Electronic Journal of Statistics
Vol. 10 (2016) 3247–3264
ISSN: 1935-7524
DOI: 10.1214/16-EJS1183

Rate-adaptive Bayesian independent component analysis

Weining Shen

Department of Statistics, University of California, Irvine, CA, 92697, USA
e-mail: weinings@uci.edu

Jing Ning and Ying Yuan

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA
e-mail: jning@mdanderson.org*;* yyuan@mdanderson.org

Abstract: We consider independent component analysis (ICA) using a Bayesian approach. The latent sources are allowed to be block-wise independent while the underlying block structure is unknown. We consider prior distributions on the block structure, the mixing matrix and the marginal density functions of latent sources using a Dirichlet mixture and random series priors. We obtain a minimax-optimal posterior contraction rate of the joint density of the latent sources. This finding reveals that Bayesian ICA adaptively achieves the optimal rate of convergence according to the unknown smoothness level of the true marginal density functions and the unknown block structure. We evaluate the empirical performance of the proposed method by simulation studies.

Keywords and phrases: Adaptive estimation, Dirichlet mixture prior, independent component analysis, nonparametric Bayes, posterior contraction rate.

Received December 2015.

1. Introduction

Independent component analysis (ICA) refers to the problem of recovering unknown independent source signals given observations of their linear combinations. More specifically, letting $\boldsymbol{S} = (S_1, \dots, S_d)^T$ be the independent unknown source signals and $\boldsymbol{W}$ be an unknown $d \times d$ matrix, ICA aims to estimate $\boldsymbol{W}$ and the distribution of $\boldsymbol{S}$ given n independent and identically distributed (i.i.d.) observations of $\boldsymbol{X}$ generated from the following model,

$$\boldsymbol{X}_{d\times 1} = \boldsymbol{W}_{d\times d}\boldsymbol{S}_{d\times 1}, \quad \text{or equivalently,} \quad \boldsymbol{S}_{d\times 1} = \boldsymbol{A}_{d\times d}\boldsymbol{X}_{d\times 1}, \tag{1}$$

where $\boldsymbol{A} = \boldsymbol{W}^{-1}$ is usually called the mixing matrix.

In recent years, ICA has been widely used in signal processing, machine learning and brain imagining, among many other areas of application (Roberts and Everson, 2001; Hyvärinen, Karhunen and Oja, 2001). There are two sets of common approaches in solving ICA problems. The first set, which builds on parametric assumptions of the marginal densities of $\boldsymbol{S}$, includes the maximum likelihood approach (Bell and Sejnowski, 1995; Lee, Girolami and Sejnowski, 1999), minimizing mutual information (Cardoso, 1999) and more generally optimizing contrast functions such as Kullback–Leibler divergence, entropy and non-Gaussian measures (Comon, 1994; Hyvärinen, 1999). However, as the distribution of $\boldsymbol{S}$ is usually unknown in practice, it may be more appealing to consider an alternative class of methods by viewing ICA as a semiparametric model without the requirement of any parametric assumptions on S. Popular methods include the kernel method (Bach and Jordan, 2002), maximum likelihood (Hastie and Tibshirani, 2003), B-spline approximation (Chen and Bickel, 2006) and log-concave ICA projection (Samworth and Yuan, 2012).

Bayesian ICA has gained popularity due to its flexibility in incorporating prior information and its easy use in making inference (e.g., chapter 20 of Hyvärinen, Karhunen and Oja, 2001). It has particular application in biomedical image processing when there is a need to impose mathematical constraints on the mixing matrix. For example, in electroencephalography analysis, it may be helpful to restrict all elements in the mixing matrix to be non-negative such that the ongoing potential from the cortex will have the same sign after being observed at the scalp (Roberts and Choudrey, 2003, 2005). Another example is in biological neural network studies such as modeling the visual cortex. It is commonly believed that only a small proportion of neural activity is actually connected (Olshausen and Field, 1996; Bell and Sejnowski, 1997). In this situation, Bayesian ICA is particularly useful to impose the prior knowledge of sparsity on the mixing matrix in the modeling process (Hyvärinen and Karthikesh, 2000).

Recent progress has been made in the development of computation algorithms for Bayesian ICA, such as variational Bayes, mean field approximation and other methods (Winther and Petersen, 2007; Højen-Sørensen, Winther and Hansen, 2002). However, there are considerable gaps in the theoretical properties of the proposed Bayesian methods and their connections to existing theory in the frequentist literature. In this paper, we fill this gap by considering two commonly used priors on marginal densities, namely a Dirichlet mixture and a random series prior, and establishing their asymptotic properties. More specifically, we show that the proposed estimation procedure will lead to a posterior estimate of the joint density of S that converges to the frequentist truth at the optimal minimax rate (up to logarithmic factors). The rate is equivalent to the nonparametric rate of simultaneously estimating d one-dimensional density functions, and is determined by the worst smoothness level of the marginal densities. Consequently, Bayesian ICA can be viewed as a sufficient dimension reduction technique that avoids "the curse of dimensionality" in an asymptotic sense. These results connect to the existing work in the frequentist literature, e.g., Samarov and Tsybakov (2004); Chen and Bickel (2006); Samworth and Yuan (2012). An additional advantage of the Bayesian method is that no tuning process is required as the prior will automatically adapt to the unknown smoothness levels. Of practical relevance, we also consider the block ICA, an extension of the classical ICA, in which the latent sources are allowed to be block-wise independent while the block structure is unknown. This problem is sometimes referred to as multi-dimensional ICA (Cardoso, 1998).

The rest of the paper is organized as follows. We propose a Bayesian approach for block ICA and discuss the choices of the prior in Section 2. We present the main results on posterior contraction rates in Section 3. In Section 4, we discuss the posterior computation and give simulation examples to illustrate the empirical performance of the proposed method. Proofs are given in the Appendix.

2. Method

2.1. Statistical setting

We consider a Bayesian ICA model with an unknown block structure:

$$\boldsymbol{S}_{d\times 1} = \boldsymbol{A}_{d\times d}\boldsymbol{X}_{d\times 1}, \tag{2}$$

where $\boldsymbol{A}$ is the mixing matrix, $\boldsymbol{S}$ are source signals and we observe n number of i.i.d. copies of $\boldsymbol{X}$. We assume that $\boldsymbol{S}$ is block-wise independent with respect to a partition $\mathcal{I} = I_1 \cup \cdots \cup I_t$ of $\{1, \dots, d\}$, i.e., S_i and S_j are independent if i and j belong to different blocks in $\mathcal{I}$. We denote the i-th row of $\boldsymbol{A}$ by $\boldsymbol{A}_i^T$; then the joint density function of $\boldsymbol{S}$ can be written as follows,

$$p(\boldsymbol{S}) = |\det \boldsymbol{A}| \prod_{j=1}^{t} g_j(\boldsymbol{A}_i^T \boldsymbol{X}, i \in I_j), \tag{3}$$

where $\boldsymbol{A}_i$ is a $d \times |I_j|$ matrix if I_j contains multiple indexes ($|I_j| > 1$). Clearly, when $\mathcal{I} = \{1\} \cup \{2\} \cup \cdots \cup \{d\}$, the proposed model reduces to the classical ICA where all components of $\boldsymbol{S}$ are mutually independent. Our goal is to propose appropriate prior distributions on the partition $\mathcal{I}$, the mixing matrix $\boldsymbol{A}$ and the marginal distributions of $\boldsymbol{S}$, denoted by $\boldsymbol{g} = (g_1, \dots, g_t)$.

2.2. Prior construction

We construct the prior in a hierarchical way, first on the block partition, then on the mixing matrix and corresponding marginal density functions. The following steps provide more details.

(A1) Prior on the block partition Π_P: The assignment of priors on the block structure is equivalent to assigning prior distributions on each partition of $\{1, \dots, d\}$. We use a uniform prior, i.e., for every possible partition $\mathcal{I}$ of $\{1, \dots, d\}$, its prior probability $\Pi_P(\mathcal{I}) = B_d^{-1}$, where B_d is the Bell number of d. Clearly, if it is known that there is no block structure, i.e., $d_1 = \cdots = d_d = 1$, then this step is no longer needed.

