are given by $\alpha = 0.05$ and $\beta = 0.93$.

	CL		InCL		IPCL*		CL		InCL		IPCL*	
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$
	N=100											
400	.018	.071	.017	.036	.017	.066	.018	.074	.017	.036	.017	.067
200	.030	.156	.028	.084	.027	.148	.031	.173	.028	.085	.028	.156
150	.037	.184	.035	.105	.034	.185	.039	.211	.035	.108	.034	.201
100	.046	.208	.055	.150	.045	.253	.050	.257	.055	.156	.046	.301
75	.047	.207	.052	.142	.049	.271	.054	.276	.052	.149	.050	.335
	N=50											
400	.018	.057	.017	.038	.017	.043	.018	.059	.017	.038	.017	.044
200	.030	.148	.029	.087	.026	.139	.030	.164	.029	.088	.027	.146
150	.037	.183	.040	.115	.035	.198	.039	.209	.040	.118	.035	.217
100	.042	.200	.051	.150	.042	.264	.047	.249	.051	.157	.043	.314
75	.051	.209	.056	.159	.052	.282	.056	.274	.056	.168	.053	.350
	N=25											
400	.017	.061	.016	.024	.017	.058	.017	.062	.016	.024	.017	.058
200	.031	.151	.029	.073	.027	.149	.031	.166	.029	.074	.028	.157
150	.039	.200	.034	.119	.035	.211	.040	.231	.034	.123	.036	.233
100	.045	.215	.048	.125	.044	.263	.048	.270	.048	.130	.045	.319
75	.053	.206	.060	.182	.056	.277	.057	.282	.061	.194	.058	.351

Table 4.8: Sample standard deviation (left panel) and root mean squared error (right panel) for $\hat{\alpha}$ and $\hat{\beta}$ by Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL) and Integrated Pseudo CL (IPCL). Based on 500 replications of GARCH panels (exhibiting strong cross-section dependence) for varying T and N where true parameter values are given by $\alpha = 0.05$ and $\beta = 0.93$.

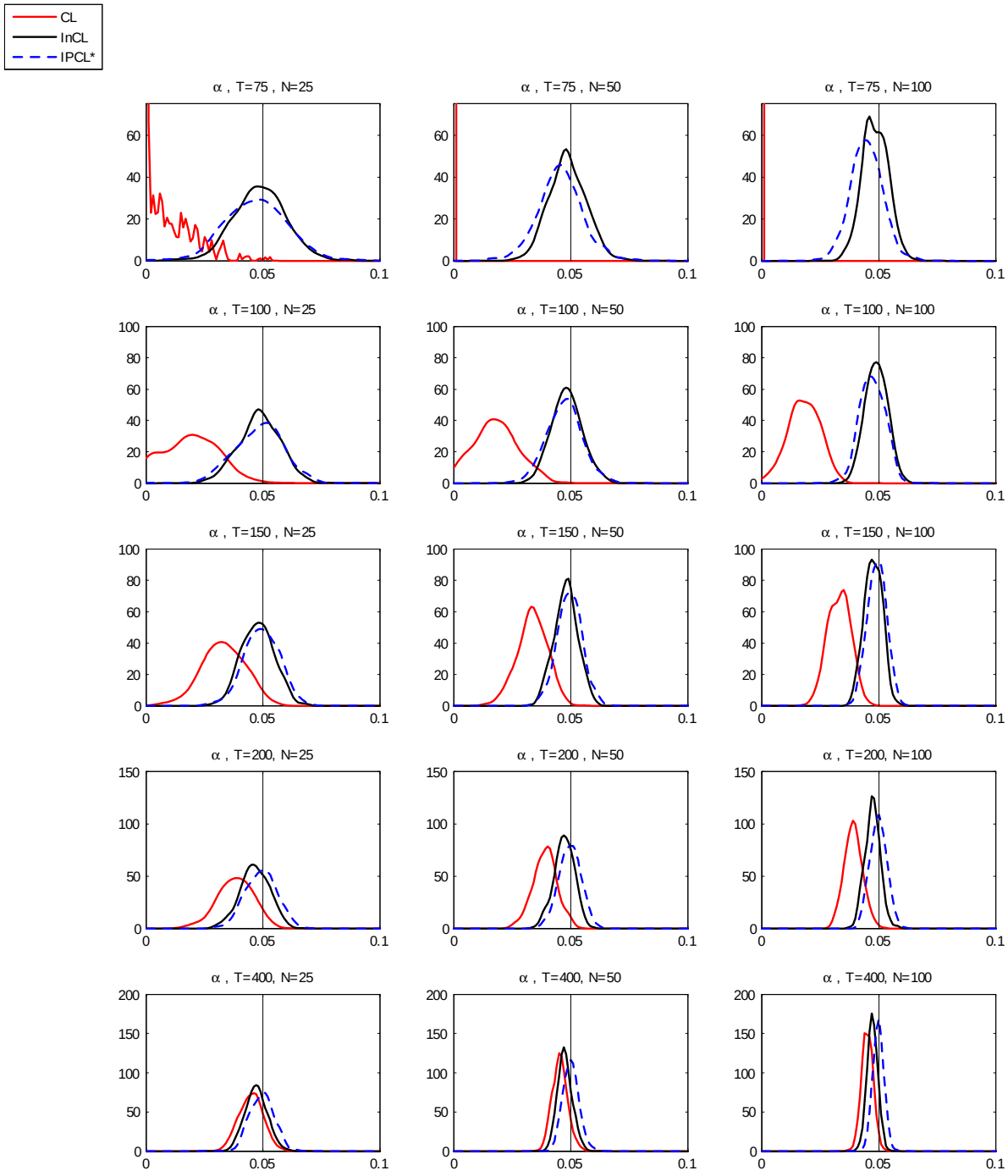

Figure 4.4: Sample distributions of $\hat{\alpha}$ using the composite likelihood (CL), infeasible CL (InCL) and integrated pseudo CL (IPCL*) methods. IPCL* is estimated by integrating both the initial value and the individual-specific parameter out. The vertical line is drawn at the true parameter value. Based on 500 replications under cross-section independence where $(\alpha, \beta) = (0.05, 0.93)$.

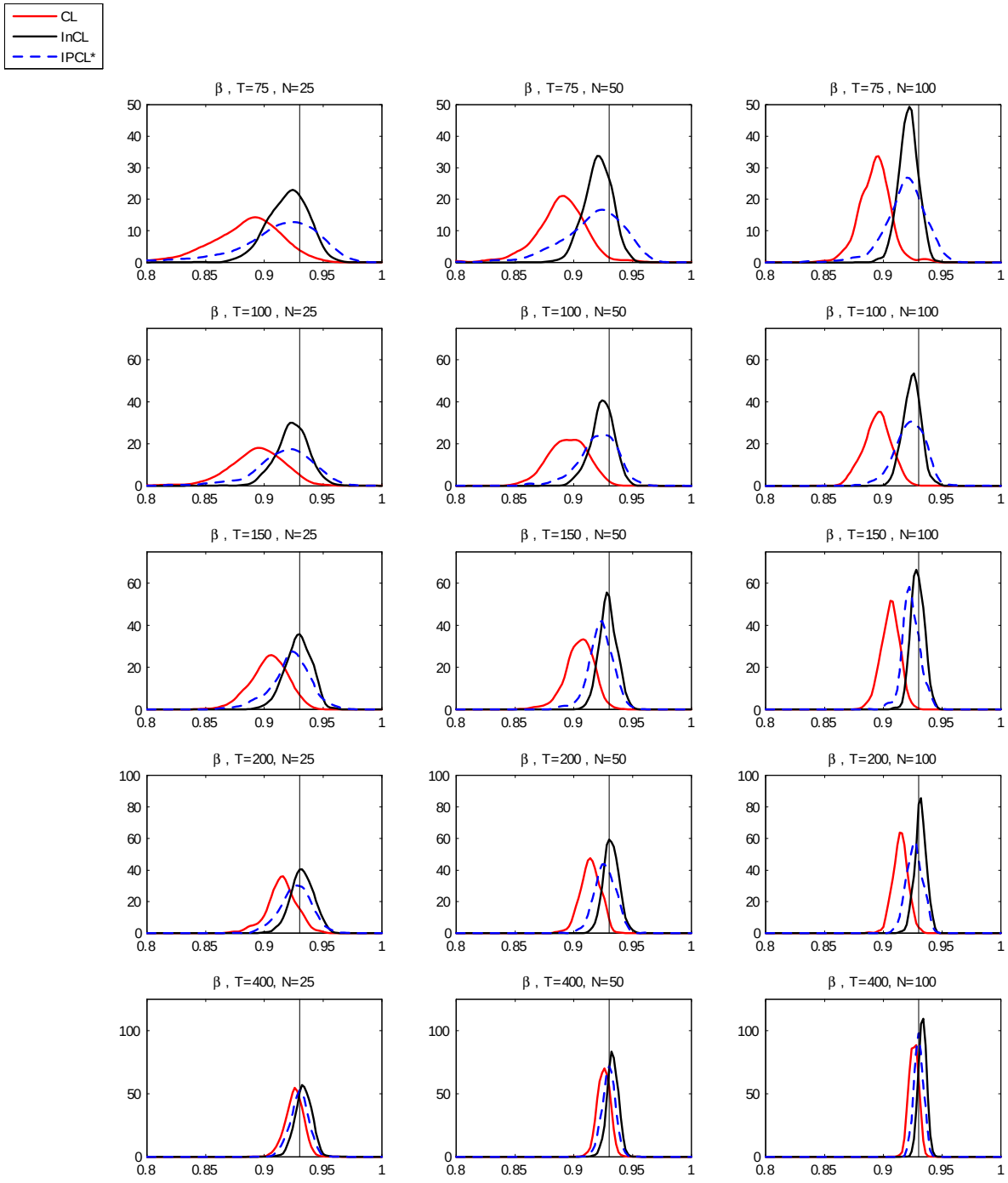

Figure 4.5: Sample distributions of $\hat{\beta}$ using the composite likelihood (CL), infeasible CL (InCL) and integrated pseudo CL (IPCL*) methods. IPCL* is estimated by integrating both the initial value and the individual-specific parameter out. The vertical line is drawn at the true parameter value. Based on 500 replications under cross-section independence where $(\alpha, \beta) = (0.05, 0.93)$.

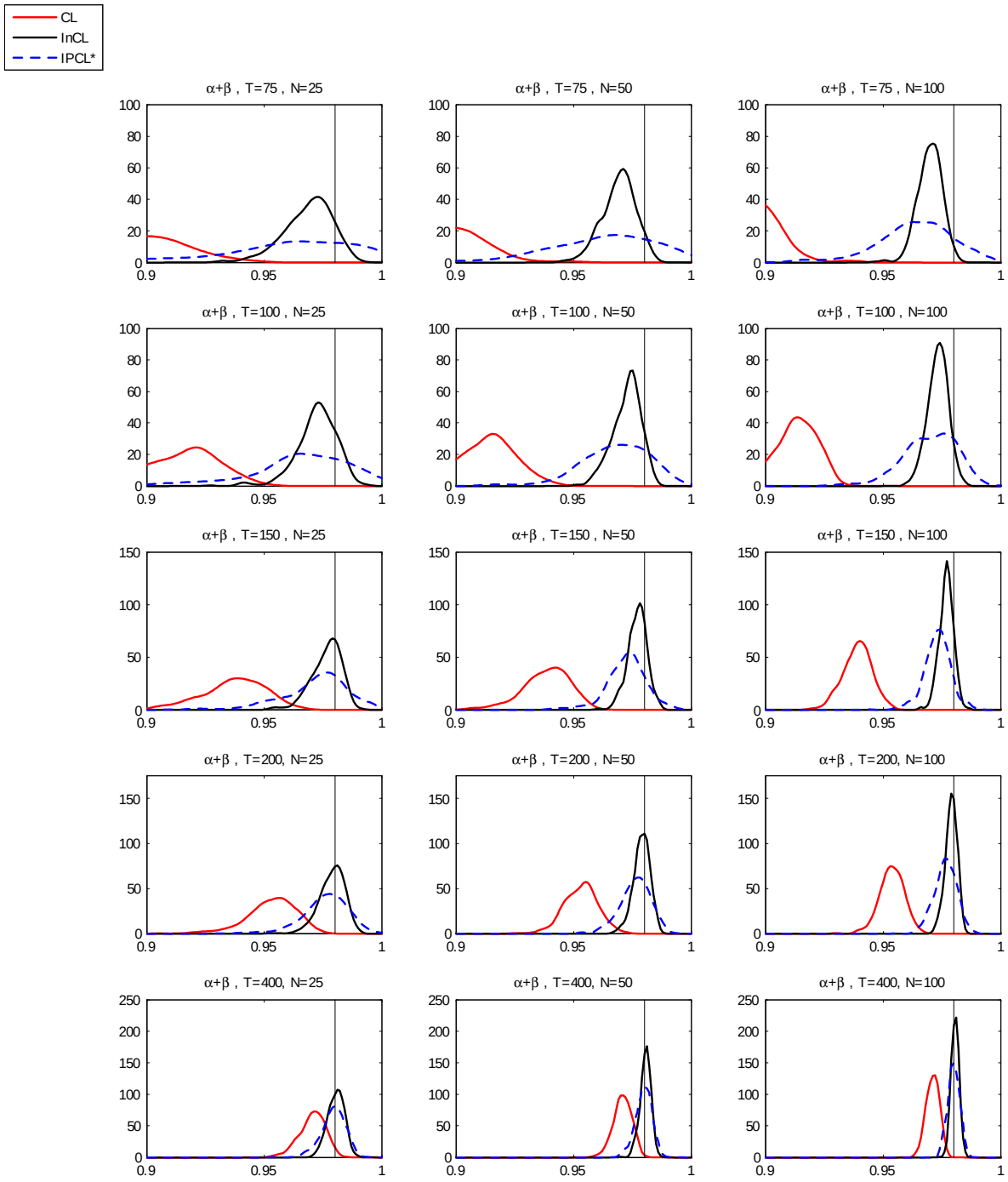

Figure 4.6: Sample distributions of $\hat{\alpha}+\hat{\beta}$ using the composite likelihood (CL), infeasible CL (InCL) and integrated pseudo CL (IPCL*) methods. IPCL* is estimated by integrating both the initial value and the individual-specific parameter out. The vertical line is drawn at the true parameter value. Based on 500 replications under cross-section independence where $(\alpha, \beta) = (0.05, 0.93)$.

