PROCEEDINGS OF SCIENCE

Developing a New Nonlinear Independent Component Analysis Scheme

Jun'ichi Yokoyama[a,*]

[a] *Research Center for the Early Universe (RESCEU) and Department of Physics, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan*

E-mail: yokoyama@resceu.s.u-tokyo.ac.jp

A new method of nonlinear Independent Component Analysis is developed. This method is useful to subtract various nonlinearly coupled noises to improve the sensitivity of gravitational wave detectors.

38th International Cosmic Ray Conference (ICRC2023)
26 July - 3 August, 2023
Nagoya, Japan

*Speaker

PoS(ICRC2023)1572

1. Introduction

The independent component analysis (ICA) is a method of signal separation making use of a number of output data [1–3]. It occupies a unique position among methods of signal processing because it makes use of non-Gaussianity of signals and noises instead of treating it as an obstacle. We have been studying it toward application to gravitational-wave data analysis. By effectively removing non-Gaussian noises, ICA can support conventional matched filtering techinique for gravitational wave data analysis which is optimal for Gaussian noises [4].

The most famous problem to which ICA can be used is what is called the cocktail party problem. A number of people a chatting in a party and we monitor their voices by a number of microphones, which receive superposed voices of some or all of the attendees. The ICA separates the sound of each source, or the voice of each participant making use of the statistical independence of each source. Similarly, ICA can be used to separate gravitational-wave signal from non-Gaussian noises by simultaneously monitoring both the strain channel and physical environmental monitor (PEM) channels.

In this paper we propose a new method of nonlinear ICA which is strikingly different from existent methods [5, 6]. Our approach uses a variational method deriving a master equation analogous to the Euler-Lagrange equation in analytical mechanics.

2. Linear Model

Here we first introduce the concept of ICA using a simple model where there are $n+1$ independent sources of signal and noises, $\boldsymbol{s}(t) =^t (s_0(t), s_1(t), ..., s_n(t))$ and observables $\boldsymbol{x}(t) =^t (x_0(t), x_1(t), ..., x_n(t))$ which are interrelated by an instantaneous linear relation

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t) \tag{1}$$

where A is assumed to be a time independent matrix. Our ultimate goal is to reconstruct $\boldsymbol{s}(t)$ out of observables $\boldsymbol{x}(t)$, but it is not possible to do so in full as we do not know each component of A. What we do here to implement ICA is to try to find another set of variables $\boldsymbol{y}(t)$ which are given by a linear transformation of $\boldsymbol{x}(t)$,

$$\boldsymbol{y}(t) = W\boldsymbol{x}(t) \tag{2}$$

in such a way that each component of $\boldsymbol{y}(t)$ is statistically independent. The ICA can achieve it if signals and noises have non-Gaussian distributions except for one Gaussian variable.

The mutual independence of statistical variables may be judged by introducing a cost function $L(W)$ which represents a "distance" in the space of statistical distribution functionals. As an example, we adopt the Kullback-Leibler (KL) divergence [7] defined between two arbitrary PDFs $p(\boldsymbol{y})$ and $q(\boldsymbol{y})$ as

$$D[p(\boldsymbol{y}); q(\boldsymbol{y})] = \int p(\boldsymbol{y}) \ln \frac{p(\boldsymbol{y})}{q(\boldsymbol{y})} d^{n+1}y = E_p \left[\ln \frac{p(\boldsymbol{y})}{q(\boldsymbol{y})} \right], \tag{3}$$

where $E_p[\cdot]$ denotes an expectation value with respect to a PDF p.

We examine the distance between the real distribution function of statistically independent variables $\boldsymbol{s}$, $r(\boldsymbol{s}) = \prod_{i=0}^{n} r_i[s_i(t)]$, and a distribution of $\boldsymbol{y}$, p_y, constructed from the observed distribution function of $\boldsymbol{x}$ through the linear transformation $\boldsymbol{y} = W\boldsymbol{x}$ as

$$p_y(\boldsymbol{y}) \equiv ||W^{-1}||p_x(\boldsymbol{x}), \tag{4}$$

where $||W^{-1}||$ denotes the determinant of W^{-1}.

