Detection of Sources in Non-Negative Blind Source Separation by Minimum Description Length Criterion

Chia-Hsiang Lin, Chong-Yung Chi, *Senior Member, IEEE*, Lulu Chen, David J. Miller, *Senior Member, IEEE*, and Yue Wang, *Fellow, IEEE*

***Abstract*— While non-negative blind source separation (nBSS) has found many successful applications in science and engineering, model order selection, determining the number of sources, remains a critical yet unresolved problem. Various model order selection methods have been proposed and applied to real-world data sets but with limited success, with both order over- and under-estimation reported. By studying existing schemes, we have found that the unsatisfactory results are mainly due to invalid assumptions, model oversimplification, subjective thresholding, and/or to assumptions made solely for mathematical convenience. Building on our earlier work that reformulated model order selection for nBSS with more realistic assumptions and models, we report a newly and formally revised model order selection criterion rooted in the minimum description length (MDL) principle. Adopting widely invoked assumptions for achieving a unique nBSS solution, we consider the mixing matrix as consisting of deterministic unknowns, with the source signals following a multivariate Dirichlet distribution. We derive a computationally efficient, stochastic algorithm to obtain approximate maximum-likelihood estimates of model parameters and apply Monte Carlo integration to determine the description length. Our modeling and estimation strategy exploits the characteristic geometry of the data simplex in nBSS. We validate our nBSS-MDL criterion through extensive simulation studies and on four real-world data sets, demonstrating its strong performance and general applicability to nBSS. The proposed nBSS-MDL criterion consistently detects the true number of sources, in all of our case studies.**

***Index Terms*— Dirichlet distribution, minimum description length (MDL), model order selection, Monte Carlo integration, non-negative blind source separation (nBSS).**

Manuscript received July 20, 2016; revised February 1, 2017 and August 4, 2017; accepted August 31, 2017. Date of publication October 3, 2017; date of current version August 20, 2018. This work was supported in part by the Ministry of Science and Technology under Grant MOST 104-2221-E-007-069-MY3 and in part by the U.S. National Institute of Health under Grant CA184902 and Grant ES024988. *(Corresponding author: Chia-Hsiang Lin.)*

C.-H. Lin and C.-Y. Chi are with the Department of Electrical Engineering, Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: chiahsiang.steven.lin@gmail.com; cychi@ee.nthu.edu.tw).

L. Chen and Y. Wang are with the Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203 USA (e-mail: luluchen@vt.edu; yuewang@vt.edu).

D. J. Miller is with the School of Electrical Engineering and Computer Science, Pennsylvania State University, University Park, PA 16802 USA (e-mail: djmiller@engr.psu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2749279

I. Introduction

NON-NEGATIVE blind source separation (nBSS), an unsupervised learning problem to recover non-negative source signals, has found many successful applications in both science and engineering [1]–[4], and the frameworks for addressing this problem have early roots in neural networks, leading to the seminal nBSS method known as non-negative independent component analysis [5]. Other popular nBSS methods include non-negative matrix factorization (NMF) [6]–[9], which is of great interest in machine learning [10], and convex analysis of mixtures (CAM) [11], [12], which has received attention from the Neural Networks and Learning Systems (NNLS) community in recent years [13]. However, model order selection [14], determining the number of sources, remains a critical yet unresolved problem (just as estimating the number of clusters remains a largely unresolved problem in unsupervised data clustering)—it may be imperative to achieve source separation (or clustering) solutions that closely correspond to the underlying physical sources. In fact, even many recent nBSS works published in TRANSACTIONS ON NNLS unrealistically assume the number of sources is known (see [13], [15]). Various model order selection methods have been proposed and applied to real-world data sets, but with limited success, suffering from either order over- or under-estimation [14], [16]. By examining existing schemes, we have found that these unsatisfactory results may be due to invalid assumptions, model oversimplification, subjective thresholding, and to assumptions made for mathematical convenience.

One group of model order selection methods is based on a combination of eigenvalue analysis and sequential hypothesis testing. The null hypothesis is that there is no significant difference in explaining the observed data between two models with different orders, and with a sequential search, the statistically significant alternative model with the minimum number of sources is selected. Representative methods in this group include Neyman–Pearson virtual dimensionality (VD) [16], Bartlett test [17], geometry-based estimation of the number of endmembers (GENE) [18], and principal convex hull analysis [19]. While these methods perform well when the signal-to-noise ratio is high, they require subjective thresholding on the significance level, which is often data-dependent.

Another group of methods is rooted in information theory, or more specifically the minimax entropy principle,

without involving any subjective thresholding [20]. For example, Akaike information criterion (AIC) [21] is an unbiased estimator of Kullback–Leibler divergence between modeled and estimated data distributions. AIC uses many approximations and assumptions and thus is, in general, a heuristic criterion [22]. Nevertheless, it has been proposed to detect the number of sources in nBSS [23]. On the other hand, the minimum description length (MDL) criterion [and very similar methods such as the Bayesian information criterion (BIC)] [22] is proposed to find the model most likely in the Bayesian sense [24], [25]. MDL minimizes the total description code-length, comprised of both the data likelihood and the model complexity, over competing models. MDL has been reformulated for detecting the number of sources in various signal processing applications with consistent model order estimation, where both sources and observations are assumed to be stationary/identical and statistically independent Gaussian random vectors with zero mean, and thus the family of models is necessarily described by the corresponding data covariance matrix [16], [26]. This MDL formulation has been conveniently applied to detect the number of sources in BSS, even when some of the assumptions were clearly violated [27].

To address the aforementioned problems, we reformulated MDL-based selection specifically for nBSS and validated its performance on both the simulated and real data sets. We derived an MDL criterion to detect the number of tissue compartments (i.e., the sources) in multitissue compartment modeling and applied our method to analyze *in vivo* dynamic contrast-enhanced magnetic resonance imaging of breast cancers, where the statistical model and related assumptions or parameterization are well justified by the underlying pharmacokinetics principles [28]–[30]. We also derived an MDL criterion to detect the number of cell types in gene expression data deconvolution and applied our method to computationally dissect tissue heterogeneity in complex tissues [3], [31].

While our initial MDL approach for nBSS has shown promising performance in determining the number of sources in many real-world applications, here we report a newly and formally revised nBSS-MDL model and model order selection criterion that is supported by comprehensive theoretical analysis and experimental assessment, and which we will experimentally demonstrate to have wide applicability to a variety of different data types (one of the hallmarks of machine learning approaches), with less likelihood of suffering from overparameterization. Adopting widely invoked assumptions for achieving a unique nBSS solution [3], [4], [11], [32], we have previously shown that in a linear mixing model, when the source signals are non-negative, the scatter simplex of source signals is compressed/expanded and rotated to form the scatter simplex of observed signals whose vertices coincide with the column vectors of the mixing matrix, i.e., every observed data point is confined within the simplex (a convex hull) defined by the column vectors of the mixing matrix [3], [33]. Accordingly, in our nBSS-MDL model herein, we consider the mixing matrix as consisting of deterministic unknowns, with the source signals following a multivariate Dirichlet distribution [34]. Our modeling and estimation strategy exploits the characteristic geometry of the data simplex in nBSS. We derive a computationally efficient, stochastic algorithm to obtain approximate maximum-likelihood (ML) estimates of model parameters based on the connection between stochastic ML estimation and the Craig estimator [35] over simplex geometry. We then apply Monte Carlo integration to determine the description length. We validate our nBSS-MDL criterion through extensive simulation studies and demonstrate its performance and applicability on four real-world data sets. The proposed nBSS-MDL criterion consistently detects the true number of underlying sources, in all of our case studies.

This paper is organized as follows. After the description and formulation of the model order selection problem in Section II, an MDL criterion for determining the number of sources in nBSS is introduced in Section III. The estimation of the model parameters and determination of the total code length are presented in Section IV. Experimental results that validate and illustrate the performance of the proposed nBSS-MDL criterion on both synthetic and real-word data sets are described in Section V. Major conclusions are presented in Section VI. Detailed proofs and derivations are summarized in the appendixes.

This paper adopts the following notations. $\mathbf{e}_i$ is the ith unit vector. $\mathbf{1}_N$ and $\mathbf{0}_N$ are all-one and all-zero N-vectors, respectively. The convex hull, affine hull, and conic hull of a set $\mathcal{S}$ are denoted by $\text{conv}\mathcal{S}$, $\text{aff}\mathcal{S}$, and $\text{conic}\mathcal{S}$, respectively [36]. The relative interior $\text{int}\mathcal{S}$ of a set $\mathcal{S}$ is the interior of $\mathcal{S}$ with respect to (w.r.t.) $\text{aff}\mathcal{S}$ [36]. $\mathcal{I}_Z \triangleq \{1, \ldots, Z\}$, for any positive integer Z. $\succeq$ and $\succ$ are the componentwise inequality and strictly componentwise inequality, respectively.

II. Formulation of the Problem

The observation vector in many nBSS problems such as gene expression deconvolution [3], [37], hyperspectral remote sensing (HRS) unmixing [33], [38], and multitissue compartment analysis [28], [30], denoted by the $M \times 1$ vector $\mathbf{x}[n]$, can be accurately described by the following linear mixing model:

$$\mathbf{x}[n] = \sum_{k=1}^{K} \mathbf{a}_k s_k[n] + \mathbf{w}[n], \quad n \in \mathcal{I}_L \tag{1}$$

where $\mathbf{s}_k = [s_k[1], \ldots, s_k[L]]$ is a $1 \times L$ signal profile referred to as the kth source, $\mathbf{a}_k$ is an $M \times 1$ vector associated with the kth source, $\mathbf{w}[\cdot]$ is an $M \times 1$ additive noise vector, K is the number of sources, M is the number of samples (channels), and L is the length of the signal profile. Using matrix-vector notation, we can rewrite (1) as $\mathbf{x}[n] = \mathbf{A}\mathbf{s}[n] + \mathbf{w}[n]$, where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$ is the $M \times K$ mixing matrix; $\mathbf{s}[n] = [s_1[n], \ldots, s_K[n]]^T$ is the nth $K \times 1$ column vector of the $K \times L$ source matrix $\mathbf{S} = [\mathbf{s}[1], \ldots, \mathbf{s}[L]]$.

Within the context of nBSS, a crucial task associated with the model described in (1) is determining the number of sources K from a finite set of observations $\mathbf{X} = [\mathbf{x}[1], \ldots, \mathbf{x}[L]]$.

A promising approach to this problem is based on the geometric structure of the scatter simplex of the source vectors $\mathbf{s}[\cdot]$ and its relationship to that of the observation vectors $\mathbf{x}[\cdot]$.

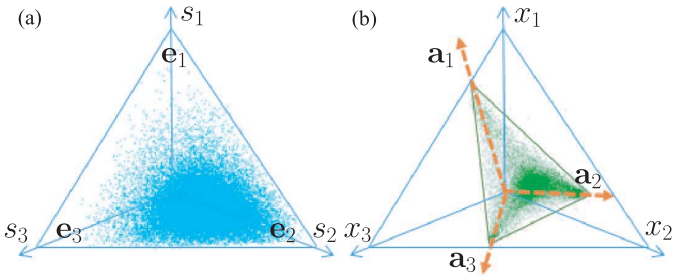


Fig. 1. (a) Resampled Dirichlet distribution with the parameter values estimated from real gene expression profiles $\mathbf{S}$ of pure tissue types (GSE19830). (b) Linear transformation of $\mathbf{S}$ produces a compressed/expanded and rotated scatter simplex of $\mathbf{X}$ whose vertices coincide with the column vectors of the mixing matrix $\mathbf{A}$.

To introduce this approach, we make the following assumptions.

(A1) $\mathbf{A}$ is of full column rank and $L \geq K$.[1]

(A2) $\mathbf{s}[n] \succeq \mathbf{0}_K$, $\sum_{k=1}^{K} s_k[n] = 1$, $\forall n \in \mathcal{I}_L$.

Here, (A1) is a widely adopted baseline requirement, and (A2) nicely captures the unique geometric structure, i.e., the scatter simplex, of the non-negative source vectors widely observed in many nBSS problems [3], [4]. In (A2), the sources are non-negative by nature in nBSS, while the full-additivity $\mathbf{1}_K^T \mathbf{s}[n] = 1$ can be easily enforced (see Remark 2 in Section IV-E). Note that the simplex structure implied in (A2), that is

$$\mathbf{s}[n] \in \{\mathbf{s} \in \mathbb{R}^K \mid \mathbf{s} \succeq \mathbf{0}_K, \Sigma_{k=1}^{K} s_k = 1\} \triangleq \mathcal{T}_e \tag{2}$$

has been widely adopted and has led to some seminal NMF theories/methods in recent machine learning research (see [42], [43] and the references therein). As a result, every observation vector is confined within the simplex (a convex hull) defined by the K column vectors of the mixing matrix, as we have previously shown [3, Th. 1]. Fig. 1 shows the scatter simplex of a benchmark real gene expression data set. In light of Fig. 1, the assumption of $\mathbf{s}[n]$ being statistically independent Gaussian random vectors with zero mean, previously used in formulating MDL criteria for detecting the number of sources [26], [27], appears to be invalid and thus unsuitable for nBSS.