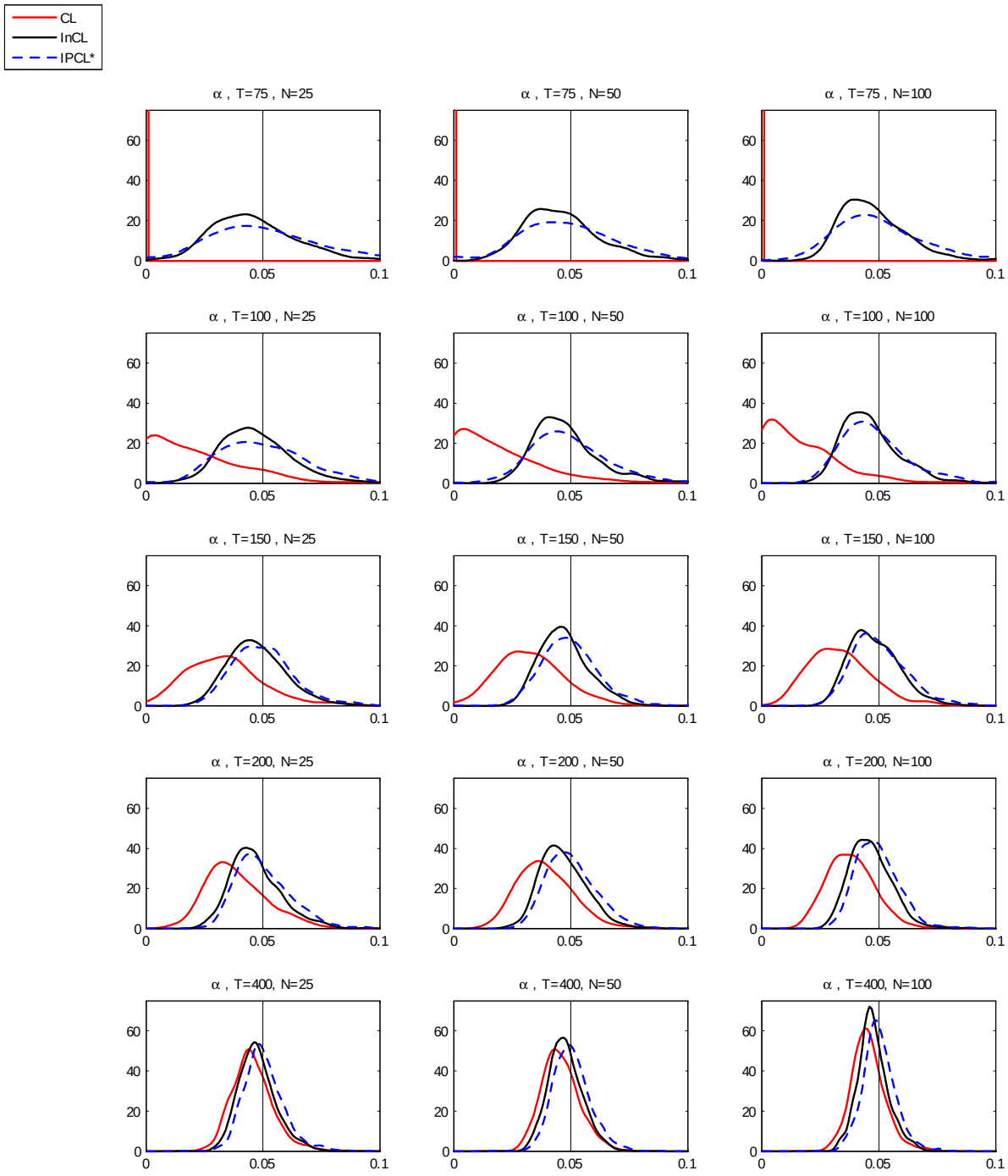

Figure 4.7: Sample distributions of $\hat{\alpha}$ using the composite likelihood (CL), infeasible CL (InCL) and integrated pseudo CL (IPCL*) methods. IPCL* is estimated by integrating both the initial value and the individual-specific parameter out. The vertical line is drawn at the true parameter value. Based on 500 replications under mild cross-section dependence where $(\alpha, \beta) = (0.05, 0.93)$.

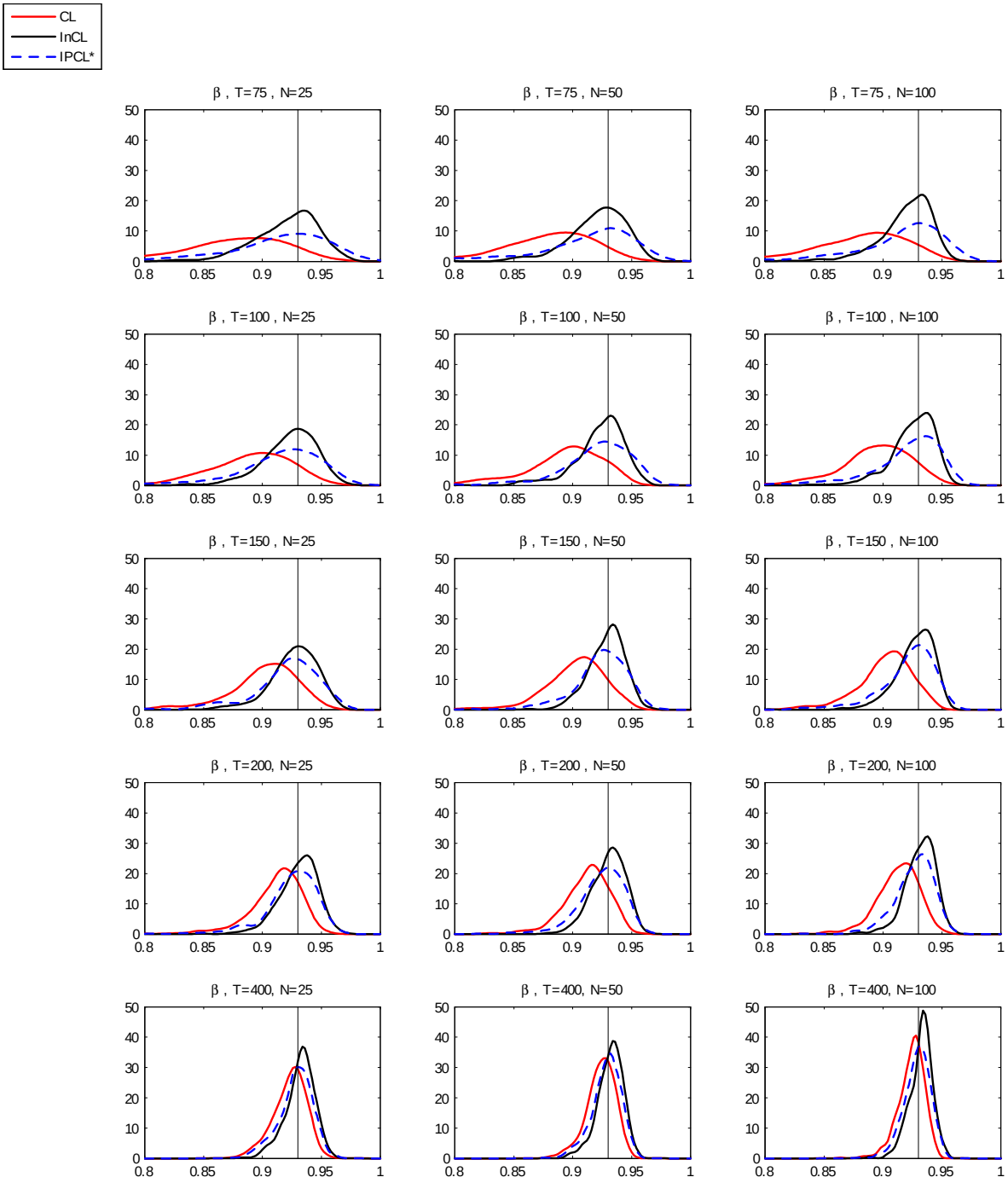

Figure 4.8: Sample distributions of $\hat{\beta}$ using the composite likelihood (CL), infeasible CL (InCL) and integrated pseudo CL (IPCL*) methods. IPCL* is estimated by integrating both the initial value and the individual-specific parameter out. The vertical line is drawn at the true parameter value. Based on 500 replications under mild cross-section dependence where $(\alpha, \beta) = (0.05, 0.93)$.

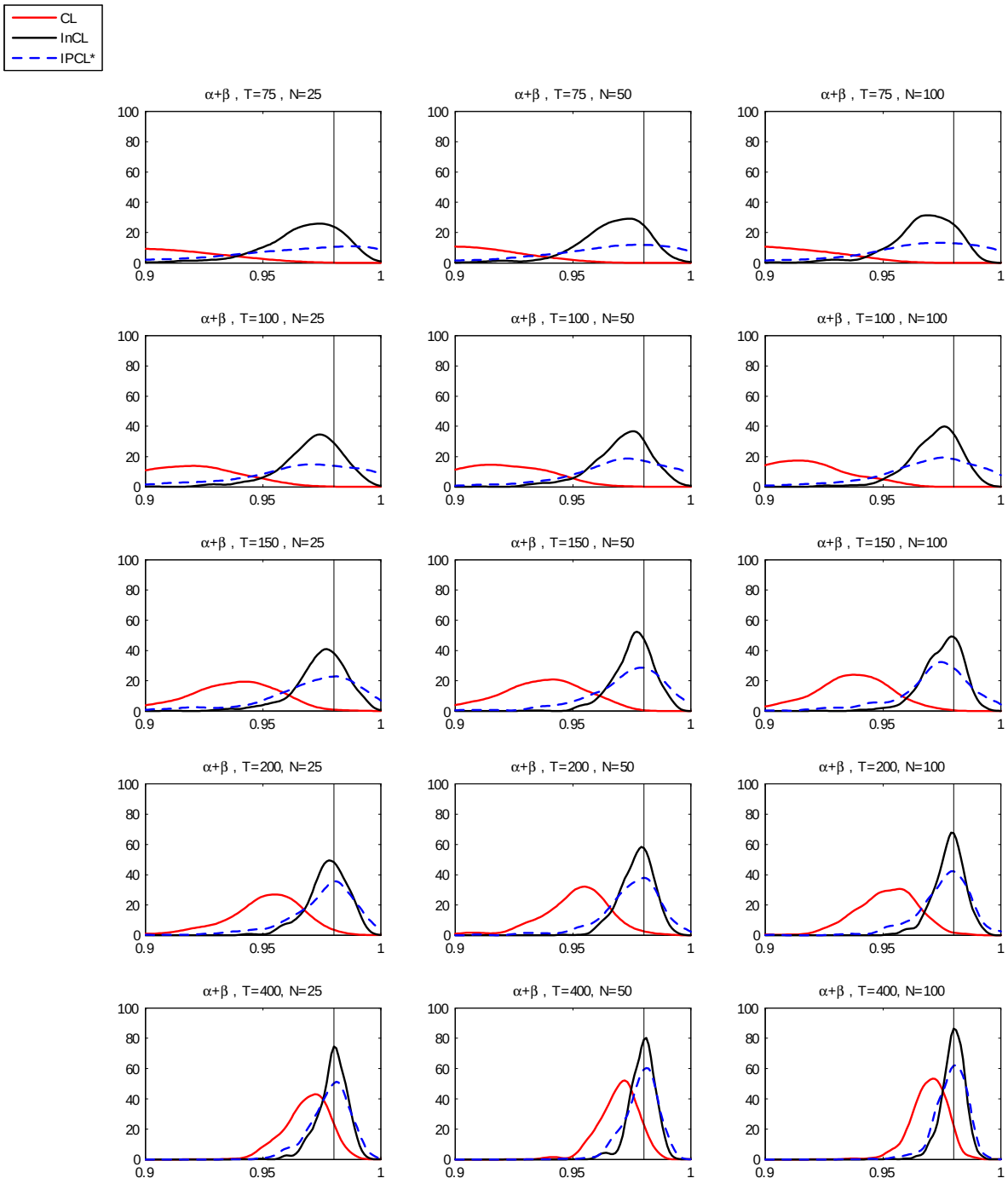

Figure 4.9: Sample distributions of $\hat{\alpha}+\hat{\beta}$ using the composite likelihood (CL), infeasible CL (InCL) and integrated pseudo CL (IPCL*) methods. IPCL* is estimated by integrating both the initial value and the individual-specific parameter out. The vertical line is drawn at the true parameter value. Based on 500 replications under mild cross-section dependence where $(\alpha, \beta) = (0.05, 0.93)$.

the theoretical observation that bias reduction would not lead to higher standard deviation, simulation results disagree. Comparison of the sample standard deviations for IPCL and IPCL* in Tables 4.2 and 4.6 shows that the new initial value selection method leads to larger variance. The increase is substantial for $\hat{\beta}$ when T is small. However, the difference between the two methods diminishes as T increases. For $\hat{\alpha}$, the gains in bias are large enough to offset the increase in variance which leads to lower $\mathcal{R}_{\hat{\alpha}}$ for IPCL*. However, the increase in variance for $\hat{\beta}$ is substantial enough to inflate $\mathcal{R}_{\hat{\beta}}$, as well. Curiously, the increase is so large that it even exceeds the RMSE of the CL estimator of β. However, notice that CL suffers heavily from an identification issue when T is small, as a large portion of $\hat{\alpha}$ is arbitrarily close to zero. Therefore, despite better variance performance, CL is suffering from more serious issues. The issues subside when T is around 150, by when IPCL*'s RMSE diminishes to reasonable levels. Therefore, the second message of the simulation results is that the better bias performance of IPCL* comes at a cost of higher variance when T is small and this difference is significant for $\hat{\beta}$. This could imply that the seemingly simple modification to the estimation process actually has an important influence on the theoretical bias-reduction properties. Of course, a more optimistic scenario is that the higher variance is due to numerical issues related to the optimisation algorithm or to the software. This is a research question that certainly has to be addressed in the future.

Finally, the effect of changing the level of cross-section dependence is investigated. This time the methods under comparison are CL, InCL and IPCL*. The left panels of Tables 4.4, 4.6 and 4.8 show that increasing cross-section dependence unambiguously leads to higher sample standard deviation. The higher dispersion in sample distributions for all methods is also clearly observed in Figures 4.4-4.12. This is not surprising noting that the cross-section dependence affects the convergence rate of asymptotic variance. The more interesting question is whether dependence affects the properties of the first-order bias. Remember from Chapter 3 that cross-section dependence leads to an extra, different type of small sample bias, the magnitude of which is closely linked to the level of dependence. The simulation setup makes it possible to study the small sample characteristics from that perspective.