The cost function of $p_y(\boldsymbol{y})$ from $r(\boldsymbol{s})$ is given by

$$\begin{aligned} L_r(W) &= D[p_y(\boldsymbol{y}); r(\boldsymbol{y})] = E_{p_y}[\ln p_y(\boldsymbol{y})] - E_{p_y}[\ln r(\boldsymbol{y})] \\ &= \int ||W^{-1}||p_x(\boldsymbol{x}) \ln\left[||W^{-1}||p_x(\boldsymbol{x})\right] d^{n+1}y - E_{p_y}[\ln r(\boldsymbol{y})] \\ &= -\ln ||W|| + \int p_x(\boldsymbol{x}) \ln [p_x(\boldsymbol{x})]\, d^{n+1}x - E_{p_y}[\ln r(\boldsymbol{y})] \\ &= -H[x] - E_{p_y}[||W|| \ln r(\boldsymbol{y})] = -H[x] - E_{p_x}[\ln p(\boldsymbol{x}, W)], \end{aligned} \tag{5}$$

with

$$p(\boldsymbol{x}, W) \equiv ||W||r(\boldsymbol{y}), \tag{6}$$

and

$$H[x] \equiv -\int p_x(\boldsymbol{x}) \ln [p_x(\boldsymbol{x})]\, d^{n+1}x. \tag{7}$$

The PDF of $\boldsymbol{x}$ in the last expression of (5) has W dependence because $p(\boldsymbol{x}, W)$ is a PDF of $\boldsymbol{x}$ which is made out of the PDF of $\boldsymbol{y}$ (= $\boldsymbol{s}$ in this particular case) through the relation $\boldsymbol{y} = W\boldsymbol{x}$. The above formula shows that the matrix W which minimizes the cost function $L_r(W)$ also maximizes the log-likelihood ratio of $\boldsymbol{x}$.

Since we do not know $r(\boldsymbol{y})$ a priori, we instead adopt an arbitrary mutually independent distribution $q(\boldsymbol{y}) = \prod_{i=0}^{n} q_i(y_i)$ in the cost function. Defining a PDF consisting of marginal distribution functions

$$\tilde{p}(\boldsymbol{y}) \equiv \prod_{i=0}^{n} \int p_y(y_0, y_1, ..., y_n) \prod_{j\neq i} dy_j = \prod_{i=0}^{n} \tilde{p}_i(y_i), \tag{8}$$

we find the following relation

$$L_q(W) = D[p_y(\boldsymbol{y}); q(\boldsymbol{y})] = D[p_y(\boldsymbol{y}); \tilde{p}(\boldsymbol{y})] + D[\tilde{p}(\boldsymbol{y}); q(\boldsymbol{y})] \tag{9}$$

holds. Since the Kullback-Leibler divergence is known to be positive semi-definite, a distribution that minimizes the first term in the right-hand-side yields the desired linear transformation $\boldsymbol{y} = W\boldsymbol{x}$ for which this term vanishes. In this case the second term gives a discrepancy due to the possible incorrect choice of $q(\boldsymbol{y})$. In this sense it would be better to choose a realistic trial function $q(\boldsymbol{y})$ as much as possible.

It is known in fact that even for an arbitrary choice of $q(\boldsymbol{y})$, the correct W gives an extremum of $L_q(W)$. Hence we solve

$$\frac{\partial L_q(W)}{\partial w_{ij}} = 0. \tag{10}$$

From

$$L_q(W) = -H[x] - \ln||W|| - E_{p_y}[\ln q(\boldsymbol{y})] \equiv -H[x] - \ln||W|| - E_{p_y}[f(\boldsymbol{y})], \tag{11}$$

$$f(\boldsymbol{y}) \equiv \ln q(\boldsymbol{y}), \tag{12}$$

$$\begin{aligned} d_W \ln||W|| &\equiv \ln||W + dW|| - \ln||W|| = \ln||\mathbf{1} + dWW^{-1}|| \\ &= \mathrm{Tr}(dWW^{-1}) = (W^{-1})_{ji} dw_{ij}, \end{aligned}$$

and

$$d_W f(\boldsymbol{y}) \equiv f((W + dW)\boldsymbol{x}) - f(W\boldsymbol{x}) = \frac{\partial f}{\partial y_i} dw_{ij} x_j,$$

we find

$$d_W L_q(W) = E_{p_y}\left[-(W^{-1})_{ji} - x_j \frac{\partial f}{\partial y_i}\right] dw_{ij}. \tag{13}$$

