Identifiable Markov Switching Models with Instantaneous Effects and Exponential Families

ICML 2026

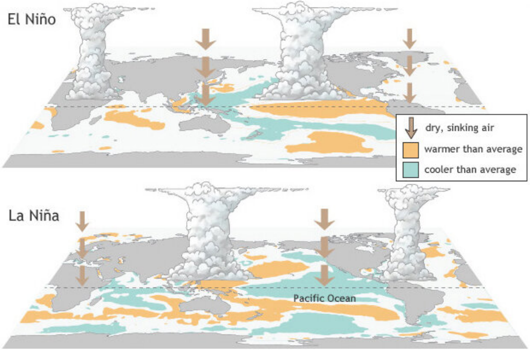
Roel Hulsman[1,2], Carles Balsells-Rodas[3], Sara Magliacane[1,4]

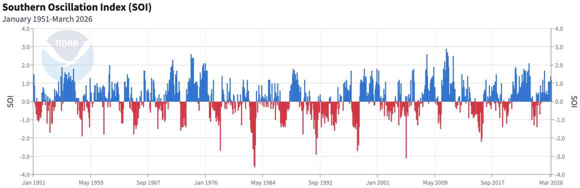
[1]University of Amsterdam [2]Adyen [3]Imperial College London [4]Saarland University

Adyen – June 18th, 2026

More info at `https://roelhulsman.nl/flowmsm/blog`

Regimes in El Niño-Southern Oscillation (ENSO)[1,2,3]





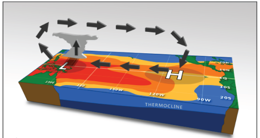


Figure 1
NEUTRAL CONDITIONS

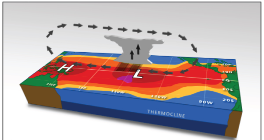


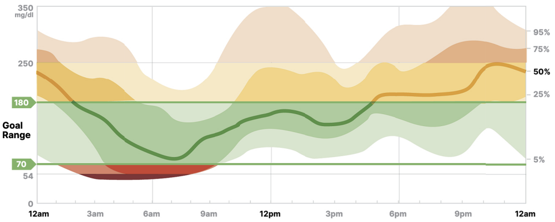
Figure 2
EL NIÑO CONDITIONS

[1] Michelle L'Heureux. *What is the El Niño–Southern Oscillation (ENSO) in a nutshell?* 2014. www.climate.gov/blogs/what-el-nino-nutshell.
[2] https://www.ncei.noaa.gov/access/monitoring/enso/soi.
[3] Emily Becker. *EN... SO?*. 2014. www.climate.gov/blogs/en-so.

Instantaneous effects in glucose levels of type-1 diabetes patients[4]

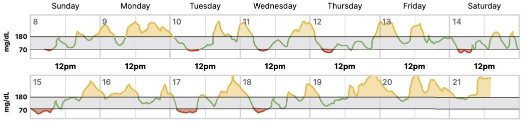
Ambulatory Glucose Profile (AGP)

AGP is a summary of glucose values from the report period, with median (50%) and other percentiles shown as if they occurred in a single day.



Daily Glucose Profiles

Each daily profile represents a midnight-to-midnight period.



[4] American Diabetes Association. *Glycemic targets: standards of medical care in diabetes*. Diabetes Care, 2025.

Our work in short

Main contribution: We establish *identifiability* for a broad class of regime-switching (nonlinear) structural causal models (SCMs) under independent exponential family noise.

Identifiability theory characterises when the data likelihood *uniquely determines* latent regimes.

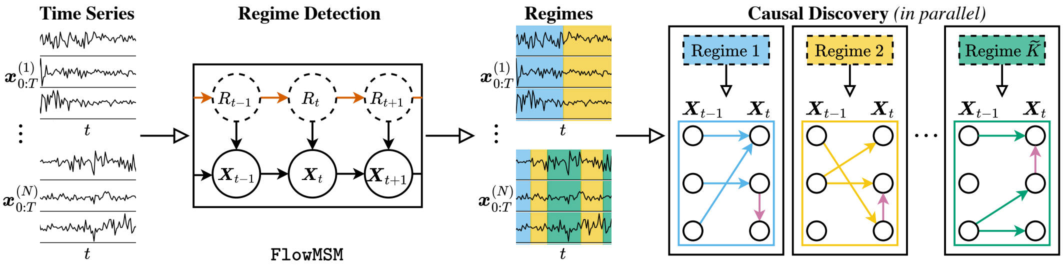


Figure 1: We detect *latent regimes* $R_{0:T}$ and *window graphs* from time series $X_{0:T}$, with dependencies between regimes (**orange** edges) and *instantaneous effects* between observed variables (**pink** edges).

Fama-French five-factor asset-pricing model – Regime detection

Five risk factors that capture systemic patterns in stock returns: size, value, market risk, profitability and investment[5]. This is supplemented by excess returns of Apple's stock (`AAPL`)[6].

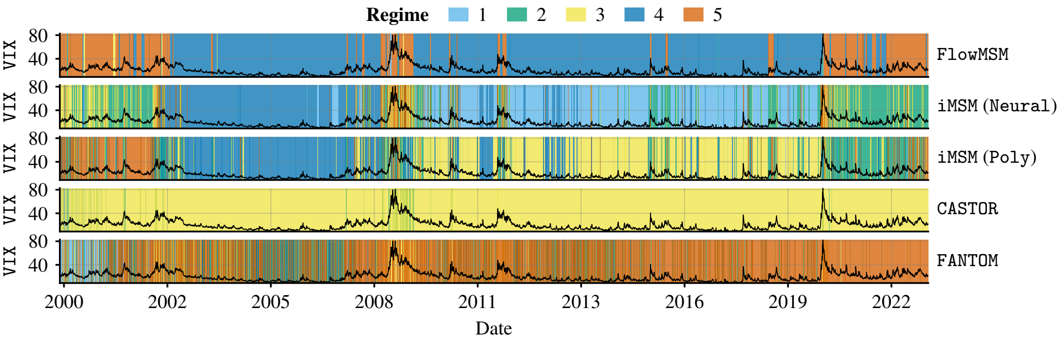


Figure 2: Estimated regimes on daily data, overlayed by the `VIX` volatility index (not used in training).