The analysis of average bias for the three methods yields interesting observations. First, the change in bias between "independence" and "mild dependence" settings is generally

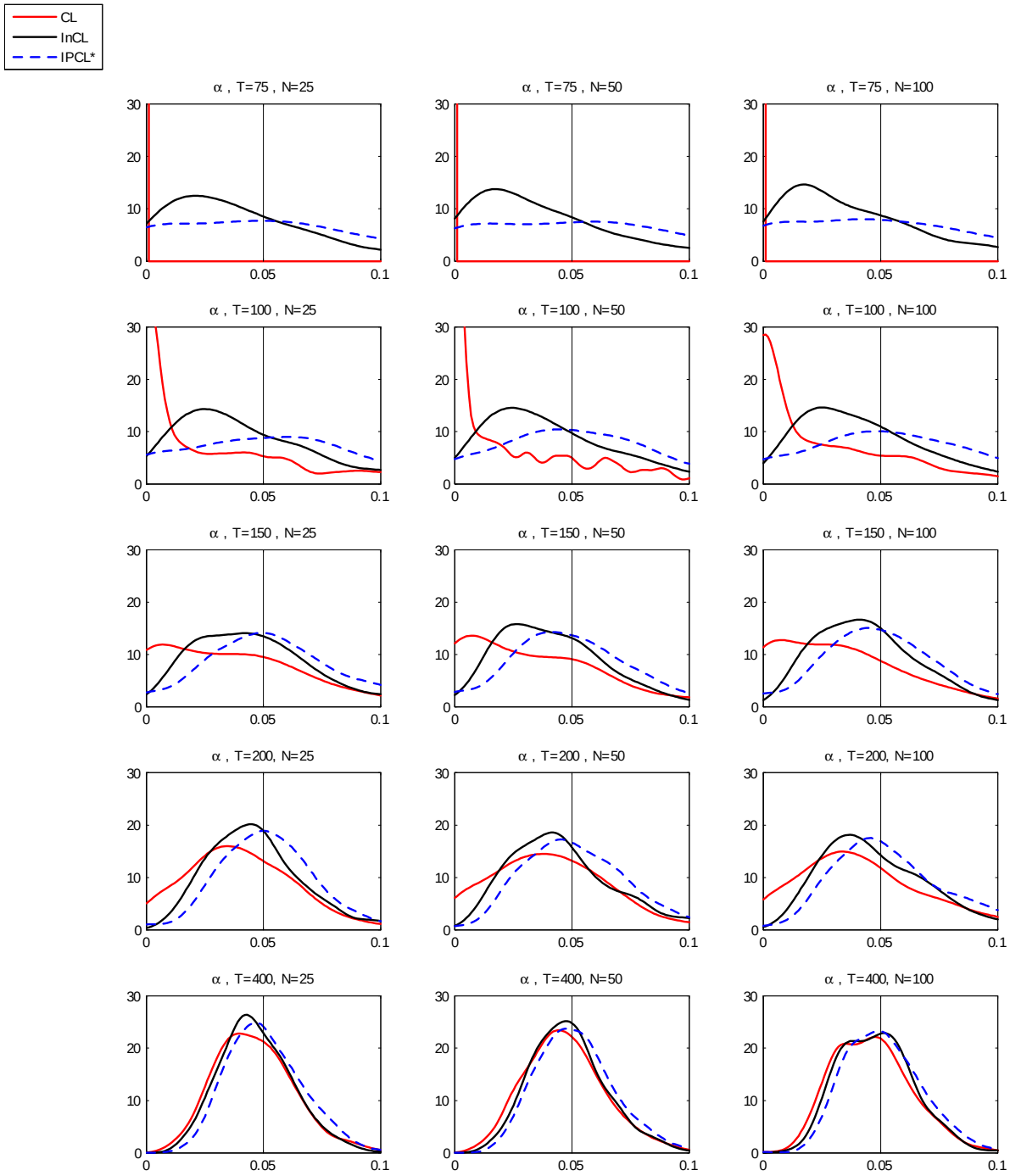

Figure 4.10: Sample distributions of $\hat{\alpha}$ using the composite likelihood (CL), infeasible CL (InCL) and integrated pseudo CL (IPCL*) methods. IPCL* is estimated by integrating both the initial value and the individual-specific parameter out. The vertical line is drawn at the true parameter value. Based on 500 replications under strong cross-section dependence where $(\alpha, \beta) = (0.05, 0.93)$.

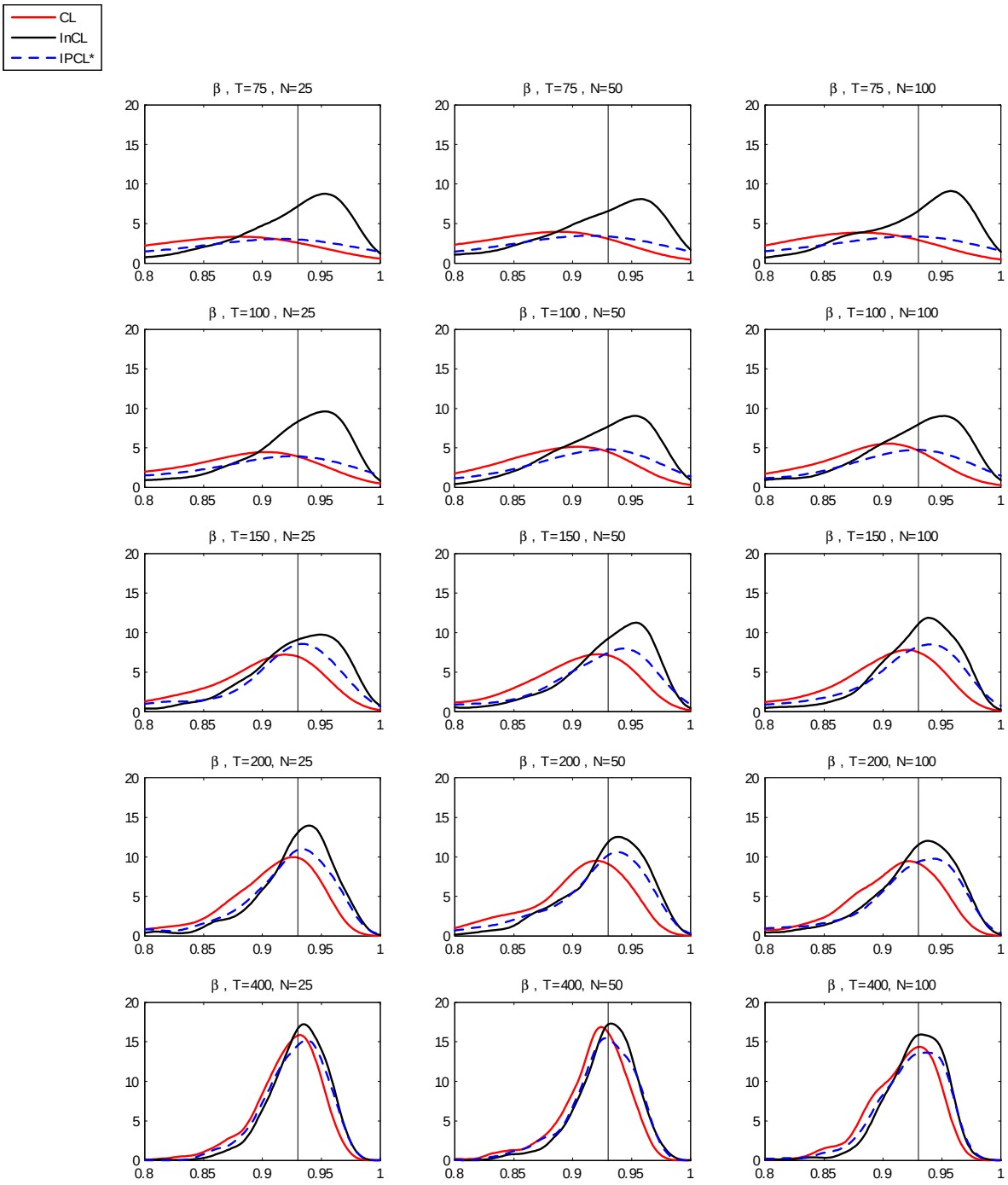

Figure 4.11: Sample distributions of $\hat{\beta}$ using the composite likelihood (CL), infeasible CL (InCL) and integrated pseudo CL (IPCL*) methods. IPCL* is estimated by integrating both the initial value and the individual-specific parameter out. The vertical line is drawn at the true parameter value. Based on 500 replications under strong cross-section dependence where $(\alpha, \beta) = (0.05, 0.93)$.

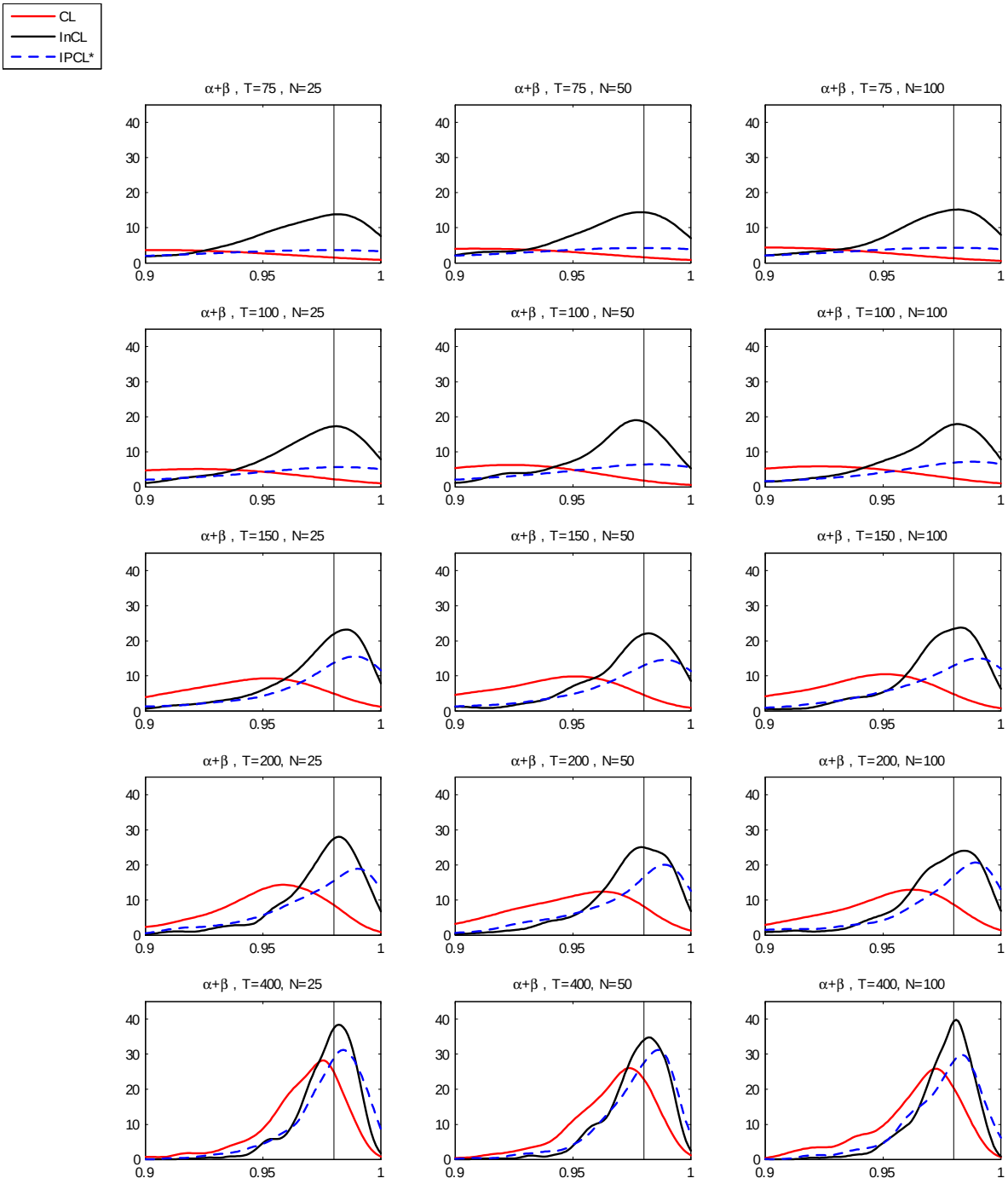

Figure 4.12: Sample distributions of $\hat{\alpha} + \hat{\beta}$ using the composite likelihood (CL), infeasible CL (InCL) and integrated pseudo CL (IPCL*) methods. IPCL* is estimated by integrating both the initial value and the individual-specific parameter out. The vertical line is drawn at the true parameter value. Based on 500 replications under strong cross-section dependence where $(\alpha, \beta) = (0.05, 0.93)$.

negligible (compare Tables 4.3 and 4.5). Most of the difference is observed for $T = 75$ which quickly vanishes as T increases. This is in line with the theoretical results. Second, the transition from "mild" to "strong" dependence yields a striking picture (compare Tables 4.5 and 4.7). All methods suffer from bias and in the case of CL and IPCL the bias is still severe when T is as large as 200. It is important to underline that even InCL, which is not much damaged from "mild" dependence, suffers visibly from higher dependence when estimating β. This supports the theoretical findings of Chapter 3 in the case of GARCH panels, and suggests that higher cross-section dependence leads to small sample bias, even in the absence of the incidental parameter issue, as demonstrated by the less than satisfactory performance of InCL under "strong" dependence.

An intriguing observation is that CL actually performs very well in terms of the average bias of $\hat{\alpha}$, in the presence of "strong" dependence. However, these results are deceptive, as a quick look at Figure 4.10 reveals. Actually, CL is suffering hugely from higher dependence. There are now a much larger number of cases where $\hat{\alpha} \approx 0$ and these do not entirely vanish even when $T = 200$. Clearly, the average bias results do not reflect this issue.

4.3.5 Analysis of Likelihoods

Finally, average likelihood plots, based on the 500 replications, for several panel dimensions are provided in Figure 4.13. Since ICL and IPCL behave similarly, only the plots for ICL are presented. Average likelihood for varying values of α are plotted by fixing the likelihood with respect to the true value of β (and similarly for the average likelihood for varying values of β). The plots for CL are based on estimated values of the nuisance parameters, while infeasible CL plots are based on the true nuisance parameter values.

