Probabilistic Simplex Component Analysis by Importance Sampling

Nerya Granot, Tzvi Diskin, Nicolas Dobigeon, Ami Wiesel

Probabilistic Simplex Component Analysis by Importance Sampling

Nerya Granot, Tzvi Diskin, Nicolas Dobigeon and Ami Wiesel

***Abstract*—In this paper we consider the problem of linear unmixing hidden random variables defined over the simplex with additive Gaussian noise, also known as probabilistic simplex component analysis (PRISM). Previous solutions to tackle this challenging problem were based on geometrical approaches or computationally intensive variational methods. In contrast, we propose a conventional expectation maximization (EM) algorithm which embeds importance sampling. For this purpose, the proposal distribution is chosen as a simple surrogate distribution of the target posterior that is guaranteed to lie in the simplex. It is based on fitting the Dirichlet parameters to the linear minimum mean squared error (LMMSE) approximation, which is accurate at high signal-to-noise ratio. Numerical experiments in different settings demonstrate the advantages of this adaptive surrogate over state-of-the-art methods.**

***Index Terms*—Expectation maximization, importance sampling, simplex-structured matrix factorization**

I. Introduction

This letter considers the problem of linear unmixing hidden random variables lying on the simplex corrupted by an additive Gaussian noise. The problem, recently coined as Probabilistic Simplex Component Analysis (PRISM) [1], is a variant of Non-negative Matrix Factorization (NMF) [2, 3] that assumes an underlying Dirichlet prior distribution on the mixing coefficients. This leads to a well defined and identifiable parameter estimation problem under the maximum likelihood paradigm. The main challenge is then to design a numerical solution to the underlying optimization that involves high dimensional marginalization over the latent variables. In line with other approaches already proposed in the literature, we propose to solve this problem by resorting to a particular instance of the popular Expectation Maximization (EM) algorithm. In particular, the a E-step is approximated by a Monte Carlo integrator based on importance sampling [4, 5] with a carefully designed proposal distribution.

PRISM has a rich history in the signal processing, machine learning, remote sensing, statistics or chemometrics literatures where it is referred to under various names such as soft classification [6], unmixing [7], compositional data analysis [8] or multivariate curve resolution [9]. Some geometry-inspired approaches formulates this task as recovering the simplex with the minimal volume that covers all of the samples [10–12]. Others propose to identify the "purest" observations associated with the vertices of the simple [13, 14]. Methods have been derived for the noise-free case, for additive Gaussian noise and for more challenging scenarios involving outliers [15]. There is also a family of Bayesian solutions to this problem [16–18]. More advanced models also allow random mixing matrices to be characterized by different types of distributions [19, 20].

Closest to our letter is the recent PRISM paper which adopted a maximum likelihood formalism and derived its properties [1]. PRISM suggested two numerical solutions. The first ISA method based on importance sampling [21] was shown to be highly accurate but non-scalable. The second VIA method relied on variational inference using surrogate Dirichlet distributions, performed well in terms of accuracy and scalability, but was suboptimal at high signal-to-noise ratios (SNRs). These two methods motivate the present letter and are the building blocks to our proposed approach that unifies their ideas.

The main contribution of this letter is a normalized importance sampling approach to PRISM. First, we revisit ISA and show that using a simple surrogate based on the prior distribution, the resulting so-called SISA performs well even in large problems. Second, following VIA, we develop LISA, an adaptive importance sampling method. LISA uses Dirichlet surrogates based on the closed-form Linear Minimum Mean Squared Error (LMMSE) estimates. In a low SNR regime, LISA is shown to behave as SISA which is near optimal. At high SNR, LISA mimics the LMMSE estimate and provides its samples around the estimate. Both SISA and LISA embed sampling schemes and are therefore computationally intensive. However, contrary to previous methods, their samples are guaranteed to lie within the simplex and thus are never rejected and ensure scalability. Numerical experiments using synthetic simulations demonstrate the advantages of the proposed methods. Results show that SISA can serve as a promising initialization to VIA and that LISA provides the best performance (especially in high SNR where VIA is theoretically suboptimal).