[5] Fama, French. *A five-factor asset pricing model.* Journal of Financial Economics, 2015

[6] Sadeghi, Gopal, Fesanghary. *Causal discovery from nonstationary time series.* International Journal of Data Science and Analytics, 2024.

SCMs and MSMs

What is a structural causal model[7] (SCM)?

A **structural causal model** encodes the causal relations between

- endogenous variables $\boldsymbol{X} = (X_1, \dots, X_D) \in \mathcal{X}$ *(in this work continuous: $\mathcal{X} \subseteq \mathbb{R}^D$)*;
- their causal parents $\mathrm{Pa}_i \subset \boldsymbol{X} \backslash X_i$ for $i = 1, \dots, D$; and
- exogenous noises $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_D) \in \mathcal{X}_\epsilon$ with $\boldsymbol{\epsilon} \sim p_{\boldsymbol{\epsilon}}$.

We assume the data generating process can be described by a set of **structural equations**:

$$X_1 \leftarrow f_1(\mathrm{Pa}_1, \epsilon_1), \\ \vdots \\ X_D \leftarrow f_D(\mathrm{Pa}_D, \epsilon_D). \tag{1}$$

Typically, we assume causal parents are **acyclic**, such that the structural equations are *recursive*.

[7] Pearl. *Causality: Models, Reasoning, and Inference*. 2009.

The causal structure can be visualised in a **causal graph**

Assuming **causal sufficiency**, we exclude latent confounders between observed variables.
That is, all relevant variables are observed.

It allows us to visualise the causal structure in a **directed acyclic graph** (DAG).

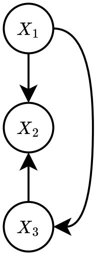


Figure 3: Example DAG with $D = 3$ variables.

(But we partly relax this later...)

A temporal SCM

Say we observe time series $\boldsymbol{X}_{0:T} \in \mathcal{X}^{\times(T+1)}$.

In vector-notation, for (instantaneous) causal parents $\mathbf{Pa}^0 \subset \boldsymbol{X}_0$, we obtain initial equations,

$$\boldsymbol{X}_0 \leftarrow \boldsymbol{f}^0\left(\mathbf{Pa}^0, \epsilon_0\right), \tag{2}$$

and transition equations for (instantaneous) parents $\mathbf{Pa} \subset \boldsymbol{X}_{t-1:t}$ and $t \in \{1, \dots, T\}$,

$$\boldsymbol{X}_t \leftarrow \boldsymbol{f}\left(\mathbf{Pa}, \epsilon_t\right). \tag{3}$$

For simplicity, causal parents are of maximum lag $L = 1$, but all our results generalise.

We typically assume **causal stationarity**, meaning functions $\boldsymbol{f}$ and causal parents $\mathbf{Pa}$ are invariant across time. *(But not in this work...)*

A temporal causal structure

Under **causal stationarity**, we observe a repeating causal structure. *(But not in this work...)*

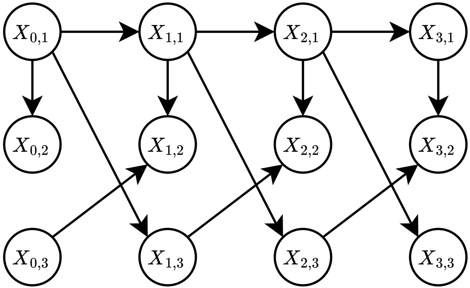


Figure 4: Example temporal DAG with $D = 3$ variables.

A regime-switching SCM

Consider time series $\boldsymbol{X}_{0:T}$ that is causally stationary only in **discrete segments of time**.

This is commonly modelled using discrete latent regimes $R_t \in \mathcal{A}_K = \{1, \ldots, K\}$, where $K < \infty$. The set $\mathcal{A}_K$ is a finite subset of $\mathcal{A}$, *e.g.* $\mathcal{A} = \mathbb{N}$, indexing all regimes in a model class.

The SCM becomes

$$\begin{aligned} \boldsymbol{X}_0 &\leftarrow \boldsymbol{f}^0\left(\mathbf{Pa}^0(R_0), R_0, \epsilon_0\right), \\ \boldsymbol{X}_t &\leftarrow \boldsymbol{f}\left(\mathbf{Pa}\left(R_t\right), R_t, \epsilon_t\right), \qquad t \in \{1, \ldots, T\}. \end{aligned} \tag{4}$$

The equations governing regimes $\boldsymbol{R}_{0:T}$ can be anything, although they cannot depend on $\boldsymbol{X}_{0:T}$.

We often write shorthand $\boldsymbol{f}_a \triangleq \boldsymbol{f}|_{R_t=a}$ when we fix the regime to some value $a \in \mathcal{A}$.

A regime-dependent temporal causal graph

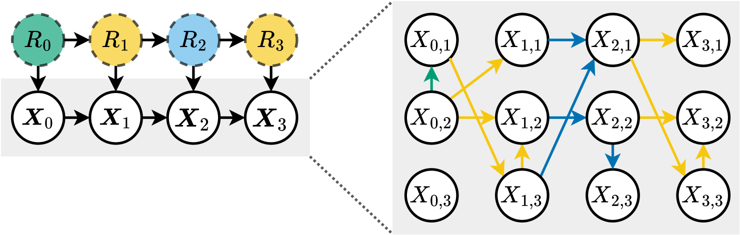


Figure 5: Example DAG given regimes $\boldsymbol{R}_{0:3}$, with coloured edges belonging to initial and window graphs[8].

