Journal of Multivariate Analysis 181 (2021) 104689

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On the estimation of entropy in the FastICA algorithm

Elena Issoglio, Paul Smith *, Jochen Voss

School of Mathematics, University of Leeds, UK

ARTICLE INFO

Article history:
Received 3 June 2019
Received in revised form 21 September 2020
Accepted 21 September 2020
Available online 9 October 2020

AMS 2010 subject classifications:
primary 62-04
secondary 65C60

Keywords:
Approximation
Blind source separation
Convergence
Counterexample
FastICA
Independent component analysis
Projection pursuit
Projections

ABSTRACT

The fastICA method is a popular dimension reduction technique used to reveal patterns in data. Here we show both theoretically and in practice that the approximations used in fastICA can result in patterns not being successfully recognised. We demonstrate this problem using a two-dimensional example where a clear structure is immediately visible to the naked eye, but where the projection chosen by fastICA fails to reveal this structure. This implies that care is needed when applying fastICA. We discuss how the problem arises and how it is intrinsically connected to the approximations that form the basis of the computational efficiency of fastICA.

1. Introduction

Independent Component Analysis (ICA) is a well-established and popular dimension reduction technique that finds an orthogonal projection of data onto a lower-dimensional space, while preserving some of the original structure. ICA is also used as a method for blind source separation and is closely connected to projection pursuit. We refer the reader to Hyvärinen et al. [13], Hyvärinen [12] and Stone [21] for a comprehensive overview of the mathematical principles underlying ICA and its applications in a wide variety of practical examples.

In ICA, the projections that are determined to be "interesting" are those that maximise the non-Gaussianity of the data, which can be measured in several ways. One quantity for this measurement that is used frequently in the ICA literature is entropy. For distributions with a given variance, the Gaussian distribution is the one which maximises entropy, and all other distributions have strictly smaller entropy. Therefore, our aim is to find projections which minimise the entropy of the projected data. Different methods are available for both the estimation of entropy and the optimisation procedure, and have different speed–accuracy trade-offs.

A widely used method to perform ICA in higher dimensions is fastICA [14]. This method has found applications in areas as wide ranging as facial recognition [6], epileptic seizure detection [26] and fault detection in wind turbines [7]. Recent works on extensions of the algorithm can be seen in Miettinen et al. [18], Ghaffarian and Ghaffarian [8] and He et al. [10]. The fastICA method uses a series of substitutions and approximations of the projected density and its entropy. It then applies an iterative scheme for optimising the resulting contrast function (which is an approximation to negentropy).

* Corresponding author.
E-mail address: mmpws@leeds.ac.uk (P. Smith).

https://doi.org/10.1016/j.jmva.2020.104689

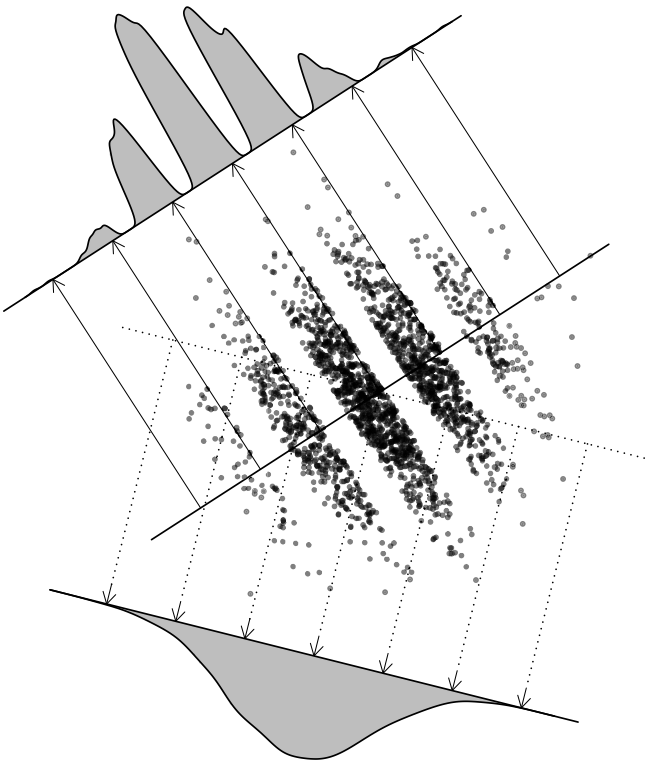


Fig. 1. Scatter plot of original data with densities of the projected data in the direction obtained by m-spacing ICA (solid line) and fastICA (dotted line). Kernel density estimation was used to obtain the marginal densities shown.

Because of its popularity in many areas, analysis and evaluation of the strengths and weaknesses of the fastICA algorithm is crucially important. In particular, we need to understand both how well the contrast function estimates entropy and the performance of the optimisation procedure.

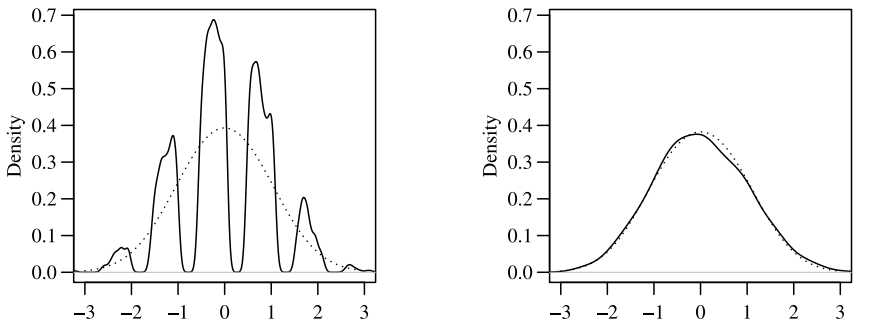
The main strength of the fastICA method is its speed, which is considerably higher than many other methods. Furthermore, if the data is a mixture of a small number of underlying factors, fastICA is often able to correctly identify these factors. However, fastICA also has some drawbacks, which have been pointed out in the literature. Learned-Miller and Fisher [16] use test problems from Bach and Jordan [2] with performance measured by the Amari error [1] to compare fastICA to other ICA methods. They find that these perform better than fastICA on many examples. Focussing on a different aspect, Wei [23] investigates issues with the convergence of the iterative scheme employed by fastICA to optimise the contrast function. In Wei [24] it is shown that the two most common fastICA contrast functions fail to de-mix certain bimodal distributions with Gaussian mixtures, although some other contrast function choices (related to classical kurtosis estimation) may give reliable results within the fastICA framework.

In this article we identify and discuss a more fundamental problem with fastICA. We demonstrate that the approximations used in fastICA can lead to a contrast function where the optimal points no longer correspond to directions of low entropy.

To illustrate the effect discussed in this paper, we consider the example shown in Fig. 1. In this example, two-dimensional samples are generated from a two-dimensional normal distribution, conditioned on avoiding a collection of parallel bands (see Section 5 for details). This procedure produces a pattern which is obvious to the bare eye, and indeed the projection which minimises entropy (solid lines) exposes this pattern. In contrast, the fastICA contrast function seems unable to resolve the pattern of bands and prefers a direction with higher entropy which does not expose the obvious pattern (dotted lines).

We remark that this failure by fastICA to recover the obvious structure is relatively robust in this example. Changing the parameters used in the fastICA method does not significantly change the outcome, and the underlying structure is still lost. It is also worth mentioning here that the example dataset was very simple to obtain and no optimisation was performed to make the fastICA method perform poorly.

To obtain the projection indicated by the solid lines in Fig. 1 we used the m-spacing method [3] for entropy estimation, combined with a standard optimisation technique. The m-spacing entropy approximation is shown to be consistent in Hall [9] and converges to true entropy for large sample size. While the m-spacing entropy approximation theoretically makes for an excellent contrast function for use in ICA, it is relatively slow to evaluate.



(a) Densities along the optimal direction found using the m-spacing method

(b) Densities along the optimal direction found using the `fastICA` method

Fig. 2. Plots showing an estimate of the density of the projected data (solid line), and the surrogate density f_0 used in the fastICA method (dotted line), for two different projections of the data. Panel 2(a) corresponds to the direction of highest entropy, found using m-spacing, and Panel 2(b) corresponds to the direction found by fastICA. These two directions are shown by the solid line and dotted line respectively in Fig. 1.

Our main theoretical contribution helps explain why fastICA in the example in Fig. 1 performs poorly. To obtain the contrast function in the fastICA method a surrogate to the true density is first obtained, and then it is approximated through several steps to increase computational speed. In Section 4 we show convergence results for the approximation steps, and conclude that the accuracy loss occurs at the initial stage where the real density is replaced by the surrogate one. This is highlighted in Fig. 2(a), which shows the estimated density (solid line) along the direction that exposes the pattern (i.e. the solid line in Fig. 1). The dotted line shows the surrogate density in this same direction that the fastICA method uses in its approximation to entropy. Fig. 2(b) shows the analogue for the direction found by fastICA (i.e. the dotted line in Fig. 1). This figure motivates this paper by highlighting the area where the approximations used in fastICA diverge from the true values. The two densities in Fig. 2(a) are very different to one-another, and this error propagates through the fastICA method to the approximation used for entropy. Note that the solid lines in Fig. 2 show the same estimated densities as given in Fig. 1.

This paper is structured as follows. In Section 2, entropy and negentropy are introduced alongside associated estimates. In Section 3 we describe the fastICA method. Section 4 contains some proofs which help to understand where errors are introduced in the fastICA method. Section 5 contains details on the example given in Fig. 1. Some concluding remarks can be found in Section 6. The code to produce the figures in this paper can be found at https://github.com/pws3141/fastICA_code.

2. Entropy and negentropy

The aim of the fastICA method is to efficiently find a projection of given data which minimises entropy. Suppose we have a one-dimensional random variable X with density $f:\mathbb{R}\to[0,\infty)$. Then the entropy H of the distribution of X is defined to be

$$H[f] := -\int_{\mathbb{R}} f(x)\log f(x)\,\mathrm{d}x, \tag{1}$$

whenever this integral exists. We use square brackets to indicate that H is a functional, taking the function f as its argument. In the special case of a Gaussian random variable with variance σ^2, the entropy can be calculated explicitly and it takes the value $\eta(\sigma^2)$ given by

$$\eta(\sigma^2) := \frac{1}{2}\big(1+\log(2\pi\sigma^2)\big). \tag{2}$$

It is known that this is an upper bound for entropy, namely the entropy of any random variable with variance σ^2 will belong to the interval $(-\infty, \eta(\sigma^2)]$ [see, for example,4]. The negentropy J is defined as

$$J[f] := \eta(\sigma^2) - H[f],$$

where $\eta(\sigma^2)$ is given by (2). This implies that $J[f]\in[0,\infty)$. Negentropy is zero when the density is Gaussian, and strictly greater than zero otherwise.

