CFS WORKING PAPER

No. 450

Efficient Iterative Maximum Likelihood Estimation of High-Parameterized Time Series Models

Nikolaus Hautsch, Ostap Okhrin, Alexander Ristig

Center for Financial Studies
Goethe University ■ House of Finance
Grüneburgplatz 1 ■ 60323 Frankfurt am Main

Tel: +49 69 798-30050
Fax: +49 69 798-30077
www.ifk-cfs.de ■ info@ifk-cfs.de

CFS Working Paper Series

The *Center for Financial Studies*, located in Goethe University's House of Finance in Frankfurt, is an independent non-profit research center, funded by the non-profit-making organisation Gesellschaft für Kapitalmarktforschung e.V. (GfK). The CFS is financed by donations and by contributions of the GfK members, as well as by national and international research grants. The GfK members comprise major players in Germany's financial industry. Established in 1967 and closely affiliated with the University of Frankfurt, it provides a strong link between the financial community and academia. CFS is also a contributor to policy debates and policy analyses, building upon relevant findings in its research areas.

The CFS Working Paper Series presents the result of scientific research on selected topics in the field of money, banking and finance. The authors were either participants in the Center´s Research Fellow Program or members of one of the Center´s Research Projects.

If you would like to know more about the Center for Financial Studies, please let us know of your interest.

Prof. Michalis Haliassos, Ph.D. Prof. Dr. Jan Pieter Krahnen Prof. Dr. Uwe Walz

Center for Financial Studies
Goethe University ■ House of Finance
Grüneburgplatz 1 ■ 60323 Frankfurt am Main

Tel: +49 69 798-30050
Fax: +49 69 798-30077
www.ifk-cfs.de ■ info@ifk-cfs.de

Efficient Iterative Maximum Likelihood Estimation of High-Parameterized Time Series Models

20th January 2014

Nikolaus Hautsch[*] **Ostap Okhrin**[†] **Alexander Ristig**[‡]

Abstract

We propose an iterative procedure to efficiently estimate models with complex log-likelihood functions and the number of parameters relative to the observations being potentially high. Given consistent but inefficient estimates of sub-vectors of the parameter vector, the procedure yields computationally tractable, consistent and asymptotic efficient estimates of all parameters. We show the asymptotic normality and derive the estimator's asymptotic covariance in dependence of the number of iteration steps. To mitigate the curse of dimensionality in high-parameterized models, we combine the procedure with a penalization approach yielding sparsity and reducing model complexity. Small sample properties of the estimator are illustrated for two time series models in a simulation study. In an empirical application, we use the proposed method to estimate the connectedness between companies by extending the approach by Diebold and Yilmaz (2014) to a high-dimensional non-Gaussian setting.

JEL classification: C13, C32, C50
Keywords: Multi-Step estimation, Sparse estimation, Multivariate time series, Maximum likelihood estimation, Copula.

1. Introduction

Statistical inference for models including many parameters is of growing interest in various fields in econometrics and statistics. Examples include high-dimensional vector autoregressive moving average

[*]Department of Statistics and Operations Research, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria, nikolaus.hautsch@univie.ac.at

[†]Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany, ostap.okhrin@hu-berlin.de

[‡]Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany, alexander.ristig@hu-berlin.de

The research was supported by the Deutsche Forschungsgemeinschaft through the CRC 649 "Economic Risk", Humboldt-Universität zu Berlin. Hautsch acknowledges support by the Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF).

(VARMA) models, multivariate generalized autoregressive conditional heteroscedasticity (GARCH) models, vector multiplicative error models (VMEMs) or corresponding copula approaches. Such models are mostly estimated using multi-step approaches constructed from parts of the log-likelihood. Such estimators are typically inefficient and their asymptotic distributions are difficult to compute as asymptotic results for multi-step likelihood procedures are generally widely missing.

In this paper, we address the situation of a complex, possibly highly parameterized log-likelihood function (in terms of the number of parameters relative to the sample size) whose first- and second-order derivatives cannot necessarily be derived analytically. Complexity can arise from nonlinearities in the underlying model and/or if the number of parameters r is high relative to the number of observations. In such a situation, a one-step optimization of the log-likelihood is typically (computationally or numerically) not possible and parameters have to be estimated in multiple steps. The contribution of this paper is to propose an asymptotically efficient and computationally tractable iterative estimation algorithm and to derive the asymptotic distribution of the estimates in dependence of the number of underlying iteration steps.

Our approach rests on the assumption of the existence of a consistent but (eventually highly) inefficient estimator of a vector of parameters ϑ of a log-likelihood function $\mathcal{L}(\vartheta)$. For example, ϑ might be consistently but inefficiently estimated by a 2-stage procedure. To obtain an efficient and computationally feasible estimator, we suggest splitting the estimation problem into appropriate computationally tractable sub-problems. In particular, we decompose ϑ into G sub-vectors $\vartheta_1, \ldots, \vartheta_G$ of arbitrary size, and maximize $\mathcal{L}(\cdot)$ iteratively with respect to ϑ_g, $g = 1, \ldots, G$, holding fixed all other parameters which have been updated in previous iteration steps. We show the consistency and asymptotic normality of the resulting estimator ϑ_n^h in dependence of the number of iterations h and show that it is asymptotically efficient as $h \to \infty$. Moreover, we illustrate how to combine the procedure with penalization techniques as, e.g., the smoothly clipped absolute deviation (SCAD) penalty introduced by Fan and Li (2001). This step yields sparse estimates and allows diminishing the curse of dimensionality arising from highly parameterized models.

Our major focus is on time series models where the number of parameters relative to the number of observations is high and thus it is computationally challenging or virtually impossible to optimize the entire log-likelihood in one step. The algorithm and corresponding asymptotic theory, however, can also be applied to other estimation and inference problems. The asymptotic distribution of the iterative estimation procedure in dependence of the exact number of iterations is particularly useful since researchers limit the latter in realistic applications. As illustrated in the paper, these results can be used, among others, to easily establish the asymptotic efficiency of the feasible generalized least squares (FGLS) estimator.

Closest to our approach is the procedure proposed by Song, Fan, and Kalbfleisch (2005) who suggest decomposing the log-likelihood into a so-called (simple) working and a (complicated) error part. While the analytical first- and second-order derivatives can be computed for the working part, there is no analytical second-order derivative available for the error part. Then, the log-likelihood's first order condition – evaluating the error part at the estimate from previous step – is solved to update the estimator. Our approach, however, differs in two important respects: Firstly, our algorithm relies on the decomposition of the parameter space into G sub-spaces and thus is more flexible if ϑ is large. Secondly, we do not require the analytical first-order derivative which makes it more tractable if the

underlying model is complex.

However, the drawback of relying on a derivative-free optimization of $\mathcal{L}(\cdot)$ is that each sub-vector, ϑ_g, $g = 1, \ldots, G$, should realistically only consist of a few parameters, say up to 10, inducing a curse of dimensionality if ϑ is large. To address the latter and to keep the number of sub-vectors G small, we combine the underlying log-likelihood with a non-concave penalization function, as, for instance, the least absolute shrinkage and selection operator (LASSO), see Tibshirani (1996, 2011), or the SCAD penalty function see Fan and Li (2001). This step makes our approach applicable in high dimensions and thus useful for many comprehensive applications. We derive the asymptotic properties of the resulting *sparse* iterative procedure building on the results by Fan and Li (2001).

The small-sample performance of the procedure is illustrated in two comprehensive simulation studies. The first one investigates the properties of ϑ_n^h for a 5-dimensional VARMA model including 24 parameters based on 50 observations. In the second simulation study, we analyze the performance of our estimator for a 15-dimensional VMEM containing 375 parameters based on a sample size of 500. We illustrate that our proposed procedure significantly simplifies the underlying estimation problem and performs sufficiently well even in these inherently high-dimensional settings. Finally, we apply our approach to measure volatility connections between 30 companies by extending the connectedness measure introduced by Diebold and Yilmaz (2014) to a high-dimensional and non-Gaussian setting. This requires estimating prediction error variance decompositions based on a 30-dimensional MA(∞) process of realized volatilities. To allow for a non-Gaussian joint distribution, we model the joint dependence using Vine copulae, see Kurowicka and Joe (2011), and compute the final connectedness measure based on simulated (generalized) prediction error variance decompositions. The resulting model consists of 1860 parameters which are efficiently and sparsely estimated using our approach. Overall, the examples show that the proposed estimation technique performs well even in challenging settings and can serve as a working horse for parameter estimation in complex situations.

The paper is organized as follows. Section 2 and 3 discuss the estimation details. Section 4 illustrates an application of the procedure in a generalized least squares setting. Section 5 shows the performance of the estimator in two simulation studies. Section 6 presents the empirical application and Section 7 concludes. Proofs are moved to Appendix A.

2. Efficient Multi-Step Estimation

Let the observed data x be a realization of the finite history $X \overset{\text{def}}{=} (X_1^\top, \ldots, X_n^\top)^\top$ of the d-dimensional stochastic process $\{X_i : \Omega \to \mathbb{R}^d, d \in \mathbb{N}, i = 1, 2, \ldots\}$, which is defined on a complete probability space $(\Omega, \mathcal{F}, \mathrm{P})$ with $X_i \overset{\text{def}}{=} (X_{i1}, \ldots, X_{id})^\top$. Let the absolutely continuous probability measure P describe the complete probabilistic behavior of X. Equivalently, the stochastic behavior of X can also be characterized by the measurable Radon-Nikodým density of P denoted by $f(X)$. Based on the σ-field $\mathcal{F}_{i-1} \overset{\text{def}}{=} \{X_l : l \leq i-1\}$ and conditional density $f_i(\cdot) \overset{\text{def}}{=} f_{X_i|\mathcal{F}_{i-1}}(X_{i1}, \ldots, X_{id})$, we rewrite the density as $f(X) = \prod_{i=1}^n f_i(X_{i1}, \ldots, X_{id})$. Unless stated differently, we assume that $\mathrm{P} \in \mathcal{P}$, with $\mathcal{P} = \{\mathrm{P}_\vartheta : \vartheta \in \Theta \subseteq \mathbb{R}^r, r \in \mathbb{N}\}$, so that the density of P is given by $f(\cdot; \vartheta) = \prod_{i=1}^n f_i(\cdot; \vartheta)$, which is assumed to be measurable for each $\vartheta \in \Theta$ and absolutely continuous on the parameter space Θ.

Assume that the parameter vector $\vartheta \in \Theta = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_G$ can be split up into G sub-vectors

ϑ_g, $g = 1, \ldots, G$, each consisting of r_g components, $g = 1, \ldots, G$, with $r = \sum_{g=1}^{G} r_g$ denoting the total number of parameters. Let $\vartheta_0 = \mathfrak{v}(\vartheta_{1,0}, \ldots, \vartheta_{G,0})$ denote the (true) parameter vector where the $\mathfrak{v}(\cdot)$ operator vectorizes vectors of possibly different dimension, i.e., $\mathfrak{v}(\vartheta_1, \ldots, \vartheta_G) \stackrel{\text{def}}{=} (\vartheta_1^\top, \ldots, \vartheta_G^\top)^\top$. The underlying log-likelihood function is given by $\mathcal{L}(\vartheta) \stackrel{\text{def}}{=} \mathcal{L}(\vartheta; X) = \sum_{i=1}^{n} \ell_i(\vartheta)$ with $\ell_i(\vartheta) \stackrel{\text{def}}{=} \log f_i(X_{i1}, \ldots, X_{id}; \vartheta)$. According to Sklar's Theorem, see Sklar (1959), each d-dimensional distribution function can be decomposed into its conditional marginal distribution functions and a conditional dependence component – the copula function. Consequently, $\ell_i(\vartheta)$ can be decomposed into the log copula density $\ell_i^c(\vartheta_1, \ldots, \vartheta_G) \stackrel{\text{def}}{=} \ell_i^c(\vartheta)$ depending on ϑ and the sum of log marginal densities $\ell_i^m(\vartheta_1, \ldots, \vartheta_k) \stackrel{\text{def}}{=} \sum_{j=1}^{d} \log f_{X_{ij}|\mathcal{F}_{i-1}}(\cdot; \vartheta_1, \ldots, \vartheta_k)$ depending on $p = \sum_{g=1}^{k} r_g$ parameters split into $k < G$ groups, $\vartheta_1, \ldots, \vartheta_k$. See Joe (1997), Nelsen (2006), and Jaworski, Durante, and Härdle (2013) for comprehensive overviews of copulae and several examples. The log-likelihood can be then written as

$$\mathcal{L}(\vartheta) = \sum_{i=1}^{n} \{\ell_i^m(\vartheta_1, \ldots, \vartheta_k) + \ell_i^c(\vartheta)\} = \mathcal{L}^m(\vartheta_1, \ldots, \vartheta_k) + \mathcal{L}^c(\vartheta), \tag{1}$$

where $\mathcal{L}^m(\cdot) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \ell_i^m(\cdot)$ denotes the marginal and $\mathcal{L}^c(\cdot) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \ell_i^c(\cdot)$ the copula part. To keep the notation simple, we define

$$\dot{\mathcal{L}}(\vartheta_0) \stackrel{\text{def}}{=} \left.\frac{\partial \mathcal{L}(\vartheta)}{\partial \vartheta}\right|_{\vartheta=\vartheta_0} \quad \text{and} \quad \ddot{\mathcal{L}}(\vartheta_0) \stackrel{\text{def}}{=} \left.\frac{\partial^2 \mathcal{L}(\vartheta)}{\partial \vartheta \partial \vartheta^\top}\right|_{\vartheta=\vartheta_0}$$

and accordingly

$$\dot{\mathcal{L}}_{\vartheta_g}(\vartheta_0) \stackrel{\text{def}}{=} \left.\frac{\partial \mathcal{L}(\vartheta)}{\partial \vartheta_g}\right|_{\vartheta=\vartheta_0} \quad \text{and} \quad \ddot{\mathcal{L}}_{\vartheta_g,\vartheta_l}(\vartheta_0) \stackrel{\text{def}}{=} \left.\frac{\partial^2 \mathcal{L}(\vartheta)}{\partial \vartheta_g \partial \vartheta_l^\top}\right|_{\vartheta=\vartheta_0},$$

with $g, l = 1, \ldots, G$. An analogous notation is used for the components $\mathcal{L}^m(\cdot)$ and $\mathcal{L}^c(\cdot)$. Expectations are taken with respect to the measure P and defined as $\mathrm{E}(\cdot) \stackrel{\text{def}}{=} \mathrm{E}_{\vartheta_0}(\cdot) = \mathrm{E}_{\mathrm{P}}(\cdot)$. For a sequence of random variables, vectors or matrices U_n is said to be bounded in probability, the notation $U_n = \mathcal{O}_p(1)$ is used. The notation $U_n = o_p(V_n)$ implies for two sequences of compatible random variables, vectors or matrices U_n and V_n, $U_n V_n^{-1} \stackrel{\mathrm{P}}{\to} 0$.

2.1. Iterative Estimation

In non-linear models, first- and second-order derivatives of $\mathcal{L}(\cdot)$ are typically complicated, making the maximization of (1) with respect to ϑ challenging. If, moreover, the number of underlying parameters, r, is high (either absolutely or relative to the sample size n), one-step estimation is often numerically impossible. In these situations, it is inevitable to simplify the estimation problem by breaking it up into lower-dimensional and/or less demanding problems which can be solved individually. In most situations, however, the resulting estimators are inefficient since the dependence between the sub-components is neglected in the estimation. Addressing this shortcoming makes it necessary to apply a multi-step estimation procedure which iterates through all sub-model estimations and thus

allows estimates to be successively updated exploiting information from the other steps. A well-known example of this proceeding is the FGLS estimation of a heteroscedastic linear regression model which is efficiently estimated by iterating various times between (covariance-based weighted) least squares estimations of slope parameters and corresponding covariance estimations.

