Bayesian anti-sparse coding

Clément Elvira, Pierre Chainais, Nicolas Dobigeon

▶ **To cite this version:**

Clément Elvira, Pierre Chainais, Nicolas Dobigeon. Bayesian anti-sparse coding. IEEE Transactions on Signal Processing, 2016, 10.1109/TSP.2016.2645543 . hal-01433706

HAL Id: hal-01433706
https://hal.science/hal-01433706

Submitted on 12 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bayesian anti-sparse coding

Clément Elvira, *Student Member, IEEE*, Pierre Chainais, *Senior Member, IEEE*
and Nicolas Dobigeon, *Senior Member, IEEE*

***Abstract*—Sparse representations have proven their efficiency in solving a wide class of inverse problems encountered in signal and image processing. Conversely, enforcing the information to be spread uniformly over representation coefficients exhibits relevant properties in various applications such as robust encoding in digital communications. Anti-sparse regularization can be naturally expressed through an ℓ_∞-norm penalty. This paper derives a probabilistic formulation of such representations. A new probability distribution, referred to as the democratic prior, is first introduced. Its main properties as well as three random variate generators for this distribution are derived. Then this probability distribution is used as a prior to promote anti-sparsity in a Gaussian linear model, yielding a fully Bayesian formulation of anti-sparse coding. Two Markov chain Monte Carlo (MCMC) algorithms are proposed to generate samples according to the posterior distribution. The first one is a standard Gibbs sampler. The second one uses Metropolis-Hastings moves that exploit the proximity mapping of the log-posterior distribution. These samples are used to approximate maximum a posteriori and minimum mean square error estimators of both parameters and hyperparameters. Simulations on synthetic data illustrate the performances of the two proposed samplers, for both complete and over-complete dictionaries. All results are compared to the recent deterministic variational FITRA algorithm.**

***Index Terms*—democratic distribution, anti-sparse representation, proximal operator.**

I. Introduction

SPARSE representations have been widely advocated for as an efficient tool to address various problems encountered in signal and image processing. As an archetypal example, they were the core concept underlying most of the lossy data compression schemes, exploiting compressibility properties of natural signals and images over appropriate bases. Sparse approximations, generally resulting from a *transform coding* process, lead for instance to the famous image, audio and video compression standards JPEG, MP3 and MPEG [2], [3]. More recently and partly motivated by the advent of both the compressive sensing [4], respectively, and dictionary learning paradigms [5], sparsity has been intensively exploited to regularize (e.g., linear) ill-posed inverse problems. The ℓ_0-norm and the ℓ_1-norm as its convex relaxation are among the most popular sparsity promoting penalties. Following the ambivalent interpretation of penalized regression optimization [6], Bayesian inference naturally offers an alternative and flexible framework to derive estimators associated with sparse coding problems. For instance, it is well known that a straightforward Bayesian counterpart of the LASSO shrinkage operator [7] can be obtained by adopting a Laplace prior [8]. Designing other sparsity inducing priors has motivated numerous research works. They generally rely on hierarchical mixture models [9]–[12], heavy tail distributions [13]–[15] or Bernoulli-compound processes [16]–[18].

In contrast, the use of the ℓ_∞-norm within an objective criterion has remained somehow confidential in the signal processing literature. One may cite the minimax or Chebyshev approximation principle, whose practical implementation has been made possible thanks to the Remez exchange algorithm [19] and leads to a popular design method of finite impulse response digital filters [20], [21]. Besides, when combined with a set of linear equality constraints, minimizing a ℓ_∞-norm is referred to as the minimum-effort control problem in the optimal control framework [22], [23]. Much more recently, a similar problem has been addressed by Lyubarskii *et al.* in [24] where the *Kashin's representation* of a given vector over a tight frame is introduced as the expansion coefficients with the smallest possible dynamic range. Spreading the information over representation coefficients in the most uniform way is a desirable feature in various applicative contexts, e.g., to design robust analog-to-digital conversion schemes [25], [26] or to reduce the peak-to-average power ratio (PAPR) in multi-carrier transmissions [27], [28]. Resorting to an uncertainty principle (UP), Lyubarskii *et al.* have also introduced several examples of frames yielding computable Kashin's representations, such as random orthogonal matrices, random subsampled discrete Fourier transform (DFT) matrices, and random sub-Gaussian matrices [24]. The properties of the alternate optimization problem, which consists of minimizing the maximum magnitude of the representation coefficients for an upper-bounded ℓ_2-reconstruction error, have been deeply investigated in [29], [30]. In these latest contributions, the optimal expansion is called the *democratic representation* and some bounds associated with archetypal matrices ensuring the UP are derived. In [31], the constrained signal representation problems considered in [24] and [30] are converted into their penalized counterpart. More precisely, inferring a so-called *spread* or *anti-sparse* representation $\mathbf{x}$ of an observation vector $\mathbf{y}$ under the linear model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e} \tag{1}$$

Clément Elvira and Pierre Chainais are with Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France (e-mail: {Clement.Elvira, Pierre.Chainais}@ec-lille.fr).

Nicolas Dobigeon is with the University of Toulouse, IRIT/INP-ENSEEIHT, CNRS, 2 rue Charles Camichel, BP 7122, 31071 Toulouse cedex 7, France (e-mail: Nicolas.Dobigeon@enseeiht.fr).

Part of this work has been funded thanks to the BNPSI ANR project no. ANR-13-BS-03-0006-01.

Part of this work was presented during IEEE SSP Workshop 2016 [1].
(c) 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

where $\mathbf{H}$ is the coding matrix and $\mathbf{e}$ is a residual can be formulated as a variational optimization problem where the admissible range of the coefficients has been penalized through an ℓ_∞-norm

$$\min_{\mathbf{x}\in\mathbb{R}^N} \frac{1}{2\sigma^2} \|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_\infty . \tag{2}$$

In (2), $\mathbf{H}$ defines the $M \times N$ representation matrix, and σ^2 stands for the variance of the error resulting from the approximation. In particular, the error term $\mathbf{y}-\mathbf{H}\mathbf{x}$ is referred to as the residual term throughout the paper. Again, the anti-sparse property brought by the ℓ_∞-norm penalization enforces the information brought by the measurement vector $\mathbf{y}$ to be evenly spread over the representation coefficients in $\mathbf{x}$ with respect to the dictionary $\mathbf{H}$. Note that the minimizer of (2) is generally not unique. Such representation can be desired when the coding matrix $\mathbf{H}$ is overcomplete, i.e. $N > M$. It is worth noting that recent applications have capitalized on these latest theoretical and algorithmic advances, including approximate nearest neighbor search [32] and PAPR reduction [33].

Surprisingly, up to our knowledge, no probabilistic formulation of these democratic representations has been proposed in the literature. The present paper precisely attempts to fill this gap by deriving a Bayesian formulation of the anti-sparse coding problem (2) considered in [31]. Note that this objective differs from the contribution in [34] where a Bayesian estimator associated with an ℓ_∞-norm loss function has been introduced. Instead, we merely introduce a Bayesian counterpart of the variational problem (2). The main motivations for deriving the proposed Bayesian strategy for anti-sparse coding are threefold. Firstly, Bayesian inference is a flexible methodology that may allow other parameters and hyperparameters (e.g., residual variance σ^2, regularization parameter λ) to be jointly estimated with the parameter of interest $\mathbf{x}$. Secondly, through the choice of the considered Bayes risk, it permits to make use of Bayesian estimators, beyond the standard penalized maximum likelihood estimator resulting from the solution of (2). Finally, within this framework, Markov chain Monte Carlo algorithms can be designed to generate samples according to the posterior distribution and, subsequently, approach these estimators. Contrary to deterministic optimization algorithms which provide only one point estimate, these samples can be subsequently used to build a comprehensive statistical description of the solution.

To this purpose, a new probability distribution as well as its main properties are introduced in Section II. In particular, we show that $\mathrm{p}(\mathbf{x}) \propto \exp(-\lambda \|\mathbf{x}\|_\infty)$ properly defines a probability density function (pdf), which leads to tractable computations. In Section III, this so-called *democratic distribution* is used as a prior distribution in a linear Gaussian model, which provides a straightforward equivalent of the problem (2) under the maximum *a posteriori* paradigm. Moreover, exploiting relevant properties of the democratic distribution, this section describes two Markov chain Monte Carlo (MCMC) algorithms as alternatives to the deterministic solvers proposed in [30], [31]. The first one is a standard Gibbs sampler which sequentially generates samples according to the conditional distributions associated with the joint posterior distribution. The second MCMC algorithm relies on a proximal Monte Carlo step recently introduced in [35]. This step exploits the proximal operator associated with the logarithm of the target distribution to sample random vectors asymptotically distributed according to this non-smooth density. Section IV illustrates the performances of the proposed algorithms on numerical experiments. Concluding remarks are reported in Section V.

TABLE I
LIST OF SYMBOLS.

Symbol	Description
N, n	Dimension, index of representation vector
M, m	Dimension, index of observed vector
$\mathbf{x}$, x_n	Representation vector, its n^{th} component
$\mathbf{y}$, y_m	Observation vector, its m^{th} component
$\mathbf{H}$	Coding matrix
$\mathbf{e}$	Additive residual vector
λ	Parameter of the democratic distribution
μ	Re-parametrization of λ such that $\lambda = N\mu$
$\mathcal{D}_N(\lambda)$	Democratic distribution of parameter λ over $\mathbb{R}^N$
$C_N(\lambda)$	Normalizing constant of the distribution $\mathcal{D}_N(\lambda)$
$\mathcal{K}_J$	A J-element subset $\{i_1 \dots i_J\}$ of $\{1,\dots,N\}$
$\mathcal{U}$, $\mathcal{G}$, $\mathcal{IG}$	Uniform, gamma and inverse gamma distributions
$d\mathcal{G}$	Double-sided gamma distribution
$\mathcal{N}_\mathcal{I}$	Truncated Gaussian distribution over $\mathcal{I}$
$\mathcal{C}_n$	Double convex cones partitioning $\mathbb{R}^N$
c_n, $\mathcal{I}_n$	Weights and intervals defining the conditional distribution $p\left(x_n \vert \mathbf{x}_{\setminus n}\right)$
g, g_1, g_2	Negative log-distribution ($g = g_1 + g_2$)
δ	Parameter of the proximity operator
ε_j, d_j, $\phi_\delta(\mathbf{x})$	Family of distinct values of $\vert\mathbf{x}\vert$, their respective multiplicity and family of local maxima of $\mathrm{prox}^{\delta}_{\lambda\|\cdot\|_\infty}$
$q(\cdot\vert\cdot)$	Proposal distribution
$\omega_{in}, \mu_{in}, s_n^2, \mathcal{I}_{in}$ $i=1,2,3$	Weights, parameters and intervals defining the conditional distribution $p\left(x_n \vert \mathbf{x}_{\setminus n}, \mu, \sigma^2, \mathbf{y}\right)$

II. DEMOCRATIC DISTRIBUTION

This section introduces the democratic distribution and the main properties related to its marginal and conditional distributions. Finally, two random variate generators are proposed. Note that, for the sake of conciseness, the details of the proofs associated with the following results are reported in the technical report [36].

A. Probability density function

Lemma 1. *Let $\mathbf{x} \in \mathbb{R}^N$ and $\lambda \in \mathbb{R}_+ \setminus \{0\}$. The integral of the function $\exp(-\lambda \|\mathbf{x}\|_\infty)$ over $\mathbb{R}^N$ is properly defined and the following equality holds (see proof in Appendix A)*

$$\int_{\mathbb{R}^N} \exp(-\lambda \|\mathbf{x}\|_\infty)\, d\mathbf{x} = N! \left(\frac{2}{\lambda}\right)^N .$$

As a corollary of Lemma 1, the democratic distribution can be defined as follows.

