Bias Reduction under Dependence, in a Nonlinear and Dynamic Panel Setting: the Case of GARCH Panels*

Cavit Pakel†
Department of Economics & Oxford-Man Institute
University of Oxford

December 6, 2011

Job Market Paper

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. 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; however, it inflates estimator variance. Finally, I consider two empirical illustrations; an analysis of hedge fund volatility characteristics and a test of predictive ability using stock volatility forecasts.

*This paper has previously been circulated under the title "Bias Reduction in GARCH Panels, with an Application to Hedge Fund Volatility." 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**: Oxford-Man Institute, Eagle House, Walton Well Road, Oxford, OX2 6ED, United Kingdom. **Email address**: cavit.pakel@economics.ox.ac.uk.

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 an unbiased estimator of A, say $\hat{A}$, exists, then $\tilde{\theta} = \hat{\theta} - \hat{A}/T$ will be a first-order unbiased estimator, since $\mathbb{E}[\tilde{\theta} - \theta_0] = O(1/T^2)$. For moderate T, the remaining $O(1/T^2)$ 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).

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, 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 the extreme case of $\sqrt{T}$- rather than $\sqrt{NT}$-convergence. In other words, cross-section dependence is taken to be so strong that cross-sectional variation has no contribution to the speed of convergence. Under this worst-case scenario, 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,

[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).

it has to be corrected for separately. Of course, this type of assumption on dependence might possibly be too extreme in reality. It is shown by an intuitive argument that, if the cross-section dimension is allowed to contribute to convergence at a mild rate, the cross-section dependence bias becomes $O(1/T^{3/2})$.Then, the $O(1/T)$ bias becomes identical to the one characterised by Arellano and Bonhomme (2009) and their robust priors can be used to reduce the magnitude of bias to $O(1/T^{3/2})$. The analysis in this part is the main contribution of this study to the panel data literature. Importanly, the analysis of first-order bias under cross-section dependence has not been considered in the analytical bias reduction literature before.

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[4], can be employed; but this is beyond the scope of this 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 can also shed light on the properties of estimators under neglected cross-section dependence.

Estimation of financial volatility in a panel (rather than the traditional time-series) setting is the application of interest in this paper and can therefore be considered as an extended example. Volatility modelling is based on the Generalised Autoregressive Conditional Heteroskedasticity (GARCH) type models due to Engle (1982) and Bollerslev (1986), which implies a nonlinear dynamic panel setting. It must be underlined that all theoretical results are given in terms of likelihood-based general expressions and are not model specific.[5] In fact, few assumptions that are specific to the GARCH model are made and these can be modified to accommodate other models. Therefore, the bias characterisations and asymptotic expansions derived in this paper do potentially apply to a wider array of nonlinear dynamic panel models and provide important insights into bias reduction under cross-section dependence.[6]

[4]For recent important examples of this literature, see, among others, Bai and Ng (2002, 2004), Phillips and Sul (2003), Pesaran (2006), Bai (2009), Chudik Pesaran and Tosetti (2011), Kapetanios Pesaran and Yamagata (2011) and Pesaran and Tosetti (2011).

[5]Indeed, for the GARCH case this is a necessity rather than a luxury, as closed-form expressions for likelihood derivatives do not exist for the GARCH model.

[6]Some careful thinking might be required on a case-by-case basis. For example, the quasi maximum like-

Following the theoretical analysis, I conduct a Monte Carlo study of panel volatility estimation. This reveals that a substantial portion of bias is removed with as little as 150-200 time-series observations. This is a significant improvement, as consistent GARCH estimation by standard time-series approaches requires around 1,000-1,500 observations. This is the main contribution of this paper to the financial econometrics literature. 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 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 hedge fund volatility using a consolidated database. This dataset is 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. As GARCH analysis of hedge fund returns has hitherto been virtually impossible under standard methods, this empirical illustration is another contribution to 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 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 and main theoretical results are given in Section 3. Section 4 provides a detailed simulation analysis 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 5. Section 6 concludes. Proofs and additional discussions are given in the Appendix.

lihood theory for GARCH is well-established (Bollerslev and Wooldridge (1992)), although for a different nonlinear dynamic model there may be issues.

[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.

2 Main Concepts and Notation

The next subsection outlines the main points of panel GARCH estimation. After that, the analysis proceeds on the basis of general likelihood-based terms.

2.1 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.[8] 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.[9] 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

[8] 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.

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

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 volatilites will evolve according 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

$$\begin{aligned} \varepsilon_{it} &= \sigma_{it}\eta_{it}, \quad \eta_{it} \sim F, \quad \mathbb{E}[\eta_{it}] = 0, \quad Var(\eta_{it}) = 1, && (2) \\ \sigma^2_{it} &= \lambda_i (1 - \alpha - \beta) + \alpha\varepsilon^2_{i,t-1} + \beta\sigma^2_{i,t-1}, && (3) \\ \lambda_i &> 0 \ \forall i; \quad \alpha, \beta \geq 0 \quad \text{and} \quad \alpha + \beta < 1, && (4) \end{aligned}$$

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 (3) implies that $\mathbb{E}[y^2_{it}] = \lambda_i$.[10] Therefore, a simple method of moments estimator for λ_i is provided by

$$\tilde{\lambda}_i = T^{-1} \sum_{t=1}^{T} y^2_{it}. \quad (5)$$

[10] 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.

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).[11] Hence, the composite likelihood method offers a convenient way of pooling information, while keeping the computational burden at a minimum.[12]

2.2 Likelihood Concepts and Notation

In what follows, λ_{i0} and $\theta_0 = (\alpha_0, \beta_0)$ are the true parameter values. Define

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

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

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

Assuming that the expectations exist, centred likelihood derivatives with respect to λ_i are defined as

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

[11] 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 results.

[12] 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.

The three likelihood concepts under the focus of this study are the concentrated, integrated and target likelihoods. The most familiar of these is the concentrated likelihood, given by

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

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

The main idea is to center the likelihood function at the likelihood estimator for λ_i, for some given value of θ. In large samples this estimator has good properties. 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 bias) in $\hat{\lambda}_i(\theta)$ is accumulated across strata, and contaminates the estimation of θ_0 (see, for example, McCullagh and Tibshirani (1990) and Sartori (2003)). Consequently, $\hat{\theta}$ is inconsistent for θ_0. More formally, $\hat{\theta}_T = \arg\max_\theta p\lim_{N\to\infty}(NT)^{-1}\ell_{NT}(\theta, \hat{\lambda}_i(\theta)) \neq \theta_0$.This is the well-known incidental parameter issue (Neyman and Scott (1948)).

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

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

The choice of weights/priors, $\pi_i(\lambda_i|\theta)$, is central to successfully removing the incidental parameter bias. Following Arellano and Bonhomme (2009), who investigated 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. Note that the specification of the robust priors depend entirely on the characterisation of the incidental parameter bias. Therefore, no subjective prior has to be specified and one can refer to the robust prior as a robust weighting scheme.

Bias correction by integrated likelihood is a common approach in the Bayesian literature. Severini (1999, 2007) analyses the links between the use of integrated likelihood in the Bayesian literature and key contributions of the frequentist literature. In particular, Severini (1999) shows that the adjusted profile likelihood function (see 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 (see Barndorff-Nielsen (1983)) functions are locally equivalent to second order, the latter is asymptotically Bayesian to second order. The adjusted and modified profile log-likelihood functions are important contributions in the frequentist literature and therefore, these observations suggest an important link between the frequentist and Bayesian approaches. Moreover, Severini (2007) analyses the issue of selecting appropriate priors that would ensure that integrated like-

lihood is appropriate under the frequentist approach, as well. In a recent work, Severini (2010) analyses the integrated log-likelihood ratio statistic and compares it to the standard log-likelihood ratio statistic. All these contributions support the use of integrated likelihood within the frequentist framework.

The benchmark likelihood function 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 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.[13] 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 theoretical benchmark to analyse theoretical properties of the incidental parameter bias. Note that, under the iid assumption, it can be shown that $\bar{\lambda}_{iT}(\theta)$ is equivalent to the large-T consistent concentrated likelihood estimator of λ_{i0}.[14] Hence, this gives the target likelihood the more intuitive meaning of a likelihood free from the incidental parameter bias.

In what follows, the following notational convention will be used for sake of conciseness: whenever a likelihood function is evaluated at $(\theta, \bar{\lambda}_i(\theta))$, the argument will be omitted. Moreover, if the likelihood is evaluated at $(\psi, \bar{\lambda}_i(\psi))$ for some ψ, then the likelihood is written as a function of ψ only. Specifically,

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

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. Moreover, $\hat{\lambda}_i$ and $\bar{\lambda}_i$ are used as shorthand for $\hat{\lambda}_{iT}(\theta)$ and $\bar{\lambda}_i(\theta)$; therefore, the dependence of $\bar{\lambda}_{iT}(\theta)$ on T will be implicit. 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

[13] See Severini and Wong (1992) and Severini (2000, Chapter 4). A lucid discussion is given by Pace and Salvan (2006).

[14] To see this, notice that

$$\bar{\lambda}_{iT}(\theta) = \arg\max_{\lambda_i} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{it}(\theta, \lambda_i)] = \arg\max_{\lambda_i} p\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \ell_{it}(\theta, \lambda_i) = \arg\max_{\lambda_i} \mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{it}(\theta, \lambda_i)],$$

where $\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{it}(\theta, \lambda_i)] = \mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{is}(\theta, \lambda_i)]$ $\forall t, s$ due to the iid assumption. Then, $\bar{\lambda}_i(\theta)$ does not depend on T and the target likelihood $\ell_{it}(\theta, \bar{\lambda}_i(\theta))$ is clearly free from the incidental parameter bias issue as $\bar{\lambda}_i(\theta)$ is the same as the estimator when $T \to \infty$.

true parameter values, respectively.

3 Bias Correction by Integrated Likelihood

3.1 Main Results

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 a large-T expansion of the bias of the integrated likelihood with respect to the concentrated likelihood is derived. This provides a characterisation of the bias in a given stratum. Based on this expression, the robust prior that removes the first-order bias in the score function will be specified.[15] Arellano and Bonhomme (2009) have already studied the time-series and cross-section independence case, so results presented here extend their analysis to time-series dependence. Next, cross-sectional dynamics will be incorporated into the analysis by obtaining a large-T, large-N double-asymptotic expansion. The central assumption here is that cross-section dependence is such that, cross-sectional information does not contribute to the Central Limit Theorem, implying $\sqrt{T}$-convergence. This is the worst-possible case, and an intuitive discussion on how a relaxation of this assumption would change the bias characteristics is provided at the end of this Section.

The Assumptions are given next.

Assumption 3.1 *The support of $\pi_i(\lambda_i|\theta)$ contains an open neighbourhood of the true parameters λ_{i0} and θ_0.*

Assumption 3.2 *θ and λ_i belong to the interior of Θ and Λ_i, respectively, where Θ and Λ_i are compact parameter spaces.*

Assumption 3.3 *$N, T \to \infty$ jointly and, for $0 < c < \infty$, $N/T \to c$.*

Assumption 3.4 $\sup_{\theta\in\Theta} \sup_i \left|\hat{\lambda}_i(\theta) - \bar{\lambda}_{iT}(\theta)\right| = O_p(T^{-1/2})$.

Assumption 3.5 *For each $\theta \in \Theta$, $\ell_{iT}(\theta, \lambda_i)$ has a unique maximum at $\hat{\lambda}_i(\theta)$ for all i.*

Assumption 3.6 *For $0 \leq m \leq 3$ and $0 \leq n \leq 4$,*

$$\sup_{\theta\in\Theta} \sup_{i,t} \mathbb{E}\left|\frac{d^{(m+n)}}{d\lambda_i^m d\theta_{j_1}...d\theta_{j_n}} \ell_{it}(\theta, \bar{\lambda}_i(\theta))\right| = O(1),$$

where $j_1, ..., j_n \in \{1, 2\}$, $\theta_1 = \alpha$ and $\theta_2 = \beta$.

[15] As mentioned by Arellano and Bonhomme (2009), removing the bias of the score is equivalent to removing the bias of the estimator.

