Bias Reduction in Nonlinear and Dynamic Panels in the Presence of Cross-Section Dependence, with a GARCH Panel Application[*]

Cavit Pakel[†]
Department of Economics
Bilkent University

Abstract

In nonlinear dynamic panels where the time-series dimension, T, is small relative to the cross-section dimension, N, fixed effect models are subject to the incidental parameter bias. Considering a general setting where dependence across both T and N is allowed, I use the integrated likelihood method to characterise this bias and obtain bias-reduced estimators. Under large-T, large-N asymptotics, I show that time-series dependence leads to an extra incidental parameter bias term, which is not present in the iid case. Moreover, due to cross-section dependence, a second type of bias emerges, the magnitude of which depends on the level of dependence. Likelihood-based analytical expressions are provided for both terms. Next, the particular case of spatial dependence with clustered individuals is considered. It turns out that, under certain conditions, the bias due to cross-section dependence is negligible in this setting. I then utilise these results to fit GARCH models using a panel structure. Monte Carlo analysis reveals that the proposed method successfully fits GARCH with little bias and no increase in variance using 150-200 time-series observations, compared to around 1,000-1,500 observations required for successful GARCH estimation by standard methods. Simulation results further indicate that the effect of cross-section dependence on bias is negligible, although it leads to higher estimator variance. Finally, I consider two empirical illustrations; an analysis of monthly hedge fund volatility characteristics and a test of predictive ability using daily stock volatility forecasts.

[*] This paper has previously been circulated under the title "Bias Reduction under Dependence, in a Nonlinear and Dynamic Panel Setting: the Case of GARCH Panels" and is based on two chapters of my DPhil thesis at the University of Oxford. I would like to thank Neil Shephard for his support and guidance. I would also like to express my gratitude to Manuel Arellano, Stéphane Bonhomme, Fitnat Banu Demir, Shin Kanaya, Kasper Lund-Jensen, Bent Nielsen, Andrew Patton, Anders Rahbek, Enrique Sentana, Kevin Sheppard and Michael Streatfield for stimulating discussions. I also thank seminar participants at CEMFI, Oxford-Man Institute, 2011 Spring Meeting of Young Economists in the University of Groningen, and University of Oxford. The hedge fund dataset used in this paper has been consolidated by Michael Streatfield and Sushant Vale. Part of this work has been undertaken during my visit to CEMFI and their unbounded hospitality is greatly acknowledged. All errors are mine.

[†] **Contact address**: Department of Economics, Bilkent University, 06800 Ankara, Turkey **Email address**: `cavit.pakel@bilkent.edu.tr`.

1 Introduction

A substantial body of research in econometrics has been dedicated to controlling for unobserved individual 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 nonlinear 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-section observations, i.e. fixed-T large-N asymptotics. 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 parameters to be estimated in a maximum likelihood framework.[2] However, this approach is known to be subject to the incidental parameter issue, first studied by Neyman and Scott (1948) (see also the excellent survey by Lancaster (2000)). 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 $L_i(\theta, \lambda_i) = L_i(\theta, \lambda_i; y_i)$ be the likelihood function for the i^{th} individual ($i = 1, ..., N$) and $L(\theta, \lambda_1, ..., \lambda_N)$ be the joint likelihood. Here y_i is the data vector for the i^{th} individual, 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, $\hat{\lambda}_i(\theta)$ will be subject to estimation error. This estimation error will be inherited by the corresponding concentrated likelihood function, which will be incorrectly centred. Consequently, the resulting fixed-T, large-N estimator $\hat{\theta}_T = \arg\max_\theta p\lim_{N\to\infty} L(\theta, \hat{\lambda}_1(\theta), ..., \hat{\lambda}_N(\theta))$ 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.

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

[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).

of $1/T$ and removing the leading $O(1/T)$ bias term. In other words, for

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

if a consistent estimator of A, say $\hat{A} = A + o_p(1)$, exists, then $\tilde{\theta} = \hat{\theta} - \hat{A}/T$ will be a first-order unbiased estimator, since $\mathbb{E}[\tilde{\theta} - \theta_0] = o(1/T)$. For moderate T, the remaining $o(1/T)$ term is expected to be negligible. Based on this principle, the analytical bias-correction methods attack the first order bias of either (i) the estimator $\hat{\theta}$ (Hahn and Kuersteiner (2002, 2011), Hahn and Newey (2004), Hahn and Moon (2006), Fernández-Val (2009)); or (ii) the likelihood (or the objective) function (Arellano and Hahn (2006), Arellano and Bonhomme (2009), Bester and Hansen (2009) and Kristensen and Salanie (2010)); or (iii) the score (or the estimating) function (Woutersen (2002), Arellano (2003), Carro (2007), 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).

The aforementioned literature is in general based on the assumption of cross-section independence.[4] However, many interesting macroeconomic and financial panels will almost certainly violate this assumption. This can, for example, be due to a common shock process which hits all individuals simultaneously, but with different magnitudes, as in the case of factor models. The main contribution of this study is extension of bias reduction in nonlinear dynamic panels to the case of cross-section dependence.

In a recent study, Arellano and Bonhomme (2009) consider the integrated likelihood function as a unifying framework for likelihood-based estimation. This is given by

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

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 effect 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 this likelihood function; the *robust priors*. In this paper, I extend their analysis and study the bias properties of (1) under serial and cross-section dependence to obtain likelihood-based general characterisations of extra bias terms. The theoretical analysis reveals that,

[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. 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), Pace and Salvan (2006) and Dhaene and Jochmans (2010).

[4] For examples of studies outside this literature, where bias in the presence of a factor structure is analysed, see Phillips and Sul (2007) and Bai (2009, 2012).

time-series dependence leads to an extra $O(1/T^{3/2})$ incidental parameter bias term which is not present under serial independence. Then, without specifying an explicit structure for cross-section dependence, I consider a flexible structure where the speed of convergence is assumed to be $\sqrt{N^{\rho}T}$ where ρ varies between 0 and 1. Therefore, the two extremes are $\sqrt{NT}$-convergence (cross-section independence) and $\sqrt{T}$-convergence (cross-section dependence of strong type). Intuitively, the first polar case corresponds to independence or weak dependence across cross-section, while in the latter case dependence is so strong that cross-sectional variation does not contribute to convergence at all.[5] Hence, ρ measures the strength of cross-section dependence. In a recent study, Bailey, Kapetanios and Pesaran (2012) measure cross-section dependence in the same fashion. The theoretical analysis reveals that, depending on the strength of cross-section dependence, a second type of bias term, due to cross-section dependence, emerges. This extra bias is not related to the incidental parameter issue and so, it has to be corrected for separately. Based on these findings it is shown that, if the cross-section dimension is allowed to contribute to convergence at a mild rate, then the cross-section dependence induced bias becomes $O(1/T^{3/2})$. Consequently, the $O(1/T)$ bias of the estimator will be identical to the one characterised by Arellano and Bonhomme (2009) and their robust priors can be used for bias correction, despite the presence of cross-section dependence.

Next, I consider the particular setting of spatial dependence and clustered individuals. This part of the paper mainly aims to establish a connection between the analytical bias reduction and spatial dependence/clustering literatures. Unlike the usual clustering setting where independence between clusters is assumed, the spatial dependence framework allows for some weak dependence between clusters, thus generating richer interaction possibilities. Interestingly, it turns out that the extra bias due to cross-section dependence becomes $O(1/T)$ when the number of clusters and the number of members of each cluster increase at the rate $O(\sqrt{N})$.

It must be noted that this study is based on a pseudo-likelihood function, called the "composite-likelihood" function (see Lindsay (1988), Cox and Reid (2004) and Varin Reid and Firth (2011)). Estimation by maximum likelihood under cross-section dependence and time-series heteroskedasticity requires specification of an $(N \times N)$ covariance matrix at each t. This entails complications in both computation (inversion of a large dimensional matrix) and statistical modelling, even when N is modestly large. The composite-likelihood method is used here to side-step these issues, which is based on the idea of approximating the joint density by averaging univariate marginal densities. This is equivalent to treating data as if there were no cross-section dependence. Engle, Shephard and Sheppard (2008) show that under mild conditions this ensures consistency, possibly at some efficiency loss. More sophisticated methods for dealing with cross-section dependence, notably factor modelling[6], can be employed; but this is beyond the scope of this

[5] This second case is not very realistic but it nevertheless provides a useful tool for understanding the effects of dependence on small sample bias.

[6] For recent important examples of this literature, see, among others, Bai and Ng (2002, 2004), Phillips

study and not pursued here. In addition, intuition on the effect of cross-section dependence on bias can still be understood in a pseudo-likelihood setting. Importantly, results of this study also shed light on the properties of estimators under neglected cross-section dependence.

All theoretical results are given in terms of likelihood-based general expressions and are not model specific. Therefore, the bias characterisations and asymptotic expansions derived in this paper do potentially apply to a wide array of nonlinear dynamic panel models. The nonlinear dynamic panel application considered here is estimation of financial volatility in a panel (rather than the traditional time-series) setting. This is based on the Generalised Autoregressive Conditional Heteroskedasticity (GARCH) type models (Engle (1982) and Bollerslev (1986)). In general, consistent GARCH estimation by standard time-series approaches requires around 1,000-1,500 observations for the small sample bias to vanish. However, simulation analysis reveals that when a bias-corrected panel approach is employed, a substantial portion of the small-sample bias is removed with as little as 150-200 time-series observations. This is a significant improvement. As such, suggesting a new approach for volatility modelling in small samples is the main contribution of this paper to the field of financial econometrics. In line with the rest of the literature, bias reduction does *not* come at the cost of higher variance. In fact, variance is reduced. Simulation results further indicate that the effect of cross-section dependence on bias is negligible. However, compared to the case of cross-section independence, it leads to inflated estimator variance.

Finally, two empirical illustrations are considered. The first one is a comparison of out-of-sample predictive ability using stock market data, where the bias-corrected GARCH panel model attains superior forecasting performance in comparison to its alternatives. This is followed by an analysis of monthly hedge fund volatility. This type of data are a typical example of short panels, as fund returns are recorded at monthly frequency and observations are available for the last 18 years only. The results indicate that funds' volatility characteristics show variation both across and, more interestingly, within different investment strategies. Furthermore, sample distributions of volatility across funds are asymmetric, skewed to the right and react to major economic events, such as the credit crunch. To the best of my knowledge, this is the first example of GARCH-based hedge fund volatility modelling in the literature.

An indirect and appealing feature of modelling conditionally heteroskedastic errors in a panel framework is that it offers a mechanism to induce time-varying heterogeneity in panel data.[7] One possibility to control for time-varying common shocks is to assume year effects. However, without further modelling, this implies that all individuals are affected identically by the common shocks. The GARCH panel approach offers an alternative and

and Sul (2003), Pesaran (2006), Bai (2009), Chudik Pesaran and Tosetti (2011), Kapetanios Pesaran and Yamagata (2011) and Pesaran and Tosetti (2011).

[7] Fernández-Val and Vella (2009) list possible examples where both individual-specific (time-invariant) and time-varying heterogeneity is present and analyse bias-reduction under this setting.

flexible mechanism through which time-varying heteroskedasticity can be induced without making such assumptions. Of course, the number of observations required for GARCH estimation, even after bias-reduction, might be too large for some microeconometric datasets. However, this study makes an initial step towards a more flexible heterogeneity structure.

The rest of this study is organised as follows: Section 2 introduces the notation and briefly discusses relevant concepts. Key assumptions are listed and discussed in Section 3. The main theoretical results concerning the bias of the integrated likelihood are given in Sections 4 and 5. In Section 6, the specific case of spatial dependence for clustered individuals is considered. The Panel GARCH application is introduced in Section 7, where a detailed simulation analysis is provided to investigate the small sample properties and the bias-reduction performance of the integrated likelihood method. This is followed by two empirical applications in Section 8. Section 9 concludes. Proofs and additional discussions are given in the Appendix.

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:

$$(\hat{\theta}, \hat{\lambda}_1, ..., \hat{\lambda}_N) = \arg \max_{\theta, \lambda_1, ..., \lambda_n} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \varphi_{it}(\theta, \lambda_i). \tag{2}$$

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, (2) has a composite likelihood function interpretation, where the joint density is approximated by the average of marginal densities. Engle, Shephard and Sheppard (2008) consider a similar approach where they approximate the joint density by combining many bivariate marginal densities and their consistency results can easily be adopted to the framework of this study.

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

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

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,

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

The centred likelihood derivatives with respect to λ_i are defined as

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

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

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

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.[8] 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, $\hat{\theta}$

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

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

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

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.

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

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

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, $\pi_i(\lambda_i|\theta)$ 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.[9] 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,

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

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.

[9] 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 Assumptions

As outlined in Section 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 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. Formal definitions of these concepts are given in Definition A.1 in the Appendix. The mixing concept is commonly used to impose weak dependence on economic time-series.[10] A convenient property of mixing sequences is that for any measureable function $g(\cdot)$, if a sequence $(x_{it}, x_{i,t-1}, x_{i,t-2}, ...)$ is mixing, then $g(x_{it}, x_{i,t-1}, ..., x_{i,t-\tau})$ is also mixing of the same size (see Davidson (1994) and White (2001)).

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

Assumption 3.1 *$N, T \to \infty$ jointly and, for $0 < c < \infty$, $N/T \to c$.*

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

Assumption 3.4 *The support of $\pi_i(\lambda_i|\theta)$ contains an open neighbourhood of the true parameters λ_{i0} and θ_0.*

Assumption 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 3.6 *For each $\theta \in \Theta$, $\ell_{iT}(\theta, \lambda_i)$ has a unique maximum at $\hat{\lambda}_i(\theta)$ for all i.*

[10] 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 3.7 *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 3.8 *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 3.9 *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 3.10 *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 $k \in \{0, 1, 2, 3, 4\}$*

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

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

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

Assumption 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 3.1 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 3.2 characterises the structure of time-series dependence. As mentioned previously, 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, all derivatives of the objective function up to and including the eighth order are guaranteed to exist and to be continuous.[11] Consequently, Assumptions 3.2 and 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 3.10 which states that the necessary moment conditions hold. The proofs for the results given in Section 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 3.4 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 3.5 is a standard regularity condition on the parameter space. Assumption 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 3.7 controls the convergence rate of $\hat{\lambda}_i(\theta)$ to $\bar{\lambda}_{iT}(\theta)$, the benchmark "least-favourable" estimator. Intuitively, this ensures that the concentrated likelihood estimator is never "too" far away from the benchmark estimator. Assumption 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 3.9.

Assumption 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 3.11 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 3.1 *Cross-section independence, which, under standard regularity conditions, implies $\sqrt{NT}$-convergence.*

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

Case 3.2 *Cross-section dependence of strong type, such that there is no gain from cross-section size. Hence, $\sqrt{T}$-convergence.*

Case 3.1 is the setting considered invariably in the analytical bias reduction literature. Case 3.2, 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 $\rho_1 = \rho_2 = \rho_3 = 1$. Bailey, Kapetanios and Pesaran (2012) use a similar approach and define the "exponent of cross-sectional dependence" to measure the strength of cross-section dependence. The parameters ρ_i, $i = 1, 2, 3$ are closely related to the exponent of cross-sectional dependence.

