Supplementary Materials for *Independent Innovation Analysis for Nonlinear Vector Autoregressive Process*

A Proof of Theorem 1

The log-pdf of $\tilde{\mathbf{x}}$ is given by, using the probability transformation formula,

$$\begin{aligned}\log p(\tilde{\mathbf{x}}(t)) &= \log p(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{u}_t) \\ &= \log p_{\tilde{\mathbf{s}}}(\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})|\mathbf{u}_t) + \log p(\mathbf{u}_t) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})| \\ &= \log p_{\tilde{\mathbf{s}}}(\mathbf{g}(\mathbf{x}_t, \mathbf{x}_{t-1}), \mathbf{x}_{t-1}|\mathbf{u}_t) + \log p(\mathbf{u}_t) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})| \\ &= \log p_{\mathbf{s}}(\mathbf{g}(\mathbf{x}_t, \mathbf{x}_{t-1})|\mathbf{u}_t) + \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t) + \log p(\mathbf{u}_t) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})| \end{aligned} \tag{11}$$

where $p_{\tilde{\mathbf{s}}}$, $p_{\mathbf{s}}$, and $p_{\mathbf{x}}$ are the conditional pdfs of $(\mathbf{s}, \mathbf{x})$, $\mathbf{s}$, and $\mathbf{x}$, respectively, $\mathbf{J}$ denotes the Jacobian, and $s_i = g_i(\mathbf{x}_t, \mathbf{x}_{t-1})$ by definition. The third equation is from the structure of the augmented demixing model (Eq. 3), and the last equation is from the temporal independence of $\mathbf{s}_t$ (assumption 2).

By well-known theory (Gutmann and Hyvärinen, 2012; Hastie et al., 2001), after convergence of logistic regression, with infinite data and a function approximator with universal approximation capability, the regression function (Eq. 6) will equal the difference of the log-pdfs in the two classes:

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{k}\psi_{ij}(h_i(\mathbf{x}_t, \mathbf{x}_{t-1}))\mu_{ij}(\mathbf{u}_t) + \phi(\mathbf{x}_{t-1}, \mathbf{u}_t) + \alpha(\mathbf{u}_t) + \beta(\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1})) + \gamma(\mathbf{x}_{t-1}) \\ &= \log p_{\mathbf{s}}(\mathbf{g}(\mathbf{x}_t, \mathbf{x}_{t-1})|\mathbf{u}_t) + \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t) + \log p(\mathbf{u}_t) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})| \\ &- \log p_{\tilde{\mathbf{s}}}(\mathbf{g}(\mathbf{x}_t, \mathbf{x}_{t-1})) - \log p_{\tilde{\mathbf{x}}}(\mathbf{x}_{t-1}) - \log p(\mathbf{u}_t) - \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})| \\ &= \sum_{i=1}^{n}\left[Q_i(g_i(\mathbf{x}_t, \mathbf{x}_{t-1})) - Z_i(\mathbf{u}_t) + \sum_{j=1}^{k} q_{ij}(g_i(\mathbf{x}_t, \mathbf{x}_{t-1}))\lambda_{ij}(\mathbf{u}_t)\right] + \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t) \\ &- \log p_{\tilde{\mathbf{s}}}(\mathbf{g}(\mathbf{x}_t, \mathbf{x}_{t-1})) - \log p_{\tilde{\mathbf{x}}}(\mathbf{x}_{t-1}) \end{aligned} \tag{12}$$

where $p_{\tilde{\mathbf{s}}}$ and $p_{\tilde{\mathbf{x}}}$ are the marginal pdfs of the innovations and observations when $\mathbf{u}$ is integrated out, and the last equation came from the conditional exponential pdf model of $\mathbf{s}$ (A1). The Jacobians and marginals $\log p(\mathbf{u})$ cancel out here. Considering its factorization form and the distinctive dependency of each term on $\mathbf{x}_t$, $\mathbf{x}_{t-1}$, and $\mathbf{u}_t$, the approximation solution is possible as

$$\begin{aligned}\psi_{ij}(h_i(\mathbf{x}_t, \mathbf{x}_{t-1})) &= q_{ij}(g_i(\mathbf{x}_t, \mathbf{x}_{t-1})) \\ \mu_{ij}(\mathbf{u}_t) &= \lambda_{ij}(\mathbf{u}_t) \\ \phi(\mathbf{x}_{t-1}, \mathbf{u}_t) &= \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t) \\ \alpha(\mathbf{u}_t) &= -\sum_{i=1}^{n} Z_i(\mathbf{u}_t) \\ \beta(\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1})) &= \sum_{i=1}^{n} Q_i(g_i(\mathbf{x}_t, \mathbf{x}_{t-1})) - \log p_{\tilde{\mathbf{s}}}(\mathbf{g}(\mathbf{x}_t, \mathbf{x}_{t-1})) \\ \gamma(\mathbf{x}_{t-1}) &= -\log p_{\tilde{\mathbf{x}}}(\mathbf{x}_{t-1}). \end{aligned} \tag{13}$$

Next, we have to prove that this is the only solution up to the indeterminacies given in the Theorem. Let $\mathbf{u}_0, \ldots, \mathbf{u}_{nk}$ be the points given by assumption 3 in the Theorem. We plug each of those $\mathbf{u}_l$ to obtain $nk + 1$ equations. By collecting those equations into rows, with subtracting the first equation for $\mathbf{u}_0$ from the remaining nk equations:

$$\mathbf{M}^T\boldsymbol{\psi}(\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1})) + \boldsymbol{\phi}(\mathbf{x}_{t-1}) + \boldsymbol{\alpha} = \mathbf{L}^T\mathbf{q}(\mathbf{s}_t) + \mathbf{p}(\mathbf{x}_{t-1}) + \mathbf{z}, \tag{14}$$

where $\mathbf{M} \in \mathbb{R}^{nk\times nk}$ is a matrix of $\mu_{ij}(\mathbf{u}_l) - \mu_{ij}(\mathbf{u}_0)$, with the product of i, j giving row index and l column index, $\mathbf{L}$ is a matrix of $\lambda_{ij}(\mathbf{u}_l) - \lambda_{ij}(\mathbf{u}_0)$ given in the assumption 3 in the Theorem, $\boldsymbol{\psi}(\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1})) = (\psi_{11}(h_1(\mathbf{x}_t, \mathbf{x}_{t-1})), \ldots, \psi_{nk}(h_n(\mathbf{x}_t, \mathbf{x}_{t-1})))^T$, $\mathbf{q}(\mathbf{s}_t) = (q_{11}(s_1(t)), \ldots, q_{nk}(s_n(t)))^T$, and the other vectors are

the collection of the corresponding terms in Eq. 12 at the nk points with all subtracting the one with $l = 0$; $\boldsymbol{\phi}(\mathbf{x}_{t-1}) = (\phi(\mathbf{x}_{t-1}, \mathbf{u}_1), \ldots, \phi(\mathbf{x}_{t-1}, \mathbf{u}_{nk}))^T - \mathbf{1}\phi(\mathbf{x}_{t-1}, \mathbf{u}_0)$, $\mathbf{1}$ is a $nk \times 1$ vector of ones, $\boldsymbol{\alpha} = (\alpha(\mathbf{u}_1), \ldots, \alpha(\mathbf{u}_{nk}))^T - \mathbf{1}\alpha(\mathbf{u}_0)$, $\mathbf{p}(\mathbf{x}_{t-1}) = (\log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_1), \ldots, \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_{nk}))^T - \mathbf{1} \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_0)$, and $\mathbf{z} = (-\sum_{i=1}^n Z_i(\mathbf{u}_1), \ldots, -\sum_{i=1}^n Z_i(\mathbf{u}_{nk}))^T + \mathbf{1}\sum_{i=1}^n Z_i(\mathbf{u}_0)$. In both sides of the equation, the terms not depending on $\mathbf{u}_t$ disappeared by the subtraction with $l = 0$. Let a compound demixing-mixing function $\mathbf{v}(\mathbf{s}_t, \mathbf{x}_{t-1}) = \mathbf{h} \circ \tilde{\mathbf{f}}(\mathbf{s}_t, \mathbf{x}_{t-1})$, and change variables to $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2] = [\mathbf{s}_t, \mathbf{x}_{t-1}]$, we then have

$$\mathbf{M}^T \boldsymbol{\psi}(\mathbf{v}(\mathbf{y})) + \boldsymbol{\phi}(\mathbf{y}_2) + \boldsymbol{\alpha} = \mathbf{L}^T \mathbf{q}(\mathbf{y}_1) + \mathbf{p}(\mathbf{y}_2) + \mathbf{z}. \tag{15}$$

Firstly, we will show that $\mathbf{M}$ is invertible. From the definition of $\mathbf{q}(\mathbf{y}_1)$, its partial derivative with respect to y_{1i} is $\mathbf{q}'(y_{1i}) = (0, \ldots, 0, q'_{i1}(y_{1i}), \ldots, q'_{ik}(y_{1i}), 0, \ldots, 0)^T$. According to Lemma 3 of Khemakhem et al. (2020), for y_{1i} which satisfies A1, there exist k points $(\bar{y}^1_{1i}, \ldots, \bar{y}^k_{1i})$ such that $(\mathbf{q}'(\bar{y}^1_{1i}), \ldots, \mathbf{q}'(\bar{y}^k_{1i}))$ are linearly independent. By differentiating Eq. 15 with respect to y_{1i} and collecting their evaluations at such k distinctive points for all i, we get

$$\mathbf{M}^T \tilde{\mathbf{Q}} = \mathbf{L}^T \mathbf{Q}, \tag{16}$$

where $\mathbf{Q} \in \mathbb{R}^{nk \times nk}$ is a matrix collecting $\mathbf{q}'(\bar{y}^l_{1i})$ to the columns indexed by (i, l), and $\tilde{\mathbf{Q}}$ is a collection of partial derivatives of $\boldsymbol{\psi}(\mathbf{v}(\mathbf{y}))$ evaluated at the same points. $\mathbf{Q}$ is invertible (through a combination of Lemma 3 of Khemakhem et al. (2020) and the fact that each component of $\mathbf{q}$ is univariate), and thus the right-hand side is invertible because $\mathbf{L}$ is invertible as well (assumption 3). The invertibility of the right-hand side implies the invertibility of $\mathbf{M}$ and $\tilde{\mathbf{Q}}$.