Definition 1. *A N-real-valued random vector $\mathbf{x} \in \mathbb{R}^N$ is said to be distributed according to the democratic distribution $\mathcal{D}_N(\lambda)$, namely $\mathbf{x} \sim \mathcal{D}_N(\lambda)$, when the corresponding pdf is*

$$\mathrm{p}(\mathbf{x}) = \frac{1}{C_N(\lambda)} \exp(-\lambda \|\mathbf{x}\|_\infty) \tag{3}$$

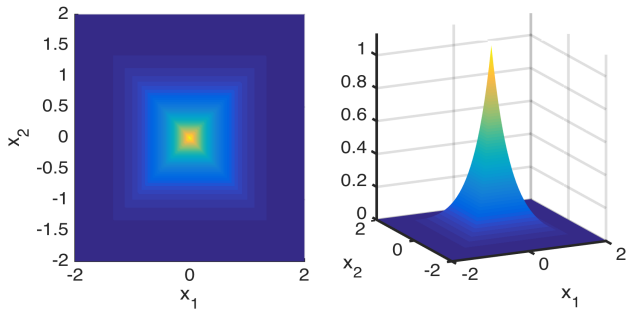

Fig. 1. The democratic pdf $\mathcal{D}_N(\lambda)$ for $N = 2$ and $\lambda = 3$.

with $C_N(\lambda) \triangleq N! \left(\frac{2}{\lambda}\right)^N$.

Fig. 1 illustrates the pdf of the bidimensional democratic pdf for $\lambda = 3$.

Remark 1. *Note that the democratic distribution belongs to the exponential family. Indeed, its pdf can be factorized as*

$$\mathrm{p}(\mathbf{x}) = a(\mathbf{x})\, b(\lambda)\, \exp\left(\eta(\lambda) T(\mathbf{x})\right) \tag{4}$$

where $a(\mathbf{x}) = 1$, $b(\lambda) = 1/C_N(\lambda)$, $\eta(\lambda) = -\lambda$ *and* $T(\mathbf{x}) = \|\mathbf{x}\|_\infty$ *defines sufficient statistics.*

B. Moments

The first two moments of the democratic distribution are available through the following property [36].

Property 1. *Let* $\mathbf{x} = [x_1, \ldots, x_N]^T$ *be a random vector obeying the democratic distribution* $\mathcal{D}_N(\lambda)$. *The mean and the covariance matrix are given by:*

$$\mathrm{E}[x_n] = 0 \qquad \forall n \in \{1, \ldots, N\} \tag{5}$$

$$\mathrm{var}[x_n] = \frac{(N+1)(N+2)}{3\lambda^2} \qquad \forall n \in \{1, \ldots, N\} \tag{6}$$

$$\mathrm{cov}[x_i, x_j] = 0 \qquad \forall i \neq j. \tag{7}$$

Note that components are pairwise decorrelated.

C. Marginal distributions

The marginal distributions of any democratically distributed vector $\mathbf{x}$ are given by the following Lemma

Lemma 2. *Let* $\mathbf{x} = [x_1, \ldots, x_N]^T$ *be a random vector obeying the democratic distribution* $\mathcal{D}_N(\lambda)$. *For any positive integer* $J < N$, *let* $\mathcal{K}_J$ *denote a* J*-element subset of* $\{1, \ldots, N\}$ *and* $\mathbf{x}_{\setminus \mathcal{K}_J}$ *the sub-vector of* $\mathbf{x}$ *whose* J *elements indexed by* $\mathcal{K}_J$ *have been removed. Then the marginal pdf of the sub-vector* $\mathbf{x}_{\setminus \mathcal{K}_J} \in \mathbb{R}^{N-J}$ *is given by (see proof in Appendix B)*

$$\mathrm{p}\left(\mathbf{x}_{\setminus \mathcal{K}_J}\right) = \frac{2^J}{C_N(\lambda)} \sum_{j=0}^{J} \binom{J}{j} \frac{(J-j)!}{\lambda^{J-j}} \left\|\mathbf{x}_{\setminus \mathcal{K}_J}\right\|_\infty^j \times \exp\left(-\lambda \left\|\mathbf{x}_{\setminus \mathcal{K}_J}\right\|_\infty\right). \tag{8}$$

In particular, as a straightforward corollary of this lemma, two specific marginal distributions of $\mathcal{D}_N(\lambda)$ are given by the following property [36].

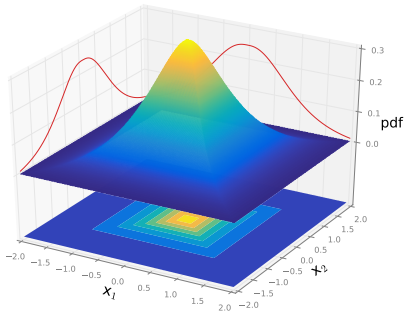

Fig. 2. Marginal distribution of $\mathbf{x}_{\setminus 3}$ when $\mathbf{x} \sim \mathcal{D}_N(\lambda)$ and $\lambda = 3$, when $N = 3$. The two red curves are the 2D marginal distributions of x_1 and x_2.

Property 2. *Let* $\mathbf{x} = [x_1, \ldots, x_N]^T$ *be a random vector obeying the democratic distribution* $\mathcal{D}_N(\lambda)$. *The components* x_n $(n = 1, \ldots, N)$ *of* $\mathbf{x}$ *are identically and marginally distributed according to the following* N*-component mixture of double-sided Gamma distributions*[1]

$$x_n \sim \frac{1}{N} \sum_{j=1}^{N} d\mathcal{G}(j, \lambda). \tag{9}$$

Moreover, the pdf of the sub-vector $\mathbf{x}_{\setminus n}$ *of* $\mathbf{x}$ *whose* n*th element has been removed is*

$$\mathrm{p}\left(\mathbf{x}_{\setminus n}\right) = \frac{1 + \lambda \left\|\mathbf{x}_{\setminus n}\right\|_\infty}{N\, C_{N-1}(\lambda)} \exp\left(-\lambda \left\|\mathbf{x}_{\setminus n}\right\|_\infty\right). \tag{10}$$

Fig. 2 illustrates the marginal distributions $\mathrm{p}\left(\mathbf{x}_{\setminus n}\right)$ and $\mathrm{p}(x_n)$.

Remark 2. *It is worth noting that the distribution in* (9) *can be rewritten as*

$$\mathrm{p}(x_n) = \frac{\lambda}{2N} \left(\sum_{j=0}^{N-1} \frac{\lambda^j}{j!} |x_n|^j \right) \exp\left(-\lambda |x_n|\right) = \frac{\lambda}{2N!} \Gamma(N, \lambda |x_n|)$$

where $\Gamma(a, b)$ *is the upper incomplete Gamma function. It is easy to prove that the random variable associated with the marginal distribution rescaled by a factor* $\frac{\lambda}{N}$ *converges in distribution to the uniform distribution* $\mathcal{U}([-1, 1])$. *The interested reader may find details in [36].*

D. Conditional distributions

Before introducing conditional distributions associated with any democratically distributed random vector, let us partition $\mathbb{R}^N$ into a set of N double cones $\mathcal{C}_n \subset \mathbb{R}^N$ $(n = 1, \ldots, N)$ defined by

$$\mathcal{C}_n \triangleq \left\{ \mathbf{x} = [x_1, \ldots, x_N]^T \in \mathbb{R}^N \ : \ \forall j \neq n,\ |x_n| \geq |x_j| \right\}. \tag{11}$$

Strictly speaking, the $\mathcal{C}_n$ are not a partition but only a covering of $\mathbb{R}^N$. However, the intersections between cones are null sets.

[1]The double-sided Gamma distribution $d\mathcal{G}(a, b)$ is defined as a generalization over $\mathbb{R}$ of the standard Gamma distribution $\mathcal{G}(a, b)$ with the pdf $\mathrm{p}(x) = \frac{b^a}{2\Gamma(a)} |x|^{a-1} \exp(-b|x|)$. See [37] for an overview.

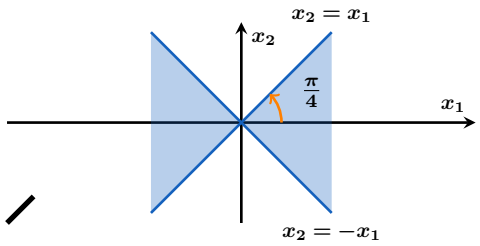

Fig. 3. The double-cone $\mathcal{C}_1$ of $\mathbb{R}^2$ appears as the shaded area while the complementary double-cone $\mathcal{C}_2$ is the uncolored area.

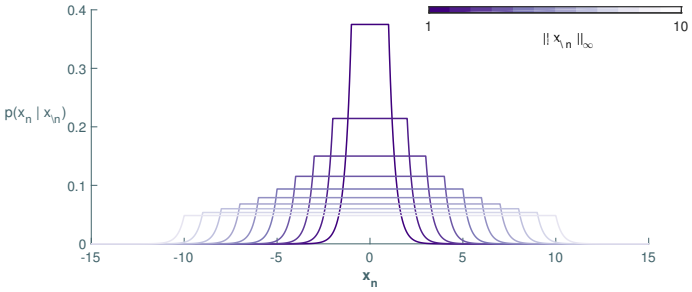

Fig. 4. Conditional distribution of $x_n|\mathbf{x}_{\backslash n}$ when $\mathbf{x} \sim \mathcal{D}_N(\lambda)$ for $N = 3$, $\lambda = 3$ and $\|\mathbf{x}_{\backslash n}\|_\infty = 1 \dots 10$.

These sets are directly related to the index of the so-called *dominant component* of a given democratically distributed vector $\mathbf{x}$. More precisely, if $\|\mathbf{x}\|_\infty = |x_n|$, then $\mathbf{x} \in \mathcal{C}_n$ and the n^{th} component x_n of $\mathbf{x}$ is said to be the dominant component. Note that this dominant component can be referred to as $x_n|\mathbf{x} \in \mathcal{C}_n$.

An example is given in Fig. 3 where $\mathcal{C}_1 \subset \mathbb{R}^2$ is depicted. These double-cones partition $\mathbb{R}^N$ into N equiprobable sets with respect to (w.r.t.) the democratic distribution, as stated in the following property [36].

Property 3. *Let $\mathbf{x} = [x_1, \dots, x_N]^T$ be a random vector obeying the democratic distribution $\mathcal{D}_N(\lambda)$. Then the probability that this vector belongs to a given double-cone is*

$$\mathrm{P}\left[\mathbf{x} \in \mathcal{C}_n\right] = \frac{1}{N}. \tag{12}$$

Remark 3. *This property can be simply proven using the intuitive intrinsic symmetries of the democratic distribution: the dominant component of a democratically distributed vector is located with equal probabilities in any of the N cones $\mathcal{C}_n$.*

Moreover, the following lemma yields some results on conditional distributions related to these sets.

Lemma 3. *Let $\mathbf{x} = [x_1, \dots, x_N]^T$ be a random vector obeying the democratic distribution $\mathcal{D}_N(\lambda)$. Then the following results hold (see proof in Appendix C-A)*

$$x_n|\mathbf{x} \in \mathcal{C}_n \sim d\mathcal{G}\left(N, \lambda\right) \tag{13}$$

$$\mathbf{x}_{\backslash n}|\mathbf{x} \in \mathcal{C}_n \sim \mathcal{D}_{N-1}(\lambda) \tag{14}$$

$$\mathbf{x}_{\backslash n}|x_n, \mathbf{x} \in \mathcal{C}_n \sim \prod_{j \neq n} \mathcal{U}\left(-|x_n|, |x_n|\right) \tag{15}$$

$$\mathrm{P}\left[\mathbf{x} \in \mathcal{C}_n|\mathbf{x}_{\backslash n}\right] = \frac{1}{1 + \lambda \left\|\mathbf{x}_{\backslash n}\right\|_\infty} \tag{16}$$

$$\mathrm{p}\left(\mathbf{x}_{\backslash n}|\mathbf{x} \notin \mathcal{C}_n\right) = \frac{\lambda}{N-1} \frac{\left\|\mathbf{x}_{\backslash n}\right\|_\infty}{C_{N-1}(\lambda)} e^{-\lambda \left\|\mathbf{x}_{\backslash n}\right\|_\infty}. \tag{17}$$

Remark 4. *According to* (13)*, the marginal distribution of the dominant component is a double-sided Gamma distribution. Conversely, according to* (14)*, the vector of the non-dominant components is marginally distributed according to a democratic distribution. Conditioned to the dominant component, the non-dominant components are independently and uniformly distributed on the admissible set, as shown in* (15)*. Equation* (16) *shows that the probability that the nth component dominates increases when the other components are of low amplitude.*

Finally, based on Lemma 3, the following property related to the conditional distributions of $\mathcal{D}_N(\lambda)$ can be stated.

