Nonlinear ICA using auxiliary variables and generalized contrastive learning
AISTATS 2019

Supplementary Material

A Proof of Theorem 1

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

$$\sum_{i=1}^{n} \psi_i(h_i(\mathbf{x}), \mathbf{u}) = \sum_i q_i(g_i(\mathbf{x}), \mathbf{u}) + \log p(\mathbf{u}) + \log|\det \mathbf{Jg}(\mathbf{x})| - \log p_{\mathbf{s}}(\mathbf{g}(\mathbf{x})) - \log p(\mathbf{u}) - \log|\det \mathbf{Jg}(\mathbf{x})|$$

where the $\log p_{\mathbf{s}}$ is the marginal log-density of the components when $\mathbf{u}$ is integrated out (as pointed above, it does not need to be factorial), $\log p(\mathbf{u})$ is the marginal density of the auxiliary variables, $\mathbf{g} = \mathbf{f}^{-1}$, and the $\mathbf{Jg}$ are the Jacobians of the inverse mixing—which nicely cancel out. Also, the marginals $\log p(\mathbf{u})$ cancel out here.

Now, change variables to $\mathbf{y} = \mathbf{h}(\mathbf{x})$ and define $\mathbf{v}(\mathbf{y}) = \mathbf{g}(\mathbf{h}^{-1}(\mathbf{y}))$, which is possible by the assumption of invertibility of $\mathbf{h}$. We then have

$$\sum_i \psi_i(y_i, \mathbf{u}) = \sum_i q_i(v_i(\mathbf{y}), \mathbf{u}) - \log p_{\mathbf{s}}(\mathbf{v}(\mathbf{y})) \quad (17)$$

What we need to prove is that this can be true for all $\mathbf{y}$ and $\mathbf{u}$ only if the v_i depend on only one of the y_i.

Denote $\bar{q}(\mathbf{y}) = \log p_{\mathbf{s}}(\mathbf{v}(\mathbf{y}))$. Taking derivatives of both sides of (17) with respect to y_j, denoting the derivatives by a superscript as

$$q_i^1(s, \mathbf{u}) = \partial q_i(s, \mathbf{u})/\partial s \quad (18)$$

$$q_i^{11}(s, \mathbf{u}) = \partial^2 q_i(s, \mathbf{u})/\partial s^2 \quad (19)$$

and likewise for ψ, and $v_i^j(\mathbf{y}) = \partial v_i(\mathbf{y})/\partial y_j$, we obtain

$$\psi_j^1(y_j, \mathbf{u}) = \sum_i q_i^1(v_i(\mathbf{y}), \mathbf{u}) v_i^j(\mathbf{y}) - \bar{q}^j(\mathbf{y}) \quad (20)$$

Taking another derivative with respect to $y_{j'}$ with $j' \neq j$, the left-hand-side vanishes, and we have

$$\sum_i q_i^{11}(v_i(\mathbf{y}), \mathbf{u})\, v_i^j(\mathbf{y})\, v_i^{j'}(\mathbf{y}) + q_i^1(v_i(\mathbf{y}), \mathbf{u})\, v_i^{jj'}(\mathbf{y}) - \bar{q}^{jj'}(\mathbf{y}) = 0 \quad (21)$$

where the $v_i^{jj'}$ are second-order cross-derivatives. Collect all these equations in vector form by defining $\mathbf{a}_i(\mathbf{y})$ as a vector collecting all entries $v_i^j(\mathbf{y})\, v_i^{j'}(\mathbf{y}), j = 1, ..., n, j' = 1, ..., j-1$ (we omit diagonal terms, and by symmetry, take only one half of the indices). Likewise, collect all the entries $v_i^{jj'}(\mathbf{y}), j = 1, ..., n, j' = 1, ..., j-1$ in the vector $\mathbf{b}(\mathbf{y})$, and all the entries $\bar{q}^{jj'}(\mathbf{y}), j = 1, ..., n, j' = 1, ..., j-1$ in the vector $\mathbf{c}(\mathbf{y})$. We can thus write the $n(n-1)/2$ equations above as a single system of equations

$$\sum_i \mathbf{a}_i(\mathbf{y}) q_i^{11}(v_i(\mathbf{y}), \mathbf{u}) + \mathbf{b}_i(\mathbf{y}) q_i^1(v_i(\mathbf{y}), \mathbf{u}) = \mathbf{c}(\mathbf{y}) \quad (22)$$