R_t is a **latent confounder** for the variables $\boldsymbol{X}_t$, inducing regime-dependent causal structures.

The causal graphs need *not* be unique across regimes.

[8]Assaad, Devijver, Gaussier. *Survey and evaluation of causal discovery methods for time series.* Journal of Artificial Intelligence Research, 2022.

An overview of our assumptions

To summarise, we assume

- *Conditional* causal stationarity: Regimes fully govern structural changes over time;
- *Conditional* causal sufficiency: No latent confounders *within each regime*;
- Acyclicity: Structural equations are recursive;
- Causal Markov & faithfulness: Conditional independency in data $\iff$ d-separation in DAG.

For tractable estimation, we assume *(i)* regimes $\boldsymbol{R}_{0:T}$ follow a *first-order stationary Markov chain* and *(ii)* the mappings $\boldsymbol{f}_a^0, \boldsymbol{f}_a$ are *contractive* to guarantee stability.

Markov Switching Models (MSMs)

The dynamic Bayesian network induced by a regime-switching SCM is an MSM[9].

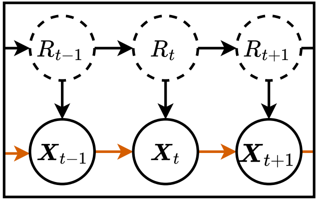


Figure 6: An MSM is an autoregressive Hidden Markov Model (HMM).

Classic identifiability results for finite-state HMMs[10,11,12] do not trivially extend to MSMs.

[9] Hamilton. *A new approach to the economic analysis of nonstationary time series and the business cycle.* Econometrica, 1989.
[10] Kruskal. *Three-way arrays: Rank and uniqueness of trilinear decompositions, ...* Linear Algebra and its Applications, 1977.
[11] Allman, Matias, Rhodes. *Identifiability of parameters in latent structure models with many observed variables.* The Annals of Statistics, 2009.
[12] Gassiat, Cleynen, Robin. *Inference in finite state space non parametric hidden Markov models and applications.* Statistics and Computing, 2016.

MSMs as finite mixture models[13]

The joint distribution of $\boldsymbol{X}_{0:T}$ can be written as a **finite mixture** over K^{T+1} regime sequences,

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T}) = \sum_{\boldsymbol{r}_{0:T} \in \mathcal{A}_K^{\times(T+1)}} \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{r}_{0:T})}_{\text{coefficients}} \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T} \mid \boldsymbol{r}_{0:T})}_{\text{components}}. \tag{5}$$

We factorise components into initial and transition distributions, drawn from families $\mathcal{P}_{\mathcal{A}}^{0}$, $\mathcal{P}_{\mathcal{A}}$,

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T} \mid \boldsymbol{r}_{0:T}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid r_0) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, r_t). \tag{6}$$

[13] Frühwirth-Schnatter. *Finite mixture and Markov switching models*. 2006.

Example SCMs (initial equations left out for brevity)

Let $\boldsymbol{W}_{a,0}, \boldsymbol{W}_{a,1}$ be weighted adjacency matrices, where $\boldsymbol{W}_{a,0}$ is acyclic and $\rho(\boldsymbol{W}_{a,0}), \rho(\boldsymbol{W}_{a,1}) < 1$.

Linear SVAR

$$\boldsymbol{X}_t \leftarrow \boldsymbol{W}_{a,0}\boldsymbol{X}_t + \boldsymbol{W}_{a,1}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t.$$

Nonlinear ANM[14]

$$\boldsymbol{X}_t \leftarrow \tanh(\boldsymbol{W}_{a,0}\boldsymbol{X}_t + \boldsymbol{W}_{a,1}\boldsymbol{X}_{t-1}) + \boldsymbol{\epsilon}_t.$$

LSNM[15]

$$\boldsymbol{X}_t \leftarrow \big(1 + \delta - \tanh(\boldsymbol{W}_{a,0}\boldsymbol{X}_t)\big) \circ \big(1 + \delta - \tanh(\boldsymbol{W}_{a,0}\boldsymbol{X}_t + \boldsymbol{W}_{a,1}\boldsymbol{X}_{t-1})\big) \circ \boldsymbol{\epsilon}_t,$$

where $\circ$ denotes the Hadamard product and $\delta > 0$ is a tiny constant.

[14] Hoyer, Janzing, Mooij, Peters, Schölkopf. *Nonlinear causal discovery with additive noise models.* NeurIPS, 2008.

[15] Immer, Schultheiss, Vogt, Schölkopf, Bühlmann, Marx. *On the identifiability and estimation of causal location-scale noise models.* ICML, 2023.

A visualisation of the generated time series

MSMs with (non-)Gaussian noises and (non)linear transitions may be hard to distinguish...

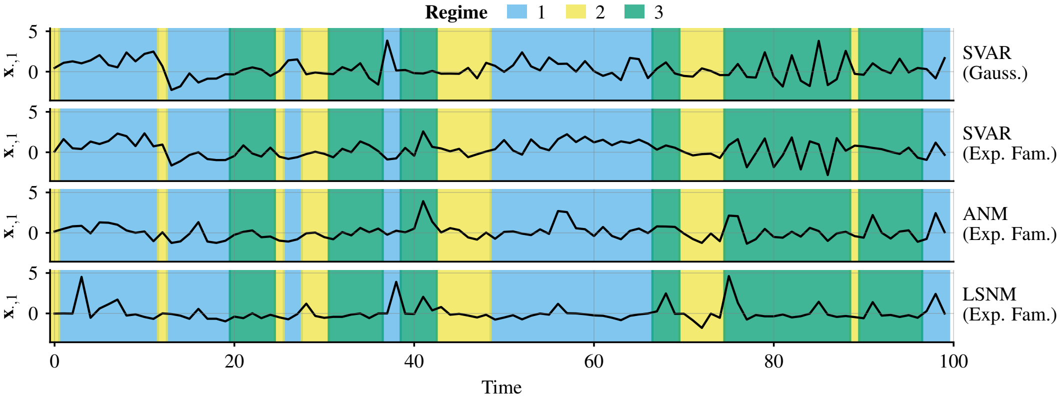


Figure 7: The only differences between the rows amounts to the structural equations and exogenous noises, using *(Gauss)* $\epsilon_{t,d} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and *(Exp. Fam.)* $\epsilon_{t,d} \overset{i.i.d.}{\sim} \text{Gamma}(0.25, 2) - 0.5$.