II. Dirichlet preliminaries

This section provides basic properties related to the simplex and the Dirichlet distribution that will be used throughout this paper. The $(k-1)$-dimensional simplex is defined as

$$\mathbb{S}_{k-1} = \left\{ \mathbf{z} \in R^k : z_i \geq 0,\ \mathbf{1}^T \mathbf{z} = 1 \right\}, \tag{1}$$

where $\mathbf{1}$ is a length-k vector of ones. A popular multivariate distribution over this simplex is the Dirichlet distribution whose probability density function (pdf) writes $\mathrm{Dir}(\mathbf{z}; \boldsymbol{\alpha}) \propto \prod_{n=1}^{k} z_n^{\alpha_n - 1}$ $(\mathbf{z} \in \mathbb{S}_{k-1})$ where $\boldsymbol{\alpha} > 0$ is the concentration

The first two authors equally contributed to this letter. The research was partially supported by ISF grant number 2672/21, the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI) under grant agreement ANITI ANR-19-PI3A-0004 and the ANR Active Molecular Imaging and Unmixing (ANR IMAGIN) Project under Grant ANR-21-CE29-0007.

parameter (the inequalities are element-wise). The vector $\boldsymbol{\alpha}$ controls how "probability mass" is allocated over the components by adjusting the mean and covariance given by

$$\mathbf{m} = \frac{\boldsymbol{\alpha}}{\mathbf{1}^T\boldsymbol{\alpha}} \in \mathbb{S}_{k-1} \text{ and } \mathbf{C} = \frac{\mathrm{diag}(\mathbf{m}) - \mathbf{m}\mathbf{m}^T}{\mathbf{1}^T\boldsymbol{\alpha} + 1} \quad (2)$$

and satisfying

$$\mathbf{m} = \mathbf{P}\mathbf{m} - \frac{1}{k}\mathbf{1}, \quad \mathbf{C} = \mathbf{P}\mathbf{C}, \quad \mathbf{P} = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T\mathbf{1}}. \quad (3)$$

Because of the linear dependence between the vector components, the covariance $\mathbf{C}$ is singular.

III. Problem formulation

We consider linear mixing with random hidden variables, also known as PRISM [1]

$$\mathbf{y}_i = \mathbf{H}\mathbf{z}_i + \mathbf{w}_i \quad i = 1, \cdots, N, \quad (4)$$

where $\mathbf{H}$ is a deterministic unknown matrix of size $d \times k$, $\mathbf{z}_i \sim p(\mathbf{z}) = \mathrm{Dir}(\mathbf{z}; \boldsymbol{\alpha})$ are independent and identically distributed (i.i.d.) hidden random vectors from a Dirichlet distribution with a known[1] deterministic parameter $\boldsymbol{\alpha}$, and $\mathbf{w}_i$ are i.i.d. noise vectors $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with a known variance σ^2. We assume that $\mathbf{z}_i$ and $\mathbf{w}_i$ are independent. The goal is then to estimate $\mathbf{H}$ given an observed set of measurements $\{\mathbf{y}_i\}_{i=1}^N$.

MCEM algorithm – A standard approach consists in maximizing the log-likelihood with respect to (w.r.t.) the unknown parameter [22]

$$\hat{\mathbf{H}} = \arg\max_{\mathbf{H}} \frac{1}{N}\sum_{i=1}^{N} \log p_{\mathbf{H}}(\mathbf{y}_i). \quad (5)$$

The distribution of $\mathbf{y}$ is defined through the hidden variable $\mathbf{z}$ and requires marginalization

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \quad (6)$$

Computing this high dimensional integral or its gradient is often impossible. A popular alternative is the EM algorithm that iteratively maximizes a lower bound [23]. The EM algorithm can be cast as a minimization-majorization strategy [24] and, under regular technical conditions, it is shown to converge to a stationary point of the log-likelihood. Each iteration of the overall algorithm consists of two steps. Given a current estimate $\mathbf{H}'$ of the parameter, the first E-step computes a conditional expectation of the complete log-likelihood

$$\text{(E)} \quad Q(\mathbf{H}; \mathbf{H}') = \sum_{i=1}^{N} \mathrm{E}[\log p(\mathbf{y}_i|\mathbf{z}_i) + \log p(\mathbf{z}_i)|\mathbf{y}_i; \mathbf{H}'], \quad (7)$$

where $\mathrm{E}[\cdot|\mathbf{y}; \mathbf{H}']$ denotes the conditional expectation given $\mathbf{y}$ and $\mathbf{H}'$. In the context of PRISM, the quantity (7) can be explicitly derived as

$$Q(\mathbf{H}; \mathbf{H}') = \sum_{i=1}^{N} \frac{\mathrm{E}[\|\mathbf{H}\mathbf{z}_\mathrm{i} - \mathbf{y}_\mathrm{i}\|^2|\mathbf{y}_\mathrm{i}; \mathbf{H}']}{-2\sigma^2}$$
$$= \sum_{i=1}^{N} \frac{\mathrm{Tr}(\mathbf{H}^T\mathbf{H}\mathrm{E}[\mathbf{z}_\mathrm{i}\mathbf{z}_\mathrm{i}^\mathrm{T}|\mathbf{y}_\mathrm{i}; \mathbf{H}']) - 2\mathbf{y}_\mathrm{i}^\mathrm{T}\mathbf{H}\mathrm{E}[\mathbf{z}_\mathrm{i}|\mathbf{y}_\mathrm{i}; \mathbf{H}']}{-2\sigma^2}. \quad (8)$$

