# NONLINEAR INDEPENDENT COMPONENT ANALYSIS FOR CONTINUOUS-TIME SIGNALS

HARALD OBERHAUSER AND ALEXANDER SCHELL

*Mathematical Institute*
*University of Oxford*

Abstract. We study the classical problem of recovering a multidimensional source process from observations of nonlinear mixtures of this process. Assuming statistical independence of the coordinate processes of the source, we show that this recovery is possible for many popular models of stochastic processes (up to order and monotone scaling of their coordinates) if the mixture is given by a sufficiently differentiable, invertible function. Key to our approach is the combination of tools from stochastic analysis and recent contrastive learning approaches to nonlinear ICA. This yields a scalable method with widely applicable theoretical guarantees for which our experiments indicate good performance.

## 1. Introduction

A common problem in science and engineering is that an *observed quantity*, $X$, is determined by an *unobserved source*, $S$, which one is interested in. Denoting by $f$ the deterministic relationship between $X$ and $S$, one thus arrives at the equation

$$X = f(S) \tag{1}$$

where $X$ is known but both the relation $f$ and the source $S$ are unknown.

The premise that the data $X$ is determined by its source $S$ reflects in the assumption that $f$ is a deterministic function, while the premise that $S$ can be completely inferred from $X$ — i.e. that no information be lost in the process of going from $S$ to $X$ — is reflected in the assumption that the function $f$ is one-to-one; for simplicity, it is typically also assumed that $f$ is onto. Any function $f$ of this kind will be referred to as a *mixing transformation*.

The central challenge, known as the problem of *Blind Source Separation (BSS)*, then becomes to infer — or 'identify' — the hidden source $S$ from the given data $X$:

(2) *Under which assumptions is it possible to recover the source data $S$ in (1) if only its mixture $X$ is observed? To what extent can such a recovery be achieved and how can it be performed in practice?*

It is clear that without additional assumptions, the above problem of inference (2) is severely underdetermined: If $X$ and equation (1) is the only information available but both $f$ and $S$ are unknown, then we may generally find infinitely many possible 'explanations' $(\tilde{S}, \tilde{f})$ for $X$ which all satisfy (1) but are not otherwise meaningfully related to the *true* explanation $(S, f)$ underlying the data. In many cases, however, this 'indeterminacy of $S$ given $X$ with $f$ unknown' can be controlled by imposing certain *statistical conditions* on the source $S$.

The following simple example illustrates this situation.

*E-mail address*: harald.oberhauser@maths.ox.ac.uk and alexander.schell@maths.ox.ac.uk.
*MSC2020 subject classification:* Primary 62H25, 62M99; secondary 62H05, 60L10, 62M45, 62R10.
*Keywords and Phrases:* Blind Source Separation, Independent Component Analysis, inverse problem, statistical independence, latent variable model, functional data analysis, nonlinear BSS, nonlinear ICA.

HO is supported by the DataSig grant [EP/S026347/1], the Alan Turing Institute, and the Oxford-Man Institute. AS acknowledges support from the Oxford-Cocker Graduate Scholarship. The authors would like to thank Aapo Hyvärinen for helpful comments and suggestions.

*Example* 1.1. Suppose that you are on a video-call and want to follow the simultaneous speeches of two speakers $S^1$ and $S^2$. As the propagation of sound adheres to the superposition principle, the acoustic signals $X^1$ and $X^2$ that reach your left and right ear, respectively, may be modelled as linear mixtures $X^i = a_{i1}S^1 + a_{i2}S^2$ of the individual speech signals $S^1$ and $S^2$. Denoting $X \equiv (X^1, X^2)^\intercal$ and $S \equiv (S^1, S^2)^\intercal$ and $A \equiv (a_{ij}) \in \mathbb{R}^{2\times 2}$, the relation between the audio data $X$ and its underlying sources $S$ can hence be expressed by the model equation $X = A \cdot S$, which for $A$ invertible is a special case of (1) for the linear map $f := A$. The above problem (2) then becomes to recover the constituent speeches $S^1$ and $S^2$ from their observed mixtures $X^1, X^2$ alone, given that the relationship between $X$ and $S$ is linear.

Without further assumptions, the true explanation $(S, A)$ of the data $X$ cannot be distinguished from any of its 'alternative explanations' $\{(\tilde{S}, \tilde{A}) \equiv (B\cdot S, AB^{-1}) \mid B \in \mathbb{R}^{2\times 2} \text{ invertible}\}$. But if the speech signals $S^1$ and $S^2$ were assumed to be uncorrelated, say, then the above family of best-approximations of $(S, A)$ reduced to $\{(\tilde{S}, \tilde{A}) \equiv (B\Lambda \cdot S, A\Lambda^{-1}B^\intercal) \mid \Lambda \in \mathbb{R}^{2\times 2} \text{ (invertible) diagonal}, B \in \mathbb{R}^{2\times 2} \text{ orthogonal}\}$;[1] hence if they are uncorrelated, $S^1$ and $S^2$ may be recovered from $X$ uniquely up to scale and a rotation. ◆

This simple observation can be significantly improved by way of the classical Darmois-Skitovich theorem [21, 77] which implies that for $f$ *linear*, the original source $S$ may be identified from $X$ even up to scaling and a permutation of its components if $S$ is modelled as a random vector whose coordinates $S^i$ are not only uncorrelated but *statistically independent*. This mathematical insight, elaborated in P. Comon's seminal framework [18], quickly became the theoretical foundation of *Independent Component Analysis (ICA)*, a popular statistical method that has since seen far-reaching theoretical investigations and extensions, e.g. [2, 74], and has been successfully implemented in numerous widely-applied algorithms, e.g. [4, 12, 36, 40]; see e.g. [27, 41, 58] as well as the monographs [19, 42] for an overview.

Comon's contribution provided the first and arguably the conceptionally most influential answer to the above inference task (2) to date that was both practically relevant and mathematically rigorous. However, Comon's approach applies to *linear* relationships (1) between $X$ and $S$ only, because among nonlinear mixing functions on $\mathbb{R}^d$ there are many 'non-trivial' transformations that preserve the mutual statistical independence of their input vectors [45]. Given this substantial limitation, the past twenty-six years have seen various attempts of establishing alternative identifiability conditions to recover multivariate data from their nonlinear transformations. Prominent ideas in this direction include the optimisation of mutual information over outputs of (adversarial) neural networks, e.g. [1, 10, 37, 47, 78], or the idea of 'linearising' the generative relation (1) by mapping the observable $X$ into a high-dimensional feature space where it is then subjected to a linear ICA-algorithm [35].

More recently, the works of Hyvärinen et al. [43, 44, 46] achieved significant progress regarding the recovery of nonlinearly mixed sources with *temporal structure* (e.g. time series, instead of random vectors in $\mathbb{R}^d$) by first augmenting the observed mixture of these sources with an auxiliary variable such as time [43] or its history [44], and then training logistic regression to discriminate ('contrast') between the thus-augmented observable and some additional 'variation' of the data. This variation is obtained by augmenting the observable with a randomized auxiliary variable of the same type as before, thus linking the asymptotical recovery of the source $S = f^{-1}(X)$ to a trainable optimisation problem, namely the convergence of a universal function approximator (e.g. a neural network) learning a classification task. These identifiability results were extended and embedded into the context of variational autoencoders in [47].

[1] Indeed: The assumption of uncorrelatedness complements the original model equation (1) by the additional (statistical) source condition $\mathrm{Cov}(\tilde{S}, \tilde{S}) = \mathrm{Cov}(S, S) = \mathrm{I}_2$, which implies that $B^\intercal B = \mathrm{Cov}(\tilde{S}, \tilde{S}) = \mathrm{I}_2$ (where the components of $\tilde{S}$ are assumed to be scaled to unit variance).

Motivated by the classical ICA framework of Comon [18] and the recent contrastive learning breakthrough [44], we revisit the inference problem (2) for stochastic processes $X = (X_t)$ and $S = (S_t)$ with recent tools from stochastic analysis. We believe the following to be our main contributions to the existing literature:

**Identifiability for Stochastic Processes.** We provide general identifiability results that extend Comon's classical identifiability criterion from linear mixtures of random vectors to nonlinear mixtures of continuous-time stochastic processes; cf. Theorems 1, 2, 3.

On a theoretical level, working with infinite-dimensional (i.e. path-valued) random variables poses new challenges that we address by using rough path theory. From an applied perspective, many models are naturally formulated in continuous time rather than in discrete time (e.g. in biology, physics, medicine or finance), which our approach accounts for by naturally including Stochastic Differential Equations (SDEs) in particular. *Nevertheless,* discrete-time models are immediately covered as well since these can be naturally identified with continuous-time processes via piecewise-linear interpolation.

**Signature Cumulants as Contrast Functions.** Unlike for vector-valued data, statistical dependence between stochastic processes can manifest itself inter-temporally, in the sense that different coordinates of the processes may exhibit statistical dependencies over different points in time. To quantify such complex dependency relations, we use so-called signature cumulants [7] as our contrast function. These signature cumulants can be seen as generalising the concept of cumulants from vector-valued data to path-valued data. Analogous to classical cumulants, signature cumulants then provide a graded, parsimonious, and computable quantification of the degree of statistical (in)dependence between stochastic processes. This turns our theoretical identifiability characterization into a usable algorithm based on independence maximization by means of a suitable contrast function; cf. (Corollary 1 and) Theorem 4.

**Consistency With Respect to Time Discretization and Sample Size.** When applying our methodology in practice, the following issues arise: Firstly, although the underlying stochastic model is often formulated in continuous time, in practice one usually has access to time-discretized samples only, often taken over non-equally spaced time grids. Secondly, often only a single (time-discretized) sample path of the process is available rather than many independent realisations, for example, in the classical cocktail party problem. We address both of these issues and show that our method is statistically consistent even if only a single, time-discretized sample path of the observable is given; cf. Theorem 5. This is also the setting in which our experiments are carried out in Section 9.

This article is structured as follows. The exposition of our approach towards the identifiability of nonlinearly mixed independent sources begins by recalling the main results of [18] as conceptional points of reference (Section 2). The core of our identifiability theory is developed in the subsequent two sections: advocating for the incorporation of time as an integral dimension of our source model (Section 3), we show how sources admitting a non-degenerate 'temporal structure' harbour sufficient mathematical richness to encode any nonlinear action performed upon them as a sort of 'intrinsic statistical fingerprint', based on which the constituent relation (1) may then be inverted up to an optimal deviation by maximizing an independence criterion (Section 4). Modifying the underlying idea of proof allows an extension of our approach to sources of various types of statistical regularity (Section 5), including popular time series models, various Gaussian processes and Geometric Brownian Motion (Section 6). The practical applicability of our ICA-method is enabled by a novel independence criterion for time-dependent data (Section 7) that allows for a statistically consistent estimation (Section 8) and is demonstrated in a series of numerical experiments (Section 9).

1.1. **Notation.** Below is some of the notation that we use throughout.

| Symbol | Meaning | Page |
|---|---|---|
| $[k]$ | $:= \{1,\dots,k\}$, and $[k]_0 := [k] \cup \{0\}$ $(k \in \mathbb{N})$. | 4 |
| $S_d$ | $:= \{\tau : [d] \to [d] \mid \tau$ is bijective$\}$; the group of all permutations of $[d]$. | 4 |
| $\Delta_d$ | $:= \{\Lambda = (\lambda_i \cdot \delta_{ij}) \in \mathrm{GL}_d \mid \lambda_1,\dots,\lambda_d \in \mathbb{R}\setminus\{0\}\}$; the group of (real) invertible diagonal $d \times d$ matrices. | 4 |
| $\mathrm{P}_d$ | $:= \{(\delta_{\sigma(i),j})_{i,j\in[d]} \in \mathrm{GL}_d \mid \sigma \in S_d\}$; the $d \times d$ permutation matrices. | 4 |
| $\mathrm{GL}_d$ | $:= \{A \in \mathbb{R}^{d\times d} \mid \det(A) \neq 0\}$; the general linear group of degree $d$ over $\mathbb{R}$. | 5 |
| $\mathrm{M}_d$ | $:= \{M \in \mathrm{GL}_d \mid M = D \cdot P$ for $D \in \mathrm{GL}_d$ diagonal and $P \in \mathrm{P}_d\}$; the group of (real) monomial matrices of degree $d$. | 5 |
| $\pi^I_J$ | the canonical projection from $E_I := \{(u_i)_{i\in I} \mid u_i \in E_i$ for all $i \in I\}$, with $(E_i \mid i \in I)$ some indexed family of sets, onto $E_J$ $(J \subseteq I)$, that is $\pi^I_J((u_i)_{i\in I}) = (u_i)_{i\in J}$ where the tuple-indexation follows the order of $I$ and $J$, resp. The superscript $I$ will be omitted if the domain of $\pi^I_J$ is clear. E.g.: $\pi_{\{1,3,5\}}(x_1, x_2, \cdots, x_6) = (x_1, x_3, x_5)$, and $\pi_i := \pi_{\{i\}}$. | 6 |
| $\Delta_2(\mathbb{I})$ | $:= \{(s,t) \in \mathbb{I}^{\times 2} \mid s < t\}$; the (relatively) open 2-simplex on $\mathbb{I} \times \mathbb{I}$. | 7 |
| $\mathrm{int}(A)$ | the topological interior of a set $A \subseteq \mathbb{R}^d$ (w.r.t. the Euclidean topology). | 8 |
| $J_\varphi$ | $:= \left(\frac{\partial}{\partial x_j}\varphi_i\right)_{ij}$; the Jacobian of $\varphi \equiv (\varphi_1,\cdots,\varphi_d)^\intercal \in C^1(G;\mathbb{R}^d)$. | 8 |
| $\varphi\big\vert_{\tilde{A}}$ | the restriction of a map $\varphi : A \to B$ to a subdomain $\tilde{A} \subseteq A$. | 9 |
| $f_1 \times f_2$ | $: \mathbb{R}^{k_1+k_2} \to \mathbb{R}^{\ell_1} \times \mathbb{R}^{\ell_2}$, $(z_1,\dots,z_{k_1+k_2}) \mapsto (f_1(z_1,\dots,z_{k_1}), f_2(z_{k_1+1},\dots,z_{k_1+k_2}))$; the Cartesian product of $f_1 : \mathbb{R}^{k_1} \to \mathbb{R}^{\ell_1}$ and $f_2 : \mathbb{R}^{k_2} \to \mathbb{R}^{\ell_2}$. | 11 |
| $C^{k,k}(D)$ | $:= \{h : D \to \mathbb{R}^d \mid h \in \mathrm{Diff}^k(G)$ for some open $G \supseteq D\}$, $D \subseteq \mathbb{R}^d$; the set of all $C^k$-invertible transformations on $D$. | 10 |
| $\mathrm{Diff}^k(G)$ | $:= \{h : G \to \mathbb{R}^d \mid h : G \twoheadrightarrow h(G)$ is a $C^k$-diffeomorphism$\}$; the set of all functions $h \in C^k(G;\mathbb{R}^d)$ which are one-to-one with $h^{-1} \in C^k(h(G);\mathbb{R}^d)$, for $G \subseteq \mathbb{R}^d$ some open subset. | 11 |
| $\mathrm{diag}_{i\in[d]}[a_i]$ | $:= (a_i \cdot \delta_{ij})_{ij}$; the diagonal matrix with main diagonal $(a_1,\dots,a_d)$. | 15 |

## 2. Comon's Framework of Linear Independent Component Analysis

Our approach to the Problem of nonlinear Blind Source Separation (2) for stochastic processes can be regarded as a natural extension of Comon's classical identifiability framework [18]. We therefore briefly recall the main results of this framework as conceptional points of reference.

**Theorem 1** (Comon [18, Theorem 11])**.** *Let $S = (S^1,\cdots,S^d)^\intercal$ be a random vector in $\mathbb{R}^d$ with mutually independent, non-deterministic components $S^1,\dots,S^d$ of which at most one is Gaussian. Let further $X = C \cdot S$ for an orthogonal matrix $C \in \mathbb{R}^{d\times d}$. Then for each orthogonal matrix $\theta \in \mathbb{R}^{d\times d}$ it holds that*

$$(\tilde{S}^1,\cdots,\tilde{S}^d) := \theta \cdot X \;=\; \Lambda P \cdot S \quad \textit{for some} \quad (\Lambda, P) \in \Delta_d \times \mathrm{P}_d \qquad (3)$$
$$\textit{if and only if} \quad \tilde{S}^1,\dots,\tilde{S}^d \quad \textit{are mutually independent.}$$

The significance of Theorem 1 is that it characterises—up to *an optimal deviation*, namely their scaling and re-ordering—the independent sources $S^1,\dots,S^d$ underlying an observable linear mixture $X = A\cdot(S^1,\dots,S^d)^\intercal$ as *precisely those transformations* $\theta_\star \cdot X =: (X^1_{\theta_\star},\dots,X^d_{\theta_\star})$ of the data whose components $X^i_{\theta_\star}$ are *mutually independent.*

*Remark* 2.1. (i) The orthogonality constraint of Theorem 1 imposes no loss of generality with regards to general linear mixtures since any invertible linear relation $X = A \cdot S$, $A \in \mathrm{GL}_d$, between $X$ and $S$ can be reduced to an orthogonal one by performing a principal component analysis on $X$.

(ii) The proof of Theorem 1 is based on the remarkable probabilistic fact that any two linear combinations of a family of statistically independent random variables can themselves be statistically independent only if each random variable of this family which has a non-zero coefficient in both of the linear combinations is Gaussian. (A result which is known as the Darmois-Skitovich theorem, see [21, 77].) This accounts for the theorem's somewhat curious 'non-Gaussianity' condition.

Theorem 1 enables the recovery of $S$ from $X$ by way of solving an optimisation problem.

**Corollary 1** ([18]). *Let $X$ and $S$ be as in Theorem 1. Then for any function*[2] $\phi : \mathcal{M}_1(\mathbb{R}^d) \to \mathbb{R}_+$ *such that $\phi(\mu) = 0$ iff $\mu = \mu^1 \otimes \cdots \otimes \mu^d$, it holds that*[3]

$$\left[\arg\min_{\theta\in\Theta} \phi(\theta \cdot X)\right] \cdot X \;\subseteq\; \mathrm{M}_d \cdot S \tag{4}$$

*where* $\mathrm{M}_d := \{\Lambda \cdot P \mid (\Lambda, P) \in \Delta_d \times \mathrm{P}_d\}$ *is the subgroup of* monomial matrices *and* $\Theta \subset \mathrm{GL}_d$ *is the subgroup of orthogonal matrices.*

In other words: For $f$ linear and $S = (S^1, \cdots, S^d)^\intercal$ a random vector with mutually independent, non-Gaussian components, the constituent relationship (1) between the observable $X$ and its source $S$ can be inverted (up to a minimal deviation) by optimizing some *independence criterion* $\phi$ over a set of candidate transformations $\Theta$ applied to $X$.

Partially driven by their applicability (4) to ICA, a variety of such criteria $\phi$, referred to in [18] as a *contrast functions*, have been developed.

The 'original' independence criterion $\phi_c$ proposed in [18] quantifies the statistical dependence between the components $Y^i$ of a random vector $Y = (Y^1, \cdots, Y^d)$ in $\mathbb{R}^d$ via the sum of the squares of all *standardized cross-cumulants* $\kappa^Y_{i_1\cdots i_j}$ of $Y$ up to $r^{\text{th}}$-order (see [18, Sect. 3.2] and cf. (191)), i.e. via the quantity

$$\phi_c(Y) := \sum_{j=2}^{r} \sum_{i_1,\ldots,i_j}^{\times} (\kappa^Y_{i_1\cdots i_j})^2 \qquad (r \geq 2) \tag{5}$$

(where the inner sum runs over the indices $i_1, \ldots, i_j \in [d]$ corresponding to (192)).
Initially proposed in [18], the statistic (5) originates from a truncated Edgeworth-expansion of mutual information in terms of the standardized cumulants of its argument.

A variety of alternatives to (5) soon followed, including kernel-based independence measures [2, 33], a variety of (quasi-) maximum-likelihood objectives, e.g. [3, 59, 64], as well as mutual information and approximations thereof, e.g. [11, 18, 38, 39].

While successfully achieving the separability of linear mixtures, Theorem 1 has its limitations: being based on somewhat of a probabilistic curiosity (Rem. 2.1 (ii)), it might not be surprising that the characterisation (3) cannot be generalised to guarantee the recovery of independent scalar sources from substantially more general *nonlinear* mixtures of them [45]. Roughly speaking, the reason for this is that for a single random vector in $\mathbb{R}^d$, the statistical property

[2] Here and in the following, $\mathcal{M}_1(V) := \{\mu : \mathcal{B}(V) \to [0, 1] \mid \mu$ is a (Borel) probability measure$\}$ denotes the space of probability measures over the Borel $\sigma$-algebra $\mathcal{B}(V) := \sigma(\mathcal{T})$ of a topological space $(V, \mathcal{T})$.

[3] We write $\mu^i := \mu \circ \pi_i^{-1}$ for the $i^{\text{th}}$ marginal of a (Borel) measure $\mu$ on $\mathbb{R}^d$. We further abuse notation by writing $\phi(Z) := \phi(\mathbb{P}_Z)$ for any random vector $Z : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

of componental independence is *too weak* to characterise the nonlinear mixing transformations preserving this property as 'trivial'[4]. The following example illustrates this.[5]

*Example* 2.1 (Comon's Criterion (3) Does Not Apply to Nonlinearly Mixed Vectors in $\mathbb{R}^d$). Let $S^1$ and $S^2$ be *independent* with $S^1$ Rayleigh-distributed of scale 1 and $S^1$ uniformly distributed over $(-\pi, \pi)$, and consider the nonlinear mixing transformation $f$ given by $f(u, v) := (u\cos(v), u\sin(v))$ (transformation from polar to Cartesian coordinates). Then even though their functional relation $f$ to $S^1, S^2$ is 'non-trivial', i.e. $f$ is not *monomial* in the sense of Definition 5, the mixed variables $X^1$ and $X^2$ defined by $(X^1, X^2) := f(S^1, S^2)$ are (normally distributed and) *statistically independent*.[6] ◆

## 3. Modelling Sources as Stochastic Processes

A central direction along which the 'blind recovery' of the source $S$ from its nonlinear mixture $X$ can be controlled is the amount of statistical structure that $S$ carries: If the source $S$ is deterministic, then no additional information is given and a meaningful recovery of $S$ from $X$ is generally impossible (cf. Example 1.1). If, on the other hand, the source $S$ were to be described merely as a random vector in $\mathbb{R}^d$, then a recovery of $S$ from $X$ is possible but in general only if $X$ is a linear function of $S$ (cf. [18, 45], Example 2.1). A key insight from [44] is to go for the middle ground: demanding the source $S$ to have a 'non-degenerate *temporal* structure' and exploiting it, the recovery of $S$ from even its nonlinear mixtures can be possible. To formalize such temporal statistical dependencies requires us to model the source $S$ as a stochastic process. In this section, we to this end briefly recall basic notions from stochastic analysis and provide some lemmas that we will use for our identifiability results in Section 4.

### 3.1. **Stochastic Processes Interpolate Statistical Extremes.**

Here and throughout, let $\mathbb{I}$ be a compact interval, $d \in \mathbb{N}$ be some fixed integer, write $\mathcal{C}_d \equiv C(\mathbb{I}; \mathbb{R}^d) := \{x : \mathbb{I} \to \mathbb{R}^d \mid \text{the map } \mathbb{I} \ni t \mapsto x(t) =: x_t \text{ is continuous}\}$ for the space of continuous paths in $\mathbb{R}^d$, and let $(\Omega, \mathscr{F}, \mathbb{P})$ denote a fixed probability space.

*Definition* 1 (Source Model). We call a *continuous stochastic process in* $\mathbb{R}^d$ any map

$$S : \Omega \to \mathcal{C}_d \quad \text{s.t.} \quad \omega \mapsto S(\omega) \equiv (S_t(\omega))_{t\in\mathbb{I}} \quad \text{is} \quad (\mathscr{F}, \mathcal{B}(\mathcal{C}_d))\text{-measurable},$$

where $\mathcal{B}(\mathcal{C}_d) = \sigma(\pi_t \mid t \in \mathbb{I})$ denotes the Borel $\sigma$-algebra on the Banach space $(\mathcal{C}_d, \|\cdot\|_\infty)$. Writing $S_t(\omega) \equiv (S_t^1(\omega), \cdots, S_t^d(\omega))^\intercal \in \mathbb{R}^d$ for each $\omega \in \Omega$, the scalar processes $S^i \equiv (S_t^i)_{t\in\mathbb{I}}$ ($i \in [d]$) are called *the coordinate processes* or *the components* of $S \equiv (S^1, \cdots, S^d)$. We say that a stochastic process $S = (S^1, \cdots, S^d)$ *has independent components*, or that $S$ *is IC*, if its distribution $\mathbb{P}_S := \mathbb{P} \circ S^{-1}$ satisfies the factor-identity[7]

$$\mathbb{P}_{(S^1,\cdots,S^d)} = \mathbb{P}_{S^1} \otimes \cdots \otimes \mathbb{P}_{S^d}. \tag{6}$$

*Remark* 3.1. From a more local perspective, Definition 1 is equivalent to the description of a continuous stochastic process $S$ as an $\mathbb{I}$-indexed family $S = (S_t)_{t\in\mathbb{I}}$ of random vectors[8] $S_t$ in

---

[4] In a sense made precise by Definition 5 below.

[5] Ex. 2.1 is based on the 'Box-Muller transform', a well-known subroutine from computational statistics. For a systematic procedure of constructing 'unidentifiable' nonlinear mixtures of IC random vectors in $\mathbb{R}^d$, see [45].

[6] Note that since the density $p_S$ of $(S^1, S^2)$ reads $p_S(s_1, s_2) = \frac{1}{2\pi} s_2 e^{-s_2^2/2}$, the (joint) density $p_X = (p_S \circ f) \cdot |\det J_f|^{-1}$ of $(X^1, X^2)$ factorizes, implying the independence of $X^1$ and $X^2$ as claimed.

[7] Strictly speaking, (6) reads $\mathbb{P}_{(S^1,\cdots,S^d)} = \big(\mathbb{P}_{S^1} \otimes \cdots \otimes \mathbb{P}_{S^d}\big) \circ \psi^{-1}$, an identity of measures on $\mathcal{B}(\mathcal{C}_d)$, where $\psi : \mathcal{C}_1^{\times d} \to \mathcal{C}_d$ is a canonical isometry defining the Cartesian identification $\mathcal{C}_d \cong \mathcal{C}_1^{\times d}$ (Rem. A.1).

[8] For us every random vector in $\mathbb{R}^d$ is Borel, i.e. $(\mathscr{F}, \mathcal{B}(\mathbb{R}^d))$-measurable.

$\mathbb{R}^d$ such that the map $S(\omega) : \mathbb{I} \ni t \mapsto S_t(\omega) \in \mathbb{R}^d$ is continuous for each $\omega \in \Omega$; e.g. [70, Sect. II.27]. Consequently (also Remark A.1), the independence condition (6) is equivalent to

$$\big(S^1_{t^{(1)}_1}, \cdots, S^1_{t^{(1)}_{k_1}}\big), \big(S^2_{t^{(2)}_1}, \cdots, S^2_{t^{(2)}_{k_2}}\big), \cdots, \big(S^d_{t^{(d)}_1}, \cdots, S^d_{t^{(d)}_{k_d}}\big) \quad \text{mutually } \mathbb{P}\text{-independent}$$

for any finite selection of time-points $t^{(1)}_1, \ldots, t^{(1)}_{k_1}, \ldots, t^{(d)}_1, \ldots, t^{(d)}_{k_d} \in \mathbb{I}$, $k_1, \ldots, k_d \in \mathbb{N}_0$.

Stochastic processes can be given a prominent role in the BSS-context, namely as *natural interpolants* between deterministic signals and random vectors. While the first type of signal is the unidentifiable default model for the source in (1), the latter is the predominant source model in classical ICA-approaches. More specifically, the following is easy to see.

*Remark* 3.2 (Stochastic Processes Interpolate Between Extremal Source Models).
Let $S = (S_t)_{t\in\mathbb{I}}$ be a continuous stochastic process in $\mathbb{R}^d$ such that either

(a) $S_s$ and $S_t$ are independent for each $s, t \in \tilde{\mathbb{I}}$ with $s \neq t$, or
(b) $S_s = S_t$ almost surely for each $s, t \in \tilde{\mathbb{I}}$ with $s \neq t$, for some $\tilde{\mathbb{I}} \subset \mathbb{I}$ dense.

Then $S$ is either a single path in $\mathcal{C}_d$ almost surely (i.e. $S$ is deterministic; *'statistically trivial'*)[9] namely iff (a) holds, or the sample-paths of $S$ are constant almost surely (i.e. $S$ is a random vector; *'temporally trivial'*) namely iff (b) holds.

Remark 3.2 asserts that both deterministic signals (a) as well as random vectors (b) can be seen as degenerate stochastic processes, and that for a given stochastic process $S = (S_t)_{t\in\mathbb{I}}$ this degeneracy manifests on the level of its *$2^{nd}$-order finite-dimensional distributions*, i.e. on

$$\text{the distributions of} \qquad \big\{(S_s, S_t) \;\big|\; (s,t) \in \Delta_2(\mathbb{I})\big\} \tag{7}$$

where the index set $\Delta_2(\mathbb{I}) := \{(s,t) \in \mathbb{I}^{\times 2} \mid s < t\}$ is the (relatively) open 2-simplex on $\mathbb{I} \times \mathbb{I}$. In the following, we refer to (7) as the *temporal structure* of a stochastic process $S = (S_t)_{t\in\mathbb{I}}$.

As mentioned above and illustrated in the next section, if the temporal structure of the IC source $S$ in (1) is *degenerate* in the sense of Remark 3.2 (a), (b), then $S$ is *unidentifiable* from $X$ unless $f$ is of a very specific form, e.g. linear (cf. Theorem 1). Conversely, we will argue that if the source $S$ has a temporal structure which is *non-degenerate* and satisfies some additional regularity assumptions, then $S = (S^1, \cdots, S^d)$ will be *identifiable* from even its nonlinear mixtures up to a permutation and monotone scaling of its components $S^i$ (Theorems 2, 3, 4).

3.2. **Sources as Stochastic Processes: Basic Notions and Assumptions.** Recall that the problem of BSS (2) concerns *the recovery of the source $S$ from its image $X$ under some mixing transformation $f$ on $\mathbb{R}^d$*. It is thus clear that given $X$, the transformation $f$ can be analysed only on *that part* of its domain that is actually reached by $S$ during the time $X$ is observed. With this in mind, we introduce the *spatial support* of a stochastic process as the smallest closed subset of $\mathbb{R}^d$ on which this process is 'spread' during its evolution.[10]

*Definition* 2 (Spatial Support). For $Y = (Y_t)_{t\in\mathbb{I}}$ a (continuous) stochastic process in $\mathbb{R}^d$, the *spatial support* of $Y$ is defined as the set

$$D_Y = \overline{\bigcup_{t\in\mathbb{I}} \operatorname{supp}(Y_t)}^{|\cdot|} \tag{8}$$

with $\operatorname{supp}(Y_t) \equiv \operatorname{supp}(\mathbb{P}_{Y_t}) =: D_{Y_t}$ denoting the support of the distribution of $Y_t$, and where the closure is taken w.r.t. the Euclidean topology on $\mathbb{R}^d$.

[9] This implication is obtained from Kolmogorov's zero-one law (applied after a straightforward subsequence argument) and the sample continuity of $S$.

[10] Analogous to how the support $D_Z := \operatorname{supp}(Z)$ of a random vector $Z$ in $\mathbb{R}^d$ is the smallest closed subset of $\mathbb{R}^d$ within which $Z$ is contained with probability one.

(Readers uncomfortable with (8) may for simplicity assume that $D_S = \mathbb{R}^d$ throughout.)

The following elementary properties of the set (8) will be useful to us.

**Lemma 1.** *Let $Y = (Y_t)_{t\in\mathbb{I}}$ be a stochastic process in $\mathbb{R}^d$ which is continuous with spatial support $D_Y$. Then the following holds:*

(i) *if $f : D_Y \to \mathbb{R}^d$ is a homeomorphism onto $f(D_Y)$, then $f(D_Y) = D_{f(Y)}$;*
(ii) *the traces $\operatorname{tr}(Y(\omega)) := \{Y_t(\omega) \mid t \in \mathbb{I}\}$ are contained in $D_Y$ for $\mathbb{P}$-almost each $\omega \in \Omega$;*
(iii) *for each open subset $U$ of $D_Y$ there is some $t^\star \in \mathbb{I}$ with $\mathbb{P}(Y_{t^\star} \in U) > 0$;*
(iv) *if each random vector $Y_t$, $t \in \mathbb{I}$, admits a continuous Lebesgue density on $\mathbb{R}^d$, then $D_Y$ is the closure of its interior.*
(v) *if each random vector $Y_t$, $t \in \mathbb{I}$, admits a density $\upsilon_t$ such that the function $\upsilon^x : \mathbb{I} \ni t \mapsto \upsilon_t(x)$ is continuous for each $x \in D_Y$, we for $\dot{D}_t := \{\upsilon_t > 0\}$ have that the set*

$$\bigcup_{(s,t)\in\Delta_2(\mathbb{I})} \dot{D}_s \cap \dot{D}_t \quad \textit{is} \quad \textit{dense} \quad \textit{in} \quad D_Y.$$

*Proof.* See Appendix A.2. □

Given the above, we can describe the mixing transformation $f$ sending $S$ to $X$ via (1) as

$$\text{a homeomorphism}^{11} \quad f : D_S \to D_X, \tag{9}$$

with the action of $f$ outside of $D_S$ and $D_X$ being irrelevant (and inaccessible) to us.

We now introduce smoothness conditions on the density which we require later on.

*Definition* 3. A random vector $Z$ in $\mathbb{R}^n$ will be called $C^k$*-distributed*, $k \in \mathbb{N}_0$, if its distribution admits a Lebesgue density $\varsigma \in C^k(G)$ for $G := \operatorname{int}(\operatorname{supp}(\varsigma))$; if $\varsigma$ is $C^1$ on some open neighbourhood of $x_0 \in \mathbb{R}^n$, then $Z$ will be called $C^k$*-distributed around* $x_0$.

*Remark* 3.3. We recall that for $\vartheta : \mathbb{R}^n \to \mathbb{R}^n$ a $C^\ell$-diffeomorphism, the classical *transformation formula* for densities asserts that the image $\tilde{Z} := \vartheta(Z)$ of a $C^k$-distributed random vector $Z$ with density $\varsigma$ is itself $C^{k\wedge\ell}$-distributed with density $\tilde{\varsigma}$ given by

$$\tilde{\varsigma} = (\varsigma \circ \vartheta^{-1}) \cdot |\det J_{\vartheta^{-1}}|. \tag{10}$$

The action of the mixing transformation (9) on the source $S$ can be profitably captured by imposing the temporal structure (7) of $S$ to meet the following analytical regularity condition:

In the following, a stochastic process $Y = (Y_t)_{t\in\mathbb{I}}$ in $\mathbb{R}^d$ will be called

$$C^k\textit{-regular at } (s,t) \in \Delta_2(\mathbb{I}) \quad \text{if} \quad \text{the random vector } (Y_s, Y_t) \text{ is } C^k\text{-distributed;}$$

the process $Y$ will be called $C^k$*-regular at* $((s,t), y_0) \in \Delta_2(\mathbb{I}) \times \mathbb{R}^{2d}$ if the random vector $(Y_s, Y_t)$ is $C^k$-distributed around $y_0 \in \mathbb{R}^n$ and its density at $y_0$ is positive.

*Remark* 3.4. Note that if $Y$ is $C^k$-regular at $(s,t)$, then the boundary of the support of the joint density of $(Y_s, Y_t)$ is a Lebesgue nullset. (A direct consequence of Sard's theorem.)

The theory of ICA knows two prominent 'exceptional cases' for which the recovery of an IC random vector $S$ in $\mathbb{R}^d$ from even its linear mixtures $X$ cannot be guaranteed without further assumptions, namely the cases in which

(i) more than one of the components of $S$ is Gaussian (cf. Theorem 1), or

[11] Note that while the assumption of invertibility of $f$ is canonical, the additionally imposed bi-continuity of the mixing transformation $f$ is a technical condition to ensure that the sample-continuity of the considered processes is preserved under any of the operations that follow.

(ii) the source $S$ is 'statistically trivial' in the sense of Remark 3.2 (a).

As it turns out, a generalised version of these pathologies carries over to the first and more 'static' of our separation principles (Theorem 2), owing to the fact that certain analytical forms of the joint distributions constituting (7) will be 'too simple' to guarantee nonlinear identifiability even for sources whose temporal structure (7) is not otherwise degenerate.

Generalising (i) and (ii) from 'spatial' to 'inter-temporal statistics', these exceptional types of joint distributions[12] will be named 'pseudo-Gaussian' and 'separable', respectively:

*Definition* 4 (Non-Gaussian, (Regularly) Non-Separable). A function $\varsigma : G \to \mathbb{R}$, $G \subseteq \mathbb{R}^2$ open, will be called *pseudo-Gaussian* if there are functions $\varsigma_1, \varsigma_2, \varsigma_3 : \mathbb{R} \to \mathbb{R}$ for which

$$\varsigma(x, y) = \varsigma_1(x) \cdot \varsigma_2(y) \cdot \exp(\pm\, \varsigma_3(x) \cdot \varsigma_3(y))$$

holds on all of $G$; the function $\varsigma$ will be called *separable* if the above holds for $\varsigma_3 \equiv 0$. The function $\varsigma : G \to \mathbb{R}$ will be called *strictly non-Gaussian* if it is such that

$$\varsigma|_{\mathcal{O}} \quad \text{is not pseudo-Gaussian} \quad \text{for every open subset } \mathcal{O} \text{ of } G\,;$$

the property of $\varsigma$ being *strictly non-separable* is declared mutatis mutandis. Furthermore, the function $\varsigma : G \to \mathbb{R}$ will be called *almost everywhere non-Gaussian* if

$$\text{there is a closed nullset} \quad \mathcal{N} \subset U \quad \text{s.t.} \quad \varsigma|_{(U \setminus \mathcal{N})} \quad \text{is strictly non-Gaussian}\,;$$

the notion of $\varsigma$ being *a.e. non-separable* is defined analogously.
Finally, a twice continuously differentiable function $\varsigma : \tilde{U} \times \tilde{U} \to \mathbb{R}_{>0}$, with $\tilde{U} \subseteq \mathbb{R}$ open, will be called *regularly non-separable* if

$$\varsigma \ \text{ is a.e. non-separable} \quad \text{and} \quad \big(\partial_x\partial_y \log \varsigma\big)\big|_{\Delta_{\tilde{U}}} \neq 0 \ \text{ a.e. on } \Delta_{\tilde{U}} \tag{11}$$

where $\Delta_{\tilde{U}} := \{(x, x) \mid x \in \tilde{U}\}$ denotes the diagonal over $\tilde{U}$.

(Clearly, if $\varsigma$ is [strictly/a.e.] non-Gaussian then it is also [strictly/a.e.] non-separable.)

*Remark* 3.5. (i) In light of Lemma 2 (ii), the assumption of regular non-separability can be regarded as a *minimal* extension of the above notion of strict non-separability. The necessity to complement our assumptions on the source's temporal structure (7) by this extension will become clear in Section 4.2.
(ii) The log-derivative condition of (11) is non-vacuous as there are (strictly) non-separable functions whose mixed log-derivatives vanish on the diagonal, see Example A.4.

It will be convenient to have an analytical characterisations of these 'pathological' types of densities (Lemma 2). To this end, we first declare what we mean by a symmetric set:

Writing $(u, v) \equiv (u_1, \cdots, u_d, v_1, \cdots, v_d)$ for the coordinates on $\mathbb{R}^{2d} \cong \mathbb{R}^d \times \mathbb{R}^d$, a given subset $A \subseteq \mathbb{R}^{2d}$ will be called *symmetric* if

$$A = \tau(A) \qquad \text{for the transposition} \qquad \tau(u, v) := (v, u).$$

A function $\varphi : G \to \mathbb{R}$, $G \subseteq \mathbb{R}^{2d}$, will be called *symmetric* if $\varphi \circ \tau = \varphi$.

(Since $\tau^2 = \mathrm{id}$, it is clear that $A$ is symmetric iff $\tau(A) \subseteq A$. Also, if $A \subseteq \mathbb{R}^{2d}$ is symmetric then $A \subseteq \pi_{[d]}(A) \times \pi_{[d]}(A)$.)

**Lemma 2.** *Let $\zeta : G \to \mathbb{R}_{>0}$, with $G \subseteq \mathbb{R}^2$ open, be twice continuously differentiable. Then the following holds.*

[12] Similar distributional pathologies have been first described in [44]. More specifically, the above notions of (strict) non-separability and pseudo-Gaussianity generalise the notions [44, Def. 1 and Def. 2], respectively.

(i) *Provided that $G$ is convex, we have that:*

$$\partial_x\partial_y \log\zeta \equiv 0 \qquad \textit{if and only if} \qquad \zeta \ \ \textit{is separable};$$

(ii) $\zeta$ *is strictly non-separable if and only if the open set*

$$G' := \{z \in G \mid \partial_x\partial_y \log\zeta(z) \neq 0\} \ \ \textit{is a dense subset of } G;$$

(iii) *provided that $\mathcal{O} \subseteq G'$ is symmetric, open and convex, we have that:*

$$[\partial_x\partial_y \log\zeta]\big|_{\mathcal{O}} \ \ \textit{is separable and symmetric} \quad \textit{iff} \quad \zeta\big|_{\mathcal{O}} \ \ \textit{is pseudo-Gaussian.}$$

*Proof.* We use the global abbreviations $\xi := \partial_x\partial_y \log\zeta$ and $\phi := \log\zeta$.

(i): The 'if'-direction is clear, so suppose that $\partial_x\partial_y\phi = 0$. Then, as $G$ is convex, $\phi \equiv \phi(x,y) = \phi_1(x) + \phi_2(y)$ and hence $\zeta = \exp(\phi) = \zeta_1(x)\cdot\zeta_2(y)$ for $\zeta_i := \exp(\phi_i)$, as claimed.

(ii): Since $\xi$ is continuous, the set $\{\xi = 0\}$ is closed, whence the set $G' = G \cap \{\xi = 0\}^c$ is open. To see that $G'$ is dense in $G$, take any $z \in G$ and note that, as $G$ is open, there is some $z$-centered open ball $B_z \subseteq G$. Since $\zeta$ is strictly non-separable, $\zeta|_{B'_z}$ is not separable for *any* open $z$-centered sub-ball $B'_z \subseteq B_z$, whence by (i) there must be some $z' \in B'_z$ with $\xi(z') \neq 0$, implying $B'_z \cap G' \neq \emptyset$.

The (contrapositive of the) converse implication in (ii) follows via (i).

(iii): Let $\mathcal{O} \subseteq G'$ be symmetric, open and convex. ($\Leftarrow$) is clear by Def. 4.

($\Rightarrow$): Suppose that $\tilde{\xi} := \xi\big|_{\mathcal{O}}$ is separable and symmetric, i.e. assume that

$$\tilde{\xi} \equiv \tilde{\xi}(x,y) = f(x)\cdot g(y) \quad \text{and} \quad \tilde{\xi}\circ\tau = \tilde{\xi}$$

for some $f, g : \mathcal{O}_1 \to \mathbb{R}$, with $\mathcal{O}_1 := \pi_1(\mathcal{O})$. Then $\tilde{\xi} \equiv \tilde{\xi}(x,y) = \operatorname{sgn}(\tilde{\xi})\cdot\eta(x)\cdot\eta(y)$ for some function $\eta : \mathcal{O}_1 \to \mathbb{R}$, where $\epsilon := \operatorname{sgn}(\tilde{\xi})$ denotes the sign of $\tilde{\xi}$ (i.e., $\operatorname{sgn}(\tilde{\xi}) = \mathbb{1}_{(0,\infty)}(\tilde{\xi}) - \mathbb{1}_{(-\infty,0)}(\tilde{\xi})$). Indeed, the symmetry of $\tilde{\xi}$ implies that $\tilde{\xi}^2 = \tilde{\xi}(x,y)\cdot\tilde{\xi}(y,x) = \tilde{\eta}(x)\cdot\tilde{\eta}(y)$ for the map $\tilde{\eta}(z) := f(z)\cdot g(z)$; consequently, $\tilde{\xi} = \epsilon\cdot\sqrt{\tilde{\xi}^2} = \epsilon\cdot\eta(x)\cdot\eta(y)$ for the map $\eta \equiv \eta(z) := \sqrt{|\tilde{\eta}(z)|}$. Now since $\mathcal{O}$ is a connected subset of $G'$, the sign of $\tilde{\xi}$ is constant, i.e. $\epsilon = \pm 1$. Integrating $\xi = \partial_x\partial_y\phi$ thus implies that

$$\begin{aligned} \phi \ &= \int \partial_y\phi \,\mathrm{d}y + f_1(x) = \iint \tilde{\xi}\,\mathrm{d}x + f_2(y)\,\mathrm{d}y + f_1(x) \\ &= \epsilon\cdot f_3(x)\cdot f_3(y) + \tilde{f}_2(y) + \tilde{f}_1(x) \end{aligned} \tag{12}$$

for $f_3 \equiv f_3(z) := \int_{z_0}^{z}\eta(s)\,\mathrm{d}s$ (some priorly fixed $z_0 \in \mathcal{O}_1$) and some additional continuous functions $f_i, \tilde{f}_i : \mathcal{O}_1 \to \mathbb{R}$. Note that since $\mathcal{O}$ is convex, the integrated identity of functions (12) holds pointwise on all of $\mathcal{O}$. Exponentiating (12) now yields the claim. □

## 4. An Identifiability Theorem for Nonlinearly Mixed Independent Sources

We now present the core of our identifiability result for nonlinearly mixed independent sources with non-degenerate temporal structure.

Throughout, let $S$ and $X$ be two continuous stochastic processes in $\mathbb{R}^d$ that are related via

$$X = f(S) \tag{13}$$

for a mixing transformation $f$ which is $C^3$-invertible on some open superset of $D_S$.

Here, a map $f : \mathbb{R}^d \to \mathbb{R}^d$ will be called $C^k$-*invertible* on an open set $G \subseteq \mathbb{R}^d$ , in symbols: $f \in C^k(G)$, if $f\big|_G$ is a $C^k$-diffeomorphism (with $C^k$-inverse $f^{-1} : f(G) \to G$).

4.1. **Overview.** Starting from (13) with the coordinates $(S_t^1)_{t\in\mathbb{I}}, \ldots, (S_t^d)_{t\in\mathbb{I}}$ of the source $S = (S_t^1, \cdots, S_t^d)_{t\in\mathbb{I}}$ assumed mutually independent, we seek to identify $S$ from $X$ by exploiting the main dimensions of our model, *space* and *time*, by way of their statistical synthesis (7), the *temporal structure* of $S$. This will be done along the following lines.