Now, let an augmented compound demixing-mixing function $\tilde{\mathbf{v}}(\mathbf{y}) = [\tilde{\mathbf{v}}_1(\mathbf{y}), \tilde{\mathbf{v}}_2(\mathbf{y})] = \tilde{\mathbf{h}} \circ \tilde{\mathbf{f}}(\mathbf{y})$, where $\tilde{\mathbf{h}}$ is the augmented function defined in the assumption 5 in the Theorem. The $\tilde{\mathbf{v}}_1(\mathbf{y})$ corresponds to $\mathbf{v}(\mathbf{y})$ defined above. Note that $\tilde{\mathbf{v}}$ is invertible because both $\tilde{\mathbf{h}}$ and $\tilde{\mathbf{f}}$ are invertible. What we need to prove is that $\tilde{\mathbf{v}}$ is a block-wise invertible point-wise function, in the sense that $\tilde{v}_{1i}$ is a function of only one y_{1j_i} and not of any of y_{2j_i}, and vise versa. This can be done by showing that the product of any two distinct partial derivatives of any component is always zero, and the Jacobian $\mathbf{J}_{\tilde{\mathbf{v}}} \in \mathbb{R}^{2n \times 2n}$ is block diagonal; the upper and lower block correspond to $\mathbf{y}_1$ and $\mathbf{y}_2$ respectively. Along with invertibility, this means that each component depends exactly on one variable of the corresponding block ($\mathbf{y}_1$ or $\mathbf{y}_2$). Below, we show that separately for $\mathbf{J}_{\mathbf{v}} \in \mathbb{R}^{n \times 2n}$ and $\mathbf{J}_{\tilde{\mathbf{v}}_2} \in \mathbb{R}^{n \times 2n}$. Firstly, this is obviously true for $\mathbf{J}_{\tilde{\mathbf{v}}_2}$ because $\tilde{\mathbf{v}}_2(\mathbf{y})$ is just an identity mapping of $\mathbf{y}_2$ from the definitions of $\tilde{\mathbf{h}}$ and $\tilde{\mathbf{f}}$, and does not depend on $\mathbf{y}_1$; the lower non-zero block of $\mathbf{J}_{\tilde{\mathbf{v}}}$ is an identity matrix. Next, we will show that for $\mathbf{J}_{\mathbf{v}}$. We differentiate Eq. 15 with respect to $y_c, 1 \leq c \leq n$ (an element of $\mathbf{y}_1 = \mathbf{s}_t$), and $y_d, c < d \leq 2n$, and get

$$\mathbf{M}^T \frac{\partial^2}{\partial y_c \partial y_d} \boldsymbol{\psi}(\mathbf{v}(\mathbf{y})) = 0. \tag{17}$$

From the invertibility of $\mathbf{M}$ and the calculation of differentials, we get

$$\frac{\partial^2}{\partial y_c \partial y_d} \boldsymbol{\psi}(\mathbf{v}(\mathbf{y})) = \boldsymbol{\Psi}(\mathbf{y})^T \boldsymbol{\upsilon}(\mathbf{y}) = 0, \tag{18}$$

where $\boldsymbol{\Psi}(\mathbf{y}) = (\mathbf{e}^{(1,1)}(y_1), \ldots, \mathbf{e}^{(1,k)}(y_1), \ldots, \mathbf{e}^{(n,1)}(y_n), \ldots, \mathbf{e}^{(n,k)}(y_n)) \in \mathbb{R}^{2n \times nk}$, $\mathbf{e}^{(a,b)} = (0, \ldots, 0, \psi'_{ab}(v_a), \psi''_{ab}(v_a), 0, \ldots, 0)^T \in \mathbb{R}^{2n}$, such that the non-zero entries are at indices $(2a-1, 2a)$, $\boldsymbol{\upsilon}(\mathbf{y}) = (v_1^{c,d}(\mathbf{y}), v_1^c(\mathbf{y}) v_1^d(\mathbf{y}), \ldots, v_n^{c,d}(\mathbf{y}), v_n^c(\mathbf{y}) v_n^d(\mathbf{y}))^T \in \mathbb{R}^{2n}$, $v_i^c = \frac{\partial v_i}{\partial y_c}(\mathbf{y})$, and $v_i^{c,d} = \frac{\partial^2 v_i}{\partial y_c \partial y_d}(\mathbf{y})$. From Lemma 4 and 5 of Khemakhem et al. (2020), assumption 6 implies that $\boldsymbol{\Psi}(\mathbf{y})$ has full row rank $2n$, and thus the pseudo-inverse of $\boldsymbol{\Psi}(\mathbf{y})^T$ fulfils $\boldsymbol{\Psi}(\mathbf{y})^{+T} \boldsymbol{\Psi}(\mathbf{y})^T = \mathbf{I}$. We multiply the equation above from the left by such pseudo-inverse and obtain

$$\boldsymbol{\upsilon}(\mathbf{y}) = 0. \tag{19}$$

In particular, $v_a^c(\mathbf{y}) v_a^d(\mathbf{y}) = 0$ for all $1 \leq a \leq n$, $1 \leq c \leq n$, and $c < d \leq 2n$. This means that a row of $\mathbf{J}_{\mathbf{v}} \in \mathbb{R}^{n \times 2n}$ at each $\mathbf{y}$ has either 1) only one non-zero entry somewhere in the former half block (corresponding to the partial derivatives by $\mathbf{y}_1$) or 2) non-zero entries only in the latter half block (corresponding to the partial derivatives by $\mathbf{y}_2$). The latter case is contradictory because it means that the component v_i is a function of only $\mathbf{y}_2 = \mathbf{x}_{t-1}$, and cannot hold Eq 15, which right-hand side is a function of all components of $\mathbf{y}_1$ (and $\mathbf{y}_2$). Therefore, $\mathbf{J}_{\mathbf{v}}$ should

have only one non-zero entry in the former half block for each row. From the results of $\mathbf{J}_{\mathbf{v}}$ and $\mathbf{J}_{\tilde{\mathbf{v}}_2}$, we deduce that $\mathbf{J}_{\tilde{\mathbf{v}}}$ is a block diagonal matrix. Now, by invertibility and continuity of $\mathbf{J}_{\tilde{\mathbf{v}}}$, we deduce that the location of the non-zero entries are fixed and do not change as a function of $\mathbf{y}$. This proves that $\tilde{\mathbf{v}} = \tilde{\mathbf{h}} \circ \tilde{\mathbf{f}}(\mathbf{y})$ is a block-wise invertible point-wise function, and v_i $(= h_i(\mathbf{x}_t, \mathbf{x}_{t-1}))$ is represented by only one y_{1j_i} $(= s_{j_i}(t))$ up to a scalar (component-specific) invertible transformation, and the Theorem is proven.

B Proof of Theorem 2

The conditional joint log-pdf of a data point $(\mathbf{x}_t, \mathbf{x}_{t-1})$ is given by, using the probability transformation formula,

$$\begin{aligned}
\log p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{u}_t) &= \log p_{\tilde{\mathbf{s}}}(\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})|\mathbf{u}_t) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})| \\
&= \log p_{\mathbf{s}}(\mathbf{g}(\mathbf{x}_t, \mathbf{x}_{t-1})|\mathbf{u}_t) + \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})| \\
&= \sum_{i=1}^{n} \left[Q_i(g_i(\mathbf{x}_t, \mathbf{x}_{t-1})) - Z_i(\mathbf{u}_t) + \sum_{j=1}^{k} q_{ij}(g_i(\mathbf{x}_t, \mathbf{x}_{t-1}))\lambda_{ij}(\mathbf{u}_t) \right] \\
&+ \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})| \qquad (20)
\end{aligned}$$

where $p_{\tilde{\mathbf{s}}}$, $p_{\mathbf{s}}$, and $p_{\mathbf{x}}$ are the conditional pdfs of $(\mathbf{s}, \mathbf{x})$, $\mathbf{s}$, and $\mathbf{x}$, respectively, $\mathbf{J}$ denotes the Jacobian, and $s_i = g_i(\mathbf{x}_t, \mathbf{x}_{t-1})$ by definition. The second equation is from the structure of the augmented demixing model (Eq. 3) and the temporal independence of $\mathbf{s}$ (assumption 2), and the last equation is from the conditional exponential family model of the innovation (A1). On the other hand, by applying Bayes rule on the optimal discrimination relation given by Eq. 8, after dividing all the exponential term by the one of $\tau = 1$ to avoid the well-known indeterminacy of the softmax function,

$$\begin{aligned}
\log p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{u}_t = \tau) &= \sum_{i=1}^{n}\sum_{j=1}^{k}(w_{ij\tau} - w_{ij1})\psi_{ij}(h_i(\mathbf{x}_t, \mathbf{x}_{t-1})) + \phi(\mathbf{x}_{t-1}, \mathbf{u}_t = \tau) \\
&- \phi(\mathbf{x}_{t-1}, \mathbf{u}_t = 1) + \log p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{u}_t = 1) + \alpha_\tau, \qquad (21)
\end{aligned}$$

where $\alpha_\tau = b_\tau - b_1 - \log p(\mathbf{u}_t = \tau) + \log p(\mathbf{u}_t = 1)$. Substituting Eq. 20 with $\mathbf{u}_t = 1$ into Eq. 21, we have;

$$\begin{aligned}
\log p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{u}_t = \tau) &= \sum_{i=1}^{n}\sum_{j=1}^{k}\left[(w_{ij\tau} - w_{ij1})\psi_{ij}(h_i(\mathbf{x}_t, \mathbf{x}_{t-1})) + q_{ij}(g_i(\mathbf{x}_t, \mathbf{x}_{t-1}))\lambda_{ij}(\mathbf{u}_t = 1)\right] \\
&+ \sum_{i=1}^{n}\left[Q_i(g_i(\mathbf{x}_t, \mathbf{x}_{t-1})) - Z_i(\mathbf{u}_t = 1)\right] + \phi(\mathbf{x}_{t-1}, \mathbf{u}_t = \tau) - \phi(\mathbf{x}_{t-1}, \mathbf{u}_t = 1) \\
&+ \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t = 1) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_t, \mathbf{x}_{t-1})| + \alpha_\tau \qquad (22)
\end{aligned}$$

Setting Eq. 22 and Eq. 20 with $\mathbf{u}_t = \tau$ to be equal for arbitrary τ, we have:

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{k}(w_{ij\tau} - w_{ij1})\,\psi_{ij}(h_i(\mathbf{x}_t, \mathbf{x}_{t-1})) + \phi(\mathbf{x}_{t-1}, \mathbf{u}_t = \tau) - \phi(\mathbf{x}_{t-1}, \mathbf{u}_t = 1) + \alpha_\tau \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}(\lambda_{ij}(\mathbf{u}_t = \tau) - \lambda_{ij}(\mathbf{u}_t = 1))\,q_{ij}(g_i(\mathbf{x}_t, \mathbf{x}_{t-1})) + \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t = \tau) - \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t = 1) + z_\tau \qquad (23)
\end{aligned}$$

where $z_\tau = \sum_{i=1}^{n} Z_i(\mathbf{u}_t = 1) - Z_i(\mathbf{u}_t = \tau)$. By collecting this equation for all the T labels into rows, except $\tau = 1$, which makes both-sides zero;

$$\mathbf{W}^T\boldsymbol{\psi}(\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1})) + \boldsymbol{\phi}(\mathbf{x}_{t-1}) + \boldsymbol{\alpha} = \mathbf{L}^T\mathbf{q}(\mathbf{s}_t) + \mathbf{p}(\mathbf{x}_{t-1}) + \mathbf{z}, \qquad (24)$$

where $\mathbf{W} \in \mathbb{R}^{nk\times(T-1)}$ is a matrix of $w_{ij\tau} - w_{ij1}$, with the product of i, j giving row index and τ column index, $\mathbf{L}$ is a matrix of $\lambda_{ij}(\mathbf{u}_t = \tau) - \lambda_{ij}(\mathbf{u}_t = 1)$ given in the assumption 4 in the Theorem, $\boldsymbol{\psi}(\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1})) = (\psi_{11}(h_1(\mathbf{x}_t, \mathbf{x}_{t-1})), \ldots, \psi_{nk}(h_n(\mathbf{x}_t, \mathbf{x}_{t-1})))^T$, $\mathbf{q}(\mathbf{s}_t) = (q_{11}(s_1(t)), \ldots, q_{nk}(s_n(t)))^T$, $\boldsymbol{\phi}(\mathbf{x}_{t-1}) = (\phi(\mathbf{x}_{t-1}, \mathbf{u}_t = 2), \ldots, \phi(\mathbf{x}_{t-1}, \mathbf{u}_t = T))^T - \mathbf{1}\phi(\mathbf{x}_{t-1}, \mathbf{u}_t = 1)$, $\mathbf{1}$ is a $(T-1)\times 1$ vector of ones, $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_T)^T$,