[1]We assume that $\boldsymbol{\alpha}$ is known, but EM can be used to estimate it too.

The second M-step searches for the parameter that maximizes this quantity

$$\text{(M)} \quad \mathbf{H} \leftarrow \max_{\mathbf{H}} \mathrm{Q}(\mathbf{H}; \mathbf{H}'). \quad (9)$$

Combining the (E) and (M) steps, the EM iteration boils down to the updating rule

$$\mathbf{H} \leftarrow \sum_{i=1}^{N} \mathbf{y}_i\mathrm{E}^T[\mathbf{z}_i|\mathbf{y}_i; \mathbf{H}'] \left(\sum_{i=1}^{N} \mathrm{E}[\mathbf{z}_i\mathbf{z}_i^T|\mathbf{y}_i; \mathbf{H}']\right)^{-1}. \quad (10)$$

The main challenge with this strategy lies in the computation of $\mathrm{E}[\mathbf{z}_\mathrm{i}|\mathbf{y}_\mathrm{i}; \mathbf{H}']$ and $\mathrm{E}[\mathbf{z}_\mathrm{i}\mathbf{z}_\mathrm{i}^\mathrm{T}|\mathbf{y}_\mathrm{i}; \mathbf{H}']$ efficiently for each sample at each iteration. One solution consists in resorting to a Monte Carlo (MC) integration, resulting in a overall algorithm which is an archetypal instance of Monte Carlo EM [21]. The MC approximation adopted in this work is detailed in what follows.

Importance sampling – Performing the EM updates (10) requires to compute expectations of the form

$$\mathrm{E}[d(\mathbf{z}, \mathbf{y})|\mathbf{y}, \mathbf{H}'] = \int d(\mathbf{z}, \mathbf{y})p(\mathbf{z}|\mathbf{y}, \mathbf{H}')d\mathbf{z} \quad (11)$$
$$= \frac{1}{C(\mathbf{y}, \mathbf{H}')} \int d(\mathbf{z}, \mathbf{y})p(\mathbf{y}|\mathbf{z}, \mathbf{H}')p(\mathbf{z})d\mathbf{z},$$

where $C(\mathbf{y}, \mathbf{H}') = \int p(\mathbf{y}|\mathbf{z}, \mathbf{H}')p(\mathbf{z})d\mathbf{z}$ and $d(\cdot, \mathbf{y})$ specifies the quantity of interest. To approximate such expectations, a conventional MC integration technique is referred to as importance sampling (IS) and relies on a surrogate (or proposal) distribution $q(\cdot)$ it is easier to sample from. Under generic assumptions about the proposal $q(\cdot)$, the expectation in (11) can be rewritten as

$$\mathrm{E}[d(\mathbf{z}, \mathbf{y})|\mathbf{y}, \mathbf{H}'] = \frac{1}{C} \int d(\mathbf{z}, \mathbf{y})p(\mathbf{y}|\mathbf{z}, \mathbf{H}')p(\mathbf{z}) \frac{q(\mathbf{z}|\mathbf{y}, \mathbf{H}')}{q(\mathbf{z}|\mathbf{y}, \mathbf{H}')} d\mathbf{z}.$$

Then, for a given set of M i.i.d. samples $\mathbf{z}_m$ drawn from $q(\mathbf{z}|\mathbf{y}, \mathbf{H}')$, IS proceeds with the Monte Carlo approximation

$$\mathrm{E}[d(\mathbf{z}, \mathbf{y})|\mathbf{y}, \mathbf{H}'] \approx \frac{1}{\tilde{C}} \sum_{m=1}^{M} \tilde{w}_m d(\mathbf{z}_m, \mathbf{y}) \quad (12)$$

with $\tilde{w}_m = \frac{p(\mathbf{y}|\mathbf{z}_m, \mathbf{H}')p(\mathbf{z}_m)}{q(\mathbf{z}_m|\mathbf{y}, \mathbf{H}')}$ and $\tilde{C} = \sum_{m=1}^{M} \tilde{w}_m$. The quality of the approximation (12) is governed by the similarity between the target distribution and its surrogate. The goal is therefore to choose a surrogate distribution $q(\mathbf{z}|\mathbf{y}, \mathbf{H}')$ which is a good approximation to $p(\mathbf{z}|\mathbf{y}, \mathbf{H}')$ and easy to sample from. This point is the core of the next section.

