The Annals of Statistics
2023, Vol. 51, No. 2, 487–518
https://doi.org/10.1214/23-AOS2256

NONLINEAR INDEPENDENT COMPONENT ANALYSIS FOR DISCRETE-TIME AND CONTINUOUS-TIME SIGNALS

BY ALEXANDER SCHELL[a] AND HARALD OBERHAUSER[b]

Mathematical Institute, University of Oxford, [a]*schell@maths.ox.ac.uk,* [b]*oberhauser@maths.ox.ac.uk*

We study the classical problem of recovering a multidimensional source signal from observations of nonlinear mixtures of this signal. We show that this recovery is possible (up to a permutation and monotone scaling of the source's original component signals) if the mixture is due to a sufficiently differentiable and invertible but otherwise arbitrarily nonlinear function and the component signals of the source are statistically independent with 'non-degenerate' second-order statistics. The latter assumption requires the source signal to meet one of three regularity conditions which essentially ensure that the source is sufficiently far away from the nonrecoverable extremes of being deterministic or constant in time. These assumptions, which cover many popular time series models and stochastic processes, allow us to reformulate the initial problem of nonlinear blind source separation as a simple-to-state problem of optimisation-based function approximation. We propose to solve this approximation problem by minimizing a novel type of objective function that efficiently quantifies the mutual statistical dependence between multiple stochastic processes via cumulant-like statistics. This yields a scalable and direct new method for nonlinear Independent Component Analysis with widely applicable theoretical guarantees and 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—that is, 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:

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 (2)
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

Received June 2021; revised December 2022.
MSC2020 subject classifications. Primary 62H25, 62M99; secondary 62H05, 60L10, 62M45, 62R10.
Key words and phrases. Blind source separation, independent component analysis, inverse problem, latent variable model, statistical independence, unsupervised learning, statistical inference for stochastic processes, functional data analysis, nonlinear BSS, nonlinear ICA.

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.

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, modelled as real-valued time series each. 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. Now without further assumptions, the true explanation (S, A) of the data X cannot be distinguished from any of its 'alternative explanations' $\{(\tilde{S}, \tilde{A}) \equiv (B \cdot S, AB^{-1}) | B \in \mathbb{R}^{2\times 2} \text{ invertible}\}$. 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 $\{(\tilde{S}, \tilde{A}) \equiv (B\Lambda \cdot S, A\Lambda^{-1}B^\intercal) | \Lambda \in \mathbb{R}^{2\times 2}$ (invertible) diagonal, $B \in \mathbb{R}^{2\times 2}$ 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 [17, 49, 54] 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 [14], quickly became the theoretical foundation of *Independent Component Analysis (ICA)*, a popular statistical method that has since seen far-reaching theoretical investigations and extensions, for example, [3, 51], and has been successfully implemented in numerous widely-applied algorithms, for example, [5, 9, 26, 31]; see, for instance, [22, 32, 43] and the monographs [15, 33] for an overview.

Comon's contribution is arguably the most conceptionally 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 'nontrivial' transformations that preserve the mutual statistical independence of their input vectors [36]. This is a substantial limitation not only from a theoretical perspective but also in applications, where real-world data is often assumed to depend nonlinearly on certain nonredundant (independent) explanatory source signals and the instantaneous invertible nonlinear model (1) is deemed an adequate description of this dependence. See, for instance, [2, 16, 19, 27, 38, 40, 46] and the references therein for a few according example applications of nonlinear BSS ranging from the analysis of star clusters in interstellar gas clouds and biomedical tissue monitoring during surgery over electroencephalography and molecular simulation to statistical process monitoring, vibration analysis and stock market prediction.

[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).

Overcoming the traditional confinement to linearity has thus been a long-standing scientific endeavour, and the past twenty-six years have seen various attempts of establishing alternative identifiability approaches 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, for example, [1, 7, 28, 39, 55], 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 [25].

More recently, the works of Hyvärinen et al. [34, 35, 37] 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 [34] or its history [35], and then training logistic regression to discriminate ('contrast') between the thus-augmented observable and some additional 'variation' of the data. This variation links 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 results were extended and embedded into the context of variational autoencoders in [39].

Motivated by the classical ICA framework of Comon [14] and the recent contrastive learning breakthrough [35], we revisit the inference problem (2) for stochastic processes[2] $X = (X_t)$ 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 generalise Comon's classical independence-based identifiability criterion from linear mixtures of random vectors to nonlinear mixtures of discrete- and 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 covering both discrete-time and continuous-time models alike, including Stochastic Differential Equations (SDEs) in particular.

Blind Source Separation via Signature Cumulants Our identifiability theory allows us to reformulate the problem of nonlinear blind source separation as an easy-to-state optimisation problem which involves the minimisation of statistical dependence between multiple stochastic processes, see Theorem 4. 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 both instantaneously and over different points in time. We propose to quantify such complex dependency relations by using so-called signature cumulants [6] as objective functions. 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 efficiently computable quantification of the degree of statistical (in)dependence between stochastic processes. Joined with our optimisation approach, this combines to a widely applicable new and robust statistical method for the nonlinear blind source separation of time-dependent signals, see Theorem 4 and Section E.4 in [53].

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 nonequally spaced time grids. Secondly,

[2]Throughout, 'stochastic process' means 'continuous-time stochastic process' unless mentioned otherwise.

oftentimes 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 and finite sample of the observable is given [53], Section E. This is also the setting in which our experiments are carried out in Section 8.

This article is structured as follows. We precede our statistical analysis with an informal yet concise summary of this paper's main contributions (Section 2). The formal exposition of our approach towards the recovery of nonlinearly mixed independent sources begins thereafter by recalling the main results of [14] as conceptional points of reference (Section 3). 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 4), we show how sources admitting a nondegenerate '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 a minimal deviation by maximizing an independence criterion (Section 5). Our approach covers sources of various types of statistical regularity, 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) which leads to a practical and statistically consistent separation algorithm that we demonstrate in a series of numerical experiments (Section 8). The paper ends with a brief conclusion and an outlook on future directions (Section 9). Due to space constraints, most of the proofs, along with a consistency analysis for our method, are presented in the Supplementary Material [53]. This supplement further contains some technical auxiliaries and additional remarks, including an explication of how, as promised in the title, the results and methods in this paper are directly applicable to the separation of discrete-time signals as well ([53], Section G).

2. Summary of contribution. Motivated by recent breakthroughs of Hyvärinen and Morioka [34, 35], we propose a new approach to the problem of nonlinear blind source separation (2) for multidimensional time-dependent signals that leverages modern tools from stochastic analysis: For an unknown discrete- or continuous-time signal $S = (S_t)$ in $\mathbb{R}^d$ and an unknown function $f : \mathbb{R}^d \to \mathbb{R}^d$, a new statistical method to recover S from its transformation $X = f(S) \equiv (f(S_t))$ via 'signature cumulants' is presented.

In essence, we provide a new algorithm[3] that performs the inversion, or 'retransformation',

$$X \longmapsto S \tag{3}$$

of the generative relation (1) in the case that f and S are not explicitly known and f is sufficiently differentiable and (by necessity) invertible[4] but otherwise arbitrarily nonlinear.

Finding ways to achieve this 'blind inversion' (3) has been of long-standing scientific interest, and efforts in this direction gave rise to an established area of specialised statistical research that has been very active for nearly three decades now. Apart from only a small number of exceptions, however, related works were predominantly confined to the very limiting assumption that the hidden relation f be a linear map on $\mathbb{R}^d$—the few existing approaches towards the blind inversion of nonlinear causal relations were either heuristic or required f to belong to very narrowly defined function classes only, and it was not until the recent breakthroughs of Hyvärinen et al. that the first mathematically justified ideas for the blind inversion of general nonlinear relationships between X and S have emerged. Our work is a contribution to the dawning research on nonlinear blind inversion.

[3]That is, an explicitly computable map—or estimator, in the statistical sense—that takes in [a realisation of] the mixture X and returns an 'optimal' approximation of [the corresponding realisation of] S as an output.

[4]Invertible at least on the smallest subset of $\mathbb{R}^d$ which is actually reached by S, but see Definition 2 and (13).

2.1. *Identifiability* (*Theorems* 2, 3). To achieve a meaningful recovery (3) of the source S from X, we need to compensate for the blindness regarding f and S by imposing some additional assumptions on the latter. The most established such assumption, and arguably the most relevant one in practice, is that the component signals of S are statistically independent; we adopt this assumption throughout.

Many of the conceptional issues that arise in the nonlinear blind reconstruction of an independent-component source S from X can then be anticipated from the classical, that is, linear case $f \in \mathbb{R}^{d\times d}$ already. Similar to the classical case (cf. Theorem 1):

- the blindness[5] underlying (3) makes an exact recovery of S impossible, but statistical prior information on the source allows to identify S from X up to a minimal ambiguity, namely up to a permutation and monotone scaling of the source's original component signals;
- these minimally ambiguous (in the above sense) estimates $\hat{S}$ of the original source S preserve the initial condition of intercomponental independence (IC), but under some natural assumptions on S the converse is also true: those retransformations of X which are IC must be minimally ambiguous to S.

These insights into the blind inversion (3), which are rigorously discussed in Section 5, are the mathematical heart of our approach. Especially the equivalence stated in the last point, which is made precise in Theorems 2 and 3, is a central new finding:

Under some mild statistical conditions on the source S, we can show that the assumed IC property of the source is strong enough to trivialise[6] the action of any spatial diffeomorphism which preserves this property; in other words: their property of having minimal intercomponental statistical dependence distinguishes the minimally ambiguous estimates $\hat{S}$ of S from any other invertible nonlinear transformations of X.

This makes 'minimisation of intercomponent-dependence' an illuminating optimisation principle for the initially blind search for S, which immediately translates into the following strategy for the desired inversion (3):

(4) as an estimate $\hat{S}$ for S, choose $\hat{S} = \theta_\star(X)$, $\theta_\star$ invertible, s.t. $\theta_\star(X)$ is IC;

that is, the right retransformations of X are those that minimise intercomponental dependence.

As mentioned, the sources S for which this strategy works are those that 'carry their IC property well enough' for this property to characterise them, up to minimal ambiguity, among their (invertible) nonlinear transformations. But not every source is of this kind, as becomes particularly clear from considering two 'unrecoverable' statistical extreme cases: If the source S is deterministic,[7] then the IC property is void and a meaningful blind inversion (3) of the source's mixtures is generally impossible. If S is constant in time, that is, $S = (Z)_{t\in[0,1]}$ for some random vector Z in $\mathbb{R}^d$, then the IC property on S cannot manifest cross-componentally over different time-points and is then generally too weak to support the strategy (4) for nonlinear mixtures, see [36] and Example 3.1.

These unidentifiable source types can be seen as degenerate extremes that are naturally interpolated by the mathematical model class of continuous-time stochastic processes, and said interpolation can be controlled at the level of the second-order finite-dimensional distributions (fdds) of such processes, see Section 4. In fact, we can formulate three regularity assumptions on the family of fdds of a source S which enable the IC-based identifiability

[5]That is, the fact that the inverse problem (3) is inherently underdetermined since the constituents f and S of the RHS in (1) are both unknown.

[6]Here, 'trivialise' means reduce to the composition of a permutation and a componentwise monotone scaling.

[7]That is, if S attains exactly one sample path with probability one.

(4) of the source by ensuring that it is sufficiently far away from the above degeneracies (Section 5). More specifically, our nondegeneracy assumptions on the source require that sufficiently many of its fdds admit a probability density which is sufficiently complex in that it satisfies one of the following conditions:

(a) the density avoids local factorisations and is not of a certain 'pathological' Gaussian-like shape, as is made precise in Definition 6;

(b) the density has locally nonvanishing mixed log-derivatives that lie outside certain nullsets, as specified in Definition 7.

While the nonfactorizability and nonvanishing-log-derivative conditions ensure that the source is 'stable enough' to make its IC property unfold[8] into its component signals in such a way that the ('residual') action inflicted upon S by the composition of the mixing transformation f with an IC-enforcing retransformation [as in (4)] does not collapse when considered jointly at different points in time, the exclusion of Gaussian-like shapes or algebraically degenerate density configurations ensures that this residual action on S is 'expressive' enough (as per implying a nondegenerate eigenspectrum of a Jacobian). All of this is made precise in Section 5.1 and the proofs of Theorems 2 and 3.

The source conditions (a) and (b) again generalise classical theory in a natural way (cf. Section 3 and the remarks on p. 499 and Remark C.2), and in Section 6 we illustrate their broad applicability by compiling a set of widely used signal classes to which these conditions apply.

