Supplementary Material for Identifiable Autoregressive Variational Autoencoders for Nonlinear and Nonstationary Spatio-Temporal Blind Source Separation ⋆

Mika Sipilä[1](✉), Klaus Nordhausen[2], and Sara Taskinen[1]

[1] Department of Mathematics and Statistics, University of Jyvaskyla, Finland
[2] Department of Mathematics and Statistics, University of Helsinki, Finland

A Lemmas for the autoregressive exponential families

In this section, some useful Lemmas are given for univariate autoregressive exponential family distributions.

Definition 3 (Autoregressive models). *A generative model of x is considered to be autoregressive, if it can be written as*

$$x(\boldsymbol{\theta}^t) = \mu(\boldsymbol{\theta}^t) + \sum_{r=1}^{R} \gamma_r(\boldsymbol{\theta}^t)\Big(x^{t-r} - \mu(\boldsymbol{\theta}^{t-r})\Big) + \omega(\boldsymbol{\theta}^t), \tag{14}$$

where $\boldsymbol{\theta}^t \in \mathbb{R}^m$ is the parameter vector at time step t, μ is a trend function, $\gamma_1, \dots, \gamma_R$ are the functions for autoregressive coefficients and ω is white noise so that $E(\omega(\boldsymbol{\theta}^t)) = 0$ and $Var(\omega(\boldsymbol{\theta}^t)) < \infty$ for all $t = 1, \dots, T$, and $Cov(\omega(\boldsymbol{\theta}^t), \omega(\boldsymbol{\theta}^{t'})) = 0$ for all $t \neq t'$. To ensure local weak-sense stationarity for each t, the (complex) roots y_i of the polynomial $1 - \sum_{i=1}^{R} \gamma_i(\theta_{t-i}) y^i$ must satisfy $|y_i| > 1$.

Definition 4 (Autoregressive exponential family). *Assume an autoregressive model defined by Definition 3. The univariate distribution $p(x^t | \{x^{t-1:t-R}; \boldsymbol{\theta}^t\})$ belongs in univariate autoregressive exponential family, if its probability distribution can be written as*

$$p(x^t | \{x^{t-1:t-R}; \boldsymbol{\theta}^t\}) = \frac{Q(x^t, \{x^{t-1:t-R}\})}{Z(\boldsymbol{\theta}^t)} e^{\sum_{j=1}^{k} T_j(x^t, \{x^{t-1:t-R}\}) \lambda_j(\boldsymbol{\theta}^t)}, \tag{15}$$

where Q is a base measure, Z is a normalizing constant, $T_1, \dots, T_k$ are sufficient statistics and $\boldsymbol{\theta}^t$ is a parameter vector at time point t. The dimension $k \in \{1, 2, \dots\}$ is assumed to be minimal, meaning that the distribution p cannot be written in form (15) using a smaller $k' < k$.

Lemma 1. *Consider autoregressive exponential family distribution. The components of sufficient statistics $\boldsymbol{T}$ of the distribution are linearly independent. In other words, if there exists $\boldsymbol{\alpha} \in \mathbb{R}^k$ so that $\alpha_1 T_1(x^t, \{x^{t-1:t-R}\}) + \dots + \alpha_k T_k(x^t, \{x^{t-1:t-R}\}) = \mathbf{0}$, then $\boldsymbol{\alpha} = \mathbf{0}$.*

Proof: Assume that the components of $\boldsymbol{T}$ are not linearly independent. Then, there exists $\boldsymbol{\alpha} \in \mathbb{R}^k$, $\boldsymbol{\alpha} \neq \mathbf{0}$, meaning that for some $i \in \{1, \dots, k\}$, $\alpha_i \neq 0$. By reordering the indices, we can assume that $\alpha_k \neq 0$. Then, we can write $T_k(x^t, \{x^{t-1:t-R}\}) = \sum_{j=1}^{k-1} \frac{\alpha_i}{\alpha_k} T_k(x^t, \{x^{t-1:t-R}\})$. Let $\lambda_j^*(\boldsymbol{\theta}^t) := (\lambda_j(\boldsymbol{\theta}^t) + \frac{a_j}{a_k} \lambda_k(\boldsymbol{\theta}^t))$. Then, the term in the exponent of (15) can be written as

$$\sum_{j=1}^{k} T_j(x^t, \{x^{t-1:t-R}\}) \lambda_j(\boldsymbol{\theta}^t) = \sum_{j=1}^{k-1} T_j(x^t, \{x^{t-1:t-R}\}) \lambda_j(\boldsymbol{\theta}^t) + \sum_{j=1}^{k-1} \frac{\alpha_i}{\alpha_k} T_k(x^t, \{x^{t-1:t-R}\}) \tag{16}$$

$$= \sum_{j=1}^{k-1} T_j(x^t, \{x^{t-1:t-R}\}) \left(\lambda_j(\boldsymbol{\theta}^t) + \frac{a_j}{a_k} \lambda_k(\boldsymbol{\theta}^t)\right) \tag{17}$$

$$= \sum_{j=1}^{k-1} T_j(x^t, \{x^{t-1:t-R}\}) \lambda_j^*(\boldsymbol{\theta}^t), \tag{18}$$

which contradicts the minimality of k in Definition 4. □

⋆ This research was supported by the Research Council of Finland (363261, 453691) and the Vilho, Yrjö and Kalle Väisälä foundation.

Definition 5 (Strongly exponential autoregressive distributions). *Exponential autoregressive distribution is considered strongly exponential if the following holds:*

$$(\exists\, \boldsymbol{\theta}^t \in \mathbb{R}^m \,|\, \forall\, x^t, ..., x^{t-R} \in \mathcal{X},\ \sum_{j=1}^{k} T_j(x^t, \{x^{t-1:t-R}\})\lambda_j(\boldsymbol{\theta}^t) = const) \implies l(\mathcal{X}) = 0 \ or\ \boldsymbol{\lambda}(\boldsymbol{\theta}^t) = \mathbf{0}, \tag{19}$$

where l is a Lebesgue measure.

Definition 5 says that strongly exponential distribution has the exponential component in its expression almost surely, and the distribution can be reduced only to base measure and normalizing constant on a set of measure zero.