IV. Surrogate posterior distributions

This section discusses several choices of cheap surrogates $q(\mathbf{z}|\mathbf{y})$ approximating $p(\mathbf{z}|\mathbf{y})$ for $\mathbf{z} \in \mathbb{S}_{k-1}$.

A. Dirichlet prior

The simplest surrogate distribution, denoted by Simple ISA (SISA), ignores $\mathbf{y}$ and approximates the posterior by the prior

$$q(\mathbf{z}|\mathbf{y}) = p(\mathbf{z}) = \mathrm{Dir}(\mathbf{z}; \boldsymbol{\alpha}). \quad (13)$$

This approach should be optimal for low SNR where $p(\mathbf{z}|\mathbf{y}) \approx p(\mathbf{z})$. Otherwise, it seems wasteful as it ignores the information brought by $\mathbf{y}$. SISA can also be derived as a Sample Average Approximation or naive Monte Carlo averaging [15].

B. Gaussian posterior

The target posterior distribution $p(\mathbf{z}|\mathbf{y})$ is a multivariate Gaussian distribution truncated on the simplex. One solution to generate samples from this distribution consists in resorting to rejection sampling. Such a strategy is shown to performed poorly for large values of k due to a high rejection rate. In particular, the authors of [1] stated, in settings identical to those in our experiments, this method generated almost no samples. One alternative would rely on more advanced Monte Carlo techniques, e.g., Markov Chain Monte Carlo (MCMC) algorithms [25]. However, such strategies are generally computationally demanding and can be hardly embedded into the iterative scheme of EM.

One alternative is the conditional Gaussian distribution [26, Sec 10.6], denoted by $\mathcal{N}(\mathbf{z}; \overline{\mathbf{m}}(\mathbf{y}), \overline{\mathbf{C}})$ with

$$\begin{aligned}\overline{\mathbf{m}}(\mathbf{y}) &= \mathbf{m} + \mathbf{C}\mathbf{H}^T(\mathbf{H}\mathbf{C}\mathbf{H}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{H}\mathbf{m}) \\ \overline{\mathbf{C}} &= \mathbf{C} - \mathbf{C}\mathbf{H}^T(\mathbf{H}\mathbf{C}\mathbf{H}^T + \sigma^2\mathbf{I})^{-1}\mathbf{H}\mathbf{C},\end{aligned} \tag{14}$$

where $\mathbf{m}$ and $\mathbf{C}$ are the prior Dirichlet moments. This approximation is near optimal in high SNR regimes where the $\overline{\mathbf{m}}(\mathbf{y}) \approx \mathbf{z}$ is accurate and it makes sense to sample around it. Unfortunately, with even small noise, samples from this distribution do not necessarily lie in the simplex and this approach leads to a high rejection rate. Recent contributions on importance sampling from a truncated Gaussian were provided in [27, 28].

Interestingly, $\overline{\mathbf{m}}(\mathbf{y})$ can also be derived as the linear minimum mean squared error estimator (LMMSE) that holds for any distribution. The matrix $\overline{\mathbf{C}}$ is its corresponding mean squared error (MSE) [26, Sec. 12.5].

C. Dirichlet posterior

A more promising approximation, denoted by LISA, relies on the Dirichlet distribution but adjusts it according to the LMMSE detailed above. We define

$$\overline{\mathbf{z}} \sim q(\overline{\mathbf{z}}|\mathbf{y}) = \mathrm{Dir}(\overline{\mathbf{z}}; \overline{\boldsymbol{\alpha}}(\mathbf{y})), \tag{15}$$

which is guaranteed to lie in the simplex and choose $\overline{\boldsymbol{\alpha}}(\mathbf{y})$ to fit the moments in (14), i.e.,

$$\mathrm{E}[\overline{\mathbf{z}}] = \widetilde{\mathbf{m}}(\mathbf{y}) \tag{16}$$

$$\mathrm{Tr}[\mathrm{cov}[\overline{\mathbf{z}}]] = \mathrm{Tr}[\overline{\mathbf{C}}], \tag{17}$$

