The Annals of Statistics
2012, Vol. 40, No. 6, 2973–3002
DOI: 10.1214/12-AOS1060

INDEPENDENT COMPONENT ANALYSIS VIA NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION

BY RICHARD J. SAMWORTH[1] AND MING YUAN[2]

University of Cambridge and Georgia Institute of Technology

Independent Component Analysis (ICA) models are very popular semiparametric models in which we observe independent copies of a random vector $X = AS$, where A is a non-singular matrix and S has independent components. We propose a new way of estimating the unmixing matrix $W = A^{-1}$ and the marginal distributions of the components of S using nonparametric maximum likelihood. Specifically, we study the projection of the empirical distribution onto the subset of ICA distributions having log-concave marginals. We show that, from the point of view of estimating the unmixing matrix, it makes no difference whether or not the log-concavity is correctly specified. The approach is further justified by both theoretical results and a simulation study.

1. Introduction. In recent years, Independent Component Analysis (ICA) has seen an explosion in its popularity in diverse fields such as signal processing, machine learning and medical imaging, to name a few. For a wide-ranging list of algorithms and applications of ICA, see the monograph by Hyvärinen, Karhunen and Oja (2001). In the ICA paradigm, one observes a random vector $X \in \mathbb{R}^d$ that can be expressed as a non-singular linear transformation of d mutually independent latent factors $S_1, \ldots, S_d$; thus, $X = AS$, where $S = (S_1, \ldots, S_d)^\mathsf{T}$ and A is a $d \times d$ full rank matrix often referred to as the mixing matrix. As such, ICA postulates the following model for the probability distribution P of X: for any Borel set B in $\mathbb{R}^d$,

$$P(B) = \prod_{j=1}^{d} P_j(w_j^\mathsf{T} B),$$

where $W = (w_1, \ldots, w_d)^\mathsf{T} = A^{-1}$ is the so-called unmixing matrix and $P_1, \ldots, P_d$ are the univariate probability distributions of the latent factors $S_1, \ldots, S_d$, respectively.

The goal of ICA, as in other blind source separation problems, is to infer, from a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of independent observations of X, the independent factors

Received June 2012; revised October 2012.

[1]Supported in part by a Leverhulme Research Fellowship and an EPSRC Early Career Fellowship.

[2]Supported in part by NSF Career Award DMS-0846234.

MSC2010 subject classifications. 62G07.

Key words and phrases. Blind source separation, density estimation, independent component analysis, log-concave projection, nonparametric maximum likelihood estimator.

$\mathbf{s}_1 = W\mathbf{x}_1, \ldots, \mathbf{s}_n = W\mathbf{x}_n$ or, equivalently, the unmixing matrix W. This task is typically accomplished by first postulating a certain parametric family for the marginal probability distributions $P_1, \ldots, P_d$, and then optimising a contrast function involving $(W, P_1, \ldots, P_d)$, for example, Karvanen and Koivunen (2002). The contrast functions are often chosen to represent the mutual information as measured by Kullback–Leibler divergence or maximum entropy, or non-Gaussianity as measured by kurtosis or negentropy. Alternatively, in recent years, methods for ICA have also been developed which assume $P_1, \ldots, P_d$ have smooth (log) densities, for example, Bach and Jordan (2002), Hastie and Tibshirani (2003b), Samarov and Tsybakov (2004), Chen and Bickel (2006) and Ilmonen and Paindaveine (2011). Although more flexible than their aforementioned parametric peers, there remain unsettling questions about what happens if the smoothness assumptions on the marginal densities are violated, which may occur, in particular, when some of the marginal probability distributions $P_1, \ldots, P_d$ have atoms. Another issue is that, in common with most other smoothing methods, a choice of tuning parameters is required to balance the fidelity to the observed data and the smoothness of the estimated marginal densities, and it is notoriously difficult to select these tuning parameters appropriately in practice.

In this paper, we argue that these assumptions and tuning parameters are unnecessary, and propose a new paradigm for ICA, based on the notion of nonparametric maximum likelihood, that is free of these burdens. In fact, we show that the usual nonparametric (empirical) likelihood approach does not work in this context, and instead we proceed under the working assumption that the marginal distributions of $S_1, \ldots, S_d$ are log-concave. More specifically, we propose to estimate W by maximising

$$\log|\det W| + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\log f_j(w_j^{\mathsf{T}}\mathbf{x}_i)$$

over all $d \times d$ non-singular matrices $W = (w_1, \ldots, w_d)^{\mathsf{T}}$ and univariate log-concave densities $f_1, \ldots, f_d$. Remarkably, from the point of view of estimating the unmixing matrix W, it turns out that it makes no difference whether or not this hypothesis of log-concavity is correctly specified.

The key to understanding how our approach works is to study what we call the *log-concave ICA projection* of a distribution on $\mathbb{R}^d$ onto the set of densities that satisfy the ICA model with log-concave marginals. In Section 2.1 below, we define this projection carefully and give necessary and sufficient conditions for it to make sense. In Section 2.2, we prove that the log-concave projection of a distribution from the ICA model preserves both the ICA structure and the unmixing matrix. Finally, in Section 2.3, we derive a continuity property of log-concave ICA projections, which turns out to be important for understanding the theoretical properties of our ICA procedure.

Our ICA estimating procedure uses the log-concave ICA projection of the empirical distribution of the data, and is studied in Section 3. After explaining why the usual empirical likelihood approach cannot be used, we prove the consistency of our method. We also present an iterative algorithm for the computation of our estimator. Our simulation studies in Section 4 confirm our theoretical results and show that the proposed method compares favourably with existing methods. Proofs are deferred to Section 5.

To conclude this section, we remark that in addition to the previous literature already cited, further approaches to ICA have been proposed that use a choice of two scatter (or shape) matrices [Nordhausen, Oja and Ollila (2011), Oja, Sirkiä and Eriksson (2006), Ollila, Oja and Koivunen (2008)]. To uniquely define the unmixing matrix, one scatter matrix (often chosen based on fourth moments) must take distinct values on its main diagonal, which rules out situations where two of the marginal distributions $P_1, \ldots, P_d$ are the same. Nevertheless, under further (e.g., moment) assumptions, root-n consistency and asymptotic normality results for estimates of the unmixing matrix under correct model specification have been obtained [e.g., Ilmonen, Nevalainen and Oja (2010)].

2. Log-concave ICA projections.

2.1. *Notation and overview.* Let $\mathcal{P}_k$ be the set of probability distributions P on $\mathbb{R}^k$ satisfying $\int_{\mathbb{R}^k} \|x\| \, dP(x) < \infty$ and $P(H) < 1$ for all hyperplanes H, that is, the probability measures in $\mathbb{R}^k$ that have finite mean and are not supported in a translate of a lower-dimensional linear subspace of $\mathbb{R}^k$. Here and throughout, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^k$, and we will be interested in the cases $k = 1$ and $k = d$. Further, let $\mathcal{W}$ denote the set of non-singular $d \times d$ real matrices. We use upper case letters to denote matrices in $\mathcal{W}$, and the corresponding lower case letters with subscripts to denote rows: thus, w_j^T is the jth row of $W \in \mathcal{W}$. Let $\mathcal{B}_k$ denote the class of Borel sets on $\mathbb{R}^k$. Then the ICA model $\mathcal{P}_d^{\mathrm{ICA}}$ is defined to be the set of $P \in \mathcal{P}_d$ of the form

$$P(B) = \prod_{j=1}^{d} P_j(w_j^\mathsf{T} B) \qquad \forall B \in \mathcal{B}_d \tag{1}$$

for some $W \in \mathcal{W}$ and $P_1, \ldots, P_d \in \mathcal{P}_1$. As shown by Dümbgen, Samworth and Schuhmacher [(2011), Theorem 2.2], the condition $P \in \mathcal{P}_d$ is necessary and sufficient for the existence of a unique upper semi-continuous and log-concave density that is the closest to P in the Kullback–Leibler sense. More precisely, let $\mathcal{F}_k$ denote the class of all upper semi-continuous, log-concave densities with respect to the Lebesgue measure on $\mathbb{R}^k$. Then the projection $\psi^* : \mathcal{P}_d \to \mathcal{F}_d$ given by

$$\psi^*(P) = \operatorname*{arg\,max}_{f \in \mathcal{F}_d} \int_{\mathbb{R}^d} \log f \, dP$$

is well defined and surjective. In what follows, we refer to ψ^* as the *log-concave projection operator* and $f^* := \psi^*(P)$ as the *log-concave projection* of P. By a slight abuse of notation, we also use ψ^* to denote the log-concave projection from $\mathcal{P}_1$ to $\mathcal{F}_1$.

Although the log-concave projection operator does play a role in this paper, our main interest is in a different projection, onto the subset of $\mathcal{F}_d$ consisting of those densities that belong to the ICA model. This class is given by

$$\mathcal{F}_d^{\mathrm{ICA}} = \Bigg\{ f \in \mathcal{F}_d : f(x) = |\det W| \prod_{j=1}^{d} f_j\big(w_j^{\mathsf{T}} x\big) \quad \text{with } W \in \mathcal{W}, f_1, \ldots, f_d \in \mathcal{F}_1 \Bigg\}. \tag{2}$$

Note that, in this representation, if X has density $f \in \mathcal{F}_d^{\mathrm{ICA}}$, then $w_j^{\mathsf{T}} X$ has density f_j. The corresponding log-concave ICA projection operator $\psi^{**}(\cdot)$ is defined for any distribution P on $\mathbb{R}^d$ by

$$\psi^{**}(P) = \operatorname*{arg\,max}_{f \in \mathcal{F}_d^{\mathrm{ICA}}} \int_{\mathbb{R}^d} \log f \, dP.$$

We also write $L^{**}(P) = \sup_{f \in \mathcal{F}_d^{\mathrm{ICA}}} \int_{\mathbb{R}^d} \log f \, dP$.

PROPOSITION 1. (1) *If* $\int_{\mathbb{R}^d} \|x\| \, dP(x) = \infty$, *then* $L^{**}(P) = -\infty$ *and* $\psi^{**}(P) = \mathcal{F}_d^{\mathrm{ICA}}$.

(2) *If* $\int_{\mathbb{R}^d} \|x\| \, dP(x) < \infty$, *but* $P(H) = 1$ *for some hyperplane* H, *then* $L^{**}(P) = \infty$ *and* $\psi^{**}(P) = \varnothing$.

(3) *If* $P \in \mathcal{P}_d$, *then* $L^{**}(P) \in \mathbb{R}$ *and* $\psi^{**}(P)$ *defines a non-empty, proper subset of* $\mathcal{F}_d^{\mathrm{ICA}}$.

In view of Proposition 1, and to avoid lengthy discussion of trivial exceptional cases, we henceforth consider $\psi^{**}(\cdot)$ as being defined on $\mathcal{P}_d$. In contrast to $\psi^*(P)$, which defines a unique element of $\mathcal{F}_d$, the log-concave ICA projection operator $\psi^{**}(P)$ may not define a unique element of $\mathcal{F}_d^{\mathrm{ICA}}$, even for $P \in \mathcal{P}_d$. For instance, consider the situation where P is the uniform distribution on the closed unit disk in $\mathbb{R}^2$ equipped with the Euclidean norm. Here, the spherical symmetry means that the choice of $W \in \mathcal{W}$ is not uniquely defined. In fact, after a straightforward calculation, it can be shown that $\psi^{**}(P)$ includes all $f \in \mathcal{F}_2^{\mathrm{ICA}}$, where, in the representation (2), W is an orthogonal matrix and $f_1, f_2 \in \mathcal{F}_1$ are given by $f_1(x) = f_2(x) = \frac{2}{\pi}(1 - x^2)^{1/2} \mathbb{1}_{\{x \in [-1,1]\}}$. It is certainly possible to make different choices of W that yield different elements of $\mathcal{F}_2^{\mathrm{ICA}}$. This example shows that, in general, we must think of $\psi^{**}(P)$ as defining a subset of $\mathcal{F}_d^{\mathrm{ICA}}$.

The relationship between the spaces introduced above and the projection operators is illustrated in the diagram below:

$$\begin{array}{ccc} \mathcal{P}_d & \xrightarrow{\psi^*} & \mathcal{F}_d \\ & \overset{\psi^{**}}{\searrow} & \\ \mathcal{P}_d^{\mathrm{ICA}} & \xrightarrow{\psi^{**}|_{\mathcal{P}_d^{\mathrm{ICA}}}} & \mathcal{F}_d^{\mathrm{ICA}}. \end{array}$$

Our next subsection studies the restriction of ψ^{**} to $\mathcal{P}_d^{\mathrm{ICA}}$, denoted $\psi^{**}|_{\mathcal{P}_d^{\mathrm{ICA}}}$; Section 2.2 examines ψ^{**} more generally as a map on $\mathcal{P}_d$.

2.2. *Log-concave projections of the ICA model.* Our first result in this subsection characterises $\psi^{**}|_{\mathcal{P}_d^{\mathrm{ICA}}}$.

THEOREM 2. *If $P \in \mathcal{P}_d^{\mathrm{ICA}}$, then $\psi^{**}(P)$ defines a unique element of $\mathcal{F}_d^{\mathrm{ICA}}$. The map $\psi^{**}|_{\mathcal{P}_d^{\mathrm{ICA}}}$ is surjective, and coincides with $\psi^*|_{\mathcal{P}_d^{\mathrm{ICA}}}$. Moreover, suppose that $P \in \mathcal{P}_d^{\mathrm{ICA}}$, so that*

$$P(B) = \prod_{j=1}^{d} P_j(w_j^{\mathsf{T}} B) \qquad \forall B \in \mathcal{B}_d$$

for some $W \in \mathcal{W}$ and $P_1, \ldots, P_d \in \mathcal{P}_1$. Then $f^{**} = \psi^{**}(P)$ can be written as

$$f^{**}(x) = |\det W| \prod_{j=1}^{d} f_j^*(w_j^{\mathsf{T}} x),$$

where $f_j^* = \psi^*(P_j)$.