Given $(s,t) \in \Delta_2(\mathbb{I})$, we first double the available spatial degrees of freedom by *lifting* the mixing identity (13) to an *associated identity in the factor-space* $\mathbb{R}^d \times \mathbb{R}^d$ via

$$(X_s, X_t) = (f \times f)(S_s, S_t). \tag{14}$$

The lifted mixing identity (14), which directly involves the temporal structure (7) of the source, now allows for the following *statistical comparison* in the spirit of [44]:

For $X_t^*$ an independent copy of $X_t$, consider the *intertemporal features* $Y := (X_s, X_t)$ and $Y^* := (X_s, X_t^*)$ of the observable $X$ at $(s,t)$ together with their random convex combination

$$\bar{Y} := C \cdot Y + (1 - C) \cdot Y^*$$

for an equiprobable $\{0,1\}$-valued random variable $C$ independent of $Y, Y^*$. Combining (14) with the fact that $S$ is IC, we for the (deterministic) functional $L(Y, Y^*) := \psi \circ \rho$ with $\rho(y) := \mathbb{E}[C \mid \bar{Y} = y]$ and $\psi(p) := \log(p/(1-p))$ obtain a *contrast identity* of the form

$$L(Y, Y^*) = R(f, (S_s, S_t)) \tag{15}$$

for a function $R \equiv R(f, (S_s, S_t))$ which depends exclusively on $f$ and *the distribution* of $(S_s, S_t)$. In other words, (15) relates $X$ to $S$ by way of the source's temporal structure (7).

Since the LHS $L \equiv L(Y, Y^*)$ in (15) is a function of the (joint) distribution of $(Y, Y^*)$—and thus of the observable data $X$—only, we for *any alternative pair* $(\tilde{f}, \tilde{S})$ with $\tilde{f}(\tilde{S}) = X$ and $\tilde{f} \in C^3$ and $\tilde{S}$ IC analogously obtain that $L(Y, Y^*) = \tilde{R}(\tilde{f}, (\tilde{S}_s, \tilde{S}_t))$ and hence

$$\tilde{R}(\tilde{f}, (\tilde{S}_s, \tilde{S}_t)) = R(f, (S_s, S_t)) \tag{16}$$

by (15), where again $\tilde{R} \equiv \tilde{R}(\tilde{f}, (\tilde{S}_s, \tilde{S}_t))$ is some function which depends only on $\tilde{f}$ and the distribution of $(\tilde{S}_s, \tilde{S}_t)$. Using the $C^3$-invertibility of $\tilde{f}$, the IC-properties of both $\tilde{S}$ and $S$ allow us to from (16) via (10) derive a (deterministic) system of functional equations

$$\Gamma(\varrho, (\tilde{S}_s, \tilde{S}_t), (S_s, S_t)) = 0 \qquad \text{for} \qquad \varrho := (\tilde{f}^{-1} \circ f)\Big|_{D_S} \tag{17}$$

which involves the partial derivatives of the 'mixing residual' $\varrho$ and is otherwise completely determined by the distributions of $(\tilde{S}_s, \tilde{S}_t)$ and $(S_s, S_t)$.

The assumed distributional properties of $(S_s, S_t)$, i.e. the temporal structure of $S$ as specified by Definition 6, together with the required IC-property of $\tilde{S}$ are then sufficient to from (17) infer that the residual $\varrho$ must be *monomial* in the sense of Definition 5.

In other words, we obtained that for every $C^3$-invertible map $\tilde{f}$ it holds that

$$\begin{gathered}(\tilde{S}^1, \cdots, \tilde{S}^d) \equiv \tilde{S} = \tilde{f}^{-1}(X) \;=\; \big[P \circ (h_1 \times \cdots \times h_d)\big](S) \\ \text{for some } P \in \mathrm{P}_d \;\text{ and monotone }\; h_1, \ldots, h_d \qquad \textit{if and only if} \\ \text{the coordinate processes} \qquad \tilde{S}^1, \ldots, \tilde{S}^d \qquad \text{are mutually independent.}\end{gathered} \tag{18}$$

The characterisation (18), formulated as Theorem 2, can thus be read as *a natural extension* of Comon's classical independence criterion (3) to nonlinear mixtures of IC stochastic processes whose temporal structure is sufficiently regular.

Additional source conditions that qualify $S$ for the characterisation (18) are obtained by 'unfreezing' the above time pair $(s,t) \in \Delta_2(\mathbb{I})$, see Theorem 3.

Analogous to how Comon's criterion (3) became practically applicable by way of (4), our extended criterion (18) is clearly equivalent to the optimisation-based procedure (cf. Thm. 4)

$$\left[\arg\min_{\tilde{g}\in\Theta} \phi(\tilde{g}\cdot X)\right]\cdot X \;\subseteq\; \mathrm{DP}_d\cdot S \quad \text{for any} \quad \phi:\mathcal{M}_1(\mathcal{C}_d)\to\mathbb{R}_+ \\ \text{such that:} \quad \phi(\mu)=0 \quad \text{iff} \quad \mu=\mu^1\otimes\cdots\otimes\mu^d, \tag{19}$$

for $\Theta$ some 'large enough' family of $C^3$-invertible candidate transformations, and $\mathrm{DP}_d$ a nonlinear analogon of the family of monomial matrices $\mathrm{M}_d$ (Definition 5).

Based on a 'moment-like' coordinate description for (the distribution of) stochastic processes, we propose an efficiently computable such $\phi$ that generalises Comon's original contrast (5) from random vectors to stochastic processes (Section 7).

4.2. **Main Theorem.** This section forms the heart of our identifiability results.

We seek to recover the source $S=(S^1,\cdots,S^d)$ from its nonlinear mixture $X$ in (13) up to a *minimal deviation*, namely a permutation and monotone scaling of its coordinates $S^1,\ldots,S^d$.

The following generalisation of the family of monomial matrices makes this precise.

*Definition* 5 (Monomial Transformations). Given a subset $G$ of $\mathbb{R}^d$, a map $\varrho:\mathbb{R}^d\to\mathbb{R}^d$ will be called *monomial on* $G$ if for each connected component $\tilde{G}$ of $G$ we have that

$$\varrho\big|_{\tilde{G}} \;=\; P\circ(h_1\times\cdots\times h_d) \quad \text{for} \quad P\in\mathrm{P}_d \quad \text{and} \quad h_i\in\mathrm{Diff}^1(\pi_i(\tilde{G})).$$

(The above differentiability condition is considered void at isolated points of $\pi_i(\tilde{G})$.) We write $\mathrm{DP}_d(G)$ for the family of all functions on $\mathbb{R}^d$ which are monomial on $G$.

Accordingly, we say that any two paths $\tilde{x}$ and $x$ in $\mathcal{C}_d$ coincide *up to a permutation and monotone scaling of their coordinates*, in symbols:

$$\tilde{x} \;\in\; \mathrm{DP}_d\cdot x, \tag{20}$$

if $(\tilde{x}_t)_{t\in\mathbb{I}}=(\varrho(x_t))_{t\in\mathbb{I}}$ for some $\varrho\in\mathrm{DP}_d(\mathrm{tr}(x))$, where $\mathrm{tr}(x)\equiv\bigcup_{t\in\mathbb{I}}x_t$ is the *trace* of $x$.

The next lemma will help us to infer the desired recovery of $S$ from the existence of a 'well-behaved' subset of its spatial support $D_S$.

**Lemma 3.** *Let $\varrho\in C^1(G)$ for some $G\subseteq\mathbb{R}^d$ open. Then the following holds.*

(i) *Let $G$ also be connected. Then $\varrho\in\mathrm{DP}_d(G)$ if and only if the Jacobian $J_\varrho$ of $\varrho$ is monomial on $G$, i.e. such that $\{J_\varrho(u)\mid u\in G\}\subseteq\mathrm{M}_d$.*

(ii) *Let $D$ be a subset of $G$ with the property that $D$ is the closure of an open subset $\mathcal{O}$ of $D$. If $J_\varrho$ is monomial on $\mathcal{O}$, then $\varrho$ is monomial on $D$.*

*Proof.* See Appendix A.5. ☐

Definition 4 describes analytical forms that 'sufficiently many' of the distributions constituting the temporal structure (7) of $S$ need to *avoid* in order for $S$ to be identifiable from $X$ up to a monomial transformation. The following regularity assumption accounts for this.

*Definition* 6 ($\alpha$-Contrastive). A continuous stochastic process $S\equiv(S_t^1,\cdots,S_t^d)^{\intercal}_{t\in\mathbb{I}}$ in $\mathbb{R}^d$ with spatial support $D_S$ will be called *$\alpha$-contrastive* if $S$ is IC and there exists a collection $\mathcal{P}\subseteq\Delta_2(\mathbb{I})$ together with an associated collection of open subsets $(D_\mathfrak{p})_{\mathfrak{p}\in\mathcal{P}}\subseteq\mathbb{R}^d$ such that

(i) the union $\bigcup_{(s,t)\in\mathcal{P}}D_{(s,t)}$ is dense in $D_S$, and

(ii) for each $(i,(s,t))\in[d]\times\mathcal{P}$ it holds that $S^i$ is $C^2$-regular at $(s,t)$ with density $\zeta^i_{s,t}$ s.t.

$$\zeta^i_{s,t}\big|_{D^{\times 2}_{(s,t)}} \text{ is regularly non-separable for all } i\in[d], \quad \text{and}$$
$$\zeta^i_{s,t}\big|_{D^{\times 2}_{(s,t)}} \text{ is almost everywhere non-Gaussian for all but at most one } i\in[d],$$

where the above restrictions of the densities are understood w.r.t. the abuse of notation $\zeta^i_{s,t}(x) := \zeta^i_{s,t}(x_i, x_{i+d})$ for $x = (x_\nu) \in \mathbb{R}^{2d}$. (For notational convenience, this abuse of notation is kept throughout the following.)

Notice that the conditions in Def. 6 (ii) reflect the classical pathologies (ii) & (i) from p. 9. We will see that the assumption of Definition 6 are satisfied for a number of popular copula-based time series models (Section 6.1).

**Theorem 2.** *Let the process $S$ in (13) be $\alpha$-contrastive. Then, for any transformation $h$ which is $C^3$-invertible on some open superset of $D_X$, we have with probability one that:*

$$h \cdot X \;\in\; \mathrm{DP}_d \cdot S \quad \textit{if and only if} \quad h \cdot X \quad \textit{has independent coordinates.} \tag{21}$$

*Proof.* The 'only-if'-direction in (21) is clear, so we only need to show the converse implication. To this end, we in total prove the slightly stronger assertion that

$$\text{If} \quad h \cdot X \text{ is IC} \quad \text{and} \quad D \equiv D_{(s,t)} \text{ as in Def. 6,} \qquad \text{then} \qquad \{J_{h\circ f}(u) \mid u \in D\} \subseteq \mathrm{M}_d\,. \tag{22}$$

Given (22) (and Definition 6 (i)), the assertion (21) follows by way of Lemma 3 (ii) and the fact that the trace of almost every realisation of $S$ is contained in a connected subset of $D_S$ (Lemma 1 (ii)).

Let now $(s,t) \in \Delta_2(\mathbb{I})$ be as in Definition 6 (ii), i.e. suppose that $(S_s, S_t) = \pi_{(s,t)}(S)$ admits a (joint) $C^2$-density $\zeta = \zeta_1 \cdots \zeta_d$ (where $\zeta_i \equiv \zeta^i_{s,t}$) with a support $\bar{D} := \mathrm{supp}(\zeta) \subseteq \mathbb{R}^{2d}$ whose boundary $\partial\bar{D}$ is a Lebesgue nullset (cf. Remark 3.4).

Moreover, let $X^*_t$ be a copy of $X_t$ which is independent of $(X_s, X_t)$, and denote

$$Y := (X_s, X_t) \qquad \text{and} \qquad Y^* := (X_s, X^*_t). \tag{23}$$

For $C \sim \mathrm{Ber}(1/2)$ Bernoulli and independent of $Y$ and $Y^*$, consider further

$$\bar{Y} := C \cdot Y + (1 - C) \cdot Y^*$$

(so that $\mathbb{P}_{\bar{Y}} = \frac{1}{2}\mathbb{P}_Y + \frac{1}{2}\mathbb{P}_{Y^*}$) together with the associated regression function

$$\rho \,:\, \mathbb{R}^{2d} \to [0,1] \quad \text{given by} \quad \rho(y) := \mathbb{E}[C \,|\, \bar{Y} = y]. \tag{24}$$

The function $\rho$ then satisfies the following central equation.

**Lemma 4.** *For $\mu$ the probability density of $Y$, and $\mu^*$ the probability density of $Y^*$,*

$$\psi \circ \rho = \log\mu - \log\mu^* \quad \textit{a.e. on} \quad \tilde{D} := \mathrm{supp}(\mu) \tag{25}$$

*for the logit-function $\psi(p) := \log(p/(1-p))$.*

The proof of Lemma 4 is given in Appendix A.6. Recalling now that the components of $S$ are mutually independent, we obtain from the transformation formula for densities (10) that for the inverse $g \equiv (g_1, \cdots, g_d) := f^{-1}$ and the density $\zeta^i_1$ of $S^i_s$, resp. the density $\zeta^i_2$ of $S^i_t$,

$$\log\mu - \log\mu^* = \sum_{i=1}^{d} \big[\log\zeta_i \circ (g_i \times g_i) - \log\zeta^i_1 \circ g_i(u) - \log\zeta^i_2 \circ g_i(v)\big].$$

almost everywhere on $\tilde{D}\,(= (f \times f)(\bar{D}))$. Using (25), it follows that

$$\psi \circ \rho = \sum_{i=1}^{d} P_i \circ (g_i \times g_i) \qquad \text{for} \qquad P_i := \log\zeta_i - \sum_{\nu=1,2} \log\zeta^i_\nu \circ \pi_\nu. \tag{26}$$

Let now $h \equiv (h_1, \cdots, h_d) \in \mathrm{Diff}^3(\mathcal{O}_X)$, for some $\mathcal{O}_X \supseteq D_X$ open, be such that the process $S' := h \cdot X$ has independent components. Using that the above function $\psi \circ \rho$ depends on the observable $X$ only, we due to $(S'_s, S'_t) = (h \times h)(X_s, X_t)$ and (10) obtain that

$$\psi \circ \rho = \sum_{i=1}^{d} Q_i \circ (h_i \times h_i) \quad \text{a.e. on} \quad \tilde{D} \tag{27}$$

analogous to (26), where the functions[13] $Q_i \in C^2(\bar{D}')$, $i \in [d]$, are given as

$$Q_i := \log \tilde{\zeta}_i - \sum_{\nu=1,2} \log \tilde{\zeta}^i_\nu \circ \pi_\nu \quad \text{with} \quad \tilde{\zeta}_i := \frac{\mathrm{d}\mathbb{P}_{(S'^i_s, S'^i_t)}}{\mathrm{d}(u,v)} \quad \text{and} \quad \tilde{\zeta}^i_\nu := \frac{\mathrm{d}\mathbb{P}_{S'^i_{r_\nu}}}{\mathrm{d}u} \tag{28}$$

for $r_1 := s$ and $r_2 := t$, and where $\bar{D}' \subseteq \mathbb{R}^{2d}$ denotes the support of $\tilde{\zeta} \equiv \tilde{\zeta}_1 \cdots \tilde{\zeta}_d$.

Note that, by (124), the marginal densities $\tilde{\zeta}_i$, $\tilde{\zeta}^i_\nu$, and hence the $Q_i$, are indeed twice continuously differentiable since, by (10),

$$\tilde{\zeta} = \frac{\mathrm{d}\mathbb{P}_{(S'_s, S'_t)}}{\mathrm{d}(u,v)} = |\det(J_\phi)| \cdot \left[\zeta \circ \phi\right] \;\in\; C^2(\bar{D}')$$

for the diffeomorphism $\phi := ((h \circ f) \times (h \circ f))^{-1} \in \mathrm{Diff}^3(\mathcal{O}_{S'}^{\times 2}; \mathcal{O}_S^{\times 2})$, with $\mathcal{O}_{S'} := h(\mathcal{O}_S)$.

Combining the identities (26) and (27) yields that

$$\sum_{i=1}^{d} Q_i \circ (h_i \times h_i) = \sum_{i=1}^{d} P_i \circ (g_i \times g_i) \tag{29}$$

everywhere on the dense open subset $D_\mu := \{\mu > 0\}$ of $\tilde{D}$.

Therefore, the desired implication (22) – and hence the assertion of the theorem (see the initial remarks of this proof) – holds if we can show (29) to imply that for $\varrho := h \circ f$ we have

$$\{J_\varrho(u) \mid u \in D\} \;\subseteq\; \mathrm{M}_d \quad \text{for each open} \quad D \subseteq D_S \quad \text{as in Def. 6 (ii)}, \tag{30}$$

i.e. for any (non-empty) open subset $D$ of $\mathbb{R}^d$ for which $\zeta^i\big|_{D^{\times 2}}$ is regularly non-separable for all $i \in [d]$, and a.e. non-Gaussian for all but at most one $i \in [d]$. Let any such $D$ be fixed.

The remainder of this proof is aimed at deriving (30) from (29). To this end, notice that since (29) can be equivalently written as

$$Q \circ (h \times h) = P \circ (g \times g)$$

for $Q := \varsigma \circ (Q_1 \times \cdots \times Q_d) \circ \tau$ and $P := \varsigma \circ (P_1 \times \cdots \times P_d) \circ \tau$ with $\varsigma(y_1, \ldots, y_d) := \sum_{i=1}^d y_i$ and $\tau(x_1, \ldots, x_{2d}) := (x_1, x_{d+1}, x_2, x_{d+2}, \ldots, x_d, x_{2d})$, we obtain that (29) is equivalent to $Q \circ (\varrho \times \varrho) = P$, i.e. to the ($D_\zeta := \{\zeta > 0\}$–everywhere) identity[14]

$$\sum_{i=1}^{d} Q_i \circ (\varrho_i \times \varrho_i) = \sum_{i=1}^{d} P_i. \tag{31}$$

The above is an identity between two twice-continuously-differentiable functions in the arguments $(u_1, \ldots, u_d, v_1, \ldots, v_d) \in D_\zeta \subseteq \mathbb{R}^{2d}$, so we can apply the cross-derivatives $\partial_{u_j}\partial_{v_k}$ to both sides of (31) to arrive at the identities

$$\sum_{i=1}^{d} \left[q_i \circ (\varrho_i \times \varrho_i)\right] \cdot \partial_{u_j}\varrho_i \cdot \partial_{v_k}\varrho_i = \sum_{i=1}^{d} \xi_i \cdot \delta_{ijk} \qquad (j,k \in [d]) \tag{32}$$

where the $\varrho_i$ are the components of (30) and the functions $q_i$ and $\xi_i$ are given as

$$q_i \;:=\; \partial_{u_i}\partial_{v_i} Q_i \qquad \text{and} \qquad \xi_i \;:=\; \partial_{u_i}\partial_{v_i} P_i = \partial_{u_i}\partial_{v_i} \log \zeta_i, \tag{33}$$

[13] Note that here, we employ the abuse of notation $Q_i(x) \equiv Q_i(x_i, x_{i+d})$ for $x = (x_\nu) \in \bar{D}'$.

[14] Once more, we abuse notation by writing $P_i(x) \equiv P_i(x_i, x_{i+d})$ $(x \in D)$ for the RHS of (31).

respectively. (Note that $\partial_{u_j}\partial_{v_k} R_i = r_i \cdot \delta_{ijk}$ $((R,r) \in \{(Q,q),(P,\xi)\})$ by the Cartesian product-form of the functions (26) and (28).) Observe now that the system of equations (32) can be equivalently expressed as the congruence relation

$$J_\varrho^\intercal \cdot \Lambda_q \cdot J_\varrho = \Lambda_\xi \qquad \big(:\Leftrightarrow\ J_\varrho^\intercal(u)\cdot\Lambda_q(u,v)\cdot J_\varrho(v) = \Lambda_\xi(u,v)\big)$$

for $J_\varrho$ the Jacobian of $\varrho$ and for $\Lambda_q, \Lambda_\xi$ defined as the matrix-valued functions

$$\Lambda_q := \operatorname{diag}_{i=1,\dots,d}[q_i \circ (\varrho_i \times \varrho_i)] \quad \text{and} \quad \Lambda_\xi := \operatorname{diag}_{i=1,\dots,d}[\xi_i].$$

Since $\varrho$ is a diffeomorphism over $\bar{D}$, its Jacobian $J_\varrho$ is invertible and hence

$$\Lambda_q = B_\varrho^\intercal \cdot \Lambda_\xi \cdot B_\varrho \quad \text{on} \quad D_\zeta, \qquad \text{for} \quad B_\varrho := J_\varrho^{-1}. \tag{34}$$

Since $B_\varrho = J_{\varrho^{-1}} \circ \varrho$ by the inverse function theorem, the matrix-valued function $B_\varrho$ is clearly continuous. Hence[15] we can apply Lemma 5 below to from (34) and the assumptions of Definition 6 (ii) obtain as desired that

$$\{J_\varrho(u) \mid u \in D\} \ \subseteq\ \mathrm{M}_d\,. \tag{35}$$

Indeed, since the above open set $D \subseteq D_S$ has been chosen such that the (positive) functions $\zeta^i\big|_{D^{\times 2}}$ are regularly non-separable for each $i \in [d]$ and a.e. non-Gaussian for all but at most one $i \in [d]$ (Definition 6 (ii)), Lemma 5 is clearly applicable to the system (34), providing (35) as required. But since the above set $D$ was chosen without further restrictions, (35) amounts to (30) and hence proves Theorem 2 as desired. □

**Lemma 5.** *Let $U \subseteq \mathbb{R}^d$ be open and $\varphi_i \in C^2(U^{\times 2};\mathbb{R}_{>0})$, $i \in [d]$, with $\varphi_i(x) \equiv \varphi_i(x_i, x_{i+d})$, be a family of regularly non-separable, positive functions all of which but at most one are a.e. non-Gaussian. Set $\xi_i := \partial_{x_i}\partial_{x_{i+d}} \log\varphi_i$ for each $i \in [d]$. Then for any continuous $B : U \to \mathrm{GL}_d(\mathbb{R})$ for which the composition $\Lambda : U^{\times 2} \to \mathbb{R}^{d\times d}$ given by*

$$\Lambda(u,v) := B(u)^\intercal \cdot \operatorname{diag}_{i\in[d]}\big[\xi_i(u_i,v_i)\big]\cdot B(v) \tag{36}$$

*(in the coordinates $(u,v) \equiv (u_1,\dots,u_d,v_1,\dots,v_d) \in U^{\times 2}$) has identically-vanishing off-diagonal elements, it holds that the function $B$ is monomial on $U$, i.e. that*

$$B(u) \ \in\ \mathrm{M}_d \qquad \text{for each } \ u \in U. \tag{37}$$

*Proof of Lemma 5.* Set $\check{U} := U \times U$, and for a given $(u,v) \in \check{U}$, denote $\hat{\Lambda}_{u,v} := \operatorname{diag}_{i\in[d]}\big[\xi_i(u_i,v_i)\big]$ and $\Lambda_{u,v} := \Lambda(u,v)$ and $B_u := B(u)$, and assume (wlog, upon re-enumeration) that $\varphi_i$ is a.e. non-Gaussian for each $i \in [d-1]$.

Note that by the fact that $B$ is $\mathrm{GL}_d$-valued and continuous, the identity (37) holds if

$$\exists\, \tilde{U} \subseteq U \text{ dense} \quad \text{s.t.} \quad B_u \in \mathrm{M}_d \quad \text{for all } u \in \tilde{U} \tag{38}$$

(cf. the argument around (134) for details). Our proof consists of constructing a set $\tilde{U}$ for which (38) holds. Let to this end $i \in [d]$ be fixed, and recall that $\varphi_i$ being regularly non-separable implies that there is a closed nullset[16] $\mathcal{N} \subset \check{U}$ s.t. for the open and dense[17] subset

[15] Notice that $D^{\times 2} \subset \bar{D} \equiv \operatorname{supp}(\zeta)$ (and hence $D^{\times 2} \subseteq D_\zeta$, as $D$ is open) since $\zeta\big|_{D^{\times 2}} > 0$ a.e. by the fact that $\zeta\big|_{D^{\times 2}}$ is a.e. non-separable (and hence a.e. non-zero in particular).

[16] Notice that if $\varphi_i$ is regularly non-separable and a.e. non-Gaussian, there (by Definition 4) will be a closed nullset $\tilde{\mathcal{N}}_i \subset \pi_{(i,i+d)}(\check{U}) \subseteq \mathbb{R}^2$ (s.t. $\tilde{\mathcal{N}}_i \cap \{(x,x) \mid x \in \mathbb{R}\}$ has Hausdorff-measure zero on the diagonal $\Delta_\mathbb{R} := \{(x,x) \mid x \in \mathbb{R}\}$) on whose complement $\varphi_i$ is strictly non-Gaussian and non-separable and s.t. $\varphi_i\big|_{\Delta_\mathbb{R}\setminus\tilde{\mathcal{N}}_i}$ vanishes nowhere. Hence for the (relatively) closed nullsets $\mathcal{N}_i := \pi^{-1}_{(i,i+d)}(\tilde{\mathcal{N}}_i) \cap \check{U} \subset \mathbb{R}^{2d}$, the (relatively) closed union $\mathcal{N} := \bigcup_{i\in[d]} \mathcal{N}_i$ works as desired.

[17] Recall that for (any) $\check{U} \subseteq \mathbb{R}^m$ open and $\mathcal{N} \subseteq \mathbb{R}^m$ a Lebesgue nullset, the complement $\check{U} \setminus \mathcal{N}$ is dense in $\check{U}$. Indeed: If for $\check{U}_\circ := \check{U} \setminus \mathcal{N}$ we had $\operatorname{clos}(\check{U}_\circ) \subsetneq \check{U}$, then there would be $u \in \check{U}$ with $B_\delta(u) \subseteq \check{U} \setminus \operatorname{clos}(\check{U}_\circ) \subseteq \mathcal{N}$ for some $\delta > 0$, contradicting that the Lebesgue measure of $\mathcal{N}$ is zero.

$\check{U}_\circ := \check{U} \setminus \mathcal{N}$ of $\check{U}$, each restriction $\varphi_i\big|_{\check{U}_\circ}$ is s.t.

$$\varphi_i\big|_{\check{U}_\circ} \text{ is strictly non-Gaussian for } i \neq d, \quad \text{and for each } i \in [d]: \tag{39}$$

$$\varphi_i\big|_{\check{U}_\circ} \text{ is strictly non-separable} \quad \text{with} \quad \xi_i\big|_{(\Delta_U \cap \check{U}_\circ)} \neq 0 \text{ everywhere.} \tag{40}$$

Given (40), Lemma 2 (ii) (together with the elementary topological facts that $(a)$ : a subset which lies densely inside a dense subspace is itself dense again, and $(b)$ : the intersection of two open dense subsets is again an open dense subset) implies that the intersection

$$\check{U}_* := \bigcap_{i\in[d]} \big\{z \in \check{U}_\circ \;\big|\; \xi_i(z) \neq 0\big\} \quad \text{is an open dense subset of } \check{U}.$$

Consequently, the coordinate-projections $U' := \pi_{[d]}(\check{U}_*)$ and $V' := \pi_{(d+1,\dots,2d)}(\check{U}_*)$ are open and dense subsets of $U\ (= \pi_{[d]}(\check{U}))$. We now claim that the sets

$$\tilde{U}_i := \Big\{u \in U' \;\big|\; \exists\,(\emptyset \neq)\,\mathcal{V}_u \subseteq V' \text{ open}^{18} \;:\; q_u^i\big|_{\mathcal{V}_u} \text{ is non-constant}\Big\}, \quad \text{defined by the function} \quad q_u^i(v) := \frac{\xi_i(u,u)\cdot\xi_i(v,v)}{\xi_i(u,v)^2}, \tag{41}$$

are dense in $U'$ — and hence (cf. fact $(a)$) also are dense in $U$ — for each $i \in [d-1]$.

To see that this holds, we proceed via proof by contradiction and assume that $\tilde{U}_i$ is not dense in $U'$. In this case, there exists $(\bar{u}, r) \in U' \times \mathbb{R}_{>0}$ with $B_r(\bar{u}) \subset U' \setminus \tilde{U}_i$ (recall that $U'$ is open). Moreover: Since we have $\Delta_U \cap \check{U}_\circ \subset \check{U}_*$ (by (40)) and $\check{U}_*$ is open, we can even find a convex open neighbourhood $\mathcal{V}_{\bar{u}} \subseteq B_r(\bar{u})$ of $\bar{u}$ such that the whole square $\mathcal{Q}_{\bar{u}} \equiv \mathcal{V}_{\bar{u}} \times \mathcal{V}_{\bar{u}}$ is contained in $\check{U}_*$. Now by construction, we for each slice $\{u\} \times \mathcal{V}_{\bar{u}} \subset \mathcal{Q}_{\bar{u}}$, $u \in \mathcal{V}_{\bar{u}}$, must have that $q_u^i\big|_{\mathcal{V}_{\bar{u}}}$ is a constant function, say $q_u^i\big|_{\mathcal{V}_{\bar{u}}} \equiv c_u$ for some $c_u \in \mathbb{R}$, so that in particular (for $\varrho : \mathcal{V}_{\bar{u}} \ni u \mapsto c_u \in \mathbb{R}$)

$$\xi_i(u,u)\cdot\xi_i(v,v) = \varrho(u)\cdot\xi_i(u,v)^2, \quad \forall\,(u,v) \in \mathcal{Q}_{\bar{u}}. \tag{42}$$

But since the square $\mathcal{Q}_{\bar{u}}$ contains its diagonal, i.e.: $\mathcal{Q}_{\bar{u}} \supset \{(u,u) \mid u \in \mathcal{V}_{\bar{u}}\} =: \Delta_{\mathcal{V}_{\bar{u}}}$, we can evaluate the relation (42) for the points in $\Delta_{\mathcal{V}_{\bar{u}}}$, yielding that $\varrho \equiv 1$. Consequently,

$$\xi_i\big|_{\mathcal{Q}_{\bar{u}}}(u,v) = \epsilon\cdot\varsigma(u)\cdot\varsigma(v), \quad \forall\,(u,v) \in \mathcal{Q}_{\bar{u}},$$

for $\varsigma : \mathcal{V}_{\bar{u}} \ni x \mapsto \varsigma(x) := \sqrt{|\xi_i(x,x)|}$ and with $\epsilon$ denoting the sign of $\xi_i\big|_{\mathcal{Q}_{\bar{u}}}$. Notice that since $\xi_i$ is continuous and $\mathcal{Q}_{\bar{u}}$ is connected, $\epsilon$ will be constant (i.e. $\epsilon \equiv \pm 1$). But since $\mathcal{Q}_{\bar{u}} \subset \check{U}_*$ is symmetric, open and convex, Lemma 2 (iii) then implies that

$$\varphi_i\big|_{\mathcal{Q}_{\bar{u}}} \quad \text{is pseudo-Gaussian,} \quad \text{in contradiction to (39).}$$

This proves that each of the above sets $\tilde{U}_i$, $i \in [d-1]$, must be dense in $U$.

But since each of the dense subsets $\tilde{U}_i$ is also open by the fact that the quotients in (41) are continuous in $(u,v)$, we (once more by the above fact (b)) find that their intersection

$$\tilde{U} := \bigcap_{i\in[d-1]} \tilde{U}_i \quad \text{is a dense subset of } U.$$

We claim that the above set $\tilde{U}$ satisfies (38).

[18] Note that for $\tilde{U}_i$ to be well-defined, we in addition require each $\mathcal{V}_u^i$ to be s.t. $\{u\} \times \mathcal{V}_u^i \subset \check{U}_*$.

To see this, note first that (36) yields[19] the system of matrix equations

$$\begin{cases} B_u^\intercal \cdot \hat{\Lambda}_{u,u} \cdot B_u &= \Lambda_{u,u} \\ B_u^\intercal \cdot \hat{\Lambda}_{u,v} \cdot B_v &= \Lambda_{u,v} \\ B_v^\intercal \cdot \hat{\Lambda}_{v,v} \cdot B_v &= \Lambda_{v,v} \end{cases} \qquad \text{for each} \quad (u,v) \in U^{\times 2}; \tag{43}$$

we prove (38) by defining a map $\eta : \tilde{U} \to V'$ with the property that (43) when evaluated at $(u,v) := (u, \eta(u))$ yields $B_u \in \mathrm{M}_d$ by necessity. Let to this end $u \in \tilde{U}$ be arbitrary. Then by the definition of $\tilde{U}$ (recalling (41)), we for each $i \in [d]$ can find some open $\mathcal{V}_u^i \subset V'$ such that the intersection $\mathcal{V}_u := \bigcap_{i\in[d]} \mathcal{V}_u^i$ is non-empty and the continuous map

$$q_u := q_u^1 \times \cdots \times q_u^d \;:\; \mathcal{V}_u \;\to\; \mathbb{R}^d \qquad (q_u^i \text{ as in } (41)) \tag{44}$$

is non-constant in its first $(d-1)$ components. Hence by the intermediate-value theorem, the set $\pi_{[d-1]}(q_u(\mathcal{V}_u)) \subseteq \mathbb{R}^{d-1}$ has non-empty interior, which implies that for $\nabla^\times := \{(v_i) \in \mathbb{R}^d \mid \exists\, i,j \in [d], i \neq j \,:\, |v_i| = |v_j|\}$ (the closed nullset of all vectors in $\mathbb{R}^d$ having two components differing at most up to a sign), the preimage

$$\tilde{\mathcal{V}}_u := q_u^{-1}(\mathbb{R}^d \setminus \nabla^\times) \;\subset\; \mathcal{V}_u \tag{45}$$

will be non-empty. This observation gives rise to maps of the form

$$\eta \,:\, \tilde{U} \to V', \quad u \mapsto \eta(u) \in \tilde{\mathcal{V}}_u,$$

and as we will now see, any such map is of the desired type that we announced above. Indeed: Taking any $(u,v) \in \bigcup_{\tilde{u}\in\tilde{U}} \{\tilde{u}\} \times \tilde{\mathcal{V}}_u \subseteq \check{U}_*$, we obtain from (43) that[20]

$$B_u^{-1} \cdot \bar{\Lambda}_{u,v} \cdot B_u = \tilde{\Lambda} \qquad \text{for} \quad \bar{\Lambda}_{u,v} \equiv \mathrm{diag}_{i\in[d]}[\lambda_{u,v}^i] := \hat{\Lambda}_{u,u} \cdot \hat{\Lambda}_{u,v}^{-2} \cdot \hat{\Lambda}_{v,v} \tag{46}$$

and the diagonal matrix $\tilde{\Lambda} := \Lambda_{u,v}^{-1} \cdot \Lambda_{v,v} \cdot \Lambda_{u,v}^{-1} \cdot \Lambda_{u,u}$. Observing now (recalling (41)) that

$$(\lambda_{u,v}^1, \cdots, \lambda_{u,v}^d) \;=\; q_u(v) \qquad \text{for the function} \;\; q_u \;\; \text{defined in (44)},$$

we by the definition (45) of $\tilde{\mathcal{V}}_u$ immediately obtain that

$$\text{the eigenvalues} \quad \lambda_{u,v}^1, \ldots, \lambda_{u,v}^d \;\; \text{of} \;\; \bar{\Lambda}_{u,v} \quad \text{are pairwise distinct.} \tag{47}$$

Hence by the elementary fact that diagonal matrices with pairwise distinct eigenvalues are stabilised by monomial matrices only, observation (47) by way of the similarity equation (46) finally implies $B_u \in \mathrm{M}_d$ — and hence (38) — as desired. □

The next section extends the above lines of argumentation to additional types of sources.

## 5. An Extension to Source Types of Alternative Temporal Structures

The argumentation behind Theorem 2 can be generalised by 'unfreezing' its usage of the temporal structure (7), i.e. by allowing the considered time-pairs $(s,t)$ to 'vary more freely' across $\Delta_2(\mathbb{I})$. This qualifies additional source classes for nonlinear identification via (21).

**Lemma 6.** *Suppose that in addition to (13) there are $\mathfrak{p}_0, \mathfrak{p}_1, \mathfrak{p}_2 \in \Delta_2(\mathbb{I})$ and $(u,v) \in \mathbb{R}^{2d}$ such that $S$ is $C^2$-regular at $(\mathfrak{p}_0, (u,v))$, $(\mathfrak{p}_1, (u,u))$ and $(\mathfrak{p}_2, (v,v))$ with density $\zeta_{\mathfrak{p}_0}$, $\zeta_{\mathfrak{p}_1}$ and $\zeta_{\mathfrak{p}_2}$, respectively. Then for any map $h$ which is $C^3$-invertible on some open superset of $D_X$ and such that $h \cdot X$ has independent components, we have that*

$$B_\varrho(u)^{-1} \cdot \bar{\Lambda}_{\mathfrak{p}_0,\mathfrak{p}_1,\mathfrak{p}_2}(u,v) \cdot B_\varrho(u) = \tilde{\Lambda}_{\mathfrak{p}_0,\mathfrak{p}_1,\mathfrak{p}_2}(u,v) \tag{48}$$

[19] Cf. the proof of Lemma 6 for details.
[20] Note that the invertibility of $\Lambda_{u,v}$ is obtained from the choice of $(u,v)$.

*for $B_\varrho$ the inverse Jacobian of $\varrho := h\circ f$ and the diagonal matrices*

$$\begin{aligned} \bar{\Lambda}_{\mathfrak{p}_0,\mathfrak{p}_1,\mathfrak{p}_2}(u,v) &:= \Lambda_{\xi_{\mathfrak{p}_1}}(u,u)\cdot\Lambda_{\xi_{\mathfrak{p}_0}}(u,v)^{-2}\cdot\Lambda_{\xi_{\mathfrak{p}_2}}(v,v) \qquad \textit{and} \\ \tilde{\Lambda}_{\mathfrak{p}_0,\mathfrak{p}_1,\mathfrak{p}_2}(u,v) &:= \Lambda_{q_{\mathfrak{p}_1}}(u,u)\cdot\Lambda_{q_{\mathfrak{p}_0}}(u,v)^{-2}\cdot\Lambda_{q_{\mathfrak{p}_2}}(v,v), \end{aligned} \tag{49}$$

*where $\Lambda_{\xi_{\mathfrak{p}_\nu}}(x) := \operatorname{diag}[\xi^1_{\mathfrak{p}_\nu}(x),\cdots,\xi^d_{\mathfrak{p}_\nu}(x)]$ for $\xi^i_{\mathfrak{p}_\nu}(x) := \partial_{x_i}\partial_{x_{i+d}}\log\zeta_{\mathfrak{p}_\nu}(x)$ and $\Lambda_{q_{\mathfrak{p}_\nu}}(x) :=$ $\operatorname{diag}[q^1_{\mathfrak{p}_\nu}(x),\cdots,q^d_{\mathfrak{p}_\nu}(x)]$ given by the LHS of (33) (in dependence of $\mathfrak{p}_\nu$).*

*Proof.* Copying the argumentation that led to (34), we obtain the congruence relations

$$\Lambda_{q_{\mathfrak{p}_\nu}}(\tilde{u},\tilde{v}) = B_\varrho^\intercal(\tilde{u})\cdot\Lambda_{\xi_{\mathfrak{p}_\nu}}(\tilde{u},\tilde{v})\cdot B_\varrho(\tilde{v}) \quad \text{for each } (\tilde{u},\tilde{v})\in\{\zeta_{\mathfrak{p}_\nu}>0\}$$

and $\nu = 0,1,2$. Evaluating these at the points $(u,v),(u,u)$ and $(v,v)$, we arrive at the system

$$B_\varrho(u)^\intercal\cdot\Lambda_{\xi_{\mathfrak{p}_0}}(u,v)\cdot B_\varrho(v) = \Lambda_{q_{\mathfrak{p}_0}}(u,v) \tag{50}$$

$$B_\varrho(u)^\intercal\cdot\Lambda_{\xi_{\mathfrak{p}_1}}(u,u)\cdot B_\varrho(u) = \Lambda_{q_{\mathfrak{p}_1}}(u,u) \tag{51}$$

$$B_\varrho(v)^\intercal\cdot\Lambda_{\xi_{\mathfrak{p}_2}}(v,v)\cdot B_\varrho(v) = \Lambda_{q_{\mathfrak{p}_2}}(v,v). \tag{52}$$

From (50) we then find that

$$B_\varrho(v) = \Lambda^{-1}_{\xi_{\mathfrak{p}_0}}(u,v)\cdot\left(B_\varrho(u)^\intercal\right)^{-1}\cdot\Lambda_{q_{\mathfrak{p}_0}}(u,v),$$

which, when plugged into (52), yields

$$B_\varrho(u)^{-1}\cdot\Lambda_{\xi_{\mathfrak{p}_2}}(v,v)\cdot\Lambda^{-2}_{\xi_{\mathfrak{p}_0}}(u,v)\cdot\left(B_\varrho(u)^\intercal\right)^{-1} = \Lambda_{q_{\mathfrak{p}_2}}(v,v)\cdot\Lambda_{q_{\mathfrak{p}_0}}(u,v)^{-2} \tag{53}$$

and hence, upon left-multiplying both sides of (51) by the matrix product (53),

$$B_\varrho(u)^{-1}\cdot\Lambda_{\xi_{\mathfrak{p}_2}}(v,v)\Lambda^{-2}_{\xi_{\mathfrak{p}_0}}(u,v)\Lambda_{\xi_{\mathfrak{p}_1}}(u,u)\cdot B_\varrho(u) = \Lambda_{q_{\mathfrak{p}_2}}(u,u)\Lambda_{q_{\mathfrak{p}_2}}(v,v)\Lambda_{q_{\mathfrak{p}_0}}(u,v)^{-2}.$$

This last equation is identical to (48), as desired. □

Define $\psi(x,y,z) := x^{-2}yz$, and denote by $\nabla^\times := \left\{(\lambda_\nu)\in\mathbb{R}^d \;\middle|\; \exists\, i,j\in[d],\, i\neq j \;:\; \lambda_i=\lambda_j\right\}$ the set of all vectors in $\mathbb{R}^d$ whose coordinates are not pairwise distinct.

*Definition* 7 ($\{\beta,\gamma\}$-Contrastive). A continuous stochastic process $S = (S^1_t,\ldots,S^d_t)_{t\in\mathbb{I}}$ in $\mathbb{R}^d$ with independent components and spatial support $D_S$ will be called

- *$\beta$-contrastive* if $D_S$ is the closure of its interior and for any open subset $U$ of $D_S$ there is

  $$\begin{gathered}\text{an open subset } \tilde{U} \text{ of } U \quad\text{and}\quad \mathfrak{p}\equiv(s,t),\, \mathfrak{p}'\in\Delta_2(\mathbb{I}) \quad\text{such that,}\quad \text{for all } i\in[d],\\ \text{the density } \zeta^i_{s,t} \text{ of } (S^i_s,S^i_t), \text{ likewise } \zeta^i_{\mathfrak{p}'}, \text{ exists with } \zeta^i_{\mathfrak{p}},\zeta^i_{\mathfrak{p}'}\in C^2(\tilde{U}^{\times 2}) \text{ and}\end{gathered}$$

  $$\xi^{i\,|\,\tilde{U}}_{s,t} := \left[\partial_x\partial_y\log\zeta^i_{s,t}\right]\circ\iota_{\tilde{U}} \neq 0 \quad\text{and}\quad \xi^{i\,|\,\tilde{U}}_{\mathfrak{p}'}\neq 0 \quad\text{(a.e.),}\qquad\text{and} \tag{54}$$

  $$\xi^{i\,|\,\tilde{U}}_{\mathfrak{p}'} \notin \langle\xi^{i\,|\,\tilde{U}}\rangle_{\mathbb{R}} := \left\{c\cdot\xi^{i\,|\,\tilde{U}}_{\mathfrak{p}} \;\middle|\; c\in\mathbb{R}\right\} \tag{55}$$

  with $\iota_{\tilde{U}} : \tilde{U}\ni u\mapsto(u,u)\in\Delta_{\tilde{U}}$ and both $U,\tilde{U}$ non-empty;[21]

- *$\gamma$-contrastive* if there is a dense open subset $\mathcal{U}$ of $D_S$ such that

  $$\begin{gathered}\text{for each } u\in\mathcal{U} \quad \text{there exists } (v,\mathfrak{p}_0,\mathfrak{p}_1,\mathfrak{p}_2)\in\mathbb{R}^d\times\Delta_2(\mathbb{I})^{\times 3} \quad\text{such that}\\ S \text{ is } C^2\text{-regular around } (\mathfrak{p}_0,(u,v)),(\mathfrak{p}_1,(u,u)) \text{ and } (\mathfrak{p}_2,(v,v)), \quad\text{and}\\ \left(\psi(\xi^i_{\mathfrak{p}_0}(u,v),\xi^i_{\mathfrak{p}_1}(u,u),\xi^i_{\mathfrak{p}_2}(v,v))\right)_{i\in[d]}\in\left(\mathbb{R}^d\setminus\nabla^\times\right),\end{gathered} \tag{56}$$

  where $\xi^i_{\mathfrak{p}} := \partial_{x_i}\partial_{x_{i+d}}\log\zeta_{\mathfrak{p}}$ is the mixed log-derivatives of the $C^2$-density $\zeta^i_{\mathfrak{p}}$ of $(S^i_s,S^i_t)$.

We will see that the assumptions of $\gamma$-contrastivity are satisfied for a number of popular stochastic processes (Section 6.2).

[21] Here as before, we abused notation by writing $\zeta^i_{s,t}(x) = \zeta^i_{s,t}(x_i,x_{i+d})$ for $x=(x_\nu)\in\mathbb{R}^{2d}$.

*Remark* 5.1 (Relation Between $\alpha$-, $\beta$- and $\gamma$-Contrastive Sources). Notice that every $\alpha$-contrastive process is also $\gamma$-contrastive (for $\mathfrak{p}_0 = \mathfrak{p}_1 = \mathfrak{p}_2$, as the proof of Theorem 2 shows), while $\beta$-contrastivity does not imply—nor is it implied by—either $\alpha$- or $\gamma$-contrastivity.

**Theorem 3.** *Let the process $S$ in (13) be $\beta$- or $\gamma$-contrastive. Then for any transformation $h$ which is $C^3$-invertible on an open superset of $D_X$, we have with probab. one that:*

$$h \cdot X \;\in\; \mathrm{DP}_d \cdot S \quad \textit{if and only if} \quad h \cdot X \textit{ has independent coordinates.} \tag{57}$$

*Proof.* Let $h$ be $C^3$-invertible on some open superset of $D_X$ and such that $h \cdot X$ has independent coordinates; the proof of (57) is an extension of the proof of Theorem 2, so let us adopt the set-up and notation of the latter (as done in Lemma 6). Recall from there (cf. (30)) that (57) follows if we can find a dense open subset $\mathcal{D}$ of $D_S$ such that

$$B_\varrho(u) \,\in\, \mathrm{M}_d \quad \text{ for each } \; u \in \mathcal{D}. \tag{58}$$

Let us first suppose that $S$ is $\beta$-contrastive. In this case, we consider the set

$$\mathcal{D}_0 := \{u \in \mathrm{int}(D_S) \mid \exists\, (\delta, \mathfrak{p}) \in \mathbb{R}_{>0} \times \Delta_2(\mathbb{I}) \;\; \text{satisfying } \; (60) \text{ and } (61)\} \tag{59}$$

with properties (60), (61) that for a given $(\delta, \mathfrak{p}) \in \mathbb{R}_{>0} \times \Delta_2(\mathbb{I})$ are defined as

$$B_\delta(u) \subset D_S \quad \text{with} \quad \Lambda_{\xi_\mathfrak{p}}\big|_{B_\delta(u) \times B_\delta(u)} \subset \mathrm{GL}_d(\mathbb{R}), \qquad \text{and} \tag{60}$$

$$\begin{array}{c}\text{there is } \; \mathfrak{p}' \in \Delta_2(\mathbb{I}) \;\; \text{s.t. for } \; \mathcal{U} := B_\delta(u) \text{ and each } i \in [d], \;\text{ the diagonal restriction} \\ \xi^{i\,|\,\mathcal{U}}_{\mathfrak{p}'} \;\; \text{vanishes nowhere and is such that } \; \xi^{i\,|\,\mathcal{U}}_{\mathfrak{p}'} \notin \langle \xi^{i\,|\,\mathcal{U}}_{\mathfrak{p}} \rangle_{\mathbb{R}}.\end{array} \tag{61}$$

Let us show first that $\mathcal{D}_0$ is dense in the interior $D_S^\circ := \mathrm{int}(D_S)$. Indeed: Since $D_S^\circ$ is open, assuming that $\mathcal{D}_0$ is not dense in $D_S^\circ$ implies that there exists $(u_*, r) \in D_S^\circ \times \mathbb{R}_{>0}$ with $B_r(u_*) \subseteq D_S^\circ \setminus \mathcal{D}_0$. Now since $S$ is $\beta$-contrastive, there will be some $(\tilde{u}_*, r_1) \in B_r(u_*) \times (0, r)$ with $B_{r_1}(\tilde{u}_*) \subseteq B_r(u_*)$ such that both (54) and (55) hold everywhere on $\tilde{U} := B_{r_1}(\tilde{u}_*)$ for some $\mathfrak{p}, \tilde{\mathfrak{p}}' \in \Delta_2(\mathbb{I})$; thus also (61) holds for $\mathcal{U} = \tilde{U}$ and $(\mathfrak{p}, \mathfrak{p}') := (\tilde{\mathfrak{p}}, \tilde{\mathfrak{p}}')$. And since the functions $\xi^i_{\tilde{\mathfrak{p}}}$ are continuous at $(\tilde{u}_*, \tilde{u}_*)$, there (due to (54)) will further be some $r_2 > 0$ such that $\xi^i_{\tilde{\mathfrak{p}}}$ vanishes nowhere on $B_{r_2}(\tilde{u}_*)^{\times 2} \subset D_S^{\times 2}$ for each $i \in [d]$; hence also (60) holds for $(\delta, \mathfrak{p}) := (r_2, \tilde{\mathfrak{p}})$. But this yields that both (60) and (61) hold for $u := \tilde{u}_*$ and $\delta := \min(r_1, r_2)$ and $(\mathfrak{p}, \mathfrak{p}') := (\tilde{\mathfrak{p}}, \tilde{\mathfrak{p}}')$, which implies that $\tilde{u}_* \in D_S^\circ \setminus \mathcal{D}_0$ is an element of $\mathcal{D}_0$.