Assumption 3.7 *For* $0 \leq k \leq 4$ *and* $2 \leq j \leq 4$,

$$\sup_{\theta\in\Theta}\sup_{i} Var\left[\nabla_{\theta^{(k)}}\ell^{\lambda}_{iT}(\theta,\bar{\lambda}_i(\theta))\right] = O\left(\frac{1}{T}\right),$$
$$\sup_{\theta\in\Theta}\sup_{i} Var\left[\nabla_{\theta^{(k)}}V^{\lambda\lambda}_{iT}(\theta,\bar{\lambda}_i(\theta))\right] = O\left(\frac{1}{T}\right),$$
$$\sup_{\theta\in\Theta} Var[\nabla_{\theta}\ell_{NT}(\theta,\bar{\lambda}_i(\theta))] = O\left(\frac{1}{T}\right),$$
$$\sup_{\theta\in\Theta} Var\left(\nabla_{\theta^{(j)}}\ell_{NT}(\theta,\bar{\lambda}_i(\theta)) - \mathbb{E}[\nabla_{\theta^{(j)}}\ell_{NT}(\theta,\bar{\lambda}_i(\theta))]\right) = O\left(\frac{1}{T}\right).$$

Assumption 3.8 *For* $0 \leq k \leq 4$,

$$0 < \sup_{\theta\in\Theta}\sup_{i}\left|\nabla_{\theta^{(k)}}\mathbb{E}[\ell^{\lambda\lambda}_{iT}(\theta,\bar{\lambda}_i(\theta))]\right| = O(1),$$
$$\sup_{\theta\in\Theta}\sup_{i}\left|\nabla_{\theta^{(k)}}\mathbb{E}[\ell^{\lambda\lambda\lambda}_{iT}(\theta,\bar{\lambda}_i(\theta))]\right| = O(1),$$
$$\sup_{\theta\in\Theta,\lambda_i\in\Lambda_i}\sup_{i}\left|\nabla_{\theta^{(k)}}\frac{d^{(n)}\ln\pi_i(\lambda_i|\theta)}{d\lambda_i^n}\right| = O(1), \qquad \textit{where } n\in\{0,1\}.$$

Assumption 3.1 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 precluded. Assumption 3.2 is a standard regularity condition on the parameter space. Assumption 3.3 implies that N and T converge to infinity at the same rate, hence a large-T, large-N setting. This is appropriate for financial panels where T and N are of comparable magnitudes. Assumption 3.4 controls the convergence rate of $\hat{\lambda}_i(\theta)$ to $\bar{\lambda}_{iT}(\theta)$, the benchmark "least-favourable" estimator. Assumption 3.5 is required for the existence of a Laplace approximation to the integrated likelihood function and is a mild condition. The conditions set by Assumptions 3.6, 3.7 and 3.8 are central to the double asymptotic expansions. The first of these puts a uniform bound on several likelihood derivatives that appear in the asymptotic expansions. This implies that the likelihood is smooth. Observe that all expressions in the variance operator in Assumption 3.7 are zero mean for any θ. Then, the uniform bound of $O(T^{-1})$ implies that all these zero-mean terms are $O_p(T^{-1/2})$. Intuitively, this means that all the centred likelihood derivatives under consideration exhibit convergence in distribution across T but not N.[16] More primitive conditions ensuring these Assumptions can be specified. This is not pursued here and instead left for future research.

It must be stressed that the following theoretical results are based on general, likelihood-based expressions. In fact, this is a necessity for the GARCH model as closed form expressions for the likelihood derivatives of the GARCH model do not exist. Therefore, this study has a more general scope than GARCH and the following theoretical results can

[16] This is the same in spirit as the setting proposed in both Engle, Shephard and Sheppard (2008) and Pakel, Shephard and Sheppard (2011).

be applied to any model satisfying the underlying assumptions. To put it differently, the analysis in this section is mainly a contribution to the panel data literature.

Finally, Assumptions have to be made on the time-series dependence of data. This is central to obtaining large-T asymptotic expansions for the incidental parameter bias in each stratum. Hahn and Kuersteiner (2011) provide a general set of conditions for the time-series dependence and cross-section independence case. In particular, their Condition 3 limits serial dependence sufficiently. A similar Assumption can be used here, as well, to use the bias-reduction method in a more general context other than GARCH estimation. For the GARCH model, explicit assumptions which ensure mixing-type serial dependence are as follows.

Assumption 3.9 *The GARCH(1,1) process outlined in (2) to (4) satisfies*

$$-\infty < \mathbb{E}[\log(\beta + \alpha\eta_{it}^2)] < 0 \qquad \forall i, t.$$

Assumption 3.10 *The distribution of η_{it} is such that η_{i0} is absolutely continuous with strictly positive Lebesgue density in a neighbourhood of zero. Moreover, there exists some $0 < p < \infty$ satisfying $\mathbb{E}\,|\eta_{i0}|^p < \infty$.*

Assumptions 3.9 and 3.10, due to a result by Boussama (1998), imply that both σ_{it}^2 and ε_{it}^2 are β-mixing with geometric rate. It can be shown that these assumptions ensure existence of asymptotic convergence results for likelihood derivatives (see Francq and Zakoïan (2010)). Intuitively, this follows from the property that functions of mixing processes are mixing, as well. Now, the likelihood is a function of σ_{it}^2 and ε_{it}^2; therefore, it is also mixing. This property will also be retained by the derivatives. Notice that, since $\alpha + \beta < 1$ (as assumed in (4)), Assumption 3.9 holds, by Jensen's Inequality. Assumption 3.10 is a technical assumption on the distribution of the innovation process.[17]

The first main result of this study characterises the bias of the integrated likelihood and reveals that introduction of time-series dependence leads to an extra bias term of order $O(T^{-3/2})$. This is an extension to the corresponding result derived by Arellano and Bonhomme (2009), under serial independence.

Proposition 3.1

$$\mathbb{E}_{\theta_0,\lambda_{i0}}\left[\ell_{iT}^I(\theta) - \bar{\ell}_{iT}(\theta)\right] = C + \frac{\mathcal{B}_{iT}^{(1)}(\theta)}{T} + \frac{\mathcal{B}_{iT}^{(2)}(\theta)}{T^{3/2}} + O\left(\frac{1}{T^2}\right), \tag{6}$$

[17] See Francq and Zakoïan (2006) for more general versions of Assumption 3.10. Their results, combined with Theorem 3.7 of Bradley (2005) also show the β-mixing property of σ_{it}^2 and ε_{it}^2. For a more detailed discussion of asymptotic properties of GARCH processes, see Lindner (2009) and Francq and Zakoïan (2010).

where

$$\begin{aligned}\mathcal{B}_{iT}^{(1)}(\theta) &= \frac{1}{2}\{\mathbb{E}_{\theta_0,\lambda_{i0}}[-\ell_{iT}^{\lambda\lambda}]\}^{-1}\mathbb{E}_{\theta_0,\lambda_{i0}}[T(\ell_{iT}^{\lambda})^2] \\ &\quad -\frac{1}{2}\ln\mathbb{E}_{\theta_0,\lambda_{i0}}[-\ell_{iT}^{\lambda\lambda}] + \ln\pi_i(\bar{\lambda}_i), \qquad (7) \\ \mathcal{B}_{iT}^{(2)}(\theta) &= T^{3/2}\frac{1}{2}\frac{\mathbb{E}_{\theta_0,\lambda_{i0}}\left[V_{iT}^{\lambda\lambda}(\ell_{iT}^{\lambda})^2\right]}{\{\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda}]\}^2} - T^{3/2}\frac{1}{6}\frac{\mathbb{E}_{\theta_0,\lambda_{i0}}[(\ell_{iT}^{\lambda})^3]\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda\lambda}]}{\{\mathbb{E}_{\theta_0,\lambda_{i0}}[\ell_{iT}^{\lambda\lambda}]\}^3}, \qquad (8)\end{aligned}$$

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

Remark 3.1 *Notice that, by standard arguments,* $\mathcal{B}_i^{(1)}(\theta)$ *and* $\mathcal{B}_i^{(2)}(\theta)$ *are both* $O(1)$. *An important difference between the result presented here and the corresponding result in Arellano and Bonhomme (2009) is the extra* $O(T^{-3/2})$ *term, given by* $\mathcal{B}_i^{(2)}(\theta)/T^{3/2}$. *This is due to the presence of serial dependence. When the error term is assumed to be serially independent, this term is actually* $O(T^{-2})$, *which leaves the* $O(T^{-1})$ *bias term only. A proof of this is given in Lemma A.3 in the Mathematical Appendix. However, in the case at hand, error terms are dependent, and without further assumptions, the extra bias term might remain.*

By taking the derivative of (6) with respect to θ, one can derive an expression for $\pi_i(\bar{\lambda}_i)$ that removes the first order bias of the score (see Arellano and Bonhomme (2009)). This leads to two specifications of the bias-reducing priors.

Proposition 3.2 *The robust prior that cancels the bias term of order* $O(T^{-1})$ *only is given by*

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

which is valid in a likelihood setting while

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

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

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

A further characterisation of bias-reducing priors is available.

Proposition 3.3 *Assuming* $T\widehat{Var}(\hat{\lambda}_i(\theta))$ *is a consistent estimator of asymptotic variance of* $\sqrt{T}\left[\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta)\right]$ *as* $T \to \infty$*, a further characterisation of (P1) is given by*

$$\pi_i^R\left(\hat{\lambda}_i(\theta)|\theta\right) \propto \left(\sqrt{\widehat{Var}(\hat{\lambda}_i(\theta))}\right)^{-1}\left(1 + O_p\left(\frac{1}{T^{1/2}}\right)\right). \tag{P3}$$

Moreover, any non-dogmatic prior that satisfies (P3) will also correct the $O(T^{-1})$ *bias.*

Remark 3.2 *Proposition 3.3 is useful in getting intuition about the mechanism underlying the bias-reducing priors. This reveals that the robust prior favours cases where* $\hat{\lambda}_i(\theta)$ *is precisely estimated for a given* θ*. The resulting integrated likelihood will concentrate around such cases. If, on the other hand,* $\hat{\lambda}_i(\theta)$ *is imprecise, a lower weight will be assigned (Arellano and Bonhomme (2011)).*

Priors (P1), (P2) and (P2*) follow directly from Proposition 3.1. See Arellano and Bonhomme for the details of the derivation of (P1) and (P2); these are not reproduced here. In particular, (P2) and (P2*) are analogous and follow by simple inspection. 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.

Next, implications of cross-section dependence are analysed. As mentioned previously, the case of cross-section dependence has so far not been analysed in the bias-reduction literature.[18] In this study, cross-section dependence is assumed to be such that the average score and, hence, the estimator are characterised by $\sqrt{T}$-convergence. In other words, cross-section dependence is so strong that cross-section size does not contribute to convergence (see Assumption 3.7). Intuitively, this means that there is no central limit theorem working across the cross-section. It is likely that this assumption is too strict, yet it also provides a "worst-case" benchmark to obtain some intuition about the effects of dependence. Define

$$\begin{aligned}
&\hat{\theta}_{IL} = \arg\max_{\theta} \sum_{i=1}^{N} \ell_{iT}^{I}(\theta)\,, \\
&S = \nabla_\theta \ell_{NT}(\theta_0) = \left\{\ell_{NT}^{\theta}(\theta_0) - \ell_{NT}^{\lambda}(\theta_0)\frac{\mathbb{E}[\ell_{NT}^{\lambda\theta}(\theta_0)]}{\mathbb{E}[\ell_{NT}^{\lambda\lambda}(\theta_0)]}\right\}, \\
&H = \nabla_{\theta\theta}\ell_{NT}(\theta_0)\,, \qquad \nu = \mathbb{E}[\nabla_{\theta\theta}\ell_{NT}(\theta_0)], \\
&Z_i = \mathbb{E}\left[\nabla_{\theta\theta}\frac{d\ell_{NT}(\theta)}{d\theta_i}\right]\Big|_{\theta=\theta_0}, \qquad \text{where } i = 1, 2,\ \theta_1 = \alpha \text{ and } \theta_2 = \beta,
\end{aligned}$$

[18] 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.

and $\qquad M = \begin{bmatrix} S'\nu^{-1}Z_1\nu^{-1}S \\ S'\nu^{-1}Z_2\nu^{-1}S \end{bmatrix}.$

The following Proposition gives the double-asymptotic expansion for $(\hat{\theta}_{IL} - \theta)$ as $T, N \to \infty$.

Proposition 3.4

$$\begin{aligned}(\hat{\theta}_{IL} - \theta_0) &= -\nu^{-1}S \\ &\quad -\nu^{-1}\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta\left\{\frac{1}{T}\ln\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta,\bar{\lambda}_i(\theta))] + \frac{1}{T}\ln\pi_i(\bar{\lambda}_i(\theta)|\theta) - \frac{\left[\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta))\right]^2}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta,\bar{\lambda}_i(\theta))]}\right\}\Bigg|_{\theta=\theta_0} \\ &\quad +\nu^{-1}\left(H-\nu\right)S\nu^{-1} - \frac{1}{2}\nu^{-1}M + O_p\left(\frac{1}{T^{3/2}}\right). \qquad (9)\end{aligned}$$

The first term on the right-hand side of (9) is the average efficient score with respect to θ and drives the convergence in distribution. The second term contains the now familiar term of the first-order incidental parameter bias of the score. Remember that if the robust prior is used to construct $T^{-1}\ln\pi_i(\bar{\lambda}_{iT}(\theta)|\theta)$, then, by the definition of robust priors, the expectation of this term vanishes in the $O(T^{-3/2})$ remainder term. The orders of magnitude of the third and fourth terms are determined by Assumption 3.7 and these are both $O_p(T^{-1})$. The below Corollary follows immediately.