It is interesting to observe that the log-concave projection operator ψ^* preserves the ICA structure. But perhaps the most important aspect of this result is the fact that the same unmixing matrix W can be used to represent both the original ICA model and its log-concave projection. This observation lies at the heart of the rationale for our approach to ICA.

A remaining concern is that the unmixing matrix may not be identifiable. For instance, applying the same permutation to the rows of W and the vector of marginal distributions $(P_1, \ldots, P_d)$ leaves the distribution unchanged; similarly, the same effect occurs if we multiply any of the rows of W by a scaling factor and apply the corresponding scaling factor to the relevant marginal distribution. The question of identifiability for ICA models was first addressed by Comon (1994), who assumed that W is orthogonal, and was settled in the general case by Eriksson

and Koivunen (2004). One way to state their result is as follows: suppose that a probability measure P on $\mathbb{R}^d$ has two representations as

$$P(B) = \prod_{j=1}^{d} P_j(w_j^\top B) = \prod_{j=1}^{d} \tilde{P}_j(\tilde{w}_j^\top B) \qquad \forall B \in \mathcal{B}_d, \tag{3}$$

where $W, \tilde{W} \in \mathcal{W}$ and $P_1, \ldots, P_d, \tilde{P}_1, \ldots, \tilde{P}_d$ are probability measures on $\mathbb{R}$. Then the pair of conditions that $P_1, \ldots, P_d$ are not Dirac point masses and not more than one of $P_1, \ldots, P_d$ is Gaussian is necessary and sufficient for the existence of a permutation π of $\{1, \ldots, d\}$ and scaling vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d) \in (\mathbb{R} \setminus \{0\})^d$ such that $\tilde{P}_j(B_j) = P_{\pi(j)}(\varepsilon_j B_j)$ for all $B_j \in \mathcal{B}_1$, and $\tilde{w}_j = \varepsilon_j^{-1} w_{\pi(j)}$. When such a permutation and scaling factor exist for any two ICA representations of P, we say that *the ICA representation of P is identifiable*, or simply that *P is identifiable*. By analogy, we define $f \in \mathcal{F}_d^{\mathrm{ICA}}$ to be identifiable if not more than one of $f_1, \ldots, f_d$ in the representation (2) is Gaussian.

Our next result shows that ψ^{**} preserves the identifiability of the ICA model. Together with Theorem 2, we see that if $P \in \mathcal{P}_d^{\mathrm{ICA}}$ is identifiable, then the unmixing matrices of P and $\psi^{**}(P)$ are identical up to the permutation and scaling transformations described above.

THEOREM 3. *Let $P \in \mathcal{P}_d^{\mathrm{ICA}}$. Then $\psi^{**}(P)$ is identifiable if and only if P is identifiable.*

2.3. *General log-concave ICA projections.* We now consider the general log-concave ICA projection ψ^{**} defined on $\mathcal{P}_d$. Define the Mallows distance d (also known as the Wasserstein distance) between probability measures P and $\tilde{P}$ on $\mathbb{R}^d$ with finite mean by

$$d(P, \tilde{P}) = \inf_{(X, \tilde{X}) \sim (P, \tilde{P})} \mathbb{E}\|X - \tilde{X}\|,$$

where the infimum is taken over all pairs $(X, \tilde{X})$ of random vectors $X \sim P$ and $\tilde{X} \sim \tilde{P}$ on a common probability space. Recall that $d(P^n, P) \to 0$ if and only if both $P^n \xrightarrow{d} P$ and $\int_{\mathbb{R}^d} \|x\| \, dP^n(x) \to \int_{\mathbb{R}^d} \|x\| \, dP(x)$. We are interested in the continuity of ψ^{**}.

PROPOSITION 4. *Let $P, P^1, P^2, \ldots$ be probability measures in $\mathcal{P}_d$ with $d(P^n, P) \to 0$ as $n \to \infty$. Then $L^{**}(P^n) \to L^{**}(P)$. Moreover,*

$$\sup_{f^n \in \psi^{**}(P^n)} \inf_{f \in \psi^{**}(P)} \int_{\mathbb{R}^d} |f^n - f| \to 0$$

as $n \to \infty$.

The second part of this proposition says that any element of $\psi^{**}(P^n)$ is arbitrarily close in total variation distance to some element of $\psi^{**}(P)$ once n is sufficiently large. In the special case where $\psi^{**}(P)$ consists of only a single element, we can say more. It is convenient to let Π_d denote the set of permutations of $\{1,\dots,d\}$, and write $(W, f_1, \dots, f_d) \overset{\text{ICA}}{\sim} f$ if $W \in \mathcal{W}$ and $f_1, \dots, f_d \in \mathcal{F}_1$ can be used to give an ICA representation of $f \in \mathcal{F}_d^{\text{ICA}}$ in (2). Similarly, we write $(W, P_1, \dots, P_d) \overset{\text{ICA}}{\sim} P$ if $W \in \mathcal{W}$ and $P_1, \dots, P_d \in \mathcal{P}_1$ represent $P \in \mathcal{P}_d^{\text{ICA}}$ in (1).

THEOREM 5. *Suppose that $P \in \mathcal{P}_d^{\text{ICA}}$, and write $f^{**} = \psi^{**}(P)$. If $P^1, P^2, \dots \in \mathcal{P}_d$ are such that $d(P^n, P) \to 0$, then*

$$\sup_{f^n \in \psi^{**}(P^n)} \int_{\mathbb{R}^d} |f^n - f^{**}| \to 0.$$

Suppose further that P is identifiable and that $(W, P_1, \dots, P_d) \overset{\text{ICA}}{\sim} P$. Then

$$\sup_{f^n \in \psi^{**}(P^n)} \sup_{(W^n, f_1^n, \dots, f_d^n) \overset{\text{ICA}}{\sim} f^n} \inf_{\pi^n \in \Pi_d} \inf_{\varepsilon_1^n, \dots, \varepsilon_d^n \in \mathbb{R}\setminus\{0\}} \left\{ \left\| (\varepsilon_j^n)^{-1} w_{\pi^n(j)}^n - w_j \right\| + \int_{-\infty}^{\infty} \left| |\varepsilon_j^n| f_{\pi^n(j)}^n(\varepsilon_j^n x) - f_j^*(x) \right| dx \right\} \to 0$$

for each $j = 1, \dots, d$, where $f_j^* = \psi^*(P_j)$. As a consequence, for sufficiently large n, every $f^n \in \psi^{**}(P^n)$ is identifiable.

The convergence statement in Theorem 5 is quite complicated, partly because of the need to deal with possible reordering and/or scaling of the rows of the unmixing matrices. An alternative approach, as adopted in Ilmonen and Paindaveine (2011), for instance, would be to study a particular representative of the equivalence class of unmixing matrices that can be used to represent a given distribution in $\mathcal{F}_d^{\text{ICA}}$. Our main reason for choosing this presentation was to make clear that the same permutation and scaling yields simultaneous convergence of the log-concave projections of the marginal densities.

The first part of Theorem 5 shows that if $P \in \mathcal{P}_d^{\text{ICA}}$ and $\tilde{P} \in \mathcal{P}_d$ are close in Mallows distance, then every $\tilde{f} \in \psi^{**}(\tilde{P})$ is close to the corresponding (unique) log-concave ICA projection $f = \psi^{**}(P)$ in total variation distance. The second part shows further that if P is identifiable, then up to permutation and scaling, every $\tilde{f} \in \psi^{**}(\tilde{P})$ and every choice of unmixing matrix $\tilde{W}$ and marginal densities $\tilde{f}_1, \dots, \tilde{f}_d$ in the ICA representation of $\tilde{f}$ is close to the unmixing matrix W and marginal densities $f_1, \dots, f_d$ in the ICA representation of f.

To conclude this subsection, we remark that, by analogy with the situation when $P \in \mathcal{P}_d^{\mathrm{ICA}}$ described in Theorem 2, if $P \in \mathcal{P}_d$ and $X \sim P$, any $f^{**} \in \psi^{**}(P)$ can be written as

$$f^{**}(x) = |\det W| \prod_{j=1}^{d} f_j^*(w_j^{\mathsf{T}} x)$$

for some $W \in \mathcal{W}$, where $f_j^* = \psi^*(P_j)$ and P_j is the marginal distribution of $w_j^{\mathsf{T}} X$. This observation reduces the maximisation problem involved in computing $\psi^{**}(P)$ to a finite-dimensional one (over $W \in \mathcal{W}$), and follows because

$$\begin{aligned}
&\sup_{f \in \mathcal{F}_d^{\mathrm{ICA}}} \int_{\mathbb{R}^d} \log f \, dP \\
&\quad = \sup_{W \in \mathcal{W}} \sup_{f_1, \ldots, f_d \in \mathcal{F}_1} \left\{ \log |\det W| + \sum_{j=1}^{d} \int_{\mathbb{R}^d} \log f_j(w_j^{\mathsf{T}} x) \, dP(x) \right\} \\
&\quad = \sup_{W \in \mathcal{W}} \left\{ \log |\det W| + \sum_{j=1}^{d} \int_{\mathbb{R}^d} \log f_j^*(w_j^{\mathsf{T}} x) \, dP(x) \right\}.
\end{aligned}$$

3. Nonparametric maximum likelihood estimation for ICA models. We are now in position to study the proposed nonparametric maximum likelihood estimator.

3.1. *Estimating procedure and theoretical properties.* Assume $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are independent copies of a random vector $X \in \mathbb{R}^d$ satisfying the ICA model. Thus, $X = AS$, where $A = W^{-1} \in \mathcal{W}$ and $S = (S_1, \ldots, S_d)^{\mathsf{T}}$ has independent components. In this section, we study a nonparametric maximum likelihood estimator of W and the marginal distributions $P_1, \ldots, P_d$ of $S_1, \ldots, S_d$ based on $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $n \geq d + 1$.

We start by noting that the usual nonparametric maximum likelihood estimate does not work. Indeed, in the spirit of empirical likelihood [Owen (1990)], it would suffice to consider, for a given $W = (w_1, \ldots, w_d)^{\mathsf{T}} \in \mathcal{W}$, estimates $\tilde{P}_j$ of the marginal distribution P_j, supported on $w_j^{\mathsf{T}} \mathbf{x}_1, \ldots, w_j^{\mathsf{T}} \mathbf{x}_n$. This leads to the nonparametric likelihood

$$L(W, \tilde{P}_1, \ldots, \tilde{P}_d) = \prod_{i=1}^{n} \prod_{j=1}^{d} \tilde{p}_{ij}, \tag{4}$$

where $\tilde{p}_{ij} = \tilde{P}_j(w_j^{\mathsf{T}} \mathbf{x}_i)$. Let J denote a subset of $(d + 1)$ distinct indices in $\{1, \ldots, n\}$, and let $\mathbf{X}_J$ denote the $d \times (d + 1)$ matrix obtained by extracting the columns of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with indices in J. Now let $\mathbf{X}_{(-j)}$ denote the $d \times d$ matrix obtained by removing the jth column of $\mathbf{X}_J$. Let $W_J \in \mathcal{W}$ have jth row

$w_j = (\mathbf{X}_{(-j)}^{-1})^\top \mathbf{1}_d$, for $j = 1, \ldots, d$, where $\mathbf{1}_d$ is a d-vector of ones. We say that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are in *general position* if, whenever we take a $n \times r$ matrix $\mathbf{M}$ of full rank, where every column of $\mathbf{M}$ contains exactly two non-zero entries, namely, a 1 and a -1, and define $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_r)$ by $\mathbf{Y} = \mathbf{X}\mathbf{M}$, then $\mathbf{Y}$ has full rank. Our next result shows that if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are in general position, then every W_J corresponds to a maximiser of the nonparametric likelihood (4).

PROPOSITION 6. *Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are in general position. Then for any choice J of $(d+1)$ distinct indices in $\{1, \ldots, n\}$, there exist $\hat{P}_1, \ldots, \hat{P}_d \in \mathcal{P}_1$ such that $(W_J, \hat{P}_1, \ldots, \hat{P}_d)$ maximises $L(\cdot)$.*

If X has a density with respect to the Lebesgue measure on $\mathbb{R}^d$, then $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are in general position with probability 1. On the other hand, there is no reason for different choices of J to yield similar estimates W_J, so we cannot hope for such an empirical likelihood-based procedure to be consistent.

As a remedy, we propose to estimate $P^0 \in \mathcal{P}_d^{\mathrm{ICA}}$ by $\psi^{**}(\hat{P}^n)$, where $\hat{P}^n$ denotes the empirical distribution of $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim P^0$. More explicitly, we estimate the unmixing matrix and the marginals by maximising the log-likelihood

$$\begin{aligned} \ell^n(W, f_1, \ldots, f_d) &= \ell^n(W, f_1, \ldots, f_d; \mathbf{x}_1, \ldots, \mathbf{x}_n) \\ &= \log|\det W| + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} \log f_j(w_j^\top \mathbf{x}_i) \end{aligned} \tag{5}$$

over $W \in \mathcal{W}$ and $f_1, \ldots, f_d \in \mathcal{F}_1$. Note from Proposition 1 that $\psi^{**}(\hat{P}^n)$ exists as a proper subset of $\mathcal{F}_d^{\mathrm{ICA}}$ once the convex hull of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is d-dimensional, which happens with probability 1 for sufficiently large n. As a direct consequence of Theorem 5 and the fact that $d(\hat{P}^n, P^0) \xrightarrow{\text{a.s.}} 0$, we have the following consistency result.