As this is obviously a contradiction, the set (59) must be dense in $D_S^\circ$.

Now since the interior $D_S^\circ$ is dense in $D_S$ by assumption, the theorem's assertion follows if we can show that (58) holds for $\mathcal{D} := D_S^\circ$. But since in turn $\mathcal{D}_0$ is dense in $D_S^\circ$, we obtain that (58) holds for $\mathcal{D} := D_S^\circ$ if we can show that (58) holds for $\mathcal{D} := \mathcal{D}_0$.[22]

Let to this end $u \in \mathcal{D}_0$ be fixed with $\mathfrak{p}, \mathfrak{p}' \in \Delta_2(\mathbb{I})$ and $\mathcal{U} \equiv B_\delta(u) \subset D_S$ as in (60) and (61). Then by Lemma 6 we have that

$$B_\varrho(\tilde{u})^{-1} \cdot \bar{\Lambda}_\nu(\tilde{u}, \tilde{v}) \cdot B_\varrho(\tilde{u}) = \tilde{\Lambda}_\nu(\tilde{u}, \tilde{v}) \qquad \text{for each } \; (\tilde{u}, \tilde{v}) \in \mathcal{U} \times \mathcal{U} \tag{62}$$

with $\nu = 1, 2$ and diagonal matrices $\bar{\Lambda}_1, \bar{\Lambda}_2, \tilde{\Lambda}_1, \tilde{\Lambda}_2 \in \mathrm{GL}_d(\mathbb{R})$ given by

$$\bar{\Lambda}_1 := \bar{\Lambda}_{\mathfrak{p},\mathfrak{p},\mathfrak{p}}, \quad \tilde{\Lambda}_1 := \tilde{\Lambda}_{\mathfrak{p},\mathfrak{p},\mathfrak{p}} \qquad \text{and} \qquad \bar{\Lambda}_2 := \bar{\Lambda}_{\mathfrak{p},\mathfrak{p},\mathfrak{p}'}, \quad \tilde{\Lambda}_2 := \tilde{\Lambda}_{\mathfrak{p},\mathfrak{p},\mathfrak{p}'}$$

with matrices $\bar{\Lambda}_{\mathfrak{p}_0,\mathfrak{p}_1,\mathfrak{p}_2}$ and $\bar{\Lambda}_{\mathfrak{p}_0,\mathfrak{p}_1,\mathfrak{p}_2}$ as defined in (49). Combining the cases $\nu = 1$ and $\nu = 2$ of (62), we for any $C \in \mathbb{R}$ obtain that

$$B_\varrho(\tilde{u})^{-1} \cdot \left[\bar{\Lambda}_1(\tilde{u}, \tilde{v}) + C \cdot \bar{\Lambda}_2(\tilde{u}, \tilde{v})\right] \cdot B_\varrho(\tilde{u}) = \tilde{\Lambda}_1(\tilde{u}, \tilde{v}) + C \cdot \tilde{\Lambda}_2(\tilde{u}, \tilde{v}) \tag{63}$$

[22] Cf. (38) and the argument around (134) for details.

for each $(\tilde{u},\tilde{v}) \in \mathcal{U}\times\mathcal{U}$. Hence (and since $u \in \mathcal{D}_0$ was chosen arbitrarily), the identity (63) implies (58) if there is a pair $(C,\tilde{v}) \in \mathbb{R}\times\mathcal{U}$ for which the diagonal entries of $[\bar{\Lambda}_1(u,\tilde{v}) + C\cdot\bar{\Lambda}_2(u,\tilde{v})] =: \operatorname{diag}[\lambda^1_{u,\tilde{v}},\cdots,\lambda^d_{u,\tilde{v}}]$ are pairwise distinct. We now prove this, i.e. we show that

$$\text{there is } \; C\in\mathbb{R} \; \text{ for which we can find some } \; \tilde{v}\in\mathcal{U} \; \text{ s.t. the diagonal entries}$$
$$\lambda^i_{u,\tilde{v}} = \frac{\xi^i_{\mathfrak{p}}(u,u)}{\xi^i_{\mathfrak{p}}(u,\tilde{v})^2}\cdot\left(\xi^i_{\mathfrak{p}}(\tilde{v},\tilde{v}) + C\cdot\xi^i_{\mathfrak{p}'}(\tilde{v},\tilde{v})\right),\; i\in[d], \quad \text{are pw. distinct.} \tag{64}$$

Notice that, as detailed in the proof of Lemma 5, the fact that by construction each of the functions $q_i : \mathcal{U}\times\mathcal{U} \ni (\tilde{u},\tilde{v}) \mapsto \lambda^i_{\tilde{u},\tilde{v}}$ $(i\in[d])$ are continuous implies that (64) holds if

$$\exists\, C\in\mathbb{R} \; \text{ such that } \; \vartheta_i\,:\mathcal{U}\ni\tilde{v}\,\mapsto\, q_i(u,\tilde{v}) \; \text{ is non-constant for each } \; i\in[d]. \tag{65}$$

To prove (65), notice that since for each $i\in[d]$ we have the decomposition

$$\vartheta_i = \theta_i + C\cdot\theta_i' \qquad \text{with} \qquad \theta_i(\tilde{v}) := \frac{\xi^i_{\mathfrak{p}}(u,u)\xi^i_{\mathfrak{p}}(\tilde{v},\tilde{v})}{\xi^i_{\mathfrak{p}}(u,\tilde{v})^2}$$

and $\theta_i' : \mathcal{U}\times\mathcal{U}\to\mathbb{R}$ defined likewise but with the right factor in the above enumerator replaced by $\xi^i_{\mathfrak{p}'}(\tilde{v},\tilde{v})$, we find that if, for each $i\in[d]$, the functions

$$\theta_i \;\text{ or }\; \theta_i' \;\text{ are non-constant,} \qquad \text{then} \qquad (65)\;\text{ holds.} \tag{66}$$

Indeed: If *either* $\theta_i$ *or* $\theta_i'$ is non-constant in $\tilde{v}$, then clearly their linear combination $\vartheta_i$ will be non-constant in $\tilde{v}$ for each $C\neq 0$. If $\theta_i$ or $\theta_i'$ are *both* non-constant in $\tilde{v}$, then there might be some $C_i\in\mathbb{R}$ such that $\theta_i + C_i\cdot\theta_i'$ is constant in $\tilde{v}$ (define $C_i := 1$ otherwise); in this case, setting $C := \max_{i\in[d]} C_i + 1$ implies that $\vartheta_i = \theta_i + C\cdot\theta_i' = (\theta_i + C_i\cdot\theta_i') + (C - C_i)\cdot\theta_i'$ is non-constant in $\tilde{v}$ for each $i\in[d]$, as desired.

To see that the premise of (66) holds, assume otherwise that there is $i\in[d]$ for which the function $\theta_i : \mathcal{U}\to\mathbb{R}$ is constant in $\tilde{v}$, say

$$\theta_i(\tilde{v}) =: \varsigma_i \quad \text{ for all } \; \tilde{v}\in\mathcal{U}.$$

Then, as $\theta_i$ vanishes nowhere in consequence of (60), we find that

$$\left[\xi^i_{\mathfrak{p}}(u,\cdot)\right]^2 = \frac{\xi^i_{\mathfrak{p}}(u,u)\cdot\xi^{i\,|\,\mathcal{U}}_{\mathfrak{p}}}{\theta_i} = c_i\cdot\eta \qquad \text{on} \quad \mathcal{U} \tag{67}$$

for the constant $c_i := \xi^i_{\mathfrak{p}}(u,u)\cdot\varsigma_i^{-1}$ and the function $\eta : \mathcal{U}\to\mathbb{R}$ given by $\eta(\tilde{v}) := \xi^i_{\mathfrak{p}}(\tilde{v},\tilde{v})$. Now if the function $\theta_i'$ were constant as well, say $\theta_i' \equiv \varsigma_i'\,(\neq 0)$, then we would likewise obtain that $\left[\xi^i_{\mathfrak{p}}(u,\cdot)\right]^2 = c_i'\cdot\eta'$ on $\mathcal{U}$, for the non-zero constant $c_i' := \xi^i_{\mathfrak{p}}(u,u)\cdot(\varsigma_i')^{-1}$ and the function $\eta' : \mathcal{U}\to\mathbb{R}$ given by $\eta'(\tilde{v}) := \xi^i_{\mathfrak{p}'}(\tilde{v},\tilde{v})$. Combined with (67), we find that

$$c_i'\cdot\eta' = c_i\cdot\eta \quad \text{ and hence } \quad \xi^{i\,|\,\mathcal{U}}_{\mathfrak{p}'} = \text{const.}\cdot\xi^{i\,|\,\mathcal{U}}_{\mathfrak{p}},$$

the latter contradicting (61). This proves the premise of (66) and hence (58) for $\mathcal{D} = \operatorname{int}(D_S)$.

Suppose next that $S$ is $\gamma$-contrastive. In this case, we claim that (58) holds for the dense subset $\mathcal{D}\equiv\mathcal{U}$ of $D_S$ postulated by Def. 7. To see that this is true, fix any $u\in\mathcal{U}$ and recall that, by Lemma 6 and the previous discussions, the Jacobian $B_\varrho(u)$ of $\varrho\equiv h\circ f$ at $u$ is monomial if there is $v\in\mathbb{R}^d$ and $\mathfrak{p}_0,\mathfrak{p}_1,\mathfrak{p}_2\in\Delta_2(\mathbb{I})$ with $(u,v)\in\{\xi^1_{\mathfrak{p}_0}\neq 0,\ldots,\xi^d_{\mathfrak{p}_0}\neq 0\}$ for which the diagonal matrix $\bar{\Lambda}_{u,v}\equiv\bar{\Lambda}_{\mathfrak{p}_0,\mathfrak{p}_1,\mathfrak{p}_2}(u,v)$ given by (49) has pairwise distinct eigenvalues.[23] Since the diagonal of $\bar{\Lambda}_{u,v}$ equals the vector $(\psi(\xi^i_{\mathfrak{p}_0}(u,v),\xi^i_{\mathfrak{p}_1}(u,u),\xi^i_{\mathfrak{p}_2}(v,v))_{i\in[d]}$, choosing $(v,\mathfrak{p}_0,\mathfrak{p}_1,\mathfrak{p}_2)$ as in (7), (56) thus yields $B_\varrho(u)\in\mathrm{M}_d$ as claimed. As $u\in\mathcal{U}$ was arbitrary, we obtain $B_\varrho\big|_{\mathcal{U}}\subset\mathrm{M}_d$ as desired in (58). This finishes the proof. □

[23] Note that $(u,v)\in\{\xi^1_{\mathfrak{p}_0}\neq 0,\ldots,\xi^d_{\mathfrak{p}_0}\neq 0\}$ if $\{\psi(\xi^i_{\mathfrak{p}_0}(u,v),\alpha,\beta)\mid i\in[d]\}\subset\mathbb{R}$ for some $\alpha,\beta\in\mathbb{R}$.

## 6. Examples of Applicable Sources

The statistical assumptions of $\alpha$-, $\beta$- or $\gamma$-contrastivity are satisfied by a number of popular models for stochastic signals, among them many copula-based time series models (Section 6.1) as well as a variety of Gaussian processes and Geometric Brownian Motion (Section 6.2).

6.1. **Popular Copula-Based Source Models Are $\alpha$-Contrastive.** It is well-known (e.g. [61, Sect. 2.10], [22]) that the temporal structure (7) of a scalar stochastic process $S = (S_t)_{t\in\mathbb{I}}$ can be given an analytical representation of the form

$$\zeta_{s,t}(x,y) = \zeta_s(x)\zeta_t(y)\cdot c_{s,t}(F_s^S(x), F_t^S(y)) \qquad \big((s,t)\in\Delta_2(\mathbb{I})\big), \tag{68}$$

where $\zeta_{s,t}$ is the probability density of $(S_s, S_t)$, $F_r^S$ is the cdf of the vector $S_r$ with $\zeta_r$ its density, and $c_{s,t} : [0,1]^{\times 2}\to\mathbb{R}$ is the uniquely determined *copula density* of $(S_s, S_t)$.

**Proposition 1.** *Let $S \equiv (S_t)_{t\in\mathbb{I}} \equiv (S^1,\cdots,S^d)$ be an IC stochastic process in $\mathbb{R}^d$ such that $S_t$ admits a $C^2$-density $\zeta_t$ for each $t\in\mathbb{I}$ with the property that $t\mapsto\zeta_t(x)$ is continuous for each $x\in\mathbb{R}^d$. Suppose further that for some $\mathcal{P}\subseteq\Delta_2(\mathbb{I})$ with $\bigcup_{(s,t)\in\mathcal{P}}\{\zeta_s\cdot\zeta_t>0\}$ dense in $D_S$,[24] it holds that the copula densities $\{c^i_{s,t} \mid (s,t)\in\mathcal{P}\}$ of $S^i$ (cf. (68)) are such that*

$$c^i_{s,t} \ \text{ are positive and strictly non-Gaussian} \quad \text{and} \quad \partial_x\partial_y \log c^i_{s,t} \ \text{vanishes nowhere} \tag{69}$$

*for each $i\in[d]$. Then the process $S$ is $\alpha$-contrastive.*

*Proof.* We verify that $S$ satisfies the conditions of Definition 6. Take any $i\in[d]$ and $(s,t)\in\mathcal{P}$. Since by assumption the density $\zeta^i_{s,t}$ of $(S^i_s, S^i_t)$ exists and satisfies (68), we find

$$\xi^i_{s,t} := \partial_x\partial_y\log\zeta^i_{s,t} = \zeta^i_s\cdot\zeta^i_t\cdot\left[(\partial_x\partial_y\log c_i)\circ\phi^i_{s,t}\right] \quad \text{on} \quad \{\zeta^i_{s,t}>0\}\supseteq\tilde{D}^{\times 2}_{s,t} \tag{70}$$

for $\tilde{D}_{s,t} := \dot{D}_s\cap\dot{D}_t$ and the map $\phi^i_{s,t} := \mathfrak{s}^i_s\times\mathfrak{s}^i_t : \tilde{D}^{\times 2}_{s,t}\to[0,1]^{\times 2}$ with $\mathfrak{s}^i_r := F^{S^i}_r$.[25] Notice that $\phi^i_{s,t}$ is a differentiable injection since the function $\mathfrak{s}^i_r = \mathfrak{s}^i_r(x) \overset{\text{def}}{=} \int_{-\infty}^{x_i}\zeta^i_r(u)\,\mathrm{d}u$ $(r\in\mathbb{I})$ has positive derivative on $\dot{D}_r$. Hence and since $(\zeta^i_s\cdot\zeta^i_t)\big|_{\tilde{D}^{\times 2}_{s,t}} > 0$ by construction, the $\alpha$-contrastivity of $S$ follows by way of (68) and (70) and assumption (69). Indeed: Setting $D_{(s,t)} := \tilde{D}_{s,t}$ for $(s,t)\in\mathcal{P}$, we see that Def. 6 (i) holds by the assumption on $\mathcal{P}$, while Def. 6 (ii) is immediate by (68), (70) and (69) and the above-noted fact that $\phi^i_{s,t} : \tilde{D}^{\times 2}_{s,t}\to\phi^i_{s,t}(\tilde{D}^{\times 2}_{s,t})$ is a diffeomorphism. □

A popular approach in finance, insurance economy and other fields is to read (68) as *a semi-parametric stationary model* for $S = (S_t)_{t\in\mathbb{I}}$ by assuming the existence of some $\mathcal{I}\subset\mathbb{I}$ discrete ('set of observations') for $\zeta_r\equiv\zeta$ with cdf $F_\zeta$ (each $r\in\mathcal{I}$) and $D_S = \operatorname{supp}(\zeta)$ and $c_{s,t}\equiv c_\theta$ uniformly parametrized for all $(s,t)\in\mathcal{P} := \mathcal{I}^{\times 2}\cap\Delta_2(\mathbb{I})$, e.g. [14, Sect. 2], [26]:

$$\zeta_{s,t}(x,y) = \zeta(x)\zeta(y)\cdot c_\theta(F_\zeta(x), F_\zeta(y)), \qquad (s,t)\in\mathcal{P}. \tag{71}$$

We verify exemplarily that a source $S = (S^1,\cdots,S^d)$ in $\mathbb{R}^d$ whose components $S^i$ are modelled according to (71) is $\alpha$-contrastive for a number of popular copula densities $c_\theta$.

**Corollary 2.** *Let $S = (S^1,\cdots,S^d)$ be a stochastic process whose independent components $S^i$ are modelled according to (71) for each $i\in[d]$ with copula-density $c_i$ belonging to one of the following popular classes:*

(i) (Clayton) $\qquad c_i(x,y) = (1+\theta)(xy)^{(-1-\theta)}(-1+x^{-\theta}+y^{-\theta})^{(-2-1/\theta)}$

*where $\theta\in(-1,\infty)\setminus\{0,-\tfrac{1}{2}\}$;*

---

[24] Lemma 1 (v) guarantees that such a set $\mathcal{P}$ exists.

[25] Recall that, by convention, we write $\zeta^i_{s,t}(x)\equiv\zeta^i_{s,t}(x_i, x_{i+d})$ for $x = (x_1,\cdots,x_{2d})\in\mathbb{R}^{2d}$.

(ii) (Gumbel) $c_i(x,y) = 1 + \theta(1-2x)(1-2y), \qquad \theta \in [-1,1] \setminus \{0\};$

(iii) (Frank) $c_i(x,y) = \dfrac{\theta e^{\theta(x+y)}(e^\theta - 1)}{(e^\theta - e^{\theta x} - e^{\theta y} + e^{\theta(x+y)})^2}, \qquad \theta \in \mathbb{R} \setminus \{0\}.$

*Then $S$ is $\alpha$-contrastive.*

*Proof.* This is a direct consequence of Proposition 1 upon checking that each of the copula densities (i), (ii) and (iii) satisfies (69). This, however, follows from inspection and a straightforward computational verification. □

6.2. **Popular Gaussian Processes and Geometric Brownian Motion are $\gamma$-Contrastive.** Given an interval $\mathbb{I}$ and functions $\mu : \mathbb{I} \to \mathbb{R}^d$ and $\kappa : \mathbb{I}^{\times 2} \to \mathrm{GL}_d(\mathbb{R})$, we write $S \sim \mathcal{GP}_{\mathbb{I}}(\mu,\kappa)$ to denote that $S = (S_t)_{t\in\mathbb{I}}$ is a Gaussian Process in $\mathbb{R}^d$ with mean $\mu = (\mu_i)$ and covariance $\kappa = (\kappa^{ij})$. In the following we assume that any pair $(\mu,\kappa)$ considered is such that each process $S \sim \mathcal{GP}(\mu,\kappa)$ admits a version with continuous sample paths.

**Lemma 7.** *Let $S \sim \mathcal{GP}_{\mathbb{I}}(\mu,\kappa)$ be a (continuous) Gaussian process in $\mathbb{R}^d$ with diagonal covariance function $\kappa \equiv (\kappa^{ij}) = (\kappa^{ij}\delta_{ij})$. Then $S$ is $\gamma$-contrastive if and only if there exist pairs $\mathfrak{p}_0, \mathfrak{p}_1, \mathfrak{p}_2 \in \Delta_2(\mathbb{I})$ such that*

$$\left(\frac{\kappa^i_{\mathfrak{p}_1} \cdot \kappa^i_{\mathfrak{p}_2} \cdot [k^i_{\mathfrak{p}_0} - (\kappa^i_{\mathfrak{p}_0})^2]^2}{[k^i_{\mathfrak{p}_1} - (\kappa^i_{\mathfrak{p}_1})^2] \cdot [k^i_{\mathfrak{p}_2} - (\kappa^i_{\mathfrak{p}_2})^2] \cdot (\kappa^i_{\mathfrak{p}_0})^2}\right)_{i\in[d]} \in \ (\mathbb{R}^d \setminus \nabla^\times) \tag{72}$$

*for the auxiliary functions*

$$\kappa^i_{s,t} := \kappa^{ii}(s,t) \quad \text{and} \quad k^i_{s,t} := \kappa^{ii}(s,s)\cdot\kappa^{ii}(t,t).$$

*Proof.* Let $i \in [d]$ be fixed, and $(s,t) \in \Delta_2(\mathbb{I})$ be arbitrary. Since $S^i \sim \mathcal{GP}(\mu_i, \kappa^{ii})$, we find that[26] $(S^i_s, S^i_t) \sim \mathcal{N}(\mu^i_{s,t}, \Sigma^i_{s,t})$ with $\mu^i_{s,t} := (\mu_i(s), \mu_i(t))$ and $\Sigma^i_{s,t} := (\kappa^{ii}(t_\nu, t_{\tilde\nu}))_{\nu,\tilde\nu=1,2}$ for $t_\nu := s$ and $t_{\tilde\nu} := t$, so the density $\zeta^i_{s,t}$ of $(S^i_s, S^i_t)$ exists, is smooth on $\mathbb{R}^2$ and reads

$$\zeta^i_{s,t}(x,y) = c_i(s,t)\cdot\exp\big(\varphi_i(s,t,x,y)\big) \qquad \text{with}$$
$$\varphi_i(s,t,x,y) = \frac{\rho_{s,t}}{(1-\rho^2_{s,t})\sigma_s\sigma_t}\cdot xy \ + \ \eta_i(s,t,x) \ + \ \tilde\eta_i(s,t,y)$$

for $c_i(s,t) = (4\pi^2\sigma_s^2\sigma_t^2(1-\rho^2_{s,t}))^{-1/2}$ and $-\frac{1}{2}((x,y)-\mu^i_{s,t})^\intercal\cdot[\Sigma^i_{s,t}]^{-1}\cdot((x,y)-\mu^i_{s,t}) - \frac{\rho_{s,t}}{(1-\rho^2_{s,t})\sigma_s\sigma_t}\cdot xy =: \eta_i(s,t,x) + \tilde\eta_i(s,t,y)$ and with the (auto-)correlations

$$\sigma_r := \sqrt{\kappa^{ii}(r,r)} \qquad \text{and} \qquad \rho_{s,t} := \frac{\kappa^{ii}(s,t)}{\sqrt{\kappa^{ii}(s,s)\kappa^{ii}(t,t)}} = \frac{\kappa^i_{s,t}}{\sqrt{k^i_{s,t}}}.$$

Consequently, the mixed log-derivatives $\xi^i_{s,t} := \partial_x\partial_y \log(\zeta^i_{s,t})$ are given as

$$\xi^i_{s,t}(x,y) = \partial_x\partial_y\varphi_i(s,t,x,y) = \frac{\kappa^i_{s,t}}{k^i_{s,t} - (\kappa^i_{s,t})^2}. \tag{73}$$

Since by Definition 7 the process $S$ is $\gamma$-contrastive iff there are $\mathfrak{p}_0, \mathfrak{p}_1, \mathfrak{p}_2 \in \Delta_2(\mathbb{I})$ with $\big(\psi(\xi^i_{\mathfrak{p}_0}, \xi^i_{\mathfrak{p}_1}, \xi^i_{\mathfrak{p}_2})\big)_{i\in[d]} = \left(\frac{\xi^i_{\mathfrak{p}_1}\xi^i_{\mathfrak{p}_2}}{(\xi^i_{\mathfrak{p}_0})^2}\right)_{i\in[d]} \in (\mathbb{R}^d \setminus \nabla^\times)$, the lemma now follows from (73). □

We verify the above contrastivity condition for a number of popular Gaussian processes.

**Proposition 2.** *Let $S = (S_t)_{t\in\mathbb{I}} = (S^1, \cdots, S^d)$ be an IC stochastic process in $\mathbb{R}^d$ with $S^i \sim \mathcal{GP}(\mu_i, \kappa_i)$ for each $i \in [d]$. Then $S$ is $\gamma$-contrastive in each of these classical cases:*

[26] We write $Z \sim \mathcal{N}(\mu, \Sigma)$ to say that $Z$ is normally distributed with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$.

(i) *Including the family of $\gamma$-exponential processes,*[27] *suppose that for each $i \in [d]$ it holds*

$$\kappa_i(s,t) = \exp\biggl(-\biggl[\frac{|t-s|}{\alpha_i}\biggr]^{\gamma_i}\biggr)$$

*with $\gamma \equiv (\gamma_i)_{i\in[d]} \in (0,2]^d$ and $\alpha \equiv (\alpha_i)_{i\in[d]} \in (\mathbb{R}_\times)^{\times d} \setminus \mathcal{N}_\gamma$ for one of the Lebesgue nullsets $\mathcal{N}_\gamma \subset \mathbb{R}^d$ defined in the proof below.*

(ii) *Each of the $S^i$ is an Ornstein-Uhlenbeck process,*

$$\mathrm{d}S^i_t = \theta_i \cdot (\mu_i - S^i_t)\,\mathrm{d}t \,+\, \sigma_i\,\mathrm{d}B^i_t, \quad S^i_0 = a_i \qquad (i\in[d]) \tag{74}$$

*with $a_i, \mu_i \in \mathbb{R}$ and $\sigma \equiv (\sigma_i)_{i\in[d]} \in \mathbb{R}^d_{>0}$ and $\theta \equiv (\theta_i)_{i\in[d]} \in \mathbb{R}^d_{>0} \setminus \tilde{\mathcal{N}}$ for one of the Lebesgue nullsets $\tilde{\mathcal{N}} \subset \mathbb{R}^d$ defined in the proof below.*

(iii) *The coordinate processes of $S$ are fractional Brownian motions with pairwise distinct Hurst indices, that is*

$$\kappa^i(s,t) = \frac{1}{2}(|t|^{2H_i} + |s|^{2H_i} - |t-s|^{2H_i}) \qquad (i\in[d]);$$

*for some $(H_i)_{i\in[d]} \in (0,1)^d \setminus \nabla^\times$.*

(iv) *Denoting $s\wedge t := \min(s,t)$, it holds that*

$$\kappa_i(s,t) = \int_0^{s\wedge t} \eta_i(r)\,\mathrm{d}r \qquad \text{for each} \quad i\in[d],$$

*with functions $\eta_1,\ldots,\eta_d : \mathbb{I}\to\mathbb{R}$ for which there are $r_0, r_1 \in \mathbb{I}$ such that the products $\{\eta_i(r_0)\cdot\eta_j(r_1) \mid i,j\in[d]\}$ are pairwise distinct. This is includes deterministic signals perturbed by white noise, i.e. signals $S = (S^1_t,\cdots,S^d_t)_{t\in\mathbb{I}}$ which, for $(B^i_t)_{t\geq 0}$ some standard Brownian motion in $\mathbb{R}^d$, are given by*

$$\mathrm{d}S^i_t = \mu_i(t)\,\mathrm{d}t \,+\, \sigma_i(t)\,\mathrm{d}B^i_t \qquad \text{for each} \quad i\in[d]$$

*with $\mu_i, \sigma_i : \mathbb{I}\to\mathbb{R}$ integrable and continuous such that the entries of $(\sigma_i^2(r_0)\cdot\sigma_j^2(r_1))_{i,j\in[d]}$ are pairwise distinct for some $r_0, r_1 \in \mathbb{I}$.*

*Proof.* We apply Lemma 7 by showing that for each case there are $\mathfrak{p}_0, \mathfrak{p}_1, \mathfrak{p}_2 \in \Delta_2(\mathbb{I})$ for which (72) holds. Write $\Xi_i$ for the $i^{\text{th}}$ component of (72) and let $|(s,t)| := |t-s|$.

(i): Fix any $\mathfrak{p}_0, \mathfrak{p}_1 \in \Delta_2(\mathbb{I})$ with $|\mathfrak{p}_0| \neq |\mathfrak{p}_1|$ and take $\mathfrak{p}_2 := \mathfrak{p}_0$. Then for each $i\in[d]$ we have $\Xi_i = \tilde{\Xi}^{(1)}_i / \tilde{\Xi}^{(0)}_i$ for the factors

$$\tilde{\Xi_i}^{(\nu)} = \frac{\kappa^i_{\mathfrak{p}_\nu}}{1-(\kappa^i_{\mathfrak{p}_\nu})^2} = \bigl([\kappa^i_{\mathfrak{p}_\nu}]^{-1} - \kappa^i_{\mathfrak{p}_\nu}\bigr)^{-1} = \tfrac{1}{2}\left(\sinh\biggl(\biggl[\frac{|\mathfrak{p}_\nu|}{\alpha_i}\biggr]^{\gamma_i}\biggr)\right)^{-1}.$$

Hence we have the parametrisation $\alpha_i \mapsto \Xi_i \equiv \Xi_i(\alpha_i)$ given by

$$\Xi_i = \sinh\biggl(\biggl[\frac{|\mathfrak{p}_0|}{\alpha_i}\biggr]^{\gamma_i}\biggr)\cdot\biggl[\sinh\biggl(\biggl[\frac{|\mathfrak{p}_1|}{\alpha_i}\biggr]^{\gamma_i}\biggr)\biggr]^{-1},$$

which for $|\mathfrak{p}_1| \neq |\mathfrak{p}_0|$ and $\gamma_i$ fixed is differentiable and strictly monotone in $\alpha_i > 0$. Denoting by $\phi_i$ the associated (differentiable) inverse of $\alpha_i \mapsto \Xi_i(\alpha_i)$ (for $\gamma$ and $\mathfrak{p}_0, \mathfrak{p}_1$ fixed), we find that $(\Xi_i(\alpha_i))_{i\in[d]} \in (\mathbb{R}^d \setminus \nabla^\times)$ for any $(\alpha_i)_{i\in[d]}$ not contained in $\mathcal{N}_\gamma := (\phi_1\times\cdots\times\phi_d)(\nabla^\times)$.

The proofs of (ii) and (iii) are similar and hence deferred to Appendix A.7.

(iv): As the numbers $\theta_{ij} := \eta_i(r_0)\cdot\eta_j(r_1)$, $(i,j)\in[d]\times[d]$, are pairwise distinct and (thus) non-zero, the continuity of the functions $\vartheta_{ij} : \mathbb{I}^{\times 2} \ni (s,t) \mapsto \eta_i(s)\cdot\eta_j(t)$ allows us to find pairs

[27] Cf. [66, Sect. 4.2 (pp. 84 ff.)].

$(s_0,t_0),(s_1,t_1) \in \Delta_2(\mathbb{I})$ such that for the rectangle $R := [s_0,t_0]\times[s_1,t_1] \subseteq \mathbb{I}^{\times 2}$ the associated integrals

$$\int_R \vartheta_{ij}\,\mathrm{d}s\,\mathrm{d}t, \quad (i,j)\in[d]^{\times 2}, \quad \text{are pairwise distinct and non-zero.} \tag{75}$$

Clearly then, (75) implies that for $\iota_i(s,t) := \int_s^t \eta_i(r)\,\mathrm{d}r$, the numbers

$$\iota_i(s_0,t_0)\cdot\iota_j(s_1,t_1),\ (i,j)\in[d]\times[d], \quad \text{are pairwise dinstinct} \tag{76}$$

(Note further that by (75), $s_0$ and $s_1$ may be chosen such that in addition to (76) it holds $\iota_i(0,s_\nu)\neq 0$ for each $i\in[d]$.) Now by setting $\mathfrak{p}_2 := \mathfrak{p}_0$ with $\mathfrak{p}_\nu := (s_\nu,t_\nu)$ for $\nu = 0,1$ (notice that $\mathfrak{p}_0 \neq \mathfrak{p}_1$ by (76)), we once more find that $\Xi_i = \tilde{\Xi}_i^{(1)}/\,\Xi_i^{(0)}$ for each $i\in[d]$, this time for the factors

$$\tilde{\Xi}_i^{(\nu)} = \frac{\kappa^i_{\mathfrak{p}_\nu}}{k^i_{\mathfrak{p}_\nu} - (\kappa^i_{\mathfrak{p}_\nu})^2} = \frac{\iota_i(0,s_\nu)}{\iota_i(0,s_\nu)\cdot\iota_i(0,t_\nu) - \iota_i(0,s_\nu)^2} = (\iota_i(s_\nu,t_\nu))^{-1}.$$

Consequently, the entries of $(\Xi_i)_{i\in[d]} = \left(\frac{\iota_i(s_0,t_0)}{\iota_i(s_1,t_1)}\right)_{i\in[d]}$ are pairwise distinct. Indeed, assuming otherwise that $\Xi_i = \Xi_j$ for some $i\neq j$, we find that

$$\iota_i(s_0,t_0)\cdot\iota_j(s_1,t_1) = \iota_j(s_0,t_0)\cdot\iota_i(s_1,t_1), \qquad \text{contradicting (76).}$$

Hence $(\Xi_i)_{i\in[d]} \in (\mathbb{R}^d\setminus\nabla^\times)$ as desired. □

The proposition below concludes our short compilation of applicable source models.

**Proposition 3.** *Let $S = (S_t)_{t\geq 0} = (S^1,\cdots,S^d)$ be an IC geometric Brownian motion in $\mathbb{R}^d$, i.e. suppose that there is a standard Brownian motions $B = (B^1_t,\cdots,B^d_t)_{t\geq 0}$ such that*

$$\mathrm{d}S^i_t = S^i_t\cdot\big(\mu_i(t)\,\mathrm{d}t \ + \ \sigma_i(t)\,\mathrm{d}B^i_t\big), \quad S^i_0 = s^i_0 \qquad (i\in[d])$$

*for some $s^i_0 > 0$ and continuous functions $\mu_i : \mathbb{I}\to\mathbb{R}$ and $\sigma_i : \mathbb{I}\to\mathbb{R}_{>0}$. Then $S$ has spatial support $D_S = \mathbb{R}^d_+$, and $S$ is $\gamma$-contrastive if there are $r_0,r_1\in\mathbb{I}$ for which the numbers $\{\sigma_i^2(r_0)\cdot\sigma_j^2(r_1) \mid (i,j)\in[d]\times[d]\}$ are pairwise distinct.*

*Proof.* The proof is similar to the ones above and hence deferred to Appendix A.8. □

## 7. Signature Cumulants as Contrast Function

Many results in statistics, including Corollary 1 via (5), are based on the well-known facts that laws of $\mathbb{R}^d$-valued random variables are often characterised by their *moments*, and that statistical independence turns into simple algebraic relations when expressed in terms of *cumulants*. Our main object of interest are $\mathcal{C}_d$-valued random variables (stochastic processes), for which the so-called expected signature provides a natural generalisation of the classical moment sequence. Similar to classical moments, these *signature moments* can characterize the laws of stochastic processes and also give rise to *signature cumulants* that quantify statistical dependence of stochastic processes (that is, between their coordinates and over time). Building on these results, we propose a generalisation of Comon's classical contrast function (5) and use it to recover an IC source from its nonlinear mixtures by turning our above identifiability results (Theorem 2 and Theorem 3) into a practical optimization procedure in the spirit of Corollary 1 (Theorem 4).

*Remark* 7.1. In this section, we restrict our exposition to stochastic processes whose sample paths are smooth (i.e., of bounded variation[28]), and further assume that the expected signature

[28] A path $x = (x_t)_{t\in[0,1]} \in \mathcal{C}_d$ is called *of bounded variation* if its variation norm $\|x\|_{1\text{-var}} := |x_0| + \sup\sum|x_{t_{i+1}} - x_{t_i}|$ is finite, where the supremum is taken over all finite partitions $\{0\leq t_1\leq\cdots\leq t_n\leq 1\}$ $(n\in\mathbb{N})$ of $[0,1]$; cf. also Section B.2.

of these processes (defined below) exists and characterizes their law. These assumptions can be avoided by using rough integration and tensor normalization, but since this requires background in rough path theory and is not central to our methodology, we simply refer the interested reader to [30, 52] and [17, 16], respectively. Further, we assume $\mathbb{I} = [0,1]$ without loss of generality.

Denote by $[d]^\star := \bigcup_{m\geq 0}[d]^{\times m}$ the set of all multi-indices[29] with entries in $[d] = \{1,\dots,d\}$.

*Definition* 8 (Expected Signature). For $Y = (Y_t^1,\cdots,Y_t^d)_{t\in[0,1]}$ a stochastic process in $\mathbb{R}^d$ with sample-paths of bounded variation, the collection of real numbers (if it exists) $\mathfrak{S}(Y) := (\sigma_{\boldsymbol{i}}(Y))_{\boldsymbol{i}\in[d]^\star}$ defined by the expected iterated Stieltjes integrals

$$\sigma_{\boldsymbol{i}}(Y) := \mathbb{E}\left[\int_{0\leq t_1\leq t_2\leq\cdots\leq t_m\leq 1} \mathrm{d}Y_{t_1}^{i_1}\mathrm{d}Y_{t_2}^{i_2}\cdots\mathrm{d}Y_{t_m}^{i_m}\right] \quad \text{for} \quad \boldsymbol{i} = (i_1,\dots,i_m), \tag{77}$$

with $\sigma_\emptyset(Y) := 1$, is called *the expected signature of* $Y$.

The expected signature of a stochastic process plays a role analogous to the one played by the sequence of moments for a vector-valued random variable, and analogous to the case of classical moments, for many statistical purposes the concept of cumulants is better suited. (See Appendix C for details.) This leads to the notion of signature cumulants [7] below.

*Definition* 9 (Signature Cumulants). For $Y$ a stochastic process in $\mathbb{R}^d$ with sample-paths of bounded variation, the collection of real numbers[30]

$$(\kappa_{\boldsymbol{i}}(Y))_{\boldsymbol{i}\in[d]^\star} := \log[\mathfrak{S}(Y)] \tag{78}$$

is called *the signature cumulant* of $Y$. We further define

$$\bar{\kappa}_{\boldsymbol{i}}(Y) := \frac{\kappa_{\boldsymbol{i}}(Y)}{\kappa_{11}(Y)^{\eta_1(\boldsymbol{i})/2}\cdot\ldots\cdot\kappa_{dd}(Y)^{\eta_d(\boldsymbol{i})/2}} \quad \text{for} \quad \boldsymbol{i} = (i_1,\dots,i_m)\in[d]^\star, \tag{79}$$

where $\eta_\nu(\boldsymbol{i})$ denotes the number of times the index-entry $\nu$ appears in $\boldsymbol{i}$. We refer to $(\bar{\kappa}_{\boldsymbol{i}}(Y))_{\boldsymbol{i}\in[d]^\star}$ as the *standardized signature cumulant* of $Y$.

Just as for standardized classical cumulants, the normalisation (79) yields the benefit of *scale invariance* which facilitates our usage of signature cumulants as a contrast function.

*Notation* 7.1. For convenience, we denote by $[d]^\star_+$ the family of all finite sums[31] of indices in $[d]^\star$, and for any such sum $\boldsymbol{i} \equiv \boldsymbol{i}_1 + \ldots + \boldsymbol{i}_\ell \in [d]^\star_+$ define $\kappa_{\boldsymbol{i}} := \kappa_{\boldsymbol{i}_1} + \ldots + \kappa_{\boldsymbol{i}_\ell}$.

Recall that a random vector in $\mathbb{R}^d$ has independent components if and only if all of its cross-cumulants vanish. It was shown in [7] that signature cumulants generalise this classical relation to an algebraic characterisation of statistical independence between stochastic processes. This is particularly useful in our context as it yields a natural and explicitly computable contrast function for path-valued random variables, as desired for nonlinear ICA.

To this end, we define the *shuffle product* of two multi-indices $\boldsymbol{i} = (i_1,\dots,i_m)$ and $\boldsymbol{j} = (i_{m+1},\dots,i_{m+n})$ in $[d]^\star$ as the element of $[d]^\star_+$ given by

$$\boldsymbol{i} \sqcup\!\sqcup \boldsymbol{j} := \sum_\tau (i_{\tau(1)},\dots,i_{\tau(m+n)}) \ \in \ [d]^\star_+ \tag{80}$$

where the sum is taken over the family of permutations

$$\{\tau\in S_{m+n} \mid \tau(1)<\cdots<\tau(m) \ \text{ and } \ \tau(m+1)<\cdots<\tau(m+n)\}.$$

[29] We define $[d]^{\times 0} := \{\emptyset\}$, with $\emptyset$ the empty set.

[30] The log in (78) denotes the logarithm on the space of formal power series, see Remark C.2 and [7].

[31] Mathematically, $[d]^\star_+$ is an additive subgroup of the free algebra over the monoid $[d]^\star$, see Remark C.2.

**Proposition 4.** *For any stochastic process $Y = (Y^1, \cdots, Y^d)$ in $\mathbb{R}^d$ whose expected signature exists, the coordinate processes $Y^1, \ldots, Y^d$ are mutually independent if and only if*

$$\bar{\kappa}_{\mathrm{IC}}(Y) := \sum_{k=2}^{d} \sum_{\boldsymbol{q}\in\mathcal{W}_k} \bar{\kappa}_{\boldsymbol{q}}(Y)^2 = 0 \tag{81}$$

*where* $\mathcal{W}_k := \big\{\boldsymbol{i} \sqcup \boldsymbol{j} \;\mid\; \boldsymbol{i} \in [k-1]^\star \setminus \{\emptyset\},\ \boldsymbol{j} \in \{k\}^{\times m},\ m \geq 1\big\} \;\subset\; [d]^\star_+$.

*Proof.* Observe that the coordinate processes $Y^1, \ldots, Y^d$ are mutually independent iff:

$$\text{for each } \; 2 \leq k \leq d, \quad \text{the process } \; Y^k \; \text{ is independent of } \; (Y^1, \cdots, Y^{k-1}).$$

The asserted characterisation is a direct consequence of this and [7, Theorem 1.1 (iii)]. □

Combining Proposition 4 with Theorems 2 and 3 leads to the following instance of (19) for the inversion '$X \mapsto S$' desired in (2) (cf. Corollary 1).

**Theorem 4.** *Let the process $S$ in (13) be $\alpha$-, $\beta$- or $\gamma$-contrastive with sample-paths of bounded variation. Then it holds with probability one that*

$$\left[\underset{h\in\Theta}{\arg\min}\; \bar{\kappa}_{\mathrm{IC}}(h\cdot X)\right]\cdot X \;\subseteq\; \mathrm{DP}_d \cdot S \tag{82}$$

*for any family of transformations $\Theta \subseteq C^{3,3}(D_X)$ with $\Theta\big|_{D_X} \cap \big(\mathrm{DP}_d(D_S)\cdot f^{-1}\big)\big|_{D_X} \neq \emptyset$.*

*Remark* 7.2.

- (i) For $\Theta \subseteq \mathrm{GL}_d$ and under the temporally degenerate hypothesis of Theorem 1, the algorithm (82) reduces to Comon's optimisation (4) for $\phi = \phi_c$ since
$$\bar{\kappa}_{\mathrm{IC}}\big((Y\cdot t)_{t\in[0,1]}\big) \;=\; \phi_c(Y) \quad \text{if } Y \text{ is a random vector in } \mathbb{R}^d,$$
cf. Remark C.3.
- (ii) Regarding implementations of (82), one may choose to realise the above domain $\Theta$ by way of an Artificial Neural Network, see e.g. Section 9.3. This choice is mathematically justified by the fact that neural networks can be designed as universal approximators to $C^{3,3}(D_X)$ [79] with a favourable convergence topology [63] (cf. also Remark 8.1).
- (iii) In practice, only discrete-time observations $(X_t)_{t\in\mathcal{I}}$ of $X$ for a finite $\mathcal{I} \subset \mathbb{I}$ are available. By identifying such data with a continuous, bounded variation process via piecewise linear interpolation of the points $\{X_t \mid t \in \mathcal{I}\}$, discrete time-series sampled from continuous-time stochastic processes are covered by the above framework, cf. Section 8. A simple adaptation of Theorems 2 & 3 further allows to extend our identifiability approach to discrete time-series that are not necessarily generated from continuous-time processes, with Theorem 4 remaining applicable as stated upon piecewise-linear interpolation of this time-discrete data; see Appendix A.10 for details.
- (iv) The contrast function $\bar{\kappa}_{\mathrm{IC}}$ can be efficiently approximated by restricting the summation in (81) to multindices $(i_1, \ldots, i_m)$ up to a maximal length $m \leq \hat{\mu}$ and estimating the remaining summands using the unbiased minimum-variance estimators for signature cumulants introduced in [7, Section 4]. For the latter, a more naive but straightforward approach that is sufficient for our experiments is to just use the Monte-Carlo estimator. See Sections 8.1 & 8.3 for details.

## 8. Statistical Consistency

In practical applications, realisations of the mixture $X$ are usually not observed as continuous signals in $\mathcal{C}_d$ but rather as discrete, sequentially ordered collections of data points in $\mathbb{R}^d$. Formally, such data can be modelled as a *discrete time-series*

$$\mathfrak{x} := (X_t(\omega))_{t\in\mathcal{I}} \quad \text{with} \quad \mathcal{I} := \{0 = t_0 < t_1 < \ldots < t_{n-1} = 1\},^{32} \tag{83}$$

where in empirical language the dissection $\mathcal{I}$ of $[0,1]$ can be regarded as a 'protocol' describing a sequence of measurements of $X$ carried out per unit time with frequency $n$.

The time series data (83) then is typically collected over not just one but several ($\nu \in \mathbb{N}$) units of time, and different observations of $X$ ($k \in \mathbb{N}$) may vary in the frequency at which measurements are made. This gives rise to the data scheme

$$\mathfrak{x}^{(k)} \equiv \big(\mathfrak{x}_1^{(k)}, \mathfrak{x}_2^{(k)}, \cdots\big) := (X_t(\omega))_{t\in\mathcal{J}_k}, \quad \text{for} \quad \mathcal{J}_k := \mathcal{I}_1^{(k)} \sqcup \mathcal{I}_2^{(k)} \sqcup \ldots = \bigsqcup\nolimits_{\nu\in\mathbb{N}} \mathcal{I}_\nu^{(k)} \tag{84}$$

a dissection[33] of $[0,\infty)$ such that $\mathcal{I}_1^{(k)} < \mathcal{I}_2^{(k)} < \ldots$ and $|\mathcal{I}_\nu^{(k)}| = |\mathcal{I}_1^{(k)}| < \infty$ for each $\nu \in \mathbb{N}$, with $\mathcal{I}_1^{(k)}$ a dissection of $[0,1]$. Any such family $(\mathcal{J}_k)_{k\in\mathbb{N}}$ will be called a *protocol*, where the index $k \in \mathbb{N}$ enumerates observations of $X$ with varying 'sampling resolution' $|\mathcal{I}_1^{(k)}| =: n_k$.