$\mathbf{p}(\mathbf{x}_{t-1}) = (\log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t = 2), \ldots, \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t = T))^T - \mathbf{1} \log p_{\mathbf{x}}(\mathbf{x}_{t-1}|\mathbf{u}_t = 1)$, and $\mathbf{z} = (z_2, \ldots, z_T)^T$. Let a compound demixing-mixing function $\mathbf{v}(\mathbf{s}_t, \mathbf{x}_{t-1}) = \mathbf{h} \circ \tilde{\mathbf{f}}(\mathbf{s}_t, \mathbf{x}_{t-1})$, and change variables to $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2] = [\mathbf{s}_t, \mathbf{x}_{t-1}]$, we then have

$$\mathbf{W}^T \boldsymbol{\psi}(\mathbf{v}(\mathbf{y})) + \boldsymbol{\phi}(\mathbf{y}_2) + \boldsymbol{\alpha} = \mathbf{L}^T \mathbf{q}(\mathbf{y}_1) + \mathbf{p}(\mathbf{y}_2) + \mathbf{z}. \tag{25}$$

Firstly, we will show that $\mathbf{W}$ has full row rank nk. From the definition of $\mathbf{q}(\mathbf{y}_1)$, its partial derivative with respect to y_{1i} is $\mathbf{q}'(y_{1i}) = (0, \ldots, 0, q'_{i1}(y_{1i}), \ldots, q'_{ik}(y_{1i}), 0, \ldots, 0)^T$. According to Lemma 3 of Khemakhem et al. (2020), for y_{1i} which satisfies A1, there exist k points $(\bar{y}^1_{1i}, \ldots, \bar{y}^k_{1i})$ such that $(\mathbf{q}'(\bar{y}^1_{1i}), \ldots, \mathbf{q}'(\bar{y}^k_{1i}))$ are linearly independent. By differentiating Eq. 25 with respect to y_{1i} and collecting their evaluations at such k distinctive points for all i, we get

$$\mathbf{W}^T \tilde{\mathbf{Q}} = \mathbf{L}^T \mathbf{Q}, \tag{26}$$

where $\mathbf{Q} \in \mathbb{R}^{nk \times nk}$ is a matrix collecting $\mathbf{q}'(\bar{y}^l_{1i})$ to the columns indexed by (i, l), and $\tilde{\mathbf{Q}}$ is a collection of partial derivatives of $\boldsymbol{\psi}(\mathbf{v}(\mathbf{y}))$ evaluated at the same points. $\mathbf{Q}$ is invertible (through a combination of Lemma 3 of Khemakhem et al. (2020) and the fact that each component of $\mathbf{q}$ is univariate), and thus the right-hand side has full column rank nk because $\mathbf{L}$ has full row rank nk (assumption 4). The full column rank of the right-hand side implies the full row rank of $\mathbf{W}$ and the invertibility of $\tilde{\mathbf{Q}}$.

Now, let an augmented compound demixing-mixing function $\tilde{\mathbf{v}}(\mathbf{y}) = [\tilde{\mathbf{v}}_1(\mathbf{y}), \tilde{\mathbf{v}}_2(\mathbf{y})] = \tilde{\mathbf{h}} \circ \tilde{\mathbf{f}}(\mathbf{y})$, where $\tilde{\mathbf{h}}$ is the augmented function defined in the assumption 6 in the Theorem. The $\tilde{\mathbf{v}}_1(\mathbf{y})$ corresponds to $\mathbf{v}(\mathbf{y})$ defined above. Note that $\tilde{\mathbf{v}}$ is invertible because both $\tilde{\mathbf{h}}$ and $\tilde{\mathbf{f}}$ are invertible. What we need to prove is that $\tilde{\mathbf{v}}$ is a block-wise invertible point-wise function, in the sense that $\tilde{v}_{1i}$ is a function of only one y_{1j_i} and not of any of y_{2j_i}, and vise versa. This can be done by showing that the product of any two distinct partial derivatives of any component is always zero, and the Jacobian $\mathbf{J}_{\tilde{\mathbf{v}}} \in \mathbb{R}^{2n \times 2n}$ is block diagonal; the upper and lower block correspond to $\mathbf{y}_1$ and $\mathbf{y}_2$ respectively. Along with invertibility, this means that each component depends exactly on one variable of the corresponding block ($\mathbf{y}_1$ or $\mathbf{y}_2$). Below, we show that separately for $\mathbf{J}_{\mathbf{v}} \in \mathbb{R}^{n \times 2n}$ and $\mathbf{J}_{\tilde{\mathbf{v}}_2} \in \mathbb{R}^{n \times 2n}$. Firstly, this is obviously true for $\mathbf{J}_{\tilde{\mathbf{v}}_2}$ because $\tilde{\mathbf{v}}_2(\mathbf{y})$ is just an identity mapping of $\mathbf{y}_2$ from the definitions of $\tilde{\mathbf{h}}$ and $\tilde{\mathbf{f}}$, and does not depend on $\mathbf{y}_1$; the lower non-zero block of $\mathbf{J}_{\tilde{\mathbf{v}}}$ is an identity matrix. Next, we will show that for $\mathbf{J}_{\mathbf{v}}$. We differentiate Eq. 25 with respect to $y_c, 1 \leq c \leq n$ (an element of $\mathbf{y}_1 = \mathbf{s}_t$), and $y_d, c < d \leq 2n$, and get

$$\mathbf{W}^T \frac{\partial^2}{\partial y_c \partial y_d} \boldsymbol{\psi}(\mathbf{v}(\mathbf{y})) = 0. \tag{27}$$

From the full row rank of $\mathbf{W}$ and the calculation of differentials, we get

$$\frac{\partial^2}{\partial y_c \partial y_d} \boldsymbol{\psi}(\mathbf{v}(\mathbf{y})) = \boldsymbol{\Psi}(\mathbf{y})^T \boldsymbol{\upsilon}(\mathbf{y}) = 0, \tag{28}$$

where $\boldsymbol{\Psi}(\mathbf{y}) = (\mathbf{e}^{(1,1)}(y_1), \ldots, \mathbf{e}^{(1,k)}(y_1), \ldots, \mathbf{e}^{(n,1)}(y_n), \ldots, \mathbf{e}^{(n,k)}(y_n)) \in \mathbb{R}^{2n \times nk}$, $\mathbf{e}^{(a,b)} = (0, \ldots, 0, \psi'_{ab}(v_a), \psi''_{ab}(v_a), 0, \ldots, 0)^T \in \mathbb{R}^{2n}$, such that the non-zero entries are at indices $(2a-1, 2a)$, $\boldsymbol{\upsilon}(\mathbf{y}) = (v_1^{c,d}(\mathbf{y}), v_1^c(\mathbf{y}) v_1^d(\mathbf{y}), \ldots, v_n^{c,d}(\mathbf{y}), v_n^c(\mathbf{y}) v_n^d(\mathbf{y}))^T \in \mathbb{R}^{2n}$, $v_i^c = \frac{\partial v_i}{\partial y_c}(\mathbf{y})$, and $v_i^{c,d} = \frac{\partial^2 v_i}{\partial y_c \partial y_d}(\mathbf{y})$. From Lemma 4 and 5 of Khemakhem et al. (2020), assumption 7 implies that $\boldsymbol{\Psi}(\mathbf{y})$ has full row rank $2n$, and thus the pseudo-inverse of $\boldsymbol{\Psi}(\mathbf{y})^T$ fulfils $\boldsymbol{\Psi}(\mathbf{y})^{+T} \boldsymbol{\Psi}(\mathbf{y})^T = \mathbf{I}$. We multiply the equation above from the left by such pseudo-inverse and obtain

$$\boldsymbol{\upsilon}(\mathbf{y}) = 0. \tag{29}$$

In particular, $v_a^c(\mathbf{y}) v_a^d(\mathbf{y}) = 0$ for all $1 \leq a \leq n$, $1 \leq c \leq n$, and $c < d \leq 2n$. This means that a row of $\mathbf{J}_{\mathbf{v}} \in \mathbb{R}^{n \times 2n}$ at each $\mathbf{y}$ has either 1) only one non-zero entry somewhere in the former half block (corresponding to the partial derivatives by $\mathbf{y}_1$) or 2) non-zero entries only in the latter half block (corresponding to the partial derivatives by $\mathbf{y}_2$). The latter case is contradictory because it means that the component v_i is a function of only $\mathbf{y}_2 = \mathbf{x}_{t-1}$, and cannot hold Eq 25, which right-hand side is a function of all components of $\mathbf{y}_1$ (and $\mathbf{y}_2$). Therefore, $\mathbf{J}_{\mathbf{v}}$ should have only one non-zero entry in the former half block for each row. From the results of $\mathbf{J}_{\mathbf{v}}$ and $\mathbf{J}_{\tilde{\mathbf{v}}_2}$, we deduce that $\mathbf{J}_{\tilde{\mathbf{v}}}$ is a block diagonal matrix. Now, by invertibility and continuity of $\mathbf{J}_{\tilde{\mathbf{v}}}$, we deduce that the location of the non-zero entries are fixed and do not change as a function of $\mathbf{y}$. This proves that $\tilde{\mathbf{v}} = \tilde{\mathbf{h}} \circ \tilde{\mathbf{f}}(\mathbf{y})$ is a block-wise invertible point-wise function, and v_i $(= h_i(\mathbf{x}_t, \mathbf{x}_{t-1}))$ is represented by only one y_{1j_i} $(= s_{j_i}(t))$ up to a scalar (component-specific) invertible transformation, and the Theorem is proven.

C Discussion on the identifiability of IIA-HMM

We obtain the following Theorem on identifiability of IIA-HMM.

Theorem 3. *Assume the following:*

1. *We obtain observations from an NVAR model (Eq. 1), whose augmented model (Eq. 2) is invertible and sufficiently smooth.*
2. *The latent innovations of the process follow the assumption A1 with $k \geq 2$, and the sufficient statistics q_{ij} are twice differentiable.*
3. *The $\mathbf{u}$ are unobserved (in contrast to the previous frameworks), and follow A2, where the transition matrix $\mathbf{A}$ has full rank with non-zero diagonal entries, and induces irreducible stationary Markov chain with a unique stationary state distribution $\boldsymbol{\pi}$.*
4. *The conditional distributions $p(\cdot|\mathbf{x}_{t-1}, \mathbf{u}_t), \mathbf{u}_t = 1, \ldots, C$ are all generically distinct for any $\mathbf{x}_{t-1}$, meaning that the set of points for which this doesn't hold is measure zero.*
5. *The modulation matrix of size $nk \times (C-1)$*

$$\mathbf{L} = (\boldsymbol{\lambda}(2) - \boldsymbol{\lambda}(1), \ldots, \boldsymbol{\lambda}(T) - \boldsymbol{\lambda}(1)) \tag{30}$$

   has full row rank nk, where $\boldsymbol{\lambda}(c) = (\lambda_{11}(\mathbf{u} = c), \ldots, \lambda_{nk}(\mathbf{u} = c))^T \in \mathbb{R}^{nk}$.

6. *We estimate the transition matrix, parameters of the innovation model, latent state at each data point, and demixing model $\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1}) : \mathbb{R}^{2n} \to \mathbb{R}^n$ with universal approximation capability, by maximizing the likelihood of the observations.*
7. *The augmented function $\tilde{\mathbf{h}}(\mathbf{x}_t, \mathbf{x}_{t-1}) = [\mathbf{h}(\mathbf{x}_t, \mathbf{x}_{t-1}), \mathbf{x}_{t-1}] : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is invertible.*

Then, in the limit of infinite data, $\mathbf{h}$ provides a consistent estimator of the IIA model: The functions $h_i(\mathbf{x}_t, \mathbf{x}_{t-1})$ give the independent innovations, up to permutation and scalar (component-wise) invertible transformations.