Corollary 3.1 *When the robust prior, $\pi_i^R\left(\bar{\lambda}_{iT}(\theta_0)|\theta_0\right)$, is used,*

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

Remark 3.3 *Corollary 3.1 reveals that under cross-section dependence, there are two types of bias. The first is the incidental parameter bias, which is successfully removed by the robust prior. The second bias is due to cross-section dependence: in the case where cross-section size does not contribute to convergence, extra terms that do not vanish as fast as a $O\left(T^{-3/2}\right)$ term arise. However, this setting might be more strict than necessary; as will be discussed below, simulation results reveal that the average bias of the integrated likelihood estimator does change only marginally between the cross-section dependence and independence settings.*

Remark 3.4 *That N does not contribute to convergence speed leads to another difference with the literature. In the usual microeconometric setting (assuming cross-section independence), when $N, T \to \infty$ jointly and $N/T \to c$, the classical result for the concentrated likelihood estimator, $\hat{\theta}$, is*

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

where $\mathcal{B}$ is some first-order bias and Ω is the asymptotic variance matrix. Then, clearly it is not enough that T is large; one must either ensure that $N/T \to 0$ or remove the first-order bias. In this paper, on the other hand, $p\lim_{T\to\infty} \mathbb{E}\left[\hat{\theta}_{IL} - \theta_0\right] = 0$ independent of at what rate N and T go to infinity. Hence, here the bias problem is purely a small-T issue. The situation will be different if N has some contribution to convergence. This is explained next.

Corollary 3.2 *Consider the following modifications to Assumption 3.7:*

$$Var[\nabla_\theta \ell_{NT}(\theta, \bar{\lambda}_T(\theta))] = O\left(\frac{1}{N^{\rho_1}T}\right)$$

and $$\sup_{\theta\in\Theta} Var\left(\nabla_{\theta\theta}\ell_{NT}(\theta, \bar{\lambda}_i(\theta)) - \mathbb{E}[\nabla_{\theta\theta}\ell_{NT}(\theta, \bar{\lambda}_i(\theta))]\right) = O\left(\frac{1}{N^{\rho_2}T}\right),$$

where $1/2 \leq \rho_1 \leq 1$, $0 \leq \rho_2 \leq 1$ and $1 \leq \rho_1 + \rho_2 \leq 2$. Then, using the robust prior,

$$\sqrt{N^{\rho_1}T}(\hat{\theta}_{IL} - \theta_0) \xrightarrow{d} \mathcal{N}\left(\sqrt{\frac{N^{\rho_1}}{T^2}}\mathcal{B}, \Omega\right),$$

where $T^{-3/2}\mathcal{B}$ is a $O\left(T^{-3/2}\right)$ remainder, after the removal of the first-order bias.

Remark 3.5 *Corollary 3.2 follows from Corollary 3.1 by observing the updated convergence rates. This result indicates that the remainder term will be of a negligible magnitude if $N^{\rho_1}/T^2 \to 0$ as $N, T \to \infty$. For $\rho_1 = 1/2$, this implies that N can grow at the same rate as T^4, a realistic case for financial panels.*

Remark 3.6 *A final remark on the extra bias terms derived in this paper is in order. Although there would be gains in estimating the extra $O(T^{-3/2})$ incidental parameter bias in (6) and the second type of bias due to cross-section dependence appearing in (9), this is not pursued here. There are several reasons for that. First, these expressions include non-standard terms such as third order moments of centred likelihood expressions and third order likelihood derivatives. As closed-form expressions for GARCH likelihood derivatives do not exist, the third-order derivative has to be calculated numerically. Due to associated numerical accuracy issues, it is not clear whether estimating this terms will be useful. Second, to estimate the cross-section dependence bias, one has to know the correct order of magnitude of this term. If it is $O(T^{-3/2})$, then assuming wrongly that it is $O(T^{-1})$ and removing a wrongly normalised term might introduce further bias into the system. Moreover, simulation results reported in the next section suggest that the effect of neglected cross-section dependence on average bias is not significant. For these reasons, the extra bias terms will not further be analysed in this study.*

3.2 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.

4 Simulation Analysis

4.1 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.

Data are generated by using,

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

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

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

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%. [19]

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

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

as in Shephard and Sheppard (2010), where $\lceil T^{1/2} \rceil$ is obtained by rounding $T^{1/2}$ up to the nearest integer.[20]

[19] 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.

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

4.2 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 $\alpha+\beta$ 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.[21] 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 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).[22]

To summarise, simulation results show that, in line with the theoretical results, bias

[21] 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.

[22] 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.

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.

4.3 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.

5 Empirical Analysis

This section presents two empirical studies of the bias-reduced GARCH panel estimator. The first is a comparison of predictive ability, based on stock return volatility forecasts by different methods. The second is an analysis of hedge fund volatility using a consolidated database of hedge fund returns. Hedge fund returns are rarely available at higher than monthly frequency and the maximum number of observations for any fund is around 200. This makes it virtually impossible to analyse hedge fund volatility using standard GARCH estimation techniques. Hence, this empirical analysis is a novel contribution to the literature. In both applications, naturally, a pseudo-likelihood setting is assumed and the integrated likelihood functions are constructed using the pseudo-likelihood Prior given

in (P2).

5.1 Analysis of Predictive Ability

5.1.1 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.[23] 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.[24]

For a more detailed explanation on the features of the dataset and estimation of the realised variances, see Noureldin, Shephard and Sheppard (2011). In particular, they report that both the returns and the realised variances are open-to-close due to market microstructure noise. In addition, the first and last 15 minutes of trading are dropped from the sample in order to deal with overnight effects. Lastly, realised variances are based on 5-minute returns with subsampling.

5.1.2 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

$$\begin{aligned} \mathbb{E}[\varepsilon_{it}^2|\mathcal{F}_{i,t-1}] = \sigma_{it}^2 &= \lambda_i(1-\alpha-\beta) + \alpha\varepsilon_{i,t-1}^2 + \beta\sigma_{i,t-1}^2, \\ \hat{\sigma}_{it}^2 &= \hat{\lambda}_i(1-\hat{\alpha}-\hat{\beta}) + \hat{\alpha}\varepsilon_{i,t-1}^2 + \hat{\beta}\sigma_{i,t-1}^2. \end{aligned}$$

[23] 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).

[24] 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.

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 (3). In the first step, λ_i are estimated by method of moments using (5). $\tilde{\lambda}_1, ..., \tilde{\lambda}_N$ 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.[25]

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 (5) 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

$$\hat{\lambda}_i^c(\hat{\theta}_{IL}) = \arg\max_{\lambda_i \in \Lambda_i} \frac{1}{T} \sum_{t=1}^{T} \ell_{it}(\hat{\theta}_{IL}, \lambda_i). \tag{10}$$

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.[26] 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.[27]

5.1.3 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

[25] 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 $.8 \times (\min_{i,t} r_{it}^2)$.

[26] 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.

[27] Remember that CL uses $\tilde{\lambda}_i = T^{-1} \sum_{t=1}^{T} y_{it}^2$ to construct $(NT)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N} \ell_{it}(\theta, \tilde{\lambda}_i)$ which is not necessarily the same as $(NT)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N} \ell_{it}(\theta, \hat{\lambda}_i(\theta))$ where $\hat{\lambda}_i(\theta) \equiv \arg\max_{\lambda_i} T^{-1} \sum_{t=1}^{T} \ell_{it}(\theta, \lambda_i)$.

(GARCH, Exponential GARCH etc.) this test, rather than the Diebold-Mariano-West[28] 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 better predictive ability (and similarly for negative $\Delta L_{i,t+1}$).

5.1.4 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

[28] 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.

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 (10). 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 four of those being statistically significant. The difference is, without much surprise, more striking between QML and IPCL where the GW-test favours IPCL six 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.[29]

5.2 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. The global assets under management are now expected to reach $2.25 billion by the end of 2011, despite capital outflows following the credit crunch episode.[30] 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 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

[29] 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.

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

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))[31]. 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.

5.2.1 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.[32] Importantly, funds are classified into ten vendor-reported invest-

[31] 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.

[32] 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.

ment 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.

5.2.2 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.

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.[33] 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

[33] 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.

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 $T = 150$.[34] 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 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

[34] Intercept parameters have been estimated using the concentrated likelihood method as in (10).

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.

6 Conclusion

This paper has analysed the properties of first-order bias correction by the integrated likelihood in a nonlinear dynamic panel estimation setting under both serial and cross-section dependence. Analysis of bias-reduction of the integrated likelihood method has been extended to time-series and cross-section dependence and analytical expressions for the resulting extra terms have been provided. The latter contribution is a general one, as such dependence has not yet been considered in the bias-reduction literature. The provided double-asymptotic expansions and bias characterisations can be used in a general setting, not necessarily confined to the particular application of interest in this paper. These results have been used to model GARCH effects using panels with a limited number of time-series observations. 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. This empirical analysis is another novel contribution, as such analysis has hitherto been impossible due to hedge fund data being available at monthly frequency. Moreover, in a test of predictive ability using stock volatility forecasts, the proposed estimation method achieved superior forecasting performance compared to its alternatives.

Several extensions are in order. The results derived in this paper should be extended to other panel data models, to attain a better understanding of underlying dependence mechanisms. Also, further research has to be undertaken in order to develop a criterion to find those series that indeed satisfy the assumption of a common set of parameters. This is a challenging issue, as a large N would imply large-dimensional hypothesis testing, where controlling the rejection rate is problematic. Insights from the empirical Bayes literature and recent research on microarrays can be used to make advances in this direction.

Understanding the effects of time-series and cross-section dependence is crucial when the interest is in modelling macro or financial panels. Especially financial series are generally observed to exhibit co-movements of some degree. For example, correlation between stock returns increases during financial downturns and high volatility episodes. It is reasonable to assume some underlying factor structure where units are hit by the same shock, but react in different ways. Bai (2009) lists several examples where the error term in a linear regression may exhibit a factor structure: common technological or financial shocks that impact countries differently (macroeconometrics); individual-specific time-invariant unobservable characteristics and skills that have a time-varying common price (microeconometrics); factor models with unobservable factor returns and individual-specific factor loadings (finance). Although a significant amount of effort has gone into analysing cross-section dependence in the panel data literature, it has not yet been considered in the growing bias reduction literature. Extensions in that direction would be fruitful. This might be somewhat challenging for GARCH panels as the bias expressions are based on likelihood derivatives, and these do not exist in closed form for GARCH. A factor structure will also be useful in reducing the parameter dimension when cross-section dependence is not neglected but modelled explicitly.

A Mathematical Appendix

A.1 A Preliminary Result

The following Lemmas will be useful in proving some of the results mentioned in this study. These are not novel and relevant general results have already been analysed in detail (see, among others, McCullagh (1987) and Pace and Salvan (1997)).