By (A2), a valid probability model for $\mathbf{s}[n]$ should have the domain $\mathcal{T}_e$ [see (2)], i.e., the standard simplex (or probability simplex) [36]. Given the fact that the well-known Dirichlet distribution provides K degrees of freedom to fit the sources on its domain $\mathcal{T}_e$, it is a very reasonable choice here and is, therefore, adopted to model $\mathbf{s}[n]$. Specifically, the Dirichlet distribution is defined as in [34], with concentration parameters $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]^T \succ \mathbf{0}_K$, that is

$$\text{Dir}(\mathbf{s}; \boldsymbol{\alpha}) \triangleq \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \cdot \prod_{k=1}^{K} s_k^{\alpha_k - 1}, \quad \forall \mathbf{s} \in \text{int } \mathcal{T}_e \tag{3}$$

[1]Note that (A1) does not assume non-negativity of $\mathbf{A}$, which is not needed in the ensuing theoretical development. However, $\mathbf{A} \in \mathbb{R}^{M \times K}$ is assumed to be of full column rank, implying $M \geq K$ (i.e., there are more observed samples than sources), which is well satisfied by a wide range of nBSS applications [37], [39]–[41]. The possibility of extension for the underdetermined case $M < K$ is left for future work.

where $\Gamma(\alpha) \triangleq \int_0^\infty x^{\alpha-1} e^{-x}\, dx$, $\alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$, and $\mathbf{s} = [s_1, \ldots, s_K]^T$. It is noted that the Dirichlet distribution has been extensively studied, and there is abundant theory applicable for effectively and elegantly deriving the induced MDL. If a user has prior knowledge about the underlying distribution that is non-Dirichlet, then one should derive the MDL based on such distribution—the order selection principle and associated steps (to be developed next) should remain valid. In the ensuing development, we will assume that $\mathbf{s}[n]$ can be well modeled or approximated by a multivariate Dirichlet distribution.

Assuming that the noise vector is from a stationary Gaussian process, independent of the sources, with zero-mean and covariance matrix given by $\sigma^2 \mathbf{I}_M$, where σ^2 is an unknown scalar constant and $\mathbf{I}_M$ is the $M \times M$ identity matrix, it follows that the probability density function (p.d.f.) for $\mathbf{x}[n]$ can be derived in the following lemma.

Lemma 1: The p.d.f. of the random vector $\mathbf{x}[n]$, w.r.t. the probability model parameterized by $\Theta^{(K)} = [\sigma^2, \mathbf{a}_1^T, \ldots, \mathbf{a}_K^T, \alpha_1, \ldots, \alpha_K]^T$ [see (1) and (3)], is given by

$$f(\mathbf{x}|\Theta^{(K)}) = \int_{\mathbf{y} \in \text{dom } h} g(\mathbf{x} - \mathbf{y}) \cdot h(\mathbf{y})\, d\mathbf{y} \tag{4}$$

in which $g(\cdot)$ and $h(\cdot)$ are, respectively, the p.d.f. of $\mathbf{w}[n]$ and $\mathbf{x}_0[n] = \mathbf{A}\mathbf{s}[n]$, that is

$$g(\mathbf{w}) = \frac{1}{\sqrt{(2\pi\sigma^2)^M}} \cdot \exp\left(-\frac{\mathbf{w}^T\mathbf{w}}{2\sigma^2}\right), \quad \forall \mathbf{w} \in \mathbb{R}^M \tag{5}$$

$$h(\mathbf{y}) = J(K, \mathbf{A}) \cdot \text{Dir}(\mathbf{A}^\dagger \mathbf{y}; \boldsymbol{\alpha}), \quad \forall \mathbf{y} \in \text{dom } h \tag{6}$$

where

$$\text{dom } h \triangleq \text{int conv}\{\mathbf{a}_1, \ldots, \mathbf{a}_K\} \subseteq \mathbb{R}^M \tag{7}$$

is the domain of $h(\cdot)$ [int $\mathcal{S}$ denotes the relative interior (w.r.t. aff$\mathcal{S}$) of the set $\mathcal{S}$] [36], $\mathbf{A}^\dagger \triangleq (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is the Moore–Penrose pseudoinverse of $\mathbf{A}$ [44], and $J(K, \mathbf{A})$ is the scale factor such that

$$\int_{\mathbf{y} \in \text{dom } h} h(\mathbf{y})\, d\mathbf{y} = 1; \tag{8}$$

note that the measures associated with the integrals in both (4) and (8) are the Lebesgue measure [45] w.r.t. aff(dom h) = aff$\{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$.

The proof of Lemma 1 is given in Appendix B. Then, assuming that $\mathbf{x}[1], \ldots, \mathbf{x}[L]$ are statistically independent [26], [28], the joint p.d.f. of $\mathbf{X}$ is given by

$$f(\mathbf{X}|\Theta^{(K)}) = \prod_{n=1}^{L} \left\{ \int_{\mathbf{y} \in \text{dom } h} \frac{J(K, \mathbf{A})}{\sqrt{(2\pi\sigma^2)^M}} \cdot \text{Dir}(\mathbf{A}^\dagger \mathbf{y}; \boldsymbol{\alpha}) \cdot \exp\left(-\frac{\|\mathbf{x}[n] - \mathbf{y}\|^2}{2\sigma^2}\right) d\mathbf{y} \right\} \tag{9}$$

where $\|\cdot\|$ denotes the Euclidean norm.

By (A1) and [3, Th. 1], it follows that the number of vertices defining the scatter simplex of $\mathbf{X}$ is K, or equivalently, is equal to the length of $\boldsymbol{\alpha}$ defining the Dirichlet distribution of $\mathbf{S}$. The number of sources K can, hence, be determined from the size of the smallest simplex of $\mathbf{X}$. The problem is that

the $\boldsymbol{\alpha}$ defining the Dirichlet distribution of $\mathbf{S}$ is unknown in practice. When estimated from a finite sample size, the resulting concentration parameters are all nonzero, thus making it difficult to determine the number of sources merely by "observing" the simplex of $\mathbf{X}$. In this paper, we pose the source detection problem as a model order selection problem and then formally derive an nBSS-MDL criterion.

III. Minimum Description Length Criterion

Given a set of L observations $\mathbf{X} = [\mathbf{x}[1], \ldots, \mathbf{x}[L]]$ and a parameterized family of probability densities $f(\mathbf{X}|\Theta^{(K)})$, the MDL criterion for model order selection, introduced by Rissanen [24], selects the model that best explains the data. Since each competing model can be used to encode the observed data, the two-part code length version of MDL selects the model that yields the minimum code length, given by [24], [25]

$$\mathrm{MDL}(K) = -\log\left(f\left(\mathbf{X} \mid \Theta_{\mathrm{ML}}^{(K)}\right)\right) + \tfrac{1}{2} D(\Theta^{(K)}) \log(L)$$

where $\Theta_{\mathrm{ML}}^{(K)}$ is the ML estimate of the parameter vector $\Theta^{(K)}$, and $D(\Theta^{(K)})$ is the number of freely adjusted parameters in $\Theta^{(K)}$. The first term is the well-known negative log-likelihood given the ML model parameters, corresponding to the code length for the data given the model. The second term is the penalty on model complexity, which is the code length for the model parameters.

Accurate detection of the number of sources via MDL heavily relies on the suitability of the chosen family of competing models under consideration. In our earlier work on multitissue compartment modeling of *in vivo* dynamic imaging data, we used the so-called latent variable model that was derived from the underlying pharmacokinetics equations [28]–[30], where $\mathbf{A}(\lambda_{1,2,3}, \beta_{1,2}, \beta_{\mathrm{ep},k})$ (describing the time activity curves and the tracer concentration in plasma) is parameterized by the flux rate constants in tissue-type $k \in \mathcal{I}_{K-1}$. While the initial applications of this modeling strategy have produced biologically plausible results, because of both $\mathbf{A}$ (parameterized) and $\mathbf{S}$ being considered deterministic unknowns, when the number of pixels L is large, the overparameterization associated with treating $\mathbf{S}$ as parameters can potentially lead to an underestimate of the number of sources. In our other earlier work on unsupervised deconvolution of tissue heterogeneity using mixed gene expression data [3], [46], we considered the widely adopted linear mixing model (before log-transform) [47], where both $\mathbf{A}$ and $\mathbf{S}$ are considered as deterministic unknowns. Again, when the number of genes L in $\mathbf{S}$ is large, it can lead to an underestimate of the number of sources. Although we have tried to parameterize $\mathbf{S}$ by gene clustering, the number of clusters may still be too large and order underestimation may still ensue. These approaches also require accurate estimates of $\mathbf{A}$ and $\mathbf{S}$, which may not be possible when the data are noisy or the identifiability condition is not met [3], [12].

These earlier efforts motivate the consideration of a parameterized statistical model for $\mathbf{S}$ (instead of considering $\mathbf{S}$ as model parameters) involving the Dirichlet distribution and its useful properties uniquely suited to nBSS. Specifically, our proposed MDL criterion can be expressed as

$$\mathrm{MDL}(K) = -\log\left(f\left(\mathbf{X} \mid \sigma_{\mathrm{ML}}^2, \{\mathbf{a}_{k,\mathrm{ML}}\}_{k=1}^K, \{\alpha_{k,\mathrm{ML}}\}_{k=1}^K\right)\right) + \frac{K(M+1)}{2}\log(L) \quad (10)$$

where the specific form of the likelihood function will be detailed in the next section, and the terms independent of K are omitted. While the Dirichlet distribution is not the only parametric model for describing non-negative source data, it captures the overall simplex data structure present in many nBSS applications, leading to the highly successful model order selection convincingly shown in our experimental results on real benchmark data sets [3]. We have also opted not to parameterize the mixing matrix $\mathbf{A}$, because $M, K \ll L$ in many nBSS applications means the improved data likelihood fit that comes from choosing all MK entries in $\mathbf{A}$ may compensate [in (10)] for the descriptive penalty associated with these parameters. Nevertheless, when sufficient prior knowledge about the structure of the mixing matrix is available, appropriate parameterization should be explored as we have done previously in modeling the pharmacokinetics of contrast-enhanced dynamic imaging data [29].

IV. Maximum Likelihood Estimation and Code Length Calculation

The MDL criterion is based on the ML estimates of the model parameters, but the ML estimator may need to be approximated if it induces a computationally intractable optimization problem. For the joint density function model given in (4), obtaining locally optimal ML estimates for the parameters will require a computationally complex joint optimization procedure and is further complicated by the convolution integral in (4). For the sake of computational efficiency, here we propose a greedy *approximate* ML estimation procedure, with the parameters estimated sequentially, rather than jointly. This greatly reduces the complexity of the parameter estimation and, as will be shown, does not in practice compromise the accuracy of the resulting MDL model order selection.

A. ML Estimation of the Noise Variance σ^2

To obtain the ML estimate of σ^2, we exploit the relationship between the population covariance matrix $\Sigma_{\mathbf{x}[n]}$, and its unbiased estimate $\widehat{\Sigma}_{\mathbf{x}[n]} \triangleq [1/(L-1)]\ \mathbf{U}\mathbf{U}^T$ given by [48], where

$$\mathbf{U} = [\mathbf{x}[1] - \mathbf{d}, \ldots, \mathbf{x}[L] - \mathbf{d}] \quad \text{and } \mathbf{d} = \frac{1}{L}\mathbf{X}\mathbf{1}_L \quad (11)$$

as outlined in the following lemma.

Lemma 2 *[48, Th. 2]:* Let $\tilde{\lambda}_i$ be the ith eigenvalue of $\widehat{\Sigma}_{\mathbf{x}[n]}$, where $\tilde{\lambda}_M > \cdots > \tilde{\lambda}_1$ with probability 1 (w.p.1). Also, let $\lambda_1, \ldots, \lambda_{M'}$ ($M' \leq M$) be the eigenvalues of the population covariance matrix $\Sigma_{\mathbf{x}[n]}$, with m_i denoting the multiplicity of λ_i. Assuming that $\lambda_{M'} > \cdots > \lambda_1$, the ML estimate of λ_k is then given by

$$\lambda_k^{(\mathrm{ML})} = \frac{L-1}{L\ m_k} \cdot \sum_{i=m_1+\cdots+m_{k-1}+1}^{m_1+\cdots+m_k} \tilde{\lambda}_i, \quad \forall k \in \mathcal{I}_{M'}. \quad (12)$$

From Lemma 2, the ML estimate of σ^2 is derived in the following corollary.

Corollary 1: The ML estimation of σ^2 is given by

$$\sigma^2_{\mathrm{ML}} = \frac{(L-1)\cdot(\tilde{\lambda}_1 + \cdots + \tilde{\lambda}_{M-K+1})}{L\cdot(M-K+1)} \tag{13}$$

where $\tilde{\lambda}_i$ is defined in Lemma 2.

The proof of Corollary 1 is given in Appendix A.

B. Estimation of $\mathbf{a}_1, \ldots, \mathbf{a}_K$

To estimate $\{\mathbf{a}_k\}$, we first write the explicit form of its log-likelihood function $\log \mathcal{L}(\mathbf{A}|\mathbf{X})$ according to (9) as

$$\begin{aligned}&\log \mathcal{L}(\mathbf{A}|\mathbf{X})\\ &= C_{\sigma^2} + L\cdot\log(J(K,\mathbf{A}))\\ &\quad + \sum_{n=1}^{L}\log\left\{\int_{\mathbf{y}\in\mathrm{dom}\ h}\mathrm{Dir}(\mathbf{A}^\dagger\mathbf{y};\boldsymbol{\alpha})\cdot\exp\left(\frac{\|\mathbf{x}[n]-\mathbf{y}\|^2}{-2\sigma^2}\right)d\mathbf{y}\right\}\end{aligned} \tag{14}$$

where C_{σ^2} is a constant independent of $\mathbf{A}$. However, the resulting ML problem, that is

$$\max_{\mathbf{a}_1,\ldots,\mathbf{a}_K\in\mathbb{R}^M}\ \log\ \mathcal{L}(\mathbf{A}|\mathbf{X}) \tag{15}$$

is difficult due to the complex, integral form of the nonconvex objective function defined by (14).