As the definition of entropy involves the integral of the density, the estimation of entropy or negentropy from data is non-trivial. For a survey of different methods to estimate entropy from data, see Beirlant et al. [3]. As an example, we consider here the m-spacing estimator, originally given in Vasicek [22]. Suppose we have a sample of one-dimensional points, $y_1, y_2, \ldots, y_n \in \mathbb{R}$, from a distribution with density f, and $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ is the ordering such that $y_{(1)} \le y_{(2)} \le$

$\cdots \leq y_{(n)}$. Define the m-spacing difference to be $\Delta_m y_i = y_{(i+m)} - y_{(i)}$ for $m \in \{3, \ldots, n-1\}$ and $i \in \{1, 2, \ldots, n-m\}$. The m-spacing approximation for entropy $H[f]$ for the sample $y = (y_1, y_2, \ldots, y_n)$ is given by

$$H_{m,n}(y) = \frac{1}{n} \sum_{i=1}^{n-m} \log\Big(\frac{n}{m}\, \Delta_m y_i\Big) - \digamma(m) + \log(m), \tag{3}$$

where $\digamma(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\Gamma(x)$ is the digamma function. This is a realisation of the general m-spacing formula given in Hall [9]. This approximation tends to the true value of entropy under certain conditions and so for a "large enough" number of points should be comparable to the true value. This method has been used previously within an ICA method by Learned-Miller and Fisher [16]. While the methods provide consistent estimates for the entropy, it is computationally expensive. The main contribution to computational cost comes from the need to sort the sample y in increasing order.

The fastICA method provides a more efficient way to estimate negentropy $J[f]$ by using a series of approximations and substitutions both for f and for $J[\cdot]$ to obtain a surrogate for negentropy $J[f]$ which is then subsequently maximised. The reason behind these substitutions is to reduce computational cost, but the drawback is that the resulting approximation may be very different from the true contrast function.

3. The fastICA algorithm

In this section we describe the fastICA method of Hyvärinen and Oja [14]. The theory behind this method was originally introduced in Hyvärinen [11], although here we adjust the notation to match the rest of this paper. We will mention explicitly where our notation differs from Hyvärinen [11] and Hyvärinen and Oja [14]. We will write 'fastICA' when we are discussing the theoretical method, and '`fastICA`' when we are discussing the `R` implementation from the `fastICA` CRAN package [17].

The fastICA method to obtain the first loading from data $\tilde{D} \in \mathbb{R}^{n\times\tilde{p}}$ follows the steps given below. Following the usual convention, the rows of $\tilde{D}$ denote observations, the columns denote variables.

(i) Whiten the data to obtain $D \in \mathbb{R}^{n\times p}$ with $p = \min(\tilde{p}, n-1)$, such that $\frac{1}{n-1} D^\top D = I_p$ [14, Section 5.2];

(ii) Iteratively find the optimal projection w^*, given by

$$w^* = \operatorname*{argmax}_{w\in\mathbb{R}^p,\, w^\top w = 1} \hat{J}^*(Dw), \tag{4}$$

where $\hat{J}^*$ is an approximation to negentropy, given in Eq. (11).

If more than one loading is required, Step (ii) is repeated for each subsequent new direction, with the added constraint that w must be orthogonal to the previously found directions. This can be implemented within the fastICA framework using Gram–Schmidt orthogonalisation [14, Section 6.2]. This is known as the deflation fastICA method. There is also a parallel fastICA method that finds all loadings concurrently, although in this paper we only consider the deflation approach.

In the literature regarding fastICA it is often the convergence of the iterative method to solve (4) that is examined. It can be shown, for example in Wei [23], that in certain situations this iterative step fails to find a good approximation for w^*. In contrast, here we consider the mathematical substitutions and approximations used in the derivation of $\hat{J}^*(Dw)$. Assumption 1 introduces the technical assumptions given in Hyvärinen [11, Sections 4 and 6], using slightly adjusted notation.

Assumption 1. Let G_i, $i \in \{1, \ldots, I\}$ be functions that do not grow faster than quadratically. Let $\varphi(\cdot)$ denote the density of a standard Gaussian random variable and assume that there are $\alpha_i, \beta_i, \gamma_i, \delta_i$, $i \in \{1, \ldots, I\}$, such that the functions

$$K_i(x) := \frac{G_i(x) + \alpha_i x^2 + \beta_i x + \gamma_i}{\delta_i} \tag{5}$$

satisfy

$$\int_{\mathbb{R}} K_i(x) K_j(x) \varphi(x)\, \mathrm{d}x = \mathbb{1}_{\{i=j\}}; \tag{6a}$$

$$\int_{\mathbb{R}} K_i(x) x^k \varphi(x)\, \mathrm{d}x = 0,\ k \in \{0, 1, 2\}, \tag{6b}$$

for $i, j \in \{1, \ldots, I\}$, where $\mathbb{1}_{\{i=j\}} = 1$ if $i = j$ and zero otherwise.

The functions G_i are given as $\bar{G}_i$ in Hyvärinen [11] and as G_i in Hyvärinen and Oja [14]. The functions K_i are described in Hyvärinen [11, Section 6] and are called G_i there.

The `fastICA` algorithm only implements the case $I = 1$. In this case, the function G_1 can be chosen nearly arbitrarily so long as it does not grow faster than quadratically: It is easy to show that for every G which is not exactly equal to a second order polynomial, a function K_1 can be found that satisfies the conditions given in (6) by choosing suitable α_1,

β_1, γ_1 and δ_1. For general $I \in \mathbb{N}$, specific G_i, $i \in \{1, \dots, I\}$ must be chosen for the conditions (6) to hold. With $I = 2$, the functions $G_1(x) = x^3$ and $G_2(x) = x^4$ are proposed in the literature [11, Section 7] and seem to be useful in practice, even though these functions violate the growth condition from Assumption 1. We have not found any examples of specific functions G_i that satisfy (6) for $I > 2$ in the fastICA literature.

Let $w \in \mathbb{R}^p$ with $\|w\| = 1$ and let $y = (y_1, y_2, \dots, y_n) = Dw \in \mathbb{R}^n$ be the data projected onto w. Since the data has been whitened, y has sample mean 0 and sample variance 1. Further, let $f\colon \mathbb{R} \to \mathbb{R}$ be the unknown density of the population-level-whitened and projected data. Then f satisfies $\int f(x)\,dx = 1$, $\int xf(x)\,dx = 0$ and $\int x^2 f(x)\,dx = 1$. We need to estimate the negentropy $J[f]$ using the data $y_1, \dots, y_n$. Define

$$c_i := \mathsf{E}_f K_i(X) = \int f(x) K_i(x)\,\mathrm{d}x \tag{7}$$

for all $i \in \{1, \dots, I\}$. For $I = 1$, setting $K(x) := K_1(x)$, $G(x) := G_1(x)$ and $c := c_1$, the derivation of the contrast function used in the fastICA method then consists of the following steps:

1. Replace f by a density f_0 given by
$$f_0(x) = A \exp\Big(\kappa x + \zeta x^2 + aK(x)\Big), \tag{8}$$
for all $x \in \mathbb{R}$. The constants A, κ, ζ and a are chosen to minimise negentropy (and hence maximise entropy) under the constraints $\int f_0(x)K(x)\,\mathrm{d}x = c$. In Proposition 3 we will show that $J[f_0] \le J[f]$.
2. Approximate f_0 by $\hat{f}_0$ defined as
$$\hat{f}_0(x) = \varphi(x)\big(1 + cK(x)\big) \tag{9}$$
for all $x \in \mathbb{R}$. In Theorem 8 we will show $J[\hat{f}_0] \approx J[f_0]$.
3. Approximate $J[\hat{f}_0]$ by second order Taylor expansion,
$$\hat{J}[\hat{f}_0] = \frac{1}{C}\big(\mathsf{E}_f G(Y) - \mathsf{E}_\varphi G(Z)\big)^2, \tag{10}$$
where Y is a random variable with density f, $Z \sim \mathcal{N}(0, 1)$, and C some constant. Note that, maybe surprisingly, Y has density f, not f_0. In Proposition 11 we will show that $\hat{J}[\hat{f}_0] \approx J[\hat{f}_0]$.
4. Use Monte-Carlo approximation for the expectations in (10), i.e. use
$$\hat{J}^*(y) = \Big(\frac{1}{n}\sum_{j=1}^{n} G(y_j) - \frac{1}{L}\sum_{j=1}^{L} G(z_j)\Big)^2, \tag{11}$$
where $z_1, \dots, z_L$ are samples from a standard Gaussian and L is large. Here $\hat{J}^*(y) \approx C\hat{J}[\hat{f}_0]$.

The restriction to $I = 1$ here removes a summation from Step 1. and Step 2. therefore simplifying Step 3. and the associated estimations in Step 4. Theoretically these steps can be completed for arbitrary $I \in \mathbb{N}$, although in this case a closed-form version equivalent to Step 3. is much more complicated.

The approximation (11) to the negentropy used in fastICA dramatically decreases the computational time needed to find ICA projections. Unlike the m-spacing estimator introduced in Section 2, the approximation $\hat{J}^*(Dw)$ is a simple Monte-Carlo estimator and does not require sorting of the data. The algorithm to solve (4) also benefits from the fact that an approximate derivative of $w \mapsto \hat{J}^*(Dw)$ can be derived analytically.

The steps in this chain of approximations are illustrated in Fig. 3 and we will investigate the approximation more formally in Section 4. In Step 1. of the procedure, we do not obtain a proper approximation, but have an inequality instead: f is replaced with a density f_0 such that $J[f_0] \le J[f]$. As a result, the w which maximises $J[f_0]$ can be very different from the one which maximises $J[f]$. In contrast, Steps 2. and 3. are proper approximations and in Section 4 we prove convergence of $\hat{f}_0$ to f_0 for Step 2. and of $\hat{J}[\hat{f}_0]$ to $J[f_0]$ for Step 3. in the limit $\|c\| \to 0$, where $c = (c_1, \dots, c_I)$. Step 4. is a simple Monte-Carlo approximation exhibiting well-understood behaviour. From the above discussion, it seems sensible to surmise that the loss of accuracy in fastICA is due to the surrogate used in Step 1. above.