Lemma 2. *Consider a strongly exponential autoregressive family distribution whose sufficient statistics $\boldsymbol{T}$ are differentiable almost everywhere. Then, $T_j' \neq 0$ for all $j = 1, \ldots, k$ almost everywhere on $\mathbb{R}$.*

Proof: Assume that p is strongly exponential autoregressive distribution. Let $\mathcal{X} = \cup_j \{x \in \mathbb{R}, T_j'(x) = 0\}$ and select any $\boldsymbol{\theta}$ for which $\boldsymbol{\lambda}(\boldsymbol{\theta}^t) \neq \mathbf{0}$. Then, for all $x \in \mathcal{X}$, it holds that

$$\sum_{j=1}^{k} T_j'(x^t, \{x^{t-1:t-R}\})\lambda_j(\boldsymbol{\theta}^t) = 0 \tag{20}$$

$$\implies \sum_{j=1}^{k} T_j(x^t, \{x^{t-1:t-R}\})\lambda_j(\boldsymbol{\theta}^t) = \text{const.} \tag{21}$$

By Definition 5, this means that $l(\mathcal{X}) = 0$. □

Lemma 3. *Consider a strongly exponential autoregressive family distribution of size $k \geq 2$ so that the sufficient statistics $\boldsymbol{T}$ are differentiable almost everywhere. Then, there exist k distinct points $(x_1^t, \ldots, x_1^{t-R})$, $\ldots, (x_k^t, \ldots, x_k^{t-R})$ such that the vectors $\boldsymbol{T}'(x_1^t, \{x_1^{t-1:t-R}\}), \ldots, \boldsymbol{T}'(x_k^t, \{x_k^{t-1:t-R}\})$ are linearly independent in $\mathbb{R}^k$.*

Proof: Suppose that for any choice of such k points, the vectors $\boldsymbol{T}'(x_1^t, \{x_1^{t-1:t-R}\}), \ldots, \boldsymbol{T}'(x_k^t, \{x_k^{t-1:t-R}\})$ are not linearly independent, meaning that there are a subspace of $\mathbb{R}^k$ of dimension ar most $k-1$ in which $\boldsymbol{T}'(\mathbb{R}^R)$ is included in. Thus, there exists $\boldsymbol{\theta}^t$ such that $\boldsymbol{\lambda}(\boldsymbol{\theta}) \in \mathbb{R}^k$ is a non-zero vector that is orthogonal to $\boldsymbol{T}'(\mathbb{R}^R)$. Because of the orthogonality, it holds for all $x_t, \ldots, x^{t-R} \in \mathbb{R}$ that $\sum_{j=1}^{k} T_j'(x^t, \{x^{t-1:t-R}\})\lambda_j(\boldsymbol{\theta}^t) = 0$. By integrating, we find that $\sum_{j=1}^{k} T_j(x^t, \{x^{t-1:t-R}\})\lambda_j(\boldsymbol{\theta}^t) = \text{const}$. Since $\boldsymbol{\lambda}(\boldsymbol{\theta}^t)) \neq \mathbf{0}$ and $l(\mathbb{R}) \neq 0$, the distribution cannot be strongly exponential, which contradicts the hypothesis.

Lemma 4. *Consider a strongly exponential autoregressive distribution of size $k \geq 2$ for which the sufficient statistics $\boldsymbol{T}$ are twice differentiable almost everywhere. Then it holds that*

$$rank\left((T_1'(x^t, \{x^{t-1:t-R}\}), T_1''(x^t, \{x^{t-1:t-R}\})^\top, \ldots, (T_k'(x^t, \{x^{t-1:t-R}\}), T_k''(x^t, \{x^{t-1:t-R}\})^\top\right) \geq 2 \tag{22}$$

almost everywhere on $\mathbb{R}$.

Proof: Suppose there exists a set $\mathcal{X}$ so that $l(\mathcal{X}) > 0$, but the equation (22) does not hold. In other words, for all $j \in \{1, \ldots, k\}$ and $x \in \mathcal{X}$, the vectors $(T_j'(x^t, \{x^{t-1:t-R}\}), T_j''(x^t, \{x^{t-1:t-R}\})^\top$ are collinear. This means that there exists a vector $\boldsymbol{\alpha} \in \mathbb{R}^k$, $\boldsymbol{\alpha} \neq 0$, so that $\sum_{j=1}^{k} \alpha_j T_j'(x^t, \{x^{t-1:t-R}\}) = 0$. By integrating, we get $\sum_{j=1}^{k} \alpha_j T_j(x^t, \{x^{t-1:t-R}\}) = \text{const}$ for all $x \in \mathcal{X}$. Since $l(\mathcal{X}) > 0$, this contradicts the hypothesis.

Lemma 5. *Consider P strongly exponential autoregressive distributions of size $k \geq 2$ for which the sufficient statistics $\boldsymbol{T}_j$, $j = 1, \ldots, P$ are twice differentiable almost everywhere. Let $\boldsymbol{x} := (\boldsymbol{x}_1, \ldots, x_P) \in \mathbb{R}^P$ and $\boldsymbol{e}^{(j,i)}(x_i) = (0, \ldots, 0, T_{j,i}'(x_i), T_{j,i}''(x_i), 0, \ldots, 0) \in \mathbb{R}^{2P}$, so that the non-zero entries are at indices $(2j, 2j+1)$. Then the matrix $\boldsymbol{E}(\boldsymbol{z}) = (\boldsymbol{e}^{(1,1)}(x_1), \ldots, \boldsymbol{e}^{(1,k)}(x_1), \ldots, \boldsymbol{e}^{(P,1)}(x_P), \ldots, \boldsymbol{e}^{(P,k)}(x_P)) \in \mathbb{R}^{2P \times Pk}$ has rank $2P$ almost everywhere on $\mathbb{R}^P$.*

Proof: As the non-zero entries are at indices $(2j, 2j+1)$, and there are k columns in the matrix $\boldsymbol{E}$ for each $j = 1, \ldots, P$, the matrix $\boldsymbol{E}$ has at least the rank of P. By using Lemma 4, it can be deduced that for each $j = 1, \ldots, P$, the submatrix $\boldsymbol{E}_j = (\boldsymbol{e}^{(j,1)}(x_j), \ldots, \boldsymbol{e}^{(j,k)}(x_j)$ has rank greater or equal to 2 almost everywhere on $\mathbb{R}$. Thus, it can be concluded that the rank of $\boldsymbol{E}$ is $2P$ almost everywhere on $\mathbb{R}^P$. □

B Proofs

In this section, the proofs are provided for the main identifiability theorems and for all propositions. The proofs of Theorems 1 and 2 closely follow the approach of [1], where the identifiability was proved for the exponential family without autoregressive structure.