Therefore, instead of directly solving the ML problem (15), we reformulate it via some careful approximations and eventually approximate the ML problem by the following tractable geometry problem, detailed in Appendix C:

$$\begin{aligned}\mathbf{a}_{k,\mathrm{ML}} \approx \arg\min_{\mathbf{a}_k\in\mathcal{A}(\mathbf{C},\mathbf{d})}\ &\mathrm{vol}(\mathrm{conv}\{\tilde{\mathbf{a}}_1,\ldots,\tilde{\mathbf{a}}_K\})\\ \text{s.t.}\quad &\tilde{\mathbf{x}}[n]\in\mathrm{conv}\{\tilde{\mathbf{a}}_1,\ldots,\tilde{\mathbf{a}}_K\},\quad \forall n\in\mathcal{I}_L\\ &\tilde{\mathbf{a}}_k \triangleq \mathbf{C}^\dagger(\mathbf{a}_k-\mathbf{d}),\quad \forall k\in\mathcal{I}_K\end{aligned} \tag{16}$$

where $\mathcal{A}(\mathbf{C},\mathbf{d})$ denotes the $(K-1)$-dimensional affine hull that best fits the data cloud in the sense of least-squares fitting error [see (28)], in which $\mathbf{C}\triangleq[\boldsymbol{q}_1,\ldots,\boldsymbol{q}_{K-1}]\in\mathbb{R}^{M\times(K-1)}$ with $\boldsymbol{q}_i\in\mathbb{R}^M$ denoting the ith principal eigenvector (with unity norm) of $\mathbf{U}\mathbf{U}^T$ [see (11) for the definitions of $\mathbf{U}$ and $\mathbf{d}$], $\tilde{\mathbf{x}}[n]$ is the dimension-reduced affine representation of $\mathbf{x}[n]$ w.r.t. $\mathcal{A}(\mathbf{C},\mathbf{d})$ [11] [see (30)], and $\mathrm{vol}(\cdot)$ denotes simplex volume (cf. (34)).

The minimum volume simplex in (16) is known as the Craig simplex in the nBSS context [35], and a fast alternating direction method of multipliers [36]-based algorithm developed in [49] can be used to solve (16). The vertices of the Craig simplex approximate $\mathbf{a}_{k,\mathrm{ML}}$ well under certain practical conditions. Specifically, we show the Craig estimator (16) is theoretically equivalent to the ML estimator (15), when the source density is uniform [i.e., condition 1) in Theorem 1] [34], or when well-grounded points (WGPs) exist [i.e., condition 2) in Theorem 1; readers are referred to [28, Definition 3] for a rigorous introduction to the concept of WGP], as outlined in Theorem 1 given below.

Theorem 1: Assuming (A1), (A2), and the noiseless case, the problems (15) and (16) are equivalent when one of the following conditions holds true: 1) $\boldsymbol{\alpha}=\mathbf{1}_K$ and 2) $\forall k\in\mathcal{I}_K$, there exists an $n\in\mathcal{I}_L$ such that $\mathbf{s}[n]=\mathbf{e}_k$.

The proof of condition 1) in Theorem 1 is given in Appendix D. The proof of condition 2) in Theorem 1 follows directly from the observation that the Craig simplex is uniquely given by the data convex hull itself [under condition 2)], and is omitted here due to space limitations. From our experimental results on the benchmark real data sets, we have observed that even when the data are noisy with nonuniform density $\boldsymbol{\alpha}\neq\mathbf{1}_K$, the Craig estimator can still well-approximate the simplex structure of the data [50]. Practically, the $\{\mathbf{a}_{k,\mathrm{ML}}\}$ can also be approximated using the clustered CAM under alternative conditions [3], [29], [38], [50].

C. Estimation of Concentration Parameters $\alpha_1, \ldots, \alpha_K$

The estimation of $\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_K]^T$, from a finite set of source samples $\{\mathbf{s}[1],\ldots,\mathbf{s}[L]\}$, can be obtained using some benchmark techniques, including gradient ascent search, the expectation-maximization (EM) algorithm, and the Newton–Raphson method [51]. Here, a modified Newton–Raphson method is used to estimate $\boldsymbol{\alpha}_{\mathrm{ML}}$. An estimate of standardized $\{\mathbf{s}[1],\ldots,\mathbf{s}[L]\}$ is first obtained by solving the following fully constrained non-negative least squares problem [52] [see (1) and (2)]:

$$\begin{aligned}\hat{\mathbf{s}}[n] = \min_{\mathbf{s}'\in\mathbb{R}^K}\ &\|\mathbf{x}[n]-\mathbf{A}_{\mathrm{ML}}\cdot\mathbf{s}'\|^2\\ \text{s.t. }&\mathbf{s}'\succeq\mathbf{0}_K,\quad \mathbf{1}_K^T\mathbf{s}'=1\end{aligned} \tag{17}$$

where $\mathbf{A}_{\mathrm{ML}}\triangleq[\mathbf{a}_{1,\mathrm{ML}}\cdots\mathbf{a}_{K,\mathrm{ML}}]$; note that (17) is a convex optimization problem [36] and can be efficiently solved [52].

Then, the log-likelihood can be derived from (3) as

$$\begin{aligned}\log\mathcal{L}(\boldsymbol{\alpha}|\widehat{\mathbf{S}}) = L\log\Gamma(\alpha_0) - L\sum_{k=1}^{K}\log\Gamma(\alpha_k)\\ + \sum_{k=1}^{K}(\alpha_k-1)\sum_{n=1}^{L}\log(\hat{s}_k[n])\end{aligned} \tag{18}$$

whose gradient $\mathbf{g}$ and inverse Hessian $\mathbf{H}^{-1}$ (w.r.t. the reference point $\boldsymbol{\alpha}^{\mathrm{old}}$) can be verified as

$$[\mathbf{g}]_k = L\cdot\Psi\left(\mathbf{1}_K^T\boldsymbol{\alpha}^{\mathrm{old}}\right) - L\cdot\Psi([\boldsymbol{\alpha}^{\mathrm{old}}]_k) + \sum_{n=1}^{L}\log(\hat{s}_k[n])$$

$$\mathbf{H}^{-1} = \mathbf{Q}^{-1} - \frac{\mathbf{Q}^{-1}\mathbf{1}_K\mathbf{1}_K^T\mathbf{Q}^{-1}}{\left(L\cdot\Psi'\left(\mathbf{1}_K^T\boldsymbol{\alpha}^{\mathrm{old}}\right)\right)^{-1}+\mathbf{1}_K^T\mathbf{Q}^{-1}\mathbf{1}_K}$$

where $\Psi(x)\triangleq(d\ \log\Gamma(x)/d\ x)$ is the digamma function, $\Psi'(x)\triangleq(d\ \Psi(x)/d\ x)$ is the trigamma function, and $\mathbf{Q}$ is a diagonal matrix with its kth diagonal entry being $[\mathbf{Q}]_{kk}\triangleq -L\cdot\Psi'([\boldsymbol{\alpha}^{\mathrm{old}}]_k)$. Then, the Newton–Raphson method can be adopted to obtain a stationary point of (18) [53, p. 2] via the following iterative updating rule [54]:

$$\boldsymbol{\alpha}^{\mathrm{new}} = \boldsymbol{\alpha}^{\mathrm{old}} - \mathbf{H}^{-1}\mathbf{g}. \tag{19}$$

Note that as (18) is a unimodal concave function of $\boldsymbol{\alpha}$ [51], the only stationary point is its maximum [53, p. 2], so the sequence generated by (19) must converge to the ML solution.

Moreover, (19) can be initialized with an arbitrary $\boldsymbol{\alpha} \succ \mathbf{0}_K$ [thanks to the concavity of (18)], and is computationally very efficient since $\mathbf{Q}^{-1}$ in the inverse Hessian can be easily obtained [thanks to the diagonality of $\mathbf{Q}$].

D. Code Length Calculation

Based on the estimates of $\Theta^{(K)}$, the code length associated with the MDL criterion [see (10)] can be theoretically determined. However, due to the complexity of the log-likelihood term requiring a high dimensional integral [see (9)], an approximation of this integral is proposed, which is computationally efficient and exploits Monte Carlo integration [55].

We observe that the support of $h(\mathbf{y}|\Theta_{\rm ML}^{(K)})$ (i.e., $\text{conv}\{\mathbf{a}_{1,\rm ML}, \ldots, \mathbf{a}_{K,\rm ML}\}$) is equivalent to the domain of the integral, allowing a convenient adoption of Monte Carlo integration. Accordingly, we recognize from Lemma 1 that

$$f(\mathbf{x}[n] \mid \Theta^{(K)}) = \mathbb{E}_{\mathbf{Y}}[g(\mathbf{x}[n] - \mathbf{Y}) \mid \Theta^{(K)}] \tag{20}$$

where $\mathbb{E}_{\mathbf{Y}}[\cdot|\cdot]$ denotes the conditional expectation, and $\mathbf{Y}$ is a random vector w.r.t. the p.d.f. $h(\mathbf{y}|\Theta_{\rm ML}^{(K)})$. The integral in (9) is then approximated using the sample average estimate of (20)

$$I_n^{(m)} \triangleq \frac{1}{m}\sum_{i=1}^{m} g\big(\mathbf{x}[n] - \mathbf{y}_i \mid \Theta_{\rm ML}^{(K)}\big) \tag{21}$$

where $\mathbf{y}_1, \ldots, \mathbf{y}_m$ are independent and identically distributed (i.i.d.) generated by $h(\mathbf{y}|\Theta_{\rm ML}^{(K)})$, and m is the (large) number of trials (see the discussions in Remark 1 below). To generate the sequence of $\mathbf{y}_i$ in (21), we notice the one-to-one correspondence between the standard simplex $\text{conv}\{\mathbf{e}_1, \ldots, \mathbf{e}_K\} \subseteq \mathbb{R}^K$ (i.e., the closure of the domain of Dirichlet distribution) and the ML endmembers' simplex $\text{conv}\{\mathbf{a}_{1,\rm ML}, \ldots, \mathbf{a}_{K,\rm ML}\} \subseteq \mathbb{R}^M$ (i.e., the support of $h(\mathbf{y}|\Theta_{\rm ML}^{(K)})$), under the linear mapping defined by the matrix $\mathbf{A}_{\rm ML} \in \mathbb{R}^{M\times K}$. Thus, by (A1), one can first generate an i.i.d. sequence $\mathbf{s}_1, \ldots, \mathbf{s}_m$ following the Dirichlet distribution with parameter vector $\boldsymbol{\alpha}_{\rm ML}$, and then obtain the desired sequence $\mathbf{y}_1, \ldots, \mathbf{y}_m$ by linearly mapping each $\mathbf{s}_i$ to $\mathbf{y}_i \triangleq \mathbf{A}_{\rm ML}\mathbf{s}_i$. Note that we have to generate the Dirichlet samples $\mathbf{s}_i$, instead of using the readily available $\hat{\mathbf{s}}[n]$ by (17), because $\hat{\mathbf{s}}[n]$ estimated from the real-world data set may not be Dirichlet distributed as desired. Fortunately, the Dirichlet random vector $\mathbf{s}_i$ can be efficiently generated in a typical numerical computing environment such as MATLAB, by normalizing a K-vector whose kth entry is gamma-distributed with shape parameter α_k and scale parameter 1 [34]. The proposed MDL algorithm is summarized in Algorithm 1.

E. Discussion

Before we finish this section, some analysis of the proposed MDL algorithm, preprocessing strategies, and theoretical imperfection are discussed in the following remarks.

Remark 1: According to the Law of Large Numbers [56], one can see that $I_n \triangleq f(\mathbf{x}[n] \mid \Theta_{\rm ML}^{(K)}) = \lim_{m\to\infty} I_n^{(m)}$ [see (21)]. Moreover, by the central limit theorem (CLT), or, more precisely, the Lindeberg–Levy CLT [56], one can show that $\sqrt{m}\cdot(I_n^{(m)} - I_n)$ converges in distribution to a zero-mean Gaussian variable with the same variance as the random variable $g(\mathbf{x}[n] - \mathbf{Y} \mid \Theta_{\rm ML}^{(K)})$ [56], and hence, the error term in the approximation of (21) (i.e., $\epsilon_m \triangleq I_n^{(m)} - I_n$) is of the order $\mathcal{O}(1/\sqrt{m})$, where a constant term independent of m is dropped; note that we are analyzing how ϵ_m decreases as m increases. Hence, for a desired precision ϵ, the required m is of the order $\mathcal{O}\left(1/\epsilon^2\right)$—the precision of the approximation in (21) [for computing the code length of (20)] is solely controlled by the number of trials m—a remarkable property of Monte Carlo integration. To understand such efficiency, one should note that to achieve a desired precision, the required m usually grows exponentially with the dimension M if we use naive approximation, such as by Riemann integration [57].

Algorithm 1 Pseudo-Code for the MDL Algorithm

Given nBSS data matrix $\mathbf{X}$.
for $K = K_{\min} : K_{\max}$ **do**
  Approximate $\Theta_{\rm ML}^{(K)} = [\sigma^2_{\rm ML}, \mathbf{a}_{1,\rm ML}^T, \ldots, \mathbf{a}_{K,\rm ML}^T, \boldsymbol{\alpha}_{\rm ML}^T]^T$ by (13), (16) and (19).
  Generate i.i.d. sequence of $\mathbf{s}_i \sim \text{Dir}(\mathbf{s}; \boldsymbol{\alpha}_{\rm ML})$ and obtain $\mathbf{y}_i = \mathbf{A}_{\rm ML}\mathbf{s}_i,\ \forall i \in \mathcal{I}_m$.
  Approximate $-\log(f(\mathbf{X} \mid \Theta_{\rm ML}^{(K)}))$ by (9) and (21).
  Calculate $\text{MDL}(K)$ by (10).
end for
Output $\widehat{K} = \arg\min\{\text{MDL}(K) \mid K_{\min} \le K \le K_{\max}\}$.