The iterative algorithm proposed in this paper builds on the idea of iterative multi-step estimation relying on arbitrary decompositions of ϑ into k sub-vectors associated with the marginals and $G-k$ sub-vectors associated with the copula function. Assuming consistency though inefficiency of the (initial) estimator in *Step h = 1* below, we propose the following algorithm:

Algorithm 1.
Step $h = 1$*:*

(1) $\mathfrak{v}(\vartheta_{1,n}^1, \ldots, \vartheta_{k,n}^1) = \arg \underset{\mathfrak{v}(\vartheta_1,\ldots,\vartheta_k)}{\text{zero}} \dot{\mathcal{L}}^m(\vartheta_1, \ldots, \vartheta_k)$

(2) $\mathfrak{v}(\vartheta_{k+1,n}^1, \ldots, \vartheta_{G,n}^1) = \arg \underset{\mathfrak{v}(\vartheta_{k+1},\ldots,\vartheta_G)}{\text{zero}} \dot{\mathcal{L}}^c_{\mathfrak{v}(\vartheta_{k+1},\ldots,\vartheta_G)}(\vartheta_{1,n}^1, \ldots, \vartheta_{k,n}^1, \vartheta_{k+1}, \ldots, \vartheta_G)$

Step $h > 1$*:*

(1) $\vartheta_{1,n}^h = \arg \max_{\vartheta_1} \mathcal{L}(\vartheta_1, \vartheta_{2,n}^{h-1}, \ldots, \vartheta_{G,n}^{h-1})$

(2) $\vartheta_{2,n}^h = \arg \max_{\vartheta_2} \mathcal{L}(\vartheta_{1,n}^h, \vartheta_2, \vartheta_{3,n}^{h-1}, \ldots, \vartheta_{G,n}^{h-1})$

$\vdots$

(G) $\vartheta_{G,n}^h = \arg \max_{\vartheta_G} \mathcal{L}(\vartheta_{1,n}^h, \ldots, \vartheta_{G-1,n}^h, \vartheta_G)$

The 2-stage procedure at *Step 1* of Algorithm 1 is well known as inference functions for margins and a simple way to obtain consistent estimators of parametric copula-based models, see Joe and Xu (1996). Starting with the initial estimates in *Step 1*, the Algorithm 1 builds on an iterative estimation of the parameters of each ϑ_g, given the parameters of the other groups ϑ_l, $l \neq g$, $g = 1, \ldots, G$, estimated in the (instantaneous) previous steps. For a discussion in the context of non-consistent initial estimators, we refer to Song *et al.* (2005) and the references therein.

2.2. Asymptotic Properties

We assume that the maximum likelihood (ML) estimator $\vartheta_n = \mathfrak{v}(\vartheta_{1,n}, \ldots, \vartheta_{G,n})$ of ϑ can be formulated as the maximizer of $\mathcal{L}(\vartheta)$ obtained from solving $\dot{\mathcal{L}}(\vartheta) = 0$. To show the consistency of ϑ_n^h $\forall h$ in Theorem 1 below we need the following set of assumptions:

Assumption 1. *The model is identifiable and the true value* ϑ_0 *is an interior point of the compact parameter space* Θ*. We assume that the model is correctly specified in the sense that* $\mathrm{E}_\vartheta\{\partial \ell_i(\vartheta)/\partial \vartheta_g\} = 0$ *and information equality holds,*

$$\mathcal{I}_{i,gl}(\vartheta) \overset{\text{def}}{=} \mathrm{E}_\vartheta \left\{ \frac{\partial \ell_i(\vartheta)}{\partial \vartheta_g} \frac{\partial \ell_i(\vartheta)}{\partial \vartheta_l^\top} \right\} = -\mathrm{E}_\vartheta \left\{ \frac{\partial^2 \ell_i(\vartheta)}{\partial \vartheta_g \partial \vartheta_l^\top} \right\},$$

for $g, l = 1, \ldots, G$ *and* $i = 1, \ldots, n$.

Assumption 2. *The information matrix is* $\mathcal{I}(\vartheta) = \sum_{i=1}^{n} \mathcal{I}_i(\vartheta)$, *with* $\mathcal{I}_i(\vartheta) = \{\mathcal{I}_{i,gl}(\vartheta)\}_{g,l=1}^{G}$. *Let the limit of* $n^{-1}\mathcal{I}(\vartheta) \xrightarrow{\mathrm{P}} \mathcal{J}(\vartheta)$ *be the asymptotic information matrix, which is finite and positive definite at* ϑ_0 *and* $n^{-1}\ddot{\mathcal{L}}(\vartheta) \xrightarrow{\mathrm{P}} \mathcal{H}(\vartheta)$ *be the asymptotic Hessian, which is finite and negative definite for* $\vartheta \in \{\vartheta : ||\vartheta - \vartheta_0|| < \delta\}$, $\delta > 0$.

Then, we can state the following theorem:

Theorem 1. *Let the random variables of the sequence* X *have an identical conditional density* $f_i(\cdot; \vartheta)$ *for which Assumptions 1-2 hold. If* $\vartheta_n^1 \xrightarrow{\mathrm{P}} \vartheta_0$, *then* $\vartheta_n^h \xrightarrow{\mathrm{P}} \vartheta_0$, $\forall\ h = 2, 3, \ldots$.

For deriving a compact formulation of the asymptotic covariance matrix of $n^{1/2}(\vartheta_n^h - \vartheta_0)$, define the number of copula parameters as $q = r - p$ and the matrices

$$\mathcal{T}_1 = \begin{pmatrix} I_p & 0_{pp} & 0_{pq} \\ 0_{qp} & 0_{qp} & I_q \end{pmatrix} \quad \text{and} \quad \mathcal{T}_2 = \begin{pmatrix} I_p & I_p & 0_{pq} \\ 0_{qp} & 0_{qp} & I_q \end{pmatrix}, \tag{2}$$

with p-dimensional identity matrix I_p and $(p \times q)$ null matrix 0_{pq}. Moreover, define $\mathcal{H}^1(\vartheta_0)$ via the relationship

$$n^{-1} \begin{Bmatrix} \ddot{\mathcal{L}}^m(\vartheta_{1,0}, \ldots, \vartheta_{k,0})\ 0_{pq} \\ \ddot{\mathcal{L}}^c_{\mathfrak{v}(\vartheta_{k+1}, \ldots, \vartheta_G), \vartheta}(\vartheta_0) \end{Bmatrix} = \mathcal{H}^1(\vartheta_0) + \mathcal{O}_p(1),$$

where $\mathcal{H}^1(\vartheta_0)$ denotes the (partial) Hessian resulting from *Step 1*. To show the asymptotic distribution of the estimator in dependence of the number of iteration steps h, we make the following additional assumptions:

Assumption 3. *The score* $s(\vartheta_0) = \mathfrak{v}\{\dot{\mathcal{L}}^m(\vartheta_{1,0}, \ldots, \vartheta_{k,0}), \dot{\mathcal{L}}^c(\vartheta_0)\}$ *of the decomposed log-likelihood* $\mathcal{L}(\vartheta) = \mathcal{L}^m(\vartheta_1, \ldots \vartheta_k) + \mathcal{L}^c(\vartheta)$, *with* $\left\{n^{-1} s(\vartheta_0) s(\vartheta_0)^\top\right\} \xrightarrow{\mathrm{P}} \Sigma(\vartheta_0)$, *obeys*

$$n^{-1/2} s(\vartheta_0) \quad \xrightarrow{\mathcal{L}} \quad \mathrm{N}\{0, \Sigma(\vartheta_0)\}. \tag{3}$$

If X is the finite history of a stationary and ergodic stochastic process, Assumption 3 is then satisfied by "Gordin's conditions" as follows: Based on the observation-specific score contributions $s_i(\vartheta_0) \overset{\text{def}}{=} \partial \ell_i(\vartheta)/\partial \vartheta|_{\vartheta = \vartheta_0}$, denote the long-run covariance by $\Sigma(\vartheta_0) = \sum_{i=-\infty}^{\infty} \mathrm{E}\left\{s_i(\vartheta_0) s_i(\vartheta_0)^\top\right\}$. According to Gordin (1969), assuming (i) $\Sigma(\vartheta_0)$ existing and being finite, (ii) $\mathrm{E}\left\{s_i(\vartheta_0) | s_{i-j}(\vartheta_0), s_{i-j-1}(\vartheta_0), \ldots\right\} \xrightarrow{\mathrm{P}} 0$ as $j \to \infty$ and (iii) $\sum_{j=0}^{\infty} \mathrm{E}(\nu_{ij}^\top \nu_{ij})^{1/2}$ being finite with

$$\nu_{ij} = \mathrm{E}\left\{s_i(\vartheta_0) | s_{i-j}(\vartheta_0), s_{i-j-1}(\vartheta_0), \ldots\right\} - \mathrm{E}\left\{s_i(\vartheta_0) | s_{i-j-1}(\vartheta_0), s_{i-j-2}(\vartheta_0), \ldots\right\},$$

is sufficient to guarantee (3).

Assumption 4. *Define the lower block and upper block triangular matrix of* $-n^{-1}\ddot{\mathcal{L}}(\vartheta_0)$ *as* L_n *and* U_n, *respectively, such that* $-n^{-1}\ddot{\mathcal{L}}(\vartheta_0) = L_n - U_n$ *with* $L_{gl,n} = 0$ *for* $g < l \leq G$ *and* $U_{gl,n} = 0$ *for*

$l \leq g \leq G$. *For the probability limits* L *and* U *of* L_n *and* U_n*, respectively, we assume* $\rho(\Gamma) < 1$*, where* $\rho(\cdot)$ *denotes the spectral radius and* $\Gamma \overset{\text{def}}{=} \mathrm{L}^{-1}\,\mathrm{U}$.

Using these assumptions, we can state the following theorem:

Theorem 2. *Let the random variables of the sequence* X *have an identical conditional density* $f_i(\cdot;\vartheta)$ *for which Assumptions 1-4 hold. Then,*

$$\begin{aligned} n^{1/2}(\vartheta_n^h - \vartheta_0) \quad &\xrightarrow{\mathcal{L}} \mathrm{N}\left\{0, \mathcal{B}_h \Sigma(\vartheta_0) \mathcal{B}_h^\top\right\}, \\ \textit{where} \qquad &\mathcal{B}_h = \Gamma^{h-1}\left[\mathcal{K}\mathcal{T}_1 - \{-\mathcal{H}(\vartheta_0)\}^{-1}\mathcal{T}_2\right] + \{-\mathcal{H}(\vartheta_0)\}^{-1}\mathcal{T}_2, \\ \textit{and} \qquad &\mathcal{K} = \left\{-\mathcal{H}^1(\vartheta_0)\right\}^{-1}. \end{aligned}$$

The theorem shows that the asymptotic covariance of ϑ_n^h has a sandwich form consisting of the covariance of the "decomposed" score $s(\vartheta_0)$, $\Sigma(\vartheta_0)$, and matrices $\mathcal{B}_h$. The latter can be computed based on $\Sigma(\vartheta_0)$ exploiting information equality (Assumption 1) and the fact that $\mathcal{J}(\vartheta_0) = \mathcal{T}_2\Sigma(\vartheta_0)\mathcal{T}_2^\top = -\mathcal{H}(\vartheta_0)$. Since $\mathcal{T}_1\Sigma(\vartheta_0)\mathcal{T}_1^\top$ is the expectation of the outer score product obtained from the 2-stage procedure in *Step 1*, the asymptotic covariance matrix of $n^{1/2}(\vartheta_n^1 - \vartheta_0)$ after the first iteration step ($h = 1$) collapses to the well known form

$$\{\mathcal{H}^1(\vartheta_0)^{-1}\}\mathcal{T}_1\Sigma(\vartheta_0)\mathcal{T}_1^\top\{\mathcal{H}^1(\vartheta_0)^{-1}\}^\top. \tag{4}$$

Moreover, an important implication of Theorem 2 is that the estimator is asymptotically efficient if $h \to \infty$. This is due to the fact that, by Assumption 1 and 4, $\lim_{h\to\infty}\mathcal{B}_h = \mathcal{J}(\vartheta_0)^{-1}\mathcal{T}_2$ and thus the asymptotic covariance matrix of $n^{1/2}(\vartheta_n^h - \vartheta_0)$ is $\mathcal{J}(\vartheta_0)^{-1}$:

Corollary 1. *Under the assumptions of Theorem 2,*

$$\lim_{h\to\infty} n^{1/2}(\vartheta_n^h - \vartheta_0) \xrightarrow{\mathcal{L}} \mathrm{N}\left\{0, \mathcal{J}(\vartheta_0)^{-1}\right\}.$$

While Assumptions 1-3 are standard, Assumption 4 is usually not imposed in the context of ML estimation. From a mathematical point of view, Assumption 4 ensures the convergence of Algorithm 1, but it is unclear whether $\rho(\Gamma) < 1$ is guaranteed for arbitrary decompositions of $\vartheta = \mathfrak{v}(\vartheta_1, \ldots, \vartheta_G)$. Using the terminology of Song *et al.* (2005), if U is "larger" than L, then $\rho(\Gamma) \not< 1$ and thus the asymptotic normality of the estimator is not guaranteed anymore. Such a situation, however, is unlikely, as $\mathcal{J}(\vartheta_0)$ can be decomposed as $\mathcal{J}(\vartheta_0) = (\mathrm{D} - \mathrm{U}^\top) - \mathrm{U}$, where $\mathrm{D} = -\operatorname{diag}\{\mathcal{H}_{11}(\vartheta_0), \ldots, \mathcal{H}_{GG}(\vartheta_0)\}$ is a block diagonal matrix. We can neither verify that $\rho(\Gamma) < 1$ generally holds, nor find a theoretical or numerical counter-example. Consider for illustration the trivial case of $r = 2$ and $r_1 = r_2 = 1$. Then, the smallest eigenvalue of

$$\mathrm{L}^{-1}\,\mathrm{U} = \begin{pmatrix} 0 & -\mathrm{L}_{21}\,/\,\mathrm{L}_{11} \\ 0 & \mathrm{L}_{21}^2\,/(\mathrm{L}_{11}\,\mathrm{L}_{22}) \end{pmatrix},$$

is zero and the largest eigenvalue is smaller than one, since the information matrix is positive definite. The case $r = 3$, however, is already more elaborate and the conclusion $\rho(\Gamma) < 1$ cannot be drawn

straightforwardly due to possibly complex eigenvalues of Γ. A stronger condition implying $\rho(\Gamma) < 1$, is given by $\|\Gamma\| < 1$, where $\|\cdot\|$ denotes a matrix norm. Yet, upper bounds constructed from standard inequalities are too rough and it can be generally shown that $\|\Gamma\| \not< 1$.

The condition $\rho(\Gamma) < 1$ is closely related to the dependence of the group-specific estimators $\vartheta_{g,n}^h$, $g = 1, \ldots, G$. Two sub-vectors ϑ_g and ϑ_l are said to be orthogonal for $g \neq l$, if all elements of the corresponding information matrix $\mathcal{J}_{gl}(\vartheta_0)$ are zero, c.f., Lehmann and Casella (1998). According to the structure of L and U, respectively, the blocks of Γ associated with the vectors ϑ_g and ϑ_l, are given by $\Gamma_{gl} = (\mathrm{L}^{-1})_{g\bullet}\,\mathrm{U}_{\bullet l}$, where $(\mathrm{L}^{-1})_{g\bullet}$ refers to rows related to $\vartheta_{g,n}$. If all pairs of $\{\vartheta_{g,n}^h, \vartheta_{l,n}^h\}_{g\neq l}$ are (almost) independent of each other, Assumption 4 will be fulfilled, since U will be close to 0_{rr} and the inverse of L will be mainly driven by the blocks of the main diagonal. Hence, we have a strong conjecture that the condition will most likely be fulfilled if the dependence between the estimates $\vartheta_{g,n}^h$, $g = 1, \ldots, G$, is not too strong. The latter condition typically holds if the number of sub-vectors G (relative to r) does not become too high or (strong) dependence can be ruled out by construction of the appropriate sub-vectors.

2.3. Properties under Misspecification

Consider the case where $\mathrm{P} \notin \mathcal{P}$ and let the true probability measure G be characterized by an absolutely continuous distribution function defined on $\mathbb{R}^d$ with $g(X)$ denoting its measurable Radon-Nikodým density. The observed trajectory X stems from a stochastic process defined on $(\Omega, \mathcal{F}, \mathrm{G})$. For the remainder of this subsection, expectations are taken with respect to G, so that $\mathrm{E}(\cdot) \stackrel{\text{def}}{=} \mathrm{E}_{\mathrm{G}}(\cdot)$. Then, the quasi log-likelihood is given by $\mathcal{Q}(\vartheta) \stackrel{\text{def}}{=} n^{-1}\mathcal{L}(\vartheta; X)$. White (1982) builds on the inference of Akaike (1973) that the maximizer of $\mathcal{Q}(\vartheta)$ estimates the minimizer of the Kullback-Leibler discrepancy between G and P, denoted by $\vartheta_n^\star$, and shows that it converges almost surely to $\vartheta_n^\star$. The latter is sometimes referred to as the pseudo-true parameter. Suppose analogously to Theorem 1 that $\vartheta_n^1 - \vartheta_n^\star = \mathcal{O}_g(1)$, where $\mathcal{O}_g(\cdot)$ refers to the probability measure G. Then, under the assumptions stated in White (1994), convergence in probability of ϑ_n^h, $h > 1$, to $\vartheta_n^\star$ can be established by recursively applying White (1994, Theorem 3.11), which intrinsically establishes the consistency of the 2-stage quasi ML estimator.