(A2) Prior on the mixing matrix Π_A: We consider i.i.d. continuous distributions Π_A^1 on each element a_{ij} of $\boldsymbol{A}$ satisfying

$$\Pi_A^1(|a_{ij}| > M) \leq c_1 \exp\{-c_1' M^{\tau_1}\} \text{ for a sufficiently large } M \tag{4}$$

for some constants $c_1, c_1', \tau_1 > 0$. This condition is easily satisfied for distributions such as exponential, Gamma and Laplace. It is possible to impose a sparsity structure (as discussed by (Hyvärinen and Karthikesh, 2000)) by considering $\Pi_A^1(a_{ij}) \propto \exp\{-G(a_{ij})\}$, where G is a positive, convex function, say $G(x) = |x|$. It can be easily verified that these sparse priors satisfy the proposed conditions when $G(x)$ is a polynomial of $|x|$. In Roberts and Choudrey (2005), the authors proposed using a rectified Gaussian distribution (Gaussian distribution restricted on $[0, \infty)$) as the prior for every element of A. Clearly, that prior also satisfies condition (4).

It is well known that the solution to ICA is unique up to block-wise permutation and scaling (Theis, 2005). Therefore, in practice, one may also consider scaling restrictions on the mixing matrix, e.g., each row of $\boldsymbol{A}$ belongs to $\Omega = \{\boldsymbol{x} = (x_1, \ldots, x_d)^T : \boldsymbol{x} \in \mathbb{R}^d, x_1 \geq 0, \|\boldsymbol{x}\|_2 = 1\}$. Then we may consider i.i.d. priors Π_A^2 on $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_d$ satisfying

$$\Pi_A^2(\|\boldsymbol{A}_i - \boldsymbol{x}\|_2 < \epsilon) \geq c_1'' \epsilon^{\tau_2} \tag{5}$$

for every $\boldsymbol{x} \in \Omega$, sufficiently small $\epsilon > 0$ and some constants $c_1'' > 0, \tau_2 \geq 0$. This can be done by first constructing prior distributions on each element of A as in Π_A^1, and then performing a transformation (standardization, change sign) if needed.

Given the partition chosen as $\mathcal{I} = \{I_1, \ldots, I_t\}$ and the fixed mixing matrix, the induced joint density function $p(\boldsymbol{S})$ can be obtained as in (3). Our last step is to build prior distributions on the marginal density functions $g_1, \ldots, g_t$. Denote the size of block i by $d_i = |I_i|$. We consider two sets of priors based on the input $\boldsymbol{X}$. If $\boldsymbol{X}$ is bounded, then we consider a random series prior based on the tensor-product B-spline expansion. On the other hand, if $\boldsymbol{X}$ is unbounded, then we consider a Dirichlet mixture prior with Gaussian kernels.

(A3.1) Random series prior Π_g^S: We consider the B-spline basis (tensor-product if $d_i \geq 2$) of order q with fixed, equally spaced knots; see de Boor (2001) for an introduction. We assign independent priors on $g_1, \ldots, g_t$; in particular, we rewrite g_i as

$$g_i(s) = \Psi\left(\sum_{j=1}^{J_i} \theta_{i,j} B_j(s)\right) \Big/ \int \Psi\left(\sum_{j=1}^{J_i} \theta_{i,j} B_j(\boldsymbol{u})\right) d\boldsymbol{u}, i = 1, \ldots, t, \tag{6}$$

where $\Psi \in C^\infty$ is a prechosen, nonnegative and monotonic link function (e.g., exponential) that ensures the validity of g_i, and $B_j(s)$ is the B-spline (tensor-product of B-splines if $d_i \geq 2$) basis function. The number of basis terms J_i controls the accuracy and complexity of the model.

For $J_1, \dots, J_t$, we consider i.i.d priors Π_J that satisfy

$$\exp\{-c_2 j(\log j)^{\kappa_1}\} \leq \Pi_J(J_i = j) \leq \exp\{-c_2' j(\log j)^{\kappa_2}\}, i = 1, \dots, t. \quad (7)$$

for some fixed constants $c_2, c_2' > 0, 0 \leq \kappa_2 \leq \kappa_1 \leq 1$ and any positive integer j. This condition is satisfied for discrete distributions such as Poisson and geometric distributions (Shen and Ghosal, 2015). For tensor-product B-splines, i.e., $d_i \geq 2$, we can take $\lfloor J_i^{1/d_i} \rfloor$ as the number of basis expansion terms for each direction.

Given fixed $J_1, \dots, J_t$, we consider independent J_i-dimensional priors on the corresponding coefficients $\boldsymbol{\theta_i} = (\theta_{i,1}, \dots, \theta_{i,J_i})^T$ satisfying

$$\Pi_\theta(\|\boldsymbol{\theta_i} - \boldsymbol{\theta_0}\|_2 \leq \epsilon) \geq \exp\{-c_3 J_i \log(1/\epsilon)\} \quad (8)$$

$$\Pi_\theta(\boldsymbol{\theta_i} \notin [-M, M]^{J_i}) \leq J_i \exp\{-c_3' M^{\kappa_3}\}, i = 1, \dots, t \quad (9)$$

for any finite $\boldsymbol{\theta_0}$, sufficiently small $\epsilon > 0$, large M and some positive constants c_3, c_3', κ_3. Examples include independent Gaussian, Laplace priors on each element of $\boldsymbol{\theta_i}$ and joint distributions such as the Dirichlet distribution on $\boldsymbol{\theta_i}$.

(A3.2) Dirichlet mixture prior Π_g^K: For each $i = 1, \dots, t$, we write

$$g_i(\boldsymbol{s}) = \int_{\boldsymbol{z} \in \mathbb{R}^{d_i}} \boldsymbol{\phi}_{\boldsymbol{\Sigma}_i}(\boldsymbol{s} - \boldsymbol{z}) dF_i(\boldsymbol{z}), \ \boldsymbol{s} \in \mathbb{R}^{d_i}, \quad (10)$$

where $\phi_{\boldsymbol{\Sigma}}$ is a normal kernel function with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. When $d_i = 1$, $\boldsymbol{\Sigma}_i$ is just a scalar. We consider Dirichlet process priors $F_i \overset{\text{iid}}{\sim} D_{\boldsymbol{\alpha_i}}$, where $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_t$ are mutually independent positive base measures. Let $\bar{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i / \boldsymbol{\alpha}_i(\mathbb{R}^{d_i})$. We assume each $\boldsymbol{\alpha}_i$ satisfies condition (1) in Shen, Tokdar and Ghosal (2013), that is, $1 - \bar{\boldsymbol{\alpha}}_i([-x, x]^d) \lesssim \exp(-x^{a_1})$ for some $a_1 > 0$ and any sufficiently large x. We assign independent priors Π_Σ on $\boldsymbol{\Sigma}_i$ satisfying conditions (2)–(4) in Shen, Tokdar and Ghosal (2013), i.e., for every $i = 1, \dots, t$,

$$\Pi_\Sigma\{\boldsymbol{\Sigma}_i : \lambda_d(\boldsymbol{\Sigma}_i^{-1}) \geq x\} \lesssim \exp(x^{a_2}), \ \Pi_\Sigma\{\boldsymbol{\Sigma}_i : \lambda_1(\boldsymbol{\Sigma}_i^{-1}) \geq x^{-1}]\} \lesssim x^{-a_3}$$

for sufficiently large x, $a_1, a_2 > 0$, and

$$\Pi_\Sigma\left\{\boldsymbol{\Sigma}_i : s_j < \lambda_j(\boldsymbol{\Sigma}_i^{-1}) < s_j(1+t), j = 1, \dots, d_i\right\} \gtrsim s_1^{a_4} t^{a_5} \exp(-s_d^{\kappa/2})$$

for any $0 < s_1 \leq \cdots \leq s_{d_i}$ and $t \in (0,1)$, $\kappa > 0$, where $\lambda_1(\boldsymbol{\Sigma}) \leq \cdots \leq \lambda_d(\boldsymbol{\Sigma})$ are eigenvalues of matrix $\boldsymbol{\Sigma}$. The commonly used inverse Wishart (Gamma if $d_i = 1$) distribution satisfies these conditions for $\kappa = 2$.

