**Supplementary Material for**
***Nonlinear ICA of Temporally Dependent Stationary Sources***
**by Aapo Hyvärinen and Hiroshi Morioka, AISTATS2017**

## Analysis of Processes in Eq. (8) and Eq. (9)

For the nonlinear AR model in Eq. (9), quasi-Gaussianity and uniform dependence are easy to see (under the assumptions given in the main text) since we have $\frac{\partial^2 \log p_{x,y}(x,y)}{\partial x \partial y} = 2\lambda r'(y)$. This implies uniform dependence by the strict monotonicity of $r$, and non-quasi-Gaussianity by its functional form.

For the non-Gaussian AR model in Eq. (8) we proceed as follows: First, we have

$$\frac{\partial^2 \log p_{x,y}(x,y)}{\partial x \partial y} = -\rho\, G''(x - \rho y). \tag{19}$$

This is always non-zero by the assumption on $G''$, and non-zero $\rho$, so uniform dependence holds. Assume a factorization as in (4) holds:

$$-G''(x - \rho y) = c\alpha(x)\alpha(y). \tag{20}$$

By assumption, $-G''$ is always positive, so we can take logarithms on both sides of (20), and again cross-derivatives. We necessarily have $\frac{\partial \log -G''(x-\rho y)}{\partial x \partial y} = 0$, since the RHS is separable. This can be evaluated as $(\log -G'')''(x - \rho y) = 0$ which implies $\log -G''(u) = du + b$ and

$$G''(u) = -\exp(du + b) \tag{21}$$

for some real parameters $d, b$. Now, if we have $d = 0$ and thus $G''(u)$ constant, we have a Gaussian process. On the other hand, if we have $d \neq 0$, we can plug this back in (20) and see that it cannot hold because the exponents for $x$ and $y$ would be different unless $\rho = -1$, which was excluded by assumption (as is conventional to ensure stability of the process). Thus, only a Gaussian linear AR process can be quasi-Gaussian under the given assumptions.

## Proof of Theorem 1

Denote by $\mathbf{g}$ the (true) inverse function of $\mathbf{f}$ which transforms $\mathbf{x}$ into $\mathbf{s}$, i.e. $\mathbf{s}(t) = \mathbf{g}(\mathbf{x}(t))$. We can easily derive the log-pdf of an observed $(\mathbf{x}(t), \mathbf{x}(t-1))$ as

$$\log p(\mathbf{x}(t), \mathbf{x}(t-1)) = \sum_{i=1}^{n} \log p_i^{\tilde{s}}(g_i(\mathbf{x}(t)), g_i(\mathbf{x}(t-1))) + \log |\mathbf{J}\mathbf{g}(\mathbf{x}(t))| + \log |\mathbf{J}\mathbf{g}(\mathbf{x}(t-1))| \tag{22}$$

where $p_i^{\tilde{s}}$ is the pdf of $(s_i(t), s_i(t-1))$, and $\mathbf{J}\mathbf{g}$ denotes the Jacobian of $\mathbf{g}$; its log-determinant appears twice because the transformation is done twice, separately for $\mathbf{x}(t)$ and $\mathbf{x}(t-1)$.

On the other hand, according to well-known theory, when training logistic regression we will asymptotically have

$$r(\mathbf{y}) = \log p_{\mathbf{y}}(\mathbf{y}) - \log p_{\mathbf{y}^*}(\mathbf{y}) \tag{23}$$

i.e. the regression function will asymptotically give the difference of the log-probabilities in the two classes. This holds in our case in the limit of an infinitely long stochastic process due to the assumption of a stationary ergodic process (Assumption 1).