Proof. Assume equality of joint-data distributions for $2T + 1$ observations from the IIA-HMM model with two different sets of parameters $\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}$:

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_T, \ldots, \mathbf{x}_{2T+1}; \boldsymbol{\theta}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_T, \ldots, \mathbf{x}_{2T+1}; \hat{\boldsymbol{\theta}})$$

With the Assumptions 1, 3, and 4, we can apply Lemma 1 below and identify the following:

$$\mathbf{A} = \hat{\mathbf{A}} \tag{31}$$

$$\boldsymbol{\pi} = \hat{\boldsymbol{\pi}} \tag{32}$$

$$p(\mathbf{x}_t|\mathbf{u}_t = c; \boldsymbol{\theta}) = p(\mathbf{x}_t|\mathbf{u}_t = \sigma(c); \hat{\boldsymbol{\theta}}) \tag{33}$$

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_T|\mathbf{x}_{T+1}, \mathbf{u}_{T+1} = c; \boldsymbol{\theta}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_T|\mathbf{x}_{T+1}, \mathbf{u}_{T+1} = \sigma(c); \hat{\boldsymbol{\theta}}). \tag{34}$$

where $\sigma(\cdot)$ accounts for permutation of labels. For rest of the proof, without loss of generality, we assume label ordering matches. Since joint distributions identify their marginals uniquely, equation (34) implies

$$p(\mathbf{x}_T|\mathbf{x}_{T+1}, \mathbf{u}_{T+1}; \boldsymbol{\theta}) = p(\mathbf{x}_T|\mathbf{x}_{T+1}, \mathbf{u}_{T+1}; \hat{\boldsymbol{\theta}}) \tag{35}$$

$$\implies \frac{p(\mathbf{x}_T, \mathbf{x}_{T+1}|\mathbf{u}_{T+1}; \boldsymbol{\theta})}{p(\mathbf{x}_{T+1}|\mathbf{u}_{T+1}; \boldsymbol{\theta})} = \frac{p(\mathbf{x}_T, \mathbf{x}_{T+1}|\mathbf{u}_{T+1}; \hat{\boldsymbol{\theta}})}{p(\mathbf{x}_{T+1}|\mathbf{u}_{T+1}; \hat{\boldsymbol{\theta}})}. \tag{36}$$

This, with (33), implies that the following is identified:

$$p(\mathbf{x}_T, \mathbf{x}_{T+1}|\mathbf{u}_{T+1}; \boldsymbol{\theta}) = p(\mathbf{x}_T, \mathbf{x}_{T+1}|\mathbf{u}_{T+1}; \hat{\boldsymbol{\theta}}) \tag{37}$$

Finally, notice that

$$\sum_{k=1}^{C} p(\mathbf{x}_T|\mathbf{u}_T = k;\boldsymbol{\theta})p(\mathbf{u}_T = k|\mathbf{u}_{T+1};\mathbf{A}) = \sum_{k=1}^{C} p(\mathbf{x}_T|\mathbf{u}_T = k;\hat{\boldsymbol{\theta}})p(\mathbf{u}_T = k|\mathbf{u}_{T+1};\hat{\mathbf{A}})$$
$$\implies p(\mathbf{x}_T|\mathbf{u}_{T+1};\boldsymbol{\theta}) = p(\mathbf{x}_T,|\mathbf{u}_{T+1};\hat{\boldsymbol{\theta}}) \tag{38}$$

Writing out the log-likelihoods in equation (37), we get:

$$\begin{aligned}
&\log p_{\tilde{\mathbf{s}}}(\tilde{\mathbf{g}}(\mathbf{x}_T,\mathbf{x}_{T+1})|\mathbf{u}_{T+1}) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_T,\mathbf{x}_{T+1})| = \log \hat{p}_{\tilde{\mathbf{s}}}(\hat{\tilde{\mathbf{g}}}(\mathbf{x}_T,\mathbf{x}_{T+1})|\mathbf{u}_{T+1}) + \log|\det \mathbf{J}\hat{\tilde{\mathbf{g}}}(\mathbf{x}_T,\mathbf{x}_{T+1})| \\
\implies &\log p_{\mathbf{s}}(\mathbf{g}(\mathbf{x}_T,\mathbf{x}_{T+1})|\mathbf{u}_{T+1}) + \log p_{\mathbf{x}}(\mathbf{x}_T|\mathbf{u}_{T+1}) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_T,\mathbf{x}_{T+1})| \\
&= \log \hat{p}_{\mathbf{s}}(\hat{\mathbf{g}}(\mathbf{x}_T,\mathbf{x}_{T+1})|\mathbf{u}_{T+1}) + \log \hat{p}_{\mathbf{x}}(\mathbf{x}_T|\mathbf{u}_{T+1}) + \log|\det \mathbf{J}\hat{\tilde{\mathbf{g}}}(\mathbf{x}_T,\mathbf{x}_{T+1})|,
\end{aligned}$$

where $p_{\tilde{\mathbf{s}}}$, $p_{\mathbf{s}}$, and $p_{\mathbf{x}}$ are the conditional pdfs of $(\mathbf{s},\mathbf{x})$, $\mathbf{s}$, and $\mathbf{x}$, respectively, and $\mathbf{J}$ denotes the Jacobian. Using the result in (38), gives us

$$\log p_{\mathbf{s}}(\mathbf{g}(\mathbf{x}_T,\mathbf{x}_{T+1})|\mathbf{u}_{T+1}) + \log|\det \mathbf{J}\tilde{\mathbf{g}}(\mathbf{x}_T,\mathbf{x}_{T+1})| = \log \hat{p}_{\mathbf{s}}(\hat{\mathbf{g}}(\mathbf{x}_T,\mathbf{x}_{T+1})|\mathbf{u}_{T+1}) + \log|\det \mathbf{J}\hat{\tilde{\mathbf{g}}}(\mathbf{x}_T,\mathbf{x}_{T+1})|.$$

Remainder of the proof follows as in Hälvä and Hyvärinen (2020) and is not shown here for brevity. The general idea is to take the above equation for different values of $\mathbf{u}_{T+1}$ and use one of them as a 'pivot' in order to get rid of the Jacobians. Finally, the exponential family distribution properties, as done in the earlier proofs of this paper, are used to show identifiability. □

C.1 Lemmas

Set-up: These Lemmas follow, in general, those of Alexandrovich et al. (2016) but with substantial modifications made to accomodate our model. We first define some relevant notation. Let $(X_t)_{t\in\mathbb{N}}$ denote the observed process and $(U_t)_{t\in\mathbb{N}}$ the discrete latent first-order Markov chain. Assume these processes are time-homogeneous. K is the cardinality of the state-space of U_t, that is, the number of latent states. The first-order Markov chain for U_t is governed by transition matrix $\mathbf{A} = (\alpha_{j,k})_{j,k=1,\dots,K}$. Define $\mathcal{S} \subset \mathbb{R}^q$ as any subset of Euclidean space. Suppose that X_t takes values in $\mathcal{S}$, and its distribution depends on its most recent past X_{t-1} and the current latent state U_t – this distribution function is denoted by $F_{U_t,X_{t-1}}(X_t)$ and is time-homogeneous. Notice that subsequently, X_t is independent of all X_{t+s} for $|s| \geq 2$ given X_{t-1}, X_{t+1} and U_t. $F_{U_t}(X_t)$ is used to denote the conditonal distribution of X_t on U_t alone; that is, all other variables have been integrated out. $\pi = (\pi_1,\dots,\pi_K)$ denotes a stationary distribution of $\mathbf{A}$. In the following proofs, $\mathbf{x}_t$ is not a random variable, but represents a point in $\mathcal{S}$.

Let $\dim(V)$ denote the dimension of vector space V. For $\mathbf{v}_1,\dots,\mathbf{v}_n \in V$ let $\text{span}\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ denote the subspace of V spanned by $\mathbf{v}_1,\dots,\mathbf{v}_n$. For scalars $x_1,\dots,x_n \in \mathbb{R}$ let $\text{diag}(x_1,\dots,x_n)$ denote n-dimensional diagonal matrix with $x_1,\dots,x_n$ along the diagonal. $\mathbf{1}_K$ is a K-dimensional vector of ones. Let $\mathbf{M}_i \in \mathbb{R}^{K\times n_i}$ $(n_i \in \mathbb{N}; i = 1,2,3)$, then $[\mathbf{M}_i,\mathbf{M}_j]$ denotes the $K \times (n_i + n_j)$ matrix made by joining the two matrices at their columns. $(\mathbf{M})_{m,n}$ denotes the element of matrix $\mathbf{M}$ on the m-th row and n-th column. Finally, let's define three-way arrays, indexed by (i_1,i_2,i_3), where the corresponding element is given by:

$$\langle \mathbf{M}_1,\mathbf{M}_2,\mathbf{M}_3\rangle_{(i_1,i_2,i_3)} = \sum_{k=1}^{K} (\mathbf{M}_1)_{k,i_1}(\mathbf{M}_2)_{k,i_2}(\mathbf{M}_3)_{k,i_3} \;\; (i_j = 1,\dots,n_j) \tag{39}$$

Define kruskal rank $R_\kappa(\mathbf{M})$ as the maximal j such that any selection of j rows of $\mathbf{M}$ are linearly independent. Theorem 4a of Kruskal (1977) states that if:

$$R_\kappa(\mathbf{M}_1) + R_\kappa(\mathbf{M}_2) + R_\kappa(\mathbf{M}_3) \geq 2K + 2 \tag{40}$$

and

$$\langle \mathbf{M}_1,\mathbf{M}_2,\mathbf{M}_3\rangle = \langle \mathbf{N}_1,\mathbf{N}_2,\mathbf{N}_3\rangle$$

then there exist permutation matrix $\mathbf{P}$ and diagonal matrices $\boldsymbol{\Lambda}_i$, such that $\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_3 = I_K$ and $\mathbf{N}_i = \boldsymbol{\Lambda}_i\mathbf{P}\mathbf{M}_i$.

