Journal of Machine Learning Research 27 (2026) 1-50 Submitted 03/23; Revised 9/24; Published 1/26

Online Detection of Changes in Moment–Based Projections: When to Retrain Deep Learners or Update Portfolios?

Ansgar Steland STELAND@STOCHASTIK.RWTH-AACHEN.DE
Institute of Statistics and AI Center
RWTH Aachen University
52062 Aachen, Germany

Editor: Aurélion Garivier

Abstract

Training deep learning neural networks often requires massive amounts of computational ressources. We propose to sequentially monitor network predictions to trigger retraining only if the predictions are no longer valid. This can reduce drastically computational costs and opens a door to green deep learning. Our approach is based on the relationship to projected second moments monitoring, a problem also arising in other areas such as computational finance. Various open–end as well as closed–end monitoring rules are studied under mild assumptions on the training sample and the observations of the monitoring period. The results allow for high–dimensional non-stationary time series data and thus, especially, non–i.i.d. training data. Asymptotics is based on Gaussian approximations of projected partial sums allowing for an estimated projection vector. Estimation of projection vectors is studied both for classical non–ℓ_0–sparsity as well as under sparsity. For the case that the optimal projection depends on the unknown covariance matrix, hard– and soft–thresholded estimators are studied. The method is analyzed by simulations and supported by synthetic data experiments.

Keywords: Change–point, Deep learning, Gaussian approximation, Sparsity, Time Series.

1. Introduction

Despite the success of fast algorithms such as ADAM, training of deep learners requires massive amounts of computational resources, and when the data distribution changes, called *concept drift*, the network predictions may be outdated and render the network useless. This asks for a monitoring approach to trigger retraining, if (and only if) the predictions are no longer valid. A crucial observation is that neural networks with linear output layer compute inner products between the features derived by the hidden layers and the connection weights of that output layer yielding the prediction for some input $\boldsymbol{x}$. If the distribution of the input samples changes and thus differs from the distribution in the learning sample used to train the net, these predictions will be outdated and may lead to biased net outputs. This suggests the following approach: When using the deep learner to compute predictions for input data, a monitoring algorithm is used to detect

when the predictions are no longer compatible with the training sample thus indicating the need to retrain the net. Our solution monitors efficiently in an online fashion a second moment functionals associated to the predictions, which evaluates the features derived by the deep learner, and signals the need to retrain the network.

Such second moment functionals, even in a much simpler form, also arise in computational finance when determining optimal portfolios. Thus, before proceeding with our primary deep learning application, let us briefly discuss this example from finance: The risk of a portfolio $\boldsymbol{v}$ of d risky assets with log returns $\boldsymbol{Y}$ is usually measured by the variance $\boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}$ of the portfolio return $\boldsymbol{v}^\top \boldsymbol{Y}$, where $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{Y})$ is the $d \times d$ covariance matrix. It is well known that the Markowitz–optimal portfolio explicitly depends on the structure of the first two moments and requires to compute the inverse $\boldsymbol{\Sigma}^{-1}$. It should be updated if and only if its risk has changed, since for a large investment universe such an update leads to substantial computational costs and, even more important, substantial trading costs.

The problem to examine neural network predictions breaks down to a similar quadratic form. Anticipating the main message behing the detailed discussion in Section 5, the basic idea is as follows: Recall that deep neural networks with H hidden layers and linear output process inputs $\boldsymbol{X}$ and linearly combine derived features $\boldsymbol{Z} = f_H(\boldsymbol{X})$ by some network function f_H to determine the net output $\boldsymbol{\beta}^\top \boldsymbol{Z}$ where $\boldsymbol{\beta}$ denotes the vector of output weights. This applies regardless of the network topology. Here, for simplicity of presentation, we consider a single output neuron. Since the network prediction for a new input $\boldsymbol{X}^*$ is given by the projection $\hat{Y}^* = \hat{\boldsymbol{\beta}}_m^\top f_H(\boldsymbol{X}^*; \hat{\boldsymbol{\theta}}_m)$, where $(\hat{\boldsymbol{\theta}}_m, \hat{\boldsymbol{\beta}}_m)$ are the trained network parameters regarded as estimators of the true but unknown optimal weights $(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)$, it is clear that the net needs retraining using an updated training sample, only if the change is relevant for that projection. This can be assessed by monitoring the second moment $\mathbb{E}(\boldsymbol{\beta}^\top f_H(\boldsymbol{X}^*))^2 = \boldsymbol{\beta}^\top \mathbf{M}^H \boldsymbol{\beta}$, which is a quadratic form with coefficient matrix $\mathbf{M}_n^H = \mathbb{E}(f_H(\boldsymbol{X}^*; \boldsymbol{\theta}_0) f_H(\boldsymbol{X}^*; \boldsymbol{\theta}_0)^\top)$ given by second moment matrix of the last hidden layer, which represents the features learned by the net. The practical algorithms discussed later will be based on associated quantities where unknowns are estimated from the training sample and compared to current data.

We approach this machine learning problem from a sequential statistical perspective, which allows to draw on what is known on sequential detection. It is assumed that after having trained the deep learner, the trained net is used to generate predictions of a sequential input data stream. We use these predictions to monitor the stream, in order to derive a signal when the current data are no longer compatible with the training sample indicating that the data distribution has changed and the neural network needs to be retrained. Our solution employs the open-end monitoring framework usually addressed to Chu et al. (1996), which incorporates such a training sample of size m satisfying the no–change null hypothesis of interest and allows for open–end monitoring. Open–end means that the stream may be monitored inifinitely long. Since often neural networks are retrained on a coarse regular grid, we also consider the closed–end setting where monitoring stops latest at a fixed time horizon. The training sample is used to estimate unknowns and to compare current data of the monitoring period with stable training data. We study several detectors based on a cumulated sum (CUSUM) statistic

following Horváth et al. (2004) to monitor regression residuals and elaborated in Chen et al. (2010) to monitor squared residuals for a variance change of regression errors, as well as detectors which draw on classical results on boundary-crossing probabilities for Brownian motion, specifically Erdös and Kac (1946) and Robbins and Siegmund (1970). A common basic feature is that the data from the monitoring period is successively cumulated, such that information increases as monitoring proceeds, and the unknown no–change second moment functional is replaced by a suitable estimator calculated from the training sample. The associated sequential stream of detector statistics is then compared with a boundary function to decide at each time instant whether or not a signal should be raised and the no-change null hypothesis be rejected.

This paper contributes by establishing asymptotic theory for high–dimensional nonlinear and nonstationary time series of growing dimension, and we particularly address the case that the projection vector is estimated from the training sample. The results are based on strong Gaussian approximations of partial sums which extend previous results of Mies and Steland (2023) and Steland (2020). We show that any estimator with convergence rate $O_{\mathbb{P}}(\sqrt{\frac{r_d}{m}})$, $r_d \geq 0$, can be used, provided the dimension d ensures $\sqrt{dr_d} = O(m^{\zeta})$ for $0 < \zeta < \frac{1}{2}$, so that for $\sqrt{m}$–consistent estimators the dimension d can be almost as large as $\sqrt{m}$. For highdimensional settings where d is large compared to m or even much larger, it is common to impose sparseness assumptions and employ sparse estimators. Examples include estimation of eigenvectors. If $\boldsymbol{v}$ is s–sparse and the estimator $\hat{\boldsymbol{v}}_m$ is concentrated on the support of $\boldsymbol{v}$ with sufficiently large probability for large m, then the condition can be relaxed to $\sqrt{sr_d} = O(m^{\zeta})$. Under sparsity, the dimension an grow almost exponentially, see our discussion in Section 3.1.

For the above two examples, Markowitz portfolios and deep learning networks, the optimal projection vector is a function of the unknown covariance resp. precision matrix. Thus, we study (soft–) thresholding estimates and contribute consistency guarantees under the general assumptions imposed in this paper, thus generalizing the results of Bickel and Levina (2008). Especially, we contribute a non–probabilistic bound on the maximum norm of the difference between a thresholded matrix and its estimation target, which may be of independent interest.

We elaborate on both applications discussed above, with a focus on neural net monitoring. Here, as a side–product of independent interest, we contribute an improved result on the Lipschitz constant of a (deep) neural network wth respect to the network parameters.

The organization of the rest of the paper is as follows. Section 2 describes the framework and the proposed detection methods. Asymptotic results justifying the procedure and providing theoretical insights into its properties are provided in Section 3. Section 4 studies (soft–) thresholding estimators of projections depending on the covariance matrix and the mean vector suitable for high–dimensional data. Applications to deep learning and financial portfolio optimization are discussed in Section 5, and the approach is illustrated by monitoring synthetic data and investigated by simulations. Proofs of the main results are provided in Section 6.

2. Monitoring Framework

In view of the fact that machine learning approaches for prediction are typically applied whenever new input data arrives, we consider an idealized setting where data vectors are sequentially observed, one after the other, and at each time point a $\{0,1\}$–valued detector is applied where 1 stands for a signal indicating a change and the need for a reaction, such as retraining the deep learner, whereas 0 means that we keep using the given neural network specification. We observe random vectors

$$\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m, \boldsymbol{Y}_{m+1}, \cdots$$

of dimension $d = d_m$, $m \in \mathbb{N}$, defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ represents a training sample of size m, and starting at time $m+1$ the observations $\boldsymbol{Y}_{m+i}$, $i \geq 1$, arrive one after the other. Going beyond the classical assumption of independent observations, we consider below a general vector time series model allowing for serial correlations as arising in real–world applications, see below and Section 3.1 for details.

Our goal is to sequentially decide at each time $n \geq 1$ whether or not the second moment structure of the observations $\boldsymbol{Y}_{m+1}, \ldots, \boldsymbol{Y}_{m+n}$ has changed in comparison to the training sample. We reduce dimensionality by projecting the data onto a vector $\boldsymbol{v} \in \mathbb{R}^d$. Precisely, the detector proposed in this paper monitors the functional

$$\theta(\mathbb{P}_{\boldsymbol{Y}_i}) = \boldsymbol{v}^\top [\mathbb{E}(\boldsymbol{Y}_i \boldsymbol{Y}_i^\top)] \boldsymbol{v}, \qquad i \geq 1,$$

where $\mathbb{P}_{\boldsymbol{Y}}$ denotes the law of a random vector $\boldsymbol{Y}$. The parameter $\theta(\mathbb{P}_{\boldsymbol{Y}})$ reacts to a change in the mean as well as to a change in the covariance matrix, if these changes are visible through the bilinear form induced by $\boldsymbol{v}$ and do not cancel. There may be problems where methods are preferable which are able to explicitly decouple changes in the mean and the variances; here we refer to methods proposed by Steland and Rafajłowicz (2014). If $\boldsymbol{v}$ is unknown, we replace it by an estimator calculated from the training sample as discussed in greater detail below, where we consider the classical non–ℓ_0–sparse setting as well as ℓ_0–sparse vectors and associated sparse estimators which typically yield much better convergence rates.

The sequential testing problem of interest is formalized as a at–most–one–change (AMOC), since our goal is to detect a change quickly to initiate measurements instead of finding and dating all change-points possibly present in the data. We test the no-change null hypothesis

$$H_0 : \boldsymbol{v}^\top \mathbf{M}_i \boldsymbol{v} = \theta_0 \text{ for all } i \geq 1$$

which asserts that the second moments – viewed through the projection vector $\boldsymbol{v}$ – is constant and, especially, coincides with the second moments of the training sample, against the AMOC alternative hypothesis,

$$\begin{aligned} H_A : &\text{ there is } k^* \geq m\delta \text{ with } \boldsymbol{v}^\top \mathbf{M}_i \boldsymbol{v} = \theta_0,\ 1 \leq i < m + k^*, \\ &\text{and } \boldsymbol{v}^\top \mathbf{M}_i \boldsymbol{v} = \theta_A,\ i \geq m + k^* \ , \end{aligned}$$

for some $\theta_A \neq \theta_0$. Under H_A there is a change at time k^* in terms of the second moments. In our treatment, θ_0 and θ_A may be unknown positive constants, namely the pre- and after–change values of the functional $\boldsymbol{v}^\top \mathbf{M}_i \boldsymbol{v}$, and the change–point k^* is assumed unknown and non–random. More general functional alternative hypotheses are considered in the next section. If the underlying vector time series is assumed to be centered, then $\mathbf{M}_t = \boldsymbol{\Sigma}_t$ and the above change–point problem considers a change in the variance of the projection $\boldsymbol{v}^\top \boldsymbol{Y}_t$.

Our time series model follows the widely used physical systems approach introduced by Wu (2005), which hosts many specific time series models as special cases, including autoregressive and moving average models, conditional heteroskedastic series and recursive stochastic models. We assume that $\boldsymbol{Y}_t = (Y_{tj})_{j=1}^d$, $t \geq 1$, is a nonlinear and possibly nonstationary vector time series with mean vectors $\boldsymbol{\mu}_t = (\mu_{t1}, \ldots, \mu_{td})^\top$, second moment matrix $\mathbf{M}_t = \mathbb{E}(\boldsymbol{Y}_t \boldsymbol{Y}_t^\top)$, covariance matrices $\boldsymbol{\Sigma}_t = (\Sigma_{t,ij})_{i,j=1,\ldots,d}$ and coordinates given by

$$Y_{tj} = G_{tj}(\epsilon_t, \epsilon_{t-1}, \ldots),$$

for a measurable mapping $\boldsymbol{G}_t = (G_{tj})_{j=1}^d : \mathbb{R}^\infty \to \mathbb{R}^d$ and i.i.d. $U(0,1)$–distributed random variables ϵ_t. To formulate the physical weak dependence assumptions, let

$$\boldsymbol{\epsilon}_t = (\epsilon_t, \epsilon_{t-1}, \ldots), \qquad \boldsymbol{\epsilon}_{t,j} = (\epsilon_t, \ldots, \epsilon_{j+1}, \epsilon'_j, \epsilon_{j-1}, \ldots)$$

for $j \leq t$ and $t \geq 1$, where $\{\epsilon'_t\}$ is an independent i.i.d. sequence with $\{\epsilon'_t\} \stackrel{d}{=} \{\epsilon_t\}$. For a vector $\boldsymbol{a} = (a_1, \ldots a_n)^\top \in \mathbb{R}^n$, $\|\boldsymbol{a}\|_r$ denotes the vector–r norm, $r \geq 1$, and $\|\boldsymbol{a}\|_\infty = \max_{1 \leq i \leq n} |a_i|$ the maximum norm.

2.1 Open-end monitoring

The open–end monitoring procedure essentially works as follows: At time $m + k$ of the monitoring period $m + 1, m + 2, \ldots$ the current sum of the most recent k transformed observations is compared to an average calculated from the training sample. To avoid that decisions are made based on too few data points of the monitoring period, the constraint $k \geq m\delta$ for some (small) $\delta > 0$ is imposed, so that monitoring starts at observation $m(1 + \delta)$ and continues either infinitely long or until a signal in favor of a change is given. We focus on detectors using different strategies to threshold a suitable cumulated sum centered at the corresponding average in the learning sample. They are constructed such that the probability of a false alarm is controlled. Such a procedure has a significance level comparable with the significance level of classical fixed sample Neyman–Pearson type hypothesis tests, which eases interpretation of results. But, clearly, the method differs from such a classical test in that it tries to come to the conclusion that the null hypothesis needs to be rejected as soon as the detector statistic provides enough evidence. This approach rules out classical methods studied in the field of statistical quality control which have the unpleasant property that in infinite monitoring an alarm is raised with probability one.

If we define the random matrices $\mathbf{S}_i = \boldsymbol{Y}_i \boldsymbol{Y}_i^\top$, then $\mathbb{E}(\mathbf{S}_i) = \mathbf{M}_i$. Therefore, to detect a change at the kth time point, $m + k$, of the monitoring period, we may consider for

known $\boldsymbol{v}$ the detector statistics

$$\left| \sum_{m<i\leq m+k} \boldsymbol{v}^\top \mathbf{S}_i \boldsymbol{v} - \frac{k}{m} \sum_{j=1}^{m} \boldsymbol{v}^\top \mathbf{S}_j \boldsymbol{v} \right|, \qquad k \geq \delta m.$$

Since usually $\boldsymbol{v}$ is unknown, it is replaced by some estimator $\hat{\boldsymbol{v}}_m$ computed from the training sample and safisfying Assumption A–iv (cf. next section), leading to the detector

$$Q(m,k) = \left| \sum_{m<i\leq m+k} \hat{\boldsymbol{v}}_m^\top \mathbf{S}_i \hat{\boldsymbol{v}}_m - \frac{k}{m} \sum_{j=1}^{m} \hat{\boldsymbol{v}}_m^\top \mathbf{S}_j \hat{\boldsymbol{v}}_m \right|.$$

All procedures signal at time $i \geq \delta m$ of the monitoring period and reject the null hypothesis of no change, if the trajectory of the standardized detector statistic, $\{Q(m,j)/\hat{\sigma}_{0,\infty} : j \geq \delta m\}$, crosses a suitable multiple of a boundary function, $\{cg(m,j) : j \geq \delta m\}$, at the ith time instant for the first time, where $\hat{\sigma}^2_{0,\infty}$ is an estimator of the asymptotic variance $\sigma^2_{0,\infty}$ discussed below. This means,

$$\frac{Q(m,i)}{\hat{\sigma}_{0,\infty} g(m,i)} > c, \quad \text{but} \quad \frac{Q(m,j)}{\hat{\sigma}_{0,\infty} g(m,j)} \leq c, \ m\delta \leq j < i. \tag{1}$$

$c > 0$ is a critical value, whose choice is discussed below. Note that the signal is given at the random stopping time

$$\tau_m = \inf\{k \geq m\delta : Q(m,k)/\hat{\sigma}_{0,\infty} > cg(m,k)\},$$

with the convention $\inf \emptyset = \infty$.

Erdös-Kac detector: The first procedure compares $Q(m,k)$ with a multiple of the simple boundary function

$$b(m,k) = m^{\frac{1}{2}} \left(\frac{m+k}{m} \right).$$

As we will formally show, the probability that $m^{-1/2}(1+k/m)^{-1} Q(m,k)/\sigma_{0,\infty}$ does not cross $b(m,k)$ converges to $\mathbb{P}(\sup_{\delta/(1+\delta)\leq t\leq 1} |B(t)| \leq c)$ for some Brownian motion B. Thus, comparing $Q(m,k)$ with a multiple of the boundary $b(m,k)$ asymptotically corresponds to comparing the absolute value of B with the constant c. For small δ one may select c such that

$$\mathbb{P}\left(\sup_{t\in[0,1]} |B(t)| \leq c \right) = \frac{4}{\pi} \sum_{i=0}^{\infty} \frac{(-1)^i}{2i+1} \exp\left(\frac{-(2i+1)^2}{\pi^2/8c^2} \right) = 1-\alpha,$$

see Erdös and Kac (1946) and (Shorack and Wellner, 1986, ch. 2).

γ–detector: The increasing variance of Brownian motion suggests to replace the latter constant by a curved boundary function. The choice $t^{1/2}$ would standardize B, but due to the law of the iterated logarithm the corresponding supremum is not well defined.

Following Horváth et al. (2004) and generalizing the Erdös–Kac detector, the γ–detector uses a multiple of the boundary

$$g(m,k) = m^{\frac{1}{2}} \left(\frac{m+k}{m}\right) \left(\frac{k}{m+k}\right)^{\gamma}$$

where γ is a tuning constant with $0 \leq \gamma < 1/2$. For a late change $k^*/(m+k^*) \approx 1$, so that the choice of γ has a negligible effect, but the closer the change to m, the smaller $k^*/(m+k^*)$, so that large values of γ close to $1/2$ are preferable for quick detection. Since monitoring may last forever and then analyzes infinitely many observations, we need to ensure that asymptotically the null hypothesis is never rejected with probability $1-\alpha$, if it is true. To achieve the latter, we select c such that

$$\lim_{m\to\infty} \mathbb{P}\left(\sup_{k \geq \delta m} \frac{Q(m,k)}{\hat{\sigma}_{0,\infty} g(m,k)} > c \right) = \alpha.$$

Our theoretical results, see Section 3, justify to select c as the (simulated) $1-\alpha$ quantile of the distribution function

$$\mathcal{D}_\infty(x) = \mathbb{P}\left(\sup_{\frac{\delta}{1+\delta} \leq t \leq 1} \frac{|B(t)|}{t^\gamma} \leq x \right), \qquad x \in \mathbb{R},$$

where $B(t)$ denotes a standard Brownian motion. Even for $\delta = 0$ explicit formulas for $\mathcal{D}_\infty(x)$ are not kown except $\gamma = 0$.

Robbins–Siegmund detector: Lastly, we consider a detector which does not require to simulate a critical value. It makes use of the fact that $|B(t)|$ crosses the boundary

$$g_a(t) = \sqrt{(t+1)(a^2 + \log(t+1))}, \qquad a > 0,$$

for some $t > 0$ with probability $e^{-a^2/2}$, i.e., $\mathbb{P}(|B(t)| \geq g_a(t), \exists t > 0) = e^{-a^2/2}$, see Robbins and Siegmund (1970) and Chu et al. (1996). This result suggests to use $a = \sqrt{2\log(\alpha)}$ to ensure a crossing probability of $\alpha \in (0,1)$. For random processes converging weakly to a standard Brownian motion $B(t)$, $t \geq 0$, this boundary–crossing probability carries over. However, this result is not directly applicable to $Q(m,k)$. But applying a suitable time transformation and interpreting the resulting rule in terms of the physical time scale $k \in \mathbb{N}$ yields the following detection scheme: Raise a signal at time $m+i$, $i \geq \delta m$, if

$$Q(m,i) > \hat{\sigma}_{0,\infty} g_a(m,i), \quad g_a(m,i) := \sqrt{m+i^*} \frac{m}{m-i^*} \sqrt{a^2 + \log \frac{m+i^*}{m}} \tag{2}$$

for the first time, where $i^* = \lceil i \frac{m}{m+i} \rceil$.

The estimator $\hat{\sigma}_{0,\infty}$ of the asymptotic variance $\sigma^2_{0,\infty} = \lim \mathrm{Var}(m^{-1/2} \sum_{i=1}^m \boldsymbol{v}^\top \mathbf{S}_i \boldsymbol{v})$ can be obtained as follows. If the training data is an i.i.d. sample, use the sample

variance of $\boldsymbol{v}^\top \mathbf{S}_i \boldsymbol{v}$, $1 \le i \le m$. For time series data under Assumptions A–i – A–iii of the next section we propose

$$\hat{\sigma}^2_{0,\infty} = \frac{1}{m-b+1} \sum_{j=0}^{m-b} (S_j(b) - \bar{S}(b))^2,$$

where $S_j(b) = \frac{1}{\sqrt{b}} \sum_{i=j-b+1}^{j} \boldsymbol{v}^\top \mathbf{S}_i \boldsymbol{v}$, $0 \le j \le m-b$, are subsampled scaled partial sums, $\bar{S}(b) = \frac{1}{m-b+1} \sum_{j=0}^{m-b} S_j(b)$, and b is a bandwidth parameter. Consistency of this estimator, which can be dated back to works of Carlstein (1986) and Peligrad and Shao (1995), has been shown in (Mies and Steland, 2023, Theorem 4.1), provided Assumptions A–i and A–ii hold and the bandwidth is selected such that $b = O(m^\rho)$ for some $0 < \rho < \frac{1}{2}$.

2.2 Closed–end monitoring

The above detection schemes can also be applied as closed–end schemes where monitoring stops when a prespecified time horizon parameterized as Tm, for some fixed $T > 0$, is reached. Let us briefly discuss the closed–end scheme for the γ–detector. Here, one is interested in detecting a change point $m\delta \le k^* \le Tm$ and thus considers the stopping time

$$\tau_m^T = \min\{m\delta \le k \le mT : Q(m,k)/\hat{\sigma}_{0,\infty} > c g(m,k)\}$$

with the convention that $\min \emptyset = \infty$ (no change detected). If the monitoring procedure examines the partial sums from δm to Tm, one has more freedom to choose a boundary function $g(m,k)$ including case $\gamma = 1/2$. The decision threshold c is now selected as the (simulated) $1-\alpha$ quantile of the distribution function

$$\mathcal{D}_T(x) = \mathbb{P}\left(\sup_{t \in [\delta, T]} \frac{|B_1(t) - t B_2(1)|}{(1+t)\left(\frac{t}{1+t}\right)^\gamma} \le x \right), \qquad x \in \mathbb{R},$$

for two independent standard Brownian motions B_1 and B_2.

2.3 Application to deep learners

To monitor a deep learner to decide when it needs retraining, the above algorithms are applied with

$$\boldsymbol{Y}_i \leftarrow f_H(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}_m), \boldsymbol{v} \leftarrow \hat{\boldsymbol{\beta}}_m, \qquad \text{(detector } D\text{, monitoring of network predictions)}$$

for $i \ge 1$, where $(\hat{\boldsymbol{\theta}}_m, \hat{\boldsymbol{\beta}}_m)$ are the trained connection weights of the deep learning.

The above detector considers a single output neuron. For a network with q outputs, one can apply a Bonferroni correction. Here, we monitor each output neuron separately, using a common critical value c corresponding to a nominal significance level α/q, and give a signal at time $m+k$ if at least one of the q detectors signals. If we denote by $\tau_m(i)$ the time points when the ith detectors gives a signal, the overall detector signals at time $\min(\tau_m(1), \ldots, \tau_m(q))$.