B.1 Proof of Proposition 1

Since $(\boldsymbol{f}, \boldsymbol{T}, \boldsymbol{\lambda})$ is identifiable up to block-affine transformation and $R = 0$, we have $\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{z}}) = \boldsymbol{A}\boldsymbol{T}(\boldsymbol{z}) + \boldsymbol{c}$, where $\boldsymbol{A}$ is a block-permutation matrix and $\boldsymbol{c}$ is a constant vector.

Let π be the permutation of $\{1, \dots, P\}$ induced by the block structure of $\boldsymbol{A}$. For each i, the ith block equation of the above is: $\tilde{\boldsymbol{T}}_i(\tilde{z}_i) = \boldsymbol{A}_{i,\pi(i)}\boldsymbol{T}_{\pi(i)}(z_{\pi(i)}) + \boldsymbol{c}_i$ where $\boldsymbol{A}_{i,\pi(i)}$ is the $k \times k$ submatrix of $\boldsymbol{A}$ corresponding to the transformation from the $\pi(i)$th to the ith component, and $\boldsymbol{c}_i \in \mathbb{R}^k$ is the corresponding subvector of $\boldsymbol{c}$.

By applying $\tilde{g}_i$ to both sides for each i and using assumption (ii), we have $a_i\tilde{z}_i = \tilde{g}_i(\boldsymbol{A}_{i,\pi(i)}\boldsymbol{T}_{\pi(i)}(z_{\pi(i)}) + \boldsymbol{c}_i)$. Let $g_{\pi(i)}(z_{\pi(i)}) = \frac{1}{a_i}\tilde{g}_i(\boldsymbol{A}_{i,\pi(i)}\boldsymbol{T}_{\pi(i)}(z_{\pi(i)}) + \boldsymbol{c}_i)$. Then, we have $\tilde{z}_i = g_{\pi(i)}(z_{\pi(i)})$.

The permutation π defines a permutation matrix $\boldsymbol{P}$, giving us
$\tilde{\boldsymbol{z}} = \boldsymbol{P}(g_1(z_1), \dots, g_P(z_P))^\top$. □

B.2 Proof of Theorem 1

Step 1. Let us denote $\boldsymbol{x}^- = \{\boldsymbol{x}^{t-1:t-R}\}$, $\boldsymbol{x} = \boldsymbol{x}^t$ and $\boldsymbol{z} = \boldsymbol{z}^t$. Suppose there are two sets of parameters $\boldsymbol{\theta} = (\boldsymbol{f}, \boldsymbol{T}, \boldsymbol{\lambda})$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{f}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{\lambda}})$ such that $p_{\boldsymbol{f},\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{x}|\boldsymbol{x}^-, \boldsymbol{u}) = p_{\tilde{\boldsymbol{f}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}}}(\boldsymbol{x}|\boldsymbol{x}^-, \boldsymbol{u})$ for all $(\boldsymbol{x}|\boldsymbol{x}^-, \boldsymbol{u})$. Then

$$\int_{\mathcal{Z}} p_{\boldsymbol{f},\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{x}^-, \boldsymbol{u})d\boldsymbol{z} = \int_{\mathcal{Z}} p_{\tilde{\boldsymbol{f}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}}}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{x}^-, \boldsymbol{u})d\boldsymbol{z} \tag{23}$$

$$\Longrightarrow \quad \int_{\mathcal{Z}} p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}^-, \boldsymbol{u})p_{\boldsymbol{f}}(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z} = \int_{\mathcal{Z}} p_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}}}(\boldsymbol{z}|\boldsymbol{x}^-, \boldsymbol{u})p_{\tilde{\boldsymbol{f}}}(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z} \tag{24}$$

$$\overset{(i)}{\Longrightarrow} \quad \int_{\mathcal{Z}} p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}^-, \boldsymbol{u})p_{\boldsymbol{\epsilon}}(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{z}))d\boldsymbol{z} = \int_{\mathcal{Z}} p_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}}}(\boldsymbol{z}|\boldsymbol{x}^-, \boldsymbol{u})p_{\boldsymbol{\epsilon}}(\boldsymbol{x} - \tilde{\boldsymbol{f}}(z))d\boldsymbol{z} \tag{25}$$

$$\Longrightarrow \int_{\mathcal{X}} p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{q}(\bar{\boldsymbol{x}})|\boldsymbol{x}^-, \boldsymbol{u})p_{\boldsymbol{\epsilon}}(\boldsymbol{x} - \bar{\boldsymbol{x}})|\det(J_{\boldsymbol{q}}(\bar{\boldsymbol{x}}))|d\bar{\boldsymbol{x}} = \int_{\mathcal{X}} p_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}}}(\tilde{\boldsymbol{q}}(\boldsymbol{x})|\boldsymbol{x}^-, \boldsymbol{u})p_{\boldsymbol{\epsilon}}(\boldsymbol{x} - \bar{\boldsymbol{x}}')|\det(J_{\tilde{\boldsymbol{q}}}(\bar{\boldsymbol{x}}'))|d\bar{\boldsymbol{x}} \tag{26}$$

$$\Longrightarrow \quad \int_{\mathbb{R}^S} \tilde{p}_{\boldsymbol{T},\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{u},\boldsymbol{x}^-}(\boldsymbol{q}(\bar{\boldsymbol{x}}))p_{\boldsymbol{\epsilon}}(\boldsymbol{x} - \bar{\boldsymbol{x}})d\bar{\boldsymbol{x}} = \int_{\mathbb{R}^S} \tilde{p}_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{f}},\boldsymbol{u},\boldsymbol{x}^-}(\tilde{\boldsymbol{q}}(\boldsymbol{x}))p_{\boldsymbol{\epsilon}}(\boldsymbol{x} - \bar{\boldsymbol{x}}')d\bar{\boldsymbol{x}} \tag{27}$$

$$\Longrightarrow \quad (\tilde{p}_{\boldsymbol{T},\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{u},\boldsymbol{x}^-} * p_{\boldsymbol{\epsilon}})(\bar{\boldsymbol{x}}) = (\tilde{p}_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{f}},\boldsymbol{u},\boldsymbol{x}^-} * p_{\boldsymbol{\epsilon}})(\bar{\boldsymbol{x}}') \tag{28}$$