We conclude this section with a few simple observations: Using (6), (5) and the fact that X and Z are standardised we find

$$\begin{aligned} c_i = \mathsf{E}_f K_i(X) = \mathsf{E}_f K_i(X) - \mathsf{E}_\varphi K_i(Z) &= \mathsf{E}_f\left(\frac{G_i(X) + \alpha_i X^2 + \beta_i X + \gamma_i}{\delta_i}\right) - \mathsf{E}_\varphi\left(\frac{G_i(Z) + \alpha_i Z^2 + \beta_i Z + \gamma_i}{\delta_i}\right) \\ &= \frac{\mathsf{E}_f G_i(X) + \alpha_i 1 + \beta_i 0 + \gamma_i}{\delta_i} - \frac{\mathsf{E}_\varphi G_i(Z) + \alpha_i 1 + \beta_i 0 + \gamma_i}{\delta_i} = \frac{\mathsf{E}_f G_i(X) - \mathsf{E}_\varphi G_i(Z)}{\delta_i}. \end{aligned}$$

Thus, the fastICA objective function (ignoring the final Monte Carlo approximation) satisfies $\hat{J}[\hat{f}_0] \propto c^2$ for the case $I = 1$, considered above, and $\hat{J}[\hat{f}_0] \propto \sum_{i=1}^{I} c_i^2$ in the general case. Thus, fastICA can only see the data through the c_i. If the data are approximately Gaussian, we have $\mathsf{E}_f G_i(X) \approx \mathsf{E}_\varphi G_i(Z)$ and $c_i \approx 0$ for all i and thus $\hat{J}[\hat{f}_0] \approx 0$, but the opposite implication does not hold. This is in contrast to the true negentropy, which satisfies $J[f] = 0$ if and only if f is Gaussian.

$$\begin{array}{ccccccccc} f & & f_0 & \approx & \hat{f}_0 & & & & \\ \downarrow & & \downarrow & & \downarrow & & & & \\ J[f] & \geq & J[f_0] & \approx & J[\hat{f}_0] & \approx & \hat{J}[\hat{f}_0] & \approx & \frac{1}{C}\hat{J}^*(y) \\ & & & & & & \| & & \| \\ & & & & & & \frac{1}{C}(\mathsf{E}_f G(Y) - \mathsf{E}_\varphi G(Z))^2 & & \frac{1}{C}(\frac{1}{n}\sum_{j=1}^n G(y_j) - \frac{1}{L}\sum_{j=1}^L G(z_j))^2 \end{array}$$

Fig. 3. Approximations used in fastICA: The fastICA contrast function $\hat{J}^*(y)$ is used in place of negentropy $J[f]$. Note that the first step involves an inequality rather than an approximation.

A first consequence of this argument is that projections where the true distribution is Gaussian will look 'uninteresting' to fastICA: for these directions w the objective function $\hat{J}^*(Dw)$ will be small and the search for the maximum in (4) will be driven away from these directions. This is particularly relevant since for high dimensional data, where the search volume is vast, projections along most directions are close to Gaussian [5,25], so fastICA will be able to exclude much of the search volume. Conversely, if $\hat{J}[\hat{f}_0]$ and thus $\|c\|$ is large, the projected density f is not Gaussian and by maximising (an approximation to) $\hat{J}[\hat{f}_0]$, the fastICA method will find directions which are 'interesting'. But the above discussion also shows that optima can be missed when $\hat{J}[\hat{f}_0]$ is small, but the projected density f is still far from Gaussian. This is the case we are concerned with in this paper and thus we assume $\|c\| \approx 0$ when we consider the fastICA approximations in detail in the next section.

4. Approximations used in the fastICA method

In this section, we investigate the validity of the approximation given in Section 3. We consider Step 1. in Proposition 3, Step 2. in Theorem 8, and Step 3. in Proposition 11. Throughout this section, we consider arbitrary $I \in \mathbb{N}$ for completeness.

We first introduce some assumptions, in addition to Assumption 1, that are required for the mathematics in this section to hold.

Assumption 2. There exists $\varepsilon > 0$ such that for all $h \in \mathbb{R}^I$ with $h^\top h < \varepsilon$, we have

$$h^\top K(x) \geq -\frac{1}{2} \tag{12}$$

for all $x \in \mathbb{R}$, where $K(x) = \big(K_1(x), K_2(x), \ldots, K_I(x)\big)$. In addition, there exists a function $M\colon \mathbb{R} \to \mathbb{R}$ such that

$$\sum_{i=1}^I \sum_{j=1}^I \sum_{k=1}^I |K_i(x)K_j(x)K_k(x)| \leq M(x) \text{ for all } x \in \mathbb{R}, \tag{13a}$$

$$\int_{\mathbb{R}} \varphi(x)M(x)\,\mathrm{d}x =: \tilde{M} < \infty. \tag{13b}$$

Note that under the condition that each G_i does not grow faster than quadratically (given in Assumption 1), we can always find some positive constants B_i, $i \in \{1, \ldots, I\}$ such that

$$|K_i(x)| \leq B_i(1 + x^2), \tag{14}$$

for all $x \in \mathbb{R}$. Note also that for $I = 1$ the condition given by (12) that there exists an $\varepsilon > 0$ such that for all $h \in [0, \varepsilon)$, we have $hK(x) \geq -1/2$ is satisfied as follows. Let $\alpha, \beta, \gamma, \delta$ be parameters for which (6) holds. Then, (6) holds also for $\alpha, \beta, \gamma, -\delta$. Moreover, since G does not grow faster than quadratically, αx^2 is the dominant term in $K(x)$ as $x \to \pm\infty$. Therefore, to ensure that (12) holds it is enough to choose δ or $-\delta$ such that the sign is the same as that of α.

4.1. Step 1

We start our discussion by considering Step 1. of the approximations described in Section 3. We prove that the distribution which maximises entropy for given values of $c_1, \ldots, c_I$ is indeed of the form (8) and thus that we indeed have $J[f_0] \leq J[f]$.

Proposition 3. *Let f be the density of the population-level-whitened data projected in some direction (thus with zero mean and unit variance). Recall c_i is defined by (7). The density f_0 that maximises entropy in the set*

$$\left\{g\colon \mathbb{R} \to \mathbb{R}\,;\ g \text{ is a density function, and } \int_{\mathbb{R}} g(x)K_i(x)\,dx = c_i,\ i \in \{1, \ldots, I\}\right\},$$

is given by,

$$f_0(x) = A\exp\left(\kappa x + \zeta x^2 + \sum_{i=1}^{I} a_i K_i(x)\right) \tag{15}$$

for some constants κ, ζ, A and a_i, $i \in \{1,\dots,I\}$ that depend on c_i, $i \in \{1,\dots,I\}$. It follows from this that $J[f_0] \le J[f]$.

Proof. We use the method of Lagrange multipliers in the calculus of variations [see, for example,15] to find a necessary condition for the density that maximises entropy given the constraints on mean and variance, and in (7). Let $F[\cdot]\colon C^2 \to \mathbb{R}$ be a functional of the function $g\colon \mathbb{R}\to\mathbb{R}$, with $g \in C^2$, where C^2 is the set of all twice continuously differentiable functions. Then, the functional derivative $\delta F/\delta g\colon \mathbb{R}\to\mathbb{R}$ is explicitly defined by

$$\int_{\mathbb{R}} \frac{\delta F}{\delta g}(x)\phi(x)\,\mathrm{d}x := \frac{\mathrm{d}}{\mathrm{d}\varepsilon}F[g+\varepsilon\phi]\Big|_{\varepsilon=0} = \lim_{\varepsilon\downarrow 0}\Big(\frac{F[g+\varepsilon\phi]-F[g]}{\varepsilon}\Big), \tag{16}$$

for any function $\phi \in C^2$. The right-hand side of (16) is known as the Gâteaux differential $\mathrm{d}F(g;\phi)$. Define the inner product of two functions by $\langle g,h\rangle := \int_{\mathbb{R}} g(x)h(x)\,\mathrm{d}x$, with norm $\|g\|_{L^2} := \langle g,g\rangle^{\frac{1}{2}} = \big(\int_{\mathbb{R}} g(x)^2\,\mathrm{d}x\big)^{\frac{1}{2}}$. We want to solve the following system of equations

$$\begin{cases} U[g](x) & := \frac{\delta}{\delta g}H[g] + \lambda_1\frac{\delta}{\delta g}V[g] + \lambda_2\frac{\delta}{\delta g}P[g] + \lambda_3\frac{\delta}{\delta g}Q[g] + \sum_{i=1}^{I}\nu_i\frac{\delta}{\delta g}R_i[g] = 0;\\ V[g] & = 0;\\ P[g] & = 0;\\ Q[g] & = 0;\\ R_i[g] & = 0, \end{cases}$$

where $\lambda_1,\lambda_2,\lambda_3,\nu_i$ are some real numbers, $i\in\{1,\dots,I\}$, $H[g]$ is entropy as given in (1), and

$$V[g] := \mathrm{Var}[g] - 1 = \int_{\mathbb{R}} g(x)x^2\,\mathrm{d}x - \Big(\int_{\mathbb{R}} g(x)x\,\mathrm{d}x\Big)^2 - 1; \quad P[g] := \int_{\mathbb{R}} g(x)\,\mathrm{d}x - 1;$$

$$Q[g] := \int_{\mathbb{R}} g(x)x\,\mathrm{d}x; \quad R_i[g] := \int_{\mathbb{R}} g(x)K_i(x)\,\mathrm{d}x - c_i.$$

Using (1) and (16) the term with H gives,

$$\begin{aligned}\langle\frac{\delta H}{\delta g},\phi\rangle &= -\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\int \big(g(x)+\varepsilon\phi(x)\big)\log\big(g(x)+\varepsilon\phi(x)\big)\,\mathrm{d}x\Big|_{\varepsilon=0}\\ &= -\int\Big(g(x)\frac{\phi(x)}{g(x)+\varepsilon\phi(x)} + \phi(x)\log\big(g(x)+\varepsilon\phi(x)\big) + \varepsilon\phi(x)\frac{\phi(x)}{g(x)+\varepsilon\phi(x)}\Big)\,\mathrm{d}x\Big|_{\varepsilon=0}\\ &= -\int\big(1+\log g(x)\big)\phi(x)\,\mathrm{d}x = \langle -1-\log g(x),\phi\rangle.\end{aligned}$$