Thus far, our work has established the dependence-minimising approach (4) as a successful mathematical strategy to achieve the nonlinear blind source separation task (3), see Theorem 2 and Theorem 3: We identified natural probabilistic conditions (a) and (b) on the source which guarantee that its IC property manifests strongly enough to characterise that source among any invertible (re)transformations of X up to some inevitable ambiguity.[9]

2.2. *Blind inversion via optimisation* (*Theorem* 4). In the second part of the paper, we propose a way to turn this theoretical strategy into a ready-to-use statistical method that can be easily implemented in practice. What we need to do for this is provide the observer of the mixture X with three things, namely:

– a set Θ of invertible candidate demixing transformations on $\mathbb{R}^d$ which is 'large enough' to include approximations of the original inverse f^{-1} up to permutation and scale, and for consistency is endowed with a suitable approximation topology;[10]
– a 'pair of goggles' ϕ that allows the observer to gauge the degree of intercomponental statistical dependence of any given (re)transformation of X: the weaker the statistical dependence between the component signals of a process Y, the smaller shall be $\phi(Y) \in \mathbb{R}_{\geq 0}$; the desired inversion (3) is then performed [via (4)] by choosing those transformations $\theta(X)$, $\theta \in \Theta$, of X for which the value $\phi(\theta(X))$ is minimal;
– an automatable optimisation procedure that combines Θ and ϕ and returns

$$\theta_\star \in \operatorname*{arg\,min}_{\theta\in\Theta} \phi\big(\theta(X)\big) \quad \text{and then} \quad \hat{S} = \theta_\star(X) \tag{5}$$

as the desired [minimally ambiguous] estimate of S, in accordance with (4).

[8] Instead of holding it merely within its fixed-time marginals, as in the generally unidentifiable case of IC random vectors in $\mathbb{R}^d$.

[9] That is, as we recall, up to a permutation and monotone scaling of the source's original component signals.

[10] For a specification of 'large enough' see the hypothesis on Θ that is formulated in Theorem 4, and for a sufficient assumption on the approximation topology on Θ see [53], Assumption 3 (on p. 39).

The above is formalised in Theorem 4. A natural choice in practice is to implement Θ as the realisation space of an invertible artificial neural network (NN) with d input nodes; cf. Section 8.3 and [53], Remark C.3(ii). Adding ϕ as a loss function to the NN, the optimisation (5) can then be performed efficiently via backpropagation; details in Sections 7, 8, and E.

Intuitively speaking, in the course of the optimisation (5) the observer gradually performs the desired inversion (3) directly by comparing different transformations of the data and choosing as most akin to the true inverse those that minimize the ϕ-quantified statistical dependence of X. For nonlinear f the theoretical justification of this procedure is new, while the underlying idea of source separation via quantified dependence minimisation is a well-established concept for the recovery of linearly mixed random vectors in $\mathbb{R}^d$, see, for example, Corollary 1.

Inspired by another classical concept (cf. (9) on page 496), in Section 7 we propose as dependence quantification ϕ a 'cross-cumulant'-based energy functional of the form

$$\phi(Y) = \sum_{m=2}^{\infty} \sum_{\boldsymbol{q}_m} \bar{\kappa}_{\boldsymbol{q}_m}(Y)^2, \tag{6}$$

where $\bar{\kappa}_{\boldsymbol{q}}(Y)$ denotes 'the (standardised) signature cumulant at index $\boldsymbol{q}$' of a stochastic process Y in $\mathbb{R}^d$, see Definition 9 and Notation 7.1, and the inner sums run over all 'cross-shuffles' of word-length m (see (141) on page 33). The entirety of all signature cumulants $(\bar{\kappa}_{\boldsymbol{q}}(Y))$, which can be thought of as a carefully chosen 'coordinate vector' for the distribution of the multidimensional stochastic process Y, provides a hierarchical and parsimous description of the statistical dependence relations within Y, which may occur simultaneously between coordinates and over different points in time. The functional (6) summarises the aspects of this description that are most central for us, namely 'how much' of this dependence there is between the multiple component signals of Y. Since the above ϕ vanishes over exactly those processes that are IC (Proposition 4), the functional (6) is well suited to operationalise the inversion strategy (4) via the optimisation scheme (5), as described in Theorem 4; further aspects are discussed in Sections E and 8.

2.3. *Consistency* ([53], *Theorem E.*1). Up to this point, we discussed the method (5) in a setting where the whole distribution of X is assumed to be known. This idealisation is of course difficult to uphold in practice, where mixtures are typically not available as continuous-time stochastic processes and only discrete-time sample trajectories of X, that is, finite sequences of data points in $\mathbb{R}^d$, are observed.

The statistical guarantees of Theorem E.1 in [53] ensure that our method remains applicable under these practical constraints. More specifically, a statistical consistency analysis of the procedure (5) requires to simultaneously deal with:

- time-discretization: if S, and hence X, are continuous-time signals, then 'full' sample observations of the underlying model X (i.e., continuous paths in $\mathbb{R}^d$) are not available in real-world applications, where only discrete-time data can be used;[11]
- finite samples: typically, one of two situations arise in applications. One is that n presumably independent [discrete-time] sample trajectories of the observable are recorded, for example, medical recordings of n patients. The other situation is that only one [discrete-time] sample trajectory of X is given and ergodicity or mixing assumptions are invoked to make inference about the underlying distribution; for example, this situation is common in finance and economics.

[11] In spirit, this is similar to the well-developed statistical question of parameter estimation for stochastic differential equations where also only time-discretized sample trajectories are observed.

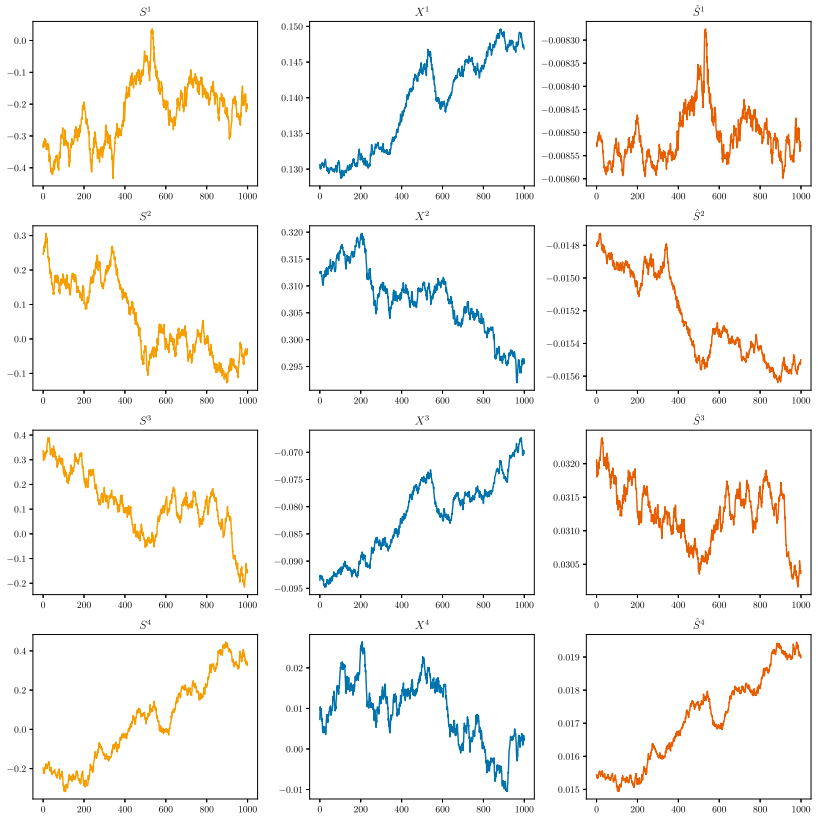

FIG. 1. *A source S with four components $S^1, \dots, S^4$ (orange) is mapped under some nonlinear transformation f, resulting in the observed mixture X (blue). We present a new method to recover the original source S from its mixture X up to a minimal deviation: Given X, this approach returns an estimate $\hat{S} = (\hat{S}^1, \dots, \hat{S}^4)$ (brown) that approximates S up to the original order of its channels (corrected here for ease of comparison) and a monotone scaling.*

We show that under general conditions, which for example are satisfied by many classical SDE models, our method (5) is (strongly) consistent in a sense that addresses both of these points: As the grid of observational time-points gets finer and the length of the observed time series increases, our method (5) produces a signal $\hat{S}$ that gets closer to the unobserved source signal S, even when the model for S is formulated in continuous time; see Theorem E.1 for the precise statement. Additionally, Theorem E.1 shows that our method is robust under approximations of the contrast function ϕ (for computational purposes, the series (6) of signature cumulants needs to be truncated in practice). The key ingredients to establish this result are tools from stochastic analysis, natural assumptions on the topology of function approximators (e.g., deep neural networks), and statistical arguments on the optimality of extremum estimators. Practitioners may find the displayed algorithm in Section E.4 a useful summary.

Our exposition is complemented by a number of numerical examples (Section 8) which further illustrate the practical utility of our method by applying it to a series of nonlinear blind inversion problems (3) with multidimensional source signals in discrete and continuous time.

A practical illustration of our blind inversion approach (5) and its underlying procedure, contextualizing Figure 1 above, is given as Example A.1 in the Supplementary Material [53].

We emphasize that the above methodology in its entirety, including any of our definitions or theorems, applies to both continuous-time and discrete-time signals alike.[12] The latter type includes signals that are 'genuinely discrete', that is, generated from a discrete-time process, and signals that are of continuous origin but 'discretely observed', that is, obtained from sampling a continuous-time process at a discrete set of time points. These cases are treated in detail in Section E and [53], Section G, which are referenced accordingly throughout the text.[13]

In total, the contents of this paper combine to a general and flexible new statistical method for the nonlinear blind source separation of multidimensional time-dependent signals.

3. 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 identifiability framework [14]. This section briefly recalls the main results of this classical framework as conceptional points of reference.

THEOREM 1 (Comon [14], Theorem 11). *Let $S = (S^1, \dots, S^d)^\intercal$ be a random vector in $\mathbb{R}^d$ with mutually independent, nondeterministic components $S^1, \dots, S^d$ of which at most one is Gaussian. Let further $X = C \cdot S$ for an orthogonal[14] matrix $C \in \mathbb{R}^{d\times d}$. Then, for any orthogonal matrix $\theta \in \mathbb{R}^{d\times d}$, we have the following characterisation*:

$$\begin{aligned}(\tilde{S}^1, \dots, \tilde{S}^d) := \theta \cdot X = \Lambda P \cdot S \quad &\textit{for some } (\Lambda, P) \in \Delta_d \times \mathrm{P}_d \\ \textit{if and only if} \quad \tilde{S}^1, \dots, \tilde{S}^d \quad &\textit{are mutually independent.}\end{aligned} \tag{7}$$

The significance of Theorem 1 is that it characterises—up to some minimal deviation, namely their scaling and reordering—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.

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

COROLLARY 1 ([14]). *Let X and S be as in Theorem 1. Then for any function[15] $\phi : \mathcal{M}_1(\mathbb{R}^d) \to \mathbb{R}_+$ such that $\phi(\mu) = 0$ iff $\mu = \mu^1 \otimes \cdots \otimes \mu^d$, it holds that*[16]

$$\Big[\arg\min_{\theta\in\Theta} \phi(\theta \cdot X)\Big] \cdot X \subseteq \mathrm{M}_d \cdot S, \tag{8}$$

where $\mathrm{M}_d := \{\Lambda \cdot P | (\Lambda, P) \in \Delta_d \times \mathrm{P}_d\}$ 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 $S = (S^1, \dots, S^d)^\intercal$ a random vector with mutually independent, non-Gaussian components, the constituent relationship (1) between the observable

[12] For the case of discrete-time signals, everything basically applies as in the continuous case up to very minor modifications necessitated by the change from (path-)connected to discrete realisations of the underlying signals.

[13] For overview: [53], Section G.1, explicates our identifiability theory (Sections 4 to 7) for the exact inversion of genuinely discrete mixtures, while the (asymptotic) recovery of signals from samples of their discretely observed nonlinear mixtures is developed as part of Section E (Theorem E.1 in particular) and in [53], Section G.2.

[14] This orthogonality constraint imposes no loss of generality w.r.t. linear mixtures; cf. [53], Remark C.1(i).

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

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

X and its source S can be inverted (up to a minimal deviation) by optimizing some independence criterion ϕ over a set of candidate transformations Θ applied to X.

Partially driven by their applicability (8) to ICA, a variety of such criteria ϕ, referred to in [14] as contrast functions, have been developed.

The 'original' independence criterion ϕ_c proposed in [14] quantifies the statistical dependence between the components Y^i of a random vector $Y = (Y^1, \dots, Y^d)$ 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 rth-order (see [14] and cf. [53], equation (131), Section 3.2), that is, via the quantity

$$\phi_c(Y) := \sum_{j=2}^{r} \sum_{i_1,\dots,i_j}{}^{\times} \big(\kappa^Y_{i_1 \cdots i_j}\big)^2 \quad (r \geq 2), \tag{9}$$

where the inner sum runs over the indices $i_1, \dots, i_j \in [d]$ corresponding to (54).

Initially proposed in [14], the statistic (9) originates from a truncated Edgeworth-expansion of mutual information in terms of the standardized cumulants of its argument. A variety of alternatives to (9) soon followed, including kernel-based independence measures [3, 24], a variety of (quasi-) maximum-likelihood objectives, for example, [4, 44, 47], as well as mutual information and approximations thereof, for example, [8, 14, 29, 30].