3. Asymptotics

The theoretical results justifying the proposed methods hold true under a general nonlinear time series model, whose assumptions are discussed in the detail in the next subsection. The results comprise Gaussian approximations of projected partial sums, on which the detectors rely and that is of independent interest, non–asymptotic bounds and convergence rates of estimators of means, second moments and covariance matrices, and the asymptotic theory of the proposed detectors. Here we provide the required asymptotic null distributions as well as asymptotics under local alternatives for the γ–detector. The class of local alternatives consists of functional alternatives going beyond the classical case of jump– or trend–like local alternatives.

3.1 Assumptions

We will now list and discuss all assumptions for the general nonlinear time series model $\boldsymbol{Y}_t = \boldsymbol{G}_t(\boldsymbol{\epsilon}_t)$, which substantially goes beyond typical settings assuming independent data. Besides classical moment assumptions, the assumptions limit the degree of dependence (A–i) and non-stationarity (A–ii), formalize the training sample (A–iii), and provide the required assumptions of proper estimators of the projection vector (A–iv).

Assumption A–i For some $\beta > 2$ and $q > 4$, $\|\boldsymbol{\mu}_t\|_\infty \leq K$,

$$(\mathbb{E}|G_{tj}(\boldsymbol{\epsilon}_0) - \mu_{tj}|^q)^{\frac{1}{q}} \leq K,$$

$$(\mathbb{E}|G_{tj}(\boldsymbol{\epsilon}_t) - G_{tj}(\boldsymbol{\epsilon}_{t,t-\ell})|^q)^{\frac{1}{q}} \leq C_1 \ell^{-\beta},$$

for $1 \leq j \leq d$, $\ell \geq 1$ and $t \geq 1$, for constants K and C_1.

If $\boldsymbol{Y}_t = \boldsymbol{Y}_t^{(m)}$, $t, m \geq 1$, is an array indexed by $m \in \mathbb{N}$, then the constants K, C_1 arising in Assumption A–i and all assumptions to follow are required to be universal, i.e., independent of d and m.

The first condition of A–i requires existence of the central moments of order $4+\varepsilon$, $\varepsilon > 0$ arbitrarily small. The second condition is a uniform coordinate–wise weak dependence assumption in the sense of Wu's physical dependence measure, so that the impact when replacing in Y_{tj} the lag ℓ innovation $\epsilon_{t-\ell}$ by some i.i.d. copy is of magnitude $\ell^{-\beta}$ in the L_q–norm.

Assumption A–ii There exists some $L > 0$ such that the inequality

$$V(\bar{q}) = \max_{1 \leq j \leq d} \sum_{t=2}^{n} \left(\mathbb{E}|G_{tj}(\boldsymbol{\epsilon}_0) - G_{t-1,j}(\boldsymbol{\epsilon}_0)|^{\bar{q}}\right)^{\frac{1}{\bar{q}}} \leq KL$$

holds for $\bar{q} = 4$, for all $n \geq 1$.

Assumption A–ii constrains the degree of nonstationarity by assuming that the cumulated increments $Y_{tj} - Y_{t-1,j}$ remain bounded when they are measured in the L_4–norm. This hold if the mapping $t \mapsto G_{tj}$ is smooth in a Lipschitz sense. The following example serves as a non–trivial illustration and discusses the case of random curves, such as temperature in climate studies, which are discretely sampled at time instants $t_k = \frac{k}{m}$, $1 \leq k \leq m$.

Example 1. Suppose that the time series Y_{tj} is obtained by discretely sampling d random curves $G_j(u, \boldsymbol{\epsilon}_t)$, $u \in [0,\infty)$, $G_j : [0,\infty) \times \mathbb{R}^\infty \to \mathbb{R}$, at equdistant time instants $u \in \{\frac{t}{m} : 1 \le t \le m\}$, such that $Y_{tj} = G_{tj}(\boldsymbol{\epsilon}_t) = G_j(\frac{t}{m}, \boldsymbol{\epsilon}_t)$, $1 \le j \le d$. For example, $G_j(\frac{t}{m}, \boldsymbol{\epsilon}_t)$ can be the temperature at day j of year t, and the learning sample consists of m years of daily temperature data. Here, the influence of time, t/m, on the functional dependence of the randomness $\boldsymbol{\epsilon}_t$ on the temperature may be nonlinear.

Let us assume that the functions $G_j(u, \boldsymbol{x})$ satisfy a uniform Lipschitz property

$$|G_j(u_1, \boldsymbol{x}) - G_j(u_2, \boldsymbol{x})| \le \mathrm{Lip}(G)|u_2 - u_1|H(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^\infty, u_1, u_2 \in [0,\infty), 1 \le j \le d,$$

for some Lipschitz constant $\mathrm{Lip}(G)$ and a function $H : \mathbb{R}^\infty \to \mathbb{R}$ with $\mathbb{E}(H^4(\boldsymbol{\epsilon}_0)) < \infty$. Then

$$\left(\mathbb{E}|G_{tj}(\boldsymbol{\epsilon}_0) - G_{t-1,j}(\boldsymbol{\epsilon}_0)|^4\right)^{\frac{1}{4}} \le \mathrm{Lip}(G)\frac{\left(\mathbb{E}H^4(\boldsymbol{\epsilon}_0)\right)^{\frac{1}{4}}}{m}, \qquad 1 \le j \le d, t \ge 1,$$

such that A–ii holds with $L = \mathrm{Lip}(G)(\mathbb{E}H^4(\boldsymbol{\epsilon}_0))^{1/4}/K$.

In our framework the first m observations represent a training sample, also called non–contamination period. In real world applications the training sample is sometimes carefully selected and assumed to satisfy much stronger assumptions than imposed here. Typical assumptions are that the training sample is a second–order stationary vector time series, or that it consists of i.i.d. random vectors of dimension d with covariance matrix $\boldsymbol{\Sigma}$, perhaps even assuming that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y}_i$ have i.i.d. (Gaussian) coordinates with finite fourth moments. These restrictive assumptions limit the applicability of results, since they are in practice often not fulfilled. They are included in the following assumption on the training sample.

Assumption A–iii The training sample $\boldsymbol{Y}_t$, $1 \le t \le m$, satisfies the no–change (stability) hypothesis

$$\theta(\mathbb{P}_{\boldsymbol{Y}_t}) = \boldsymbol{v}^\top \mathbf{M}_t \boldsymbol{v} = \theta_0, \qquad 1 \le t \le m, \tag{3}$$

for a constant θ_0, there exists a measurable mapping $\boldsymbol{G}_0 = (G_{0j})_{j=1}^d : \mathbb{R}^\infty \to \mathbb{R}^d$ such that for all $1 \le t \le m$, $1 \le j,k \le d$ and $h \ge 0$

$$\left|\frac{\mathrm{Cov}(G_{tj}(\boldsymbol{\epsilon}_t), G_{tk}(\boldsymbol{\epsilon}_{t+h}))}{\mathrm{Cov}(G_{0j}(\boldsymbol{\epsilon}_t), G_{0k}(\boldsymbol{\epsilon}_{t+h}))} - 1\right| \le \frac{C_2}{m}. \tag{4}$$

for some constant C_2, and

$$\inf_{i \ge 1} \lambda_{\min}(\boldsymbol{\Sigma}_{i,\infty}) > 0 \qquad \text{where} \qquad \boldsymbol{\Sigma}_{i,\infty} = \sum_{h=-\infty}^{\infty} \mathrm{Cov}(\boldsymbol{G}_i(\boldsymbol{\epsilon}_0)\boldsymbol{G}_i(\boldsymbol{\epsilon}_h)^\top). \tag{5}$$

Condition (3) of A–iii requires that $\theta(\mathbb{P}_{\boldsymbol{Y}_t})$ attains a constant value in the training sample, which is required to be known, but it nevertheless allows for nonstationarity to some extent with constraints on the process mean and the covariance structure. Clearly,

if $\boldsymbol{Y}_t$, $1 \leq t \leq m$, have mean zero, then $\theta(\mathbb{P}_{\boldsymbol{Y}_t}) = \boldsymbol{v}^\top \boldsymbol{\Sigma}_t \boldsymbol{v}$ and (3) collapses to the null hypothesis that $\mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{Y}_1) = \cdots = \mathrm{Var}(\boldsymbol{v}^\top \boldsymbol{Y}_m)$. Further, a sufficient condition for constancy of $\theta(\mathbb{P}_{\boldsymbol{Y}_i})$, $1 \leq i \leq m$, is to assume that

$$\boldsymbol{v}^\top \boldsymbol{\Sigma}_t \boldsymbol{v} = \sigma_0^2, \qquad \boldsymbol{v}^\top \boldsymbol{\mu}_t = m_0, \qquad 1 \leq t \leq m,$$

for constants $\sigma_0^2 \geq 0$ and $m_0 \in \mathbb{R}$.

Condition (4) of Assumption A–iii requires that the relative error when replacing any pseudo–lag h cross–covariance of $\boldsymbol{Y}_t = \boldsymbol{G}_t(\boldsymbol{\epsilon}_t)$ with $\boldsymbol{G}_t(\boldsymbol{\epsilon}_{t+h})$ by the lag h cross–covariance of $\boldsymbol{G}_0(\boldsymbol{\epsilon}_t)$ is of the order $O(\frac{1}{m})$. It limits the nonstationarity of the covariances but not necessarily those of other nuisance characteristics such as skewness or tail behavior. Condition (4) can be replaced by the assumption

$$|\operatorname{Cov}(G_{tj}(\boldsymbol{\epsilon}_t), G_{tk}(\boldsymbol{\epsilon}_{t+h})) - \operatorname{Cov}(G_{0j}(\boldsymbol{\epsilon}_t), G_{0k}(\boldsymbol{\epsilon}_{t+h}))| \leq \frac{k_h}{m} \tag{6}$$

for some summable sequence k_h, which can be formulated as

$$\|\operatorname{Cov}(\boldsymbol{G}_t(\boldsymbol{\epsilon}_t), \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t+h})) - \operatorname{Cov}(\boldsymbol{G}_0(\boldsymbol{\epsilon}_t), \boldsymbol{G}_0(\boldsymbol{\epsilon}_{t+h}))\|_\infty \leq \frac{k_h}{m}. \tag{7}$$

Condition (5) requires that the smallest eigenvalues of the matrices $\boldsymbol{\Sigma}_{i,\infty}$ are bounded away from zero. Combined with (4) this will ensure that all summands of the cumulated sum arising in the CUSUM detector of interest have asymptotically the same scale, such that a well defined limiting law exists which is governed by a standard Brownian motion. If these assumption are omitted, then one could still construct Gaussian approximations using Mies and Steland (2023), but then there may be no well defined asymptotic limit distribution of the detection scheme. Assumptions (4) and (5) ensure that, asymptotically, the procedure is governed by Brownian motion, and the asymptotic distribution is related to previous works on related detectors.

The question arises whether the assumption on the second moments is compatible with nonstationarity and the assumed form of $\boldsymbol{Y}_t$.

Example 2. Consider the vector time series

$$\boldsymbol{Y}_t = \sqrt{\frac{\nu_t - 2}{\nu_t}} \tilde{\boldsymbol{Y}}_t + e_t, \qquad t \geq 1,$$

where $e_t = \sum_{j=0}^{\infty} \rho^j \zeta_{t-j}$ for i.i.d. $\zeta_s \sim (0,1)$, $s \in \mathbb{Z}$, and some $\rho \in (-1,1)$, is a stationary autoregressive process of order 1 and $\tilde{\boldsymbol{Y}}_t \sim t(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu_t)$ with varying degrees of freedom specified as $\nu_t = \nu_0 + t$, $t \geq 1$, for some $\nu_0 > 4$. Here, the multivariate t–distribution, $t(\boldsymbol{\mu}, \Sigma, \nu)$, is defined as the law of $\boldsymbol{\mu} + \boldsymbol{Z}^*/\sqrt{Q/\nu}$ for independent random variables $Q \sim \chi^2(\nu)$ and $\boldsymbol{Z}^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Its covariance matrix is $\frac{\nu}{\nu-2}\boldsymbol{\Sigma}$. Thus, $\boldsymbol{Y}_t$ has stationary first and second order moments. However, the higher order moments are nonstationary, since the mixed moments of a $t(\boldsymbol{0}, \boldsymbol{\Sigma}, \nu_0 + t)$–distribution are given by

$$\mathbb{E}(Y_{t1}^{n_1} \cdot \ldots \cdot Y_{td_m}^{n_{d_m}}) = [(t-1)/(t+1)]^k (\nu_0 + t)^{k/2} \frac{\Gamma((\nu_0 + t - k)/2)}{2^k \Gamma((\nu_0 + t)/2)} \prod_{i=1}^{d_m} \frac{n_i!}{(n_i/2)!},$$

for $(n_1, \ldots, n_{d_m})^\top \in \mathbb{N}_0^{d_m}$, with $k = \sum_{i=1}^{d_n} n_i$, if all n_i are even and $\nu_0 + t > k$, and vanishes if at least one n_i is odd, see, e.g., Sutradhar (1986).

More generally, the question arises whether a specific admissible moment structure for the whole vector process can be realized in the assumed form $\boldsymbol{G}_t(\boldsymbol{\epsilon}_t)$. This question is related to two classical probems: Firstly, to the classical moment problem, which is not solved satisfactorily, especially in terms of algorithms, and, secondly, the problem to construct stochastic processes from prescribed marginals. However, under regularity conditions addressing these two basic problems, such a construction is possible, even on the original probability space, provided it carries uniform random variables ϵ_t, and this is a major motivation to study such nonlinear processes governed by i.i.d. innovations. For given real numbers $\mu_{\boldsymbol{n}}$, $\boldsymbol{n} \in \mathbb{N}_0^{kd_m}$, introduce the linear Riesz functional R on the space of polynomials in kd_m variables via $R(\boldsymbol{x}^{\boldsymbol{n}}) = \mu_{\boldsymbol{n}}$ for monomials $\boldsymbol{x}^{\boldsymbol{n}} = x_{11}^{n_{11}} \cdot \ldots \cdot x_{kd_m}^{n_{kd_m}}$, $\boldsymbol{n} \in \mathbb{N}_0^{kd_m}$.

Proposition 1. *Suppose that $\mu_{\boldsymbol{t},\boldsymbol{n}}^{(m)}, \boldsymbol{t} \in \mathbb{N}^k, \boldsymbol{n} \in \mathbb{N}_0^{kd_m}, k \geq 1$, are real numbers such that for all $\boldsymbol{t} \in \mathbb{N}^k$, $\mu_{\boldsymbol{t},\boldsymbol{n}}^{(m)}, \boldsymbol{n} \in \mathbb{N}_0^{kd_m}$, satisfy the Riesz–Haviland condition of positive semidefiniteness of the Riesz functional, i.e, $R(p) \geq 0$ for all polynomials p with $p(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$, equivalently,*

$$\sum_{\boldsymbol{n} \in \mathbb{N}_0^{kdm}} a_{\boldsymbol{n}} \mu_{\boldsymbol{t},\boldsymbol{n}}^{(m)} \geq 0, \qquad \forall \{a_{\boldsymbol{n}} : \boldsymbol{n} \in \mathbb{N}_0^{kd_m}\} \subset \mathbb{R}. \tag{8}$$

Then, for any $1 \leq t_1 < \cdots < t_k$, $k \in \mathbb{N}$, there exists a distribution $Q_{t_1,\ldots,t_k}$ and a random vector

$$(X_{t_1 1}, \ldots, X_{t_1 d_m}, \ldots, X_{t_k 1}, \ldots, X_{t_k d_m})^\top \sim Q_{t_1,\ldots,t_k}$$

with the prescribed moments $\mathbb{E}(X_{t_1 1}^{n_{11}} \cdot \ldots \cdot X_{t_k d_m}^{n_{t_k d_m}}) = \mu_{(t_1,\ldots,t_k),(n_{11},\ldots,n_{t_k d_m})}$, for all $\boldsymbol{n} = (n_{11}, \ldots, n_{t_k d_m})^\top \in \mathbb{N}_0^{kd_m}$. If $Q_{t_1,\ldots,t_k}$, $t_1 < \cdots < t_k$, $k \geq 1$, satisfy Kolmogorov's consistency conditions, then one can construct, on a new probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, a d_m–dimensional process, $\tilde{\boldsymbol{X}}_t$, $t \in \mathbb{N}$, whose finite dimensional distributions are given by the $Q_{t_1,\ldots,t_k}$. Given i.i.d. $\epsilon_t \sim U(0,1)$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$, one can construct, on $(\Omega, \mathcal{F}, \mathbb{P})$, a process $\boldsymbol{X}_t$, $t \geq 1$, of the form $\boldsymbol{X}_t = \boldsymbol{G}_t(\boldsymbol{\epsilon}_t)$ for Borel functions $\boldsymbol{G}_t$, with $\{\boldsymbol{X}_t : t \geq 1\} \stackrel{d}{=} \{\tilde{\boldsymbol{X}}_t : t \geq 1\}$.

Condition (8) cannot be weakened, since it is a necessary and sufficient condition to ensure that a probability measure with the given numbers as moments exists, see Haviland (1936) and Ghasemi et al. (2016) for recent extensions. The assumption $\boldsymbol{v}^\top \mathbf{M}_t \boldsymbol{v} = \theta_0$, $1 \leq t \leq m$, puts a constraint only on the elements $\mu_{(1,\ldots,m),\boldsymbol{n}}$ with $\sum_{i=1}^{d_m} n_{ti} = 2$ for $1 \leq t \leq m$, but the remaining free parameters allow to specify nonstationary higher order moments. However, to best of our knowledge, specific algorithms to construct distributions with prescribed moments are constraint to dimension one (and hence to multivariate distributions with independent coordinates). For example, Mead and Papanicolaou (1984) and Erdogmus et al. (2004) study algorithms maximizing entropy and John et al. (2007) an approach using splines. Clearly, one can combine these

methods with copulas, but this is certainly feasible in practice only to generate sequences of independent d_m–vectors.

If $\boldsymbol{v}$ is unknown, it is natural to estimate it from the training sample of size m. We consider the non–ℓ_0–sparse case and *s–sparseness*. The latter means that the true projection $\boldsymbol{v} = (v_1, \ldots, d)^\top$ has active set $A = \{j : v_j \neq 0\}$ of cardinality s.

Assumption A–iv Under s–sparseness there is a sparse estimator $\hat{\boldsymbol{v}}_m$ and constants $r_d \geq 1$, such that , with $\hat{A} = \{j : \hat{v}_j \neq 0\}$,

$$\mathbb{P}(\hat{A} \neq A) = O_{\mathbb{P}}\left(\sqrt{\frac{sr_d}{m}}\right), \quad \|\hat{\boldsymbol{v}}_m - \boldsymbol{v}\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{r_d}{m}}\right)$$

and it holds $\sqrt{sr_d} = O(m^\zeta)$ for some $0 \leq \zeta < \frac{1}{2}$. In the non–$\ell_0$–sparse setting, the estimator $\hat{\boldsymbol{v}}_m$ ensures the above convergence rate and it holds $\sqrt{dr_d} = O(m^\zeta)$.

This assumption includes the typical cases $r_d = d$, $r_d = d\log(d)$ and $r_d = s\log(d)$. We discuss this issue and some special cases of interest in a separate section. In the non–ℓ_0–sparse setting, the dimension d can grow at best almost at the rate $\sqrt{m}$. This growth of the dimension can be attained when $\boldsymbol{v}$ is ℓ_1–sparse with coordinates satisfying a decay condition.

Under s–sparseness the situation is much better. Especially, the constraint $d = O(\sqrt{m})$ on the dimension is no longer required. If, for example, $r_d = s\log d$, then the number of active variables can be almost of the order $O(\sqrt{m})$. But for the more important case that the number s of relevant variables grows slowly or is bounded, the constraint on the dimension d becomes very mild. For example, if $r_d = s\log(d)$ with $s = O(1)$, the dimension can be of the order $O(\exp(m^{2\eta}))$, i.e., d can grow almost exponentially in m.

3.2 Basic estimation under nonstationarity: Non–asymptotic guarantees

In view of the general assumptions the following theorem provides non–asymptotic consistency guarantees (i.e., for all $m \geq 1$) for sample moments. Let

$$\hat{\boldsymbol{\Sigma}}_m = \frac{1}{m}\sum_{t=1}^{m}(\boldsymbol{Y}_t - \hat{\boldsymbol{\mu}}_m)(\boldsymbol{Y}_t - \hat{\boldsymbol{\mu}}_m)^\top, \qquad \hat{\mathbf{M}}_m = \frac{1}{m}\sum_{t=1}^{m}\boldsymbol{Y}_t\boldsymbol{Y}_t^\top,$$

where $\hat{\boldsymbol{\mu}}_m = \frac{1}{m}\sum_{t=1}^{m}\boldsymbol{\mu}_t$. Further put $\bar{\boldsymbol{\mu}}_m = \mathbb{E}(\hat{\boldsymbol{\mu}}_m), \bar{M}_m = \mathbb{E}\hat{\mathbf{M}}_m$ and $\bar{\boldsymbol{\Sigma}}_m = \bar{M}_m - \bar{\boldsymbol{\mu}}_m\bar{\boldsymbol{\mu}}_m^\top$. In general, these population quantities are averages and depend on m, since we do not assume stationary data. For example, $\bar{\boldsymbol{\mu}}_m = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\mu}_i$, etc. Throughout, we indicate this by a bar–notation. Denote the elements of $\hat{\boldsymbol{\mu}}_m$ by $\hat{\mu}_{mj}$ and those of $\hat{\boldsymbol{\Sigma}}_m$ by $\hat{\Sigma}_{m,jk}$, etc. Let $\zeta(z) = \sum_{j=1}^{\infty} j^{-z}$, $z > 0$, denote Riemann's zeta function. The following theorem summarizes basic consistency results of the estimators $\hat{\boldsymbol{\Sigma}}_m, \hat{\boldsymbol{\mu}}_m$ and $\hat{\mathbf{M}}_m$. Threshold estimators of covariance matrices are discussed in greater detail in the next section.

Theorem 1. *Under Assumption A–i the sample mean $\hat{\boldsymbol{\mu}}_m$, sample covariance matrix $\hat{\boldsymbol{\Sigma}}_m$ and sample second moments, $\hat{\mathbf{M}}_m$, are elementwise $\sqrt{m}$–consistent in L_r and $L_{\frac{r}{2}}$,*

respectively, for any $2 \leq r \leq q$*, and hence in* L_2*, with*

$$\max_{1\leq j\leq d} (\mathbb{E}|\hat{\mu}_{mj} - \bar{\mu}_{mj}|^r)^{\frac{1}{r}} \leq \frac{c\zeta(\beta)}{\sqrt{m}},$$

$$\max_{1\leq j,k\leq d} \left(\mathbb{E}|\hat{M}_{m,jk} - \bar{M}_{jk}|^{\frac{r}{2}}\right)^{\frac{2}{r}} \leq \frac{c\zeta(\beta)}{\sqrt{m}},$$

$$\max_{1\leq j,k\leq d} \left(\mathbb{E}|\hat{\Sigma}_{m,jk} - \bar{\Sigma}_{m,jk}|^{\frac{r}{2}}\right)^{\frac{2}{r}} \leq \frac{c\zeta(\beta)}{\sqrt{m}},$$

for some universal constant c *(only depending on* q*) and all* $m \geq 1$*.*

The above results provide the following convergence rates with respect to the maximum and Frobenius norm.

Corollary 1. *Under the assumptions of Theorem 1*

$$\|\hat{\boldsymbol{\mu}}_m - \bar{\boldsymbol{\mu}}_m\|_\infty = O_{\mathbb{P}}\left(\frac{d^{\frac{2}{q}}}{\sqrt{m}}\right),\ \|\hat{\boldsymbol{\Sigma}}_m - \bar{\boldsymbol{\Sigma}}_m\|_\infty = O_{\mathbb{P}}\left(\frac{d^{\frac{4}{q}}}{\sqrt{m}}\right),\ \|\hat{\mathbf{M}}_m - \bar{\mathbf{M}}_m\|_\infty = O_{\mathbb{P}}\left(\frac{d^{\frac{4}{q}}}{\sqrt{m}}\right).$$

and

$$\|\hat{\boldsymbol{\mu}}_m - \bar{\boldsymbol{\mu}}_m\|_2 = O_{\mathbb{P}}\left(\frac{d^{\frac{2-q}{q}}}{\sqrt{m}}\right),\ \|\hat{\boldsymbol{\Sigma}}_m - \bar{\boldsymbol{\Sigma}}_m\|_F = O_{\mathbb{P}}\left(\frac{d^{2+\frac{4}{q}}}{\sqrt{m}}\right),\ \|\hat{\mathbf{M}}_m - \bar{\mathbf{M}}_m\|_F = O_{\mathbb{P}}\left(\frac{d^{2+\frac{4}{q}}}{\sqrt{m}}\right).$$

3.3 A sequential Gaussian approximation

The following theorem extends sequential Gaussian approximations for projected partial sums studied in Mies and Steland (2023) and Mies and Steland (2024) to estimated projections and provides approximations in terms of Brownian motion.