4 Bias of the Integrated Likelihood in the Presence of Time-Series Dependence

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

Theorem 4.1 *Under Assumptions 3.1-3.8,*

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

where

$$\begin{aligned}\mathcal{B}_{iT}^{(1)}(\theta) &= \frac{1}{2}\{\mathbb{E}_{\theta_0,\lambda_{i0}}[-\ell_{iT}^{\lambda\lambda}]\}^{-1}\mathbb{E}_{\theta_0,\lambda_{i0}}[T(\ell_{iT}^{\lambda})^2] \\ &\quad -\frac{1}{2}\ln \mathbb{E}_{\theta_0,\lambda_{i0}}[-\ell_{iT}^{\lambda\lambda}] + \ln \pi_i(\bar{\lambda}_i|\theta), \qquad (4)\\ \mathcal{B}_{iT}^{(2)}(\theta) &= T^{3/2}\frac{1}{2}\frac{\mathbb{E}_{\theta_0,\lambda_{i0}}\left[V_{iT}^{\lambda\lambda}(\ell_{iT}^{\lambda})^2\right]}{\{\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda}]\}^2} - T^{3/2}\frac{1}{6}\frac{\mathbb{E}_{\theta_0,\lambda_{i0}}[(\ell_{iT}^{\lambda})^3]\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda\lambda}]}{\{\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda}]\}^3}, \qquad (5)\end{aligned}$$

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 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 $O(T^{-2})$. Another contribution is the derivation of a likelihood based characterisation of this extra bias term, given in (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 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 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) 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 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*

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

Priors (P1), (P2) and (P2*) follow directly from Theorem 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) 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.

5 Bias of the Integrated Likelihood Estimator in the Presence of Cross-Section Dependence

The case of cross-section dependence has so far not been analysed in the analytical bias-reduction literature.[12] The question of interest is whether cross-section dependence introduces extra bias terms in addition to $\mathcal{B}_{iT}^{(1)}\left(\theta\right)/T$. As discussed in Section 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 3.11 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.

[12] 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 5.1 *Under Assumptions 3.1-3.11*

$$\begin{aligned}(\hat{\theta}_{IL} - \theta_0) \;=\; & -\nu^{-1}S \\ & -\nu^{-1}\frac{1}{N}\sum_{i=1}^{N}\nabla_{\theta}\left\{\frac{1}{T}\ln\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta,\bar{\lambda}_i(\theta))]\right. \\ & \left.+\frac{1}{T}\ln\pi_i(\bar{\lambda}_i(\theta)|\theta) - \frac{[\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta))]^2}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta,\bar{\lambda}_i(\theta))]}\right\}\Bigg|_{\theta=\theta_0} \\ & +\nu^{-1}(H-\nu)S\nu^{-1} - \frac{1}{2}\nu^{-1}M + O_p\left(\frac{1}{T^{3/2}}\right). \end{aligned} \tag{6}$$

The first term on the right-hand side of (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 3.11 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 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 5.1 and Corollary 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 (6) and is corrected by the robust prior. The (potential) second type of bias is due to the third and fourth terms in (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 5.3 *If* ρ_1 *and* ρ_2 *are assumed to be such that*

$$1/2\leq\rho_1\leq 1,\qquad 0\leq\rho_2\leq 1\qquad \textit{and}\qquad 1\leq\rho_1+\rho_2\leq 2, \tag{7}$$

then, using the robust prior gives

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

Hence, (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 (6)*

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

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

Corollary 5.3 can be confirmed by inspection, using the results of Theorem 5.1 and Corollary 5.2. The first part of Corollary 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.[13] 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.

6 Analysis of Bias in the Presence of Spatial Dependence and Clustering

6.1 Discussion on Modelling Cross-Section Dependence

This section considers a spatial dependence/clustering based approach to model cross-section dependence and analyses 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 analytical bias reduction and clustering/spatial dependence literatures. The motivation for this is quite pragmatic. Theorem 5.1 and Corollary 5.2 suggest that now that one knows what the bias due to cross-section dependence looks like, one can simply remove these terms and reduce the bias. However, this requires exact knowledge of the magnitudes of the extra terms, that is exact knowledge on the values of ρ_1 and ρ_2. If, for example, one mistakenly assumes that both of these terms are $O(T^{-1})$ while in reality they are $O(T^{-3/2})$, then the bias-correction operation will actually introduce a $O(T^{-1})$ bias. A possible solution, then, is to model dependence explicitly and obtain the values of ρ_1 and

[13] 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.

ρ_2 for that particular type of dependence. This is useful from an applied perspective as well, as it opens the way for bias reduction in spatially dependent and, as will be illustrated in a moment, clustered samples.

A particularly popular approach to modelling cross-section dependence is factor modelling. To illustrate by a simple example, let ε_{it} be the variable of interest where

$$\varepsilon_{it} = \delta_i \phi_t + \zeta_{it}, \qquad \phi_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\phi^2) \quad \text{and} \quad \zeta_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\zeta^2). \tag{8}$$

Then, ε_{it} exhibits contemporaneous cross-section dependence due to the presence of the common factor, ϕ_t, where the factor loading term, δ_i, ensures that the common factor impacts individuals differently.[14] Using this approach in the present context is not convenient for several reasons. First, the entire analysis is based on the likelihood function, rather than some random variable of interest, so the objective is to model the dependence of the likelihood function itself. 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. This works best for analytically tractable models. 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. However, given that the analysis here considers high 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 will be very difficult as higher order derivatives enter the analysis.

The more convenient alternative for our purposes is to consider a spatial dependence setting for the likelihood function itself. The idea in this setting is that the degree of dependence between individuals is related to their "distance."[15] This type of dependence is meaningful for various fields, including urban, agricultural, development and labour economics and economic growth. The main difficulty in modelling cross-section dependence in this fashion is that cross-section data do not possess some convenient properties of time-series data. To start with, 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. In addition, data arrive in a natural order: observations one week apart are less dependent than observations for two consecutive days.

[14] Important recent contributions in this literature include, but are certainly not limited to, Bai (2003, 2009, 2012), 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).

[15] 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).

Such a natural ordering and distance does not exist for cross-section data. Naturally, the definition of distance or, more formally, the choice of a "distance metric" depends on the application under consideration.[16] These are all concerns of applied econometric analysis and they have to be addressed fully in any application. However, as the analysis here considers the theoretical aspects of spatial dependence, we abstract away from such issues.

Then, the setting is as follows: the panel is spatially dependent and characterised by cross-sectional clustering. It is assumed that the dataset has already been divided into appropriate clusters by the econometrician. The task is to find out whether cross-section dependence will lead to extra bias terms. The cluster structure considered here is characterised by an increasing number of clusters and an increasing number of members in each cluster, as N 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.

6.2 Notation and the Dependence Setting

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.[17] 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.

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. When Z_{iT} is replaced by the relevant centred log-likelihood derivatives, the relevance of this approach in relation with Assumption 3.11 becomes immediately obvious.

As mentioned, individuals are assumed to have already been grouped into clusters by

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

[17] 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).

the researcher. Both the number of groups/clusters,[18] 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.[19] First, a general d-dimensional discussion is provided.[20] 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

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

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

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

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.[21] 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.

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

Definition 6.1 *For $D'_N \subseteq D$ and $D''_N \subseteq D$, define the random fields $Y' = \{Y_{iT,N} : i \in$*

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

[19] The following discussion closely follows Section 2 of Jenish and Prucha (2009) and Section 3 of Bester, Conley and Hansen (2011).

[20] 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).

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

$D_N'\}$ 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

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

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. This is because, 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.[22] 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.

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

Assumption 6.3 *Groups are mutually exclusive, exhaustive and contiguous in the maximum coordinatewise distance metric.*

Assumption 6.4 $\sup_{i,T} \mathbb{E}[|Z_{iT}|^{\tilde{\varepsilon}}] < \infty$ *where $\tilde{\varepsilon} > 2 + \delta$ for some $\delta > 0$.*

Assumption 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 $\varepsilon > 0$.*

[22] See the discussion in Jenish and Prucha (2011).

Assumption 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 6.1-6.3 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 6.4-6.6 are technical conditions for mixing processes which are used to invoke Theorem 1 of Jenish and Prucha (2009).

6.3 Theoretical Results

Under the notation introduced in the previous section,

$$\begin{aligned} Var\left(\frac{1}{N}\sum_{i=1}^{N} Z_{iT}\right) &= \frac{1}{N^2}\sum_{i=1}^{N} Var\left(Z_{iT}\right) + \frac{1}{N^2}\sum^{N}\sum^{N}_{i\neq j} Cov(Z_{iT}, Z_{jT}) \\ &= \frac{1}{N^2}\sum_{g=1}^{G}\sum_{i\in\mathcal{G}_g} Cov\left(Z_{iT}, Z_{jT}\right) \qquad (9) \\ &\quad + \frac{1}{N^2}\sum^{G}\sum^{G}_{g\neq h}\sum_{i\in\mathcal{G}_g}\sum_{j\in\mathcal{G}_h} Cov(Z_{iT}, Z_{jT}). \qquad (10) \end{aligned}$$

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 (9) and (ii) the normalised sum of covariances of all pairs from *different* clusters (10). The main result now follows.

Theorem 6.2 *Under Assumptions 6.1-6.6,*

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

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 6.4-6.6. 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 6.2 and 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 6.2.

To illustrate the significance of Theorem 6.2, first notice that in this study $d = 1$. Then, the following corollary yields the updated convergence rates for Assumption 3.11.

Corollary 6.3 *Assume that $L_N = O(\sqrt{N})$. If the sampling setting and the three likelihood derivatives considered in Assumption 3.11 satisfy Assumptions 6.1-6.6, 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 $N, T \to \infty$. Therefore, $\rho_1 = \rho_2 = \rho_3 = 1$.

The results follow by observing that, for example,

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

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

Not surprisingly, weak dependence across both time and cross-section leads to the faster convergence rate of $\sqrt{NT}$. As a consequence of Corollary 6.3, the cross-section dependence induced bias term will be negligible. Importantly, these results are based on the assumption that both the number of groups and the number of members of each group grow at rate $\sqrt{N}$. Of course, there can be many other settings, e.g. the number of groups might grow at a much slower rate. A more detailed analysis is certainly desirable but outside the scope of this paper.

7 Application: Modelling Financial Volatility in Small Samples

7.1 Panel Estimation of Volatility

The literature on ARCH-type models starts with Engle (1982) and Bollerslev (1986) who modelled the conditional variance of returns. Consider some variable of interest y_t where $t = 1, ..., T$, such that

$$y_t = \mu_t + \varepsilon_t, \quad \mu_t = \mathbb{E}[y_t|\mathcal{F}_{t-1}] \quad \text{and} \quad \varepsilon_t|\mathcal{F}_{t-1} \sim F(0, \sigma_t^2),$$

where $\mathcal{F}_t$ is the information set at time t and $F(0, \sigma_t^2)$ is some zero-mean distribution with variance σ_t^2. To keep the analysis simple, and since the focus of this study is on conditional variance, henceforth it is assumed that $\mu_t = \mathbb{E}[y_t|\mathcal{F}_{t-1}] = 0$. This is a reasonable assumption for, for example, daily stock returns. Then, the GARCH(1,1) model is given by

$$\sigma_t^2 = \omega_i + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \quad \text{where} \quad \omega > 0;\ \alpha,\beta \geq 0 \text{ and } \alpha + \beta < 1.$$

Hence, other things being equal, high/low past shocks, ε_{t-1}, lead to high/low conditional variance today. Similarly, high/low past conditional variance, σ_{t-1}^2, causes high/low conditional variance today. The common approach to parameter estimation is to conduct "Quasi Maximum Likelihood estimation" (QMLE) by using the Normal distribution instead of the unknown true distribution F. As shown by Bollerslev and Wooldridge (1992), this gives consistent estimators even if the normality assumption is wrong, as long as the conditional mean and conditional variance are correctly specified.

ARCH-type univariate models of volatility are based on the analysis of financial time-series individually, while multivariate volatility modelling focuses on the covariance structure between many financial series.[23] Estimation of GARCH parameters in a panel rather than the standard time-series setting was suggested by Pakel, Shephard and Sheppard (2011), which they call the GARCH Panel method. Their main motivation is that consistent estimation of GARCH parameters by standard time-series methods typically requires 1,000-1,500 observations, due to the nonlinear dynamics of the GARCH model and the high levels of persistence in the conditional variance.[24] For financial or macro variables such as hedge fund returns, inflation and industrial production, which are recorded at monthly frequency, a long record of observations may not exist. This virtually rules GARCH modelling out for such datasets. As a remedy for insufficient time-series variation, they propose utilising the cross-section information, as well; hence, the panel approach. This they achieve by applying the results of Engle, Shephard and Sheppard (2008) to univariate volatility modelling. Their simulation and empirical analyses suggest that, although the GARCH Panel method leads to gains both in- and out-of-sample, it suffers from the incidental parameter issue. This, however, is not investigated theoretically. The current study, although motivated by their results, is concerned with analysing the first-order bias properties of nonlinear and dynamic panels under time-series and cross-section dependence. As such, although the GARCH Panel method is used as some sort of an extended example, this paper has a wider scope than GARCH modelling.

A GARCH panel is defined as a collection of N individual financial time-series that are characterised by GARCH(1,1) dynamics. Crucially, it is assumed that the parameters of interest that govern the conditional variance dynamics (α and β) are common to all series while the intercept parameters are allowed to vary across cross-section. It can be shown that this implies individual-specific long-run (unconditional) variances. Hence stock X can, on average, be more volatile than stock Y, although their volatilities will evolve according

[23] An introductory survey of univariate models is given by Teräsvirta (2009), while a detailed analysis of multivariate GARCH models is provided by Bauwens, Laurent and Rombouts (2006). See Francq and Zakoïan (2010) for a detailed textbook treatment of GARCH type models.

[24] For example, GARCH parameter estimates for stock market volatility usually imply high level of persistence, close to being unit-root (Nelson (1991))

to the same dynamics. Specifically, let the variable of interest, e.g. stock returns, for series i at time t be given by

$$y_{it} = \mathbb{E}[y_{it}|\mathcal{F}_{i,t-1}] + \varepsilon_{it} \quad \text{where} \quad t = 1, ..., T \quad \text{and} \quad i = 1, ..., N.$$

Here, $\mathcal{F}_{i,t}$ is the information set for individual i at time t and, again, it is assumed that $\mathbb{E}[y_{it}|\mathcal{F}_{i,t-1}] = 0$. The specification of the conditional variance follows along common lines where

$$\begin{aligned} \varepsilon_{it} &= \sigma_{it}\eta_{it}, \quad \eta_{it} \sim F, \quad \mathbb{E}[\eta_{it}] = 0, \quad Var(\eta_{it}) = 1, && (11) \\ \sigma_{it}^2 &= \lambda_i (1 - \alpha - \beta) + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2, && (12) \\ \lambda_i &> 0 \ \forall i; \quad \alpha, \beta \geq 0 \quad \text{and} \quad \alpha + \beta < 1, && (13) \end{aligned}$$

where F is, again, some distribution, such as the Standard Normal. It must be underlined that η_{it} are not assumed to be iid across i as it is reasonable to assume that financial time-series are characterised by some degree of cross-sectional dependence. Possible examples of this could be returns of firms operating in the same industry or of funds following similar investment strategies.