While successfully achieving the separability of linear mixtures, Theorem 1 has its limitations: Being based on somewhat of a probabilistic curiosity (Rem. C.1(ii)), it might not be surprising that the characterisation (7) cannot be generalised to guarantee the recovery of independent scalar sources from substantially more general nonlinear mixtures of them [36]. Roughly speaking, the reason for this is that for a single random vector in $\mathbb{R}^d$, the statistical property of componental independence is too weak to characterise the nonlinear mixing transformations preserving this property as 'trivial' in a sense made precise by Definition 5 below. The following example illustrates this.[17]

EXAMPLE 3.1 (Comon's criterion (7) 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 'nontrivial' (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.[18]

4. 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. [14, 36] and Example 3.1. A key insight from [35] is to go for the middle ground (see Remark 4.2): if we demand the source S to have a 'nondegenerate temporal structure' and exploit this in a suitable manner, then the recovery of S from even its nonlinear mixtures is possible. To formalize such temporal statistical dependencies requires us to model the source S as a

[17] Example 3.1 is based on the 'Box–Muller transform', a well-known subroutine from computational statistics. For a systematic way of constructing 'unidentifiable' nonlinear mixtures of IC random vectors in $\mathbb{R}^d$, see [36].

[18] 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.

stochastic process. To this end, we use this section to briefly recall foundational notions from stochastic analysis (Section 4.1) and provide some basic notions and lemmas (Section 4.2) that we will use for our subsequent identifiability results in Section 5.

4.1. *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

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

where $\mathcal{B}(\mathcal{C}_d) = \sigma(\pi_t | t \in \mathbb{I})$ denotes the Borel σ-algebra on the Banach space $(\mathcal{C}_d, \|\cdot\|_\infty)$. Writing $S_t(\omega) \equiv (S_t^1(\omega), \ldots, S_t^d(\omega))^\intercal \in \mathbb{R}^d$ for each $\omega \in \Omega$, the scalar processes $S^i \equiv (S_t^i)_{t\in\mathbb{I}}$ ($i \in [d]$) are called *the component processes* or *the components* of $S \equiv (S^1, \ldots, S^d)$. We say that a stochastic process $S = (S^1, \ldots, S^d)$ *has independent components*, or that S *is IC*, if its distribution $\mathbb{P}_S := \mathbb{P} \circ S^{-1}$ satisfies the factor-identity[19]

$$\mathbb{P}_{(S^1,\ldots,S^d)} = \mathbb{P}_{S^1} \otimes \cdots \otimes \mathbb{P}_{S^d}. \tag{10}$$

REMARK 4.1. 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[20] S_t in $\mathbb{R}^d$ such that the map $S(\omega) : \mathbb{I} \ni t \mapsto S_t(\omega) \in \mathbb{R}^d$ is continuous for each $\omega \in \Omega$, for example, [50], Section II.27. Consequently (cf. also [53], Section C.2), condition (10) is equivalent to

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

for any finite selection of time-points $t_1^{(1)}, \ldots, t_{k_1}^{(1)}, \ldots, t_1^{(d)}, \ldots, t_{k_d}^{(d)} \in \mathbb{I}$, $k_1, \ldots, k_d \in \mathbb{N}_0$.

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 4.2 (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 $s, t \in \tilde{\mathbb{I}}$,

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')[21] 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.

[19] Strictly speaking, (10) reads $\mathbb{P}_{(S^1,\ldots,S^d)} = (\mathbb{P}_{S^1} \otimes \cdots \otimes \mathbb{P}_{S^d}) \circ \psi^{-1}$, 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 $\mathcal{C}_d \cong \mathcal{C}_1^{\times d}$ ([53], Remark C.2).

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

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

Remark 4.2 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 2nd-order finite-dimensional distributions, that is, on

$$\text{the distributions of} \quad \big\{(S_s, S_t)|(s,t) \in \Delta_2(\mathbb{I})\big\}, \tag{11}$$

where the index set $\Delta_2(\mathbb{I}) := \{(s,t) \in \mathbb{I}^{\times 2} | s < t\}$ is the (relatively) open 2-simplex on $\mathbb{I}\times\mathbb{I}$. In the following, we refer to (11) as the *temporal structure* of a stochastic process $S = (S_t)_{t\in\mathbb{I}}$.

The following is essential: 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 4.2(a), (b), then S is unidentifiable from X unless f is of a very specific form, for example, linear (cf. Theorem 1). Conversely, we will argue that if the source S has a temporal structure which is 'nondegenerate' (in some specified sense) and satisfies some additional regularity assumptions, then $S = (S^1, \ldots, S^d)$ will be identifiable from even its nonlinear mixtures up to a permutation and monotone scaling of its components S^i (Theorems 2, 3, 4).

4.2. *Stochastic processes as sources*: *Basic notions and assumptions*. Recall that the BSS problem (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 map 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 $\mathbb{R}^d$ which contains (the trace of) $\mathbb{P}$-almost each sample path of the process.[22]

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

$$D_Y = \overline{\bigcup_{t\in\mathbb{I}} \operatorname{supp}(Y_t)} \tag{12}$$

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 $\mathbb{R}^d$.

A few useful elementary properties of the set (12) are compiled in [53], Lemma C.1.

(Readers uncomfortable with (12) may for simplicity assume that $D_S = \mathbb{R}^d$ throughout.)

Given the above, we can describe the mixing transformation f mapping S to X via[23] (1) as

$$\text{a homeomorphism}[24] \quad f : D_S \to D_X, \tag{13}$$

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

We now introduce some 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 ς is C^k on some open neighbourhood of $x_0 \in \mathbb{R}^n$, then Z will be called *C^k-distributed around x_0*.

[22] 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.

[23] Recall that $X = f(S)$ means: $X_t = f(S_t)$ for each $t \in \mathbb{I}$.

[24] 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.

REMARK 4.3. We recall that for $\vartheta : \mathbb{R}^n \to \mathbb{R}^n$ a C^ℓ-diffeomorphism, $\ell \geq 1$, the classical transformation formula for densities asserts that the image $\tilde{Z} := \vartheta(Z)$ of a C^k-distributed random vector Z with density ς is itself $C^{k\wedge(\ell-1)}$-distributed with density $\tilde{\varsigma}$ given by

$$\tilde{\varsigma} = \left(\varsigma \circ \vartheta^{-1}\right) \cdot |\det J_{\vartheta^{-1}}|. \tag{14}$$

The action of the mixing transformation (13) on the source can be profitably captured by imposing the temporal structure (11) 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

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

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 $y_0 \in \mathbb{R}^{2d}$ and its density at y_0 is positive.

REMARK 4.4. 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
(ii) the source S is 'statistically trivial' in the sense of Remark 4.2.

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 (11) will be 'too simple' to guarantee nonlinear identifiability even for sources whose temporal structure (11) is not otherwise degenerate.

Generalising (i) and (ii) from 'spatial' to 'inter-temporal statistics', these exceptional types of joint distributions[25] will be named 'pseudo-Gaussian' and 'separable', respectively:

DEFINITION 4 (Non-Gaussian, (regularly) nonseparable). 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

$$\varsigma(x,y) = \varsigma_1(x) \cdot \varsigma_2(y) \cdot \exp\left(\pm\varsigma_3(x) \cdot \varsigma_3(y)\right) \tag{15}$$

holds on all of G; the function ς 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

$$\varsigma|_{\mathcal{O}} \text{ is not pseudo-Gaussian, for every open subset } \mathcal{O} \text{ of } G; \tag{16}$$

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

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

the notion of ς being *a.e. nonseparable* 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 nonseparable* if

$$\varsigma \text{ is a.e. nonseparable and } (\partial_x \partial_y \log \varsigma)|_{\Delta_{\tilde{U}}} \neq 0 \text{ a.e. on } \Delta_{\tilde{U}}, \tag{17}$$

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

[25] Distributional pathologies similar to Definition 4 have been first described in [35], see Section C.13. More specifically, the above notions of (strict) nonseparability and pseudo-Gaussianity generalise the notions [35], Def. 1 and Def. 2, respectively.

(Clearly, if ς is [strictly/a.e.] non-Gaussian then it is also [strictly/a.e.] nonseparable.)

It will be convenient for us to have an analytical characterisation of these 'pathological' types of densities at hand. Such a characterisation is provided by Lemma C.2 in [53], Section C.5.

REMARK 4.5.

(i) In light of [53], Lemma C.2(ii), the assumption of regular nonseparability can be regarded as a minimal extension of the above notion of strict nonseparability. The necessity of this extension will become clear in Section 5.2.

(ii) The log-derivative condition of (17) is nonvacuous as there are (strictly) nonseparable functions whose mixed log-derivatives vanish on the diagonal, see [53], Example C.6.

5. An identifiability theorem for nonlinearly mixed independent sources. We are now ready to present the core ideas behind our identifiability results for nonlinearly mixed time-dependent sources. Following an outline of our strategy (Section 5.1), we state and prove our main results (Sections 5.2 and 5.3) and conclude by comparing to related work.

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

$$X = f(S) \tag{18}$$

for a mixing transformation f which is C^2-invertible on some open superset of D_S.

Here, we say that $f : \mathbb{R}^d \to \mathbb{R}^d$ is *C^k-invertible* on an open set G of $\mathbb{R}^d$, in symbols: $f \in C^k(G)$, if the restriction $f|_G$ is a C^k-diffeomorphism (with C^k-inverse $f^{-1} : f(G) \to G$).

Throughout the rest of this paper, we operate under the following convenience assumption:

ASSUMPTION 1. For the source S in (18), every connected component of D_S is convex.

REMARK 5.1. While Assumption 1 holds for most conventional process models (including the examples in Sections 6 and 8 below), it can be dropped immediately at the only price that for a given realisation $\mathfrak{s} \equiv S(\omega)$ of S the scales $\alpha_i \equiv \alpha_i(\mathfrak{s})$ in (23) may[26] then vary with each maximally convex subset of D_S that the trace of $\mathfrak{s}$ passes through, see [53], Section C.12.[27]

5.1. *Overview.* Starting from (18) with the components $(S_t^1)_{t\in\mathbb{I}}, \dots, (S_t^d)_{t\in\mathbb{I}}$ of the source $S = (S_t^1, \dots, S_t^d)_{t\in\mathbb{I}}$ assumed mutually independent, we seek to identify S from X by exploiting the main dimensions of our model, space and time, via their statistical synthesis (11), the temporal structure of S. This is 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 (18) to an associated identity in the factor-space $\mathbb{R}^d \times \mathbb{R}^d$, namely

$$(X_s, X_t) = (f \times f)(S_s, S_t). \tag{19}$$

The lifted mixing identity (19), which directly involves the temporal structure (11) of the source, now allows for the following statistical comparison in the spirit of [35]:

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 fixed (s, t) together with their random combination

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

[26] Even for nonconvex geometries of D_S this price is not necessarily incurred, see [53], Example C.1.

[27] Even without Assumption 1 and except for the incurred $\mathfrak{s}$-dependence of the multiples α_i, the last identity in (23) below continues to hold as stated with the permutation P depending on only the connected component of D_S that the realisation $\mathfrak{s}$ of S is [almost surely] contained in, cf. [53], Theorem C.1.

for an equiprobable $\{0, 1\}$-valued random variable C independent of Y, Y^*. Combining (19) with the fact that S is IC, we obtain for the (deterministic) functional $L(Y, Y^*) := \psi \circ \rho$ with $\rho(y) := \mathbb{E}[C|\bar{Y} = y]$ and $\psi(p) := \log(p/(1-p))$ a contrast identity of the form

$$L(Y, Y^*) = R(f, (S_s, S_t)) \tag{20}$$

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, (20) relates X to S by way of the source's temporal structure (11).

Since the LHS $L \equiv L(Y, Y^*)$ in (20) is a function of the (joint) distribution of (Y, Y^*)—and thus of the mixture X—only, we for any alternative pair $(\tilde{f}, \tilde{S})$ with $\tilde{f}(\tilde{S}) = X$ and $\tilde{f} \in C^2$ and $\tilde{S}$ IC analogously obtain that $L(Y, Y^*) = \tilde{R}(\tilde{f}, (\tilde{S}_s, \tilde{S}_t))$ and hence

$$\tilde{R}(\tilde{f}, (\tilde{S}_s, \tilde{S}_t)) = R(f, (S_s, S_t)) \tag{21}$$

by (20), 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^2-invertibility of $\tilde{f}$, the IC-properties of both $\tilde{S}$ and S allow us to derive from (21) via (14) a (deterministic) system of functional equations

$$\Gamma(\varrho, (\tilde{S}_s, \tilde{S}_t), (S_s, S_t)) = 0 \quad \text{for } \varrho := (\tilde{f}^{-1} \circ f)\big|_{D_S}, \tag{22}$$

which involves the partial derivatives of the 'mixing residual' ϱ 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), that is, the temporal structure of S as specified by Definition 6, together with the required IC-property of $\tilde{S}$ are then sufficient to infer from (22) that the residual ϱ must be 'monomial' in the sense of Definition 5.