COROLLARY 7. *Suppose that $P^0 \in \mathcal{P}_d^{\mathrm{ICA}}$ is identifiable and is represented by $W^0 \in \mathcal{W}$ and $P_1^0, \ldots, P_d^0 \in \mathcal{P}_1$. Then for any maximiser $(\hat{W}^n, \hat{f}_1^n, \ldots, \hat{f}_d^n)$ of $\ell^n(W, f_1, \ldots, f_d)$ over $W \in \mathcal{W}$ and $f_1, \ldots, f_d \in \mathcal{F}_1$, there exist a permutation $\hat{\pi}^n$ of $\{1, \ldots, d\}$ and scaling factors $\hat{\varepsilon}_1^n, \ldots, \hat{\varepsilon}_d^n \in \mathbb{R} \setminus \{0\}$ such that*

$$(\hat{\varepsilon}_j^n)^{-1} \hat{w}_{\hat{\pi}^n(j)}^n \xrightarrow{a.s.} w_j^0 \quad \text{and} \quad \int_{-\infty}^{\infty} \big|\,|\hat{\varepsilon}_j^n| \hat{f}_{\hat{\pi}^n(j)}^n(\hat{\varepsilon}_j^n x) - f_j^*(x)\big|\,dx \xrightarrow{a.s.} 0$$

for $j = 1, \ldots, d$, where $f_j^* = \psi^*(P_j^0)$.

3.2. *Pre-whitening.* Pre-whitening is a standard pre-processing technique in the ICA literature; see Hyvärinen, Karhunen and Oja [(2001), pages 140 and 141] or Chen and Bickel (2005). In this subsection, we explain the rationale for pre-whitening and the simplifications it provides.

Assume for now that $P \in \mathcal{P}_d^{\mathrm{ICA}}$ and $\int_{\mathbb{R}^d} \|x\|^2 \, dP(x) < \infty$, and let Σ denote the (positive-definite) covariance matrix corresponding to P. Consider the ICA model $X = AS$, where $X \sim P$, the mixing matrix A is non-singular and $S = (S_1, \ldots, S_d)$ has independent components with $S_j \sim P_j$. Assuming without loss of generality that each component of S has unit variance, we can write $\Sigma^{-1/2} X = \Sigma^{-1/2} AS \equiv \tilde{A} S$, say, where $\tilde{A}$ belongs to the set $O(d)$ of orthogonal $d \times d$ matrices. Thus, the unmixing matrix W belongs to the set $O(d)\Sigma^{-1/2} = \{O\Sigma^{-1/2} : O \in O(d)\}$.

It follows that, if Σ were known, we could maximise ℓ^n with the restriction that $W \in O(d)\Sigma^{-1/2}$. In practice, Σ is typically unknown, but we can estimate it using the sample covariance matrix $\hat{\Sigma}$. For n large enough that the convex hull of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is d-dimensional, we can therefore consider maximising

$$\ell^n(W, f_1, \ldots, f_d; \mathbf{x}_1, \ldots, \mathbf{x}_n)$$

over $W \in O(d)\hat{\Sigma}^{-1/2}$ and $f_1, \ldots, f_d \in \mathcal{F}_1$. Denote such a maximiser by $(\hat{\hat{W}}^n, \hat{\hat{f}}_1^n, \ldots, \hat{\hat{f}}_d^n)$. The corollary below shows that, under a second moment condition, $\hat{\hat{W}}^n$ and $\hat{\hat{f}}_1^n, \ldots, \hat{\hat{f}}_d^n$ have the same asymptotic properties as the original estimators $\hat{W}^n$ and $\hat{f}_1^n, \ldots, \hat{f}_d^n$.

COROLLARY 8. *Suppose that $P^0 \in \mathcal{P}_d^{\mathrm{ICA}}$ is identifiable, is represented by $W^0 \in \mathcal{W}$ and $P_1^0, \ldots, P_d^0 \in \mathcal{P}_1$ and that $\int_{\mathbb{R}^d} \|x\|^2 \, dP^0(x) < \infty$. Then with probability 1 for sufficiently large n, a maximiser $(\hat{\hat{W}}^n, \hat{\hat{f}}_1^n, \ldots, \hat{\hat{f}}_d^n)$ of $\ell^n(W, f_1, \ldots, f_d)$ over $W \in O(d)\hat{\Sigma}^{-1/2}$ and $f_1, \ldots, f_d \in \mathcal{F}_1$ exists. Moreover, for any such maximiser, there exist a permutation $\hat{\hat{\pi}}^n$ of $\{1, \ldots, d\}$ and scaling factors $\hat{\hat{\varepsilon}}_1^n, \ldots, \hat{\hat{\varepsilon}}_d^n \in \mathbb{R} \setminus \{0\}$ such that*

$$(\hat{\hat{\varepsilon}}_j^n)^{-1} \hat{\hat{w}}^n_{\hat{\hat{\pi}}^n(j)} \xrightarrow{a.s.} w_j^0 \quad \text{and} \quad \int_{-\infty}^{\infty} \big|\hat{\hat{\varepsilon}}_j^n \big| \hat{\hat{f}}^n_{\hat{\hat{\pi}}^n(j)}\big(\hat{\hat{\varepsilon}}_j^n x\big) - f_j^*(x)\big| \, dx \xrightarrow{a.s.} 0,$$

where $f_j^* = \psi^*(P_j^0)$.

An alternative, equivalent way of computing $(\hat{\hat{W}}^n, \hat{\hat{f}}_1^n, \ldots, \hat{\hat{f}}_d^n)$ is to *pre-whiten* the data by replacing $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{z}_1 = \hat{\Sigma}^{-1/2}\mathbf{x}_1, \ldots, \mathbf{z}_n = \hat{\Sigma}^{-1/2}\mathbf{x}_n$, and then maximise

$$\ell^n(O, g_1, \ldots, g_d; \mathbf{z}_1, \ldots, \mathbf{z}_n)$$

over $O \in O(d)$ and $g_1, \ldots, g_d \in \mathcal{F}_1$. If $(\hat{O}^n, \hat{g}_1^n, \ldots, \hat{g}_d^n)$ is such a maximiser, we can then set $\hat{\hat{W}}^n = \hat{O}^n \hat{\Sigma}^{-1/2}$ and $\hat{\hat{f}}_j^n = \hat{g}_j^n$. Note that pre-whitening breaks down the estimation of the d^2 parameters in W into two stages: first, we use $\hat{\Sigma}$ to estimate the $d(d+1)/2$ free parameters of the symmetric, positive definite matrix Σ, leaving only the maximisation over the $d(d-1)/2$ free parameters of $O \in O(d)$ at the second stage. The advantage of this approach is that it facilitates more stable maximisation algorithms, such as the one described in the next subsection.

3.3. *Computational algorithm.* In this subsection, we address the challenge of maximising

$$\ell^n(W, f_1, \ldots, f_d; \mathbf{x}_1, \ldots, \mathbf{x}_n)$$

over $W \in O(d)$ and $f_1, \ldots, f_d \in \mathcal{F}_1$; thus, we are assuming that our data have already been pre-whitened. As a starting point, we choose W to be randomly distributed according to the Haar measure on the set $O(d)$ of $d \times d$ orthogonal matrices. A simple way of generating W with this distribution is to generate a $d \times d$ matrix Z whose entries are independent $N(0, 1)$ random variables, compute the QR-factorisation $Z = QR$, and let $W = Q$.

Our proposed algorithm then alternates between maximising the log-likelihood over $f_1, \ldots, f_d$ for fixed W, and then over W for fixed $f_1, \ldots, f_d$. The first of these steps is straightforward given Theorem 2 and the recent work on log-concave density estimation: we set f_j to be the log-concave maximum likelihood estimator of the data $w_j^\top \mathbf{x}_1, \ldots, w_j^\top \mathbf{x}_n$. This can be computed using the Active Set algorithm implemented in the `R` package `logcondens` [Dümbgen and Rufibach (2011), Rufibach and Dümbgen (2006)]. This fast algorithm exploits two basic facts: first, the logarithm of the log-concave maximum likelihood estimator is piecewise linear and continuous between the smallest and largest order statistics, with "knots" at the observations; and second, that the likelihood maximiser for a given (typically small) set of knots can be computed very efficiently. The algorithm therefore varies the set of knots appropriately until, after finitely many steps, the global optimum is attained.

This leaves the challenge of updating $W \in O(d)$. In common with other ICA algorithms [e.g., Plumbley (2005)], we treat $O(d)$ as a Riemannian manifold, and use standard techniques from differential geometry, as well as some features particular to our problem, to construct our proposal. Recall that the set $O(d)$ is a $d(d-1)/2$-dimensional submanifold of $\mathbb{R}^{d^2}$. The tangent space at $W \in O(d)$ is $T_W O(d) := \{WY : Y = -Y^\top\}$. In fact, if we define the natural inner product $\langle \cdot, \cdot \rangle$ on $T_W O(d) \times T_W O(d)$ by $\langle U, V \rangle = \operatorname{tr}(UV^\top)$, then $O(d)$ becomes a Riemannian manifold. (Note that if we think of U and V as vectors in $\mathbb{R}^{d^2}$, then this inner product is simply the Euclidean inner product.)

There is no loss of generality in assuming W belongs to the Riemannian manifold $\mathrm{SO}(d)$, the set of special orthogonal matrices having determinant 1. We can now define geodesics on $\mathrm{SO}(d)$, recalling that the matrix exponential is given by

$$\exp(Y) = I + \sum_{r=1}^{\infty} \frac{Y^r}{r!}.$$

The unique geodesic passing through $W \in \mathrm{SO}(d)$ with tangent vector WY (where $Y = -Y^\top$) is the map $\alpha : [0, 1] \to \mathrm{SO}(d)$ given by $\alpha(t) = W \exp(tY)$.

We update W by moving along a geodesic in $\mathrm{SO}(d)$, but need to choose an appropriate skew-symmetric matrix Y, which ideally should (at least locally)

give a large increase in the log-likelihood. The key to finding such a direction is Proposition 9 below. To set the scene for this result, observe that for $x \in [\min(w_j^\top \mathbf{x}_1, \ldots, w_j^\top \mathbf{x}_n), \max(w_j^\top \mathbf{x}_1, \ldots, w_j^\top \mathbf{x}_n)]$, we can write

$$\log f_j(x) = \min_{k=1,\ldots,m_j} (b_{jk}x - \beta_{jk}) \tag{6}$$

for some $b_{jk}, \beta_{jk} \in \mathbb{R}$ [e.g., Cule, Samworth and Stewart (2010)]. Since we may assume that $b_{j1}, \ldots, b_{jm_j}$ are strictly decreasing, the minimum in (6) is attained in either one or two indices. It is convenient to let

$$\mathcal{K}_{ij} = \operatorname*{arg\,min}_{k=1,\ldots,m_j} \bigl(b_{jk} w_j^\top \mathbf{x}_i - \beta_{jk}\bigr).$$

PROPOSITION 9. *Consider the map $g : \mathrm{SO}(d) \to \mathbb{R}$ given by*

$$g(W) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \min_{k=1,\ldots,m_j} \bigl(b_{jk} w_j^\top \mathbf{x}_i - \beta_{jk}\bigr).$$

Let Y be a skew-symmetric matrix and let c_j denote the jth row of WY. If $|\mathcal{K}_{ij}| = 1$, let k_{ij} denote the unique element of $\mathcal{K}_{ij}$. If $|\mathcal{K}_{ij}| = 2$, write $\mathcal{K}_{ij} = \{k_{ij1}, k_{ij2}\}$. If $c_j^\top \mathbf{x}_i \geq 0$, let $k_{ij} = k_{ijl}$, where $l = \arg\min_{l=1,2} b_{k_{ijl}}$; if $c_j^\top \mathbf{x}_i < 0$, let $k_{ij} = k_{ijl}$, where $l = \arg\max_{l=1,2} b_{k_{ijl}}$. Then the one-sided directional derivative of g at W in the direction WY is

$$\nabla_{WY} g(W) := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} b_{jk_{ij}} c_j^\top \mathbf{x}_i.$$

Note that while the one-sided directional derivatives of g exist, the function is not differentiable, so we cannot apply a basic gradient descent algorithm. Instead, for $1 < s < r < d$, let $Y_{r,s}$ denote the $d \times d$ matrix with $Y_{r,s}(r,s) = 1/\sqrt{2}$, $Y_{r,s}(s,r) = -1/\sqrt{2}$ and all other entries equal to zero. Then $\mathcal{Y}^+ = \{Y_{r,s} : 1 < s < r < d\}$ forms an orthonormal basis for the set of skew-symmetric matrices. Let $\mathcal{Y}^- = \{-Y : Y \in \mathcal{Y}^+\}$. We choose $Y^{\max} \in \mathcal{Y}^+ \cup \mathcal{Y}^-$ to maximise $\nabla_{WY} g(W)$.

We therefore update W with $W \exp(\varepsilon Y^{\max})$, and it remains to select ε. This we propose to choose by means of a backtracking line search. Specifically, we fix $\alpha \in (0,1)$ and $\varepsilon = 1$, and if

$$g\bigl(W \exp(\varepsilon Y^{\max})\bigr) > g(W) + \alpha\varepsilon \nabla_{WY^{\max}} g(W), \tag{7}$$

we accept a move from W to $W \exp(\varepsilon Y^{\max})$. Otherwise, we successively reduce ε by a factor of $\gamma \in (0,1)$ until (7) is satisfied, and then move to $W \exp(\varepsilon Y^{\max})$. In our implementation, we used $\alpha = 0.3$ and $\gamma = 1/2$.