Now, based on (22), the probability in the real data class is of the form

$$\log p_{\mathbf{y}}(\mathbf{y}) = \sum_{i=1}^{n} Q_i(g_i(\mathbf{y}^1), g_i(\mathbf{y}^2)) + \log |\mathbf{J}\mathbf{g}(\mathbf{y}^1)| + \log |\mathbf{J}\mathbf{g}(\mathbf{y}^2)| \tag{24}$$

where we denote $Q_i(a, b) = \log p_i^{\tilde{s}}(a, b)$, while in the permuted (time-shuffled) data class the time points are i.i.d., which means that the log-pdf is of the form

$$\log p_{\mathbf{y}^*}(\mathbf{y}) = \sum_{i=1}^{n} \bar{Q}_i(g_i(\mathbf{y}^1)) + \bar{Q}_i(g_i(\mathbf{y}^2)) + \log |\mathbf{J}\mathbf{g}(\mathbf{y}^1)| + \log |\mathbf{J}\mathbf{g}(\mathbf{y}^2)| \tag{25}$$

for some functions $\bar{Q}_i$ which are simply the marginal log-pdf's.

The equality in (23) means the regression function (12) is asymptotically equal to the difference of (24) and (25), i.e.

$$\sum_{i=1}^{n} B_i(h_i(\mathbf{y}^1), h_i(\mathbf{y}^2)) = \sum_{i=1}^{n} Q_i(g_i(\mathbf{y}^1), g_i(\mathbf{y}^2)) - \bar{Q}_i(g_i(\mathbf{y}^1)) - \bar{Q}_i(g_i(\mathbf{y}^2)) \tag{26}$$

where we see that the Jacobian terms vanish because we "contrast" two data sets with the same Jacobian terms.

We easily notice that one solution to this is given by $h_i(\mathbf{x}) = g_i(\mathbf{x})$, $B_i(x, y) = Q_i(x, y) - \bar{Q}_i(x) - \bar{Q}_i(y)$. In fact, due to the assumption of the universal approximation capability of $B$ and $h$, such a solution can be reached by the learning process. Next we prove that this is the only solution, up to permutation of the $h_i$ and element-wise transformations.

Make the change of variables

$$\mathbf{z}^1 = \mathbf{g}(\mathbf{y}^1), \;\; \mathbf{z}^2 = \mathbf{g}(\mathbf{y}^2) \tag{27}$$

and denote the compound function

$$\mathbf{k} = \mathbf{h} \circ \mathbf{f} = \mathbf{h} \circ \mathbf{g}^{-1} \tag{28}$$

This is the compound transformation of the attempted demixing by $\mathbf{h}$ and the original mixing by $\mathbf{f}$. Such a compound function is of main interest in the theory of ICA, since it tells how well the original sources were separated. Our goal here is really to show that this function is a permutation with component-wise nonlinearities. So, we consider the transformed version of (26) given by

$$\sum_{i=1}^{n} B_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2)) = \sum_{i=1}^{n} Q_i(z_i^1, z_i^2) - \bar{Q}_i(z_i^1) - \bar{Q}_i(z_i^2) \quad (29)$$

Take cross-derivatives of both sides of (29) with respect to $z_j^1$ and $z_k^2$. This gives

$$\sum_{i=1}^{n} \frac{\partial^2 B_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2))}{\partial z_j^1 \partial z_k^2} = \sum_{i=1}^{n} \frac{\partial^2 Q_i(z_i^1, z_i^2)}{\partial z_j^1 \partial z_k^2}. \quad (30)$$

Denoting cross-derivatives as

$$b_i(a,b) := \frac{\partial^2 B_i(a,b)}{\partial a \partial b}, \quad q_i(a,b) := \frac{\partial^2 Q_i(a,b)}{\partial a \partial b} \quad (31)$$

this gives further

$$\sum_{i=1}^{n} b_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2)) \frac{\partial k_i}{\partial z_j^1}(\mathbf{z}^1) \frac{\partial k_i}{\partial z_k^2}(\mathbf{z}^2) = \sum_{i=1}^{n} q_i(z_i^1, z_i^2) \delta_{ij} \delta_{ik}$$

which must hold for all $j, k$. We can collect these equations in a matrix form as

$$\mathbf{Jk}(\mathbf{z}^1)^T \mathrm{diag}_i[b_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2))] \mathbf{Jk}(\mathbf{z}^2) = \mathrm{diag}_i[q_i(z_i^1, z_i^2)] \quad (32)$$