Likelihood plots immediately confirm that the problem with CL is that the likelihood for α is wrongly centred. As a result, estimates of α are always close to the boundary. As T increases, the mode of the average likelihood moves towards the true value of α. For β, on the other hand, the major problem is that the likelihood is almost flat, implying that β is not identified. This is not surprising, since, as mentioned previously, β is not identified when $\alpha = 0$. Only when T increases does the average likelihood show some improvement. Moreover, it is clear that ICL is effective in correcting the location of the likelihood. This also solves the identification problem for β, as can be seen from the average ICL for β,

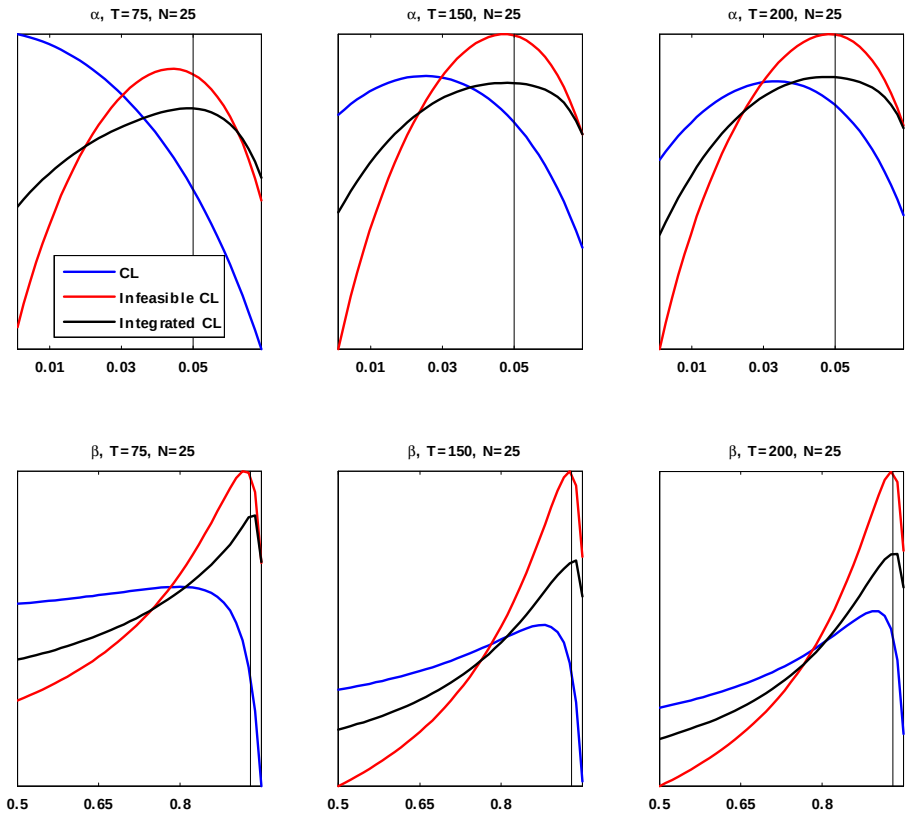

Figure 4.13: Average likelihood plots for α and β. Based on likelihood averages over 500 replications (with cross-sectional dependence). In the upper panel, β is fixed at 0.93 while the lower panel is based on $\alpha = 0.05$. CL is evaluated at the sample estimates of λ_i, while Infeasible CL is evaluated at the true values of λ_i. In calculating Integrated CL, priors for each replication are evaluated at the parameter estimates from the penultimate iteration for that particular replication. Vertical lines are drawn at the true parameter values of $\alpha = 0.05$ and $\beta = 0.93$.

which is not flat and its shape is similar to that of the average infeasible CL. These findings further attest the effectiveness of robust priors in removing the first-order bias.

4.3.6 Summary

The simulation analysis presented here confirms that the integrated likelihood based bias reduction approach successfully removes a substantial proportion of the small sample bias. This method offers two strategies where the researcher can choose whether to incorporate the selection of the initial value into the bias-reduction mechanism or not. When initial value selection is incorporated into the analysis, simulation results reveal that sample distributions of bias-reduced estimators achieve a better match with those of the infeasible estimation method. However, this also leads to an increase in variance for $\hat{\beta}$, especially when T is small.

In addition, it has also been demonstrated that in the specific case of GARCH panels, realistic levels of cross-section dependence affect bias properties negligibly when T is reasonably large. However, as the level of dependence increases, all methods suffer from bias which does not entirely vanish even when $T \geq 200$.

Finally, the analysis of average likelihood plots also confirms that the bias-reduction method leads to correctly centred likelihood functions. This also solves the issue with the identification of β in small samples.

These observations confirm that the integrated likelihood method is effective in estimating volatility when T is as small as 150-200. Armed with these results, the next step is to take the model to real data.

4.4 Empirical Analysis

This section presents two empirical studies of the bias-reduced GARCH panel estimator. The first is a comparison of predictive ability, based on stock return volatility forecasts by different methods. The second is an analysis of hedge fund volatility using a consolidated database of hedge fund returns. Hedge fund returns are rarely available at higher than monthly frequency and the maximum number of observations for any fund is around 200. This makes it virtually impossible to analyse hedge fund volatility using standard

GARCH estimation techniques. Hence, this empirical analysis is a novel contribution to the literature. In both applications, naturally, a pseudo-likelihood setting is assumed and the integrated likelihood functions are constructed using the pseudo-likelihood Prior given in (P2).

4.4.1 Analysis of Predictive Ability

Dataset

The analysis of predictive ability is based on daily data on returns to nine stocks traded in the Dow Jones Industrial Average. The dataset has been downloaded from the *Oxford-Man Institute's Realized Library* (produced by Heber, Lunde, Shephard and Sheppard (2009)) and is based on data used by Noureldin, Shephard and Sheppard (2011). The dataset covers the period between 1 February 2001 and 28 September 2009 and is from the TAQ database. The included stocks are Alcoa, American Express, Bank of America, Coca Cola, Du Pont, Exxon Mobil, General Electric, IBM and Microsoft.

The comparison of predictive ability is based on the Giacomini-White test of predictive ability, which has been used and outlined in detail in Chapter 2 (see Section 2.B for an overview). An important advantage of the chosen dataset is that it includes realised variances for each stock, in addition to daily returns. As discussed previously, volatility is latent and a proxy is required in order to measure the loss due to a particular volatility forecast. Realised variance is a well-known accurate proxy for volatility and therefore, is preferable to the simple but inaccurate squared returns as volatility proxy. This is the main motivation behind using this dataset.[5]

For a more detailed explanation on the features of the dataset and estimation of the realised variances, see Noureldin, Shephard and Sheppard (2011). In particular, they report that both the returns and the realised variances are open-to-close due to market microstructure noise. In addition, the first and last 15 minutes of trading are dropped from the sample in order to deal with overnight effects. Lastly, realised variances are based on 5-minute returns with subsampling.

[5] It would be desirable to base the analysis on panels with a larger cross-section dimension. However, estimation of realised variances for a random selection of stocks is a non-trivial and highly time-consuming task. In addition, a given stock may not be liquidly traded to start with, which implies complications for realised variance estimation. For these reasons, this is not pursued here.

Forecast Construction

This study focuses on one-step ahead forecasts only, for sake of brevity. The one-step ahead forecasts for a given set of estimators $(\hat{\alpha}, \hat{\beta}, \hat{\lambda}_1, ..., \hat{\lambda}_N)$ are obtained by using

$$\begin{aligned} \mathbb{E}[\varepsilon_{it}^2|\mathcal{F}_{i,t-1}] = \sigma_{it}^2 &= \lambda_i(1-\alpha-\beta) + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2, \\ \hat{\sigma}_{it}^2 &= \hat{\lambda}_i(1-\hat{\alpha}-\hat{\beta}) + \hat{\alpha}\varepsilon_{i,t-1}^2 + \hat{\beta}\sigma_{i,t-1}^2. \end{aligned}$$

The three methods under consideration are the Quasi Maximum Likelihood (QML), Composite Likelihood (CL) and Integrated Pseudo Composite Likelihood (IPCL) methods. QML is the standard way of fitting the GARCH model, where GARCH parameters are estimated individually for each time-series under consideration. QML and CL are based on the two-step estimation framework which has been investigated in detail in Chapter 2 (see Section 2.3). To summarise briefly, this method exploits the variance targeting version of GARCH given in (4.1). In the first step, λ_i are estimated by the method of moments estimator,

$$\tilde{\lambda}_i = T^{-1}\sum_{t=1}^{T} y_{it}^2. \tag{4.6}$$

The first-step estimators are then plugged into the likelihood function in order to estimate the parameters of interest in the second step. As for the integrated likelihood method, the particular parameterisation of the intercept parameter is of no consequence as the intercept is integrated out anyway. The only consideration that matters is that the integral is evaluated over an interval which includes the true parameter value.[6]

An interesting question is how to estimate the intercept parameter for the IPCL method. When the main objective is to obtain consistent and bias-corrected estimators of parameters of interest, the individual effects are not of direct importance and they are indeed nuisance parameters in the literal sense. However, when the interest is in making predictions, the intercept has to be estimated, as well. This is an important distinction from the traditional bias-reduction literature. For all methods under consideration, the method of moments estimator $\tilde{\lambda}_i$ is consistent and valid independent of how α and β are estimated. However, remember that the integrated likelihood function is in essence an approximation to a spe-

[6] To achieve this, when calculating the integrated likelihood the lower and upper limits of the integral are set to $.8 \times (\min_{i,t} r_{it}^2)$ and $2 \times (\max_{i,t} r_{it}^2)$, respectively.

cial concentrated likelihood function, the target likelihood. Therefore, a natural intercept estimator is given by

$$\hat{\lambda}_i^c(\hat{\theta}_{IL}) = \arg\max_{\lambda_i \in \Lambda_i} \frac{1}{T} \sum_{t=1}^{T} \ell_{it}(\hat{\theta}_{IL}, \lambda_i). \tag{4.7}$$

As λ_i are estimated for each time-series individually, estimation by the concentrated likelihood method comes at little cost in terms of computation time.[7] For QML and CL, why λ_i would be estimated by a similar method is less obvious as these methods do not estimate θ by concentrated likelihood to begin with.[8]

The Test Procedure

Comparisons of the predictive ability of the three methods are done using the Giacomini and White (2006) unconditional predictive ability test (GW-test henceforth). Forecasts are constructed using a rolling window scheme, where the in-sample size is fixed at 150. Specifically, the first forecast is calculated using estimates that are based on observations $t = 1$ to $t = 150$. The second forecast is then calculated using estimates that are based on observations $t = 2$ to $t = 151$, and so on. Therefore, successive forecasts are always based on the most recent 150 observations.[9] The dataset consists of $2,176$ observations, implying a total of $2,026$ forecasts for each of the nine stocks.

The test procedure is identical to the one used in Chapter 2. To briefly describe it, suppose $\hat{\sigma}^2_{1,i,t+1}$ and $\hat{\sigma}^2_{2,i,t+1}$ are the one-step ahead forecasts for stock i calculated at time t by two different methods. Accuracy of these forecasts is measured by using the QLIKE loss function,

$$L(\sigma^2_{i,t+1}, \hat{\sigma}^2_{i,t+1}) = \log \hat{\sigma}^2_{i,t+1} + \frac{\sigma^2_{i,t+1}}{\hat{\sigma}^2_{i,t+1}}.$$

For the bias-reduction method, results for both estimation strategies considered in the simulation analysis are reported. These are given by IPCL and IPCL*, as before. In addition, results for the two possible intercept estimators (4.6) and (4.7) are also reported.

[7] From a theoretical perspective, both this and the method of moments estimators are consistent and valid. However, there might be different implications in small samples, as will be illustrated below.

[8] Remember that CL uses $\tilde{\lambda}_i = T^{-1}\sum_{t=1}^{T} y_{it}^2$ to construct $(NT)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N} \ell_{it}(\theta, \tilde{\lambda}_i)$ which is not necessarily the same as $(NT)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N} \ell_{it}(\theta, \hat{\lambda}_i(\theta))$ where $\hat{\lambda}_i(\theta) = \arg\max_{\lambda_i} T^{-1}\sum_{t=1}^{T} \ell_{it}(\theta, \lambda_i)$.

[9] Remember that this is done in order to make sure that the test of predictive ability compares estimation methods and not models. The latter option would be problematic because all considered methods estimate the same model.

Results

The test results are given in Tables 4.9 and 4.10. All CL and QML forecasts are based on the intercept parameter estimates by $\tilde{\lambda}_i$ given by (4.6). For IPCL and IPCL*, results for forecasts based on the concentrated likelihood estimator of λ_i by (4.7) are given in Table 4.9 while the respective results based on (4.6) are presented in Table 4.10. Each table contains the t-statistics and the result of the GW-test. A dash signifies that the test result is inconclusive. All tests are done at 5% level of significance.

GW-test results reported in Table 4.9 indicate that bias-reduction based forecasts achieve the best forecasting performance compared to both QML and CL. In addition, integrating the initial parameter out improves the performance of the integrated likelihood method: IPCL beats CL in four cases, whereas IPCL* is the preferred method in six cases in the same comparison. Similarly, IPCL* is preferred to ML in seven cases while IPCL in the chosen against ML in six cases. These results have two implications. First, using bias-corrected estimators leads to superior forecasting performance. Second, integrating the initial parameter out improves the performance of bias-corrected estimators. In the comparison between CL and bias-reduced estimation, this improvement is by 50%.

Among all three methods, not surprisingly, QML always leads to higher loss, as indicated by the test statistics. Another interesting observation is that CL is preferred to QML only three times, whereas the superiority of IPCL* to QML is certain. These observations suggest that, although CL stands out as an alternative to standard QML estimation, it is hard to claim that it is unambiguously superior. However, test results confirm without doubt that the integrated likelihood method provides the necessary extra power to panel GARCH estimation. Hence, IPCL* emerges as the best performer and bias-reduction clearly improves the performance of panel-based estimation.