Remark 2 (Standardization of nBSS Data:) By (A2), the full-additivity (i.e., $\mathbf{1}_K^T\mathbf{s}[n] = 1$) is assumed to validate the adoption of the Dirichlet modeling, which elegantly captures the simplex structure observed in many nBSS data. Although full-additivity may not be satisfied by nature for some nBSS applications, it can be easily enforced via a simple and judicious preprocessing strategy, so that the proposed MDL algorithm can still be applied. The technique is recalled here; for illustration purpose, the noiseless scenario is assumed. In particular, if full-additivity is violated (or not ensured), one can standardize each $\mathbf{x}[n]$ to be

$$\boldsymbol{x}[n] \triangleq \frac{\mathbf{x}[n]}{\mathbf{1}_M^T\mathbf{x}[n]} = \sum_{k=1}^{K} \tilde{s}_k[n]\boldsymbol{a}_k \tag{22}$$

where $\tilde{s}_k[n] \triangleq [(\mathbf{1}_M^T\mathbf{a}_k)/(\mathbf{1}_M^T\mathbf{x}[n])]\cdot s_k[n]$ is the standardized source vector and $\boldsymbol{a}_k \triangleq \mathbf{a}_k/(\mathbf{1}_M^T\mathbf{a}_k)$. Now, the source vector

$$\tilde{\mathbf{s}}[n] \triangleq [\tilde{s}_1[n], \ldots, \tilde{s}_K[n]]^T \in \mathbb{R}^K$$

corresponding to the standardized data $\boldsymbol{x}[n]$ can be easily shown to satisfy full-additivity, i.e., $\mathbf{1}_K^T\tilde{\mathbf{s}}[n] = 1$. The geometrical meaning of (22) is to perform *perspective projection* [36] of the data, which originally locate in the conic hull [36]

$$\text{conic}\{\mathbf{a}_1, \ldots, \mathbf{a}_K\} \triangleq \left\{\mathbf{x} = \sum_{k=1}^{K} s_k\mathbf{a}_k \mid s_k \ge 0,\ \forall k \in \mathcal{I}_K\right\}$$

onto the simplex $\text{conv}\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K\}$. This standardization procedure can be easily implemented and has been often used to reveal the simplex structure of nBSS data (see the normalization procedure performed in the non-negative dependent source separation [11, eq. (4)] and the perspective mapping

performed in pharmacokinetic analysis for prostate tumor characterization [58, eq. (16)].

Remark 3 The Craig estimator is a very good approximation to the ML estimator of $\mathbf{A}$ when the sources distribute uniformly enough on the standard simplex, or when WGPs exist (by Theorem 1), but the two estimators are not equivalent in general. Although it seems feasible to further improve the ML estimate of $\mathbf{A}$, e.g., by employing an EM algorithm, this may result in a much higher computational complexity. Fortunately, this does not seem necessary in our case studies. Our experiments will demonstrate that the Craig estimator is able to capture the simplex structure embedded in nBSS data.

Remark 4 For complex-valued $\mathbf{A}$, we can stack its real and imaginary parts as $\overline{\mathbf{A}} = [\mathbf{A}_R^T,\ \mathbf{A}_I^T]^T \in \mathbb{R}^{2M\times K}$ (and $\overline{\mathbf{X}} = [\mathbf{X}_R^T,\ \mathbf{X}_I^T]^T \in \mathbb{R}^{2M\times L}$ is similarly defined). Then, it can be verified that if $(\mathbf{X}, \mathbf{A}, \mathbf{S})$ satisfies (A1) and (A2), so does $(\overline{\mathbf{X}}, \overline{\mathbf{A}}, \mathbf{S})$. In other words, the proposed MDL method can still be applied (with input $\overline{\mathbf{X}}$), without affecting the efficiency of Monte Carlo integration in the code length calculation (even though the data size is doubled; see Remark 1).

V. Experimental Results

In this section, we demonstrate the superior efficacy of the proposed MDL algorithm (i.e., Algorithm 1) using both synthetic and real benchmark nBSS data. For performance comparisons, three benchmark model order selection algorithms, including Wax' MDL algorithm [26], the eigenvalue thresholding-based VD algorithm [16], and the convex geometry-based GENE algorithm [18], are also used to estimate the number of sources for these data sets; we implemented Wax's MDL algorithm, and the source codes for VD and GENE were provided by their authors. Note that there are two versions of the GENE algorithm: one estimating the convex hull of the data (GENE-CH) and another estimating the affine hull of the data (GENE-AH) [18]. Since GENE-CH tends to overestimate the number of sources, we used the GENE-AH version in our experiments [18].

In Theorem 1, we have proven that the Craig estimator is equivalent to the ML estimator when the sources follow a uniform Dirichlet distribution. In our experiments on real-world data sets, where the sources are estimated to be nonuniformly distributed, the Craig estimator is still capable of capturing the simplex structure embedded in the nBSS data under test, as seen next.

A. Synthetic Data With Isotropic Gaussian Noise

In this section, we evaluate the performance of the proposed MDL algorithm using synthetic data sets composed of $K = 5$ sources. Specifically, the source vectors $\mathbf{s}[n] \in \mathbb{R}^K$ are i.i.d. generated following the Dirichlet distribution with parameter $\boldsymbol{\alpha} = \mathbf{1}_K$ [see (3) and (A2)]. Then, the sources are used to generate $L = 1000$ noise-free synthetic data $\mathbf{x}_0[n] \triangleq \mathbf{A}\mathbf{s}[n]$ according to the nBSS model in (1), where the mixing matrix $\mathbf{A}$ of $M = 100$ bands is randomly generated with each entry normally distributed by $\mathcal{N}(0, 1)$, which enforces (A1) w.p.1. Finally, we add i.i.d. zero-mean Gaussian noise $\mathbf{w}[n]$ with variance σ^2 to $\mathbf{x}_0[n]$ for different values of SNR defined as

$$\text{SNR} \triangleq \frac{\sum_{n=1}^{L} \|\mathbf{x}_0[n]\|^2}{\sigma^2 ML}. \tag{23}$$

For each SNR $\in \{0, 5, 10, 15\}$ (dB), we synthesized 50 synthetic data sets that were then processed by Wax's MDL algorithm [26], the VD algorithm [16], the GENE-AH algorithm [18], and the proposed MDL algorithm [with the number of trials set to $m = 1000$ in the Monte Carlo integration; see (21)], respectively. As the VD and GENE-AH algorithms are developed based on the Neyman–Pearson detection theory, a well-tuned false alarm probability P_{FA} is critical to their performance. In view of this, we tested the efficacy of these two algorithms using the range of $P_{\text{FA}} \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, so as to yield their best performances. Moreover, the two MDL algorithms and the GENE algorithm require a preset value of maximum possible number of sources, which was set as $K_{\max} = 16$ for all three algorithms. The mean and standard deviation of the detected number of sources $\widehat{K}$ over the 50 independent realizations are displayed in the top-left block of Table I, where each boldface number denotes the best performance for a particular SNR among the four algorithms under test.

Apparently, the two MDL algorithms can unsupervisedly yield perfect estimates of the number of sources $\widehat{K} = 5$ for each scenario under test. While the VD algorithm tends to underestimate the number of sources, the GENE-AH algorithm can yield a good estimate of the number of sources with a properly selected P_{FA}. Nevertheless, the performance of GENE-AH appears to be sensitive to the false alarm probability; actually, it can be verified that, for given data, K estimated by GENE-AH gets smaller as P_{FA} decreases [18]. Note that the best setting of P_{FA} is often data-dependent and unknown ahead of time. The performance of the two MDL algorithms, by contrast, does not sensitively depend upon the proper tuning of any hyperparameters.

B. Synthetic Data With Nonisotropic Gaussian Noise

Next, we study robustness to nonisotropic noise for these model order selection algorithms. To this end, we corrupted the data $\mathbf{x}_0[n]$ by nonisotropic Gaussian noise generated following a standard procedure [18]. Specifically, the noise variance σ_i^2 for the ith observation (or the ith band) is given by [18]

$$\sigma_i^2 = \sigma^2 \frac{\exp(-(i - M/2)^2/2\tau^2)}{\sum_{j=1}^{M} \exp(-(j - M/2)^2/2\tau^2)}, \quad \forall i = 1, \ldots, M$$

where σ^2 is defined by (23), and $\tau > 0$ is a parameter that controls the degree of nonisotropy of the noise. The larger the value of τ, the more isotropic the noise becomes. When $\tau = \infty$, it corresponds to perfectly isotropic noise, i.e., the scenario studied in Section V-A. The simulation results for $\tau \in \{30, 20, 10\}$ in terms of the mean and the standard deviation of the detected number of sources $\widehat{K}$ (over 50 independent realizations) are displayed in Table I.

From Table I, one can observe that Wax's MDL algorithm tends to overestimate the number of sources $\widehat{K}$ as the noise

TABLE I

MEAN ± STANDARD DEVIATION OF THE ESTIMATED NUMBER OF (DIRICHLET) SOURCES $\widehat{K}$ FOR TRUE $K = 5$, OVER 50 INDEPENDENT RUNS, FOR VARIOUS MODEL ORDER SELECTION ALGORITHMS, WITH DIFFERENT VALUES OF SNR AND τ (A MEASURE OF ISOTROPY OF THE NOISE DISTRIBUTION)

Methods	P_{FA}	isotropic noise ($\tau = \infty$)				slightly non-isotropic ($\tau = 30$)			
		SNR (dB)				SNR (dB)			
		0	5	10	15	0	5	10	15
VD	10^{-3}	2.86±0.70	3.04±0.78	3.14±0.78	3.18±0.80	2.88±0.72	3.02±0.77	3.16±0.82	3.16±0.77
	10^{-4}	2.56±0.54	2.60±0.61	2.64±0.63	2.72±0.67	2.54±0.54	2.58±0.57	2.62±0.64	2.72±0.67
	10^{-5}	2.40±0.49	2.52±0.54	2.52±0.54	2.48±0.50	2.38±0.49	2.52±0.54	2.48±0.50	2.48±0.50
	10^{-6}	2.22±0.42	2.28±0.45	2.28±0.45	2.28±0.45	2.18±0.39	2.30±0.46	2.28±0.45	2.26±0.44
GENE-AH	10^{-3}	7.62±1.21	7.42±1.18	7.42±1.26	7.22±1.13	7.20±1.23	7.30±0.93	6.68±0.94	6.98±0.94
	10^{-4}	6.44±0.93	6.20±0.95	6.04±0.88	6.06±0.89	6.18±1.06	5.98±0.82	5.78±0.68	5.76±0.69
	10^{-5}	5.72±0.70	5.54±0.61	5.38±0.60	5.42±0.57	5.56±0.76	5.40±0.53	5.34±0.48	5.32±0.51
	10^{-6}	5.36±0.53	5.16±0.37	5.12±0.33	5.10±0.30	5.22±0.42	5.14±0.35	5.08±0.27	5.08±0.27
MDL (Wax)	–	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**
MDL	–	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**
Methods	P_{FA}	moderately non-isotropic ($\tau = 20$)				highly non-isotropic ($\tau = 10$)			
		SNR (dB)				SNR (dB)			
		0	5	10	15	0	5	10	15
VD	10^{-3}	2.84±0.68	3.06±0.77	3.12±0.77	3.18±0.77	14.48±2.52	14.78±2.58	14.84±2.63	14.90±2.65
	10^{-4}	2.52±0.54	2.58±0.57	2.64±0.66	2.72±0.67	9.22±1.95	9.40±2.02	9.46±2.09	9.50±2.08
	10^{-5}	2.40±0.49	2.48±0.54	2.48±0.50	2.46±0.50	6.56±1.45	6.68±1.48	6.72±1.46	6.74±1.43
	10^{-6}	2.18±0.39	2.30±0.46	2.30±0.46	2.28±0.45	**4.94±1.28**	5.04±1.32	5.06±1.33	5.04±1.31
GENE-AH	10^{-3}	7.40±0.99	6.76±1.00	6.56±0.97	6.52±0.95	7.98±1.25	6.98±1.17	6.48±1.11	6.32±1.00
	10^{-4}	6.20±0.76	5.92±0.85	5.70±0.71	5.68±0.62	6.84±1.25	6.02±0.87	5.70±0.91	5.66±0.77
	10^{-5}	5.62±0.75	5.28±0.45	5.26±0.44	5.26±0.44	6.20±1.16	5.60±0.73	5.30±0.61	5.16±0.37
	10^{-6}	5.22±0.58	5.14±0.35	5.08±0.27	5.08±0.27	5.86±0.93	5.38±0.64	5.12±0.33	5.10±0.30
MDL (Wax)	–	5.72±0.86	5.68±0.82	5.68±0.82	5.68±0.82	16.00±0.00	16.00±0.00	16.00±0.00	16.00±0.00
MDL	–	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	**5.00±0.00**	7.66±0.66	**5.02±0.14**	**5.00±0.00**	**5.00±0.00**

distribution gets more nonisotropic. It fails to correctly detect the number of sources when the data are corrupted by highly nonisotropic noise (i.e., $\tau = 10$), even for the scenario of high SNR = 15 (dB) (see the bottom-right block of Table I). This should be partly attributable to the fact that Wax's MDL algorithm was developed without considering the "simplex structure" of the nBSS data, making its performance rely more on the isotropic noise assumption [26], and should also be partly attributable to the fact that the synthetic data are generated based on a Dirichlet distribution (instead of a Gaussian distribution as assumed in [26]).

To be exact, Wax and Kailath [26] assumed that the data $\mathbf{x}[n]$ simply follow a Gaussian distribution, based on which the data cloud would configure as an ellipsoid. Alternatively, by (A2), the data cloud would be expected to shape like the simplex $\text{conv}\{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$, and hence it may be advisable to incorporate a (linearly transformed) Dirichlet distribution to depict how the data distribute on this simplex. Even when the data are corrupted by nonisotropic noise, the noisy data cloud may still preserve the simplex structure to some degree.

Since our MDL algorithm takes this simplex structure into account, it appears to be more robust against the nonisotropic noise effect, except for the scenario of very low SNR and highly nonisotropic noise [i.e., $(\tau, \text{SNR}) = (10, 0\text{ dB})$]. It is worth mentioning that the convex geometry-based GENE-AH algorithm is also devised under a framework wherein the noiseless data are assumed to form a simplex (see [18]). That is why its performance does not degrade too much as the noise becomes more nonisotropic. The importance of considering the simplex structure of the data when performing model order selection will be further demonstrated next by our experimental studies involving real data sets.