Lemma A.1 *Under Assumption 3.4,*

$$\begin{aligned}
\delta_i &= \frac{1}{E_{iT}}\Bigg\{-\ell^{\lambda}_{iT}+\frac{V^{\lambda\lambda}_{iT}\ell^{\lambda}_{iT}}{E_{iT}}-\frac{1}{2}\frac{\ell^{\lambda\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^2}{E^2_{iT}}+\frac{-V^{\lambda\lambda}_{iT}}{E_{iT}}\left[\frac{V^{\lambda\lambda}_{iT}\ell^{\lambda}_{iT}}{E_{iT}}-\frac{1}{2}\frac{\ell^{\lambda\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^2}{E^2_{iT}}\right] \\
&\quad -\frac{1}{2}\frac{\ell^{\lambda\lambda\lambda}_{iT}}{E^2_{iT}}\left[-2\frac{V^{\lambda\lambda}_{i}\left(\ell^{\lambda}_{iT}\right)^2}{E_{iT}}+\frac{\ell^{\lambda\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^3}{E^2_{iT}}\right]-\frac{1}{6}\frac{\ell^{\lambda\lambda\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^3}{E^3_{iT}}\Bigg\}+O_p(T^{-2}), \qquad (11)\\
\delta^2_i &= \frac{1}{E^2_{iT}}\left[\left(\ell^{\lambda}_{iT}\right)^2-2\frac{V^{\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^2}{E_{iT}}+\frac{\ell^{\lambda\lambda\lambda}_{iT}\left(\ell^{\lambda}_{iT}\right)^3}{E^2_{iT}}\right]+O_p(T^{-2}), \qquad (12)\\
\delta^3_i &= -\frac{1}{E^3_{iT}}\left(\ell^{\lambda}_{iT}\right)^3+O_p(T^{-2}). \qquad (13)
\end{aligned}$$

Proof of Lemma A.1. 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

$$\begin{aligned}
\delta_i &= \frac{1}{E_{iT}}\left[-\ell^{\lambda}_{iT}-\frac{V^{\lambda\lambda}_{iT}}{E_{iT}}\left(-\ell^{\lambda}_{iT}-V^{\lambda\lambda}_{iT}\delta_i-\frac{1}{2}\ell^{\lambda\lambda\lambda}_{iT}\delta^2_{iT}\right)-\frac{1}{2}\ell^{\lambda\lambda\lambda}_{iT}\delta^2_i-\frac{1}{6}\ell^{\lambda\lambda\lambda\lambda}_{iT}\delta^3_i\right]+O_p(T^{-2}) \\
&= \frac{1}{E_{iT}}\Bigg\{-\ell^{\lambda}_{iT}-\frac{V^{\lambda\lambda}_{iT}}{E_{iT}}\left[-\ell^{\lambda}_{iT}-\frac{V^{\lambda\lambda}_{iT}}{E_{iT}}\left(-\ell^{\lambda}_{iT}\right)-\frac{1}{2}\ell^{\lambda\lambda\lambda}_{iT}\delta^2_i\right] \\
&\quad -\frac{1}{2}\ell^{\lambda\lambda\lambda}_{iT}\delta^2_i-\frac{1}{6}\ell^{\lambda\lambda\lambda\lambda}_{iT}\delta^3_i\Bigg\}+O_p(T^{-2}) \qquad (14)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\delta^2_i &= \frac{1}{E^2_{iT}}\left[\left(\ell^{\lambda}_{iT}\right)^2+2\ell^{\lambda}_{iT}V^{\lambda\lambda}_{iT}\delta_i+\ell^{\lambda}_{iT}\ell^{\lambda\lambda\lambda}_{iT}\delta^2_i\right]+O_p(T^{-2}) \\
&= \frac{1}{E^2_{iT}}\left[\left(\ell^{\lambda}_{iT}\right)^2-2\frac{\left(\ell^{\lambda}_{iT}\right)^2V^{\lambda\lambda}_{iT}}{E_i}+\ell^{\lambda}_{iT}\ell^{\lambda\lambda\lambda}_{iT}\delta^2_i\right]+O_p(T^{-2}), \qquad (15)
\end{aligned}$$

where (14) is used to obtain (15). Substituting δ^2_i back into (15) yields (12), while observing $\delta^3_i=\delta_i\delta^2_i$ gives (13). Finally, using (12) and (13) in (14), (11) follows. ■

A.2 Proof of Proposition 3.1

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

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

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

Lemma A.2

$$\ell^{I}_{iT}(\theta)-\ell^{c}_{iT}(\theta)=\frac{1}{2T}\ln\left(\frac{2\pi}{T}\right)-\frac{1}{2T}\ln[-\ell^{\lambda\lambda}_{iT}(\theta,\hat{\lambda}_i(\theta))]+\frac{1}{T}\ln\pi_i(\hat{\lambda}_i(\theta))+O\left(\frac{1}{T^2}\right). \qquad (16)$$

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

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

and likewise for higher order derivatives. Then, expanding $\ell_i(\theta,\lambda_i)$ and $\pi_i(\lambda_i|\theta)$ around $\hat{\lambda}_i(\theta)$ and using Assumption 3.5, 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[35]

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

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 $\exp x = 1 + x + x^2/2 + x^3/6 + ...$,

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

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

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

powers of z_i implying that their expectations will all be $O(T^{-2})$. Hence,

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

■

Next, a series of Taylor approximations will be used to derive an expression for $E[\ell_{iT}^{I}(\theta) - \ell_{iT}^{c}(\theta)]$ using (16). As mentioned at the beginning of the Appendix, it is assumed that all relevant moments exist and are finite, and that appropriate laws of large numbers (LLN) and central limit theorems (CLT) for second through fourth order derivatives of the composite log-likelihood with respect to λ_i exist as $T \to \infty$.

Lemma A.3

$$\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta) = \frac{A_i}{\sqrt{T}} + O_p\left(\frac{1}{T}\right), \tag{17}$$

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

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

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

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

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

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

Lemma A.4

$$\ell_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta)) = \ell_{iT}^{\lambda\lambda}(\theta,\bar{\lambda}_i(\theta)) + \frac{B_i}{\sqrt{T}} + O_p\left(\frac{1}{T}\right), \tag{18}$$

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

Proof. Since $\ell_{iT}^{\lambda\lambda\lambda}$ and $\ell_{iT}^{\lambda\lambda\lambda\lambda}$ are both $O_p(1)$, expanding $\ell_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta))$ around $\hat{\lambda}_i(\theta) = \bar{\lambda}_i(\theta)$, and using (17) yields

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

Using an appropriate LLN and CLT for β-mixing processes, as $T \to \infty$, $T^{-1}\sum_{t=1}^{T}\ell_{it}^{\lambda\lambda\lambda} - T^{-1}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda\lambda}] = O_p(T^{-1/2})$. 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.5

$$\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta)) = \frac{C_i}{\sqrt{T}} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta))] + O_p\left(\frac{1}{T}\right), \tag{19}$$

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

Proof. Using appropriate LLN and CLT for β-mixing processes, as $T \to \infty$ one can obtain $V_i^{\lambda\lambda} = T^{-1}\sum_{t=1}^{T}\ell_{it}^{\lambda\lambda} - T^{-1}\sum_{t=1}^{T}\mathbb{E}[l_{it}^{\lambda\lambda}] = O_p(T^{-1/2})$. Then, using (18),

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

Lastly,

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

■

Lemma A.6

$$\mathbb{E}\ln[-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))] = \ln\{-\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta, \bar{\lambda}_i(\theta))]\} + O\left(\frac{1}{T}\right). \tag{20}$$

Proof. Using (19),

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

Then

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

and

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

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

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

Hence,

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

and

$$\ln\{-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))\} = \ln\{-\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\} + \frac{C_i}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\sqrt{T}} + O_p(T^{-1}), \tag{21}$$

implying

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

■

Lemma A.7

$$\mathbb{E}_{\theta_0,\lambda_{i0}} \ln \pi_i(\hat{\lambda}_i(\theta)) = \ln \pi_i(\bar{\lambda}_i(\theta)) + O\left(\frac{1}{T}\right). \tag{22}$$

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

$$\ln \pi_i(\hat{\lambda}_i(\theta)) = \ln \pi_i(\bar{\lambda}_i(\theta)) + \frac{\partial \ln \pi_i(\bar{\lambda}_i(\theta))}{\partial \lambda_i}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta)) + O_p(T^{-1}), \tag{23}$$

which implies that,

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

■

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

$$\mathbb{E}_{\theta_0,\lambda_{i0}}\left[\ell_i^I(\theta) - \ell_i^c(\theta)\right] = C - \frac{1}{2T}\ln\left\{-T^{-1}\sum\nolimits_{t=1}^{T}\mathbb{E}[\ell_{it}^{\lambda\lambda}]\right\} + \frac{1}{T}\ln \pi_i\left(\bar{\lambda}_i(\theta)\right) + O\left(T^{-2}\right). \tag{24}$$

Proof. Taking the expectation of (16) gives

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

Using $C = (2T)^{-1}\ln\left(2\pi T^{-1}\right)$ and substituting (20) and (22), (24) follows. ■

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

$$\mathbb{E}\left[\ell_{iT}^c(\theta, \hat{\lambda}_i) - \ell_{iT}\left(\theta, \bar{\lambda}_i\right)\right] = -\frac{1}{2}\frac{\mathbb{E}[(\ell_{iT}^{\lambda})^2]}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]} + \frac{1}{2}\frac{\mathbb{E}[V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^2} - \frac{1}{6}\frac{\mathbb{E}[(\ell_{iT}^{\lambda})^3]\mathbb{E}[\ell_{iT}^{\lambda\lambda\lambda}]}{\left\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}]\right\}^3} + O_p(T^{-2}). \tag{25}$$

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

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

Using Lemma A.1,

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

which implies that

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

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

■

Finally, the proof of Proposition 3.1 follows.

Proof. (Proposition 3.1) Using (24) and (25) gives (6), (7) and (8). ■

A.3 Characterisation of the Bias under the iid Assumption

It was mentioned previously in Remark 3.1 that when the error terms are iid, the bias term given by $\mathcal{B}_i^{(2)}(\theta)/T^{3/2}$ can be shown to be $O(T^{-2})$ rather than $O(T^{-3/2})$. This is proved below. See also Severini (2005, Theorem 4.18) and Pace and Salvan (1997).

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

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

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

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

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

For $j_1 + j_2 = j$, let $cum(X_1^{j_1}, X_2^{j_2})$ be a generic j^{th} order cumulant and assume that for $j \leq 3$ all cumulants exist. Theorem 2.1 in Hall (1992, pp. 53-54) establishes that

$$
cum(X_1^{j_1} X_2^{j_2}) = T^{-(j-2)/2}(k_{j,1} + T^{-1}k_{j,2} + T^{-2}k_{j,3} + ...), \tag{26}
$$

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

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

where the third line follows from $cum(\sqrt{T}\tilde{\ell}_{iT}^{\lambda}) = \mathbb{E}[\sqrt{T}\tilde{\ell}_{iT}^{\lambda}] = 0$ and the final result is due to (26). Then, $\mathbb{E}\left[(\tilde{\ell}_{iT}^{\lambda})^3\right] = O(T^{-2})$ and $\mathbb{E}\left[\left(\tilde{\ell}_{iT}^{\lambda}\right)^3\right] = O(T^{-2})$. Similarly,

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

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

A.4 Proof of Proposition 3.3

The following follow closely the proof of *Proposition 3* in Arellano and Bonhomme (2009). The only difference is due to the extra terms of order $O(T^{-3/2})$.

Proof of Proposition 3.3. By Lemma A.1

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

and

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

Then,

$$\begin{aligned}\widehat{Var}(\hat{\lambda}_i(\theta)) &= \mathbb{E}\left[\left(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta)\right)^2\right] - \left\{\mathbb{E}[\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta)]\right\}^2 \\ &= \frac{\mathbb{E}[\left(\ell_{iT}^{\lambda}\right)^2]}{\left\{\mathbb{E}\left[\ell_{iT}^{\lambda\lambda}\right]\right\}^2} + O_p\left(\frac{1}{T^{3/2}}\right) = \frac{1}{\left[\pi_i^R(\bar{\lambda}_i(\theta))\right]^2} + O_p\left(\frac{1}{T^{3/2}}\right),\end{aligned}$$

by the definition of $\pi_i^R(\bar{\lambda}_i(\theta))$. Note that $\mathbb{E}[\left(\ell_{iT}^{\lambda}\right)^2]\left\{\mathbb{E}\left[\ell_{iT}^{\lambda\lambda}\right]\right\}^{-2}$ is $O(1)$ which is consistent with the fact that $T\widehat{Var}(\hat{\lambda}_i(\theta))$ is a consistent estimator of the asymptotic variance of $\sqrt{T}(\hat{\lambda}_i(\theta) - \bar{\lambda}_i(\theta))$. These imply that

$$\left[\pi_i^R(\bar{\lambda}_i(\theta))\right]^2 = \frac{1}{\widehat{Var}(\hat{\lambda}_i(\theta))} + O_p\left(T^{1/2}\right) = \frac{1}{\widehat{Var}(\hat{\lambda}_i(\theta))}\left(1 + O_p\left(\frac{1}{T^{1/2}}\right)\right)$$

since $T\widehat{Var}(\hat{\lambda}_i(\theta)) = O_p(1)$. Observe that

$$\pi_i^R(\bar{\lambda}_i(\theta)) \propto \frac{1}{\sqrt{\widehat{Var}(\hat{\lambda}_i(\theta))}}\left(1 + O_p\left(\frac{1}{T^{1/2}}\right)\right).$$