Now, looking at $V[g]$ and using the constraint $Q[g]=0$ we get,

$$\begin{aligned}\langle\frac{\delta V}{\delta g},\phi\rangle &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big(\int\big(g(x)+\varepsilon\phi(x)\big)x^2\,\mathrm{d}x - \Big(\int\big(g(x)+\varepsilon\phi(x)\big)x\,\mathrm{d}x\Big)^2 - 1\Big)\Big|_{\varepsilon=0}\\ &= \int\phi(x)x^2\,\mathrm{d}x - 2\Big(\int\phi(x)x\,\mathrm{d}x\cdot\int g(x)x\,\mathrm{d}x\Big) = \langle x^2,\phi\rangle - 2\langle x,\phi\rangle\cdot Q[g] = \langle x^2,\phi\rangle.\end{aligned}$$

Let $L[\cdot]\colon C^2\to\mathbb{R}$ be of the form $L[g]=\int g(x)l(x)\,\mathrm{d}x - k$ for some function $l\colon\mathbb{R}\to\mathbb{R}$, and some constant $k\in\mathbb{R}$. Then it is easy to check that $\langle\frac{\delta L}{\delta g},\phi\rangle = \langle l,\phi\rangle$ and therefore $\frac{\delta P}{\delta g}=1$, $\frac{\delta Q}{\delta g}=x$ and $\frac{\delta R_i}{\delta g}=K_i$. Putting this into the equation for $U[g]$, we have

$$U[g](x) = -1-\log g(x) + \lambda_1 + \lambda_2 x^2 + \lambda_3 x + \sum_{i=1}^{I}\nu_i K_i(x).$$

Setting $U[g]=0$ and solving for g gives, $g(x)=f_0(x)=\exp[\lambda_1 - 1 + \lambda_2 x^2 + \lambda_3 x + \sum_{i=1}^{I}\nu_i K_i(x)]$ which is (15) with $A=\exp(\lambda_1-1)$, $\kappa=\lambda_3$, $\zeta=\lambda_2$ and $a_i=\nu_i$, $i=1,\dots,I$. Note that the constants A, κ, ζ, and a_i depend on c_i indirectly through the constraints on the K_i expressed as $R_i[g]=0$. □

Remark 4. It is possible to specify a density f such that in some limit, $H[f]\to\infty$ whilst $H[f_0]$ remains bounded and thus $\big|J[f]-J[f_0]\big|\to\infty$, with f_0 the density given in (8). That is, in Step 1. of the fastICA method given in Section 3,

the difference between the true negentropy and the surrogate negentropy can be arbitrarily large. For example, set the density f to be a mixture of two independent uniform densities, i.e.

$$f(x) = \frac{1}{2}\big(g(x;\ -1-\varepsilon, -1) + g(x;\ 1, 1+\varepsilon)\big)$$

where $\varepsilon \in \mathbb{R}$ and $g(\cdot\,;\ a, b)$ is the density function of a Uniform distribution in the interval $[a, b]$. Then we have expectation and variance given by

$$\mathsf{E}_f X = 0; \quad \mathrm{Var}_f X = 1 + \varepsilon + \frac{\varepsilon^2}{3}.$$

As the support of $g(\cdot\,;\ -1-\varepsilon, -1)$ is disjoint from that of $g(\cdot\,;\ 1, 1+\varepsilon)$, the entropy is given by,

$$H[f] = \frac{1}{2}\big(H[g(\cdot\,;\ -1-\varepsilon, -1)] + H[g(\cdot\,;\ 1, 1+\varepsilon)]\big) - \log(2).$$

We have $H[f] \to -\infty$ as $\varepsilon \to 0$, since f tends to a pair of Dirac deltas. Also,

$$\mathsf{E}_f K_i(x) =: c_i \to \frac{1}{2}\big(K_i(-1) + K_i(1)\big), \tag{17}$$

as $\varepsilon \to 0$. With f_0 as in (8),

$$c_i = \int K_i(x) f_0(x)\,\mathrm{d}x, \tag{18}$$

and,

$$\begin{aligned} H[f_0] &= \int f_0(x)\log(A)\,\mathrm{d}x + \int f_0(x)\big(\eta x + \kappa x^2 + \sum_{i=1}^{I} a_i K_i(x)\big)\,\mathrm{d}x \\ &= \log(A) + \eta \mathsf{E}_{f_0} X + \kappa \mathsf{E}_{f_0} X^2 + \sum_{i=1}^{I} a_i \mathsf{E}_{f_0} K_i(x) = \log(A) + \kappa + \sum_{i=1}^{I} a_i c_i. \end{aligned}$$

Therefore, for $H[f_0]$ to be unbounded from below as $\varepsilon \to 0$ we would require some $\kappa \to -\infty$, $a_i \to -\infty$ or $A \to 0$, as c_i is bounded by (17) and Assumption 2. However, this cannot occur whilst f_0 satisfies (18).

4.2. Step 2

We now switch our attention to Step 2. of the approximations. As discussed in Section 3, we consider the case where $c \to 0$. The first step of our analysis is to identify the behaviour of the constants in the definition of f_0 as $c \to 0$. We then prove some auxiliary results before concluding our discussion of Step 2. in Theorem 8.

Proposition 5. *Suppose Assumption 1 is satisfied, and let $A, \kappa, \zeta, a_1, \ldots, a_I$ be defined as in Proposition 3, as functions of c. Then*

$$A - \frac{1}{\sqrt{2\pi}} = \mathcal{O}(\|c\|^2),\ \kappa = \mathcal{O}(\|c\|^2),\ \zeta + \frac{1}{2} = \mathcal{O}(\|c\|^2),\ a_i - c_i = \mathcal{O}(\|c\|^2), \quad i \in \{1, \ldots, I\},$$

as $\|c\| \to 0$.

Proof. Define $x = (c_1, \ldots, c_I)^\top \in \mathbb{R}^I$ and $y = (A, \kappa, \zeta, a_1, \ldots, a_I)^\top \in \mathbb{R}^{I+3}$. Furthermore, let $F\colon \mathbb{R}^I \times \mathbb{R}^{I+3} \to \mathbb{R}^{I+3}$ be given by

$$F(x, y) = \begin{pmatrix} \int f_0(x)\,\mathrm{d}x - 1 \\ \int f_0(x) x\,\mathrm{d}x \\ \int f_0(x) x^2\,\mathrm{d}x - 1 \\ \int f_0(x) K_1(x)\,\mathrm{d}x - c_1 \\ \vdots \\ \int f_0(x) K_I(x)\,\mathrm{d}x - c_I \end{pmatrix},$$

where f_0 is given in (15) and K_i in (5). Then, for the points $x_1 = (0, \ldots, 0)^\top$ and $y_1 = (\frac{1}{\sqrt{2\pi}}, 0, -\frac{1}{2}, 0, \ldots, 0)^\top$, we have $F(x_1, y_1) = 0$.

Assuming F is twice differentiable, we use the Implicit Function Theorem around (x_1, y_1). First, we need to show $D_yF(x_1, y_1)$ is invertible. We have

$$D_yF(x_1, y_1) = \begin{pmatrix} M & 0 \\ 0 & -I_I \end{pmatrix}, \quad M = \begin{pmatrix} \sqrt{2} & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 4 \end{pmatrix}.$$

Therefore, $D_yF(x_1, y_1)$ is non-singular, and so the Implicit Function Theorem holds. There exist some open set $\mathcal{U} \subset \mathbb{R}^I$ and a unique continuously differentiable function $g: \mathcal{U} \to \mathbb{R}^{I+3}$ such that $g(x_1) = y_1$ and $F\big(x,\ g(x)\big) = 0$ for all $x \in \mathcal{U}$. Then,

$$Dg(x) = -D_yF\big(x,\ g(x)\big)^{-1} D_xF\big(x,\ g(x)\big). \tag{19}$$

As g is continuous in the set $\mathcal{U}$, there exists some $\varepsilon > 0$, such that for all $c \in \mathcal{U}$ with $\|c\| < \varepsilon$, $g(x_1 + c) = y_1 + d$ for some $d \in \mathbb{R}^{I+3}$. Using Taylor series we can expand g around $x_1 = 0 \in \mathbb{R}^I$ to obtain $g(x_1 + c) = g(x_1) + Dg(x_1)\, c + \mathcal{O}(\|c\|^2)$, and

$$Dg(x_1) = \frac{d + \mathcal{O}(\|c\|^2)}{c}.$$

Putting this together with (19) at $x = x_1$ and rearranging gives,

$$d = -D_yF(x_1, y_1)^{-1} D_xF(x_1, y_1)\, c + \mathcal{O}(\|c\|^2).$$

Now, since

$$D_xF(x_1, y_1) = \begin{pmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ & I_I & \end{pmatrix} \in \mathbb{R}^{(I+3)\times I},$$

one easily obtains that

$$d = \begin{pmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ & I_I & \end{pmatrix} c + \mathcal{O}(\|c\|^2),$$

and so,

$$y_1 + d = \begin{pmatrix} \frac{1}{\sqrt{2\pi}} & 0 & -\frac{1}{2} & c_1 & \cdots & c_I \end{pmatrix}^\top + \mathcal{O}(\|c\|^2), \text{ as } c \to 0.$$

This completes the proof. □

We now define the following functions $y(\cdot)$ and $r(\cdot)$ for future use. Let $y: \mathbb{R} \to \mathbb{R}$ be given by

$$y(x) := \kappa x + (\zeta + \frac{1}{2})x^2 + \sum_{i=1}^{I} a_i K_i(x), \tag{20}$$

and $r: \mathbb{R} \to \mathbb{R}$ given by

$$r(x) := e^x - 1 - x. \tag{21}$$

Using these definitions, we can write f_0, given in Proposition 3, as

$$f_0(x) = \varphi(x) \cdot \sqrt{2\pi} A e^{y(x)}. \tag{22}$$

The following lemmas are two technical results needed in the proof of Theorem 8.

Lemma 6. *Let $g: \mathbb{R} \to \mathbb{R}$ and $l: \mathbb{R} \to \mathbb{R}$ be any functions and $h: \mathbb{R} \to \mathbb{R}_+$ be convex with $h(0) = 0$. Then,*

$$\sup_{x\in\mathbb{R}} \big|l(x)h(\varepsilon g(x))\big| \le \varepsilon \sup_{x\in\mathbb{R}} \big|l(x)h(g(x))\big|$$

for all $\varepsilon \in [0, 1]$.