Lemma 1. *Assume that:*

1. *The latent state transition matrix* $\mathbf{A}$ *has full rank and is ergodic.*

2. *The conditional emission distributions* $F_{u_t,\mathbf{x}_{t-1}}(\mathbf{x}_t)$ *for* $k = 1, \ldots, K$ *are generically distinct for each given* $\mathbf{x}_{t-1}$*. That is, the set of points* $(\mathbf{x}_{t-1}, \mathbf{x}_t)$ *for which this doesn't hold is measure zero.*

3. *First-order Markov chain* (U_t) *is stationary with starting distribution* π*, which is thus the stationary distribution of* $\mathbf{A}$

then the marginal emission distributions $F_{u_t}(\mathbf{x}_t)$*, the transition matrix* $\mathbf{A}$*, the initial state probabilities* π *are all identified from the joint distribution of the observation process* $(X_1, \ldots, X_{2T+1})$ *where* $T \geq K - 1$*, up to label swapping.*

Proof. ***Step 1*** *(factorizing likelihood into blocks by conditional independence)*: Consider we have $2T + 1$ observations from our model. The likelihood of the model can be factored by conditioning on the variables at the central time point $T + 1$ as per below:

$$\begin{aligned}\Pr(X_{1:2T+1} \leq \mathbf{x}_{1:2T+1}) = \sum_k \Pr(X_{1:T} \leq \mathbf{x}_{1:T} | U_{T+1} = k, X_{T+1} \leq \mathbf{x}_{T+1}) \pi_k \\ \times \Pr(X_{T+1} \leq \mathbf{x}_{T+1} | U_{T+1} = k) \Pr(X_{T+2:2T+1} \leq \mathbf{x}_{T+2:2T+1} | U_{T+1} = k, X_{T+1} \leq \mathbf{x}_{T+1}),\end{aligned} \tag{41}$$

where notation $\mathbf{x}_{1:2T+1} = (\mathbf{x}_1, \ldots, \mathbf{x}_{2T+1})$ is used, and π_k represents the stationary distribution $\Pr(U_{T+1} = k)$. Assume $T \geq K - 1$. The factorial structure of the likelihood allows us to consider two random variables

$$V_T = X_{1:T} = (X_1, \ldots, X_T) \quad \text{and} \quad W_T = X_{T+2:2T+1} = (X_{T+2}, \ldots, X_{2T+1}).$$

The conditional distribution of W_T, evaluated at some $\mathbf{y}_{1:T} \in \mathcal{S}^T$, given $X_{T+1} = \mathbf{y}_0 \in \mathcal{S}$ and $U_{T+1} = k$ can be written as:

$$G_T(\mathbf{y}_{0:T}; k) = \Pr(W_T \leq \mathbf{y}_{1:T} | U_{T+1} = k, X_{T+1} \leq \mathbf{y}_0) = \sum_{k_1 \cdots k_T} \alpha_{k,k_1} \prod_{t=2}^{T} \alpha_{k_{t-1},k_t} \prod_{t=1}^{T} F_{k_t,\mathbf{y}_{t-1}}(\mathbf{y}_t).$$

For the conditional likelihood of V_T on U_{T+1} and X_{T+1}, we must consider time reversal:

$$\begin{aligned}\tilde{\mathbf{A}} &= (\tilde{\alpha}_{j,k})_{j,k=1,\ldots,K}, \quad \tilde{\alpha}_{j,k} = \frac{\pi_k \alpha_{k,j}}{\pi_j}, \\ \tilde{F}_{k,\mathbf{x}}(\mathbf{x}_t) &= \Pr(X_t \leq \mathbf{x}_t | U_t = k, X_{t+1} \leq \mathbf{x}).\end{aligned}$$

Then for $\mathbf{y}_{T:1} = (\mathbf{y}_T, \ldots, \mathbf{y}_1) \in \mathcal{S}^T$, and given $X_{T+1} = \mathbf{y}_0 \in \mathcal{S}$ and $U_{T+1} = k$ can be written as:

$$\begin{aligned}H_T(\mathbf{y}_{T:0}; k) &= \Pr(V_T \leq \mathbf{y}_{T:1} | U_{T+1} = k, X_{T+1} \leq \mathbf{y}_0) \\ &= \sum_{k_1 \cdots k_T} \tilde{\alpha}_{k,k_1} \prod_{t=2}^{T} \tilde{\alpha}_{k_{t-1},k_t} \prod_{t=1}^{T} \tilde{F}_{k_t,\mathbf{y}_{t-1}}(\mathbf{y}_t).\end{aligned}$$

Now, take any arbitrary points $\bar{\mathbf{x}} \in \mathcal{S}$ and $\mathbf{z}_j, \tilde{\mathbf{z}}_j \in \mathcal{S}^T$ for $j = 1, \ldots, K$. Define $\mathbf{z}_j^+ = (\bar{\mathbf{x}}, \mathbf{z}_j)$ and $\tilde{\mathbf{z}}_j^+ = (\bar{\mathbf{x}}, \tilde{\mathbf{z}}_j)$. The likelihood in (41), at these arbitrary points, for some j, can thus be formulated as:

$$\Pr(X_{1:2T+1} \leq (\tilde{\mathbf{z}}_j, \bar{\mathbf{x}}, \mathbf{z}_j)) = \sum_k H_T(\tilde{\mathbf{z}}_j^+; k) \pi_k F_k(\bar{\mathbf{x}}) G_T(\mathbf{z}_j^+; k). \tag{42}$$

Note the correspondence of the above equation to (39). Now consider $K \times K$ matrix:

$$\mathbf{G}_1 = (G_T(\mathbf{z}_j^+; k))_{k,j=1,\ldots,K} = (G_T((\bar{\mathbf{x}}, \mathbf{z}_j); k))_{k,j=1,\ldots,K}. \tag{43}$$

From Lemma 3 below we have that there exist $\mathbf{z}_1, \ldots, \mathbf{z}_K \in \mathcal{S}^T$ such that $\mathbf{G}_1$ is full rank, for any $\bar{\mathbf{x}}$. Similarly, we form matrix:

$$\mathbf{H}_1 = (H_T(\tilde{\mathbf{z}}_j^+; k))_{k,j=1,\ldots,K} = (H_T((\bar{\mathbf{x}}, \tilde{\mathbf{z}}_j); k))_{k,j=1,\ldots,K}. \tag{44}$$

Again, by Lemma 3, there exists $\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_K \in \mathcal{S}^T$ for which $\mathbf{H}_1$ has full rank.

Step 2 *(Identifying the distribution as a three-way array):* Now, let $\mathbf{v}, \tilde{\mathbf{v}} \in \mathcal{S}^T$ be *any* arbitrary points (*c.f.* (43),(44) focused on the existence of *some* points). Let $\bar{\mathbf{x}} \in \mathcal{S}$ also be any point but such that it matches always the one in $\mathbf{G}_1$ and $\mathbf{H}_1$. From Assumption 2, we have that $K \times 2$ matrix

$$\mathbf{M}_2 = [(F_i(\bar{\mathbf{x}}))_{i=1,\ldots,K}, \mathbf{1}_K] \tag{45}$$

has Kruskal rank of 2. From Step 1, the $K \times (K+2)$-matrices

$$\begin{aligned} &\mathbf{M}_3 = [\mathbf{G}_1, (G_T((\bar{\mathbf{x}}, \mathbf{v}); k))_{k=1,\ldots,K}, \mathbf{1}_K], \quad \mathbf{M}_1 = [\mathbf{H}_1, (H_T((\bar{\mathbf{x}}, \tilde{\mathbf{v}}); k))_{k=1,\ldots,K}, \mathbf{1}_K] \\ &\tilde{\mathbf{M}}_1 = \mathrm{diag}(\pi)\mathbf{M}_1, \end{aligned} \tag{46}$$

all have full ranks, K, where $\pi_k > 0 \ \ (k = 1, \ldots, K)$, and hence the Kruskal rank condition

$$R_\kappa(\tilde{\mathbf{M}}_1) + R_\kappa(\mathbf{M}_2) + R_\kappa(\mathbf{M}_3) \geq 2K + 2 \tag{47}$$

is satisfied for the three-way array

$$\mathbf{M}^\star = \langle \tilde{\mathbf{M}}_1, \mathbf{M}_2, \mathbf{M}_3 \rangle. \tag{48}$$

The question now is whether the distribution of $(X_1, \ldots, X_{2T+1})$ is alone sufficient to identify $\mathbf{M}^\star$. To see that this is the case, consider the following, exhaustive, possibilities:

For $i < K+2$; $j = 1$; $r < K+2$

$$\mathbf{M}^\star_{(i,1,r)} = \sum_{k=1}^{K} \pi_k H_T((\bar{\mathbf{x}}, \tilde{\mathbf{v}}_i); k) F_k(\bar{\mathbf{x}}) G_T((\bar{\mathbf{x}}, \mathbf{v}_r); k) = \Pr(X_{1:T} \leq \tilde{\mathbf{v}}_i, X_{T+1} \leq \bar{\mathbf{x}}, X_{T+2:2T+1} \leq \mathbf{v}_r)$$

For $i = K+2$; $j = 1$; $r < K+2$

$$\mathbf{M}^\star_{(K+2,1,r)} = \sum_{k=1}^{K} \pi_k F_k(\bar{\mathbf{x}}) G_T((\bar{\mathbf{x}}, \mathbf{v}_r); k) = \Pr(X_{T+1} \leq \bar{\mathbf{x}}, X_{T+2:2T+1} \leq \mathbf{v}_r)$$

For $i < K+2$; $j = 1$; $r = K+2$

$$\mathbf{M}^\star_{(i,1,K+2)} = \sum_{k=1}^{K} \pi_k H_T((\bar{\mathbf{x}}, \tilde{\mathbf{v}}_i); k) F_k(\bar{\mathbf{x}}) = \Pr(X_{1:T} \leq \tilde{\mathbf{v}}_i, X_{T+1} \leq \bar{\mathbf{x}})$$

For $i = K+2$; $j = 1$; $r = K+2$

$$\mathbf{M}^\star_{(K+2,1,K+2)} = \sum_{k=1}^{K} \pi_k F_k(\bar{\mathbf{x}}) = \Pr(X_{T+1} \leq \bar{\mathbf{x}}).$$

For $i < K+2$; $j = 2$; $r < K+2$

$$\begin{aligned} \mathbf{M}^\star_{(i,2,r)} &= \sum_{k=1}^{K} \pi_k H_T((\bar{\mathbf{x}}, \tilde{\mathbf{v}}_i); k) G_T((\bar{\mathbf{x}}, \mathbf{v}_r); k) = \Pr(X_{1:T} \leq \tilde{\mathbf{v}}_i, X_{T+2:2T+1} \leq \mathbf{v}_r | X_{T+1} \leq \bar{\mathbf{x}}) \\ &= \frac{\Pr(X_{1:T} \leq \tilde{\mathbf{v}}_i, X_{T+1} \leq \bar{\mathbf{x}}, X_{T+2:2T+1} \leq \mathbf{v}_r)}{\Pr(X_{T+1} \leq \bar{\mathbf{x}})} \end{aligned}$$

For $i = K+2$; $j = 2$; $r < K+2$

$$\mathbf{M}^\star_{(K+2,2,r)} = \sum_{k=1}^{K} \pi_k G_T((\bar{\mathbf{x}}, \mathbf{v}_r); k) = \Pr(X_{T+2:2T+1} \leq \mathbf{v}_r | X_{T+1} \leq \bar{\mathbf{x}})$$

For $i < K+2$; $j = 2$; $r = K+2$

$$\mathbf{M}^\star_{(i,2,K+2)} = \sum_{k=1}^{K} \pi_k H_T((\bar{\mathbf{x}}, \tilde{\mathbf{v}}_i); k) = \Pr(X_{1:T} \leq \tilde{\mathbf{v}}_i | X_{T+1} \leq \bar{\mathbf{x}})$$

For $i = K+2$; $j = 2$; $r = K+2$

$$\mathbf{M}^\star_{(K+2,2,K+2)} = 1$$

These are all uniquely determined by the joint distribution of $(X_1, \ldots, X_{2T+1})$ (joint distribution uniquely defines marginals).