where $\widetilde{\mathbf{m}}(\mathbf{y})$ is the (approximate) projection[2] of $\overline{\mathbf{m}}(\mathbf{y})$ onto $\mathbb{S}_{k-1}$. Indeed, the Dirichlet mean is always within the simplex and since it has $k-1$ degrees of freedom, imposing (16) and (17) boils down to fit k parameters to k moments constraints . The first moment constraint

$$\mathrm{E}[\overline{\mathbf{z}}] = \frac{\overline{\boldsymbol{\alpha}}}{\mathbf{1}^T\overline{\boldsymbol{\alpha}}} = \widetilde{\mathbf{m}}(\mathbf{y}), \tag{18}$$

leads to $\overline{\boldsymbol{\alpha}} = \mu\widetilde{\mathbf{m}}(\mathbf{y})$ with some $\mu > 0$. The scaling factor μ controls the variance. It is adjusted to enforce the covariance

$$\mathrm{Tr}[\mathrm{cov}[\overline{\mathbf{z}}]] = \mathrm{Tr}\left[\frac{\mathrm{diag}(\widetilde{\mathbf{m}}(\mathbf{y})) - \widetilde{\mathbf{m}}(\mathbf{y})\widetilde{\mathbf{m}}^T(\mathbf{y})}{\mu + 1}\right] = \mathrm{Tr}(\overline{\mathbf{C}}),$$

which yields

$$\mu = \frac{1 - \|\overline{\mathbf{m}}(\mathbf{y})\|^2}{\mathrm{Tr}(\overline{\mathbf{C}})} - 1. \tag{19}$$

To summarize, the LISA proposal distribution is defined as

$$q(\overline{\mathbf{z}}|\mathbf{y}) = \mathrm{Dir}\left(\overline{\mathbf{z}}; \left(\frac{1 - \|\widetilde{\mathbf{m}}(\mathbf{y})\|^2}{\mathrm{Tr}(\overline{\mathbf{C}})} - 1\right)\widetilde{\mathbf{m}}(\mathbf{y})\right). \tag{20}$$

Capitalizing on the properties stated in Section II, one can easily characterize the asymptotic behavior of this proposal wrt to the noise level. In low SNR, LISA depends only on the prior and we get $\widetilde{\mathbf{m}}(\mathbf{y}) \approx \mathbf{m}$ and $\overline{\mathbf{C}} \approx \mathbf{C}$. After some algebraic manipulations this yields

$$q(\overline{\mathbf{z}}|\mathbf{y}) \overset{\text{low SNR}}{\to} \mathrm{Dir}\left(\overline{\mathbf{z}}; \boldsymbol{\alpha}\right), \tag{21}$$

so that LISA converges to SISA in low SNR. Conversely, in high SNR, LISA does not depend on the prior $\boldsymbol{\alpha}$. The moments reduce to (see the Appendix for proof):

$$\begin{aligned}\widetilde{\mathbf{m}}(\mathbf{y}) &\overset{\text{high SNR}}{\to} (\mathbf{H}\mathbf{P})^\dagger\mathbf{y} + \mathbf{v}_\mathbf{H} \\ \overline{\mathbf{C}} &\overset{\text{high SNR}}{\to} \sigma^2(\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{P})^\dagger,\end{aligned} \tag{22}$$

where $\mathbf{v}_\mathbf{H} = \frac{1}{k}\left(\mathbf{I} - (\mathbf{H}\mathbf{P})^\dagger\mathbf{H}\right)\mathbf{1}$ and we get

$$q(\overline{\mathbf{z}}|\mathbf{y}) \overset{\text{high SNR}}{\to} \mathrm{Dir}\left(\overline{\mathbf{z}}; \frac{c}{\sigma^2}[(\mathbf{H}\mathbf{P})^\dagger\mathbf{y} + \mathbf{v}_\mathbf{H}]\right), \tag{23}$$

where $c > 0$ is a constant. As expected, this yield samples which are concentrated around the LMMSE estimate with a small variance.