Our algorithm produces a sequence $(W^{(1)}, f_1^{(1)}, \ldots, f_d^{(1)}), (W^{(2)}, f_1^{(2)}, \ldots, f_d^{(2)}), \ldots$. We terminate the algorithm once

$$\frac{\ell^n(W^{(t)}, f_1^{(t)}, \ldots, f_d^{(t)}) - \ell^n(W^{(t-1)}, f_1^{(t-1)}, \ldots, f_d^{(t-1)})}{|\ell^n(W^{(t-1)}, f_1^{(t-1)}, \ldots, f_d^{(t-1)})|} < \eta,$$

where, in our implementation, we chose $\eta = 10^{-7}$. As with other ICA algorithms, global convergence is not guaranteed, so we used 10 random starting points and took the solution with the highest log-likelihood.

4. Numerical experiments. To illustrate the practical merits of our proposed nonparametric maximum likelihood estimation method for ICA models, we conducted several sets of numerical experiments. To fix ideas, we focus on two-dimensional signals, that is, $d = 2$. The components of the signal were generated independently, and then rotated by $\pi/3$, so the mixing matrix is

$$A = \begin{pmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{pmatrix}.$$

Our goal is to reconstruct the signal and estimate A or, equivalently, $W = A^{-1}$, based on $n = 200$ observations of the rotated input.

We first consider a typical example in the ICA literature where the density of each component of the true signal is uniform on the interval $[-0.5, 0.5]$. The top left panel of Figure 1 plots the 200 simulated signal pairs, while the top right panel gives the rotated observations. The bottom left panel plots the recovered signal using the proposed nonparametric maximum likelihood method. Also included in the bottom right panel of the figure are the estimated marginal densities of the two sources of signal.

Figure 2 gives corresponding plots when the marginals have an $\mathrm{Exp}(1) - 1$ distribution. We note that both uniform and exponential distributions have log-concave densities and, therefore, our method not only recovers the mixing matrix but also accurately estimates the marginal densities, as can be seen in Figures 1 and 2.

To investigate the robustness of the proposed method when the marginal components do not have log-concave densities, we repeated the simulation in two other cases, with the true signal simulated first from a t-distribution with two degrees of freedom scaled by a factor of $1/\sqrt{2}$ and second from a mixture of normals distribution $0.7N(-0.9, 1) + 0.3N(2.1, 1)$. Figures 3 and 4 show that, in both cases, the misspecification of the marginals does not affect the recovery of the signal. Also, the estimated marginals represent estimates of the log-concave projection of the true marginals (a standard Laplace density in the case Figure 3), as correctly predicted by our theoretical results.

As discussed before, one of the unique advantages of the proposed method over existing ones is its general applicability. For example, the method can be used even

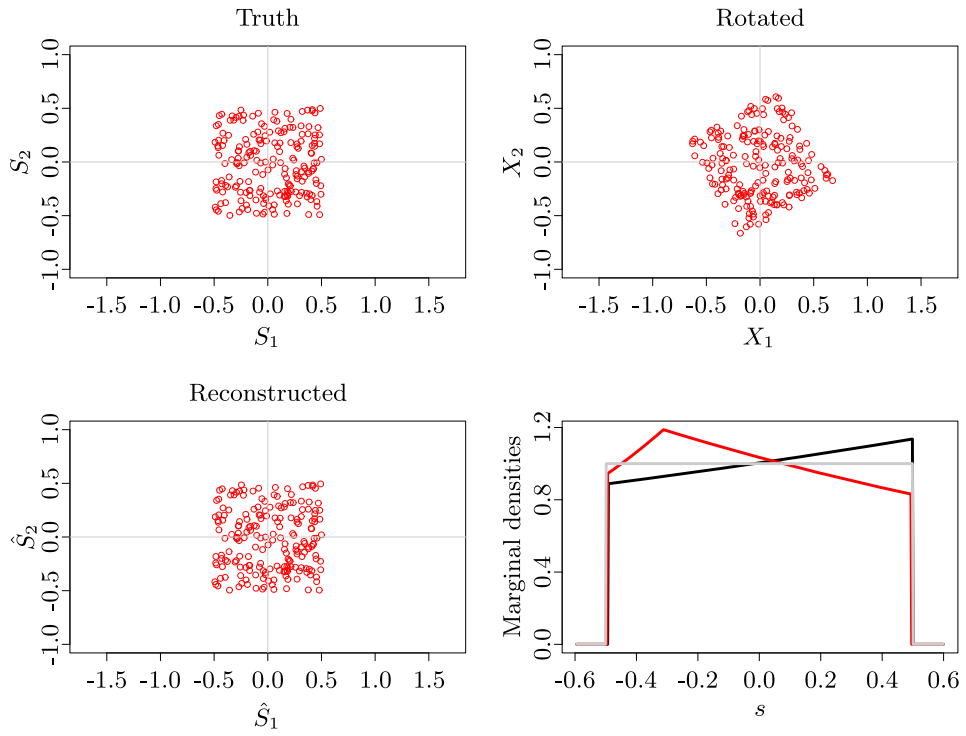

FIG. 1. *Uniform signal*: *top left panel*, *top right panel and bottom left panel give the true signal*, *rotated observations and the reconstructed signal*, *respectively. The bottom right panel gives the estimated marginal densities along with the true marginal* (*grey line*).

when the marginal distributions of the true signal do not have densities. To demonstrate this property, we now consider simulating signals from a $\mathrm{Bin}(3, 1/2) - 1.5$ distribution. To the best of our knowledge, none of the existing ICA methods are applicable for these types of problems. The simulation results presented in Figure 5 suggest that the method works very well in this case.

To further conduct a comparative study, we repeated each of the previous simulations 200 times and computed our estimate along with those produced by FastICA and ProDenICA methods. FastICA algorithms [e.g., Hyvärinen and Oja (2000), Nordhausen et al. (2011), Ollila (2010)] are popular ICA methods that traditionally proceed by maximising an approximation to the negentropy; ProDenICA is a nonparametric ICA method proposed by Hastie and Tibshirani (2003b), and has been shown to enjoy the best performance among a large collection of existing ICA methods [Hastie, Tibshirani and Friedman (2009)]. Both the FastICA and ProDenICA methods were implemented using the `R` package `ProDenICA` [Hastie and Tibshirani (2003a)]. In the former case, we used the `Gfunc=G1` option to the `ProDenICA` function, corresponding to cosh negentropy [Hyvärinen and Oja (2000)]; in the latter case, we used the `Gfunc=GPois`

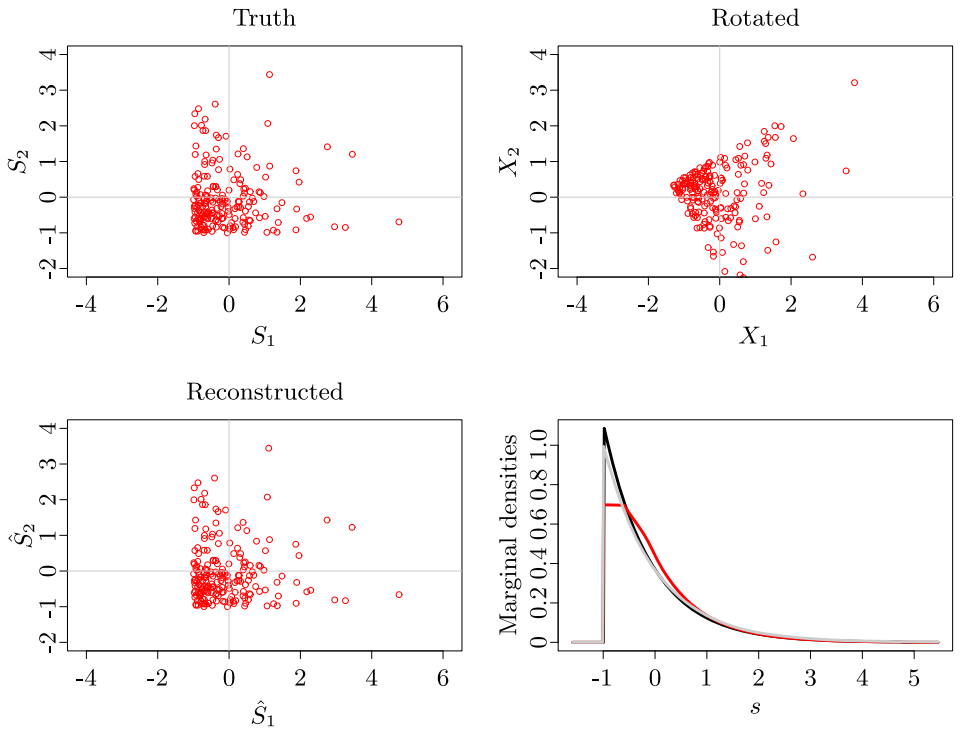

FIG. 2. *Exponential signal*: *top left panel*, *top right panel and bottom left panel give the true signal*, *rotated observations and the reconstructed signal*, *respectively. The bottom right panel gives the estimated marginal densities along with the true marginal* (*grey line*).

option, which fits a tilted Gaussian density using a Poisson generalised additive model. To compare the performance of these methods, we follow convention [Hyvärinen, Karhunen and Oja (2001)] and compute the Amari metric between the true unmixing matrix W and its estimates. The Amari metric between two $d \times d$ matrices is defined as

$$\rho(A, B) = \frac{1}{2d} \sum_{i=1}^{d} \left(\frac{\sum_{j=1}^{d} |C_{ij}|}{\max_{1 \le j \le d} |C_{ij}|} - 1 \right) + \frac{1}{2d} \sum_{j=1}^{d} \left(\frac{\sum_{i=1}^{d} |C_{ij}|}{\max_{1 \le i \le d} |C_{ij}|} - 1 \right), \tag{8}$$

where $C = (C_{ij})_{1 \le i,j \le d} = AB^{-1}$. Boxplots of the Amari metric for all three methods are given in Figure 6.

It is clear that both our proposed method (LogConICA) and ProDenICA outperform the FastICA method. For both uniform and exponential marginals, LogConICA improves upon ProDenICA, which might be expected since both distributions have log-concave densities. It is, however, interesting to note the robustness of LogConICA to misspecification of log-concavity, as it still outperforms ProDenICA for t_2 marginals, and remains competitive for the mixture of normal

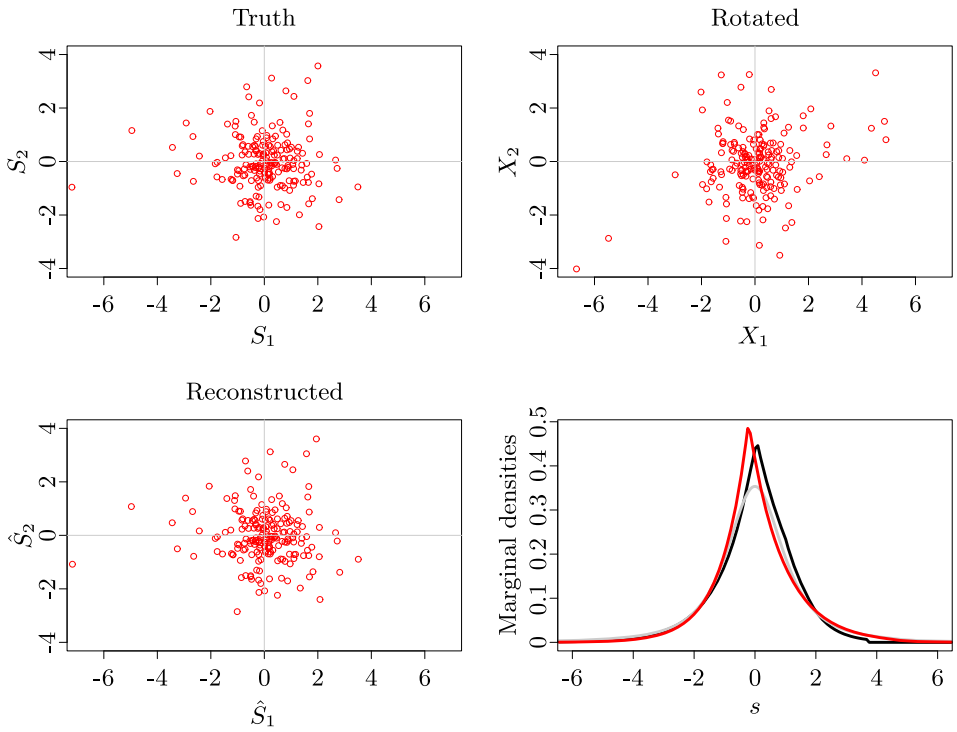

FIG. 3. *t_2 signal: top left panel, top right panel and bottom left panel give the true signal, rotated observations and the reconstructed signal, respectively. The bottom right panel gives the estimated marginal densities along with the true marginal (grey line).*

marginals. The most significant advantage of the proposed method, however, is displayed when the marginals are binomial. Recall that ProDenICA, in common with other nonparametric methods, assumes that the log density is smooth. This assumption is not satisfied with the binomial distribution and, as a result, ProDenICA performs rather poorly. In contrast, LogConICA works fairly well in this setting even though the true marginal does not have a log-concave density with respect to the Lebesgue measure. All these observations confirm our earlier theoretical development.

5. Proofs.

PROOF OF PROPOSITION 1. (1) Suppose that $\int_{\mathbb{R}^d} \|x\| \, dP(x) = \infty$. Fix an arbitrary $f \in \mathcal{F}_d^{\mathrm{ICA}}$, and find $\alpha > 0$ and $\beta \in \mathbb{R}$ such that $f(x) \le e^{-\alpha\|x\|+\beta}$. Then

$$\int_{\mathbb{R}^d} \log f \, dP \le -\alpha \int_{\mathbb{R}^d} \|x\| \, dP(x) + \beta = -\infty.$$

Thus, $L^{**}(P) = -\infty$ and $\psi^{**}(P) = \mathcal{F}_d^{\mathrm{ICA}}$.