Theorem 2. *(Sequential Gaussian Approximation)*
Suppose that Assumptions A–i – A–iii are satisfied and $\|\boldsymbol{v}\|_1 \leq C_{\boldsymbol{v}}$ *for some constant* $C_{\boldsymbol{v}}$*. Let*

$$\xi(q,\beta) = \begin{cases} \frac{q-2}{6q-4}, & \beta \geq 3, \\ \frac{(\beta-2)(q-2)}{(4\beta-6)q-4}, & \frac{3+\frac{2}{q}}{1+\frac{2}{q}} < \beta < 3, \\ \frac{1}{2} - \frac{1}{\beta}, & 2 < \beta \leq \frac{3+\frac{2}{q}}{1+\frac{2}{q}}. \end{cases}$$

For some standard Brownian motion $\{B_t : t \geq 0\}$ *the following assertions hold.*

(i) We have

$$\max_{k\leq m} \left| \sum_{i=1}^{k} \frac{\boldsymbol{v}^\top(\mathbf{S}_i - \mathbf{M}_i)\boldsymbol{v}}{\sigma_{i,\infty}} - B_k \right| = O_{\mathbb{P}}\left(m^{\frac{1}{\nu}}\right)$$

as well as

$$\max_{k\leq m} \left| \sum_{i=1}^{k} \frac{\boldsymbol{v}^\top(\mathbf{S}_i - \mathbf{M}_i)\boldsymbol{v}}{\sigma_{0,\infty}} - B_k \right| = O_{\mathbb{P}}\left(m^{\frac{1}{\nu}}\right)$$

with $\nu = \frac{1}{\frac{1}{2}-\xi(q,\beta)+\varepsilon} > 2$ *for* $0 < \varepsilon < \xi(q,\beta)$*, where*

$$\sigma_{i,\infty}^2 = \sum_{h\in\mathbb{Z}} \boldsymbol{v}^\top \operatorname{Cov}(\boldsymbol{G}_i(\boldsymbol{\epsilon}_0), \boldsymbol{G}_i(\boldsymbol{\epsilon}_h))\boldsymbol{v},\ i \geq 0.$$

(ii) If Assumption A–iv holds, then

$$\max_{k\leq m}\left|\sum_{i=1}^{k} \frac{\hat{\boldsymbol{v}}_m^\top(\mathbf{S}_i - \mathbf{M}_i)\hat{\boldsymbol{v}}_m}{\sigma_{i,\infty}} - B_k\right| = O_{\mathbb{P}}\left(m^{\frac{1}{\nu}}\right)$$

and

$$\max_{k\leq m}\left|\sum_{i=1}^{k} \frac{\hat{\boldsymbol{v}}_m^\top(\mathbf{S}_i - \mathbf{M}_i)\hat{\boldsymbol{v}}_m}{\sigma_{0,\infty}} - B_k\right| = O_{\mathbb{P}}\left(m^{\frac{1}{\nu}}\right)$$

with $\nu = \min\left(\frac{1}{\frac{1}{2}-\xi(q,\beta)+\varepsilon}, \frac{1}{\zeta+\varepsilon}\right) > 2$ *for some small* $\varepsilon > 0$.

The first assertion of part (i) of Theorem 2 refines Mies and Steland (2024), where related Gaussian approximations (not in terms of a Brownian motion) are derived under a slightly different assumption and used to study shrinkage covariance matrix estimators.

3.4 Detector asymptotics under H_0 and local alternatives

The following theorem entails the detector's limiting distribution using the boundary function $g(m,k)$, when no change is present.

Theorem 3. *Suppose that Assumptions A–i – A–iii are fulfilled and* $\|\boldsymbol{v}\|_1 \leq C_{\boldsymbol{v}}$ *for some constant* $C_{\boldsymbol{v}}$*. If* $\boldsymbol{v}$ *is estimated, additionally assume that A–iv holds, otherwise formally put* $\hat{\boldsymbol{v}}_m = \boldsymbol{v}$*. Under the null hypothesis* H_0 *the detector statistic converges in distribution,*

$$\lim_{m\to\infty} \mathbb{P}\left(\sup_{k\geq\delta m} \frac{Q(m,k)}{\hat{\sigma}_{0,\infty} g(m,k)} \leq x\right) = \mathbb{P}\left(\sup_{\frac{\delta}{1+\delta}\leq t\leq 1} \frac{|B(t)|}{t^\gamma} \leq x\right), \qquad x \in \mathbb{R},$$

where $B(t)$*,* $0 \leq t < \infty$*, is a standard Brownian motion.*

Remark 1. Note that the $\mathcal{S}_{\delta,\gamma} = \sup_{\frac{\delta}{1+\delta}\leq t\leq 1} \frac{|B(t)|}{t^\gamma} \to \mathcal{S}_{0,\gamma} = \sup_{0\leq t\leq 1} \frac{|B(t)|}{t^\gamma}$, $\delta \to 0$, so that for small δ one can work with $\mathcal{S}_{0,\gamma}$. Explicit formulas for $\mathcal{S}_{0,\gamma}$ are, however, only known for $\gamma = 0$, but one can simulate quantiles or rely on quantiles provided in the literature.

For the Robbins–Siegmund detector, (2), we have the following result.

Theorem 4. *Suppose that Assumptions A–i – A–iii hold true. Then*

$$\lim_{m\to\infty} \mathbb{P}\left(Q(m,k) > \hat{\sigma}_{0,\infty}\sqrt{m+k^*}\frac{m}{m-k^*}\sqrt{a^2 + \log\frac{m+k^*}{m}}, \exists \delta m \leq k < m\right) \uparrow \alpha,$$

as $\delta \downarrow 0$*, where* $k^* = \lceil k\frac{m}{m+k}\rceil$ *and* $a = \sqrt{2\log(1/\alpha)}$.

Let us briefly discuss how these results are shown. The derivation of Theorem 3 relies on the Gaussian approximation of Theorem 2. That result also yields the asymptotics of the related closed–end procedures using weighting functions of the form

$$g(m,k) = g\left(\frac{k}{m}\right), \qquad g : [0,\infty) \to (0,\infty).$$

One obtains, by the continuous mapping theorem,

$$\max_{m\delta \le k \le mT} \frac{Q(m,k)}{\hat{\sigma}_{0,\infty} g(k/m)} \xrightarrow{d} \sup_{\delta \le t \le T} \frac{|B_1(t) - tB_2(t)|}{g(t)},$$

as $m \to \infty$, where B_1, B_2 are two independent Brownian motions, and the fact that $B_1(t) - tB_2(t)$ is equal in distribution to $(1+t)B(t/(t+1))$ then leads to the result. The proof of Theorem 4 applies a time transformation so that the latter process becomes a standard Brownian motion. Careful inspection and interpretation of the related boundary crossing result for the time transformed $Q(m,k)$ then yield the rule (2) and Theorem 4.

In the sequel we study asymptotic properties under alternatives and focus on the procedure (1). It is well known that the γ–detector based on $Q(m,k)$ has asymptotic power 1 for a fixed alternative for the mean of the cumulated random variables. For very early changes $k^* \ll \sqrt{m}$ the behaviour of such weighted CUSUMs for changes in the mean has been studied in Aue and Horváth (2004) assuming independent errors. Here we want to study the behaviour under local functional alternatives which converge to 0, if m increases, and allow for a functional shape after the change point for the problem under investigation. For that purpose, we consider the closed–end monitoring setting where monitoring stops at the time horizon Tm, for some fixed $T > 0$.

We formulate the local functional alternative model directly in terms of the moment functional and thus assume that for $k \ge 1$

$$\boldsymbol{v}^\top \mathbf{M}_{m+k} \boldsymbol{v} = \theta_0 + \frac{\Delta\left(\frac{k-k^*}{m}\right)}{\sqrt{m}} \mathbf{1}_{k \ge k^*} \tag{9}$$

for some integrable function $\Delta : [0,T] \to \mathbb{R}$ which is right-continuous at 0 and satisfies $\Delta(0) \neq 0$. The function $\Delta(x)$ models the shape of the moment functional after the change-point k^* which is assume to satisfy

$$\frac{k^*}{m} \to \vartheta, \qquad m \to \infty,$$

for some $\vartheta \in (0,T)$.

Theorem 5. *Suppose that the assumptions of Theorem 3 are fulfilled. If Δ has bounded total variation on $[0,T]$, then under the sequence of local functional alternatives (9)*

$$\lim_{m\to\infty} \mathbb{P}\left(\max_{\delta m \le k \le Tm} \frac{Q(m,k)}{\hat{\sigma}_{0,\infty} g(m,k)} \le x \right) = \mathcal{D}_T(x;\Delta)$$

where

$$\mathcal{D}_T(x;\Delta) = \mathbb{P}\left(\sup_{t\in[\delta,T]} \frac{|B_1(t) - tB_2(1) + \mathbf{1}_{t\geq\vartheta} \int_0^{t-\vartheta} \Delta(s)\, ds|}{(1+t)\left(\frac{t}{1+t}\right)^\gamma} \leq x \right),$$

$x \in \mathbb{R}$, *where* B_1 *and* B_2 *are independent standard Brownian motions.*

4. Estimating projections depending on covariance and precision matrices

As discussed in detail in the next section, in finance and deep learning one would like to select the projection by optimizing some criterion, which leads to an optimal projection, $\boldsymbol{v}$, depending on the covariance or second moment matrix (or its inverse) and the mean. As a general and widely applicable approach we consider threshold estimators dating back to the works of Bickel and Levina (2008) and Rothman et al. (2009) for i.i.d. samples. Here we show that the results can be generalized to nonlinear nonstationary vector time series. The results apply to covariance matrices as well as second moment matrices and their estimator. For brevity of presentation, we formulate the results for covariance matrices. Introduce the uniformity class

$$\mathcal{U}(r, s_0, M, \varepsilon_0) = \left\{ \mathbf{\Sigma} : \Sigma_{ii} \leq M, \sum_{j=1}^{d} |\Sigma_{ij}|^r \leq s_0, 1 \leq i \leq d, \lambda_{\min}(\mathbf{\Sigma}) \geq \varepsilon_0 > 0 \right\},$$

for some constant $s_0 = s_0(d)$ and some $0 \leq r < 1$. Matrices of this class have bounded variances and their spectrum is contained in $[\varepsilon_0, M^{1-r}c_0(d)]$, since the general upper bound $\max_i \sum_j |\Sigma_{ij}|$ for the maximal eigenvalue $\lambda_{\max}(\mathbf{\Sigma})$ is bounded by $M^{1-r}s_0$ for each $\mathbf{\Sigma} \in \mathcal{U}(r, s_0, M, \varepsilon_0)$. The thresholded estimator is defined by

$$\hat{\mathbf{\Sigma}}_{th} = \mathcal{T}_t \hat{\mathbf{\Sigma}}_m = (\hat{\Sigma}_{m,ij} \mathbf{1}_{|\hat{\Sigma}_{m,ij}| \geq t}) \, {}_{1\leq j\leq d}^{1\leq i\leq d},$$

for some threshold $t \geq 0$, where $\hat{\mathbf{\Sigma}}_m$ is the sample covariance matrix. The operator $\mathcal{T}_t : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ performs a hard–thresholding operation. Soft–thresholding also nulls entries which are smaller in magnitude than t, but allows to shrink entries towards zero, if their magnitude is small but larger than the threshold. Any operator $\mathcal{S}_t : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ which operates element–wise, i.e., $\mathcal{S}_t\mathbf{\Gamma} = (\mathcal{S}_t(\Gamma_{ij}))_{ij}$, $\mathbf{\Gamma} = (\Gamma_{ij})_{ij} \in \mathbb{R}^{d\times d}$, and satisfies

(i) $|\mathcal{S}_t(x)| \leq |x|$,

(ii) $\mathcal{S}_t(x) = 0$ for $|x| \leq t$,

(iii) $|\mathcal{S}_t(x) - x| \leq t$,

is called a soft–thresholding operator. Important special cases are hard–thresholding, lasso soft–thresholding defined by

$$\mathcal{S}_t^l(x) = \text{sign}(x)(|x| - t)_+,$$

and SCAD thresholding, which is between hard–thresholding and lasso thresholding. The resulting error when approximating some covariance matrix $\mathbf{\Sigma}$ by a (soft–) thresholded version, $\mathcal{S}_t\mathbf{\Gamma}$, of an arbitrary $d \times d$ matrix $\mathbf{\Gamma}$ attains the following crucial bound with respect to the spectral norm $\|\cdot\|_{op}$, when $\mathbf{\Sigma}$ is constrained to $\mathcal{U}(r, s_0, M)$.

Theorem 6. *Suppose that $\mathbf{\Sigma} \in \mathcal{U}(r, s_0, M)$. Then for any symmetric $d \times d$ matrix $\mathbf{\Gamma} = (\Gamma_{ij})_{ij}$ the soft–threshold operator $\mathcal{S}_t$ satisfies the bound*

$$\|\mathcal{S}_t\mathbf{\Gamma} - \mathbf{\Sigma}\|_{op} \leq 2t^{1-r}s_0 + \|\mathbf{\Gamma} - \mathbf{\Sigma}\|_\infty \left(\#(|\Gamma_{ij} - \Sigma_{ij}| > (1-\gamma)t) + \frac{1}{(\gamma t)^r}s_0 + \frac{2}{t^r}s_0\right)$$

for any $0 < \gamma < 1$.

The following theorem extends the result of Bickel and Levina (2008) and Rothman et al. (2009) for i.i.d. $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ with uniformly bounded moments of order q to the nonlinear nonstationary times series model satisfying Assumptions A–i and A–ii.

Theorem 7. *Under Assumptions A–i and A–ii the thresholded sample covariance matrix is consistent for $\bar{\mathbf{\Sigma}}_m \in \mathcal{U}(r, s_0, M, \varepsilon_0)$ in the operator norm with*

$$\|\mathcal{S}_t\hat{\mathbf{\Sigma}}_m - \bar{\mathbf{\Sigma}}_m\|_{op} = O_{\mathbb{P}}\left(\left(\frac{d^{\frac{4}{q}}}{\sqrt{m}}\right)^{1-r}\right)$$

if the threshold is selected as

$$t_{th} = C_{th}\frac{d^{\frac{4}{q}}}{\sqrt{m}}$$

for some large enough constant C_{th} and $d = o\left(m^{\frac{q}{8}}\right)$.

A related problem is estimation of the precision matrix $\bar{\Phi}_m = \bar{\mathbf{\Sigma}}_m^{-1}$. Here one may invert the thresholded estimator $\mathcal{S}_{t_m}\hat{\mathbf{\Sigma}}_m$ and thus define the estimator

$$\hat{\Phi}_m = (\mathcal{S}_{t_m}\hat{\mathbf{\Sigma}}_m)^{-1}.$$

This allows estimation of projection vectors which are smooth functions of the precision matrix $\bar{\Phi}_m$ and the expectation $\bar{\boldsymbol{\mu}}_m$ as required when computing optimal portfolios.

Theorem 8. *Under the assumptions of Theorem 7 we have*

$$\|\hat{\Phi}_m - \bar{\Phi}_m\|_{op} = O_{\mathbb{P}}\left(\left(\frac{d^{\frac{4}{q}}}{\sqrt{m}}\right)^{1-r}\right).$$

A projection given by $\boldsymbol{v} = \boldsymbol{v}_m = f(\bar{\mathbf{\Sigma}}_m^{-1}, \bar{\boldsymbol{\mu}}_m)$, for some Lipschitz continuous function f, can be estimated by

$$\hat{\boldsymbol{v}}_m = f((\mathcal{S}_{t_m}\hat{\mathbf{\Sigma}}_m)^{-1}, \hat{\boldsymbol{\mu}}_m),$$

at the same convergence rate provided the estimator $\hat{\boldsymbol{\mu}}_m$ attains the same convergence rate.

It is worth discussing the rate of convergence obtained in the above results and compare it with the classical parametric rate of convergence $O(1/\sqrt{m})$. For small r the convergence rate is almost $O_{\mathbb{P}}\left(\frac{r_d}{\sqrt{m}}\right)$ with $r_d = d^{\frac{4}{q}}$, and for $q \to \infty$ one obtains almost the parametric rate $O_{\mathbb{P}}(\frac{1}{\sqrt{m}})$.

5. Applications and Simulations

5.1 Deep neural networks

We consider a (deep) artificial neural network with H hidden layers for a d–dimensional input $\boldsymbol{x} \in \mathcal{D} \subset \mathbb{R}^d$ and a univariate output y defined recursively as

$$\boldsymbol{f}_1(\boldsymbol{x}) = \sigma_1(\mathbf{W}_1\boldsymbol{x}), \qquad \boldsymbol{f}_j(\boldsymbol{x}) = \sigma_j(\mathbf{W}_j\boldsymbol{f}_{j-1}(\boldsymbol{x})), \quad 2 \le j \le H,$$

for weight matrices $\mathbf{W}_j \in \mathbb{R}^{n_j \times n_{j-1}}$ and (nonlinear) activation functions σ_j, $1 \le j \le H$, where $n_0 = d$ is the input dimension and $n_j \in \mathbb{N}$ is the width of the jth layer, $1 \le j \le H$. The linear output layer linearly combines the features generated by the H layers and computes the output

$$y = f_{net}(\boldsymbol{x}) = \boldsymbol{\beta}^\top f_H(\boldsymbol{x}),$$

with $\boldsymbol{\beta} \in \mathbb{R}^{n_H}$. For multivariate output of dimension n_o one takes a coefficient matrix $\mathbf{B} \in \mathbb{R}^{n_o \times n_H}$ and calculates $\boldsymbol{y} = \mathbf{B}f_H(\boldsymbol{x})$. The network can be compactly written as

$$\boldsymbol{y} = \mathbf{B}\sigma_H^{\mathbf{W}_H} \circ \cdots \circ \sigma_1^{\mathbf{W}_1}(\boldsymbol{x}), \qquad \sigma_j^{\mathbf{W}_j}(\boldsymbol{x}) = \sigma_j(\mathbf{W}_j\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^{n_{j-1}}.$$

And if we put $\sigma_{H+1}(x) = x$ and $\mathbf{W}_{H+1} = \mathbf{B}$, then

$$\boldsymbol{y} = \sigma_{H+1}^{\mathbf{W}_{H+1}} \circ \cdots \circ \sigma_1^{\mathbf{W}_1}(\boldsymbol{x}).$$

For simplicity of presentation, however, we consider a univariate output in what follows.

The activation functions, $\sigma_j : \mathbb{R} \to \mathbb{R}$, act elemenwise on vectors. Common choices of the activation functions, σ, such as the ReLU given by $\max(x, 0)$, the sigmoid $\frac{1}{1+e^{-x}}$ or the tanh activation $\frac{e^x - e^{-x}}{e^x + e^{-x}}$, are Lipschitz continuous and satisfy $\sigma(0) = 0$. Indeed, Lipschitz continuity has beneficial effects for deep learning. Whereas the latter choices are differentiable, the widely used ReLU activation is not differentiable at 0, where it has the subdifferential $[0, 1]$. However, one may approximate it by $\sigma_k(x) = \frac{1}{2k} log(1 + e^{2kx})$ for some large $k > 0$.

Denote by $\boldsymbol{\theta} = (\mathbf{W}_1, \ldots, \mathbf{W}_H)$ the parameters of the hidden layers; the full Euclidean parameter is then $\boldsymbol{\eta} = (\mathrm{vec}(\boldsymbol{\theta}), \boldsymbol{\beta})$. Ocassionally, we write $\boldsymbol{f}_j(\boldsymbol{x}; \boldsymbol{\theta})$ and $f_{net}(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\theta})$ to stress the dependence on the weights.

The following result establishes the Lipschitz property of a deep neural network, which plays an important role in the generalization capabilities of a network, Bartlett et al. (2017). The result is similar to (Lederer, 2024, Prop. 6), but our proof is much simpler and the Lipschitz constant is typically smaller (note that typically $L_H \ll 1$). For

bounds on the Lipschitz constant for specific hidden layers and a discussion of related generalzation bounds see Gouk et al. (2021). Define for $1 \leq k \leq H$

$$\boldsymbol{\theta}(\widetilde{\mathbf{W}}_k) = (\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_k, \mathbf{W}_{k+1}, \ldots, \mathbf{W}_H),$$

and $\boldsymbol{\theta}(\widetilde{\mathbf{W}}_0) = (\mathbf{W}_1, \ldots, \mathbf{W}_H)$.

Proposition 2. *Consider a deep learning network given by* $\sigma_H^{\mathbf{W}_H} \circ \cdots \circ \sigma_1^{\mathbf{W}_1}(\boldsymbol{x})$ *with* ρ_i*–Lipschitz continuous activation functions* $\sigma_i : \mathbb{R} \to \mathbb{R}$ *satisfying* $\sigma_i(0) = 0$, $1 \leq i \leq H$. *Suppose that* $\mathcal{D} \times \Theta$ *is bounded or* σ_H *is bounded. Then*

$$\|\boldsymbol{f}_j(\boldsymbol{x}; \boldsymbol{\theta}(\widetilde{\mathbf{W}}_k)) - \boldsymbol{f}_j(\boldsymbol{x}; \boldsymbol{\theta}(\mathbf{W}_k))\|_2 \leq L_{jk} \|\boldsymbol{x}\|_2 \|\widetilde{\mathbf{W}}_k - \mathbf{W}_k\|_{op},$$

for $1 \leq k \leq H$, *where* $L_{jk} = \prod_{i=1, i \neq k}^{j} \rho_i \|\mathbf{W}_i\|_{op} \rho_k$.

The neural network has the Lipschitz property

$$\|\boldsymbol{f}_H(\boldsymbol{x}; \boldsymbol{\theta}) - \boldsymbol{f}_H(\boldsymbol{x}; \widetilde{\boldsymbol{\theta}})\|_2 \leq L_H \sqrt{H} \|\boldsymbol{x}\|_2 \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_F,$$

where $L_H = \max_{1 \leq k \leq H} L_{Hk}$.

The deep learning regression model assumes that before a change–point, k^*,

$$Y_t = \boldsymbol{\beta}_0^\top \boldsymbol{f}_H(\boldsymbol{X}_t; \boldsymbol{\theta}_0) + e_t, \qquad t \leq k^*,$$

for true parameters $\boldsymbol{\beta}_0, \boldsymbol{\theta}_0$ (see below), and a zero mean noise process $\{e_t\}$.

Whereas there is a vast of literature on performance guarantees for neural networks, results about parameter–consistency are quite limited. Under ideal conditions, nonlinear least squares estimators of fixed–topology neural networks are consistent and asymptotically normal, White (1992). We confine here to this setting, mainly because recent and ongoing work on statistical guarantees for networks with multiple layers is not yet settled, although the results of Loh and Wainwright (2015) for penalized M–estimators under restricted strong convexity (RSC) are potentially applicable to wide classes of deep networks, see also the recent result about RSC for the least squares loss of Banerjee et al. (2022), where local minima in layer–wise ℓ_2–balls around a random initialization are considered. Further, for the perceptron with additive i.i.d. noise Yang et al. (2016) have shown that the ℓ_1–penalized least squares estimator, $\hat{\boldsymbol{\beta}}_m$, attains under suitable conditions with high probability the optimal rate of convergence $O\left(\sqrt{\frac{s \log d}{m}}\right)$, where s is the number of active entries of the true parameter.

Let us now fix a multiple layered network as above and consider minimization of the least squares loss in the training sample,

$$\ell_m(\boldsymbol{\eta}) = \frac{1}{m} \sum_{i=1}^{m} \ell((Y_i, \boldsymbol{X}_i); \boldsymbol{\eta}), \qquad \ell((y, \boldsymbol{x}), \boldsymbol{\eta}) = (y - f_{net}(\boldsymbol{x}; \boldsymbol{\eta}))^2, y \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^d,$$

over a compact set $\mathcal{B} \times \Theta$ around a random initialization.

In practice, an iterative numerical solver is applied which ideally converges to the local minimum of the error landscape closest to the initialization. Common algorithms used in practice are classical backpropagation, stochastic gradient descent, Robbins and Monro (1951), and ADAM, Kingma and Ba (2015), which is the state–of–art algorithm from a practical point of view due its superior running time. Let us assume that the numerical solver succeeds in finding a stationary point of $\ell_m(\boldsymbol{\eta})$ located in $\mathcal{B} \times \Theta$, and thus yields an estimator $\hat{\boldsymbol{\eta}}_m = (\hat{\boldsymbol{\theta}}_m, \hat{\boldsymbol{\beta}}_m)$ of the true value $\boldsymbol{\eta}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) = \operatorname{argmin}_{(\boldsymbol{\beta},\boldsymbol{\theta})\in\mathcal{B}\times\Theta} \mathbb{E}\ell_m(\boldsymbol{\beta}, \boldsymbol{\theta})$, assumed to be an interior point and unique. Suitable conditions for smooth activation functions such that weakly consistent solutions exist, a.s., can be formulated in terms of the estimating function $G_m(\boldsymbol{\eta}) = \nabla\ell_m(\boldsymbol{\eta})$, see Jacod and Soerensen (2018):

(i) $G_m(\boldsymbol{\eta}_0) \xrightarrow{\mathbb{P}} 0$, $m \to \infty$.

(ii) G and all G_m, $m \geq 1$, are a.s. differentiable on $M = \mathcal{B} \times \Theta$ and ∇G_m converges uniformly in probability,

$$\sup_{\boldsymbol{\eta}\in M} \|\nabla G_m(\boldsymbol{\eta}) - \nabla G(\boldsymbol{\eta})\|_2 \xrightarrow{\mathbb{P}} 0.$$

(iii) The Hessian satisfies $\nabla^2 G(\boldsymbol{\eta}_0) > 0$.