Property 4. *Let $\mathbf{x} = [x_1, \dots, x_N]^T$ be a random vector obeying the democratic distribution $\mathcal{D}_N(\lambda)$. The pdf of the conditional distribution of a given component x_n given $\mathbf{x}_{\backslash n}$ is (See proof in Appendix C-B)*

$$\mathrm{p}\left(x_n|\mathbf{x}_{\backslash n}\right) = (1 - c_n) \frac{1}{2\left\|\mathbf{x}_{\backslash n}\right\|_\infty} \mathbb{1}_{\mathcal{I}_n}(x_n) + c_n \frac{\lambda}{2} e^{-\lambda\left(|x_n| - \left\|\mathbf{x}_{\backslash n}\right\|_\infty\right)} \mathbb{1}_{\mathbb{R}\backslash\mathcal{I}_n}(x_n) \tag{18}$$

where $c_n = \mathrm{P}\left[\mathbf{x} \in \mathcal{C}_n|\mathbf{x}_{\backslash n}\right]$ is given by (16)*, $\mathbb{1}_A(x)$ is the indicator function whose value is 1 if $x \in A$ and 0 otherwise, and $\mathcal{I}_n$ is defined as follows*

$$\mathcal{I}_n \triangleq \left(-\left\|\mathbf{x}_{\backslash n}\right\|_\infty, \left\|\mathbf{x}_{\backslash n}\right\|_\infty\right). \tag{19}$$

Remark 5. *The pdf in* (18) *defines a mixture of one uniform distribution and two shifted exponential distributions with probabilities $1 - c_n$ and $c_n/2$, respectively. An example of this pdf is depicted in Fig. 4.*

E. Proximity operator of the negative log-pdf

The pdf of the democratic distribution $\mathcal{D}_N(\lambda)$ can be written as $\mathrm{p}(\mathbf{x}) \propto \exp\left(-g_1\left(\mathbf{x}\right)\right)$ with

$$g_1\left(\mathbf{x}\right) = \lambda \left\|\mathbf{x}\right\|_\infty. \tag{20}$$

This subsection introduces the proximity mapping operator associated with the negative log-distribution $g_1(\mathbf{x})$ (defined up to a multiplicative constant). This proximal operator will be subsequently used to implement Monte Carlo algorithms to draw samples from the democratic distribution $\mathcal{D}_N(\lambda)$ (see Section II-F) as well as posterior distributions derived from a democratic prior (see Section III-B3b). In this context, it is convenient to define the proximity operator of $g_1\left(\cdot\right)$ as [38]

$$\mathrm{prox}_{g_1}(\mathbf{x}) = \underset{\mathbf{u} \in \mathbb{R}^N}{\mathrm{argmin}}\ \lambda \left\|\mathbf{u}\right\|_\infty + \frac{1}{2}\|\mathbf{x} - \mathbf{u}\|_2^2. \tag{21}$$

Since g_1 is a norm, its proximity mapping can be simply linked to the projection over the ball generated by the dual norm [39], *i.e.*, the ℓ_1-norm here. Thus, one has

$$\mathrm{prox}_{\delta g_1}(\mathbf{x}) = \mathbf{x} - \lambda\delta\Pi_{\|\mathbf{x}/\lambda\delta\|_1 \leq 1}(\mathbf{x}) \tag{22}$$

where $\Pi_{\|\mathbf{x}\|_1 \leq 1}$ is the projector into the unit ℓ_1-ball. Although this projection cannot be performed directly, fast numerical techniques have been recently investigated. For an overview and an implementation, see for instance [40].

F. Random variate generation

This section introduces a random variate generator that permits to draw samples according to the democratic distribution. Two others methods are also suggested.

1) Exact random variate generator: Property 3 combined with Lemma 3 permits to rewrite the joint distribution of a democratically distributed vector using the chain rule

$$\begin{aligned} \mathrm{p}(\mathbf{x}) &= \sum_{n=1}^{N} \mathrm{p}\left(\mathbf{x}_{\setminus n} | x_n, \mathbf{x} \in \mathcal{C}_n\right) \mathrm{p}\left(x_n | \mathbf{x} \in \mathcal{C}_n\right) \mathrm{P}\left[\mathbf{x} \in \mathcal{C}_n\right] \\ &= \sum_{n=1}^{N} \left[\prod_{j \neq n} \mathrm{p}\left(x_j | x_n, \mathbf{x} \in \mathcal{C}_n\right) \right] \mathrm{p}\left(x_n | \mathbf{x} \in \mathcal{C}_n\right) \mathrm{P}\left[\mathbf{x} \in \mathcal{C}_n\right] \end{aligned} \tag{23}$$

where $\mathrm{P}\left[\mathbf{x} \in \mathcal{C}_n\right]$, $\mathrm{p}\left(x_n | \mathbf{x} \in \mathcal{C}_n\right)$ and $\mathrm{p}\left(x_j | x_n, \mathbf{x} \in \mathcal{C}_n\right)$ are given in (12), (13) and (15), respectively. Note that the conditional joint distribution of the non-dominant components is decomposed as a product, see (15) and Remark 4 right after. This finding can be fully exploited to design an efficient and exact random variate generator for the democratic distribution, see Algo. 1.

Algorithm 1: Democratic random variate generator using an exact sampling scheme.

Input: Parameter $\lambda > 0$, dimension N

```
% Drawing the cone of the dominant component
Sample n_dom uniformly on the set {1...N};
% Drawing the dominant component
Sample x_{n_dom} according to (13);
% Drawing the non-dominant components
for j ← 1 to N (j ≠ n_dom) do
    Sample x_j according to (15);
end
```

Output: $\mathbf{x} = [x_1, \ldots, x_N]^T \sim \mathcal{D}_N(\lambda)$

2) Other strategies: Although exact sampling is by far the most elegant and efficient method to sample according to the democratic distribution, two others strategies can be evoked. Firstly, one may want to exploit Property 4 to design a Gibbs sampling scheme by successively drawing the components x_n according to the conditional distributions. Secondly, the proximal operator of the negative log-pdf can be used within the proximal Metropolis-adjusted Langevin algorithm (P-MALA) introduced in [35]. See [36] for a comparison between the three samplers. These findings pave the way to extended schemes for sampling according to a posterior distribution resulting from a democratic prior when possibly no exact sampler is available, see Section III.

III. Democratic prior in a linear Gaussian model

This section aims to provide a Bayesian formulation of the model underlying the problem described by (2). From a Bayesian perspective, the solution of (2) can be straightforwardly interpreted as the MAP estimator associated with a linear observation model characterized by an additive Gaussian residual and complemented by a democratic prior assumption. Assuming a Gaussian residual results in a quadratic discrepancy measure as in (2). Setting the anti-sparse coding problem into a fully Bayesian framework paves the way to a comprehensive statistical description of the solution. The resulting posterior distribution can be subsequently sampled to approximate the Bayesian estimators, e.g., not only the MAP but also the MMSE estimators.

A. Hierarchical Bayesian model

Let $\mathbf{y} = [y_1 \ldots y_M]^T$ denote an observed measurement vector. The problem addressed in this work consists of recovering an anti-sparse code $\mathbf{x} = [x_1, \ldots, x_N]^T$ of these observations given the coding matrix $\mathbf{H}$ according to the linear model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e}. \tag{24}$$

The residual vector $\mathbf{e} = [e_1 \ldots e_M]^T$ is assumed to be distributed according to a centered multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}_M, \sigma^2 \boldsymbol{I}_M)$, where $\mathbf{0}_M$ is a M-dimensional vector of 0 and $\boldsymbol{I}_M$ is the identity matrix of size $M \times M$. The choice and the design of the coding matrix H should ensure the existence of a democratic representation with a small dynamic range [24]. The proposed Bayesian model relies on the definition of the likelihood function associated with the observation vector $\mathbf{y}$ and on the choice of prior distributions for the unknown parameters, i.e., the representation vector $\mathbf{x}$ and the residual variance σ^2, assumed to be a priori independent.

1) Likelihood function: the Gaussian property of the additive residual term yields the following likelihood function

$$f(\mathbf{y}|\mathbf{x}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{M}{2}} \exp\left[-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2\right]. \tag{25}$$

2) Residual variance prior: a noninformative Jeffreys prior distribution is chosen for the residual variance σ^2

$$f\left(\sigma^2\right) \propto \frac{1}{\sigma^2}. \tag{26}$$

3) Description vector prior: as motivated earlier, the democratic distribution is elected as the prior distribution of the N-dimensional vector $\mathbf{x}$

$$\mathbf{x} \mid \lambda \sim \mathcal{D}_N(\lambda). \tag{27}$$

In the following, the hyperparameter λ is set as $\lambda = N\mu$, where μ is assumed to be unknown. Enforcing the parameter of the democratic distribution to depend on the dimension of the problem permits the prior to be scaled with this dimension. Indeed, as stated in (13), the absolute value of the dominant component is distributed according to the Gamma distribution $\mathcal{G}(N, \lambda)$, whose mean and variance are N/λ and N/λ^2, respectively. With the proposed scalability, the prior mean is constant w.r.t. the dimension

$$\mathrm{E}\left[|x_n| \mid \mathbf{x} \in \mathcal{C}_n, \mu\right] = 1/\mu \tag{28}$$

and the variance tends to zero

$$\text{var}\left[|x_n| \mid \mathbf{x} \in \mathcal{C}_n, \mu\right] = 1/(N\mu^2). \tag{29}$$

4) Hyperparameter prior: the prior modeling introduced in the previous section is complemented by assigning prior distribution to the unknown hyperparameter μ, introducing a second level in the Bayesian hierarchy. More precisely, a conjugate Gamma distribution is chosen as a prior for μ

$$\mu \sim \mathcal{G}(a, b) \tag{30}$$

since conjugacy allows the posterior distribution to be easily derived. The parameters a and b will be chosen to obtain a flat prior[2].

5) Posterior distribution: the posterior distribution of the unknown parameter vector $\boldsymbol{\theta} = \{\mathbf{x}, \sigma^2, \mu\}$ can be computed from the following hierarchical structure:

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{x}, \sigma^2) f(\mathbf{x}|\mu) f(\mu) f(\sigma^2) \tag{31}$$

where $f(\mathbf{y}|\mathbf{x}, \sigma^2)$, $f(\sigma^2)$, $f(\mathbf{x}|\mu)$ and $f(\mu)$ have been defined in (25) to (30), respectively. Thus, this posterior distribution can be written as

$$\begin{aligned} f(\mathbf{x}, \sigma^2, \mu|\mathbf{y}) \propto & \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{H}\mathbf{x}\right\|_2^2\right) \\ & \times \frac{1}{C_N(\mu N)} \exp\left(-\mu N \left\|\mathbf{x}\right\|_\infty\right) \\ & \times \left(\frac{1}{\sigma^2}\right)^{\frac{M}{2}+1} \mathbf{1}_{\mathbb{R}_+}(\sigma^2) \\ & \times \frac{b^a}{\Gamma(b)} \mu^{a-1} \exp\left(-b\mu\right) \mathbf{1}_{\mathbb{R}_+}(\mu). \end{aligned} \tag{32}$$

As expected, for given values of the residual variance σ^2 and the democratic parameter $\lambda = \mu N$, maximizing the posterior (32) can be formulated as the optimization problem in (2), for which some algorithmic strategies have been for instance introduced in [30] and [31]. In this paper, a different route has been taken by deriving inference schemes relying on MCMC algorithms. This choice permits to include the nuisance parameters σ^2 and μ into the model and to estimate them jointly with the representation vector $\mathbf{x}$. Moreover, since the proposed MCMC algorithms generate a collection $\left\{\left(\mathbf{x}^{(t)}, \mu^{(t)}, \sigma^{2(t)}\right)\right\}_{t=1}^{N_{\text{MC}}}$ asymptotically distributed according to the posterior of interest (31), they provide a good knowledge of the statistical distribution of the solutions.