Proof. As h is convex, for all $\lambda \in [0, 1]$ and for all $x, y \in \mathbb{R}$, we have $h\big(\lambda x + (1-\lambda)y\big) \le \lambda h(x) + (1-\lambda)h(y)$. Let $\varepsilon \in [0, 1]$. Then, substituting $\lambda = \varepsilon$, $x = g(x)$ and $y = 0$, we have $h\big(\varepsilon\, g(x)\big) \le \varepsilon\, h\big(g(x)\big)$, for all $g(x) \in \mathbb{R}$, as $h(0) = 0$. Noticing that h maps to the positive real line allows to conclude. □

Lemma 7. *Let $r: \mathbb{R} \to \mathbb{R}_+$ be given as in (21). Then,*

$$r(\varepsilon\, y) \le \varepsilon^2 r(y), \text{ for all } y \ge 0, \text{ and for all } \varepsilon \in [0, 1]. \tag{23}$$

Moreover, for any function $l\colon \mathbb{R} \to \mathbb{R}$, *we have*

$$\sup_{x\in\mathbb{R}} \left| l(x) r\big(\varepsilon(1+x^2)\big) \right| \leq \varepsilon^2 \sup_{x\in\mathbb{R}} \left| l(x) r(1+x^2) \right|.$$

Proof. We will use the Taylor expansion of the exponential around 0 for both the left-hand and right-hand side of (23). The left-hand side gives,

$$r(\varepsilon y) = \exp(\varepsilon y) - 1 - \varepsilon y = \sum_{n=0}^{\infty} \frac{\varepsilon^n}{n!} y^n - 1 - \varepsilon y = \varepsilon^2 \Big(\sum_{n=2}^{\infty} \frac{\varepsilon^{n-2}}{n!} y^n \Big), \quad \text{absolutely convergent for all } \varepsilon y \in \mathbb{R}$$

and the right-hand side of (23) gives,

$$\varepsilon^2 r(y) = \varepsilon^2 \Big(\sum_{n=0}^{\infty} \frac{1}{n!} y^n - 1 - y \Big) = \varepsilon^2 \Big(\sum_{n=2}^{\infty} \frac{1}{n!} y^n \Big).$$

Putting these two results together,

$$r(\varepsilon y) - \varepsilon^2 r(y) = \varepsilon^2 \Big(\sum_{n=2}^{\infty} \frac{1}{n!} y^n (\varepsilon^{n-2} - 1) \Big) \leq 0,$$

as $\varepsilon^n - 1 \leq 0$ for all $\varepsilon \in [0,1]$ and $n \in \mathbb{N}_+$. This proves (23).

Let $l\colon \mathbb{R} \to \mathbb{R}$ be some function. Then, as r maps to the positive real line and using (23) with $y = 1 + x^2$, we have $|l(x) r\big(\varepsilon(1+x^2)\big)| \leq \varepsilon^2 |l(x) r(1+x^2)|$, for all $x \in \mathbb{R}$. Taking the supremum over the real line we conclude. □

We now consider the error term between the density f_0 that maximises entropy, and its estimate $\hat{f}_0$.

Theorem 8. *Suppose we have functions* K_i, $i \in \{1, \dots, I\}$ *that satisfy Assumptions 1 and 2. Let* f_0 *be given as in Proposition 3, and let* $\hat{f}_0$ *be given by*

$$\hat{f}_0(x) = \varphi(x) \Big(1 + \sum_{i=1}^{I} c_i K_i(x) \Big).$$

Then for all $\delta < 1/2$

$$\sup_{x\in\mathbb{R}} \left| e^{\delta x^2} \big(f_0(x) - \hat{f}_0(x) \big) \right| = \mathcal{O}(\|c\|^2) \text{ as } c \to 0.$$

Proof. Let $\varphi(x) = (2\pi)^{1/2} e^{-x^2/2}$ be the density of a standard Gaussian random variable and let the function $g\colon \mathbb{R} \to \mathbb{R}$ be defined by

$$g(x) := \frac{f_0(x) - \hat{f}_0(x)}{\varphi(x)}.$$

Then, with $y\colon \mathbb{R} \to \mathbb{R}$ as defined in (20) and using (22) we get,

$$\begin{aligned} g(x) &= \sqrt{2\pi} A \exp\big(y(x)\big) - \big(1 + \sum_{i=1}^{I} c_i K_i(x)\big) \\ &= \sqrt{2\pi} A \Big(\exp\big(y(x)\big) - 1 - y(x) \Big) + \sqrt{2\pi} A \big(1 + y(x)\big) - (1 + \sum_{i=1}^{I} c_i K_i(x)) + \sqrt{2\pi} A \Big(\sum_{i=1}^{I} c_i K_i(x) - \sum_{i=1}^{I} c_i K_i(x) \Big). \end{aligned}$$

Rearranging this using the function $r\colon \mathbb{R} \to \mathbb{R}$ given in (21) and by expanding $y(x)$ gives,

$$g(x) = \sqrt{2\pi} A \cdot r\big(y(x)\big) + \sqrt{2\pi} A \cdot \Big(\kappa x + \big(\zeta + \frac{1}{2}\big) x^2 \Big) + \big(\sqrt{2\pi} A - 1\big) \sum_{i=1}^{I} c_i K_i(x) + \sqrt{2\pi} A \sum_{i=1}^{I} (a_i - c_i) K_i(x) + (\sqrt{2\pi} A - 1).$$

Note that the absolute value of $g(x)$ can be bounded by the following terms,

$$\begin{aligned} |g(x)| \leq{}& \sqrt{2\pi} A \, |r\big(y(x)\big)| + \sqrt{2\pi} A \, |\kappa x| + \sqrt{2\pi} A \, |\zeta + \frac{1}{2}| x^2 \\ &+ \sqrt{2\pi} A \, \Big| \sum_{i=1}^{I} (a_i - c_i) K_i(x) \Big| + |\sqrt{2\pi} A - 1| \, \Big| \sum_{i=1}^{I} c_i K_i(x) \Big| + |\sqrt{2\pi} A - 1|. \end{aligned}$$

We have,

$$|f_0(x)-\hat{f}_0(x)| = |\varphi(x)\cdot g(x)| = \varphi(x)\Big|\sqrt{2\pi}Ar\big(y(x)\big) + \sqrt{2\pi}A\Big(\kappa x + \big(\zeta+\frac{1}{2}\big)x^2\Big) + (\sqrt{2\pi}A-1)\sum_{i=1}^{I} c_iK_i(x) + \sqrt{2\pi}A\sum_{i=1}^{I}(a_i-c_i)K_i(x) + (\sqrt{2\pi}A-1)\Big|.$$

We now multiply both sides by $e^{\delta x^2}$ and setting $\tilde{\delta} = \frac{1}{2}-\delta$, so that $e^{\delta x^2}\varphi(x) = (2\pi)^{-1/2}e^{-\tilde{\delta}x^2}$, we have

$$\begin{aligned}\big|e^{\delta x^2}\big(f_0(x)-\hat{f}_0(x)\big)\big| &= (2\pi)^{-1/2}e^{-\tilde{\delta}x^2}\Big|\sqrt{2\pi}Ar\big(y(x)\big) + \sqrt{2\pi}A\Big(\kappa x + \big(\zeta+\frac{1}{2}\big)x^2\Big) + (\sqrt{2\pi}A-1)\sum_{i=1}^{I} c_iK_i(x)\\ &\quad + \sqrt{2\pi}A\sum_{i=1}^{I}(a_i-c_i)K_i(x) + (\sqrt{2\pi}A-1)\Big| \le T_1(x)+T_2(x)\\ &\quad + \frac{1}{\sqrt{2\pi}}\cdot T_3(x) + T_4(x) + \frac{1}{\sqrt{2\pi}}\cdot T_5(x), \end{aligned} \tag{24}$$

where,

$$T_1(x) := \big|Ae^{-\tilde{\delta}x^2}r\big(y(x)\big)\big|; \quad T_2(x) := \big|Ae^{-\tilde{\delta}x^2}\big(\kappa x + (\zeta+\frac{1}{2})x^2\big)\big|;$$
$$T_3(x) := \big|(\sqrt{2\pi}A-1)e^{-\tilde{\delta}x^2}\sum_{i=1}^{I} c_iK_i(x)\big|; \quad T_4(x) := \big|Ae^{-\tilde{\delta}x^2}\sum_{i=1}^{I}(a_i-c_i)K_i(x)\big|;$$
$$T_5(x) := |e^{-\tilde{\delta}x^2}(\sqrt{2\pi}A-1)|.$$

If we show that $\|T_i\|_\infty$ is at least of order $\|c\|^2$ as $c\to 0$ for $i\in\{1,\dots,5\}$, then we can conclude the proof by taking the supremum of (24) over $x\in\mathbb{R}$, which gives,

$$\sup_{x\in\mathbb{R}}\big|e^{\delta x^2}(f_0(x)-\hat{f}_0(x))\big| = \mathcal{O}(\|c\|^2),$$

as $c\to 0$.

For T_1 first note that

$$\big|e^{-\tilde{\delta}x^2}r(y(x))\big| \le \max_{\sigma\in\{-1,1\}}\big|e^{-\tilde{\delta}x^2}r(\sigma\cdot|y(x)|)\big|, \text{ for all } x\in\mathbb{R},$$

and thus,

$$\sup_{x\in\mathbb{R}}|T_1(x)| \le A\cdot \sup_{\substack{x\in\mathbb{R}\\ \sigma\in\{-1,1\}}}|e^{-\tilde{\delta}x^2}r(\sigma\cdot|y(x)|)|. \tag{25}$$

Next we choose γ such that,

$$q_1 := \sup_{\substack{x\in\mathbb{R}\\ \sigma\in\{-1,1\}}}\big|e^{-\tilde{\delta}x^2}r\big(\sigma\cdot\gamma(1+x^2)\big)\big| < \infty. \tag{26}$$

This is always possible for some $\gamma\in(-\tilde{\delta},\tilde{\delta})$, as $r(0)=0$, and since $e^{-\tilde{\delta}x^2}r\big(\pm\gamma(1+x^2)\big)$ is continuous and $r\big(\pm\gamma(1+x^2)\big)$ grows no faster that $e^{\gamma x^2}$ as $x\to\pm\infty$, it is beaten by $e^{-\tilde{\delta}x^2}$ in the tails.