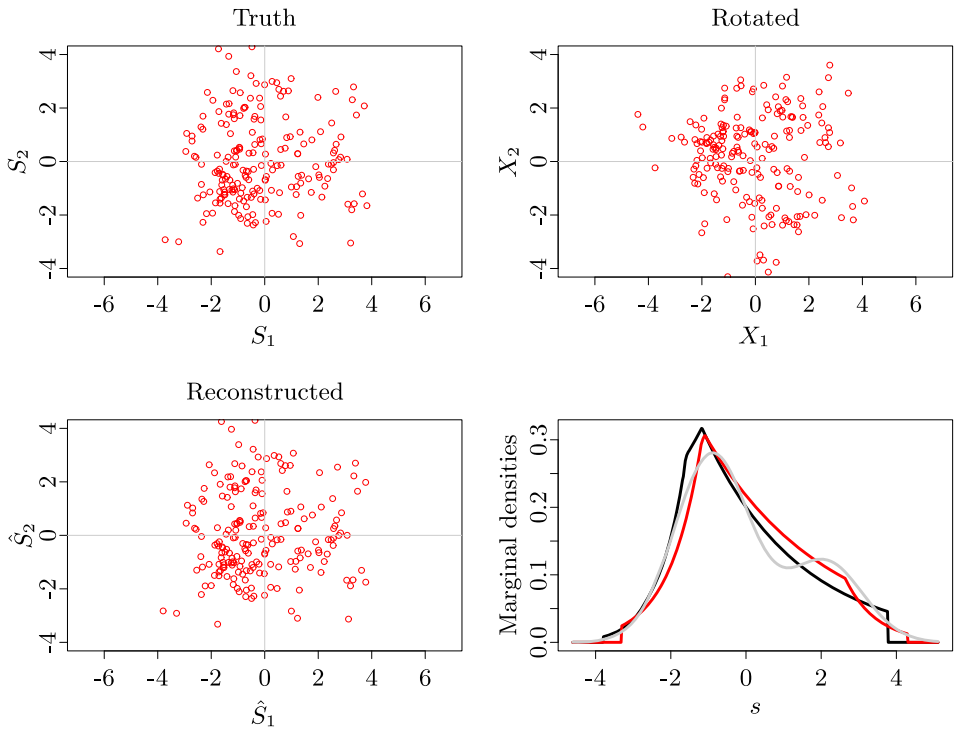

FIG. 4. *Mixture of normals signal*: *top left panel*, *top right panel and bottom left panel give the true signal*, *rotated observations and the reconstructed signal*, *respectively. The bottom right panel gives the estimated marginal densities along with the true marginal* (*grey line*).

(2) Now suppose that $\int_{\mathbb{R}^d} \|x\| \, dP(x) < \infty$, but $P(H) = 1$ for some hyperplane $H = \{x \in \mathbb{R}^d : a_1^\top x = \alpha\}$, where a_1 is a unit vector in $\mathbb{R}^d$ and $\alpha \in \mathbb{R}$. Find $a_2, \ldots, a_d$ such that $a_1, \ldots, a_d$ is an orthonormal basis for $\mathbb{R}^d$. Define the family of density functions

$$f_\sigma(x) = \frac{1}{2\sigma} e^{-|a_1^\top x - \alpha|/\sigma} \prod_{j=2}^{d} \frac{e^{-|a_j^\top x|}}{2}.$$

Then $f_\sigma \in \mathcal{F}_d^{\mathrm{ICA}}$, and

$$\begin{aligned}\int_{\mathbb{R}^d} \log f_\sigma(x) \, dP(x) &= -\log(\sigma) - d \log 2 - \sum_{j=2}^{d} \int_H |a_j^\top x| \, dP(x) \\ &\geq -\log(\sigma) - d \log 2 - \sum_{j=2}^{d} \int_H \|x\| \, dP(x) \to \infty\end{aligned}$$

as $\sigma \to 0$.

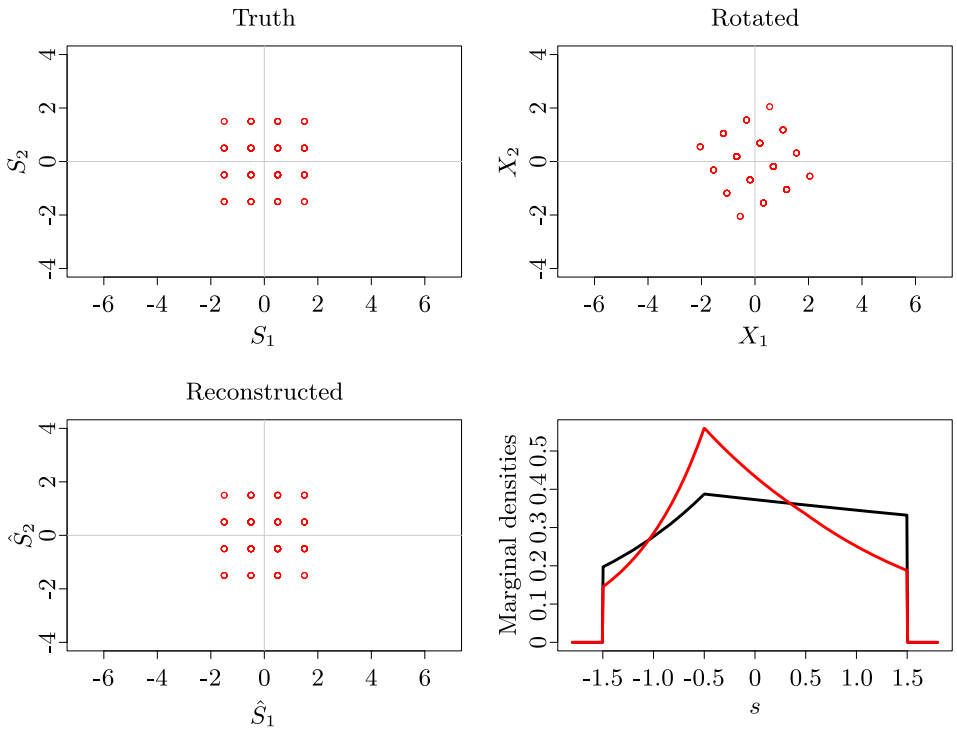

FIG. 5. *Binomial signal*: *top left panel*, *top right panel and bottom left panel give the true signal*, *rotated observations and the reconstructed signal*, *respectively. The bottom right panel gives the estimated marginal densities*.

(3) Now suppose that $P \in \mathcal{P}_d$. Notice that the density $f(x) = 2^{-d} \prod_{j=1}^{d} e^{-|x_j|}$ belongs to $\mathcal{F}_d^{\mathrm{ICA}}$ and satisfies

$$\int_{\mathbb{R}^d} \log f \, dP = -\sum_{j=1}^{d} \int_{\mathbb{R}^d} |x_j| \, dP(x) - d \log 2 > -\infty.$$

Moreover,

$$\sup_{f \in \mathcal{F}_d^{\mathrm{ICA}}} \int_{\mathbb{R}^d} \log f \, dP \le \sup_{f \in \mathcal{F}_d} \int_{\mathbb{R}^d} \log f \, dP < \infty,$$

where the second inequality follows from the proof of Theorem 2.2 of Dümbgen, Samworth and Schuhmacher (2011). We may therefore take a sequence $f^1, f^2, \ldots \in \mathcal{F}_d^{\mathrm{ICA}}$ such that

$$\int_{\mathbb{R}^d} \log f^n \, dP \nearrow \sup_{f \in \mathcal{F}_d^{\mathrm{ICA}}} \int_{\mathbb{R}^d} \log f \, dP.$$

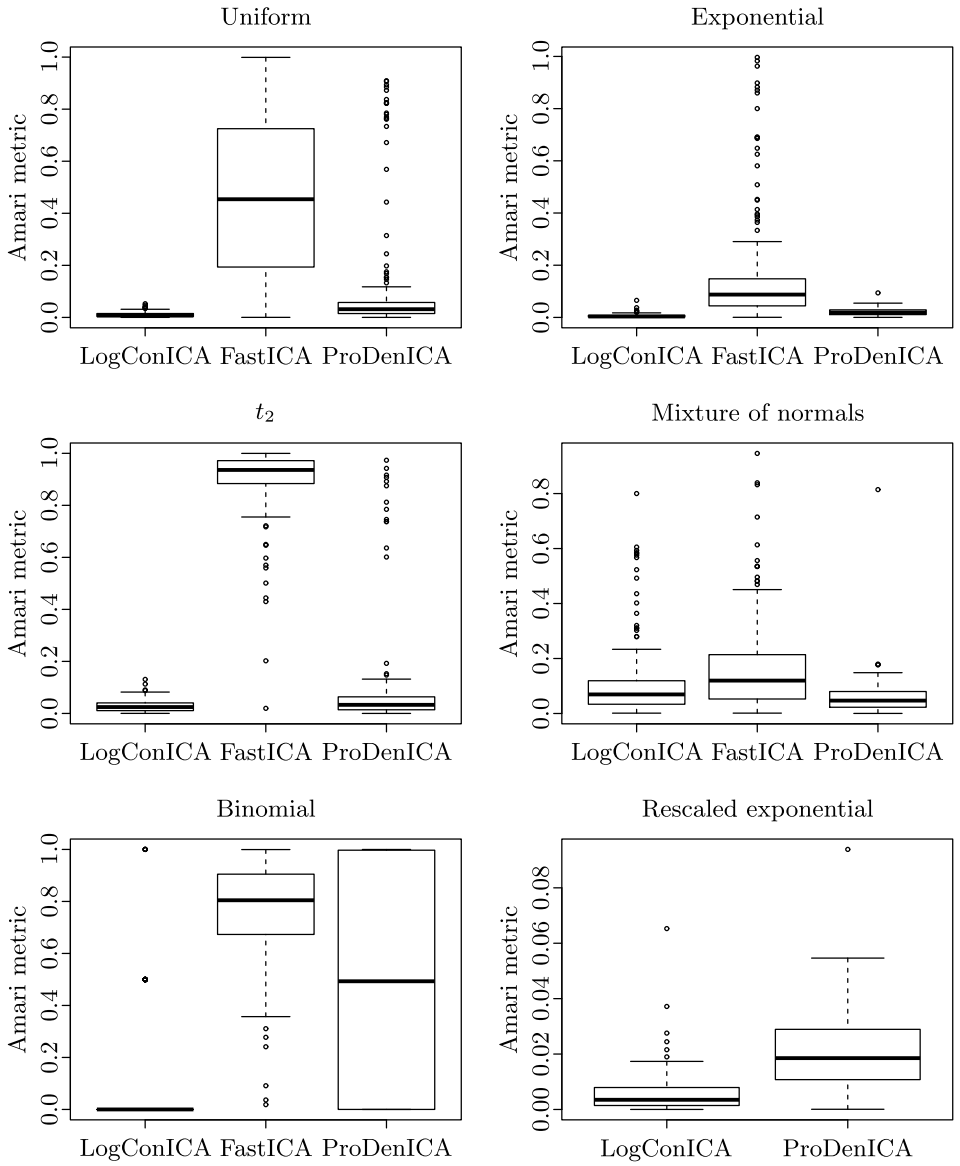

FIG. 6. *Comparison between LogConICA, FastICA and ProDenICA. The bottom right panel gives the Amari distances of the LogConICA and ProDenICA methods for the exponential example shown in the top right plot, but with a rescaled y-axis.*

Let csupp(P) denote the convex support of P, that is, the intersection of all closed, convex sets having P-measure 1. The hypothesis $P \in \mathcal{P}_d$ implies that csupp(P) is d-dimensional [e.g., Dümbgen, Samworth and Schuhmacher (2011), Lemma 2.1]. Following the arguments in the proof of Theorem 2.2 of Dümbgen,

Samworth and Schuhmacher (2011), there exist $\alpha > 0$ and $\beta \in \mathbb{R}$ such that $\sup_{n\in\mathbb{N}} f^n(x) \le e^{-\alpha\|x\|+\beta}$ for all $x \in \mathbb{R}^d$. Moreover, these arguments [see also the proof of Theorem 4 of Cule and Samworth (2010)] yield the existence of a closed, convex set $C \supseteq \mathrm{int}(\mathrm{csupp}(P))$, a log-concave density $f^{**} \in \mathcal{F}_d$ with $\{x \in \mathbb{R}^d : f^{**}(x) > 0\} = C$ and a subsequence (f^{n_k}) such that

$$f^{**}(x) = \lim_{k\to\infty} f^{n_k}(x) \qquad \text{for all } x \in \mathrm{int}(C) \cup (\mathbb{R}^d \setminus C).$$

Since the boundary of C has zero Lebesgue measure, we deduce from Fatou's lemma applied to the non-negative functions $x \mapsto -\alpha\|x\| + \beta - \log f^{n_k}(x)$ that

$$\int_{\mathbb{R}^d} \log f^{**}\,dP \ge \limsup_{k\to\infty} \int_{\mathbb{R}^d} \log f^{n_k}\,dP = \sup_{f\in\mathcal{F}_d^{\mathrm{ICA}}} \int_{\mathbb{R}^d} \log f\,dP.$$