Using the argument in Arellano and Bonhomme (2009),

$$\pi_i^R(\hat{\lambda}_i(\theta)) = \pi_i^R(\bar{\lambda}_i(\theta))\left(1 + O_p\left(\frac{1}{T}\right)\right),$$

which implies that

$$\pi_i^R(\hat{\lambda}_i(\theta)) \propto \frac{1}{\sqrt{\widehat{Var}(\hat{\lambda}_i(\theta))}}\left(1 + O_p\left(\frac{1}{T^{1/2}}\right)\right).$$

To complete the proof, it only remains to show that any non-dogmatic prior that is proportional to

$$\left(\sqrt{\widehat{Var}(\hat{\lambda}_i(\theta))}\right)^{-1}\left(1 + O_p\left(T^{-1/2}\right)\right)$$

is robust. Take such a prior, $\pi_i(\hat{\lambda}_i(\theta)) \propto \left(\sqrt{\widehat{Var}(\hat{\lambda}_i(\theta))}\right)^{-1}\left(1 + O_p\left(T^{-1/2}\right)\right)$. Looking at the difference between the integrated and concentrated likelihood,

$$\ell_{iT}^I(\theta) - \ell_{iT}^c = \frac{1}{2T}\ln\left(\frac{2\pi}{T}\right) - \frac{1}{2T}\ln\left[-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))\right] + \frac{1}{T}\ln\pi_i(\hat{\lambda}_i(\theta)) + O_p\left(\frac{1}{T^{3/2}}\right), \tag{27}$$

the only quantity involving the prior is $T^{-1}\ln\pi_i(\hat{\lambda}_i(\theta))$. Moreover, both $\pi_i^R(\hat{\lambda}_i(\theta))$ and $\pi_i(\hat{\lambda}_i(\theta))$ are proportional to

$\left(\sqrt{\widehat{Var}(\hat{\lambda}_i(\theta))}\right)^{-1}$ up to a remainder term of the same order. Therefore,

$$\ln \pi_i(\hat{\lambda}_i(\theta)) = \ln \pi_i^R(\hat{\lambda}_i(\theta)) + O_p\left(\frac{1}{T^{1/2}}\right),$$

and, finally, (27) is equal to

$$\frac{1}{2T}\ln\left(\frac{2\pi}{T}\right) - \frac{1}{2T}\ln\left[-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))\right] + \frac{1}{T}\ln \pi_i^R(\hat{\lambda}_i(\theta)) + O_p\left(\frac{1}{T^{3/2}}\right),$$

which shows that $\pi_i(\hat{\lambda}_i(\theta))$ is a robust prior. ■

A.5 Proof of Proposition 3.4

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 2×1 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.5.1 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 $r = 1, 2$; $\theta_1 = \alpha$ and $\theta_2 = \beta$. In the following, to make the notation less cumbersome, indices and subscripts are dropped whenever variables can be distinguished by context. For example, instead of $V_{iT}^{\lambda\lambda}$, simply V is used. Also, for a given function $f(\phi)$, and a P dimensional parameter vector $\phi = [\phi_p]$, $p = 1, ..., P$, define the generic m^{th} order derivative as

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

Then, for example,

$$\frac{d^m V_{iT}^{\lambda\lambda}}{d\theta_{r_1} ... d\theta_{r_m}} = \frac{d^m V}{d\theta_{r_1} ... d\theta_{r_m}} = V_{r_1,...,r_m}, \quad \text{where } r_1, ..., r_m \in \{1, 2\},$$

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.

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

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

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

where $r, s \in \{1,2\}$, $\theta^1 = \alpha$ and $\theta^2 = \beta$. Lastly, define

$$\delta_I^r = \left(\hat{\theta}_{IL} - \theta_0\right)^r \quad \text{where } r = 1, 2,$$

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.5.2 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 3.4.

Lemma A.8

$$\begin{aligned}
\nabla_\theta \ln(-E_{iT}) &= -\frac{E_{r_1}}{E} = O(1), \\
\nabla_{\theta\theta} \ln(-E_{iT}) &= -\frac{E_{r_1,r_2}}{E} + \frac{E_{r_1}E_{r_2}}{E^2} = O(1), \\
\nabla_\theta \left(\frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\right) &= \frac{V_{r_1}}{E} - \frac{VE_{r_1}}{E^2} = O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_{\theta\theta} \left(\frac{V_{iT}^{\lambda\lambda}}{E_{iT}}\right) &= \frac{V_{r_1,r_2}}{E} - \frac{V_{r_1}E_{r_2}[2] + VE_{r_1,r_2}}{E^2} + 2\frac{VE_{r_1}E_{r_2}}{E^3} = O_p\left(\frac{1}{\sqrt{T}}\right), \\
\nabla_\theta \left\{\frac{\ell_{iT}^{\lambda} F_{iT}}{E_{iT}^2}\right\} &= \frac{\ell_{r_1}F + \ell F_{r_1}}{E^2} - 2\frac{\ell F E_{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{\ell_{r_1,r_2}F + \ell_{r_1}F_{r_2}[2] + \ell F_{r_1,r_2}}{E^2} - 2\frac{\ell_{r_1}E_{r_2}F[2] + \ell E_{r_2}F_{r_1}[2] + \ell F E_{r_1,r_2}}{E^3} + 6\frac{\ell F E_{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}^{\lambda}}{E_{iT}}\right) &= \frac{\ell_{r_1}\bar{\pi} + \ell\bar{\pi}_{r_1}}{E} - \frac{\ell\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}^{\lambda}}{E_{iT}}\right) &= \frac{\ell_{r_1,r_2}\bar{\pi} + \ell_{r_1}\bar{\pi}_{r_2}[2] + \ell\bar{\pi}_{r_1,r_2}}{E} - \frac{\ell_{r_1}\bar{\pi}E_{r_2}[2] + \ell\bar{\pi}_{r_1}E_{r_2}[2] + \ell\bar{\pi}E_{r_1,r_2}}{E^2} + 2\frac{\ell\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{\ell\ell_{r_1}}{E} - \frac{\ell^2 E_{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{\ell_{r_2}\ell_{r_1} + \ell\ell_{r_1,r_2}}{E} - \frac{2\ell\ell_{r_2}E_{r_1}[2] + \ell^2 E_{r_1,r_2}}{E^2} + 2\frac{\ell^2 E_{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}\ell^2 + 2V\ell\ell_{r_1}}{E^2} - 2\frac{V\ell^2 E_{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}\ell^2 + 2V_{r_1}\ell\ell_{r_2}[2] + 2V\ell_{r_2}\ell_{r_1} + 2V\ell\ell_{r_1,r_2}}{E^2} \\
&\quad -2\frac{V_{r_1}\ell^2 E_{r_2}[2] + 2V\ell\ell_{r_1}E_{r_2}[2] + V\ell^2 E_{r_1,r_2}}{E^3} + 6\frac{V\ell^2 E_{r_1}E_{r_2}}{E^4} \\
&= O_p\left(\frac{1}{T^{3/2}}\right), \\
\nabla_\theta \left\{\frac{(\ell_{iT}^{\lambda})^3 F_{iT}}{E_{iT}^3}\right\} &= \frac{3\ell^2\ell_{r_1}F + \ell^3 F_{r_1}}{E^3} - 3\frac{\ell^3 F E_{r_1}}{E^4} = O_p\left(\frac{1}{T^{3/2}}\right),
\end{aligned}$$

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

Moreover, the third and fourth derivatives satisfy,

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

A.5.3 The Proof

Proof (Proposition 3.4). The starting point is

$$\begin{aligned}\ell_{iT}^{I}(\theta) &= \ell_{iT}(\theta,\bar{\lambda}_i(\theta))+C-\frac{1}{2T}\left[\ln(-E_{iT})+\frac{V_{iT}^{\lambda\lambda}}{E_{iT}}-\frac{\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta))F_{iT}}{E_{iT}^2}\right]+\frac{1}{T}\left[\ln\pi_{iT}(\bar{\lambda}_{iT}(\theta))-\frac{\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta))\bar{\pi}_{iT}^{\lambda}}{E_{iT}}\right]\\ &\quad -\frac{1}{2}\frac{\left(\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta))\right)^2}{E_{iT}}+\frac{1}{2}\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta))\right)^2}{(E_{iT})^2}-\frac{1}{6}\frac{(\ell_{iT}^{\lambda}(\theta,\bar{\lambda}_i(\theta)))^3F_{iT}}{E_{iT}^3}+O_p(T^{-2}),\end{aligned}\tag{28}$$

where $C=(2T)^{-1}\ln(2\pi/T)$. The first line follows from using (21) and (23) to substitute for $\ln[-\ell_{iT}^{\lambda\lambda}(\theta,\hat{\lambda}_i(\theta))]$ and $\ln\pi_i(\hat{\lambda}_i(\theta))$, respectively, in (16). Notice that the expression given by (23) is a function of $[\hat{\lambda}_i(\theta)-\bar{\lambda}_i(\theta)]$, which has to be substituted by (11), up to a $O_p(T^{-2})$ term. The second line is obtained by adding $\ell_{iT}(\theta,\hat{\lambda}_i(\theta))-\ell_{iT}$, which is calculated by using the arguments in the Proof of Proposition A.2. ■

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

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\ell_{r_1}^{I}(\hat{\theta}_{IL}) &= \frac{1}{N}\sum_{i=1}^{N}\ell_{r_1}^{I}(\theta_0)+\left[\frac{1}{N}\sum_{i=1}^{N}\ell_{r_1,r_2}^{I}(\theta_0)\right]\delta_I^{r_2}+\frac{1}{2}\left[\frac{1}{N}\sum_{i=1}^{N}\ell_{r_1,r_2,r_3}^{I}(\theta_0)\right]\delta_I^{r_2}\delta_I^{r_3}\\ &\quad +\frac{1}{6}\left[\frac{1}{N}\sum_{i=1}^{N}\ell_{r_1,r_2,r_3,r_4}^{I}(\theta_0)\right]\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}+\frac{1}{24}\left[\frac{1}{N}\sum_{i=1}^{N}\ell_{r_1,r_2,r_3,r_4,r_5}^{I}(\bar{\theta})\right]\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\delta_I^{r_5},\end{aligned}\tag{29}$$

where $r_1,...,r_5=1,2$ and $\bar{\theta}\in[\min(\hat{\theta}_{IL},\theta_0),\max(\hat{\theta}_{IL},\theta_0)]$. In the worst possible case of $\sqrt{T}$-convergence (rather than the$\sqrt{NT}$-convergence observed under cross-section dependence)

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

Notice that the expansion gives a vector.

The integrated likelihood is not a familiar concept. Instead, the concentrated likelihood would be much more convenient

intuitive to work with. This is achieved by using (28) to obtain target-likelihood based approximations for integrated-likelihood derivatives appearing on the right-hand side of (29). These approximations are then substituted for relevant integrated likelihood derivatives in (29). This leads to the next Lemma.

Lemma A.9

$$\begin{aligned}
-\delta_I^{r_2}\nu_{r_1,r_2} &= \tilde{\ell}_{r_1} + \mathcal{D}_{1;r_1} + \delta_I^{r_2}\mathcal{H}_{r_1,r_2} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\nu_{r_1,r_2,r_3} + \mathcal{D}_{3;r_1} \\
&\quad + \delta_I^{r_2}\mathcal{D}_{2;r_1,r_2} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\mathcal{H}_{r_1,r_2,r_3} + \frac{1}{6}\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\upsilon_{r_1,r_2,r_3,r_4} + O_p\left(\frac{1}{T^2}\right). \qquad (30)
\end{aligned}$$

where

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

Proof. First, derivatives of (28) with respect to θ have to obtained. This is achieved by simply substituting the results given in Lemma A.8 as necessary. Then,

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

Substituting these expansions for the integrated likelihood derivatives into (29) gives

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

Noting that $\tilde{\ell}^I_{r_1}(\hat{\theta}_{IL}) = 0$ for $r_1 \in \{1, 2\}$ and rearranging terms according to their stochastic orders of magnitude yields

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

Then, using Assumption 3.7,

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

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

or, more concisely,

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

which is the desired result. ■

Notice that, by definition, $\left(\hat{\theta}_{IL}-\theta_0\right)=[\delta_I^{r_2}]$, where $r_2=1,2$. The expansion given by (30) is, intuitively, a polynomial of $\left(\hat{\theta}_{IL}-\theta_0\right)$. To obtain an expansion for $\left(\hat{\theta}_{IL}-\theta_0\right)$ that is not a function of itself, (30) 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.10

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

where $a,b,c,d,e,f,g,h,i,m\in\{1,2\}$.

Proof. In what follows, only when $\nu_{r_1,r_2,\ldots}$ is concerned, superscripts indicate the corresponding entry of the inverse of $\nu_{r_1,r_2,\ldots}$. 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 $(\nu'')^{-1}=[\nu^{r_1,r_2}]$.