Now, the $q_i$ are non-zero for all $\mathbf{z}^1, \mathbf{z}^2$ by assumption of uniform dependence. Since the RHS of (32) is invertible at any point, also $\mathbf{Jk}$ must be invertible at any point. We can thus obtain

$$[\mathbf{Jk}(\mathbf{z}^1)^{-1}]^T \mathrm{diag}_i[q_i(z_i^1, z_i^2)] \mathbf{Jk}(\mathbf{z}^2)^{-1} = \mathrm{diag}_i[b_i(k_i(\mathbf{z}^1), k_i(\mathbf{z}^2))] \quad (33)$$

Next, we use the assumption of non-quasi-Gaussianity, in the form of the following Lemma (proven below):

**Lemma 2** *Assume the continuous functions $q_i(a,b)$ are non-zero everywhere, and not factorizable as in Eq. (4) in the definition of quasi-Gaussianity.*[5] *Assume $\mathbf{M}$ is any continuous matrix-valued function $\mathbb{R}^n \to \mathbb{R}^{n \times n}$, such that the matrix $\mathbf{M}(\mathbf{u})$ is non-singular for any $\mathbf{u}$. Assume we have*

$$\mathbf{M}(\mathbf{u}^1)^T \, diag_i[q_i(u_i^1, u_i^2)] \, \mathbf{M}(\mathbf{u}^2) = \mathbf{D}(\mathbf{u}^1, \mathbf{u}^2) \quad (34)$$

*for any $\mathbf{u}^1, \mathbf{u}^2$ in $\mathbb{R}^n$, and for some unknown matrix-valued function $\mathbf{D}$ which takes only diagonal values. Then, the function $\mathbf{M}(\mathbf{u})$ is such that every row and column has exactly one non-zero entry, and the locations and signs of the non-zero entries are the same for all $\mathbf{u}$.*

We apply this Lemma on Eq. (33) with $\mathbf{M}(\mathbf{z}) = \mathbf{Jk}(\mathbf{z})^{-1}$. The assumptions of the Lemma are included in the assumptions of the Theorem, except for the non-singularity of $\mathbf{M}$ which was just proven above, and the continuity of $\mathbf{M}$. If $\mathbf{Jk}(\mathbf{z})^{-1}$ were not continuous, the fact that the diagonal matrix on the LHS of (33) is continuous would imply that the diagonal matrix on the RHS is discontinuous, and this contradicts the assumptions on smoothness of $\mathbf{h}$, $\mathbf{g}$ and $B_i$.

Thus the Lemma shows that $\mathbf{Jk}(\mathbf{z})^{-1}$ must be a rescaled permutation matrix for all $\mathbf{z}$, with the same locations of the non-zero elements; the same applies to $\mathbf{Jk}(\mathbf{z})$. Thus, by (28), $\mathbf{g}$ and $\mathbf{h}$ must be equal up to a permutation and element-wise functions, plus a constant offset which can be absorbed in the element-wise functions. The fact that the signs of the elements in $\mathbf{M}$ stay the same implies the transformations are strictly monotonic, which proves the Theorem.

**Proof of Lemma 2**

Consider (34) for two different points $\bar{\mathbf{u}}^1$ and $\bar{\mathbf{u}}^2$ in $\mathbb{R}^n$. Denote for simplicity

$$\mathbf{M}_p = \mathbf{M}(\bar{\mathbf{u}}^p), \qquad \mathbf{D}_{pq} = \mathrm{diag}_i[q_i(\bar{u}_i^p, \bar{u}_i^q)] \quad (35)$$

with $p, q \in \{1, 2\}$. Evaluating (34) with all the possible combinations of setting $\mathbf{u}^1$ and $\mathbf{u}^2$ to $\bar{\mathbf{u}}^1$ and $\bar{\mathbf{u}}^2$, that is the four combinations $\mathbf{u}^1 := \bar{\mathbf{u}}^1, \mathbf{u}^2 := \bar{\mathbf{u}}^2$; $\mathbf{u}^1 := \bar{\mathbf{u}}^2, \mathbf{u}^2 := \bar{\mathbf{u}}^1$; $\mathbf{u}^1 := \bar{\mathbf{u}}^1, \mathbf{u}^2 := \bar{\mathbf{u}}^1$; and $\mathbf{u}^1 := \bar{\mathbf{u}}^2, \mathbf{u}^2 := \bar{\mathbf{u}}^2$, we have three different equations (the first one being obtained twice):