For $y(x)$ as given in (20) and using (14) we can find an upper bound by

$$\begin{aligned}|y(x)| &\le |\kappa|\cdot\Big(\frac{1+x^2}{2}\Big) + |\zeta+\frac{1}{2}|\cdot(1+x^2) + \sum_{i=1}^{I}|a_i|B_i(1+x^2) = \gamma(1+x^2)\cdot\frac{1}{\gamma}\Big(\frac{1}{2}|\kappa| + |\zeta+\frac{1}{2}| + \sum_{i=1}^{I}|a_i|B_i\Big)\\ &=: \gamma(1+x^2)\cdot\varepsilon_1, \end{aligned} \tag{27}$$

where γ is such that (26) holds. As $c\to 0$, we have by Proposition 5, $\kappa\to 0$, $\zeta\to -1/2$ and $a_i\to c_i$. Therefore, we can choose c small enough (and depending on γ) such that $\varepsilon_1\in[0,1]$. Now, from (25), (27), the fact that r is convex with a minimum at zero, and by Lemma 7 we get

$$\sup_{x\in\mathbb{R}}\big|T_1(x)\big| \le A\sup_{\substack{x\in\mathbb{R}\\ \sigma\in\{-1,1\}}}\big|e^{-\tilde{\delta}x^2}r\big(\sigma\gamma(1+x^2)\varepsilon_1\big)\big| \le \varepsilon_1^2 A\sup_{\substack{x\in\mathbb{R}\\ \sigma\in\{-1,1\}}}\big|e^{-\tilde{\delta}x^2}r\big(\sigma\gamma(1+x^2)\big)\big| = \varepsilon_1^2 A\, q_1.$$

By Proposition 5, we have $A \to 1/\sqrt{2\pi}$ as $c \to 0$ and,

$$\varepsilon_1 = \frac{1}{\gamma}\Big(\frac{1}{2}|\kappa| + |\zeta + \frac{1}{2}| + \sum_{i=1}^{I}|a_i|B_i\Big) = \mathcal{O}(\|c\|), \text{ as } c \to 0,$$

and therefore $\varepsilon_1^2 = \mathcal{O}(\|c\|^2)$ as $c \to 0$, and $\|T_1\|_\infty = \mathcal{O}(\|c\|^2)$ as $c \to 0$.

For T_2 we proceed similarly as for T_1, and look for some $\varepsilon_2 \in [0, 1]$ such that $|\kappa x + (\zeta + \frac{1}{2})x^2| \le \varepsilon_2(1 + x^2)$. We have,

$$\big|\kappa x + (\zeta + \frac{1}{2})x^2\big| \le |\kappa|(\frac{1+x^2}{2}) + |\zeta + \frac{1}{2}|(1+x^2) = \big(\frac{1}{2}|\kappa| + |\zeta + \frac{1}{2}|\big)(1+x^2).$$

Setting $\varepsilon_2 := (\frac{1}{2}|\kappa| + |\zeta + \frac{1}{2}|)$, by Proposition 5, $\varepsilon_2 = \mathcal{O}(\|c\|^2)$ as $c \to 0$, and thus we can choose c sufficiently small such that $\varepsilon_2 \le 1$. Let,

$$q_2 := \sup_{x\in\mathbb{R}} \big|e^{-\tilde{\delta}x^2}(1+x^2)\big| < \infty,$$

where $q_2 < \infty$ since $e^{-\tilde{\delta}x^2}(1+x^2)$ is continuous and tends to zero in the tails. From this, for $\varepsilon_2 \in [0, 1]$ as above, we can apply Lemma 6 and get

$$\sup_{x\in\mathbb{R}} \big|e^{-\tilde{\delta}x^2}(\kappa x + (\zeta + \frac{1}{2})x^2)\big| \le A \sup_{x\in\mathbb{R}} \big|e^{\tilde{\delta}x^2}\varepsilon_2(1+x^2)\big| \le \varepsilon_2 \sup_{x\in\mathbb{R}} \big|e^{-\tilde{\delta}x^2}(1+x^2)\big| = \varepsilon_2\, q_2.$$

Then,

$$\sup_{x\in\mathbb{R}} \big|T_2(x)\big| = A \sup_{x\in\mathbb{R}} \big|e^{\tilde{\delta}x^2}(\kappa x + (\zeta + \frac{1}{2})x^2)\big| \le A\,\varepsilon_2\, q_2.$$

Therefore, we have $\|T_2\|_\infty = \mathcal{O}(\|c\|^2)$, as $c \to 0$.

For T_3, as with the T_2 term we want an $\varepsilon_3 \in [0, 1]$, such that $|\sum_{i=1}^{I} c_i K_i(x)| \le \varepsilon_3(1+x^2)$, so that we can apply Lemma 6 to show

$$\sup_{x\in\mathbb{R}} \big|e^{-\tilde{\delta}x^2}\sum_{i=1}^{I} c_i K_i(x)\big| \le \varepsilon_3 \sup_{x\in\mathbb{R}} \big|e^{-\tilde{\delta}x^2}(1+x^2)\big| < \infty.$$

First, note that by (14),

$$\big|\sum_{i=1}^{I} c_i K_i(x)\big| \le \big|\sum_{i=1}^{I} c_i B_i(1+x^2)\big| = \big|\sum_{i=1}^{I} c_i B_i\big| \cdot (1+x^2),$$

and thus we set $\varepsilon_3 := \big|\sum_{i=1}^{I} c_i B_i\big|$. Clearly, $\varepsilon_3 = \mathcal{O}(\|c\|)$ as $c \to 0$. Now, with c sufficiently small such that $\varepsilon_3 \in [0, 1]$, we have by Lemma 6,

$$\sup_{x\in\mathbb{R}} \big|e^{-\tilde{\delta}x^2}\sum_{i=1}^{I} c_i K_i(x)\big| \le \sup_{x\in\mathbb{R}} \big|e^{-\tilde{\delta}x^2}\big(|\sum_{i=1}^{I} c_i B_i|\big)(1+x^2)\big| \le |\sum_{i=1}^{I} c_i B_i| \cdot \sup_{x\in\mathbb{R}} \big|e^{-\tilde{\delta}x^2}(1+x^2)\big| \le \varepsilon_3\, q_2.$$

Therefore,

$$\sup_{x\in\mathbb{R}} \big|T_3(x)\big| \le |\sqrt{2\pi}A - 1|\varepsilon_3\, q_2$$

Thus, $\|T_3\|_\infty = \mathcal{O}(\|c\|^3)$, as $c \to 0$, since $|\sqrt{2\pi}A - 1| = \mathcal{O}(\|c\|^2)$ and $\varepsilon_3 = \mathcal{O}(\|c\|)$ as $c \to 0$.

Similar to the T_2 and T_3 terms, for T_4 we want an $\varepsilon_4 \in [0, 1]$ such that $\sum_{i=1}^{I}(a_i - c_i)K_i(x) \le \varepsilon_4(1+x^2)$. Note that

$$\big|\sum_{i=1}^{I}(a_i - c_i)K_i(x)\big| \le \big|\sum_{i=1}^{I}(a_i - c_i)B_i\big| \cdot (1+x^2),$$

by (14) and thus we set $\varepsilon_4 := |\sum_{i=1}^{I}(a_i - c_i)B_i|$, and by Proposition 5, $\varepsilon_4 = \mathcal{O}(\|c\|^2)$ as $c \to 0$. Choose c small enough such that $\varepsilon_4 \in [0, 1]$. Then, by Lemma 6,

$$\sup_{x\in\mathbb{R}}|T_4(x)| \le A \sup_{x\in\mathbb{R}} \big|e^{-\tilde{\delta}x^2}\sum_{i=1}^{I}(a_i - c_i)B_i(1+x^2)\big| = \varepsilon_4\, A\, q_2,$$

and since $\varepsilon_4 = \mathcal{O}(\|c\|^2)$ as $c \to 0$, we have $\|T_4\|_\infty = \mathcal{O}(\|c\|^2)$, as $c \to 0$.

For T_5 we can use $e^{\tilde{\delta}x^2} \le 1$ for all $x \in \mathbb{R}$, and from Proposition 5 we have

$$T_5(x) \le |\sqrt{2\pi}A - 1| = \mathcal{O}(\|c\|^2), \text{ as } c \to 0.$$

This completes the proof. □

We have therefore shown that for sufficiently small c, the approximation $\hat{f}_0$ for the density that maximises entropy given the constraints in (7) is 'close to' f_0. We have also shown that the speed of convergence is of order $\|c\|^2$.

4.3. Step 3

We now turn our attention to Step 3. of the approximations, where we find approximations for the entropy and negentropy of $\hat{f}_0$. For these proofs we require that $\hat{f}_0(x) \geq 0$ for all $x \in \mathbb{R}$, and thus $\hat{f}_0$ is a density.

Lemma 9 (*Approximation of Entropy*). *Suppose Assumptions 1 and 2 hold, and let $\hat{f}_0$ be given as in Theorem 8. Suppose also that $\hat{f}_0(x) \geq 0$ for all $x \in \mathbb{R}$. Then the entropy of $\hat{f}_0$ satisfies*

$$H[\hat{f}_0] = \hat{H}[\hat{f}_0] + R(\hat{f}_0),$$

where,

$$\hat{H}[\hat{f}_0] := \eta(1) - \frac{1}{2}\|c\|^2,$$

with $\eta(\cdot)$ given in (2), $c = (c_1, c_2, \ldots, c_l)^\top$, with the c_i defined in Proposition 3 and the remainder term bounded by

$$|R(\hat{f}_0)| \leq C\,\tilde{M} \cdot \|c\|^3,$$

for some constant $C \in \mathbb{R}$, and $\tilde{M}$ given in Assumption 2.