In order to satisfy (10) the above expectation value should vanish for each index. Multiplying w_{kj} to the argument of the expectation value, we find it equivalent to

$$E_{p_y}\left[y_k \frac{\partial f}{\partial y_i}\right] + \delta_{ki} = 0. \tag{14}$$

That is, we require

$$E_{p_y}[\varphi_i(y_i) y_j] = \delta_{ij} \tag{15}$$

with

$$\varphi_i(y_i) \equiv -\frac{d}{dy_i} \ln q_i(y_i). \tag{16}$$

We determine W so that (16) is satisfied for each component choosing plausible forms of $q_i(y_i)$. For $q_1(y_1)$ we take

$$q_1(y_1) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_1 - h(t,\theta))^2}{2\sigma^2}\right], \tag{17}$$

so that $\varphi_1(y_1) = (y_1 - h(t,\theta))/\sigma^2$. In real experiments, we do not know $h(t,\theta)$ a priori. However, as found later, when we take temporal average, the contributions from gravitational waves can be neglected. In fact, we can set $h(t,\theta) = 0$ when we apply $q_1(y_1)$ to real analysis. As for $\varphi_2(y_2)$, it is recommended to take

$$\varphi_2(y_2) = c_2 \tanh y_2 \tag{18}$$

to model a super-Gaussian distribution [8]. Using these expressions in (16) we determine W which relates each component of $\boldsymbol{y}$ and $\boldsymbol{x}$ as $y_1 = w_{11}x_1 + w_{12}x_2$ and $y_2 = w_{22}x_2$. In doing so we replace the ensemble average $E[\cdot]$ by temporal average of observed values of $\boldsymbol{x}$ which we denote by brackets.

3. Nonlinear case

We now extend the above analysis to the case observables $\boldsymbol{x}(t)$ and sources $\boldsymbol{s}(t)$ are nonlinearly related. Our reconstruction problem is now to find a set of functions $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ such that each component of $\boldsymbol{y}(t)$ is statistically independent with others. For the moment, we assume that this relation holds at any time and the probability distribution functions (PDFs) of $\boldsymbol{x}$ and $\boldsymbol{y}$ are related with each other by

$$P_y(\boldsymbol{y})d^{n+1}y = P_x(\boldsymbol{x})d^{n+1}x = P_x(\boldsymbol{x}(\boldsymbol{y}))\left|\frac{\partial(\boldsymbol{x})}{\partial(\boldsymbol{y})}\right| d^{n+1}y, \tag{19}$$

which means

$$P_y(\boldsymbol{y}) = \prod_{i=0}^{n} \tilde{P}_i(y_i) = P_x(\boldsymbol{x}(\boldsymbol{y}))\left|\frac{\partial(\boldsymbol{x})}{\partial(\boldsymbol{y})}\right|. \tag{20}$$

As before, we wish to minimize the KL divergence

$$L_q(\boldsymbol{x}(\boldsymbol{y})) = D[P_y(\boldsymbol{y}), q(\boldsymbol{y})] = E_{P_y}[\ln P_y(\boldsymbol{y})] - E_{P_y}[\ln q(\boldsymbol{y})], \quad q(\boldsymbol{y}) \equiv \prod_{k=0}^{n} q_k(y_k), \tag{21}$$