Adopting the ergodicity perspective common in time-series analysis and signal processing, the relevant statistical information of $X$, in our case: the signature coordinates (78), may be averaged from the discretely-sampled observation $\mathfrak{x}^{(k)}$ of a *single*, sufficiently long realisation of $X$.

For this, we may generalise established (i.e. typically equispaced; e.g. [28, Section 3] and [84] for an overview) observation schemes by assuming that the data $\mathfrak{x}^{(k)}$ be obtained according to nothing but the assumptions (for $\|\mathcal{J}_k\|$ the mesh-size of $\mathcal{J}_k$; cf. (95))

$$X = (X_t)_{t\geq 0} \quad \text{and} \quad \lim_{k\to\infty} \|\mathcal{J}_k\| = 0 \tag{85}$$

i.e. the requirements that the continuous observable $X$ be 'infinitely long' (i.e., defined over $[0,\infty)$) and the frequency $|\mathcal{I}_1^{(k)}|$ of observations per unit time be going to infinity.

A protocol $(\mathcal{J}_k)_{k\in\mathbb{N}}$ as in (84) & (85) will be called *exhaustive* with *base lengths* $n_k := |\mathcal{I}_1^{(k)}|$.

Given data $(\mathfrak{x}^{(k)})$ associated to an exhaustive observation $\big(X, (\mathcal{J}_k)\big)$ via (84), it is *the aim of this section* to establish conditions under which the optimisation scheme of Theorem 4 when applied to $(\mathfrak{x}^{(k)})$ (in lieu of $X$) yields a sequence of approximations $(\hat{\theta}_k)$ of the true demixing transformation $f^{-1}$ which is *statistically consistent* in the sense that

$$\lim_{k\to\infty} \operatorname{dist}(\hat{\theta}_k \cdot X,\, \mathrm{DP}_d \cdot S) = 0 \tag{86}$$

almost surely or in probability (where the distance is taken with respect to the uniform norm on $\mathcal{C}_d$). Since the original optimisation (82) is composed of three ('limiting') operations that each involve an 'infinite amount of information', namely the *infinite series* $\bar{\kappa}_{\mathrm{IC}}$ from (81) whose summands (79) are each defined by taking *expectations* of nonlinear functionals (77) of the *continuous-time stochastic processes* $Y = \theta \cdot X$, one may expect the consistency (86) to result as a combination of the following three sublimits:

[32] For $\omega \in \Omega$ any fixed elementary event in the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ underlying $X$.

[33] We call $\mathcal{J}_T := \{t_0 < t_1 < \ldots\}$ a *dissection of* $[0,\infty)$ if $t_0 = 0$ and $t_j \nearrow \infty$ as $j \to \infty$; note that this guarantees $\|\mathcal{J}_k\| < \infty$. Also, we then assume $X$ to be defined over the full positive time-axis $[0,\infty)$.

- *Capping Limit* (Section 8.1). In practice, only finitely many summands of the infinite statistics $\bar{\kappa}_{\mathrm{IC}}$ from (81) can be computed from the data, which is to say that the series

$$\bar{\kappa}_{\mathrm{IC}}(Y) = \sum_{m=2}^{\infty} \sum_{\boldsymbol{q}\in\mathfrak{C}_m} \bar{\kappa}_{\boldsymbol{q}}(Y)^2 \tag{87}$$

with $\mathfrak{C}_m \subset [d]^\star_+$ denoting the set of all cross-shuffles $\mathfrak{C} := \bigsqcup_{k=2}^{d} \mathcal{W}_k$ of word-length $m \in \mathbb{N}$, needs to be *capped* at some index $m = m_0$. Denoting this capped series by

$$\bar{\kappa}_{\mathrm{IC}}^{[m_0]}(Y) := \sum_{m=2}^{m_0} \sum_{\boldsymbol{q}\in\mathfrak{C}_m} \bar{\kappa}_{\boldsymbol{q}}(Y)^2 \tag{88}$$

we show that in the *capping limit* $m_0 \to \infty$ the minimizers of $\theta \mapsto \bar{\kappa}_{\mathrm{IC}}^{[m_0]}(\theta\cdot X)$ approach those of (87) with respect to a naturally chosen topology on $\Theta$, providing the first ingredient for the consistency limit (86).

- *Interpolation Limit* (Section 8.2). As mentioned in (83), the mixture $X$ is usually observed along a discrete set of time-points $\mathcal{I}$ rather than continuously over time. By way of their piecewise-linear interpolation $\hat{X}_{\mathcal{I}}$, these discrete observations $(X_t)_{t\in\mathcal{I}}$ can be reinterpreted as $\mathcal{C}_d$-valued data, which then allows to approximate the summands in (88) via

$$\bar{\kappa}_{\boldsymbol{q}}(\theta\cdot X) \;\approx\; \bar{\kappa}_{\boldsymbol{q}}(\hat{X}^\theta_{\mathcal{I}}), \qquad \text{for}\;\; \hat{X}^\theta_{\mathcal{I}}\;\; \text{the linear interpolant of}\;\; (\theta\cdot X_t)_{t\in\mathcal{I}}. \tag{89}$$

By showing that (89) defines a $\Theta$-uniform approximation as $\|\mathcal{I}\| \to 0$, we obtain that the minimizers of $\theta \mapsto \bar{\kappa}_{\mathrm{IC}}^{[m_0]}(\hat{X}^\theta_{\mathcal{I}})$ converge to those of (88); our second ingredient for (86).

- *Ergodicity Limit* (Section 8.3). Finally, as the data (84) that is actually available is but a single realisation of the discrete time-series $(X_t)_{t\in\mathcal{J}_k}$, we propose to approximate the above approximations (89) by estimating their constituent signature moments (77) via

$$\sigma_{\boldsymbol{i}}\big(\hat{X}^\theta_{\mathcal{I}_1^{(k)}}\big) \;\approx\; \frac{1}{T}\sum_{\nu=1}^{T} \mathfrak{sig}_{\boldsymbol{i}}\big(\hat{\mathfrak{x}}^{\theta|k}_\nu\big), \qquad \text{for}\;\; \hat{\mathfrak{x}}^{\theta|k}_\nu\;\; \text{the linear interpolant of}\;\; \theta\cdot\mathfrak{x}^{(k)}_\nu \tag{90}$$

and where $\mathfrak{sig}_{\boldsymbol{i}}(Y)$ denotes the iterated integrals inside the expectation (77) (see also Appendix C). Showing that many popular time-series models and stochastic processes are covered by the class of processes for which the (90)-based estimation scheme for $\bar{\kappa}_{\boldsymbol{q}}(\hat{X}^\theta_{\mathcal{I}})$ is $\Theta$-uniformly consistent as $T \to \infty$, we obtain our third and final ingredient for (86).

In Section 8.4, these three sublimits are then combined to prove the statistical consistency of our nonlinear ICA method (Theorem 4) in the form of Theorem 5, which can be seen to contain the above consistency limit (86) as a special case.

(The majority of the proofs for this section are deferred to Appendix B as they are mostly technical and independent of the argumentation developed in the main body of this work.)

8.1. **The Capping Limit.** Let the subset $\mathfrak{C}_m$ of $[d]^\star_+$ denote the set of all cross-shuffles $\mathfrak{C} := \bigsqcup_{k=2}^{d} \mathcal{W}_k$ of fixed word-length $m$,[34] for $\mathcal{W}_2, \ldots, \mathcal{W}_d$ as in Proposition 4 and $m \in \mathbb{N}$.

Let further $\Theta$ be a given set of nonlinearities (as specified below), and for $(\kappa_q(Y) \mid q \in [d]^\star)$ as in (78) and $\theta \in \Theta$, consider the (in)finite cumulant series

$$Q(\theta) := \sum_{\nu=2}^{\infty} \sum_{\boldsymbol{q}\in\mathfrak{C}_\nu} c_{\boldsymbol{q}}^{-1}\cdot\kappa_{\boldsymbol{q}}(\theta\cdot X)^2 \qquad \text{and} \qquad Q_m(\theta) := \sum_{\nu=2}^{m} \sum_{\boldsymbol{q}\in\mathfrak{C}_\nu} c_{\boldsymbol{q}}^{-1}\cdot\kappa_{\boldsymbol{q}}(\theta\cdot X)^2 \tag{91}$$

[34] The *word-length* of an element $\boldsymbol{i} \in [d]^\star_+$ is defined as the maximal order of the (finitely many) indices in $[d]^\star_+$ whose formal sum is $\boldsymbol{i}$ (cf. Notation 7.1). Thus $\mathfrak{C}_m = V_m \cap \bigsqcup_{k=2}^{d} \mathcal{W}_k$ in the language of Section C.2.2.

for $m \geq 2$, where $c_{\boldsymbol{q}}$ denotes the number of monomials in $\boldsymbol{q}$ (cf. Remark A.9).

We make following technical compatibility assumptions on $\Theta$ and $X$.

(See Appendix C.2.2 for notation.)

*Assumption* 1. Let $\Theta \subseteq C(D_X; \mathbb{R}^d)$ be equipped with the topology of compact convergence and suppose that $\Theta$ is compact, satisfies $\Theta \cdot X \subseteq \mathcal{BV}$ and is such that it holds with probability one that for every convergent sequence $(\theta_j)_{j\geq 1}$ in $\Theta$ we have

$$\sup_{j\geq 1} \|\theta_j \cdot X\|_{p\text{-var}} \; < \; \infty \qquad \text{for some} \;\; p \geq 1, \tag{92}$$

where $\|\cdot\|_{p\text{-var}}$ denotes the $p$-variation seminorm (196). On side of the signature moments (77), suppose that the expected signatures $\mathfrak{S}(\theta \cdot X) \equiv \mathbb{E}[\mathfrak{sig}(\theta \cdot X)]$, $\theta \in \Theta$, exist and characterize the law of their arguments, that their collection $\{\mathfrak{S}(\theta \cdot X) \mid \theta \in \Theta\}$ is $|\!|\!|\cdot|\!|\!|_\lambda$-bounded[35] for some $\lambda > 2$, and that for each $m \geq 1$ (with $\mathfrak{sig}_m := \pi_m \circ \mathfrak{sig}$, for $\mathfrak{sig}$ as in (187)),[36]

$$\mathbb{E}\left[\sup_{\theta\in\Theta} \left\|\mathfrak{sig}_m(\theta \cdot X)\right\|_m\right] \; < \; \infty. \tag{93}$$

(To avoid potential measurability problems, we may as well replace (93) by the (weaker) requirement that $\left[\mathbb{E}[\sup_{\theta\in\mathcal{F}} \|\mathfrak{sig}_m(\theta \cdot X)\|_m] \; < \; \infty, \;\; \forall \, \text{countable } \mathcal{F} \subseteq \Theta\right]$ if desired.)

*Remark* 8.1. Notice that the above conditions on $\Theta$ and $X$ are quite natural and well-established in the contexts of Artificial Neural Networks (ANNs) and Stochastic Analysis. Indeed: The topological requirement of compact convergence is typically met if $\Theta$ is given as the realisation space of an ANN, see e.g. [5, 63], while the assumptions $\Theta \cdot X \subseteq \mathcal{BV}$ resp. (92) hold for instance if $\Theta \subseteq C^1(D_X; \mathbb{R}^d)$ resp. if the elements of $\Theta$ are continuously differentiable with uniformly bounded Jacobians. The growth assumptions on the signature coordinates (including (93)), on the other hand, have been extensively studied, established and applied in the context of rough path analysis and statistics, see e.g. [15] and [7, 16].

**Lemma 8.** *Let $X$ and $\Theta$ be as described in Assumption 1. Then the following holds:*

(i) *the functions $Q, Q_m : \Theta \to \mathbb{R}$ given in (91) are continuous;*

(ii) *the capped objectives $Q_m$ approximate $Q$ uniformly as $m$ goes to infinity, in symbols:*

$$\lim_{m\to\infty} \|Q - Q_m\|_\Theta = 0 \qquad \textit{for} \qquad \|q\|_\Theta := \sup_{\theta\in\Theta} |q(\theta)| \, ;$$

(iii) *if $Q$ is uniquely minimized at $\theta_\star \in \Theta$, i.e. such that $Q(\theta) > Q(\theta_\star)$ if $\theta \neq \theta_\star$, then any ('minimising') sequence $(\theta^\star_m)$ in $\Theta$ such that $Q_m(\theta^\star_m) \leq \inf_{\theta\in\Theta} Q_m(\theta) + \eta_m$ for some $\eta_m \geq 0$ with $\lim_{m\to\infty} \eta_m = 0$ a.s., converges to $\theta_\star$ almost surely as $m \to \infty$.*

*Proof.* See Appendix B.1. □

The following result is but a reformulation of Lemma 8 in terms of the standardized signature cumulants (81). It is also illustrative of the consistency assertion in Theorem 5 below.

[35] As the source $S$ can be recovered up to (a componental permutation and) a monotone scaling only, we may assume wlog (cf. Lemma C.1 (vi)) that the set $\{\mathfrak{S}(\theta \cdot X) - 1 \mid \theta \in \Theta\} \subseteq V_{(0)}$ is $|\!|\!|\cdot|\!|\!|_\lambda$-bounded by 1.

[36] Some of the $\sup_\theta$-related (or $\operatorname{dist}(\,\cdot\,, \mathrm{DP}_d \cdot S)$-related) expressions in the following sections may be non-measurable, in which case any probability statements involving these expressions are to be understood in terms of outer measure (cf. [80, 82]).

**Proposition 5** (Capping Limit). *Let $X$ and $\Theta \subseteq C^{3,3}(D_X)$ fulfil Assumption 1, and suppose that there is a unique $\theta_\star \in \Theta$ such that $\theta_\star \cdot X$ is IC. Let $\bar{\kappa}_{\mathrm{IC}}^{[m]}$ be as in (88). Then for any sequence of minimizers $(\theta_m^\star)$ in $\Theta$ such that*

$$\bar{\kappa}_{\mathrm{IC}}^{[m]}(\theta_m^\star \cdot X) \;\leq\; \min_{\theta\in\Theta} \bar{\kappa}_{\mathrm{IC}}^{[m]}(\theta \cdot X) \;+\; \eta_m$$

*for some $(\eta_m) \subset \mathbb{R}_+$ with $\lim_{m\to\infty}\eta_m = 0$ a.s., it holds with probability one that*

$$\lim_{m\to\infty} \mathrm{dist}_{\|\cdot\|_\infty}\big(\theta_m^\star \cdot X,\, \mathrm{DP}_d \cdot S\big) = 0. \tag{94}$$

*Proof.* Since for any $\boldsymbol{q} \in \mathfrak{C}$ we have that $\kappa_{\boldsymbol{q}}(Y) = 0$ iff $\bar{\kappa}_{\boldsymbol{q}}(Y) = 0$ (recall (79) and Notation 7.1), it holds that $\arg\min_{\theta\in\Theta} Q_m(\theta) = \arg\min_{\theta\in\Theta} \bar{\kappa}_{\mathrm{IC}}^{[m]}(\theta\cdot X)$ for each $m \geq 2$. The convergence (94) is thus an immediate consequence Lemma 8 (iii) and Prop. 4/ Thm. 4. □

### 8.2. The Interpolation Limit.

Let $\mathbb{I}$ be a compact interval; say $\mathbb{I} = [0,1]$ wlog as above.

A finite subset $\mathcal{I}$ of $\mathbb{I}$ is called a *dissection of* $\mathbb{I}$ if it contains the boundary points of $\mathbb{I}$, and a sequence $(\mathcal{I}_\mu)_{\mu\geq 1}$ of dissections $\mathcal{I}_\mu \equiv \big\{t_0^{(\mu)}, \ldots, t_{n_\mu-1}^{(\mu)} \;\big|\; t_0^{(\mu)} < t_1^{(\mu)} < \ldots < t_{n_\mu-1}^{(\mu)}\big\}$ of $\mathbb{I}$ is called *refined* if the maximal distance $\|\mathcal{I}_\mu\|$ between two successive points in $\mathcal{I}_\mu$, the so-called *mesh-size* of $\mathcal{I}_\mu$, goes to zero as $\mu \to \infty$; in symbols:

$$\|\mathcal{I}_\mu\| := \max_{j\in[n_\mu-1]} \big|t_j^{(\mu)} - t_{j-1}^{(\mu)}\big| \;\longrightarrow\; 0 \qquad \text{as} \quad \mu \to \infty. \tag{95}$$

Writing $\hat{X}_{\mathcal{I}}^\theta$ for the piecewise linear interpolant[37] of the transformed data $X_{\mathcal{I}}^\theta := \big(\theta \cdot X_t\big)_{t\in\mathcal{I}}$, $\theta \in \Theta$, the next lemma shows that as $\|\mathcal{I}\| \to 0$ the statistic (cf. (91))

$$\widehat{Q}_m(Y) := \sum_{\nu=2}^{m} \sum_{\boldsymbol{q}\in\mathfrak{C}_\nu} c_{\boldsymbol{q}}^{-1} \cdot \kappa_{\boldsymbol{q}}(Y)^2 \quad \text{with} \quad Y := \hat{X}_{\mathcal{I}}^\theta \qquad (m \geq 2) \tag{96}$$

yields a $\Theta$-uniform approximation of the contrast $Q_m$ from (91).

**Lemma 9** (Interpolation Limit). *Let $\Theta$ and $X$ be as in Assumption 1, $\widehat{Q}_m$ as in (96) and $Q_m$ as in (91). Then for $(\mathcal{I}_n)_{n\in\mathbb{N}}$ any refined sequence of dissections of $\mathbb{I}$ and any $m \in \mathbb{N}_{\geq 2}$,*

$$Q_m(\theta) = \lim_{n\to\infty} \widehat{Q}_m(\hat{X}_{\mathcal{I}_n}^\theta) \quad \textit{uniformly on} \;\; \Theta. \tag{97}$$

*Proof.* See Appendix B.3. □

### 8.3. The Ergodicity Limit.

We formalise the estimation scheme (90) and show that it holds uniformly on $\Theta$ for a large class of time-series models and stochastic processes.

*Notation* 8.1. Let $Z := \mathbb{R}^d$. Given $z \equiv (z_j)_{j\in\mathbb{N}}$ and $J \subset \mathbb{N}$, we write $z_J := (z_j)_{j\in J}$ and $z_{(\ell_1:\ell_2]} := (z_{\ell_1+1}, z_{\ell_1+2}, \ldots, z_{\ell_2})$ for $\ell_1, \ell_2 \in \mathbb{N}_0$ with $\ell_1 < \ell_2$, and denote by $\hat{z}_{\mathcal{E}_n} \equiv \hat{\iota}_{\mathcal{E}_n}(z)$ the piecewise-linear interpolation of $z \equiv (z_j)_{j\in[n]} \in Z^{\times n}$ along the equidistant dissection $\mathcal{E}_n := \{(\nu-1)/(n-1) \mid \nu \in [n]\}$ of $[0,1]$ (cf. Appendix B.2). For $X_* \equiv (X_j)_{j\in\mathbb{N}}$ a discrete time-series in $\mathbb{R}^d$, we denote by $D_{X_*} := \overline{\bigcup_{j\in\mathbb{N}} \mathrm{supp}(X_j)}^{|\cdot|}$ its *spatial support.*

Set further $\mathfrak{sig}_{[m]} := \pi_{[m]} \circ \mathfrak{sig}$ for the signature capped at level $m \geq 2$, cf. Section C.2.2.

As the signature transform (187), and thereby its cumulants (78), (79), are invariant under time-domain reparametrisations of $X$ (Lemma C.1 (iii)), the statistics (88) of an interpolant $Y \equiv \hat{X}_{\mathcal{I}}$ depend only on the time series $X_{\mathcal{I}}$ — i.e. on the random variables

[37] For a formal definition of this operation see Appendix B.2, where a unified notation for the projection of continuous-time data to discrete time series—and, conversely, the embedding (via interpolation) of the latter type of data into $C(\mathbb{I}; \mathbb{R}^d)$—is provided.

$Z_1 := X_{t_0}, \ldots, Z_n := X_{t_{n-1}}$ and their *sequential order* — and not on the dissection along which $X_{\mathcal{I}} = (Z_j)_{j\in[n]}$ is interpolated. In symbols, see Appendix B.2 (152) for notation,

$$\mathfrak{sig}(\hat{X}_{\mathcal{I}}) = \mathfrak{sig}(\hat{\iota}_{\mathcal{J}}(Z_1, \ldots, Z_n)) \qquad \text{for any dissection } \mathcal{J} \tag{98}$$

with cardinality $|\mathcal{J}| = |\mathcal{I}|$. This justifies to in (84) abstract from the topology of the time-indices $t \in \mathcal{I}$, as done in the formulation of Definition 10 below.

All expectations in the following definition are assumed to exist.

*Definition* 10 (Signature Ergodicity). Let $X_* = (X_j)_{j\in\mathbb{N}}$ be a discrete time-series in $\mathbb{R}^d$ and $n, m \in \mathbb{N}$. We call $X_*$ $m^{\text{th}}$*-order signature ergodic to length* $n$ if, almost surely,

$$\mathbb{E}\big[\phi(X_{[n]})\big] = \lim_{T\to\infty} T^{-1} \sum_{j=1}^{T} \phi(X_{(n(j-1):\, nj]}) \quad \text{for} \quad \phi(z) := \mathfrak{sig}_{[m]}(\hat{z}_{\mathcal{E}_n}) \,, \tag{99}$$

and $X_*$ will be called *weakly* $m^{\text{th}}$-order signature ergodic to length $n$ if (99) holds in probability. We call the process $X_*$ [weakly] *signature ergodic* to length $n$ if $X$ is [weakly] $m^{\text{th}}$-order signature ergodic to length $n$ for every $m \geq 1$.

Given $\Theta \subseteq C(D_{X_*}; \mathbb{R}^d)$, we call $X_*$ [weakly/ $m^{\text{th}}$-order] signature ergodic to length $n$ *on* $\Theta$ if the respective property holds for each $\theta \cdot X_* := (\theta \cdot X_j)_{j\in\mathbb{N}}$, $\theta \in \Theta$.

We refer to the LHS of (99) as the $[m]^{\text{th}}$*-signature moment* of the batch $(X_1, \ldots, X_n)$.

*Remark* 8.2. In other words, the time-series $X = (X_j)_{j\in\mathbb{N}}$ is [weakly] $m^{\text{th}}$-order signature ergodic to length $n$ iff the sequence of empirical path-space measures (on $\mathcal{B}(\mathcal{C}_d)$)

$$\hat{\mu}_T := \frac{1}{T} \sum_{j=1}^{T} \delta_{\hat{X}_j} \qquad \text{for} \qquad \hat{X}_j := \hat{\iota}_{\mathcal{E}_n}(X_{n(j-1)+1}, \ldots, X_{nj})$$

yields a consistent estimator for the expected signature $\mathfrak{S}_{[m]}(\hat{X}_1)$ of $\hat{X}_1$, that is iff

$$\mathfrak{S}_k(\hat{X}_1) = \lim_{T\to\infty} \int_{\mathcal{C}_d} \mathfrak{sig}_k(x)\, \hat{\mu}_T(\mathrm{d}x) \quad \text{a.s. [in probab.]} \tag{100}$$

for each $1 \leq k \leq m$. Notice that due to (98), the equidistant dissection $\mathcal{E}_n$ in (99) may be replaced by any other $[0,1]$-dissection of the same cardinality.

Let as before the space $C(D_{X_*}; \mathbb{R}^d)$ be endowed with the compact-open topology.

Thanks to the 'universality' of the signature transform (Lemma C.1 (vii)), we find that the [weak] signature ergodicity of $X_* = (X_j)_{j\in\mathbb{N}}$ is passed onto $\theta \cdot X_* = (\theta \cdot X_j)_{j\in\mathbb{N}}$ for any $\theta \in C(D_{X_*}; \mathbb{R}^d)$.

**Proposition 6.** *Let $\Theta \subseteq C(D_{X_*}; \mathbb{R}^d)$ and $X_* = (X_j)_{j\in\mathbb{N}}$ be a discrete time-series in $\mathbb{R}^d$ with compact spatial support and such that for each $\theta \in \Theta$ the expectations*

$$\mathbb{E}[\mathfrak{sig}_m(\hat{X}_1^\theta)] \;\; \textit{exist for all } m \geq 1, \;\; \textit{with } \hat{X}_1^\theta \textit{ the interpolant of } \theta \cdot X_1, \ldots, \theta \cdot X_n.$$

*It then holds that: if $X_*$ is [weakly] signature ergodic to length $n$, then $X_*$ is [weakly] signature ergodic to length $n$ on $\Theta$.*

*Proof.* See Appendix B.4. □

Using a Glivenko-Cantelli type result yields the following observation of uniform convergence.

For the lemma below, let $\Theta$ be as in Assumption 1 and $\kappa_{\text{IC}}^{[m]}$ as in (88), and for $n \in \mathbb{N}$ denote by $(\mathcal{I}_{n|j})_{j\in\mathbb{N}}$ any fixed sequence of $[0,1]$-dissections with $|\mathcal{I}_{n|j}| = n$ for all $j \in \mathbb{N}$.

**Lemma 10** (Ergodicity Limit)**.** *Let $X_* = (X_j)_{j\in\mathbb{N}}$ be a discrete time-series which for some $m$ is [weakly] $m^{\text{th}}$-order signature ergodic to some length $n \in \mathbb{N}$ on $\Theta$, and denote*

$$\mathfrak{K}^{m|n|T}(\theta) := \log_{[m]}(\hat{\mathfrak{S}}^{m|n}_T(\theta)) \quad \textit{for} \quad \hat{\mathfrak{S}}^{m|n}_T(\theta) := \frac{1}{T}\sum_{j=1}^{T} \mathfrak{sig}_{[m]}(\hat{X}^\theta_j),$$

$$\textit{and}^{38} \quad \bar{\mathfrak{K}}^{m|n|T}_{\boldsymbol{i}}(\theta) := \frac{\mathfrak{K}^{m|n|T}_{\boldsymbol{i}}(\theta)}{\big(\mathfrak{K}^{m|n|T}_{11}(\theta)\big)^{\eta_1(\boldsymbol{i})/2}\cdot\ldots\cdot\big(\mathfrak{K}^{m|n|T}_{dd}(\theta)\big)^{\eta_d(\boldsymbol{i})/2}}, \quad \boldsymbol{i} \in [d]^\star,$$

*for each $\theta \in \Theta$, where $\hat{X}^\theta_j$ is the interpolant of $\theta\cdot X_{n(j-1)+1}, \ldots, \theta\cdot X_{nj}$ along $\mathcal{I}_{n|j}$. For any $m \geq 2, n, T \in \mathbb{N}$ and $\theta \in \Theta$, denote further (recalling Notation 7.1)*

$$\hat{\kappa}^{m|n}_T(\theta) := \sum_{\nu=2}^{m}\sum_{\boldsymbol{q}\in\mathfrak{C}_\nu} \bar{\mathfrak{K}}^{m|n|T}_{\boldsymbol{q}}(\theta)^2. \tag{101}$$

*Provided that $\mathbb{E}\big[\sup_{\theta\in\Theta}\big\|\mathfrak{sig}_k(\hat{X}^\theta_1)\big\|_k\big] < \infty$ for each $k \in [m]$, it then holds that*

$$\bar{\kappa}^{[m]}_{\text{IC}}(\hat{X}^\theta_1) = \lim_{T\to\infty}\hat{\kappa}^{m|n}_T(\theta) \quad \textit{uniformly on} \quad \Theta \quad \textit{a.s.} \quad [\textit{in probability}]. \tag{102}$$

*Proof.* See Appendix B.5. □

A detailed study of the class of (weakly) signature-ergodic stochastic processes is beyond the scope of this article, but the examples below show that the ergodicity assumption (100) is met for many popular time series models and stochastic processes.

We say that a time series $(X_j)_{j\in\mathbb{N}}$ has $(m,n)$-*stationary sigmoments* if the $[m]^{\text{th}}$-signature moments of the batches $(X_1,\ldots,X_n), (X_{n+1},\ldots,X_{2n}), \ldots$ exist and are equal, i.e. if for $\phi = \phi_m$ as in (99) we have: $\mathbb{E}[\phi_m(X_{[n]})] = \mathbb{E}[\phi_m(X_{(n(j-1):nj]})]$ for each $j \in \mathbb{N}$.

**Lemma 11.** *For $X_* \equiv (X_j)_{j\in\mathbb{N}}$ uniformly integrable and $n \in \mathbb{N}$, the following holds.*[39]

(i) *If $X_*$ is $\alpha$-mixing and has $(m,n)$-stationary sigmoments ($m \in \mathbb{N}$), then $X_*$ is $m^{\text{th}}$-order weakly signature-ergodic to length $n$;*
(ii) *if $X_*$ is $\phi$-mixing with $\sum_{\nu=1}^{\infty}\phi^{1/2}_{1+(\nu-1)n}(X_*)\frac{\log\nu}{\nu} < \infty$ and has $n$-seasonal increments, then $X_*$ is signature-ergodic to length $n$.*

*(i) and (ii) persist if $X_*$ is replaced by $\theta\cdot X_* = (\theta\cdot X_j)_{j\in\mathbb{N}}$ for any measurable $\theta : D_{X_*} \to \mathbb{R}^d$.*

*Proof.* See Appendix B.6. □

*Definition* 11 (Ergodic Observations)*.* For $X = (X_t)_{t\geq 0}$ a continuous stochastic process in $\mathbb{R}^d$ and $\mathcal{J} = (\mathcal{J}_k)_{k\in\mathbb{N}} \subset 2^{[0,\infty)}$ an exhaustive protocol with base lengths $(n_k)_{k\in\mathbb{N}}$,

(i) the pair $(X, \mathcal{J})$ will be called an *ergodic observation* if for almost all $k \in \mathbb{N}$,

$$(X_t)_{t\in\mathcal{J}_k} \quad \text{is} \quad \text{signature ergodic to length } n_k, \tag{103}$$

and $(X, \mathcal{J})$ will be called *ergodic** if in addition the spatial support of $X$ is compact;

(ii) the pair $(X, \mathcal{J})$ will be called a *weakly* ergodic observation if for almost all $k \in \mathbb{N}$,

$$(X_t)_{t\in\mathcal{J}_k} \quad \text{is} \quad \text{weakly signature ergodic to length } n_k \tag{104}$$

and the running maximum of $|X|$ has finite expectation,[40] i.e. $\mathbb{E}[\sup_{t\in[0,1]}|X_t|] < \infty$.

We say that $(X, \mathcal{J})$ is a [weakly] ergodic$^{[*]}$ *observation on* $\Theta \subseteq C(D_X; \mathbb{R}^d)$ if the respective property holds for each $(\theta\cdot X, \mathcal{J})$, $\theta \in \Theta$, individually.

[38] Where $\eta_\nu(\boldsymbol{i})$ denotes the number of times the index-entry $\nu$ appears in $\boldsymbol{i}$; cf. (79).
[39] The definitions of $\{\alpha,\phi\}$-mixing and '$n$-seasonal increments' are given in Appendix B.6.
[40] The integrability of the running maximum of $|X|$ is discussed in, e.g., [8, Chapter 13] and [54]

*Examples* 8.1. Lemma 11 implies that the (strong resp. weak) ergodicity assumption (99) is satisfied by a large number of time series and continuous stochastic processes $X$ (and with it $\theta \cdot X$, $\theta \in \Theta$), for the latter by way of (103) resp. (104) via any protocol $\mathcal{J}$ chosen such that $X_{\mathcal{J}}$ is appropriately stationary. These include, adequate stationarity provided,

1. (trivially) all $q$-dependent time series (e.g. all moving-average processes of finite degree);
2. various linear and related processes such as certain [MC]ARMA, ARCH and GARCH models (see, e.g., [31, 55, 65, 76]);
3. many Markov processes, diffusions and stochastic dynamical systems (e.g. [23, 56, 71, 83]);

see e.g. [9] for an overview. In practice however, infringements of the above (sufficient) conditions for signature ergodicity may typically be innocuous, cf. Section 9.

8.4. **The Consistency Limit.** The considerations of Subsections 8.1 to 8.3 combine to the following consistency result for our ICA-method (Theorem 4).

**Theorem 5** (Consistency)**.** *Let $X, S$ and $\Theta$ be as in Theorem 4 and Assumption 1, and suppose that there is a unique $\theta_\star \in \Theta$ such that $\theta_\star \cdot X$ is IC. Suppose further that $(X, (\mathcal{J}_k)_{k\in\mathbb{N}})$ is an ergodic* [resp. weakly ergodic] observation [on $\Theta$] with base lengths $(n_k)_{k\in\mathbb{N}}$. Then for any error bound $\varepsilon > 0$ there exists a capping threshold $m_0 \geq 2$ such that for any fixed $m \geq m_0$ the following holds: There is a mesh-index $k_0 = k_0(m) \in \mathbb{N}$ such that for any $k \geq k_0$ and any sequence $(\hat{\theta}^\star_T)$ in $\Theta$ with the property that, for $\hat{\kappa}^{m|n_k}_T$ as in (101),*

$$\hat{\kappa}^{m|n_k}_T(\hat{\theta}^\star_T \cdot X) \;\leq\; \min_{\theta\in\Theta} \hat{\kappa}^{m|n_k}_T(\theta \cdot X) \;+\; \eta_T \qquad (T \in \mathbb{N}) \tag{105}$$

*for some $(\eta_T) \subset \mathbb{R}_+$ with $\lim_{T\to\infty} \eta_T = 0$ almost surely [resp. in probability], it holds that*

$$\lim_{\tau\to\infty} \max\left\{ \sup_{T\geq\tau} \left[ \mathrm{dist}_{\|\cdot\|_\infty}\big(\hat{\theta}^\star_T \cdot X,\, \mathrm{DP}_d \cdot S\big)\right],\, \varepsilon \right\} = \varepsilon \tag{106}$$

*almost surely [resp. in probability]. If $(X, (\mathcal{J}_k)_{k\in\mathbb{N}})$ is ergodic on $\Theta$ and the spatial support of $X$ is not necessarily compact, then (106) holds almost surely with the above threshold $m_0$ depending on the realisation of $X$.*

With the maximum norm and the Euclidean norm on $\mathbb{R}^d$ being equivalent, Theorem 5 guarantees that in the infinite-data limit $T \to \infty$, the optimization (105) of the empirical signature contrasts $\hat{\kappa}^{m|n}_T$ from (101) allows us to (up to order and monotone scaling) recover each component of the source to an arbitrarily high $\mathbb{I}$-uniform precision given that the hyperparameters $m$ (capping threshold) and $k$ (observation frequency) are chosen large enough.[41]

*Proof of Theorem 5.* For the sake of a compact exposition, only the statement for $(X, (\mathcal{J}_k))$ ergodic* is proved here; a proof of the remaining non-compact [weakly] ergodic case is given in Appendix B.7. So let $(X, (\mathcal{J}_k))$ be an ergodic observation with $D_X$ compact.

Making the $(m, k, T)$-dependence of each minimizer $\hat{\theta}^\star_T$ in (105) explicit by writing $\hat{\theta}^\star_T =: \theta^{m|k}_T$, observe that assertion (106) follows from the claim that

$$\begin{gathered}\forall\, \tilde{\varepsilon} > 0 \,:\, \exists\, m_0 \geq 2 \,:\, \text{for each } m \geq m_0 \text{ there is } k_0 \equiv k_0(m) \text{ such that}: \\ \lim_{\tau\to\infty} \alpha^{m|k}_\tau \vee \tilde{\varepsilon} = \tilde{\varepsilon} \quad \text{a.s.} \quad \text{with} \quad \alpha^{m|k}_\tau := \sup\nolimits_{T\geq\tau} \tilde{d}(\theta^{m|k}_T, \theta_\star), \quad \text{for each } k \geq k_0;\end{gathered} \tag{107}$$

here, $a \vee b := \max\{a, b\}$ and $\tilde{d}$ denotes the topology-inducing metric (173) on $\Theta$.

(We import the setting and notation of Subsection B.7.1 for usage below.)

[41] Note that (86) is contained in Theorem 5 as the asymptotical special case $(m_0, k, T) = (\infty, k, \infty)$.

Indeed: Provided that (107) holds almost surely, we find that for every $\varepsilon > 0$ there is $m_0 \geq 2$ with the property that each capping index $m \geq m_0$ comes with a mesh-threshold $k_0 = k_0(m) \in \mathbb{N}$ such that for each observation at mesh-level $k \geq k_0$ we almost surely have

$$\|\theta_T^{m|k} \cdot X - \theta_\star \cdot X\|_\infty \;\leq\; \varepsilon \quad \text{ for almost all } T \in \mathbb{N}, \quad \text{where } \;\theta_\star \cdot X \in \mathrm{DP}_d \cdot S \tag{108}$$

with probability one (Theorem 4). To derive (108) from (107), let $\varepsilon > 0$ be arbitrary, assuming $\varepsilon < 1$ wlog, and note that $D_X \subseteq K_{\nu_0}$ for some $\nu_0 \in \mathbb{N}$ as $D_X$ is compact.

Observe that (174) provides the inclusion $B_\varepsilon^{d_{\nu_0}}(\theta_\star) \supseteq B_{\varepsilon/2}^{\rho_{\nu_0}}(\theta_\star)$ while the definition of $\tilde{d}$ yields $\tilde{B}_{\tilde{\varepsilon}}(\theta_\star) \subseteq B_{\varepsilon/2}^{\rho_{\nu_0}}(\theta_\star)$ for $\tilde{\varepsilon} := 2^{-\nu_0}\varepsilon/2$, cf. (173). Given (107), this $\tilde{\varepsilon}$ comes with associated $m_0, m, k_0, k \in \mathbb{N}$ and a $\mathbb{P}$-full set $\Omega' \equiv \Omega'_{m,k} \in$ [42] $\mathscr{F}$ such that

$$\alpha^{m|k}(\omega) := \lim_{\tau\to\infty} \sup\nolimits_{T\geq\tau} \tilde{d}(\theta_T^{m|k}(\omega), \theta_\star) \;\leq\; \tilde{\varepsilon}/2 \qquad \text{for each } \;\omega \in \Omega'$$

(note: $\alpha^{m|k}$ exists as $(\alpha_\tau^{m|k})_{\tau\in\mathbb{N}}$ is monotone and bounded). Thus, for each $\omega \in \Omega' \cap \Omega''$ (for $\Omega'' \in \mathscr{F}$ for the $\mathbb{P}$-full set on which the traces of $X$ are all contained in $D_X$; Lemma 1 (ii)) there is $\tau_0 \,(= \tau_0(\omega)) \in \mathbb{N}$ with $(\theta_T^{m|k}(\omega))_{T\geq\tau_0} \subset \tilde{B}_{\tilde{\varepsilon}}(\theta_\star) \subseteq B_{\varepsilon/2}^{\rho_{\nu_0}}(\theta_\star) \subseteq B_\varepsilon^{d_{\nu_0}}(\theta_\star)$ and hence

$$\left\|\theta_T^{m|k}(\omega) \cdot X(\omega) - \theta_\star \cdot X(\omega)\right\|_\infty \leq \left\|\theta_T^{m|k}(\omega) - \theta_\star\right\|_{K_{\nu_0}} = d_{\nu_0}(\theta_T^{m|k}(\omega), \theta_\star) \;\leq\; \varepsilon \tag{109}$$

for each $T \geq \tau_0$, which gives (108) as desired.

Let us now prove (107), for which we let $\tilde{\varepsilon} > 0$ be arbitrary.

Then for $\kappa_{m,k}(\theta) := \widehat{Q}_m(\hat{X}^\theta_{\mathcal{I}_1^{(k)}})$ as in (96) with $\mathscr{J} =: (\mathcal{J}_k \equiv \sqcup_{\nu=1}^\infty \mathcal{I}_\nu^{(k)} \mid k \in \mathbb{N})$ the given protocol under consideration, there is $m_0 \geq 2$ such that for each $m \geq m_0$ it holds that

$$\exists\, k_0 \,(\equiv k_0(m)) \in \mathbb{N} \;\;:\;\; \operatorname*{arg\,min}_{\theta\in\Theta} \kappa_{m,k}(\theta) \;\subseteq\; \tilde{B}_{\tilde{\varepsilon}/2}(\theta_\star) \quad \text{ for each } \;k \geq k_0. \tag{110}$$

Indeed: The uniqueness of $\theta_\star$ implies that $\zeta_{\tilde{\varepsilon}} := \min_{\theta\in\Theta\setminus C_{\tilde{\varepsilon}}} Q(\theta) > 0$ for[43] $C_{\tilde{\varepsilon}} := \tilde{B}_{\tilde{\varepsilon}/2}(\theta_\star) \cap \Theta$ and $Q$ as in (91), while Lemma 8 (ii) provides an $m_0 \geq 2$ with $\sup_{m\geq m_0} \|Q - Q_m\|_\Theta < \zeta_{\tilde{\varepsilon}}/2$. Fixing any $m \geq m_0$ and using that $(\mathcal{J}_k)_{k\in\mathbb{N}}$ is exhaustive (whence the sequence $(\mathcal{I}_1^{(k)})_{k\in\mathbb{N}}$ is refined), Lemma 9 yields some $k_0 \equiv k_0(m) \in \mathbb{N}$ with $\sup_{k\geq k_0} \|Q_m - \kappa_{m,k}\|_\Theta < \zeta_{\tilde{\varepsilon}}/2$. Hence for any fixed $k \geq k_0$ and each $\tilde{\theta} \in \Theta$ with $\kappa_{m,k}(\tilde{\theta}) = \min_{\theta\in\Theta} \kappa_{m,k}(\theta)$, it holds

$$Q(\tilde{\theta}) = \kappa_{m,k}(\tilde{\theta}) + (Q - \kappa_{m,k})(\tilde{\theta}) \;<\; \zeta_{\tilde{\varepsilon}}/2 + \zeta_{\tilde{\varepsilon}}/2 = \zeta_{\tilde{\varepsilon}} \quad \text{and hence} \quad \tilde{\theta} \in \tilde{B}_{\tilde{\varepsilon}/2}(\theta_\star),$$

where we used (the triangle inequality and) the fact that $0 \leq \min_{\theta\in\Theta} \kappa_{m,k}(\theta) \leq \kappa_{m,k}(\theta_\star) = 0$ (i.e., the argmins of $\kappa_{m,k}$ on $\Theta$ coincide with its roots), where this last property follows from Proposition 4 and the obvious fact that if $\theta_\star \cdot X$ is IC then so is $\hat{X}^{\theta_\star}_{\mathcal{I}_1^{(k)}}$.

Together with (110) the identity $\arg\min_\Theta(\kappa_{m,k}) = \{\theta \in \Theta \mid \kappa_{m,k}(\theta) = 0\} =: \mathcal{N}(\kappa_{m,k})$ now implies that, for $\bar{\kappa}_{m,k}(\theta) := \bar{\kappa}^{[m]}_{\mathrm{IC}}\big(\hat{X}^\theta_{\mathcal{I}_1^{(k)}}\big)$ as in (88),

$$\mathcal{M} := \operatorname*{arg\,min}_{\theta\in\Theta} \bar{\kappa}_{m,k}(\theta) \;\subseteq\; \tilde{B}_{\tilde{\varepsilon}/2}(\theta_\star) \quad \text{ for each } \;k \geq k_0, \tag{111}$$

because $\arg\min_\Theta \bar{\kappa}_{m,k} = \mathcal{N}(\bar{\kappa}_{m,k})$ (as above) and the zero sets $\mathcal{N}(\bar{\kappa}_{m,k})$ and $\mathcal{N}(\kappa_{m,k})$ coincide (by Definition 9). Next we claim that, for $m$ and $k$ as above,

$$\lim_{\tau\to\infty} \sup\nolimits_{T\geq\tau} \operatorname{dist}\big(\theta_T^{m|k}, \mathcal{M}\big) \;=\; 0 \qquad \text{almost surely,} \tag{112}$$

---

[42] For convenience, we may assume the underlying probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to be complete.

[43] Provided that $\tilde{\varepsilon}$ is small enough such that $C_{\tilde{\varepsilon}} \subsetneq \Theta$, which can be assumed without loss of generality.

which by way of (111) implies (107) as desired. To see (112), observe first that

$$\lim_{T\to\infty} \bar{\kappa}_{m,k}(\theta^\star_T) = 0 \quad \text{almost surely,} \quad \text{with} \;\; (\theta^*_T) \equiv (\theta^{m|k}_T) \tag{113}$$

as in (112), which due to (102) follows by the same arguments that led to (148). Invoking a proof by contradiction, assume now that (112) does not hold. Then, pointwise on an event of positive probability, there will be $\delta_0 > 0$ together with a subsequence $(T_j)_{j\in\mathbb{N}} \subseteq \mathbb{N}$ such that $\operatorname{dist}(\theta^\star_{T_j}, \mathcal{M}) \geq \delta_0$ for each $j \in \mathbb{N}$. But as $\Theta$ is compact, we (upon passing to a convergent subsequence) may assume that $(\theta^\star_{T_j})_{j\in\mathbb{N}}$ converges to some $\theta_0 \in \Theta$. Then by continuity $\bar{\kappa}_{m,k}(\theta_0) = \lim_{j\to\infty} \bar{\kappa}_{m,k}(\theta^\star_{T_j})$, whence $\bar{\kappa}_{m,k}(\theta_0) = 0$ by (113) and thus $\theta_0 \in \mathcal{M}$. The latter is a contradiction, however, as $(\theta^\star_{T_j})_{j\in\mathbb{N}}$ is bounded away from $\mathcal{M}$ (by $\delta_0$), proving (112). □

## 9. Numerical Experiments

We conclude with a series of numerical experiments. A complete description of the following experiments and results, including their full parameter settings and all relevant implementations, is provided on the GitHub repository [75].

### 9.1. A Performance Index for Nonlinear ICA.

In order to assess how close an estimated source $\hat{S}$ is to the true source $S$ of (13), we propose to quantify the distance between $\hat{S}$ and the orbit $\mathrm{DP}_d \cdot S$ (cf. (20)), whose elements we recall to be in a minimal distance from $S$, by way of the following performance statistic (whose applicability is due to Remark 7.2 (iii)).

*Definition* 12 (Monomial Discordance). Given two time series $\mathcal{X} := (X^1_t, \cdots, X^d_t)_{t\in\mathcal{I}}$ and $\mathcal{Y} := (Y^1_t, \cdots, Y^d_t)_{t\in\mathcal{I}}$ in $\mathbb{R}^d$ for $\mathcal{I}$ finite, define the *concordance matrix of* $(\mathcal{X}, \mathcal{Y})$ as

$$\mathcal{C}(\mathcal{X},\mathcal{Y}) := \left(\frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} |\rho_{\mathrm{K}}(X^i_t, Y^j_t)|\right)_{(i,j)\in[d]^2} \in [0,1]^{d\times d}$$

where $\rho_{\mathrm{K}}$ is the Kendall rank correlation coefficient.[44] Furthermore, we define

$$\varrho(\mathcal{X},\mathcal{Y}) := \frac{1}{\sqrt{d(d-1)}} \min_{P\in\mathrm{P}_d} \|C(\mathcal{X},\mathcal{Y}) - P\|_2 \in [0,1] \tag{114}$$

and call this quantity the *monomial discordance* of $\mathcal{X}$ and $\mathcal{Y}$.