3. Main results

3.1. Identifiability and uniqueness

Identifiability, uniqueness and separability results play a central role in ICA problems since they allow ICA algorithms to uniquely (up to the changes of scale and permutation) identify the mixing matrix and to recover the source signals. These results have been obtained by Comon (1994); Eriksson and Koivunen (2004) for standard ICA, and then extended to block ICA by Theis (2004). Here, we consider the model defined in (2), and say its solution is *identifiable* and *unique* if the following two conditions hold for any two pair of solutions $(\boldsymbol{A}_1, \boldsymbol{S}_1, \mathcal{I}_1)$ and $(\boldsymbol{A}_2, \boldsymbol{S}_2, \mathcal{I}_2)$: (1) $\boldsymbol{A}_1$ (resp. $\boldsymbol{A}_2$) can be obtained by a linear column transformation of $\boldsymbol{A}_2$ (resp. $\boldsymbol{A}_1$); and (2) the source signals $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ have the same distribution up to the changes of scale and permutations. If these conditions are satisfied for the two solutions $(\boldsymbol{A}_1, \boldsymbol{S}_1, \mathcal{I}_1)$ and $(\boldsymbol{A}_2, \boldsymbol{S}_2, \mathcal{I}_2)$, we say they are equivalent, and define $\mathcal{T}$ as the transformation such that $\mathcal{T}(\boldsymbol{A}_1, \boldsymbol{S}_1, \mathcal{I}_1) = (\boldsymbol{A}_2, \boldsymbol{S}_2, \mathcal{I}_2)$. Let p_1 and p_2 be the density functions of the joint distribution produced by $(\boldsymbol{A}_1, \boldsymbol{S}_1, \mathcal{I}_1)$, and $(\boldsymbol{A}_2, \boldsymbol{S}_2, \mathcal{I}_2)$. Then p_1 and p_2 are also equivalent under $\mathcal{T}$. We write $\mathcal{T}(p_1) = p_2$.

We assume the following conditions on the true data generating process.

(B1) True model: Suppose that there exists a partition $\mathcal{I}_0 = \{I_1^0, \ldots, I_{t_0}^0\}$ of the index set $\{1, \ldots, d\}$ and a non-singular mixing matrix $\boldsymbol{A}_0 = (\boldsymbol{A}_{10}, \ldots, \boldsymbol{A}_{d0})^T$ such that the true density function p_0 can be written in a product form:

$$p_0(x_1, \ldots, x_d) = |\det \boldsymbol{A}_0| \prod_{i=1}^{t_0} g_i^0(\boldsymbol{A}_{j0}^T \boldsymbol{x}, j \in I_i^0),\ \boldsymbol{x} = (x_1, \ldots, x_d)^T,$$

where g_i^0 is a d_i-dimensional marginal density with $d_i = |I_i^0|$, $i = 1, \ldots, t_0$.

(B2) Source densities: For every $i = 1, \ldots, t_0$, assume that g_i^0 is not a degenerating point mass and does not follow a normal distribution (joint normal distribution if $d_i > 1$).

(B3) Mixing matrix: Let $d^* = \max_{i=1}^{t_0} d_i$. We assume that every sub-matrix of size $d^* \times d^*$ of $\boldsymbol{A}^{-1}$ is either invertible or zero.

Conditions (B1)–(B3) are essentially equivalent to those assumed in Theis (2004, 2005). Condition (B2) rules out a joint normal distribution, but still allows for a marginal normal distribution for sources in a block of size greater than one. Condition (B3) is often called d^*-admissible, and is trivially satisfied for standard ICA when $d = 1$. The following lemma asserts that these conditions are sufficient for obtaining identifiability and uniqueness of block ICA.

Lemma 1. *Suppose that (B1)–(B3) hold for model* (2)*, then its solution is identifiable and unique.*

The proof of Lemma 1 is essentially the same as that of Theorem 5.1 in Theis (2004) except that the size of the blocks (d_i) can differ. The necessity of (B1)–(B3) is not clear. If $d^* = 1$, (B2) requires that every marginal density not follow a normal distribution, which is stronger than the necessary condition of allowing at most one Gaussian signal for the standard ICA model (Eriksson and Koivunen, 2004).

3.2. Posterior contraction rate for random series priors

We first consider random series priors in (A1), (A2) and (A3.1). The following assumptions are needed.

(C1) We assume that g_i^0 belongs to a Hölder class $\mathcal{C}^{\alpha_i}$ for some unknown smoothness values $\alpha_i \in (0, q]$, $i = 1, \ldots, t_0$ for some fixed constant $q > 0$.

(C2) We assume that the true joint density p_0 is defined on $[0, 1]^d$ without loss of generality, and is lower bounded by a constant $\underline{m}_p > 0$.

(C3) Denote the support of prior distribution Π_A by S_A. We assume that the true mixing matrix $\boldsymbol{A}_0$ belongs to a known compact set $\mathcal{A}^0 \subset S_A$, such that the density function of Π_A is lower bounded by a constant $\underline{m}_A > 0$ on $\mathcal{A}^0$.

Conditions (C1) and (C2) are commonly used in the literature; see de Jonge and van Zanten (2012); Shen and Ghosal (2015) for example. Condition (C3) is applicable for both unscaled and scaled priors of the mixing matrix as described in (A2). If there is a scaling constraint, then an easy choice for Π_A is to first use i.i.d. continuous distributions truncated between $-M$ and M for each element of $\boldsymbol{A}$, whose density is bounded below by $\underline{m}_A > 0$ and M is a pre-chosen large constant; then rescale each row of $\boldsymbol{A}$ to $\Omega = \{\boldsymbol{x} = (x_1, \ldots, x_d)^T : \boldsymbol{x} \in \mathbb{R}^d, x_1 \geq 0, \|\boldsymbol{x}\|_2 = 1\}$. Doing so automatically satisfies (C3) with $\mathcal{A}^0 = \Omega^d$.

We define the Hellinger distance between two density functions f and g by $d_H(f, g) = \left\{\int (f^{1/2} - g^{1/2})^2 d\mu\right\}^{1/2}$ with respect to the Lebesgue measure μ. Let Π_n be the posterior distribution of p given the observed data. The following theorem obtains the posterior contraction rate for the use of random series priors. The proof is given in the Appendix.

Theorem 1. *(Random series prior) Suppose that conditions (B1)–(B3) and (C1)–(C3) hold. If the prior is constructed as in (A1), (A2) and (A3.1), then there exists a column transformation of the mixing matrix and a scaling transformation of the source signals, together denoted by $\mathcal{T}_0$, such that for any $M_n \to \infty$,*

$$\lim_{n\to\infty} \Pi_n \left[\{p : d_H(\mathcal{T}_0(p_0), p) \leq M_n \epsilon_n\}\right] = 1 \quad \textit{almost surely}, \tag{11}$$

where $\epsilon_n = \max_{i=1}^{t_0} n^{-\alpha_i/(2\alpha_i+d_i)} (\log n)^{\alpha_i/(2\alpha_i+d_i)+(1-\kappa_2)/2}$ is the contraction rate.

Theorem 1 states that the posterior distribution of the joint density p contracts around the true p_0 within an equivalent class under permutation/scaling transformation. For the classical ICA estimation with no block structure, the contraction rate reduces to $n^{-\alpha^*/(2\alpha^*+1)}$ up to a logarithmic factor, in which $\alpha^* = \min(\alpha_1, \ldots, \alpha_d)$ is the worst smoothness level among all directions. Note that this rate corresponds to the classical nonparametric estimation rate for one-dimensional density functions without the logarithmic factor; and it is faster than the usual rate of estimating a d-dimensional function $n^{-\alpha'/(2\alpha'+d)}$ with $\alpha' = d/(\sum \alpha_i^{-1})$ as the harmonic mean of smoothness levels. The assumption $0 < \alpha_1, \ldots, \alpha_{t_0} \leq q$ is needed to ensure sufficient approximation ability of the

B-spline functions being used in the prior (de Boor, 2001). In the prior construction, we do not assume any prior knowledge about the true block structure and the smoothness parameters. Hence the proposed Bayesian estimation procedure is rate-adaptive to the smoothness levels in $(0, q]$. Note that the rate also depends on κ_2, which reflects the tail decay rate in the prior distribution of J_i. A Poisson prior satisfies $\kappa_2 = 1$, hence will help improve the contraction rate.

It is possible to extend our result by considering anisotropic smoothness levels within each block, i.e., condition (C1) can be replaced by

(C1') Assume that g_i^0 belongs to a tensor-Sobolev class with smoothness levels $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{i,d_i})^T$ for some unknown smoothness values $\alpha_i \in (0, q] \cap \mathbb{N}$, $i = 1, \dots, t_0$ for some fixed constant $q > 0$.