$$\Longrightarrow \quad F[\tilde{p}_{\boldsymbol{T},\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{u},\boldsymbol{x}^-}](\omega)\varphi_{\boldsymbol{\epsilon}}(\omega) = F[\tilde{p}_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{f}},\boldsymbol{u},\boldsymbol{x}^-}](\omega)\varphi_{\boldsymbol{\epsilon}}(\omega) \tag{29}$$

$$\overset{(i)}{\Longrightarrow} \quad F[\tilde{p}_{\boldsymbol{T},\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{u},\boldsymbol{x}^-}](\omega) = F[\tilde{p}_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{f}},\boldsymbol{u},\boldsymbol{x}^-}](\omega) \tag{30}$$

$$\Longrightarrow \quad \tilde{p}_{\boldsymbol{T},\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{u},\boldsymbol{x}^-}(\boldsymbol{x}) = \tilde{p}_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{f}},\boldsymbol{u},\boldsymbol{x}^-}(\boldsymbol{x}) \tag{31}$$

- In equation (26), J denotes Jacobian, a variable change $\bar{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{z})$ is introduced left hand side and $\bar{\boldsymbol{x}}' = \tilde{\boldsymbol{f}}(\boldsymbol{z})$ to right hand side.
- In equation (27), $\tilde{p}_{\boldsymbol{T},\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{u},\boldsymbol{x}^-} = p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{q}(\bar{\boldsymbol{x}}^t)|\boldsymbol{x}^-, \boldsymbol{u})|\det(J_{\boldsymbol{q}}(\bar{\boldsymbol{x}}))|\mathbb{1}_{\mathcal{X}}(\boldsymbol{x})$ is introduced left hand side and similarly to right hand side. The indicator function $\mathbb{1}_{\mathcal{X}}(\boldsymbol{x})$ is defined as $\mathbb{1}_{\mathcal{X}}(\boldsymbol{x}) = \begin{cases} 1, & \text{when } \boldsymbol{x} \in \mathcal{X}, \\ 0, & \text{otherwise.} \end{cases}$
- In equation (28), $*$ denotes a convolution operator.
- In equation (29), F denotes Fourier transform, and $\varphi_{\boldsymbol{\epsilon}} = F[p_{\boldsymbol{\epsilon}}]$.
- In equation (30), $\varphi_{\boldsymbol{\epsilon}}$ is dropped from both sides because of assumption (i) ($\varphi_{\boldsymbol{\epsilon}}$ is non-zero almost everywhere).

The step 1 guarantees that if the distributions with noise $\boldsymbol{\epsilon}$ are the same, then the noise-free distributions have to be the same.

Step 2. By starting from equation (31) and replacing the conditioning variable $\boldsymbol{x}^-$ with $\boldsymbol{q}(\boldsymbol{x}^-) = \{\boldsymbol{q}_x^{t-1}, \dots, \boldsymbol{q}_x^{t-R}\}$ (this can be done because $\boldsymbol{f}(\boldsymbol{q}(\boldsymbol{x})) = \boldsymbol{x}$, meaning that $\boldsymbol{q}(\boldsymbol{x})$ contains the same information as $\boldsymbol{x}$), denoting that the transformation $\boldsymbol{q}$ is applied to all $\boldsymbol{x}^i$, $i = t-1, \dots, t-R$, we get the following form:

$$\tilde{p}_{\boldsymbol{T},\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{u},\boldsymbol{x}^-}(\boldsymbol{x}) = \tilde{p}_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{f}},\boldsymbol{u},\boldsymbol{x}^-}(\boldsymbol{x}) \tag{32}$$

$$\implies p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{q}(\boldsymbol{x})|\boldsymbol{q}(\boldsymbol{x}^-),\boldsymbol{u})|\det(J_{\boldsymbol{q}}(\boldsymbol{x}))|\mathbb{1}_{\mathcal{X}}(\boldsymbol{x}) = p_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}}}(\tilde{\boldsymbol{q}}(\boldsymbol{x})|\boldsymbol{q}(\boldsymbol{x}^-),\boldsymbol{u})|\det(J_{\tilde{\boldsymbol{q}}}(\boldsymbol{x}))|\mathbb{1}_{\mathcal{X}}(\boldsymbol{x}) \tag{33}$$

By taking a logarithm on both sides of equation (33) and replacing $p_{\boldsymbol{T},\boldsymbol{\lambda}}$ and $p_{\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\lambda}}}$ with the form in equation (5), we get:

$$\begin{aligned}\log|\det(J_{\boldsymbol{q}}(\boldsymbol{x}))| + \sum_{i=1}^{P}(\log Q_i(q_i(\boldsymbol{x}), q_i(\boldsymbol{x}^-)) - \log Z_i(\boldsymbol{u}) + \sum_{j=1}^{k} T_{i,j}(q_i(\boldsymbol{x}), q_i(\boldsymbol{x}^-))\lambda_{i,j}(\boldsymbol{u})) = \\ \log|\det(J_{\tilde{\boldsymbol{q}}}(\boldsymbol{x}))| + \sum_{i=1}^{P}(\log\tilde{Q}_i(\tilde{q}_i(\boldsymbol{x}), \tilde{q}_i(\boldsymbol{x}^-)) - \log\tilde{Z}_i(\boldsymbol{u}) + \sum_{j=1}^{k} \tilde{T}_{i,j}(\tilde{q}_i(\boldsymbol{x}), \tilde{q}_i(\boldsymbol{x}^-))\tilde{\lambda}_{i,j}(\boldsymbol{u}))\end{aligned} \tag{34}$$

Let $\boldsymbol{u}_0, \dots, \boldsymbol{u}_{Pk}$ be the distinct points in assumption (iv). Then, we have $Pk+1$ equations as in (34), one for each point. By subtracting the first equation from the others, for point $\boldsymbol{u}_l$, $l = 1, \dots, Pk$, we have

$$\begin{aligned}\sum_{i=1}^{P}\log\frac{Z_i(\boldsymbol{u}_0)}{Z_i(\boldsymbol{u}_l)} + \sum_{i=1}^{P}\sum_{j=1}^{k}(T_{i,j}(q_i(\boldsymbol{x}), q_i(\boldsymbol{x}^-))(\lambda_{i,j}(\boldsymbol{u}_l) - \lambda_{i,j}(\boldsymbol{u}_0)) = \\ \sum_{i=1}^{P}\log\frac{\tilde{Z}_i(\boldsymbol{u}_0)}{\tilde{Z}_i(\boldsymbol{u}_l)} + \sum_{i=1}^{P}\sum_{j=1}^{k}(\tilde{T}_{i,j}(\tilde{q}_i(\boldsymbol{x}), \tilde{q}_i(\boldsymbol{x}^-))(\tilde{\lambda}_{i,j}(\boldsymbol{u}_l) - \tilde{\lambda}_{i,j}(\boldsymbol{u}_0))\end{aligned} \tag{35}$$