C. Synthetic Data From a Non-Dirichlet Source

We have assumed that the source vectors can be well-modeled or approximated by a Dirichlet distribution. To understand the efficacy of our MDL criterion when this hypothesis is violated, we study the scenario where the synthetic data are generated using non-Dirichlet sources. Specifically, we follow the same data generation procedure as described in Section V-A and corrupt the data by isotropic or nonisotropic Gaussian noise as described in Section V-B. However, the sources are obtained from $K = 6$ remote sensing images (i.e., materials' abundance maps [33]) as displayed in Fig. 2; each image contains $L = 100 \times 100$ pixels and represents one of the rows in $\mathbf{S} \in \mathbb{R}^{6\times 10000}$ after vectorization. This set of images has been widely used in generating synthetic remote sensing data [59] and is regarded as a non-Dirichlet source [38, Sec. IV.D].

The simulation results are summarized in Table II. For the scenario with isotropic noise, the two MDL algorithms perform best, and GENE-AH also works well if the false alarm probability is properly chosen. For nonisotropic cases (i.e., $\tau \in \{30, 20, 10\}$), the proposed MDL criterion still perfectly detects the number of sources, while GENE-AH performs second best; for the highly nonisotropic case, the VD algorithm and Wax's MDL algorithm fail to correctly detect the number of sources. These simulation results indicate that the proposed MDL criterion is potentially insensitive to nonisotropic noise and non-Dirichlet sources.

TABLE II

MEAN ± STANDARD DEVIATION OF THE ESTIMATED NUMBER OF (NON-DIRICHLET) SOURCES $\widehat{K}$ FOR TRUE $K = 6$, OVER 50 INDEPENDENT RUNS, FOR VARIOUS MODEL ORDER SELECTION ALGORITHMS, WITH DIFFERENT VALUES OF SNR AND τ (A MEASURE OF ISOTROPY OF THE NOISE DISTRIBUTION)

Methods	P_{FA}	isotropic noise ($\tau = \infty$) SNR (dB)				slightly non-isotropic ($\tau = 30$) SNR (dB)			
		0	5	10	15	0	5	10	15
VD	10^{-3}	5.90±0.30	5.90±0.30	5.90±0.30	5.92±0.27	5.90±0.30	5.90±0.30	5.90±0.30	5.92±0.27
	10^{-4}	5.84±0.37	5.86±0.35	5.86±0.35	5.84±0.37	5.84±0.37	5.88±0.33	5.88±0.33	5.86±0.35
	10^{-5}	5.84±0.37	5.82±0.39	5.82±0.39	5.82±0.39	5.80±0.40	5.86±0.35	5.82±0.39	5.82±0.39
	10^{-6}	5.76±0.43	5.82±0.39	5.82±0.39	5.82±0.39	5.78±0.42	5.80±0.40	5.82±0.39	5.82±0.39
GENE-AH	10^{-3}	6.92±0.72	6.90±0.79	6.78±0.71	6.70±0.71	7.04±0.67	6.86±0.64	6.64±0.60	6.58±0.61
	10^{-4}	6.16±0.37	6.14±0.35	6.14±0.35	6.08±0.27	6.42±0.50	6.26±0.44	6.14±0.35	6.12±0.33
	10^{-5}	6.02±0.14	**6.00±0.00**	6.02±0.14	6.02±0.14	6.12±0.33	6.04±0.20	6.02±0.14	**6.00±0.00**
	10^{-6}	**6.00±0.00**	**6.00±0.00**	6.02±0.14	**6.00±0.00**	6.02±0.14	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**
MDL (Wax)	–	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	16.00±0.00	16.00±0.00	16.00±0.00	16.00±0.00
MDL	–	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**

Methods	P_{FA}	moderately non-isotropic ($\tau = 20$) SNR (dB)				highly non-isotropic ($\tau = 10$) SNR (dB)			
		0	5	10	15	0	5	10	15
VD	10^{-3}	6.24±0.72	6.68±0.82	7.26±0.94	7.88±1.22	40.66±2.63	45.52±2.80	49.20±2.87	51.78±2.84
	10^{-4}	5.98±0.55	6.18±0.66	6.46±0.73	6.58±0.81	36.56±2.90	41.32±3.04	44.66±3.22	46.56±3.07
	10^{-5}	5.84±0.42	5.96±0.49	6.06±0.62	6.20±0.76	33.22±2.51	37.82±2.83	40.46±3.20	42.24±3.32
	10^{-6}	5.78±0.42	5.80±0.40	5.90±0.51	5.92±0.53	30.84±2.41	34.98±2.46	37.72±2.66	39.10±2.73
GENE-AH	10^{-3}	7.52±0.86	7.26±0.72	6.88±0.69	6.66±0.66	9.88±1.44	8.82±1.12	7.92±1.16	7.30±0.93
	10^{-4}	6.78±0.71	6.44±0.50	6.24±0.48	6.24±0.43	9.00±1.05	8.12±1.00	7.32±1.02	6.76±0.85
	10^{-5}	6.26±0.49	6.10±0.30	6.02±0.14	6.04±0.20	8.40±1.05	7.44±1.20	6.68±0.74	6.40±0.57
	10^{-6}	6.02±0.14	6.02±0.14	**6.00±0.00**	**6.00±0.00**	7.80±1.07	6.94±1.00	6.44±0.64	6.22±0.42
MDL (Wax)	–	16.00±0.00	16.00±0.00	16.00±0.00	16.00±0.00	16.00±0.00	16.00±0.00	16.00±0.00	16.00±0.00
MDL	–	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**	**6.00±0.00**

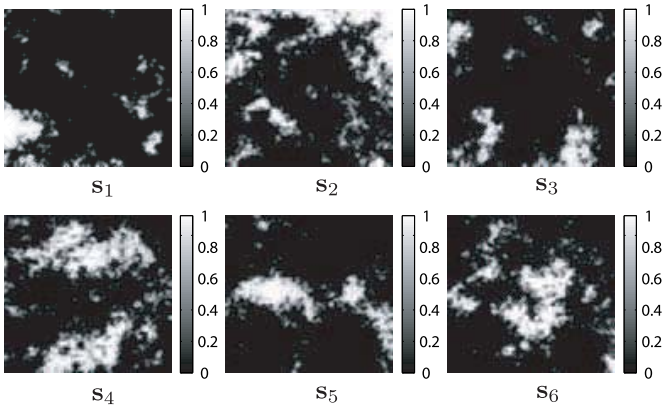


Fig. 2. Remote sensing images for generating non-Dirichlet sources, where the kth image $\mathbf{s}_k$ [defined below (1)] has $L = 100 \times 100$ pixels and corresponds to the kth row of $\mathbf{S} \in \mathbb{R}^{6\times 10000}$, $\forall k = 1, \ldots, 6$.

However, to rigorously justify its robustness to nonisotropic noise and non-Dirichlet sources, it may involve nontrivial theoretical analysis that is left as our future research.

D. Real Benchmark Rat Cell Type-Specific Gene Expression Data (RD1)

In this paper, we test the proposed MDL algorithm on a benchmark gene expression data—known as GSE19830 [37]—that is used for cell type-specific significance analysis of microarrays from the rat genome. In particular, it serves as a benchmark data set for analyzing the contribution of each cell type to the total measured gene expression in a given biological sample. The mathematical model for characterizing this data set can be found in [37], which matches our signal model (1) and is briefly next described.

This data set contains $K = 3$ sources, corresponding to three different rat cell types—brain, liver, and lung [37]. There are 11 biologically mixed heterogeneous samples, with three replicates for each sample, resulting in a total of $M = 33$ channels in this data set. In each sample, there are $L = 31042$ probes; note that each gene may correspond to multiple probes. Then, the ith entry of $\mathbf{x}[n]$ denotes the measured expression value of probe n for sample i, and the ith entry of $\mathbf{a}_k$ is the mixing abundance of cell type k in sample i. Moreover, $s_k[n]$ represents the gene expression measured from probe n for cell-type k.

Obviously, there is no reason to require that the three gene expressions (for brain, liver, and lung) measured by probe n satisfy $s_{\text{brain}}[n] + s_{\text{liver}}[n] + s_{\text{lung}}[n] = 1$, and hence full-additivity is expected to be violated for most probes in this data set. In view of this, we standardized this data set by the technique given in Remark 2 (i.e., normalizing each column of $\mathbf{X}$), so as to enforce source full-additivity, before it was processed by the proposed MDL algorithm. The description length for this data set versus different values of K is shown in Fig. 3, with model-order $K = 3$, yielding the shortest description length. Hence, the proposed MDL algorithm has correctly detected the number of sources $\widehat{K} = 3$, even though the corresponding Dirichlet distribution is estimated to be nonuniform, i.e., $\boldsymbol{\alpha} = [2.66,\ 1.70,\ 3.06]^T$. We also processed this data set using Wax's MDL algorithm, which detects $\widehat{K} = 27$ sources, grossly overestimating the ground-truth value. One of the major reasons is that the noise powers for

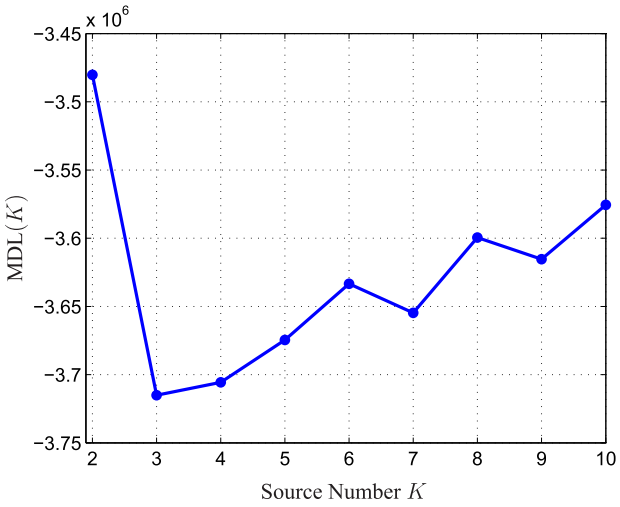


Fig. 3. MDL curve for real benchmark rat cell type-specific gene expression data (GSE19830), where the detected number of sources is $\widehat{K} = 3$.

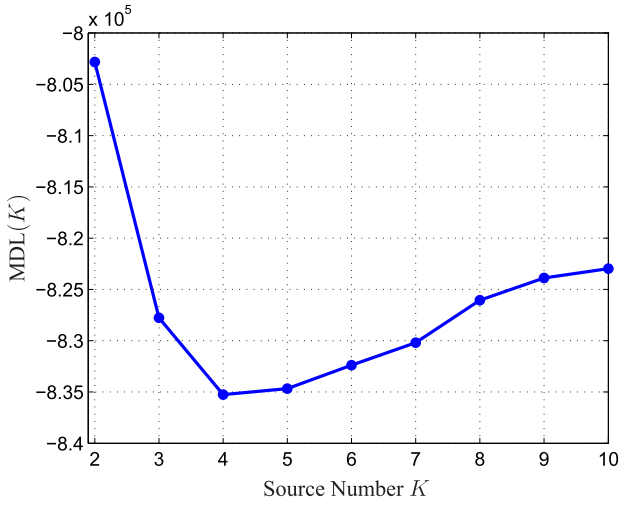


Fig. 4. MDL curve for real benchmark human blood microarray data (GSE11058), where the detected number of sources is $\widehat{K} = 4$.

different biologically mixed samples could be nonisotropic, as alluded to in our simulation studies (see Section V-B). Another is that Wax's model does not capture the simplex structure that may be intrinsic to the data. This shows that simply using a Gaussian distribution is not sufficient to capture the characteristic of this gene expression (nBSS) data. Furthermore, both VD and GENE yield model-order estimates that are grossly in error (for a wide range of $P_{\text{FA}} \in \{10^{-1}, 10^{-2}, \ldots, 10^{-6}\}$), and hence we do not even show their results. A recently developed algorithm, called GLAD [60], also estimates the number of cell types for this data set based on BIC; it gives a slightly overestimated result of $\widehat{K} = 5$.

E. Real Benchmark Human Blood Microarray Data (RD2)

Next, we test the proposed MDL algorithm on a benchmark data set, termed GSE11058 [39], used for studying Systemic Lupus Erythematosus (SLE) disease, a systemic autoimmune disease. The immune systems of patients suffering from SLE mistakenly attack healthy tissues and damage multiple organs. The microarray expression deconvolution technique (MEDT) has been applied to blindly analyze the SLE disease [39], but correct model order selection is crucial for MEDT to reliably characterize changes in mixed populations of blood cells. Let us describe how the nBSS model (1) fits this data set [39].

In this data set, there are $K = 4$ constituent subpopulations (i.e., sources) that correspond to four (phenotypically very similar) transformed cell lines of immune origin in blood—Raji (from B-cell), IM-9 (from B-cell), Jurkat (from T-cell), and THP-1 (from monocyte) [39]. This data set contains four sample profiles, each with three replicates, resulting in $M = 12$ biologically mixed expression profiles of the four subpopulations. Around a third of the probes, corresponding to genes with too low or too high signal intensity, are reported to be unreliable [29] and hence are eliminated from the data set; finally, $L = 35498$ probes are retained in each sample profile. Then, the ith entry of $\mathbf{x}[n]$ denotes the measured microarray data from probe n in the ith biological sample profile, and the ith entry of $\mathbf{a}_k$ is the mixing abundance of the kth constituent subpopulation in the ith mixed sample profile. Moreover, $s_k[n]$ is the expression level associated with probe n for the kth subpopulation.

Since there is no literature supporting the validity of source full-additivity for this data set, we again performed standardization before processing this data set by the proposed MDL algorithm. The obtained MDL curve for this data set is shown in Fig. 4; it successfully indicates the presence of $\widehat{K} = 4$ constituent subpopulations in the mixed biological samples. Note that the estimated Dirichlet distribution is highly nonuniform with $\boldsymbol{\alpha} = [7.41,\ 7.93,\ 9.07,\ 9.08]^T$, but the proposed nBSS-MDL criterion still successfully captures the (presumably) embedded simplex structure of this data set. We also performed model order selection for this data set using Wax's MDL algorithm, which incorrectly detected $\widehat{K} = 10$. Again, both VD and GENE yield model-order estimates that are grossly in error for a wide range of P_{FA}, and so their results are omitted.