To derive the limiting distribution of $n^{1/2}(\vartheta_n^h - \vartheta_n^\star)$, the quantities L and U are re-defined based on $\mathcal{H}(\vartheta_n^\star) = \mathrm{E}\{\ddot{\mathcal{Q}}(\vartheta_n^\star)\}$ whith $\rho(\mathrm{L}^{-1}\,\mathrm{U}) = \rho(\Gamma) < 1$. Moreover, $\mathcal{H}^1(\cdot)$ has to be re-defined as

$$\mathcal{H}^1(\vartheta_n^\star) = \mathrm{E}\left\{ \begin{array}{c} \ddot{\mathcal{Q}}^m(\vartheta_{1,n}^\star, \ldots, \vartheta_{k,n}^\star)\; 0_{pq} \\ \ddot{\mathcal{Q}}^c_{\mathfrak{v}(\vartheta_{k+1},\ldots,\vartheta_G),\vartheta}(\vartheta_n^\star) \end{array} \right\}.$$

Since the structure of the procedure is preserved and only the probability limits are changed,

$$\left[\Gamma^{h-1}\{\mathcal{H}(\vartheta_n^\star)^{-1}\mathcal{T}_2 - \mathcal{H}^1(\vartheta_n^\star)^{-1}\mathcal{T}_1\} - \mathcal{H}(\vartheta_n^\star)^{-1}\mathcal{T}_2\right]^{-1} n^{1/2}(\vartheta_n^h - \vartheta_n^\star) \tag{5}$$

converges in distribution to an r-dimensional normally distributed random variable with covariance matrix $\Sigma(\vartheta_n^\star)$. As in Theorem 2, the asymptotic covariance matrix for *Step 1* is identical to that obtained from a 2-stage procedure. For $h \to \infty$, it collapses to

$$\left\{\mathcal{H}(\vartheta_n^\star)^{-1}\right\} \mathcal{T}_2 \Sigma(\vartheta_n^\star) \mathcal{T}_2^\top \left\{\mathcal{H}(\vartheta_n^\star)^{-1}\right\}^\top,$$

corresponding to the robust covariance matrix as if $n^{1/2}(\vartheta_n^h - \vartheta_n^\star)$ is estimated in one step. Accordingly, the robust covariance matrix leads to valid statistical inference under misspecification if the step-wise increments of the log-likelihood given by Algorithm 1 converge to zero for increasing h. Furthermore, it collapses to the inverse of the information matrix if $f_i(\cdot)$ is correctly specified and information equality holds.

3. Sparse and Efficient Estimation

The iterative estimation approach proposed in the previous section rests on the idea of $\mathcal{L}(\cdot)$ being a complicated function and analytical expressions of $\dot{\mathcal{L}}(\cdot)$ are not available. These properties require derivative-free optimization methods to obtain $\vartheta_{g,n}^h$, $g = 1. \ldots, G$, at *Step h* of the algorithm. However, derivative-free optimization routines do not lead to reliable results for a large number of parameters r_g in group g, $g = 1, \ldots, G$. Therefore, the number of parameters per sub-vector should be small, which, however, leads to a large number of sub-vectors G, and thus increases the computational burden in each iteration step. Moreover, as discussed in Section 2.1, a large number of sub-vectors makes it more difficult to satisfy the condition $\rho(\Gamma) < 1$ since the dependence between the sub-vectors generally rises. In contrast, grouping non-orthogonal parameters in one sub-vector leads to a small G and by construction reduces inter-group dependencies. To address the resulting tradeoff between the reliability of derivative-free optimization procedures (suggesting a high G) and the requirement of keeping the dependence between sub-vectors small (suggesting a low G), we propose combining our estimation algorithm with a suitable penalization procedure reducing the model complexity in the first step and providing sparse (though inefficient) estimates as starting point for the iteration steps. Hence, the idea is to replace *Step 1* of Algorithm 1 by a penalized 2-stage procedure.

3.1. Penalized 2-Stage ML Estimation

Though alternative forms of penalization are possible, we formulate the procedure based on a SCAD penalization of the parameters of $\mathcal{L}^m(\cdot)$ and $\mathcal{L}^c(\cdot)$ according to Fan and Li (2001). They suggest a penalty function which is zero at the origin and whose first derivative is given by

$$p'_{\lambda,a}(|\gamma|) = \lambda \mathbf{I}(|\gamma| \leq \lambda) + \frac{\max(a\lambda - |\gamma|, 0)}{(a-1)} \mathbf{I}(|\gamma| > \lambda) \tag{6}$$

for $a > 2$. Fan and Li (2001) show that this form of penalization function yields unbiased ML estimators which are sparse, i.e., the procedure serves as a thresholding rule setting small estimated coefficients to zero, and are continuous in the data.

The penalized parameters of the marginals and the copula function are collected in $\vartheta_{p_m} \overset{\text{def}}{=} \mathfrak{v}(\vartheta_1, \vartheta_2)$ and $\vartheta_{p_c} \overset{\text{def}}{=} \mathfrak{v}(\vartheta_{G-1}, \vartheta_G)$, respectively. Conversely, $\vartheta_m \overset{\text{def}}{=} \mathfrak{v}(\vartheta_3, \ldots, \vartheta_k)$ and $\vartheta_c \overset{\text{def}}{=} \mathfrak{v}(\vartheta_{k+1}, \ldots, \vartheta_{G-2})$ are non-penalized. Fan and Li (2001) penalize all parameters for the sake of simplicity, but point out that their theoretical results also apply for decomposing parameters into penalized and non-penalized components. Such a separation is necessary in our multi-step estimation context. While existing theory mostly discuss shrinking parameters to zero, we introduce so called penalization targets denoted by $\breve{\vartheta}_1, \breve{\vartheta}_2, \breve{\vartheta}_{G-1}, \breve{\vartheta}_G$. The latter are user-specific and should imply a reduction of model complexity.

This is the case, for instance, (i) if the penalization target of the parameter in a linear model is the corresponding null vector yielding a more parsimonious model or, (ii) if the penalization target of a copula parameter reflects the independence copula yielding a reduction of model complexity. For ease of notation, define the centered SCAD penalty as $\breve{p}_{\lambda,a}(\gamma) = p_{\lambda,a}(|\gamma - \breve{\gamma}|)$, where $\breve{\gamma}$ denotes the penalization target of γ.

In the following analysis, we assume that the independence copula exists as a special case of the considered copula family. Without loss of generality suppose $\vartheta_{1,0} = \breve{\vartheta}_1$ and $\vartheta_{G,0} = \breve{\vartheta}_G$, i.e., the true parameters coincide with the penalization target. The aim is to group as many parameters in ϑ_{p_m} and ϑ_{p_c} as possible, so that shrinking (some of) them implies that $f_i(\cdot;\vartheta_0)$ has a less complicated functional form than $f_i(\cdot;\vartheta)$ with $\vartheta \neq \vartheta_0$. Equivalently, appropriately selected penalization targets imply a centering of the penalty $p_{\lambda,a}(|\cdot|)$ around zero, which finally leads to a simpler functional form, for instance, a more parsimonious regression model. Based on the penalized log-likelihoods

$$\mathcal{L}^{p_m}(\vartheta_1,\ldots,\vartheta_k) = \mathcal{L}^m(\vartheta_1,\ldots,\vartheta_k) - n\sum_{l=1}^{r_1+r_2} \breve{p}_{\lambda_n^m,a^m}(\vartheta_{l,p_m}), \tag{7}$$

$$\mathcal{L}^{p_c}(\vartheta) = \mathcal{L}^c(\vartheta) - n\sum_{l=1}^{r_{G-1}+r_G} \breve{p}_{\lambda_n^c,a^c}(\vartheta_{l,p_c}), \tag{8}$$

we formulate a penalized 2-stage ML estimation procedure as

(1) $\mathfrak{v}(\vartheta^1_{1,n},\ldots,\vartheta^1_{k,n}) = \arg \underset{\mathfrak{v}(\vartheta_1,\ldots,\vartheta_k)}{\text{zero}} \dot{\mathcal{L}}^{p_m}(\vartheta_1,\ldots,\vartheta_k)$

(2) $\mathfrak{v}(\vartheta^1_{k+1,n},\ldots,\vartheta^1_{G,n}) = \arg \underset{\mathfrak{v}(\vartheta_{k+1},\ldots,\vartheta_G)}{\text{zero}} \dot{\mathcal{L}}^{p_c}_{\mathfrak{v}(\vartheta_{k+1},\ldots,\vartheta_G)}(\vartheta^1_{1,n},\ldots,\vartheta^1_{k,n},\vartheta_{k+1},\ldots,\vartheta_G).$

In general, the penalties are permitted to be different for each of the penalized coefficients, but we assume for simplicity one penalty for each of the log-likelihoods. Even though we suggest a data driven choice of the penalty tuning parameters a^m and a^c, we do not index them by the sample size as they are irrelevant for the asymptotic analysis. To formulate the asymptotic properties for the penalized (first-step) ϑ^1_n, define $\mathrm{a}^m_n = \| \mathrm{b}^m_n \|_\infty$ and $\mathrm{a}^c_n = \| \mathrm{b}^c_n \|_\infty$, where $\|\cdot\|_p$ denotes the L_p-norm, with maximum norm for $p = \infty$, and

$$\mathrm{b}^m_n \overset{\text{def}}{=} \left\{ \breve{p}'_{\lambda_n^m,a^m}(\vartheta_{21,0}),\ldots,\breve{p}'_{\lambda_n^m,a^m}(\vartheta_{2r_2,0}) \right\}^\top,$$
$$\mathrm{b}^c_n \overset{\text{def}}{=} \left\{ \breve{p}'_{\lambda_n^c,a^c}(\vartheta_{(G-1)1,0}),\ldots,\breve{p}'_{\lambda_n^c,a^c}(\vartheta_{(G-1)r_{G-1},0}) \right\}^\top.$$

Theorem 3 below gives the consistency of the penalized 2-stage procedure in the first step, ϑ^1_n. It mainly relies on Fan and Li (2001, Lemma 1), whose extension to the modified penalty $\breve{p}_{\lambda,a}(\cdot)$ is trivial and therefore not proved here. However, it additionally requires the penalization target being an interior point of the feasible parameter space. Likewise, while Fan and Li (2001) formulate the proof for i.i.d. data, we apply Lemma 1 in a time series context, as the extension is straightforward due to Assumption 3. Additionally, we impose Assumption 5 bounding the third-order derivative of $\ell_i(\vartheta)$:

Assumption 5. *There exists an open subset θ of Θ containing the true parameter ϑ_0 such that for almost all X_i, $i = 1, \ldots, n$, the density $f_i(\cdot; \vartheta)$ admits all third derivatives $\partial f_i(X_{i1}, \ldots, X_{id}; \vartheta)/\partial\vartheta_u\partial\vartheta_v\partial\vartheta_w$ for all $\vartheta \in \theta$. Furthermore, there exist functions $M_{uvw}(\cdot)$ such that*

$$\left| \frac{\partial \ell_i(\vartheta)}{\partial\vartheta_u\partial\vartheta_v\partial\vartheta_w} \right| \leq M_{uvw}(X_i) \quad \textit{for all} \quad \vartheta \in \theta,$$

where $\mathrm{E}\{M_{uvw}(X_i)\} < \infty$ *for* $u, v, w = 1, \ldots, r$.

Theorem 3. *Let the random variables of the sequence X have an identical conditional density $f_i(\cdot; \vartheta)$ for which Assumptions 1-3 and 5 hold. Let* $\max\{|\breve{p}''_{\lambda_n^m, a^m}(\vartheta_{2l,0})| : \vartheta_{2l,0} \neq \breve{\vartheta}_{2l}\} \to 0$, $l = 1, \ldots, r_2$, *and* $\max\{|\breve{p}''_{\lambda_n^c, a^c}(\vartheta_{(G-1)l,0})| : \vartheta_{(G-1)l,0} \neq \breve{\vartheta}_{(G-1)l}\} \to 0$, $l = 1, \ldots, r_{G-1}$, *be satisfied. If* $\lambda_n^m, \lambda_n^c \to 0$, $n^{1/2}\lambda_n^m \to \infty$ *and* $n^{1/2}\lambda_n^c \to \infty$ *as* $n \to \infty$, *then,*

(a) $\vartheta_{1,n}^1 \xrightarrow{a.s.} \breve{\vartheta}_1$ *and* $\vartheta_{G,n}^1 \xrightarrow{a.s.} \breve{\vartheta}_G$,

(b) $\vartheta_{2,n}^1 + \mathcal{O}(\mathrm{a}_n^m) \xrightarrow{\mathrm{P}} \vartheta_{2,0}$ *and* $\vartheta_{G-1,n}^1 + \mathcal{O}(\mathrm{a}_n^c) \xrightarrow{\mathrm{P}} \vartheta_{G-1,0}$, *with* $\mathrm{a}_n^m, \mathrm{a}_n^c \to 0$ *for* $\lambda_n^m, \lambda_n^c \to 0$ *as* $n \to \infty$,

(c) $\vartheta_{m,n}^1 \xrightarrow{\mathrm{P}} \vartheta_{m,0}$ *and* $\vartheta_{c,n}^1 \xrightarrow{\mathrm{P}} \vartheta_{c,0}$.

Note, however, that the penalization of certain parameters – particularly copula parameters – can also be counterproductive. For example, the penalization of the non-diagonal parameters of a correlation matrix does not ensure its invertibility or a meaningfully chosen penalization target for the parameter of the Gumbel or Clayton copula lies on the boundary of the feasible parameter space, which does not support Theorem 3.

3.2. Iterative Efficient and Sparse Parameter Estimation

According to Theorem 3 *(a)*, the estimators of ϑ_1 and ϑ_G do not need to be updated within the iterative procedure as by assumption their asymptotic limit $\breve{\vartheta}_1$ and $\breve{\vartheta}_G$ imply a simplified form of $\mathcal{L}(\cdot)$ with probability tending to one. Re-estimating the parameters would again lead to a more complex form of $\mathcal{L}(\cdot)$. Consequently, we propose a modification of Algorithm 1 by replacing *Step 1* by the penalized 2-stage ML estimation procedure and ϑ_1 and ϑ_G being replaced by their penalization targets $\breve{\vartheta}_1$ and $\breve{\vartheta}_G$ in the subsequent steps $h > 1$. Hence, the resulting algorithm benefits from a reduced number of parameters to be re-estimated, especially if r_1 and r_G are large:

Algorithm 2.
Step $h = 1$

(1) $\mathfrak{v}(\vartheta_{1,n}^1, \ldots, \vartheta_{k,n}^1) = \arg \underset{\mathfrak{v}(\vartheta_1, \ldots, \vartheta_k)}{\operatorname{zero}} \dot{\mathcal{L}}^{p_m}(\vartheta_1, \ldots, \vartheta_k)$

(2) $\mathfrak{v}(\vartheta_{k+1,n}^1, \ldots, \vartheta_{G,n}^1) = \arg \underset{\mathfrak{v}(\vartheta_{k+1}, \ldots, \vartheta_G)}{\operatorname{zero}} \dot{\mathcal{L}}^{p_c}_{\mathfrak{v}(\vartheta_{k+1}, \ldots, \vartheta_G)}(\vartheta_{1,n}^1, \ldots, \vartheta_{k,n}^1, \vartheta_{k+1}, \ldots, \vartheta_G)$

Step $h > 1$*:*

(1) {*blank step*}

(2) $\quad \vartheta_{2,n}^h = \arg\max_{\vartheta_2} \mathcal{L}(\breve{\vartheta}_1, \vartheta_2, \vartheta_{3,n}^{h-1}, \ldots, \vartheta_{G-1,n}^{h-1}, \breve{\vartheta}_G)$

$\vdots$

$(G-1)$ $\vartheta_{G-1,n}^h = \arg\max_{\vartheta_{G-1}} \mathcal{L}(\breve{\vartheta}_1, \vartheta_{2,n}^h, \ldots, \vartheta_{G-2,n}^h, \vartheta_{G-1}, \breve{\vartheta}_G)$

(G) $\quad \{\textit{blank step}\}$

For the non-shrunken components of ϑ_n^h, define $\tilde{\vartheta} \overset{\text{def}}{=} \mathfrak{v}(\vartheta_2, \ldots, \vartheta_{G-1})$ and $\tilde{q} = \tilde{r} - \tilde{p}$, where $\tilde{r} = \sum_{g=2}^{G-1} r_g$ and $\tilde{p} = \sum_{g=2}^{k} r_g$. The corollary below shows that the consistency of the iterative estimator $\tilde{\vartheta}_n^h$ proved in Theorem 1 also holds in case of Algorithm 2:

Corollary 2. *Under the assumptions of Theorem 3, if* $\lambda_n^m, \lambda_n^c \to 0$, $n^{1/2}\lambda_n^m \to \infty$ *and* $n^{1/2}\lambda_n^c \to \infty$ *as* $n \to \infty$, $\tilde{\vartheta}_n^h \xrightarrow{\text{P}} \tilde{\vartheta}_0 \; \forall\, h = 2, 3, \ldots$.