If, additionally, the estimating function satisfies a central limit theorem,

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \nabla\ell((Y_i, \boldsymbol{X}_i), \boldsymbol{\eta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}(\boldsymbol{\eta}_0)),$$

as $m \to \infty$, for some matrix $\mathbf{V}(\boldsymbol{\eta}_0) > 0$, then the estimator $\hat{\boldsymbol{\eta}}_m$ attains the convergence rate $O_{\mathbb{P}}(\frac{1}{\sqrt{m}})$. If the training sample $(Y_1, \boldsymbol{X}_1), \ldots, (Y_m, \boldsymbol{X}_m)$ forms a strictly stationary nonlinear vector time series of the form $\boldsymbol{H}(\boldsymbol{\epsilon}_t)$ satisfying Assumptions A–i – A–ii, then the above conditions (i) and (ii) typically hold, especially if Y_t and $\boldsymbol{X}_t$ are bounded and the activation functions are C^2, since $M = \mathcal{B} \times \Theta$ is bounded, cf. Theorem 1. A central limit theorem for the estimating function $G_m(\boldsymbol{\eta}) = \nabla\ell_m(\boldsymbol{\eta})$ holds as well, where the positive definiteness of the limiting covariance matrix depends on the neural network and the dependence structure of the training sample and thus needs to be assumed, cf. Theorem 2.

Denote $\boldsymbol{Z}_i(\boldsymbol{\beta}) = \boldsymbol{f}_H(\boldsymbol{X}_i; \boldsymbol{\theta})$, $1 \leq i \leq m$. Observe that $\boldsymbol{\beta}_0$ is a solution of the normal equations of the output layer,

$$\boldsymbol{\Sigma}_Z \boldsymbol{\beta}_0 = \mathbb{E}(\boldsymbol{Z}_1(\boldsymbol{\theta}_0) Y_1), \qquad \text{with } \boldsymbol{Z}_i(\boldsymbol{\beta}) = \boldsymbol{f}_H(\boldsymbol{X}_i; \boldsymbol{\theta}).$$

It attains the explicit representation $\boldsymbol{\beta}_0 = \boldsymbol{\Sigma}_Z^{-1} \mathbb{E}(\boldsymbol{Z}_1(\boldsymbol{\theta}_0) Y_1)$ if $\boldsymbol{\Sigma}_Z > 0$. $\hat{\boldsymbol{\beta}}_m$ solves the corresponding sample normal equations,

$$\hat{\boldsymbol{\Sigma}}_{Zm}(\hat{\boldsymbol{\theta}}_m) \hat{\boldsymbol{\beta}}_m = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{Z}_i(\hat{\boldsymbol{\theta}}_m) Y_i$$

where $\hat{\boldsymbol{\Sigma}}_{Zm}(\hat{\boldsymbol{\theta}}_m) = \frac{1}{m}\sum_{i=1}^m \boldsymbol{Z}_i(\hat{\boldsymbol{\theta}}_m)\boldsymbol{Z}_i(\hat{\boldsymbol{\theta}}_m)^\top$. Thus, under the assumptions discussed above, $\hat{\boldsymbol{\beta}}_m = \boldsymbol{\beta}_0 + O_{\mathbb{P}}(\frac{1}{\sqrt{m}})$ with $\boldsymbol{\beta}_0 = \boldsymbol{\Sigma}_Z(\boldsymbol{\theta}_0)^{-1}\mathbb{E}(\boldsymbol{Z}_1(\boldsymbol{\theta}_0)Y_1)$. The optimal weights encode the second moment structure of the last hidden layer and its covariances with the responses.

For a new input $\boldsymbol{X}^*$ the true network prediction is $\boldsymbol{\beta}_0^\top \boldsymbol{f}_H(\boldsymbol{X}^*;\boldsymbol{\theta}_0)$ and the prediction of the trained network is $\hat{\boldsymbol{\beta}}_m^\top \boldsymbol{f}_H(\boldsymbol{X}^*;\hat{\boldsymbol{\theta}}_m)$. As long as the projection $\boldsymbol{\beta}_0^\top \boldsymbol{f}_H(\boldsymbol{X}^*;\boldsymbol{\theta}_0)$ does not change, there is no need to retrain the network. But, contrary, if that projection changes, which will be indicated sooner or later when cumulating information contained in the stream $f_H(\boldsymbol{X}_{m+i};\hat{\boldsymbol{\theta}}_m)$, $i \geq 1$, the network predictions may no longer be valid. This leads to the detector, D, with

$$\begin{aligned} \boldsymbol{Y}_{m+i} &\leftarrow f_H(\boldsymbol{X}_{m+i};\hat{\boldsymbol{\theta}}_m), \qquad i \geq 1, \\ \hat{\boldsymbol{v}}_m &\leftarrow \hat{\boldsymbol{\beta}}_m, \end{aligned}$$

in order to detect a change in the regressors pointing to the necessity to retrain the network.

5.2 Financial portfolio optimization

In financial portfolio optimization one aims at determining a portfolio vector $\boldsymbol{w}$ of the proportions of money to invest in each asset, such that $\sum_j w_j = 1$. $w_j > 0$ is a long position and $w_j < 0$ a short position in asset j. If a portfolio has many short positions, the gross–exposure $\|\boldsymbol{w}\|_1$ may be large. It is now well understood that portfolios should have bounded gross–exposure $\|\boldsymbol{w}\|_1$. However, the celebrated mean–variance Markowitz theory, see Markowitz (1952, 1959) and Sharpe (1964), which establishes the well known relationship between investment risk measured by the variance and mean return, does not guarantee bounded gross-exposure. It selects $\boldsymbol{w}$ by minimizing the variance $\mathrm{Var}(\boldsymbol{w}^\top \boldsymbol{Y})$ of the portfolio return under the constraint $\boldsymbol{w}^\top \boldsymbol{\mu} = \mu_0$ of a target portfolio return μ_0, where $\boldsymbol{Y}$ is a d–vector of asset returns with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} > 0$. The optimal solution, $\boldsymbol{w}^*$, is the unique portfolio realizing the target return μ_0 and having the smallest variance possible, typically with short positions. It is a linear combination of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}\mathbf{1}$, so that the results of our Section 4 can be applied.

The oracle risk, $R(\boldsymbol{w}^*)$, the empirical risk, $R_m(\boldsymbol{w}^*)$ and the associated estimation error $|R_m(\boldsymbol{w}^*) - R(\boldsymbol{w}^*)|$, where $R(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}$ and $R_m(\boldsymbol{x}) = \boldsymbol{x}^\top \hat{\boldsymbol{\Sigma}}_m \boldsymbol{x}$ with an arbitrary estimator $\hat{\boldsymbol{\Sigma}}_m$ of $\boldsymbol{\Sigma}$, are, however, not controlled by the Markowitz solution. For large dimension the error of the risk estimate may suffer from substantial cumulation of estimation errors, especially due to the need to estimate $d(d-1)/2$ covariances. One popular and theoretically sound approach to mitigate this problem is to minimize the Markowitz mean–variance function $M = \boldsymbol{w}^\top \boldsymbol{\mu} + \lambda \boldsymbol{w}^\top \boldsymbol{\Sigma}\boldsymbol{w}$, for some risk–aversion parameter $\lambda > 0$, under a gross–exposure constraint $\|\boldsymbol{w}\|_1 \leq c$. Then, for any portfolio $\boldsymbol{w}$, the sensitivity of M can be bounded by $\|\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|_\infty \|\boldsymbol{w}\|_1 + \lambda \|\hat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}\|_\infty \|\boldsymbol{w}\|_1^2$, and this bound holds for arbitrary estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. It follows that if one can estimate all entries of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ equally well, as guaranteed by our Corollary 1, the sensitivity is controlled by the maximum estimation error and the gross–error exposure $\|\boldsymbol{w}\|_1$. This

phenomenon has been observed in the finance literature by Jagannathan and Ma (2003) for the choice $c = 1$ yielding the optimal no–short–sales (i.e., long–only) portfolio and has been recently discussed in greater detail by Fan et al. (2012). Further, constraining $\|\boldsymbol{w}\|_1$ is equivalent to shrinking of the estimated covariance matrix, see Fan et al. (2012) and Zhao et al. (2021) for a recent discussion and further references.

Even if $\boldsymbol{w}$ is determined by another approach, for example by tracking an index by a sparse portfolio leading to a sparse portfolio vector $\boldsymbol{w}$, i.e., with $\|\boldsymbol{w}\|_1 \leq c$ or even with finite $\|\cdot\|_0$–norm, as done by many exchange traded funds (ETFs), the portfolio variance $\boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w}$ is a crucial risk functional, and one is faced with the following problem: If the portfolio risk changes, the weights should be updated. For a large universe of assets, however, the frequent precise (re–) calculations of $\boldsymbol{w}$ is a challenging and costly issue from a numerical and statistical viewpoint as well as from a practical viewpoint, since this may lead to massive trading costs and can even be infeasible for highly captitalize funds, which therefore tend to update weights only twice a year. If the ETF implements a certain strategy, e.g. momentum, more frequent updates triggered by a suitable monitoring procedure could, however, lead to more profitable short–term positions.

Given a sequence of returns $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ denote by $\hat{\boldsymbol{v}}_m$ an estimator of an unknown true (preferably bounded–gross–exposure) portfolio $\boldsymbol{v}$, estimated from the learning sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$. Recall that the risk of the unknown portfolio $\boldsymbol{v}$ is given by $R(\boldsymbol{v}) = \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}$. As long as the distribution of the returns changes without altering the true risk $R(\boldsymbol{v})$, there is no need to update the portfolio, exactly the changes affecting $R(\boldsymbol{v})$ are of interest. Under Assumption A–iii and the common zero mean assumption for financial returns under normal market conditions, unbiased estimates of $R(\boldsymbol{v})$ are given by $\boldsymbol{v}^\top \boldsymbol{Y}_i \boldsymbol{Y}_i^\top \boldsymbol{v}$, $1 \leq i \leq m$. A change of the expectation of $\boldsymbol{v}^\top \boldsymbol{Y}_{m+k} \boldsymbol{Y}_{m+k}^\top \boldsymbol{v}$, $k \geq 1$, indicates the necessity to update the porfolio. This motivates to apply the proposed detector, D, which monitors the sequence $\hat{\boldsymbol{v}}_m^\top \boldsymbol{Y}_{m+k} \boldsymbol{Y}_{m+k}^\top \hat{\boldsymbol{v}}_m$, $k \geq 1$.

5.3 Experimental support and simulations

Experiment: To check whether the proposed approach works we applied it to synthetic data following the model

$$Y_t = \begin{cases} \cos(10\pi x_t) + e_t, & 1 \leq t < 2,000, \\ \cos(10\pi x_t^*) + e_t, & 2,000 \leq t <, 4000, \\ \cos(4\pi x_t^*) + e_t^*, & 4,000 \leq t, \end{cases}$$

where $x_t \sim U(-1/2, 1/2)$, $x_t^* \sim U(0,1)$, $e_t \sim N(0, 0.1)$ and $e_t^* \sim N(0, 0.05)$. At $k_1^* = 2,000$ the distribution of the regressor changes and at $k_2^* = 4,000$ the regression coefficient changes. Two independent uniformly distributed regressors were added to mimic the situation that the regression function depends only on a subset of the regressors. For this data a neural network needs to be trained three times: Once at the beginning and then after the two change–points.

A neural network with three fully–connected hidden layers consisting of 4, 2, and 1 neurons, respectively, relu activations for the hidden layers and a linear output layer was trained using the first $m = 1,000$ observations using ADAM for 100 epochs with

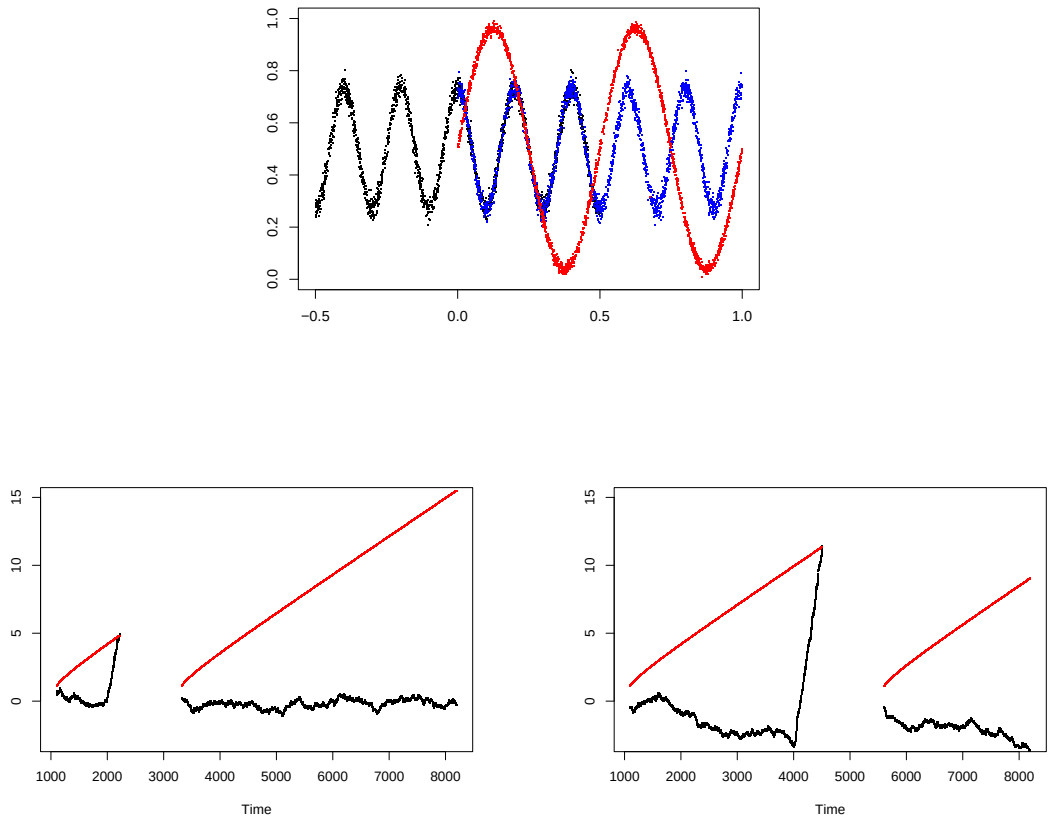

Figure 1: Experiments: The top panel shows the synthetic data set. Before the first change data points are black, after the first and before the second one blue, and data after both changes are in red. The left (right) panel depicts the detector statistic $D_{proj,k}$ ($D_{res,k}$) and the boundary B_k. The method reacts quickly and there are no false alarms.

batch size 32, and a validation split of the training data of 0.2. For the simulation study below training was based on 30 epochs. The fitted network was then used for prediction and only retrained if a signal was given. The procedure was rolled over after a signal as follows: All data before the signal were discarded, and the next m observations were used as a fresh training sample from which the neural network was retrained. Then, the detector was restarted after the training sample. To avoid non–termination of the detector, a sample of length $T = 50,000$ was simulated and the stopping time was set to ∞ if there was no signal up to T. The experiment was implemented in R using tensorflow and keras on a Mac M1 Max with GPU support.

Figure 1 depicts the synthethic data (Y_t scaled to the unit interval), simulated according to the above model, and, for the first 10,000 data points, the scaled detector statistic, $D_{proj,k} = \frac{1}{\sqrt{m}}Q(m,k)$, and the associated scaled boundary, $B_k = c_\alpha g(m,k)$, where the constant c_α is selected to ensure an asymptotic false detection rate α. In additiona, a residuals detector, $D_{res,k}$, was run obtained from $D_{proj,k}$ by replacing all $\hat{\boldsymbol{\beta}}_m^\top \mathbf{S}_t \hat{\boldsymbol{\beta}}_m$ by $Y_{m+\ell} - \hat{\boldsymbol{\beta}}_m^\top \mathbf{S}_t \hat{\boldsymbol{\beta}}_m$, $t \geq 1$. This detector is designed to detect a change in the regression coefficients, e.g. when using a different neural network. For the simulated synthetic data the detector $D_{proj,k}$ gives a signal at obs. $2,117$ corresponding to a delay of 117 and no further signal. The detector $D_{resid,k}$ signals a change at obs. $4,508$ (delay: 508) and gave no further signal as well. Both detectors react quickly to the changes and do not suffer from false signals. We can summarize that for this example the proposed detection algorithm operates optimally and minimizes the required computational costs of network training by invoking it only three times.

Simulations: A simulation study was conducted to investigate the statistical behaviour of the proposed detectors in terms of the probability to get i) a signal (to reject H_0) at all and ii) to get a correct signals (i.e., after the change–point). A further characteristic of interest is the delay when a signal was given. For that purpose, both detectors were applied to samples simulated according to the above model for varying tuning parameter γ which controls the weighting function. The simulation model has one change–point of interest, k_1^*, which should be detected by the proposed prediction detector D, and one change–point for which D should be agnostic. Analyzing the simulation results reveiled that indeed the detector D only reacts for the change–point k_1^* and gives a signal at k_2^* only in rare cases. Therefore results for k_2^* are not reported.

The empirical rejection rates and average delays were simulated under the null hypothesis and under parameterized alternatives $H_1(f)$, for $f \in \{0.05, 0.1, 0.15\}$, where the concept drift change–point k_1^*, at which the regressor distribution changes from a uniform law on $[-1/2, 1/2]$ to a uniform law on $[-1/2+f, 1/2+f]$, is located at $k_1^* = (1+f)m$. That is, samples

$$Y_t = \begin{cases} \cos(10\pi x_t) + e_t, & 1 \leq t < (1+f)1,000, \\ \cos(10\pi x_t^*) + e_t, & (1+f)1,000 \leq t <, 4000, \\ \cos(4\pi x_t^*) + e_t^*, & 4,000 \leq t, \end{cases}$$

were generated, where $x_t \sim U(-1/2, 1/2)$, $x_t^* \sim U(-1/2+f, 1/2+f)$, $e_t \sim N(0, 0.1)$ and $e_t^* \sim N(0, 0.05)$. Again, two additional uniform regressors were added and fed into the neural network. At $k_1^* = 2,000$ the distribution of the regressor changes and at $k_2^* = 4,000$ the regression coefficient changes.

The results are provided in Table 1. All methods perform quite well in terms of the type I error rate, but the Robbins–Siegmund detector turned out to result in quite conservative type I error rates, which leads to a loss of power to detect alternatives. This observation is in agreement with simulation results obtained in Chu et al. (1996), althought the boundary was used there in a quite different setting. To allow better comparisons, the value a was selected to match the type I error rate of the γ–detector, studied for and $\gamma \in \{0.25, 0.45, 0.49\}$, when $\gamma = 0.49$. In terms of the power to detect

γ	Level	f	Power	Correct Signal	Delay
Erdös–Kac detector					
	7.6	0.05	54.7	1	1041
		0.10	75.2	1	662
		0.15	82.6	1	626
γ–detector					
0.25	10.1	0.05	58.3	1	995
		0.10	76.0	1	594
		0.15	82	99.9	463
0.45	9.7	0.05	53.5	1	809
		0.10	75.4	99.3	464
		0.15	82	98.3	394
0.49	8.7	0.05	63.6	1	609
		0.10	76.5	98.9	514
		0.15	80.6	98.3	385
Robbins-Siegmund detector					
	8.7	0.05	56.5	1	937
		0.10	74.8	1	688
		0.15	83.5	1	559

Table 1: Simulated level and power under alternative $H_1(f)$, the conditional probabilties that a signal is correct, i.e., given after the change k_1^*, and the mean delay.

the concept drifts parameterized by f, the results also indicate that for larger changes the methods are quite similar. But for small changes the γ–detector with $\gamma = 0.49$ (i.e., close to $1/2$) outperforms the competitors. Except rare cases, the signals are correct, i.e., after the change–point. However, in terms of the delay the procedures differ. Both the Erdös–Kac detector and the Robbins–Siegmund detector have much larger delays than the γ–detector, and this holds even for large changes where the detection rate is around 80%. One can summarize the findings as follows: The Robbins–Siegmund detector is easy to apply but operates on a smaller–than–nominal type I error rate when using the theoretical value for a. The γ–detector with γ close to $1/2$ has high power and leads to quick signals compared to its competitors.

6. Proofs

Proof of Proposition 1: Assumption (8) yields the existence of a probability measure $Q = Q_{t_1,\ldots,t_k}$ on $\mathbb{R}^{kd_m}$, not necessarily unique, such that $\int x_{11}^{n_{11}} \cdot \ldots \cdot x_{kd_m}^{n_{kd_m}} \, dQ = \mu_{\boldsymbol{t},\boldsymbol{n}}$, for any $\boldsymbol{t} = (t_1, \ldots, t_k) \in \mathbb{N}^k$ with pairwise different entries and all $\boldsymbol{n} \in \mathbb{N}_0^{kd_m}$. By Kolmogorov's extension theorem, one can construct on a new probability space a stochastic process $\tilde{\boldsymbol{X}}_t$, $t \geq 1$, whose finite dimensional distributions with respect to $t_1, \ldots, t_k$, are given by $Q_{t_1,\ldots,t_k}$, for all time points $1 \leq t_1 < \cdots < t_k$ and all $k \in \mathbb{N}$. The explicit

construction of an equivalent process on the original probability space is as follows. Suppose we already have constructed $\sigma = (\boldsymbol{X}_1, \dots \boldsymbol{X}_{t-1})$ with $\boldsymbol{X}_i = \boldsymbol{G}_i(\boldsymbol{\epsilon}_i)$, $1 \le i \le t-1$. Apply (Billingsley, 1999, Lem. 1, p.212) with $\nu = Q_{1,\dots,t}$, $\mu = \mathcal{L}(\sigma) = Q_{1,\dots,t-1}$ and $U = \epsilon_t \sim U(0,1)$ to obain a random variable τ defined on $(\Omega, \mathcal{F}, \mathbb{P})$, which is a function of $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_{t-1})$ and ϵ_t, such that $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_{t-1}, \tau) \sim \nu$. Thus, we can take $\boldsymbol{X}_t = \tau(\boldsymbol{X}_1(\boldsymbol{\epsilon}_1), \dots, \boldsymbol{X}_{t-1}(\boldsymbol{\epsilon}_{t-1}), \epsilon_t) =: \boldsymbol{G}_t(\boldsymbol{\epsilon}_t)$. It remains to construct $\boldsymbol{X}_1 = (X_{11}, \dots, X_{1d_m})$. Use ϵ_1 as a seed to define i.i.d. uniform random variables $\epsilon_{11}, \dots, \epsilon_{1d_m}$ (which are functions of ϵ_1). Put $X_{11} = X_{11}(\epsilon_{11}) := F^{-1}_{Q_{11}}(\epsilon_{11})$, where $F_{Q_{11}}(x) = F_{Q_1}(x, \infty, \dots, \infty)$ is the 1st marginal d.f. of Q_1. Next, apply iteratively (Billingsley, 1999, Lem. 1, p.212): Having already constructed $\sigma = (X_{11}, \dots, X_{1,i-1})$, put $\nu = Q_{11,\dots,1i}$ (the marginal distribution of Q_1 with respect to the first i coordinates) and $U = \epsilon_{1i}$, to obtain τ, a function of $(X_{11}, \dots, X_{1,i-1})$ and ϵ_{1i} with $(\sigma, \tau) = (X_{11}, \dots, X_{1,i-1}, \tau) \sim \nu = Q_{11,\dots,qi}$. Thus, put $X_{1i} = \tau(X_{11}(\epsilon_{i1}), \dots, X_{1,i-1}(\epsilon_{1,i-1}), \epsilon_{1i})$. This gives us $\boldsymbol{X}_1 = (X_{11}, \dots, X_{1d_m})$, a function $\boldsymbol{G}_1(\boldsymbol{\epsilon}_1)$ of $\boldsymbol{\epsilon}_1 = (\epsilon_1, \epsilon_0, \dots)$. □

The proofs of Theorems 2 and 3 rely on several auxiliary results.

Lemma 1. *Suppose that ξ'_i, $i \ge 1$, are mean zero and unit variance random variables defined on a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ on which there exists a standard Brownian motion B'_1, such that with $S'_n = \sum_{i=1}^n \xi'_i$ the strong approximation*

$$|S'_n - B'_1(n)| = o(n^{1/\nu}), \qquad a.s., \tag{10}$$

holds for some $\nu > 2$. Define the standard Brownian motion $B'_{2,m}(t) = B'_1(m+t) - B'_1(m)$, $t \ge 0$, which is independent of $B'_1(t)$, $0 \le t \le m$. Then for any $\delta > 0$

$$\sup_{k \ge \delta m} k^{-1/\nu} |S'_{m+k} - S'_m - B'_{2,m}(k)| = O_{\mathbb{P}}(1), \qquad m \to \infty.$$

Proof of Lemma 1: Observe that

$$\begin{aligned} k^{-\frac{1}{\nu}} |S'_{m+k} - S'_m - B'_{2,m}(k)| \le & \left(\frac{m+k}{k}\right)^{\frac{1}{\nu}} \frac{1}{(m+k)^{\frac{1}{\nu}}} |S'_{m+k} - B'_1(m+k)| \\ & + \left(\frac{m}{k}\right)^{\frac{1}{\nu}} \frac{1}{m^{\frac{1}{\nu}}} |S'_m - B'_1(m)|. \end{aligned}$$

The first term can be bounded by

$$\frac{1+\delta}{\delta} \sup_{k \ge \delta m} (m+k)^{-\frac{1}{\nu}} |S'_{m+k} - B'_1(m+k)| \le \frac{1+\delta}{\delta} \sup_{N \ge 1} N^{-\frac{1}{\nu}} |S'_N - B'_1(N)|. \tag{11}$$

By assumption, there exists an event A with $P(A) = 1$ such that for each $\omega \in A$ the random variable $Z_N(\omega) = N^{-\frac{1}{\nu}} |S'_N - B'_1(N)|(\omega)$ is $o(1)$ and hence bounded. Thus, $C(\omega) = \sup_{N \ge 1} Z_N(\omega) < \infty$. We obtain with probability one

$$\sup_{k \ge \delta m} \left(\frac{m+k}{k}\right)^{\frac{1}{\nu}} \frac{1}{(m+k)^{\frac{1}{\nu}}} |S'_{m+k} - B'_1(m+k)| \le \frac{1+\delta}{\delta} C$$

Since the random variable $\frac{1+\delta}{\delta}C$ is stochastically bounded, the first term is $O_{\mathbb{P}}(1)$. The second term can be bounded by

$$\sup_{k\geq\delta m}\left(\frac{m}{k}\right)^{\frac{1}{\nu}}\frac{1}{m^{\frac{1}{\nu}}}|S'_m-B'_1(m)|\leq\frac{1}{\delta^{\frac{1}{\nu}}}\frac{1}{m^{\frac{1}{\nu}}}|S'_m-B'_1(m)|=o_{\mathbb{P}}(1).$$

This proves the assertion. □

The following lemma checks for several series derived from $\boldsymbol{Y}_t$ that they satisfy Assumptions A-i and A-ii with $V(2)$ and determines appropriate powers and constants.