with

$$\begin{aligned} E_{P_y}[\ln P_y(\boldsymbol{y})] &= \int \frac{\partial(\boldsymbol{x})}{\partial(\boldsymbol{y})} P_x(\boldsymbol{x}(\boldsymbol{y})) \ln\left[\frac{\partial(\boldsymbol{x})}{\partial(\boldsymbol{y})} P_x(\boldsymbol{x}(\boldsymbol{y}))\right] d^{n+1}y \\ &= \int P_x(\boldsymbol{x}(\boldsymbol{y})) \ln\left[\frac{\partial(\boldsymbol{x})}{\partial(\boldsymbol{y})}\right] d^{n+1}x + \int P_x(\boldsymbol{x}(\boldsymbol{y})) \ln P_x(\boldsymbol{x}(\boldsymbol{y})) d^{n+1}x, \end{aligned} \tag{22}$$

and

$$\begin{aligned} E_{P_y}[\ln P_y(\boldsymbol{y})] &= \int \frac{\partial(\boldsymbol{x})}{\partial(\boldsymbol{y})} q(\boldsymbol{x}(\boldsymbol{y})) \ln\left[\frac{\partial(\boldsymbol{x})}{\partial(\boldsymbol{y})} P_x(\boldsymbol{x}(\boldsymbol{y}))\right] d^{n+1}y \\ &= \int P_x(\boldsymbol{x}(\boldsymbol{y})) \ln\left[\frac{\partial(\boldsymbol{x})}{\partial(\boldsymbol{y})}\right] d^{n+1}x + \int P_x(\boldsymbol{x}(\boldsymbol{y})) \ln\left[\frac{\partial(\boldsymbol{y})}{\partial(\boldsymbol{x})} q(\boldsymbol{y})\right] d^{n+1}x, \end{aligned} \tag{23}$$

so that

$$L_q(\boldsymbol{x}(\boldsymbol{y})) = \int P_x(\boldsymbol{x}(\boldsymbol{y})) \ln P_x(\boldsymbol{x}(\boldsymbol{y})) d^{n+1}x - \int P_x(\boldsymbol{x}(\boldsymbol{y})) \ln\left[\frac{\partial(\boldsymbol{y})}{\partial(\boldsymbol{x})} q(\boldsymbol{y})\right] d^{n+1}x. \tag{24}$$

We wish to minimize the second term of the right-hand-side of (24) with respect to the function $\boldsymbol{y}(\boldsymbol{x})$. Neglecting the first term of the right-hand-side of (24) hereafter, we may rewrite the minization problem by an action principle

$$L_q(\boldsymbol{x}(\boldsymbol{y})) = -\int d^{n+1}x \mathcal{L}\left(y_i(\boldsymbol{x}), \frac{\partial_i}{\partial x_j}(\boldsymbol{x})\right) \tag{25}$$

with the Lagrangian

$$\mathcal{L} = P_x(\boldsymbol{x}(\boldsymbol{y})) \ln\left[\frac{\partial(\boldsymbol{y})}{\partial(\boldsymbol{x})} q(\boldsymbol{y}(\boldsymbol{x}))\right]. \tag{26}$$

The action (25) is minimized by a solution of the Euler-Lagrange equation:

$$\frac{\delta L_q(\boldsymbol{y})}{\delta y_k(\boldsymbol{x})} = \sum_\ell \frac{d}{dx_\ell} P_x(\boldsymbol{x}) \frac{\partial x_\ell}{\partial y_k} - P_x(\boldsymbol{x}) \frac{\partial}{\partial y_k} \ln q_k(y_k) = 0. \tag{27}$$

Multiplying this by x_j and integrating over $d^{n+1}x$ we find

$$E_{P_x}\left[-\frac{\partial x_\ell}{\partial y_k} - x_j \frac{\partial}{\partial y_l} \ln q_k(y_k) \right] = 0, \tag{28}$$

which is a direct extension of (13). Indeed, $\partial x_\ell / \partial y_k$ corresponds to $(W^{-1})_{\ell k}$ in the case the transformation is linear.