It remains to show that $f^{**} \in \mathcal{F}_d^{\mathrm{ICA}}$. We can write

$$f^{n_k}(x) = |\det W^k| \prod_{j=1}^d f_j^k\big((w_j^k)^{\mathsf{T}} x\big),$$

where $W^k \in \mathcal{W}$ and $f_j^k \in \mathcal{F}_1$ for each $k \in \mathbb{N}$ and $j = 1, \dots, d$. Let X^k be a random vector with density $f^{n_k} \in \mathcal{F}_d^{\mathrm{ICA}}$, and let X be a random vector with density $f^{**} \in \mathcal{F}_d$. We know that $X^k \xrightarrow{d} X$ as $k \to \infty$, and that $(w_1^k)^{\mathsf{T}} X^k, \dots, (w_d^k)^{\mathsf{T}} X^k$ are independent for each k. Let $\tilde{w}_j^k = w_j^k/\|w_j^k\|$ and $\tilde{f}_j^k(x) = \|w_j^k\| f_j^k(\|w_j^k\| x)$. Then we have

$$f^{n_k}(x) = |\det \tilde{W}^k| \prod_{j=1}^d \tilde{f}_j^k\big((\tilde{w}_j^k)^{\mathsf{T}} x\big), \tag{9}$$

where the matrix $\tilde{W}^k$ has jth row $\tilde{w}_j^k$. Moreover, $\tilde{W}^k \in \mathcal{W}$ and $\tilde{f}_1^k, \dots, \tilde{f}_d^k \in \mathcal{F}_1$, so (9) provides an alternative, equivalent representation of the density f^{n_k}, in which each row of the unmixing matrix has unit Euclidean length. By reducing to a further subsequence if necessary, we may assume that for each $j = 1, \dots, d$, there exists $\tilde{w}_j \in \mathbb{R}^d$ such that $\tilde{w}_j^k \to \tilde{w}_j$ as $k \to \infty$. By Slutsky's theorem, it then follows that

$$\big((\tilde{w}_1^k)^{\mathsf{T}} X^k, \dots, (\tilde{w}_d^k)^{\mathsf{T}} X^k\big) \xrightarrow{d} (\tilde{w}_1^{\mathsf{T}} X, \dots, \tilde{w}_d^{\mathsf{T}} X).$$

Thus, for any $t = (t_1, \dots, t_d)^{\mathsf{T}} \in \mathbb{R}^d$,

$$\begin{aligned}\mathbb{E}\big(e^{it^{\mathsf{T}}(\tilde{w}_1^{\mathsf{T}} X, \dots, \tilde{w}_d^{\mathsf{T}} X)}\big) &= \lim_{k\to\infty} \mathbb{E}\big(e^{it^{\mathsf{T}}((\tilde{w}_1^k)^{\mathsf{T}} X^k, \dots, (\tilde{w}_d^k)^{\mathsf{T}} X^k)}\big) \\ &= \lim_{k\to\infty} \prod_{j=1}^d \mathbb{E}\big(e^{it_j(\tilde{w}_j^k)^{\mathsf{T}} X^k}\big) = \prod_{j=1}^d \mathbb{E}\big(e^{it_j \tilde{w}_j^{\mathsf{T}} X}\big).\end{aligned}$$

We conclude that $\tilde{w}_1^\mathsf{T} X, \ldots, \tilde{w}_d^\mathsf{T} X$ are independent. Now, the fact that the support of f^{**} is a d-dimensional convex set means that none of $\tilde{w}_1^\mathsf{T} X, \ldots, \tilde{w}_d^\mathsf{T} X$ is almost surely constant, and each of these random variables has a log-concave density, by Theorem 6 of Prékopa (1973). Finally, since $\|\tilde{w}_j\| = 1$ for all j, we deduce further that $\tilde{W} = (\tilde{w}_1, \ldots, \tilde{w}_d)^\mathsf{T}$ is non-singular. This shows that $f^{**} \in \mathcal{F}_d^{\mathrm{ICA}}$, as required. □

PROOF OF THEOREM 2. Suppose that $P \in \mathcal{P}_d^{\mathrm{ICA}}$ satisfies

$$P(B) = \prod_{j=1}^{d} P_j(w_j^\mathsf{T} B)$$

for some $W \in \mathcal{W}$ and $P_1, \ldots, P_d \in \mathcal{P}_1$. Consider maximising

$$\int_{\mathbb{R}^d} \log f(x)\, dP(x)$$

over $f \in \mathcal{F}_d$. Letting $s = Wx$ and $\tilde{f}(s) = f(As)$, where $A = W^{-1}$, we can equivalently maximise

$$\int_{\mathbb{R}^d} \log \tilde{f}(s)\, d\Big(\bigotimes_{j=1}^{d} P_j(s_j)\Big)$$

over $\tilde{f} \in \mathcal{F}_d$. But, by Theorem 4 of Chen and Samworth (2012), the unique solution to this maximisation problem is to choose $\tilde{f}(s) = \prod_{j=1}^d f_j^*(s_j)$, where $f_j^* = \psi^*(P_j)$. This shows that $f^* := \psi^*(P)$ can be written as

$$f^*(x) = |\det W| \prod_{j=1}^{d} f_j^*(w_j^\mathsf{T} x).$$

Since $f^* \in \mathcal{F}_d^{\mathrm{ICA}}$ also, we deduce that f^* is also the unique maximiser of $\int_{\mathbb{R}^d} \log f\, dP$ over $f \in \mathcal{F}_d^{\mathrm{ICA}}$, so $\psi^{**}(P) = \psi^*(P)$. □

PROOF OF THEOREM 3. Suppose that $P \in \mathcal{P}_d^{\mathrm{ICA}}$. Let $X \sim P$, so there exists $W \in \mathcal{W}$ such that WX has independent components. Writing P_j for the marginal distribution of $w_j^\mathsf{T} X$, note that $P_1, \ldots, P_d \in \mathcal{P}_1$. By Theorem 2 and the identifiability result of Eriksson and Koivunen (2004), it therefore suffices to show that $P_j \in \mathcal{P}_1$ has a Gaussian density if and only if $\psi^*(P_j)$ is a Gaussian density. If P_j has a Gaussian density f_j^*, then since f_j^* is log-concave, we have $f_j^* = \psi^*(P_j)$. Conversely, suppose that P_j does not have a Gaussian density. Since $f_j^* = \psi^*(P_j)$ satisfies $\int_{-\infty}^{\infty} x\, dP_j(x) = \int_{-\infty}^{\infty} x f_j^*(x)\, dx$ [Dümbgen, Samworth and Schuhmacher (2011), Remark 2.3], we may assume without loss of generality that P_j and f_j^* have mean zero. We consider maximising

$$\int_{-\infty}^{\infty} \log f\, dP_j$$

over all mean zero Gaussian densities f. Writing ϕ_{σ^2} for the mean zero Gaussian density with variance σ^2, we have

$$\int_{-\infty}^{\infty} \log \phi_{\sigma^2} \, dP_j = -\frac{1}{2\sigma^2} \int_{-\infty}^{\infty} x^2 \, dP_j(x) - \frac{1}{2} \log(2\pi\sigma^2).$$

This expression is maximised uniquely in σ^2 at $\sigma_*^2 = \int_{-\infty}^{\infty} x^2 \, dP_j(x)$. But Chen and Samworth (2012) show that the only way a distribution P_j and its log-concave projection $\psi^*(P_j)$ can have the same second moment is if P_j has a log-concave density, in which case P_j has density $\psi^*(P_j)$. We therefore conclude that the only way $\psi^*(P_j)$ can be a Gaussian density is if P_j has a Gaussian density, a contradiction. □

PROOF OF PROPOSITION 4. The proof of this proposition is very similar to the proof of Theorem 4.5 of Dümbgen, Samworth and Schuhmacher (2011), so we only sketch the argument here. For each $n \in \mathbb{N}$, let $f^n \in \psi^{**}(P^n)$, and consider an arbitrary subsequence (f^{n_k}). By reducing to a further subsequence if necessary, we may assume that $L^{**}(P^{n_k}) \to \lambda \in [-\infty, \infty]$. Observe that

$$\begin{aligned} \lambda &\geq \lim_{k\to\infty} \int_{\mathbb{R}^d} \log\bigl(2^{-d} e^{-\sum_{j=1}^d |x_j|}\bigr) \, dP^{n_k}(x) \\ &= -d \log 2 - \sum_{j=1}^d \int_{\mathbb{R}^d} |x_j| \, dP(x) > -\infty. \end{aligned}$$

Arguments from convex analysis can be used to show that the sequence (f^{n_k}) is uniformly bounded above, and $\liminf_{k\in\mathbb{N}} f^{n_k}(x_0) > -\infty$ for all $x_0 \in \operatorname{int}(\operatorname{csupp}(P))$. From this it follows that there exist $a > 0$ and $b \in \mathbb{R}$ such that $\sup_{k\in\mathbb{N}} \sup_{x\in\mathbb{R}^d} f^{n_k}(x) \leq e^{-a\|x\|+b}$. Thus, by reducing to a further subsequence if necessary, we may assume there exists $f^{**} \in \mathcal{F}_d$ such that

$$\begin{aligned} &\limsup_{k\to\infty, x\to x_0} f^{n_k}(x) = f^{**}(x_0) \qquad \text{for all } x_0 \in \mathbb{R}^d \setminus \partial\bigl\{x \in \mathbb{R}^d : f^{**}(x) > 0\bigr\}, \\ &\limsup_{k\to\infty, x\to x_0} f^{n_k}(x) \leq f^{**}(x_0) \qquad \text{for all } x_0 \in \partial\bigl\{x \in \mathbb{R}^d : f^{**}(x) > 0\bigr\}. \end{aligned} \tag{10}$$

Note from this that

$$\lambda = \lim_{k\to\infty} \int_{\mathbb{R}^d} \log f^{n_k} \, dP^{n_k} \leq -a \int_{\mathbb{R}^d} \|x\| \, dP(x) + b < \infty.$$

In fact, we can use the argument from the proof of Proposition 1 to deduce that $f^{**} \in \mathcal{F}_d^{\mathrm{ICA}}$. Skorokhod's representation theorem and Fatou's lemma can then be used to show that $\lambda \leq \int_{\mathbb{R}^d} \log f^{**} \, dP \leq L^{**}(P)$.

We can obtain the other bound $\lambda \geq L^{**}(P)$ by taking any element of $\psi^{**}(P)$, approximating it from above using Lipschitz continuous functions, as in the proof of Theorem 4.5 of Dümbgen, Samworth and Schuhmacher (2011), and using

monotone convergence. From these arguments, we conclude that $L^{**}(P^n) \to L^{**}(P)$ and $f^{**} \in \psi^{**}(P)$.

We can see from (10) that $f^{n_k} \xrightarrow{\text{a.e.}} f^{**}$, so $\int_{\mathbb{R}^d} |f^{n_k} - f^{**}| \to 0$, by Scheffé's theorem. Thus, given any $f^n \in \psi^{**}(P^n)$ and any subsequence (f^{n_k}), we can find $f^{**} \in \psi^{**}(P)$ and a further subsequence of (f^{n_k}) which converges to f^{**} in total variation distance. This yields the second part of the proposition. □

PROOF OF THEOREM 5. The first part of the theorem is a special case of Proposition 4. Now suppose $P \in \mathcal{P}_d^{\mathrm{ICA}}$ is identifiable and is represented by $W \in \mathcal{W}$ and $P_1, \ldots, P_d \in \mathcal{P}_1$. Suppose without loss of generality that $\|w_j\| = 1$ for all $j = 1, \ldots, d$ and let $f^{**} = \psi^{**}(P)$. Recall from Theorem 2 that if X has density f^{**}, then $w_j^{\mathsf{T}} X$ has density $f_j^* = \psi^*(P_j)$.

Suppose for a contradiction that we can find $\varepsilon > 0$, integers $1 \le n_1 < n_2 < \cdots$, $f^k \in \psi^{**}(P^{n_k})$ and $(W^k, f_1^k, \ldots, f_d^k) \overset{\mathrm{ICA}}{\sim} f^k$ such that

$$\inf_{k \in \mathbb{N}} \inf_{\varepsilon_j^k \in \mathbb{R} \setminus \{0\}} \inf_{\pi^k \in \Pi_d} \left\{ \left\| (\varepsilon_j^k)^{-1} w_{\pi^k(j)}^k - w_j \right\| + \int_{-\infty}^{\infty} \left| |\varepsilon_j^k| f_{\pi^k(j)}^k (\varepsilon_j^k x) - f_j^*(x) \right| dx \right\} \ge \varepsilon.$$

We can find a subsequence $1 \le k_1 < k_2 < \cdots$ such that $w_j^{k_l} / \|w_j^{k_l}\| \to \tilde{w}_j$, say, as $l \to \infty$, for all $j = 1, \ldots, d$. The argument toward the end of the proof of case (3) of Proposition 1 shows that $\tilde{W}$ can be used to represent the unmixing matrix of f^{**}, so by the identifiability result of Eriksson and Koivunen (2004) and the fact that $\|\tilde{w}_j\| = 1$, there exist $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_d \in \{-1, 1\}$ and a permutation π of $\{1, \ldots, d\}$ such that $\tilde{\varepsilon}_j \tilde{w}_{\pi(j)} = w_j$. Setting $\pi^n = \pi$ and $\varepsilon_j^n = \tilde{\varepsilon}_j^{-1} \|w_{\pi^n(j)}^n\|$, we deduce that

$$(\varepsilon_j^{k_l})^{-1} w_{\pi^{k_l}(j)}^{k_l} = \tilde{\varepsilon}_j \frac{w_{\pi(j)}^{k_l}}{\|w_{\pi(j)}^{k_l}\|} \to w_j$$

for $j = 1, \ldots, d$. Now observe that if X^{k_l} has density f^{k_l}, then by Slutsky's theorem, $(\varepsilon_j^{k_l})^{-1} (w_{\pi^{k_l}(j)}^{k_l})^{\mathsf{T}} X^{k_l} \xrightarrow{d} w_j^{\mathsf{T}} X$. It therefore follows from Proposition 2(c) of Cule and Samworth (2010) that

$$\int_{-\infty}^{\infty} \left| |\varepsilon_j^{k_l}| f_{\pi^{k_l}(j)}^n (\varepsilon_j^{k_l} x) - f_j^*(x) \right| dx \to 0$$

for $j = 1, \ldots, d$. This contradiction establishes that

$$\sup_{f^n \in \psi^{**}(P^n)} \sup_{(W^n, f_1^n, \ldots, f_d^n) \overset{\mathrm{ICA}}{\sim} f^n} \inf_{\pi^n \in \Pi_d} \inf_{\varepsilon_1^n, \ldots, \varepsilon_d^n \in \mathbb{R} \setminus \{0\}} \left\{ \left\| (\varepsilon_j^n)^{-1} w_{\pi^n(j)}^n - w_j \right\| + \int_{-\infty}^{\infty} \left| |\varepsilon_j^n| f_{\pi^n(j)}^n (\varepsilon_j^n x) - f_j^*(x) \right| dx \right\} \to 0 \tag{11}$$

for each $j = 1, \ldots, d$.