Step 3 *(identifying parameters from three-way arrays):* Next assume we have an alternate set of parameters to those above; transition matrix $\widehat{\mathbf{A}}$, arbitrary initial state distribution $\hat{\pi}$ (not necessarily stationary), and distribution function $\widehat{F}_{u,\mathbf{x}}$ defined analogously to above. These parameters define matrices $\mathbf{N}_i (i = 1, 2, 3)$, which are defined, and evaluated at the same points, as $\mathbf{M}_i$ from above. Further, $\tilde{\mathbf{N}}_1 = \mathrm{diag}(\hat{\pi}\widehat{\mathbf{A}}^T)\mathbf{N}_1$, where $\hat{\pi}\widehat{\mathbf{A}}^T$ is the marginal distribution of U_{T+1}. If the two sets of parameters induce the same joint distribution $(X_1, \ldots, X_{2T+1})$ then Step 2 ensures that

$$\langle \tilde{\mathbf{M}}_1, \mathbf{M}_2, \mathbf{M}_3 \rangle = \langle \tilde{\mathbf{N}}_1, \mathbf{N}_2, \mathbf{N}_3 \rangle$$

And, due to Theorem 4a Kruskal (1977), since $\tilde{\mathbf{M}}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$ satisfy (40), there are $K \times K$ permutation matrix $\mathbf{P}$ and scaling matrices $\mathbf{\Lambda}_i, (i = 1, 2, 3)$ with $\mathbf{\Lambda}_1\mathbf{\Lambda}_2\mathbf{\Lambda}_3 = I_K$, such that

$$\mathbf{M}_i = \mathbf{\Lambda}_i \mathbf{P} \mathbf{N}_i \ (i = 2, 3) \quad \text{and} \quad \tilde{\mathbf{M}}_1 = \mathbf{\Lambda}_1 \mathbf{P} \tilde{\mathbf{N}}_1. \tag{49}$$

Since $\mathbf{M}_i, \mathbf{N}_i$ $(i = 2, 3)$ have only ones in the last column, $\mathbf{\Lambda}_2 = \mathbf{\Lambda}_3 = I_K$ and thus also $\mathbf{\Lambda}_1 = I_K$. The first consequence of this is that $H_T((\bar{\mathbf{x}}, \tilde{\mathbf{v}}); k)$, $F_u(\bar{\mathbf{x}})$, and $G_T((\bar{\mathbf{x}}, \mathbf{v}); k)$ are identified, up to simultaneous permutation of labels, for arbitrary $\mathbf{v}, \tilde{\mathbf{v}} \in \mathcal{S}$ and given $\bar{\mathbf{x}}$. But notice that we can construct above argumentation for any $\bar{\mathbf{x}}$.

Further, as $\tilde{\mathbf{M}}_1$ and $\mathbf{M}_3$ are full rank, then so must be $\tilde{\mathbf{N}}_1$ and $\mathbf{N}_3$. This in turn means that $\mathbf{P}$ is uniquely determined and $\pi = \hat{\pi}\widehat{\mathbf{A}}^T$ as they are both in the last columns of $\tilde{\mathbf{M}}_1 = \tilde{\mathbf{N}}_1$.

Step 4 *(identifying the transition matrix):* We show this for $T = K - 1$. In *Step 1*, we considered the matrix

$$\mathbf{G}_1 = (G_T((\mathbf{x}_0, \mathbf{z}_j); k))_{k,j=1,\ldots,K}. \tag{50}$$

Now consider instead a one time-step longer sequence, with only the first observation different, keeping labeling fixed:

$$\mathbf{G} = (G_{T+1}((\mathbf{x}, \mathbf{x}_0, \mathbf{z}_j); k))_{k,j=1,\ldots,K}. \tag{51}$$

From *Step 2*, $H_{T+1}(\cdot; k), F_k, G_{T+1}(\cdot; k)$ are identified up to joint label swapping and thus so is $\mathbf{G}$. $\mathbf{G}_1$ and $\mathbf{A}$ are related by

$$\mathbf{G} = \mathbf{A}\mathbf{D}_{\mathbf{x}}(\mathbf{x}_0)\mathbf{G}_1,$$

where $\mathbf{D}_{\mathbf{x}}(\mathbf{x}_0) = \mathrm{diag}(F_{1,\mathbf{x}}(\mathbf{x}_0), \ldots, F_{K,\mathbf{x}}(\mathbf{x}_0))$, and therefore

$$\mathbf{A} = \mathbf{G}\mathbf{G}_1^{-1}\mathbf{D}_{\mathbf{x}}(\mathbf{x}_0)^{-1}.$$

Thus $\mathbf{A}$ can be identified from above (for a large enough $\mathbf{x}_0$ as to avoid issues in the inverse), as all the constituents are identified, so $\mathbf{A} = \hat{\mathbf{A}}$. Also, as $\mathbf{A}$ is invertible and from above we can now get that $\pi = \hat{\pi}\mathbf{A}^T$, which combined with $\pi\mathbf{A}^{-1} = \pi$ gives $\pi = \hat{\pi}$. □

Lemma 2. *Let $t \leq K - 1$ and $\mathbf{B}_1, \ldots, \mathbf{B}_t$ be full-rank matrices in $\mathbb{R}^{K \times K}$ such that $\mathbf{B}_1\mathbf{1}_K, \ldots, \mathbf{B}_t\mathbf{1}_K$ are linearly independent vectors. Let $\mathbf{A}$ be a $K \times K$ full rank transition matrix, and $F_{1,\mathbf{x}_0}(\mathbf{x}), \ldots, F_{K,\mathbf{x}_0}(\mathbf{x})$ distribution functions satisfying Assumption 2. Then, for every $\mathbf{x}_0$, there exists some $\mathbf{x}^\star \in \mathcal{S}$ and $j \in \{1, \ldots, t\}$ for which the $K \times (t + 1)$-matrix*

$$[\mathbf{B}_1\mathbf{A}\mathbf{1}_K, \ldots, \mathbf{B}_t\mathbf{A}\mathbf{1}_K, \mathbf{B}_j\mathbf{A}\mathbf{D}_{\mathbf{x}_0}(\mathbf{x}^\star)\mathbf{1}_K]$$

is full rank.

Proof. Since $\mathbf{A}$ is a proper transition matrix, we have that

$$\mathbf{M} = [\mathbf{B}_1\mathbf{1}_K, \ldots, \mathbf{B}_t\mathbf{1}_K] = [\mathbf{B}_1\mathbf{A}\mathbf{1}_K, \ldots, \mathbf{B}_t\mathbf{A}\mathbf{1}_K],$$

and therefore $\mathbf{B}_1\mathbf{A}\mathbf{1}_K, \ldots, \mathbf{B}_t\mathbf{A}\mathbf{1}_K$ are linearly independent, with $S_1 = \mathrm{span}\{\mathbf{B}_1\mathbf{A}\mathbf{1}_K, \ldots, \mathbf{B}_t\mathbf{A}\mathbf{1}_K\}$ and $\dim(S_1) = t$. The Lemma can now be proven by contradiction. Assume that for any j, $\mathbf{B}_j\mathbf{A}\mathbf{D}_{\mathbf{x}_0}(\mathbf{x}^\star)\mathbf{1}_K$ is in the span S_1. We can write $\mathbf{Q}_j = \mathbf{B}_j\mathbf{A}$, and notice that this is full-rank for all j. Hence

$$\mathbf{B}_j\mathbf{A}\mathbf{D}_{\mathbf{x}_0}(\mathbf{x}^\star)\mathbf{1}_K = \mathbf{Q}_j\mathbf{D}_{\mathbf{x}_0}(\mathbf{x}^\star)\mathbf{1}_K = \sum_{i=1}^{K} F_{i,\mathbf{x}_0}(\mathbf{x}^\star)\mathbf{q}_{j,i},$$

where $\mathbf{q}_{j,i}$ denotes the i-th column vector of $\mathbf{Q}_j$, and we thus have a conic (positive) combination of K linearly independent vectors. If we consider all feasible $\mathbf{x}^\star$, this defines a subspace of conical hull in K dimensions. This contradicts the assumption of $\mathbf{B}_j\mathbf{A}\mathbf{D}_{\mathbf{x}_0}(\mathbf{x}^\star)\mathbf{1}_K$ being in the span S_1 for all $\mathbf{x}^\star$ and thus concludes the proof. □

Lemma 3. *Under Assumption 2 (of Lemma 1), for $T \geq K-1$ the conditional distributions of W_T given $U_{T+1} = k (k = 1, \ldots, K)$ and $X_{T+1} = \mathbf{x}_0 \in \mathcal{S}$, that is the functions $G_T((\mathbf{x}_0, \cdot); k)$, are linearly independent over $k = 1, \ldots, K$ for any fixed $\mathbf{x}_0$, and furthermore, there exist $\mathbf{z}_1, \ldots, \mathbf{z}_K \in \mathcal{S}^T$ such that the matrix*

$$\mathbf{G}_1 = (G_T(\mathbf{z}_j^+; k))_{k,j=1,\ldots,K} = (G_T(\mathbf{x}_0, \mathbf{z}_j; k))_{k,j=1,\ldots,K},$$

has full rank K.

Proof. Recall:

$$G_t((\mathbf{x}_0, \mathbf{x}_{1:t}); k) = \sum_{k_1 \cdots k_t} \alpha_{k,k_1} \prod_{s=2}^{t} \alpha_{k_{s-1},k_s} \prod_{s=1}^{t} F_{k_s,\mathbf{x}_{s-1}}(\mathbf{x}_s). \tag{52}$$

Define $K \times K$ stochastic diagonal matrix $\mathbf{D}_{\mathbf{x}_{t-1}}(\mathbf{x}_t) = \text{diag}(F_{k_t=1,\mathbf{x}_{t-1}}(\mathbf{x}_t), \ldots, F_{k_t=K,\mathbf{x}_{t-1}}(\mathbf{x}_t))$, and $\boldsymbol{\alpha}_l$ the l-th row vector of the transition matrix $\mathbf{A}$. We can then write:

$$G_t((\mathbf{x}_0, \mathbf{x}_{1:t}); k) = \boldsymbol{\alpha}_k \mathbf{D}_{\mathbf{x}_0}(\mathbf{x}_1) \mathbf{A} \mathbf{D}_{\mathbf{x}_1}(\mathbf{x}_2) \mathbf{A} \cdots \mathbf{A} \mathbf{D}_{\mathbf{x}_{t-1}}(\mathbf{x}_t) \mathbf{1}_K. \tag{53}$$

And define:

$$\tilde{G}_t((\mathbf{x}_0, \mathbf{x}_{1:t}); k) = F_{k,\mathbf{x}_0}(\mathbf{x}_1) \boldsymbol{\alpha}_k \mathbf{D}_{\mathbf{x}_1}(\mathbf{x}_2) \mathbf{A} \cdots \mathbf{A} \mathbf{D}_{\mathbf{x}_{t-1}}(\mathbf{x}_t) \mathbf{1}_K. \tag{54}$$

It follows that

$$\mathbf{G}_1 = \mathbf{A}(\tilde{G}_{K-1}((\mathbf{x}_0, \mathbf{z}_j); k)_{k,j=1,\ldots,K} = \mathbf{A}\tilde{\mathbf{G}}_1,$$

and therefore it suffices to prove the lemma for $\tilde{\mathbf{G}}_1$.

Proof by induction is used to show that there exist

$$\mathbf{z}_1^{(t)}, \ldots, \mathbf{z}_{t+1}^{(t)} \in \mathcal{S}^t \ (t = 1, \ldots, K-1), \tag{55}$$

for which the vectors (i.e. columns of $\tilde{\mathbf{G}}_1^{(t)}$)

$$\mathbf{v}_j^{(t)} = \left[\tilde{G}_t((\mathbf{x}_0, \mathbf{z}_j^{(t)}; 1) \quad \cdots \quad \tilde{G}_t((\mathbf{x}_0, \mathbf{z}_j^{(t)}; K)\right]' \ (j = 1, \ldots, t+1), \tag{56}$$

are linearly independent, and $\mathbf{v}_1^{(t)}$ has strictly positive entries. Note that the superscipts (t) are used just to keep track of the t being considered. The case $t = K-1$ will establish the lemma. In other words, we will only prove the theorem up to $T = K-1$. Since marginal distributions of linearly dependent distributions remain linearly dependent, linear independence follow for any $T \geq K-1$, and the existence of corresponding points $\mathbf{z}_1, \ldots, \mathbf{z}_K \in \mathcal{S}^T$ follows from Lemma 17 in Allman et al. (2009).