B. MCMC algorithm

This section introduces two MCMC algorithms to generate samples according to the posterior (32). There are two specific instances of Gibbs samplers which generate samples according to the conditional distributions associated with the posterior (32), see Algo. 2. As shown below, the steps for sampling according to the conditional distributions of the residual variance $f(\sigma^2|\mathbf{y}, \mathbf{x})$ and the democratic parameter $f(\mu|\mathbf{x})$ are straightforward. In addition, generating samples from $f(\mathbf{x}|\mu, \mathbf{y})$ can be achieved component-by-component using N Gibbs moves. However, for high dimensional problems, such a crude strategy may suffer from poor mixing properties, leading to slow convergence of the algorithm. To alleviate this issue, it is also possible to use an alternative approach consisting of sampling the full vector $\mathbf{x}|\mu, \mathbf{y}$ using a P-MALA step [35]. These two strategies are detailed in the following paragraphs. Note that the implementation has been validated using the sampling procedure proposed by Geweke in [41]. For more details about this experiment, see [36, Section IV-A].

[2]Typically $a = b = 10^{-6}$ in the experiments reported in Section IV.

Algorithm 2: Gibbs sampler

Input: Observation vector $\mathbf{y}$, coding matrix $\mathbf{H}$, hyperparameters a and b, number of burn-in iterations T_{bi}, total number of iterations T_{MC}, algorithmic parameter δ, initialization $\mathbf{x}^{(0)}$

for $t \leftarrow 1$ **to** T_{MC} **do**
- *% Drawing the residual variance*
- Sample $\sigma^{2(t)}$ according to (33). ;
- *% Drawing the democratic parameter*
- Sample $\mu^{(t)}$ according to (35). ;
- *% Drawing the representation vector*
- Sample $\mathbf{x}^{(t)}$ using, either (see Section III-B3)
  - **for** $n \leftarrow 1$ **to** N **do**
    Gibbs sample x_n, see (36) ;
  - **end**
  - or
  - **for** $t^* \leftarrow 1$ **to** $T_{\text{P-MALA}}$ **do**
    P-MALA step, sample $\mathbf{x}^*$ according to (40) and accept it with probability given by (41);
  - **end**

end

Output: A collection of samples $\left\{\mu^{(t)}, \sigma^{2(t)}, \mathbf{x}^{(t)}\right\}_{t=T_{\text{bi}}+1}^{T_{\text{MC}}}$ asymptotically distributed according to (32).

1) Sampling the residual variance: Sampling according to the conditional distribution of the residual variance can be conducted according to the following inverse-gamma distribution

$$\sigma^2|\mathbf{y}, \mathbf{x} \sim \mathcal{IG}\left(\frac{M}{2}, \frac{1}{2}\left\|\mathbf{y} - \mathbf{H}\mathbf{x}\right\|_2^2\right). \tag{33}$$

2) Sampling the democratic hyperparameter: Looking carefully at (32), the conditional posterior distribution of the democratic parameter μ is

$$f(\mu|\mathbf{x}) \propto \mu^N \exp\left(-\mu N \left\|\mathbf{x}\right\|_\infty\right) \mu^{a-1} \exp\left(-b\mu\right). \tag{34}$$

Therefore, sampling according to $f(\mu|\mathbf{x})$ is achieved as follows

$$\mu|\mathbf{x} \sim \mathcal{G}(a + N, b + N \left\|\mathbf{x}\right\|_\infty). \tag{35}$$

3) Sampling the description vector: Following the technical developments of Section II-F, two strategies can be considered to generate samples according to the conditional posterior distribution of the representation vector $f(\mathbf{x}|\mu, \sigma^2, \mathbf{y})$. They are detailed below.

a) Component-wise Gibbs sampling: A first possibility to draw a vector $\mathbf{x}$ according to $f(\mathbf{x}|\mu,\sigma^2,\mathbf{y})$ is to successively sample according to the conditional distribution of each component given the others, namely, $f(x_n|\mathbf{x}_{\setminus n},\mu,\sigma^2,\mathbf{y})$. More precisely, straightforward computations yield the following 3-mixture of truncated Gaussian distributions for this conditional

$$x_n|\mathbf{x}_{\setminus n},\mu,\sigma^2,\mathbf{y} \sim \sum_{i=1}^{3} \omega_{in}\mathcal{N}_{\mathcal{I}_{in}}\left(\mu_{in},s_n^2\right) \tag{36}$$

where $\mathcal{N}_{\mathcal{I}}(\cdot,\cdot)$ denotes the Gaussian distribution restricted to $\mathcal{I}$ and the intervals are defined as

$$\begin{aligned}\mathcal{I}_{1n} &= \left(-\infty,-\left\|\mathbf{x}_{\setminus n}\right\|_\infty\right]\\ \mathcal{I}_{2n} &= \left(-\left\|\mathbf{x}_{\setminus n}\right\|_\infty,\left\|\mathbf{x}_{\setminus n}\right\|_\infty\right)\\ \mathcal{I}_{3n} &= \left[\left\|\mathbf{x}_{\setminus n}\right\|_\infty,+\infty\right).\end{aligned} \tag{37}$$

The probabilities ω_{in} $(i=1,2,3)$ as well as the (hidden) means μ_{in} $(i=1,2,3)$ and variance s_n^2 of these truncated Gaussian distributions are given in Appendix D. This specific nature of the conditional distribution is intrinsically related to the nature of the conditional prior distribution stated in Property 4, which has already exhibited a 3-component mixture: one uniform distribution and two (shifted) exponential distributions defined over $\mathcal{I}_{2n}$, $\mathcal{I}_{1n}$ and $\mathcal{I}_{3n}$, respectively (see Remark 5). Note that sampling according to truncated distributions can be achieved using the strategy proposed in [42].

b) P-MALA: Sampling according to the conditional distribution $f(\mathbf{x}|\mu,\sigma^2,\mathbf{y})$ can be achieved using a P-MALA step [35]. P-MALA uses the proximity mapping of the negative log-posterior distribution. In this case, the distribution of interest can be written as

$$f(\mathbf{x}|\mu,\sigma^2,\mathbf{y}) \propto \exp\left(-g\left(\mathbf{x}\right)\right)$$

where $g\left(\mathbf{x}\right)$ derives from the Gaussian (negative log-) likelihood function and the (negative log-) distribution of the democratic prior so that

$$g(\mathbf{x}) = \frac{1}{2\sigma^2}\left\|\mathbf{y}-\mathbf{H}\mathbf{x}\right\|_2^2 + \lambda\left\|\mathbf{x}\right\|_\infty \tag{38}$$

with $\lambda = \mu N$. However, up to the authors' knowledge, the proximal operator associated with $g(\cdot)$ in (38) has no closed-form solution. To alleviate this problem, a first order approximation is considered[3], as recommended in [35]

$$\text{prox}_{\frac{\delta}{2}g}(\mathbf{x}) \approx \text{prox}_{\frac{\delta}{2}g_1}\left(\mathbf{x}+\delta\,\nabla\left[\frac{1}{2\sigma^2}\left\|\mathbf{y}-\mathbf{H}\mathbf{x}\right\|_2^2\right]\right) \tag{39}$$

where $g_1(\cdot)=\lambda\left\|\cdot\right\|_\infty$ has been defined in Section II-E, δ is a control parameter and the corresponding proximity mapping is described in Section II-E.

Finally, at iteration t of the main algorithm, sampling according to the conditional distribution $f(\mathbf{x}|\mu,\sigma^2,\mathbf{y})$ is performed by drawing a candidate

$$\mathbf{x}^*|\mathbf{x}^{(t-1)} \sim \mathcal{N}\left(\text{prox}_{\frac{\delta}{2}g}\left(\mathbf{x}^{(t-1)}\right),\delta\mathbf{I}_N\right) \tag{40}$$

[3]Note that a similar step is involved in the fast iterative truncation algorithm (FITRA) [33], a deterministic counterpart of the proposed algorithm and considered in the next section for comparison.

and either keep $\mathbf{x}^{(t)}=\mathbf{x}^{(t-1)}$ or accept this candidate $\mathbf{x}^*$ as the new state $\mathbf{x}^{(t)}$ with probability

$$\alpha = \min\left(1,\frac{f\left(\mathbf{x}^*|\mu,\sigma^2,\mathbf{y}\right)}{f\left(\mathbf{x}^{(t-1)}|\mu,\sigma^2,\mathbf{y}\right)}\frac{\mathrm{q}\left(\mathbf{x}^{(t-1)}|\mathbf{x}^*\right)}{\mathrm{q}\left(\mathbf{x}^*|\mathbf{x}^{(t-1)}\right)}\right). \tag{41}$$

Note that the first order approximation made in (39) has no impact on the posterior distribution, since the proposition is then adjusted by a Metropolis-Hastings scheme.

The hyperparameter δ required in (40) is dynamically tuned to reach an average acceptance rate for the Metropolis Hastings step between 0.4 and 0.6, as suggested in [35].

C. Inference

The sequences $\left\{\mathbf{x}^{(t)},\sigma^{2(t)},\mu^{(t)}\right\}_{t=1}^{T_{\text{MC}}}$ generated by the MCMC algorithms proposed in Section III-B are used to approximate Bayesian estimators. After a burn-in period of N_{bi} iterations, the set of generated samples

$$\mathcal{X} \triangleq \left\{\mathbf{x}^{(t)}\right\}_{t=T_{\text{bi}}+1}^{T_{\text{MC}}} \tag{42}$$

is asymptotically distributed according to the marginal posterior distribution $f\left(\mathbf{x}|\mathbf{y}\right)$, resulting from the marginalization of the joint posterior distribution $f\left(\mathbf{x},\sigma^2,\mu|\mathbf{y}\right)$ in (32) over the nuisance parameters σ^2 and μ

$$f\left(\mathbf{x}|\mathbf{y}\right) = \int f\left(\mathbf{x},\sigma^2,\mu|\mathbf{y}\right)\mathrm{d}\sigma^2\mathrm{d}\mu \tag{43}$$

$$\propto \left\|\mathbf{y}-\mathbf{H}\mathbf{x}\right\|_2^{-\frac{M}{2}}\left(b+N\left\|\mathbf{x}\right\|_\infty\right)^{-(a+N)}. \tag{44}$$

As a consequence, the minimum mean square error (MMSE) estimator of the representation vector $\mathbf{x}$ can be approximated as an empirical average over the set $\mathcal{X}$

$$\hat{\mathbf{x}}_{\text{MMSE}} = \mathrm{E}\left[\mathbf{x}|\mathbf{y}\right] \tag{45}$$

$$\simeq \frac{1}{T_{\text{MC}}-T_{\text{bi}}}\sum_{t=T_{\text{bi}}+1}^{T_{\text{MC}}}\mathbf{x}^{(t)}. \tag{46}$$