4. Memory effect

We can extend the analysis in the case with memory effect dealing with the PDFs over the entire time span of interest and assuming invertibility. That is, the PDF P_x is a functional of $\boldsymbol{x}(t_*)$ at all times t_* in the relevant range and it is related to that of $\boldsymbol{y}(t_*)$ as

$$P_x[\boldsymbol{x}(t_*)][d^{n+1}x(t_*)] = P_x[\boldsymbol{x}(t_*)] \prod_\alpha d^{n+1}x(t_\alpha) = P_y[\boldsymbol{y}(t_*)] \prod_\beta d^{n+1}x(t_\beta) = P_y[\boldsymbol{y}(t_*)][d^{n+1}y(t_*)]. \tag{29}$$

In terms of Fourier transformed modes,

$$\tilde{x}_i(f_\alpha) = \int x_i(t) e^{2\pi i f_\alpha t} dt, \tag{30}$$

the above relation is expressed as

$$P_y[\tilde{\boldsymbol{y}}(f_*)] \prod_\beta d^{n+1}y(f_\beta) = P_x[\tilde{\boldsymbol{x}}(f_*)] \prod_\alpha d^{n+1}x(f_\alpha) = P_x[\tilde{\boldsymbol{x}}(f_*)], \tag{31}$$

where f_* denotes all the frequencies collectively.

The KL divergence is minimized by the solution of the following Euler-Lagrange equation.

$$\frac{L_q[\tilde{\boldsymbol{y}}[\tilde{\boldsymbol{x}}(f_*)]]}{\delta \tilde{y}_k(f_\alpha)} = \sum_\beta \sum_\ell \frac{d}{dx_\ell(f_\beta)} P_x[\tilde{\boldsymbol{x}}(f_*)] \frac{\partial \tilde{x}_\ell(f_\beta)}{\partial \tilde{y}_k(f_\alpha)} - P_x[\tilde{\boldsymbol{x}}(f_*)] \frac{\partial}{\partial \tilde{y}_k(f_\alpha)} \ln q_k[y_k(f_\alpha)] = 0. \tag{32}$$

Once again, by multiplying $x_k(f_\gamma)$ and integrating over $\prod_\alpha d^{n+1}x(f_\alpha)$, we obtain a formula

$$E_{P_x}\left[-\frac{\partial \tilde{x}_\ell(f_\gamma)}{\partial \tilde{y}_k(f_\alpha)} - x_j(f_\gamma) \frac{\partial}{\partial y_k(f_\alpha)} \ln q_k[y_k(f_\gamma)] \right] = 0, \tag{33}$$

which is to be compared with (13) and (28).

5. Example

Let us consider a case there are two statistically independent noises, $w_1(f)$ and $w_2(f)$ that can be measured by some sensors x_1 and x_2 as $\tilde{x}_1(f) = w_1(f)$ and $\tilde{x}_2(f) = w_2(f)$ and the strain channel x_0 measureing the gravitational wave $h(f)$ is affected by these two noises nonlinearly as

$$\tilde{x}_0 = h(f) + n(f) + \int df_1 df_2 K_{12} w_1(f_1) w_2(f_2) \delta(f - f_1 - f_2), \tag{34}$$

where $K_{12}(f_1, f_2)$ is an unkown function and $n(f)$ is residual noises. We define

$$\begin{aligned} \tilde{y}_0(f) =& \tilde{x}_0(f) - \int df_1 K_{12}(f_1, f - f_1) \tilde{x}_1(f_1) \tilde{x}(f - f_1), \\ \tilde{y}_1(f) =& \tilde{x}_1(f), \\ \tilde{y}_2(f) =& \tilde{x}_2(f). \end{aligned} \tag{35}$$

Multiplying $\tilde{x}(f_\mu)\tilde{x}(f_\nu)$ the Euler-Lagrange equation (33) by $\tilde{x}(f_\mu)\tilde{x}(f_\nu)$ and integrating over the phase space, we find

$$\left\langle -\tilde{x}_2(f_\nu) \frac{\partial \tilde{x}_1(f_\mu)}{\partial \tilde{y}_k(f_\alpha)} - \tilde{x}_1(f_\mu) \frac{\partial \tilde{x}_2(f_\nu)}{\partial \tilde{y}_k(f_\alpha)} - \tilde{x}(f_\mu)\tilde{x}(f_\nu) \frac{\partial}{\partial \tilde{y}_k(f_\alpha)} \ln q_k[t y_k(f_\alpha)] \right\rangle = 0 \tag{36}$$