Proof by induction – base case: Set $t = 1$. In this instance, $\tilde{\mathbf{G}}_1^{(1)}$ is $K \times 2$, with columns given by:

$$\mathbf{v}_j^{(1)} = \left[F_{1,\mathbf{x}_0}(\mathbf{z}_j^{(1)}), \quad \cdots, \quad F_{K,\mathbf{x}_0}(\mathbf{z}_j^{(1)})\right]' \ (j = 1, 2).$$

By Assumption 2, the K density functions with fixed $\mathbf{x}_0$ are distinct and therefore $\mathbf{v}_1^{(1)}$ and $\mathbf{v}_2^{(1)}$ are linearly independent with $\mathbf{v}_1^{(1)}$ strictly positive.

Proof by induction – induction step: For induction, suppose that the claim (55)-(56) holds for some $t < K-1$. Equation (56) can be rewritten as:

$$\begin{aligned}
\mathbf{v}_j^{(t)} &= [F_{1,\mathbf{x}_{0,j}}(\mathbf{x}_{1,j}) \boldsymbol{\alpha}_1 \mathbf{D}_{\mathbf{x}_{1,j}}(\mathbf{x}_{2,j}) \mathbf{A} \cdots \mathbf{A} \mathbf{D}_{\mathbf{x}_{t-1,j}}(\mathbf{x}_{t,j}) \mathbf{1}_K, \cdots, \\
&\quad F_{K,\mathbf{x}_{0,j}}(\mathbf{x}_{1,j}) \boldsymbol{\alpha}_K \mathbf{D}_{\mathbf{x}_{1,j}}(\mathbf{x}_{2,j}) \mathbf{A} \cdots \mathbf{A} \mathbf{D}_{\mathbf{x}_{t-1,j}}(\mathbf{x}_{t,j}) \mathbf{1}_K]' \\
&= [\underbrace{\mathbf{D}_{\mathbf{x}_{0,j}}(\mathbf{x}_{1,j}) \mathbf{A} \mathbf{D}_{\mathbf{x}_{1,j}}(\mathbf{x}_{2,j}) \mathbf{A} \cdots \mathbf{A} \mathbf{D}_{\mathbf{x}_{t-1,j}}(\mathbf{x}_{t,j})}_{\mathbf{B}_j(\mathbf{z}_j^{(t)})} \mathbf{1}_K],
\end{aligned} \tag{57}$$

and thus we have

$$\tilde{\mathbf{G}}_1^{(t)} = [\mathbf{v}_1^{(t)}, \cdots, \mathbf{v}_{t+1}^{(t)}] = [\mathbf{B}_1(\mathbf{z}_1^{(t)}) \mathbf{1}_K, \cdots, \mathbf{B}_{t+1}(\mathbf{z}_{t+1}^{(t)}) \mathbf{1}_K].$$

All $\mathbf{B}_j (j = 1, \ldots, t+1)$ are full rank, and by the inductive assumption all the vectors are linearly independent. It follows from Lemma 2 that there exists $j \in (1, \ldots, t+1)$ and $\mathbf{x}^\star$ for which the $K \times (t+2)$ matrix:

$$\mathbf{M} = \left[\mathbf{B}_1(\mathbf{z}_1^{(t)})\mathbf{A}\mathbf{1}_K, \ldots, \mathbf{B}_{t+1}(\mathbf{z}_{t+1}^{(t)})\mathbf{A}\mathbf{1}_K, \mathbf{B}_j(\mathbf{z}_j^{(t)})\mathbf{A}\mathbf{D}_{\mathbf{x}_t}(\mathbf{x}^\star)\mathbf{1}_K\right] \tag{58}$$

has full rank $t+2$,and hence a $(t+2) \times (t+2)$ submatrix of non-zero determinant. Since $\mathbf{D}_{\mathbf{x}_{t-1}}(\mathbf{x}_t) \to I$ when $\mathbf{x}_t \to \infty$,

$$\left[\mathbf{B}_1(\mathbf{z}_1^{(t)})\mathbf{A}\mathbf{D}_{\mathbf{x}_t}(\mathbf{x})\mathbf{1}_K, \ldots, \mathbf{B}_{t+1}(\mathbf{z}_{t+1}^{(t)})\mathbf{A}\mathbf{D}_{\mathbf{x}_t}(\mathbf{x})\mathbf{1}_K, \mathbf{B}_j(\mathbf{z}_j^{(t)})\mathbf{A}\mathbf{D}_{\mathbf{x}_t}(\mathbf{x}^\star)\mathbf{1}_K\right] \to \mathbf{M}, \quad \mathbf{x} \to \infty \tag{59}$$

and hence the corresponding submatrix will also have non-zero determinant in above, for an appropriate $\mathbf{x} \in \mathcal{S}$. Notice also how above defines $\mathbf{v}_j^{(t+1)}$ $(j = 1, \ldots, t+2)$, as per equation (57). Therefore the claim for $t+1$ is satisfied by setting

$$\mathbf{z}_s^{(t+1)} = \left[\mathbf{z}_s^{(t)}, \mathbf{x}\right] \ (s = 1, \ldots, t+1) \quad \mathbf{z}_{t+2}^{(t+1)} = \left[\mathbf{z}_j^{(t)}, \mathbf{x}^\star\right] \tag{60}$$

and so the proof concludes.

□

D Implementation Detail for Simulation 1

We give here more detail on the data generation, training, and evaluation for IIA-GCL in Simulation 1 (Section 4.1).

Data Generation We generated data from an artificial NVAR process with time-index-parameterized nonstationary innovations. The nonstationary innovations were randomly generated from a Gaussian distribution by modulating its mean and standard deviation across time t; i.e., the auxiliary variable $\mathbf{u}_t = t$, and $\log p(s_i(t)) \propto -\lambda_{i1}(t)s_i(t)^2 - \lambda_{i1}(t)\lambda_{i2}(t)s_i(t)$, where $\lambda_{i1}(t)$ and $\lambda_{i2}(t)$ control the standard deviation and mean of the i-th component at time point t, respectively. Each of $\lambda_{i1}(t)$ and $\lambda_{i2}(t)$ was modeled to be temporally smooth and continuous, by 1) obtaining a combination of Fourier basis functions spanning the whole time series (sine and cosine bases with 64 frequencies), which weights were randomly selected from uniform distribution, 2) normalizing to $[-2, 2]$, and 3) (only for $\lambda_{i1}(t)$) putting into exponential function. The dimensions of the observations and innovations (n) were 20. As the NVAR model, we used a multilayer perceptron we call NVAR-MLP, which takes a concatenation of $\mathbf{x}_{t-1}$ and $\mathbf{s}_t$ as an input, then outputs $\mathbf{x}_t$. To guarantee the invertibility, we fixed the number of units of each layer to n, and used leaky ReLU units for the nonlinearity except for the last layer which has no-nonlinearity.

Training Considering the innovation model with $\mathbf{u}_t = t$, we here used IIA-GCL for the estimation of the latent innovations. We adopted MLPs as the nonlinear scalar functions in Eq. 6. The MLP for $\mathbf{h}$ (*h-MLP*) outputs n-dimensional feature values from an input $(\mathbf{x}_t, \mathbf{x}_{t-1})$, which is supposed to represent the latent innovations after the training. The number of layers was selected to be the same as that of the NVAR-MLP, and the number of node in each layer was $4n$ except for the output layer (n), so as to make it have enough number of parameters as the demixing model. A *maxout* unit was used as the activation function in the hidden layers, which was constructed by taking the maximum across two affine fully connected weight groups, while no-nonlinearity was applied at the last layer. The scalar functions ψ_{ij}, μ_{ij}, and $\alpha(\mathbf{u}_t)$ were modeled to be consistent with the NVAR model; i.e., we incorporated the information into the model that 1) the innovations were generated based on the Gaussian distribution with mean and std modulations by the log-pdf shown above, and 2) λ_{i1} and λ_{i2} were generated through a combination of Fourier basis functions with 64 frequencies, while their weights have to be estimated from the data. For ϕ, which has dependency on $\mathbf{u}_t$, we used the same structure as the combination of $\mathbf{h}$, ψ_{ij}, and μ_{ij} explained above, which we call ϕ-MLP, except that the ϕ-MLP takes a single data point $(\mathbf{x}_{t-1})$ as an input, instead of a set of the consecutive points $(\mathbf{x}_t, \mathbf{x}_{t-1})$. The regression function also needs additional terms representing the marginal distributions of $\mathbf{s}$ and $\mathbf{x}$ (β and γ), which were here modeled by the weighted squared sum of the output units of the h-MLP and ϕ-MLP, respectively.

The nonlinear regression function was trained by back-propagation with a momentum term so as to discriminate the real dataset from its $\mathbf{u}_t$-randomized version. The initial parameters were randomly drawn from a uniform

distribution. The performance was evaluated by the Pearson correlation between the true innovations and the estimated feature values $\mathbf{h}$. It was averaged over 10 runs, for each setting of the complexity (number of layers) $L \in [1, 3, 5]$ of the NVAR-MLP and the number of data points.

For comparison, we also applied NICA based on GCL (NICA-GCL; Hyvärinen et al. (2019)), an NVAR with additive innovation model (AD-NVAR), and variational autoencoder (VAE; Kingma and Welling (2014)) to the same data. For all of them, we fixed the number of layers of the demixing model to be the same as that of the NVAR-MLP. We fixed $L \in [1, 2]$ exceptionally for VAE because of the instability of training in high layer models. See Supplementary Material I for the details of the baseline methods.

E Implementation Detail for Simulation 2

We give here more detail on the training for IIA-TCL in Simulation 2 (Section 4.2).

Training We applied IIA-TCL to the same data used in Simulation 1. For IIA-TCL, we first divided the time series into 256 equally-sized segments, and used the segment label as the auxiliary variable $\mathbf{u}_t$; i.e., we assume that the data are segment-wise stationary, Although this assumption is not consistent with the real innovation model (Simulation 1), it is approximately true because the modulations were temporally smooth and continuous; we thus consider here data with a realistic deviation from model assumptions. We adopted MLPs as the nonlinear scalar functions in the regression function (Eq. 8). The architecture of the MLP for $\mathbf{h}$ (*h-MLP*) was the same as that in Simulation 1. Considering the log-pdf of the innovation, we fixed $\psi_{i1}(y_i) = y_i^2$, and $\psi_{i2}(y_i) = y_i$. For ϕ, which has dependency on $\mathbf{u}_t$, we used the same structure as the combination of $\mathbf{h}$, ψ_{ij}, and $w_{ij\tau}$, except that ϕ takes a single data point ($\mathbf{x}_{t-1}$) as an input, instead of a set of consecutive points ($\mathbf{x}_t, \mathbf{x}_{t-1}$). The training and evaluation methods follow those in Simulation 1. We discarded the cases of small data sets (2^{10} and 2^{12}, corresponding to 4 and 16 samples in a segment) because of the instability of training.

For comparison, we also applied NICA (TCL; Hyvärinen and Morioka (2016)). See Supplementary Material I for the details of the baseline methods.

F Simulation 2 in two-dimensional space

We conducted an additional simulation to visually demonstrate the advantage of the IIA framework. The settings were the same to Simulation 4.2 (see Supplementary Material E) except that the dimensions of the observations and the innovations were two, the number of layers was 5, and the number of data points was 2^{18}. The estimated innovations by IIA-TCL looks clearly better demixed compared to the baseline methods (AD-NVAR and NICA-TCL; see Supplementary Material I for the details).