F. Real Benchmark Brain Disease-Related Molecular Data (RD3)

Our next experiment is conducted on a benchmark data set, whose accession code is GSE19380 [40], used for studying brain molecular changes (or histological abnormalities) caused by Huntington's disease. Characterizing molecular changes in the diseased brain plays a pivotal role in revealing pathophysiological mechanisms and developing associated targeted drugs, and this raises the so-called differential expression analysis problem (DEAP) [40], where the very first task for DEAP is to choose a correct model order in order to address a major confounding factor called tissue heterogeneity. This problem is caused by limited imaging resolution and can be modeled as an nBSS problem, where (1) can be utilized for the mathematical modeling of this data set as next discussed [40].

This data set is generated from $K = 4$ messenger ribonucleic acids (mRNAs) (i.e., sources), which are neuronal

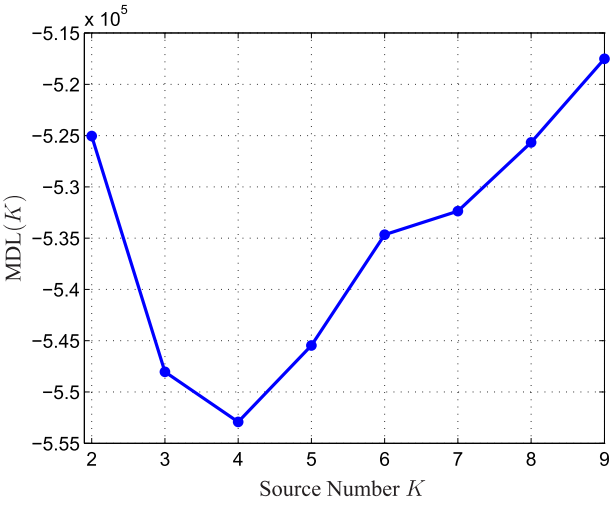


Fig. 5. MDL curve for real benchmark brain disease-related molecular data (GSE19380), where the detected number of sources is $\widehat{K} = 4$.

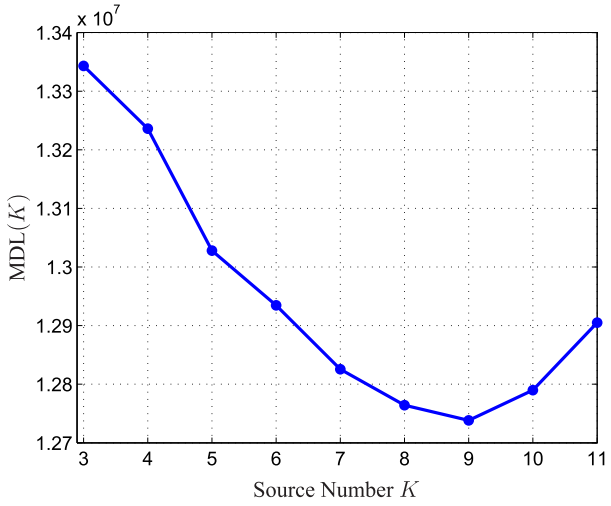


Fig. 6. MDL curve for real benchmark HRS data (collected from the Cuprite mining site, Nevada), where the detected number of sources is $\widehat{K} = 9$.

mRNA, astrocytic mRNA, oligodendrocytic mRNA, and microglial mRNA [40], and these reference mRNAs are biologically mixed to generate $M = 10$ composite samples, where there are $L = 31042$ probes in each sample. Here, the ith entry of $\mathbf{x}[n]$ (w.r.t. the ith entry of $\mathbf{a}_k$) denotes the measured expression level of probe n (w.r.t. the mixing abundance of the kth mRNA) for the ith mixed sample, and $s_k[n]$ stands for the expression level of the kth mRNA from probe n.

We processed the standardized molecular data set using our MDL algorithm, and the obtained description length is shown in Fig. 5, where $K = 4$ yields the shortest description length. Even in the presence of some impure samples [40] and the estimated nonuniform Dirichlet distribution ($\boldsymbol{\alpha} = [5.92,\ 6.00,\ 5.42,\ 7.21]^T$), our MDL algorithm still successfully determines the correct model-order for DEAP. However, Wax's algorithm again overestimates the number of sources as $\widehat{K} = 8$. Moreover, both VD and GENE yield meaningless model-order estimates of $\widehat{K} = 0$ for a wide range of P_{FA}.

G. Real Benchmark Hyperspectral Remote Sensing Data (RD4)

Our final experiment is conducted on a benchmark HRS data set, taken over the Cuprite mining site, Nevada, in 1997 [41]. Analyzing HRS data has been challenging due to the limited spatial resolution of the hyperspectral sensor (usually equipped on satellites or aircraft), under which arises the so-called hyperspectral unmixing (HU) problem, and correct model order selection plays an important role in yielding meaningful HU results [38]. The linear mixing model (1) can be used to characterize this data set as next discussed [38].

We use the same region of interest (ROI) from this mining site as used in [38], and for this ROI there are $K = 9$ minerals present—Muscovite, Alunite, Desert Varnish, Hematite, Montmorillonite, Kaolinite #1, Kaolinite #2, Buddingtonite, and Chalcedony. This ROI is composed of 150×150 pixels, but it contains ten outliers as reported in [38], so there are $L = 22490$ pixels in this data set. On the other hand, the hyperspectral sensor used to record this data set has 224 spectral bands, but with the bands 1–4, 107–114, 152–170, and 215–224 reported to be corrupted by water-vapor absorption, which hence were eliminated from the data set [38]; so, a total of $M = 183$ bands were used in this experiment. Then, the ith entry of $\mathbf{x}[n]$ denotes the measured solar electromagnetic radiation in pixel n from the ith spectral band, and $\mathbf{a}_k$ is the spectral signature of the kth mineral. Moreover, $s_k[n]$ represents the proportion of the kth mineral present in pixel n [38]; source full-additivity has been satisfied by nature—so there is no need to perform source standardization in this case.

We performed model order selection for this data set using our MDL algorithm, and the obtained MDL curve is shown in Fig. 6. One can observe that it successfully estimates the correct number of sources as the description length attains the minimum at $K = 9$, while the corresponding Dirichlet parameter vector is estimated as $\boldsymbol{\alpha} = [1.68,\ 2.44,\ 2.07,\ 2.65,\ 3.57,\ 2.68,\ 3.15,\ 3.71,\ 3.85]^T$. Wax's MDL algorithm again overestimates the number of sources present in this data set as $\widehat{K} = 18$. On the other hand, we note that geometry-based model order selection algorithms tend to underestimate K for this data set, due to the high similarity among some spectral signatures $\mathbf{a}_k$, making them not easily discernible. For instance, the signatures of Kaolinite #1 and Kaolinite #2 hold high resemblance [38] and the GENE-AH algorithm [18] underestimates the number of minerals to be $\widehat{K} = 7$, for all $P_{\text{FA}} \in \{10^{-4}, 10^{-5}, 10^{-6}\}$.

The above real experimental results were obtained by a computer equipped with Core-i7-4790K CPU with 4-GHz speed and 16-GB RAM. Let us conclude this section with a summary of the performances and computational efficiencies of the two MDL algorithms as shown in Table III. Though Wax's Gaussian modeling does allow very efficient MDL calculation, it seriously overestimates the number of sources. By contrast, though the proposed MDL requires greater computational expense due to the complicated Gaussian–Dirichlet convolution modeling, it correctly detects the number of sources.

TABLE III

ESTIMATED NUMBER OF SOURCES $\widehat{K}$ AND COMPUTATIONAL TIME [IN SECOND (sec.) OR MINUTE (min.)], FOR THE FOUR REAL-WORLD DATA SETS RD1 TO RD4 IN SECTIONS V-D TO V-G

Methods	Performance Measures			
	RD1	RD2	RD3	RD4
	$K=3$	$K=4$	$K=4$	$K=9$
MDL (Wax)	$\widehat{K}=27$ 0.19 (sec.)	$\widehat{K}=10$ 0.11 (sec.)	$\widehat{K}=8$ 0.10 (sec.)	$\widehat{K}=18$ 6.64 (sec.)
MDL	$\widehat{K}=3$ 14.3 (min.)	$\widehat{K}=4$ 10.4 (min.)	$\widehat{K}=4$ 8.31 (min.)	$\widehat{K}=9$ 19.8 (min.)

VI. DISCUSSION AND CONCLUSION

We have devised a model order selection algorithm for nBSS based on the MDL criterion, summarized in Algorithm 1. Some noteworthy characteristics and concluding remarks are as follows.

1) It employs the Gaussian–Dirichlet convolution model, much more consistent with the simplex structure of nBSS data than the Gaussian source model [26] considered in some prior works and hence should be more suitable for model order selection for a variety of nBSS domains.
2) It efficiently approximates ML parameter estimates of the Gaussian–Dirichlet density by linking the stochastic-oriented ML estimation problem to the simplex geometry-oriented Craig estimator, with the latter widely studied in the nBSS context.
3) We reformulated the high-dimensional integral (appearing in the calculation of the code length) into a form that can be efficiently approximated by Monte Carlo integration.
4) We performed substantial experimental comparisons, not only on simulated data sets but on multiple real-world data domains, and in comparison with some of the most well-known peer methods. Not only did we demonstrate superior performance of the proposed method, but we also demonstrated that the peer methods in general gave wildly inaccurate estimates of the number of sources present for the real-world data domains—thus it is not simply a matter of achieving "better" results than the peer methods (in fact, determining the true source number, which the proposed MDL algorithm accomplished, on all the tested real-world domains). None of the peer methods produced even remotely acceptable results on the real-world domains that we considered here.
5) For application domains where there is a natural physically motivated parameterization of the mixing matrix $\mathbf{A} \in \mathbb{R}^{M\times K}$ [28]–[30], an extension of the method proposed here could be developed employing such parameterization. However, we have demonstrated here that for many practical application domains, where $M, K \ll L$, and where there is no such obvious parameterization, full representation of the mixing matrix (and its estimation) does not prevent our method from achieving accurate model order estimates.
6) In the future work, we can investigate development of an approach that achieves locally optimal (rather than approximate) ML estimates of parameters and seek to identify scenarios where such optimal (but substantially more computationally intensive) parameter estimation is in fact needed to accurately estimate the number of sources. Another important yet unresolved line is to consider MDL-based model order selection in the underdetermined scenario (i.e., $M < K$), which appears to be quite challenging.

APPENDIX

A. Proof of Corollary 1

The covariance matrix of the noiseless counterpart of $\mathbf{x}[n]$, i.e., $\mathbf{x}_0[n] \triangleq \mathbf{A}\mathbf{s}[n]$, is given by

$$\begin{aligned}\Sigma_{\mathbf{x}_0[n]} &\triangleq \mathbb{E}[(\mathbf{x}_0[n]-\mathbb{E}[\mathbf{x}_0[n]])(\mathbf{x}_0[n]-\mathbb{E}[\mathbf{x}_0[n]])^T] \\ &= \mathbf{A}\ \Sigma_{\mathbf{s}[n]}\ \mathbf{A}^T \end{aligned} \tag{24}$$

where $\Sigma_{\mathbf{s}[n]} \triangleq \mathbb{E}[(\mathbf{s}[n]-\mathbb{E}[\mathbf{s}[n]])(\mathbf{s}[n]-\mathbb{E}[\mathbf{s}[n]])^T]$ is the covariance matrix of the random vector $\mathbf{s}[n]$. By (3), it can be verified that $\mathbb{E}[\mathbf{s}[n]] = (\boldsymbol{\alpha}/\alpha_0)$ [34], and hence we have

$$\begin{aligned}\Sigma_{\mathbf{s}[n]} &= \mathbb{E}\left[\left(\mathbf{s}-\frac{\boldsymbol{\alpha}}{\alpha_0}\right)\left(\mathbf{s}-\frac{\boldsymbol{\alpha}}{\alpha_0}\right)^T\right] \\ &= \frac{1}{\alpha_0^2(\alpha_0+1)}\cdot(\alpha_0\cdot\mathbf{DIAG}(\boldsymbol{\alpha})-\boldsymbol{\alpha}\boldsymbol{\alpha}^T)\ \ [34] \end{aligned} \tag{25}$$

where $\mathbf{DIAG}(\boldsymbol{\alpha})$ is the diagonal matrix whose kth diagonal element is α_k. One can observe, from (25), that $\Sigma_{\mathbf{s}[n]}$ is a positive semidefinite (PSD) matrix with exactly one zero eigenvalue (whose corresponding eigenvector is $\mathbf{1}_K$, i.e., $\Sigma_{\mathbf{s}[n]}\mathbf{1}_K = 0$), implying that $\mathrm{rank}(\Sigma_{\mathbf{s}[n]}) = K-1$, which, together with (A1), (24) and the Sylvester's rank inequality [44], yields $\mathrm{rank}(\Sigma_{\mathbf{x}_0[n]}) = K-1$. Finally, by further noting that $\Sigma_{\mathbf{x}[n]} = \Sigma_{\mathbf{x}_0[n]} + \sigma^2\mathbf{I}_M$ and that $\Sigma_{\mathbf{x}_0[n]}$ is a PSD matrix (i.e., the eigenvalues of $\Sigma_{\mathbf{x}_0[n]}$ are all non-negative), we observe that the smallest eigenvalue of $\Sigma_{\mathbf{x}[n]}$ is σ^2 with a multiplicity of $M-\mathrm{rank}(\Sigma_{\mathbf{x}_0[n]}) = M-(K-1)$. Then, the proof of Corollary 1 directly follows from (12). ■

B. Proof of Lemma 1

By (3) and (A1), one can observe that the p.d.f. of $\mathbf{x}_0[n] = \mathbf{A}\mathbf{s}[n]$ has a support of $\{\mathbf{A}\mathbf{s} \mid \mathbf{s} \in \mathrm{int}\mathcal{T}_e\}$, which is exactly the relative interior [36] of the convex hull $\mathrm{conv}\{\mathbf{a}_1,\ldots,\mathbf{a}_K\}$ (by the fact that $\mathcal{T}_e = \mathrm{conv}\{\mathbf{e}_1,\ldots,\mathbf{e}_K\}$ [see (2)], namely

$$\mathrm{dom}\ h = \mathrm{int}\ \mathrm{conv}\{\mathbf{a}_1,\ldots,\mathbf{a}_K\}.$$