Now, collect the $\mathbf{a}$ and $\mathbf{b}$ into a matrix $\mathbf{M}$:

$$\mathbf{M}(\mathbf{y}) = \big(\mathbf{a}_1(\mathbf{y}), ..., \mathbf{a}_n(\mathbf{y}), \mathbf{b}_1(\mathbf{y}), ..., \mathbf{b}_n(\mathbf{y})\big) \quad (23)$$

Equation (22) takes the form of the following linear system

$$\mathbf{M}(\mathbf{y})\mathbf{w}(\mathbf{y}, \mathbf{u}) = \mathbf{c}(\mathbf{y}) \quad (24)$$

where $\mathbf{w}$ is defined in the Assumption of Variability, Eq. (9). This must hold for all $\mathbf{y}$ and $\mathbf{u}$. Note that the size of $\mathbf{M}$ is $n(n-1)/2 \times 2n$.

Now, fix $\mathbf{y}$. Consider the $2n+1$ points $\mathbf{u}_j$ given for that $\mathbf{y}$ by the Assumption of Variability. Collect the equations (24) above for the $2n$ points starting from index 1:

$$\mathbf{M}(\mathbf{y}) \big(\mathbf{w}(\mathbf{y}, \mathbf{u}_1), ..., \mathbf{w}(\mathbf{y}, \mathbf{u}_{2n})\big) = \big(\mathbf{c}(\mathbf{y}), \ldots, \mathbf{c}(\mathbf{y})\big) \quad (25)$$

and collect likewise the equation for index 0 repeated $2n$ times:

$$\mathbf{M}(\mathbf{y}) \big(\mathbf{w}(\mathbf{y}, \mathbf{u}_0), ..., \mathbf{w}(\mathbf{y}, \mathbf{u}_0)\big) = \big(\mathbf{c}(\mathbf{y}), \ldots, \mathbf{c}(\mathbf{y})\big) \quad (26)$$

Now, subtract (26) from (25) to obtain

$$\mathbf{M}(\mathbf{y}) \begin{pmatrix} \mathbf{w}(\mathbf{y}, \mathbf{u}_1) - \mathbf{w}(\mathbf{y}, \mathbf{u}_0), ..., \\ \mathbf{w}(\mathbf{y}, \mathbf{u}_{2n}) - \mathbf{w}(\mathbf{y}, \mathbf{u}_0) \end{pmatrix} = \mathbf{0} \quad (27)$$

The matrix consisting of the $\mathbf{w}$ here has, by the Assumption of Variability, linearly independent columns. It is square, of size $2n \times 2n$, so it is invertible. This implies $\mathbf{M}(\mathbf{y})$ is zero, and thus by definition in (23), the $\mathbf{a}_i(\mathbf{y})$ and $\mathbf{b}_i(\mathbf{y})$ are all zero.

In particular, $\mathbf{a}_i(\mathbf{y})$ being zero implies no row of the Jacobian of $\mathbf{v}$ can have more than one non-zero entry. This holds for any $\mathbf{y}$. By continuity of the Jacobian and its invertibility, the non-zero entries in the Jacobian must be in the same places for all $\mathbf{y}$: If they switched places, there would have to be a point where the Jacobian is singular, which would contradict the assumption of invertibility of $\mathbf{h}$.

This means that each v_i is a function of only one y_i. The invertibility of $\mathbf{v}$ also implies that each of these

scalar functions is invertible. Thus, we have proven the convergence of our method, as well as provided a new identifiability result for nonlinear ICA.

B Proof of Theorem 2

For notational simplicity, consider just the case $n = 2, k = 3$; the results are clearly simple to generalize to any dimensions. Furthermore, we set $Q_i \equiv 1$; again, the proof easily generalizes. The assumption of conditional exponentiality means

$$\begin{aligned} q_1(s_1, \mathbf{u}) &= \tilde{q}_{11}(s_1)\lambda_{11}(\mathbf{u}) + \tilde{q}_{12}(s_1)\lambda_{12}(\mathbf{u}) \\ &\quad + \tilde{q}_{13}(s_1)\lambda_{13}(\mathbf{u}) - \log Z_1(\mathbf{u}) \qquad (28) \\ q_2(s_2, \mathbf{u}) &= \tilde{q}_{21}(s_2)\lambda_{21}(\mathbf{u}) + \tilde{q}_{22}(s_2)\lambda_{22}(\mathbf{u}) \\ &\quad + \tilde{q}_{23}(s_2)\lambda_{23}(\mathbf{u}) - \log Z_2(\mathbf{u}) \qquad (29) \end{aligned}$$