Whether estimating λ_i by concentrated likelihood, rather than the method of moments, leads to a difference in the performance of bias-corrected estimators is an interesting question in its own right. In large samples, not much difference would be expected as both estimators are consistent. However, results reported in Table 4.10 suggest that in this small sample exercise this is not the case. Although both IPCL and IPCL* still outperform QML, they do so less decisively, while the comparison between CL and IPCL is not clear as the GW-test

	QML vs CL		CL vs IPCL		QML vs IPCL		CL vs IPCL*		QML vs IPCL*	
Stock	t-stat	Result	t-stat	Result	t-stat	Result	t-stat	Result	t-stat	Result
Alcoa	0.993	-	2.563	IPCL	2.144	IPCL	2.668	IPCL*	2.217	IPCL*
American Express	1.451	-	3.728	IPCL	3.920	IPCL	3.168	IPCL*	3.444	IPCL*
Bank of America	1.109	-	2.550	IPCL	2.968	IPCL	3.862	IPCL*	3.271	IPCL*
Du Pont	2.288	CL	1.168	-	2.044	IPCL	2.587	IPCL*	3.030	IPCL*
General Electric	0.960	-	2.176	IPCL	2.428	IPCL	2.142	IPCL*	2.529	IPCL*
IBM	1.815	-	1.419	-	2.036	IPCL	2.067	IPCL*	2.493	IPCL*
Coca Cola	2.870	CL	-1.372	-	1.093	-	-0.932	-	1.541	-
Microsoft	2.755	CL	-1.381	-	0.844	-	-0.427	-	1.879	-
Exxon Mobil	1.404	-	0.986	-	1.650	-	1.920	-	2.404	IPCL*

Table 4.9: Giacomini-White test results for GARCH panels. The level of significance is 5%. The methods under consideration are quasi maximum likelihood (QML), composite likelihood (CL) and integrated pseudo composite likelihood. Forecasts for IPCL have been calculated using two different methods. The first version uses the initial value given in (4.5) (IPCL) whereas the second version integrates the initial values out along with λ_i (IPCL*). The result of each test is given in the 'Result' column while t-statistics are reported in the 't-stat' column. A dash signifies that the test is inconclusive. Realised volatility is used as volatility proxy. λ_i is estimated using the method of moments estimator for the CL and QMLE methods while the intercept parameter for ICL is estimated using the concentrated likelihood method, as defined in (4.7).

results in a draw. IPCL*'s performance also deteriorates: this time, it is preferred in three cases only and is even beaten by CL in one case. The majority of the t-statistics is still in favour of IPCL and IPCL*, but not large enough to force rejection of the hypothesis of equal predictive ability. A final important observation is that even in this less favourable case, IPCL* still performs better than IPCL in all comparisons. This is another motivation in favour of using IPCL* in empirical applications.

To conclude, this empirical exercise indicates that the bias reduction method attains superior predictive performance. Moreover, test results suggest that the forecasts should be based on (i) the concentrated likelihood estimator of λ_i and (ii) integration of both the individual specific parameter and the initial value out.

4.4.2 Hedge Fund Analysis

Hedge funds are alternative investment vehicles comprising one of the fastest growing industries: the total value of assets under management has increased from $50 billion in 1990 to $1 trillion in 2004. In April 2011, the global assets under management were expected to reach $2.25 billion by the end of 2011, despite capital outflows following the credit crunch episode.[10] Some of the peculiar features of hedge funds are that they are less regulated and less transparent. For example, it is entirely up to a given fund whether to supply data or not. Moreover, often there are mandatory lockup periods whereby investors cannot withdraw their investment before a certain period which could be as long as a few years.

Hedge fund returns are usually reported at monthly frequency. As databases generally start around 1994, the maximum number of time-series observations for any given fund is around 200 (and possibly much smaller than that). Clearly, this is well below what is necessary for traditional GARCH estimation to be successful. However, as the simulation results indicate, the bias-corrected GARCH Panel model is well-suited to the task.

Estimation of hedge fund volatility is interesting for a number of reasons. First, the ability to model volatility using the GARCH model is a novel capability which opens up potential research avenues for the analysis of hedge fund returns. Due to limitations of data, such analysis has hitherto been virtually impossible. One relevant analysis is by Huggler (2004) who argues that modelling hedge fund portfolio returns is problematic due to the

[10] Sources: *The Economist*, June 10, 2004; *Financial Times*, March 10, 2011.

	CL vs IPCL		QML vs IPCL		CL vs IPCL*		QML vs IPCL*	
Stock	t-stat	Result	t-stat	Result	t-stat	Result	t-stat	Result
Alcoa	0.828	-	1.306	-	1.772	-	1.629	-
American Express	1.211	-	2.464	IPCL	3.862	IPCL*	4.242	IPCL*
Bank of America	2.141	IPCL	2.190	IPCL	4.128	IPCL*	3.419	IPCL*
Du Pont	0.345	-	1.846	-	0.921	-	2.003	IPCL*
General Electric	-0.008	-	0.741	-	1.351	-	1.714	-
IBM	0.955	-	1.683	-	1.834	-	2.279	IPCL*
Coca Cola	-1.227	-	2.179	IPCL	-1.518	-	1.460	-
Microsoft	-2.064	CL	1.814	-	-2.057	CL	1.384	-
Exxon Mobil	1.339	-	1.888	-	2.453	IPCL*	2.722	IPCL*

Table 4.10: Giacomini-White test results for GARCH panels. The level of significance is 5%. The methods under consideration are quasi maximum likelihood (QML), composite likelihood (CL) and integrated pseudo composite likelihood. For all methods under consideration, the intercept parameter has been estimated using the method of moments estimator, as defined in 4.6. See Table 4.9 for more details.

shortness and low quality of available data. Instead, he considers constructing representative proxies for hedge fund portfolios, where he uses the standard univariate GARCH approach to model the error terms. To the best of my knowledge, the empirical illustration presented here is the only other example of hedge fund volatility modelling using GARCH errors.

Even when the volatility itself is not of direct interest, an accurate estimator of volatility can still be instrumental in analysing characteristics of hedge fund returns. For example, a popular question is how much of a fund's excess return can be attributed to manager skills, the so called *alpha*. Alpha is a measure of the manager's contribution to fund returns, in excess of the portion that is attributed to economy-wide common or systemic factors. The popular way to model excess returns is to use the seven-factor model due to Fung and Hsieh (2004).[11] As datasets are short, incorporation of serial dependence and heteroskedasticity in the specification of error terms is generally not possible, requiring the use of bootstrapped standard errors. The GARCH Panel estimator would be useful here, as it is specifically designed to model this type of dependence in short panels. A further use of volatility estimators is related to the use of volatility as a control factor. For example, Agarwal, Daniel and Naik (2011) study the case of funds that report substantially higher returns during December, compared to the rest of the year. Arguing that it is difficult to consider a time-series approach to model risk exposure (due to data being available at monthly frequency), they control for volatility by using the cross-sectional sample standard deviation of monthly returns. Again, fitted monthly volatilities for all funds individually can be obtained by using the methods proposed here. Finally, as empirical results will also attest, even within the same investment strategy, funds can vary in their levels of volatilities due to, e.g., market characteristics or manager's risk appetites (Huggler (2004)). In such a case, the integrated likelihood method provides an appropriate estimator of standard deviations, which can then be used to obtain standardised returns.

[11] See, among others, Bollen and Whaley (2009), Teo (2009) and Patton and Ramadorai (2011). These seven factors are (1) the excess returns on the S&P500 stock index; the excess returns on portfolios of lookback straddle options on (2) currencies, (3) commodities and (4) bonds; (5) the change in the credit spread of Moody's BAA bond over the 10-year Treasury bond; (6) a small minus big factor; and (7) the yield spread of the US 10-year treasury bond over the three-month Treasury bill.

Data Description

The dataset consists of monthly returns for 27,396 funds for the period between February 1994 and April 2011, implying 207 monthly returns at most for any given fund. This database of funds is a consolidation of data in the TASS, HFR, CISDM, Barclay-Hedge and Morningstar databases.[12] Importantly, funds are classified into ten vendor-reported investment strategies. These are, Security Selection, Global Macro, Relative Value, Directional Trading, Fund of Funds, Multi-Process, Emerging Markets, Fixed Income, Commodity Trading Advisors (CTA) and Other. This provides a convenient criterion for grouping funds into separate panels.

Results

The fund panels are generated as follows. First, funds which have been reporting in the last T periods are selected, where T is some chosen panel length, say $T = 150$. Then, one has to deal with the inherent biases in hedge fund data (Fung and Hsieh (2000)). Firstly, it is common for many funds to undergo an incubation period where they do not accept outside investors and build a track record on their own. Only when they have been successful for a period, they take other investors on board. Naturally, this implies that returns are biased upwards as funds that have been unsuccessful and went out of the market during incubation are not observed. A second cause of upward bias is the backfill bias. When a fund decides to list returns in a database, it has the option to report returns prior to the listing date, as well. This incentive is high for those funds with a good returns history, and low for those with a less impressive track record. The result is an upward bias in returns. To deal with these issues, funds with less than 12 months' history prior to the start date of the chosen sub-sample are dropped. Lastly, to deal with possible performance smoothing by hedge fund managers, returns for each fund are filtered using an MA(2) model, following Getmansky, Lo and Makarov (2004). Specifically, instead of raw returns, residuals from an MA(2) model are used. The resulting returns are then grouped according to the fund-reported investment strategies. By default, this implies that only live funds are considered

[12] The data consolidation process is the same as that followed in e.g. Patton and Ramadorai (2011), Ramadorai (2011) and Ramadorai and Streatfield (2011). See Appendix B in Ramadorai and Streatfield (2011) for more information on the consolidation process.

	$T = 150$			$T = 175$			$T = 195$		
Strategy	#	$\hat{\alpha}$	$\hat{\beta}$	#	$\hat{\alpha}$	$\hat{\beta}$	#	$\hat{\alpha}$	$\hat{\beta}$
Security selection	52	.202	.788	34	.174	.820	26	.179	.815
Macro	25	.114	.884	17	.093	.907	15	.105	.893
Directional Traders	51	.208	.771	24	.153	.840	16	.161	.832
Fund of funds	78	.153	.847	41	.143	.857	25	.152	.836
Multi-process	28	.176	.824	19	.165	.835	15	.230	.770
Emerging	19	.220	.772	11	.176	.794	7	.176	.801
Fixed income	13	.249	.751	8	.195	.805	5	.229	.768
CTA	41	.090	.910	22	.061	.939	15	.072	.928

Table 4.11: Integrated pseudo composite likelihood parameter estimates for hedge fund data. Estimation is based on the following three samples periods: (*i*) November 1998 - April 2011 (150 time-series observations) given in columns $2-4$, (*ii*) October 1996 - April 2011 (175 time-series observations) given in columns $5-7$ and (*iii*) February 1994 - April 2011 (195 time-series observations) given in columns $8-10$. Number of funds included in the analysis given in the '#' column.

in the analysis. Finally, all fund returns are either in or converted into US Dollars.

The maximum panel length is then 195. Clearly, longer panels will produce more reliable estimates. However, as the consolidated database in not balanced, there is a trade-off as collection of a larger cross-section of funds is only possible by considering shorter panels, and vice-versa. In fact, the strategies Global Macro and Other had to be dropped from the analysis as only a handful of funds are available even when $T = 150$. Therefore, although parameter estimates for $T \in \{150, 175, 195\}$ are reported, the analysis will focus on $T = 150$ only, to achieve maximum cross-section variation.

Parameter estimates and the number of included funds for the three sample sizes are reported in Table 4.11.[13] Estimates of α vary between .061 and .249, while $\hat{\beta}$ takes on values between .751 and .939. All strategies exhibit high memory as $\hat{\alpha} + \hat{\beta}$ is generally close to 1, across all T.[14] Moreover, values of the estimates tend to change as T varies. However, this should not entirely be attributed to changes in the sample size. The composition of the panel changes, as well, as funds with less than the necessary number of observations are dropped from the sample. Results suggest that Fixed Income, Emerging Markets and Security Selection are the strategies that are most responsive to past shocks (high $\hat{\alpha}$). CTA, Macro and Fund of Funds, on the other hand, stand out as those strategies with the lowest sensitivity to past shocks and higher responsiveness to past conditional variance (high $\hat{\beta}$).

[13] Estimation is based on the IPCL* method. However, IPCL yields similar results.

[14] Note that, technically, $\hat{\alpha} + \hat{\beta}$ is always restricted to be less than one. However, practically, they may be close to one, differing only marginally from it.

These observations hold generally, independent of the panel length.

Figure 4.14 gives an overview of fitted conditional volatilities for $T = 150$.[15] Generally, varying degrees of volatility clustering is present across all strategies. The clustering is more pronounced for, for example, Security Selection, Directional Traders and Emerging Markets. Another observation is that, even within the same strategy, there is a lot of variation between funds in terms of volatility. For almost all strategies it is possible to spot funds with volatility rarely going above, say, 5%, while some other funds are characterised by higher volatility across the whole sampling period. A few random examples of both cases are highlighted in Figure 4.14, where high-volatility funds are plotted in thick solid lines while low-volatility funds are plotted in thick broken lines. This non-uniform behaviour within strategies could be attributed either to the fact that the strategies do not comprise an objective criterion as they are self-reported or that, despite following the same strategy, some funds' specific investment strategies are more liable to be volatile due to specific market conditions, manager characteristics etc.

To have a better idea about volatility characteristics, quantiles of the sample distribution of fitted volatility across funds are plotted at each point in time in Figure 4.15. With the exception of Emerging Markets and Directional Traders, median volatility is around or less than 5%. Moreover, across all strategies, the sample distribution of volatility is asymmetric and skewed to the right. Another interesting observation is that the two important economic events in 2000s, the burst of the dotcom bubble (2000) and the credit crunch (2007-2008), have clearly had an effect on the tail behaviour of volatility distributions. This is most discernible for the 90% and 100% quantiles, although other quantiles exhibit some reaction, as well. The Fund of Funds and Macro strategies are two extreme examples where the difference between the 90% and 100% quantiles becomes enormous during these two periods. Similar changes are observed for the Macro, Multi-Process, Fixed Income and CTA strategies, as well. The Macro strategy is an interesting case, as its volatility distribution becomes skewed only during the two aforementioned periods while it is characterised by symmetry otherwise. It must nevertheless be remembered that the volatility behaviour does not necessarily have a direct implication on how well a given fund has performed. This is because GARCH is a symmetric model in the sense that it does not distinguish between

[15] Intercept parameters have been estimated using the concentrated likelihood method as in (4.7).