Gaussian and non-Gaussian mixtures

Only for affine instantaneous effects and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 \boldsymbol{I})$, we obtain a **classic Gaussian mixture model**:

$$p_\theta(\boldsymbol{x}_0 \mid a) = \mathcal{N}\big((\boldsymbol{I} - \boldsymbol{W}_{a,0})^{-1} \boldsymbol{\mu}_a, \Sigma_a\big), \tag{7}$$
$$p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, a) = \mathcal{N}\big((\boldsymbol{I} - \boldsymbol{W}_{a,0})^{-1} \boldsymbol{W}_{a,1} \boldsymbol{x}_{t-1}, \Sigma_a\big),$$

where $\Sigma_a \triangleq \boldsymbol{\sigma}^2 (\boldsymbol{I} - \boldsymbol{W}_{a,0})^{-1} (\boldsymbol{I} - \boldsymbol{W}_{a,0})^{-T}$. Acyclicity guarantees invertibility of $\boldsymbol{I} - \boldsymbol{W}_{a,0}$.

Identifiability is guaranteed when the Gaussian parameters are distinct[16].

Problem: Many real-world scenarios are much more complex, *e.g.*, if instantaneous effects are not affine, or the noise is non-Gaussian, yet we lose these guarantees...

Key question: When are regimes identifiable in such non-Gaussian mixtures?

[16] Yakowitz, Spragins. *On the identifiability of finite mixtures*. The Annals of Mathematical Statistics, 1968.

Identifiability theory

Identifiability (up to permutation)

MSMs are **identifiable (up to permutation)** if the data likelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T})$ *uniquely determines* the regime prior $p_{\boldsymbol{\theta}}(\boldsymbol{r}_{0:T})$ and mixture components $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T} \mid \boldsymbol{r}_{0:T})$, *up to a regime relabelling*.

For brevity, in this talk we leave out *"up to permutation"*.

Thm. 3.5 – Identifiable Regime-Switching SCMs

Consider an acyclic regime-switching SCM that satisfies conditional causal stationarity, conditional causal sufficiency, and **Ass. 3.1** to **3.4**. Then the induced MSM is **identifiable**.

In the remainder, we introduce our assumptions **3.1** to **3.4** and provide a *proof sketch*.

A necessary and sufficient global condition for identifiability

Yakowitz and Spragins (1968)[17]

A necessary and sufficient condition for identifiability is **linear independence** (under finite mixtures) of the functions in the family of mixture components, *i.e.*, for any finite $\mathcal{A}_K \subset \mathcal{A}$,

$$\sum_{a \in \mathcal{A}_K} \lambda_a p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid a) = 0 \quad a.e. \quad \Longrightarrow \quad \lambda_a = 0 \quad \forall a \in \mathcal{A}_K.$$

Thus, identifiability reduces to linear independence of PDFs in the **product family** $\mathcal{P}^0_{\mathcal{A}} \otimes {\mathcal{P}_{\mathcal{A}}}^{\otimes T}$.

[17] Yakowitz, Spragins. *On the identifiability of finite mixtures*. The Annals of Mathematical Statistics, 1968.

Is linear independence in $\mathcal{P}_{\mathcal{A}}^{0}$ and $\mathcal{P}_{\mathcal{A}}$ sufficient?

Short answer: It is **necessary**, but **not sufficient**.

An overlapping variable space challenges linear independence.

For $T = 1$, we have one overlapping variable $\boldsymbol{x}_0$, *i.e.*,

$$\mathcal{P}_{\mathcal{A}}^{0} \otimes \mathcal{P}_{\mathcal{A}} = \left\{ p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid r_0) p_{\boldsymbol{\theta}}(\boldsymbol{x}_1 \mid \boldsymbol{x}_0, r_1) \right\}. \quad (8)$$

Coupling distributions through $\boldsymbol{x}_0$ might accidentally create linear dependencies in the joint space.

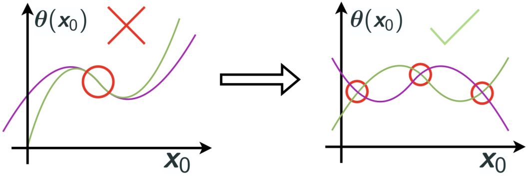


Figure 8: We allow linear dependence only on zero-measure subsets of the overlapping variable space[18]. A sufficiently regular space (*e.g.*, *real-analytic*) forces discontinuities to arise solely from regime changes.

[18]Balsells-Rodas, Wang, Li. *On the identifiability of switching dynamical systems.* ICML, 2024.

Sufficient local conditions for identifiability

So, if the PDFs in the initial and transition families are linearly independent, and the joint variable space is sufficiently regular, then we obtain **identifiability up to permutation**.

Our technical contributions answer two questions:

- Can we ensure linear independence in the initial and transition families beyond Gaussians?
- Linear independence is an abstract notion, can we translate this to more fine-grained conditions on regime-switching SCMs?

For simplicity, we focus on the transition family in this talk, omitting the initial family.