Based on the consistency of $\tilde{\vartheta}_n^h$, its asymptotic normality can be derived similarly as in Theorem 2. Let $\mathcal{T}_1$, $\mathcal{T}_2$ be as in (2), where p and q are replaced by $\tilde{p}$ and $\tilde{q}$. Define $\mathrm{b}_n \overset{\text{def}}{=} \mathfrak{v}(\mathrm{b}_n^m, 0_s, \mathrm{b}_n^c)$, with $s = \sum_{g=3}^{G-2} r_g$, and let the matrices $\Sigma(\tilde{\vartheta})$, $\mathcal{H}^1(\tilde{\vartheta})$, $\mathcal{H}(\tilde{\vartheta})$, $\ddot{\mathcal{L}}(\tilde{\vartheta})$ and $\mathcal{J}(\tilde{\vartheta})$ depend on $\tilde{\vartheta}$. These matrices are the corresponding sub-matrices of $\Sigma(\vartheta)$, $\mathcal{H}^1(\vartheta)$, $\mathcal{H}(\vartheta)$, $\ddot{\mathcal{L}}(\vartheta)$ and $\mathcal{J}(\vartheta)$. For instance, $\Sigma(\tilde{\vartheta}) = \Sigma(\breve{\vartheta}_1, \vartheta_2, \ldots, \vartheta_{G-1}, \breve{\vartheta}_G)$. To impose Assumption 4 based on the sub-vectors ϑ_g, $g = 2, \ldots, G-1$, we accordingly re-define the limit of $L_n^{-1}U_n \xrightarrow{\text{P}} \mathrm{L}^{-1}\,\mathrm{U} = \widetilde{\Gamma}$, where the lower block triangular matrix L_n and the strict upper block triangular matrix U_n are arranged according to $-n^{-1}\ddot{\mathcal{L}}(\tilde{\vartheta}_0) = L_n - U_n$. Furthermore, since the asymptotic covariance of $\tilde{\vartheta}_n^h$ also involves expressions of the second derivative of the penalty, $\breve{p}''_{\lambda_n,a}(\cdot)$, denote

$$\begin{aligned}
\Psi_n^m &= \mathrm{diag}\left\{\breve{p}''_{\lambda_n^m,a^m}(\vartheta_{21,0}), \ldots, \breve{p}''_{\lambda_n^m,a^m}(\vartheta_{2r_2,0})\right\}, \\
\Psi_n^c &= \mathrm{diag}\left[\breve{p}''_{\lambda_n^c,a^c}\{\vartheta_{(G-1)1,0}\}, \ldots, \breve{p}''_{\lambda_n^c,a^c}\{\vartheta_{(G-1)r_{G-1},0}\}\right].
\end{aligned}$$

Corollary 3. *Under the assumptions of Theorem 2 and Theorem 3, if* $\lambda_n^m, \lambda_n^c \to 0$, $n^{1/2}\lambda_n^m \to \infty$ *and* $n^{1/2}\lambda_n^c \to \infty$ *as* $n \to \infty$, *then,*

$$\begin{aligned}
& n^{1/2}\mathcal{B}_{h,n}^{-1}\left\{(\tilde{\vartheta}_n^h - \tilde{\vartheta}_0) + \widetilde{\Gamma}^{h-1}\mathcal{K}_n\,\mathrm{b}_n\right\} \xrightarrow{\mathcal{L}} \mathrm{N}\left\{0, \Sigma(\tilde{\vartheta}_0)\right\}, \\
& \textit{with} \quad \mathcal{B}_{h,n} = \widetilde{\Gamma}^{h-1}\left[\mathcal{K}_n\mathcal{T}_1 - \{-\mathcal{H}(\tilde{\vartheta}_0)\}^{-1}\mathcal{T}_2\right] + \{-\mathcal{H}(\tilde{\vartheta}_0)\}^{-1}\mathcal{T}_2, \text{[1]} \\
& \qquad\quad \mathcal{K}_n = \left\{\Psi_n - \mathcal{H}^1(\tilde{\vartheta}_0)\right\}^{-1}, \\
& \textit{and} \quad \Psi_n = \mathrm{diag}\left(\Psi_n^m, 0_{ss}, \Psi_n^c\right).
\end{aligned}$$

Hence, compared with Theorem 2, we observe that the first-stage penalization induces two differences: Firstly, the penalization generates a bias $\widetilde{\Gamma}^{h-1}\mathcal{K}_n\,\mathrm{b}_n$ depending on the first and second derivatives of

[1] Since $\mathcal{B}_{h,n}$ is a non-square matrix, $\mathcal{B}_{h,n}^{-1}$ refers to the generalized inverse.

the penalty function and vanishing for $h \to \infty$. Secondly, while in the non-penalization case, $\mathcal{K}_n$ just equals the inverse (partial) Hessian from *Step 1*, it is now adjusted by the diagonal matrix of second derivatives of the penalty. Likewise, after the first iteration step, the asymptotic covariance matrix of $n^{1/2}(\tilde{\vartheta}_n^1 - \tilde{\vartheta}_0)$ is given by

$$\left[\left\{\Psi_n - \mathcal{H}^1(\tilde{\vartheta}_0)\right\}^{-1}\right] \mathcal{T}_1 \Sigma(\tilde{\vartheta}_0) \mathcal{T}_1^\top \left[\left\{\Psi_n - \mathcal{H}^1(\tilde{\vartheta}_0)\right\}^{-1}\right]^\top, \tag{9}$$

which is different from that provided by Fan and Li (2001) since $-\mathcal{H}^1(\tilde{\vartheta}_0) \neq \mathcal{J}(\tilde{\vartheta}_0)$ and $\mathcal{T}_1 \Sigma(\tilde{\vartheta}_0) \mathcal{T}_1^\top \neq \mathcal{J}(\tilde{\vartheta}_0)$. The "sandwich structure" follows from the (inefficient) 2-stage procedure in *Step 1* and can be well approximated by (4), if $\lambda_n^m, \lambda_n^c \to 0$. As in Fan and Li (2001), if $\lambda_n^m, \lambda_n^c \to 0$, $n^{1/2}\lambda_n^m \to \infty$ and $n^{1/2}\lambda_n^c \to \infty$ as $n \to \infty$, the estimator $\tilde{\vartheta}_n^h$ enjoys the oracle property, i.e., $\tilde{\vartheta}_n^h$ performs as well as the corresponding sub-vector $\mathfrak{v}(\vartheta_{2,n}^h, \ldots, \vartheta_{G-1,n}^h)$ in Theorem 2. In other words, the asymptotic properties of $\tilde{\vartheta}_n^h$ are the same as if we knew that $\vartheta_{1,0} = \breve{\vartheta}_1$ and $\vartheta_{G,0} = \breve{\vartheta}_G$, since all elements of b_n and Ψ_n converge to zero if $\lambda_n^m, \lambda_n^c \to 0$, $n^{1/2}\lambda_n^m \to \infty$, and $n^{1/2}\lambda_n^c \to \infty$ as $n \to \infty$.

Finally, if $\rho(\widetilde{\Gamma}) < 1$ and information equality holds, $\lim_{h,n\to\infty} \mathcal{B}_{h,n} = \mathcal{J}(\tilde{\vartheta}_0)^{-1}\mathcal{T}_2$ and therefore,

$$\lim_{h\to\infty} n^{1/2}(\tilde{\vartheta}_n^h - \tilde{\vartheta}_0) \xrightarrow{\mathcal{L}} \mathrm{N}\{0, \mathcal{J}(\tilde{\vartheta}_0)^{-1}\}. \tag{10}$$

Result (10) is a crucial implication of Corollary 3, showing that also the sparse estimator $\tilde{\vartheta}_n^h$ is efficient as $h \to \infty$. Hence, if the iteration-specific increments of the log-likelihood given by Algorithm 2 are sufficiently small for a certain h, the finite sample covariance of $\tilde{\vartheta}_n^h$ can be well estimated by $n^{-1}\mathcal{J}(\tilde{\vartheta}_n^h)^{-1}$ and is independent of the tuning parameters λ_n^m, λ_n^c and a^m, a^c.

4. Iterative Generalized Least Squares Estimation

Besides complex likelihood-based models, Algorithm 1 and 2 can also be advantageous for maximizing simple(r) log-likelihoods, whose parameters $\vartheta_1, \ldots, \vartheta_G$, are non-orthogonal to each other. Consider, for example, a d-dimensional VAR(q) model under the assumption of heteroscedastic and/or autocorrelated errors of the form

$$x_i = c + \sum_{l=1}^{q} A_l x_{i-l} + \varepsilon_i, \tag{11}$$

where $c = (c_1, \ldots, c_d)^\top$ is a vector of constants and A_l is a $(d \times d)$ matrix. To compactly rewrite (11), define $\mathrm{Y} \overset{\text{def}}{=} \mathrm{vec}(x_1, \ldots, x_n)$, $Z_i \overset{\text{def}}{=} (1, x_{i-1}^\top, \ldots, x_{i-q}^\top)^\top$, $\mathrm{Z} \overset{\text{def}}{=} (Z_1, \ldots, Z_n)$ and $\varepsilon \overset{\text{def}}{=} \mathrm{vec}(\varepsilon_1, \ldots, \varepsilon_n)$. Then, (11) can be rewritten as $\mathrm{Y} = (\mathrm{Z}^\top \otimes I_d)\beta + \varepsilon$, where $\beta \overset{\text{def}}{=} \mathrm{vec}(c, A_1, \ldots, A_q)$. In a situation, where ε_i is assumed to be homoscedastic Gaussian with covariance matrix $\Sigma_\varepsilon = \mathrm{E}(\varepsilon_i \varepsilon_i^\top)$, Algorithm 1 and 2 are not beneficial, as β is consistently and efficiently estimated by equation-by-equation OLS and the estimators for β and Σ_ε are independent of each other.

However, as soon as we allow the sequence $\{\varepsilon_i\}_{i=1}^n$ being autocorrelated and/or heteroscedastic, i.e., $\varepsilon \sim \mathrm{N}(0, \Sigma)$, with $\Sigma = \mathrm{E}(\varepsilon\varepsilon^\top) \neq I_n \otimes \Sigma_\varepsilon$, equation-by-equation OLS estimation is not efficient anymore.

In this case, the relevant log-likelihood is given by

$$\mathcal{L}(\beta, \Sigma) \propto -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\left\{\mathrm{Y} - (\mathrm{Z}^\top \otimes I_d)\beta\right\}^\top \Sigma^{-1}\left\{\mathrm{Y} - (\mathrm{Z}^\top \otimes I_d)\beta\right\}, \tag{12}$$

where $\beta = \mathrm{vec}(c, A_1, \ldots, A_p)$ and efficient estimation of β usually requires maximizing (12) with respect to β and $\mathrm{vech}(\Sigma)$ in one step. While this is nearly impossible in practice in case of a non-small d, the iterative FGLS estimator, constructed from applying Algorithm 1 to (12), approaches the Cramér-Rao bound according to Corollary 1.

To illustrate the application of our iterative procedure in such a situation, assign $\mathfrak{v}(\vartheta_1, \vartheta_2) = \beta$ and $\vartheta_3 = \mathrm{vech}(\Sigma)$, and let $\mathfrak{v}(\vartheta^0_{1,n}, \vartheta^0_{2,n})$ denote the (consistent) OLS estimator for β, where $\mathrm{vech}(\cdot)$ denotes half-vectorization of a (symmetric) matrix. As in Algorithm 2, we assume that some VAR parameters are penalized which are without loss of generality collected in $\vartheta_{1,0} = \breve{\vartheta}_1 = 0$. The vector ϑ_3 reflects the imposed (parametric) structure of Σ causing autocorrelation and/or heteroscedasticity. Then, Algorithm 3 yields a sparse estimator for $\vartheta = \mathfrak{v}(\vartheta_1, \vartheta_2, \vartheta_3)$, which is asymptotically efficient as $h \to \infty$:

Algorithm 3.
Step $h = 1$:

(1) $\mathfrak{v}(\vartheta^1_{1,n}, \vartheta^1_{2,n}) = \left[\left\{(\mathrm{Z}\,\mathrm{Z}^\top + n\,\mathrm{B}_{\lambda^m_n, a^m}(\vartheta^0_{1,n}, \vartheta^0_{2,n})\right\}^{-1}\mathrm{Z} \otimes I_d\right]\mathrm{Y}$

(2) $\vartheta^1_{3,n} = \mathrm{vech}\left[\left\{\mathrm{Y} - (\mathrm{Z}^\top \otimes I_d)\mathfrak{v}(\vartheta^1_{1,n}, \vartheta^1_{2,n})\right\}\left\{\mathrm{Y} - (\mathrm{Z}^\top \otimes I_d)\mathfrak{v}(\vartheta^1_{1,n}, \vartheta^1_{2,n})\right\}^\top\right]$

Step $h > 1$:

(1) {*blank step*}

(2) $\vartheta^h_{2,n} = \left\{(\mathrm{Z} \otimes I_d)(\Sigma^{h-1}_n)^{-1}(\mathrm{Z}^\top \otimes I_d)\right\}^{-1}(\mathrm{Z} \otimes I_d)(\Sigma^{h-1}_n)^{-1}\,\mathrm{Y}$

(3) $\vartheta^h_{3,n} = \mathrm{vech}\left[\left\{\mathrm{Y} - (\mathrm{Z}^\top \otimes I_d)\vartheta^h_{2,n}\right\}\left\{\mathrm{Y} - (\mathrm{Z}^\top \otimes I_d)\vartheta^h_{2,n}\right\}^\top\right],$

where

$$\mathrm{B}_{\lambda^m_n, a^m}(\vartheta_1, \vartheta_2) = \mathrm{diag}\left\{p'_{\lambda^m_n, a^m}(|\vartheta_{11}|)/|\vartheta_{11}|, \ldots, p'_{\lambda^m_n, a^m}(|\vartheta_{2r_2}|)/|\vartheta_{2r_2}|\right\},$$

see Fan and Li (2001). The variable selection at *Step 1(1)* rests on the assumption of homoscedastic noise $\Sigma = I_n \otimes \Sigma_\varepsilon$ and corresponds to a ridge regression, which is iteratively computed until the estimator converges. Similar to Algorithm 2, only the regressors of the active set of parameters are kept for the computations in *Step $h > 1$*. Corollary 3 then yields straightforward statistical inference for a fixed h. If a non-sparse estimator is considered, the consequences of a misspecified covariance structure can be inferred from the arguments in Section 2.3.

5. Simulation Study

We illustrate the finite-sample properties of Algorithm 1 and 2 in two simulation studies. The first one applies Algorithm 1 for a $d = 5$-dimensionl VARMA process based on $r = 24$ parameters using $n = 50$ and $n = 150$ observations. Though the number of parameters is comparably small and we do not incorporate any penalization, the model's dimension is high as r/n approaches 0.5. The second study is based on a VMEM process of the dimension $d = 15$ incorporating 375 (partly penalized) parameters and shows the performance of Algorithm 2. All results rely on $w = 500$ Monte Carlo replications.