The marginal maximum a posteriori (mMAP) estimator can be approximated as

$$\hat{\mathbf{x}}_{\text{mMAP}} = \underset{\mathbf{x}\in\mathbb{R}^N}{\text{argmax}}\, f\left(\mathbf{x}|\mathbf{y}\right) \tag{47}$$

$$\simeq \underset{\mathbf{x}^{(t)}\in\mathcal{X}}{\text{argmax}}\, f\left(\mathbf{x}^{(t)}|\mathbf{y}\right). \tag{48}$$

IV. Experiments

This section reports several simulation results to illustrate the performance of the Bayesian anti-sparse coding algorithms introduced in Section III. Section IV-A evaluates the performances of the two versions of the samplers (i.e., using Gibbs or P-MALA steps) on a toy example, by considering measurements resulting from a representation vector whose coefficients are democratically distributed. Finally, Section IV-B compares the performances of the proposed algorithm and its deterministic counterpart introduced in [33], as well as the method proposed in [32]. For all experiments, the coding matrices $\mathbf{H}$ have been chosen as randomly subsampled columnwise DCT matrices since they have shown to yield democratic

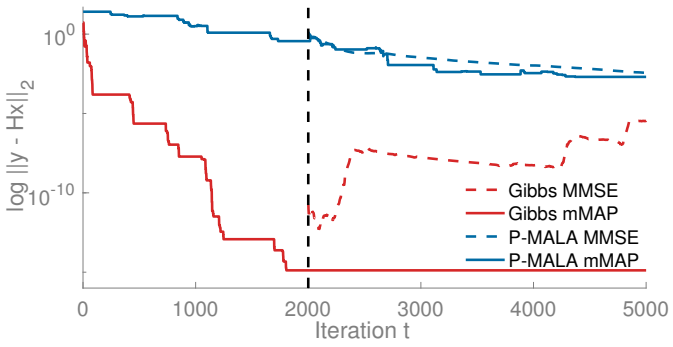

Fig. 5. As functions of the iteration number, approximation errors associated with mMAP and MMSE estimates computed using the two proposed algorithms. The end of the burn-in period is localized with a vertical black line. Results are averaged over 100 Monte Carlo simulations.

representations with small ℓ_∞-norm and good democracy bounds [30]. However, note that a deep investigation of these bounds is out of the scope of the paper. In all experiments, we will denote "Gibbs" or "full Gibbs" the algorithm where the vector $\mathbf{x}$ is updated component-wisely (see (36)), and "P-MALA" our second algorithm where $\mathbf{x}$ is updated with a P-MALA step (see (40) and (41)).

A. A toy example

This section focuses on a toy example to study the convergence of the two versions of the proposed algorithm. Experiments are carried out in dimensions $M = 12$ and $N = 15$. For each set of parameters, 100 Monte Carlo simulations are run. For each Monte-Carlo run, data are generated as follows. A coding matrix $\mathbf{H}$ is generated and a coding vector $\mathbf{x}$ is sampled according to a democratic distribution of parameter $\mu = 2$. The observation $\mathbf{y}$ is then simulated according to $\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{n}$, where $\boldsymbol{n}$ is an additive Gaussian noise. The variance of n has been adjusted to reach an average noise level of 20dB, *i.e.*, $10\log_{10}\frac{\|Hx\|_2^2}{M\sigma^2} = 20$. Recall that the purpose will be to estimate a democratic code $\hat{\mathbf{x}}$ that ensures a good description of $\mathbf{y}$. Two criteria have been used to evaluate the performance of the estimators

$$\text{SNR}_{\mathbf{y}} = 10\log_{10}\frac{\|\mathbf{y}\|_2^2}{\|\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}\|_2^2} \tag{49}$$

$$\text{PAPR} = \frac{N\|\hat{\mathbf{x}}\|_\infty^2}{\|\hat{\mathbf{x}}\|_2^2} \tag{50}$$

where $\hat{\mathbf{x}}$ refers to the MMSE or mMAP estimator of $\mathbf{x}$. The signal-to-noise ratio $\text{SNR}_{\mathbf{y}}$ measures the quality of the approximation. Conversely, the peak-to-average power ratio PAPR quantifies anti-sparsity by measuring the ratio between the crest of a signal and its average energy. Note that the proposed algorithms do not aim at directly minimizing the PAPR: the use of a democratic distribution prior should promote anti-sparsity and therefore anti-sparse representations with low PAPR.

Fig. 5 shows the evolution of the reconstruction error for all estimators, seen as a function of the number of iterations. The plots show that all algorithms converge to solutions that ensure reconstruction error of the observation vector lower than 1 while the lowest is reached by the full Gibbs sampler. Note that the average value of $\|\mathbf{y}\|_2$ is 4.2.

Finally, with a personal computer equipped with a 2.8GHz Intel i5 processor, the simulation of 5000 samples requires 2 minutes using Gibbs sampling and only 15 seconds using P-MALA steps. These observations highlight the fact that the algorithm based on P-MALA steps is much faster, even though the reconstruction error decreases slower compared to the Gibbs version.

To alleviate this limitation, the strategy adopted in the next experiments performs $T_{\text{P-MALA}} = 20$ Metropolis-Hastings moves (40) and (41) within a single iteration of the MCMC algorithm (as recommended in [35]).

TABLE II
RESULTS IN TERMS OF APPROXIMATION ERROR, PAPR AND ESTIMATED μ AND σ^2.

	$\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2$	PAPR	μ	σ^2
Original code	4.2×10^{-1}	2.8	2.0	3.9×10^{-1}
Gibbs MMSE	1.1×10^{-2}	2.9	1.6	2.8×10^{-2}
Gibbs mMAP	1.9×10^{-6}	2.9	*idem*	*idem*
P-MALA MMSE	2.1×10^{-1}	2.7	1.2	2.4×10^{-1}
P-MALA mMAP	1.6×10^{-1}	2.6	*idem*	*idem*

Table II shows the average performance of all estimators compared to the original values of parameters. The PAPR of all estimates are either close to or even lower than the original one for a better approximation of $\mathbf{y}$. We emphasize that the main objective here is to estimate an anti-sparse coding vector $\hat{\mathbf{x}}$ of observations $\mathbf{y}$. In the present experiment, even though $\mathbf{y}$ are built from noise-free vectors $\mathbf{H}\mathbf{x}$ corrupted by a Gaussian noise, a coding vector $\hat{\mathbf{x}}$ of $\mathbf{y}$ with similar or even lower PAPR than the original code $\mathbf{x}$ can be inferred with a lower approximation error as well. This apparent paradox can be easily explained by a behavior akin to over-fitting: the noise component itself is also democratically encoded by $\mathbf{H}$. This is not a problem here since the purpose is not to recover some hidden true $\mathbf{x}$ as would be the case for a denoising task.

B. Application to spread representation

1) Experimental setup: In this experiment, the observation vector $\mathbf{y}$ is composed of coefficients independently and identically distributed according to a Gaussian distribution, as in [30]. The proposed MCMC algorithm is applied to infer the anti-sparse representation $\mathbf{x}$ of this measurement vector $\mathbf{y}$ w.r.t. the $M \times N$ coding matrix $\mathbf{H}$ for two distinct scenarios. Again, $\mathbf{H}$ corresponds to randomly subsampled columnwise DCT matrices. Scenario 1 considers a small dimension problem with $M = 40$ and $N = 60$. In scenario 2, a higher dimension problem has been addressed, i.e., with $M = 128$ and N ranging from 128 to 256, which permits to evaluate the performance of the algorithm as a function of the ratio N/M. In Scenario 1 (resp., Scenario 2), the proposed mMAP and MMSE estimators are computed from a total of $T_{\text{MC}} = 12 \times 10^3$ (resp., $T_{\text{MC}} = 55 \times 10^3$) iterations, including $T_{\text{bi}} = 10 \times 10^3$ (resp., $T_{\text{bi}} = 50 \times 10^3$) burn-in iterations.

For this latest scenario, the algorithm based on Gibbs steps (see Section III-B3a) has not been considered because of its computational burden, which experimentally justifies the interest of the proximal MCMC-based approach for large scale problems.

The proposed algorithm is compared with a recent PAPR reduction technique, detailed in [33] and an anti-sparse coding scheme proposed in [32]. The fast iterative truncation algorithm (FITRA) proposed in [33] is a deterministic counterpart of the proposed MCMC algorithm and solves the ℓ_∞-penalized least-squares problem (2) using a forward backward-method. Note that FITRA could be algorithmically improved since the recent work of Condat [40]. In [32] the authors propose a path following inspired-algorithm (PFA) that also solves the variational counterpart of the considered Bayesian anti-sparse coding problem. Note that the PFA-oriented scheme is derived in [32] to solve the constrained version of (2) while the codes provided by the authors[4] solve the corresponding variational problem. Similarly to various variational techniques, FITRA needs the prior knowledge of the hyperparameters λ (anti-sparsity level) and σ^2 (residual variance) or, equivalently, of the regularization parameter β defined (up to a constant) as the product of the two hyperparameters, i.e., $\beta \triangleq 2\lambda\sigma^2$. As a consequence, in the following experiments, this parameter β has been chosen according to 3 distinct rules. The first one, denoted FITRA-mmse, consists of applying FITRA with $\beta = 2\hat{\lambda}_{\text{MMSE}}\hat{\sigma}^2_{\text{MMSE}}$, where $\hat{\lambda}_{\text{MMSE}}$ and $\hat{\sigma}^2_{\text{MMSE}}$ are the MMSE estimates obtained with the proposed P-MALA based algorithm. In the second and third configurations, the regularization parameter β has been tuned to reach two solutions with the same figure-of-merits as P-MALA mMAP, either in terms of reconstruction error $\text{SNR}_{\mathbf{y}}$ (and free PAPR) or anti-sparsity level PAPR (and free $\text{SNR}_{\mathbf{y}}$). These two solutions are denoted FITRA-snr and FITRA-papr, respectively. For all these configurations, FITRA has been run with a maximum of 500 iterations. Following the implementation provided with [32], PFA also needs the prior knowledge of a hyperparameter h. As for FITRA, two versions are proposed, with $h = 2\hat{\lambda}_{\text{MMSE}}\hat{\sigma}^2_{\text{MMSE}}$ (PFA-mmse) and h tuned to reach the same $\text{SNR}_{\mathbf{y}}$ as P-MALA mMAP (PFA-snr). Note that no version of PFA with a targeted PAPR is presented, since PFA systematically produces solutions with smaller PAPR than P-MALA. Note that we do note expect to perform better than FITRA or PFA, since both algorithms are supervised while the two proposed methods are fully unsupervised.

Moreover, to illustrate the regularizing effect of the democratic prior (or, similarly, the ℓ_∞-penalization), the proposed algorithm and the 3 configurations of FITRA have been finally compared with the least-squares[5] (LS) solution as well as the MMSE and mMAP estimates resulting from a Bayesian model based on a Gaussian prior (or, similarly, an ℓ_2-penalization). Algorithm performances have been evaluated over 50 Monte Carlo simulations in terms of reconstruction error $\text{SNR}_{\mathbf{y}}$ and PAPR, respectively given by (49) and (50).

[4]Available at http://gforge.inria.fr/projects/antisparse/

[5]The LS solution has been computed from the pseudo-inverse of $\mathbf{H}$.

TABLE III
SCENARIO 1: RESULTS IN TERMS OF $\text{SNR}_{\mathbf{y}}$ AND PAPR FOR VARIOUS ALGORITHMS.