The "variance-targeting" representation in (12) implies that $\mathbb{E}[y_{it}^2] = \lambda_i$.[25] Therefore, a simple method of moments estimator for λ_i is provided by

$$\tilde{\lambda}_i = T^{-1} \sum_{t=1}^{T} y_{it}^2. \tag{14}$$

Pakel, Shephard and Sheppard (2011) use this to estimate $\lambda_1, ..., \lambda_N$ in a first step. In the second step, estimators of the intercept parameters are plugged into the pseudo-likelihood function to obtain an estimator for θ. This two-step estimation method allows for estimation of the GARCH parameters under large cross-section dimensions.

What remains is to construct the joint likelihood function for $\{y_{it}\}_{i=1,...,N;t=1,...,T}$. Define $\theta = (\alpha, \beta)$ and let $\ell_{it}(\theta, \lambda_i) \equiv \ell_{it}(\theta, \lambda_i; y_{it}|\mathcal{F}_{i,t-1})$ be the conditional log-likelihood for y_{it}. To side-step the computational and statistical issues in modelling the full joint likelihood, a composite likelihood function is used as an approximation to the joint likelihood. This is achieved by averaging the univariate marginal (conditional) likelihoods. Then, the composite likelihood function given by $(NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \ell_{it}(\theta, \lambda_i)$ will still deliver consistent estimators, albeit with some efficiency loss, depending on the specific dependence structure (see Cox and Reid (2004) and Engle, Shephard and Sheppard (2008) for theoretical details).[26] Hence, the composite likelihood method offers a convenient way

[25] Using $\lambda_i(1 - \alpha - \beta)$ instead of ω_i is a monotonic transformation of the model which does not affect its properties. See Engle and Mezrich (1996) who introduced this parameterisation.

[26] It is possible to account for covariation between conditional densities by using bivariate conditional densities, as well. However, this approach will not be taken here, as it will increase the computational burden further, which is already high when bias-reduction methods are employed. Moreover, simulation results in Pakel, Shephard and Sheppard (2011) suggest that this simple approximation delivers satisfactory

of pooling information, while keeping the computational burden at a minimum.[27]

7.2 Calculation of the Priors

For (P1) and (P2), estimates of population moments are obtained by using

$$\begin{aligned}
\widehat{\mathbb{E}}\left[-\ell_i^{\lambda\lambda}(\theta,\lambda_i)\right] &= -\frac{1}{T}\sum_{t=1}^{T}\ell_{it}^{\lambda\lambda}(\theta,\lambda_i), \\
\widehat{\mathbb{E}}\{\left[\ell_i^{\lambda_i}(\theta,\lambda_i)\right]^2\} &= 2\sum_{l=0}^{T^{1/3}}\left(1-\frac{l}{1+T^{1/3}}\right)\Omega_l(\theta,\lambda_i), \\
\Omega_l(\theta,\lambda_i) &= \frac{1}{T}\sum_{t=\max(1,l+1)}^{\min(T,T+l)}\left[\ell_{it}^{\lambda}(\theta,\lambda_i)\times\ell_{i,t-l}^{\lambda}(\theta,\lambda_i)\right].
\end{aligned}$$

Calculation of $\widehat{\mathbb{E}}\{[\ell_i^{\lambda}(\theta,\lambda_i)]^2\}$ is based on heteroskedasticity and autocorrelation consistent (HAC) covariance estimation by Newey and West (1987) (see also Arellano and Hahn (2006)). Derivatives of the log-likelihood are not available in closed form for the GARCH(1,1) process and are calculated by using numerical optimisation methods.

7.3 Simulation Analysis

7.3.1 Simulation Setting

In this section, small sample performance of the integrated likelihood method using priors (P1) and (P2) is analysed. The baseline estimation method is the Composite Likelihood (CL) method suggested by Pakel, Shephard and Sheppard (2011) to estimate the GARCH panel model. The Infeasible Composite Likelihood (InCL) method, where true values of λ_i are used in estimation, is used as the theoretical benchmark. Lastly, integrated likelihood methods using prior (P1) and (P2) are designated as the integrated composite likelihood (ICL) and integrated pseudo composite likelihood (IPCL) methods, respectively.

In light of the simulation results in Pakel, Shephard and Sheppard (2011), who observe that the incidental parameter problem is most acute when T is around or less than 250, this section focuses on $T \in \{75, 100, 150, 200, 400\}$ and $N \in \{25, 50, 100\}$. Data are generated for $\theta_0 = (0.05, 0.93)$ and the nuisance parameters are drawn from a uniform distribution such that the corresponding annual volatility is between 15% and 80%, which provides a reasonable interval for most stock returns.

results.

[27] Utilisation of cross-sectional information in modelling conditional variance, by focusing on a collection of GARCH processes, has previously also been considered by e.g. Engle and Mezrich (1996), Bauwens and Rombouts (2007), Engle, Shephard and Sheppard (2008) and Engle (2009). However, this study follows a different approach and models conditional variance explicitly within a panel structure. Hospido (2010) also considers GARCH errors in analysing earning dynamics using the PSID dataset; however, she assumes cross-section independence and does not analyse the effects of time-series dependence on the incidental parameter bias.

Data are generated by using,

$$\begin{aligned} y_{it} &= \mu_{it} + \varepsilon_{it}, \qquad \mu_{it} = E[y_{it}|\mathcal{F}_{i,t-1}] = 0, \qquad \varepsilon_{it} = \sigma_{it}\eta_{it}, \\ \sigma_{it}^2 &= \lambda_i(1-\alpha-\beta) + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2, \qquad \sigma_{i0}^2 = \lambda_{i0}, \end{aligned}$$

where the unconditional variance, λ_{i0}, is used as the initial value for the conditional variance, σ_{i0}^2. Following Engle, Shephard and Sheppard (2008), cross-sectional dependence is generated by using a single-factor model where

$$\begin{aligned} &\eta_{it} = \rho_i u_t + \sqrt{1-\rho_i^2}\tau_{it}, \\ &u_t \overset{iid}{\sim} N(0,1), \qquad \tau_{it} \overset{iid}{\sim} N(0,1). \end{aligned}$$

This implies that

$$\begin{aligned} E[\eta_{it}|\rho_i] &= 0 \quad \forall i, t, \\ cov\left[\begin{pmatrix} \eta_{it} \\ \eta_{jt} \end{pmatrix} \Big| \rho_i, \rho_j\right] &= \begin{bmatrix} 1 & \rho_i\rho_j \\ \rho_i\rho_j & 1 \end{bmatrix} \quad \forall i \neq j \text{ and } \forall t \\ cov(\eta_{it}, \eta_{js}|\rho_i, \rho_j) &= 0 \quad \forall t \neq s \text{ and } \forall i, j. \end{aligned}$$

For this purpose, ρ_i are drawn from a Uniform distribution where $\rho_i \sim U(0.5, 0.9)$. Therefore, the correlation between any two given series will be between 25% and 81%. [28]

Estimation is conducted in Matlab. The optimisation procedure supplied by this software requires user-supplied starting values for the parameters of interest. In order to prevent any bias in estimation performance due to the selection of starting values, starting values for α and β are drawn randomly from a Uniform distribution, for each replication, 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 basic quadrature method. It is possible to use different and more sophisticated numerical integration methods. However, to keep the analysis simple, these will not be investigated here. The integrated composite likelihood estimator is obtained by using iterated updating. Iteration stops either at the tenth iteration or convergence of the estimator, whichever happens first. 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. Lastly, an initial value for conditional variance, σ_{i0}^2, has to be specified to construct the composite likelihood. This is done by using

$$\sigma_{i0}^2 = \frac{1}{\lceil T^{1/2} \rceil} \sum_{t=1}^{\lceil T^{1/2} \rceil} y_{it}^2,$$

[28] It is important to ensure that cross-sectional dependence is not too high as that will lead to inconsistency of the composite likelihood estimator (see Cox and Reid (2004)). Seen from a different perspective, high levels of cross-sectional dependence will imply that there is not much point in considering a panel structure as there is not much cross-sectional variation.

as in Shephard and Sheppard (2010), where $\lceil T^{1/2} \rceil$ is obtained by rounding $T^{1/2}$ up to the nearest integer.[29]

7.3.2 Analysis of Estimation Performance

Simulation results are based on 500 replications. The following results and illustrations are provided for the cross-sectional dependence case: Average parameter estimates, calculated over all replications, are given in Table 1. Also, the sample standard deviations of parameter estimates ($\bar{\sigma}_{\hat{\alpha}}$ and $\bar{\sigma}_{\hat{\beta}}$) and the root mean square errors ($\mathcal{R}_{\hat{\alpha}}$ and $\mathcal{R}_{\hat{\beta}}$) are given on the left and right panels of Table 2, respectively. Finally, sample distributions of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\alpha} + \hat{\beta}$ are given in Figures 1, 2 and 3, respectively. Several results for panels with cross-section independence are also provided for comparison: Average parameter estimates are given in Table 3, while Table 4 presents the sample standard errors and root mean square errors.

Result in Table 1 suggest that using the integrated likelihood and the robust priors leads to substantial reductions in the bias of the CL estimators. In some cases, the reduction in bias is enormous: for example, for $T = 100$ and $N = 100$, ICL reduces 52% of the bias in $\hat{\alpha}$ due to CL, while the bias in $\hat{\beta}$ is reduced by 73%, 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 reveal 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 clear trend 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 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 $\alpha + \beta$. However, without further theoretical analysis, which is beyond the scope of this study, this remains a speculation.

Figures 1, 2 and 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 $\hat{\alpha} + \hat{\beta}$, 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

[29] A more detailed discussion of the estimation procedure, which is standard, is given in Appendix B for possible replication purposes.

$\alpha + \beta$ is a slightly different story where, in line with the previous discussion, CL is slow in converging to InCL.

Sample distribution of $\hat{\alpha}$ given in Figure 1 gives some more idea about the behaviour of $\hat{\alpha}$. Results for CL are omitted in the first row, as in almost all cases $\hat{\alpha} \approx 0$, implying that β is not identified. Although the situation for ICL and IPCL is not as severe, in a substantial proportion of cases $\hat{\alpha} \approx 0$, nevertheless. However, the ratio of such cases diminishes as T increases. Moreover, for a given T, increasing N also leads to a substantial decrease in the number of instances of $\hat{\alpha} \approx 0$, for ICL and IPCL. One example is panels with 75 time-series observations. Clearly, increasing the number of cross-sectional observations from 25 to 100 makes almost all cases where $\hat{\alpha} \approx 0$ disappear. The same is not observed for CL, which is not surprising. Increasing T provides more time-series variation, leading to better estimation of the incidental parameter. Increasing N, on the other hand, implies more cross-sectional variation, which would improve the estimation of the common parameter but not the nuisance parameter. Simulation results are in line with this argument, since the problem in estimation of $\hat{\alpha}$ by CL can only be solved by increasing T as what is missing is time-series information. ICL and IPCL, on the other hand, are based on the bias reduction mechanism, implying that the small-T issue is much less severe. Adding more cross-sectional information is, thus, enough to improve the estimation of $\hat{\alpha}$.

In line with the rest of the bias-reduction literature, bias-reduction does *not* come at a cost of increased variance. The left panel of Table 2 reveals that bias-reduction by robust priors does not increase the variance of the estimators in comparison to CL; instead it leads to lower standard deviation.[30] As before, as T increases, standard deviations of different methods become similar. Also, for a 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, as can be observed in the right panel of Table 2.

Finally, it is an interesting question whether neglecting the presence of cross-sectional correlation when constructing bias-reducing priors might have a non-negligible effect on parameter estimates. For example, one unpleasant scenario could be such that bias is reduced not because of bias-reducing priors directly, but because of possible interaction between the prior, the bias term and the extra term that appears due to cross-sectional dependence. However, simulation results reveal that the effect of neglected cross-sectional dependence is not on the bias of the estimator but on its variance. Comparison of Tables 1 and 3 indicates that for all methods the change in average bias due to neglected cross-section dependence is not significant. One observation for IPCL is that under cross-sectional independence $\hat{\alpha} + \hat{\beta}$ is estimated with less bias even when using shorter panels, while $\hat{\beta}$ is slightly more biased. In general, these minor differences between the two dependence structures tend to lessen as T increases. The change in sample standard

[30] The only exception to this observation occurs for $\hat{\alpha}$ when $T = 50$. However, it must be remembered that in this case, in the majority of replications, $\hat{\alpha} \approx 0$ for CL, which implies very low variance.

deviation, on the other hand, is striking. In some cases, introduction of cross-sectional dependence leads to a three-fold increase in standard deviation (see Tables 2 and 4).[31]

To summarise, simulation results show that, in line with the theoretical results, bias reduction using robust priors removes a substantial portion of the bias. Moreover, bias-reduction does not entail an increase in the standard deviation of the estimators and, instead, leads to lower standard deviation compared to CL. Crucially, robust priors achieve good small sample properties when T is around 150, which suggests that they can be used to model conditional volatility for short GARCH panels. Importantly, simulation results indicate that the effect of neglected cross-sectional dependence is clearly on standard deviation while it has little or no effect on average bias.

7.3.3 Analysis of Likelihoods

Finally, average likelihood plots, based on the 500 replications, for several panel dimensions are provided in Figure 4. 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. Lastly, in order to calculate the integrated CL, a value for α and β at which the robust prior has to be evaluated should be chosen for each replication. For a given replication, integrated likelihood estimates from the penultimate iteration are used for that purpose.

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, 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.

8 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

[31] Phillips and Sul (2007) analyse the Nickell bias under neglected cross-sectional dependence and show that, in such a setting, the probability limit of the estimator becomes a random variable. This could be considered similar in spirit to the results obtained here, which suggest that neglected cross-sectional dependence leads to higher dispersion of the average bias.

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).

8.1 Analysis of Predictive Ability

8.1.1 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 comparison of forecast loss due to competing estimators, where the forecast loss is computed with respect to the variable of interest; the conditional variance. However, conditional variance is not observable, even ex-post, and a proxy has to be used instead. A convenient proxy is squared returns. However, this is a very noisy proxy, potentially leading to misleading results (Patton and Sheppard (2009) and Patton (2011)). A better alternative is realised variance, which is an estimator of ex-post volatility based on high-frequency intra-daily data.[32] An important advantage of the chosen dataset is that it includes realised variances for each stock, in addition to daily returns. This is the main motivation behind using this dataset, as the ability to base forecast comparison on a more accurate proxy is a crucial one.[33]

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.

[32] See, for example, Andersen, Bollerslev, Diebold and Labys (2001), Barndorff-Nielsen and Shephard (2002), and Barndorff-Nielsen, Lunde, Hansen and Sheppard (2008). Reviews include Barndorff-Nielsen and Shephard (2007) and Andersen, Bollerslev and Diebold (2009).

[33] 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, a more detailed analysis is left for future research.

8.1.2 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. This setting also allows for a comparison of the forecasting performances of the standard QML method against the panel-based methods (CL and IPCL). QML and CL are based on a two-step estimation framework which uses the variance-tracking version of GARCH as specified in (12). In the first step, λ_i are estimated by method of moments using (14). $\tilde{\lambda}_1, ..., \tilde{\lambda}_N$ 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 support of the integrand of the integrated likelihood (as set by the researcher) includes the true parameter value.[34]

An important concern is estimation of the intercept parameter. 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 given in (14) is consistent and valid independent of how α and β are estimated. However, remember that the integrated likelihood estimators are in essence concentrated likelihood estimators. Therefore, a natural intercept estimator is given by

$$\hat{\lambda}_i^c(\hat{\theta}_{IL}) = \arg\max_{\lambda_i \in \Lambda_i} \frac{1}{T}\sum_{t=1}^{T} \ell_{it}(\hat{\theta}_{IL}, \lambda_i). \tag{15}$$

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.[35] 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.[36]

[34] In this study, when calculating the integrated likelihood, the upper and lower limits of the integral are set to $2 \times (\max_{i,t} r_{it}^2)$ and $.8 \times (\min_{i,t} r_{it}^2)$.