V. Numerical experiments

This section compares the performance of the different algorithms using numerical experiments. The simulations are reproduction of the synthetic experiments in [1] with the exact settings. The data were generated synthetically based on the linear unmixing model in (4) with $\boldsymbol{\alpha} = \mathbf{1}$. The matrix $\mathbf{H}$ of dimensions $d = 50$ and $k = 20$ was generated once per experiment with i.i.d. elements uniformly distributed in $[0, 1]$. Performance was measured by mean squared error over the best permutation:

$$\mathrm{MSE} = \min_{\pi\in\Pi}\sum_i \|\mathbf{H}_i - \hat{\mathbf{H}}_{\pi_i}\|^2, \tag{24}$$

where $\mathbf{H}_i$ is the $i$the column of $\mathbf{H}$ and Π is the set of all indices permutations. Four competing algorithms are compared

- VCA: A simple and fast baseline [29].
- SISA: An EM method initialized by VCA and using standard importance sampling as detailed in Sec. IV-A. The EM has 100 iterations and SISA is based on $M = 500$ samples.
- LISA: A similar EM method where the last 50 iterations use an LMMSE surrogate as detailed in Sec. IV-C (The first 50 are identical to SISA).
- VIA: A variational approach due to [1]. Following [30], we implemented VIA using Torch with a line search for the learning rate. In order to achieve good accuracy, we initialized VIA with SISA.

[2]The approximation was done by zeroing negative terms and dividing by the sum.

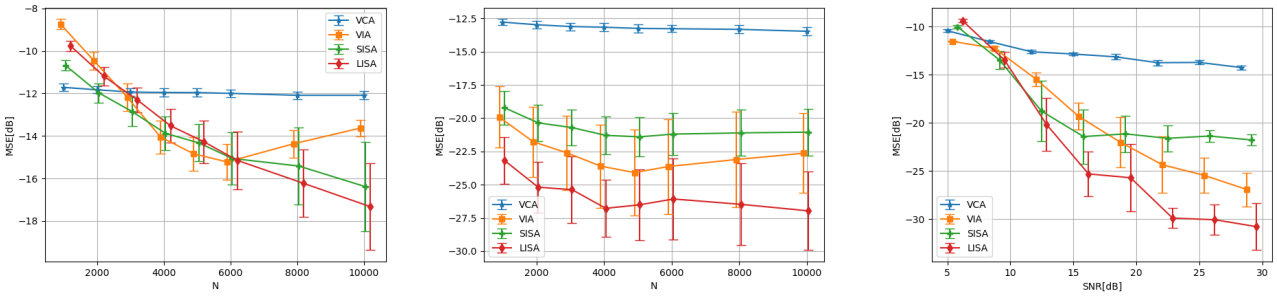

Fig. 1. MSE as a function of the number of samples in SNR $= 10$dB **(left)**, the number of samples in SNR $= 20$dB **(middle)** and the SNR with $N = 5000$ **(right)**. The SNR is given by $\mathrm{Tr}\left(\mathbf{HCH}^T\right)/\sigma^2$.

The first experiment considered performance as a function of the number of samples in low SNR. The results are provided in Fig. 1. As expected from the theory, SISA was near optimal in low SNR. LISA behaved similarly and outperformed it when the number of samples is large.

The second experiment repeated the experiment in higher SNR. The results are provided in Fig. 1. This setting is more challenging for SISA which is outperformed by VIA. LISA was significantly better than the rest of the algorithms throughout this graph.

The third experiment in Fig. 1 examined the performance for a fixed number of samples 5000 as a function of the SNR. Here too the advantages of LISA are apparent. It is only to see the expected degradation in performance of VIA in high SNR. In terms of computational complexity, VCA is the fastest algorithm. SISA and LISA are significantly higher because of the sampling. LISA is slightly more expensive than SISA because of its data dependent concentration parameters. Finally, VIA is more tricky. The original implementation in [1] is quite complicated, but the Torch implementation of [30] is very fast. However, in order to get the performance detailed above, we had to initialize VIA with SISA and this slowed it down considerably. Future work can consider the use of adaptive importance sampling [31] for smoother transition from SISA to LISA.

Appendix
Derivation of LISA in high SNR (22)

First, we show that if $\mathbf{H}$ is full rank and $d \geq k$, then

$$\mathbf{M} = \lim_{\sigma^2 \to 0} \mathbf{CH}^T \left(\mathbf{HCH}^T + \sigma^2 \mathbf{I}\right)^{-1} = (\mathbf{HP})^\dagger . \tag{25}$$

$\mathbf{C}$ is positive semi-definite and there exists a matrix $\mathbf{B}$ such that $\mathbf{BB} = \mathbf{C}$:

$$\mathbf{M} = \lim_{\sigma^2 \to 0} \mathbf{BBH}^T \left(\mathbf{HBBH}^T + \sigma^2 \mathbf{I}\right)^{-1} = \mathbf{B}\left(\mathbf{HB}\right)^\dagger . \tag{26}$$