**Proposition 7.** *Let $X$ and $S$ be as in (13) with $S$ IC, and $h$ be $C^1$-invertible on some open superset of $D_X$. Then for $\mathcal{I} \subset \mathbb{I}$ finite and $\varrho$ as in (114), we have that:*

$$(h\cdot X_t)_{t\in\mathcal{I}} \in \mathrm{DP}_d \cdot (S_t)_{t\in\mathcal{I}} \quad \textit{iff} \quad \varrho\big((h\cdot X_t)_{t\in\mathcal{I}}, (S_t)_{t\in\mathcal{I}}\big) = 0. \tag{115}$$

*Proof.* See Appendix A.11. □

Hence the smaller the monomial discordance between $S$ and a transformation $h(X)$ of its observable, the closer to optimal will be the deviation between $h(X)$ and $S$.

Below we provide a brief synopsis of our experiments and the results that we obtained. For brevity, the approximation of the contrast function (81) (Remark 7.2 (iv)) will be denoted $\phi_{\hat{\mu}}$.

### 9.2. Nonlinear Mixings With Explicitly Parametrized Inverses.

First we consider three families of $C^3$-diffeomorphisms on the plane whose inverses are explicitly parametrized.

More specifically: We sample two types of IC processes in $\mathbb{R}^2$, namely an Ornstein-Uhlenbeck process $S_{\mathrm{ou}} = (S^1_{\mathrm{ou}}, S^2_{\mathrm{ou}})$ and a copula-based time-series $S_{\mathrm{cy}} = (S^1_{\mathrm{cy}}, S^2_{\mathrm{cy}})$ following the dependence model (68).[45] Both $S_{\mathrm{ou}}$ and $S_{\mathrm{cy}}$ are contrastive (cf. Prop. 2 (ii) and Cor. 2 (i)), and

[44] If preferred, $\rho_{\mathrm{K}}$ might alternatively be chosen as Spearman's rank correlation coefficient.

[45] With $F^{S^i_{\mathrm{cy}}}_t$ chosen as the cdf of $\mathcal{N}(0,1)$ and $c$ chosen as the Clayton-density (cf. Proposition 1 (i)).

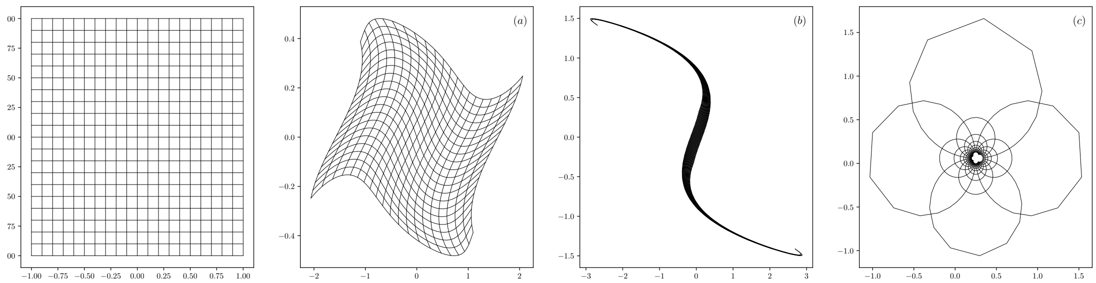


FIGURE 1. *The image of the spatial support $D_S = [-1,1]^2$ (leftmost) under three increasingly nonlinear mixing transformations $f_1, f_2, f_3$, namely conjugates of the Hénon map ($f_1$ and $f_2$) and of the Möbius transformation ($f_3$). The images $f_1(D_S)$ and $f_2(D_S)$ are shown in panels (a) and (b) respectively, and $f_3(D_S)$ is shown in panel (c).*

centering and scaling them to unit amplitude gives $D_{S_{\mathrm{ou}}} = D_{S_{\mathrm{cy}}} = [-1,1]^2$. Figure 2 shows the spatial trace of a sample of $S_{\mathrm{ou}}$ and $S_{\mathrm{cy}}$, next to an excerpt of its components. The sources $S_{\mathrm{ou}}$ and $S_{\mathrm{cy}}$ are then transformed by one of three mixing maps $f_j : \mathbb{R}^2 \to \mathbb{R}^2$ with increasing degree of 'nonlinearity', see Figure 1. (For a definition of the $f_j$, see [75]). The mixtures are denoted $X_\eta^{(j)} := f_j(S_\eta)$ for $j = 1,2,3$ and $\eta \in \{\text{'ou'}, \text{'cy'}\}$, see Figure 2.

Using that each 'demixing transform' $g^j := f_j^{-1}$ is contained in an injectively parametrized family $\Theta_j \equiv \{g_\theta^j \in C^3(\mathbb{R}^2) \mid \theta \in \tilde{\Theta}_j\}$ of candidate de-mixing transformations for some $\tilde{\Theta}_j \subseteq \mathbb{R}^2$ open, we consider the data-based objective functions

$$\Phi_\eta^j \,:\, \tilde{\Theta}_j \to \mathbb{R}, \qquad \theta \;\mapsto\; \phi_{\hat{\mu}_j}(g_\theta^j(X_\eta^{(j)})) \tag{116}$$

with $\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_3 = 6$, and compare their topography to that of the monotone discordances

$$\delta_\eta^j \,:\, \tilde{\Theta}_j \to \mathbb{R}, \qquad \theta \;\mapsto\; \varrho(g_\theta^j(X_\eta^{(j)}), S_\eta) \tag{117}$$

(cf. (114)); recall that the latter are 'distance functions' that quantify how much an estimated source $\hat{S}_\eta^\theta := g_\theta^j(X_\eta^{(j)})$ deviates from the monomial orbit $\mathrm{DP}_d \cdot S_\eta$ of $S_\eta$.

The results, which for $\eta = \mathrm{ou}$ are displayed in the first three columns of Figure 3 (with $\Phi_{\mathrm{ou}}^{1|2|3}$ shown in the first-row panels and the $\delta_{\mathrm{ou}}^{1|2|3}$ shown in the second-row panels),[46] show clearly that within the given families of candidate transformations $\Theta_j$, those candidate nonlinearities which map the data $X_\eta^{(j)}$ to a best-approximation of its source $S_\eta$ are precisely those that minimise the contrast (116), as asserted by Theorem 4.

An analogous experiment ($j = 4$) is performed for a mixing transformation $f_4 : \mathbb{R}^3 \to \mathbb{R}^3$, see Figures 4 and 5, and the results, obtained for the maximal index-length $\hat{\mu}_4 = 7$, are again affirmative of Theorem 4 (see the rightmost column of Figure 3).

9.3. **Nonlinear Mixings With Inverses Approximated By Neural Networks.** The practical applicability of our ICA-method is illustrated by running the optimisation (82) over (approximate) demixing-transformations which are modelled by an artificial neural network.

More specifically: We subject two Ornstein-Uhlenbeck sources $S^{(1)}$ and $S^{(2)}$ with two resp. four independent components to a two- resp. four-dimensional nonlinear mixing transform

[46] The results for the case $\eta = \mathrm{cy}$ can be found in [75].

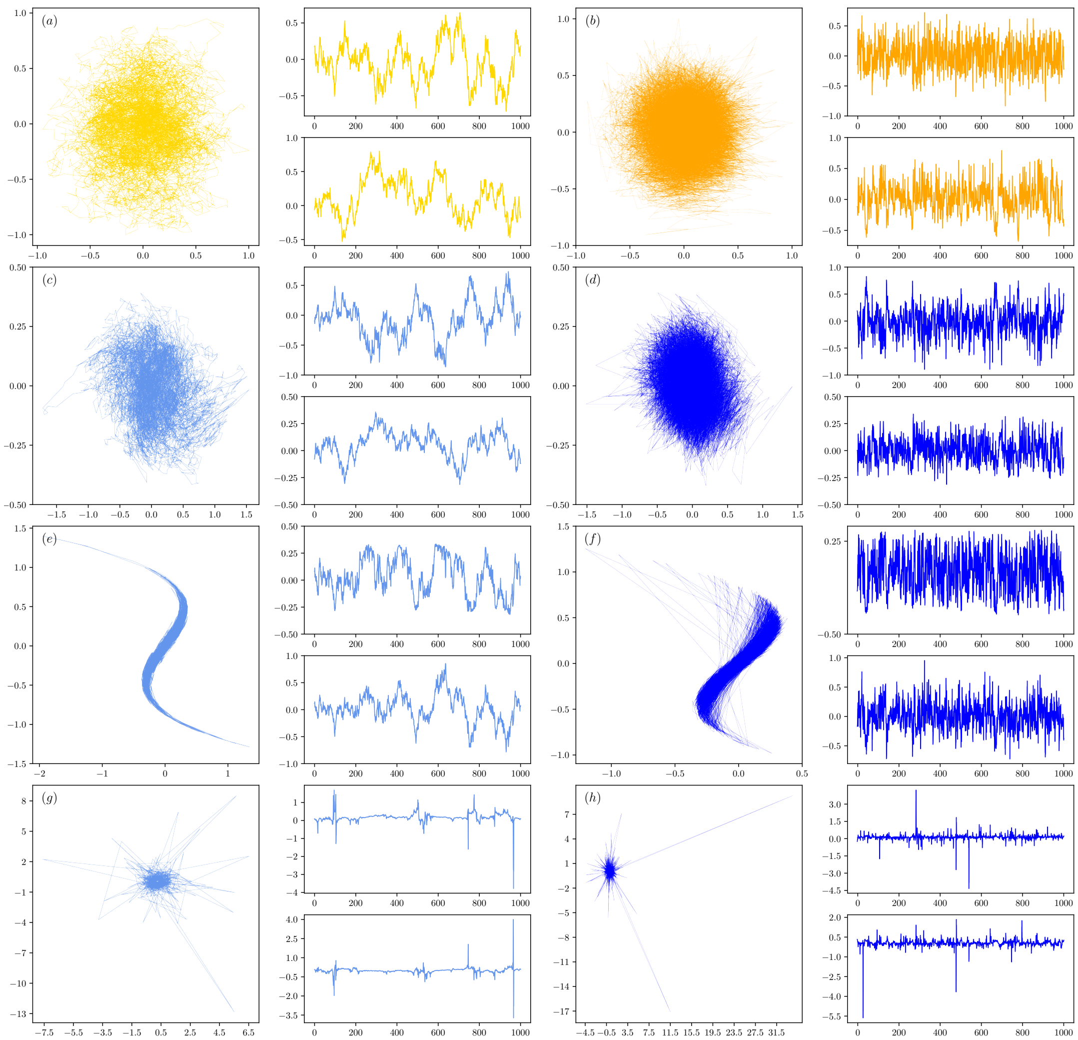

FIGURE 2. *Spatial traces and sampled components of three nonlinear mixtures of the sources $S_{\mathrm{ou}}$ (panel (a)) and $S_{\mathrm{cy}}$ (panel (b)). Depicted are the mixtures $X^{(1)}_{\mathrm{ou}}$ and $X^{(1)}_{\mathrm{cy}}$ ((c) and (d)), $X^{(2)}_{\mathrm{ou}}$ and $X^{(2)}_{\mathrm{cy}}$ ((e) and (f)), and $X^{(3)}_{\mathrm{ou}}$ and $X^{(3)}_{\mathrm{cy}}$ ((g) and (h)). The sampled components of each mixture, excerpted over 1000 data points each, are shown to the right of each panel.*

(see [75] for details). The resulting mixtures $X^{(1)}$ and $X^{(2)}$ are then passed to candidate demixing-nonlinearities $g^\nu_\theta \in \Theta_\nu$ ($\nu = 1, 2$) given as elements of the parametrized families

$$\Theta_\nu := \left\{ g^\nu_\theta : \mathbb{R}^{2\nu} \to \mathbb{R}^{2\nu} \;\middle|\; g^\nu_\theta \;\text{ is an ANN with weights }\; \theta \in \tilde{\Theta}_\nu \right\} \tag{118}$$

of transformations that are spanned by the configurations of some artificial neural network (ANN) instantiated over weight-vectors $\theta$ that are chosen from a given parameter set $\tilde{\Theta}_\nu$ in $\mathbb{R}^{m_\nu}$, where the number of weights $m_\nu$ is part of the pre-defined architecture of the ANN.

Given these candidate-inverses, the optimisations (82) are run by

$$\text{minimizing} \qquad \tilde{\Theta}_\nu \;\ni\; \theta \quad\longmapsto\quad \phi_{\hat{\mu}_\nu}(g^\nu_\theta \cdot X^{(\nu)}), \tag{119}$$

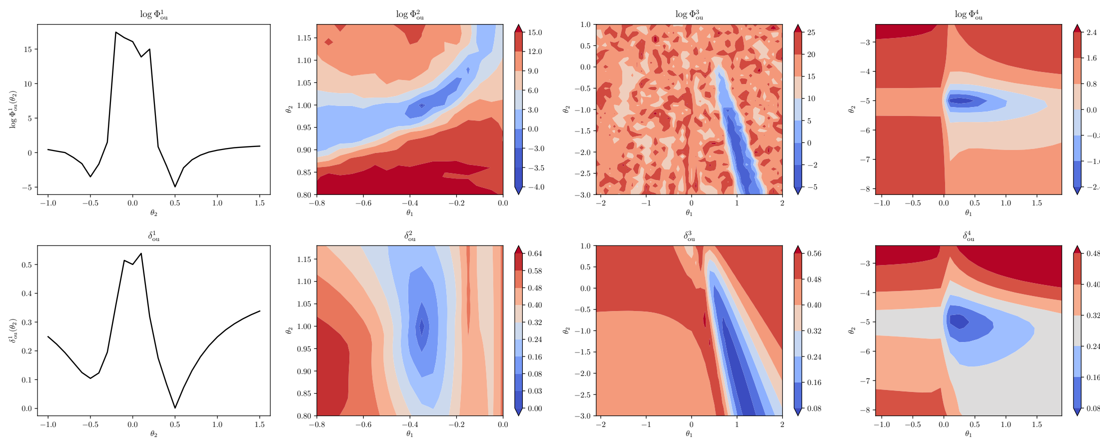

FIGURE 3. *Contour plot (leftmost column) and heatmaps of the log-transformed contrast functions ((116), first row) and the associated discordance functions ((117), second row) for the above mixings $X^{(j)}_{\mathrm{ou}} = f_j(S_{\mathrm{ou}})$ with $j = 1, \ldots, 4$. The parameters $(\theta_1^{(j)}, \theta_2^{(j)})$ of the true inverses $f_j^{-1} \in \Theta_j$ are $\theta_2^{(1)} = 0.5$, $(\theta_1^{(2)}, \theta_2^{(2)}) = (-0.35, 1)$, $(\theta_1^{(3)}, \theta_2^{(3)}) = (1, -2)$, and $(\theta_1^{(4)}, \theta_2^{(4)}) = (0.2, -5)$. Notice that: (a) by definition of $\Theta_1$, the function $\Phi^1_{\mathrm{ou}}$ depends on the one-dimensional parameter $\theta_2$ only; (b) as the concordance matrix of $\hat{S} := g^{(1)}_{-0.5}(X^{(1)}_{\mathrm{ou}})$ and $S_{\mathrm{ou}}$ is $\left(\begin{smallmatrix} 0.053 & 0.929 \\ 0.834 & 0.099 \end{smallmatrix}\right)$ (indicating a close proximity between $\hat{S}$ and $\mathrm{DP}_d \cdot S_{\mathrm{ou}}$, cf. Prop. 7), the observation of $\Phi^1_{\mathrm{ou}}$ attaining a low local minimum at $-0.5$ is in accordance with Theorem 4.*

i.e. by training each ANN (118) with the truncated contrast $\phi_{\hat{\mu}_\nu}$ as its loss function, where the optimization steps are done via backpropagation along the weights of the ANN (118).

For the case $\nu = 1$, we applied the mixing transformation depicted in Figure 6 (leftmost panel), and as the parametrising ANN $\Theta_1$ we chose a feedforward neural network with a two-nodal in- and output layer and two hidden layers consisting of 4 resp. 32 neurons with tanh activation each; the cumulant series (81) was capped at maximal index-length $\hat{\mu}_1 = 6$. For the case $\nu = 2$, we followed the simulations of [43, 44] in as a mixing transformation using an invertible feedforward-neural network with four-nodal in- and output layers and two four-nodal hidden layers with tanh activation each, and as the parametrising ANN chose a feedforward network with a four-nodal in- and output layer and one hidden layer of 1024 neurons and uniformly-weighted Leaky ReLU activations; the contrast function (81) was capped at the maximal index-length $\hat{\mu}_2 = 5$. For both $\nu = 1, 2$, the resulting loss functions (119) were optimised using stochastic gradient descent (Adam) with non-vanishing $\ell_2$-penalty.

Denoting by $\theta^*_\nu \in \tilde{\Theta}_\nu$ the (local) optimum obtained by the minimisation of the objective (119), and setting $\hat{S}^{(\nu)} := g^{\nu}_{\theta^*_\nu}(X^{(\nu)})$ for the associated estimate of the source $S^{(\nu)}$ (cf. (82)), we as a

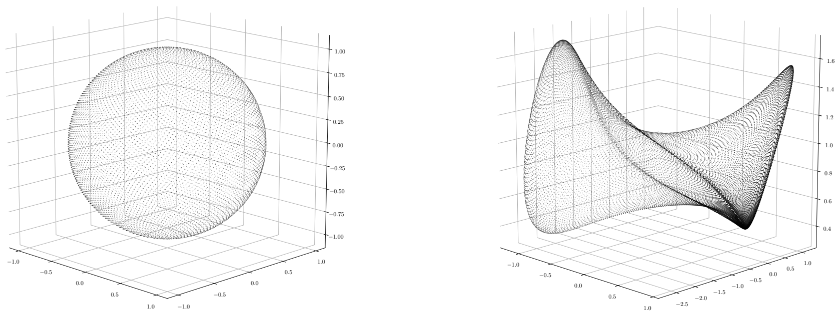

FIGURE 4. *Illustration of the three-dimensional mixing transform $f_4 : \mathbb{R}^3 \to \mathbb{R}^3$ via its action $f_4(S^2)$ (right panel) on the 2-sphere $S^2 \equiv \{x \in \mathbb{R}^3 \mid |x| = 1\}$ (left panel).*

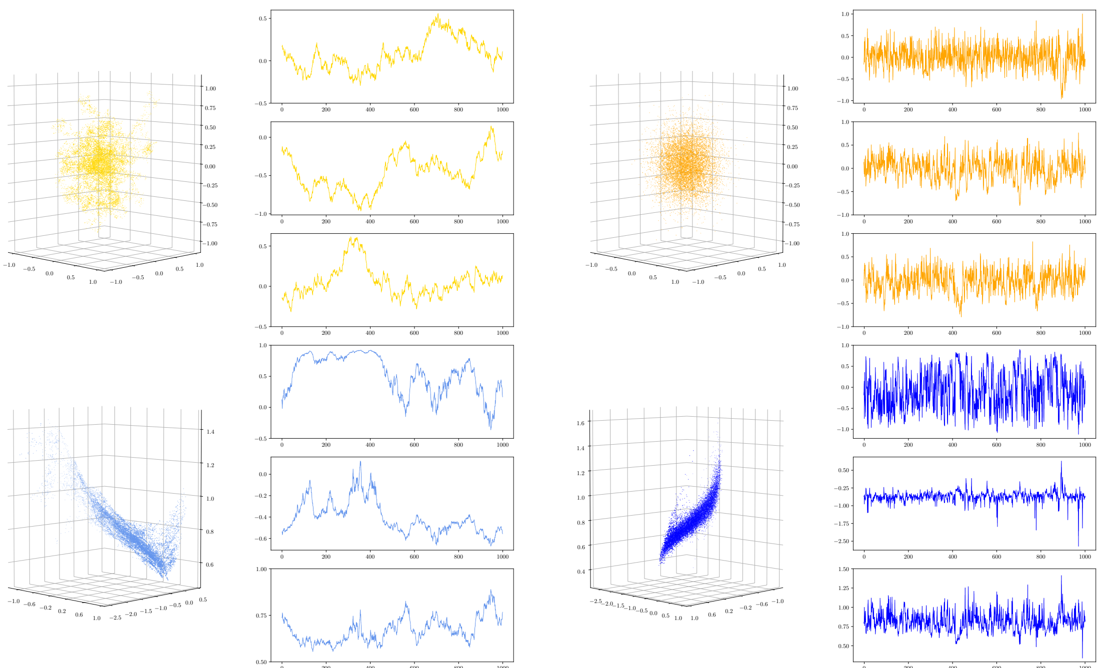

FIGURE 5. *Spatial trace and sampled components of a three-dimensional IC Ornstein-Uhlenbeck $\tilde{S}_{\mathrm{ou}}$ process (top left) and an IC copula-based time-series model $\tilde{S}_{\mathrm{cy}}$ (top right) and their respective nonlinear mixtures $f_4(\tilde{S}_{\mathrm{ou}})$ (bottom left) and $f_4(\tilde{S}_{\mathrm{cy}})$ (bottom right).*

result of these experiments obtained the concordance matrices (cf. Definition 12)

$$\mathcal{C}(\hat{S}^{(1)}, S^{(1)}) \doteq \begin{pmatrix} 0.079 & \mathbf{0.930} \\ \mathbf{0.853} & 0.065 \end{pmatrix} \quad \text{and} \tag{120}$$

$$\mathcal{C}(\hat{S}^{(2)}, S^{(2)}) \doteq \begin{pmatrix} 0.148 & \mathbf{0.725} & 0.109 & 0.069 \\ 0.037 & 0.034 & \mathbf{0.803} & 0.265 \\ \mathbf{0.834} & 0.003 & 0.037 & 0.016 \\ 0.077 & 0.131 & 0.072 & \mathbf{0.787} \end{pmatrix}. \tag{121}$$

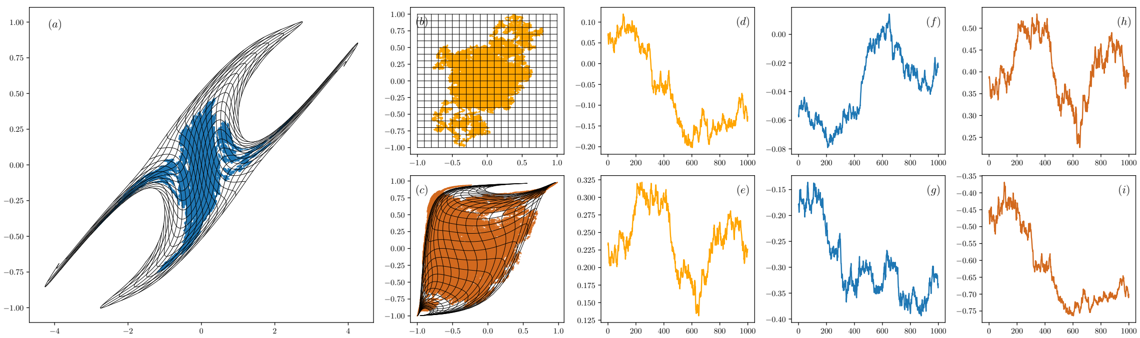

FIGURE 6. *Nonlinear mixture $X$ (with sample trace ((a), blue) and components (f) and (g)) of an IC Ornstein-Uhlenbeck process $S$ (shown in ((b), orange), with components (d) and (e)). Further shown is the residual $g \circ f\big|_{[-1,1]^2}$ for an approximation $g$ of $f^{-1}\big|_{D_X}$ ((c); cf. (122)) which is obtained by optimising an artificial neural network $(g_\theta)$ via the loss function (119). Trace and time-plots of the components of the resulting estimate $\hat{S} := g(X)$ of $S$ are shown in (c), (e) and (f) respectively. Notice that to a good approximation, $\hat{S}$ and $S$ coincide up to a transposition and a monotone scaling of their components, as quantified by (120).*

Both (120) and (121) indicate a *good fit between $\hat{S}^{(\nu)}$ and $S^{(\nu)}$* in the sense that, to a good approximation, $\hat{S}^{(\nu)}$ and $S^{(\nu)}$ differ only up to an inevitable permutation and monotone scaling of their components, as described in Theorem 4. (Recall that an optimal deviation $\hat{S}^{(\nu)} \in \mathrm{DP}_d \cdot S^{(\nu)}$ between $\hat{S}^{(\nu)}$ and $S^{(\nu)}$ is achieved iff (120) and (121) are permutation matrices (Proposition 7).) A visual comparison of the original samples $S^{(1)}, S^{(2)}$ and their estimates $\hat{S}^{(1)}, \hat{S}^{(2)}$, see Figures 6 and 7, confirms these results. To reaffirm that our findings are not due to chance, we ran our experiments repeatedly with randomly chosen realisations and initial configurations of the data and the ANN learning process, cf. [75] and Figure 8.

These experiments underline the practical applicability of our proposed ICA-method. In addition, we note the following experimental findings:

(i) Given an observable $X = f(S)$ together with a family $\Theta$ of candidate transformations on $\mathbb{R}^d$, the technical compatibility condition $\big(\mathrm{DP}_d(D_S) \cdot f^{-1}\big)\big|_{D_X} \cap \Theta\big|_{D_X} \neq \emptyset$ of Theorem 4 can in practice typically not be guaranteed a priori. However, as indicated by the above findings (120) and (121), infringements of this (sufficient) technical condition might typically be innocuous in practice, provided that at least

$$\big(\mathrm{DP}_d \cdot g\big)\big|_{D_X} \cap \Theta\big|_{D_X} \neq \emptyset \quad \text{for some } g \text{ with } g\big|_{D_X} \text{ `\textit{close enough}' to } f^{-1}\big|_{D_X}, \tag{122}$$

which will be satisfied if $\Theta$ is chosen large enough, say as a suitable ANN or another universal approximator.[47]

(ii) We emphasize that the configurations of the neural networks and their backpropagation that we used in our experiments were ad hoc and not tuned for approximational optimality. Since the loss functions (119) are typically non-convex with their topography crucially depending on the choice of (118) (cf. Figure 3), we expect that the accuracy and efficiency of our estimates may be *significantly* improved by applying

[47] In a similar vein, our experiments indicate that the regularity condition $\Theta \subseteq C^3(D_X)$ may in practice be softened by merely requiring that the 'approximate inverse' $g$ in (122) be '$C^3$-invertible on most of $D_X$' (cf. e.g. Figure 6, panel (c)) and the parametrization of $\Theta$ be 'continuous' at (some) point $\tilde{g} \in \Theta$ with $\tilde{g}\big|_{D_X} \in \mathrm{DP}_d \cdot g\big|_{D_X}$, though this a priori reduces the optimisation (82) to the search for a (low) *local* minimum.

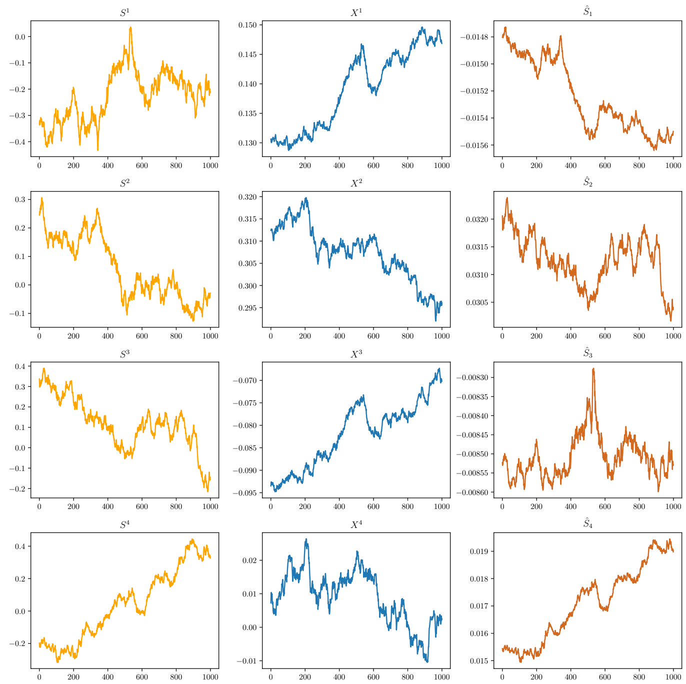

FIGURE 7. *Coordinate processes (over 1000 data points each) of an IC Ornstein-Uhlenbeck process* $S = (S^1, S^2, S^3, S^4)$ *(orange), a nonlinear mixture* $X = (X^1, X^2, X^3, X^4)$ *(blue) of* $S$*, and an estimate* $\hat{S} = (\hat{S}^1, \hat{S}^2, \hat{S}^3, \hat{S}^4)$ *of* $S$ *(brown). The estimate* $\hat{S}$ *is obtained as* $\hat{S} = g^{(2)}_{\theta_*}(X)$ *where* $(g^{(2)}_\theta)$ *is an ANN and* $\theta_*$ *is a (local) minimum of the associated contrast function* (119)*. To a good approximation, the processes* $\hat{S}$ *and* $S$ *coincide up to a permutation and a monotone scaling of their components, in accordance with Theorem 4 and as quantified by* (121)*.*

our ICA-method to ANN-based approximation schemes (118), (119) which are more carefully designed.

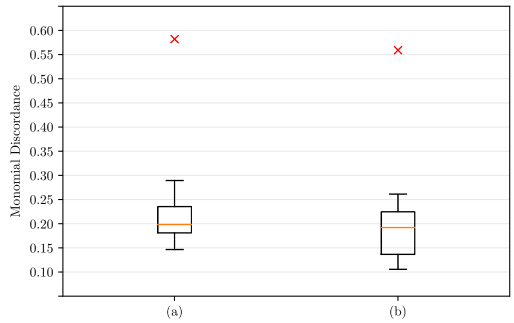


FIGURE 8. *Boxplots of the monomial discordance (114) between the true source $S$ and its (119)-based approximations $\hat{S}$ from the nonlinear mixture $X$ introduced in Figure 6, for $S$ an IC Ornstein-Uhlenbeck source (a) and an IC copula-based time series (b), respectively; shown in addition (red crosses) are the average deviations between $S$ and $\hat{S}$ at the start of the learning process. The discordances for each boxplot are estimated from ten realisations of $S$ and ten initial configurations $g_{\theta_0}$ of the demixing ANNs (118), with any of these realisations chosen independently and uniformly at random. The red crosses, at heights 0.59 and 0.56 respectively, indicate the average monomial discordance between $S$ and the initial guesses $g_{\theta_0}(X)$ at the start of the iterative learning procedure (119). The obtained discordances have mean 0.21 and standard deviation $45 \cdot 10^{-3}$ for (a), and mean 0.18 and st. dev. $55 \cdot 10^{-3}$ for (b). The average concordance matrix of $(S, \hat{S})$ is $\left(\begin{smallmatrix} 0.78 & 0.10 \\ 0.11 & 0.89 \end{smallmatrix}\right)$ for (a), and $\left(\begin{smallmatrix} 0.83 & 0.11 \\ 0.11 & 0.88 \end{smallmatrix}\right)$ for (b). (All numbers are rounded to two significant digits.)*

## APPENDIX A. TECHNICAL PROOFS AND REMARKS

A.1. **The Path-Space $\mathcal{C}_d$ is Cartesian.** Let us remark that the space $\mathcal{C}_d$ of paths from $\mathbb{I}$ to $\mathbb{R}^d$ (for $\mathbb{I}$ some fixed compact interval) inherits the Cartesian structure of $\mathbb{R}^d = \mathbb{R} \times \cdots \times \mathbb{R}$ via the canonical Banach-isometry $\psi : \mathcal{C}_1 \times \cdots \times \mathcal{C}_1 \to \mathcal{C}_d$ which is given by

$$\mathcal{C}_1^{\times d} \ \ni \ \left((\gamma_t^1)_{t\in\mathbb{I}}, \cdots, (\gamma_t^d)_{t\in\mathbb{I}}\right) \quad \longmapsto \quad \left[\mathbb{I} \ni t \ \mapsto \ \sum_{i=1}^{d} \gamma_t^i \cdot e_i\right] \ \in \ \mathcal{C}_d \tag{123}$$

for $(e_i)_{i\in[d]}$ the standard basis of $\mathbb{R}^d$. Indeed: It is clear that $\psi$ is a linear isometry with respect to the Banach norms $\|(\gamma^1, \cdots, \gamma^d)\|_\alpha := \max_{i\in[d]} \|\gamma^i\|_\infty$ and $\|\gamma\|_\beta := \max_{i\in[d]} \|\pi_i \circ \gamma\|_\infty$ on $\mathcal{C}_1^{\times d}$ and $\mathcal{C}_d$, respectively; as the norms $\|\cdot\|_\beta$ and $\|\cdot\|_\infty$ are equivalent on $\mathcal{C}_d$ (and hence induce the same Borel-structure on $\mathcal{C}_d$), the Banach isometry (123) extends to an isomorphism of Borel spaces

$$\left(\mathcal{C}_d, \|\cdot\|_\infty\right) \ \cong \ \left(\mathcal{C}_1 \times \cdots \times \mathcal{C}_1, \|\cdot\|_\alpha\right).$$

But since the factors $\mathcal{C}_1$ are separable metric spaces, so is $\mathcal{C}_1^{\times d}$, implying that (as $\|\cdot\|_\alpha$ induces the topology of componentwise convergence) $\mathcal{B}(\mathcal{C}_1^{\times d}) = \mathcal{B}(\mathcal{C}_1)^{\otimes d}$, see e.g. [6, Appendix M (p. 244)]. Using now that $\mathcal{B}(\mathcal{C}_d)$ is generated by the family of cylinder sets, we (upon testing on those) obtain that the factor identity (6) holds as desired.

A.2. **Proof of Lemma 1.**

**Lemma 1.** *Let $Y = (Y_t)_{t\in\mathbb{I}}$ be a stochastic process in $\mathbb{R}^d$ which is continuous with spatial support $D_Y$. Then the following holds:*

(i) *if* $f : D_Y \to \mathbb{R}^d$ *is a homeomorphism onto* $f(D_Y)$*, then* $f(D_Y) = D_{f(Y)}$*;*
(ii) *the traces* $\operatorname{tr}(Y(\omega)) := \{Y_t(\omega) \mid t \in \mathbb{I}\}$ *are contained in* $D_Y$ *for* $\mathbb{P}$*-almost each* $\omega \in \Omega$*;*
(iii) *for each open subset* $U$ *of* $D_Y$ *there is some* $t^\star \in \mathbb{I}$ *with* $\mathbb{P}(Y_{t^\star} \in U) > 0$*;*
(iv) *if each random vector* $Y_t$*,* $t \in \mathbb{I}$*, admits a continuous Lebesgue density on* $\mathbb{R}^d$*, then* $D_Y$ *is the closure of its interior.*
(v) *if each random vector* $Y_t$*,* $t \in \mathbb{I}$*, admits a density* $\upsilon_t$ *such that the function* $\upsilon^x : \mathbb{I} \ni t \mapsto \upsilon_t(x)$ *is continuous for each* $x \in D_Y$*, we for* $\dot{D}_t := \{\upsilon_t > 0\}$ *have that the set*

$$\bigcup_{(s,t)\in\Delta_2(\mathbb{I})} \dot{D}_s \cap \dot{D}_t \quad \textit{is} \quad \textit{dense} \quad \textit{in} \quad D_Y.$$

*Proof.* (i): Recalling the support $\operatorname{supp}(\mu)$ of a Borel-measure $\mu : \mathcal{B}(\mathbb{R}^d) \to [0, 1]$ to be defined as the smallest closed subset $C \subseteq \mathbb{R}^d$ having total mass $\mu(C) = 1$, it is easy to see that $\operatorname{supp}(\mu \circ f^{-1}) = f(\operatorname{supp}(\mu))$ for any Borel-measurable $f$. This implies that $D_{f(Y_t)} = \operatorname{supp}(\mathbb{P}_{Y_t} \circ f^{-1}) = f(D_{Y_t})$ for each $t \in \mathbb{I}$, from which the assertion $f(D_Y) = D_{f(Y)}$ is readily obtained because of (8) and the fact that $f : D_Y \to f(D_Y)$ is a homeomorphism.

(ii): By definition of $D_{Y_t}$, the event $\tilde{\Omega}_t := Y_t^{-1}(D_{Y_t})$ has probability one for each $t \in \mathbb{I}$, and hence so does the countable intersection $\tilde{\Omega} := \bigcap_{t\in\mathbb{I}\cap\mathbb{Q}} \tilde{\Omega}_t$. Now taking any $\tilde{\omega} \in \tilde{\Omega}$, we find by construction that the sample path $\gamma^{\tilde{\omega}} = Y(\tilde{\omega})$ satisfies $\operatorname{im}\big(\gamma^{\tilde{\omega}}|_{\tilde{\mathbb{I}}}\big) \subset D_Y$ for $\tilde{\mathbb{I}} := \mathbb{I} \cap \mathbb{Q}$. But since $\tilde{\mathbb{I}}$ is dense and the sample path $\gamma^{\tilde{\omega}}$ is continuous, we obtain $\operatorname{tr}(Y(\tilde{\omega})) = \operatorname{im}\big(\gamma^{\tilde{\omega}}\big) \subseteq D_Y$ by the fact that $D_Y$ is closed.

(iii): We proceed by contradiction: If $\mathbb{P}(Y_t \in U) = 0$ for each $t \in \mathbb{I}$, then $\mathbb{P}(Y_t \in U^c) = 1$ and thus (as $U^c$ is closed) $\operatorname{supp}(\mathbb{P}_{Y_t}) \subseteq U^c$ for each $t \in \mathbb{I}$, yielding $D_Y \equiv \overline{\bigcup_{t\in\mathbb{I}} \operatorname{supp}(\mathbb{P}_{Y_t})} \subseteq U^c$ in contradiction to $U \subset D_Y$.

(iv): Recall that in our notation, $D_Y = \overline{\bigcup_{t\in\mathbb{I}} D_{Y_t}}^{|\cdot|}$ for $D_{Y_t} := \operatorname{supp}(Y_t)$. Since $Y_t$ admits a continuous Lebesgue density $\chi_t \in C(\mathbb{R}^d)$ (vanishing identically outside of $D_{Y_t}$) by assumption, each of the sets $D_{Y_t} \overset{\text{def}}{=} \overline{\chi_t^{-1}(\{0\}^c)}^{|\cdot|}$, $t \in \mathbb{I}$, is the closure of an open set. Denote $D'_Y := \bigcup_{t\in\mathbb{I}} D_{Y_t}$. Then for each $u \in D_Y$ there is a sequence $(u_n)_n \subset D'_Y$ with $\lim_{n\to\infty} u_n = u$, and for each $n \in \mathbb{N}$ there is a sequence $(u_{n,m})_m \subset \operatorname{int}(D_{Y_{t_n}}) \subseteq \operatorname{int}(D_Y)$ (some $t_n \in \mathbb{I}$) with $\lim_{m\to\infty} u_{n,m} = u_n$ by the fact that each element of $\langle D_{Y_t}\rangle_{t\in\mathbb{I}}$ is the closure of its interior. With this, it is easy to see that there is a subsequence $(m_n)_n \subset \mathbb{N}$ such that $\lim_{n\to\infty} u_{n,m_n} = u$, proving $u \in \overline{\operatorname{int}(D_Y)}$ as claimed. The remaining inclusion $\overline{\operatorname{int}(D_Y)} \subseteq D_Y$ is clear as $\operatorname{int}(D_Y) \subset D_Y$ and $D_Y$ is closed.

(v): Suppose that the set $\tilde{D} := \bigcup_{(s,t)\in\Delta_2(\mathbb{I})} \dot{D}_s \cap \dot{D}_t$ is not dense in $D_S$. Then by (iv) the set $\tilde{D}$ is not dense in the interior of $D_S$, whence there exists $x_0 \in D_S$ and $\varepsilon > 0$ such that $B_\varepsilon(x_0) \subset D_S \setminus \tilde{D}$. Hence by (iii), there must then be some $t^\star \in \mathbb{I}$ such that $\mathcal{O} := \dot{D}_{t^\star} \cap B_\varepsilon(x_0) \neq \emptyset$. Now since $\mathcal{O} \subset \tilde{D}^c$, we for each $x \in \mathcal{O} \subseteq \dot{D}_{t^\star}$ have that $x \in \dot{D}_{t^\star} \cap (\dot{D}_s^c)$ for all $s \neq t^\star$, the latter implying that $x \in \dot{D}_{t^\star} \setminus \big(\bigcup_{s\neq t^\star} \dot{D}_s\big)$ and hence $\zeta_{t^\star}(x) > 0$ and $\zeta_s(x) = 0$ for all $s \neq t\star$, contradicting the continuity of $s \mapsto \zeta_s(x)$. □

A.3. **Probability Density of Projections of a Random Vector.** Given a $C^k$-distributed random vector $Z = (Z^1, \cdots, Z^n)$ in $\mathbb{R}^n$ with density $\varsigma$ together with some fixed subset $I \subseteq [n]$, say $I = \{i_1, \ldots, i_k\}$ with $k := |I|$, denote by $Z_I := \pi_I(Z) = (Z^{i_1}, \cdots, Z^{i_k})$ the random vector in $\mathbb{R}^k$ which is given by the projection of $Z$ to its $I$-indexed subcoordinates. Then $Z_I$ is $C^k$-distributed with Lebesgue density $\varsigma_I$ given by

$$\varsigma_I = \int_{\mathbb{R}^{n-k}} \varsigma \; \mathrm{d}x_1 \cdots \widehat{\mathrm{d}x_{i_1}} \cdots \widehat{\mathrm{d}x_{i_k}} \cdots \mathrm{d}x_n. \tag{124}$$

As an immediate consequence, we have the inclusion

$$\text{supp}(\varsigma) \;\subseteq\; \text{supp}(\varsigma_{[k]}) \times \text{supp}(\varsigma_{[n]\setminus[k]}) \qquad \text{for each} \quad k \in [n]. \tag{125}$$

Indeed, setting $C_k := \text{supp}(\varsigma_{[k]})$ and $C'_k := \text{supp}(\varsigma_{[n]\setminus[k]})$, we note that $\mathbb{P}(Z \in C_k \times C'_k) = \mathbb{P}(Z_{[k]} \in C_k,\, Z_{[n]\setminus[k]} \in C'_k) \geq \mathbb{P}(Z_{[k]} \in C_k) + \mathbb{P}(Z_{[n]\setminus[k]} \in C'_k) - 1 = 1$ and hence $C_k \times C'_k \supseteq \text{supp}(\mathbb{P}_Z) = \text{supp}(\varsigma)$, as claimed.

### A.4. **A Function Which is Strictly Non-Separable but Not Regularly Non-Separable.**

Consider the function $\varphi_0 \equiv \varphi_0(x,y)$ given by

$$\varphi_0(x,y) = \begin{cases} \exp\big(-1/(x-y)^2\big), & x<y \\ 0, & x=y \\ \exp\big(-1/(x-y)^2\big), & y<x, \end{cases} \qquad \text{and define} \qquad \varphi := e^{\varphi_0}.$$

Then clearly $\varphi \in C^2(\tilde{U}^{\times 2}; \mathbb{R}_{>0})$ for $\tilde{U} := (0,1)$, and as the mixed-log-derivatives $\partial_x\partial_y \log(\varphi) = \partial_x\partial_y\varphi_0$ vanish nowhere on the dense subset $\tilde{U}^2 \setminus \Delta_{\tilde{U}} \subset \tilde{U}^{\times 2}$ the function $\varphi$ is also strictly non-separable on $\tilde{U}^2$ by Lemma 2 (ii). However, since $(\partial_x\partial_y\varphi_0)|_{\Delta_{\tilde{U}}} = 0$ everywhere on $\tilde{U}$, the function $\varphi$ is clearly not regularly non-separable.

### A.5. **Proof of Lemma 3.**

**Lemma 3.** *Let $\varrho \in C^1(G)$ for some $G \subseteq \mathbb{R}^d$ open. Then the following holds.*

- (i) *Let $G$ also be connected. Then $\varrho \in \text{DP}_d(G)$ if and only if the Jacobian $J_\varrho$ of $\varrho$ is monomial on $G$, i.e. such that $\{J_\varrho(u) \mid u \in G\} \subseteq \text{M}_d$.*
- (ii) *Let $D$ be a subset of $G$ with the property that $D$ is the closure of an open subset $\mathcal{O}$ of $D$. If $J_\varrho$ is monomial on $\mathcal{O}$, then $\varrho$ is monomial on $D$.*

*Proof.* (i): The 'only-if' direction is clear, so let us prove that $\varrho \in \text{DP}_d(G)$ if

$$J_\varrho(u) = \big(\beta_\mu(u_\mu)\cdot\delta_{\nu,\sigma^u(\mu)}\big)_{\mu,\nu\in[d]} \;\in\; \text{M}_d, \qquad \forall\, u \equiv (u_\mu) \in G, \tag{126}$$

for some $\beta_\mu : \pi_\mu(G) \to \mathbb{R}_\times$ and $\{\sigma^u \mid u \in G\} \subseteq S_d$. To this end, we first note that

$$\sigma^u = \sigma^{u'} \quad \text{for any} \quad u, u' \in G. \tag{127}$$

Indeed: Fix any $u_0 \in G$ and note that since the Jacobian $J_\varrho \equiv (J^{ij}_\varrho)_{ij} : G \to \text{GL}_d$ of $\varrho$ is continuous, we for each $\varepsilon_{u_0} > 0$ can find a $\delta_{u_0} > 0$ such that ($B_{\delta_{u_0}}(u_0) \subseteq G$ and)

$$\|J_\varrho(u_0) - J_\varrho(u)\| < \varepsilon_{u_0} \quad \text{for all} \quad u \in B_{\delta_{u_0}}(u_0).$$

Taking $\|\cdot\|$ as the infinity-norm $\|A\| \equiv \|A\|_\infty := \max_{1\leq i\leq d}\sum_{j=1}^d |a_{ij}|$ and choosing $\varepsilon_{u_0} := \min_{1\leq 1\leq d}\sum_{j=1}^d |J^{ij}_\varrho(u_0)| > 0$, the fact that $J_\varrho(u_0), J_\varrho(u) \in \text{M}_d$ then readily implies that

$$\sigma^u = \sigma^{u_0} \quad \text{for all} \quad u \in B_{\delta_{u_0}}(u_0). \tag{128}$$

Let now $u, u' \in G$ be arbitrary. Since, as a connected subset of $\mathbb{R}^d$, the domain $G$ is also path-connected, the points $u$ and $u'$ can be joined by a continuous path $\gamma$ with $u = \gamma(0)$ and $u' = \gamma(1)$ and trace $\bar{\gamma} := \gamma([0,1]) \subseteq G$. We claim that

$$\sigma^{u_1} = \sigma^{u_2} \quad \text{for any} \quad u_1, u_2 \in \bar{\gamma}, \tag{129}$$

yielding (127) in particular. And indeed: Since $\bar{\gamma} \subset \bigcup_{u\in\bar{\gamma}} B_{\delta_u}(u)$ for $\{\delta_u\}$ as in (128), the compactness of $\bar{\gamma}$ yields that, for an $m \in \mathbb{N}$,

$$\bar{\gamma} \;\subseteq\; \bigcup_{j\in[m]} B_{\delta_{u_j}}(u_j) \qquad \text{for some} \;\; u_1, \ldots, u_m \in \bar{\gamma}. \tag{130}$$

But since the trace $\bar{\gamma}$ is connected, the $\{u_j\}$ from (130) can be renumerated such that $B_{\delta_{u_i}}(u_i) \cap B_{\delta_{u_{i+1}}}(u_{i+1}) \neq \emptyset$ for each $1 \leq i < m$, which implies (129) (and hence (127)) by way of (128).