Let us define $\bar{\boldsymbol{\lambda}}(\boldsymbol{u}) = \boldsymbol{\lambda}(\boldsymbol{u}) - \boldsymbol{\lambda}(\boldsymbol{u}_0)$, and subtract $\sum_{i=1}^{P}\log\frac{\tilde{Z}_i(\boldsymbol{u}_0)}{\tilde{Z}_i(\boldsymbol{u}_l)}$ both sides. Then we have

$$\sum_{i=1}^{P}\sum_{j=1}^{k}(T_{i,j}(q_i(\boldsymbol{x}), q_i(\boldsymbol{x}^-))(\bar{\lambda}_{i,j}(\boldsymbol{u}_l)) = \sum_{i=1}^{P}\log\frac{Z_i(\boldsymbol{u}_0)\tilde{Z}_i(\boldsymbol{u}_0)}{Z_i(\boldsymbol{u}_l)\tilde{Z}_i(\boldsymbol{u}_l)} + \sum_{i=1}^{P}\sum_{j=1}^{k}(\tilde{T}_{i,j}(\tilde{q}_i(\boldsymbol{x}), \tilde{q}_i(\boldsymbol{x}^-))(\bar{\tilde{\lambda}}_{i,j}(\boldsymbol{u}_l)) \tag{36}$$

Let us write $b_l = \sum_{i=1}^{P}\log\frac{Z_i(\boldsymbol{u}_0)\tilde{Z}_i(\boldsymbol{u}_0)}{Z_i(\boldsymbol{u}_l)\tilde{Z}_i(\boldsymbol{u}_l)}$ and set $\boldsymbol{b} = (b_1, \dots, b_{Pk})$. Let $\boldsymbol{L}$ be the matrix in assumption (iv), and $\tilde{\boldsymbol{L}}$ similar matrix for $\tilde{\boldsymbol{\lambda}}$. By expressing (36) in matrix form for all point b_l, $l = 1, \dots, Pk$, we have:

$$\boldsymbol{L}^\top\boldsymbol{T}(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{q}(\boldsymbol{x}^-)) = \tilde{\boldsymbol{L}}^\top\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{q}}(\boldsymbol{x}), \boldsymbol{q}(\boldsymbol{x}^-)) + \boldsymbol{b} \tag{37}$$

$$\implies \boldsymbol{T}(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{q}(\boldsymbol{x}^-)) = (\boldsymbol{L}^\top)^{-1}\tilde{\boldsymbol{L}}^\top\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{q}}(\boldsymbol{x}), \boldsymbol{q}(\boldsymbol{x}^-)) + (\boldsymbol{L}^\top)^{-1}\boldsymbol{b} \tag{38}$$

$$\implies \boldsymbol{T}(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{q}(\boldsymbol{x}^-)) = \boldsymbol{A}\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{q}}(\boldsymbol{x}), \boldsymbol{q}(\boldsymbol{x}^-)) + \boldsymbol{c}, \tag{39}$$

where $\boldsymbol{A} = (\boldsymbol{L}^\top)^{-1}\tilde{\boldsymbol{L}}^\top$ and $\boldsymbol{c} = (\boldsymbol{L}^\top)^{-1}\boldsymbol{b}$.

Step 3. By assumption (iii), Jacobian of $\boldsymbol{T}$ exists and is a $Pk \times P$ matrix of rank P. Because equation (39) holds, it also holds that $J(\boldsymbol{T}(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{q}(\boldsymbol{x}^-)) = \boldsymbol{A}J(\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{q}}(\boldsymbol{x}), \tilde{\boldsymbol{q}}(\boldsymbol{x}^-)))$ and that $\text{rank}\Big(J(\boldsymbol{T}(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{q}(\boldsymbol{x}^-))\Big) = \text{rank}\Big(\boldsymbol{A}J(\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{q}}(\boldsymbol{x}), \tilde{\boldsymbol{q}}(\boldsymbol{x}^-)))\Big)$. This leads to the fact that both $\boldsymbol{A}$ and $J(\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{q}}(\boldsymbol{x}), \tilde{\boldsymbol{q}}(\boldsymbol{x}^-)))$ are of rank P.

- If $k = 1$, then A is invertible since it is a $P \times P$ matrix of rank P.
- If $k \geq 2$, define $\bar{\boldsymbol{z}} = \boldsymbol{q}(\boldsymbol{x})$, $\bar{\boldsymbol{z}}^- = \boldsymbol{q}(\boldsymbol{x}^-)$ and $\boldsymbol{T}_i = (T_{i,1}(\bar{z}_i, \bar{z}_i^-), \dots, T_{i,k}(\bar{z}_i, \bar{z}_i^-))$. Based on Lemma 3, it holds that for each $i = 1, \dots, P$, there exists k points $(\bar{z}_i^j, \bar{z}_i^{-,j})$, $j = 1, \dots, k$ such that $(\boldsymbol{T}_i'(\bar{z}_i^1, \bar{z}_i^{-,1}), \dots, \boldsymbol{T}_i'(\bar{z}_i^k, \bar{z}_i^{-,k}))$ are linearly independent. Let us define $\boldsymbol{Q} = (J(\boldsymbol{T}(\bar{\boldsymbol{z}}^1, \bar{\boldsymbol{z}}^{-,1})), \dots, J(\boldsymbol{T}(\bar{\boldsymbol{z}}^k, \bar{\boldsymbol{z}}^{-,k})))$, where each Jacobian is $Pk \times P$ matrix calculated with respect to $\bar{\boldsymbol{z}}^i$, and the vector $\bar{\boldsymbol{z}}^l$ and $\bar{\boldsymbol{z}}^{-,l}$ are defined as $\bar{\boldsymbol{z}}^l = (\bar{z}_1^l, \dots, \bar{z}_P^l)$ and $\bar{\boldsymbol{z}}^{-,l} = (\bar{z}_1^{-,l}, \dots, \bar{z}_P^{-,l})$. Similarly define matrix $\tilde{\boldsymbol{Q}}$ for Jacobians of $\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{q}}(\boldsymbol{f}(\bar{\boldsymbol{x}}^l)), \tilde{\boldsymbol{q}}(\boldsymbol{f}(\bar{\boldsymbol{x}}^{-,l}))$ for the same points $l = 1, \dots, k$. Then, by differentiating the equation (39) for each $\boldsymbol{x}_l$, we get the following in matrix form:

$$\boldsymbol{Q} = \boldsymbol{A}\tilde{\boldsymbol{Q}}. \tag{40}$$

The matrix $\boldsymbol{Q}$ is invertible based on Lemma 3, and hence also $\boldsymbol{A}$ and $\tilde{\boldsymbol{Q}}$ are invertible. As we have invertible $\boldsymbol{A}$, the equation (39) says that the sufficient statistics are identifiable up to linear transformation and a constant. □

B.3 Proof of Theorem 2

Step 1. The assumptions of theorem 1 holds, so we have

$$\boldsymbol{T}(\boldsymbol{q}(\boldsymbol{x}), \boldsymbol{q}(\boldsymbol{x}^-)) = \boldsymbol{A}\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{q}}(\boldsymbol{x}), \tilde{\boldsymbol{q}}(\boldsymbol{x}^-)) + \boldsymbol{c}, \tag{41}$$

where $\boldsymbol{c}$ is a constant vector and $\boldsymbol{A}$ is an invertible $Pk \times Pk$ matrix. Let (i, l, a, b) be four indices so that $1 \leq i \leq P$, $1 \leq l \leq k$ refer to the rows of the matrix $\boldsymbol{A}$, and $1 \leq a \leq P$, $1 \leq b \leq k$ refer to the columns of $\boldsymbol{A}$. Let $\boldsymbol{v}(\boldsymbol{z}) = \tilde{\boldsymbol{q}}(\boldsymbol{f}(\boldsymbol{z})) : \mathcal{Z} \to \mathcal{Z}$. The function $\boldsymbol{v}$ is bijective as $\tilde{\boldsymbol{f}} : \mathcal{Z} \to \mathcal{X}$ and $\boldsymbol{f} : \mathcal{Z} \to \mathcal{X}$ are injective functions, and $\tilde{\boldsymbol{q}}(\tilde{\boldsymbol{f}}(\boldsymbol{z})) = \boldsymbol{z}$. Further, let there be two other indices $c, d \in \{1, \ldots, P\}$, $c < d$ and denote $v_i^c = \frac{\partial v_i}{\partial v_c}$ and $v_i^{c,d} = \frac{\partial v_i}{\partial v_c \partial v_d}$. By differentiating (41) with respect to z_c, we get for each $1 \leq i \leq P$ and $1 \leq l \leq k$ the following:

$$\frac{\partial T_{i,l}(z_i, z_i^-)}{\partial z_c} = \sum_{a,b} A_{i,l,a,b} \left(\frac{\partial \tilde{T}_{a,b}(v_a(\boldsymbol{z}), v_a(\boldsymbol{z}^-))}{\partial v_a(\boldsymbol{z})} \frac{\partial v_a(\boldsymbol{z})}{\partial z_c} + \sum_{r=1}^{R} \frac{\partial \tilde{T}_{a,b}(v_a(\boldsymbol{z}), v_a(\boldsymbol{z}^-))}{\partial v_a(\boldsymbol{z}^{-r})} \frac{\partial v_a(\boldsymbol{z}^{-r})}{\partial z_c} \right). \tag{42}$$

It holds that $\frac{\partial v_a(\boldsymbol{z}^{-r})}{\partial z_c} = 0$ for all $r = 1, \ldots, R$, as the values of previous time points do not depend on the value of current time point. Thus, we have

$$\frac{\partial T_{i,l}(z_i, z_i^-)}{\partial z_c} = \sum_{a,b} A_{i,l,a,b} \left(\frac{\tilde{T}_{a,b}}{\partial v_a(\boldsymbol{z})} \frac{\partial v_a(\boldsymbol{z})}{\partial z_c} \right). \tag{43}$$

By differentiating (43) with respect to z_d, we get

$$0 = \sum_{a,b} A_{i,l,a,b} \left(\frac{\partial \tilde{T}_{a,b}(v_a(\boldsymbol{z}), v_a(\boldsymbol{z}^-))}{\partial z_d} \frac{\partial v_a(\boldsymbol{z})}{\partial z_c \partial z_d} + \frac{\partial \tilde{T}_{a,b}(v_a(\boldsymbol{z}))}{\partial^2 v_a(\boldsymbol{z})} \frac{\partial v_a(\boldsymbol{z})}{\partial z_c} \frac{\partial v_a(\boldsymbol{z})}{\partial z_d} \right). \tag{44}$$

Let us define $\boldsymbol{r}_a^1(\boldsymbol{z}) = (v_a^{1,2}(\boldsymbol{z}), \ldots, v_a^{P-1,P}) \in \mathbb{R}^{\frac{P(P-1)}{2}}$, $\boldsymbol{r}_a^2(\boldsymbol{z}) = (v_a^1(\boldsymbol{z})v_a^2(\boldsymbol{z}), \ldots, v_a^{P-1}(\boldsymbol{z})v_a^P(\boldsymbol{z})) \in \mathbb{R}^{\frac{P(P-1)}{2}}$, $\boldsymbol{M}(\boldsymbol{z}) = (\boldsymbol{r}_1^1(\boldsymbol{z}), \boldsymbol{r}_1^2(\boldsymbol{z}), \ldots, \boldsymbol{r}_P^1(\boldsymbol{z}), \boldsymbol{r}_P^2(\boldsymbol{z})) \in \mathbb{R}^{\frac{P(P-1)}{2} \times \frac{P(P-1)}{2}}$ and $\boldsymbol{e}^{(a,b)}(z_i) = (0, \ldots, 0, T'_{a,b}(z_i), T''_{a,b}(z_i), 0, \ldots, 0) \in \mathbb{R}^{2P}$, so that the non-zero entries are at indices $(2a, 2a+1)$ and $\boldsymbol{E}(\boldsymbol{z}) = (\boldsymbol{e}^{(1,1)}(z_1), \ldots, \boldsymbol{e}^{(1,k)}(z_1), \ldots, \boldsymbol{e}^{(P,1)}(z_P), \ldots, \boldsymbol{e}^{(P,k)}(z_P)) \in \mathbb{R}^{2P \times Pk}$. Finally, let $A_{i,l}$ be the (i,l)th row of the matrix $\boldsymbol{A}$. Then, by gathering the equation (43) for all pairs (c, d), $c < d$ and pairs (i, l) to a matrix form, we get

$$\boldsymbol{M}(\boldsymbol{z})\boldsymbol{E}(\boldsymbol{z})\boldsymbol{A} = \boldsymbol{0}. \tag{45}$$