It remains to prove that for sufficiently large n, every $f^n \in \psi^{**}(P^n)$ is identifiable. Recall from the identifiability result of Eriksson and Koivunen (2004) and Theorem 3 that not more than one of $f_1^*, \ldots, f_d^*$ is Gaussian. Let $\phi_{\mu,\sigma^2}(\cdot)$ denote the univariate normal density with mean μ and variance σ^2. Let J denote the index set of the non-Gaussian densities among $f_1^*, \ldots, f_d^*$, so the cardinality of J is at least $d-1$, and consider, for each $j \in J$, the problem of minimising $g(\mu,\sigma) = \int_{-\infty}^{\infty} |\phi_{\mu,\sigma^2} - f_j^*|$ over $\mu \in \mathbb{R}$ and $\sigma > 0$. Observe that g is continuous with $g(\mu,\sigma) < 2$ for all μ and σ, that $\inf_{\mu\in\mathbb{R}} g(\mu,\sigma) \to 2$ as $\sigma \to 0, \infty$ and $\inf_{\sigma>0} g(\mu,\sigma) \to 2$ as $|\mu| \to \infty$. It follows that g attains its infimum, and there exists $\eta > 0$ such that

$$\inf_{j\in J} \inf_{\mu\in\mathbb{R}} \inf_{\sigma>0} \int_{-\infty}^{\infty} |\phi_{\mu,\sigma^2} - f_j^*| \geq \eta. \tag{12}$$

Comparing (11) and (12), we see that, for sufficiently large n, whenever $f^n \in \psi^{**}(P^n)$ and $(W^n, f_1^n, \ldots, f_d^n) \overset{\mathrm{ICA}}{\sim} f^n$, at most one of the densities $f_1^n, \ldots, f_d^n$ can be Gaussian. It follows that when n is large, every $f^n \in \psi^{**}(P^n)$ is identifiable. $\square$

PROOF OF PROPOSITION 6. It is well known that for fixed $W \in \mathcal{W}$, the nonparametric likelihood $L(\cdot)$ defined in (4) is maximised by choosing

$$\hat{P}_j^W = \frac{1}{n}\sum_{i=1}^{n} \delta_{w_j^\top \mathbf{x}_i}, \qquad j = 1, \ldots, d.$$

For $i = 1, \ldots, n$, $W \in \mathcal{W}$ and $j = 1, \ldots, d$, let

$$n_{w_j}(i) = \{\tilde{i} \in \{1, \ldots, n\} : w_j^\top \mathbf{x}_{\tilde{i}} = w_j^\top \mathbf{x}_i\}.$$

The binary relation $i \sim \tilde{i}$ if $n_{w_j}(i) = n_{w_j}(\tilde{i})$ defines an equivalence relation on $\{1, \ldots, n\}$, so we can let I_j^W denote a set of indices obtained by choosing one element from each equivalence class. Then

$$\begin{aligned} L(W, \hat{P}_1^W, \ldots, \hat{P}_d^W) &= \prod_{j=1}^{d} \frac{|n_{w_j}(1)||n_{w_j}(2)| \cdots |n_{w_j}(n)|}{n^n} \\ &= \prod_{j=1}^{d} n^{-n} \prod_{i\in I_j^W} |n_{w_j}(i)|^{|n_{w_j}(i)|}. \end{aligned}$$

Note that $|n_{w_j}(i)| \leq d$, because otherwise there would exist a subset of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of cardinality at least $d+1$ lying on a $(d-1)$-dimensional hyperplane of the form $\{x \in \mathbb{R}^d : w_j^\top x = \beta\}$, for some $\beta \in \mathbb{R}$, contradicting the hypothesis that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are in general position. In fact, we claim that $\sum_{i\in I_j^W} (|n_{w_j}(i)| - 1) \leq d-1$. Indeed,

suppose for a contradiction that this were not the case. Then, by reordering the observations if necessary, we could find $K \leq |I_j^W|$, integers $n_1, \ldots, n_K \geq 2$ with $n_1 + \cdots + n_K \geq d + K$ and $\beta_1, \ldots, \beta_K \in \mathbb{R}$ such that $w_j^\mathsf{T} \mathbf{x}_i = \beta_k$ for $i = n_1 + \cdots + n_{k-1} + 1, \ldots, n_1 + \cdots + n_k$ and $k = 1, \ldots, K$. Now set $\mathbf{y}_{ik} = \mathbf{x}_i - \mathbf{x}_{n_1+\cdots+n_k}$ for $i = n_1 + \cdots + n_{k-1} + 1, \ldots, n_1 + \cdots + n_k - 1$ and $k = 1, \ldots, K$. Note that there at least d vectors $\{\mathbf{y}_{ik}\}$, all of which are non-zero, and any subset of cardinality d is linearly independent because $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are in general position. On the other hand, $w_j^\mathsf{T} \mathbf{y}_{ik} = 0$ for all i, k, which establishes our contradiction. Thus, the best we can hope for in aiming to maximise the likelihood is to be able to choose $|n_{w_j}(i_0)| = d$ for some $i_0 \in I_j^W$, and $|n_{w_j}(i)| = 1$ for $i \in I_j^W \setminus \{i_0\}$. It follows that $L(W, \hat{P}_1^W, \ldots, \hat{P}_d^W) \leq (d^d/n^n)^d$.

Moreover, for any choice J of distinct indices in $\{1, \ldots, n\}$, if we construct the matrix $W_J \in \mathcal{W}$ as described just before the statement of Proposition 6, then for each $j = 1, \ldots, d$ and $i \in J \setminus \{j\}$, we have $w_j^\mathsf{T} \mathbf{x}_i = \mathbf{1}_d^\mathsf{T} \mathbf{X}_{(-j)}^{-1} \mathbf{x}_i = 1$, so $|n_{w_j}(i)| = d$ for such i, and $L(W_J, \hat{P}_1^{W_J}, \ldots, \hat{P}_d^{W_J}) = (d^d/n^n)^d$. $\square$

As a preliminary to the proof of Corollary 8, it is convenient to define some more notation. Let $\mathcal{F}_{d,0}^{\mathrm{ICA}}$ denote the set of $f \in \mathcal{F}_d^{\mathrm{ICA}}$ that can be represented using an orthogonal unmixing matrix. Let $\psi_0^{**} : \mathcal{P}_d \to \mathcal{F}_{d,0}^{\mathrm{ICA}}$ denote the log-concave ICA projection operator onto $\mathcal{F}_{d,0}^{\mathrm{ICA}}$, so that

$$\psi_0^{**}(P) = \operatorname*{arg\,max}_{f \in \mathcal{F}_{d,0}^{\mathrm{ICA}}} \int_{\mathbb{R}^d} \log f \, dP.$$

The projection operator ψ_0^{**} has many properties in common with ψ^{**}. Indeed, the analogue of Proposition 1 is immediate (in fact, the proof is slightly simpler because all rows of unmixing matrices have unit length). Let $\mathcal{P}_{d,0}^{\mathrm{ICA}}$ denote the set of $P \in \mathcal{P}_d^{\mathrm{ICA}}$ such that any $X \sim \mathcal{P}_d^{\mathrm{ICA}}$ has identity covariance matrix. Then the analogue of Theorem 2 states that the restriction $\psi_0^{**}|_{\mathcal{P}_{d,0}^{\mathrm{ICA}}}$ coincides with $\psi^{**}|_{\mathcal{P}_{d,0}^{\mathrm{ICA}}}$. This follows because if the distribution of X belongs to $\mathcal{P}_{d,0}^{\mathrm{ICA}}$, so that $X = AS$, where S has independent components, then there is no loss of generality in assuming each component of S has unit variance, and then $I = \mathrm{Cov}(X) = AA^\mathsf{T}$. Thus, the unmixing matrix of P may be assumed to be orthogonal, and the result follows from Theorem 2. Analogues of Theorem 3 and Proposition 4 for ψ_0^{**} are immediate.

The proof of Corollary 8 is based on an analogue of part of Theorem 5 for $\mathcal{P}_{d,0}^{\mathrm{ICA}}$, which is stated below. Its proof is virtually identical to that of Theorem 5, and is omitted.

PROPOSITION 10. *Suppose that $P \in \mathcal{P}_{d,0}^{\mathrm{ICA}}$ is identifiable and that $(W, P_1, \ldots, P_d) \overset{\mathrm{ICA}}{\sim} P$ with $W \in O(d)$. If $P^1, P^2, \ldots \in \mathcal{P}_d$ are such that $d(P^n, P) \to 0$, then*

$$\sup_{f^n \in \psi_0^{**}(P^n)} \sup_{(W^n, f_1^n, \ldots, f_d^n) \overset{\mathrm{ICA}}{\sim} f^n} \inf_{\pi^n \in \Pi_d} \inf_{\varepsilon_1^n, \ldots, \varepsilon_d^n \in \{-1,1\}} \biggl\{ \bigl\| (\varepsilon_j^n)^{-1} w_{\pi^n(j)}^n - w_j \bigr\| + \int_{-\infty}^{\infty} \bigl| |\varepsilon_j^n| f_{\pi^n(j)}^n(\varepsilon_j^n x) - f_j^*(x) \bigr| \, dx \biggr\} \to 0$$

for each $j = 1, \ldots, d$, where $f_j^* = \psi^*(P_j)$. As a consequence, for sufficiently large n, every $f^n \in \psi_0^{**}(P^n)$ is identifiable.

Notice that the scaling factors here may be assumed to belong to $\{-1, 1\}$, again because the rows of W^n and W have unit length.

PROOF OF COROLLARY 8. Let $\mathbf{z}_1 = \hat{\Sigma}^{-1/2}\mathbf{x}_1, \ldots, \mathbf{z}_n = \hat{\Sigma}^{-1/2}\mathbf{x}_n$, and let $\hat{P}^{n,\mathbf{z}}$ denote their empirical distribution. Writing $\bar{\mathbf{z}} = n^{-1}\sum_{i=1}^n \mathbf{z}_i$ and $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i$, note that the covariance matrix corresponding to $\hat{P}^{n,\mathbf{z}}$ is

$$\frac{1}{n}\sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^{\mathsf{T}} = \frac{1}{n}\sum_{i=1}^n \hat{\Sigma}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathsf{T}}\hat{\Sigma}^{-1/2} = I.$$

Now $\hat{P}^{n,\mathbf{z}} \in \mathcal{P}_d$ provided the convex hull of $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is d-dimensional, which occurs with probability 1 for sufficiently large n. It follows by the analogue of Proposition 1 for ψ_0^{**} that there then exists a maximiser $(\hat{O}^n, \hat{g}_1^n, \ldots, \hat{g}_d^n)$ of $\ell^n(O, g_1, \ldots, g_d; \mathbf{z}_1, \ldots, \mathbf{z}_n)$ over $O \in O(d)$ and $g_1, \ldots, g_d \in \mathcal{F}_1$.