Hence on $G$, the Jacobian (126) of $\varrho$ in fact is of the form

$$J_\varrho\big|_G = \big(\beta_\mu \cdot \delta_{\nu,\sigma(\mu)}\big)_{\mu,\nu\in[d]} \quad \text{for some} \quad \beta_\mu \in C(\pi_\mu(G); \mathbb{R}_\times) \quad \text{and} \quad \sigma \in S_d. \tag{131}$$

The assertion that $\varrho \equiv (\varrho_i) \in \mathrm{DP}_d(G)$ now follows from (131) and the mean value theorem (MVT): Given any $u_0 = (u_0^1, \ldots, u_0^d) \in G$, the MVT implies[48] that for each $u = (u_1, \ldots, u_d) \in G$ which is connected to $u_0$ via the line segment $\overline{u_0, u} \equiv \{u_0 + t \cdot (u - u_0) \mid t \in [0,1]\} \subset G$ and any $v \in \mathbb{R}^d$, there exists a point $\xi \in \overline{u_0, u}$ such that

$$v \cdot (\varrho(u) - \varrho(u_0)) = v \cdot J_\varrho(\xi) \cdot (u - u_0). \tag{132}$$

Hence if for $i \in [d]$ we take $u$ with $u_{\sigma(i)} = u_0^{\sigma(i)}$ and choose $v = e_i$ (for $(e_i)_{i\in[d]}$ the standard basis of $\mathbb{R}^d$), then by way of (132) and (131) we find that

$$\varrho_i(u) - \varrho_i(u_0) = \big[J_\varrho(\xi) \cdot (u - u_0)\big]_i = 0 \qquad (i \in [d]).$$

This implies that for any given $u_0 \equiv (u_0^1, \ldots, u_0^d) \in G$ we have $\varrho_i(u) = \varrho_i(u_0^{\sigma(i)})$ for all $u \in G_{u_0} \cap \pi_{\sigma(i)}^{-1}(\{u_0^{\sigma(i)}\})$ $(= \{(u_i) \in G_{u_0} \mid u_{\sigma(i)} = u_0^{\sigma(i)}\})$, where

$$G_{u_0} := \{u \in G \mid \exists \text{ polygonal path connecting } u \text{ and } u_0\}$$

denotes the polygonally-connected component of $G$ containing $u_0$. But since (as a connected open subset of $\mathbb{R}^d$) the domain $G$ is polygonally-connected, we have that $G_u = G$ for each $u \in G$ and hence find that

$$\varrho_i(u) = \varrho_i(u_{\sigma(i)}) \quad \text{for each} \quad u \equiv (u_1, \ldots, u_d) \in G \qquad (i \in [d]),$$

hence the diffeomorphism $\varrho \equiv (\varrho_1, \cdots, \varrho_d)$ is monomial on $G$ as claimed.

(ii) : Let $\mathcal{O}$ be an open and dense subset of $D$ with the property that

$$J_\varrho(u) \in \mathrm{M}_d \quad \text{for each} \quad u \in \mathcal{O}.$$

Since each $u \in \mathcal{O}$ has a connected open neighbourhood (as $\mathcal{O}$ is open), statement (i) then gives that for every $u \in \mathcal{O}$ there is a permutation $\sigma^u \in S_d$ such that

$$\varrho_i(u) = \varrho_i(u_{\sigma^u(i)}) \quad \text{for each} \quad i \in [d]. \tag{133}$$

Let now $Z \subseteq D$ be connected and $z^* \equiv (z_1^*, \ldots, z_d^*) \in Z$ be arbitrary but fixed. Then, as $\mathcal{O}$ is dense in $\mathcal{O}$, there will be a sequence $(u^{(k)})_{k\in\mathbb{N}} \subset \mathcal{O}$ with $\lim_{k\to\infty} u^{(k)} = z^*$, implying that

$$J_\varrho(z^*) = \lim_{k\to\infty} J_\varrho(u^{(k)}) \tag{134}$$

due to $\varrho$ being continuously differentiable on $G$. But since $J_\varrho(z^*) \in \mathrm{GL}_d$, we then obtain that $J_\varrho(z^*) \in \mathrm{M}_d$ by (134) and the fact that the subset $\mathrm{M}_d$ is closed in $\mathrm{GL}_d$. Moreover, denoting the permutation of $J_\varrho(z^*)$ by $\sigma_{z^*} \in S_d$, the argument behind (128) gives that, by way of (133),

$$\varrho_i(z^*) = \lim_{k\to\infty} \varrho_i(u^{(k)}_{\sigma_{z^*}(i)}) = \varrho_i(\lim_{k\to\infty} u^{(k)}_{\sigma_{z^*}(i)}) = \varrho_i(z^*_{\sigma_{z^*}(i)}) \tag{135}$$

for each $i \in [d]$. Let now $z \in Z$ be arbitrary. Then by the (path-)connectedness of $Z$, the points $z$ and $z^*$ can be joined by a continuous path with trace $\bar{\gamma} \subseteq Z \subseteq G$. As we have just shown that $J_\varrho$ is monomial on $\bar{\gamma}$, the same argument that led us to (129) now yields that, by way of (135), we for each $i \in [d]$ have

$$\varrho_i(z) = \varrho_i(z_{\sigma(i)}) \quad \text{for each } z \in \bar{\gamma}, \quad \text{with } \sigma = \sigma_{z^*}.$$

[48] Indeed: For $u, u_0, v$ as above, define $\varphi(t) := v \cdot \varrho(u_0 + t \cdot \eta)$ for $t \in I_\delta \equiv (-\delta, 1 + \delta)$ and $\eta := u - u_0$ and $\delta > 0$ s.t. $\{u_0 + t \cdot \eta \mid t \in I_\delta\} \subset G$ (such a $\delta$ exists as $G$ is open). Then $\varphi \in C^1(I_\delta)$, whence (132) follows from the (classical) MVT applied to the difference $\varphi(1) - \varphi(0)$.

But since $J_\varrho(z) \equiv (\beta_\mu(z_{\sigma(\mu)}) \cdot \delta_{\nu,\sigma(\mu)})_{\mu,\nu\in[d]} \in \mathrm{M}_d$ for each $z \in \bar{\gamma}$, the derivatives $\frac{\mathrm{d}}{\mathrm{d}z_{\sigma(i)}}\varrho_i = \beta_i$ vanish nowhere on $\pi_{\sigma(i)}(\bar{\gamma}) \subset \mathbb{R}$, giving that $\varrho\big|_{\bar{\gamma}} \in \mathrm{DP}_d(\bar{\gamma})$ which yields the claim. ☐

A.6. **Proof of Lemma 4.** Recall that $Y, Y^*$ are defined by (23) and $\rho$ is given by (24).

**Lemma 4.** *For $\mu$ the probability density of $Y$, and $\mu^*$ the probability density of $Y^*$,*

$$\psi \circ \rho = \log \mu - \log \mu^* \quad a.e.\ on \quad \tilde{D} := \operatorname{supp}(\mu) \tag{25}$$

*for the logit-function $\psi(p) := \log(p/(1-p))$.*

*Proof.* We note first that since the support $\operatorname{supp}(\nu)$ of a (Borel) probability measure $\nu : \mathcal{B}(E) \to [0,1]$ is defined as the smallest closed set $C \subseteq E$ having total mass $\nu(C) = 1$, it is easy to see that $\operatorname{supp}(\tilde{f}_*\nu) = \tilde{f}(\operatorname{supp}(\nu))$ for any Borel-measurable $\tilde{f}$. For the above case, this implies $\operatorname{supp}(\mu) = (f \times f)(\operatorname{supp}(\zeta)) = (f \times f)(\bar{D})$, whence $\partial(\operatorname{supp}(\mu))$ is a Lebesgue-nullset (as is $\partial\bar{D}$, by assumption) and hence $\mu > 0$ a.e. on $\operatorname{supp}(\mu)$. Since also $\operatorname{supp}(\mu^*) = \operatorname{supp}(\mathbb{P}_{X_s}) \times \operatorname{supp}(\mathbb{P}_{X_t})$, we further obtain $\operatorname{supp}(\mu) = \operatorname{supp}(\mathbb{P}_{(X_s,X_t)}) \subseteq \operatorname{supp}(\mu^*)$ by (125), which implies that the RHS of (25) is defined a.e. on $\operatorname{supp}(\mu)$ indeed.

Note now that since by definition the function $\rho$ equals the conditional probability of the event $\{C = 1\}$ given $\bar{Y}$, we have

$$\rho \cdot \frac{\mathrm{d}\mathbb{P}_{\bar{Y}}}{\mathrm{d}y} = \mathbb{P}(C = 1\,|\,\bar{Y}) \cdot \frac{\mathrm{d}\mathbb{P}_{\bar{Y}}}{\mathrm{d}y} = \frac{\mathrm{d}\mathbb{P}_{\bar{Y}}(\,\cdot\,|\,C = 1)}{\mathrm{d}y} \cdot \mathbb{P}(C = 1) \tag{136}$$

almost everywhere, where the first factor on the RHS of (136) denotes (a regular version of) the conditional density of $\bar{Y}$ given $C = 1$.[49] Next we observe that

$$\frac{\mathrm{d}\mathbb{P}_{\bar{Y}}(\,\cdot\,|\,C = 1)}{\mathrm{d}y} \cdot \mathbb{P}(C = 1) = \tfrac{1}{2}\mu \qquad \text{(a.e.)}. \tag{137}$$

Indeed, denote by $\eta$ the LHS of (137) and let $A \in \mathcal{B}(\mathbb{R}^{2d})$ be arbitrary. Then, since by construction $\mathbb{P}_{(C,\bar{Y})} = \mathbb{P}_C \otimes \mathbb{P}_C^{\bar{Y}}$ and $\mathbb{P}_{C=1}^{\bar{Y}} \equiv \mathbb{P}_{\bar{Y}}(\,\cdot\,|\,C = 1) = \mathbb{P}_Y$ and $\mathbb{P}(C = 1) = \frac{1}{2}$,

$$\begin{aligned}\int_{\mathbb{R}^{2d}} \eta \cdot \mathbb{1}_A \,\mathrm{d}y \;&=\; \mathbb{P}(\bar{Y} \in A\,|\,C = 1) \cdot \mathbb{P}(C = 1) = \mathbb{P}(\bar{Y} \in A,\ C = 1)\\ &=\; \mathbb{P}_{(C,\bar{Y})}(\{1\} \times A) = \mathbb{P}(C = 1) \cdot \mathbb{P}_Y(A) = \int_{\mathbb{R}^{2d}} \tfrac{1}{2}\mu \cdot \mathbb{1}_A \,\mathrm{d}y\end{aligned}$$

from which (137) follows by the fundamental lemma of calculus of variations. Combining (136) with the fact that $\frac{\mathrm{d}\mathbb{P}_{\bar{Y}}}{\mathrm{d}y} = \frac{1}{2}\mu + \frac{1}{2}\mu^*$ and (137) now yields the identity $(\mu + \mu^*) \cdot \rho = \mu$ (a.e.), from which equation (25) follows immediately. ☐

A.7. **Proof of Proposition 2 (ii), (iii).**

*Proof.* (ii): It is well-known that the covariance function $\kappa^i$ of (74) reads

$$\kappa^i(s,t) = \gamma_i\big(e^{-\theta_i|s-t|} - e^{-\theta_i(s+t)}\big) \quad \text{for} \quad \gamma_i := \frac{\sigma_i^2}{2\theta_i}.$$

---

[49] Indeed, abbreviating $\ell := \rho \cdot \frac{\mathrm{d}\mathbb{P}_{\bar{Y}}}{\mathrm{d}y}$ and $r := \frac{\mathrm{d}\mathbb{P}(\,\cdot\,|\,C=1)}{\mathrm{d}y} \cdot \mathbb{P}(C = 1)$, we for any $A \in \mathcal{B}(\mathbb{R}^{2d})$ find

$$\begin{aligned}\int_{\mathbb{R}^{2d}} r \cdot \mathbb{1}_A \,\mathrm{d}y \;&= \mathbb{P}(C = 1)\int_A \mathbb{P}_{\bar{Y}}(\mathrm{d}y\,|\,C = 1) = \mathbb{P}((C, \bar{Y}) \in \{1\} \times A)\\ &= \mathbb{P}((\bar{Y}, C) \in A \times \{1\}) = \int_A \mathbb{P}(C = 1\,|\,\bar{Y} = y)\,\mathbb{P}_{\bar{Y}}(\mathrm{d}y) = \int_{\mathbb{R}^{2d}} \ell \cdot \mathbb{1}_A \,\mathrm{d}y\end{aligned}$$

which implies $r = \ell$ (a.e.) by the fundamental lemma of calculus of variations. (Note that the second and the fourth of the above equations hold by definition of conditional distributions.)

Suppose $\mathbb{I} = [0,1]$ wlog. Then for $\mathfrak{p}_\nu \equiv (t_\nu, 2t_\nu) \in \Delta_2(\mathbb{I})$ $(\nu = 0,1)$ with $t_0 \neq t_1$ and $\mathfrak{p}_2 := \mathfrak{p}_0$, we obtain $\kappa^i_{\mathfrak{p}_\nu} = \gamma_i(e^{-\theta_i t_\nu} - e^{-3\theta_i t_\nu})$ and $k^i_{\mathfrak{p}_\nu} = \kappa^i_{\mathfrak{p}_\nu} \cdot \gamma_i e^{\theta_i t_\nu} \cdot (1 - e^{-4\theta_i t_\nu})$ and, thus, $\Xi_i = \tilde{\Xi}_i^{(1)} / \tilde{\Xi}_i^{(0)}$ for each $i \in [d]$, with the factors

$$\tilde{\Xi}_i^{(\nu)} = \frac{\kappa^i_{\mathfrak{p}_\nu}}{k^i_{\mathfrak{p}_\nu} - (\kappa^i_{\mathfrak{p}_\nu})^2} = \left(\gamma_i e^{\theta_i t_\nu} \cdot (1 - e^{-4\theta_i t_\nu}) - \kappa^i_{\mathfrak{p}_\nu}\right)^{-1} = \left(\frac{\sigma_i^2}{\theta_i} \sinh(\theta_i t_\nu)\right)^{-1}.$$

We hence have the parametrisation $\theta_i \mapsto \Xi_i \equiv \Xi_i(\theta_i) = \frac{\sinh(\theta_i t_0)}{\sinh(\theta_i t_1)}$, which due to $t_0 \neq t_1$ is strictly monotone and differentiable in $\theta_i > 0$. Denoting by $\tilde{\phi}_i$ the associated (differentiable) inverse of $\theta_i \mapsto \Xi_i(\theta_i)$, we obtain that $(\Xi_i(\theta_i))_{i\in[d]} \in (\mathbb{R}^d \setminus \nabla^\times)$ provided the parameter vector $(\theta_i)_{i\in[d]} \in \mathbb{R}^d_{>0}$ is not contained in the nullset $\tilde{\mathcal{N}} := (\tilde{\phi}_1 \times \cdots \times \tilde{\phi}_d)(\nabla^\times)$.

(iii): Choosing again $\mathfrak{p}_\nu \equiv (t_\nu, 2t_\nu) \in \Delta_2(\mathbb{I})$ $(\nu = 0,1)$ with $t_0 \neq t_1$ and $\mathfrak{p}_2 := \mathfrak{p}_0$, we for each $i \in [d]$ find that $\Xi_i = \tilde{\Xi}_i^{(1)} / \Xi_i^{(0)}$ for the factors

$$\tilde{\Xi}_i^{(\nu)} = \frac{4^{H_i - \frac{1}{2}} \cdot t_\nu^{2H_i}}{4^{H_i} \cdot t_\nu^{4H_i} - 4^{2H_i - 1} \cdot t_\nu^{4H_i}} = \left(2(1 - 4^{H_i - 1}) \cdot t_\nu^{2H_i}\right)^{-1},$$

whence it holds that

$$\Xi_i = \left(\frac{t_0}{t_1}\right)^{2H_i} \qquad \text{for each} \quad i \in [d].$$

But since due to $t_0 \neq t_1$ the assignment $h \mapsto (\frac{t_0}{t_1})^{2h}$ is clearly injective, we clearly obtain that $(\Xi_i)_{\in[d]}$ is not in $\nabla^\times$ whenever $(H_i)_{i\in[d]}$ is not in $\nabla^\times$, as claimed. ☐

A.8. **Proof of Proposition 3.**

**Proposition 3.** *Let $S = (S_t)_{t\geq 0} = (S^1, \cdots, S^d)$ be an IC geometric Brownian motion in $\mathbb{R}^d$, i.e. suppose that there is a standard Brownian motions $B = (B^1_t, \cdots, B^d_t)_{t\geq 0}$ such that*

$$\mathrm{d}S^i_t = S^i_t \cdot \left(\mu_i(t)\,\mathrm{d}t \; + \; \sigma_i(t)\,\mathrm{d}B^i_t\right), \quad S^i_0 = s^i_0 \qquad (i \in [d])$$

*for some $s^i_0 > 0$ and continuous functions $\mu_i : \mathbb{I} \to \mathbb{R}$ and $\sigma_i : \mathbb{I} \to \mathbb{R}_{>0}$. Then $S$ has spatial support $D_S = \mathbb{R}^d_+$, and $S$ is $\gamma$-contrastive if there are $r_0, r_1 \in \mathbb{I}$ for which the numbers $\{\sigma_i^2(r_0) \cdot \sigma_j^2(r_1) \mid (i,j) \in [d] \times [d]\}$ are pairwise distinct.*

*Proof.* A straightforward application of Itô's lemma yields that for any $(s,t) \in \Delta_2(\mathbb{I})$, the density $\zeta^i_{s,t}$ of $(S^i_s, S^i_t)$ is given by

$$\zeta^i_{s,t}(x,y) = \rho^i_{s,t}(\log(x), \log(y)) \cdot (xy)^{-1} = c_i(s,t,x,y) \cdot \exp\bigl(\varphi_i(s,t,x,y)\bigr)$$

for the functions $c_i : \Delta_2(\mathbb{I}) \times \mathbb{R}^2_{>0} \to \mathbb{R}$ and $\varphi_i : \Delta_2(\mathbb{I}) \times \mathbb{R}^2_{>0} \to \mathbb{R}$ defined by

$$c_i(s,t,x,y) = \frac{(4\pi^2 \cdot \det(\mathfrak{s}^i_{s,t}))^{-1/2}}{xy} \qquad \text{and}$$

$$\begin{aligned} \varphi_i(s,t,x,y) &= -\frac{1}{2}\left([\phi^{-1}(x,y) - \mathfrak{m}^i_{s,t}]^\intercal \cdot (\mathfrak{s}^i_{s,t})^{-1} \cdot [\phi^{-1}(x,y) - \mathfrak{m}^i_{s,t}]\right) \\ &= \beta^i_{s,t} \log(x)\log(y) + \eta_i(s,t,x) + \tilde{\eta}_i(s,t,y) \end{aligned}$$

with $\eta_i, \tilde{\eta}_i$ given by $\eta_i(s,t,x) + \tilde{\eta}_i(s,t,y) := \varphi_i(s,t,x,y) - \beta^i_{s,t}$ and

$$\beta^i_{s,t} := \frac{\kappa_i(s,t)}{\kappa_i(s,s)\kappa_i(t,t) - \kappa_i^2(s,t)} = \frac{\int_0^s \sigma_i^2(r)\,\mathrm{d}r}{\left(\int_0^s \sigma_i^2(r)\,\mathrm{d}r\right)\left(\int_0^t \sigma_i^2(r)\,\mathrm{d}r\right) - \left(\int_0^s \sigma_i^2(r)\,\mathrm{d}r\right)^2}.$$

Consequently, the spatial support of $S$ is $D_S = \mathbb{R}^d_+ = \pi_{[d]}\big(\mathrm{supp}\big[\mathbb{R}^{2d} \ni (u,v) \mapsto \zeta^i_{s,t}(u_i, v_i)\big]\big)$ (any $i \in [d]$), and the mixed log-derivatives of $\zeta^i_{s,t}$ read

$$\xi^i_{s,t} := \partial_x\partial_y \log(\zeta^i_{s,t}) = \partial_x\partial_y\varphi^i(s,t,\cdot,\cdot) = \frac{\beta^i_{s,t}}{xy}.$$

Hence by Definition 7, the process $S$ is $\gamma$-contrastive iff

$$\exists\, \mathfrak{p}_0, \mathfrak{p}_1, \mathfrak{p}_2 \in \Delta_2(\mathbb{I}) \quad \text{with} \quad (\Xi_i)_{i\in[d]} := \left(\frac{\beta^i_{\mathfrak{p}_1}\beta^i_{\mathfrak{p}_2}}{(\beta^i_{\mathfrak{p}_0})^2}\right)_{i\in[d]} \in (\mathbb{R}^d \setminus \nabla^\times).$$

Having $\mathfrak{p}_\nu \equiv (s_\nu, t_\nu) \in \Delta_2(\mathbb{I})$ $(\nu = 0,1)$ arbitrary and $\mathfrak{p}_2 := \mathfrak{p}_0$ hence yields that

$$\Xi_i = \frac{\beta^i_{\mathfrak{p}_1}}{\beta^i_{\mathfrak{p}_0}} = \frac{\int_{s_0}^{t_0}\sigma_i^2(r)\,\mathrm{d}r}{\int_{s_1}^{t_1}\sigma_i^2(r)\,\mathrm{d}r} \qquad \text{for each } \ i \in [d].$$

Thus by choosing $\mathfrak{p}_0$ and $\mathfrak{p}_1$ as done in the proof of Proposition 2 (iv), we obtain $(\Xi_i)_{i\in[d]} \in (\mathbb{R}^d \setminus \nabla^\times)$ as desired. ☐

A.9. **Basic Cross-Shuffle Combinatorics.** Using the notation of Sections 7, 8.1 and C.2 throughout this remark, consider the family of *cross-shuffles* $\mathfrak{C} \equiv \bigsqcup_{k=2}^d \mathcal{W}_k = \bigsqcup_{\nu=2}^\infty \mathfrak{C}_\nu$.

By definition (80) of the shuffle product, each element $\boldsymbol{q} \in \mathfrak{C}_\nu$ (seen as an element of (189)) is a homogeneous polynomial of degree $\nu$ whose monomial coefficients are all 1, i.e. there is $c_{\boldsymbol{q}} \in \mathbb{N}$ such that $\boldsymbol{q} = \boldsymbol{q}_1 + \cdots + \boldsymbol{q}_{c_{\boldsymbol{q}}}$ with $\boldsymbol{q}_j \in [d]^*$ for each $j \in [c_{\boldsymbol{q}}]$. Partitioning

$$\mathfrak{C}_\nu = \bigsqcup_{k=2}^d \mathfrak{C}_{\nu|k} \quad \text{for} \quad \mathfrak{C}_{\nu|k} := \mathfrak{C}_\nu \cap \mathcal{W}_k,$$

the definition of $\mathcal{W}_k$ yields that for any given $\boldsymbol{q} \in \mathfrak{C}_\nu$ the pair $\vartheta_{\boldsymbol{q}} \equiv (k_{\boldsymbol{q}}, \mu_{\boldsymbol{q}}) \in [d]_{\geq 2} \times [m-1]$, with $k_{\boldsymbol{q}} := \max\{i \in [d] \mid i \in \boldsymbol{q}_1\}$ being the largest letter contained in $\boldsymbol{q}$ and $\mu_{\boldsymbol{q}} := \sum_{i\in\boldsymbol{q}_1}\delta_{i,k_{\boldsymbol{q}}}$ denoting the number of times this largest letter appears in one (and hence any) of the monomials of $\boldsymbol{q}$, determines $\boldsymbol{q}$ uniquely (in $\mathfrak{C}_\nu$) up to a $\sqcup\!\sqcup$-left-factor of word length $\nu - \mu_{\boldsymbol{q}}$. (Indeed: Given $\boldsymbol{q} \in \mathfrak{C}_\nu$, the number $k_{\boldsymbol{q}} \in \mathbb{N}$ is the (unique) index s.t. $\boldsymbol{q} \in \mathfrak{C}_{\nu|k_{\boldsymbol{q}}} \subset \mathcal{W}_{k_{\boldsymbol{q}}}$, whence $\boldsymbol{q} = \boldsymbol{w} \sqcup\!\sqcup (k_{\boldsymbol{q}})^{*\mu_{\boldsymbol{q}}}$ for some $\boldsymbol{w} \in [k_{\boldsymbol{q}} - 1]^*$ with $|\boldsymbol{w}| = \nu - \mu_{\boldsymbol{q}}$.)

Since the shuffle product (80) of two words $\boldsymbol{i}, \boldsymbol{j} \in [d]^*$ is precisely the sum over the $c_{\boldsymbol{i}\sqcup\!\sqcup\boldsymbol{j}}$ $\big(= \frac{(|\boldsymbol{i}|+|\boldsymbol{j}|)!}{|\boldsymbol{i}|!|\boldsymbol{j}|!}\big)$ ways of *interleaving* $\boldsymbol{i}$ and $\boldsymbol{j}$, any two monomials in $\boldsymbol{i}\sqcup\!\sqcup\boldsymbol{j}$ are composed of exactly the same letters and differ only in the order in which their letters appear. Consequently,

(a) any two $\boldsymbol{q}, \boldsymbol{q}' \in \mathfrak{C}_\nu$ have a monomial in common iff $\boldsymbol{q} = \boldsymbol{q}'$ ;
(b) given any $\boldsymbol{q} \in \mathfrak{C}_\nu$ with its unique[50] decomposition $\boldsymbol{q} = \boldsymbol{q}_1 + \ldots + \boldsymbol{q}_{c_{\boldsymbol{q}}}$ into monic monomials $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{c_{\boldsymbol{q}}} \in [d]^*$, these monomials $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{c_{\boldsymbol{q}}}$ are pairwise distinct.

(Note that point (b) follows inductively: Let $\boldsymbol{q} \equiv \boldsymbol{w} \sqcup\!\sqcup (k_{\boldsymbol{q}})^{*\mu_{\boldsymbol{q}}} \in \mathfrak{C}_\nu$ with $(k_{\boldsymbol{q}}, \mu_{\boldsymbol{q}}) = \vartheta_{\boldsymbol{q}}$. The assertion clearly holds if $\mu_{\boldsymbol{q}} = 1$ or (by symmetry) $|\boldsymbol{w}| = 1$. Fixing any $\boldsymbol{q}$ as above, assume that (b) holds for any $\boldsymbol{r} \equiv \boldsymbol{w}' \sqcup\!\sqcup (k_{\boldsymbol{q}})^{*\mu_{\boldsymbol{r}}} \in \mathfrak{C}$ with $\mu_{\boldsymbol{r}} = \mu_{\boldsymbol{q}} - 1$ or $|\boldsymbol{w}'| = |\boldsymbol{w}| - 1$. Note that

$$\boldsymbol{q} = (\boldsymbol{w}' * \mathtt{i}) \sqcup\!\sqcup (k_{\boldsymbol{q}})^{*(\mu'_{\boldsymbol{q}}+1)} = \boldsymbol{r}_0 * \mathtt{i} + \boldsymbol{r}_1 * k_{\boldsymbol{q}} \qquad (\boldsymbol{w} =: \boldsymbol{w}' * \mathtt{i},\ \mathtt{i} \in [k_{\boldsymbol{q}} - 1] \setminus \{\epsilon\})$$

for the polynomials $\boldsymbol{r}_0 := \boldsymbol{w}' \sqcup\!\sqcup (k_{\boldsymbol{q}})^{*\mu_{\boldsymbol{q}}}$ and $\boldsymbol{r}_1 := \boldsymbol{w} \sqcup\!\sqcup (k_{\boldsymbol{q}})^{*(\mu_{\boldsymbol{q}}-1)}$ (by the recursive formulation of the shuffle product, e.g. [69, p. 25 f.]). Now since the monomials of $\boldsymbol{r}_0$ and $\boldsymbol{r}_1$ are all monic and pairwise distinct by induction hypothesis, it is clear that the same applies to $\boldsymbol{r}_0 * \mathtt{i}$ and $\boldsymbol{r}_1 * k_{\boldsymbol{q}}$ and, hence (as $\mathtt{i} \neq k_{\boldsymbol{q}}$), to $\boldsymbol{q}$. Thus by induction, assertion (b) holds for $\boldsymbol{q}$ as desired.)

[50] Up to the order of summands.

A.10. **Nonlinear ICA for Discrete-Time Signals.** As detailed in this section, our approach towards the identifiability of nonlinearly mixed stochastic processes also covers the case of discrete-time signals with almost no further modifications.

With the robustness of Theorem 4 under the 'sampling-based' discretization (83) of a continuous-time observable $X = (X_t)_{t\in\mathbb{I}}$ being established via Proposition 5 and Lemma 9, assume here that $X_* \equiv (X_j)_{j\in\mathbb{Z}}$ is any discrete time series in $\mathbb{R}^d$ with the property that

$$X_* = f(S_*) \equiv (f(S_j))_{j\in\mathbb{Z}} \tag{138}$$

for $S_* \equiv (S_j)_{j\in\mathbb{Z}}$ some IC discrete time-series in $\mathbb{R}^d$ and $f \in C^{3,3}(D_{S_*};\mathbb{R}^d)$.

Here, a time series $Y_* \equiv (Y_j)_{j\in\mathbb{Z}}$ in $\mathbb{R}^d$, with $Y_j \equiv (Y_j^1,\dots,Y_j^d)$ for each $j\in\mathbb{Z}$, is called *IC* if its componental time-series $(Y_j^1)_{j\in\mathbb{Z}},\dots,(Y_j^d)_{j\in\mathbb{Z}}$ are mutually independent; we further denote by $D_{Y_*} := \overline{\bigcup_{j\in\mathbb{Z}}\mathrm{supp}(Y_j)}^{|\cdot|_2}$ the *spatial support* of $Y_*$.

Denote $\Delta_2(\mathbb{Z}) := \{(j_1,j_2)\in\mathbb{Z}^2 \mid j_1 < j_2\}$.

*Definition* A.1 ($\bar\alpha,\bar\beta,\bar\gamma$-Contrastive). A discrete time-series $S_* \equiv (S_j)_{j\in\mathbb{Z}}$ in $\mathbb{R}^d$ with spatial support $D_{S_*}$ will be called *$\bar\alpha$-contrastive* if $S_*$ is IC and there exists $\mathcal{P}\subseteq\Delta_2(\mathbb{Z})$ together with a collection $(D_\mathfrak{p})_{\mathfrak{p}\in\mathcal{P}}$ of open subsets in $\mathbb{R}^d$ such that

(i) the union $\bigcup_{\mathfrak{p}\in\mathcal{P}} D_\mathfrak{p}$ is dense in $D_{S_*}$, and

(ii) for each $(i,(j_1,j_2))\in[d]\times\mathcal{P}$, the vector $(S^i_{j_1},S^i_{j_2})$ is $C^2$-regular with density $\zeta^i_{j_1,j_2}$ s.t.

$$\zeta^i_{j_1,j_2}\big|_{D^{\times 2}_{(j_1,j_2)}} \text{ is regularly non-separable for all } i\in[d], \quad \text{and}$$
$$\zeta^i_{j_1,j_2}\big|_{D^{\times 2}_{(j_1,j_2)}} \text{ is almost everywhere non-Gaussian for all but at most one } i\in[d]$$

(cf. Definition 6). The notions of $\bar\beta$- and $\bar\gamma$-*contrastive time series* are defined analogously in adaptation of Definition 7.

**Theorem A.1.** *For $X_*$ and $S_*$ as in (138) with spatial supports $D_{X_*}$ and $D_{S_*}$ respectively, let the time-series $S_*$ in be $\bar\alpha$-, $\bar\beta$- or $\bar\gamma$-contrastive. Then, for any transformation $h$ which is $C^3$-invertible on some open superset of $D_{X_*}$, we have with probability one that:*

$$(h\circ f)\big|_{\tilde Z} \in \mathrm{DP}_d(\tilde Z),\ \forall\, Z\subseteq D_{X_*} \text{ connected} \quad \text{if and only if} \quad h\cdot X_* \text{ is IC}, \tag{139}$$

*where for any $Z\subseteq D_{X_*}$ connected we denoted $\tilde Z := f^{-1}(Z)$.*

*Proof.* Let $S_*$ be $\bar\alpha$-contrastive, and $h\in C^{3,3}(D_{X_*};\mathbb{R}^d)$ be such that $h\cdot X_*$ is IC. Then

$$(h\times h)(X_{j_1},X_{j_2}) \text{ is IC for any fixed } (j_1,j_2)\in\mathcal{P},$$

which in consequence of Definition A.1 (ii) implies that

$$\text{the Jacobian of } \varrho := h\circ f \text{ is monomial on } D_{(j_1,j_2)},$$

as detailed in the proof of Theorem 2. The equivalence (139) thus follows from Def. A.1 (i) and Lemma 3. The case of $S_*$ being $\bar\beta$- or $\bar\gamma$-contrastive follows similarly via Theorem 3. ☐

Since a discrete time-series $Y_*$ in $\mathbb{R}^d$ is IC if and only if its piecewise-linear interpolation[51] $\hat Y_*$ is IC in $\mathcal{C}_d$, the assertion of Theorem 4 remains valid as stated[52] if $(S,\alpha,\beta,\gamma,X)$ is replaced by $(S_*,\bar\alpha,\bar\beta,\bar\gamma,X_*)$ and the argument $h\cdot X$ in (82) is replaced by the piecewise-linear interpolation of $h\cdot X_*$. This shows that the identifiability theory of Sections 4, 5 and 7 directly applies to the discrete-time setting (138).

[51] … along any (countable) dissection of, say, $[0,1]$.

[52] With the addition that in (82), the monomial transformations $\alpha$ with $h\cdot X = \alpha\cdot S$ then depend on $(j,\omega)$ via the connected component of $D_{S_*}$ that the given realisation of $S_j = S_j(\omega)$ is contained in.

### A.11. **Proof of Proposition 7.**

**Proposition 7.** *Let $X$ and $S$ be as in (13) with $S$ IC, and $h$ be $C^1$-invertible on some open superset of $D_X$. Then for $\mathcal{I} \subset \mathbb{I}$ finite and $\varrho$ as in (114), we have that:*

$$(h \cdot X_t)_{t\in\mathcal{I}} \;\in\; \mathrm{DP}_d \cdot (S_t)_{t\in\mathcal{I}} \quad \textit{iff} \quad \varrho\big((h \cdot X_t)_{t\in\mathcal{I}},\, (S_t)_{t\in\mathcal{I}}\big) = 0. \tag{115}$$

*Proof.* This is a direct consequence of the fact that Kendall's (and Spearman's) rank correlation coefficient $\rho_{\mathrm{K}}$ attains its extreme values $\pm 1$ iff one of its arguments is a monotone transformation of the other (cf. e.g. [25, Theorem 3 (7.), (8.)]), combined with the fact that $\rho_{\mathrm{K}}(U,V) = 0$ if $U$ and $V$ are independent (cf. e.g. [25, Theorem 3 (2.)]).

Indeed, note for the 'if'-direction in (115) that $\varrho\big((h(X_t))_{t\in\mathcal{I}},\, (S_t)_{t\in\mathcal{I}}\big) = 0$ implies that there is $\sigma \in S_d$ with $|\rho_{\mathrm{K}}(h_i(X_t), S_t^j)| = \delta_{j,\sigma(i)}$ for each $t \in \mathcal{I}$ and $i \in [d]$, which by the above-mentioned property of $\rho_{\mathrm{K}}$ yields that $(h_i \circ f)(S_t) = \alpha_{i|t}(S_t^{\sigma(i)})$, and hence $(h_i \circ f)\big|_{\mathrm{supp}(S_t)} = \alpha_{i|t}\big|_{\mathrm{supp}(S_t)}$, for some function $\alpha_{i|t} : \mathrm{supp}(S_t) \to \mathbb{R}$ with $\mathrm{supp}(S_t^{\sigma(i)}) \ni x_{\sigma(i)} \mapsto \alpha_{i|t}(x_{\sigma(i)}) \equiv \alpha_{i|t}(x)$ monotone. But since for each $i \in [d]$ we have $h_i \circ f \in C^1(\mathcal{O}_S)$ for some $\mathcal{O}_S \supset D_S$ open, the classical pasting lemma[53] guarantees that the functions $(\alpha_{i|t} \mid t \in \mathcal{I})$ can be 'glued together' to an injective $C^1$-map $\alpha_i : \bigcup_{t\in\mathcal{I}} \mathrm{supp}(S_t^i) \to \mathbb{R}$ with $\alpha_i\big|_{\mathrm{supp}(S_t^i)} = \alpha_{i|t}\big|_{\mathrm{supp}(S_t^i)}$ for each $t \in \mathcal{I}$, implying that $(h(X_t))_{t\in\mathcal{I}} = \big(P \cdot (\alpha_{\sigma^{-1}(1)} \times \cdots \times \alpha_{\sigma^{-1}(d)})(S_t)\big)_{t\in\mathcal{I}}$ for $P = (\delta_{\sigma(i),j})_{ij} \in \mathrm{P}_d$, as claimed. □

## Appendix B. Proofs and Remarks for Section 8

The following subsections make tacit use of the notation introduced in Appendix C.

### B.1. **Proof of Lemma 8.**

**Lemma 8.** *Let $X$ and $\Theta$ be as described in Assumption 1. Then the following holds:*

(i) *the functions $Q, Q_m : \Theta \to \mathbb{R}$ given in (91) are continuous;*
(ii) *the capped objectives $Q_m$ approximate $Q$ uniformly as $m$ goes to infinity, in symbols:*
$$\lim_{m\to\infty} \|Q - Q_m\|_\Theta = 0 \qquad \textit{for} \qquad \|q\|_\Theta := \sup_{\theta\in\Theta} |q(\theta)|\,;$$
(iii) *if $Q$ is uniquely minimized at $\theta_\star \in \Theta$, i.e. such that $Q(\theta) > Q(\theta_\star)$ if $\theta \neq \theta_\star$, then any ('minimising') sequence $(\theta_m^\star)$ in $\Theta$ such that $Q_m(\theta_m^\star) \leq \inf_{\theta\in\Theta} Q_m(\theta) + \eta_m$ for some $\eta_m \geq 0$ with $\lim_{m\to\infty} \eta_m = 0$ a.s., converges to $\theta_\star$ almost surely as $m \to \infty$.*

*Proof.* *(i)*: For $\nu \geq 2$ fixed, consider the function $q_\nu : \Theta \to V_\nu$ given by

$$q_\nu(\theta) \;:=\; \sum_{\boldsymbol{q}\in\mathfrak{C}_\nu} \kappa_{\boldsymbol{q}}(\theta \cdot X) \cdot \frac{\boldsymbol{q}}{\sqrt{c_{\boldsymbol{q}}}} \tag{140}$$

with $c_{\boldsymbol{q}}$ the number of monomials in $\boldsymbol{q}$ (cf. Remark A.9). As the sets $\{c_{\boldsymbol{q}}^{-1/2} \cdot \boldsymbol{q} \mid \boldsymbol{q} \in \mathfrak{C}_\nu\}$ are each finite and orthonormal w.r.t. the Euclidean structure on $V_\nu$ (cf. (193)), we have that $Q_m = \sum_{\nu=2}^m \|q_\nu\|_\nu^2$ for each $m \geq 2$ and thus obtain the continuity of $Q_m$ from the continuity of (140). To convince ourselves of the latter, fix any index $\boldsymbol{q} \in \mathfrak{C}_\nu$ and let $(\theta_j)_{j\in\mathbb{N}}$ be an arbitrary convergent sequence in $\Theta$, with limit $\lim_{j\to\infty} \theta_j =: \tilde{\theta}$. By a classical interpolation inequality (see e.g. [30, Proposition 5.5. (i)]) we for any $p' > p$ have that

$$\|\tilde{\theta} \cdot X - \theta_j \cdot X\|_{p'\text{-var}} \;\leq\; C \cdot \|\tilde{\theta} \cdot X - \theta_j \cdot X\|_\infty^{1-p/p'} \;\longrightarrow\; 0 \quad \text{a.s.} \quad (\text{as } \; j \to \infty) \tag{141}$$

[53] See, for instance, [51, Corollary 2.8].

for the a.s. finite (by (92)) random variable $C := 2^{1-p/p'}\sup_{j\geq 1}\big[\|\tilde{\theta}\cdot X - \theta_j\cdot X\|_{p\text{-var}}\big]^{p/p'}$, where the convergence in (141) then follows by the compact convergence $\theta_j \to \tilde{\theta}$ and the fact that almost every realisation of $X$ has a compact trace in $D_X$ (Lemma 1 (ii)).

Hence by the $p'$-variation continuity of $\mathfrak{sig}$ (Lemma C.1 (ii)) followed by dominated convergence (cf. (93)) and the fact that $\log_{[\nu]} \equiv \pi_{[\nu]}\circ\log = \log_{[\nu]}\circ\pi_{[\nu]}$ is continuous (Lemma C.1 (iv)), we see that the convergence (141) implies that, for any $\boldsymbol{q}\in\mathfrak{C}_\nu$,

$$\begin{aligned}\kappa_{\boldsymbol{q}}(\theta_j\cdot X) &\overset{\text{def}}{=} \big\langle \log\big[\mathbb{E}[\mathfrak{sig}(\theta_j\cdot X)]\big], \boldsymbol{q}\big\rangle = \big\langle \log_{[\nu]}\big[\mathbb{E}[(\pi_{[\nu]}\circ\mathfrak{sig})(\theta_j\cdot X)]\big], \boldsymbol{q}\big\rangle \\ &\longrightarrow\ \big\langle \log_{[\nu]}\big[\mathbb{E}[(\pi_{[\nu]}\circ\mathfrak{sig})(\tilde{\theta}\cdot X)]\big], \boldsymbol{q}\big\rangle \ =\ \kappa_{\boldsymbol{q}}(\tilde{\theta}\cdot X) \quad \text{as } j\to\infty.\end{aligned} \tag{142}$$

Since our topology on $\Theta$ is metrizable, the sequential convergence (142) characterizes the continuity of $\Theta\ni\theta\mapsto\kappa_{\boldsymbol{q}}(\theta\cdot X)$, which yields that (140) (hence $Q_m$) is continuous as desired.

The continuity of $Q$ thus follows from assertion (ii) of this lemma, i.e. from the claim that

$$Q_m \to Q \quad \text{uniformly on } \Theta \qquad \text{as } m\to\infty. \tag{143}$$

(ii): To see that (143) holds, observe that since $\sup_{\theta\in\Theta}|||\mathfrak{S}(\theta\cdot X)-1|||_\lambda \leq 1$ for some $\lambda>2$ by assumption, Lemma C.1 (iv) yields that the set $\mathcal{L} := \log\big(\{\mathfrak{S}(\theta\cdot X)\mid\theta\in\Theta\}\big) \equiv \{\ell(\theta)\mid\theta\in\Theta\}$ of signature cumulants is $|||\cdot|||_\rho$-bounded for some $\rho>1$. Hence by Lemma C.1 (v),

$$\varsigma_m := \sup_{\boldsymbol{\ell}\in\mathcal{L}}\sum_{\nu>m}\|\pi_\nu(\boldsymbol{\ell})\|_\nu \ \longrightarrow\ 0 \qquad \text{as } m\to\infty. \tag{144}$$

Writing now $\boldsymbol{q} = \boldsymbol{w}_1(\boldsymbol{q}) + \ldots + \boldsymbol{w}_{c_{\boldsymbol{q}}}(\boldsymbol{q})$ for the decomposition of a polynomial $\boldsymbol{q}\in\mathfrak{C}_\nu$ into its (monic) monomials $\boldsymbol{w}_j(\boldsymbol{q})\in[d]^\star_\nu$ (cf. Remark A.9 (b)), we have that for each $\boldsymbol{q}\in\mathfrak{C}_\nu$ the monomials $\boldsymbol{w}_1(\boldsymbol{q}),\ldots,\boldsymbol{w}_{c_{\boldsymbol{q}}}(\boldsymbol{q})$ are pairwise distinct (Rem. A.9 (b)), and further that the union $\bigcup_{\boldsymbol{q}\in\mathfrak{C}_\nu}\{\boldsymbol{w}_1(\boldsymbol{q}),\ldots,\boldsymbol{w}_{c_{\boldsymbol{q}}}(\boldsymbol{q})\}\subset[d]^*_\nu$ is disjoint (Rem. A.9 (a)). Hence and since we have[54]

$$\kappa^2_{\boldsymbol{q}} \overset{\text{def}}{=} \Big(\kappa_{\boldsymbol{w}_1(\boldsymbol{q})} + \ldots + \kappa_{\boldsymbol{w}_{c_{\boldsymbol{q}}}(\boldsymbol{q})}\Big)^2 \ \leq\ c_{\boldsymbol{q}}\cdot\sum_{j=1}^{c_{\boldsymbol{q}}}\kappa^2_{\boldsymbol{w}_j(\boldsymbol{q})} \tag{145}$$

by the Cauchy-Schwarz inequality, we for each $m\geq 2$ obtain the estimate

$$\begin{aligned}\|Q-Q_m\|_\Theta \ &\leq\ \sup_{\theta\in\Theta}\sum_{\nu>m}\sum_{\boldsymbol{q}\in\mathfrak{C}_\nu} c^{-1}_{\boldsymbol{q}}\cdot\kappa_{\boldsymbol{q}}(\theta\cdot X)^2 \\ &\leq\ \sup_{\theta\in\Theta}\sum_{\nu>m}\sum_{\boldsymbol{w}\in[d]^*_\nu}\kappa_{\boldsymbol{w}}(\theta\cdot X)^2 = \sup_{\boldsymbol{\ell}\in\mathcal{L}}\sum_{\nu>m}\|\pi_\nu(\boldsymbol{\ell})\|^2_\nu.\end{aligned} \tag{146}$$

Hence, and since $\lim_{\nu\to\infty}\|\pi_\nu(\boldsymbol{\ell})\|_\nu = 0$ uniformly on $\mathcal{L}$ by (144), there will be an $m_0\geq 2$ such that $\sup_{\boldsymbol{\ell}\in\mathcal{L},\,\nu\geq m_0}\|\pi_\nu(\boldsymbol{\ell})\|_\nu < 1$ and therefore, by (146), $\|Q-Q_m\|_\Theta\leq\varsigma_m$ for all $m\geq m_0$, implying (143) as claimed.

(iii): Let $\theta_\star\in\Theta$ be as above, and $\varepsilon>0$ be arbitrary. Since $\Theta$ is compact so is its closed subset[55] $C_\varepsilon := \Theta\setminus B_\varepsilon(\theta_\star)$, and for $\zeta_\varepsilon := \inf_{\theta\in C_\varepsilon}Q(\theta) = Q(\theta_\epsilon) > Q(\theta_\star)$ (for some $\theta_\varepsilon\in C_\varepsilon$; recall that $Q$ is continuous) and any $\theta\in\Theta$ we have the obvious implication that:

$$\text{if} \quad Q(\theta)\ <\ \zeta_\varepsilon, \qquad \text{then} \quad \theta\ \in\ B_\varepsilon(\theta_\star). \tag{147}$$

Let now $(\theta^\star_m)\subset\Theta$ be a minimising sequence of the required kind. As then $Q(\theta_\star)\leq Q(\theta^\star_m)$ and $Q_m(\theta^\star_m)\leq Q_m(\theta_\star)+\eta_m$ for each $m\geq 2$, we find that $Q(\theta_\star)\leq Q_m(\theta^\star_m)+\big(Q(\theta^\star_m)-Q_m(\theta^\star_m)\big)\leq$

[54] To ease notation, we in (145) drop the argument of the cumulants, i.e. denote $\kappa_q\equiv\kappa_q(\theta\cdot X)$.

[55] The topology (of compact convergence) on $\Theta$ is metrizable (cf. Appendix B.7.1), and $B_\varepsilon(\theta_\star)$ denotes the open ball of radius $\varepsilon$ defined w.r.t. any applicable metric on $\Theta$.