5.1. VARMA$(1, 1)$

We assume a VARMA(1,1) process with the conditional mean corresponding to the fourth of the data generating processes (DGPs) considered by Kascha (2012), who investigates the quality of the parameter estimates of

$$x_i = A\, x_{i-1} + B\, \varepsilon_{i-1} + \varepsilon_i, \tag{13}$$

for different estimation algorithms. Several elements of A and B are constrained to be zero a priori in order to avoid identification problems, see Assumption 1, Kascha (2012) or Lütkepohl (2006) for details.

While Kascha (2012) assumes $\varepsilon_i \sim \mathrm{N}(0, \Sigma)$, we assume the errors to be t-distributed, i.e., $\varepsilon_{ij} \sim t_{\nu_j}$, which are linked by a Gaussian copula with correlation matrix R. Hence, $\Sigma_{\ell\ell} = \nu_\ell/(\nu_\ell - 2)$, for $\nu_\ell > 2$, and $\Sigma_{k\ell} = \mathrm{R}_{k\ell}\sqrt{\Sigma_{kk}\Sigma_{\ell\ell}}$, $k, \ell = 1, \ldots, d$, where $\Sigma_{k\ell}$ denotes the $k\ell$-th element of the error term covariance Σ. To emphasize the importance of using the complete log-likelihood and to challenge the estimation, we assume a strong dependence structure with

$$\mathrm{R} = \begin{pmatrix} 1.00 & 0.31 & 0.57 & 0.10 & 0.74 \\ 0.31 & 1.00 & 0.53 & 0.51 & 0.78 \\ 0.57 & 0.53 & 1.00 & 0.10 & 0.78 \\ 0.10 & 0.51 & 0.10 & 1.00 & 0.33 \\ 0.74 & 0.78 & 0.78 & 0.33 & 1.00 \end{pmatrix}, \qquad \nu = \begin{pmatrix} 9 \\ 14 \\ 6 \\ 7 \\ 14 \end{pmatrix}.$$

While Kascha (2012) sets the starting values of the optimization procedure to the true parameter values, we choose the start value for the elements of ν as 10, for the elements of R as 0.35, and for the non-zero parameters of A and B as 0. The $r = 24$-dimensional parameter vector ϑ is decomposed into $G = 4$ sub-vectors: $\vartheta_1 = \nu$, $\vartheta_2 = \mathrm{vec}(A_{k\ell})$ and $\vartheta_3 = \mathrm{vec}(B_{k\ell})$, for $A_{k\ell} \neq 0$, $B_{k\ell} \neq 0$ respectively, $k, \ell = 1, \ldots, d$, and $\vartheta_4 = \mathrm{vech}(\mathrm{R})$. Covariance stationarity and invertibility of (13) is ensured with spectral radius $\rho(A) \approx 0.57$ and $\rho(B) \approx 0.78$.

To evaluate the estimation performance after *Step h* relative to that after *Step 1*, we compute the ratio of the corresponding absolute estimation errors (henceforth, relative absolute estimation errors, RAE) as given by

$$\mathrm{RAE}_g^h \overset{\mathrm{def}}{=} \frac{\|\vartheta_{g,0} - \vartheta_{g,n}^h\|_1}{\|\vartheta_{g,0} - \vartheta_{g,n}^1\|_1}.$$

	RAE_1^h		RAE_2^h		RAE_3^h		RAE_4^h	
n	50	150	50	150	50	150	50	150
$h = 2$	0.98 (0.15)	0.88	0.80 (0.16)	0.68	0.82 (0.14)	0.65	0.94 (0.12)	0.94
$h = 4$	0.94 (0.20)	0.76	0.60 (0.26)	0.41	0.62 (0.25)	0.41	0.90 (0.21)	0.93
$h = 6$	0.93 (0.23)	0.73	0.53 (0.28)	0.37	0.53 (0.28)	0.39	0.90 (0.25)	0.95
$h = 10$	0.90 (0.27)	0.72	0.50 (0.29)	0.36	0.50 (0.34)	0.39	0.93 (0.30)	0.95
$h = 15$	0.92 (0.28)	0.71	0.51 (0.31)	0.37	0.51 (0.32)	0.39	0.93 (0.31)	0.96
$h = 20$	0.92 (0.29)	0.72	0.52 (0.31)	0.37	0.51 (0.32)	0.39	0.92 (0.32)	0.98

Table 5.1: Medians of RAE_g^h for the sample sizes $n = 50$ and $n = 150$, with $g = 1, 2, 3, 4$, and for 500 replications. The MAD (in parentheses) is given only for $n = 50$ as the corresponding findings for $n = 150$ are very similar.

Table 5.1 reports the median of the h-specific sampled RAE_g^h together with the corresponding median absolute deviations (MAD, in parantheses). We observe distinct improvements in terms of the RAE for the first steps of the procedure, e.g., from $h = 2$ to $h = 4$ and from $h = 4$ to $h = 6$ respectively. Specifically for $n = 50$, the RAE is larger for $\vartheta_{g,n}^h$, $h \leq 2$, than for all other estimators $\vartheta_{g,n}^h$, $h > 2$, $g = 1, 2, 3, 4$. For higher values of h the performance gains generally become smaller and even become (slightly) negative for some parameters.

Figure 5.1 shows kernel density estimates (KDE) of the RAE_g^h for different values of h. We identify three major effects: (i) The distribution of RAEs generally shifts to the left if h increases. This confirms the statistics shown in Table 5.1 and is true for all sub-vectors. The pattern is most distinct for the parameters of the time series model and less pronounced for the parameters of the distribution model. (ii) The RAE distributions of the sub-vectors of time series parameters, ϑ_2 and ϑ_3, become right-skewed and thus reflect clear performance gains in most cases but also a higher risk of (rare but distinct) deteriorations. This effect is not shown for the distribution and copula parameters, RAE_1^h and RAE_4^h, for which we observe smaller performance gains (on average). This is also confirmed by Table 5.1 reporting a slight deterioration of the quality of the estimates of copula parameters when moving from $h = 6$ to $h = 10$. (iii) Particularly for the copula parameter, the KDE becomes more dispersed for increasing h. Hence, for this parameter, there exists a higher risk to obtain a worse performing estimate over the course of iterations, which is not the case for the other three sub-vectors.

The performance differences between distribution (and copula) parameters and time series parameters are obviously due to the strong correlation between the errors ε_i. These mutual correlations induce a strong dependence between the estimators $\vartheta_{g,n}$, $g = 1, 2, 3$, which is not accounted for in *Step 1* but only if $h > 1$. Consequently, we observe significant improvements in the quality of estimates if h increases. Conversely, the dependence between $\vartheta_{4,n}$ and each $\vartheta_{g,n}$, $g = 1, 2, 3$, is mostly captured directly at *Step 1*. Consequently, for these parameters, additional iteration steps cannot generate strong additional improvements. Overall, the results show a significant superior performance of Algorithm 1 compared to the 2-stage procedure.

The findings above are also supported by corresponding improvements of the log-likelihood as depicted by Figure 5.2. The median of log-likelihood values strongly increases during the first iterations and then stabilizes at the final level. The graph also illustrates that the distribution of log-likelihood values

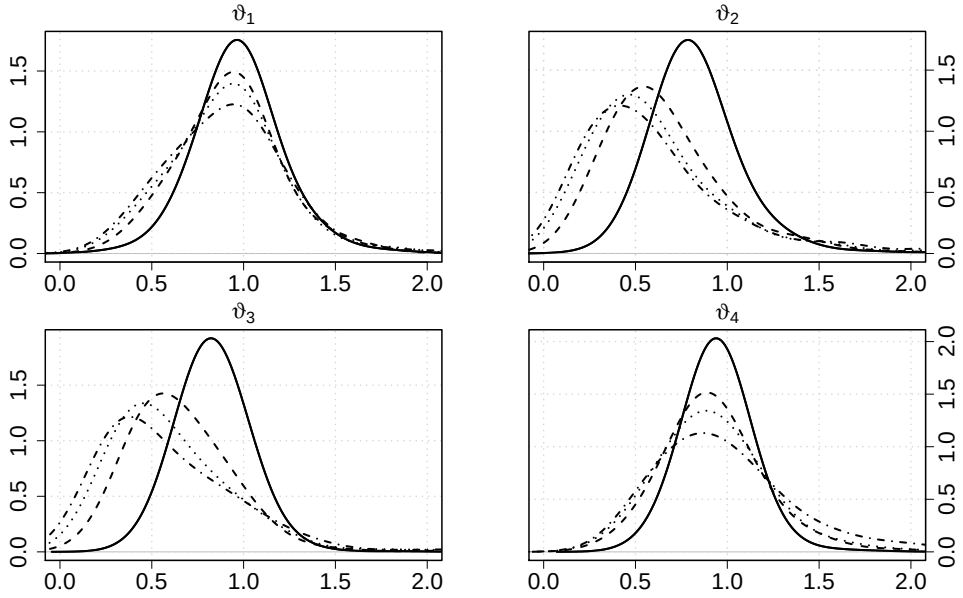

Figure 5.1: Kernel density estimates for RAE_g^h, where $h = 2$ (solid), $h = 4$ (dashed), $h = 6$ (dotted) and $h = 20$ (dashed-dotted). Although the selected bandwidth of 0.15 leads most likely to a slight overfitting of all curves, we keep it constant for all estimates to guarantee a comparable bias.

becomes slightly right-skewed confirming the findings on the RAEs above.

5.2. VMEM$(1, 1)$

In this section, we apply Algorithm 2 to a vector multiplicative error model (VMEM), which is a working horse to describe the dynamics of multivariate positive-valued time series x, such as financial trading volumes, market depth or volatilities. The model has a multiplicative structure given by

$$\begin{aligned} x_i &= \mu_i \odot \varepsilon_i, \\ \mu_i &= \omega + A\, x_{i-1} + B\, \mu_{i-1}, \end{aligned} \tag{14}$$

where $\mu_i \overset{\text{def}}{=} \mathrm{E}(x_i | x_{i-1}, \ldots)$ denotes a d-dimensional vector of conditional means, ε_i is a d-dimensional vector of i.i.d. error terms with $\mathrm{E}(\varepsilon_{ij}) = 1$, $j = 1, \ldots, d$. Moreover, A and B are d-dimensional parameter matrices and "$\odot$" denotes the Hadamard product. For more details on VMEM processes, see, e.g., Hautsch (2012). To challenge our proposed estimation procedure and illustrate its applicability to VMEM processes of higher dimensions, we set $d = 15$. This dimension is significantly higher than typically used in extant studies and thus causes numerical challenges induced by a high number of parameters. To limit the complexity in such a high-dimensional process, we, nevertheless, restrict B being a diagonal matrix $B = \mathrm{diag}(B_{11}, \ldots, B_{dd})$. The errors ε_{ij} are assumed to follow Weibull(γ_j) distributions, whose parameters are randomly chosen from $\mathrm{U}(0.8, 10)$.

Capturing mutual dependencies between the components of ε_i is not straightforward as multivariate

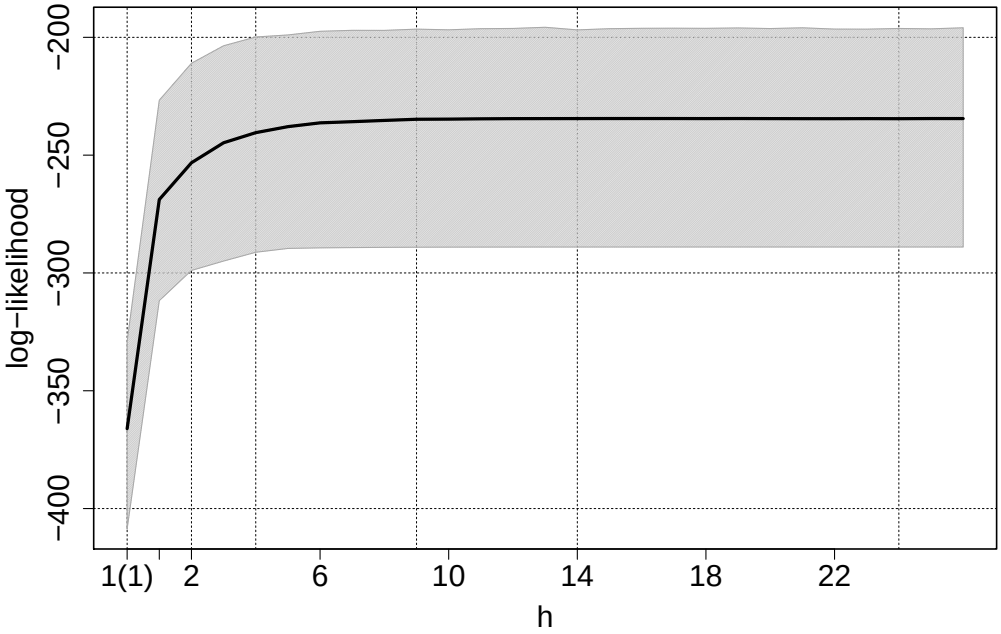

Figure 5.2: Median of log-likelihood values for each iteration step. The gray area covers 95% of the estimates.

extensions of standard distributions for positive-valued random variables do not exist or require strong restrictions. Therefore, extant literature captures the contemporaneous dependence between ε_i by copulas, see, e.g., Bodnar and Hautsch (2012) or Hautsch, Okhrin, and Ristig (2013). Here, we induce dependence between the errors through an R-vine copula, which can generate a broad range of dependence structures including non-linearities, asymmetries and tail dependence. On the other hand, R-vines are not necessarily parsimonious in their representation, as the copula density is split into the product of $d(d-1)/2$ parametric (conditional) bivariate copulae. As the particular choice of the copula is not in the major focus of the present simulation study – the copula parameter is not penalized/decomposed – we refrain from going into more details and refer the reader to Bedford and Cooke (2001), Aas, Czado, Frignessi, and Bakken (2009), Kurowicka and Joe (2011) and Hobæk Haff (2013). To ensure a realistic simulation setup, the R-vine copula is specified based on estimates of empirical distribution functions of financial returns. This allows capturing typical dependencies being present in financial data.

The off-diagonal elements $A_{k\ell}$, $k \neq \ell$, $k, \ell = 1, \ldots, d$, are penalized. Out of the 210 off-diagonal elements of A, we set 180 elements equal to zero and keep 30 elements as non-zero, being randomly chosen from $\mathrm{U}(0.08, 0.2)$. The diagonal elements of A are sampled from $\mathrm{U}(0.05, 0.15)$. The elements of B are chosen such that $\mathrm{E}(x_i) = (I_d - A - B)^{-1}\omega = 1_d$ holds, where $\omega_j = 0.05$, $j = 1, \ldots, d$. The model is covariance stationary as $\rho(A + B) = 0.95$. We construct the low dimensional vectors $\mathfrak{v}(\gamma_j, \omega_j, A_{j\bullet}, B_{jj})$ for the re-estimation with $A_{j\bullet} \neq 0$, $j = 1, \ldots, d$ and $A_{j\bullet}$ refering to the j-th row of A. The parameters shrunken to zero are not re-estimated in the iteration steps h, $h > 1$. Each replication is based on a sample size of $n = 500$ with $r = 375$ parameters to be estimated (including

h	Parameter	RAE^h		SC^h	
1(1)	$A_{k\ell},\ k \neq \ell$	0.35	(0.09)	169	(10.38)
	$A_{k\ell},\ k \neq \ell$	0.34	(0.10)	169	(10.38)
2	$\omega_j, A_{jj}, B_{jj}\ \forall\, j$	0.88	(0.17)		-
	γ	0.60	(0.15)		-
	$A_{k\ell},\ k \neq \ell$	0.32	(0.10)	169	(11.86)
4	$\omega_j, A_{jj}, B_{jj}\ \forall\, j$	0.82	(0.18)		-
	γ	0.46	(0.16)		-
	$A_{k\ell},\ k \neq \ell$	0.31	(0.09)	169	(10.38)
11	$\omega_j, A_{jj}, B_{jj}\ \forall\, j$	0.80	(0.18)		-
	γ	0.43	(0.16)		-

Table 5.2: Median values of RAE^h and SC^h for different parameters. The MAD is given in parentheses. The results are based on 500 replications.

the penalized ones) in total.