Lemma 2. *Suppose that Assumptions A–i and A–ii hold.*

(i) The time series $\boldsymbol{v}^\top\mathbf{S}_i\boldsymbol{v}-\boldsymbol{v}^\top\mathbf{M}_i\boldsymbol{v}$, $i\geq 1$, satisfies Assumption A–i with power $q/2$ instead of q, $4K^2\|\boldsymbol{v}\|_1^2$ instead of K, and $4K\|\boldsymbol{v}\|_1^2C_1$ instead of C_1. A–ii holds with $4K^2\|\boldsymbol{v}\|_1^2L$ instead of KL for $V(2)$.

(ii) The vector time series $\boldsymbol{v}^\top\boldsymbol{Y}_t\boldsymbol{Y}_t^\top$, $t\geq 1$, satisfies Assumption A–i with power $q/2$ instead of q, $2K^2\|\boldsymbol{v}\|_1$ instead of K, and $2K\|\boldsymbol{v}\|_1(\|\boldsymbol{v}\|_1+1)C_1\sqrt{d}$ instead of C_1. Especially,

$$\left(\mathbb{E}\left\|\boldsymbol{G}_t(\boldsymbol{\epsilon}_t)\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_t)-\boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})\right\|_2^{\frac{q}{2}}\right)^{\frac{2}{q}}\leq 2K\|\boldsymbol{v}\|_1(\|\boldsymbol{v}\|_1+1)C_1\sqrt{d}j^{-\beta}. \tag{12}$$

Assumption A-ii holds for $V(2)$ with $4\|\boldsymbol{v}\|_1^2K^2L$ instead of KL.

(iii) $(\mathbb{E}\|\boldsymbol{Y}_t\|_2^q)^{\frac{1}{q}}\leq K^{\frac{1}{q}}\sqrt{d}$.

(iv) $(\mathbb{E}\|\boldsymbol{G}_t(\boldsymbol{\epsilon}_t)-\boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})\|_2^q)^{\frac{1}{q}}\leq C_1^{\frac{1}{q}}\sqrt{d}j^{-\beta}$, $j\geq 0$.

(v) The (vectorized) series $\boldsymbol{Y}_t\boldsymbol{Y}_t^\top$ satisfies Assumption A–i with power $q/2$ and constants Kd and C_1d.

Proof Lemma 2: As a preparation, let $X,Y\in L_q$ be random variables with expectations μ_X and μ_Y. The decomposition

$$XY=(X-\mu_X)(Y-\mu_Y)+(X-\mu_X)\mu_Y+\mu_X(Y-\mu_Y)+\mu_X\mu_Y$$

implies the bound $\|XY\|_{L_r}\leq\|X-\mathbb{E}X\|_{L_q}\|Y-\mathbb{E}Y\|_{L_q}+|\mathbb{E}Y|\|X-\mathbb{E}X\|_{L_q}+|\mathbb{E}X|\|Y-\mathbb{E}Y\|_{L_q}+|\mathbb{E}X||\mathbb{E}Y|$ for $1\leq r\leq q$. Therefore, $|\mathbb{E}Y_{tj}Y_{tk}|\leq 4K^2$ under Assumption A-i, such that $\|\mathbf{M}_t\|_\infty\leq 4K^2$, since $q>4$. Further $\|\boldsymbol{\Sigma}_t\|_\infty=\max_{1\leq j,k\leq d}|\mathbb{E}(Y_{tj}-\mu_{tj})(Y_{tk}-\mu_{tk})|\leq K^2$, and $\max_{1\leq j\leq d}\|Y_{tj}\|_{L_r}\leq\max_{1\leq j\leq d}\|Y_{tj}-\mu_{tj}\|_{L_r}+K\leq 2K$, for any $2\leq r\leq q$.

Let us first show (ii). Rearranging the above decomposition and denoting $\Sigma_{XY}=\mathrm{Cov}(X,Y)$ and $M_{XY}=\mathbb{E}(XY)=\Sigma_{XY}+\mu_X\mu_Y$, one has the representation

$$\overline{XY}=XY-M_{XY}=(X-\mu_X)(Y-\mu_Y)+(X-\mu_X)\mu_Y+\mu_X(Y-\mu_Y)-\Sigma_{XY}.$$

Thus,

$$\|\overline{XY}\|_{L_{q/2}} \le \|X-\mu_X\|_{L_q}\|Y-\mu_Y\|_{L_q} + |\mu_Y|\|X-\mu_X\|_{L_{q/2}} + |\mu_X|\|Y-\mu_Y\|_{L_{q/2}} + |\Sigma_{XY}|.$$

Applying this inequality with $X = Y_{tk}$ and $Y = \boldsymbol{Y}_t^\top \boldsymbol{v}$ gives for the kth coordinate $Y_{tk}\boldsymbol{Y}_t^\top \boldsymbol{v}$ of $\boldsymbol{Y}_t\boldsymbol{Y}_t^\top \boldsymbol{v}$, $1 \le k \le d$,

$$\begin{aligned}\|Y_{tk}\boldsymbol{Y}_t^\top \boldsymbol{v} - \mathbb{E}(Y_{tk}\boldsymbol{Y}_t^\top \boldsymbol{v})\|_{L_{q/2}} &\le \|Y_{tk}-\mu_{tk}\|_{L_q}\|(\boldsymbol{Y}_t-\boldsymbol{\mu}_t)^\top \boldsymbol{v}\|_{L_q} + \|\boldsymbol{\mu}_t\|_\infty\|\boldsymbol{v}\|_1\|Y_{tk}-\mu_{tk}\|_{L_{q/2}} \\ &\quad + \|\boldsymbol{\mu}_t\|_\infty\|\boldsymbol{v}^\top(\boldsymbol{Y}_t-\boldsymbol{\mu}_t)\|_{L_{q/2}} + |\operatorname{Cov}(Y_{tk}, \boldsymbol{v}^\top \boldsymbol{Y}_t)|.\end{aligned}$$

To estimate the terms on the right side, notice that for any $2 \le r \le q$

$$\left\|(\boldsymbol{Y}_t-\boldsymbol{\mu}_t)^\top \boldsymbol{v}\right\|_{L_r} \le \sum_{\ell=1}^d |v_\ell|\|Y_{t\ell}-\mu_{t\ell}\|_{L_r} \le K\|\boldsymbol{v}\|_1$$

and

$$|\operatorname{Cov}(Y_{tk}, \boldsymbol{v}^\top \boldsymbol{Y}_t)| \le \sqrt{\mathbb{E}(Y_{tk}-\mu_{tk})^2}\sqrt{\mathbb{E}\left(\boldsymbol{v}^\top(\boldsymbol{Y}_t-\boldsymbol{\mu}_t)\right)^2} \le K^2\|\boldsymbol{v}\|_1.$$

Therefore,

$$\|Y_{tk}\boldsymbol{Y}_t^\top \boldsymbol{v} - \mathbb{E}(Y_{tk}\boldsymbol{Y}_t^\top \boldsymbol{v})\|_{L_{q/2}} \le 4K^2\|\boldsymbol{v}\|_1,$$

Observe that by the triangle inequality and Jensen's inequality, for any $1 \le r \le q$,

$$\begin{aligned}\mathbb{E}|\boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_t)|^r &\le \|\boldsymbol{v}\|_1^r \mathbb{E}\left|\sum_{j=1}^d \frac{|v_j|}{\|\boldsymbol{v}\|_1}|G_{tj}(\boldsymbol{\epsilon}_t)|\right|^r \\ &\le \|\boldsymbol{v}\|_1^r \sum_{j=1}^d \frac{|v_j|}{\|\boldsymbol{v}\|_1}\mathbb{E}|G_{tj}(\boldsymbol{\epsilon}_t)|^r \\ &\le \|\boldsymbol{v}\|_1^r \max_{1\le j\le d}\mathbb{E}|G_{tj}(\boldsymbol{\epsilon}_t)|^r,\end{aligned}$$

so that

$$\|\boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_t)\|_{L_r} \le \|\boldsymbol{v}\|_1 \max_{1\le j\le d}\|G_{tj}(\boldsymbol{\epsilon}_t)\|_{L_r} \le 2K\|\boldsymbol{v}\|_1. \tag{13}$$

Using this fact, the lag j (coordinate–wise) physical weak dependence measure can now be estimated by

$$\begin{aligned}&\left(\mathbb{E}|G_{tk}(\boldsymbol{\epsilon}_t)\boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - G_{tk}(\boldsymbol{\epsilon}_{t,t-j})\boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})|^{\frac{q}{2}}\right)^{\frac{2}{q}} \\ &\quad\le \left(\mathbb{E}|(G_{tk}(\boldsymbol{\epsilon}_t) - G_{tk}(\boldsymbol{\epsilon}_{t,t-j}))\boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_t)|^{\frac{q}{2}}\right)^{\frac{2}{q}} + \left(\mathbb{E}|G_{tk}(\boldsymbol{\epsilon}_{t,t-j})(\boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - \boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j}))|^{\frac{q}{2}}\right)^{\frac{2}{q}} \\ &\quad\le \left(\mathbb{E}|\boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_t)|^q\right)^{\frac{1}{q}}\left(\mathbb{E}|G_{tk}(\boldsymbol{\epsilon}_t) - G_{tk}(\boldsymbol{\epsilon}_{t,t-j})|^q\right)^{\frac{1}{q}} \\ &\qquad + \left(\mathbb{E}|G_{tk}(\boldsymbol{\epsilon}_{t,t-j})|^q\right)^{\frac{1}{q}}\left(\mathbb{E}|\boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - \boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})|^q\right)^{\frac{1}{q}}\end{aligned}$$

$$\leq 2K\|\boldsymbol{v}\|_1 C_1 j^{-\beta} + 2K\|\boldsymbol{v}\|_1^2 \max_{1\leq k\leq d} (\mathbb{E}|G_{tk}(\boldsymbol{\epsilon}_t) - G_{tk}(\boldsymbol{\epsilon}_{t,t-j})|^q)^{\frac{1}{q}}$$
$$\leq 2K\|\boldsymbol{v}\|_1(\|\boldsymbol{v}\|_1 + 1)C_1 j^{-\beta}.$$

This also yields for the associated vector version

$$\left(\mathbb{E}\left\|\boldsymbol{G}_t(\boldsymbol{\epsilon}_t)\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})\right\|_2^{\frac{q}{2}}\right)^{\frac{2}{q}}$$
$$= \left(\mathbb{E}\left(\sum_{k=1}^d (G_{tk}(\boldsymbol{\epsilon}_t)\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - G_{tk}(\boldsymbol{\epsilon}_{t,t-j})\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j}))^2\right)^{\frac{q}{4}}\right)^{\frac{2}{q}}$$
$$\leq \left(d^{\frac{q}{4}-1}\sum_{k=1}^d \mathbb{E}(G_{tk}(\boldsymbol{\epsilon}_t)\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - G_{tk}(\boldsymbol{\epsilon}_{t,t-j})\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j}))^{\frac{q}{2}}\right)^{\frac{2}{q}}$$
$$\leq 2K\|\boldsymbol{v}\|_1(\|\boldsymbol{v}\|_1 + 1)C_1\sqrt{d}j^{-\beta},$$

verifying (12). Assumption A–ii can be shown as follows. We have

$$\sum_{t=1}^n \|\boldsymbol{v}^\top(\boldsymbol{G}_t(\boldsymbol{\epsilon}_0) - \boldsymbol{G}_{t-1}(\boldsymbol{\epsilon}_0))\|_{L_4} \leq \sum_{j=1}^d |v_j| \sum_{t=1}^n \|G_{tj}(\boldsymbol{\epsilon}_0) - G_{t-1,j}(\boldsymbol{\epsilon}_0)\|_{L_4}$$
$$\leq \|\boldsymbol{v}\|_1 KL.$$

This gives, using $\|\boldsymbol{v}^\top\boldsymbol{Y}_t\|_{L_4} \leq 2K\|\boldsymbol{v}\|_1$,

$$\|G_{tk}(\boldsymbol{\epsilon}_0)\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_0) - G_{t-1,k}(\boldsymbol{\epsilon}_0)\boldsymbol{v}^\top\boldsymbol{G}_{t-1}(\boldsymbol{\epsilon}_0)\|_{L_2}$$
$$\leq \|\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_0)\|_{L_4}\|G_{tk}(\boldsymbol{\epsilon}_0) - G_{t-1,k}(\boldsymbol{\epsilon}_0)\|_{L_4} + \|G_{t-1,k}(\boldsymbol{\epsilon}_0)\|_{L_4}\|\boldsymbol{v}^\top(\boldsymbol{G}_t(\boldsymbol{\epsilon}_0) - \boldsymbol{G}_{t-1}(\boldsymbol{\epsilon}_0))\|_{L_4}$$
$$\leq 2K\|\boldsymbol{v}\|_1\|G_{tk}(\boldsymbol{\epsilon}_0) - G_{t-1,k}(\boldsymbol{\epsilon}_0)\|_{L_4} + 2K\|\boldsymbol{v}^\top(\boldsymbol{G}_t(\boldsymbol{\epsilon}_0) - \boldsymbol{G}_{t-1}(\boldsymbol{\epsilon}_0))\|_{L_4},$$

and therefore

$$\sum_{t=1}^n \|G_{tk}(\boldsymbol{\epsilon}_0)\boldsymbol{v}^\top\boldsymbol{G}_t(\boldsymbol{\epsilon}_0) - G_{t-1,k}(\boldsymbol{\epsilon}_0)\boldsymbol{v}^\top\boldsymbol{G}_{t-1}(\boldsymbol{\epsilon}_0)\|_{L_2} \leq 4\|\boldsymbol{v}\|_1 K^2 L.$$

This verifies A–ii for $V(2)$ with constant $4\|\boldsymbol{v}\|_1 K^2 L$ instead of KL.

To show (i) use the fact that

$$\boldsymbol{Y}_t\boldsymbol{Y}_t^\top - \mathbf{M}_t = (\boldsymbol{Y}_t - \boldsymbol{\mu}_t)(\boldsymbol{Y}_t - \boldsymbol{\mu}_t)^\top + (\boldsymbol{Y}_t - \boldsymbol{\mu}_t)\boldsymbol{\mu}_t^\top + \boldsymbol{\mu}_t(\boldsymbol{Y}_t - \boldsymbol{\mu}_t)^\top - \boldsymbol{\Sigma}_t$$

yielding

$$\|\boldsymbol{v}^\top(\boldsymbol{Y}_t\boldsymbol{Y}_t^\top - \mathbf{M}_t)\boldsymbol{v}\|_{L_{q/2}} \leq \|\boldsymbol{v}^\top(\boldsymbol{Y}_t - \boldsymbol{\mu}_t)\|_{L_q}^2 + 2|\boldsymbol{\mu}_t^\top\boldsymbol{v}|\|\boldsymbol{v}^\top(\boldsymbol{Y}_t - \boldsymbol{\mu}_t)\|_{L_q} + \max_{1\leq j,k\leq d}|\boldsymbol{\Sigma}_{t,jk}|\|\boldsymbol{v}\|_1^2$$
$$\leq \|\boldsymbol{v}\|_1^2 \max_{1\leq j\leq d}\|Y_{tj} - \mu_{tj}\|_{L_q}^2 + 2K\|\boldsymbol{v}\|_1^2 \max_{1\leq j\leq d}\|Y_{tj} - \mu_{tj}\|_{L_q} + K^2\|\boldsymbol{v}\|_1^2$$

$$\leq 4K^2\|\boldsymbol{v}\|_1^2.$$

Further, observing that $\boldsymbol{v}^\top \boldsymbol{Y}_t\boldsymbol{Y}_t^\top \boldsymbol{v} - \boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})\boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})^\top \boldsymbol{v}$ can be written as the sum of $\boldsymbol{v}^\top(\boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j}))\boldsymbol{G}_t(\boldsymbol{\epsilon}_t)^\top \boldsymbol{v}$ and $\boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})(\boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j}))^\top \boldsymbol{v}$, we obtain

$$\begin{aligned}
\|\boldsymbol{v}^\top \boldsymbol{Y}_t\boldsymbol{Y}_t^\top \boldsymbol{v} - \boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})\boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})^\top \boldsymbol{v}\|_{L_{q/2}} &\leq 2\|\boldsymbol{v}^\top(\boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j}))\boldsymbol{v}^\top \boldsymbol{G}_t(\boldsymbol{\epsilon}_t)\|_{L_{q/2}} \\
&\leq 2\|\boldsymbol{v}^\top \boldsymbol{Y}_t\|_{L_q}\|\boldsymbol{v}^\top(\boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j}))\|_{L_q} \\
&\leq 4K\|\boldsymbol{v}\|_1^2 \max_{1\leq k\leq d}\|G_{tk}(\boldsymbol{\epsilon}_t) - G_{tk}(\boldsymbol{\epsilon}_{t,t-j})\|_{L_q} \\
&\leq 4K\|\boldsymbol{v}\|_1^2 C_1 j^{-\beta},
\end{aligned}$$

which verifies A–i with constant $4K\|\boldsymbol{v}\|_1^2 C_1$. Similarly, one shows that A–ii holds for $V(2)$ with KL replaced by $4\|\boldsymbol{v}\|_1^2 K^2 L$: By symmetry,

$$\begin{aligned}
\boldsymbol{v}^\top \boldsymbol{Y}_t\boldsymbol{Y}_t^\top \boldsymbol{v} - \boldsymbol{v}^\top \boldsymbol{Y}_{t-1}\boldsymbol{Y}_{t-1}^\top \boldsymbol{v} &= \boldsymbol{v}^\top(\boldsymbol{Y}_t - \boldsymbol{Y}_{t-1})\boldsymbol{Y}_t^\top \boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{Y}_{t-1}(\boldsymbol{Y}_t - \boldsymbol{Y}_{t-1})^\top \boldsymbol{v} \\
&= 2\boldsymbol{v}^\top(\boldsymbol{Y}_t - \boldsymbol{Y}_{t-1})\boldsymbol{Y}_t^\top \boldsymbol{v},
\end{aligned}$$

such that by A–ii

$$\begin{aligned}
\sum_{t=1}^n \|\boldsymbol{v}^\top \boldsymbol{Y}_t\boldsymbol{Y}_t^\top \boldsymbol{v} - \boldsymbol{v}^\top \boldsymbol{Y}_{t-1}\boldsymbol{Y}_{t-1}^\top \boldsymbol{v}\|_{L_2} &\leq 2\|\boldsymbol{v}\|_1 \sum_{t=1}^n \max_{1\leq j\leq d}\|Y_{tj}\|_{L_4}\|\boldsymbol{v}^\top(\boldsymbol{Y}_t - \boldsymbol{Y}_{t-1})\|_{L_4} \\
&\leq 4\|\boldsymbol{v}\|_1 K \sum_{t=1}^n \|Y_{tj} - Y_{t-1,j}\|_{L_4} \\
&\leq 4\|\boldsymbol{v}\|_1^2 K^2 L.
\end{aligned}$$

Assertions (iii) and (iv) follow by Jensen's inequality,

$$E\|\boldsymbol{Y}_t\|_2^q \leq d^{\frac{q}{2}}\mathbb{E}\left(\frac{1}{d}\sum_{j=1}^d Y_{tj}^2\right)^{\frac{q}{2}} \leq d^{\frac{q}{2}} \max_{1\leq j\leq d}\mathbb{E}(Y_{tj}^q),$$

and

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{G}_t(\boldsymbol{\epsilon}_t) - \boldsymbol{G}_t(\boldsymbol{\epsilon}_{t,t-j})\|_2^q &= \mathbb{E}\left(\sum_{k=1}^d \mathbb{E}(G_{tk}(\boldsymbol{\epsilon}_t) - G_{tk}(\boldsymbol{\epsilon}_{t,t-j}))^2\right)^{\frac{q}{2}} \\
&\leq d^{\frac{q}{2}} d^{-1}\sum_{k=1}^d \mathbb{E}\,|G_{tk}(\boldsymbol{\epsilon}_t) - G_{tk}(\boldsymbol{\epsilon}_{t,t-j})|^q \\
&\leq d^{\frac{q}{2}} \max_{1\leq k\leq d}\mathbb{E}\,|G_{tk}(\boldsymbol{\epsilon}_t) - G_{tk}(\boldsymbol{\epsilon}_{t,t-j})|^q
\end{aligned}$$

Let us now verify (v): Write $\boldsymbol{H}_t(\boldsymbol{\epsilon}_t) = (H_{tk}(\boldsymbol{\epsilon}_t))_{k=1}^{\bar{d}}$ for the $\bar{d} = d^2$–dimensional vectorized series. By Jensen's inequality we may estimate the jth physical dependence

measure with respect to the vector–2 norm in L_q as follows:

$$\begin{aligned}(\mathbb{E}\|\boldsymbol{H}_t(\boldsymbol{\epsilon}_t)-\boldsymbol{H}_t(\boldsymbol{\epsilon}_{t,t-j})\|_2^q)^{\frac{1}{q}} &= \left(\bar{d}^{\frac{q}{2}}\mathbb{E}\left(\frac{1}{\bar{d}}\sum_{k=1}^{\bar{d}}[H_{tk}(\boldsymbol{\epsilon}_t)-H_{tk}(\boldsymbol{\epsilon}_{t,t-j})]^2\right)^{\frac{q}{2}}\right)^{\frac{1}{q}}\\ &\le \left(\bar{d}^{\frac{q}{2}-1}\sum_{k=1}^{\bar{d}}\mathbb{E}|H_{tk}(\boldsymbol{\epsilon}_t)-H_{tk}(\boldsymbol{\epsilon}_{t,t-j})|^q\right)^{\frac{1}{q}}\\ &\le \bar{d}^{\frac{1}{2}-\frac{1}{q}}\left(\bar{d}C_1^q j^{-\beta q}\right)^{\frac{1}{q}}\\ &= C_1 d j^{-\beta}.\end{aligned}$$

The moment bound $(\mathbb{E}\|\boldsymbol{H}_t(\boldsymbol{\epsilon}_t)\|_2^q)^{\frac{1}{q}} \le Kd$ follows analogously.