The null space of $\mathbf{B}$ is the same as $\mathbf{C}$ and $\mathbf{P}$, and therefore $\mathbf{B}^\dagger\mathbf{B} = \mathbf{P}$. It remains to prove that $\mathbf{M}'^\dagger = \mathbf{HBB}^\dagger$ is the pseudo-inverse of $\mathbf{M} = \mathbf{B}\left(\mathbf{HB}\right)^\dagger$ using the four Moore-Penrose conditions [32]:

(I) Because $\mathbf{H}$ is full rank and $n \geq k$, $\mathbf{H}^\dagger\mathbf{H} = \mathbf{I}_{k\times k}$ and:

$$\mathbf{MM}'^\dagger = \mathbf{B}\left(\mathbf{HB}\right)^\dagger \mathbf{HBB}^\dagger = \mathbf{H}^\dagger\mathbf{HB}\left(\mathbf{HB}\right)^\dagger \mathbf{HBB}^\dagger = \mathbf{BB}^\dagger, \tag{27}$$

which is clearly a symmetric orthogonal projection. (II) Similarly,

$$\mathbf{M}'^\dagger\mathbf{M} = \mathbf{HBB}^\dagger\mathbf{B}\left(\mathbf{HB}\right)^\dagger = \mathbf{HB}\left(\mathbf{HB}\right)^\dagger , \tag{28}$$

which is again symmetric. (III) Using (27),

$$\mathbf{MM}'^\dagger\mathbf{M} = \mathbf{BB}^\dagger\mathbf{B}\left(\mathbf{HB}\right)^\dagger = \mathbf{B}\left(\mathbf{HB}\right)^\dagger = \mathbf{M}. \tag{29}$$

(IV) Finally, using (28), we have

$$\mathbf{M}'^\dagger\mathbf{MM}'^\dagger = \mathbf{HB}\left(\mathbf{HB}\right)^\dagger \mathbf{HBB}^\dagger = \mathbf{HBB}^\dagger = \mathbf{M}'^\dagger. \tag{30}$$

Next, we show that the mean terms yield $\mathbf{v_H}$ which is independent of the prior $\boldsymbol{\alpha}$. Plugging (25) into (14) gives:

$$\overline{\mathbf{m}}(\mathbf{y}) = (\mathbf{HP})^\dagger \mathbf{y} + \mathbf{m} - (\mathbf{HP})^\dagger \mathbf{Hm}. \tag{31}$$

Now using the fact that $\mathbf{m}$ satisfies (3), we get:

$$(\mathbf{HP})^\dagger \mathbf{Hm} = (\mathbf{HP})^\dagger \mathbf{HPm} + \frac{1}{k}(\mathbf{HP})^\dagger \mathbf{H1}. \tag{32}$$

We note that because $\mathbf{H}$ is full rank, the null space of $\mathbf{HP}$ is the same as of $\mathbf{P}$ and thus $(\mathbf{HP})^\dagger \mathbf{HP} = \mathbf{P}$. Therefore, $(\mathbf{HP})^\dagger \mathbf{HPm} = \mathbf{m} - \frac{1}{k}$ Plugging it into (32) and then to (31) gives:

$$\overline{\mathbf{m}}(\mathbf{y}) = (\mathbf{HP})^\dagger \mathbf{y} + \frac{1}{k}\left(\mathbf{I} - (\mathbf{HP})^\dagger \mathbf{H}\right)\mathbf{1}. \tag{33}$$

Finally, the covariance of the error is given by:

$$\begin{aligned}\overline{\mathbf{C}} &= \mathrm{cov}(\hat{\mathbf{z}} - \mathbf{z}) = \mathrm{cov}\left((\mathbf{HP})^\dagger (\mathbf{y} - \mathbf{H}v_{\mathbf{H}})) - \mathbf{z}\right) \\ &= \left((\mathbf{HP})^\dagger \mathbf{H} - \mathbf{I}\right)\mathbf{C}\left((\mathbf{HP})^\dagger \mathbf{H} - \mathbf{I}\right)^T \\ &\quad + \sigma^2 (\mathbf{HP})^\dagger (\mathbf{HP})^{\dagger T}.\end{aligned} \tag{34}$$

Now we note that $\left((\mathbf{HP})^\dagger \mathbf{H} - \mathbf{I}\right)\mathbf{C} = \mathbf{0}$ and thus:

$$\overline{\mathbf{C}} = \sigma^2 (\mathbf{HP})^\dagger (\mathbf{HP})^{\dagger T} = \sigma^2 \left(\mathbf{PH}^T\mathbf{HP}\right)^\dagger . \tag{35}$$

References

[1] R. Wu, W.-K. Ma, Y. Li, A. M.-C. So, and N. D. Sidiropoulos, "Probabilistic simplex component analysis," *IEEE Trans. Signal Processing*, vol. 70, pp. 582–599, 2022.