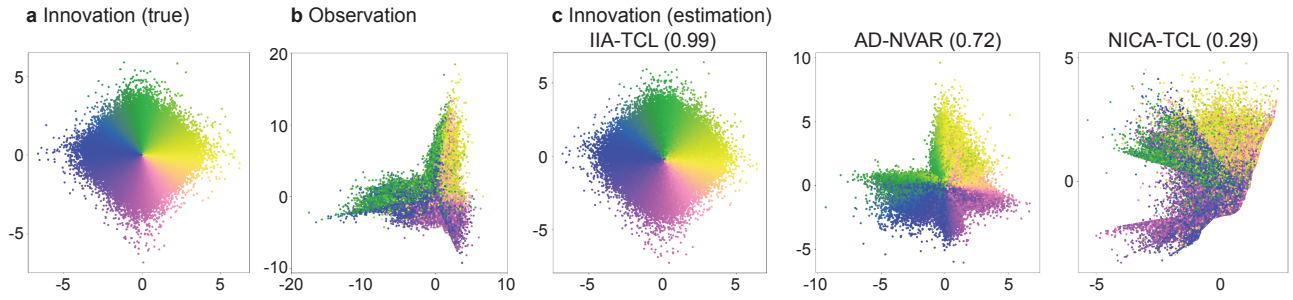

Figure 3: Estimation of the latent innovations from unknown artificial two dimensional NVAR process. (**a**) Scatter plot of the true innovations. (**b**) Observations. (**c**) Innovations estimated by IIA-TCL, and for comparison, by AD-NVAR, and NICA-TCL. The values show the mean absolute correlation coefficients between innovations and their estimates.

G Implementation Detail for Simulation 3

We give here more detail on the training for IIA-HMM in Simulation 3 (Section 4.3).

Data Generation We generated data from an artificial NVAR with hidden Markov chain. The innovations were generated based on the method used in Hälvä and Hyvärinen (2020). Briefly, the innovations were generated by Gaussian emission distributions of an HMM with C discrete states, where the means and the variances of the Gaussian distribution were selected to be distinctive across components/states. The transition matrix was defined to have 99% probability to stay at the current state, and 1% probability to switch to the next state, in cyclic manner. We fixed the dimension of the innovations (n) to 5, and the number of latent states was set to $C = 2n + 1$. The observations were then obtained by randomly generated NVAR-MLP (see Supplementary Material D), using the generated innovations.

Training We used here EM algorithm to maximize the likelihood for estimating the demixing model $\mathbf{h}$ based on MLP (h-MLP), the transition probability matrix $\mathbf{A}$, the latent state at each data point, and the mean and the variance parameters of each state. The implementation is based on that of NICA-HMM (Hälvä and Hyvärinen (2020); github.com/HHalva/hmnlica), with some differences such as the demixing model and the incorporation of the margianl distribution $p(\mathbf{x}_0)$ (see Eq. 10). Although the likelihood includes the determinant of the Jacobian, which is widely considered difficult to compute, we can numerically calculate its gradient by utilizing recent developments of the numerical calculation of gradients (here, JAX library). The number of layers of h-MLP was selected to be the same as that of the NVAR-MLP, and the number of node in each layer was $2n$ except for the output layer (n). A smooth version of leaky ReLU was used as the activation function in the hidden layers; $y = ax + (1 - a)\log(1 + \exp^x)$, where x is the input, y is the output, and a is the leak coefficient. This type of differentiable function is useful for the stable estimation by the EM algorithm. No-nonlinearity was applied at the last layer. For better initialization of the h-MLP parameters than the random values, we firstly applied IIA-TCL to the observation with assuming segment-wise stationarity (length of segments was 32), then used it as the initial values of the h-MLP. Due to the sensitivity of the algorithm to the initial values of the parameters, we repeated the estimation 20 times with different initializations, then selected the one with the highest likelihood. The evaluation methods follow those in Simulation 1. For comparison, we also applied NICA based on HMM (NICA-HMM; Hälvä and Hyvärinen (2020)), an NVAR with additive innovation model (AD-NVAR), and IIA-TCL which was also used as the initialization. For all of them, we fixed the number of layers of the demixing model to be the same as that of the NVAR-MLP. See Supplementary Material I for the details of the baseline methods.

H Detail for Experiments on Real Brain Imaging Data

Data and Preprocessing We used a publicly available MEG dataset (Westner et al. (2018); https://doi.org/10.17605/OSF.IO/M25N4). Briefly, the participants were presented with a random word selected from 420 unrelated German nouns (duration = 697 ± 119 ms) either visually (projected centrally on a screen) or auditorily (via nonferromagnetic tubes to both ears) randomly for each trial. The stimulus was followed by a visual fixation cross until the end of the trial (2000 ms after the stimulus onset). MEG signals were measured from twenty healthy volunteers by a 148-channel magnetometer (MAGNES 2500 WH, 4D Neuroimaging, San Diego, USA) inside a magnetically shielded room. The data were downsampled to 300 Hz, and epoched into trials. The contaminated trials were rejected by visual inspections, and thereafter the blinks, eye movements, and cardiac artifacts were corrected using ICA (see Westner et al. (2018) for more details of the preprocessing). We further band-pass filtered the data between 4 Hz and 125 Hz, normalized them to have zero-mean and unit variance at the base line period ($-1{,}000$ ms to 0 ms) for each channel and trial, and then cropped from -300 ms to 2,000 ms after the onset for each trial. The dimension of the data was reduced to 30 by PCA. There were 219.1 ± 22.4 trials (110.4 ± 11.5 for auditory and 108.7 ± 11.9 for visual) for each subject, and in total, 2,207 auditory and 2,174 visual trials in the whole dataset.

IIA Settings We used IIA-TCL for the training, by assuming a third-order NVAR model (NVAR(3)) and the segment-wise-stationarity of the latent innovations. The trial data were divided into 84 equally sized segments of length of 8 samples (26.7 ms), and the segment label was used as the auxiliary variable $\mathbf{u}_t$. The same segment labels were given across the trials; however, considering the possible stimulus-specific dynamics of the brain, we assigned different labels for the auditory and visual trials. In total, there are 168 segments (classes) to be discriminated by MLR. The network architectures of the MLPs are the same with those in Simulation 2, except that $\mathbf{h}$ and ϕ take $\mathbf{x}_{t:t-3}$ and $\mathbf{x}_{t-1:t-3}$ as inputs, respectively, the number of units of each layer was fixed to 30, and that of the last layer (number of components) was 5. The smaller number of components than the data dimension can be justified by assuming the stationarity of the remaining components (the remaining innovations

do not depend on $\mathbf{u}$; Hyvärinen and Morioka (2016)). Considering the fast sampling rate of the data (300 Hz), we fixed the time lag between two consecutive samples to 3 (10 ms). The other settings were as in Simulation 2. The training of a four-layer model by IIA-TCL took about 2 hours (Intel Xeon 3.5 GHz 16 core CPUs, 376 GB Memory, NVIDIA Tesla V100 GPU).

Evaluation Methods For evaluation, we performed classification of the stimulus modality (auditory or visual) by using the estimated innovations. The classification was performed using a linear support vector machine (SVM) classifier trained on the stimulation label and sliding-window-averaged innovations (width=16 and stride=8 samples) obtained for each trial. The performance was evaluated by the generalizability of a classifier across subjects, i.e., one-subject-out cross-validation (OSO-CV); the feature extractor and the classifier were trained only from the training subjects, and then applied to the held-out subject. The hyperparameters of the SVM were determined by nested OSO-CV without using the test data. For comparison, we also applied NICA based on TCL (Hyvärinen and Morioka, 2016) and AD-NVAR(3) (See Supplementary Material I for the details of the baseline methods, with changing $\mathbf{x}_{t-1}$ to $\mathbf{x}_{t-1:t-3}$). We additionally applied principal component analysis (PCA) to the estimations by AD-NVAR(3) before applying linear ICA to reduce the dimension to 5 for fair comparisons. We omitted $L = 1$ for IIA-TCL because of the instability of training.

We also visualized the spatial characteristics of each innovation component by estimating the optimal (maximal and minimal) input $\mathbf{x}_t$ while fixing $\mathbf{x}_{t-1:t-3}$ to zero. This method is commonly used in deep learning studies to visualize the input specificities of a hidden node of a neural network. We used l_2 regularization on the input to avoid overfitting.

I Details of the baseline methods

NVAR with additive innovation model (AD-NVAR) AD-NVAR assumes NVAR with additive innovation model:

$$\mathbf{x}_t = \mathbf{f}_{\mathrm{ad}}(\mathbf{x}_{t-1}) + \mathbf{s}_t, \tag{61}$$

where $\mathbf{f}_{\mathrm{ad}} : \mathbb{R}^n \to \mathbb{R}^n$ is an unknown mixing model, and $\mathbf{s}_t \in \mathbb{R}^n$ is the latent innovations to be estimated. For the estimation, we firstly estimate the mixing model from the observable time series, which can be done practically by training a nonlinear predictor which takes $\mathbf{x}_{t-1}$ as an input and then outputs the estimation of $\mathbf{x}_t$ so as to minimize the mean squared prediction errors. The error term was then used as the estimation of the additive innovation $\mathbf{s}_t$. Since the obtained components are not guaranteed to be mutually independent, we additionally applied linear ICA based on nonstationarity of variance (NSVICA; Hyvärinen (2001)) to the estimated additive innovations for fair comparisons. For the mixing model $\mathbf{f}_{\mathrm{ad}}$, we used an MLP with the similar architecture as IIA, except for the difference of the dimension of the input.

Variational auto encoder (VAE) We used VAE (Kingma and Welling, 2014) as a baseline of unsupervised representation learning frameworks. VAE assumes that the latent variables have spherical Gaussian distribution, then embed data into the latent space in an unsupervised manner by training an encoder, which embeds the data into the latent space, and a decoder, which reconstructs the input from the latent variables, so as to minimize the reconstruction error. In the simulations, we trained an encoder based on an MLP, which nonlinearly embeds an input $(\mathbf{x}_t, \mathbf{x}_{t-1}) \in \mathbb{R}^{2n}$ into an n-dimensional feature space representing the estimation of the innovation. The number of nodes in each layer was designed to linearly decrease from input $(2n)$ to the output (n). We additionally applied linear ICA based on nonstationarity of variance (NSVICA; Hyvärinen (2001)) to the estimated innovations for fair comparisons because VAE does not assume the independence on the estimations.

Nonlinear Independent component analysis (NICA) NICA assumes instantaneous nonlinear mixture model:

$$\mathbf{x}_t = \mathbf{f}_{\mathrm{ICA}}(\mathbf{s}_t), \tag{62}$$

where $\mathbf{f}_{\mathrm{ICA}} : \mathbb{R}^n \to \mathbb{R}^n$ is the instantaneous (nonlinear) mixture function, and $\mathbf{s}_t$ is the latent components. Since general NICA problem has indeterminacy (Hyvärinen and Pajunen, 1999), we need some assumptions on the latent components to guarantee the identifiability, similarly to the IIA frameworks. We here used NICA-GCL (Hyvärinen et al., 2019), NICA-TCL (Hyvärinen and Morioka, 2016), and NICA-HMM (Hälvä and Hyvärinen, 2020) for comparison, which the basic proofs of IIA are based on. In the simulations, we estimated the independent components by the similar architecture as IIA (e.g., latent components assumptions, MLPs, and so on), except that it assumed the instantaneous mixture model for the observation.

References

G. Alexandrovich, H. Holzmann, and A. Leister. Nonparametric identification and maximum likelihood estimation for hidden markov models. *Biometrika*, 103(2):423–434, 2016.

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37:3099–3132, 2009.

J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95 – 138, 1977.