$$\mathbf{M}_1^T \mathbf{D}_{12} \mathbf{M}_2 = \mathbf{D} \quad (36)$$

$$\mathbf{M}_2^T \mathbf{D}_{22} \mathbf{M}_2 = \mathbf{D}' \quad (37)$$

$$\mathbf{M}_1^T \mathbf{D}_{11} \mathbf{M}_1 = \mathbf{D}'' \quad (38)$$

for some diagonal matrices $\mathbf{D}, \mathbf{D}', \mathbf{D}''$.

We will show that for any given $\bar{\mathbf{u}}^1$, it is always possible to find a $\bar{\mathbf{u}}^2$ such that the conditions (36–38) lead to an eigenvalue problem which has only a trivial solution consisting of a scaled permutation matrix.

By the assumption that $q_i$ is non-zero, $\mathbf{D}_{12}$ is invertible, which also implies $\mathbf{D}$ is invertible. By elementary

[5] In this lemma, the $q_i$ need not have anything to do with pdf's, so we do not directly use the assumption of quasi-Gaussianity, but the conditions on $q$ are identical.

linear algebra, we can thus solve from the first equation (36)

$$\mathbf{M}_2 = \mathbf{D}_{12}^{-1}\mathbf{M}_1^{-T}\mathbf{D} \tag{39}$$

and plugging this into the second equation (37) we have

$$\mathbf{M}_1^{-1}\mathbf{D}_{22}\mathbf{D}_{12}^{-2}\mathbf{M}_1^{-T} = \mathbf{D}^{-1}\mathbf{D}'\mathbf{D}^{-1} \tag{40}$$

Next we multiply both sides of (38) by the respective sides of (40) from the left, and denoting $\mathbf{D}''' = \mathbf{D}^{-1}\mathbf{D}'\mathbf{D}^{-1}\mathbf{D}''$ we have

$$\mathbf{M}_1^{-1}[\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}]\mathbf{M}_1 = \mathbf{D}''' \tag{41}$$

Here, we see a kind of eigenvalue decomposition.

The rest of the proof of this lemma is based on the uniqueness of the eigenvalue decomposition, which requires that the eigenvalues are distinct (i.e. no two of them are equal). So, next we show that the assumption of non-factorizability of $q_i$ implies that for any given $\bar{\mathbf{u}}^1$ we can find a $\bar{\mathbf{u}}^2$ such that the diagonal entries in $\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}$ are distinct. The diagonal entries are given by the function $\psi$ defined as

$$\psi(\bar{u}_i^1, \bar{u}_i^2) = \frac{q_i(\bar{u}_i^1, \bar{u}_i^1)q_i(\bar{u}_i^2, \bar{u}_i^2)}{q_i^2(\bar{u}_i^1, \bar{u}_i^2)}. \tag{42}$$

For simplicity of notation, drop the index $i$ and denote $a := \bar{u}_i^1, b = \bar{u}_i^2$. The diagonal entries in $\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}$ can be chosen distinct if $\psi$ is not a function of $a$ alone (which was fixed above since $\bar{\mathbf{u}}^1$ was fixed). Suppose $\psi$ is a function of $a$ alone: Then we would have

$$\frac{q(a,a)q(b,b)}{q^2(a,b)} = f(a) \tag{43}$$

for some function $f$. Since this holds for any $b$, we can set $b = a$, we see that $f$ must be identically equal to one. So, we would have

$$q^2(a,b) = q(a,a)q(b,b) \tag{44}$$

or

$$q(a,b) = c\sqrt{|q(a,a)|}\sqrt{|q(b,b)|} \tag{45}$$

with the constant $c = \pm 1$. But a factorizable form in (45) with $\alpha(y) = \sqrt{|q(y,y)|}$ is exactly the same as in (4) in the definition of quasi-Gaussianity, or, equivalently, in the assumptions of the Lemma, and thus excluded by assumption.