[2] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of

error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[3] 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, 1999.

[4] A. B. Owen, *Monte Carlo theory, methods and examples*, 2013.

[5] V. Elvira and L. Martino, "Advances in importance sampling," *arXiv preprint arXiv:2102.05407*, 2021.

[6] G. Wahba, "Soft and hard classification by reproducing kernel Hilbert space methods," *Proc. Nat. Academy Sci.*, vol. 99, no. 26, pp. 16 524–16 530, 2002.

[7] W.-K. Ma, J. M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C.-Y. Chi, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 67–81, 2014.

[8] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[9] A. de Juan and R. Tauler, "Multivariate curve resolution: 50 years addressing the mixture analysis problem – a review," *Analytica Chimica Acta*, vol. 1145, pp. 59–78, 2021.

[10] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geoscience and Remote Sensing*, vol. 32, no. 3, pp. 542–552, 1994.

[11] J. Li and J. M. Bioucas-Dias, "Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data," in *Proc. Int. Geosci. Remote Sensing Symposium (IGARSS)*, vol. 3. IEEE, 2008, pp. III–250.

[12] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "Convex analysis based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 1089–1092.

[13] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications." *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, 2019.

[14] M. E. Winter, "N-findr: An algorithm for fast autonomous spectral end-member determination in hyperspectral data," in *Imaging Spectrometry V*, vol. 3753. SPIE, 1999, pp. 266–275.

[15] R. Wu, W.-K. Ma, and X. Fu, "A stochastic maximum-likelihood framework for simplex structured matrix factorization," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2557–2561.

[16] J. M. Nascimento and J. M. Bioucas-Dias, "Hyperspectral unmixing based on mixtures of dirichlet components," *IEEE Trans. Geoscience and Remote Sensing*, vol. 50, no. 3, pp. 863–878, 2011.

[17] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero, "Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Trans. Signal Processing*, vol. 57, no. 11, pp. 4355–4368, 2009.

[18] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[19] O. Eches, N. Dobigeon, C. Mailhes, and J.-Y. Tourneret, "Bayesian estimation of linear mixtures using the normal compositional model. application to hyperspectral imagery," *IEEE Trans. Image Processing*, vol. 19, no. 6, pp. 1403–1413, 2010.

[20] Y. Woodbridge, U. Okun, G. Elidan, and A. Wiesel, "Unmixing k-gaussians with application to hyperspectral imaging," *IEEE Trans. Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 7281–7293, 2019.

[21] G. C. Wei and M. A. Tanner, "A Monte Carlo implementation of the em algorithm and the poor man's data augmentation algorithms," *J. Am. Stat Assoc.*, vol. 85, no. 411, pp. 699–704, 1990.

[22] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[24] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Processing*, vol. 65, no. 3, pp. 794–816, 2016.

[25] Y. Altmann, S. McLaughlin, and N. Dobigeon, "Sampling from a multivariate Gaussian distribution truncated on a simplex: a review," in *Proc. IEEE Workshop on Statistical Signal Processing (SSP)*, Gold Coast, Australia, July 2014, pp. 113–116, invited paper.

[26] S. M. Kay and S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-hall Englewood Cliffs, NJ, 1993, vol. 1.

[27] A. B. Owen, Y. Maximov, and M. Chertkov, "Importance sampling the union of rare events with an application to power systems analysis," *arXiv preprint arXiv:1710.06965*, 2018.

[28] V. Elvira and I. Santamaria, "Multiple importance sampling for symbol error rate estimation of maximum-likelihood detectors in mimo channels," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1200–1212, 2021.

[29] J. Nascimento and J. Dias, "Vertex component analysis: a fast algorithm to unmix hyperspectral data," *IEEE Trans. Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, 2005.

[30] C. Huang, M. Shao, W.-K. Ma, and A. M.-C. So, "SISAL revisited," *SIAM Journal on Imaging Sciences*, vol. 15, no. 2, pp. 591–624, 2022.

[31] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric, "Adaptive importance sampling: The past, the present, and the future," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 60–79, 2017.

[32] G. H. Golub and C. F. Van Loan, *Matrix computations*.

JHU press, 2013.