	$\text{SNR}_{\mathbf{y}}$ (dB)	PAPR	time (s)
P-MALA mMAP	29.3	2.8	3.2×10^1
P-MALA MMSE	19.3	3.9	*idem*
Gibbs mMAP	8.8	3.0	1.4×10^2
Gibbs MMSE	4.3	2.9	*idem*
FITRA-mmse	34.4	1.7	2.6×10^{-2}
FITRA-snr	29.3	1.9	*idem*
FITRA-papr	83.5	2.9	*idem*
PFA-mmse	29.6	1.6	5.3×10^{-2}
PFA-snr	29.3	1.8	*idem*
LS	∞	6.6	6.1×10^{-2}
Gibbs mMAP (Gaussian)	∞	5.9	1.9×10^1
Gibbs MMSE (Gaussian)	73.1	6.8	*idem*

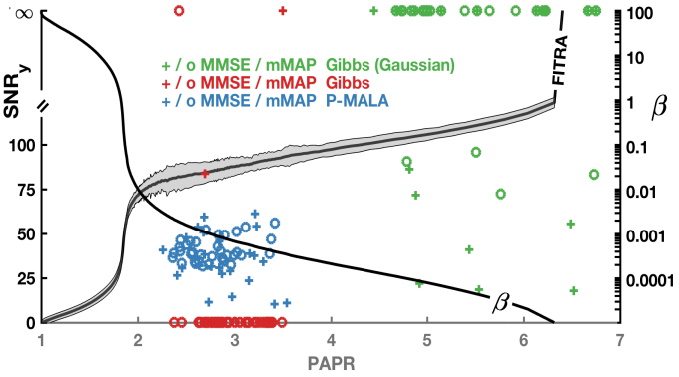

Fig. 6. Scenario 1: $\text{SNR}_{\mathbf{y}}$ as a function of PAPR. The black line associated with the left axis represent all the solutions of FITRA for various value of the parameter β. The path for the FITRA parameter β is also depicted as black line with scale in the right y-axis. The green, blue and red points are the estimators proposed respectively by the Gaussian model, the full Gibbs sampler and P-MALA, associated with the left axis. For all three algorithms, MMSE and mMAP estimators are depicted respectively by crosses and circles.

2) Results: Table III shows the results in Scenario 1 ($M = 40$ and $N = 60$) for all considered algorithms in terms of $\text{SNR}_{\mathbf{y}}$ and PAPR. For this scenario, the full Gibbs method needs approximately 2.5 minutes while P-MALA needs 32 seconds only. The mMAP and the MMSE estimates provided by P-MALA reach reconstruction errors of $\text{SNR}_{\mathbf{y}} = 29.3$dB and $\text{SNR}_{\mathbf{y}} = 19.3$dB, respectively. The mMAP estimate obtained using the full Gibbs sampler provides unsatisfying results compared to P-MALA, with low $\text{SNR}_{\mathbf{y}}$ and PAPR similar to P-MALA. This behaviour will be investigated in the next paragraph. Compared to our Bayesian algorithms, FITRA-mmse, recovers solutions with lower PAPR for a given $\text{SNR}_{\mathbf{y}}$. Both PFA-mmse and PFA-snr recover solutions comparable to FITRA-snr. All three algorithms MCMC, PFA and FITRA have provided anti-sparse representations with lower PAPR than LS or ℓ_2-penalized solutions, which confirms the interest of the democratic prior or, equivalently, the ℓ_∞-penalization.

Fig. 6 displays the results for all realizations of the measurement vector $\mathbf{y}$ where the $\text{SNR}_{\mathbf{y}}$ is plotted as a function of PAPR. To provide a whole characterization of FITRA and illustrate the trade-off between the expected reconstruction error and anti-sparsity level, the solutions provided by FITRA

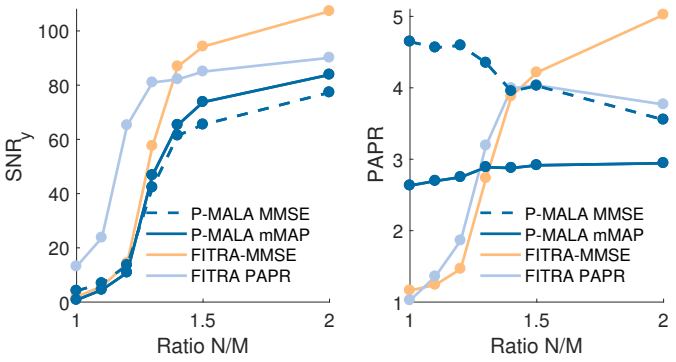

Fig. 7. Scenario 2: $\text{SNR}_{\mathbf{y}}$ (left) and PAPR (right) as a function of N/M.

corresponding to a wide range of regularization parameter β are shown[6]. The mMAP and MMSE solutions recovered by the two versions of the proposed algorithm are also reported in this $\text{SNR}_{\mathbf{y}}$ vs. PAPR plot. Note that the the critical region refers to the area around the phase transition of the blue line, *i.e.*, where FITRA abruptly moves from solutions with low PAPR and $\text{SNR}_{\mathbf{y}}$ to solutions with high PAPR and $\text{SNR}_{\mathbf{y}}$. The P-MALA estimates are located close to the critical region between solutions with low PAPR and $\text{SNR}_{\mathbf{y}}$ and solutions with high PAPR and $\text{SNR}_{\mathbf{y}}$: the proposed method recovers relevant solutions in a quasi non parametric way. However, Gibbs estimates reach either solutions with very high $\text{SNR}_{\mathbf{y}}$ or solution with very low $\text{SNR}_{\mathbf{y}}$. This means that for several runs, the Gibbs sampler has been stuck in a local maximum of (32) and more iterations are required for the Gibbs sampler to escape these local maximizers. This betrays a relatively unstable behaviour therefore less robust. Moreover, the Gibbs sampler does not scale to high dimensions due to its prohibitive computational cost.

Scenario 2 permits to evaluate the performances of the algorithms as a function of the ratio N/M. For measurement vectors of fixed dimension $M = 128$, the anti-sparse coding algorithms aim at recovering representation vectors of increasing dimensions $N = 128, \ldots, 256$. As it has been empirically shown in [29] for randomly subsampled DCT matrices, the $\text{SNR}_{\mathbf{y}}$ is expected to be an increasing function of N/M for a given PAPR level of anti-sparsity. The rational is that the number of combinations of $\pm\|\boldsymbol{y}\|_2/N$ increases when N grows, leading to more chance of finding a better representation for a given PAPR, due to increased redundancy. Fig. 7 shows the evolution of both $\text{SNR}_{\mathbf{y}}$ and PAPR as a function of the ratio N/M. For the two estimates from P-MALA, the $\text{SNR}_{\mathbf{y}}$ is increasing with the ratio, while the PAPR is constant for the mMAP estimate and (slightly) decreasing for the MMSE. As in the previous scenario, all versions of FITRA exhibit better $\text{SNR}_{\mathbf{y}}$. However, it is noticeable on Fig. 7 (right) that both FITRA-mmse and FITRA-snr provide higher PAPR when the ratio N/M increases beyond 1.3.

[6]Note that $\text{SNR}_{\mathbf{y}} = 0$ for PAPR $= 1$ with FITRA since for large value of β, the proximity operator given by (39) tends to the null vector.

V. Conclusion

This paper introduced a fully Bayesian framework for anti-sparse coding of a given measurement vector on a known and potentially over-complete dictionary. To derive a Bayesian formulation of the problem, a new probability distribution was introduced. Various properties of this so-called *democratic distribution* were exhibited, which permitted to design an exact random variate generator as well as two MCMC-based methods. This distribution was used as a prior for the representation vector in a linear Gaussian model, a probabilistic version of the anti-sparse coding problem. The residual variance as well as the anti-sparsity level were included into a fully Bayesian model and estimated jointly with the anti-sparse code. A Gibbs sampler was derived to generate samples distributed according to the joint posterior distribution of the coefficients of representation, the residual variance and the anti-sparse level. A second sampler was also proposed to scale to higher dimensions. To this purpose, the proximity mapping of the ℓ_∞-norm was considered to design a P-MALA within Gibbs algorithm. The generated samples were used to approximate two Bayesian estimators of the representation vector, namely the MMSE and mMAP estimators.

The validity of the proposed algorithms was assessed and evaluated through various experiments, and compared with FITRA a variational counterpart of the proposed algorithms. They produced solutions comparable to FITRA in terms of reconstruction error and PAPR, with the noticeable advantage to be fully unsupervised. In all experiments, as expected, the democratic prior distribution, was able to promote anti-sparse solutions of the coding problem. For that specific task, the mMAP estimator generally provided more relevant solutions than the MMSE estimator. Moreover, the P-MALA-based algorithm seemed to be more robust than the full Gibbs sampler and had the ability to scale to higher dimensions, both in term of computational times and performances.

Future works include the unsupervised estimation of the coding matrix jointly with the sparse code. This would open the door to the design of encoding matrices that would ensure equal spreading of the information over their atoms. Furthermore, since the P-MALA based sampler showed promising results, it would be relevant to investigate the geometric ergodicity of the chain. Unlike most of the illustrative examples considered in [35], this property can not be easily stated for the democratic distribution since it is not continuously differentiable, but only continuous. Finally, it could be interesting to investigate practical applications such as exact PAPR oriented scheme [43]. One may consider for instance a prior on the coding vector that takes values in $\{-\alpha, +\alpha\}$, with an hyperprior on α. Such a model would be appropriate for binarization, as considered in [32] for approximate nearest neighbor search.

Acknowledgments

The authors would like to thank the reviewers and the Associate Editor for their constructive comments on this article. They are also grateful to Dr. Vincent Mazet, Université de Strasbourg, France, for providing the code to draw according to truncated Gaussian distributions following [42].

APPENDIX A
PROPERNESS OF THE DEMOCRATIC DISTRIBUTION

The parity of f, as well as its symmetries w.r.t. the set of cones $\mathcal{C}_n$ introduced in (11) lead to, by noting $\mathcal{C}_n^+$ the half positive part of $\mathcal{C}_n$

$$\begin{aligned} C_N(\lambda) &= 2^N \int_{\cup_{n=1}^N \mathcal{C}_n^+} \exp\left(-\lambda \left\|\mathbf{x}\right\|_\infty\right) \mathrm{d}\mathbf{x} \\ &= 2^N N \int_{\mathcal{C}_1^+} \exp\left(-\lambda x_1\right) \mathrm{d}\mathbf{x} \\ &= 2^N N \int_{\mathbb{R}_+} x_1^{N-1} \exp\left(-\lambda x_1\right) \mathrm{d}x_1 = \frac{2^N N!}{\lambda^N}. \end{aligned}$$

APPENDIX B
MARGINAL DISTRIBUTIONS

Lemma 2 is proved by induction. To that aim, let one consider the assertion, indexed by J and denoted $\mathcal{P}(J)$: *"For any J-element subset $\mathcal{K}_J$ of $\{1,\dots,N\}$, the marginal distribution given in Lemma 2 holds"*.

a) initialization: for $J = 0$ the marginal distribution is nothing more the pdf of the democratic distribution.

b) inductive step : let J be an integer of $\{0,\dots,N-1\}$, and suppose $\mathcal{P}(J)$ is true. Let k be any integer of $\{0,\dots,N\}\setminus\mathcal{K}_J$ and consider the set $\mathcal{K}_{J+1} = \mathcal{K}_J \cup \{k\}$. Since $\mathcal{P}(J)$ holds, the marginal distribution $\mathrm{p}\left(\mathbf{x}_{\setminus\mathcal{K}_{J+1}}\right)$ can be computed as follows

$$\begin{aligned} \mathrm{p}\left(\mathbf{x}_{\setminus\mathcal{K}_{J+1}}\right) =& 2\int_{\mathbb{R}_+} \frac{2^J}{C_N(\lambda)} \sum_{j=0}^{J} \binom{J}{j} \frac{(J-j)!}{\lambda^{J-j}} \left\|\mathbf{x}_{\setminus\mathcal{K}_J}\right\|_\infty^j \\ &\times \exp\left(-\lambda \left\|\mathbf{x}_{\setminus\mathcal{K}_J}\right\|_\infty\right) \mathrm{d}x_k. \end{aligned}$$

Inverting integral and series leads to the integration of $J+1$ similar functions. Partitioning $\mathbb{R}_+$ as $\left[0, \left\|\mathbf{x}_{\setminus\mathcal{K}_{J+1}}\right\|_\infty\right) \cup \left[\left\|\mathbf{x}_{\setminus\mathcal{K}_{J+1}}\right\|_\infty, +\infty\right)$ allows one to rewrite the ℓ_∞-norm in terms of either $\left\|\mathbf{x}_{\setminus\mathcal{K}_{J+1}}\right\|_\infty$ or x_k. Thus

$$\begin{aligned} \mathrm{p}\left(\mathbf{x}_{\setminus\mathcal{K}_{J+1}}\right) =& \frac{2^{J+1}}{C_N(\lambda)} e^{-\lambda\left\|\mathbf{x}_{\setminus\mathcal{K}_{J+1}}\right\|_\infty} \\ &\times \sum_{j=0}^{J} \binom{J}{j} \frac{(J-j)!}{\lambda^{J-j}} \left(\left\|\mathbf{x}_{\setminus\mathcal{K}_{J+1}}\right\|_\infty^{j+1} \right. \\ &\left. + \sum_{l=1}^{j+1} \frac{j!}{(j+1-l)!} \frac{1}{\lambda^l} \left\|\mathbf{x}_{\setminus\mathcal{K}_{J+1}}\right\|_\infty^{j+1-l} \right). \end{aligned}$$

The last step consists in gathering all the terms of same degree in the polynomial function of $\left\|\mathbf{x}_{\setminus\mathcal{K}_{J+1}}\right\|_\infty$.