Thus, we have proven by contradiction that $\psi$ cannot be a function of $a$ alone. The functions involved are continuous by assumption, so since $\psi$ takes more than one value for any given $a$, it takes an infinity of values for any given $a$. Thus, it is possible to choose $\bar{\mathbf{u}}^2$ (corresponding to $n$ choices of $b$ for given $n$ values of $a$) so that the diagonal entries in $\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}$ are all distinct, for any given $\bar{\mathbf{u}}^1$.

Since the entries in $\mathbf{D}_{11}\mathbf{D}_{12}^{-2}\mathbf{D}_{22}$ can be assumed to be distinct, the eigenvectors of the (product) matrix on the LHS of (41) are equal to the columns of $\mathbf{M}_1^{-1}$, and uniquely defined up to a multiplication by a scalar constant which is always indetermined for eigenvectors. The diagonal entries on both sides are equal to the eigenvalues of the corresponding matrices, because eigenvalues are invariant to change of basis by $\mathbf{M}_1$, so we have $d_i''' = d_i^{11}d_i^{22}/(d_i^{12})^2$, up to permutation. On the other hand, the eigenvectors on the RHS of (41) are equal to the canonical basis vectors, and they are also uniquely defined (up to scalar multiplication) since the $d_i'''$ are also distinct. The eigenvectors on both sides must be equal, and thus, $\mathbf{M}(\bar{\mathbf{u}}^1)$ must be equal to a permutation matrix, up to multiplication of each row by a scalar which depends on $\bar{\mathbf{u}}^1$.

Since $\bar{\mathbf{u}}^1$ could be freely chosen, $\mathbf{M}(\mathbf{u})$ is equal to such a rescaled permutation matrix everywhere. By continuity the non-zero entries in $\mathbf{M}(\mathbf{u})$ must be in the same locations everywhere; if they switched locations, $\mathbf{M}(\mathbf{u})$ would have to be singular at one point at least, which is excluded by assumption. With the same logic, we see the signs of the entries cannot change. Thus the Lemma is proven.

**Proof of Theorem 2**

First, since we have a restricted form of regression function, we have to prove that it can actually converge to the optimal theoretical regression function in (23). This is true because the regression function in (13) can still approximate all quasi-Gaussian densities which have uniform dependence, after suitable transformation. Namely, uniform dependence together with quasi-Gaussianity implies that $\bar{\alpha}$ must be monotonic. Thus, by a pointwise transformation inverting such monotonic $\bar{\alpha}$, we can transform the data so that $\bar{\alpha}$ is linear, and the regression function in the Theorem can be learned to be optimal.

The proof of Theorem 1 is then valid all the way until (33), since we didn't use non-quasi-Gaussianity up to that point. We have from (33), (13), and the definition of quasi-Gaussianity

$$[\mathbf{Jk}(\mathbf{z}^1)^{-1}]^T\mathrm{diag}_i[\alpha_i(z_i^1)]\mathrm{diag}_i[c_i]\mathrm{diag}_i[\alpha_i(z_i^2)]\mathbf{Jk}(\mathbf{z}^2)^{-1} = \mathrm{diag}_i[a_i] \tag{46}$$

which must hold for any $\mathbf{z}^1, \mathbf{z}^2$. The matrices in this equation are invertible by the proof of Theorem 1. Now, define

$$\mathbf{V}(\mathbf{z}) = \mathrm{diag}_i[\alpha_i(z_i)]\mathbf{Jk}(\mathbf{z})^{-1} \tag{47}$$

so the condition above takes the form

$$\mathbf{V}(\mathbf{z}^1)^T\mathrm{diag}_i[c_i]\mathbf{V}(\mathbf{z}^2) = \mathrm{diag}_i[a_i]. \tag{48}$$