Proof. Set $K(x) = (K_1(x), K_2(x), \ldots, K_l(x))^\top$, for $x \in \mathbb{R}$. Now, with $\hat{f}_0$ as in Theorem 8, expanding $H[\hat{f}_0]$ gives,

$$\begin{aligned}
H[\hat{f}_0] &= -\int \hat{f}_0(x)\log\hat{f}_0(x)\,\mathrm{d}x = -\int \varphi(x)\Big(1 + c^\top K(x)\Big)\Big(\log\varphi(x) + \log\big(1 + c^\top K(x)\big)\Big) \\
&= -\int \varphi(x)\log\varphi(x)\,\mathrm{d}x - \int \varphi(x)c^\top K(x)\log\varphi(x)\,\mathrm{d}x - \int \varphi(x)\big(1 + c^\top K(x)\big)\log\big(1 + c^\top K(x)\big)\,\mathrm{d}x \\
&= \eta(1) - \int \varphi(x)c^\top K(x)\Big(-\frac{1}{2}\log(2\pi) - \frac{1}{2}x^2\Big)\,\mathrm{d}x - \int \varphi(x)\big(1 + c^\top K(x)\big)\log\big(1 + c^\top K(x)\big)\,\mathrm{d}x \\
&= \eta(1) - 0 - \int \varphi(x)\big(1 + c^\top K(x)\big)\log\big(1 + c^\top K(x)\big)\,\mathrm{d}x,
\end{aligned}$$

using the constraints given in (6). To obtain the approximation $\hat{H}[\hat{f}_0]$ and remainder $R(\hat{f}_0)$ terms, we consider the expansion of $\big(1 + c^\top K(x)\big)\log\big(1 + c^\top K(x)\big)$ around $c = 0$ using the Taylor series. Let $q(y) = y\log(y)$, $y \in \mathbb{R}$. Then, we have

$$q'(y) = \log(y) + 1; \quad q''(y) = \frac{1}{y}; \quad q'''(y) = -\frac{1}{y^2}.$$

and thus using Taylor series around y_0 gives $q(y_0 + h) = h + \frac{1}{2}h^2 + R_1(y_0, h)$, where $R_1(y_0, h)$ is the remainder term given by

$$R_1(y_0, h) = \int_{y_0}^{y_0+h} \frac{(y_0 + h - \tau)^2}{2}\Big(\frac{-1}{\tau^2}\Big)\,\mathrm{d}\tau = -h^3\int_0^1 \frac{(1-t)^2}{2(1+th)^2}\,\mathrm{d}t$$

with the change of variables $\tau = (y_0 + th)$.

Now let us pick $y_0 = 1$ and $h = c^\top K(x)$ and denote by $R_2(x)$ the corresponding remainder $R_2(x) = R_1(1,\ c^\top K(x))$. Then,

$$H[\hat{f}_0] = \eta(1) - \int \varphi(x)\Big(c^\top K(x) + \frac{1}{2}\big(c^\top K(x)\big)^2 + R_2(x)\Big)\,\mathrm{d}x, \tag{28}$$

where the remainder term $R_2(x)$ is given explicitly by

$$R_2(x) = -\big(c^\top K(x)\big)^3\int_0^1 \frac{(1-t)^2}{2\big(1 + tc^\top K(x)\big)^2}\,\mathrm{d}t.$$

Now using (6) and setting

$$R(\hat{f}_0) := -\int_{\mathbb{R}} \varphi(x)\,R_2(x)\,\mathrm{d}x \tag{29}$$

we get from (28),

$$H[\hat{f}_0] = \eta(1) + 0 - \frac{1}{2}\sum_{i=1}^{l} c_i^2 + R(\hat{f}_0) = \hat{H}[\hat{f}_0] + R(\hat{f}_0),$$

as needed to be shown. It remains to prove the bound for $R(\hat{f}_0)$.

From Assumption 2 there exists some $\varepsilon > 0$ such that $c^\top K(x) \geq -1/2$ for all c with $c^\top c \leq \varepsilon$ for all $x \in \mathbb{R}$, and therefore,

$$|R_2(x)| = \left|\left(c^\top K(x)\right)^3 \int_0^1 \frac{(t-1)^2}{2\cdot(1+tc^\top K(x))^2}\,\mathrm{d}t\right| \leq \left|\left(c^\top K(x)\right)\right|^3 \cdot \left|\int_0^1 \frac{(t-1)^2}{2\cdot(1-t/2)^2}\,\mathrm{d}t\right| = C\cdot\left|\left(c^\top K(x)\right)\right|^3,$$

where $C \in \mathbb{R}$, as the integral is of a continuous function over a compact set.

Now, there exists some $\delta > 0$ such that for all $c^\top c \leq \delta$, $c_i \leq \|c\|$ for all $i \in \{1, \dots, I\}$. Then, with $c^\top c \leq \min(\varepsilon, \delta)$, we have

$$|R_2(x)| \leq C \sum_{i,j,k=1}^{I} \left|K_i(x)K_j(x)K_k(x)\right| \cdot \|c\|^3 \leq C \cdot M(x) \cdot \|c\|^3$$

having used (13a) from Assumption 2. Putting this all together we obtain the bound for $R(\hat{f}_0)$,

$$\left|R(\hat{f}_0)\right| \leq \int_{\mathbb{R}} \varphi(x)|R_2(x)|\,\mathrm{d}x \leq C\tilde{M}\|c\|^3,$$

where $\tilde{M}$ is given in (13b), as required. □

Remark 10. Note that the density $\hat{f}_0$ has unit variance. Indeed, by (6),

$$\int \hat{f}_0(x)x^2\,\mathrm{d}x = \int \varphi(x)\big(1+\sum_{i=1}^{I} c_iK_i(x)\big)\,\mathrm{d}x = \int \varphi(x)x^2\,\mathrm{d}x + \sum_{i=1}^{I} c_i \int \varphi(x)K_i(x)x^2\,\mathrm{d}x = 1.$$

Therefore, the negentropy equivalent of the entropy approximation given in Lemma 9 is $J[\hat{f}_0] = \hat{J}[\hat{f}_0] + R(\hat{f}_0)$ with $R(\hat{f}_0)$ given as in (29) and

$$\hat{J}[\hat{f}_0] = \frac{1}{2}\|c\|^2 = \frac{1}{2}\sum_{i=1}^{I} c_i^2. \tag{30}$$

Proposition 11. *With the same assumptions as in Lemma 9. Set $I = 1$. Then,*

$$\hat{J}[\hat{f}_0] \propto \left(\mathsf{E}_f G(Y) - \mathsf{E}_\varphi G(Z)\right)^2,$$

where Y is a random variable with density f and $Z \sim \mathcal{N}(0,1)$.

Proof. By the constraints that need to be satisfied by K, given in Assumption 2, we have $\int \varphi(x)K(x)x^k\,\mathrm{d}x = 0$ for $k = 0, 1, 2$. Substituting (5) for $K(x)$ in (6) and solving these three equations gives an explicit expression for α, β, γ in terms of G, given by,

$$\alpha = \frac{1}{2}\Big(\int \varphi(x)G(x)\,\mathrm{d}x - \int \varphi(x)G(x)x^2\,\mathrm{d}x\Big); \quad \beta = -\int \varphi(x)G(x)x\,\mathrm{d}x; \tag{31}$$
$$\gamma = \frac{1}{2}\Big(\int \varphi(x)G(x)x^2\,\mathrm{d}x - 3\int \varphi(x)G(x)\,\mathrm{d}x\Big).$$

Recall that $c = \mathsf{E}K(Y) = \frac{1}{\delta}\big(\mathsf{E}G(Y) + \alpha\mathsf{E}Y^2 + \beta\mathsf{E}Y + \gamma\big)$. Now using (31) and the fact that $\mathsf{E}Y = 0$ and $\mathsf{E}Y^2 = 1$ (since Y has density f), we get $c = \frac{1}{\delta}\big(\mathsf{E}G(Y) - \mathsf{E}G(Z)\big)$. From (30) with $I = 1$, we have $\hat{J}[\hat{f}_0] = \frac{1}{2}c^2$, hence,

$$\hat{J}[\hat{f}_0] = \frac{\big(\mathsf{E}G(Y) - \mathsf{E}G(Z)\big)^2}{2\,\delta^2}.$$

This completes the proof, with $C = 2\,\delta^2$ in Step 3. of Section 3. Note that δ can be found by solving the additional constraint $\int \varphi(x)K(x)^2\,\mathrm{d}x = 1$. □

This concludes our discussion of the approximations used in fastICA. We have shown that under certain conditions, the approximations given in Steps 2., 3. and 4. in Section 3 are "close" to the true values. We will now give an example where these approximations are indeed close to one-another, but the surrogate density of the projections, f_0 from Step 1. is not close to the true density f.

5. Example

We now highlight the approximation steps as explained in Section 3 on a toy example. In this section we use example data as illustrated in Fig. 1, which was intentionally created in a very simplistic manner to further emphasise the ease at

which false optima are found using the contrast function $\hat{J}^*(y)$. The data was obtained by pre-selecting vertical columns where no data points are allowed. An iterative scheme was then employed, as explained below:

1. Sample n points from a standard two-dimensional Gaussian distribution;
2. Remove all points that lie in the pre-specified columns;
3. Whiten the remaining $\tilde{n}$ points;
4. Sample $n - \tilde{n}$ points from a standard two-dimensional Gaussian distribution.

Repeat 2. - 4. until we have a sample of size n with no points lying in the pre-specified columns. No optimisation was done to the distribution of these points to attempt to force the fastICA contrast function to have a false optimum.

We will use the m-spacing approximation (3) to obtain a contrast function that can be compared to the fastICA contrast function (11). Following Learned-Miller and Fisher [16], we chose $m = \sqrt{n}$, where $n \in \mathbb{N}$ is the number of observations. This was chosen so that the condition $m/n \to 0$ as $n \to \infty$ is satisfied [3,22]. This approximation to entropy is a direct approximation to $H[f]$, and therefore does not involve an equivalent Step 1. from Section 3 where f is substituted by a new density f_0.

Using the m-spacing method to find the first independent component loading, we want to obtain the direction $w^* := \text{argmin}_{w\in\mathbb{R}^p, w^\top w=1} H_{m,n}(Dw)$. In the example of this paper, numerical minimisation is used to obtain w^* and the associated projection Dw^*. The contrast function to compare against the fastICA contrast function (11) is given by the m-spacing negentropy approximation, $J_{m,n}(y) = \eta(1) - H_{m,n}(y)$ for directions $w \in \mathbb{R}^n$ on the half-sphere. Note that $w^* = \text{argmax}_{w\in\mathbb{R}^p, w^\top w=1} J_{m,n}(Dw)$. In general this contrast function is not very smooth, although a method to attempt to overcome this non-smoothness (and the resulting local optima, which can cause numerical optimisation issues) is given in Learned-Miller and Fisher [16], and involves replicating the data with some added Gaussian noise.