To evaluate the performance of the penalization procedure, we employ the RAE statistics introduced above and furthermore check sign consistency by computing

$$\text{SC}^h \stackrel{\text{def}}{=} \sum_{k \neq \ell} \mathbf{I}\left\{\text{sign}(A_{k\ell,0}) = \text{sign}(A^h_{k\ell,n})\right\}.$$

This statistic determines the number of elements of $A_{k\ell}$, $k \neq \ell$ being correctly estimated (un)equal to zero. The results are presented in Table 5.2. Note that the true values of the (non-zero) penalized coefficients are relatively small, making it difficult to discriminate between relevant and non-relevant coefficients. Nevertheless, just around 20% out of the 210 penalized parameters are either estimated unequal zero although they are zero, or estimated zero although they are non-zero. This fraction remains constant in the course of the algorithm. An explanation for this failure rate is the selection of the tuning parameters discussed below. Moreover, RAE^h reveals remarkable improvements of the quality of the estimates – noticeably for the penalized parameters and the parameters of the marginal distributions whose RAE^hs are significantly smaller than 1. Note that we take the maximizer of $\mathcal{L}^m(\vartheta_1, \ldots, \vartheta_k)$ as reference value in the denominator of the RAE^h in order to evaluate the performance of the penalization procedure. Therefore, RAE^h for $A_{k\ell}$, $k \neq \ell$, in Table 5.2 is already smaller than one at *Step 1(1)*. To maximize $\mathcal{L}^{p_m}(\cdot)$ we use the ordinary ML estimator as starting value. Zou and Li (2008) provide a comprehensive overview concerning the maximization of non-concavely penalized log-likelihood functions.

Figure 5.3 descriptively illustrates the convergence of Algorithm 2. The very first range of sample quantiles refers to $\mathcal{L}(\cdot)$ evaluated at the ordinary ML estimator. Consequently, the values of the log-likelihood decline because the values of $\mathcal{L}(\vartheta^1_n)$ must be smaller than the values of $\mathcal{L}(\cdot)$ evaluated at the ordinary ML estimator. Moreover, we observe that the range of sample quantiles is wider for *Step 1* than for each *Step* $h > 1$ and the procedure converges after a few iterations.

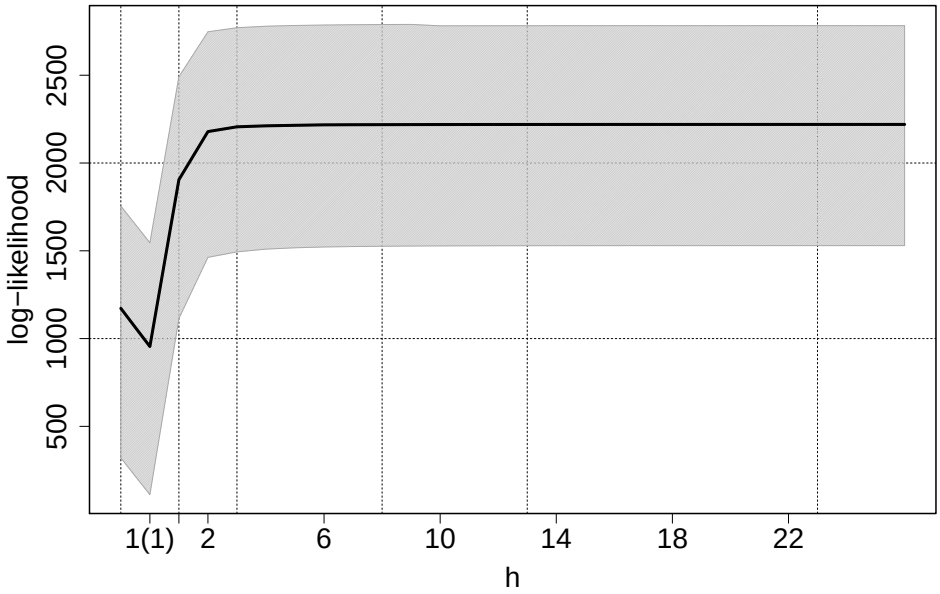

Figure 5.3: Median values of the log-likelihood for each step of the iteration. The gray area includes 95% of the observations.

Remarks on Selecting the Penalization Parameters

The possibly complex functional form of $\mathcal{L}^m(\cdot)$ makes the data driven choice of the tuning parameters of the penalty, λ_n^m and a^m, via cross-validation computationally demanding. Furthermore, in a time series setting, the log-likelihood contributions $\ell_i(\vartheta)$ might be serially correlated. Hence, instead of using classical cross-validation, we split the sample into two parts, S_1 and S_2, containing, for instance, 80% and 20% of the data. Then, we maximize the non-penalized log-likelihood built from S_2 in λ_n^m and a^m, while the estimator $\mathfrak{v}\{\vartheta_{1,n}(\lambda, a), \vartheta_{2,n}(\lambda, a)\}$ defined through *Step 1(1)* of Algorithm 2 is estimated from S_1, see, e.g., Sun (2011). Formally, we follow the data driven choice

$$(\lambda_n^m, a^m)^\top = \arg \max_{(\lambda,a)^\top} \mathcal{L}^m \left\{\vartheta_{1,n}(\lambda, a), \vartheta_{2,n}(\lambda, a), \vartheta_{3,n}, \ldots, \vartheta_{k,n}\right\}, \tag{15}$$

where the non-penalized parameters are fixed at values from the ML estimator of the partial log-likelihood $\mathcal{L}^m(\cdot)$ denoted by $\vartheta_{3,n}, \ldots, \vartheta_{k,n}$. Fitting the tuning parameters of the penalized estimator based on the non-penalized log-likelihood can be motivated by the fact that the asymptotic properties of the penalized estimator hold, if $\lambda_n^m \to 0$ as $n \to \infty$, and that the penalized estimator is the ML estimator for $\lambda_n^m = 0$. Therefore, training the tuning parameters via (15) ensures $\lambda_n^m \to 0$ as $n \to \infty$ and leads to a small λ_n^m for a finite n. Figure 5.4 indicates the distribution of the fitted tuning parameters, confirming our expectation that $a^m > 2$ and $\lambda_n^m \geq 0$ due to the definition of the SCAD penalty. On average, the values of a^m are significantly smaller than the traditional value $a^m = 3.7$ suggested by Fan and Li (2001). For the sake of simplicity, we select just one pair of tuning parameters, which, however, implicitly requires that the log densities $\ell_{ij}^m(\cdot) = \log f_{X_{ij}|\mathcal{F}_{i-1}}(\cdot)$, $j = 1, \ldots, d$, are similar.

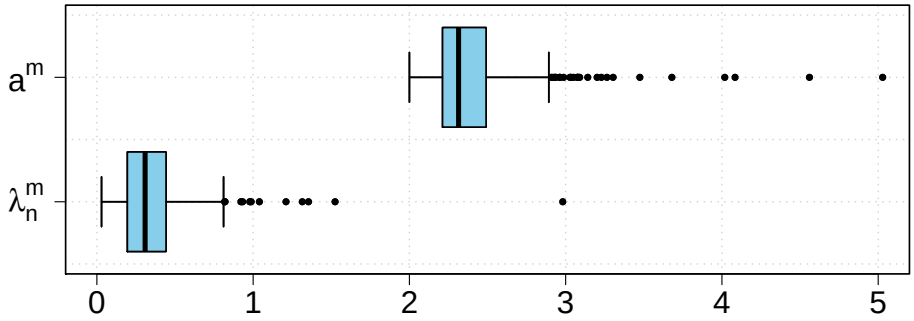

Figure 5.4: Boxplots for the tuning parameters of the penalization function.

For example, if two marginals would be fundamentally different, selecting the same tuning parameters would deteriorate the statistical performance of the procedure. The choice of the tuning parameters becomes even more relevant, if additionally the number of penalized parameters per margin is very large.

The incorrect shrinkage in roughly 20% of the cases as reported above can also be explained by selecting just one pair of tuning parameters λ_n^m, a^m for different marginal distribution functions Weibull(γ_j), $j = 1, \ldots, d$. While $\varepsilon_{i3} \sim$ Weibull(1.41) and $\varepsilon_{i4} \sim$ Weibull(0.82), the other components of ε_i follow Weibull(γ_j) distributions with significantly larger parameters than 1.41 and 0.82, which induce a different shape for the corresponding log densities $\ell_{ij}^m(\cdot)$. Consequently, the selected tuning parameters lead to an inappropriate penalization of the parameters associated with the third and fourth time series leading to a negative bias for the penalized elements of $A_{3\bullet}$ and $A_{4\bullet}$.

6. Measuring Volatility Connectedness

Our empirical study builds on Diebold and Yilmaz (2014) who proposed measuring the connectedness between financial firms based on generalized forecast error variance decompositions (GVDs) stemming from the covariance stationary MA(∞) representation of a linear time series model for daily realized asset price volatilities. Given the importance of such connectedness/spillover measures in recent discussions of systemic risk and financial networks, we illustrate an application of our procedure to an extension of the underlying framework as follows: Firstly, while Diebold and Yilmaz (2014) apply the measure only to a few stocks in order to keep the parameterization of the underlying process tractable, we study 30 U.S. companies making the setup more realistic but also significantly more challenging. Secondly, given that we aim at modeling (realized) volatilities, we specify the underlying time series model not as a VAR process for *log* volatilities but as a VMEM process for plain values thereof. Researchers often model log volatilities instead of the plain series for reasons of tractability and convenience. Indeed, logarithmic transformations ensure positiveness of volatilities by construction and reduce the impact of large outliers. However, when measuring volatility connectedness, it makes a difference whether we measure dependencies in terms of logarithmic or plain series. Therefore, we suggest a parameterization which allows modeling the non-transformed series while ensuring the non-

negativeness of the volatility processes. Thirdly, we allow for deviations from multivariate normality in terms of non-normal marginals (to ensure non-negativeness) which are coupled together with an R-vine copula. The resulting framework is statistically more flexible and realistic but obviously more challenging due to its high parameterization and departures from multivariate normality. However, note that for a dimension of 30, even a Gaussian VAR(3) parameterization as used by Diebold and Yilmaz (2014) cannot be easily estimated by OLS. Depending on the sample size, the need of shrinkage methods is very likely as the number of parameters easily exceeds the number of observations.

Denote the d-dimensional positive-valued time series by x and define the zero-mean martingale difference sequence by $\eta_i \overset{\text{def}}{=} x_i - \mu_i$, $i = 1, \ldots, n$, with $\Sigma_\eta = \mathrm{E}(\eta_i \eta_i^\top)$. Then, under the assumption $\rho(A+B) < 1$, the VMEM$(1,1)$ model in (14) can be rewritten in terms of a MA(∞) parameterization, i.e.,

$$y_i = \eta_i + \sum_{l=1}^{\infty} \left\{ (A+B)^l - (A+B)^{l-1} B \right\} \eta_{i-l} = \eta_i + \sum_{l=1}^{\infty} \Psi_l \eta_{i-l}, \tag{16}$$

with $y_i = x_i - \{I_d - (A+B)\}^{-1} \omega$. While in the given context the insights from interpreting the single elements of the matrix Ψ_l are rather low, the components of the GVDs "summarize" the effect from shocking the ℓ-th element of η_i on the k-th time series. Define the H-step prediction error as $\nu_i(H) = \sum_{l=0}^{H-1} \Psi_l \eta_{i+H-l}$ and conditional on $\eta_{\ell,i+H-l} = \delta$, $l = 0, \ldots, H-1$ as $\nu_{i,\ell}(H) = \sum_{l=0}^{H-1} \Psi_l \{\eta_{i+H-l} - \mathrm{E}(\eta_{i+H-l} | \eta_{\ell,i+H-l} = \delta)\}$. Then, the elements of the GVD are defined as

$$\tilde{v}_{k\ell,H} = \frac{e_k^\top \left[\mathrm{Var}\{\nu_i(H)\} - \mathrm{Var}\{\nu_{i,\ell}(H)\} \right] e_k}{e_k^\top \mathrm{Var}\{\nu_i(H)\} e_k}, \tag{17}$$

where $e_k = (0, \ldots, 0_{k-1}, 1_k, 0_{k+1}, \ldots, 0)^\top$ is a $(d \times 1)$ vector.

Building on the components of the standardized H-step GVD, whose elements are given by $v_{k\ell,H} = \tilde{v}_{k\ell,H} / \sum_{\ell=1}^{d} \tilde{v}_{k\ell,H}$, Diebold and Yilmaz (2014) propose three types of aggregated connected measures: (i) the (net) pairwise directional connectedness' from ℓ to k are defined as $C_{k \leftarrow \ell,H} = v_{k\ell,H}$ and $C_{k\ell,H} = C_{\ell \leftarrow k,H} - C_{k \leftarrow \ell,H}$, respectively. (ii) The total directional connectedness from others to k is given by $C_{k \leftarrow \bullet,H} = \sum_{\ell \neq k} v_{k\ell,H}$, the total directional connectedness to others from ℓ by $C_{\bullet \leftarrow \ell,H} = \sum_{k \neq \ell} v_{k\ell,H}$ and the net total directional connectedness by $C_{\ell,H} = C_{\bullet \leftarrow \ell,H} - C_{\ell \leftarrow \bullet,H}$. (iii) Accordingly, the total connectedness in the system is given by $C_H = \sum_{k \neq \ell} v_{k\ell,H}$.

As a closed form expression for $\tilde{v}_{k\ell,H}$ can only be derived for η_i being Gaussian white noise, the components $\tilde{v}_{k\ell,H}$ are simulated. In particular, the GVD is constructed based on 250 Monte Carlo simulations, where the (conditional centered) moments in (17) are replaced by corresponding sample averages. We conduct this study for $\delta = \sqrt{\Sigma_{\ell\ell,\eta}}$ though the simulation-based estimation of the GVDs also supports alternative specifications of δ. For instance, constructing the measures based on extreme shocks, we might consider $\delta = \kappa_\ell$ with κ_ℓ denoting the fourth standardized moment of $\eta_{\ell,i}$. This can be particularly insightful when copulae are used which incorporate tail dependence in contrast to the Gaussian distribution.

The sequence $\{\Psi_{l,n}\}_{l=1}^{H}$ can be computed from A_n and B_n for arbitrary $H > 0$. Analogously to the simulation study, we assume $\varepsilon_{ij} \sim \text{Weibull}(\gamma_j)$, with $\mathrm{E}(\varepsilon_{ij}) = 1$, but restrict the bivariate (un)conditional

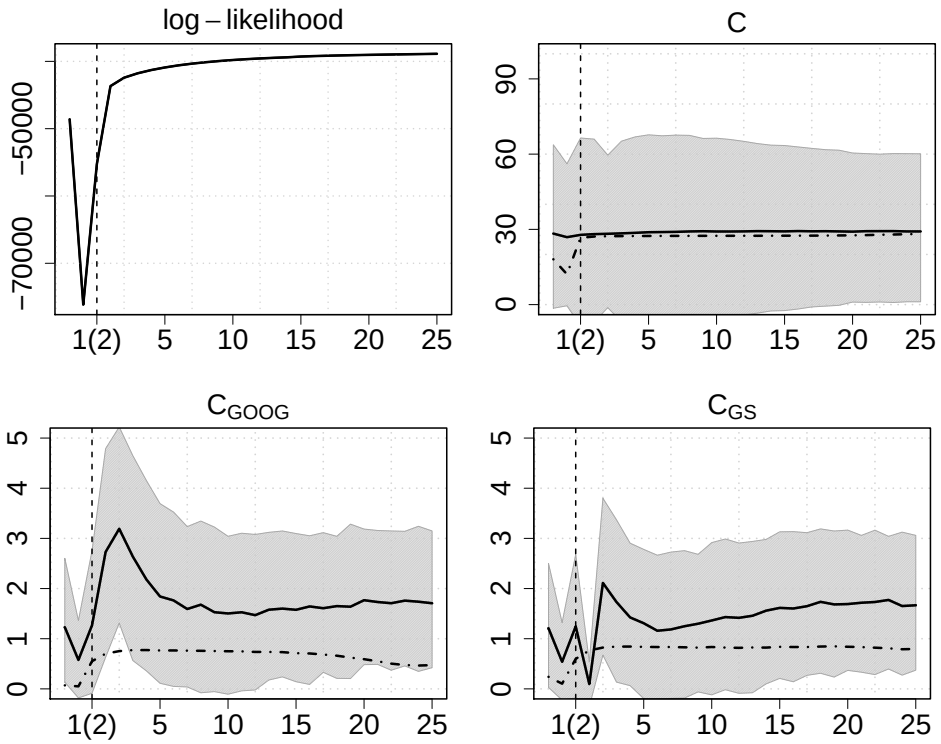

Figure 6.1: Upper panel: log-likelihood values and total systemic connectedness C_{12} in dependence of h. Lower panel: volatility contagion from Google $C_{\bullet \leftarrow \text{GOOG},12}$ and Goldman Sachs $C_{\bullet \leftarrow \text{GS},12}$ in dependence of h. The solid lines refer to the median of the simulated connectedness measures where the gray areas contain 90% of the respective Monte Carlo sample. The dotted-dashed lines refer to the connectedness measures under the assumption of η_i being Gaussian white noise.

copulae of the R-vine to be the t-copulae. Since the resulting specification involves $r = 1860$ parameters to be estimated, an efficient and sparse estimation procedure is expected to induce substantial efficiency gains.