In other words, we obtained the following: Given a C^2-invertible map $\tilde{f}$, we have that

$$(\tilde{S}^1, \ldots, \tilde{S}^d) \equiv \tilde{S} = \tilde{f}^{-1}(X) = [P \circ (\alpha_1 \times \cdots \times \alpha_d)](S) \tag{23}$$

$$\begin{aligned} &\text{for some } P \in \mathrm{P}_d \text{ and monotone } \alpha_1, \ldots, \alpha_d \quad \textit{if and only if} \\ &\qquad \text{the component processes } \tilde{S}^1, \ldots, \tilde{S}^d \text{ are mutually independent.} \end{aligned} \tag{24}$$

The characterisation (24), formulated as Theorem 2, can thus be read as a natural extension of Comon's classical independence criterion (7) to nonlinear mixtures of IC stochastic processes whose temporal structure is sufficiently regular.

Additional source conditions that qualify S for the characterisation (24) are obtained by 'unfreezing' the above time pair $(s, t) \in \Delta_2(\mathbb{I})$; see Theorem 3 in Section 5.3.

Analogous to how Comon's criterion (7) became practically applicable by way of (8), our extended criterion (24) is clearly equivalent to the optimisation-based procedure (cf. Theorem 4)

$$\begin{aligned} &\Big[\operatorname*{arg\,min}_{\tilde{g} \in \Theta} \phi(\tilde{g}(X))\Big] \cdot X \subseteq \mathrm{DP}_d \cdot S \quad \text{for any } \phi : \mathcal{M}_1(\mathcal{C}_d) \to \mathbb{R}_+ \\ &\qquad \text{such that:} \quad \phi(\mu) = 0 \quad \text{iff} \quad \mu = \mu^1 \otimes \cdots \otimes \mu^d, \end{aligned} \tag{25}$$

for Θ some 'large enough' family of C^2-invertible candidate transformations, and DP_d a nonlinear analogon of the family of monomial matrices M_d (Definition 5).

Based on a 'moment-like' coordinate description for (the laws of) stochastic processes, we propose an efficiently computable such objective ϕ that generalises Comon's original contrast (9) from random vectors to stochastic processes (Section 7).

5.2. *Main theorem.* This section forms the heart of our identifiability theory.

We seek to recover the source $S = (S^1, \dots, S^d)$ from its nonlinear mixture X in (18) up to a minimal deviation, namely a permutation and monotone scaling of its coordinates $S^1, \dots, S^d$.

The following nonlinear analogue 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

$$\varrho\big|_{\tilde{G}} = P \circ (\alpha_1 \times \cdots \times \alpha_d) \quad \text{for } P \in \mathrm{P}_d \text{ and } \alpha_i \in \mathrm{Diff}^1\big(\pi_i(\tilde{G})\big).$$

(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:

$$\tilde{x} \in \mathrm{DP}_d \cdot x, \tag{26}$$

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.

Definition 4 describes analytical forms that need to be avoided by 'sufficiently many' of the distributions constituting its temporal structure (11) if the source S is to be identifiable from X up to a monomial transformation. Sources for which this is the case will be given the following label of regularity (or 'nondegeneracy').

DEFINITION 6 (α-contrastive). A continuous stochastic process $S \equiv (S_t^1, \dots, S_t^d)_{t\in\mathbb{I}}^\intercal$ in $\mathbb{R}^d$ with spatial support D_S will be called *α-contrastive* if S is IC and there is a collection of time-pairs $\mathcal{P}$ in $\Delta_2(\mathbb{I})$ and an associated collection $(D_{\mathfrak{p}})_{\mathfrak{p}\in\mathcal{P}}$ of open subsets of $\mathbb{R}^d$ 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 $\zeta^i_{s,t}$, and

$$\zeta^i_{s,t}\Big|_{D^{\times 2}_{(s,t)}} \quad \text{is regularly nonseparable for all } i \in [d], \text{ and}$$

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

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 $x = (x_v) \in \mathbb{R}^{2d}$.

Notice that the conditions in Definition 6(ii) reflect the classical pathologies (ii) and (i) from page 499. Further below we will see how the assumptions of Definition 6 are linked to related works (Section C.13) and that they are satisfied for a number of popular copula-based time series models (Section 6.1). Recall that the following result operates under Assumption 1.

THEOREM 2. *Let the process S in (18) be α-contrastive. Then for any transformation h which is C^2-invertible on some open superset of D_X, we have with probability one that*:

$$h(X) \in \mathrm{DP}_d \cdot S \quad \textit{if and only if} \quad h(X) \textit{ has independent components}. \tag{27}$$

PROOF. The 'only-if'-direction in (27) is clear, so we only need to show the converse implication. To this end, we in total prove the slightly stronger assertion that

$$\text{If } h(X) \text{ is IC and } D \equiv D_{(s,t)} \text{ as in Definition 6, then } \{J_{h\circ f}(u) | u \in D\} \subseteq \mathrm{M}_d\,. \tag{28}$$

Given (28) (and Definition 6(i)), the assertion (27) follows by way of [53], Lemma C.3(ii), and the fact that the trace of almost every realisation of S is contained in a connected component of D_S ([53], Lemma C.1(ii)) which in turn is convex by Assumption 1.

Let now $(s,t) \in \Delta_2(\mathbb{I})$ be as in Definition 6(ii), that is, 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 4.4).

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

$$Y := (X_s, X_t) \quad \text{and} \quad Y^* := (X_s, X^*_t). \tag{29}$$

For $C \sim \mathrm{Ber}(1/2)$ 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

$$\rho : \mathbb{R}^{2d} \to [0,1] \quad \text{given by } \rho(y) := \mathbb{E}[C|\bar{Y} = y]. \tag{30}$$

The function ρ then satisfies the following central equation.

LEMMA 1. *For μ the probability density of Y, and μ^* the probability density of Y^*,*

$$\psi \circ \rho = \log\mu - \log\mu^* \quad \textit{a.e. on } \tilde{D} := \mathrm{supp}(\mu) \tag{31}$$

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

The proof of Lemma 1 is given in [53], Section C.8. Recalling now that the components of S are mutually independent, we obtain from the transformation formula for densities (14) that for the inverse $g \equiv (g_1, \ldots, g_d) := f^{-1}$ and the density ζ^i_1 of S^i_s, resp. the density ζ^i_2 of S^i_t,

$$\log\mu - \log\mu^* = \sum_{i=1}^{d} \big[\log\zeta_i \circ (g_i \times g_i) - \log\zeta^i_1 \circ g_i(u) - \log\zeta^i_2 \circ g_i(v)\big] \tag{32}$$

almost everywhere on $\tilde{D} (= (f \times f)(\bar{D}))$. Using (31), it follows that

$$\psi \circ \rho = \sum_{i=1}^{d} P_i \circ (g_i \times g_i) \quad \text{for } P_i := \log\zeta_i - \sum_{v=1,2} \log\zeta^i_v \circ \pi_v. \tag{33}$$

Let now $h \equiv (h_1, \ldots, h_d) \in \mathrm{Diff}^2(\mathcal{O}_X)$, for some $\mathcal{O}_X \supseteq D_X$ open, be such that the process $\tilde{S} := h(X)$ has independent components. Using that the above function $\psi \circ \rho$ depends on the observable X only, we due to $(\tilde{S}_s, \tilde{S}_t) = (h \times h)(X_s, X_t)$ and (14) obtain that

$$\psi \circ \rho = \sum_{i=1}^{d} Q_i \circ (h_i \times h_i) \quad \text{a.e. on } \tilde{D} \tag{34}$$

analogous to (33), where the functions[28] $Q_i \in C^2(\bar{D}')$, $i \in [d]$, are given as

$$Q_i := \log\tilde{\zeta}_i - \sum_{v=1,2} \log\tilde{\zeta}^i_v \circ \pi_v \quad \text{with } \tilde{\zeta}_i := \frac{\mathrm{d}\mathbb{P}_{(\tilde{S}^i_s, \tilde{S}^i_t)}}{\mathrm{d}(u,v)} \text{ and } \tilde{\zeta}^i_v := \frac{\mathrm{d}\mathbb{P}_{\tilde{S}^i_{r_v}}}{\mathrm{d}u} \tag{35}$$

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

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

Note that the Q_i are indeed twice continuously differentiable: By (14) we have

$$\tilde{\zeta} = \frac{\mathrm{d}\mathbb{P}_{(\tilde{S}_s,\tilde{S}_t)}}{\mathrm{d}(u,v)} = \big|\det(J_\phi)\big| \cdot [\zeta \circ \phi] \in C^1\big(\bar{D}'\big)$$

for the C^2-density ζ and for $\phi := ((h \circ f) \times (h \circ f))^{-1} \in \mathrm{Diff}^2(\mathcal{O}_{\tilde{S}}^{\times 2}; \mathcal{O}_S^{\times 2})$, with $\mathcal{O}_{\tilde{S}} := h(\mathcal{O}_S)$; reading off the marginal densities $\tilde{\zeta}_i$, $\tilde{\zeta}_\nu^i$ (cf. [53], equation (C.4)), we see that the Jacobians appearing in (35) cancel out as they did in (32), giving us $Q_i \in C^2(\bar{D}')$ as desired.

Combining the identities (33) and (34) yields that

$$\sum_{i=1}^{d} Q_i \circ (h_i \times h_i) = \sum_{i=1}^{d} P_i \circ (g_i \times g_i) \tag{36}$$

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

Therefore, the desired implication (28)—and hence the assertion of the theorem (see the initial remarks of this proof)—holds if we can show (36) to imply that for $\varrho := h \circ f$ we have

$$\big\{J_\varrho(u) | u \in D\big\} \subseteq \mathrm{M}_d \quad \text{for each open } D \subseteq D_S \text{ as in Definition 6(ii)}, \tag{37}$$

that is, for any (nonempty) open subset D of $\mathbb{R}^d$ for which $\zeta^i\big|_{D^{\times 2}}$ is regularly nonseparable 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 (37) from (36). To this end, notice that since (36) 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 (36) is equivalent to $Q \circ (\varrho \times \varrho) = P$, that is, to the ($D_\zeta := \{\zeta > 0\}$–everywhere) identity[29]

$$\sum_{i=1}^{d} Q_i \circ (\varrho_i \times \varrho_i) = \sum_{i=1}^{d} P_i. \tag{38}$$

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 (38) to arrive at the identities

$$\sum_{i=1}^{d} \big[q_i \circ (\varrho_i \times \varrho_i)\big] \cdot \partial_{u_j}\varrho_i \cdot \partial_{v_k}\varrho_i = \sum_{i=1}^{d} \xi_i \cdot \delta_{ijk} \quad \big(j,k \in [d]\big), \tag{39}$$

where the ϱ_i are the components of (37) and the functions q_i and ξ_i are given as

$$q_i := \partial_{u_i}\partial_{v_i} Q_i \quad \text{and} \quad \xi_i := \partial_{u_i}\partial_{v_i} P_i = \partial_{u_i}\partial_{v_i} \log \zeta_i, \tag{40}$$

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 (33) and (35).) Observe now that the system of equations (39) can be equivalently expressed as the congruence relation

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

for J_ϱ the Jacobian of ϱ and for Λ_q, Λ_ξ defined as the matrix-valued functions

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

[29] 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 (38).

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

$$\Lambda_q = B_\varrho^\intercal \cdot \Lambda_\xi \cdot B_\varrho \quad \text{on } D_\zeta, \text{ for } B_\varrho := J_\varrho^{-1}. \tag{41}$$

Since $B_\varrho = J_{\varrho^{-1}} \circ \varrho$ by the inverse function theorem, the matrix-valued function B_ϱ is clearly continuous. Hence[30] we can apply [53], Lemma C.4, to from (41) and the assumptions of Definition 6(ii) obtain as desired that

$$\{J_\varrho(u) | u \in D\} \subseteq \mathrm{M}_d\,. \tag{42}$$

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 nonseparable for each $i \in [d]$ and a.e. non-Gaussian for all but at most one $i \in [d]$ (Definition 6(ii)), [53], Lemma C.4, is clearly applicable to the system (41), providing (42) as required. But since the above set D was chosen without further restrictions, (42) amounts to (37) and hence proves Theorem 2 as desired. □

5.3. *An extension to sources of alternative temporal structures.* We can generalise the strategy behind Theorem 2 by 'unfreezing' its usage of the temporal structure (11), that is by allowing the considered time-pairs (s,t) to 'vary more freely' across $\Delta_2(\mathbb{I})$; see [53], Lemma C.5. This qualifies additional source classes for nonlinear identification via the characterisation (27). As before)cf. [53], (41)), the technical key for this is to have the Jacobian of the mixing residual serve as change of basis for a source-dependent matrix function with nondegenerate eigenspectrum. The next definition describes two sufficient conditions for this.