The objective is to obtain an expression for a generic element of $\left(\hat{\theta}_{IL}-\theta_0\right)$, 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 (30). 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 (30) by $\nu^{r_1,m}$

$$\begin{aligned} \delta_I^m &= -\Bigg[\tilde{\ell}_{r_1}\nu^{r_1,m} + \mathcal{D}_{1;r_1}\nu^{r_1,m} + \delta_I^{r_2}\mathcal{H}_{r_1,r_2}\nu^{r_1,m} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\nu_{r_1,r_2,r_3}\nu^{r_1,m} + \mathcal{D}_{3;r_1}\nu^{r_1,m} \\ &\quad + \delta_I^{r_2}\mathcal{D}_{2;r_1,r_2}\nu^{r_1,m} + \frac{1}{2}\delta_I^{r_2}\delta_I^{r_3}\mathcal{H}_{r_1,r_2,r_3}\nu^{r_1,m} + \frac{1}{6}\delta_I^{r_2}\delta_I^{r_3}\delta_I^{r_4}\nu_{r_1,r_2,r_3,r_4}\nu^{r_1,m}\Bigg] + O_p\left(\frac{1}{T^2}\right). \end{aligned} \tag{31}$$

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 (31).

$$\begin{aligned} \delta_I^{r_2} &= -\Bigg[\tilde{\ell}_a\nu^{a,r_2} + \mathcal{D}_{1;a}\nu^{a,r_2} + \delta_I^b\mathcal{H}_{a,b}\nu^{a,r_2} + \frac{1}{2}\delta_I^b\delta_I^c\nu_{a,b,c}\nu^{a,r_2} + \mathcal{D}_{3;a}\nu^{a,r_2} \\ &\quad + \delta_I^b\mathcal{D}_{2;a,b}\nu^{a,r_2} + \frac{1}{2}\delta_I^b\delta_I^c\mathcal{H}_{a,b,c}\nu^{a,r_2} + \frac{1}{6}\delta_I^b\delta_I^c\delta_I^d\nu_{a,b,c,d}\nu^{a,r_2}\Bigg] + O_p\left(\frac{1}{T^2}\right), \\ \delta_I^{r_3} &= -\Bigg[\tilde{\ell}_e\nu^{e,r_3} + \mathcal{D}_{1;e}\nu^{e,r_3} + \delta_I^f\mathcal{H}_{e,f}\nu^{e,r_3} + \frac{1}{2}\delta_I^f\delta_I^g\nu_{e,f,g}\nu^{e,r_3} + \mathcal{D}_{3;e}\nu^{e,r_3} \\ &\quad + \delta_I^f\mathcal{D}_{2;e,f}\nu^{e,r_3} + \frac{1}{2}\delta_I^f\delta_I^g\mathcal{H}_{e,f,g}\nu^{e,r_3} + \frac{1}{6}\delta_I^f\delta_I^g\delta_I^h\nu_{e,f,g,h}\nu^{e,r_3}\Bigg] + O_p\left(\frac{1}{T^2}\right), \end{aligned} \tag{32}$$

$$\begin{aligned} \delta_I^{r_4} &= -\Bigg[\tilde{\ell}_i\nu^{i,r_4} + \mathcal{D}_{1;i}\nu^{i,r_4} + \delta_I^j\mathcal{H}_{i,j}\nu^{i,r_4} + \frac{1}{2}\delta_I^j\delta_I^k\nu_{i,j,k}\nu^{i,r_4} + \mathcal{D}_{3;i}\nu^{i,r_4} \\ &\quad + \delta_I^j\mathcal{D}_{2;i,j}\nu^{i,r_4} + \frac{1}{2}\delta_I^j\delta_I^k\mathcal{H}_{i,j,k}\nu^{i,r_4} + \frac{1}{6}\delta_I^j\delta_I^k\delta_I^l\nu_{i,j,k,l}\nu^{i,r_4}\Bigg] + O_p\left(\frac{1}{T^2}\right). \end{aligned} \tag{33}$$

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

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

Next, (32) and (33) are substituted for $\delta_I^{r_3}$ and $\delta_I^{r_4}$, respectively, which yields

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

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,

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

which proves Lemma A.10. ■

Based on these results, the Proof of Proposition 3.4 now follows.

Proof (Proposition 3.4). Follows directly from Lemma A.10, by observing that, $-\tilde{\ell}_a\nu^{a,m}$ is $O_p(T^{-1/2})$, $\tilde{\ell}_a\nu^{a,b}\mathcal{H}_{c,b}\nu^{c,m} - \mathcal{D}_{1;a}\nu^{a,m} - \frac{1}{2}\tilde{\ell}_a\nu^{a,b}\tilde{\ell}_c\nu^{c,d}\nu_{e,b,d}\nu^{e,m}$ is $O_p(T^{-1})$ and the remaining terms up to the $O_p(T^{-2})$ remainder are all $O_p\left(T^{-3/2}\right)$. Then, writing the first two lines in matrix notation finally gives (9). ■

Proof (Lemma A.8). The proof of Lemma A.8 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,

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

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,

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

The third order derivative is then given by

$$\begin{aligned}
\nabla_{\theta\theta\theta}\left[\frac{V_{iT}^{\lambda\lambda}\left(\ell_{iT}^{\lambda}\right)^2}{E_{iT}^2}\right] &= \frac{V_{r_1,r_2,r_3}\ell^2+2V_{r_1,r_2}\ell\ell_{r_3}[3]+2V_{r_1}\ell_{r_3}\ell_{r_2}[3]}{E^2}+2\frac{V_{r_1}\ell\ell_{r_2,r_3}[3]+V\ell_{r_2,r_3}\ell_{r_1}[3]+V\ell\ell_{r_1,r_2,r_3}}{E^2}\\
&\quad -2\frac{V_{r_1,r_2}\ell^2E_{r_3}[3]+2V_{r_1}\ell\ell_{r_2}E_{r_3}[6]+2V\ell_{r_2}\ell_{r_1}E_{r_3}[3]+2V\ell\ell_{r_1,r_2}E_{r_3}[3]}{E^3}\\
&\quad -2\frac{V_{r_1}\ell^2E_{r_2,r_3}[3]+2V\ell\ell_{r_1}E_{r_2,r_3}[3]+V\ell^2E_{r_1,r_2,r_3}}{E^3}\\
&\quad +6\frac{V_{r_1}\ell^2E_{r_2}E_{r_3}[3]+2V\ell\ell_{r_1}E_{r_2}E_{r_3}[3]}{E^4}+6\frac{V\ell^2E_{r_1,r_2}E_{r_3}[3]}{E^4}\\
&\quad -24\frac{V\ell^2E_{r_1}E_{r_2}E_{r_3}}{E^5}.
\end{aligned}$$

Lastly,

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

■

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. (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, vol. 5, pp. 3229–3296. North-Holland, Amsterdam.

Bai, J. (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77, 1229–1279.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

——— (2004): "A Panic Attack on Unit Roots and Cointegration," *Econometrica*, 72, 1127–1177.

Barndorff-Nielsen, O. E. (1983): "On a Formula for the Distribution of the Maximum Likelihood Estimator," *Biometrika*, 70, 343–65.

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 Realised 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.

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.

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.

Boussama, F. (1998): "Ergodicite, Melange et Estimation dans les Modeles GARCH," PhD dissertation, Paris-7 University.

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

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.

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.

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.

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 (2006): "Mixing Properties of a General Class of GARCH(1,1) Models without Moment Assumptions on the Observed Process," *Econometric Theory*, 22, 815–834.

——— (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*, forthcoming.

Hahn, J., and H. R. Moon (2006): "Reducing Bias of MLE in a Dynamic Panel Model," *Econometric Theory*, 22, 499–512.

Hahn, J., and W. K. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72(4), 1295–1319.

Hall, P. G. (1992): *The Bootstrap and Edgeworth Expansion*. Springer, New York.

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.

Islam, N. (1995): "Growth Empirics: A Panel Data Approach," *Quarterly Journal of Economics*, 110, 1127–1170.

Kapetanios, G., M. H. Pesaran, and T. Yamagata (2011): "Panels with Non-Stationary Multifactor Error Structures," *Journal of Econometrics*, 160, 326–348.

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.

Lindner, A. M. (2009): "Stationarity, Mixing, Distributional Properties and Moments of GARCH(p,q)-Processes," in *Handbook of Financial Time Series*, ed. by T. G. Andersen, R. A. Davis, J. P. Kreiss, and T. Mikosch, pp. 43–69. Springer-Verlag.

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.

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.

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 Dixed 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 Verlag.

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.

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

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 $\beta = 0.93$.

T	CL		InCL		ICL		IPCL		CL		InCL		ICL		IPCL	
	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$
	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 $\beta = 0.93$.

	CL		InCL		IPCL		CL		InCL		IPCL	
T	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$\bar{\sigma}_{\hat{\alpha}}$	$\bar{\sigma}_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$	$R_{\hat{\alpha}}$	$R_{\hat{\beta}}$
	N=100											
400	.002	.004	.002	.003	.002	.004	.005	.006	.004	.005	.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 $\beta = 0.93$.

	QML vs CL		CL vs ICL		QML vs ICL	
Stock	t-stat	Result	t-stat	Result	t-stat	Result
Alcoa	0.993	-	2.563	ICL	2.144	ICL
American Express	1.451	-	3.728	ICL	3.920	ICL
Bank of America	1.109	-	2.550	ICL	2.968	ICL
Du Pont	2.288	CL	1.168	-	2.044	ICL
General Electric	0.960	-	2.176	ICL	2.428	ICL
IBM	1.815	-	1.419	-	2.036	ICL
Coca Cola	2.870	CL	-1.372	-	1.093	-
Microsoft	2.755	CL	-1.381	-	0.844	-
Exxon Mobil	1.404	-	0.986	-	1.650	-

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 (10).

	$T=150$			$T=175$			$T=207$		
Strategy	#	$\hat{\alpha}$	$\hat{\beta}$	#	$\hat{\alpha}$	$\hat{\beta}$	#	$\hat{\alpha}$	$\hat{\beta}$
Security selection	52	.202	.788	34	.174	.820	26	.179	.815
Macro	25	.114	.884	17	.093	.907	15	.105	.893
Directional Traders	51	.208	.771	24	.153	.840	16	.161	.832
Fund of funds	78	.153	.847	41	.143	.857	25	.152	.836
Multi-process	28	.176	.824	19	.165	.835	15	.230	.770
Emerging	19	.220	.772	11	.176	.794	7	.176	.801
Fixed income	13	.249	.751	8	.195	.805	5	.229	.768
CTA	41	.090	.910	22	.061	.939	15	.072	.928

Table 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.