Our analysis employs daily variances estimated using realized kernels as proposed by Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) for 30 companies listed in Table B.1 over the period 01/01/2007 to 31/12/2008. The list contains several constitutes of the Dow Jones Industrial Average and is completed by various large financial institutions. This allows assessing the connectedness between financial companies and important components of the major U.S. industrial stock index.

Figure 6.1 summarizes the main estimation results for a forecast horizon $H = 12$ as in Diebold and Yilmaz (2014).The upper left panel of Figure 6.1 shows the convergence of the log-likelihood function. As in the simulation study, the log-likelihood function sharply decreases as soon as the significant variables are selected and the irrelevant parameters are set to zero. Even though the procedure visually converged, the increments of the log-likelihood at $h = 25$ are still around 30. They become sufficiently tiny for $h \approx 80$, but the insights from the remaining iteration steps are rather low for illustration purposes. The slower rate of convergence compared to the simulation study is mainly caused by the

large-dimensional parameter vector. The penalization shrinks 487 out of the 870 penalized off-diagonal elements of A to zero. Therefore, a couple of sub-vectors are comparably large for using derivative-free optimization techniques which additionally slows down the rate of convergence. Moreover, estimates of the R-vine copula reveal strong dependencies between the components of the random vector ε_i, which also increases the number of steps h until convergence of Algorithm 2. As illustrated by Figure B.1, this cross-sectional dependence is due to strong co-movements of the 30 time series. Song *et al.* (2005) detect similar features of their algorithm in such a setting. In any case, given the separation of the sub-vectors $\mathfrak{v}(\gamma_j, \omega_j, A_{j\bullet}, B_{jj})$, $j = 1, \ldots, d$, they are merely implicitly dependent through the copula and thus Assumption 4 should be fulfilled.

The directional connectedness from Google (GOOG) and Goldman Sachs (GS) to other companies of the sample, $C_{\bullet \leftarrow \text{GOOG},12}$ and $C_{\bullet \leftarrow \text{GS},12}$, are exemplarily investigated. Based on these measures for volatility contagion, we have identified Google and Goldman Sachs as driving factors for volatility in our sample, as the medians of $C_{\bullet \leftarrow \text{GOOG},12}$ and $C_{\bullet \leftarrow \text{GS},12}$ in Figure 6.1 are significantly larger than 1. The lower panel illustrates how the estimates change in dependence of the number of iterations h. As in Section 5.2, the values at $h = 0$ refer to the non-penalized likelihood estimation at which the medians of $C_{\bullet \leftarrow \text{GOOG},12}$ and $C_{\bullet \leftarrow \text{GS},12}$ are already larger than 1. Then, the shrinkage procedure sets several elements of the matrix A to zero and thus the estimated volatility connectedness to the other companies declines as expected. Incoporating the copula into the estimation procedure at *Step 1(2)* increases both measures (to 1.26) since the components of η_{i+H-l}, $l = 0, \ldots, H-1$ are not mutually independent anymore. Both connectedness estimates stabilize after some fluctuations on a level around 1.69, which is supported by widely symmetric sample quantiles.

Comparing our estimates with the corresponding connectedness measures built on the GVD under the (misleading) assumption of Gaussian white noise errors η_i, reveals that our more flexible approach produces on average larger values for $C_{\bullet \leftarrow \text{GOOG},12}$ and $C_{\bullet \leftarrow \text{GS},12}$. Hence, imposing (multivariate) normality yields an under-estimation of volatility connectedness. Nevertheless, also the estimates based on the Gaussian GVD react to the penalization and the incorporation of the copula in a similar (and expected) way.

The total connectedness presented in the upper right panel of Figure 6.1 does not vary with increasing h, but a smaller sample range can be observed. Values of C_{12} close to 30 indicate strong connectedness for the considered network. Likewise, we also find stable results for the connectedness from others as h increases. Overall, it is well illustrated that our iterative algorithm can be used as a valuable workhorse for the estimation of complex high-dimensional time series models. Moreover, it is shown that inefficient 2-stage estimation procedures may yield significantly different estimates resulting in different interpretations of the underlying effects.

7. Conclusion

In this paper, we have proposed an iterative algorithm for maximizing complicated log-likelihood functions and have established the asymptotic properties of the resulting estimator. We have shown that the resulting estimator is asymptotically efficient as the number of iteration steps tends to infinity. As a valuable by-product, we have derived the exact (asymptotic) distribution of the estimator in dependence of the number of iteration steps.

To deal with highly parameterized models, we have combined the procedure with a non-concave penalty reducing model complexity and the curse of dimensionality. While we have focused on multivariate time series models and have illustrated the finite-sample performance of our estimator in a simulation study, the procedure and asymptotic theory can be straightforwardly carried over to several other estimation and inference problems. For example, some are listed in Joe and Xu (1996) and include the multivariate Poisson-Lognormal distribution, multivariate extreme value models and in general copula-based models. Further applications comprise the limited information estimator for the simultaneous probit model and similar models for binary choice variables like the multivariate and the recursive probit model. To illustrate the applicability of the method in realistic but challenging settings, we have estimated volatility connectedness measures constructed from the generalized forecast error variance decomposition of a highly-parameterized non-linear 30-dimensional MA(∞) process of realized volatilities.

A. Mathematical Appendix

This appendix presents the mathematical proofs of the theorems and corollaries given in Section 2 and 3. Expectations are taken with respect to the (true) measure P and defined as $\mathrm{E}(\cdot) \overset{\text{def}}{=} \mathrm{E}_{\vartheta_0}(\cdot) = \mathrm{E}_{\mathrm{P}}(\cdot)$.

Proof of Theorem 1.
Assume $\vartheta^1_{g,n}$ is consistent, so that $\vartheta^1_{g,n} = \vartheta_{g,0} + \mathcal{O}_p(1)$ for $g = 1, \ldots, G$. Note that $\vartheta^2_{1,n}$ satisfies $\dot{\mathcal{L}}_{\vartheta_1}(\vartheta^2_{1,n}, \vartheta^1_{2,n}, \ldots, \vartheta^1_{G,n}) = 0$. Then, by a Taylor expansion of $\dot{\mathcal{L}}_{\vartheta_1}(\cdot)$ around $\vartheta_{1,0}$ and utilizing the mean value theorem it follows that

$$0 = \dot{\mathcal{L}}_{\vartheta_1}(\vartheta_{1,0}, \vartheta^1_{2,n}, \ldots, \vartheta^1_{G,n}) + \left\{ \ddot{\mathcal{L}}_{\vartheta_1\vartheta_1}(\bar{\vartheta}_1, \vartheta^1_{2,n}, \ldots, \vartheta^1_{G,n}) \right\} (\vartheta^2_{1,n} - \vartheta_{1,0}),$$

where $\bar{\vartheta}_1$ lies between $\vartheta^2_{1,n}$ and $\vartheta_{1,0}$. This leads directly to

$$(\vartheta^2_{1,n} - \vartheta_{1,0}) = \left\{ -n^{-1} \ddot{\mathcal{L}}_{\vartheta_1\vartheta_1}(\bar{\vartheta}_1, \vartheta^1_{2,n}, \ldots, \vartheta^1_{G,n}) \right\}^{-1} n^{-1} \dot{\mathcal{L}}_{\vartheta_1}(\vartheta_{1,0}, \vartheta^1_{2,n}, \ldots, \vartheta^1_{G,n}). \tag{18}$$

The first term of the right hand side of (18) converges in probability to a bounded matrix by Assumption 2. Since ϑ^1_n is consistent, we obtain

$$n^{-1} \dot{\mathcal{L}}_{\vartheta_1}(\vartheta_{1,0}, \vartheta^1_{2,n}, \ldots, \vartheta^1_{G,n}) = \lim_{n\to\infty} n^{-1} \mathrm{E}\{\dot{\mathcal{L}}_{\vartheta_1}(\vartheta_0)\} + \mathcal{O}_p(1),$$

as $n \to \infty$ and since the derivatives of all log-likelihood contributions have a mean zero at ϑ_0 by Assumption 1, the second term on the right hand side of (18) converges in probability to zero. Hence, the product of the two random quantities converge in probability to zero by applying Slutsky's theorem and the consistency of $\vartheta^2_{1,n}$ can be deduced. Given the consistency of $\vartheta^2_{1,n}$, the consistency of $\vartheta^2_{g,n}$, $g = 2, \ldots, G$, can be shown in a similar manner. As all sub-vectors $\vartheta^2_{1,n}, \ldots, \vartheta^2_{G,n}$ are consistent, ϑ^2_n is consistent. □

Proof of Theorem 2.
Using a Taylor expansion of $\dot{\mathcal{L}}_{\vartheta_g}(\cdot)$, $g = 1 \ldots, G$, around ϑ_0, the estimator ϑ^h_n satisfies the following equations:

$$0 = \dot{\mathcal{L}}_{\vartheta_g}(\vartheta_0) + \sum_{l \le g} \left\{ \ddot{\mathcal{L}}_{\vartheta_g\vartheta_l}(\vartheta_0) \right\} (\vartheta^h_{l,n} - \vartheta_{l,0}) + \sum_{l > g} \left\{ \ddot{\mathcal{L}}_{\vartheta_g\vartheta_l}(\vartheta_0) \right\} (\vartheta^{h-1}_{l,n} - \vartheta_{l,0}).$$

Rewriting this system of equations in matrix notation leads to

$$n^{1/2}(\vartheta^h_n - \vartheta_0) = L_n^{-1} n^{-1/2} \dot{\mathcal{L}}(\vartheta_0) + L_n^{-1} U_n n^{1/2} (\vartheta^{h-1}_n - \vartheta_0). \tag{19}$$

Note that $L_n^{-1} \xrightarrow{\mathrm{P}} \mathrm{L}^{-1}$, $U_n \xrightarrow{\mathrm{P}} \mathrm{U}$ and L is invertible. Similarly to Song *et al.* (2005), iterating the

recursive system of equations (19) results in

$$
\begin{aligned}
n^{1/2}(\vartheta_n^h - \vartheta_0) &= \left(L_n^{-1}U_n\right)^{h-1} n^{1/2}(\vartheta_n^1 - \vartheta_0) + \sum_{l=0}^{h-2} \left(L_n^{-1}U_n\right)^l L_n^{-1} n^{-1/2} \dot{\mathcal{L}}(\vartheta_0) \\
&= \left(L_n^{-1}U_n\right)^{h-1} n^{1/2}(\vartheta_n^1 - \vartheta_0) + \left[\left\{I_r - \left(L_n^{-1}U_n\right)^{h-1}\right\}\right. \\
&\quad \left.\left(I_r - L_n^{-1}U_n\right)^{-1} L_n^{-1} n^{-1/2} \dot{\mathcal{L}}(\vartheta_0)\right].
\end{aligned} \tag{20}
$$

Following basic matrix algebra, we get $\left(I_r - L_n^{-1}U_n\right)^{-1} L_n^{-1} = \left\{-n^{-1}\ddot{\mathcal{L}}(\vartheta_0)\right\}^{-1}$, so that (20) simplifies to

$$
\begin{aligned}
n^{1/2}(\vartheta_n^h - \vartheta_0) &= \left(L_n^{-1}U_n\right)^{h-1} n^{1/2}(\vartheta_n^1 - \vartheta_0) \\
&\quad + \left[\left\{I_r - \left(L_n^{-1}U_n\right)^{h-1}\right\} \left\{-n^{-1}\ddot{\mathcal{L}}(\vartheta_0)\right\}^{-1} n^{-1/2} \dot{\mathcal{L}}(\vartheta_0)\right].
\end{aligned} \tag{21}
$$

A local approximation of the left hand side of $\mathcal{T}_1 s(\vartheta_n^1) = 0$ around ϑ_0, with $\mathcal{T}_1 s(\vartheta) = \mathfrak{v}\{\dot{\mathcal{L}}^m(\vartheta_1, \ldots, \vartheta_k), \dot{\mathcal{L}}^c_{\vartheta_{k+1}}(\vartheta), \ldots, \dot{\mathcal{L}}^c_{\vartheta_G}(\vartheta)\}$ and $\mathcal{T}_1$ from (2), leads to

$$
0 = \begin{Bmatrix} \dot{\mathcal{L}}^m(\vartheta_{1,0}, \ldots, \vartheta_{k,0}) \\ \dot{\mathcal{L}}^c_{\vartheta_{k+1}}(\vartheta_0) \\ \vdots \\ \dot{\mathcal{L}}^c_{\vartheta_G}(\vartheta_0) \end{Bmatrix} + \underbrace{\begin{Bmatrix} \ddot{\mathcal{L}}^m(\vartheta_{1,0}, \ldots, \vartheta_{k,0}) & 0_{pq} \\ \ddot{\mathcal{L}}^c_{\vartheta_{k+1},\vartheta_1}(\vartheta_0) & \ldots \quad \ddot{\mathcal{L}}^c_{\vartheta_1,\vartheta_G}(\vartheta_0) \\ \vdots & \ddots \quad \vdots \\ \ddot{\mathcal{L}}^c_{\vartheta_G,\vartheta_1}(\vartheta_0) & \ldots \quad \ddot{\mathcal{L}}^c_{\vartheta_G,\vartheta_G}(\vartheta_0) \end{Bmatrix}}_{\stackrel{\text{def}}{=} K} \left(\vartheta_n^1 - \vartheta_0\right).
$$

Based on the matrix $\left(-n^{-1}K\right)^{-1} \xrightarrow{\text{P}} \left\{-\mathcal{H}^1(\vartheta_0)\right\}^{-1}$, a closed form expression for $n^{1/2}(\vartheta_n^1 - \vartheta_0)$ can directly be derived and (21) can be reformulated as

$$
\begin{aligned}
n^{1/2}(\vartheta_n^h - \vartheta_0) &= \left(L_n^{-1}U_n\right)^{h-1} \left(-n^{-1}K\right)^{-1} \mathcal{T}_1 n^{-1/2} s(\vartheta_0) \\
&\quad - \left[\left(L_n^{-1}U_n\right)^{h-1} \left\{-n^{-1}\ddot{\mathcal{L}}(\vartheta_0)\right\}^{-1} - \left\{-n^{-1}\ddot{\mathcal{L}}(\vartheta_0)\right\}^{-1}\right] \mathcal{T}_2 n^{-1/2} s(\vartheta_0),
\end{aligned}
$$

with $\mathcal{T}_2 s(\vartheta_0) = \dot{\mathcal{L}}(\vartheta_0)$ and $\mathcal{T}_2$ from (2). The statement of the theorem follows, as $n \to \infty$, by applying Slutsky's theorem to the right hand side of the latter equation and factorizing the outcome. □