Identifiable mixtures of exponential family distributions

Recall, the PDF of a continuous variable $\boldsymbol{\epsilon} \in \mathcal{X}_\epsilon$ from a minimal regular exponential family is

$$p_{\boldsymbol{\eta}}(\boldsymbol{\epsilon}) = h(\boldsymbol{\epsilon}) \exp\left(\boldsymbol{\eta} \cdot \boldsymbol{\tau}(\boldsymbol{\epsilon}) - A(\boldsymbol{\eta})\right), \tag{9}$$

for P-dimensional natural parameters $\boldsymbol{\eta}$, $P \geq 1$, sufficient statistic $\boldsymbol{\tau}$, base measure h and log-partition A. The support $\mathcal{X}_\epsilon \subseteq \mathbb{R}^D$ does not depend on $\boldsymbol{\eta}$.

Barndorff-Nielsen (1965)[19]

(Finite) mixtures of a continuous minimal regular exponential family are **identifiable** if *(i)* the sufficient statistic is continuous with *(ii)* an image that contains an open set on its support.

Note mixtures are over the natural parameters $\boldsymbol{\eta}$, keeping the sufficient statistic $\boldsymbol{\tau}$ and support $\mathcal{X}_\epsilon$ fixed across regimes. Finite mixtures of multivariate Gaussian families are a special case.

[19] Barndorff-Nielsen. *Identifiability of mixtures of exponential families*. Journal of Mathematical Analysis and Applications, 1965.

Exponential family noise

We derive conditions on regime-switching SCMs such that we can use Barndorff-Nielsen (1965).

Ass. 3.1 – Exponential Family Noise

The exogenous noise $\boldsymbol{\epsilon}_t$ is from a continuous minimal regular exponential family that satisfies:

(a1) *Real-analytic sufficient statistic*: The sufficient statistic $\boldsymbol{\tau}$ is a real-analytic function *a.e.*.

(a2) *Rich image of the sufficient statistic*: The image $\{\boldsymbol{\tau}(\boldsymbol{\epsilon}_t) \mid h(\boldsymbol{\epsilon}_t) > 0, \boldsymbol{\epsilon}_t \in \mathcal{X}_\epsilon\}$ contains a (non-empty) open set.

We strengthen *continuity* of the sufficient statistic to *real-analyticity* to avoid linear dependence in the overlapping variable space.

The condition *(a2)* excludes certain degenerate and curved exponential families where the image of the sufficient statistic is confined to a lower-dimensional manifold, *e.g.*, when $P > D$.

Each regime represents a unique pushforward of the noise distribution

We aim to establish that each regime represents a unique pushforward of the noise distribution.

Ass. 3.2 – Functional Model Restrictions

The mappings $\boldsymbol{f}_a$ satisfy:

($b1$) *Unique regimes*: For all $a, a' \in \mathcal{A}$,

$$\boldsymbol{f}_a = \boldsymbol{f}_{a'} \quad a.e. \quad \Longrightarrow \quad a = a'; \tag{10}$$

($b2$) *Pointwise diffeomorphisms*: For almost every $\boldsymbol{x}_{t-1} \in \mathcal{X}$, and for all $a \in \mathcal{A}$, the mapping $\boldsymbol{f}_a|_{\boldsymbol{X}_{t-1}=\boldsymbol{x}_{t-1}}$ is a diffeomorphism $a.e.$ in $\boldsymbol{\epsilon}_t$;

($b3$) *Jointly real-analytic transitions*: For all $a \in \mathcal{A}$, the mapping $\boldsymbol{f}_a$ is jointly real-analytic in $(\boldsymbol{x}_{t-1}, \boldsymbol{\epsilon}_t)$ $a.e.$.

Condition *(b3)*, in combination with *(a1)*, ensures the overlapping variable space is real-analytic.

Some ambiguous unidentifiable cases

We rule out unidentifiable cases that exploit symmetries, *e.g.*, rotations of isotropic Gaussians.

Ass 3.3 – Trivial Automorphisms

At least one of the following holds:

(c1) *Trivial automorphisms of the noise*: The noise distribution has a trivial automorphism class, *i.e.*, for any invertible mapping Φ,

$$\Phi(\epsilon) \stackrel{d}{=} \epsilon \quad \Longrightarrow \quad \Phi = \epsilon \quad a.s.; \tag{11}$$

(c2) *Monotone canonicalisation*: There exists a canonical variable order *s.t.* $\forall a \in \mathcal{A}$, $\partial \mathbf{f}_a / \partial \epsilon_t$ can be permuted to a lower-triangular matrix, and it has strictly positive diagonal *a.e.* in ϵ_t.

The stronger condition $(c2)$ guarantees a *unique* monotone triangular transport known as the Knothe-Rosenblatt rearrangement[20,21].

[20] Knothe. *Contributions to the theory of convex bodies*. Michigan Mathematical Journal, 1957.

[21] Rosenblatt. *Remarks on a multivariate transformation*. The Annals of Mathematical Statistics, 1952.

The exponential family structure is preserved

The structural equations define mappings $\Phi_a|_{\mathbf{x}_{t-1}} = \mathbf{x}_{t-1} : \mathcal{X}_\epsilon \to \mathcal{X}$ from the noise space to the observed space at each time step, for a fixed regime $a \in \mathcal{A}$.

Under **Ass. 3.1-3.3**, using the change of variables formula and the exponential family structure,

$$p_{\boldsymbol{\theta}}(\mathbf{x}_t \mid \mathbf{x}_{t-1}, a) = h\big(\Phi_a^{-1}(\mathbf{x}_t, \mathbf{x}_{t-1})\big) \cdot \big|\det J_{\Phi_a^{-1}(\cdot, \mathbf{x}_{t-1})}(\mathbf{x}_t)\big| \cdot \exp\Big(\boldsymbol{\eta} \cdot \boldsymbol{\tau}\big(\Phi_a^{-1}(\mathbf{x}_t, \mathbf{x}_{t-1})\big) - A(\boldsymbol{\eta})\Big),$$

where $\Phi_a^{-1}(\cdot, \mathbf{x}_{t-1})$ denotes the inverse of $\Phi_a(\mathbf{x}_{t-1}, \cdot)$ with respect to $\boldsymbol{\epsilon}_t$ for fixed $\mathbf{x}_{t-1}$.