□

Proof Theorem 1: The product series $Y_{tj}Y_{tk} = G_{tj}(\boldsymbol{\epsilon}_t)G_{tk}(\boldsymbol{\epsilon}_t)$, $t \ge 1$, satisfies Assumption A–i with power $q/2$ instead of q: Firstly, $\mathbb{E}|Y_{ij}Y_{ik}|^{\frac{q}{2}} \le \|Y_{ij}\|_{L_2}\|Y_{ik}\|_{L_q} \le 4K^2$, see the proof of Lemma 2. Secondly, $\mathbb{E}(G_{tj}(\boldsymbol{\epsilon}_t)G_{tk}(\boldsymbol{\epsilon}_t)) = \mathbb{E}(G_{tj}(\boldsymbol{\epsilon}_{t,t-\ell})G_{tk}(\boldsymbol{\epsilon}_{t,t-\ell}))$, for $\ell \ge 1$, such that

$$\begin{aligned}\|G_{tj}(\boldsymbol{\epsilon}_t)G_{tk}(\boldsymbol{\epsilon}_t)-G_{tj}(\boldsymbol{\epsilon}_{t,t-\ell})G_{tk}(\boldsymbol{\epsilon}_{t,t-\ell})\|_{L_{\frac{q}{2}}} &\le \|G_{tk}(\boldsymbol{\epsilon}_0)\|_{L_q}\|G_{tj}(\boldsymbol{\epsilon}_t)-G_{tj}(\boldsymbol{\epsilon}_{t,t-\ell})\|_{L_q}\\ &\quad+\|G_{tj}(\boldsymbol{\epsilon}_0)\|_{L_q}\|G_{tk}(\boldsymbol{\epsilon}_t)-G_{tk}(\boldsymbol{\epsilon}_{t,t-\ell})\|_{L_q}\\ &\le 4K^2C_1\ell^{-\beta},\end{aligned}$$

Thus, condition (G.1) of Mies and Steland (2023) holds with $\Theta = \max(4K^2, 4KC_1)$, such that Theorem 3.1 therein yields the existence of a universal constant c_1 such that

$$\max_{1\le j,k\le d}\left(\mathbb{E}\left|\sum_{t=1}^m(Y_{tj}Y_{tk}-\mathbb{E}(Y_{tj}Y_{tk}))\right|^{\frac{q}{2}}\right)^{\frac{2}{q}} \le c_1\sqrt{m}\zeta(\beta),$$

i.e., by monotonicity of the norms, $\max_{1\le j,k\le d}\left(\mathbb{E}\left|\hat{M}_{m,jk}-\bar{M}_{m,jk}\right|^{r/2}\right)^{2/r} \le \frac{c_1}{\sqrt{m}}\zeta(\beta)$, for $2 \le r \le q$. Similarly, one shows that $\max_{1\le j\le d}(\mathbb{E}|\hat{\mu}_{mj}-\bar{\mu}_{mj}|^r)^{1/r} \le \frac{c_2}{\sqrt{m}}\zeta(\beta)$ for some constant c_2, $2 \le r \le q$. This also implies

$$\begin{aligned}\|\hat{\mu}_{mj}\hat{\mu}_{m,k}-\bar{\mu}_{mj}\bar{\mu}_{mk}\|_{L_{\frac{r}{2}}} &\le \|\hat{\mu}_{mk}\|_{L_r}\|\hat{\mu}_{mj}-\bar{\mu}_{mj}\|_{L_r}+|\mu_j|\|\hat{\mu}_{mk}-\bar{\mu}_{mk}\|_{L_r}\\ &\le \left(\frac{c_2}{\sqrt{m}}\,\zeta(\beta)+K\right)\frac{c_2}{\sqrt{m}}\zeta(\beta)+\frac{Kc_2}{\sqrt{m}}\zeta(\beta)\\ &\le \frac{c_3}{\sqrt{m}}\zeta(\beta),\end{aligned}$$

with $c_3 = (c_2\zeta(\beta)+K)c_2+Kc_2$. Noting that $\bar{\mathbf{\Sigma}}_m = \bar{\mathbf{M}}_m+\bar{\boldsymbol{\mu}}_m\bar{\boldsymbol{\mu}}_m^\top$ and $\hat{\mathbf{\Sigma}}_m = \hat{\mathbf{M}}_m+\hat{\boldsymbol{\mu}}_m\hat{\boldsymbol{\mu}}_m^\top$ we get from the triangle inequality $\max_{1\le j,k\le d}\left(\mathbb{E}|\hat{\Sigma}_{m,jk}-\bar{\Sigma}_{m,jk}|^{r/2}\right)^{2/r} \le \frac{c_4}{\sqrt{m}}\zeta(\beta)$ for some constant c_4. Now the theorem follows with $c = \max(c_1,\dots,c_4)$. □

Lemma 3. *Suppose that Assumptions A–i and A–ii as well as A–iv hold, so that particularly* $\|\boldsymbol{v}\|_1 \le C_{\boldsymbol{v}}$ *and*

$$\|\hat{\boldsymbol{v}}_m - \boldsymbol{v}\| = O_{\mathbb{P}}\left(\sqrt{\frac{r_d}{m}}\right),$$

with $\sqrt{sr_d} = O(m^\zeta)$ *for some* $0 \le \zeta < \frac{1}{2}$ *under s–sparsity, whereas* $\sqrt{dr_d} = O(m^\zeta)$ *for non–*ℓ_0*–sparse estimation. Then*

$$R_m = \max_{k\le m}\left|\sum_{i\le k}\frac{\hat{\boldsymbol{v}}_m^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)\hat{\boldsymbol{v}}_m}{\sigma_{0,\infty}} - \sum_{i\le k}\frac{\boldsymbol{v}^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)\boldsymbol{v}}{\sigma_{0,\infty}}\right| = O_{\mathbb{P}}(\sqrt{dr_d}),$$

and under s–sparseness

$$R_m = \max_{k\le m}\left|\sum_{i\le k}\frac{\hat{\boldsymbol{v}}_m^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)\hat{\boldsymbol{v}}_m}{\sigma_{0,\infty}} - \sum_{i\le k}\frac{\boldsymbol{v}^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)\boldsymbol{v}}{\sigma_{0,\infty}}\right| = O_{\mathbb{P}}(\sqrt{sr_d}).$$

Proof of Lemma 3: To simplify presentation assume $\sigma_{0,\infty}^2 = 1$. We have the decomposition, by symmetry of $\boldsymbol{Y}_i\boldsymbol{Y}_i^\top - \mathbf{M}_i$,

$$R_m \le \max_{k\le m}\left|\sum_{i\le k}(\hat{\boldsymbol{v}}_m-\boldsymbol{v})^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)(\hat{\boldsymbol{v}}_m-\boldsymbol{v})\right| + 2\max_{k\le m}\left|\sum_{i\le k}\boldsymbol{v}^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)(\hat{\boldsymbol{v}}_m-\boldsymbol{v})\right|.$$

We shall estimate the terms on the right hand side separately. By Lemma 2 (ii) the series $\boldsymbol{Y}_i\boldsymbol{Y}_i^\top\boldsymbol{v}$ satisfies Assumptions A–i and A–ii. Consequently, since $q > 4$, $d = O(\sqrt{m})$ and the physical dependence measure of $\boldsymbol{Y}_i\boldsymbol{Y}_i^\top\boldsymbol{v}$ is $O(\sqrt{d})$, see (12), (Mies and Steland, 2023, Th. 3.2) with $r = 2$ yields

$$\mathbb{E}\max_{k\le m}\left\|\sum_{i\le k}\boldsymbol{v}^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)\right\|_2^2 \le c_1 dm$$

for some constant c_1. By Assumption A–iv, for any given $\varepsilon > 0$ we can find a constant $C > 0$ such that

$$\mathbb{P}\left(\|\hat{\boldsymbol{v}}_m-\boldsymbol{v}\|_2 > C\sqrt{\frac{r_d}{m}}\right) < \frac{\varepsilon}{2}. \tag{14}$$

Using the elementary inequality $|\boldsymbol{x}^\top\mathbf{A}\boldsymbol{x}| \le \|\mathbf{A}\boldsymbol{x}\|_2\|\boldsymbol{x}\|_2$, $\mathbf{A}\in\mathbb{R}^{d\times d}$, $\boldsymbol{x}\in\mathbb{R}^d$, we obtain for the second term of the above decomposition, for any $B > 0$,

$$\mathbb{P}\left(\frac{2}{\sqrt{dr_d}}\max_{k\le m}\left|\sum_{i\le k}\boldsymbol{v}^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)(\hat{\boldsymbol{v}}_m-\boldsymbol{v})\right| > B\right)$$

$$\leq \mathbb{P}\left(2C\sqrt{\frac{1}{dm}}\max_{k\leq m}\left\|\sum_{i\leq k}\boldsymbol{v}^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)\right\|_2>B\right)+\mathbb{P}\left(\|\hat{\boldsymbol{v}}_m-\boldsymbol{v}\|_2>C\sqrt{\frac{r_d}{m}}\right)$$
$$\leq \frac{4C^2\mathbb{E}\max_{k\leq m}\left\|\sum_{i\leq k}\boldsymbol{v}^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)\right\|_2^2}{B^2dm}+\frac{\varepsilon}{2}$$

Now choose B large enough to ensure that the last bound is less than ε. Therefore,

$$\max_{k\leq m}\left\|2\boldsymbol{v}^\top\sum_{i\leq k}(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)(\hat{\boldsymbol{v}}_m-\boldsymbol{v})\right\|_2=O_{\mathbb{P}}(\sqrt{dr_d}).$$

Further, by Lemma 2 (ii) and (Mies and Steland, 2023, Th. 3.2), for some constant c_2,

$$\mathbb{E}\max_{k\leq m}\left\|\sum_{i\leq k}\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i\right\|_2^2\leq c_2d^2m,$$

since $d=O(\sqrt{m})$ by assumption. Hence,

$$\mathbb{P}\left(\frac{\sqrt{m}}{dr_d}\max_{k\leq m}\left|\sum_{i\leq k}(\hat{\boldsymbol{v}}_m-\boldsymbol{v})^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)(\hat{\boldsymbol{v}}_m-\boldsymbol{v})\right|>B\right)$$
$$\leq \mathbb{P}\left(\frac{C^2}{d\sqrt{m}}\max_{k\leq m}\left\|\sum_{i\leq k}\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i\right\|_2>B\right)+\frac{\varepsilon}{2}$$
$$\leq \frac{4C^2\mathbb{E}\max_{k\leq m}\left\|\sum_{i\leq k}\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i\right\|_2^2}{Bd^2m}+\frac{\varepsilon}{2}$$

By increasing B, if necessary, this upper bound is less than ε. Thus,

$$\max_{k\leq m}\left|\sum_{i\leq k}(\hat{\boldsymbol{v}}_m-\boldsymbol{v})^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top-\mathbf{M}_i)(\hat{\boldsymbol{v}}_m-\boldsymbol{v})\right|=O_{\mathbb{P}}\left(\frac{dr_d}{\sqrt{m}}\right)=O_{\mathbb{P}}(\sqrt{dr_d}),$$

since $\sqrt{dr_d}=o(\sqrt{m})$ by assumption. Let us now consider the ℓ_0–sparse setting. Denote $A=\{j:v_j\neq 0\}$, $\hat{A}=\{j:\hat{v}_j\neq 0\}$ and let $\boldsymbol{Y}_{iA}=(Y_{ij})_{j\in A},\boldsymbol{v}_A=(v_j)_{j\in A}$ as well as $\hat{\boldsymbol{v}}_A=(\hat{v}_j)_{j\in A}$. Notice that on the event $\{\hat{A}=A\}$ quadratic forms in the remainder R_m as well as in the terms of the upper bound we need to estimate now collapse to quadratic forms in terms of the s–dimensional (random) vectors $\hat{\boldsymbol{v}}_A$, $\boldsymbol{v}_A$ and $\boldsymbol{Y}_{iA}$. The arguments given for the non–ℓ_0–sparse case apply with $d=s$. For example, now we have

$$\mathbb{E}\max_{k\leq m}\left\|\sum_{i\leq k}\boldsymbol{Y}_{iA}\boldsymbol{Y}_{iA}^\top-\mathbf{M}_{iA}\right\|_2^2\leq c_2s^2m,$$

since $s = O(\sqrt{m})$ by assumption. On the event $\{\hat{A} = A\}$ it holds $(\hat{\boldsymbol{v}}_m - \boldsymbol{v})^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top - \mathbf{M}_i)(\hat{\boldsymbol{v}}_m - \boldsymbol{v}) = (\hat{\boldsymbol{v}}_A - \boldsymbol{v}_A)^\top(\boldsymbol{Y}_{iA}\boldsymbol{Y}_{iA}^\top - \mathbf{M}_i)(\hat{\boldsymbol{v}}_A - \boldsymbol{v}_A)$, and therefore

$$\begin{aligned}
&\mathbb{P}\left(\frac{\sqrt{m}}{sr_d}\max_{k\leq m}\left|\sum_{i\leq k}(\hat{\boldsymbol{v}}_m - \boldsymbol{v})^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top - \mathbf{M}_i)(\hat{\boldsymbol{v}}_m - \boldsymbol{v})\right| > B\right) \\
&\leq \mathbb{P}\left(\frac{\sqrt{m}}{sr_d}\max_{k\leq m}\left|\sum_{i\leq k}(\hat{\boldsymbol{v}}_A - \boldsymbol{v}_A)^\top(\boldsymbol{Y}_{iA}\boldsymbol{Y}_{iA}^\top - \mathbf{M}_{iA})(\hat{\boldsymbol{v}}_A - \boldsymbol{v}_A)\right| > B\right) + \mathbb{P}(A \neq \hat{A})
\end{aligned}$$

yielding

$$\max_{k\leq m}\left|\sum_{i\leq k}(\hat{\boldsymbol{v}}_m - \boldsymbol{v})^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top - \mathbf{M}_i)(\hat{\boldsymbol{v}}_m - \boldsymbol{v})\right| = O_{\mathbb{P}}\left(\frac{sr_d}{\sqrt{m}}\right) = O_{\mathbb{P}}(\sqrt{sr_d}),$$

if $\sqrt{sr_d} = o(\sqrt{m})$. The first term is treated analogously. This completes the proof. □

Lemma 4. *Under Assumptions A–i and A–ii it holds for all $1 \leq j, k \leq d$, $t, s \geq 1$ and all d, m*

$$\operatorname{Cov}(G_{tj}(\boldsymbol{\epsilon}_t), G_{sk}(\boldsymbol{\epsilon}_{s+h})) \leq c\zeta(\beta)$$

for some constant c. Especially, if $\beta > 2$ then $\sum_{h\in Z}|\operatorname{Cov}(G_{tj}(\boldsymbol{\epsilon}_t), G_{sk}(\boldsymbol{\epsilon}_{s+h}))| < \infty$.

Proof of Lemma 4: The proof goes along the lines of the proof of (Mies and Steland, 2023, Prop. 5.4). For completeness, we provide the details. Let $\bar{\boldsymbol{\epsilon}}_{h,0} = (\epsilon_h, \ldots, \epsilon_1, \epsilon_0', \epsilon_1', \ldots)$. By independence of $G_{tj}(\boldsymbol{\epsilon}_0)$ and $G_{sk}(\bar{\boldsymbol{\epsilon}}_{h,0})$,

$$\begin{aligned}
\operatorname{Cov}(G_{tj}(\boldsymbol{\epsilon}_t), G_{sk}(\boldsymbol{\epsilon}_{s+h})) &= \operatorname{Cov}(G_{tj}(\boldsymbol{\epsilon}_0), G_{sk}(\boldsymbol{\epsilon}_h)) \\
&= \operatorname{Cov}(G_{tj}(\boldsymbol{\epsilon}_0), G_{sk}(\boldsymbol{\epsilon}_h) - G_{sk}(\bar{\boldsymbol{\epsilon}}_{h,0})) + \operatorname{Cov}(G_{tj}(\boldsymbol{\epsilon}_0), G_{sk}(\bar{\boldsymbol{\epsilon}}_{h,0})) \\
&= \operatorname{Cov}(G_{tj}(\boldsymbol{\epsilon}_0), G_{sk}(\boldsymbol{\epsilon}_h) - G_{sk}(\bar{\boldsymbol{\epsilon}}_{h,0})).
\end{aligned}$$

Using $\|G_{tj}(\boldsymbol{\epsilon}_0) - \mathbb{E}G_{tj}(\boldsymbol{\epsilon}_0)\|_{L_2} \leq K$ and $\|G_{sk}(\boldsymbol{\epsilon}_h) - G_{sk}(\bar{\boldsymbol{\epsilon}}_{0,h})\|_{L_2} = O(\sum_{\ell=h}^{\infty}\ell^{-\beta})$ (by a telescoping argument and applying Assumption A-ii to each term), the assertion follows from the Cauchy-Schwarz inequality. □

Lemma 5. *Under Assumptions A–i - A–iii it holds*

$$|\sigma_{i,\infty}^2 - \sigma_{0,\infty}^2| \leq \frac{CK^2\|\boldsymbol{v}\|_1^2}{m}\zeta(\beta - 1),$$

uniformly in $i \geq 1$, for some constant C.

Proof of Lemma 5: Observe that by Assumption A–iii there exists some constant C_2 such that for all $i \geq 1$ and $1 \leq j, k \leq d$

$$|\operatorname{Cov}(G_{ij}(\boldsymbol{\epsilon}_0), G_{ik}(\boldsymbol{\epsilon}_h)) - \operatorname{Cov}(G_{0j}(\boldsymbol{\epsilon}_0), G_{0k}(\boldsymbol{\epsilon}_h))| \leq \frac{C_2}{m}|\operatorname{Cov}(G_{0j}(\boldsymbol{\epsilon}_0), G_{0k}(\boldsymbol{\epsilon}_h))|.$$

Therefore, using Lemma 4

$$\begin{aligned}
|\sigma_{i,\infty}^2 - \sigma_{0,\infty}^2| &\le \sum_{h\in\mathbb{Z}} |\operatorname{Cov}(\boldsymbol{v}^\top \boldsymbol{G}_i(\boldsymbol{\epsilon}_0), \boldsymbol{v}^\top \boldsymbol{G}_i(\boldsymbol{\epsilon}_h)) - \operatorname{Cov}(\boldsymbol{v}^\top \boldsymbol{G}_0(\boldsymbol{\epsilon}_0), \boldsymbol{v}^\top \boldsymbol{G}_0(\boldsymbol{\epsilon}_h))| \\
&\le \sum_{h\in\mathbb{Z}} \sum_{j,k=1}^{d} |v_j||v_k||\operatorname{Cov}(G_{ij}(\boldsymbol{\epsilon}_0), G_{ik}(\boldsymbol{\epsilon}_h)) - \operatorname{Cov}(G_{0j}(\boldsymbol{\epsilon}_0), G_{0k}(\boldsymbol{\epsilon}_h))| \\
&\le \frac{C_2\|\boldsymbol{v}\|_1^2}{m} \sum_{h\in\mathbb{Z}} \max_{1\le j,k\le d} |\operatorname{Cov}(G_{0j}(\boldsymbol{\epsilon}_0), G_{0k}(\boldsymbol{\epsilon}_h))| \\
&\le \frac{C_2K^2\|\boldsymbol{v}\|_1^2}{m} \sum_{h\in\mathbb{Z}} \sum_{\ell=h}^{\infty} \ell^{-\beta} \\
&\le \frac{CK^2\|\boldsymbol{v}\|_1^2}{m} \sum_{\ell=0}^{\infty} \ell^{-\beta+1}
\end{aligned}$$

uniformly in $i \ge 1$, for some constant C. □

Proof of Theorem 2: (i): Observe that by Assumption A–iii

$$\inf_{i\ge1} \sigma_{i,\infty}^2 \ge \inf_{i\ge1} \boldsymbol{v}^\top \boldsymbol{\Sigma}_{i,\infty} \boldsymbol{v} > 0.$$

Define the random variables

$$\xi_i = \boldsymbol{v}^\top (\boldsymbol{Y}_i\boldsymbol{Y}_i^\top - \mathbf{M}_i)\boldsymbol{v}/\sigma_{i,\infty} \quad \text{with} \quad \sigma_{i,\infty}^2 = \sum_{h\in\mathbb{Z}} \boldsymbol{v}^\top \operatorname{Cov}(\boldsymbol{G}_i(\boldsymbol{\epsilon}_0), \boldsymbol{G}_i(\boldsymbol{\epsilon}_h))\boldsymbol{v},\ i \ge 1,$$

and let $S_n = \sum_{i=1}^n \xi_i$. Denote the jth physical dependence coefficients of $\{\xi_i\}$ by $\vartheta_{j,q}$, $j \ge 1$, and the tail dependence measure by $\Theta_{i,q} = \sum_{j\ge i} \vartheta_{j,q}$. By Lemma 2 (i) the random variables ξ_i have a summable tail dependence measure, since $\Theta_{j,q} \le 4K\|\boldsymbol{v}\|_1^2 C_1 j^{-\beta+1}$, and satisfy the assumptions of (Mies and Steland, 2023, Th. 3.1). Hence, after redefining the vector time series on a new probability space, there exist i.i.d. random variables $G_i \sim N(0,1)$, such that for fixed $0 < \varepsilon < \xi(q,\beta)$

$$\max_{k\le m} \left| S_k - \sum_{i=1}^k G_i \right| = O_{\mathbb{P}}(\sqrt{\log(m)} m^{\frac{1}{2}-\xi(q,\beta)}) = o_{\mathbb{P}}(m^{\frac{1}{2}-\xi(q,\beta)+\varepsilon}) = o_{\mathbb{P}}(m^{\frac{1}{\nu_1}}), \tag{15}$$

where $\nu_1 = \frac{1}{\frac{1}{2}-\xi(q,\beta)+\varepsilon} > 2$. Next we show that we can use $\sigma_{0,\infty}^2 = \boldsymbol{v}^\top \boldsymbol{\Sigma}_{0,\infty}\boldsymbol{v}$ instead of $\sigma_{i,\infty}^2 = \boldsymbol{v}^\top \boldsymbol{\Sigma}_{i,\infty}\boldsymbol{v}$ for standardization: Using $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x-y|}$, we obtain by Assumption A–iii, Lemma 2 (i), Lemma 5 and (Mies and Steland, 2023, Th. 3.2)

$$\begin{aligned}
&\mathbb{E}\max_{k\le m} \left| \sum_{i=1}^k \frac{\boldsymbol{v}^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top - \mathbf{M}_i)\boldsymbol{v}}{\sigma_{0,\infty}} - \sum_{i=1}^k \frac{\boldsymbol{v}^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top - \mathbf{M}_i)\boldsymbol{v}}{\sigma_{i,\infty}} \right|^2 \\
&\le \mathbb{E}\max_{k\le m} \sum_{i=1}^k \left| \frac{\sqrt{|\sigma_{0,\infty}^2 - \sigma_{i,\infty}^2|}}{\sigma_{0,\infty}\sigma_{i,\infty}} \right|^2 |\boldsymbol{v}^\top(\boldsymbol{Y}_i\boldsymbol{Y}_i^\top - \mathbf{M}_i)\boldsymbol{v}|^2
\end{aligned}$$

$$\leq \sup_{i\geq 1} \frac{|\sigma^2_{0,\infty} - \sigma^2_{i,\infty}|}{\sigma^2_{0,\infty} \inf_{\ell\geq 1} \sigma^2_{\ell,\infty}} \mathbb{E} \max_{k\leq m} \sum_{i=1}^{k} |\boldsymbol{v}^\top (\boldsymbol{Y}_i \boldsymbol{Y}_i^\top - \mathbf{M}_i)\boldsymbol{v}|^2$$
$$= O(1),$$

Therefore,

$$\max_{k\leq m} \left| \sum_{i=1}^{k} \frac{\boldsymbol{v}^\top (\boldsymbol{Y}_i \boldsymbol{Y}_i^\top - \mathbf{M}_i)\boldsymbol{v}}{\sigma_{0,\infty}} - S_k \right| = O_{\mathbb{P}}(1) = o_{\mathbb{P}}(m^{\frac{1}{\nu}}).$$

We can assume that $G_i = B_i - B_{i-1}$, $i \in \mathbb{N}$, for some standard Brownian motion B_t, $t \geq 0$, such that

$$\max_{k\leq m} |S_k - B_k| = o_{\mathbb{P}}(m^{\frac{1}{\nu_1}}),$$

which verifies (i) with $\nu = \nu_1$.