Now let $\Sigma = \mathrm{Cov}(\mathbf{x}_1)$, and note that the distribution $P^{0,\mathbf{z}}$ of $\Sigma^{-1/2}\mathbf{x}_1$ belongs to $\mathcal{P}_{d,0}^{\mathrm{ICA}}$. Suppose that $(O^0, P_1^{0,\mathbf{z}}, \ldots, P_d^{0,\mathbf{z}}) \overset{\mathrm{ICA}}{\sim} P^{0,\mathbf{z}}$, with $O^0 \in O(d)$. To show that $d(\hat{P}^{n,\mathbf{z}}, P^{0,\mathbf{z}}) \overset{\mathrm{a.s.}}{\to} 0$, note first that

$$\begin{aligned}
&\biggl| \int_{\mathbb{R}^d} \|x\| \, d\bigl(\hat{P}^{n,\mathbf{z}} - P^{0,\mathbf{z}}\bigr)(x) \biggr| \\
&\quad \le \biggl| \frac{1}{n}\sum_{i=1}^n \bigl\{ \bigl\| \hat{\Sigma}^{-1/2}\mathbf{x}_i \bigr\| - \bigl\| \Sigma^{-1/2}\mathbf{x}_i \bigr\| \bigr\} \biggr| + \biggl| \frac{1}{n}\sum_{i=1}^n \bigl\| \Sigma^{-1/2}\mathbf{x}_i \bigr\| - \mathbb{E}\bigl\| \Sigma^{-1/2}\mathbf{x}_1 \bigr\| \biggr| \\
&\quad \le \frac{1}{n}\sum_{i=1}^n \bigl\| \bigl(\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}\bigr)\mathbf{x}_i \bigr\| + \biggl| \frac{1}{n}\sum_{i=1}^n \bigl\| \Sigma^{-1/2}\mathbf{x}_i \bigr\| - \mathbb{E}\bigl\| \Sigma^{-1/2}\mathbf{x}_1 \bigr\| \biggr| \\
&\quad \le |\lambda|_{\max}\bigl(\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}\bigr) \frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i\| + \biggl| \frac{1}{n}\sum_{i=1}^n \bigl\| \Sigma^{-1/2}\mathbf{x}_i \bigr\| - \mathbb{E}\bigl\| \Sigma^{-1/2}\mathbf{x}_1 \bigr\| \biggr|,
\end{aligned}$$

where $|\lambda|_{\max}(\hat{\Sigma}^{-1/2} - \Sigma^{-1/2})$ denotes the largest absolute value of the eigenvalues of $\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}$. Now $|\lambda|_{\max}(\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}) \xrightarrow{\text{a.s.}} 0$, by the strong law of large numbers and the continuous mapping theorem, while $n^{-1}\sum_{i=1}^n \|\mathbf{x}_i\| \xrightarrow{\text{a.s.}} \mathbb{E}\|\mathbf{x}_1\|$, also by the strong law. Another application of the strong law shows that the second term in the sum converges almost surely to zero. Moreover, if $h : \mathbb{R}^d \to [-1, 1]$ has Lipschitz constant at most 1, then

$$\begin{aligned}
&\left|\int_{\mathbb{R}^d} h\, d\big(\hat{P}^{n,\mathbf{z}} - P^{0,\mathbf{z}}\big)\right| \\
&\quad\le \left|\frac{1}{n}\sum_{i=1}^n \big\{h\big(\hat{\Sigma}^{-1/2}\mathbf{x}_i\big) - h\big(\Sigma^{-1/2}\mathbf{x}_i\big)\big\}\right| \\
&\qquad + \left|\frac{1}{n}\sum_{i=1}^n h\big(\Sigma^{-1/2}\mathbf{x}_i\big) - \mathbb{E}h\big(\Sigma^{-1/2}\mathbf{x}_1\big)\right| \\
&\quad\le |\lambda|_{\max}\big(\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}\big)\frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i\| \\
&\qquad + \left|\frac{1}{n}\sum_{i=1}^n h\big(\Sigma^{-1/2}\mathbf{x}_i\big) - \mathbb{E}h\big(\Sigma^{-1/2}\mathbf{x}_1\big)\right| \\
&\quad\xrightarrow{\text{a.s.}} 0.
\end{aligned}$$

This shows that $d(\hat{P}^{n,\mathbf{z}}, P^{0,\mathbf{z}}) \xrightarrow{\text{a.s.}} 0$, so by Proposition 10, there exist $\hat{\pi}^n \in \Pi_d$ and $\hat{\varepsilon}_1^n, \ldots, \hat{\varepsilon}_d^n \in \{-1, 1\}$ such that

$$\big\|\big(\hat{\varepsilon}_j^n\big)^{-1}\hat{o}^n_{\hat{\pi}^n(j)} - o_j^0\big\| + \int_{-\infty}^{\infty} \big||\hat{\varepsilon}_j^n|\hat{g}^n_{\hat{\pi}^n(j)}\big(\hat{\varepsilon}_j^n x\big) - g_j^*(x)\big|\, dx \xrightarrow{\text{a.s.}} 0$$

for each $j = 1, \ldots, d$, where $g_j^* = \psi^*(P_j^{0,\mathbf{z}})$. Now set $\hat{\hat{W}}^n = \hat{O}^n\hat{\Sigma}^{-1/2}$ and $\hat{\hat{f}}_j^n = \hat{g}_j^n$, and observe that $(\hat{\hat{W}}^n, \hat{\hat{f}}_1^n, \ldots, \hat{\hat{f}}_d^n)$ maximises $\ell^n(W, f_1, \ldots, f_d; \mathbf{x}_1, \ldots, \mathbf{x}_n)$ over $W \in O(d)\hat{\Sigma}^{-1/2}$ and $f_1, \ldots, f_d \in \mathcal{F}_1$. Since $(O^0\Sigma^{-1/2}, P_1^{0,\mathbf{z}}, \ldots, P_d^{0,\mathbf{z}}) \overset{\text{ICA}}{\sim} P^0$, there exist $\pi \in \Pi_d$ and scaling factors $\varepsilon_1, \ldots, \varepsilon_d \in \mathbb{R} \setminus \{0\}$ such that $o_j^0\Sigma^{-1/2} = \varepsilon_j^{-1}w^0_{\pi(j)}$ and $P_j^{0,\mathbf{z}}(B_j) = P^0_{\pi(j)}(\varepsilon_j B_j)$ for all $B_j \in \mathcal{B}_1$. It follows that, setting $\hat{\hat{\pi}} = \hat{\pi}^n \circ \pi^{-1}$ and $\hat{\hat{\varepsilon}}_j^n = \varepsilon^{-1}_{\pi^{-1}(j)}\hat{\varepsilon}^n_{\pi^{-1}(j)}$, we have

$$\begin{aligned}
\big(\hat{\hat{\varepsilon}}_j^n\big)^{-1}\hat{\hat{w}}^n_{\hat{\hat{\pi}}^n(j)} &= \big(\hat{\varepsilon}_j^n\big)^{-1}\hat{o}^n_{\hat{\pi}^n(j)}\hat{\Sigma}^{-1/2} \\
&= \varepsilon_{\pi^{-1}(j)}\big(\hat{\varepsilon}^n_{\pi^{-1}(j)}\big)^{-1}\hat{o}^n_{\hat{\pi}^n(\pi^{-1}(j))}\hat{\Sigma}^{-1/2} \\
&\xrightarrow{\text{a.s.}} \varepsilon_{\pi^{-1}(j)}o^0_{\pi^{-1}(j)}\Sigma^{-1/2} = w_j^0
\end{aligned}$$

for $j = 1, \ldots, d$. Now,

$$f^*_{\pi(j)}(x) = \psi^*\big(P^0_{\pi(j)}\big)(x) = |\varepsilon_j^{-1}| g^*_j(\varepsilon_j^{-1}x),$$

where the second equality follows by the affine equivariance of log-concave projections [Dümbgen, Samworth and Schuhmacher (2011), Remark 2.4]. Thus,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \big||\hat{\varepsilon}^n_j| \hat{f}^n_{\hat{\pi}^n(j)}(\hat{\varepsilon}^n_j x) - f^*_j(x)\big|\,dx \\
&\quad = \int_{-\infty}^{\infty} \big||\hat{\varepsilon}^n_{\pi^{-1}(j)}\varepsilon^{-1}_{\pi^{-1}(j)}| \hat{g}^n_{\hat{\pi}^n(\pi^{-1}(j))}\big(\hat{\varepsilon}^n_{\pi^{-1}(j)}\varepsilon_{\pi^{-1}(j)}x\big) \\
&\qquad\qquad - |\varepsilon^{-1}_{\pi^{-1}(j)}| g^*_{\pi^{-1}(j)}\big(\varepsilon^{-1}_{\pi^{-1}(j)}x\big)\big|\,dx \\
&\quad = \int_{-\infty}^{\infty} \big||\hat{\varepsilon}^n_{\pi^{-1}(j)}| \hat{g}^n_{\hat{\pi}^n(\pi^{-1}(j))}\big(\hat{\varepsilon}^n_{\pi^{-1}(j)}y\big) - g^*_{\pi^{-1}(j)}(y)\big|\,dy \xrightarrow{\text{a.s.}} 0
\end{aligned}$$

as required. □

PROOF OF PROPOSITION 9. For $\varepsilon > 0$, let $W_\varepsilon = W\exp(\varepsilon Y)$, and let $w_{j,\varepsilon}$ denote the jth row of W_ε. Notice that

$$w^{\mathsf{T}}_{j,\varepsilon}\mathbf{x}_i = w^{\mathsf{T}}_j\mathbf{x}_i + \varepsilon c^{\mathsf{T}}_j\mathbf{x}_i + O\big(\varepsilon^2\big)$$

as $\varepsilon \searrow 0$. It follows that for sufficiently small $\varepsilon > 0$,

$$\begin{aligned}
\frac{g(W_\varepsilon) - g(W)}{\varepsilon} &= \frac{1}{\varepsilon}\sum_{i=1}^n \sum_{j=1}^d \Big\{ \min_{k=1,\ldots,m_j} \big(b_{jk} w^{\mathsf{T}}_{j,\varepsilon}\mathbf{x}_i - \beta_{jk}\big) \\
&\qquad\qquad - \min_{k=1,\ldots,m_j} \big(b_{jk} w^{\mathsf{T}}_j\mathbf{x}_i - \beta_{jk}\big)\Big\} \\
&= \frac{1}{\varepsilon}\sum_{i=1}^n \sum_{j=1}^d b_{jk_{ij}}\big(w^{\mathsf{T}}_{j,\varepsilon}\mathbf{x}_i - w^{\mathsf{T}}_j\mathbf{x}_i\big) \\
&\to \sum_{i=1}^n \sum_{j=1}^d b_{jk_{ij}} c^{\mathsf{T}}_j\mathbf{x}_i
\end{aligned}$$

as $\varepsilon \searrow 0$. □

Acknowledgements. We are very grateful to the Associate Editor and three anonymous referees for their helpful comments and suggestions.

REFERENCES

BACH, F. R. and JORDAN, M. I. (2002). Kernel independent component analysis. *J. Mach. Learn. Res.* **3** 1–48. MR1966051

CHEN, A. and BICKEL, P. J. (2005). Consistent independent component analysis and prewhitening. *IEEE Trans. Signal Process.* **53** 3625–3632. MR2239886

CHEN, A. and BICKEL, P. J. (2006). Efficient independent component analysis. *Ann. Statist.* **34** 2825–2855. MR2329469

CHEN, Y. and SAMWORTH, R. J. (2012). Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sinica*. To appear.

COMON, P. (1994). Independent component analysis, a new concept? *Signal Proc.* **36** 287–314.

CULE, M. and SAMWORTH, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat.* **4** 254–270. MR2645484

CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 545–607. MR2758237

DÜMBGEN, L. and RUFIBACH, K. (2011). `logcondens`: Computations related to univariate log-concave density estimation. *J. Statist. Software* **39** 1–28.

DÜMBGEN, L., SAMWORTH, R. and SCHUHMACHER, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.* **39** 702–730. MR2816336

ERIKSSON, J. and KOIVUNEN, V. (2004). Identifiability, separability and uniqueness of linear ICA models. *IEEE Signal Processing Letters* **11** 601–604.

HASTIE, T. and TIBSHIRANI, R. (2003a). `ProDenICA`: Product density estimation for ICA using tilted Gaussian density estimates `R` package version 1.0. Available at http://cran.r-project.org/web/packages/ProDenICA/.

HASTIE, T. and TIBSHIRANI, R. (2003b). Independent component analysis through product density estimation. In *Advances in Neural Information Processing Systems* 15 (S. Becker and K. Obermayer, eds.) 649–656. MIT Press, Cambridge, MA.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

HYVÄRINEN, A., KARHUNEN, J. and OJA, E. (2001). *Independent Component Analysis*. Wiley, New York.

HYVÄRINEN, A. and OJA, E. (2000). Independent component analysis: Algorithms and applications. *Neural Netw.* **13** 411–430.

ILMONEN, P., NEVALAINEN, J. and OJA, H. (2010). Characteristics of multivariate distributions and the invariant coordinate system. *Statist. Probab. Lett.* **80** 1844–1853. MR2734250

ILMONEN, P. and PAINDAVEINE, D. (2011). Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *Ann. Statist.* **39** 2448–2476. MR2906874

KARVANEN, J. and KOIVUNEN, V. (2002). Blind separation methods based on Pearson system and its extensions. *Signal Proc.* **82** 663–673.

NORDHAUSEN, K., OJA, H. and OLLILA, E. (2011). Multivariate models and the first four moments. In *Nonparametric Statistics and Mixture Models* 267–287. World Sci. Publ., Hackensack, NJ. MR2838731

NORDHAUSEN, K., ILMONEN, P., MANDAL, A., OJA, H. and OLLILA, E. (2011). Deflation-based FastICA reloaded. In *Proceedings of* 19*th European Signal Processing Conference* 2011 (*EUSIPCO* 2011) 1854–1858.

OJA, H., SIRKIÄ, S. and ERIKSSON, J. (2006). Scatter matrices and independent component analysis. *Austrian J. Statist.* **35** 175–189.

OLLILA, E. (2010). The deflation-based FastICA estimator: Statistical analysis revisited. *IEEE Trans. Signal Process.* **58** 1527–1541. MR2758026

OLLILA, E., OJA, H. and KOIVUNEN, V. (2008). Complex-valued ICA based on a pair of generalized covariance matrices. *Comput. Statist. Data Anal.* **52** 3789–3805. MR2427381

OWEN, A. (1990). *Empirical Likelihood*. Chapman & Hall, London.

PLUMBLEY, M. D. (2005). Geometrical methods for non-negative ICA: Manifolds, lie groups and toral subalgebras. *Neurocomputing* **67** 161–197.

PRÉKOPA, A. (1973). On logarithmic concave measures and functions. *Acta Sci. Math.* (*Szeged*) **34** 335–343. MR0404557

RUFIBACH, K. and DÜMBGEN, L. (2006). `logcondens`: Estimate a log-concave probability density from *i.i.d.* observations `R` package version 2.01. Available at http://cran.r-project.org/web/packages/logcondens/.

SAMAROV, A. and TSYBAKOV, A. (2004). Nonparametric independent component analysis. *Bernoulli* **10** 565–582. MR2076063

STATISTICAL LABORATORY
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD
CAMBRIDGE
CB3 0WB
UNITED KINGDOM
E-MAIL: r.samworth@statslab.cam.ac.uk
URL: http://www.statslab.cam.ac.uk/~rjs57

H. MILTON STEWART SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332
USA
E-MAIL: myuan@isye.gatech.edu
URL: http://www2.isye.gatech.edu/~myuan/