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

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

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

  an open subset $\tilde{U}$ of U and $\mathfrak{p} \equiv (s,t), \mathfrak{p}' \in \Delta_2(\mathbb{I})$ such that, for all $i \in [d]$,

  the density $\zeta^i_{s,t}$ of (S^i_s, S^i_t), likewise $\zeta^i_{\mathfrak{p}'}$, exists with $\zeta^i_{\mathfrak{p}}, \zeta^i_{\mathfrak{p}'} \in C^2(\tilde{U}^{\times 2})$ and

  $$\xi^{i|\tilde{U}}_{s,t} := \left[\partial_{x_i}\partial_{x_{i+d}} \log \zeta^i_{s,t}\right] \circ \iota_{\tilde{U}} \neq 0 \quad \text{and} \quad \xi^{i|\tilde{U}}_{\mathfrak{p}'} \neq 0 \quad \text{(a.e.)}, \quad \text{and} \tag{43}$$

  $$\xi^{i|\tilde{U}}_{\mathfrak{p}'} \notin \langle \xi^{i|\tilde{U}}_{\mathfrak{p}} \rangle_{\mathbb{R}} := \{c \cdot \xi^{i|\tilde{U}}_{\mathfrak{p}} | c \in \mathbb{R}\} \tag{44}$$

  with $\iota_{\tilde{U}} : \tilde{U} \ni u \mapsto (u,u) \in \Delta_{\tilde{U}}$ and both $U, \tilde{U}$ nonempty;[31]
- *γ-contrastive* if there is a dense open subset $\mathcal{U}$ of D_S for which the following holds:

  for each $u \in \mathcal{U}$ there exists $(v, \mathfrak{p}_0, \mathfrak{p}_1, \mathfrak{p}_2) \in \mathbb{R}^d \times \Delta_2(\mathbb{I})^{\times 3}$ such that

  S is C^2-regular around $(\mathfrak{p}_0, (u,v))$, $(\mathfrak{p}_1, (u,u))$ and $(\mathfrak{p}_2, (v,v))$, and

  $$\left(\psi\left(\xi^i_{\mathfrak{p}_0}(u,v), \xi^i_{\mathfrak{p}_1}(u,u), \xi^i_{\mathfrak{p}_2}(v,v)\right)\right)_{i\in[d]} \in \left(\mathbb{R}^d \setminus \nabla^\times\right), \tag{45}$$

  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).

[30] Notice that $D^{\times 2} \subset \bar{D} \equiv \mathrm{supp}(\zeta)$ (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. nonseparable (and hence a.e. nonzero in particular).

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

Notice that every α-contrastive process is also γ-contrastive (for $\mathfrak{p}_0 = \mathfrak{p}_1 = \mathfrak{p}_2$, as the proof of Theorem 2 shows), while β-contrastivity does not imply—nor is it implied by—either α- or γ-contrastivity. We will see in Section 6.2 that the assumptions of γ-contrastivity are satisfied for a number of popular stochastic processes.

Recall that the following theorem operates under Assumption 1.

THEOREM 3. *Let the process S in (18) be β- or γ-contrastive. Then for any transformation h which is C^2-invertible on an open superset of D_X, we have with probab. one that*

$$h(X) \in \mathrm{DP}_d \cdot S \quad \textit{if and only if} \quad h(X) \textit{ has independent components}. \tag{46}$$

A comparison to related work can be found in Section C.13 of the Supplementary Material [53].

6. Examples of applicable sources. The statistical nondegeneracy assumptions of α-, β- or γ-contrastivity hold for a number of well-established models for stochastic signals, among them most popular 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 α-contrastive.* It is well-known (e.g., [45], Section 2.10, [18]) that the temporal structure (11) of a scalar stochastic process $S = (S_t)_{t\in\mathbb{I}}$ can be given an analytical representation of the form

$$\zeta_{s,t}(x, y) = \zeta_s(x)\zeta_t(y) \cdot c_{s,t}\big(F_s^S(x), F_t^S(y)\big) \quad \big((s, t) \in \Delta_2(\mathbb{I})\big), \tag{47}$$

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 ζ_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 $S \equiv (S_t)_{t\in\mathbb{I}} \equiv (S^1, \dots, S^d)$ be an IC stochastic process in $\mathbb{R}^d$ such that S_t admits a C^2-density ζ_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,[32] it holds that the copula densities $\{c^i_{s,t} | (s, t) \in \mathcal{P}\}$ of S^i (cf. (47)) are such that*

$$c^i_{s,t} \textit{ are positive and strictly non-Gaussian and } \partial_x\partial_y \log c^i_{s,t} \textit{ vanishes nowhere}, \tag{48}$$

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

A popular approach in finance, insurance economy and other fields is to read (47) 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') such that $\zeta_r \equiv \zeta$ with cdf F_ζ for each $r \in \mathcal{I}$, and $D_S = \mathrm{supp}(\zeta)$ and $c_{s,t} \equiv c_\theta$ uniformly parametrized for all $(s, t) \in \mathcal{P} := \mathcal{I}^{\times 2} \cap \Delta_2(\mathbb{I})$; see, for example, [10], Section 2, [21]:

$$\zeta_{s,t}(x, y) = \zeta(x)\zeta(y) \cdot c_\theta\big(F_\zeta(x), F_\zeta(y)\big), \quad (s, t) \in \mathcal{P}. \tag{49}$$

We verify exemplarily that a source $S = (S^1, \dots, S^d)$ in $\mathbb{R}^d$ whose components S^i are modelled according to (49) is α-contrastive for a number of popular copula densities c_θ.

COROLLARY 2. *Let $S = (S^1, \dots, S^d)$ be a stochastic process whose independent components S^i are modelled according to (49) for each $i \in [d]$ with copula-density c_i belonging to one of the following popular classes*:

[32] The existence of such a set $\mathcal{P}$ is guaranteed by [53], Lemma C.1(v).

(i) (Clayton)

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

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

(ii) (Gumbel)

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

(iii) (Frank)

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

Then S is α-contrastive.

PROOF. This is a direct consequence of Proposition 1 upon checking that each of the copula densities (i), (ii) and (iii) satisfies (48). This, however, follows from inspection and a straightforward computational verification. □

6.2. *Popular Gaussian processes and geometric Brownian motion are γ-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 $\kappa=(\kappa^{ij})$. We assume that any pair (μ,κ) we consider in the following is such that each process $S\sim\mathcal{GP}(\mu,\kappa)$ admits a version with continuous sample paths.

We demonstrate contrastivity for a number of popular Gaussian processes.

PROPOSITION 2. *Let $S=(S_t^1,\dots,S_t^d)_{t\in\mathbb{I}}$ 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 γ-contrastive in each of these four classical cases.*

(i) *For each $i\in[d]$, the componental autocovariance functions* (2) *of S are of the form*

$$\kappa^i(s,t) = \exp\left(-\left[\frac{|t-s|}{\alpha_i}\right]^{\gamma_i}\right)$$

with $\gamma\equiv(\gamma_i)_{i\in[d]}\in(0,2]^d$ and $\alpha\equiv(\alpha_i)_{i\in[d]}\in(\mathbb{R}_\times)^{\times d}\setminus\mathcal{N}_\gamma$, where $\mathcal{N}_\gamma\subset\mathbb{R}^d$ is a Lebesgue nullset defined in the proof below.[33]

(ii) *Each component process S^i of S is an Ornstein–Uhlenbeck process*

$$\mathrm{d}S_t^i = \theta_i\cdot\big(\mu_i - S_t^i\big)\,\mathrm{d}t + \sigma_i\,\mathrm{d}B_t^i,\ S_0^i = a_i, \quad \big(i\in[d]\big) \tag{50}$$

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

(iii) *The component processes of S are fractional Brownian motions with pairwise distinct Hurst indices, that is their autocovariance functions* (2) *take the form*

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

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

[33] This includes the family of γ-exponential processes; cf. [48], Section 4.2 (pp. 84 ff.).

(iv) *Denoting* $s \wedge t := \min(s,t)$, *the autocovariance functions* (2) *of the* S^i *are of the form*

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

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

$$\mathrm{d}S_t^i = \mu_i(t)\,\mathrm{d}t + \sigma_i(t)\,\mathrm{d}B_t^i \quad \textit{for each } i \in [d]$$

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* $r_0, r_1 \in \mathbb{I}$.

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

PROPOSITION 3. *Let* $S = (S_t)_{t\geq 0} = (S^1, \dots, S^d)$ *be an IC geometric Brownian motion in* $\mathbb{R}^d$, *that is, suppose that there is a standard Brownian motion* $B = (B_t^1, \dots, B_t^d)_{t\geq 0}$ *such that*

$$\mathrm{d}S_t^i = S_t^i \cdot \big(\mu_i(t)\,\mathrm{d}t + \sigma_i(t)\,\mathrm{d}B_t^i\big), \qquad S_0^i = s_0^i \quad \big(i \in [d]\big)$$

for some $s_0^i > 0$ *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* γ*-contrastive if there are* $r_0, r_1 \in \mathbb{I}$ *for which the numbers* $\{\sigma_i^2(r_0) \cdot \sigma_j^2(r_1) | (i,j) \in [d] \times [d]\}$ *are pairwise distinct.*

7. Signature cumulants as contrast function. This section uses the identifiability results of Section 5 to reformulate the problem of nonlinear blind source separation as an optimisation task in the spirit of Corollary 1. Central to this is the concept of an IC-characterising contrast function on stochastic processes. We propose such a function by means of signature cumulants, which we introduce as a natural extension of classical (multivariate) cumulants to multidimensional stochastic processes.

REMARK 7.1. In this section, we restrict our exposition to stochastic processes whose sample paths are smooth (i.e., of bounded variation[34]), and further assume that the expected signature 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 [23, 42] and [11, 12], respectively. Let further $\mathbb{I} = [0, 1]$ w.l.o.g.

7.1. *Signature cumulants.* Many results in statistics, including Corollary 1 via (9), 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 [11] provides a natural generalisation of the classical moment sequence. Similar to classical moments, these signature moments form multi-indexed collections of numbers that can characterize the laws of stochastic processes. Similar still, upon their 'logarithmic compression' these number

[34]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 definition (135) and Section F.7 in the Supplementary Material [53].

collections give rise to signature cumulants that quantify the statistical dependencies within multidimensional stochastic processes (that is, between their coordinates and over time).

Denote by $[d]^\star := \bigcup_{m\geq 0}[d]^{\times m}$ the set of all multi-indices[35] with entries in $[d] = \{1,\dots,d\}$.

DEFINITION 8 (Expected signature). For $Y = (Y_t^1,\dots,Y_t^d)_{t\in[0,1]}$ 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

$$\sigma_{\boldsymbol{i}}(Y) := \mathbb{E}\left[\int_{0\leq t_1\leq t_2\leq\cdots\leq t_m\leq 1} \mathrm{d}Y_{t_1}^{i_1}\,\mathrm{d}Y_{t_2}^{i_2}\cdots\mathrm{d}Y_{t_m}^{i_m}\right] \quad \text{for } \boldsymbol{i} = (i_1,\dots,i_m), \tag{51}$$

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

The expected signature is to a stochastic process roughly what the sequence of moments is to a vector-valued random variable, and analogous to the case of classical moments, for many statistical purposes the concept of cumulants is better suited. This leads to the notion of signature cumulants [6] below. (See Remark 7.2 and [53], Sections C.18 and D, for details.)

DEFINITION 9 (Signature cumulants). For Y a stochastic process in $\mathbb{R}^d$ with sample-paths of bounded variation, the collection of real numbers[36]

$$\big(\kappa_{\boldsymbol{i}}(Y)\big)_{\boldsymbol{i}\in[d]^\star} := \log\big[\mathfrak{S}(Y)\big] \tag{52}$$

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

$$\bar{\kappa}_{\boldsymbol{i}}(Y) := \frac{\kappa_{\boldsymbol{i}}(Y)}{\kappa_{11}(Y)^{\eta_1(\boldsymbol{i})/2}\cdot\ldots\cdot\kappa_{dd}(Y)^{\eta_d(\boldsymbol{i})/2}} \quad \text{for } \boldsymbol{i} = (i_1,\dots,i_m)\in[d]^\star, \tag{53}$$

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

REMARK 7.2. The signature cumulant of a process Y is an efficiently computable [41], informationally condensed and hierarchically graded (cf. [53], Section D.2) compression of the statistical information contained in [the distribution of] Y (cf. [53], Sections C.18, D.1), which enjoys a broad variety of excellent practical and theoretical features [12]. Just as for standardized classical cumulants, the normalisation (53) brings the additional benefit of scale invariance which facilitates our below usage of signature cumulants as a contrast function.

7.2. *Signature contrasts for nonlinear ICA.* Similar to how classical cumulants are traditional in linear ICA (cf. page 496), the usage of signature cumulants in our present ICA-context is due to the following observation: Recall that a random vector Y in $\mathbb{R}^d$ has independent components if and only if all of its cross-cumulants vanish, that is iff, in the notation of (9) and for $*$ the concatenation of indices,

$$\kappa_{\boldsymbol{q}}^Y = 0 \quad \text{for all } \boldsymbol{q}\in\bigsqcup_{k=2}^{d}\big\{\boldsymbol{i}*\boldsymbol{j}\,|\,\boldsymbol{i}\in[k-1]^\star\setminus\{\varnothing\},\,\boldsymbol{j}\in\{k\}^\star\setminus\{\varnothing\}\big\}. \tag{54}$$

Now in the same way that the expected signature generalises the classical concept of moments (cf. [53], Remark C.18), it was shown in [6] that signature cumulants generalise this classical relation (54) to an algebraic characterisation of statistical independence between [the

[35] We define $[d]^{\times 0} := \{\varnothing\}$ with $\varnothing$ the empty set, and let $\{k\}^\star := \bigcup_{m\geq 0}\{k\}^{\times m} (= \{\varnothing, k, kk, kkk, \ldots\})$.