[35] 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.

[36] 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

8.1.3 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). As the objective of this analysis is to compare methods (QML, CL and IPCL) rather than models (GARCH, Exponential GARCH etc.) this test, rather than the Diebold-Mariano-West[37] type tests, is better suited to the analysis.

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. The dataset consists of $2,176$ observations, implying a total of $2,026$ forecasts for each of the nine stocks.

To briefly describe the test procedure, 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 measure by using the QLIKE loss function:

$$L(\sigma^2_{i,t+1}, \hat{\sigma}^2_{i,t+1}) = \log \hat{\sigma}^2_{i,t+1} + \frac{\sigma^2_{i,t+1}}{\hat{\sigma}^2_{i,t+1}}.$$

A particular advantage of QLIKE is that it is robust to noisy proxies (Patton (2011)). In other words, on average, it is expected to provide the same ranking between two forecasts independent of whether the true conditional variance or a conditionally unbiased proxy is used.

Defining RV_{it} as the realised variance for stock i at time t, the difference between the loss functions when RV_{it} is used as the proxy is given by $\Delta L_{i,t+1} = L(RV_{i,t+1}, \hat{\sigma}^2_{1,i,t+1}) - L(RV_{i,t+1}, \hat{\sigma}^2_{2,i,t+1})$. Assuming that forecasts are made at periods $\underline{T}$ to $\overline{T}$, the test setup is given by

$$\begin{aligned} H_0 &: \mathbb{E}[\Delta L_{i,t}] = 0 \quad \text{for} \quad t = \underline{T}, \underline{T}+1, ..., \overline{T}, \\ H_1 &: \left|\mathbb{E}[\Delta \bar{L}_{i.n}]\right| \geq \delta > 0 \quad \text{for all } n \text{ sufficiently large,} \end{aligned}$$

where $\Delta \bar{L}_{i,n} = n^{-1} \sum_{t=\underline{T}}^{\overline{T}} \Delta L_{i.t}$ and $n = \overline{T} - \underline{T} + 1$. The relevant test statistic is $t_{i,n} = \sqrt{n}\Delta \bar{L}_{i,n}/\hat{\sigma}_n$, where $\hat{\sigma}_n$ is an estimator for $\sigma^2_n = var\left(\sqrt{n}\Delta \bar{L}_{i,n}\right)$, obtained by using a HAC estimator. Under H_0, $t_{i,n}$ converges in distribution to $\mathcal{N}(0,1)$ as $n \to \infty$. See Giacomini and White (2006) for details. Intuitively, if H_0 is rejected, a positive $\Delta L_{i,t+1}$ implies relatively higher loss due to the first method, suggesting that the second method has

necessarily the same as $(NT)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N} \ell_{it}(\theta, \hat{\lambda}_i(\theta))$ where $\hat{\lambda}_i(\theta) \equiv \arg\max_{\lambda_i} T^{-1} \sum_{t=1}^{T} \ell_{it}(\theta, \lambda_i)$.

[37] See the seminal works by Diebold and Mariano (1995) and West (1996). Basically, the structure of these tests is such that the null hypothesis is based on the probability limits of the estimators. Therefore, they are not suited to comparing different methods that all produce consistent estimators of the same parameter. Under the GW-test framework, on the other hand, the in-sample size is not allowed to increase asymptotically, which allows for comparison of different methods, even if they are based on the same model.

better predictive ability (and similarly for negative $\Delta L_{i,t+1}$).

8.1.4 Results

The test results are given in Table 5, which contains the t-statistics and the result of the GW-test. Loss functions are based on realised variances, RV_{it}. A dash signifies that the test result is inconclusive. All tests are done at 5% level of significance.

Forecasts for QML and CL are based on intercept parameter estimates by the method of moments, while IPCL forecasts are based on intercept estimates by the concentrated likelihood estimator for the nuisance parameter, as in (15). The GW-test indicates that IPCL achieves a better forecasting performance compared to both QML and CL. Except for two cases (Coca Cola and Microsoft), IPCL delivers less loss relative to CL, with six of those being statistically significant. The difference is larger between QML and IPCL where the GW-test favours IPCL seven out of nine times. Furthermore, Columns 2 and 3 of Table 5 indicate that QML always leads to a higher loss in comparison to CL, as all t-statistics are positive. The difference is statistically significant in three out of nine cases where the test decides in favour of CL. These results suggest that in the given sample the panel-based methods perform better that the standard QML method in forecasting one-step ahead volatility. Moreover, IPCL emerges as the best performer and bias-reduction clearly improves the performance of panel-based estimation in comparison to QML.[38]

8.2 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. By the end of 2011, the global assets under management were expected to reach \$2.25 billion , despite capital outflows following the credit crunch episode.[39] 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 lower than that). Clearly, this is well below what is

[38] Whether estimating λ_i by concentrated likelihood, rather than the method of moments, leads to a difference in small samples is an interesting question. In large samples, not much difference would be expected as both estimators are consistent. However, in small samples things might be different. Results not reported here show that although IPCL still outperforms QML, it does so less decisively, while the comparison between CL and IPCL results in a draw. The majority of the t-statistics is still in favour of IPCL, but not large enough to force rejection of the hypothesis of equal predictive ability. These results are available upon request. A thorough analysis of the effects of the intercept estimator on predictive ability is left for future research.

[39] Sources: *The Economist*, June 10, 2004; *Financial Times*, March 10, 2011.

necessary for traditional GARCH estimation to be successful. However, as the simulation results indicate, the 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. The only relevant analysis known to me is by Huggler (2004) who argues that modelling hedge fund portfolio returns is problematic due to the 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), (see, for example, Bollen and Whaley (2009), Teo (2009) and Patton and Ramadorai (2011))[40]. 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.

[40] 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.

8.2.1 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.[41] 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.

8.2.2 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 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.

[41] 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.

Parameter estimates and the number of included funds for the three sample sizes are reported in Table 6. 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.[42] 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 $\hat{\beta}$). These observations hold generally, independent of the panel length.

Figure 5 gives an overview of fitted conditional volatilities for $T = 150$.[43] 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 5, 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 6. 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 provides one extreme example 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

[42] 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.

[43] Intercept parameters have been estimated using the concentrated likelihood method as in (15).

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 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 7 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) is observed. For the Security Selection, Directional Traders and Emerging Markets 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.

9 Conclusion

This paper has analysed the first-order bias in nonlinear dynamic panel data models in the presence of both time-series and cross-section dependence. Extending the analytical bias reduction literature to the case of cross-section dependence is the main contribution of this study to the panel data literature. In doing so, the extra bias terms that appear due to dependence in the time-series and cross-section dimensions have been characterised. The theoretical investigation reveals that time-series dependence leads to an extra yet negligible bias term. However, crucially, the extra bias term due to cross-section dependence might not be negligible, depending on the strength of cross-section dependence. These results are also useful in establishing the conditions under which the Arellano-Bonhomme priors can still be safely used for bias-correction, despite the presence of two-dimensional dependence. Furthermore, the specific case of spatial cross-section dependence for clustered individuals has also been analysed. It has been shown that, under certain assumptions on the cluster characteristics, the bias due to cross-section dependence is asymptotically negligible.

The theoretical analysis has a general scope in the sense that characterisations of extra bias terms are provided in terms of a general nonlinear and dynamic likelihood function, with no specific model in mind. Therefore, the results presented here are potentially applicable to a wide array of models. As a particular application, modelling of GARCH effects using panels with a limited number of time-series observations has been considered. Simulations indicate that the proposed approach can successfully reduce a substantial portion of the incidental parameter bias with 150-200 time series observations, without increasing the standard errors. This is in stark contrast with around 1,000-1,500 observations which would be required for consistent estimation of GARCH parameters using standard time-series methods. In an empirical analysis, hedge fund volatility characteristics have been analysed by focusing on groups of funds following different investment strategies. By analysing sample distributions of volatility across funds, it has been shown that hedge fund volatilities are in general characterised by an asymmetric right-skewed distribution and that the size of the right tail reacts to important economic events such as the burst of the dotcom bubble and the credit crunch. Moreover, in a test of predictive ability using stock volatility forecasts, the proposed estimation method achieved superior forecasting performance compared to its alternatives.

Suggestions for possible extensions are in order. The clustering example was based on one of many possible settings. It would be useful to extend the analysis to other settings where the number of groups and the membership size for each group are allowed to grow at different rates (or perhaps remain fixed). Naturally, such assumptions are very much linked to the particular application at hand. Therefore, a better understanding of bias under different settings would be beneficial to both theoretical and applied econometricians. The other possibility is to employ a factor structure. However, it is not clear how this can be done when nothing is known about the likelihood's functional form, which was the case in this paper. On the other hand, when the underlying model is known, this is a very effective approach which can deliver closed form expressions for the small sample bias. Especially for popular models such as the dynamic autoregressive panel or panel probit, this research avenue would be extremely fruitful. Last but not least, more simulation and empirical analyses have to be considered for a large variety of models in order to attain a better understanding of the behaviour of bias and the bias-correction performance.

A Mathematical Appendix

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

A.1 Definition of α-and ϕ-mixing

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

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

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

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

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).

A.2 A Preliminary Lemma

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

Lemma A.2 *Under Assumption 3.7,*

$$\begin{aligned}
\delta_i &= \frac{1}{E_{iT}} \Bigg\{ -\ell^{\lambda}_{iT} + \frac{V^{\lambda\lambda}_{iT}\ell^{\lambda}_{iT}}{E_{iT}} - \frac{1}{2}\frac{\ell^{\lambda\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^2}{E^2_{iT}} + \frac{-V^{\lambda\lambda}_{iT}}{E_{iT}}\left[\frac{V^{\lambda\lambda}_{iT}\ell^{\lambda}_{iT}}{E_{iT}} - \frac{1}{2}\frac{\ell^{\lambda\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^2}{E^2_{iT}}\right] \\
&\quad -\frac{1}{2}\frac{\ell^{\lambda\lambda\lambda}_{iT}}{E^2_{iT}}\left[-2\frac{V^{\lambda\lambda}_{i}\left(\ell^{\lambda}_{iT}\right)^2}{E_{iT}} + \frac{\ell^{\lambda\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^3}{E^2_{iT}}\right] - \frac{1}{6}\frac{\ell^{\lambda\lambda\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^3}{E^3_{iT}}\Bigg\} + O_p(T^{-2}), && (16) \\
\delta^2_i &= \frac{1}{E^2_{iT}}\left[\left(\ell^{\lambda}_{iT}\right)^2 - 2\frac{V^{\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^2}{E_{iT}} + \frac{\ell^{\lambda\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^3}{E^2_{iT}}\right] + O_p(T^{-2}), && (17) \\
\delta^3_i &= -\frac{1}{E^3_{iT}}\left(\ell^{\lambda}_{iT}\right)^3 + O_p(T^{-2}). && (18)
\end{aligned}$$

Proof of Lemma A.2. Expanding $\ell^{\lambda}_{iT}(\theta, \hat{\lambda}_i(\theta))$ around $\hat{\lambda}_i(\theta) = \bar{\lambda}_i(\theta)$ yields

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

Then

$$\begin{aligned}
\delta_i &= \frac{1}{E_{iT}}\left[-\ell^{\lambda}_{iT} - \frac{V^{\lambda\lambda}_{iT}}{E_{iT}}\left(-\ell^{\lambda}_{iT} - V^{\lambda\lambda}_{iT}\delta_i - \frac{1}{2}\ell^{\lambda\lambda\lambda}_{iT}\delta^2_{iT}\right) - \frac{1}{2}\ell^{\lambda\lambda\lambda}_{iT}\delta^2_i - \frac{1}{6}\ell^{\lambda\lambda\lambda\lambda}_{iT}\delta^3_i\right] + O_p(T^{-2}) \\
&= \frac{1}{E_{iT}}\Bigg\{ -\ell^{\lambda}_{iT} - \frac{V^{\lambda\lambda}_{iT}}{E_{iT}}\left[-\ell^{\lambda}_{iT} - \frac{V^{\lambda\lambda}_{iT}}{E_{iT}}\left(-\ell^{\lambda}_{iT}\right) - \frac{1}{2}\ell^{\lambda\lambda\lambda}_{iT}\delta^2_i\right] \\
&\quad -\frac{1}{2}\ell^{\lambda\lambda\lambda}_{iT}\delta^2_i - \frac{1}{6}\ell^{\lambda\lambda\lambda\lambda}_{iT}\delta^3_i\Bigg\} + O_p(T^{-2}) && (19)
\end{aligned}$$

Similarly,

$$\begin{aligned}\delta_i^2 &= \frac{1}{E_{iT}^2}\left[\left(\ell_{iT}^{\lambda}\right)^2 + 2\ell_{iT}^{\lambda}V_{iT}^{\lambda\lambda}\delta_i + \ell_{iT}^{\lambda}\ell_{iT}^{\lambda\lambda\lambda}\delta_i^2\right] + O_p(T^{-2}) \\ &= \frac{1}{E_{iT}^2}\left[\left(\ell_{iT}^{\lambda}\right)^2 - 2\frac{\left(\ell_{iT}^{\lambda}\right)^2 V_{iT}^{\lambda\lambda}}{E_i} + \ell_{iT}^{\lambda}\ell_{iT}^{\lambda\lambda\lambda}\delta_i^2\right] + O_p(T^{-2}), \qquad (20)\end{aligned}$$

where (19) is used to obtain (20). Substituting δ_i^2 back into (20) yields (17), while observing $\delta_i^3 = \delta_i\delta_i^2$ gives (18). Finally, using (17) and (18) in (19), (16) follows. ■

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

A.3 Proof of Theorem 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 A.3

$$\ell_{iT}^I(\theta) - \ell_{iT}^c(\theta) = \frac{1}{2T}\ln\left(\frac{2\pi}{T}\right) - \frac{1}{2T}\ln[-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))] + \frac{1}{T}\ln\pi_i(\hat{\lambda}_i(\theta)) + O\left(\frac{1}{T^2}\right). \qquad (21)$$

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

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

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 3.6, 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[44]

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

[44] 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 (21). 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. This property, along with the moment conditions in Assumption 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 A.4

$$\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta) = \frac{A_i}{\sqrt{T}} + O_p\left(\frac{1}{T}\right), \tag{22}$$

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 $\hat{\lambda}_i(\theta) = \bar{\lambda}_i(\theta)$.