Then the posterior contraction rate in Theorem 1 becomes

$$\epsilon_n = \max_{i=1,\dots,t_0} (n/\log n)^{-\alpha_i^*/(2\alpha_i^*+d_i)} (\log n)^{(1-\kappa_2)/2},$$

where $\alpha_i^* = d_i/(\sum_{j=1}^{d_i} \alpha_{i,j}^{-1})$ is the harmonic mean of the elements in $\boldsymbol{\alpha}_i$. This rate can be viewed as the worst rate of estimating t_0 independent anisotropic density functions. If we ignore the block structure and assume that the mixing matrix is known, then the rate agrees with those obtained in the literature for estimating a multi-dimensional anisotropic density function (de Jonge and van Zanten, 2012; Arbel, Gayraud and Rousseau, 2013; Belitser and Serra, 2014; Shen and Ghosal, 2015). Here, the smoothness levels have to be natural numbers to ensure good approximation of the tensor-product B-splines; similar assumptions appeared in Shen and Ghosal (2016).

3.3. Posterior contraction rate for Dirichlet mixture priors

Next, we consider the Dirichlet mixture priors as specified in (A1), (A2) and (A3.2). The density functions are now defined on $\mathbb{R}^d$. We need the following assumptions.

(C4) We consider the joint density function defined on $\mathbb{R}^d$, which has tails that decay exponentially fast for some $\tau, c_4, c_4' > 0$:

$$p_0(\boldsymbol{x}) \leq c_4 \exp(-c_4' \|\boldsymbol{x}\|_2^\tau), \text{ for any } \boldsymbol{x} \in \mathbb{R}^d \text{ that } \|\boldsymbol{x}\|_2 \text{ is sufficiently large.}$$

(C5) Let $\alpha_1, \dots, a_{t_0}$ be positive smoothness levels. For every $i = 1, \dots, t_0$, and every multi-index $\boldsymbol{k}(i) = (k_1, \dots, k_{d_i})$ with $|\boldsymbol{k}(i)| \leq \alpha_i$ and $|\boldsymbol{k}(i)| = k_1 + \dots + k_{d_i}$, assume that

$$\int \frac{\partial^{|\boldsymbol{k}(i)|} g_i^0(x_1, \dots, x_{d_i})}{\partial x_1^{k_i} \cdots \partial x_{d_i}^{k_{d_i}}} (g_i^0)^{(2\alpha_i+\epsilon)/|\boldsymbol{k}(i)|-1} < \infty,$$

and

$$\left| \frac{\partial^{|\boldsymbol{k}(i)|}}{\partial x_1^{k_i} \cdots \partial x_{d_i}^{k_{d_i}}} g_i^0(\boldsymbol{x}+\boldsymbol{y}) - \frac{\partial^{|\boldsymbol{k}(i)|}}{\partial x_1^{k_i} \cdots \partial x_{d_i}^{k_{d_i}}} g_i^0(\boldsymbol{x}) \right| \leq \exp(\|\boldsymbol{y}\|_2^2) \|\boldsymbol{y}\|_2^{\alpha_i - \lfloor \alpha_i \rfloor}$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{d_i}$ and some $\epsilon > 0$, where $\lfloor \alpha \rfloor$ is defined as the largest integer strictly smaller than α.

Condition (C4) requires the tail of p_0 to decay exponentially fast. Condition (C5) imposes smoothness on $g_1, \ldots, g_{t_0}$. We then obtain the convergence result for the Dirichlet mixture prior as follows.

Theorem 2. *(Dirichlet mixture prior) Suppose that the true density function p_0 satisfies conditions (B1)–(B3), (C4)–(C5) and the prior is constructed as in (A1), (A2) and (A3.2). Then there exists a column transformation of the mixing matrix and a scaling transformation of the source signals, together denoted by $\mathcal{T}_0$, such that for any $M_n \to \infty$,*

$$\lim_{n\to\infty} \Pi_n \left[\{p : d_H(\mathcal{T}_0(p_0), p) \leq M_n \epsilon_n\}\right] = 1 \quad \textit{almost surely}, \tag{12}$$

where $\epsilon_n = \max_{i=1}^{t_0} n^{-\alpha_i/(2\alpha_i+d_i^*)}(\log n)^{\gamma_i}$ is the contraction rate with $\gamma_i > d_i^*(1 + \tau^{-1} + \alpha_i^{-1})/(2 + d_i^*/\alpha_i)$ and $d_i^* = \max(d_i, \kappa)$.

Here, the optimal rate is only obtained if $\kappa \leq \min_i d_i$, in which case $d_i^* = d_i$. This puts some restrictions on the prior distribution of the covariance kernel $\boldsymbol{\Sigma_i}$. For example, if $d_i = 1$ for some i, then one may need to use the squared inverse gamma (instead of inverse gamma) prior on g_i to obtain the optimal rate of posterior convergence. Similar arguments appeared in Theorem 1 of Shen, Tokdar and Ghosal (2013).

Note that consistency of the joint distribution of the signals does not necessarily imply consistency of the block structure or marginal densities. Our results can be viewed as a "prediction consistency" consequence. Intuitively, one would expect that the marginal density function estimates do not deviate from the truth (up to the permutation of indexes). It will be interesting to establish posterior consistency results of these quantities in future work, building on the techniques developed by Juditsky, Lepski and Tsybakov (2009), for example. In this paper, we only consider random series and Dirichlet mixture priors on marginal density functions. We believe that optimal posterior contraction rates can also be obtained by using other type of priors on the marginal densities, such as a Gaussian process or Pitman-Yor process (Bhatacharya, Pati and Dunson, 2014; Scricciolo, 2014). In addition, it will be of interest to consider convergence under the sup-norm using the results from Castillo (2014).

4. Simulation study

We give two simulation examples. In the first example, we consider a three-dimensional source signal $\boldsymbol{S} = (S_1, S_2, S_3)^T$ with a block structure $\mathcal{I} = \{\{1\}, \{2, 3\}\}$, i.e., S_1 is independent of (S_2, S_3). We generate $S_1 \sim 2\text{Beta}(5, 2) - 1$ on $[-1, 1]$, and (S_2, S_3) from a mixture of two Gaussian distributions truncated between -1 and 1,

$$\frac{1}{2} N\left(\begin{pmatrix} 1/2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/9 & 0 \\ 0 & 1/9 \end{pmatrix}\right) + \frac{1}{2} N\left(\begin{pmatrix} 0 \\ -1/3 \end{pmatrix}, \begin{pmatrix} 1/16 & 0 \\ 0 & 1/8 \end{pmatrix}\right).$$

We generate 300 samples of incoming signals S_1, S_2, S_3 and choose the mixing matrix as

$$\boldsymbol{A} = \begin{pmatrix} .36 & -.8 & -.48 \\ .48 & .6 & -.64 \\ .8 & 0 & .6 \end{pmatrix}. \tag{13}$$

For posterior computation, we use the prior in (A1), (A2) and (A3.1). For the prior (A1) on the block structure, we exclude the case when there is no block structure, and consider four cases, $\mathcal{I}_1 = \{\{1\}, \{2, 3\}\}$, $\mathcal{I}_2 = \{\{2\}, \{1, 3\}\}$, $\mathcal{I}_3 = \{\{3\}, \{1, 2\}\}$, $\mathcal{I}_4 = \{\{1\}, \{2\}, \{3\}\}$, i.e., $\mathcal{I}_4$ means mutual independence and $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$ mean two independent densities of dimensions one and two. The prior probability is then 1/4 for each partition. For (A2), we put a normal prior on each element of $\boldsymbol{A}$ independently, and rescale each row of $\boldsymbol{A}$ to have a norm of one. For (A3.1), we consider an identity link function, and use the standardized B-spline basis. Then there is no need to include the integral as long as the coefficient vector $\boldsymbol{\theta}_i$ belongs to a J_i-dimensional simplex for each i (Shen and Ghosal, 2015). In other words, if $d_i = 1$, then the corresponding marginal density can be written as $g_i = \sum_{j=1}^{J_i} \theta_j B_j$ given $\sum \theta_j = 1$. If $d_i = 2$, then $g_i(s_1, s_2) = \sum_{j=1}^{J} \sum_{k=1}^{K} \theta_{jk} B_j(s_1) B_k(s_2)$ given $\sum \theta_{ij} = 1$. We choose the basis to be cubic spline and fix $J_i = 10^{d_i}$ for computational convenience. We use $\mathrm{Dir}(1, \ldots, 1)$ as the prior for θ.