Setting $\mathbf{z}^2 = \mathbf{z}^1$, we can solve

$$\mathbf{V}(\mathbf{z}^1)^T = \mathrm{diag}_i[a_i]\mathbf{V}(\mathbf{z}^1)^{-1}\mathrm{diag}_i[1/c_i]. \quad (49)$$

Plugging this back into (48), we have

$$\mathrm{diag}_i[a_i]\mathbf{V}(\mathbf{z}^1)^{-1}\mathrm{diag}_i[1/c_i]\mathrm{diag}_i[c_i]\mathbf{V}(\mathbf{z}^2) = \mathrm{diag}_i[a_i] \quad (50)$$

which gives equivalently

$$\mathbf{V}(\mathbf{z}^1) = \mathbf{V}(\mathbf{z}^2). \quad (51)$$

That is, $\mathbf{V}(\mathbf{z})$ does not depend on $\mathbf{z}$. Denote its constant value by $\mathbf{V}$.

Solving for $\mathbf{Jk}(\mathbf{z})$ in (47) with such a constant $\mathbf{V}$, we have

$$\mathbf{Jk}(\mathbf{z}) = \mathbf{V}^{-1}\mathrm{diag}_i[\alpha_i(z_i)]. \quad (52)$$

Now, substitute, by (28), $\mathbf{J}(\mathbf{h}\circ\mathbf{f})(\mathbf{z})$ for the LHS, and change the dummy variable $\mathbf{z}$ to $\mathbf{s}$. Then we can integrate both sides to obtain

$$(\mathbf{h}\circ\mathbf{f})(\mathbf{s}) = \mathbf{h}(\mathbf{x}) = \mathbf{V}^{-1}\begin{pmatrix}\bar{\alpha}_1(s_1)\\ \bar{\alpha}_2(s_2)\\ \vdots\\ \bar{\alpha}_n(s_n)\end{pmatrix} + \mathbf{d} \quad (53)$$

for some integration constant vector $\mathbf{d}$. Thus we get the form given in the Theorem, with $\mathbf{B} = \mathbf{V}^{-1}$.

### Theory and Proof for Multiple Time Lags

In the case of multiple lags, the assumptions in a theorem corresponding to Theorem 1 are apparently identical to those in Theorem 1, but we use the general definition of quasi-Gaussianity in Definition 3, and the general definition of uniform dependence, which is that the cross-derivative $q_{j,k}(\mathbf{x})$ is non-zero for any $j, k$ and any $\mathbf{x}$. We further define the discrimination problem using (18) and use the obvious generalization of the regression function given by

$$r(\mathbf{y}) = \sum_{i=1}^{m} B_i(h_i(\mathbf{y}^1), h_i(\mathbf{y}^2), \ldots, h_i(\mathbf{y}^m)). \quad (54)$$

We can then use the proof of Theorem 1 with minimal changes. Non-quasi-Gaussianity implies that for some $j, k$, factorizability is impossible. Fix $j, k$ to those values. Fix $\mathbf{y}^p$ for $p \neq j, k$ to any arbitrary values. The proof proceeds in the same way, largely ignoring any $\mathbf{y}^p$ with $p$ not equal to $j$ or $k$. In particular, the derivative in (30) is taken with respect to those $j, k$. Furthermore, (33) has the form

$$[\mathbf{Jk}(\mathbf{z}^j)^{-1}]^T\mathrm{diag}_i[q_i(z_i^1, \ldots, z_i^m))]\mathbf{Jk}(\mathbf{z}^k)^{-1} = \mathrm{diag}_i[b_i(k_i(\mathbf{z}^1), \ldots, k_i(\mathbf{z}^m))] \quad (55)$$

where both $q_i$ and $b_i$ are functions of $\mathbf{z}^j$ and $\mathbf{z}^k$ (or, equivalently, of $\mathbf{y}^j$ and $\mathbf{y}^k$) only, since all the other $\mathbf{z}^p$ (or $\mathbf{y}^p$) are fixed.

A version of Theorem 2 for multiple time lags is left as a question for future research.