By Lemma 5, the matrix $\boldsymbol{E}$ is of rank $2P$ almost surely on $\mathcal{Z}$. Since the matrix $\boldsymbol{A}$ is full rank $Pk \times Pk$ matrix, we have rank$(\boldsymbol{EA}) = 2P$ almost surely on $\mathcal{Z}$. Hence, by multiplying (45) from right with the pseudo-inverse of $(\boldsymbol{EA})$ we have

$$\boldsymbol{M}(\boldsymbol{z}) = \boldsymbol{0}. \tag{46}$$

Particularly, $\boldsymbol{r}_a^2 = \boldsymbol{0}$ for all $a = 1, \ldots, P$. This means that at each $\boldsymbol{z} \in \mathcal{Z}$, the Jacobian of $\boldsymbol{v}$, $J_{\boldsymbol{v}}$ has at most one non-zero entry in each row. Because $J_{\boldsymbol{v}}$ is invertible and continuous, the locations of the non-zero entries are fixed and do not change as function of $\boldsymbol{z}$. This proves that the function $\tilde{\boldsymbol{q}}(\boldsymbol{f}(\boldsymbol{z}))$ is a composition of a permutation and a point-wise nonlinearity.

Step 2. Without loss of generality, we assume that the permutation in $\boldsymbol{v}$ is identity. Let $\bar{\boldsymbol{T}}(\boldsymbol{z}) = \tilde{\boldsymbol{T}}(\boldsymbol{v}(\boldsymbol{z})) + \boldsymbol{A}^{-1}\boldsymbol{c}$. In particular, $\bar{\boldsymbol{T}}$ is then a point-wise nonlinearity. Then, the equation (41) can be written as

$$\boldsymbol{T}(\boldsymbol{z}, \boldsymbol{z}^-) = \boldsymbol{A}\bar{\boldsymbol{T}}(\boldsymbol{z}, \boldsymbol{z}^-). \tag{47}$$

Let $\boldsymbol{W} = \boldsymbol{A}^{-1}$. Then, the equation (47) can be written for each component $1 \leq i \leq P$ and sufficient statistic $1 \leq l \leq k$ as

$$\bar{T}_{i,l} = \sum_{a,b} D_{i,l,a,b} T_{a,b}(z_a, z_a^-). \tag{48}$$

By differentiating both sides with respect to z_c, $c \neq i$, we get

$$0 = \sum_{b} D_{i,l,c,b} \frac{\partial T'_{c,b}(z_a, z_a^-)}{\partial z_c}. \tag{49}$$

By Lemma 1, we know that $D_{i,l,c,b} = 0$ for all $1 \leq b \leq k$, and since (49) holds for all l and $c \neq i$, the matrix $\boldsymbol{D}$ must have a block diagonal form

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 & & \\ & \ddots & \\ & & \boldsymbol{D}_P \end{pmatrix}, \tag{50}$$

where each submatrix $D_1, \ldots, D_P$ is a $k \times k$ matrix. Then, also the matrix $\boldsymbol{A}$ has the same block diagonal form, meaning that each submatrix $\boldsymbol{A}_i$ transforms $\boldsymbol{T}_i(\boldsymbol{z}, \boldsymbol{z}^-)$ into $\bar{\boldsymbol{T}}_i(\boldsymbol{z}, \boldsymbol{z}^-)$. Since $\bar{\boldsymbol{T}}$ is a point-wise nonlinearity, $\boldsymbol{A}$ has to be a permutation matrix. □

B.4 Proof of Proposition 2

Based on the assumptions we have the following equalities

$$\tilde{z}_j = a_{11} z_i + a_{12} z_i^2 + c_1,$$
$$\tilde{z}_j^2 = a_{21} z_i + a_{22} z_i^2 + c_2,$$

for some constants $a_{11}, a_{12}, a_{21}, a_{22}, c_1$ and c_2. By squaring the first equation, we have $(a_{11} z_i + a_{12} z_i^2 + c_1)^2 = a_{21} z_i + a_{22} z_i^2 + c_2$. In order the equation to hold for all $z_i \in \mathcal{Z}$, it must hold that $a_{12} = 0$. Hence, we have that $\tilde{z}_j = a_{11} z_i + c_1$. □

B.5 Proof of Proposition 3

Since $(\boldsymbol{f}, \boldsymbol{T}, \boldsymbol{\lambda})$ is identifiable up to block-affine transformation, we have $\tilde{\boldsymbol{T}}(\tilde{\boldsymbol{z}}) = \boldsymbol{A}\boldsymbol{T}(\boldsymbol{z}) + \boldsymbol{c}$, where $\boldsymbol{A}$ is a block-permutation matrix and $\boldsymbol{c}$ is a constant vector.

Let π be the permutation of $\{1, \ldots, P\}$ induced by the block structure of $\boldsymbol{A}$, and $j = \pi_i$. Then we have that $\tilde{\boldsymbol{T}}_j(\tilde{z}_j^t, \ldots, \tilde{z}_j^{t-R}) = \boldsymbol{A}_{i,j} \boldsymbol{T}_i(z_j^t, \ldots, z_j^{t-R})$, where $\boldsymbol{A}_{i,j}$ is a $k \times k$ submatrix of $\boldsymbol{A}$ corresponding the indices i and j. Because of Gaussian AR form (1), we have

$$p(\boldsymbol{z} | \{\boldsymbol{z}^{t-1:t-R}\}, \boldsymbol{u}^t, \ldots, \boldsymbol{u}^{t-R}) =$$
$$\prod_{i=1}^{P} \frac{1}{2\pi\sigma_i(\boldsymbol{u}^t)} \exp\left[\frac{\left(z_i - \mu_i(\boldsymbol{u}^t) - \sum_{r=1}^{R} (\gamma_r(\boldsymbol{u}^t) z_i^{t-r} - \mu_i(\boldsymbol{u}^{t-r})) \right)^2}{2\sigma^2(\boldsymbol{u}^t)} \right]. \tag{51}$$