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

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

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 A.5

$$\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta)) = \ell_{iT}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta)) + \frac{B_i}{\sqrt{T}} + O_p\left(\frac{1}{T}\right), \tag{23}$$

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 (22) 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 A.6

$$\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta)) = \frac{C_i}{\sqrt{T}} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta))] + O_p\left(\frac{1}{T}\right), \tag{24}$$

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 (23),

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

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 A.7

$$\mathbb{E}_{\theta_0,\lambda_{i0}}\left\{\ln[-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))]\right\} = \ln\{-\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta))]\} + O\left(\frac{1}{T}\right). \tag{25}$$

Proof. Using (24),

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

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

$$\ln\{-\ell_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta))\} = \ln\{-\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\} + \frac{C_i}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\sqrt{T}} + O_p(T^{-1}), \tag{26}$$

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 A.8

$$\mathbb{E}_{\theta_0,\lambda_{i0}}\left[\ln\pi_i(\hat{\lambda}_i(\theta)|\theta)\right] = \ln\pi_i(\bar{\lambda}_i(\theta)|\theta) + O\left(\frac{1}{T}\right). \tag{27}$$

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

$$\ln\pi_i(\hat{\lambda}_i(\theta)|\theta) = \ln\pi_i(\bar{\lambda}_i(\theta)|\theta) + \frac{\partial\ln\pi_i(\bar{\lambda}_i(\theta)|\theta)}{\partial\lambda_i}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta)) + O_p(T^{-1}), \tag{28}$$

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 A.1

$$\begin{aligned}\mathbb{E}_{\theta_0,\lambda_{i0}}\left[\ell_i^I(\theta) - \ell_i^c(\theta)\right] &= C - \frac{1}{2T}\ln\left\{-T^{-1}\sum\nolimits_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda}]\right\} \\ &\quad + \frac{1}{T}\ln\pi_i\left(\bar{\lambda}_i(\theta)|\theta\right) + O\left(\frac{1}{T^2}\right).\end{aligned} \tag{29}$$

Proof. Taking the expectation of (21) 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 (25) and (27), (29) follows. ■

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

$$\begin{aligned}
\mathbb{E}\left[\ell^{c}_{iT}(\theta, \hat{\lambda}_i) - \ell_{iT}(\theta, \bar{\lambda}_i)\right] &= -\frac{1}{2}\frac{\mathbb{E}[(\ell^{\lambda}_{iT})^2]}{\mathbb{E}[\ell^{\lambda\lambda}_{iT}]} + \frac{1}{2}\frac{\mathbb{E}[V^{\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^2]}{\left\{\mathbb{E}[\ell^{\lambda\lambda}_{iT}]\right\}^2} \\
&\quad -\frac{1}{6}\frac{\mathbb{E}[(\ell^{\lambda}_{iT})^3]\mathbb{E}[\ell^{\lambda\lambda\lambda}_{iT}]}{\left\{\mathbb{E}[\ell^{\lambda\lambda}_{iT}]\right\}^3} + O_p(\frac{1}{T^2}). \qquad (30)
\end{aligned}$$

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

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

Using Lemma A.2,

$$\begin{aligned}
\ell^{\lambda}_{iT}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta)) &= -\frac{\left(\ell^{\lambda}_{iT}\right)^2}{\mathbb{E}[\ell^{\lambda\lambda}_{iT}]} + \frac{V^{\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^2}{\left\{\mathbb{E}[\ell^{\lambda\lambda}_{iT}]\right\}^2} - \frac{1}{2}\frac{\mathbb{E}[\ell^{\lambda\lambda\lambda}_{iT}]\left(\ell^{\lambda}_{iT}\right)^3}{\left\{\mathbb{E}[\ell^{\lambda\lambda}_{iT}]\right\}^3} + O_p(T^{-2}), \\
\ell^{\lambda\lambda}_{iT}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta))^2 &= \frac{\left(\ell^{\lambda}_{iT}\right)^2}{\mathbb{E}[\ell^{\lambda\lambda}_{iT}]} - \frac{\left(\ell^{\lambda}_{iT}\right)^2 V^{\lambda\lambda}_{iT}}{\left\{\mathbb{E}[\ell^{\lambda\lambda}_{iT}]\right\}^2} + \frac{\left(\ell^{\lambda}_{iT}\right)^3 \mathbb{E}[\ell^{\lambda\lambda\lambda}_{iT}]}{\left\{\mathbb{E}[\ell^{\lambda\lambda}_{iT}]\right\}^3} + O_p(T^{-2}), \\
\ell^{\lambda\lambda\lambda}_{iT}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta))^3 &= -\frac{\left(\ell^{\lambda}_{iT}\right)^3 E[\ell^{\lambda\lambda\lambda}_{iT}]}{\left\{E[\ell^{\lambda\lambda}_{iT}]\right\}^3} + O_p(T^{-2}),
\end{aligned}$$

which implies that

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

■

Finally, the proof of Theorem 4.1 follows.

Proof. (Theorem 4.1) Using (29) and (30) gives (3), (4) and (5). ■

A.4 Proof of Theorem 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.4.1 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

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

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.4.2 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 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 A.9

$$\begin{aligned}
\nabla_\theta \ln(-E_{iT}) &= -\frac{E_{r_1}}{E} = O(1), \\
\nabla_{\theta\theta} \ln(-E_{iT}) &= -\frac{E_{r_1,r_2}}{E} + \frac{E_{r_1}E_{r_2}}{E^2} = O(1),
\end{aligned}$$

$$
\begin{aligned}
\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),\\
\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,

$$
\begin{aligned}
\nabla_{\theta\theta\theta}\ln(-E_{iT}) = O\left(1\right), &\qquad \nabla_{\theta\theta\theta\theta}\ln(-E_{iT}) = O\left(1\right),\\
\nabla_{\theta\theta\theta}\left(\frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\right)=O_p\left(\frac{1}{\sqrt{T}}\right), &\qquad \nabla_{\theta\theta\theta\theta}\left(\frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\right)=O_p\left(\frac{1}{\sqrt{T}}\right),
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{\theta\theta\theta}\left(\frac{\ell^{\lambda}_{iT}F_{iT}}{E^2_{iT}}\right) &= O_p\left(\frac{1}{\sqrt{T}}\right), \quad & \nabla_{\theta\theta\theta\theta}\left(\frac{\ell^{\lambda}_{iT}F_{iT}}{E^2_{iT}}\right) &= O_p\left(\frac{1}{\sqrt{T}}\right),\\
\nabla_{\theta\theta\theta}\left(\frac{\ell^{\lambda}_{iT}\bar{\Pi}_{iT}}{E_{iT}}\right) &= O_p\left(\frac{1}{\sqrt{T}}\right). \quad & \nabla_{\theta\theta\theta\theta}\left(\frac{\ell^{\lambda}_{iT}\bar{\Pi}_{iT}}{E_{iT}}\right) &= O_p\left(\frac{1}{\sqrt{T}}\right),\\
\nabla_{\theta\theta\theta}\left[\frac{\left(\ell^{\lambda}_{iT}\right)^2}{E_{iT}}\right] &= O_p\left(\frac{1}{T}\right), \quad & \nabla_{\theta\theta\theta\theta}\left[\frac{\left(\ell^{\lambda}_{iT}\right)^2}{E_{iT}}\right] &= O_p\left(\frac{1}{T}\right),\\
\nabla_{\theta\theta\theta}\left[\frac{V^{\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^2}{E^2_{iT}}\right] &= O_p\left(\frac{1}{T^{3/2}}\right), \quad & \nabla_{\theta\theta\theta\theta}\left[\frac{V^{\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^2}{E^2_{iT}}\right] &= O_p\left(\frac{1}{T^{3/2}}\right),\\
\nabla_{\theta\theta\theta}\left\{\frac{(\ell^{\lambda}_{iT})^3F_{iT}}{E^3_{iT}}\right\} &= O_p\left(\frac{1}{T^{3/2}}\right), \quad & \nabla_{\theta\theta\theta\theta}\left\{\frac{(\ell^{\lambda}_{iT})^3F_{iT}}{E^3_{iT}}\right\} &= O_p\left(\frac{1}{T^{3/2}}\right).
\end{aligned}
$$

A.4.3 The Proof

The starting point is

$$
\begin{aligned}
\ell^I_{iT}(\theta) \quad = \quad & \ell_{iT}(\theta,\bar{\lambda}_i(\theta)) + C - \frac{1}{2T}\left[\ln(-E_{iT}) + \frac{V^{\lambda\lambda}_{iT}}{E_{iT}} - \frac{\ell^{\lambda}_{iT}(\theta,\bar{\lambda}_i(\theta))F_{iT}}{E^2_{iT}}\right]\\
&+\frac{1}{T}\left[\ln\pi_{iT}(\bar{\lambda}_{iT}(\theta)|\theta) - \frac{\ell^{\lambda}_{iT}(\theta,\bar{\lambda}_i(\theta))}{E_{iT}}\frac{\partial\ln\pi_{iT}(\lambda_i|\theta)}{\partial\lambda_i}\Big|_{\lambda_i=\bar{\lambda}_{iT}(\theta)}\right]\\
&-\frac{1}{2}\frac{\left(\ell^{\lambda}_{iT}(\theta,\bar{\lambda}_i(\theta))\right)^2}{E_{iT}} + \frac{1}{2}\frac{V^{\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}(\theta,\bar{\lambda}_i(\theta))\right)^2}{\left(E_{iT}\right)^2}\\
&-\frac{1}{6}\frac{(\ell^{\lambda}_{iT}(\theta,\bar{\lambda}_i(\theta)))^3F_{iT}}{E^3_{iT}} + O_p(T^{-2}), \qquad (31)
\end{aligned}
$$

where $C = (2T)^{-1}\ln(2\pi/T)$. The first line follows from using (26) and (28) to substitute for $\ln[-\ell^{\lambda\lambda}_{iT}(\theta,\hat{\lambda}_i(\theta))]$ and $\ln\pi_i(\hat{\lambda}_i(\theta))$, respectively, in (21). Notice that the expression given by (28) is a function of $[\hat{\lambda}_i(\theta)-\bar{\lambda}_i(\theta)]$, which has to be substituted by (16), 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 A.2.

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

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \ell^I_{r_1}(\hat{\theta}_{IL}) \quad = \quad & \frac{1}{N}\sum_{i=1}^N \ell^I_{r_1}(\theta_0) + \left[\frac{1}{N}\sum_{i=1}^N \ell^I_{r_1,r_2}(\theta_0)\right]\delta^{r_2}_I + \frac{1}{2}\left[\frac{1}{N}\sum_{i=1}^N \ell^I_{r_1,r_2,r_3}(\theta_0)\right]\delta^{r_2}_I\delta^{r_3}_I\\
&+\frac{1}{6}\left[\frac{1}{N}\sum_{i=1}^N \ell^I_{r_1,r_2,r_3,r_4}(\theta_0)\right]\delta^{r_2}_I\delta^{r_3}_I\delta^{r_4}_I\\
&+\frac{1}{24}\left[\frac{1}{N}\sum_{i=1}^N \ell^I_{r_1,r_2,r_3,r_4,r_5}(\bar{\theta})\right]\delta^{r_2}_I\delta^{r_3}_I\delta^{r_4}_I\delta^{r_5}_I, \qquad (32)
\end{aligned}
$$

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

$$
\ell^I_{r_1,r_2,r_3,r_4,r_5}(\bar{\theta})\delta^{r_2}_I\delta^{r_3}_I\delta^{r_4}_I\delta^{r_5}_I = O_p(T^{-2}).
$$

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 (31) to obtain target-likelihood based approximations for integrated-likelihood derivatives appearing on the right-hand side of (32). These approximations are then substituted for relevant integrated likelihood derivatives in (32). This leads to the next lemma.

Lemma A.10

$$
\begin{aligned}
-\delta_I^{r_2}\nu_{r_1,r_2} &= \tilde{\ell}_{r_1} + \mathcal{D}_{1;r_1} + \delta_I^{r_2}\mathcal{H}_{r_1,r_2} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\nu_{r_1,r_2,r_3} + \mathcal{D}_{3;r_1} \\
&\quad + \delta_I^{r_2}\mathcal{D}_{2;r_1,r_2} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\mathcal{H}_{r_1,r_2,r_3} + \frac{1}{6}\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\nu_{r_1,r_2,r_3,r_4} + O_p\left(\frac{1}{T^2}\right). \qquad (33)
\end{aligned}
$$

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 (31) with respect to θ have to obtained. This is achieved by simply substituting the results given in Lemma A.9 as necessary. Then,

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

Substituting these expansions for the integrated likelihood derivatives into (32) 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

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

Then, using Assumption 3.11,

$$\begin{aligned}
-\delta_I^{r_2}\nu_{r_1,r_2} &= \tilde{\ell}_{r_1}\left(\theta_0,\bar{\lambda}_i(\theta_0)\right)\\
&+\delta_I^{r_2}\mathcal{H}_{r_1,r_2} + \frac{1}{T}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{E_{r_1}}{2E} + \frac{1}{N}\sum_{i=1}^{N}\Pi_{r_1}\right] - \frac{1}{N}\sum_{i=1}^{N}\frac{UU_{r_1}}{E}\\
&+\frac{1}{N}\sum_{i=1}^{N}\frac{U^2E_{r_1}}{2E^2} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\nu_{r_1,r_2,r_3} + \delta_I^{r_2}\delta_I^{r_3}\frac{1}{2}\mathcal{H}_{r_1,r_2,r_3}
\end{aligned}$$

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

or, more concisely,

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

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 (33) 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, (33) 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 A.11

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

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

Remark 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 (33). 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 (33) by $\nu^{r_1,m}$

$$\begin{aligned} \delta_I^m &= -\Bigg[\tilde{\ell}_{r_1}\nu^{r_1,m} + \mathcal{D}_{1;r_1}\nu^{r_1,m} + \delta_I^{r_2}\mathcal{H}_{r_1,r_2}\nu^{r_1,m} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\nu_{r_1,r_2,r_3}\nu^{r_1,m} + \mathcal{D}_{3;r_1}\nu^{r_1,m} \\ &\quad + \delta_I^{r_2}\mathcal{D}_{2;r_1,r_2}\nu^{r_1,m} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\mathcal{H}_{r_1,r_2,r_3}\nu^{r_1,m} + \frac{1}{6}\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\nu_{r_1,r_2,r_3,r_4}\nu^{r_1,m}\Bigg] \\ &\quad + O_p\left(\frac{1}{T^2}\right). \end{aligned} \tag{34}$$

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 (34).

$$\begin{aligned} \delta_I^{r_2} &= -\Bigg[\tilde{\ell}_a\nu^{a,r_2} + \mathcal{D}_{1;a}\nu^{a,r_2} + \delta_I^b\mathcal{H}_{a,b}\nu^{a,r_2} + \frac{1}{2}\delta_I^b\delta_I^c\nu_{a,b,c}\nu^{a,r_2} + \mathcal{D}_{3;a}\nu^{a,r_2} \\ &\quad + \delta_I^b\mathcal{D}_{2;a,b}\nu^{a,r_2} + \frac{1}{2}\delta_I^b\delta_I^c\mathcal{H}_{a,b,c}\nu^{a,r_2} + \frac{1}{6}\delta_I^b\delta_I^c\delta_I^d\nu_{a,b,c,d}\nu^{a,r_2}\Bigg] + O_p\left(\frac{1}{T^2}\right), \\ \delta_I^{r_3} &= -\Bigg[\tilde{\ell}_e\nu^{e,r_3} + \mathcal{D}_{1;e}\nu^{e,r_3} + \delta_I^f\mathcal{H}_{e,f}\nu^{e,r_3} + \frac{1}{2}\delta_I^f\delta_I^g\nu_{e,f,g}\nu^{e,r_3} + \mathcal{D}_{3;e}\nu^{e,r_3} \\ &\quad + \delta_I^f\mathcal{D}_{2;e,f}\nu^{e,r_3} + \frac{1}{2}\delta_I^f\delta_I^g\mathcal{H}_{e,f,g}\nu^{e,r_3} + \frac{1}{6}\delta_I^f\delta_I^g\delta_I^h\nu_{e,f,g,h}\nu^{e,r_3}\Bigg] + O_p\left(\frac{1}{T^2}\right), \end{aligned} \tag{35}$$

$$\begin{aligned} \delta_I^{r_4} &= -\Bigg[\tilde{\ell}_i\nu^{i,r_4} + \mathcal{D}_{1;i}\nu^{i,r_4} + \delta_I^j\mathcal{H}_{i,j}\nu^{i,r_4} + \frac{1}{2}\delta_I^j\delta_I^k\nu_{i,j,k}\nu^{i,r_4} + \mathcal{D}_{3;i}\nu^{i,r_4} \\ &\quad + \delta_I^j\mathcal{D}_{2;i,j}\nu^{i,r_4} + \frac{1}{2}\delta_I^j\delta_I^k\mathcal{H}_{i,j,k}\nu^{i,r_4} + \frac{1}{6}\delta_I^j\delta_I^k\delta_I^l\nu_{i,j,k,l}\nu^{i,r_4}\Bigg] + O_p\left(\frac{1}{T^2}\right). \end{aligned} \tag{36}$$