For $k = 0$ we find

$$\frac{\partial}{\partial \tilde{y}_0(f_\alpha)} \ln q_0(\tilde{y}_k(f_\alpha)) = -\frac{1}{\sigma^2} \left(\tilde{x}_0(f_\alpha) - \int df_1 K_{12}((f_1, f_\alpha - f_1) \tilde{x}_1(f_1) \tilde{x}(f_\alpha - f_1) \right). \tag{37}$$

Multiplying the above equation by $\tilde{x}_1(f_\mu)\tilde{x}_2(f_\nu)$ and taking the statistical average, we find

$$\begin{aligned} &\left\langle \tilde{x}_1(f_\mu)\hat{x}_2(f_\nu) \frac{\partial}{\partial \tilde{y}_0(f_\alpha)} \ln q_0(\tilde{y}_k(f_\alpha)) \right\rangle \\ &= \left\langle \tilde{x}_0(f_\alpha)\tilde{x}_1(f_\mu)\tilde{x}_2(f_\nu) - \int df_1 K_{12}((f_1, f_\alpha - f_1)\tilde{x}_1(f_1)\tilde{x}_1(f_\mu)\tilde{x}_2(f_\nu)\tilde{x}(f_\alpha - f_1) \right\rangle = 0 \end{aligned} \tag{38}$$

Assuming that $\tilde{x}_1$ and $\tilde{x}_2$ are stationary independent noises, we find

$$\langle \tilde{x}_i(f)\tilde{x}_j(f') \rangle = \langle |\tilde{x}_i(f)|^2 \rangle \delta_{ij} \delta(f - f'). \tag{39}$$

We can therefore determine the kernel function as

$$K_{12}(f_1, f_2) = \frac{\langle \tilde{x}_0(f_1 + f_2)\tilde{x}_1(f_1)\tilde{x}_2(f_2) \rangle}{\langle |\tilde{x}_1(f_1)|^2 \rangle \langle |\tilde{x}_2(f_2)|^2 \rangle}. \tag{40}$$

Thus this nonlinear model can be solved in this method.

6. Conclusion

The linear ICA has been applied to both iKAGRA data [9] and O3GK data [10] and proven to be effective for gravitational wave data analysis. The application of the nonlinear ICA proposed here is now underway.

7. Acknowledgements

I would like to thank members of KAGRA PEM team for discussions. This work was partially supported by JSPS KAKENHI Grant No. 20H05639.

References

[1] Jutten, C. and Herault, J. (1991) Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture. Signal Processing **24**, 1.

[2] Common, P. (1994) Independent component analysis, a new concept ? Signarl Processing **36**, 287.

[3] Amari, S. and Cardoso, J.-F. (1997) Blind Source Separation—Semiparametric Statistical Approach. IEEE Trans. Signal Processing **45**, 2692.

[4] S. Morisaki, J. Yokoyama, K. Eda and Y. Itoh, Proc. Japan Acad. B **92** (2016), 336 doi:10.2183/pjab.92.336 [arXiv:1605.01983 [gr-qc]].

[5] A. Hyvärinen and P. Pajunen, Neural Networks **12** (1999), 429

[6] K. Zhang and L. Chan, J. Machine Learning Res. **9** (2008), 2455

[7] Kullback, S. and Leibler, R.A. (1952) On Information and Sufficiency. Annals of Mathematical Statistics **22**, 79.

[8] S. Amari, T.P. Chen, and A. Cichocki, Neural Networks **10** (1997), 1345

[9] T. Akutsu *et al.* [KAGRA], PTEP **2020**, no.5, 053F01 (2020) doi:10.1093/ptep/ptaa056 [arXiv:1908.03013 [astro-ph.IM]].

[10] H. Abe *et al.* [KAGRA], Class. Quant. Grav. **40**, no.8, 085015 (2023) doi:10.1088/1361-6382/acc0cb [arXiv:2206.05785 [astro-ph.IM]].