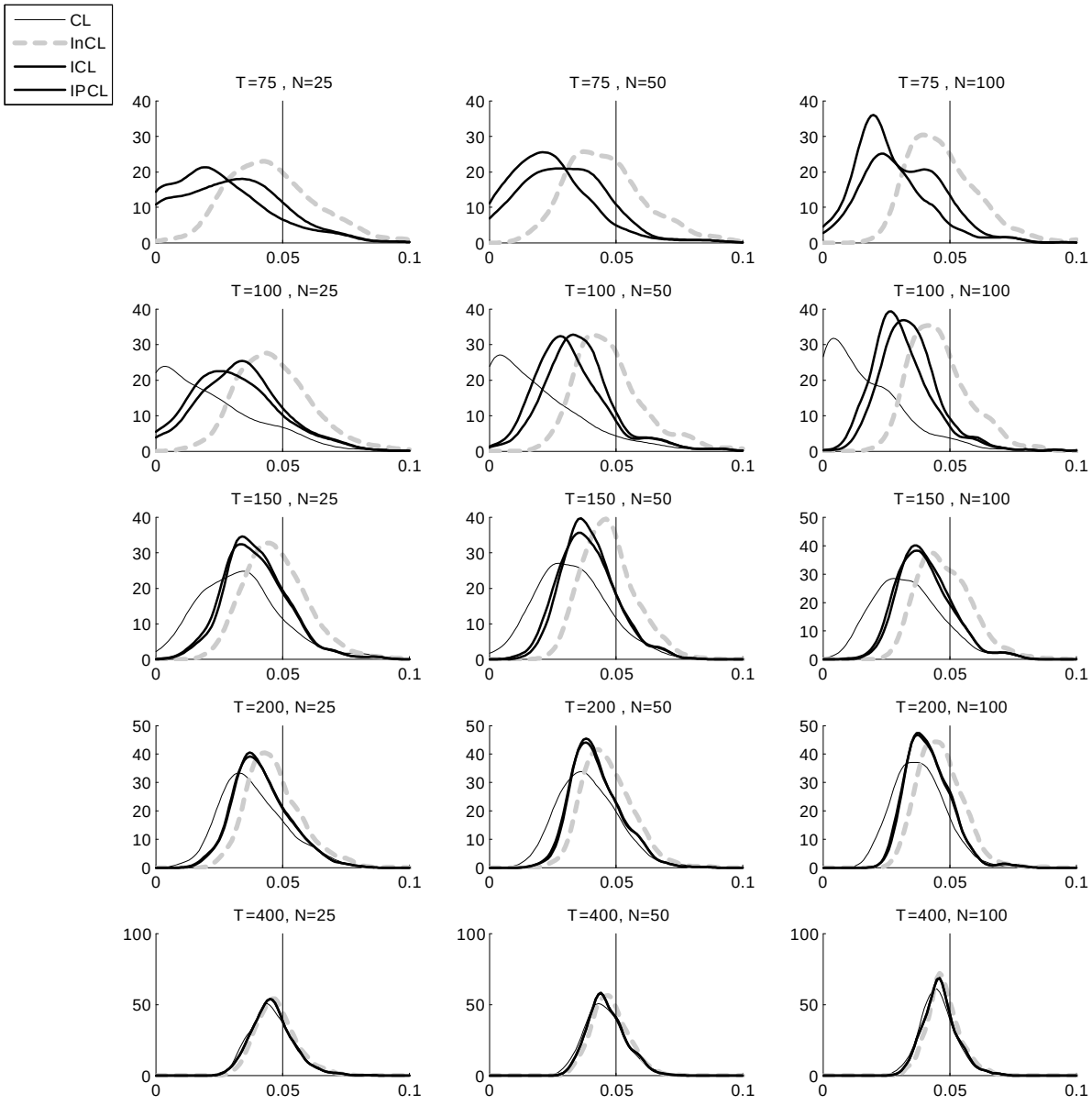

Figure 1: Sample distributions of $\hat{\alpha}$ using the Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL), and Integrated Pseudo CL (IPCL). The vertical line is drawn at the true parameter value. Based on 500 replications under cross-sectional dependence where $(\alpha, \beta) =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

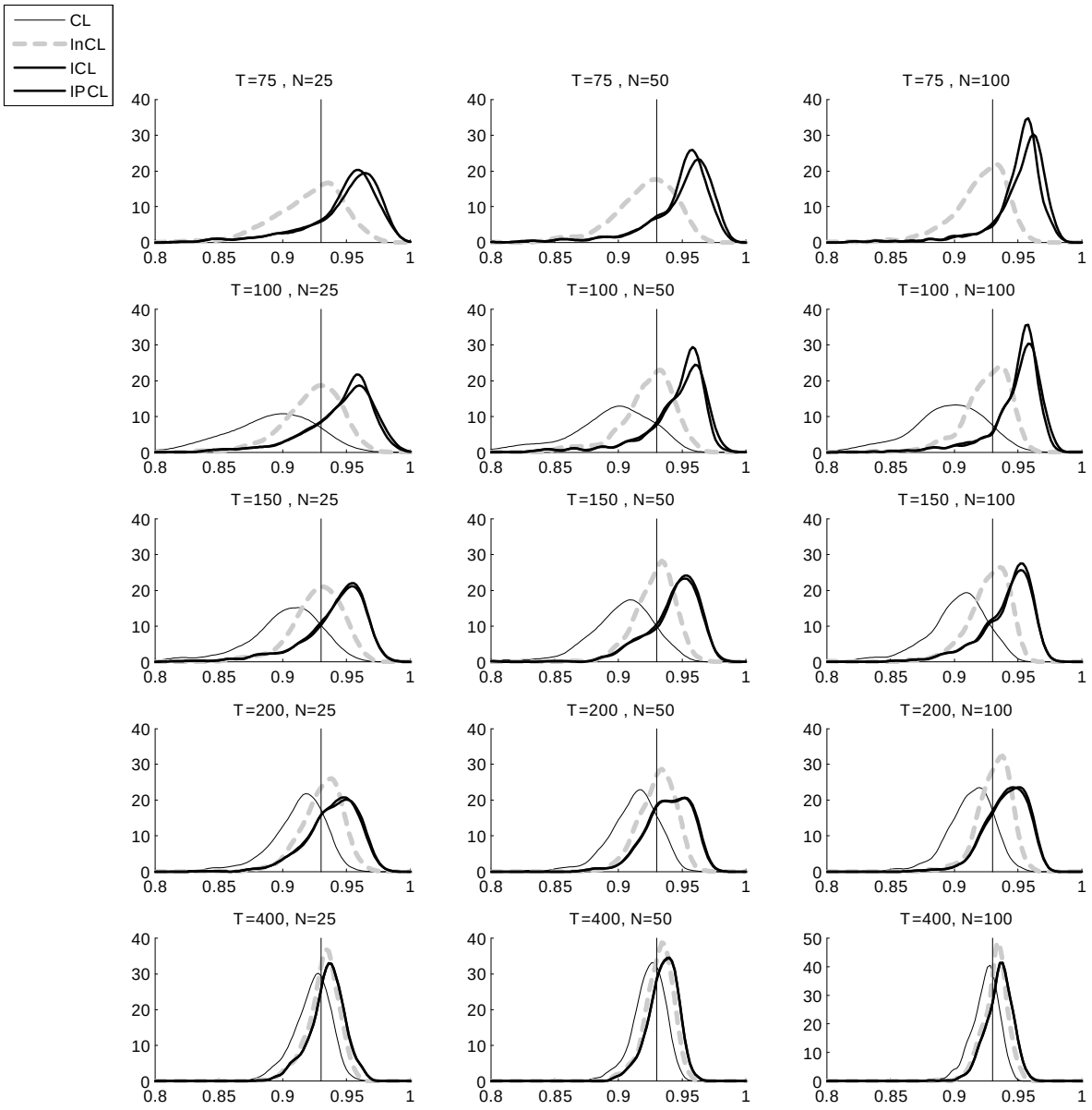

Figure 2: Sample distributions of $\hat{\beta}$ using the Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL), and Integrated Pseudo CL (IPCL). The vertical line is drawn at the true parameter value. Based on 500 replications under cross-sectional dependence where $(\alpha, \beta) =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

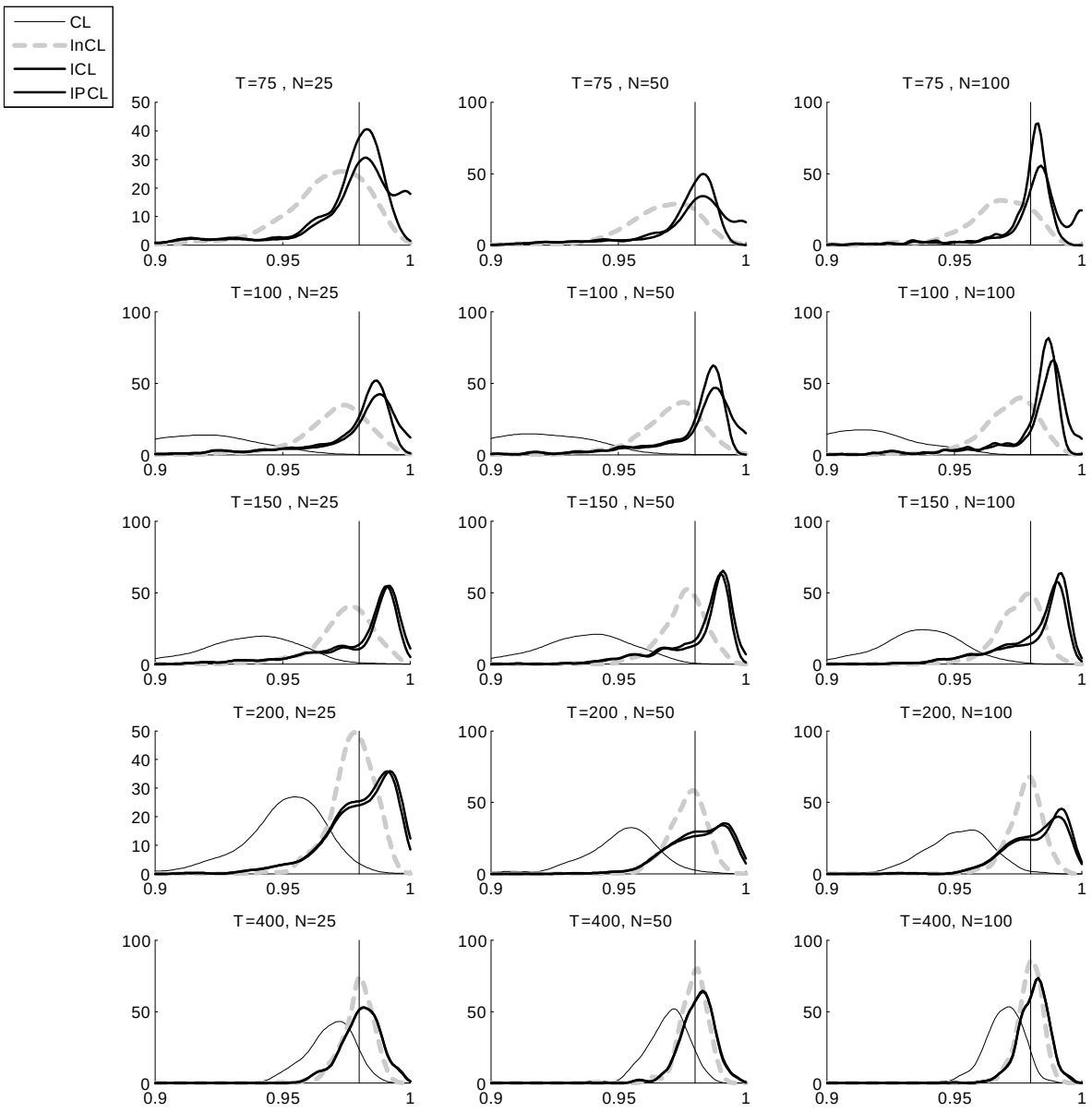

Figure 3: Sample distributions of $\hat{\alpha} + \hat{\beta}$ using the Composite Likelihood (CL), Infeasible CL (InCL), Integrated CL (ICL), and Integrated Pseudo CL (IPCL). The vertical line is drawn at the true parameter value. Based on 500 replications under cross-sectional dependence where $(\alpha,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

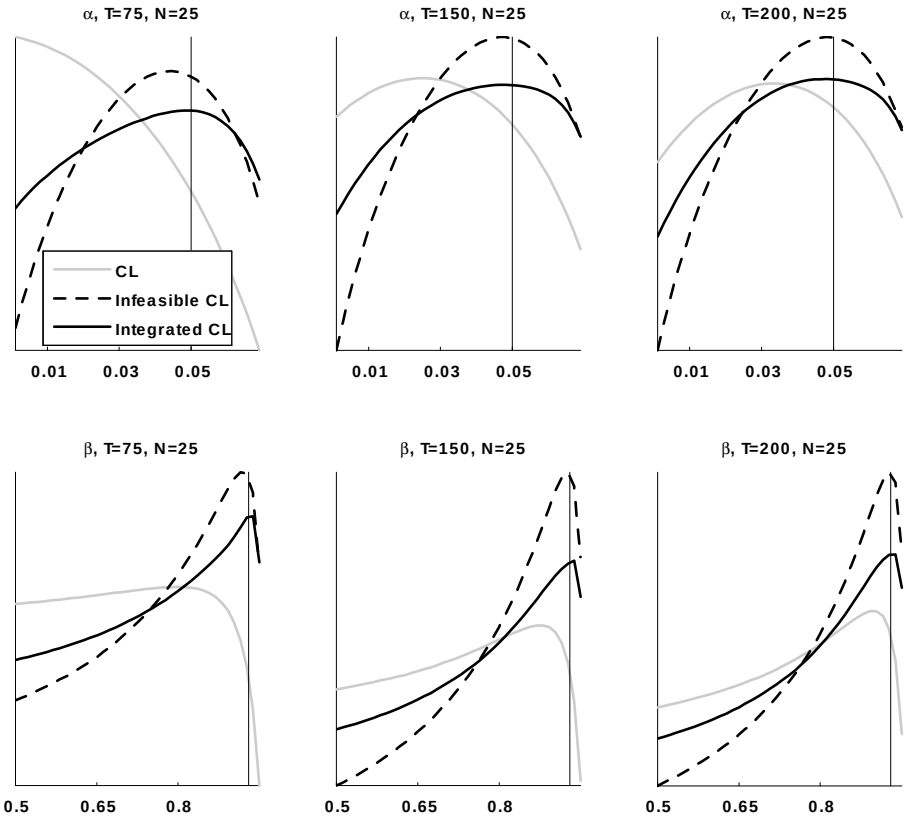

Figure 4: Average likelihood plots for α and β. Based on likelihood averages over 500 replications (with cross-sectional dependence). In the upper panel, β is fixed at 0.93 while the lower panel is based on $\alpha = 0.05$. CL is evaluated at the sample estimates of λ_i, while Infeasible CL is evaluated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