Proof of Theorem 3.
As $\mathcal{L}(\vartheta_{1,n}^1, \ldots, \vartheta_{k,n}^1, \vartheta_{k+1}, \ldots, \vartheta_G)$ is measurable for each $\mathfrak{v}(\vartheta_{k+1}, \ldots, \vartheta_G) \in \Theta_{k+1} \times \ldots \times \Theta_G$ and $\mathfrak{v}(\vartheta_{k+1}, \ldots, \vartheta_G)$ can be chosen to induce the product copula, the existence of the penalized estimator as maximizer of $\mathcal{L}^{p_m}(\cdot)$ from (7) is ensured by Fan and Li (2001, Theorem 1). Therefore, the first statement of part *(a)* follows from Fan and Li (2001, Lemma 1), since the expectation of the third derivative of the log-likelihood contributions are bounded by Assumption 5, $n^{-1/2}\dot{\mathcal{L}}^m(\vartheta_{1,0}, \ldots, \vartheta_{k,0}) = \mathcal{O}_p(1)$ by Assumption 3 and $n^{-1}\ddot{\mathcal{L}}^m(\vartheta_1, \ldots, \vartheta_k) \xrightarrow{\text{P}} \mathcal{H}^m(\vartheta_1, \ldots, \vartheta_k)$ by Assumption 2.
To prove the first statements of *(b)* and *(c)*, treat $\mathcal{L}^m(\vartheta_2, \vartheta_m) = \mathcal{L}^m(\breve{\vartheta}_1, \vartheta_2, \ldots, \vartheta_k)$ and its derivatives

as functions of $\mathfrak{v}(\vartheta_2, \vartheta_m)$ and apply the mean value theorem around $\mathfrak{v}(\vartheta_{2,0}, \vartheta_{m,0})$ to the right hand side of

$$0 = \dot{\mathcal{L}}^m(\vartheta^1_{2,n}, \vartheta^1_{m,n}) - n \left\{ \breve{p}'_{\lambda_n^m, a^m}(\vartheta^1_{21,n}), \ldots, \breve{p}'_{\lambda_n^m, a^m}(\vartheta^1_{2r_2,n}), 0^\top_{s_m} \right\}^\top,$$

with $s_m \stackrel{\text{def}}{=} \sum_{g=3}^k r_g$. Based on $\bar{\vartheta}_m$ lying between $\vartheta^1_{m,n}$ and $\vartheta_{m,0}$ and $\bar{\vartheta}_2$ between $\vartheta^1_{2,n}$ and $\vartheta_{2,0}$, the latter equation can be rewritten as

$$\begin{aligned} \begin{pmatrix} \vartheta^1_{2,n} \\ \vartheta^1_{m,n} \end{pmatrix} - \begin{pmatrix} \vartheta_{2,0} \\ \vartheta_{m,0} \end{pmatrix} = & \left\{ -n^{-1} \ddot{\mathcal{L}}^m(\bar{\vartheta}_2, \bar{\vartheta}_m) + \bar{\Psi}^m_n \right\}^{-1} n^{-1} \dot{\mathcal{L}}^m(\vartheta_{2,0}, \vartheta_{m,0}) \\ & - \left\{ -n^{-1} \ddot{\mathcal{L}}^m(\bar{\vartheta}_2, \bar{\vartheta}_m) + \bar{\Psi}^m_n \right\}^{-1} \begin{pmatrix} \mathrm{b}^m_n \\ 0_{s_m} \end{pmatrix}, \end{aligned} \tag{22}$$

where $\bar{\Psi}^m_n = \operatorname{diag}\{\breve{p}''_{\lambda_n^m, a^m}(\bar{\vartheta}_{21}), \ldots, \breve{p}''_{\lambda_n^m, a^m}(\bar{\vartheta}_{2r_2}), 0^\top_{s_m}\}$ and $\breve{p}''_{\lambda,a}(\cdot)$ denotes the derivative of $\breve{p}'_{\lambda,a}(\cdot)$. The statement follows as $n \to \infty$, since the first term on the right hand side of (22) is $\mathcal{O}_p(1)$ and the second of order $\mathcal{O}(\mathrm{a}_n)$.

The second statement of part *(a)* follows from Fan and Li (2001, Lemma 1) as $\mathfrak{v}(\vartheta^1_{1,n}, \ldots, \vartheta^1_{k,n})$ is consistent for $\lambda^m_n \to 0$ as $n \to \infty$. The second statements of *(b)* and *(c)* follow straightforwardly by similar arguments as for proving the first statements of *(b)* and *(c)*. □

Proof of Corollary 2.
The consistency of $\tilde{\vartheta}^1_{2,n}$ and $\tilde{\vartheta}^1_{G-1,n}$ follows from Theorem 3 for $\lambda^m_n, \lambda^c_n \to 0$ as $n \to \infty$. Thus, $\tilde{\vartheta}^1_n$ can be consistently estimated. Applying Theorem 1 leads to the consistency statement for $\tilde{\vartheta}^h_n$. □

Proof of Corollary 3.
To show the asymptotic normality, a closed form expression for $n^{1/2}(\tilde{\vartheta}^1_n - \tilde{\vartheta}_0)$ has to be derived. For this purpose, treat $\mathcal{L}(\tilde{\vartheta}) = \mathcal{L}(\breve{\vartheta}_1, \vartheta_2, \ldots, \vartheta_{G-1}, \breve{\vartheta}_G)$ and $\mathcal{L}^c(\tilde{\vartheta}) = \mathcal{L}^c(\breve{\vartheta}_1, \vartheta_2, \ldots, \vartheta_{G-1}, \breve{\vartheta}_G)$ as a function of $\tilde{\vartheta}$. Similarly to the proof of Theorem 2, Taylor's expansion around $\tilde{\vartheta}_0$ leads to

$$0 = \begin{Bmatrix} \dot{\mathcal{L}}^m(\breve{\vartheta}_1, \vartheta_{2,0}, \ldots, \vartheta_{k,0}) \\ \dot{\mathcal{L}}^c_{\vartheta_{k+1}}(\tilde{\vartheta}_0) \\ \vdots \\ \dot{\mathcal{L}}^c_{\vartheta_{G-1}}(\tilde{\vartheta}_0) \end{Bmatrix} + \underbrace{\left[\begin{Bmatrix} \ddot{\mathcal{L}}^m(\breve{\vartheta}_1, \vartheta_{2,0} \ldots, \vartheta_{k,0}) & 0_{\tilde{p}\tilde{q}} \\ \ddot{\mathcal{L}}^c_{\vartheta_{k+1}, \vartheta_2}(\tilde{\vartheta}_0) & \ldots \ddot{\mathcal{L}}^c_{\vartheta_2, \vartheta_{G-1}}(\tilde{\vartheta}_0) \\ \vdots & \ddots \vdots \\ \ddot{\mathcal{L}}^c_{\vartheta_{G-1}, \vartheta_2}(\tilde{\vartheta}_0) & \ldots \ddot{\mathcal{L}}^c_{\vartheta_{G-1}, \vartheta_{G-1}}(\tilde{\vartheta}_0) \end{Bmatrix} - \frac{\Psi_n}{n} \right]}_{\stackrel{\text{def}}{=} K} (\tilde{\vartheta}^1_n - \tilde{\vartheta}_0) - \frac{\mathrm{b}_n}{n},$$

with $\left(-n^{-1}K\right)^{-1} \xrightarrow{\mathrm{P}} \left\{ \Psi_n - \mathcal{H}^1(\tilde{\vartheta}_0) \right\}^{-1}$. Replacing $n^{1/2}(\tilde{\vartheta}^1_n - \tilde{\vartheta}_0)$ in the corresponding expression of

(21) results in

$$
\begin{aligned}
n^{1/2}\left\{(\tilde{\vartheta}_n^h - \tilde{\vartheta}_0) \;+\; \left(L_n^{-1}U_n\right)^{h-1}\left(-n^{-1}K\right)^{-1}\mathrm{b}_n\right\} \qquad (23)\\
= \left[\left(L_n^{-1}U_n\right)^{h-1}\left[\left(-n^{-1}K\right)^{-1}\mathcal{T}_1 - \{-n^{-1}\ddot{\mathcal{L}}(\tilde{\vartheta}_0)\}^{-1}\mathcal{T}_2\right]\right.\\
\left.+\{-n^{-1}\ddot{\mathcal{L}}(\tilde{\vartheta}_0)\}^{-1}\mathcal{T}_2\right] n^{-1/2}s(\tilde{\vartheta}_0),
\end{aligned}
$$

with $s(\tilde{\vartheta}_0) = \mathfrak{v}\{\dot{\mathcal{L}}^m(\breve{\vartheta}_1, \vartheta_{2,0}, \ldots, \vartheta_{k,0}), \dot{\mathcal{L}}^c_{\vartheta_2}(\tilde{\vartheta}_0), \ldots, \dot{\mathcal{L}}^c_{\vartheta_{G-1}}(\tilde{\vartheta}_0)\}$. Given that $n^{-1/2}s(\tilde{\vartheta}_0) \xrightarrow{\mathcal{L}} \mathrm{N}\{0, \Sigma(\tilde{\vartheta}_0)\}$, applying Slutsky's theorem to the product on the right hand side of (23), as $n \to \infty$, completes the proof. □

B. List of Companies

This appendix presents a list of the 30 companies used in Section 6.

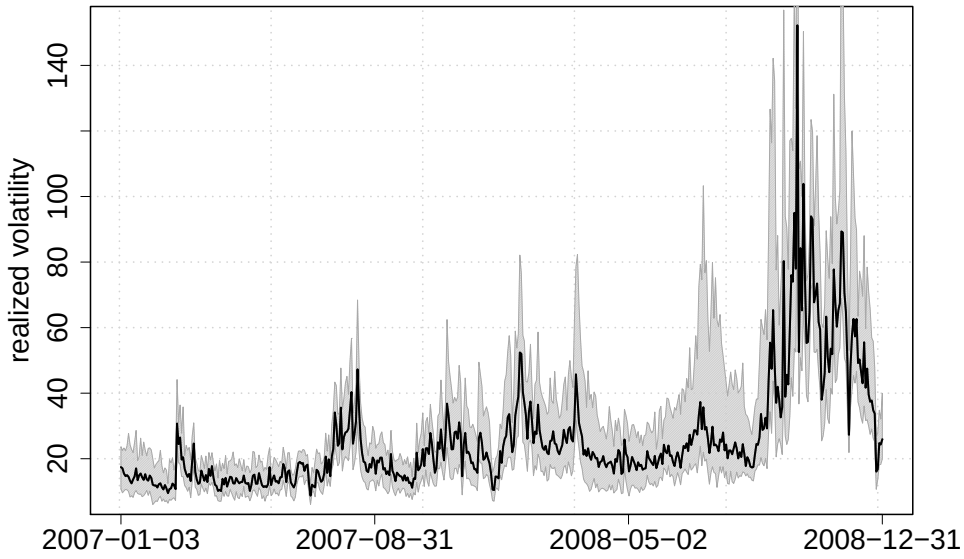

Figure B.1: Median of the realized volatilities over the companies presented in Table B.1. The gray area includes 0.90% of the observations.

Company	Ticker	Sector
3M	MMM	Conglomerate
AT&T	T	Telecommunications
ConocoPhillips	COP	Oil & Gas
Caterpillar	CAT	Heavy Equipment
Chevron	CVX	Oil & Gas
Cisco Systems	CSCO	Networking Equipment
Coca-Cola	KO	Beverage
DuPont	DD	Chemicals
ExxonMobil	XOM	Oil & Gas
General Electric	GE	Conglomerate
Goldman Sachs	GS	Banking
Google	GOOG	IT
Hewlett-Packard	HPQ	IT
Home Depot	HD	Retailing
IBM	IBM	IT
Intel	INTC	IT
Johnson & Johnson	JNJ	Medical Equipment
JPMorgan Chase	JPM	Banking
Kraft Foods	KFT	Food Processing
McDonald's	MCD	Restaurants
Merck & Co	MRK	Pharmaceuticals
MetLife	MET	Financial Services
Microsoft	MSFT	IT
Pfizer	PFE	Pharmaceutical
Procter & Gamble	PG	Consumer Goods
United Technologies	UTX	Conglomerate
U.S. Bancorp	USB	Banking
Walmart	WMT	Retail
Walt Disney	DIS	Mass Media
Wells Fargo	WFC	Banking

Table B.1: Basic information of the companies used in Section 6.

References

Aas K, Czado C, Frignessi A, Bakken H (2009). "Pair-Copula Constructions of Multiple Dependence." *Insurance, Mathematics and Economics*, **8**(2), 182–198.

Akaike H (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In

BN Petrov, F Csáki (eds.), *Second International Symposium on Information Theory*, pp. 267–281. Akadémiai Kiado.

Barndorff-Nielsen O, Hansen P, Lunde A, Shephard N (2008). "Designing Realized Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise." *Econometrica*, **76**, 1481–1536.

Bedford T, Cooke RM (2001). "Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines." *Annals of Mathematical and Artificial Intelligence*, **32**, 245–268.

Bodnar T, Hautsch N (2012). "Modeling Time-Varying Covariances of Trading Processes: Copula-Based Dynamic Conditional Correlation Multiplicative Error Processes." *SFB 649 Discussion Paper 2012-44*, Sonderforschungsbereich 649, Humboldt-Universität zu Berlin, Germany.

Diebold FX, Yilmaz K (2014). "On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms." *Journal of Econometrics, forthcoming*.

Fan J, Li R (2001). "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association*, **96**(456), 1348–1360.

Gordin MI (1969). "The Central Limit Theorem for Stationary Processes." *Soviet Math. Dokl.*, **10**, 1174–1176.

Hautsch N (2012). *Econometrics of Financial High-Frequency Data*. 1 edition. Springer, Berlin.

Hautsch N, Okhrin O, Ristig A (2013). "Modeling Time-Varying Dependencies between Positive-Valued High-Frequency Time Series." In P Jaworski, F Durante, WK Härdle (eds.), *Copulae in Mathematical and Quantitative Finance*. Springer.

Hobæk Haff I (2013). "Parameter Estimation for Pair-Copula Constructions." *Bernoulli*, **19**(2), 462–491.

Jaworski P, Durante F, Härdle WK (2013). *Copulae in Mathematical and Quantitative Finance*, volume 213 of *Lecture Notes in Statistics*. Springer.

Joe H (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

Joe H, Xu JJ (1996). "The Estimation Method of Inference Functions for Margins for Multivariate Models." *Technical Report 166*, Department of Statistics, University of British Columbia.

Kascha CJ (2012). "A Comparison of Estimation Methods for Vector Autoregressive Moving-Average Models." *Econometric Reviews*, **31**(3), 297–324.

Kurowicka D, Joe H (2011). *Dependence Modeling: Vine Copula Handbook*. World Scientific Publishing Company, Incorporated.

Lehmann E, Casella G (1998). *Theory of point estimation*. 2 edition. Springer.

Lütkepohl H (2006). *New Introduction to Multiple Time Series Analysis*. Springer.

Nelsen RB (2006). *An Introduction to Copulas*. Springer, New York.

Sklar A (1959). "Fonctions de Répartition à n Dimension et Leurs Marges." *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 299–231.

Song PX, Fan Y, Kalbfleisch JD (2005). "Maximization by Parts in Likelihood Inference." *Journal of the American Statistical Association*, **100**, 1145–1158.

Sun Y (2011). "Regularization for High Dimensional Time Series Models." PhD-thesis, University of Cincinnati.

Tibshirani R (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. 267–288.

Tibshirani R (2011). "Regression Shrinkage and Selection via the Lasso: a Retrospective." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(3), 273–282.

White H (1982). "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, **50**, 1–25.

White H (1994). *Estimation, Inference and Specification Analysis*. 1 edition. Cambridge University Press, Cambridge.

Zou H, Li R (2008). "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models." *The Annals of Statistics*, **36**(4), 1509–1533.

CENTER FOR FINANCIAL STUDIES

CFS WORKING PAPER SERIES

No.	Author(s)	Title
449	**Engelbert J. Dockner** **Manuel Mayer** **Josef Zechner**	Sovereign Bond Risk Premiums
448	**Johann Reindl** **Neal Stoughton** **Josef Zechner**	Market Implied Costs of Bankruptcy
447	**Jaewon Choi** **Dirk Hackbarth** **Josef Zechner**	Granularity of Corporate Debt
446	**Giuseppe Bertola** **Winfried Koeniger**	Hidden Insurance in a Moral Hazard Economy
445	**Rong Hai** **Dirk Krueger** **Andrew Postlewaite**	On the Welfare Cost of Consumption Fluctuations in the Presence of Memorable Goods
444	**Jill E. Fisch** **Tess Wilkinson-Ryan**	Why Do Retail Investors Make Costly Mistakes? An Experiment on Mutual Fund Choice
443	**Christiane Baumeister** **Pierre Guérin** **Lutz Kilian**	Do High-Frequency Financial Data Help Forecast Oil Prices? The MIDAS Touch at Work
442	**Brigitte Haar**	Investor protection through model case procedures – implementing collective goals and individual rights under the 2012 Amendment of the German Capital Markets Model Case Act (KapMuG)