(ii): By Lemma 3 (iii), under Assumption A–iv, which provides $d = O(\sqrt{m})$ and $\sqrt{dr_d} = O(m^\zeta)$ for some $0 \leq \zeta < \frac{1}{2}$, one can replace $\boldsymbol{v}$ by its estimator $\hat{\boldsymbol{v}}_m$, because

$$\max_{k\leq m} \left| \sum_{i\leq k} \frac{\hat{\boldsymbol{v}}_m^\top (\boldsymbol{Y}_i \boldsymbol{Y}_i^\top - \mathbf{M}_i)\hat{\boldsymbol{v}}_m}{\sigma_{0,\infty}} - \sum_{i\leq k} \frac{\boldsymbol{v}^\top (\boldsymbol{Y}_i \boldsymbol{Y}_i^\top - \mathbf{M}_i)\boldsymbol{v}}{\sigma_{0,\infty}} \right| = O_{\mathbb{P}}\left(\sqrt{dr_d}\right) = o_{\mathbb{P}}\left(m^{\frac{1}{\nu_2}}\right), \tag{16}$$

where $\nu_2 = \frac{1}{\zeta+\varepsilon} > 2$, if ε is decreased, when necessary, to ensure $\zeta + \varepsilon < \frac{1}{2}$. Hence, by the triangle inequality, (15) remains true when replacing the ξ_i's by

$$\hat{\xi}_i = \frac{\hat{\boldsymbol{v}}_m^\top (\boldsymbol{Y}_i \boldsymbol{Y}_i^\top - \mathbf{M}_i)\hat{\boldsymbol{v}}_m}{\sigma_{0,\infty}}, \qquad 1 \leq i \leq m,$$

and ν_1 by $\nu = \min(\nu_1, \nu_2) > 2$. Thus, if $\hat{S}_k = \sum_{i\leq k} \hat{\xi}_i$, (16) shows that $\max_{k\leq m} |\hat{S}_k - S_k| = o_{\mathbb{P}}\left(m^{\frac{1}{\nu}}\right)$, and we obtain

$$\max_{k\leq m} \left| \hat{S}_k - B_k \right| = o_{\mathbb{P}}(m^{\frac{1}{\nu}}).$$

It remains to consider the case of s–sparseness. Under Assumption A–iv the quadratic forms appearing in (16) collapse to quadratic forms in terms of $\hat{\boldsymbol{v}}_A, \boldsymbol{v}_A$ and $\boldsymbol{Y}_{iA}$. Thus, all properties derived from the vector time series can assume that the dimension is s. Note that the assumption of (Mies and Steland, 2023, Th. 3.1) that the dimension is $O(m)$ is void, since we apply it to a univariate series, and the dimension only appears in the constants. Lemma 3 (iii) provides the bound $O_{\mathbb{P}}(\sqrt{sr_d})$ which is of the order $o_{\mathbb{P}}(m^{\frac{1}{\nu_2}})$, if $\sqrt{sr_d} = O(m^\zeta)$ as assumed in Assumption A–iv. Now the proof can be completed as for non–ℓ_0–sparse estimation. □

Proof of Theorem 3: Under the null hypothesis $\boldsymbol{v}^\top \mathbf{M}_i \boldsymbol{v} = \theta_0$, $i \geq 1$, we have

$$\mathbb{E}_0(\boldsymbol{v}^\top \mathbf{S}_i \boldsymbol{v}) = \boldsymbol{v}^\top \mathbb{E}_0(\mathbf{S}_i)\boldsymbol{v} = \theta_0$$

where $\mathbb{E}_0$ indicates that the expectation is taken under H_0. Therefore, centering all summands under $\mathbb{E}_0$ yields

$$\sum_{m<i\leq m+k} \boldsymbol{v}^\top \mathbf{S}_i \boldsymbol{v} - \frac{k}{m}\sum_{j=1}^{m} \boldsymbol{v}^\top \mathbf{S}_j \boldsymbol{v} = \sum_{m<i\leq m+k} \boldsymbol{v}^\top (\mathbf{S}_i - \mathbf{M}_i)\boldsymbol{v} - \frac{k}{m}\sum_{j=1}^{m} \boldsymbol{v}^\top (\mathbf{S}_j - \mathbf{M}_j)\boldsymbol{v}.$$

Therefore, for known $\boldsymbol{v}$ we can be write

$$\max_{\delta m\leq k} \frac{Q(m,k)}{\hat{\sigma}_{0,\infty} g(m,k)} = \max_{\delta m\leq k<\infty} \frac{\left|\sum_{m<i\leq m+k} \boldsymbol{v}^\top (\mathbf{S}_i - \mathbf{M}_i)\boldsymbol{v} - \frac{k}{m}\sum_{j=1}^{m} \boldsymbol{v}^\top (\mathbf{S}_j - \mathbf{M}_j)\boldsymbol{v}\right|}{\hat{\sigma}_{0,\infty} g(m,k)}.$$

As in the proof of Theorem 2, denote $\xi_i = \boldsymbol{v}^\top(\mathbf{S}_i - \mathbf{M}_i)\boldsymbol{v}/\sigma_{0,\infty}$, $\hat{\xi}_i = \hat{\boldsymbol{v}}_m^\top(\mathbf{S}_i - \mathbf{M}_i)\hat{\boldsymbol{v}}_m/\sigma_{0,\infty}$, $i \geq 1$, and $S_k = \sum_{i=1}^k \xi_i$, $\hat{S}_k = \sum_{i=1}^k \hat{\xi}_i$, $k \geq 1$. By Theorem 2, after redefining the vector time series on a richer probability space, there exists a standard Brownian motion $\{B_t : t \geq 0\}$ such that

$$\max_{k\leq m} |S_k - B_k| = o_{\mathbb{P}}(m^{\frac{1}{\nu}}) \qquad \text{and} \qquad \max_{k\leq m} \left|\hat{S}_k - B_k\right| = o_{\mathbb{P}}(m^{\frac{1}{\nu}})$$

with ν as specified in the theorem. Next, invoking the Skorohod representation theorem shows that one can define, on a new probability space, random variables $\hat{\xi}_i', i \geq 1$, together with a standard Brownian motion B_m' such that $\{\hat{\xi}_i : i \geq 1\} \stackrel{d}{=} \{\hat{\xi}_i' : i \geq 1\}$ and with $\hat{S}_k' = \sum_{i=1}^k \hat{\xi}_i'$

$$m^{-\frac{1}{\nu}} \left|\hat{S}_m' - B_m'\right| = o(1), \qquad a.s..$$

Therefore, the assumptions of Lemma 1 are satisfied, and we can assume that there exist, for each m, two independent standard Brownian motions $B_{1,m}$ and $B_{2,m}$, such that

$$|\hat{S}_m' - B_{1,m}| = o(m^{\frac{1}{\nu}})$$

and

$$\sup_{k\geq \delta m} k^{-\frac{1}{\nu}} |\hat{S}_{m+k}' - \hat{S}_m' - B_{2,m}| = O_{\mathbb{P}}(1).$$

The rest of the proof can be carried out along the lines of Horváth et al. (2004) which some modifications. For sake of clarity, we provide details and repeat some arguments. Consider

$$\begin{aligned}
&\sup_{k\geq m\delta} \frac{\left|\sum_{m<i\leq m+k} \hat{\xi}_i' - \frac{k}{m}\sum_{i=1}^{m} \hat{\xi}_i' - \left(B_{1,m}(k) - \frac{k}{m} B_{2,m}(m)\right)\right|}{g(m,k)} \\
&= \sup_{k\geq m\delta} \frac{O_{\mathbb{P}}(k^{\frac{1}{\nu}}) + \frac{k}{m} O_{\mathbb{P}}(m^{\frac{1}{\nu}})}{m^{\frac{1}{2}}\left(1+\frac{k}{m}\right)\left(\frac{k}{m+k}\right)^{\gamma}}
\end{aligned}$$

$$= O_{\mathbb{P}}(1) \sup_{k \geq \delta m} \frac{k^{\frac{1}{\nu}} + \frac{k}{m} m^{\frac{1}{\nu}}}{m^{\frac{1}{2}} \left(1 + \frac{k}{m}\right) \left(\frac{k}{m+k}\right)^{\gamma}}$$
$$= o_{\mathbb{P}}(1),$$

Let $\hat{Q}'(m,k) = \left| \sum_{m<i \leq m+k} \hat{\xi}'_i - \frac{k}{m} \sum_{j=1}^{m} \hat{\xi}'_j \right|$ and define

$$\tilde{Q}(m,k) = \left| B_{1,m}(k) - \frac{k}{m} B_{2,m}(m) \right|.$$

We have shown

$$\left| \sup_{k \geq m\delta} \frac{\hat{Q}'(m,k)}{\sigma_{0,\infty} g(m,k)} - \sup_{k \geq m\delta} \frac{\tilde{Q}(m,k)}{g(m,k)} \right| = o_{\mathbb{P}}(1). \tag{17}$$

Let us clarify the distributional limit of $\sup_{k \geq m\delta} \frac{\tilde{Q}(m,k)}{g(m,k)}$. We obtain for each fixed $T > 0$

$$\max_{\delta m \leq k \leq Tm} \frac{\tilde{Q}(m,k)}{g(m,k)} = \max_{t \in \left\{ \frac{k}{m} : \delta m \leq k \leq Tm \right\}} \frac{\left| \frac{1}{\sqrt{m}} B_{1,m}(tm) - \frac{k}{m} \frac{1}{\sqrt{m}} B_{2,m}(m) \right|}{(1+t) \left(\frac{t}{t+1}\right)^{\gamma}}$$
$$= \sup_{t \in [\delta, T]} \frac{|\frac{1}{\sqrt{m}} B_{1,m}(tm) - t \frac{1}{\sqrt{m}} B_{2,m}(m)|}{(1+t) \left(\frac{t}{1+t}\right)^{\gamma}} + o(1),$$

as $m \to \infty$, a.s., by Levy's continuity modulus. Since $\frac{1}{\sqrt{m}}(B_{1,m}(tm), B_{2,m}(m)) \stackrel{d}{=} (B_1(t), B_2(1))$ for all $m \geq 1$, where B_1, B_2 are independent standard Brownian motions, we may conclude that

$$\mathbb{P}\left(\left| \max_{\delta m \leq k \leq Tm} \frac{\tilde{Q}(m,k)}{g(m,k)} - \sup_{t \in [\delta, T]} \frac{|B_1(t) - tB_2(1)|}{(1+t) \left(\frac{t}{1+t}\right)^{\gamma}} \right| \to 0 \right) = 1.$$

The law of the iterated logarithm yields

$$\sup_{k > mT} \frac{\left| B_1\left(\frac{k}{m}\right) - \frac{k}{m} B_2\left(\frac{k}{m}\right) \right|}{\left(1 + \frac{k}{m}\right) \left(\frac{k}{k+m}\right)^{\gamma}} = o_{\mathbb{P}}(1),$$

and therefore we arrive at

$$\max_{\delta m \leq k < \infty} \frac{\tilde{Q}(m,k)}{g(m,k)} = \sup_{\delta \leq t < \infty} \frac{|B_1(t) - tB_2(1)|}{(1+t) \left(\frac{t}{1+t}\right)^{\gamma}} + o_{\mathbb{P}}(1).$$

Combining that results with (17) gives

$$\max_{\delta m \leq k < \infty} \frac{Q(m,k)}{\sigma_{0,\infty} g(m,k)} = \sup_{\delta \leq t < \infty} \frac{|B_1(t) - tB_2(1)|}{(1+t) \left(\frac{t}{1+t}\right)^{\gamma}} + o_{\mathbb{P}}(1),$$

and an application of the second half of the Skorohod representation theorem, yields, on the original probability space,

$$\max_{\delta m \le k < \infty} \frac{Q(m,k)}{\sigma_{0,\infty} g(m,k)} \xrightarrow{d} \sup_{\delta \le t < \infty} \frac{|B_1(t) - tB_2(1)|}{(1+t)\left(\frac{t}{1+t}\right)^{\gamma}},$$

as $m \to \infty$. Since $\{B_1(t) - tB_2(1) : t \ge 0\} \stackrel{d}{=} \{(1+t)B(t/(1+t)) : t \ge 0\}$, we may conclude that

$$\sup_{\delta \le t < \infty} \frac{|B_1(t) - tB_2(1)|}{(1+t)\left(\frac{t}{1+t}\right)^{\gamma}} \stackrel{d}{=} \sup_{\frac{\delta}{1+\delta} \le t \le 1} \frac{|B(t)|}{t^{\gamma}},$$

where $B(t)$, $0 \le t \le 1$, is a standard Brownian motion. Therefore,

$$\sup_{k \ge m\delta} \frac{Q(m,k)}{\sigma_{0,\infty} g(m,k)} \xrightarrow{d} \sup_{\frac{\delta}{1+\delta} \le t \le 1} \frac{|B(t)|}{t^{\gamma}},$$

and if $\hat{\sigma}^2_{0,\infty} \xrightarrow{\mathbb{P}} \sigma^2_{0,\infty}$, $m \to \infty$, then we also have

$$\sup_{k \ge m\delta} \frac{Q(m,k)}{\hat{\sigma}_{0,\infty} g(m,k)} \xrightarrow{d} \sup_{\frac{\delta}{1+\delta} \le t \le 1} \frac{|B(t)|}{t^{\gamma}}.$$

This completes the proof. □

Proof of Theorem 4: Define

$$\mathcal{Q}_m(t) = \frac{Q(m, \lfloor mt \rfloor)}{\sqrt{m}\hat{\sigma}_{0,\infty}(1+t)}, \qquad t \ge 0.$$

It follows from the proof of Theorem 3 that

$$\left\{ \frac{\mathcal{Q}_m(t)}{g_a(t/(1+t))} : t \ge \delta \right\} \Rightarrow \left\{ \frac{B(t/(1+t))}{g_a(t/(1+t))} : t \ge \delta \right\},$$

as $m \to \infty$, for some standard Brownian motion B, where $g_a(x) = \sqrt{(x+1)(a^2 + \log(x+1))}$. Therefore, by transforming time using the inverse $h^{-1} : [\delta/(1+\delta), 1) \to [\delta, \infty)$, $t \mapsto t/(1-t)$, of $h : [\delta, \infty) \to [\delta/(1+\delta), 1)$, $t \mapsto t/(1+t)$, we obtain

$$\left\{ \frac{\mathcal{Q}_m(t/(1-t))}{g_a(t)} : t \in [\delta/(1+\delta), 1) \right\} \Rightarrow \left\{ \frac{B(t)}{g_a(t)} : t \ge \delta \right\},$$

such that we can conclude

$$\begin{aligned} \lim_{m \to \infty} \mathbb{P}(\mathcal{Q}_m(k/(m-k)) > g_a(k/m), \exists \delta m \le k < m) &= \mathbb{P}(B(t) > g_a(t), \exists t \ge \delta) \\ &\nearrow \mathbb{P}(B(t) > g_a(t), \exists t > 0) \\ &= e^{-a^2/2}, \end{aligned}$$

as $\delta \searrow 0$. Note that $1 + t/(1-t) = 1/(1-t)$, so that

$$\mathcal{Q}_m\left(\frac{k}{m-k}\right) = \left(1-\frac{k}{m}\right)\frac{1}{\sqrt{m}\hat{\sigma}_{0,\infty}}\left(\sum_{i=m+1}^{m+\lfloor m\frac{k}{m-k}\rfloor}\xi_i - \frac{\lfloor m\frac{k}{m-k}\rfloor}{m}\sum_{j=1}^{m}\xi_j\right)$$
$$= \frac{m-k}{m}\frac{1}{\sqrt{m}\hat{\sigma}_{0,\infty}}\left(\sum_{i=m+1}^{m+\lfloor m\frac{k}{m-k}\rfloor}\xi_i - \frac{\lfloor m\frac{k}{m-k}\rfloor}{m}\sum_{j=1}^{m}\xi_j\right)$$

Hence, the event that an alarm is raised at physical time $k \geq \delta m$ where

$$\left|\mathcal{Q}_m\left(\frac{k}{m-k}\right)\right| > \sqrt{\frac{m+k}{m}\left(a^2 + \log(\frac{m+k}{m})\right)}$$

is equivalent to raising an alarm at transformed time $\delta m \leq k < m$ where

$$\left|\sum_{i=m+1}^{m+\lfloor m\frac{k}{m-k}\rfloor}\xi_i - \frac{\lfloor m\frac{k}{m-k}\rfloor}{m}\sum_{j=1}^{m}\xi_j\right| > \hat{\sigma}_{0,\infty}\sqrt{m}\frac{m}{m-k}\sqrt{\frac{m+k}{m}}\sqrt{a^2+\log\frac{m+k}{m}},$$

i.e., after rearranging terms,

$$\left|\sum_{i=m+1}^{m+\lfloor k\frac{m}{m-k}\rfloor}\xi_i - \frac{\lfloor k\frac{m}{m-k}\rfloor}{m}\sum_{j=1}^{m}\xi_j\right| > \hat{\sigma}_{0,\infty}\sqrt{m+k}\frac{m}{m-k}\sqrt{a^2+\log\frac{m+k}{m}}.$$

This means, when for (transformed time) k the threshold is exceeded for the first time, this occurs when the centered partial sum using the first $\ell = \lfloor m\frac{k}{m-k}\rfloor \geq k$ observations after m crosses the boundary. Vice versa, when a signal is given when the centered partial sum using the first $k \geq \delta m$ observations (physical time) crosses the boundary, this corresponds to $k^* = \lceil k\frac{m}{m+k}\rceil$ (solve $k \leq k^*\frac{m}{m-k^*} < k+1$ for k^*), i.e.,

$$\left|\sum_{i=m+1}^{m+k}\xi_i - \frac{k}{m}\sum_{j=1}^{m}\xi_j\right| > \hat{\sigma}_{0,\infty}\sqrt{m+k^*}\frac{m}{m-k^*}\sqrt{a^2+\log\frac{m+k^*}{m}},$$

which verifies (2). □

Proof of Theorem 5: By Assumption A–iii we have $\mathbf{M}_i = \mathbf{M}_i^0$ for $1 \leq i \leq m$, where $\mathbf{M}_i^0$ guarantees $\boldsymbol{v}^\top \mathbf{M}_i^0 \boldsymbol{v} = \theta_0$, and, under the sequence of local alternatives for the moment functional,

$$\boldsymbol{v}^\top \mathbf{M}_{m+k}\boldsymbol{v} = \theta_0 + \frac{\Delta\left(\frac{k-k^*}{m}\right)}{\sqrt{m}}\mathbf{1}_{k\geq k^*}, \qquad k \geq 1.$$

Therefore

$$\sum_{i=m+1}^{m+k}(\boldsymbol{v}^\top \mathbf{S}_i \boldsymbol{v}-\theta_0)=\sum_{i=m+1}^{m+k}\boldsymbol{v}^\top(\mathbf{S}_i-\mathbf{M}_i)\boldsymbol{v}+\mathbf{1}_{k\geq k^*}\frac{1}{\sqrt{m}}\sum_{i=k^*}^{k}\Delta\left(\frac{i-k^*}{m}\right)$$

and

$$\begin{aligned}\sum_{i=m+1}^{m+k}\boldsymbol{v}^\top\mathbf{S}_i\boldsymbol{v}-\frac{k}{m}\sum_{j=1}^{m}\boldsymbol{v}^\top\mathbf{S}_j\boldsymbol{v}&=\sum_{i=m+1}^{m+k}(\boldsymbol{v}^\top\mathbf{S}_i\boldsymbol{v}-\theta_0)-\frac{k}{m}\sum_{j=1}^{m}(\boldsymbol{v}^\top\mathbf{S}_j\boldsymbol{v}-\theta_0)\\&=\sum_{i=m+1}^{k}\boldsymbol{v}^\top(\mathbf{S}_i-\mathbf{M}_i)\boldsymbol{v}-\frac{k}{m}\sum_{j=1}^{m}\boldsymbol{v}^\top(\mathbf{S}_j-\mathbf{M}_j)\boldsymbol{v}\\&\quad+\mathbf{1}_{k\geq k^*}\frac{1}{\sqrt{m}}\sum_{i=k^*}^{k}\Delta\left(\frac{i-k^*}{m}\right).\end{aligned}$$

Since Δ has bounded total variation $TV(\Delta;[0,T])$ on $[0,T]$, we have for $k^*\leq k$

$$\left|\frac{1}{\sqrt{m}}\sum_{i=k^*}^{k}\Delta\left(\frac{i-k^*}{m}\right)-\sqrt{m}\int_0^{\frac{k-k^*}{m}}\Delta(s)\,ds\right|\leq\frac{TV(\Delta;[0,T])}{\sqrt{m}}$$

and thus

$$\max_{\delta m\leq k\leq Tm}\left|\frac{\frac{\mathbf{1}_{k\geq k^*}}{\sqrt{m}}\sum_{i=k^*}^{k}\Delta\left(\frac{i-k^*}{m}\right)-\sqrt{m}\mathbf{1}_{k\geq k^*}\int_0^{\frac{k-k^*}{m}}\Delta(s)\,ds}{\sqrt{m}\left(1+\frac{k}{m}\right)\left(\frac{k}{k+m}\right)^\gamma}\right|\leq\frac{TV(\Delta;[0,T])}{(1+\delta)\frac{\delta}{T+1}m}.$$

Arguing as in the proof of Theorem 3 we obtain for $\hat{Q}'(m,k)=\sum_{m<i\leq m+k}\hat{\xi}_i'-\frac{k}{m}\sum_{j=1}^{m}\hat{\xi}_j'$

$$\left|\max_{\delta m\leq k\leq Tm}\frac{\hat{Q}'(m,k)}{\sigma_{0,\infty}^2 g(m,k)}-\max_{\delta m\leq k\leq Tm}\frac{B_{1,m}(k)-\frac{k}{m}B_{2,m}(m)+\sqrt{m}\mathbf{1}_{k\geq k^*}\int_0^{\frac{k-k^*}{m}}\Delta(s)\,ds}{g(m,k)}\right|=o_{\mathbb{P}}(1).$$

Further,

$$\begin{aligned}&\max_{\delta m\leq k\leq Tm}\frac{B_{1,m}(k)-\frac{k}{m}B_{2,m}(m)+\sqrt{m}\int_0^{\frac{k-k^*}{m}}\Delta(s)\,ds}{g(m,k)}\\&\quad\stackrel{d}{=}\max_{\delta m\leq k\leq Tm}\frac{\left|B_1\left(\frac{k}{m}\right)-\frac{k}{m}B_2(1)+\mathbf{1}_{k\geq k^*}\int_0^{(k-k^*)/m}\Delta(s)\,ds\right|}{\left(1+\frac{k}{m}\right)\left(\frac{k}{k+m}\right)^\gamma}\\&\quad\stackrel{d}{\to}\sup_{t\in[\delta,T]}\frac{|B_1(t)-tB_2(1)+\mathbf{1}_{t\geq\vartheta}\int_0^{t-\vartheta}\Delta(s)\,ds|}{(1+t)\left(\frac{t}{1+t}\right)^\gamma},\end{aligned}$$

which completes the proof. □

The following proofs draw to some extent on Bickel and Levina (2008).

Proof of Theorem 6: First observe that

$$\|\mathcal{T}_t\mathbf{\Sigma} - \mathbf{\Sigma}\|_{op} \leq \max_{1\leq i\leq d}\sum_{j=1}^{d}|\Sigma_{ij}|\mathbf{1}(|\ \Sigma_{ij}| \leq t) \leq \max_{1\leq i\leq d}\sum_{j=1}^{d}|\Sigma_{ij}|^r t^{1-r} \leq t^{1-r}s_o$$

Clearly, $\|\mathcal{T}_t\mathbf{\Gamma} - \mathcal{T}_t\mathbf{\Sigma}\|_{op} \leq \max_{1\leq i\leq d}\sum_{j=1}^{d}|\Gamma_{ij}\mathbf{1}_{|\Gamma_{ij}|\geq t} - \Sigma_{ij}\mathbf{1}_{|\Sigma_{ij}|\geq t}|$. We have the decomposition

$$\begin{aligned}
\Gamma_{ij}\mathbf{1}_{|\Gamma_{ij}|\geq t} - \Sigma_{ij}\mathbf{1}_{|\Sigma_{ij}|\geq t} &= (\Gamma_{ij}-\Sigma_{ij})\mathbf{1}_{|\Gamma_{ij}|\geq t,|\Sigma_{ij}|<t} + (\Gamma_{ij}-\Sigma_{ij})\mathbf{1}_{|\Gamma_{ij}|\geq t,|\Sigma_{ij}|\geq t} \\
&\quad + \Sigma_{ij}\left(\mathbf{1}_{|\Gamma_{ij}|\geq t,|\Sigma_{ij}|<t} - \mathbf{1}_{|\Sigma_{ij}|\geq t} + \mathbf{1}_{|\Gamma_{ij}|\geq t,\Sigma_{ij}\geq t}\right) \\
&= (\Gamma_{ij}-\Sigma_{ij})\mathbf{1}_{|\Gamma_{ij}|\geq t,|\Sigma_{ij}|<t} + (\Gamma_{ij}-\Sigma_{ij})\mathbf{1}_{|\Gamma_{ij}|\geq t,|\Sigma_{ij}|\geq t} \\
&\quad + \Sigma_{ij}\left(\mathbf{1}_{|\Gamma_{ij}|\geq t,|\Sigma_{ij}|<t} - \mathbf{1}_{|\Gamma_{ij}|<t,|\Sigma_{ij}|\geq t}\right) \\
&= U_{ij}^1 + U_{ij}^2 + U_{ij}^3.
\end{aligned}$$

Fix $0<\gamma<1$. Since $|\Gamma_{ij}-\Sigma_{ij}| > (1-\gamma)t$, if $|\Gamma_{ij}|\geq t$ and $|\Sigma_{ij}|<\gamma t$,

$$\begin{aligned}
\max_{1\leq i\leq d}\sum_{j=1}^{d}|U_{ij}^1| &\leq \max_{1\leq i\leq d}\sum_{j=1}^{d}|\Gamma_{ij}-\Sigma_{ij}|\mathbf{1}_{|\Gamma_{ij}|\geq t,|\Sigma_{ij}|<\gamma t} + \max_{1\leq i\leq d}\sum_{j=1}^{d}|\Gamma_{ij}-\Sigma_{ij}|\mathbf{1}_{|\Gamma_{ij}|\geq t,\gamma t\leq|\Sigma_{ij}|<t} \\
&\leq \|\mathbf{\Gamma}-\mathbf{\Sigma}\|_\infty\left(\#(|\Gamma_{ij}-\Sigma_{ij}|>(1-\gamma)t) + \max_{1\leq i\leq d}\sum_{j=1}^{d}\frac{|\Sigma_{ij}|^r}{(\gamma t)^r}\mathbf{1}(|\Sigma_{ij}|\geq\gamma t)\right) \\
&\leq \|\mathbf{\Gamma}-\mathbf{\Sigma}\|_\infty\left(\#(|\Gamma_{ij}-\Sigma_{ij}|>(1-\gamma)t) + \frac{1}{(\gamma t)^r}s_0\right).
\end{aligned}$$

Further,

$$\max_{1\leq i\leq d}\sum_{j=1}^{d}|U_{ij}^2| \leq \|\mathbf{\Gamma}-\mathbf{\Sigma}\|_\infty \max_{1\leq i\leq d}\sum_{j=1}^{d}\frac{|\Sigma_{ij}|^r}{t^r}\mathbf{1}_{|\Sigma_{ij}|>t} \leq \frac{s_0}{t^r}\|\mathbf{\Gamma}-\mathbf{\Sigma}\|_\infty.$$

To bound $\max_{1\leq i\leq d}\sum_{j=1}^{d}|U_{ij}^3|$ first observe that

$$\max_{1\leq i\leq d}\sum_{j=1}^{d}|\Sigma_{ij}|\mathbf{1}_{|\Gamma_{ij}|\geq t,|\Sigma_{ij}|<t} \leq \max_{1\leq i\leq d}\sum_{j=1}^{d}|\Sigma_{ij}|^r|\Sigma_{ij}|^{1-r}\mathbf{1}_{|\Sigma_{ij}|<t} \leq t^{1-r}s_0.$$

Lastly,

$$\begin{aligned}
\max_{1\leq i\leq d}\sum_{j=1}^{d}|\Sigma_{ij}|\mathbf{1}_{|\Gamma_{ij}|<t,|\Sigma_{ij}|\geq t} &\leq \max_{1\leq i\leq d}\sum_{j=1}^{d}|\Gamma_{ij}-\Sigma_{ij}|\mathbf{1}_{|\Gamma_{ij}|<t,|\Sigma_{ij}|\geq t} + \max_{1\leq i\leq d}\sum_{j=1}^{d}|\Gamma_{ij}|\mathbf{1}_{|\Gamma_{ij}|<t,|\Sigma_{ij}|\geq t} \\
&\leq \|\mathbf{\Sigma}-\mathbf{\Sigma}\|_\infty \max_{1\leq i\leq d}\sum_{j=1}^{d}\frac{|\Sigma_{ij}|^r}{t^r}\mathbf{1}_{|\Sigma_{ij}|\geq t} + \max_{1\leq i\leq d}\sum_{j=1}^{d}t\frac{|\Sigma_{ij}|^r}{t^r}\mathbf{1}_{|\Sigma_{ij}|>t}
\end{aligned}$$

$$\leq \|\mathbf{\Gamma} - \mathbf{\Sigma}\|_\infty \frac{s_0}{t^r} + t^{1-r} s_0.$$