[36] The log in (52) denotes the logarithm on the space of formal power series, see [53], Section D.2.1, and [6].

components of] stochastic processes, cf. also [53], Remark D.3. This is particularly useful in our context as it yields a natural and explicitly computable contrast function for path-valued random variables (Proposition 4) as desired for nonlinear ICA.

Algebraically (cf. [53], Remark D.2), the (52)-based extension of the characterisation (54) to stochastic processes requires us to replace the simple operation $*$ of index concatenation by a slightly more involved combinatorial operation on $[d]^\star$. This operation is defined next.

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

The *shuffle product* of two multi-indices $\boldsymbol{i} = (i_1, \ldots, i_m)$ and $\boldsymbol{j} = (i_{m+1}, \ldots, i_{m+n})$ in $[d]^\star$ is defined as the element of $[d]^\star_+$ which is given by

$$\boldsymbol{i} \sqcup\!\sqcup \boldsymbol{j} := \sum_{\tau} (i_{\tau(1)}, \ldots, i_{\tau(m+n)}) \in [d]^\star_+, \tag{55}$$

where the sum is taken over the family of permutations

$$\big\{\tau \in S_{m+n} | \tau(1) < \cdots < \tau(m) \text{ and } \tau(m+1) < \cdots < \tau(m+n)\big\}. \tag{56}$$

This enables us to formulate the following central observation.

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

$$\bar{\kappa}_{\mathrm{IC}}(Y) := \sum_{k=2}^{d} \sum_{\boldsymbol{q} \in \mathcal{W}_k} \bar{\kappa}_{\boldsymbol{q}}(Y)^2 = 0, \tag{57}$$

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

PROOF. Observe that the component processes $Y^1, \ldots, Y^d$ are mutually independent iff

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

The asserted characterisation is a direct consequence of this and [6], Theorem 1.2(iii). □

We may now combine Proposition 4 with Theorems 2 and 3 to obtain the following instance of (25) for the inversion '$X \mapsto S$' that is desired in (2) (cf. Corollary 1).

(Recall Remark 7.1 for the well-definedness of the signature statistics featured in (58).)

THEOREM 4. *Let the process S in (18) be α-, β- or γ-contrastive with sample-paths of bounded variation. Then it holds with probability one that*

$$\Big[\arg\min_{h \in \Theta} \bar{\kappa}_{\mathrm{IC}}\big(h(X)\big)\Big] \cdot X \subseteq \mathrm{DP}_d \cdot S \tag{58}$$

for any family of transformations $\Theta \subseteq C^{2,2}(D_X)$ with $\Theta \cap (\mathrm{DP}_d(D_S) \cdot f^{-1})\big|_{D_X} \neq \varnothing$.

This theorem states that the initial problem (2) of nonlinear blind source separation can be reformulated as a problem of optimisation-based function approximation. More specifically, statement (58) says that the desired demixing transformations of the data can be found as minimizers of the energy-like functional (57). In [53], Section C.20, we complement Theorem 4 with a few practical and contextualising remarks, and in Section E of [53] we derive from Theorem 4 a practical blind inversion algorithm with strong consistency guarantees.

8. Numerical experiments. We present a series of numerical examples to illustrate the practical applicability of our ICA method on discrete- and continuous-time signals. A complete account of the following experiments and results, including their full parameter settings and all relevant implementations and estimates, is provided on the public repository [52].

8.1. *A performance index for nonlinear ICA.* As before, we consider stochastic processes X and S in continuous or discrete[37] time such that

$$X = f(S) \quad \text{for some } f \in C^{2,2}(D_S). \tag{59}$$

In order to assess how close an estimate $\hat{S} \equiv \hat{S}(X)$ of S is to the true source S in (59), we propose to quantify the distance between $\hat{S}$ and the orbit[38] $\mathrm{DP}_d \cdot S$ by way of the following intuitive[39] performance statistic (cf. [53], Remark C.3(iii), for applicability).

DEFINITION 10 (Monomial discordance). Given two time series $\mathcal{X} := (X_t^1, \dots, X_t^d)_{t\in\mathcal{I}}$ and $\mathcal{Y} := (Y_t^1, \dots, Y_t^d)_{t\in\mathcal{I}}$ in $\mathbb{R}^d$ for $\mathcal{I}$ finite, define the *concordance matrix of* $(\mathcal{X}, \mathcal{Y})$ as

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

where ρ_{K} is the Kendall[40] rank correlation coefficient. Furthermore, we define

$$\varrho(\mathcal{X}, \mathcal{Y}) := \frac{1}{\sqrt{d(d-1)}} \min_{P\in\mathrm{P}_d} \left\| C(\mathcal{X}, \mathcal{Y}) - P \right\|_2 \in [0,1] \tag{60}$$

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

PROPOSITION 5. *Let X and S be as in (59) 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 ϱ as in (60), we have that*

$$\big(h(X_t)\big)_{t\in\mathcal{I}} \in \mathrm{DP}_d \cdot (S_t)_{t\in\mathcal{I}} \quad \textit{iff} \quad \varrho\big(\big(h(X_t)\big)_{t\in\mathcal{I}}, (S_t)_{t\in\mathcal{I}}\big) = 0. \tag{61}$$

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 truncated approximations (142) of the above contrast $\bar{\kappa}_{\mathrm{IC}}$ will be denoted ϕ_{m_0}.

8.2. *Nonlinear mixings with explicitly parametrized inverses.* First, we consider three families of C^2-diffeomorphisms on the plane whose inverses are explicitly parametrized.

More specifically: We sample two types of source processes in $\mathbb{R}^2$, namely: an IC Ornstein–Uhlenbeck process $S_{\mathrm{ou}} = (S_{\mathrm{ou}}^1, S_{\mathrm{ou}}^2)$, and an IC copula-based time-series $S_{\mathrm{cy}} = (S_{\mathrm{cy}}^1, S_{\mathrm{cy}}^2)$ that follows the dependence model (47).[41] Both S_{ou} and S_{cy} are contrastive by Proposition 2(ii) and Corollary 2(i), respectively.[42] These sources are first mapped to the

[37] See [53], Section C.3(iii), for an explicated treatment of the latter.

[38] See (26) for notation, and recall that the elements of $\mathrm{DP}_d \cdot S$ are in a minimal distance from S.

[39] Recall the classical facts (e.g., [20]) that Kendall's (and Spearman's) rank correlation coefficient ρ_{K} attains its extreme values ± 1 iff one of its arguments is a monotone transformation of the other, with $\rho_{\mathrm{K}}(U, V) = 0$ if its arguments U and V are independent.

[40] If preferred, ρ_{K} might alternatively be chosen as Spearman's rank correlation coefficient.

[41] With $F_t^{S_{\mathrm{cy}}^i}$ chosen as the cdf of $\mathcal{N}(0,1)$ and c chosen as the Clayton-density (cf. Proposition 1(i)).

[42] Note further that the Ornstein–Uhlenbeck processes S_{ou}, while continuous-time by nature, are processed as discrete-time observations according to their classical Euler–Maruyama approximation. The copula-based time-series S_{cy}, on the other hand, are simulated at their observation frequency and thus showcase the applicability of our method to discrete-time signals (in accordance with Section G in [53]).

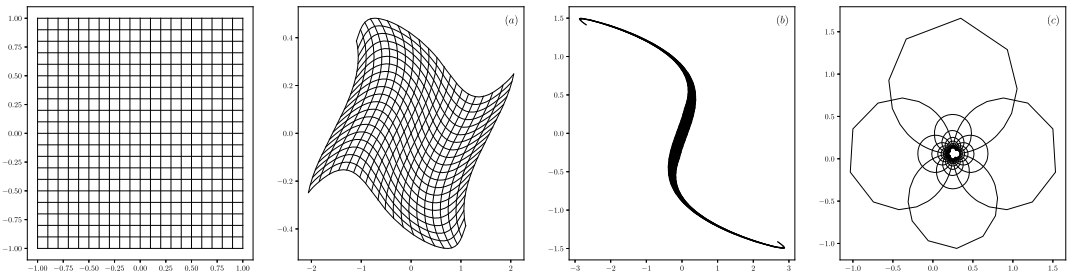

FIG. 2. *The image of the square* $[-1,1]^2$ (*leftmost*) *under three increasingly nonlinear mixing transformations* f_1, f_2, f_3, *namely conjugates of the Hénon map* (f_1 *and* f_2; *panels* (*a*) *and* (*b*), *respectively*) *and of the Möbius transformation* (f_3; *panel* (*c*)).

square $[-1,1]^2$ upon centering and scaling them to unit amplitude, and then transformed by one of three mixing maps $f_j : \mathbb{R}^2 \to \mathbb{R}^2$ $(j = 1,2,3)$ with increasing degree of 'nonlinearity', see Figure 2. (For an explicit definition of the f_j, see [52].) [53], Figure 2, shows the spatial trace of a sample realisation of S_{ou} and S_{cy} (panels (a) and (b)) next to an excerpt of the time-parametrised components of these realisations, together with their nonlinear mixtures $X_\eta^{(j)} := f_j(S_\eta)$ for $j = 1,2,3$ and $\eta =$ 'ou' (panels (c), (e), (g)) and $\eta =$ 'cy' (panels (d), (f), (h)).

Each of the 'true' inverses $g^j := f_j^{-1}$ $(j = 1,2,3)$ are contained in an (injectively parametrized) family $\Theta_j \equiv \{g_\theta^j \in C^2(\mathbb{R}^2) | \theta \in \tilde{\Theta}_j\}$ of candidate de-mixing transformations g_θ^j, where $\tilde{\Theta}_j \subseteq \mathbb{R}^2$ is some open parameter set. On these parameter sets, we consider the data-based objective functions

$$\Phi_\eta^j : \tilde{\Theta}_j \to \mathbb{R}, \qquad \theta \mapsto \phi_{m_j}\big(g_\theta^j\big(X_\eta^{(j)}\big)\big), \tag{62}$$

with $\phi_m := \bar{\kappa}_{\mathrm{IC}}^{[m]}$ as in (142) and capped at the cumulant orders $m_1 = m_2 = m_3 = 6$, and compare the topography of the functions (62) to that of the monotone discordances

$$\delta_\eta^j : \tilde{\Theta}_j \to \mathbb{R}, \qquad \theta \mapsto \varrho\big(g_\theta^j\big(X_\eta^{(j)}\big), S_\eta\big) \quad \text{(cf. (60))}. \tag{63}$$

Recall that the latter are 'distance functions' that quantify how much a candidate source estimate $\hat{S}_\eta^\theta := g_\theta^j(X_\eta^{(j)})$ deviates [from the monomial orbit $\mathrm{DP}_d \cdot S_\eta$ of S_η, that is] from the true source S_η up to order and monotone scaling of its components.

The results are displayed in the first three columns of Figure 5, with the 'estimator's view' $\Phi_{\mathrm{ou}}^{1|2|3}$ of the demixing performance shown in the top-row panels and the 'true view' $\delta_{\mathrm{ou}}^{1|2|3}$ of

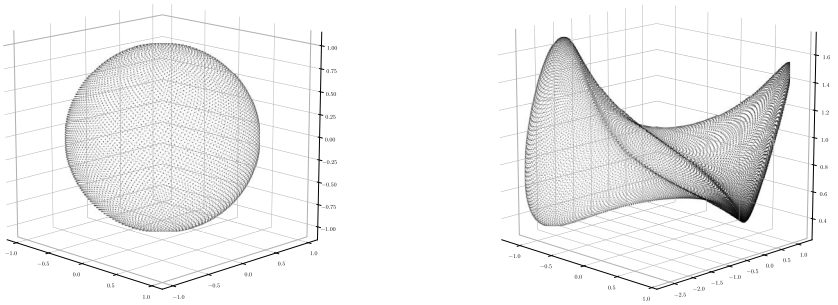

FIG. 3. *Illustration of the three-dimensional mixing transform* $f_4 : \mathbb{R}^3 \to \mathbb{R}^3$ *via its action* $f_4(S^2)$ (*right panel*) *on the* 2*-sphere* $S^2 \equiv \{x \in \mathbb{R}^3 | |x| = 1\}$ (*left panel*).