$Q_m(\theta_\star)+\big(Q(\theta_m^\star)-Q_m(\theta_m^\star)+\eta_m\big) = Q(\theta_\star)+r_m$ for $r_m := \big(Q_m(\theta_\star)-Q(\theta_\star)+Q(\theta_m^\star)-Q_m(\theta_m^\star)+\eta_m\big)$. Hence $Q(\theta_\star) \leq Q(\theta_m^\star) \leq Q(\theta_\star) + r_m$ and therefore

$$\lim_{m\to\infty} |Q(\theta_\star) - Q(\theta_m^\star)| \;\leq\; \lim_{m\to\infty} r_m = 0 \qquad \text{(a.s.)}, \tag{148}$$

where the last identity is due to the uniform convergence (ii) (and our assumption on $(\eta_m)$). Consequently $Q(\theta_m^\star) < \zeta_\varepsilon$ for almost all $m$, which in light of (147) implies that $\theta_m^\star \in B_\varepsilon(\theta_\star)$ for almost all $m$ (a.s.), as desired. ☐

B.2. **Linear Interpolation of Discrete-Time Data.** Let $\mathbb{I}$ be a compact interval; say $\mathbb{I} = [0,1]$ wlog. Any dissection $\mathcal{I} \equiv \{t_0, \ldots, t_{n-1} \mid t_0 < \ldots < t_{n-1}\}$ of $\mathbb{I}$ can be uniquely assigned the family of $\mathcal{I}$-*centered hat functions* $\tau_0, \ldots, \tau_{n-1} \in C(\mathbb{I};\mathbb{R})$ characterised by:

$$\tau_j \;\text{ is }\; \mathcal{I}\text{-piecewise affine} \qquad \text{and} \qquad \tau_j(t_\nu) \;=\; \delta_{j\nu} \;\text{ for each } \nu \in [n-1]_0 \tag{149}$$

for all $j \in [n-1]_0$. A path in $\mathcal{C} \equiv C(\mathbb{I};\mathbb{R}^d)$ will be called $\mathcal{I}$-*piecewise linear* if it lies in

$$\mathcal{C}_\mathcal{I} := \big\{v_0 \cdot \tau_0 + \ldots + v_{n-1} \cdot \tau_{n-1} \;\big|\; v_0, \ldots, v_{n-1} \in \mathbb{R}^d\big\} \tag{150}$$

(the 'vectorial span' of (149)). Clearly, the set (150) is a closed linear subspace of $(\mathcal{C}, \|\cdot\|_\infty)$, in fact of $(\mathcal{BV}, |\!|\!|\cdot|\!|\!|_{p\text{-var}})$ (cf. below), and each element $\hat{x} = (\hat{x}_t) \in \mathcal{C}_\mathcal{I}$ is of the form

$$\hat{x}_t = \hat{x}_{t_{j-1}} + \frac{t - t_{j-1}}{t_j - t_{j-1}} \cdot (\hat{x}_{t_j} - \hat{x}_{t_{j-1}}) \quad \text{ for } \quad t \in [t_{j-1}, t_j] \qquad (j \in [n-1]).$$

The space $\mathcal{C}_\mathcal{I}$ is the co-domain of two natural operators, namely the linear projection

$$\hat{\pi}_\mathcal{I} \;:\; \mathcal{C} \twoheadrightarrow \mathcal{C}_\mathcal{I}, \quad \hat{\pi}_\mathcal{I}(x) := x_{t_0} \cdot \tau_0 + \ldots + x_{t_{n-1}} \cdot \tau_{n-1} \equiv \hat{x}_\mathcal{I} \tag{151}$$

as well as the (continuous w.r.t. both $\|\cdot\|_\infty$ and $|\!|\!|\cdot|\!|\!|_p$; set $Z := \mathbb{R}^d$) linear injection

$$\hat{\imath}_\mathcal{I} \;:\; Z^{\times n} \hookrightarrow \mathcal{C}_\mathcal{I}, \quad \hat{\imath}_\mathcal{I}(v_0, \ldots, v_{n-1}) := v_0 \cdot \tau_0 + \ldots + v_{n-1} \cdot \tau_{n-1}. \tag{152}$$

It is clear that the linear operator $\hat{\pi}_\mathcal{I}$ is bounded on $\mathcal{C}$ with operator norm $\|\hat{\pi}_\mathcal{I}\| = 1$.

*Remark* B.1. (i) As any two points in $\mathbb{R}^d$ uniquely determine the affine path-segment that joins them, the $\mathcal{I}$-piecewise linear projection $\hat{x}_\mathcal{I}$ of a path $x$ can be seen as the *unbiased continuous-time approximation* of $x$ given the observations $(x_t \mid t \in \mathcal{I})$.[56]

(ii) For any $(z_j)_{j\in[n]} \in Z^{\times n}$ and any $\mathbb{I}$-dissection of cardinality $|\mathcal{I}| = n$,

$$\big\|\hat{\imath}_\mathcal{I}(z_1, \ldots, z_n)\big\|_{1\text{-var}} = \sum_{j=0}^{n-1} |z_{j+1} - z_j| \;\leq\; 2\|(z_1, \ldots, z_n)\|_1.$$

Denote by $\mathcal{BV}_p := \{x \in \mathcal{C} \mid (196) \text{ is finite}\}$ the space of all continuous paths of bounded $p$-variation ($p \geq 1$), and remark that each $\mathcal{BV}_p$ is a Banach space w.r.t. the norm $|\!|\!|x|\!|\!|_{p\text{-var}} := \|x\|_{p\text{-var}} + \|x(0)\|$ (e.g. [30, Thm. 5.25 (i)]).

**Lemma B.1.** *For $(\mathcal{I}_n)_{n\in\mathbb{N}}$ a refined sequence of dissections of a compact interval $\mathbb{I}$,*

$$\lim_{n\to\infty} \hat{\pi}_{\mathcal{I}_n} = \mathrm{id}_\mathcal{C} \qquad \textit{pointwise on} \;\; \mathcal{C}(\mathbb{I};\mathbb{R}^d) \tag{153}$$

*where for each argument the above convergence is understood to take place in $(\mathcal{C}, \|\cdot\|_\infty)$. In addition, the family of operators $(\hat{\pi}_{\mathcal{I}_n} \mid n \in \mathbb{N})$ is equicontinuous, whence in particular the convergence (153) is uniform on compact subsets of $\mathcal{C}$. The family of operators $(\hat{\pi}_{\mathcal{I}_n} \mid n \in \mathbb{N})$ remains equicontinuous if $(\mathcal{C}, \|\cdot\|_\infty)$ is replaced by $(\mathcal{BV}_p, |\!|\!|\cdot|\!|\!|_{p\text{-var}})$ for any $p \geq 1$.*

[56] Likewise, the injection (152) can be seen as the 'unbiased $\mathcal{I}$-centered *continuous-time localisation*' of a sequence $(v_1, \ldots, v_n) \in Z^{\times n}$.

*Proof.* The pointwise convergence (153) is an easy consequence of definition (151) and the fact that every element of $\mathcal{C}$ is uniformly continuous on $\mathbb{I}$. The equicontinuity of the family of linear operators $\big(\hat{\pi}_{\mathcal{I}_n} \mid n \in \mathbb{N}\big)$ is immediate by the fact that this family is uniformly bounded (by 1) in the operator norm. That a pointwise convergent sequence of equicontinuous functions on a metric space (with values in a complete metric space) converges uniformly on compact subsets of its domain is a well-known fact from real analysis (e.g. [73, Exercise 7.16]). As shown in [30, Prop. 5.20], the operator family $(\hat{\pi}_{\mathcal{I}} : \mathcal{BV}_p \to \mathcal{BV}_p \mid n \in \mathbb{N})$ remains uniformly bounded in the operator norm (and hence is equicontinuous) if the latter is defined w.r.t. the $p$-variation norm $\vert\!\vert\!\vert\cdot\vert\!\vert\!\vert_{p\text{-var}}$ on the Banach space $\mathcal{BV}_p$. ☐

## B.3. **Proof of Lemma 9.**

**Lemma 9** (Interpolation Limit)**.** *Let $\Theta$ and $X$ be as in Assumption 1, $\widehat{Q}_m$ as in (96) and $Q_m$ as in (91). Then for $(\mathcal{I}_n)_{n\in\mathbb{N}}$ any refined sequence of dissections of $\mathbb{I}$ and any $m \in \mathbb{N}_{\geq 2}$,*

$$Q_m(\theta) = \lim_{n\to\infty} \widehat{Q}_m(\hat{X}^\theta_{\mathcal{I}_n}) \quad \textit{uniformly on} \;\; \Theta. \tag{97}$$

*Proof.* Recalling that $\operatorname{tr}(X) \subseteq D_X$ with probability one (Lemma 1 (ii)), notice that

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \left\| \hat{X}^\theta_{\mathcal{I}_n} - \theta \cdot X \right\|_{\tilde{p}\text{-var}} = 0 \qquad \text{almost surely} \tag{154}$$

for any $\tilde{p} > p$ with $p \geq 1$ as in (92). Indeed: Denoting by $x := (X_t(\omega))_{t\in\mathbb{I}} \subseteq D_X$ (inclusion with probab. one) a given realisation of $X$, we remark first that (cf. Section B.2 for notation)

$$\Theta_x := \left\{ x^\theta \equiv (\theta \cdot x_t)_{t\in\mathbb{I}} \;\middle|\; \theta \in \Theta \right\} \quad \text{is a compact subset of} \quad (\mathcal{BV}_{\tilde{p}}, \vert\!\vert\!\vert\cdot\vert\!\vert\!\vert_{\tilde{p}\text{-var}}) \tag{155}$$

for any $\tilde{p} > p$. To see that (155) holds, observe that [30, Lemma 5.27 (i)] (together with (92)) implies that, for any $\tilde{p} > p$, each of the functions

$$\alpha_n, \alpha \;:\; \Theta \to \mathcal{BV}_{\tilde{p}}, \quad \alpha_n(\theta) := \hat{\pi}_{\mathcal{I}_n}(\theta \cdot x) \;\text{ and }\; \alpha(\theta) := \theta \cdot x \qquad (n \in \mathbb{N}) \tag{156}$$

are continuous. In particular, $\Theta_x = \alpha(\Theta)$ is compact (as continuous image of a compact set).

In addition, [30, Lemma 5.27 (i)] (by virtue of Lemma B.1 (153) and [30, Proposition 5.20 (5.13)]) implies that $\lim_{n\to\infty} \alpha_n = \alpha$ pointwise on $\Theta$. Hence by (155) and the last assertion of Lemma B.1 (which implies that $(\hat{\pi}_{\mathcal{I}_n})$ converges uniformly on $\Theta_x$; see the proof of Lemma B.1 for details), we obtain that $\lim_{n\to\infty} \alpha_n = \alpha$ uniformly on $\Theta$, which in turn yields (154) by the fact that $\hat{X}^\theta_{\mathcal{I}_n} = \hat{\pi}_{\mathcal{I}_n}(\theta \cdot X)$ for each $\theta \in \Theta$.

Given (154) for any fixed $\tilde{p} > p$, the $\tilde{p}$-variation continuity of $\mathfrak{sig}$ (Lemma C.1 (ii)) together with the equicontinuity of $(\hat{\pi}_{\mathcal{I}_n} : \mathcal{BV}_{\tilde{p}} \to \mathcal{BV}_{\tilde{p}} \mid n \in \mathbb{N})$ (Lemma B.1) yields that

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \left\| \mathfrak{sig}_m(\hat{X}^\theta_{\mathcal{I}_n}) - \mathfrak{sig}_m(\theta \cdot X) \right\|_m = 0 \quad \text{almost surely} \qquad (m \in \mathbb{N}).$$

Indeed, the above holds path-wise, with probability one, by Lemma B.2 (applied to $\Theta$ as above, $B = \mathcal{BV}_{\tilde{p}}$, $V = V_{[m]}$, $\Psi = \mathfrak{sig}_m$, $\alpha$ and $\alpha_n$ as in (156) and $\tau = \alpha$).

Thus for $\mathfrak{S}_{m|n}(\theta) := \mathbb{E}\big[\mathfrak{sig}_m(\hat{X}^\theta_{\mathcal{I}_n})\big]$ and $\mathfrak{S}_m(\theta) := \mathbb{E}\big[\mathfrak{sig}_m(\theta \cdot X)\big]$ we have that

$$\lim_{n\to\infty} \mathfrak{S}_{m|n}(\theta) = \mathfrak{S}_m(\theta) \qquad \text{uniformly on} \;\; \Theta \tag{157}$$

due to [32, Theorem 22 (p. 241)] (note that the hypothesis in loc.cit. of $(\mathfrak{S}_{m|n})_n$ to be "absolutely continuous uniformly" is met in light of [32, Thm. 11 (p. 192)] and assumption (93)).

Finally, the fact that $\log_{[m]} \equiv \pi_{[m]} \circ \log$ is continuous (Lemma C.1 (iv)) together with the uniform convergence (157) of $\mathfrak{S}_{[m]|n} := \sum_{\nu=0}^m \mathfrak{S}_{\nu|n}$ towards $\mathfrak{S}_{[m]} := \sum_{\nu=0}^m \mathfrak{S}_\nu$, yields that

$$\kappa_n^{[m]} := \log_{[m]} \circ \mathfrak{S}_{[m]|n} \quad \overset{n\to\infty}{\longrightarrow} \quad \log_{[m]} \circ \mathfrak{S}_{[m]} =: \kappa^{[m]} \qquad \text{uniformly on} \;\; \Theta.$$

In particular, $\kappa_{\boldsymbol{q}}(\hat{X}^{\theta}_{\mathcal{I}_n}) = \langle\kappa^{[m]}_n(\theta), \boldsymbol{q}\rangle \to \langle\kappa^{[m]}(\theta), \boldsymbol{q}\rangle = \kappa_{\boldsymbol{q}}(\theta\cdot X)$ uniformly on $\Theta$ for each $\boldsymbol{q}\in V_{[m]}$, which by definitions (96) (of $\widehat{Q}_m$) and (91) (of $Q_m$) yields (97) as desired. □

**Lemma B.2.** *Let $\Theta$ be a compact metric space, $B$ and $V$ be Banach spaces, $\Psi : B \to V$ be a continuous map, and $\alpha$, $\alpha_n$, $\tau : \Theta \to B$, $n\in\mathbb{N}$, be continuous functions such that*

$$\alpha_n = p_n\circ\tau, \quad n\in\mathbb{N}, \quad \textit{with} \quad \big(p_n : \tau(\Theta)\to B \;\big|\; n\in\mathbb{N}\big) \;\; \textit{equicontinuous}$$

*and $\lim_{n\to\infty}\alpha_n = \alpha$ pointwise on $\Theta$. Then $\lim_{n\to\infty}\Psi\circ\alpha_n = \Psi\circ\alpha$ uniformly on $\Theta$.*

*Proof.* Let $(\theta_n)_{n\in\mathbb{N}}$ be any convergent sequence in $\Theta$, say $\theta_n\to\theta$ for some $\theta\in\Theta$. Then

$$\lim_{n\to\infty}\Phi_n(\theta_n) = \Phi(\theta) \qquad \text{for} \qquad \Phi_n := \Psi\circ\alpha_n \;\; \text{and} \;\; \Phi := \Psi\circ\alpha, \tag{158}$$

since $\|\Phi(\theta)-\Phi_n(\theta_n)\|_V \le \|\Phi(\theta)-\Phi_n(\theta)\|_V + \|\Phi_n(\theta)-\Phi_n(\theta_n)\|_V$ with $\lim_{n\to\infty}\|\Phi(\theta)-\Phi_n(\theta)\|_V = 0$ and $\lim_{n\to\infty}\|\Phi_n(\theta)-\Phi_n(\theta_n)\|_V = 0$. For the latter convergence, take any $\varepsilon>0$ and let $\delta_1>0$ be such that $\sup_{b\in B_{\delta_1}(\alpha(\theta))}\|\Psi(\alpha(\theta))-\Psi(b)\|_V\le\varepsilon$, and $\delta_2>0$ be such that $\rho_n := \sup_{b\in B_{\delta_2}(\tau(\theta))\cap\tau(\Theta)}\|p_n(\tau(\theta))-p_n(b)\|_B\le\delta_1$ for all $n\in\mathbb{N}$. Taking $n_0\ge 1$ such that $\sup_{n\ge n_0}\|\tau(\theta)-\tau(\theta_n)\|_B\le\delta_2$ then implies that

$$\sup_{n\ge n_0}\|\alpha_n(\theta)-\alpha_n(\theta_n)\|_B = \sup_{n\ge n_0}\|p_n(\tau(\theta))-p_n(\tau(\theta_n))\|_B \;\le\; \sup_{n\ge n_0}\rho_n\le\delta_1$$

and therefore $\sup_{n\ge n_0}\|\Phi_n(\theta)-\Phi_n(\theta_n)\|_V\le\varepsilon$, as required.

Conclude by observing that (158) implies $\Phi_n\to\Phi$ uniformly on $\Theta$, as desired.

Indeed, assume otherwise that $\Phi_n\nrightarrow\Phi$ uniformly, i.e. that there is $\tilde{\varepsilon}>0$ such that

$$\forall\, k\in\mathbb{N} \;:\; \exists\, n_k\in\mathbb{N}_{\ge n} \;\; \text{with} \;\; \sup_{\theta\in\Theta}\|\Phi(\theta)-\Phi_{n_k}(\theta)\|_V > \tilde{\varepsilon}. \tag{159}$$

Then (159) informs the choice of a subsequence $(\theta_{n_k})_k\subseteq\Theta$, with $(n_k)_k\subseteq\mathbb{N}$ increasing, s.t.

$$\|\Phi(\theta_{n_k})-\Phi_{n_k}(\theta_{n_k})\|_V \;>\; \tilde{\varepsilon} \quad \text{for each} \;\; k\in\mathbb{N}. \tag{160}$$

As $\Theta$ is compact, we may assume, by passing to a further subsequence if necessary, that this subsequence converges, say to $\tilde{\theta}\in\Theta$. The continuity of $\Phi$ then implies $\lim_{k\to\infty}\Phi(\theta_{n_k}) = \Phi(\tilde{\theta})$, while property (158) combined with a doubling argument (as in [68, Sect. 3.5*: remark on p. 98]) yields $\lim_{k\to\infty}\Phi_{n_k}(\theta_{n_k}) = \Phi(\tilde{\theta})$. Hence $\lim_{k\to\infty}\|\Phi(\theta_{n_k})-\Phi_{n_k}(\theta_{n_k})\|_V = 0$, in contradiction to (160). □

## B.4. **Proof of Proposition 6.**

**Proposition 6.** *Let $\Theta\subseteq C(D_{X_*};\mathbb{R}^d)$ and $X_* = (X_j)_{j\in\mathbb{N}}$ be a discrete time-series in $\mathbb{R}^d$ with compact spatial support and such that for each $\theta\in\Theta$ the expectations*

$$\mathbb{E}[\mathfrak{sig}_m(\hat{X}^\theta_1)] \;\; \textit{exist for all } m\ge 1, \;\; \textit{with } \hat{X}^\theta_1 \textit{ the interpolant of } \theta\cdot X_1,\ldots,\theta\cdot X_n.$$

*It then holds that: if $X_*$ is [weakly] signature ergodic to length $n$, then $X_*$ is [weakly] signature ergodic to length $n$ on $\Theta$.*

*Proof.* For $\tilde{m}\ge 1$ and $\theta\in\Theta$ and $w\in V_{\tilde{m}}$ all arbitrary but fixed, let $\phi=\phi_{\tilde{m}}$ be as in (99) and set $\xi := \langle\phi\circ\theta^{\times n}, w\rangle$. Set further $\hat{\xi}_T(z) := \frac{1}{T}\sum_{j=1}^{T}\xi(z_{n(j-1)+1},\ldots,z_{nj})$ for any given sequence $z=(z_\nu)_{\nu\in\mathbb{N}}$ in $\mathbb{R}^d$. The lemma then asserts that, under the given integrability and ergodicity conditions,

$$\mathbb{E}\big[\xi(X_1,\ldots,X_n)\big] \;=\; \lim_{T\to\infty}\hat{\xi}_T(X_*) \quad \text{a.s.} \quad [\text{resp.}^{57} \text{ in probab.}]. \tag{161}$$

To see that (161) holds, note first that for $\hat{X}_1 := \hat{\imath}_{\mathcal{E}_n}(X_1,\ldots,X_n)$ (cf. Def. 10 and (152)),

$$\xi(X_{[n]}) \,=\, \varphi(\hat{X}_1) \qquad \text{for} \qquad \varphi(x) := \Big\langle \mathfrak{sig}_{[\tilde{m}]}\big(\tilde{\pi}_{\mathcal{E}_n}(\tilde{\theta}\cdot x)\big),\, w\Big\rangle \tag{162}$$

where $\tilde{\theta}$ is any fixed continuous extension of $\theta$ to $\widehat{D} := \operatorname{conv}(D_{X_*})$, the convex hull of $D_{X_*}$. (Recall that such a $\tilde{\theta}$ exists by Tietze's extension theorem.) Since the function $\varphi : \widehat{\mathcal{BV}}_n \to \mathbb{R}$ defined by (162) on the compact[58] subset

$$\widehat{\mathcal{BV}}_n \,:=\, \Big\{ x \in \mathcal{C}_{\mathcal{E}_n} \,\Big|\, x_t \in \widehat{D} \;\text{ for each }\; t \in \mathcal{E}_n \Big\} \,\subset\, \mathcal{BV} \qquad \text{(cf. (150))}$$

is continuous (by [30, Prop. 5.20] and Lemma C.1 (ii)), the universality property of the signature (e.g. Lemma C.1 (vii)) implies that there is a sequence $(\boldsymbol{\ell}_j)_{j\in\mathbb{N}}$ in $V^\circ$ such that

$$\varphi = \lim_{j\to\infty} \langle \mathfrak{sig}(\cdot),\, \boldsymbol{\ell}_j\rangle \qquad \text{in} \quad \big(C(\widehat{\mathcal{BV}}_n),\, \|\cdot\|_\infty\big). \tag{163}$$

For $\big(\hat{\mathbb{E}}_T^{(m)}(X_*)\big)_{T\in\mathbb{N}} := \big(\frac{1}{T}\sum_{\nu=1}^T \mathfrak{sig}_{[m]}(\hat{X}_\nu)\big)_{T\in\mathbb{N}}$ with $\hat{X}_\nu := \hat{\imath}_{\mathcal{E}_n}(X_{n(\nu-1)+1},\ldots,X_{n\nu})$, our assumption on $X_*$ gives that, for each $m\in\mathbb{N}$,

$$\mathbb{E}\big[\mathfrak{sig}_{[m]}(\hat{X}_1)\big] = \lim_{T\to\infty} \hat{\mathbb{E}}_T^{(m)}(X_*) \quad \text{a.s.} \quad \text{in } \operatorname{conv}(\mathfrak{sig}_{[m]}(\widehat{\mathcal{BV}}_n)). \tag{164}$$

Hence upon combining (162) and (163), and using that dominated convergence applies as both sides of (163) are bounded (cf. Lemma C.1 (ii)), we find that with probability one,

$$\begin{aligned} \mathbb{E}\big[\xi(X_1,\ldots,X_n)\big] &= \lim_{j\to\infty} \big\langle \mathbb{E}\big[\mathfrak{sig}(\hat{X}_1)\big],\, \boldsymbol{\ell}_j\big\rangle = \lim_{j\to\infty}\lim_{T\to\infty} \big\langle \hat{\mathbb{E}}_T^{(d_{\ell_j})}(X_*),\, \boldsymbol{\ell}_j\big\rangle \\ &= \lim_{T\to\infty}\lim_{j\to\infty} \frac{1}{T}\sum_{\nu=1}^T \big\langle \mathfrak{sig}_{[d_{\boldsymbol{\ell}_j}]}(\hat{X}_\nu),\, \boldsymbol{\ell}_j\big\rangle \\ &= \lim_{T\to\infty} \frac{1}{T}\sum_{\nu=1}^T \lim_{j\to\infty} \big\langle \mathfrak{sig}(\hat{X}_\nu),\, \boldsymbol{\ell}_j\big\rangle \overset{(163)}{=} \lim_{T\to\infty}\frac{1}{T}\sum_{\nu=1}^T \xi(\hat{X}_\nu), \end{aligned} \tag{165}$$

where we denoted $d_{\boldsymbol{\ell}}$ for the degree of the index-polynomial $\boldsymbol{\ell}\in V^\circ$. Notice that the interchange of limits in the second line of (165) is permissible as the convergence in (163) is uniform (see, e.g., [72, Theorem 7.11]). This shows the almost-sure case of (161).

To prove that (161) holds in probability if (164) holds in probability for each $m\in\mathbb{N}$ (which is true by assumption if $X_*$ is weakly signature ergodic), we resort to a subsequence argument, recalling that (as the topology of weak convergence is metrizable) a sequence converges in probability iff each of its subsequences admits yet another subsequence that converges almost surely. To this end, abbreviate $\mu_{m,T} := \hat{\mathbb{E}}_T^{(m)}(X_*)$ and assume that

$$\mu_m := \mathbb{E}[\mathfrak{sig}_{[m]}(\hat{X}_1)] \,=\, \lim_{T\to\infty} \mu_{m,T} \quad \text{in probability} \qquad \text{for each} \quad m\in\mathbb{N}. \tag{166}$$

Then for any fixed subsequence $(T_k)_{k\in\mathbb{N}} \subset \mathbb{N}$, there is a subsequence $T_k^{(1)} < T_{k+1}^{(1)}$ of $(T_k)$ such that $\lim_{k\to\infty}\mu_{1,T_k^{(1)}} = \mu_1$ almost surely. But since, by (166), $\lim_{k\to\infty}\mu_{2,T_k^{(1)}} = \mu_2$ in probability, there will be a subsequence $T_k^{(2)} < T_{k+1}^{(2)}$ of $(T_k^{(1)})$ such that $\lim_{k\to\infty}\mu_{2,T_k^{(2)}} = \mu_2$ almost surely (thus $\lim_{k\to\infty}\mu_{1,T_k^{(2)}} = \mu_1$ a.s. in particular). Iterating this procedure, Cantor's diagonal trick (e.g. [67, Proof of Theorem I.24]) thus allows for the choice of a subsequence $T_k^{(\infty)} < T_{k+1}^{(\infty)}$ of $(T_k)$ such that $\lim_{k\to\infty}\mu_{m,T_k^{(\infty)}} = \mu_m$ almost surely for each $m\in\mathbb{N}$.

---

[57] For simplicity of exposition, we present the case of almost sure convergence first and give the changes necessary for the case of convergence in probability at the end of this proof.

[58] By [52, Prop. 1.7] and the facts that: (a) the convex hull operator on $\mathbb{R}^d$ preserves compactness, and (b) the Cartesian product of compact sets is compact (noting that $\widehat{\mathcal{BV}}_n \cong \widehat{D}^{\times n}$).

Repeating the above calculation (165) then shows that the subsequence $(\hat{\xi}_{T_k^{(\infty)}}(X_*))_{k\in\mathbb{N}}$ of $(\hat{\xi}_{T_k}(X_*))_{k\in\mathbb{N}}$ converges almost surely to $\mathbb{E}[\xi(X_{[n]})]$, as desired. □

B.5. **Proof of Lemma 10.**

**Lemma 10** (Ergodicity Limit)**.** *Let $X_* = (X_j)_{j\in\mathbb{N}}$ be a discrete time-series which for some $m$ is [weakly] $m^{\text{th}}$-order signature ergodic to some length $n \in \mathbb{N}$ on $\Theta$, and denote*

$$\mathfrak{K}^{m|n|T}(\theta) := \log_{[m]}(\hat{\mathfrak{S}}_T^{m|n}(\theta)) \quad \text{for} \quad \hat{\mathfrak{S}}_T^{m|n}(\theta) := \frac{1}{T}\sum_{j=1}^{T}\mathfrak{sig}_{[m]}(\hat{X}_j^\theta),$$

$$\text{and}^{59} \quad \bar{\mathfrak{K}}_{\boldsymbol{i}}^{m|n|T}(\theta) := \frac{\mathfrak{K}_{\boldsymbol{i}}^{m|n|T}(\theta)}{\left(\mathfrak{K}_{11}^{m|n|T}(\theta)\right)^{\eta_1(\boldsymbol{i})/2}\cdot\ldots\cdot\left(\mathfrak{K}_{dd}^{m|n|T}(\theta)\right)^{\eta_d(\boldsymbol{i})/2}}, \quad \boldsymbol{i} \in [d]^\star,$$

*for each $\theta \in \Theta$, where $\hat{X}_j^\theta$ is the interpolant of $\theta\cdot X_{n(j-1)+1},\ldots,\theta\cdot X_{nj}$ along $\mathcal{I}_{n|j}$. For any $m \geq 2, n, T \in \mathbb{N}$ and $\theta\in\Theta$, denote further (recalling Notation 7.1)*

$$\hat{\kappa}_T^{m|n}(\theta) := \sum_{\nu=2}^{m}\sum_{\boldsymbol{q}\in\mathfrak{C}_\nu}\bar{\mathfrak{K}}_{\boldsymbol{q}}^{m|n|T}(\theta)^2. \tag{101}$$

*Provided that $\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|\mathfrak{sig}_k(\hat{X}_1^\theta)\right\|_k\right] < \infty$ for each $k\in[m]$, it then holds that*

$$\bar{\kappa}_{\text{IC}}^{[m]}(\hat{X}_1^\theta) = \lim_{T\to\infty}\hat{\kappa}_T^{m|n}(\theta) \quad \textit{uniformly on } \Theta \quad \textit{a.s. [in probability]}. \tag{102}$$

*Proof.* Let $Z := \mathbb{R}^d$ and $\mathcal{E}_n$ be as in Def. 10, and for every $\theta\in\Theta$ denote

$$\xi_\theta := \mathfrak{sig}_{[m]} \circ \hat{\iota}_{\mathcal{E}_n} \circ \theta^{\times n} \ : \ Z^{\times n} \longrightarrow V_{[m]} \cap V_{(1)}. \tag{167}$$

The parametrisation-invariance of $\mathfrak{sig}$ (Lemma C.1 (iii)) gives that

$$\hat{\mathfrak{S}}_T^{m|n}(\theta) = \frac{1}{T}\sum_{j=1}^{T}\xi_\theta(\bar{X}_j) =: \hat{\mathbb{E}}_T[\xi_\theta(X_*)] \quad \text{for} \quad \bar{X}_j := \left(X_{n(j-1)+1},\ldots,X_{nj}\right),$$

and the continuity of $\log_{[m]}$ (Lemma C.1 (iv)) yields that (102) follows from the convergence

$$\lim_{T\to\infty}\sup_{\theta\in\Theta}\left\|\mathbb{E}[\xi_\theta(\bar{X}_1)] - \hat{\mathbb{E}}_T[\xi_\theta(X_*)]\right\|_{[m]} = 0 \qquad \text{a.s.} \quad \text{[in probab.]} \tag{168}$$

for the norm $\|\cdot\|_{[m]} := \sum_{\nu=1}^{m}\|\cdot\|_\nu$, followed by an application of the continuous mapping theorem (e.g. [80, Theorem 2.3]). As (168) is equivalent to the coordinatewise convergences

$$\lim_{T\to\infty}\sup_{w\in[d]_k^*}\sup_{\theta\in\Theta}\left|\mathbb{E}[\langle\xi_\theta(\bar{X}_1), w\rangle] - \langle\hat{\mathbb{E}}_T[\xi_\theta(X_*)], w\rangle\right| = 0 \qquad \text{for} \ \ k\in[m] \tag{169}$$

almost surely (resp. in prob.), we can see that (169) holds by fixing any $w\in[d]_k^*$ and showing

$$\lim_{T\to\infty}\sup_{\theta\in\Theta}\left|\mathbb{E}\big[\tilde{\xi}_\theta(\bar{X}_1)\big] - \hat{\mathbb{E}}_T\big[\tilde{\xi}_\theta(X_*)\big]\right| = 0 \quad \text{[a.s./in prob.]} \quad \text{for} \quad \tilde{\xi}_\theta := \langle\xi_\theta, w\rangle \tag{170}$$

and $\hat{\mathbb{E}}_T[\tilde{\xi}_\theta(X_*)] := \langle\hat{\mathbb{E}}_T[\xi_\theta(X_*)], w\rangle = T^{-1}\sum_{j=1}^{T}\tilde{\xi}_\theta(\bar{X}_j)$. To this end, note that the function

$$\tilde{\xi} \ : \ Z^{\times n}\times\Theta \longrightarrow \mathbb{R}, \quad (z,\theta) \mapsto \tilde{\xi}_\theta(z),$$

is continuous in $\theta$ for every $z\in Z^{\times n}$, as is seen directly from (167) (recalling the continuity of $\hat{\iota}_{\mathcal{E}_n} : Z^{\times n} \to \mathcal{BV}$ (Rem. B.1 (ii)) and Lemma C.1 (ii)). Also, by assumption, $\Theta$ is compact

[59] Where $\eta_\nu(\boldsymbol{i})$ denotes the number of times the index-entry $\nu$ appears in $\boldsymbol{i}$; cf. (79).

with $\mathbb{E}[\sup_{\theta\in\Theta}|\tilde{\xi}_\theta(\bar{X}_1)|] < \infty$ and $\hat{\mathbb{E}}_T[\xi_\theta(X_*)] \to \mathbb{E}[\xi_\theta(\bar{X}_1)]$ pointwise, which altogether implies that $\mathcal{F} := \{\tilde{\xi}(\cdot, \theta) \mid \theta \in \Theta\}$ is Glivenko-Cantelli via [82, Lem. 6.1, Thm. 6.1], i.e. that

$$\lim_{T\to\infty} \sup_{\varphi\in\mathcal{F}} \left| \mathbb{E}[\varphi(\bar{X}_1)] - \frac{1}{T}\sum_{j=1}^{T} \varphi(\bar{X}_j) \right| = 0, \tag{171}$$

where the mode of the convergence in (171) (almost surely or in probability) coincides with the mode of the pointwise convergence $\hat{\mathbb{E}}_T[\xi_\theta(X_*)] \to \mathbb{E}[\xi_\theta(\bar{X}_1)]$ on $\Theta$ (cf. [82, (Proof of) Theorem 6.1]). As (171) is identical to (170), the proof is finished. ☐

B.6. **Ergodicity Conditions and Proof of Lemma 11.** Let $X_* \equiv (X_j)_{j\in\mathbb{N}}$ be a sequence of $\mathbb{R}^d$-valued random variables.

*Definition* B.1. The sequence $X_*$ is called *$\alpha$-mixing* if for the sub-$\sigma$-algebras $\mathcal{X}_k^\ell := \sigma\big(X_\nu \mid k \leq \nu \leq \ell\big)$ it holds that $\lim_{\nu\to\infty} \alpha_\nu(X_*) = 0$ for the sequence

$$\alpha_\nu(X_*) := \sup_{A\in\mathscr{X}_1^j,\, B\in\mathscr{X}_{j+\nu}^\infty,\, j\in\mathbb{N}} \big|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)\big|,$$

and $X_*$ is called *$\phi$-mixing* if it holds that $\lim_{\nu\to\infty} \phi_\nu(X_*) = 0$ for the sequence

$$\phi_\nu(X_*) := \sup_{A\in\mathscr{X}_1^j,\, B\in\mathscr{X}_{j+\nu}^\infty,\, \mathbb{P}(A)>0,\, j\in\mathbb{N}} \big|\mathbb{P}(B\,|\,A) - \mathbb{P}(B)\big|.$$

Note that $\phi$-mixing implies $\alpha$-mixing, and see e.g. [9] for further information.

*Definition* B.2. The sequence $X_*$ will be said to have *$n$-seasonal increments*, $n \in \mathbb{N}$, if the sequence $\Delta(X_*) := (X_{j+1} - X_j)_{j\in\mathbb{N}}$ of its increments is suff. integrable and such that

$$\Delta(X_*)_{[n]} = \Delta(X_*)_{(n(j-1)+1,\, nj]} \qquad \text{for each } j \in \mathbb{N}.$$

**Lemma 11.** *For $X_* \equiv (X_j)_{j\in\mathbb{N}}$ uniformly integrable and $n \in \mathbb{N}$, the following holds.*

(i) *If $X_*$ is $\alpha$-mixing and has $(m, n)$-stationary sigmoments ($m \in \mathbb{N}$), then $X_*$ is $m^{\text{th}}$-order weakly signature-ergodic to length $n$;*

(ii) *if $X_*$ is $\phi$-mixing with $\sum_{\nu=1}^\infty \phi_{1+(\nu-1)n}^{1/2}(X_*)\frac{\log\nu}{\nu} < \infty$ and has $n$-seasonal increments, then $X_*$ is signature-ergodic to length $n$.*

*(i) and (ii) persist if $X_*$ is replaced by $\theta \cdot X_* = (\theta \cdot X_j)_{j\in\mathbb{N}}$ for any measurable $\theta : D_{X_*} \to \mathbb{R}^d$.*

*Proof.* Starting from definition (187), a direct calculation yields that for any $\ell_1 < \ell_2$,

$$\mathfrak{sig}_m(\hat{\iota}_{\mathcal{E}}(X_{\ell_1}, \ldots, X_{\ell_2})) = \sum_{(i_1,\ldots,i_m)\in(\ell_1:\,\ell_2]^{\times m}} c_{i_1\cdots i_m} \cdot \Delta_{i_1} \otimes \cdots \otimes \Delta_{i_m} \tag{172}$$

for certain $c_{i_1\cdots i_m} \in \mathbb{R}$ and increments $\Delta_j := X_j - X_{j-1}$, where $\mathcal{E} \equiv \mathcal{E}_{\ell_1,\ell_2}$ is the equidistant (or any other) $\mathbb{I}$-dissection of cardinality $\ell_2 - \ell_1 + 1$. Let now $n \in \mathbb{N}$ be fixed. If we introduce the shift-map $\vartheta(i) := i + n$ (with $\vartheta^0 :=$ id and $\vartheta^j := \vartheta \circ \vartheta^{j-1}$) for convenience and denote

$$Y_j := \mathfrak{sig}_m(\hat{\iota}_{\mathcal{E}}(\theta \cdot X_{n(j-1)+1}, \ldots, \theta \cdot X_{nj})) \qquad (j \in \mathbb{N})$$

for brevity, then the above shows that each $Y_j$ is a measurable function of the arguments $X_{\vartheta^{j-1}(1)}, \ldots, X_{\vartheta^{j-1}(n)}$. This in turn implies the inclusion of $\sigma$-algebras

$$\mathscr{Y}_p^q := \sigma(Y_p, \ldots, Y_q) \quad\subseteq\quad \sigma(X_{\vartheta^{p-1}(1)}, \ldots, X_{\vartheta^{p-1}(n)}, \ldots, X_{\vartheta^{q-1}(1)}, \ldots, X_{\vartheta^{q-1}(n)})$$

for any $p \leq q$, whence in particular $\mathscr{Y}_1^j \subseteq \mathscr{X}_1^{\vartheta^{j-1}(n)}$ and $\mathscr{Y}_{j+\nu}^\infty \subseteq \mathscr{X}_{\vartheta^{j+\nu-1}(1)}^\infty$ for all $\nu \in \mathbb{N}$. Since $\vartheta^{j-1}(n) = jn$ and $\vartheta^{j+\nu-1}(1) = jn + \vartheta^{\nu-1}(1)$, we can use Definition B.1 to for $Y_* := (Y_j)_{j\in\mathbb{N}}$ and $\gamma \in \{\alpha, \phi\}$ conclude that

$$\gamma_\nu(Y_*) \;\leq\; \gamma_{\vartheta^{\nu-1}(1)}(X_*) = \gamma_{1+(\nu-1)n}(X_*) \qquad \text{for each } \nu \in \mathbb{N},$$

which shows that if $X_*$ is $\alpha$-mixing ($\phi$-mixing) then so is $Y_*$.

The proof of statement (i) is finished by a coordinatewise application of the weak law of large numbers for non-stationary $\alpha$-mixing time series given in [81, Theorem 7.15].

As to (ii), we note similarly that if $X_*$ has $n$-seasonal increments and is $\phi$-mixing at the assumed rate, then $Y_*$ is stationary (by (172)) and $\phi$-mixing with

$$\sum_{\nu=1}^{\infty}\phi_\nu^{1/2}(Y_*)\frac{\log\nu}{\nu} \;\le\; \sum_{\nu=1}^{\infty}\phi_{1+(\nu-1)n}^{1/2}(X_*)\frac{\log\nu}{\nu} \;<\; \infty\,,$$

whence assertion (ii) follows from a coordinatewise application of [50, Corollary 1].

This proof of the statements (i) and (ii) goes through without changes if the sequence $(X_j)_{j\in\mathbb{N}}$ is replaced by $(\theta\cdot X_j)_{j\in\mathbb{N}}$ for any (Borel-)measurable map $\theta : D_{X_*}\to\mathbb{R}^d$. □

B.7. **Complementary Remarks and Proofs for Theorem 5.** Throughout this subsection, the setting and notation from the proof of Theorem 5 (pp. 33) applies.

B.7.1. *The Compact-Open Topology on $\Theta$ is Metrizable.* Since $D_X$ is a closed subset of $\mathbb{R}^d$, there are $\{K_\nu\}\subseteq D_X$ compact with $K_\nu\subseteq K_{\nu+1}$ and $D_X=\bigcup_{\nu\in\mathbb{N}_0}K_\nu$, and the topology of compact convergence on $C(D_X;\mathbb{R}^d)$ coincides with the compact-open topology on $C(D_X;\mathbb{R}^d)$, e.g. [60, Theorem 46.8]. Defining $\|\theta\|_K:=\sup_{u\in K}|\theta(u)|$, this topology is induced by the metric (see, e.g., [20, Proposition VII.1.6])

$$\tilde d(\theta,\tilde\theta)\;:=\;\sum_{\nu=0}^{\infty}2^{-\nu}\rho_\nu(\theta,\tilde\theta)\quad\text{with}\quad \rho_\nu(\theta,\tilde\theta):=\frac{\|\theta-\tilde\theta\|_{K_\nu}}{1+\|\theta-\tilde\theta\|_{K_\nu}}\,;\tag{173}$$

we choose $K_\nu:=\overline{B_{r_\nu}(0)}$ for any $r_\nu\uparrow\infty$ monotonously with $r_0:=0$ for convenience.

Note that the metrics (on $C(K_\nu;\mathbb{R}^d)$) $\rho_\nu$ and $d_\nu(\theta,\tilde\theta):=\|\theta-\tilde\theta\|_{K_\nu}$ are equivalent for all $\nu\in\mathbb{N}_0$. Specifically, for each $\nu\in\mathbb{N}_0$ we have $d_\nu(\theta,\tilde\theta)\le d_{\nu+1}(\theta,\tilde\theta)$ for any $\theta,\tilde\theta\in\Theta$, and

$$d_\nu(\theta,\tilde\theta)\;\le\;2\rho_\nu(\theta,\tilde\theta)\qquad\text{if}\quad\rho_\nu(\theta,\tilde\theta)\le\tfrac12.\tag{174}$$

For $\eta=d_\nu,\rho_\nu,\tilde d$, denote $B_r^\eta(\theta_*):=\{\theta\in C(D_X;\mathbb{R}^d)\mid\eta(\theta,\theta_*)<r\}$ and $\tilde B_r(\theta_*):=B_r^{\tilde d}(\theta_*)$.

Below are the proofs of Theorem 5 for the cases $(X,\mathcal{J})$ ergodic resp. weakly ergodic.

B.7.2. *Proof of Theorem 5 for Ergodic Observations.* Let $(X,\mathcal{J})$ be an ergodic observation such that $D_X$ is not necessarily compact. In this case, Theorem 5 asserts that each $\varepsilon>0$ comes with a $\mathbb{P}$-full set $\tilde\Omega_\varepsilon\in\mathscr{F}$ such that for each $\omega\in\tilde\Omega_\varepsilon$ the following holds:

$$\begin{gathered}\exists\,m_0\equiv m_0(\omega)\ge2\;:\;\forall\,m\ge m_0\;\;\text{there is}\;\;k_0\equiv k_0(m)\in\mathbb{N}\;\;\text{s.t.}\;\;\forall\,k\ge k_0:\\ \lim_{\tau\to\infty}\max\left\{\sup_{T\ge\tau}\Big[\operatorname{dist}_{\|\cdot\|_\infty}\big(\hat\theta_T^\star\cdot X(\omega),\,\mathrm{DP}_d\cdot S(\omega)\big)\Big],\,\varepsilon\right\}=\varepsilon\\ \text{for any}\;\;(\theta_T^\star)_{T\in\mathbb{N}}\equiv(\theta_T^\star(m,k,\omega))_{T\in\mathbb{N}}\equiv(\theta_T^{m|k}(\omega))_{T\in\mathbb{N}}\subset\Theta\;\;\text{as in (105)}.\end{gathered}\tag{175}$$

The above proof of (107) which did not involve any compactness assumption on $D_X$, remains valid without any changes, so that (175) holds if it can be derived from (107).