This is still an exponential family, but each transition distribution might have a different sufficient statistic and base measure...

How can we still leverage Barndorff-Nielsen (1965)?

A reparametrisation to an exponential family with a shared sufficient statistic

Ass. 3.4 – Sufficient variability across regimes in a finite polynomial Subspace

The mappings Φ_a satisfy for some finite $O < \infty$ and all $a \in \mathcal{A}$ and almost every $\boldsymbol{x}_{t-1} \in \mathcal{X}$:

($d1$) *Common support:* The base measure $h \circ \Phi_a^{-1}$ can be separated into a common base measure $\widetilde{h}$ and a scaling function $b_a(\boldsymbol{x}_{t-1}) > 0$;

($d2$) *Finite polynomial reparametrisation:* There exists a common sufficient statistic $\widetilde{\boldsymbol{\tau}} : \mathcal{X} \to \mathbb{R}^{\widetilde{P}}$, with $\widetilde{P} \geq P$, whose components are monomials up to order O, and matrices $\boldsymbol{C}_a(\boldsymbol{x}_{t-1}) \in \mathbb{R}^{P \times \widetilde{P}}$ with full row rank P, such that

$$\boldsymbol{\tau} \circ \Phi_a^{-1}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) = \boldsymbol{C}_a(\boldsymbol{x}_{t-1}) \widetilde{\boldsymbol{\tau}}(\boldsymbol{x}_t) + \mathcal{R}_O(\boldsymbol{x}_t) \quad a.e.,$$

where the remainder is regime-invariant and can be absorbed into the base measure;

($d3$) *Injectivity of polynomial coefficients:* For all $a \neq a' \in \mathcal{A}$,

$$\boldsymbol{C}_a(\boldsymbol{x}_{t-1})^T \boldsymbol{\eta} = \boldsymbol{C}_{a'}(\boldsymbol{x}_{t-1})^T \boldsymbol{\eta} \quad \forall \boldsymbol{\eta} \in \mathbb{R}^P \quad \Longrightarrow \quad \Phi_a = \Phi_{a'} \quad a.e..$$

A reparametrisation to an exponential family with a shared sufficient statistic

Ass. 3.4 guarantees an equivalent representation as a continuous minimal regular exponential family with distinct natural parameters and common sufficient statistic.

All regime- and history-dependent effects can be absorbed into parameters $\boldsymbol{\theta} : \mathcal{X} \times \mathcal{A} \to \Theta$, *i.e.*,

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, a) = \underbrace{\widetilde{h}(\boldsymbol{x}_t) \exp\left(\mathcal{R}_O(\boldsymbol{x}_t)\right)}_{\overline{h}(\boldsymbol{x}_t)} \exp\left(\underbrace{\left(\boldsymbol{C}_a^T(\boldsymbol{x}_{t-1})\boldsymbol{\eta}\right)}_{\boldsymbol{\theta}_a(\boldsymbol{x}_{t-1})} \cdot \widetilde{\boldsymbol{\tau}}(\boldsymbol{x}_t) - \underbrace{\log b_a(\boldsymbol{x}_{t-1}) - \widetilde{A}\left(\boldsymbol{C}_a^T(\boldsymbol{x}_{t-1})\boldsymbol{\eta}\right)}_{\overline{A}_a\left(\boldsymbol{\theta}_a(\boldsymbol{x}_{t-1}), \boldsymbol{x}_{t-1}\right)} \right),$$

Thus, the transition distribution remains a minimal regular exponential family.

The same argument holds for the initial family.

Putting it all together

We established finite mixtures of the initial and transition distribution families are identifiable[22], so the families contain linearly independent functions[23].

Throughout, we made sure the overlapping variable space is sufficiently regular, *i.e.*, real-analytic.

Thm. 3.5 – Identifiable Regime-Switching SCMs

Consider an acyclic regime-switching SCM that satisfies conditional causal stationarity, conditional causal sufficiency, and **Ass. 3.1** to **3.4**. Then the induced MSM is **identifiable**.

[22] Barndorff-Nielsen. *Identifiability of mixtures of exponential families*. Journal of Mathematical Analysis and Applications, 1965.
[23] Yakowitz, Spragins. *On the identifiability of finite mixtures*. The Annals of Mathematical Statistics, 1968.

Identifiable initial and window graphs

Upon identification of latent regimes, **known causal theory applies** to identify stationary causal graphs from the disentangled transition distributions.

For example, each window graph is identifiable up to a Markov equivalence class (MEC) of conditional independencies under *faithfulness*[24].

Alternatively, the transition distributions may correspond to a single window graph when the function class of $\boldsymbol{f}_a^0, \boldsymbol{f}_a$ is restricted[25].

[24] Spirtes, Glymour, Scheines. *Causation, Prediction, and Search*. 2000.

[25] Peters, Mooij, Janzing, Schölkopf. *Identifiability of causal graphs using functional models*. UAI, 2011.