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

$$\begin{aligned} \delta_I^m &= -\tilde{\ell}_{r_1}\nu^{r_1,m} - \mathcal{D}_{1;r_1}\nu^{r_1,m} \\ &\quad + \left(\tilde{\ell}_a\nu^{a,r_2} + \mathcal{D}_{1;a}\nu^{a,r_2} + \delta_I^b\mathcal{H}_{a,b}\nu^{a,r_2} + \frac{1}{2}\delta_I^b\delta_I^c\nu_{a,b,c}\nu^{a,r_2}\right)\mathcal{H}_{r_1,r_2}\nu^{r_1,m} \\ &\quad + \frac{1}{2}\left(\tilde{\ell}_a\nu^{a,r_2} + \mathcal{D}_{1;a}\nu^{a,r_2} + \delta_I^b\mathcal{H}_{a,b}\nu^{a,r_2} + \frac{1}{2}\delta_I^b\delta_I^c\nu_{a,b,c}\nu^{a,r_2}\right)\delta_I^{r_3}\nu_{r_1,r_2,r_3}\nu^{r_1,m} \end{aligned}$$

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

Next, (35) and (36) 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 A.11. ■

Based on these results, the proof of Theorem 5.1 now follows.

Proof (Theorem 5.1). Follows directly from Lemma A.11, 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 (6). ■

Finally, this section ends with the proof of Lemma A.9.

Proof (Lemma A.9). The proof of Lemma A.9 is tedious but straightforward. To save space, proofs for $V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2\left(E_{iT}\right)^{-2}$ 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

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

$$
\begin{aligned}
& -2\frac{V_{r_1}U^2E_{r_2,r_3}[3] + 2VUU_{r_1}E_{r_2,r_3}[3] + VU^2E_{r_1,r_2,r_3}}{E^3} \\
& +6\frac{V_{r_1}U^2E_{r_2}E_{r_3}[3] + 2VUU_{r_1}E_{r_2}E_{r_3}[3]}{E^4} \\
& +6\frac{VU^2E_{r_1,r_2}E_{r_3}[3]}{E^4} - 24\frac{VU^2E_{r_1}E_{r_2}E_{r_3}}{E^5} \\
= \; & O_p\left(\frac{1}{T^{3/2}}\right)
\end{aligned}
$$

Lastly,

$$
\begin{aligned}
\nabla_{\theta\theta\theta\theta}\left[\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^2}\right] = \; & \frac{V_{r_1,r_2,r_3,r_4}U^2 + 2V_{r_1,r_2,r_3}UU_{r_4}[4] + 2VUU_{r_1,r_2,r_3,r_4}}{E^2} \\
& +2\frac{V_{r_1,r_2}U_{r_4}U_{r_3}[6] + V_{r_1,r_2}UU_{r_3,r_4}[6] + V_{r_1}U_{r_3,r_4}U_{r_2}[12]}{E^2} \\
& +2\frac{V_{r_1}UU_{r_2,r_3,r_4}[4] + VU_{r_2,r_3,r_4}U_{r_1}[4] + VU_{r_2,r_3}U_{r_1,r_4}[3]}{E^2} \\
& -2\frac{V_{r_1,r_2,r_3}U^2E_{r_4}[4] + 2V_{r_1,r_2}UU_{r_3}E_{r_4}[12] + 2V_{r_1}U_{r_3}U_{r_2}E_{r_4}[12]}{E^3} \\
& -4\frac{V_{r_1}UU_{r_2,r_3}E_{r_4}[12] + VU_{r_2,r_3}U_{r_1}E_{r_4}[12] + VUU_{r_1,r_2,r_3}E_{r_4}[4]}{E^3} \\
& -2\frac{V_{r_1,r_2}U^2E_{r_3,r_4}[6] + V_{r_1}U^2E_{r_2,r_3,r_4}[4] + VU^2E_{r_1,r_2,r_3,r_4}}{E^3} \\
& -4\frac{V_{r_1}UU_{r_2}E_{r_3,r_4}[12] + VU_{r_2}U_{r_1}E_{r_3,r_4}[6]}{E^3} \\
& -4\frac{VUU_{r_1,r_2}E_{r_3,r_4}[6] + VUU_{r_1}E_{r_2,r_3,r_4}[4]}{E^3} \\
& +6\frac{V_{r_1,r_2}U^2E_{r_3}E_{r_4}[6] + 2V_{r_1}UU_{r_2}E_{r_3}E_{r_4}[12]}{E^4} \\
& +6\frac{2VU_{r_2}U_{r_1}E_{r_3}E_{r_4}[6] + 2VUU_{r_1,r_2}E_{r_3}E_{r_4}[6]}{E^4} \\
& +6\frac{V_{r_1}U^2E_{r_2,r_3}E_{r_4}[12] + 2VUU_{r_1}E_{r_2,r_3}E_{r_4}[12]}{E^4} \\
& +6\frac{VU^2E_{r_1,r_2,r_3}E_{r_4}[4] + VU^2E_{r_1,r_2}E_{r_3,r_4}[3]}{E^4} \\
& -24\frac{V_{r_1}U^2E_{r_2}E_{r_3}E_{r_4}[4] + 2VUU_{r_1}E_{r_2}E_{r_3}E_{r_4}[4]}{E^5} \\
& -24\frac{VU^2E_{r_1,r_2}E_{r_3}E_{r_4}[6]}{E^5} \\
& +120\frac{VU^2E_{r_1}E_{r_2}E_{r_3}E_{r_4}}{E^6},
\end{aligned}
$$

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

A.5 Proof of Theorem 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,

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}[|Z_{iT}|^{2+\delta} |Z_{iT}|^{\varepsilon} |Z_{iT}|^{-\varepsilon}] &< \mathbb{E}_{\mathcal{A}}[|Z_{iT}|^{2+\delta} |Z_{iT}|^{\varepsilon} k^{-\varepsilon}] \\ &= k^{-\varepsilon}\mathbb{E}_{\mathcal{A}}[|Z_{iT}|^{2+\delta} |Z_{iT}|^{\varepsilon}] \\ &\leq k^{-\varepsilon}\mathbb{E}[|Z_{iT}|^{2+\delta+\varepsilon}].\end{aligned}$$

By Assumption 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 $\delta > 0$:

$$\lim_{k\to\infty} \sup_{i,T} \mathbb{E}[|Z_{iT}|^{2+\delta} \mathbf{1}_{(|Z_{iT}|)>k}] = 0. \tag{37}$$

In addition, Assumption 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 6.5(b)-(c) and are directly verified. Finally, Assumption 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,

$$\begin{aligned}\frac{1}{G_N^2}\sum_{g=1}^{G}\frac{1}{L_N^2}\sum\sum_{i,j\in\mathcal{G}_g} Cov(Z_{iT}, Z_{jT}) &= \frac{1}{G_N^2}G_N O\left(\frac{1}{L_N}\right) \\ &= O\left(\frac{1}{G_N L_N}\right) \\ &= O\left(\frac{1}{N}\right),\end{aligned}$$

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 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 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 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 6.5(a). Then,

$$\begin{aligned}\frac{1}{N^2}\sum^{G_N}\sum^{G_N}_{g\neq h}\sum_{i\in\mathcal{G}_g}\sum_{j\in\mathcal{G}_h} Cov(Z_{iT}, Z_{jT}) &\leq \frac{1}{N^2}\sum^{G_N}\sum^{G_N}_{g\neq h} O\left(L_N^{(d-1)/d}\right) \\ &= \frac{1}{N^2} O\left(G_N^2 L_N^{(d-1)/d}\right) \\ &= \frac{1}{N^2} O\left(N^2 L_N^{-(d+1)/d}\right) \\ &= O\left(L_N^{-(d+1)/d}\right),\end{aligned}$$

which proves the second result. ■

B Details of the Simulation Analysis

In order to numerically evaluate the integrated likelihood, $\pi_i(\lambda_i|\theta)$ is evaluated at 15 equally distant points on a grid between $(0.05)^2/252$ and $(0.87)^2/252$, which are the daily variances corresponding to annual volatilities of 5% and 87%. These boundaries were chosen randomly and different choices can be used as long as the interval contains the true parameter values, which, by design, take on a value between $(0.15)^2/252$ and $(0.80)^2/252$. Similarly, the integral can be calculated using a larger number of draws within the interval. The reason for choosing 15 values for this purpose is to keep the computation time at a reasonable length.

Iterated updating is done as follows: first some consistent estimates of α and β have to be obtained. The composite likelihood method is used here for this purpose. Define these initial estimates as $\hat{\theta}^{(1)} = (\hat{\alpha}, \hat{\beta})$. Then, $\hat{\theta}^{(1)}$ is used to calculate the value of the prior values at each λ_i, $\pi_i(\lambda_i|\hat{\theta}^{(1)})$.Define each value of λ_i that is used to evaluate the integral as $\lambda_i^{(j)}$, $j = 1, ..., 15$. This gives

$$\pi_i(\lambda_i^{(j)}|\hat{\theta}^{(1)}) \quad \text{for} \quad j = 1, ..., 15.$$

In the next step, these priors are used to calculate the integrated likelihood,

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

Note that $\hat{\theta}^{(1)}$ does not vary in this step. The integrated composite likelihood estimator of θ_0 is then given by

$$\hat{\theta}_{IL} = \arg\max_{\theta} \frac{1}{NT} \sum_{i=1}^{N} \ln \int \exp\left[T\ell_{iT}\left(\theta, \lambda_i\right)\right] \pi_i(\lambda_i|\hat{\theta}^{(1)}) d\lambda_i.$$

Define now $\hat{\theta}^{(2)} = \hat{\theta}_{IL}$. In the next step, $\hat{\theta}^{(2)}$ is used to calculate the priors and a new estimate of θ_0, $\hat{\theta}^{(3)}$, is obtained by maximising the new integrated likelihood, $(NT)^{-1} \sum_{i=1}^{N} \ln \int \exp\left[T\ell_{iT}\left(\theta, \lambda_i\right)\right] \pi_i(\lambda_i|\hat{\theta}^{(2)}) d\lambda_i$. This procedure continues until $\theta^{(n)} \approx \theta^{(n-1)}$. The minimum necessary number of iterations to attain convergence will depend on the model and estimation method at hand. In this study, optimisation continues until either $n = 10$ or $\theta^{(n)} - \theta^{(n-1)} < (0.003, 0.01)'$.Again, the choice of $(0.003, 0.01)'$ as a cut-off point here is for illustration purposes and is not determined by a specific criterion. This, along with the specific numerical integration method, the density of the grid for λ_i and the number of iterations have to be determined depending on the model and data at hand.

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.

Andersen, E. B. (1970): "Properties of Conditional Maximum-Likelihood Estimators," *Journal of the Royal Statistical Society, Series B*, 32, 283–301.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2009): "Parametric and Nonparametric Measurement of Volatility," in *Handbook of Financial Econometrics*, ed. by Y. Aït-Sahalia, and L. P. Hansen, pp. 67–137. North-Holland, Amsterdam.

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.

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. (2012): "Fixed-Effects Dynamic Panel Models, A Factor Analytical Method," *Econometrica*, forthcoming.

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.

Bailey, N., G. Kapetanios, and M. H. Pesaran (2012): "Exponent of Cross-Sectional Dependence: Estimation and Inference," 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., P. R. Hansen, A. Lunde, and N. Shephard (2008): "Designing Realised Kernels to Measure the ex-post Variation of Equity Prices in the Presence of Noise," *Econometrica*, 76, 1481–1536.

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.

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.

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.

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.

Bradley, R. C. (2005): "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions," *Probability Surveys*, 2, 107–144.

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.

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.

Conley, T. G. (1999): "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 92, 1–45.

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.

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.

Doukhan, P. (1994): *Mixing, Properties and Examples.* Springer.

Engle, R. F. (1982): "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of the United Kingdom Inflation," *Econometrica*, 50, 987–1007.

——— (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 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.

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.

Fernández-Val, I., and F. Vella (2009): "Bias Correction for Two-Step Fixed Effects Panel Data Estimators," working paper.

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.

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.

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.

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.

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.

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.

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.

Manova, K., and Z. Zhang (2012): "Export Prices across Firms and Destinations," *Quarterly Journal of Economics*, 127, 379–436.

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.

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 K. D. West (1987): "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

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.

Pakel, C., N. Shephard, and K. K. Sheppard (2011): "Nuisance Parameters, Composite Likelihoods and a Panel of GARCH Models," *Statistica Sinica*, 21, 307–329.

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): "Evaluating Volatility and Correlation Forecasts," in *Handbook of Financial Time Series*, ed. by T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch, pp. 801–838. Springer.

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.

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.

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.

Teo, M. (2009): “Geography of Hedge Funds,” *Review of Financial Studies*, 22, 3531–3561.

Teräsvirta, T. (2009): “An Introduction to Univariate GARCH Models,” in *Handbook of Financial Time Series*, ed. by T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch, pp. 17–42. Springer-Verlag.

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.

Varin, C., N. Reid, and D. Firth (2011): “An Overview of Composite Likelihood Methods,” *Statistica Sinica*, 21, 5–42.

Wansbeek, T., and E. Meijer (2000): *Measurment Error and Latent Variables in Econometrics.* North-Holland.

West, K. D. (1996): “Asymptotic Inference about Predictive Ability,” *Econometrica*, 64, 1067–1084.

White, H. (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.

	$\alpha = 0.05,\quad \beta = 0.93,\quad \alpha+\beta = 0.98$											
	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 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 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 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 cross-section dependence) for varying T and N where true parameter values are given by $\alpha = 0.05$ and $\beta = 0.93$.

$\alpha = 0.05, \quad \beta = 0.93, \quad \alpha + \beta = 0.98$

	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	.046	.936	.982
200	.039	.914	.954	.047	.932	.979	.042	.940	.982
150	.033	.906	.939	.048	.929	.977	.040	.944	.984
100	.018	.895	.913	.049	.924	.973	.033	.950	.983
75	.003	.892	.895	.048	.921	.970	.026	.954	.980
	N=50								
400	.045	.925	.971	.047	.933	.980	.046	.935	.982
200	.039	.914	.953	.047	.932	.979	.042	.940	.982
150	.034	.905	.938	.048	.929	.977	.040	.943	.983
100	.018	.895	.913	.048	.924	.972	.033	.949	.982
75	.005	.884	.889	.048	.921	.969	.025	.953	.979
	N=25								
400	.045	.926	.970	.047	.933	.980	.046	.936	.981
200	.038	.915	.953	.047	.932	.979	.041	.941	.982
150	.033	.904	.937	.048	.929	.976	.039	.942	.981
100	.019	.891	.911	.048	.924	.972	.032	.947	.979
75	.007	.874	.881	.049	.920	.969	.025	.951	.976

Table 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 $\beta = 0.93$.