and by definition of $\mathbf{w}$ in (9), we get

$$\mathbf{w}(\mathbf{s}, \mathbf{u}) = \begin{pmatrix} \tilde{q}'_{11}(s_1)\lambda_{11}(\mathbf{u}) + \tilde{q}'_{12}(s_1)\lambda_{12}(\mathbf{u}) + \tilde{q}'_{13}(s_1)\lambda_{13}(\mathbf{u}) \\ \tilde{q}'_{21}(s_2)\lambda_{21}(\mathbf{u}) + \tilde{q}'_{22}(s_2)\lambda_{22}(\mathbf{u}) + \tilde{q}'_{23}(s_2)\lambda_{23}(\mathbf{u}) \\ \tilde{q}''_{11}(s_1)\lambda_{11}(\mathbf{u}) + \tilde{q}''_{12}(s_1)\lambda_{12}(\mathbf{u}) + \tilde{q}''_{13}(s_1)\lambda_{13}(\mathbf{u}) \\ \tilde{q}''_{21}(s_2)\lambda_{21}(\mathbf{u}) + \tilde{q}''_{22}(s_2)\lambda_{22}(\mathbf{u}) + \tilde{q}''_{23}(s_2)\lambda_{23}(\mathbf{u}) \end{pmatrix} \qquad (30)$$

Now we fix $\mathbf{s}$ like in the Assumption of Variability, and drop it from the equation. The $\mathbf{w}(\mathbf{s}, \mathbf{u})$ above can be written as

$$\begin{aligned} &\begin{pmatrix} \tilde{q}'_{11} \\ 0 \\ \tilde{q}''_{11} \\ 0 \end{pmatrix} \lambda_{11}(\mathbf{u}) + \begin{pmatrix} \tilde{q}'_{12} \\ 0 \\ \tilde{q}''_{12} \\ 0 \end{pmatrix} \lambda_{12}(\mathbf{u}) + \begin{pmatrix} \tilde{q}'_{13} \\ 0 \\ \tilde{q}''_{13} \\ 0 \end{pmatrix} \lambda_{13}(\mathbf{u}) \\ &+ \begin{pmatrix} 0 \\ \tilde{q}'_{21} \\ 0 \\ \tilde{q}''_{21} \end{pmatrix} \lambda_{21}(\mathbf{u}) + \begin{pmatrix} 0 \\ \tilde{q}'_{22} \\ 0 \\ \tilde{q}''_{22} \end{pmatrix} \lambda_{22}(\mathbf{u}) + \begin{pmatrix} 0 \\ \tilde{q}'_{23} \\ 0 \\ \tilde{q}''_{23} \end{pmatrix} \lambda_{23}(\mathbf{u}) \end{aligned} \qquad (31)$$

So, we see that $\mathbf{w}(\mathbf{s}, \mathbf{u})$ for fixed $\mathbf{s}$ is basically given by a linear combination of nk fixed "basis" vectors, with the λ's giving their coefficients.

If $k = 1$, it is impossible to obtain the $2n$ linearly independent vectors since there are only n basis vectors. On the other hand, if $k > 1$, the vectors k vectors for each i span a 2D subspace by assumption. For different i, they are clearly independent since the non-zero entries are in different places. Thus, the nk basis vectors span a $2n$-dimensional subspace, which means we will almost surely obtain $2n$ linearly independent vectors $\mathbf{w}(\mathbf{s}, \mathbf{u}_i), i = 1, \ldots, 2n$ by this construction for λ_{ij} independently and randomly chosen from a set of non-zero measure (this is a sufficient but by no means a necessary condition). Subtraction of $\mathbf{w}(\mathbf{s}, \mathbf{u}_0)$ does not reduce the independence almost surely, since it is simply redefining the origin, and does not change the linear independence.