The main challenge in posterior computation is to update the block structure and the corresponding coefficients $\boldsymbol{\theta}$. To accommodate a varying-dimensional parameter space, we use a reversible jump Markov chain Monte Carlo (MCMC) approach (Green, 1995). In particular, if the block structure in the current stage is $\mathcal{I}_i (i = 1, 2, 3)$, then we let $\mathcal{I}$ in the next stage be either the same or $\mathcal{I}_4$ with equal transition probability, 1/2. If the current block structure is $\mathcal{I}_4$, then $\mathcal{I}$ in the next step can be any value of $\mathcal{I}_i, i = 1, \ldots, 4$ with equal probability. To illustrate how dimension matching works, we first consider an example of moving from "lower-dimension" $\mathcal{I}_4$ to "higher-dimension" $\mathcal{I}_3$. The coefficients under $\mathcal{I}_4$ are $\theta_1^{(k)}, \ldots, \theta_J^{(k)}$ for $k = 1, 2, 3$, i.e., coefficients for each marginal density. We keep the coefficients for the marginal density of S_3 the same and update the coefficients for the joint density of S_1 and S_2, denoted by θ_{ij} for $i, j = 1, \ldots, J$. We generate i.i.d. random variables $\eta_{11}, \eta_{12}, \ldots, \eta_{(J-1),(J-1)}$ from the uniform distribution on $[0, 1]$. Then we define $\theta_{ij} = \eta_{ij} \theta_i^{(1)} \theta_j^{(2)}$ for every $i, j = 1, \ldots, (J - 1)$, and solve the values of other θ_{ij} with either $i = J$ or $j = J$ such that $\sum_{j=1}^{J} \theta_{ij} = \theta_i^{(1)}$ and $\sum_{i=1}^{J} \theta_{ij} = \theta_j^{(2)}$. On the other hand, to move from "higher-dimension" $\mathcal{I}_3$ to "'lower-dimension" $\mathcal{I}_4$, we simply let the coefficients for the marginal density of S_1 and S_2 be $\theta_i^{(1)} = \sum_{j=1}^{J} \theta_{ij}$ and $\theta_j^{(2)} = \sum_{i=1}^{J} \theta_{ij}$.

We run the model for 25000 MCMC iterations and discard the first 5000. The results are summarized in Figure 1. The first two subplots show that the original signals (S_2 on the x-axis and S_3 on the y-axis) and the reconstructed signals agree well. The third subplot gives the posterior selection frequency of 4 block structures with $\mathcal{I}_1$ being chosen over 67% of the time. The last subplot

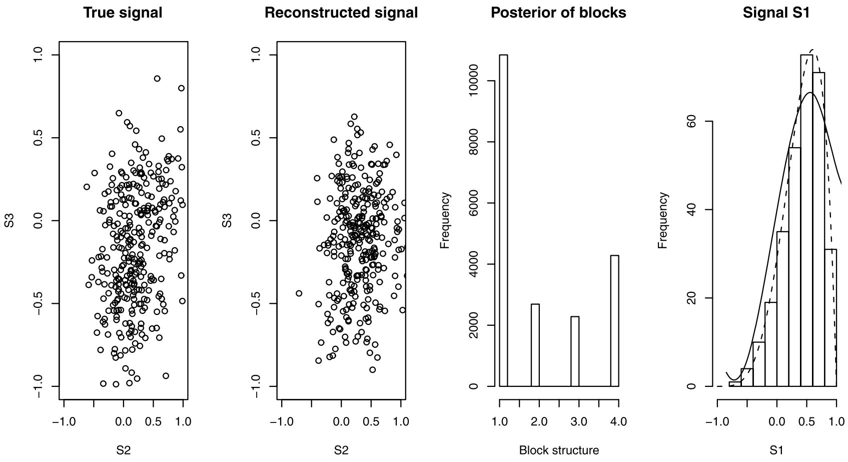

FIG 1. *From left to right: true signals for S_2 and S_3, reconstructed signals for S_2 and S_3, posterior frequency of block structure, and histogram/recovered marginal density (dashed)/true marginal density (solid) for S_1.*

gives the posterior mean of the marginal density of S_1 (dashed line) and the true marginal density (solid line). We find that the reconstructed density fits the data reasonably well. The estimated mixing matrix A^* is fairly close to the true $\boldsymbol{A}$.

$$\boldsymbol{A}^* = \begin{pmatrix} .42 & -.79 & -.42 \\ .43 & .61 & -.63 \\ .83 & -.10 & .49 \end{pmatrix}, \qquad \boldsymbol{A} = \begin{pmatrix} .36 & -.8 & -.48 \\ .48 & .6 & -.64 \\ .8 & 0 & .6 \end{pmatrix}.$$

We also run the model for a larger sample size $n = 1000$, and find similar patterns in the results. The true block structure has been correctly chosen for over 69% of the time. When computing $\boldsymbol{A}^*$, we have used the posterior mean for each matrix element. As a result, each row does not have exactly unit length as desired. One alternative is to consider using the Karcher mean instead of arithmetic mean.

In the second example, we compare the performance of the proposed method with two other popular ICA methods. The first is called FastICA, which is based upon minimizing approximations to entropy (Hyvärinen, Karhunen and Oja, 2001). The second is called ProDenICA, which uses semi-parametric density estimation with cubic splines (Hastie and Tibshirani, 2003). Both methods are implemented in R package “ProDenICA”. We generate data from a three-dimensional source signal with no block structure, i.e., the sources are mutually independent. We use the same mixing matrix as (13). The marginal densities are uniform$(-1, 1)$, Beta$(2, 5)$ rescaled to $(-1, 1)$ and t(3) truncated between -1 and 1. For each method, we compute the Amari metric, which takes values in $[0, 1]$ (Hyvärinen, Karhunen and Oja, 2001), between the estimated mixing matrix and the truth $\boldsymbol{A}$. For FastICA and ProDenICA, we assume the block structure is known and solve the classical ICA problem. For our method, we

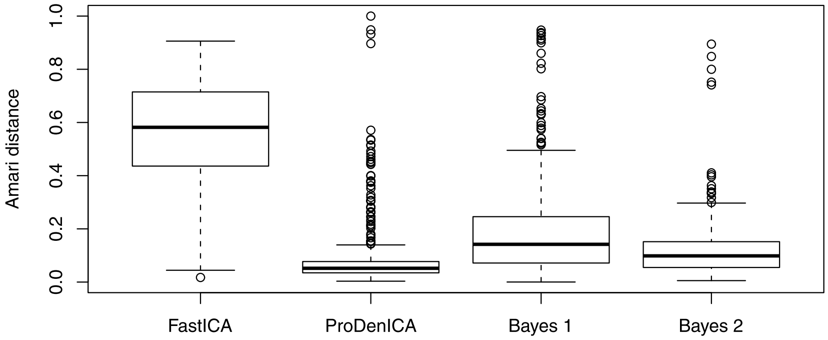

Fig 2. *Boxplot of the Amari distance between the mixing matrix and its estimate for FastICA, ProDenICA, the proposed Bayes method with unknown/known block structure*

consider both scenarios, unknown block structure ("Bayes 1") and known block structure("Bayes 2"). Boxplots of the Amari metric based on 500 replications are summarized in Figure 2. It can be seen that our method has a better performance than FastICA, and performs nearly as well as ProDenICA if the block structure is assumed known. By comparing Bayes 1 with Bayes 2, we find that there is a significant improvement with the use of the true block structure, which suggests the room for future work on improving the estimation accuracy of the block structure.

5. Discussion

In this paper, we study the posterior contraction rate of the Bayesian ICA with an unknown block structure. In practice, a common extension is to include random noise in the output, i.e., $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{S} + \boldsymbol{E}$, where random noise $\boldsymbol{E}$ can be Gaussian or non-Gaussian (Eloyan and Ghosh, 2013). This problem is closely connected to Bayesian density deconvolution. It will be of interest to extend our method to accommodate random noise and obtain the posterior contraction rate of the joint density using the recent results in Sarkar et al. (2014) and Donnet et al. (2015).

In the posterior computation, we use reversible jump MCMC, which only allows for splitting and merging when updating the block structure. This may lead to a low acceptance probability when the number of source signals becomes larger. It is of interest to explore more efficient computational algorithms in a future work.