Moreover, one can infer from the full column rank of $\mathbf{A}$ that the p.d.f. of $\mathbf{x}_0[n]$ is proportional to the p.d.f. of $\mathbf{s}[n] = \mathbf{A}^\dagger\mathbf{x}_0[n]$ [by (A1) and $\mathbf{x}_0[n] = \mathbf{A}\mathbf{s}[n]$], namely [by (3)]

$$h(\mathbf{y}) \propto \mathrm{Dir}(\mathbf{A}^\dagger\mathbf{y};\boldsymbol{\alpha}), \quad \forall \mathbf{y} \in \mathrm{dom}\ h$$

and hence (6) follows. As $\mathbf{w}[n]$ is assumed to be a zero-mean additive white Gaussian noise, its p.d.f. is given by (5). Finally, by [56, Th. 6.1.1] and (1), the p.d.f. $f(\mathbf{x}|\Theta^{(K)})$ of $\mathbf{x}[n]$ is given by the convolution of the density of $\mathbf{x}_0[n]$ and the density of $\mathbf{w}[n]$, and therefore the proof of Lemma 1 is completed. ■

C. Derivation of Approximate ML Estimation for $\mathbf{a}_1, \ldots, \mathbf{a}_K$

We begin by observing that as the function $\exp(-\|\mathbf{z}\|^2)$ is a symmetric unimodal function centered at the origin and decreases sharply as $\|\mathbf{z}\|$ increases, the exponential term in the integrand in (14) looks like an impulse function centered at $\mathbf{x}[n]$, thereby leading to the following approximation [see (3) and (6)]:

$$\int_{\mathbf{y}\in\text{dom } h} \text{Dir}(\mathbf{A}^\dagger\mathbf{y};\boldsymbol{\alpha})\cdot\exp\left\{\frac{-\|\mathbf{x}[n]-\mathbf{y}\|^2}{2\sigma^2}\right\}\,d\mathbf{y} \approx \int_{\mathbf{y}\in\text{dom } h} \text{Dir}(\mathbf{A}^\dagger\mathbf{y};\boldsymbol{\alpha})\cdot V_{\text{Dirac}}\cdot\text{Dirac}(\mathbf{y}-\mathbf{x}[n])\,d\mathbf{y} \quad (26)$$

where $\text{Dirac}(\cdot)$ is the Dirac delta function [w.r.t. $\text{aff}(\text{dom } h)$], and V_{Dirac} is the normalization constant for $\text{Dirac}(\cdot)$ defined as

$$V_{\text{Dirac}} \triangleq \int_{\mathbf{y}\in\text{aff}(\text{dom } h)} \exp\left\{\frac{-\|\mathbf{x}[n]-\mathbf{y}\|^2}{2\sigma^2}\right\}\,d\mathbf{y} = (2\pi\sigma^2)^{\frac{K-1}{2}}, \quad \text{if } \mathbf{x}[n]\in\text{dom } h \quad (27)$$

in which the last equality can be verified from the fact that the affine dimension of $\text{aff}(\text{dom } h) = \text{aff}\{\mathbf{a}_1,\ldots,\mathbf{a}_K\}$ is $K-1$ [see (7) and (A1)]. Furthermore, under assumptions (A1) and the source full-additivity $\mathbf{1}_K^T\mathbf{s}[n]=1$ [by (A2)], the affine hull of dom h can be approximately given by the $(K-1)$-dimensional affine hull that best fits the data cloud in the sense of least-squares fitting error [11], namely

$$\text{aff}(\text{dom } h) \approx \mathcal{A}(\mathbf{C},\mathbf{d}) \triangleq \{\mathbf{C}\tilde{\mathbf{x}}+\mathbf{d} \mid \tilde{\mathbf{x}}\in\mathbb{R}^{K-1}\} \quad (28)$$

where $\mathbf{C}$ is defined below (16), and $\mathbf{d}$ is defined in (11). By (7) and (28), the condition "$\mathbf{x}[n]\in\text{dom } h\subseteq\mathbb{R}^M$" can be explicitly and approximately written in the lower-dimension space $\mathbb{R}^{K-1}$ as

$$\text{"}\mathbf{x}[n]\in\text{dom } h\text{"} \;\approx\; \text{"}\tilde{\mathbf{x}}[n]\in\text{conv}\{\tilde{\mathbf{a}}_1,\ldots,\tilde{\mathbf{a}}_K\}\text{"} \quad (29)$$

where

$$\begin{cases}\tilde{\mathbf{x}}[n] \triangleq \mathbf{C}^\dagger(\mathbf{x}[n]-\mathbf{d})\in\mathbb{R}^{K-1}\\ \tilde{\mathbf{a}}_k \triangleq \mathbf{C}^\dagger(\mathbf{a}_k-\mathbf{d})\in\mathbb{R}^{K-1}.\end{cases} \quad (30)$$

Since the affine set fitting in (28) also serves as noise suppression [11], we have the approximation of $\mathbf{P}_{\mathcal{A}(\mathbf{C},\mathbf{d})}\mathbf{x}[n]\approx\mathbf{x}_0[n]$, where $\mathbf{P}_{\mathcal{A}}$ denotes the orthogonal projector on to the affine hull $\mathcal{A}$. Then, when $\mathbf{x}[n]\in\text{dom } h$, we see that $\mathbf{x}[n] = \mathbf{P}_{\text{aff}(\text{dom } h)}\mathbf{x}[n] \approx \mathbf{x}_0[n]$ [see the approximation in (28)], i.e., $\mathbf{A}^\dagger\mathbf{x}[n]\approx\mathbf{A}^\dagger\mathbf{x}_0[n]=\mathbf{s}[n]$, from which, together with (7) and (26)–(30) and the *sifting property* of the Dirac delta function, i.e., $\tilde{f}(\mathbf{c})=\int_{\mathbf{y}\in\text{aff}(\text{dom } h)}\text{Dirac}(\mathbf{c}-\mathbf{y})\cdot\tilde{f}(\mathbf{y})\,d\mathbf{y}$, we have that

$$\int_{\mathbf{y}\in\text{dom } h}\text{Dir}(\mathbf{A}^\dagger\mathbf{y};\boldsymbol{\alpha})\cdot\exp\left\{\frac{-\|\mathbf{x}[n]-\mathbf{y}\|^2}{2\sigma^2}\right\}\,d\mathbf{y} \approx \begin{cases}(2\pi\sigma^2)^{\frac{K-1}{2}}\cdot\text{Dir}(\mathbf{s}[n];\boldsymbol{\alpha}), & \text{if } \tilde{\mathbf{x}}[n]\in\text{conv}\{\tilde{\mathbf{a}}_1,\ldots,\tilde{\mathbf{a}}_K\}\\ 0, & \text{otherwise.}\end{cases} \quad (31)$$

From (31), one can see that if $\tilde{\mathbf{x}}[n]\notin\text{conv}\{\tilde{\mathbf{a}}_1,\ldots,\tilde{\mathbf{a}}_K\}$ for some n, the objective function of problem (15) approaches $-\infty$ [see (14)], preventing (15) from reaching its maximum, and hence we can reasonably impose the constraint $\tilde{\mathbf{x}}[n]\in\text{conv}\{\tilde{\mathbf{a}}_1,\ldots,\tilde{\mathbf{a}}_K\}$ to the unconstrained maximization problem (15). Then, by (14), (30), and (31), problem (15) can be approximated as

$$\begin{aligned}&\max_{\mathbf{a}_k}\; L\cdot\log(J(K,\mathbf{A}))+\sum_{n=1}^{L}\log\left\{(2\pi\sigma^2)^{\frac{K-1}{2}}\cdot\text{Dir}(\mathbf{s}[n];\boldsymbol{\alpha})\right\}\\ &\text{s.t. } \tilde{\mathbf{x}}[n]\in\text{conv}\{\tilde{\mathbf{a}}_1,\ldots,\tilde{\mathbf{a}}_K\},\quad \forall n\in\mathcal{I}_L,\ \text{(cf. (29))}\\ &\quad\tilde{\mathbf{a}}_k\triangleq\mathbf{C}^\dagger(\mathbf{a}_k-\mathbf{d}),\quad \text{for some } \mathbf{a}_k\in\mathcal{A}(\mathbf{C},\mathbf{d}),\ \forall k\in\mathcal{I}_K.\end{aligned} \quad (32)$$

Next, from (3), (6), and (8), one can verify that $J(K,\mathbf{A})$ can be expressed by the ratio of the volumes of two simplices, $\text{conv}\{\mathbf{a}_1,\ldots,\mathbf{a}_K\}$ and $\mathcal{T}\triangleq\text{conv}\{\mathbf{e}_1,\ldots,\mathbf{e}_{K-1},\mathbf{0}_K\}$ (note that the integral of the Dirichlet density over $\mathcal{T}$ equals to one [34]); to be precise, we have

$$J(K,\mathbf{A})=\frac{\text{vol}(\text{conv}\{\mathbf{e}_1,\ldots,\mathbf{e}_{K-1},\mathbf{0}_K\})}{\text{vol}(\text{conv}\{\mathbf{a}_1,\ldots,\mathbf{a}_K\})} \quad (33)$$

where

$$\text{vol}(\text{conv}\{\mathbf{b}_1,\ldots,\mathbf{b}_K\})\triangleq\frac{1}{(K-1)!}\sqrt{\det(\mathbf{B}^T\mathbf{B})} \quad (34)$$

in which $\mathbf{B}=[\,\mathbf{b}_1-\mathbf{b}_K,\mathbf{b}_2-\mathbf{b}_K,\ldots,\mathbf{b}_{K-1}-\mathbf{b}_K\,]\in\mathbb{R}^{M'\times(K-1)}$ (here, $M'\geq K-1$, and $\{\mathbf{b}_1,\ldots,\mathbf{b}_K\}$ is an affinely independent set). Then, by substituting (33) into (32), using the approximation of $\text{vol}(\text{conv}\{\mathbf{a}_1,\ldots,\mathbf{a}_K\})\approx\text{vol}(\text{conv}\{\tilde{\mathbf{a}}_1,\ldots,\tilde{\mathbf{a}}_K\})$ (by (28), (30), and the fact that $\mathbf{C}$ is semi-unitary), and dropping terms not dependent on $\mathbf{a}_k$ [note that the p.d.f. $\text{Dir}(\mathbf{s}[n];\boldsymbol{\alpha})\approx\text{Dir}(\mathbf{A}^\dagger\mathbf{x}[n];\boldsymbol{\alpha})$ is approximately inversely proportional to the volume of its support, i.e., $\text{vol}(\text{conv}\{\mathbf{a}_1,\ldots,\mathbf{a}_K\})$], the log-likelihood maximization problem (15) can then be approximately and explicitly expressed as the geometry-oriented simplex volume minimization problem (16). ■

D. Proof of Condition 1) in Theorem 1

By (1), (A2), the noiseless scenario and the observation that $\mathcal{T}_e=\text{conv}\{\mathbf{e}_1,\ldots,\mathbf{e}_K\}$, we can infer that $\mathbf{x}[n]\in\text{conv}\{\mathbf{a}_1,\ldots,\mathbf{a}_K\}$, that is

$$f(\mathbf{X}|\Theta^{(K)})=0,\quad \text{if } \mathbf{x}[n]\notin\text{conv}\{\mathbf{a}_1,\ldots,\mathbf{a}_K\}\text{ for some } n$$

indicating that the candidates of the ML estimates $\mathbf{a}_{k,\text{ML}}$, $\forall k\in\mathcal{I}_K$ [i.e., solutions to (15)], must form a simplex that encloses all the data points $\mathbf{x}[n]$, that is

$$\mathbf{x}[n]\in\text{conv}\{\mathbf{a}_{1,\text{ML}},\ldots,\mathbf{a}_{K,\text{ML}}\}. \quad (35)$$

On the other hand, under circumstance of (35), it can be inferred from (1), (6), and (33) that the likelihood function of $\mathbf{A}$, for a given noiseless observation matrix $\mathbf{X}$, is given by

$$\mathcal{L}(\mathbf{A}\mid\mathbf{X})=\prod_{n=1}^{L}J(K,\mathbf{A})\cdot\text{Dir}(\mathbf{A}^\dagger\mathbf{x}[n];\boldsymbol{\alpha}). \quad (36)$$

By [61, Lemma 1], we have, in the noiseless case, that

$$\mathcal{A}(\mathbf{C},\mathbf{d})=\text{aff}\{\mathbf{x}[1],\ldots,\mathbf{x}[L]\}=\text{aff}\{\mathbf{a}_1,\ldots,\mathbf{a}_K\} \quad (37)$$

which, together with (35), implies that the maximum of the likelihood function in (36) can occur only when $\mathbf{A}$ satisfies the constraints of problem (16). Hence, one can infer that $\mathbf{A}_{\mathrm{ML}}^{\dagger}\mathbf{x}[n]$ belongs to the domain of the Dirichlet distribution [see (A1), (3), and (2)]. Then, we have from the premise of $\boldsymbol{\alpha} = \mathbf{1}_K$ that $\mathrm{Dir}(\mathbf{A}^{\dagger}\mathbf{x}[n]; \boldsymbol{\alpha}) = (K-1)!$ [see (3)], which, together with (33) and (36), yields that

$$\mathcal{L}(\mathbf{A} \mid \mathbf{X}) \propto \left(\frac{1}{\mathrm{vol}(\mathrm{conv}\{\mathbf{a}_1, \ldots, \mathbf{a}_K\})} \right)^L. \tag{38}$$

From (30), (37), and the fact that $\mathbf{C}$ is semi-unitary, we see

$$\mathrm{vol}(\mathrm{conv}\{\mathbf{a}_1, \ldots, \mathbf{a}_K\}) = \mathrm{vol}(\mathrm{conv}\{\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_K\}). \tag{39}$$

Finally, combining (35), (37), (38), and (39), the problem of maximizing the likelihood of $\mathcal{L}(\mathbf{A}|\mathbf{X})$ [i.e., the ML problem (15)] is equivalent to the problem of minimizing the volume of the simplex $\mathrm{conv}\{\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_K\}$ with the constraint of (16) (i.e., the Craig simplex identification problem). Therefore, the proof of Theorem 1 is completed. ■

ACKNOWLEDGMENT

The authors would like to express our sincere gratitude to knowledgeable reviewers for providing many insightful comments that significantly improve the quality of this paper.