Degree $J+1$: the only term of degree $J+1$ is for $j = J$

$$\binom{J}{J} \frac{(J-J)!}{\lambda^{J-J}} = 1 = \binom{J+1}{J+1} \frac{(J+1-J-1)!}{\lambda^{J+1-J-1}}. \quad (51)$$

Degree 0: the term of degree 0 appears for all values of j where $l = j+1$. Thus

$$\begin{aligned} \sum_{j=0}^{J} \binom{J}{j} \frac{(J-j)!}{\lambda^{J-j}} \times \frac{j!}{\lambda^{j+1}} =& \frac{1}{\lambda^{J+1}} \sum_{j=0}^{J} \binom{J}{j} (J-j)!j! \\ =& \binom{J+1}{0} \frac{(J+1)!}{\lambda^{J+1}}. \end{aligned} \quad (52)$$

Degree $0 < p < J+1$: this term comes from all $j \geq p$ and $l = j+1-p$

$$\begin{aligned} &\binom{J}{p-1} \frac{(J-p+1)!}{\lambda^{J-p+1}} + \sum_{j=p}^{J} \binom{J}{j} \frac{(J-j)!}{\lambda^{J-j}} \frac{j!}{p!} \frac{1}{\lambda^{j+1-p}} \\ =& \frac{(J+1-p)!}{\lambda^{J+1-p}} \left(\binom{J}{p-1} + \binom{J}{p} \right) \\ =& \frac{(J+1-p)!}{\lambda^{J+1-p}} \binom{J+1}{p}. \end{aligned} \quad (53)$$

Hence $\mathcal{P}(J+1)$ is true and the assertion is proven by induction.

APPENDIX C
CONDITIONAL DISTRIBUTIONS

A. Conditional distributions

a) Equation (13) is obtained by means of a Bayes rule, $\mathrm{p}\left(x_n|\mathbf{x}\in\mathcal{C}_n\right) = \frac{\mathrm{p}(x_n,\, \mathbf{x}\in\mathcal{C}_n)}{\mathrm{P}[\mathbf{x}\in\mathcal{C}_n]}$, where $\mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n\right]$ is given in Property 3. By marginalizing over all other variables, one has

$$\begin{aligned} \mathrm{p}\left(x_n|\mathbf{x}\in\mathcal{C}_n\right) &= \frac{N2^{N-1}}{C_N(\lambda)} \int_0^{|x_n|} \exp\left(-\lambda\left|x_n\right|\right) \mathrm{d}\, x_{i\neq n} \\ &= \frac{\lambda^N}{2(N-1)!} \left|x_n\right|^{N-1} \exp\left(-\lambda\left|x_n\right|\right). \end{aligned}$$

b) Equation (14) results from the Bayes rule followed by a marginalization over x_n

$$\begin{aligned} \mathrm{p}\left(\mathbf{x}_{\setminus n}|\mathbf{x}\in\mathcal{C}_n\right) &= \frac{2N}{C_N(\lambda)} \int_{\left\|x_{\setminus n}\right\|_\infty}^{+\infty} \exp\left(-\lambda\left|x_n\right|\right) \mathrm{d}\, x_n \\ &= \frac{1}{C_{N-1}(\lambda)} \exp\left(-\lambda\left\|x_{\setminus n}\right\|_\infty\right). \end{aligned}$$

c) Equation (15) states that conditionally to a cone and the value of the dominant component, the non-dominant components are $(N-1)$ i.i.d uniform random variables. First

$$\mathrm{p}\left(\mathbf{x}_{\setminus n}|x_n, \mathbf{x}\in\mathcal{C}_n\right) = \frac{\mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n|\mathbf{x}\right] \mathrm{p}\left(\mathbf{x}\right)}{\mathrm{p}\left(x_n|\mathbf{x}\in\mathcal{C}_n\right)\ \mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n\right]}$$

where $\mathrm{p}\left(x_n|\mathbf{x}\in\mathcal{C}_n\right)$ and $\mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n\right]$ are respectively given by equations (13) and (3). Note that $\mathrm{p}\left(\mathbf{x}\right)$ is the democratic pdf. Finally, $\mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n|\mathbf{x}\right]$ is the indicator function of the set $\mathcal{C}_n$ or, in other words, $\mathbb{1}_{\forall j\neq n,\ |x_j|\leq|x_n|}(.)$. Thus

$$\begin{aligned} &\mathrm{p}\big(\mathbf{x}_{\setminus n}|x_n, \mathbf{x}\in\mathcal{C}_n\big) \\ &\quad= \frac{2\,\Gamma(N)}{\lambda^N \left|x_n\right|^{N-1} e^{-\lambda|x_n|}} \frac{Ne^{-\lambda\|\mathbf{x}\|_\infty}}{C_N(\lambda)} \mathbb{1}_{\substack{|x_j|\leq|x_n| \\ \forall j\neq n}}(\mathbf{x}) \\ &\quad= \frac{1}{2^{N-1}\left|x_n\right|^{N-1}} \mathbb{1}_{\substack{|x_j|\leq|x_n| \\ \forall j\neq n}}(\mathbf{x}). \end{aligned} \quad (54)$$

d) Equation (16) results from the Bayes rule

$$\begin{aligned} \mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n|\mathbf{x}_{\setminus n}\right] =& \frac{1}{\mathrm{p}\left(\mathbf{x}_{\setminus n}\right)} \times \int_{\left\|x_{\setminus n}\right\|_\infty}^{\infty} \mathrm{p}\left(\mathbf{x}|\mathbf{x}\sim\mathcal{D}_N(\lambda)\right) \mathrm{d}x_n \\ =& \frac{1}{\mathrm{p}\left(\mathbf{x}_{\setminus n}\right)} \times \frac{\lambda^{N-1}}{2^{N-1}N!} \exp\left(-\lambda\left\|x_{\setminus n}\right\|_\infty\right) \end{aligned}$$

where the marginal distribution $\mathrm{p}\left(\mathbf{x}_{\setminus n}\right)$ has been derived in (10). Once plugged, the computation directly leads to (16).

e) Equation (17) is computed as previously

$$\begin{aligned}\mathrm{p}\left(\mathbf{x}_{\backslash n}|\mathbf{x}\notin\mathcal{C}_n\right) &= \frac{2}{\mathrm{P}\left[\mathbf{x}\notin\mathcal{C}_n\right]}\int_0^{\left\|x_{\backslash n}\right\|_\infty}\frac{\lambda^N}{2^N N!}e^{-\lambda\|x\|_\infty}\,\mathrm{d}\,x_n\\ &= \frac{1}{\mathrm{P}\left[\mathbf{x}\notin\mathcal{C}_n\right]}\frac{\lambda^N}{2^{N-1}N!}\left\|x_{\backslash n}\right\|_\infty e^{-\lambda\left\|x_{\backslash n}\right\|_\infty}.\end{aligned}$$

Then, $\mathrm{P}\left[\mathbf{x}\notin\mathcal{C}_n\right]=1-\mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n\right]=\frac{N-1}{N}$ using Property 3.

B. Full conditional distributions

This appendix describes how to compute the conditional distribution $\mathrm{p}\left(x_n|\mathbf{x}_{\backslash n}\right)$. The proposed strategy consists in conditioning the probability by the event $x_n\in\mathcal{C}_n$. Hence

$$\begin{aligned}\mathrm{p}\left(x_n|\mathbf{x}_{\backslash n}\right) = \mathrm{p}\left(x_n|\mathbf{x}_{\backslash n},\mathbf{x}\in\mathcal{C}_n\right)\ \mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n|\mathbf{x}_{\backslash n}\right]\\ +\ \mathrm{p}\left(x_n|\mathbf{x}_{\backslash n},\mathbf{x}\notin\mathcal{C}_n\right)\ \mathrm{P}\left[\mathbf{x}\notin\mathcal{C}_n|\mathbf{x}_{\backslash n}\right].\end{aligned}\tag{55}$$

a) $\mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n|\mathbf{x}_{\backslash n}\right]$ is given by equation (16). Thus, $\mathrm{P}\left[\mathbf{x}\notin\mathcal{C}_n|\mathbf{x}_{\backslash n}\right]$ follows.

b) $\mathrm{p}\left(x_n|\mathbf{x}_{\backslash n},\mathbf{x}\in\mathcal{C}_n\right)$ can be computed using two nested Bayes rules

$$\mathrm{p}\left(\mathbf{x}_n|\mathbf{x}_{\backslash n},\mathbf{x}\in\mathcal{C}_n\right)=\frac{\mathrm{p}\left(\mathbf{x}\right)}{\mathrm{p}\left(\mathbf{x}_{\backslash n}|\mathbf{x}\in\mathcal{C}_n\right)\mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n\right]}\mathbf{1}_{\mathcal{C}_n}(\mathbf{x})\tag{56}$$

where $\mathrm{p}\left(\mathbf{x}_{\backslash n},x_n\right)$ is the pdf of the democratic distribution. Since $\mathbf{x}$ belongs to the cone $\mathcal{C}_n$, $\|\mathbf{x}\|_\infty$ can be replaced by $|x_n|$. $\mathrm{P}\left[\mathbf{x}\in\mathcal{C}_n\right]$ is given in (12). Then

$$\begin{aligned}\mathrm{p}\big(\mathbf{x}_{\backslash n}|\mathbf{x}\in\mathcal{C}_n\big) &= 2N\int_{\left\|x_{\backslash n}\right\|_\infty}^{+\infty}\frac{\lambda^N}{2^N N!}\exp\left(-\lambda\left|x_n\right|\right)\mathrm{d}\,x_n\\ &= \frac{\lambda^{(N-1)}}{2^{(N-1)}(N-1)!}\exp\left(-\lambda\left\|\mathbf{x}_{\backslash n}\right\|_\infty\right).\end{aligned}\tag{57}$$

Combining (3), (3) and (57) in (56) leads to

$$\mathrm{p}\left(x_n|\mathbf{x}_{\backslash n},\mathbf{x}\in\mathcal{C}_n\right)=\frac{\lambda}{2}e^{-\lambda(|x_n|-\left\|\mathbf{x}_{\backslash n}\right\|_\infty)}\,\mathbf{1}_{|x_n|\geq\left\|\mathbf{x}_{\backslash n}\right\|_\infty}(x_n).\tag{58}$$

c) Calculations are the same as in the previous paragraph

$$\mathrm{p}\big(x_{\backslash n}|\mathbf{x}\notin\mathcal{C}_n\big)=\frac{\lambda\left\|x_{\backslash n}\right\|_\infty}{N-1}\frac{\lambda^{(N-1)}}{2^{(N-1)}(N-1)!}e^{-\lambda\left\|\mathbf{x}_{\backslash n}\right\|_\infty}.$$

Thus

$$\mathrm{p}\left(x_n|\mathbf{x}_{\backslash n},\mathbf{x}\notin\mathcal{C}_n\right)=\frac{1}{2\left\|\mathbf{x}_{\backslash n}\right\|_\infty}\,\mathbf{1}_{|x_n|\leq\left\|\mathbf{x}_{\backslash n}\right\|_\infty}(x_n).\tag{59}$$

Finally the conditional (18) is obtained by combining (12), (58) and (59) as suggested in (55).