Appendix: Proofs

Proof of Theorem 1. Throughout the proof, we use Π as a generic notation for the prior on p, and C for a universal positive constant, the value of which may change depending on the context. For any $a, b \in \mathbb{R}$, we say $a \lesssim b$ if $a \leq Cb$, and

$a \gtrsim b$ if $a \geq Cb$. In view of the definition of identifiability and uniqueness for p, we drop the transformation $\mathcal{T}_0$ from the statement and directly work with p_0 without loss of generality. The proof proceeds by verifying the set of conditions given by Theorem 1 of Ghosal, Ghosh and van der Vaart (2000) and Theorem 2.1 of Ghosal and van der Vaart (2001) as listed below,

$$\Pi(\mathcal{F}_n^c) \lesssim \exp(-n\epsilon_n^2), \tag{14}$$

$$\log D(\bar{\epsilon}_n, \mathcal{F}_n, d_H) \lesssim n\bar{\epsilon}_n^2, \tag{15}$$

$$\Pi(p:\ K(p_0,p) \leq \epsilon_n^2, V(p_0,p) \leq \epsilon_n^2) \gtrsim \exp(-n\epsilon_n^2), \tag{16}$$

where $\mathcal{F}_n$ is called a sieve, which is a subset of the parameter space of p, $K(p_0,p) = \int p_0 \log(p_0/p)$ and $V(p_0,p) = \int p_0 \log^2(p_0/p)$ are first- and second-order Kullback-Leibler divergences, and $\epsilon_n, \bar{\epsilon}_n > 0$ are two sequences of numbers going to 0. In particular, we let

$$\begin{aligned} \epsilon_n &= \max_{i=1}^{t_0} (\log n/n)^{\alpha_i/(2\alpha_i+d_i)}, \\ \bar{\epsilon}_n &= \max_{i=1}^{t_0} n^{-\alpha_i/(2\alpha_i+d_i)} (\log n)^{\alpha_i/(2\alpha_i+d_i)+(1-\kappa_2)/2}. \end{aligned} \tag{17}$$

Note that $\kappa_2 \in [0,1]$, hence $\bar{\epsilon}_n$ is the posterior contraction rate because $\bar{\epsilon}_n \geq \epsilon_n$. We define $\mathcal{F}_n$ by considering sieves on the block structure, mixing matrix and marginal densities,

$$\mathcal{F}_n = \mathcal{F}_{\mathcal{I}} \times \mathcal{F}_A \times \mathcal{F}_{g|\mathcal{I},A}, \tag{18}$$

where $\mathcal{F}_{\mathcal{I}}$ is the collection of all possible partitions of $\{1,\ldots,d\}$, and $\mathcal{F}_A$ is defined by

$$\mathcal{F}_A = \{\boldsymbol{A} = (a_{ij})_{d\times d} : |a_{ij}| \leq n^{1/\tau_1}, i,j = 1,\ldots,d\}.$$

Given any $\mathcal{I}$ and $\boldsymbol{A}$, suppose that there are t blocks, with sizes $d_1,\ldots,d_t$. We can then form a sieve on the marginal densities $\boldsymbol{g} = (g_1,\ldots,g_t)$ as

$$\begin{aligned} &\mathcal{F}_{g|\mathcal{I},A} = \mathcal{F}_{g_1|\mathcal{I},A} \times \cdots \times \mathcal{F}_{g_t|\mathcal{I},A}, \\ &\mathcal{F}_{g_i|\mathcal{I},A} = \left\{ g_i(s) \propto \Psi\left(\sum_{j=1}^{J_i} \theta_{i,j} B_j(s)\right) : J_i \leq \bar{J}_i, |\theta_{i,j}| \leq n^{1/\kappa_3} \text{ for } j = 1,\ldots,J_i \right\}, \\ &\bar{J}_i = n^{d_i/(2\alpha_i+d_i)} (\log n)^{2\alpha_i/(2\alpha_i+d_i)-\kappa_2}. \end{aligned}$$

To verify condition (14), note that

$$\Pi(\mathcal{F}_n^c) \leq \Pi_P(\mathcal{F}_{\mathcal{I}}^c) + \Pi_A(\mathcal{F}_A^c) + \Pi_g^S(\mathcal{F}_{g|\mathcal{I},A}^c),$$

where $\Pi_P(\mathcal{F}_{\mathcal{I}}^c) = 0$, $\Pi_A(\mathcal{F}_A^c) \lesssim \exp(-Cn)$. For the third term, the following holds for any partition $\mathcal{I}$ and mixing matrix $\boldsymbol{A}$,

$$\begin{aligned} \Pi_g^S(\mathcal{F}_{g_i|\mathcal{I},A}^c) &\leq \Pi_J(J_i > \bar{J}_i) + \Pi_\theta(\boldsymbol{\theta} \notin [-n^{1/\kappa_3}, n^{1/\kappa_3}]^{\bar{J}_i}) \\ &\leq \exp(-c_2' \bar{J}_i (\log n)^{\kappa_2}) + \bar{J}_i \exp(-c_3' n) \\ &\lesssim \exp(-C n^{d_i/(2\alpha_i+d_i)} (\log n)^{2\alpha_i/(2\alpha_i+d_i)}). \end{aligned}$$

Considering all possible indexes i, we obtain $\Pi_g^S(\mathcal{F}^c_{g|\mathcal{I},A}) \lesssim \exp(-n\epsilon_n^2)$ for any partition $\mathcal{I}$ and mixing matrix $\boldsymbol{A}$. Hence (14) holds.

Next we check condition (15). Consider a partition $\mathcal{I} = \{I_1, \dots, I_t\}$, and its corresponding $g_1, \dots, g_t$. Assume $|I_1| = 1$ without loss of generality. For a mixing matrix $\boldsymbol{A} = (\boldsymbol{A}_1, \dots, \boldsymbol{A}_d)^T$ and its ϵ-perturbation $\boldsymbol{A}_\epsilon = (\boldsymbol{A}_1^*, \boldsymbol{A}_2, \dots, \boldsymbol{A}_d)^T$, with the first row satisfying $\|\boldsymbol{A}_1 - \boldsymbol{A}_1^*\|_\infty < \epsilon$. Then it can be shown that

$$|\det \boldsymbol{A} - \det \boldsymbol{A}_\epsilon| \le C'\epsilon, \;\; |g_1(\boldsymbol{A}_1^T \boldsymbol{X}) - g_1(\boldsymbol{A}_1^{*T}\boldsymbol{X})| \le C'\epsilon$$

for some constant $C' > 0$ since g_1 is Lipschitz continuous and $\boldsymbol{X}$ is bounded. This calculation holds similarly for $g_2, \dots, g_t$. Now, given $\mathcal{I}$ and $\boldsymbol{A}$ as chosen for marginal density functions, their entropy calculation can be obtained in the same way as in Shen and Ghosal (2016), that is,

$$D(\bar{\epsilon}_n, \mathcal{F}_{g_i|\mathcal{I},A}, d_H) \le D(\bar{\epsilon}_n^2, \mathcal{F}_{g_i|\mathcal{I},A}, \|\cdot\|_1) \le \bar{J}_i \left(\frac{3}{\bar{\epsilon}_n^2}\right)^{\bar{J}_i}.$$

Since there are B_d (Bell number of d) possible partitions in $\mathcal{I}$, and the entropy associated with $\mathcal{F}_A$ is bounded by a constant multiple of $(1/\bar{\epsilon}_n^2)^{d^2}$, we have

$$\log D(\bar{\epsilon}_n, \mathcal{F}_n, d_H) \lesssim \log B_d + d^2 \log(1/\bar{\epsilon}_n) + \max_i \left\{\bar{J}_i \log(1/\bar{\epsilon}_n)\right\} \lesssim n\bar{\epsilon}_n^2. \quad (19)$$

This shows that (15) holds.