To illustrate the kind of problems which can occur during the approximation from f to $\hat{f}_0$ and from $J[f]$ to $\hat{J}^*(y)$, we construct an example where the density f in the direction of maximum negentropy is significantly different to $\hat{f}_0$ in the same direction. This results in `fastICA` selecting a sub-optimal projection, as shown below. Here we just consider the case $I = 1$ in Assumption 1, with one $G = G_1$ and thus one $K = K_1$. Moreover, in `fastICA` there is a choice of two functions to use, $G(x) := (1/\alpha)\log\cosh(\alpha x)$, $\alpha \in [1, 2]$, and $G(x) := -\exp(-x^2/2)$. We have considered these two functions with varying alpha, as well as the fourth moment contrast function given in Miettinen et al. [19]. Here, the function for Step 3. in Section 3 is $|\mathsf{E}_f X^4 - 3|$, and the empirical approximation of the expectation is used for Step 4., such that the approximate contrast function is $|\frac{1}{n}\sum_{i=1}^n y_i^4 - 3|$. In the example of this paper all choices give very similar results and thus we only show the fastICA contrast function resulting from $G(x) = (1/\alpha)\log\cosh(\alpha x)$, with $\alpha = 1$ for simplicity.

With the data distributed as in Fig. 1, the negentropy over projections in the directions $w_\theta = (\sin(\theta), \cos(\theta))$ with $\theta \in [0, \pi)$ found by the m-spacing approximation and used in the `fastICA` method is shown in Fig. 4. The contrast function obtained by approximating $J[f_0]$ directly is also included as the dashed line. The three contrast functions have been placed below one-another in the order of approximations given in Fig. 3 and so the y-axis is independent for each. The search is only performed on the half unit circle, as projections in directions $w_1 = (\sin(\theta), \cos(\theta))$ and $w_2 = (\sin(\theta + \pi), \cos(\theta + \pi))$ for any $\theta \in [0, \pi)$ have a reflected density with the same entropy. It is clear from Fig. 4 that the `fastICA` result $\hat{J}^*$ is poor, with the `fastICA` contrast function missing the peak of negentropy that appears when using m-spacing. The contrast function used in the `fastICA` method clearly differentiates between the direction of the maximum and other directions, and thus in this example it is both confident and wrong (since there is a clear and unique peak). This is also true of the direct approximation to $J[f_0]$, showing that issues occur at the first step of approximations, when $J[f_0]$ is used instead of $J[f]$.

As is shown in Section 4, for sufficiently small c, the approximation for the density $\hat{f}_0$ (given in (9)) is "close to" f_0 (given in (8)), and the speed of convergence is of order c^2 for $c \to 0$. Therefore, it is our belief (backed up by computational experiments) that the majority of the loss of accuracy occurs in the approximation step where the surrogate f_0 is used instead of f, rather than in the later estimation steps for $J[\hat{f}_0]$ and $\hat{J}^*(y)$. This can be seen by comparing numerically the contrast functions $J[f]$, $J[f_0]$ and $\hat{J}^*(y)$ (shown in Fig. 4), and by comparing the densities f, f_0 and $\hat{f}_0$. Here, $J[f_0]$ and $\hat{J}^*(y)$ give similar directions for the maximum, and these differ significantly from the location of the maximum of $J[f]$. This is a fundamental theoretical problem with the fastICA method, and is not a result of computational or implementation issues with `fastICA`. In particular, the fact that the dotted vertical line in Fig. 4 is at the maximum of $\hat{J}^*(y)$ indicates that the effect is not a convergence problem in the `fastICA` implementation.

6. Conclusions

In this paper we have given an example where the fastICA method misses structure in the data that is obvious to the naked eye. Since this example is very simple, the fastICA result is concerning, and this concern is magnified when working in high dimensions as visual inspection is no longer easy. There is clearly some issue with the contrast function (surrogate negentropy) used in fastICA. Indeed, this surrogate has the property of being an approximation of a lower bound for negentropy, and this does not necessarily capture the actual behaviour of negentropy over varying projections since we want to maximise negentropy. To strengthen the claim that accuracy is lost when substituting the density with the surrogate, we have shown convergence results for all the approximation steps used in the method.

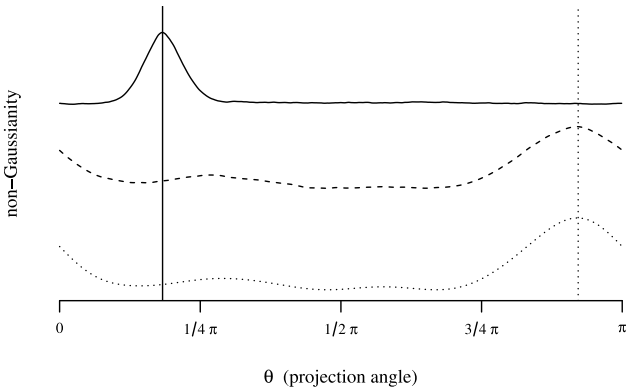


Fig. 4. Objective functions of m-spacing (solid line), $J[f_0]$ (dashed line) and `fastICA` method (dotted line) for projections of the data given in Fig. 1 in the directions $\theta \in [0, \pi)$. These correspond to $J[f]$, $J[f_0]$ and $\hat{J}^*(y)$ in Fig. 3. The vertical lines give the directions which maximise the contrast functions for m-spacing (solid line) and `fastICA` (dotted line).

To conclude this paper, we ask the following questions which could make for interesting future work: Is there a way, a priori, to know whether fastICA will work? This is especially pertinent when fastICA is used with high dimensional data. The trade-off in accuracy for the fastICA method comes at the point where the density f is substituted with f_0. Therefore one could also ask: Are there other methods similar to that of fastICA but that use a different surrogate density which more closely reflects the true projection density?

If these two options are not possible, then potentially a completely different method for "fast" ICA is needed, one that either gives a "good" approximation for all distributions, or where it is known when it breaks down. An initial step in this direction can be found in Smith et al. [20]. In this work the authors propose a new ICA method, known as clusterICA, using the m-spacing approximation for entropy discussed in this paper, combined with a clustering procedure.

CRediT authorship contribution statement

Elena Issoglio: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Supervision. **Paul Smith:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Jochen Voss:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision.

Acknowledgments

We thank the Editor, Associate Editor and referees for their helpful comments. P. Smith was funded by NERC DTP SPHERES, Grant Number NE/L002574/1.

References

[1] S. Amari, A. Cichocki, H. Yang, A new learning algorithm for blind signal separation, in: Adv. Neural Inf. Process. Syst., MIT Press, 1996, pp. 757–763.
[2] F. Bach, M. Jordan, Kernel independent component analysis, J. Mach. Learn. Res. 3 (Jul) (2002) 1–48.
[3] J. Beirlant, E.J. Dudewicz, L. Györfi, E.C. Van der Meulen, Nonparametric entropy estimation: An overview, Int. J. Math. Stat. Sci. 6 (1) (1997) 17–39.
[4] T. Cover, J. Thomas, Elements of Information Theory, John Wiley & Sons, New York, 2012.
[5] P. Diaconis, D. Freedman, Asymptotics of graphical projection pursuit, Ann. Statist. 12 (3) (1984) 793–815.
[6] B. Draper, K. Baek, M. Bartlett, J. Beveridge, Recognizing faces with PCA and ICA, Comput. Vis. Image Underst. 91 (1–2) (2003) 115–137.
[7] M. Farhat, Y. Gritli, M. Benrejeb, Fast-ICA for mechanical fault detection and identification in electromechanical systems for wind turbine applications, Int. J. Adv. Comput. Sci. Appl. 8 (7) (2017) 431–439.
[8] S. Ghaffarian, S. Ghaffarian, Automatic building detection based on Purposive FastICA (PFICA) algorithm using monocular high resolution Google Earth images, ISPRS J. Photogramm. Remote Sens. 97 (2014) 152–159.
[9] P. Hall, Limit theorems for sums of general functions of m-spacings, Mathematical Proceedings of the Cambridge Philosophical Society 96 (3) (1984) 517–532.
[10] X. He, F. He, T. Zhu, Large-scale super-Gaussian sources separation using fast-ICA with rational nonlinearities, Internat. J. Adapt. Control Signal Process. 31 (3) (2017) 379–397.
[11] A. Hyvärinen, New approximations of differential entropy for independent component analysis and projection pursuit, in: Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10, MIT Press, 1998, pp. 273–279.
[12] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. Neural Netw. 10 (3) (1999) 626–634.
[13] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, New York, 2001.
[14] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Netw. 13 (4) (2000) 411–430.
[15] E. Lawrence, Partial Differential Equations, American Mathematical Society, New York, 1998.

[16] E. Learned-Miller, J. Fisher, ICA using spacings estimates of entropy, J. Mach. Learn. Res. 4 (Dec) (2003) 1271–1295.
[17] J. Marchini, C. Heaton, B. Ripley, FastICA: FastICA algorithms to perform ICA and projection pursuit, 2013, R package version 1.2-0.
[18] J. Miettinen, K. Nordhausen, H. Oja, S. Taskinen, Deflation-based fastICA with adaptive choices of nonlinearities, IEEE Trans. Signal Process. 62 (21) (2014) 5716–5724.
[19] J. Miettinen, S. Taskinen, K. Nordhausen, H. Oja, et al., Fourth moments and independent component analysis, Stat. Sci. 30 (3) (2015) 372–390.
[20] P. Smith, J. Voss, E. Issoglio, ClusterICA, 2020, in preparation.
[21] J. Stone, Independent Component Analysis: A Tutorial Introduction, MIT Press, Massachusetts, 2004.
[22] O. Vasicek, A test for normality based on sample entropy, J. R. Stat. Soc. Ser. B Stat. Methodol. 38 (1) (1976) 54–59.
[23] T. Wei, On the spurious solutions of the fastICA algorithm, in: 2014 IEEE Workshop on Statistical Signal Processing (SSP), 2014, pp. 161–164.
[24] T. Wei, A study of the fixed points and spurious solutions of the deflation-based FastICA algorithm, Neural Comput. Appl. 28 (1) (2017) 13–24.
[25] H. von Weizsäcker, Sudakov's typical marginals, random linear functionals and a conditional central limit theorem, Probab. Theory Related Fields 107 (3) (1997) 313–324.
[26] C.-H. Yang, Y.-H. Shih, H. Chiueh, An 81.6 μW fastICA processor for epileptic seizure detection, IEEE Trans. Biomed. Circuits Syst. 9 (1) (2015) 60–71.