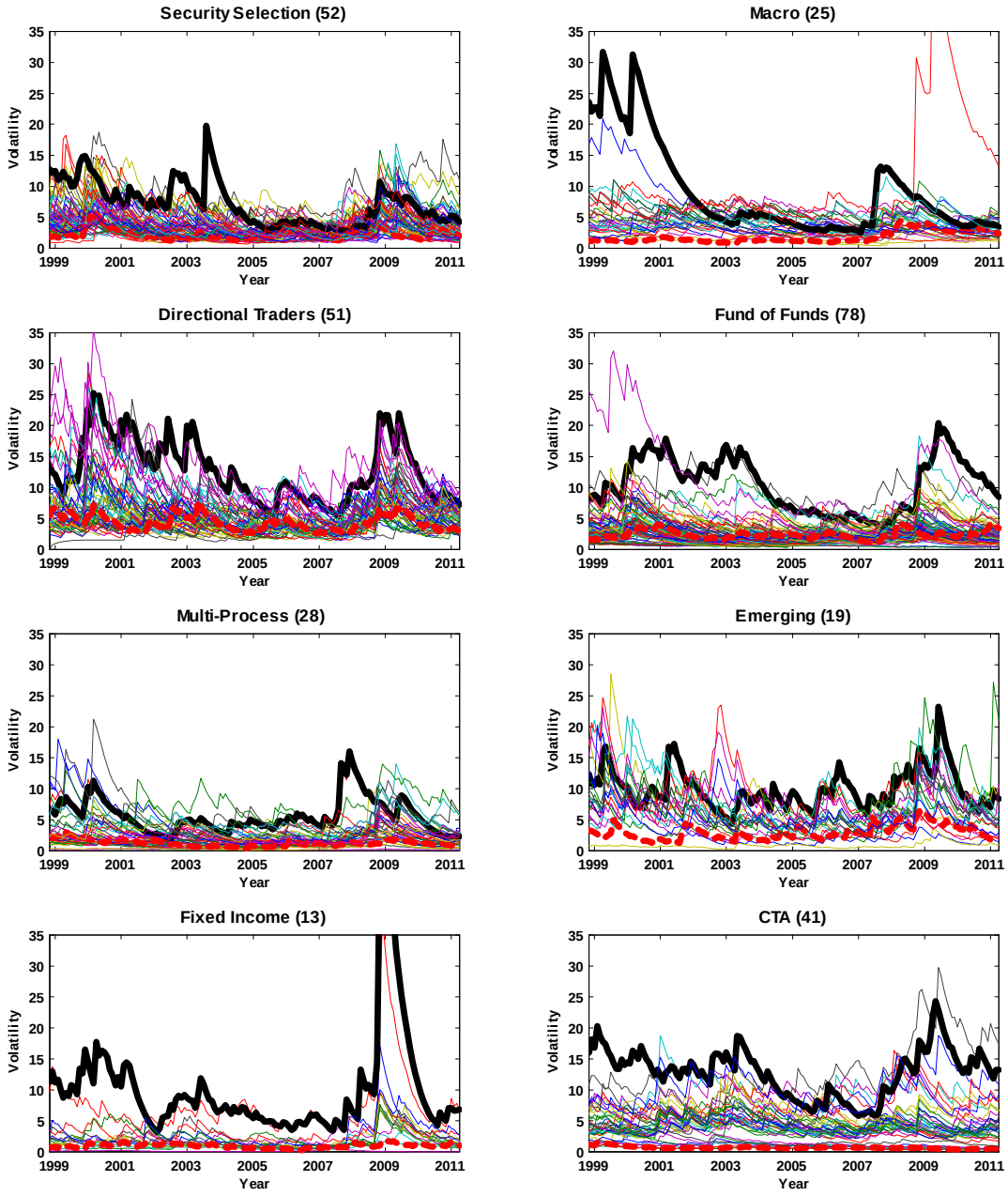

Figure 4.14: Conditional volatility plots. Based on parameters estimates by the integrated likelihood method using panels of funds that have reported non-zero returns between November 1998 and April 2011 (150 observations). Number of funds in each strategy-panel is given in parentheses. Random examples of high-volatility funds are given by thick solid lines, while the thick broken lines belong to random examples of low volatility funds.

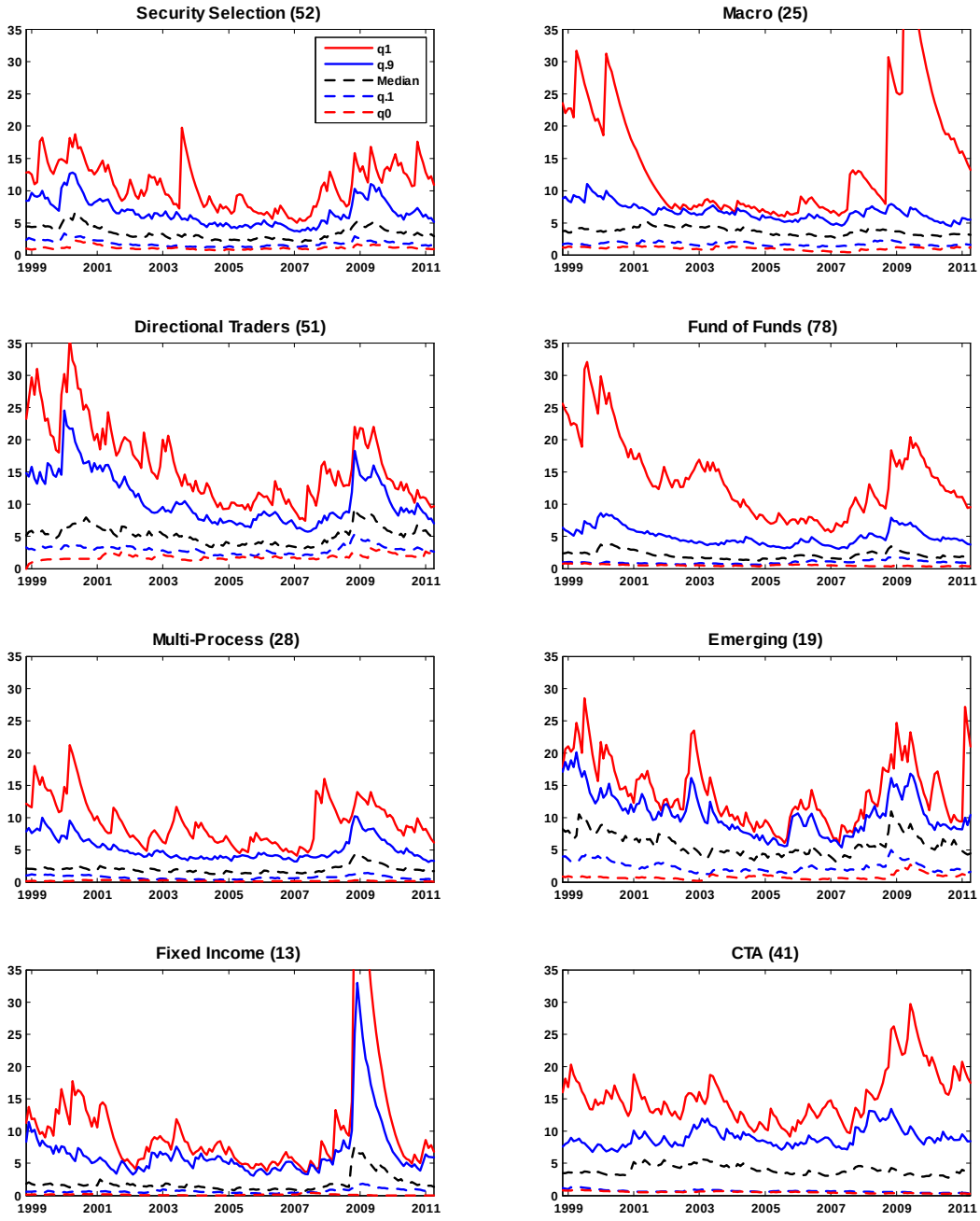

Figure 4.15: Plots of the 0%, 10%, 50% (median), 90% and 100% quantiles of the sample distribution of volatility across funds. Based on fitted conditional volatilities displayed in Figure 4.14. Number of funds in each strategy is given in parentheses.

positive and negative shocks. So, large volatility does not necessarily imply negative returns, although that would not be counter-intuitive.

The 90% quantile also exhibits variation across time, while the 10% quantile is relatively more stable. Especially for the Security Selection, Directional Traders, Multi-Process, Fixed Income and CTA strategies, the sample distributions are marked by higher volatility during economic downturns.

Lastly, Figure 4.16 presents plots of quantiles normalised by the median. This reveals some important points. First, with the exception of the Fixed Income strategy, the %90 quantile always takes on values between two to four times the median. Therefore, the dispersion of volatility distribution is more or less stable with respect to the fluctuations in the median. Second, two types of patterns for the behaviour of extreme values (100% quantile) are observed. For the Security Selection, Directional Traders and Emerging Marketsf strategies, the size of the right-tail does not change much once normalised by median. However, even after adjusting for the median, an increase in the right-tail is observed during one or both of the dotcom bubble and credit crunch periods for the remaining strategies. An extreme case is the Fund of Funds strategy which is fat-tailed throughout the whole sample even after normalisation. Therefore, although the relative dispersion of volatility remains more or less stable for almost all strategies, for some strategies it is more likely to observe extremely high volatilities, even after adjusting for fluctuations in the median.

To conclude, empirical results show that the volatility behaviour of funds exhibits variation, both within and between strategies. Some strategies, such as Multi-Process and Fixed Income generally tend to have lower volatility. Moreover, even within the same strategy, funds are characterised by different levels of volatility. The analysis of the volatility sample distribution reveals that for almost all strategies, volatility distribution exhibits large right tails, which tend to become larger during the dotcom bubble and credit crunch episodes. Nevertheless, normalised quantiles reveal that when adjusted for the median volatility, quantiles become more stable and behave uniformly across all strategies. Interestingly, while for the Macro, Fund of Funds and CTA strategies the right tail becomes heavier during economic downturns, the 90% quantile remains relatively stable. This suggests that, while higher levels of volatility were not necessarily more probable, "bad surprises" were more likely to happen.

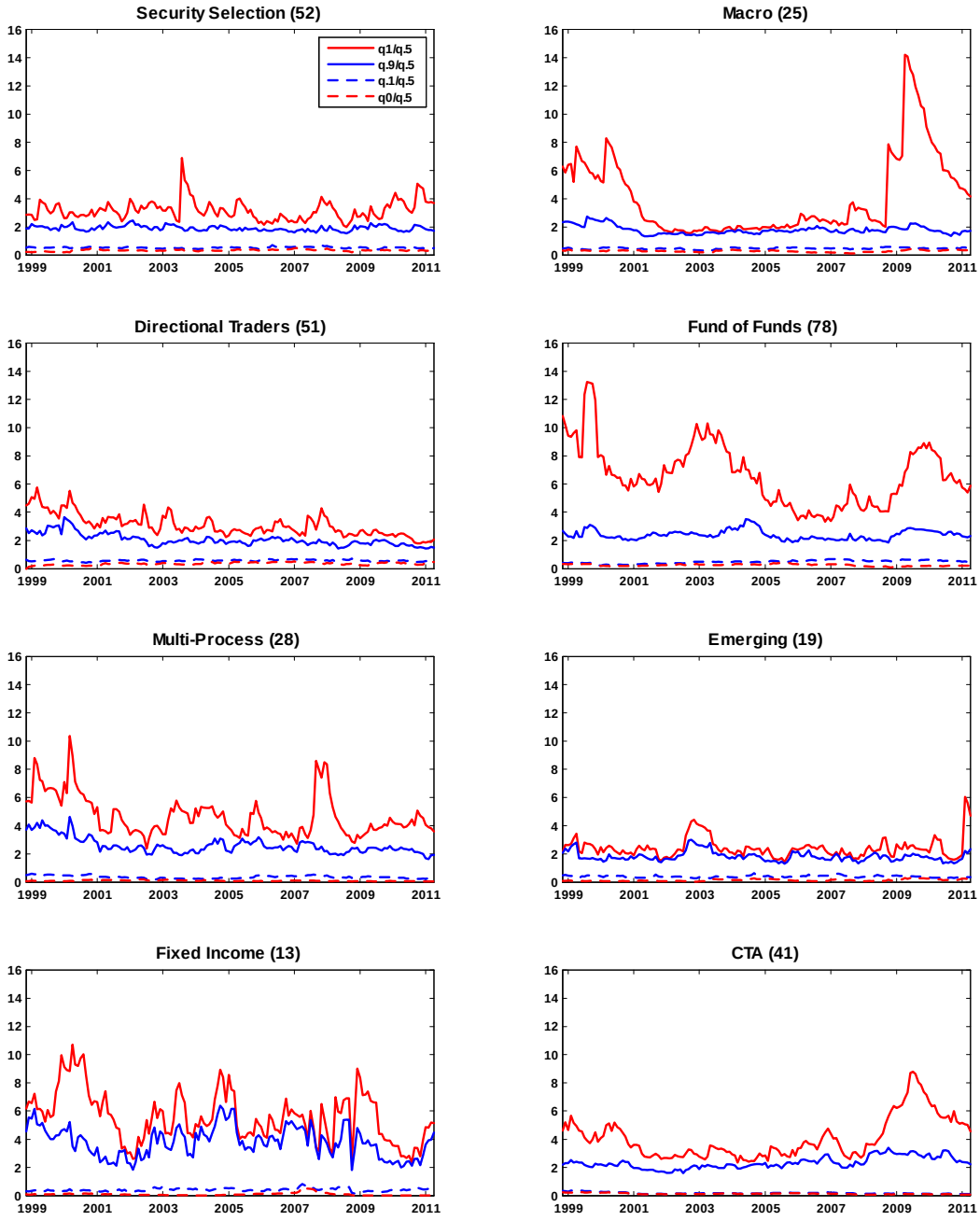

Figure 4.16: Plots of $0\%, 10\%$, 90% and 100% quantiles (normalised by the median) of the sample distribution of volatility across funds. Based on fitted conditional volatilities displayed in Figure 4.14. Number of funds in each strategy is given in parentheses.

4.5 Conclusion

This chapter combined the panel estimation and first-order bias reduction methods studied in Chapters 2 and 3 in order to conduct volatility modelling in small samples. The performance of this approach has been analysed using simulation and empirical analyses. Simulations indicate that the proposed approach can successfully reduce a substantial portion of the incidental parameter bias with 150-200 time series observations. In addition it is demonstrated that the integrated likelihood function offers a natural way of incorporating initial value selection into the bias-reduction mechanism. Simulation results clearly show that this alternative estimation approach improves the performance of the integrated likelihood method even further and is better able to match the sample distribution of the infeasible estimator. Interestingly, the superior bias performance is accompanied by higher sample standard deviation. This is contrary to the fact that bias is of a smaller order than variance, which implies that bias reduction should not affect variance. In fact, in line with this reasoning, the standard bias-reduction algorithm leads to a decrease in variance. Clearly, something more complex is taking place in the background. This poses an intriguing further research question. However, without doubt, the satisfactory performance of the bias reduction approach in small samples is in contrast with the ineffectiveness of standard GARCH tools in such samples.

In the empirical analysis, hedge fund volatility characteristics have been analysed by focusing on groups of funds following different investment strategies. The analysis reveals that the sample distribution of hedge fund volatility is generally asymmetric, skewed to right, and varies across time and in response to major economic events such as the burst of the dotcom bubble and the credit crunch. This empirical analysis is another novel contribution, as such analysis has hitherto been impossible due to hedge fund data being available at monthly frequency. Moreover, in a test of predictive ability using stock volatility forecasts, the proposed estimation method achieved superior forecasting performance compared to its alternatives. Interestingly, the alternative bias-correction method which is based on integrating both the initial value and the individual-specific parameter out, performed distinctively better compared to all other methods including the standard bias-reduced estimator.