Method

Regime detection and causal discovery

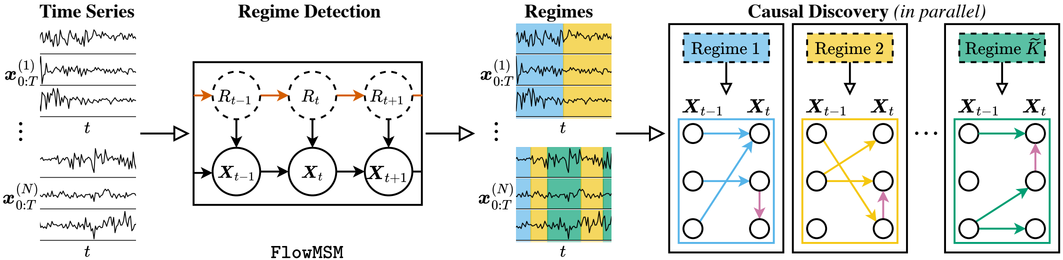


Figure 9: `FlowMSM` detects *latent regimes* $\boldsymbol{R}_{0:T}$, while a stationary causal method subsequently discovers *window graphs* $\boldsymbol{G}_{1:K}$ from time series $\boldsymbol{X}_{0:T}$.

Regime detection – `FlowMSM`

Input: Dataset $\{\boldsymbol{x}_{0:T}^{(n)}\}_{n=1}^{N}$ of N *i.i.d.* realisations of the time series $\boldsymbol{X}_{0:T}$.

Output: Estimates $\widehat{p}_{\boldsymbol{\theta}}(r_t \mid \boldsymbol{x}_{0:T}^{(n)})$ of the regime posterior likelihood.

We use **Generalised Expectation-Maximisation** (GEM), where a **conditional normalising flow** models the transition distributions, using shared parameters ψ and regime-embeddings $\boldsymbol{\mathcal{E}}_{1:\widetilde{K}}$.

The hyperparameter $\widetilde{K}$ models the number of regimes K. In practice we overshoot the oracle value, and redundant regimes are effectively ignored.

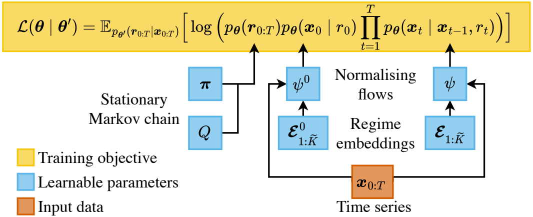


Figure 10: GEM updates $\boldsymbol{\theta}^{\text{new}} \leftarrow \boldsymbol{\theta}' + \alpha \, \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}$, guaranteed to converge to local optimum.

Sample splitting scheme

We assign the sequences $\boldsymbol{x}_{t-1:t}^{(n)}$, $t \in \{1, \ldots, T\}$, $n \in \{1, \ldots, N\}$, to the MAP regime

$$\widehat{r}_t^{(n)} = \arg\max_{r_t \in \mathcal{A}_{\widetilde{K}}} \widehat{p}_{\boldsymbol{\theta}}(r_t \mid \boldsymbol{x}_{0:T}^{(n)}). \tag{12}$$

This creates $\widetilde{K}$ clusters with partially overlapping sliding windows.

In theory, a perfect regime assignment would lead to clusters of causally stationary windows, since the causal parents of variables $\boldsymbol{x}_t^{(n)}$ depend solely on the oracle regime $r_t^{(n)}$.

In practice, we show `FlowMSM` exhibits high accuracy in most settings.

Causal discovery – Covered by established methods

To estimate window graphs from the clustered samples, `FlowMSM` can be paired with any stationary causal discovery method that allows for instantaneous effects.

- `VARLiNGAM`[26];
- `DYNOTEARS`[27];
- `PCMCI`$^{+}$[28];
- `Rhino`[29].

[26] Hyvärinen, Zhang, Shimizu, Hoyer. *Estimation of a structural vector autoregression model using non-gaussianity*. JMLR, 2010.
[27] Pamfil, Sriwattanaworachai, Desai, ..., Aragam. *DYNOTEARS: Structure learning from time-series data*. AISTATS, 2020.
[28] Runge. *Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets*. UAI, 2020.
[29] Gong, Jennings, Zhang, Pawlowski. *Rhino: Deep causal temporal relationship learning with history-dependent noise*. ICLR, 2023.

Experiments

Synthetic SCMs (initial equations left out for brevity)

Let $\boldsymbol{W}_{a,0:1}$ be weighted adjacency matrices, where $\boldsymbol{W}_{a,0}$ is acyclic and $\rho(\boldsymbol{W}_{a,0}) < 1, \rho(\boldsymbol{W}_{a,1}) < 1$.

Linear SVAR

$$\boldsymbol{X}_t \leftarrow \boldsymbol{W}_{a,0}\boldsymbol{X}_t + \boldsymbol{W}_{a,1}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t.$$

Nonlinear ANM[30]

$$\boldsymbol{X}_t \leftarrow \tanh(\boldsymbol{W}_{a,0}\boldsymbol{X}_t + \boldsymbol{W}_{a,1}\boldsymbol{X}_{t-1}) + \boldsymbol{\epsilon}_t.$$

LSNM[31]

$$\boldsymbol{X}_t \leftarrow \big(1 + \delta - \tanh(\boldsymbol{W}_{a,0}\boldsymbol{X}_t)\big) \circ \big(1 + \delta - \tanh(\boldsymbol{W}_{a,0}\boldsymbol{X}_t + \boldsymbol{W}_{a,1}\boldsymbol{X}_{t-1})\big) \circ \boldsymbol{\epsilon}_t,$$

where $\circ$ denotes the Hadamard product and $\delta > 0$ is a tiny constant.

[30] Hoyer, Janzing, Mooij, Peters, Schölkopf. *Nonlinear causal discovery with additive noise models.* NeurIPS, 2008.
[31] Immer, Schultheiss, Vogt, Schölkopf, Bühlmann, Marx. *On the identifiability and estimation of causal location-scale noise models.* ICML, 2023.