APPENDIX D
POSTERIOR DISTRIBUTION OF THE COEFFICIENTS

The parameters of the truncated Gaussian distributions involved in the mixture distribution (36) are given by

$$\begin{aligned}\mu_{1n}&=\frac{1}{\|\mathbf{h}_n\|^2}\left(\mathbf{h}_n^T\mathbf{e}_n+\sigma^2\lambda\right)\\ \mu_{2n}&=\frac{1}{\|\mathbf{h}_n\|^2}\left(\mathbf{h}_n^T\mathbf{e}_n\right) \qquad\qquad s_n^2=\frac{\sigma^2}{\|\mathbf{h}_n\|_2^2}\\ \mu_{3n}&=\frac{1}{\|\mathbf{h}_n\|^2}\left(\mathbf{h}_n^T\mathbf{e}_n-\sigma^2\lambda\right)\end{aligned}$$

where $\mathbf{h}_i$ denotes the ith column of $\mathbf{H}$ and $\mathbf{e}_n=\mathbf{y}-\sum_{i\neq n}x_i\mathbf{h}_i$. Moreover, the weights associated with each mixture component are

$$\omega_{in}=\frac{u_{in}}{\sum_{j=1}^3 u_{jn}}\tag{60}$$

with

$$\begin{aligned}u_{1n}&=\exp\left(\frac{\mu_{1n}^2}{2s_n^2}+\lambda\left\|\mathbf{x}_{\backslash n}\right\|_\infty\right)\phi_{\mu_{1n},s_n^2}\left(-\left\|\mathbf{x}_{\backslash n}\right\|_\infty\right)\\ u_{2n}&=\exp\left(\frac{\mu_{2n}^2}{2s_n^2}\right)\left[\phi_{\mu_{2n},s_n^2}\left(\left\|\mathbf{x}_{\backslash n}\right\|_\infty\right)-\phi_{\mu_{2n},s_n^2}\left(-\left\|\mathbf{x}_{\backslash n}\right\|_\infty\right)\right]\\ u_{3n}&=\exp\left(\frac{\mu_{3n}^2}{2s_n^2}+\lambda\left\|\mathbf{x}_{\backslash n}\right\|_\infty\right)\times\left(1-\phi_{\mu_{3n},s_n^2}\left(\left\|\mathbf{x}_{\backslash n}\right\|_\infty\right)\right)\end{aligned}$$

where $\phi_{\mu,s^2}(\cdot)$ is the cumulated distribution function of the normal distribution $\mathcal{N}(\mu,s^2)$.

REFERENCES

[1] C. Elvira, P. Chainais, and N. Dobigeon, "Democratic prior for anti-sparse coding," in *Proc. IEEE-SP Workshop Stat. and Signal Process. (SSP)*, Palma de Mallorca, Spain, Jun. 2016.

[2] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[3] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, "Introduction to compressed sensing," in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge, UK: Cambridge University Press, 2012, ch. 1, pp. 1–64.

[4] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118–121, Jul. 2007.

[5] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, Mar. 2011.

[6] R. Gribonval, "Should penalized least squares regression be interpreted as maximum a posteriori estimation?" *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2405–2410, May 2011.

[7] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[8] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 53–63, Jan. 2010.

[9] F. Caron and A. Doucet, "Sparse Bayesian nonparametric regression," in *Proc. Int. Conf. Machine Learning (ICML)*, Helsinki, Finland, Jul. 2008.

[10] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learning Research*, vol. 1, pp. 211–244, 2001.

[11] A. Lee, F. Caron, A. Doucet, and C. Holmes, "A hierarchical Bayesian framework for constructing sparsity-inducing priors," *arXiv.org*, Sep. 2010.

[12] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "A novel hierarchical Bayesian approach for sparse semisupervised hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 585–599, Feb. 2012.

[13] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 2008.

[14] D. Tzikas, A. Likas, and N. Galatsanos, "Variational Bayesian sparse kernel-based blind image deconvolution with Student's-t priors," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 753–764, Apr. 2009.

[15] C. Févotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2174–2188, Nov. 2006.

[16] N. Dobigeon, A. O. Hero, and J.-Y. Tourneret, "Hierarchical Bayesian sparse image reconstruction with application to MRFM," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 2059–2070, Sep. 2009.

[17] C. Soussen, J. Idier, D. Brie, and J. Duan, "From Bernoulli-Gaussian deconvolution to sparse signal restoration," *IEEE Trans. Signal Process.*, vol. 29, no. 10, pp. 4572–4584, Oct. 2011.

[18] L. Chaari, H. Batatia, N. Dobigeon, and J.-Y. Tourneret, "A hierarchical sparsity-smoothness Bayesian model for $\ell_0-\ell_1-\ell_2$ regularization," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 1901–1905.

[19] E. Remes, "Sur une propriété extrémale des polynômes de Tchebychef," *Communications de l'Institut des Sciences Mathématiques et Mécaniques de l'Université de Kharkoff et de la Société Mathématique de Kharkoff*, vol. 13, no. 1, pp. 93–95, 1936.

[20] T. W. Parks and J. H. McClellan, "Chebyshev approximation for nonrecursive digital filters with linear phase," *IEEE Trans. Circ. Theory*, vol. 19, no. 2, pp. 189–194, Mar. 1972.

[21] J. H. McClellan and T. W. Parks, "A personal history of the Parks-McClellan algorithm," *IEEE Signal Process. Mag.*, vol. 22, no. 2, pp. 82–86, Mar. 2005.

[22] L. W. Neustadt, "Minimum effort control systems," *J. SIAM Control*, vol. 1, no. 1, pp. 16–31, 1962.

[23] J. A. Cadzow, "Algorithm for the minimum-effort problem," *IEEE Trans. Autom. Contr.*, vol. 16, no. 1, pp. 60–63, 1971.

[24] Y. Lyubarskii and R. Vershynin, "Uncertainty principles and vector quantization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3491–3501, Jul. 2010.

[25] Z. Cvetković, "Resilience properties of redundant expansions under additive noise and quantization," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 644–656, Mar. 2003.

[26] A. R. Calderbank and I. Daubechies, "The pros and cons of democracy," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1721–1725, Jun. 2002.

[27] B. Farrell and P. Jung, "A Kashin approach to the capacity of the discrete amplitude constrained Gaussian channel," in *Proc. Int. Conf. Sampling Theory and Applications (SAMPTA)*, Marseille, France, May 2009.

[28] J. Ilic and T. Strohmer, "PAPR reduction in OFDM using Kashin's representation," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Comm.*, Perugia, Italy, 2009, pp. 444–448.

[29] C. Studer, Y. Wotao, and R. G. Baraniuk, "Signal representations with minimum ℓ_∞-norm," in *Proc. Ann. Allerton Conf. Comm. Control Comput. (Allerton)*, 2012, pp. 1270–1277.

[30] C. Studer, T. Goldstein, W. Yin, and R. G. Baraniuk, "Democratic representations," *IEEE Trans. Inf. Theory*, 2014. [Online]. Available: http://arxiv.org/abs/1401.3420/

[31] J.-J. Fuchs, "Spread representations," in *Proc. IEEE Asilomar Conf. Signals, Systems, Computers*, 2011.

[32] H. Jegou, T. Furon, and J.-J. Fuchs, "Anti-sparse coding for approximate nearest neighbor search," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, 2012, pp. 2029–2032.

[33] C. Studer and E. G. Larsson, "PAR-aware large-scale multi-user MIMO-OFDM downlink," *IEEE J. Sel. Areas Comm.*, vol. 31, no. 2, pp. 303–313, Feb. 2013.

[34] J. Tan, D. Baron, and L. Dai, "Wiener filters in Gaussian mixture signal estimation with ℓ_∞-norm error," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6626–6635, Oct. 2014.

[35] M. Pereyra, "Proximal markov chain monte carlo algorithms," *Statistics and Computing*, vol. 26, no. 4, pp. 745–760, 2016.

[36] C. Elvira, P. Chainais, and N. Dobigeon, "Bayesian anti-sparse coding – Complementary results and supporting materials," University of Toulouse, IRIT/INP-ENSEEIHT, and Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, France, Tech. Rep., Nov. 2016. [Online]. Available: http://pierrechainais.ec-lille.fr/PUB/ANTISPARSE/supp_mat_democratic.pdf

[37] P. Kabal, "Quantizers for the gamma distribution and other symmetrical distributions," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, vol. 32, no. 4, pp. 836–841, Aug. 1984.

[38] J.-J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *CR Acad. Sci. Paris Sér. A Math*, vol. 255, pp. 2897–2899, 1962.

[39] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, Jan. 2014.

[40] L. Condat, "Fast projection onto the simplex and the l1 ball," *Mathematical Programming*, pp. 1–11, 2015.

[41] J. Geweke, "Getting it right: Joint distribution tests of posterior simulators," *J. Amer. Stat. Assoc.*, vol. 99, pp. 799–804, Sep. 2004.

[42] N. Chopin, "Fast simulation of truncated Gaussian distributions," *Statistics and Computing*, vol. 21, no. 2, pp. 275–288, 2010.

[43] R. Balu, T. Furon, and H. Jégou, "Beyond "project and sign" for cosine estimation with binary codes," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, May 2014, pp. 68 884–6888.

Clément Elvira received his Eng. degree from Centrale Lille, France, and the M.Sc. degree in applied Mathematics from the university of Lille, France, both in September 2014. He is currently pursuing his PhD studies at CRIStAL, Lille, France, under the supervision of Pierre Chainais and Nicolas Dobigeon. His research interests are centered around Bayesian nonparametrics and Markov Chain Monte Carlo (MCMC) methods with applications to signal processing.

Pierre Chainais received his Ph.D. in Physics in 2001 from the Ecole Normale Supérieure de Lyon (France). He joined the University Blaise Pascal at Clermont-Ferrand as an Assistant Professor in signal processing in 2002. He moved to Centrale Lille in 2011 where he currently is an Associate Professor in Signal Processing at CRIStAL Lab. His research interests lie in statistical signal processing (dictionary learning, Bayesian non parametrics...) and machine learning with applications to physical systems.

Nicolas Dobigeon (S'05–M'08–SM'13) was born in Angoulême, France, in 1981. He received the Eng. degree in electrical engineering from ENSEEIHT, Toulouse, France, and the M.Sc. degree in signal processing from INP Toulouse, both in 2004, the Ph.D. degree and the Habilitation à Diriger des Recherches in signal processing from INP Toulouse in 2007 and 2012, respectively.

From 2007 to 2008, he was a Post-Doctoral Research Associate with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor. Since 2008, he has been with INP-ENSEEIHT Toulouse, University of Toulouse, where he is currently a Professor. He conducts his research within the Signal and Communications Group, IRIT Laboratory, and he is also an Affiliated Faculty Member of the TeSA Laboratory. His recent research activities have been focused on statistical signal and image processing, with a particular interest in Bayesian inverse problems and applications to remote sensing and biomedical imaging.