To this end, let $\Omega''\in\mathscr{F}$ be the $\mathbb{P}$-full set on which $\Omega''\in\mathscr{F}$ be the $\mathbb{P}$-full set on which the traces of $X$ are all contained in $D_X$ and (82) holds, and for each $n\in\mathbb{N}$ denote by $\Omega_n$ the $\mathbb{P}$-full set on which (107) holds for $\tilde\varepsilon=\frac1n$. Set $\tilde\Omega:=\Omega''\cap\bigcap_{n\in\mathbb{N}}\Omega_n$ (another $\mathbb{P}$-full set) and let $\varepsilon>0$ be arbitrary. Take any $\omega\in\tilde\Omega$. Then $\operatorname{tr}(X(\omega))\subset K_{\nu_0}$ for some $\nu_0\in\mathbb{N}$, whence for any

$n_0 \in \mathbb{N}$ with $n_0^{-1} \leq 2^{-\nu_0}\varepsilon/4$ we have that

$$\exists\, m_0 \equiv m_0(n_0) \geq 2 \;:\; \forall\, m \geq m_0 \;\text{ there is }\; k_0 \equiv k_0(m) \in \mathbb{N} \;\text{ s.t. }\; \forall\, k \geq k_0 :$$
$$\alpha^{m|k}(\omega) \;\leq\; n_0^{-1} \quad \text{and hence} \quad \sup_{T\geq\tau_0} \left\|\theta_T^{m|k}(\omega)\cdot X(\omega) - \theta_\star \cdot X(\omega)\right\|_\infty \;\leq\; \varepsilon$$

(for some $\tau_0\,(\equiv \tau_0(\omega)) \in \mathbb{N}$) by the exact same argumentation that led us to (109). □

B.7.3. *Proof of Theorem 5 for Weakly Ergodic Observations.* Let $(X, \mathcal{J})$ be a weakly ergodic observation. Adopting the setting and notation from pp. 33, suppose now that

$$\forall\,\tilde{\varepsilon} > 0 \;:\; \exists\, m_0 \geq 2 \;:\; \text{for each } m \geq m_0 \text{ there is } k_0 \equiv k_0(m) \text{ such that}: \tag{176}$$
$$\lim_{\tau\to\infty} \alpha_\tau^{m|k} \vee \tilde{\varepsilon} = \tilde{\varepsilon} \quad \text{in probability,}^{60} \quad \text{for each } k \geq k_0.$$

Spelled out, (176) implies that for any given $(\varepsilon', \delta') \in (0,\infty)^2$ and $(m,k)\,(\equiv (m,k)_{\tilde{\varepsilon}}$ as in (176) for $\tilde{\varepsilon} := \varepsilon'/2)$

$$\text{there is} \quad \tau_* \equiv \tau_*(\varepsilon', \delta') \in \mathbb{N} \quad \text{such that} \quad \sup_{\tau\geq\tau_*} \mathbb{P}(\alpha_\tau^{m|k} \geq \varepsilon') \;\leq\; \delta'. \tag{177}$$

(Indeed: for any $m, k$ as in (176) with $\tilde{\varepsilon} := \varepsilon'/2$, it holds $\mathbb{P}(\alpha_\tau^{m|k} \geq \varepsilon') \leq \mathbb{P}((\alpha_\tau^{m|k} \vee \frac{\varepsilon'}{2}) \geq \varepsilon') = \mathbb{P}(|(\alpha_\tau^{m|k} \vee \frac{\varepsilon'}{2}) - \frac{\varepsilon'}{2}| \geq \frac{\varepsilon'}{2}) \to 0$ as $\tau \to \infty$.) In particular, for any given $\mathfrak{p} \equiv (\varepsilon, \delta) \in (0,\infty)^2$ there will be $m_{\mathfrak{p}} \in \mathbb{N}$ such that for every $m' \geq m_{\mathfrak{p}}$ there is $k_{\mathfrak{p}} \equiv k_{\mathfrak{p}}(m')$ with the property that: for any $k' \geq k_{\mathfrak{p}}$ there is $\tau'_{\mathfrak{p}} \equiv \tau'_{\mathfrak{p}}(k') \in \mathbb{N}$ with

$$\varrho_{\mathfrak{p}} \;:=\; \sup_{\tau\geq\tau'_{\mathfrak{p}}} \mathbb{P}(\sup_{T\geq\tau} \|\theta_T^{m'|k'} \cdot X - \theta_\star \cdot X\|_\infty \geq \varepsilon) \;\leq\; \delta, \tag{178}$$

which due to $\sup_{\tau\geq\tau_{\mathfrak{p}}'} \mathbb{P}\big(\sup_{T\geq\tau} \operatorname{dist}_{\|\cdot\|_\infty}(\theta_T^{m'|k'}\cdot X, \mathrm{DP}_d\cdot S) \geq \varepsilon\big) \leq \varrho_{\mathfrak{p}}$ implies that the asserted convergence (106) holds in probability. To see that (178) holds, fix $\varepsilon, \delta > 0$ and note

$$\big\{\sup_{T\geq\tau} \|\theta_T \cdot X - \theta_\star \cdot X\|_\infty \geq \varepsilon\big\} \cap \Omega' \quad\subseteq\quad \bigcup\nolimits_{\nu\in\mathbb{N}} A_\nu^{\hat{\theta}_\tau} \cap B_\nu \tag{179}$$

for any given sequence $\hat{\theta} \equiv (\theta_T)$ of $\Theta$-valued random variables, $\tau \in \mathbb{N}$, and for the events[61] $A_\nu^{\hat{\theta}_\tau} := \big\{\sup_{T\geq\tau} d_\nu(\theta_T, \theta_\star) \geq \varepsilon\big\}$ and $B_\nu := \{\sup_{t\in\mathbb{I}} |X_t| \geq r_{\nu-1}\}$, where $r_\nu$ denotes the radius of the 0-centered closed ball $K_\nu$. Noting that $A_\nu^{\hat{\theta}_\tau} \subseteq A_{\nu+1}^{\hat{\theta}_\tau}$ and $B_{\nu+1} \subseteq B_\nu$ for all $\nu \in \mathbb{N}$, we from (179) obtain that

$$\mathbb{P}(\sup_{T\geq\tau} \|\theta_T \cdot X - \theta_\star \cdot X\|_\infty \geq \varepsilon) \;\leq\; \mathbb{P}(A_{\nu_0}^{\hat{\theta}_\tau}) \;+\; \mathbb{P}(B_{\nu_0+1}) \tag{180}$$

for any fixed $\nu_0 \in \mathbb{N}$. Denoting $\mu_X := \mathbb{E}[\sup_{t\in\mathbb{I}} |X_t|]$, Markov's inequality implies that

$$\mathbb{P}(B_{\nu_0+1}) \;\leq\; \frac{\mu_X}{r_{\nu_0}} \quad\longrightarrow\quad 0 \qquad (\nu_0 \to \infty), \tag{181}$$

while (174) implies $B_\varepsilon^{d_\nu}(\theta_\star) \supseteq B_{\varepsilon/2}^{\rho_\nu}(\theta_\star)$ (if $\varepsilon < 1$, assumable wlog) and hence yields

$$\mathbb{P}(A_{\nu_0}^{\hat{\theta}_\tau}) \;\leq\; \mathbb{P}(\sup_{T\geq\tau} \rho_{\nu_0}(\theta_T, \theta_\star) \geq \varepsilon/2) \;\leq\; \mathbb{P}(\sup_{T\geq\tau} \tilde{d}(\theta_T, \theta_\star) \geq 2^{-\nu_0}\varepsilon/2). \tag{182}$$

Given (181) and (182), we may now fix an $\nu_0 \in \mathbb{N}$ large enough such that $\mathbb{P}(B_{\nu_0+1}) \leq \delta/2$, and for this choice of $\nu_0$ obtain an $m_{\mathfrak{p}} \in \mathbb{N}$, as guaranteed by (176) for $\tilde{\varepsilon} = \tilde{\varepsilon}_\star$ with $\tilde{\varepsilon}_\star := 2^{-\nu_0}\varepsilon/4$, such that for every $m \geq m_{\mathfrak{p}}$ there is $k_{\mathfrak{p}}\,(\equiv k_{\mathfrak{p}}(m))$ with the property that: for any $k \geq k_{\mathfrak{p}}$ there is $\tau_* \equiv \tau_*(2\tilde{\varepsilon}_\star, \delta/2) \in \mathbb{N}$, as guaranteed by (177), such that

$$\sup_{\tau\geq\tau_*} \mathbb{P}(A_{\nu_0}^{\hat{\theta}_\tau}) \overset{(182)}{\leq} \sup_{\tau\geq\tau_*} \mathbb{P}(\alpha_\tau^{m|k} \geq \tilde{\varepsilon}_\star) \;\leq\; \delta/2 \qquad \text{for } \;\hat{\theta} = (\theta_T^{m|k}).$$

60 Remark that the (usual) notion of convergence in probability is well-defined for $\Theta$-valued random variables since the topology of compact convergence on $\Theta$ is metrizable, second-countable (e.g. [57]) and, hence, separable.

61 As $X$ has continuous realisations, we have $\sup_{t\in\mathbb{I}} |X_t| = \sup_{t\in\mathbb{I}\cap\mathbb{Q}} |X_t|$ so that $B_\nu^\theta$ is measurable.

Taken altogether, the estimate (180) then allows us to conclude that

$$\sup_{\tau\geq\tau_*}\mathbb{P}(\sup_{T\geq\tau}\|\theta_T^{m|k}\cdot X-\theta_\star\cdot X\|_\infty\geq\varepsilon)\ \leq\ \delta/2+\delta/2\ \leq\ \delta,$$

which (via Thm. 4) yields the conclusion (106) of Theorem 5 for the weakly ergodic case.

It hence remains to prove (176), for which we may follow the previous lines of pp. 34 with only slight adaptations. Indeed: Since in the weakly ergodic case the Θ-uniform estimator convergence (102) holds in probability, we – by way of the very same argumentation as for (113) – obtain that

$$\lim_{T\to\infty}\bar{\kappa}_{m,k}(\theta_T^\star)\ =\ 0\quad\text{in probability,}\quad\text{with}\ \ (\theta_T^\star)\equiv(\theta_T^{m|k}) \tag{183}$$

as in (105) for $m,k$ as in (110) for some (arbitrary but) fixed $\tilde{\varepsilon}>0$. From this we obtain that in the present context, the convergence (112) holds in probability. Indeed: Assuming otherwise implies the existence of $\varepsilon_0,\delta_0>0$ such that

$$\mathbb{P}(\operatorname{dist}(\theta_{T_j}^\star,\mathcal{M})\geq\varepsilon_0)\ \geq\ \delta_0\qquad\text{for each } j\in\mathbb{N}, \tag{184}$$

for some sequence $(T_j)_{j\in\mathbb{N}}\subset\mathbb{N}$. As a subsequence of $(\bar{\kappa}_{m,k}(\theta_T^\star))_{T\in\mathbb{N}}$, we by way of (183) find that $(\bar{\kappa}_{m,k}(\theta_{T_j}^\star))_{j\in\mathbb{N}}$ converges to 0 in probability, whence there is yet another subsequence $(T_{j_\ell})_{\ell\in\mathbb{N}}$ of $(T_j)_{j\in\mathbb{N}}$ such that $\lim_{\ell\to\infty}\bar{\kappa}_{m,k}(\theta_{T_{j_\ell}}^\star)=0$ almost surely. Applying the (essentially) same argument which brought '(113) ⇒ (112)' now yields that $\lim_{\ell\to\infty}\operatorname{dist}(\theta_{T_{j_\ell}}^\star,\mathcal{M})=0$ almost surely and hence in probability, contradicting (184).

As this proves $\lim_{\tau\to\infty}\sup_{T\geq\tau}\operatorname{dist}(\theta_T^{m|k},\mathcal{M})=0$ in probability, we for any $\epsilon>0$ obtain

$$\mathbb{P}\big(|\alpha_\tau^{m|k}\vee\tilde{\varepsilon}-\tilde{\varepsilon}|\geq\epsilon\big)\ \leq\ \mathbb{P}\big(\alpha_\tau^{m|k}\geq\tilde{\varepsilon}\big)\ \leq\ \mathbb{P}\big(\sup_{T\geq\tau}\operatorname{dist}(\theta_T^{m|k},\mathcal{M})\geq\tilde{\varepsilon}/2\big)\ \to\ 0$$

as $\tau\to\infty$, where the last inequality is due to (111). This shows (176) as required. □

## Appendix C. A 'Moment-Like' Coordinate Description for the Law of Stochastic Processes

C.1. **The Expected Signature: A Coordinate Vector for Stochastic Processes.** Many results in statistics, including Corollary 1 via (5), are based on the well-known fact that the distribution of a random vector $Z=(Z^1,\cdots,Z^d)$ in $\mathbb{R}^d$ can be characterised by a set of coordinates with respect to a *basis of nonlinear functionals on* $\mathbb{R}^d$. More specifically, any such vector $Z$ can be assigned its *moment coordinates* $(\mathfrak{m}_{\boldsymbol{i}}(Z))_{\boldsymbol{i}\in[d]^\star}\subset\bar{\mathbb{R}}$ defined by

$$\mathfrak{m}_{i_1\cdots i_m}(Z)\ :=\ \mathbb{E}\left[Z^{i_1}\cdots Z^{i_m}\right]=\int_{\mathbb{R}^d}x_{i_1}\cdots x_{i_m}\,\mathbb{P}_Z(\mathrm{d}x). \tag{185}$$

As the linear span of the monomials $\{x_{\boldsymbol{i}}\equiv x_{i_1}\cdots x_{i_m}\mid\boldsymbol{i}\equiv(i_1,\ldots,i_m)\in[d]^\star\}$ is uniformly dense in the spaces of continuous functions over compact subsets of $\mathbb{R}^d$, the coordinatisation (185) is *faithful* in the sense that, under certain conditions [49], the (coefficients of) the moment vector $\mathfrak{m}(Z)\equiv(m_{\boldsymbol{i}}(Z))_{\boldsymbol{i}\in[d]^\star}$ determine the distribution of $Z$ uniquely.

Now, if instead of a random *vector in* $\mathbb{R}^d$ one seeks to find a convenient coordinatisation for the distribution of a stochastic process $Y$ in $\mathbb{R}^d$, i.e. a random *path in* $\mathcal{C}_d$, then one can – perhaps surprisingly – resort to a natural generalisation of (185), which is known as the *expected signature of* $Y$: Analogous to how the monomials $\{x_{\boldsymbol{i}}\mid\boldsymbol{i}\in[d]^\star\}$ are a basis[62] of nonlinear functionals on $\mathbb{R}^d$ that provides coordinates $(\mathfrak{m}_{i_1\cdots i_m}\mid(185))\subset\bar{\mathbb{R}}$ for a random vector in $\mathbb{R}^d$, there is a basis $\{\chi_{i_1\cdots i_m}\mid(i_1,\ldots,i_m)\in[d]^\star\}$ of nonlinear functionals on (regular enough subspaces of) $\mathcal{C}_d$ which provides coordinates $(\sigma_{\boldsymbol{i}})_{\boldsymbol{i}\in[d]^\star}\subset\bar{\mathbb{R}}$ for (the law of) a random path $Y$ in $\mathcal{C}^d$. This path-space basis is defined as follows:

[62] Cf. the trivial fact the monomials $x_1=\langle\cdot\,,e_1\rangle,\ldots,x_d=\langle\cdot\,,e_d\rangle$ determine each vector in $\mathbb{R}^d$ uniquely.

Given a path $x = (x_t^1, \cdots, x_t^d)_{t\in\mathbb{I}} \in \mathcal{C}_d$ of bounded variation in $\mathbb{R}^d$ (assuming $\mathbb{I} = [0,1]$ wlog), consider the *noncommutative moments of* $x$, that is the iterated Stieltjes-integrals

$$\chi_{i_1\cdots i_m}(x) := \int_{0\leq t_1\leq t_2\leq\cdots\leq t_m\leq 1} \mathrm{d}x_{t_1}^{i_1}\mathrm{d}x_{t_2}^{i_2}\cdots\mathrm{d}x_{t_m}^{i_m}, \qquad (i_1,\ldots,i_m)\in[d]^\star, \tag{186}$$

of $(x^{i_1}, \cdots, x^{i_m}) \in \mathcal{C}_m$ over the standard $m$-simplex $\{(t_1,\ldots,t_m)\in\mathbb{I}^m \mid t_1\leq\cdots\leq t_m\}$ (for $\epsilon$ the empty index in $[d]^\star$, we set $\chi_\epsilon \equiv 1$).[63] Then, the nonlinear functionals $x \mapsto \chi_{i_1\cdots i_m}(x)$ define *a dual basis for the vector* $x \in \mathcal{C}_d$, in the sense that the coefficients $(\chi_{i_1\cdots i_m}(x) \mid (186))$ *determine the path* $x$ *uniquely*; see [13, 29, 34].[64] The resulting family of coordinates

$$\mathfrak{sig}(x) := \big(\chi_{i_1\cdots i_m}(x) \;\big|\; (i_1,\ldots,i_d)\in[d]^\star\big) \tag{187}$$

is known as the *signature* of the path $x$.

*Remark* C.1. Similarly still to the monomial dual basis $\{x_{i_1}\cdots x_{i_m}\}$ on $\mathbb{R}^d$, the linear span of the above functionals $\{\chi_{i_1\cdots i_m} \mid (186)\}$ is closed under pointwise multiplication and hence forms an algebra over the space of applicable paths in $\mathcal{C}_d$, from which one obtains that their linear span is uniformly dense in the space of continuous functions over (certain) compact subsets of $\mathcal{C}_d$ (Stone-Weierstrass), see e.g. [52, Thm. 2.15] (and Lemma C.1 (vii)).

As a consequence of Remark C.1,[65] one can infer in analogy to (185) that the dual coefficients

$$\begin{aligned}\sigma_{i_1\cdots i_m}(Y) &:= \int_{\mathcal{C}_d}\chi_{i_1\cdots i_m}(x)\,\mathbb{P}_Y(\mathrm{d}x) \qquad\qquad (i_1,\ldots,i_m\in[d],\ m\geq 0)\\ &= \mathbb{E}\Big[\int_{0\leq t_1\leq t_2\leq\cdots\leq t_m\leq 1}\mathrm{d}Y_{t_1}^{i_1}\mathrm{d}Y_{t_2}^{i_2}\cdots\mathrm{d}Y_{t_m}^{i_m}\Big]\end{aligned} \tag{188}$$

define a *complete set of coordinates* for the distribution of a stochastic process $Y = (Y_t^1, \cdots, Y_t^d)_{t\in\mathbb{I}}$ in $\mathbb{R}^d$ that has compact support (and sample paths of bounded variation).

The *signature-based coordinatisation* (188) of a random path in $\mathcal{C}_d$ can thus be regarded as a natural generalisation of the moment-based coordinatisation (185) of a random vector in $\mathbb{R}^d$.

The assumption of compact support is of course much too restrictive on a non-locally compact space like $\mathcal{C}_d$, but under additional decay conditions [17] or by using a normalization [16, Theorem 5.6] it can be shown that the coefficients $(\sigma_{i_1\cdots i_k}(Y) \mid (188))$ indeed characterize the distribution of $Y$ uniquely even if the compactness assumption is dropped. Extending the definition of (187) to paths less regular ('rougher') than of bounded variation is at the centre of the Theory of Rough Paths ([30, 52, 53]).

The first application of the coordinates (188) in statistics was given in [62] for SDE parameter estimation, with more recent applications including the development of non-commutative cumulants [7] and Hurst parameter estimation [24].

[63] For $x$ defined on a general compact interval $\mathbb{I}\subset\mathbb{R}$, set $\chi_{i_1\cdots i_m}(x) := |\mathbb{I}|^{-m}\int_{\Delta_m(\mathbb{I})}\mathrm{d}x_{t_1}^{i_1}\mathrm{d}x_{t_2}^{i_2}\cdots\mathrm{d}x_{t_m}^{i_m}$, with the $m$-simplex $\Delta_m(\mathbb{I})$ over $\mathbb{I}$ defined as above.

[64] Up to a negligible indeterminacy known as 'tree-like equivalence', see [34].

[65] Recall that by Riesz representation theorem, a (signed) Borel measure on a compact metric space $K$ acts as a continuous linear functional over the space $C(K)$ of continuous functions on $K$ and is hence uniquely determined by its (dual) functional action on a dense subset of $C(K)$.

C.2. **A Coordinate Space for the Laws of Stochastic Processes.** In order to make the information provided by (186) and (188) amenable to mathematical analysis, it will be convenient to regard $\mathfrak{sig}(x)$ and $(\sigma(Y)_{\boldsymbol{i}} \mid \boldsymbol{i} \in [d]^\star)$ as elements of a suitable topological space.

To this end, we denote by $[d]^*$ the *free monoid* on the alphabet $[d] = \{1, \ldots, d\}$, and identify each multiindex $(i_1, \cdots, i_m) \in [d]^\star$ in (77) with the *word* $\mathtt{i}_1 \cdots \mathtt{i}_m \in [d]^*$ it defines.

From this view, both $\mathfrak{sig}(x)$ and $\mathfrak{S}(Y) \equiv (\sigma_{\boldsymbol{i}})_{\boldsymbol{i}\in[d]^\star}$ can then be treated as *formal power series in the variables* $\{1, \ldots, d\}$, i.e. as elements of the *free algebra*

$$\mathbb{R}[d]^* := \{\boldsymbol{t} : [d]^* \to \mathbb{R} \mid \boldsymbol{t} \text{ is a map}\} \equiv \left\{ \sum\nolimits_{w\in[d]^*} \boldsymbol{t}(w) \cdot w \;\middle|\; \mathtt{t} \in \mathbb{R}[d]^* \right\}; \tag{189}$$

indeed: $\mathfrak{S}(Y) \cong \sum_{w\in[d]^\star} \boldsymbol{t}_\sigma(w) \cdot w \in \mathbb{R}[d]^*$ with $\boldsymbol{t}_\sigma(\mathtt{i}_1 \cdots \mathtt{i}_m) := \sigma_{i_1\cdots i_m}$. For convenience, we may henceforth write $\boldsymbol{t}(w) =: \langle \boldsymbol{t}, w\rangle$ $(w \in [d]^\star)$ for a given $\boldsymbol{t} \in \mathbb{R}[d]^\star$.

The space $\mathbb{R}[d]^*$ thus serves as a graded *coordinate space* for (the laws of applicable) continuous stochastic processes in $\mathbb{R}^d$.

*Remark* C.2*.* The coordinate space $\mathbb{R}[d]^*$ is not just an $\mathbb{R}$-vector space but a twofold *bialgebra* (in fact: a *Hopf algebra*), namely w.r.t. the two multiplications given by (a) the concatenation product $*$ (the $\mathbb{R}$-bilinear extension of the word-concatenation on $[d]^*$), and (b) the shuffle product $\sqcup\!\sqcup$ from (80), see [69, pp. 29 and 31] for details. The bi-algebra structures associated to these two products are an algebraic reflection of the duality between (185) and (186), cf. also (192).

C.2.1. *The Log Transform.* Accordingly, the expected signature $\mathfrak{S}(Y)$ of $Y$ can be seen as *a coordinate vector* of $Y$ *w.r.t. the monomial standard basis* $\mathfrak{B} := \{\mathtt{i}_1 \cdots \mathtt{i}_k \mid i_1, \ldots, i_k \in [d],\ k \geq 0\}$ of $\mathbb{R}[d]^*$. The vector $\mathfrak{S}(Y)$ itself, however, is contained in a *nonlinear* subspace of $\mathbb{R}[d]^\star$; more specifically, $\mathfrak{S}(Y)$ is 'close to an exponential'.[66]

It is hence reasonable to expect *a more parsimonious* coordinatisation of $Y$ w.r.t. $\mathfrak{B}$ to be achieved by, instead of the vector $\mathfrak{S}(Y)$, considering the $\mathfrak{B}$-coordinates of the faithful *linearisation* $\Phi(\mathfrak{S}(Y))$ that is effected by the log-transform $\Phi(\boldsymbol{t}) \equiv \log(\boldsymbol{t})$ defined by

$$\log(\boldsymbol{t}) := \sum_{m\geq 1} \frac{(-1)^{m-1}}{m} (\boldsymbol{t} - \epsilon)^{*m} \tag{190}$$

for $\boldsymbol{t} \in \mathbb{R}[d]^*$ with $\langle \boldsymbol{t}, \epsilon\rangle = 1$; this linearised coordinate description is accounted for by Def. 9.

*Remark* C.3 (Signature Cumulants Generalise Classical Cumulants)*.* In the same way that the expected signature generalises the classical concept of moments, the signature cumulant generalises the classical concept of cumulants from vector-valued to path-valued random variables, cf. [7]: The (classical) cumulants of a random vector $Z$ in $\mathbb{R}^d$ read[67]

$$\kappa_{i_1\cdots i_m}(Z) \;=\; \langle \pi_{\mathrm{Sym}}(\log[\mathfrak{m}(Z)]),\, \mathtt{i}_1 \cdots \mathtt{i}_m\rangle \tag{191}$$

and hence are identical to the signature cumulant of the linear process $Y := (Z \cdot t)_{t\in[0,1]}$. Given this relation between (191) and (78), Proposition 4 appears as a natural generalisation of the well-known fact that the (classical) cumulant relations

$$\kappa_{\tilde{w}}(Z) = 0 \quad \text{for all} \quad \tilde{w} \in \bigsqcup_{k=2}^{d} \left\{ \tilde{u} * \tilde{v} \;\mid\; \tilde{u} \in [k-1]^* \setminus \{\epsilon\},\ \tilde{v} \in \{k\}^* \setminus \{\epsilon\} \right\} \tag{192}$$

[66] Algebraically, $\mathfrak{S}(Y)$ lies in the convex hull of the Lie-group $\{\sum_{w\in[d]^\star} \chi_w(x) \cdot w \mid x \in \mathcal{BV}\} \subset \mathbb{R}[d]^*$, where $\mathcal{BV} := \{x \in \mathcal{C}_d \mid \|x\|_{1\text{-var}} < \infty\}$ and $\chi_w : \mathcal{BV} \to \mathbb{R}$ are the functionals in (186) (e.g. [69, Cor. 3.5]).

[67] For $\pi_{\mathrm{Sym}}(\mathtt{i}_1 \cdots \mathtt{i}_m) := \sum_{\tau\in S_m} \mathtt{i}_{\tau(1)} \cdots \mathtt{i}_{\tau(m)}$ the projection onto the ($[d]$-adic closure of) the subspace spanned by all symmetric polynomials in $\mathbb{R}[d]^*$.

are characteristic of a random vector $Z$ in $\mathbb{R}^d$ to have mutually independent components.

C.2.2. *The Coordinate Space and Its Topology.* As of yet, the coordinate space (189) provides only a 'purely algebraic container' for the coordinate tuples (187) and (188). Statistical analysis, however, typically concerns convergence and thus requires a topology.

A convenient such topology on $\mathbb{R}[d]^*$ can be defined by

$$\text{identifying} \quad \mathbb{R}[d]^* \quad \text{with the tensor algebra} \quad V^\infty := \prod_{m=0}^{\infty} V_m \tag{193}$$

where $V_0 := \mathbb{R}$ and $V_m := V_1^{\otimes m}$ for $V_1 \equiv (\mathbb{R}^d, |\cdot|_2)$, via $[d]^* \ni \mathtt{i}_1 \cdots \mathtt{i}_m \leftrightarrow e_{i_1} \otimes \cdots \otimes e_{i_m} \in V_m$ and $\epsilon \leftrightarrow 1 \in \mathbb{R}$ (with $(e_i)_{i\in[d]}$ the standard basis in $V_1$); in other words, we identify the free algebra (189) with the Cartesian product $V^\infty$ which we then endow with its natural tensor algebra structure with 1 (cf. e.g. [52, Sect. 2.2.1, Rem. 1.24 f.] for details). Denote by $\|\cdot\|_m$ the Euclidean (i.e., $|\cdot|_2$-induced) tensor norm on $V_m$, and write $\pi_m : V^\infty \to V_m$ ($\hookrightarrow V$), $\pi_m((v_j)_{j\geq 0}) = v_m$, for the canonical projection of $V^\infty$ onto its $m^{\text{th}}$ factor. Let further $V_{[m]} := \prod_{\nu=0}^{m} V_\nu$ be the truncated tensor algebra, and $\pi_{[m]} := \sum_{\nu=0}^{m} \pi_\nu$ the truncation map.

*Remark* C.4 (Truncation). Notice that $V_{[m]}$ comes equipped with a natural algebra structure, namely the one realised as the quotient of $V^\infty$ by the ideal $\prod_{\nu>m} V_\nu$; the map $\pi_{[m]}$ is then the canonical quotient epimorphism. Note in particular that

$$\pi_{[m]}\big(\log(\boldsymbol{t})\big) = \sum_{k=1}^{m} \frac{(-1)^{k-1}}{k}\big[\pi_{[m]}(\boldsymbol{t}-\epsilon)\big]^{\underline{*}k} =: \log_{[m]}(\boldsymbol{t})\,, \tag{194}$$

defining a (bijective) polynomial map $\log_{[m]} : V_{(1)} \to V_{[m]}$ ($\hookrightarrow V^\infty$; the space $V_{[m]}$ is embedded as a (closed) linear subspace of $V^\infty$ but not as a subalgebra). In the above, $\underline{*}$ denotes the multiplication in the algebra $V_{[m]}$, i.e.: $\pi_{[m]}(\boldsymbol{t}_1) \,\underline{*}\, \pi_{[m]}(\boldsymbol{t}_2) \overset{\text{def}}{=} \pi_{[m]}(\boldsymbol{t}_1 * \boldsymbol{t}_2)$, $\forall\, \boldsymbol{t}_1, \boldsymbol{t}_2 \in V^\infty$.

Our *topological coordinate space* (for (random) paths and their laws) is

$$V := \Big\{\boldsymbol{t} \in V^\infty \;\Big|\; |\!|\!|\boldsymbol{t}|\!|\!|_\lambda := \textstyle\sum_{m\geq 0} \|\pi_m(\boldsymbol{t})\|_m \cdot \lambda^m \,<\, \infty,\ \forall\, \lambda > 0\Big\} \tag{195}$$

equipped with the locally convex topology induced by the (fundamental) family of norms $(|\!|\!|\cdot|\!|\!|_\lambda \mid \lambda > 0)$; cf. [17, Section 2], where the locally $m$-convex algebra (195) was first introduced to the analysis of signatures and their expectation. (Note that the subspace topology on $V_m \subset V$ coincides with the (Euclidean) topology on $(V_m, \|\cdot\|_m)$.) The factorial decay

$$\big|\chi_{i_1\cdots i_m}(x)\big| \;\lesssim\; \|x\|_{1\text{-var}}^m / m!$$

of the functionals (186) implies that $\mathfrak{sig}(x), \mathfrak{S}(Y) \in V$, cf. also Lemma C.1 below.

For convenience, we also introduce the dilation maps

$$\delta_\lambda \;:\; V \to V, \quad (v_m)_{m\geq 0} \;\mapsto\; (\lambda^m \cdot v_m)_{m\geq 0}, \qquad (\lambda > 0)$$

as well as the subspaces $V_{(c)} := \{\boldsymbol{t} \in V \mid \pi_0(\boldsymbol{t}) = c\}$ and $V_{(c)}^\infty$ (defined analogously), and recall that the space $\mathcal{BV} := \mathcal{C}_d \cap \mathrm{BV}$ of continuous $\mathbb{R}^d$-valued paths of bounded variation can be endowed with the *p-variation topology* (any $p \geq 1$) defined via the $p$-variation seminorm

$$\|x\|_{p\text{-var}} := \left[\sup_{\mathcal{D}} \sum_{(t_k)\in\mathcal{D}} \big|x_{t_k} - x_{t_{k-1}}\big|^p\right]^{1/p} \tag{196}$$

where the sup is taken over the set $\mathcal{D}$ of all dissections of $[0,1]$; e.g. [52, Sect. 1.2] for details.

The next lemma collects basic facts on $V$, (187) and (190) that are useful for Section 8.

**Lemma C.1.** *Let $V$ and $\mathfrak{sig}$ and $\log$ be as above, and $\rho > 1$. Then the following holds:*

(i) *the space $V$ is a separable and metrizable Hausdorff space;*

(ii) *the signature transform $x \mapsto \mathfrak{sig}(x)$ defines a map $\mathfrak{sig} : \mathcal{BV} \to V$ which for any $p \geq 1$ is continuous w.r.t. the $p$-variation topology on $\mathcal{BV}$;*

(iii) *the signature is invariant under order-preserving time-domain reparametrisations of its arguments, i.e. $\mathfrak{sig}(x) = \mathfrak{sig}(x_\varphi)$ for $x_\varphi \equiv (x_{\varphi(t)})_{t\in\mathbb{J}}$ with $\varphi \in C(\mathbb{J};\mathbb{I})$ strictly monotone;*

(iv) *the capped logarithm $\log_{[m]} : V_{(1)} \to V$ from (194) satisfies $\log_{[m]} = \log_{[m]} \circ \pi_{[m]}$ and is continuous for each $m \geq 0$, and $\log$ from (190) maps subsets of $\{\boldsymbol{t} \in V_{(1)} \mid |||\boldsymbol{t} - 1|||_\lambda \leq 1\}$ to subsets of $\{\boldsymbol{\ell} \in V^\infty \mid |||\boldsymbol{\ell}|||_\rho \leq \sum_{m\geq 0}(2\rho/\lambda)^m\}$ for any $\lambda > 2\rho$;*

(v) *for each $m \geq 0$, it holds that on $|||\cdot|||_\rho$-bounded subsets the projections $\pi_{[m]} : V^\infty \to V^\infty$ converge uniformly w.r.t. $|||\cdot|||_1$ to the identity operator on $V^\infty$;*

(vi) *for each $\lambda > 0$, we have that $\delta_\lambda \circ \log = \log \circ\, \delta_\lambda$ and $\delta_\lambda[\mathfrak{sig}(x)] = \mathfrak{sig}(\lambda \cdot x)$, any $x \in \mathcal{BV}$;*

(vii) *for each $\varphi \in C(\mathcal{K})$ with $\mathcal{K} \subset \mathcal{BV}$ compact, there is a sequence of index-polynomials $(\boldsymbol{\ell}_j)_{j\in\mathbb{N}}$ in $V^\circ := \bigoplus_{m\geq 0} V_m \subset V^\infty$ such that $\varphi = \lim_{j\to\infty}\langle \mathfrak{sig}(\cdot), \boldsymbol{\ell}_j\rangle$ w.r.t. $\|\cdot\|_\infty$.*

*Proof.* (i), (ii) and (iii) are well-known, see e.g. [17, Cor. 2.4, Cor. 5.5] and [34, Thm. 4].

(iv): As is immediate from (194), the map $\log_{[m]}$ is a polynomial and hence $V$-valued and continuous, the latter by the fact that (both $\pi_{[m]}$ and) the multiplication $*$ ($\cong$ tensor multiplication $\otimes$; (193)) on $V$ is continuous (e.g. [17, Section 3]). The commutativity of $\log_{[m]}$ and $\pi_{[m]}$ is clear again from (194). As to the boundedness assertion, let $\lambda > 2\rho$ and denote $B_\lambda := \{\boldsymbol{t} \in V_{(1)} \mid |||\boldsymbol{t} - 1|||_\lambda \leq 1\}$. Then in particular $\sup_{\boldsymbol{t}\in B_\lambda} \|\pi_m(\boldsymbol{t})\|_m \leq \lambda^{-m}$ for every $m \geq 0$, whence for each $\boldsymbol{t} \in B_\lambda$ and $\boldsymbol{\ell} := \log(\boldsymbol{t})$ we have that, for all $m \geq k \geq 1$,

$$\left\|\pi_m\left[(\boldsymbol{t} - 1)^{*k}\right]\right\|_m \;\leq\; \sum_{\substack{m_1+\ldots+m_k=m \\ m_\nu\geq 1}} \|\pi_{m_1}(\boldsymbol{t}) * \cdots * \pi_{m_k}(\boldsymbol{t})\|_m \;\leq\; \binom{m-1}{k-1}\cdot\lambda^{-m}$$

(as the tensor norms $\|\cdot\|_m$ are each submultiplicative), and hence find from (190) that

$$\|\pi_m(\boldsymbol{\ell})\|_m \;\leq\; \sum_{k=1}^{m} \frac{1}{k}\binom{m-1}{k-1}\lambda^{-m} \;\leq\; 2^m\lambda^{-m} \qquad \text{for each} \quad m \geq 1,$$

implying that $\sup_{\ell\in\log(B_\lambda)} |||\boldsymbol{\ell}|||_\rho \leq \sum_{m\geq 1}(2\rho/\lambda)^m < \infty$, as desired.

(v): Let $B \subset V^\infty$ be bounded w.r.t. $|||\cdot|||_\rho$, i.e. suppose that $\beta_\rho := \sup_{\boldsymbol{t}\in B} |||\boldsymbol{t}|||_\rho < \infty$. Then there will be some $0 < q < 1$ together with an index $m_0 \geq 1$ such that

$$\sup_{\boldsymbol{t}\in B} \|\pi_m(\boldsymbol{t})\|_m \;\leq\; q^m \qquad \text{for each} \quad m \geq m_0. \tag{197}$$

Indeed: Assuming otherwise that the above does not hold, we for any given $q \in (0,1)$ obtain the existence of a sequence $(\boldsymbol{t}^{(n)})_{n\in\mathbb{N}} \subset B$ with the property that

$$\|\pi_{m_n}(\boldsymbol{t}^{(n)})\|_{m_n} \;>\; q^{m_n} \qquad \text{for each} \quad n \in \mathbb{N}$$

for some strictly increasing sequence $(m_n)_{n\in\mathbb{N}} \subset \mathbb{N}$. But choosing $q > \rho^{-1}$ then implies that

$$\beta_\rho \;\geq\; \sup_{n\in\mathbb{N}} \left|\left|\left|\boldsymbol{t}^{(n)}\right|\right|\right|_\rho \;\geq\; \sup_{n\in\mathbb{N}}(\rho\cdot q)^{m_n} = \infty$$

in contradiction to the $|||\cdot|||_\rho$-boundedness of $B$. Thus (197) holds, and with it (by convergence of the geometric series) the claimed uniform $|||\cdot|||_1$-convergence of $(\pi_{[m]})_{m\in\mathbb{N}}$.

(vi): This is clear by inspection of (190) and (186), respectively.

(vii): This approximation property, which is sometimes referred to as the *universality* of the signature, is an immediate consequence of the Stone-Weierstrass theorem and the fact that the set $\{\langle \mathfrak{sig}(\cdot), \boldsymbol{\ell}\rangle \mid \boldsymbol{\ell} \in V^\circ\}$ is a subalgebra of $C(\mathcal{K})$ which contains the constants and separates points, see e.g. [16, Thm. 5.6. (2)]. □

## References


[1] L.B. Almeida. Misep – linear and nonlinear ica based on mutual information. *J. Mach. Learn. Res.*, **4**:1297–1318, 2003.

[2] F. R. Bach and M. I. Jordan. Kernel Independent Component Analysis. *J. Mach. Learn. Res.*, **3**:1–48, 2002.

[3] A. J. Bell and T. Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, **7**.6:1129–1159, 1995.

[4] A. Belouchrani, K. A. Meraim, J. F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, **45**.2:434–444, 1997.

[5] J. Berner, D. Elbrächter, and P. Grohs. How degenerate is the parametrization of neural networks with the ReLU activation function? *arXiv: 1905. 09803*, 2019.

[6] P. Billingsley. *Convergence of Probability Measures*. Second Edition, John Wiley & Sons, 1999.

[7] Patric Bonnier and Harald Oberhauser. Signature cumulants, ordered partitions, and independence of stochastic processes. *Bernoulli*, **26**.4:2727–2757, 2020.

[8] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

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

[10] P. Brakel and Y. Bengio. Learning independent features with adversarial nets for non-linear ICA. *Preprint*, page arXiv:1710.05050[stat.ML], 2017.

[11] J. F. Cardoso. High-order Contrasts for Independent Component Analysis. *Neural Computation*, **11**:157–192, 1999.

[12] J. F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, **140**.6:362–370, 1993.

[13] Kuo-Tsai Chen et al. Iterated path integrals. *Bulletin of the American Mathematical Society*, 83(5):831–879, 1977.

[14] X. Chen and Y. Fan. Estimation of Copula-Based Semiparametric Time Series Models. *Journal of Econometrics*, **130**.2:307–335, 2006.

[15] I. Chevyrev and T. Lyons. Characteristic functions of measures on geometric rough paths. *Ann. Probab.*, **44**.6:4049–4082, 2016.

[16] I. Chevyrev and H. Oberhauser. Signature Moments to Characterize Laws of Stochastic Processes. *Preprint*, page arXiv:1810.1097, 2018.

[17] Ilya Chevyrev and Terry Lyons. Characteristic functions of measures on geometric rough paths. *Annals of Probability*, 44(6):4049–4082, 2016.

[18] P. Comon. Independent Component Analysis, a new concept? *Signal Processing*, **36**.3:287–314, 1994.

[19] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic press, 2010.

[20] J. B. Conway. *Functions of One Complex Variable*. Second Edition. Graduate Texts in Mathematics **11**, Springer, 1978.

[21] G. Darmois. Analyse générale des liaisons stochastiques. *Rev. Inst. Intern. Stat.*, **21**:2–8, 1953.

[22] W. F. Darsow, B. Nguyen, and T. Olsen. Copulas and Markov Processes. *Illinois J. Math.*, **36**.4:600–642, 1992.

[23] M. Denker. The central limit theorem for dynamical systems. *Dynamical Systems and Ergodic Theory, (K. Krzyzewski, ed.)*, pages 33–62, Banach Center Publications, Polish Scientific Publishers, Warsaw, 1989.

[24] Joscha Diehl, Peter Friz, Hilmar Mai, et al. Pathwise stability of likelihood estimators for diffusions via rough paths. *The Annals of Applied Probability*, 26(4):2169–2192, 2016.

[25] P. Embrechts, A. McNeil, and D. Straumann. Correlation and Dependence in Risk Management: Properties and Pitfalls. *Risk Management: Value at Risk and Beyond*, **1**:176–223, 2002.

[26] T. Emura, T.-H. Long, and L.-H. Sun. R routines for performing estimation and statistical process control under copula-based time series models. *Communications in Statistics - Simulation and Computation*, **46**.4:3067–3087, 2017.

[27] J. Eriksson and V. Koivunen. Identifiability, separability and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, **11**:601–604, 2004.

[28] J. Fan. A Selective Overview of Nonparametric Methods in Financial Econometrics. *Statistical Science*, **20**.4:317–337, 2005.

[29] Michel Fliess. Fonctionnelles causales non linéaires et indéterminées non commutatives. *Bulletin de la société mathématique de France*, 109:3–40, 1981.

[30] P. K. Friz and N. B. Victoir. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*. Cambridge Studies in Advanced Mathematics **120**, Cambridge University Press, 2010.

[31] P. Fryzlewicz and S. S. Rao. Mixing properties of ARCH and time-varying ARCH processes. *Bernoulli*, **17**.1:320–346, 2011.

[32] L. M. Graves. *Theory of functions of real variables*. McGraw-Hill, 1946.

[33] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *J. Mach. Learn. Res.*, **6**:2075–2129, 2005.

[34] B. Hambly and T. Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Ann. of Math.*, **171**.1:109–167, 2010.

[35] S. Harmeling, A. Ziehe, M. Kawanabe, and K.R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, **15**.5:1089–1124, 2003.

[36] T. Hastie and R. Tibshirani. Independent component analysis through product density estimation. *Advances in Neural Information Processing Systems*, **15**:649–656, 2003.

[37] R. D. Hjelm et al. Learning deep representations by mutual information estimation and maximization. *Preprint*, page arXiv:1808.06670v5[stat.ML], 2018.

[38] A. Hyvärinen. *Independent Component Analysis by Minimization of Mutual Information*. Technical Report (Report A46), Helsiniki University of Technology, Department of Computer Science and Engineering, September 1997.

[39] A. Hyvärinen. New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit. *Adv. Neural Inf. Process Syst.*, pages 273–279, 1998.

[40] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, **10**.3:626–634, 1999.

[41] A. Hyvärinen. Independent component analysis: recent advances. *Phil. Trans. R. Soc. A*, **371**.1984:20110534, 2013.

[42] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[43] A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *NeurIPS2016*, pages 3765–3773, 2016.

[44] A. Hyvärinen and H. Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. *PMLR*, **54**:460–469, 2017.

[45] A. Hyvärinen and P. Pajunen. Nonlinear Independent Component Analysis: Existence and Uniqueness Results. *Neural Networks*, **12**.3:429–439, 1999.

[46] A. Hyvärinen, H. Sasaki, and R. Turner. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. *AISTATS*, 2019.
[47] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. *Proc. Artificial Intelligence and Statistics (AISTATS2020)*, 2020.
[48] P. Kidger and T. Lyons. Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU. *arXiv:2001.00706 and https://github.com/patrick-kidger/signatory*, 2020.
[49] C. Kleiber and J. Stoyanov. Multivariate distributions and the moment problem. *J. Multivariate Anal.*, **113**:7–18, 2013.
[50] A. Kuczmaszewska. On the Strong Law of Large Numbers for $\phi$-Mixing and $\rho$-Mixing Random Variables. *Acta Math. Hungar.*, **132**.1-2:174–189, 2011.
[51] J. M. Lee. *Introduction to Smooth Manifolds.* Graduate Texts in Mathematics **218**, Springer, 2013.
[52] T. J. Lyons, M. Caruana, and T. Lévy. *Differential Equations Driven by Rough Paths.* Springer, 2007.
[53] T. J. Lyons and Z. Qian. *System Control and Rough Paths.* Oxford Mathematical Monographs, Oxford University Press, 2002.
[54] M. B. Marcus and L. A. Shepp. Sample behavior of gaussian processes. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 423–441, 1972.
[55] H. Masuda. Mixing properties of ARMA processes. *Stochastic Process. Appl.*, **19**:297–303, 1988.
[56] H. Masuda. Ergodicity and exponential $\beta$-mixing bounds for multidimensional diffusions with jumps. *Stochastic Process. Appl.*, **117**.1:35–56, 2007.
[57] R. A. McCoy. Second countable and separable function spaces. *The American Mathematical Monthly*, **85**.6:487–489, 1978.
[58] J. Miettinen, K. Nordhausen, and S. Taskinen. Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp. *Journal of Statistical Software*, **76**.2, 2017.
[59] E. Moulines, J. F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, page 3617–3620, 1997.
[60] J. R. Munkres. *Topology.* Second Edition. Prentice Hall, 2000.
[61] R. B. Nelsen. *An Introduction to Copulas.* Springer Series in Statistics. Second Edition. Springer, 2006.
[62] C. Papavasiliou, A; Ladroue. Parameter Estimation for Rough Differential Equations. *Ann. Stat.*, **39**.4:2047–2073, 2011.
[63] P. Petersen, M. Raslan, and F. Voigtlaender. Topological Properties of the Set of Functions Generated by Neural Networks of Fixed Size. *Found. Comput. Math*, **21**:375–444, 2021.
[64] D. T. Pham and P. Garrat. Blind Separation of Mixture of Independent Sources Through a Quasi-Maximum Likelihood Approach. *IEEE Trans. Signal Process.*, **45**.7:1712–1725, 1997.
[65] T. D. Pham and L. T. Tran. Some mixing properties of time series models. *Stochastic Process. Appl.*, **19**:297–303, 1985.
[66] C. E. Rasmussen and C-K. I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, 2006.
[67] M. Reed and B. Simon. *Methods of Modern Mathematical Physics: Vol. I, Functional Analysis.* Academic Press, 1972.

[68] R. Remmert. *Theory of Complex Functions.* (Fourth corrected printing.) Springer, 1998.

[69] C. Reutenauer. *Free Lie Algebras.* London Mathematical Society Monographs, New Series. Oxford Science Publications **7**, The Clarendon Press, 1993.

[70] L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales.* Volume 1, Cambridge University Press, 2000.

[71] M. Rosenblatt. *Markov Processes, Structure and Asymptotic Behavior.* Springer-Verlag, New York, 1971.

[72] W. Rudin. *Principles of Mathematical Analysis.* (3rd ed.) McGraw-Hill, 1976.

[73] W. Rudin. *Real and Complex Analysis.* (3rd ed.) McGraw-Hill, 1987.

[74] R.J. Samworth and M. Yuan. Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.*, **40**.6:2973–3002, 2012.

[75] A. Schell. *SigNICA.* GitHub repository, https://github.com/alexander-schell/SigNICA.git Code for Section 8 (Jupyter Notebooks and Python files)., 2021.

[76] E. Schlemm and R. Stelzer. Multivariate CARMA processes, continuous-time state space models and complete regularity of the innovations of the sampled processes. *Bernoulli*, **18**.1:46–63, 2012.

[77] V. P. Skitovich. On a Property of a Normal Distribution. *Doklady Akad. Nauk. SSSR*, **89**:217–219 (in Russian), 1953.

[78] Y. Tan, J. Wang, and J. M. Zurada. Nonlinear blind source separation using a radial basis function network. *IEEE Trans. Neural Netw.*, **12**.1:124–134, 2001.

[79] T. et al. Teshima. Coupling-based invertible neural networks are universal diffeomorphism approximators. *arXiv: 2006. 11469, accepted at NeurIPS 2020*, 2020.

[80] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics, 1998.

[81] A. W. van der Vaart. *Time Series.* Version 242013, Lecture Notes, Universiteit Leiden, 2013. (Available at http://www.math.leidenuniv.nl/~avdvaart/timeseries/dictaat.pdf.).

[82] J. A. Wellner. *Empirical Processes: Theory and Applications.* Special Topics Course Spring 2005, Delft Technical University, June 2005. (Available at https://www.stat.washington.edu/people/jaw/RESEARCH/TALKS/Delft/emp-proc-delft-big.pdf).

[83] C. Xiaohong, L. P. Hansen, and M. Carrasco. Nonlinearity and temporal dependence. *Econometrics*, **155**.2:155–169, 2010.

[84] Q. Zhou and J. Yu. Asymptotic theory for linear diffusions under alternative sampling schemes. *Ecnonomics Letters*, **128**:1–5, 2015.



Mathematical Institute
University of Oxford
AWB, ROQ, Woodstock Road
Oxford, OX2 6GG, UK


25 June 2021