Synthetic experiments

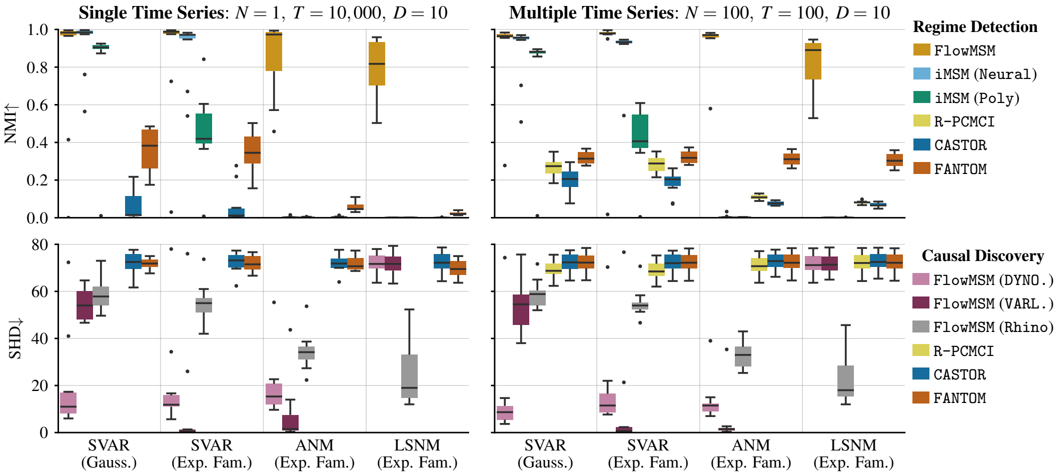


Figure 11: Performance on (*top*) regime detection and (*bottom*) causal discovery on (*left*) a single long time series and (*right*) multiple smaller time series, with $K = 3$ regimes. Each boxplot covers 10 seeds.

Fama-French five-factor asset-pricing model – Regime detection

Five risk factors that capture systemic patterns in stock returns: size, value, market risk, profitability and investment[32]. This is supplemented by excess returns of Apple's stock (`AAPL`)[33].

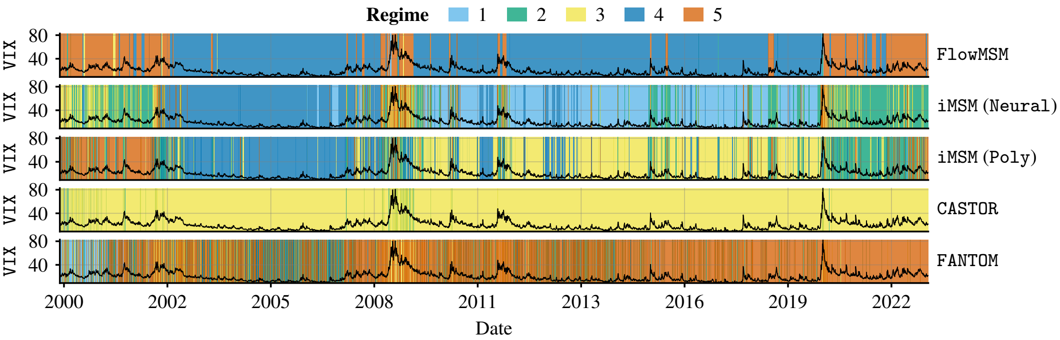


Figure 12: Estimated regimes on daily data, overlayed by the `VIX` volatility index (not used in training).

[32] Fama, French. *A five-factor asset pricing model.* Journal of Financial Economics, 2015

[33] Sadeghi, Gopal, Fesanghary. *Causal discovery from nonstationary time series.* International Journal of Data Science and Analytics, 2024.

Fama-French five-factor asset-pricing model – Causal discovery

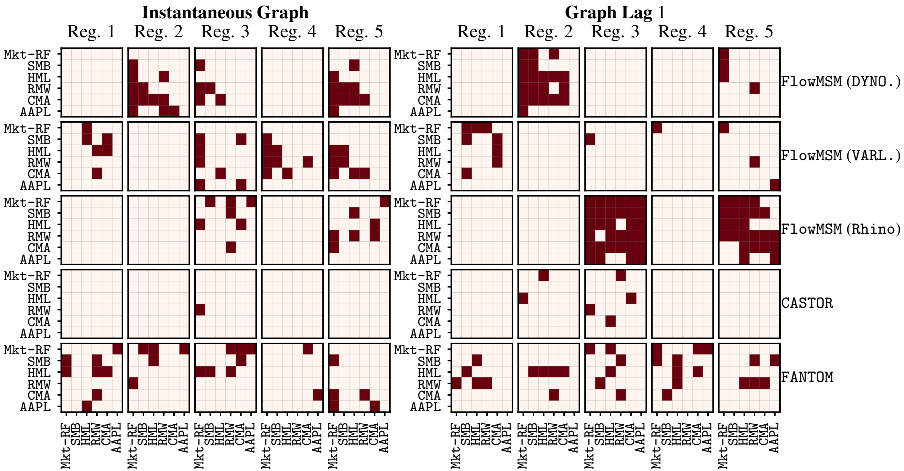


Figure 13: Adjacency matrices of the window graphs corresponding to the estimated regimes. The index (i, j, k, l) indicates the (lagged) edge $X_{t-l,i} \to X_{t,j}$ in regime k.

Discussion

Discussion

We explored regime detection and causal discovery in non-stationary time series.

We proposed novel theory on identifiability of latent regimes.

Our regime detection framework, `FlowMSM`, readily extends to regime-dependent causal discovery.

Future work:

- We require the regime variables $\boldsymbol{R}_{0:T}$ to be independent of the observations $\boldsymbol{X}_{0:T}$. Recent work partly relaxes this in Gaussian settings[34], but extending beyond is nontrivial;
- Our identifiability theory critically relies on exponential family noise and real-analytic diffeomorphisms. How to preserve linear independence in more relaxed settings?;
- Non-stationary time series might involve latent confounders other than regime variables.

[34] Balsells-Rodas, Sumba, Narendra, Tu, Schweikert, Kjellstrom, Li. *Causal discovery from conditionally stationary time series*. ICML, 2025.