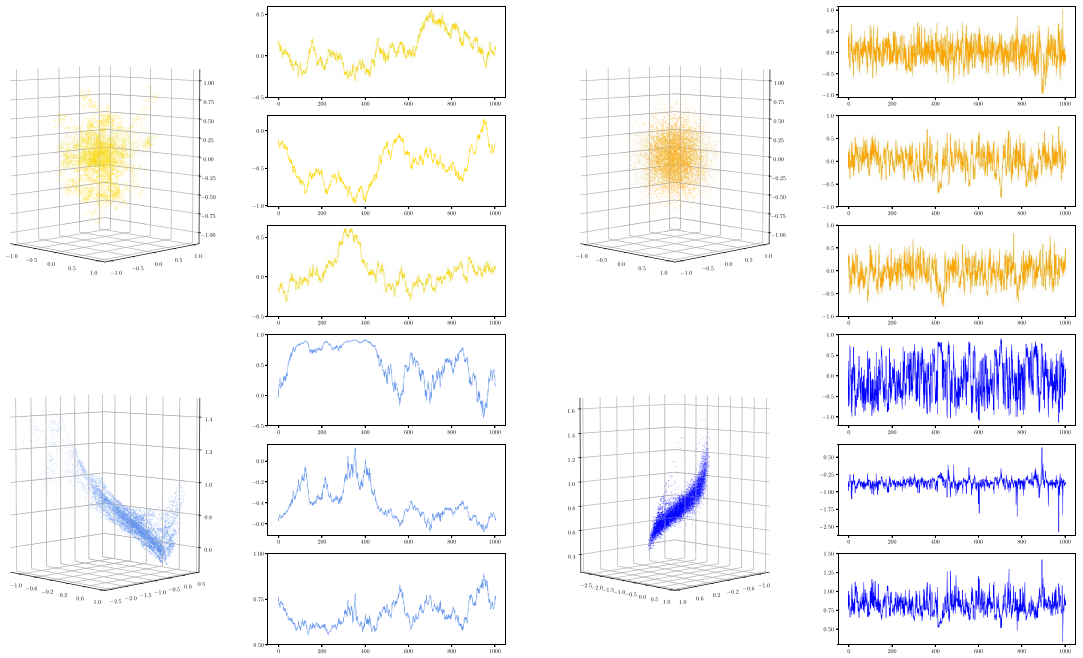

FIG. 4. *Spatial trace and sampled components of a three-dimensional IC Ornstein–Uhlenbeck process* $\tilde{S}_{\mathrm{ou}}$ (*top left*) *and an IC copula-based time-series model* $\tilde{S}_{\mathrm{cy}}$ (*top right*) *and their respective nonlinear mixtures* $f_4(\tilde{S}_{\mathrm{ou}})$ (*bottom left*) *and* $f_4(\tilde{S}_{\mathrm{cy}})$ (*bottom right*).

the demixing performance shown in the bottom-row panels.[43] This shows clearly that within the given families Θ_j of candidate transformations, those candidate nonlinearities which map the data $X_\eta^{(j)}$ to a best-approximation of its source S_η are precisely those that minimise the contrast (62), 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 3 and 4. The results, obtained for a contrast capped at cumulant order $m_4 = 7$ and shown as the rightmost column of Figure 5, are again affirmative of Theorem 4.

8.3. *Nonlinear mixings with inverses approximated by neural networks.* The practical applicability of our ICA-method is illustrated by running the optimisation (58) 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 (see [52] for details). The resulting mixtures $X^{(1)}$ and $X^{(2)}$ are then passed on to candidate demixing-nonlinearities $g_\theta^\nu \in \Theta_\nu$ which are given as elements of the parametrized families

$$\Theta_\nu := \{g_\theta^\nu : \mathbb{R}^{2\nu} \to \mathbb{R}^{2\nu} | g_\theta^\nu \text{ is an ANN with weights } \theta \in \tilde{\Theta}_\nu\} \quad (\nu = 1, 2). \tag{64}$$

Here, the families of transformations Θ_ν are spanned by the various configurations of some artificial neural network (ANN) instantiated over weight-vectors θ which are chosen from a given parameter set $\tilde{\Theta}_\nu$ in $\mathbb{R}^{m_\nu}$, where the number of weights m_ν is part of the predefined architecture of the ANN. Given these candidate-inverses, the optimisations (58) are run by

$$\text{minimizing} \quad \tilde{\Theta}_\nu \ni \theta \longmapsto \phi_{m_\nu}\big(g_\theta^\nu(X^{(\nu)})\big), \tag{65}$$

[43] For brevity, Figure 5 shows the case $\eta = \mathrm{ou}$ only; the results for the case $\eta = \mathrm{cy}$ can be found in [52].

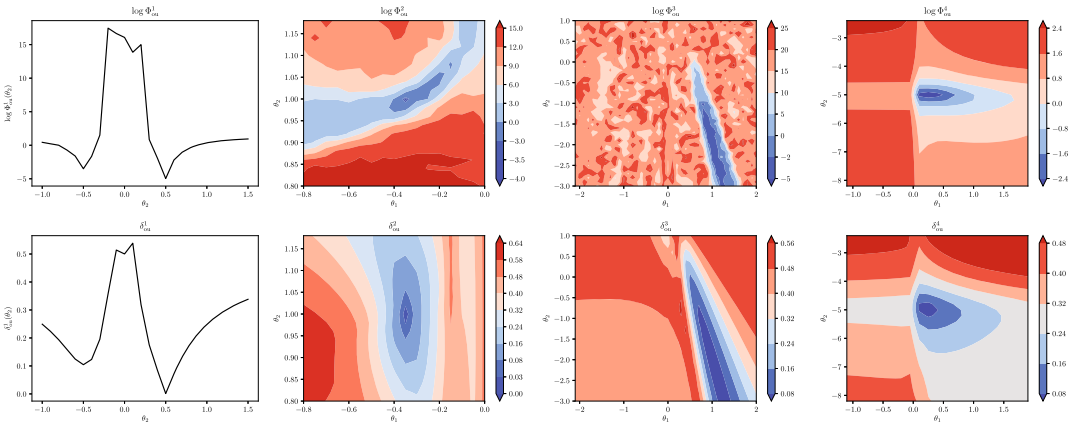

FIG. 5. *Contour plot (leftmost column) and heatmaps of the log-transformed contrast functions (62) (top row) and of the associated discordance functions (63) (bottom row) for the mixings* $X_{\mathrm{ou}}^{(j)} = f_j(S_{\mathrm{ou}})$, $j = 1, \ldots, 4$. *The parameters* $\theta_\star^{(j)} \equiv (\theta_1^{(j)}, \theta_2^{(j)})$ *of the true inverses* $f_j^{-1} \equiv g_{\theta_\star^{(j)}}^j \in \Theta_j$ *are* $\theta_\star^{(1)} = 0.5$,[44] $\theta_\star^{(2)} = (-0.35, 1)$, $\theta_\star^{(3)} = (1, -2)$, *and* $\theta_\star^{(4)} = (0.2, -5)$.

that is, by training each constituent ANN (64) with the truncated contrast $\phi_{m_\nu} = \bar{\kappa}_{\mathrm{IC}}^{[m_\nu]}$ (cf. (142)) as its loss function, where the optimization steps are computed via backpropagation along the weights of the ANN. For technical details behind the setups of (64) and (65), see [52, 53].

For the case $\nu = 1$ we applied the mixing transformation depicted in Figure 6 (leftmost panel), and for the case $\nu = 2$ we followed the simulations of [34, 35] in using as a mixing transformation an invertible feedforward-neural network with four-nodal in- and output layers and two four-nodal hidden layers with tanh activation each.

Denoting by $\theta_\nu^* \in \tilde{\Theta}_\nu$ the (local) optimum obtained by the minimisation of the objective (65) and setting $\hat{S}^{(\nu)} := g_{\theta_\nu^*}^\nu(X^{(\nu)})$ for the associated estimate of the source $S^{(\nu)}$ (cf. (58)), we obtained as results to these experiments the concordance matrices (cf. Definition 10)

$$\mathcal{C}(\hat{S}^{(1)}, S^{(1)}) \doteq \begin{pmatrix} \mathbf{0.853} & 0.065 \\ 0.079 & \mathbf{0.930} \end{pmatrix} \quad \text{and} \tag{66}$$

$$\mathcal{C}(\hat{S}^{(2)}, S^{(2)}) \doteq \begin{pmatrix} \mathbf{0.834} & 0.003 & 0.037 & 0.016 \\ 0.148 & \mathbf{0.725} & 0.109 & 0.069 \\ 0.037 & 0.034 & \mathbf{0.803} & 0.265 \\ 0.077 & 0.131 & 0.072 & \mathbf{0.787} \end{pmatrix}, \tag{67}$$

where we corrected for the permutation ambiguity between $\hat{S}$ and S to simplify comparison.

Both (66) and (67) 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,[45] as stated by Theorem 4. A visual comparison of the original samples $S^{(1)}$, $S^{(2)}$ and their estimates $\hat{S}^{(1)}$, $\hat{S}^{(2)}$, see Figures 1 and 6, confirms these results.

[44] Notice that: (a) by definition of Θ_1, the function Φ_{ou}^1 depends on the one-dimensional parameter θ_2 only; (b) as the concordance matrix of $\hat{S}_{-0.5} := g_{-0.5}^{(1)}(X_{\mathrm{ou}}^{(1)})$ and S_{ou} is $\left(\begin{smallmatrix} 0.053 & 0.929 \\ 0.834 & 0.099 \end{smallmatrix}\right)$ (indicating a close proximity between $\hat{S}_{-0.5}$ and $\mathrm{DP}_d \cdot S_{\mathrm{ou}}$, cf. Proposition 5), the observation of Φ_{ou}^1 attaining a low local minimum at -0.5 is in accordance with Theorem 4.

[45] Recall that the optimal deviation $\hat{S}^{(\nu)} \in \mathrm{DP}_d \cdot S^{(\nu)}$ between $\hat{S}^{(\nu)}$ and $S^{(\nu)}$ is achieved iff (66) and (67) are permutation matrices (Proposition 5).

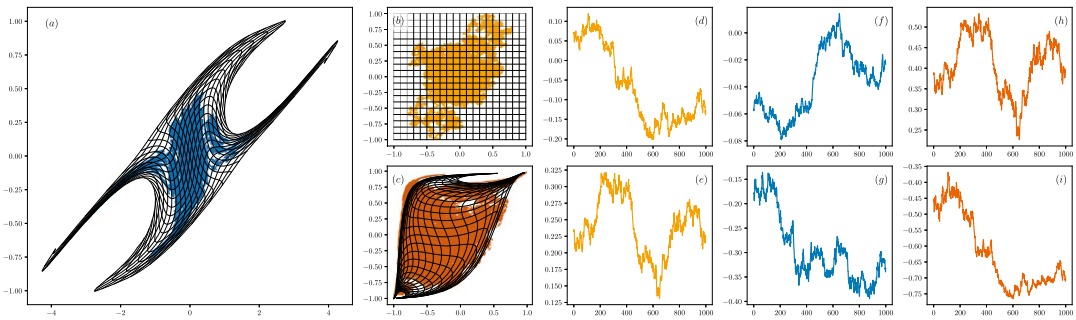

FIG. 6. *Nonlinear mixture X (sampled trace (a) and components (f), (g)) of an IC Ornstein–Uhlenbeck source S ((b) and (d), (e)). Further shown is the residual $g \circ f|_{[-1,1]^2}$ ((c); cf. (68)) for an estimate g of $f^{-1}|_{D_X}$. The function g is found by optimising an artificial neural network (g_θ) via the loss function (65), and the resulting estimate $\hat{S} := g(X)$ of S is shown in brown ((c) and (h), (i)). To a good approximation, the source S and its estimate coincide up to (a transposition and) a monotone scaling of their components, as quantified by (66).*

Specifically, Figure 1 shows the components, excerpted over 1000 data points each, of: the source $S^{(2)}$ (orange), its nonlinear mixture $X^{(2)}$ (blue), and its estimate $\hat{S}^{(2)}$ (brown).

To reaffirm that the above results of finding good approximations to the source are not simply due to chance, we ran our experiments repeatedly with randomly chosen realisations and initial configurations for the data and the learning process (64) and (65), see [52], [53], Section 3.

These experiments underline the practical applicability of our proposed ICA-method.

To conclude, we note the following empirical findings.

REMARK 8.1 (Empirical comments).

(i) Given an observable $X = f(S)$ together with a family Θ of candidate transformations on $\mathbb{R}^d$, the technical compatibility condition $(\mathrm{DP}_d(D_S) \cdot f^{-1})\big|_{D_X} \cap \Theta\big|_{D_X} \neq \varnothing$ of Theorem 4 can in practice typically not be guaranteed a priori. However, as indicated by the above findings (66) and (67), infringements of this (sufficient) technical condition might typically be innocuous, provided that at least

$$(68) \qquad (\mathrm{DP}_d \cdot g)\big|_{D_X} \cap \Theta\big|_{D_X} \neq \varnothing \quad \text{for some } g \text{ with } g\big|_{D_X} \text{ 'close enough' to } f^{-1}\Big|_{D_X},$$

which will be satisfied if Θ is chosen large enough, say as a suitable ANN or another universal approximator. In a similar vein, our experiments indicate that the regularity condition $\Theta \subseteq C^2(D_X)$ may in practice be softened by merely requiring that the 'approximate inverse' g in (68) be 'C^2-invertible on most of D_X' (cf. e.g., Figure 6, panel (c)) and the parametrization of Θ 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 (58) to the search for a (low) local minimum.

(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 (65) are typically nonconvex with their topography crucially depending on the choice of (64) (cf. e.g., Figure 5), we expect that the accuracy and efficiency of our estimates may be significantly improved by applying our ICA-method to ANN-based approximation schemes (64), (65) which are more carefully designed.