C Proof of Theorem 3

Denote by $\bar{q}_i(s_i)$ the marginal log-density of s_i. As in the proof of Theorem 1, assuming infinite data, well-known theory says that the regression function will converge to

$$\begin{aligned} &\sum_{i=1}^{n} \psi_i(h_i(\mathbf{x}), \mathbf{u}) = \log p(\mathbf{s}, \mathbf{u}) + \log |\mathbf{Jg}(\mathbf{x})| - \log p(\mathbf{s}) \\ &\qquad - \log p(\mathbf{u}) - \log |\mathbf{Jg}(\mathbf{x})| \\ &= \sum_i \log Q_i(s_i) + [\sum_j \tilde{q}_{ij}(s_i)\lambda_{ij}(\mathbf{u})] - \log Z_i(\mathbf{u}) - q_0(\mathbf{s}) \end{aligned} \qquad (32)$$

provided that such a distribution can be approximated by the regression function. Here, we define $q_0(\mathbf{s}) = \log p_{\mathbf{s}}(\mathbf{s})$. In fact, the approximation is clearly possible since the difference of the log-pdf's is linear in the same sense as the regression function. In other words, a solution is possible as

$$\begin{aligned} \sum_{ij} \tilde{h}_{ij}(\mathbf{x})^T v_{ij}(\mathbf{u}) + a(\mathbf{x}) + b(\mathbf{u}) &= \sum_{ij} \tilde{q}_{ij}(s_i)\lambda_{ij}(\mathbf{u}) \\ &\quad + \sum_i \log Q_i(s_i) - q_0(\mathbf{s}) - \log Z_i(\mathbf{u}) \qquad (33) \end{aligned}$$

with

$$\tilde{h}_{ij}(\mathbf{x}) = \tilde{q}_{ij}(\mathbf{x}) \qquad (34)$$

$$v_{ij}(\mathbf{u}) = \lambda_{ij}(\mathbf{u}) \qquad (35)$$

$$a(\mathbf{x}) = \sum_i \log Q_i(s_i) - q_0(\mathbf{s}) \qquad (36)$$

$$b(\mathbf{u}) = \sum_i - \log Z_i(\mathbf{u}) \qquad (37)$$

Thus, we can have the special form for the regression function in (11). Next, we have to prove that this is the only solution up to the indeterminacies given in the Theorem.

Collect these equations for all the $\mathbf{u}_k$ given by Assumption 3 in the Theorem. Denote by $\mathbf{L}$ a matrix of the $\lambda_{ij}(\mathbf{u}_k)$, with the product of i, j giving row index and k column index. Denote a vector of all the sufficient statistics of all the independent components as $\tilde{\mathbf{q}}(\mathbf{x}) = (\tilde{q}_{11}(s_1), ..., \tilde{q}_{nk}(s_n))^T$.Collect all the $\mathbf{v}(\mathbf{u}_k)^T$ into a matrix $\mathbf{V}$ with again k as the column index. Collect the terms $\sum_i \log Z_i(\mathbf{u}_k) + b(\mathbf{u}_k)$ for all the different k into a vector $\mathbf{z}$.

Expressing (33) for all the time points in matrix form, we have

$$\mathbf{V}^T\tilde{\mathbf{h}}(\mathbf{x}) = \mathbf{L}^T\tilde{\mathbf{q}}(\mathbf{s}) - \mathbf{z} + \mathbf{1}[\sum_i \log Q_i(s_i) - q_0(\mathbf{s}) - a(\mathbf{x})] \quad (38)$$

where $\mathbf{1}$ is a $T \times 1$ vector of ones. Now, on both sides of the equation, subtract the first row from each of the other rows. We get

$$\bar{\mathbf{V}}^T\tilde{\mathbf{h}}(\mathbf{x}) = \bar{\mathbf{L}}^T\tilde{\mathbf{q}}(\mathbf{s}) - \bar{\mathbf{z}} \quad (39)$$

where the matrices with bars are such differences of the rows of $\mathbf{V}^T$ and $\mathbf{L}^T$, and likewise for $\mathbf{z}$. We see that the last term in (38) disappears.

Now, the matrix $\bar{\mathbf{L}}$ is indeed the same as in Assumption 3 of the Theorem, which says that the modulations of the distributions of the s_i are independent in the sense that $\bar{\mathbf{L}}$ is invertible. Then, we can multiply both sides by the inverse of $\bar{\mathbf{L}}$ and get

$$\mathbf{A}\tilde{\mathbf{h}}(\mathbf{x}) = \tilde{\mathbf{q}}(\mathbf{s}) - \tilde{\mathbf{z}} \quad (40)$$

with an unknown matrix $\mathbf{A} = \bar{\mathbf{L}}^{-1}\bar{\mathbf{W}}$, and a constant vector $\tilde{\mathbf{z}} = \bar{\mathbf{L}}^{-1}\bar{\mathbf{z}}$.

Thus, just like in TCL, we see that the hidden units give the sufficient statistics $\tilde{\mathbf{q}}(\mathbf{s})$, up to a linear transformation $\mathbf{A}$, and the Theorem is proven.