This chapter finishes the research project of modelling volatility in small samples. The

major contribution of this chapter has been to propose a novel approach for volatility modelling in samples that are too small for standard GARCH tools to work. This potentially makes GARCH modelling available to a wide range of new datasets.

Chapter 5

Conclusion

This thesis has investigated volatility estimation using panels of financial data and bias-reduction in non-linear and dynamic panel data models in the presence of cross-section dependence. The main contributions are as follows:

1. A novel estimation method to model univariate GARCH volatilities using panels of financial data has been proposed and its large sample theory has been developed. Simulation results and empirical analysis reveal that this method has good in- and out-of-sample properties. In particular, due to pooling both time-series and cross-sectional information, it is able to fit volatility accurately using a smaller number of time-series observations compared to the standard estimation method. It has also been illustrated by simulations that this structure suffers from the incidental parameter issue.

2. Correction of the incidental parameter bias in non-linear and dynamic panel data models have been investigated in the presence of both time-series and cross-section dependence. The latter type of dependence has not yet been considered in the bias reduction literature. Therefore, extension towards this direction is the main theoretical contribution to the panel data literature. Theoretical analysis reveals that time-series dependence leads to an extra incidental bias term, which is negligible. Cross-section dependence, on the other hand, leads to an additional bias term, which is not related to the incidental parameter issue. Instead, it is due to cross-section dependence and its severity depends on the strength of this dependence. Likelihood-based expressions for the both bias terms are derived. In addition, the relevance of the analysis has

been illustrated in a spatial dependence/clustering based setting. This analysis has a general scope, in the sense that it is based on a generic likelihood function, rather than a specific model or application. Therefore, the results are applicable to a potentially large set of models and applications. In addition, some of the higher order asymptotic expansions derived in this study can be used in different applications.

3. Finally, the two research projects are combined in order to model volatility in small sample using the bias-corrected GARCH panel estimators. This approach has been shown to have good small sample properties. In addition, a forecasting exercise indicates that it also possesses superior predictive ability. This suggests that GARCH modelling is now available for a potentially large collection of datasets that are not long enough for standard GARCH tools to work. Finally, an empirical analysis of monthly hedge fund has been conducted. Due to the typical shortness of hedge fund data, GARCH-based volatility modelling has hitherto been impossible. Therefore, this empirical analysis is another novel contribution of this study. From a wider perspective, this study attempted to build a bridge between the panel data and financial econometrics literatures.

In the remainder, several possible extensions and future research ideas will be outlined. One possibility is to extend the GARCH panel approach to other GARCH-type model such as the exponential GARCH (EGARCH) model. Unlike GARCH, this model allows for asymmetric effects, where positive and negative shocks of the same magnitude can have different effects on volatility. A related idea is to extend this approach to volatility models that use high-frequency data. In a recent paper, Shephard and Sheppard (2010) suggest a realised variance based volatility model which closely resembles the standard GARCH model, is easy to estimate and delivers good results. Given the popularity of high-frequency data in financial econometrics, this would be a very promising research direction.

For analytical bias-reduction, most of the immediate extensions pertain to the cross-section dependence aspect. An obvious idea is to model dependence using specific dependence structures. The advantage of this approach is that it will make it possible to derive exact convergence rate. This is much more preferable than assuming flexible convergence rates, as was done in this study. Conceptualising dependence in terms of flexible conver-

gence rates is very convenient for theoretical analysis, but if one aims to remove the small sample bias, then exact knowledge of its order of magnitude has to be known. Otherwise, using an incorrectly normalised correction term would introduce extra bias into the system. Two possible options are the factor modelling and spatial dependence approaches. The latter has already been considered in a small exercise which yields interesting results. A third option for conceptualising cross-section dependence could be Copula modelling. Another extension is to consider simple and tractable models. This would make it easier to derive exact closed-form bias expressions. The virtue of using closed-form bias expressions is that they are easier to calculate (as opposed to numerically evaluated terms) and the literature generally agrees that bias correction delivers highly accurate results when closed-form bias expressions are available. This would also facilitate the modelling of cross-section dependence and so, these two extension ideas can form the basis of a major research project.

Another possible avenue is a theoretical analysis of whether the GARCH Panel model satisfies the high level assumption derived in Chapter 3 for analytical bias reduction. Simulation results clearly demonstrate that the bias-reduction approach works. Moreover, recent research shows that , under regularity conditions, GARCH variances and squared errors (σ_{it}^2 and ε_{it}^2) are β-mixing, which is a strong type of mixing implying α-mixing (e.g. Boussama (1998) and Carrasco and Chen (2002)). Initially, one might reason that the GARCH likelihood is a function of these two variables and so the GARCH likelihood is also mixing. However, the stationary solution for σ_{it}^2 depends on the infinite past of the GARCH process. Therefore, as the mixing property only carries to functions of a finite sequence of mixing random variables, mixing-type arguments are not immediately obvious. It would be an interesting project to further analyse this and eventually derive low level assumptions for bias reduction in GARCH panels.

It was illustrated by simulations that the selection of initial values has a substantial effect on the performance of bias-corrected estimators. When initial value selection is incorporated into bias-reduction, the bias performance improves significantly; however, at the same time, the variance increases. This is surprising, given that bias-reduction is not supposed to lead to higher variance. Clearly, selection of the initial value interacts with higher order variance terms, somehow. This would be very intriguing to further analyse. In small samples, the initial value is likely to have a non-negligible effect on estimates and forecasts. Therefore,

research in this direction would appeal to an applied audience.

Finally, given the lack of empirical analysis in the bias reduction literature, more empirical analysis is needed. Given that the bias-reduction approach has a general scope, these methods can be used in many areas including labour economics, financial economics, economic growth and, as demonstrated here, financial econometrics. Certainly, a large catalogue of simulation and empirical studies is necessary in order to attain a better grasp of the practical issues and advantages of analytical bias reduction.

References

Agarwal, V., N. D. Daniel, and N. Y. Naik (2011): "Do Hedge Funds Manage Their Reported Returns?," *Review of Financial Studies*, 24, 3281–3320.

Aielli, G. P., and M. Caporin (2011): "Variance Clustering Improved Dynamic Conditional Correlation MGARCH Estimators," working paper.

Andersen, E. B. (1970): "Properties of Conditional Maximum-Likelihood Estimators," *Journal of the Royal Statistical Society, Series B*, 32, 283–301.

Andersen, T., and L. Benzoni (2009): "Realized Volatility," in *Handbook of Financial Time Series*, ed. by T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch, pp. 555–575. Springer-Verlag.

Andersen, T. G., and T. Bollerslev (1998): "Deutsche Mark-Dollar Volatility: Iintraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies," *Journal of Finance*, 53, 219–265.

Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold (2001): "Volatility and Correlation Forecasting," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. W. J. Granger, and A. Timmermann, pp. 777–878. North-Holland.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001): "The Distribution of Exchange Rate Volatility," *Journal of the American Statistical Association*, 96, 42–55.

——— (2003): "Modeling and Forecasting Realized Volatility," *Econometrica*, 71, 579–625.

Andersen, T. G., T. Bollerslev, and S. Lange (1999): "Forecasting Financial Market Volatility: Sample Frequency vis-á-vis Forecast Horizon," *Journal of Empirical Finance*, 6, 457–477.

Andersen, T. G., T. Bollerslev, and N. Meddahi (2004): "Analytic Evaluation of Volatility Forecasts," *International Economic Review*, 45, 1079–1110.

Andersen, T. G., R. A. Davis, J.-P. Kreiss, and T. Mikosch (eds.) (2009): *Handbook of Financial Time Series*. Springer.

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Andrews, D. W. K., and J. C. Monahan (1992): "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953–966.

Arellano, M. (1987): "Computing Robust Standard Errors for Within-Group Estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431–434.

——— (2003): "Discrete Choices with Panel Data," *Investigaciones Economicas*, 27, 423–458.

Arellano, M., and S. Bonhomme (2009): "Robust Priors in Nonlinear Panel Data Models," *Econometrica*, 77, 489–536.

——— (2011): "Nonlinear Panel Data Analysis," *Annual Review of Economics*, 3, 395–424.

Arellano, M., and J. Hahn (2006): "A Likelihood-Based Approximate Solution To The Incidental Parameter Problem In Dynamic Nonlinear Models With Multiple Effects," working paper.

——— (2007): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress - Volume III*, ed. by R. Blundell, W. Newey, and T. Persson, pp. 381–409. Cambridge University Press.

Arellano, M., and B. Honore (2001): "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. Leamer, pp. 3229–3296. North-Holland, Amsterdam.

Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171.

——— (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77, 1229–1279.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

——— (2004): "A Panic Attack on Unit Roots and Cointegration," *Econometrica*, 72, 1127–1177.

——— (2006): "Evaluating Latent and Observed Factors in Macroeconomics and Finance," *Journal of Econometrics*, 131, 507–537.

Barigozzi, M., C. T. Brownlees, G. M. Gallo, and D. Veredas (2010): "Disentangling Systemic and Idiosyncratic Volatility for Large Panels of Assets: a Seminonparametric Vector Multiplicative Error Model," working paper.

Barndorff-Nielsen, O. E. (1983): "On a Formula for the Distribution of the Maximum Likelihood Estimator," *Biometrika*, 70, 343–65.

Barndorff-Nielsen, O. E., and D. R. Cox (1994): *Inference and Asymptotics*. Chapman & Hall, London.

Barndorff-Nielsen, O. E., and N. Shephard (2002): "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society, Series B*, 64, 253–280.

——— (2004): "Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics," *Econometrica*, 72, 885–925.

Bauwens, L., S. Laurent, and J. V. K. Rombouts (2006): "Multivariate GARCH Models: A Survey," *Journal of Applied Econometrics*, 21, 79–109.

Bauwens, L., and J. V. K. Rombouts (2007): "Bayesian Clustering of Many GARCH Models," *Econometric Reviews*, 26, 365–386.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates," *Quarterly Journal of Economics*, 119, 249–275.

Besag, J. (1974): "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Bester, A. C., T. G. Conley, and C. B. Hansen (2011): "Inference with Dependent Data Using Cluster Covariance Estimators," *Journal of Econometrics*, 165, 137–151.

Bester, C. A., and C. Hansen (2009): "A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects," *Journal of Business and Economic Statistics*, 27, 131–148.

Billingsley, P. (1995): *Probability and Measure*. Wiley, New York, 3 edn.

Bollen, N. P. B., and R. E. Whaley (2009): "Hedge Fund Risk Dynamics: Implications for Performance Appraisal," *Journal of Finance*, 64, 985–1035.

Bollerslev, T. (1986): "Generalised Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 51, 307–327.

——— (1990): "Modelling the Coherence in Short-Run Nominal Exchange Rates: a Multivariate Generalized ARCH Approach," *Review of Economics and Statistics*, 72, 498–505.

Bollerslev, T., R. Y. Chou, and K. F. Kroner (1992): "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence," *Journal of Econometrics*, 52, 5–59.

Bollerslev, T., R. F. Engle, and D. B. Nelson (1994): "ARCH Models," in *The Handbook of Econometrics*, ed. by R. F. Engle, and D. McFadden, pp. 2959–3038. North-Holland.

Bollerslev, T., R. F. Engle, and J. M. Wooldridge (1988): "A Capital Asset Pricing Model with Time Varying Covariances," *Journal of Political Economy*, 96, 116–131.

Bollerslev, T., and J. M. Wooldridge (1992): "Quasi Maximum Likelihood Estimation and Inference in Dynamic Models with Time Varying Covariances," *Econometric Reviews*, 11, 143–172.

Bolthausen, E. (1982): "On the Central Limit Theorem for Stationary Mixing Random Fields," *The Annals of Probability*, 10, 1047–1050.

Boussama, F. (1998): "Ergodicite, Melange et Estimation dans les Modeles GARCH," PhD dissertation, Paris-7 University.

Bradley, R. C. (2005): "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions," *Probability Surveys*, 2, 107–144.

Brownlees, C. (2010): "On the Relation Between Firm Characteristics and Volatility Dynamics with an Application to the 2007-2009 Financial Crisis," working paper.

Cappiello, L., R. F. Engle, and K. K. Sheppard (2006): "Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns," *Journal of Financial Econometrics*, 4, 537–572.

Carrasco, M., and X. Chen (2002): "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," *Econometric Theory*, 18, 17–39.

Carro, J. M. (2007): "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects," *Journal of Econometrics*, 140, 503–528.

Caselli, F., G. Esquivel, and F. Lefort (1996): "Reopening the Convergence Debate: A New Look at Cross-Country Growth Empirics," *Journal of Economic Growth*, 1, 363–389.

Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, 225–38.

——— (1984): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intrilligator, pp. 1248–1318. North-Holland, Amsterdam.

Chao, J. C., V. Corradi, and N. R. Swanson (2001): "An Out-Of-Sample Test for Granger Causality," *Macroeconomic Dynamics*, 5, 598–620.

Chudik, A., M. H. Pesaran, and E. Tosetti (2011): "Weak and Strong Cross-Section Dependence and Estimation of Large Panels," *Econometrics Journal*, 14, C45–C90.

Clark, T. E., and M. W. McCracken (2001): "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85–110.

Clements, M. P. (2005): *Evaluating Econometric Forecasts of Economic and Financial Variables*. Palgrave MacMillan.

Conley, T. G. (1999): "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 92, 1–45.

Corradi, V., and N. R. Swanson (2006): "Predictive Density Evaluation," in *Handbook of Economic Forecasting*, ed. by G. Elliot, C. W. J. Granger, and A. Timmermann, pp. 197–284. North-Holland.

Corradi, V., N. R. Swanson, and C. Olivetti (2001): "Predictive Ability with Cointegrated Variables," *Journal of Econometrics*, 104, 315–358.

Cox, D. R., and N. Reid (1987): "Parameter Orthogonality and Approximate Conditional Inference (with discussion)," *Journal of the Royal Statistical Society, Series B*, 49, 1–39.

——— (2004): "A Note on Pseudolikelihood Constructed from Marginal Densities," *Biometrika*, 91, 729–737.

Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, Oxford.

Davison, A. C. (2003): *Statistical Models*. Cambridge University Press, Cambridge.

den Haan, W. J., and A. Levin (1996): "Inferences from Parametric and Non-Parametric Covariance Matrix Estimation Procedures," NBER working paper.

Dhaene, G., and K. Jochmans (2010): "Split-Panel Jackknife Estimation of Fixed-Effects Models," working paper.