T	CL		InCL		IPCL		CL		InCL		IPCL	
	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$
	N=100											
400	.002	.004	.002	.003	.002	.004	.005	.006	.004	.005	.005	.007
200	.004	.006	.003	.005	.004	.007	.012	.017	.004	.005	.008	.012
150	.005	.008	.004	.006	.004	.008	.018	.025	.004	.006	.011	.016
100	.007	.011	.005	.007	.006	.008	.033	.037	.005	.009	.018	.021
75	.005	.020	.005	.008	.006	.007	.047	.043	.006	.012	.025	.025
	N=50											
400	.003	.005	.003	.005	.003	.005	.006	.007	.004	.006	.005	.008
200	.005	.008	.004	.006	.005	.009	.012	.018	.005	.007	.009	.014
150	.006	.012	.005	.007	.006	.011	.018	.028	.005	.007	.012	.017
100	.009	.016	.006	.010	.007	.012	.033	.039	.007	.011	.019	.022
75	.007	.036	.008	.011	.009	.012	.046	.059	.008	.015	.026	.026
	N=25											
400	.005	.007	.005	.007	.005	.007	.007	.009	.005	.008	.007	.009
200	.008	.013	.006	.009	.007	.013	.014	.020	.007	.010	.011	.017
150	.009	.016	.007	.011	.008	.015	.019	.030	.007	.011	.014	.019
100	.012	.032	.009	.013	.011	.017	.033	.050	.009	.015	.021	.024
75	.010	.059	.011	.017	.014	.020	.044	.081	.011	.020	.028	.029

Table 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 $\beta = 0.93$.

Stock	QML vs CL		CL vs IPCL		QML vs IPCL	
	t-stat	Result	t-stat	Result	t-stat	Result
Alcoa	0.993	-	2.668	IPCL	2.217	IPCL
American Express	1.451	-	3.168	IPCL	3.444	IPCL
Bank of America	1.109	-	3.862	IPCL	3.271	IPCL
Du Pont	2.288	CL	2.587	IPCL	3.030	IPCL
General Electric	0.960	-	2.142	IPCL	2.529	IPCL
IBM	1.815	-	2.067	IPCL	2.493	IPCL
Coca Cola	2.870	CL	-0.932	-	1.541	-
Microsoft	2.755	CL	-0.427	-	1.879	-
Exxon Mobil	1.404	-	1.920	-	2.404	IPCL

Table 5: Giacomini-White test results for GARCH panels. The level of significance is 5%. Results for the following comparisons are reported: quasi maximum likelihood vs composite likelihood (columns $2-3$), composite likelihood vs integrated composite likelihood (columns $4-5$) and quasi maximum likelihood vs integrated composite likelihood (columns $6-7$). Loss functions are based on realised covariance, RV_{it}. 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. 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 (15).

Strategy	$T = 150$			$T = 175$			$T = 207$		
	#	$\hat{\alpha}$	$\hat{\beta}$	#	$\hat{\alpha}$	$\hat{\beta}$	#	$\hat{\alpha}$	$\hat{\beta}$
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 6: Integrated 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 $5-8$ and (*iii*) February 1994 - April 2011 (207 time-series observations) given in columns $8-10$. Number of funds included in the analysis given in the '#' column.

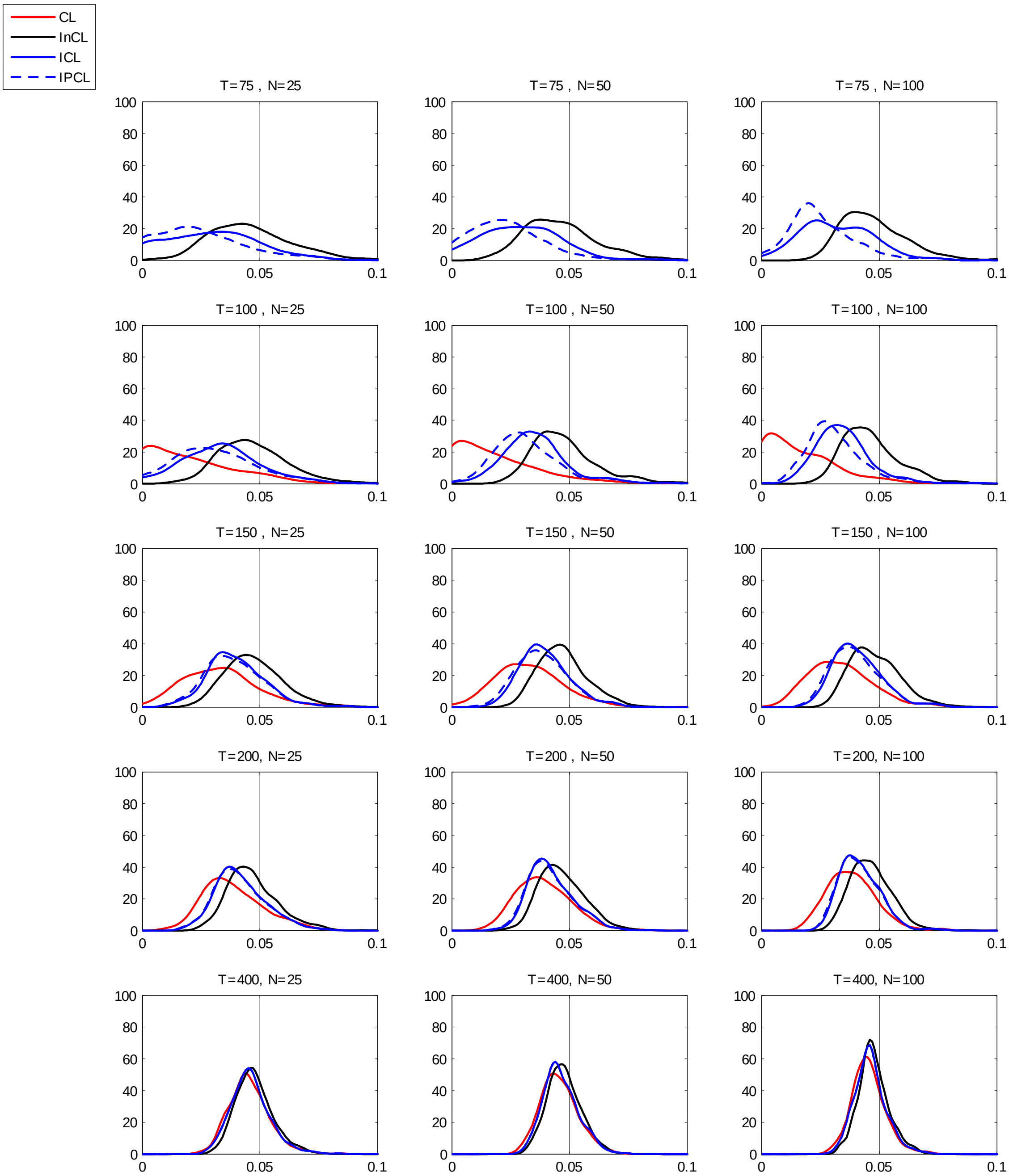

Figure 1: Sample distributions of $\hat{\alpha}$ using the Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL), and Integrated Pseudo CL (IPCL). The vertical line is drawn at the true parameter value. Based on 500 replications under cross-sectional dependence where $(\alpha, \beta) = (0.05, 0.93)$.

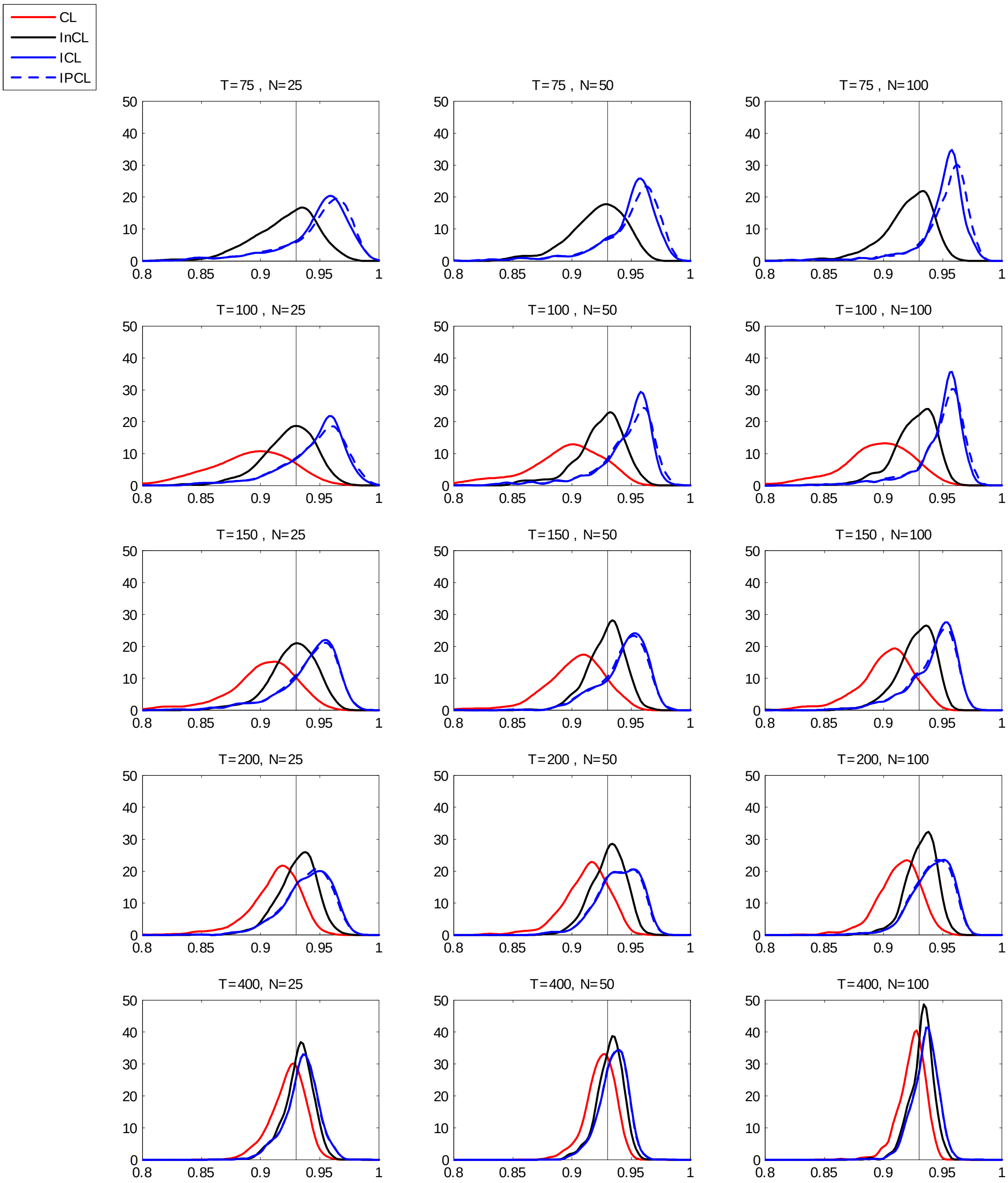

Figure 2: Sample distributions of $\hat{\beta}$ using the Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL), and Integrated Pseudo CL (IPCL). The vertical line is drawn at the true parameter value. Based on 500 replications under cross-sectional dependence where $(\alpha, \beta) = (0.05, 0.93)$.

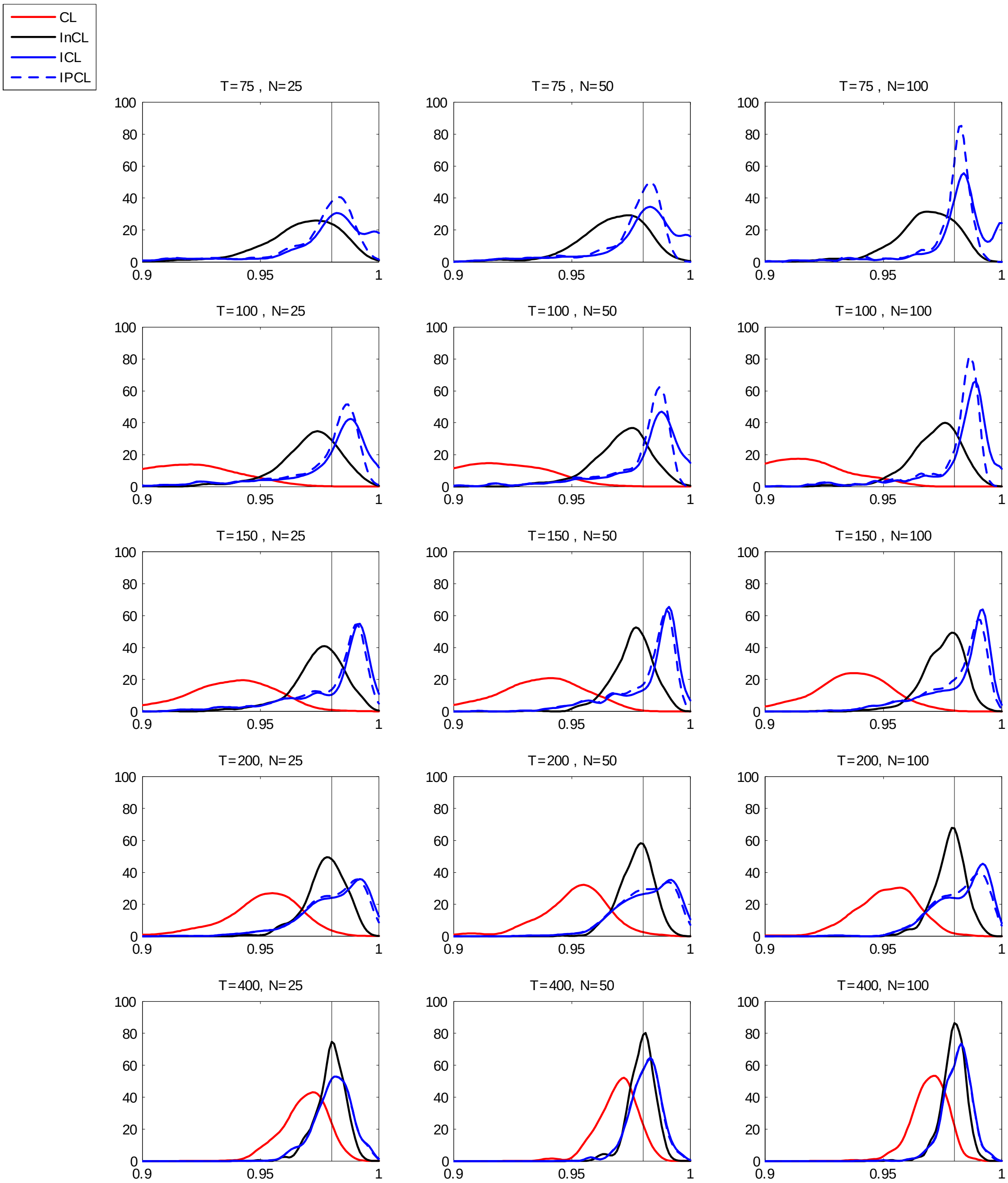

Figure 3: Sample distributions of $\hat{\alpha} + \hat{\beta}$ using the Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL), and Integrated Pseudo CL (IPCL). The vertical line is drawn at the true parameter value. Based on 500 replications under cross-sectional dependence where $(\alpha, \beta) = (0.05, 0.93)$.