D Alternative formulation of the Assumption of Variablity

To further strengthen our theory, we provide an alternative formulation of the Assumption of Variability. We define the following alternative:

> [**Alternative Assumption of Variability**] Assume $\mathbf{u}$ is continuous-valued, and that there exist $2n$ values for $\mathbf{u}$, denoted by $\mathbf{u}_j, j = 1...2n$ such that the $2n$ vectors in $\mathbb{R}^{2n}$ given by
>
> $$(\tilde{\mathbf{w}}(\mathbf{y}, \mathbf{u}_1), \tilde{\mathbf{w}}(\mathbf{y}, \mathbf{u}_2), ..., \tilde{\mathbf{w}}(\mathbf{y}, \mathbf{u}_{2n})) \quad (41)$$
>
> with
>
> $$\tilde{\mathbf{w}}(\mathbf{s}, \mathbf{u}) = (\frac{\partial^2 q_1(s_1, \mathbf{u})}{\partial s_1 \partial u_j}, \ldots, \frac{\partial^2 q_n(s_n, \mathbf{u})}{\partial s_n \partial u_j}, \frac{\partial^3 q_1(s_1, \mathbf{u})}{\partial s_1^2 \partial u_j}, \ldots, \frac{\partial^3 q_n(s_n, \mathbf{u})}{\partial s_n^2 \partial u_j}) \quad (42)$$
>
> are linearly independent, for some choice of the auxiliary variable index j.

Theorem 1 holds with with this alternative assumption as well. In the proof of the Theorem, take derivatives of both sides of (25) with respect to the u_j in the Theorem. Then, the right-hand-side vanishes, and we have an equation similar to (25) but with $\tilde{\mathbf{w}}$. All the logic after (27) applies to that equation.

E Using a function of x as auxiliary variable

We provide an informal proof without full generality to show why defining $\mathbf{u}$ as a direct deterministic function of $\mathbf{x}$ is likely to violate the assumption of conditional independent. Consider a simple linear mixing $x_1 = s_1 + s_2$ (with something similar for x_2), and define tentatively $u = x_1$. Conditioning s_1 on u will now create the dependence $s_1 = x_1 - s_2 = u - s_2$ which violates conditional independence. (This example would be more realistic with additive noise $u = x_1 + n$ to avoid degenerate pdf's, but the same logic applies anyway.) In fact, if we could make the model identifiable by such $\mathbf{u}$ defined as a function of $\mathbf{x}$, we would have violated the basic unidentifiability theory by Darmois. Thus, conditional independence implies that $\mathbf{u}$ must bring new information in addition to $\mathbf{x}$, and this information must be, in some very loose intuitive sense, "sufficiently independent" of the information in $\mathbf{x}$.

F Additional discussion to Section 5.2

In (Hyvärinen and Morioka, 2017a), the model was proven to be identifiable under two assumptions: First, the joint log-pdf of two consecutive time points is not "factorizable" in the conditionally exponential form of order one, A variant of such dependency was called "quasi-Gaussianity" in (Hyvärinen and Morioka, 2017a). However, here we use a different terminology to highlight the connection to the exponential family important in our theory as well as TCL. There is also a slight difference between the two definitions, since in (Hyvärinen and Morioka, 2017a), it was only necessary to exclude the case where the two functions in the factorization are equal, i.e. $\tilde{q}_1 = \lambda_1$ in the current notation. The second assumption was that there is a rather strong kind of temporal dependency between the time points, which was called uniform dependency. Here, we need no such latter condition, essentially because here we constrain $\mathbf{h}$ to be invertible, which was not done in (Hyvärinen and Morioka, 2017a), but seems to have a somewhat similar effect.

G Additional discussion to Section 5.3

One might ask whether it would better to randomize t and $\mathbf{x}(t-1)$ separately, by using two independent random indices t^* and $\mathbf{x}(t^{**}-1)$. The choice between these two should be made based on how to modulate the conditional distribution $p(s_i|t, \mathbf{x}(t-1))$ as strongly as possible. In practice, we would intuitively assume it is usually best to use a single time index as above, because then the dependency in t^* and $\mathbf{x}(t^*-1)$ will make the modulation stronger. Moreover, the Theo-

rems above would not apply directly to a case where we have two different random indices, although the results might be easy to reformulate for such a case as well.

H Acknowledgments

A.H. was supported by CIFAR and the Gatsby Charitable Foundation. H.S. was supported by JSPS KAKENHI 18K18107. R.E.T. thanks EPSRC grant EP/M026957/1.