Combining the above bounds and collecting term shows that

$$\begin{aligned} \| \mathcal{T}_t\mathbf{\Gamma} - \mathbf{\Sigma}\|_{op} &\leq \|\mathcal{T}_t\mathbf{\Sigma} - \mathbf{\Sigma}\|_{op} + \|\mathcal{T}_t\mathbf{\Gamma} - \mathcal{T}_t\mathbf{\Sigma}\|_{op} \\ &\leq 2t^{1-r} s_0 + \|\mathbf{\Gamma} - \mathbf{\Sigma}\|_\infty \left(\#(|\Gamma_{ij} - \Sigma_{ij}| > (1-\gamma)t) + \frac{1}{(\gamma t)^r} s_0 + \frac{2}{t^r} s_0 \right). \end{aligned}$$

For the soft-thresholding operator we have

$$|\mathcal{S}_t(\Gamma_{ij}) - \mathcal{S}_t(\Sigma_{ij})| \leq \begin{cases} |\mathcal{S}_t(\Gamma_{ij})| \leq |\Gamma_{ij}| \leq |\Gamma_{ij} - \Sigma_{ij}| + |\Sigma_{ij}|, & |\Gamma_{ij}| \geq t, |\Sigma_{ij}| < t, \\ |\Sigma_{ij}|, & |\Gamma_{ij}| < t, |\Sigma_{ij}| \geq t, \\ 2t + |\Gamma_{ij} - \Sigma_{ij}|, & |\Gamma_{ij}| \geq t, |\Sigma_{ij}| \geq t, \end{cases}$$

where in the last case we used the estimate

$$\begin{aligned} |\mathcal{S}(\Gamma_{ij}) - \mathcal{S}(\Sigma_{ij})| &\leq |\mathcal{S}_t(\Gamma_{ij}) - \Gamma_{ij}| + |\Gamma_{ij} - \Sigma_{ij}| + |\mathcal{S}_t(\Sigma_{ij}) - \Sigma_{ij}| \\ &\leq 2t + |\Gamma_{ij} - \Sigma_{ij}|. \end{aligned}$$

Therefore,

$$\begin{aligned} \max_{1\leq i\leq d} \sum_{j=1}^d |\mathcal{S}_t(\Gamma_{ij}) - \mathcal{S}_t(\Sigma_{ij})| &\leq \max_{1\leq i\leq d} \sum_{j=1}^d |\Gamma_{ij} - \Sigma_{ij}| (\mathbf{1}_{|\Gamma_{ij}|\geq t, |\Sigma_{ij}|<t} + \mathbf{1}_{|\Gamma_{ij}|\geq t, |\Sigma_{ij}|\geq t}) \\ &\quad + \max_{1\leq i\leq d} \sum_{j=1}^d |\Sigma_{ij}| (\mathbf{1}_{|\Gamma_{ij}|\geq t, |\Sigma_{ij}|<t} + \mathbf{1}_{|\Gamma_{ij}|<t, |\Sigma_{ij}|\geq t}) \\ &= U_{ij}^1 + U_{ij}^2 + U_{ij}^3, \end{aligned}$$

such that the same bound results. □

Proof of Theorem 7 and Corollary 1: We need to control the upper bound on $\|T_s\hat{\mathbf{\Sigma}}_m - \bar{\mathbf{\Sigma}}_m\|_{op}$. By Lemma 1, Markov's inequality yields for $t > 0$

$$\mathbb{P}(|\hat{\Sigma}_{m,jk} - \bar{\Sigma}_{m,jk}| \geq t) \leq \frac{\mathbb{E}|\hat{\Sigma}_{m,jk} - \bar{\Sigma}_{m,jk}|^{\frac{q}{2}}}{t^{\frac{q}{2}}} \leq \frac{c m^{-\frac{q}{4}}}{t^{\frac{q}{2}}}$$

such that the union bound yields

$$\mathbb{P}\left(d^{-\frac{4}{q}} m^{\frac{1}{2}} \max_{1\leq j,k\leq d} |\hat{\Sigma}_{m,jk} - \bar{\Sigma}_{m,jk}| \geq t \right) \leq \frac{c}{t^{\frac{q}{2}}}.$$

Therefore,

$$\|\hat{\mathbf{\Sigma}}_m - \bar{\mathbf{\Sigma}}_m\|_\infty = O_{\mathbb{P}}\left(\frac{d^{\frac{4}{q}}}{\sqrt{m}} \right).$$

Clearly, the proof carries over to $\hat{\mathbf{M}}_m$, and since $\max_{1\leq j\leq d}\mathbb{E}|\hat{\mu}_{mj}-\bar{\mu}_{mj}|^q = O(m^{q/2})$, one gets $\mathbb{P}(\frac{\sqrt{m}}{d^{2/q}}\|\hat{\boldsymbol{\mu}}_m-\bar{\boldsymbol{\mu}}_m\|_\infty \geq t) = O(t^{-q/2})$, so that $\|\hat{\boldsymbol{\mu}}_m-\bar{\boldsymbol{\mu}}_m\|_\infty = O_{\mathbb{P}}(\frac{d^{2/q}}{\sqrt{m}})$ follows. Bounding the Frobenius norm by d times the supnorm gives

$$\begin{aligned}\mathbb{P}\left(d^{-\frac{4+q}{q}}m^{\frac{1}{2}}\|\hat{\Sigma}_{m,jk}-\bar{\Sigma}_{m,jk}\|_F \geq t\right) &\leq \mathbb{P}\left(d^{-\frac{4}{q}}m^{-\frac{1}{2}}\max_{1\leq j,k\leq d}|\hat{\Sigma}_{m,jk}-\bar{\Sigma}_{m,jk}| \geq t\right)\\ &\leq c\frac{d^2m^{-\frac{q}{4}}}{t^{\frac{q}{2}}m^{-\frac{q}{4}}d^2} = \frac{c}{t^{\frac{q}{2}}},\end{aligned}$$

and $\mathbb{P}(d^{\frac{2-q}{q}}m^{\frac{1}{2}}\|\hat{\boldsymbol{\mu}}_m-\bar{\boldsymbol{\mu}}_m\|_2 \geq t) \leq d\mathbb{P}(d^{2/q}m^{1/2}\|\hat{\boldsymbol{\mu}}_m-\bar{\boldsymbol{\mu}}_m\|_\infty \geq t) = O(t^{-q/2})$. Thus, Corollary 1 is shown. Next, we have

$$p_0 = \mathbb{P}\left(\#\left(|\hat{\Sigma}_{m,ij}-\bar{\Sigma}_{m,ij}| > (1-\gamma)t\right) > 0\right) = \mathbb{P}(\|\hat{\boldsymbol{\Sigma}}_m-\bar{\boldsymbol{\Sigma}}_m\|_\infty > (1-\gamma)t)$$

Thus, if one selects the threshold as a multiple of the convergence rate of $\|\hat{\boldsymbol{\Sigma}}_m-\bar{\boldsymbol{\Sigma}}_m\|_\infty$, i.e., $t_{th} = C_{th}\frac{d^{\frac{4}{q}}}{\sqrt{m}}$, for some constant C_{th}, then p_0 is arbitrarily small if C_{th} is large enough. Now, recalling that $0 \leq r < 1$, it easily follows by noting that $\|\hat{\boldsymbol{\Sigma}}_m-\bar{\boldsymbol{\Sigma}}_m\|_\infty = O_P(t_{th}^{1-r})$ that

$$\|\mathcal{T}_{th}\hat{\boldsymbol{\Sigma}}_m-\bar{\boldsymbol{\Sigma}}_m\|_{op} = O(t_{th}^{1-r}) + O_{\mathbb{P}}(\|\hat{\boldsymbol{\Sigma}}_m-\bar{\boldsymbol{\Sigma}}_m\|_\infty) = O_{\mathbb{P}}(t_{th}^{1-r})$$

which establishes Theorem 7. □

Proof of Theorem 8: Observe that the thresholded matrix $\mathcal{S}_{t_m}\hat{\boldsymbol{\Sigma}}_m$, for some sequence $t_m = o(1)$ of thresholds, has maximal eigenvalue bounded by $\max_i\sum_j|\bar{\Sigma}_{m,ij}| \leq M^{1-r}s_0$ as well, and that $\lambda_{\min}(\mathcal{S}_{t_m}\hat{\boldsymbol{\Sigma}}_m) \geq \frac{1}{2}\lambda_{\min}(\bar{\boldsymbol{\Sigma}}_m) \geq \frac{\varepsilon_0}{2}$ if t_m is small enough to ensure that $\|\mathcal{S}_{t_m}-\mathcal{S}_0\|_{op} < \inf_{m\geq 1}\lambda_{\min}(\bar{\boldsymbol{\Sigma}}_m)$, since the arguments of (Bickel and Levina, 2008, p.2580) carry over. Therefore $\mathcal{S}_{t_m}\hat{\boldsymbol{\Sigma}}_m \in \mathcal{U}(s,s_0,M,\varepsilon_0)$. Consequently, the well known inequality

$$\|\mathbf{A}^{-1}-\mathbf{B}^{-1}\|_{op} \leq \frac{\|\mathbf{A}-\mathbf{B}\|_{op}}{\lambda_{\min}(\mathbf{A})\lambda_{\min}(\mathbf{B})}$$

for regular square matrices $\mathbf{A},\mathbf{B}$ of the same dimension, yields for the inverses, under the assumptions of Theorem 7,

$$\|(\mathcal{S}_{t_m}\hat{\boldsymbol{\Sigma}}_m)^{-1}-\bar{\boldsymbol{\Sigma}}_m^{-1}\|_{op} = O_{\mathbb{P}}\left(\left(\frac{d^{\frac{4}{q}}}{\sqrt{m}}\right)^{1-r}\right).$$

□

Proof of Proposition 2: If $\sigma:\mathbb{R}\to\mathbb{R}$ is ρ–Lipschitz, then $\boldsymbol{x}\mapsto\sigma(\boldsymbol{x})$, $\boldsymbol{x}$ some vector, is ρ–Lipschitz with respect to any p–vector norm. Hence, for a ρ–Lipschitz activation σ with $\sigma(0)=0$, $N\times r$ weighting matrices $\mathbf{W},\widetilde{\mathbf{W}}$ and $\boldsymbol{x}\in\mathbb{R}^r$, $N,r\in\mathbb{N}$,

(i) $|\sigma^{\widetilde{\mathbf{W}}}(\boldsymbol{x})-\sigma^{\mathbf{W}}(\boldsymbol{x})| \leq \rho\|(\widetilde{\mathbf{W}}-\mathbf{W})\boldsymbol{x}\|_2 \leq \rho\|\widetilde{\mathbf{W}}-\mathbf{W}\|_{op}\|\boldsymbol{x}\|_2$,

(ii) $\|\sigma^{\mathbf{W}}(\widetilde{\boldsymbol{x}}) - \sigma^{\mathbf{W}}(\boldsymbol{x})\|_2 \le \rho\|\mathbf{W}(\widetilde{\boldsymbol{x}} - \boldsymbol{x})\|_2 \le \rho\|\mathbf{W}\|_{op}\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}\|_2$,

(iii) $\|\sigma^{\mathbf{W}}(\boldsymbol{x})\|_2 = \|\sigma^{\mathbf{W}}(\boldsymbol{x}) - \sigma(0)\|_2 \le \rho\|\mathbf{W}\|_{op}\|\boldsymbol{x}\|_2$, and therefore

(iv) $\|\boldsymbol{f}_j(\boldsymbol{x})\|_2 = \|\sigma_j^{\mathbf{W}_j} \circ \cdots \circ \sigma_1^{\mathbf{W}_1}(\boldsymbol{x})\|_2 \le \prod_{i=1}^{j} \rho_i\|\mathbf{W}_i\|_{op}\|\boldsymbol{x}\|_2$.

We have for $k < j \le H$

$$
\begin{aligned}
&\|\boldsymbol{f}_j(\boldsymbol{x};\boldsymbol{\theta}(\widetilde{\mathbf{W}}_k)) - \boldsymbol{f}_j(\boldsymbol{x};\boldsymbol{\theta}(\widetilde{\mathbf{W}}_{k-1}))\|_2 \\
&\quad= \|\sigma_j^{\mathbf{W}_j} \circ \cdots \circ \sigma_{k+1}^{\mathbf{W}_{k+1}} \circ \sigma_k^{\widetilde{\mathbf{W}}_k} \circ \cdots \circ \sigma_1^{\widetilde{\mathbf{W}}_1}(\boldsymbol{x}) - \sigma_j^{\mathbf{W}_j} \circ \cdots \circ \sigma_k^{\mathbf{W}_k} \circ \sigma_{k-1}^{\widetilde{\mathbf{W}}_{k-1}} \circ \cdots \circ \sigma_1^{\widetilde{\mathbf{W}}_1}(\boldsymbol{x})\|_2 \\
&\quad\overset{(ii)}{\le} \rho_j\|\mathbf{W}_j\|_{op}\|\sigma_{j-1}^{\mathbf{W}_{j-1}} \circ \cdots \circ \sigma_{k+1}^{\mathbf{W}_{k+1}} \circ \sigma_k^{\widetilde{\mathbf{W}}_k} \circ \cdots \circ \sigma_1^{\widetilde{\mathbf{W}}_1}(\boldsymbol{x}) \\
&\qquad\qquad - \sigma_{j-1}^{\mathbf{W}_{j-1}} \circ \cdots \circ \sigma_k^{\mathbf{W}_k} \circ \sigma_{k-1}^{\widetilde{\mathbf{W}}_{k-1}} \circ \cdots \circ \sigma_1^{\widetilde{\mathbf{W}}_1}(\boldsymbol{x})\|_2 \\
&\quad\le \cdots \le \prod_{i=k+1}^{j} \rho_i\|\mathbf{W}_i\|_{op}\|\sigma_k^{\widetilde{\mathbf{W}}_k} \circ \cdots \circ \sigma_1^{\widetilde{\mathbf{W}}_1}(\boldsymbol{x}) - \sigma_k^{\mathbf{W}_k} \circ \sigma_{k-1}^{\widetilde{\mathbf{W}}_{k-1}} \cdots \circ \sigma_1^{\mathbf{W}_1}(\boldsymbol{x})\|_2 \\
&\quad\overset{(i)}{\le} \left(\prod_{i=k+1}^{j} \rho_i\|\mathbf{W}_i\|_{op}\right) \rho_k\|\widetilde{\mathbf{W}}_k - \mathbf{W}_k\|_{op}\|\sigma_{k-1}^{\widetilde{\mathbf{W}}_{k-1}} \circ \cdots \circ \sigma_1^{\widetilde{\mathbf{W}}_1}(\boldsymbol{x})\|_2 \\
&\quad\overset{(iv)}{\le} \prod_{i=1, i\neq k}^{j} \rho_i\|\mathbf{W}_i\|_{op}\rho_k\|\widetilde{\mathbf{W}}_k - \mathbf{W}_k\|_{op}\|\boldsymbol{x}\|_2 \\
&\quad= L_{jk}\|\widetilde{\mathbf{W}}_k - \mathbf{W}_k\|_{op}\|\boldsymbol{x}\|_2
\end{aligned}
$$

where $L_{jk} = \prod_{i=1, i\neq k}^{j} \rho_i\|\mathbf{W}_i\|_{op}\rho_k$. Notice that $\boldsymbol{\theta}(\widetilde{\mathbf{W}}_0) = \boldsymbol{\theta}$ and $\boldsymbol{\theta}(\widetilde{\mathbf{W}}_H) = \widetilde{\boldsymbol{\theta}}$. By telescoping and estimating all spectral norms by the corresponding Frobenius matrix norms, we obtain

$$
\begin{aligned}
\|\boldsymbol{f}_H(\boldsymbol{x};\widetilde{\boldsymbol{\theta}}) - \boldsymbol{f}_H(\boldsymbol{x};\boldsymbol{\theta})\|_2 &= \left\| \sum_{k=1}^{H} f_H(\boldsymbol{x};\theta(\widetilde{\mathbf{W}}_k)) - f_H(\boldsymbol{x};\theta(\widetilde{\mathbf{W}}_{k-1})) \right\|_2 \\
&\le \sum_{h=1}^{H} L_{Hk}\|\widetilde{\mathbf{W}}_k - \mathbf{W}_k\|_{op}\|\boldsymbol{x}\|_2 \\
&\le H \max_{1\le k\le H} L_{Hk}\|\boldsymbol{x}\|_2 \frac{1}{H}\sum_{k=1}^{H} \|\widetilde{\mathbf{W}}_k - \mathbf{W}_k\|_F \\
&\le H L_H\|\boldsymbol{x}\|_2 \sqrt{\frac{1}{H}\sum_{k=1}^{H} \|\widetilde{\mathbf{W}}_k - \mathbf{W}_k\|_F^2} \\
&= L_H\sqrt{H}\|\boldsymbol{x}\|_2 \sqrt{\sum_{k=1}^{H} \|\widetilde{\mathbf{W}}_k - \mathbf{W}_k\|_F^2} \\
&= L_H\sqrt{H}\|\boldsymbol{x}\|_2\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_F,
\end{aligned}
$$

where the last inequality follows from Jensen's inequality. □

Acknowledgements

The author acknowledges support from Deutsche Forschungsgemeinschaft (DFG, grant STE 1034/11–2). He thanks M.Sc. Florian Scholze for carefully proof–reading an earlier version of the manuscript.

References

Alexander Aue and Lajos Horváth. Delay time in sequential detection of change. *Statist. Probab. Lett.*, 67(3):221–231, 2004. ISSN 0167-7152. doi: 10.1016/j.spl.2004.01.002. URL https://doi.org/10.1016/j.spl.2004.01.002.

Arindam Banerjee, Pedro Cisneros-Velarde, Libin Zhu, and Mikhail Belkin. Restricted strong convexity of deep learning models with smooth activations, 2022. URL https://arxiv.org/abs/2209.15106.

P.L. Bartlett, D.J. Foster, and M.J. Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604, 2008. ISSN 0090-5364.

Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1999. ISBN 0-471-19745-9. doi: 10.1002/9780470316962. URL http://dx.doi.org/10.1002/9780470316962. A Wiley-Interscience Publication.

Edward Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.*, 14(3):1171–1179, 1986. ISSN 0090-5364. doi: 10.1214/aos/1176350057. URL https://doi.org/10.1214/aos/1176350057.

Zhan Shou Chen, Zheng Tian, Rui Bing Qin, and Cheng Cai Leng. Sequential monitoring variance change in linear regression model. *J. Math. Res. Exposition*, 30(4):610–618, 2010. ISSN 1000-341X.

Chia-Shang James Chu, Maxwell Stinchcombe, and Halbert White. Monitoring structural change. *Econometrica*, 64(5):1045–1065, 1996.

Deniz Erdogmus, Kenneth E Hild, Yadunandana N Rao, and Jose C Principe. Minimax mutual information approach for independent component analysis. *Neural Computation*, 16(6):1235–1252, 2004.

P. Erdös and M. Kac. On certain limit theorems of the theory of probability. *Bulletin of the American Mathematical Society*, 52(4):292 – 302, 1946.

Jianqing Fan, Jingjin Zhang, and Ke Yu. Vast portfolio selection with gross-exposure constraints. *J. Amer. Statist. Assoc.*, 107(498):592–606, 2012. ISSN 0162-1459. doi:

10.1080/01621459.2012.682825. URL https://doi.org/10.1080/01621459.2012.682825.

Mehdi Ghasemi, Salma Kuhlmann, and Murray Marshall. Moment problem in infinitely many variables. *Israel J. Math.*, 212(2):989–1012, 2016. ISSN 0021-2172,1565-8511. doi: 10.1007/s11856-016-1318-5. URL https://doi.org/10.1007/s11856-016-1318-5.

H. Gouk, E. Frank, B. Pfahringer, and M.J. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Mach. Learning*, 110:393–416, 2021.

E. K. Haviland. On the Momentum Problem for Distribution Functions in More Than One Dimension. II. *Amer. J. Math.*, 58(1):164–168, 1936. ISSN 0002-9327,1080-6377. doi: 10.2307/2371063. URL https://doi.org/10.2307/2371063.

Lajos Horváth, Marie Hušková, Piotr Kokoszka, and Josef Steinebach. Monitoring changes in linear models. *J. Statist. Plann. Inference*, 126(1):225–251, 2004. ISSN 0378-3758. doi: 10.1016/j.jspi.2003.07.014. URL https://doi.org/10.1016/j.jspi.2003.07.014.

Jean Jacod and Michael Soerensen. A review of asymptotic theory of estimating functions. *Stat. Inference Stoch. Process.*, 21(2):415–434, 2018. ISSN 1387-0874. doi: 10.1007/s11203-018-9178-8. URL https://doi.org/10.1007/s11203-018-9178-8.

R Jagannathan and TS Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58(4):1651–1683, AUG 2003. ISSN 0022-1082. doi: 10.1111/1540-6261.00580.

V. John, I. Angelov, A.A. Öncül, and D. Thévenin. Techniques for the reconstruction of a distribution from a finite number of its moments. *Chemical Engineering Science*, 62(11):2890–2904, 2007. ISSN 0009-2509. doi: https://doi.org/10.1016/j.ces.2007.02.041. URL https://www.sciencedirect.com/science/article/pii/S0009250907002072.

D. Kingma and J. Ba. Adam: A method for stochatic optimization. In *Proceedings of the 3rd International Conference on Learning Representation (ICLR 2015), San Diego.*, 2015.

Johannes Lederer. Statistical guarantees for sparse deep learning. *AStA Advances in Statistical Analysis*, 108:231–258, 2024. doi: 10.1007/s10182-022-00467-3.

Po-Ling Loh and Martin J. Wainwright. Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.*, 16:559–616, 2015. ISSN 1532-4435.

H.M. Markowitz. Portfolio selection. *Journal of Finance*, pages 77–91, 1952.

H.M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments.* John Wiley & Son, 1959.

Lawrence R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *J. Math. Phys.*, 25(8):2404–2417, 1984. ISSN 0022-2488,1089-7658. doi: 10.1063/1.526446. URL https://doi.org/10.1063/1.526446.

Fabian Mies and Ansgar Steland. Sequential Gaussian approximation for nonstationary time series in high dimensions. *Bernoulli*, 29(4):3114 – 3140, 2023. doi: 10.3150/22-BEJ1577. URL https://doi.org/10.3150/22-BEJ1577.

Fabian Mies and Ansgar Steland. Projection inference for high-dimensional covariance matrices with structured shrinkage targets. *Electronic Journal of Statistics*, 18(1): 1643 – 1676, 2024. doi: 10.1214/24-EJS2225. URL https://doi.org/10.1214/24-EJS2225.

Magda Peligrad and Qi Man Shao. Estimation of the variance of partial sums for ρ-mixing random variables. *J. Multivariate Anal.*, 52(1):140–157, 1995. ISSN 0047-259X. doi: 10.1006/jmva.1995.1008. URL https://doi.org/10.1006/jmva.1995.1008.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Herbert Robbins and David Siegmund. Boundary Crossing Probabilities for the Wiener Process and Sample Sums. *The Annals of Mathematical Statistics*, 41(5):1410 – 1429, 1970. doi: 10.1214/aoms/1177696787. URL https://doi.org/10.1214/aoms/1177696787.

Adam J. Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009. doi: 10.1198/jasa.2009.0101.

W. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risks. *Journal of Finance*, pages 425–442, 1964.

Galen R. Shorack and Jon A. Wellner. *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. ISBN 0-471-86725-X.

Ansgar Steland. Testing and estimating change-points in the covariance matrix of a high-dimensional time series. *J. Multivariate Anal.*, 177:104582, 24, 2020. ISSN 0047-259X. doi: 10.1016/j.jmva.2019.104582. URL https://doi.org/10.1016/j.jmva.2019.104582.

Ansgar Steland and Ewaryst Rafajłowicz. Decoupling change-point detection based on characteristic functions: methodology, asymptotics, subsampling and application. *J. Statist. Plann. Inference*, 145:49–73, 2014. ISSN 0378-3758. doi: 10.1016/j.jspi.2013.08.009. URL https://doi.org/10.1016/j.jspi.2013.08.009.

Brajendra C. Sutradhar. On the characteristic function of multivariate Student t-distribution. *Canad. J. Statist.*, 14(4):329–337, 1986. ISSN 0319-5724,1708-945X. doi: 10.2307/3315191. URL https://doi.org/10.2307/3315191.

Halbert White. *Artificial Neural Networks*. Blackwell Publishers, 1992.

Wei Biao Wu. Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102(40):14150–14154, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506715102. URL https://doi.org/10.1073/pnas.0506715102.

Zhuoran Yang, Zhaoran Wang, Han Liu, Yonina Eldar, and Tong Zhang. Sparse nonlinear regression: Parameter estimation under nonconvexity. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2472–2481, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/yangc16.html.

Zhao Zhao, Olivier Ledoit, and Hui Jiang. Risk reduction and efficiency increase in large portfolios: Gross-exposure constraints and shrinkage of the covariance matrix. *Journal of Financial Econometrics*, 2021. doi: 10.1093/jjfinec/nbab001.