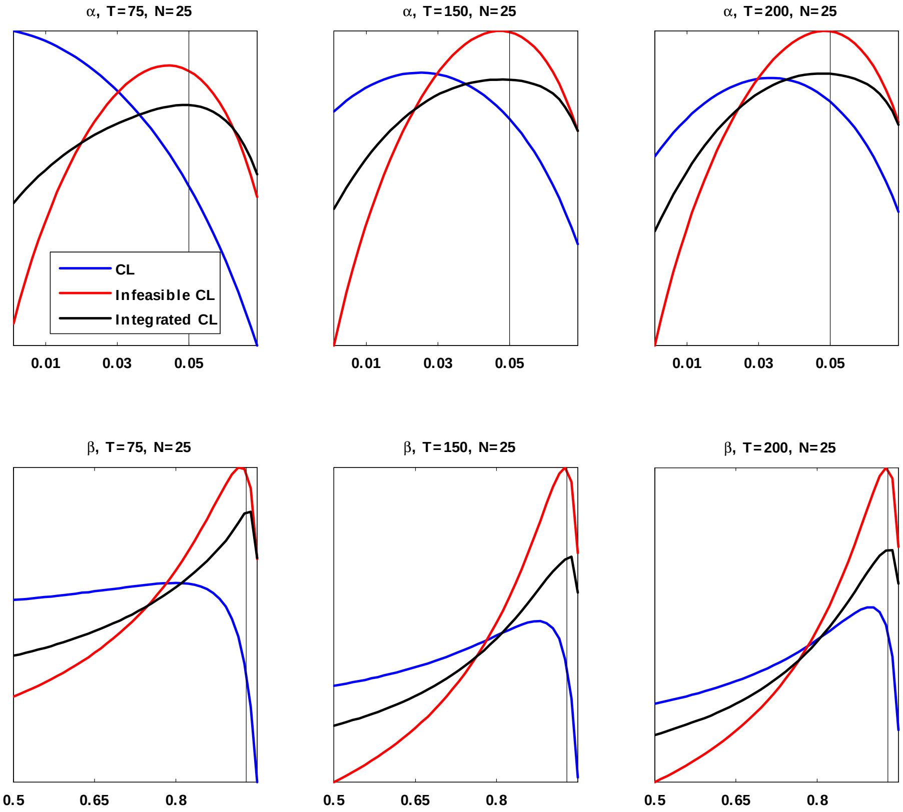

Figure 4: Average likelihood plots for α and β. Based on likelihood averages over 500 replications (with cross-sectional dependence). In the upper panel, β is fixed at 0.93 while the lower panel is based on $\alpha = 0.05$. CL is evaluated at the sample estimates of λ_i, while Infeasible CL is evaluated at the true values of λ_i. Integrated CL is calculated using prior (P1) where 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 $\beta = 0.93$.

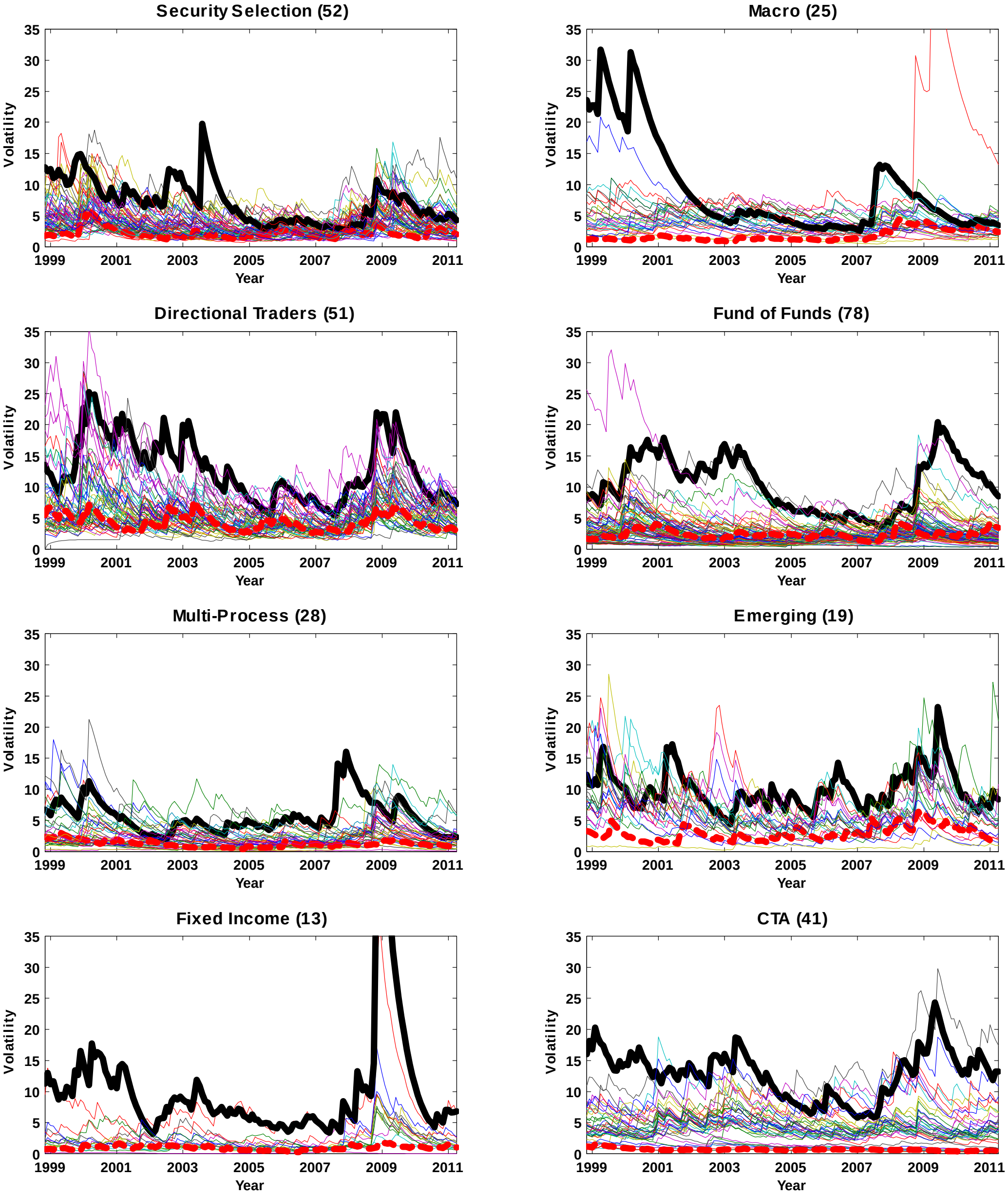

Figure 5: Conditional volatility plots. Based on parameters estimates by the integrated likelihood method using panels of funds that have reported non-zero returns between November 1998 and April 2011 (150 observations). Number of funds in each strategy-panel is given in parentheses. Random examples of high-volatility funds are given by thick solid lines, while the thick broken lines belong to random examples of low volatility funds.

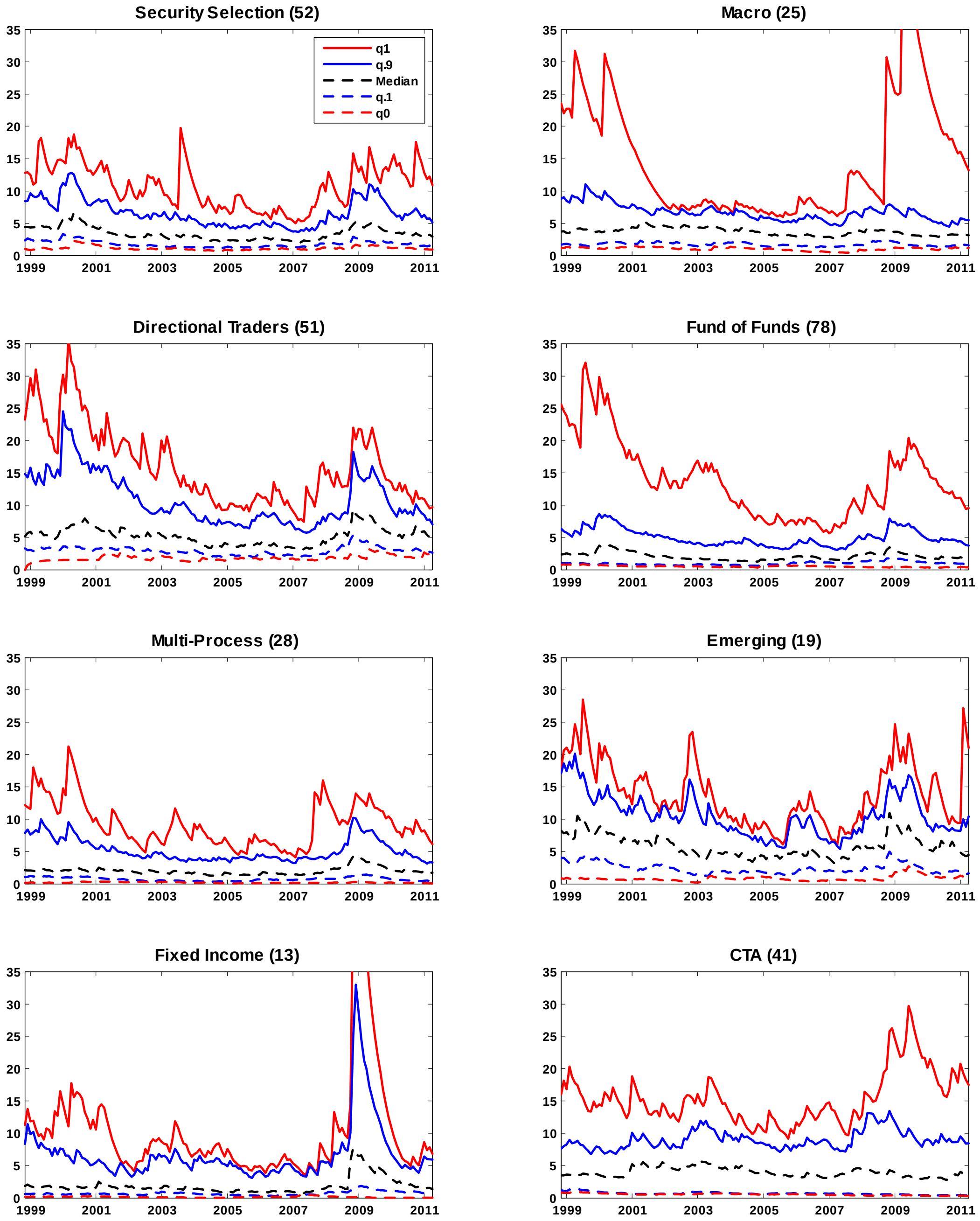

Figure 6: Plots of the 0%, 10%, 50% (median), 90% and 100% quantiles of the sample distribution of volatility across funds. Based on fitted conditional volatilities displayed in Figure 5. Number of funds in each strategy is given in parentheses.

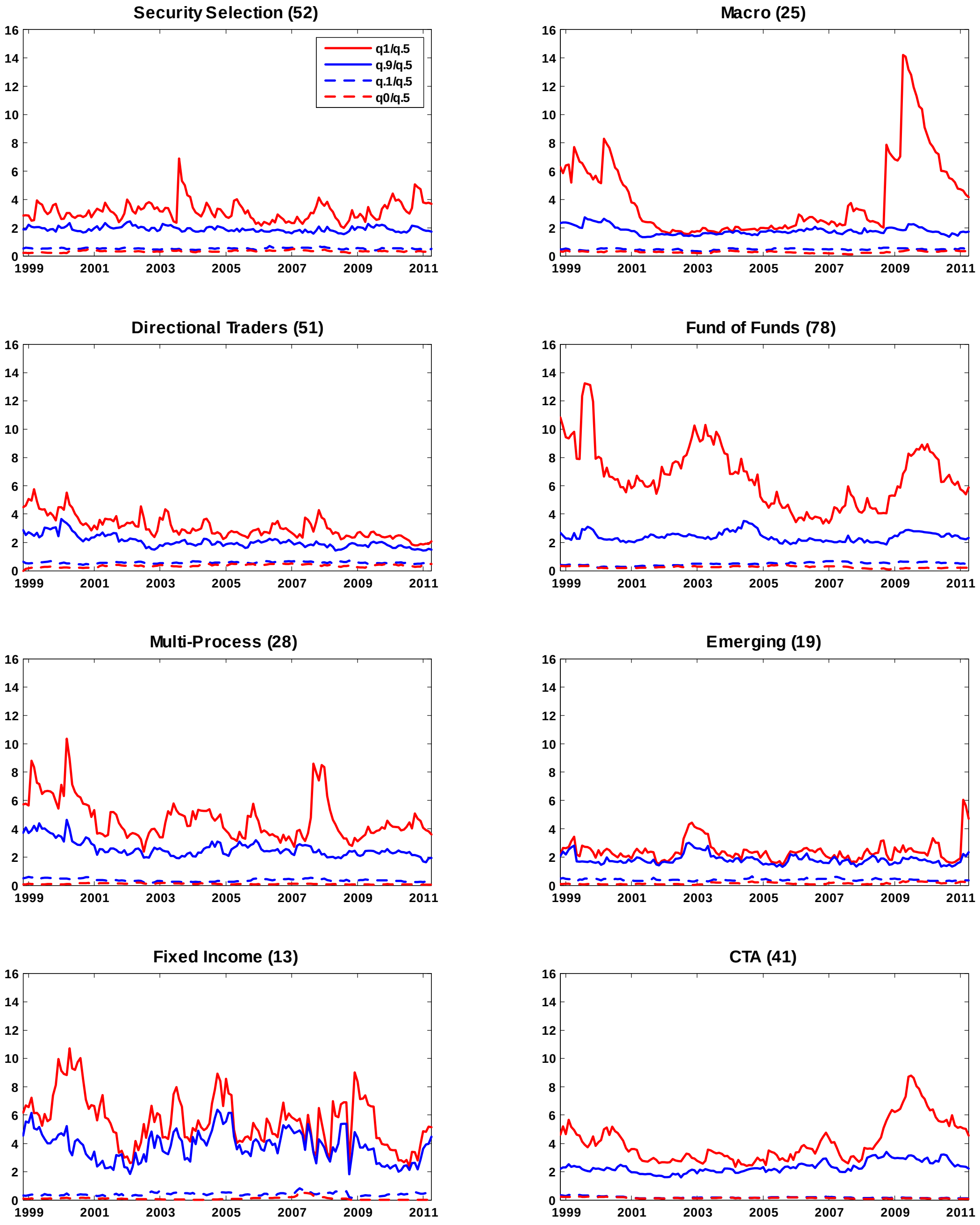

Figure 7: Plots of 0%, 10%, 90% and 100% quantiles (normalised by the median) of the sample distribution of volatility across funds. Based on fitted conditional volatilities displayed in Figure 5. Number of funds in each strategy is given in parentheses.