In order to verify condition (16), we first need to find an approximation of g_0. Using some existing approximation result of the (tensor-product) B-spline, e.g., Lemma 2.1 of de Jonge and van Zanten (2012), for every $i = 1, \dots, t$, there exist sequences $J_i^* = \lfloor C\epsilon_n^{-d_i/\alpha_i} \rfloor$, such that for every $J_i \ge J_i^*$, there exists a vector of coefficients with good approximation $\boldsymbol{\theta}^*_{i,J_i} = \left(\theta^*_{i,1}(J_i), \dots, \theta^*_{i,J_i}(J_i)\right)^T$ satisfying $\|\Psi^{-1} g_i^0(\boldsymbol{s}) - \sum_{j=1}^{J_i} \theta^*_{i,j}(J_i) B_j(\boldsymbol{s})\|_\infty \le J_i^{-\alpha_i/d_i} \le \epsilon_n$. Because Ψ is Lipschitz continuous, $\left\|g_i^0(\boldsymbol{s}) - \Psi\left\{\sum_{j=1}^{J_i} \theta^*_{i,j}(J_i) B_j(\boldsymbol{s})\right\}\right\|_\infty \le \epsilon_n$. Define $g_i^*(\boldsymbol{s}; J_i) = C_i^{-1}\Psi\left\{\sum_{j=1}^{J_i} \theta^*_{i,j}(J_i) B_j(\boldsymbol{s})\right\}$, where C_i is the normalizing constant that ensures g_i^* is a valid density function of $\boldsymbol{s}$. Then $|C_i^{-1} - 1| \le \epsilon_n$, and $\|g_i^*(\boldsymbol{s}; J_i) - g_i^0(\boldsymbol{s})\|_\infty \le \epsilon_n$. Define $g_i^{**}(\boldsymbol{s}; J; \boldsymbol{\theta}_i) = C_i'^{-1}\Psi\left\{\sum_{j=1}^{J_i} \theta^{**}_{i,j} B_j(\boldsymbol{s})\right\}$ with C_i' being the normalizing constant and $\boldsymbol{\theta}_i^{**} = (\theta^{**}_{i,1}, \dots, \theta^{**}_{i,J_i})^T$ being an arbitrary element of the support of the prior. Let $\mathcal{G}$ be the collection of $\boldsymbol{g} = (g_1^{**}(\boldsymbol{s}; J_1; \boldsymbol{\theta}_1^{**}), \dots, g_t^{**}(\boldsymbol{s}; J_t; \boldsymbol{\theta}_t^{**}))$ with $J_i \in (J_i^*, C^* J_i^*) \cap \mathbb{N}$ for some large constant C^*, $i = 1, \dots, t$, and $\|\boldsymbol{\theta}_i^{**} - \boldsymbol{\theta}^*_{i,J_i}\|_2 \le J_i^{-1}\epsilon_n$.

For the mixing matrix, we consider a neighborhood around $\boldsymbol{A}_0 = (\boldsymbol{A}_{10}, \dots, \boldsymbol{A}_{d0})^T$. In other words, define

$$\mathcal{A} = \{\boldsymbol{A} = (\boldsymbol{A}_1, \dots, \boldsymbol{A}_d)^T : \|\boldsymbol{A}_i - \boldsymbol{A}_{i0}\|_2 \le C_A \epsilon_n, i = 1, \dots, d\},$$

where C_A is a small constant that satisfies

$$|\det \boldsymbol{A}_0 - \det \boldsymbol{A}| \le \epsilon_n, \quad \|\boldsymbol{A}_{i0} - \boldsymbol{A}_i\|_\infty \le \epsilon_n,$$

for every $\boldsymbol{A} \in \mathcal{A}$. Then under the true partition $\mathcal{I}_0 = \{I_1^0, \dots, I_{t_0}^0\}$, for any mixing matrix $\boldsymbol{A} \in \mathcal{A}$, and any marginal densities $\boldsymbol{g} = (g_1, \dots, g_{t_0}) \in \mathcal{G}$, where $g_i \propto \Psi\left\{\sum_{j=1}^{J_i} \theta_{i,j} B_j(\boldsymbol{s})\right\}$ and $\boldsymbol{\theta}_{i,J_i} = (\theta_{i,1}, \dots, \theta_{i,J_i})^T$, define

$$p_{\boldsymbol{g},\boldsymbol{A},\theta_{i,J_i}}(\boldsymbol{x}) = \det \boldsymbol{A} \prod_{i=1}^{t_0} g_i(\boldsymbol{A}_j^T \boldsymbol{x}, j \in I_j^0; \theta_{i,J_i}),$$

$$p^*_{\boldsymbol{g},\boldsymbol{A},\theta^*_{i,J_i}}(\boldsymbol{x}) = \det \boldsymbol{A} \prod_{i=1}^{t_0} g_i^*(\boldsymbol{A}_j^T \boldsymbol{x}, j \in I_j^0; J_i, \theta^*_{i,J_i}).$$

Then

$$\begin{aligned}
&\|p_0(\boldsymbol{x}) - p_{\boldsymbol{g},\boldsymbol{A}}(\boldsymbol{x})\|_\infty \\
&\quad\le \left\| p_0(\boldsymbol{x}) - \det \boldsymbol{A} \prod_{i=1}^{t_0} g_i^0(\boldsymbol{A}_j^T \boldsymbol{x}, j \in I_j^0) \right\|_\infty \\
&\qquad + \left\| \det \boldsymbol{A} \prod_{i=1}^{t_0} g_i^0(\boldsymbol{A}_j^T \boldsymbol{x}, j \in I_j^0) - p^*_{\boldsymbol{g},\boldsymbol{A},\theta^*_{i,J_i}}(\boldsymbol{x}) \right\|_\infty \\
&\qquad + \left\| p^*_{\boldsymbol{g},\boldsymbol{A},\theta^*_{i,J_i}}(\boldsymbol{x}) - p_{\boldsymbol{g},\boldsymbol{A},\theta_{i,J_i}}(\boldsymbol{x}) \right\|_\infty \\
&\quad\lesssim \epsilon_n,
\end{aligned} \tag{20}$$

because p_0 only differs with $\det \boldsymbol{A} \prod_{i=1}^{t_0} g_i^0(\boldsymbol{A}_j^T \boldsymbol{x}, j \in I_j^0)$ in the mixing matrix, $\det \boldsymbol{A} \prod_{i=1}^{t_0} g_i^0(\boldsymbol{A}_j^T \boldsymbol{x}, j \in I_j^0)$ differs with $p^*_{\boldsymbol{g},\boldsymbol{A},\theta^*_{i,J_i}}(\boldsymbol{x})$ in the marginal density functions $\boldsymbol{g}$, $p^*_{\boldsymbol{g},\boldsymbol{A},\theta^*_{i,J_i}}(\boldsymbol{x})$ differs with $p_{\boldsymbol{g},\boldsymbol{A},\theta_{i,J_i}}(\boldsymbol{x})$ by the B-spline coefficients θ, and all these approximation errors are bounded by ϵ_n given $\boldsymbol{A} \in \mathcal{A}$ and $\boldsymbol{g} \in \mathcal{G}$.

Under condition (B3), given ϵ_n sufficiently small, $p_{\boldsymbol{g},\boldsymbol{A}}(\boldsymbol{x})$ is also lower bounded by a positive constant. This implies that $p_0/p_{\boldsymbol{g},\boldsymbol{A}}$ is always finite. Hence $d_H(p_0, p_{\boldsymbol{g},\boldsymbol{A}}) \lesssim \epsilon_n$ and

$$K(p_0, p_{\boldsymbol{g},\boldsymbol{A}}) \lesssim \epsilon_n^2, \quad V(p_0, p_{\boldsymbol{g},\boldsymbol{A}}) \lesssim \epsilon_n^2$$

due to Lemma 8 of Ghosal and van der Vaart (2007). Thus it is good enough to obtain a lower bound for the prior probability of $\mathcal{I}_0 \times \mathcal{A} \times \mathcal{G}$. We have

$$\begin{aligned}
&\Pi_P(\mathcal{I}_0) = B_d^{-1}, \quad \Pi_A^2(\mathcal{A}) \gtrsim \epsilon_n^{d\tau_2}, \\
&\Pi_g^S(\mathcal{G}) \\
&\quad\gtrsim \prod_{i=1}^{t_0} \left\{ \Pi_J(J_i^* \le J_i \le C^* J_i^*) \times \min_{J_i \in (J_i^*, C^* J_i^*) \cap \mathbb{N}} \Pi_\theta(\|\boldsymbol{\theta}_{i,J_i} - \boldsymbol{\theta}^*_{i,J_i}\|_2 \le J_i^{-1} \epsilon_n) \right\} \\
&\quad\gtrsim \prod_{i=1}^{t_0} \exp(-C J_i^* \log n).
\end{aligned}$$

Therefore $\Pi(p : K(p_0, p) \le \epsilon_n^2, V(p_0, p) \le \epsilon_n^2) \ge \Pi_P(\mathcal{I}_0) \times \Pi_A^2(\mathcal{A}) \times \Pi_g^S(\mathcal{G}) \gtrsim \exp(-n\epsilon_n^2)$. This completes the proof. □