——— (2011): "An Adjusted Profile Likelihood for Non-Stationary Panel Data Models with Fixed Effects," working paper.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263.

Ding, Z., R. F. Engle, and C. W. J. Granger (1993): "A Long Memory Property of Stoc Market Returns and a New Model," *Journal of Empirical Finance*, 1, 83–106.

Doukhan, P. (1994): *Mixing, Properties and Examples*. Springer.

Enders, W. (2010): *Applied Econometric Time Series*. John Wiley and Sons, 3 edn.

Engle, R. F. (1982): "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of the United Kingdom Inflation," *Econometrica*, 50, 987–1007.

——— (1995): *ARCH: Selected Readings*. Oxford University Press, Oxford.

——— (2002): "Dynamic Conditional Correlation - A Simple Class of Multivariate GARCH Models," *Journal of Business and Economic Statistics*, 20, 339–350.

——— (2009): "High Dimensional Dynamic Correlations," in *The Methodology and Practice of Econometrics: Papers in Honour of David F Hendry*, ed. by J. L. Castle, and N. Shephard, pp. 122–148. Oxford University Press.

Engle, R. F., and K. F. Kroner (1995): "Multivariate Simultaneous Generalized ARCH," *Econometric Theory*, 11, 122–150.

Engle, R. F., and J. Mezrich (1996): "GARCH for Groups," *Risk*, 9, 36–40.

Engle, R. F., N. Shephard, and K. K. Sheppard (2008): "Fitting Vast Dimensional Time-Varying Covariance Models," working paper.

Engle, R. F., and K. K. Sheppard (2001): "Theoretical and Empirical Properties of Dynamic Conditional Correlation Multivariate GARCH," working paper.

Erdélyi, A. (1956): *Asymptotic Expansions*. Dover Publications, New York.

Fernández-Val, I. (2009): "Fixed Effects Estimation of Structural Parameters and Marginal Effects in Panel Probit Models," *Journal of Econometrics*, 150, 71–85.

Firth, D. (1993): "Bias Reduction of Maximum Likelihood Estimates," *Biometrika*, 80, 27–38.

Francq, C., L. Horváth, and J. M. Zakoïan (2011): "Merits and Drawbacks of Variance Targeting in GARCH Models," *Journal of Financial Econometrics*, 9, 619–656.

Francq, C., and J.-M. Zakoïan (2004): "Maximum Likelihood Estimation of Pure GARCH and ARMA-GARCH Processes," *Bernoulli*, 10, 605–637.

Francq, C., and J.-M. Zakoïan (2010): *GARCH Models: Structure, Statistical Inference and Financial Applications*. Wiley.

Fung, D. A., and W. Hsieh (2000): "Performance Characteristics of Hedge Funds and Commodity Funds: Natural vs. Spurious Biases," *Journal of Financial and Quantitative Analysis*, 35, 291–307.

——— (2004): "Hedge Fund Benchmarks: A Risk Based Approach," *Financial Analyst Journal*, 60, 65–80.

Getmansky, M., A. W. Lo, and I. Makarov (2004): "An Econometric Model of Serial Correlation and Illiquidity in Hedge Fund Returns," *Journal of Financial Economics*, 74, 529–610.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578.

Glosten, L. R., R. Jagannathan, and D. Runkle (1993): "Relationship between the Expected Value and the Volatility of the Nominal Excess Return on Stocks," *Journal of Finance*, 48, 1779–1802.

Gourieroux, C., and J. Jasiak (2001): *Financial Econometrics: Problems, Models, and Methods*. Princeton University Press.

Gourieroux, C., A. Monfort, and A. Trognon (1984): "Pseudo-Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681–700.

Hahn, J., and G. Kuersteiner (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects when both n and T are Large," *Econometrica*, 70, 1639–1657.

——— (2011): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," *Econometric Theory*, 27, 1152–1191.

Hahn, J., and H. R. Moon (2006): "Reducing Bias of MLE in a Dynamic Panel Model," *Econometric Theory*, 22, 499–512.

Hahn, J., and W. K. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72(4), 1295–1319.

Hall, A. R. (2005): *Generalized Method of Moments*. Oxford University Press, Oxford.

Hall, P. G. (1992): *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hansen, B. E. (1994): "Autoregressive Conditional Density Estimation," *International Economic Review*, 35, 705–730.

Hansen, C. B. (2007): "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When t is Large," *Journal of Econometrics*, 141, 597–620.

Hansen, P. R., and A. Lunde (2005): "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?," *Journal of Applied Econometrics*, 20, 873–889.

——— (2006): "Realized Variance and Market Microstructure Noise (with Discussion)," *Journal of Business and Economic Statistics*, 24, 127–218.

Heber, G., A. Lunde, N. Shephard, and K. K. Sheppard (2009): *Oxford Man Institute's Realized Library*. Oxford-Man Institute: University of Oxford, Version 0.1.

Honoré, B. E. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Models with Fixed Effects," *Econometrica*, 60, 533–565.

Honoré, B. E., and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 95, 839–874.

Horowitz, J. L., and S. Lee (2004): "Semiparametric Estimation of a Panel Data Proportional Hazard Model with Fixed Effects," *Journal of Econometrics*, 119, 155–198.

Hospido, L. (2010): "Modelling Heterogeneity and Dynamics in the Volatility of Individual Wages," forthcoming.

Huggler, B. (2004): "Modelling Hedge Fund Returns," University of Zurich masters thesis.

Ibragimov, I. A., and Y. V. Linnik (1971): *Independent and Stationary Random Variables*. Wolters-Noordhoff.

Islam, N. (1995): "Growth Empirics: A Panel Data Approach," *Quarterly Journal of Economics*, 110, 1127–1170.

Jenish, N., and I. R. Prucha (2009): "Central Limit Theorems and Uniform Laws of Large Numbers for Arrays of Random Fields," *Journal of Econometrics*, 150, 86–98.

——— (2010): "On Spatial Processes and Asymptotic Inference under Near-Epoch Dependence," working paper.

Kapetanios, G., M. H. Pesaran, and T. Yamagata (2011): "Panels with Non-Stationary Multifactor Error Structures," *Journal of Econometrics*, 160, 326–348.

Kawakatsu, H. (2006): "Matrix Exponential GARCH," *Journal of Econometrics*, 134, 95–128.

Kelejian, H. H., and I. R. Prucha (2007): "HAC Estimation in a Spatial Framework," *Journal of Econometrics*, 140, 131–154.

Kristensen, D., and B. Salanié (2010): "Higher Order Improvements for Approximate Estimators," working papers.

Kroner, K., and V. Ng (1998): "Modeling Asymmetric Comovements of Assets Returns," *Review of Financial Studies*, 11, 817–844.

Lancaster, T. (2000): "The Incidental Parameter Problem since 1948," *Journal of Econometrics*, 95, 391–413.

Lee, L.-F. (2004): "Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Econometric Models," *Econometrica*, 72, 1899–1926.

——— (2007): "GMM and 2SLS Estimation of Mixed Regressive, Spatial Autoregressive Models," *Journal of Econometrics*, 137, 489–514.

Li, H., B. G. Lindsay, and R. P. Waterman (2003): "Efficiency of Projected Score Methods in Rectangular Array Asymptotics," *Journal of the Royal Statistical Society, Series B*, 65, 191–208.

Liang, K. Y., and S. L. Zeger (1986): "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lindsay, B. G. (1988): "Composite Likelihood Methods," in *Statistical Inference from Stochastic Processes*, ed. by N. U. Prabhu, pp. 221–239. Amercian Mathematical Society, Providence, RI.

Magnus, J. R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, New York.

Manova, K., and Z. Zhang (2012): "Export Prices across Firms and Destinations," *Quarterly Journal of Economics*, 127, 379–436.

McCracken, M. W. (2000): "Robust Out-Of-Sample Inference," *Journal of Econometrics*, 99, 195–223.

McCullagh, P. (1984): "Tensor Notation and Cumulants of Polynomials," *Biometrika*, 71, 461–476.

——— (1987): *Tensor Methods in Statistics*. Chapman & Hall, London.

McCullagh, P., and R. Tibshirani (1990): "A Simple Method for the Adjustment of Profile Likelihoods," *Journal of the Royal Statistical Society. Series B (Methodological)*, 52, 325–344.

Meddahi, N. (2002): "A Theoretical Comparison between Integrated and Realized Volatilities," *Journal of Applied Econometrics*, 17, 479–508.

Moon, H. R., and B. Perron (2004): "Testing for a Unit Root in Panels with Dynamic Factors," *Journal of Econometrics*, 122, 81–126.

Nelson, D. B. (1991): "Conditional Heteroskedasticity in Asset Pricing: A New Approach," *Econometrica*, 59, 347–370.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *The Handbook of Econometrics*, ed. by R. F. Engle, and D. McFadden, pp. 2111–2245. North-Holland.

Newey, W. K., and K. D. West (1987): "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

——— (1994): "Automatic Lag Selection in Covariance Matrix Estimation," *The Review of Economic Studies*, 61, 631–653.

Neyman, J., and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–16.

Nickell, S. J. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49, 1417–1426.

Noureldin, D., N. Shephard, and K. K. Sheppard (2011): "Multivariate High-Frequency-Based Volatility (HEAVY) Models," *Journal of Applied Econometrics*, forthcoming.

Pace, L., and A. Salvan (1997): *Principles of Statistical Inference from a Neo-Fisherian Perspective*. World Scientific, Singapore.

——— (2006): "Adjustments of the Profile Likelihood from a New Perspective," *Journal of Statistical Planning and Inference*, 136, 3554–3564.

Patton, A. J. (2011): "Volatility Forecast Comparison using Imperfect Volatility Proxies," *Journal of Econometrics*, 160, 246–256.

Patton, A. J., and T. Ramadorai (2011): "On the High Frequency Dynamics of Hedge Fund Risk Exposures," working paper.

Patton, A. J., and K. K. Sheppard (2009): "Good Volatility, Bad Volatility: Signed Jumps and Persistence of Volatility," Unpublished paper: Oxford-Man Institute, University of Oxford.

Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74, 967–1012.

Pesaran, M. H., and E. Tosetti (2011): "Large Panels with Common Factors and Spatial Correlation," *Journal of Econometrics*, 161, 182–202.

Phillips, P. C. B., and D. Sul (2003): "Dynamic Panel Estimation and Homogeneity Testing under Cross Section Dependence," *Econometrics Journal*, 6, 217–259.

——— (2007): "Bias in Dynamic Panel Estimation with Fixed Effects, Incidental Trends and Cross Section Dependence," *Journal of Econometrics*, 137, 162–188.

Poon, S.-H., and C. W. J. Granger (2003): "Forecasting Volatility in Financial Markets: A Review," *Journal of Economic Literature*, 41, 478–539.

Ramadorai, T. (2011): "Capacity Constraints, Investor Information, and Hedge Fund Returns," *Journal of Financial Economics*, forthcoming.

Ramadorai, T., and M. Streatfield (2011): "Money for Nothing? Understanding Variation in Reported Hedge Fund Fees," working paper.

Sartori, N. (2003): "Modified Profile Likelihoods in Models with Stratum Nuisance Parameters," *Biometrika*, 90, 533–549.

Severini, T. A. (1999): "On the Relationship between Bayesian and Non-Bayesian Elimination of Nuisance Parameters," *Statistica Sinica*, 9, 713–724.

——— (2000): *Likelihood Methods in Statistics*. Oxford University Press, New York.

——— (2005): *Elements of Distribution Theory*. Cambridge University Press, New York.

——— (2007): "Integrated Likelihood Functions for Non-Bayesian Inference," *Biometrika*, 94, 529–542.

——— (2010): "Likelihood Ratio Statistics Based on An Integrated Likelihood," *Biometrika*, 97, 481–496.

Severini, T. A., and W. H. Wong (1992): "Profile Likelihood and Conditionally Parametric Models," *Annals of Statistics*, 20, 1768–1802.

Shephard, N., and K. K. Sheppard (2010): "Realising the Future: Forecasting with High-Frequency-Based Volatility (HEAVY) Models," *Journal of Applied Econometrics*, 25, 197–231.

Sheppard, K. K. (2010): "Financial Econometrics Notes," University of Oxford Lecture Notes.

Silvennoinen, A., and T. Teräsvirta (2009): "Multivariate GARCH Models," in *Handbook of Financial Time Series*, ed. by T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch, pp. 201–229. Springer-Verlag.

Stein, C. (1956): "Efficient Nonparametric Testing and Estimation," in *Proceedings of the Third Berkeley Symposium on Mathematical and Statistical Probability*, vol. 1, pp. 187–195. University of California Press, Berkeley.

Taylor, S. J. (2005): *Asset Price Dynamics, Volatility, and Prediction*. Princeton University Press.

Teo, M. (2009): "Geography of Hedge Funds," *Review of Financial Studies*, 22, 3531–3561.

Tierney, L., R. E. Kass, and J. B. Kadane (1989): "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, 84, 710–716.

Tsay, R. (2002): *Analysis of Financial Time Series*. Wiley, New York, 2 edn.

Varin, C. (2008): "On Composite Marginal Likelihoods," *Advances in Statistical Analysis*, 92, 1–28.

Varin, C., N. Reid, and D. Firth (2011): "An Overview of Composite Likelihood Methods," *Statistica Sinica*, 21, 5–42.

Varin, C., and P. Vidoni (2005): "A Note on Composite Likelihood Inference and Model Selection," *Biometrika*, 92, 519–528.

Wansbeek, T., and E. Meijer (2000): *Measurment Error and Latent Variables in Econometrics*. North-Holland.

Waterman, R. P., and B. G. Lindsay (1996): "Projected Score Methods for Approximating Conditional Scores," *Biometrika*, 83, 1–13.

West, K. D. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067–1084.

——— (2006): "Forecast Evaluation," in *Handbook of Economic Forecasting*, ed. by G. Elliot, C. W. J. Granger, and A. Timmermann, pp. 99–134. North-Holland.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

——— (1994): *Estimation, Inference and Specification Analysis*. Cambridge University Press, Cambridge.

——— (2000): "A Reality Check on Data Snooping," *Econometrica*, 68, 1097–1126.

——— (2001): *Asymptotic Theory for Econometricians*. Academic Press, Orlando, 2 edn.

Wooldridge, J. M. (2003): "Cluster-Sample Methods in Applied Econometrics," *American Economic Review*, 93, 133–188.

——— (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2 edn.

Woutersen, T. (2002): "Robustness against Incidental Parameters," working paper.

Zakoïan, J.-M. (1994): "Threshold Heteroskedastic Models," *Journal of Economic Dynamics and Control*, 18, 931–55.