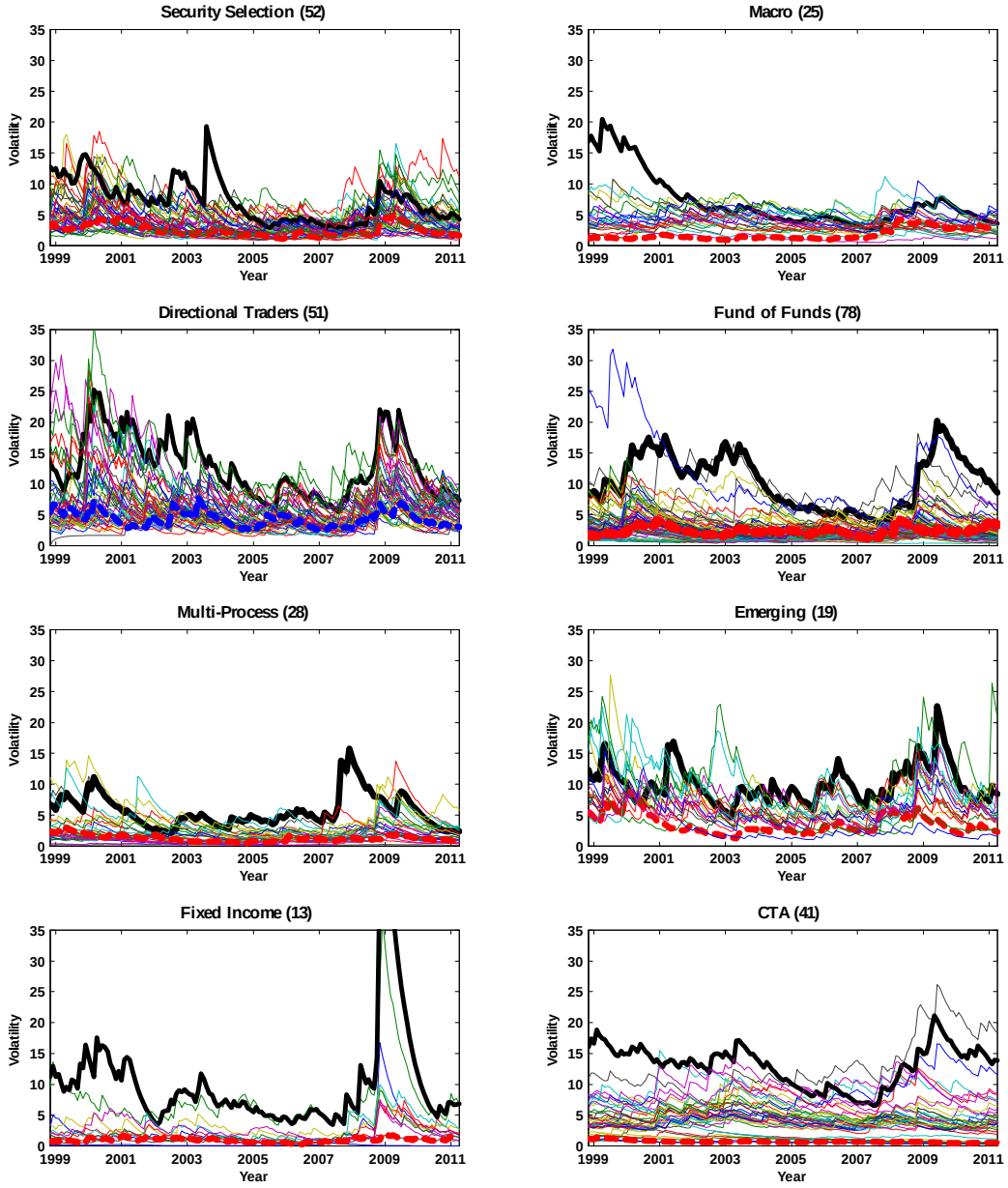

Figure 5: Conditional volatility plots. Based on parameters estimates by the integrated likelihood method using panels of funds that have reported non-zero returns between November 1998 and April 2011 (150 observations). Number of funds in each strategy-panel is given in parentheses. Random examples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

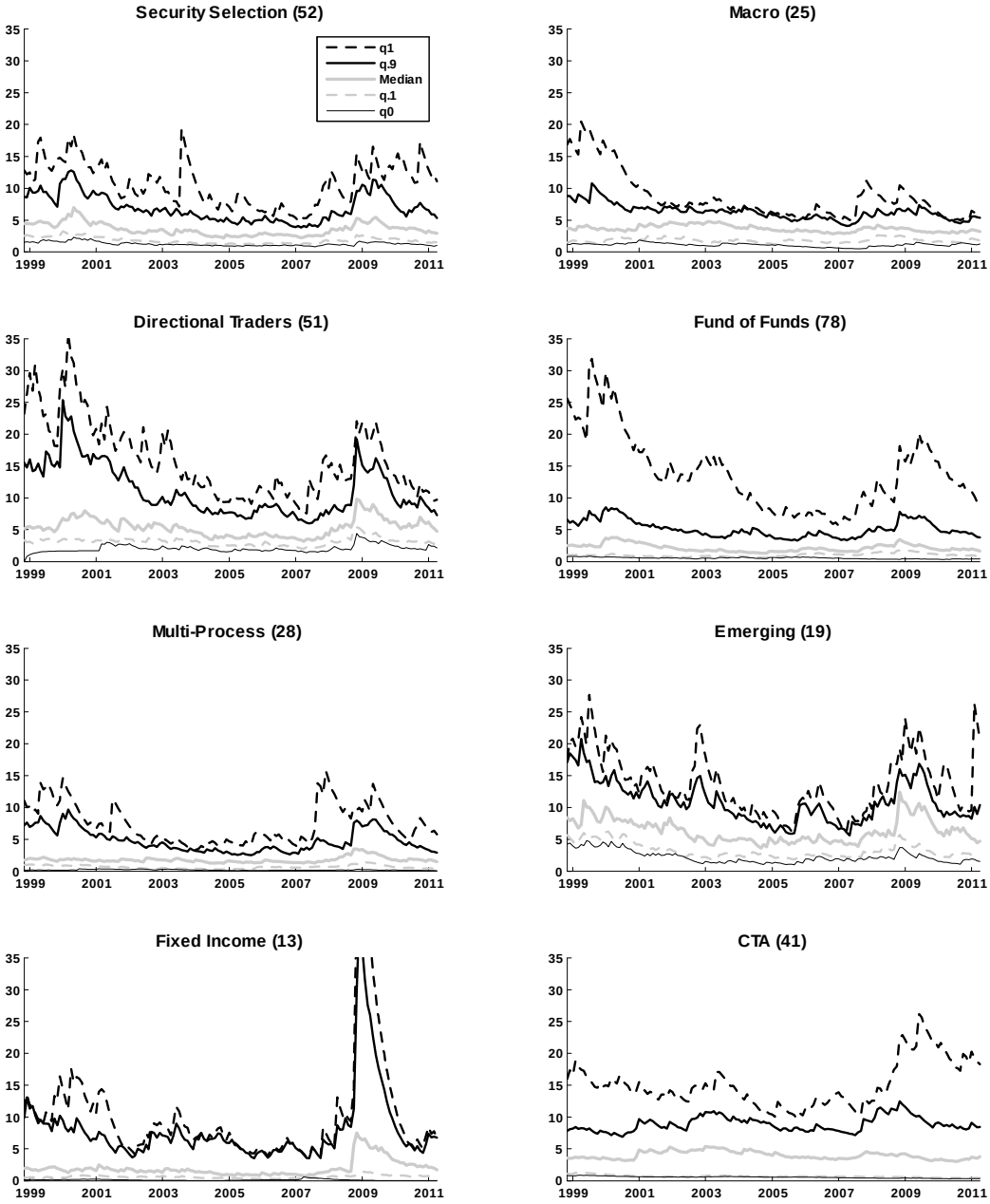

Figure 6: Plots of the $0\%, 10\%$, 50% (median), 90% and 100% quantiles of the sample distribution of volatility across funds. Based on fitted conditional volatilities displayed in Figure 5. Number of funds in each strategy is given in parentheses.

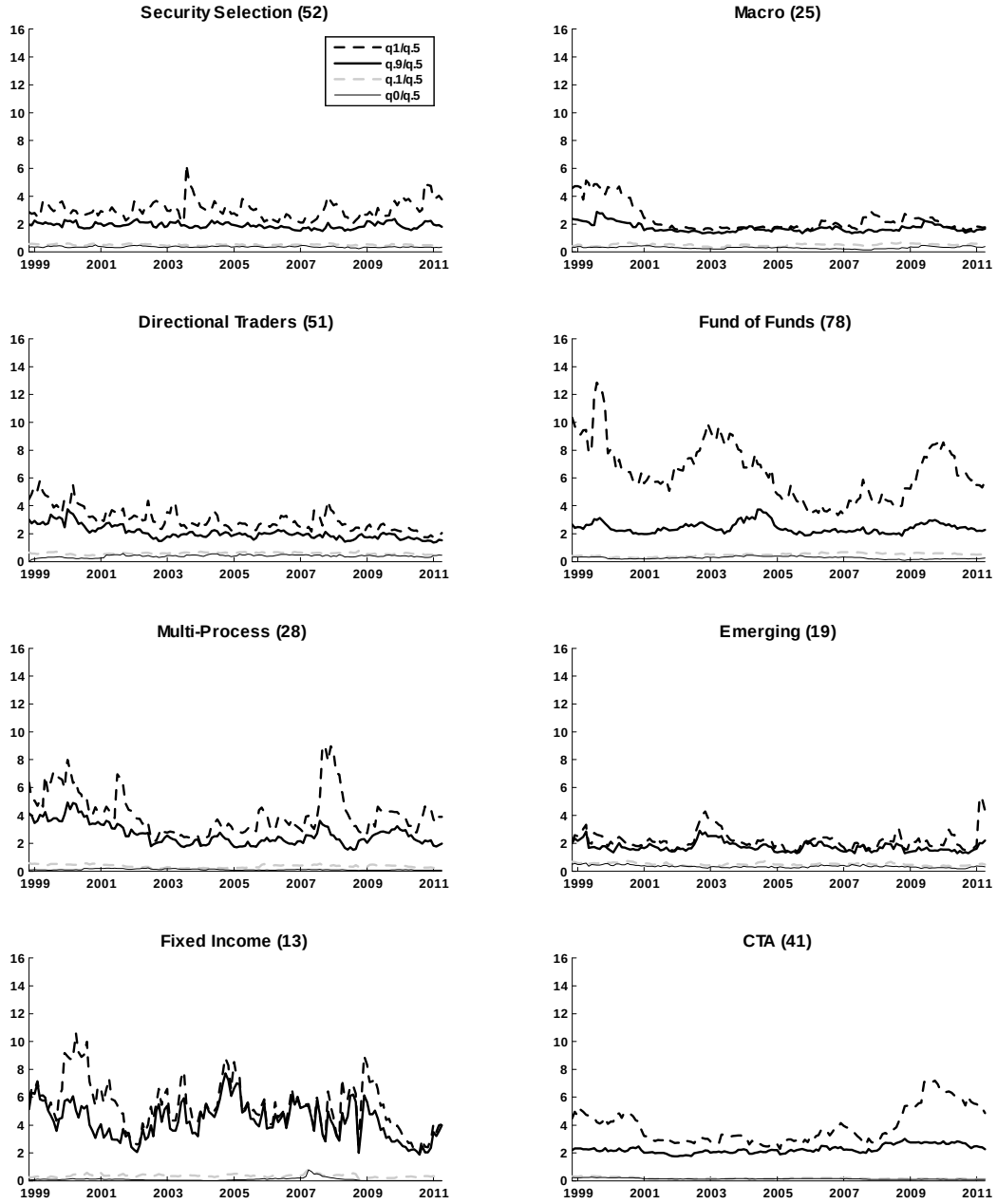

Figure 7: Plots of $0\%, 10\%$, 90% and 100% quantiles (normalised by the median) of the sample distribution of volatility across funds. Based on fitted conditional volatilities displayed in Figure 5. Number of funds in each strategy is given in parentheses.