Proof of Theorem 2. The proof proceeds in the same way as Theorem 1. We briefly describe the main differences here. First, in order to verify conditions (14) and (15), we need an alternative definition of $\mathcal{F}_g$. For each $i = 1, \ldots, t$,let $\tilde{\epsilon}_{n,i} = n^{-\alpha_i/(2\alpha_i+d_i^*)}(\log n)^{\gamma_{i0}}$ with $\gamma_{i0} = d_i^*(1+\tau^{-1}+\alpha_i^{-1})/(2+d_i^*/\alpha_i)$ and $d_i^* = \max(d_i, \kappa)$. Let $\epsilon_{n,i} = n^{-\alpha_i/(2\alpha_i+d_i^*)}(\log n)^{\gamma_i}$ with $\gamma_i > \gamma_{i0}$. For $\mathcal{F}_g$, we consider the sieve as described in Proposition 2 of Shen, Tokdar and Ghosal (2013) for each of $g_1, \ldots, g_t$, denoted by $\mathcal{Q}_1, \ldots, \mathcal{Q}_t$, respectively. Then by Theorem 5 of Shen, Tokdar and Ghosal (2013),

$$\log D(\tilde{\epsilon}_{n,i}, \mathcal{Q}_i, d_H) \lesssim n\epsilon_{n,i}^2, \quad \Pi_g(\mathcal{Q}_i^c) \lesssim \exp(-n\epsilon_{n,i}^2).$$

Combining these results for each i, we obtain (14) and (15).

Second, to verify condition (16), we consider the true partition and construct the approximation of $g_1^0, \ldots, g_{d_0}^0$ as in Proposition 1 of Shen, Tokdar and Ghosal (2013). For the mixing matrix, we consider a neighborhood around $\boldsymbol{A}_0 = (\boldsymbol{A}_{10}, \ldots, \boldsymbol{A}_{d0})^T$. In other words, we define

$$\mathcal{A} = \{\boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_d)^T : \|\boldsymbol{A}_i - \boldsymbol{A}_{i0}\|_2 \le C_A\epsilon_n, i = 1, \ldots, d\},$$

where Ω is defined in (A2.2). Given C_A sufficiently small, we have

$$|\det \boldsymbol{A}_0 - \det \boldsymbol{A}| \le \epsilon_n, \quad \|\boldsymbol{A}_{i0} - \boldsymbol{A}_i\|_\infty \le \epsilon_n$$

for every $\boldsymbol{A} \in \mathcal{A}$. This ensures that the approximation to p_0 under the Hellinger distance is still within a multiple of ϵ_n. By condition (B5), every element of $\boldsymbol{A}_0$ belongs to $\mathcal{A}^0$, and so does every element of $\boldsymbol{A}$ for every $\boldsymbol{A} \in \mathcal{A}$. Hence the prior probability of $\mathcal{A}$ is lower bounded by a constant multiple of $\epsilon_n^{d^2} = \exp(-d^2\log(1/\epsilon_n))$, which will not affect the rate calculation in (16). □

Acknowledgments

The authors are grateful to the Editor, an Associate Editor and two Referees for their valuable comments which led to a substantial improvement of paper. Weining Shen would like to thank Subhashis Ghoshal and Sujit Ghosh for helpful discussions. Shen's research is partially supported by UCI CORCL award.

References

Arbel, J., Gayraud, G. and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.* **40** 549–570. MR3091697

Bach, F. R. and Jordan, M. I. (2002). Kernel Independent Component Analysis. *J. Mach. Learn. Res.* **3** 1–48. MR1966051

Belitser, E. and Serra, P. (2014). Adaptive Priors Based on Splines with Random Knots. *Bayesian Anal* **9** 859–882. MR3293959

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation* **7** 1129–1159.

Bell, A. J. and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research* **37** 3327–3338.

Bhatacharya, A., Pati, D. and Dunson, D. B. (2014). Anisotropic function estimation with multi-bandwidth Gaussian process. *Ann. Statist.* **32** 352–381. MR3189489

Cardoso, J. F. (1998). Multidimensional independent component analysis. In *Proc. of ICASSP '98*.

Cardoso, J. F. (1999). High-order contrasts for independent component analysis. *Neural Computation* **11** 157–192.

Castillo, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091. MR3262477

Chen, A. and Bickel, P. (2006). Efficient Independent component analysis. *Ann. Statist.* **34** 2825–2855. MR2329469

Comon, P. (1994). Independent component analysis. A new concept? *Signal Processing* **36** 287–314.

de Boor, C. (2001). *A Practical Guide to Splines.* Springer.

de Jonge, R. and van Zanten, H. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electron. J. Stat.* **6** 1984–2001. MR3020254

Donnet, S., Rivoirard, V., Rousseau, J. and Scricciolo, C. (2015). Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures. Technical Report, arXiv:1406.4406.

Eloyan, A. and Ghosh, S. K. (2013). A semiparametric approach to source separation using independent component analysis. *Comput. Stat. Data. Anal.* **58** 383–396. MR2997950

Eriksson, J. and Koivunen, V. (2004). Identifiability, Separability, and Uniqueness of Linear ICA Models. *IEEE Signal Processing Letters* **11** 601–604.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. (2000). Convergence Rates of Posterior Distributions. *Ann. Statist.* **28** 500–531. MR1790007

Ghosal, S. and van der Vaart, A. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329

Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* **82** 711–732. MR1380810

Hastie, T. and Tibshirani, R. (2003). Independent component analysis through product density estimation. In *Advances in Neural Information Processing Systems 15* 649–656.

Højen-Sørensen, P. A., Winther, O. and Hansen, L. K. (2002). Mean-field approaches to independent component analysis. *Neural Computation* **14** 889–918.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks* **10** 626–634.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001). *Independent Component Analysis.* Wiley.

Hyvärinen, A. and Karthikesh, R. (2000). Sparse Priors On The Mixing Matrix In Independent Component Analysis. In *Proc. Int. Workshop on ICA2000* 477–452.

Juditsky, A. B., Lepski, O. V. and Tsybakov, A. B. (2009). Nonparametric Estimation of Composite Functions. *Ann. Statist.* **37** 1360–1404. MR2509077

Lee, T. W., Girolami, M. and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation* **11** 417–441.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381** 607–609.

Roberts, S. and Choudrey, R. (2003). Data decomposition using independent component analysis with prior constraints. *Pattern Recognition* **36** 1813–1825.

Roberts, S. and Choudrey, R. (2005). Bayesian independent component analysis with prior constraints: an application in biosignal analysis. In *First international conference on Deterministic and Statistical Methods in Machine Learning* 159–179.

Roberts, S. and Everson, R. (2001). *Independent Component Analysis: Principles and Practice.* Cambridge University Press.

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

Samworth, R. J. and Yuan, M. (2012). Independent Component Analysis via Nonparametric Maximum Likelihood Estimation. *Ann. Statist.* **40** 2973–3002. MR3097966

Sarkar, A., Mallick, B. K., Staudenmayer, J., Pati, D. and Carroll, R. J. (2014). Bayesian Semiparametric Density Deconvolution in the Presence of Conditionally Heteroscedastic Measurement Errors. *J. Comp. Graph. Stat.* **24** 1101–1125. MR3270713

Scricciolo, C. (2014). Adaptive Bayesian Density Estimation in Lp-metrics with Pitman-Yor or Normalized Inverse-Gaussian Process Kernel Mixtures. *Bayesian Anal.* **9** 475–520. MR3217004

Shen, W. and Ghosal, S. (2015). Adaptive Bayesian procedures using random series priors. *Scand. J. Stat.* **42** 1194–1213. MR3426318

Shen, W. and Ghosal, S. (2016). Adaptive Bayesian density regression for high dimesnional data. *Bernoulli* **22** 396–420. MR3449788

Shen, W., Tokdar, S. T. and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441

Theis, F. J. (2004). Uniqueness of complex and multidimensional independent component analysis. *Signal Processing* **84** 951–956.

Theis, F. J. (2005). Multidimensional independent component analysis using characteristic functions. In *Proc. of EUSIPCO.*

Winther, O. and Petersen, K. B. (2007). Bayesian independent component analysis: Variational methods and non-negative decompositions. *Digital Signal Processing* **17** 858–872.