REFERENCES

[1] R. A. Moffitt *et al.*, "Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma," *Nature Genet.*, vol. 47, no. 10, pp. 1168–1178, 2015.

[2] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 12, pp. 4164–4169, 2004.

[3] N. Wang *et al.*, "Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues," *Sci. Rep.*, vol. 6, p. 18909, Jan. 2016.

[4] Y. Hart *et al.*, "Inferring biological tasks using Pareto analysis of high-dimensional data," *Nature Methods*, vol. 12, no. 3, pp. 233–235, 2015.

[5] M. D. Plumbley, "Algorithms for nonnegative independent component analysis," *IEEE Trans. Neural Netw.*, vol. 14, no. 3, pp. 534–543, May 2003.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[7] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

[8] Z. Yang, Y. Xiang, K. Xie, and Y. Lai, "Adaptive method for nonsmooth nonnegative matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 4, pp. 948–960, Apr. 2017.

[9] Z. Yang, G. Zhou, S. Xie, S. Ding, J.-M. Yang, and J. Zhang, "Blind spectral unmixing based on sparse nonnegative matrix factorization," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 1112–1125, Apr. 2011.

[10] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Dec. 2004.

[11] T. H. Chan, W. K. Ma, C. Y. Chi, and Y. Wang, "A convex analysis framework for blind separation of non-negative sources," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5120–5134, Oct. 2008.

[12] F.-Y. Wang, C.-Y. Chi, T.-H. Chan, and Y. Wang, "Nonnegative least-correlated component analysis for separation of dependent sources by volume maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 875–888, May 2010.

[13] Z. Yang, Y. Xiang, Y. Rong, and K. Xie, "A convex geometry-based blind source separation method for separating nonnegative sources," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1635–1644, Aug. 2014.

[14] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach*. New York, NY, USA: Springer-Verlag, 2002.

[15] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Online nonnegative matrix factorization with robust stochastic approximation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1087–1099, Jul. 2012.

[16] C.-I. Chang and Q. Du, "Estimation of number of spectrally distinct signal sources in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 3, pp. 608–619, Mar. 2004.

[17] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers, "How many principal components? Stopping rules for determining the number of non-trivial axes revisited," *Comput. Stat., Data Anal.*, vol. 49, no. 4, pp. 974–997, 2005.

[18] A. Ambikapathi, T.-H. Chan, C.-Y. Chi, and K. Keizer, "Hyperspectral data geometry-based estimation of number of endmembers using p-norm-based pure pixel identification algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2753–2769, May 2013.

[19] M. Mørup and L. K. Hansen, "Archetypal analysis for machine learning and data mining," *Neurocomputing*, vol. 80, pp. 54–63, Mar. 2012.

[20] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[21] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[22] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[23] A. Banerji, J. H. Naish, Y. Watson, G. C. Jayson, G. A. Buonaccorsi, and G. J. Parker, "DCE-MRI model selection for investigating disruption of microvascular function in livers with metastatic disease," *J. Magn. Reson. Imag.*, vol. 35, no. 1, pp. 196–203, 2012.

[24] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[25] P. D. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA, USA: MIT Press, 2007.

[26] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 387–392, Apr. 1985.

[27] Y.-O. Li, T. Adalı, and V. D. Calhoun, "Estimating the number of independent components for functional magnetic resonance imaging data," *Human Brain Mapping*, vol. 28, no. 11, pp. 1251–1266, 2007.

[28] L. Chen, P. L. Choyke, T.-H. Chan, C.-Y. Chi, G. Wang, and Y. Wang, "Tissue-specific compartmental analysis for dynamic contrast-enhanced MR imaging of complex tumors," *IEEE Trans. Med. Imag.*, vol. 30, no. 12, pp. 2044–2058, Dec. 2011.

[29] L. Chen *et al.*, "Unsupervised deconvolution of dynamic imaging reveals intratumor vascular heterogeneity and repopulation dynamics," *PLoS ONE*, vol. 9, no. 11, p. e112143, 2014.

[30] L. Chen *et al.*, "CAM-CM: A signal deconvolution tool for *in vivo* dynamic contrast-enhanced imaging of complex tissues," *Bioinformatics*, vol. 27, no. 18, pp. 2607–2609, 2011.

[31] N. Wang *et al.*, "The CAM software for nonnegative blind source separation in R-Java," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2899–2903, 2013.

[32] R. Schwartz and S. E. Shackney, "Applying unmixing to gene expression data for tumor phylogeny inference," *BMC Bioinform.*, vol. 11, no. 1, p. 42, 2010.

[33] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 44–57, Jan. 2002.

[34] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the Dirichlet distribution and related processes," Dept. Elect. Eng., Univ. Washington, Seattle, WA, USA, Tech. Rep. UWEETR-2010-0006, 2010. [Online]. Available: http://www.semanticsearchart.com/downloads/UWEETR-2010-0006.pdf

[35] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 542–552, May 1994.

[36] C.-Y. Chi, W.-C. Li, and C.-H. Lin, *Convex Optimization for Signal Processing and Communications: From Fundamentals to Applications*. Boca Raton, FL, USA: CRC Press, 2017.

[37] S. S. Shen-Orr *et al.*, "Cell type–specific gene expression differences in complex tissues," *Nature Methods*, vol. 7, no. 4, pp. 287–289, 2010.
[38] C.-H. Lin, C.-Y. Chi, Y.-H. Wang, and T.-H. Chan, "A fast hyperplane-based minimum-volume enclosing simplex algorithm for blind hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 64, no. 8, pp. 1946–1961, Apr. 2016.
[39] A. R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan, and H. F. Clark, "Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus," *PLoS ONE*, vol. 4, no. 7, p. e6098, 2009.
[40] A. Kuhn, D. Thu, H. J. Waldvogel, R. L. Faull, and R. Luthi-Carter, "Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain," *Nature Methods*, vol. 8, no. 11, pp. 945–947, 2011.
[41] *AVIRIS Free Standard Data Products*. Accessed: Aug. 1, 2014. [Online]. Available: http://aviris.jpl.nasa.gov/html/aviris.freedata.html
[42] N. Gillis and S. A. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 698–714, Apr. 2014.
[43] S. Arora *et al.*, "A practical algorithm for topic modeling with provable guarantees," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 280–288.
[44] S. Friedberg, A. Insel, and L. Spence, *Linear Algebra*, 4th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2003.
[45] R. L. Wheeden and A. Zygmund, *Measure and Integral: An Introduction to Real Analysis*. New York, NY, USA: Marcel Dekker Inc., 1977.
[46] N. Wang *et al.*, "UNDO: A bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples," *Bioinformatics*, vol. 31, no. 1, pp. 137–139, 2015.
[47] Y. Zhong and Z. Liu, "Gene expression deconvolution in linear space," *Nature Methods*, vol. 9, no. 1, pp. 8–9, 2012.
[48] T. W. Anderson, "Asymptotic theory for principal component analysis," *Ann. Math. Stat.*, vol. 34, no. 1, pp. 122–148, 1963.
[49] J. M. Bioucas-Dias, "A variable splitting augmented Lagrangian approach to linear spectral unmixing," in *Proc. IEEE WHISPERS*, Grenoble, France, Aug. 2009, pp. 1–4.
[50] C.-H. Lin, W.-K. Ma, W.-C. Li, C.-Y. Chi, and A. Ambikapathi, "Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5530–5546, Oct. 2015.
[51] G. Ronning, "Maximum likelihood estimation of Dirichlet distributions," *J. Stat. Comput. Simul.*, vol. 32, no. 4, pp. 215–221, 1989.
[52] D. Heinz and C.-I. Chang, "Fully constrained least squares linear mixture analysis for material quantification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 3, pp. 529–545, Mar. 2001.
[53] T. P. Minka. (2012). *Estimating a Dirichlet Distribution*. Massachusetts Institute of Technology. [Online]. Available: https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf
[54] A. Narayanan, "Algorithm AS 266: Maximum likelihood estimation of the parameters of the Dirichlet distribution," *Appl. Stat.*, vol. 40, no. 2, pp. 365–374, 1991.
[55] M. Newman and G. Barkema, *Monte Carlo Methods in Statistical Physics*. New York, NY, USA: Oxford Univ. Press, 1999.
[56] K. L. Chung, *A Course in Probability Theory*. San Diego, CA, USA: Academic, 2001.
[57] T. M. Apostol, *Mathematical Analysis*, 2nd ed. Reading, MA, USA: Addison Wesley, 1974.
[58] A. Ambikapathi, T.-H. Chan, C.-H. Lin, F.-S. Yang, C.-Y. Chi, and Y. Wang, "Convex optimization-based compartmental pharmacokinetic analysis for prostate tumor characterization using DCE-MRI," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 707–720, Apr. 2016.
[59] J. Chen, C. Richard, and P. Honeine, "Nonlinear estimation of material abundances in hyperspectral images with ℓ_1-norm spatial regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2654–2665, May 2014.
[60] H. Saddiki, J. McAuliffe, and P. Flaherty, "GLAD: A mixed-membership model for heterogeneous tumor subtype classification," *Bioinformatics*, vol. 31, no. 2, pp. 225–232, 2014.
[61] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4418–4432, Nov. 2009.

Chia-Hsiang Lin received the B.S. degree in electrical engineering and the Ph.D. degree in communications engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2010 and 2016, respectively. From 2015 to 2016, he was a Visiting Student with Virginia Tech, Blacksburg, VA, USA.

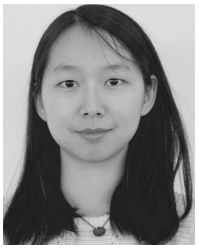
He is currently a Post-Doctoral Researcher with the Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. He was also a Post-Doctoral Researcher with NTHU from August 2016 to July 2017, and a Visiting Scholar with the Chinese University of Hong Kong, Hong Kong in 2014 and 2017. His research interests include network science, game theory, convex geometry and optimization, blind source separation, and imaging science.

Dr. Lin recieved the "Outstanding Doctoral Dissertation Award" from the Chinese Image Processing and Pattern Recognition Society in 2016 and the "Best Doctoral Dissertation Award" from the IEEE Geoscience and Remote Sensing Society in 2016.

Chong-Yung Chi (S'83–M'83–SM'89) received the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1983.

From 1983 to 1988, he was with the Jet Propulsion Laboratory, Pasadena, CA, USA. He has been a Professor with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan, since 1989, where he has also been a Professor with the Institute of Communications Engineering (ICE) since 1999 and the Chairman of ICE from 2002 to 2005. He has authored more than 220 technical papers, including more than 80 journal papers, more than 130 peer-reviewed conference papers, 4 book chapters, and 2 books, including a new textbook *Convex Optimization for Signal Processing and Communications from Fundamentals to Applications* (CRC Press, 2017) (popularly used in major universities in China). His current research interests include signal processing for wireless communications, convex analysis and optimization for blind source separation, and biomedical and hyperspectral image analysis.

Dr. Chi was an Associate Editor of IEEE TRANSACTIONS IN SIGNAL PROCESSING from 2001 to 2005 and 2012 to 2015, and IEEE SIGNAL PROCESSING LETTERS from 2006 to 2010, etc. He is a member of the Sensor Array and Multichannel Technical Committee since 2013, IEEE Signal Processing Society.



Lulu Chen received the B.S. degree in electronic information engineering and the M.S. degree in signal and information processing from the University of Science and Technology of China, Hefei, China, in 2010 and 2013, respectively. She is currently pursuing the Ph.D. degree in computer engineering with the Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

Her current research interests include latent variable model, blind source separation, and data visualization, with applications to computational bioinformatics.

David J. Miller (S'86–M'87–SM'07) received the B.S.E. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 1987, the M.S.E. degree in electrical engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 1990, and the Ph.D. degree in electrical engineering from the University of California, Santa Barbara, CA, USA, in 1995.

From 1988 to 1990, he was with General Atronics Corporation, Wyndmoor, PA, USA. He joined Pennsylvania State University, University Park, State College, PA, USA, in 1995, as an Assistant Professor and has been a Full Professor since 2007. His current research interests include statistical pattern recognition, machine learning, source coding, bioinformatics, and network security.

Dr. Miller was a member of the Machine Learning for Signal Processing Technical Committee within the IEEE Signal Processing Society from 2000 to 2010, the Chair of the committee from 2007 to 2009, and has been a member since 2015. He was the General Chair for the 2001 IEEE Workshop on Neural Networks for Signal Processing. He was a recipient of the National Science Foundation CAREER Award in 1996. From 2004 to 2007, he was an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Yue Wang (S'93–M'95–SM'03–F'16) received the B.S. and M.S. degrees in electrical and computer engineering from Shanghai Jiao Tong University, Shanghai, China, in 1984 and 1987, respectively, and the Ph.D. degree in electrical engineering from the University of Maryland Graduate School, Baltimore, MD, USA, in 1995.

In 1996, he was a Post-Doctoral Fellow with the Georgetown University School of Medicine, Washington, DC, USA. From 1996 to 2003, he was an Assistant and later an Associate Professor of electrical engineering with The Catholic University of America, Washington, DC, USA. In 2003, he joined Virginia Tech, Blacksburg, VA, USA, where he is currently the Grant A. Dove Professor of electrical and computer engineering and the Director of the Computational Bioinformatics and Bio-imaging Laboratory. His current research interests include machine learning, data science, and signal and image processing, with applications to bioinformatics, computational biology, and biomedical imaging.