9. Conclusion. This paper has addressed the problem of Blind Source Separation via the classical approach of Independent Component Analysis. As our main contribution, we have formulated and proved a statistical method to recover multidimensional stochastic processes (in both continuous- as well as discrete-time) from observations of their nonlinear

mixtures. Conceptually, our method assumes a source process with independent component processes and, by exploiting the temporal structure of this source, characterises its nonlinear transformations by the degree of intercomponental statistical dependence that they inflict on the source. Quantifying the latter by way of an efficiently computable contrast function derived from the signature cumulants of a stochastic process, the initial source separation problem may then be reformulated as a provably robust problem of optimisation-based function approximation which in practice can be conveniently implemented by, for example, contemporary neural network-based learning schemes. A comprehensive consistency analysis [53], Section E, ensures that the resulting method is usable in real-world situations (discretized time, one sample trajectory), which is further illustrated by a number of theoretical and numerical examples.

The mathematics of the identifiability theory established in this work appears flexible enough to allow for extensions in various further directions. For instance, by considering third-order in place of second-order finite-dimensional distributions it may be adapted to infer the identifiability of stochastic processes from their time-dependent nonlinear mixing transformations ('invertible flows'). By adapting the ideas of this paper further, it does now also seem within reach to prove the identifiability of stochastic sources from more general nonlinear relations, such as for instance in the setting of controlled differential equations where one may be interested to recover an (independent-component) stochastic control from its nonlinear response. As with most methods involving an optimisation over flexibly parametrisable nonlinearities, however, a significant practical caveat of our approach is the occurence of spurious local minima in the approximation of the demixing transformation. This leaves room for improvement that future research might explore: In addition to practical deliberations such as spanning the optimisation domain by more carefully designed learning architectures, or amplifying the contrast function by the addition of tunable hyperparameters such as weights attached to its summands, one may attempt to tame the critical optimisation task by adjusting it to (localised) polynomial approximations of the mixing nonlinearity and harvesting the additional algebraic structure that then results from the fact ([13]) that the signature transform 'dualises' the action of polynomial transformations on its arguments.

Acknowledgments. The authors would like to extend their gratitude to the Associate Editor, four anonymous referees, the Editor, and Aapo Hyvärinen for their very helpful comments and suggestions which helped to significantly improve the original version of this paper and its presentation.

Funding. AS was financially supported by an Oxford-Cocker Graduate Scholarship and a Mathematical Institute Scholarship. HO is supported by the DataSig Program [EP/S026347/1] and the CIMDA collaboration between the City University of Hong Kong and the University of Oxford.

SUPPLEMENTARY MATERIAL

Supplement to "Nonlinear independent component analysis for discrete-time and continuous-time signals" (DOI: 10.1214/23-AOS2256SUPP; .pdf). Due to page number limitations, we have moved some of the more technical aspects of this paper to several appendices that we provide as supplementary material [53]. This includes:

- a guided use case for our method and a table of frequent notation (Appendices A, B);
- all omitted proofs along with auxiliary results and further remarks (Appendices C, D, G);
- a theoretical consistency analysis of our ICA-method (Appendices E, F) together with a readily implementable algorithm to perform this method in practice (Appendix E.4);
- a code documentation for the numerical experiments from Section 8 (Appendix H).

REFERENCES

[1] ALMEIDA, L. B. (2003). MISEP—linear and nonlinear ICA based on mutual information. *J. Mach. Learn. Res.* **4** 1297–1318.

[2] ARDIZZONE, L., KRUSE, J., WIRKERT, S., RAHNER, D., PELLEGRINI, E. W., KLESSEN, R. S., MAIER-HEIN, L., ROTHER, C. and KÖTHE, U. (2018). Analyzing Inverse Problems With Invertible Neural Networks. Published at ICLR 2019. Preprint. Available at arXiv:1808.04730.

[3] BACH, F. R. and JORDAN, M. I. (2003). Kernel independent component analysis. *J. Mach. Learn. Res.* **3** 1–48. MR1966051 https://doi.org/10.1162/153244303768966085

[4] BELL, A. J. and SEJNOWSKI, T. (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Comput.* **7** 1129–1159.

[5] BELOUCHRANI, A., MERAIM, K. A., CARDOSO, J. F. and MOULINES, E. (1997). A blind source separation technique using second-order statistics. *IEEE Trans. Signal Process.* **45** 434–444.

[6] BONNIER, P. and OBERHAUSER, H. (2020). Signature cumulants, ordered partitions, and independence of stochastic processes. *Bernoulli* **26** 2727–2757. MR4140527 https://doi.org/10.3150/20-BEJ1205

[7] BRAKEL, P. and BENGIO, Y. (2017). Learning independent features with adversarial nets for non-linear ICA. Preprint. Available at arXiv:1710.05050 [stat.ML].

[8] CARDOSO, J. F. (1999). High-order contrasts for independent component analysis. *Neural Comput.* **11** 157–192. https://doi.org/10.1162/089976699300016863

[9] CARDOSO, J. F. and SOULOUMIAC, A. (1993). Blind beamforming for non Gaussian signals. *IEE Proc. F* **140** 362–370.

[10] CHEN, X. and FAN, Y. (2006). Estimation of copula-based semiparametric time series models. *J. Econometrics* **130** 307–335. MR2211797 https://doi.org/10.1016/j.jeconom.2005.03.004

[11] CHEVYREV, I. and LYONS, T. (2016). Characteristic functions of measures on geometric rough paths. *Ann. Probab.* **44** 4049–4082. MR3572331 https://doi.org/10.1214/15-AOP1068

[12] CHEVYREV, I. and OBERHAUSER, H. (2018). Signature Moments to Characterize Laws of Stochastic Processes. Preprint. Available at arXiv:1810.1097.

[13] COLMENAREJO, L. and PREISS, R. (2020). Signatures of paths transformed by polynomial maps. *Beitr. Algebra Geom.* **61** 695–717. MR4160818 https://doi.org/10.1007/s13366-020-00493-9

[14] COMON, P. (1994). Independent component analysis, a new concept? *Signal Process.* **36** 287–314.

[15] COMON, P. and JUTTEN, C., eds. (2010). *Handbook of Blind Source Separation*: *Independent Component Analysis and Applications*. Academic Press, San Diego.

[16] CRANMER, K., BREHMER, J. and LOUPPE, G. (2020). The frontier of simulation-based inference. *Proc. Natl. Acad. Sci. USA* **117** 30055–30062. MR4263287 https://doi.org/10.1073/pnas.1912789117

[17] DARMOIS, G. (1953). Analyse générale des liaisons stochastiques. Etude particulière de l'analyse factorielle linéaire. *Rev. Inst. Int. Stat.* **21** 2–8. MR0061322

[18] DARSOW, W. F., NGUYEN, B. and OLSEN, E. T. (1992). Copulas and Markov processes. *Illinois J. Math.* **36** 600–642. MR1215798

[19] DING, H., WANG, Y., YANG, Z. and PFEIFFER, O. (2019). Nonlinear blind source separation and fault feature extraction method for mining machine diagnosis. *Appl. Sci.* **9** 1852.

[20] EMBRECHTS, P., MCNEIL, A. J. and STRAUMANN, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. In *Risk Management*: *Value at Risk and Beyond* (*Cambridge*, 1998) 176–223. Cambridge Univ. Press, Cambridge. MR1892190 https://doi.org/10.1017/CBO9780511615337.008

[21] EMURA, T., LONG, T.-H. and SUN, L.-H. (2017). R routines for performing estimation and statistical process control under copula-based time series models. *Comm. Statist. Simulation Comput.* **46** 3067–3087. MR3640123 https://doi.org/10.1080/03610918.2015.1073303

[22] ERIKSSON, J. and KOIVUNEN, V. (2004). Identifiability, separability and uniqueness of linear ICA models. *IEEE Signal Process. Lett.* **11** 601–604.

[23] FRIZ, P. K. and VICTOIR, N. B. (2010). *Multidimensional Stochastic Processes as Rough Paths*: *Theory and Applications*. *Cambridge Studies in Advanced Mathematics* **120**. Cambridge Univ. Press, Cambridge. MR2604669 https://doi.org/10.1017/CBO9780511845079

[24] GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. and SCHÖLKOPF, B. (2005). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6** 2075–2129. MR2249882

[25] HARMELING, S., ZIEHE, A., KAWANABE, M. and MÜLLER, K. R. (2003). Kernel-based nonlinear blind source separation. *Neural Comput.* **15** 1089–1124.

[26] HASTIE, T. and TIBSHIRANI, R. (2003). Independent component analysis through product density estimation. *Adv. Neural Inf. Process. Syst.* **15** 649–656.

[27] HE, Q. P. and WANG, J. (2018). Statistical process monitoring as a big data analytics tool for smart manufacturing. *J. Process. Control* **67** 35–43.

[28] HJELM, R. D. et al. (2018). Learning deep representations by mutual information estimation and maximization. Preprint. Available at arXiv:1808.06670v5 [stat.ML].
[29] HYVÄRINEN, A. (1997). Independent Component Analysis by Minimization of Mutual Information. Technical Report (Report A46), Helsiniki Univ. Technology.
[30] HYVÄRINEN, A. (1998). New approximations of differential entropy for independent component analysis and projection pursuit. *Adv. Neural Inf. Process. Syst.* 273–279.
[31] HYVÄRINEN, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10** 626–634.
[32] HYVÄRINEN, A. (2013). Independent component analysis: Recent advances. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **371** 20110534. MR3005670 https://doi.org/10.1098/rsta.2011.0534
[33] HYVÄRINEN, A., KARHUNEN, J. and OJA, E. (2001). *Independent Component Analysis*. Wiley, New York.
[34] HYVÄRINEN, A. and MORIOKA, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *NeurIPS*2016 3765–3773.
[35] HYVÄRINEN, A. and MORIOKA, H. (2017). Nonlinear ICA of temporally dependent stationary sources. *PMLR* **54** 460–469. Supplementary Material at http://proceedings.mlr.press/v54/hyvarinen17a/hyvarinen17a-supp.pdf.
[36] HYVÄRINEN, A. and PAJUNEN, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw.* **12** 429–439.
[37] HYVÄRINEN, A., SASAKI, H. and TURNER, R. (2019). Nonlinear ICA using auxiliary variables and generalized contrastive learning. *AISTATS*.
[38] İCAN, Ö. and ÇELIK, T. B. (2017). Stock market prediction performance of neural networks: A literature review. *Int. J. Econ. Finance* **9** 100–108.
[39] KHEMAKHEM, I., KINGMA, D. P., MONTI, R. P. and HYVÄRINEN, A. (2020). Variational autoencoders and nonlinear ICA: A unifying framework. *AISTATS*2020.
[40] KHOSHNEVIS, S. A. and SANKAR, R. (2019). Applications of higher order statistics in electroencephalography signal processing: A comprehensive survey. *IEEE Rev. Biomed. Eng.* **13** 169–183.
[41] KIDGER, P. and LYONS, T. Signatory: Differentiable computations of the signature and logsignature transforms, on both CPU and GPU. *ICLR* 2021.
[42] LYONS, T. J., CARUANA, M. and LÉVY, T. (2007). *Differential Equations Driven by Rough Paths. Lecture Notes in Math.* **1908**. Springer, Berlin. MR2314753
[43] MIETTINEN, J., NORDHAUSEN, K. and TASKINEN, S. (2017). Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp. *J. Stat. Softw.* **76**.
[44] MOULINES, E., CARDOSO, J. F. and GASSIAT, E. (1997). 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) 3617–3620.
[45] NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2197664 https://doi.org/10.1007/s11229-005-3715-x
[46] NOÉ, F., TKATCHENKO, A., MÜLLER, K.-R. and CLEMENTI, C. (2020). Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **71** 361–390. https://doi.org/10.1146/annurev-physchem-042018-052331
[47] PHAM, D. T. and GARRAT, P. (1997). Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. Signal Process.* **45** 1712–1725.
[48] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435
[49] REIERSØL, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica* **18** 375–389. MR0038054 https://doi.org/10.2307/1907835
[50] ROGERS, L. C. G. and WILLIAMS, D. (2000). *Diffusions, Markov Processes, and Martingales. Vol.* 2. *Cambridge Mathematical Library*. Cambridge Univ. Press, Cambridge. MR1780932 https://doi.org/10.1017/CBO9781107590120
[51] SAMWORTH, R. J. and YUAN, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.* **40** 2973–3002. MR3097966 https://doi.org/10.1214/12-AOS1060
[52] SCHELL, A. (2021). SigNICA. GitHub repository, https://github.com/alexander-schell/SigNICA.git Code for Section 8 (Jupyter Notebooks and Python files).
[53] SCHELL, A. and OBERHAUSER, H. (2023). Supplement to "Nonlinear Independent Component Analysis For Discrete-Time and Continuous-Time Signals." https://doi.org/10.1214/23-AOS2256SUPP
[54] SKITOVIČ, V. P. (1953). On a property of the normal distribution. *Dokl. Akad. Nauk SSSR* **89** 217–219. MR0055597
[55] TAN, Y., WANG, J. and ZURADA, J. M. (2001). Nonlinear blind source separation using a radial basis function network. *IEEE Trans. Neural Netw.* **12** 124–134.