and similar form for $\tilde{z}_j$ with parameter functions $\tilde{\mu}_j, \tilde{\sigma}_j, \tilde{\gamma}_{j,1}, \ldots, \tilde{\gamma}_{j,R}$. Let $\gamma_{i,r} \coloneqq \gamma_{i,r}(\boldsymbol{u}^t)$, $\mu_{i,r} \coloneqq \mu_i(\boldsymbol{u}^t - r)$ and $\sigma_i \coloneqq \sigma_i(\boldsymbol{u}^t)$. By expanding the nominator in the exponential term, we have

$$(z_i^t)^2 - 2z_i^t \mu_{i,0} - 2z_i^t \sum_{r=1}^{R} \gamma_{i,r} z_i^{t-r} + 2z_i^t \sum_{r=1}^{R} \gamma_{i,r}\mu_{i,r} + \mu_{i,0}^2 + 2\mu_{i,0} \sum_{r=1}^{R} \gamma_{i,r} z_i^{t-r} +$$
$$(\sum_{r=1}^{R} \gamma_{i,r} z_i^{t-r})^2 - 2(\sum_{r=1}^{R} \gamma_{i,r} z_i^{t-r})(\sum_{r=1}^{R} \gamma_{i,r}\mu_{i,r}) + (\sum_{r=1}^{R} \gamma_{i,r}\mu_{i,r})^2. \tag{52}$$

From this form, it is easy to see that the minimal sufficient statistics are $T_{i,1} = (z_i^t)^2$, $T_{i,2} = z_i^t$, $T_{i,3,r} = z_i^t z_i^{t-r}$, $T_{i,4,r} = z_i^{t-r}$, $T_{i,5,r_1,r_2} = z_i^{t-r_1} z_i^{t-r_2}$, $r, r_1, r_2 \in \{1, \dots, R\}$. Similarly, we have the sufficient statistics $\tilde{\boldsymbol{T}}_j(\tilde{z}_j^t, \dots, \tilde{z}_j^{t-R})$. Because of the block-affine identifiability, we have for each $k_1 \in \{1, \dots, k\}$ that

$$\tilde{T}_{k_1,j} = \sum_{k_2=1}^{k} a_{k_2,k_1,i} T_{k_2,i} + c_{i,k_1}, \tag{53}$$

where $a_{k_1,k_2,i}$ and c_{i,k_1} are constants. Importantly, we have for all $r = 0, \dots, R$ that $\tilde{z}_j^{t-r} = \sum_{k_2=1}^{k} a_{k_2,r_1,i} T_{k_2,i} + c_i$ and $(\tilde{z}_j^{t-r})^2 = \sum_{k_2=1}^{k} a_{k_2,r_2,i} T_{k_2,i} + c_i$. By squaring the first equation, we have that

$$(\sum_{k_2=1}^{k} a_{k_2,r_1,i} T_{k_2,i} + c_{i,r_1})^2 = \sum_{k_2=1}^{k} a_{k_2,r_2,i} T_{k_2,i} + c_{i,r_2}. \tag{54}$$

This equation holds only if the coefficients of the third order and above in the left hand side are zero, meaning that $a_{1,r_1,i}, a_{(3,r),r_1,i}, a_{(5,r),r_1,i} = 0$. Hence, we have for all $r_1 = 0, \dots, R$ and $t = R+1, \dots, T$ that

$$\tilde{z}_j^{t-r_1} = \sum_{r_2=0}^{R} b_{r_1,r_2,i} z_i^{t-r_2} + c_{r_1,i}, \tag{55}$$

where $b_{r_1,r_2,i}$ are constants. Since (55) holds for all $t = R+1, \dots, T$, we also have the following equations:

$$\begin{aligned} \tilde{z}_j^t &= \sum_{r=0}^{R} b_{0,r,i} z_i^{t-r} + c_{0,i}, \\ \tilde{z}_j^t &= \sum_{r=0}^{R} b_{R,r,i} z_i^{t+R-r} + c_{R,i}, \end{aligned} \tag{56}$$

where the second equation is obtained by shifting (55), for $r_1 = R$, R time steps forward. From (56) we can deduce that all coefficients $b_{0,r,i}$, $r \neq 0$, have to be zero in order for the equations to hold for all $t \in \{R+1, T\}$. Hence, we obtain $\tilde{z}_j^t = b_{0,0,i} z_i^t + c_{0,i}$, which concludes the proof. □

B.6 Proof of Theorem 3

The lower bound of the data log likelihood (ELBO) (9) can also be written in the following format:

$$\mathrm{ELBO} = E_{q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{u})}\big(\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{x}^-, \boldsymbol{u}) + \mathrm{KL}(\log q_{\boldsymbol{\theta_g}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{x}^-, \boldsymbol{u}) || p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{x}^-, \boldsymbol{u}))\big), \tag{57}$$

where KL is Kullback-Leibler divergence and the set $(\tilde{\boldsymbol{f}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{\lambda}})$ are parametrized by $\boldsymbol{\theta}$. Minimizing ELBO given in (9) with respect to the parameters $(\boldsymbol{\theta}, \boldsymbol{\theta_g})$ is equivalent to minimizing (57), which means that in the limit of infinite data, the KL term eventually reaches zero, making the loss equal to the data log likelihood. Hence in this case, minimizing ELBO is equivalent to maximum likelihood estimation (MLE). As we assume that Theorem 1 or Theorem 2 hold, the consistency of MLE guarantees that the estimation converges to the corresponding identifiability class of the true set $(\boldsymbol{f}, \boldsymbol{T}, \boldsymbol{\lambda})$ in the limit of infinite data. □

C Additional simulation details

The parameters used in all simulation settings of Section 4.1, are provided in Table 1.

References

1. Khemakhem, I., Kingma, D., Monti, R., Hyvärinen, A.: Variational autoencoders and nonlinear ICA: A unifying framework. In: International Conference on Artificial Intelligence and Statistics. pp. 2207–2217. PMLR (2020)

Table 1. The parameters for the Matern covariance function in all simulation settings.

	IC1	IC2	IC3	IC4	IC5	IC6
ϕ	0.20	0.15	0.10	0.30	0.05	0.25
ν	0.50	1.00	0.25	2.00	0.75	1.50